Laura Castaño Pujalte, práctica de Big Data Arquitecture, bootcamp BD.

Diseño de un Data lake

Definición la estrategia del DAaaS

Definir el catálogo de servicios que proporcionará la plataforma DAaaS, que incluye incorporación de datos, limpieza de datos, transformación de datos, datapedias, bibliotecas de herramientas analíticas y otros.

Se trata de un servicio de pago para acceder fácilmente a los productos mejor evaluados por los clientes de Shein. Está basado en la cantidad de comentarios positivos de cada uno de los artículos y extrae solo los 30 mejores de cada una de las categorías.

Arquitectura DAaaS

- Necesitamos un crawler en una VM para extraer, una vez al día en un CSV, los productos (título + link), el numero total de comentarios y los comentarios de la web www.Shein.com

- Un bucket de google storage para almacenar los datos del CSV en la nube. En el que se hará la ingesta cada día de forma automática.

- Un cluster de Hadoop:

    - Un HDFS para almacenar los datos que se le ingestan diariamente.
    - Un Scheduler de Yarn que ejecute el resto de contenedores del cluster y los supervise.
    - Spark con Python donde se procesarán los datos, se filtrarán y se obtendrán los 30 mejores de cada categoría.
    - Hive donde almacenar el archivo resultado.

- Un bucket de Cloud Storage para guardar los datos de Hive fuera del Cluster de Hadoop.
- Un backend que conecte el Bucket con una web.
- Un servicio web /app que hará las consultas al bucket de resultados.

DAaaS Operating Model Design and Rollout

1. Ejecutar el crawler cada día a la misma hora desde una VM con una Cloud Function programada por el Scheduler de Google Cloud.

2. Guardar los datos del crawler en un bucket de Cloud Storage con otra cloud Function como respuesta a un evento.

3. Levantar automáticamente el Datarproc (cluster) cuando se termine la ingesta de datos en el Storage con una Cloud Function condicionada por un evento.

4. Con Yarn se planificará y supervisará la ejecución de Spark y Hive.

5. Mediante un script de python en spark se procesarán los datos.

6. Hive almacenará los datos resultado del script de Spark.

7. Un Bucket para sacar los datos del Cluster.

8. Una Cloud Function detendrá el Cluster de Hadoop cuando se hayan volcado los datos en el bucket de resultados.

9. Un backend para pasar los datos del bucket a la web.

10. En un servidor, habrá un servicio web desde donde poder hacer las consultas.

Desarrollo de la plataforma DaaaS.

- Ejecutar el crawler cada día a la misma hora por una Cloud Function programada por el Scheduler de Google Cloud.

- Se ingestará automáticamente el CSV obtenido por el crawler a un bucket de Cloud Storage con otra cloud funtion local dependiendo de un eventarc del Scheduler de Google Cloud. Cuando detecte que el crawler a terminado de descargar el archivo.

- Se levantará el cluster de Hadoop con una function con eventarc del Scheduler cuando se termine la ingesta del CSV en el Storage.

- El Scheduler del Resource Manager de Yarn planificará las colas por orden de llegada y el Aplications Manager supervisará la ejecución del resto de aplicaciones en los contenedores exclavos, Spark y Hive.

- Mediante un script de python en spark se procesarán los datos. Con método short() ordenaremos los datos de mayor a menor cantidad de comentarios e imprimiremos los 50 primeros. Con el comando re.search() se filtrarán los comentarios de estos 50 artículos por palabras clave positivas y se imprimirán los 30 con más comentarios positivos en una lista. Finalmente con el comando shorted() se imprimirá una lista ordenada en un CSV para ser almacenada en el Hive y de alguna manera se guardarán los resultados en un bucket externo.

- Una Cloud Function detendrá el Cluster de Hadoop cuando se hayan volcado los datos en el bucket de resultados. De nuevo como respuesta al evento de eventarc.

- Necesitamos un backend para poder consultar los datos del bucket con la web.

- En un servidor, habrá un servicio web desde donde poder hacer consultas del CSV que hay en el Bucket de resultados a través del backend.

Link a Diagrama

https://docs.google.com/drawings/d/1waap0hBxdyP_-uJD9tYfPM14OLimN_L2D8IWdd7Onp0/edit?usp=sharing