NLTK Module

December 10, 2010

1 Module name

Hadoop Map-reduce

2 Scope

Hadoop Map-Reduce is a software framework for writing applications for processing large amounts of data in parallel on commodity hardware.

3 Learning objectives

These are the learning objectives for this module:

a. Be confident in the utilization and application of the *Hadoop Map-reduce* package.

b. Understand how *Hadoop Map-reduce* relates to information retrieval topics.

4 5S characteristics of the module

- Stream: Hadoop Map-Reduce is media agnostic. In principle, any kind of media that can be split into reasonable sized chunks is applicable.
- Structure: Hadoop is a number of programs that aid in working with large amounts of data in a distributes manner. The Hadoop submodule "Map-Reduce" is a package written in Java.
- Space: Hadoop Map-Reduce consists of two steps. The map step outputs key-value pairs, and the reduce step merges keys into the final key-value representation.
- Scenario: Hadoop Map-Reduce can be used in any scenario that can be reasonably mapped to the key-value data model.
- Society: Hadoop Map-Reduce is a good tool for production processing needs as well as researchers needing to work on large amounts of data.

5 Relationships with other modules

6 Prerequisite knowledge/skills required

- Java: Since Hadoop Map-Reduce is written in Java, Java knowledge is a preferable in order to use Map-Reduce for custom applications. It is not strictly required, as other languages can be used.

- Linux basis:
  - Shell basics: Running shell commands etc.
  - Filesystem basics: Navigating the filesystem.

7 Introductory remedial instruction

None

8 Body of knowledge

- Getting Started With Hadoop: MapReduce
  - Running the examples
    http://hadoop.apache.org/common/docs/current/single_node_setup.html#Execution - "Execution"
  - Understanding basic Java examples
    http://wiki.apache.org/hadoop/WordCount - "WordCount example"
  - Distrubuted indexing example
    An Introduction to Information Retrieval: Chapter 20

9 Resources

- Official Hadoop Map-Reduce Website
  http://hadoop.apache.org/mapreduce/
- Map-Reduce Tutorial on the Official Website
  http://hadoop.apache.org/common/docs/r0.17.0/mapred_tutorial.html
- An Introduction to Information Retrieval chapter 20
  By Christopher D. Manning, Prabhakar Raghavan & Hinrich Schtze

10 Exercises / Learning activities

We have prepared the exercise material on the IBM instance. The namenode required has beenformated and Hadoop Common has been started. The data files are in `"/home/idcuser/data/input/"` and the WordCount sourcefile can be found here: `"/home/idcuser/application/WordCount.java"`

1. Use "hadoop fs" to create directory for your team, also create "input" and "output" directory in their team directory.
2. Copy the local files into your "input" directory in Hadoop filesystem.
3. Create team directory in the cloud home directory, copy WordCount sourcefile into it.
4. Compile WordCount.
5. Run WordCount application with Hadoop using the "input" directory in the Hadoop filesystem as input directory and "output" as output directory.
6. Copy the output from Hadoop filesystem back to the cloud team directory.

11 Evaluation of learning objective achievement

In evaluation the learning objectives, you will be evaluated on your ability to comprehend the Map-Reduce computing model, as well as your ability to set up simple applications using Hadoop Map-Reduce.

12 Glossary

None

13 Additional useful links

- Map-Reduce paper from Google
  http://labs.google.com/papers/mapreduce.html
- Map-Reduce Wikipedia Article
  http://en.wikipedia.org/wiki/MapReduce

14 Contributors

Prepared for class CS5604 at Virginia Tech

- Initial Authors: Team 3
  - Xiaokui Shu
  - Ron Cohen
- Reviewer
  - Dr. Edward A. Fox