Data Engineering Case Study

Product Recommentation

- Pooja Kishore K (AA.SC.P2MCA2107479)

Problem Statement

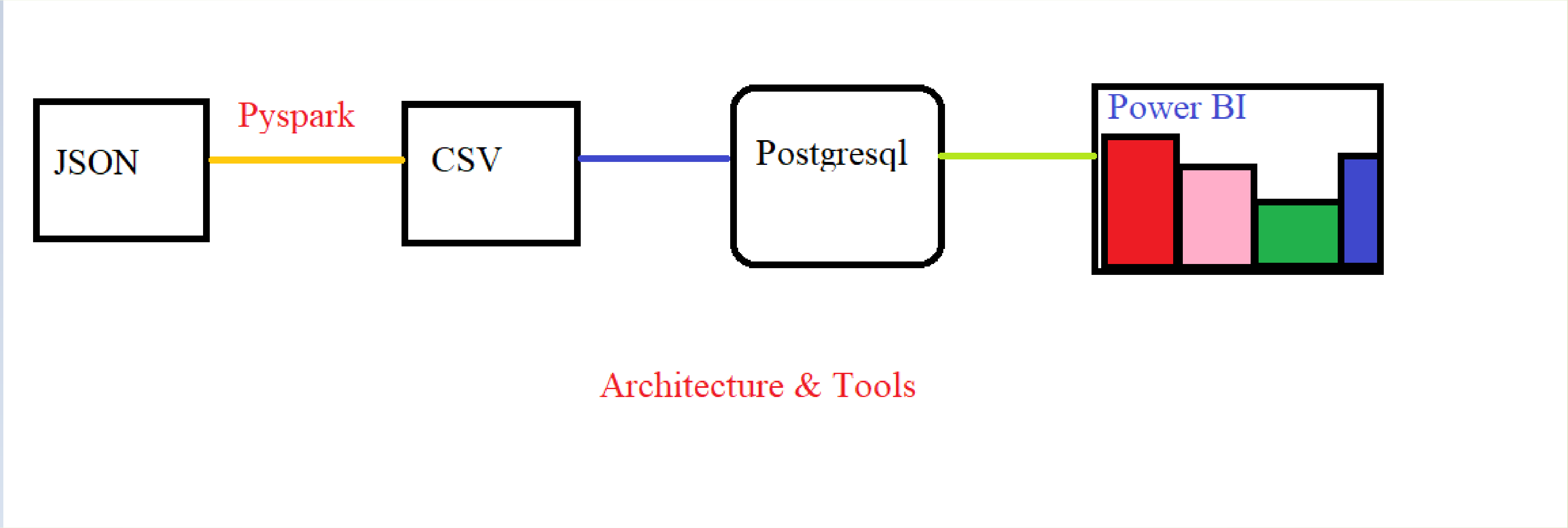
- To implement ETL (Extract , Transform , Load) process by transforming a json file to csv using pyspark and loading all the data to postgresql and finally showing a data visualization in visualization tool (Power BI).

Data

- Here the data is from a json file. The dataset consists of more than 1000 data. It is an Amazon product data with below columns:

  product_id,
  product_name,
  category,
  discounted_price,
  actual_price,
  discounted_percentage,
  rating,
  rating_count,
  about_product,
  user_id,
  user_name,
  review_id,
  review_title,
  review_content,
  image_link,
  product_link

Architecture and tools



JSON

Pyspark

CSV

Postgresql

Power BI

Architecture & Tools

Conclusion

- Successfully converted Extract json format
- Formatted json to CSV file
- Load data from CSV to postgresql
- Data Visualization in PowerBI