BDT
MINI PROJECT REPORT

Topic: Movie Analysis
BDT-3

Group 4 members:

PA38 Kamakshi Sarbhai 1032211505
PA39 Shivam Gupta 1032211510
PH46 Shreya Deshpande 1032211616
PA29 Nicole Lobe 1032211344

Contents:

Overview of the Tech Stack Used:

Pig, Hive, and Hue are three key components in the Cloudera environment, offering different functionalities for working with big data in Hadoop. Here's a detailed overview of each and how they operate both individually and together in the Cloudera ecosystem.

Pig:

Overview:

Pig is a high-level platform for processing and analyzing large datasets in Apache Hadoop. It provides a scripting language called Pig Latin, which abstracts the complexity of MapReduce programming. Pig Latin allows developers to focus on the logic of their data analysis tasks rather than the intricacies of the underlying Hadoop framework.

Functionality:

- Data Flow Language: Pig Latin is a procedural language that describes data transformations and analysis steps. Users write scripts to perform data operations like loading, filtering, transforming, and storing data.
- Optimization: Pig optimizes the execution of operations by converting Pig Latin scripts into sequences of MapReduce jobs, which are then executed on the Hadoop cluster.
- Extensibility: It allows for user-defined functions (UDFs) written in Java, Python, or other languages, enabling custom data processing logic to be integrated into Pig scripts.

Use Cases:
- ETL (Extract, Transform, Load) processes
- Data cleaning and transformation
- Data processing for analytics

Hive:

Overview:
Hive is a data warehouse infrastructure built on top of Hadoop. It provides a SQL-like interface (HiveQL) to query and manage large datasets stored in Hadoop's distributed storage (HDFS) or other compatible file systems. Hive translates HiveQL queries into MapReduce or Tez jobs for execution on the Hadoop cluster.

Functionality:
- SQL Interface: HiveQL allows users to write queries similar to SQL for data querying, manipulation, and analysis.
- Schema on Read: Hive provides a schema on read approach, allowing users to apply a structure to data during query execution rather than at the time of data ingestion.
- Metastore: It utilizes a metastore to store metadata information such as table schemas, partitions, and storage location, improving query performance by reducing metadata retrieval overhead.

Use Cases:
- Ad-hoc querying and analysis
- Business intelligence and reporting
- Data exploration and analysis using SQL-like syntax

Hue:

Overview:
Hue (Hadoop User Experience) is a web-based graphical user interface (GUI) that simplifies working with Hadoop and related technologies. It acts as a comprehensive interface to interact with various components of the Hadoop ecosystem, providing a user-friendly experience for developers, data analysts, and administrators.

Functionality:
- Query Editor: Hue includes editors for Hive queries, Pig scripts, and other tools, offering a comfortable environment for writing and executing code.
- Workflow Management: It allows users to create workflows (using Oozie) for orchestrating tasks and dependencies across different Hadoop components.

- File Browser: Users can navigate, upload, and manage files stored in Hadoop Distributed File System (HDFS) through Hue's interface.
- Job Dashboard: Monitoring and managing Hadoop jobs and clusters through a centralized dashboard.

Use Cases:
- Interactive data exploration and analysis
- Query development and execution
- Job and workflow management

- **Integration in Cloudera Environment:**

In the Cloudera ecosystem, these tools work together to offer a comprehensive data processing and management platform:

- Integration with Cloudera Manager: Cloudera Manager provides centralized management and monitoring of the Hadoop cluster, allowing administrators to configure and manage Pig and Hive services.
- Usage in Workflows: Hue's workflow capabilities can orchestrate tasks that involve both Pig and Hive, allowing users to create complex data pipelines.
- Compatibility: Pig and Hive can utilize Hue as an interface for writing scripts, executing queries, and managing jobs, providing a unified experience for users.

Pig, Hive, and Hue are integral parts of the Cloudera ecosystem, each serving specific purposes in processing, querying, and managing big data. Their integration offers users a powerful suite of tools for handling large datasets in Hadoop environments, simplifying development, analysis, and management tasks while leveraging the capabilities of the underlying distributed computing framework.

```
grunt> most_liked_movie = FOREACH (GROUP average_movie_ratings ALL) {
>> filtered = FILTER average_movie_ratings BY rating_average >= (float)4.6;
>> ordered_movies = ORDER filtered BY num_ratings DESC;
>> most_rated_movie_with_specified_avg = LIMIT ordered_movies 1;
>>     GENERATE FLATTEN(most_rated_movie_with_specified_avg);
>> }
2023-11-30 02:54:12,193 [main] WARN  org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 1 time(s).
2023-11-30 02:54:12,193 [main] WARN  org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_FLOAT 1 time(s).
2023-11-30 02:54:12,193 [main] WARN  org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_LONG 1 time(s).
grunt> DUMP most_liked_movie
2023-11-30 02:54:29,612 [main] WARN  org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_FLOAT 1 time(s).
2023-11-30 02:54:29,612 [main] WARN  org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_LONG 1 time(s).
2023-11-30 02:54:29,614 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2023-11-30 02:54:29,617 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune,
DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach,
NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier,
PartitionFilterOptimizer]}
2023-11-30 02:54:29,767 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic?
false
2023-11-30 02:54:29,806 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.CombinerOptimizer - Choosing to move algebraic foreach to
combiner
2023-11-30 02:54:29,818 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 3
2023-11-30 02:54:29,818 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 3
2023-11-30 02:54:29,829 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.AccumulatorOptimizer - Reducer is to run in accumulative
mode.
2023-11-30 02:54:29,924 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2023-11-30 02:54:29,956 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2023-11-30 02:54:30,045 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler -
mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2023-11-30 02:54:30,046 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating #
of required reducers.
2023-11-30 02:54:30,050 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator:
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2023-11-30 02:54:30,070 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000
```

```
grunt> group_users = GROUP cleaned_movies_nd_ratings BY userId;
2023-11-30 02:19:52,612 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-11-30 02:19:52,612 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use
mapreduce.jobtracker.address
grunt> average_user_rating = FOREACH group_users GENERATE group As userId, SUM(cleaned_movies_nd_ratings.rating) as num_ratings,
AVG(cleaned_movies_nd_ratings.rating) AS average_rating;
grunt> user_highest_avg = FOREACH (GROUP average_user_rating ALL) {
>> filtered = FILTER average_user_rating BY average_rating == (float)5.0;
>> ordered = ORDER filtered by num_ratings DESC;
>> limited = LIMIT ordered 1;
>> GENERATE FLATTEN(limited);
>> }
2023-11-30 02:20:44,966 [main] WARN  org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 1 time(s).
grunt> dump user_highest_avg
2023-11-30 02:21:21,481 [main] WARN  org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 1 time(s).
2023-11-30 02:21:21,482 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2023-11-30 02:21:21,484 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune,
DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach,
NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier,
PartitionFilterOptimizer]}
2023-11-30 02:21:21,541 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic
false
2023-11-30 02:21:21,563 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.CombinerOptimizer - Choosing to move algebraic foreach to
combiner
2023-11-30 02:21:21,575 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 2
2023-11-30 02:21:21,575 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 2
2023-11-30 02:21:21,611 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2023-11-30 02:21:21,616 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2023-11-30 02:21:21,640 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler -
mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2023-11-30 02:21:21,658 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating #
of required reducers.
```

```
grunt> read_ratings = LOAD '/bdtProjectFinal/ratings.csv' using PigStorage(',') AS (userId:int, movieId:int, rating:int, timestamp:int);
2023-11-30 01:32:41,222 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-11-30 01:32:41,222 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use
mapreduce.jobtracker.address
grunt> comma_del_ratings = FILTER read_ratings BY (userId IS NOT NULL) AND (movieId IS NOT NULL);
grunt> STORE comma_del_ratings INTO 'output/piped_ratings' using PigStorage('|');
2023-11-30 01:33:41,985 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: FILTER
2023-11-30 01:33:41,987 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune,
DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach,
NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier,
PartitionFilterOptimizer]}
2023-11-30 01:33:42,048 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimist
false
2023-11-30 01:33:42,050 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization:
2023-11-30 01:33:42,051 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization:
2023-11-30 01:33:42,113 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2023-11-30 01:33:42,122 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2023-11-30 01:33:42,166 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler -
mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2023-11-30 01:33:44,242 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - creating jar file
Job7440349418838190862.jar
2023-11-30 01:33:57,049 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - jar file Job7440349418838190862.ja
created
2023-11-30 01:33:57,134 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2023-11-30 01:33:57,136 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2023-11-30 01:33:57,136 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2023-11-30 01:33:57,136 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2023-11-30 01:33:57,214 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for
submission.
2023-11-30 01:33:57,214 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use
mapreduce.jobtracker.address
2023-11-30 01:33:57,222 [JobControl] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2023-11-30 01:33:57,251 [JobControl] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-11-30 01:33:58,036 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
```

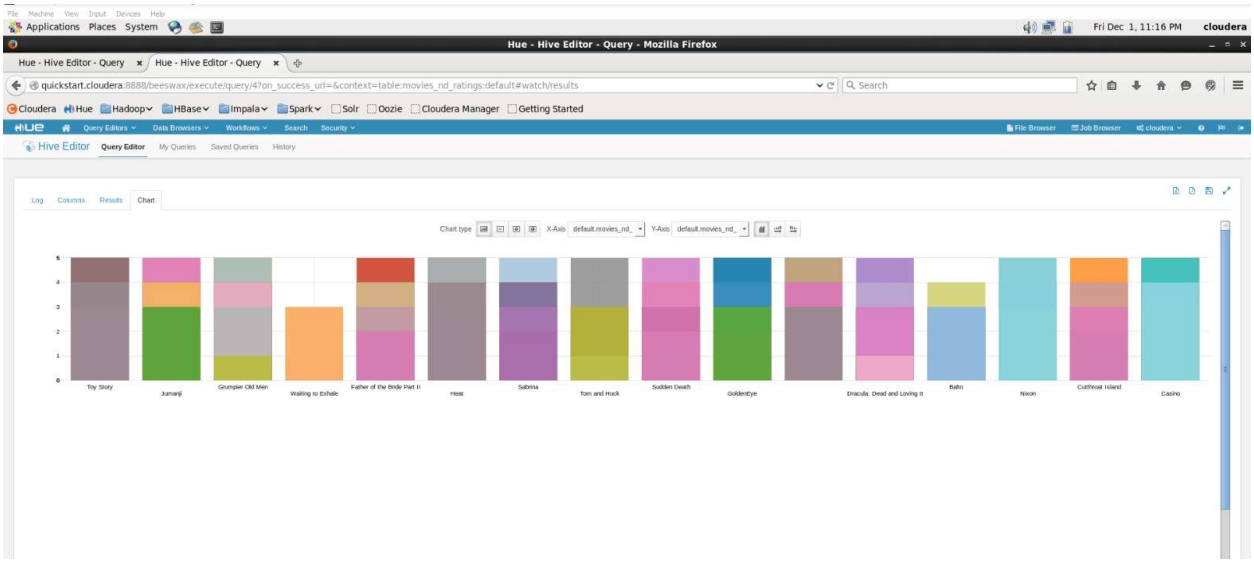
```
grunt> movies_with_pipe = LOAD 'output/piped_movies' using PigStorage('|') AS (movieId:int, title:chararray, genres:chararray);
2023-11-30 01:39:43,391 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-11-30 01:39:43,391 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use
mapreduce.jobtracker.address
grunt> part_cleaned_movies = FOREACH movies_with_pipe GENERATE movieId AS movieId, REPLACE (title, '@@@', ',') AS title, genres AS genres;
grunt> fully_cleaned_movies = FOREACH part_cleaned_movies GENERATE
>>     movieId,
>>     REGEX_EXTRACT(title, '([\\S ]+) \\(\\d{4}|\\d{4}-?\\d{4})\\)', 1) AS title,
>>     REGEX_EXTRACT(title, '\\(\\d{4}|(\\d{4}-\\d{4}))\\)', 1) AS year,
>>     genres;
grunt> STORE fully_cleaned_movies INTO 'output/cleaned_movies' using PigStorage('|');
2023-11-30 01:40:59,092 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2023-11-30 01:40:59,094 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune,
DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach,
NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier,
PartitionFilterOptimizer]}
2023-11-30 01:40:59,130 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic?
false
2023-11-30 01:40:59,134 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2023-11-30 01:40:59,134 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2023-11-30 01:40:59,183 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2023-11-30 01:40:59,189 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2023-11-30 01:40:59,212 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler -
mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2023-11-30 01:41:00,986 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - creating jar file
Job7433603926229813548.jar
2023-11-30 01:41:11,836 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - jar file Job7433603926229813548.jar
created
2023-11-30 01:41:11,889 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2023-11-30 01:41:11,892 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2023-11-30 01:41:11,892 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2023-11-30 01:41:11,896 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2023-11-30 01:41:11,989 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for
submission.
```

```
//ADDING DATASET TO HDFS

[cloudera@quickstart ~]$ hadoop fs -ls /
Found 8 items
drwxr-xr-x   - cloudera supergroup          0 2023-11-29 06:58 /bdt_project
drwxr-xr-x   - hbase    supergroup          0 2023-11-26 03:09 /hbase
drwxr-xr-x   - cloudera supergroup          0 2023-11-29 06:23 /input
drwxr-xr-x   - cloudera supergroup          0 2023-11-26 07:10 /project
drwxr-xr-x   - solr     solr                0 2015-06-09 03:38 /solr
drwxrwxrwx   - hdfs     supergroup          0 2023-11-29 08:13 /tmp
drwxr-xr-x   - hdfs     supergroup          0 2015-06-09 03:38 /user
drwxr-xr-x   - hdfs     supergroup          0 2015-06-09 03:36 /var
[cloudera@quickstart ~]$ hadoop fs -mkdir /bdtProjectFinal
[cloudera@quickstart ~]$ hadoop fs -put /home/cloudera/Desktop/data/* /bdtProjectFinal
[cloudera@quickstart ~]$ hadoop fs -ls /bdtProjectFinal
Found 2 items
-rw-r--r--   1 cloudera supergroup     494431 2023-11-30 01:15 /bdtProjectFinal/movies.csv
-rw-r--r--   1 cloudera supergroup    2483723 2023-11-30 01:15 /bdtProjectFinal/ratings.csv



//STARTING THE PIG
[cloudera@quickstart ~]$ pig -x local
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
2023-11-30 01:20:30,144 [main] INFO  org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.4.2 (rexported) compiled May 19 2015, 17:03:41
2023-11-30 01:20:30,151 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/cloudera/pig_1701336029924.log
2023-11-30 01:20:30,352 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/cloudera/.pigbootup not found
2023-11-30 01:20:32,350 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-11-30 01:20:32,351 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use
mapreduce.jobtracker.address
2023-11-30 01:20:32,360 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
```


Hue - Hive Editor - Query - Mozilla Firefox
Hive Editor
Query Editor
Chart
Toy Story
Jumanji
Grumpier Old Men
Waiting to Exhale
Father of the Bride Part II
Heat
Sabrina
Tom and Huck
Sudden Death
GoldenEye
Dracula: Dead and Loving It
Balto
Nixon
Cutthroat Island
Casino

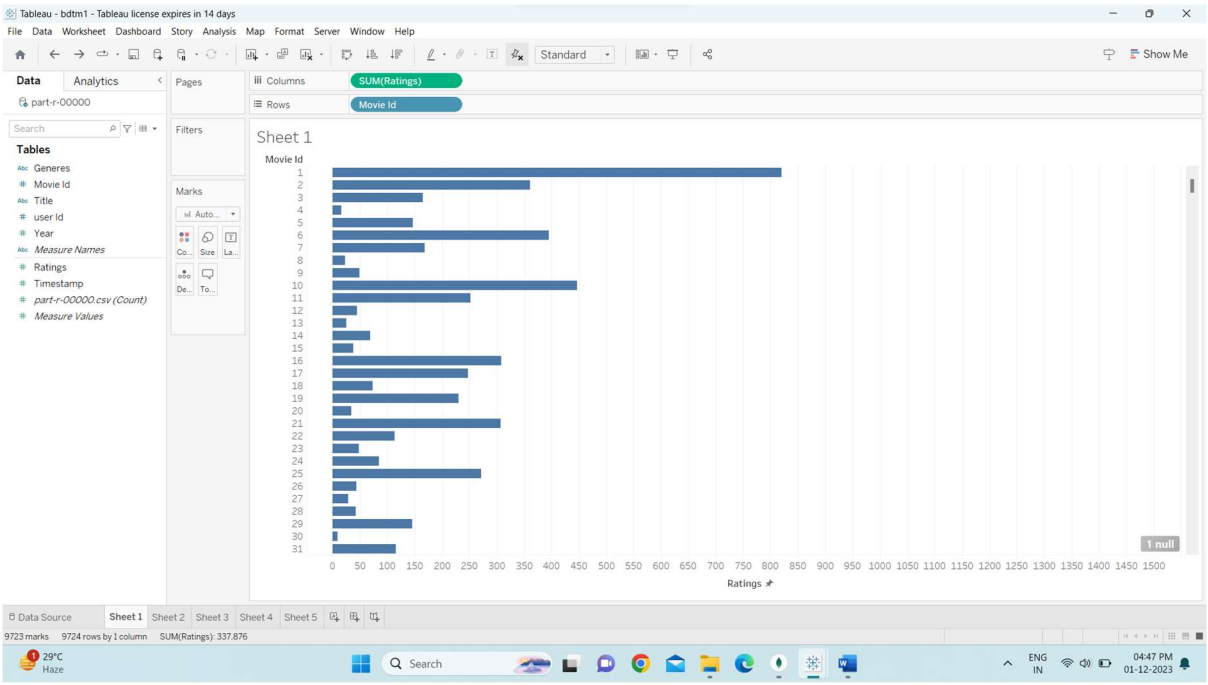


Tableau - bdtm1 - Tableau license expires in 14 days
Data
Analytics
part-r-00000
Tables
Generes
Movie Id
Title
user Id
Year
Measure Names
Ratings
Timestamp
part-r-00000.csv (Count)
Measure Values
Pages
Filters
Marks
Columns
SUM(Ratings)
Rows
Movie Id
Sheet 1
Ratings
1 null
Data Source
Sheet 1
Sheet 2
Sheet 3
Sheet 4
Sheet 5
9723 marks
9724 rows by 1 column
SUM(Ratings): 337,876

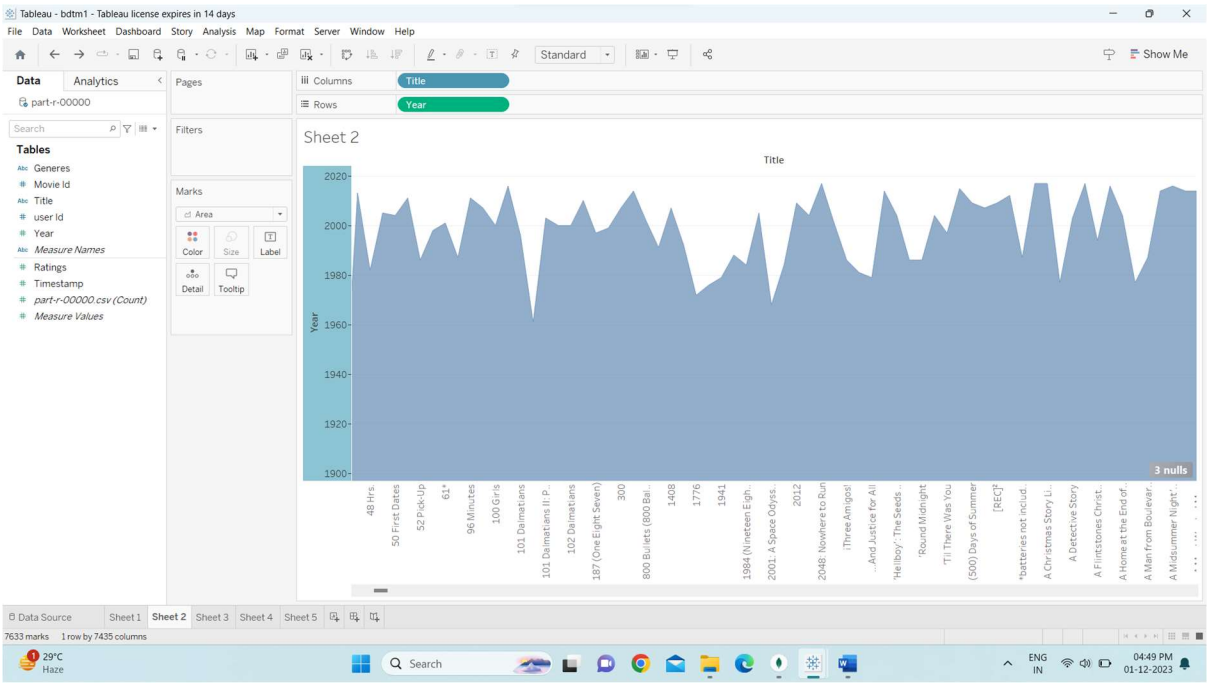


Tableau - bdtm1 - Tableau license expires in 14 days
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Data Analytics
part-r-00000
Tables
Generes
Movie Id
Title
user Id
Year
Measure Names
Ratings
Timestamp
part-r-00000.csv (Count)
Measure Values
Pages
Filters
Marks
Area
Color Size Label
Detail Tooltip
Columns Title
Rows Year
Sheet 2
Title
2020
2000
1980
1960
1940
1920
1900
Year
3 nulls
48 Hrs.
50 First Dates
52 Pick-Up
61*
96 Minutes
100 Girls
101 Dalmatians
101 Dalmatians II: P...
102 Dalmatians
187 (One Eight Seven)
300
800 Bullets (800 Bal...
1408
1776
1941
1984 (Nineteen Eigh...
2001: A Space Odyss...
2012
2048: Nowhere to Run
'Three Amigos!
...And Justice for All
'Hellboy': The Seeds...
'Round Midnight
'Til There Was You
(500) Days of Summer
[REC]²
*batteries not includ...
A Christmas Story Li...
A Detective Story
A Flintstones Christ...
A Home at the End of...
A Man from Boulevar...
A Midsummer Night's...
Data Source
Sheet 1
Sheet 2
Sheet 3
Sheet 4
Sheet 5
7633 marks
1 row by 7435 columns
29°C
Haze
Search
ENG
IN
04:49 PM
01-12-2023

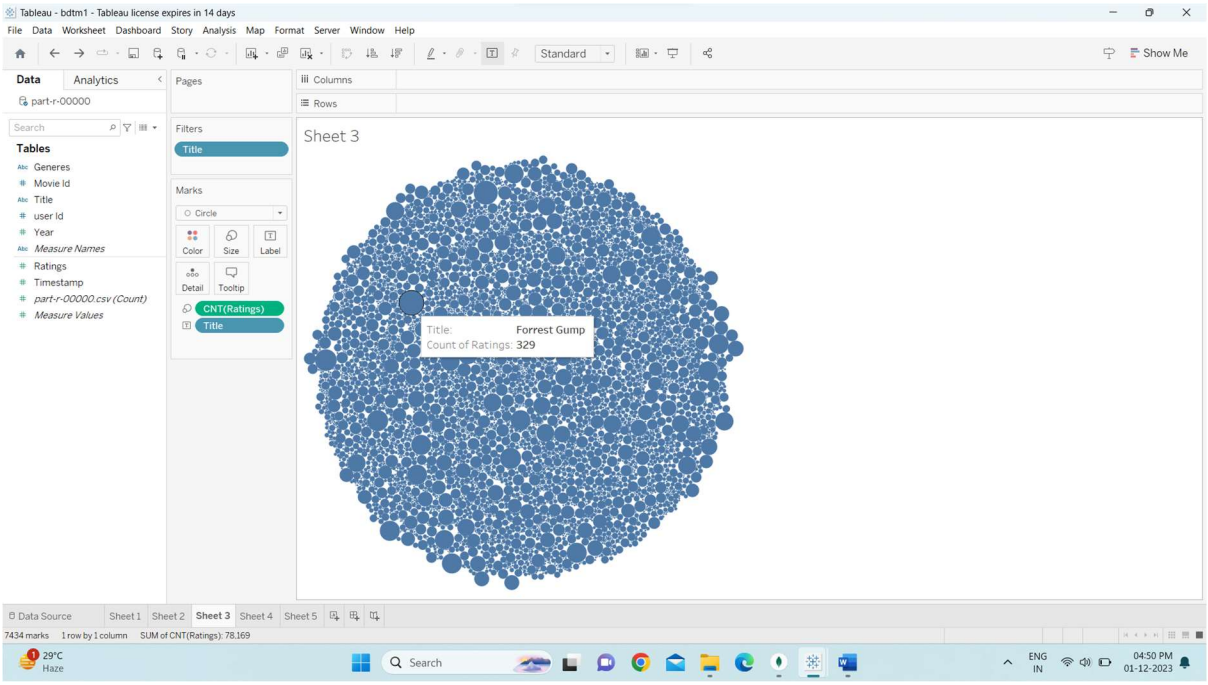


Tableau - bdtm1 - Tableau license expires in 14 days
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Data Analytics
part-r-00000
Tables
Generes
Movie Id
Title
user Id
Year
Measure Names
Ratings
Timestamp
part-r-00000.csv (Count)
Measure Values
Pages
Filters
Title
Marks
Circle
Color Size Label
Detail Tooltip
CNT(Ratings)
Title
Columns
Rows
Sheet 3
Title: Forrest Gump
Count of Ratings: 329
Data Source
Sheet 1
Sheet 2
Sheet 3
Sheet 4
Sheet 5
7434 marks
1 row by 1 column
SUM of CNT(Ratings): 78,169
29°C
Haze
Search
ENG
IN
04:50 PM
01-12-2023

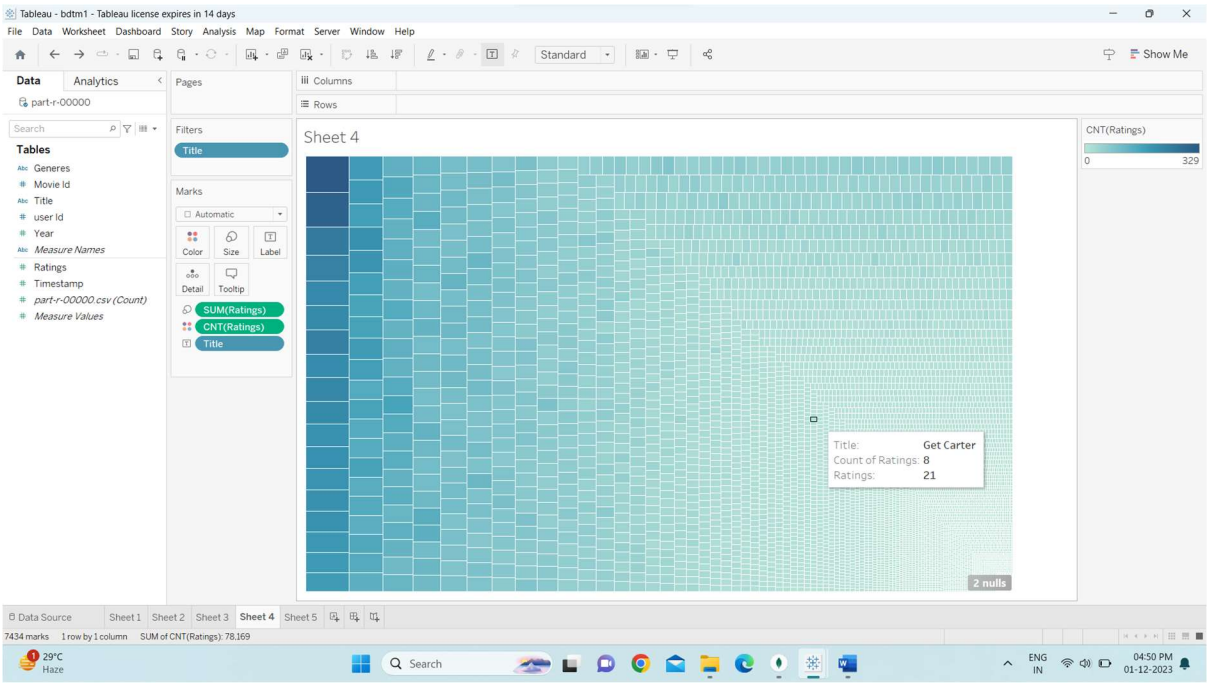


Tableau - bdtm1 - Tableau license expires in 14 days
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Standard
Show Me
Data
Analytics
part-r-00000
Search
Tables
Generes
Movie Id
Title
user Id
Year
Measure Names
Ratings
Timestamp
part-r-00000.csv (Count)
Measure Values
Pages
Filters
Title
Marks
Automatic
Color
Size
Label
Detail
Tooltip
SUM(Ratings)
CNT(Ratings)
Title
Columns
Rows
Sheet 4
CNT(Ratings)
0
329
Title: Get Carter
Count of Ratings: 8
Ratings: 21
2 nulls
Data Source
Sheet 1
Sheet 2
Sheet 3
Sheet 4
Sheet 5
7434 marks
1 row by 1 column
SUM of CNT(Ratings): 78.169
29°C
Haze
Search
ENG
IN
04:50 PM
01-12-2023

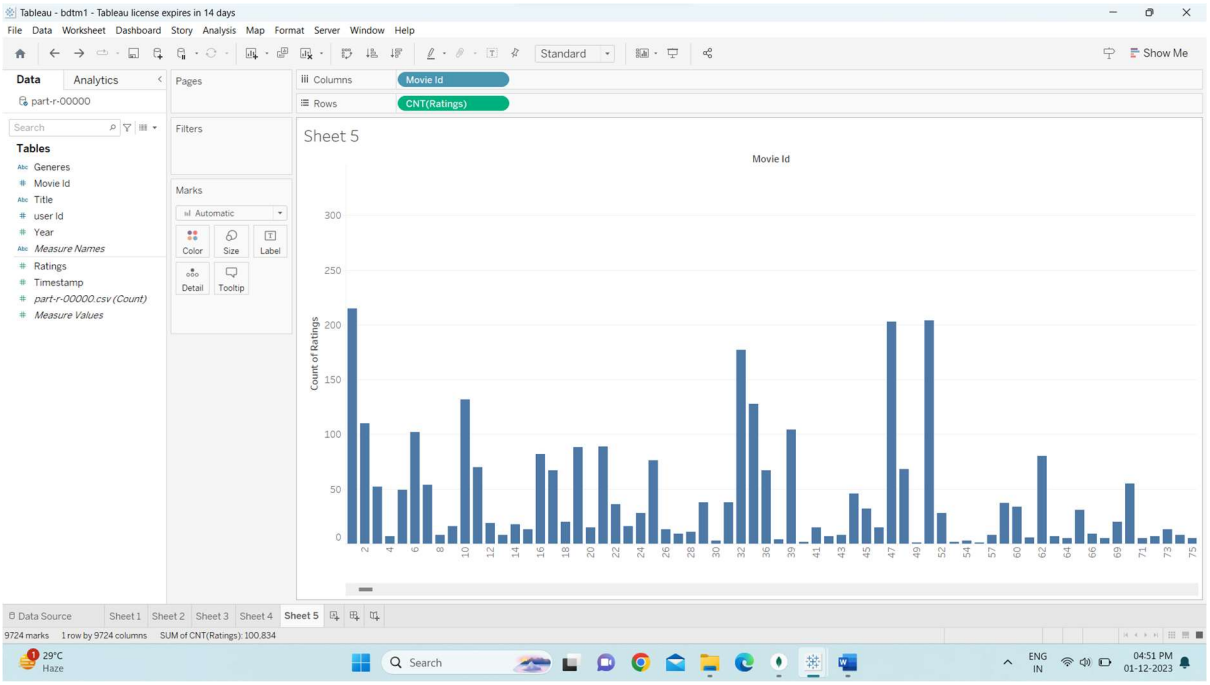
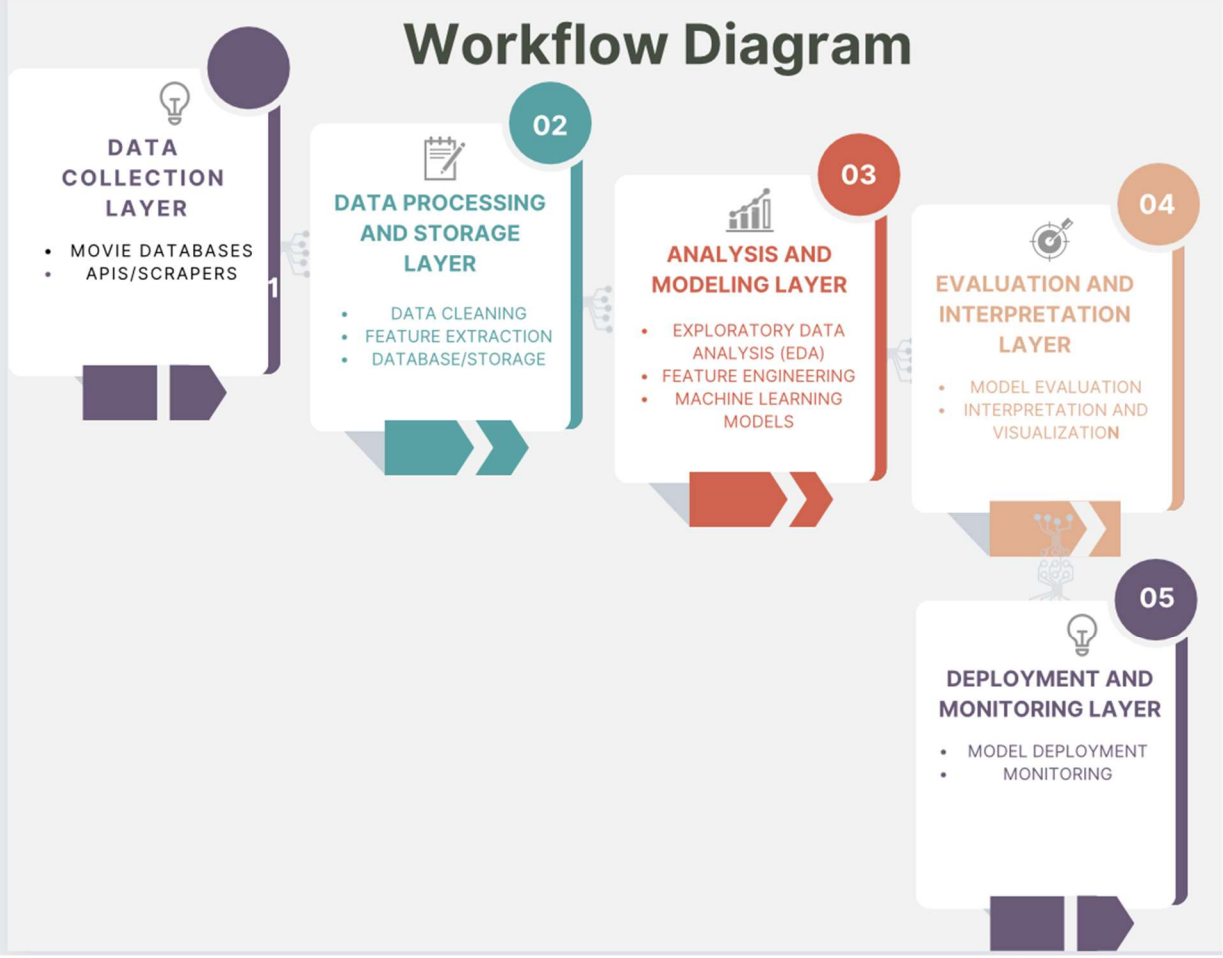


Tableau - bdtm1 - Tableau license expires in 14 days
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Standard
Show Me
Data
Analytics
part-r-00000
Search
Tables
Generes
Movie Id
Title
user Id
Year
Measure Names
Ratings
Timestamp
part-r-00000.csv (Count)
Measure Values
Pages
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Columns
Movie Id
Rows
CNT(Ratings)
Sheet 5
Movie Id
Count of Ratings
300
250
200
150
100
50
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
36
39
41
43
45
47
49
52
54
57
60
62
64
66
69
71
73
75
Data Source
Sheet 1
Sheet 2
Sheet 3
Sheet 4
Sheet 5
9724 marks
1 row by 9724 columns
SUM of CNT(Ratings): 100,834
29°C
Haze
Search
ENG
IN
04:51 PM
01-12-2023

Workflow/Architecture Diagram:



Future Scope:

Enhanced Data Processing with Pig:

- **Complex Data Transformation:** Leverage Pig for complex data transformations and custom processing of movie-related data. Utilize its scripting language to refine and prepare data for analysis, enabling more sophisticated feature engineering.

Advanced Analytics with Hive:

- **Performance Optimization:** Explore Hive's optimizations like partitioning, indexing, and bucketing to enhance query performance on movie datasets, particularly as data volumes grow.
- **Integration of External Data:** Incorporate external datasets into Hive, such as demographic or geographic information, to enrich movie analysis and provide a broader context for ratings.

Expanded Visualization with Hue:

- **Interactive Dashboards:** Enhance Hue's visualization capabilities by creating interactive dashboards that allow stakeholders to dynamically explore movie ratings, trends, and correlations.
- **Real-time Monitoring:** Develop visual monitoring interfaces in Hue to track real-time changes in ratings, user sentiments, or box office performance.

Sentiment Analysis Integration:

- **Text Mining Techniques:** Implement sentiment analysis using Pig or Hive to extract sentiment from user reviews and social media data related to movies. Integrate these sentiments into rating predictions for more nuanced insights.

Machine Learning Improvements:

- **Advanced Modeling:** Experiment with more advanced machine learning models within Hive or with external integrations to predict ratings more accurately. Consider ensemble methods or deep learning architectures for improved predictions.

Collaboration with Streaming Platforms:

- **Real-time Data Integration:** Collaborate with streaming platforms to integrate real-time data, allowing the models to adapt swiftly to changing trends and user preferences.

Conclusion:

The movie rating analysis project utilizing Hive, Pig, and Hue showcased the potential of Hadoop's ecosystem in handling large-scale movie-related datasets and deriving insights:

- **Data Processing Efficiency:** Hive facilitated SQL-like querying and analysis over large volumes of structured data, optimizing performance through storage optimizations.
- **Custom Data Processing:** Pig's scripting language enabled custom data transformations, refining data for analysis and preparing it for modeling.
- **User-Friendly Interface:** Hue offered an intuitive interface for data exploration, basic visualization, and job monitoring within the Hadoop environment.

Key Findings:

- Identified influential factors affecting movie ratings based on comprehensive data analysis.
- Developed predictive models leveraging machine learning techniques to forecast movie ratings.

Challenges and Recommendations:

- **Scalability Concerns:** As the dataset grows, optimizing queries and processing with Hive and Pig becomes crucial for maintaining performance.
- **Integration Complexity:** Integrating real-time data streams and external sources requires robust data pipelines and continual monitoring.

Future Direction:

- The project lays the groundwork for future enhancements, including sentiment analysis integration, advanced analytics, and collaborations with streaming platforms, aiming for more accurate and real-time movie rating predictions.

References:

1. https://www.javatpoint.com/hive
2. https://www.geeksforgeeks.org/introduction-to-apache-pig/
3. https://github.com/Crone1/Pig-and-Hive-MovieLens-Analysis/blob/main/Pig_movieLens_analysis.pig