Procesamiento y análisis de una base de datos masiva de libros

Máster en Ciencia de Datos e Ingeniería de Datos en la Nube

Autor: Juan Jesús Fernández Fernández

Tutor: Luis de la Ossa

Junio, 2023

Agenda

- Introducción al supuesto práctico
- Conceptos relacionados
  - Data Lakehouse
  - Databricks
  - Arquitectura de medallón
- Metodología y Desarrollo
  - Esquema del proceso de ETL
  - Decisiones de diseño
  - Pasos del proceso
- Bibliografía y enlaces a GitHub

1. Introducción al supuesto práctico

Introducción al supuesto práctico

- **Resumen:** El departamento de marketing de la empresa necesita obtener información sobre los libros más populares entre los lectores para utilizarlos como reclamo en sus campañas de marketing.

- **Petición:** Un conjunto de datos con los 1000 libros más reseñas y mejores valoraciones con información relevante y métricas útiles para la toma de decisiones.

- **Problemática:** Los archivos disponibles son demasiado grandes para su análisis directo en Excel, por lo que requieren de ayuda para su procesamiento

2. Conceptos relacionados

Del Data Warehouse al Data Lakehouse

Data Warehouse

- Pro: Datos mejor estructurados
- Contra: Limitación en el uso de datos semi y no estructurados

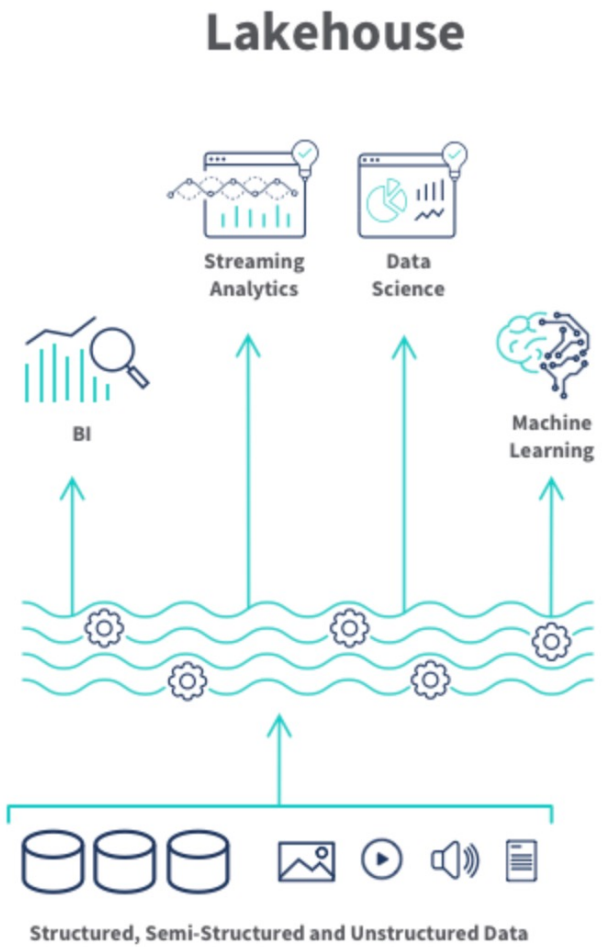
Data Lake

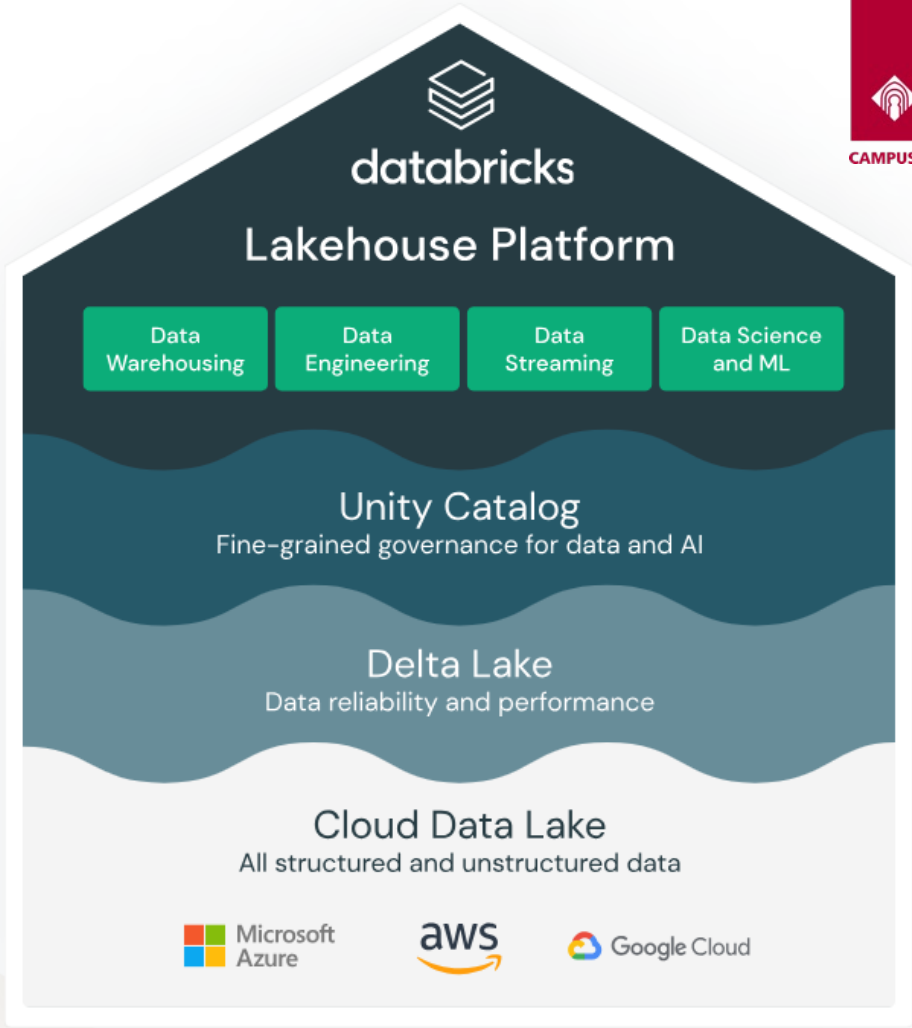
- Pro: Mayor acceso a todo tipo de estructura de datos
- Contra: Menor control sobre la calidad de los datos
- Contra: Falta de soporte para datos transaccionales
- Datos en formato raw

Data Lakehouse

- Pro (Data Lake): Mayor acceso a todo tipo de estructura de datos
- Pro (Data Warehouse): Datos mejor procesados, con procesos que aseguran su calidad

Data Lakehouse





Databricks

- Plataforma basada en procesamiento con Spark
- Notebooks compatibles con SQL, Python, Scala, R
- Usos:
  - Data Engineering (ETL)
  - Machine Learning
  - Data Analysis

databricks

Arquitectura de medallón

Bronce (Bronze):

Carga de datos no procesados (raw data)

Plata (Silver):

Limpieza, procesamiento, enriquecimiento, filtrado, deduplicado

Oro (Gold)

Datos agregados preparados para analítica (reportes, ML)

3. Metodología y Desarrollo

Esquema del proceso de ETL

E

1. Descarga de los datos en archivos de texto comprimidos

L

2. Carga de los archivos al Data Lake (S3 / ADLS2)

T

3. Procesamiento de los datos en Databricks con Spark

4. Modelado de datos en arquitectura de medallón

L

5. Exportación de la tabla Oro al Data Lake (S3) en ficheros parquet

E

6. Lectura de los datos en local con Python

7. Extracción de datos de la API Open Library sobre las obras

T

8. Enriquecimiento del conjunto de datos con datos.

9. Transformación al conjunto de datos final

L

10. Carga del archivo final al Data Lake (S3 / ADLS2)

Decisiones de diseño

- Los conjuntos de datos iniciales se extraen en archivos completos actualizados cada mes. Las tablas se espera que sean refrescadas por completo en cada ciclo de procesamiento mensual.
- Una vez se ha realizado la reducción del conjunto de datos inicial, se puede trabajar en local para reducir costes de procesamiento.
- El uso de la API está muy limitado, por lo que solo se completan los datos para los libros seleccionados.

Carga de archivos al Data Lake

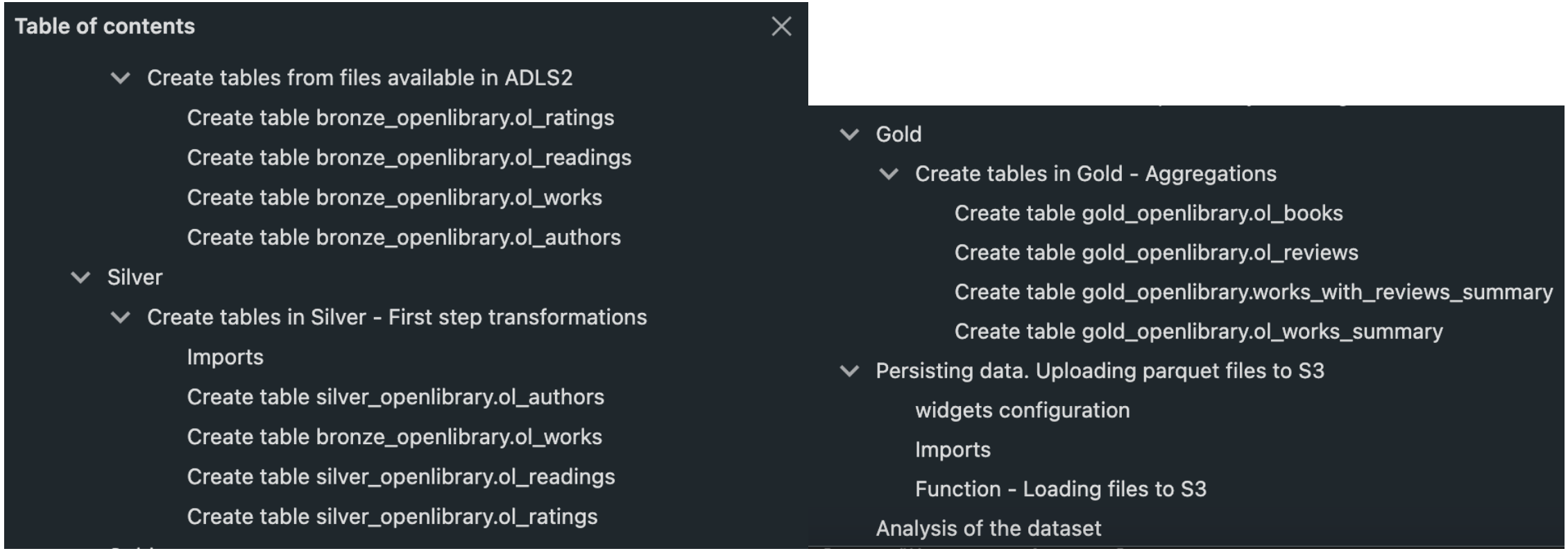
- Archivos de texto comprimidos:

Nombre del archivo	Tamaño	Número de filas
ol_dump_authors_2023-01-10.txt.gz	505.0 MB	11,316,199
ol_dump_ratings_2023-01-10.txt.gz	3.1 MB	321,435
ol_dump_reading-log_2023-01-10.txt.gz	46.0 MB	4,882,308
ol_dump_works_2023-01-10.txt.gz	2.7 GB	30,852,119

- Carga de datos al Data Lake (S3)

Procesamiento de datos

- Procesamiento de datos: Estructura del notebook en Databricks



- Carga de los datos de la tabla Oro a S3 en archivos parquet

Enriquecimiento del conjunto de datos

- Para completar el conjunto de datos de los 1000 libros, se llama a la API de Open Library para extraer mas detalles
  - Carga de los datos exportados de la tabla Oro
  - Llamadas a la API para las referencias de obras del conjunto
  - Agregación de los datos por obra

	works_reference	author_name	title	rating_count	rating_average
148364	/works/OL82563W	J. K. Rowling	Harry Potter and the Philosopher's Stone	55	4.30
95120	/works/OL262758W	J.R.R. Tolkien	The Hobbit	54	4.20
131277	/works/OL5720023W	Stephenie Meyer	Twilight	52	4.07
148326	/works/OL82536W	J. K. Rowling	Harry Potter and the Prisoner of Azkaban	51	4.34
147435	/works/OL81613W	Stephen King	It	50	4.05

Enriquecimiento del conjunto de datos

- Unión de los datos de las obras con los datos de la API

Conjunto principal

	works_reference	author_name	title	rating_count	rating_average
148364	/works/OL82563W	J. K. Rowling	Harry Potter and the Philosopher's Stone	55	4.30
95120	/works/OL262758W	J.R.R. Tolkien	The Hobbit	54	4.20
131277	/works/OL5720023W	Stephenie Meyer	Twilight	52	4.07
148326	/works/OL82536W	J. K. Rowling	Harry Potter and the Prisoner of Azkaban	51	4.34
147435	/works/OL81613W	Stephen King	It	50	4.05

API

work_reference	bs_want_to_read	bs_currently_reading	bs_already_read	rating_average	rating_count	rating_1	rating_2	rating_3	rating_4	rating_5
/works/OL3759085W	43	1	40	4.250000	16	1	0	1	6	8
/works/OL261196W	53	4	44	4.000000	22	0	1	7	5	9
/works/OL262460W	288	14	45	4.043478	23	0	0	8	6	9
/works/OL453743W	45	2	38	4.388889	18	0	0	2	7	9
/works/OL17043626W	180	5	32	3.750000	24	1	2	6	8	7

Cálculo de la métrica de NPS

- **NPS - Net Promoter Score:** Métrica de satisfacción del cliente y promoción de la marca

- Agrupación según escala

Escala	Detractores	Neutrales	Promotores
1-5	1-3	4	5
1-10	1-6	7-8	9-10

- Fórmula de cálculo del NPS

```
NPS(%) = (Promotores - Detractores) / Total de valoraciones
```

4. Bibliografía y enlaces a GitHub

Bibliografía

- https://www.qlik.com/us/data-lake/data-lakehouse
- https://www.databricks.com/glossary/medallion-architecture

Enlaces al código disponible en GitHub

Parte 1: Exportación en Jupyter notebook .ipynb del notebook de Databricks .DBC

https://github.com/fernandezj-cjro/open-library/blob/main/Cidaen-OpenLibrary%20(AWS).ipynb

Parte 2: Jupyter notebook .ipynb

https://github.com/fernandezj-cjro/open-library/blob/main/tfm_openlibrary_part2.ipynb