EECS219 Homework III: Iterative Mapreduce on Hbase

Due June 5, 11pm

Spring 2015

1. HW3 Project Assignment

In the assignment, you are to firstly create a schema within Hbase to store the data, and then implement clustering algorithm that can cluster dataset into k groups by using the map and reduce primitives. You are open to choose the algorithm you like, but you need to considering the feasibility and complexity of the work.

The dataset is from UCI machine learning repository. Basically, it is collected to analysis the energy efficiency of building with different shapes. There are totally 768 data points, and each of these points has 8 features and 2 responses.

DataSet Description & Download:

http://archive.ics.uci.edu/ml/datasets/Energy+efficiency

https://www.dropbox.com/s/9nctlvqjg3inedk/dataset.txt?dl=0

K-means Tutorials:

https://dl.dropboxusercontent.com/u/73182465/eecs219/kmeans.pdf

Data Schema:

Energy --> table name

Area | Property --> column family

X1, X5, X6, Y1, Y2 | X2, X3, X4, X7, X8

Key of each data point is rowi, like row1, row2.... row10000....

1) You program should accept four parameters:

1. hdfs/S3 path of data input
2. Number of cluster to calculate

2) Your program need to firstly delete the HTables you want to use, create data table (for storing data) and **center** table for result.

3) You need to insert K initial data points based on required schema in the **center** table.

4) You need to store these data into the Hbase through using a map-reduce job.

5) You need to implement an iterative map reduce to cluster these 768 data points into k group with the condition that data points in the same group are as similar as possible and points in different group are as different as possible.

6) You need to generate the center points of these k group data subset into a HTable called **center**.

7) Make sure to clean all the intermediate data in the end.

Hints

You may firstly need to put dataset.txt into HDFS, and has a mapper to load the data into a Hbase table call **input**. In another table **center**, it stores the initial K center points. Your mapper should load **center** as lookup table to determine which group a data point from **input** table belongs to. Then, your reduce should override the points in **center with new calculated mean.** After that, there should be iterative operation to adjust the center points in each round, but you need think about when should the program terminates.

Look Up Table Example

https://hbase.apache.org/book.html

~~Amazon EMR Test~~

- ~~**Setup EMR cluster with hbase as application to be installed**~~
- ~~Compile and package your source code to code.jar~~
- ~~Create a S3 bucket for input data, create table input and center for processing~~
- ~~Upload code.jar and input file into S3 bucket~~
- ~~Start a EMR cluster to execute the program as a job.~~

Submission Instructions

The following are requirement on your submissions, Points may be deducted if they are not followed.

- Write a report contains ~~the answers of the questions in homework and~~ the basic idea of your algorithm for this lab assignment.
- Your main class should called **Program** for the convenience of grading.
- Please organize your project in the following directory hierarchy:
  project2-studentID/codebase/code.jar
  /codebase/*.java (you can have multiple here)
  /{readme.txt, report}

- Compress project3-studentID into a single ZIP file, with the name "project3-studentID.zip"

Lab Setup

This lab helps you to setup a local instance of HBase, for the cluster based configuration please refer to Hbase's website. Before proceed, please make sure you install Java in your Linux machine or Linux virtual machine.

(1) HBase & Hadoop Version Compliance Matrix

These open source projects are actively developed, so different versions may have some functional gap. In this assignment, please use Hadoop-2.6.0, and Hbase-0.98.9

(2) Download HBase

You have to download Hbase from Apache Download Mirrors and extract the contents of the Hbase Package to a location of your choice. I am using /opt/hbase. Please use the version 0.98.9 rather than other former versions, otherwise there will some unexpected issues when integrate it with the Hadoop, which will be introduced in next two labs. Download it from the link below:

http://archive.apache.org/dist/hbase/hbase-0.98.9/

```
$ cd /opt/hbase

$ sudo tar xzf hbase-0.98.9.tar.gz

$ sudo mv hbase-0.98.9 hbase

$ sudo chown -R {username} hbase
```

(3) Configure Data Source

It is identify where to store the data. In the local configuration, we can simply use a local file folder. When we integrate HBase with Hadoop, then we can use HDFS path as the storage file path.

```
$ cd /opt/hbase/hbase/conf
```

Look into the hbase-site.xml configuration file, replace the DIRECTORY to a folder path of your local HDFS. (like hdfs://localhost:8020/hbase)

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
  <property>
    <name>hbase.rootdir</name>
    <value>file:///DIRECTORY/hbase</value>
  </property>
  <property>
    <name>hbase.zookeeper.property.dataDir</name>
    <value>/DIRECTORY/zookeeper</value>
  </property>
</configuration>
```

(4) Start HBase

$ ./opt/hbase/hbase/start-hbase.sh

you should see "starting Master, logging to logs/hbase-user-master-exmaple.org.out".

Other Reference

(1) Hbase Map Reduce Example

http://sujee.net/tech/articles/hadoop/hbase-map-reduce-freq-counter/

(2) Data Clustering Algorithms

http://en.wikipedia.org/wiki/Category:Data_clustering_algorithms