High Level Design Document: Movie analytics Using PySpark

CONTENTS:

|  |  |
| --- | --- |
|  | Page Number |
| **Introduction** | 1 |
| **Architecture** | 1 |
| **Detailed Procedure** | 2 |
| **Benefits** | 2 |
| **Conclusion** | 3 |

1. **Introduction:**

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.

1. **Architecture:**

The architecture of the project is :

Data Source: The Data has to be kept in the Hadoop file system(Hdfs):

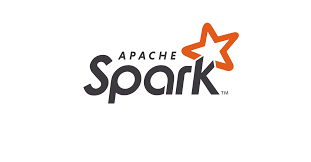
We can keep the data in Hadoop file system using the command:

hdfs dfs -put sample.txt /

Processing Engine: We can use spark to process the data that is present in the hdfs and extract meaningful insights from the data.

Storage:Once the data is processed we will be storing the required data in the hive metastore.

Image:



1. **Detailed Explanation:**

**Data Source:**

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.

**Processing Engine:**

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.

1. **Benefits:**

**Scalability**: The proposed pipeline is highly scalable and can handle large volumes of customer reviews data.

**Real-time analysis**: The pipeline can be scheduled to run iteratively after each hour, which ensures real-time feedback on customer satisfaction.

**Efficiency**: The use of Spark enables the pipeline to process and analyze the data efficiently, which leads to faster results

1. **Conclusion:**

We can read the data from the hdfs and perform analysis of data in spark and write that data

into hive.