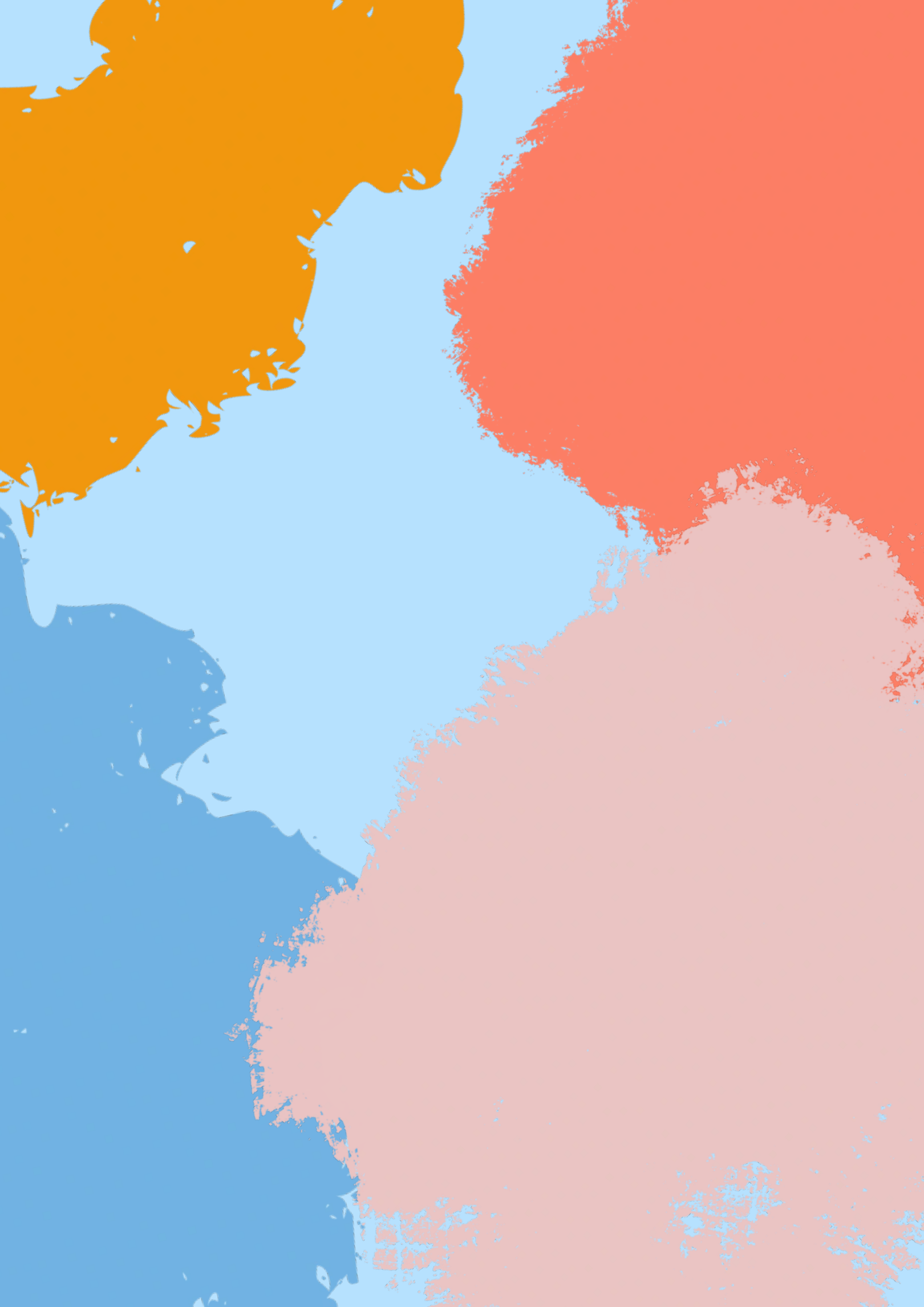

Spark Project-Real World E-commerce data

In this project, we will embark on a detailed exploration of a real-world e-commerce dataset

https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

Project Modules

1. Data ingestion and exploration

→ HDFS

- Set up our Hadoop and spark environment
- Import csv into hdfs
- load data into spark dataframe
- Examining the schema and data types
- Perform EDA on rob of our data

Always NULL

2. Data cleaning and Transformation

address quality issue and transform the data in a structured format

→ handling of null
→ Standardize date formats
→ Numerize & scale numerical values
→ Create new features

3. Data Integration and Aggregation

Combine data from multiple tables to create a unified datasets.

→ Perform join

→ Aggregate data to compute metrics

→ Resolve any inconsistency arising from data integration.

4. Performance Optimization

enhance the efficiency of data processing task

→ data partitioning to optimize spark jobs

→ Utilize caching for iterative operations

→ Optimizing the spark configuration

5. Data Serving

Make the processed data available

→ export transformed data to an external system / databases

→ Create visualization to represent data trends & patterns

Exploration and Understanding of Data

Brazilian E-commerce : Olist dataset

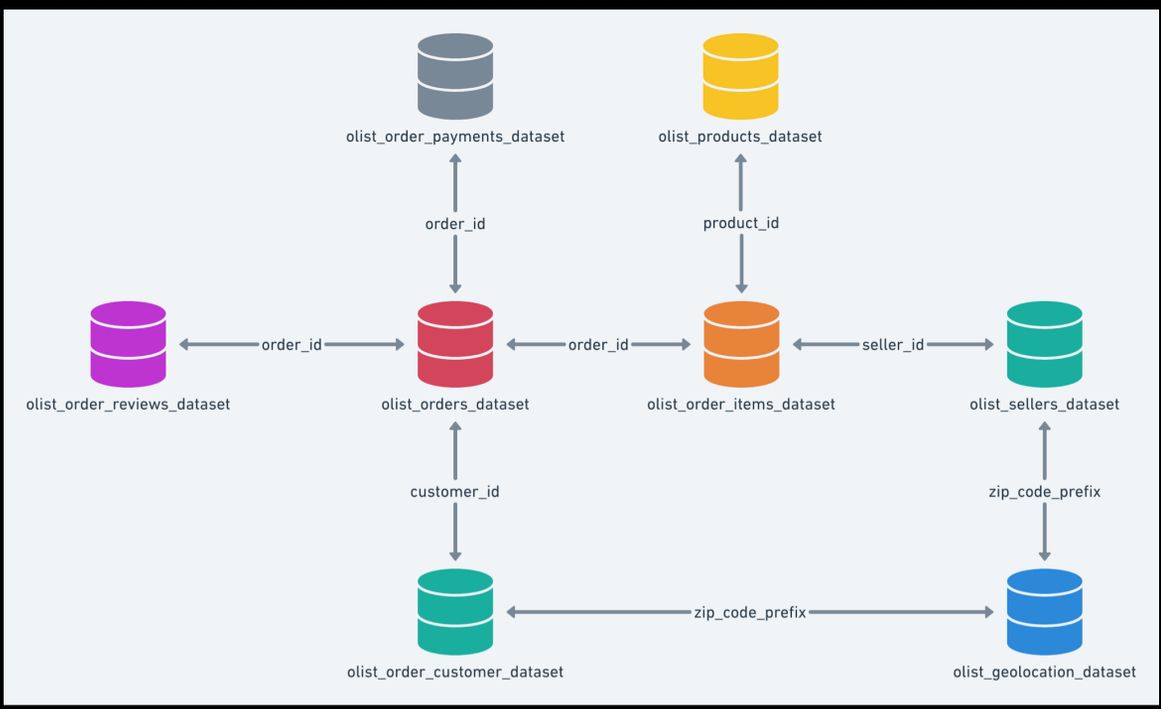

Table Name	Primary Key	Description
`olist_orders_dataset.csv`	`order_id`	Contains details of customer orders (timestamps, status, etc.)
`olist_order_items_dataset.csv`	`order_id`, `product_id`, `seller_id`	Links orders to products and sellers (price, shipping info)
`olist_customers_dataset.csv`	`customer_id`	Contains customer info like city, state, and unique ID
`olist_order_payments_dataset.csv`	`order_id`	Shows payment methods, installment plans, and values
`olist_order_reviews_dataset.csv`	`order_id`, `review_id`	Stores customer reviews and ratings
`olist_products_dataset.csv`	`product_id`	Includes product details like category, weight, and dimensions
`olist_sellers_dataset.csv`	`seller_id`	Seller information (location, ZIP code)
`olist_geolocation_dataset.csv`	`geolocation_zip_code_prefix`	Provides geographic data to map locations
`product_category_name_translation.csv`	`product_category_name`	Translates product categories from Portuguese to English

Module 1: Data Ingestion and Exploration

We will be following a productionize industry approach.

→ Storing data into hdfs / object storage
→ using spark df for scalable operation
→ exploration will focus on practical business use case.
→ we have to make sure to document everything.

Module 2 - Data Cleaning & Transformation

This module will focus on real world data engineering practices for cleaning and transforming data at scale using PySpark

Steps in data cleaning and Transformation

1. Identify issues	missing values, duplicate, invalid data
2. Handle missing value	drop or fill null values, impute
3. Standardize Formats	Convert date time, normalize categorical fields
4. Data type Correction	Correct data type is there for each column
5. Deduplication	remove duplicate records
6. Data Transformation	Apply feature engineering
7. Store cleaned data	data in HDFS, in parquet

Module 3 - Data Integration & Aggregation

Spend lots of time and understand data in depth

1. Join datasets efficiently
2. Optimizing Joins
3. Aggregations
4. Caching and optimizing queries for performance