Retail Data Analysis using Sqoop, HDFS, and Hive

Business Overview

Retail analytics is the process of delivering analytical data on inventory levels, supply chain movement, customer demand, sales, and other important factors for marketing and procurement choices. Demand and supply data analytics may be utilized to manage procurement levels as well as make marketing decisions. Retail analytics provides us with precise consumer insights and insights into the organization's business and procedures, as well as the scope and need for development.
Aside from inventory management, many retailers employ analytics to determine customer patterns and shifting preferences by merging various sources. Businesses may discover developing trends, and better predict them by combining sales data with a range of criteria. This is strongly related to marketing functions, which also benefit from analytics.
Companies may use retail analytics to strengthen their marketing strategies by better grasping individual preferences and gaining more detailed data. They may design strategies that focus on people and have a greater success rate by combining demographic data with information such as purchasing patterns, preferences, and purchase history.
In this, we will be utilizing Walmart store sales data to perform analysis and answer the following questions:

- Which store has a minimum and maximum sales?
- Which store has a maximum standard deviation?
- Which store/s has an excellent quarterly growth rate in Q3'2012?
- Find out holidays which have higher sales than the mean sales in non-holiday season for all stores together.

Data Pipeline

A data pipeline is a technique for transferring data from one system to another. The data may or may not be updated, and it may be handled in real-time (or streaming) rather than in batches. The data pipeline encompasses everything from harvesting or acquiring data using various methods to storing raw data, cleaning, validating, and transforming data into a query-worthy format, displaying KPIs, and managing the above process.

Dataset Description

The Walmart Store sales dataset consists of 6436 data points, including the following parameters:

- Store
- Date
- Weekly_Sales
- Holiday_Flag
- Temperature
- Fuel_Price
- CPI
- Unemployment

Tech Stack

➔

Language: SQL, Bash

➔

Services: AWS EC2, Docker, MySQL, Sqoop, Hive, HDFS

AWS EC2

A virtual server on Amazon's Elastic Compute Cloud (EC2) for executing applications on the Amazon Web Services (AWS) architecture is known as an Amazon EC2 instance. The Amazon Elastic Compute Cloud (EC2) service allows corporate customers to run application applications in a computer environment. Using Amazon EC2 eliminates the need to invest in hardware up front, allowing users to create and deploy apps quickly. Amazon EC2 allows users to launch as many or as few virtual servers as they want, set security and networking, and manage storage.

Docker

Docker is a containerization platform that is a free source. It allows developers to bundle programs into containers. These standardized executable components combine application source code with the operating system (OS) libraries and dependencies necessary to run that code in any environment.

MySQL

MySQL is a SQL (Structured Query Language) based relational database management system. Data warehousing, e-commerce, and logging applications are just a few of the uses of the platform.

Sqoop

Sqoop is a tool for transferring data between Hadoop and relational database servers. It is used to import data from relational databases such as MySQL and Oracle into Hadoop HDFS, Hive, and export data from the Hadoop file system to relational databases.

Hive

Apache Hive is a fault-tolerant distributed data warehousing solution that enables massive-scale analytics. Using SQL, Hive allows users to read, write, and manage petabytes of data.

Hive is based on Apache Hadoop, an open-source system for storing and processing massive information. As a result, Hive is tightly linked with Hadoop and is built to handle petabytes of data fast. The ability to query massive datasets with a SQL-like interface, using Apache Tez or MapReduce, distinguishes Hive.

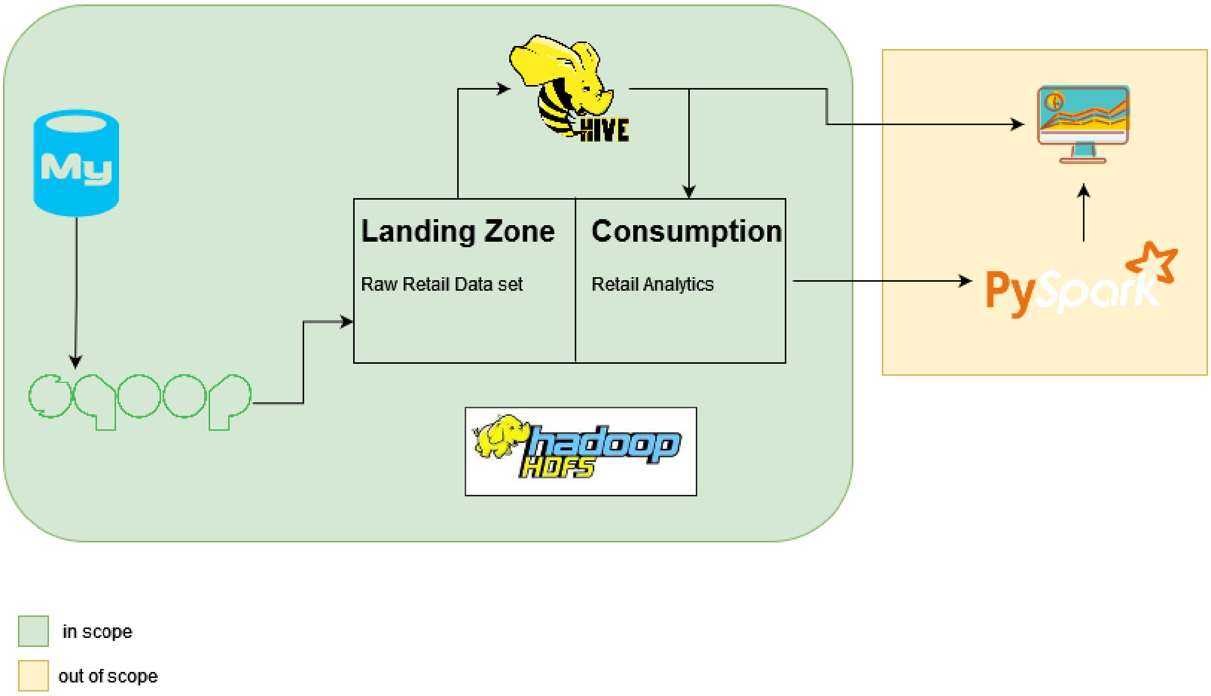
Approach

- Containers for all the services are spun up using Docker.
- Setup for MySQL is performed for Table creation using the dataset.
- Data is imported using Sqoop into Hive.
- Data transformation is performed for analysis and reporting.

Key Takeaways

- Understanding the project and how to use AWS EC2 Instance
- Introduction to Docker
- Visualizing the complete Architecture of the system
- Usage of docker-composer and starting all tools
- Understanding HFDS and various file formats
- Understanding the use of different HDFS command
- Understanding Sqoop Jobs and valuable tools
- Introduction to Hive architecture
- Understanding Hive Joins and Views
- Performing various transformation tasks in Hive
- Setting up MySQL for table creation
- Migrating from RDBMS to Hive warehouse

Architecture


My
sqoop
HIVE
Landing Zone
Raw Retail Data set
Consumption
Retail Analytics
hadoop
HDFS
PySpark
in scope
out of scope