DSC-650: Big Data Final Project

Joelle Rudolph

Introduction

For my final project for DSC-650, I chose the more simple approach of the two options provided to examine the factors that may influence the total burned area of forest fires in Portugal. Forest fires pose significant environmental and economic challenges, making it essential to understand the underlying variables that contribute to their severity. To conduct this project, I utilized a robust data pipeline comprising several key technologies: NiFi for data ingestion, HDFS for data storage, Hive for data warehousing, and Spark for data processing and querying.

Dataset

The dataset I chose was the Forest Fires dataset, created by Paulo Cortez and Anbal Morais, and I obtained it through the UCI Machine Learning Repository (Link Here). The dataset has 13 variables and 517 instances on the burned area of forest fires in northeast Portugal. These 13 variables consisted of:

X: x-axis spatial coordinate within the Montesinho park map: 1 to 9

Y: y-axis spatial coordinate within the Montesinho park map: 2 to 9

Month: month of the year: 'jan' to 'dec'

Day: day of the week: 'mon' to 'sun'

FFMC: FFMC index from the FWI system: 18.7 to 96.20

DMC: DMC index from the FWI system: 1.1 to 291.3

DC: DC index from the FWI system: 7.9 to 860.6

ISI: ISI index from the FWI system: 0.0 to 56.10

Temp: temperature: 2.2 to 33.30 (Celsius)

RH: relative humidity: 15.0 to 100 (%)

Wind: wind speed: 0.40 to 9.40 (km/h)

Rain: outside rain: 0.0 to 6.4 (mm/m2)

Area: the burned area of the forest: 0.00 to 1090.84 (ha)

Pipeline Overview

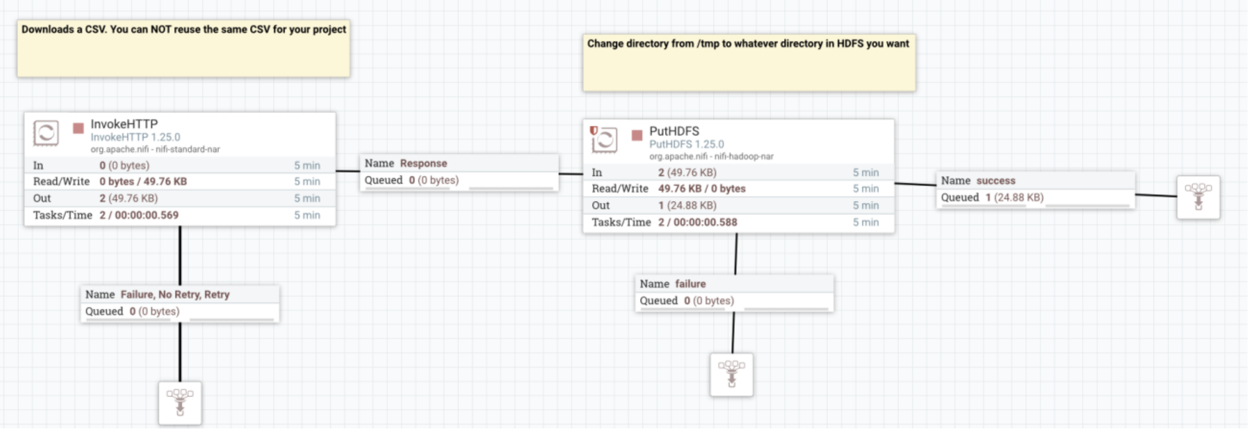
I first downloaded the data and uploaded it to my own GitHub repository, then utilized the NiFi template provided in the class GitHub repository to load in the data using a link to the raw data. I changed the /tmp output to my own directory path. From there, I closed out NiFi and started a Hive session to create my Hive table structure before proceeding. I started a PySpark session with Hive support enabled and showed all tables to make sure that the 'fires' table was present before running my queries. I pulled the first few rows of data, then looked at the number of fires where over 10 ha were burned. I also found the average temperature (in Celsius) for each month represented in the data.

Issues Encountered

The issue that took me the longest to work out was the switch from data in HDFS to Hive table. I had some problems with finding the file path and as a result, had some difficulty with pulling the data from HDFS to Hive for the table queries. I also had an issue with a null row in my data once imported into the Hive table but was still able to run my Spark queries.

Screenshots and Code

NiFi UI:



Downloads a CSV. You can NOT reuse the same CSV for your project

Change directory from /tmp to whatever directory in HDFS you want

InvokeHTTP
InvokeHTTP 1.25.0
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 49.76 KB	5 min
Out	2 (49.76 KB)	5 min
Tasks/Time	2 / 00:00:00.569	5 min

Name Response
Queued 0 (0 bytes)

PutHDFS
PutHDFS 1.25.0
org.apache.nifi - nifi-hadoop-nar

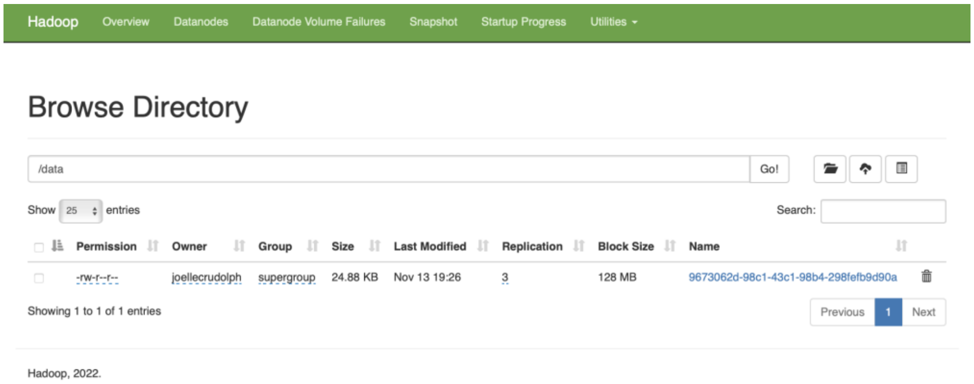
In	2 (49.76 KB)	5 min
Read/Write	49.76 KB / 0 bytes	5 min
Out	1 (24.88 KB)	5 min
Tasks/Time	2 / 00:00:00.588	5 min

Name success
Queued 1 (24.88 KB)

Name Failure, No Retry, Retry
Queued 0 (0 bytes)

Name failure
Queued 0 (0 bytes)

Dataset in Hadoop Directory:



Browse Directory

/data

Show 25 entries

Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
□	-rw-r--r--	joellecrudolph	supergroup	24.88 KB	Nov 13 19:26	3	128 MB	9673062d-98c1-43c1-98b4-298fefb9d90a

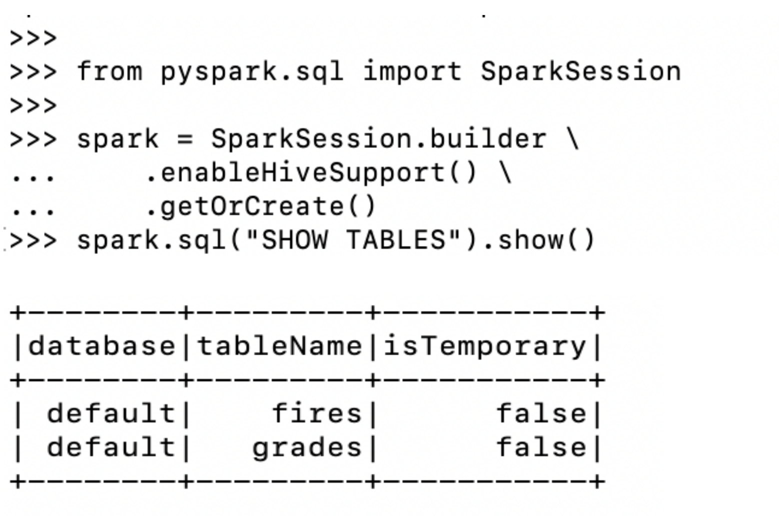
Showing 1 to 1 of 1 entries

Hadoop, 2022.

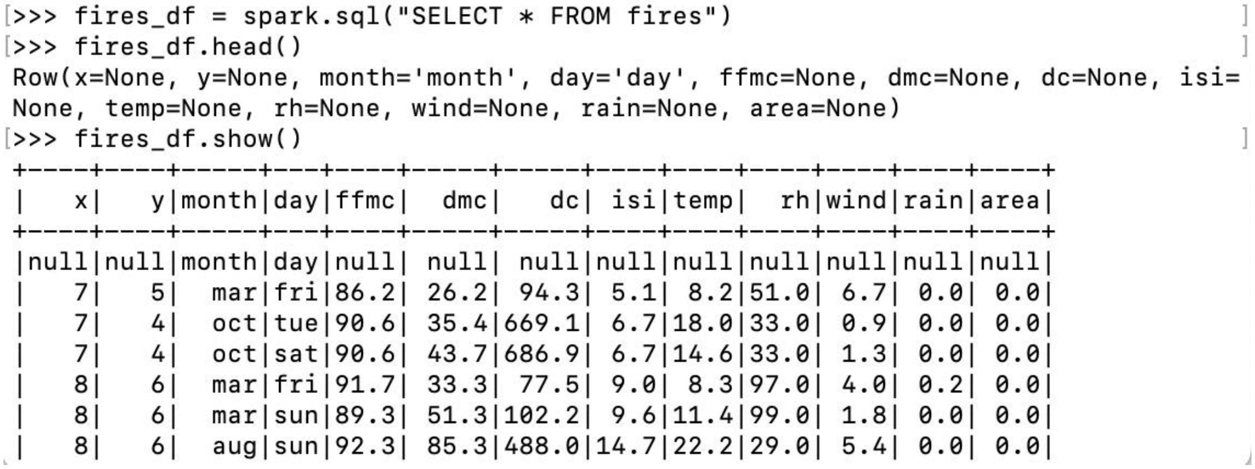
Table Structure in Hive:

```
CREATE TABLE fires(
   > `X` INT,
   > `Y` INT,
   > `month` STRING,
   > `day` STRING,
   > `FFMC`DOUBLE,
   > `DMC` DOUBLE,
   > `DC` DOUBLE,
   > `ISI` DOUBLE,
   > `temp` DOUBLE,
   > `RH` DOUBLE,
   > `wind` DOUBLE,
   > `rain` DOUBLE,
   > `area` DOUBLE)
   > ROW FORMAT DELIMITED
   > FIELDS TERMINATED BY ','
   > STORED AS TEXTFILE
   > tblproperties("skip.header.line.count"="1");
```

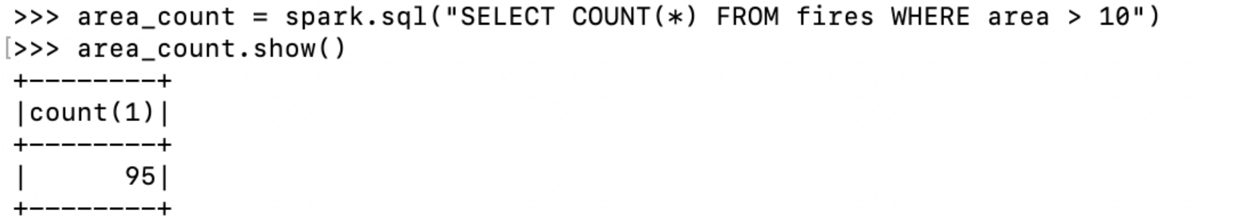
PySpark Queries:



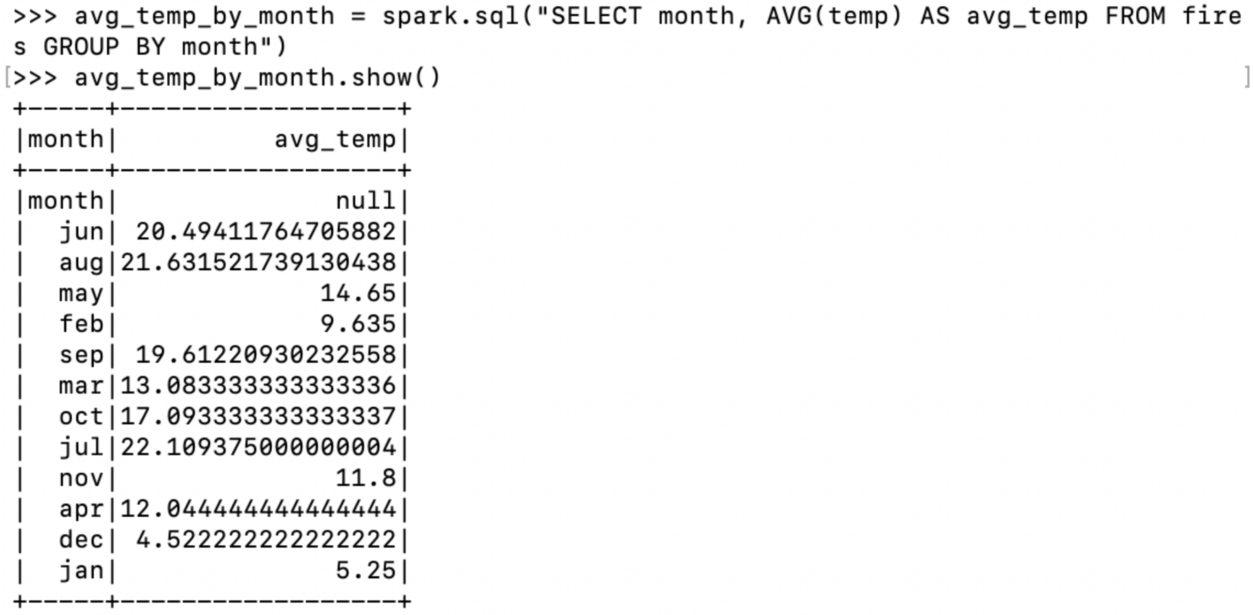
```
>>>
>>> from pyspark.sql import SparkSession
>>>
>>> spark = SparkSession.builder \
...     .enableHiveSupport() \
...     .getOrCreate()
>>> spark.sql("SHOW TABLES").show()

+--------+---------+-----------+
|database|tableName|isTemporary|
+--------+---------+-----------+
| default|    fires|      false|
| default|   grades|      false|
+--------+---------+-----------+
```



```
>>> fires_df = spark.sql("SELECT * FROM fires")
>>> fires_df.head()
Row(x=None, y=None, month='month', day='day', ffmc=None, dmc=None, dc=None, isi=
None, temp=None, rh=None, wind=None, rain=None, area=None)
>>> fires_df.show()
+----+----+-----+---+----+-----+-----+----+----+----+----+----+----+
|   x|   y|month|day|ffmc|  dmc|   dc| isi|temp|  rh|wind|rain|area|
+----+----+-----+---+----+-----+-----+----+----+----+----+----+----+
|null|null|month|day|null| null| null|null|null|null|null|null|null|
|   7|   5|  mar|fri|86.2| 26.2| 94.3| 5.1| 8.2|51.0| 6.7| 0.0| 0.0|
|   7|   4|  oct|tue|90.6| 35.4|669.1| 6.7|18.0|33.0| 0.9| 0.0| 0.0|
|   7|   4|  oct|sat|90.6| 43.7|686.9| 6.7|14.6|33.0| 1.3| 0.0| 0.0|
|   8|   6|  mar|fri|91.7| 33.3| 77.5| 9.0| 8.3|97.0| 4.0| 0.2| 0.0|
|   8|   6|  mar|sun|89.3| 51.3|102.2| 9.6|11.4|99.0| 1.8| 0.0| 0.0|
|   8|   6|  aug|sun|92.3| 85.3|488.0|14.7|22.2|29.0| 5.4| 0.0| 0.0|
```



```
>>> area_count = spark.sql("SELECT COUNT(*) FROM fires WHERE area > 10")
>>> area_count.show()
+---------+
|count(1)|
+---------+
|       95|
+---------+
```



```
>>> avg_temp_by_month = spark.sql("SELECT month, AVG(temp) AS avg_temp FROM fire
s GROUP BY month")
>>> avg_temp_by_month.show()
+-----+------------------+
|month|          avg_temp|
+-----+------------------+
|month|              null|
|  jun| 20.49411764705882|
|  aug|21.631521739130438|
|  may|             14.65|
|  feb|             9.635|
|  sep| 19.61220930232558|
|  mar|13.083333333333336|
|  oct|17.093333333333337|
|  jul|22.109375000000004|
|  nov|              11.8|
|  apr|12.044444444444444|
|  dec| 4.522222222222222|
|  jan|              5.25|
+-----+------------------+
```

Conclusion

By leveraging tools like NiFi, HDFS, Hive, and Spark, I was able to effectively manage and analyze the Portuguese Forest Fires dataset. Despite encountering challenges, particularly in transitioning data from HDFS to Hive and addressing null values, I navigated these issues and was able to successfully perform meaningful queries using Spark. This project not only enhanced my technical skills but also deepened my understanding of data processing workflows utilizing a multi-tool pipeline.

Works Cited

Cortez, Paulo and Anbal Morais. "Forest Fires." UCI Machine Learning Repository, 2007, https://doi.org/10.24432/C5D88D.