Leopold–Franzens–University
Innsbruck

Institute of Computer Science
Distributed and Parallel Systems

Apache Hadoop Karwendel Homework

Advanced Cloud and Grid Technologies (WS 2011/2012)

Supervisor: Dr. Radu Prodan

Martin Illecker

Innsbruck, January 17, 2012

Apache Hadoop Karwendel Homework

Martin Illecker

martin.illecker@student.uibk.ac.at

1 Exercise

Get familiar with Hadoop and HDFS. You should run the WordCount.java from the Tutorial `http://hadoop.illecker.at/Tutorial.pdf` on karwendel.dps.uibk.ac.at with different configurations.
Different configurations includes various input data and a different amount of slaves. Calculate the speedup and explain possible deviations.
Change input data, submit slaves and measure the results from the JobTracker Webinterface.

1.1 Input Data

Use the books from `http://www.gutenberg.org` which are available for free in plaintext.
For example:
`http://www.gutenberg.org/cache/epub/20417/pg20417.txt`
`http://www.gutenberg.org/cache/epub/5000/pg5000.txt`
`http://www.gutenberg.org/cache/epub/4300/pg4300.txt`

1.2 Additional Slaves

Option 4 of the HadoopSystem submits one slave. You can see the current slaves also in the JobTracker Webinterface. You have to kill them by qdel at the end.

```
$ ./hadoop 50001 50002
HadoopSystem started @ /home/lab406/hadoop_system
HadoopSystem config complete.
User: lab406
HDFS: hdfs//karwendel.dps.uibk.ac.at:50001
JobTracker: karwendel.dps.uibk.ac.at:50002
TempDir: /home/lab406/hadoop_system/hadoop-1.0.0/tmp
============================
|   HADOOP MENU SELECTION  |
============================
| Options:                 |
|  0. Exit                 |
|  1. Start Master         |
```

```
|  2. Stop  Master         |
|  3. Master Web PORTS     |
|  4. Start Slave          |
|  You have to kill Slaves |
|  by running qdel         |
============================
Select option:
```

1.3 Configuration 1

1. Input Data: 5 books, each about 2 Megabyte
2. Slave Amount: Only one Slave which is already started with the Master

1.4 Configuration 2

1. Input Data: 5 books, each about 2 Megabyte
2. Slave Amount: 5 Slaves (by submitting these via HadoopSystem)

1.5 Configuration 3

1. Input Data: 5 books, each about 20 Megabyte (concatenate books, content is irrelevant)
2. Slave Amount: Only one Slave which is already started with the Master

1.6 Configuration 4

1. Input Data: 5 books, each about 20 Megabyte (concatenate books, content is irrelevant)
2. Slave Amount: 5 Slaves (by submitting these via HadoopSystem)

1.7 Configuration 5

1. Input Data: 10 books, each about 10 Megabyte (concatenate books, content is irrelevant)
2. Slave Amount: 5 Slaves (by submitting these via HadoopSystem)