**GOOGLE PLAYSTORE DATA ANALYSIS**

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| ***TITLE*** | ***PAGE NUMBER*** |
| ***ABSTRACT*** | **1** |
| ***INTRODUCTION*** | **3** |
| ***METHOD*** | **6** |
| ***RESULTS*** | **11** |
| ***SUMMARY*** | **19** |

**ABSTRACT**

The google play store is one of the largest and most pop-

ular Android app stores. It has an enormous amount of data

that can be used to make an optimal model. We have used a

raw data set of Google Play Store from the Kaggle website.

This data set contains 13 different features that can be used

for predicting whether an app will be successful or not using

different features

The google play store is one of the largest and most pop-

ular Android app stores. It has an enormous amount of data

that can be used to make an optimal model. We have used a

raw data set of Google Play Store from the Kaggle website.

This data set contains 13 different features that can be used

for predicting whether an app will be successful or not using

different features

Analysis of Google Play Store Data set and predict the popularity of an app on

Google Play Store

Rimsha Maredia

Texas A&M University

College Stattion, Texas

rimsha.maredia@tamu.edu

Abstract

The google play store is one of the largest and most pop-

ular Android app stores. It has an enormous amount of data

that can be used to make an optimal model. We have used a

raw data set of Google Play Store from the Kaggle website.

This data set contains 13 different features that can be used

for predicting whether an app will be successful or not using

different features. This data set is scraped from the Google

Play Store.

Analysis of Google Play Store Data set and predict the popularity of an app on

Google Play Store

Rimsha Maredia

Texas A&M University

College Stattion, Texas

rimsha.maredia@tamu.edu

Abstract

The google play store is one of the largest and most pop-

ular Android app stores. It has an enormous amount of data

that can be used to make an optimal model. We have used a

raw data set of Google Play Store from the Kaggle website.

This data set contains 13 different features that can be used

for predicting whether an app will be successful or not using

different features. This data set is scraped from the Google

Play Store.

Analysis of Google Play Store Data set and predict the popularity of an app on

Google Play Store

Rimsha Maredia

Texas A&M University

College Stattion, Texas

rimsha.maredia@tamu.edu

Abstract

The google play store is one of the largest and most pop-

ular Android app stores. It has an enormous amount of data

that can be used to make an optimal model. We have used a

raw data set of Google Play Store from the Kaggle website.

This data set contains 13 different features that can be used

for predicting whether an app will be successful or not using

different features. This data set is scraped from the Google

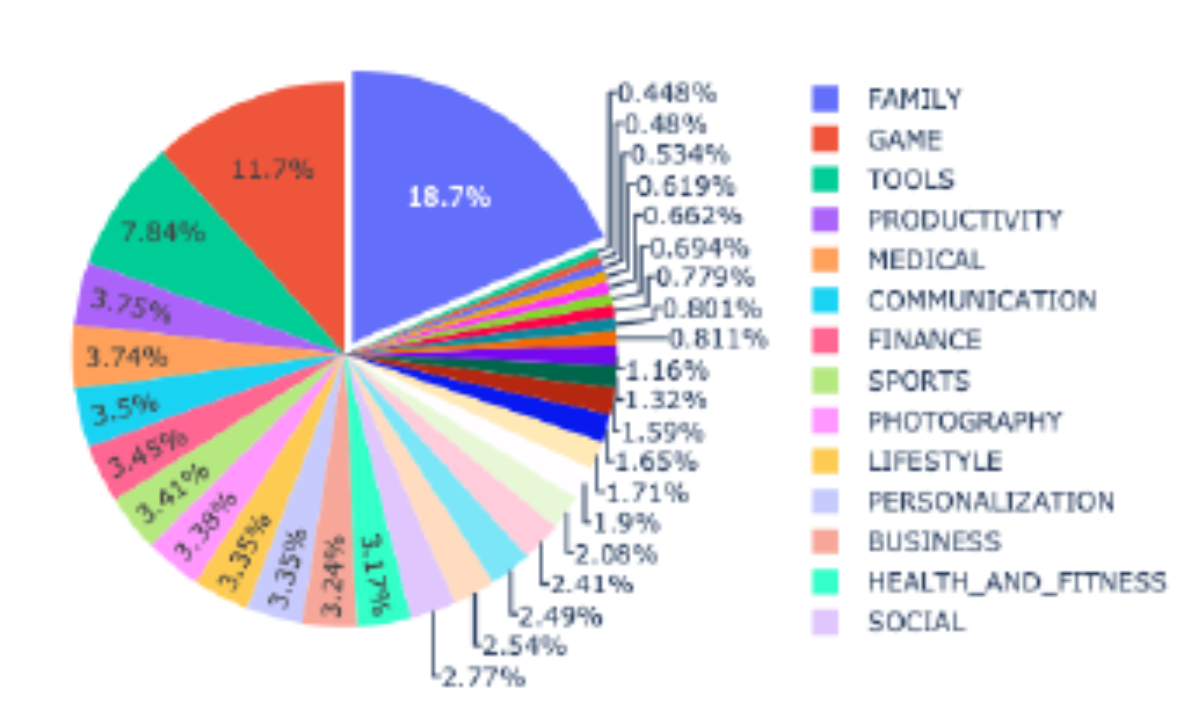
Play Store.

The google play store is one of the largest and most popular Android app stores. It has an enormous amount of data that can be used to make an optimal model. We have used a raw data set of Google Play Store from the Kaggle website. This data set contains 10 different features that can be used for predicting whether an app will be successful or not using different features. This data set is scraped from the GooglePlay Store. The analysis is done using Apache hive and Hadoop tools.

The main aim of the analysis is to find out the average rating of all categories apps in the playstore and number of reviews and downloads of each category apps and the best app among all the categories , the smallest and largest app, the most reviewed app, the most installed app. The dataset is preprocessed and cleaned using MS Excel.

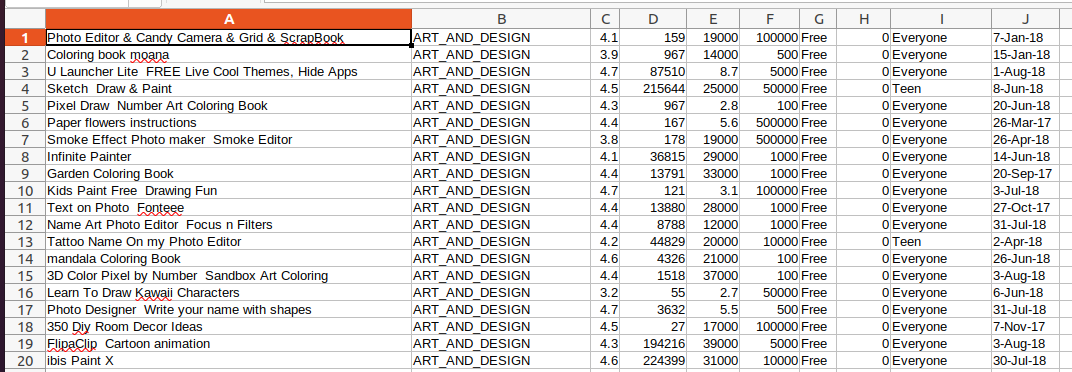
**INTRODUCTION**

The Internet is a true gold mine of data. E-commerce and review sites are brimming with a lot of untapped data with a prominent potential to convert into meaningful insights that can help with robust decision making. Mobile applications are one of the fastest-growing segments of downloadable software application markets. Out of all of the markets we choose Google Play store due to its increasing popularity and recent fast growth. One of the main reasons for this popularity is the fact that about 81% of the apps are free of cost. The market has in-creased to over 845900 Apps and 226,500 unique sellers in April 2013. This rapid market has, in turn, led to over 500 million users downloading around 40 billion Apps all over the world .Developers and users play key roles in determining the impact that market interactions have on future technology. However, the lack of a clear understanding of the inner working and dynamic of popular app markets impacts both the developers and users. In this article, we seek to shed light on the dynamics of the Google Play Store and how we can use different features from this data set for prediction purposes.

****

The dataset contains 10 attributes app category, app name, app rating, reviews, size of the app, number of downloads, type ,price, usertype, Date of release. Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.

The dataset



**METHOD**

The analysis is done by hive in hiveql.

**Analysis 1:**

AVERAGE RATING OF ALL APPS IN THE DATASE.

THIS QUERY WILL FIND THE AVERAGE RATING OF ALL APPLICATIONS IN THE DATASET.

**Code:**

select avg(rating) from apps;

**Analysis 2:**

AVERAGE REVIEWS OF ALL APPS

THIS QUERY WILL FIND THE AVERAGE REVIEW OF ALL APPLICATIONS IN THE DATASET.

**Code** : select avg(review) from apps;

**Analysis 3:**

LIST THE COMICS APP WHICH HAS RATING MORE THAN 4.0

**Code :**

select app from apps where category="COMICS" and rating>4.0;

**Analysis 4:**

User wants to dowload an education app for his/her maths course confused with Khan academy /Coursera: Online courses ?

**Code :**

select app,rating from apps where app="Coursera: Online courses";select app,rating from apps where app="Khan academy";

**Analysis 5:**

AVERAGE RATING OF EACH CATEGORY APPS

THIS QUERY WILL GIVE AVERAGE RATING FOR EACH CATEGORY

**Code :**

select category,avg(rating) from apps group by category;

**Analysis 6:**

AVERAGE Review OF EACH CATEGORY APPS-

**Code :**

select category,avg(review) from apps group by category;

**Analysis 7:**

Apps which have rating==5

THIS QUERY WILL FIND THE APPS WHICH HAVE FULL RATING IN THE PLAYSTORE

**Code :**

select app,rating from apps where rating=5.0;

**Analysis 8:**

MOST REVIEWD APP

THE APP WHICH HAS LARGEST NUMBER OF FEEDBACK. A SEPARATE QUERY IS EXECUTED TO FIND THE LARESGT REVIEW NUMBER AND IT IS 69LAKHS.

**Code :**

select DISTINCT app from apps where review=69119316;

**Analysis 9:**

Largest app

THIS QUERY WILL FIND OUT THE APP WHITH LARGEST DOWNLOAD SIZE IN THE PLAYSTORE.

**Code :**

select max(size) from apps;

select app,category from apps where size=7820209;

**Analysis 10:**

SMALLEST APP

THIS QUERY WILL FIND OUT THE APP WHITH SMALLEST DOWNLOAD SIZE IN THE PLAYSTORE. THE SMALLEST IS IN KB SIZE SO HERE IT IS REFERRED AS 0.

**Code :**

select app,category from apps where size=0.3;

**Analysis 11**

Best game app

**Code:**

select app , category from apps where rating=4.9 and category="GAME";

**Analysis 12**

which category has maxmium no.of apps?

**Code:**

select category,count(app) from apps group by category;

**Analysis 13**

Best Education app

**Code:**

select app from apps where category="EDUCATION" and rating>4.8;

**TABLE STRUCTURE IN HIVE**

DATABASE NAME: project

Analysis of Google Play Store Data set and predict the popularity of an app on

Google Play Store

Rimsha Maredia

Texas A&M University

College Stattion, Texas

rimsha.maredia@tamu.edu

Abstract

The google play store is one of the largest and most pop-

ular Android app stores. It has an enormous amount of data

that can be used to make an optimal model. We have used a

raw data set of Google Play Store from the Kaggle website.

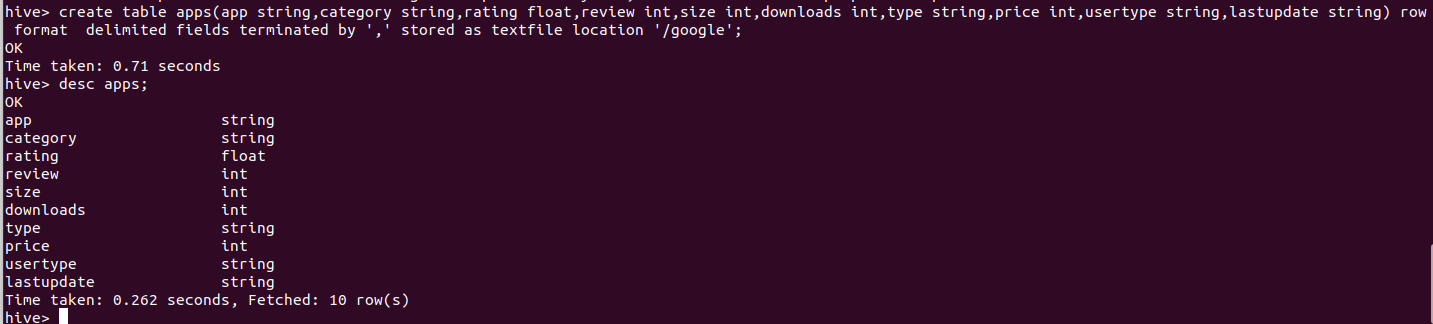
This data set contains 13 different features that can be used

for predicting whether an app will be successful or not using

different features. This data set is scraped from the Google

Play

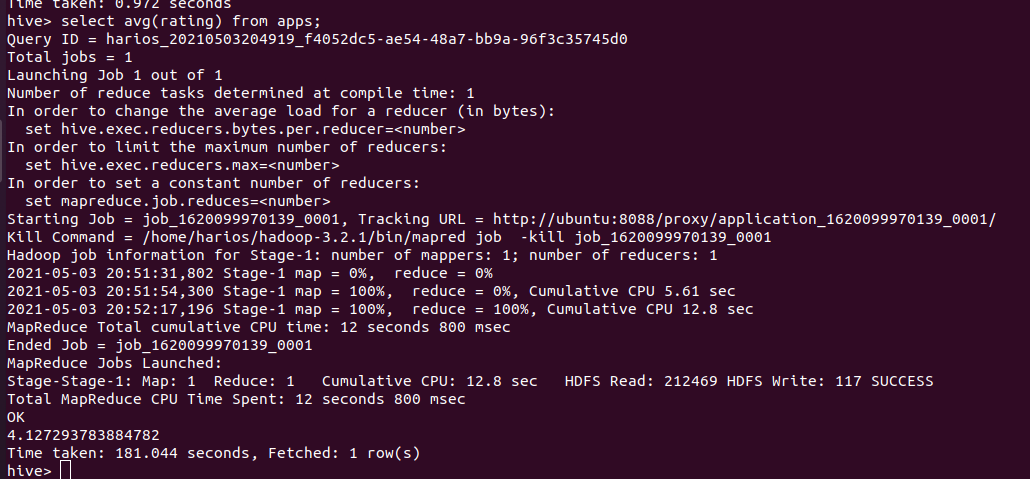
TABLE NAME: apps



**RESULTS**

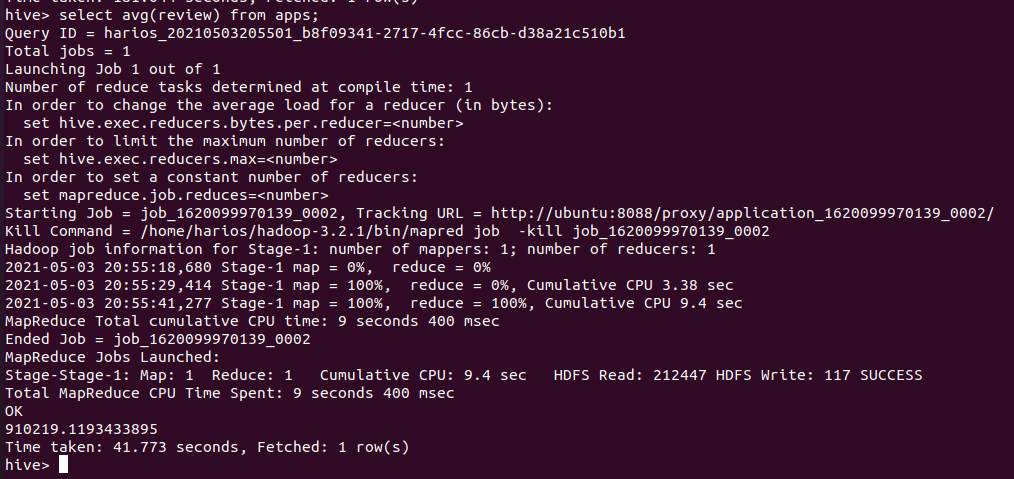
**Analysis 1 output :**

***AVERAGE RATING OF ALL APPS – 4.13***

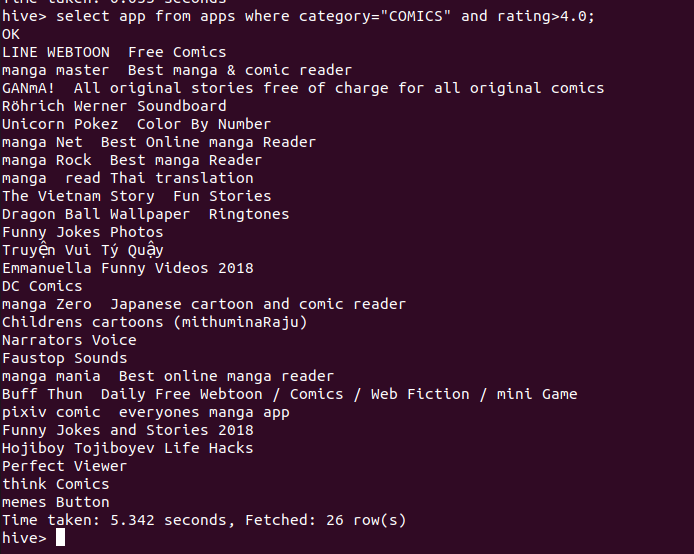
******

***Analysis 2 output:***

***AVERAGE REVIEWS OF ALL APPS-- 910219.1***

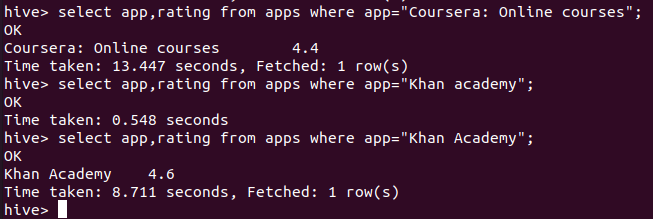
******

***Analysis 3 output:***

******

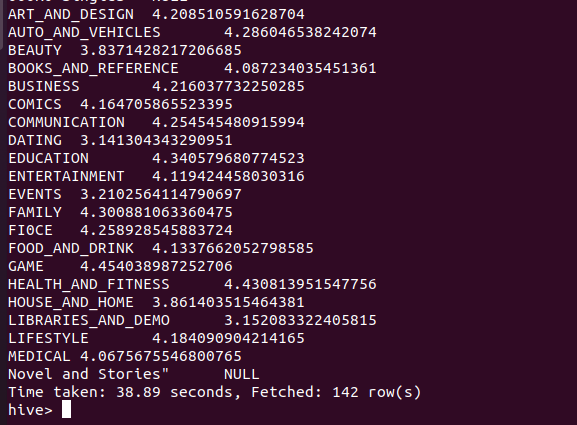
***Analysis 4 output:***

***Best app between courser and khan Academy – Khan Academy***

******

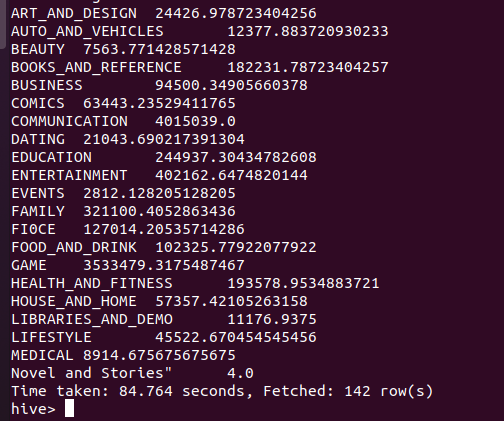
***Analysis 5 output:***

***AVERAGE RATING OF APPS BY CATEGORY***

******

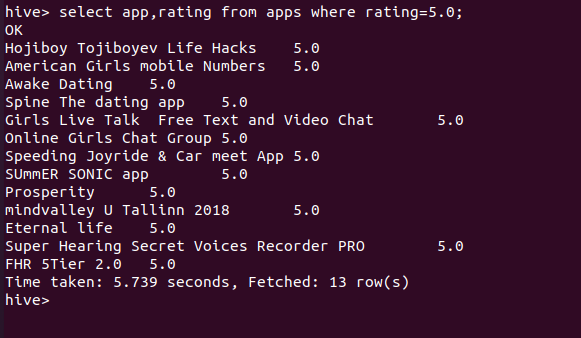
***Analysis 6 output:***

***Average reviews of all apps***

******

***Analysis 7 output:***

***HOW MANY APPS HAVE RATING =5; -- 13 Apps***



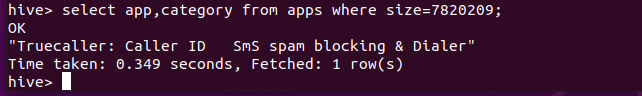
***Analysis 8 output:***

***Most reviewed app -- WHATSAPP***



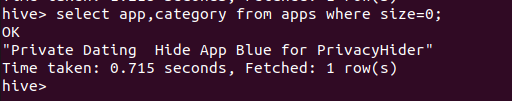
***Analysis 9 output:***

***LARGEST APP- TRUECALLER***

******

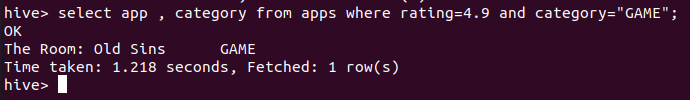
***Analysis 10 output:***

***Smallest App—Private Dating app***

******

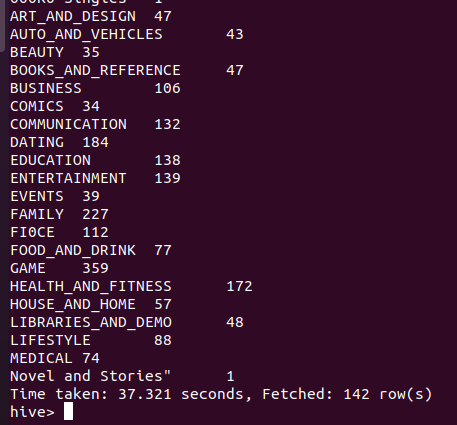
***Analysis 11 output:***

***Best Gaming app—The room :Old sins***

******

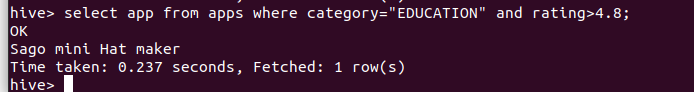
***Analysis 12 output:***

***CATEGORY WITH MAXIMUM NO.OF APPS-FAMILY -- GAME***

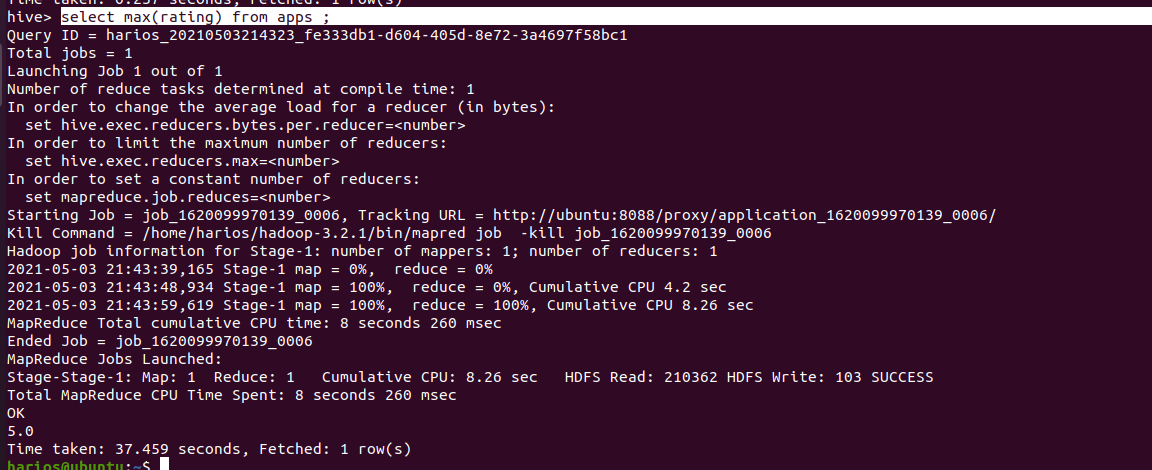
******

***Analysis 13 output:***

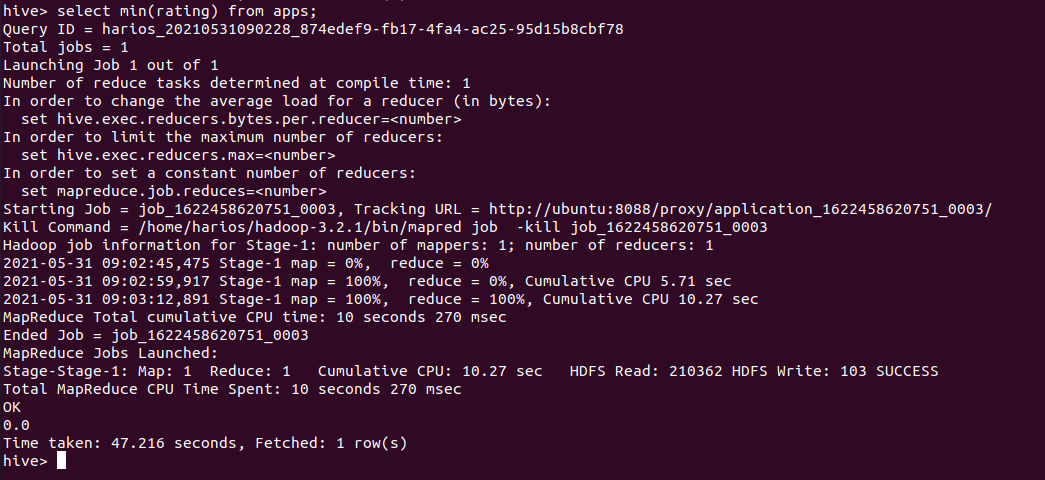
***BEST EDUCATION APP – Sago mini hat maker***



***MAX RATING OF APP IN THE DATASET***



**LOWEST RATING IN THE PLAYSTORE**

******

***SUMMARY***

* *APPS OF CATEGORY GAME HAS HIGHEST AVERAGE RATING =4.45*
* *APPS OF CATEGORY LIBRARIES HAS LOWEST AVERAGE RATING =3.15*
* *AVERAGE RATING OF ALL CATEGORY APP =4.12*
* *AVERAGE REVIEW OF ALL CATEGORY APP =910220*
* *LARGEST APP IS TRUECALLER ID*
* *SMALLEST APP IS PRIVATE DATING APP*
* *BEST GAME =THE ROOM :OLD SIN RATING =5*
* *MOST REVIEWED APP = WHATSAPP*
* *GAME CATEGORY HAS MAXIMUM NUMBER OF APPS*
* *SAGO MINI HAT MAKER IS THE BEST EDUCATION /LEARNING APP*
* *APPS OF CATEGORY GAME HAS HIGHEST REVIEWS.*