Exploring Play Store Application Trends

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology/Master of Technology

In

Computer Science and Engineering School

of Engineering and Sciences

Submitted by

Candidate Name

Dheeraj T (AP21110010022)

Sricharan G (AP21110010029)

Mounish Sai M (AP21110010031)

Under the Guidance of

Rajiv Senapati

SRM University–AP

Neerukonda, Mangalagiri, Guntur Andhra

Pradesh – 522 502

[November, 2024]

Certificate

Date: 25/11/2024

This is to certify that the work present in this Project entitled "**Exploring Play Store Application Trends**" has been carried out by **Dheeraj, Sricharan and Mounish** under my/our supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology/Master of Technology in **School of Engineering and Sciences**.

Supervisor

Dr. Rajiv Senapati

Assistant Professor,

Computer Science and Engineering.

Acknowledgements

We express our sincere thanks to everyone who contributed to the successful completion of this "Exploring Play Store Application Trends" project. First we express our deep gratitude to our respected supervisor Dr. Rajiv Senapati sir. His exceptional mentorship, unwavering support and expert guidance was the cornerstone of this project. His insightful suggestions and encouragement made us achieve our goals. We are also extremely grateful to SRM University - AP for providing us with the enabling environment, resources and academic infrastructure required for this endeavor.

The university's commitment to fostering a culture of learning and innovation was key to the successful implementation of this research. In addition, we would like to express our heartfelt thanks to our family and friends for their constant support, belief in our abilities and understanding during this journey. Their motivation was a constant source of motivation.

We would also like to thank all the individuals, researchers and institutions whose work paved the way for this project. Their commitment to the development of data analysis was an inspiration.

Table of Contents

Abstract

A few thousand new applications are regularly uploaded on Google play store. A huge number of designers are working freely on designing apps and making them successful. With the enormous challenge from everywhere throughout the globe, it is important for a developer to know whether he/she is continuing the correct way or not. To navigate the competitive environment of the Google Play Store and help developers make informed decisions, our experiment explores the complex approaches to get the better decision by analyzing the content. The constant flow of thousands of new apps underscores the need for developers to ensure they are on track in the global competition. Since a significant number of developers work independently to create and optimize applications, it is very important for developers to measure the progress of their efforts. The challenges are multifaceted, especially given the uncertainty surrounding the revenue model. Most apps in the Play Store are free, so it's hard to figure out the impact of in-app purchases, ads, and subscriptions on an app's success. Thus, traditional app and wellness metrics were based solely on installs and user ratings, rather than a clear understanding of monetization. We try to identify patterns, such as the impact of a free or paid app on its success, how user reviews correlate with overall ratings, and what factors affect an app and its popularity over its lifetime. By creating an understanding of customer requirements and preferences, we aim to provide developers with a valuable tool to improve their products and improve visibility in the competitive application market. With this research, we aim to contribute to a broader understanding of the factors that influence the success of the application in the Google Play Store, empowering developers to make informed decisions and optimize their strategies for greater impact.

Abbreviations

SQL	Structured Query Language
CSV	Comma separated by Values
HDFS	Hadoop Distributed File System

1. Introduction

In today's scenario we can see that mobile apps play an important role in any individual's life. With enormous challenges from everywhere throughout the globe, it is important for a designer to realize that he/she is continuing in the right way or not. 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 the downloadable software application market. Out of all the marketplaces, we choose the Google Play Store because of its growing popularity and recent rapid growth. One of the main reasons for its popularity is that about 81% of the programs are free. As of April 2013, the marketplace has grown to over 845,900 apps and 226,500 unique sellers. This booming market has in turn resulted in over 500 million users who have downloaded around 40 billion apps worldwide. Developers and users have a key role in influencing future technology through market interactions.

Central to this project is the utilization of Hive, an open-source data warehousing framework that seamlessly integrates with the Hadoop ecosystem. Hive empowers users with a SQL-like interface, known as HiveQL, MapReduce also provides a familiar and intuitive means to query and analyse data stored within distributed storage systems, such as the Hadoop Distributed File System (HDFS).

In the field of Big Data, where the amount, speed and variety of data is enormous, Hive is an essential tool. It works with the Hadoop distributed computing framework and efficiently processes and analyses huge data sets. Hive and#039; architecture allows complex queries to be turned into multiple parallel jobs that run across a Hadoop cluster. From the point of view of this company, the creation of a safe environment is very important. This requires careful configuration of Hive settings, ensuring seamless integration with the underlying Hadoop cluster. The most important steps include setting up Hive configuration files, setting environment variables, and using the Hadoop Distributed File System (HDFS).

1.1 Hadoop Ecosystem

Hadoop Distributed File System (HDFS), is one of the largest Apache projects and primary storage system of Hadoop. It employs a NameNode and DataNode architecture. It is a distributed file system able to store large files running over the cluster of commodity hardware.

Apache Hadoop ecosystem refers to the various components of the Apache Hadoop software library; it includes opensource projects as well as a complete range of complementary tools. Some of the most well-known tools of the Hadoop ecosystem include HDFS, Hive, Pig, YARN, MapReduce, Spark, HBase, Oozie, Sqoop, Zookeeper, etc.

The Hadoop ecosystem architecture is made up of four main components: data storage, data processing, data access, and data management.

1. Data Storage

The first step to explaining the Hadoop ecosystem is where all your raw data is stored. It could be on a local hard drive or in the cloud.

2. Data Processing

The second phase of the Hadoop ecosystem in Big Data involves analysing your data and transforming it into something meaningful that can be used for further analysis.

3. Data Access

In this third phase of the Hadoop ecosystem, you can use tools like Hive or Pig to query your data sets and perform actions like filtering out specific rows, sorting them by certain columns or values within them such as location.

4. Data Management

Finally, the last phase of the Hadoop ecosystem architecture involves taking all the work we've done on data sets in previous phases and storing it safely somewhere so we can return to it later if needed.

1.2 Working of Map-Reduce in HDFS environment

MapReduce is a programming model or pattern within the Hadoop framework that is used to access big data stored in the Hadoop File System (HDFS). It is a core component, integral to the functioning of the Hadoop framework.

With MapReduce, rather than sending data to where the application or logic resides, the logic is executed on the server where the data already resides, to expedite processing. Data access and storage is disk-based—the input is usually stored as files containing structured, semi-structured, or unstructured data, and the output is also stored in files.

MapReduce was once the only method through which the data stored in the HDFS could be retrieved, but that is no longer the case. Today, there are other query-based systems such as Hive and Pig that are used to retrieve data from the HDFS using SQL-like statements. However, these usually run along with jobs that are written using the MapReduce model. That's because MapReduce has unique advantages.

Map Function:

During the map phase, input data is divided into smaller chunks, and a map function is applied to each chunk independently.

The map function emits a set of key-value pairs, creating an intermediate dataset.

The key-value pairs are typically chosen based on the specific requirements of the problem.

Shuffle and Sort Phase:

The output of the map phase is shuffled and sorted based on keys. All values for a particular key are grouped together.

This phase is critical for ensuring that data with the same key ends up on the same node during the reduce phase.

Reduce Function:

During the reduce phase, the sorted and shuffled data is input to a reduce function. The reduce

function processes the grouped data for each key and produces the final output.

1.3 HIVE Environment

Hive is an open-source data warehousing framework that facilitates querying, and analysis of large data sets stored in distributed storage systems such as Hadoop Distributed File System (HDFS).

It provides a SQL-like interface (called HiveQL) for querying data, making it accessible to SQL-savvy users. Here's how Hive works with big data:

Data processing: Data is first entered into the Hadoop Distributed File System (HDFS) or another supported distributed storage system. This can be structured or semi-structured data, often stored in files such as CSV, JSON, Avro, Parquet, etc.

Metadata storage: Hive maintains a metadata repository that contains information about the structure of data files, including their schemas and locations. This metadata is stored in a database (default is Apache Derby but can be configured to use other databases such as MySQL).

HiveQL Queries: Users interact with Hive through the SQL-like language HiveQL. They write queries to retrieve and process data stored in HDFS.

Query Execution: When a HiveQL query is submitted, Hive's query compiler transforms the query into a series of MapReduce or Tez jobs (depending on the execution engine) that will be executed on the Hadoop cluster.

MapReduce/Tez execution: Generated jobs are then sent to the Hadoop cluster for execution. A MapReduce or Tez framework processes data in parallel across cluster nodes.

Results Retrieval: When the job is done, Hive collects the results and returns them to the user. To pass through the Hive environment, the necessary settings and variables must be set so that Hive can communicate with the underlying Hadoop cluster.

2. Methodology

2.1 Data Set:

The dataset contains 13 attributes App, Category, Rating, Reviews, Size, Installs, Type, Price, Content Rating**,** Genres, Last Updated, Current Ver, Android Ver.

App: Name of the apps

Category: To which Category the app belongs to

Rating: Overall user rating of the app by the user.

Reviews: Number of user reviews for the apps.

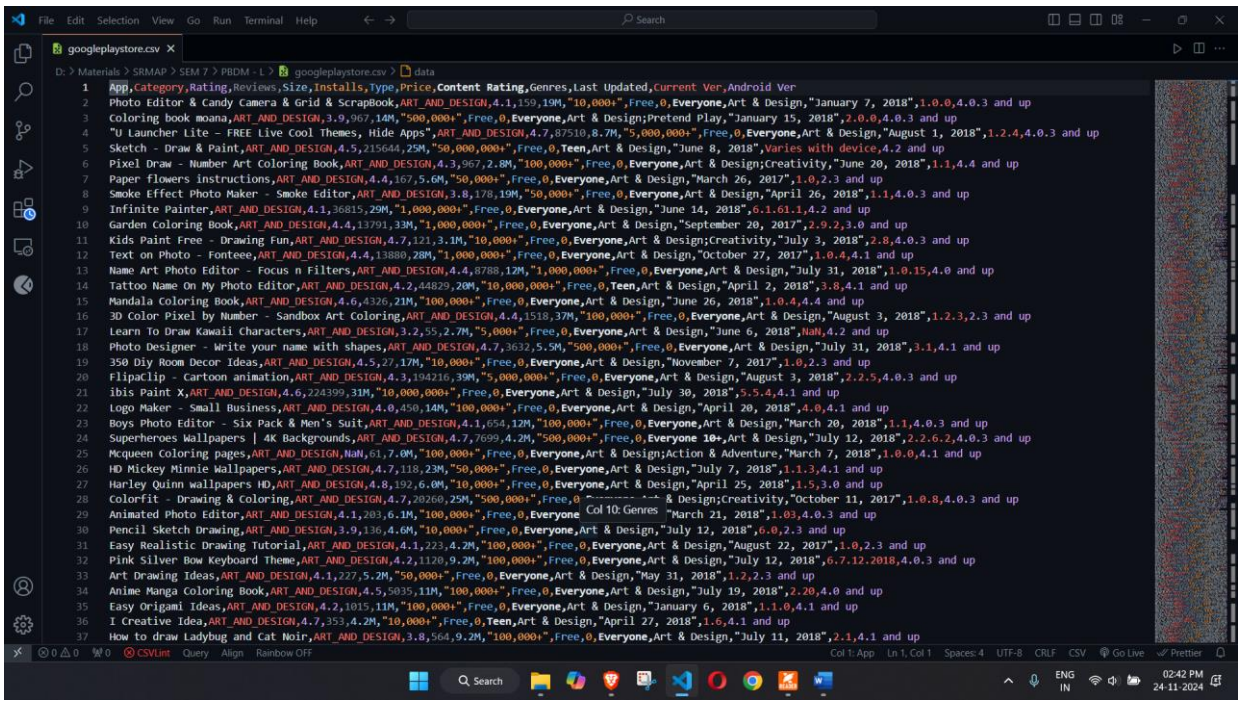
size of the app: Size of the apps **Installs:**

Number of installs of the apps **Type:** If the

app is free or paid

Price: Price of the apps

number of downloads: Number of user downloads/installs for the app



2.2 HIVE QUERIES:

The analysis is done using HIVE Queries in Hadoop Ecosystem.

1) create hive table:

Query:

```
CREATE TABLE googleplaystore (
   App STRING,
   Category STRING,
   Rating FLOAT,
   Reviews INT,
   Size STRING,
   Installs STRING,
   Type STRING,
   Price FLOAT,
   Content_Rating STRING,
   Genres STRING,
   Last_Updated STRING,
   Current_Ver STRING,
   Android_Ver STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;
```

Output:

```
hive> SELECT App, Rating, Reviews, Size, Price
    > FROM googleplaystore
    > WHERE Category = 'ART_AND_DESIGN';
OK
Photo Editor & Candy Camera & Grid & ScrapBook  4.1     159     19M     0.0
Coloring book moana     3.9     967     14M     0.0
Sketch - Draw & Paint   4.5     215644  25M     0.0
Pixel Draw - Number Art Coloring Book   4.3     967     2.8M    0.0
Paper flowers instructions      4.4     167     5.6M    0.0
Smoke Effect Photo Maker - Smoke Editor 3.8     178     19M     0.0
Infinite Painter        4.1     36815   29M     0.0
Garden Coloring Book    4.4     13791   33M     0.0
Kids Paint Free - Drawing Fun   4.7     121     3.1M    0.0
Text on Photo - Fonteee 4.4     13880   28M     0.0
Name Art Photo Editor - Focus n Filters 4.4     8788    12M     0.0
Tattoo Name On My Photo Editor  4.2     44829   20M     0.0
Mandala Coloring Book   4.6     4326    21M     0.0
3D Color Pixel by Number - Sandbox Art Coloring 4.4     1518    37M     0.0
Learn To Draw Kawaii Characters 3.2     55      2.7M    0.0
Photo Designer - Write your name with shapes    4.7     3632    5.5M    0.0
350 Diy Room Decor Ideas        4.5     27      17M     0.0
FlipaClip - Cartoon animation   4.3     194216  39M     0.0
ibis Paint X    4.6     224399  31M     0.0
Logo Maker - Small Business     4.0     450     14M     0.0
Boys Photo Editor - Six Pack & Men's Suit       4.1     654     12M     0.0
Superheroes Wallpapers | 4K Backgrounds 4.7     7699    4.2M    0.0
Mcqueen Coloring pages  0.0     61      7.0M    0.0
HD Mickey Minnie Wallpapers     4.7     118     23M     0.0
Harley Quinn wallpapers HD      4.8     192     6.0M    0.0
Colorfit - Drawing & Coloring   4.7     20260   25M     0.0
Animated Photo Editor   4.1     203     6.1M    0.0
Pencil Sketch Drawing   3.9     136     4.6M    0.0
Easy Realistic Drawing Tutorial 4.1     223     4.2M    0.0
Pink Silver Bow Keyboard Theme  4.2     1120    9.2M    0.0
Art Drawing Ideas       4.1     227     5.2M    0.0
Anime Manga Coloring Book       4.5     5035    11M     0.0
Easy Origami Ideas      4.2     1015    11M     0.0
I Creative Idea 4.7     353     4.2M    0.0
How to draw Ladybug and Cat Noir        3.8     564     9.2M    0.0
UNICORN - Color By Number & Pixel Art Coloring  4.7     8145    24M     0.0
Floor Plan Creator      4.1     36639   70      0.0
PIP Camera - PIP Collage Maker  4.7     158     11M     0.0
How To Color Disney Princess - Coloring Pages   4.0     591     9.4M    0.0
Drawing Clothes Fashion Ideas   4.2     117     15M     0.0
```

2) Load Data into Hive Table:

Query:

LOAD DATA LOCAL INPATH '/path/to/googleplaystore.csv' INTO TABLE googleplaystore;

Analysis 1:

Get All Apps in a Specific Category:

SELECT App, Rating, Reviews, Size, Price

FROM googleplaystore

WHERE Category = 'ART_AND_DESIGN';

Output:

```
BQ Scan 4.6     TOOLS   18      0.0
Software Update 4.4     TOOLS   8259    0.0
RoboPad 4.6     FAMILY  41      0.0
Bar-B-Q Rib House       5.0     FOOD_AND_DRINK  2       0.0
Gold Wallpapers 4.5     PERSONALIZATION 43      0.0
[Substratum] M5 Theme   4.4     PERSONALIZATION 16      0.0
Zowi App        4.5     FAMILY  516     0.0
Bar-B-Q Recipes 4.8     FOOD_AND_DRINK  18      0.0
System Update Free      4.3     TOOLS   3068    0.0
Camera MX - Free Photo & Video Camera   4.3     PHOTOGRAPHY     244302  0.0
Launcher Oreo 8.1       4.5     PERSONALIZATION 13466   0.0
Free antivirus and VPN  4.3     TOOLS   27749   0.0
RoboPad++       4.4     FAMILY  77      0.0
Simple Gallery  4.5     TOOLS   28030   0.0
Left vs Right: Brain Training   4.5     FAMILY  75719   0.0
Infinity Dungeon VIP    4.3     FAMILY  21804   0.99
Carros Rebaixados BR    4.3     GAME    20691   0.0
Cardboard       4.2     LIBRARIES_AND_DEMO      130287  0.0
QR Code Pro     4.8     PRODUCTIVITY    5865    4.49
Br. Parking - Busy road Parking 3D 2018 4.5     FAMILY  41      0.0
B R Telco FCU Mobile Banking    4.7     FINANCE 413     0.0
Br Browser      4.9     SOCIAL  29      0.0
TV Guide BR Gold        4.4     FAMILY  544     1.49
Dr B R Ambedkar (Jai Bhim)      4.7     SOCIAL  2068    0.0
Br Shafi        4.9     FAMILY  1288    0.0
Dr. B.R.Ambedkar        4.8     SOCIAL  1902    0.0
Texas HoldEm Poker Deluxe (BR)  4.5     GAME    1418    0.0
Companion for Fortnite & Fortnite Battle Royale 4.6     FAMILY  2736    0.0
Mu Mobile BR    4.5     BUSINESS        43      0.0
Brick Breaker BR        5.0     GAME    7       0.0
CJ - BR MEMES   4.3     COMICS  109     0.0
Vlogger Go Viral - Tuber Game   4.8     FAMILY  1304467 0.0
Next Portuguese(BR) Langpack    4.3     TOOLS   1320    0.0
BR Ambedkar Biography & Quotes  4.6     BOOKS_AND_REFERENCE     156     0.0
Black Commando | Special Ops | FPS Shooting     4.1     GAME    1167    0.0
GearBest Online Shopping        4.5     SHOPPING        245104  0.0
iPlayIT for YouTube VR Player   4.3     VIDEO_PLAYERS   5879    0.0
BSPlayer FREE   4.3     VIDEO_PLAYERS   138337  0.0
Block Strike    4.5     GAME    947515  0.0
Loved by King Bs        4.6     FAMILY  5546    0.0
BombSquad Remote        4.1     GAME    13304   0.0
BSPlayer        4.2     VIDEO_PLAYERS   4585    5.99
BS CS IT & SE   4.6     FAMILY  9       0.0
BSPlayer ARMv7 VFP CPU support  4.3     VIDEO_PLAYERS   9966    0.0
Marked by King Bs       4.6     FAMILY  70335   0.0
COMSATS BOOK STORE FOR BS(CS)   5.0     FAMILY  15      0.0
BS-Mobile       5.0     COMMUNICATION   1       0.0
BS Calendar / Patro / पा हरे    4.2     PRODUCTIVITY    218     0.0
BS Generator    4.2     FAMILY  5       0.0
```

Analysis 2:

Get Apps with Rating Greater than 4:

```
SELECT App, Rating, Category, Reviews, Price
FROM googleplaystore
WHERE Rating > 4.0;
```

Output:

```
hive> SELECT App, Rating, Category, Reviews, Price
    > FROM googleplaystore
    > WHERE Rating > 4.0;
```

Analysis 3:

Find the App with the Highest Number of Reviews:

```
WITH max_reviews_cte AS (
   SELECT MAX(Reviews) AS max_reviews
   FROM googleplaystore
)
SELECT App, Reviews AS max_reviews
FROM googleplaystore, max_reviews_cte
WHERE Reviews = max_reviews_cte.max_reviews;D
```

Output:

```
hive> WITH max_reviews_cte AS (
    >     SELECT MAX(Reviews) AS max_reviews
    >     FROM googleplaystore
    > )
    > SELECT App, Reviews AS max_reviews
    > FROM googleplaystore, max_reviews_cte
    > WHERE Reviews = max_reviews_cte.max_reviews;
Query ID = raj_ops_20241119174821_408f02e5-a764-4ed7-8676-23b763b18f21
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1732036815546_0006)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Map 3 ..........   SUCCEEDED      1          1        0        0       0       0
Reduce 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 33.49 s
--------------------------------------------------------------------------------
OK
Facebook        78158306
Time taken: 39.406 seconds, Fetched: 1 row(s)
```

Analysis 4:

Calculate the Average Rating by Content Rating:

SELECT Content_Rating, AVG(Rating) AS avg_rating

FROM googleplaystore

GROUP BY Content_Rating;

Output:

```
hive> SELECT Content_Rating, AVG(Rating) AS avg_rating
    > FROM googleplaystore
    > GROUP BY Content_Rating;
Query ID = raj_ops_20241119175157_1e01804c-b592-4092-8ef5-a19bdaef68f4
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1732036815546_0006)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1                 INITED      1          0        0        1       0       0
Reducer 2             INITED      2          0        0        2       0       0
--------------------------------------------------------------------------------
VERTICES: 00/02  [>>--------------------------] 0%    ELAPSED TIME: 4.20 s
--------------------------------------------------------------------------------
```

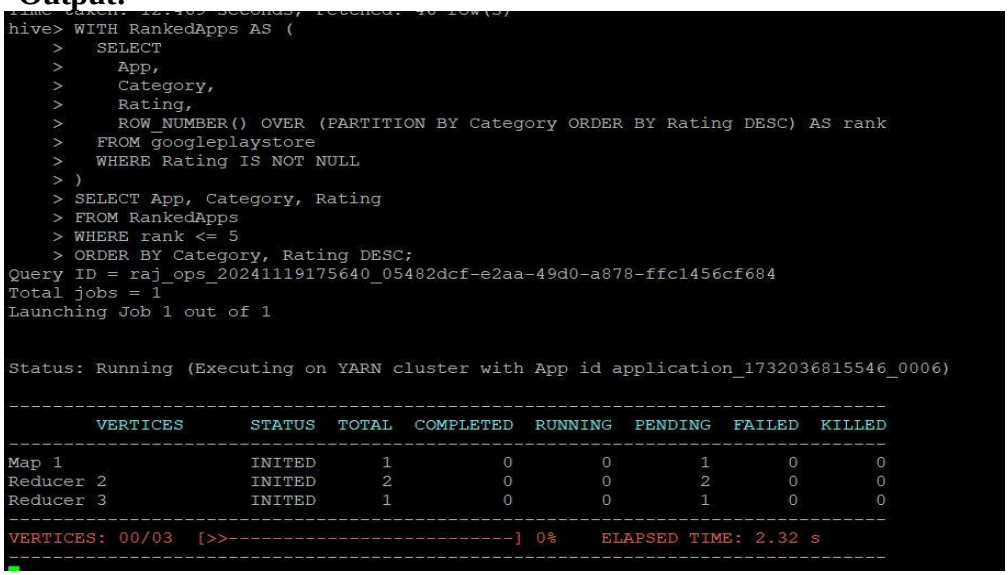
```
0.0     19.0
0.99    NULL
1.2     NULL
1.49    NULL
1.99    NULL
10000   2.0
100000  NULL
1000000 NULL
10000000        NULL
100000000       NULL
12.99   NULL
13M     NULL
17882   NULL
2.49    NULL
2.8M    NULL
2.99    NULL
3.3     NULL
3.49    NULL
3.99    NULL
32M     NULL
34      NULL
3776    NULL
37M     NULL
4.0     NULL
4.7     NULL
4.99    NULL
44M     NULL
50000   NULL
500000  NULL
5000000 NULL
50000000        NULL
500000000       NULL
60M     NULL
70      NULL
74359   NULL
76      NULL
Adults only 18+ 4.299999952316284
Content Rating  NULL
Everyone        3.5055659527964886
Everyone 10+    4.046814404696309
Free    NULL
Mature 17+      3.770680626649507
Paid    NULL
Teen    3.7525252477087156
Unrated 2.049999952316284
```

Analysis 5:

Find Top 5 Apps by Rating in Each Category:

```
WITH RankedApps AS (
  SELECT
    App,
    Category,
    Rating,
    ROW_NUMBER() OVER (PARTITION BY Category ORDER BY Rating DESC) AS rank
  FROM googleplaystore
  WHERE Rating IS NOT NULL
)
SELECT App, Category, Rating
FROM RankedApps
WHERE rank <= 5
ORDER BY Category, Rating DESC;
```

Output:



```
Meme Generator  ENTERTAINMENT   4.6
ColorFul - Adult Coloring Book  ENTERTAINMENT   4.6
My Talking Pet  ENTERTAINMENT   4.6
Motorola Spotlight Player™      ENTERTAINMENT   4.6
FAST EO EVENTS  5.0
EK Bailey Preaching Conference  EVENTS  5.0
Mindvalley U Tallinn 2018       EVENTS  5.0
EF Events       EVENTS  5.0
SUMMER SONIC app        EVENTS  5.0
Master E.K      FAMILY  5.0
Pyaar Ek Dhoka  FAMILY  5.0
Ek Bander Ne Kholi Dukan        FAMILY  5.0
Ek Qissa He Quran Se (Qurani Waqiyat)   FAMILY  5.0
Lyrics of Ek Paheli Leela       FAMILY  5.0
BK Gold App     FINANCE 5.0
DN Calculators  FINANCE 5.0
CT Checkout     FINANCE 5.0
BxPort - Bitcoin Bx (Thailand)  FINANCE 5.0
Free coupons and vouchers       FINANCE 5.0
Bar-B-Q Rib House       FOOD_AND_DRINK  5.0
Food-Aw - Order Food Online in Aruba    FOOD_AND_DRINK  5.0
SarashpazPapion (Cooking with Chef Bowls)       FOOD_AND_DRINK  4.8
Bar-B-Q Recipes FOOD_AND_DRINK  4.8
Simple Recipes  FOOD_AND_DRINK  4.7
Ra Ga Ba        GAME    5.0
Monster Ride Pro        GAME    5.0
CP Trivia       GAME    5.0
Axe Champs! Wars        GAME    5.0
211:CK  GAME    5.0
CF Townsville   HEALTH_AND_FITNESS      5.0
Bacterial vaginosis Treatment - Sexual disease  HEALTH_AND_FITNESS      5.0
MI-BP   HEALTH_AND_FITNESS      5.0
CL Strength     HEALTH_AND_FITNESS      5.0
EF Academy      HEALTH_AND_FITNESS      5.0
Ruler   HOUSE_AND_HOME  4.7
DIY On A Budget HOUSE_AND_HOME  4.7
Redfin Real Estate      HOUSE_AND_HOME  4.6
Houzz Interior Design Ideas     HOUSE_AND_HOME  4.6
Living Smart Home       HOUSE_AND_HOME  4.6
Eternal life    LIBRARIES_AND_DEMO      5.0
Nur təfsiri 1-ci cild   LIBRARIES_AND_DEMO      5.0
Girls Nancy Ajram Without Net   LIBRARIES_AND_DEMO      4.7
Blackpink as if it's your last  LIBRARIES_AND_DEMO      4.7
Call Recorder   LIBRARIES_AND_DEMO      4.7
The Divine Feminine App: the DF App     LIFESTYLE       5.0
Overcomers CF - GA      LIFESTYLE       5.0
BP Log lite     LIFESTYLE       5.0
Helping BD      LIFESTYLE       5.0
AC DC Power Monitor     LIFESTYLE       5.0
```

Analysis 6:

Find Apps with the Most Frequent Size Range (e.g., Apps under 50MB):

```
WITH SizeCategory AS (
  SELECT
    App,
    CASE
      WHEN Size LIKE '%M' THEN CAST(SUBSTRING(Size, 1, LENGTH(Size)-1) AS FLOAT) * 1024
      WHEN Size LIKE '%K' THEN CAST(SUBSTRING(Size, 1, LENGTH(Size)-1) AS FLOAT)
      ELSE NULL
    END AS size_kb
  FROM googleplaystore
  WHERE Size IS NOT NULL
)
SELECT
  CASE
    WHEN size_kb < 51200 THEN 'Under 50MB'
    WHEN size_kb BETWEEN 51200 AND 102400 THEN '50MB to 100MB'
    WHEN size_kb BETWEEN 102400 AND 204800 THEN '100MB to 200MB'
    ELSE 'Above 200MB'
  END AS size_range,
  COUNT(*) AS num_apps
FROM SizeCategory
GROUP BY
  CASE
    WHEN size_kb < 51200 THEN 'Under 50MB'
    WHEN size_kb BETWEEN 51200 AND 102400 THEN '50MB to 100MB'
    WHEN size_kb BETWEEN 102400 AND 204800 THEN '100MB to 200MB'
    ELSE 'Above 200MB'
  END
ORDER BY num_apps DESC;
```

Output:

```
hive> WITH SizeCategory AS (
    >     SELECT
    >         App,
    >         CASE
    >             WHEN Size LIKE '%M' THEN CAST(SUBSTRING(Size, 1, LENGTH(Size)-1) AS FLOAT) * 1024
    >             WHEN Size LIKE '%K' THEN CAST(SUBSTRING(Size, 1, LENGTH(Size)-1) AS FLOAT)
    >             ELSE NULL
    >         END AS size_kb
    >     FROM googleplaystore
    >     WHERE Size IS NOT NULL
    > )
    > SELECT
    >     CASE
    >         WHEN size_kb < 51200 THEN 'Under 50MB'
    >         WHEN size_kb BETWEEN 51200 AND 102400 THEN '50MB to 100MB'
    >         WHEN size_kb BETWEEN 102400 AND 204800 THEN '100MB to 200MB'
    >         ELSE 'Above 200MB'
    >     END AS size_range,
    >     COUNT(*) AS num_apps
    > FROM SizeCategory
    > GROUP BY
    >     CASE
    >         WHEN size_kb < 51200 THEN 'Under 50MB'
    >         WHEN size_kb BETWEEN 51200 AND 102400 THEN '50MB to 100MB'
    >         WHEN size_kb BETWEEN 102400 AND 204800 THEN '100MB to 200MB'
    >         ELSE 'Above 200MB'
    >     END
    > ORDER BY num_apps DESC;
Query ID = raj_ops_20241119180041_69176eb2-2739-4ab9-a4fd-7637454d5797
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1732036815546_0006)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1                RUNNING      1          0        1        0       0       0
Reducer 2             INITED      2          0        0        2       0       0
Reducer 3             INITED      1          0        0        1       0       0
--------------------------------------------------------------------------------
VERTICES: 00/03  [>>--------------------------] 0%    ELAPSED TIME: 2.98 s
--------------------------------------------------------------------------------
OK
Under 50MB      7128
Above 200MB     2194
50MB to 100MB   1037
Time taken: 8.328 seconds, Fetched: 3 row(s)
hive>
```

Analysis 7:

Calculate the Total Revenue for All Paid Apps:

```
WITH Revenue AS (
  SELECT
    App,
    CAST(REGEXP_REPLACE(Installs, ',', '') AS BIGINT) AS installs,
    CAST(REGEXP_REPLACE(Price, '\\$', '') AS FLOAT) AS price
  FROM googleplaystore
  WHERE Price LIKE '$%' -- Ensures only priced apps are considered
)
SELECT
    SUM(installs * price) AS total_revenue
FROM Revenue;
```

Output:

```
hive> WITH Revenue AS (
    >   SELECT
    >     App,
    >     CAST(REGEXP_REPLACE(Installs, ',', '') AS BIGINT) AS installs,
    >     CAST(REGEXP_REPLACE(Price, '\\$', '') AS FLOAT) AS price
    >   FROM googleplaystore
    >   WHERE Price LIKE '$%' -- Ensures only priced apps are considered
    > )
    > SELECT
    >     SUM(installs * price) AS total_revenue
    > FROM Revenue;
Query ID = raj_ops_20241119181027_5206e02b-873a-421e-8835-c52eb5c89646
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1732036815546_0008)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1                 INITED      1          0        0        1       0       0
Reducer 2             INITED      1          0        0        1       0       0
--------------------------------------------------------------------------------
VERTICES: 00/02  [>>--------------------------] 0%    ELAPSED TIME: 2.32 s
--------------------------------------------------------------------------------
```

Analysis 8:

Find the Most Popular App in Terms of Installs:

SELECT App, Installs
FROM googleplaystore
ORDER BY CAST(REGEXP_REPLACE(Installs, ',', '') AS BIGINT) DESC
LIMIT 1;

Output:

```
hive> SELECT App, Installs
    > FROM googleplaystore
    > ORDER BY CAST(REGEXP_REPLACE(Installs, ',', '') AS BIGINT) DESC
    > LIMIT 1;
Query ID = raj_ops_20241119180532_6cfd9f7c-1859-4c77-a280-063dc2d377d9
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1732036815546_0007)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1                 INITED      1          0        0        1       0       0
Reducer 2             INITED      1          0        0        1       0       0
--------------------------------------------------------------------------------
VERTICES: 00/02  [>>--------------------------] 0%    ELAPSED TIME: 2.10 s
```

```
OK
Google Play Books       1000000000
Time taken: 12.548 seconds, Fetched: 1 row(s)
hive>
```

MapReduce Analysis:

Average Rating per Category:

Mapper Class:

```java
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.Mapper;

import java.io.IOException;

public class AverageRatingMapper extends Mapper<LongWritable, Text, Text, DoubleWritable> {
    @Override
    protected void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
        String line = value.toString();
        String[] fields = line.split(",");

        if (!fields[1].isEmpty() && !fields[2].isEmpty() && !fields[2].equals("Rating")) {
            String category = fields[1];
            try {
                double rating = Double.parseDouble(fields[2]);
                context.write(new Text(category), new DoubleWritable(rating));
            } catch (NumberFormatException e) {
                // Skip invalid ratings
            }
        }
    }
}
```

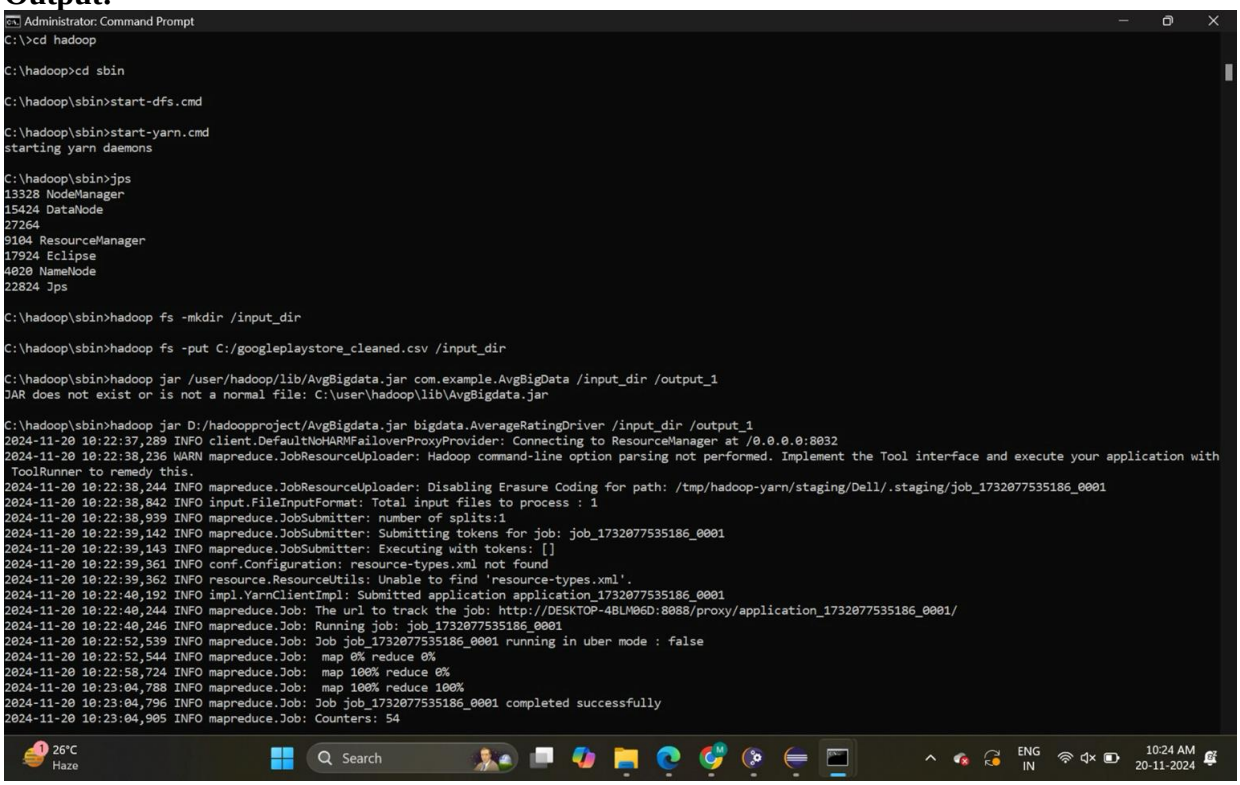
Reducer Class:

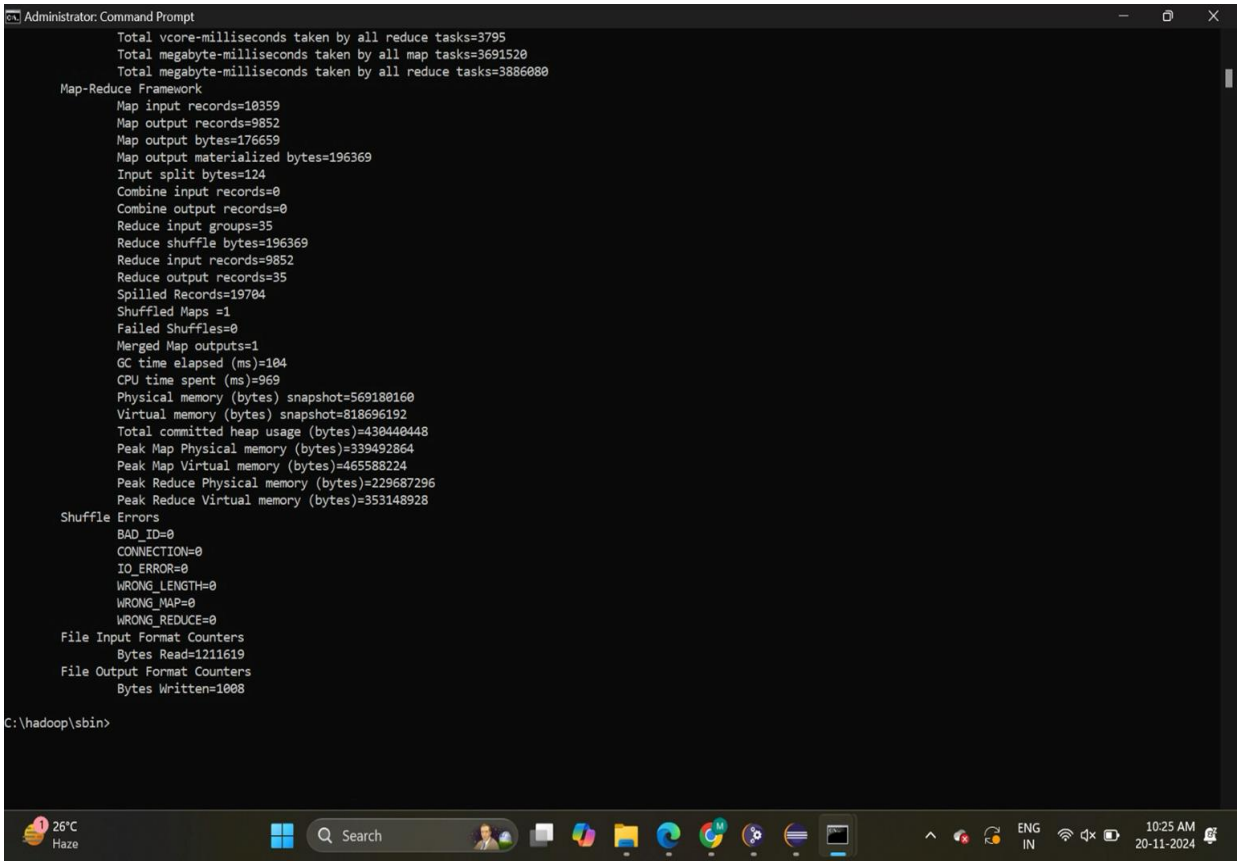
```java
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.Reducer;

import java.io.IOException;

public class AverageRatingReducer extends Reducer<Text, DoubleWritable, Text, DoubleWritable> {
    @Override
    protected void reduce(Text key, Iterable<DoubleWritable> values, Context context) throws IOException, InterruptedException {
        double sum = 0;
        int count = 0;

        for (DoubleWritable value : values) {
            sum += value.get();
            count++;
        }

        if (count > 0) {
            context.write(key, new DoubleWritable(sum / count));
        }
    }
}
```

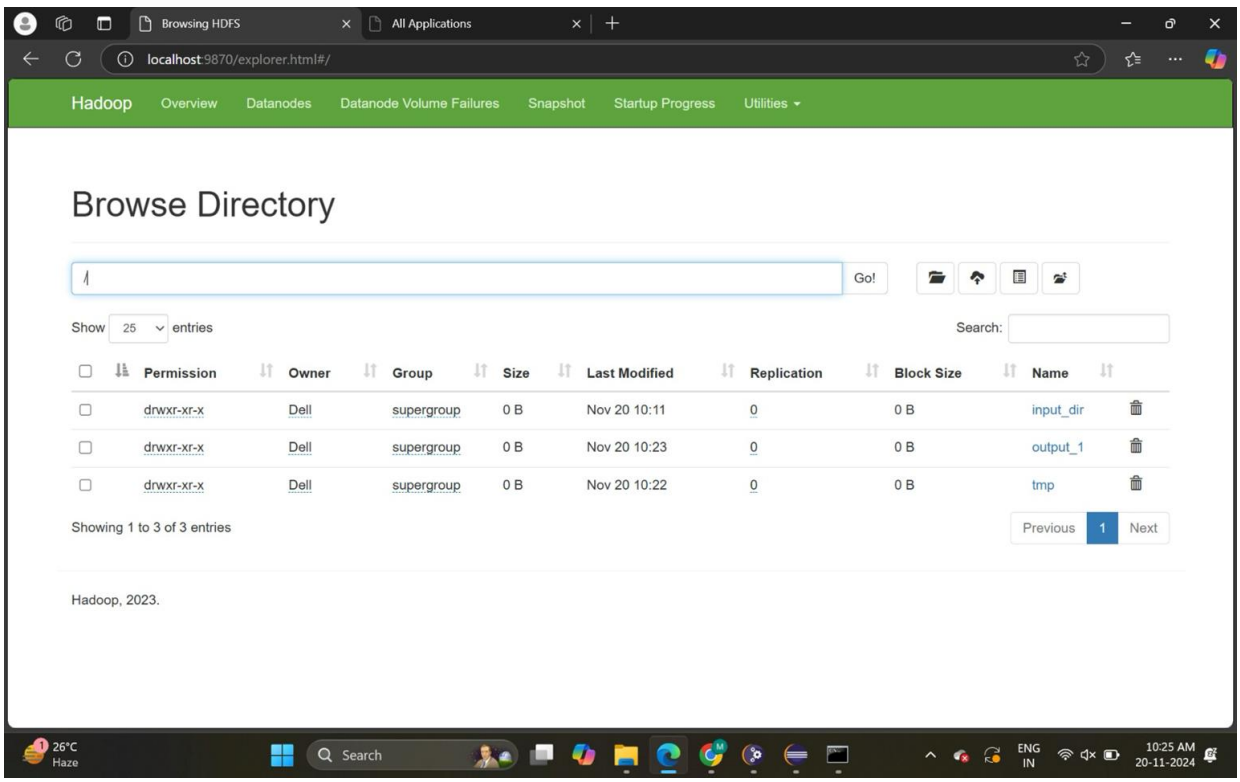
Driver Class:

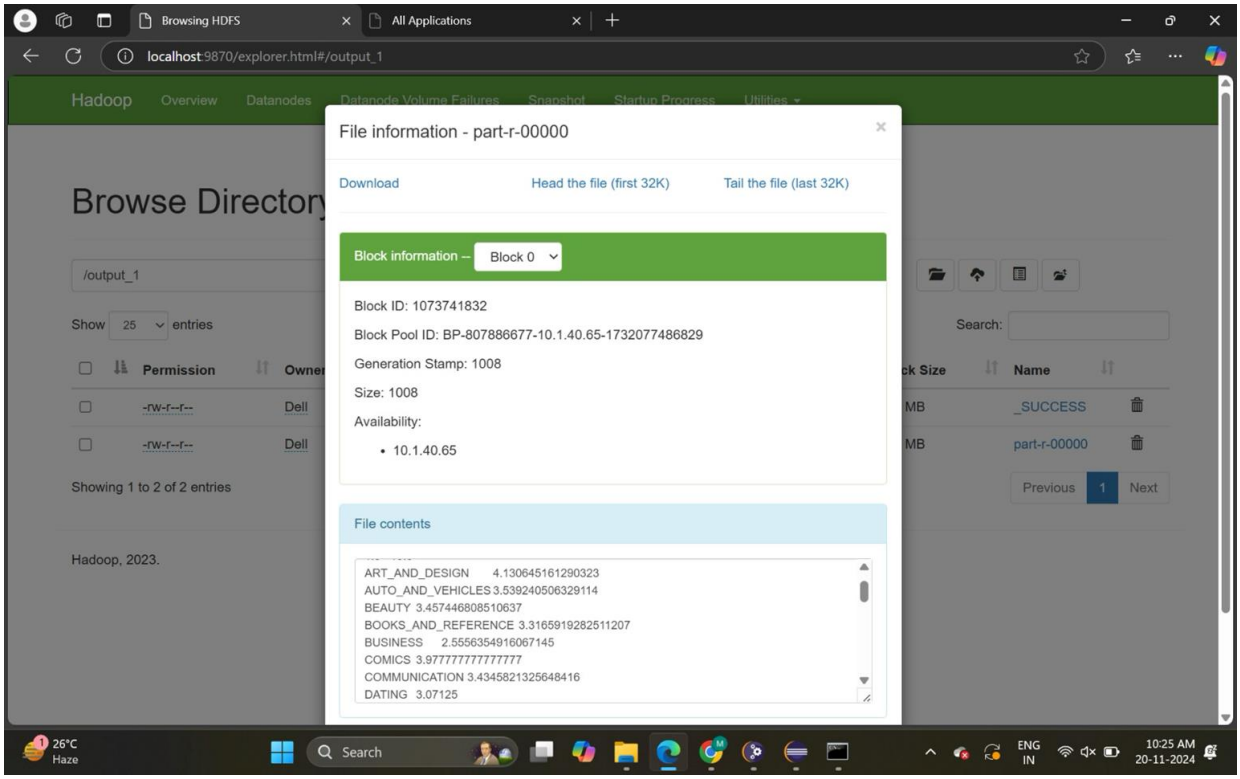
```java
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.DoubleWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class AverageRatingDriver {
    public static void main(String[] args) throws Exception {
        if (args.length != 2) {
            System.err.println("Usage: AverageRatingDriver <input path> <output path>");
            System.exit(-1);
        }

        Configuration conf = new Configuration();
        Job job = Job.getInstance(conf, "Average Rating per Category");
        job.setJarByClass(AverageRatingDriver.class);
        job.setMapperClass(AverageRatingMapper.class);
        job.setReducerClass(AverageRatingReducer.class);

        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(DoubleWritable.class);

        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));

        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }
}
```

Output:




Administrator: Command Prompt
Total vcore-milliseconds taken by all reduce tasks=3795
Total megabyte-milliseconds taken by all map tasks=3691520
Total megabyte-milliseconds taken by all reduce tasks=3886080
Map-Reduce Framework
Map input records=10359
Map output records=9852
Map output bytes=176659
Map output materialized bytes=196369
Input split bytes=124
Combine input records=0
Combine output records=0
Reduce input groups=35
Reduce shuffle bytes=196369
Reduce input records=9852
Reduce output records=35
Spilled Records=19704
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=104
CPU time spent (ms)=969
Physical memory (bytes) snapshot=569180160
Virtual memory (bytes) snapshot=818696192
Total committed heap usage (bytes)=430440448
Peak Map Physical memory (bytes)=339492864
Peak Map Virtual memory (bytes)=465588224
Peak Reduce Physical memory (bytes)=229687296
Peak Reduce Virtual memory (bytes)=353148928
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=1211619
File Output Format Counters
Bytes Written=1008
C:\hadoop\sbin>
26°C Haze
Search
ENG IN
10:25 AM 20-11-2024


Browsing HDFS
All Applications
localhost:9870/explorer.html#/
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
drwxr-xr-x Dell supergroup 0 B Nov 20 10:11 0 0 B input_dir
drwxr-xr-x Dell supergroup 0 B Nov 20 10:23 0 0 B output_1
drwxr-xr-x Dell supergroup 0 B Nov 20 10:22 0 0 B tmp
Showing 1 to 3 of 3 entries
Previous
1
Next
Hadoop, 2023.
26°C Haze
Search
ENG IN
10:25 AM 20-11-2024


Browsing HDFS
All Applications
localhost:9870/explorer.html#/output_1
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/output_1
Show 25 entries
Search:
Permission
Owner
Block Size
Name
-rw-r--r--
Dell
MB
_SUCCESS
-rw-r--r--
Dell
MB
part-r-00000
Showing 1 to 2 of 2 entries
Previous
1
Next
Hadoop, 2023.
File information - part-r-00000
Download
Head the file (first 32K)
Tail the file (last 32K)
Block information --
Block 0
Block ID: 1073741832
Block Pool ID: BP-807886677-10.1.40.65-1732077486829
Generation Stamp: 1008
Size: 1008
Availability:
10.1.40.65
File contents
ART_AND_DESIGN 4.130645161290323
AUTO_AND_VEHICLES 3.539240506329114
BEAUTY 3.457446808510637
BOOKS_AND_REFERENCE 3.3165919282511207
BUSINESS 2.5556354916067145
COMICS 3.977777777777777
COMMUNICATION 3.4345821325648416
DATING 3.07125
26°C
Haze
Search
ENG
IN
10:25 AM
20-11-2024

3. Concluding Remarks

- 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.

4. Future Work

For our future work on the Exploring Play Store application Trends using Hive, we aim to delve into more advanced analytics, including the exploration of machine learning models for predictive insights into app success. Additionally, we plan to conduct sentiment analysis on user reviews, implement time series analysis to identify trends, and integrate geospatial data for a more nuanced understanding of regional preferences. Feature engineering techniques will be experimented with to enhance model performance, and the optimization of Hive queries is on our agenda to improve overall efficiency. The development of interactive visualization dashboards will aid in presenting our findings, and we'll explore additional user engagement metrics to gain a comprehensive understanding of app performance. Cross-platform analysis, benchmarking against industry standards, and thorough documentation of methodologies are key components of our future strategy, and we'll actively seek feedback from peers and domain experts to refine and enhance our analytical approach.

5. References

[1] Amit Chile, Dr. P. R. Gundalwar.(2019). Anal-ysis of Google Play Store Application.[online]http://ijraset.com/fileserve.php?FID=24134

[2] Kaggle.com.(2018). Google Play Store Apps.[online]https://www.kaggle.com/lava18/google-play-store-apps [Accessed 3 Mar. 2020].

[3] Google's original MapReduce paper: "MapReduce: Simplified Data Processing on Large Clusters" by Jeffrey Dean and Sanjay Ghemawat.

[4] "Programming Hive" by Edward Capriolo, Dean Wampler, and Jason Rutherglen.

[5] Google play store: number of apps2018(2018). [online] https://www.statista.com/statistics/266210/number-of-available-applications-in-the- google-play-store/

[6] https://www.tutorialspoint.com/hive/index.htm