Low Level Design
Movie Analytics

Contents

1. Introduction:

The aim of this document is to provide a low-level design for a scalable system that uses Hadoop, Pyspark and hive metastore for movie analytics. Data will be read in CSV format fromhadoop storage, and we use pyspark to perform analysis on the data.

2. Tech Stack Used:

The following technologies will be used in the pipeline:

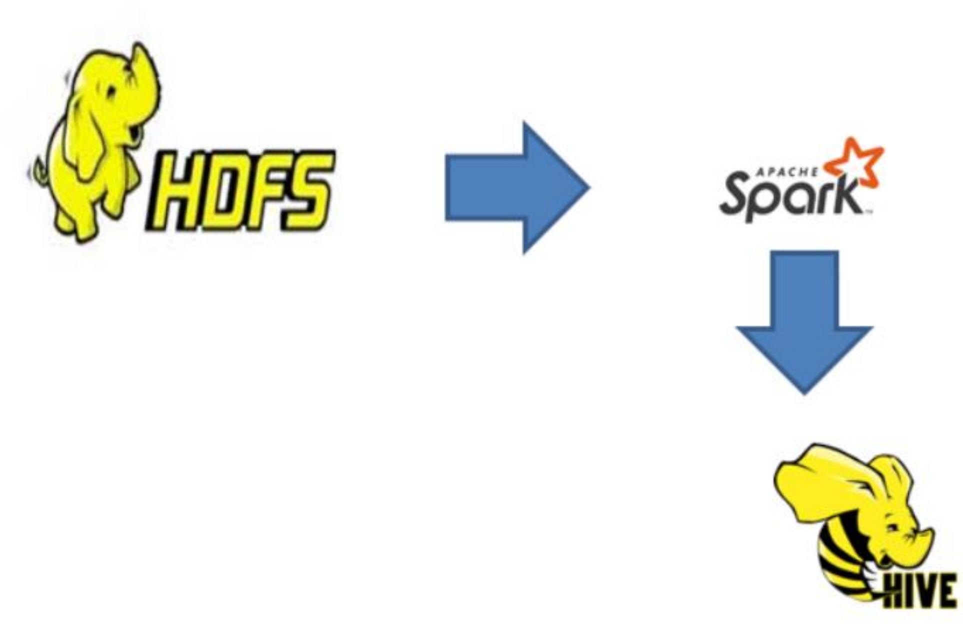
Apache Spark

Hadoop Distributed File System (HDFS)

Hive Metastore

3. Architecture:

The architecture is as follows:



Architecture Explanation:

For this project there are 3 data sources named as ratings.csv,users.csv,movies.csv. In the ratings.csv there are 4 columns userid , movieid , rating, timestamp. The userid gives the detail about the user,the movieid gives the information of the movie to which user gave rating. The rating will be in between 1 to 5. In the users.csv we have 4 columns userid,gender,occupation,zip-code.The userid gives information about the user.the gender gives information regarding the gender of the user and the occupation gives the info regarding occupation of the user and zipcode gives the information about the zipcode of the user. In the movies.csv we have 3 columns. we have information regarding movieid,title genre.the moviid gives the id of the movie,the title represents the title for the movieand genre gives the information regarding to which genre it belongs to this columnis again delimted by '|'.it has various values in that column like action, Thriller etc.

Our data is picked from the hdfs layer and read by the spark engine. After reading the data We can perform various analytical queries on the data and extract useful business insights like top 10 viewed movies,distinct list of genres,to which genre audience are giving more ratings etc.So that the movie industry can analyze the trends and can know the audience interests. Storage: Finally we can store the data as table in hive meatstore if required we can also create real time dashboards by using some dashboarding tools.

4. Detailed Procedure:

The following steps will be taken to implement:

- Dump the CSV files into Hadoop storage.
- Read the data from hdfs and make spark connection.
- Store the data as Dataframes using pyspark.
- Create multiple functions which will be used to query data and call all these functions in the main function.
- Use hive metastore to store the data in hive without DDL.
- Use pyspark and spark sql to query and analyse the data.
- Use hive queries to analyse the data stored in Hive metastore.
- Schedule the Pipeline(to automate the pipleline):

5. Conclusion:

This low-level documentation provides a step-by-step guide for designing a scalable pipeline using Spark to read customer reviews on movies from Hadoop storage and store it into HDFS and hive metastore. The pipeline uses pyspark to perform analysis on the data. We can use Airflow to schedule the pipeline as further enhancement.