cloudera®

Ask Bigger Questions

CAP - Developing with Spark and Hadoop:

Homework Assignment Guide for Students

Optional Homework: Partition Data Files Using Spark

Files and Data Used in this Homework

Exercise Directory: `$DEV1/exercises/spark-partfile`

Data files:

`/loudacre/devicestatus_etl/*` (HDFS)

`$DEV1DATA/status-regions.txt` (local)

Stubs: `partition-status.pyspark` / `.scalaspark`

Output directory (HDFS): `/loudacre/devstatusByRegion`

In this exercise you will use Spark to create a dataset for device status data, partitioned by region.

This exercise brings together what you have learned about using Spark for data processing with the earlier chapter on Data File Partitioning.

The Task

In the previous exercise, you calculated the five k-means data points for the locations in a data set of device status records.

Now you will use those five locations to define five *regions*: A, B, C, D and E. For each region, the location from the previous exercise is the center point.

- Save the five data points you calculated above, or use the provided data file: `$DEV1DATA/status-regions.txt` (professor VM only), which is in the format:

```
region,latitude,longitude
region,latitude,longitude
…
```

- Start with the stub files in the exercise directory (professor VM only), which provides code to read the region location datafile a dictionary with region as key, and (latitude,longitude) pairs as value. It also provides a function to parse a line of the device status data, and return the region whose center point is closest to the location for the status data.
- Read the device status data and write out the same data such that the files are in partitioned directories; that is, all status for devices in the `A` region is written in a directory called `region=A`, and so on.
- In Hive or Impala, define a table for the device status data that is partitioned by region.
- If you are using Impala, manually add the five partitions to the table using `ADD PARTITION`.

```
$ alter table devstatus_by_region
    add partition (region='A');
```

- If you are using Hive, you can use the `MSCK REPAIR TABLE` command to automatically add the partitioned directories to the table.

This is the end of the Homework