Hadoop Standard Group-by

Java implimentation of standard Group-by for Apache Hadoop/MapReduce2.

Prerequisites

To run the algorithms you will need Java version 8, Hadoop version 2.9.1, and the latest version of Eclipse.

Local Mode Walkthrough

1. Download both the java files and the dimension 200, SF1 through SF5 files.
2. Create a new project in Eclipse and place the java files in the `src` folder.
3. Add the Hadoop jars to your project.
4. In your run configuration, copy and paste the following parameter values.

Parameters:

```
200
1
28
path/to/the/input
path/to/the/output
```

Refer to lines 37-42 for a description of the input parameters.

Since Hadoop is running locally, simply specify the input path as the location where you choose to store the files on your system. The output path can be whatever you specify given that the directory does not already exist.

5. Hit run in Eclipse.

Cluster Mode Walkthrough

1. Download both the java files and the dimension 200, SF1 through SF5 files.
2. Create a new project in Eclipse and place the java files in the `src` folder.
3. Add the Hadoop jars to your project.
4. Export the project as a jar and add place the jar in your cluster.
5. As part of the submission, copy and paste the following parameter values.

Parameters:

```
200
1
28
```

```
path/to/the/input
path/to/the/output
```

Refer to lines 37-42 for a description of the input parameters.

Since Hadoop is running on a cluster, place the data files into your cluster and specficy the location as part of your input values. The output path can be whatever you specify given that the directory does not already exist.

6. Submit the job.