Project 1 Demo Outline

<https://github.com/KumarRevatureReston/220314---UTA-SH---Scala-Big-Data/blob/master/00%20Trainer%20Material's%20-%20Reference/Projects/Project%201/P1_HDFS_HIVE_Sample%20report.pdf>

|  |  |
| --- | --- |
| Topic: Analyze MovieLens data |  |
|  | Dataset: http://grouplens.org/datasets/ |
|  | Use the ml-100k.zip file dataset. |
|  |  |
|  | Problem Statement: |
|  | Import csv data about movies and their ratings into HDFS, then create tables from the imported files in Hive. |
|  | Once imported, create queries to analyze and display data about the movies and their ratings. |
|  |  |
|  | **Example Queries to analyze:** |
|  | **1.Display first 5 entries from movies table (Brandon)** |
|  | **2.Count number of unique movies (Melissa)** |
|  | **3.Which movie received the highest number of ratings (Brandon)** |
|  | **4.Top 25 most rated movies etc. (Melissa)**  **5. Which movie received the lowest number of ratings** |
|  |  |
|  | Step 1: Verify the Dataset [https://grouplens.org/datasets/movielens/] |
|  | Step 2: Load Data into HDFS [https://cloudxlab.com/assessment/displayslide/326/hive-movielens-assignment] |
|  | Step 3: Create tables in Apache Hive |
|  | Step 4: Load data into Hive table |
|  | Step 5: Query to analyze and display the result [https://github.com/Crone1/Pig-and-Hive-MovieLens-Analysis/blob/main/Hive\_movieLens\_analysis.hive] |
|  |  |
|  | Ref Links: https://www.youtube.com/watch?v=4tW9HY0Q460 |
|  | https://github.com/Crone1/Pig-and-Hive-MovieLens-Analysis/blob/main/Hive\_movieLens\_analysis.hive |
|  | https://towardsdatascience.com/getting-started-with-hive-ad8a93862f1a |
|  | https://medium.com/@notesharsha/13-steps-to-import-data-into-hive-from-your-local-drive-using-ambari-sandbox-hiveql-bonus-a44bffae89d7 |
|  |  |

Citation:

CITATION

==============================================

To acknowledge use of the dataset in publications, please cite the

following paper:

F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets:

History and Context. ACM Transactions on Interactive Intelligent

Systems (TiiS) 5, 4, Article 19 (December 2015), 19 pages.

DOI=http://dx.doi.org/10.1145/2827872

ACKNOWLEDGEMENTS

==============================================

Thanks to Al Borchers for cleaning up this data and writing the

accompanying scripts.

PUBLISHED WORK THAT HAS USED THIS DATASET

==============================================

Herlocker, J., Konstan, J., Borchers, A., Riedl, J.. An Algorithmic

Framework for Performing Collaborative Filtering. Proceedings of the

1999 Conference on Research and Development in Information

Retrieval. Aug. 1999.

FURTHER INFORMATION ABOUT THE GROUPLENS RESEARCH PROJECT

==============================================

The GroupLens Research Project is a research group in the Department

of Computer Science and Engineering at the University of Minnesota.

Members of the GroupLens Research Project are involved in many

research projects related to the fields of information filtering,

collaborative filtering, and recommender systems. The project is lead

by professors John Riedl and Joseph Konstan. The project began to

explore automated collaborative filtering in 1992, but is most well

known for its world wide trial of an automated collaborative filtering

system for Usenet news in 1996. The technology developed in the

Usenet trial formed the base for the formation of Net Perceptions,

Inc., which was founded by members of GroupLens Research. Since then

the project has expanded its scope to research overall information

filtering solutions, integrating in content-based methods as well as

improving current collaborative filtering technology.

Further information on the GroupLens Research project, including

research publications, can be found at the following web site:

http://www.grouplens.org/

GroupLens Research currently operates a movie recommender based on

collaborative filtering:

http://www.movielens.org/

Discord Instructions

Flow chart for p1:

1. Form team of 6 members

2. frame 4/5 questions to query

3. Load the data in hdfs

4. create tables from the imported files in Hive

5. create queries to analyze

6. display through o/p

7. Update Github repo

8. Prep a 5 slide Powerpoint

9. Rehearse the presentation

SELECT u\_data.movieid, SUM(u\_data.rating) AS rating, u\_item.movietitle FROM u\_data INNER JOIN u\_item ON u\_data.movieid=u\_item.movieid GROUP BY u\_data.movieid, u\_item.movietitle ORDER BY rating DESC LIMIT 30;

SELECT u\_data.movieid, SUM(u\_data.rating) AS rating, u\_item.movietitle FROM u\_data INNER JOIN u\_item ON u\_data.movieid=u\_item.movieid GROUP BY u\_data.movieid, u\_item.movietitle ORDER BY rating ASC LIMIT 30;