**Week 3 Assignment**

Guruprasad Velikadu Krishnamoorthy

College of Science and Technology, Bellevue University

[DSC650-T301: Big](https://cyberactive.bellevue.edu/webapps/blackboard/execute/courseMain?course_id=_512542_1) Data

Professor. Nasheb Ismaily

December 10, 2023

1. **Loading the grades.csv file into Hive:**

The Screenshot below shows grades.csv is loaded into the Hive table and a select query is executed:

A screenshot of a computer

Description automatically generated

1. **Screenshot of Commands executed on grades table:**

Commands used are:

1. Select count(\*) from grades;

***Significance:*** *To check the count of rows in the table*

1. Show databases;

***Significance:*** *Returns the list of databases.*

1. Show tables;

***Significance:*** *Returns the list of tables.*

1. Describe grades;

***Significance:*** *Returns the table structure with the name of the column and its datatype.*

1. Select sum(test1) from grades;

***Significance:*** *returns the sum of all values of test1 variable in the grades table.*

1. Select grade,count(\*) from grades group by grade;

***Significance:*** *returns the total count of each grade scored by the students.*

A screenshot of a computer

Description automatically generated

A computer screen with icons on it

Description automatically generated

1. **Loading a new Dataset into Hive:**

For this exercise, the Hollywood movies dataset from GitHub was downloaded from https://raw.githubusercontent.com/reisanar/datasets/master/HollywoodMovies.csv.

Commands used are listed below:

* *# To download the file from GitHub into Google Cloud VM.*

wget <https://raw.githubusercontent.com/reisanar/datasets/master/HollywoodMovies.csv>

* *# After starting the docker, copying the file into master container from the VM.*

docker compose cp /home/KRISUNICORN/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase/data/HollywoodMovies.csv master:/data/

* *# Copying the file from Master Container into HDFS filesystem in the / directory*

hdfs dfs -put /data/HollywoodMovies.csv /

A screenshot of a computer program

Description automatically generated

1. **The Reason for choosing the dataset?**

The movie dataset contains 970 Rows and 16 columns that contain the details of the profitable Hollywood movies released from 2007 to 2013. As the number of records wasn’t too small, it seemed perfect for this assignment project. Also, I am a movie enthusiast, so I thought it may be a little fun to do a Data Science project on Movie related datasets.

1. **The description of some of the commonly used columns is explained below:**

* ***Movie:*** Name of the movie
* ***LeadStudio:*** TheStudio that produced the movie.
* ***RottenTomatoes:*** Represents the score rated by Rotten Tomatoes critics.
* ***AudienceScore***: Represents the score rated by Audience critics.
* ***Genre***: Represents the Genre of the movie
* ***TheatersOpenWeek***: Number of theatres where the movie was released in the opening week in the US and Canada. This is an integer value.
* ***OpeningWeekend*:** Box Office collection in the opening week in the US and Canada in USD. This is a floating number value.
* ***DomesticGross:*** *Total domestic collection in the US and Canada represented in USD.*
* ***ForeignGross:*** *Total collection from countries other than the US and Canada represented in USD.*
* ***WorldGross:*** Sum of Domestic and Foreign Gross collection represented in USD.
* ***Budget:*** Total movie budget represented in USD.
* ***Profitability:*** Total profits made from the movie represented in USD
* ***Year***: The year in which the movie was released.

1. **What is expected to be achieved by analyzing the dataset?**

The questions that we are trying to answer by analyzing the dataset are listed below:

1. How many movies had a perfect Rotten Tomatoes score of 95% and above in the dataset?
2. How many Profitable movies were made each year?
3. What was the average profit made each year in each of the Genres?
4. What are the variance, Mean, and Standard Deviation values of Domestic Gross each year?
5. What was the most profitable movie made each year and by how much?
6. **Creating tables and loading data in Hive:**

*# Creating Hive table and defining the datatype of each column*

CREATE TABLE movies(

`Movie` STRING,

`LeadStudio` STRING,

`RottenTomatoes` INT,

`AudienceScore` INT,

`Story` STRING,

`Genre` STRING,

`TheatersOpenWeek` INT,

`OpeningWeekend` DOUBLE,

`BOAvgOpenWeekend` DOUBLE,

`DomesticGross` DOUBLE,

`ForeignGross` DOUBLE,

`WorldGross` DOUBLE,

`Budget` DOUBLE,

`Profitability` DOUBLE,

`OpenProfit` DOUBLE,

`Year` INT

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE

tblproperties("skip.header.line.count"="1");

*# Loading the movies table with the csv file data.*

LOAD DATA INPATH '/HollywoodMovies.csv' INTO TABLE movies;

A screenshot of a computer

Description automatically generated

1. **Running Hive Queries:**

* **Command 1:** *select Movie,RottenTomatoes,Genre,Year from movies where RottenTomatoes >=95;*
* **Significance:** The query returns the movie name, Genre name and Year along with Rotten Tomates where the score is more than or equal to 95.

A screen shot of a computer

Description automatically generated

* **Command 2:** *select Year,count(Movie) from movies where Year >=2007 group by Year;*
* **Significance:** The query returns the count of number of movies from each year in the dataset.

**A screenshot of a computer

Description automatically generated**

* **Command 3:** *select Genre,Year,avg(Profitability) from movies where Year >=2007 and Year is not Null and Genre is not Null group by Genre,Year order by Year,Genre;*
* **Significance:** The Query returns the average Profitability of the movies from the years 2007-2013 per Genre.

A screenshot of a computer

Description automatically generated

* **Command 4:** *select avg(DomesticGross),var\_pop(DomesticGross),stddev\_pop(DomesticGross),Year from movies where Year >=2007 group by Year;*
* **Significance:** The Query returns statistical measures of Domestic Gross such as Variance, Standard deviation, and mean.

A screenshot of a computer

Description automatically generated

**Command 5**: *SELECT Movie,Year,Profitability from movies where Profitability in (select max(Profitability) from movies where Year >=2007 group by Year) order by Year;*

**Significance**: The query returns the most profitable movie of each year from 2007 to 2013.

A screenshot of a computer

Description automatically generated