PySpark Training Notes

Up front code to load pyspark in Jupyter

* import pyspark
* from pyspark.sql import SparkSession
* spark = SparkSession.builder.appName("Jupyter Notebook").getOrCreate()
* spark

Import csv – two ways

* odf = spark.read.format("csv").option("header", "true").load("original.csv") # imports original csv and says first row contains column names
* odf = spark.read.csv("original.csv", header=True) #same as above but shorter and preferred

Order of Operations for PySpark

1. filter first to remove unnessary rows
2. then group
3. then aggregate
4. then select the columns for the final output
5. Then sort the rows

This is the Pyspark Code

odf.filter(col("Price in US Dollars") > 3) \

.groupBy("Year#", "Country") \

.agg(avg("Price in US Dollars").alias("Avg Price")) \

.select("Country", "Year", "Avg Price") \

.orderBy("Country", "Year") \

.show()

# created before Pyspark query: odf = odf.withColumn("year", substring("Date",1,4))

This is the equivalent SQL code

odf\_sql = spark.sql("""

SELECT Country, substr(Date, 1, 4) AS Year, AVG(`Price in US Dollars’) AS `Avg Price`

FROM bigmac

WHERE CAST(`Price in US Dollars` AS FLOAT) > 3

GROUP BY Country, Year

ORDER BY Year, Country

""")