In [1]:
```
import configparser
import os
import pyspark.sql.functions as F

from datetime import datetime
from pyspark.sql import SparkSession
from pyspark.sql.window import Window
```

```
VBox()
Starting Spark application
```

| ID | YARN Application ID | Kind | State | Spark UI | Driver log | Current session? |
|---|---|---|---|---|---|---|
| 2 | application_1664445079255_0003 | pyspark | idle | Link | Link | ✔ |

```
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layou
t(height='25px', width='50%'),…
SparkSession available as 'spark'.
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layou
t(height='25px', width='50%'),…
```

In [2]:
```
config = configparser.ConfigParser()
config.read('dl.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']
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layou
t(height='25px', width='50%'),…
```

In [3]:
```
def create_spark_session():
    spark = SparkSession \
        .builder \
        .config("spark.jars.packages", "org.apache.hadoop:hadoop-aws:2.7.0") \
        .getOrCreate()
    return spark
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layou
t(height='25px', width='50%'),…
```

In [4]:

```
def process_song_data(spark, input_data, output_data):
    """ Reading song data and create songs and artists table

        Arguments:
            spark {object}: SparkSession object
            input_data {object}: Source S3 endpoint
            output_data {object}: Target S3 endpoint
        Returns:
            None
    """
    # get filepath to song data file
    song_data = input_data + "song_data/*/*/*/"

    # read song data file
    df = spark.read.json(song_data)
    df.count()

    # extract columns to create songs table
    songs_table = df.select(["song_id", "title", "artist_id", "year", "duration"
    print(songs_table.show(5, False))

    # write songs table to parquet files partitioned by year and artist

    songs_table.write.mode("overwrite").parquet(output_data+'songs/'+'songs.parq

    # extract columns to create artists table
    artists_table = df.select(["artist_id", "artist_name", "artist_location", "a
    print(artists_table.show(5, truncate = False))
    # write artists table to parquet files

    artists_table.write.mode("overwrite").parquet(output_data + 'artists/' + 'ar
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layou
t(height='25px', width='50%'),…
```

In [10]:

```
def process_log_data(spark, input_data, output_data):
    """ Reading log data and create songs and artists table

        Arguments:
            spark {object}: SparkSession object
            input_data {object}: Source S3 endpoint
            output_data {object}: Target S3 endpoint
        Returns:
            None
    """
    # get filepath to log data file
    log_data =input_data + "log_data/"

    # read log data file
    log_df = spark.read.json(log_data)
    print(log_df.show(2))

    # filter by actions for song plays
    log_df = log_df.where(log_df['page'] == 'NextSong')
    print(log_df.show(2))

    # extract columns for users table
    users_table = log_df.select('userId', 'firstName', 'lastName', 'gender', 'le
    print(users_table.show(5, truncate = False))

    # write users table to parquet files
    users_table.write.mode("overwrite").parquet(output_data + 'users/' + 'users.

    # create timestamp column from original timestamp column
    log_df = log_df.withColumn('timestamp',( (log_df.ts.cast('float')/1000).cast

    # extract columns to create time table
    time_table = log_df.select(
                    F.col("timestamp").alias("start_time"),
                    F.hour("timestamp").alias('hour'),
                    F.dayofmonth("timestamp").alias('day'),
                    F.weekofyear("timestamp").alias('week'),
                    F.month("timestamp").alias('month'),
                    F.year("timestamp").alias('year'),
                    F.date_format(F.col("timestamp"), "E").alias("weekday")
                )


    time_table.show(5, False)

    # write time table to parquet files partitioned by year and month

    time_table.write.mode("overwrite").parquet(output_data + 'time/' + 'time.par


    # read in song data to use for songplays table
    song_df = spark.read.json(input_data + "song_data/*/*/*/")

    # extract columns from joined song and log datasets to create songplays tabl
    songplays_table = log_df.join(song_df, (log_df.song == song_df.title) & (log
    songplays_table = songplays_table.distinct() \
                        .select("userId", "timestamp", "song_id", "artist_id", "
                        .withColumn("songplay_id", F.row_number().over( Window.p
                        .withColumnRenamed("userId","user_id")        \
                        .withColumnRenamed("timestamp","start_time")  \
                        .withColumnRenamed("sessionId","session_id")  \
                        .withColumnRenamed("userAgent", "user_agent") \
    # write songplays table to parquet files partitioned by year and month
```

```
    # write songplays table to parquet files partitioned by year and month
    print(songplays_table.show(5))
    songplays_table.write.mode("overwrite").parquet(output_data + 'songplays/' +
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layou
t(height='25px', width='50%'),…
```

In [11]:
```
spark = create_spark_session()
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layou
t(height='25px', width='50%'),…
```

In [12]:
```
input_data = "s3a://udacity-dend-dl-lake/"
output_data = "s3a://sparkify-udacity-data-lake/"
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layou
t(height='25px', width='50%'),…
```

In [13]:
```
process_song_data(spark, input_data, output_data)
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layou
t(height='25px', width='50%'),…
+------------------+----------------------------------------------------+--------
|song_id           |title                                               |artist_i
+------------------+----------------------------------------------------+--------
|SOGOSOV12AF72A285E|¿Dónde va Chichi?                                   |ARGUVEV1
|SOTTDKS12AB018D69B|It Wont Be Christmas                                |ARMBR4Y1
|SOBBUGU12A8C13E95D|Setting Fire to Sleeping Giants                     |ARMAC4T1
|SOIAZJW12AB01853F1|Pink World                                          |AR8ZCNI1
|SONYPOM12A8C13B2D7|I Think My Wife Is Running Around On Me (Taco Hell)|ARDNS031
+------------------+----------------------------------------------------+--------
only showing top 5 rows

None
+------------------+---------------+---------------+---------------+------------
|artist_id         |artist_name    |artist_location|artist_latitude|artist_longi
+------------------+---------------+---------------+---------------+------------
|AR3JMC51187B9AE49D|Backstreet Boys|Orlando, FL    |28.53823       |-81.37739
|AR0IAWL1187B9A96D0|Danilo Perez   |Panama         |8.4177         |-80.11278
|ARWB3G61187FB49404|Steve Morse    |Hamilton, Ohio |null           |null
|AR47JEX1187B995D81|SUE THOMPSON   |Nevada, MO     |37.83721       |-94.35868
|ARHHO3O1187B989413|Bob Azzam      |               |null           |null
+------------------+---------------+---------------+---------------+------------
only showing top 5 rows

None
```

In [14]:
```
process_log_data(spark, input_data, output_data)
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layou
t(height='25px', width='50%'),…
```

```
+-----------+---------+---------+------+-------------+--------+---------+-----+-
|     artist|     auth|firstName|gender|itemInSession|lastName|   length|level|
+-----------+---------+---------+------+-------------+--------+---------+-----+-
|   Harmonia|Logged In|     Ryan|     M|            0|   Smith|655.77751| free|S
|The Prodigy|Logged In|     Ryan|     M|            1|   Smith|260.07465| free|S
+-----------+---------+---------+------+-------------+--------+---------+-----+-
only showing top 2 rows

None
+-----------+---------+---------+------+-------------+--------+---------+-----+-
|     artist|     auth|firstName|gender|itemInSession|lastName|   length|level|
+-----------+---------+---------+------+-------------+--------+---------+-----+-
|   Harmonia|Logged In|     Ryan|     M|            0|   Smith|655.77751| free|S
|The Prodigy|Logged In|     Ryan|     M|            1|   Smith|260.07465| free|S
+-----------+---------+---------+------+-------------+--------+---------+-----+-
only showing top 2 rows

None
+------+---------+--------+------+-----+
|userId|firstName|lastName|gender|level|
+------+---------+--------+------+-----+
|57    |Katherine|Gay     |F     |free |
|84    |Shakira  |Hunt    |F     |free |
|22    |Sean     |Wilson  |F     |free |
|52    |Theodore |Smith   |M     |free |
|80    |Tegan    |Levine  |F     |paid |
+------+---------+--------+------+-----+
only showing top 5 rows

None
+-----------------------+----+---+----+-----+----+-------+
|start_time             |hour|day|week|month|year|weekday|
+-----------------------+----+---+----+-----+----+-------+
|2018-11-15 00:29:39.712|0   |15 |46  |11   |2018|Thu    |
|2018-11-15 00:40:35.072|0   |15 |46  |11   |2018|Thu    |
|2018-11-15 00:44:57.216|0   |15 |46  |11   |2018|Thu    |
|2018-11-15 03:44:05.12 |3   |15 |46  |11   |2018|Thu    |
|2018-11-15 05:48:36.224|5   |15 |46  |11   |2018|Thu    |
+-----------------------+----+---+----+-----+----+-------+
only showing top 5 rows

+-------+--------------------+------------------+------------------+-----+------
|user_id|          start_time|           song_id|         artist_id|level|sessio
+-------+--------------------+------------------+------------------+-----+------
|     15|2018-11-21 21:56:...|SOZCTXZ12AB0182364|AR5KOSW1187FB35FF4| paid|
+-------+--------------------+------------------+------------------+-----+------

None
```

In [ ]: