

Analyzing TikTok Music Trends with Hadoop Cloud and Tableau

Kenny Duong, Karina Gonzalez, Erick Gomez, Javier Machuca

Introduction

In this tutorial, we will learn how to import files onto github, create multiple directories using HDFS, and list the files on HDFS to ensure that they were created correctly. Furthermore, we'll learn how to import data onto the created directories. Then we will use Beeline to create tables for all directories. Additionally we will clean the data by getting rid of all null values. Finally we will show the tables to ensure that the data is correct. The last step will be to download the data on our personal PC and visualize it on Tableau.

List of technologies in this tutorial

- Hadoop Cloud Big Data
- Tableau

Pre-requisites

- Hadoop Big Data Account
- Download and install Tableau

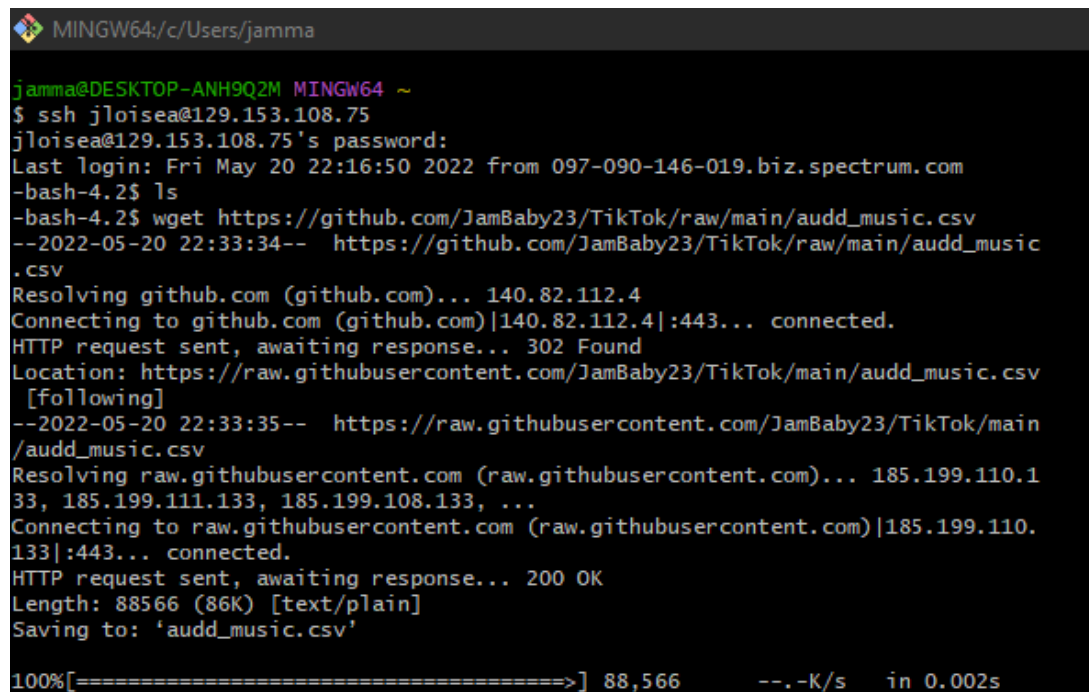
Outline

- Sign into hadoop cloud using SSH
- Install Tableau Desktop for free
- Analyze TikTok music data in Hive
- Import HDFS Data into tableau for visualization

1.

Download the listed files from GitHub.

```
wget
https://github.com/JamBaby23/TikTok/raw/main/audd\_music.csv
wget
https://github.com/JamBaby23/TikTok/raw/main/audd\_music\_apple\_music.csv
wget
https://github.com/JamBaby23/TikTok/raw/main/audd\_music\_spotify\_music.csv
wget
https://github.com/JamBaby23/TikTok/raw/main/audd\_music\_spotify\_music\_artists.csv
```



```
MINGW64/c/Users/jamma

jamma@DESKTOP-ANH9Q2M MINGW64 ~
$ ssh jloisea@129.153.108.75
jloisea@129.153.108.75's password:
Last login: Fri May 20 22:16:50 2022 from 097-090-146-019.biz.spectrum.com
-bash-4.2$ ls
-bash-4.2$ wget https://github.com/JamBaby23/TikTok/raw/main/audd_music.csv
--2022-05-20 22:33:34-- https://github.com/JamBaby23/TikTok/raw/main/audd_music.csv
Resolving github.com (github.com)... 140.82.112.4
Connecting to github.com (github.com)|140.82.112.4|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://raw.githubusercontent.com/JamBaby23/TikTok/main/audd_music.csv [following]
--2022-05-20 22:33:35-- https://raw.githubusercontent.com/JamBaby23/TikTok/main/audd_music.csv
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.110.133, 185.199.111.133, 185.199.108.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.110.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 88566 (86K) [text/plain]
Saving to: 'audd_music.csv'

100%[=====>] 88,566 --.-K/s in 0.002s
```

```

2022-05-20 22:33:35 (45.8 MB/s) - 'audd_music.csv' saved [88566/88566]

-bash-4.2$ wget https://github.com/JamBaby23/TikTok/raw/main/audd_music_apple_mu
sic.csv
--2022-05-20 22:33:45-- https://github.com/JamBaby23/TikTok/raw/main/audd_music
_apple_music.csv
Resolving github.com (github.com)... 140.82.112.4
Connecting to github.com (github.com)|140.82.112.4|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://raw.githubusercontent.com/JamBaby23/TikTok/main/audd_music_app
le_music.csv [following]
--2022-05-20 22:33:45-- https://raw.githubusercontent.com/JamBaby23/TikTok/main
/audd_music_apple_music.csv
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.109.1
33, 185.199.108.133, 185.199.111.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.109.
133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 185432 (181K) [text/plain]
Saving to: 'audd_music_apple_music.csv'
100%F... 185.432... K/s... in 0.002s

```

1a.

List files to confirm files were downloaded correctly

Ls

```

2022-05-20 22:34:00 (96.8 MB/s) - 'audd_music_spotify_music_artists.csv' saved [
83144/83144]

-bash-4.2$ ls
audd_music_apple_music.csv  audd_music_spotify_music_artists.csv
audd_music.csv             audd_music_spotify_music.csv
-bash-4.2$

```

2.

Reads all the files

```

tail -3 audd_music.csv
tail -3 audd_music_apple_music.csv
tail -3 audd_music_spotify_music.csv
tail -3 audd_music_spotify_music_artists.csv

```

```

MINGW64:/c/Users/jamma
audd_music_apple_music.csv audd_music_spotify_music_artists.csv
audd_music.csv audd_music_spotify_music.csv
-bash-4.2$ tail -3 audd_music.csv
6829505139569608705,Lady Gaga,Rain On Me (with Ariana Grande),Rain On Me (with A
riana Grande),2020-05-22,Interscope Records,02:34,https://lis.tn/EvcBk,,24yS12hO
PGCDcx8xFIqWBU
6836709783203105541,Variance,Inborn,Monodream,2015-10-28,Breakdown Records,00:46
,https://lis.tn/Inborn,FR2X41510836,2b0JMyCUHttZPUgM6zY09o
57419989,Aly & AJ,Potential Breakup Song,Insomniatic,2007-10-22,Charisma,00:26,h
ttps://lis.tn/PotentialBreakupSong,USHR10723111,11dxtPJKR4E0w15r0A0t47
-bash-4.2$ tail -3 audd_music_apple_music.csv
USSM12006058,Kina & Mishaal,https://music.apple.com/us/album/tell-me-about-you/1
536219696?app=music&at=1000133QU&i=1536219700&mt=1,1.0,"['Electronic', 'Music']"
,194400.0,2020-11-20,Tell Me About You,Tell Me About You - Single,1.0,Kina & Mis
haal Tamer,3000.0,3000.0,https://is3-ssl.mzstatic.com/image/thumb/Music114/v4/7f
/77/b9/7f77b959-8543-dbb5-47d8-f879c774bf98/886448822899.jpg/{w}{h}bb.jpeg,0101
04,b6b0d0,a8a7dd,918da7,8686b1,1536219700,song
FR2X41510836,Variance,https://music.apple.com/us/album/inborn/1052868622?app=mus
ic&at=1000133QU&i=1052869303&mt=1,1.0,"['Rock', 'Music']",75001.0,2015-10-28,Inb
orn,Monodream - EP,1.0,Dedo & Dony,1440.0,1440.0,https://is3-ssl.mzstatic.com/im
age/thumb/Music62/v4/5c/c6/ad/5cc6ad2b-6124-a119-0618-ab7c9edd543f/mzm.mqavrszs.
jpg/{w}{h}bb.jpeg,dedede,131313,222222,3b3b3b,484848,1052869303,song
USHR10723111,Aly & AJ,https://music.apple.com/us/album/potential-breakup-song/14
44199068?app=music&at=1000133QU&i=1444199217&mt=1,1.0,"['Pop', 'Music']",219773.
0,2007-06-26,Potential Breakup Song,Insomniatic,1.0,"Aly Michalka, AJ Michalka,
Antonina Arnato & Tim James",1423.0,1421.0,https://is3-ssl.mzstatic.com/image/th
umb/Music128/v4/ef/9b/63/ef9b63fd-f512-3061-e1d0-19abe4451dd2/00720616264220.rgb
.jpg/{w}{h}bb.jpeg,030306,d1cab0,c8c7bf,a7a28e,ala09a,1444199217,song
-bash-4.2$ tail -3 audd_music_spotify_music.csv
24yS12hOPGCDcx8xFIqWBU,83.0,True,,1.0,182200.0,False,https://api.spotify.com/v1
/tracks/24yS12hOPGCDcx8xFIqWBU,Rain On Me (with Ariana Grande),1.0,spotify:trac
k:24yS12hOPGCDcx8xFIqWBU,Rain On Me (with Ariana Grande),single,4TqgXMSSTwP3RCO
3MMSR6t,spotify:album:4TqgXMSSTwP3RCO3MMSR6t,https://api.spotify.com/v1/albums/
4TqgXMSSTwP3RCO3MMSR6t,"[{ 'height': 640, 'width': 640, 'url': 'https://i.scdn.co
/image/ab67616d0000b273c8583f0bd97d3042d4971acf'}, { 'height': 300, 'width': 300,
'ur1': 'https://i.scdn.co/image/ab67616d00001e02c8583f0bd97d3042d4971acf'}, { 'h
eight': 64, 'width': 64, 'url': 'https://i.scdn.co/image/ab67616d00004851c8583f0
bd97d3042d4971acf'}]\"",https://open.spotify.com/album/4TqgXMSSTwP3RCO3MMSR6t,2020
-05-22,day,USUM72004304,https://open.spotify.com/track/24yS12hOPGCDcx8xFIqWBU,"1
HY2Jd0NmPumShAr6Kms,66CXWjxzNUsdJx32Jdwvnr"

```

3.

Create the directories for the required files. Check that the directories were correctly made.

```

hdfs dfs -ls
hdfs dfs -mkdir tiktokmusic
hdfs dfs -mkdir applemusic
hdfs dfs -mkdir spotifymusic
hdfs dfs -mkdir spotifyartist
hdfs dfs -ls

```

```

-bash-4.2$ hdfs dfs -ls
Found 1 items
drwx----- - jloisea hdfs          0 2022-05-20 22:18 .Trash
-bash-4.2$ hdfs dfs -mkdir tiktokmusic
-bash-4.2$ hdfs dfs -mkdir applemusic
-bash-4.2$ hdfs dfs -mkdir spotifymusic
-bash-4.2$ hdfs dfs -mkdir spotifyartist
-bash-4.2$ hdfs dfs -ls
Found 5 items
drwx----- - jloisea hdfs          0 2022-05-20 22:18 .Trash
drwxr-xr-x - jloisea hdfs          0 2022-05-20 22:44 applemusic
drwxr-xr-x - jloisea hdfs          0 2022-05-20 22:45 spotifyartist
drwxr-xr-x - jloisea hdfs          0 2022-05-20 22:44 spotifymusic
drwxr-xr-x - jloisea hdfs          0 2022-05-20 22:44 tiktokmusic

```

4.

Inserts all the data from the file into the appropriate directory

```
hdfs dfs -put audd_music.csv tiktokmusic
hdfs dfs -put audd_music_apple_music.csv applemusic
hdfs dfs -put audd_music_spotify_music.csv spotifymusic
hdfs dfs -put audd_music_spotify_music_artists.csv spotifyartist
```

```
drwxr-xr-x  - jloisea hdfs          0 2022-05-20 22:44 tiktokmusic
-bash-4.2$ hdfs dfs -put audd_music.csv tiktokmusic
-bash-4.2$ hdfs dfs -put audd_music_apple_music.csv applemusic
-bash-4.2$ hdfs dfs -put audd_music_spotify_music.csv spotifymusic
-bash-4.2$ hdfs dfs -put audd_music_spotify_music_artists.csv spotifyartist
-bash-4.2$ hdfs dfs -ls tiktokmusic
Found 1 items
-rw-r--r--  3 jloisea hdfs      88566 2022-05-20 22:45 tiktokmusic/audd_music.csv
-bash-4.2$ hdfs dfs -ls applemusic
Found 1 items
-rw-r--r--  3 jloisea hdfs     185432 2022-05-20 22:45 applemusic/audd_music_apple_music.csv
-bash-4.2$ hdfs dfs -ls spotifymusic
Found 1 items
-rw-r--r--  3 jloisea hdfs     392272 2022-05-20 22:45 spotifymusic/audd_music_spotify_music.csv
-bash-4.2$ hdfs dfs -ls spotifyartist
Found 1 items
-rw-r--r--  3 jloisea hdfs      83144 2022-05-20 22:46 spotifyartist/audd_music_spotify_music_artists.csv
-bash-4.2$ beeline
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/odh/1.1.0.351/hive/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/odh/1.1.0.351/hadoop/lib/slf4j-log4j12-1.7.25.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]
Connecting to jdbc:hive2://bigdaiwn0.sub02180640120.trainingvcn.oraclevcn.com:2181,bigdaiwn0.sub02180640120.trainingvcn.oraclevcn.com:2181,bigda
erNamespace=hiveserver2
22/05/20 22:47:41 [main-EventThread]: ERROR imps.EnsembleTracker: Invalid config event received: {server.1=bigdaiwn0.sub02180640120.trainingvcn.
nt, server.2=bigdaiwn0.sub02180640120.trainingvcn.oraclevcn.com:2888:3888:participant}
22/05/20 22:47:41 [main-EventThread]: ERROR imps.EnsembleTracker: Invalid config event received: {server.1=bigdaiwn0.sub02180640120.trainingvcn.
nt, server.2=bigdaiwn0.sub02180640120.trainingvcn.oraclevcn.com:2888:3888:participant}
22/05/20 22:47:42 [main]: INFO jdbc.HiveConnection: Connected to bigdaiwn0.sub02180640120.trainingvcn.oraclevcn.com:10000
Connected to: Apache Hive (version 3.1.2)
Driver: Hive JDBC (version 3.1.2)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 3.1.2 by Apache Hive
0: jdbc:hive2://bigdaiwn0.sub02180640120.tra: use jloisea;
INFO : Compiling command(queryId=hive_20220520224746_425c8035-ca86-4273-b4bf-ea045e4cd5ae): use jloisea
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(FieldSchemas:null, properties:null)
INFO : Completed compiling command(queryId=hive_20220520224746_425c8035-ca86-4273-b4bf-ea045e4cd5ae); Time taken: 0.039 seconds
INFO : Executing command(queryId=hive_20220520224746_425c8035-ca86-4273-b4bf-ea045e4cd5ae): use jloisea
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20220520224746_425c8035-ca86-4273-b4bf-ea045e4cd5ae); Time taken: 0.009 seconds
INFO : OK
No rows affected (0.1 seconds)
0: jdbc:hive2://bigdaiwn0.sub02180640120.tra: show tables;
```

5.

Lists all the files in the specified directory.

```
hdfs dfs -ls tiktokmusic
hdfs dfs -ls applemusic
hdfs dfs -ls spotifymusic
hdfs dfs -ls spotifyartist
```

6.

Goes into Beeline to begin creating the tables. Creates a table for the 'tiktokmusic' directory. Gives each attribute a NAME or INT, gives it a location in the file system and assigns the table properties. (**Make sure to use your own username**)

```
beeline
use username

CREATE EXTERNAL TABLE IF NOT EXISTS tiktokmusic
(id bigint, artist string, title string, album string, release_date
date, label string, timecode string, song_link string,
apple_music_isrc string, spotify_id string)
ROW FORMAT DELIMITED FIELDS TERMINATED BY "," LOCATION
"/user/jloisea/tiktokmusic" TBLPROPERTIES ('skip.header.line.count' =
'1');
```

6a.

Creates a table for the 'applemusic' directory. Gives each attribute a NAME or INT, and gives it a location in the file system.

```
CREATE EXTERNAL TABLE IF NOT EXISTS applemusic
(isrc string, artistName string, url string, discNumber float,
genreNames string, durationInMillis double, releaseDate date, name
string, albumName string, trackNumber float, composerName string,
artwork_width double, artwork_height double, artwork_url string,
artwork_bgColor string, artwork_textColor1 string, artwork_textColor2
string, artwork_textColor3 string, artwork_textColor4 string,
playParams_id double, playParams_kind string)
ROW FORMAT DELIMITED FIELDS TERMINATED BY "," LOCATION
"/user/jloisea/applemusic" TBLPROPERTIES ('skip.header.line.count' =
'1');
```

6b.

Creates a table for the 'spotifymusic' directory. Gives each attribute a NAME or INT, gives it a location in the file system and assigns the table properties.

```
CREATE EXTERNAL TABLE IF NOT EXISTS spotifymusic
(id string, popularity double, is_playable boolean, linked_from
string, available string, disc_number float, duration_ms double,
explicit boolean, href string, name string, preview_url string,
track_number float, uri string, album_name string, album_album_group
```

```

string, album_album_type string, album_uri string,
album_available_markets string, album_herf string, album_images
string, album_external_urls_spotify string, album_release_date
string, album_release_date_precision string, external_ids_isrc
string, external_urls_spotify string, artists_ids string )
ROW FORMAT DELIMITED FIELDS TERMINATED BY "," LOCATION
"/user/jloisea/spotifymusic" TBLPROPERTIES ('skip.header.line.count'
= '1');

```

Creates a table for the 'spotifyartist' directory.Gives each attribute a NAME or INT, and gives it a location in the file system.

```

CREATE EXTERNAL TABLE IF NOT EXISTS spotifyartist (id string, name
string, url string, href string, external_urls string) ROW FORMAT
DELIMITED FIELDS TERMINATED BY "," LOCATION
"/user/jloisea/spotifyartist" TBLPROPERTIES ('skip.header.line.count'
= '1');

```

6c.

Show tables to see if the tables were made correctly

```

... 4 more
[Vertex killed, vertexName=Reducer 3, vertexId=vertex_1652074645349_0150_1_02, diagnostics=[Vertex received kill in INITED state, Vertex vertex_1652074645349_0150_1_02 [Reducer 3] killed/failed due to:OTHER_VERTEX_FAILURE]Vertex killed, vertexName=Reducer 2, vertexId=vertex_1652074645349_0150_1_01, diagnostics=[Vertex received kill in INITED state, Vertex vertex_1652074645349_0150_1_01 [Reducer 2] killed/failed due to:OTHER_VERTEX_FAILURE]DAG did not succeed due to:VERTEX_FAILURE, failedVertices:1 killedVertices:2
INFO : Completed executing command(queryId=hive_20220520225446_7d5106a9-88f4-4d12-b9e2-a7276d8b943e): Error while processing statement: java.lang.RuntimeException: Error: return code 2 from org.apache.hadoop.hiveql.exec.tez.TezTask, Vertex failed, vertexName=Map 1, vertexId=vertex_1652074645349_0150_1_00, diagnostics=[Vertex vertex_1652074645349_0150_1_00 [Map 1] killed/failed due to:ROOT_INPUT_FAILURE, Vertex input: spotifyartist_initializer failed, vertex=vertex_1652074645349_0150_1_00 [Map 1], java.lang.RuntimeException: ORC split generation failed with exception: org.apache.orc.FileFormatException: Malformed ORC File hdfs://bigdataim0.sdb02180640120.trainingvcn.oraclecn.com:8020/user/jloisea/spotifyartist/audd_music_spotify_music_artists.csv. Invalid postscript.
at org.apache.hadoop.hiveql.io.orc.OrcInputFormat.generateSplitInfo(OrcInputFormat.java:1833)
at org.apache.hadoop.hiveql.io.orc.OrcInputFormat.getSplit(OrcInputFormat.java:1939)
at org.apache.hadoop.hiveql.io.HiveInputFormat.addSplitToGroup(OrcInputFormat.java:519)
at org.apache.hadoop.hiveql.io.HiveInputFormat.getSplit(HiveInputFormat.java:765)
at org.apache.hadoop.hiveql.exec.tez.HiveSplitGenerator.initialize(HiveSplitGenerator.java:243)
at org.apache.tez.dag.app.dag.RootInputInitializerManager.initialize(HiveSplitGenerator$RootInputInitializer$1(RootInputInitializerManager.java:200)
at java.security.AccessController.doPrivileged(Native Method)
at org.apache.hadoop.hiveql.io.HiveSplitGenerator.doAs(HiveSplitGenerator.java:179)
at org.apache.hadoop.hiveql.io.HiveSplitGenerator.run(HiveSplitGenerator.java:193)
at org.apache.tez.dag.app.dag.RootInputInitializerManager.run(HiveSplitGenerator$RootInputInitializerManager.java:174)
at org.apache.tez.dag.app.dag.RootInputInitializerManager.initialize(HiveSplitGenerator$RootInputInitializerManager.java:168)
at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511)
at java.util.concurrent.FutureTask.run(FutureTask.java:266)
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
at java.lang.Thread.run(Thread.java:750)
Caused by: java.util.concurrent.ExecutionException: org.apache.orc.FileFormatException: Malformed ORC file hdfs://bigdataim0.sdb02180640120.trainingvcn.oraclecn.com:8020/user/jloisea/spotifyartist/audd_music_spotify_music_artists.csv. Invalid postscript.
at java.util.concurrent.FutureTask.report(FutureTask.java:122)
at java.util.concurrent.FutureTask.get(FutureTask.java:192)
at org.apache.hadoop.hiveql.io.orc.OrcInputFormat.generateSplitInfo(OrcInputFormat.java:1845)
... 16 more
Caused by: org.apache.orc.FileFormatException: Malformed ORC file hdfs://bigdataim0.sdb02180640120.trainingvcn.oraclecn.com:8020/user/jloisea/spotifyartist/audd_music_spotify_music_artists.csv. Invalid postscript.
at org.apache.orc.impl.ReaderImpl.ensureReader(ReaderImpl.java:173)
at org.apache.orc.impl.ReaderImpl.extractFile(ReaderImpl.java:570)
at org.apache.orc.impl.ReaderImpl.<init>(ReaderImpl.java:343)
at org.apache.hadoop.hiveql.io.orc.ReaderImpl.<init>(ReaderImpl.java:61)
at org.apache.hadoop.hiveql.io.orc.OrcFile.createReader(OrcFile.java:190)
at org.apache.hadoop.hiveql.io.orc.OrcInputFormat$SplitGenerator.openLocalAndCacheStripedDetails(OrcInputFormat.java:1847)
at org.apache.hadoop.hiveql.io.orc.OrcInputFormat$SplitGenerator.callInternal(OrcInputFormat.java:1533)
at org.apache.hadoop.hiveql.io.orc.OrcInputFormat$SplitGenerator.access$57$NORCInputFormat.java:1329)
at org.apache.hadoop.hiveql.io.orc.OrcInputFormat$SplitGenerator$1.run(OrcInputFormat.java:1513)
at org.apache.hadoop.hiveql.io.orc.OrcInputFormat$SplitGenerator$1.run(OrcInputFormat.java:1510)
at java.security.AccessController.doPrivileged(Native Method)
at java.security.AccessController.doAs(AccessController.java:420)
at org.apache.hadoop.hiveql.io.orc.OrcInputFormat$SplitGenerator.call(OrcInputFormat.java:1510)
at org.apache.hadoop.hiveql.io.orc.OrcInputFormat$SplitGenerator.call(OrcInputFormat.java:1329)
... 4 more
[Vertex killed, vertexName=Reducer 3, vertexId=vertex_1652074645349_0150_1_02, diagnostics=[Vertex received kill in INITED state, Vertex vertex_1652074645349_0150_1_02 [Reducer 3] killed/failed due to:OTHER_VERTEX_FAILURE]Vertex killed, vertexName=Reducer 2, vertexId=vertex_1652074645349_0150_1_01, diagnostics=[Vertex received kill in INITED state, Vertex vertex_1652074645349_0150_1_01 [Reducer 2] killed/failed due to:OTHER_VERTEX_FAILURE]DAG did not succeed due to:VERTEX_FAILURE, failedVertices:1 killedVertices:2 (state=0800),code=2)
INFO : Compiling command(queryId=hive_20220520225446_7d5106a9-88f4-4d12-b9e2-a7276d8b943e): show tables
INFO : Semantic Analysis Completed (retail = false)
INFO : Returning Hive schema: Schema(fieldsSchema:[FieldsSchema(name=tab_name, type=string, comment=from deserialiser]], properties=[])
INFO : Completed compiling command(queryId=hive_20220520225446_7d5106a9-88f4-4d12-b9e2-a7276d8b943e): Time taken: 0.03 seconds
INFO : Executing command(queryId=hive_20220520225446_7d5106a9-88f4-4d12-b9e2-a7276d8b943e): show tables
INFO : Starting Task (Stage=000): in serial mode
INFO : Completed executing command(queryId=hive_20220520225446_7d5106a9-88f4-4d12-b9e2-a7276d8b943e): Time taken: 0.008 seconds
INFO : OK

+-----+
| tab_name |
+-----+
| applemusic |
| spotifyartist |
| spotifymusic |
| tiktokmusic |
+-----+

4 rows selected (0.954 seconds)
01: jdbc:hive2://bigdataim0.sdb02180640120.traini

```

show tables;

6d. Selected 5 rows from the specified directory and displayed them in a descending order. (Figure above)

```
select * from tiktokmusic desc limit 5;
select * from applemusic desc limit 5;
select * from spotifymusic desc limit 5;
select * from spotifyartist desc limit 5;
```


7.

Creates a new table named 'cleanspotifyartist', only this time it gets rid of all the null values. This way the data is cleaned.

```
CREATE TABLE IF NOT EXISTS cleanspotifyartist AS SELECT id, name, url, count(id) as cnt from spotifyartist where id is not null AND id != "" AND id != "null" group by id, name, url order by cnt DESC;
```

7b.

Creates a new table named 'cleanspotifymusic', only this time it gets rid of all the null values. This way the data is cleaned.

```
CREATE TABLE IF NOT EXISTS cleanspotifymusic AS SELECT id, popularity, duration_ms, name, album_name, count(id) as cnt from spotifymusic where id is not null AND id != "" AND id != "null" group by id, popularity, duration_ms, name, album_name order by cnt DESC;
```

7c.

Creates a new table named 'cleanapplemusic', only this time it gets rid of all the null values. This way the data is cleaned.

```
CREATE TABLE IF NOT EXISTS cleanapplemusic AS SELECT isrc, artistName, durationInMillis, releaseDate, name, albumName, count(isrc) as cnt from applemusic where isrc is not null AND isrc != "" AND isrc != "null" group by isrc, artistName, durationInMillis, releaseDate, name, albumName order by cnt DESC;
```

8.

Verifies the tables were created correctly and lists them in a descending order with a limit of 5 attributes.

```
select * from cleanapplemusic DESC LIMIT 5;
select * from cleanspotifymusic DESC LIMIT 5;
select * from cleanspotifyartist DESC LIMIT 5;
```

```

INFO : Starting task [Stage-2:DEPENDENCY_COLLECTION] in serial mode
INFO : Starting task [Stage-0:MOVE] in serial mode
INFO : Starting task [Stage-4:DDL] in serial mode
INFO : Starting task [Stage-3:STATS] in serial mode
INFO : Completed executing command(queryId=hive_20220521042502_653ec63b-33cd-4c62-ad8e-fc66de630409); Time taken: 18.248 seconds
INFO : OK
No rows affected (18.469 seconds)
0: jdbc:hive2://bigdaiwn0.sub02180640120.trai> select * from cleanapplemusic DESC LIMIT 5;
INFO : Compiling command(queryId=hive_20220521042557_a6a9f3fc-2e57-4a32-891d-d8b30a6966d6): select * from cleanapplemusic DESC LIMIT 5
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(FieldSchemas:[FieldSchema(name:desc.isrc, type:string, comment:null), FieldSchema(name:desc.artistname, type:string, type:date, comment:null), FieldSchema(name:desc.name, type:string, comment:null), FieldSchema(name:desc.albumname, type:string, comment:null), FieldSchema(name:desc.durationinmillis, type:bigint, comment:null)], properties:[]); Time taken: 0.301 seconds
INFO : Executing command(queryId=hive_20220521042557_a6a9f3fc-2e57-4a32-891d-d8b30a6966d6): select * from cleanapplemusic DESC LIMIT 5
INFO : Completed executing command(queryId=hive_20220521042557_a6a9f3fc-2e57-4a32-891d-d8b30a6966d6); Time taken: 0.0 seconds
INFO : OK

```

desc.isrc	desc.artistname	desc.durationinmillis	desc.releasedate	desc.name	desc.albumname	desc.cnt
USUM72021500	Billie Eilish	NULL	NULL	2020-11-12	Therefore I Am	12
GBKQU1777771	Pascal Letoublon	NULL	NULL	242016.0	2017-09-05	5
GBX2M1300001	Studio Killers	NULL	NULL	2013-05-03	Jenny (I Wanna Ruin Our Friendship)	5
USSM12006586	The Kid LAROI	NULL	NULL	2020-11-06	WITHOUT YOU	5
USHR10723111	Aly & AJ	NULL	NULL	2007-06-26	Potential Breakup Song	5

```

5 rows selected (0.362 seconds)
0: jdbc:hive2://bigdaiwn0.sub02180640120.trai> select * from cleanspotifymusic DESC LIMIT 5;
INFO : Compiling command(queryId=hive_20220521042750_9879bd83-0db8-4080-b1bf-1213defdf5f10): select * from cleanspotifymusic DESC LIMIT 5
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(FieldSchemas:[FieldSchema(name:desc.id, type:string, comment:null), FieldSchema(name:desc.popularity, type:double, comment:null), FieldSchema(name:desc.album_name, type:string, comment:null), FieldSchema(name:desc.cnt, type:bigint, comment:null)], properties:[]); Time taken: 0.141 seconds
INFO : Executing command(queryId=hive_20220521042750_9879bd83-0db8-4080-b1bf-1213defdf5f10): select * from cleanspotifymusic DESC LIMIT 5
INFO : Completed executing command(queryId=hive_20220521042750_9879bd83-0db8-4080-b1bf-1213defdf5f10); Time taken: 0.001 seconds
INFO : OK

```

desc.id	desc.popularity	desc.duration_ms	desc.name	desc.album_name	desc.cnt
54bFM56PmE4YLRnqpW6Tha	91.0	174321.0	Therefore I Am	Therefore I Am	12
11dxtPJKR4E0w1Sr0A0t47	72.0	219773.0	Potential Breakup Song	Insomniatic	5
3S2XkZJHSP3AqYk8ChYsWB	74.0	242015.0	Friendships	Friendships	5
4JLB0UY5a0MPYND0iWeSWQ	78.0	215280.0	Jenny (I Wanna Ruin Our Friendship)	Jenny (I Wanna Ruin Our Friendship)	5
270eeYzk6k1gBh83TSvGMA	94.0	161384.0	WITHOUT YOU	F*CK LOVE (SAVAGE)	5

```

5 rows selected (0.199 seconds)
0: jdbc:hive2://bigdaiwn0.sub02180640120.trai>

```

9.

```
INSERT OVERWRITE DIRECTORY '/user/jloisea/data/' ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',' SELECT * FROM cleanspotifymusic;
```

```
INSERT OVERWRITE DIRECTORY '/user/jloisea/data1/' ROW FORMAT
DELIMITED FIELDS TERMINATED BY ',' SELECT * FROM cleanspotifyartist;
```

```
INSERT OVERWRITE DIRECTORY '/user/jloisea/data2/' ROW FORMAT
DELIMITED FIELDS TERMINATED BY ',' SELECT * FROM cleanapplemusic;
```

10.

Return to HDFS.

Shows the directory of the newly created tables

```
hdfs dfs -ls data
hdfs dfs -ls data1
hdfs dfs -ls data2
```

```
drwxr-xr-x - jloisea hdfs 0 2022-05-22 00:35 data/hive-staging_hive_
2022-05-22_00-34-52.582.4644288677458444776-32
rwxr-xr-x 3 jloisea hdfs 20501 2022-05-22 04:46 data/000000_0
-bash-4.2$
-bash-4.2$ hdfs dfs -cat data/000000_0 | tail -n 2
VN,VN,spotify:artist:7uBCumake2U6No8rWbFzr,1
VN,VN,spotify:artist:7uBCumake2U6No8rWbFzr,1
-bash-4.2$ hdfs dfs -ls data
Found 2 items
drwxr-xr-x - jloisea hdfs 0 2022-05-22 00:35 data/hive-staging_hive_
2022-05-22_00-34-52.582.4644288677458444776-32
rwxr-xr-x 3 jloisea hdfs 20501 2022-05-22 04:46 data/000000_0
-bash-4.2$ hdfs dfs -ls clean
Found 1 items
rwxr-xr-x 3 jloisea hdfs 26055 2022-05-22 04:45 clean/000000_0
-bash-4.2$ hdfs dfs -rm -r clean/000000_0
22/05/22 04:47:52 INFO Fs.TrashPolicyDefault: Moved: 'hdfs://bigdata1m0.sub021806
40120.trainingcn.oraclecn.com:8020/user/jloisea/clean/000000_0' to trash at: h
dfs://bigdata1m0.sub02180640120.trainingcn.oraclecn.com:8020/user/jloisea/Tras
h/Current/user/jloisea/clean/000000_0
-bash-4.2$ hdfs dfs -rm -r data/000000_0
22/05/22 04:48:01 INFO Fs.TrashPolicyDefault: Moved: 'hdfs://bigdata1m0.sub021806
40120.trainingcn.oraclecn.com:8020/user/jloisea/data/000000_0' to trash at: h
dfs://bigdata1m0.sub02180640120.trainingcn.oraclecn.com:8020/user/jloisea/Trash
/Current/user/jloisea/data/000000_061334881011
-bash-4.2$ INSERT OVERWRITE DIRECTORY '/user/jloisea/data/' ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',' SELECT * FROM cleanspotifymusic;
-bash: INSERT: command not found
-bash-4.2$ INSERT OVERWRITE DIRECTORY '/user/jloisea/data/' ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',' SELECT * FROM cleanspotifymusic;
-bash: INSERT: command not found
-bash-4.2$ INSERT OVERWRITE DIRECTORY '/user/jloisea/data/' ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',' SELECT * FROM cleanspotifymusic;
-bash: INSERT: command not found
-bash-4.2$ hdfs dfs -ls data
Found 2 items
drwxr-xr-x - jloisea hdfs 0 2022-05-22 00:35 data/hive-staging_hive_
2022-05-22_00-34-52.582.4644288677458444776-32
rwxr-xr-x 3 jloisea hdfs 26051 2022-05-22 04:48 data/000000_0
-bash-4.2$ hdfs dfs -cat data/000000_0 | tail -n 2
PVxdu8t9t6zTr97pKw,tl,0.238826,0,Can I Get an Outlaw,Can I Get an Outlaw,1
ACOSTIVicinaliniDp,72,0.34973,0,Where Is My Mind? - Remastered,Surfer Rosa,
-bash-4.2$ hdfs dfs -ls data
Found 2 items
drwxr-xr-x - jloisea hdfs 0 2022-05-22 00:35 data/hive-staging_hive_
2022-05-22_00-34-52.582.4644288677458444776-32
rwxr-xr-x 3 jloisea hdfs 20501 2022-05-22 04:50 data/000000_0
-bash-4.2$ hdfs dfs -ls data
Found 2 items
drwxr-xr-x - jloisea hdfs 0 2022-05-22 00:35 data/hive-staging_hive_
2022-05-22_00-34-52.582.4644288677458444776-32
rwxr-xr-x 3 jloisea hdfs 26051 2022-05-22 04:50 data/000000_0
-bash-4.2$ hdfs dfs -cat data/000000_0 | tail -n 2
VN,VN,spotify:artist:7uBCumake2U6No8rWbFzr,1
VN,VN,spotify:artist:7uBCumake2U6No8rWbFzr,1
-bash-4.2$ hdfs dfs -rm -r data/000000_0
22/05/22 04:52:18 INFO Fs.TrashPolicyDefault: Moved: 'hdfs://bigdata1m0.sub021806
40120.trainingcn.oraclecn.com:8020/user/jloisea/data/000000_0' to trash at: h
dfs://bigdata1m0.sub02180640120.trainingcn.oraclecn.com:8020/user/jloisea/Trash
/Current/user/jloisea/data/000000_0613353504
-bash-4.2$ hdfs dfs -ls data
Found 2 items
drwxr-xr-x - jloisea hdfs 0 2022-05-22 00:35 data/hive-staging_hive_2022-05-22_00-34-52.582.4644288677458444776-12
rwxr-xr-x 3 jloisea hdfs 26053 2022-05-22 04:52 data/000000_0
-bash-4.2$ hdfs dfs -ls data1
Found 1 items
rwxr-xr-x 3 jloisea hdfs 20501 2022-05-22 04:53 data1/000000_0
-bash-4.2$ hdfs dfs -ls data2
Found 1 items
rwxr-xr-x 3 jloisea hdfs 18149 2022-05-22 04:53 data2/000000_0
-bash-4.2$
```

11.
Verifies that the data was properly moved to see if we obtained the desired results. Then reads the files outside of linux to see if they are properly downloaded.

```
hdfs dfs -cat data/000000_0 | tail -n 2
hdfs dfs -cat data1/000000_0 | tail -n 2
hdfs dfs -cat data2/000000_0 | tail -n 2

hdfs dfs -get data/000000_0 spotifym.csv
hdfs dfs -get data1/000000_0 spotifya.csv
hdfs dfs -get data2/000000_0 apple.csv
```

```
MINGW64:/c/Users/kenny
3b-b3eb-0871b9b44a92); Time taken: 3.736 seconds
INFO : OK
No rows affected (4.046 seconds)
0: jdbc:hive2://bigdaiwn0.sub02180640120.traig> hdfs dfs -ls data
. . . . .> Permission denied
Closing: 0: jdbc:hive2://bigdaiwn0.sub02180640120.trainingvcn.oraclevcn.com:2181
bigdaimn0.sub02180640120.trainingvcn.oraclevcn.com:2181,bigdaiun0.sub0218064012
0.trainingvcn.oraclevcn.com:2181/default;password=kduong31;serviceDiscoveryMode=
zooKeeper;user=kduong31;zooKeeperNamespace=hiveserver2
-bash-4.2$ hdfs dfs -ls data
Found 1 items
-rw-r--r-- 3 kduong31 hdfs 26055 2022-05-22 06:05 data/000000_0
-bash-4.2$ hdfs dfs -cat data/000000_0 | tail -n 2
7vVsDu0gTg6oZtrKy7pXcW,61.0,238826.0,Can I Get an Outlaw,Can I Get an Outlaw,1
7wCmS9TTVUcIhRa1DYFgPy,78.0,234973.0,Where Is My Mind? - Remastered,Surfer Rosa,
1
-bash-4.2$ hdfs dfs -cat data1/000000_0 | tail -n 2
7wbCwmaAe2U6NoBrWBfeTz,Sickddellz,spotify:artist:7wbCwmaAe2U6NoBrWBfeTz,1
7xTcuBOIAAIGDOSvwYFPzk,Daniel Powter,spotify:artist:7xTcuBOIAAIGDOSvwYFPzk,1
-bash-4.2$ hdfs dfs -cat data2/000000_0 | tail -n 2
USWD10730703,Hannah Montana,\N,\N,2007-03-20,Nobody's Perfect,1
ZA82Y2000076,Master KG,\N,\N,2019-11-29,Jerusalema (feat. Nomcebo Zikode) [Edit]
,1
-bash-4.2$ |
```

12.

List the cvs files and then reads the specified files.

```
ls
cat spotifym.csv | tail -n 2
cat spotifya.csv | tail -n 2
cat apple.csv | tail -n 2
```

```
-bash-4.2$ cat spotifym.csv | tail -n 2
7vVsDu0gTg6oZtrKy7pXcW,61.0,238826.0,Can I Get an Outlaw,Can I Get an Outlaw,1
7wCmS9TTVUcIhRa1DYFgPy,78.0,234973.0,Where Is My Mind? - Remastered,Surfer Rosa,
1
-bash-4.2$ cat spotifya.csv | tail -n 2
7wbCwmaAe2U6NoBrWBfeTz,Sickddellz,spotify:artist:7wbCwmaAe2U6NoBrWBfeTz,1
7xTcuBOIAAIGDOSvwYFPzk,Daniel Powter,spotify:artist:7xTcuBOIAAIGDOSvwYFPzk,1
-bash-4.2$ cat apple.csv | tail -n 2
USWD10730703,Hannah Montana,\N,\N,2007-03-20,Nobody's Perfect,1
ZA82Y2000076,Master KG,\N,\N,2019-11-29,Jerusalema (feat. Nomcebo Zikode) [Edit]
,1
-bash-4.2$ |
```

Downloads the files onto personal computer **(use your own username)**

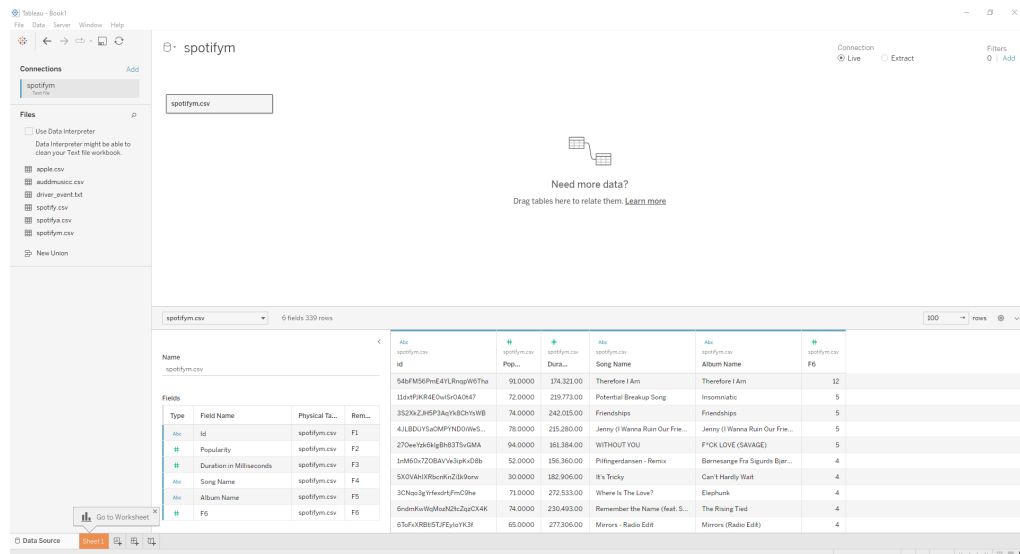
Open Tableau. Choose 'Text File' on the left hand side. Search through your Local disk, then navigate to your user file and select the name of the computer. There you will find the files that we downloaded from Git Bash.

The screenshot shows the Microsoft Excel interface with the following elements:

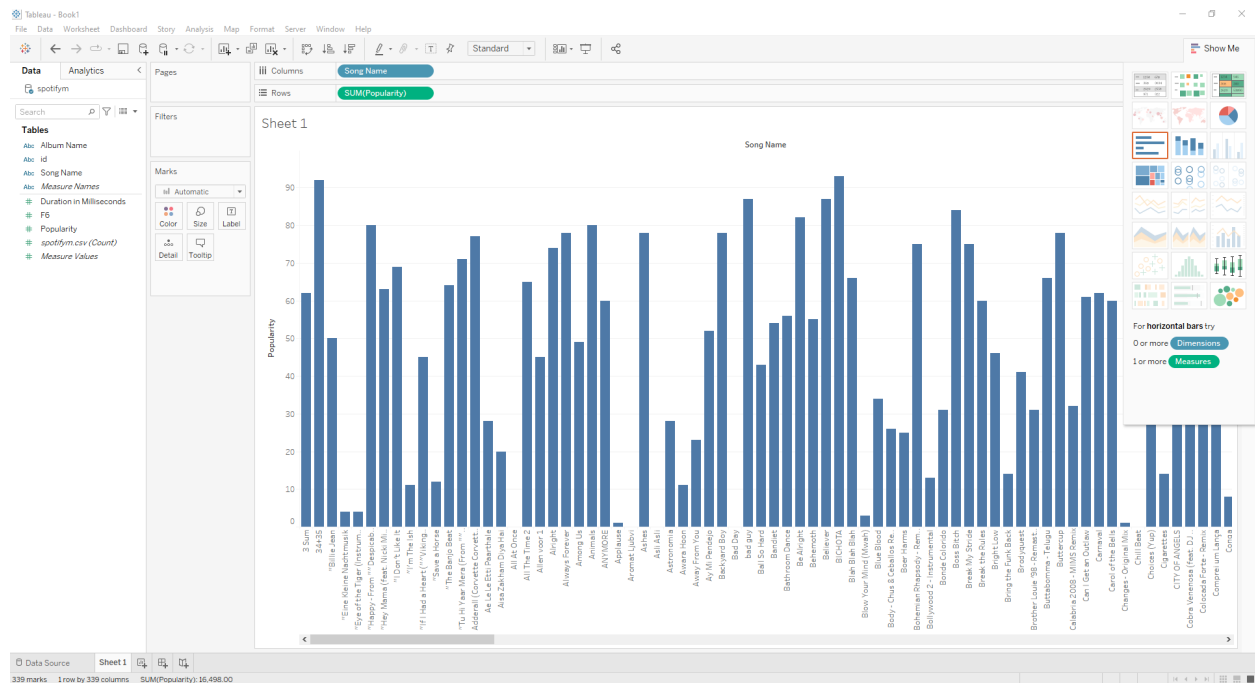
- Connections Pane (Left):** Shows a connection to 'spotify.csv'.
- Files Pane (Left):** Lists files including 'apple.csv', 'audio.midi.csv', 'driver_event.txt', 'spotify.csv', 'spotifya.csv', and 'spotifym.csv'.
- Main Data Table:**

Name	Type	Field Name	Value
spotify.csv	F1	spotify.csv	F2
spotify.csv	F3	spotify.csv	F4
spotify.csv	F5	spotify.csv	F6
spotify.csv	F6	spotify.csv	F6
- Context Menu (Over Name Column):**
 - Rename
 - Copy Values
 - Hide
 - Aliases...
 - Create Calculated Field...
 - Create Group...
 - Split
 - Custom Split...
 - Pivot (select multiple fields)
- Go to Worksheet Button:** Located at the bottom left of the main data table.

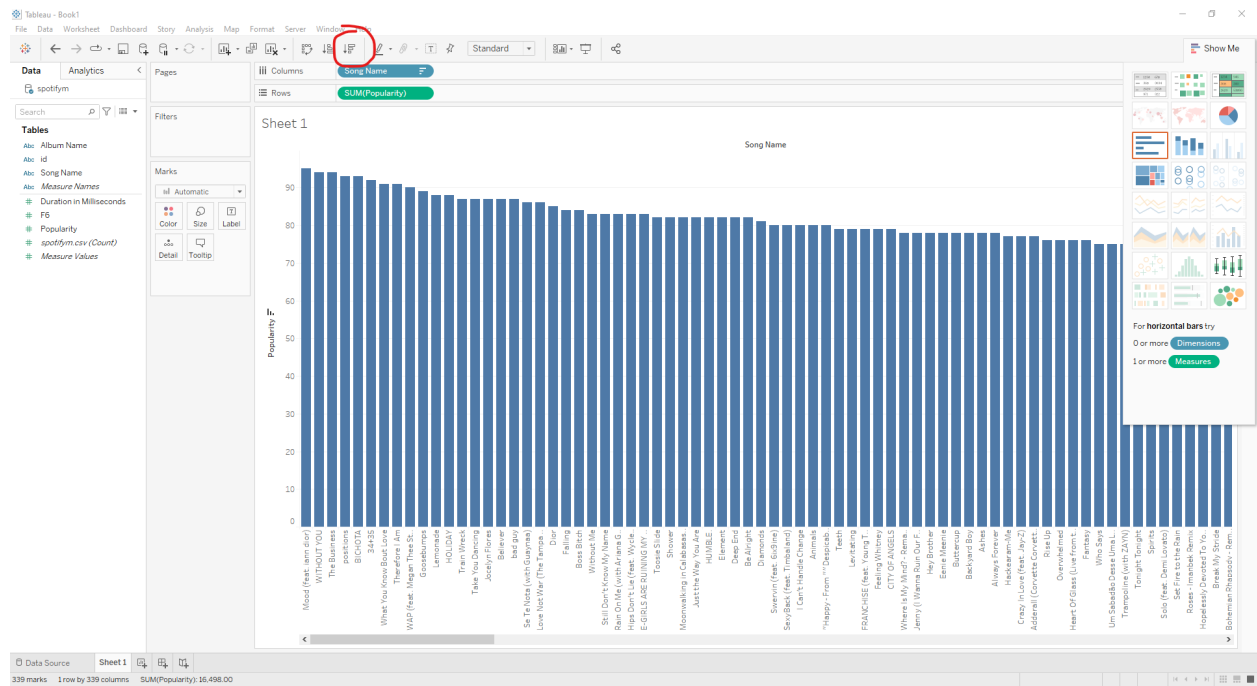
14b. This should be your result:



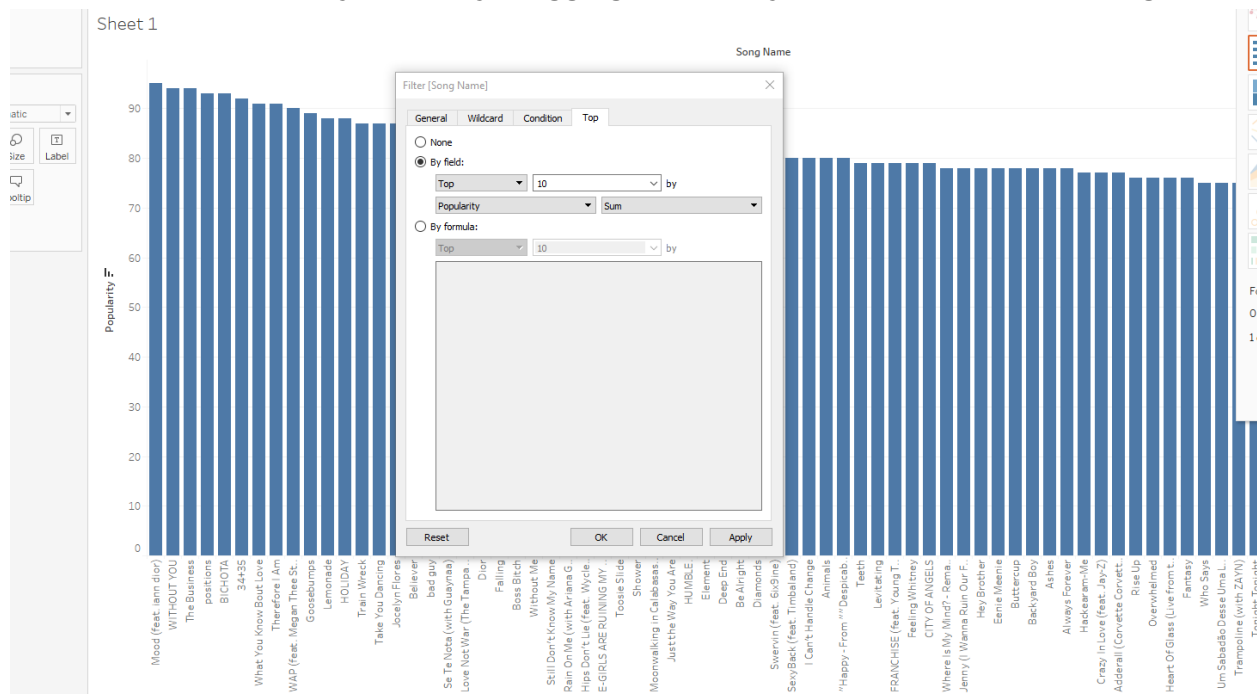
15. Drag 'song name' from the left hand side to the columns box and drag 'popularity' to rows.



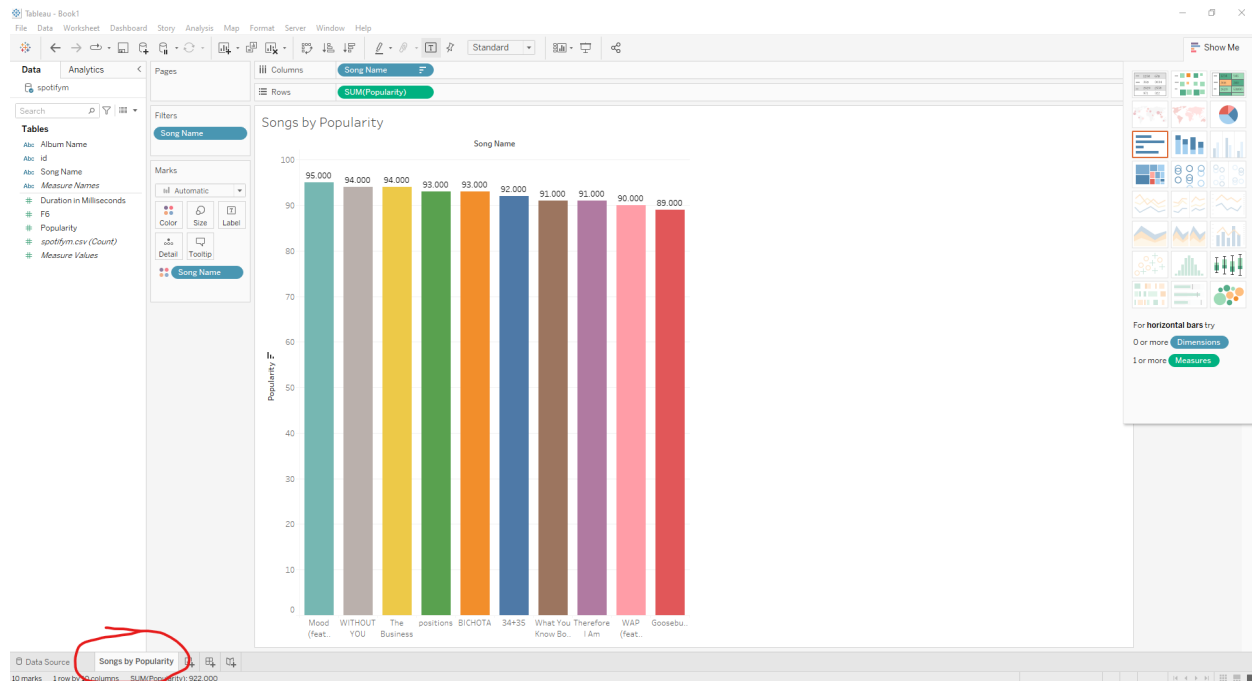
15b. You can sort by 'ascending' icon circled in red.



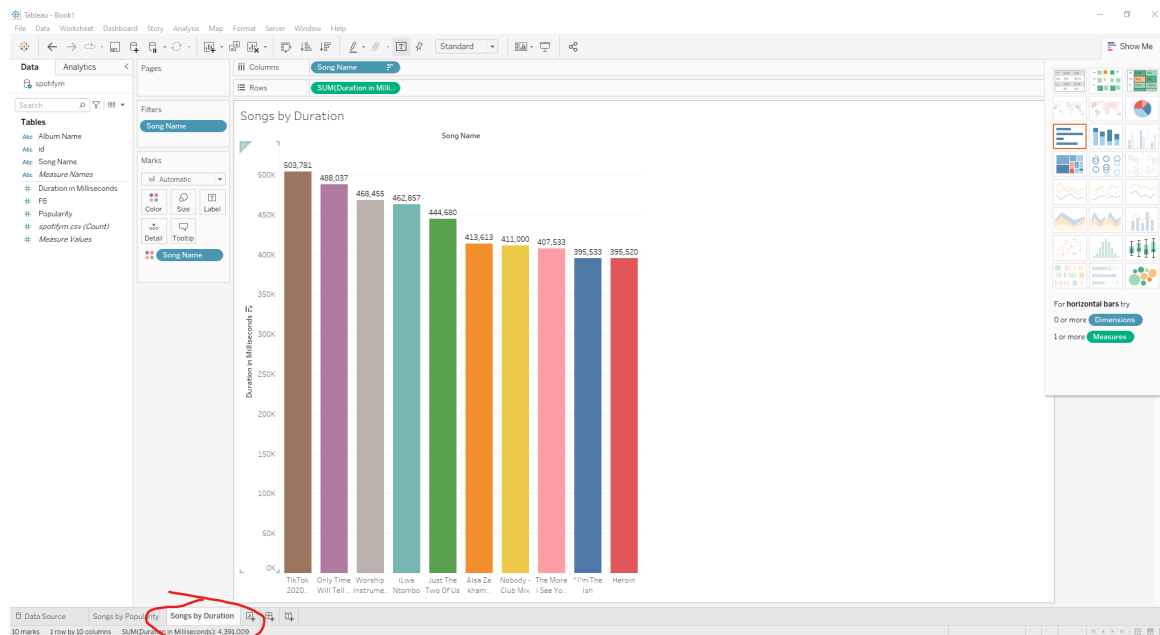
15c. You can also sort by top 10 by dragging 'Popularity' into 'filters' box on the right.



15d. Your graph should look something like this. Drag Song Name to color to add color to the graph. Make sure to change the sheet name to 'Songs by Popularity' on the bottom left corner.



16. You're going to be doing the exact same thing with the previous graph as with this graph but you will be replacing 'Popularity' with duration to create a time based graph. Make sure to change the name of the graph to 'Songs by Duration' (circled in red at the bottom).



Congratulations!!!

This is the end of the Lab! Make sure to screenshot the last two graphs as well as adding your initials at the top by saving the book under a different name to get full credit.