1 Spark Installation

1.1 MAC Terminal

```
spark-3.4.0-bin-hadoop3 — java ◂ python3 — 123×29
Last login: Mon May 15 14:25:19 on ttys000
(base) deleonv@Vicentes-MacBook-Air ~ % pwd
/Users/deleonv
(base) deleonv@Vicentes-MacBook-Air ~ % cd Desktop
(base) deleonv@Vicentes-MacBook-Air Desktop % cd Spark
(base) deleonv@Vicentes-MacBook-Air Spark % cd spark-3.4.0-bin-hadoop3
(base) deleonv@Vicentes-MacBook-Air spark-3.4.0-bin-hadoop3 % bin/pyspark
Python 3.9.12 (main, Jun  1 2022, 06:36:29)
[Clang 12.0.0 ] :: Anaconda, Inc. on darwin
Type "help", "copyright", "credits" or "license" for more information.
23/05/16 17:57:05 WARN Utils: Your hostname, Vicentes-MacBook-Air.local resolves to a loopback address: 127.0.0.1; using 10
.0.0.110 instead (on interface en0)
23/05/16 17:57:05 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/05/16 17:57:06 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
es where applicable
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.4.0
      /_/

Using Python version 3.9.12 (main, Jun  1 2022 06:36:29)
Spark context Web UI available at http://10.0.0.110:4040
Spark context available as 'sc' (master = local[*], app id = local-1684274226853).
SparkSession available as 'spark'.
>>>
```

1.1.1 Code used for MAC Terminal

- pwd
- cd Desktop
- cd Spark
- cd spark-3.4.0-bin-hadoop3
- bin/pyspark
- quit()

1.2 PySpark + Google Colab Installation:

I really like using Google Colab and after doing a couple of research I came across this link:

https://www.youtube.com/watch?v=Ev_mwYGAbcg

Due to this, I decided to try and to Homework 1 using Google Colab.

In [ ]:
```
!pip install pyspark py4j
```

In [2]:
```
from pyspark.sql import SparkSession
```

In [3]:
```
spark = SparkSession.builder.appName('DataFrame').getOrCreate()
```

In [4]:
```
# https://stackoverflow.com/questions/69553072/reading-a-xml-file-in-pyspark

spark_xml = SparkSession.builder.appName('ReadXML').getOrCreate()
```

2 Working with Text Files

Reading Text Files:

This was the best source (youtube link) I came across for text files. It was extremely useful for me:

- https://www.youtube.com/watch?v=9Dpng3bDsPI
- https://www.geeksforgeeks.org/read-text-file-into-pyspark-dataframe/
- https://spark.apache.org/docs/latest/sql-data-sources-text.html

In [73]:
```
# https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/api/pyspark.sql.functions.col.html
# https://spark.apache.org/docs/3.1.2/api/python/reference/api/pyspark.sql.functions.when.html

from pyspark.sql.functions import col
from pyspark.sql.functions import when
```

In [64]:
```
df1 = spark.read.option('header', 'true') \
      .option('delimiter', '\t') \
      .option('inferSchema', 'true') \
      .csv('Raisin_Dataset.txt')
```

In [65]:
```
# https://sparkbyexamples.com/spark/spark-show-full-column-content-dataframe/

df1.show(n = 5, truncate = False)
```

```
+-----+---------------+---------------+------------+----------+-----------+---------+-------+
|Area |MajorAxisLength|MinorAxisLength|Eccentricity|ConvexArea|Extent     |Perimeter|Class  |
+-----+---------------+---------------+------------+----------+-----------+---------+-------+
|87524|442.2460114    |253.291155     |0.819738392 |90546     |0.758650579|1184.04  |Kecimen|
|75166|406.690687     |243.0324363    |0.801805234 |78789     |0.68412957 |1121.786 |Kecimen|
|90856|442.2670483    |266.3283177    |0.798353619 |93717     |0.637612812|1208.575 |Kecimen|
|45928|286.5405586    |208.7600423    |0.684989217 |47336     |0.699599385|844.162  |Kecimen|
|79408|352.1907699    |290.8275329    |0.56401133  |81463     |0.792771926|1073.251 |Kecimen|
+-----+---------------+---------------+------------+----------+-----------+---------+-------+
only showing top 5 rows
```

Lets print out the schema:

In [66]:
```
df1.printSchema()
```

```
root
 |-- Area: integer (nullable = true)
 |-- MajorAxisLength: double (nullable = true)
 |-- MinorAxisLength: double (nullable = true)
 |-- Eccentricity: double (nullable = true)
 |-- ConvexArea: integer (nullable = true)
 |-- Extent: double (nullable = true)
 |-- Perimeter: double (nullable = true)
 |-- Class: string (nullable = true)
```

Changing Column Name

In [67]:
```
new_df1 = df1.withColumnRenamed("Class", "Categorical Class")
new_df1.show(n = 5, truncate=False)
```

```
+-----+---------------+---------------+------------+----------+-----------+---------+-----------------+
|Area |MajorAxisLength|MinorAxisLength|Eccentricity|ConvexArea|Extent     |Perimeter|Categorical Class|
+-----+---------------+---------------+------------+----------+-----------+---------+-----------------+
|87524|442.2460114    |253.291155     |0.819738392 |90546     |0.758650579|1184.04  |Kecimen          |
|75166|406.690687     |243.0324363    |0.801805234 |78789     |0.68412957 |1121.786 |Kecimen          |
|90856|442.2670483    |266.3283177    |0.798353619 |93717     |0.637612812|1208.575 |Kecimen          |
|45928|286.5405586    |208.7600423    |0.684989217 |47336     |0.699599385|844.162  |Kecimen          |
|79408|352.1907699    |290.8275329    |0.56401133  |81463     |0.792771926|1073.251 |Kecimen          |
+-----+---------------+---------------+------------+----------+-----------+---------+-----------------+
only showing top 5 rows
```

In [72]:
```
new_df1.select(['Categorical Class']).show(n=5)
```

```
+-----------------+
|Categorical Class|
+-----------------+
|          Kecimen|
|          Kecimen|
|          Kecimen|
|          Kecimen|
|          Kecimen|
+-----------------+
only showing top 5 rows
```

Adding Columns

In [74]:
```
# https://towardsdatascience.com/4-different-ways-of-creating-a-new-column-with-pyspark-1a6782fe764e

final_df = new_df1.withColumn('Numerical Class', when(col('Categorical Class') == 'Kecimen', 1).otherwise(0))
```

In [80]:
```
final_df.show(n = 5, truncate = False)
```

```
+-----+---------------+---------------+------------+----------+-----------+---------+-----------------+---------------+
|Area |MajorAxisLength|MinorAxisLength|Eccentricity|ConvexArea|Extent     |Perimeter|Categorical Class|Numerical Class|
+-----+---------------+---------------+------------+----------+-----------+---------+-----------------+---------------+
|87524|442.2460114    |253.291155     |0.819738392 |90546     |0.758650579|1184.04  |Kecimen          |1              |
|75166|406.690687     |243.0324363    |0.801805234 |78789     |0.68412957 |1121.786 |Kecimen          |1              |
|90856|442.2670483    |266.3283177    |0.798353619 |93717     |0.637612812|1208.575 |Kecimen          |1              |
|45928|286.5405586    |208.7600423    |0.684989217 |47336     |0.699599385|844.162  |Kecimen          |1              |
|79408|352.1907699    |290.8275329    |0.56401133  |81463     |0.792771926|1073.251 |Kecimen          |1              |
+-----+---------------+---------------+------------+----------+-----------+---------+-----------------+---------------+
only showing top 5 rows
```

In [84]:
```
# https://sparkbyexamples.com/pyspark/pyspark-select-distinct/

KvsB = final_df.select('Categorical Class', 'Numerical Class').distinct()
KvsB.show()
```

```
+-----------------+---------------+
|Categorical Class|Numerical Class|
+-----------------+---------------+
|          Kecimen|              1|
|            Besni|              0|
+-----------------+---------------+
```

3 Working with Json Files

- https://www.youtube.com/watch?v=6XdXHgFMgf4

In [133…
```
df2 = spark.read.json('quize.json')
df2.show(truncate = False)
```

```
+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
|quiz
|
+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
|{{{12, [10, 11, 12, 13], 5 + 7 = ?}, {4, [1, 2, 3, 4], 12 - 8 = ?}}, {{Huston Rocket, [New York Bulls, Los Angeles Kings, Golden State Warriros, Huston Rocket], Which one is correct team name in NBA?}}}|
+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
```

In [134…
```
df2.printSchema()
```

```
root
 |-- quiz: struct (nullable = true)
 |    |-- maths: struct (nullable = true)
 |    |    |-- q1: struct (nullable = true)
 |    |    |    |-- answer: string (nullable = true)
 |    |    |    |-- options: array (nullable = true)
 |    |    |    |    |-- element: string (containsNull = true)
 |    |    |    |-- question: string (nullable = true)
 |    |    |-- q2: struct (nullable = true)
 |    |    |    |-- answer: string (nullable = true)
 |    |    |    |-- options: array (nullable = true)
 |    |    |    |    |-- element: string (containsNull = true)
 |    |    |    |-- question: string (nullable = true)
 |    |-- sport: struct (nullable = true)
 |    |    |-- q1: struct (nullable = true)
 |    |    |    |-- answer: string (nullable = true)
 |    |    |    |-- options: array (nullable = true)
 |    |    |    |    |-- element: string (containsNull = true)
 |    |    |    |-- question: string (nullable = true)
```

4 Working with XML Files

Working with XML was the hardest task in this homework for me. Since I wanted to use Google Colab to read files using PySpark, I ended up facing many issues for question 4.

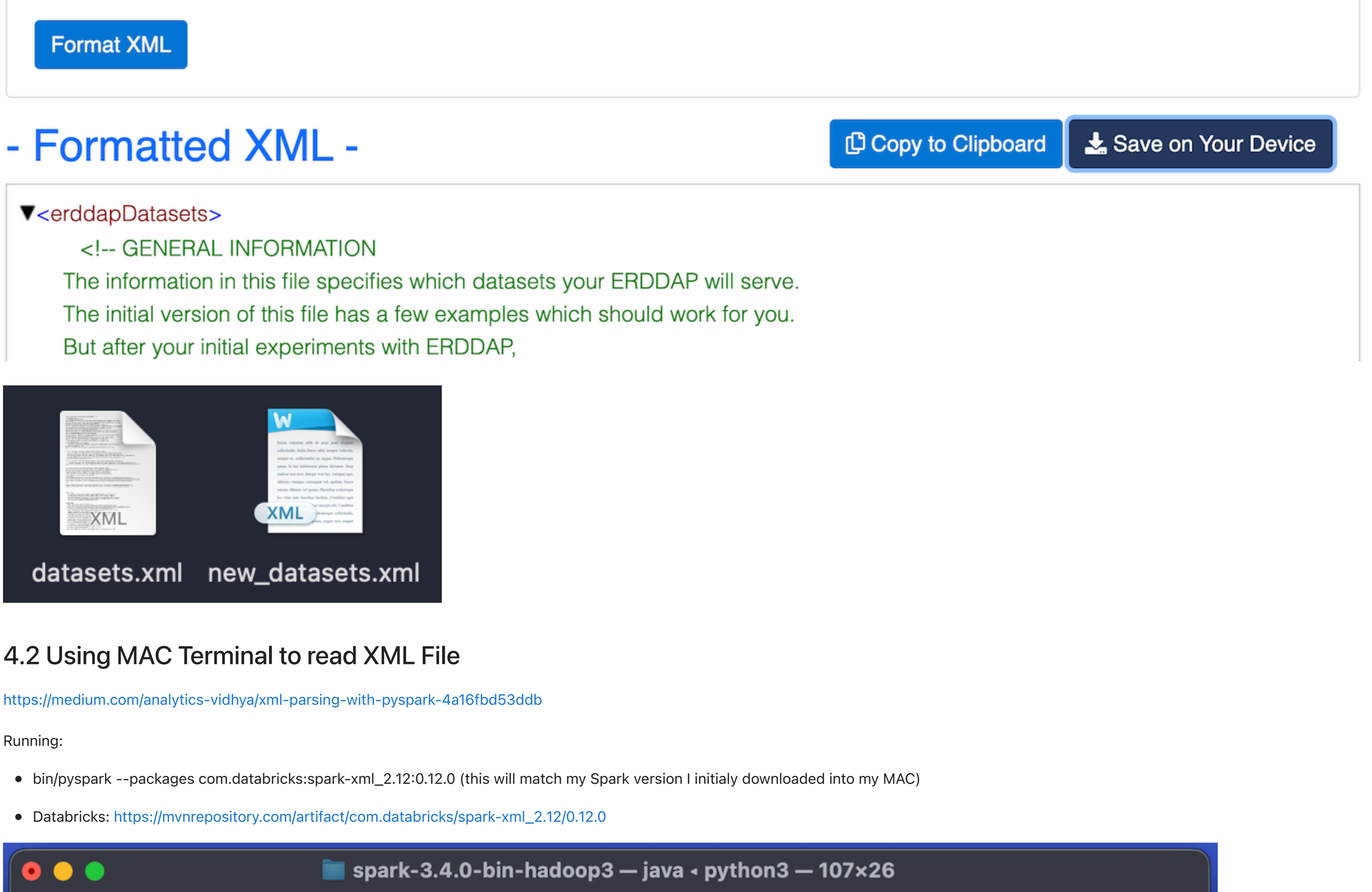
I tried downloading databricks API many times, but it ended up failing multiple times. Due to this I managed to run this question using my MAC Terminal. The images and codes below will provide each step I took in order to succesfully run this task.

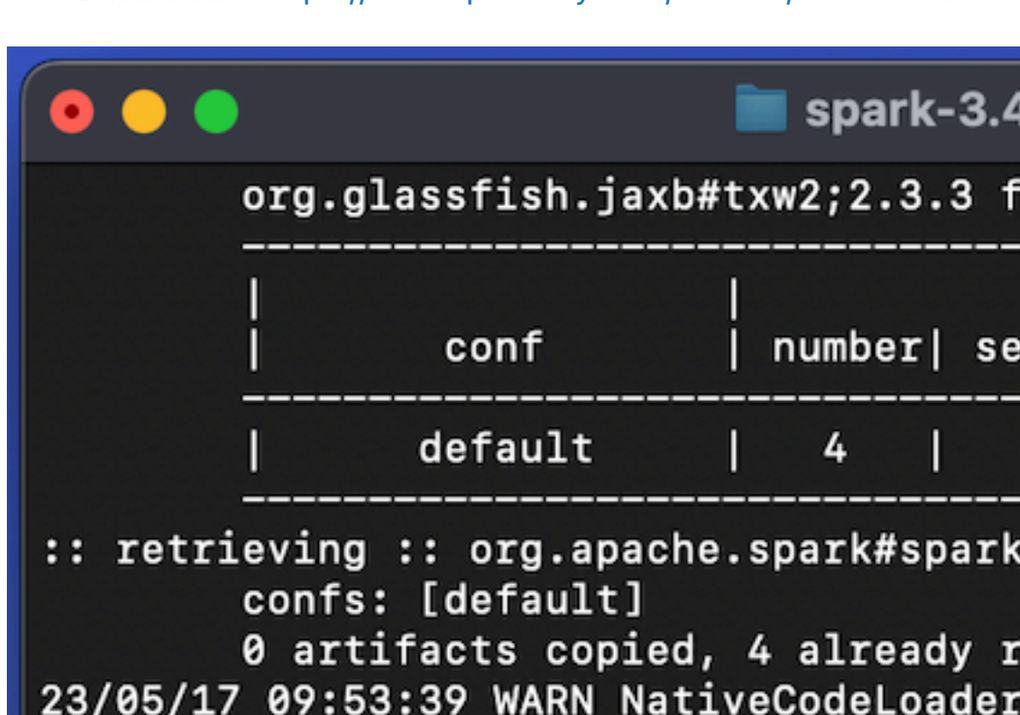
Important: I will try to submit all other suplemental material into Canvas.

4.1 Google Chrome XML Viewer

I noticed there was something wrong when I tried to open the XML File. I've never worked with xml before, so I decided to open it and noticed that it had an odd formatting. During office hours, the TA kindly helped me getting through the json file structure and we saw it had an odd formatting. I decided to go ahead and used Google Chrome XML Viewer to properly view the "datasets.xml" file.

https://www.freecodecamp.org/news/how-to-open-an-xml-file/#:~:text=If%20you%20want%20to%20open,the%20%22Open%20With%22%20option.





4.2 Using MAC Terminal to read XML File

https://medium.com/analytics-vidhya/xml-parsing-with-pyspark-4a16fbd53ddb

Running:

- bin/pyspark --packages com.databricks:spark-xml_2.12:0.12.0 (this will match my Spark version I initialy downloaded into my MAC)
- Databricks: https://mvnrepository.com/artifact/com.databricks/spark-xml_2.12/0.12.0

```
spark-3.4.0-bin-hadoop3 — java ◂ python3 — 107×26
        org.glassfish.jaxb#txw2;2.3.3 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   4   |   0   |   0   |   0   ||   4   |   0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-3ffc83f4-5c8e-471b-a0cf-cdeec8a2f47a
        confs: [default]
        0 artifacts copied, 4 already retrieved (0kB/3ms)
23/05/17 09:53:39 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using bu
iltin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.4.0
      /_/

Using Python version 3.9.12 (main, Jun  1 2022 06:36:29)
Spark context Web UI available at http://10.0.0.110:4040
Spark context available as 'sc' (master = local[*], app id = local-1684331620407).
SparkSession available as 'spark'.
>>>
```

Running:

- spark_xml = SparkSession.builder.appName('ReadXML').getOrCreate()
- df3= spark_xml.read.format('xml').option('rowTag', 'dataset').load('/Users/deleonv/Desktop/Spark/HW1/new_datasets.xml') df3.printSchema()

```
SparkSession available as 'spark'.
>>> spark_xml = SparkSession.builder.appName('ReadXML').getOrCreate()
23/05/17 10:10:16 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take effect.
>>> df3= spark_xml.read.format('xml').option('rowTag', 'dataset').load('/Users/deleonv/Desktop/Spark/HW1/new_datasets.xml')
>>> df3.printSchema()
root
 |-- _active: boolean (nullable = true)
 |-- _datasetID: string (nullable = true)
 |-- _type: string (nullable = true)
 |-- addAttributes: struct (nullable = true)
 |    |-- att: array (nullable = true)
 |    |    |-- element: struct (containsNull = true)
 |    |    |    |-- _VALUE: string (nullable = true)
 |    |    |    |-- _name: string (nullable = true)
 |-- columnNameForExtract: string (nullable = true)
 |-- dataVariable: array (nullable = true)
 |    |-- element: struct (containsNull = true)
 |    |    |-- addAttributes: struct (nullable = true)
 |    |    |    |-- att: array (nullable = true)
 |    |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |    |-- _VALUE: string (nullable = true)
 |    |    |    |    |    |-- _name: string (nullable = true)
 |    |    |    |    |    |-- _type: string (nullable = true)
 |    |    |-- dataType: string (nullable = true)
 |    |    |-- destinationName: string (nullable = true)
 |    |    |-- sourceName: string (nullable = true)
 |-- extractRegex: string (nullable = true)
 |-- fileDir: string (nullable = true)
 |-- fileNameRegex: string (nullable = true)
 |-- fileTableInMemory: boolean (nullable = true)
 |-- metadataFrom: string (nullable = true)
 |-- postExtractRegex: string (nullable = true)
 |-- preExtractRegex: string (nullable = true)
 |-- recursive: boolean (nullable = true)
 |-- reloadEveryNMinutes: long (nullable = true)
 |-- sortFilesBySourceNames: string (nullable = true)
 |-- sortedColumnSourceName: string (nullable = true)
```

Running:

- df3.show(truncate=False)

```
ceNames|sortedColumnSourceName|
+-------+---------------------+------------------------------------------------------------------------------------------
|true   |test_2788_304d_47c8|EDDTableFromNcFiles|{[{CF-1.6, COARDS, Unidata Dataset Discovery v1.0, Conventions}, {OOI, institution}, {bars, based, conductivity, ctd41cp, ctdgv, ctdgvm000, cycle, data, dataset, degrees, density, deployment, depth, derived, driver, elapsed, electrical, engine, gl363, glider, gp05moas, gp05moas-gl363-04-ctdgvm000-recovered, host, host-ctdgv, ingestion, instrument, internal, m-3, mission, observations, ocean, oceans,\nOceans > Ocean Temperature > Water Temperature,\nOceans > Salinity/Density > Conductivity,\nOceans > Salinity/Density > Density,\nooi, port, practical, preferred, pressure, produced, provenance, pss, pss-78, recovered, s/m, salinity, sample, science, sea, sea_water_density, sea_water_electrical_conductivity, sea_water_practical_salinity, sea_water_temperature, seawater, secs, start, stream, temperature, time, timestamp, unitless, version, water, keywords}, {GCMD Science Keywords, keywords_vocabulary}, {[standard], license}, {CF-1.6, COARDS, Unidata Dataset Discovery v1.0, Metadata_Conventions}, {(local files), sourceUrl}, {OOI dataset, summary}, {Data produced by Stream Engine version 0.8.4 for GP05MOAS-GL363-04-CTDGVM000-recovered_host-ctdgv_m_glider_instrument_recovered (Sample OOI Dataset), title}, {deployment, lon, lat, cdm_timeseries_variables}]}|                |[{{[{Obs, long_name, null}]}, long, obs, obs}, {{[{40.0, colorBarMaximum, double}, {30.0, colorBarMinimum, double}, {null, coordinates, null}, {sea_water_electrical_conductivity, standard_name, null}]}, double, sci_water_cond, sci_water_cond}, {{[{null, coordinates, null}, {time, standard_name, null}]}, double, sci_m_present_time, sci_m_present_time}, {{[{1.0, colorBarMaximum, double}, {0.0, colorBarMinimum, double}, {null, coordinates, null}]}, double, sci_water_pracsal, sci_water_pracsal}, {{[{null, coordinates, null}, {time, standard_name, null}]}, double, driver_timestamp, driver_timestamp}, {{[{null, coordinates, null}]}, String, id, id}, {{[{null, coordinates, null}]}, String, provenance, provenance}, {{[{longitude, standard_name, null}]}, double, lon, lon}, {{[{null, coordinates, null}, {time, standard_name, null}]}, double, internal_timestamp, internal_timestamp}, {{[{null, coordinates, null}]}, double, m_present_secs_into_mission, m_present_secs_into_mission}, {{[{null, coordinates, null}, {time, standard_name, null}]}, double, m_present_time, m_present_time}, {{[{null, coordinates, null}, {time, standard_name, null}]}, double, ingestion_timestamp, ingestion_timestamp}, {{[{null, coordinates, null}, {time, standard_name, null}]}, double, port_timestamp, port_timestamp}, {{[{null, coordinates, null}]}, double, sci_water_pressure, sci_water_pressure}, {{[{28.0, colorBarMaximum, double}, {20.0, colorBarMinimum, double}, {null, coordinates, null}]}, double, sci_seawater_density, sci_seawater_density}, {{[{null, coordinates, null}]}, long, deployment, deployment}, {{[{latitude, standard_name, null}]}, double, lat, lat}, {{[{null, coordinates, null}, {time, standard_name, null}]}, double, sci_ctd41cp_timestamp, sci_ctd41cp_timestamp}, {{[{1.0, colorBarMaximum, double}, {0.0, colorBarMinimum, double}, {null, coordinates, null}]}, String, preferred_timestamp, preferred_timestamp}, {{[{null, coordinates, null}]}, double, sci_m_present_secs_into_mission, sci_m_present_secs_into_mission}, {{[{32.0, colorBarMaximum, double}, {0.0, colorBarMinimum, double}, {null, coordinates, null}, {sea_water_temperature, standard_name, null}]}, double, sci_water_temp, sci_water_temp}, {{null}, double, time, time}, {{[{null, coordinates, null}, {L0 Provenance Data, long_name, null}]}, String, l0_provenance_information, l0_provenance_information}]|            |/Users/lcampbell/data/ooi/test/|.*\.nc       |false            |last        |                |                |true     |5              |sci_m_present_time    |sci_m_present_time    |
+-------+---------------------+------------------------------------------------------------------------------------------
```

5 Working with CSV Files

In [111…
```
df4 = spark.read.option('header', 'true').csv('/content/Social_Network_Ads.csv')
```

In [112…
```
df4.printSchema()
```

```
root
 |-- User ID: string (nullable = true)
 |-- Gender: string (nullable = true)
 |-- Age: string (nullable = true)
 |-- EstimatedSalary: string (nullable = true)
 |-- Purchased: string (nullable = true)
```

Print 10 rows

In [115…
```
df4.show(n = 10, truncate = False)
```

```
+--------+------+---+---------------+---------+
|User ID |Gender|Age|EstimatedSalary|Purchased|
+--------+------+---+---------------+---------+
|15624510|Male  |19 |19000          |0        |
|15810944|Male  |35 |20000          |0        |
|15668575|Female|26 |43000          |0        |
|15603246|Female|27 |57000          |0        |
|15804002|Male  |19 |76000          |0        |
|15728773|Male  |27 |58000          |0        |
|15598044|Female|27 |84000          |0        |
|15694829|Female|32 |150000         |1        |
|15600575|Male  |25 |33000          |0        |
|15727311|Female|35 |65000          |0        |
+--------+------+---+---------------+---------+
only showing top 10 rows
```

Select Columns 2 and 3 (Gender and Age respectively) and print them

In [117…
```
df4.select('Gender', 'Age').show()
```

```
+------+---+
|Gender|Age|
+------+---+
|  Male| 19|
|  Male| 35|
|Female| 26|
|Female| 27|
|  Male| 19|
|  Male| 27|
|Female| 27|
|Female| 32|
|  Male| 25|
|Female| 35|
|Female| 26|
|Female| 26|
|  Male| 20|
|  Male| 32|
|  Male| 18|
|  Male| 29|
|  Male| 47|
|  Male| 45|
|  Male| 46|
|Female| 48|
+------+---+
only showing top 20 rows
```

Rename the column "EstimatedSalary" to "Annual Salary" and "Purchased" to "Score".

In [121…
```
new_df4 = df4.withColumnRenamed('EstimatedSalary', 'Annual Salary')
new_df4 = new_df4.withColumnRenamed('Purchased', 'Score')

new_df4.show(n = 10, truncate=False)
```

```
+--------+------+---+-------------+-----+
|User ID |Gender|Age|Annual Salary|Score|
+--------+------+---+-------------+-----+
|15624510|Male  |19 |19000        |0    |
|15810944|Male  |35 |20000        |0    |
|15668575|Female|26 |43000        |0    |
|15603246|Female|27 |57000        |0    |
|15804002|Male  |19 |76000        |0    |
|15728773|Male  |27 |58000        |0    |
|15598044|Female|27 |84000        |0    |
|15694829|Female|32 |150000       |1    |
|15600575|Male  |25 |33000        |0    |
|15727311|Female|35 |65000        |0    |
+--------+------+---+-------------+-----+
only showing top 10 rows
```