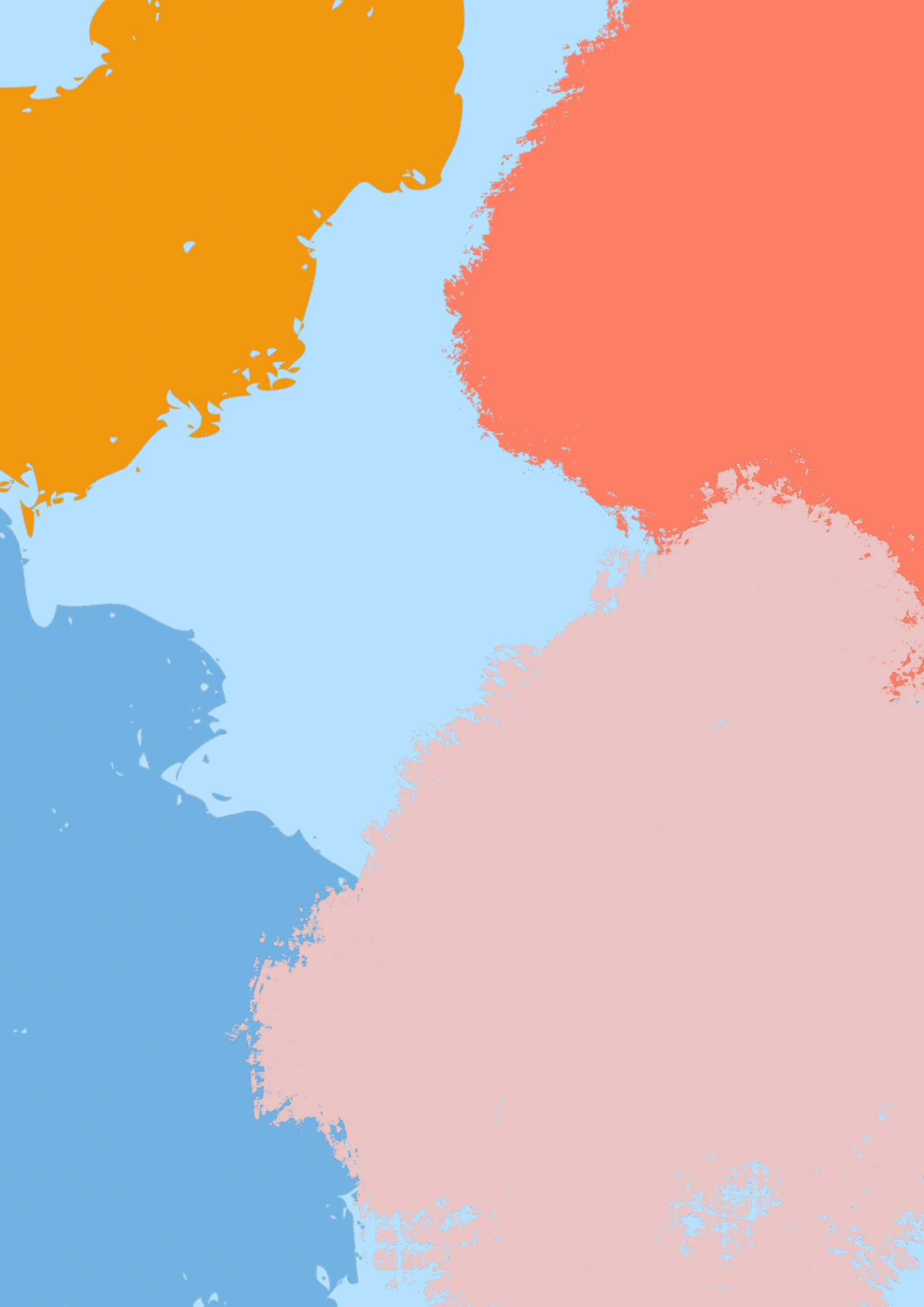


Spark Project - Real World E-commerce Data

Data engineering ✓

Processing ⇒ sparK

1. Internal Codebase

VS code

end — end

2. ~~Cloud Environment~~

AWS / Azure / Oracle

by next week have vscode / pyspark set up.

/ Logging

Project Anatomy / Modules

1. Data ingestion and exploration

   → hdfs

   lake

   → Set up our hadoop & spark environment

   → import all the data

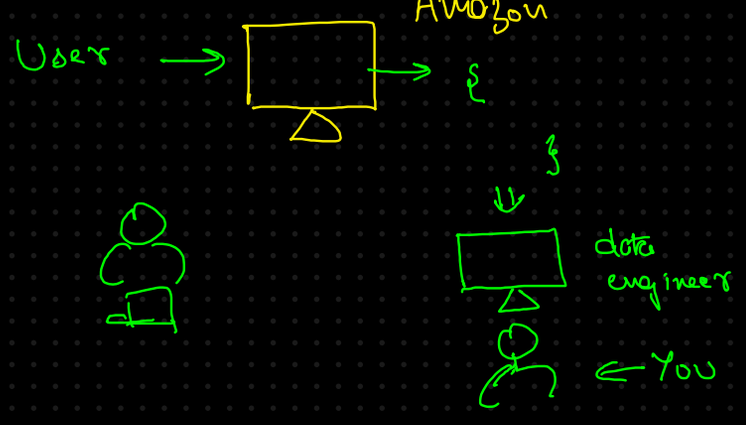
   → load data into spark df.

   ★ → Examine the schema and data types

   → Perform EDA on top of our data

2. Data cleaning & transformation

   addressing the quality issues & transform the data into a structured format

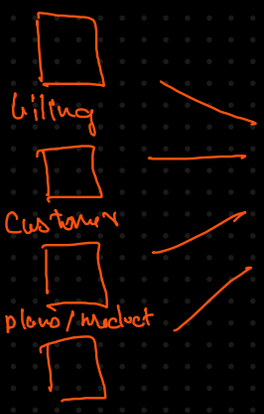
   → handling of nulls

   → standardize column types & date formats

   → Normalize & scale numerical values

   ★ → Create new features

3. Data integration and aggregation



→ Perform Join

→ Aggregating data to compute metrics

→ Resolve any inconsistency arising from data integration

4. Performance Optimization

Tue ⟶ Fri

~ 80%

text → cancel/confirm

5 weeks then I worked on this

enhance efficiency to scale

★ don't unnecessary overengineer

→ data partitioning for optimizing

→ utilize caching

→ optimize our spark config.

5. Data Serving

Make the processed data available for downstream

→ export all the transformed data to the company DB / central loc

→ visualization created for trends & pattern

A person who tries to do everything do nothing

* Document code and learning

→ 1st Hack

AWS/Azure
→

10 ~ 15 min

MongoDB/Cassandra

reviewed

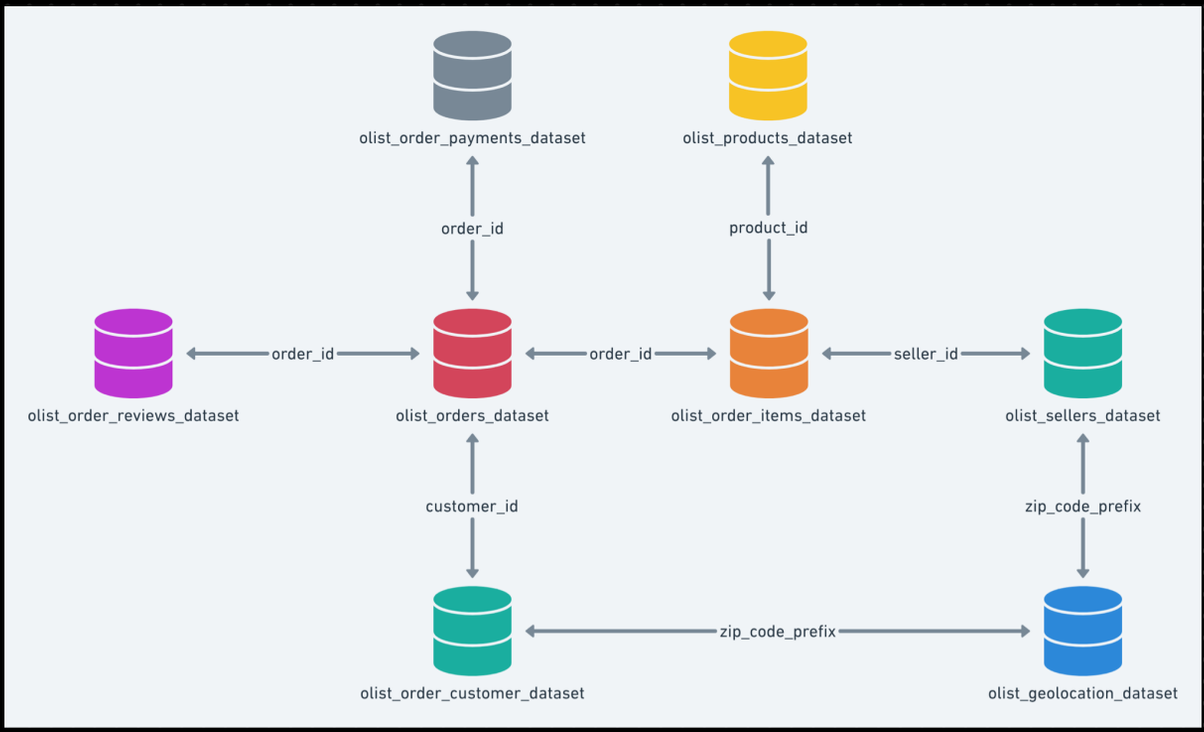
1. 5k amazon voucher
2. Course fee refund
3. Knish AI full access vala course ↴ 15k

⇒ Assignment

⇒ LinkedIn

Dataset ⇒ Olist e commerce

30min - 1 hr on data understanding



Ideally in a company, you should document & write everything about your data

Module 2 - Data Cleaning & Transformation

In this module we will focus on real data engineering practices for cleaning & transforming at scale using Databricks

Steps in data cleaning & transformation

1. Identify issues — null, missing value, invalid data
2. Handle missing values — drop or fill null values, impute
3. Standardize formats — Convert date time, normalize categorical fields.
4. Data Type correction
5. Deduplicate
6. Data Transformation — Feature Engineering
7. Save data