Universidad CENFOTEC

Maestría en Tecnología de Bases de Datos

Documento final de Proyecto de Investigación Aplicada 2

DISEÑAR UN DATA LAKE EN AWS LAKE FORMATION PARA MEJORAR EL ALMACENAMIENTO Y EL ACCESO DE LOS DATOS

Tencio Zúñiga Karina Francini

Octubre, 2021

Declaratoria de derechos de autor

Agradecimientos

Al profesor tutor José A. Cabezas Jaikel:

Por haber aceptado ser mi tutor durante este proyecto, un tutor que tiene un amplio conocimiento en los temas relacionados con base de datos, compromiso y paciencia. Muchas gracias por luchar conmigo para lograr un proyecto bien elaborado.

A la familia:

Me apoyaron durante todo el proyecto desde el día 1, dándome mucha motivación para lograr concluir con éxito.

-

TRIBUNAL EXAMINADOR

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

JOSE ALBERTO CABEZAS JAIKEL (FIRMA) Digitally signed by JOSE ALBERTO CABEZAS JAIKEL (FIRMA) Date: 2021.07.09 19:46:35 -06'00'

MBD. José Cabezas Jaikel
Tutor

MARCO ANTONIO HERNANDEZ VASQUEZ (FIRMA) Firmado digitalmente por MARCO ANTONIO HERNANDEZ VASQUEZ (FIRMA) Fecha: 2021.07.13 16:10:38 -06'00'

MBD. Marco Hernández Vázquez
Lector 1

IGNACIO TREJOS ZELAYA (FIRMA) Firmado digitalmente por IGNACIO TREJOS ZELAYA (FIRMA) Fecha: 2021.07.15 21:29:53 -06'00'

M. Sc. Ignacio Trejos Zelaya
Lector 2

San José, Costa Rica, 08 de julio de 2021

Firmada digitalmente, de conformidad con la Ley de Certificados, Firmas Digitales y Documentos Electrónicos N° 8454, destacando el artículo 9°-

1. Tabla de Contenido

Lista de tablas

Lista de Figuras

RESUMEN EJECUTIVO

El objetivo principal del presente proyecto es diseñar un Data Lake en AWS Lake Formation para mejorar el almacenamiento y el acceso de los datos. En este se muestra una

propuesta para resolver los diferentes problemas que se enfrentan con el manejo de los datos, la principal causa se centra en los riesgos que representan almacenar volúmenes de datos de diferentes fuentes, no tener un acceso rápido a los datos para todos los usuarios, falta de confianza y seguridad y al mismo tiempo que aumenta la cantidad de datos también aumenta la cantidad de equipos que se necesitan.

Al implementarlo, se va poder dar más valor a los datos para tomar mejores decisiones, porque cualquier tipo de usuario podrá acceder a los datos, no se necesita tener conocimiento técnico para visualizarlos.

En este sentido, los datos utilizados para implementar dicho proyecto son demográficos de doctores en USA, son más de 1 millón usados para realizar la propuesta ((CMS), 2018).

Asimismo, se va a demostrar por qué se escogió Cloud para almacenamiento de los datos y por qué se incluyó Amazon S3 como proveedor en la nube para esta implementación.

En esta propuesta se indica la necesidad de la creación de un Data Lake, los beneficios que ofrece y los prerrequisitos necesarios para lograr implementar este proyecto.

Además, para lograr esta implementación se utilizó la técnica de análisis ICAV, basada en 4 fases que facilitaron el estudio de los datos y el logro de los objetivos.

El diseño de la investigación es aplicado, se va a resolver un problema presente, para lo cual se necesitó investigación para ser desarrollado.

Esta propuesta se realizó con una empresa ficticia.

Palabras Clave: Data Lake, AWS Lake Formation, Amazon S3, AWS Glue, Athena, QuickSight, Big Data, datos, seguridad en Amazon S3.

1 Capítulo 1. Introducción

1.1 Generalidades

Son muchos los problemas para almacenar los datos de forma centralizada, segura y tener una sencilla y acelerada búsqueda de datos. Es por esto que sugerir la creación de un Data Lake,

viene a mejorar muchos de los procesos actuales de la empresa y ayudaría a darle más valor a los datos para tomar mejores decisiones.

La creación del Data Lake por parte del autor de este proyecto en conjunto con la empresa trae beneficios como: capacidad de obtener valor a partir de tipos ilimitados de datos, posibilidad de almacenar todo tipo de datos estructurados y no estructurados, es útil para todo tipo de usuarios.

La creación del Data Lake se va a ejecutar con datos demográficos de doctores especialistas ((CMS), 2018), no hay ninguna cláusula de confidencialidad, por cuanto puede ser público.

1.2 Antecedentes del problema

Esta propuesta se va a realizar con una empresa ficticia, el nombre de esta se conocerá como Zuteka. Esta cumple con la función de garantizar la calidad de los algoritmos que entregan a sus clientes. Actualmente, se está enfrentado a distintos problemas relacionados con el manejo de los datos, desde el almacenamiento por su aumento hasta su acceso. Este problema está evitando tener una absoluta confianza en la versión final de sus algoritmos.

Existen diversas propuestas que la empresa estuvo evaluando para lograr un mejor manejo de datos, entre esas está la creación de un Data Lake con AWS Lake Formation.

AWS ofrece un conjunto integrado de servicios que proveen todo lo necesario para crear y administrar de manera rápida y sencilla un Data Lake (AWS, AWS, 2019). Los Data Lakes con tecnología de AWS pueden manejar el nivel de escala, agilidad y flexibilidad requerido para combinar diferentes tipos de datos y enfoques analíticos con el fin de obtener información más detallada y almacena cualquier tipo de datos de manera segura.

1.3 Definición y descripción del problema.

El problema se centra en los riesgos que representa almacenar volúmenes de datos de diferentes fuentes, no tener un acceso ágil a los datos para todos los usuarios, falta de confianza en la seguridad.

1.4 Justificación

Este proyecto surgió por la tarea que tienen las organizaciones de administrar mayores volúmenes de datos, de más fuentes y contener más tipos de datos que nunca. En vista de los volúmenes de datos masivos y heterogéneos, muchas organizaciones están descubriendo que, para brindar información comercial oportuna, necesitan una solución de almacenamiento y análisis que ofrezca más velocidad y flexibilidad que los sistemas heredados. De ahí que la utilización de un Data Lake desarrollado en este proyecto pretendía abastecer a la empresa con las herramientas necesarias para combatir los problemas que enfrentan con los datos.

Entre las principales ventajas de utilizar un Data Lake, está el almacenar todos sus datos de forma estructurada o no estructurada en un repositorio centralizado y así tener un acceso rápido y lograr obtener más valor de los datos (Services, Data Lake Solution, 2016).

Este proyecto tuvo como iniciativa dotar a la empresa de una herramienta para almacenar todos sus datos, tener un acceso fácil y ágil y así lograr obtener más valor sobre ellos.

1.5 Viabilidad

El proyecto se hará con apoyo del personal experto en el área de datos, pero no lo es en el tema de Data Lake, por lo que se vio como una oportunidad de aprendizaje.

1.5.1 Punto de vista técnico.

Al ser parte del equipo de datos, se cuenta con la experiencia necesaria para realizar dicho proyecto.

Se va a utilizar AWS Lake Formation, para la implementación. Los datos se van a almacenar en Amazon S3. Este es un servicio de almacenamiento de objetos, creado para almacenar y recuperar cualquier volumen de datos.

Como herramienta para catalogar los datos, AWS Lake Formation utiliza crawlers de AWS Glue para obtener los metadatos y crear un catálogo con ellos. El usuario administrador podrá crear los accesos que considere convenientes, y AWS Lake Formation se encargará de bloquear o permitir el acceso a los datos a otros usuarios.

Para realizar el análisis de datos, se va a utilizar Amazon Athena, el cual es un servicio de consultas interactivo que facilita el análisis de datos en Amazon S3 con SQL estándar (Admin, 2019).

La razón principal porque se escogió AWS por encima de otros proveedores, es el actual líder de los servicios en la nube, Azure es el segundo en la lista (Gartner, 2001). AWS ofrece los servicios por separado lo que permite construir el proyecto a la medida (construir la casa a la medida) mientras que Azure vende la herramienta ya construida (casas prefabricadas), por lo tanto, AWS se adapta más a las necesidades de este proyecto de poder elegir los servicios que se requieren para desarrollarlo.

1.5.2 Punto de vista operativo.

Para la realización de este proyecto se utilizaron datos demográficos de doctores especialistas. Estos fueron limpiados para la creación del Data Lake y sus diferentes análisis. Desde el punto de vista operativo, el funcionamiento de la empresa no se verá afectado o interrumpido durante su implementación. En el ámbito técnico, la empresa cuenta con el personal capacitado, en caso de requerirse algún cambio dentro de los requerimientos iniciales establecidos.

1.5.3 Punto de vista económico.

Implementar este proyecto es realmente viable. El Data Lake se va a desarrollar con un tiempo máximo de 4 meses. Se va a utilizar la cuenta de AWS Lake Formation, para el desarrollo. Esta cuenta tiene todo lo necesario para su implementación, el único equipo físico que se requiere es la computadora; adicional a eso, están disponibles los diferentes servicios en la cuenta. En la tabla 1, se detallan los costos de cada servicio requerido para el desarrollo del proyecto.

Tabla 1.1. Costo de elaboración del proyecto

Servicio	Descripción	Costo por mes	Costo Total Estimado
Amazon S3	Almacenando hasta 6 GB de datos por mes	₡ 7 800	₡ 31 200
AWS Glue	Incluye el catálogo de datos y la creación de los ETLs	₡ 100 000	₡ 400 000
Amazon Athena	Realizando 3100 consultas por día y cada consulta escanea 500 MB de datos	₡ 4 350	₡ 17 400
	Total	₡112 150	₡448 600

Fuente: Elaboración propia

Los siguientes costos corresponden al pago del ingeniero que se requiere para la implementación del proyecto. Esos costos se definieron, comparando salarios de personas con las mismas habilidades que se necesitan para el desarrollo del trabajo y haciendo una búsqueda de salarios por especialidad en el Ministerio de Trabajo.

Tabla 1.2. Costo de elaboración del recurso del proyecto

Puesto	Salario por hora	Tiempo Estimado	Costo Estimado
Senior Data Engineer	₡ 7 500	160 horas	₡ 1 200 000
	Total	160 horas	₡ 1 200 000

Fuente: Elaboración propia

1.6 Objetivos

Los objetivos planteados a continuación se definen siguiendo la taxonomía de Bloom, la cual orienta hacia dónde se quiere llegar con el proyecto.

1.6.1 Objetivo general

Diseñar un Data Lake en AWS Lake Formation para mejorar el almacenamiento y el acceso de los datos.

1.6.2 Objetivos específicos.

A continuación, se detallan los objetivos específicos:

- Preparar los datos para ser consumidos por el Data Lake.
- Proponer un almacenamiento que se ajuste al problema que se quiere resolver.
- Desplegar el Data Lake para que pueda seguir recibiendo datos actualizados y logre cumplir con su función.
- Proponer una herramienta para la visualización de los datos, para mejorar el acceso de los datos para todo tipo de usuario.

- Definir las políticas de seguridad para los diferentes usuarios y así lograr obtener más valor a los datos.

1.7 Alcances y limitaciones

1.7.1 Alcances.

Este proyecto tiene como alcance el Diseño y desarrollo de una guía para la implementación de un Data Lake en un ambiente Big Data en AWS Lake Formation, para una empresa consultora que ofrece diferentes soluciones. Lo que comprende a su vez los siguientes alcances:

- Se va a implementar una guía para que sea de fácil uso para todos los usuarios.
- El presente proyecto abarca únicamente el uso de AWS Lake Formation para el desarrollo del Data Lake
- Lograr que los datos estén disponibles en Athena.

1.7.2 Limitaciones.

- El Data Lake estará disponible únicamente en AWS Lake Formation
- Los datos son ficticios
- El acceso a los datos para realizar diferentes análisis se podrá hacer únicamente en Athena

1.8 Marco de referencia organizacional y socioeconómico

La empresa que se utilizó para el presente proyecto es ficticia.

1.8.1 Historia.

Zuteka es una consultora global que se enfoca en resolver problemas concernientes a la administración estratégica. Trabaja prestando sus servicios a empresas, gobiernos e instituciones.

1.8.2 Tipo de negocio y mercado meta.

Es una consultora que sirve a empresas, líderes, gobiernos, organizaciones no gubernamentales y organizaciones sin fines de lucro.

Ayuda a los clientes a realizar mejoras duraderas en su desempeño y lograr sus metas más importantes. No importa el reto, se enfocan en ofrecer resultados prácticos y duraderos.

1.8.3 Misión, Visión y Valores.

La misión es ayudar a los clientes a realizar mejoras distintivas, duraderas y sustanciales en su desempeño.

La visión es ayudar a las principales empresas e instituciones públicas afrontar sus retos más importantes y a lograr mejoras destacadas y sostenibles en su desempeño.

Nuestros valores:

- Poner los intereses del cliente por delante de la empresa
- Gestionar los recursos del cliente y de la empresa de manera rentable
- Mejorar el rendimiento de nuestros clientes significativamente
- Crear un ambiente incomparable para personas excepcionales
- Desarrollarse mutuamente a través del aprendizaje y la tutoría

1.8.4 Políticas institucionales.

Es una empresa ficticia, por lo cual ninguna política incide de manera directa con el presente proyecto a desarrollar.

1.9 Estado de la cuestión

Las revisiones sistemáticas se hacen para identificar, evaluar e interpretar las investigaciones de un tema o problema. El presente apartado se va a desarrollar tomando como base la plantilla publicada en el libro “Systematic Review in Software Engineering”. (Biolchini, 2017)

1.9.1 Planificación de la revisión.

1.9.1.1 Formulación de la pregunta.

Localizar diferentes proyectos de Data Lakes que hayan sido implementados en un ambiente Big Data en AWS Lake Formation

Foco de la pregunta

En esta revisión sistemática se pretende localizar proyectos centrados en el uso de Data Lakes en un ambiente Big Data en AWS Lake Formation

Amplitud y calidad de la pregunta

Para definir la amplitud y calidad de la pregunta, se basaron en la respuesta a una serie de apartados en los que se analiza el problema a tratar, se propone la pregunta de investigación de los resultados que esperan obtener y cómo serán analizados.

- **Problema**: Se necesita una solución de almacenamiento y análisis que ofrezca más velocidad y flexibilidad que los sistemas heredados.
- **Pregunta de investigación: ¿**Qué proyectos aplicados en Big Data se han llevado a cabo la implementación de Data Lakes en AWS Lake Formation?
- **Palabras clave y sinónimos:**
    - Data Lake en Big Data con AWS: Data Lakes, Big Data, AWS, AWS Lake Formation
- **Intervención:** Propuestas existentes sobre la implementación de Data Lakes en un ambiente Big Data en AWS Lake Formation
- **Resultado:** El resultado de esta investigación fue la identificación de propuestas existentes de Data Lakes que se ajustara de manera confiable a la implementación de este proyecto en la empresa.
- **Medida de salida:** índice de confiabilidad de la implementación de un Data Lake en el contexto de Big Data en AWS Lake Formation
- **Población:** Publicaciones sobre la implementación de Data Lakes en ambiente Big Data en AWS Lake Formation
- **Aplicación:** El beneficiario directo de esta revisión sistemática fueron los usuarios que van a utilizar el Data Lake.
- **Diseño Experimental:** Ninguno fue aplicado

1.9.1.2 Selección de fuentes.

Definición del criterio de selección de fuentes

El criterio para la selección de las fuentes está basado en la opinión de los autores de este proyecto basándose en su experiencia profesional; estos recomendaron la lista de fuentes, así como la accesibilidad web usando motores de búsqueda con palabras claves y consultas avanzadas.

Lenguaje de estudio

La búsqueda de fuentes fue en inglés y español pero los estudios primarios fueron en inglés.

Identificación de fuentes

Método de selección de fuentes: Investigación por medio de consultas en buscadores web.

Lista de fuentes:

- IEEE Explore
- AWS
- Udemy

Cadenas de búsqueda: Data Lakes in Big Data with AWS

Selección de fuentes después de la evaluación

Todas las fuentes cumplieron con el criterio de calidad

Comprobación de las fuentes

Todas las fuentes fueron aprobadas

1.9.1.3 Selección de los estudios.

Definición de estudios:

Definición del criterio de inclusión y exclusión de estudios: Los estudios debían presentar proyectos relacionados con Data Lakes en un ambiente Big Data en AWS Lake Formation, por lo cual, el análisis del contenido de las publicaciones por medio del índice y del resumen ejecutivo con base a las palabras clave dio la relevancia para la revisión sistemática.

Procedimiento para la selección de los estudios: El procedimiento inició con la aplicación de la cadena de búsqueda en las fuentes seleccionadas. Se refinó la selección con una lectura más profunda de las fuentes seleccionadas.

Definición de tipos de estudio: Se seleccionaron los tipos de estudio que cumplieran con el criterio de selección de fuente y que pasaran el procedimiento definido para la selección de estudio.

1.9.2 Ejecución de la revisión

1.9.2.1 Selección de la ejecución.

Los siguientes estudios fueron seleccionados:

Tabla 1.3. Selección de estudios iniciales

Fuente	Estudio	Autor	Año
AWS	Data Lake Solution	Amazon Web Services	2016
AWS	Data Lakes and Analytics on AWS	Amazon Web Services	2020
IEEE Explore	Managing Data Lakes in big data	Huang Fang	2015
AWS	Introduction to AWS Lake Formation	Amazon Web Services	2018
AWS	Building Your Data Lake on AWS	Amazon Web Services	2019
Udemy	Big Data in the AWS (Amazon Web Services)	AWS Data Consultant	2019

Fuente: Información obtenida de diferentes fuentes

1.9.2.2 Extracción de la información.

Inclusión de la información y exclusión del criterio de definición: La información extraída de los estudios debía contener demos, conceptos generales o cualquier otra cosa relacionada con la implementación de Data Lakes en Big Data con AWS Lake Formation

Formularios de extracción de datos

Tabla 1.4. Formulario de extracción de datos

Extracción de resultados objetivos: Estos detalles se pueden obtener de cada estudio
• Identificación del estudio: incluye título de publicación, autores y año.
• Metodología del estudio: método utilizado para conducir el estudio.
• Resultados del estudio: efectos obtenidos por medio de la ejecución del estudio.
• Problemas de estudio: limitaciones de estudio.
Extracción de resultados subjetivos: son apreciaciones personales
• Información mediante autores: información obtenida directamente del autor.
• Impresiones generales y abstracciones: información de otros lectores.

Fuente: Elaboración propia

Extracción de la ejecución

A continuación, se va analizar cada fuente seleccionada

Tabla 1.5. Primera fuente investigada

Extracción de resultados objetivos
Identificación del estudio: Data Lake Solution, Amazon Web Services, 2016

Metodología del estudio: El apartado busca referenciar la creación de Data Lakes desde la etapa de inicio.
Resultados del estudio: Da conocimiento para la creación de un Data Lake en la plataforma de AWS
Problemas de estudio: N/A
Extracción de resultados subjetivos
Información mediante autores: N/A
Impresiones generales y abstracciones: El estudio muestra cualidades esenciales para el desarrollo de una Data Lake en AWS6

Fuente: AWS

Tabla 1.6. Segunda fuente investigada

Extracción de resultados objetivos
Identificación del estudio: Data Lakes and Analytics on AWS, Amazon Web Services, 2020
Metodología del estudio: El apartado muestra los conceptos de forma detallada de que es un Data Lakes y que lo compone.
Resultados del estudio: Da conocimiento de que es un Data Lake con Big Data en AWS
Problemas de estudio: N/A
Extracción de resultados subjetivos

Información mediante autores: N/A
Impresiones generales y abstracciones: El estudio muestra cualidades esenciales teóricas para lograr crear el Data Lakes.

Fuente: AWS

Tabla 1.7. Tercera fuente investigada.

Extracción de resultados objetivos
Identificación del estudio: Managing Data Lakes in big data, Huang Fang, 2016
Metodología del estudio: El apartado muestra las diferentes formar de manejar Data Lakes con Big Data
Resultados del estudio: Da conocimiento de cómo crear un Data Lakes en diferentes ambientes.
Problemas de estudio: N/A
Extracción de resultados subjetivos
Información mediante autores: N/A
Impresiones generales y abstracciones: El estudio muestra diferentes tecnologías para la creación de Data Lakes

Fuente: IEEE Explore

Tabla 1.8. Cuarta fuente investigada

Extracción de resultados objetivos
Identificación del estudio: Introduction to AWS Lake Formation, Amazon Web Services, 2018.
Metodología del estudio: El apartado muestra cómo crear un Data Lake en AWS Lake Formation.
Resultados del estudio: Explica de forma detallada el uso de AWS Lake Formation.
Problemas de estudio: N/A
Extracción de resultados subjetivos
Información mediante autores: N/A
Impresiones generales y abstracciones: El estudio muestra AWS Lake Formation como tecnología para crear el Data Lake

Fuente: AWS

Resolución de divergencias entre los revisores: No existe divergencia alguna.

1.9.3 Análisis de los resultados.

Resultado calculo estadístico: No se utilizó ningún cálculo estadístico.

Presentación de los resultados:

Tabla 1.9. Resultados de las fuentes

Fuente	Estudios	Relevantes	Excluidos	Primarios
AWS	15	4	11	3
Udemy	5	1	4	0
IEEE Xplore	10	1	9	1

Fuente: Elaboración propia

Análisis de sensibilidad: No aplicable

Comentarios finales:

- Número de estudios: 30 estudios encontrados, solo 5 fueron seleccionados.
- Sesgo de búsqueda: No fue definido.
- Sesgo de publicación: No fue definido.
- Variación entre revisores: No hay variaciones
- Aplicación de resultados: Los estudios mostraron formas de desarrollar Data Lakes en un ambiente Big Data en AWS.
- Recomendaciones: Ninguna

2 Capítulo 2. Marco Conceptual

A continuación, se presenta una serie de elementos teóricos que permiten sustentar la investigación, específicamente en los conceptos principales de estudio “Data Lake”,” AWS Lake Formation”, “Big Data”, la definición de términos básicos y la funcionalidad de cada variable aplicada y relacionada a la investigación.

2.1 Data Lake en AWS

Es un repositorio centralizado que le permite almacenar todos sus datos estructurados y no estructurados. Puede almacenar sus datos tal cual, sin tener que estructurar los datos primero y luego ejecutar diferentes tipos de análisis (Services, Data Lake Solution, 2016).

El Data Lake almacena cualquier tipo de datos de manera segura, desde gigabytes a exabytes.

2.2 Big Data

Big Data es un término en el que se agrupan toda clase de técnicas de tratamiento de grandes volúmenes de datos, fuera de los análisis y herramientas clásicas. Este concepto engloba muchas ideas y aproximaciones, pero todas con un objetivo común: extraer información de valor de los datos, de forma que pueda ser de ayuda para las decisiones y procesos de negocio (Big Data, Consulting, 2012).

Las características más importantes del Big Data perfectamente se pueden clasificar en cuatro magnitudes, más conocidas como las cuatro *V* del Big Data, relativas a volumen, variedad, velocidad y veracidad. A estas cuatro *V*, se pueden añadir tres más, como son la de Viabilidad y

Visualización. Pero si se habla de *V* en Big Data no se debe dejar pasar la principal característica del análisis de datos que es la *V* de Valor de los datos. En la actualidad se empieza a hablar, ya no de las tradicionales cuatro V de Big Data, sino de las 7 "V" del Big Data (Big Data, Consulting, 2012). A continuación, se explican las 7 "V":

1. Volumen : se refiere a la cantidad de datos que son generados cada segundo, minuto y días en nuestro entorno.

2. Velocidad: se refiere a los datos en movimiento por las constantes interconexiones que realizamos, es decir, a la rapidez en la que son creados, almacenados y procesados en tiempo real.

3. Variedad de los datos: se refiere a las formas, tipos y fuentes en las que se registran los datos. Estos datos pueden ser datos estructurados y no estructurados.

4. Veracidad de los datos: hace referencia a la incertidumbre de los datos, es decir, al grado de fiabilidad de la información recibida.

5. Viabilidad: la inteligencia empresarial es un componente fundamental para la viabilidad de un proyecto y el éxito para la empresa. Se trata de la capacidad que tienen las compañías en generar un uso eficaz del gran volumen de datos que manejan.

6. Visualización de los datos: esto hace referencia al modo en el que los datos son presentados. Una vez que estos son procesados, se necesita representarlos visualmente de manera que sean legibles y accesibles, para encontrar patrones y claves ocultas en el tema a investigar.

7. Valor de los datos: el dato no es valor. El valor se obtiene de datos que se transforman en información; esta a su vez se convierte en conocimiento, y este en acción o en decisión. El valor de los datos está en que sean accionables, es decir, que los responsables de las empresas puedan tomar una decisión (la mejor decisión) en base a estos datos.

2.3 AWS Lake Formation

AWS Lake Formation es un servicio que facilita la configuración de un Data Lake seguro en cuestión de días.

Con Lake Formation, puede trasladar, almacenar, catalogar y limpiar los datos más rápido.

AWS Lake Formation recopila y cataloga los datos de bases de datos y almacenamiento de objetos, los traslada al nuevo Data Lake de Amazon S3, los limpia y los clasifica mediante algoritmos de aprendizaje automático y aporta seguridad al acceso a su información confidencial. Sus usuarios pueden acceder a un catálogo de datos centralizado que describe los conjuntos de datos disponibles y su uso adecuado. Luego, aprovechan estos conjuntos de datos con los servicios de análisis y aprendizaje automático. Lake Formation se basa en las capacidades disponibles en AWS Glue.

Lake Formation tiene aprendizaje automático integrado para realizar la de duplicación y encontrar registros coincidentes (dos entradas que se refieren a lo mismo) a fin de aumentar la calidad de los datos (Services, Data Lakes and Analytics on AWS, 2020).

2.4 Catálogo de datos

El catálogo de datos de AWS Glue es un repositorio central que almacena metadatos estructurales y operativos para sus recursos de datos. Para un conjunto de datos determinado, puede almacenar la definición de la tabla y la ubicación física, agregar atributos relevantes para la empresa y realizar un seguimiento de cómo los datos han cambiado con el tiempo.

AWS Glue es un servicio de extracción, transformación y carga (ETL) completamente administrado y automatiza la ardua tarea de la preparación de datos para el análisis. AWS Glue encuentra y clasifica los datos automáticamente mediante el catálogo de datos de Glue, genera código ETL para transformar sus datos de origen en esquemas de destino y ejecuta los trabajos ETL en un entorno Apache Spark escalable y completamente administrado para cargar los datos en su destino (AWS, AWS, 2019).

2.5 Amazon S3

Amazon Simple Storage Service (Amazon S3) es un servicio de almacenamiento de objetos que ofrece escalabilidad, disponibilidad de datos, seguridad y rendimiento líderes en el sector. Esto significa que clientes de todos los tamaños y sectores pueden utilizarlo para almacenar y proteger cualquier cantidad de datos para diversos casos de uso, como sitios web, aplicaciones móviles, procesos de copia de seguridad y restauración, operaciones de archivado, aplicaciones empresariales, dispositivos IoT y análisis de big data (AWS, AWS, 2019).

2.6 Athena

Amazon Athena es un servicio de consultas interactivo que facilita el análisis de datos en Amazon S3 con SQL estándar.

Athena no tiene servidor, de manera que no es necesario administrar infraestructura y solo paga por las consultas que ejecuta.

La mayoría de los resultados se proporciona en cuestión de segundos. Con Athena, no es necesario realizar trabajos complejos de ETL para preparar los datos para el análisis. Por ello, cualquier persona con habilidades SQL puede analizar conjuntos de datos a gran escala de forma rápida y sencilla (Amazon Athena, 2019).

2.7 Tipos de datos

Datos estructurados

Los datos estructurados es la información que se suele encontrar en la mayoría de las bases de datos relacionales (RDBMS). Suelen ser archivos de texto que se almacenan en formato tabla, hojas de cálculo o bases de datos relacionales con títulos para cada categoría que permite identificarlos.

Ejemplo de datos estructurados:

Ilustración 2.1. Ejemplo de datos estructurados

	nombre	color	edad	altura	peso	puntuacion
1:	Paco	Rojo	24	182	74.8	83
2:	Juan	Green	30	170	70.1	500
3:	Andres	Amarillo	41	169	60.0	20
4:	Natalia	Green	22	183	75.0	865
5:	Vanesa	Verde	31	178	83.9	221
6:	Miriam	Rojo	35	172	76.2	413
7:	Juan	Amarillo	22	164	68.0	902

Fuente: Elaboración propia

Datos no estructurados

Los datos no estructurados se caracterizan por no tener un formato específico. Se trata de un cúmulo de información que deben identificarse y almacenarse de forma organizada a través de una base de datos no relacional (NoSQL).

Se almacenan en múltiples formatos como documentos PDF o Word, correos electrónicos, ficheros multimedia de imagen, audio o video.

3 Capítulo 3. Marco Metodológico

En este capítulo se describen los aspectos metodológicos de este proyecto. Se detalla el tipo de investigación, la técnica de recolección de datos y metodologías aplicadas.

3.1 Tipo de investigación

La investigación es un conjunto de procesos sistemáticos, críticos y empíricos que se aplican al estudio de un fenómeno o problema.

Esta investigación, requiere un enfoque cualitativo porque los datos que se reúnen son recolectados en forma de texto, basados en análisis que la empresa va a implementar y se realiza un periodo de análisis para obtener únicamente los datos relevantes.

El diseño de esta investigación es aplicado, se va a resolver un problema presente que vive la empresa; para lograr resolverlo, se necesita de un estudio para ser desarrollado.

3.2 Alcance investigativo

El alcance de una investigación indica el resultado que se obtendrá a partir de ella y condiciona el método que se seguirá para obtener dichos resultados.

El alcance investigativo del presente proyecto es exploratorio, presenta muchas dudas para ser desarrollarlo y es una tecnología nueva.

3.3 Enfoque

Se utilizará el enfoque alternativo.

- Base ontológica: el tema permite conocer cómo manejar grandes cantidades de datos, cómo crear y utilizar un Data Lake.
- Base epistemológica: el investigador está involucrado y se encargará de desarrollar el Data Lake
- Base axiológica: requiere un especial tratamiento en cuanto a la facilidad de implementación del objetivo para la empresa, la facilidad de implementación será evaluada a la escala de valores mostrada en la siguiente tabla.

Tabla 3.1. Evaluación de escala de valores para determinar la facilidad de implementación

Descripción	Valor	Opciones
Alternativas de implementación	15%	Opción única: 5% Dos opciones:10% Más de dos opciones: 15%
Preparación de los datos	20%	2 semanas: 20% 3 semanas: 15% 4 semanas: 10%
Cantidad de líneas de código	35%	Más de 2000 lineas15% De 1000 a 2000 líneas: 25% Menos de 1000 líneas: 35%

Preparación con las nuevas tecnologías	15%	2 semanas: 20% 3 semanas: 15% 4 semanas: 10%

Fuente: Elaboración propia

Cuanto mayor sea el porcentaje asignado, más fácil es de implementar la herramienta. Se considerará que al obtener un 80% o más, significa que el algoritmo es fácil de implementar, de un 50% a un 80% significa que es medianamente complejo de implementar y menos de un 50% significa que es de implementación muy compleja.

3.4 Diseño

El diseño guía al investigador en lo que se debe hacer para alcanzar los objetivos propuestos. Se va a utilizar diseño de campo, van a obtener información de las fuentes seleccionadas en el estado de la cuestión.

Se va a utilizar el diseño experimental, someterán a un grupo de personas expertas en el tema para que los guíe de cómo deben realizar la implementación del objetivo general de este proyecto

3.5 Población y muestreo

Se debe tener en cuenta, que nunca es posible estudiar la totalidad de la población, por ello el muestreo es una parte importante que se realiza en el proceso de investigación.

El muestreo como lo define Piergiorgio (2017) es el procedimiento por el cual de un conjunto de unidades que forman el objeto de estudio (población), se elige un número reducido de

unidades (muestra) aplicando criterios tales que permiten generalizar los resultados obtenidos del estudio de la muestra de la población.

Para este caso específico, la muestra la integran los usuarios responsables de cargar los datos, analizarlos y hacer uso del Data Lake; al ser una población pequeña, la muestra incluye a toda la población, siendo esta realmente significativa.

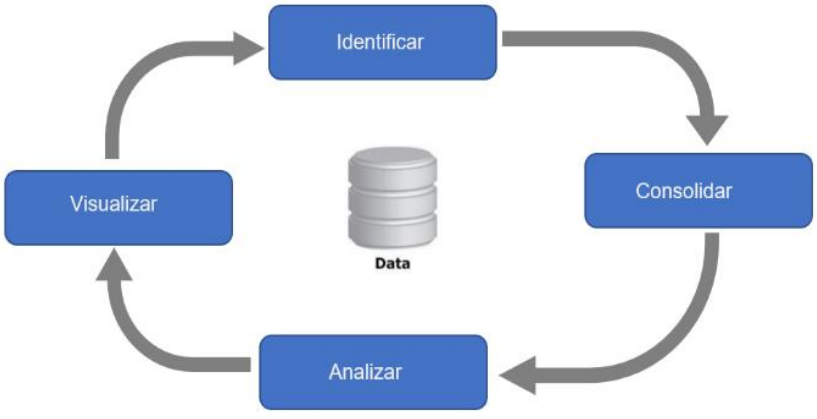
3.6 Instrumentos para la recolección de datos

Los datos utilizados están a disposición en el sitio web de datos abiertos “The Centers for Medicare & Medicaid Services”, tienen información demográfica de doctores y están disponibles para descarga.

3.7 Técnicas para análisis de información

A continuación, se muestra un gráfico con la técnica de análisis que se utilizó (ICAV)

Ilustración 3.1. Ejemplo de datos estructurados



Fuente: Elaboración propia

Tabla 3.2. Definición de la técnica de análisis (ICAV)

Identificar	Identificar claramente las necesidades del negocio. Identificar usuarios finales. Conocer las preguntas que requieren responder para tomar mejores decisiones empresariales
Consolidar	Ubicar donde están las fuentes de información requeridas para responder las preguntas del negocio Identificar que fuentes de información no estructuradas se requieren para hacer el análisis
Analizar	Realizar el análisis de grandes volúmenes de información utilizando técnicas avanzadas de análisis predictivo y minería de datos

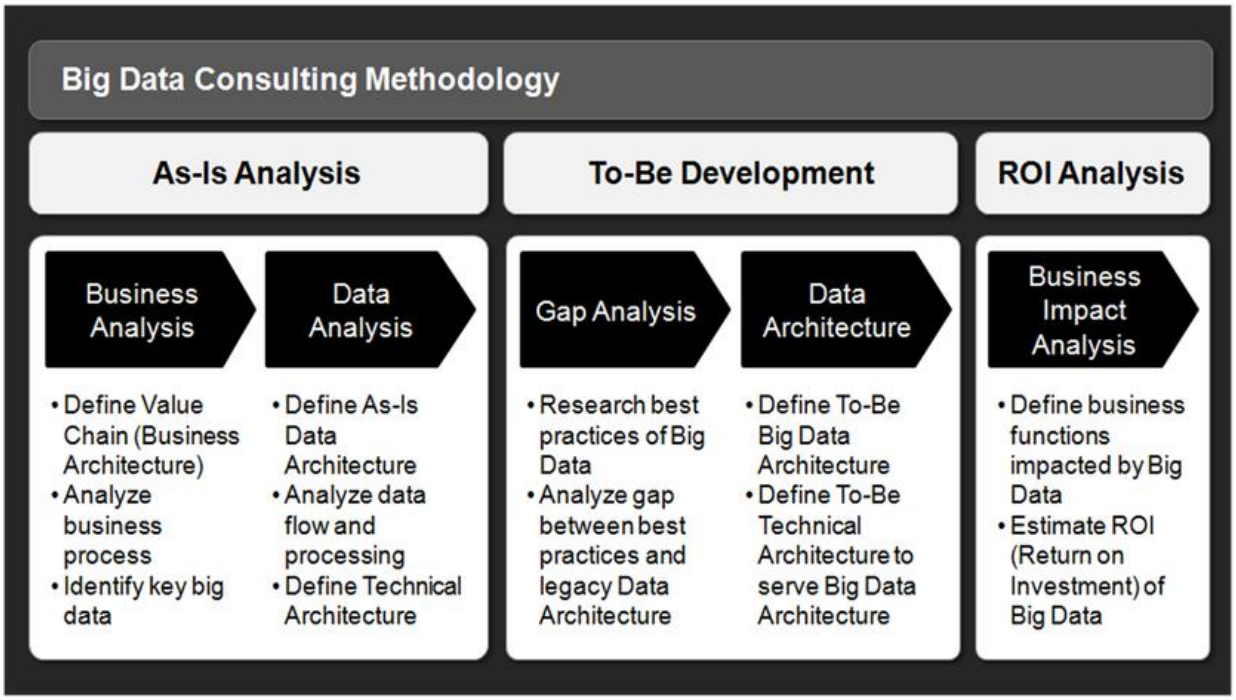
Visualizar	Muestra la visualización de los análisis.

Fuente: Elaboración propia.

3.8 Estrategia de desarrollo de la propuesta

A continuación, se muestra la estrategia que se desarrolló para lograr con éxito el proyecto. (Big Data Consulting Methodology)

Ilustración 3.2. Big Data Consulting Methodology



Fuente (Big Data, Consulting, 2012)

1. **Análisis de negocio**:
    - Analizar los procesos de negocio.
    - Identificar clave de Big Data.

2. **Análisis de datos**:
    - Definir la arquitectura de datos.
    - Analizar el flujo de datos y el procesamiento.
    - Definir la arquitectura técnica.
3. **Análisis Gap**:
    - Investigar las mejores prácticas para el tratamiento de grandes volúmenes de datos.
4. **Arquitectura de datos:**
    - Definir arquitectura de datos.
    - Definir Arquitectura técnica para soportar la arquitectura de datos.
5. **Análisis del impacto comercial**
    - Definir las funciones del negocio afectadas por Big Data.
    - Estimar el ROI (retorno de inversión).

4 Capítulo 4. Análisis del diagnóstico

4.1 Actualidad de los datos

En la actualidad hay muchas empresas que manejan los datos sin caer en soluciones de Big Data, enfrentan problemas de rendimiento, el uso de datos no estructurados, redundancia de datos y mucho trabajo manual.

En algunas entidades estatales de nuestro país, no tienen tecnología Big Data y manejan las bases de datos por medio de Excel; por ello la actualización de esta, la realizan manual, tienen varios archivos con los datos, como resultado no los tienen centralizados y solo 1 usuario tiene acceso a esa base de datos. Para realizar algún reporte, deben hacerlo manual y semanalmente.

Otro problema frecuente es lograr que las bases de datos cuenten con un desempeño óptimo. Cuando se diseña una base de datos, se debe asegurar de que realiza todas las operaciones importantes de forma rápida y correcta. Algunos problemas de rendimiento se pueden resolver una vez que la base de datos se encuentra en producción. Sin embargo, otros pueden ser el resultado de un diseño inadecuado y se pueden solucionar mediante el cambio de la estructura y el diseño de la base de datos.

Las empresas buscan almacenar los datos de forma rápida, segura y en tiempo real, y a la vez necesitan tener una fácil y acelerada búsqueda de datos es por este motivo que utilizar un Data Lake con Big Data, es una solución perfecta a los problemas que se enfrentan.

4.2 Almacenamiento de los datos

Muchas empresas almacenan sus datos en disco y no en la nube, porque en algunos casos deben cumplir normativas internas, hoy en día hay muchas opciones para almacenar los datos y con resultados altamente positivos, pero el que una empresa siga almacenando en disco no significa que sea una opción mala.

Entre las ventajas para almacenar los datos en la nube son las siguientes:

1. Protección de datos: la empresa configura qué usuarios van a tener acceso y qué tipo de acceso. También es posible proteger los datos confidenciales con control de los usuarios.
2. Disminución de costos de almacenamiento
3. Facilidad en el intercambio de la información: cuando los datos están concentrados y accesibles en internet, es más fácil el intercambio de información y actualización de archivos en tiempo real, al que las personas tendrán acceso, aunque estén físicamente en diferentes ubicaciones.
4. Almacenamiento flexible: en la nube es posible aumentar o disminuir el espacio de almacenamiento, dependiendo de la cantidad y tamaño de archivos, se contrata solamente el servicio que será realmente utilizado.

En este proyecto se utilizó Amazon S3 para el almacenamiento de datos, es un servicio que ofrece escalabilidad, disponibilidad de datos, seguridad y facilidad en el intercambio de información.

Amazon S3, proporciona características de administración fáciles de utilizar, por lo que permite que cualquier empresa pueda organizar los datos y configurar los accesos de sus usuarios. Con Amazon S3, se paga únicamente por lo que se utiliza, no hay un cargo mínimo.

No se requieren cuotas de configuración ni compromisos para comenzar a utilizar el servicio. El pago por uso le permite adaptarse con facilidad a las cambiantes necesidades de la empresa sin comprometerse a dedicar presupuestos excesivos y, además, mejorar la capacidad de respuesta ante los cambios. Con el modelo de pago por uso, puede adaptar su empresa en función de las necesidades y no de previsiones, con lo que se reduce el riesgo de aprovisionar capacidad insuficiente o en exceso.

4.3 Para qué se necesita un Data Lake

Actualmente se está en la era digital y de transformación tecnológica donde las tecnologías de la información tienen cada vez más un papel muy importante en nuestras tareas del día a día. Todo esto implica un crecimiento desproporcionado de los datos, datos valiosos y que en un futuro pueden ser de gran utilidad para la toma de decisiones en las empresas, pero el problema es saber qué hacemos con la cantidad de datos, como los gestionamos y los organizamos para no desperdiciar su valor. Datos que en la fecha actual pueden carecer de utilidad para la empresa o negocio, por lo que si no se usan o se gestionan bien se está perdiendo el valor de los datos.

Por este motivo las empresas optan por conservar todos los datos que generan sus diferentes fuentes de información y si en un futuro se necesitan se pueda tener acceso a ellos. Viendo estas necesidades, la opción es un almacenamiento de datos de forma indefinida.

Una vez que se pone en marcha el análisis de los datos almacenados en el Data Lake, se pueden realizar muchas acciones.

4.4 Beneficios de un Data Lake

El principal beneficio de un Data Lake es la centralización de datos de diferentes fuentes (soporta cualquier tipo de datos). Al iniciar un proyecto de Data Lake, es necesario tener una alineación muy fuerte con el negocio.

Las medidas de seguridad en el Data Lake pueden ser asignadas de manera que se otorga acceso por usuarios.

Los datos se preparan según sea necesario, lo que reduce los costos de preparación sobre el procesamiento inicial (tal como seria requerido por los Data Warehouses). Una estructura de Big Data permite escalar este procesamiento para incluir los conjuntos de datos más grandes posibles.

Los usuarios de diferentes departamentos, sin importar su ubicación física, van a tener acceso al Data Lake de manera fácil. Esto ayuda a la organización para impulsar las decisiones empresariales.

4.5 Caso de uso de un Data Lake con tecnología en Big Data en Amazon S3

Grandes compañías alrededor del mundo han implementado y están utilizando Data Lakes. A continuación, se explica un caso de uso que describe el problema y la solución y por qué tuvieron que elegir esta tecnología.

La empresa Zalando, fundada en 2008, es la plataforma online líder en Europa de moda con más de 32 millones de clientes activos.

Problema: Zalando estaba generando muchos datos, necesitaban una opción de almacenamiento que pudiera hacer frente a las crecientes cantidades de información y la opción a elegir, debía ser escalable y confiable. La base de datos que manejaban era transaccional o analítica, al mismo tiempo que aumentaba la cantidad de material, también aumentaba la cantidad de equipos necesarios.

Zalando necesitaba obtener información valiosa sobre los datos, almacenarlos de forma rápida, segura y a la vez tener acceso fácil a estos.

Solución: Zalando, tomo la decisión de usar Amazon S3 como el pilar principal de la infraestructura de datos y así lograr construir el Data Lake. Amazon S3 evolucionó con el tiempo para la empresa, desde brindar acceso a los empleados a los datos hasta lograr optimizar los costos de almacenamiento. Con esta solución, Zalando logró obtener buenas prácticas para administrar una empresa basada en datos de varios *petabytes*.

La creación de un Data Lake en Amazon S3 ha permitido a los empleados de toda la organización acceder sobre datos a los que anteriormente no habrían tenido acceso. Se mejoró la experiencia de los clientes, cuando compran.

5 Capítulo 5 Propuesta de solución

Se recomienda proporcionar una solución oportuna para la empresa Zuteka, fundamentada en aspectos claves que combinan, almacenamiento, integración de los datos, seguridad y una acelerada búsqueda de ellos.

Los datos que se utilizaron en esta propuesta son datos demográficos de doctores en USA. Esos datos se pueden usar de entrada para realizar diferentes análisis como calcular la red óptima para una aseguradora de salud.

5.1 Técnicas para análisis de información

En este proyecto se utilizó la técnica de análisis de información, basado en 4 etapas con el fin de lograr la realización de este. (ICAV)

5.1.1 Identificar.

La empresa cada vez que inicia un nuevo proyecto con sus clientes, recibe nuevos datos para crear el Data Warehouse. Al iniciar un nuevo proyecto, se establecen los entregables desde el inicio, por lo tanto, la empresa solo recibe y carga únicamente los datos que están dentro de los entregables. El cliente normalmente envía más datos de los que se apegan a los entregables, pero la empresa no cuenta con el almacenamiento disponible para guardarlos todos. Cuando pasa este problema y el cliente solicita una nueva entrega y no se encuentran almacenados esos datos, la empresa debe estimar un tiempo adicional para ese nuevo entregable porque tiene que limpiarlos y almacenarnos en el Data Warehouse para que estén disponibles.

Se propone la creación del Data Lake para resolver el problema que enfrentan anteriormente mencionado. No tendrán más restricciones en el tamaño de los datos, almacenar más

de ellos, no agregará más tiempo en las entregas de sus clientes y podrán almacenar todos los que envía el cliente desde el principio sin tener que realizar una limpieza.

5.1.2 Consolidar.

Con cada cliente, los datos necesarios son diferentes porque no siempre se realizan los mismos análisis. El cliente le envía a la empresa los datos de diferentes fuentes, Zuteka solo es capaz de almacenar datos de una sola fuente, porque almacenarlos de varias fuentes conlleva mucho trabajo manual; incluso consideran que en una fuente vienen datos importantes y en otras, los restantes, entonces deben identificar cuáles son las fuentes donde están los datos requeridos y hacer una unión de estas, para finalmente almacenarlos.

Al proponerle a la empresa la creación del Data Lake, se logra eliminar ese trabajo manual, se podrá almacenar datos de diferentes fuentes y centralizar todo, sin trabajo manual.

5.1.3 Analizar.

No todos los usuarios tienen interacción con los datos y no es tarea fácil acceder a ellos, hay una cantidad limitada de usuarios para ingresar a consultarlos. Para solucionar esto, en la presente propuesta se decidió almacenar los datos en la nube, porque esto resuelve el tema de seguridad y la interacción de datos entre usuarios. Ahora estos pueden acceder a ciertos objetos donde tienen acceso y realizar una acelerada y sencilla búsqueda de datos, sin problemas en el Data Lake centralizado.

5.1.4 Visualizar.

Actualmente en la empresa, muchos usuarios no tienen el conocimiento técnico para acceder a los datos, porque se requiere saber el lenguaje SQL, Hive o PySpark. En muchas ocasiones, ellos andaban buscando una respuesta fácil de responder y deben esperar que algún

usuario técnico los ayude. Para resolver este problema, en esta propuesta se recomienda el uso de QuickSight para ver la visualización de los datos sin tener conocimiento técnico. El proponer usar QuickSight también ayuda a Zuteka a encontrar más valor a los datos, porque más usuarios tienen interacción sobre estos.

5.2 Prerrequisitos para la implementación de esta propuesta

Para lograr la implementación de esta propuesta, es necesario tener una alineación muy fuerte con el negocio. El Data Lake necesita proporcionar el valor que el negocio no está recibiendo de sus fuentes actuales.

Se debe crear un plan de todas las tareas que conlleva la creación del Data Lake, cada tarea debe tener un dueño y cada tarea debe tener una fecha de finalización. Se debe priorizar las tareas, es importante realizar llamadas de *check-in*, mínimo una vez a la semana para ver el progreso de la tarea asignada y discutir si se necesita ayuda para poder lograrla. Se debe asegurar que todo el equipo tenga una cuenta en AWS.

Como requisitos técnicos, el equipo involucrado debe tener una *laptop* para realizar la implementación, en conocimientos técnicos es necesario dominar aspectos relacionados con los datos, cómo manipularlos, Big Data, poseer preparación general sobre Cloud y SQL. No es necesario conocer de Data Lake en esta propuesta, los involucrados aprenderán qué es.

5.3 El desarrollo e implementación de la solución

5.3.1 ¿Cuáles son los beneficios que ofrece al negocio esta propuesta?

La empresa podrá almacenar volumen de datos de forma rápida, segura y de diferentes fuentes, sin restricción de su tipo (estructurados y no estructurados) o volúmenes de estos y, por lo tanto, va a lograr reducir la fase de almacenamiento de los datos con cada uno de sus clientes, no necesita dedicar mucho tiempo en limpieza, o filtrar información porque podrá almacenar los datos en su estado original y acceder a ellos en cualquier momento, aunque el objetivo inicial de sus proyectos cambie. Con ello, se van a reducir muchas tareas manuales, después de crear un proyecto, el próximo podrá reutilizar los pasos que siguió para almacenar los datos y así automatizar esta fase.

El acceso a los datos será más ágil, no es necesario que todos los usuarios tengan conocimiento técnico para poder analizarlos, van a tener la oportunidad de poder visualizarlos y así van a lograr darle más valor a estos para tomar mejores decisiones.

Asimismo, se ahorrarán costos de almacenamiento, solo se va a pagar por lo que usa, logra eliminar todo el equipo físico que usa para almacenar datos y reducirá el equipo que le da soporte a al hardware donde se almacenan los datos.

La forma de asignar permisos será más fácil, pues logrará reducir tiempo en la asignación de permisos a los usuarios. Además, permitirá a todos los empleados de la empresa acceder sobre los datos a los que anteriormente no tenía acceso. También, se mejorará la experiencia de los clientes, cuando reciben el análisis final.

5.3.2 ¿Por qué se propone la creación del Data Lake en AWS Lake Formation?

Este es un servicio que facilita la configuración de un Data Lake seguro en cuestión de días. En la actualidad, la configuración y la administración de un Data Lake implica muchas tareas manuales y complejas que llevan mucho tiempo. AWS Lake Formation simplifica y automatiza muchos de los pasos manuales complejos que suelen ser necesarios para crear el Data Lake. Estos pasos incluyen la recopilación, limpieza, traslado y catalogación de datos, y hacer que esos datos estén disponibles de forma segura para análisis y aprendizaje automático. Lake Formation proporciona su propio modelo de permisos el: AWS Identity and Access Management (IAM). Debido a la facilidad y ahorro de tiempo que ofrece AWS Lake Formation se decidió crear el Data Lake ahí. Se tomó la decisión de no analizar otro proveedor porque AWS Lake Formation, ofrece todo lo que se necesita para esta propuesta.

5.3.3 Creación del Data Lake.

Zuteka es una empresa consultora que tiene su ambiente de IT corriendo *on-premises*. Al mismo tiempo que aumenta el volumen de los datos, también aumenta la cantidad de equipos que necesitan agregar. El primer reto que se asumió en esta propuesta consistió en buscar la opción adecuada para el almacenamiento de los datos. Considerando las necesidades del negocio en la actualidad y el futuro, la decisión más lógica para la empresa era mudarse a Cloud.

¿Selecciones de proveedores en Cloud?

Los 2 principales proveedores de servicios en Cloud son (Forbes, 2021), Microsoft y Amazon S3. En esta propuesta se decidió analizar solo estos 2 proveedores, basados en ese análisis, se seleccionó 1 proveedor.

A continuación, se muestra una tabla comparativa entre AWS S3 y Microsoft Azure:

Tabla 5.1. Tabla comparativa entre AWS S3 y Microsoft Azure

	AWS S3	Microsoft Azure
Ventajas	• Estrategia de precios competitivos y flexibles. Se paga únicamente por lo que se usa. • Amazon puede ser ideal para colocar grandes bases de datos en la nube • AWS es el más progresivo en análisis de datos • La implementación es sencilla • Tiene muchas funciones y configuraciones • Más de 200 servicios • El proveedor de la nube más maduro • Domina el market share • La seguridad de AWS la proporciona mediante roles definidos con función de control de permisos.	• Microsoft es ideal para subir aplicaciones a la nube • Más de 100 servicios • La implementación es sencilla • Es excelente para aplicaciones en Windows, Active Directory y System Center. • Excelente para el servicio de nube hibrida. • Cada vez más abierto a las tecnologías de código abierto

Desventajas	• El modelo de precios es un poco complicado • Enfoque en la nube publica	• Los costos son más agresivos que AWS • Carece de madurez de soporte técnico en comparación con AWS • Proporciona seguridad al ofrecer permisos en toda la cuenta
Clientes que lo utilizan	• Netflix, Airbnb, Unilever, BMW, Samsung, Facebook, etc.	• Polycom, HP, Honeywell, Apple, eBay, Adobe, etc.

Fuente: Elaboración propia

Se propone seleccionar Amazon S3 como proveedor en la nube debido a la durabilidad, escalabilidad, tiene costos más económicos, se paga únicamente por lo que se usa, es experto en el análisis de datos, es ideal para almacenar grandes cantidades de datos y el proveedor de la nube más maduro en el mercado. También se tomó en cuenta la cantidad de servicios que ofrece, Zuteka podría aprovechar esto en el futuro.

Se descarta Microsoft porque los costos son más elevados y la forma de cómo se manejan los permisos de seguridad.

A continuación, se muestra con números y datos reales en detalle, porque se descarta Microsoft Azure en esta propuesta:

- Comparación de precios: Amazon es más accesible que Microsoft con un 28%, almacenando cualquier cantidad de datos

Tabla 5.2. Comparación de precios entre Amazon S3 y Microsoft Azure

Servicio	Almacenaje	Costo por mes
Amazon S3	50 TB	$0.023 por-GB
	450 TB	$0.022 por GB
	Más de 500 TB	$0.021 por GB
Microsoft	Más de 500 TB	$0.18 por-GB

- Seguridad:

En AWS, la seguridad se proporciona mediante roles definidos con función de control de permisos y Azure proporciona seguridad al ofrecer permisos en toda la cuenta. Por lo tanto, la seguridad es más fácil y sencilla de configurarla en AWS.

¿Por qué cambiar la infraestructura a Cloud?

La mayoría de las empresas prefieren los centros de datos físicos, los cuales requieren una inversión significativa en equipos de hardware, software, instalaciones y un personal capacitado

para la instalación y mantenimiento de todo el Data Center. Esto representa un impedimento para los proyectos que no cuentan con grandes presupuestos; sin embargo, existen grandes compañías que ya están apostando por las tecnológicas de virtualización para el procesamiento de datos a gran escala. Las tecnologías Cloud ha hecho accesible disponer de la información en cualquier momento y desde cualquier lugar, pues no reside en una única máquina. Se paga por lo que realmente se usa, la infraestructura se encuentra bajo demanda.

La computación en la nube o "Cloud Computing" viene a ocupar un importante lugar en la actualidad. Esto ha traído importantes ventajas, como por ejemplo la posibilidad de hacer despegar bajo demanda complejos servicios que antes necesitaban días o incluso semanas en estar listos.

Zuteka tiene limitaciones para almacenar todos los datos de sus clientes, al mismo tiempo que aumenta la cantidad de almacenamiento, también aumenta la cantidad de equipos requeridos. Se le propone a la empresa moverse a Cloud para eliminar la restricción de almacenamiento de datos.

Al proponerle a Zuteka, cambiar la infraestructura tecnológica (moverse a Cloud) tuvo un impacto directo en el panorama de datos de su empresa. Las bases de datos centrales a las que acceden muchos componentes estaban descentralizadas. El almacén central de datos de la empresa, con conexiones directas a los almacenes de datos transaccionales tuvo que hacer frente a una producción de datos descentralizados sin accesibilidad directa. A continuación, se enumeran los incentivos que se tomaron en cuenta para la creación del Data Lake en Zuteka.

- Almacenar volúmenes de datos de diferentes fuentes sin importar el tipo de datos (estructurados y no estructurados).
- Almacenar los datos en su estado original.

- Conserva todos los datos.
- Acceder a los datos en tiempo real.
- Útil para todo tipo de usuario, no es necesario tener conocimiento técnico para realizar consultas en los datos.
- Los datos se preparan "según sea necesario", lo que reduce los costos de preparación sobre el procesamiento inicial.

A continuación, se muestra el diagrama de arquitectura que se propone para la implementación:

Ilustración 5.1. Diagrama de arquitectura

Fuente: Elaboración propia y AWS

A continuación, se explica cada elemento del diagrama de arquitectura:

- Relational DB Instance, NoSQL DB Instance: se pueden tener diferentes fuentes de datos y diferentes tipos de datos.

- Amazon S3: almacena los datos en la nube y los habilita en AWS Lake Formation.
- AWS Lake Formation: crea el repositorio centralizado seguro, almacena los datos en forma original, cataloga los datos, aplica las políticas de seguridad para lograr el intercambio y acceso a los datos.
- Amazon Athena: permite el acceso a los datos para realizar diferentes análisis, pero se debe tener conocimiento del lenguaje SQL.
- QuickSight: permite la visualización de los datos sin conocimientos técnicos.

Lo siguiente es un resumen de los pasos que se debe seguir del diagrama de arquitectura:

1. Identificar las fuentes de datos (RDBMS, archivos planos, transacciones.)
2. Almacenar los datos en Amazon S3.
3. Rastrear los datos de Amazon S3 al Data Lake de AWS Lake Formation.
4. Catalogar los datos.
5. Crear usuarios y roles asignando políticas de seguridad.
6. Acceder los datos por medio de Athena.
7. Visualizar los datos en QuickSight.

Con esta propuesta, se le da a la empresa una serie de beneficios, tales como la eliminación de restricciones de tamaño de datos, la posibilidad de utilizar diferentes fuentes; además, le ahorra el tiempo de creación de un Data Warehouse. Asimismo, puede darse cuenta de que la mayor parte

de sus datos son potencialmente valiosos sin saberlo antes y ahorrar costo en almacenamiento porque ahora solo paga por lo que utiliza.

A continuación, se detallarán los complementos que se deben utilizar en esta propuesta, los cuales son de importancia para el proyecto, estos serán enlistados con su nombre original que en su mayoría están en inglés.

5.3.4 Amazon S3.

En esta propuesta, el servicio se encarga de almacenar los datos en la nube y da la opción de habilitarlos para el Data Lake. Se eligió, dada la facilidad y la libertad que ofrece en el almacenamiento de datos, ya que brinda la opción de utilizar diferentes fuentes y no tiene restricciones de tamaño de los datos. Al almacenar los datos en la nube, la empresa logra ahorrar costos, almacena todo el material y le da el valor que antes no le daba.

La empresa consideraba que los datos significaban un costo más porque no era tarea fácil lograr almacenarlos, pero con S3, logra entender que son de gran utilidad para el negocio y son un gran activo.

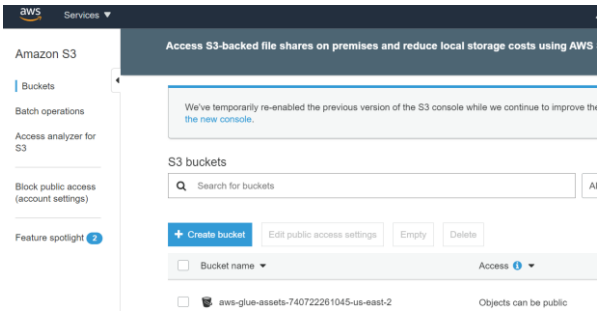
Está claro que Amazon S3 ofrece un gran modelo de precios, pero en un futuro se va convertir en un objetivo de optimización de costos porque se puede tener guardado datos de *petabytes* y se va convertir en una parte importante en la factura de la nube.

Por eso se recomienda utilizar 0, porque permite comprender qué datos se poseen realmente; es imposible comprender el almacenamiento solo a partir de la intervención humana. S3 Invetory permite ver la información de cada objeto almacenado en Amazon S3, cosas simples como el tamaño de los objetos, la última vez que se accedió y así puede ver la antigüedad de sus datos y decidir si es necesario realizar una limpieza.

Gracias a S3 Invetory, se logra identificar cuáles son los datos que no se usan continuamente y así tomar la decisión de realizar operaciones de limpieza y ahorro de costos impactantes.

En la siguiente imagen, se puede observar cómo se visualiza Amazon S3

Ilustración 5.2. visualización de Amazon S3



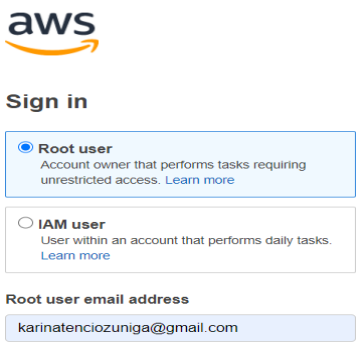
Fuente: AWS

5.3.5 AWS Lake Formation.

Se decidió la creación del Data Lake en AWS Lake Formation, porque es un servicio que facilita la configuración del Data Lake seguro en días. Se obtiene un repositorio centralizado seguro, que almacena sus datos en su forma original y fácil de catalogar los datos, aplica políticas de seguridad y logra el intercambio y acceso a los datos para sus diferentes usuarios.

En la siguiente imagen, se puede observar una captura de pantalla de la ventana de ingreso

Ilustración 5.3. Ventana de Inicio de Sesión



Fuente: AWS

5.3.6 AWS Glue.

En esta propuesta se tomó la decisión de utilizar AWS Glue para catalogar los datos, aparte de ser un servicio que ofrece AWS Lake Formation, ayuda a automatizar la ardua tarea de la preparación de datos para los análisis. AWS Glue ayuda a catalogar los datos automáticamente. Fue fácil la utilización después de haber creado el Data Lake, muestra beneficios como reducir el trabajo manual de sus empleados.

5.3.7 Security.

El acceso a los datos en cuanto a seguridad e intercambio de ellos es un problema que enfrenta la empresa, temía a la seguridad, pero también se dio cuenta que no tenía sentido almacenar grandes cantidades de datos si nadie los estaba usando, porque se está perdiendo el valor de los datos.

Después de la creación del Data Lake en AWS Lake Formation, la mayoría de los usuarios ya tienen sus propias cuentas de AWS que inmediatamente establecen requisitos para el intercambio de datos entre cuentas. AWS Lake Formation proporciona un modelo de permisos que se basan en un mecanismo simple de GRANT/REVOKE. Los permisos de Lake Formation se

combinan con los de AWS Identity and Access Management (IAM) para controlar el acceso a los datos almacenados en el Data Lake y a los metadatos que los describen.

Se recomendó usar IAM y tener un usuario root que tiene permisos ilimitados y este usuario crea otros usuarios administradores que se les asigna políticas de seguridad para tener acceso a ciertos objetos del Data Lake, a AWS Glue y Athena

Esta solución le dio varios beneficios a la empresa, como intercambio de los datos de forma segura, encontrar más valor a estos, porque más usuarios pueden accederlos y es un mecanismo sencillo y confiable de utilizar.

5.3.8 Athena.

Se le propone a la empresa, utilizar Athena para tener acceso a sus datos y lograr la realización de los algoritmos para sus clientes. Es una herramienta que les permite a sus usuarios una acelerada búsqueda de los datos, no se pega en el momento de la consulta. Athena simplemente lee los datos almacenados en el Data Lake. Un beneficio importante es que los usuarios pueden ver la estructura de las tablas, sin ver los datos, para la empresa es algo positivo porque a veces se trabaja con clientes que no permiten que trabajen desde la ubicación donde están, pero para la empresa son personas importantes en el proyecto por lo que pueden ir haciendo el código del análisis del cliente sin tener acceso a los datos.

5.3.9 QuickSight.

La compañía tenía problemas para darle valor a sus datos para tomar mejores decisiones. En esta propuesta se propone la utilización de QuickSight. La solución fue sencilla de aplicar, los datos ya están disponibles para accederlos en Athena solo era realizar una conexión de Athena a

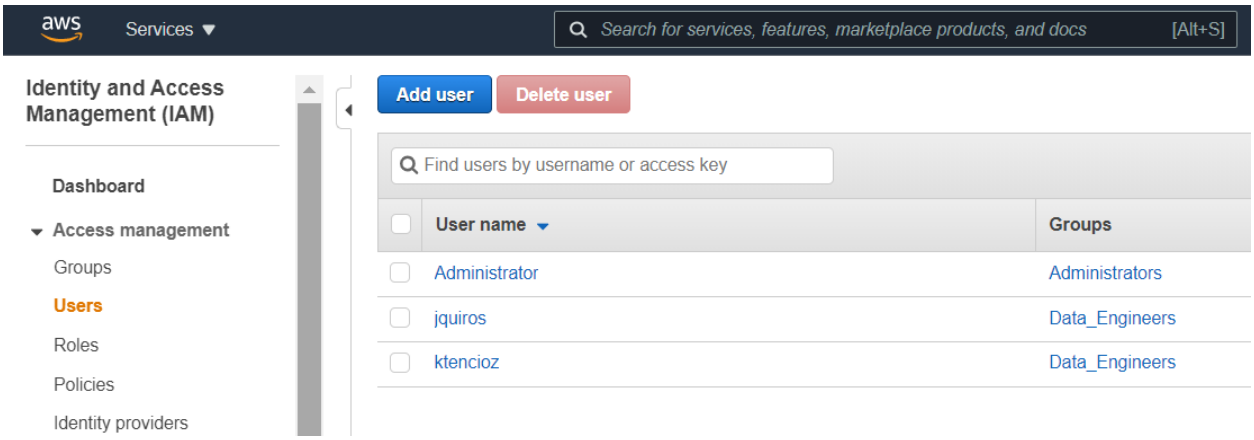
QuickSight para visualizar los datos. En los usuarios que ya están creados, se les asignó una política de seguridad, para que puedan conectarse en QuickSight. Teniendo los datos disponibles en QuickSight, le permite a Zuteka realizar diferentes análisis, pero de forma visual para así lograr obtener más valor a los datos.

Después de haber implementado QuickSight, se logra encontrar una solución que no se había pensado para la empresa, muchos usuarios no tienen el conocimiento técnico para acceder los datos por medio de Athena porque no conocen el lenguaje SQL, entonces QuickSight viene a solucionar ese problema para los usuarios que no tienen conocimiento técnico. Además de ayudar a encontrar valor a los datos, también da la opción a usuarios no técnicos que puedan acceder a estos. En esta propuesta se propuso utilizar QuickSight para la visualización, pero esta no es la única opción disponible, existen varias herramientas en el mercado que funcionan correctamente para visualización y también tienen la función de recibir una tabla externa o un CSV, como input para visualizar los datos como Tableau y Domo.

5.4 Pasos que se siguieron del diagrama de arquitectura para lograr la implementación de esta propuesta

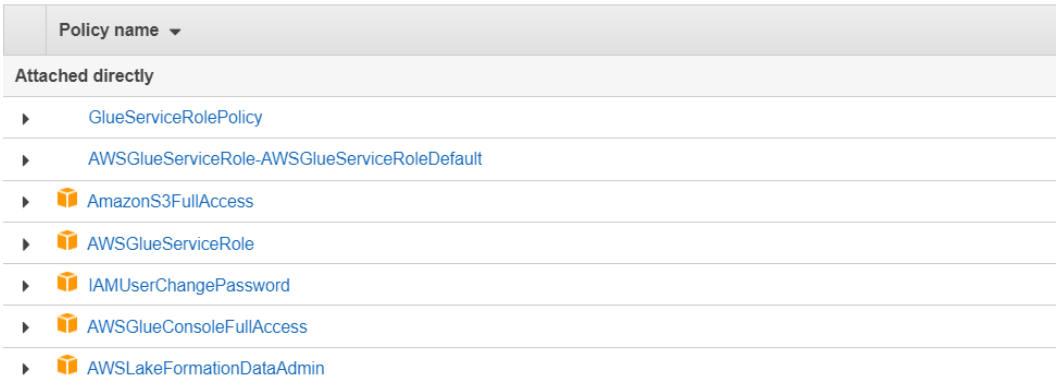
Es necesario tener un usuario administrador aparte del usuario root, con la política de seguridad: “AWSLakeFormationDataAdmin” para poder realizar la implementación y ese usuario es el responsable de crear los usuarios, roles y políticas para que otros usuarios puedan acceder al Data Lake.

Ilustración 5.4. Ventana de IAM Dashboard



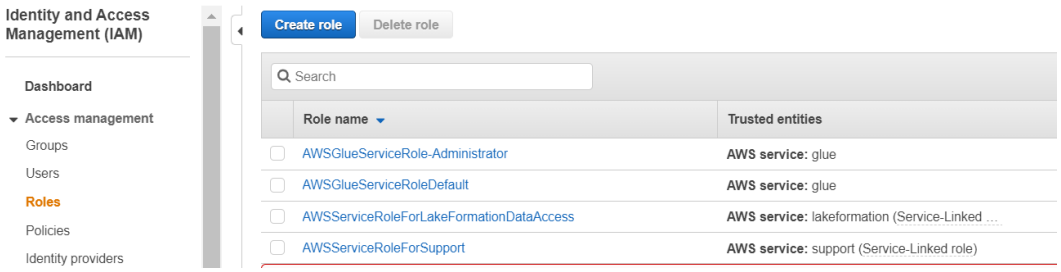
Fuente: AWS

Ilustración 5.5. Ventana de las políticas de seguridad asignadas al usuario



Fuente: AWS

Ilustración 5.6. Ventana de los roles creados

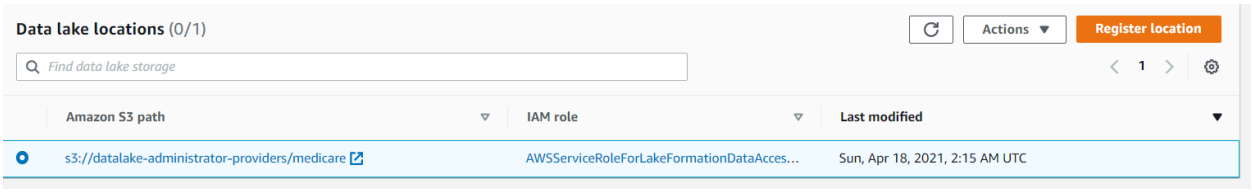


Fuente: AWS

Identificar las fuentes de datos: para este caso de uso, se utilizaron los datos de medicare enrollment, estos son datos demográficos de los doctores de USA, con estos se puede ayudar a las aseguradoras de salud en Estados Unidos a crear una red óptima de doctores. Estos serían la entrada para ese análisis, también se pueden usar para identificar el tipo de especialidad de cada doctor, si no se tienen los claims. Dichos datos son públicos y son actualizados 2 veces al mes.

Se debe almacenar los datos en Amazon S3, rastrearlos al Data Lake y utilizar AWS Glue Crawler para catalogarlos y crear la metadata de la base de datos.

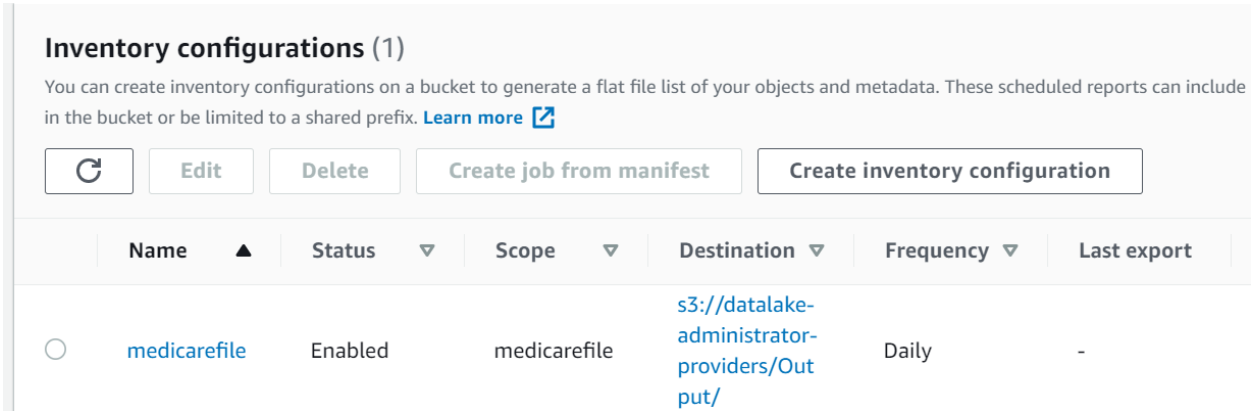
Ilustración 5.7. Ventana donde se muestra la ubicación de los datos



Fuente: AWS

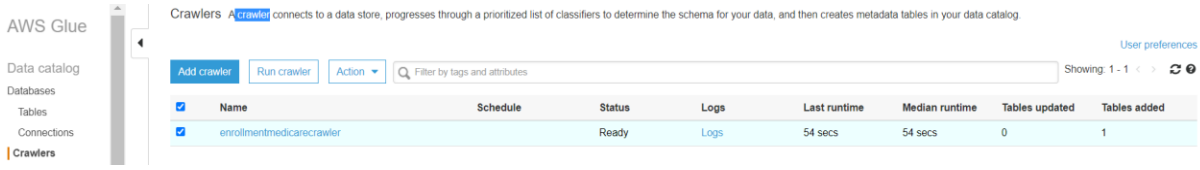
Es importante activar Amazon S3 Inventory para verificar si realmente los datos almacenados, se están usando de lo contrario archivarlos para no pagar almacenaje por algo que no se usa. El reporte se programa diario o semanal.

Ilustración 5.8. Ventana donde se muestra que se configuro Inventory.



Fuente: AWS

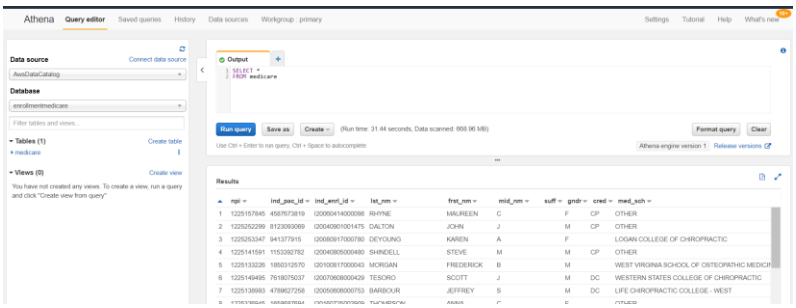
Ilustración 5.9. Ventana donde se muestra la creación de la base de datos y la tabla por medio de AWS Glue Crawlers



Fuente: AWS

Finalmente, los datos se pueden acceder por medio de Athena utilizando el código de SQL.

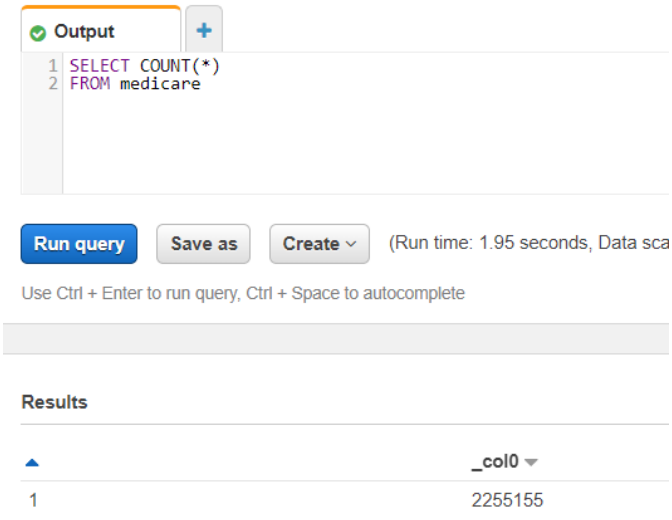
Ilustración 5.10. Ventana del resultado de la consulta en SQL por medio de Athena



Fuente: AWS

Se puede observar que se almacenaron más de 2 millones de datos, no se tuvo ningún problema para cargarlos y accederlos.

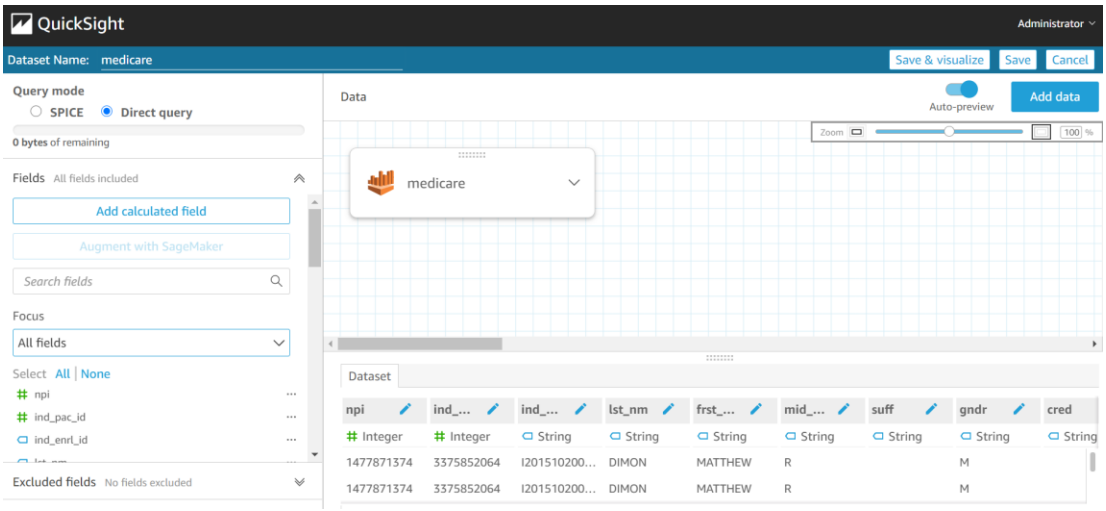
Ilustración 5.11 Cantidad de datos cargados



Fuente: AWS

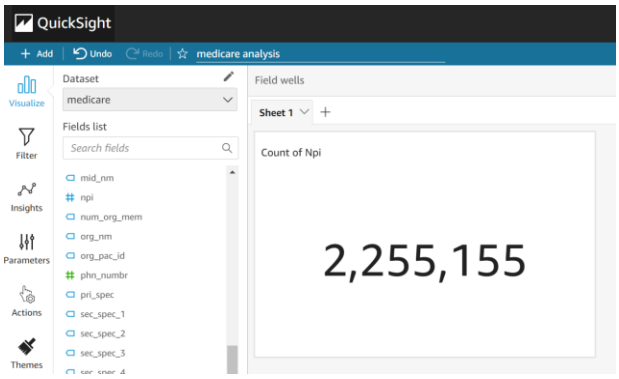
Los usuarios, que no son técnicos, utilizan QuickSight para acceder a los datos. Los usuarios administradores deben realizar la conexión a Amazon S3 o Athena para acceder a los datos y así dejar la conexión lista para los usuarios no técnicos. Esos usuarios no técnicos van a poder explorar los datos, realizar análisis y dashboard.

Ilustración 5.12 Ventana de visualización de datos por medio de QuickSight



Fuente: AWS

Ilustración 5.13 Ventana de visualización de datos por medio de QuickSight. Los usuarios igual pueden realizar consultas con funciones de agregación.



Fuente: AWS

5.5 Evaluación de la propuesta

Para la implementación de esta propuesta, se debe tener en cuenta que presenta beneficios positivos, negativos y no se puede implementar en cualquier escenario, a continuación, se explican esos beneficios y mencionar en cuáles escenarios se puede implementar.

Esta propuesta se puede implementar en escenarios donde el almacenar datos, acceso a los datos, seguridad a los datos, no encontrar valor a los datos, es un problema. Se mencionan escenarios específicos donde se puede implementar esta propuesta:

- Cuando los datos se distribuyen entre varias bases de datos desconectadas, se dificulta su análisis y se restringe el acceso a estos. Para solucionar este problema, se puede crear un Data Lake en AWS Lake Formation, esto permite agregar datos de diferentes fuentes a Amazon S3 donde se va catalogar y asignar políticas de seguridad por medio de AWS Lake Formation.

- Crear una plataforma de datos con la capacidad de administrar la seguridad para todas las diferentes aplicaciones en el entorno. Con AWS Lake Formation, se define las políticas.

- Falta de datos confiables cuando se realiza análisis de datos provenientes de múltiples fuentes. La limpieza de datos es un paso crítico en el análisis de datos y puede tener un gran impacto en los resultados comerciales y la toma de decisiones. Las funciones de AWS Lake Formation permite abordar el desafío de la veracidad de los datos y asegurar el acceso al a los datos.

Al implementar esta propuesta, se encuentra los siguientes aspectos positivos:

- No hay necesidad de descartar datos.
- Proporciona un punto de control centrar para cargar, limpiar, proteger y catalogar fácilmente los datos de miles de clientes en el Data Lake.
- Automatiza mucho trabajo manual.
- Es fácil agregar nuevos datos y permite cargar datos de diferentes fuentes
- Es útil para todo tipo de usuario.
- Se lograr encontrar más valor en los datos.
- Se paga únicamente lo que se utiliza, va reducir costos.

Se recomienda implementar esta propuesta, cuando se tienen grandes cantidades de datos, es requisito tomar la decisión de migrarse a Cloud para implementarla y es requisito activar S3 Invetory para evitar tener datos almacenados, si realmente no se están utilizando.

Uno de los aspectos más importantes a la hora de realizar una propuesta, es identificar y comprender el problema que se desea resolver.

Al lograr implementar esta propuesta, se logra resolver problemas existentes y se identifica un nuevo beneficio para los usuarios no técnicos, van a poder acceder a los datos sin tener conocimiento en SQL.

El proponer almacenar los datos en Amazon S3, les ha permitido a los usuarios de toda la empresa acceder sobre datos a los que antes no tenían acceso. La empresa logró reducir el costo de almacenamiento mediante el uso de varios servicios de Amazon S3.

6 Conclusiones y recomendaciones

El objetivo principal de este proyecto era “Diseñar un Data Lake en AWS Lake Formation para mejorar el almacenamiento y el acceso de los datos”, que al ser completado pudiera resultar clave para el manejo de datos.

A continuación, se detallarán las conclusiones a las que se llegaron después de cumplir con los objetivos que se plantearon al inicio del proyecto.

Seguidamente, se mencionan las recomendaciones consideradas útiles para mejorar el proceso investigativo realizado.

6.1 Conclusiones

1. Actualmente los datos se generan de diferentes formatos, es imposible encontrar una sola fuente de estos y si es así, la información no estará completa. Por lo tanto, evaluar las fuentes de datos y que la empresa deba elegir por una, no debería de ser una opción, se debe tener la oportunidad de recibir diferentes fuentes de datos y almacenarlos. Cuanto más completos están, mejor van a ser los resultados.

2. Los datos están creciendo constantemente y son un gran activo para las empresas, pero hoy en día muchas siguen almacenando los datos *on-premises*, al mismo tiempo que aumenta su volumen, también aumenta la cantidad de equipos requeridos. Para solucionar este problema, se concluye que la mejor opción es almacenar los datos en la nube. Solo se pagará por el uso de almacenamiento y esto tendrá como resultado un ahorro en su costo. Además, todos los usuarios van a poder acceder a los datos sin importar el lugar donde se encuentren.

3. El manejo de seguridad de los datos hacia los diferentes usuarios de la empresa es importante, pero también hay que tener claro que no tiene sentido almacenar muchos datos, si no hay interacción sobre estos; por lo tanto, aplicar políticas de seguridad es principal pero las empresas no deben complicarse con este proceso, hay muchos servicios que facilitan esto, se requiere solo de la asignación de políticas de seguridad a diferentes usuarios o grupos para tener acceso a los mismos y con este mecanismos garantizan la seguridad a sus datos e interacción de estos hacia la población.

4. En la actualidad, las empresas tienen usuarios técnicos y no técnicos, es importante que ambos usuarios logren acceder a los datos sin ninguna limitante de conocimiento, por lo cual se debe tener solo una herramienta para acceder los datos y se necesita conocimiento técnico aunque sea mínimo; si no es suficiente, se debe tener mínimo 2 opciones que se adapten a todos los usuarios para lograr encontrar más valor a los datos. Con ello el trabajo será más eficiente porque no hay dependencia entre sí, hay que tener claro que ciertos análisis requieren de conocimiento técnico pero para realizar búsquedas.

5. Solo tener acceso a los datos no es suficiente, es importante tener una alternativa para visualizarlos, eso ayudará a encontrar más valor a lo almacenado.

6. Hoy en día, muchas empresas realizan mucho trabajo manual, porque no tienen sus datos centralizados, o solo una persona lleva la última versión y si esta no tiene disponibilidad, no se puede acceder a los datos porque están guardados de forma local; por lo tanto, la creación del Data Lake logra resolver este problema y eliminará errores humanos.

7. Para crear un Data Lake, se debe manejar grandes cantidades de datos si no es así, no tiene mucho sentido su creación.

8. En la actualidad hay muchas empresas que manejan los datos sin caer en soluciones de Big Data, enfrentan problemas de rendimiento, el uso de datos no estructurados, redundancia de estos y mucho trabajo manual, por eso es importante analizar y utilizar soluciones de Big Data.

9. La creación del Data Lake en AWS Lake Formation facilita la configuración en cuestión de días y es seguro. AWS Lake Formation simplifica y automatiza muchos de los pasos manuales complejos que suelen ser necesarios para crear el Data Lake.

10. En la actualidad hay muchos proveedores en la nube. Realizando una comparación entre los 2 principales proveedores de la nube según Forbes: Amazon S3 y Microsoft Azure, se decide elegir Amazon S3, por diferentes razones como las siguientes: ofrece precios más flexibles, es el más progresivo en análisis de datos, la implementación es sencilla y es el proveedor de la nube más maduro, etc. Pero esto no quiere decir que siempre se deba elegir Amazon S3, eso depende del problema por resolver y las diferentes necesidades de la empresa.

11. AWS Lake Formation respalda la importante iniciativa de dedicar menos tiempo a administrar datos para que las organizaciones puedan dedicar más tiempo en obtener información valiosa de los datos y así tomar mejores decisiones.

12. Hay muchas herramientas de visualización de datos en el mercado que se adaptan a cualquier necesidad como Tableau y Domo, no es necesario utilizar QuickSight para visualización.

13. Es importante tener principios de administración de datos, son mandatorios y no son negociables para que la empresa, que se dedica a análisis, sea exitosa.

6.2 Recomendaciones

Para alcanzar el éxito en este tipo de proyectos, es necesario tener una alineación muy fuerte con el negocio; de lo contrario, no se va a tener éxito, porque el Data Lake debe proporcionar el valor que el negocio no está recibiendo de sus fuentes actuales.

Se recomienda la creación de un Data Lake, cuando se requiere tener todos los datos centralizados y se manejan grandes volúmenes de datos, si la cantidad de datos que se utilizan es pequeña no tiene sentido crear un Data Lake.

Se recomienda utilizar AWS Lake Formation para la creación del Data Lake, debido a que la configuración se realiza en cuestión de días, es segura, simplifica y automatiza muchos de los pasos manuales complejos que suelen ser necesarios para crear el Data Lake y al elegir AWS no impone servicios que realmente no se necesitan para implementación del proyecto.

Es importante tomar en cuenta todo tipo de usuarios para tener a disponibilidad varias herramientas para acceder los datos. Se recomienda una herramienta de visualización para los usuarios no técnicos, así ellos van a poder accederlos sin ningún requisito técnico. Se recomienda utilizar QuickSight para visualización, es fácil de usar y es parte de los servicios de AWS, pero no es la única disponible en el mercado, hay muchas como Tableau y Domo y también son sencillas de usar. Cada proyecto debe elegir su herramienta de visualización, de acuerdo con sus necesidades y no preferir solo con una.

Se recomienda almacenar los datos en la nube porque ofrece resultados altamente positivos como pagar únicamente por utilizado y no tener que agregar equipo cada vez que se aumenta el almacenaje, pero el que una empresa siga almacenando en disco, no significa una opción mala.

Almacenar los datos en Amazon S3, ayuda a reducir costos, porque solo se paga por usado, pero es importante tomar en cuenta que si se guardan muchos datos y esos no están siendo usados por ninguna persona, se debe hacer limpieza, por lo cual se recomienda utilizar el Amazon S3 Inventory que indica si los datos son útiles, están siendo usados o no, para evitar un costo adicional a la factura.

Se recomienda manejar el acceso a los datos de los usuarios, por medio de políticas sencillas y seguras.

7 Reflexiones finales

No hay una respuesta definitiva que indique cuál es la tecnología correcta por utilizar, se debe tener claro que cada proveedor es distinto y cada uno tiene sus ventajas y desventajas; el proveedor por elegir, debe estar muy alineado con el negocio y la lógica que se quiere implementar para ser seleccionado.

Es importante tener conocimientos técnicos para la implementación del proyecto, pero no es requisito ser experto en el tema de Data Lakes; asimismo, estudiando y realizando una investigación se logran implementar modificaciones adecuadas. No obstante, debe tenerse cuidado, cuando se está haciendo la implementación porque la factura puede llegar alta, al no saber utilizar bien los servicios seleccionados.

Cada vez que existe la opción de eliminar trabajo manual, hay que buscar una solución y aplicarla, tal y como se logró en esta propuesta.

Es bueno tener datos almacenados, pero no tiene sentido almacenarlos, si no hay interacción de datos, no se va a encontrar valor a los datos si no hay usuarios que puedan accederlos.

En definitiva, realizar este proyecto fue de gran interés para el equipo, por cuanto el Data Lake era un tema completamente nuevo para todos los interesados y ver todas las soluciones que se le dio a Zuteka fue de mucha satisfacción.

Bibliografía

(CMS), C. f. (2018). Obtenido de CMS: https://data.cms.gov/provider-data/dataset/mj5m-pzi6

(2019). Obtenido de Talend: https://www.talend.com/es/resources/what-is-data-lake/

Admin, A. (2019). Obtenido de AWS: https://aws.amazon.com/es/blogs/aws-spanish/como-extraer-valor-de-sus-datos-con-data-lakes-y-analisis-en-aws/

AWS. (2019). Obtenido de AWS: https://aws.amazon.com/lake-formation/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc*

AWS. (2019). Obtenido de AWS: https://aws.amazon.com/es/solutions/implementations/data-lake-solution/

AWS. (2019). Obtenido de AWS: https://docs.aws.amazon.com/AmazonS3/latest/userguide/acl-overview.html

Big Data, Consulting. (2012). *Big Data Consulting*. Obtenido de https://bigdataconsulting.wordpress.com/consulting/methodology/

Biolchini, J. (2017). *Systematic Review in Software Engineering.*

Fiz, J. M. (2010). *Think Big*. Obtenido de Think Big: https://www.paradigmadigital.com/dev/comparativa-servicios-cloud-aws-azure-gcp/

Forbes. (20 de March de 2021). Obtenido de Forbes: https://www.c-sharpcorner.com/article/top-10-cloud-service-providers/

Gartner. (2001). Obtenido de Gartner: https://www.gartner.com/en/information-technology/glossary/big-data

Google. (2020). Obtenido de Google Cloud: https://cloud.google.com/architecture/build-a-data-lake-on-gcp?hl=es9

Goswami, A. (2020). Obtenido de AWS: https://aws.amazon.com/blogs/big-data/enforce-column-level-authorization-with-amazon-quicksight-and-aws-lake-formation/

Heinrich, G. (2017). Obtenido de AWS: https://aws.amazon.com/blogs/big-data/build-a-data-lake-foundation-with-aws-glue-and-amazon-s3/

información, T. d. (2007). *Tecnologías de información*. Obtenido de https://www.tecnologias-informacion.com/data-lake.html

Latam, T. (2020). Obtenido de Tivit Latam: https://latam.tivit.com/blog/aws-o-microsoft-azure-cual-es-la-mejor-opcion-para-mi-empresa

Learning, I. (2020). Obtenido de Ingenio Learning: https://ingenio.edu.pe/aws-vs-azure-vs-google-cual-es-la-mejor-opcion/

Lingam, C. (2021). Obtenido de Udemy: https://www.udemy.com/course/data-lake-in-aws/?utm_source=adwords&utm_medium=udemyads&utm_campaign=LongTail_la.EN_cc.ROW&utm_content=deal4584&utm_term=_._ag_77879423894_._ad_437497333812_._kw__._de_c_._dm__._pl__._ti_dsa-1007766171032_._li_9068562_._pd__.

Mishra, G. K. (2020). Obtenido de AWS: https://aws.amazon.com/blogs/big-data/how-to-delete-user-data-in-an-aws-data-lake/

Piergiorgio,Corbetta. (2017). Metologia y Tecnicas de Investigacion Social Madrid. McGraw-Hill. Obtenido de https://diversidadlocal.files.wordpress.com/2012/09/metodologc3ada-y-tc3a9cnicas-de-investigacic3b3n-social-piergiorgio-corbetta.pdf

Samata, M. (2020). Obtenido de AWS: https://aws.amazon.com/blogs/big-data/anonymize-and-manage-data-in-your-data-lake-with-amazon-athena-and-aws-lake-formation/

Sasmal, A. K. (2019). Obtenido de AWS: https://aws.amazon.com/blogs/big-data/etl-and-elt-design-patterns-for-lake-house-architecture-using-amazon-redshift-part-1/

Service, S. A. (2018). Obtenido de Software Advisory Service: https://www.softwareadvisoryservice.com/es/blog/amazon-web-services-vs-microsoft-azure-qu%C3%A9-servicio-en-la-nube-es-mejor/

Services, A. W. (2016). *Data Lake Solution.* Obtenido de Data Lake Solution: https://s3.amazonaws.com/solutions-reference/data-lake-solution/latest/data-lake-solution-on-aws.pdf

Services, A. W. (2020). *Data Lakes and Analytics on AWS*. Obtenido de https://www.youtube.com/watch?v=hXXytLly7tw