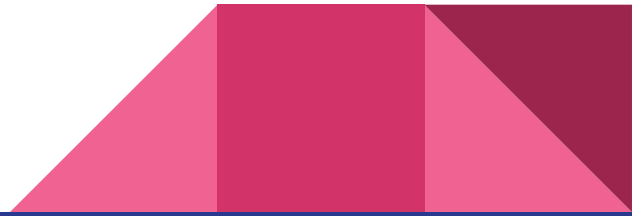
Amazon Product Review - Case Study

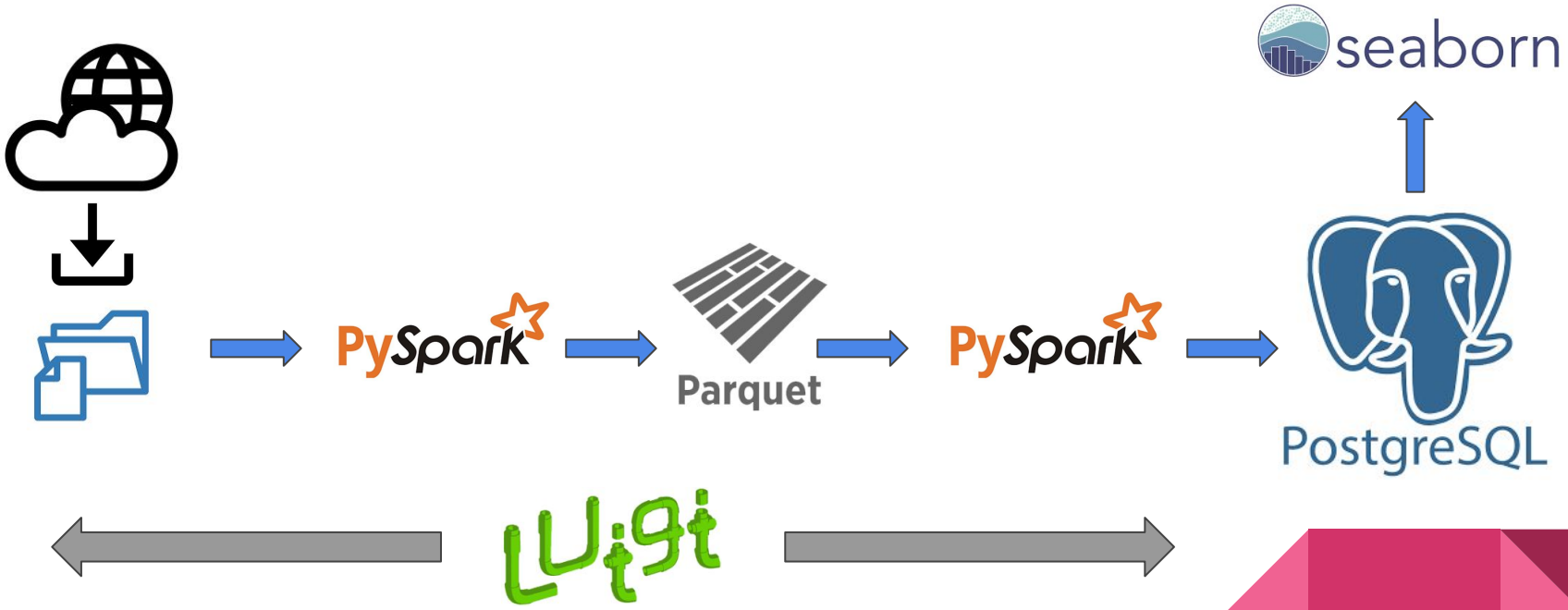
Prepared by - Som Debnath

Motivation

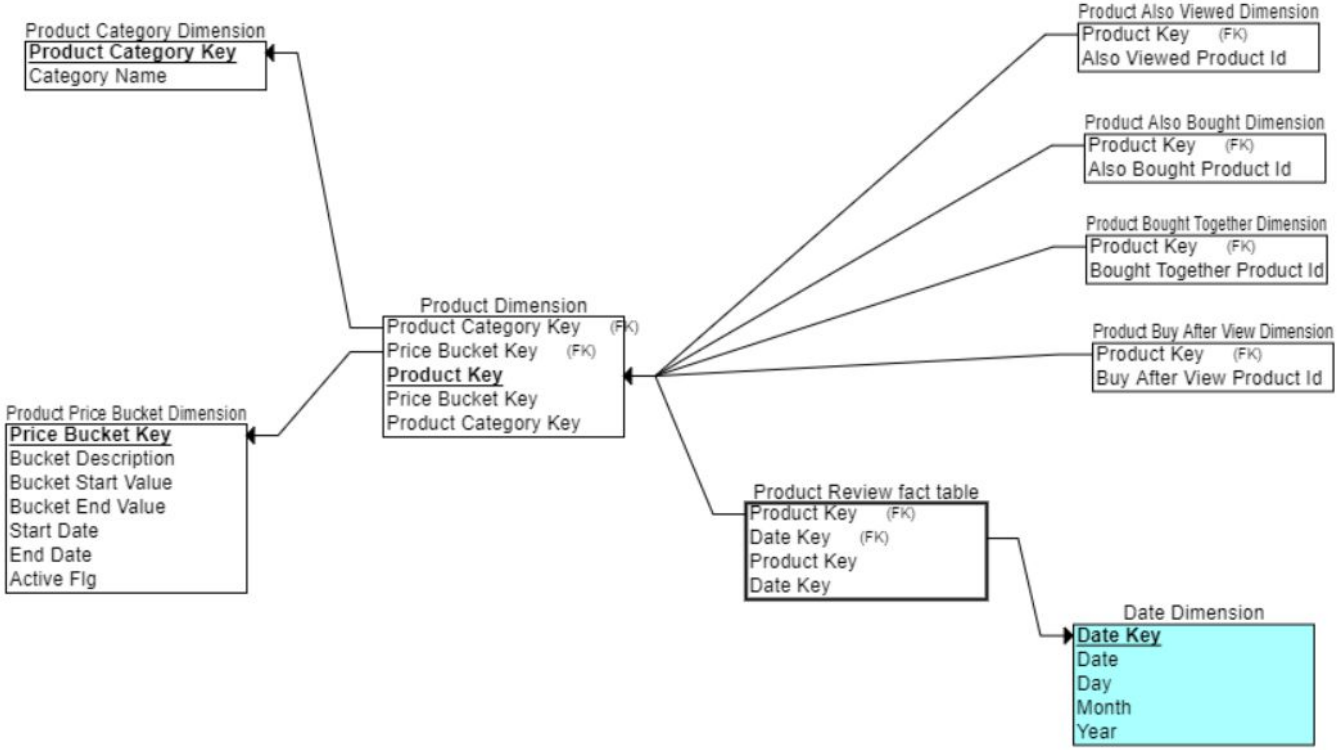
- Data analytics is indispensable in all areas of interest that are related with data.
- Data Warehouse & Data Science are the core parts within the Data Analytics.
- Efficient data analysis is a must for business performance in commercial sectors.
- Data warehouse provides the way to analyse and consume data efficiently by further systems (e.g. Predictive analytics)



Architecture

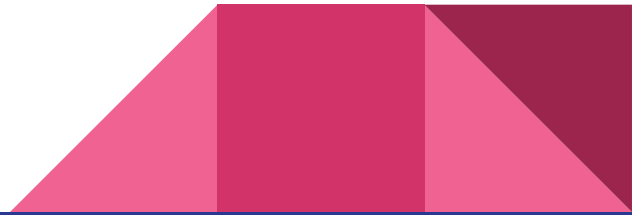


Dimensional Data model



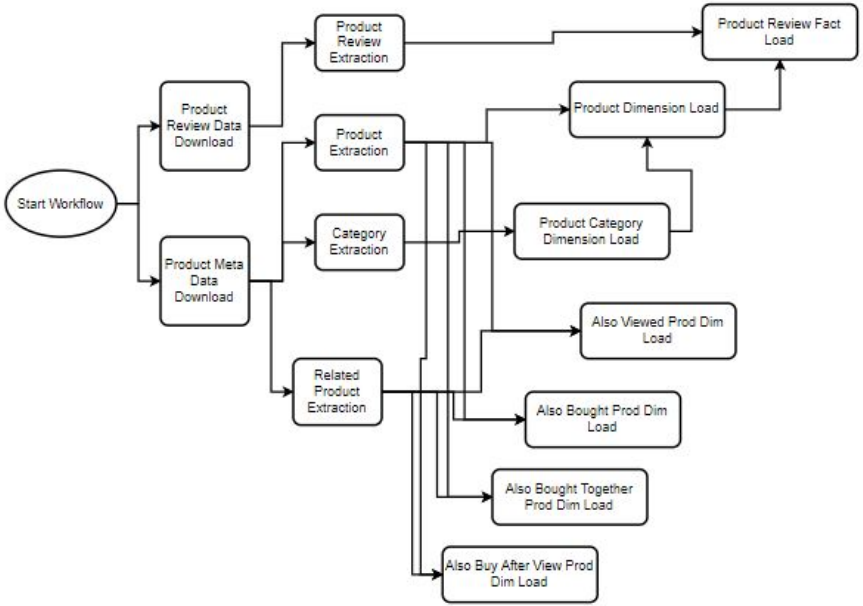
Data Quality

- Product data file had null product identifiers. These were filtered out before reading and processing into parquet format.
- The duplicate value checking in Product Review dataset has been implemented based on Reviewer Id, Product Id and Unix Review Timestamp.
- The related product identifiers are not covered fully in the metadata file. For that reason, the related product ids were not converted to product surrogate key relation in the dimensional modelling.



Job Orchestration

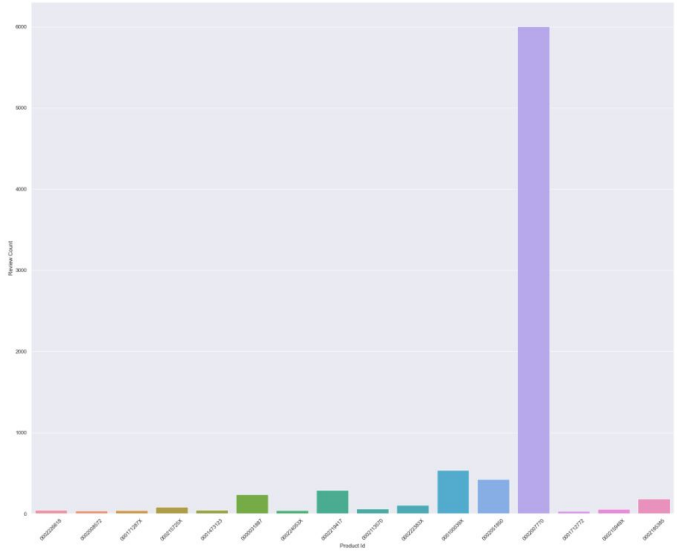
- The job orchestration has been implemented in open source luigi software.

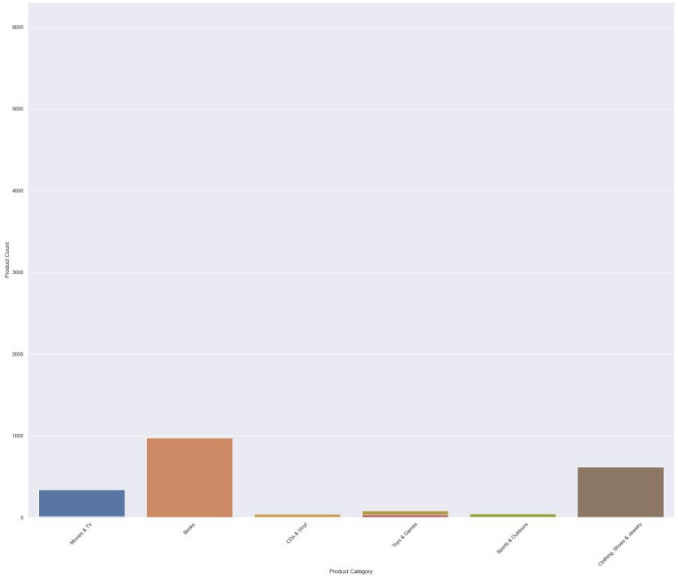


Alternative data model

- The source data can be stored into Enterprise Data Warehouse or Operational Data Store first.
- This will enable to help us produce operational reporting which requires highly granular and normalized data.
- Then we can build individual data mart to produce dimensional model based on the end business requirements.
- We can also leverage Data vault modeling to fulfill the requirements.

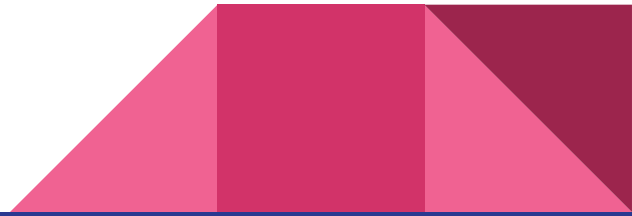
Data analysis Insights - Example





Challenges Faced

- Processing high volume of data using my personal computer. The spark was not able to process the full dataset even with single thread.
- Time constraint has created obstacles to design, develop and test the solutions comprehensively.
- Setting up the orchestration platform.
- Could not implement the solution in container because of time constraint.



Tools & Technologies used

- ❏ PostGres Database
- ❏ Spark (PySpark)
- ❏ Seaborn & Pandas (Python)
- ❏ DBeaver (DB client tool)
- ❏ Luigi (Job Orchestration)
- ❏ Python as programming language
- ❏ PgPLSQL - Postgres DB programming language