Lab 6

Introduction

In this lab, you will practice Structured APIs in the Spark shell.

Objectives

- Practice Spark's Structured APIs.

Preparation

Transfer iot_devices.json to the Peel cluster, unzip it, and store it in HDFS. Please note that all data are fictional.

Your tasks

Task 1

Use Spark's Structured APIs in the Spark shell to play around this data set.

Define a Scala case class named `DeviceIoTData` that will map to a Scala Dataset:

```
case class DeviceIoTData(
  battery_level: Long,
  co2_level: Long,
  cca2: String,
  cca3: String,
  cn: String,
  device_id: Long,
  device_name: String,
  humidity: Long,
  ip: String,
  latitude: Double,
  lcd: String,
  longitude: Double,
  scale: String,
  temp: Long,
  timestamp: Long
)
```

Read `iot_devices.json` with device information.

```
val ds = spark.read.json("iot_devices.json").as[DeviceIoTData]
```

Try the following commands:

```
ds.printSchema
ds.show(5, false)
ds.first()
```

Task 2

Use Spark's Structured APIs in the Spark shell to perform the following queries.

1. Detect failing devices whose battery levels are zero. How many are there in total? Show five of them.
2. Identify the top 5 countries with the highest levels of average CO2 emissions.

Submission

To receive full credit, please compress all of the following items into a single file named `YourName_NetID_lab6.zip` and upload it to NYU Brightspace.

- Screenshot(s) showing the commands executing.

Tips

Please use Discord if you experience any difficulties. The graders and I will help you get your environment working.

Don't procrastinate. The Hadoop cluster tends to get crowded near the due date.

Sample input data from https://github.com/databricks/LearningSparkV2.