13 – 10 -2022

Proyecto Final Sistema de recomendación

Documentación – Semana 2

**Integrantes:**

* Franco Pietrokovsky
* Joaquín Gómez
* María Pía Ruiz Jaimes
* Marcos Calixto López

Tabla de contenido

Arquitectura del sistema………………………………………………………………… 2

Integración de datos.……………………………………………………………………… 2

Datos recepcionados ……………………..……………………………………………… 2

Almacenamiento en la nube …………….…………………………………………… 3

Datawarehouse ……………………………………………………………………………... 3

Consultas a la BD ……………..…………………………………………………………… 4

# Arquitectura del sistema

Para el presente proyecto se adopta un esquema basado en tres etapas, que interactúan con el objetivo de poder insertar los datos recepcionados en un modelo relacional, además de poder realizar consultas en la nube utilizando *herramientas de big data*.

Diagrama

Descripción generada automáticamente

**Figura 1.0.** Diagrama de la arquitectura para sistema de recomendación.

# Pipeline

## **Datos recepcionados**

Los archivos proporcionados por el cliente se introducen en la primera etapa del Pipeline, es la base fundamental para poder iniciar con el proceso en general.

Estos archivos son obtenidos mediante los enlaces de descarga proporcionados por el cliente, a continuación, algunos ejemplos:

* Libros: <http://snap.stanford.edu/data/amazon/productGraph/categoryFiles/reviews_Books_5.json.gz>
* Oficina: <http://snap.stanford.edu/data/amazon/productGraph/categoryFiles/reviews_Office_Products_5.json.gz>
* Salud y cuidado personal: <http://snap.stanford.edu/data/amazon/productGraph/categoryFiles/reviews_Health_and_Personal_Care_5.json.gz>

Existen en total 25 archivos originales, los cuales fueron proporcionados por el cliente.

## **Almacenamiento en la nube**

Una vez tengamos la data almacenada, la cargamos en un servidor externo, para realizar esto, daremos una previa descripción de los servicios de Google Cloud.

### Google Cloud Service

Google Cloud (Nube de Google) es una plataforma que ha reunido todas las aplicaciones de desarrollo web que Google estaba ofreciendo por separado. Es utilizada para crear ciertos tipos de soluciones a través de la tecnología almacenada en la nube y permite por ejemplo destacar la rapidez y la escalabilidad de su infraestructura en las aplicaciones del buscador.

De lo mencionado anteriormente, utilizamos para esta segunda etapa los servicios de *Google Cloud Storage*, el cual almacena los archivos en depósitos denominados buckets.

Interfaz de usuario gráfica, Aplicación, Tabla, Excel

Descripción generada automáticamente

**Figura 2.0.** Archivos almacenados en un bucket de GCS.

## **Datawarehouse**

Para la implementación del Datawarehouse se identifican los campos clave de cada tabla, obteniendo finalmente un diagrama como el que se muestra a continuación:

Diagrama

Descripción generada automáticamente

**Figura 3.0.** Diagrama relacional para el proyecto utilizando 8 tablas (2 de hechos y 6 de dimensiones).

## **Consultas a la BD**

Una vez tengamos la data almacenada en un bucket de Google Cloud Storage, creamos las tablas utilizando un servicio adicional de Google Cloud Service, se trata de *Google Big Query*.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

**Figura 4.0.** Tablas almacenadas en la nube con GBQ.

Para realizar consultas en GBQ se hace uso del interprete SQL implementado por defecto en la interfaz. Por ejemplo, si quisiéramos obtener “los primeros 5 productos con mejor puntaje de toda la tabla”.

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

**Figura 5.0.** Ejecución de una Query con GBQ.