Data Pipeline Using PySpark

Introduction

Walmart is the largest retailer that has both **physical stores**, **online platforms**, and different types of **distribution centers** along with warehouses, it possesses **billions and trillions** of historical data points that are available for machine learning and data analysis.

From such an open sea of data, the data scientists at Walmart need to carry out exploratory analysis of the data and select features that will be relevant to the model fitting for **demand forecasting** down the whole pipeline. It is **time-consuming** and **resource-limiting** to generate features using traditional methods such as python pandas and numpy libraries.

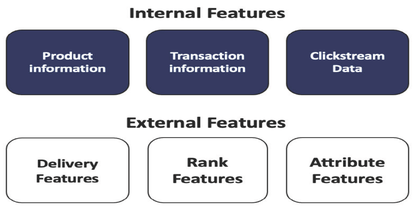

Figure 1. Forecasting Methodology

Why PySpark?

PySpark supports the interaction between Python and Apache Spark, which is an **open-source** framework for processing large datasets in **a distributed fashion**. It brings more **cost-effective** ways to handle **millions and trillions** of data to run machine learning applications on distributed clusters **faster** than traditional python applications and compared to running queries directly in Google BigQuery.

Vertex AI

- Different pricing for functions such as AutoML models and Custom-trained models
- $0.2186 per hour per node; prices change based on the machine type and data center location

DataProc

- Prices based on cluster size, and duration it runs
- $0.01 * (1 master - 4 vCPUs) + 20 worker nodes (80 vCPUS) = $0.84 per hour
- Storage priced separately, which is around $0.02 per GB per month in North America

BigQuery

- On-demand analysis pricing in the US: $5 per TB, including SQL queries, user-defined functions, etc.
- Active storage (tables that have been modified in the last 90 days): $0.02 per GB
- Data ingestion cost, extraction cost (streaming reads - $1.1 per TB read)

Demand Forecasting Workflow

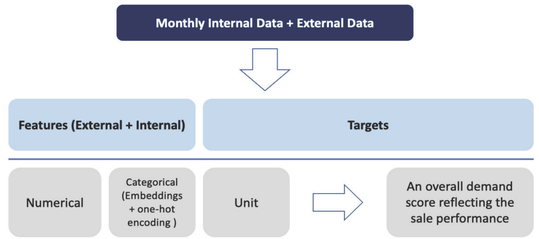

Figure 2. General Flow

- A predicted unit (monthly net quantity) and a demand score will be generated by machine learning to estimate sale performance at the Company
- Information from the past X months will be used to predict the quantity and the score, including the current month; X can be modified by needs or conditions
- Model used to train the cleaned and combined data: XGBoost

Package Design

Supporting scripts serve the facilitating purpose. They may include helper functions, special user-defined functions, and so on. **Embeddings** are variables in text forms, they will be fitted separately by pre-trained deep-learning networks.

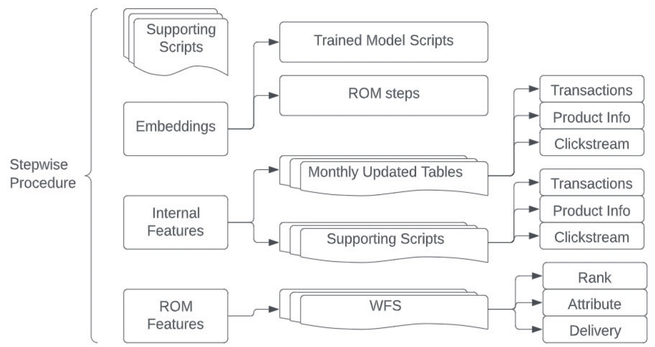

Figure 3. Structure

Internal features need maintenance on a frequent basis, some of them will be selected and transformed, and categorized into three groups - transactions, product information, and clickstream. **External features** are categorized into different three groups, rank features, delivery, and attribute features. All the features will **be combined together by some unique ID**.

Exploratory Data Analysis

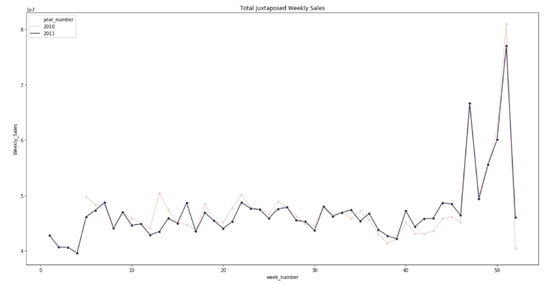

Figure 4. Weekly Sales

Generated by **random data**, mimicking the sales of a certain item. The **granularity** can be changed, for example to department level, category level, or different channels such as online or in-store. Many business decisions can be reflected in the graphs and these graphs, in turn, **help the business to make better decisions**.

Achievements

The final package was pushed to the GitHub page of the Big Org so that everyone who needs the data pipeline can use or customize it. The general structure and procedure were approved team leader and manager.

Hard Skills

- Google Cloud Platform - Vertex AI, Dataproc, Cloud Storage, BigQuery
- PySpark and Spark Configs - translated BigQuery queries to PySpark queries with UDFs
- Code Modularization - designed a python package for the whole team
- Data Analytic Skills - EDA on select features and original tables

Soft Skills

- Developed understanding of demand forecasting pipeline
- Learned how to work remotely with the team and practice in the industry setting

References

1. "PySpark Documentation." PySpark Documentation - PySpark 3.3.2 Documentation, https://spark.apache.org/docs/latest/api/python/.

2. "Pricing per Product; Google Cloud." https://cloud.google.com/pricing/list.

3. Reiff, Nathan. "How Walmart Makes Money: U.S. and e-Commerce Sales Are Growing Fastest." Investopedia, Investopedia, 27 June 2022, https://www.investopedia.com/articles/personal-finance/011815/how-walmart-model-wins-everyday-low-prices.asp.

hg136@duke.edu