Project Title: Designing a Scalable Pipeline for Movie Analytics using pyspark and Hadoop

1. **Introduction:**

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.

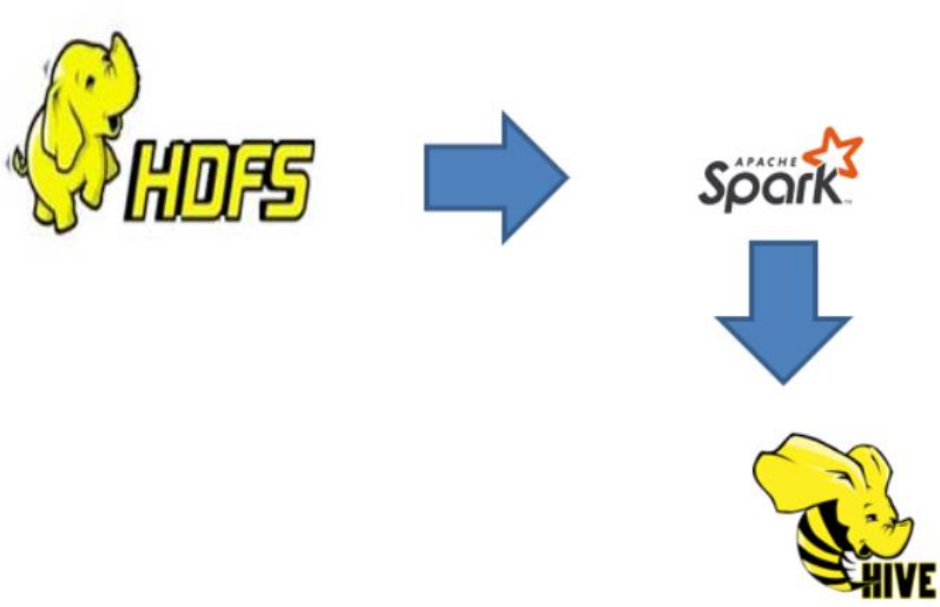
2. **Objectives:**

The main goal of the project is to Analyze the data from various data sources and extract meaningful insights from the data and to know the audience interests.the data will be present in the hdfs storage layer in csv format with delimiter(::).We will be reading data from Hdfs layer and process the data using pyspark.

3. **Architecture:**

The following architecture diagram shows the overall flow of the project:

Architecture Diagram



4. **Step-by-Step Procedure:**

Step 1: Dump the data to Hadoop storage

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.

These data will be processed and stored into Hadoop storage.

Step 2: Create Spark connection and process the data using Pyspark.

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.

Step 3: Store the data in hdfs storage and hive metastore

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.

5. Conclusion:

By following the steps outlined in this project, we were able to design a scalable pipeline for movie analytics using Spark . This 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.