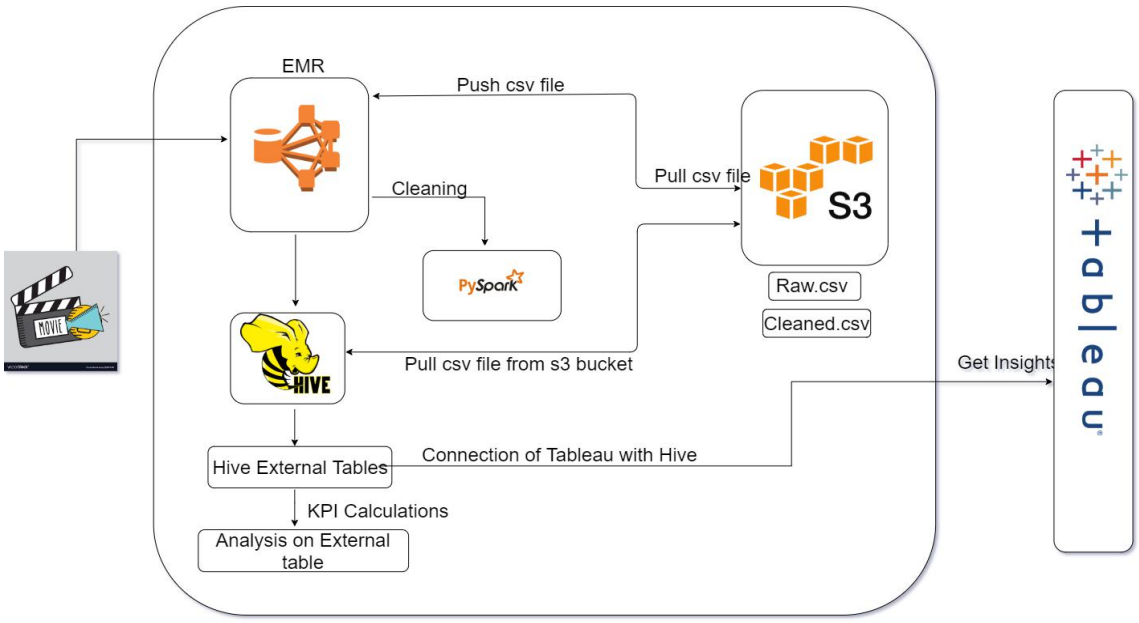
# Final Project Case Study

Title: Movielens dataset analysis using Hive for Movie Recommendations

Platform: AWS, Pyspark, Hive, Tableau, Linux

Objectives: The movieLens dataset is most often used for the purpose of recommendation systems which aim to predict user movie ratings based on other users' ratings. In other words we expect that users with similar taste will tend to rate movies with high correlation.

Description: Ingesting data from an automated process of data pulling from Rdbms and/or Files and storing the same in AWS S3 bucket using Hive. Created an Hive external table pointing to S3, then connected the Hive external table to the Bi tool ; here Tableau, for reporting and analytics.



## 1. Getting Your Data:

" Which year has the most number of ratings?"

"Which is the top rated movie each year?"

"Which movie has been rated 50% more after 5 years? "

**Our Data size = 1GB**

## 2. Data Infrastructure: - Data Lake/DWH

**Steps involved to create a Hadoop cluster on cloud:**

Create a capacity plan for cluster of 3 nodes –

RAM Size 2+1+1,

Hard Disk 10+10+10,

CPU Cores 2+1+1

## 3. Data Ingestion: Data Access

1. Create a data dictionary -

| Field | Type | Description |
|---|---|---|
| movieId | Int | represent the movie id. |
| tag | String | represent user-generated textual metadata. |
| tagId | Int | represent the user id. |
| title | String | represent the full movie title and may include the year of release. |
| genres | String | Genres are a pipe-separated list. |
| rating | Float | Ratings are made on a 5-star scale, with half-star increments (0.5 stars - 5.0 stars). |
| timestamp | Timestamp | use the epoch format (seconds since midnight of January 1, 1970 on UTC time zone) |
| relevance | Float | |
| imdbId | Int | can be used to generate a link to the IMDb site. |
| userId | Int | represent the user id. |
| tmdbId | Int | can be used to generate a link to the The Movie DB site. |

2. Extract data from source in automated fashion – WIP

3. How will you handle incremental data load in data ingestion pipeline?

There are many ways around this situation including:

a) parallelizing ingestion across tables
b) parallelizing ingestion within tables, by reading separate segments simultaneously
c) incrementally loading the changed data
d) utilizing source native paths to accelerate data unloading from the sources (e.g using TPT instead of JDBC to access Teradata systems)

4. What is SCD (slowly changing dimensions) and how will you implement in your project? - WIP

5. Preparing and Cleaning Your Data – Data Transformation.

This phase of the pipeline is very time consuming and laborious. Most of the time data comes with its own anomalies like missing parameters, duplicate values, irrelevant features, etc. So it becomes very important that we do a cleanup exercise and take only the information which is important to the problem asked, because the results and output of your machine learning model are only as good as what you put into it. Again, garbage in-garbage out.The objective should be to examine the data thoroughly to understand every feature of the data you're working with, identifying errors, filling data holes, removal of duplicate or corrupt records,throwing away the whole feature sometimes, etc. Domain level expertise is crucial at this stage to understand the impact of any feature or value.

**The more questions you ask of your data; the more insight you will get. This is how your own data yields hidden knowledge which has the potential to transform your business totally.**

## 4.Exploration/Visualization of Data

**Objective:**

Find patterns in your data through visualizations and charts
Extract features by using statistics to identify and test significant Variables. During the visualization phase, you should try to find out patterns and values your data has. You should use different types of visualizations and statistical testing techniques to back up your findings. This is where your data will start revealing the hidden secrets through various graphs, charts, and analysis. Domain-level expertise is desirable at this stage to fully understand the visualizations and their interpretations. The objective is to find out the patterns through visualizations and charts which will also lead to the feature extraction step using statistics to identify and test significant variables

## 5.Interpreting the Data

Interpreting the data is more like communicating your findings to the interested parties. If you can't explain your findings to someone believe me, whatever you have done is of no use. Hence, this step becomes very crucial. The objective of this step is to first identify the business insight and then correlate it to your data findings. You might need to involve domain experts in correlating the findings with business problems. Domain experts can help you in visualizing your findings according to the business dimensions which will also aid in communicating facts to a non-technical audience.

## 6. Data Reports

The objective of this is to do the in-depth analytics, mainly the creation of relevant reports like what happened/why happened to answer the problems related to Business. For this we are using Tableau for analytics and reporting.