Text

Description automatically generated

**A PROJECT REPORT ON**

AMAZON MUSIC REVIEWS

**For the Course:**

Big Data Analytics (AY2022- 2023: TRIMESTER-III)

Submitted in partial fulfilment of the requirements for the

Two-year full time Master’s Degree in MBA - Business Analytics

**Submitted By**

Abinash Mishra - Roll No: B003 (SAP ID: 80812200003)

Aditi Srivastava - Roll No: B004 (SAP ID: 80812200065)

Ashwin Lanka - Roll No: B014 (SAP ID: 80812200016)

Bezawada Jeevika - Roll No: B016 (SAP ID: 8081200177)

Bindu Vasini Kanney - Roll No: B018 (SAP ID: 80812200020)

(Batch of 2022- 24)

**Submitted to**

**Faculty Mentor**

Prof. Cyrus Lentin,

Centre of Excellence in Analytics & Data Science,

SVKM'S NMIMS, MUMBAI

**MARCH 2023**

**Table of Contents (TOC)**

**Content**

[Project Overview 3](#_Toc129718410)

[**Brief Overview of the Project** 3](#_Toc129718411)

[**Learning Objectives** 3](#_Toc129718412)

[Code & Command Section 3](#_Toc129718413)

[**HDFS Code:** 3](#_Toc129718414)

[**PIG Code:** 4](#_Toc129718415)

[**HIVE Code:** 5](#_Toc129718416)

[**SQOOP Code:** 9](#_Toc129718417)

[Summary 9](#_Toc129718418)

[**Experience of using Hadoop for analysing Big Data** 10](#_Toc129718419)

**List of Figures**

[**Figure 1: Creation of the folder in the HDFS directory** 4](file:///E:\NMIMS\Academics\3rd%20Trimester\Big%20Data%20Analytics\Group%20Project\Group5_BDA_Report.docx#_Toc129715859)

[**Figure 2: Output for the dump command for Pig** 5](file:///E:\NMIMS\Academics\3rd%20Trimester\Big%20Data%20Analytics\Group%20Project\Group5_BDA_Report.docx#_Toc129715860)

[**Figure 3: Output file being stored in the HDFS directory** 5](file:///E:\NMIMS\Academics\3rd%20Trimester\Big%20Data%20Analytics\Group%20Project\Group5_BDA_Report.docx#_Toc129715861)

[**Figure 4: Output for the table description** 6](file:///E:\NMIMS\Academics\3rd%20Trimester\Big%20Data%20Analytics\Group%20Project\Group5_BDA_Report.docx#_Toc129715862)

[**Figure 5: Output for Query 1** 6](file:///E:\NMIMS\Academics\3rd%20Trimester\Big%20Data%20Analytics\Group%20Project\Group5_BDA_Report.docx#_Toc129715863)

[**Figure 6: Output for query 2** 7](file:///E:\NMIMS\Academics\3rd%20Trimester\Big%20Data%20Analytics\Group%20Project\Group5_BDA_Report.docx#_Toc129715864)

[**Figure 7: Output for query 3** 7](file:///E:\NMIMS\Academics\3rd%20Trimester\Big%20Data%20Analytics\Group%20Project\Group5_BDA_Report.docx#_Toc129715865)

[**Figure 8: Output for query 4** 8](file:///E:\NMIMS\Academics\3rd%20Trimester\Big%20Data%20Analytics\Group%20Project\Group5_BDA_Report.docx#_Toc129715866)

[**Figure 9: Output for query 5** 8](file:///E:\NMIMS\Academics\3rd%20Trimester\Big%20Data%20Analytics\Group%20Project\Group5_BDA_Report.docx#_Toc129715867)

[**Figure 10: Output for query 6** 8](file:///E:\NMIMS\Academics\3rd%20Trimester\Big%20Data%20Analytics\Group%20Project\Group5_BDA_Report.docx#_Toc129715868)

# **Project Overview**

## **Brief Overview of the Project**

Amazon Music is a music streaming service that allows users to access a vast collection of songs, playlists, and podcasts. The service offers various subscription plans, including Amazon Music Unlimited, which provides access to over 70 million songs and is compatible with various devices, including desktop, mobile, and smart speakers. The project focusses on viewing the reviews posted by people for the various music tracks made available by Amazon along with the ratings provided for the same.

## **Learning Objectives**

1. Appreciating the fundamental concepts of Big Data Management and Analytics.
2. To gain hands on experience on using Hadoop architecture and develop basic proficiency in using the same.
3. To perform Big Data Analytics on mega dataset using Hadoop architecture such as: HDFS, Pig, HIVE and SQOOP.

# **Code & Command Section**

## **HDFS Code:**

**# *Starting the HDFS terminal***

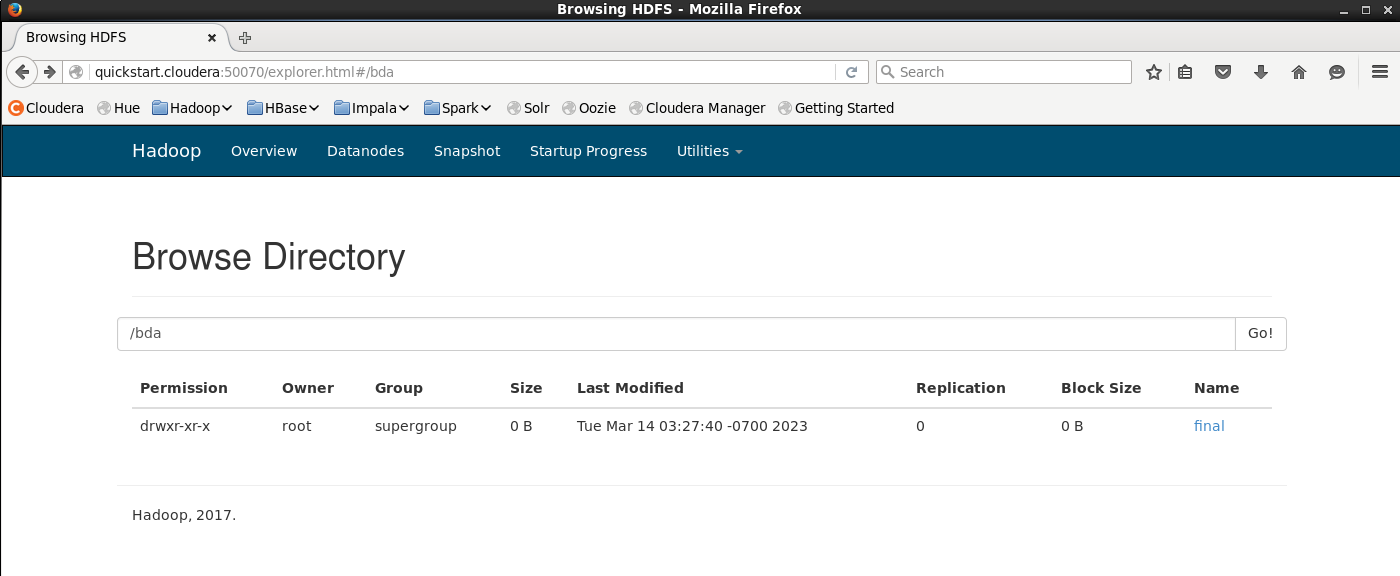
su root

Password: cloudera

***# Creating the folder in the hdfs***

>> hdfs dfs -mkdir /bda

>> hdfs dfs -ls /bda

>> hdfs dfs -mkdir /bda/final

**Figure 1: Creation of the folder in the HDFS directory**

***# Shifting the file from the local to hdfs***

>> hdfs dfs -put /home/cloudera/myfiles/review.txt /bda/final/amazon\_review.txt

>> hdfs dfs -ls /bda/final

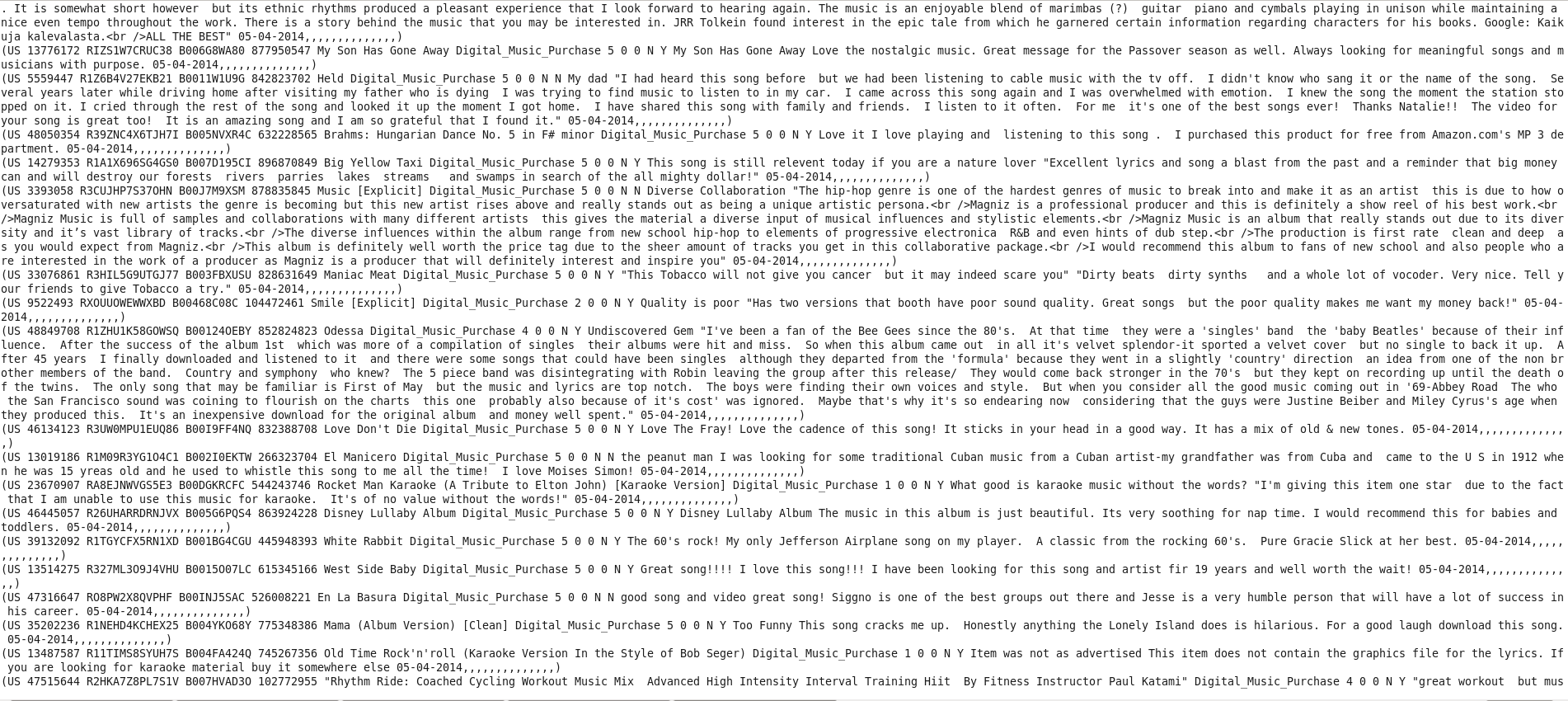
## **PIG Code:**

***# Starting the PIG terminal***

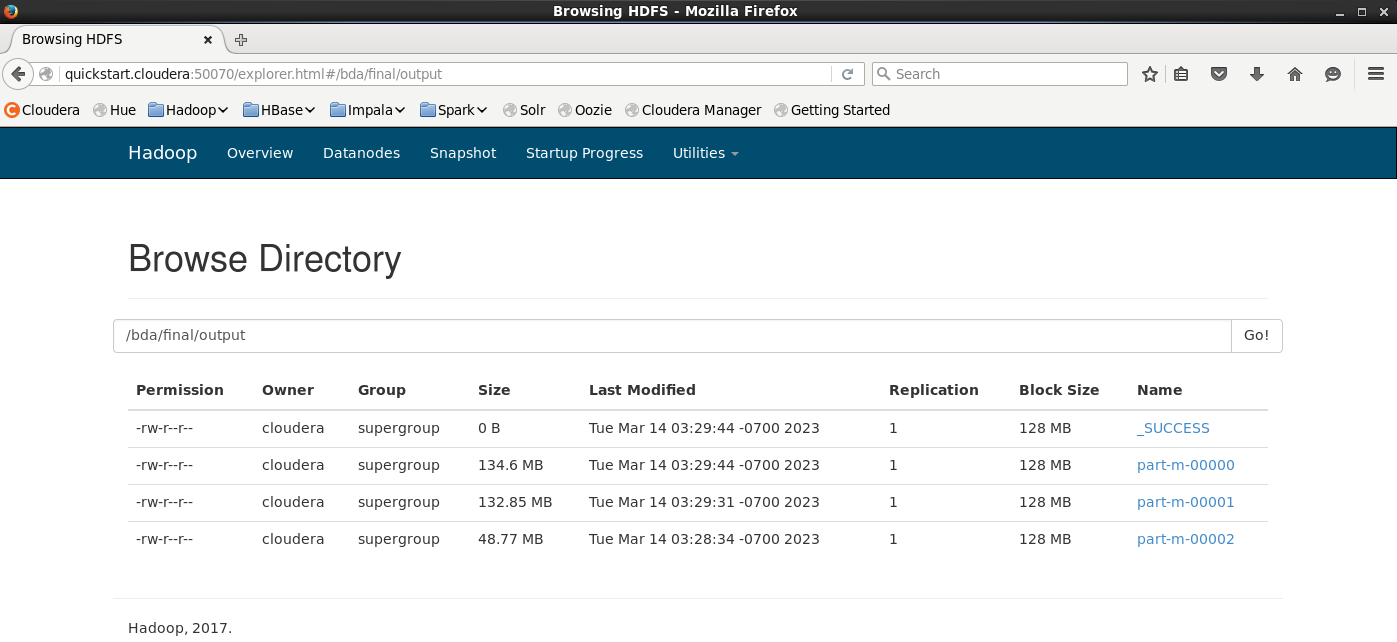
>> pig

***# Parsing the text file stored in the hdfs***

>> amazon = load ‘/bda/final’ using PigStorage (‘\t’) as ( marketplace:chararray, customer\_id:chararray, review\_id:chararray, product\_id:int, product\_parent:chararray, product\_title:chararray, product\_category:chararray, star\_rating:int, helpful\_votes:int, total\_votes:int, vine:chararray, verified\_purchase:chararray, review\_headline:chararray, review\_body:chararray, review\_date:chararray );

****>> dump amazon;

**Figure 2: Output for the dump command for Pig**

>> store amazon into '/bda/final/output' using PigStorage(',');

**Figure 3: Output file being stored in the HDFS directory**

## **HIVE Code:**

***# Starting the HIVE terminal***

>> beeline -u jdbc:hive2://

***# Creating the database***

>> create database amazon;

>> use amazon;

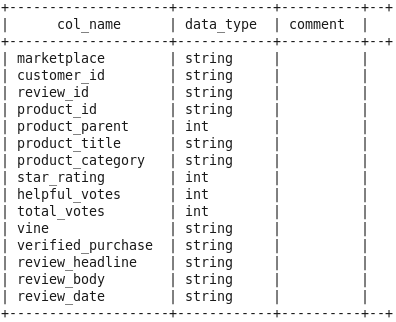
***# Creating the table***

>> create table reviews (marketplace string, customer\_id string, review\_id string, product\_id string, product\_parent int, product\_title string, product\_category string, star\_rating int, helpful\_votes int, total\_votes int, vine string, verified\_purchase string, review\_headline string, review\_body string, review\_date string) row format delimited fields terminated by ',' lines terminated by '\n';

***# Loading the data***

>> load data local inpath '/home/cloudera/myfiles/amazon\_review.csv' overwrite into table reviews;

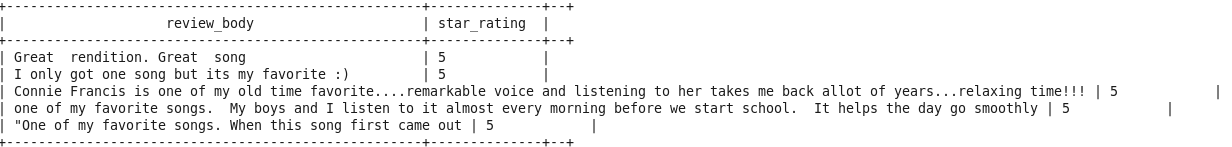
***# Checking the description of the table***

>> desc reviews;

**Figure 4: Output for the table description**

***# Running Queries to generate insights***

1. **To know the overall reviews for the review headline titled as “favorites”**

>> select review\_body from reviews where review\_headline like 'favorites';

**Figure 5: Output for Query 1**

**Insight –** It is observed that the people have provided the highest rating for the music tracks categorized as favourites under the review headline. Additionally, their response has been very positive towards the type of the music and even the music composer.

1. **Finding the music tracks which have been reviewed as great by the people**

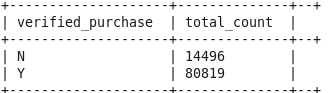
>> select product\_title from reviews where review\_body like ‘%Great%;

**Figure 6: Output for query 2**

**Insight –** It was observed that music tracks such as the one provided above in the output have been rated Great by the people. Some of the noted artists who have appeared on the list include: Kelly Clarkson, Pharell Williams etc.

1. **Finding the count of music tracks which are verified and unverified as well as reviewed as Great by the people.**

>> select verified\_purchase, count(product\_title) as Total\_Count from reviews where review\_body like '%Great%' group by verified\_purchase;

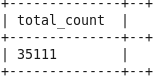


**Figure 7: Output for query 3**

**Insight –** It was observed that around 16% purchase of the music tracks have been unverified and yet have been reviewed as “Great” while the share of the verified purchase of the music tracks with “Great” rating is 84%.

1. **Finding the count of music tracks which have been loved by the people as posted in the review headline.**

>> select count(product\_title) as Total\_Count from reviews where review\_headline like '%Love%';

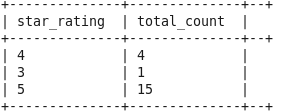


**Figure 8: Output for query 4**

**Insight –** It is observed that people have loved 35,111 music tracks provided by Amazon.

1. **Finding the count of music which have been rated as 5, reviewed as great and the music track contains “Praise the Lord”.**

>> select star\_rating, count(product\_title) as Total\_Count from reviews where product\_title like '%Praise the Lord%' group by star\_rating;

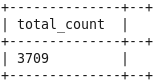


**Figure 9:Output for query 5**

**Insight –** It was observed that 20 music tracks contained the phrase “Praise the Lord”. Out of the 20 music tracks, 75% of the people voted the track to be very good i.e. 5, with 20% of the people voting the track to be average i.e. 3 and the rest 5% voting it to be good.

1. **Finding the count of music tracks which have total votes greater than 10 and has atleast helpful votes greater than 5.**

>> select count(product\_title) as Total\_Count from reviews where total\_votes > 10 and helpful\_votes >= 5;



**Figure 10: Output for query 6**

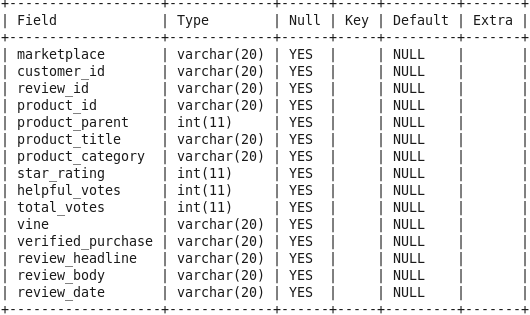
## **SQOOP Code:**

***# Creating the database in mysql***

>> create database amazon\_review;

>> use amazon\_review;

>> create table amazon\_reviews (marketplace varchar(20), customer\_id varchar(20), review\_id varchar(20), product\_id varchar(20), product\_parent int, product\_title varchar(20), product\_category varchar(20), star\_rating int, helpful\_votes int, total\_votes int, vine varchar(20), verified\_purchase varchar(20), review\_headline varchar(20), review\_body varchar(20), review\_date varchar(20));

>> desc amazon\_review;

***# Exporting the hive commands to my sql***

>> sqoop export --connect jdbc:mysql://localhost/amazon\_review --username root -P --table amazon\_reviews --export-dir /bda/final/amazon\_review.csv --input-fields-terminated-by ',' --lines-terminated-by '\n'

# **Summary**

Hadoop has been helpful in Data ingestion, Data Processing and Data Analysis. Data has been ingested into Hadoop Distributed File System. Once the Amazon music Data has been ingested into the system, we have used Hadoop's MapReduce processing framework to perform data transformations and aggregations. It has been used in multiple scenarios like:

* Counting the positive & negative reviews of the music
* Counting the music tracks which have got ratings = 5

After processing the Amazon music data, we have used Hadoop's query engines like Apache Hive and Apache Pig to perform ad-hoc queries and analysis on the data. This has allowed us to extract meaningful insights from our dataset.

## **Experience of using Hadoop for analysing Big Data**

The dataset that we have considered was around 600+ MB dataset which is tough for tools like Excel to handle. But when such a voluminous dataset has been ingested into HDFS, the processing was very smooth and analysis has been performed at the least possible time.

Given the queries that we had to execute, the processing of the information would be very time consuming in any other platforms but Hadoop made the work very easy and smooth to handle.