Parallel K-Means using Hadoop

Objective

This assignment aims to enhance your understanding of Apache Hadoop, also gaining experience with the MapReduce programming framework for large-scale data processing.

Description

A very common task in data analysis is grouping a set of unlabeled data such that all elements within a group are more similar among them than they are to the others. This falls under the field of unsupervised learning. Unsupervised learning techniques are widely used in several practical applications, e.g. analyzing the GPS data reported by a user to identify her most frequent visited locations. For any set of unlabeled observations, clustering algorithms tries to find the hidden structures in the data.

With the development of information technology, data volumes processed by many applications will routinely cross the peta-scale threshold, which would in turn increase the computational requirements. Efficient parallel clustering algorithms and implementation techniques are the key to meeting the scalability and performance requirements entailed in such scientific data analyses.

The Hadoop and the MapReduce programming model represents an easy framework to process large amounts of data where you can just implement the map and reduce functions and the underlying system will automatically parallelize the computations across large-scale clusters, handling machine failures, inter-machine communications, etc.

In this assignment you are asked to make a parallel version of the well-known and commonly used K-Means clustering algorithm using the map-reduce framework.

Specifications

You should implement a parallel version of the K-Means algorithm using the MapReduce framework. Then, evaluate your clustering algorithm using the IRIS dataset [1] as compared to the original one. In terms of run time and clustering accuracy.

Implement your algorithm in a generalized way (i.e can accept different sizes of feature vector).

Notes

- You do not have to implement the unparalleled K-Means.
- You should deliver a report that contains at least the following:
  - The unparalleled K-Means pseudo-code.
  - Your MapReduce K-Means algorithm.
  - The challenges you faced to implement it and how you solved it.
  - The evaluation results (using a 1 node cluster is enough)

Bonus

Configure hadoop on multiple cluster node using https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/ClusterSetup.html and test your algorithm using the publicly available dataset from the UCI repository for the comparison [2 a,b] (using 3 node clusters is enough).

Grading Policies

- You should work in groups of 2 or 3 students.
- No Late submission is allowed.
- Plagiarizing is not acceptable. Sharing code fragments between groups is prohibited and all the groups that are engaged in this action will be severely penalized. Not delivering the assignment will be much better than committing this offense.

References

[1] http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data
[2a] http://archive.ics.uci.edu/ml/datasets/Heterogeneity+Activity+Recognition
[2b]http://archive.ics.uci.edu/ml/datasets/US+Census+Data+%281990%29
[3] Lam, Chuck. Hadoop in action. Manning Publications Co., 2010.
[4] White, Tom. Hadoop: The definitive guide. O'Reilly Media, Inc., 2012.

Good Luck