Análisis de sentimientos del último modelo de iPhone 15

Sergio Boluda Fernandes

Dashboard Análisis de Sentimientos

TFG - Business Intelligence

Nombre Profesor Colaborador

Humberto Andrés Sanz

14 enero 2023

Universitat Oberta de Catalunya

FICHA DEL TRABAJO FINAL

Título del trabajo:	Dashboard Análisis de Sentimientos del iPhone 15
Nombre del autor:	Sergio Boluda Fernandes
Nombre del colaborador:	Humberto Andrés Sanz
Nombre del PRA:	Atanasi Daradoumis Haralabus
Fecha de entrega (mm/aaaa):	01/2024
Titulación o programa:	Grado de Ingeniería Informática
Área del Trabajo Final:	Business Intelligence
Idioma del trabajo:	Castellano
Palabras clave	Amazon Web Services, Serverless, Python, Dashboard, Sentiment.

Resumen del Trabajo

En la sociedad actual, las redes sociales se han convertido en una herramienta principal para medir la opinión pública sobre diversos temas. La consulta de reviews de productos en estas redes es la forma más habitual y confiable por los usuarios a la hora de adquirir un nuevo producto.

Apple es una de las compañías más influyentes en el mercado tecnológico que genera millones de publicaciones en redes sociales con cada lanzamiento de producto, especialmente el iPhone. Este trabajo busca analizar la opinión pública sobre el lanzamiento del iPhone 15 mediante un análisis de sentimientos de reviews de usuarios.

Se utilizará la infraestructura de Amazon Web Services (AWS) para manejar el volumen masivo de reviews, utilizando las ventajas de las soluciones serverless en términos de costos, rendimiento computacional y

sostenibilidad ambiental.

Se implementarán varios servicios de AWS, desde la recolección y almacenamiento de datos hasta su procesamiento y análisis, poniendo fin con un dashboard desarrollado en Amazon Quicksight.

Este Trabajo de Fin de Grado enfatiza la importancia de las redes sociales como herramientas de escucha y respuesta para las empresas en el mercado actual.

El objetivo es desarrollar un flujo de trabajo de análisis de datos para reviews relacionadas con el iPhone 15, utilizando servicios de AWS. Incluye desde la extracción de datos y su almacenamiento en diferentes formatos hasta el análisis de sentimientos con modelos preentrenados en Amazon Comprehend y la presentación de resultados en dashboards y reportes.

Abstract

Today, social networks have become a key tool for measuring public opinion on diverse topics. Consulting product reviews on these networks is the most common and reliable way for users to decide when buying a new product.

Apple, one of the most influential companies in the tech market, generates millions of social media posts with each product launch, especially the iPhone. This project aims to analyze public opinion on the launch of the iPhone 15 by examining user reviews' sentiments.

We will use Amazon Web Services (AWS) infrastructure to manage the massive volume of reviews, taking advantage of serverless solutions in terms of cost, computational performance, and environmental sustainability.

Several AWS services will be implemented, ranging from data collection and storage to processing and analysis, concluding with a dashboard developed in Amazon Quicksight.

This Final Degree Project highlights the importance of social networks as listening and response tools for companies in today's market.

The goal is to develop a data analysis workflow for reviews related to the iPhone 15, using AWS services. It includes everything from data extraction and storage in various formats to sentiment analysis with pre-trained models in Amazon Comprehend and the presentation of results in dashboards and reports.

Tabla de contenido

Tabla de Ilustraciones

1. Introducción

En la sociedad actual, las RRSS (Redes Sociales) se han utilizado como herramienta principal para medir la opinión pública sobre cualquier tema, desde acontecimientos mundiales hasta lanzamientos de productos, diversas redes sociales como YouTube, Instagram, Facebook, Twitter se utilizan para generar reviews sobre productos.

Dentro del mercado tecnológico, Apple se posiciona como una de las compañías más influyentes y sus lanzamientos anuales de productos, en especial el iPhone, se convierte en un evento global que generan millones de posts en las RRSS. Estas publicaciones que muestran una opinión del producto están cargadas de emociones y percepciones, que, además, ofrecen una información valiosa que, si se analiza adecuadamente, puede ofrecer resultados sobre la aceptación del producto, áreas de mejora y el pulso general del mercado.

La finalidad de este trabajo es adentrarse en el desafío de descifrar y comprender la opinión pública sobre el último lanzamiento del iPhone 15, a través de un análisis de sentimientos basado en reviews. Debido el volumen de texto analizado y la necesidad de un procesamiento eficiente y escalable, este estudio estará soportado sobre la infraestructura de Amazon Web Services (AWS).

La elección de AWS no es casual, va en línea con las tendencias tecnológicas actuales hacia soluciones serverless, las cuales no sólo proporcionan un ahorro de costes sino también optimizan el rendimiento computacional y son más responsables con el medio ambiente. Por ello, para la resolución de este proyecto, se desplegarán varios servicios de AWS, desde la recolección y almacenamiento de datos con la API de Twitter a través de funciones lambda y AWS S3, hasta el uso de herramientas serverless como Glue y Athena para el procesamiento (ETL) y análisis. Finalmente, se visualizarán los resultados a través de un dashboard desarrollado en Amazon Quicksight, proporcionando una representación gráfica y comprensible de los sentimientos y percepciones asociados al iPhone 15.

Más allá del análisis técnico y de datos, este TFG tiene como objetivo resaltar la importancia de las redes sociales como herramientas de escucha y respuesta para las empresas en el mercado actual.

1.1. Contexto y justificación del Trabajo

Desde hace unos años, las redes sociales juegan un papel crucial para los usuarios que desean obtener un nuevo producto, es imperativo para las empresas entender las opiniones, percepciones y reacciones de los usuarios hacia sus productos. En particular, Apple genera mucha controversia en cada uno de sus lanzamientos, esto genera un gran volumen de comentarios y reseñas en diversas plataformas. Estas reseñas son una necesidad crítica para

las empresas, ya que pueden analizar de manera eficiente y efectiva las impresiones de sus productos sobre sus usuarios para obtener Insights valiosos que puedan guiar las estrategias de marketing y desarrollo futuro de sus productos.

Este tema es especialmente relevante en el contexto de la ingeniería informática debido a la creciente importancia de los datos como herramienta para la toma de decisiones empresariales. Con el auge de proyectos relacionados con big data y el análisis de datos, las habilidades para recolectar, procesar y analizar grandes volúmenes de información se han vuelto indispensables.

La principal aportación de este trabajo comienza con el desarrollo de un flujo de trabajo analítico integral, empleando herramientas avanzadas de AWS como AWS Lambda, AWS Data Glue Databrew, AWS Comprehend, Athena y QuickSight. Este flujo de trabajo proporcionó una metodología robusta y escalable para el análisis de sentimientos y la extracción de insights a partir de grandes volúmenes de datos de texto. El resultado esperado es una comprensión profunda de la percepción de los usuarios analizados sobre el iPhone 15, que pueda ser utilizada para informar decisiones empresariales y estrategias de producto.

Este trabajo propuesto, no solo aborda una necesidad actual del mercado tecnológico cada vez más demandado, sino que también contribuye al campo de la ingeniería informática con una aplicación práctica y relevante de análisis de datos y herramientas de big data.

1.1. Objetivos del Trabajo

1.1.1. Objetivos inicialmente previstos

El objetivo del Trabajo de Fin de Grado es desarrollar un flujo de trabajo de análisis de datos para los tweets relacionados con el iPhone 15, utilizando servicios de AWS. Se iniciará con un Análisis Exploratorio de Datos (EDA) para evaluar la calidad de los datos para asegurar que obtendremos unos Insights valiosos.

Posteriormente, se implementarán funciones lambda para extraer tweets utilizando palabras clave definidas en el EDA realizado anteriormente Los datos recopilados se almacenarán en AWS S3 en dos formatos: RAW para datos brutos y TRANSFORMED para datos procesados y de mejor calidad, con la posibilidad de usar el formato parquet.

Se empleará AWS Glue para transformaciones de datos y mejoras en la calidad. El análisis de sentimiento se llevará a cabo con modelos preentrenados de Hugging Face, enfocados en el análisis de negocios en lugar de la generación de modelos de IA. Los resultados se almacenarán en DynamoDB y se utilizará Amazon QuickSight y Athena para crear dashboards y reportes finales.

1.1.2. Objetivos finalmente alcanzados

El objetivo final del TFG de Ingeniería Informática se centró en el desarrollo de un flujo de trabajo analítico sobre el procesamiento y análisis de datos relacionados con comentarios, reseñas o textos de usuarios sobre el iPhone 15, utilizando diversas herramientas de AWS.

Inicialmente, el proyecto se planteó para analizar tweets utilizando servicios de AWS que directamente consultaran a la API de Twitter, pero debido a restricciones de esta relacionados con la facturación, se optó por utilizar un conjunto de datos alternativo obtenido de Kaggle, que incluye reseñas de usuarios transcrita frase a frase en un fichero CSV.

Los pasos clave en el desarrollo del proyecto fueron:

- **Extracción y Preparación de Datos**: Se extrajeron y prepararon los datos del conjunto de Kaggle, utilizando AWS Data Glue Databrew. Esta herramienta permitió una preparación eficiente de los datos para el análisis ya que la interfaz visual permitía trabajar sobre los datos de forma eficiente.

- **Almacenamiento de Datos:** Los datos se almacenaron en AWS S3 en dos formatos: RAW para datos brutos y TRANSFORMED para datos procesados y de mejor calidad. Se consideró el uso del formato JSON para optimizar el almacenamiento y acceso a los datos. Estos JSON son los obtenidos desde Amazon Comprehend.

- **Análisis de Sentimiento:** Se utilizó AWS Comprehend para analizar tres elementos clave de las reviews: el sentimiento general, el sentimiento dirigido y la identificación de palabras clave. Este enfoque proporcionó Insights valiosos sobre las opiniones de estos usuarios respecto al iPhone 15.

- **Visualización de Datos y Reportes:** Finalmente, se utilizó AWS Athena y AWS QuickSight para el análisis y la visualización de los datos. Esto permitió crear dashboards interactivos y reportes detallados sobre los resultados del análisis de sentimiento y las tendencias encontradas en las reseñas de los usuarios.

Este enfoque adaptado permitió superar los desafíos surgidos durante la ejecución del proyecto y lograr un análisis efectivo y detallado de las opiniones de los usuarios que están realizando las reviews sobre el iPhone 15, proporcionando valiosos Insights para estrategias de marketing y desarrollo de productos.

No es tarea de este trabajo de Fin de Grado realizar un proyecto de gran envergadura ni que se incorpore características no esenciales que alarguen el desarrollo. Se debe entender la complejidad de iniciar un proyecto sin conocimientos previos y acotar el alcance de este.

1.2. Impacto en sostenibilidad, ético-social y de diversidad

El uso de servicios serverless de AWS como AWS Lambda, Glue, Athena y QuickSight tiene un impacto significativo en la sostenibilidad debido a la eficiencia en el uso de recursos y la reducción de la huella de carbono. Aquí analizaremos cómo cada uno de estos servicios contribuye a la sostenibilidad:

AWS Lambda:

- **Eficiencia de Recursos:** Lambda permite ejecutar código sin necesidad de aprovisionar o administrar servidores ni sistemas operativos, esto significa que solo se utilizan recursos cuando el código está en ejecución. Esto reduce el desperdicio de recursos de cómputo.

- **Escalabilidad Automática:** Se ajusta automáticamente la capacidad de cómputo asignada en función de la demanda, asegurando que solo se consuman los recursos necesarios.

- **Reducción de la Huella de Carbono:** Al optimizar el uso de recursos, AWS Lambda ayuda a reducir la cantidad de energía necesaria para mantener y operar infraestructuras de servidores físicos.

AWS Glue:

- **Gestión Eficiente de Datos:** Glue es el mejor servicio en AWS para las ETL ya que facilita la preparación y transformación de grandes volúmenes de datos. Al manejar datos de forma eficiente, minimiza la necesidad de computación.

- **Automatización de Tareas:** Reduce la necesidad de intervenciones manuales y tareas repetitivas, lo que conlleva a un uso más eficiente del tiempo y los recursos.

AWS Athena:

- **Consultas Optimizadas**: Athena permite ejecutar consultas sobre datos almacenados en S3 utilizando SQL, lo que significa que los usuarios pueden realizar análisis sin necesidad de servidores dedicados. Del mismo modo que Lambda, no requiere una administración de servidor o SO.

- **Pago por Uso:** Como el resto de los servicios serverless, Athena cobra por consulta realizada, lo que incentiva a los usuarios a optimizar sus consultas para ser más eficientes, esto también se traduce en un uso más eficiente de los recursos.

AWS QuickSight:

- **Análisis de Datos Sostenible:** QuickSight ofrece capacidades de análisis de datos y visualización sin la necesidad de servidores. Esto permite a las organizaciones realizar análisis de datos sin la carga adicional de mantener una infraestructura física, de servidor o sistema operativo, lo cual se traduce en un uso más eficiente.

- **Escalabilidad Eficiente**: Al igual que con el resto de los servicios serverless, QuickSight escala automáticamente según la demanda, lo que asegura que solo se utilicen los recursos necesarios.

En general, el uso de servicios serverless como los ofrecidos por AWS promueve la eficiencia energética y la reducción de la huella de carbono en el flujo de trabajo propuesto para este proyecto. Al eliminar la necesidad de mantener una infraestructura de servidores físicos constantemente activa, estos servicios permiten a las empresas ser más ágiles y sostenibles, contribuyendo a un menor impacto ambiental global.

1.3. Enfoque y método seguido

Se trata de la creación de un nuevo producto ya que se realiza todo el flujo de trabajo desde el inicio, no obstante, este flujo se integra con herramientas existentes proporcionadas por AWS. Las razones para elegir esta estrategia son:

- **Eficiencia y Reducción de Tiempo:** Utilizar servicios como AWS Lambda, Glue, Athena y QuickSight permite aprovechar soluciones ya probadas y eficientes, acelerando significativamente el proceso de desarrollo.

- **Flexibilidad y Escalabilidad:** Estos servicios ofrecen gran flexibilidad y escalabilidad, lo que es esencial para manejar variaciones en la carga de trabajo y el volumen de datos.

- **Costo-Efectividad:** La estructura de pago por uso de los servicios serverless de AWS hace que esta estrategia sea económicamente viable, especialmente para un proyecto de TFG donde el coste del proyecto está asumido íntegramente por el alumno.

- **Enfoque en el Análisis de Datos:** Al utilizar herramientas ya desarrolladas, se puede centrar más en el análisis de datos y menos en los aspectos técnicos de la infraestructura de TI, alineándose así directamente con los objetivos del proyecto y del ámbito del TFG elegido.

En conclusión, la integración de herramientas existentes de AWS se identifica como la estrategia más apropiada para este proyecto, ya que maximiza la eficiencia, reduce los riesgos y costos, y permite un enfoque más exhaustivo en el análisis de datos y la obtención de insights valiosos sobre las opiniones de los usuarios del iPhone 15.

1.4. Planificación del Trabajo

Este TFG se basará en el uso de dos metodologías utilizadas en el ámbito de la gestión de proyectos y el desarrollo de software. Para el inicio y gestión del proyecto se usará la metodología propuesta por PMI (Project Management Institute con la variante del desarrollo secuencial del producto Waterfall del SDLC. A continuación, se detalla el plan de trabajo con las siguientes fases: • Inicio • Planificación • Ejecución • Monitorización y Control • Cierre.

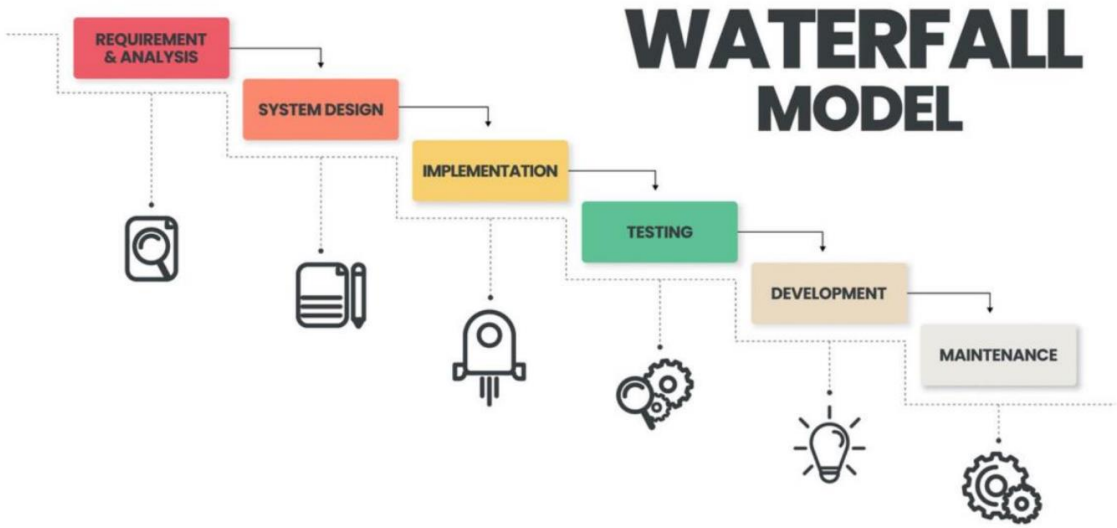

Ilustración 1. Modelo Cascada

Existe solo un recurso para el desarrollo del proyecto que deberá de ejercer los roles de Project Manager, Cloud Architecture, Data Engineer y Data Analyst.

Debido al poco conocimiento sobre el ámbito de la nube pública, las tareas se han ido desarrollando en el transcurso de la ejecución del proyecto, llegando así a una lista de tareas a cumplir para cada una de las entregas propuestas. Esto ha ocasionado que, hasta el último momento, no se puede visualizar unos datos relevantes. Las tareas realizadas han sido plasmadas en un diagrama de Gantt utilizando las herramientas de Atlassian como Jira.

1.2. Planificación y Cronograma

Para el proyecto propuesto se han desarrollado una serie de hitos a cumplir durante las muestras de las PECs que se entregarán periódicamente, estos hitos están definidos con el fin de cumplir el TFG propuesto y llevar un seguimiento.

Nombre	Fecha propuesta	Tipo	Cumplimiento
Propuesta inicial aceptada	02/10/2023	Ejecución	**DONE**
Planificación del proyecto	15/10/2023	Ejecución	**DONE**
Análisis del proyecto	15/10/2023	Ejecución	**DONE**
Análisis de datos realizado	10/11/2023	Ejecución	**DONE**
Entrega y revisión PEC2	15/11/2023	Seguimiento	**DONE**
Primera ingesta	25/11/2023	Ejecución	**DONE**
Primeras conclusiones	10/12/2023	Ejecución	**DONE**
Entrega y revisión PEC3	17/12/2023	Seguimiento	**DONE**
Desarrollo ETL	25/12/2023	Ejecución	**DONE**
Primeros Dashboards	05/01/2024	Ejecución	**DONE**
Entrega y revisión PEC4	14/01/2024	Seguimiento	**DONE**
Dashboard final	05/01/2024	Ejecución	**DONE**
Flujo completo (end to end)	14/01/2024	Ejecución	**DONE**

Para el cumplimiento de los hitos anteriores se ha basado en una planificación a través de un diagrama de Gantt para planificar el avance de las tareas, así como la desviación de las fechas de entrega de estas.

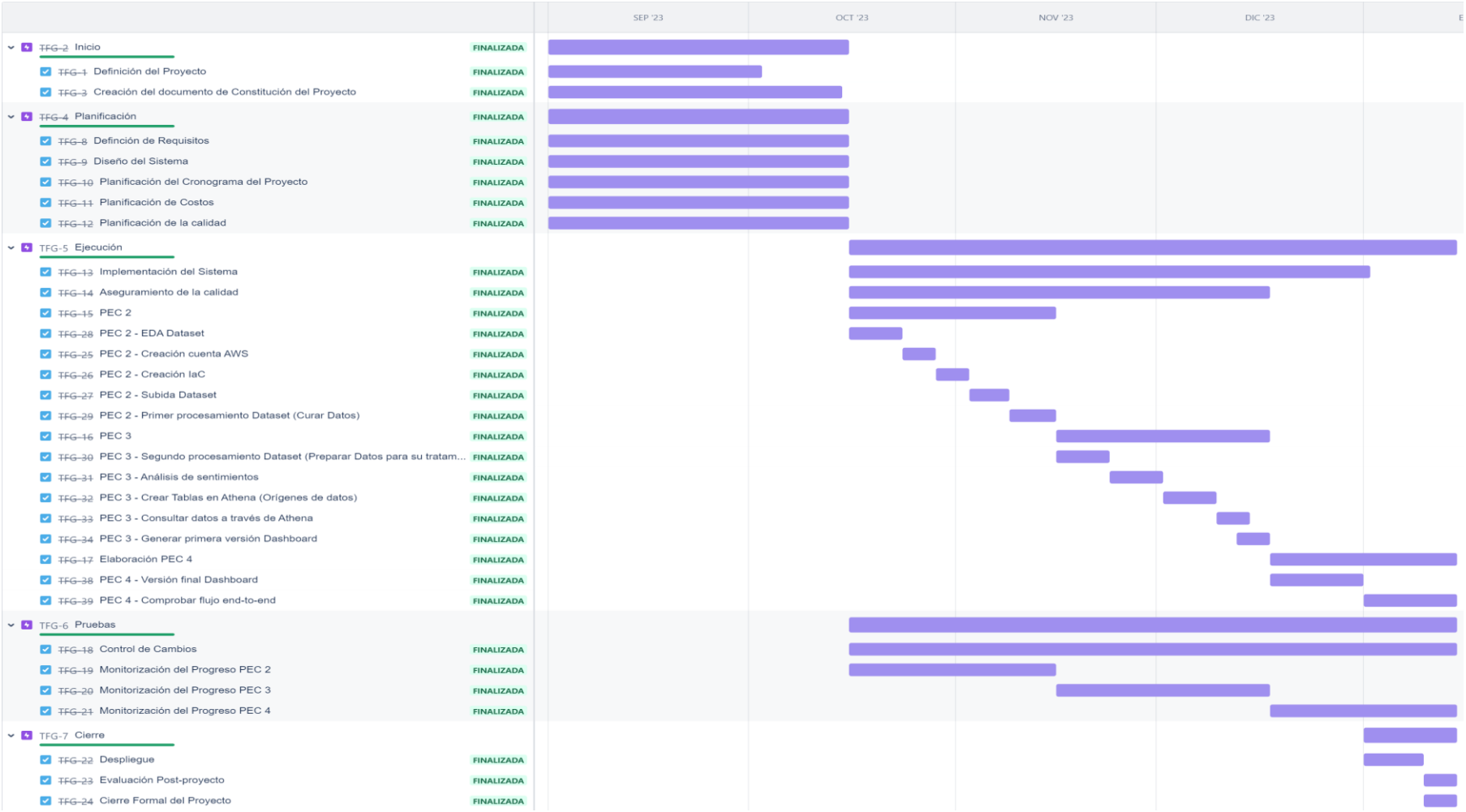

Ilustración 2. Diagrama Gantt

1.3. Gestión de Riesgos

1.4. Presupuesto y recursos asignados

Para poder hacer la estimación del equipo humano se debe conocer el salario de un ingeniero de datos en España. Este salario será utilizado para calcular los 4 meses de duración

Concepto	Valor
Salario Bruto Anual	**35.000€**
Total 4 meses	**11.666€**

La siguiente tabla muestra la estimación final del coste de realizar el proyecto, a los conceptos se le debe sumar un importe de un 15% de contingencia para posibles imprevistos.

Concepto	Valor
Estimación salario	**11.666€**
Estimación costos infraestructura	**200€**
Subtotal	**11.866€**
Contingencias 15%	**1.779€**
TOTAL	**13.645€**

1.5. Breve sumario de productos obtenidos

Análisis de Sentimientos sobre el iPhone 15: Uno de los principales productos será una base de datos o conjunto de datos que categorice y analice los sentimientos, estos datos estarán alojados en AWS S3.

Dashboard en Amazon Quicksight: Se entrega un panel de control interactivo que muestre gráficos y métricas relacionadas con la percepción del iPhone 15.

Infraestructura Serverless en AWS: Una infraestructura desplegada en AWS que incluirá el uso de funciones lambda, AWS S3, Glue, Athena y otros servicios para recolectar, almacenar, procesar y analizar los datos.

Un código que contenga un Proceso ETL: Software desarrollado en Python para las operaciones ETL estará disponible en AWS Lambda para ser consultado a través de las credenciales de acceso proporcionadas.

Conclusiones y Recomendaciones: Conclusiones basadas en los resultados obtenidos relacionados con el análisis de sentimientos y el uso de tecnologías cloud.

1.6. Breve descripción de los otros capítulos de la memoria

Introducción

Capitulo de justificación del TFG, contextualización y planificación.

Estado del arte

Capítulo del análisis de mercado de proyectos de BI, antecedentes y diferentes servicios existentes.

Arquitectura del sistema

Capítulo de explicación inicial de servicios utilizados y motivos. Propuesta del diseño de arquitectura implementada y justificación de las tareas realizadas, así como la explicación de las decisiones tomadas.

Resultados y productos obtenidos

Resultados obtenidos, explicación y conclusiones identificadas tras la finalización de la implementación del proyecto. Además, incluye las impresiones finales y las propuestas de actualizaciones de la plataforma futuras.

2. Estado del arte

2.1. Business Analytics

Los orígenes del BI tradicional, se enfocaba principalmente en la generación de informes y análisis descriptivos que se basaban en datos históricos. Con el tiempo, las herramientas de BI evolucionaron para incluir capacidades analíticas más avanzadas como el análisis predictivo. Esta evolución vino impulsada por el desarrollo de tecnologías de almacenamiento de datos.

El mercado de los proyectos de BI ha experimentado un crecimiento exponencial en los últimos años, este crecimiento viene impulsado por la necesidad de las empresas de tomar decisiones basadas en datos de un entorno empresarial cada vez más grande, competitivo y dinámico.

Las herramientas de BI se han vuelto más sofisticadas y accesibles, algunas herramientas como Microsoft Power BI, Tableau, Google Data Studio o SAS han democratizado el acceso a análisis de datos más avanzados, permitiendo a los usuarios que no tengan una gran capacidad técnica realizar análisis complejos.

Además, la integración de algoritmos de inteligencia artificial y machine learning en herramientas de BI están permitiendo realizar análisis predictivos más avanzados en función de las necesidades de la empresa, llegando a realizar predicciones en tiempo real.

El análisis en tiempo real se está convirtiendo en una norma para las empresas, cada vez es más común que las empresas busquen insights instantáneos para tomar decisiones más rápidas y efectivas.

2.2. Big Data

El Big Data surgió como respuesta a las limitaciones de las bases de datos tradicionales a la hora de manejar volúmenes de datos masivos y de gran crecimiento, con el auge de las redes sociales y aplicaciones móviles e IoT, cada vez se obtienen más datos, menos estandarizados y con un crecimiento exponencial.

Se ha observado un aumento en la adopción de lagos de datos y data warehouses modernos como AWS Lake Formation, Google BigQuery que ofrecen una capacidad de almacenamiento casi “ilimitada” para cualquier empresa ya que este almacenamiento es escalable y seguro.

Hadoop y Apache Spark fueron las primeras tecnologías en abordar el procesamiento de datos a gran escala, permitiendo democratizar el acceso al

Big Data haciéndolas más accesibles para empresas de todos los tamaños, permitiendo una mayor adopción al mercado.

Las herramientas de inteligencia artificial y machine learning, como tendencia creciente, proporcionan capacidades analíticas de mayor profundidad y con mayor sofisticación.

Además, debido a las regulaciones recientes de el GDPR, la privacidad y seguridad de los datos en Big Data han ganado relevancia en el mercado, ello conlleva a que las empresas deben invertir más en soluciones que ofrezcan una seguridad en la gestión de los datos.

2.3. Nube Pública

Inicialmente, las empresas dependían de infraestructura desplegada en local (on-premise). La nube pública emergió como una solución más flexible y escalable para las empresas, pudiendo pagar por uso sin necesidad de una inversión inicial que afectara al correcto desarrollo de esta.

Amazon Web Services, Junto Google Cloud y Azure dominan el mercado de la nube pública, estas plataformas continúan expandiendo sus servicios y capacidades para solventar las diferentes necesidades que tienen las empresas que lo utilizan.

AWS se encuentra en la vanguardia de ofrecer servicios serverless (sin servidor) que permiten a las empresas construir y desplegar aplicaciones sin gestionar la infraestructura subyacente, de esta forma, las empresas pueden enfocarse en la aplicación o servicio que ofrecen.

Estos servicios sin servidor ponen un enfoque en la sostenibilidad, enfoque fundamental para la nube pública, ya que gracias a esta tecnología se puede operar de manera más ecológica y reducir considerablemente la huella de carbono de la aplicación desarrollada por la empresa.

AWS está constantemente expandiendo la suite de herramientas que ofrecen servicios de inteligencia artificial y machine learning como Amazon Comprehend, que permiten a los desarrolladores utilizar modelos de machine learning de manera más eficiente.

En resumen, el estado del arte en estas áreas refleja una tendencia hacia la democratización del Bussiness Analytics sustentados proyectos de Big Data desplegados en nube pública. Esto permite un enfoque continuo en seguridad, privacidad, sostenibilidad y eficiencia. La innovación constante en estos campos sigue abriendo oportunidades y desafíos a las empresas.

3. Arquitectura del Sistema

3.1. Componentes del Sistema

La arquitectura está construida sobre los servicios de AWS haciendo enfoque en servicios serverless como AWS S3, AWS Lambda, AWS Glue Databrew, AWS Comprehend, AWS Athena y AWS QuickSight.

3.1.1. AWS S3 (Simple Storage Service)

AWS S3 es un servicio de almacenamiento de objetos ofrecido por la nube pública de Amazon, este servicio proporciona escalabilidad, disponibilidad de datos, seguridad y un alto rendimiento. Permite a las empresas almacenar y recuperar cualquier cantidad de datos en cualquier momento y desde cualquier lugar, esto hace de ello una solución ideal para almacenar datos de Big Data, la finalidad de nuestro proyecto.

S3 almacena los datos como objetos dentro de “buckets”. Un objeto es cualquier tipo de dato como una imagen, un texto, un video o un documento en cualquier formato. Cada objeto se identifica de manera única dentro del bucket por una key (clave) que puede tener hasta 5 TB de tamaño

Tiene una disponibilidad del >99% lo que significa que los datos están protegidos garantizando una alta disponibilidad al replicar datos en múltiples centros de datos dentro de una región en AWS.

Contiene avanzadas configuraciones para controlar el acceso a nivel de objeto y bucket, los datos se encuentran encriptados en reposo y tránsito y ofrece soporte para cumplir certificaciones de seguridad necesarias para la empresa.

Se integra perfectamente con el resto de los servicios de AWS, lo que lo hace un servicio ideal para almacenar y analizar grandes volúmenes de datos para aplicaciones de Big Data.

3.1.2. AWS Glue DataBrew

AWS Glue DataBrew es un servicio de preparación de datos a través de una interfaz visual e interactiva que facilita la limpieza y normalización de los datos sin necesidad de escribir una sola línea de código. Este servicio permite a los usuarios de cualquier nivel técnico transformar y preparar los datos para el análisis de Machine Learning de manera más eficiente.

Esta interfaz permite a los usuarios aplicar transformaciones como filtrar, formatear, sumarizar, estandarizar mediante operaciones de arrastrar y soltar

sin necesidad de escribir código, ofrece más de 250 transformaciones para las tareas más comunes de preparación de datos.

Este servicio está totalmente integrado con S3, permite automatizar los flujos de trabajo haciéndolo un servicio diseñado para manejar grandes volúmenes de datos sin degradar el rendimiento.

3.1.3.AWS Lambda

AWS Lambda permite ejecutar código sin servidor en respuesta a eventos, AWS se encargará automáticamente de los recursos de cómputo necesarios para que la ejecución sea satisfactoria. Lambda elimina la necesidad de aprovisionar servidores y se paga únicamente por el tiempo de cómputo consumido.

Lambda se activa mediante eventos específicos de otros servicios de AWS, como cambios en un bucket S3 (subida de un fichero, modificación, eliminación…) o peticiones HTTP mediante la disponibilización de una URL, esta capacidad la convierte en una solución ideal para aplicaciones impulsadas por eventos.

Los recursos están automáticamente administrados por Lambda, la infraestructura subyacente escala de manera elástica según la demanda. Esto permite a los desarrolladores centrarse en el código sin preocuparse por la gestión de los servidores.

Ofrece soporte para diversos lenguajes de programación como Node.js, Python, Java, Go, Ruy… que permite ejecutar el lenguaje que el desarrollador decida utilizar.

3.1.4.AWS Comprehend

AWS Comprehend es un servicio de procesamiento de lenguaje natural (NLP) y machine learning (ML) que facilita notablemente el análisis de textos. Utiliza modelos de aprendizaje automático para descubrir insights e interrelaciones en textos sin que el desarrollador tenga experiencia previa en machine learning.

AWS Comprehend identifica automáticamente el idioma del texto y puede extraer información valiosa como palabras clave, sentimientos dirigidos, sentimientos, relaciones…. Es capaz de determinar si el texto tiene una connotación neutral, positiva, negativa o mixta, esto es esencial para el análisis de opiniones o reviews de productos.

Además, es fácilmente integrable con otras herramientas y servicios de AWS como AWS S3 para análisis de grandes conjuntos de datos y AWS Lambda para su automatización en el procesamiento.

Es capaz de extraer insights y tendencias de conjuntos de datos de texto como transcripciones para realizar un análisis de sentimiento, esto hace que sea el algoritmo preentrenados idóneo para el proyecto.

3.1.5.AWS Athena

AWS Athena es un servicio de consulta interactiva proporcionada por AWS que facilita el análisis de datos almacenados en S3 utilizado el lenguaje SQL. Permite ejecutar consultas ad-hoc sobre grandes cantidades de datos almacenados sin necesidad de gestionar ninguna infraestructura de base de datos, lo que simplifica significativamente el proceso de análisis de datos.

Además, solo se paga por los datos que se escanean durante las consultas, este modelo de costos eficiente y flexible hace que sea totalmente recomendable para proyectos de Big Data.

AWS Athena soporta variedad de formatos de datos, incluyendo JSON, CSV, Parquet, OCR… lo que brinda una gran flexibilidad al proyecto de trabajar con diferentes tipos de conjuntos de datos.

3.1.6.AWS QuickSight

AWS QuickSight es el servicio de BI por excelencia de AWS, permite a los usuarios crear y compartir visualizaciones interactivas, dashboards y reportes de marera fácil y rápida, facilitando el análisis de datos y la toma de decisiones basada en datos.

Ofrece una interfaz de usuario intuitiva que permite a los usuarios crear visualizaciones sin necesidad de una formación técnica especializada, algo que lo diferencia de herramientas de la competencia.

QuickSight ofrece un modelo de precios basado en la suscripción que se adapta a empresas de todos los tamaños con opciones de precios según el uso o capacidades limitadas por el usuario.

3.1.7.AWS IAM (Identity and Access Management)

AWS IAM permite a los administradores de infraestructura controlar el acceso seguro a cada uno de los recursos de AWS. IAM se utiliza tanto para gestionar usuarios, como grupos, los roles que estos tienen y las políticas que aplican a cada uno de ellos.

Los roles de IAM son una forma segura de conceder permisos a entidades que necesitan realizar acciones sobre la cuenta, estas entidades pueden ser un usuario u otro servicio de AWS o incluso de terceros. Las políticas de permisos que están gestionadas como documentos JSON definen qué acciones están permitidas o denegadas para ese rol, usuario o grupo.

AWS IAM cumple con una gran cantidad de estándares y regulaciones de seguridad que hacen de este un servicio totalmente fiable para cualquier empresa que decida utilizarlo.

3.2. Visión General de la Arquitectura

La arquitectura está construida sobre los servicios previamente mencionados, donde priman los servicios serverless para crear un sistema de BI eficiente y escalable. Teniendo las siguientes consideraciones.

- Se utiliza AWS S3 como el repositorio central para almacenar los datos.
- Se utiliza AWS Lambda para ejecutar funciones de procesamiento de datos de una forma totalmente automática en función de los eventos de S3.
- Se utiliza AWS Comprehend en vez de un modelo preentrenados de Hugging Face ya que este realiza acciones de NLP y está totalmente integrado con AWS.
- AWS Athena y AWS QuickSight se utiliza para la consulta y visualización de datos directamente sobre S3.
- El flujo de trabajo seguido es el habitual en un proyecto de Big Data, como muestra la imagen siguiente.

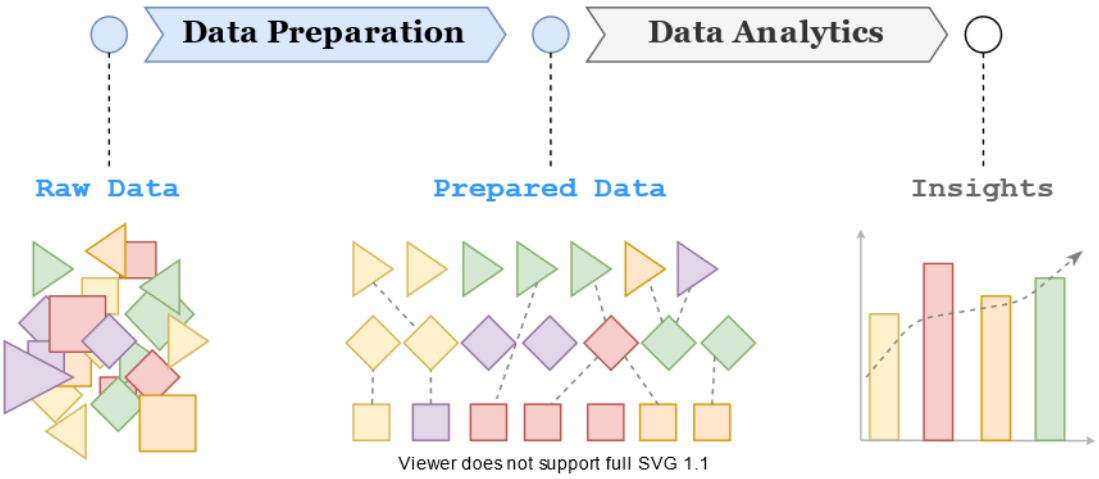

Ilustración 3. Estados del dato

La siguiente imagen recoje la infraestructura utilizada y cómo se comunican los diferentes servicios para limpiar, preparar y utilizar los datos en función de las necesidades surgidas durante el proyecto.

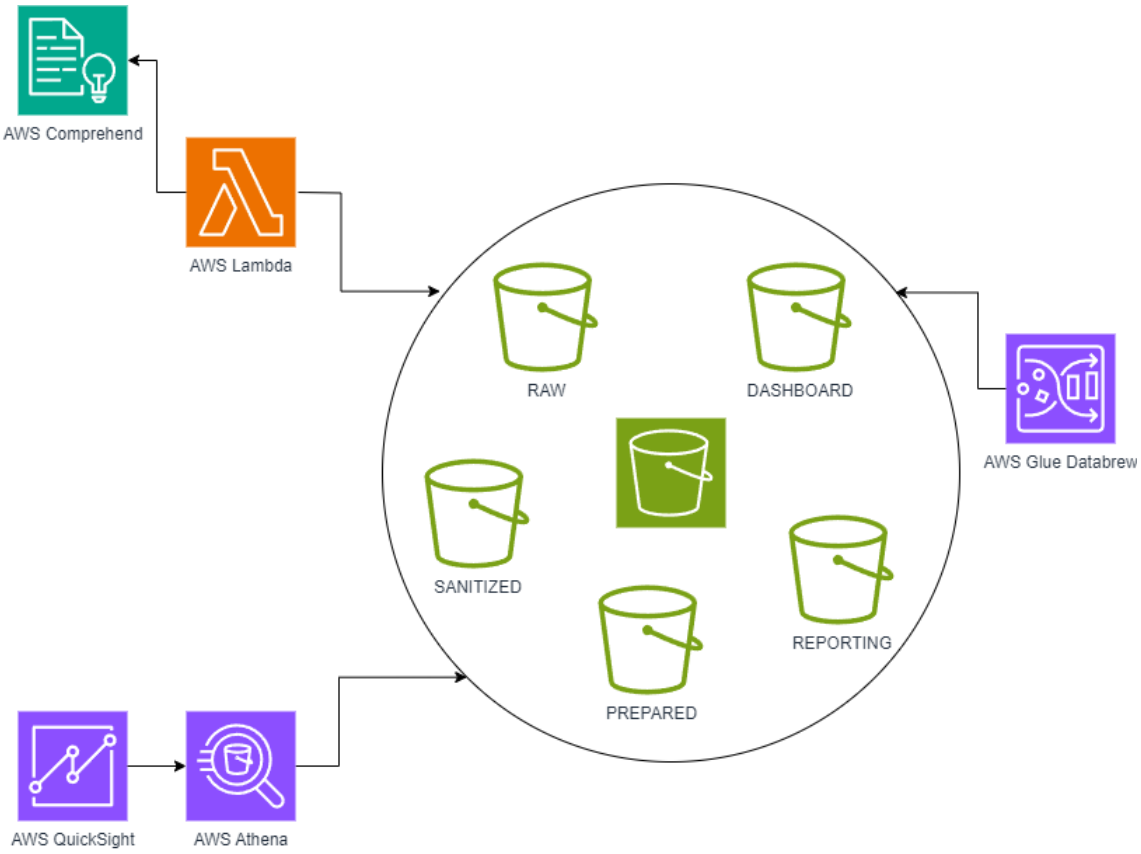

Ilustración 4. Diagrama de Arquitectura

3.3. Flujo de Datos y Procesamiento

El diseño de arquitectura garantiza un flujo de datos fluido y bien orquestado entre los diferentes componentes de AWS. Los datos se mueven desde S3 a través del procesamiento en Lambda que a su vez integra AWS Comprehend con los datos. Por último, se constituye la visualización a través de consultas en Athena y mostrándolos en QuickSight. Cada componente se integra de manera que la eficiencia es máxima en detrimento de los costos operativos.

Este diseño representa un enfoque moderno y totalmente eficiente para cualquier empresa ya que está alineado con las mejores prácticas de arquitecturas cloud y optimizado para un proyecto TFG enfocado en resultados analíticos, pero con una profunda implementación y manejo de soluciones cloud avanzadas.

La siguiente imagen muestra cómo de forma automática, cuando un fichero es subido a AWS, este automáticamente es preparado y modifcado para alimentar el dashboard que permitirá a la empresa visualizar los datos para tomar decisiones estratégicas.

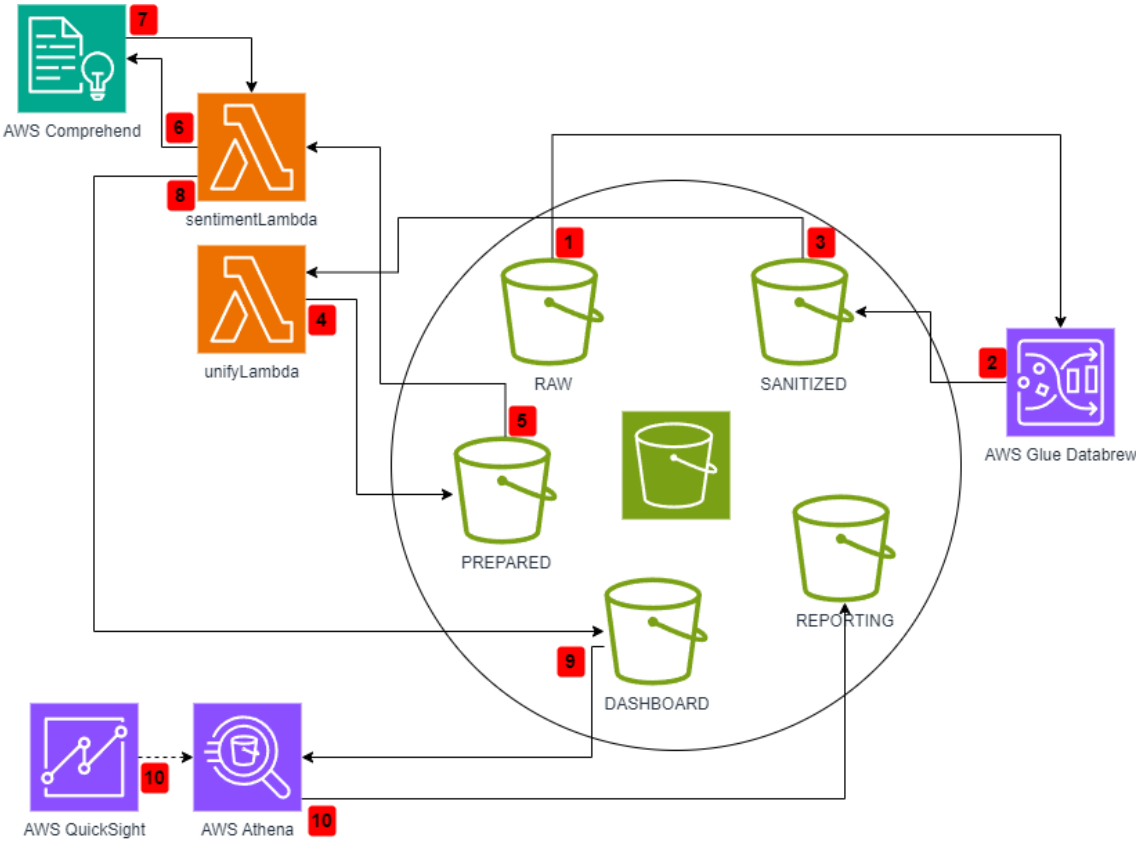

Ilustración 5. Diagrama de Flujo

3.4. Implementación de Componentes

3.4.1. Almacenamiento de datos

Los datos iniciales se encuentran en Kaggle, donde un usuario a transcrito lo que parecen vídeos en YouTube sobre usuarios que están valorando el nuevo iPhone 15, estas transcripciones están exportadas en un fichero CSV.

https://www.kaggle.com/datasets/nuhmanpk/iphone-15-15-pro-pro-max-reviews

El fichero CSV recoge cada frase en una línea, donde se muestra el texto, el segundo en el que se inicia la frase y la duración de la misma.

text,start,duration
- Hey, it's Justine and
the brand new iPhone 15.,1.71,1.74
We've got the Plus in green,3.45,1.59
and the regular iPhone 15 in pink,,5.04,2.493
which obviously we know
Now, a lot of people are saying,10.14,1.17
that this year's release was
and I actually think that this is probably,13.98,1.92
Apple's most impressive base model release,15.9,2.76
that they've had in a really long time.,18.66,1.89
And here's why.,20.55,1.21
(upbeat music),21.76,2.583
First of all, I think the

Ilustración 6. Datos en crudo

Estos datos, que están en bruto, se almacenarán en un Bucket S3 identificado como RAW, en los proyectos de Big Data, RAW hace referencia a los ficheros que no han recibido ningún tipo de tratamiento, para asegurar la integridad de estos ficheros, el bucket está protegido contra modificaciones sobre ficheros existentes o eliminación de los mismos, solamente se pueden generar nuevos ficheros. Conceptualmente esto se conoce como WORM (write once read many).

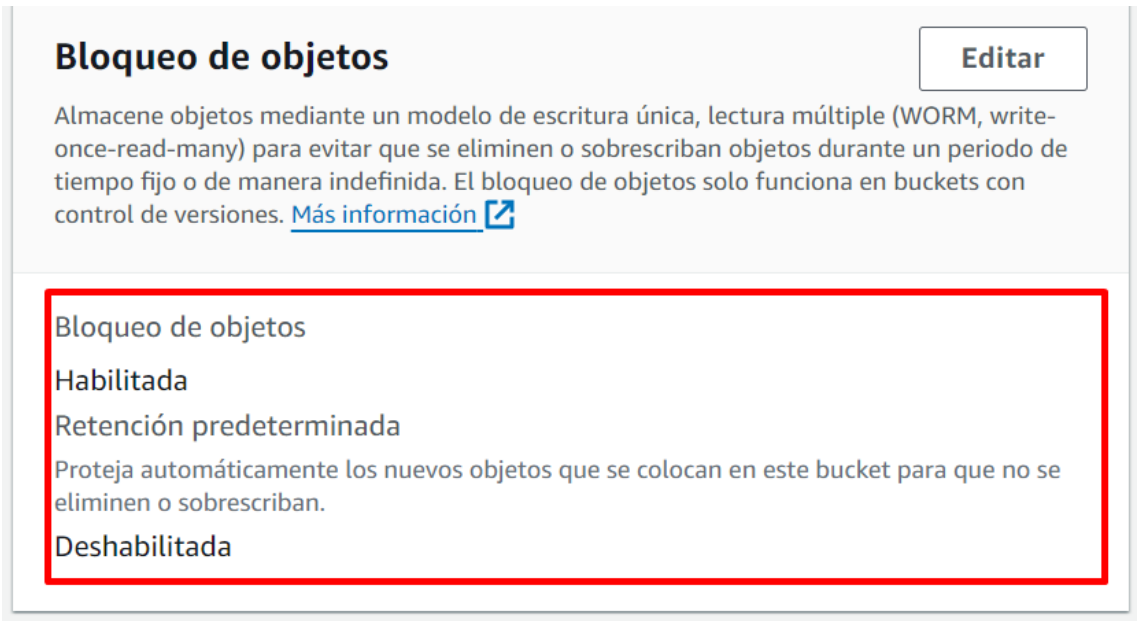

Ilustración 7. Configuración Bucket S3

El resto de Buckets S3 se denominan Sanitized, Prepared, Reporting y Dashboard donde los datos estarán almacenados en cada uno de sus estados para asegurar el flujo y el origen de los datos, estos bucket no tienen una configuración especial, por ello simplemente se mencionan que existen, en los siguientes puntos se hará referencia qué tipo de datos existen en cada uno.

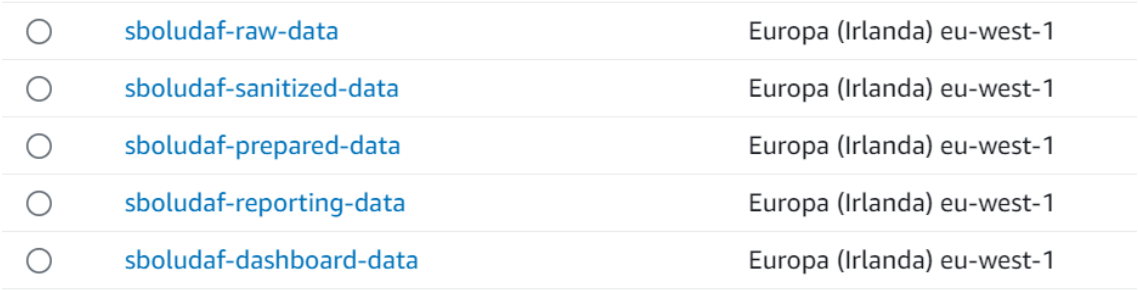

Ilustración 8. Listado de Buckets

Los bucket S3 serán los únicos elementos del proyecto donde existirán datos almacenados, el resto de servicios utilizados en AWS son de tránsito y los datos no serán retenidos.

3.4.2. Estructura de Datos

En un proyecto de BI o Big Data, la estructura de datos es un aspecto crucial. La estructura de datos hace referencia a la organización, almacenamiento y gestión de los datos dentro del proyecto. La estructuración de los datos está definida en función del uso de diferentes servicios de AWS.

Los datos estarán almacenados en bucket S3 independientes, uno para cada estado del dato:

- **RAW data:** Se encontrarán los datos en bruto que no han sufrido ningún tipo de tratamiento.
- **SANITIZED data:** Se encontrarán los datos que han sufrido una primera limpieza a través del servicio Glue DataBrew.
- **PREPARED data:** Se encontrarán los datos que han sido preparados y están listos para ser consumidos por el modelo de AWS Comprehend.
- **REPORTING data:** Se encontrarán los datos listos para ser consultados a través de AWS Athena.
- **DASHBOARD data:** Se encontrarán los datos que hacen referencia a las consultas realizadas por el servicio de AWS Athena + AWS QuickSight y serán utilizados para crear el dashboard.

Se define una nomenclatura clara para los buckets y los archivos dentro de ellos, esto ayudará a la identificación de los datos y su procedencia, esto permitirá la organización de los datos independientemente del crecimiento del proyecto teniendo las siguientes consideraciones:

- El nombre de los buckets tiene un prefijo llamado **sboludaf-** para asegurar que el nombre del bucket es único, uno de los requisitos del servicio de AWS S3.

- Los datos **RAW, SANITIZED y PREPARED** se mantiene el formato **CSV** de origen, no obstante, los datos **REPORTING** listos para ser consumidos por dashboards se encuentran en formato **JSON** al provenir de una API.

- La estructura de carpetas se mantiene con la fecha del procesamiento, Ejemplo: **sboludaf-work_12Nov2023_1699827778175/** de esta forma, se podrá obtener un histórico de procesamiento y se podrá revertir el flujo por completo si se identificara un error. De forma adicional, si los datos ya están procesados, la carpeta tendrá un prefijo **processed**_.

- Los datos listos para el reporting están totalmente separados en carpetas, además, **cada llamada de la API estará almacenado en un fichero** manteniendo el nombre del fichero para poder hacer **referencia a los ficheros RAW** que se han utilizado, Ejemplo: **sboludaf-work_12Nov2023_1699827778175_part00000_key_phrases_0_1.json**

Para los datos estructurados, se define un esquema de datos que describe la estructura de este, incluyendo los nombres de los campos y el tipo. Al utilizar Amazon Comprehend se pueden añadir campos adicionales al esquema para almacenar los resultados del análisis, el esquema de los datos es:

Sentimientos

- **Sentiment:** El sentimiento general de la entrada (como "MIXED", "POSITIVE", "NEUTRAL").
- **SentimentScore:** Un objeto que contiene puntuaciones para diferentes tipos de sentimientos (Positive, Negative, Neutral, Mixed).
- **ResponseMetadata:** Metadatos de la respuesta.

```
`Sentiment` string,
`SentimentScore` struct<
    `Positive`: double,
    `Negative`: double,
    `Neutral`: double,
    `Mixed`: double
>,
`ResponseMetadata` string
```

Ilustración 9. Esquema Sentimientos

Palabras Clave

- **KeyPhrases**: Un array de frases clave, donde cada frase incluye:
  - **Score**: Una puntuación de porcentaje de confianza para la frase.
  - **Text**: El texto de la frase.
  - **BeginOffset y EndOffset**: Posiciones de inicio y fin de la frase en el texto original.

```
`KeyPhrases` array<struct<
    `Score`: double,
    `Text`: string,
    `BeginOffset`: int,
    `EndOffset`: int
>>,
`ResponseMetadata` string
```

Ilustración 10. Esquema Palabras Clave

Sentimientos Dirigidos

- **DescriptiveMentionIndex**: Un array de índices que describen menciones.

- **Mentions**: Un array de menciones individuales, donde cada mención incluye:
  - **vScore**: Una puntuación de porcentaje de confianza para la mención.
  - **GroupScore**: Una puntuación de agrupación para la mención.
  - **Text**: El texto de la mención.
  - **Type**: Tipo de mención, a qué se refiere.
  - **MentionSentiment**: Análisis de sentimiento de la mención, que incluye:
    - **Sentiment**: Sentimiento general (como "NEUTRAL", "POSITIVE", "NEGATIVE").
    - **SentimentScore**: Puntuación de fiabilidad de la mención
  - **BeginOffset y EndOffset**: Posiciones de inicio y fin de la mención en el texto original.

```
`Entities` array<struct<
    `DescriptiveMentionIndex`: array<int>,
    `Mentions`: array<struct<
        `Score`: double,
        `GroupScore`: double,
        `Text`: string,
        `Type`: string,
        `MentionSentiment`: struct<
            `Sentiment`: string,
            `SentimentScore`: struct<
                `Positive`: double,
                `Negative`: double,
                `Neutral`: double,
                `Mixed`: double
            >
        >,
        `BeginOffset`: int,
        `EndOffset`: int
    >>
>>
```

Ilustración 11. Esquema Sentimientos Dirigidos

El control de acceso a estos datos se realiza a través de políticas de IAM asociado a los roles que utilizan estos servicios para controlar el acceso. Esto asegura que solo los usuarios y servicios autorizados pueden acceder o modificar los datos del proyecto.

Por último, los datos en S3 se encriptan tanto en tránsito como en reposo para garantizar la seguridad e integridad de estos.

Athena, a pesar de parecer un motor de base de datos, solamente se utiliza para consultar los datos sobre S3, para ello, es necesario utilizar AWS Glue para catalogar los datos y realizar las consultas directamente sobre los buckets, para que esto sea posible, es necesario crear 3 tablas externas en Athena, que son las siguientes:

Tabla Sentimientos

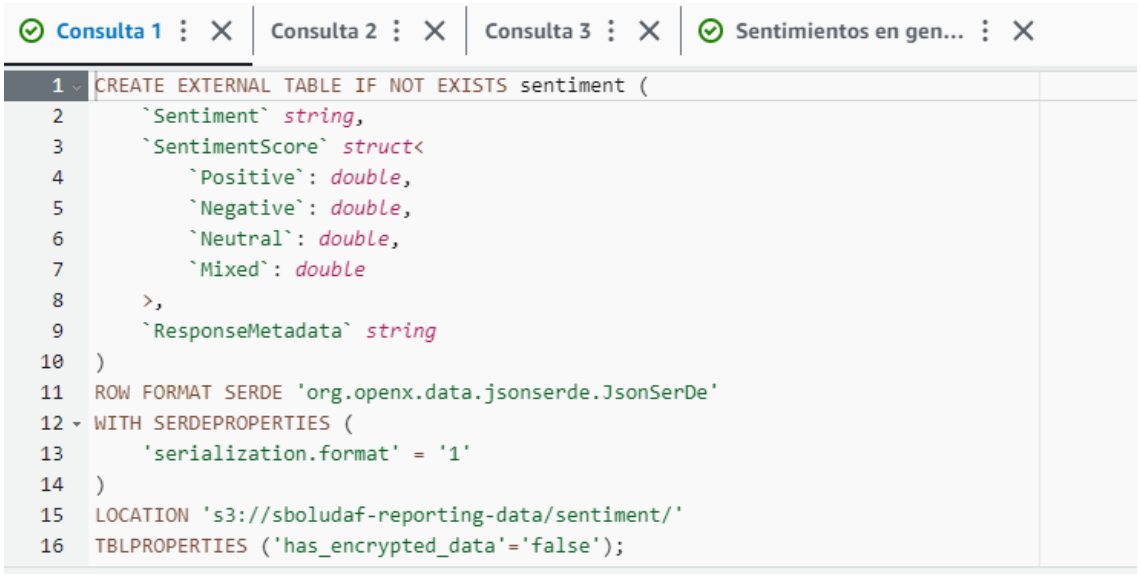

Ilustración 12. Consulta Creación Tabla Sentimientos

Tabla Palabras Clave

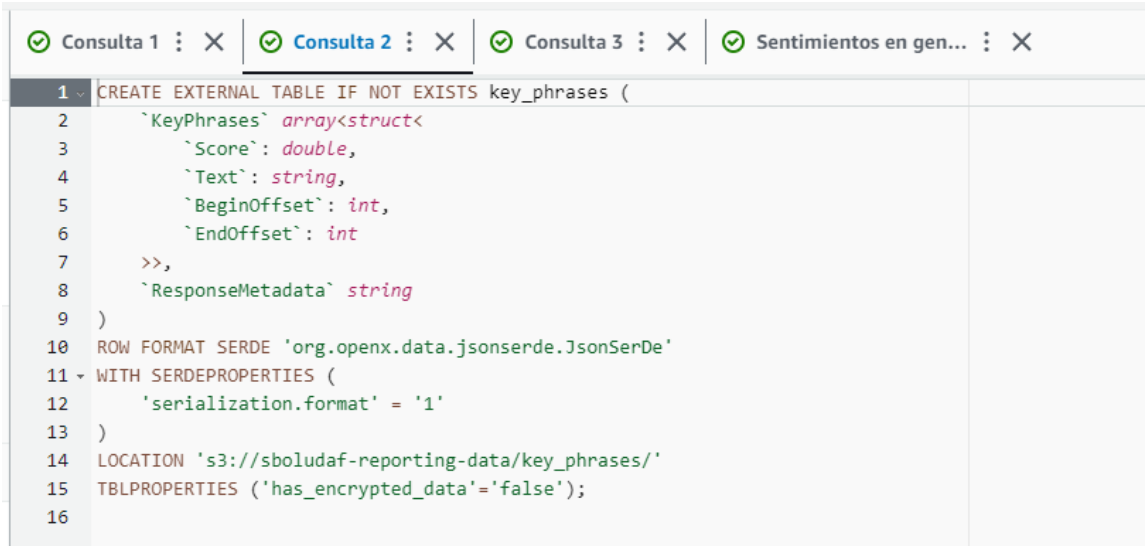

Ilustración 13. Consulta Creación Tabla Palabras Clave

Tabla Sentimientos Dirigidos

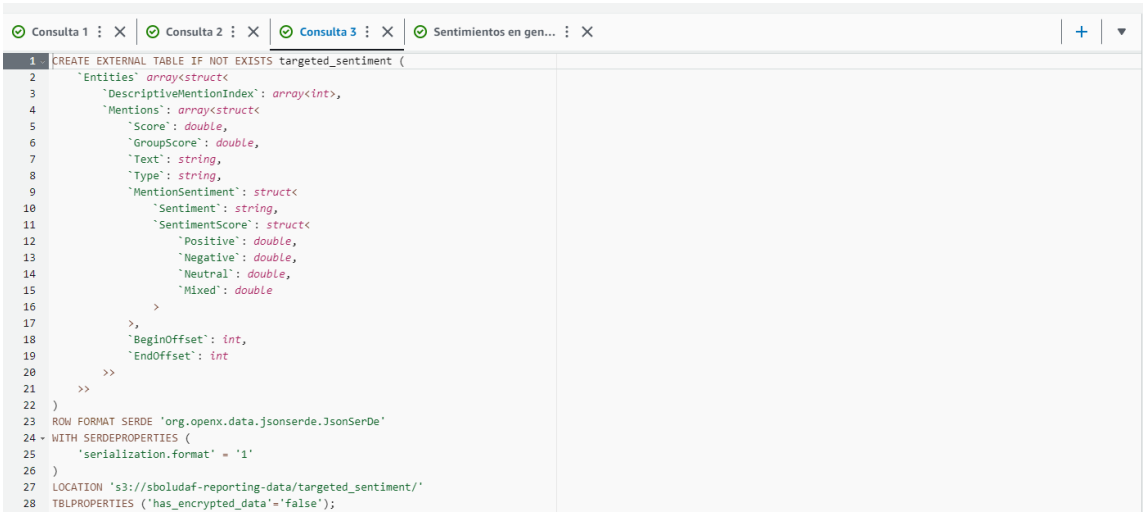

Ilustración 14. Creación Tabla Sentimientos Dirigidos

3.4.3.Procesamiento de Datos

El procesamiento de datos es un paso fundamental para el correcto funcionamiento del proyecto ya que los datos en bruto habitualmente necesitan ser tratados para poder ser consumidos. En este proyecto, el procesamiento de datos se ha realizado principalmente con AWS Glue DataBrew y AWS Lambda.

AWS Glue DataBrew

Para poder utilizar AWS Glue DataBrew se ha creado un nuevo proyecto apuntando al conjunto de datos de origen que se encuentran en el bucket RAW.

Ilustración 15. Dashboard de Trabajo AWS Glue DataBrew

Se han realizado modificaciones de forma visual, las opciones que ayudaban a la estandarización son:

- Dar formato a la columna en minúsculas
- Eliminar palabras de parada y Ampliar las contracciones
- Tokenizar el texto

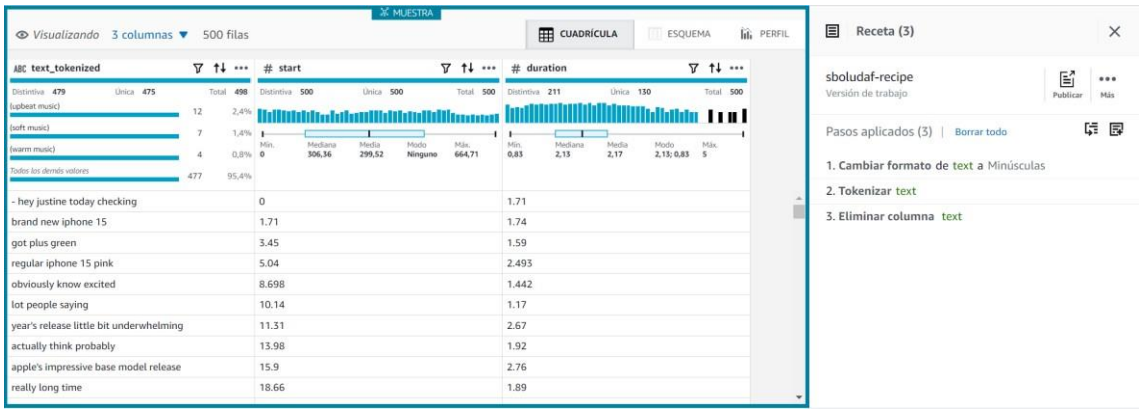

Ilustración 16. Cambios realizados AWS Glue DataBrew

Esto ayuda a la estandarización para su posterior procesamiento a través del algoritmo, esto se traduce directamente a una mayor rapidez y eficiencia.

Estos datos se almacenarán posteriormente en un bucket SANITIZED en formato CSV para que posteriormente se puedan procesar por AWS Lambda,

el objetivo inicial fue cambiarlos a parquet, pero al no tratarse de grandes cantidades de datos, se toma la decisión de mantenerlo en el formato de origen.

AWS Lambda

Las funciones de AWS Lambda pueden configurarse para activarse automáticamente ante una respuesta a ciertos eventos como si fuera una ETL, los eventos que consideramos relevantes son la llegada de nuevos datos a un bucket S3, este enfoque garantiza que el procesamiento de datos será ejecutado tan pronto como los datos sean subidos.

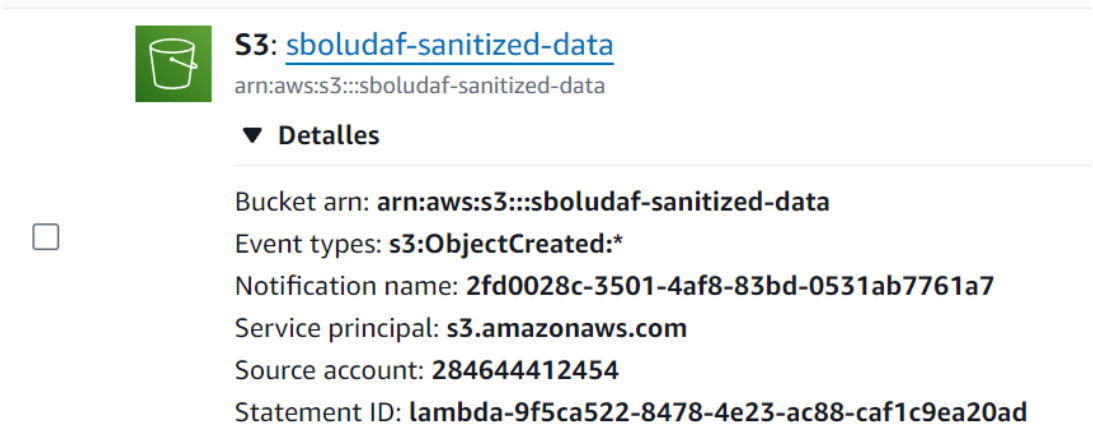

Ilustración 17. Trigger AWS Lambda

La función Lambda extraerá los datos de los bucket S3, que realiza las transformaciones necesarias a través de las librerías, se trabajará con el código de la lambda, para unificar estas reviews utilizaremos las librerías:

- **Pandas & Numpy:** Para trabajar sobre el procesamiento de datos como si se tratara de un Dataframe
- **Boto 3:** Para trabajar con la API de AWS
- **Io:** Para trabajar lectura y guardado de ficheros (Entrada / Salida)

Estas librerías no están incluidas dentro del servicio de AWS Lambda, para hacer esto posible, es necesario trabajar con *layers,* estas capas se utilizan para subir las librerías que serán utilizadas por la función, para disponibilizarlas ha sido necesario descargarlas en local a través de PIP y añadirlos a una carpeta llamada Python para posteriormente subirlo a AWS.

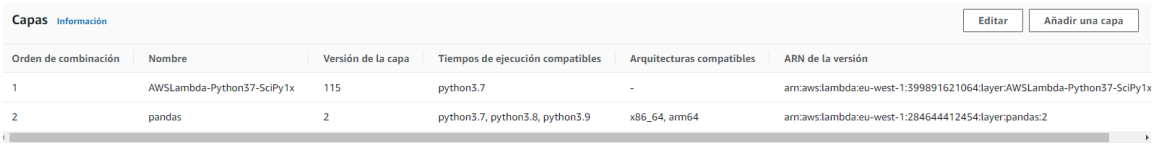

Ilustración 18. Layer Configuradas para AWS Lambda

En este caso se importan dos layers a la función, una que contiene las librerías necesarias para la función y otra que se encargará de solucionar los problemas de compatibilidad entre Numpy y AWS.

Se crean dos funciones lambdas basadas en **Python 3.8** a través de una arquitectura **x86_64** los permisos de esta función estarán declarados **en un rol que se generará en tiempo de despliegue** de la función lambda.

- **unifyLambda:** Se encarga de unificar las frases de la transcripción de reviews en una única review completa, al realizar estas acciones, obtenemos que existen 9 reviews completas en el fichero descargado de Kaggle.

- **sentimentLambda**: Se encarga de recuperar las reviews completas y procesarlas a través de AWS Comprehend para obtener los 3 atributos que serán analizados (Sentimientos, Sentimientos Dirigidos y Palabras Clave).

AWS Lambda escala automáticamente manejando el aumento en el volumen de datos sin necesidad de intervención manual, además, se configuran las políticas de IAM para garantizar que las funciones Lambda tengan los permisos necesarios para realizar su trabajo, minimizando el riesgo de acceso no autorizado a los datos.

3.4.4. Análisis de sentimientos

Este proceso se lleva a cabo a través de AWS Comprehend, un servicio de procesamiento de lenguaje natural (NLP) y machine learning (ML) que proporciona análisis de sentimientos de forma amigable y eficiente.

Después del procesamiento inicial, la limpieza y estandarización de los datos con AWS Lambda, los textos de las reviews se enviarán a AWS Comprehend para el análisis de sentimientos, de esta forma se puede obtener el tono emocional de las reviews.

Amazon Comprehend utiliza modelos de NLP para analizar el texto y proporcionar puntuación de cada uno de los sentimientos identificados, además, soporta diferentes idiomas.

Los resultados del análisis de sentimientos se reciben en un formato estructurado en JSON, estos datos de salida serán nuevamente almacenados en S3 en el bucket REPORTING para su posterior análisis y visualización. Los datos están preparados y optimizados para consultas, los insights obtenidos serán visualizados en AWS QuickSight permitiendo a los desarrolladores ver y entender las tendencias y patrones de las reviews.

Este proceso está totalmente automatizado a través de AWS Lambda y su triggers, ambos servicios están preparados para manejar un gran volumen

de datos, lo que garantiza que el sistema utilizado para el proyecto sea escalable y pueda adaptarse a un crecimiento.

AWS Comprehend está totalmente capacitado para integrarse con cualquier aplicación mediante llamadas a la API, en este caso a usar AWS Lambda y la librería Boto3, ha sido sencilla su integración. El api devolverá el siguiente resultado en función de la llamada, algunos ejemplos son:

Sentimientos


```
{
    "Sentiment": {
        "Sentiment": "NEUTRAL",
        "SentimentScore": {
            "Positive": 0.1020817756652832,
            "Negative": 0.19099242985248566,
            "Neutral": 0.5625510215759277,
            "Mixed": 0.14437474310398102
        }
    }
}
```


Palabras Clave


```
{
    "KeyPhrases": [
        {
            "Score": 0.9310188293457031,
            "Text": "Zhang Wei",
            "BeginOffset": 6,
            "EndOffset": 15
        },
        {
            "Score": 0.9990444779396057,
            "Text": "John",
            "BeginOffset": 22,
            "EndOffset": 26
        },
        {
            "Score": 0.9882776141166687,
            "Text": "Your AnyCompany Financial Services",
            "BeginOffset": 28,
            "EndOffset": 62
        },
        {
            "Score": 0.9982932806015015,
            "Text": "a great experience",
            "BeginOffset": 559,
```

```
            "EndOffset": 577
        }
    ]
}
```


Sentimientos Dirigidos


```
{
    "Entities": [
        {
            "DescriptiveMentionIndex": [
                1
            ],
            "Mentions": [
                {
                    "Score": 0.9447450041770935,
                    "GroupScore": 0.5628640055656433,
                    "Text": "your",
                    "Type": "PERSON",
                    "MentionSentiment": {
                        "Sentiment": "NEUTRAL",
                        "SentimentScore": {
                            "Positive": 0,
                            "Negative": 0,
                            "Neutral": 1,
                            "Mixed": 0
                        }
                    },
                    "BeginOffset": 175,
                    "EndOffset": 179
                },
                {
                    "Score": 0.999970018863678,
                    "GroupScore": 0.9995319843292236,
                    "Text": "John",
                    "Type": "PERSON",
                    "MentionSentiment": {
                        "Sentiment": "NEUTRAL",
                        "SentimentScore": {
                            "Positive": 0,
                            "Negative": 0,
                            "Neutral": 1,
                            "Mixed": 0
                        }
                    },
                    "BeginOffset": 22,
                    "EndOffset": 26
```

```
                },
                {
                    "Score": 0.9999939799308777,
                    "GroupScore": 0.5046669840812683,
                    "Text": "I",
                    "Type": "PERSON",
                    "MentionSentiment": {
                        "Sentiment": "NEUTRAL",
                        "SentimentScore": {
                            "Positive": 0,
                            "Negative": 0,
                            "Neutral": 1,
                            "Mixed": 0
                        }
                    },
                    "BeginOffset": 424,
                    "EndOffset": 425
                }
            ]
        },
        {
            "DescriptiveMentionIndex": [
                0
            ],
            "Mentions": [
                {
                    "Score": 0.9999470114707947,
                    "GroupScore": 1,
                    "Text": "service",
                    "Type": "ATTRIBUTE",
                    "MentionSentiment": {
                        "Sentiment": "POSITIVE",
                        "SentimentScore": {
                            "Positive": 0.9999970197677612,
                            "Negative": 0,
                            "Neutral": 9.999999974752427e-7,
                            "Mixed": 0.000001999999949504854
                        }
                    },
                    "BeginOffset": 538,
                    "EndOffset": 545
                }
            ]
        }
    ]
}
```

Estos datos, a pesar de que no son los reales de las reviews (estos se pueden encontrar en el bucket REPORTING) sirven para poder trabajar con Athena para crear las tablas y visualizar los datos correctamente.

3.4.5. Consultas y Análisis

Este proceso se realiza única y exclusivamente a través del servicio de AWS Athena, se trata de un servicio de consulta interactiva que permite realizar análisis complejos directamente sobre los datos que están almacenados en AWS S3.

AWS Athena se encuentra configurado para acceder a los datos almacenados en S3, incluyendo los datos en bruto, como procesados y los preparados, pero en este punto nos enfocaremos única y exclusivamente sobre los datos preparados.

Se utilizan esquemas de datos para definir las estructuras de los datos que hay almacenados en S3 y poder realizar consultas sobre ellos, además esto facilita la realización de consultas SQL, para que esto sea posible, se ha creado una tabla por cada tipo de dato:

Sentimientos

```
CREATE EXTERNAL TABLE IF NOT EXISTS sentiment (
    `Sentiment` string,
    `SentimentScore` struct<
        `Positive`: double,
        `Negative`: double,
        `Neutral`: double,
        `Mixed`: double
    >,
    `ResponseMetadata` string
)
ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe'
WITH SERDEPROPERTIES (
    'serialization.format' = '1'
)
LOCATION 's3://sboludaf-reporting-data/sentiment/'
TBLPROPERTIES ('has_encrypted_data'='false');
```

Sentimientos Dirigidos

```
CREATE EXTERNAL TABLE IF NOT EXISTS targeted_sentiment (
    `Entities` array<struct<
        `DescriptiveMentionIndex`: array<int>,
        `Mentions`: array<struct<
            `Score`: double,
            `GroupScore`: double,
```

```
            `Text`: string,
            `Type`: string,
            `MentionSentiment`: struct<
                `Sentiment`: string,
                `SentimentScore`: struct<
                    `Positive`: double,
                    `Negative`: double,
                    `Neutral`: double,
                    `Mixed`: double
                >
            >,
            `BeginOffset`: int,
            `EndOffset`: int
        >>
    >>
)
ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe'
WITH SERDEPROPERTIES (
    'serialization.format' = '1'
)
LOCATION 's3://sboludaf-reporting-data/targeted_sentiment/'
TBLPROPERTIES ('has_encrypted_data'='false');
```

Palabras Clave

```
CREATE EXTERNAL TABLE IF NOT EXISTS key_phrases (
    `KeyPhrases` array<struct<
        `Score`: double,
        `Text`: string,
        `BeginOffset`: int,
        `EndOffset`: int
    >>,
    `ResponseMetadata` string
)
ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe'
WITH SERDEPROPERTIES (
    'serialization.format' = '1'
)
LOCATION 's3://sboludaf-reporting-data/key_phrases/'
TBLPROPERTIES ('has_encrypted_data'='false');
```

Gracias a la importación de estos datos en Athena, se pueden desarrollar consultas SQL personalizadas para responder a las preguntas que el proyecto requiere, como identificar tendencias de la percepción del producto, sentimientos aparecidos…. Estos insights permite realizar análisis ad-hoc

donde los usuarios pueden ejecutar consultas exploratorias según sea necesario.

AWS Athena optimiza las consultas para mejorar el rendimiento y reducir los costes alineado con la estrategia medioambiental de reducción de capacidad del cómputo al máximo.

Estas consultas están totalmente integradas con la herramienta de visualización de AWS QuickSight para poder crear dashboards interactivos, estos dashboards pueden compartir para tomar decisiones empresariales.

Athena, al tratarse de un servicio serverless, es totalmente escalable y gestionado por AWS, es efectivo para manejar grandes volúmenes de datos, lo que es esencial en un entorno de Big Data.

3.4.6. Visualización de Datos

Este paso transforma los insights obtenidos desde AWS Comprehend y sus resultados en representaciones visuales comprensibles, de esta forma, podemos transformar las tablas resultantes de AWS Athena en dashboards interactivos que contienen reportes detallados.

AWS QuickSight se configura para integrarse con los datos de S3, es necesario establecer conexiones con las fuentes de datos relevantes, en este caso son los sentimientos, palabras clave y sentimientos dirigidos obtenidos de las reviews obtenidas desde AWS Comprehend.

3.4.7. Creación de Dashboards

Con AWS Quicksight, se pueden diseñar dashboards personalizados que reflejarán las métricas y KPIs claves para la obtención de conclusiones del análisis, estos dashboards interactivos permitirán filtrar, ordenar y profundizar en los datos para explorar diferentes aspectos de las reviews de los usuarios.

Se pueden utilizar diversos tipos de gráficos y visualizaciones acordes a los datos que se están representando, como gráficos de barras, mapas de calor, nubes de palabras…. De esta forma se podrán identificar rápidamente las opiniones generales de los usuarios.

Los dashboards pueden ser compartidos dentro de la organización, proporcionando un acceso realmente fácil a diferentes equipos o personas, el acceso está basado en roles utilizando AWS IAM para controlar quién puede ver o interactuar con los dashboards generados.

En este dashboard final se enfatiza en diseñar visualizaciones claras y fáciles de entender, evitando sobrecargar los dashboards con demasiada información que pueda ser abrumador para la persona que esté visualizando.

Esto, además, debe llevar una narrativa que cuente una historia o resalten los insights clave, facilitando la compresión y la tomas de decisiones basadas en los datos.

Por último, se muestran los orígenes de datos realizados a través de consultas SQL, que están configurados en AWS Quicksight para obtener los datos relevantes y poder realizar los Dashboards interactivos que el proyecto requiere:

Palabras clave más comunes y su puntuación promedio

```
SELECT kp.Text, AVG(kp.Score) AS AvgScore, COUNT(*) AS Count
FROM key_phrases
CROSS JOIN UNNEST(KeyPhrases) AS t (kp)
GROUP BY kp.Text
ORDER BY Count DESC, AvgScore DESC
```

Distribución de Puntuaciones de las Palabras Clave

```
SELECT ROUND(KeyPhrase.Score, 1) as ScoreRange, COUNT(*) as Count
FROM key_phrases
CROSS JOIN UNNEST(key_phrases.KeyPhrases) as t (KeyPhrase)
GROUP BY ROUND(KeyPhrase.Score, 1)
ORDER BY ScoreRange
```

Palabras Clave con Puntuaciones Altas

```
SELECT KeyPhrase.Text, KeyPhrase.Score
FROM key_phrases
CROSS JOIN UNNEST(key_phrases.KeyPhrases) as t (KeyPhrase)
WHERE KeyPhrase.Score > 0.9
ORDER BY KeyPhrase.Score DESC
```

Conteo de Reseñas por Sentimiento

```
SELECT mention.mentionsentiment.sentiment, COUNT(*) as TotalMentions
FROM targeted_sentiment
CROSS JOIN UNNEST(targeted_sentiment.entities) AS t(entity)
CROSS JOIN UNNEST(entity.mentions) AS t(mention)
GROUP BY mention.mentionsentiment.sentiment
```

Promedio de Puntajes de Sentimiento por Tipo de Sentimiento

```
SELECT mention.mentionsentiment.sentiment,
       AVG(mention.mentionsentiment.sentimentscore.positive) as
AvgPositiveScore,
       AVG(mention.mentionsentiment.sentimentscore.negative) as
AvgNegativeScore,
```

```
       AVG(mention.mentionsentiment.sentimentscore.neutral) as
AvgNeutralScore,
       AVG(mention.mentionsentiment.sentimentscore.mixed) as
AvgMixedScore
FROM targeted_sentiment
CROSS JOIN UNNEST(targeted_sentiment.entities) AS t(entity)
CROSS JOIN UNNEST(entity.mentions) AS t(mention)
GROUP BY mention.mentionsentiment.sentiment
```

Frecuencia de Menciones por Tipo de Entidad

```
SELECT mention.type, COUNT(*) as TotalMentions
FROM targeted_sentiment
CROSS JOIN UNNEST(targeted_sentiment.entities) AS t(entity)
CROSS JOIN UNNEST(entity.mentions) AS t(mention)
GROUP BY mention.type
```

Distribución General de Sentimientos

```
SELECT Sentiment, COUNT(*) AS Count
FROM sentiment
GROUP BY Sentiment
```

Promedio de Puntuaciones de Sentimiento

```
SELECT Sentiment, AVG(SentimentScore.Positive) AS AvgPositive,
AVG(SentimentScore.Negative) AS AvgNegative, AVG(SentimentScore.Neutral)
AS AvgNeutral, AVG(SentimentScore.Mixed) AS AvgMixed
FROM sentiment
GROUP BY Sentiment
```

3.4.8. Gestión de Identidades y acceso

En AWS, la gestión de identidades y el acceso se lleva a cabo a través de AWS IAM, donde se ha configurado un usuario para el acceso docente que pueda visualizar en tiempo real el estado de la solución, así como acceder a la plataforma de AWS.

Las políticas de seguridad que se han desarrollado para las funciones Lambda permiten el acceso a AWS S3 y AWS Comprehend para poder llevar a cabo toda la operativa necesaria para el proyecto.

El operador ha estado desplegando con unas credenciales personales donde se ha reforzado el acceso a través de una autenticación multifactor para añadir una capa adicional de seguridad a las cuentas de IAM.

Además, AWS CloudTrail es un servicio que mantiene un registro de todas las actividades y solicitudes realizadas a través de IAM, lo cual es fundamental para la auditoría de seguridad o el análisis forense en caso de una

incidencia. Como, por ejemplo, una mala ejecución de una lambda que elimine registros del bucket de sanitized.

Todo el proyecto se ha desarrollado con una metodología *credential less* esto significa que para cualquier acceso o comunicación entre los servicios no es necesario un usuario habitual con contraseña y contraseña, al utilizar roles, se evita la fuga o rotado habitual de las contraseñas.

3.4.9. Pruebas y validación

Para verificar la correcta integración y comunicación entre los diferentes servicios de AWS utilizados durante el desarrollo del proyecto como S3, Lambda, Comprehend, Athena y Quicksight se han realizado las siguientes pruebas.

Pruebas de integración

Las lambdas que interactúan con el bucket S3 llevan asignadas unas Policies administradas por AWS que permiten el acceso al servicio de AWS S3, esto permite acceder a los datos de forma íntegra.

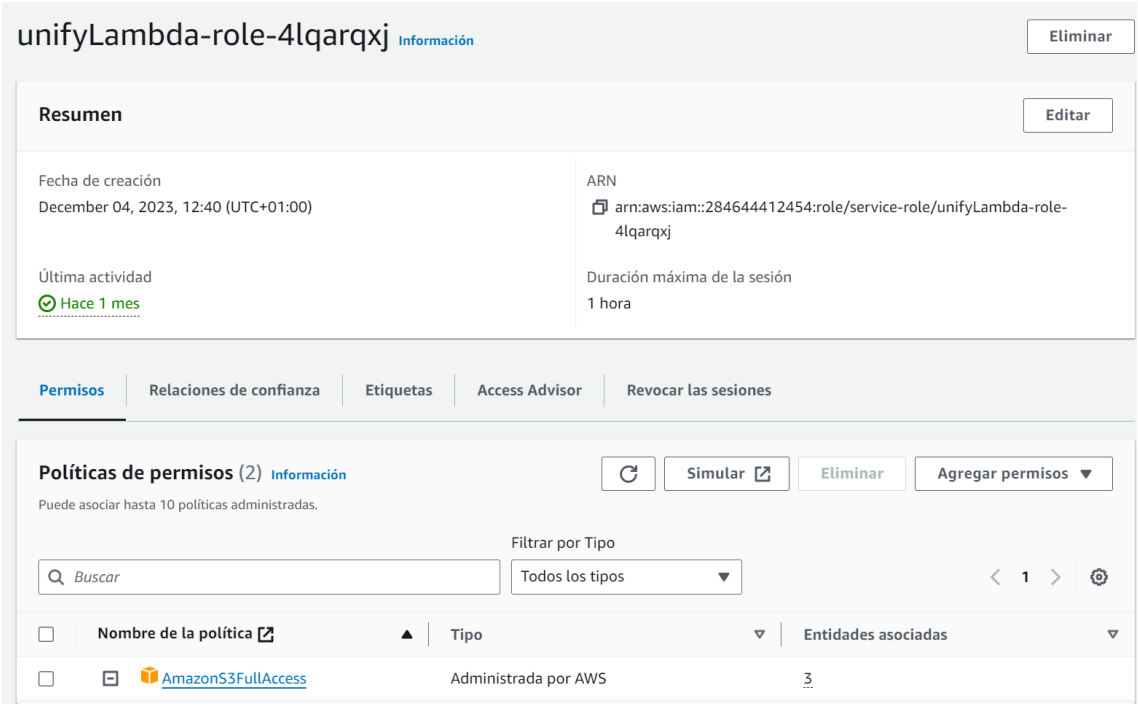

Ilustración 19. Rol Configurado AWS Lambda

De forma adicional, se ha configurado una línea en la función que se ejecuta y escribe en el log de AWS si todo ha funcionado correctamente:

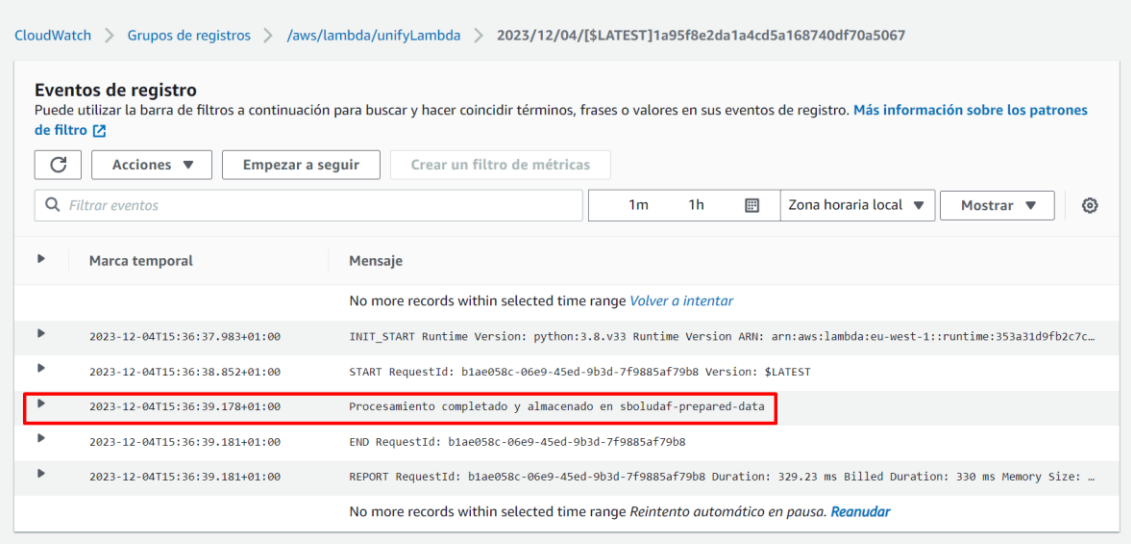

Ilustración 20. Eventos AWS Lambda

De esta forma, nos aseguramos de que todas las líneas han sido ejecutadas y el procesamiento ha funcionado correctamente.

La función lambda sentimentLambda se encarga de obtener los datos de S3 y pasarlos a través del algoritmo de AWS Comprehend, esto al tratarse de llamadas a la API de AWS, estarán recogidas en AWS CloudTrail

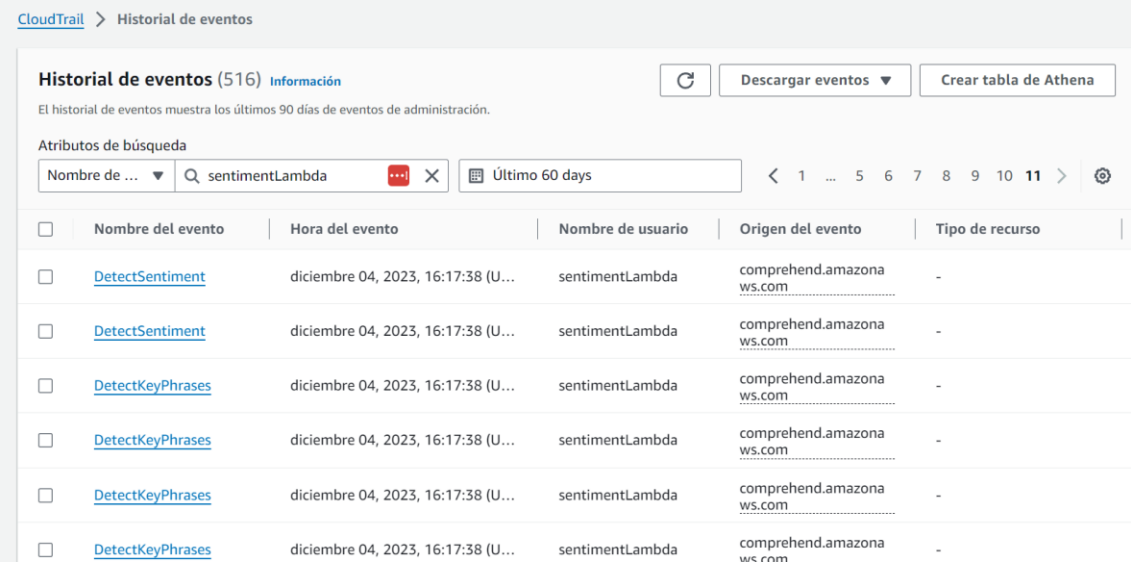

Ilustración 21. Logs de Eventos AWS CloudTrail

Donde se visualiza que la lambda está interactuando con la API de AWS Comprehend y se encarga de ejecutar el análisis de sentimientos para posteriormente almacenarlos en el bucket S3 REPORTING, donde se almacena un fichero por cada llamada a la API.

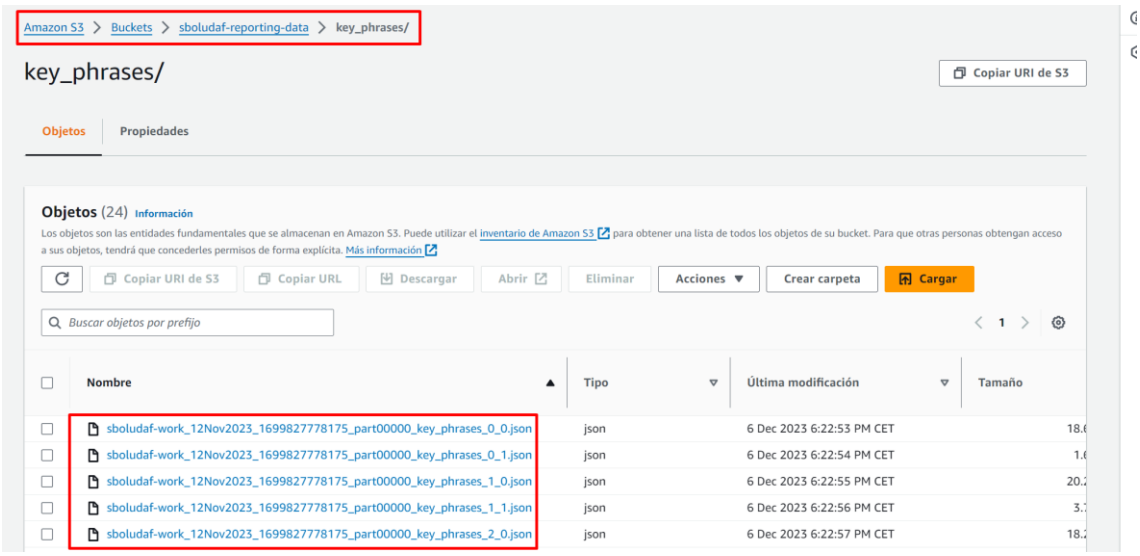

Ilustración 22. Datos Almacenados Bucket S3 REPORTING

Para comprobar la integración de AWS Quicksight con Athena y S3 se debe crear un conjunto de datos que será utilizado para visualizar los datos a través del dashboard.

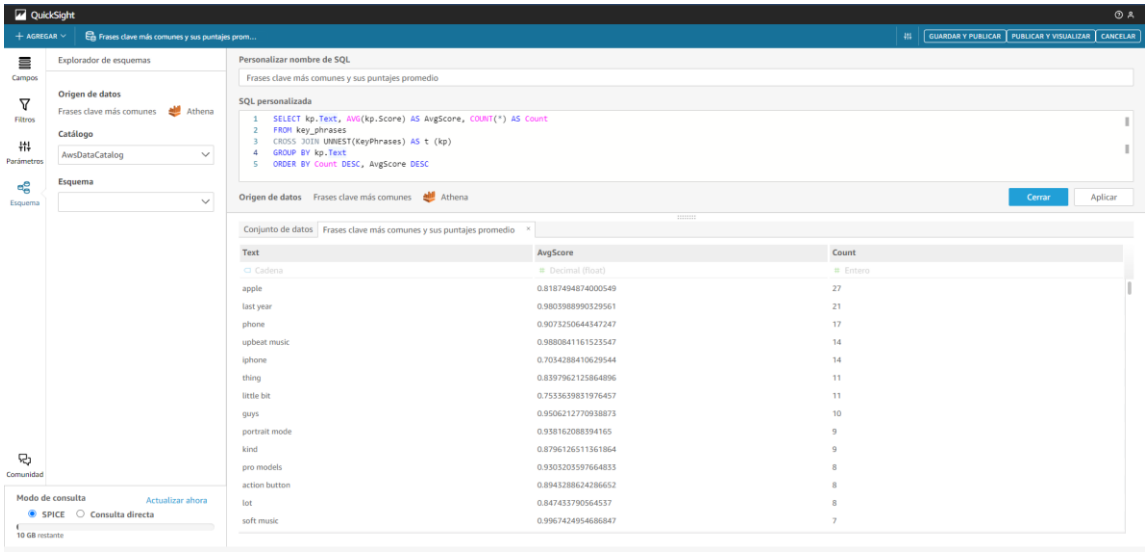

Ilustración 23. Conjuntos de datos configurados en AWS QuickSight

Pruebas de Funcionalidad

Para corroborar la correcta ejecución de las funciones lambda, se muestra el log de ejecución sin fallos.

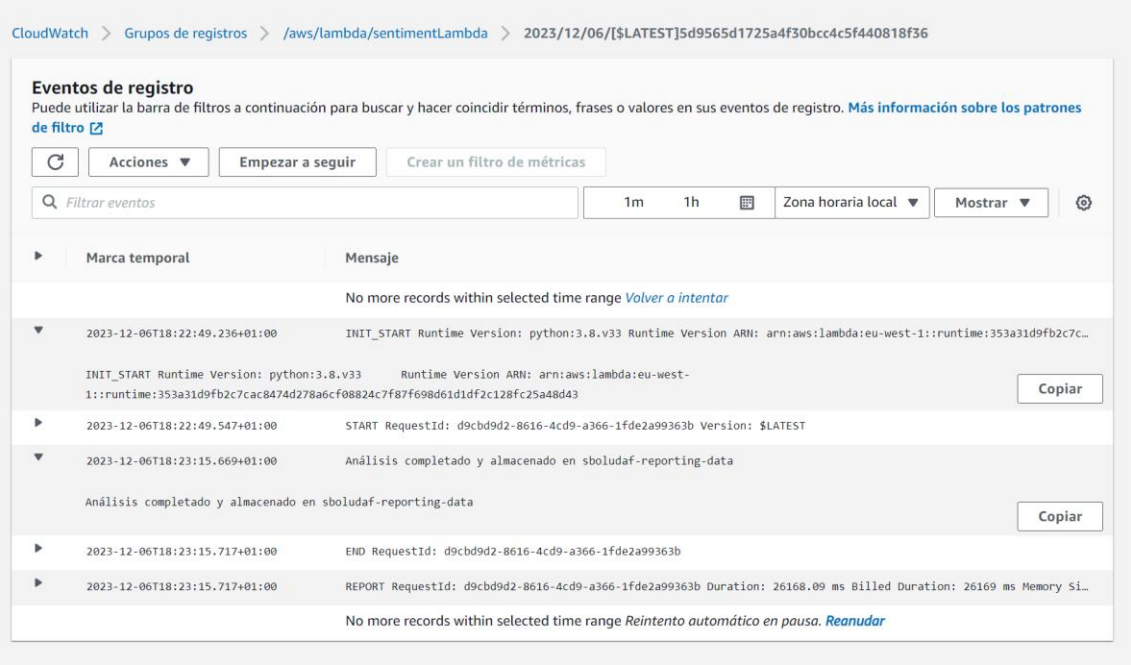

Ilustración 24. Log Ejecución AWS Lambda

En este caso, se muestra que esta función lambda ha interactuado correctamente con Amazon Comprehend y AWS S3.

Una vez creadas las tablas en AWS Athena mencionadas anteriormente, se pueden realizar consultas sobre las tablas para obtener algún dato de referencia para contrastar la información

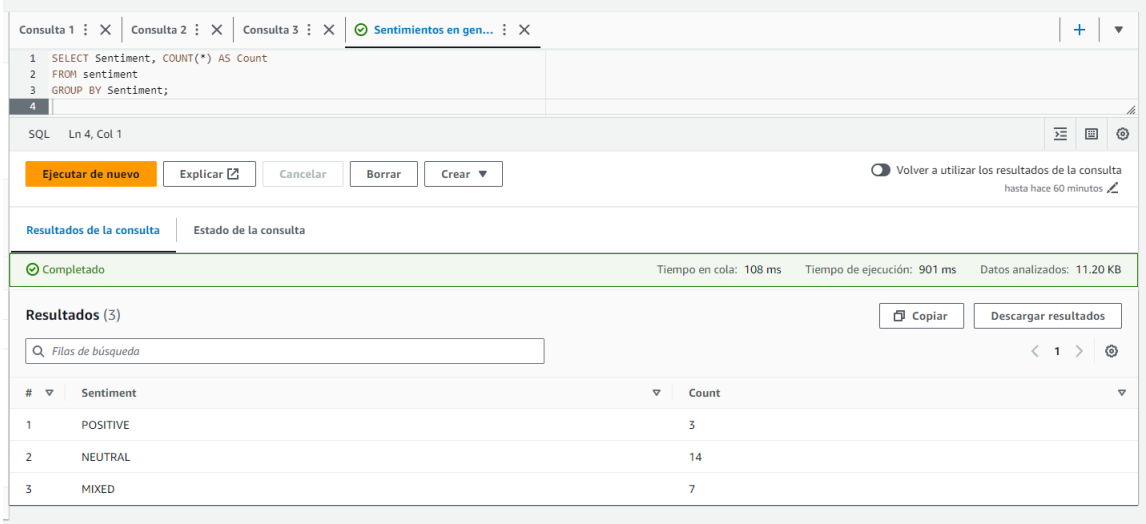

Ilustración 25. Comprobación datos en AWS Athena

De esta forma se asegura la correcta funcionalidad del entorno ya que está totalmente automatizado el fluo end-to-end de la solución del proyecto de BI.

Validación de datos

Para verificar la precisión y la calidad de los datos es necesario trabajar ya directamente sobre consultas de AWS Athena, estas arrojan información valiosa sobre el estado del arte del proyecto así como la información disponibilizada para su consumo. Algunas consultas que pueden realizarse para comprobar la veracidad de los datos son:

Temas más relevantes obtenidos de los sentimientos dirigidos

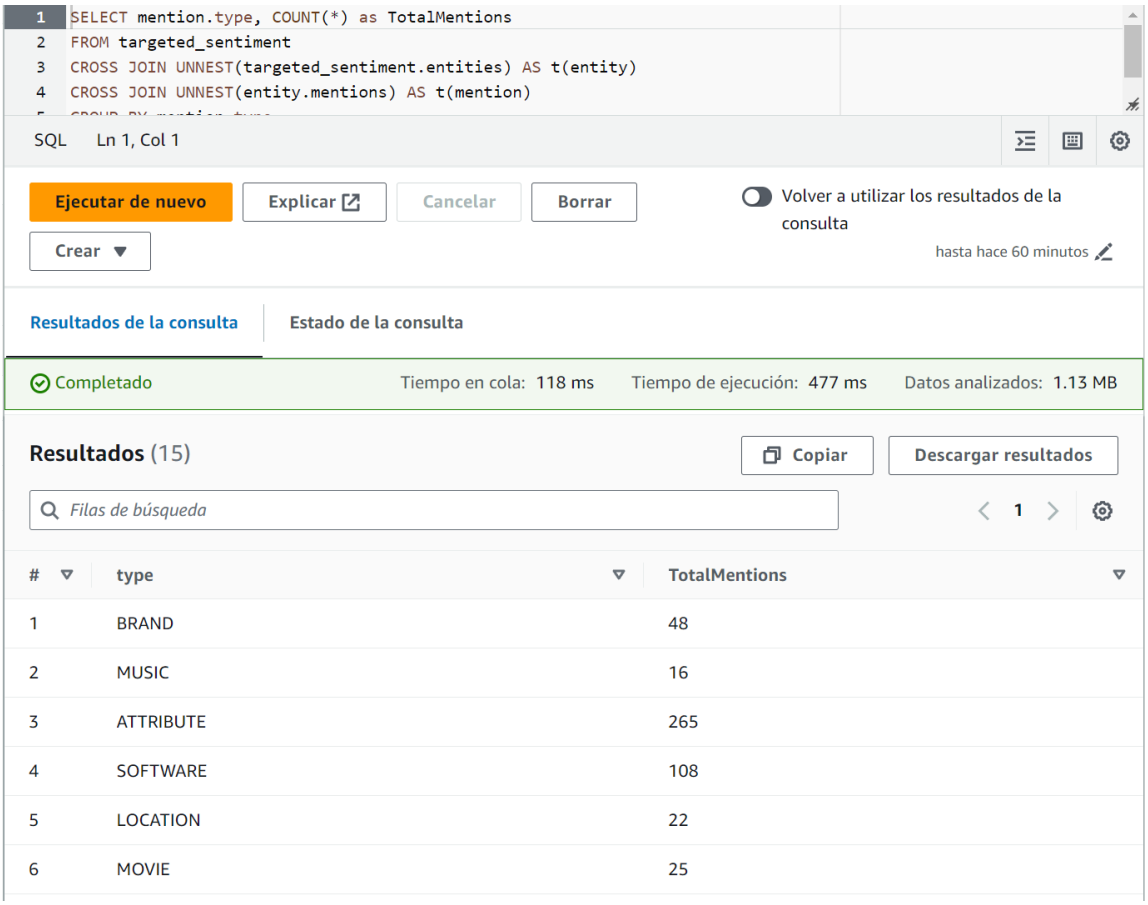

Ilustración 26. Consulta Sobre la Tabla de Sentimientos Dirigidos en AWS Athena

La fase de pruebas y validación es fundamental para asegurar que el proyecto no solo funcione según lo previsto, sino que también sea robusto, seguro y confiable. Esta fase debe ser meticulosa y adaptarse continuamente a medida que el sistema evoluciona y se desarrolla, garantizando así la calidad y la eficacia del proyecto.

4. Resultados y productos obtenidos

Se ha obtenido un Dashboard adjunto en formato PDF, el cual se pueden visualizar los siguientes diagramas que hacen referencia a los siguiente:

- **Palabras más mencionadas:** Destaca las palabras y frases más recurrentes en las reseñas, esto proporciona una idea sobre los temas más discutidos o que más aprecian los usuarios sobre el producto.

- **Puntuaciones de las frases clave:** Muestra la fiabilidad de las palabras clave identificadas, haciendo referencia que cuanto más cerca esté la puntuación a 1, más fiable es la palabra identificada como clave.

- **Distribución de palabras clave con puntuaciones altas:** Muestra la cantidad de palabras clave que aparecen identificadas con una puntuación alta superior a 0,9.

- **Frecuencia de menciones por tipo de entidad:** Muestra las entidades que más se han encontrado a la hora de identificar los sentimientos, es decir, identifica si un sentimiento o palabra hace referencia al software, precio, atributos del dispositivo…

- **Promedio de puntuación de sentimientos dirigidos / sentimientos agrupados por tipo de sentimiento:** Analiza el promedio de los puntajes de sentimiento (positivo, negativo, neutral, mixto) para cada tipo de sentimiento.

- **Distribución de sentimientos:** Muestra la distribución de sentimientos (positivo, negativo, neutral, mixto) entre las reseñas. Permite identificar la tendencia emocional general de los usuarios hacia el iPhone 15.

De los cuales se pueden obtener las siguientes conclusiones:

- Las palabras que más mencionan los usuarios son; el botón de acción, el modo retrato, la isla dinámica, la cámara y la batería. **Los usuarios hacen hincapié a las nuevas novedades del nuevo dispositivo de Apple que hacen diferenciarlo de su antecesor, dejando al margen otros elementos que son menos relevantes y los usuarios no les dan tanta importancia.**

- Las puntuaciones obtenidas de las palabras clave es bastante alta, **esto demuestra la buena calidad de los datos tratados, así como la fiabilidad del modelo NLP de clasificación de AWS Comprehend sobre los datos que estamos analizando.**

- Las palabras que tienen un mayor accuracy son similares a las más mencionadas, **esto demuestra que se los sentimientos obtenidos sobre las palabras identificadas son fiables.**

- Las entidades identificadas con mayor puntuación son: Other, Cantidad, Elemento Comercial, Atributos y cabe a destacar también

Software. **Eso demuestra que los usuarios en sus reviews prestan gran atención a elementos de los atributos del teléfono, elementos de cantidad como MPX de la cámara o elementos de** Software**.**

- En definitiva, en la distribución de los sentimientos hace especial aparición los sentimientos neutrales y positivos. **Esto significa que todas las palabras clave identificadas, así como los sentimientos identificados demuestran que los usuarios tienen en general reseñas positivas sobre dichos elementos, con una menor proporción en comentarios negativos.**

4.1. Conclusiones y trabajos futuros

Las conclusiones obtenidas a lo largo del proyecto hacen referencia a cada uno de los puntos en los que se ha ido trabajando a lo largo del TFG.

Insights de Sentimientos: Los resultados del análisis de sentimientos realizados por AWS Comprehend proporcionan una visión clara del tono general de las reviews de los usuarios sobre el nuevo iPhone 15, se puede obtener una propoción de opiniones positivas, negativas y neutrales.

Tendencias de Sentimientos: Gracias a los insights, se pueden obtener las tendencias de sentimientos a lo largo del tiempo, se pueden obtener diferentes reviews tras una actualización o a lo largo del año para comprender cómo las percepciones de los usuarios cambian.

Análisis de tendencias: Se identifican las tendencias más relevantes para los usuarios, así como los problemas frecuentes o los temas más comunes.

Comparaciones Temporales: Al tener un proceso del dato end-to-end claramente identificado, se puede comparar procesamientos de diferentes periodos de tiempo para ayudar a identificar si una campaña de marketing ha sido efectiva.

Dashboard interactivo: Un dashboard en QuickSight que muestra de forma clara las visualizaciones y los datos, facilitando a los stakeholders el acceso a insights valiosos de una manera más fácil de entender.

Decisiones Basadas en Datos: Gracias a este dashboard, los datos proveen una base sólida para la toma de decisiones de marketing o estrategias enfocadas al cliente.

Automatización de Procesos: La automatización de las funciones Lambda reduce el tiempo y el esfuerzo necesarios para obtener insights aumentando la eficiencia operativa.

Escalabilidad del sistema: El uso de servicios Serverless ofrece una escalabilidad y eficacia del sistema implementado, permitiendo al proyecto la capacidad de manejar grandes volúmenes de datos y adaptarse a futuros requerimientos de los stakeholders.

Los resultados obtenidos han sido sorprendentes ya que en función del dataset que había dispuesto para trabajar, a pesar de tener solamente 9 reviews, se han podido obtener un gran número de insights valiosos para hacer un MVP válido para presentar a los stakeholders.

No se han podido alcanzar todos los objetivos propuestos inicialmente ya que la API de Twitter pasó a ser de pago tras la obtención de la compañía por parte de Elon Musk, este cambio drástico ha cambiado el rumbo, cambiando el origen del dato de un día para otro. No obstante, los objetivos obtenidos con las transcripciones y el nuevo enfoque han sido totalmente satisfactorio.

La planificación se ha seguido correctamente, en algunas ocasiones se han entregado hitos que estaban propuestos para fechas más tardías que las entregadas inicialmente.

El impacto en la sostenibilidad ambiental del proyecto ha sido optimizado al máximo, se han utilizado todos los servicios sin servidor para pagar única y exclusivamente por uso, de esta forma existen ahorros tanto en computación como en coste.

Tras el inicio del proyecto, las vías futuras deben enfocarse en la ingesta del dato, así como su procesamiento, pudiendo ingestar datos de vídeos transcritos a través de una función lambda que se encargue de almacenar el texto en los bucket S3 y de esta forma enriquecer el dashboard con las opiniones de diversos usuarios de las diferentes plataformas de vídeos.

De forma adicional, al tratarse de un Dashboard, este puede seguir creciendo en función de las necesidades o de los objetivos propuestos por los stakeholders.

5. Glosario

AWS S3 (Simple Storage Service): Servicio de almacenamiento de objetos de Amazon Web Services (AWS).

AWS Glue DataBrew: Servicio de AWS para la preparación de datos a través de una interfaz visual e interactiva.

AWS Lambda: Servicio de AWS que permite ejecutar código sin servidor.

AWS Comprehend: Servicio de procesamiento de lenguaje natural (NLP) y machine learning (ML) de AWS.

AWS Athena: Servicio de consulta interactiva de AWS que permite el análisis de datos almacenados en AWS S3 utilizando SQL.

AWS QuickSight: Servicio de Inteligencia Empresarial de AWS para crear y compartir visualizaciones interactivas, dashboards y reportes.

AWS IAM (Identity and Access Management): Servicio de AWS para controlar de forma segura el acceso a los recursos de AWS.

Bucket: Contenedor en AWS S3 donde se almacenan datos como objetos.

Key (Clave): Identificador único para cada objeto almacenado en un bucket de AWS S3.

Serverless: Arquitectura que permite ejecutar aplicaciones y servicios sin necesidad de administrar infraestructura.

NLP (Natural Language Processing): Procesamiento de lenguaje natural, una rama de la inteligencia artificial que se enfoca en la interacción entre computadoras y lenguaje humano.

Machine Learning (ML): Campo de la inteligencia artificial que utiliza algoritmos y modelos estadísticos para que las máquinas mejoren su desempeño en una tarea específica.

ETL (Extract, Transform, Load): Proceso en el que los datos se extraen de una fuente, se transforman o procesan, y finalmente se cargan en un sistema de almacenamiento.

JSON (JavaScript Object Notation): Formato ligero de intercambio de datos, fácil de leer y escribir para humanos y fácil de analizar y generar para máquinas.

API (Application Programming Interface): Conjunto de reglas y definiciones que permiten que diferentes aplicaciones o software se comuniquen entre sí.

Big Data: Grandes conjuntos de datos que requieren tecnologías y métodos analíticos avanzados para su procesamiento y análisis eficiente.

6. Bibliografía

1. Amazon Web Services, Inc. "Amazon Simple Storage Service (S3)." Amazon Web Services, Inc., Fecha de acceso: [27 Dic 2023]. Disponible en: https://aws.amazon.com/es/s3/.

2. Amazon Web Services, Inc. "AWS Glue DataBrew – Features." Amazon Web Services, Inc. Fecha de acceso: [27 Dic 2023]. Disponible en: https://aws.amazon.com/es/glue/features/databrew/

3. Amazon Web Services, Inc. "AWS Lambda." Amazon Web Services, Inc. Fecha de acceso: [27 Dic 2023]. Disponible en: https://aws.amazon.com/es/lambda/

4. Amazon Web Services, Inc. "Amazon Comprehend." Amazon Web Services, Inc. Fecha de acceso: [27 Dic 2023]. Disponible en: https://aws.amazon.com/es/comprehend/

5. Amazon Web Services, Inc. "Amazon Athena." Amazon Web Services, Inc. Fecha de acceso: [27 Dic 2023]. Disponible en: https://aws.amazon.com/es/athena/

6. Amazon Web Services, Inc. "Amazon QuickSight." Amazon Web Services, Inc. Fecha de acceso: [27 Dic 2023]. Disponible en: Link Acortado

7. Amazon Web Services, Inc. "What is Data Preparation?" Amazon Web Services, Inc. Fecha de acceso: [04 Ene 2024]. Disponible en: https://aws.amazon.com/es/what-is/data-preparation/

7. Anexos

Debido a la metodología seguida por el proyecto, así como la contextualización y síntesis del documento, no es necesario adjuntar anexos que proporcionen información adicional más que la expuesta en este documento.