Hadoop Assignment
University of Amsterdam BSc Informatica
Concurrency & Parallel Programming

S. Koulouzis, A.S.Z. Belloum

1 Assignment Preparation

The goal of this assignment is to enable you to gain experience programming with:

- the Hadoop open source framework
- breaking down a task into a parallel distributed MapReduce model

To familiarize your self with the Hadoop components, you will install Hadoop on your account.

1.1 Installation

Install and configure Hadoop in your account. Start by installing it first on a single-node and then multi-node.

For the single-node setup follow this tutorial: `http://Hadoop.apache.org/docs/r1.0.3/single_node_setup.html`

For the multi-node setup follow this tutorial: `http://Hadoop.apache.org/docs/r1.0.3/cluster_setup.html`

1.2 Remarks

1.2.1 Notes on Installation

Make sure you have java 1.6 or higher. To add the java module in DAS you type:

```
module add java/1.6.0_17-amd64
```

Make sure that the environment variable `JAVA_HOME` is set correctly. You can check this by typing:

```
echo $JAVA_HOME
```

You can download Hadoop in you account by typing:

```
wget http://mirrors.sendthisfile.com/apache/hadoop/common/hadoop
    -1.0.X/hadoop-1.0.X-bin.tar.gz
```

To untar the archive you can type:

```
tar -xzvf Hadoop-1.0.X-bin.tar.gz
```

1.2.2 Notes on the multi-node setup

In the multi-node tutorial, you can skip the Sections starting from "Real-World Cluster Configurations" and ending at "Hadoop Startup".

in conf/core-site.xml add

```
<configuration>
     <property>
          <name>fs.default.name</name>
          <value>hdfs://fs0.das3.cs.vu.nl:7000</value>
     </property>
</configuration>
```

instead of using the port 7000 use 700 + your group num e.g. 7001, 7002, etc

in conf/hdfs-site.xml add:

```
<configuration>
     <property>
          <name>dfs.replication</name>
          <value>2</value>
     </property>
</configuration>
```

in conf/mapred-site.xml add:

```
<configuration>
     <property>
          <name>mapred.job.tracker</name>
          <value>fs0.das3.cs.vu.nl:8000</value>
     </property>
</configuration>
```

instead of using the port 8000 use 800 + your group num e.g. 8001, 8002, etc

To setup the master node, add in the conf/master file:
"`fsX.das3.xxxxx.nl`".

To setup the slave nodes, add in the conf/slaves file:
nodeXXX.das3.xxxxx.nl

1.3 Test the installation

After you have set up Hadoop, follow the tutorial in `http://Hadoop.apache.org/docs/r1.0.3/mapred_tutorial.html` and make sure you understand the code and the interfaces, you are going to need them for the assignment.

2 Assignment

In bioinformatics[2], sequence alignment is a way of arranging the sequences of a protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between sequences.

Alignment algorithms are powerful tools for searching for homologous or "similar" proteins in a dataset, providing a score for each sequence present in the dataset.

Pairwise Sequence Alignment (PSA) is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid).

Multiple Sequence Alignment is the alignment of three or more biological sequences of similar length. From the output of MSA applications, homology can be inferred and the evolutionary relationship between the sequences studied.

Considering the above, you will perform a PSA of a query protein sequence against a large dataset of known proteins using biojava[1]. The result of the pairwise alignment will provide you with the N (say 10) most relative proteins (the 10 most similar sequences). After you have the N sequences you will need to perform a MSA to see if there are any mutations or common ancestors of the query protein and the N most similar proteins. Then with the use of MapReduce parallelize this task and compare the execution times of the two approaches.

2.1 Implementation and Reporting

After you have studied and understood the sequential code, study the MapReduce skeleton code provided. You job is to parallelize the sequential code with the use of the MapReduce programming model. You are expected to hand in a small report (maximum 2 pages) describing the architecture of Hadoop and explain the functionality of its major components. Describe how the Hadoop architecture scales on large systems and how it handles large data. Also describe your approach for parallelizing the sequence alignment code and provide speed up measurements, by controlling the number of mappers that the framework will spawn. Keep in mind that you will be shearing the cluster, therefore you measurements will be influenced by each-others experiments. With this in mind you should perform multiple experiments analyze them statistically and present your findings (you should graph the speedup and include error bars). How is you implementation performing in terms of scale. What is the benefit of adding more mappers?If you would have to analyze even a larger dataset would you implementation cope with the increasing demand? Finally include a small section and discuss ways of optimizing your code.

2.2 Experimental environment

You will be running your experiments on DAS-3. DAS-3 is specifically NOT a cluster for doing production work. It is meant for people doing experimental research on parallel and distributed programming. This cluster is used by other people as well, which means that you have to respect certain rules. The default run time for jobs scheduled on DAS-3 is 15 minutes, which is also the maximum during working hours, from 08:00 to 20:00. DAS-3 policy is not to allow users to monopolize the clusters for a longer time, since that makes interactively running short jobs on a large number of nodes practically impossible.

Only during the night and in the weekend, when DAS-3 is regularly idle, it is allowed to run long jobs.

3 Assignment Deadline

Report due date: December 3rd

References

[1] The biojava web site. `http://www.biojava.org/`.

[2] Wikipedia. bioinformatics. `http://en.wikipedia.org/wiki/Bioinformatics`.