Benjamin Bartek
Bellevue University DSC 650
Final Report
November 16, 2024

DSC 650 Final Report

1. Introduction

The purpose of this project was to construct a data pipeline to process, analyze, and transform data from single data set. My pipeline uses Apache NiFi for data ingestion, Hadoop HDFS for storage, Apache Hive for data warehousing, and Apache Spark for querying and transformations. The data set and pipeline are discussed in further detail below.

2. Data Set

I selected an open-source, publicly available dataset hosted on Kaggle (https://www.kaggle.com/datasets/imls/museum-directory), by the Institute of Museum and Library Services and Abigail Larion (IMLS and Larion). This data set provides an overview of museums across the United States. It includes includes museum names, legal and alternate names, and classifications such as history museums, art museums, and botanical gardens. It also includes address data listing where the museums are located, latitude and longitude data, phone numbers, employer identification numbers (EIN), income, revenue, and tax periods.

3. Pipeline Overview

As mentioned above, my pipeline uses Apache NiFi for data ingestion, Hadoop HDFS for storage, Apache Hive for data warehousing, and Apache Spark for querying and transformations. The pipeline begins with Apache NiFi, which I used to ingest my file, titled `kagglemuseums.csv`, using an Invoke HTTP processor. A PutHDFS processor then transferred the file to HDFS. The data is then stored within HDFS. After adding the file to HDFS, I took a brief pipeline detour to use Spark Scala to preview the data set's schema before a table was created. After I understood the schema, I used Apache Hive to load the data into a table for querying and transformation. Finally, I returned to Spark Scala to demonstrate querying and transforming the data by filtering, adding derived columns (e.g., revenue in millions), grouping by state, and calculating average revenues. As we have learned in this class, these types of pipelines allow for efficient handling of data across the various tools.

4. Issues Encountered

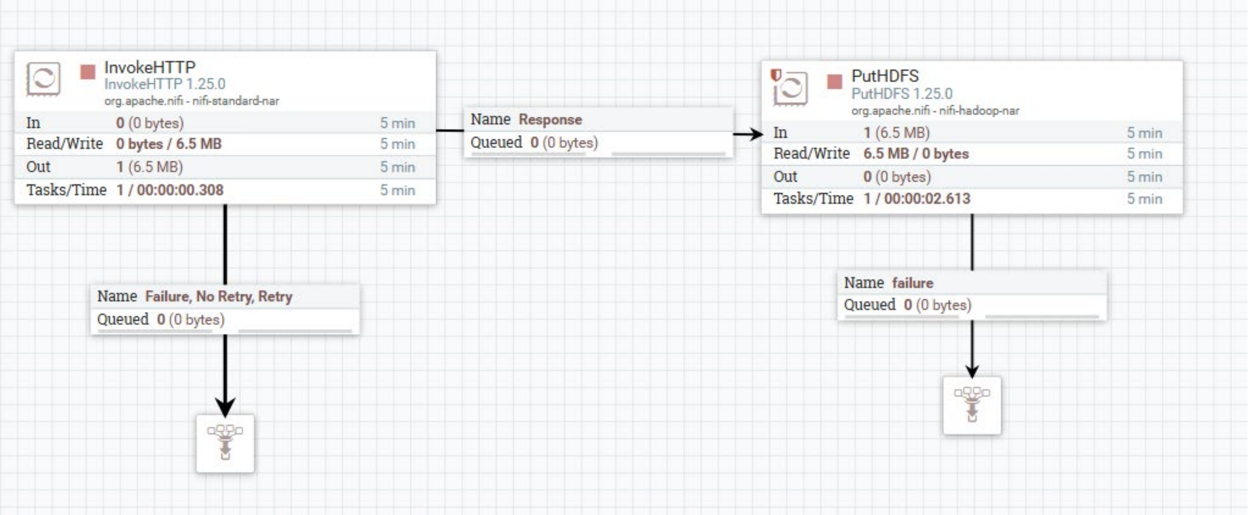
While creating this pipeline, I encountered a few challenges that required troubleshooting and adjustments. The first issue was that NiFi's PutHDFS processor consistently failed to move the file from Invoke HTTP due to a "connection refused" error until I sorted out the correct configuration because I initially had problems configuring the correct paths to the `core-site.xml` and `hdfs-site.xml` files. Locating the full HDFS file paths caused persistent problems for me in general during the project, but I learned new techniques for figuring that out. After previewing the schema in Spark Scala, some discrepancies in column data types required data type adjustments during the creation of the Hive table, but that was more along the lines of typical data cleaning. Finally, I encountered issues with Spark Scala freezing toward the end of the project, which required restarting and redoing some of my work a few times. This was sometimes, though not always, due to having a second terminal session open, which likely caused processing problems.

5. Screenshots and Code

NiFi – Data Ingestion

a. PutHDFS Configuration Resources.

```
/home/bbartek/dsc650-infra/bellevue-
bigdata/nifi/hadoopconf/core-site.xml,
/home/bbartek/dsc650-infra/bellevue-
bigdata/nifi/hadoopconf/hdfs-site.xml
```



InvokeHTTP
InvokeHTTP 1.25.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 6.5 MB 5 min
Out 1 (6.5 MB) 5 min
Tasks/Time 1 / 00:00:00.308 5 min

Name Response
Queued 0 (0 bytes)

PutHDFS
PutHDFS 1.25.0
org.apache.nifi - nifi-hadoop-nar
In 1 (6.5 MB) 5 min
Read/Write 6.5 MB / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 1 / 00:00:02.613 5 min

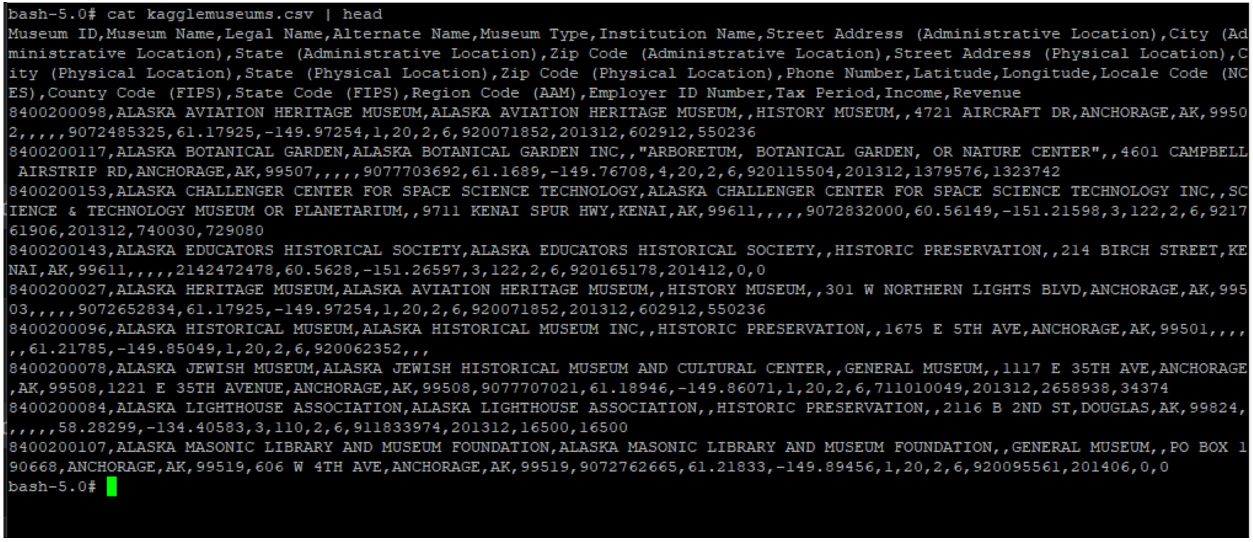
Name Failure, No Retry, Retry
Queued 0 (0 bytes)

Name failure
Queued 0 (0 bytes)

Hadoop HDFS – Data Storage

b. NiFi transferred the kagglemuseums.csv file to HDFS. View the first few rows.

```
cat kagglemuseums.csv | head
```



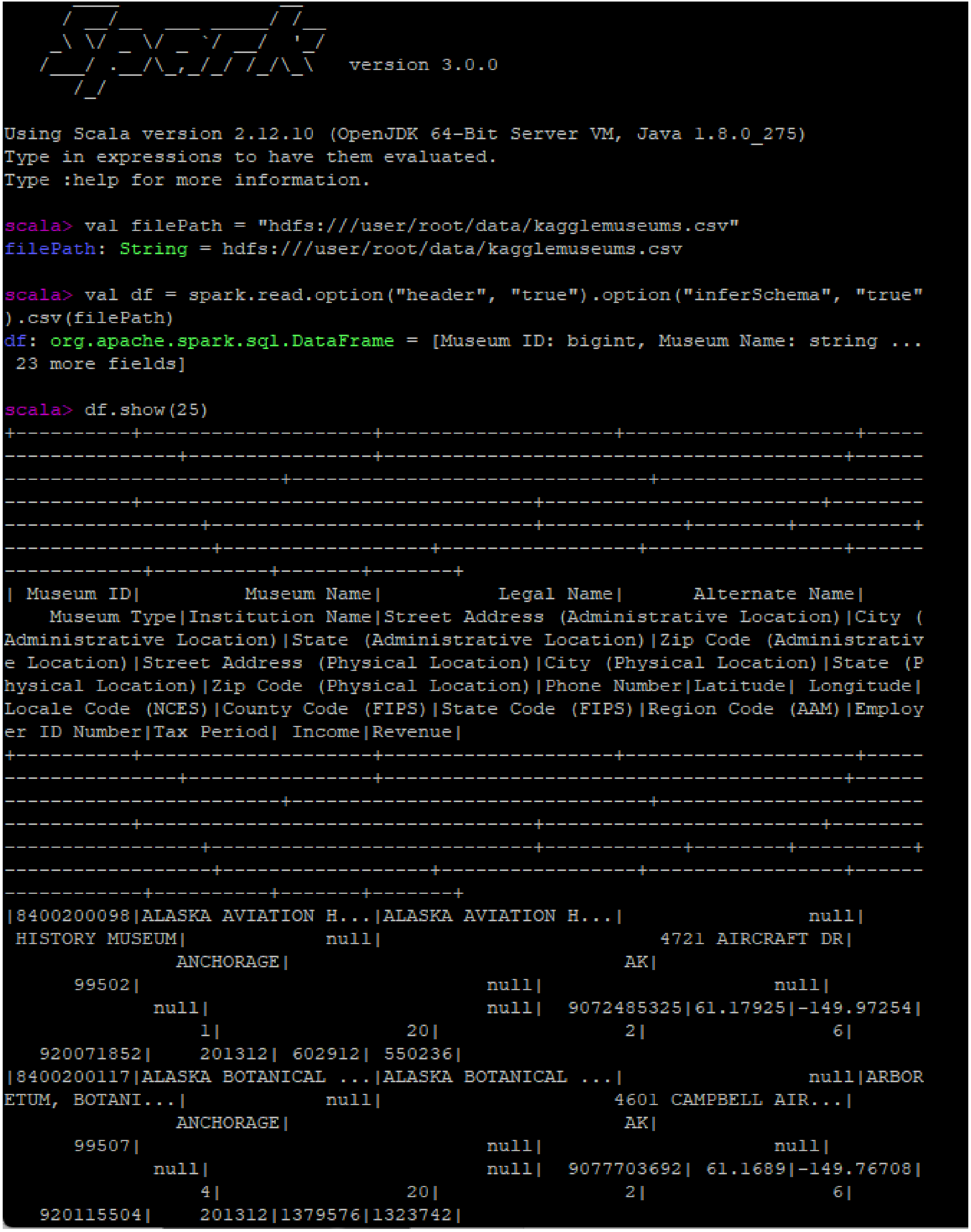
```
bash-5.0# cat kagglemuseums.csv | head
Museum ID,Museum Name,Legal Name,Alternate Name,Museum Type,Institution Name,Street Address (Administrative Location),City (Ad
ministrative Location),State (Administrative Location),Zip Code (Administrative Location),Street Address (Physical Location),C
ity (Physical Location),State (Physical Location),Zip Code (Physical Location),Phone Number,Latitude,Longitude,Locale Code (NC
ES),County Code (FIPS),State Code (FIPS),Region Code (AAM),Employer ID Number,Tax Period,Income,Revenue
8400200098,ALASKA AVIATION HERITAGE MUSEUM,ALASKA AVIATION HERITAGE MUSEUM,,HISTORY MUSEUM,,4721 AIRCRAFT DR,ANCHORAGE,AK,9950
2,,,,,9072485325,61.17925,-149.97254,1,20,2,6,920071852,201312,602912,550236
8400200117,ALASKA BOTANICAL GARDEN,ALASKA BOTANICAL GARDEN INC,,"ARBORETUM, BOTANICAL GARDEN, OR NATURE CENTER",,4601 CAMPBELL
 AIRSTRIP RD,ANCHORAGE,AK,99507,,,,,9077703692,61.1689,-149.76708,4,20,2,6,920115504,201312,1379576,1323742
8400200153,ALASKA CHALLENGER CENTER FOR SPACE SCIENCE TECHNOLOGY,ALASKA CHALLENGER CENTER FOR SPACE SCIENCE TECHNOLOGY INC,,SC
IENCE & TECHNOLOGY MUSEUM OR PLANETARIUM,,9711 KENAI SPUR HWY,KENAI,AK,99611,,,,,9072832000,60.56149,-151.21598,3,122,2,6,9217
61906,201312,740030,729080
8400200143,ALASKA EDUCATORS HISTORICAL SOCIETY,ALASKA EDUCATORS HISTORICAL SOCIETY,,HISTORIC PRESERVATION,,214 BIRCH STREET,KE
NAI,AK,99611,,,,,2142472478,60.5628,-151.26597,3,122,2,6,920165178,201412,0,0
8400200027,ALASKA HERITAGE MUSEUM,ALASKA AVIATION HERITAGE MUSEUM,,HISTORY MUSEUM,,301 W NORTHERN LIGHTS BLVD,ANCHORAGE,AK,995
03,,,,,9072652834,61.17925,-149.97254,1,20,2,6,920071852,201312,602912,550236
8400200096,ALASKA HISTORICAL MUSEUM,ALASKA HISTORICAL MUSEUM INC,,HISTORIC PRESERVATION,,1675 E 5TH AVE,ANCHORAGE,AK,99501,,,,
,,61.21785,-149.85049,1,20,2,6,920062352,,,
8400200078,ALASKA JEWISH MUSEUM,ALASKA JEWISH HISTORICAL MUSEUM AND CULTURAL CENTER,,GENERAL MUSEUM,,1117 E 35TH AVE,ANCHORAGE
,AK,99508,1221 E 35TH AVENUE,ANCHORAGE,AK,99508,9077707021,61.18946,-149.86071,1,20,2,6,711010049,201312,2658938,34374
8400200084,ALASKA LIGHTHOUSE ASSOCIATION,ALASKA LIGHTHOUSE ASSOCIATION,,HISTORIC PRESERVATION,,2116 B 2ND ST,DOUGLAS,AK,99824,
,,,,,58.28299,-134.40583,3,110,2,6,911833974,201312,16500,16500
8400200107,ALASKA MASONIC LIBRARY AND MUSEUM FOUNDATION,ALASKA MASONIC LIBRARY AND MUSEUM FOUNDATION,,GENERAL MUSEUM,,PO BOX 1
90668,ANCHORAGE,AK,99519,606 W 4TH AVE,ANCHORAGE,AK,99519,9072762665,61.21833,-149.89456,1,20,2,6,920095561,201406,0,0
bash-5.0# 
```

Spark Scala – Display Schema

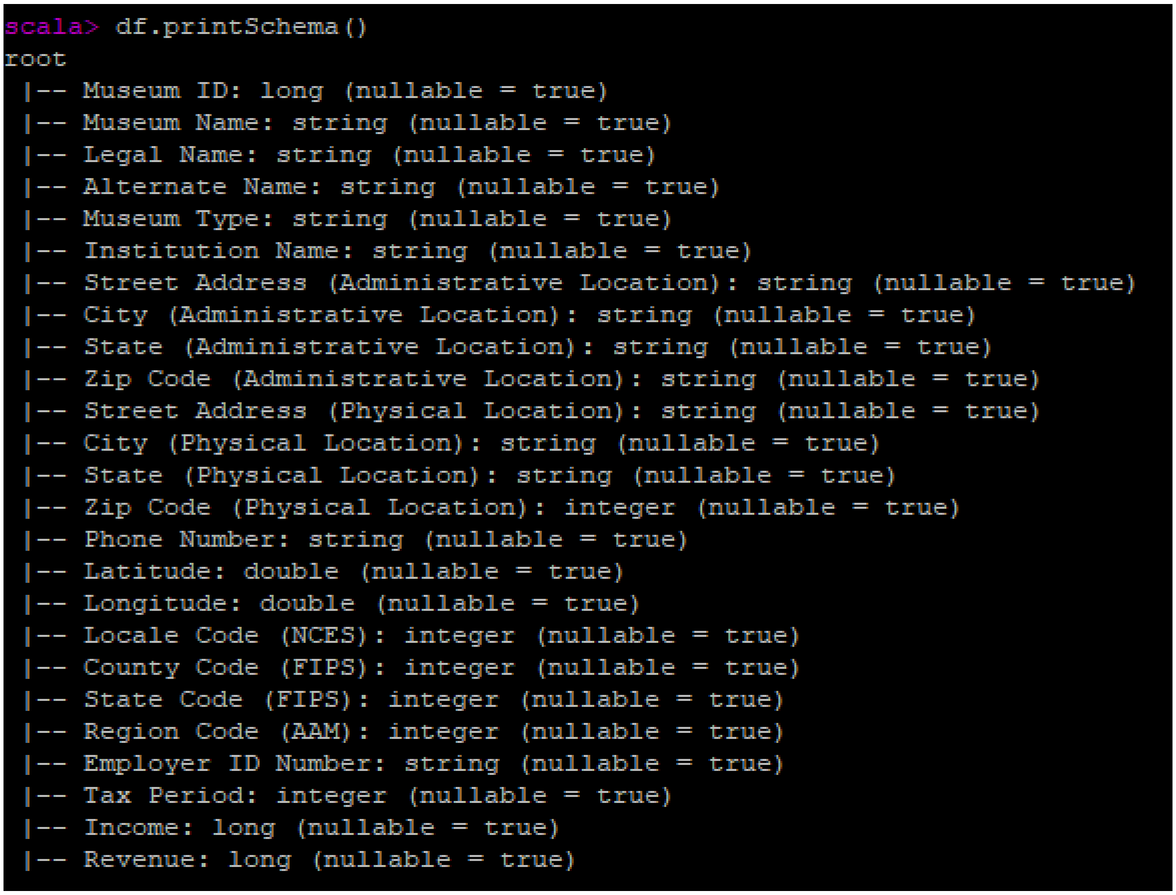
c. Preview the schema of the file using Spark Scala.

```
val filePath = "hdfs:///user/root/data/kagglemuseums.csv"

val df = spark.read.option("header",
"true").option("inferSchema", "true").csv(filePath)

df.show(25)
```

Benjamin Bartek
Bellevue University DSC 650
Final Report
November 16, 2024



```
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.0.0
      /_/

Using Scala version 2.12.10 (OpenJDK 64-Bit Server VM, Java 1.8.0_275)
Type in expressions to have them evaluated.
Type :help for more information.

scala> val filePath = "hdfs:///user/root/data/kagglemuseums.csv"
filePath: String = hdfs:///user/root/data/kagglemuseums.csv

scala> val df = spark.read.option("header", "true").option("inferSchema", "true"
).csv(filePath)
df: org.apache.spark.sql.DataFrame = [Museum ID: bigint, Museum Name: string ...
 23 more fields]

scala> df.show(25)
+----------+--------------------+--------------------+--------------------+-----
---------------+----------------+----------------------------------------+------
------------------------+-------------------------------+-----------------------
-----------+----------------------------------+------------------------+--------
-----------------+----------------------------+------------+--------+----------+
------------------+------------------+-----------------+-----------------+------
------------+----------+-------+-------+
| Museum ID|         Museum Name|          Legal Name|      Alternate Name|     
    Museum Type|Institution Name|Street Address (Administrative Location)|City (
Administrative Location)|State (Administrative Location)|Zip Code (Administrativ
e Location)|Street Address (Physical Location)|City (Physical Location)|State (P
hysical Location)|Zip Code (Physical Location)|Phone Number|Latitude| Longitude|
Locale Code (NCES)|County Code (FIPS)|State Code (FIPS)|Region Code (AAM)|Employ
er ID Number|Tax Period| Income|Revenue|
+----------+--------------------+--------------------+--------------------+-----
---------------+----------------+----------------------------------------+------
------------------------+-------------------------------+-----------------------
-----------+----------------------------------+------------------------+--------
-----------------+----------------------------+------------+--------+----------+
------------------+------------------+-----------------+-----------------+------
------------+----------+-------+-------+
|8400200098|ALASKA AVIATION H...|ALASKA AVIATION H...|                null|     
 HISTORY MUSEUM|            null|                        4721 AIRCRAFT DR|      
               ANCHORAGE|                             AK|                       
      99502|                              null|                    null|        
             null|                        null|  9072485325|61.17925|-149.97254|
                 1|                20|                2|                6|      
   920071852|    201312| 602912| 550236|
|8400200117|ALASKA BOTANICAL ...|ALASKA BOTANICAL ...|                null|ARBOR
ETUM, BOTANI...|            null|                    4601 CAMPBELL AIR...|      
               ANCHORAGE|                             AK|                       
      99507|                              null|                    null|        
             null|                        null|  9077703692| 61.1689|-149.76708|
                 4|                20|                2|                6|      
   920115504|    201312|1379576|1323742|
df.printSchema()
```

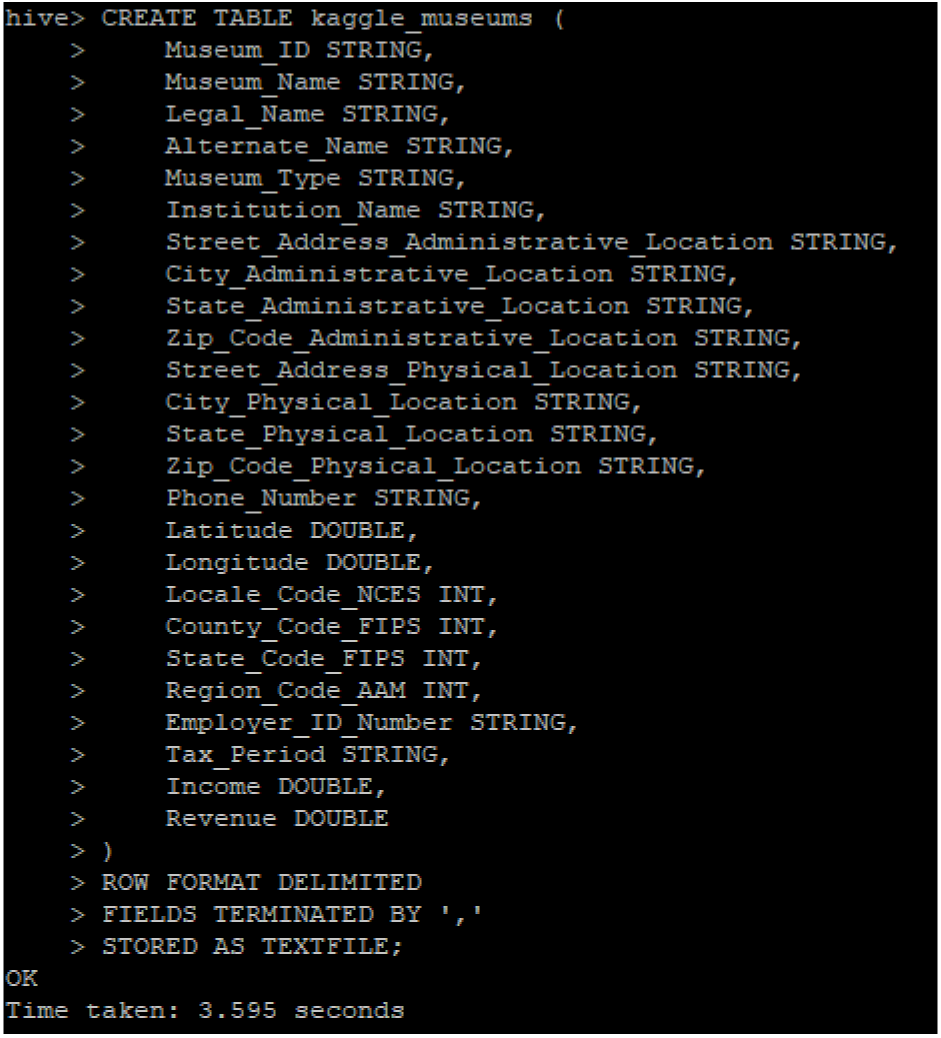


```
scala> df.printSchema()
root
 |-- Museum ID: long (nullable = true)
 |-- Museum Name: string (nullable = true)
 |-- Legal Name: string (nullable = true)
 |-- Alternate Name: string (nullable = true)
 |-- Museum Type: string (nullable = true)
 |-- Institution Name: string (nullable = true)
 |-- Street Address (Administrative Location): string (nullable = true)
 |-- City (Administrative Location): string (nullable = true)
 |-- State (Administrative Location): string (nullable = true)
 |-- Zip Code (Administrative Location): string (nullable = true)
 |-- Street Address (Physical Location): string (nullable = true)
 |-- City (Physical Location): string (nullable = true)
 |-- State (Physical Location): string (nullable = true)
 |-- Zip Code (Physical Location): integer (nullable = true)
 |-- Phone Number: string (nullable = true)
 |-- Latitude: double (nullable = true)
 |-- Longitude: double (nullable = true)
 |-- Locale Code (NCES): integer (nullable = true)
 |-- County Code (FIPS): integer (nullable = true)
 |-- State Code (FIPS): integer (nullable = true)
 |-- Region Code (AAM): integer (nullable = true)
 |-- Employer ID Number: string (nullable = true)
 |-- Tax Period: integer (nullable = true)
 |-- Income: long (nullable = true)
 |-- Revenue: long (nullable = true)
```

HIVE – Data Warehousing

d. Create a Hive Table.

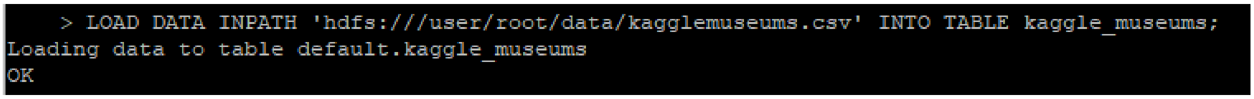
```
CREATE TABLE kaggle_museums (

Museum_ID STRING,

Museum_Name STRING,

Legal_Name STRING,

Alternate_Name STRING,

Museum_Type STRING,

Institution_Name STRING,

Street_Address_Administrative_Location STRING,

City_Administrative_Location STRING,
State_Administrative_Location STRING,
Zip_Code_Administrative_Location STRING,
Street_Address_Physical_Location STRING,
 City_Physical_Location STRING,
State_Physical_Location STRING,
Zip_Code_Physical_Location STRING,
Phone_Number STRING,
Latitude DOUBLE,
Longitude DOUBLE,
Locale_Code_NCES INT,
County_Code_FIPS INT,
State_Code_FIPS INT,
Region_Code_AAM INT,
Employer_ID_Number STRING,
Tax_Period STRING,
Income DOUBLE,
Revenue DOUBLE
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;
```



```
hive> CREATE TABLE kaggle_museums (
    >     Museum_ID STRING,
    >     Museum_Name STRING,
    >     Legal_Name STRING,
    >     Alternate_Name STRING,
    >     Museum_Type STRING,
    >     Institution_Name STRING,
    >     Street_Address_Administrative_Location STRING,
    >     City_Administrative_Location STRING,
    >     State_Administrative_Location STRING,
    >     Zip_Code_Administrative_Location STRING,
    >     Street_Address_Physical_Location STRING,
    >     City_Physical_Location STRING,
    >     State_Physical_Location STRING,
    >     Zip_Code_Physical_Location STRING,
    >     Phone_Number STRING,
    >     Latitude DOUBLE,
    >     Longitude DOUBLE,
    >     Locale_Code_NCES INT,
    >     County_Code_FIPS INT,
    >     State_Code_FIPS INT,
    >     Region_Code_AAM INT,
    >     Employer_ID_Number STRING,
    >     Tax_Period STRING,
    >     Income DOUBLE,
    >     Revenue DOUBLE
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
OK
Time taken: 3.595 seconds
```

e. **Load the data from the file in HDFS to the Hive table.**

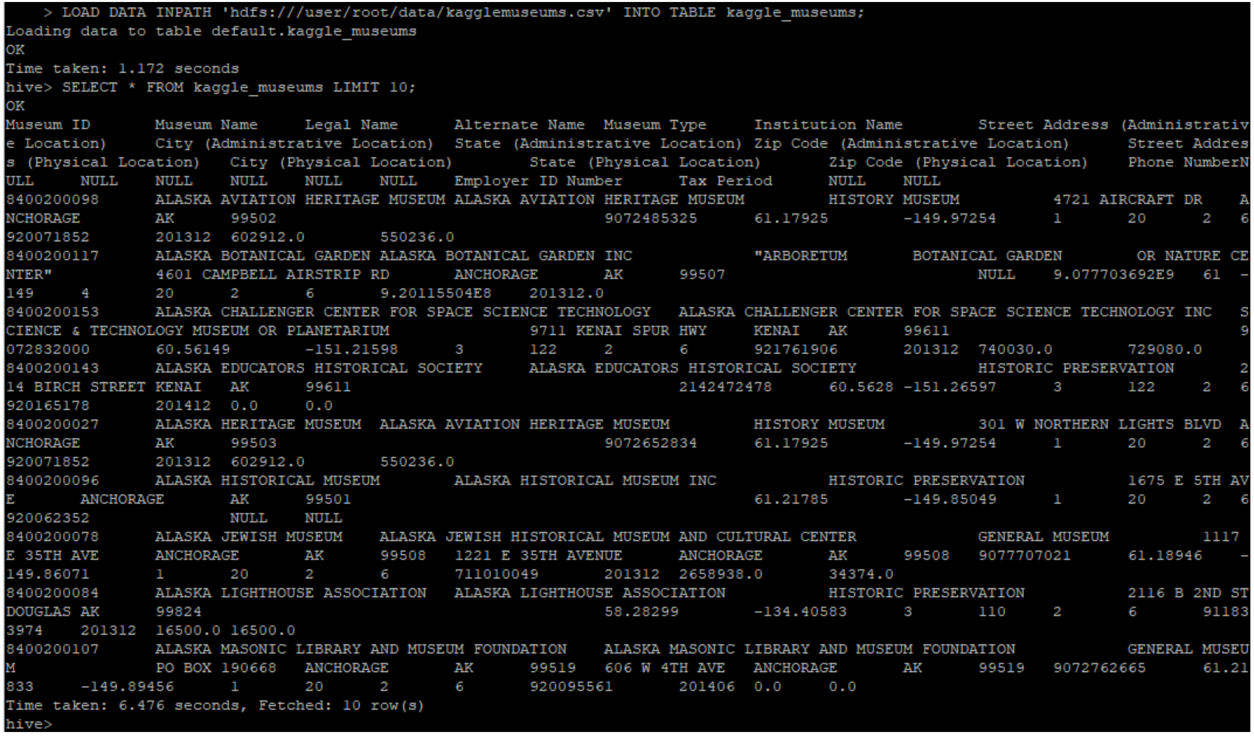
```
LOAD DATA INPATH 'hdfs:///user/root/data/kagglemuseums.csv'
INTO TABLE kaggle_museums;
```



```
    > LOAD DATA INPATH 'hdfs:///user/root/data/kagglemuseums.csv' INTO TABLE kaggle_museums;
Loading data to table default.kaggle_museums
OK
```

f. **Show the first 10 rows.**

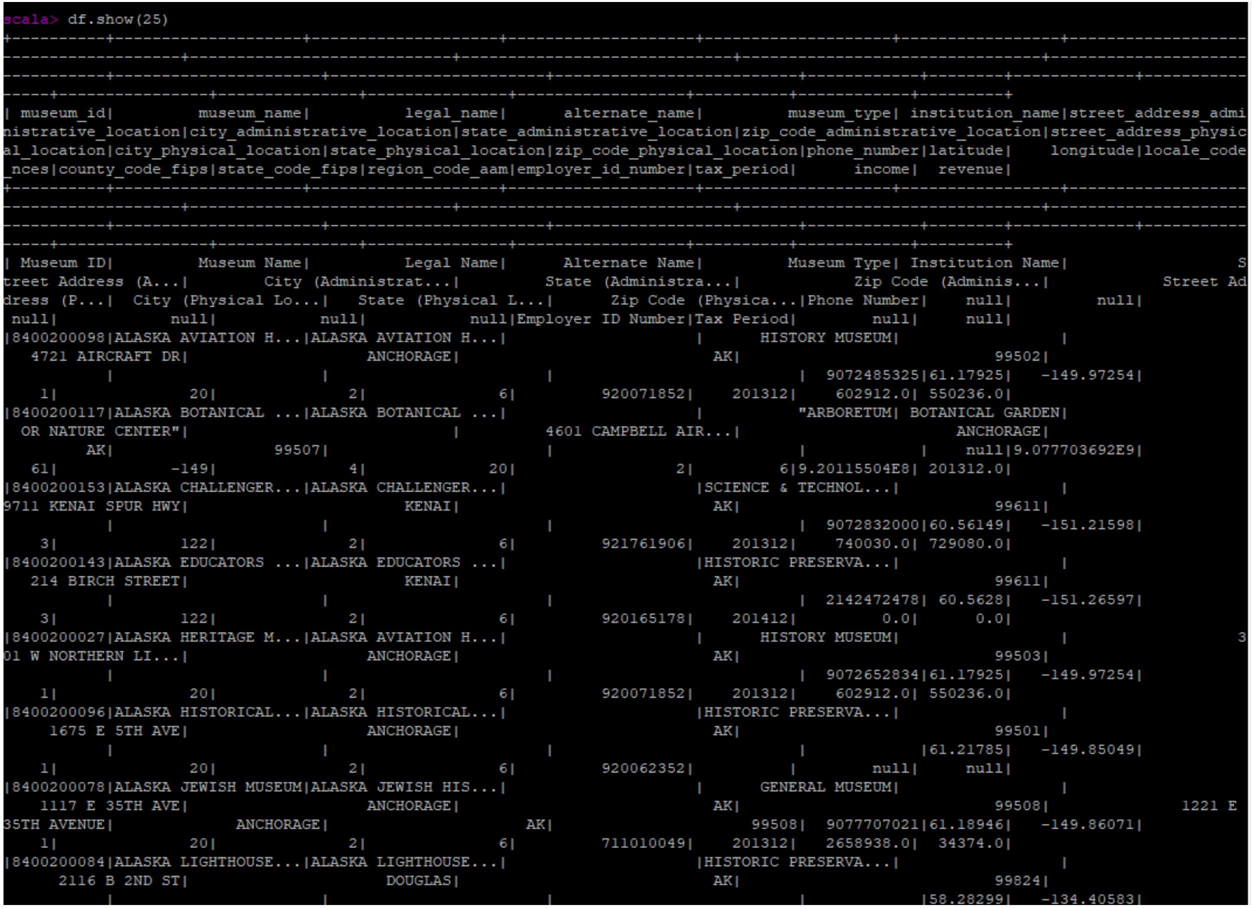
```
SELECT * FROM kaggle_museums LIMIT 10;
```



```
    > LOAD DATA INPATH 'hdfs:///user/root/data/kagglemuseums.csv' INTO TABLE kaggle_museums;
Loading data to table default.kaggle_museums
OK
Time taken: 1.172 seconds
hive> SELECT * FROM kaggle_museums LIMIT 10;
OK
Museum ID       Museum Name     Legal Name      Alternate Name  Museum Type     Institution Name        Street Address (Administrativ
e Location)     City (Administrative Location)  State (Administrative Location) Zip Code (Administrative Location)      Street Addres
s (Physical Location)   City (Physical Location)        State (Physical Location)       Zip Code (Physical Location)    Phone NumberN
ULL     NULL    NULL    NULL    NULL    NULL    Employer ID Number      Tax Period      NULL    NULL
8400200098      ALASKA AVIATION HERITAGE MUSEUM ALASKA AVIATION HERITAGE MUSEUM         HISTORY MUSEUM          4721 AIRCRAFT DR    A
NCHORAGE        AK      99502                                   9072485325      61.17925        -149.97254      1       20      2   6
920071852       201312  602912.0        550236.0
8400200117      ALASKA BOTANICAL GARDEN ALASKA BOTANICAL GARDEN INC             "ARBORETUM      BOTANICAL GARDEN        OR NATURE CE
NTER"           4601 CAMPBELL AIRSTRIP RD       ANCHORAGE       AK      99507                           NULL    9.077703692E9   61  -
149     4       20      2       6       9.20115504E8    201312.0
8400200153      ALASKA CHALLENGER CENTER FOR SPACE SCIENCE TECHNOLOGY   ALASKA CHALLENGER CENTER FOR SPACE SCIENCE TECHNOLOGY INC   S
CIENCE & TECHNOLOGY MUSEUM OR PLANETARIUM               9711 KENAI SPUR HWY     KENAI   AK      99611                               9
072832000       60.56149        -151.21598      3       122     2       6       921761906       201312  740030.0        729080.0
8400200143      ALASKA EDUCATORS HISTORICAL SOCIETY     ALASKA EDUCATORS HISTORICAL SOCIETY             HISTORIC PRESERVATION       2
14 BIRCH STREET KENAI   AK      99611                           2142472478      60.5628 -151.26597      3       122     2   6
920165178       201412  0.0     0.0
8400200027      ALASKA HERITAGE MUSEUM  ALASKA AVIATION HERITAGE MUSEUM         HISTORY MUSEUM          301 W NORTHERN LIGHTS BLVD  A
NCHORAGE        AK      99503                                   9072652834      61.17925        -149.97254      1       20      2   6
920071852       201312  602912.0        550236.0
8400200096      ALASKA HISTORICAL MUSEUM        ALASKA HISTORICAL MUSEUM INC            HISTORIC PRESERVATION           1675 E 5TH AV
E       ANCHORAGE       AK      99501                                   61.21785        -149.85049      1       20      2   6
920062352               NULL    NULL
8400200078      ALASKA JEWISH MUSEUM    ALASKA JEWISH HISTORICAL MUSEUM AND CULTURAL CENTER             GENERAL MUSEUM          1117
E 35TH AVE      ANCHORAGE       AK      99508   1221 E 35TH AVENUE      ANCHORAGE       AK      99508   9077707021      61.18946    -
149.86071       1       20      2       6       711010049       201312  2658938.0       34374.0
8400200084      ALASKA LIGHTHOUSE ASSOCIATION   ALASKA LIGHTHOUSE ASSOCIATION           HISTORIC PRESERVATION           2116 B 2ND ST
DOUGLAS AK      99824                                   58.28299        -134.40583      3       110     2       6       91183
3974    201312  16500.0 16500.0
8400200107      ALASKA MASONIC LIBRARY AND MUSEUM FOUNDATION    ALASKA MASONIC LIBRARY AND MUSEUM FOUNDATION            GENERAL MUSEU
M               PO BOX 190668   ANCHORAGE       AK      99519   606 W 4TH AVE   ANCHORAGE       AK      99519   9072762665      61.21
833     -149.89456      1       20      2       6       920095561       201406  0.0     0.0
Time taken: 6.476 seconds, Fetched: 10 row(s)
hive>
```

Spark Scala – Query and Transformation

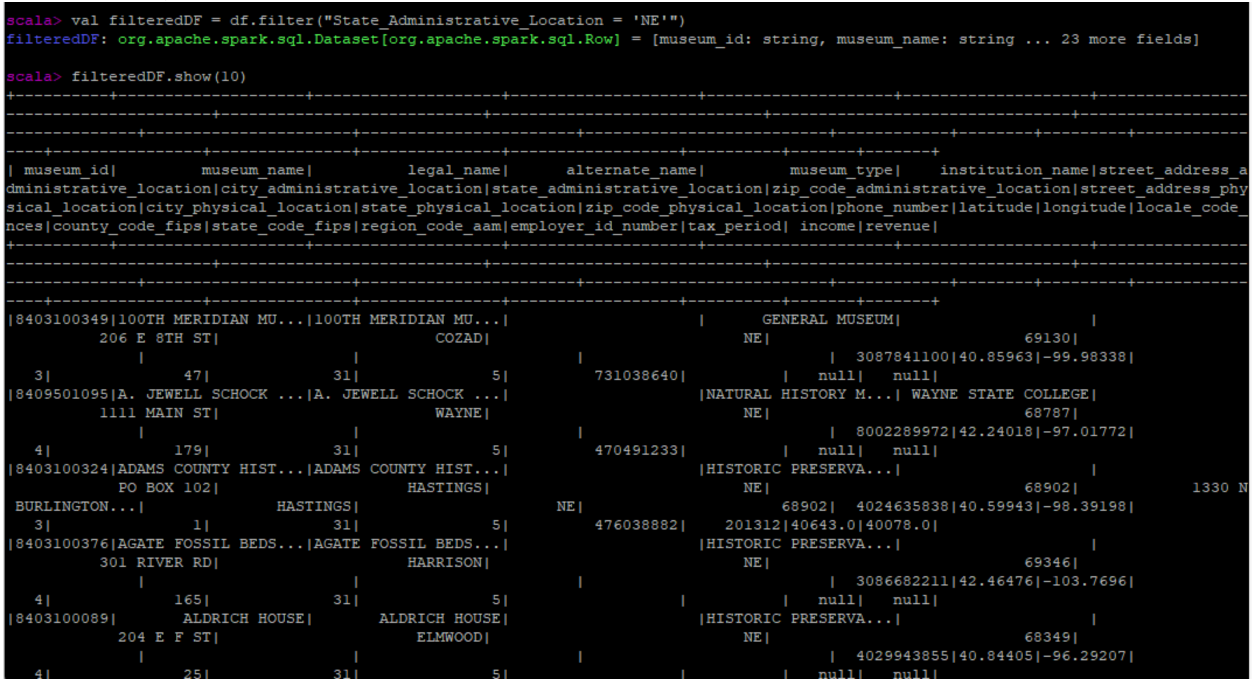
```
import org.apache.spark.sql.SparkSession

val spark = SparkSession.builder().appName("Hive
Integration").enableHiveSupport().getOrCreate()

val df = spark.sql("SELECT * FROM kaggle_museums")

df.show(25)
```



```
scala> df.show(25)
+----------+--------------------+--------------------+--------------------+--------------------+-----------------+-------------------
-------------------+----------------------------+-----------------------------+--------------------------------+---------------------
-----------+----------------------+-----------------------+--------------------------+------------+--------+-------------+-----------
-----+----------------+---------------+---------------+------------------+----------+------------+---------+
| museum_id|         museum_name|          legal_name|      alternate_name|         museum_type| institution_name|street_address_admi
nistrative_location|city_administrative_location|state_administrative_location|zip_code_administrative_location|street_address_physic
al_location|city_physical_location|state_physical_location|zip_code_physical_location|phone_number|latitude|    longitude|locale_code
_nces|county_code_fips|state_code_fips|region_code_aam|employer_id_number|tax_period|      income|  revenue|
+----------+--------------------+--------------------+--------------------+--------------------+-----------------+-------------------
-------------------+----------------------------+-----------------------------+--------------------------------+---------------------
-----------+----------------------+-----------------------+--------------------------+------------+--------+-------------+-----------
-----+----------------+---------------+---------------+------------------+----------+------------+---------+
| Museum ID|         Museum Name|          Legal Name|      Alternate Name|         Museum Type| Institution Name|                  S
treet Address (A...|        City (Administrat...|         State (Administra...|            Zip Code (Adminis...|            Street Ad
dress (P...|  City (Physical Lo...|   State (Physical L...|      Zip Code (Physica...|Phone Number|    null|         null|
 null|            null|           null|           null|Employer ID Number|Tax Period|        null|     null|
|8400200098|ALASKA AVIATION H...|ALASKA AVIATION H...|                    |      HISTORY MUSEUM|                 |
   4721 AIRCRAFT DR|                   ANCHORAGE|                           AK|                           99502|
           |                      |                       |                          |  9072485325|61.17925|   -149.97254|
    1|              20|              2|              6|         920071852|    201312|    602912.0| 550236.0|
|8400200117|ALASKA BOTANICAL ...|ALASKA BOTANICAL ...|                    |          "ARBORETUM| BOTANICAL GARDEN|
  OR NATURE CENTER"|                            |         4601 CAMPBELL AIR...|                       ANCHORAGE|
         AK|                 99507|                       |                          |            |    null|9.077703692E9|
   61|            -149|              4|             20|                 2|         6|9.20115504E8| 201312.0|
|8400200153|ALASKA CHALLENGER...|ALASKA CHALLENGER...|                    |SCIENCE & TECHNOL...|                 |
9711 KENAI SPUR HWY|                       KENAI|                           AK|                           99611|
           |                      |                       |                          |  9072832000|60.56149|   -151.21598|
    3|             122|              2|              6|         921761906|    201312|    740030.0| 729080.0|
|8400200143|ALASKA EDUCATORS ...|ALASKA EDUCATORS ...|                    |HISTORIC PRESERVA...|                 |
   214 BIRCH STREET|                       KENAI|                           AK|                           99611|
           |                      |                       |                          |  2142472478| 60.5628|   -151.26597|
    3|             122|              2|              6|         920165178|    201412|         0.0|      0.0|
|8400200027|ALASKA HERITAGE M...|ALASKA AVIATION H...|                    |      HISTORY MUSEUM|                 |                  3
01 W NORTHERN LI...|                   ANCHORAGE|                           AK|                           99503|
           |                      |                       |                          |  9072652834|61.17925|   -149.97254|
    1|              20|              2|              6|         920071852|    201312|    602912.0| 550236.0|
|8400200096|ALASKA HISTORICAL...|ALASKA HISTORICAL...|                    |HISTORIC PRESERVA...|                 |
     1675 E 5TH AVE|                   ANCHORAGE|                           AK|                           99501|
           |                      |                       |                          |            |61.21785|   -149.85049|
    1|              20|              2|              6|         920062352|          |        null|     null|
|8400200078|ALASKA JEWISH MUSEUM|ALASKA JEWISH HIS...|                    |      GENERAL MUSEUM|                 |
    1117 E 35TH AVE|                   ANCHORAGE|                           AK|                           99508|              1221 E
35TH AVENUE|             ANCHORAGE|                     AK|                     99508|  9077707021|61.18946|   -149.86071|
    1|              20|              2|              6|         711010049|    201312|   2658938.0|  34374.0|
|8400200084|ALASKA LIGHTHOUSE...|ALASKA LIGHTHOUSE...|                    |HISTORIC PRESERVA...|                 |
      2116 B 2ND ST|                     DOUGLAS|                           AK|                           99824|
           |                      |                       |                          |            |58.28299|   -134.40583|
```

g. **Filter Data by State.**

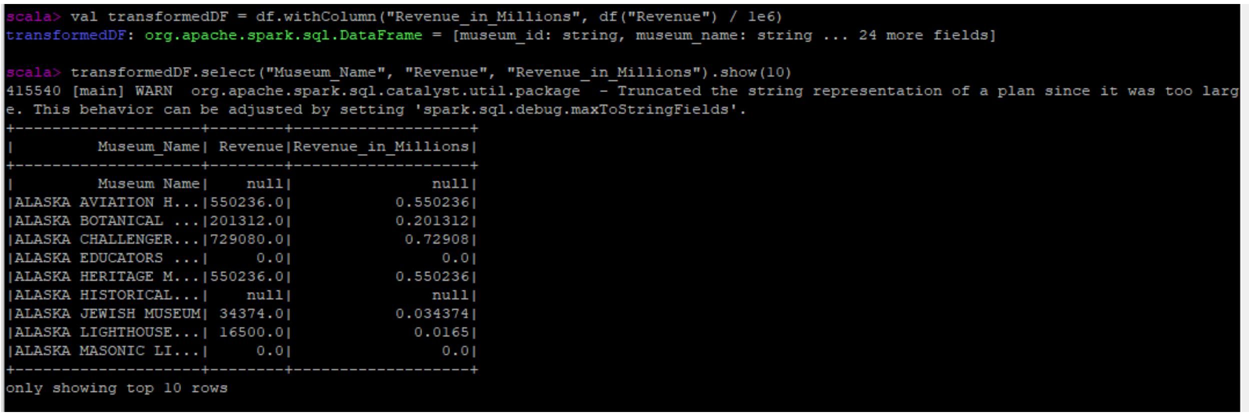
```
val filteredDF = df.filter("State_Administrative_Location =
'NE'")

filteredDF.show(10)
```



```
scala> val filteredDF = df.filter("State_Administrative_Location = 'NE'")
filteredDF: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [museum_id: string, museum_name: string ... 23 more fields]

scala> filteredDF.show(10)
+----------+--------------------+--------------------+--------------------+--------------------+--------------------+----------------
----------------------+----------------------------+-----------------------------+-------------------------------+------------------
--------------+----------------------+-----------------------+--------------------------+------------+--------+---------+------------
----+----------------+---------------+---------------+------------------+----------+-------+-------+
| museum_id|         museum_name|          legal_name|      alternate_name|         museum_type|    institution_name|street_address_a
dministrative_location|city_administrative_location|state_administrative_location|zip_code_administrative_location|street_address_phy
sical_location|city_physical_location|state_physical_location|zip_code_physical_location|phone_number|latitude|longitude|locale_code_
nces|county_code_fips|state_code_fips|region_code_aam|employer_id_number|tax_period| income|revenue|
+----------+--------------------+--------------------+--------------------+--------------------+--------------------+----------------
----------------------+----------------------------+-----------------------------+-------------------------------+------------------
--------------+----------------------+-----------------------+--------------------------+------------+--------+---------+------------
----+----------------+---------------+---------------+------------------+----------+-------+-------+
|8403100349|100TH MERIDIAN MU...|100TH MERIDIAN MU...|                    |      GENERAL MUSEUM|                    |
          206 E 8TH ST|                       COZAD|                           NE|                          69130|
              |                      |                       |                          |  3087841100|40.85963|-99.98338|
   3|              47|             31|              5|         731038640|          |   null|   null|
|8409501095|A. JEWELL SCHOCK ...|A. JEWELL SCHOCK ...|                    |NATURAL HISTORY M...| WAYNE STATE COLLEGE|
          1111 MAIN ST|                       WAYNE|                           NE|                          68787|
              |                      |                       |                          |  8002289972|42.24018|-97.01772|
   4|             179|             31|              5|         470491233|          |   null|   null|
|8403100324|ADAMS COUNTY HIST...|ADAMS COUNTY HIST...|                    |HISTORIC PRESERVA...|                    |
            PO BOX 102|                    HASTINGS|                           NE|                          68902|            1330 N
 BURLINGTON...|              HASTINGS|                     NE|                     68902|  4024635838|40.59943|-98.39198|
   3|               1|             31|              5|         476038882|    201312|40643.0|40078.0|
|8403100376|AGATE FOSSIL BEDS...|AGATE FOSSIL BEDS...|                    |HISTORIC PRESERVA...|                    |
          301 RIVER RD|                    HARRISON|                           NE|                          69346|
              |                      |                       |                          |  3086682211|42.46476|-103.7696|
   4|             165|             31|              5|                  |          |   null|   null|
|8403100089|       ALDRICH HOUSE|       ALDRICH HOUSE|                    |HISTORIC PRESERVA...|                    |
            204 E F ST|                     ELMWOOD|                           NE|                          68349|
              |                      |                       |                          |  4029943855|40.84405|-96.29207|
   4|              25|             31|              5|                  |          |   null|   null|
```

h. Add a New Column for Revenue in Millions.

```
val transformedDF = df.withColumn("Revenue_in_Millions",
df("Revenue") / 1e6)

transformedDF.select("Museum_Name", "Revenue",
"Revenue_in_Millions").show(10)
```

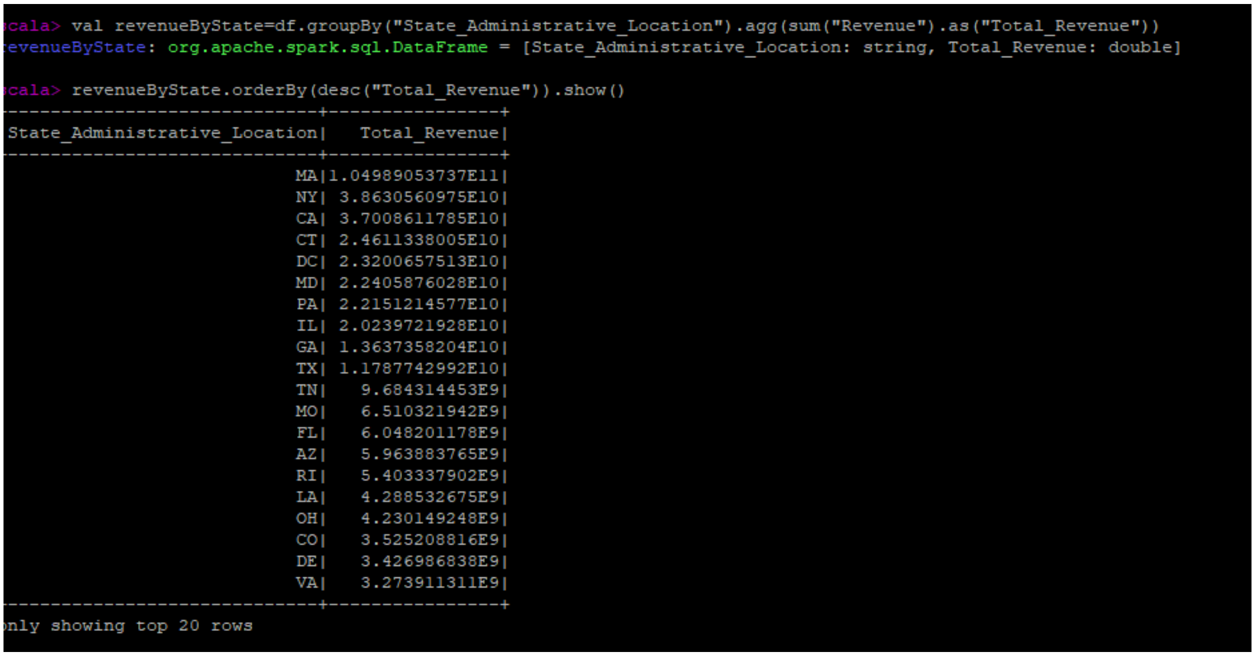


```
scala> val transformedDF = df.withColumn("Revenue_in_Millions", df("Revenue") / 1e6)
transformedDF: org.apache.spark.sql.DataFrame = [museum_id: string, museum_name: string ... 24 more fields]

scala> transformedDF.select("Museum_Name", "Revenue", "Revenue_in_Millions").show(10)
415540 [main] WARN  org.apache.spark.sql.catalyst.util.package  - Truncated the string representation of a plan since it was too larg
e. This behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.
+--------------------+--------+-------------------+
|         Museum_Name| Revenue|Revenue_in_Millions|
+--------------------+--------+-------------------+
|         Museum Name|    null|               null|
|ALASKA AVIATION H...|550236.0|           0.550236|
|ALASKA BOTANICAL ...|201312.0|           0.201312|
|ALASKA CHALLENGER...|729080.0|            0.72908|
|ALASKA EDUCATORS ...|     0.0|                0.0|
|ALASKA HERITAGE M...|550236.0|           0.550236|
|ALASKA HISTORICAL...|    null|               null|
|ALASKA JEWISH MUSEUM| 34374.0|           0.034374|
|ALASKA LIGHTHOUSE...| 16500.0|             0.0165|
|ALASKA MASONIC LI...|     0.0|                0.0|
+--------------------+--------+-------------------+
only showing top 10 rows
```

i. Group and Aggregate Data by State.

val revenueByState=df.groupBy("State_Administrative_Location").agg(sum("Revenue").as("Total_Revenue"))

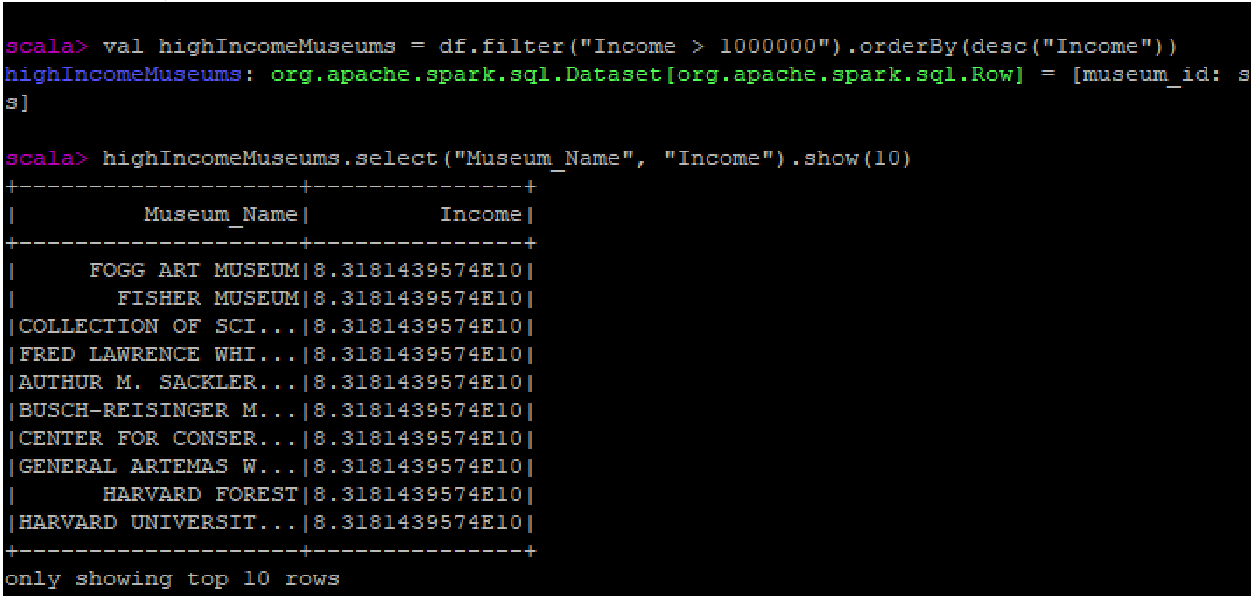
```
revenueByState.orderBy(desc("Total_Revenue")).show()
```



```
cala> val revenueByState=df.groupBy("State_Administrative_Location").agg(sum("Revenue").as("Total_Revenue"))
evenueByState: org.apache.spark.sql.DataFrame = [State_Administrative_Location: string, Total_Revenue: double]

cala> revenueByState.orderBy(desc("Total_Revenue")).show()
------------------------------+----------------+
 State_Administrative_Location|   Total_Revenue|
------------------------------+----------------+
                            MA|1.04989053737E11|
                            NY| 3.8630560975E10|
                            CA| 3.7008611785E10|
                            CT| 2.4611338005E10|
                            DC| 2.3200657513E10|
                            MD| 2.2405876028E10|
                            PA| 2.2151214577E10|
                            IL| 2.0239721928E10|
                            GA| 1.3637358204E10|
                            TX| 1.1787742992E10|
                            TN|   9.684314453E9|
                            MO|   6.510321942E9|
                            FL|   6.048201178E9|
                            AZ|   5.963883765E9|
                            RI|   5.403337902E9|
                            LA|   4.288532675E9|
                            OH|   4.230149248E9|
                            CO|   3.525208816E9|
                            DE|   3.426986838E9|
                            VA|   3.273911311E9|
------------------------------+----------------+
nly showing top 20 rows
```

j. Filter and Sort by Income.

```
val highIncomeMuseums = df.filter("Income > 1000000").orderBy(desc("Income"))

highIncomeMuseums.select("Museum_Name", "Income").show(10)
```

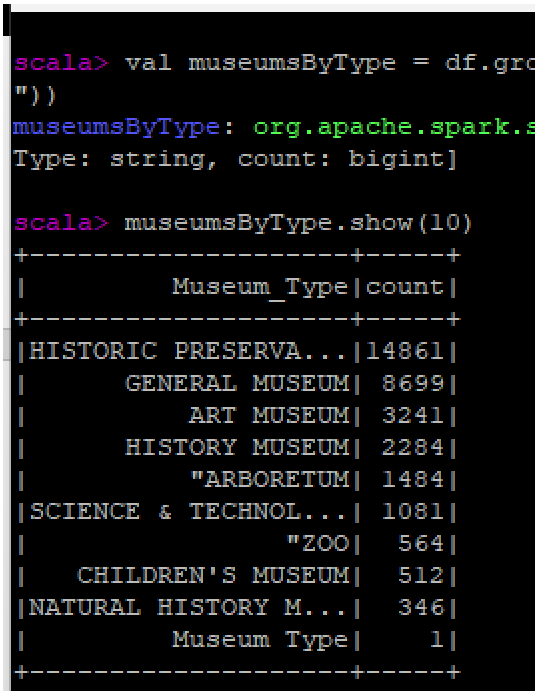


```
scala> val highIncomeMuseums = df.filter("Income > 1000000").orderBy(desc("Income"))
highIncomeMuseums: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [museum_id: s
s]

scala> highIncomeMuseums.select("Museum_Name", "Income").show(10)
+--------------------+---------------+
|         Museum_Name|         Income|
+--------------------+---------------+
|     FOGG ART MUSEUM|8.3181439574E10|
|       FISHER MUSEUM|8.3181439574E10|
|COLLECTION OF SCI...|8.3181439574E10|
|FRED LAWRENCE WHI...|8.3181439574E10|
|AUTHUR M. SACKLER...|8.3181439574E10|
|BUSCH-REISINGER M...|8.3181439574E10|
|CENTER FOR CONSER...|8.3181439574E10|
|GENERAL ARTEMAS W...|8.3181439574E10|
|      HARVARD FOREST|8.3181439574E10|
|HARVARD UNIVERSIT...|8.3181439574E10|
+--------------------+---------------+
only showing top 10 rows
```

k. Count the number of museums by type.

val museumsByType = df.groupBy("Museum_Type").count().orderBy(desc("count"))

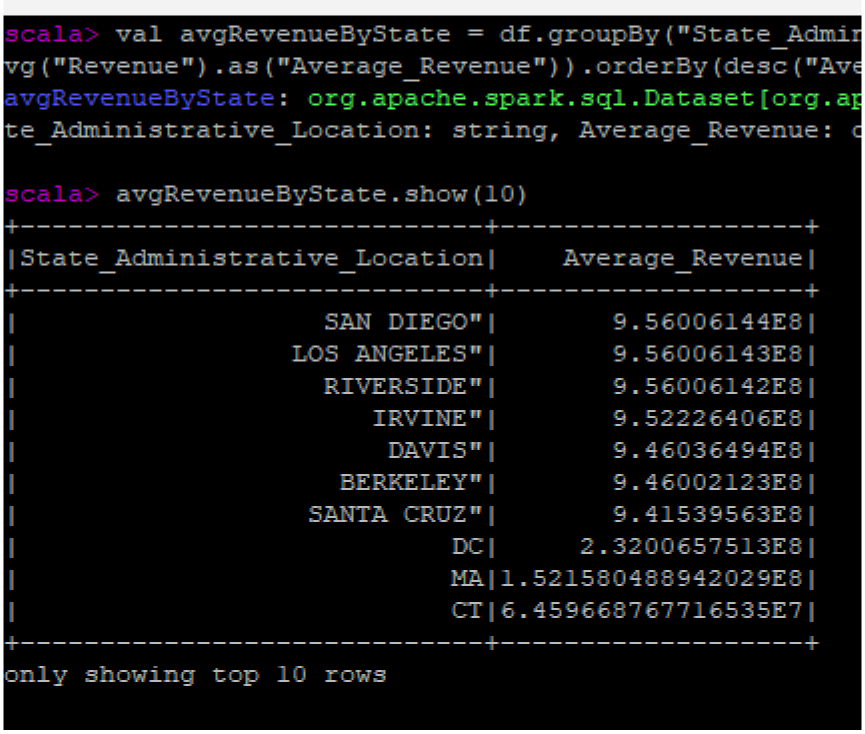
museumsByType.show(10)



```
scala> val museumsByType = df.gro
"))
museumsByType: org.apache.spark.s
Type: string, count: bigint]

scala> museumsByType.show(10)
+--------------------+-----+
|         Museum_Type|count|
+--------------------+-----+
|HISTORIC PRESERVA...|14861|
|      GENERAL MUSEUM| 8699|
|          ART MUSEUM| 3241|
|      HISTORY MUSEUM| 2284|
|          "ARBORETUM| 1484|
|SCIENCE & TECHNOL...| 1081|
|                "ZOO|  564|
|   CHILDREN'S MUSEUM|  512|
|NATURAL HISTORY M...|  346|
|         Museum Type|    1|
+--------------------+-----+
```

l. Calculate Average Revenue by State.

```
val avgRevenueByState =
df.groupBy("State_Administrative_Location").agg(avg("Revenue").as("Average_Revenue"
)).orderBy(desc("Average_Revenue"))

avgRevenueByState.show(10)
```



```
scala> val avgRevenueByState = df.groupBy("State_Admi
vg("Revenue").as("Average_Revenue")).orderBy(desc("Ave
avgRevenueByState: org.apache.spark.sql.Dataset[org.ap
te_Administrative_Location: string, Average_Revenue:

scala> avgRevenueByState.show(10)
+-----------------------------+-------------------+
|State_Administrative_Location|    Average_Revenue|
+-----------------------------+-------------------+
|                   SAN DIEGO"|       9.56006144E8|
|                 LOS ANGELES"|       9.56006143E8|
|                   RIVERSIDE"|       9.56006142E8|
|                      IRVINE"|       9.52226406E8|
|                       DAVIS"|       9.46036494E8|
|                    BERKELEY"|       9.46002123E8|
|                  SANTA CRUZ"|       9.41539563E8|
|                           DC|     2.3200657513E8|
|                           MA|1.521580488942029E8|
|                           CT|6.459668767716535E7|
+-----------------------------+-------------------+
only showing top 10 rows
```

6. Conclusion

This project demonstrates the design and implementation of a data pipeline using Apache NiFi, Hadoop HDFS, Apache Hive, and Apache Spark to ingest, store, query, and transform data from a single data set. While some challenges added to the time it took to complete the project, those challenges provided additional learning opportunities for troubleshooting and gaining a greater understanding of how these tools work. This pipeline is a small-scale example of the power and flexibility of some common big data tools that really demonstrate their power and scalability when dealing with large, complex data sets.

References

1. Institute of Museum and Library Services (IMLS) and Abigail Larion. *Museums, Aquariums, and Zoos*. Kaggle, n.d. Web. Accessed 16 Nov. 2024. https://www.kaggle.com/datasets/imls/museum-directory.