2. Data Storage and Preprocessing:

- Store the dataset in **HDFS** and/or **MongoDB**, depending on its structure.
- Perform preprocessing tasks using **Pig** for raw data cleaning and aggregation.

  sudo apt-get update, hdfs namenode –format, start-dfs.sh, start-yarn.sh, jps -> First we execute these commands in order to set up, initialize, and verify the operation of HDFS and Yarn cluster.

Switch to PowerShell Restart Manage files New session Editor Web preview Settings

```
101557069@georgebrown.ca@myVirtual:~$ sudo apt-get update
Hit:1 http://azure.archive.ubuntu.com/ubuntu jammy InRelease
Get:2 http://azure.archive.ubuntu.com/ubuntu jammy-updates InRelease [128 kB]
Hit:3 http://azure.archive.ubuntu.com/ubuntu jammy-backports InRelease
Get:4 http://azure.archive.ubuntu.com/ubuntu jammy-security InRelease [129 kB]
Hit:5 https://packages.microsoft.com/repos/microsoft-ubuntu-jammy-prod jammy InRelease
Hit:6 https://dl.google.com/linux/chrome/deb stable InRelease
Get:7 http://azure.archive.ubuntu.com/ubuntu jammy-updates/main amd64 Packages [2179 kB]
Get:8 http://azure.archive.ubuntu.com/ubuntu jammy-updates/universe amd64 Packages [1177 kB]
Fetched 3614 kB in 1s (3721 kB/s)
Reading package lists... Done
101557069@georgebrown.ca@myVirtual:~$ hdfs namenode -format
2024-11-29 03:22:25,499 INFO namenode.NameNode: STARTUP_MSG:
```

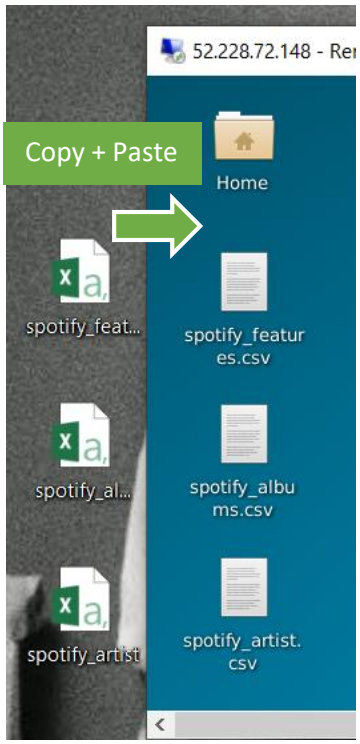
```
2024-11-29 03:22:27,477 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at myvirtual.internal.cloudapp.net/10.1.1.4
************************************************************/
101557069@georgebrown.ca@myVirtual:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [myVirtual]
101557069@georgebrown.ca@myVirtual:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
101557069@georgebrown.ca@myVirtual:~$ jps
1873 NameNode
2007 DataNode
2410 ResourceManager
2540 NodeManager
2221 SecondaryNameNode
2861 Jps
```

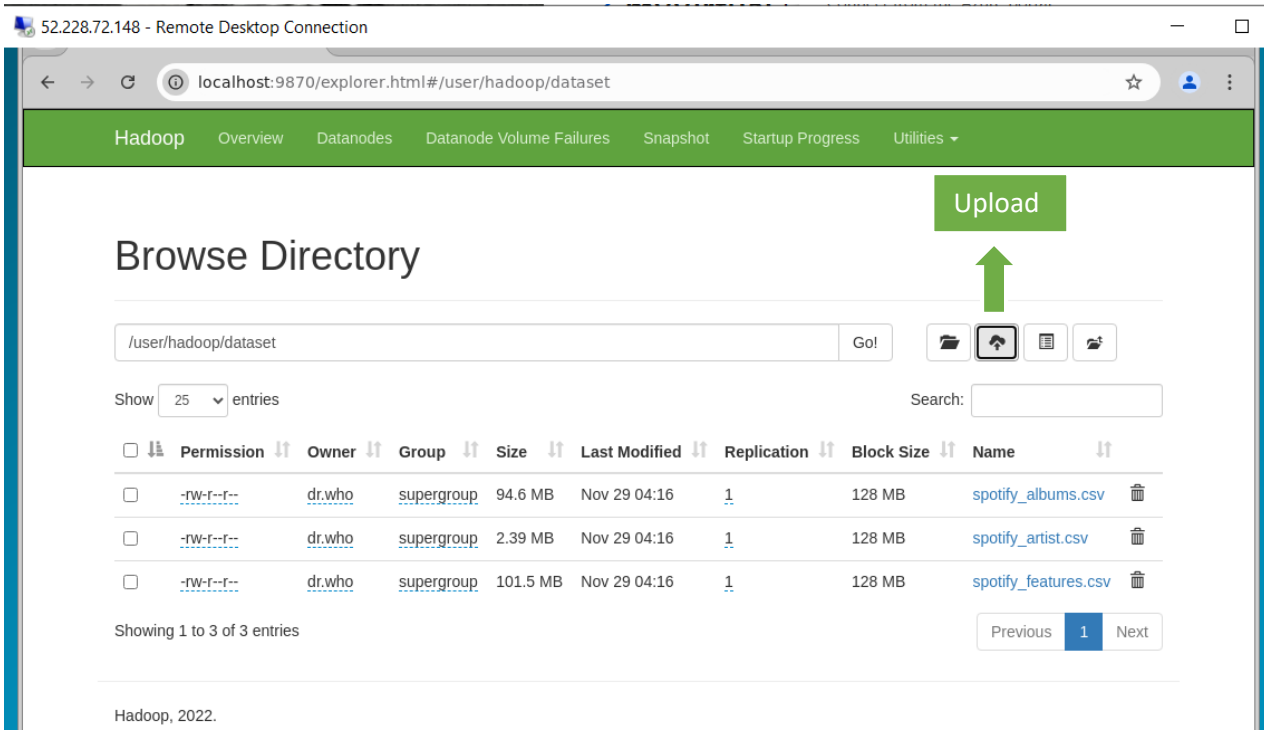
hdfs dfs –mkdir –p /user/hadoop/dataset/ , hdfs dfs –chmod 777 /user/hadoop/dataset/ -> Then we create and manage the directory and the permission within HDFS. This is

```
101557069@georgebrown.ca@myVirtual:~$ hdfs dfs -mkdir -p /user/hadoop/dataset/
101557069@georgebrown.ca@myVirtual:~$ hdfs dfs -chmod 777 /user/hadoop/dataset/
```

the path where we are going to store our .csv files.

Remote Desktop Connection -> Now we log in into our Remote Desktop in order to copy our .csv datasets from our local storage to our VM by simply ***Copy + Paste***. Then we open the browser from our VM and go to ***localhost:9870*** so we can upload our .csv files in our HDFS.





```
101557069@georgebrown.ca@myVirtual:~$ nano artist_preprocessing.pig
101557069@georgebrown.ca@myVirtual:~$ pig artist_preprocessing.pig
```

nano artist_preprocessing.pig , pig artist_preprocessing.pig -> we can close the Remote Desktop now that the work is done and return back tu SSH CLI. In order to preprocess the first file spotify_artist.csv, we create a pig file with nano command. Inside this file, firstly we load the respective dataset, replace null values with defaults because most of our genres (like genre_2) or artists (like artists_1) are missing but it would not be fair to delete them because more than half of our dataset would be erased. Then we group by artist name and calculate total followers. Cleaned data file named spotify_cleaned_artist and aggregated followers data named artist_followers_total are both saved.

```
Switch to PowerShell   Restart   Manage files   New session   Editor   ...

  GNU nano 6.2                    artist_preprocess.pig *
-- Load the dataset with schema
spotify_artist = LOAD '/user/hadoop/dataset/spotify_artist.csv'
    USING PigStorage(',')
    AS (
        id:chararray,
        name:chararray,
        artist_popularity:int,
        artist_genres:chararray,
        followers:int,
        genre_0:chararray,
        genre_1:chararray,
        genre_2:chararray,
        genre_3:chararray,
        genre_4:chararray,
        genre_5:float,
        genre_6:float
    );

-- Replace null values with defaults
cleaned_artist = FOREACH spotify_artist GENERATE
    id,
    name,
    artist_popularity,
    (artist_genres IS NULL ? '[]' : artist_genres) AS artist_genres,
    followers,
    (genre_0 IS NULL ? '' : genre_0) AS genre_0,
    (genre_1 IS NULL ? '' : genre_1) AS genre_1,
    (genre_2 IS NULL ? '' : genre_2) AS genre_2,
    (genre_3 IS NULL ? '' : genre_3) AS genre_3,

^G Help      ^O Write Out   ^W Where Is   ^K Cut       ^T Execute   ^C Location
^X Exit      ^R Read File   ^\ Replace    ^U Paste     ^J Justify   ^/ Go To Line
```

NOTE: All of preprocessing pig files, as well as cleaned data & aggregated data files are provided in the zip file for this project.

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
3.3.3   0.16.0  101557069@georgebrown.ca        2024-11-29 05:34:38     2024-11-29 05:34:45 G
ROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxRe
duceTime        MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outpu
ts
job_local286145871_0001 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a n
/a      artist_followers_total,spotify_artist,total_followers   MULTI_QUERY,COMBINER    /user
/hadoop/processed/spotify_cleaned_artist,/user/hadoop/processed/artist_followers_total,

Input(s):
Successfully read 37013 records (15767438 bytes) from: "/user/hadoop/dataset/spotify_artist.c
sv"

Output(s):
Successfully stored 37013 records (2360256 bytes) in: "/user/hadoop/processed/spotify_cleaned
_artist"
Successfully stored 36195 records (581923 bytes) in: "/user/hadoop/processed/artist_followers
_total"

Counters:
Total records written : 73208
Total bytes written : 2942179
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local286145871_0001
```

nano albums_preprocessing.pig , pig albums_preprocessing.pig, nano features_preprocessing.pig, pig features_preprocessing.pig -> the same steps and processes are repeated for spotify_albums.csv and spotify_features.csv files. Data cleaning and aggregation are repeated in similar way, saving their respective files in their respective paths.

NOTE: All of preprocessing pig files, as well as cleaned data & aggregated data files are provided in the zip file for this project.

```
101557069@georgebrown.ca@myVirtual:~$ nano albums_preprocessing.pig
101557069@georgebrown.ca@myVirtual:~$ pig albums_preprocessing.pig
```

Switch to PowerShell Restart Manage files New session Editor

```
  GNU nano 6.2                    albums_preprocess.pig *
-- Load the dataset with schema
spotify_albums = LOAD '/user/hadoop/dataset/spotify_albums.csv'
    USING PigStorage(',')
    AS (
        track_name:chararray,
        track_id:chararray,
        track_number:int,
        duration_ms:int,
        album_type:chararray,
        artists:float,
        total_tracks:int,
        album_name:chararray,
        release_date:chararray,
        label:chararray,
        album_popularity:int,
        album_id:chararray,
        artist_id:chararray,
        artist_0:chararray,
        artist_1:chararray,
        artist_2:chararray,
        artist_3:chararray,
        artist_4:chararray,
        artist_5:chararray,
        artist_6:chararray,
        artist_7:float,
        artist_8:float,
        artist_9:float,
        artist_10:float,
        artist_11:float,

^G Help      ^O Write Out  ^W Where Is   ^K Cut       ^T Execute    ^C Location
^X Exit      ^R Read File  ^\ Replace    ^U Paste     ^J Justify    ^/ Go To Line
```

Switch to PowerShell Restart Manage files New session Editor

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
3.3.3   0.16.0  101557069@georgebrown.ca        2024-11-29 05:39:40     2024-11-29 05:39:58 G
ROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxRe
duceTime        MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outpu
ts
job_local454111913_0001 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a n
/a      album_duration_avg,album_type_count,avg_duration,spotify_albums,type_count      MULTI
_QUERY,COMBINER /user/hadoop/processed/spotify_cleaned_albums,/user/hadoop/processed/album_du
ration_avg,/user/hadoop/processed/album_type_count,

Input(s):
Successfully read 438974 records (209153178 bytes) from: "/user/hadoop/dataset/spotify_albums
.csv"

Output(s):
Successfully stored 438974 records (96453108 bytes) in: "/user/hadoop/processed/spotify_clean
ed_albums"
Successfully stored 75318 records (3020490 bytes) in: "/user/hadoop/processed/album_duration_
avg"
Successfully stored 16078 records (199003 bytes) in: "/user/hadoop/processed/album_type_count
"

Counters:
Total records written : 530370
Total bytes written : 99672601
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
```

```
101557069@georgebrown.ca@myVirtual:~$ nano features_preprocessing.pig
101557069@georgebrown.ca@myVirtual:~$ pig features_preprocessing.pig
```

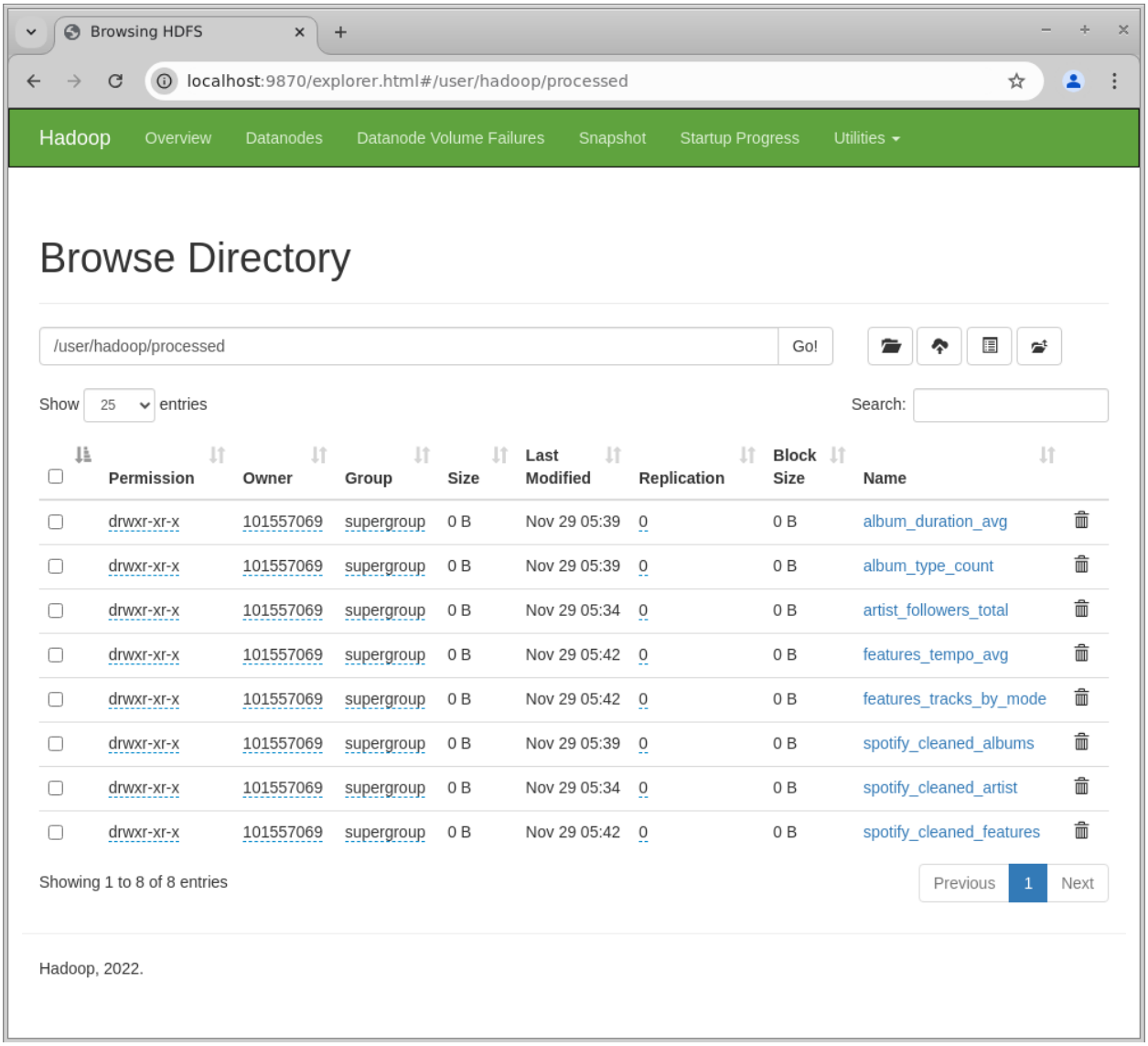
Switch to PowerShell Restart Manage files New session Editor

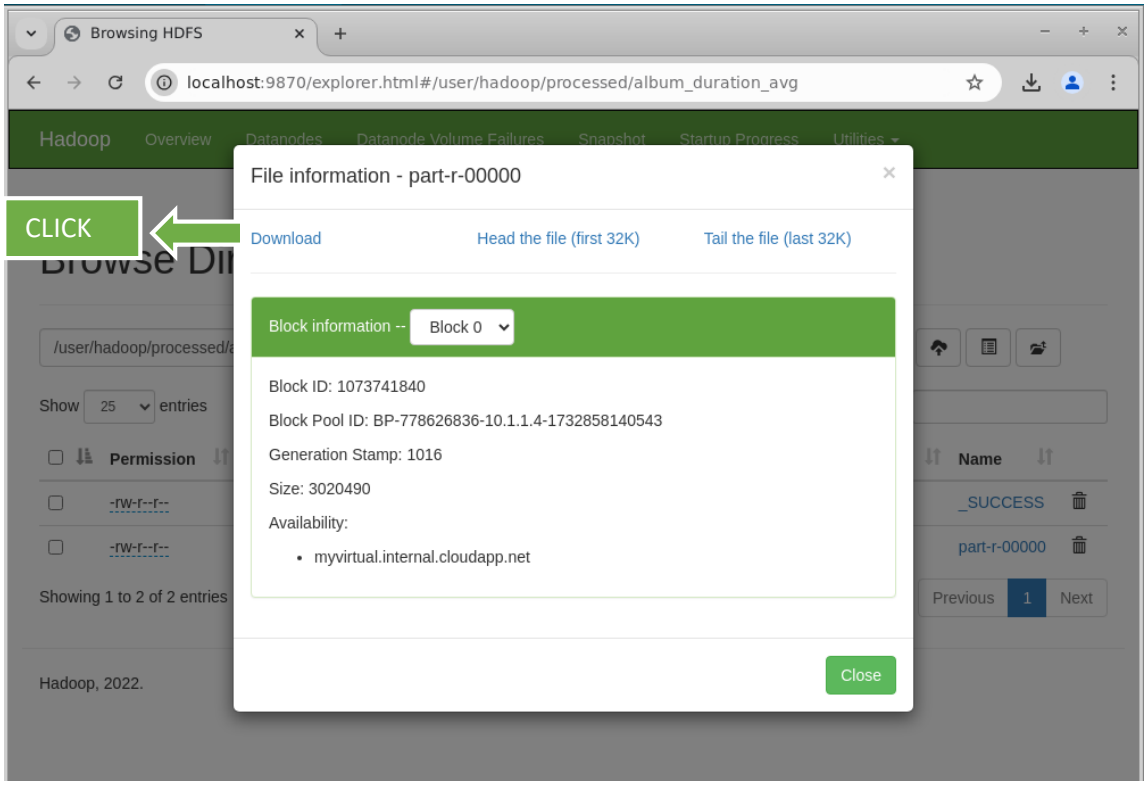
```
  GNU nano 6.2                    features_preprocess.pig *
-- Load the dataset with schema
spotify_features = LOAD '/user/hadoop/dataset/spotify_features.csv'
    USING PigStorage(',')
    AS (
        danceability:float,
        energy:float,
        key:int,
        loudness:float,
        mode:int,
        speechiness:float,
        acousticness:float,
        instrumentalness:float,
        liveness:float,
        valence:float,
        tempo:float,
        type:chararray,
        id:chararray,
        uri:chararray,
        track_href:chararray,
        analysis_url:chararray,
        duration_ms:int,
        time_signature:int
    );

-- Replace null values with defaults
cleaned_features = FOREACH spotify_features GENERATE
    (danceability IS NULL ? 0.0 : danceability) AS danceability,
    (energy IS NULL ? 0.0 : energy) AS energy,
    (key IS NULL ? -1 : key) AS key,  -- -1 indicates no key information

^G Help      ^O Write Out  ^W Where Is   ^K Cut       ^T Execute    ^C Location
^X Exit      ^R Read File  ^\ Replace    ^U Paste     ^J Justify    ^/ Go To Line
```

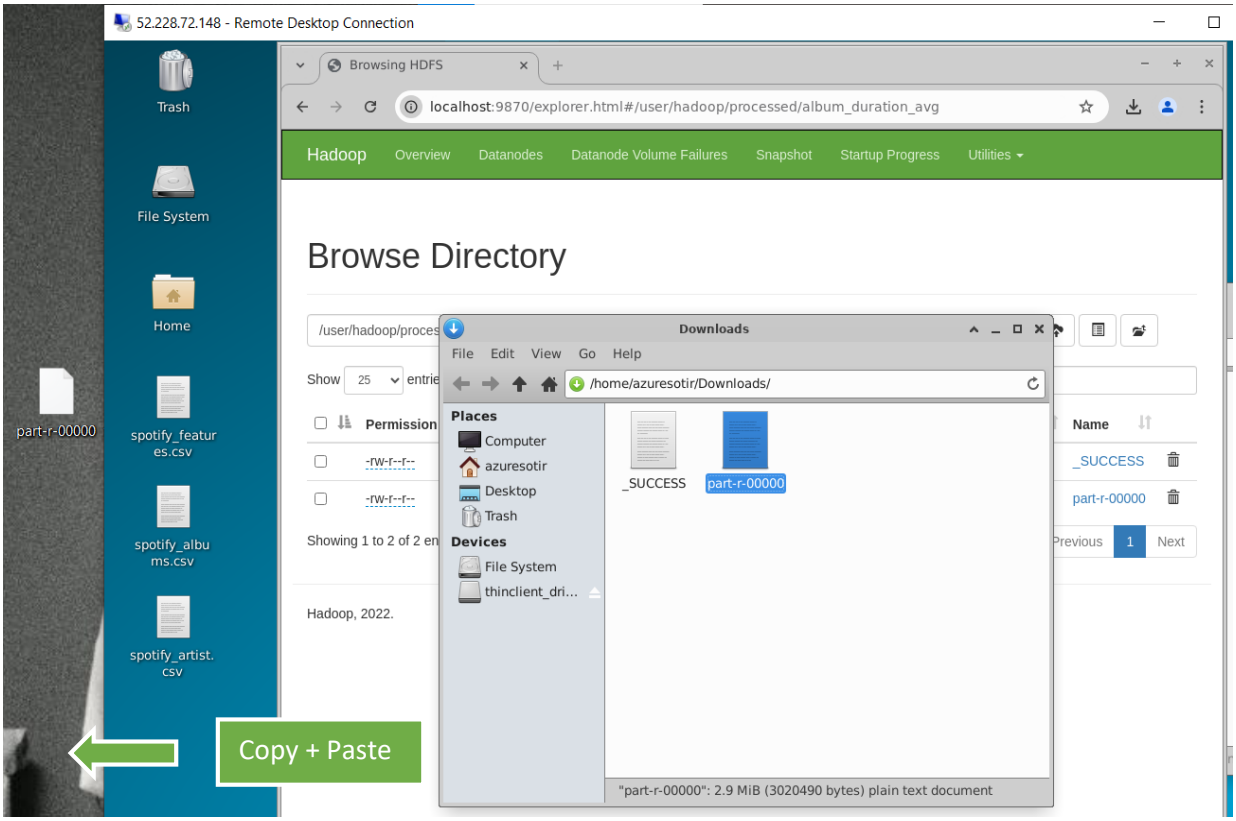
Switch to PowerShell Restart Manage files New session Editor

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
3.3.3   0.16.0  101557069@georgebrown.ca        2024-11-29 05:42:11     2024-11-29 05:42:23 G
ROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxRe
duceTime        MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outpu
ts
job_local519140208_0001 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a n
/a      avg_tempo,key_tempo_avg,mode_count,spotify_features,total_tracks_by_mode        MULTI
_QUERY,COMBINER /user/hadoop/processed/spotify_cleaned_features,/user/hadoop/processed/featur
es_tempo_avg,/user/hadoop/processed/features_tracks_by_mode,

Input(s):
Successfully read 397539 records (223617342 bytes) from: "/user/hadoop/dataset/spotify_featur
es.csv"

Output(s):
Successfully stored 397539 records (106558200 bytes) in: "/user/hadoop/processed/spotify_clea
ned_features"
Successfully stored 13 records (254 bytes) in: "/user/hadoop/processed/features_tempo_avg"
Successfully stored 3 records (21 bytes) in: "/user/hadoop/processed/features_tracks_by_mode"

Counters:
Total records written : 397555
Total bytes written : 106558475
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local519140208_0001
```

Remote Desktop Connection -> if we open our RDC again, we will see in our HDFS path that /user/hadoop/processed/ directory is created and generated files are located there. Now we manually open each generated file starting from the first one that is album_duration_avg to to the last one. When we click in each path, we see that there are 2 files: _SUCCESS and part-r-00000. The last file is the one we are looking for with the updated data. We manually download each one to our VM, and then we can Copy + Paste to our local machine.







NOTE: All of preprocessing pig files, as well as cleaned data & aggregated data files are provided in the zip file for this project.

3. Data Analysis:

- Use **Hive** to analyze the structured parts of the dataset.
- Write at least three queries to answer specific question based on the datset.

spotify_artist -> firstly we create the table for ***spotify_artist*** in Hive by executing the first query, then we start analyzing. The first query is the answer for the question: Who are the top 5 artists by popularity? ***Q1: Top 5 artists by popularity.***

```
Switch to PowerShell   Restart   Manage files   New session   Editor   ...
impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-reload4j-1.
7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = 58b233f8-c6e9-4298-9748-6562b25e14ec

Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-common-3.1.2.jar
!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider
using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Hive Session ID = 83675f22-7c10-405f-8627-448206468bab
hive> CREATE EXTERNAL TABLE spotify_artist (
    >     id STRING,
    >     name STRING,
    >     artist_popularity INT,
    >     artist_genres STRING,
    >     followers INT,
    >     genre_0 STRING,
    >     genre_1 STRING,
    >     genre_2 STRING,
    >     genre_3 STRING,
    >     genre_4 STRING,
    >     genre_5 FLOAT,
    >     genre_6 FLOAT
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE
    > LOCATION '/user/hadoop/dataset/';
OK
Time taken: 1.283 seconds
hive>
```

```
Time taken: 1.283 seconds
hive> SELECT name, artist_popularity
    > FROM spotify_artist
    > ORDER BY artist_popularity DESC
    > LIMIT 5;
Query ID = 101557069@georgebrown.ca_20241129121538_5501284f-263d-4da3-bd1e-27be83fa0bb1
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
```

```
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 416268716 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
 Sept. 18       1970
2zCruKarW8FMGbK77z2yh6  110
Taylor Swift    100
Drake   95
Bad Bunny       95
Time taken: 8.009 seconds, Fetched: 5 row(s)
hive>
```

Q2: Top 5 genres by followers

```
hive> SELECT genre_0, SUM(followers) AS total_followers
    > FROM spotify_artist
    > GROUP BY genre_0
    > ORDER BY total_followers DESC
    > LIMIT 5;
Query ID = 101557069@georgebrown.ca_20241129122004_46b882c1-9340-4e2b-90a2-9a782e1a907e
Total jobs = 2
```

```
2024-11-29 12:20:10,228 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local1420072672_0005
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 1665074864 HDFS Write: 0 SUCCESS
Stage-Stage-2:  HDFS Read: 1665074864 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
album   2943218294
pop     440522416
single  424219412
compilation     357493675
        62801471
Time taken: 6.086 seconds, Fetched: 5 row(s)
```

Q3: Top 5 most followed artists

```
hive> SELECT name AS artist_name, followers
    > FROM spotify_artist
    > ORDER BY followers DESC
    > LIMIT 5;
Query ID = 101557069@georgebrown.ca_20241129122212_4e25db54-ca6d-460c-860b-72c000b4dcb7
Total jobs = 1
Launching Job 1 out of 1
```

```
2024-11-29 12:22:16,797 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local819071062_0008
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 2497612296 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Taylor Swift    95859165
Ariana Grande   95710972
Maroon 5        41996696
A.R. Rahman     41980507
Alan Walker     39960695
Time taken: 4.746 seconds, Fetched: 5 row(s)
```

spotify_albums Hive table query

```
Switch to PowerShell   Restart   Manage files

hive> CREATE EXTERNAL TABLE spotify_albums (
    >     track_name STRING,
    >     track_id STRING,
    >     track_number INT,
    >     duration_ms INT,
    >     album_type STRING,
    >     artists FLOAT,
    >     total_tracks INT,
    >     album_name STRING,
    >     release_date STRING,
    >     label STRING,
    >     album_popularity INT,
    >     album_id STRING,
    >     artist_id STRING,
    >     artist_0 STRING,
    >     artist_1 STRING,
    >     artist_2 STRING,
    >     artist_3 STRING,
    >     artist_4 STRING,
    >     artist_5 STRING,
    >     artist_6 STRING,
    >     artist_7 FLOAT,
    >     artist_8 FLOAT,
    >     artist_9 FLOAT,
    >     artist_10 FLOAT,
    >     artist_11 FLOAT,
    >     duration_sec FLOAT
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE
    > LOCATION '/user/hadoop/dataset/';
```

Q1: Top 5 albums by popularity

```
hive> SELECT album_name, album_popularity
    > FROM spotify_albums
    > ORDER BY album_popularity DESC
    > LIMIT 5;
Query ID = 101557069@georgebrown.ca_20241129123030_3d5f9226-ea2e-4ccc-8fdf-bc42af842ff8
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2024-11-29 12:30:31,634 Stage-1 map = 0%,  reduce = 0%
2024-11-29 12:30:32,639 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local92694845_0009
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 2913881012 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
 'psychedelic rock'     19973210
 'rock' 10213575
 'rap'  10032207
 'pop emo'      8260521
 'post-grunge'  6954513
Time taken: 3.619 seconds, Fetched: 5 row(s)
```

Q2: Albums with the longest total duration

```
Switch to PowerShell   Restart   Manage files   New session   Editor   ...
hive> SELECT album_name,
    >        SUM(duration_sec) AS total_album_duration
    > FROM spotify_albums
    > GROUP BY album_name
    > ORDER BY total_album_duration DESC
    > LIMIT 5;
Query ID = 101557069@georgebrown.ca_20241129124003_afc73aed-4d6b-4dce-b159-6a7c5ed40d91
Total jobs = 2
Launching Job 1 out of 2
```

```
2024-11-29 12:40:08,410 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local502920779_0027
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 6660299456 HDFS Write: 0 SUCCESS
Stage-Stage-2:  HDFS Read: 6660299456 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Nirvana 206432.23783111572
Exodus  126574.91034889221
Future  121120.12978363037
Action  114175.48690795898
Greatest Hits   105571.85218048096
Time taken: 4.671 seconds, Fetched: 5 row(s)
```

Q3: Albums with the most tracks

```
hive> SELECT album_name,
    >        total_tracks
    > FROM spotify_albums
    > ORDER BY total_tracks DESC
    > LIMIT 10;
Query ID = 101557069@georgebrown.ca_20241129131747_fc02741d-ed04-488d-81f7-b68e36c027fe
Total jobs = 1
Launching Job 1 out of 1
```

```
2024-11-29 13:17:49,957 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1343286478_0011
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 4578955876 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
pop     115998928
detroit hip hop 79891173
reggaeton       77931484
canadian contemporary r&b       75945958
k-pop   71720409
barbadian pop   59152035
modern rock     51609418
classic rock    49194709
k-pop   47572797
canadian pop    43004737
Time taken: 2.575 seconds, Fetched: 10 row(s)
```

```
hive> show tables;
OK
orders
spotify_albums
spotify_artist
spotify_features
Time taken: 0.027 seconds, Fetched: 4 row(s)
```

spotify_features: Hive table query

```
hive> CREATE EXTERNAL TABLE spotify_features (
    >     danceability FLOAT,
    >     energy FLOAT,
    >     key INT,
    >     loudness FLOAT,
    >     mode INT,
    >     speechiness FLOAT,
    >     acousticness FLOAT,
    >     instrumentalness FLOAT,
    >     liveness FLOAT,
    >     valence FLOAT,
    >     tempo FLOAT,
    >     type STRING,
    >     id STRING,
    >     uri STRING,
    >     track_href STRING,
    >     analysis_url STRING,
    >     duration_ms INT,
    >     time_signature INT
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE
    > LOCATION '/user/hadoop/dataset/';
OK
Time taken: 0.423 seconds
```

Q1: Average tempo by key

```
hive> SELECT key, AVG(tempo) AS avg_tempo
    > FROM spotify_features
    > GROUP BY key
    > ORDER BY avg_tempo DESC;
Query ID = 101557069@georgebrown.ca_20241129132205_b38b4ecb-c2cc-4221-93b1-21779f9de6d9
Total jobs = 2
Launching Job 1 out of 2
```

```
2024-11-29 13:22:10,812 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local848287903_0013
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 4995224592 HDFS Write: 0 SUCCESS
Stage-Stage-2:  HDFS Read: 4995224592 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
77      1.997321E7
79      1.0032207E7
76      8260521.0
75      7708194.0
80      6244077.0
73      5653021.5
```

```
89      NULL
90      NULL
93      NULL
95      NULL
100     NULL
1970    NULL
Time taken: 5.211 seconds, Fetched: 97 row(s)
```

Q2: Tracks with the highest valence (positivity)

```
hive> SELECT id AS track_id,
    >        valence,
    >        track_href
    > FROM spotify_features
    > ORDER BY valence DESC
    > LIMIT 5;
Query ID = 101557069@georgebrown.ca_20241129141344_14d86d19-3d72-48c7-8104-9f99deac30e8
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2024-11-29 14:13:49,280 Stage-1 map = 0%,  reduce = 0%
2024-11-29 14:13:54,360 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1171698870_0001
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 416268716 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
philly rap      1.6387594E7     rap
rap latina      1.4374317E7     trap latino
hip hop 1.3190905E7     queens hip hop
modern rock     1.1500714E7     post-grunge
pop rap 1.1268491E7     southern hip hop
Time taken: 10.043 seconds, Fetched: 5 row(s)
```

Q3: Top 5 Tracks with the Loudest Audio Levels

```
hive> SELECT id AS track_id,
    >        loudness,
    >        track_href
    > FROM spotify_features
    > ORDER BY loudness DESC
    > LIMIT 5;
Query ID = 101557069@georgebrown.ca_20241129154939_2a0fa384-1304-45af-a02e-b489dd65b4fa
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2024-11-29 15:49:44,104 Stage-1 map = 0%,  reduce = 0%
2024-11-29 15:49:47,268 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1633156750_0001
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 416268716 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
1WZarnZpWEv7dDtjAETt4X  1.704E7
2IN85MIUl7mjSIL6dZaxXy  1.0400056E7
6paBa4y35zhmWOv0RZnzyO  7295500.0
0LyfQWJT6nXafLPZqxe9Of  7162263.0       Workout Electronica
14mErTJ0ubFVjx2zBAwjkE  5407669.0
Time taken: 8.212 seconds, Fetched: 5 row(s)
```