
SPARTAN
Data Science
PROYECTO
RestoTrends
Sprint #1

Tabla de contenido

Evaluación del Contexto Actual

En nuestra calidad de especialistas en análisis de datos, examinamos minuciosamente la información recopilada en sitios de reseñas como Yelp y Google Maps con el propósito de identificar patrones y tendencias provenientes de los comentarios de los usuarios para entender profundamente sus experiencias, necesidades y expectativas respecto a diversos servicios y negocios. Este ejercicio de evaluación de datos nos permite ofrecer consejos y sugerencias detalladas tanto a potenciales inversores interesados en este proyecto, como a empresarios existentes, siendo un recurso crucial para la toma de decisiones informadas y la mejora de la reputación y la percepción de los negocios en la mente de los clientes. Este propósito se logra mediante la implementación de métodos sofisticados de análisis de sentimientos y aprendizaje automático para ofrecer a nuestros clientes un sistema de recomendaciones eficaz y fácil de usar, así como información valiosa relacionada con tendencias y sentimientos de consumidores para la toma de decisiones informada por parte de inversionistas.

Objetivo principal

Desarrollar una plataforma comprensiva que se encargue de reunir, limpiar y presentar datos organizados y claros de comentarios de usuarios. Esto se logra mediante la realización de un análisis exploratorio, la creación de un sistema de recomendaciones fundamentado en el análisis de sentimientos, un modelo predictivo para evaluar el crecimiento empresarial y un panel interactivo para la visualización y exploración detallada de los resultados.

Objetivos Específicos

1. **Asegurar Datos Claros y Organizados:** Nos enfocamos en reunir, limpiar y presentar datos pertinentes de sitios de reseñas como Yelp y Google Maps. Este proceso garantiza la integridad y consistencia de los datos, facilitando su análisis y uso en el proyecto.

2. **Análisis Exploratorio de Comentarios de Usuarios:** Nuestra tarea es descifrar patrones, tendencias y elementos clave que afectan la percepción de los usuarios respecto a establecimientos y servicios, basándonos en los comentarios obtenidos de Yelp y Google Maps.

3. **Creación de un Modelo Avanzado de Aprendizaje Automático para Recomendaciones:** Nos proponemos desarrollar un modelo innovador que emplee el análisis de sentimientos para categorizar los comentarios y prever las preferencias de los usuarios. Esto facilitará la personalización de la experiencia del cliente, la implementación de estrategias a medida y el mejoramiento continuo de las empresas con el propósito de mejorar la percepción que sus negocios están dejando en sus clientes

4. **Desarrollo de un Modelo Predictivo:** Un segundo modelo de aprendizaje automático será elaborado para proporcionar a los empresarios predicciones precisas sobre la expansión de los negocios. Este modelo será una herramienta esencial para los inversores, permitiéndoles hacer elecciones informadas y estratégicas.

5. **Construcción de un Panel Interactivo:** Se diseñará una plataforma interactiva que ofrezca una visualización detallada de los resultados obtenidos del análisis de datos de las opiniones de los usuarios. Incluirá métricas esenciales, representaciones gráficas y estadísticas para asistir en la toma de decisiones y reconocer áreas de optimización en los servicios proporcionados.

Roles y Responsabilidades:

NOMBRE	ROL	RESPONSABILIDADES
Lucas Santos Oliveira	Data Engineer	Realizar una selección cuidadosa de las herramientas y tecnologías adecuadas que permitan cumplir con los requisitos y objetivos del proyecto, optimizando así la eficiencia y el rendimiento del sistema.
David Gonzalez	Data Engineer	Implementar medidas para garantizar la disponibilidad de datos limpios y estructurados, asegurando que la información utilizada esté completa, actualizada y libre de errores, lo que permitirá tomar decisiones informadas basadas en datos confiables y precisos.
Edgar Eduardo Barbero	Data Scientist	Desarrollar un modelo de machine learning altamente preciso y eficiente que utilice técnicas de recomendación para predecir con exactitud las preferencias individuales de los usuarios, con el objetivo de ofrecer recomendaciones personalizadas y mejorar su experiencia.
Carlos Eduardo Peña	Data Analyst	Desarrollar una interfaz interactiva que brinde la posibilidad de explorar de manera intuitiva los resultados del análisis de datos en las reseñas de los usuarios. La finalidad es facilitar la visualización y comprensión de los insights obtenidos, permitiendo una interacción dinámica con la información para una toma de decisiones más informada. Documentar el proyecto total para su entrega final
Adalber Conde Lucero	Data Analyst	Desarrollar una interfaz interactiva que brinde la posibilidad de explorar de manera intuitiva los resultados del análisis de datos en las reseñas de los usuarios. La finalidad es facilitar la visualización y comprensión de los insights obtenidos, permitiendo una interacción dinámica con la información para una toma de decisiones más informada.

Alcance

Realizar un desarrollo inicial a nivel local con solo una muestra de los datos disponibles con el propósito de desarrollar las funcionalidades necesarias y poder disponibilizar el producto final en un sistema en la nube que pueda manejar la totalidad de los datos en una etapa posterior. De la misma forma se construirá un tablero interactivo que pueda manejar cualquier tamaño de datos y genere la información relevante que finalmente va a consumir el cliente final

Dimensiones del Proyecto:

• **Duración del Proyecto:** El proyecto se extenderá por un período de tres semanas, con entregas programadas semanalmente, siguiendo la metodología SCRUM y se implementarán sprints semanales para organizar las tareas. La semana inicial se enfocará en el entendimiento del proyecto y de los datos disponibles planteando el proyecto a desarrollar, la segunda semana se enfocará en el manejo limpieza y disponibilización de los datos para su uso en el sistema propuesto y la tercera semana se dedicará al desarrollo de tableros finales, desarrollo de la función de Machine Learning y ajustes finales para la entrega del producto final

• **Presupuesto del Proyecto:** El proyecto se caracteriza por su enfoque en la practicidad y eficacia utilizando herramientas y recursos gratuitos, ya que no se dispone de un presupuesto para el desarrollo del mismo. El equipo de cinco profesionales, especializados en diferentes áreas del análisis de datos, trabajará de manera voluntaria en este proyecto no remunerado, maximizando el uso de recursos gratuitos para alcanzar los objetivos establecidos de manera efectiva.

• **Extensión Geográfica:** La iniciativa se enfocará en negocios dentro de la industria de la hospitalidad, gastronomía y ocio vinculados al sector de entretenimiento y turismo, ubicados en Estados Unidos.

• **Impacto del Proyecto:** La meta es influir positivamente tanto en los propietarios de negocios como en los consumidores, ofreciendo insights críticos para cada uno de los tipos de clientes. Para inversores, se ofrece información relevante relacionada con las percepciones de los consumidores tanto positivas como negativas para cada negocio en particular con el propósito que tengan elementos reales y actualizados que les permitan una mejor valoración de un negocio en particular. Para dueños de negocios, con base en la misma información ofrecida, les va a permitir inicialmente conocer la percepción de los consumidores en relación con su negocio y les permitirá identificar puntos clave en los cuales enfocar sus recursos y esfuerzos para mejorar su posicionamiento en la mente del consumidor. Finalmente, para público en general, se ofrece un sistema de recomendaciones de fácil acceso y uso que les permita tomar decisiones de consumo fundamentados en las opiniones uy experiencias de otros consumidores.

Análisis de los datos:

Después de revisar el diccionario de datos, determinamos que el archivo “review-estados” de la base de datos de “Google-Maps” contiene la información que vamos a tomar como punto de partida para el entendimiento de la información disponible, dado que cuenta con las opiniones de los usuarios de Google Maps. Al abrir y transformar la información relacionada a un archivo .csv podemos observar la cantidad de datos disponibles por cada estado, lo que nos da los parámetros necesarios para tomar la decisión del tamaño y tipo de muestra que vamos a seleccionar para hacer el desarrollo inicial.

Cada estado contiene información desde el año 1990 hasta el año 2021, pero los datos que aparecen clasificados durante los años 90 parece ser un contador de datos del dataset y no corresponden a datos reales de opiniones de usuarios, adicionalmente se aprecia claramente que los datos disponibles en la primera década del 2000 se cuentan solo en decenas de datos lo que es natural

pues era el inicio del internet y de los portales que permitían registrar las experiencias y opiniones de consumidores.

Durante la segunda década del 2000 el incremento datos disponibles es permanente y constante año tras año, hasta llegar al año 2020 en el cual se presenta la pandemia por covid19 y se genera un confinamiento obligatorio con el correspondiente impacto negativo en comercio, industria y la vida en general.

Del análisis anterior se toma la decisión de aplicar el desarrollo inicial para un periodo de tres años en los cuales halla mayor cantidad de datos con los cuales se pueda trabajar, este periodo corresponde a los años 2017, 2018 y 2019.

Adicionalmente se analiza la cantidad de datos disponibles por cada estado para este periodo de tres años y se crea un ranking por estado con la información disponible. Este ranking se deja como información disponible para ser cruzado con la información de Yelp y tomar decisiones relacionadas con la muestra a elegir para el desarrollo inicial.

De la misma forma, se evidencia que la información relacionada con las opiniones de consumidores incluye negocios de diferentes rubros y sectores de la industria y servicios, ante lo cual determinamos que tipo de rubro especifico es el que tiene más opiniones registra. De esta categorización se identifica que el rubro de "Restaurant" es el que más datos tiene en el la información disponible, pero adicionalmente se logra identificar que existen otros rubros similares y homólogos que podemos incluir dentro de este mismo rubro con el propósito de obtener un conjunto de datos más robusto. Con este análisis realizado, decidimos enfocarnos únicamente en negocios que se dediquen al negocio de comidas y bebidas servidas, y similares.

Por otro lado, después de abrir la información disponible en Yelp, se identifica que contiene información de negocios para 27 estados incluyendo el estado de Alberta que pertenece a Canada, país que no forma parte de nuestra extensión geográfica, con lo cual lo excluimos de nuestro espacio de trabajo, dando como resultado solo 26 estados sobre los cuales trabajar.

Este listado se ordena por cantidad de información creando un ranking de estados y se cruza con el ranking desarrollado previamente con la información proveniente de los datos de Google Maps, sumando la información disponible por cada estado en ambas bases de datos y generando un ranking definitivo con esta información final. De este ranking tomamos únicamente los 5 estados con mayor información proveniente de ambas bases de datos y lo tomamos como nuestro espacio muestral para el desarrollo inicial previsto; estos estados seleccionados son: California, Florida, Pensilvania, Illinois y New Jersey.

Planteamiento de KPI´s:

Un KPI es un indicador clave de gestión para una empresa (Key Performance Indicator) y esta orientado a medir los factores claves que determine la alta dirección para conocer el desempeño del negocio en comparación con los objetivos propuestos.

En nuestro caso en particular hemos definido tres tipos de clientes principales para el desarrollo de nuestra solución: Propietarios de negocios, inversionistas y usuarios en general. Teniendo en cuenta la definición de un KPI, debemos tener presente que el único tipo de cliente que tiene un negocio sobre el cual aplicar un KPI es el dueño del negocio. El inversor tiene la expectativa de adquirir un negocio, pero aún no lo tiene, y el usuario en general solo le interesa conocer un aspecto en particular del negocio, pero no tiene ninguna injerencia en el mismo, ante lo cual se define KPI´s para los dueños de negocios y METRICAS para inversores y usuarios en general.

KPI´s Owners

Nº	KPI	Descripción	Formula	Periodicidad	Objetivo
1	Porcentaje de clientes satisfechos en el negocio	Mide el porcentaje de clientes que están satisfechos con el negocio.	(Número de reviews positivos / Número total de reviews para el negocio) * 100	Mensual	Mantener el porcentaje de clientes satisfechos por encima del 80%
2	Porcentaje de clientes con una experiencia negativa en el negocio	Mide el porcentaje de clientes que han tenido una experiencia negativa con el negocio	(Número de reviews negativos / Número total de reviews para el negocio) * 100	Mensual	Mantener el porcentaje de experiencias negativas para el negocio por debajo del 2%
3	Promedio rating	Mide el promedio de numero de estrellas otorgadas por los usuarios.	Average stars	Mensual	Mantener el promedio de estrellas por encima de 4
4	Nivel de “Engagement” de clientes	Mide el nivel de interacción de los clientes con el negocio.	Suma total de reviews recibidos por periodo de tiempo. Se toma como periodo de comparación el total de reviews del año anterior	Mensual	Aumentar el nivel de “Engagement” de los reviews en un 12% anual o un 1% mensual

METRICAS Investors

Nº	METRICA	Descripción	Formula	Periodicidad	Objetivo
1	Top 5 de negocios con buen "review sentiment" y con valoración general media por población	Ranking de los negocios en los cuales los clientes tienen buen sentimiento hacia el mismo, pero tienen otros aspectos en los cuales se debe mejorar.	De los negocios con sentimiento positivo, se promedia el valor de los últimos 6 meses y se saca el top 5 de negocios con valoración hasta 3.5 estrellas	Mensual	Mostrar los 5 restaurantes con resultados históricos durante los últimos 6 meses de sentimiento positivo, pero valoración hasta 3.5 estrellas
2	Top 5 de Reviewers de YELP y GOOGLE con mayor actividad por poblacion	Mostrar en cada población el TOP 5 de los usuarios más activos por cada plataforma	Top 5 YELP y GOOGLE users con mayor cantidad de reviews por poblacion	Anual	Poner en contacto a Inversores con Reviewers de experiencia para que puedan hacer una crítica cruda al negocio que se busca evaluar

METRICAS Customers

Nº	METRICA	Descripción	Formula	Periodicidad	Objetivo
3	Top 5 de negocios con mayor valoración general, por categoría y por población	Ranking de los negocios mayor numero de estrellas para una categoría en específico y una población particular	Top 5 calificación general filtrado por categoría y por población	Mensual	Mostrar los 5 restaurantes con mayor calificación en estrellas por categoría y por población
4	Customer Review Metric (CRM). Sumatoría de 3 metricas: reviews, rating, sentiment	Obtiene un indicador promediando las 3 métricas descritas con un peso definido previamente para cada una	CRM = ((# reviews * 0.2) + (Average rating * 0.5) + (Sentiment Score * 0.3)) / 3	Semanal	Mostrar el CRM para cada negocio que el usuario desee consultar

Stack tecnológico

Desde la perspectiva de un Data Engineer, es necesario abordar conceptos clave, herramientas y decisiones relacionadas con el procesamiento y análisis de datos en entornos de clúster en la nube para lograr el objetivo propuesto. Como Data Engineer, es fundamental comprender los conceptos fundamentales antes de sumergirse en la práctica y tomar decisiones informadas sobre las herramientas a utilizar.

CONCEPTOS CLAVES

primero necesitamos entender algunos conceptos claves:

¿Qué es un Cluster?

Se refiere a un grupo de computadoras interconectadas que trabajan juntas para resolver tareas o procesar datos. Estas computadoras, llamadas nodos, colaboran para realizar operaciones complejas de manera más eficiente que una sola máquina. Los clusters se utilizan para abordar cargas de trabajo intensivas en procesamiento y almacenamiento de datos.

¿Qué son los piplines?

En el contexto del procesamiento de datos, un pipeline se refiere a una secuencia de etapas o pasos que se ejecutan en orden para realizar una tarea específica. En el análisis de datos, un pipeline puede consistir en la ingesta de datos, limpieza, transformación y análisis.

Relación con Clusters:

Los pipelines de procesamiento de datos pueden ser muy intensivos en recursos y, a menudo, se ejecutan en entornos de clúster. Esto significa que los pasos de un pipeline se distribuyen entre varios nodos de un clúster para acelerar el procesamiento y la capacidad de carga de trabajo.

¿Qué es Apache Spark?

Es un framework de procesamiento de datos de *código abierto* que se ha vuelto extremadamente popular en el análisis de datos y el procesamiento de big data. Spark proporciona un entorno de procesamiento distribuido que se ejecuta en “clústeres”, permitiendo el procesamiento rápido y escalable de grandes conjuntos de datos.

¿Cómo trabaja Spark con los clústers?

Apache Spark se ejecuta en clústeres de *máquinas interconectadas*. Utiliza un modelo de programación llamado Resilient Distributed Dataset (RDD) que permite distribuir automáticamente el procesamiento de datos en múltiples nodos del clúster. Esto hace que Spark sea muy eficiente para el procesamiento de datos a gran escala y lo convierte en una elección popular para aplicaciones que involucran pipelines de procesamiento de datos.

TRABAJO CON NUESTRA HERRAMIENTA CLOUD

Existen muchas herramientas para poder trabajar en la nube, estuvimos analizando trabajar con AWS pero, finalmente nos inclinamos con Google Platform Cloud (GCP) por distintas razones que consideramos muy beneficiosas:

- La red de Google: Esta es una de las razones más fuertes, ya que GCP cuenta con una de las redes más grandes y rápidas del mundo. Esto se traduce en una menor latencia y una mayor velocidad de transferencia. A demás podemos destacar que su interfaz a la hora de trabajar es muy "amigalbe".

- Capa Gratuita y pagos: Al momento de realizar el proyecto la plataforma tiene una capa gratuita muy interesante y transparente.

- Big Data y Análisis: GCP proporciona herramientas sólidas para el procesamiento y análisis de big data, como “BigQuery”, “Dataprep”, y “Dataflow”, que pueden ser utilizadas para manejar grandes volúmenes de datos de manera eficiente.

- Contenedores y Kubernetes: GCP es conocido por su soporte de contenedores y Kubernetes. Ofrece Google Kubernetes Engine (GKE) para la administración y orquestación de contenedores de manera sencilla.

- Sistemas de Almacenamiento: La plataforma ofrece soluciones de almacenamiento muy faciles de entender como Google Cloud Storage y Cloud SQL, que pueden ser utilizadas para una variedad de casos de uso, desde almacenamiento de objetos hasta bases de datos relacionales.

Es importante mencionar que muchas plataformas tienen beneficios y herramientas similares, mientras otras ofrecen menor variedad de opciones, pero finalmente se optó por GCP como decisión de grupo después del respectivo análisis de las ventajas ofrecidas.

TECNOLOGIAS A USAR

- Apache Spark en Google Cloud: En la plataforma de Google Cloud tenemos un servicio que nos permite crear clústeres de Apache Spark y Hadoop de manera rápida y sencilla. Podemos aprovechar las capacidades de Apache Spark para el procesamiento distribuido de datos.

Podemos crear un clúster de en google cloud y ejecutar aplicaciones de Apache Spark en él. Este clúster distribuirá *automáticamente* la carga de trabajo en sus nodos, lo que te permite procesar grandes volúmenes de datos de manera eficiente. Para la creación de un clúster en Google Cloud utilizamos la herramienta de Dataproc.

- Dataproc: Es un servicio que nos permite crear clústeres de Apache Spark y Hadoop de manera rápida y sencilla en la plataforma de Google Cloud. Podemos aprovechar las capacidades de Apache Spark para el procesamiento distribuido de datos.

Una las ventajas que tiene este cluster es que ya tiene preinstalada muchísimas imágenes que hace que su practicidad y uso sea mucho más ágil y sencillo.

- Dataflow: Es un servicio de procesamiento de datos que te permite construir y ejecutar pipelines de procesamiento de datos de manera escalable y gestionada. Dataflow se basa en Apache Beam, lo que permite crear pipelines de manera programática, estos pipelines nos podrían servir para la realización de diversas operaciones, como la ingesta de datos desde fuentes como Google Cloud Storage, la transformación de datos con Apache Spark, y la carga de resultados en destinos como BigQuery. Dataflow te permite definir y ejecutar estas operaciones de manera eficiente y escalable en un entorno de Google Cloud; adicionalmente tiene una abstracción más visual de definición de pipelines.

- Google Cloud Storage: Es un servicio de almacenamiento en la nube de Google que proporciona un lugar seguro y escalable para almacenar objetos y archivos que adicionalmente nos brinda escalabilidad, durabilidad y seguridad.

- BigQuery: Es un servicio de análisis de datos totalmente administrado y altamente escalable que nos permite ejecutar consultas SQL en conjuntos de datos masivos en tiempo real. Es muy potente y puede ejecutar consultas en conjuntos de datos de gran tamaño en cuestión de segundos o minutos; además es escalable porque puede manejar conjunto de datos de cualquier tamaño. Finalmente otra ventaja adicional es que Google se encarga totalmente de la administración y estabilidad de la infraestructura.

Estas herramientas las podemos utilizar de manera conjunta con Dataproc lo cual nos dará los siguientes beneficios:

- Almacenar datos de manera eficiente y segura en Cloud Storage.

- Realizar análisis de datos a gran escala y obtener información valiosa con BigQuery.

- Procesar y transformar datos en paralelo en clústeres de Dataproc utilizando Apache Spark o Hadoop.

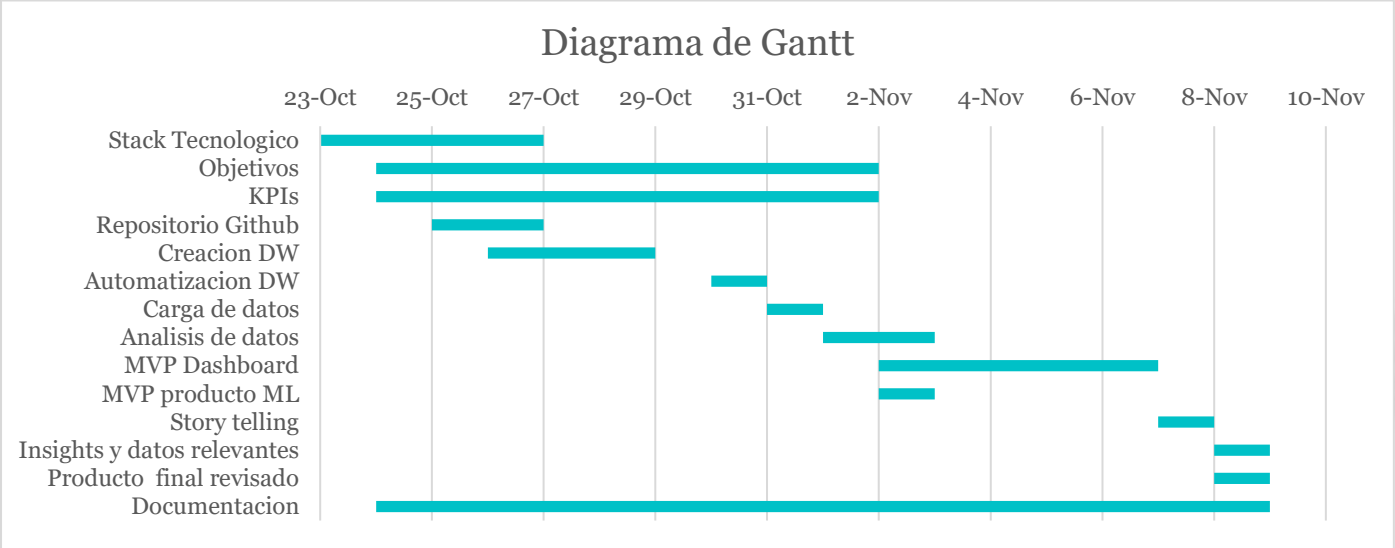
- Orquestar flujos de trabajo completos que incluyen la ingestión de datos, procesamiento y análisis.

El uso de estas herramientas nos brinda la flexibilidad y escalabilidad necesarias para abordar una amplia variedad de proyectos de procesamiento y análisis de datos en la nube.

TECNOLOGÍA		DESCRIPCIÓN
Visual Studio Code		Entorno de desarrollo integrado (IDE) popular entre científicos de datos y desarrolladores de Python
Python		El lenguaje de programación principal para la mayoría de las tareas de ciencia de datos y análisis.
	Jupyter Notebook	Para prototipar y documentar análisis y modelos
	Pandas	Librería para manipular y analizar datos
	Numpy	Librería para realizar cálculos numéricos y operación con matrices
	Seaborn	Librería para visualizar datos estadísticos
	Matplotlib	Librería para generar gráficos
	Scikit learn	Librería para implementar técnicas de machine learning y modelado
	NLTK	Librería para procesamiento de lenguaje natural (NLP), si es necesario para análisis de texto
	TensorFlow	Librería para tareas de machine learning y deep learning

	PyTorch	Librería para tareas de machine learning y deep learning
	PyArrow	Librería para interactuar con Arrow y poder trabajar con archivos Parket
	PySpark	Biblioteca oficial de Python para interactuar con Apache Spark
	Koalas	Biblioteca que proporciona una interfaz similar a Pandas para trabajar con Spark
Google Colab		Plataforma en la nube que proporciona entornos de Jupyter Notebook con acceso a recursos de Google
Google Drive		Plataforma de almacenamiento de archivos que permite interactuar como almacenamiento de Google Colab
Google Cloud Platform (GCP)		Para el almacenamiento de datos, procesamiento en la nube y despliegue de aplicaciones. Puedes usar servicios como BigQuery para consultas y procesamiento de datos a gran escala
Apache Spark		Motor de procesamiento de datos distribuido que es especialmente útil para operaciones de big data y tareas de procesamiento y análisis de datos a gran escala.
Apache Parquet		Formato de archivo columnar eficiente que es ideal para el almacenamiento y procesamiento de grandes volúmenes de datos
CSVs		Para el almacenamiento de datos estructurados en formato CSV.

Cronograma:



PROYECTO RestoTrends