## Assignment 5\_Aparna Suresh

### Question:Read the iot\_devices.json file.

#### Answer:

Start pyspark, with Hive support

pyspark --conf spark.sql.catalogImplementation=hive

iot\_devices = spark.read.json("iot\_devices.json")

A screenshot of a computer program

Description automatically generated

Question: Create a temporary view for it.

Answer:

iot\_devices.createOrReplaceTempView("iot\_devices\_view")

A screenshot of a computer program

Description automatically generated

Question: What percent of devices are showing red on the LCD? Use only SQL (not DataFrame) to come up with the answer to this question.

#### Answer:

spark.sql("""SELECT (COUNT(CASE WHEN lcd = 'red' THEN 1 END) \* 100.0 / COUNT(\*)) AS red\_percentage FROM iot\_devices\_view""").show()

A screenshot of a computer program

Description automatically generated

**Answer for the percentage of devices showing red** (as a text file)

red\_percentage= spark.sql("""SELECT (COUNT(CASE WHEN lcd = 'red' THEN 1 END) \* 100.0 / COUNT(\*)) AS red\_percentage FROM iot\_devices\_view""")

Text file didn’t support decimal so convert to string:

from pyspark.sql.functions import col

>>> red\_percentage\_str = red\_percentage.withColumn("red\_percentage", col("red\_percentage").cast("string"))

>>> red\_percentage\_str.coalesce(1).write \

.mode("overwrite") \

.format("text") \

.option("header", "true") \

.save("A:/SJSU/Sem-2/DATA-228-Big Data/Assignment\_5/red\_percentage.txt")

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

### Question: Among those "red" devices, do a **SQL query** that shows the count breakdown by country (sorted descending). What is the country with the most red devices and how many of them does it have?

spark.sql("""

SELECT cn AS country, COUNT(\*) AS red\_device\_count

FROM iot\_devices\_view

WHERE lcd = 'red'

GROUP BY cn

ORDER BY red\_device\_count DESC

LIMIT 1

""")

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Answer for the country with the most red devices and its count** (as a text file)

# Collect the result to the driver

top\_country =spark.sql("""

SELECT cn AS country, COUNT(\*) AS red\_device\_count

FROM iot\_devices\_view

WHERE lcd = 'red'

GROUP BY cn

ORDER BY red\_device\_count DESC

LIMIT 1

""")

result\_data = top\_country.collect()

# Open a text file for writing

with open("A:/SJSU/Sem-2/DATA-228-Big Data/Assignment\_5 /country\_most\_red-devices.txt", "w") as f:

for row in result\_data:

f.write(f"{row['country']}: {row['red\_device\_count']}\n")

A screenshot of a computer program

Description automatically generated

### Question: Create a managed table for this query

spark.sql("""CREATE TABLE red\_devices\_by\_country AS SELECT cn AS country, COUNT(\*) AS red\_device\_count FROM iot\_devices\_view WHERE lcd = 'red' GROUP BY cn ORDER BY red\_device\_count DESC""")

spark.sql("SHOW TABLES").show()

spark.sql("SELECT \* FROM red\_devices\_by\_country").show()

Drop table: