**MOVIE ANALYSIS**

**Objective**

* Use Sqoop to collect Movies data and to store the data to HDFS. Process various insights while analysing the data using Hive.

# **Making a directory in HDFS**

hdfs dfs -mkdir '/moviedata/'

# **Getting started with hive**

Hive

# Creating a DB

create database movie;

# Creating a table

create table movies\_dataset(id int, name string, year int, rating float, duration int) row format delimited fields terminated by ',' stored as textfile location '/moviedata/';

OK

Time taken: 1.023 seconds

# Load data into the table

# Queries

1)select \* from movies\_dataset;

OK

1 The Nightmare Before Christmas 1993 3.9 4568

2 The Mummy 1932 3.5 4388

3 Orphans of the Storm 1921 3.2 9062

4 The Object of Beauty 1991 2.8 6150

5 Night Tide 1963 2.8 5126

6 One Magic Christmas 1985 3.8 5333

7 Muriel's Wedding 1994 3.5 6323

8 Mother's Boys 1994 3.4 5733

9 Nosferatu: Original Version 1929 3.5 5651

10 Nick of Time 1995 3.4 5333 ……

Time taken: 0.193 seconds, Fetched: 49590 row(s)

2)select \* from movies\_dataset limit 10;

OK

1 The Nightmare Before Christmas 1993 3.9 4568

2 The Mummy 1932 3.5 4388

3 Orphans of the Storm 1921 3.2 9062

4 The Object of Beauty 1991 2.8 6150

5 Night Tide 1963 2.8 5126

6 One Magic Christmas 1985 3.8 5333

7 Muriel's Wedding 1994 3.5 6323

8 Mother's Boys 1994 3.4 5733

9 Nosferatu: Original Version 1929 3.5 5651

10 Nick of Time 1995 3.4 5333

Time taken: 1.102 seconds, Fetched: 10 row(s)

3) select count(\*) from movies\_dataset;

Query ID = root\_20210816181439\_d296d444-4db0-4a23-9ac3-b3b28acdadfc

Total jobs = 1

Launching Job 1 out of 1

Tez session was closed. Reopening...

Session re-established.

Status: Running (Executing on YARN cluster with App id application\_1629122499604\_0004)

--------------------------------------------------------------------------------

**VERTICES STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED**

--------------------------------------------------------------------------------

Map 1 .......... SUCCEEDED 1 1 0 0 0 0

Reducer 2 ...... SUCCEEDED 1 1 0 0 0 0

--------------------------------------------------------------------------------

**VERTICES: 02/02 [==========================>>] 100% ELAPSED TIME: 5.65 s**

--------------------------------------------------------------------------------

OK

49590

Time taken: 12.48 seconds, Fetched: 1 row(s)

4) select \* from movies\_dataset where name='Shinobi Girl';

OK

49557 Shinobi Girl 2012 2.0 NULL

Time taken: 0.192 seconds, Fetched: 1 row(s)

5) SELECT count(\*) FROM movies\_dataset WHERE year BETWEEN 1950 AND 1960;

Query ID = root\_20210816190246\_8d718739-c6e6-4097-b6f4-a6fbdc6243a5

Total jobs = 1

Launching Job 1 out of 1

Tez session was closed. Reopening...

Session re-established.

Status: Running (Executing on YARN cluster with App id application\_1629122499604\_0005)

--------------------------------------------------------------------------------

**VERTICES STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED**

--------------------------------------------------------------------------------

Map 1 .......... SUCCEEDED 1 1 0 0 0 0

Reducer 2 ...... SUCCEEDED 1 1 0 0 0 0

--------------------------------------------------------------------------------

**VERTICES: 02/02 [==========================>>] 100% ELAPSED TIME: 6.45 s**

--------------------------------------------------------------------------------

OK

547

Time taken: 12.553 seconds, Fetched: 1 row(s)

6) SELECT count(\*) FROM movies\_dataset WHERE rating > 4.0;

Query ID = root\_20210816190717\_a4ca65b3-b657-4436-89bd-9cbc0ad098a3

Total jobs = 1

Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application\_1629122499604\_0005)

--------------------------------------------------------------------------------

**VERTICES STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED**

--------------------------------------------------------------------------------

Map 1 .......... SUCCEEDED 1 1 0 0 0 0

Reducer 2 ...... SUCCEEDED 1 1 0 0 0 0

--------------------------------------------------------------------------------

**VERTICES: 02/02 [==========================>>] 100% ELAPSED TIME: 5.23 s**

--------------------------------------------------------------------------------

OK

898

Time taken: 5.993 seconds, Fetched: 1 row(s)

7) SELECT name FROM movies\_dataset where rating > 3.0 and rating < 4.0;

ok

he Square (Trailer)

My Hope America with Billy Graham

American Addict

My Hope America with Billy Graham

Greg Fitzsimmons: Life on Stage

Dave Foley: Relatively Well

Underground: The Julian Assange Story

Curious George: A Very Monkey Christmas

Lady Gaga &#38; The Muppets' Holiday Spectacular

Silver Bells

Time taken: 0.154 seconds, Fetched: 7161 row(s)

8) SELECT count(name) FROM movies\_dataset where rating > 3.0 and rating < 4.0;

Query ID = root\_20210816191316\_7ce502df-b353-4311-a360-7b585e1490fd

Total jobs = 1

Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application\_1629122499604\_0005)

--------------------------------------------------------------------------------

**VERTICES STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED**

--------------------------------------------------------------------------------

Map 1 .......... SUCCEEDED 1 1 0 0 0 0

Reducer 2 ...... SUCCEEDED 1 1 0 0 0 0

--------------------------------------------------------------------------------

**VERTICES: 02/02 [==========================>>] 100% ELAPSED TIME: 4.49 s**

--------------------------------------------------------------------------------

OK

7161

Time taken: 5.246 seconds, Fetched: 1 row(s)

9) SELECT count (\*) FROM movies\_dataset WHERE duration > 7200;

Query ID = root\_20210816191620\_f905f732-0128-40d4-a842-44014903cdcb

Total jobs = 1

Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application\_1629122499604\_0005)

--------------------------------------------------------------------------------

**VERTICES STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED**

--------------------------------------------------------------------------------

Map 1 .......... SUCCEEDED 1 1 0 0 0 0

Reducer 2 ...... SUCCEEDED 1 1 0 0 0 0

--------------------------------------------------------------------------------

**VERTICES: 02/02 [==========================>>] 100% ELAPSED TIME: 4.92 s**

--------------------------------------------------------------------------------

OK

641

Time taken: 5.679 seconds, Fetched: 1 row(s)

10)SELECT year, count (\*) FROM movies\_dataset GROUP BY (year);

Query ID = root\_20210816191914\_a662d1e4-b2fa-4912-81a8-8cf04957e741

Total jobs = 1

Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application\_1629122499604\_0005)

--------------------------------------------------------------------------------

**VERTICES STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED**

--------------------------------------------------------------------------------

Map 1 .......... SUCCEEDED 1 1 0 0 0 0

Reducer 2 ...... SUCCEEDED 1 1 0 0 0 0

--------------------------------------------------------------------------------

**VERTICES: 02/02 [==========================>>] 100% ELAPSED TIME: 4.05 s**

--------------------------------------------------------------------------------

OK

1913 3

1914 20

1915 1

1916 1

1918 1

1919 3

1920 6

1921 2

1922 2

1923 4

1924 5

1925 5

1926 2

1927 4

1928 2

1929 5

1930 5

1931 3

1932 4

1933 7

1934 8

1935 11

1936 7

1937 4

1938 5

1939 6

1940 9

1941 7

1942 3

1943 7

1944 10

1945 9

1946 6

1947 9

1948 13

1949 9

1950 10

1951 33

1952 15

1953 17

1954 17

1955 14

1956 60

1957 98

1958 73

1959 87

1960 123

1961 119

1962 124

1963 88

1964 126

1965 104

1966 103

1967 279

1968 173

1969 124

1970 141

1971 131

1972 166

1973 168

1974 178

1975 134

1976 118

1977 136

1978 231

1979 140

1980 107

1981 112

1982 153

1983 270

1984 303

1985 334

1986 287

1987 280

1988 334

1989 421

1990 470

1991 364

1992 479

1993 564

1994 517

1995 592

1996 688

1997 788

1998 843

1999 1181

2000 902

2001 1173

2002 1117

2003 1399

2004 1381

2005 1937

2006 2416

2007 2892

2008 3358

2009 4451

2010 5107

2011 5511

2012 4339

2013 981

2014 1

Time taken: 4.672 seconds, Fetched: 101 row(s)

11) SELECT count(\*) FROM movies\_dataset;

Query ID = root\_20210816192307\_32e73d2d-e9d6-4993-8298-7fcf568a4ac4

Total jobs = 1

Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application\_1629122499604\_0005)

--------------------------------------------------------------------------------

**VERTICES STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED**

--------------------------------------------------------------------------------

Map 1 .......... SUCCEEDED 1 1 0 0 0 0

Reducer 2 ...... SUCCEEDED 1 1 0 0 0 0

--------------------------------------------------------------------------------

**VERTICES: 02/02 [==========================>>] 100% ELAPSED TIME: 3.77 s**

--------------------------------------------------------------------------------

OK

49590

Time taken: 4.446 seconds, Fetched: 1 row(s)

12)desc movies\_dataset;

13) SELECT count(name) FROM movies\_dataset where rating > 3.0;

Query ID = root\_20210817101946\_b89c2b94-510f-43f5-bdb9-2b248470f7b6

Total jobs = 1

Launching Job 1 out of 1

Tez session was closed. Reopening...

Session re-established.

Status: Running (Executing on YARN cluster with App id application\_1629176045948\_0004)

--------------------------------------------------------------------------------

**VERTICES STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED**

--------------------------------------------------------------------------------

Map 1 .......... SUCCEEDED 1 1 0 0 0 0

Reducer 2 ...... SUCCEEDED 1 1 0 0 0 0

--------------------------------------------------------------------------------

**VERTICES: 02/02 [==========================>>] 100% ELAPSED TIME: 39.06 s**

--------------------------------------------------------------------------------

OK

8639

Time taken: 92.849 seconds, Fetched: 1 row(s)

14) SELECT count(\*) FROM movies\_dataset where rating > 4;

Query ID = root\_20210817110929\_ce5bb378-eabe-4e41-be37-7f52ae779a14

Total jobs = 1

Launching Job 1 out of 1

Tez session was closed. Reopening...

Session re-established.

Status: Running (Executing on YARN cluster with App id application\_1629176045948\_0005)

--------------------------------------------------------------------------------

**VERTICES STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED**

--------------------------------------------------------------------------------

Map 1 .......... SUCCEEDED 1 1 0 0 0 0

Reducer 2 ...... SUCCEEDED 1 1 0 0 0 0

--------------------------------------------------------------------------------

**VERTICES: 02/02 [==========================>>] 100% ELAPSED TIME: 22.15 s**

--------------------------------------------------------------------------------

OK

898

Time taken: 63.653 seconds, Fetched: 1 row(s)

15) SELECT count(\*) FROM movies\_dataset where year>2000;

Query ID = root\_20210817111533\_e2da8398-36d0-4f8f-87b4-1210050ebe4d

Total jobs = 1

Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application\_1629176045948\_0005)

--------------------------------------------------------------------------------

**VERTICES STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED**

--------------------------------------------------------------------------------

Map 1 .......... SUCCEEDED 1 1 0 0 0 0

Reducer 2 ...... SUCCEEDED 1 1 0 0 0 0

--------------------------------------------------------------------------------

**VERTICES: 02/02 [==========================>>] 100% ELAPSED TIME: 37.12 s**

--------------------------------------------------------------------------------

OK

36063

Time taken: 40.532 seconds, Fetched: 1 row(s)

16)select distinct name from movies\_dataset;

Ok

Zoo Baby

Zoolander

Zu Warriors

Zulu

Zulu Dawn

[Rec] 3: Genesis

deadmau5: Meowingtons Hax 2K11

eCupid

eXistenZ

iSteve

�vocateur: The Morton Downey Jr. Movie

Time taken: 25.569 seconds, Fetched: 49143 row(s)

17) select distinct name from movies\_dataset limit 5;

Query ID = root\_20210817114455\_a5c6cc3d-4fd9-4f05-9580-1b16cff238e9

Total jobs = 1

Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application\_1629176045948\_0006)

--------------------------------------------------------------------------------

**VERTICES STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED**

--------------------------------------------------------------------------------

Map 1 .......... SUCCEEDED 1 1 0 0 0 0

Reducer 2 ...... SUCCEEDED 1 1 0 0 0 0

--------------------------------------------------------------------------------

**VERTICES: 02/02 [==========================>>] 100% ELAPSED TIME: 4.14 s**

--------------------------------------------------------------------------------

OK

#1 Cheerleader Camp

'Til Death

'Til Death: Season 1

'Til Death: Season 1: Clay Date

'Til Death: Season 1: Daddy's Girl

Time taken: 4.735 seconds, Fetched: 5 row(s)

18) select name,year from movies\_dataset order by year limit 10;

Query ID = root\_20210817114744\_ac4ed6d6-50a2-428f-9a57-9eeaffe3accd

Total jobs = 1

Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application\_1629176045948\_0006)

--------------------------------------------------------------------------------

**VERTICES STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED**

--------------------------------------------------------------------------------

Map 1 .......... SUCCEEDED 1 1 0 0 0 0

Reducer 2 ...... SUCCEEDED 1 1 0 0 0 0

--------------------------------------------------------------------------------

**VERTICES: 02/02 [==========================>>] 100% ELAPSED TIME: 4.48 s**

--------------------------------------------------------------------------------

OK

Fant�mas III: The Murderous Corpse 1913

Fant�mas II: Juve vs. Fant�mas 1913

Fant�mas I: In the Shadow of the Guillotine 1913

Fant�mas V: The False Magistrate 1914

Charlie Chaplin Collection: Shorts: A Night In the Show The Bank 1914

Charlie Chaplin Collection: Shorts: By the Sea His Regeneration Mabel at the Wheel 1914

Charlie Chaplin Collection: Shorts: Behind the Screen The Knockout Between Showers 1914

Fant�mas IV: Fant�mas vs. Fant�mas 1914

Charlie Chaplin Collection: Shorts: A Night Out Kid Auto Races In Venice Cruel Cruel Love; Film Johnny 1914

Charlie Chaplin Collection: Shorts: The Immigrant Easy Street 1914

Time taken: 5.232 seconds, Fetched: 10 row(s)

19) select count(\*) year from movies\_dataset order by year;

Query ID = root\_20210817121818\_76819401-4647-4fe6-b789-ab2140eed654

Total jobs = 1

Launching Job 1 out of 1

Tez session was closed. Reopening...

Session re-established.

Status: Running (Executing on YARN cluster with App id application\_1629176045948\_0007)

--------------------------------------------------------------------------------

**VERTICES STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED**

--------------------------------------------------------------------------------

Map 1 .......... SUCCEEDED 1 1 0 0 0 0

Reducer 2 ...... SUCCEEDED 1 1 0 0 0 0

Reducer 3 ...... SUCCEEDED 1 1 0 0 0 0

--------------------------------------------------------------------------------

**VERTICES: 03/03 [==========================>>] 100% ELAPSED TIME: 4.86 s**

--------------------------------------------------------------------------------

OK

49590

Time taken: 13.79 seconds, Fetched: 1 row(s)

20) select name,year from movies\_dataset order by year desc;

Charlie Chaplin Collection: Shorts: The Tramp Police 1914

Charlie Chaplin Collection: Shorts: Shanghaied Triple Trouble 1914

Charlie Chaplin Collection: Shorts: A Woman Rival Mashers 1914

Charlie Chaplin Collection: Shorts: Work In The Park Good For Nothing 1914

Charlie Chaplin Collection: Shorts: One A.M. The Pawnshop 1914

Fant�mas III: The Murderous Corpse 1913

Fant�mas II: Juve vs. Fant�mas 1913

Fant�mas I: In the Shadow of the Guillotine 1913

Time taken: 5.142 seconds, Fetched: 49590 row(s)

21) SELECT id, COUNT(id),AVG(rating) FROM movies\_dataset GROUP BY id LIMIT 5;

Query ID = root\_20210817173053\_d047463e-d42b-41ff-8c34-2f65abe4ea9a

Total jobs = 1

Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application\_1629217855191\_0003)

--------------------------------------------------------------------------------

**VERTICES STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED**

--------------------------------------------------------------------------------

Map 1 .......... SUCCEEDED 1 1 0 0 0 0

Reducer 2 ...... SUCCEEDED 1 1 0 0 0 0

--------------------------------------------------------------------------------

**VERTICES: 02/02 [==========================>>] 100% ELAPSED TIME: 52.27 s**

--------------------------------------------------------------------------------

OK

1 1 3.9000000953674316

2 1 3.5

3 1 3.200000047683716

4 1 2.799999952316284

5 1 2.799999952316284

Time taken: 74.961 seconds, Fetched: 5 row(s)

22) SELECT id,AVG(rating) FROM movies\_dataset GROUP BY id LIMIT 5;

Query ID = root\_20210817191028\_870874d3-ce49-46f3-9f9b-45b93298f97a

Total jobs = 1

Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application\_1629217855191\_0004)

--------------------------------------------------------------------------------

**VERTICES STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED**

--------------------------------------------------------------------------------

Map 1 .......... SUCCEEDED 1 1 0 0 0 0

Reducer 2 ...... SUCCEEDED 1 1 0 0 0 0

--------------------------------------------------------------------------------

**VERTICES: 02/02 [==========================>>] 100% ELAPSED TIME: 34.16 s**

--------------------------------------------------------------------------------

OK

1 3.9000000953674316

2 3.5

3 3.200000047683716

4 2.799999952316284

5 2.799999952316284

Time taken: 37.955 seconds, Fetched: 5 row(s)

23) SELECT id,MAX(year) FROM movies\_dataset GROUP BY id LIMIT 5;

Total jobs = 1t\_20210817191718\_78848361-8723-4c87-ae30-969da07f722e

Status: Running (Executing on YARN cluster with App id application\_1629217855191\_0004)

--------------------------------------------------------------------------------

**VERTICES STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED**

--------------------------------------------------------------------------------

Map 1 .......... SUCCEEDED 1 1 0 0 0 0

Reducer 2 ...... SUCCEEDED 1 1 0 0 0 0

--------------------------------------------------------------------------------

**VERTICES: 02/02 [==========================>>] 100% ELAPSED TIME: 15.28 s**

--------------------------------------------------------------------------------

OK

2 1932

4 1991

5 1963

hive> taken: 16.937 seconds, Fetched: 5 row(s)

24) select count(\*) from movies\_dataset where year between 2000 and 2016;

Total jobs = 1t\_20210817192102\_f74798b9-44f5-498b-9d93-5e84720f1723

Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application\_1629217855191\_0004)

--------------------------------------------------------------------------------

**VERTICES STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED**

--------------------------------------------------------------------------------

Map 1 .......... SUCCEEDED 1 1 0 0 0 0

Reducer 2 ...... SUCCEEDED 1 1 0 0 0 0

--------------------------------------------------------------------------------

**VERTICES: 02/02 [==========================>>] 100% ELAPSED TIME: 17.59 s**

--------------------------------------------------------------------------------

OK

Time taken: 20.836 seconds, Fetched: 1 row(s)

25) select count(\*) from movies\_dataset where year between 2000 and 2016;

Query ID = root\_20210817193103\_bda90f37-ca5f-4958-b1fe-4df00314e835

Total jobs = 1

Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application\_1629217855191\_0004)

--------------------------------------------------------------------------------

**VERTICES STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED**

--------------------------------------------------------------------------------

Map 1 .......... SUCCEEDED 1 1 0 0 0 0

Reducer 2 ...... SUCCEEDED 1 1 0 0 0 0

--------------------------------------------------------------------------------

**VERTICES: 02/02 [==========================>>] 100% ELAPSED TIME: 44.33 s**

--------------------------------------------------------------------------------

OK------------------------------------------------------------------------------

36965

Time taken: 48.149 seconds, Fetched: 1 row(s)

26)To create an duplicate table

create table movies\_dataset1(id int, name string, year int, rating float, duration int) row format delimited fields terminated by ',' stored as textf

ile location '/moviedata/';

OK

Time taken: 0.893 seconds

27) select ascii(name) from movies\_dataset limit 5;

OK

84

84

79

84

78

Time taken: 1.852 seconds, Fetched: 5 row(s)

**SQOOP COMMANDS**

# Take data from MySQL DB to your HDFS using Sqoop

sqoop import --connect jdbc:mysql://127.0.0.1:3306/movie -username root -password hortonworks1 -table movies\_dataset -m 1 --target-di

r /movies\_dataset-table

#sqoop eval command

sqoop eval --connect jdbc:mysql://127.0.0.1:3306/movie --username root --password hortonworks1 --query "select \* from movies\_dataset

limit 5;"

# List the tables in the movie database

sqoop list-tables --connect jdbc:mysql://127.0.0.1:3306/movie --username root --password hortonworks1;

create table bucket\_movies\_dataset(id int, name string, year int, rating float, dur

ation int) clustered by (rating) into 4 buckets;

OK

Time taken: 2.546 seconds