Practice3

March 25, 2021

In [1]:
```
#%run -i 'etl.py'
```

In [2]:
```
import configparser
from datetime import datetime
import os
from pyspark.sql import SparkSession
from pyspark.sql.functions import udf, col
from pyspark.sql.functions import year, month, dayofmonth, hour, weekofyear, date_format


config = configparser.ConfigParser()
config.read('pp_test.cfg')

os.environ['AWS_ACCESS_KEY_ID']=config['AWS']['AWS_ACCESS_KEY_ID']
os.environ['AWS_SECRET_ACCESS_KEY']=config['AWS']['AWS_SECRET_ACCESS_KEY']


spark = SparkSession \
        .builder \
        .config("spark.jars.packages", "org.apache.hadoop:hadoop-aws:2.7.0") \
        .getOrCreate()
```

In [3]:
```
input_data = "s3a://udacity-dend/song_data/A/B/C/"
#input_data = "s3a://datalakepp/song_data" # my bucket

song_data = input_data + "*.json"
#song_data = "s3a://udacity-dend/song_data/A/B/C/TRABCEI128F424C983.json"
```

In [4]:
```
# use this to speed up parquet write
sc = spark.sparkContext
sc._jsc.hadoopConfiguration().set("mapreduce.fileoutputcommitter.algorithm.version", "2"
```

In [5]:
```
df = spark.read.json(song_data)
```

In [6]:
```
!pyspark --packages com.amazonaws:aws-java-sdk-pom:1.10.34,org.apache.hadoop:hadoop-aws:
```

```
Python 3.6.3 | packaged by conda-forge | (default, Dec  9 2017, 04:28:46)
[GCC 4.8.2 20140120 (Red Hat 4.8.2-15)] on linux
```

```
Type "help", "copyright", "credits" or "license" for more information.
Ivy Default Cache set to: /root/.ivy2/cache
The jars for the packages stored in: /root/.ivy2/jars
:: loading settings :: url = jar:file:/opt/spark-2.4.3-bin-hadoop2.7/jars/ivy-2.4.0.jar!/org/apa
com.amazonaws#aws-java-sdk-pom added as a dependency
org.apache.hadoop#hadoop-aws added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-a3bb0454-a5bc-4083-82c6-ea5aa9
        confs: [default]
        found com.amazonaws#aws-java-sdk-pom;1.10.34 in central
        found org.apache.hadoop#hadoop-aws;2.7.2 in central
        found org.apache.hadoop#hadoop-common;2.7.2 in central
        found org.apache.hadoop#hadoop-annotations;2.7.2 in central
        found com.google.guava#guava;11.0.2 in central
        found com.google.code.findbugs#jsr305;3.0.0 in central
        found commons-cli#commons-cli;1.2 in central
        found org.apache.commons#commons-math3;3.1.1 in central
        found xmlenc#xmlenc;0.52 in central
        found commons-httpclient#commons-httpclient;3.1 in central
        found commons-logging#commons-logging;1.1.3 in central
        found commons-codec#commons-codec;1.4 in central
        found commons-io#commons-io;2.4 in central
        found commons-net#commons-net;3.1 in central
        found commons-collections#commons-collections;3.2.2 in central
        found javax.servlet#servlet-api;2.5 in central
        found org.mortbay.jetty#jetty;6.1.26 in central
        found org.mortbay.jetty#jetty-util;6.1.26 in central
        found com.sun.jersey#jersey-core;1.9 in central
        found com.sun.jersey#jersey-json;1.9 in central
        found org.codehaus.jettison#jettison;1.1 in central
        found com.sun.xml.bind#jaxb-impl;2.2.3-1 in central
        found javax.xml.bind#jaxb-api;2.2.2 in central
        found javax.xml.stream#stax-api;1.0-2 in central
        found javax.activation#activation;1.1 in central
        found org.codehaus.jackson#jackson-core-asl;1.9.13 in central
        found org.codehaus.jackson#jackson-mapper-asl;1.9.13 in central
        found org.codehaus.jackson#jackson-jaxrs;1.9.13 in central
        found org.codehaus.jackson#jackson-xc;1.9.13 in central
        found com.sun.jersey#jersey-server;1.9 in central
        found asm#asm;3.2 in central
        found log4j#log4j;1.2.17 in central
        found net.java.dev.jets3t#jets3t;0.9.0 in central
        found org.apache.httpcomponents#httpclient;4.2.5 in central
        found org.apache.httpcomponents#httpcore;4.2.5 in central
        found com.jamesmurty.utils#java-xmlbuilder;0.4 in central
        found commons-lang#commons-lang;2.6 in central
        found commons-configuration#commons-configuration;1.6 in central
        found commons-digester#commons-digester;1.8 in central
        found commons-beanutils#commons-beanutils;1.7.0 in central
```

```
        found commons-beanutils#commons-beanutils-core;1.8.0 in central
        found org.slf4j#slf4j-api;1.7.10 in central
        found org.apache.avro#avro;1.7.4 in central
        found com.thoughtworks.paranamer#paranamer;2.3 in central
        found org.xerial.snappy#snappy-java;1.0.4.1 in central
        found org.apache.commons#commons-compress;1.4.1 in central
        found org.tukaani#xz;1.0 in central
        found com.google.protobuf#protobuf-java;2.5.0 in central
        found com.google.code.gson#gson;2.2.4 in central
        found org.apache.hadoop#hadoop-auth;2.7.2 in central
        found org.apache.directory.server#apacheds-kerberos-codec;2.0.0-M15 in central
        found org.apache.directory.server#apacheds-i18n;2.0.0-M15 in central
        found org.apache.directory.api#api-asn1-api;1.0.0-M20 in central
        found org.apache.directory.api#api-util;1.0.0-M20 in central
        found org.apache.zookeeper#zookeeper;3.4.6 in central
        found org.slf4j#slf4j-log4j12;1.7.10 in central
        found io.netty#netty;3.6.2.Final in central
        found org.apache.curator#curator-framework;2.7.1 in central
        found org.apache.curator#curator-client;2.7.1 in central
        found com.jcraft#jsch;0.1.42 in central
        found org.apache.curator#curator-recipes;2.7.1 in central
        found org.apache.htrace#htrace-core;3.1.0-incubating in central
        found javax.servlet.jsp#jsp-api;2.1 in central
        found jline#jline;0.9.94 in central
        found junit#junit;4.11 in central
        found org.hamcrest#hamcrest-core;1.3 in central
        found com.fasterxml.jackson.core#jackson-databind;2.2.3 in central
        found com.fasterxml.jackson.core#jackson-annotations;2.2.3 in central
        found com.fasterxml.jackson.core#jackson-core;2.2.3 in central
        found com.amazonaws#aws-java-sdk;1.7.4 in central
        found joda-time#joda-time;2.10.10 in central
        [2.10.10] joda-time#joda-time;[2.2,)
downloading https://repo1.maven.org/maven2/org/apache/hadoop/hadoop-aws/2.7.2/hadoop-aws-2.7.2.j
        [SUCCESSFUL ] org.apache.hadoop#hadoop-aws;2.7.2!hadoop-aws.jar (42ms)
downloading https://repo1.maven.org/maven2/org/apache/hadoop/hadoop-common/2.7.2/hadoop-common-2
        [SUCCESSFUL ] org.apache.hadoop#hadoop-common;2.7.2!hadoop-common.jar (317ms)
downloading https://repo1.maven.org/maven2/org/apache/hadoop/hadoop-annotations/2.7.2/hadoop-ann
        [SUCCESSFUL ] org.apache.hadoop#hadoop-annotations;2.7.2!hadoop-annotations.jar (25ms)
downloading https://repo1.maven.org/maven2/commons-collections/commons-collections/3.2.2/commons
        [SUCCESSFUL ] commons-collections#commons-collections;3.2.2!commons-collections.jar (59m
downloading https://repo1.maven.org/maven2/org/apache/hadoop/hadoop-auth/2.7.2/hadoop-auth-2.7.2
        [SUCCESSFUL ] org.apache.hadoop#hadoop-auth;2.7.2!hadoop-auth.jar (26ms)
:: resolution report :: resolve 7381ms :: artifacts dl 545ms
        :: modules in use:
        asm#asm;3.2 from central in [default]
        com.amazonaws#aws-java-sdk;1.7.4 from central in [default]
        com.amazonaws#aws-java-sdk-pom;1.10.34 from central in [default]
        com.fasterxml.jackson.core#jackson-annotations;2.2.3 from central in [default]
```

```
com.fasterxml.jackson.core#jackson-core;2.2.3 from central in [default]
com.fasterxml.jackson.core#jackson-databind;2.2.3 from central in [default]
com.google.code.findbugs#jsr305;3.0.0 from central in [default]
com.google.code.gson#gson;2.2.4 from central in [default]
com.google.guava#guava;11.0.2 from central in [default]
com.google.protobuf#protobuf-java;2.5.0 from central in [default]
com.jamesmurty.utils#java-xmlbuilder;0.4 from central in [default]
com.jcraft#jsch;0.1.42 from central in [default]
com.sun.jersey#jersey-core;1.9 from central in [default]
com.sun.jersey#jersey-json;1.9 from central in [default]
com.sun.jersey#jersey-server;1.9 from central in [default]
com.sun.xml.bind#jaxb-impl;2.2.3-1 from central in [default]
com.thoughtworks.paranamer#paranamer;2.3 from central in [default]
commons-beanutils#commons-beanutils;1.7.0 from central in [default]
commons-beanutils#commons-beanutils-core;1.8.0 from central in [default]
commons-cli#commons-cli;1.2 from central in [default]
commons-codec#commons-codec;1.4 from central in [default]
commons-collections#commons-collections;3.2.2 from central in [default]
commons-configuration#commons-configuration;1.6 from central in [default]
commons-digester#commons-digester;1.8 from central in [default]
commons-httpclient#commons-httpclient;3.1 from central in [default]
commons-io#commons-io;2.4 from central in [default]
commons-lang#commons-lang;2.6 from central in [default]
commons-logging#commons-logging;1.1.3 from central in [default]
commons-net#commons-net;3.1 from central in [default]
io.netty#netty;3.6.2.Final from central in [default]
javax.activation#activation;1.1 from central in [default]
javax.servlet#servlet-api;2.5 from central in [default]
javax.servlet.jsp#jsp-api;2.1 from central in [default]
javax.xml.bind#jaxb-api;2.2.2 from central in [default]
javax.xml.stream#stax-api;1.0-2 from central in [default]
jline#jline;0.9.94 from central in [default]
joda-time#joda-time;2.10.10 from central in [default]
junit#junit;4.11 from central in [default]
log4j#log4j;1.2.17 from central in [default]
net.java.dev.jets3t#jets3t;0.9.0 from central in [default]
org.apache.avro#avro;1.7.4 from central in [default]
org.apache.commons#commons-compress;1.4.1 from central in [default]
org.apache.commons#commons-math3;3.1.1 from central in [default]
org.apache.curator#curator-client;2.7.1 from central in [default]
org.apache.curator#curator-framework;2.7.1 from central in [default]
org.apache.curator#curator-recipes;2.7.1 from central in [default]
org.apache.directory.api#api-asn1-api;1.0.0-M20 from central in [default]
org.apache.directory.api#api-util;1.0.0-M20 from central in [default]
org.apache.directory.server#apacheds-i18n;2.0.0-M15 from central in [default]
org.apache.directory.server#apacheds-kerberos-codec;2.0.0-M15 from central in [default]
org.apache.hadoop#hadoop-annotations;2.7.2 from central in [default]
org.apache.hadoop#hadoop-auth;2.7.2 from central in [default]
```

```
        org.apache.hadoop#hadoop-aws;2.7.2 from central in [default]
        org.apache.hadoop#hadoop-common;2.7.2 from central in [default]
        org.apache.htrace#htrace-core;3.1.0-incubating from central in [default]
        org.apache.httpcomponents#httpclient;4.2.5 from central in [default]
        org.apache.httpcomponents#httpcore;4.2.5 from central in [default]
        org.apache.zookeeper#zookeeper;3.4.6 from central in [default]
        org.codehaus.jackson#jackson-core-asl;1.9.13 from central in [default]
        org.codehaus.jackson#jackson-jaxrs;1.9.13 from central in [default]
        org.codehaus.jackson#jackson-mapper-asl;1.9.13 from central in [default]
        org.codehaus.jackson#jackson-xc;1.9.13 from central in [default]
        org.codehaus.jettison#jettison;1.1 from central in [default]
        org.hamcrest#hamcrest-core;1.3 from central in [default]
        org.mortbay.jetty#jetty;6.1.26 from central in [default]
        org.mortbay.jetty#jetty-util;6.1.26 from central in [default]
        org.slf4j#slf4j-api;1.7.10 from central in [default]
        org.slf4j#slf4j-log4j12;1.7.10 from central in [default]
        org.tukaani#xz;1.0 from central in [default]
        org.xerial.snappy#snappy-java;1.0.4.1 from central in [default]
        xmlenc#xmlenc;0.52 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   71  |   7   |   6   |   0   ||   70  |   5   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-a3bb0454-a5bc-4083-82c6-ea5aa99d1d48
        confs: [default]
        5 artifacts copied, 65 already retrieved (4123kB/71ms)
21/03/25 15:47:40 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform.
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
21/03/25 15:47:44 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 404
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.4.3
      /_/

Using Python version 3.6.3 (default, Dec  9 2017 04:28:46)
SparkSession available as 'spark'.
>>> ^C
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/opt/spark-2.4.3-bin-hadoop2.7/python/pyspark/context.py", line 270, in signal_handler
    raise KeyboardInterrupt()
KeyboardInterrupt
```

```
>>>
```

```
In [7]: df.printSchema()
        df.show(5)
```

```
root
 |-- artist_id: string (nullable = true)
 |-- artist_latitude: double (nullable = true)
 |-- artist_location: string (nullable = true)
 |-- artist_longitude: double (nullable = true)
 |-- artist_name: string (nullable = true)
 |-- duration: double (nullable = true)
 |-- num_songs: long (nullable = true)
 |-- song_id: string (nullable = true)
 |-- title: string (nullable = true)
 |-- year: long (nullable = true)

+------------------+---------------+---------------+----------------+----------------+---------+
|         artist_id|artist_latitude|artist_location|artist_longitude|     artist_name| duration|
+------------------+---------------+---------------+----------------+----------------+---------+
|ARLTWXK1187FB5A3F8|       32.74863| Fort Worth, TX|       -97.32925|     King Curtis|326.00771|
|ARIOZCU1187FB3A3DC|           null|     Hamlet, NC|            null|   JOHN COLTRANE|220.44689|
|ARPFHN61187FB575F6|       41.88415|    Chicago, IL|       -87.63241|     Lupe Fiasco|279.97995|
|AR5S9OB1187B9931E3|       34.05349|Los Angeles, CA|      -118.24532|     Bullet Boys|156.62975|
|AR5T4OY1187B9996C6|           null|  Lulea, Sweden|            null|The Bear Quartet| 249.3122|
+------------------+---------------+---------------+----------------+----------------+---------+
only showing top 5 rows
```

```
In [ ]: print("success")
```

```
In [9]: output_data = "s3a://udacity-dend/"
        #df = spark.read.json(song_data)

            # extract columns to create songs table
            # Using dataframe property, create a new dataframe with required fields
        songs_table = df['song_id', 'title', 'artist_id', 'year', 'duration']
            # Drop duplicates
        songs_table = songs_table.dropDuplicates()
        songs_table.head()
```

```
Out[9]: Row(song_id='SOQFYBD12AB0182188', title='Intro', artist_id='ARAADXM1187FB3ECDB', year=19
```

```
In [10]:     # write songs table to parquet files partitioned by year and artist
         songs_table.write.partitionBy('year', 'artist_id').parquet(os.path.join(output_data, 's
```

```
Py4JJavaError                             Traceback (most recent call last)

<ipython-input-10-da64ebc0ad49> in <module>()
  1 # write songs table to parquet files partitioned by year and artist
----> 2 songs_table.write.partitionBy('year', 'artist_id').parquet(os.path.join(output_data,


/opt/spark-2.4.3-bin-hadoop2.7/python/pyspark/sql/readwriter.py in parquet(self, path, m
837             self.partitionBy(partitionBy)
838         self._set_opts(compression=compression)
--> 839         self._jwrite.parquet(path)
840
841     @since(1.6)


/opt/spark-2.4.3-bin-hadoop2.7/python/lib/py4j-0.10.7-src.zip/py4j/java_gateway.py in __
1255         answer = self.gateway_client.send_command(command)
1256         return_value = get_return_value(
-> 1257             answer, self.gateway_client, self.target_id, self.name)
1258
1259         for temp_arg in temp_args:


/opt/spark-2.4.3-bin-hadoop2.7/python/pyspark/sql/utils.py in deco(*a, **kw)
 61     def deco(*a, **kw):
 62         try:
---> 63             return f(*a, **kw)
 64         except py4j.protocol.Py4JJavaError as e:
 65             s = e.java_exception.toString()


/opt/spark-2.4.3-bin-hadoop2.7/python/lib/py4j-0.10.7-src.zip/py4j/protocol.py in get_re
326                 raise Py4JJavaError(
327                     "An error occurred while calling {0}{1}{2}.\n".
--> 328                     format(target_id, ".", name), value)
329             else:
330                 raise Py4JError(


Py4JJavaError: An error occurred while calling o53.parquet.
: com.amazonaws.services.s3.model.AmazonS3Exception: Status Code: 403, AWS Service: Amazon S
        at com.amazonaws.http.AmazonHttpClient.handleErrorResponse(AmazonHttpClient.java:798
        at com.amazonaws.http.AmazonHttpClient.executeHelper(AmazonHttpClient.java:421)
        at com.amazonaws.http.AmazonHttpClient.execute(AmazonHttpClient.java:232)
        at com.amazonaws.services.s3.AmazonS3Client.invoke(AmazonS3Client.java:3528)
        at com.amazonaws.services.s3.AmazonS3Client.putObject(AmazonS3Client.java:1393)
        at org.apache.hadoop.fs.s3a.S3AFileSystem.createEmptyObject(S3AFileSystem.java:1194)
```

```
        at org.apache.hadoop.fs.s3a.S3AFileSystem.createFakeDirectory(S3AFileSystem.java:117
        at org.apache.hadoop.fs.s3a.S3AFileSystem.mkdirs(S3AFileSystem.java:871)
        at org.apache.hadoop.fs.FileSystem.mkdirs(FileSystem.java:1881)
        at org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter.setupJob(FileOutputCom
        at org.apache.spark.internal.io.HadoopMapReduceCommitProtocol.setupJob(HadoopMapRedu
        at org.apache.spark.sql.execution.datasources.FileFormatWriter$.write(FileFormatWrit
        at org.apache.spark.sql.execution.datasources.InsertIntoHadoopFsRelationCommand.run(
        at org.apache.spark.sql.execution.command.DataWritingCommandExec.sideEffectResult$lz
        at org.apache.spark.sql.execution.command.DataWritingCommandExec.sideEffectResult(co
        at org.apache.spark.sql.execution.command.DataWritingCommandExec.doExecute(commands.
        at org.apache.spark.sql.execution.SparkPlan$$anonfun$execute$1.apply(SparkPlan.scala
        at org.apache.spark.sql.execution.SparkPlan$$anonfun$execute$1.apply(SparkPlan.scala
        at org.apache.spark.sql.execution.SparkPlan$$anonfun$executeQuery$1.apply(SparkPlan.
        at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:151)
        at org.apache.spark.sql.execution.SparkPlan.executeQuery(SparkPlan.scala:152)
        at org.apache.spark.sql.execution.SparkPlan.execute(SparkPlan.scala:127)
        at org.apache.spark.sql.execution.QueryExecution.toRdd$lzycompute(QueryExecution.sca
        at org.apache.spark.sql.execution.QueryExecution.toRdd(QueryExecution.scala:80)
        at org.apache.spark.sql.DataFrameWriter$$anonfun$runCommand$1.apply(DataFrameWriter.
        at org.apache.spark.sql.DataFrameWriter$$anonfun$runCommand$1.apply(DataFrameWriter.
        at org.apache.spark.sql.execution.SQLExecution$$anonfun$withNewExecutionId$1.apply(S
        at org.apache.spark.sql.execution.SQLExecution$.withSQLConfPropagated(SQLExecution.s
        at org.apache.spark.sql.execution.SQLExecution$.withNewExecutionId(SQLExecution.scal
        at org.apache.spark.sql.DataFrameWriter.runCommand(DataFrameWriter.scala:676)
        at org.apache.spark.sql.DataFrameWriter.saveToV1Source(DataFrameWriter.scala:285)
        at org.apache.spark.sql.DataFrameWriter.save(DataFrameWriter.scala:271)
        at org.apache.spark.sql.DataFrameWriter.save(DataFrameWriter.scala:229)
        at org.apache.spark.sql.DataFrameWriter.parquet(DataFrameWriter.scala:566)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java
        at java.lang.reflect.Method.invoke(Method.java:498)
        at py4j.reflection.MethodInvoker.invoke(MethodInvoker.java:244)
        at py4j.reflection.ReflectionEngine.invoke(ReflectionEngine.java:357)
        at py4j.Gateway.invoke(Gateway.java:282)
        at py4j.commands.AbstractCommand.invokeMethod(AbstractCommand.java:132)
        at py4j.commands.CallCommand.execute(CallCommand.java:79)
        at py4j.GatewayConnection.run(GatewayConnection.java:238)
        at java.lang.Thread.run(Thread.java:748)
```

In [ ]:
```
print("done")
```

In [ ]: