PROJECT PRESENTATION

Tech Titans

Ernest, Ara, Vinoth, Anjali

Tech Titan

Ernest Koh

Operations Data Analyst

Ara

Sales Analyst

Anjali

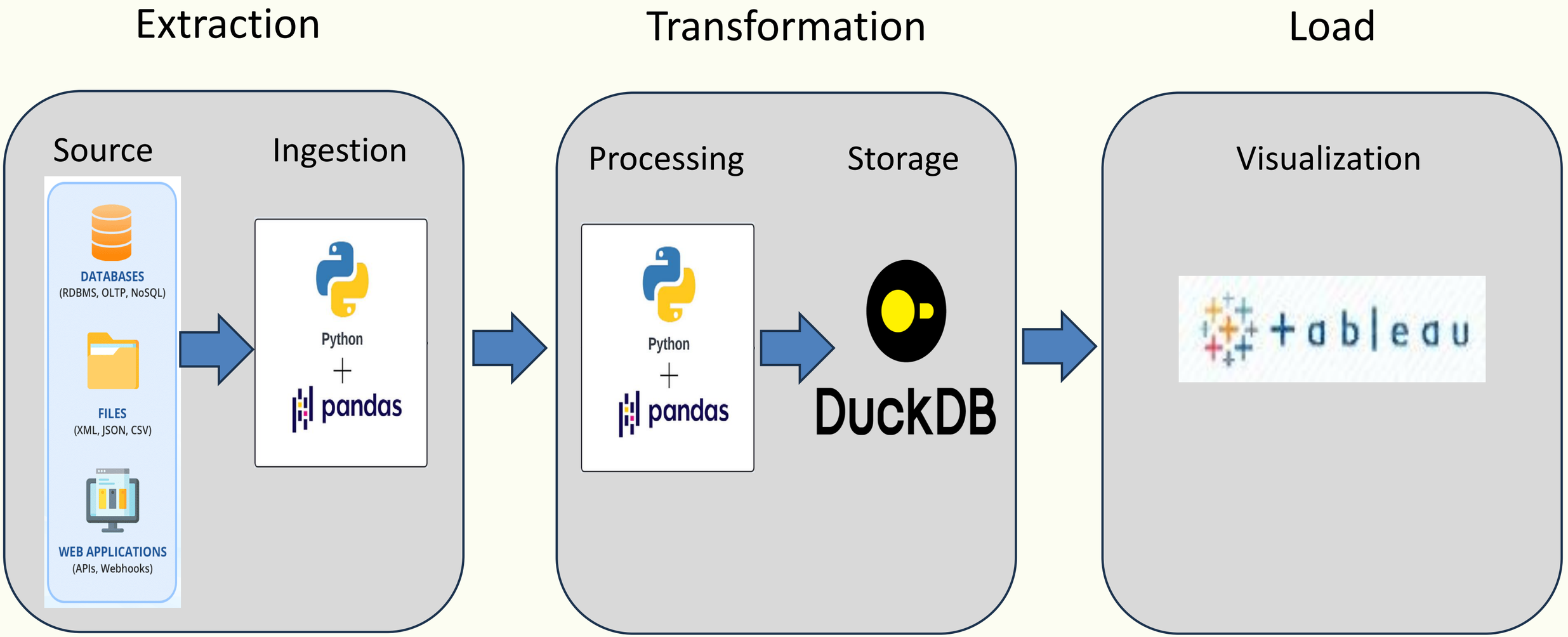
Data Engineer

Vinoth

Team Lead

The Brazilian Olist E-Commerce Dataset is a comprehensive dataset provided by Olist, a marketplace platform in Brazil. This dataset is highly valuable for exploring various aspects of e-commerce transactions, customer behaviors, seller performance and delivery in the Brazilian market.This dataset contains 100,000 records across various marketplaces for orders recorded between 2016-2018.

Global Sustainability Project - Extract Transform & Load



PROBLEM STATEMENT

Analyze the Olist E-commerce dataset to identify key factors affecting :

- customer satisfaction
- sales performance
- operational efficiency
- Order fulfillment

Provide actionable insights to:

- optimize customer experience
- streamline order processing
- improve overall profitability

ELT Pipeline on Microsoft Azure

Data Sources

Data Storage

Azure Data Lake Storage

Transformation

databricks

PySpark

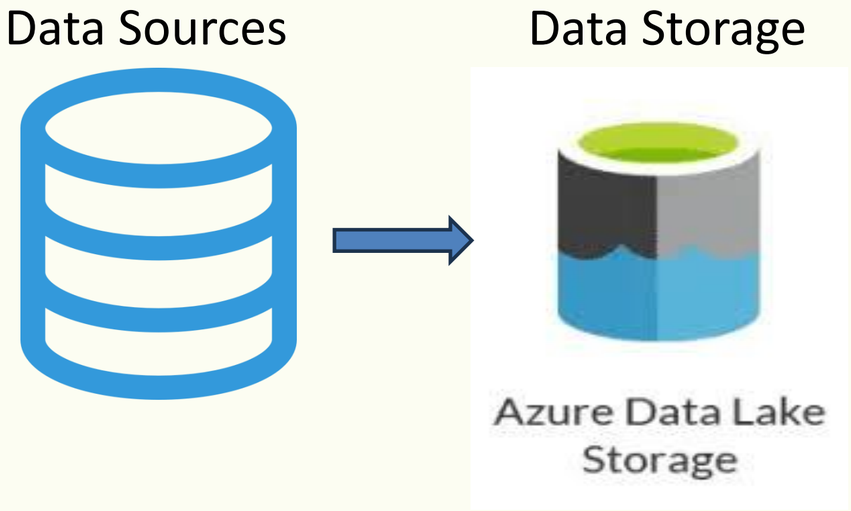
Model & Serve

Azure Synapse Analytics

Reporting

Power BI

Extract & Ingestion



- Utilized Kaggle API to retrieve the dataset - total 9 files
- Created Azure Data Lake Storage for files storage

Transform - Microsoft Azure

Transformation

- Databricks
  - cloud-based data engineering platform
  - build on top of Apache Spark

- Bronze (Raw Data):
  - Unmodified data
  - New data is appended to original data, preserving historical data

- Silver (Validated & Cleansed Data):
  - Data quality issues addressed - missing values, renaming columns
  - Organized in format ready for querying & analyzed

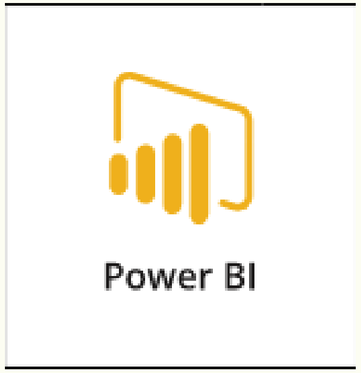
- Gold (Business-ready Data):
  - Denormalized structure for faster querying & performance
  - Data ready for project specific cases with lesser joins

- Delta Tables
  - Default table format for Databricks
  - Suitable for concurrent write operations
  - Can perform batch & streaming operations while data is available immediately for querying

ELT Pipeline on Microsoft Azure

Loading

Reporting



- Synapse Analytics
  - SQL pools for structured data analysis
  - Utilized SQL serverless database with data pulled from data lake to allow ad-hoc querying for quick analysis
- Power BI
  - Can connect to Databricks or Synapse Analytics for data visualization or dashboards
  - Easy to use interface due to drag & drop feature

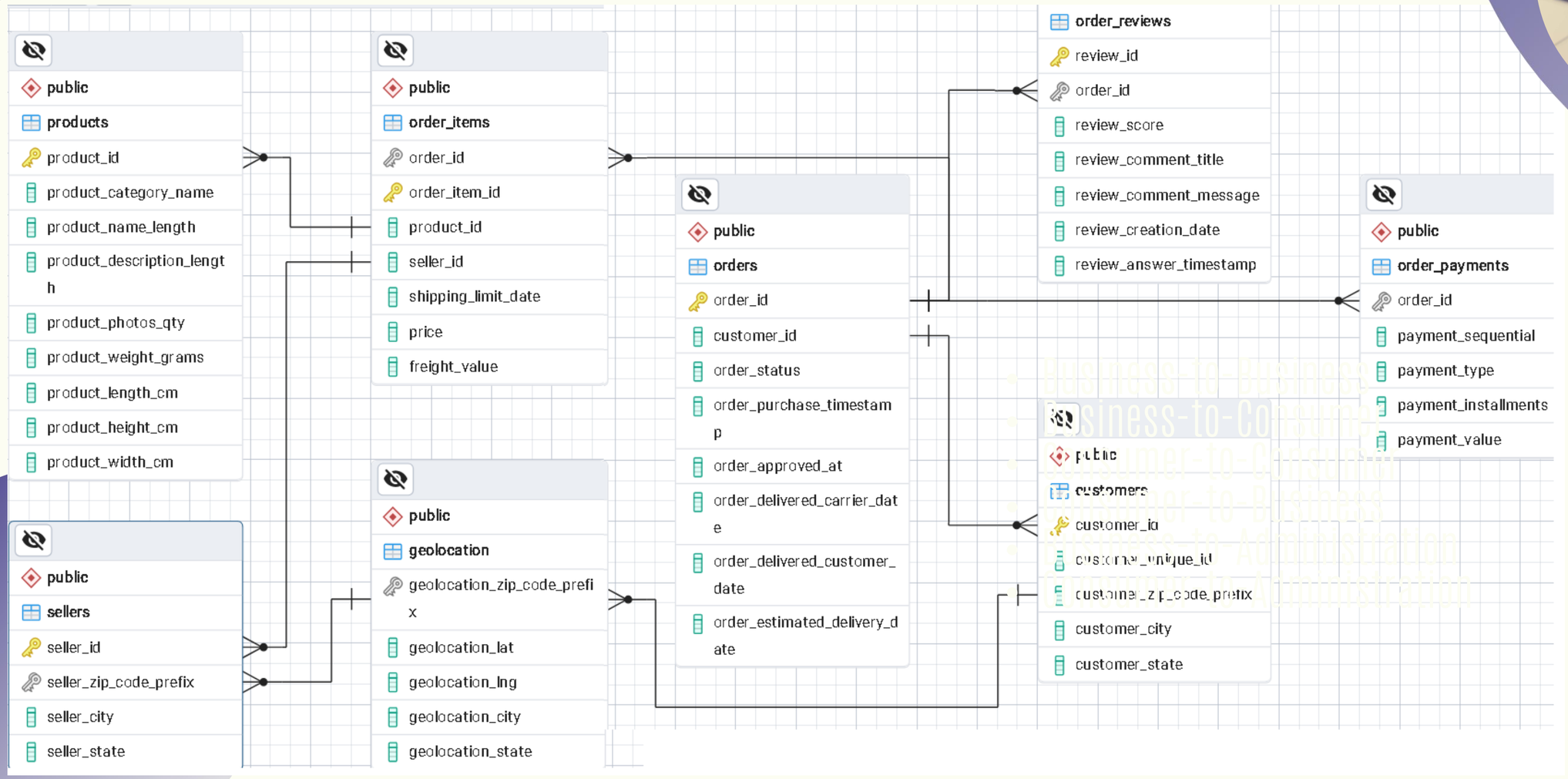
Schema Design

Total of 8 tables from our dataset.

Orders: central or fact table
Order_reviews, order_payments, order_items, products, customers, sellers, geolocation: dimension tables

- Why snowflake schema?
  - Normalized dimension tables
  - Optimizes storage space
  - Reduces data redundancy

Entity Relation Diagram



OPERATIONS (Ops) ANALYSIS

Ernest Viz

Olist Store Analysis

Sales by State

99.44K Total Unique Customers

3095 Total Sellers

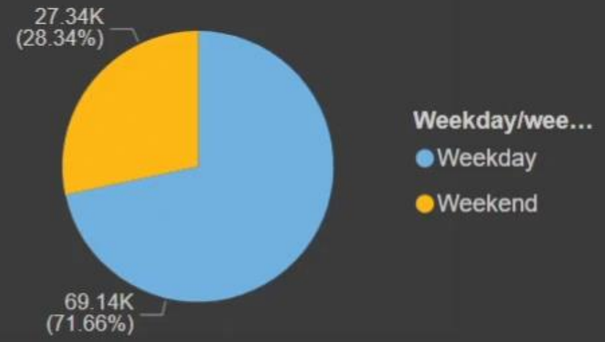
16.01M Total Sales

2.42M Total Profit

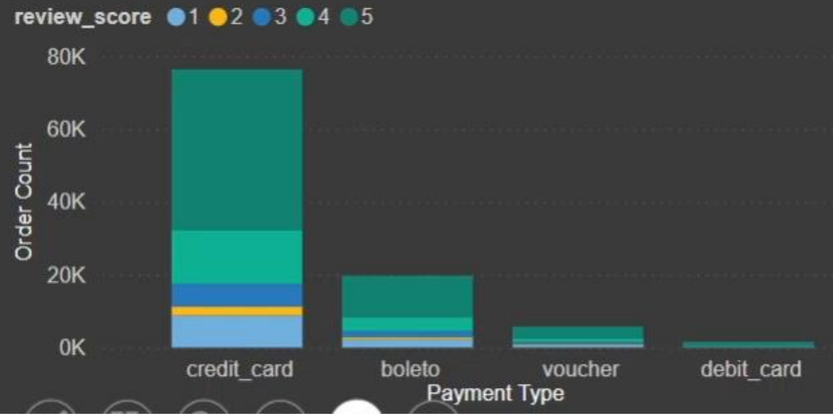
Top 5 popular product categories

Monthly Sales

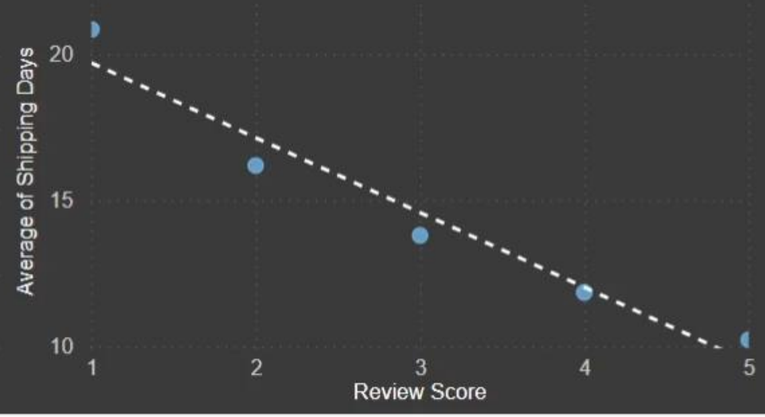
Order status on Weekday & Weekend



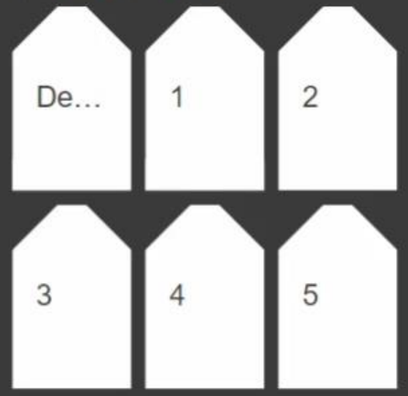
Different Payment Types



Average number of Shipping Days vs Review Score



Review Score



Filters

Ernest_powerBI_Dashboard

Live data | Data updated on 29/09/24, 14:29

Operations Insights

- **There is an increasing sales from February 2017 to Q2 of financial year 2018**

  Action: Dig into the sales product categories, identify contributing items

- **Relative to weekdays, weekends tend to bring in more sales volume**

- **Credit card was found to be the most preferred payment type**

  Action: Analyze probable reasons as to why credit card was most preferred.
  May be due to promotions, cash-back offerings etc.

- **Linear correlation: Higher review score associated with lower number of shipping days**
  - Another point to consider is the condition of shipped items when the customers receive
  - The higher review score may also be attributed by excellent state of items upon receipt

SALES PERFORMANCE

CUSTOMER SATISFACTION & ORDERS ANALYSIS

Summary

- On average, higher score given due to earlier delivery date compared to estimated date
- Credit cards are preferred compared to other modes of payment
- Although cancellation of orders are low, customers do give a low review score if deliveries are delayed.
- Taking note of highest profit generating product categories, ensure that the inventory is maintained & is aligned with peak sales period to boost revenue.

Challenges

- Identifying which schema will better suit our data
- Establishing relationships in the ER diagram
  - Identifying if the relationship is 1:1, 1:N or M:N
- Exploring Microsoft Azure services
  - Which option for files storage?
    - Blob storage/Data lake?
  - Use an all in 1 option or specific services for specific actions?
    - Databricks for transformation
    - Synapse Analytics for querying
    - Power BI for visualization & dashboarding

THANK YOU

for your attention