Discussion IST769 Unit D - Object Storage / Spark

## Agenda

1. Your Questions
2. Unit Coursework Activities

Official Spark Documentation

<https://spark.apache.org/docs/3.1.2/api/python/user_guide/index.html>

## FRONT MATTER

1. Start Your Environment:

**ps> docker-compose up -d jupyter minio**

1. Minio Web UI is here <http://localhost:9001>
2. Jupyter is here: <http://localhost:8888>

YOUR MINIO WON'T START? UPDATE IT!

(Because your prof is a moron.)

Bring it down: ps> **docker-compose down**

Remove volume: ps>

Update Image: ps> **docker-compose pull minio**

Bring back up: ps> **docker-compose up -d jupyter minio**

**minio:**

**image: minio/minio:latest**

**minio:**

**image: minio/minio:RELEASE.2023-02-10T18-48-39Z**

## 1. Your Questions

Ask any questions you have here!

* The correct URL that reads all the files to the folder
* How is http:// not a valid spark location, while “file://”, “s3a://”, “webhdfs://” “mongo://” are
  + Check maven central Hadoop 3.2 and Spark 3.12 Scala 2.12
* Difference between renameColumn(), toDF(), withColumn(), and withColumnRenamed()
  + More Examples
* To access column “bar” in a nested StructType “foo” of DataFrame x we write which spark?
  + Tables / non tabular data
* How to filter the “foo” column in the dataframe x to value “Bar”
* How to count non-null values in any column “foo” of dataframe “x”
  + Counting Null / Non null.

## 2. Coursework Activities

1. Use the minio client to accomplish the following:  
   Create an alias to the minio server <http://minio:9000> Create a Minio bucket called **D-inclass**

Add folder to the bucket called **grades**Place all the files in /datasets/grades in that folder  
Add another folder to the bucket called **other**

Place the files in the **/datasets/courses** folder in the **other** folder in the minio bucket

1. Spark list all grades (from the Minio folder) After the year 2016 that have 3 or more credits. Include the course, year and grade,
2. Spark: Using all of the UFO sightings in the datasets folder, get a count of UFO sightings by state. (Local Datasets folder)
3. Spark: Using the **US-Senators.json** file in the **json-samples** folder, create a Spark Dataframe of Republican party senators only. Include the Senafor’s name, twitter Id, phone and website. (Local Datasets Folder)
4. Use the Grades data on Minio to calculate the students GPA. The GPA calculation is the sum of the grade points times the number of credits in the course.

CODE FROM CLASS

df = spark.read.option("multiline",True).json("/home/jovyan/datasets/json-samples/US-senators.json")

df.toPandas()

from pyspark.sql.functions import col, lit, concat

df.select(concat("person.firstname",lit(" "),"person.lastname")).show(2)

## 