

AMRITSAR GROUP OF COLLEGES

**Autonomous status conferred by UGC under UGC act-1956, (2f), NAAC-A Grade,
(Formerly Known as Amritsar College of Engineering & Technology | Amritsar Pharmacy College)**

Project Report

On

“ANALYSIS OF NETFLIX DATASET”

Submitted in the Partial fulfilment of the requirement for the Award of Degree of

Bachelor of Technology

in

COMPUTER SCIENCE & ENGINEERING

Batch (2019-23)

Submitted To:

Er. Ajay Sharma

Submitted By:-

Salil Chandan (1900250)

Seijal Bhalla (1900259)

Tamanna Sharma (1900285)

Tashmeen (1900289)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Amritsar Group of Colleges, Amritsar

ACKNOWLEDGEMENT

We are extremely grateful to **Er. Ajay Sharma**, Associate Professor of Department of Computer Science for his able guidance and useful suggestions, which helped us in completing the project work on time. It would not have been possible to see through the undertaken project without his guidance.

A project is major milestone during the study period of a student. As such this project was a challenge to four of us and was an opportunity to prove our caliber.

We are very thankful to **Dr. Amarpreet Singh, Head of Department** and all the faculty members of Computer Science Department who gave us an opportunity to face real time problems.

DECLARATION

We Salil, Seijal, Tamanna and Tashmeen hereby as a team declare that the project work entitled **“ANALYSIS OF NETFLIX DATASET”** is an authentic record of our own work carried out as per the requirements of Big Data Analytics Lab (Part-B) for the award of degree of **B.Tech (CSE), Amritsar Group of Colleges, Amritsar**, under the guidance of **Er. Ajay Sharma (Associate Professor)**.

Salil Chandan

Univ Roll No: 1900250

Seijal Bhalla

Univ Roll No: 1900259

Tamanna Sharma

Univ Roll No: 1900285

Tashmeen

Univ Roll No: 1900289

TALE OF CONTENTS

Sr. No	Content	Page no.
1	Introduction to Big Data Analytics	
2	Introduction to Apache Hadoop	
3	Introduction to Apache Pig	
4	Objectives of the Project	
5	Queries along with their Snapshots	
6	Case Study 1	
7	Case Study 2	
8	References	

INTRODUCTION TO BIG DATA ANALYTICS

BIG DATA



What is Big Data Analytics?

Big data analytics definition: Big data analytics helps businesses and organisations make better decisions by revealing information that would have otherwise been hidden.

Meaningful insights about the trends, correlations and patterns that exist within big data can be difficult to extract without vast computing power. But the techniques and technologies used in big data analytics make it possible to learn more from large data sets. This includes data of any source, size and structure.

Structured vs. Unstructured Data

Before we deep dive into the nuances of Big Data, it is important to understand the different kinds of data, namely structured and unstructured data.

Structured data includes quantitative data that is stored in an organised manner. It consists of numerical and text data. It is easy to analyse and process structured data. It is generally stored in a relational database and can be queried using Structured Query Language (SQL).

Unstructured data includes qualitative data that lacks any predefined structure and can come in a variety of formats (images, mp3 files, wav files, etc.). Unstructured data is said to lack “structure”. It is stored in a non-relational database and can be queried using NoSQL.

There can be semi-structured data as well, which lies somewhat in between structured and unstructured data.

Unstructured VS Structured Data



What are the 5 Vs of Big Data?

Doug Laney introduced this concept of 3 Vs of Big Data, viz. Volume, Variety, and Velocity.

Volume

refers to the amount of data that is being collected. The data could be structured or unstructured.

Velocity

refers to the rate at which data is coming in.

Variety

refers to the different kinds of data (data types, formats, etc.) that is coming in for analysis.

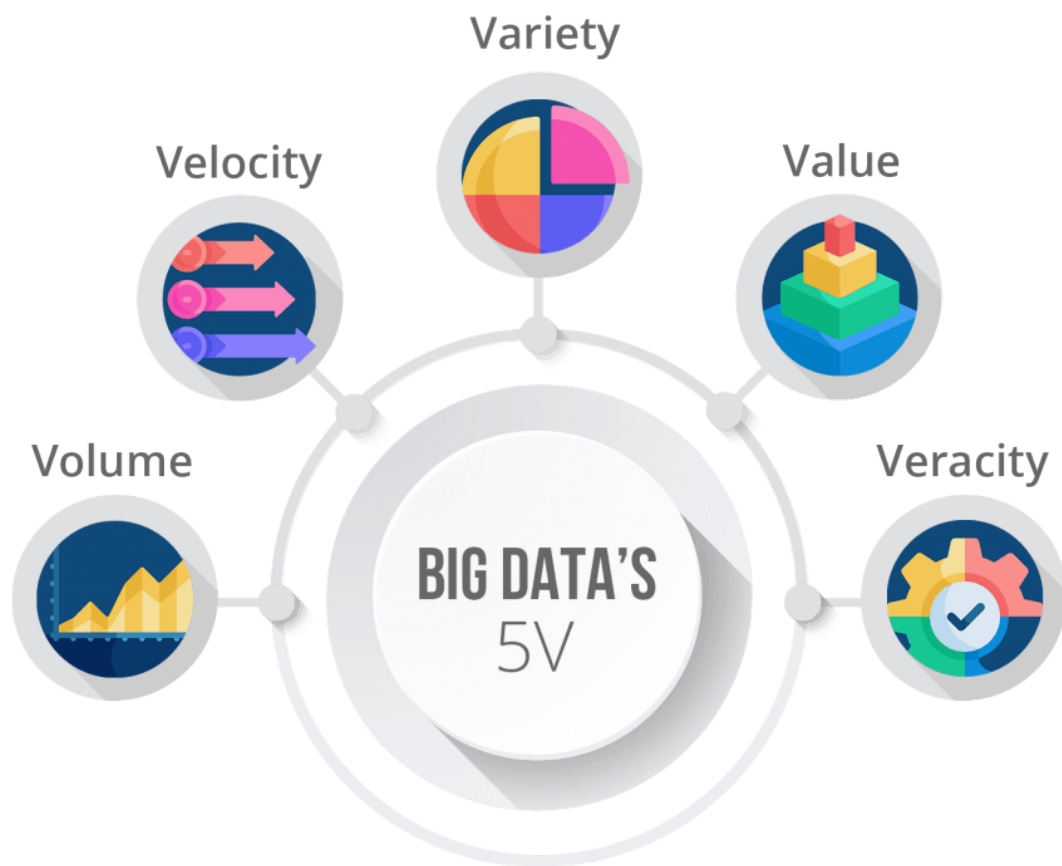
Over the last few years, 2 additional Vs of data have also emerged – value and veracity.

Value

refers to the usefulness of the collected data.

Veracity

refers to the quality of data that is coming in from different sources.



Application Areas of Big Data Analytics

Here is the list of the top 10 industries using big data applications:

1. Banking and Securities
2. Communications, Media and Entertainment
3. Healthcare Providers
4. Education
5. Manufacturing and Natural Resources

6. Government
7. Insurance
8. Retail and Wholesale trade
9. Transportation
10. Energy and Utilities

Big Data Analytics Tools⁸

NoSQL databases, (not-only SQL) or non relational, are mostly used for the collection and analysis of big data. This is because the data in a NoSQL database allows for dynamic organisation of unstructured data versus the structured and tabular design of relational databases.

Big data analytics requires a software framework for distributed storage and processing of big data. The following tools are considered big data analytics software solutions:

- Apache Kafka
- Scalable messaging system that lets users publish and consume large numbers of messages in real time by subscription.
- HBase
- Column-oriented key/value data store that runs on the Hadoop Distributed File System.
- Hive
- Open source data warehouse system for analysing data sets in Hadoop files.
- MapReduce
- Software framework for processing massive amounts of unstructured data in parallel across a distributed cluster.
- Pig
- Open source technology for parallel programming of MapReduce jobs on Hadoop clusters.
- Spark
- Open source and parallel processing framework for running large-scale data analytics applications across clustered systems.
- YARN
- Cluster management technology in second-generation Hadoop.

Tools used in the project are:

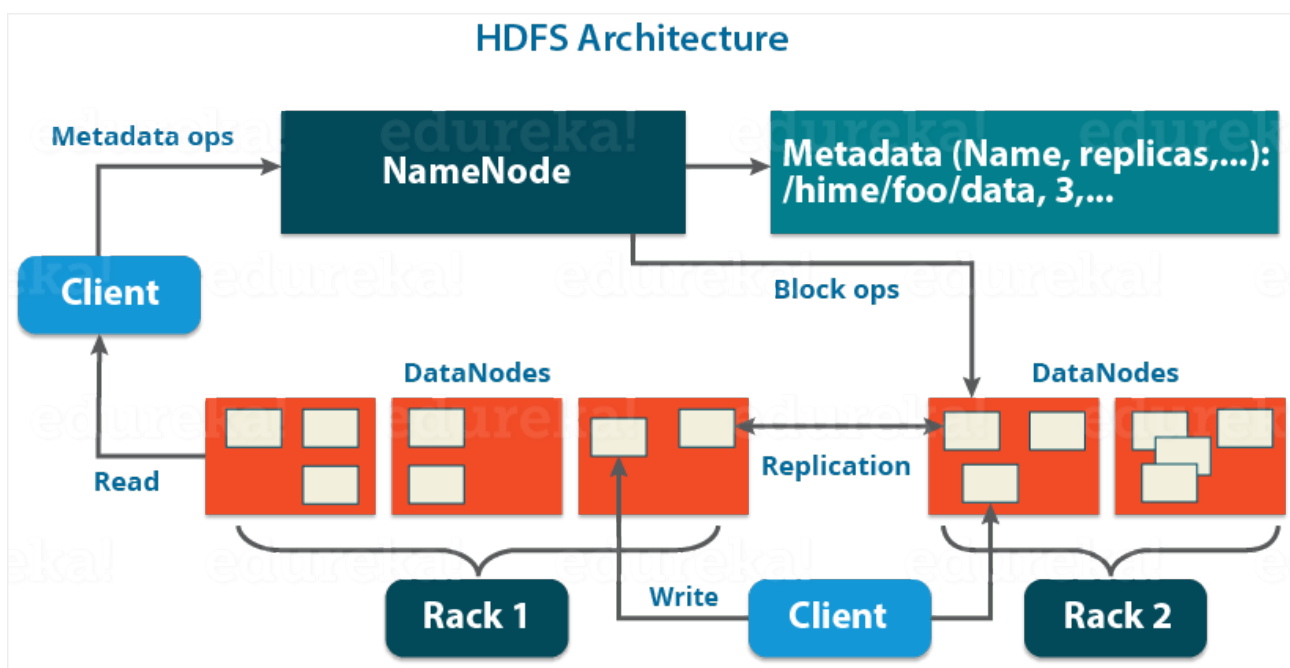
1. HDFS(Hadoop Distributed File System)
2. Apache Pig

INTRODUCTION TO APACHE HADOOP

HDFS Architecture

Apache HDFS or **Hadoop Distributed File System** is a block-structured file system where each file is divided into blocks of a pre-determined size. These blocks are stored across a cluster of one or several machines. Apache Hadoop HDFS Architecture follows a *Master/Slave Architecture*, where a cluster comprises of a single NameNode (Master node) and all the other nodes are DataNodes (Slave nodes). HDFS can be deployed on a broad spectrum of machines that support Java. Though one can run several DataNodes on a single machine, but in the practical world, these DataNodes are spread across various machines.

NameNode:



NameNode is the master node in the Apache Hadoop HDFS Architecture that maintains and manages the blocks present on the DataNodes (slave nodes). NameNode is a very highly available server that manages the File System Namespace and controls access to files by clients. I will be discussing this High Availability feature of Apache Hadoop HDFS in my next blog. The HDFS architecture is built in such a way that the user data never resides on the NameNode. The data resides on DataNodes only.

Functions of NameNode:

- It is the master daemon that maintains and manages the DataNodes (slave nodes)
- It records the metadata of all the files stored in the cluster, e.g. The location of blocks stored, the size of the files, permissions, hierarchy, etc. There are two files associated with the metadata:
 - **FsImage:** It contains the complete state of the file system namespace since the start of the NameNode.
 - **EditLogs:** It contains all the recent modifications made to the file system with respect to the most recent FsImage.

- It records each change that takes place to the file system metadata. For example, if a file is deleted in HDFS, the NameNode will immediately record this in the EditLog.
- It regularly receives a Heartbeat and a block report from all the DataNodes in the cluster to ensure that the DataNodes are live.
- It keeps a record of all the blocks in HDFS and in which nodes these blocks are located.
- The NameNode is also responsible to take care of the **replication factor** of all the blocks which we will discuss in detail later in this HDFS tutorial blog.
- In **case of the DataNode failure**, the NameNode chooses new DataNodes for new replicas, balance disk usage and manages the communication traffic to the DataNodes.

DataNode:

DataNodes are the slave nodes in HDFS. Unlike NameNode, DataNode is a commodity hardware, that is, a non-expensive system which is not of high quality or high-availability. The DataNode is a block server that stores the data in the local file ext3 or ext4.

Functions of DataNode:

- These are slave daemons or process which runs on each slave machine.
- The actual data is stored on DataNodes.
- The DataNodes perform the low-level read and write requests from the file system's clients.
- They send heartbeats to the NameNode periodically to report the overall health of HDFS, by default, this frequency is set to 3 seconds.

Till now, you must have realized that the NameNode is pretty much important to us. If it fails, we are doomed. But don't worry, we will be talking about how Hadoop solved this single point of failure problem in the next Apache Hadoop HDFS Architecture blog. So, just relax for now and let's take one step at a time.

Secondary NameNode:

Apart from these two daemons, there is a third daemon or a process called Secondary NameNode. The Secondary NameNode works concurrently with the primary NameNode as a helper daemon. And don't be confused about the Secondary NameNode being a backup NameNode because it is not.

Functions of Secondary NameNode:

- The Secondary NameNode is one which constantly reads all the file systems and metadata from the RAM of the NameNode and writes it into the hard disk or the file system.
- It is responsible for combining the EditLogs with FsImage from the NameNode.
- It downloads the EditLogs from the NameNode at regular intervals and applies to FsImage. The new FsImage is copied back to the NameNode, which is used whenever the NameNode is started the next time.

Hence, Secondary NameNode performs regular checkpoints in HDFS. Therefore, it is also called CheckpointNode.

Blocks:

Blocks are the nothing but the smallest continuous location on your hard drive where data is stored. In general, in any of the File System, you store the data as a collection of blocks.

Hadoop Ecosystem

Introduction: Hadoop Ecosystem is a platform or a suite which provides various services to solve the big data problems. It includes Apache projects and various commercial tools and solutions. There are four major elements of Hadoop i.e. HDFS, MapReduce, YARN, and Hadoop Common. Most of the tools or solutions are used to supplement or support these major elements. All these tools work collectively to provide services such as absorption, analysis, storage and maintenance of data etc.

Following are the components that collectively form a Hadoop ecosystem:

HDFS: Hadoop Distributed File System

YARN: Yet Another Resource Negotiator

MapReduce: Programming based Data Processing

Spark: In-Memory data processing

PIG, HIVE: Query based processing of data services

HBase: NoSQL Database

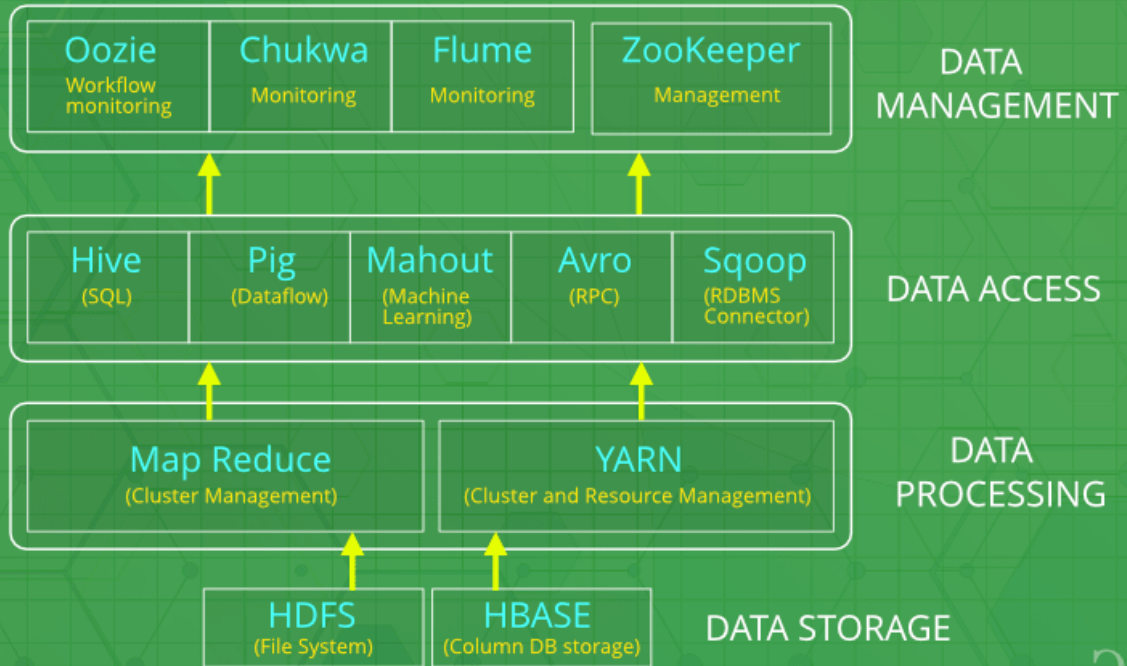
Mahout, Spark MLlib: Machine Learning algorithm libraries

Solar, Lucene: Searching and Indexing

Zookeeper: Managing cluster

Oozie: Job Scheduling

Hadoop Ecosystem



Introduction To Apache Pig

Pig is a high-level platform or tool which is used to process the large datasets. It provides a high-level of abstraction for processing over the MapReduce. It provides a high-level scripting language, known as *Pig Latin* which is used to develop the data analysis codes. First, to process the data which is stored in the HDFS, the programmers will write the scripts using the Pig Latin Language. Internally *Pig Engine* (a component of Apache Pig) converted all these scripts into a specific map and reduce task. But these are not visible to the programmers in order to provide a high-level of abstraction. Pig Latin and Pig Engine are the two main components of the Apache Pig tool. The result of Pig always stored in the HDFS.

Features of apache pig:

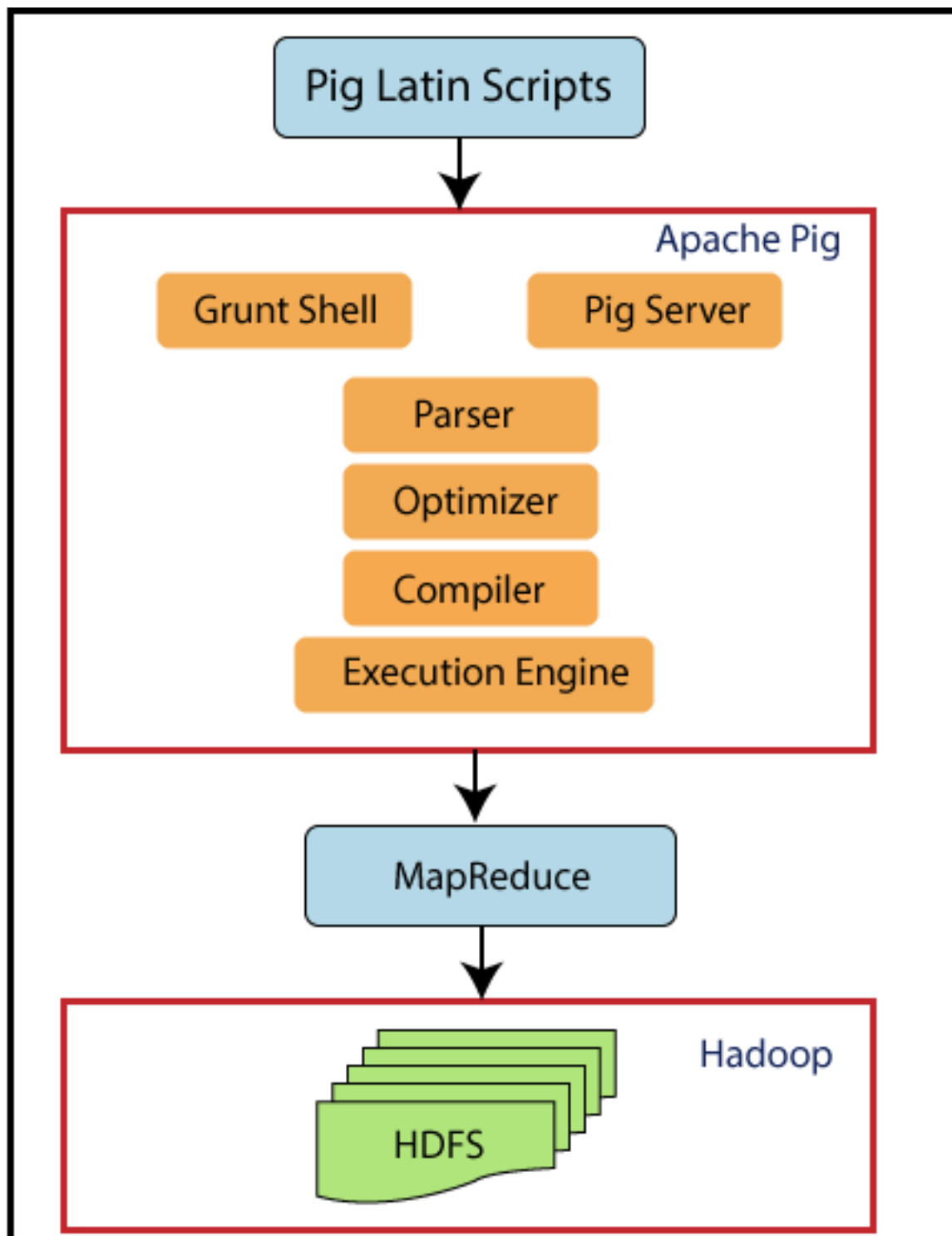
- For performing several operations Apache Pig provides rich sets of operators like the filters, join, sort, etc.
- Easy to learn, read and write. Especially for SQL-programmer, Apache Pig is a boon.
- Apache Pig is extensible so that you can make your own user-defined functions and process.
- Join operation is easy in Apache Pig.
- Fewer lines of code.
- Apache Pig allows splits in the pipeline.
- The data structure is multivalued, nested, and richer.
- Pig can handle the analysis of both structured and unstructured data.

Types of Data Models in Apache Pig: It consist of the 4 types of data models as follows:

- **Atom:** It is a atomic data value which is used to store as a string. The main use of this model is that it can be used as a number and as well as a string.
- **Tuple:** It is an ordered set of the fields.
- **Bag:** It is a collection of the tuples.
- **Map:** It is a set of key/value pairs.

Pig architecture

There are several components in the Apache Pig framework. Let's study these major components in detail:



i. Parser

At first, all the Pig Scripts are handled by the Parser. Parser basically checks the syntax of the script, does type checking, and other miscellaneous checks. Afterwards, Parser's output will be a DAG (directed acyclic graph) that represents the Pig Latin statements as well as logical operators.

The logical operators of the script are represented as the nodes and the data flows are represented as edges in DAG (the logical plan)

ii. Optimizer

Afterwards, the logical plan (DAG) is passed to the logical optimizer. It carries out the logical optimizations further such as projection and push down.

iii. Compiler

Then compiler compiles the optimized logical plan into a series of MapReduce jobs.

iv. Execution engine

Eventually, all the MapReduce jobs are submitted to Hadoop in a sorted order. Ultimately, it produces the desired results while these MapReduce jobs are executed on Hadoop.

Apache Pig Execution Modes:

Apache Pig has two execution modes:

- **Local Mode**

In 'Local Mode', the source data would be picked from the local directory in your computer system. The MapReduce mode can be specified using 'pig -x local' command.

- **MapReduce Mode:**

To run Pig in MapReduce mode, you need access to Hadoop cluster and HDFS installation. The MapReduce mode can be specified using the 'pig' command.

Apache Pig Operators

The Apache Pig Operators is a high-level procedural language for querying large data sets using Hadoop and the Map Reduce Platform. A Pig Latin statement is an operator that takes a relation as input and produces another relation as output. These operators are the main tools for Pig Latin provides to operate on the data. They allow you to transform it by sorting, grouping, joining, projecting, and filtering.

We have a huge set of Apache Pig Operators, for performing several types of Operations. Let's discuss types of Apache Pig Operators:

- Diagnostic Operators
- Grouping & Joining
- Combining & Splitting
- Filtering
- Sorting

i. Diagnostic Operators:

Basically, we use Diagnostic Operators to verify the execution of the Load statement. There are four different types of diagnostic operators –

- Dump operator
- Describe operator
- Explanation operator

ii. Grouping & Joining:

- Group Operator
- Cogroup Operator
- Join Operator
- Cross operator

iii. Combining & Splitting:

- Union
- Split

iv. Filtering:

- Filter
- Distinct
- For Each

Pig Functions

There is a huge set of Apache Pig Built in Functions available. Such as the eval, load/store, math, string, date and time, bag and tuple functions. Basically, there are two main properties which differentiate built in functions from user-defined functions (UDFs) such as:

- We do not need to register built in functions since Pig knows where they are.
- Also, we do not need to qualify built in functions, while using them, because again Pig knows where to find them.

Eval Functions

- AVG
- BagToString
- BagToTuple
- Bloom
- CONCAT
- COUNT
- COUNT_STAR
- DIFF
- IsEmpty
- MAX
- MIN
- PluckTuple
- SIZE
- SUBTRACT
- SUM
- IN
- TOKENIZE

Math Functions

- ABS
- ACOS
- ASIN
- ATAN
- CBRT
- CEIL
- COS
- COSH
- EXP
- FLOOR
- LOG

- LOG10
- RANDOM
- ROUND
- ROUND_TO
- SIN
- SINH
- SQRT
- TAN
- TANH

String Functions

- EqualsIgnoreCase
- ENDSWITH
- INDEXOF
- LAST_INDEX_OF
- LCFIRST
- LOWER
- LTRIM
- REGEX_EXTRACT
- REGEX_EXTRACT_ALL
- REGEX_SEARCH
- REPLACE
- RTRIM
- SPRINTF
- STARTSWITH
- STRSPLIT
- STRSPLITTOBAG
- SUBSTRING
- TRIM
- UCFIRST
- UPPER
- UniqueID

Datetime Functions

- AddDuration
- CurrentTime
- DaysBetween
- GetDay
- GetHour
- GetMilliSecond

- GetMinute
- GetMonth
- GetSecond
- GetWeek
- GetWeekYear
- GetYear
- HoursBetween
- MilliSecondsBetween
- MinutesBetween
- MonthsBetween
- SecondsBetween
- SubtractDuration
- ToDate
- ToMilliSeconds
- ToString
- ToUnixTime
- WeeksBetween
- YearsBetween

OBJECTIVES OF THE PROJECT

The objective of this dataset i.e. NETFLIX is to analyse the Big Data in apache pig. We have worked on HDFS cluster and also we have written different queries based on the respective dataset. Different types of functions are used in designing the queries. The purpose of analysing this dataset is to fetch our requirements in order to easily search any data which we need. Netflix is a very popular subscription-based streaming service that allows our members to watch TV shows and movies. Netflix content varies by region and may change over time. You can watch from a wide variety of award-winning Netflix Originals, TV shows, movies, documentaries, and more.

In our Dataset there are total 8807 rows and 12 columns. The columns are as follows:-

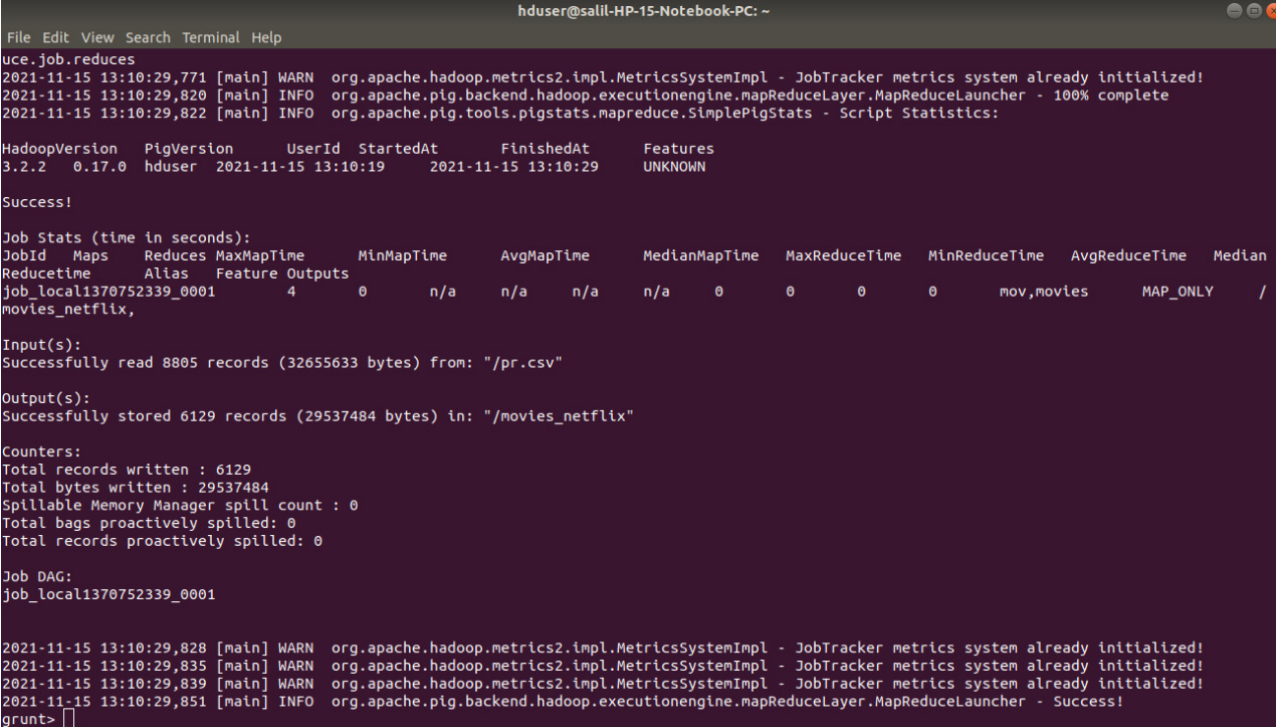
1. Id: This column contains the unique id assigned to each movie or tv show. For Example s1, s2 and so on.
2. Movie type: This column tells whether it is a movie or a tv show. For Example “Blood and Water” is a Tv show.
3. Title: This column represents the Title of the movie or tv show. For Example “Extraction”
4. Director: This column contains the name of the Director of the movie or tv show. For Example Christopher Nolan is the director of the movie “Inception”.
5. Cast: This column contains all the actors/actresses that acted in the specific movie or tv show. For example the cast of the movie titled “The Edge Of Seventeen” is Hailee Steinfeld, Woody Harrelson, Kyra Sedgwick, Haley Lu Richardson, Blake Jenner, Hayden Szeto, Alexander Calvert, Eric Keeney.
6. Country: This column represents the country of origin of the movie or tv show. For Example the country of origin of the movie titled “Taare Zameen Par” is India.
7. Dates: This column represents the date on which the movie or tv show was added in Netflix. For Example the tv show titled “Daughter From Another Mother” was added on netflix on 20-01-2021
8. Release Year: This column represents the year in which the movie or tv show was released. For Example the movie titled “Action Replay” was released in the year 2010.
9. Ratings: This column tells that which category of audience can watch the specific movie or tv show. For Example the rating TV-MA (TV Mature Audience) is specifically designed to be viewed by adults and therefore may be unsuitable for children under 17.
10. Duration: This column gives information about the running time of the movie or tv show. For Example the duration of the movie “Gunjan Saxena” is 113 minutes.

- 11.Listed_in: This column represents the category in which the movie or tv show is listed in. For Example the movie titled “Holidate” in listed_in the category of Comedies and Romantic Movies.
- 12.Description: This is the last column of our dataset which gives a brief description or an overview about the movie or tv show.

QUERIES

Query 1:- Load the given dataset ie Netflix in hdfs cluster.

```
mov= load '/pr.csv' using PigStorage(',') as
(id:chararray,mtype:chararray,title:chararray,director:chararray,cast:chararray,country:chararray,dates:chararray,
releasedate:int,rating:chararray,duration:chararray,listed_in:chararray,description:chararray);
```



```
hduser@salil-HP-15-Notebook-PC: ~
File Edit View Search Terminal Help
uce.job.reduces
2021-11-15 13:10:29,771 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,820 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2021-11-15 13:10:29,822 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion  PigVersion  UserId  StartedAt  FinishedAt  Features
3.2.2  0.17.0  hduser  2021-11-15 13:10:19  2021-11-15 13:10:29  UNKNOWN

Success!

Job Stats (time in seconds):
JobId  Maps  Reduces  MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  Median
ReduceTime  Alias  Feature  Outputs
job_local1370752339_0001  4  0  n/a  n/a  n/a  n/a  0  0  0  0  mov,movies  MAP_ONLY  /
movies_netflix,

Input(s):
Successfully read 8805 records (32655633 bytes) from: "/pr.csv"

Output(s):
Successfully stored 6129 records (29537484 bytes) in: "/movies_netflix"

Counters:
Total records written : 6129
Total bytes written : 29537484
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1370752339_0001

2021-11-15 13:10:29,828 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,839 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,851 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

Query 2:- Split the dataset into movies and Tv shows and store into hdfs cluster using pig function.

```
split mov into movies if mtype=='Movie',shows if mtype=='TV Show';
store movies into '/movies_netflix' using PigStorage(',');
store shows into '/shows_netflix' using PigStorage(',');
```

```
hduser@salil-HP-15-Notebook-PC: ~  
File Edit View Search Terminal Help  
uce.job.reduces  
2021-11-15 13:10:29,771 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!  
2021-11-15 13:10:29,820 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete  
2021-11-15 13:10:29,822 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:  
  
HadoopVersion PigVersion UserId StartedAt FinishedAt Features  
3.2.2 0.17.0 hduser 2021-11-15 13:10:19 2021-11-15 13:10:29 UNKNOWN  
  
Success!  
  
Job Stats (time in seconds):  
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime Median  
Reducetime Alias Feature Outputs  
job_local1370752339_0001 4 0 n/a n/a n/a n/a 0 0 0 0 mov,movies MAP_ONLY /  
movies_netflix,  
  
Input(s):  
Successfully read 8805 records (32655633 bytes) from: "/pr.csv"  
  
Output(s):  
Successfully stored 6129 records (29537484 bytes) in: "/movies_netflix"  
  
Counters:  
Total records written : 6129  
Total bytes written : 29537484  
Spillable Memory Manager spill count : 0  
Total bags proactively spilled: 0  
Total records proactively spilled: 0  
  
Job DAG:  
job_local1370752339_0001  
  
2021-11-15 13:10:29,828 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!  
2021-11-15 13:10:29,835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!  
2021-11-15 13:10:29,839 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!  
2021-11-15 13:10:29,851 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!  
grunt> █
```

Query 3:-Display the mtype and title whose show id is greater than 1000 and store in hdfs using pig function.

a= filter mov by id > 's1000';
m= foreach a generate mtype,title;
store m into '/query2' using PigStorage(',');

```
hduser@salil-HP-15-Notebook-PC: ~  
File Edit View Search Terminal Help  
uce.job.reduces  
2021-11-15 13:10:29,771 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!  
2021-11-15 13:10:29,820 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete  
2021-11-15 13:10:29,822 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:  
  
HadoopVersion PigVersion UserId StartedAt FinishedAt Features  
3.2.2 0.17.0 hduser 2021-11-15 13:10:19 2021-11-15 13:10:29 UNKNOWN  
  
Success!  
  
Job Stats (time in seconds):  
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime Median  
Reducetime Alias Feature Outputs  
job_local1370752339_0001 4 0 n/a n/a n/a n/a 0 0 0 0 mov,movies MAP_ONLY /  
movies_netflix,  
  
Input(s):  
Successfully read 8805 records (32655633 bytes) from: "/pr.csv"  
  
Output(s):  
Successfully stored 6129 records (29537484 bytes) in: "/movies_netflix"  
  
Counters:  
Total records written : 6129  
Total bytes written : 29537484  
Spillable Memory Manager spill count : 0  
Total bags proactively spilled: 0  
Total records proactively spilled: 0  
  
Job DAG:  
job_local1370752339_0001  
  
2021-11-15 13:10:29,828 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!  
2021-11-15 13:10:29,835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!  
2021-11-15 13:10:29,839 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!  
2021-11-15 13:10:29,851 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!  
grunt> █
```

Query 4:- Display title and director order by country and store in hdfs cluster using pig function.

q= order mov by country;

l= foreach q generate title,directer;

store l into 'query3' using PigStorage(',');

```
File Edit View Search Terminal Help
hduser@salil-HP-15-Notebook-PC: ~
uce.job.reduces
2021-11-15 13:10:29,771 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,820 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2021-11-15 13:10:29,822 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion  PigVersion  UserId  StartedAt  FinishedAt  Features
3.2.2  0.17.0  hduser  2021-11-15 13:10:19  2021-11-15 13:10:29  UNKNOWN

Success!

Job Stats (time in seconds):
JobId  Maps  Reduces  MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  Median
Reducetime  Alias  Feature  Outputs
job_local1370752339_0001  4  0  n/a  n/a  n/a  0  0  0  0  mov,movies  MAP_ONLY  /
movies_netflix,

Input(s):
Successfully read 8805 records (32655633 bytes) from: "/pr.csv"

Output(s):
Successfully stored 6129 records (29537484 bytes) in: "/movies_netflix"

Counters:
Total records written : 6129
Total bytes written : 29537484
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1370752339_0001

2021-11-15 13:10:29,828 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,839 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,851 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

Query 5:- Display the title of movies or Tv shows whose name ends with “l” and rank with release date in ascending order using dense function and limit is 500 in hdfs cluster using pig function.

p= rank mov by releasedate asc dense;

s= foreach p generate ENDSWITH(title,'l');

r= limit s 500;

store r into 'query4' using PigStorage(',');


```
hduser@salil-HP-15-Notebook-PC: ~
File Edit View Search Terminal Help
uce.job.reduces
2021-11-15 13:10:29,771 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,820 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2021-11-15 13:10:29,822 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion PigVersion UserId StartedAt FinishedAt Features
3.2.2 0.17.0 hduser 2021-11-15 13:10:19 2021-11-15 13:10:29 UNKNOWN

Success!

Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime Median
Reducetime Alias Feature Outputs
job_local1370752339_0001 4 0 n/a n/a n/a n/a 0 0 0 mov,movies MAP_ONLY /
movies_netflix,

Input(s):
Successfully read 8805 records (32655633 bytes) from: "/pr.csv"

Output(s):
Successfully stored 6129 records (29537484 bytes) in: "/movies_netflix"

Counters:
Total records written : 6129
Total bytes written : 29537484
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1370752339_0001

2021-11-15 13:10:29,828 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,839 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,851 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt> 
```

```
hduser@salil-HP-15-Notebook-PC: ~
File Edit View Search Terminal Help
uce.job.reduces
2021-11-15 13:10:29,771 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,820 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2021-11-15 13:10:29,822 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion PigVersion UserId StartedAt FinishedAt Features
3.2.2 0.17.0 hduser 2021-11-15 13:10:19 2021-11-15 13:10:29 UNKNOWN

Success!

Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime Median
Reducetime Alias Feature Outputs
job_local1370752339_0001 4 0 n/a n/a n/a n/a 0 0 0 mov,movies MAP_ONLY /
movies_netflix,

Input(s):
Successfully read 8805 records (32655633 bytes) from: "/pr.csv"

Output(s):
Successfully stored 6129 records (29537484 bytes) in: "/movies_netflix"

Counters:
Total records written : 6129
Total bytes written : 29537484
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1370752339_0001

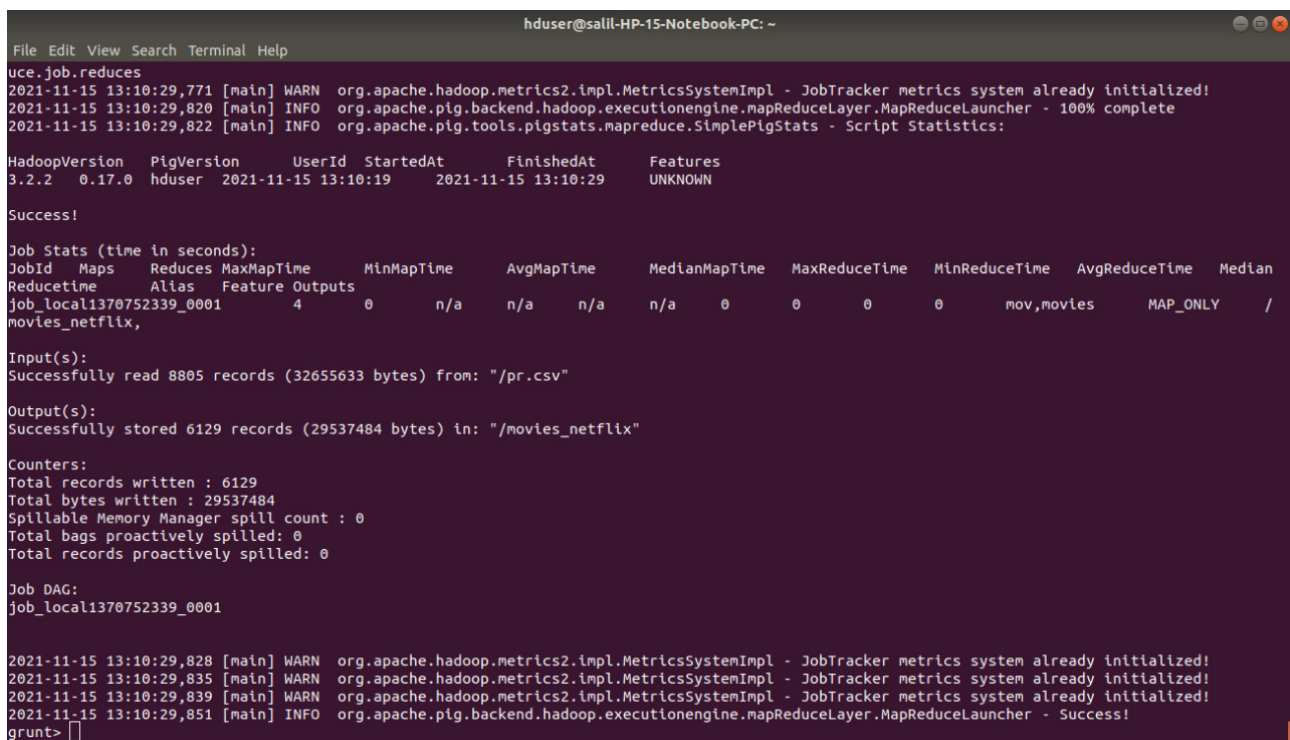
2021-11-15 13:10:29,828 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,839 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,851 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt> 
```

Query 6:- Display release date, ratings, duration and the category they are listed in whose mtype equals to movies and store in it hdfs cluster using pig function.

```
w= filter mov by mtype=='Movie';
e= foreach w generate dates,rating,duration,listed_in;
store e into '/query5' using PigStorage(',');
```

Query 7:- Display the title and show id of the movie which was released in the month of April and store it in the hdfs cluster using pig function.

```
d= foreach mov generate title,id,ToDate(dates,'dd-MM-yyyy') as dt;
t= filter d by GetMonth($2)==4;
store t into '/query6' using PigStorage(',');
```



```
hduser@salil-HP-15-Notebook-PC: ~
File Edit View Search Terminal Help
uce.job.reduces
2021-11-15 13:10:29,771 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,820 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2021-11-15 13:10:29,822 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion  PigVersion  UserId  StartedAt  FinishedAt  Features
3.2.2          0.17.0      hduser  2021-11-15 13:10:19  2021-11-15 13:10:29  UNKNOWN

Success!

Job Stats (time in seconds):
JobId  Maps  Reduces  MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  Median
ReduceTime  Alias  Feature  Outputs
job_local1370752339_0001  4  0  n/a  n/a  n/a  n/a  0  0  0  0  mov,movies  MAP_ONLY  /
movies_netflix,

Input(s):
Successfully read 8805 records (32655633 bytes) from: "/pr.csv"

Output(s):
Successfully stored 6129 records (29537484 bytes) in: "/movies_netflix"

Counters:
Total records written : 6129
Total bytes written : 29537484
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1370752339_0001

2021-11-15 13:10:29,828 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,839 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,851 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

Query 8:- Filter movies which are released in India and rating is TV-PG and sample by 10% of the dataset and store it into hdfs cluster using pig function.

```
x= filter movies by country=='India' and rating=='TV-PG';
z= sample y 0.1;
store z into '/query7' using PigStorage(',');
```

```
hduser@salil-HP-15-Notebook-PC: ~
File Edit View Search Terminal Help
uce.job.reduces
2021-11-15 13:10:29,771 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,820 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2021-11-15 13:10:29,822 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion PigVersion UserId StartedAt FinishedAt Features
3.2.2 0.17.0 hduser 2021-11-15 13:10:19 2021-11-15 13:10:29 UNKNOWN

Success!

Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime Median
Reducetime Alias Feature Outputs
job_local1370752339_0001 4 0 n/a n/a n/a n/a 0 0 0 0 mov,movies MAP_ONLY /
movies_netflix,

Input(s):
Successfully read 8805 records (32655633 bytes) from: "/pr.csv"

Output(s):
Successfully stored 6129 records (29537484 bytes) in: "/movies_netflix"

Counters:
Total records written : 6129
Total bytes written : 29537484
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1370752339_0001

2021-11-15 13:10:29,828 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,839 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,851 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

```
hduser@salil-HP-15-Notebook-PC: ~
File Edit View Search Terminal Help
uce.job.reduces
2021-11-15 13:10:29,771 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,820 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2021-11-15 13:10:29,822 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion PigVersion UserId StartedAt FinishedAt Features
3.2.2 0.17.0 hduser 2021-11-15 13:10:19 2021-11-15 13:10:29 UNKNOWN

Success!

Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime Median
Reducetime Alias Feature Outputs
job_local1370752339_0001 4 0 n/a n/a n/a n/a 0 0 0 0 mov,movies MAP_ONLY /
movies_netflix,

Input(s):
Successfully read 8805 records (32655633 bytes) from: "/pr.csv"

Output(s):
Successfully stored 6129 records (29537484 bytes) in: "/movies_netflix"

Counters:
Total records written : 6129
Total bytes written : 29537484
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1370752339_0001

2021-11-15 13:10:29,828 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,839 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,851 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```


Query 9:- Display title and movie/show which are listed in Horror and whose duration is 1 season.

ho= filter movies by listed_in matches '.*Horror.*' and duration matches '1 Season';

```

hduser@salil-HP-15-Notebook-PC: ~
File Edit View Search Terminal Help
uce.job.reduces
2021-11-15 13:10:29,771 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,820 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2021-11-15 13:10:29,822 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion  PigVersion  UserId  StartedAt  FinishedAt  Features
3.2.2  0.17.0  hduser  2021-11-15 13:10:19  2021-11-15 13:10:29  UNKNOWN

Success!

Job Stats (time in seconds):
JobId  Maps  Reduces  MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  Median
Reducetime  Alias  Feature  Outputs
job_local1370752339_0001  4  0  n/a  n/a  n/a  n/a  0  0  0  0  mov,movies  MAP_ONLY  /
movies_netflix,

Input(s):
Successfully read 8805 records (32655633 bytes) from: "/pr.csv"

Output(s):
Successfully stored 6129 records (29537484 bytes) in: "/movies_netflix"

Counters:
Total records written : 6129
Total bytes written : 29537484
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1370752339_0001

2021-11-15 13:10:29,828 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,839 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,851 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>

```

di= foreach ho generate title,mtype;

store di into 'query8' using PigStorage(',');

```

hduser@salil-HP-15-Notebook-PC: ~
File Edit View Search Terminal Help
uce.job.reduces
2021-11-15 13:10:29,771 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,820 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2021-11-15 13:10:29,822 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion  PigVersion  UserId  StartedAt  FinishedAt  Features
3.2.2  0.17.0  hduser  2021-11-15 13:10:19  2021-11-15 13:10:29  UNKNOWN

Success!

Job Stats (time in seconds):
JobId  Maps  Reduces  MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  Median
Reducetime  Alias  Feature  Outputs
job_local1370752339_0001  4  0  n/a  n/a  n/a  n/a  0  0  0  0  mov,movies  MAP_ONLY  /
movies_netflix,

Input(s):
Successfully read 8805 records (32655633 bytes) from: "/pr.csv"

Output(s):
Successfully stored 6129 records (29537484 bytes) in: "/movies_netflix"

Counters:
Total records written : 6129
Total bytes written : 29537484
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

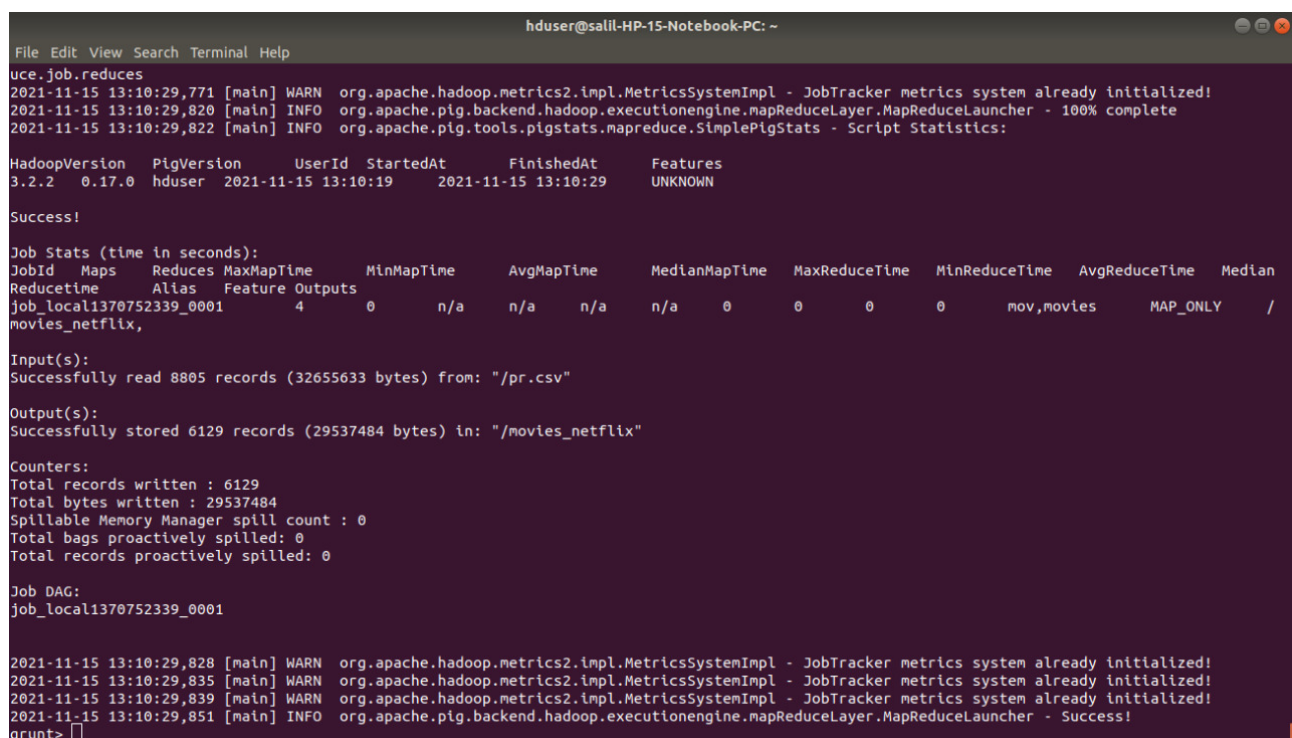
Job DAG:
job_local1370752339_0001

2021-11-15 13:10:29,828 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,839 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,851 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>

```

Query 10:- Display show id whose release date is less than 2021 and store it in hdfs cluster using pig function.

```
mx= filter mov by releasedate<2021;  
aa= foreach mx generate id;  
store aa into '/query9' using PigStorage(',');
```



```
File Edit View Search Terminal Help
hduuser@salil-HP-15-Notebook-PC: ~
uce.job.reduces
2021-11-15 13:10:29,771 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,820 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2021-11-15 13:10:29,822 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion  PigVersion  UserId  StartedAt  FinishedAt  Features
3.2.2.2  0.17.0  hduuser  2021-11-15 13:10:19  2021-11-15 13:10:29  UNKNOWN

Success!

Job Stats (time in seconds):
JobId  Maps  Reduces  MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  Median
job_local1370752339_0001  4  0  n/a  n/a  n/a  n/a  0  0  0  0  mov,movies  MAP_ONLY  /
movies_netflix,

Input(s):
Successfully read 8805 records (32655633 bytes) from: "/pr.csv"

Output(s):
Successfully stored 6129 records (29537484 bytes) in: "/movies_netflix"

Counters:
Total records written : 6129
Total bytes written : 29537484
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1370752339_0001

2021-11-15 13:10:29,828 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,839 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,851 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

Query 11:- Display id, title and description of all the movies directed by Rajkumar Hirani.

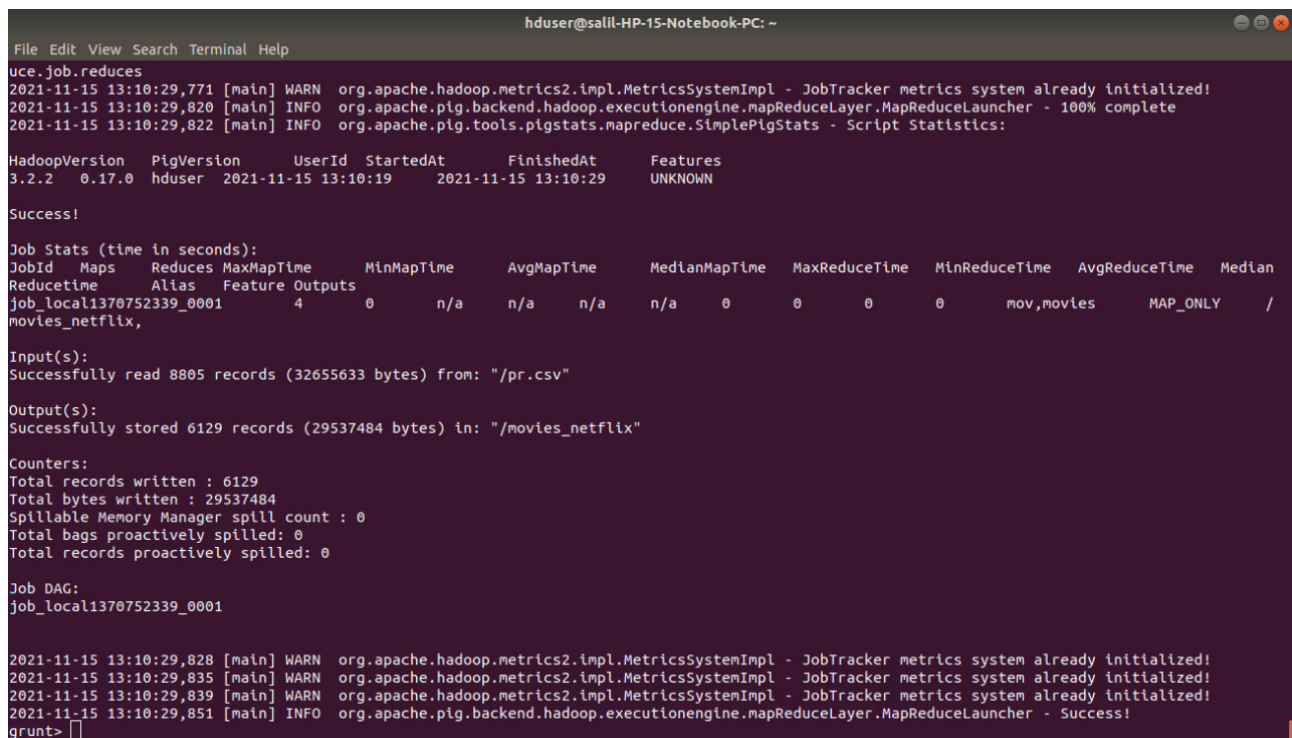
```
rj= filter mov by director matches 'Rajkumar Hirani';  
rjm= foreach rj generate id,title,description;  
store rjm into '/query10' using PigStorage(',');
```

Query 12:- Store the movies or tv shows which comes under the category of Children and family movies and group by release date and store it in hdfs cluster using pig function.

```
st= filter mov by listed_in matches 'Children & Family Movies' and rating == 'TV-Y';  
ri= group st by releasedate;  
store ri into '/query11' using PigStorage(',');
```

Query 13:- Display the title and show id of the movies or tv shows whose release year 2021 and store it into hdfs cluster using pig function.

```
daa= foreach mov generate title,id,ToDate(dates,'dd-MM-yyyy') as dt;  
t= filter daa by GetYear($2)==2021;  
store t into '/query12' using PigStorage(',');
```



The screenshot shows a terminal window titled 'hduser@salil-HP-15-Notebook-PC: ~'. It displays the output of a Pig script execution. The output includes log messages from Hadoop and Pig, a summary of job statistics, and a table of job details. The job is named 'job_local1370752339_0001' and is in a 'SUCCEEDED' state. The output file is '/movies_netflix'.

```
File Edit View Search Terminal Help  
uce.job.reduces  
2021-11-15 13:10:29,771 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!  
2021-11-15 13:10:29,820 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete  
2021-11-15 13:10:29,822 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:  
  
HadoopVersion PigVersion UserId StartedAt FinishedAt Features  
3.2.2 0.17.0 hduser 2021-11-15 13:10:19 2021-11-15 13:10:29 UNKNOWN  
  
Success!  
  
Job Stats (time in seconds):  
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime Median  
ReduceTime Alias Feature Outputs  
job_local1370752339_0001 4 0 n/a n/a n/a n/a 0 0 0 0 mov,movies MAP_ONLY /  
movies_netflix,  
  
Input(s):  
Successfully read 8805 records (32655633 bytes) from: "/pr.csv"  
  
Output(s):  
Successfully stored 6129 records (29537484 bytes) in: "/movies_netflix"  
  
Counters:  
Total records written : 6129  
Total bytes written : 29537484  
Spillable Memory Manager spill count : 0  
Total bags proactively spilled: 0  
Total records proactively spilled: 0  
  
Job DAG:  
job_local1370752339_0001  
  
2021-11-15 13:10:29,828 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!  
2021-11-15 13:10:29,835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!  
2021-11-15 13:10:29,839 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!  
2021-11-15 13:10:29,851 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!  
grunt>
```

```

hduser@salil-HP-15-Notebook-PC: ~
File Edit View Search Terminal Help
uce.job.reduce
2021-11-15 13:10:29,771 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,820 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2021-11-15 13:10:29,822 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion  PigVersion  UserId  StartedAt  FinishedAt  Features
3.2.2  0.17.0  hduser  2021-11-15 13:10:19  2021-11-15 13:10:29  UNKNOWN

Success!

Job Stats (time in seconds):
JobId  Maps  Reduces  MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  Median
ReduceTime  Alias  Feature  Outputs
job_local1370752339_0001  4  0  n/a  n/a  n/a  n/a  0  0  0  0  mov,movies  MAP_ONLY  /
movies_netflix,

Input(s):
Successfully read 8805 records (32655633 bytes) from: "/pr.csv"

Output(s):
Successfully stored 6129 records (29537484 bytes) in: "/movies_netflix"

Counters:
Total records written : 6129
Total bytes written : 29537484
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1370752339_0001

2021-11-15 13:10:29,828 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,839 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,851 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>

```

```

hduser@salil-HP-15-Notebook-PC: ~
File Edit View Search Terminal Help
uce.job.reduce
2021-11-15 13:10:29,771 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,820 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2021-11-15 13:10:29,822 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion  PigVersion  UserId  StartedAt  FinishedAt  Features
3.2.2  0.17.0  hduser  2021-11-15 13:10:19  2021-11-15 13:10:29  UNKNOWN

Success!

Job Stats (time in seconds):
JobId  Maps  Reduces  MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  Median
ReduceTime  Alias  Feature  Outputs
job_local1370752339_0001  4  0  n/a  n/a  n/a  n/a  0  0  0  0  mov,movies  MAP_ONLY  /
movies_netflix,

Input(s):
Successfully read 8805 records (32655633 bytes) from: "/pr.csv"

Output(s):
Successfully stored 6129 records (29537484 bytes) in: "/movies_netflix"

Counters:
Total records written : 6129
Total bytes written : 29537484
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1370752339_0001

2021-11-15 13:10:29,828 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,839 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,851 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>

```

Query 14:- Load the dataset which contains movies only.

movi= load '/movies_netflix' using PigStorage(',') as

(id:chararray,mtype:chararray,title:chararray,director:chararray,cast:chararray,country:chararray,dates:chararray,releasedate:int,rating:chararray,duration:chararray,listed_in:chararray,description:chararray);


```
hduser@salil-HP-15-Notebook-PC: ~
File Edit View Search Terminal Help
uce.job.reduces
2021-11-15 13:10:29,771 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,820 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2021-11-15 13:10:29,822 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion  PigVersion  UserId  StartedAt  FinishedAt  Features
3.2.2  0.17.0  hduser  2021-11-15 13:10:19  2021-11-15 13:10:29  UNKNOWN

Success!

Job Stats (time in seconds):
JobId  Maps  Reduces  MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  Median
ReduceTime  Alias  Feature  Outputs
job_local1370752339_0001  4  0  n/a  n/a  n/a  n/a  0  0  0  0  mov,movies  MAP_ONLY  /
movies_netflix,

Input(s):
Successfully read 8805 records (32655633 bytes) from: "/pr.csv"

Output(s):
Successfully stored 6129 records (29537484 bytes) in: "/movies_netflix"

Counters:
Total records written : 6129
Total bytes written : 29537484
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1370752339_0001

2021-11-15 13:10:29,828 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,839 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,851 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt> 
```

Query 15:- Generate the number of movies into the dataset and store it into hdfs cluster using pig function.

```
a= group movi all;
c= foreach a generate group ,COUNT(movi);
store c into '/query13' using PigStorage(',');
```

Query 16:-Load the dataset where mtype matches tv show.

```
showing= load '/shows_netflix' using PigStorage(',') as
```

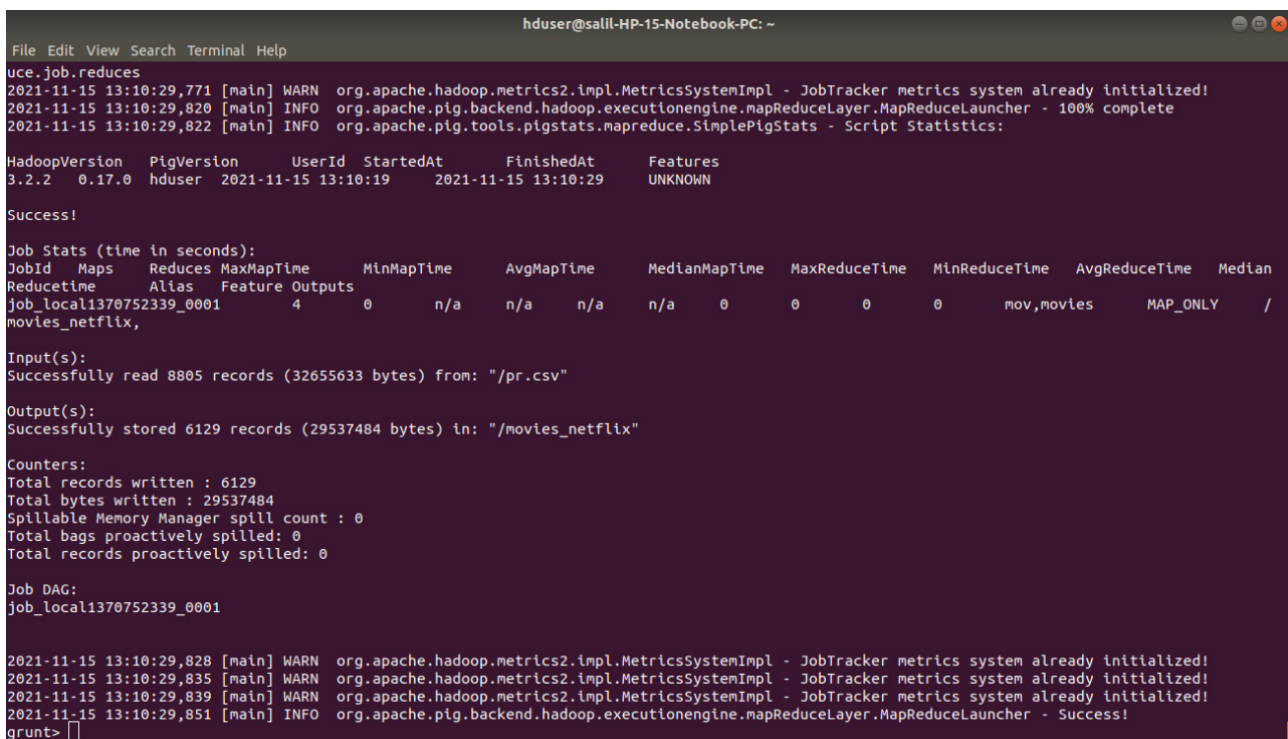
```
(id:chararray,mtype:chararray,title:chararray,director:chararray,cast:chararray,country:chararray,dates:chararray,release date:int,rating:chararray,duration:chararray,listed_in:chararray,description:chararray);
```


Query 17:- Count the number of tv shows in the dataset and store it into hdfs cluster using pig function.

```
g= group showing all;
tot= foreach g generate group,COUNT(showing);
store tot into '/query14' using PigStorage(',');
```

Query 18:- Group dataset by the date and store it into hdfs cluster using pig function.

```
mm= group mov by dates;
store mm into '/query15' using PigStorage(',');
```



```
hduser@salil-HP-15-Notebook-PC: ~
File Edit View Search Terminal Help
uce.job.reduces
2021-11-15 13:10:29,771 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,820 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2021-11-15 13:10:29,822 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion  PigVersion  UserId  StartedAt  FinishedAt  Features
3.2.2  0.17.0  hduser  2021-11-15 13:10:19  2021-11-15 13:10:29  UNKNOWN

Success!

Job Stats (time in seconds):
JobId  Maps  Reduces  MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  Median
ReduceTime  Alias  Feature  Outputs
job_local1370752339_0001  4  0  n/a  n/a  n/a  n/a  0  0  0  mov,movies  MAP_ONLY  /
movies_netflix,

Input(s):
Successfully read 8805 records (32655633 bytes) from: "/pr.csv"

Output(s):
Successfully stored 6129 records (29537484 bytes) in: "/movies_netflix"

Counters:
Total records written : 6129
Total bytes written : 29537484
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1370752339_0001

2021-11-15 13:10:29,828 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,839 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,851 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

Query 19:- Generate title and release date where mtype is Tv show and duration is 4 seasons and store it into hdfs cluster using pig function.

```
xyz= filter mov by duration =='4 Seasons' and mtype =='TV Show';
```

```
cee= foreach xyz generate title,releasedate;
```

store cee into '/query16' using PigStorage(',');

Query 20:- Display id and title of the movie/show which is released in the fourth week of the year and store it into hdfs cluster using pig function.

fo=foreach mov generate title,id,ToDate(dates,'dd-MM-yyyy') as dt;

i= filter fo by GetWeek(\$2)==4;

```
hdunder@salil-HP-15-Notebook-PC: ~
File Edit View Search Terminal Help
uce.job.reduces
2021-11-15 13:10:29,771 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,820 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2021-11-15 13:10:29,822 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion  PigVersion  UserId  StartedAt  FinishedAt  Features
3.2.2  0.17.0  hdunder  2021-11-15 13:10:19  2021-11-15 13:10:29  UNKNOWN

Success!

Job Stats (time in seconds):
JobId  Maps  Reduces  MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  Median
Reducetime  Alias  Feature  Outputs
job_local1370752339_0001  4  0  n/a  n/a  n/a  n/a  0  0  0  0  mov,movies  MAP_ONLY  /
movies_netflix,

Input(s):
Successfully read 8805 records (32655633 bytes) from: "/pr.csv"

Output(s):
Successfully stored 6129 records (29537484 bytes) in: "/movies_netflix"

Counters:
Total records written : 6129
Total bytes written : 29537484
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1370752339_0001

2021-11-15 13:10:29,828 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,839 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,851 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

store i into '/query17' using PigStorage(',');

```
hdunder@salil-HP-15-Notebook-PC: ~
File Edit View Search Terminal Help
uce.job.reduces
2021-11-15 13:10:29,771 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,820 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2021-11-15 13:10:29,822 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion  PigVersion  UserId  StartedAt  FinishedAt  Features
3.2.2  0.17.0  hdunder  2021-11-15 13:10:19  2021-11-15 13:10:29  UNKNOWN

Success!

Job Stats (time in seconds):
JobId  Maps  Reduces  MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  Median
Reducetime  Alias  Feature  Outputs
job_local1370752339_0001  4  0  n/a  n/a  n/a  n/a  0  0  0  0  mov,movies  MAP_ONLY  /
movies_netflix,

Input(s):
Successfully read 8805 records (32655633 bytes) from: "/pr.csv"

Output(s):
Successfully stored 6129 records (29537484 bytes) in: "/movies_netflix"

Counters:
Total records written : 6129
Total bytes written : 29537484
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1370752339_0001
```

Query 21:- Display top 3 movies casted by Salman and store it into hdfs cluster using pig function.

y= filter mov by cast matches 'Salman.*' and mtype=='Movie';

li= limit y 3;

store li into '/query18' using PigStorage(',');

```
hduser@salil-HP-15-Notebook-PC: ~  
File Edit View Search Terminal Help  
uce.job.reduces  
2021-11-15 13:10:29,771 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!  
2021-11-15 13:10:29,820 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete  
2021-11-15 13:10:29,822 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:  
  
HadoopVersion PigVersion UserId StartedAt FinishedAt Features  
3.2.2 0.17.0 hduser 2021-11-15 13:10:19 2021-11-15 13:10:29 UNKNOWN  
  
Success!  
  
Job Stats (time in seconds):  
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime Median  
Reducetime Alias Feature Outputs  
job_local1370752339_0001 4 0 n/a n/a n/a n/a 0 0 0 0 mov,movies MAP_ONLY /  
movies_netflix,  
  
Input(s):  
Successfully read 8805 records (32655633 bytes) from: "/pr.csv"  
  
Output(s):  
Successfully stored 6129 records (29537484 bytes) in: "/movies_netflix"  
  
Counters:  
Total records written : 6129  
Total bytes written : 29537484  
Spillable Memory Manager spill count : 0  
Total bags proactively spilled: 0  
Total records proactively spilled: 0  
  
Job DAG:  
job_local1370752339_0001  
  
2021-11-15 13:10:29,828 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!  
2021-11-15 13:10:29,835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!  
2021-11-15 13:10:29,839 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!  
2021-11-15 13:10:29,851 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!  
grunt> █
```

```
hduser@salil-HP-15-Notebook-PC: ~  
File Edit View Search Terminal Help  
uce.job.reduces  
2021-11-15 13:10:29,771 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!  
2021-11-15 13:10:29,820 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete  
2021-11-15 13:10:29,822 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:  
  
HadoopVersion PigVersion UserId StartedAt FinishedAt Features  
3.2.2 0.17.0 hduser 2021-11-15 13:10:19 2021-11-15 13:10:29 UNKNOWN  
  
Success!  
  
Job Stats (time in seconds):  
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime Median  
Reducetime Alias Feature Outputs  
job_local1370752339_0001 4 0 n/a n/a n/a n/a 0 0 0 0 mov,movies MAP_ONLY /  
movies_netflix,  
  
Input(s):  
Successfully read 8805 records (32655633 bytes) from: "/pr.csv"  
  
Output(s):  
Successfully stored 6129 records (29537484 bytes) in: "/movies_netflix"  
  
Counters:  
Total records written : 6129  
Total bytes written : 29537484  
Spillable Memory Manager spill count : 0  
Total bags proactively spilled: 0  
Total records proactively spilled: 0  
  
Job DAG:  
job_local1370752339_0001  
  
2021-11-15 13:10:29,828 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!  
2021-11-15 13:10:29,835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!  
2021-11-15 13:10:29,839 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!  
2021-11-15 13:10:29,851 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!  
grunt> █
```

```
hduser@salil-HP-15-Notebook-PC: ~
File Edit View Search Terminal Help
uce.job.reduces
2021-11-15 13:10:29,771 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,820 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2021-11-15 13:10:29,822 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion  PigVersion  UserId  StartedAt  FinishedAt  Features
3.2.2  0.17.0  hduser  2021-11-15 13:10:19  2021-11-15 13:10:29  UNKNOWN

Success!

Job Stats (time in seconds):
JobId  Maps  Reduces  MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  Median
ReduceTime  Alias  Feature  Outputs
job_local1370752339_0001  4  0  n/a  n/a  n/a  0  0  0  0  mov,movies  MAP_ONLY  /
movies_netflix,

Input(s):
Successfully read 8805 records (32655633 bytes) from: "/pr.csv"

Output(s):
Successfully stored 6129 records (29537484 bytes) in: "/movies_netflix"

Counters:
Total records written : 6129
Total bytes written : 29537484
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1370752339_0001

2021-11-15 13:10:29,828 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,839 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,851 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt> □
```

Query 22:- Display movies/shows whose name starts with the letter “A” , release date is between 2018-2020 and rating is PG-13 and store it into hdfs cluster using pig function.

redate= filter mov by STARTSWITH(title,'A') and rating matches 'PG-13' and (releasedate<2020 or releasedate>2018);

res= foreach redate generate id,title,mtype;
store res into '/query19' using PigStorage(',');

Query 23:- Count the number of items in the dataset and and store it into hdfs cluster using pig function.

fu= group mov by releasedate ;
gu= foreach fu generate group,COUNT(mov);
store gu into '/query20' using PigStorage(',');

Query 24:- Display the most occurred letter in the dataset and store it into hdfs cluster using pig function.

```
poem= load '/pr.csv' as line:chararray;  
token= foreach poem generate TOKENIZE(line) as line;  
flat= foreach token generate flatten(line) as line;  
substr= foreach flat generate SUBSTRING(line,0,1) as letter;  
gro= group substr by letter;  
con= foreach gro generate group,COUNT(substr);  
result= order con by $1 desc;  
store result into '/query21' using PigStorage(',');
```

```
File Edit View Search Terminal Help
hduser@salil-HP-15-Notebook-PC: ~
uce.job.reduces
2021-11-15 13:10:29,771 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,820 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2021-11-15 13:10:29,822 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion  PigVersion  UserId  StartedAt      FinishedAt      Features
3.2.2          0.17.0      hduser  2021-11-15 13:10:19  2021-11-15 13:10:29  UNKNOWN

Success!

Job Stats (time in seconds):
JobId  Maps  Reduces  MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  Median
Reducetime  Alias  Feature  Outputs
job_local1370752339_0001  4      0      n/a      n/a      n/a      n/a      0      0      0      0      mov,movies  MAP_ONLY  /
movies_netflix,

Input(s):
Successfully read 8805 records (32655633 bytes) from: "/pr.csv"

Output(s):
Successfully stored 6129 records (29537484 bytes) in: "/movies_netflix"

Counters:
Total records written : 6129
Total bytes written : 29537484
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1370752339_0001

2021-11-15 13:10:29,828 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,839 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-15 13:10:29,851 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

CASE STUDY 1

a) Calculate the wordcount of file shakespeare.txt

a= load 'shakespeare.txt' using PigStorage as lines;

b= foreach a generate FLATTEN(TOKENIZE(lines)) as word;

c= group b by word;

d= foreach c generate group,COUNT(b);

dump d;

```
Nov 15 17:26
hduser@sourav-VirtualBox: /home/sourav

poisonous-tongu'd      1
precious-princely     1
something-settled      1
strangely-visited      1
sweet-complaining      1
sword-and-buckler      1
thought-executing      1
thrice-victorious      1
toasts-and-butter      1
undistinguishable      2
well-accomplish'd      2
well-proportion'd      1
world-without-end      1
Jack-out-of-office     1
deep-contemplative     1
foolish-compounded     1
freestone-colour'd     1
heavenly-harness'd     1
intolerable-fright     1
lord-protectorship     1
marriage-pleasures     1
ne'er-lust-wearied     1
pepper-gingerbread     1
serving-creature's     1
always-wind-obeying    1
historical-pastoral    1
senseless-obstinate    1
shameless-desperate    1
ten-times-barr'd-up    1
tennis-court-keeper    1
tragical-historical    1
waiting-gentlewoman    4
death-counterfeiting   1
honourable-dangerous   1
one-trunk-inheriting   1
three-farthing-worth   1
thrice-puissantliege   1
wholesome-profitable   1
mustachio-purple-hued  1
to-and-fro-conflicting 1
honorificabilitudinitatibus 1
six-or-seven-times-honoured 1
tragical-comical-historical-pastoral 1

grunt>
```

b) most occurred starting letter.

a= load 'shakespeare.txt' using PigStorage as lines;

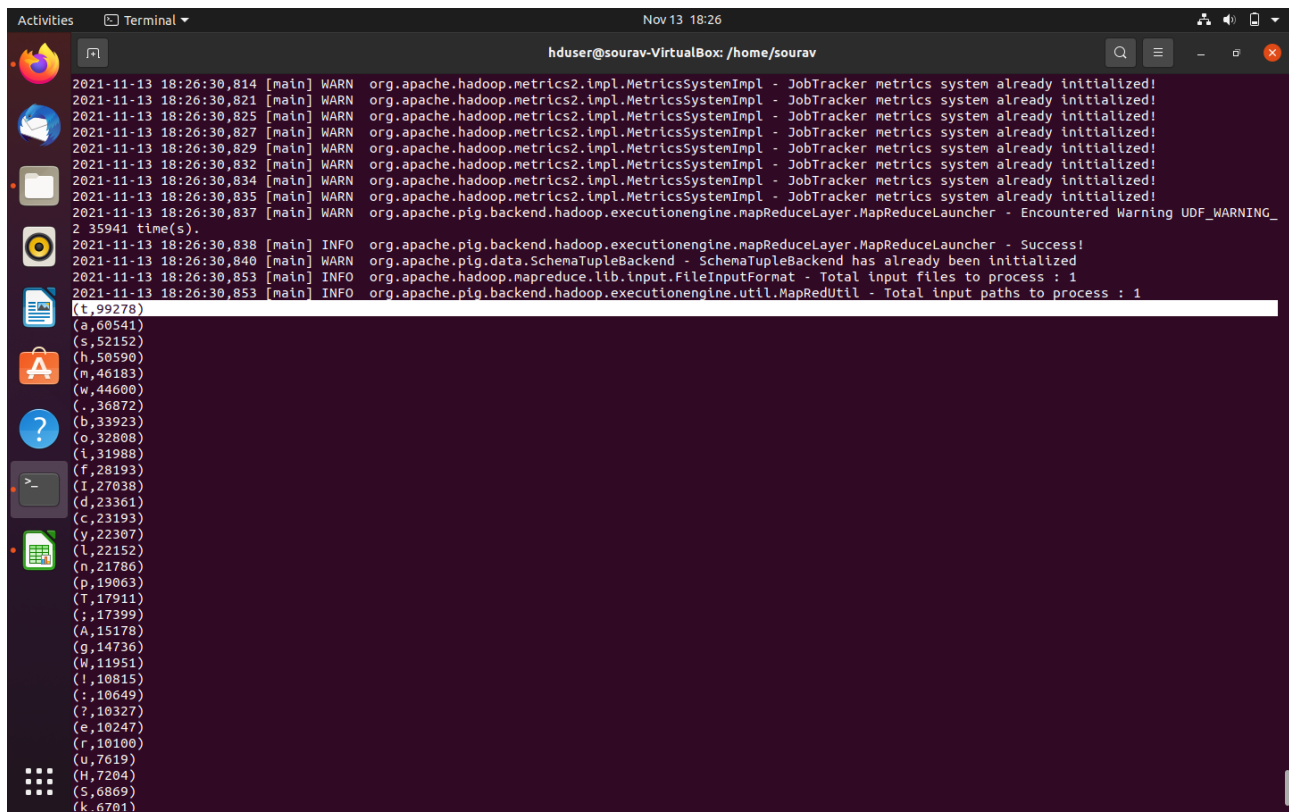
b= foreach a generate FLATTEN(TOKENIZE(lines)) as word;

c= foreach b generate SUBSTRING(word,0,1) as letter;

d= group c by letter;

e= foreach d generate group,COUNT(c);

dump d;



```
Activities Terminal Nov 13 18:26
hduser@sourav-VirtualBox: /home/sourav

2021-11-13 18:26:30,814 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-13 18:26:30,821 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-13 18:26:30,825 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-13 18:26:30,827 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-13 18:26:30,829 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-13 18:26:30,832 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-13 18:26:30,834 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-13 18:26:30,835 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-11-13 18:26:30,837 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning UDF_WARNING_
2 35941 tTime(s).
2021-11-13 18:26:30,838 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2021-11-13 18:26:30,840 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2021-11-13 18:26:30,853 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2021-11-13 18:26:30,853 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1

(t, 99278)
(a, 60541)
(s, 52152)
(h, 50590)
(m, 46183)
(w, 44600)
(., 36872)
(b, 33923)
(o, 32808)
(l, 31988)
(f, 28193)
(I, 27038)
(d, 23361)
(c, 23193)
(y, 22307)
(l, 22152)
(n, 21786)
(p, 19063)
(T, 17911)
(., 17399)
(A, 15178)
(g, 14736)
(W, 11951)
(l, 10815)
(., 10649)
(? , 10327)
(e, 10247)
(r, 10100)
(u, 7619)
(H, 7204)
(S, 6869)
(k, 6701)
```


CASE STUDY 2

Removal of stopwords and punctuation from a large text file. Calculate the occurrence of each word from the file afterwards. (related files shared with you already)

```
shakespeare= load '/shakespeare.txt' as (lineoftext:chararray);
```

```
stopwords = load '/stop-word-list.csv' using PigStorage() as (stopword:chararray);
```

```
words= foreach shakespeare generate
```

```
FLATTEN(TOKENIZE(REPLACE(LOWER(TRIM(lineoftext)),['\p{Punct},\p{Cntrl}'],''))) as word;
```

```
realwords = FILTER words by SIZE(word) > 0;
```

```
flattened_stopwords = foreach stopwords generate FLATTEN(TOKENIZE(stopword)) as  
stopwords;
```

```
right_joined = join flattened_stopwords by stopwords RIGHT OUTER,realwords by word;
```

```
meaningful_words = filter right_joined by (flattened_stopwords::stopwords IS NULL);
```

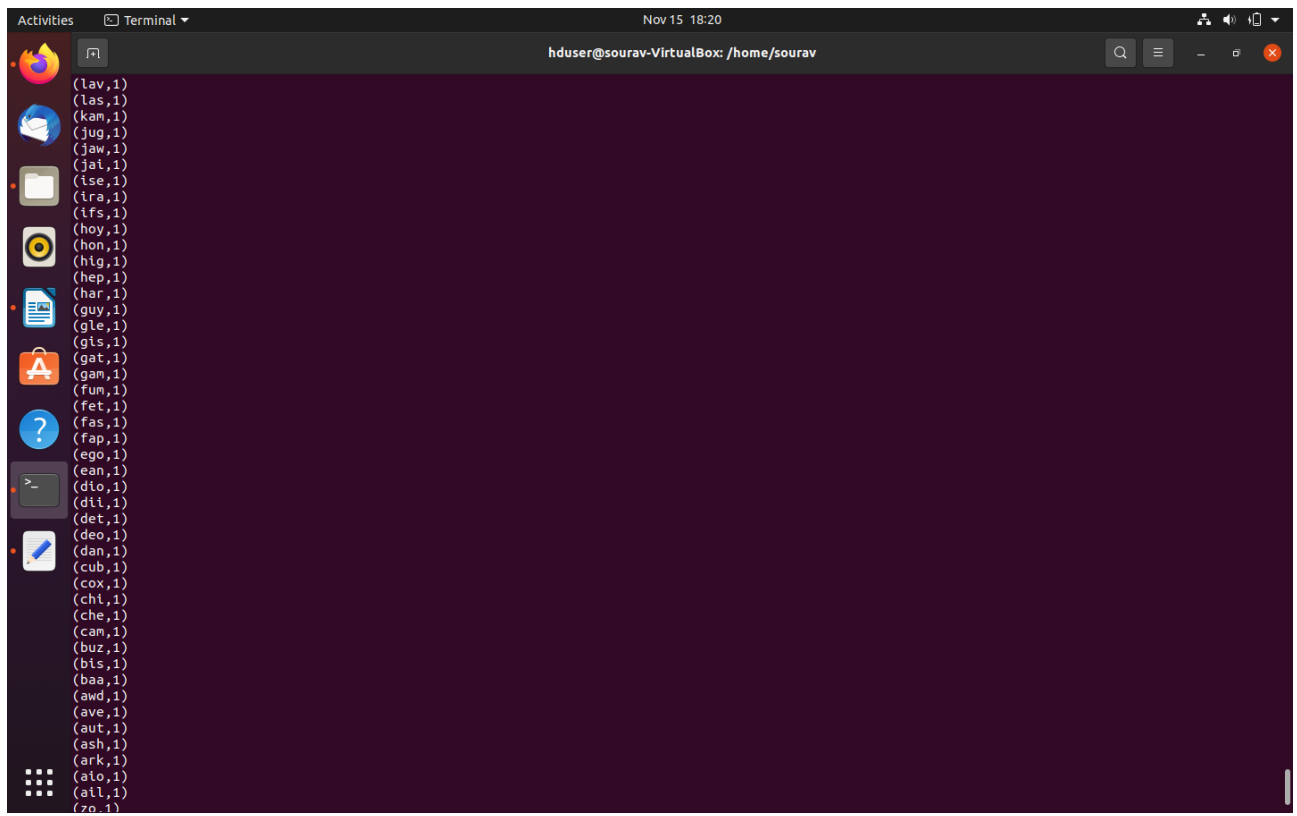
```
shakespeare_real_words = foreach meaningful_words generate realwords::word as word;
```

```
grouped = GROUP shakespeare_real_words by word;
```

```
counted = foreach grouped generate group as word, COUNT(shakespeare_real_words) as  
wordcount;
```

```
ordered = ORDER counted by wordcount desc
```

```
dump ordered;
```



References

- <https://www.omnisci.com/learn/big-data-analytics>
- <https://data-flair.training/blogs/pig-architecture/>
- <https://www.edureka.co/blog/apache-hadoop-hdfs-architecture/>
- <https://techvidvan.com/tutorials/apache-pig-operators/>
- <https://pig.apache.org/docs/r0.17.0/func.html>