Discussion IST769 Unit E - Schemaless SQL

## Agenda

1. Your Questions
2. Go over Problem Set
3. Unit Coursework Activities

## FRONT MATTER

PS>  **docker-compose up -d minio jupyter drill**

<https://drill.apache.org/docs/connect-a-data-source/>

<https://drill.apache.org/docs/installing-the-odbc-driver/>

## 1. Your Questions

Ask any questions you have here!

* Which HTTP storage configuration setting will parse the HTTP Query String?
  + Params
  + <http://someapi.com/resource?name=tony>  
    SELECT \* from http.resource WHERE name=’tony’
* When can we access COLUMNS[x]
  + When there is no schema. Drill defaults to a column array
* Can spark SQL output be converted back to a dataframe?
  + Yes. save it in a variable before you materialize it… that variable IS A DATAFRAME!

4.<https://drill.apache.org/docs/installing-the-odbc-driver/> what are the credentials?

Trying to download from HP Enterprise - I got a request in.

OpenWeatherMaps: f8dc07f223611eefcb8dc902b44a184a

## 2. Coursework Activities

Drill: <http://localhost:8047/> no authenticagtion

Minio: <http://localhost:9001> minio / SU2orange!

1. Place **datasets/customers/customers.csv** into a minio bucket, then configure drill to query the file. Provide the SQL statement
2. Drill the following file using HTTP <https://raw.githubusercontent.com/mafudge/datasets/master/crime/nys-crime-1960-2012.csv>
3. What was the worst year for violent crimes?
4. What was the best year to not have your vehicle stolen?
5. Load this file **datasets/netflix-canceled-2021/peaky-blinders.json** into a Spark Dataframe using SparkFiles
6. Spark: Extract the episodes into a Temp View **pbeps**
7. What was the best season of this show?
8. What is the name and season of the highest rated episode?