



Universidad Cenfotec

Maestría en Tecnologías de Bases de Datos

Documento final de Proyecto de Investigación Aplicada 2

**Construcción de una Plataforma de Software para la
Proyección de la Demanda en el Sistema Eléctrico Nacional
de Costa Rica basada en una Solución de Inteligencia de
Negocios**

Arias Murillo Leonardo
Hidalgo Soto José Alberto

Enero, 2015

©2015, Arias Murillo Leonardo, Hidalgo Soto José Alberto

Se autoriza la reproducción total o parcial, con fines académicos, por cualquier medio o procedimiento, incluyendo la cita bibliográfica del documento.

A Dios por todas sus bendiciones.

A mi esposa Patricia, a mis hijas Melanie, Krisly, Mariel y a mis padres.

Leonardo

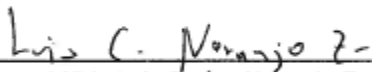
A mis padres y a los compañeros del Centro Nacional de Control de Energía

José Alberto

Agradecemos al MBA Luis Carlos Naranjo Z. por la guía brindada y sus oportunos comentarios en el desarrollo de este trabajo.

TRIBUNAL EXAMINADOR

Este proyecto fue aprobado por el Tribunal Examinador de la carrera: **Maestría en Tecnologías de Bases de Datos**, requisito para optar por el título de grado de **Maestría**.



MBA. Luis Carlos Naranjo Z.
Tutor



M. Sc. Ignacio Trejos Zelaya
Lector 1



M. Sc. Roberto Loaiza Álvarez
Lector 2

Contenido

Contenido	iv
Índice de figuras	vi
Índice de tablas	ix
Resumen ejecutivo	xi
1. Introducción	1
1.1. Antecedentes	1
1.2. Marco Institucional	3
1.3. Descripción del problema.....	8
1.4. Justificación.....	8
1.5. Objetivos	10
2. Estado de la cuestión	12
2.1. Data Marts y Modelos Multidimensionales	12
2.2. Proyección de la demanda	19
3. Marco Metodológico	38
3.1. Enfoque de la investigación	38
3.2. Diseño de la investigación.....	38
3.3. Fuentes de información.....	40
3.4. Descripción de los instrumentos	40
4. Desarrollo del data mart	41
4.1. Planificación del proyecto.....	41
4.2. Definición de requerimientos del negocio.....	47
4.3. Diseño de la arquitectura técnica	49
4.4. Selección de productos e implementación	51
4.5. Modelado dimensional.....	55

4.6. Diseño físico	65
4.7. Diseño e implementación del subsistema de ETL.....	71
4.8. Diseño de aplicaciones BI	78
4.9. Desarrollo de aplicaciones BI	83
4.10. Implementación	84
5. Proyección de la demanda eléctrica utilizando metodología CRISP-DM	86
5.1. Fase de comprensión del negocio o problema.....	86
5.2. Comprensión de los datos.....	97
5.3. Preparación de los datos	101
5.4. Modelado y evaluación	103
5.5. Implementación.....	112
Implementación de promedios móviles	112
Implementación de redes neuronales	117
6. Conclusiones	123
7. Recomendaciones.....	126
Glosario	128
Referencias	130
Apéndices.....	133
Apéndice 1. Lista de elementos claves del sistema.	133
Apéndice 2. Tablas del Modelado Dimensional.....	136
Apéndice 3. Detalle de las transformaciones.	142
Apéndice 4. Consulta Microsoft Time Series.....	162
Apéndice 5. Implementación de promedios móviles	163
Apéndice 6. Implementación de redes neuronales.....	165
Apéndice 7. Código MatLab para ejecutar la red neuronal	167

Índice de figuras

Figura 1. Procesos generales de un sistema eléctrico (elaboración propia).	4
Figura 2. Esquema de un sistema eléctrico regional, adaptación de (REE, 2014)	4
Figura 3. Organigrama del CENCE. (CENCE, 2014)	7
Figura 4. Data warehouse según Kimball (elaboración propia)	12
Figura 5. Data warehouse según Inmon (elaboración propia)	13
Figura 6. Diagrama de modelo lógico multidimensional (elaboración propia)	17
Figura 7. Proyección de la Demanda de Energía para el 12 de marzo de 2014 (elaboración propia)	22
Figura 8. Predespacho Horario de Energía para el 12 de marzo de 2014 (elaboración propia)	22
Figura 9. Modelo de pronóstico basado en promedios móviles (D'oro, Lozano, & Moreno, 2007)	29
Figura 10. Esquema de una neurona artificial (elaboración propia)	33
Figura 11. Red neuronal multicapa típica (elaboración propia)	34
Figura 12. Metodología de Kimball (Kimball, 2009)	41
Figura 13. Priorización según impacto y viabilidad (elaboración propia)	49
Figura 14. Diagrama de diseño de arquitectura técnica (elaboración propia)	51
Figura 15. Cuadrante mágico de Gartner para octubre, 2014. (Gartner, 2014)	53
Figura 16. Esquema de una solución típica de inteligencia de negocios utilizando SQL Server (elaboración propia)	54
Figura 17. Diagrama de entidad relación (elaboración propia)	57
Figura 18. Diagrama de alto nivel para predespacho (elaboración propia).	59

Figura 19. Diagrama de alto nivel para la potencia según SCADA (elaboración propia).	62
Figura 20. Diagrama de alto nivel para la energía diaria (elaboración propia).	62
Figura 21. Diagrama de alto nivel para la energía según SCADA (elaboración propia).	63
Figura 22. Diagrama de alto nivel para los niveles (elaboración propia). ...	63
Figura 23. Procedimiento para diseño físico según Kimball (elaboración propia).	65
Figura 24. Modelo físico de la base de datos (elaboración propia).	67
Figura 25. Diagrama del cubo (elaboración propia).	70
Figura 26. Plan de alto nivel (elaboración propia).	72
Figura 27. Carga de datos al área de trabajo (elaboración propia).	76
Figura 28. Carga de datos a las dimensiones (elaboración propia).	77
Figura 29. Carga de datos a las tablas de hechos (elaboración propia).	77
Figura 30. Bosquejo de la aplicación BI (elaboración propia).	82
Figura 31. Ejemplo de aplicación BI (elaboración propia).	84
Figura 32. Fases de la metodología CRISP-DM (IBM Corporation, 2012) .	86
Figura 33. Comportamiento semanal de la demanda (elaboración propia).	101
Figura 34. Distribución normal de los errores para la proyección de la demanda (elaboración propia).	102
Figura 35. Comparación del error PDE vs Promedios móviles (elaboración propia).	105
Figura 36. Comparación del error PDE vs Microsoft Time Series (elaboración propia).	109
Figura 37. Comparación del error PDE vs Microsoft Time Series (elaboración propia).	111
Figura 38. Gráfico de dispersión del método actual PDE (elaboración propia).	115

Figura 39. Gráfico de dispersión del método de promedios móviles (elaboración propia).	115
Figura 40. Comparación del método actual PDE vs Promedios Móviles (elaboración propia).	117
Figura 41. Red neuronal obtenida del entrenamiento (elaboración propia).	119
Figura 42. Gráfico de dispersión del método de redes neuronales (elaboración propia).	120
Figura 43. Gráfico comparativo de los métodos propuestos vs el actual (elaboración propia).	122

Índice de tablas

Tabla 1. Riesgos identificados (elaboración propia).....	44
Tabla 2. Lista de recursos (elaboración propia).....	46
Tabla 3. Costos estimados (elaboración propia).	46
Tabla 4. Plan del proyecto (elaboración propia).	47
Tabla 5 - Cumplimiento de requerimientos del producto para desarrollar solución de inteligencia de negocios (elaboración propia).	52
Tabla 6. Matriz de bus (elaboración propia).	57
Tabla 7 - Matriz de bus actualizada (elaboración propia).	64
Tabla 8 - Hoja de trabajo para la dimensión de empresa (elaboración propia).	64
Tabla 9 - Definición de índices para tabla de hechos (elaboración propia).	69
Tabla 10. Agregaciones (elaboración propia).	71
Tabla 11. Reporte candidatos (elaboración propia).....	81
Tabla 12. Lista de archivos Excel implementados (elaboración propia).	83
Tabla 13. Recursos identificados para el proceso de minería (elaboración propia).....	92
Tabla 14. Riesgos identificados para el proceso de minería (elaboración propia).....	94
Tabla 15. Costos estimados de minería de datos para pronóstico de demanda (elaboración propia).....	95
Tabla 16. Plan de proyecto (elaboración propia).	97
Tabla 17. Descripción de los campos a utilizar de TD_Fecha (elaboración propia).	99
Tabla 18. Descripción de los campos a utilizar de TD_Hora (elaboración propia).	99
Tabla 19. Descripción de los campos a utilizar de TH_Demanda (elaboración propia).	100
Tabla 20. Ordenes de los promedios móviles para las series de demanda horaria (elaboración propia).	104

Tabla 21. Comparación con promedios móviles (elaboración propia).	105
Tabla 22. Información del modelo Microsoft Time Series (elaboración propia).	107
Tabla 23. Parámetros del modelo (elaboración propia).	107
Tabla 24. Comparación con Microsoft Time Series (elaboración propia).	108
Tabla 25. Datos de entrada a la red neuronal (elaboración propia).	110
Tabla 26. . Comparación con Redes Neuronales (elaboración propia). ...	111
Tabla 27. Estructura de la tabla Promoviles (elaboración propia).	112
Tabla 28. Estructura de la tabla Demanda_PM (elaboración propia).	112
Tabla 29. Comparación de los resultados obtenidos con promedios móviles (elaboración propia).	116
Tabla 30. Entradas para la red neuronal (elaboración propia).	118
Tabla 31. Comparación de los resultados obtenidos con redes neuronales (elaboración propia).	121

Resumen ejecutivo

El presente trabajo tiene como objetivo la construcción de una plataforma de inteligencia de negocios que permita realizar tareas de análisis sobre el comportamiento del Sistema Eléctrico Nacional de Costa Rica.

Se desarrollará un data mart para el Centro Nacional de Control de Energía que contiene la información de la energía y potencia de las diferentes unidades generadoras que componen el Sistema Eléctrico Nacional.

Para aprovechar esta plataforma, se incluye el proceso de proyección de la demanda eléctrica nacional. Se presentará un estudio para comparar el cálculo de la proyección de la demanda eléctrica utilizando un modelo de redes neuronales, el modelo ARIMA (para series de tiempo) y otro, utilizando promedios móviles. La proyección se realizará en el corto plazo, es decir, un horizonte máximo de una semana.

Un depósito de datos potenciará a la organización para realizar análisis de datos de una manera mucho más ágil. Así se constituye la base para posteriores esfuerzos en las áreas de calidad de datos y minería de datos, también para el ciclo de descubrimiento que trae consigo la adopción de las tecnologías de inteligencia de negocios.

Por otra parte, la presentación de un mecanismo automatizado para la proyección de la demanda ayudará al Centro Nacional de Control de Energía para realizar sus labores con mayor eficacia y eficiencia, redundando en un beneficio para todos los habitantes de este país.

Palabras claves

Despacho, pronóstico de demanda, energía, potencia, data mart, electricidad, promedios móviles, redes neuronales.

1. Introducción

El proyecto para brindar a Costa Rica una herramienta de software para la proyección de la demanda en el sistema eléctrico nacional nace de la observación de las potenciales ventajas y beneficios que tendría su implementación. Sus repercusiones incluyen la mejora de procesos, la optimización de recursos y la reducción de tiempos de ejecución de tareas. Más aún, una proyección de la demanda eléctrica nacional con un menor margen de error representaría una mejora en el uso de recursos que son altamente valiosos para el país.

Se presentará un data mart aplicado a un área que no es común: la de un sistema eléctrico nacional. Además, valiéndose de este depósito de datos, se presentará la ejecución de tareas de minería de datos para la proyección de la demanda eléctrica nacional de corto plazo.

Dada la naturaleza del proyecto y su especificidad, la presentación de antecedentes y la descripción del entorno se convierten en necesidad y se describen a continuación.

1.1. Antecedentes

El Centro Nacional de Control de Energía (CENCE) tiene la responsabilidad de dirigir y coordinar la operación del Sistema Eléctrico Nacional de Costa Rica (SEN) y el Mercado Eléctrico Nacional. Su finalidad es satisfacer la demanda eléctrica del país en forma eficiente, transparente y ambientalmente sostenible.

Para poder atender estas necesidades, desde su creación en 1981, el CENCE cuenta con un sistema SCADA (Supervisory Control And Data Acquisition), el cual se ha estado renovando con los avances tecnológicos y de acuerdo con las necesidades del SEN.

Así, la obtención de datos del Centro de Control se realiza por medio de terminales remotas que están en las plantas generadoras de energía y subestaciones a lo largo de todo el territorio nacional. La comunicación entre estas terminales y el CENCE se da mediante enlaces de microonda y de fibra óptica, lo

cual, conjuntamente con los servidores que reciben y procesan esta información, constituyen el SCADA.

Además del sistema SCADA, se cuenta con un sistema Historiador que va almacenando la información de tiempo real en una base de datos especializada en series de tiempo con características de no-SQL, que optimiza el almacenamiento, por el gran volumen de datos que se maneja.

Finalmente, se cuenta con bases de datos relacionales que obtienen información del Historiador y de otras fuentes de datos, éstas constituyen los sistemas transaccionales y las fuentes de información que se publican para su consumo interno y externo al CENCE.

La mayoría de las consultas se realizan sobre estas bases de datos, pues en ellas se mantiene la información que ha sido identificada como la que normalmente requieren los clientes. En esta información se guardan los datos reales de generación de energía, potencia y las respectivas proyecciones.

A pesar de ser pocas variables, existen diversos análisis y estudios que se realizan sobre éstas, como la comparación de la generación de dos plantas arbitrarias, o la generación eólica contrastada con la térmica. En otras palabras, existe la necesidad de la utilización de cubos de datos, que potencien los estudios y así sobrepasar los reportes estáticos de los que se dispone en la actualidad.

Dentro del CENCE hay un proceso funcional (entiéndase área o departamento) que tiene, entre otras, la función de realizar proyecciones, las cuales pueden ser a corto y mediano plazo.

Cabe señalar que aquellas funciones del CENCE, que giran en torno a operar el sistema en tiempo real, sólo interesan las proyecciones en el corto y mediano plazo. El largo plazo es responsabilidad de otra dependencia ajena al CENCE, pero dentro del Grupo ICE.

Para las proyecciones energéticas de mediano plazo, mensuales y anuales, se cuenta con software especializado que satisface el requerimiento, como el SDDP (Despacho Hidrotérmico Estocástico con Restricciones de Red, por sus siglas en inglés). Las proyecciones de corto plazo, se realizan con el programa NCP (Nuevo Corto Plazo) y ambos simuladores de sistemas de potencia, no

realizan proyecciones de demanda. Así, para la proyección de la demanda, no se cuenta con software específico. Para satisfacer esta necesidad, actualmente, un grupo de ingenieros en sistemas de potencia hacen uso de Excel y de las bases de datos para realizar las proyecciones. No obstante, este proceso no está completamente sistematizado y algunos puntos quedan a criterio del experto o de quién lo realiza.

Existen diferentes opciones para enfrentar este problema, tal es el caso de las redes neuronales, las cuales han sido implementadas en países como España, Bolivia y Perú. Por otro lado, en Colombia por ejemplo, se está haciendo uso de una técnica basada en promedios móviles. Otro método se basa en series de tiempo para realizar las proyecciones. Cada uno tiene sus características propias para unos casos u otros. Por tanto, se deben evaluar para determinar cuál podría ser el modelo más apropiado para su aplicación al SEN (Mallo González., 2003; Tudela, 2011; Huerta, Quispe, Ramos, Fernández, & Molina, 2012; D'oro, Lozano, & Moreno, 2007; Carvajal, 2003).

El proyecto busca innovar incorporando algoritmos de proyección (redes neuronales, ARIMA o promedios móviles) con un data mart para resolver el problema de proyección de la demanda eléctrica a corto plazo, brindando una solución sistemática y de fácil empleo.

Otro aspecto innovador es que el resultado de la proyección, será una fuente adicional para el data mart, así existirá un ciclo de retroalimentación donde el sistema se irá actualizando con la aplicación de las mismas proyecciones.

1.2. Marco Institucional

El Centro Nacional Control de Energía es una entidad que forma parte de la corporación del Grupo ICE.

El Grupo ICE es la empresa estatal para el desarrollo y comercialización de energía y telecomunicaciones de Costa Rica.

Así se subdivide en dos sectores, a saber:

- Sector Electricidad
- Sector Telecomunicaciones

El presente proyecto tiene lugar en el sector electricidad, que a grandes rasgos se encarga de satisfacer la demanda de energía eléctrica que requiere el país, para lo cual, construye y opera los sistemas para generar, transportar y distribuir la energía en todo el país.

En la cadena del negocio de una empresa eléctrica, se observan la generalidad de sus procesos, como se muestra en la Figura 1:



Figura 1. Procesos generales de un sistema eléctrico (elaboración propia)

En primer plano, la energía se debe producir o generar, luego se debe transportar a lo largo del territorio hacia las subestaciones y finalmente, a partir de éstas, se distribuye al cliente.

Todo este proceso debe ser coordinado por el Centro de Control de Energía, que es el ente autorizado para la toma las decisiones de cómo se opera este proceso.

En la Figura 2 se presenta de forma esquemática ese proceso:

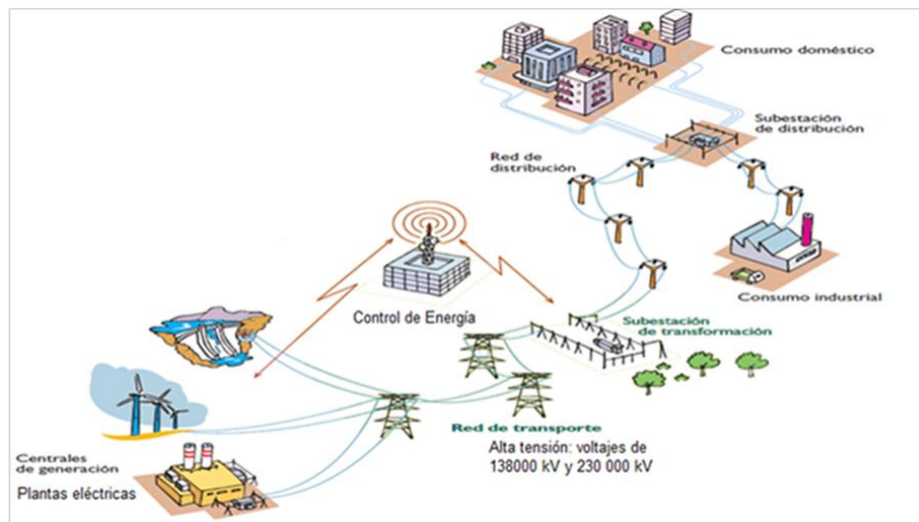


Figura 2. Esquema de un sistema eléctrico regional, adaptación de (REE, 2014)

En el caso del ICE es el Centro Nacional de Control de Energía (CENCE) quien tiene la responsabilidad de dirigir y coordinar la operación del sistema y el Mercado Eléctrico Nacional para satisfacer la demanda eléctrica del país.

El CENCE es el Operador del Sistema y el Operador de Mercado Nacional (OS&OM). En este sentido, le compete efectuar la asignación física de la oferta de energía eléctrica a la demanda nacional, es decir, optimizar el uso de la energía suministrada por generación ICE, generación privada, cooperativas y demás empresas eléctricas y si es necesario, efectuar transacciones regionales, que bien pueden ser de importación o exportación.

Cabe señalar que el ICE solo puede optimizar las plantas de generación propias, la generación privada, de cooperativas y demás empresas eléctricas se asume como generación base no optimizable, cuyas proyecciones de generación en los casos en que las entregan son indicativas y no vinculantes por legislación nacional, las variaciones que se produzcan de este grueso, son compensadas en tiempo real con generación ICE bajo los lineamientos de planeamiento operativo que ejecutan los operadores del CENCE. El CENCE define las transacciones técnico-económicas óptimas para satisfacer la demanda de electricidad de los usuarios, teniendo en cuentas los flujos de potencia en la red eléctrica y las pérdidas energéticas asociadas. Para ello, se realizan, entre otras actividades:

- Los pronósticos de la demanda diaria, mensual, semestral y anual.
- Inventario de los recursos disponibles.
- Supervisión, control y adquisición de datos.
- El control de las unidades generadoras, líneas, subestaciones e intercambio de electricidad.
- Análisis del comportamiento del sistema
- Estudios sobre seguridad operativa que respalden el despacho energético de una forma más confiable.

En años recientes y debido a la reorganización que tuvo el ICE se le agregó al CENCE un proceso o área para administrar la compra de energía a los generadores privados, esto agrega un sistema de medidores adicional para efectos comerciales, constituyéndose así el área de Comercializador Mayorista.

También, desde el año 2002, Centroamérica inició un proyecto de integración eléctrica, el Mercado Eléctrico Regional (MER), el cual entró en vigencia en junio del año 2013. Este mercado está compuesto por todos los países centroamericanos y se está en proceso de interconectar la región con México y Colombia, lo que ampliaría el mercado. Precisamente el CENCE es el representante de Costa Rica ante este mercado.

Misión y Visión

El propósito estratégico del CENCE es dirigir y coordinar la operación del sistema y el Mercado Eléctrico Nacional para satisfacer la demanda eléctrica del país en forma eficiente, transparente y ambientalmente sostenible, y participar con liderazgo en el desarrollo del mercado regional.

Futuro previsto

Ser el ente rector de la operación y del desarrollo del mercado nacional y líder regional por su competencia, tecnología y rentabilidad.

Principios rectores

- Planificar la operación del sistema. Desarrollar el recurso humano y las herramientas, para alcanzar la máxima capacidad de planificación; de la atención comercial del mercado y de la operación técnica del sistema.
- Supervisar el desempeño de los agentes del SEN y del MER. Se debe mantener un proceso integrador, óptimo de todo el SEN, para tomar decisiones oportunas y convenientes. Vigilar y controlar la calidad de los parámetros eléctricos tanto de los generadores privados nacionales, generadores ICE y regionales; así como los puntos de entrega a los distribuidores y grandes consumidores. Supervisar y garantizar el cumplimiento de los estándares de calidad y los límites de seguridad operativa.
- Satisfacer la demanda. Satisfacer las necesidades del cliente y partes interesadas que conllevan a la prestación del servicio que exige el mercado. El CENCE tiene que garantizar la energía que requieren los clientes, bajo las siguientes condiciones: continuidad en el servicio

eléctrico, calidad eléctrica, costo más bajo posible (óptimo), calidad en todos los servicios, atención e información que proporciona el CENCE.

- Realizar con excelencia la operación del sistema. Una de las razones fundamentales de la existencia del CENCE es la necesidad de garantizar el valor que le dan los clientes a las siguientes características del servicio eléctrico: continuidad en el servicio, calidad eléctrica, costo más bajo posible. Para lograr esto, es necesaria una optimización de recursos, un trabajo de calidad en equipo y un compromiso con lo que se hace.

Valores organizacionales

Los principales valores organizacionales son: la transparencia, la honestidad, el trabajo, la innovación, la rentabilidad y la gestión responsable de recursos energéticos y la excelencia en todas las actividades.

Organigrama

En la Figura 3 se muestra el organigrama del CENCE. Los recuadros amarillos del centro, muestran los cuatro procesos principales del CENCE, donde se debe resaltar el de Planeamiento y Despacho, uno de los principales clientes de este proyecto.

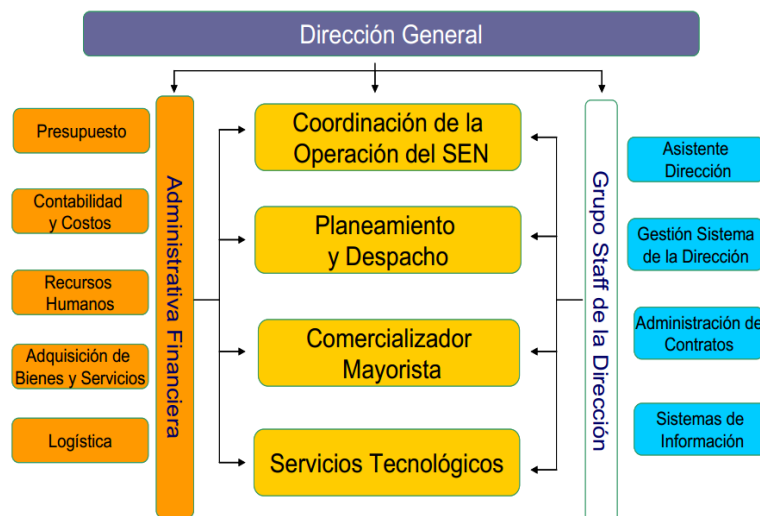


Figura 3. Organigrama del CENCE. (CENCE, 2014)

1.3. Descripción del problema

Normalmente la operación de un sistema eléctrico se fundamenta en la proyección de la demanda global. Es decir, que el objetivo del predespacho tiene como fin atender la demanda total en cada intervalo de tiempo. Si se cuenta con una herramienta para la proyección de la demanda eléctrica con un error suficientemente bajo, se logra obtener una maximización de los beneficios, reduciéndose así los costos e incrementando la estabilidad del sistema.

La proyección de la demanda de cada hora del SEN es una de las actividades necesarias y de importancia para la realización del predespacho nacional, y consiguientemente para el predespacho de las plantas que conforman el sistema. El área de planeamiento del CENACE, realiza la proyección diaria día con día, pero para mejorar la operación, se requiere aplicarlo de una manera más sistemática.

Resulta necesario realizar análisis sobre el comportamiento que tiene el sistema, para lo cual se cuenta con datos reales y el predespacho, para poder analizar las desviaciones de energía, o el impacto que pueden tener eventos acontecidos, como las fallas de elementos, apagones, disparos de carga, otros.

Para lograr estos análisis, se requiere contar con un depósito de datos y una plataforma de inteligencia de negocio que sirva de base para poder realizar las proyecciones y los análisis requeridos.

1.4. Justificación

Una solución de inteligencia de negocios constituye la base para poder realizar diversos análisis de la situación real. Particularmente, en la operación de un sistema eléctrico, estos análisis pasan a ser una tarea diaria, pues se revisa el comportamiento de las plantas: si sucedió algún evento, cómo este impactó en la operación, cómo se comportan los datos reales con respecto a lo estimado, el análisis de las desviaciones con respecto a los pronósticos, representan costos, por lo que se deben estudiar muy bien para minimizarlas. De la misma forma, existen múltiples análisis que se pueden realizar, por lo que una herramienta de inteligencia de negocios potenciará las tareas de estudio de los datos y flexibilizaría notablemente el acceso a los mismos.

Así, un depósito de datos constituiría, por su carácter histórico, la fuente primordial para la realización de proyecciones de demanda eléctrica, las cuales dentro del contexto de un sistema eléctrico son parte esencial para su operación.

Las proyecciones de la demanda eléctrica son utilizadas para (D'oro, Lozano, & Moreno, 2007):

- Programar y operar las unidades de generación diariamente.
- Reducir las desviaciones del despacho programado con respecto a la demanda real así como los redespachos, la necesidad de usar unidades de generación más costosas para suplir la demanda, como es el caso de las plantas térmicas, o por el contrario, no utilizar el total de la energía programada ya contratada. Esto permite reducir sobrecostos que repercuten en incrementos en el precio de la energía para los usuarios finales.
- Ayudar a estimar el flujo de carga y la realización de estudios de seguridad operativa. Esto finalmente ayuda a tomar decisiones que pueden prevenir las sobrecargas, evitando así eventos como la falla de equipos o apagones, que implicarían grandes pérdidas. Por tanto ayuda a mejorar la confiabilidad del sistema eléctrico.
- Establecer criterios para que los agentes generadores decidan sus ofertas de generación para el día siguiente y en los casos donde hay un mercado, se defina el precio de bolsa para la energía eléctrica.
- Evitar posibles penalizaciones por tener desviaciones en el pronóstico mayores a un margen establecido en la regulación de los mercados de energía, por ejemplo para el Mercado Eléctrico Regional de Centroamérica, las desviaciones en los intercambios de energía (exportación e importación) son penalizadas cuando son mayores a 4 MW.

La proyección de la demanda constituye la base para el predespacho de energía, el cual define cómo y cuánta energía aportará cada planta para satisfacer la demanda. De tal forma que si la proyección no es correcta, generalmente conlleva costos elevados. Por ejemplo, si la proyección fue menor a la demanda

real, el CENCE realiza un redespacho en el cual debe usar, en la mayoría de los casos, plantas o unidades de generación más costosas para atender la demanda.

Si la proyección fue mayor a la demanda real, se incurre en costos adicionales de operación pues se consideró el encendido de unidades extras, y generalmente serán las más costosas o se realizaron compras de energía que no se requerían.

Estas desviaciones en proyecciones con respecto a la realidad conducen a costos más elevados para el usuario final y disminuye la capacidad de competir con precios más bajos en el mercado de la energía.

Actualmente se realizan proyecciones de demanda diarias y semanales, ambas con resolución horaria, es decir, para las 24 horas siguientes y para las 168 horas de la semana, en este sentido, el aporte del presente trabajo es llevar a cabo la tarea de una forma menos rudimentaria, y brindar la plataforma para que permita evaluar estas proyecciones y afinarlas, e incluso, brindar un aporte que pueda ser utilizado en las reprogramaciones o redespachos.

1.5. Objetivos

Para la definición de los objetivos se utiliza la Taxonomía de Bloom, debido a que ésta es la más probada y aceptada por la academia.

Objetivo general

Construir una plataforma de inteligencia de negocios que soporte e incluya el proceso de proyección de la demanda nacional y el análisis del comportamiento del SEN.

Objetivos específicos

1. Describir el proceso de proyección de la demanda.
2. Definir el método de proyección de la demanda mediante el uso de redes neuronales u otro método, de acuerdo con los requerimientos identificados.
3. Estructurar un depósito de datos que satisfaga los requerimientos identificados para realizar la proyección de la demanda y los análisis del SEN.

4. Esbozar los mecanismos para que las instancias adecuadas resuelvan el proceso de proyección de la demanda mediante la implementación del algoritmo descrito.
5. Analizar los resultados obtenidos mediante un proceso de comparación de acuerdo con un plan de pruebas.

2. Estado de la cuestión

2.1. Data Marts y Modelos Multidimensionales

Data Mart

Para entender el significado del data mart podríamos utilizar el enfoque de Kimball (Kimball & Ross, 2002) e Inmon (Inmon, 2005).

Kimball señala que los departamentos y divisiones crearán sus propios data warehouses para responder a preguntas urgentes de negocios, así podríamos decir que los data marts son data warehouse departamentales construidos velozmente para brindar soluciones a un determinado negocio dentro de la institución, como por ejemplo, compras, ventas, inventario, otros. Luego, el data warehouse se construirá de la unión de todos estos data marts. Esquemáticamente este modelo se muestra en la Figura 4.

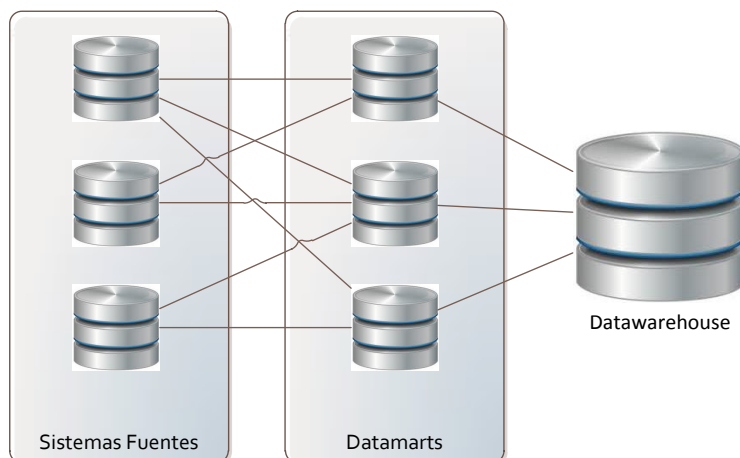


Figura 4. Data warehouse según Kimball (elaboración propia)

Ventajas:

- Más simple.
- Rápido de implementar.
- Específico a un tema.

Desventajas:

- Pueden presentarse duplicación de datos.
- Data marts podrían llegar a ser incompatibles.

Para subsanar estas desventajas se debe realizar una gobernanza de datos adecuada.

Por otro lado, Inmon parte del enfoque que los data marts se derivan a partir del data warehouse, como se aprecia en la Figura 5.

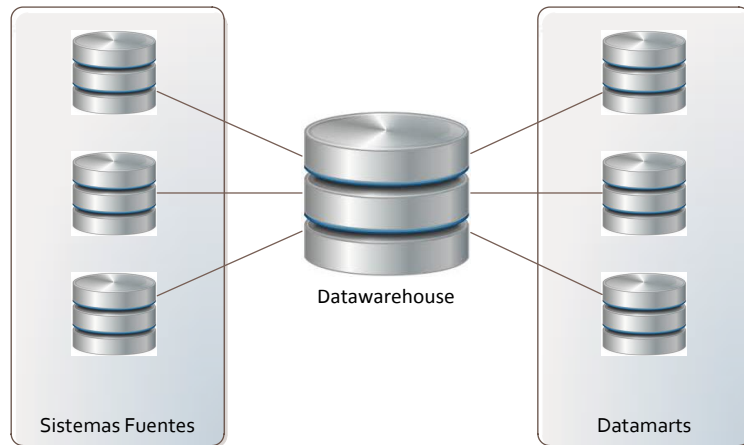


Figura 5. Data warehouse según Inmon (elaboración propia)

Ventajas:

- Fuentes comunes
- Procesamiento distribuido
- Desventajas:
- Mayor tiempo de desarrollo

Al igual que el data warehouse, los data marts tienen las mismas características de integración, no volatilidad y orientación temática (Inmon, 2005).

- Orientado temas: datos que brindan información sobre un “sujeto” del negocio en particular, en lugar de concentrarse en la dinámica de las transacciones de la organización. Aunque para el caso del data mart, el tema es aún más limitado.
- Integrados: los datos con los que se nutre vienen de diferentes fuentes y son integrados para dar una visión de un “todo” coherente.
- Variante en el tiempo: todos los datos son asociados con un periodo de tiempo específico, es decir, mantiene un histórico de los datos.
- No volátiles: los datos son estables. Se pueden agregar más datos, pero los existentes no se eliminan.

Se definen dos tipos de data mart, los dependientes y los independientes (Inmon, 2005):

Dependientes: Son los que se construyen a partir de un data warehouse central, es decir, reciben sus datos de un repositorio central, por lo que desde un mismo data warehouse, pueden surgir varios data marts.

Independientes: Son aquellos data mart que no dependen de un data warehouse central, ya que pueden recibir los datos directamente de los sistemas transaccionales.

Otro aspecto por considerar es el refrescamiento del data mart, el cual puede ser:

- A los pocos segundos de haberse actualizado los datos base
- Primero se transforman los datos y luego se los carga con base en los requerimientos de almacenamiento.
- Los datos son extraídos diariamente o con alguna sincronización.
- La decisión dependerá de los requerimientos del negocio.

De acuerdo con sus características, podemos citar que los data marts tienen ciertas ventajas por cuanto se selecciona este modelo para ser desarrollado, unido claro está, a la naturaleza del tema específico que se tratará en el presente trabajo.

- Dado que un data mart soporta menos usuarios que un data warehouse se puede optimizar para recuperar más rápidamente los datos que necesitan los usuarios.
- Menores cantidades de datos implican que se procesan antes, tanto las cargas de datos como las consultas.
- Las peticiones pueden acotarse al área o red que sirve esos datos, sin afectar al resto de los usuarios.
- La aplicación cliente, que pide la consulta es independiente del servidor que la procesa y del servidor de bases de datos que almacenan la información.
- Los costos que implica la construcción de un data mart son mucho menores a los de la implementación de un data warehouse.

La desventaja que se puede mencionar sería que no permite el manejo de grandes volúmenes de información por lo que muchas veces se debe recurrir a un

conjunto de data marts para cubrir todas las necesidades de información de la empresa.

Como lo señala Kimball, el crear un data mart, obedece a satisfacer una necesidad de un área, para el caso de la predicción de la demanda, la información de la potencia es el principal insumo para un proceso de proyección, por lo que un data mart es un punto de partida necesario para realizar esta tarea de manera sistemática (Kimball & Ross, 2002).

En resumen, la diferencia primordial entre un data warehouse y un data mart es el ámbito de acción y sus consecuencias, esto es que el alcance de un data warehouse es a nivel de todo el negocio, en tanto el data mart es dirigido a una línea específica del negocio, por lo que tiende a enfocarse en un tema, su tamaño es más reducido y demora menos tiempo su implementación, comparado con el data warehouse. Dicho de otra manera: el data mart está pensado para cubrir las necesidades de un grupo de trabajo o de un determinado departamento dentro de la organización. Es el almacén natural para los datos departamentales. En cambio, el ámbito del data warehouse es la organización en su conjunto. Es el almacén natural para los datos corporativos comunes. En la línea de pensamiento de Kimball: el data warehouse no es más que la unión de los data marts.

Sistemas de extracción, transformación y carga (ETL)

Para la carga de datos hacia el data mart se pueden utilizar herramientas de ETL (por sus siglas en inglés de Extracción, Transformación y Carga).

- Extracción de datos desde las distintas fuentes de datos
- Transformación, son tareas de limpieza y consolidación de datos
- Carga de los datos transformados al data warehouse o data mart

Según Kimball (2004) un buen diseño de ETL asegura la calidad y la consistencia de los datos. Por tanto, no sólo es vital para el almacén de datos sino que además agrega valor a los datos.

Es importante recalcar que debe existir un proceso de calidad de datos el cual debe, al menos

- Identificar anomalías; eliminar inconsistencias y detectar datos incorrectos

- Desarrollar una estrategia y metodología para mantener el sistema de detección y de limpieza de datos

Una de las herramientas comerciales más utilizadas para la carga de datos es el SQL Server Integration Services de Microsoft.

Modelo Multidimensional

El modelo de datos multidimensional forma parte integral de lo que se conoce como procesamiento analítico en línea (OLAP, por sus siglas en inglés). Al ser en línea, se espera que en sesiones interactivas, el sistema responda a complejas consultas analíticas. El modelo de datos multidimensional está diseñado para resolver tales consultas en tiempo real.

Este modelo es importante porque procura y aplica el principio de simplicidad. Kimball defiende este principio y señala que “la simplicidad es la clave fundamental que permite a los usuarios comprender las bases de datos” (Kimball & Ross, 2002).

Las hojas de cálculo tradicionales permiten hacer una representación bidimensional de datos, por filas y columnas. Por ejemplo, se puede tener a nivel de filas el tiempo (en meses) y en columnas los costos. Podrían ser varias columnas, representando ventas, costos directos, indirectos y un total. En este caso, se tendría entonces la dimensión de tiempo relacionada con varios hechos o variables. Las dimensiones ayudan a describir la organización de los datos. Si se agrega una dimensión más, por ejemplo producto, la estructura de y filas y columnas se torna insuficiente. Intuitivamente una estructura de tres dimensiones, como un cubo, resuelve el problema; sin embargo, ¿qué sucede si se considera una cuarta dimensión o más?

Para representar modelos multidimensionales de tres o más dimensiones, Thomsen sugiere las “Estructuras de Tipo Multidimensional (ETM)”, conocidas en inglés como “Multidimensional Type Structures (MTS)”. Éstas se convierten en una metáfora visual que ayuda a describir los modelos de más de cuatro dimensiones (Thomsen, E., 2002).

Los ETM permiten representar todas las combinaciones necesarias entre las variables y las dimensiones y de esta forma validar si la abstracción del problema y la solución que se construye son apropiadas.

Estas estructuras constituyen una forma sencilla de representar modelos y validarlos ante los grupos de toma de decisión, para constatar que se están comprendiendo los requerimientos de información, antes de continuar con el desarrollo de la herramienta para análisis y apoyo a la toma de decisiones.

Una estructura que de manera eficaz permita la manipulación de los datos en múltiples dimensiones es lo que se conoce como un modelo multidimensional.

El modelo de datos multidimensional se compone de cubos, medidas, dimensiones, jerarquías, niveles y atributos. La Figura 6 muestra las relaciones entre estos objetos:

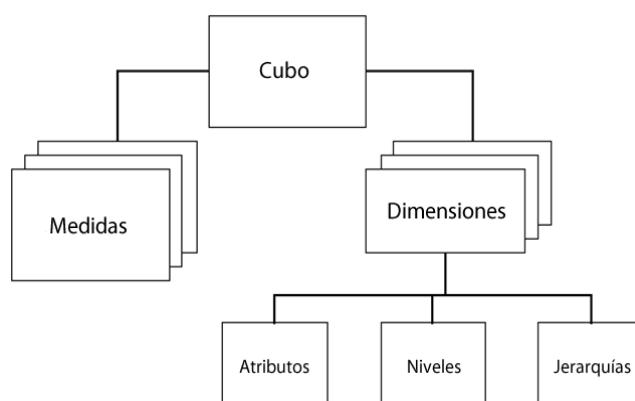


Figura 6. Diagrama de modelo lógico multidimensional (elaboración propia)

Cubos

Un cubo es una organización lógica de los datos multidimensionales. Un cubo se deriva de una tabla de hechos. Las dimensiones categorizan los datos de un cubo. A su vez, un cubo contiene medidas que comparten la misma dimensionalidad. Las medidas en un mismo cubo tienen las mismas relaciones con otros objetos lógicos y pueden ser fácilmente analizadas y se muestran juntas. Los cubos generalmente no se exponen a los usuarios finales ya que estos generalmente están más interesados en las medidas contenidas dentro de los cubos.

Medidas

Las medidas llenan las celdas de un cubo con los hechos recogidos sobre operaciones empresariales. Las medidas están organizadas en dimensiones, que suelen incluir una dimensión de tiempo.

Las medidas son como arreglos y se asocian automáticamente a la columna de la tabla física de hechos y las tablas de dimensiones relacionadas. Esta transformación de la columna de la tabla de hechos a la medida aísla al usuario de la complejidad del esquema subyacente y de tener que comprender cómo es que se unen las distintas partes del esquema.

Las medidas pueden compartir dimensiones. Así, por ejemplo, en un cubo determinado sobre ventas, el precio y el costo probablemente compartirían las mismas dimensiones: producto, canal y hora. Sin embargo, una medida como cantidad vendida puede ser dimensionada por producto, geografía, canal y tiempo.

Las medidas son estáticas y constantes mientras son usadas por los analistas y ellos a su vez las emplean para la toma de decisiones. Se suelen actualizar a intervalos regulares: semanal, diaria o periódicamente durante el día.

Una decisión crítica en la definición de una medida es el nivel de detalle más bajo, llamado también granularidad del dato. De esta decisión depende el poder contestar o no algunas preguntas con los datos del cubo y también el volumen de datos y el tamaño del cubo. Es una decisión de diseño que depende de los requerimientos del negocio.

Dimensiones

Las dimensiones contienen un conjunto de valores únicos que identifican y clasifican los datos. Ellas forman los bordes de un cubo, y por tanto, de las medidas dentro del cubo. Dado que las medidas son típicamente multidimensionales, un solo valor de una medida debe ser calificado o descrito por un miembro de cada dimensión para que tenga sentido. Por ejemplo, la medida ventas tiene cuatro dimensiones: tiempo (fecha y hora), cliente, producto y canal. Un valor de ventas en particular (43,613.50) sólo tiene sentido cuando es calificado por un período de tiempo específico (Feb-01), un cliente, un producto, y un canal (un catálogo por ejemplo).

Jerarquías y niveles

Una jerarquía es una forma de organizar los datos en diferentes niveles de agregación. Al visualizar datos, los analistas utilizan jerarquías de la dimensión para reconocer tendencias en un nivel, profundizar hasta niveles inferiores para identificar las razones de estas tendencias, y regresar hasta los niveles más altos para ver qué efecto tienen estas tendencias en un sector más amplio de la empresa.

Cada nivel representa una posición en la jerarquía. Cada nivel por encima del nivel base (el más detallado) contiene valores agregados para los niveles por debajo de él.

Jerarquías y niveles tienen una relación de muchos a muchos. Una jerarquía contiene típicamente varios niveles, y un nivel puede incluirse en más de una jerarquía.

Atributos

Un atributo proporciona información adicional acerca de los datos. Algunos atributos se utilizan para la visualización. Se podría tener atributos como colores, sabores, o tamaños. Este tipo de atributo se puede utilizar para conocer ¿cuáles colores fueron los más populares en los vestidos de mujer en el verano de 2002? ¿Cómo se compara esto con el verano anterior?

Los atributos de tiempo pueden proporcionar información sobre la dimensión de tiempo que puede ser útil en algunos tipos de análisis, tales como identificar el último día o el número de días en cada período de tiempo.

2.2. Proyección de la demanda

En la vida diaria es fácil encontrarse con una infinidad de aparatos eléctricos, por ejemplo, una bombilla, un equipo de aire acondicionado o un secador. Todos ellos consumen energía eléctrica y la transforman en un trabajo útil: iluminar, enfriar o secar. Todos ellos tienen un requerimiento de potencia, normalmente indicada en watts (vatio en español, aunque es más comúnmente utilizado el término watt o W) y esto indica la rapidez con la que un aparato eléctrico transforma o consume la energía eléctrica.

Si la potencia en cuestión es de mediana o gran potencia, normalmente se utilizan múltiplos del watt para expresar estas cantidades mayores. Uno de los más comunes es el kilowatt (kW) que equivale a 1 000 watts. Para medidas aún mayores se utiliza el megawatt (MW) que equivale a 1 000 000 watts, como podría ser en el caso de una planta o central eléctrica.

La energía consumida por estos aparatos se calcula multiplicando la potencia por el tiempo y se mide en watts hora (Wh) y sus múltiplos correspondientes.

Por tanto la potencia eléctrica está directamente relacionada con el consumo eléctrico o energía (kilowatts hora).

Por ejemplo, un cálculo del consumo o demanda de energía a partir de la potencia: supóngase que se tiene un requerimiento de potencia de 5 MW de forma constante en el día, su consumo a lo largo de un día sería 5 MW multiplicado por 24 horas, es decir, 120 MWh.

Luego si se considera la suma de todos los aparatos eléctricos del país, se encuentra con que es necesario satisfacer los requerimientos o demanda de potencia y energía a nivel nacional.

En el presente trabajo, la demanda de potencia de corto plazo se refiere a la curva de carga diaria para uno o varios días. Esta curva está compuesta por la demanda de potencia promedio, lo que equivale a la energía horaria que se requiere para satisfacer los requerimientos de energía que los usuarios o el sistema necesita. El pronóstico en el corto plazo de la potencia tiene como objetivo proyectar esta curva de carga para pronosticar los requerimientos de energía y así procurar un manejo eficiente del Sistema Eléctrico Nacional y asegurar la operación del mismo optimizando el predespacho económico/técnico horario.

El despacho de carga es una de las actividades técnicas más importantes para el Sistema Eléctrico Nacional (SEN), el cual tiene como objetivo garantizar el suministro de energía, maximizar la seguridad operativa de ese sistema y minimizar los precios mayoristas en el mercado horario de energía. Para ello el CENCE elabora un plan de despacho o predespacho económico/técnico horario, en el que parte de una curva de demanda proyectada para cubrir el requerimiento

de energía asignando la producción a aquellas unidades generadoras según un orden establecido, considerando disponibilidad, costos y contratos.

Por la naturaleza de la energía eléctrica, ésta debe ser generada en el momento en que se requiere, por lo que se debe hacer un seguimiento continuo, lo que se denomina operación en tiempo real. Éste tiene como objetivo en primera instancia, satisfacer la demanda de potencia cumpliendo, en la medida de lo posible, con el predespacho. Tiene que verificar su cumplimiento y tomar las acciones necesarias para afrontar las desviaciones que se presentan entre lo planeado y lo que realmente está sucediendo en el sistema, incluyendo el aumento o disminución de la demanda de potencia.

Para llevar a cabo el predespacho horario, las empresas generadoras de energía deben proporcionar al CENCE el programa horario de generación según sus recursos. Por otro lado el CENCE debe estimar los requerimientos de energía horario para los días posteriores, luego, con estos dos insumos se crea el predespacho, que en resumen es el cómo se va a satisfacer la demanda estimada con los recursos disponibles.

A modo de ejemplo en la Figura 7 se muestra la proyección de la demanda para un día específico, esta proyección incluye los requerimiento de energía a nivel nacional a nivel horario; la construcción de esta curva es uno de los objetivos del presente proyecto.

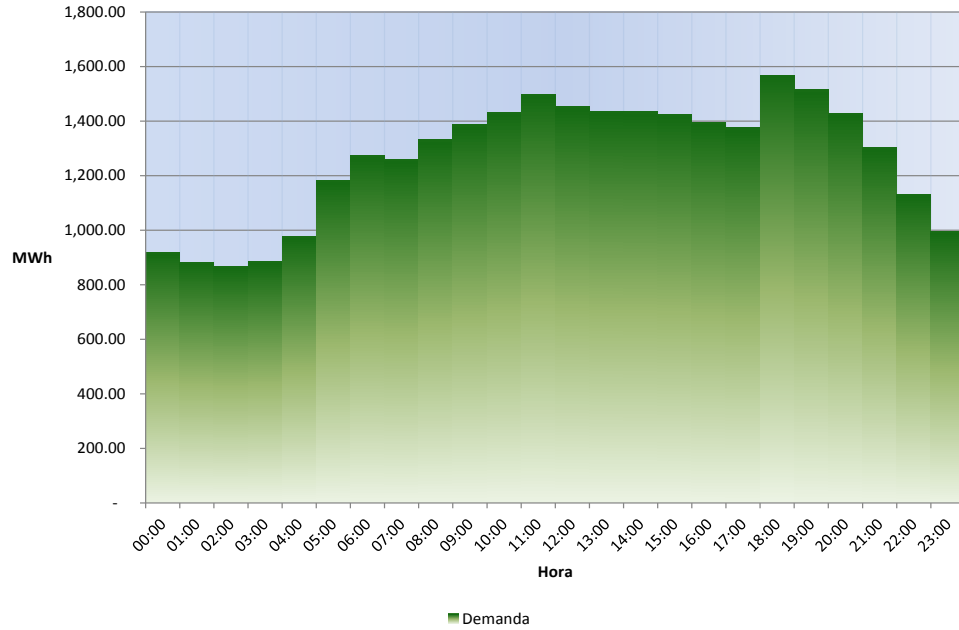


Figura 7. Proyección de la Demanda de Energía para el 12 de marzo de 2014 (elaboración propia)

A partir de esta proyección, estos datos entran en otro proceso para construir el predespacho, el cual define cómo se cumplirá los requerimientos de energía con los recursos disponibles. Un resumen de este predespacho, se muestra en la Figura 8.

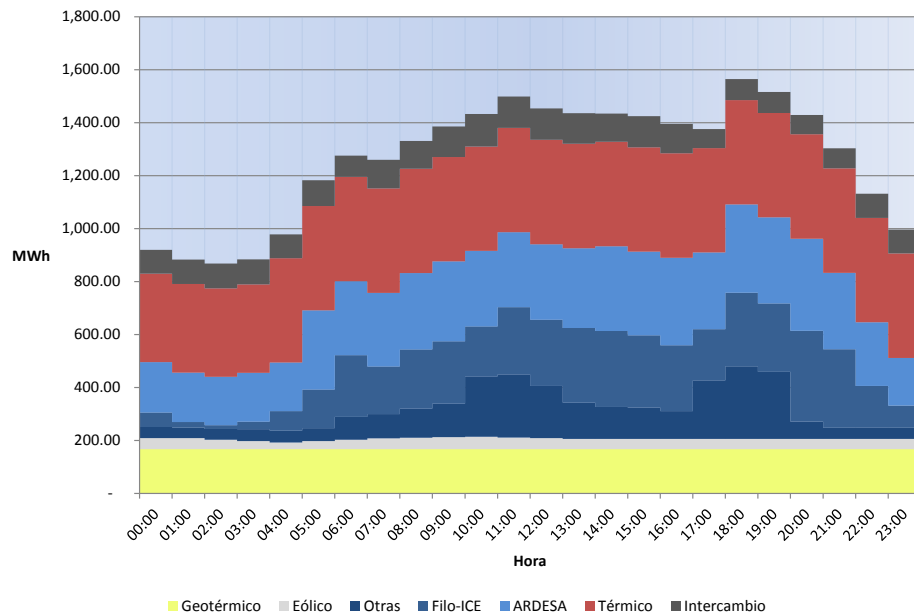


Figura 8. Predespacho Horario de Energía para el 12 de marzo de 2014 (elaboración propia)

Para la proyección de la demanda, no se cuenta con software específico. Para satisfacer la necesidad, actualmente un grupo de ingenieros en sistemas de potencia hacen uso de Microsoft Excel y de las bases de datos para realizar las proyecciones. No obstante, este proceso no está completamente sistematizado y algunos puntos quedan a criterio experto de quién lo realiza.

Proyectar la demanda consiste en realizar una estimación del consumo de energía y de potencia, que permita conocer con anticipación la demanda real.

Si se considera que una serie de tiempo es cualquier variable que conste de datos reunidos, registrados u observados sobre incrementos sucesivos de tiempo, los registros históricos de potencia y energía constituyen dos series temporales de carga, debido a esto es que el proceso de proyección de la demanda será enfrentado considerando las técnicas y algoritmos, utilizando modelos de serie de tiempo (D'oro, Lozano, & Moreno, 2007).

Existen diferentes opciones para enfrentar este problema, tal es el caso de las redes neuronales, promedios móviles y ARIMA. Cada uno tiene sus características propias para unos casos u otros. Por lo que se revisarán para buscar cual modelo sería el más apropiado para su aplicación al SEN.

Las proyecciones de la demanda eléctrica son utilizadas para (D'oro, Lozano, & Moreno, 2007):

- Programar y operar las unidades de generación diariamente.
- Reducir las desviaciones del predespacho con respecto a la demanda real así como los redespachos, evitando la necesidad de usar unidades de generación más costosas para suplir la demanda, como es el caso de las plantas térmicas, o por el contrario, no utilizar el total de la energía programada ya contratada. Esto permite reducir sobrecostos que repercuten en incrementos en el precio de la energía para los usuarios finales.
- Ayudar a estimar el flujo de carga y la realización de estudios de seguridad operativa. Esto finalmente ayuda a tomar decisiones que pueden prevenir las sobrecargas, evitando así eventos como la falla

de equipos o apagones, que implicarían grandes pérdidas. Por lo que ayuda a mejorar la confiabilidad del sistema eléctrico.

- Establecer criterios para que los agentes generadores decidan sus ofertas de generación para el día siguiente y en los casos donde hay un mercado, se defina el precio de bolsa para la energía eléctrica.
- Evitar posibles penalizaciones por tener desviaciones en el pronóstico mayores a un margen establecido en la regulación de los mercados de energía, por ejemplo, para el Mercado Eléctrico Regional de Centroamérica, las desviaciones en los intercambios de energía (exportación e importación) son penalizadas cuando son mayores a 4 MWh.

La proyección de la demanda constituye la base para el predespacho de energía, el cual define cómo y cuánta energía aportará cada planta para satisfacer la demanda. De tal forma que si la proyección no es correcta, generalmente conlleva costos elevados. Esto debido a dos características básicas que tiene un mercado eléctrico: (Mallo González., 2003)

- La energía eléctrica no puede ser almacenada, al menos no en las cantidades que puede producir una planta eléctrica.
- En la mayor parte de sus usos la electricidad juega el papel de un input específico, es un factor productivo que no puede ser reemplazado en el corto plazo, es decir, si no se dispone de energía se puede detener la producción o actividad, dado que no se puede sustituir de forma inmediata.

Para analizar el impacto de una proyección incorrecta en aspectos de costos, se puede analizar desde dos situaciones:

- Si la proyección fue menor a la demanda real, es decir, los clientes están requiriendo más energía que la que se proyectó, el CENCE realiza un redespacho, es decir, ajustar la forma en que se está despachando las plantas, e incluir en el plan de uso, en la mayoría de los casos, plantas o unidades de generación más costosas para

atender la demanda o incluso, recurrir a la importación de energía y evitar así desabastecimiento del suministro eléctrico.

- Por otro lado, si la proyección fue mayor a la demanda real, en otras palabras, se proyectó utilizar más energía que la requerida, se incurre en costos adicionales de operación pues se consideró el encendido de unidades que no se utilizarán, éstas, generalmente, serán las más costosas o se realizaron compras de energía que no se requerían y como la energía no puede ser almacenada, simplemente no se pueden utilizar esos recursos.

Estas desviaciones en las proyecciones con respecto a la realidad conducen en última instancia a costos más elevados para el usuario final y disminuye la capacidad de competir con precios más bajos en el mercado de la energía.

Actualmente se realizan proyecciones de demanda diarias con resolución horaria, es decir, para las 24 horas siguientes, en este sentido el aporte del presente trabajo es llevar a cabo la tarea de una forma más sistemática y brindar la plataforma para evaluar estas proyecciones y afinarlas. Incluso, brindar un aporte que pueda ser utilizado en las reprogramaciones o redespachos, con la posibilidad de realizar proyecciones semanales, es decir, para las 168 horas de la semana.

Series de tiempo

Una serie temporal es una sucesión de observaciones de una variable tomadas en varios instantes de tiempo. Es de interés estudiar los cambios en esa variable con respecto al tiempo y predecir sus valores futuros.

Ejemplos de series temporales se pueden encontrar en muchos campos del conocimiento:

- Economía: producto interior bruto anual, tasa de inflación, tasa de desempleo, otros.
- Demografía: nacimientos anuales, tasa de dependencia, otros.
- Meteorología: temperaturas máximas, medias o mínimas, precipitaciones diarias, otros.

- Las series temporales se pueden clasificar en:
- Estacionarias. Una serie es estacionaria cuando es estable a lo largo del tiempo, es decir, cuando la media y varianza son constantes en el tiempo. Esto se refleja gráficamente en que los valores de la serie tienden a oscilar alrededor de una media constante y la variabilidad con respecto a esa media también permanece constante en el tiempo.
- No estacionarias. Son series en las cuales la tendencia y/o variabilidad cambian en el tiempo. Los cambios en la media determinan una tendencia a crecer o decrecer a largo plazo, por lo que la serie no oscila alrededor de un valor constante.

Además, se pueden presentar efectos estacionales, es decir, que el comportamiento de la serie es parecido en ciertos tiempos periódicos en el tiempo.

Hasta antes de 1927 (Acevedo, 2004), los pronósticos de estas series se realizaban simplemente extrapolando la serie en el tiempo. El principio de lo que se puede llamar “predicción moderna” de series de tiempo o de series temporales se puede fijar en este año, cuando Udney Yule desarrolló la técnica autorregresiva para pronosticar el número anual de manchas solares, su modelo pronosticaba el siguiente valor como una suma ponderada de las observaciones previas de la serie.

En la década de los años 80, se presentaron acontecimientos importantes en la evolución de los estudios sobre series temporales, entre las que se pueden citar (Acevedo, 2004):

- el incremento en la capacidad de procesamiento de los computadores, lo que permitió el estudio de series temporales mucho más largas, la aplicación de algoritmos más complejos, y la visualización interactiva tanto de los datos como de los resultados.
- el desarrollo de la inteligencia artificial y específicamente, de las redes neuronales artificiales.

En lo que respecta a la predicción de la demanda horaria de energía eléctrica, el avance en cuanto al trabajo investigativo de la temática en las últimas

dos décadas, se debe, en primera instancia, a la transformación liberalizadora de la mayoría de los mercados de energía eléctrica del mundo, y en segunda, gracias a los mismos dos puntos citados anteriormente.

En diferentes países se han dado a la tarea de generar investigación respecto del pronóstico de demanda de energía eléctrica, y se han presentado diversos resultados, por esta razón se utilizan diferentes métodos y algoritmos en diferentes partes.

En Colombia se ha realizado una investigación con series de tiempo (Murillo, 2003) utilizando la metodología ARIMA (Autorregresive-Integrated-Moving Average), Pero más recientemente, el Centro Nacional de Despacho de Colombia, realizó investigaciones con buen éxito utilizando el Modelo de promedios móviles; donde se demuestra que el uso de promedios móviles en conjunto con filtros de tipo de día puede conducir a muy buenos resultados.

En cuanto a la predicción de demanda, las redes neuronales artificiales se presentan como una variante de gran aplicabilidad a la solución de problemas de predicción, debido a su capacidad intrínseca para aproximar funciones matemáticamente desconocidas y poder clasificar patrones. Por ejemplo, en España se han hecho estudios (Mallo González., 2003) donde se ha desarrollado un modelo de Red Neuronal Artificial.

En Bolivia se presenta un estudio donde el pronóstico, es realizado básicamente donde se utilizan modelo ARIMA x SARIMA, modelos Arma-Garch y finalmente, modelos con base en redes neuronales artificiales, logrando el mínimo error utilizando las redes neuronales artificiales (Tudela, 2011).

Un estudio en España hizo uso de otros modelos basados en sistemas de inferencia difusa, optimizados mediante un algoritmo genético, superando los resultados obtenidos con un conjunto de modelos clásicos de predicción, como los son las series de tiempo (Yanguas-Peña, y otros, 2008).

En la India presenta un método donde utiliza las redes neuronales artificiales, combinando esta técnica con algoritmos de clustering para obtener los patrones de entrenamiento y lograr mejores resultados (Jain, 2009).

A modo de resumen, para las técnicas de predicción, las referencias recopiladas muestran avances en métodos convencionales como cuando se trabaja con modelos lineales como ARIMA para predicción de demanda de corto plazo. Incluso el de promedio móviles, que pareciera de los más simples, pero han dado buen resultado en el caso de Colombia (D'oro, Lozano, & Moreno, 2007).

Por último, las redes neuronales artificiales (RNA) que son las más utilizadas para la proyección de la demanda, las cuales son modelos que intentan reproducir el comportamiento del cerebro (Calvo, 2013).

El presente trabajo se centrará en el estudio de tres métodos utilizados para los métodos de series de tiempo, a saber:

- Promedios móviles
- ARIMA
- Redes neuronales

Promedios Móviles

Este método es quizás el más simple de los que se evalúan. La idea detrás de este método es considerar que el valor por proyectar es el promedio de los n valores anteriores, así la fórmula es:

$$F_t = \frac{\sum_{i=1}^n d_{t-i}}{n}$$

Donde:

F_t = Pronóstico para el periodo

n = Número de periodos incluidos en el promedio.

d_{t-i} = Demanda del periodo $t-i$

Este diseño tiene como objetivo disminuir las desviaciones horarias en el pronóstico mediante un modelo sencillo y fácil de aplicar. Sus principales características son:

- a. Utilización de filtros de tipo de día clasificados en día festivo, lunes, martes, miércoles, jueves, viernes, sábado y domingo.
- b. Manejo especial de los días exclusivos o únicos en el año.
- c. Utilización de filtros para cada una de las 24 horas del día.

- d. Uso de promedios móviles de diferentes órdenes para cada serie de carga horaria.

La Figura 9 muestra el proceso de la proyección con el uso de promedios móviles.

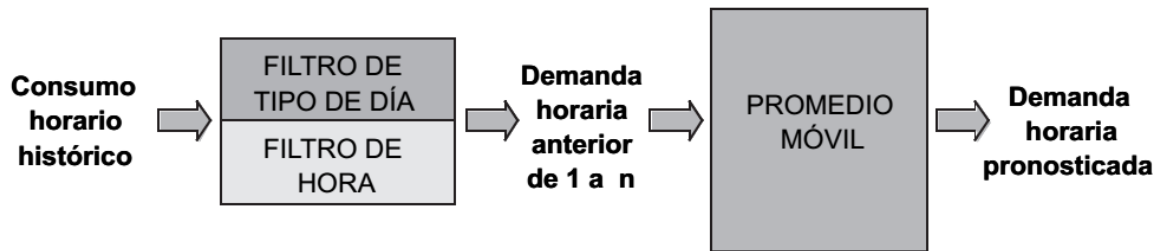


Figura 9. Modelo de pronóstico basado en promedios móviles (D'oro, Lozano, & Moreno, 2007)

Para los días normales se señala que los resultados obtenidos fueron muy buenos, pero para los días festivos en el calendario, los resultados variaron negativamente, debido a que generalmente sólo se presentan pocos días festivos repartidos en los doce meses del año, lo cual impide tener valores más recientes del comportamiento de la carga para esos días. Sin embargo, encontraron que la demanda horaria de los días festivos se modela mucho mejor si se utiliza el pronóstico del domingo anterior a ese día festivo. (D'oro, Lozano, & Moreno, 2007).

Este hallazgo se circunscribe a Colombia, por lo que se debe verificar su validez en otros sistemas.

El modelo de promedios móviles tiene las ventajas de que cuando es utilizado conjuntamente con filtros de tipo de día mejora el proceso de proyección de la demanda, además de que es un método simple de implementar, pero presenta el inconveniente con los días festivos, por lo que resulta necesario implementar un conjunto de excepciones o métodos alternativos para poder tratar estos días.

ARIMA

A comienzo de la década de los años 70, G.E.P. Box, profesor de Estadística de la Universidad de Wisconsin, y G.M. Jenkins, profesor de Ingeniería de Sistemas de la Universidad de Lancaster, introdujeron una pequeña revolución en el enfoque del análisis de series temporales, en sus trabajos sobre el comportamiento de la contaminación en la Bahía de San Francisco, con el propósito de establecer mejores mecanismos de pronóstico y control. El libro (1976) en el cual se describe la metodología, se convirtió rápidamente en un clásico, y sus procedimientos se utilizan ampliamente desde entonces en diferentes ramas de la ciencia, conociéndose como modelos ARIMA y también como modelos Box-Jenkins.

En el modelo, cada valor tomado por la variable en un instante dado, está influido por los valores de la variable en momentos anteriores y se expresa como una relación lineal en función de:

- Valores recientes de la variable
- Ruidos en valores recientes de la variable
- Valores remotos de la variable
- Ruidos en valores remotos de la variable

El esquema general del modelo es el siguiente:

$$X_t = a_1X_{t-1} + a_2X_{t-2} + \dots + a_pX_{t-p} + Z_t + b_1Z_{t-1} + \dots + b_qZ_{t-q}$$

que es la fórmula general de los modelos denominados ARMA. Está constituido por una combinación de p términos AR (proceso autorregresivo), y q términos MA (proceso de medias móviles). La parte AR modela la influencia de los valores anteriores de la serie (X_{t-1} hacia atrás) y la parte MA modela la influencia del ruido en valores anteriores de la serie (Z_{t-1} hacia atrás), junto con el término Z_t que corresponde al ruido esperado en el mismo momento t en el que se estima el nuevo valor de la variable X .

Una de las ventajas de estos modelos es su gran simplicidad (sumas de términos), frente a los modelos propuestos en la formulación clásica.

La letra I que aparece en el nombre del modelo completo -ARIMA-, corresponde al proceso último por realizar, una vez definido el tipo de modelo y

estimados los coeficientes de éste, ya que entonces se deben restablecer las características originales de la serie de datos, que fue transformada para inducir estacionalidad. A ese proceso inverso se denomina en general Integración y aporta esa letra que completa el nombre.

Así se define una serie ARIMA(p,d,q), es decir, una serie de tiempo autoregresiva integrada de media móvil. Donde p denota el número de términos autoregresivos, d el número de veces que la serie debe ser diferenciada para hacerla estacionaria y q el número de términos de la media móvil invertible.

La construcción de los modelos ARIMA(p,d,q), se lleva de manera iterativa mediante un proceso iterativo de mejora del modelo en el que se puede distinguir las siguientes cuatro etapas:

1. Identificación. Utilizando los datos ordenados cronológicamente se intentará sugerir un modelo que en primera instancia pueda aproximar la serie. El objetivo es determinar los valores que sean apropiados para reproducir la serie de tiempo. En esta etapa es posible identificar más de un modelo candidato que pueda describir la serie.
2. Estimación. Considerando el modelo apropiado para la serie de tiempo se realiza inferencia sobre los parámetros.
3. Validación. Se realizan contrastes, comparaciones para validar si el modelo seleccionado se ajusta a los datos reales que se utilizan a modo de entrenamiento, si no es así, y los datos presentan un error no aceptable, se debe escoger el próximo modelo candidato y repetir los pasos anteriores.
4. Predicción. Una vez seleccionado el mejor modelo candidato ARIMA(p,d,q), se pueden hacer pronósticos en términos probabilísticos de los valores futuros.

Este proceso interactivo, se puede realizar con las herramientas actuales de minería de datos de forma automática, por ejemplo y para el caso del presente proyecto, el SQL Server, desde su versión 2008 implementa este método y determina automáticamente los parámetros p, d y q del modelo.

Microsoft desarrolló su propio algoritmo ARTXP para utilizarlo en SQL Server 2005 para tratar las series de tiempo y basó la implementación en el algoritmo de árboles de decisión de Microsoft. Por tanto, el algoritmo ARTXP se puede describir como un modelo de árbol con regresión automática que permite representar los datos periódicos de una serie temporal. Este algoritmo relaciona un número variable de elementos pasados con cada elemento actual que se predice.

Luego, para la versión de SQL Server 2008 se agregó el algoritmo ARIMA para mejorar la exactitud de la predicción a largo plazo. Se trata de una implementación del proceso para calcular promedios de movimiento integrado de regresión automática descritos por Box y Jenkins. La metodología ARIMA permite determinar las dependencias en las observaciones tomadas secuencialmente en el tiempo y puede incorporar impactos aleatorios como parte del modelo. El método ARIMA también admite la estacionalidad multiplicativa. (MacLennan, Tang, & Crivat, 2009)

El algoritmo de serie temporal de Microsoft usa ambos métodos, ARTXP y ARIMA, mezclando los resultados para mejorar la precisión de la predicción. Pero igualmente se puede configurar los parámetros del algoritmo para usar únicamente ARTXP o únicamente ARIMA, o incluso para controlar el modo en que se combinan los resultados de los dos algoritmos.

Redes neuronales

Las Redes Neuronales Artificiales o RNAs, son modelos matemáticos inspirados en la organización y el funcionamiento de las neuronas biológicas del cerebro humano. Existen numerosas variaciones de redes neuronales que están relacionadas con la naturaleza de la tarea que se ha asignado. Igualmente existen distintas variaciones sobre cómo modelar la neurona; en algunos casos se asemejan mucho a las neuronas biológicas mientras que en otros, los modelos son muy diferentes.

En diversas fuentes se sugiere algunas características de las RNAs que las hacen especialmente útiles para realizar proyecciones en series de tiempo,

particularmente con la proyección de la demanda de energía. Fundamentalmente tienen:

- la capacidad de aproximar prácticamente cualquier función (incluso las no lineales).
- la posibilidad de hacer aproximaciones “piece-wise” o por trozos, de las funciones.

Desde el punto de vista matemático, las RNAs se pueden considerar como aproximadores universales de funciones. Esto significa que pueden automáticamente aproximar la función que mejor se ajuste a los datos, permitiendo de esta manera extraer relaciones cuando las funciones son muy complejas. Además, las RNAs son intrínsecamente no lineales (Rumelhart-1986), lo cual implica no solo que pueden estimar correctamente funciones no lineales, sino que también pueden extraer elementos no lineales de los datos, una vez extraídos los términos lineales.

El modelamiento de la neurona biológica es relativamente sencillo. Se compone de una entrada x , un valor de peso w , una función de suma Σ , una función de activación f y una salida a .

El funcionamiento de una neurona artificial está basado en este diseño. Como se ilustra en la Figura 10, básicamente consiste en aplicar un conjunto de entradas, cada una representando la salida de otra neurona, o una entrada del medio externo, luego se realiza una suma ponderada con estos valores. El valor resultante se lleva a través de f (función de activación o de transferencia), de donde sale el valor resultante, que es transmitido a la siguiente neurona o al exterior

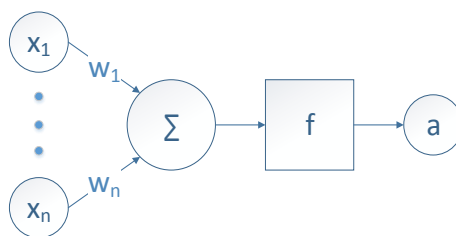


Figura 10. Esquema de una neurona artificial (elaboración propia)

Esta función de activación puede ser lineal o no. Para el caso de la implementación de redes neuronales de Microsoft Neural, se usa la tangente hiperbólica para las funciones de activación en los nodos ocultos y el sigmoide para los nodos de salida.

De acuerdo con el número y con la distribución de las neuronas artificiales, ellas conforman capas o niveles.

Capa o nivel es el conjunto de neuronas cuyas entradas provienen de la misma fuente (que puede ser otra capa de neuronas) y cuyas salidas se dirigen al mismo destino (que puede ser otra capa de neuronas)

Una red neuronal artificial es la interconexión de varias neuronas. La Figura 11 muestra una red neuronal estructurada en capas; es una de las estructuras en las cuales se pueden asociar las neuronas. En este sentido, los parámetros fundamentales de la red son: el número de capas, el número de neuronas por capa y el tipo y número de conexiones entre neuronas. No existe un método o regla que determine el número óptimo de neuronas ocultas para resolver un problema dado, generalmente se determinan por prueba y error, es decir, partiendo de una arquitectura ya entrenada, se realizan cambios aumentando y disminuyendo el número de neuronas ocultas y el número de capas, hasta conseguir la arquitectura que se ajuste a la solución del problema

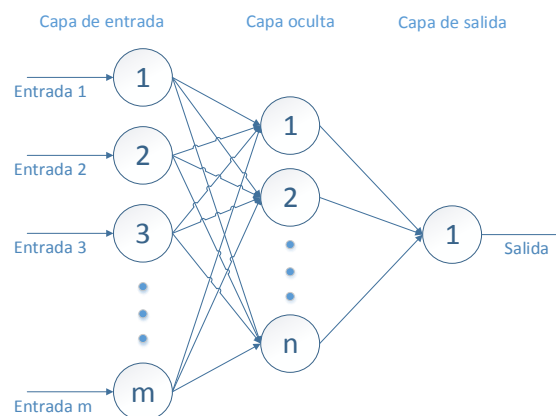


Figura 11. Red neuronal multicapa típica (elaboración propia)

Luego de conformar el esquema de la red neuronal, se procede a la etapa de aprendizaje de la red. El aprendizaje es el proceso por el cual una red neuronal

modifica sus pesos en respuesta a una información de entrada. Los cambios que se producen durante el proceso de aprendizaje se reducen a la destrucción, modificación y creación de conexiones entre las neuronas. En el caso de las redes neuronales artificiales, se puede considerar que el conocimiento se encuentra representado en los pesos de las conexiones.

En forma general, se consideran dos tipos de aprendizaje: supervisado y no supervisado. La diferencia fundamental entre ambos tipos estriba en la existencia, o no, de un agente externo (supervisor) que controle el proceso de aprendizaje de la red.

Por lo general, las redes neuronales aprenden por medio de ejemplos, los cuales comúnmente se presentan a la red en patrones de entrada y salida. El objetivo del aprendizaje o entrenamiento de la red es ajustar los parámetros de la red, pesos y umbrales, con el fin de que las entradas presentadas produzcan las salidas deseadas, es decir, con el fin de minimizar la función de error.

La parte central de una red neuronal es la retro propagación. El proceso de entrenamiento de una red neuronal es un proceso iterativo. En cada iteración, el algoritmo compara los valores de salida con los valores reales conocidos para obtener los errores para cada neurona de salida. Los pesos hacia las neuronas de salida son modificadas con base en los cálculos de error. Estas modificaciones se propagan luego de la capa de salida a través de las capas ocultas hasta la capa de entrada. En consecuencia, todos los pesos en la red neuronal se ajustan.

El proceso central de la formación de redes neuronales se describe en los pasos siguientes:

1. El algoritmo asigna aleatoriamente los valores de todos los pesos en la red en la etapa inicial (generalmente desde $-1,0$ a $1,0$).
2. Para cada conjunto de entrenamiento, el algoritmo calcula las salidas basado en los pesos actuales de la red.
3. Los errores de salida son calculados, y el proceso de retropropagación calcula los errores para cada neurona de salida y oculta en la red. Luego los pesos en la red son actualizados. En el

caso de Microsoft, este utiliza la suma de cuadrados para calcular el error.

4. Se regresa al paso 2 hasta que se cumpla la condición de parada, que puede ser por:

- Suficiente precisión en un conjunto de retención: La tasa de errores de clasificación está por debajo de un determinado umbral.
- Iteración máxima: el proceso de entrenamiento ha alcanzado el límite superior del número de iteraciones.
- Convergencia de los pesos: el cambio en los pesos después de cada iteración se mantienen por debajo de un umbral.

En el proceso de entrenamiento se debe contar con un conjunto de datos, pero no recomiendan utilizar todo el conjunto de prueba para entrenar la red, para que ésta no se limite a memorizar los resultados, por el contrario, es bueno utilizar un 80% para entrenar y un 20% para probar el desempeño. (Acevedo, 2004)

La topología de la red neuronal debe definirse claramente antes de procesar. El número de neuronas de entrada y salidas se fija con un conjunto de datos de entrenamiento. Las opciones de configuración están relacionadas principalmente con la configuración de las capas ocultas, tales como el número de capas ocultas y el número de neuronas ocultas en cada capa oculta.

Una red neuronal puede tener cualquier cantidad de capas ocultas, así como de neuronas en estas capas. Así que la definición de la capacidad de una red no es una tarea sencilla, debido a que si se tienen muchas capas ocultas puede aumentarse la capacidad de aprendizaje, pero esto aumentará el tiempo de procesamiento. Otro problema es el sobreentrenamiento. Si se tiene muchas capas ocultas y nodos ocultos, la red tiende a recordar los casos de formación en lugar de encontrar la generalización de los patrones. Se ha demostrado que, en muchos casos, es suficiente una capa oculta.

El algoritmo de red neuronal de Microsoft usa una red de tipo perceptrón multinivel, que también se denomina red de tipo regla delta de propagación hacia atrás, compuesta por tres niveles de neuronas o perceptrones.

Estos niveles son:

1. un nivel de entrada
2. un nivel oculto opcional
3. un nivel de salida

El número de neuronas en la capa oculta también es muy importante. Si se utilizan muy pocas podrían no haber suficientes recursos para suministrar a las neuronas para que estas pocas neuronas resuelvan el problema. Por el contrario, si se utilizan demasiadas neuronas aumentará el tiempo de entrenamiento.

Microsoft propone una guía muy general para elegir el número de neuronas ocultas (MacLennan, Tang, & Crivat, 2009):

$$\text{Cantidad de neuronas} = c \cdot \sqrt{m \cdot n}$$

Donde:

- n es el número de neuronas de entrada
- m es el número de neuronas de salida
- c es una constante

En la red neuronal de Microsoft, el valor predeterminado de c es 4.

El número óptimo de nodos puede variar de un problema a otro, por lo que es necesario experimentar con la cantidad de nodos por utilizar.

Un depósito de datos constituiría, por su carácter histórico, la fuente primordial para la realización de estas proyecciones de demanda eléctrica, dado que mantiene la información de las series de tiempo o los datos de entrenamiento y en producción contendrán la información de entrada para correr el modelo.

3. Marco Metodológico

3.1. Enfoque de la investigación

Este trabajo de investigación utiliza un enfoque mixto pues utiliza un componente cualitativo y otro cuantitativo. El cualitativo se aplica en la comprensión de los procesos relativos a la operación del Sistema Eléctrico Nacional. De fondo existe un proceso inductivo (Hernández, Fernández-Collado, & Baptista, 2010), que lleva a la exploración, la descripción y la generación de perspectivas teóricas. Además, el investigador construye el conocimiento teniendo en cuenta las perspectivas y subjetividades de los participantes en el proceso de recolección de datos. El diseño de un data mart requiere de un abordaje metodológico de este tipo.

También la investigación emplea un enfoque cuantitativo. En concordancia con lo establecido por (Pita Fernández & Pértegas Díaz, 2002), es considerado de este tipo porque incluye la recolección y análisis de datos. Implica la generalización de los resultados a partir de una muestra y permite inferir y realizar una predicción de la demanda eléctrica nacional. Este enfoque es utilizado en el modelaje tanto de la red neuronal como de los promedios móviles. Asimismo, la comparación con el mecanismo utilizado actualmente y la demanda energética real es realizada en términos cuantitativos.

3.2. Diseño de la investigación

Se trata de un proyecto de investigación acción:

La finalidad de la investigación-acción es resolver problemas cotidianos e inmediatos y mejorar prácticas concretas. Su propósito fundamental se centra en aportar información que guíe la toma de decisiones para programas, procesos y reformas estructurales. La investigación-acción se conceptúa como el estudio de una situación con miras a mejorar la calidad de la acción; además, este diseño se adecua a problemas prácticos vinculados con un ambiente o entorno. (Hernández, Fernández-Collado, & Baptista, 2010). En este caso se parte del problema de la proyección de la demanda dentro del ambiente del CENCE, (pero que también

impacta un ambiente nacional), se estudia la situación y se busca mejorar el proceso identificado.

La perspectiva desde la cual se desarrollará el diseño, será la visión técnico-científica. Su modelo consiste en un conjunto de decisiones en espiral, éstas se basan en ciclos repetidos de análisis para conceptualizar y redefinir el problema repetidamente. Así, la investigación-acción se compone de fases secuenciales de acción: planificación, identificación de hechos, análisis, implementación y evaluación.

Esta perspectiva se acopla muy bien a los modelos de espiral de la metodología de Kimball y de CRISP.DM. (Kimball, R., & Ross, M., 2002) (IBM, 2001)

Existen dos diseños básicos de investigación-acción, práctico y participativo.

Para el proyecto se utilizará el diseño práctico por sus características y la relación de estas con el proyecto:

- Estudia prácticas locales. Se estudia el proceso actual de la proyección.
- Involucra indagación individual o en equipo. Se estudian métodos utilizados para resolver este tipo de problemas.
- Se centra en el desarrollo y aprendizaje de los participantes. Se profundiza y desarrollan estos métodos para evaluarlos.
- Implementa un plan de acción (para resolver el problema, introducir la mejora o generar el cambio). Se realiza un proceso para resolver el problema utilizando los diferentes métodos y comparándolos para ubicar la mejor alternativa.
- El liderazgo lo ejercen conjuntamente el investigador y uno o varios miembros del grupo o comunidad. Los investigadores interactúan estrechamente con los expertos del área durante todo el proceso.

3.3. Fuentes de información

Tanto para el diseño y la aplicación metodológica como para los modelos de redes neuronales y promedios móviles como del data mart, la fuente de información principal será la recopilación bibliográfica, por medio de la consulta de libros, tesis, artículos e Internet.

Por otra parte está el criterio experto del área técnica en la que se desarrollará el proyecto, básicamente profesionales en Ingeniería Eléctrica con énfasis en Sistemas de Potencia y los especialistas en sistemas de información del Centro Nacional de Control de Energía.

3.4. Descripción de los instrumentos

Entrevista no estructurada

Con la finalidad de obtener el criterio experto sobre las variables que se van a incorporar al modelo para generar las predicciones de la demanda se utilizará este mecanismo. De igual forma se usará para lograr la comprensión necesaria para el diseño de la solución de inteligencia de negocios. En este tipo de entrevista las preguntas no están del todo predeterminadas.

Base de datos históricos

Para la construcción del depósito de datos y de los modelos predictivos se contará con acceso a las bases de datos del CENCE.

Aplicación automatizada de los modelos

Tanto para la red neuronal como para los promedios móviles se construirán modelos automatizados utilizando herramientas informáticas.

4. Desarrollo del data mart

El desarrollo del data mart se realizará utilizando la metodología de Kimball, la cual se resume en la Figura 12.

Se excluyen del presente proyecto las etapas de Crecimiento y Mantenimiento, pues estas corresponden al ciclo de vida, pero por la naturaleza del proyecto, se realizará solo la primera iteración en el ciclo.

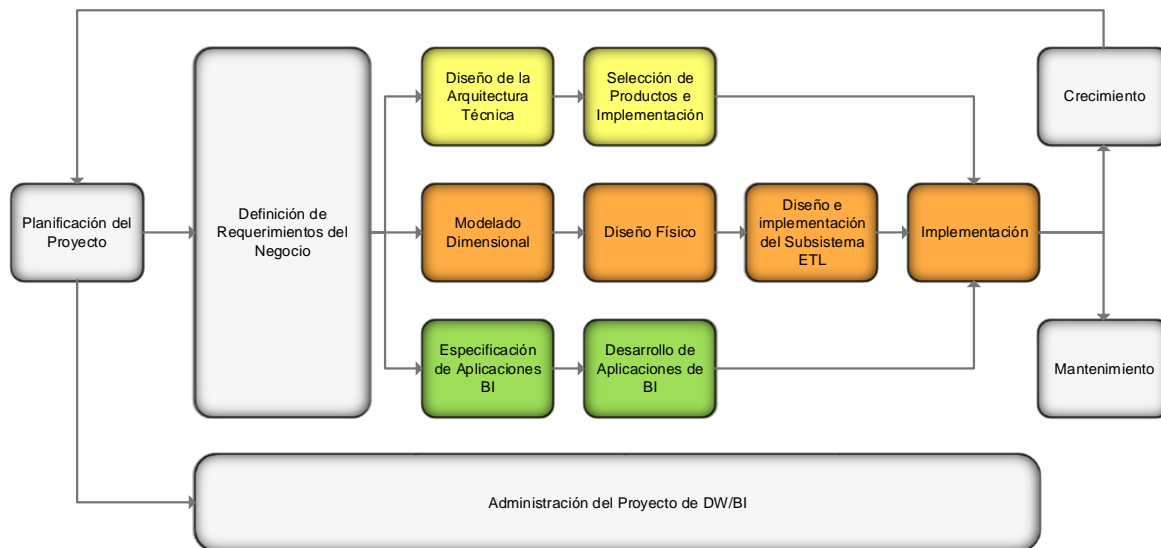


Figura 12. Metodología de Kimball (Kimball, 2009)

4.1. Planificación del proyecto

Propósito

Se establecerá un data mart sobre datos de energía y potencia del Sistema Eléctrico Nacional de Costa Rica. Este repositorio permitirá el análisis de desviaciones del predespacho nacional, realizar comparaciones y estudios sobre la generación eléctrica de diversos tipos de fuente. Será la base para posteriores estudios de minería de datos, por ejemplo, para la predicción de la demanda eléctrica nacional.

El proyecto es conocido internamente en la organización como Sistema de Análisis Histórico de Energía y Potencia (SAHEP).

Objetivos del data mart

OBJ-01: Reestructurar el análisis de los datos de generación y demanda eléctricas.

OBJ-02: Proveer una herramienta de reporteo ágil y flexible para los datos históricos del SEN.

Alcance

Consideraciones sobre el alcance del data mart:

- Se incluirán los datos históricos de 1991 en adelante, que son los datos con que se cuenta en los sistemas transaccionales.
- Se utilizará la plataforma tecnológica existente y disponible en el sitio. El acceso a los datos (cubos) para el usuario final será mediante el programa Microsoft Excel, principalmente por restricciones del negocio y por cumplir con los requisitos; esto será retomado con mayor detalle en la sección 4.7.
- El acceso será para un máximo de 10 usuarios, suficiente para las áreas del CENCE que realizan análisis sobre los mismos y para el área gerencial.

Exclusiones

- Las tareas de minería de datos relativas a la predicción de la demanda no forman parte del proyecto de construcción de un data mart. Este brindará el insumo necesario para el proceso de proyección, a saber, los datos históricos.
- Se realizará la implementación de los paquetes de los servicios de integración y Servicios de análisis de SQL Server, pero se excluye las ejecuciones automáticas de estos paquetes, la ejecución de los paquetes se realizará desde el ambiente de diseños.
- No se realizarán tareas de calidad de datos, dado que el sistema fuente tiene controles inmersos en sus procesos, los cuales ya proporcionan el grado de calidad requerido por la empresa.

Criterios de éxito

Varios criterios clave de éxito han sido definidos para el proyecto del SAHEP:

CRE-01: Proporcionar una fuente única para apoyar los análisis sobre demanda y generación eléctrica.

CRE-02: Reducir el tiempo necesario para realizar dichos análisis y los tiempos de entrega de reportes en la medida que sean solicitados.

CRE-03: Generación de reportes que permitan el uso de parámetros para su generación por parte del usuario. Se espera que el data mart permita la interacción con el usuario para generar los reportes utilizando diferentes criterios.

Riesgos

En la Tabla 1 se describe una serie de riesgos identificados que podrían afectar la consecución de los objetivos planteados:

Riesgo	Descripción	Probabilidad	Impacto	Acción de mitigación
RSG-01	El alcance del proyecto cambia.	media	alto	Mantener un estricto control del alcance y de las expectativas.
RSG-02	Falta de disponibilidad de los expertos del negocio.	media	medio	Establecer un cronograma de actividades y coordinar reuniones con antelación.
RSG-03	Plataforma tecnológica insuficiente.	baja	alto	Revisión previa de las capacidades de procesamiento y almacenamiento disponibles.
RSG-04	La solución no es capaz de satisfacer las expectativas del negocio.	media	alto	La presentación de un prototipo funcional al líder gerencial y expertos del negocio.
RSG-05	Pérdida de datos de código fuente.	baja	alto	La solución en desarrollo contará con respaldos diarios en la nube y en al menos dos medios físicos. La solución implantada estará respaldada por un servidor especializado para tal fin.
RSG-06	Falta de aceptación de la solución en la organización.	baja	alto	Una estrategia de comunicación y promoción de la solución. Involucramiento de personal clave durante el levantamiento de requerimientos e implantación de la solución.
RSG-07	Calidad de datos fuentes deficiente.	baja	media	Aplicación de reglas generales de negocio durante los procesos de carga para detectar posibles inconsistencias.

Tabla 1. Riesgos identificados (elaboración propia).

Beneficios

BNF-01 Mejor manipulación y entendimiento de los datos. Dentro del data mart los datos pueden ser sumados, promediados, otros., para satisfacer las necesidades específicas de diversas áreas de la organización. Dada la separación de los sistemas fuente, hay libertad para trabajar con los datos sin afectar el rendimiento de los sistemas relacionados.

BNF-02 Menor tiempo de respuesta para la generación de reportes. La flexibilidad de las herramientas y la facilidad de uso de los cubos de datos, permiten a los usuarios responder a las demandas y necesidades del negocio de manera más rápida y efectiva.

BNF-03 Desarrollo incremental. Los data marts pueden ser diseñados y construidos por separado para encajar en una estrategia de desarrollo incremental.

Roles

R01 - Líder Gerencial (LG): patrocinador del proyecto desde el punto de vista de la organización, el Centro Nacional de Control de Energía.

R02 - Líder del negocio (LN): proporciona guía de negocios y requerimientos para el equipo del proyecto y garantiza que la información capturada es adecuada y responde a las necesidades de la organización.

R03 - Líder de proyecto (LP): desde el punto de vista organizacional es la persona responsable de la coordinación del proyecto ante la gerencia. Esta persona se desempeña como administrador del área de sistemas de información del CENCE.

R04 - Experto del negocio (EN): persona conocedora del negocio y de las necesidades de información que debe cubrir el data mart.

R05 - Desarrollador de inteligencia de negocios (IN): encargado de analizar los requerimientos, diseñar, implementar y probar la solución de data mart.

R06 - Administrador del proyecto (AP): responsable de la gestión de tareas y actividades del proyecto, incluyendo la coordinación de recursos, el seguimiento de estado y la comunicación de los avances y problemas del proyecto, en estrecha colaboración con los líderes del proyecto y negocio.

R07- Usuario (US): persona que hará uso de la solución. También desempeñará un papel de pruebas durante el desarrollo del proyecto, interactuando con el experto del negocio para evaluar el cumplimiento de los requerimientos.

La lista de los recursos asociados a los roles descritos, se muestran en la Tabla 2.

Puesto en la organización	Roles						
	R01 LG	R02 LN	R03 LP	R04 EN	R05 IN	R06 AP	R07 US
Director del CENCE	✓	✓		✓			
Coordinador de Sistemas de Información		✓	✓				
Director del Área de PDE				✓			
Profesional en sistemas de potencia		✓		✓			
Profesional en informática				✓	✓	✓	
Profesional en informática					✓		
Colaborador de área de PDE o PCO							✓

Tabla 2. Lista de recursos (elaboración propia).

Costos

Los costos estimados se resumen en la Tabla 3, como se aprecia, los costos se resumen en horas, dado que los recursos materiales ya se encuentran disponibles y son asumidos por la organización y no por el proyecto.

Costos estimados de data mart SAHEP					
			Precio Dólar(¢)		550
Descripción	Unidad	Costo Unit.	Cant.	Total \$	Total ¢
Director del CENCE	Hora	80	4	320	¢176,000
Coordinador de Sistemas de Información	Hora	70	5	350	¢192,500
Director del Área de PDE	Hora	70	4	280	¢154,000
Profesional en sistemas de potencia	Hora	60	8	480	¢264,000
Profesional en informática (Inteligencia de Negocios)	Hora	60	80	4800	¢2,640,000
Colaborador de área de PDE o PCO	Hora	60	8	480	¢264,000
Costo total del proyecto				0	¢3,690,500

Tabla 3. Costos estimados (elaboración propia).

Plan del proyecto

El plan del proyecto se detalla en la Tabla 4, a modo de cronograma, en la cual se aprecia que la duración total del proyecto se extiende a cincuenta y tres días.

Nombre de tarea	Duración	Comienzo	Fin
Desarrollo del data mart	53 días	sáb 26/07/14	mar 16/09/14
Planificación del proyecto	5 días	sáb 26/07/14	mié 30/07/14
Definición de requerimientos del negocio	5 días	jue 31/07/14	lun 04/08/14
Informe de planificación y requerimientos	0 días	lun 04/08/14	lun 04/08/14
Tecnología	8 días	mar 05/08/14	mar 12/08/14
Diseño de la arquitectura técnica	3 días	mar 05/08/14	jue 07/08/14
Selección de productos e implementación	5 días	vie 08/08/14	mar 12/08/14
Datos	28 días	mar 05/08/14	lun 01/09/14
Modelado dimensional	3 días	mar 05/08/14	jue 07/08/14
Diseño físico	5 días	vie 08/08/14	mar 12/08/14
Diseño e implementación del subsistema de ETL	20 días	mié 13/08/14	lun 01/09/14
Aplicaciones de BI	7 días	mar 05/08/14	lun 11/08/14
Especificación de aplicaciones de BI	2 días	mar 05/08/14	mié 06/08/14
Desarrollo de aplicaciones de BI	5 días	jue 07/08/14	lun 11/08/14
Informe sobre solución del data mart	0 días	lun 01/09/14	lun 01/09/14
Implementación	15 días	mar 02/09/14	mar 16/09/14

Tabla 4. Plan del proyecto (elaboración propia).

4.2. Definición de requerimientos del negocio

El CENCE, como órgano esencial dentro de las funciones del ICE, en lo que respecta al Sistema Eléctrico Nacional, requiere generar, procesar y publicar diferentes tipos de información. Por ejemplo, algunos usuarios realizan consultas a entes como la Asamblea Legislativa, ARESEP, Ministerio de Energía, Gerencia del ICE y otros. A menudo la información es solicitada con carácter urgente. Actualmente, los usuarios tienen acceso a una serie de reportes preestablecidos a través de una página Web intranet. No obstante, estos reportes presentan limitaciones para profundizar con operaciones tipo “drill-down” sobre los datos, así como inflexibilidad para manipularlos con el objetivo de brindar explicaciones o justificaciones ante un evento ocurrido.

Con base en lo anterior, el desarrollo de un proyecto de inteligencia de negocios que proporcione a los usuarios mejores herramientas con un acceso más oportuno y flexible a los datos, potenciará sus capacidades analíticas sobre el comportamiento del Sistema Eléctrico Nacional. Esto se debe a que el usuario reducirá los tiempos de preparación de los datos y construcción de hojas de cálculo, permitiéndole dedicar más tiempo al análisis.

Además, un proyecto de esta índole ayudaría a mejorar los tiempos de respuesta a las diversas solicitudes de información, a su vez, la calidad de los reportes que los usuarios generen.

Por tanto, la creación de un data mart como parte de un proyecto de inteligencia de negocios con datos de energía y potencia sería el primer paso para optimizar el análisis de la información dentro del CENCE, para lo cual se consideran los siguientes requerimientos:

- REQ-01: proporcionar los datos tanto del predespacho como del posdespacho nacional para un periodo específico.
- REQ-02: Soportar comparaciones de predespacho contra posdespacho. Identificar las desviaciones encontradas con propósito de mejorarlas o justificarlas.
- REQ-03: Proveer capacidad para identificar el aporte de cada fuente a la satisfacción de la demanda total. Por ejemplo: ¿qué tanto aportó el térmico con respecto a la demanda total?
- REQ-04: Proveer capacidad para identificar la distribución del aporte de cada empresa a la satisfacción de demanda total durante un periodo determinado. De igual modo, se pretende obtener esta misma información por planta o por tipo de recurso (verde o térmico).
- REQ-05: Proveer capacidad para reportar la potencia durante un periodo determinado, agrupado por empresa, por planta, por fuente o por tipo de recurso (verde o térmico).

- REQ-06: Proveer capacidad de brindar los niveles de embalses por día para realizar el seguimiento de los niveles a través del tiempo.
- REQ-07: Relacionar la información del posdespacho y los niveles de agua de los embalses, con el fin de identificar y explicar eventos.
- REQ-08: Proveer capacidad de reportar la generación por diferentes periodos: la generación para un día específico, un mes específico, la generación de un año en particular, aplicando diferentes criterios de agrupamientos (agrupado por trimestre, por mes, por semana).

En la Figura 13 se muestra la priorización de los requerimientos según su impacto y viabilidad, como se aprecia, todos los requerimientos se mantienen en el cuadrante con una viabilidad de media a alta y el impacto igualmente de medio a alto, por lo que en la implementación se cubrirán todos los requerimientos.

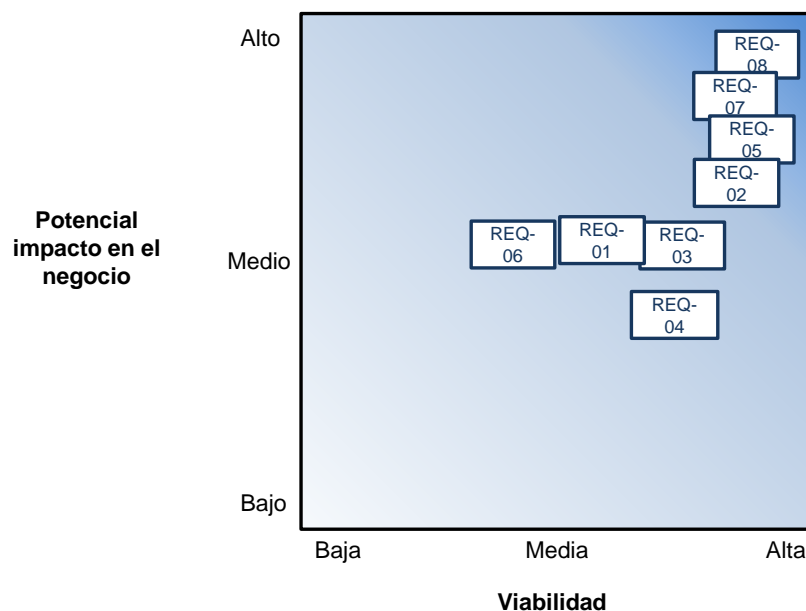


Figura 13. Priorización según impacto y viabilidad (elaboración propia)

4.3. Diseño de la arquitectura técnica

Las características propias del ambiente en el cual se desarrolla el proyecto determinan el procedimiento de diseño arquitectural. Se trata, entonces de conocer si la infraestructura disponible es suficiente para satisfacer las

necesidades de procesamiento, almacenamiento y comunicación, dado que no es factible la adquisición de equipos o herramientas de software. De igual manera, se identifica si la plataforma de software cumple o no con lo requerido.

Requerimientos generales de arquitectura

Se requiere que la carga de datos se realice diariamente. Es decir, que la carga de datos de un día esté disponible al día siguiente.

Adicionalmente existe un proceso de cierre mensual, ahí se efectúa una validación de datos y se actualiza con información proveniente de otras fuentes que actualizan y oficializan los datos recopilados diariamente. Se requiere que al cierre de este proceso se cuente con la carga del mes correspondiente.

Será necesario al menos 30 Gb de almacenamiento para la carga inicial de veintitrés años de información histórica y se prevé un crecimiento de 1.5 GB por año.

Por otra parte, se requiere también que la plataforma soporte diez usuarios analíticos concurrentes.

Infraestructura física

Se dispone de un servidor Dell PowerEdge R2900 con Windows Server 2008 R2 como sistema operativo y SQL Server 2012 como servidor dedicado a la inteligencia de negocios.

El mismo cuenta con dos procesadores Intel Xeon X5260 de dos núcleos a 3.33 GHz y 12 Gb de memoria RAM, que soportará las tareas de extracción, transformación y carga, así como las labores de análisis y minería. También alojará el data mart y los cubos de datos. Se encuentra conectado a una unidad DAS para el almacenamiento con alta disponibilidad y con 4.8 TB disponibles con discos de 300 Gb de 15k rpm. Además, se encuentra conectado a una red Ethernet de 1 Gb/s.

La infraestructura física disponible satisface las necesidades y cumple con los requerimientos de desempeño para la solución propuesta.

Diagrama de diseño de arquitectura técnica

Una vista general del flujo de los datos y de la arquitectura propuesta para una plataforma de inteligencia de negocios se muestra en la Figura 14:

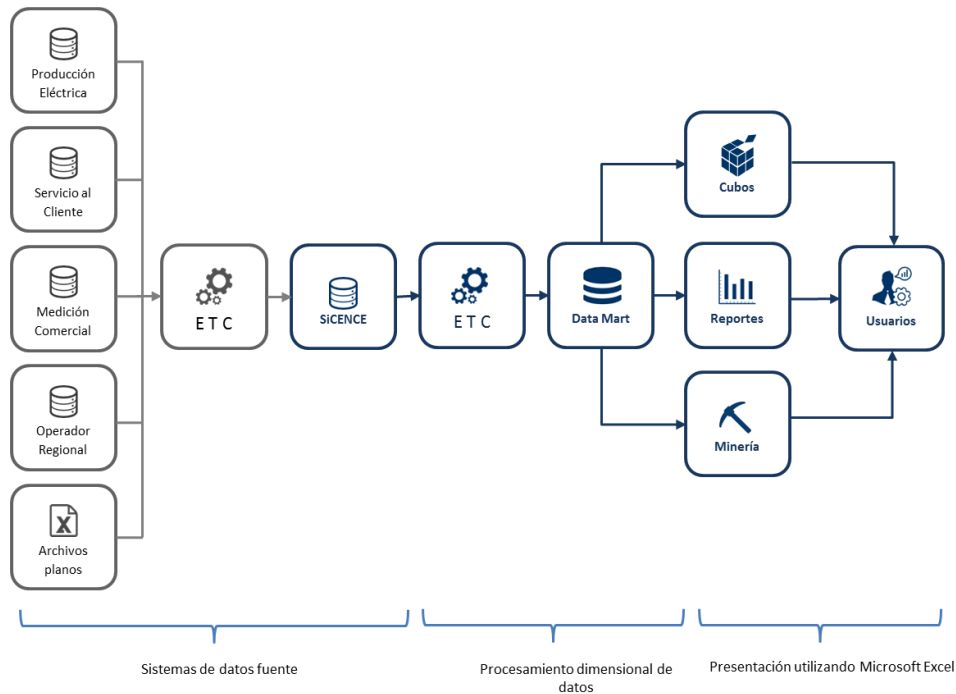


Figura 14. Diagrama de diseño de arquitectura técnica (elaboración propia)

4.4. Selección de productos e implementación

Selección de productos

Por una restricción del negocio no es posible la adquisición de nuevas herramientas. Esto reduce esta etapa de selección a una evaluación del software y hardware disponibles para verificar si cuentan con las características necesarias. La Tabla 5 muestra una serie de características necesarias para el proyecto y su cumplimiento con la herramienta disponible, Microsoft SQL Server 2012. Como se puede observar, satisface los requerimientos. La creación de prototipos ha permitido corroborar las capacidades de la herramienta y que constituye una opción segura.

Característica	Cumplimiento
Extracción de diversas fuentes de datos, particularmente con facilidades para SQL Server.	✓
Capacidades para copia y replicación de datos.	✓
Soporta funciones de validación, conversión, derivación, agregación.	✓
Soporta la carga de datos vía update, append.	✓
Soporta la calendarización de trabajos.	✓
Monitoreo, reporte y recuperación de tareas no completadas.	✓
Seguridad a nivel de usuarios.	✓
Soporte para gestión de metadatos.	✓
Facilidad de uso para usuario final e integración con herramientas existentes en la organización (Excel)	✓
Capacidades OLAP, con buen rendimiento y potencia para manejar los modelos multi-dimensionales, tareas como “slice and dice”, ordenamientos y consultar cubos.	✓
Soporte para minería de datos	✓
Generación de reportes ad hoc, con formato visual agradable, parametrizables.	✓
Calidad y consistencia general del producto aceptables.	✓

Tabla 5 - Cumplimiento de requerimientos del producto para desarrollar solución de inteligencia de negocios (elaboración propia).

Actualmente, y según la firma Gartner, Microsoft es uno de los líderes en el desarrollo de tecnologías de inteligencia de negocios, tal como se muestra en el llamado cuadrante mágico correspondiente al mes de octubre del presente año mostrado en la Figura 15:



Figura 15. Cuadrante mágico de Gartner para octubre, 2014. (Gartner, 2014)

La plataforma para inteligencia de negocios de SQL Server descrita en términos generales por (Root y Mason, 2012) y se presenta en la Figura 16:

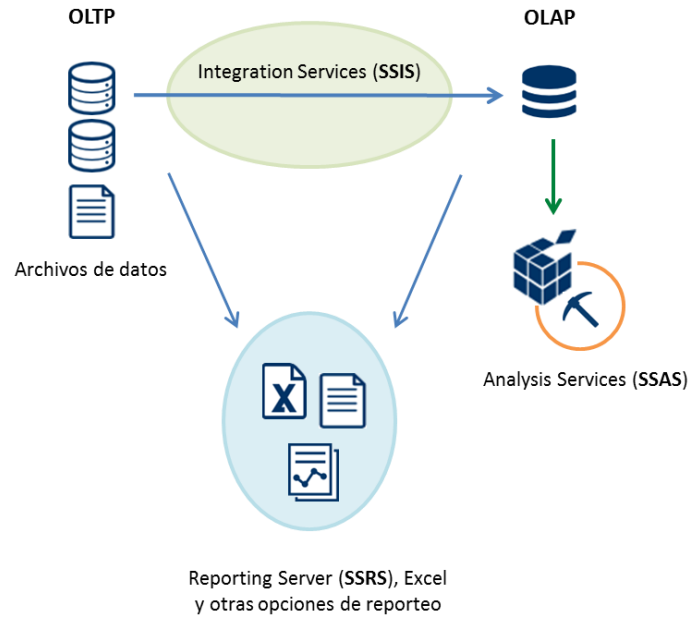


Figura 16. Esquema de una solución típica de inteligencia de negocios utilizando SQL Server (elaboración propia).

Algunos componentes importantes de la arquitectura Microsoft para inteligencia de negocios que serán utilizados en el proyecto o de los cuales podrían incorporarse algunas de sus características posteriormente son: SQL Server Integration Services, Analysis Services y Reporting Services.

SQL Server Integration Services

Los servicios de integración de Microsoft constituyen una plataforma para construir soluciones de integración y transformación de datos. Permite cargar y actualizar almacenes de datos, limpiar datos y realizar minería. Los llamados “paquetes” de Integration Services pueden extraer y transformar datos de diversos tipos de fuentes, incluyendo archivos planos, XML y fuentes de datos relacionales.

SQL Server Analysis Services

Es la tecnología Microsoft, dentro de su oferta para inteligencia de negocios, para desarrollar soluciones de procesamiento analítico en línea (OLAP) y minería de datos. Permite diseñar, crear y administrar estructuras multidimensionales que contienen datos agregados desde otros orígenes de datos, como bases de datos relacionales. En el caso de las aplicaciones de minería de datos, Analysis Services permite diseñar, crear y visualizar modelos de

minería con soporte a una variedad importante de algoritmos de minería de datos estándar.

SQL Server Reporting Services

Reporting Services es una plataforma de reportes (o informes) que proporciona una funcionalidad bastante completa. Incluye el diseño, desarrollo, pruebas, implementación y la administración de los reportes. Ofrece múltiples formatos de visualización y permite exportar informes a otras aplicaciones, como Microsoft Excel.

Implementación

De igual forma que en la selección de herramientas, no existe en este caso, como proceso formal, la implementación de tales herramientas, no se ejecuta propiamente, sino que más bien se asume el ambiente ya establecido con sus ventajas y limitaciones. Ya se ha probado que las capacidades de las herramientas de software disponibles son suficientes y suplen con creces la necesidad.

De la misma forma, las capacidades del servidor físico disponible no sólo cumplen con los requisitos mínimos descritos por el fabricante del software sino que ya ha sido objeto de pruebas con prototipos.

No existe, por tanto, una etapa de selección de herramientas ni tampoco una de implementación porque se está asumiendo un ambiente ya establecido por completo. Corresponde entonces una evaluación del mismo, que en este caso arroja resultados positivos. La base de software y hardware es adecuada para la implementación de un data mart de energía y potencia para la organización.

4.5. Modelado dimensional

Los procesos de negocio que se involucran en el modelo dimensional son:

Predespacho: Este proceso se encarga de realizar la proyección de la demanda de energía horaria que se requiere a nivel nacional, luego, con esta proyección, se revisan los recursos disponibles para generar una proyección de lo que debe generar cada elemento del sistema (planta o unidad generadora) y así, conjuntamente se pueda satisfacer la demanda nacional. Para este proceso, el

modelado se limita a mantener el registro de la generación horaria, que es el insumo para todo el proceso de proyección.

Posdespacho: Proceso empleado para la publicación de la información relacionada con el comportamiento real del sistema eléctrico, así mismo como las comparaciones del predespacho con respecto a estos datos reales.

Definidos los procesos involucrados se debe iniciar el modelado dimensional, el cual es una técnica de diseño lógico para estructurar los datos de una manera intuitiva y entendible para los usuarios de negocios y que además ofrece un alto rendimiento para las consultas.

Cómo punto de partida se debe revisar los requerimientos identificados. Luego, realizar un estudio del negocio o proceso por modelar, para comprender los datos con los cuales se trabajarán. Una herramienta importante de este proceso es la creación de una tabla en donde se identifiquen los datos para poder trabajar con sus respectivas descripciones.

La primera herramienta por utilizar es la creación de una tabla en donde se incluyan los elementos claves del modelo lógico; además de atributos físicos. Esta información se obtiene del modelo de datos del sistema fuente, que en este caso se le denomina siCENCE.

El diagrama de entidad relación presente en la Figura 17 muestra un subconjunto de las tablas que se involucran en este proyecto. Para efectos de simplicidad, se han excluido las tablas que no contienen información relacionada con los alcances definidos para el presente proyecto.

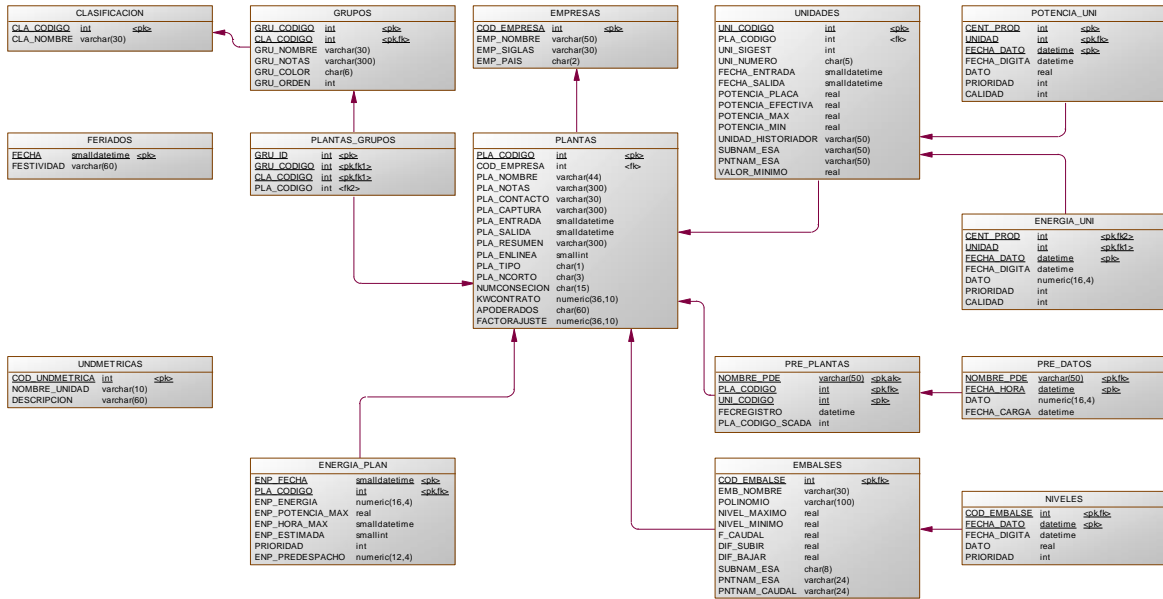


Figura 17. Diagrama de entidad relación (elaboración propia)

El resultado de identificar estos elementos claves, se muestra en el Apéndice 1.

Una vez definidos los elementos de interés, y basándose siempre en los requerimientos, se procedió a confeccionar la matriz de bus, tal como se aprecia en la Tabla 6, que será un insumo importante para el proceso posterior.

Proceso del Negocio	Empresa	Escenario	Fecha	Fuente	Hora	Plantas	Orden de despacho	Tipo de recurso	Unidades de medida	Región
Predespacho		✓	✓		✓				✓	
Postdespacho	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Tabla 6. Matriz de bus (elaboración propia).

Proceso de modelado en cuatro pasos

La creación del modelo multidimensional es un proceso altamente iterativo y dinámico. Después de unos pocos pasos de la preparación, el esfuerzo de diseño comienza con un modelo gráfico inicial derivado de la matriz del bus, se identifica el ámbito del diseño y la granularidad de las tablas de hechos propuestos. Las sesiones de diseño iniciales también deben identificar las

principales dimensiones aplicables a cada tabla de hechos, una propuesta de lista de atributos de cada dimensión y cualquier cuestión que requiera investigación adicional.

La matriz de bus identifica los procesos o elementos de interés del negocio y sus dimensiones asociadas; así se procede con los cuatro pasos siguientes:

Paso 1. Escoger el proceso del negocio

El primer paso es elegir el proceso de negocio por modelar. Esta elección en realidad ya ha sido definida a partir de los requerimientos y objetivos del proyecto, dado que normalmente se produce después de haber reunido los requisitos de negocio de alto nivel. Esta selección corresponde comúnmente a escoger una fila en la matriz de bus.

Paso 2. Declarar la granularidad

El segundo paso para crear el modelo multidimensional es declarar la granularidad, o el nivel de detalle en la tabla de hechos para el proceso de negocio seleccionado. Declarar la granularidad significa definir exactamente lo que representa una medición en la tabla de hechos. La granularidad debe ser expresada en términos del negocio.

Paso 3. Identificar las dimensiones

El tercer paso es determinar las dimensiones aplicables a la tabla de hechos según el nivel declarado de granularidad. La mayoría de las principales dimensiones caerán de forma natural una vez que se haya determinado esta granularidad. Todas las dimensiones de la matriz de bus deben probarse contra la granularidad para ver si encajan. Cualquier dimensión que conduce a un solo valor en la tabla de hechos, con la granularidad definida, es un candidato viable.

Paso 4. Identificar los datos

El paso final en el proceso de modelado es identificar los hechos o medidas de los procesos de negocio. Cuando se declara la granularidad también se cristaliza la discusión sobre los hechos numéricos medidos. En otras palabras, los hechos deben ser fieles al grano. Se debe evitar la tentación de añadir hechos que no coincidan con la granularidad de la tabla de hechos, ya que por lo general presentan complejidades y errores en las aplicaciones de BI.

Como resultado de la aplicación de estos pasos sobre la matriz de bus se obtiene:

Proceso: Predespacho

Granularidad: los datos están registrados en una fila por hora y para cada posible escenario, a saber, el predespacho (dato programado) o posdespacho (dato real).

Dimensiones. Las dimensiones involucradas son:

- Escenario
- Fecha
- Hora
- Unidades de medida

Hechos: la tabla de hecho registra el dato de energía demandada cada hora por el sistema.

Con este análisis se obtiene el diagrama en alto nivel (ver Figura 18).

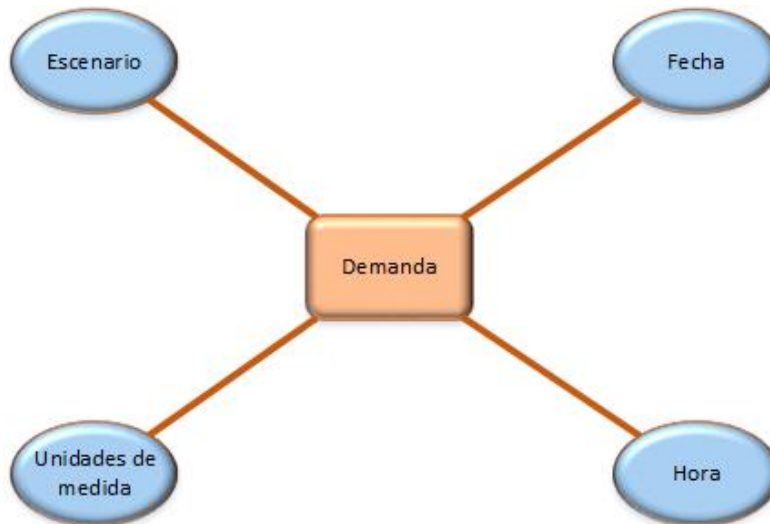


Figura 18. Diagrama de alto nivel para predespacho (elaboración propia).

Proceso: Posdespacho

Granularidad: en este proceso se identifican varias granularidades de acuerdo con los datos que se tienen, las cuales se detallan a continuación:

Para los niveles, se tiene una fila por hora y por cada planta que posee un embalse asociado.

Con respecto a la energía se obtiene un dato diario para cada planta y para cada posible escenario.

Para la energía obtenida a partir del SCADA, se cuenta con un dato por hora para cada planta y para cada escenario.

Finalmente, para la potencia, igualmente obtenida desde el SCADA, se emplea un dato por planta cada 15 minutos e igualmente se prevé para cada escenario.

Dimensiones. Las dimensiones involucradas son:

- Empresa
- Escenario
- Fecha
- Fuente
- Hora
- Plantas
- Orden de despacho
- Tipo de recurso
- Unidades de medida
- Región

Hechos: cuando se definió la granularidad, se presentaron diferentes granularidades para diferentes datos, así las dimensiones utilizadas para cada dato o hecho, impulsa la creación de tablas de hechos disímiles, es decir, si un hecho utiliza las dimensiones a, b y c, esta tabla debe ser diferente a otra que utilice las dimensiones a, d y e.

Particularmente para los datos provenientes del SCADA, se tiene que existen dos hechos que están utilizando las mismas dimensiones, pero haciendo referencia al diseño de modelos de estrella de Adamson, cuando dos o más hechos describen eventos que no tienen lugar al mismo tiempo, están describiendo procesos diferentes. Si se colocan en una sola tabla de hechos, se verá obstaculizado el análisis de los procesos individuales. La colocación de ellos en tablas de hechos separados permite que cada proceso se pueda estudiar con mayor facilidad (Adamson, 2010). Esto es precisamente lo que sucede con los datos provenientes del SCADA: la energía y la potencia, aunque tienen las mismas dimensiones, son en realidad datos diferentes y no tienen lugar al mismo tiempo; uno sucede cada hora, en tanto que el otro sucede cada 15 minutos, por esta razón se procedió a separar las tablas de hechos diferentes.

Así se obtienen los siguientes datos o hechos:

- Niveles: Nivel del embalse registrado por hora para cada planta con un embalse asociado.
- Energía: Dato de energía diaria generada para cada planta del sistema.
- Energía SCADA: Dato de energía horaria obtenida desde el sistema SCADA, cabe señalar que la mayoría de las plantas tienen medición desde el sistema SCADA, pero no todas.
- Potencia SCADA: Dato de potencia cada 15 minutos obtenido desde el sistema SCADA.

A partir de esta información, se obtienen los diagramas de alto nivel, según se muestran de la Figura 19 a la Figura 22.



Figura 19. Diagrama de alto nivel para la potencia según SCADA (elaboración propia).



Figura 20. Diagrama de alto nivel para la energía diaria (elaboración propia).



Figura 21. Diagrama de alto nivel para la energía según SCADA (elaboración propia).

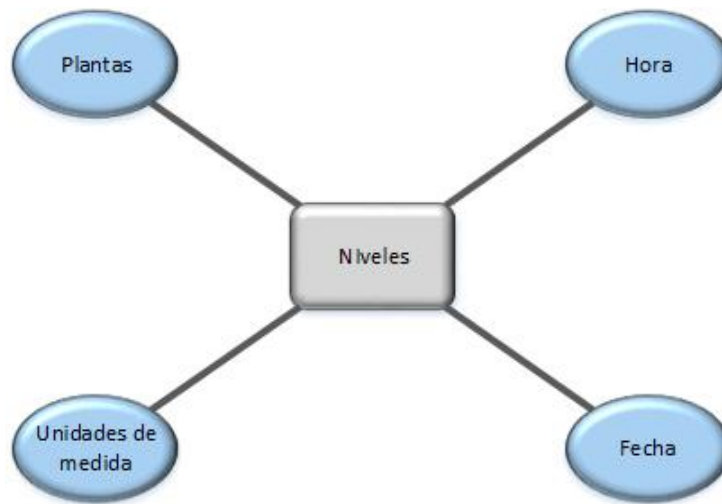


Figura 22. Diagrama de alto nivel para los niveles (elaboración propia).

Una vez identificadas las dimensiones y los hechos, se actualiza la matriz de bus para que refleje los hallazgos. Es importante mantener esta matriz actualizada porque constituye un instrumento de comunicación y planificación clave que será revisado por los promotores de proyectos y demás interesados, como diseñadores, administradores y usuarios de negocios.

La matriz actualizada se muestra en la Tabla 7:

Proceso del Negocio	Tabla de hechos	Granularidad	Hechos	Empresa	Escenario	Fecha	Fuente	Hora	Plantas	Orden de despacho	Tipo de recurso	Unidades de medida	Región
Pre despacho	Demanda	1 fila por hora	Demanda		✓	✓		✓				✓	
Postdespacho	Niveles	1 fila por hora por planta	Niveles			✓		✓	✓			✓	
	Energía	1 fila por día por planta	Energía diaria	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Energía SCADA	1 fila por hora	Energía según SCADA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Potencia SCADA	1 fila cada 15 minutos	Potencia según SCADA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Tabla 7 - Matriz de bus actualizada (elaboración propia).

Finalmente, para completar el modelado dimensional se definieron las hojas de trabajo detallado con el diseño dimensional para cada tabla de hechos y cada tabla de dimensiones. En la Tabla 8 se muestra la hoja para la dimensión de la empresa. El conjunto de tablas se adjunta en el Apéndice 2.

Nombre de la Tabla: TD_EMPRESA
 Tipo: Dimensión
 Descripción: Empresas con participación en la generación de electricidad

Columna	Descripción	Tipo	Destino			Ejemplo	Fuente						
			Tamaño	Llave	FK		Sistema	Tabla	Columna	Tipo	Comentario		
IDEMPRESA	Identificador de la empresa	entero		PK	1	ETL							Nueva llave
NOMEMPRESA	Nombre de la empresa	texto	50			Instituto Costarricense de Electricidad	siCENCE	GRUPOS	GRU_NOMBRE	varchar(30)			Se filtran con la clasificación igual a 3
ORDEN	Orden para mostrar las empresas	entero			1		siCENCE	GRUPOS	GRU_ORDEN	int			
COLOR	Color para referenciar esta empresa	texto	6			FFFFFF	siCENCE	GRUPOS	GRU_COLOR	varchar(6)			

Tabla 8 - Hoja de trabajo para la dimensión de empresa (elaboración propia).

4.6. Diseño físico

Para desarrollar el diseño físico se tomó el procedimiento propuesto por Kimball (Kimball, 2009), que se resume en el diagrama de la Figura 23, y se desarrollaron las etapas hasta el servidor de base de datos de desarrollo.

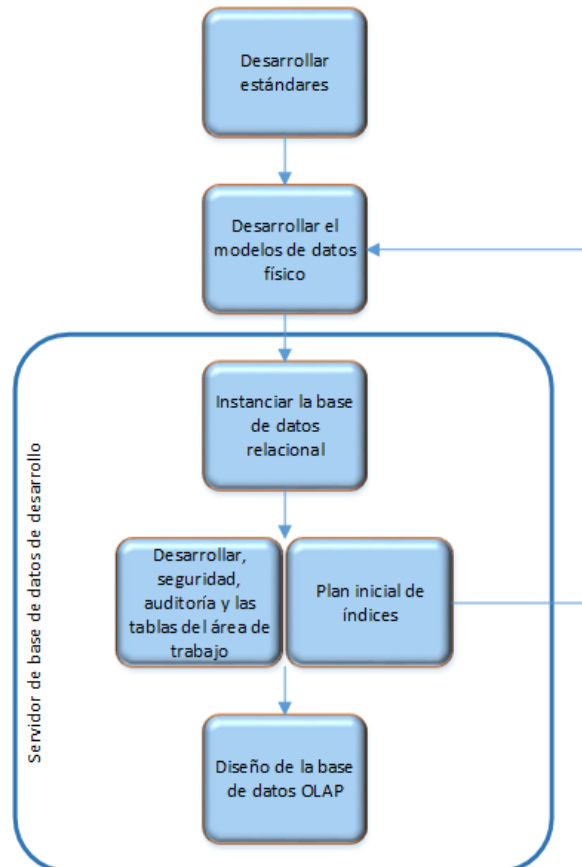


Figura 23. Procedimiento para diseño físico según Kimball (elaboración propia).

Desarrollo de estándares

Para el proyecto se tomó la decisión de seguir un estándar sencillo que no involucrara un nuevo paradigma con respecto al que se ha seguido en la empresa.

El conjunto de reglas utilizadas es el siguiente:

- El nombre de las tablas utilizarían los siguientes prefijos:
 - TD_ para las tablas de dimensiones.
 - TH_ para las tablas de hechos.

Cada tabla se completaría con el nombre de la dimensión o el hecho correspondiente, preferiblemente empleando el mismo nombre utilizado por

el sistema fuente, para crear una analogía fácil entre el data mart y el sistema fuente. Por ejemplo, en el sistema fuente la tabla que contiene el conjunto de plantas se denomina PLANTAS, entonces la dimensión correspondiente sería TD_PLANTAS.

- Para los identificadores o llaves se utilizará el prefijo ID, por ejemplo el identificador de una planta será IDPlanta.
- En las dimensiones, para definir el nombre de la dimensión se utilizará el prefijo nom, así por ejemplo el nombre de una planta será nomPlanta.
- En la dimensión de fecha hay varios elementos que tienen una representación numérica y otra de texto, entonces para el campo numérico éste se nombrará utilizando solamente el nombre de la entidad y su representación de texto se antecederá una *n* al nombre de la entidad. Por ejemplo, el campo Mes contiene el entero que corresponde al número del mes y el campo nMes contiene el texto correspondiente al nombre del mes.

Desarrollo del modelo de datos físicos

Con base en el modelo dimensional, se procedió a la creación del modelo físico de la base de datos, el cual consta de cinco tablas de hechos y de diez dimensiones.

El diagrama completo del modelo se ilustra en la Figura 24:

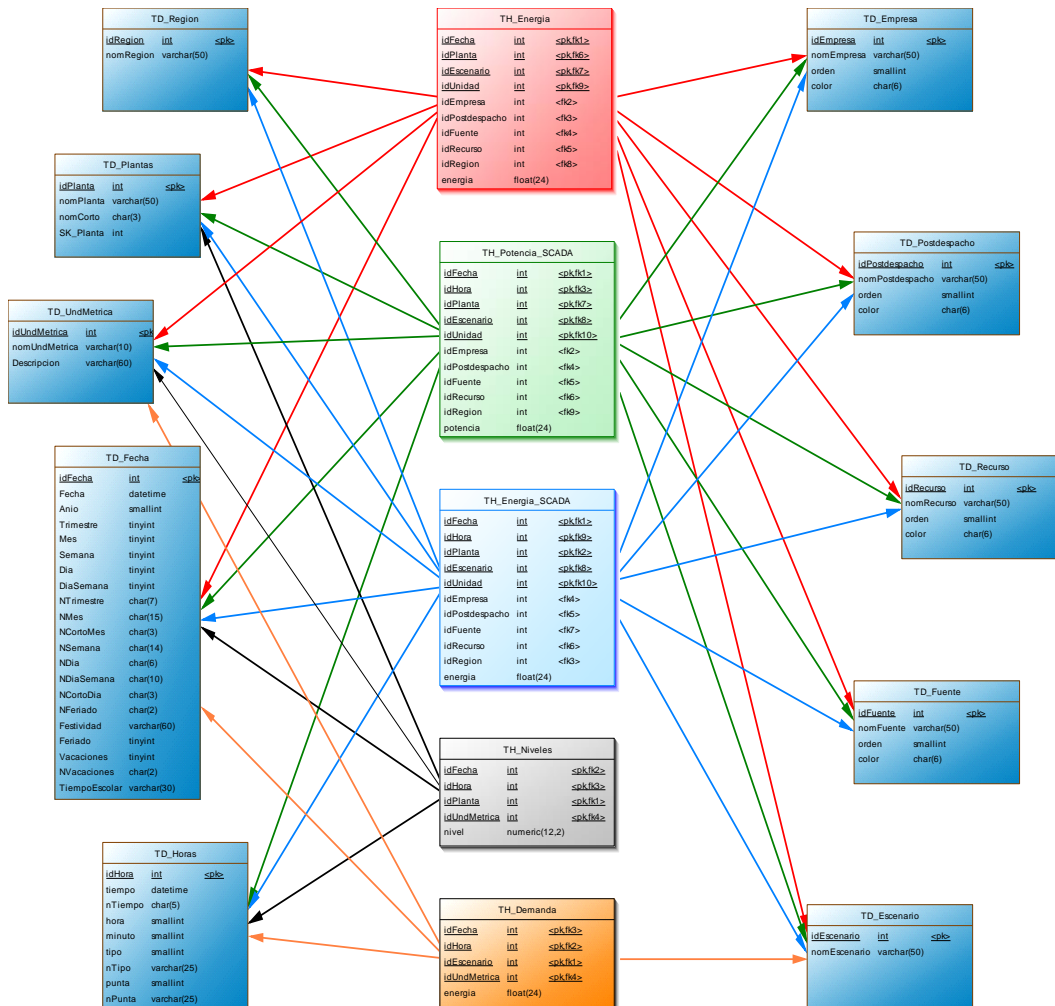


Figura 24. Modelo físico de la base de datos (elaboración propia).

El diccionario de datos para el modelo se adjunta en el Apéndice 2 con la descripción de estas tablas.

Detalle de las tablas. Cabe señalar que para las tablas de hechos, el campo correspondiente a la potencia o energía, se utilizó en este modelo un campo de tipo punto flotante, pese a que en el sistema original se utiliza un campo numérico. Esto con el fin de ayudar en la conversión de unidades y evitar los problemas de redondeo.

Instanciación de la base de datos relacional

Para instanciar esta base de datos se utilizó SQL Server, se emplearon los tipos de datos definidos en este motor y siguiendo los estándares de

almacenamiento de la empresa. A grosso modo se utiliza un DAS con un arreglo de discos, en el cual se definen varias unidades con la siguiente funcionalidad:

- Una unidad para el TempDB.
- Dos unidades para distribuir los datos.
- Una unidad para el Transaction Log.

Desarrollo de aspectos de seguridad y rendimiento

En esta etapa se procede a la creación del área de trabajo (*Staging Area*), para lo cual se realizó una copia de las estructuras de las tablas involucradas en este proceso, para que éstas reciban una copia de los datos que se procesarán; de esta forma, solamente se hará una consulta al sistema fuente para traer los datos. Luego estos serán procesados por otro servidor para no agregarle más carga de trabajo al sistema en producción.

Esta base de datos se denomina CENCE_SA.

En este repositorio se agregó una tabla FECHAS_CARGA, la cual contiene dos fechas que definen el intervalo de la información por procesar, así cuando se va a realizar una carga de datos, lo primero que se debe hacer es actualizar esta tabla con las fechas correspondientes, pues la consulta a la base de datos fuente se efectuará con estos parámetros.

Con respecto a la seguridad, básicamente se utilizaron dos usuarios particulares:

Consulta: ese usuario ya está definido en el sistema fuente siCENCE, con los permisos restringidos para realizar solamente consultas sobre esta base de datos, de modo que se utilizará éste para realizar la conexión al sistema fuente.

AgenteDM: este usuario se definió en las bases de datos del área de trabajo y en la base de datos multidimensional y cuenta con los permisos para realizar los trabajos necesarios en estas; es decir, para la ejecución de los ETL y la actualización respectiva.

Con respecto a los índices, el plan inicial es la utilización de las llaves primarias, las cuales fueron definidas en el diseño físico.

Para el caso de las tablas de dimensiones, se definió para cada tabla un identificador, declarado como llave primaria, con lo que se define un índice agrupado para estas tablas.

Luego, para las tablas de hechos, como regla general, Kimball propone definir la llave principal de las tablas de hechos como una llave compuesta. Consiste en el conjunto de llaves foráneas de las dimensiones que definen de forma unívoca una fila en la tabla de hechos. Esta regla fue seguida para las cinco tablas de hechos.

Con estas consideraciones, los índices quedaron definidos como se aprecia en la Tabla 9:

Tabla	Índice				
TD_EMPRESA	IDEMPRESA				
TD_ESCENARIO	IDESCENARIO				
TD_FECHA	IDFECHA				
TD_FUENTE	IDFUENTE				
TD_HORAS	ID_HORA				
TD_PLANTAS	IDPLANTA				
TD_POSTDESPACHO	IDPOSTDESPACHO				
TD_RECURSO	IDRECURSO				
TD_REGION	IDREGION				
TD_UNDMETRICA	IDUNIDAD				
TH_DEMANDA	IDFECHA	ID_HORA	IDESCENARIO	IDUNIDAD	
TH_ENERGIA	IDFECHA	IDPLANTA	IDESCENARIO	IDUNIDAD	
TH_ENERGIA_SCADA	IDFECHA	ID_HORA	IDPLANTA	IDESCENARIO	IDUNIDAD
TH_NIVELES	IDFECHA	ID_HORA	IDPLANTA	IDUNIDAD	
TH_POTENCIA_SCADA	IDFECHA	ID_HORA	IDPLANTA	IDESCENARIO	IDUNIDAD

Tabla 9 - Definición de índices para tabla de hechos (elaboración propia).

Conforme se vaya utilizando el data mart, es necesario realizar revisiones periódicas de los índices para validar su utilidad y verificar si es necesario su actualización.

Diseño de la base de datos OLAP

Con base en el modelo físico y lógico, se procedió a la creación de un cubo incluyendo las cinco tablas de hechos y las diez dimensiones, esto permite

relacionar los diferentes hechos cuando así se desee. El cubo resultante se muestra en el diagrama de la Figura 25:

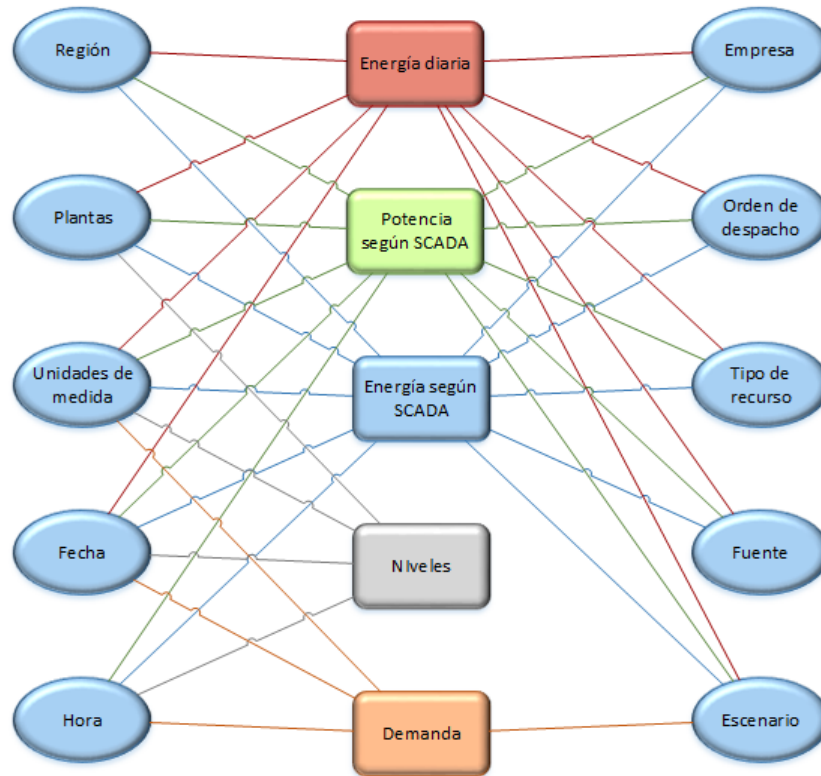


Figura 25. Diagrama del cubo (elaboración propia).

Jerarquías definidas

En la dimensión de fecha se crearon dos jerarquías, dado que existen dos formas comunes de realizar las consultas de fechas. A saber: considerando las semanas y los meses. De tal forma las dos jerarquías definidas son:

Semanal: Día de la semana -> Número de semana -> Año

Mensual: Día -> Mes -> Trimestre -> Año

Agregaciones

Con respecto a las agregaciones, según la naturaleza del dato se tiene:

Para el caso de datos de energía, las agregaciones son sumas, pues la energía de un período de tiempo es la suma de los intervalos que componen ese período, así la energía de un mes, por ejemplo, es la suma de la energía de los días del mes.

Con respecto a la potencia, las agregaciones van en función del promedio, pues la potencia es un dato instantáneo en el tiempo. La suma de estos datos no tiene sentido.

Finalmente, los niveles son también mediciones en un momento dado de cómo se encontraba el nivel del embalse. Para este caso es de interés obtener los niveles máximos, mínimos y promedio, para un período definido, para así observar su comportamiento estadístico. Además, el primero, que sería cómo estaba el nivel al inicio del período de estudio, que precisamente este es el dato más utilizado en los reportes.

Así las agregaciones quedan resumidas en la Tabla 10:

Tabla	Agregación
TH_DEMANDA	Suma
TH_ENERGIA	Suma
TH_ENERGIA_SCADA	Suma
TH_NIVELES	Mínimo, Máximo, Promedio, Primero
TH_POTENCIA_SCADA	Promedio

Tabla 10. Agregaciones (elaboración propia).

4.7. Diseño e implementación del subsistema de ETL

Antes de comenzar el diseño del sistema ETL para un modelo dimensional, se debe haber completado el diseño lógico, contar con el plan de arquitectura de alto nivel y el mapeo fuente-destino para todos los elementos de datos.

Seguidamente se propone seguir los diez pasos propuestos por Kimball para diseñar y desarrollar los ETL:

Paso 1. Dibujar el plan de alto nivel

El proceso de diseño se inicia con un esquema simple de las piezas del plan que se conoce: fuentes y destinos. El diagrama se debe mantener en un muy alto nivel, destacando las fuentes de datos y la anotación de los principales retos que se identifican inicialmente. El diagrama propuesto se muestra en la Figura 26:

Fuentes: Base de datos siCENCE

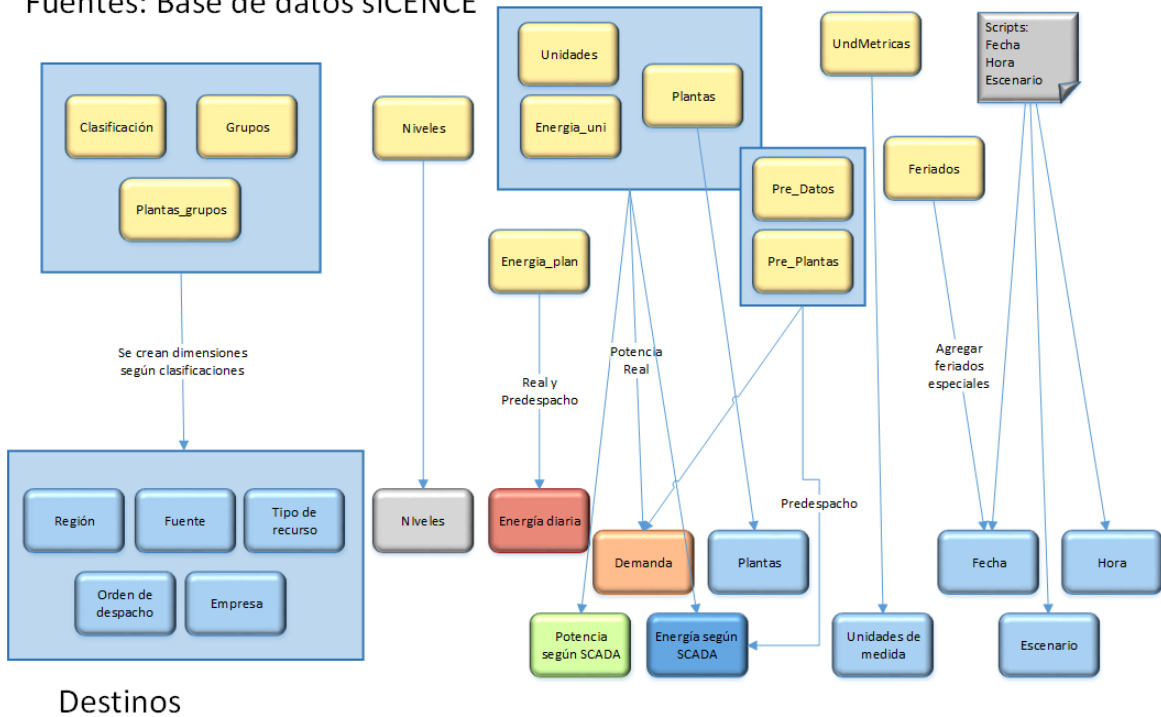


Figura 26. Plan de alto nivel (elaboración propia).

Paso 2. Elegir la herramienta ETL

Este paso, como se mencionó anteriormente en la arquitectura, es impuesto por restricciones del negocio, pues la herramienta por utilizar debe ser SQL Server, pero igualmente, cumple con los requerimientos para el proceso de ETL.

Paso 3. Desarrollar estrategias predeterminadas

Las estrategias definidas en este paso son las siguientes:

La extracción de los sistemas fuentes: se realizará una consulta lo más simple posible para evitar cargar al sistema fuente, esta consulta será básicamente SELECT y como máximo debe llevar una cláusula WHERE para restringir los datos por fecha. Los datos obtenidos se copiarán directamente al área de trabajo. Se tiene la ventaja que la base de datos fuente está en SQL Server, por lo que las consultas serán con un driver nativo.

Manejo de cambio de atributos de dimensiones: la propuesta es tratarlo como tipo 1, pero por las características del negocio, estos cambios se asume que en realidad no sucederán, esto por cuanto el nombre de las plantas no cambian,

así como sus fuentes y otras características. En caso extremo que cambie algún atributo, esto involucra en realidad que se trate como un nuevo registro, por ejemplo, si cambia la fuente, se tratará de una nueva planta pues físicamente debió cambiar toda su infraestructura, o si cambia la empresa, como ya ha sucedido, esto cambia la naturaleza de la planta, su contrato o forma de despachar, por lo que en realidad se considera en el sistema como si se tratara de una planta diferente.

Organizar el área de trabajo: para el área de trabajo, se crearon estructuras similares a las tablas del sistema fuente, con la salvedad de que se le eliminó toda la integridad referencial, esto por aspectos de rendimiento.

Calidad de los datos: con respecto a la calidad de los datos, el sistema fuente contiene controles para asegurar una adecuada calidad, por lo que los datos ya vienen suficientemente revisados.

El sistema tiene varias fuentes y cada uno con sus respectivas características de calidad:

SCADA: proporciona las mediciones en tiempo real obtenidas directamente desde los sensores de campo, su calidad no es certificada, por lo que luego de ingresar al sistema pasa por una revisión diaria para validar los datos.

SIMEC: Este sistema utiliza medidores comerciales, los cuales son auditados periódicamente, por lo que el dato suministrado, está debidamente certificado.

UEN Producción de Electricidad: para las plantas pertenecientes al ICE, está normado que el dato oficial es suministrado por esta UEN, diariamente envía los datos del día anterior y al cierre de mes, realizan los ajustes que sean necesarios y certifican los datos en sus sistemas.

UEN Servicio al Cliente: Al cierre de mes, para las plantas que venden energía al ICE, recopilan la información de sus respectivos medidores y la remite al CENCE, para iniciar con el proceso de facturación, estos datos están debidamente certificados.

EOR: Cómo ente regional, toda la información que emite, debe ser validada y certificada por cada país con los controles correspondientes.

Archivos Planos: Para las demás empresas, al cierre de mes, envían la información certificada de generación de energía, para ser incluida en los sistemas del CENCE.

Por lo tanto, un proceso adicional no cabe en esta etapa, pues no se puede aportar más elementos de juicio para una mayor calidad.

Paso 4. Profundizar en las tablas de destino

Una vez que se ha desarrollado las estrategias generales para las tareas comunes de ETL, se debe comenzar a trabajar en las transformaciones detalladas que se necesitan para llenar cada tabla de destino.

Para (Kimball, 2009), una manera realista de este aspecto, sin entrar en un exceso de documentación, es incluir en este detalle al menos lo siguiente:

- Diagrama de flujo general del proceso de carga.
- El modelo de datos de las tablas destino.
- El mapeo de campos fuente-destino.
- Una descripción de lo que se planea realizar.

Los instrumentos resultantes de este proceso se encuentran en el Apéndice 3.

Una vez que se ha creado la especificación de ETL, los pasos siguientes son:

- Paso 5. Rellenar tablas de dimensiones con datos históricos.
- Paso 6. Ejecutar la carga de la tabla de hechos con datos históricos.
- Paso 7. Procesamiento incremental de las tablas de dimensiones.
- Paso 8. Procesamiento incremental de las tablas de hechos.

Es decir, primero desarrollar el proceso ETL de la carga de datos históricos para luego seguir con el proceso de carga incremental.

Como bien lo menciona Kimball, en ocasiones se puede hacer el mismo código ETL para realizar la carga histórica inicial y las cargas incrementales

(Kimball, 2009). Precisamente en este proyecto, se está utilizando esta propuesta, dado que el diseño de los procesos de carga incremental, desarrollado en el paso 4, opera eficazmente para el proceso de carga de datos históricos.

El proceso para la carga de dimensiones y hechos se trabajó básicamente en tres etapas:

Carga de área de trabajo

En esta etapa se realiza la carga de los datos del sistema fuente al área de trabajo y se copian cada una de las tablas involucradas en el proceso a la base de datos de trabajo.

Para filtrar los datos que se procesarán, especialmente para los hechos, en la tabla FECHAS_CARGA ubicada en la base de datos de trabajo, se define el rango de fechas para crear los filtros, y de esta manera procesar sólo el rango de interés o el incremental.

El proceso general se muestra en la Figura 27, donde se aprecia, primeramente que se debe limpiar la base de datos para iniciar un proceso con las tablas limpias y evitar la duplicación de datos. Seguidamente para cada tabla del sistema fuente existe una réplica en el área de trabajo, por lo que la transformación es la más simple que podamos implementar, minimizando la carga de trabajo para el sistema fuente.

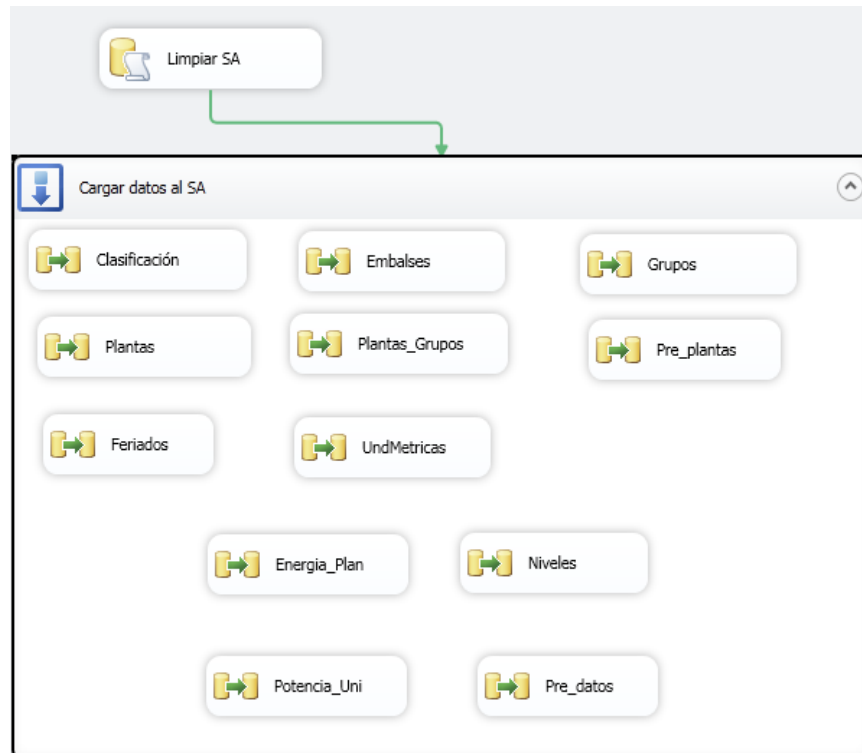


Figura 27. Carga de datos al área de trabajo (elaboración propia).

Carga de dimensiones

Las dimensiones son manejadas como de tipo 1 por sus características, éstas no se actualizan ni se eliminan, es decir, no se aplican las operaciones de UPDATE o DELETE, los cambios solamente se generan por agregación de elementos nuevos, (instrucciones INSERT), de esta manera la carga se simplifica.

El proceso general se muestra en la Figura 28, donde para cada dimensión existe una tarea que toma las tablas fuente de la dimensión, busca los registros inexistentes en la dimensión y los agrega.

Para las dimensiones estáticas, la tarea se resume en la ejecución de un script y verifica si este ha sido ejecutado.

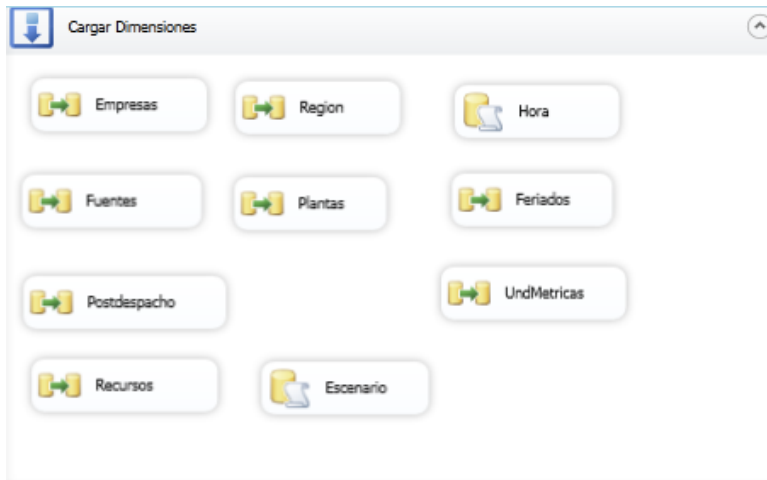


Figura 28. Carga de datos a las dimensiones (elaboración propia).

Carga de hechos

Como se muestra en la Figura 29, para la carga de hechos, el primer paso es eliminar los hechos que existen en las tablas de destinos y se procederá a cargarlos nuevamente, esto con la finalidad de evitar los registros duplicados. Luego, para cada tabla hay procesos específicos que realizarán los cálculos, búsquedas e inserciones necesarias para poblar las tablas de destino.

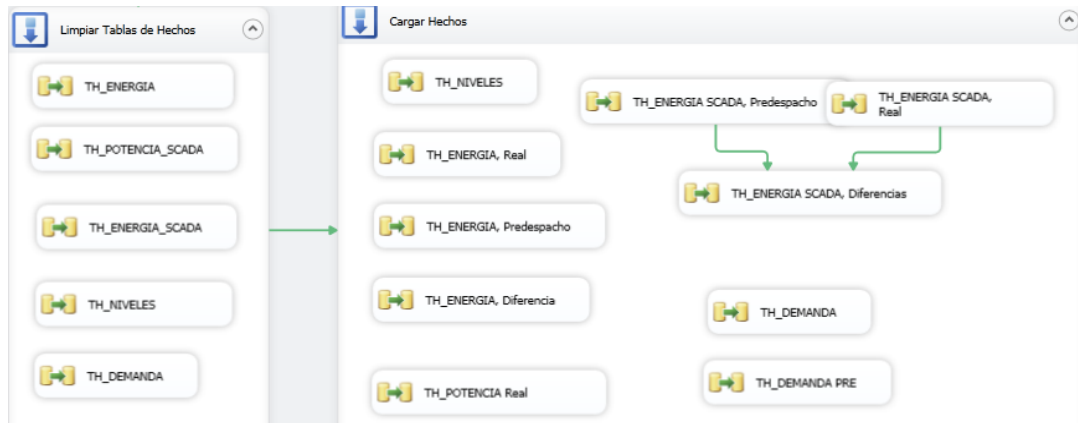


Figura 29. Carga de datos a las tablas de hechos (elaboración propia).

Paso 9. Tabla agregada y cargas OLAP

Gracias a la herramienta que se utiliza, Servicios de Análisis de SQL Server, las tablas agregadas están implementadas en un cubo OLAP, por lo que esta etapa se reduce a la ejecución de una “Tarea de procesamiento de Analysis Services”, para que se procese el cubo implementado.

Paso 10. Operación del sistema ETL y su automatización

Para el presente proyecto, no se considera la automatización de los ETL, sino una operación manual o dirigida por el usuario, esto por las características del sistema fuente.

Grosso modo, en el sistema fuente, siCENCE, se recopila la información de manera automática de las diferentes fuentes, luego, el sistema entra en un proceso de revisión dirigido por un usuario, ahí se realizan una serie de controles cruzados entre los datos a fin de encontrar inconsistencias pues no existe otra forma de tratarlas, más que de forma manual y buscando información adicional que ayude a dilucidar el origen y determinar cuál es el dato correcto. Al termina este proceso de revisión es cuando se puede ejecutar los ETL para pasar los datos al data mart. Este proceso no se puede definir en un tiempo determinado y si tomamos el peor de los datos se corre el riesgo de perder la oportunidad de la información.

De tal forma no se puede programar en el tiempo esta ejecución, la solución final es la actualización del sistema fuente, para que sea desde este que se origine la ejecución de los ETL, lo cual está fuera del alcance del presente proyecto.

4.8. Diseño de aplicaciones BI

Según los requerimientos, para el presente proyecto se utilizarán dos tipos de aplicaciones:

Consultas directas y herramientas de reportes: utilizado para cumplir con las consultas y reportes de los usuarios.

Minería de datos: esta herramienta será utilizada para el proceso de proyección de la demanda.

Por restricciones del negocio no se pueden adquirir nuevas herramientas, pero se tiene la ventaja de que la empresa cuenta con Excel, la cual será utilizada por sus características asociadas a reportes:

- Cálculos básicos con los resultados, como la obtención de subtotales por filas y columnas.
- Girar los resultados, permite fácilmente cambiar las filas por columnas y viceversa, o en términos generales, girar el cubo utilizando las diferentes dimensiones.
- Drill-down: navegar hacia niveles de detalle con los datos.
- Ordenamiento, poder ordenar el reporte utilizando diferentes criterios.
- Graficar los resultados: permite la creación de múltiples tipos de gráficos para ayudar en la visualización de resultados.
- Construir documentos complejos: el paradigma de Excel, utilizando hojas y la incorporación de tablas y gráficos en las hojas permite la construcción de reportes muy complejos.
- Filtrar resultados: Permite el uso de filtros para restringir las consultas e interactuar en el proceso para refinar aún más el filtrado de los datos.

Con respecto a la experiencia del usuario, se tiene las siguientes ventajas:

- Fácil de usar, tradicionalmente las herramientas de Microsoft abogan mucho por la facilidad al usuario final, pero en particular, en este caso, esta característica está potenciada por el hecho de que los usuarios identificados para el sistema, utilizan Excel como herramienta de trabajo, por lo que están muy familiarizados con este.
- Acceso a los metadatos, cuando se utiliza Excel con los cubos de Analysis Services, Excel brinda acceso a los metadatos con los cuales se definió el cubo, así por ejemplo, la dimensión de planta que se ubica en la tabla TD_PLANTA, Excel accede a la dimensión bajo el nombre de Planta.

- Integración con otras aplicaciones: Excel es un estándar en la industria, por lo que su integración con diversas aplicaciones es muy amplia y reconocida, pero en particular es más fuerte con los demás productos de Microsoft.
- Exportar a diferentes tipos de archivos; Excel permite exportar a diferentes tipos de archivos, como por ejemplo, a otras versiones de Excel, texto, DIF (formato de intercambio de datos), CSV, OpenDocument, XML, entre otros.

Identificación de usuarios

Los usuarios identificados inicialmente son:

Personal del “Proceso Planeamiento y Despacho de Energía”: grupo de ingenieros eléctricos con especialidad en Sistemas de potencia, quienes están involucrados en los procesos de predespacho y posdespacho de energía.

Personal del “Área Sistemas de Información”: grupo de informáticos que se encarga de la publicación de los reportes asociados a los procesos de redespacho y posdespacho de energía.

Todo el personal tiene experiencia en el uso de Excel, lo que favorece el uso de herramientas y evita la necesidad de acudir a otras, al menos en el contexto del presente proyecto.

Identificación de reportes candidatos

Con base en los requerimientos del proyecto se procedió a identificar los reportes por implementar, cabe señalar que se identificaron aquellos de prioridad alta, es decir, los que se deben implementar para cumplir con los objetivos, estos se resumen en la Tabla 11.

#	Nombre	Descripción	Categoría
1	Información de la Generación de Energía, Pre vs Post.	Reporte con los datos diarios de generación real y programada. Debe permitir realizar agrupaciones por mes, trimestre y año.	Posdespacho
2	Información de energía Horaria	Reporte con los datos horarios de generación real y/o programada. Debe permitir filtrado para obtener el detalle diario.	Posdespacho
3	Aporte a la Demanda	Reporte para visualizar el aporte de una fuente, empresa, recurso o planta a la demanda de energía total del sistema	Posdespacho
4	Información de potencia	Reporte para visualizar la curva de potencia del sistema, agrupado ya sea por posdespacho, fuente, empresa o recurso.	Posdespacho
5	Información de Niveles de Embalses	Reporte con los niveles de los embalses. Y su relación con la generación de energía de sus plantas asociadas.	Posdespacho
6	Información de la Generación de Energía.	Resumen de la generación real de energía por período, día, mes, trimestre o año.	Posdespacho

Tabla 11. Reporte candidatos (elaboración propia).

Bosquejo de la aplicación

El bosquejo de la aplicación, en este caso, de las hojas Excel se muestra en la Figura 30:

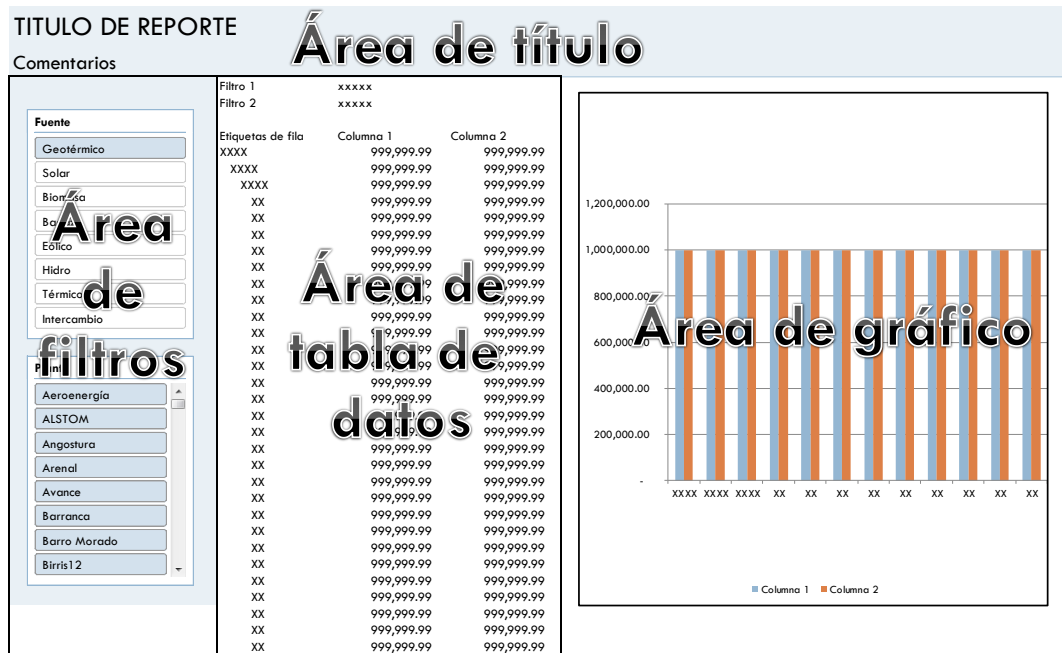


Figura 30. Bosquejo de la aplicación BI (elaboración propia).

Se definen cuatro áreas bien identificadas:

Área del título: en esta área se incluye el título del reporte, así como comentarios o instrucciones relacionadas con el reporte.

Área de filtros: en esta área se colocan las dimensiones principales, relacionadas al reporte, para facilitar el filtrado de la información.

Área de tabla de datos: en esta área se coloca la tabla dinámica con los datos del reporte.

Área de gráfico: se ubica un gráfico relacionado con la tabla de datos, para ayudar en la comprensión de los mismos. Según el tamaño de la tabla, este gráfico se puede omitir o mover a una hoja diferente.

Interacción del usuario

Con este bosquejo, la interacción con el usuario es bastante simple, pero a la vez muy flexible, pues al colocar los filtros principales a la izquierda, le permite al usuario realizar el filtrado de una manera muy intuitiva y utilizando los elementos estándar que tiene Excel.

Siguiendo hacia la derecha, se ubica la tabla dinámica con los datos. Esta tabla por defecto oculta los campos para edición, y muestra una interface más

limpia. Pero igualmente, si el usuario lo desea, puede mostrar estos campos y editar el reporte a su conveniencia.

Esta tabla también puede tener filtros adicionales, que permitan realizar filtros en busca de mayor detalle.

Cuando el usuario interactúa con los filtros, automáticamente se van actualizando los datos de la tabla y del gráfico, así la interacción es fluida e inmediata para el usuario.

Para algunas aplicaciones, se pueden tener varias pestañas que pueden contener hojas con el mismo bosquejo o un gráfico completo, por lo que se pueden construir reportes más complejos.

Conjunto de datos

Los datos se obtendrán del cubo definido en Analysis Services, que para todos los casos es el data mart definido para el CENCE, denominado CENCE DM.

4.9. Desarrollo de aplicaciones BI

Definido el conjunto de reportes y el bosquejo utilizado, se procedió al desarrollo de cada uno de los reportes.

Se crearon seis archivos Excel, para cumplir con los reportes solicitados y los requerimientos asociados, el resumen de estos se muestra en la Tabla 12:

#	Nombre	Archivo	Requerimiento asociado
1	Información de la generación de energía, Pre vs Post.	R1-R2, Pre y Post.xlsx	R1, R2
2	Información de energía horaria	R1-R2 Pre y Post Horario.xlsx	R1, R2
3	Aporte a la demanda	R3-R4, Aporte a la demanda.xlsx	R3-R4
4	Información de potencia	R5, Potencia.xlsx	R5
5	Información de niveles de embalses	R6-R7, Niveles.xlsx	R6-R7
6	Información de la generación de energía.	R8, Generación por período.xlsx	R8

Tabla 12. Lista de archivos Excel implementados (elaboración propia).

A modo de ejemplo, en la Figura 31 se muestra el archivo R6-R7, Niveles.xlsx, donde se visualiza la relación del nivel del embalse con respecto a la generación real para la planta Cachi el 1 de mayo de 2014.

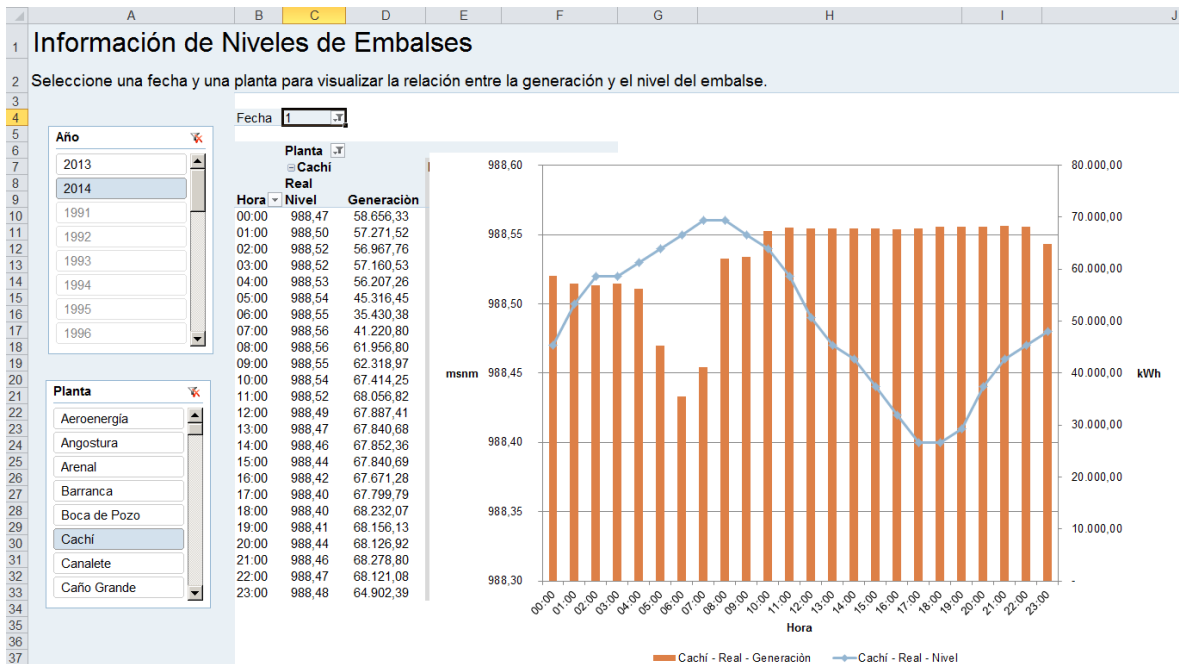


Figura 31. Ejemplo de aplicación BI (elaboración propia).

4.10. Implementación

Las vías de tecnología (arquitectura técnica), de datos (modelado dimensional) y de aplicación convergen en la implementación de la solución. Kimball hace la analogía con la actividad de “servir la cena” que ha de estar, como es de esperar con un grado adecuado de cocción (Kimball & Ross, 2013).

La implementación de la solución incluye tareas como:

El aseguramiento de la calidad de los datos. Para asegurar que el data mart y los cubos de datos arrojen resultados correctos y precisos se realizaron constantes comparaciones con los reportes existentes en la empresa. No sólo durante el proceso de construcción de la herramienta sino durante esta etapa de implementación, con apoyo y supervisión de expertos del negocio. Los cálculos fueron validados utilizando como recurso la comparación, realizando las operaciones en Excel con chequeos puntuales y continuos.

Pruebas de procesos de operación. Se verifica que los trabajos programados de ETL se ejecutan cuando deben y lo hacen correctamente. Se confirma que los cubos se procesan tal como se espera. Se monitorea el estrés del sistema fuente siCENCE, tanto durante la carga inicial de datos como en cargas incrementales y se califica de adecuada. No genera problemas de rendimiento a los servidores involucrados.

Pruebas de usabilidad. Durante la construcción se han realizado ya múltiples pruebas orientadas a la usabilidad. De la misma forma, el acompañamiento cercano de usuarios expertos que a la vez son usuarios meta permite que las pruebas de usabilidad en esta etapa no generen mayores problemas. Las pruebas con usuarios que no han estado involucrados en el desarrollo permiten validar el grado aceptable de usabilidad del producto. Claro está, dado que el front-end es Excel, desde el punto de vista de un usuario final hay bastante transparencia y familiaridad.

Empiezan a aparecer sugerencias e ideas por parte de los usuarios. Esto confirma el interés generado por la herramienta y la evolución normal de una solución de inteligencia de negocios.

Por otra parte se comprueba que los niveles de seguridad son los adecuados.

Disposición de las máquinas de escritorio y configuración. Dado que se trata de un ambiente bastante homogéneo, la implementación de los ambientes de escritorio es tarea sencilla. Las computadoras en las cuales se ejecutará la herramienta, es sabido (y comprobado) que cuentan con Microsoft Excel 2010 y con conectividad al servidor. Todos los clientes están dentro de la red institucional que es Microsoft, regidos por un Active Directory. El ambiente es controlado y las pruebas de conectividad y de ejecución son exitosas.

5. Proyección de la demanda eléctrica utilizando metodología CRISP-DM

CRISP-DM, que son las siglas de Cross-Industry Standard Process for Data Mining, es un método probado para orientar trabajos de minería de datos.

Como metodología, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas.

Organiza el desarrollo de un proyecto de Data Mining, en una serie de seis fases, como se muestra en Figura 32.

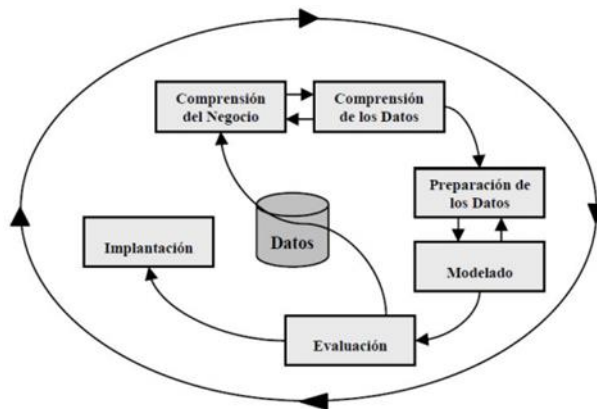


Figura 32. Fases de la metodología CRISP-DM (IBM Corporation, 2012)

La sucesión de fases no es necesariamente rígida. Cada fase es estructurada en varias tareas generales de segundo nivel. Las tareas generales se proyectan a tareas específicas, donde finalmente se describen las acciones que deben ser desarrolladas para situaciones específicas, pero en ningún momento se propone cómo realizarlas.

5.1. Fase de comprensión del negocio o problema

Esta fase inicial se centra en lograr una comprensión adecuada de los objetivos del proyecto desde una perspectiva del negocio y utilizar esa comprensión para definir el problema de minería de datos y lograr los objetivos.

Determinación de los objetivos del negocio

Descripción general

La demanda de energía eléctrica en el país, a cada instante, durante las veinticuatro horas del día, todos los días del año, tiene que ser atendida de la forma más optimizada posible. Su adecuada satisfacción significa que la estabilidad del sistema y la calidad del servicio a nivel nacional, no se vean afectados en ningún momento, ni siquiera cuando casi simultáneamente se enciendan cientos de miles de cocinas para preparar la cena o se enciendan miles de máquinas industriales para continuar la producción. Se trata también de no desperdiciar recursos, térmicos, hídricos o meramente económicos.

El dato correspondiente a la demanda constituye un insumo para el proceso de cálculo del predespacho nacional. La elaboración del llamado Predespacho Nacional es una tarea compleja, de suma importancia para el quehacer del CENCE y ulteriormente para el país en general. Tiene que considerar además la capacidad y estado de las plantas de generación, las capacidades de las líneas de transmisión, medidas de seguridad operativa del sistema y muchas otras variables. Al final se obtiene un tipo de plan de acción para que el país pueda tener la energía eléctrica en su medida adecuada para cada hora y cada día.

El cálculo de la demanda para cada hora del día tampoco carece de complejidad en sí mismo. Por lo que una herramienta que facilite esa tarea y permita lograr una mejor aproximación a la realidad se convierte en una excelente oportunidad de mejora.

Objetivo

Proveer una herramienta que ayude al usuario experto en la proyección de la demanda de energía. Con la aplicación de la metodología CRISP-DM se pretende mejorar el pronóstico de la demanda eléctrica nacional a corto plazo en Costa Rica. Entendiendo esta mejora en términos de precisión y en tiempo de realización de la tarea, con respecto a cómo es realizada actualmente por el CENCE.

Criterios de éxito

CRE – 01 Reducción de margen de error. Se considerará exitoso el proyecto de minería de datos si se logra reducir el margen de error en el pronóstico de la demanda a corto plazo para el SEN con respecto al despacho nacional, en comparación con la manera en que es realizado actualmente por el departamento de Planeamiento y Despacho de Energía del CENCE.

CRE – 02 Reducción del esfuerzo invertido en el proceso de proyección. También será considerado exitoso si, manteniendo el mismo margen de error, se reduce el tiempo necesario para calcular la demanda horaria a corto plazo del SEN. Es decir, si el número de horas hombre por semana invertidas por los profesionales encargados se reduce.

Evaluación de la situación

Uno de los objetivos generales del proyecto consiste en describir en qué consiste y cómo es elaborada actualmente la proyección de la demanda de energía. Esta descripción del proceso, constituye uno de las tareas fundamentales de cara a la aplicación de minería de datos porque implica la conceptualización y la comprensión del problema.

Proyección de la demanda de energía

Los cálculos para pronosticar la demanda son realizados por el grupo de ingenieros del Proceso de Planeamiento y Despacho de Energía del CENCE. Los expertos que conforman el grupo se turnan para ejecutar esta tarea. Se realiza dos veces por semana, el día martes se hacen los cálculos para proyectar los días siguientes, de jueves a lunes; y el día jueves se proyecta la demanda para los días martes y miércoles de la semana siguiente.

El pronóstico de la demanda se hace entonces para cada día, dividiéndolo en bloques de una hora, (veinticuatro por día). El resultado es una tabla de datos, en la que, para el o los días que calculan y para cada hora del día, se tiene una cantidad de energía que corresponde a la demanda promedio (máxima en el caso particular de las 6 p.m.) esperada durante esa hora, para la cual el CENCE debe prepararse para responder, satisfacer la demanda y mantener la estabilidad del SEN.

Someramente, el pronóstico se hace tomando la demanda del día correspondiente de la semana anterior (7 días antes) y aplicándole a este dato un factor de crecimiento que se obtiene del promedio de las demandas registradas para ese mismo día de la semana y mismo número de semana en el año, considerando los cinco años anteriores.

Descripción detallada del procedimiento

El pronóstico tiene una base diaria, es decir se realiza por bloques de 24 datos correspondiente a cada hora del día. Entonces, sea $D(s_{dh})$, la demanda máxima para el día d y la hora h , lo cual cumple con el objetivo del proceso; donde d corresponde a uno de los 7 días de la semana (L,K,M,J,V,S,D), h corresponde a una de las 24 horas del día (0..23) y s corresponde al número de semana en el año por pronosticar (se consideran 52 semanas en el año).

En general, se considera la semana actual s como el periodo objetivo y la semana anterior $s-1$ como la base para el pronóstico.

Entonces se define $D(s_{dh})$ como

$$D(s_{dh}) = \begin{cases} D((s-1)_{dh}) \cdot \frac{\overline{Pmax(s)}}{\overline{Pmax(s-1)}}, & \text{para } h = 18 \\ D((s-1)_{dh}) \cdot C_d, & \text{para } h \neq 18 \end{cases}$$

La hora de demanda máxima ($h = 18$), es la más importante en términos de criticidad y lo primero que se hace es asegurar su satisfacción. Por tanto, se calcula de manera diferente a los otros bloques horarios.

La fórmula anterior, $D((s-1)_{dh})$ corresponde a la demanda para un día y hora específicos de la semana anterior. Cabe anotar que tal dato proviene de los datos reales del sistema historiador del SCADA. Es un promedio de registros tomados cada 4 segundos.

$\overline{Pmax(s)}$ es el promedio de las demandas máximas de potencia (en MW) para la semana s en los cinco años anteriores, normalizadas (utilizando la demanda máxima de cada año como base). Es decir,

$$\overline{Pmax(s)} = \frac{1}{n} \sum_{i=1}^n \overline{Pmax(s)_i}$$

Donde n es el número de años considerados y $\overline{Pmax(s)}_i$ la potencia máxima normalizada registrada para la semana i en cada uno de los n años utilizados para el análisis.

De esta manera, el estimado de demanda a la hora pico se calcula utilizando el correspondiente en la semana $s-1$ multiplicado por la razón promedio en potencia máxima, normalizada, que se da entre las semanas $s-1$ y s para los cinco años anteriores.

Cuando h es distinto de 18, simplemente se le aplica un coeficiente C_d . Éste determina cuánto creció la demanda para el día d de la semana s , con respecto a la semana $s-1$. Entonces:

$$C_d = \frac{ETE(s_d) - D(s_{d18})}{\sum_{h=0}^{23} D((s-1)_{dh})}$$

con $h \neq 18$

C_d es igual a la energía total estimada para ese día de la semana s , o $ETE(s_d)$, dividido entre el total de energía para ese día d de la semana $s-1$ (exceptuando de nuevo la hora pico). El coeficiente C_d se convierte en el factor de cambio de energía para ese día de la semana.

La energía total estimada para el día de la semana actual s , o $ETE(s_d)$, es una proyección de la semana anterior, para el día d , considerando el crecimiento promedio presentado entre la semana $s-1$ y s , en los cinco años anteriores:

$$ETE(s_d) = \sum_{h=0}^{23} D(s_{dh}) \frac{\overline{Emax(s)}}{\overline{Emax(s-1)}}$$

Siendo $\overline{Emax(s)}$ el promedio de las demandas máximas de energía (en MWh) para la semana s en los cinco años anteriores, normalizadas (utilizando la demanda máxima de cada año como base). Calculado de forma análoga a $\overline{Pmax(s)}$.

De esta manera, para cada hora, la energía pronosticada será igual al correspondiente dato de ese día en la semana anterior ($s-1$), y modificado por el coeficiente C_d .

Nótese que el coeficiente C_d es calculado y aplicado en una base diaria, es decir, es el mismo para todas las horas del día excepto a las 6 p.m., $h=18$.

De esta forma, por ejemplo, la demanda máxima para el lunes a las 08 horas, de la semana 40 del año en curso es igual a la demanda máxima del lunes de la semana 39 a las 08 horas, multiplicado por un coeficiente C_d , que se obtiene considerando el crecimiento promedio de la semana 39 a la 40 de los cinco años anteriores.

Para los días feriados, se trata igual que el domingo anterior inmediato.

Roles

En el presente proyecto de minería de datos se identifican los siguientes roles:

R01 - Líder Gerencial (LG): patrocinador del proyecto desde el punto de vista de la organización, es el Centro Nacional de Control de Energía.

R02 - Líder del negocio (LN): proporciona la guía de negocios y los requerimientos para el equipo del proyecto; garantiza que la información capturada es adecuada y responde a las necesidades de la organización.

R03 - Líder de proyecto (LP): desde el punto de vista organizacional es la persona responsable de la coordinación del proyecto ante la gerencia. Esta persona se desempeña como administrador del área de sistemas de información del CENCE.

R04 - Experto del negocio (EN): persona conocedora del negocio y experta en los procesos de cálculo de la demanda nacional y procesos de predespacho de energía.

R05 - Desarrollador de minería de datos (MD): encargado de llevar a cabo los procesos de preparación de datos y aplicación de técnicas de minería de datos.

R06 - Administrador del proyecto (AP): responsable de la gestión de tareas y actividades del proyecto, incluyendo la coordinación de recursos, el seguimiento de estado y la comunicación de los avances y problemas del proyecto, tiene estrecha colaboración con los líderes del proyecto y negocio.

El proyecto, en su núcleo de trabajo, cuenta con dos desarrolladores de minería de datos y dos expertos del negocio. Tanto unos como otros, pueden ser considerados expertos en los datos.

Recurso humano

Los recursos humanos identificados para el proyecto de minería, se resumen en la Tabla 13.

Puesto en la organización	Roles					
	R01 LG	R02 LN	R03 LP	R04 EN	R05 MD	R06 AP
Director del CENCE	✓	✓		✓		
Coordinador de Sistemas de Información		✓	✓			
Director del Área de PDE				✓		
Profesional en sistemas de potencia		✓		✓		
Profesional en informática				✓	✓	✓
Profesional en informática					✓	

Tabla 13. Recursos identificados para el proceso de minería (elaboración propia).

Recursos de hardware y software

Para este proyecto de minería se cuenta con los mismos recursos en equipo computacional y en software que para la construcción del data mart del proyecto SAHEP. Por esto se utilizará un servidor Dell PowerEdge R2900 con Windows Server 2008 R2 como sistema operativo y SQL Server 2012 como servidor dedicado a la inteligencia de negocios.

Como herramienta de trabajo para la minería se utilizará Microsoft Excel con los complementos correspondientes para utilizar las funcionalidades de minería de datos de SQL Server, según los mismo criterios establecidos para el desarrollo del data mart.

Fuentes de datos

Se utilizará exclusivamente el data mart construido para el proyecto SAHEP, albergado en el servidor de inteligencia de negocios del CENCE.

Requerimientos

REQ-01 - Recurso Humano: se requiere la disponibilidad, en diferente medida, del recurso humano participante en el proyecto. Particularmente se hace necesaria la participación activa de los expertos de negocio y desde luego, los mineros de datos.

REQ-02 - Hardware: se requiere la disponibilidad y el acceso al servidor de inteligencia de negocios de la organización. Es preciso que el equipo provea las capacidades de procesamiento requeridas por las tareas de minería.

REQ-03 - Software: se requiere contar con las licencias respectivas de las herramientas por utilizar, tanto en la plataforma del servidor como de Microsoft Office 2010.

Restricciones

RST – 01 Variables meteorológicas: el proceso de minería ligado a pronosticar la demanda a corto plazo, no considerará factores meteorológicos dado que no se cuenta con información histórica sobre este particular.

RST – 02 Alcance temporal: La proyección será realizada en el corto plazo, es decir, un horizonte máximo de una semana.

RST – 03 Predespacho: La solución no incluye el predespacho de energía, se limita a dar la proyección de la demanda que es un insumo a este predespacho.

Riesgos

En la Tabla 14 se describe una serie de riesgos identificados que podrían afectar la consecución de los objetivos planteados.

Riesgo	Descripción	Probabilidad	Impacto	Acción de mitigación
RSG-01	Los datos disponibles no cubren todos los factores y variables relevantes.	baja	alto	Enriquecer el data mart (fuente de datos) con la información faltante requerida.
RSG-02	La granularidad de los datos impide la obtención de resultados adecuados.	baja	alto	Recargar el data mart con datos a mayor granularidad, con todas sus implicaciones
RSG-03	Falta de disponibilidad de los expertos del negocio.	media	medio	Establecer un cronograma de actividades y coordinar reuniones con antelación.
RSG-04	Plataforma tecnológica insuficiente.	baja	alto	Revisión previa de las capacidades de procesamiento y almacenamiento disponibles.
RSG-05	Pérdida de datos de código fuente.	baja	alto	La solución en desarrollo contará con respaldos diarios en la nube y en al menos dos medios físicos. La solución implantada estará respaldada por un servidor especializado para tal fin.
RSG-06	Resistencia organizacional.	baja	medio	Educación acerca de la herramienta. Trabajo cercano e involucramiento temprano de personal clave.
RSG-07	Conocimiento insuficiente de los desarrolladores sobre los datos	baja	medio	Coordinación previa para asegurar el apoyo de los expertos en la construcción de la solución y la exploración de datos.

Tabla 14. Riesgos identificados para el proceso de minería (elaboración propia).

Costos

Dado que ya se cuenta con la infraestructura tecnológica, el costo del proyecto contempla el tiempo invertido por diversos profesionales, tal como se muestra en la Tabla 15.

Descripción	Unidad	Costo Unit.	Cant	Total \$	Total ¢
Director del CENCE	Hora	80	1	80	¢44,000
Coordinador de Sistemas de Información	Hora	70	1	70	¢38,500
Director del Área de PDE	Hora	70	2	140	¢77,000
Profesional en sistemas de potencia	Hora	60	8	480	¢264,000
Profesional en informática (Minería de datos)	Hora	60	60	3600	¢1,980,000
Costo total del proyecto				0	¢2,403,500

* Precio Dólar (¢) 550

Tabla 15. Costos estimados de minería de datos para pronóstico de demanda (elaboración propia).

Beneficios

BNF – 01 Mejora en la exactitud el pronóstico de la demanda. Un pronóstico más preciso permite una mejor programación de las unidades generadoras, esto ayuda, entre otras cosas, a reducir el riesgo de redespachos de energía, donde se deban utilizar unidades más costosas, o por el contrario, que se deba prescindir de generaciones programadas. Todo ello implica mayores costos de generación. También se reduce el riesgo de sobrecargas y apagones en el SEN.

BNF – 02 Mejora los tiempos de elaboración del predespacho nacional. La proyección de demanda es uno de los insumos para la elaboración del predespacho nacional. El pronóstico de la demanda utilizando métodos de minería de datos agiliza los procesos organizacionales asociados y libera a los especialistas de esas tareas, permitiéndoles dedicar esfuerzos a otras tareas analíticas.

Determinación de los objetivos de minería de datos

Objetivos específicos de minería de datos

- Definir el método que se utilizará para la proyección de la demanda mediante el uso de redes neuronales, promedios móviles o ARIMA, de acuerdo con los requerimientos identificados.
- Implementar el algoritmo descrito para apoyar el proceso de proyección de la demanda.
- Analizar los resultados obtenidos mediante un proceso de comparación de acuerdo con un plan de pruebas.

Criterios de éxito de la minería de datos

CEM –01 Proyección de la demanda más precisa. El proceso de minería debe lograr que la proyección de la demanda sea más cercana, en la mayoría de los casos, a la demanda real, medida en el posdespacho con respecto a lo proyectado con el método actual.

CEM – 02 Facilidad de uso e integración al proceso de elaboración del predespacho. Se debe lograr que los expertos del negocio cuenten con una herramienta que resulte fácil de utilizar y de integrar al proceso del predespacho, del cual la proyección de la demanda es un insumo clave.

CEM – 03 Rendimiento. Se debe lograr que, el sistema encuentre una solución aceptable, es decir, igual o más precisa que la obtenida con el método actual, dentro de una ventana de tiempo que no supere las dos o tres horas. Se considerará excelente si encuentra la solución en una hora o menos y no será aceptable si tarda más de la duración de una noche (unas diez horas).

Realización de un plan de proyecto

Plan de proyecto

El plan de proyecto se resume en la Tabla 16

Nombre de tarea	Duración	Comienzo	Fin
Proyección de la demanda	90 días	vie 08/08/14	mié 05/11/14
Comprensión del problema	5 días	vie 08/08/14	mar 12/08/14
Comprensión de los datos	10 días	mié 13/08/14	vie 22/08/14
Preparación de los datos	5 días	mié 17/09/14	dom 21/09/14
Modelado	30 días	lun 22/09/14	mar 21/10/14
Evaluación	15 días	mar 07/10/14	mar 21/10/14
Informe métodos propuestos	0 días	mar 21/10/14	mar 21/10/14
Implementación	15 días	mié 22/10/14	mié 05/11/14

Tabla 16. Plan de proyecto (elaboración propia).

Evaluación inicial de herramientas y técnicas

Se ha investigado que las herramientas disponibles en la organización proveen soporte para las técnicas y características requeridas para el análisis de datos. SQL Server 2012, en conjunto con Excel, disponen de los instrumentos para trabajar con redes neuronales, promedios móviles y ARIMA, que son las técnicas para aplicar y comparar.

El problema por resolver es clasificado como de predicción y dado que se refiere a datos de series de tiempo, también es llamado de “previsión” o pronóstico (en inglés *forecasting*).

Parte de las actividades planteadas en el proyecto es la comparación de diversos métodos para realizar el pronóstico de la demanda. Se comparan tres métodos, en términos de exactitud, en primera instancia y en segundo plano, de facilidad de uso y versatilidad. Tales métodos serán aplicados a una muestra de los datos y se compararán los resultados obtenidos.

5.2 Comprensión de los datos

En la fase de comprensión de datos de CRISP-DM, se desarrolla el estudio de los datos disponibles para realizar el proceso de minería. Este paso es esencial para evitar problemas inesperados durante las siguientes fases. (IBM Corporation, 2012)

La comprensión de datos implica acceder a estos y explorarlos con la ayuda de herramientas, como bien podría ser tablas y gráficos, que permitan entender más en detalle los datos y su calidad.

Recopilación de los datos iniciales

El proceso de minería tiene como origen de datos el data mart desarrollado en el capítulo 4, por lo que, los datos ya han sido recopilados previamente.

Para el proceso de minería se utilizarán básicamente tres tablas, a saber:

- TD_Fecha: Dimensión de fecha, que contiene los atributos que caracterizan a las fechas en particular, por ejemplo, qué día de la semana es, si se trata de un día feriado, otro.
- TD_Hora: Dimensión de hora.
- TH_Demanda: Contiene la demanda de energía horaria para todo el sistema eléctrico.

El proceso actual de la proyección de la demanda utiliza datos de cinco años. Por otro lado, en el data mart se disponen de datos a partir del año 1996, por lo que se consideran datos suficientes para obtener conclusiones generales.

Descripción de los datos

La primera tabla es la que contiene la dimensión de fecha, a saber, TD_Fecha. La cantidad de registros es de 10193, pero de estos no todos tienen registros asociados a la demanda, pues el registro de la potencia se inició a partir de 1996 y en la base de datos de estudio finaliza en junio de 2004, por lo que en total hay 6756 registros utilizables. La tabla cuenta con veintiún atributos, pero solamente se consideran diez importantes para el proceso de minería. La descripción de esta tabla se resume en la Tabla 17.

Columna	Tipo	Descripción
IDFECHA	entero	identificador de la fecha
FECHA	fecha	fecha
ANIO	entero	año
TRIMESTRE	entero	trimestre
MES	entero	mes
SEMANA	entero	número de la semana
DIA	entero	día
DIASEMANA	entero	día de la semana
FERIADO	entero	valor para identificar si se trata de un día feriado, los posibles valores son: 0 No 1 Si
VACACIONES	entero	valor para identificar si corresponde a un día de vacaciones escolares, los posibles valores son: 0 Curso lectivo 1 Vacaciones escolares 2 Vacaciones de medio período

Tabla 17. Descripción de los campos a utilizar de TD_Fecha (elaboración propia).

Para la tabla de dimensión TD_Hora, la descripción de los campos por utilizar se resumen en la Tabla 18. En esta tabla hay noventa y seis registros y nueve atributos, pero de estos solamente se utilizarán cinco, que se consideran los relevantes para el proceso.

Columna	Tipo	Descripción
IDHORA	entero	Identificador de la hora
HORA	entero	Hora del día
MINUTO	entero	Minutos
TIPO	entero	Tipo de hora, los posibles tipos son: 0 Valle nocturno 1 Pico del Desayuno 2 Rampa de la Mañana 3 Pico del mediodía 4 Valle de la tarde 5 Pico de la tarde
PUNTA	entero	Valor para hora punta, los posibles valores son: 0 Fuera de punta 1 Punta

Tabla 18. Descripción de los campos a utilizar de TD_Hora (elaboración propia).

Finalmente, la tabla que contiene los datos para el proceso de minería es la tabla de hechos TH_Demanda, contiene más de trescientos mil registros utilizables. La descripción de esta tabla se resume en la Tabla 19.

Columna	Tipo	Descripción
IDFECHA	entero	Identificador de la fecha
ID_HORA	entero	Identificador de la hora
IDESCENARIO	entero	Identificador del escenario, los posibles valores son: 1 Real 2 Programado
IDUNIDAD	entero	Identificador de la unidad, para el estudio se considera realizarlo en MWh por lo que el valor a utilizar para filtrar este atributo será 5.
ENERGIA	numérico	Demanda de energía horaria.

Tabla 19. Descripción de los campos a utilizar de TH_Demanda (elaboración propia).

Exploración de datos

Cuando se realiza la exploración preliminar de los datos, como se observa en la Figura 33, la demanda muestra un comportamiento en ciclos, así cada día se presenta dos picos altos, uno al medio día y otro a las 18 horas. Luego, la demanda del domingo y sábado son inferiores al resto de los días, de esta revisión los atributos obvios de estudio son el día y la hora. Estos atributos constituyen el punto de partida para el proceso de minería.

La demanda constituye una serie de tiempo, con características estacionales, por lo que es un problema apropiado para ser tratado con los algoritmos propuestos.

proceso de minería, la preparación de los datos ya ha sido realizada en su mayoría.

Así, esta etapa se reduce a realizar los filtros sobre los datos según el modelo que se vaya utilizando y los datos por revisar. Por ejemplo, los datos reales para la proyección sólo están disponibles a partir de junio de 2012.

Finalmente, para cada método se realizarán los filtros o consultas sobre el data mart que mejor se ajuste al método dado.

Manejo de errores

Al realizarse el estudio de los errores encontrados con respecto a los datos reales y a los proyectos, se encuentra que estos tienen una distribución normal, como se muestra en la Figura 34.

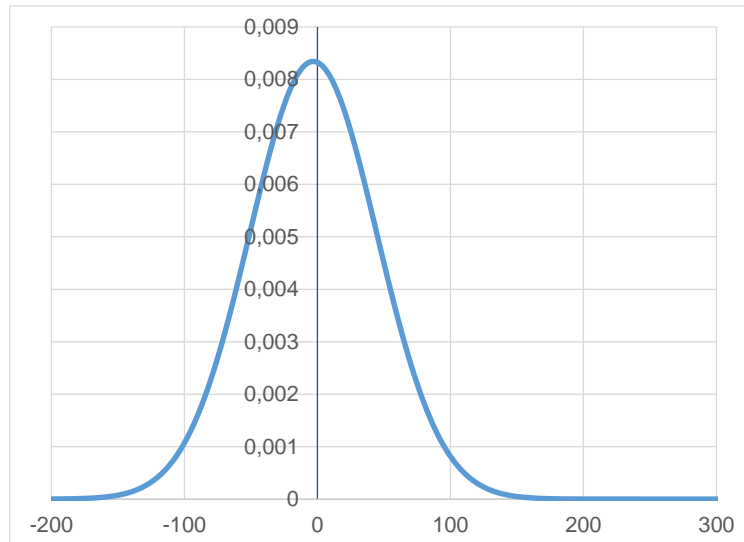


Figura 34. Distribución normal de los errores para la proyección de la demanda (elaboración propia).

Los indicadores estadísticos más utilizados para evaluar el funcionamiento de un modelo de predicción son el error absoluto medio EAM y la raíz del error cuadrático medio RECM (Fernández Jiménez, 2007).

Debido a que los errores del modelo son propensos a tener una distribución normal en lugar de una distribución uniforme, el RECM es una métrica más adecuada que el EAM para este tipo de datos. (Chai & Draxler, 2014).

De tal modo, el cálculo de los errores se realizará con el RECM, el cual se define como

$$RECM = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_r - x_p)^2}$$

Donde:

- x_r es el valor real
- x_p es la predicción del valor
- n es la cantidad de valores

Éste será el método que se utilizará para realizar las comparaciones entre el modelo actual y los métodos propuestos.

5.4. Modelado y evaluación

Promedios móviles

El primer método que se evaluará es el de promedios móviles.

Este diseño tiene como objetivo disminuir las desviaciones horarias en el pronóstico mediante un modelo sencillo y fácil de aplicar. Su implementación conlleva las siguientes tareas:

- a. Clasificar los días según el día de la semana en lunes, martes, miércoles, jueves, viernes, sábado y domingo.
- b. Se filtran los días feriados para considerar sólo los días normales.
- c. Clasificar los registros por cada una de las 24 horas del día.
- d. Uso de promedios móviles de diferentes órdenes para cada serie horaria.

Después de la aplicación de las clasificaciones descritas anteriormente se obtuvieron entre cincuenta y tres y cincuenta y ocho series de demanda horaria para cada tipo de día. Para definir el orden de los promedios o número de observaciones anteriores que mejor modelaran la demanda, se diseñó una plantilla con Excel, que calculó el error cuadrático medio (ECM) de energía para cada una de las 24 horas del día utilizando promedios móviles de los órdenes 1 al 20. A partir de esto se obtuvo el error por utilizar (RECM) de los pronósticos con respecto al valor real para el año 2013 y finalmente se seleccionó el orden que menor RECM mostró.

En la Tabla 20, se muestran los órdenes seleccionados para realizar los promedios, así para cada hora y cada día de la semana, se debe utilizar el promedio con el orden establecido en la tabla, por ejemplo, para obtener la proyección de la demanda para el día domingo a las 10 am, se deben promediar los seis domingos anteriores.

	Lunes	Martes	Miércoles	Jueves	Viernes	Sábado	Domingo
00:00	4	2	8	6	3	11	4
01:00	4	4	8	6	3	13	4
02:00	8	4	7	6	2	13	4
03:00	7	4	7	6	2	13	4
04:00	7	4	6	6	2	13	4
05:00	2	2	2	3	2	13	4
06:00	2	2	2	2	2	13	4
07:00	2	2	20	3	2	13	4
08:00	10	2	13	6	2	15	4
09:00	9	6	7	6	2	15	4
10:00	9	6	12	6	2	15	6
11:00	10	6	7	4	4	15	5
12:00	10	7	6	4	4	4	4
13:00	10	11	5	4	4	4	4
14:00	10	3	3	4	4	4	4
15:00	6	3	3	4	4	4	4
16:00	4	2	13	3	4	7	4
17:00	2	2	2	2	3	5	3
18:00	4	2	6	2	2	4	3
19:00	4	2	4	2	2	4	4
20:00	6	2	2	4	4	4	4
21:00	6	3	4	4	4	4	4
22:00	6	3	3	4	6	4	4
23:00	4	3	4	3	2	4	4

Tabla 20. Ordenes de los promedios móviles para las series de demanda horaria (elaboración propia).

Seguidamente se procedió a realizar una comparación entre los errores obtenidos por el método actualmente utilizado y el error obtenido con promedios móviles.

Los resultados de esta comparación, se observan en la Tabla 21, donde se aprecia el mejor resultado por día y el mejor resultado general; en la Figura 35 se

observan gráficamente los errores para cada día. Para los días miércoles y jueves el método actual es mejor, pero la pérdida de precisión es inferior al 5%; en tanto para los restantes días, se presentan mejoras importantes, todas superiores al 7%, 4 días con mejoras superiores al 10% y la mejora más significativa es para los días domingos con un 35% con respecto al método actualmente utilizado.

Día	PDE	Promedios Móviles	Mejora
Lunes	48,99	45,45	7,23%
Martes	41,13	35,96	12,58%
Miércoles	39,69	40,97	-3,23%
Jueves	36,86	38,41	-4,21%
Viernes	41,02	36,14	11,90%
Sábado	41,47	37,16	10,41%
Domingo	46,17	29,79	35,48%
General	42,36	37,96	10,39%

Tabla 21. Comparación con promedios móviles (elaboración propia).

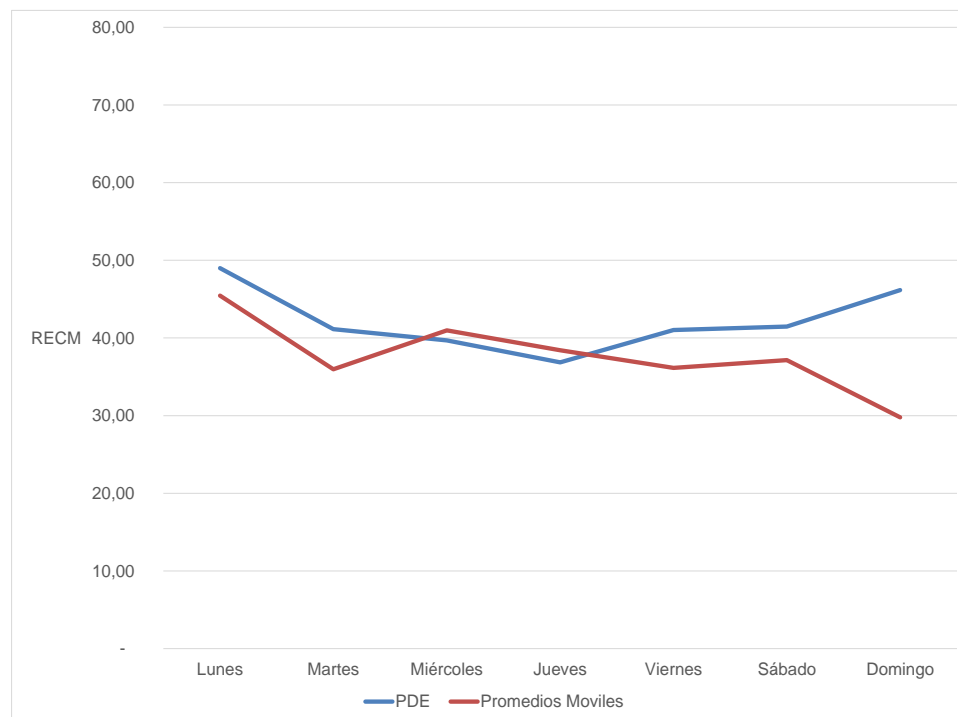


Figura 35. Comparación del error PDE vs Promedios móviles (elaboración propia).

Se debe hacer la salvedad, de que en esta evaluación, se están omitiendo los días feriados, que el modelo propone tratarlo como si fueran domingos, lo cual es recíproco con el método actualmente utilizado.

En término general, el método de promedios móviles tiene una mejora de un 10.39% por lo que su implementación ayuda a mejorar el proceso de proyección de la demanda.

Series de tiempo de Microsoft

En este modelo de minería de datos, Microsoft implementa el algoritmo ARIMA, combinado con el ARTXP.

Para modelar el problema se utilizó el Microsoft Visual Studio, se creó un proyecto de minería y el modelo seleccionado es precisamente el Microsoft Times Series.

Una vez definido el modelo se inicia una etapa de ajustes del mismo, se ajustan parámetros y se agregan o eliminan columnas de entrada. Para cada corrida se utilizó el año 2013 como conjunto de prueba.

Se emplearon dos vistas con los datos, una para la información anterior del 2013, como conjunto de entrenamiento y otra vista con los datos del 2013 como el conjunto de pruebas.

Luego de varias pruebas, se decidió construir un modelo para proyectar cada hora y no todo el día completo, esto debido a que se obtuvieron mejores resultados realizándolo de este modo; así, el modelo final se presenta en la Tabla 22.

Nombre de columna	Uso	Tipo de datos	Tipo de contenido	Valores
ANIO	Entrada	Long	Continuo	2010 - 2012
DIA	Entrada	Long	Continuo	1 - 31
DIASEMANA	Entrada	Long	Continuo	1 - 7
ENERGIA	Entrada y predicción	Double	Continuo	668 - 937
FERIADO	Entrada	Long	Continuo	0 - 1
ID Tiempo		Date	Key Time	
MES	Entrada	Long	Continuo	1 - 12
SEMANA	Entrada	Long	Continuo	1 - 54
TRIMESTRE	Entrada	Long	Continuo	1 - 4
VACACIONES	Entrada	Long	Continuo	0 - 2

Tabla 22. Información del modelo Microsoft Time Series (elaboración propia).

Como se observa, la hora fue excluida del modelo, pues este dato ingresa como un parámetro en el filtro de datos cuando el modelo es procesado.

Con respecto a los parámetros utilizados, estos son presentados en la Tabla 23.

Nombre	Valor
AUTO_DETECT_PERIODICITY	0,6
COMPLEXITY_PENALTY	0,1
FORECAST_METHOD	MIXED
HISTORIC_MODEL_COUNT	1
HISTORIC_MODEL_GAP	10
INSTABILITY_SENSITIVITY	1
MAXIMUM_SERIES_VALUE	1700
MINIMUM_SERIES_VALUE	650
MINIMUM_SUPPORT	10
MISSING_VALUE_SUBSTITUTION	None
PERIODICITY_HINT	{1}
PREDICTION_SMOOTHING	0,5

Tabla 23. Parámetros del modelo (elaboración propia).

Los parámetros que se cambiaron del default, fueron únicamente MAXIMUM_SERIES_VALUE y MINIMUM_SERIES_VALUE, para ajustarlos según los datos de las tablas de entrada.

Luego, para realizar las pruebas, el modelo se debe ir actualizando, con la información real para simular el avance del tiempo, así, cuando se va a proyectar

el día n, se agrega la información real hasta el día n-2, (se excluye el día n-1, pues corresponderá al día en que se está realizando la proyección y aún no se encontraría disponible la información), de esta forma el modelo utilizará la mayor cantidad de información real disponible.

Este proceso se simuló con un ciclo en T-SQL, el cual se muestra en el Apéndice 4.

Finalmente, se comparó el resultado obtenido por la predicción, con los datos reales de 2013. Los resultados de esta comparación, se observan en la Tabla 24, donde se muestra el mejor resultado por día y el mejor resultado general. En la Figura 36 se aprecian gráficamente los errores para cada día, así se observa que para los días sábados es el único día en que este algoritmo presenta mejoras de un 19% con respecto al actual método, pero para los restantes días, se presentan resultados negativos importantes, es decir, la predicción se degrada con respecto a la actual.

Día	PDE	Microsoft Time Series	Mejora
Lunes	48,99	71,00	-44,91%
Martes	41,13	70,75	-71,99%
Miércoles	39,69	64,06	-61,39%
Jueves	36,86	69,84	-89,50%
Viernes	41,02	59,83	-45,86%
Sábado	41,47	33,47	19,30%
Domingo	46,17	54,19	-17,37%
General	42,36	61,72	-45,70%

Tabla 24. Comparación con Microsoft Time Series (elaboración propia).

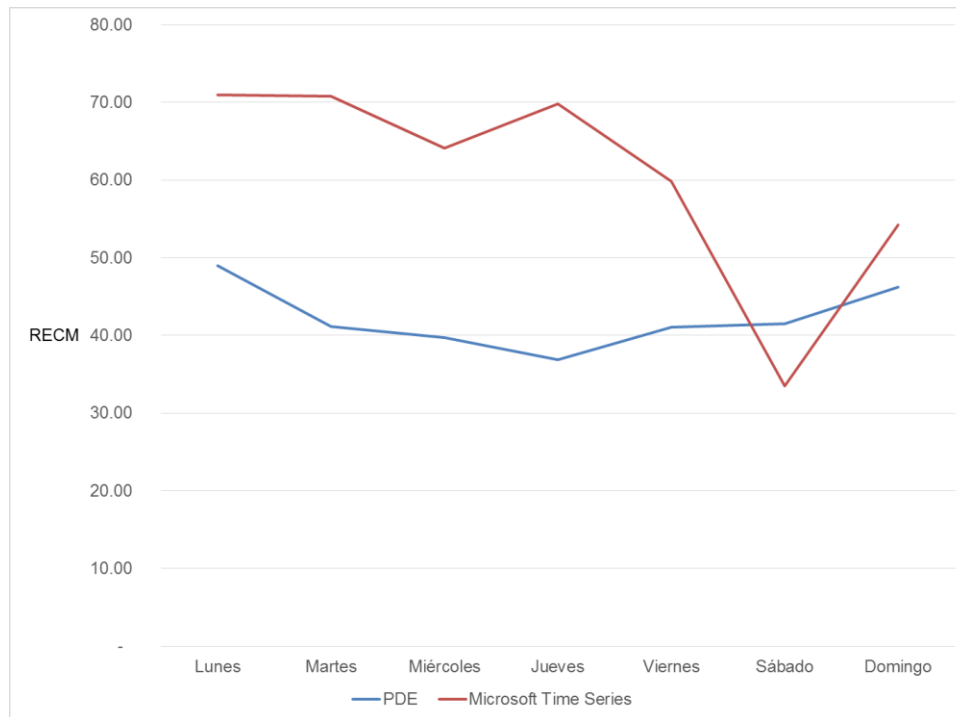


Figura 36. Comparación del error PDE vs Microsoft Time Series (elaboración propia).

Igualmente se debe aclarar, que en esta evaluación se están omitiendo los días feriados, sólo se consideran los días con un comportamiento típico.

En término general, la utilización de este método no mejora el proceso de proyección de la demanda.

Redes neuronales de Microsoft

El tercer y último método estudiado es la implementación de las redes neuronales incorporadas en el SQL Server.

Igualmente que el método anterior, se utiliza Microsoft Visual Studio para crear el modelo.

Luego de generar la red con los datos de entrada identificados en la Tabla 17, Tabla 18 y Tabla 19, se detectó que el comportamiento no era el esperado, por lo que se inició un proceso de ajustes para mejorar el rendimiento.

Luego de revisiones, se analizó tanto el método utilizado actualmente por PDE como los resultados obtenidos por el método de promedios móviles, se agregaron nuevas columnas de entrada, S1 a S10, las cuales incluyen la potencia del mismo día, a la misma hora de la semana anterior para S1, hasta diez semanas anteriores para S10.

De tal forma, las entradas de la red quedan como se muestra en la Tabla 25.

Nombre de columna	Uso	Tipo de datos	Tipo de contenido	Valores
ANIO	Entrada	Long	Continuo	2009 - 2014
DIA	Entrada	Long	Continuo	1 – 31
DIASEMANA	Entrada	Long	Cíclico	
ENERGIA	Entrada y predicción	Double	Continuo	572 - 1590
HORA	Entrada	Long	Cíclico	
ID Tiempo	Entrada	Date	Clave	
MES	Entrada	Long	Continuo	1 - 12
PUNTA	Entrada	Long	Continuo	0 - 1
S1	Entrada	Double	Continuo	561 - 1564
S10	Entrada	Double	Continuo	561 - 1590
S2	Entrada	Double	Continuo	572 - 1590
S3	Entrada	Double	Continuo	572 - 1572
S4	Entrada	Double	Continuo	561 - 1576
S5	Entrada	Double	Continuo	561 - 1590
S6	Entrada	Double	Continuo	561 - 1590
S7	Entrada	Double	Continuo	561 - 1590
S8	Entrada	Double	Continuo	561 - 1590
S9	Entrada	Double	Continuo	561 - 1590
SEMANA	Entrada	Long	Cíclico	
TIPO	Entrada	Long	Continuo	0 - 6
VACACIONES	Entrada	Long	Continuo	0 - 2

Tabla 25. Datos de entrada a la red neuronal (elaboración propia).

Se omitió la columna de trimestre, dado que no aportaba nada al modelo.

El conjunto de datos incluye cinco años, se probó variando el tamaño del conjunto, pero éste fue el de mejor desempeño. De este conjunto se consideró el 30% como datos de prueba y el 70% como datos de entrenamiento, según la recomendación de Microsoft.

Finalmente, haciendo uso del complemento de minería de datos para Excel, se realizó la proyección de la demanda del año 2013 completa, para compararla con el dato real. Los resultados de esta evaluación se muestran en la Tabla 26 y en la Figura 37.

Día	PDE	RN	Mejora
Lunes	48,99	42,25	13,76%
Martes	41,13	49,69	-20,81%
Miércoles	39,69	43,23	-8,92%
Jueves	36,86	44,86	-21,72%
Viernes	41,02	44,94	-9,55%
Sábado	41,47	33,47	19,30%
Domingo	46,17	43,65	5,45%
General	42,36	43,39	-2,44%

Tabla 26. . Comparación con Redes Neuronales (elaboración propia).

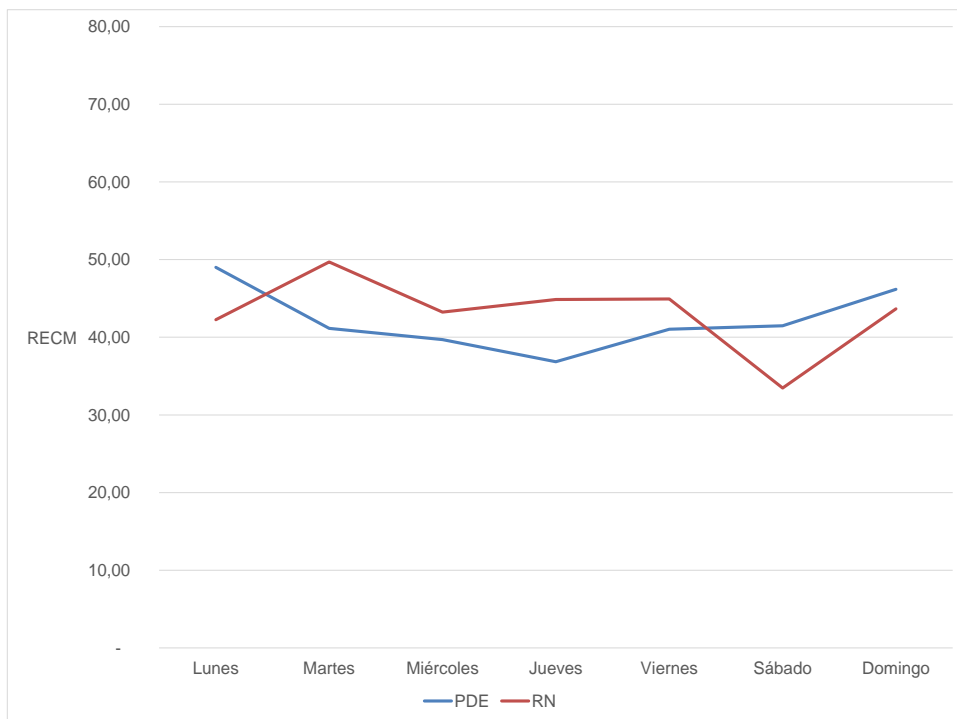


Figura 37. Comparación del error PDE vs Microsoft Time Series (elaboración propia).

Como se aprecia en tres días se notan mejoras: lunes, sábado y domingo. Donde se presenta la máxima mejora es para el día sábado, igual que el método

anterior. Pero para los días martes, miércoles y jueves se tiene un rendimiento inferior al método actual, especialmente martes y jueves.

El rendimiento general apenas es un 2.4% inferior, por lo se considera que es un método candidato para su implementación, por su cercanía con el método actual, pero mejoraría en el tiempo de su ejecución. Incluso con un mayor entrenamiento se podría mejorar aún más.

5.5. Implementación

Implementación de promedios móviles

De acuerdo con los resultados obtenidos, el algoritmo seleccionado para su implementación es el de promedios móviles. Este algoritmo, adicionalmente a las mejoras obtenidas en la revisión previa, tiene las ventajas de que es simple de implementar y no requiere ninguna licencia o costo adicional para su implementación.

Primeramente para proceder es necesario crear dos tablas adicionales que se crearán en la base de datos fuente. Estas tablas son:

Promóviles: contiene la información por día y hora, de la cantidad de datos que se deben promediar para obtener la proyección. La estructura de esta tabla se muestra en la Tabla 27.

Demanda_PM: contiene la proyección resultante utilizando el método de promedios móviles; su estructura se muestra en la Tabla 28.

Columna	Tipo	Descripción
DIA	entero	Día de la semana, 1 Lunes, 2 martes, etc.
HORA	entero	Hora del día, de 0 a 23.
DIAS_PROM	entero	Cantidad de datos por utilizar para realizar el promedio.

Tabla 27. Estructura de la tabla Promoviles (elaboración propia).

Columna	Tipo	Descripción
FECHA	Fecha	Fecha y hora correspondiente a la demanda proyectada.
DEMANDA	flotante	Demanda proyectada.
FECHA_DIGITA	fecha	Fecha y hora cuando se realizó la proyección.

Tabla 28. Estructura de la tabla Demanda_PM (elaboración propia).

El pseudocódigo para realizar la proyección con este método es:

Proyección_PM (Fecha)

EsFeriado ← Si se trata de una *fecha* correspondiente a un feriado

Si EsFeriado y es un día laboral *entonces*

Si la fecha corresponde a Día de la Madre, Navidad, Año Nuevo, Jueves Santo o Viernes Santo *entonces*
la *fecha* se trata como un domingo

En caso contrario

la *fecha* se trata como un sábado

Si existía una proyección anterior entonces

se elimina

Para cada hora de las 00 a las 23 horas del día

PM ← Se obtiene de la tabla *Promóviles* la cantidad de días para realizar el promedio, según el día de la semana de la *fecha* y la *hora*.

Se calcula el promedio de los *PM* días anteriores

Se guarda el dato de la proyección en la tabla *Demanda_PM*

El algoritmo se implementó en un procedimiento almacenado, de tal forma que está disponible para ser invocado con mayor facilidad y desde diversas fuentes, ya sea vía comando, desde una tarea programada o desde otra aplicación.

El listado completo del procedimiento se adjunta en el Apéndice 5.

Una vez implementado el método, se procedió a realizar pruebas y ajustes. En este proceso se encontró una variante para el método; inicialmente el método de promedios móviles y el método actualmente utilizado, proponen tratar los feriados como si fueran domingos, pero se observó que algunos feriados se comportan más como un sábado y otros como domingos, así se incluyó que los feriados de Año Nuevo, Día de la Madre, Navidad, Jueves y Viernes Santo se deben tratar como domingos, pero los restantes días serán tratados como sábados.

Seguidamente, para realizar la comparación de los resultados obtenidos, se tomaron varios conjuntos de datos:

- **2013-2014 Total:** todos los datos incluyendo del 1 de enero de 2013 al 30 de junio de 2014.
- **2013-2014 Feriados:** Se incluyeron solamente los feriados de Año Nuevo, Navidad, Jueves y Viernes Santo comprendidos desde el 1 de enero de 2013 al 30 de junio de 2014 (4 días al año).
- **2013-2014 Sin feriados:** los datos desde el 1 de enero de 2013 al 30 de junio de 2014, excluyendo los feriados del grupo anterior.
- **Junio 2014:** último mes que contiene datos reales en la base de datos de trabajo.
- **Última semana de junio, 2014:** última semana de la base de datos de trabajo.

En la Figura 38 se muestra el gráfico de dispersión para el método que se utiliza actualmente, y en la Figura 39 se presenta el recíproco utilizando el método implementado. La línea verde muestra el ajuste perfecto a modo de referencia, en tanto la línea roja muestra la tendencia lineal. Como se observa, la tendencia del método propuesto es más cercana a la realidad.

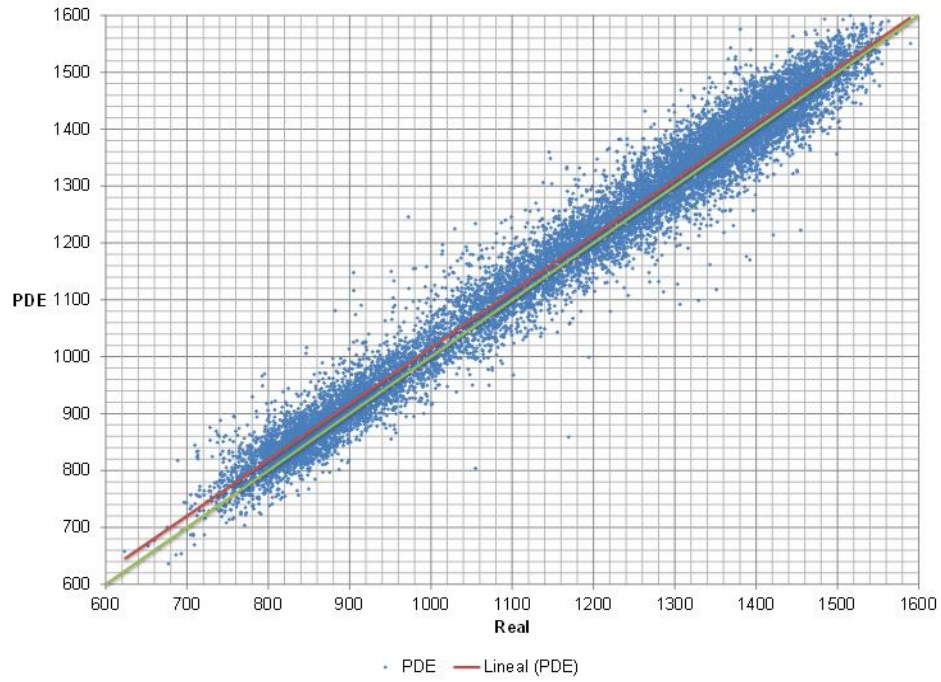


Figura 38. Gráfico de dispersión del método actual PDE (elaboración propia).

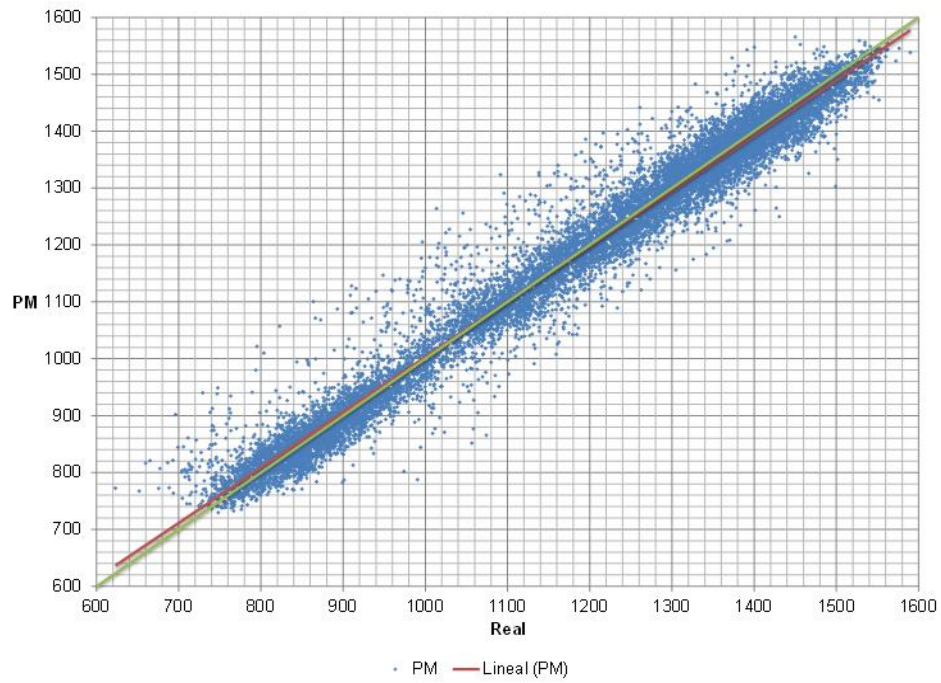


Figura 39. Gráfico de dispersión del método de promedios móviles (elaboración propia).

Adicionalmente la Tabla 29, muestra los resultados obtenidos comparando el método actual (PDE) con el de promedios móviles (PM).

Períodos	RECM			EMA		
	PDE	PM	Mejora	PDE	PM	Mejora
2013-2014 Total	41.64	39.25	2.39	2.82%	2.50%	0.32%
2013-2014 Sin Feriados	41.63	37.17	4.46	2.82%	2.39%	0.43%
2013-2014 Feriados	42.25	117.35	-75.10	3.41%	11.30%	-7.89%
Junio, 2014	47.25	36.04	11.21	3.17%	2.44%	0.73%
Última semana de Junio, 2014	66.32	49.15	17.17	4.52%	3.28%	1.25%

Tabla 29. Comparación de los resultados obtenidos con promedios móviles (elaboración propia).

Para efectos comparativos, se incluyó además del RECM, el EMA o Error Medio Absoluto, el cual brinda el error obtenido en términos de porcentaje.

Como se observa, los errores disminuyen para el conjunto total y mejora aún más eliminando los feriados identificados, pero para el caso de los feriados, este método no se recomienda, debido a que su efectividad decae de manera importante, por lo que para estos cuatro días, se recomienda utilizar el método actual o realizar un ajuste a la proyección.

De tal forma, si se excluyen los cuatro feriados, se obtendría una mejora de 0.43% con respecto al método actual según el EMA, lo que equivaldría a una mejora de 2.39 MWh según el RECM. Gráficamente la mejora en la proyección se muestra en la Figura 40.

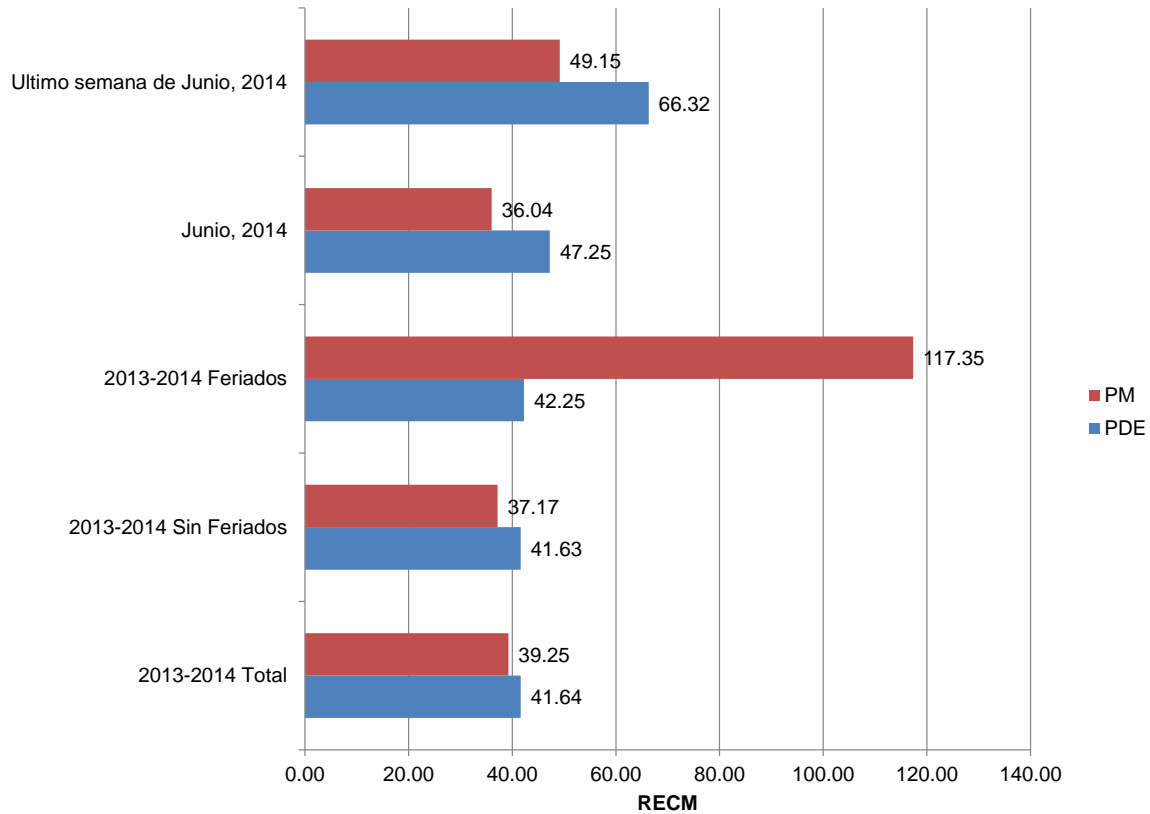


Figura 40. Comparación del método actual PDE vs Promedios Móviles (elaboración propia).

Implementación de redes neuronales

Los resultados obtenidos con las redes neuronales de Microsoft dieron un resultado cercano al actual, pero en busca de mejorar su rendimiento, se optó por utilizar otra herramienta para su modelado, a saber el MatLab.

Recientemente la empresa adquirió una licencia de MatLab, de la cual no se han explotado sus características relacionadas con redes neuronales, por lo que se optó por incluirla en el proyecto.

MatLab (abreviatura de MATrix LABoratory, "laboratorio de matrices") es un software matemático que ofrece un entorno de desarrollo integrado, que incluye un lenguaje de programación de alto nivel propio de la herramienta, denominado lenguaje M.

Entre sus características básicas se incluyen: la manipulación de matrices, la representación de datos y funciones, la implementación de algoritmos, la creación de interfaces de usuario y la comunicación con otras aplicaciones, para el

interés del proyecto, la comunicación con SQL Server. Además MatLab puede ampliar sus capacidades básicas con la inclusión de cajas de herramientas (toolboxes) en particular existe una para redes neuronales, que será la utilizada en el presente proyecto.

Para modelar la red se utilizó el asistente de MatLab llamado “*Neural Network Fitting Tool*”, el cual se utiliza para problemas donde se requiere mapear un conjunto numérico de datos de entrada contra un conjunto de datos objetivos.

Como entradas de la red se emplearon las mismas entradas utilizadas en la red neuronal de Microsoft, adicionalmente se incluyeron dos columnas, las que en el estudio anterior se descartaron por no aportar en la solución. No obstante, en este modelo si surtió efecto su inclusión, a saber, las entradas E1 y E2. La Tabla 30 resume el conjunto de entrada.

Entrada	Comentario
HORA	Hora
TIPO	Tipo de hora
PUNTA	Si se trata de hora punta o no
ANIO	Año
TRIMESTRE	Número de trimestre
MES	Número del mes
SEMANA	Número de la semana
DIASEMANA	Número del día de la semana
DIA	Día del mes
FERIADO	Bandera si se trata de un día feriado o no
VACACIONES	Bandera si se trata de un día de vacaciones escolares
S1	Energía de 1 semana antes a la misma hora
S2	Energía de 2 semanas antes a la misma hora
S3	Energía de 3 semanas antes a la misma hora
S4	Energía de 4 semanas antes a la misma hora
S5	Energía de 5 semanas antes a la misma hora
S6	Energía de 6 semanas antes a la misma hora
S7	Energía de 7 semanas antes a la misma hora
S8	Energía de 8 semanas antes a la misma hora
S9	Energía de 9 semanas antes a la misma hora
S10	Energía de 10 semanas antes a la misma hora
E1	Energía total del día de 1 semana antes
E2	Energía total del día de 2 semanas antes

Tabla 30. Entradas para la red neuronal (elaboración propia).

Este conjunto de entradas se definió a partir de los hallazgos de los promedios móviles, donde el 90% de los promedios utilizan diez o menos datos para realizar la proyección, por lo que se escogieron diez datos de energía de las semanas anteriores para el mismo día a la misma hora. Luego, el método actualmente utilizado por PDE, se vale de la energía total del día de las semanas anteriores, para determinar el crecimiento de energía, así, estos dos datos se incluyeron en las entradas. Con este conjunto de entradas se procedió al entrenamiento de la red. Cabe señalar que por la falta de experiencia en la herramienta, este entrenamiento no fue exhaustivo, más bien fue un método heurístico. Se utilizaron datos de cinco años, que es el periodo que se utiliza actualmente, de los cuales el 70% fue para entrenamiento, el 15% para validación, es decir, datos que se utilizan en el proceso de entrenamiento para verificar como va mejorando y cuando se debe detener el proceso y el restante 15%, para realizar pruebas del comportamiento que se obtuvo con la red encontrada. El diagrama de la red resultante se muestra en la Figura 41.

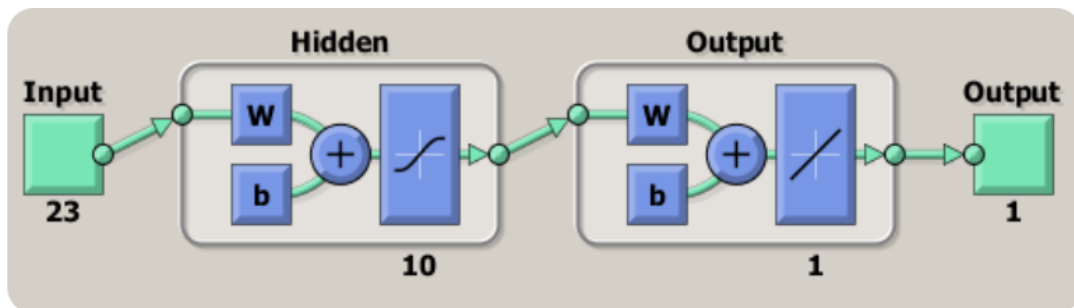


Figura 41. Red neuronal obtenida del entrenamiento (elaboración propia).

La red encontrada está compuesta de veintitrés entradas, diez neuronas ocultas y una salida.

Para implementar este método se hace uso de dos tablas adicionales:

CargaRN: tabla con los datos de entrada para la red, según las entradas definidas en la Tabla 30.

PronósticoMatLab: tabla donde se depositará el resultado de la ejecución de la red, su estructura es la misma mostrada en la Tabla 28.

Luego se sigue el siguiente pseudocódigo

```

Proyección_RN (Fecha)
-- En SQL SERVER
SELECT INTO CargaRN FROM Datos de entrada de la red
-- En MatLab
Entrada ← Cargar datos de CargaRN
Salida ← Ejecutar red neuronal (Entrada)
INSERT INTO pronosticoMatLab FROM Salida

```

Los scripts de SQL Server y de MatLab necesarios para implementar esta solución, se incluyen en los Apéndices 6 y 7.

Se procedió a realizar la evaluación respectiva. En la Figura 42 se presenta el gráfico de dispersión, el cual muestra una mejora con respecto a los promedios móviles, pero una tendencia muy similar entre ambos métodos.

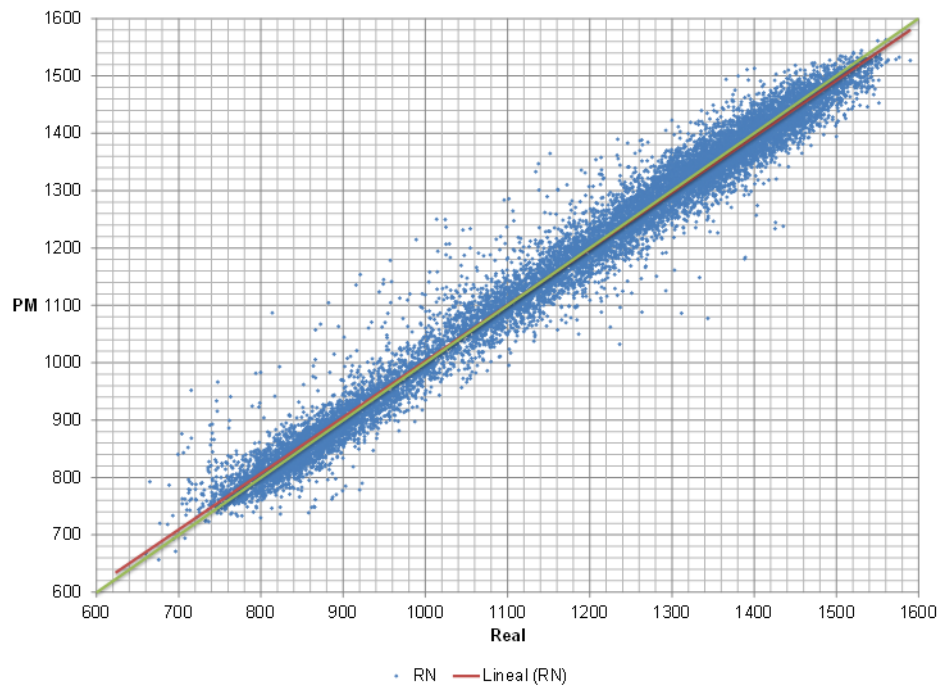


Figura 42. Gráfico de dispersión del método de redes neuronales (elaboración propia).

Se utilizaron los mismos cuatro escenarios definidos anteriormente, para determinar la mejora obtenida, los resultados obtenidos se presentan en la Tabla

31. Comparación de los resultados obtenidos con redes neuronales (elaboración propia)..

Períodos	RECM			EMA		
	PDE	RN	Mejora	PDE	RN	Mejora
2013-2014 Total	41.64	34.04	7.60	2.82%	2.18%	0.64%
2013-2014 Sin Feriados	41.63	32.74	8.89	2.82%	2.11%	0.70%
2013-2014 Feriados	42.25	88.72	-46.47	3.41%	7.37%	-3.96%
Junio, 2014	47.25	31.21	16.04	3.17%	2.08%	1.09%
Última semana de Junio, 2014	66.32	39.59	26.73	4.52%	2.52%	2.00%

Tabla 31. Comparación de los resultados obtenidos con redes neuronales (elaboración propia).

Como se observa, los errores se disminuyen para el conjunto total, e igualmente mejora aún más cuando se eliminan los feriados identificados, pero, igualmente que los promedio móviles, para el caso de los feriados, este método baja su efectividad, por lo que para estos días, es necesario un mayor trabajo, que incluso podría funcionar realizando ajustes en la red. Para estos días, trabaja mejor el método actualmente utilizado.

Con redes neuronales, excluyendo los cuatro feriados, se obtendría una mejora de 0.7% con respecto al método actual según el EMA, lo que equivaldría a una mejora de 8.89 MWh según el RECM. En la Figura 43 se muestra la mejora obtenida, además se incluye la comparación con promedios móviles para tener el panorama de las soluciones propuestas.

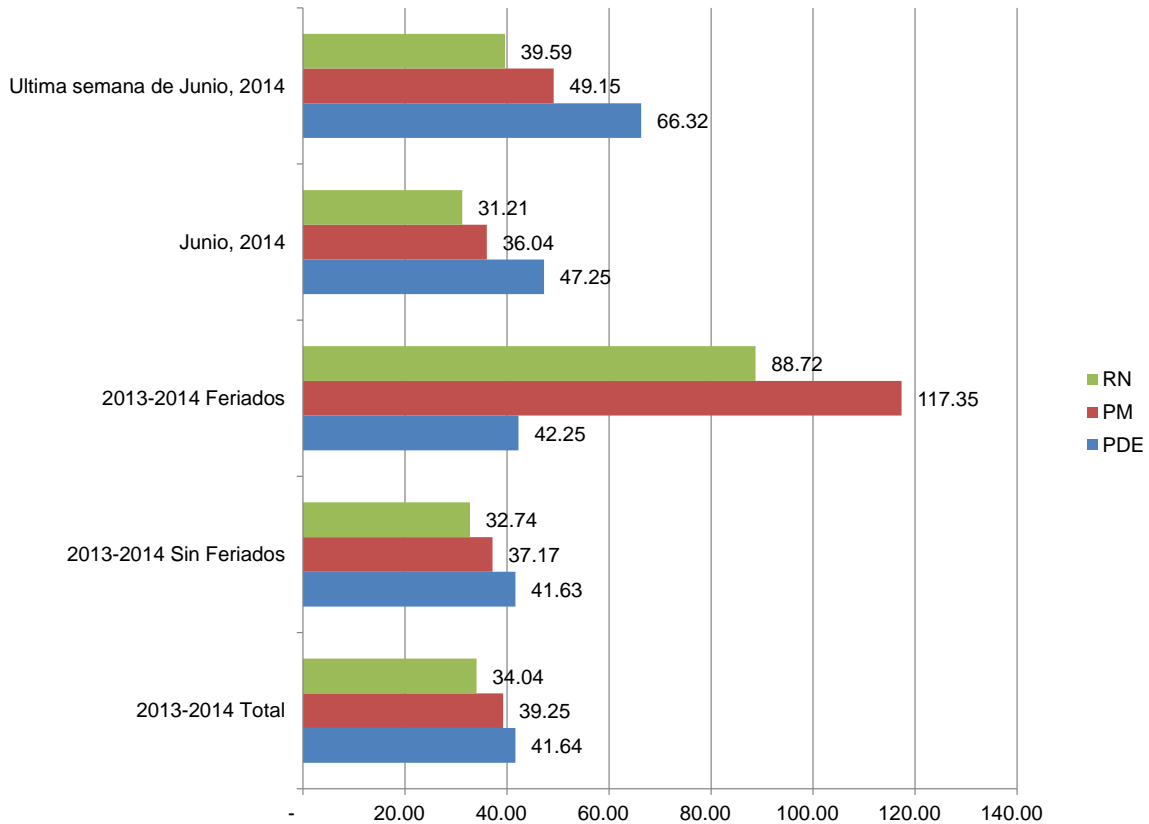


Figura 43. Gráfico comparativo de los métodos propuestos vs el actual (elaboración propia).

Como se aprecia, el uso de las redes neuronales mejora en todos los escenarios a los promedio móviles y solamente para los días feriados es que se presenta una desventaja con respecto al método actual. Pero como se había señalado anteriormente estos son sólo cuatro días en total al año, por lo que la mejora es evidente.

6. Conclusiones

Se describió de forma detallada el proceso actual de cómo se realiza la proyección de la demanda, esto resulta de suma importancia para la comprensión del problema y el cómo se puede modelar una solución. A partir de la descripción expuesta se logró identificar información clave, como por ejemplo, el uso del número de la semana y el crecimiento entre días. Por otro lado, al describir el método, se documentó de manera formal un proceso que estaba desarrollado en archivos Excel y documentado sin la rigurosidad matemática finalmente descrita.

Se definieron dos métodos para realizar la proyección de la demanda, a saber, el primero utilizando promedios móviles y un segundo con redes neuronales.

- El uso de promedios móviles demostró que este método brinda muy buenos resultados, a pesar de la simplicidad de su algoritmo, pues en términos generales aporta una mejora de 0.32% con respecto al método actual y su implementación es bastante sencilla, además de ser una alternativa económica pues no requiere de licenciamiento adicional. Razones por las cuales su implementación se considera un aporte importante para la academia.
- El segundo método definido e implementado es con redes neuronales, debido a que las redes de Microsoft se acercaron al margen de error actual, pero no lo mejoraron, se optó por utilizar un producto adicional, el MatLab, con el cual se obtuvieron los mejores resultados, una mejora general de 0.64%. Este método se constituye en la solución a implementar por el CENCE.

Se estructuró e implementó un depósito de datos con la información de energía horaria que es el requisito indispensable para el proceso de proyección de la demanda; además, en el data mart se incluyeron las variables de energía, plantas y niveles de embalses que son los insumos necesarios para realizar procesos de análisis del SEN.

Se esbozó el mecanismo para que las instancias adecuadas resuelvan el proceso de proyección de la demanda mediante la implementación del algoritmo

descrito. Con los métodos de proyección de la demanda definidos y el depósito de datos debidamente estructurado, se implementaron estos métodos con las herramientas que actualmente cuenta el CENCE, a saber, SQL Server y MatLab. Su implementación queda finalmente encapsulada en un procedimiento que se debe ejecutar ya sea de manera automática o manual, lo cual deberá ser definido según las necesidades del negocio. Así, para el usuario es un proceso sencillo o incluso podría ser transparente. Con esta implementación, se brindan los mecanismos para apoyar al personal responsable en el proceso de proyección de la demanda.

Se analizaron los resultados obtenidos mediante un proceso de comparación en el cual se definieron diversos conjuntos de datos para pruebas, obtenidos durante el año 2013 y primer semestre de 2014. Los resultados son positivos a favor de los métodos implementados, salvo en el conjunto de datos de los días feriados, compuesto por cuatro feriados, a saber: primero de enero, Jueves Santo, Viernes Santo y Navidad. Para estos días, los métodos implementados se mantuvieron inferiores al método actual. De tal forma, para estos cuatro días festivos la proyección obtenida requerirá de trabajo adicional o en su defecto, utilizar el método actual para su proyección.

En términos generales, se construyó una plataforma de inteligencia de negocios que soporta e incluye el proceso de proyección de la demanda nacional y el análisis del comportamiento del SEN. Esto se implementó utilizando SQL Server 2012 y la solución se basa un data mart que incluye los datos de generación del Sistema Eléctrico Nacional; adicionalmente se construyeron cubos con Analysis Services, lo que brinda flexibilidad y una mayor velocidad en la construcción de reportes y por ende en el análisis de la información. Como herramientas de consultas se incluyeron hojas electrónicas en Excel que ofrecen las consultas básicas identificadas, pero que se pueden ampliar aún más según las necesidades.

El data mart, al incluir toda la información histórica, se constituye en una fuente única de los datos de generación, se logra mayor consistencia en la información dentro de la organización.

Adicionalmente, en esta plataforma definida, se incluyó la implementación de dos métodos para la proyección de la demanda completamente funcionales: promedio móviles y redes neuronales.

Ambos métodos tienen las ventajas de que mejoraron el margen de error en la proyección y reducen el esfuerzo requerido por el personal encargado de manera significativa, dado que el tiempo para el proceso de proyección se reduce horas a segundos. Por otro lado, como los datos utilizados para la proyección son de una semana anterior, los métodos implementados permiten realizar la proyección hasta seis días hacia adelante sin ningún problema, ni pérdida de precisión.

7. Recomendaciones

En el data mart se incluyeron las variables que se identificaron como las más comúnmente consultadas. Existen otras que igualmente se pueden incluir, lo cual podrá realizarse en subsiguientes ciclos de la metodología de Kimball, en lo que constituye el proceso de descubrimiento. En estas etapas es necesario definir con los líderes del negocio los objetivos para el data mart y los criterios de éxito, esto permitirá que los alcances sean claros a todos los involucrados y no se presente un crecimiento del data mart simplemente por los éxitos previos.

Actualmente se están incluyendo en el data mart ciento cincuenta y dos variables del SCADA, pero en total existen cerca de cinco mil que potencialmente pueden ingresar al data mart, por lo que su inclusión debe ser analizada con sumo cuidado y atención.

Es recomendable la participación de los expertos del negocio y los usuarios de la información para validar del proceso de desarrollo o crecimiento del data mart, con el fin de asegurar el cumplimiento de los requisitos de información y en pro del logro de los objetivos.

El data mart ayuda en las tomas de decisiones, en la facilidad de la publicación, en contar con una única fuente de información. No obstante, debe estar bajo una gestión administrativa adecuada, por lo que es necesario incluir la gobernanza de datos, entendiéndose ésta como la disciplina del tratamiento de los datos como un activo empresarial. Lo que involucra el ejercicio de decisiones adecuadas para optimizar, proteger y aprovechar los datos como un activo empresarial. Se trata de la orquestación de personas, procesos, tecnología y las políticas dentro de una organización, para obtener el valor óptimo de los datos empresariales.

Es recomendable incluir un proceso de calidad de datos, pues los utilizados en el presente proyecto, por su importancia, tienen un proceso de calidad inmerso en su ciclo de vida, pero hay muchos otros, que en su mayoría provienen de fuentes automáticas de captura de datos, como medidores inteligentes o equipos de campo, que por el volumen de información no reciben un proceso de limpieza. Una opción es incluir la bandera de calidad que igualmente proviene de la fuente y

así validar la calidad con que el dato fue capturado, convirtiéndose la calidad en una dimensión más del data mart.

En los procesos de minería de datos, al utilizar la implementación propuesta de Microsoft de los métodos ARIMA y redes neuronales, se obtuvo resultados inferiores a los esperados, pero al utilizar el mismo modelo de las redes neuronales en la plataforma de MatLab, los resultados fueron mucho más precisos, por lo que, para las implementaciones actuales y al menos para redes neuronales, lo recomendable es utilizar un paquete más especializado, como el analizado en el presente proyecto, a saber MatLab, pero bien podrían analizarse otros más.

Con respecto a la proyección de la demanda, es recomendable realizar un estudio más detallado de los días feriados, pues el método propuesto tiene una menor efectividad que el actual, se podrían efectuar modificaciones en el conjunto de entrenamiento de la red, o en su defecto, crear una nueva red para tratar estos casos particulares. Esto con la finalidad de cubrir el 100% de los días del año con estos métodos y liberar al personal encargado de esta tarea.

Se recomienda realizar un procedimiento o programación de un script en el lenguaje M de MatLab, con la finalidad de realizar un entrenamiento exhaustivo de la red neuronal, encontrar una red óptima ajustando los parámetros y número de neuronas, lo cual redundaría en disminuir el error de la proyección.

Finalmente con respecto a las herramientas de consulta y publicación de información contenida en el data mart, además del Excel, es recomendable utilizar las herramientas de SharePoint y Reporting Services, para mejorar aún más la experiencia del usuario. Se tiene la ventaja de que la empresa cuenta con las licencias respectivas, por lo que no involucraría ningún costo adicional.

Glosario

CENCE. Centro Nacional Control de Energía, Unidad estratégica de negocio perteneciente al ICE Sector de electricidad, que tiene la tarea de operar el SEN.

Demanda. (*Eléctrica*). Se refiere a la cantidad de energía que requiere el país para satisfacer sus necesidades de energía en un periodo de tiempo dado.

EMA. Error medio absoluto, representa el error en términos porcentuales.

Energía no renovable. Energía obtenida a partir de combustibles fósiles (líquidos o sólidos) y sus derivados.

Energía renovable. Energía obtenida de los recursos naturales y de desechos, tanto industriales como urbanos. Incluyen la hidráulica, solar, eólica, residuos sólidos (industriales y urbanos) y biomasa.

ICE. Instituto Costarricense de Electricidad.

MER. Mercado Eléctrico Regional, mercado de electricidad compuestos por los países centroamericanos.

PDE. Proceso Planeamiento y Despacho de Energía, dependencia del CENCE que se encarga del planeamiento a corto plazo de la operación del SEN.

Período de Mercado. Intervalo de tiempo en que se divide el día para efecto del predespacho de transacciones de energía

Postdespacho. Se refiere a la operación real que se presentó en el sistema, son datos reales.

Predespacho. Programación de las transacciones de energía y de la operación del Sistema Eléctrico Nacional para el día siguiente y cumplir con la demanda del sistema, el cual se realiza por período de mercado.

RECM. Raíz del error cuadrático medio, representa el error en las unidades de medidas del muestreo.

Redespacho. Modificación de la programación efectuada en el predespacho, debido a cambios en las condiciones con las cuales se realizó el predespacho.

SCADA. Acrónimo de Supervisory Control And Data Acquisition (Supervisión, Control y Adquisición de Datos) es un software que permite controlar

y supervisar procesos industriales a distancia. Se retroalimentan en tiempo real con los dispositivos de campo (sensores y actuadores) y controla el proceso de manera automática.

SEN. Sistema Eléctrico Nacional de Costa Rica.

Referencias

- Acevedo, C. (2004). *Predicción de demanda de corto plazo empleando redes neuronales*. Facultad de Ingenierías Físico-Mecánicas. Bucaramanga, Colombia: Universidad Industrial de Santander.
- Bao, J. (2000). *Short-term Load Forecasting based on Neural network and Moving Average*. Obtenido de CiteSeer:
<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.67.3268&rep=rep1&type=pdf>
- Calvo, C. (2013). *Entrenamiento de Redes Neuronales para Agente Especializado*. Tesis para Optar al Grado de Magister Scientiae en Computación, Instituto Tecnológico de Costa Rica, Escuela de Ingeniería en Computación, Cartago.
- Carvajal, P. (2003). *Estudio de la Demanda de Energía Eléctrica, utilizando Modelos de Series de Tiempo*. Recuperado el 27 de enero de 2014, de Scientia et Technica:
http://revistas.utp.edu.co/index.php/revistaciencia/article/download/7379/4403&ei=SVTwUu2jA4iskAfdy4HIBw&usg=AFQjCNEA4FvPYI2MBtxu_o4bUBwy2NfJDw&sig2=Kss_NkoBfQEt2SzaxYunzA&bvm=bv.60444564,d.eW0
- CENCE. (2014). *Intranet CENCE*. Obtenido de
<http://sabcence04/intranet/Pages/index.aspx>
- Chai, T., & Draxler, R. (10 de febrero de 2014). *www.geosci-model-dev.net*. Obtenido de <http://www.geosci-model-dev.net/7/1247/2014/gmd-7-1247-2014.html>
- D'oro, A. L., Lozano, C. A., & Moreno, C. A. (2007). *Modelo de promedios móviles para el pronóstico horario de potencia y energía eléctrica*. Recuperado el 27 de enero de 2014, de El Hombre y la Máquina:
<http://www.redalyc.org/articulo.oa?id=47802911>
- Elliot, J. (1993). *El cambio educativo desde la investigación–acción*. Madrid: Morata.

- Fernández Jiménez, L. A. (2007). *Modelos Avanzados para la predicción a corto plazo de la producción eléctrica en parques eólicos*. Logroño, España: Universidad de La Rioja.
- Gaceta, L. (13 de Julio de 2012). Fortalecimiento de la ley fundamental de educación. *La Gaceta*.
- Gartner. (20 de febrero de 2014). *Magic Quadrant for Business Intelligence and Analytics Platforms*. Obtenido de Gartner:
<http://www.gartner.com/technology/reprints.do?id=1-1QLGACN&ct=140210&st=sb>
- Hammergren, T. C. (2009). *Data Warehousing For Dummies* (Segunda ed.). Indianapolis, Indiana: Wiley Publishing, Inc.
- Hernández, R., Fernández-Collado, C., & Baptista, P. (2010). *Metodología de la investigación* (5ed ed.). México: Ed. McGraw Hill.
- Huerta, A., Quispe, J., Ramos, E., Fernández, E., & Molina, Y. (25 de julio de 2012). *Aplicación de Redes Neuronales para el Pronóstico de Demanda a Corto Plazo*. Recuperado el 25 de enero de 2014, de SectorElectricidad: <http://www.sectorelectricidad.com/2625/aplicacion-de-redes-neuronales-para-el-pronostico-de-demanda-a-corto-plazo/>
- IBM Corporation. (2012). *Manual CRISP-DM de IBM SPSS*. USA: IBM.
- IBM Corporation, International Technical Support Organization. (2001). *Building the Operational Data Store on DB2 UDB* (Primera ed.). San Jose, California: Redbooks.
- Inmon, W. H. (2005). *Building the Data Warehouse* (Cuarta Edición ed.). Wiley Publishing, Inc.
- Jain, A. (2009). *Short Term Load Forecasting by Clustering Technique based on Daily Average and Peak Loads*. International Institute of Information Technology, Centre for Power Systems, Hyderabad. India.
- Kimball, R., & Ross, M. (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling* (Segunda Edición ed.). New York: John Wiley and Sons, Inc.

- Kimball, R., Ross, M., Thornthwaite, W., Mundy, J., & Becker, B. (2011). *The Data Warehouse Lifecycle Toolkit*. Hoboken: John Wiley & Sons.
- Korhonen, J. J. (s.f.). Recuperado el 30 de junio de 2013, de www.jannekorhonen.fi/
- MacLennan, J., Tang, Z., & Crivat, B. (2009). *Data mining with Microsoft SQL server 2008*. Indianapolis: Wiley Pub.
- Mallo González., C. (2003). *Predicción de la demanda eléctrica*. Recuperado el 24 de enero de 2014, de <http://www.uv.es/asepuma/recta/ordinarios/5/5-1.pdf>
- Navarro, O. (2013). *Diseño de un Data Mart para JASEC, modelado mediante UML Tesis de Maestría*. San José. Costa Rica: Universidad Cenfotec.
- Pablito. (2001). *dds. afre: sde*.
- Pita Fernández, S., & Pértegas Díaz, S. (2002). *Facultada de Ciencias de la Educaciòn*. Obtenido de Univesidad de Quebec: http://www.ecominga.uqam.ca/ECOMINGA_2011/PDF/BIBLIOGRAPHIE/GUIDE_LECTURE_2/4/2.Pita_Fernandez_y_Pertegas_Diaz.pdf
- REE. (2014). *Red Eléctrica de España*. Obtenido de <http://www.ree.es/es/>
- Tudela, G. N. (mayo de 2011). *Análisis y pronóstico de la demanda de potencia eléctrica en Bolivia: una aplicación de redes neuronales*. Recuperado el 24 de enero de 2014, de iisec: <http://www.iisec.ucb.edu.bo/journal/articulos/1502.pdf>
- Yanguas-Peña, A., Mendoza-Villena, M., Andrés, A. F.-D., Lara-Santillán, P., García-Garrido, E., & Zorzano-Alba, E. (2008). *Predicción a corto plazo de la demanda de energía en centros de transformación de baja tensión usando sistemas de inferencia difusa optimizados genéticamente*. Universidad de La Rioja, Departamento de Ingeniería Eléctrica, La Rioja (España).

Apéndices

Apéndice 1. Lista de elementos claves del sistema.

Tabla	Nombre	Descripción	Valores de ejemplo
CLASIFICACION	CLA_CODIGO	Código de la clasificación de las plantas	2, 6
	CLA_NOMBRE	Nombre de la clasificación	Posdespacho y Regulación, Clasificación por fuente
EMBALSES	COD_EMBALSE	Código del embalse, corresponde al mismo código de la planta	1, 2
	EMB_NOMBRE	Nombre del embalse	Garita, San Miguel
EMPRESAS	COD_EMPRESA	Código de la empresa	1, 4
	EMP_NOMBRE	Nombre de la empresa eléctrica	Instituto Costarricense de Electricidad, Compañía Nacional de Fuerza y Luz
	EMP_SIGLAS	Siglas de la empresa	ICE, CNFL
ENERGIA_PLAN	ENP_FECHA	Fecha del registro de la energía para una planta dada	01/01/2014, 02/01/2014
	PLA_CODIGO	Código de la planta de generación de energía	1, 2
	ENP_ENERGIA	Energía en kWh generada por una planta para un día dado	694 000.00, 3 809.00
	ENP_PREDESPACHO	Energía predespachada para una planta y para un día dado	700 000.00, 3 500.00
ENERGIA_UNI	CENT_PROD	Centro de producción de energía, equivale a las plantas, pero no para todas las plantas se tiene el detalle por unidad.	1, 2
	UNIDAD	Identificador de la unidad generadora	1, 2
	FECHA_DATO	Fecha correspondiente al dato registrado.	01/01/2014, 02/01/2014
	FECHA_DIGITA	Fecha y hora cuando se registró el dato	01/01/2014 06:00, 02/01/2014 06:14
	DATO	Energía en kWh generada por la unidad	213 000.00, 8 500
FERIADOS	Fecha	Fecha cuando se presentó un feriado extraordinario	28/03/1991, 29/03/1991
	Festividad	Descripción del día feriado	Jueves Santo, Viernes Santo

GRUPOS	GRU_CODIGO	Código del grupo de plantas	14, 15
	CLA_CODIGO	Clasificación bajo la cual se definió el grupo	1, 2
	GRU_NOMBRE	Nombre del grupo	Térmico, Geotérmico
	GRU_COLOR	Color que se utiliza para identificar a este grupo.	FF3333, FFE800
	GRU_ORDEN	Orden relativo del grupo dentro de la clasificación	1, 2
NIVELES	COD_EMBALSE	Código del embalse, corresponde al mismo código de la planta	1, 2
	FECHA_DATO	Fecha cuando se registró el dato	01/01/2014, 02/01/2014
	FECHA_DIGITA	Fecha y hora cuando se registró el dato	01/01/2014 06:00, 02/01/2014 06:14
	DATO	Dato del nivel de embalse registrado	546.25, 895.00
PLANTAS	PLA_CODIGO	Código de la planta de generación de energía	1, 2
	COD_EMPRESA	Código de la empresa a la que pertenece la planta	1, 4
	PLA_NOMBRE	Nombre de la planta	Garita, Arenal
	PLA_NCORTO	Nombre corto o siglas de la planta	GRT, ARN
PLANTAS_GRUPOS	GRU_ID	Identificador del grupo, es utilizado para crear la relación de muchos a muchos, entre las plantas y los grupos por clasificación	597, 598
	CLA_CODIGO	Código de la clasificación de las plantas	1, 2
	GRU_CODIGO	Código del grupo de plantas	14, 15
	PLA_CODIGO	Código de la planta de generación de energía que se está clasificando	1, 2
POTENCIA_UNI	CENT_PROD	Centro de producción de energía, equivale a las plantas, pero no para todas las plantas se tiene el detalle por unidad.	1, 2
	UNIDAD	Identificador de la unidad generadora	1, 2
	FECHA_DATO	Fecha cuando se registró el dato de potencia	01/01/2014, 02/01/2014
	FECHA_DIGITA	Fecha cuando y hora se registró el dato	01/01/2014 06:00, 01/01/2014 06:15
	DATO	Potencia en MW generada por la unidad	213 000.00, 8 500
PRE_DATOS	NOMBRE_PDE	Nombre dado por el proceso Planeamiento y Despacho de Energía al elemento que se puede despachar (puede ser a nivel de unidad o a	AeroEnergía, ARE-U1

		nivel de toda la planta).	
	FECHA_HORA	Fecha y hora correspondiente al despacho de energía que se realiza	01/01/2014 06:00, 02/01/2014 07:00
	DATO	Energía en kWh que se programa para ser generado por el elemento	4 894, 3196
	FECHA_CARGA	Fecha cuando se realizó la carga del predespacho en el sistema	16/04/2014 16:09:01, 2014-05-16 16:00:00
PRE_PLANTAS	NOMBRE_PDE	Nombre dado por el proceso Planeamiento y Despacho de Energía al elemento que se puede despachar (puede ser a nivel de unidad o a nivel de toda la planta).	AeroEnergía, ARE-U1
	PLA_CODIGO	Código de la planta de generación de energía correspondiente al elemento que se predespacha	1, 2
	UNI_CODIGO	Identificador de la unidad generadora	1, 2
	PLA_CODIGO_SCADA	Código de la planta de generación de energía correspondiente al elemento que se predespacha y que tiene medición asociada en el SCADA.	8, 10
UNDMETRICAS	COD_UNDMETRICA	Identificador de la unidad de medida	1, 2
	NOMBRE_UNIDAD	Nombre de la unidad de medida	GWh, msnm
	DESCRIPCION	Descripción de la unidad de medida	Giga Watts hora, Metros sobre el nivel del mar
UNIDADES	UNI_CODIGO	Identificador de la unidad generadora	1, 2
	PLA_CODIGO	Código de la planta de generación de energía correspondiente donde se ubica esta unidad	1, 2
	UNIDAD_HISTORIA DOR	Identificador de la medición SCADA relacionada con la unidad (no todas las unidades tienen medición)	GARITA .13.8 GEN1 P .AV, GARITA .13.8 GEN2 P .AV

Apéndice 2. Tablas del Modelado Dimensional.

Nombre de la Tabla: TD_EMPRESA
Tipo: Dimensión
Descripción: Empresas con participación en la generación de electricidad

Columna	Descripción	Tipo	Destino			Ejemplo	Fuente					
			Tamaño	Llave	FK		Sistema	Tabla	Columna	Tipo	Comentario	
IDEMPRESA	Identificador de la empresa	entero		PK	1		ETL					Nueva llave
NOMEMPRESA	Nombre de la empresa	texto	50			Instituto Costarricense de Electricidad	sICENCE	GRUPOS	GRU_NOMBRE	varchar(30)		Se filtran con la clasificación igual a 3
ORDEN	Orden para mostrar las empresas	entero			1		sICENCE	GRUPOS	GRU_ORDEN	int		
COLOR	Color para referenciar esta empresa	texto	6			FFFFFF	sICENCE	GRUPOS	GRU_COLOR	varchar(6)		

Nombre de la Tabla: TD_POSTDESPACHO
Tipo: Dimensión
Descripción: Grupos de plantas con el orden de cómo deben ser predespachadas para cumplir con la demanda del sistema.

Columna	Descripción	Tipo	Destino			Ejemplo	Fuente					
			Tamaño	Llave	FK		Sistema	Tabla	Columna	Tipo	Comentario	
IDPOSTDESPACHO	Identificador del grupo	entero		PK	1		ETL					Nueva llave
NOMPOSTDESPACHO	Nombre del grupo	texto	50			Geotérmico	sICENCE	GRUPOS	GRU_NOMBRE	varchar(30)		Se filtran con la clasificación igual a 2
ORDEN	Orden para mostrar los grupos	entero			1		sICENCE	GRUPOS	GRU_ORDEN	int		
COLOR	Color para referenciar este grupo	texto	6			FFFFFF	sICENCE	GRUPOS	GRU_COLOR	varchar(6)		

Nombre de la Tabla: TD_RECURSO
Tipo: Dimensión
Descripción: Grupos de plantas de acuerdo al tipo de recurso, que pueden ser renovables, térmico (las que utilizan combustibles fósiles) o intercambio, es decir que proviene de otro país, pero se desconoce el recurso que utilizan.

Columna	Descripción	Tipo	Destino			Ejemplo	Fuente					
			Tamaño	Llave	FK		Sistema	Tabla	Columna	Tipo	Comentario	
IDRECURSO	Identificador del recurso	entero		PK	1		ETL					Nueva llave
NOMRECURSO	Nombre del recurso	texto	50			Renovable	sICENCE	GRUPOS	GRU_NOMBRE	varchar(30)		Se filtran con la clasificación igual a 5
ORDEN	Orden para mostrar los grupos	entero			1		sICENCE	GRUPOS	GRU_ORDEN	int		
COLOR	Color para referenciar este grupo	texto	6			FFFFFF	sICENCE	GRUPOS	GRU_COLOR	varchar(6)		

Nombre de la Tabla: TD_FUENTE

Tipo: Dimensión

Descripción: Fuentes de energía utilizada por la planta dada para generar electricidad, como por ejemplo, hidroeléctrica, geotérmica, eólica, etc.

Columna	Descripción	Tipo	Destino				Ejemplo	Fuente			
			Tamaño	Llave	FK			Sistema	Tabla	Columna	Tipo
IDFUENTE	Identificador de la fuente	entero		PK	1		ETL				Nueva llave
NOMFUENTE	Nombre de la fuente	texto	50			Eólico	siCENCE	GRUPOS	GRU_NOMBRE	varchar(30)	Se filtran con la clasificación igual a 6
ORDEN	Orden para mostrar las fuentes	entero			1		siCENCE	GRUPOS	GRU_ORDEN	int	
COLOR	Color para referenciar esta fuente	texto	6			FFFFFF	siCENCE	GRUPOS	GRU_COLOR	varchar(6)	

Nombre de la Tabla: TD_ESCENARIO

Tipo: Dimensión

Descripción: Escenarios posibles donde se contextualiza el dato, los escenarios posibles son: Real, Programado o diferencia.

Columna	Descripción	Tipo	Destino				Ejemplo	Fuente			
			Tamaño	Llave	FK			Sistema	Tabla	Columna	Tipo
IDESCENARIO	Identificador del escenario	entero		PK	1		ETL				Nueva llave
NOMESCENARIO	Nombre del escenario	texto	50			Real	ETL				

Nombre de la Tabla: TD_REGION

Tipo: Dimensión

Descripción: Regiones que agrupan un conjunto de plantas, estas regiones no necesariamente son geográficas, más bien corresponden a la topología de la red del sistema eléctrico nacional.

Columna	Descripción	Tipo	Destino				Ejemplo	Fuente			
			Tamaño	Llave	FK			Sistema	Tabla	Columna	Tipo
IDREGION	Identificador de la región	entero		PK	1		ETL				Nueva llave
NOMREGION	Nombre de la región	texto	50			Norte	siCENCE	GRUPOS	GRU_NOMBRE	varchar(30)	Se filtran con la clasificación igual a 9

Nombre de la Tabla: TD_PLANTAS

Tipo: Dimensión

Descripción: Conjunto de plantas que aportan generación al país.

Columna	Descripción	Tipo	Destino				Ejemplo	Fuente			
			Tamaño	Llave	FK			Sistema	Tabla	Columna	Tipo
IDPLANTA	Identificador de la planta	entero		PK	1		ETL				Nueva llave
NOMPLANTA	Nombre de la planta	texto	50			Arenal	sICENCE	PLANTAS	PLA_NOMBRE	varchar(44)	
NOMCORTO	Nombre corto o abreviatura de la planta	entero				ARN	sICENCE	PLANTAS	PLA_NCORTO	char(3)	
SK_PLANTA	Identificador de la planta en el sistema origen.	entero				FFFFFF	sICENCE	PLANTAS	PLA_CODIGO	varchar(6)	

Nombre de la Tabla: TD_UNDMETRICA

Tipo: Dimensión

Descripción: Unidades de medidas que se utilizan para los datos registrados.

Columna	Descripción	Tipo	Destino				Ejemplo	Fuente			
			Tamaño	Llave	FK			Sistema	Tabla	Columna	Tipo
IDUNDMETRICA	Identificador de la unidad de medida	entero		PK	1		ETL				Nueva llave
NOMUNDMETRICA	Nombre de la unidad	texto	10			kWh	sICENCE	UNDMETRICAS	NOMBRE_UNIDAD	varchar(10)	
DESCRIPCION	Descripción de la unidad de medida	texto	60			Kilo watts hora	sICENCE	UNDMETRICAS	DESCRIPCION	varchar(60)	

Nombre de la Tabla: TD_FECHA
Tipo: Dimensión
Descripción: Listado de fechas

Columna	Descripción	Tipo	Destino			Ejemplo	Fuente				
			Tamaño	Llave	FK		Sistema	Tabla	Columna	Tipo	Comentario
IDFECHA	Identificador de la fecha	entero		PK		20140101	ETL				Nueva llave
FECHA	Fecha	fecha				01/01/2014	ETL				
ANIO	Año	entero				2014	ETL				
TRIMESTRE	Trimestre	entero				1	ETL				
MES	Mes	entero				1	ETL				
SEMANDA	Número de la semana	entero				1	ETL				
DIA	Día	entero				1	ETL				
DIASEMANA	Día de la semana	entero				3	ETL				
MTRIMESTRE	Nombre del trimestre	texto	7			T1/2014	ETL				
NMES	Nombre del mes	texto	15			Enero	ETL				
NCORTOMES	Nombre abreviado del mes	texto	3			Ene	ETL				
NSEMANA	Nombre de la semana	texto	14			Semana 1 /2014	ETL				
NDIA	Nombre del día	texto	6			1 Ene	ETL				
NDIASEMANA	Nombre del día de la semana	texto	10			Miércoles	ETL				
NCORTODIA	Nombre abreviado del día	texto	3			Mié	ETL				
NFERIADO	Corresponde a un feriado?	texto	2			SI	ETL				
FESTIVIDAD	Descripción de la festividad	texto	60			Año Nuevo	ETL				
FERIADO	Valor para día feriado	entero				1	ETL				
VACACIONES	Valor para día de vacaciones escolares	entero				1	ETL				
NVACACIONES	Es un día de vacaciones escolares?	texto	2			Si	ETL				
TIEMPOESCOALR	Descripción del tiempo escolar	texto	30			Vacaciones escolares	ETL				

Nombre de la Tabla: TD_HORAS
Tipo: Dimensión
Descripción: Listado de las horas del día.

Columna	Descripción	Tipo	Destino			Ejemplo	Fuente				
			Tamaño	Llave	FK		Sistema	Tabla	Columna	Tipo	Comentario
IDHORA	Identificador de la hora	entero		PK		330	ETL				Nueva llave
TIEMPO	Hora	hora				03:30	ETL				
NTIEMPO	Texto que representa la hora	texto	5			03:30	ETL				
HORA	Hora del día	entero				3	ETL				
MINUTO	Minutos	entero				30	ETL				
TIPO	Tipo de hora	entero				0	ETL				
NTIPO	Descripción del tipo	texto	25			Valle nocturno	ETL				
PUNTA	Valor para hora punta	entero				0	ETL				
NPUNTA	Descripción de hora punta	texto	25			Fuera de punta	ETL				

Nombre de la Tabla: TH_ENERGIA
Tipo: Hechos
Descripción: Contiene la energía diaria por planta

Columna	Descripción	Tipo	Destino				Ejemplo	Fuente				
			Tamaño	Llave	FK			Sistema	Tabla	Columna	Tipo	Comentario
IDFECHA	Identificador de la fecha	entero		PK	FK	20140101	Datamart	TD_FECHA	IDFECHA	entero		
IDPLANTA	Identificador de la planta	entero		PK	FK	1	Datamart	TD_PLANTAS	IDPLANTA	entero		
IDESCENARIO	Identificador del escenario	entero		PK	FK	1	Datamart	TD_ESCENARIO	IDESCENARIO	entero		
IDUNIDAD	Identificador de la unidad	entero		PK	FK	4	Datamart	TD_UNDMETRICA	IDUNIDAD	entero		
IDEMPRESA	Identificador de la empresa	entero			FK	1	Datamart	TD_EMPRESA	IDEMPRESA	entero		
IDPOSTDESPACHO	Identificador del grupo	entero			FK	7	Datamart	TD_POSTDESPACHO	IDPOSTDESPACHO	entero		
IDFUENTE	Identificador de la fuente	entero			FK	1	Datamart	TD_FUENTE	IDFUENTE	entero		
IDRECURSO	Identificador del recurso	entero			FK	1	Datamart	TD_RECURSO	IDRECURSO	entero		
IDREGION	Identificador de la región	entero			FK	3	Datamart	TD_REGION	IDREGION	entero		
ENERGIA	Dato de energía	float	24			553492,5	sICENCE	ENERGIA_PLAN	ENP_ENERGIA ENP_PREDESPACHO	numeric(12,4)	Real Predespacho	

Nombre de la Tabla: TH_POTENCIA_SCADA
Tipo: Hechos
Descripción: Contiene la potencia cada 15 minutos por planta, obtenida del SCADA

Columna	Descripción	Tipo	Destino				Ejemplo	Fuente				
			Tamaño	Llave	FK			Sistema	Tabla	Columna	Tipo	Comentario
IDFECHA	Identificador de la fecha	entero		PK	FK	20140101	Datamart	TD_FECHA	IDFECHA	entero		
ID_HORA	Identificador de la hora	entero		PK	FK	0	Datamart	TD_HORAS	ID_HORA	entero		
IDPLANTA	Identificador de la planta	entero		PK	FK	1	Datamart	TD_PLANTAS	IDPLANTA	entero		
IDESCENARIO	Identificador del escenario	entero		PK	FK	1	Datamart	TD_ESCENARIO	IDESCENARIO	entero		
IDUNIDAD	Identificador de la unidad	entero		PK	FK	1	Datamart	TD_UNDMETRICA	IDUNIDAD	entero		
IDEMPRESA	Identificador de la empresa	entero			FK	1	Datamart	TD_EMPRESA	IDEMPRESA	entero		
IDPOSTDESPACHO	Identificador del grupo	entero			FK	7	Datamart	TD_POSTDESPACHO	IDPOSTDESPACHO	entero		
IDFUENTE	Identificador de la fuente	entero			FK	1	Datamart	TD_FUENTE	IDFUENTE	entero		
IDRECURSO	Identificador del recurso	entero			FK	1	Datamart	TD_RECURSO	IDRECURSO	entero		
IDREGION	Identificador de la región	entero			FK	3	Datamart	TD_REGION	IDREGION	entero		
POTENCIA	Dato de potencia	float	24			1278,25	sICENCE	POTENCIA_UNI	DATO	numeric(12,4)		

Nombre de la Tabla: TH_ENERGIA_SCADA

Tipo: Hechos

Descripción: Contiene la energía horaria por planta, obtenida del SCADA

Columna	Descripción	Tipo	Destino			Ejemplo	Fuente				
			Tamaño	Llave	FK		Sistema	Tabla	Columna	Tipo	Comentario
IDFECHA	Identificador de la fecha	entero		PK	FK	20140101	Datamart	TD_FECHA	IDFECHA	entero	
ID_HORA	Identificador de la hora	entero		PK	FK	0	Datamart	TD_HORAS	ID_HORA	entero	
IDPLANTA	Identificador de la planta	entero		PK	FK	1	Datamart	TD_PLANTAS	IDPLANTA	entero	
IDESCENARIO	Identificador del escenario	entero		PK	FK	1	Datamart	TD_ESCENARIO	IDESCENARIO	entero	
IDUNIDAD	Identificador de la unidad	entero		PK	FK	4	Datamart	TD_UNDMETRICA	IDUNIDAD	entero	
IDEMPRESA	Identificador de la empresa	entero			FK	1	Datamart	TD_EMPRESA	IDEMPRESA	entero	
IDPOSTDESPACHO	Identificador del grupo	entero			FK	7	Datamart	TD_POSTDESPACHO	IDPOSTDESPACHO	entero	
IDFUENTE	Identificador de la fuente	entero			FK	1	Datamart	TD_FUENTE	IDFUENTE	entero	
IDRECURSO	Identificador del recurso	entero			FK	1	Datamart	TD_RECURSO	IDRECURSO	entero	
IDREGION	Identificador de la región	entero			FK	3	Datamart	TD_REGION	IDREGION	entero	
ENERGIA	Dato de energía	float	24			24378,64	siCENCE	POTENCIA_UNI PRE_DATOS	DATO	numeric(12,4)	Se utiliza una sumatoria, pues los datos viene por unidad

Nombre de la Tabla: TH_NIVELES

Tipo: Hechos

Descripción: Contienen los niveles de los embalses por hora por planta

Columna	Descripción	Tipo	Destino			Ejemplo	Fuente				
			Tamaño	Llave	FK		Sistema	Tabla	Columna	Tipo	Comentario
IDFECHA	Identificador de la fecha	entero		PK	FK	20140101	Datamart	TD_FECHA	IDFECHA	entero	
ID_HORA	Identificador de la hora	entero		PK	FK	0	Datamart	TD_HORAS	ID_HORA	entero	
IDPLANTA	Identificador de la planta	entero		PK	FK	1	Datamart	TD_PLANTAS	IDPLANTA	entero	
IDUNIDAD	Identificador de la unidad	entero		PK	FK	7	Datamart	TD_UNDMETRICA	IDUNIDAD	entero	
NIVEL	Nivel del embalse	numérico				540,65	siCENCE	NIVELES	DATO	numeric(12,2)	

Nombre de la Tabla: TH_DEMANDA

Tipo: Hechos

Descripción: Contiene la demanda de energía horaria para todo el sistema eléctrico.

Columna	Descripción	Tipo	Destino			Ejemplo	Fuente				
			Tamaño	Llave	FK		Sistema	Tabla	Columna	Tipo	Comentario
IDFECHA	Identificador de la fecha	entero		PK	FK	20140101	Datamart	TD_FECHA	IDFECHA	entero	
ID_HORA	Identificador de la hora	entero		PK	FK	0	Datamart	TD_HORAS	ID_HORA	entero	
IDESCENARIO	Identificador del escenario	entero		PK	FK	1	Datamart	TD_PLANTAS	IDPLANTA	entero	
IDUNIDAD	Identificador de la unidad	entero		PK	FK	4	Datamart	TD_UNDMETRICA	IDUNIDAD	entero	
NIVEL	Nivel del embalse	numérico				897583	siCENCE	POTENCIA_UNI	DATO	numeric(12,4)	Se utiliza una sumatoria.

Apéndice 3. Detalle de las transformaciones.

Nombre de la Tabla:	TD_EMPRESA
Tipo:	Dimensión
Descripción:	Empresas con participación en la generación de electricidad

Modelo de datos

Columna	Tipo	Descripción
IDEMPRESA	entero	Identificador de la empresa
NOMEMPRESA	texto	Nombre de la empresa
ORDEN	entero	Orden para mostrar las empresas
COLOR	texto	Color para referenciar esta empresa

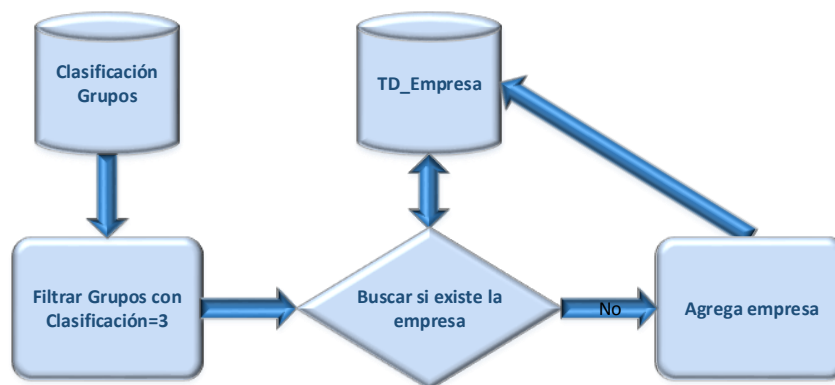
Mapeo Fuente-Destino

Sistema	Fuente			Destino
	Tabla	Columna		Columna
ETL				IDEMPRESA
siCENCE	GRUPOS	GRU_NOMBRE	→	NOMEMPRESA
siCENCE	GRUPOS	GRU_ORDEN	→	ORDEN
siCENCE	GRUPOS	GRU_COLOR	→	COLOR

Descripción

En el sistema fuente hay una tabla de CLASIFICACION que contiene el conjunto de clasificaciones, luego está la tabla de GRUPOS, que contiene el conjunto de grupos por clasificación, la dimensión de empresa se contruye a partir de los grupos que conforman la clasificación por empresa, cuyo código de clasificación es 3.

Diagrama de flujo general



Nombre de la Tabla:	TD_POSTDESPACHO
Tipo:	Dimensión
Descripción:	Grupos de plantas con el orden de cómo deben ser despachadas para cumplir con la demanda del sistema.

Modelo de datos

Columna	Tipo	Descripción
IDPOSTDESPACHO	entero	Identificador del grupo
NOMPOSTDESPACHO	texto	Nombre del grupo
ORDEN	entero	Orden para mostrar los grupos
COLOR	texto	Color para referenciar este grupo

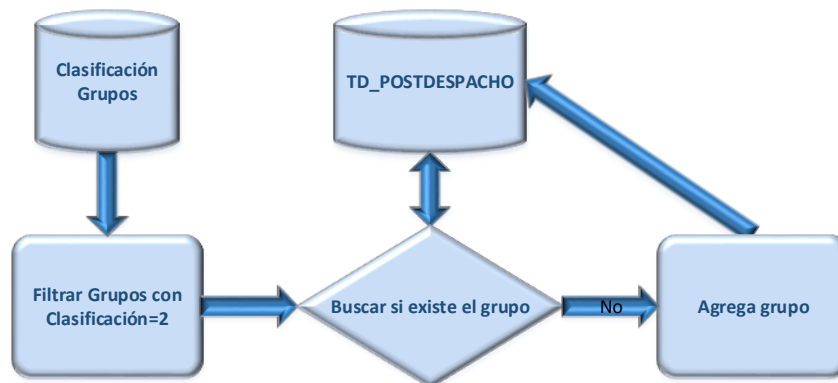
Mapeo Fuente-Destino

Sistema	Fuente			Destino
	Tabla	Columna		Columna
ETL				IDPOSTDESPACHO
siCENCE	GRUPOS	GRU_NOMBRE	→	NOMPOSTDESPACHO
siCENCE	GRUPOS	GRU_ORDEN	→	ORDEN
siCENCE	GRUPOS	GRU_COLOR	→	COLOR

Descripción

En el sistema fuente hay una tabla de CLASIFICACION que contiene el conjunto de clasificaciones, luego está la tabla de GRUPOS, que contiene el conjunto de grupos por clasificación, la dimensión de postdespacho se contruye a partir de los grupos que conforman la clasificación por postdespacho cuyo código de clasificación es 2.

Diagrama de flujo general



Nombre de la Tabla:	TD_RECURSO
Tipo:	Dimensión
Descripción:	Grupos de plantas de acuerdo al tipo de recurso, que pueden ser renovables, térmico (las que utilizan combustibles fósiles) o intercambio, es decir, que proviene de otro país, pero se desconoce el recurso que utilizan.

Modelo de datos

Columna	Tipo	Descripción
IDRECURSO	entero	Identificador del recurso
NOMRECURSO	texto	Nombre del recurso
ORDEN	entero	Orden para mostrar los grupos
COLOR	texto	Color para referenciar este grupo

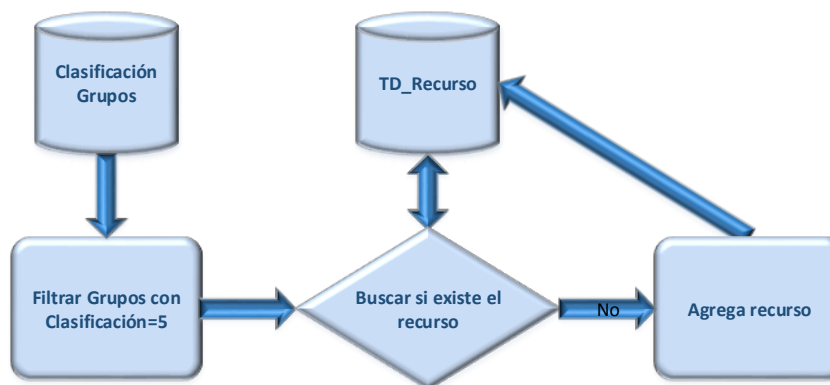
Mapeo Fuente-Destino

Sistema	Fuente			Destino
	Tabla	Columna		Columna
ETL				IDRECURSO
siCENCE	GRUPOS	GRU_NOMBRE	→	NOMRECURSO
siCENCE	GRUPOS	GRU_ORDEN	→	ORDEN
siCENCE	GRUPOS	GRU_COLOR	→	COLOR

Descripción

En el sistema fuente hay una tabla de CLASIFICACION que contiene el conjunto de clasificaciones, luego está la tabla de GRUPOS, que contiene el conjunto de grupos por clasificación, la dimensión de recurso se contruye a partir de los grupos que conforman la clasificación por recurso cuyo código de clasificación es 5.

Diagrama de flujo general



Nombre de la Tabla:	TD_FUENTE
Tipo:	Dimensión
Descripción:	Fuentes de energía utilizada por la planta dada para generar electricidad, como por ejemplo, hidroeléctrica, geotérmica, eólica, etc.

Modelo de datos

Columna	Tipo	Descripción
IDFUENTE	entero	Identificador de la fuente
NOMFUENTE	texto	Nombre de la fuente
ORDEN	entero	Orden para mostrar las fuentes
COLOR	texto	Color para referenciar esta fuente

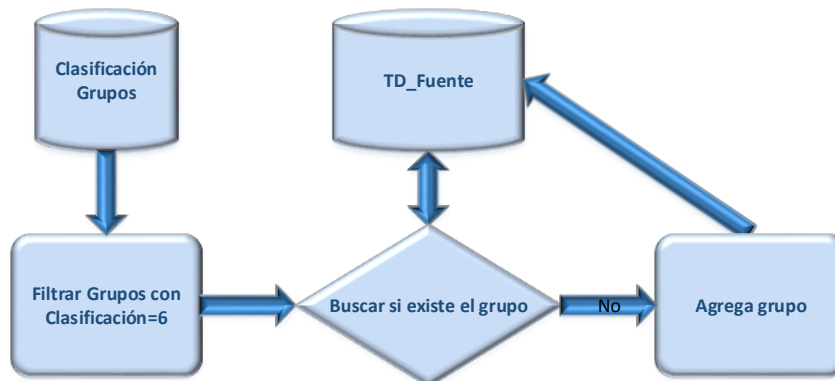
Mapeo Fuente-Destino

Sistema	Fuente			Destino
	Tabla	Columna		Columna
ETL				IDFUENTE
siCENCE	GRUPOS	GRU_NOMBRE	→	NOMFUENTE
siCENCE	GRUPOS	GRU_ORDEN	→	ORDEN
siCENCE	GRUPOS	GRU_COLOR	→	COLOR

Descripción

En el sistema fuente hay una tabla de CLASIFICACION que contiene el conjunto de clasificaciones, luego está la tabla de GRUPOS, que contiene el conjunto de grupos por clasificación, la dimensión de fuentes de energía se contruye a partir de los grupos que conforman la clasficiación por fuente cuyo código de clasificación es 6.

Diagrama de flujo general



Nombre de la Tabla:	TD_ESCENARIO
Tipo:	Dimensión
Descripción:	Escenarios posibles donde se contextualiza el dato, los escenarios posibles son: Real, Programado o diferencia.

Modelo de datos

Columna	Tipo	Descripción
IDESCENARIO	entero	Identificador del escenario
NOMESCENARIO	texto	Nombre del escenario

Mapeo Fuente-Destino

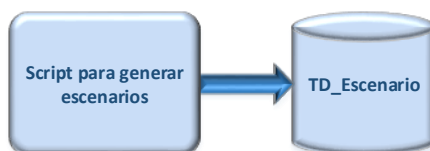
Sistema	Fuente Tabla	Columna	Destino Columna
ETL			IDESCENARIO
ETL			NOMESCENARIO

Descripción

El escenario se considera una dimensión estática y esta contiene los posibles escenarios donde se ubican los datos, a saber, si son datos reales, datos programados o una diferencia entre lo real y lo programado.

Para construirla se utilizará un script.

Diagrama de flujo general



Nombre de la Tabla:	TD_REGION
Tipo:	Dimensión
Descripción:	Regiones que agrupan un conjunto de plantas, estas regiones no necesariamente son geográficas, más bien corresponden a la topología de la red del sistema eléctrico nacional.

Modelo de datos

Columna	Tipo	Descripción
IDREGION	entero	Identificador de la región
NOMREGION	texto	Nombre de la región

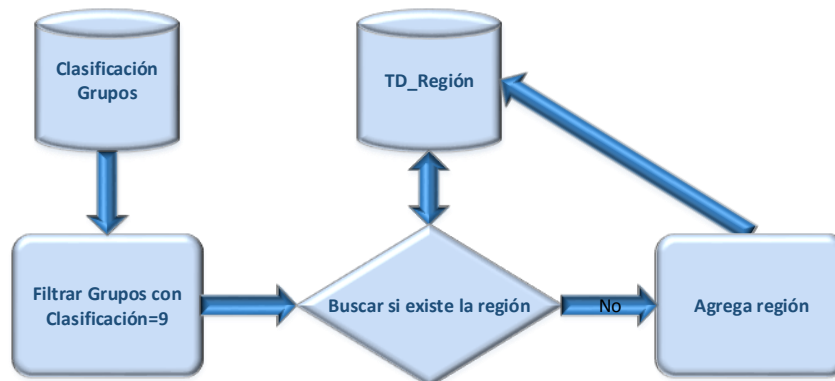
Mapeo Fuente-Destino

Sistema	Fuente			Destino
	Tabla	Columna		Columna
ETL				IDREGION
siCENCE	GRUPOS	GRU_NOMBRE	→	NOMREGION

Descripción

En el sistema fuente hay una tabla de CLASIFICACION que contiene el conjunto de clasificaciones, luego está la tabla de GRUPOS, que contiene el conjunto de grupos por clasificación, la dimensión de regiones se contruye a partir de los grupos que conforman la clasificación por región cuyo código de clasificación es 9.

Diagrama de flujo general



Nombre de la Tabla:	TD_PLANTAS
Tipo:	Dimensión
Descripción:	Conjunto de plantas que aportan generación al país.

Modelo de datos

Columna	Tipo	Descripción
IDPLANTA	entero	Identificador de la planta
NOMPLANTA	texto	Nombre de la planta
NOMCORTO	entero	Nombre corto o abreviatura de la planta
SK_PLANTA	entero	Identificador de la planta en el sistema origen

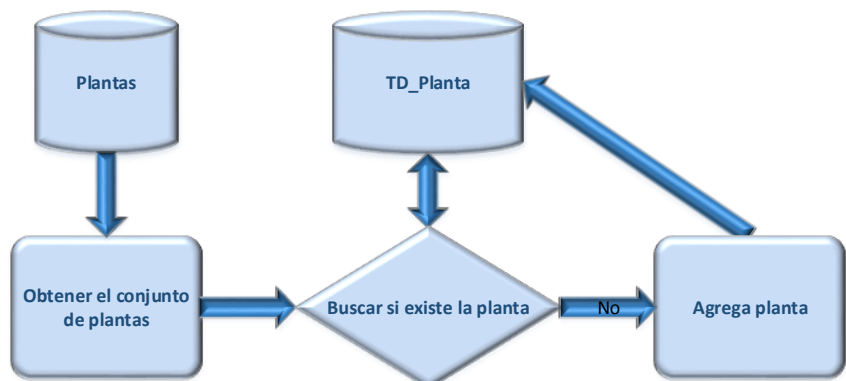
Mapeo Fuente-Destino

Fuente		Destino		
Sistema	Tabla	Columna	Columna	
ETL			IDREGION	
siCENCE	PLANTAS	PLA_NOMBRE	→	NOMREGION
siCENCE	PLANTAS	PLA_NCORTO	→	NOMCORTO
siCENCE	PLANTAS	PLA_CODIGO	→	SK_PLANTA

Descripción

En el sistema fuente hay una tabla de PLANTAS que contiene el conjunto de plantas que existen en el sistema. La dimensión de plantas se contruye a partir de esta tabla.

Diagrama de flujo general



Nombre de la Tabla:	TD_UNDMETRICA
Tipo:	Dimensión
Descripción:	Unidades de medidas que se utilizan para los datos registrados, por ejemplo kWh, MWh, kW, msnm, etc.

Modelo de datos

Columna	Tipo	Descripción
IDUNDMETRICA	entero	Identificador de la unidad de medida
NOMUNDMETRICA	texto	Nombre de la unidad
DESCRIPCION	texto	Descripción de la unidad de medida

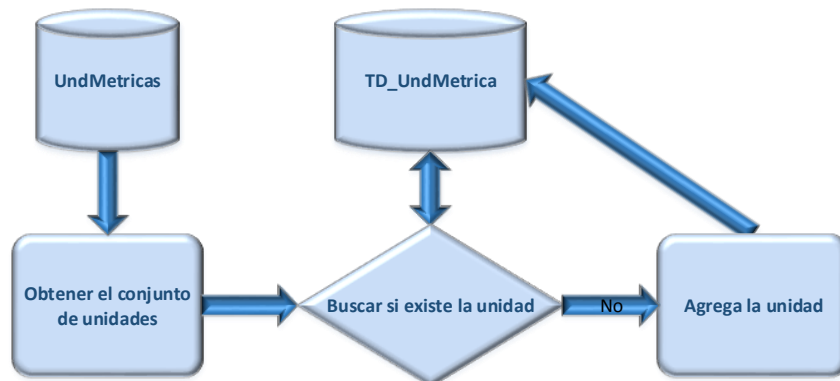
Mapeo Fuente-Destino

Sistema	Fuente			Destino
	Tabla	Columna		Columna
ETL				IDUNDMETRICA
siCENCE	UNDMETRICA	NOMBRE_UNID	→	NOMUNDMETRIC
siCENCE	UNDMETRICA	DESCRIPCION	→	DESCRIPCION

Descripción

En el sistema fuente hay una tabla de UNDMETRICAS que contiene el conjunto de unidades que se utilizan para medir los datos. La dimensión de unidades se construye a partir de esta tabla.

Diagrama de flujo general



Nombre de la Tabla:	TD_FECHA
Tipo:	Dimensión
Descripción:	Listado de fechas

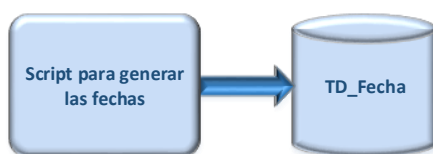
Modelo de datos

Columna	Tipo	Descripción
IDFECHA	entero	Identificador de la fecha
FECHA	fecha	Fecha
ANIO	entero	Año
TRIMESTRE	entero	Trimestre
MES	entero	Mes
SEMANA	entero	Número de la semana
DIA	entero	Día
DIASEMANA	entero	Día de la semana
MTRIMESTRE	texto	Nombre del trimestre
NMES	texto	Nombre del mes
NCORTOMES	texto	Nombre abreviado del mes
NSEMANA	texto	Nombre de la semana
NDIA	texto	Nombre del día
NDIASEMANA	texto	Nombre del día de la semana
NCORTODIA	texto	Nombre abreviado del día
NFERIADO	texto	Corresponde a un feriado?
FESTIVIDAD	texto	Descripción de la festividad
FERIADO	entero	Valor para día feriado
VACACIONES	entero	Valor para día de vacaciones escolares
NVACACIONES	texto	Es un día de vacaciones escolares?
TIEMPOESCOL	texto	Descripción del tiempo escolar

Descripción

El escenario se considera una dimensión estática y esta contiene las fechas con sus respectivos atributos.

Diagrama de flujo general



Nombre de la Tabla:	TD_HORAS
Tipo:	Dimensión
Descripción:	Listado de las horas del día.

Modelo de datos

Columna	Tipo	Descripción
IDHORA	entero	Identificador de la hora
TIEMPO	hora	Hora
NTIEMPO	texto	Texto que representa la hora
HORA	entero	hora del día
MINUTO	entero	Minutos
TIPO	entero	Tipo de hora
NTIPO	texto	Descripción del tipo
PUNTA	entero	Valor para hora punta
NPUNTA	texto	Descripción de hora punta

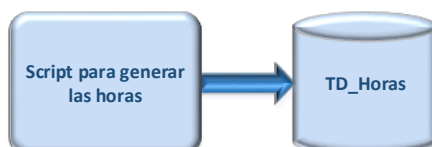
Mapeo Fuente-Destino

Sistema	Fuente Tabla	Columna	Destino Columna
ETL			IDHORA
ETL			TIEMPO
ETL			NTIEMPO
ETL			HORA
ETL			MINUTO
ETL			TIPO
ETL			NTIPO
ETL			PUNTA
ETL			NPUNTA

Descripción

El escenario se considera una dimensión estática y esta contiene las fechas con sus respectivos atributos.

Diagrama de flujo general



Nombre de la Tabla:	TH_ENERGIA
Tipo:	Hechos
Descripción:	Contiene la energía diaria por planta

Modelo de datos

Columna	Tipo	Descripción
IDFECHA	entero	Identificador de la fecha
IDPLANTA	entero	Identificador de la planta
IDESCENARIO	entero	Identificador del escenario
IDUNIDAD	entero	Identificador de la unidad
IDEMPRESA	entero	Identificador de la empresa
IDPOSTDESPACHO	entero	Identificador del grupo
IDFUENTE	entero	Identificador de la fuente
IDRECURSO	entero	Identificador del recurso
IDREGION	entero	Identificador de la región
ENERGIA	flotante	Dato de energía

Mapeo Fuente-Destino

Sistema	Fuente			Destino
	Tabla	Columna		Columna
Datamart	TD_FECHA	IDFECHA	→	IDFECHA
Datamart	TD_PLANTAS	IDPLANTA	→	IDPLANTA
Datamart	TD_ESCENARIO	IDESCENARIO	→	IDESCENARIO
Datamart	TD_UNDMETRICA	IDUNIDAD	→	IDUNIDAD
Datamart	TD_EMPRESA	IDEMPRESA	→	IDEMPRESA
Datamart	TD_POSTDESPACHO	IDPOSTDESPACHO	→	IDPOSTDESPACHO
Datamart	TD_FUENTE	IDFUENTE	→	IDFUENTE
Datamart	TD_RECURSO	IDRECURSO	→	IDRECURSO
Datamart	TD_REGION	IDREGION	→	IDREGION
siCENCE	ENERGIA_PLAN	ENP_ENERGIA ENP_PREDESPACHO	→	ENERGIA

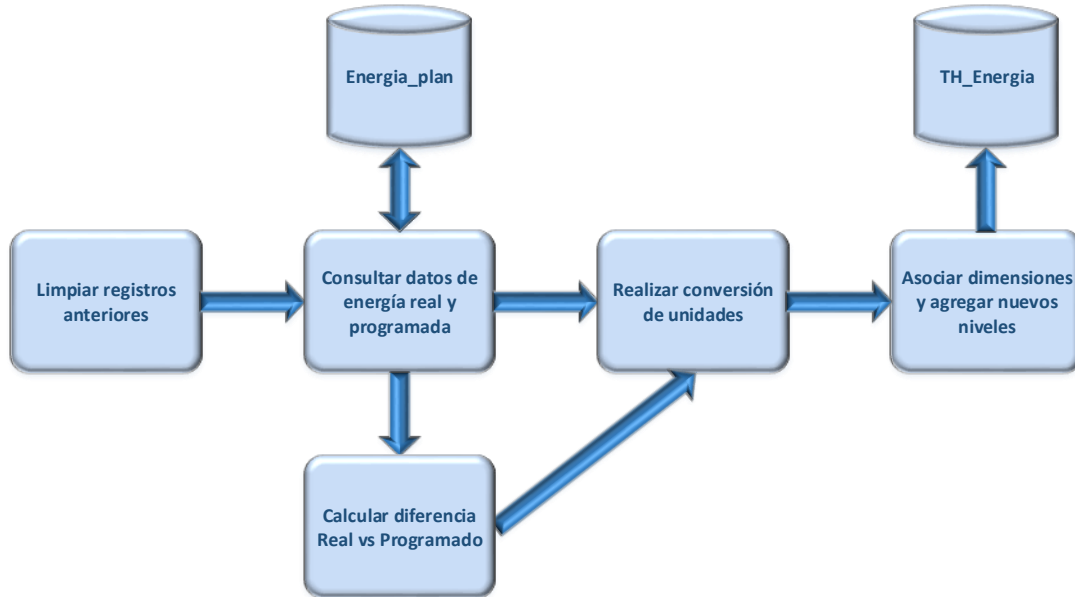
Descripción

En el sistema fuente hay una tabla de ENERGIA_PLAN que contiene la generación diaria real y programada para cada planta, estos datos están en kWh. Para cargar la tabla de hechos, se debe consultar la energía, ya sea real o programada y realizar la conversión a las unidades correspondiente.

Por otro lado, se debe calcular la diferencia e igualmente realizar las conversiones.

Finalmente se debe asociar a las dimensiones creadas para escribir el dato a la tabla de hechos. Primero se debe limpiar los hechos para evitar duplicados.

Diagrama de flujo general



Detalles	TH_POTENCIA_SCADA
Tipo:	Hechos
Descripción:	Contiene la potencia cada 15 minutos por planta, obtenida del

Modelo de datos

Columna	Tipo	Descripción
IDFECHA	entero	Identificador de la fecha
ID_HORA	entero	Identificador de la hora
IDPLANTA	entero	Identificador de la planta
IDESCENARIO	entero	Identificador del escenario
IDUNIDAD	entero	Identificador de la unidad
IDEMPRESA	entero	Identificador de la empresa
IDPOSTDESPACHO	entero	Identificador del grupo
IDFUENTE	entero	Identificador de la fuente
IDRECURSO	entero	Identificador del recurso
IDREGION	entero	Identificador de la región
POTENCIA	flotante	Dato de potencia

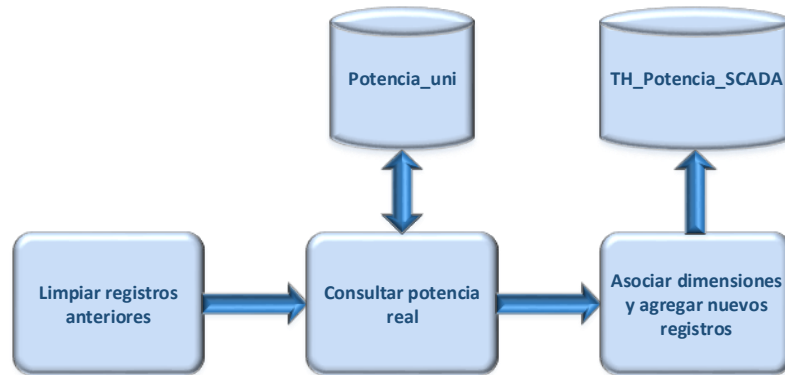
Mapeo Fuente-Destino

Sistema	Fuente			Destino
	Tabla	Columna		Columna
Datamart	TD_FECHA	IDFECHA	→	IDFECHA
Datamart	TD_HORAS	ID_HORA	→	ID_HORA
Datamart	TD_PLANTAS	IDPLANTA	→	IDPLANTA
Datamart	TD_ESCENARIO	IDESCENARIO	→	IDESCENARIO
Datamart	TD_UNDMETRICA	IDUNIDAD	→	IDUNIDAD
Datamart	TD_EMPRESA	IDEMPRESA	→	IDEMPRESA
Datamart	TD_POSTDESPACHO	IDPOSTDESPACHO	→	IDPOSTDESPACHO
Datamart	TD_FUENTE	IDFUENTE	→	IDFUENTE
Datamart	TD_RECURSO	IDRECURSO	→	IDRECURSO
Datamart	TD_REGION	IDREGION	→	IDREGION
siCENCE	POTENCIA_UNI	DATO	→	POTENCIA

Descripción

En el sistema fuente hay una tabla de POTENCIA_UNI, que contiene la potencia para cada unidad generadora del sistema, estos datos están cada 15 minutos. Primero se debe limpiar los hechos para evitar duplicados. La tabla de hechos de potencia se llenan con los datos de esta tabla, así se consultan los datos y luego se asocia a las dimensiones respectivas, para finalmente guardar el dato.

Diagrama de flujo general



Nombre de la Tabla:	TH_ENERGIA_SCADA
Tipo:	Hechos
Descripción:	Contiene la energía horaria por planta, obtenida del SCADA

Modelo de datos

Columna	Tipo	Descripción
IDFECHA	entero	Identificador de la fecha
ID_HORA	entero	Identificador de la hora
IDPLANTA	entero	Identificador de la planta
IDESCENARIO	entero	Identificador del escenario
IDUNIDAD	entero	Identificador de la unidad
IDEMPRESA	entero	Identificador de la empresa
IDPOSTDESPACHO	entero	Identificador del grupo
IDFUENTE	entero	Identificador de la fuente
IDRECURSO	entero	Identificador del recurso
IDREGION	entero	Identificador de la región
ENERGIA	flotante	Dato de energía

Mapeo Fuente-Destino

Sistema	Fuente			Destino
	Tabla	Columna		Columna
Datamart	TD_FECHA	IDFECHA	→	IDFECHA
Datamart	TD_HORAS	ID_HORA	→	ID_HORA
Datamart	TD_PLANTAS	IDPLANTA	→	IDPLANTA
Datamart	TD_ESCENARIO	IDESCENARIO	→	IDESCENARIO
Datamart	TD_UNDMETRICA	IDUNIDAD	→	IDUNIDAD
Datamart	TD_EMPRESA	IDEMPRESA	→	IDEMPRESA
Datamart	TD_POSTDESPACHO	IDPOSTDESPACHO	→	IDPOSTDESPACHO
Datamart	TD_FUENTE	IDFUENTE	→	IDFUENTE
Datamart	TD_RECURSO	IDRECURSO	→	IDRECURSO
Datamart	TD_REGION	IDREGION	→	IDREGION
siCENCE	POTENCIA_UNI PRE_DATOS	DATO	→	ENERGIA

Descripción

Para obtener la energía horaria no se cuenta con mediciones de energía, con o el caso de la energía diaria, por lo que se debe acudir a la potencia para calcular una aproximación, conciendo que la energía se calcula como la integración de la potencia en un intervalo de tiempo.

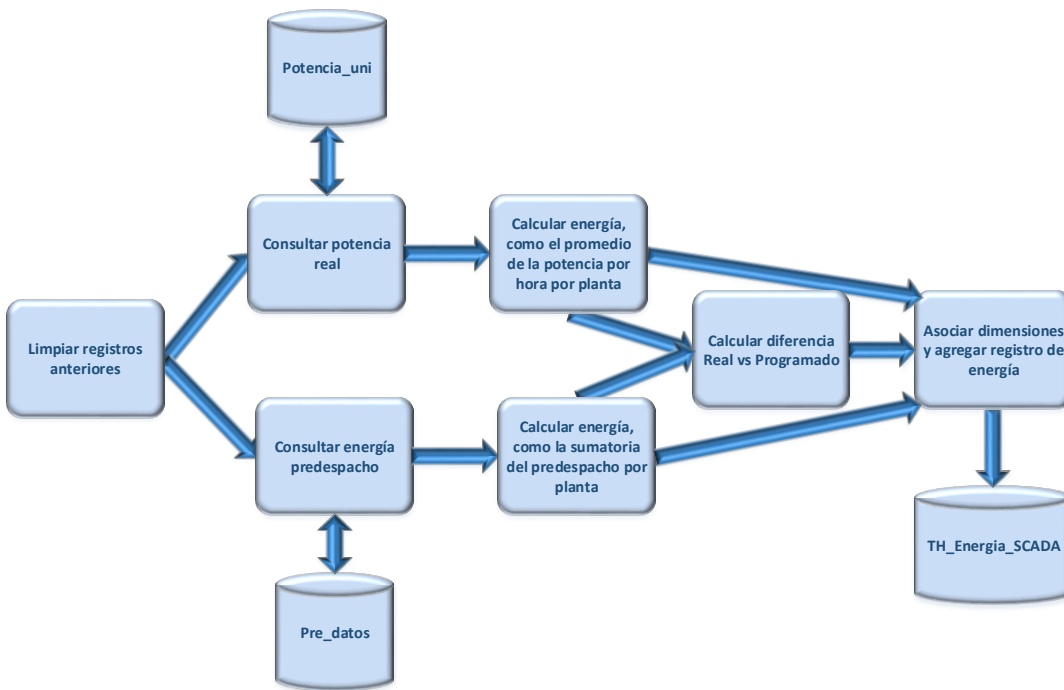
Así se utiliza la tabla de POTENCIA_UNI para calcular el promedio por hora, con lo que aproximamos el valor de la energía horaria.

Luego en la tabla PRE_DATOS se encuentra la energía predespachada por cada elemento, así se realiza la sumatoria de los elementos por planta, para obtener el predespacho horario de cada planta.

Con estos datos se procede luego a calcular la diferencia.

Finalmente se asocian las dimensiones correspondientes para guardar el dato en la tabla de hecho. Se debe previamente eliminar los hechos para evitar duplicados.

Diagrama de flujo general



Nombre de la Tabla:	TH_NIVELES
Tipo:	Hechos
Descripción:	Contienen los niveles de los embalses por hora por planta

Modelo de datos

Columna	Tipo	Descripción
IDFECHA	entero	Identificador de la fecha
ID_HORA	entero	Identificador de la hora
IDPLANTA	entero	Identificador de la planta
IDUNIDAD	entero	Identificador de la unidad
NIVEL	numérico	Nivel del embalse

Mapeo Fuente-Destino

Sistema	Fuente			Destino
	Tabla	Columna		Columna
Datamart	TD_FECHA	IDFECHA	→	IDFECHA
Datamart	TD_HORAS	ID_HORA	→	ID_HORA
Datamart	TD_PLANTAS	IDPLANTA	→	IDPLANTA
Datamart	TD_UNDMETRICA	IDUNIDAD	→	IDUNIDAD
siCENCE	NIVELES	DATO	→	NIVEL

Descripción

En el sistema fuente hay una tabla de NIVELES que contiene los datos de los niveles de los embalses. La tabla de hechos de niveles, se carga a partir de esta tabla, asociando las dimensiones respectivas.

Se deben borrar los niveles anteriores para evitar duplicados.

Diagrama de flujo general



Nombre de la Tabla:	TH_DEMANDA
Tipo:	Hechos
Descripción:	Contiene la demanda de energía horaria para todo el sistema eléctrico.

Modelo de datos

Columna	Tipo	Descripción
IDFECHA	entero	Identificador de la fecha
ID_HORA	entero	Identificador de la hora
IDESCENARIO	entero	Identificador del escenario
IDUNIDAD	entero	Identificador de la unidad
ENERGIA	numérico	Demanda de energía horaria

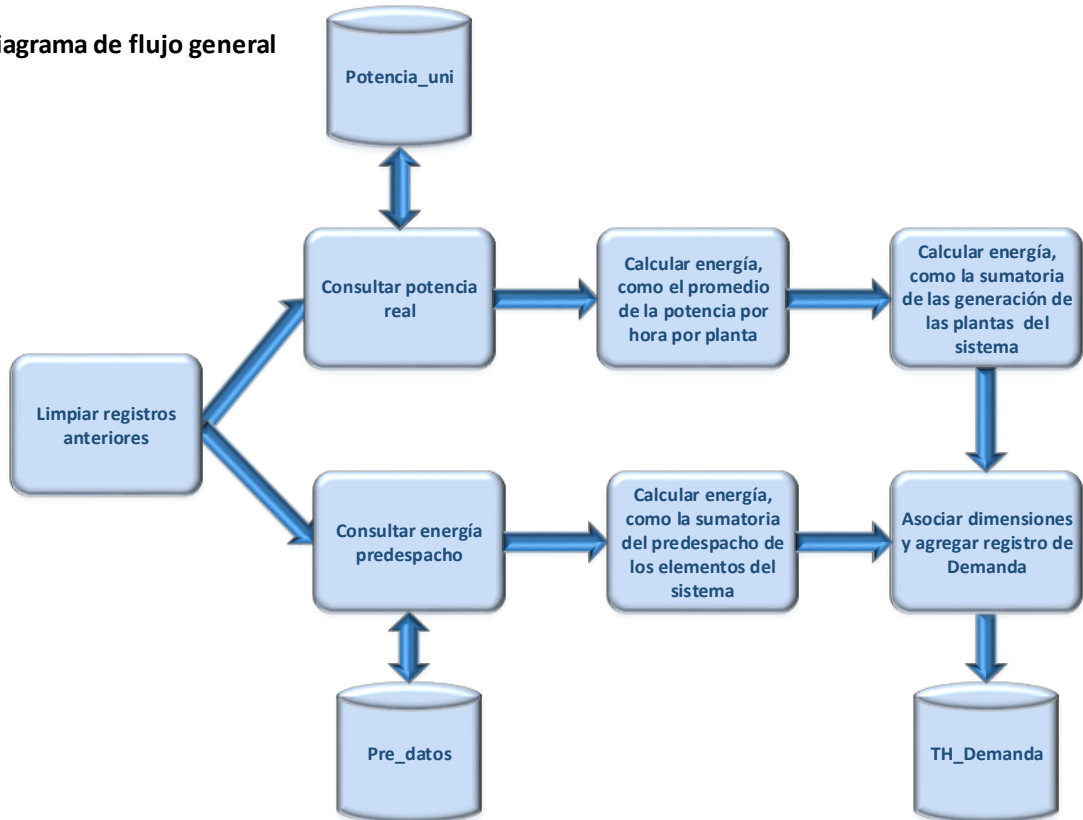
Mapeo Fuente-Destino

Fuente			Destino
Sistema	Tabla	Columna	Columna
Datamart	TD_FECHA	IDFECHA	→ IDFECHA
Datamart	TD_HORAS	ID_HORA	→ ID_HORA
Datamart	TD_PLANTAS	IDPLANTA	→ IDESCENARIO
Datamart	TD_UNDMETRICA	IDUNIDAD	→ IDUNIDAD
siCENCE	POTENCIA_UNI	DATO	→ ENERGIA

Descripción

La demanda horaria del sistema, se obtiene a partir del dato de potencia en la tabla POTENCIA_UNI, se calcula la energía horaria por planta y luego se procede a realizar la sumatoria de todas las plantas para cada hora, con lo que se obtiene la demanda real. Luego de manera recíproca, se realiza la sumatoria de todos los elementos predespachados, para obtener la demanda total predespachada. Finalmente se asocian las dimensiones correspondientes para guardar el dato. Previa limpieza de los datos anteriores, para evitar duplicados.

Diagrama de flujo general



Apéndice 4. Consulta Microsoft Time Series

```
USE CENCE_SA
TRUNCATE TABLE ARIMA2 -- Elimina el contenido de la tabla

DECLARE @dia INT
DECLARE @s1 VARCHAR(10)
DECLARE @s2 VARCHAR(10)
DECLARE @DMXQ VARCHAR(MAX)
DECLARE @statement VARCHAR(MAX)

SET @dia = 1

WHILE @DIA <= 365
BEGIN
    SET @s1 = CONCAT(CONVERT(VARCHAR(10),@dia),',',CONVERT(VARCHAR(10),@dia)) -- Día a
    proyectar
    SET @s2 = CONVERT(VARCHAR(10),(@dia-2)) -- Se incluyen datos
    históricos

    SET @DMXQ = 'SELECT FLATTENED Predict([ArimaH].[ENERGIA],'+@s1+
    ', EXTEND_MODEL_CASES) as predespacho
FROM
    [ArimaH]
PREDICTION JOIN
    OPENQUERY([CENCE DM],
    ''SELECT TOP '+@s2+
    ' [IDTiempo],
    [ENERGIA],
    [ANIO],
    [TRIMESTRE],
    [MES],
    [SEMANA],
    [DIASEMANA],
    [DIA],
    [FERIADO],
    [VACACIONES]
FROM
    [dbo].[vDemARIMAHPruebas]
    ''') AS t
[ArimaH].[ID Tiempo] = t.[IDTiempo] AND
[ArimaH].[ENERGIA] = t.[ENERGIA] AND
[ArimaH].[ANIO] = t.[ANIO] AND
[ArimaH].[TRIMESTRE] = t.[TRIMESTRE] AND
[ArimaH].[MES] = t.[MES] AND
[ArimaH].[SEMANA] = t.[SEMANA] AND
[ArimaH].[DIASEMANA] = t.[DIASEMANA] AND
[ArimaH].[DIA] = t.[DIA] AND
[ArimaH].[FERIADO] = t.[FERIADO] AND
[ArimaH].[VACACIONES] = t.[VACACIONES]'

    SET @statement = 'INSERT INTO arima2 SELECT * FROM OPENQUERY([DATA_MINING], '
    SET @Statement = @Statement + '''' + @DMXQ + ''''
    EXEC(@Statement)

    SET @dia = @dia + 1;
END;
```

Apéndice 5. Implementación de promedios móviles

```
USE [CENCE_BD]
GO
/***** Object: StoredProcedure [dbo].[PROYECCION_PM]    Script Date: 27/10/2014
19:02:23 *****/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
-- =====
-- Author:          Leonardo Arias, José Hidalgo
-- Create date:     27 octubre, 2014
-- Description:     Calcula la proyección de demanda de energía para la fecha
ingresada en el parámetros MiFecha,
-- utilizando el algoritmo de promedios móviles.
-- =====
ALTER PROCEDURE [dbo].[PROYECCION_PM] (@MiFecha datetime)
AS
BEGIN
    -- Fecha que se utilizará de base para buscar los días anteriores
    DECLARE @MiFecha2 SMALLDATETIME
    -- Día de la semana de la fecha a proyectar
    DECLARE @MiDia TINYINT
    -- Variable para indicar si la fecha corresponde a un feriado
    DECLARE @EsFeriado TINYINT
    -- Variable para obtener la festividad
    DECLARE @MiFestividad VARCHAR(60)
    -- Cantidad de días a utilizar para promedios móviles
    DECLARE @PM INT
    -- Variable para realizar el ciclo por horas
    DECLARE @MiHora TINYINT
    -- Variable el conteo de registros
    DECLARE @Registros INT
    -- Para el manejo de @@ERROR
    DECLARE @MiError INT

    -- Se obtiene si la fecha a proyectar se trata de un feriado
    SELECT @EsFeriado=FERIADO FROM [CENCE_DM].[dbo].TD_FECHA
    WHERE FECHA = @MiFecha

    -- Algunos feriados se tratan como si fuera domingo y otros como sábado
    IF (@EsFeriado = 1 AND DATEPART(weekday,@MiFecha)<6) BEGIN
        SELECT @MiFestividad=FESTIVIDAD FROM [CENCE_DM].[dbo].TD_FECHA
        WHERE FECHA = @MiFecha
        IF @MiFestividad IN ('Día de la Madre y Asunción de la Virgen',
            'Navidad',
            'Año Nuevo',
            'Jueves Santo',
            'Viernes Santo')
            SET @MiFecha2=DATEADD(DD,7-DATEPART(weekday,@MiFecha),@MiFecha)
        ELSE
            SET @MiFecha2=DATEADD(DD,6-DATEPART(weekday,@MiFecha),@MiFecha)
    END
    ELSE BEGIN
        SET @MiFecha2 = @MiFecha
    END
END
```

```
-- **** Domingo en caso de semana santa, 15 de agosto y 25 dic, los demás son modelados como sábados
```

```
-- Obtiene el día de la semana para la fecha a proyectar
```

```
SET @MiDia = DATEPART(weekday,@MiFecha2)
```

```
-- Cuenta los registros para la fecha a proyectar
```

```
SELECT @Registros=COUNT(FECHA) FROM [CENCE_BD].[dbo].[DEMANDA_PM]  
WHERE FECHA BETWEEN @MiFecha AND DATEADD(HH,23,@MiFecha)
```

```
-- Si existen registro para la fecha, se procede a borrarlos
```

```
IF @Registros>0 BEGIN
```

```
    DELETE FROM [CENCE_BD].[dbo].[DEMANDA_PM]
```

```
    WHERE FECHA BETWEEN @MiFecha AND DATEADD(HH,23,@MiFecha)
```

```
END
```

```
-- Para el ciclo de cálculo de la proyección, inicia la hora en 0
```

```
SET @MiHora = 0
```

```
WHILE @MiHora < 24 BEGIN
```

```
    -- Obtiene la cantidad de días a utilizar para el promedio móvil
```

```
    SELECT @PM=[DIAS_PROM] FROM [CENCE_BD].[dbo].[PROMOVILES]  
    WHERE DIA=@MiDia AND HORA=@MiHora
```

```
    -- Realiza el insert de la proyección a la tabla DEMANDA_PM
```

```
    INSERT INTO [CENCE_BD].[dbo].[DEMANDA_PM] ([FECHA],[DEMANDA])  
    (SELECT DATEADD(HH,@MiHora,@MiFecha), AVG(ENERGIA) FROM (  
        SELECT TOP(@PM) D.[IDFECHA], [ENERGIA]  
        FROM [CENCE_DM].[dbo].[TH_DEMANDA] D INNER JOIN  
        [CENCE_DM].[dbo].[TD_FECHA] F ON D.IDFECHA =
```

```
F.IDFECHA
```

```
        WHERE IDESCENARIO=1
```

```
        AND F.FECHA < @MiFecha
```

```
        AND (F.FERIADO = 0 OR @MiDia>5)
```

```
        AND D.IDUNDMETRICA = 5
```

```
        AND IDHORA = @MiHora*100
```

```
        AND DATEPART(weekday,F.FECHA) = @MiDia
```

```
        ORDER BY 1 DESC
```

```
    ) AS T)
```

```
    -- Si sucedió algún error, se realiza el manejo de errores
```

```
    IF @MiERROR != 0 GOTO HANDLE_ERROR
```

```
    -- Siguiete hora
```

```
    SET @MiHora = @MiHora + 1
```

```
END -- WHILE
```

```
RETURN
```

```
END
```

```
HANDLE_ERROR:
```

```
    -- Revierte cualquier transacción pendiente y termina retornando el código del error.
```

```
    ROLLBACK TRAN
```

```
    RETURN @MiERROR
```


Apéndice 6. Implementación de redes neuronales

```
CREATE VIEW [dbo].[vDemanda]
-- Vista para trabajar con el "hecho" de demanda de energía asociado a las
"dimensiones" de fecha y hora.

AS
SELECT DATEADD(hh, dbo.TD_HORAS.HORA, dbo.TD_FECHA.FECHA) AS IDTiempo,
dbo.TD_FECHA.FECHA, ROUND(dbo.TH_DEMANDA.ENERGIA, 0) AS ENERGIA,
    dbo.TD_HORAS.TIPO, dbo.TD_HORAS.PUNTA, dbo.TD_FECHA.ANIO,
    dbo.TD_FECHA.TRIMESTRE, dbo.TD_FECHA.MES, dbo.TD_FECHA.SEMANA,
    dbo.TD_FECHA.DIASEMANA, dbo.TD_FECHA.DIA, dbo.TD_FECHA.FERIADO,
    dbo.TD_FECHA.VACACIONES, dbo.TD_HORAS.HORA
FROM dbo.TD_FECHA INNER JOIN
dbo.TH_DEMANDA ON dbo.TD_FECHA.IDFECHA = dbo.TH_DEMANDA.IDFECHA INNER JOIN
dbo.TD_HORAS ON dbo.TH_DEMANDA.IDHORA = dbo.TD_HORAS.IDHORA
WHERE (dbo.TH_DEMANDA.IDESCENARIO = 1) AND (dbo.TH_DEMANDA.IDUNDMETRICA = 5)

CREATE FUNCTION [dbo].[EntradaRN](@MiFecha SMALLDATETIME)
-- Función que recibe la fecha de un día y construye las entradas necesarias para la
red neuronal.

RETURNS TABLE
AS
RETURN
(
SELECT vDemanda_1.HORA, vDemanda_1.TIPO, vDemanda_1.PUNTA, [dbo].[TD_FECHA].ANIO,
[dbo].[TD_FECHA].TRIMESTRE, [dbo].[TD_FECHA].MES, [dbo].[TD_FECHA].SEMANA,
[dbo].[TD_FECHA].DIASEMANA, [dbo].[TD_FECHA].DIA, [dbo].[TD_FECHA].FERIADO,
[dbo].[TD_FECHA].VACACIONES, vDemanda_1.ENERGIA S1, vDemanda_2.ENERGIA AS S2,
vDemanda_3.ENERGIA AS S3, vDemanda_4.ENERGIA AS S4, vDemanda_5.ENERGIA AS S5,
vDemanda_6.ENERGIA AS S6, vDemanda_7.ENERGIA AS S7, vDemanda_8.ENERGIA AS S8,
vDemanda_9.ENERGIA AS S9, vDemanda_10.ENERGIA AS S10, T1.E1, T2.E2
FROM [dbo].[TD_FECHA] INNER JOIN
dbo.vDemanda AS vDemanda_1 ON [dbo].[TD_FECHA].FECHA =
    DATEADD(DD, 7, vDemanda_1.FECHA) INNER JOIN
dbo.vDemanda AS vDemanda_2 ON [dbo].[TD_FECHA].FECHA =
    DATEADD(DD, 14, vDemanda_2.FECHA)
    AND vDemanda_1.HORA = vDemanda_2.HORA INNER JOIN
dbo.vDemanda AS vDemanda_3 ON [dbo].[TD_FECHA].FECHA =
    DATEADD(DD, 21, vDemanda_3.FECHA)
    AND vDemanda_1.HORA = vDemanda_3.HORA INNER JOIN
dbo.vDemanda AS vDemanda_4 ON [dbo].[TD_FECHA].FECHA =
    DATEADD(DD, 28, vDemanda_4.FECHA)
    AND vDemanda_1.HORA = vDemanda_4.HORA INNER JOIN
dbo.vDemanda AS vDemanda_5 ON [dbo].[TD_FECHA].FECHA =
    DATEADD(DD, 35, vDemanda_5.FECHA)
    AND vDemanda_1.HORA = vDemanda_5.HORA INNER JOIN
dbo.vDemanda AS vDemanda_6 ON [dbo].[TD_FECHA].FECHA =
    DATEADD(DD, 42, vDemanda_6.FECHA)
    AND vDemanda_1.HORA = vDemanda_6.HORA INNER JOIN
dbo.vDemanda AS vDemanda_7 ON [dbo].[TD_FECHA].FECHA =
    DATEADD(DD, 49, vDemanda_7.FECHA)
    AND vDemanda_1.HORA = vDemanda_7.HORA INNER JOIN
dbo.vDemanda AS vDemanda_8 ON [dbo].[TD_FECHA].FECHA =
    DATEADD(DD, 56, vDemanda_8.FECHA)
```

```

        AND vDemanda_1.HORA = vDemanda_8.HORA INNER JOIN
dbo.vDemanda AS vDemanda_9 ON [dbo].[TD_FECHA].FECHA =
    DATEADD(DD, 63, vDemanda_9.FECHA)
        AND vDemanda_1.HORA = vDemanda_9.HORA INNER JOIN
dbo.vDemanda AS vDemanda_10 ON [dbo].[TD_FECHA].FECHA =
    DATEADD(DD, 70, vDemanda_10.FECHA)
        AND vDemanda_1.HORA = vDemanda_10.HORA,
(SELECT SUM(ENERGIA) E1 FROM vDemanda WHERE @MiFecha =
    DATEADD(DD, 7, vDemanda.FECHA)) T1,
(SELECT SUM(ENERGIA) E2 FROM vDemanda WHERE @MiFecha =
    DATEADD(DD, 14, vDemanda.FECHA)) T2
WHERE ([dbo].[TD_FECHA].FECHA = @MiFecha)
)

```

Apéndice 7. Código MatLab para ejecutar la red neuronal

```
% Calcula el Predespacho utilizando redes neuronales
% Requisitos:
% ds_dm: es un conector ODBC definido con acceso a la base de datos fuente
% CargaRN: tabla con los datos de entrada para la red
% pronosticoMatLab: table donde se depositará el resultado

% ***** PASO 1: Leer los datos de entrada desde la BD
% Define los parámetros de cómo se devolverán los datos del SQL
s.DataReturnFormat = 'numeric';
setdbprefs(s)

% Crea la conexión con la BD
conn = database(ODBCConnection('ds_dm','sa','desa'));

% Lee datos de la BD.
e = exec(conn,['select * from CargaRN']);
e = fetch(e);
close(e);

% ***** PASO 2: Ejecutar la red
% Asigna datos a la variable de entrada.
input = e.Data;
% display(e.Data);

% Ejecuta la red
y = net(input');

% Define x como la transpuesta del resultado
x = y';

% ***** PASO 3: Prepara y almacena los datos en la BD
% Define el vector para las horas
Horas = (0:23)';

% Define en x la matriz compuesta por la hora y el pronóstico
x = [Horas x];

% Tabla de destino
tableName = 'pronósticoMatLab';

% Columnas de la tabla
fields = {'hora','valor'};

% Borra la tabla
deleteQuery = ['delete from ',tableName,'];
exec(conn,deleteQuery);

% Guarda los datos en la tabla
fastinsert(conn,tableName,fields,x)
close(conn);
```