The Trials and Tribulations and ultimate success of parallelisation using Hadoop within the SCAPE project

Alastair Duncan
STFC

Pre Coffee talk STFC July 2014
SCAPE

• Scalable Preservation Environments
  • The SCAPE project develops scalable services for planning and execution of institutional preservation strategies on an open source platform that orchestrates semi-automated workflows for large-scale, heterogeneous collections of complex digital objects.
• User Story:
  
  • As the content holder/manager of scientific data held in a local format, I wish to migrate this data into a domain standard format to reduce the risks of losing the ability to read/use and reuse the data contained within the file format.
  
  • Conversion of raw format ISIS data to NeXus format
• Taverna
  • An open source and domain-independent Workflow management system – a suite of tools used to design and execute scientific workflows and aid in silico experimentation.

• Hadoop
  • The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.
  • Hadoop was created by Doug Cutting and Mike Cafarella in 2005. Cutting, who was working at Yahoo! at the time, named it after his son's toy elephant. It was originally developed to support distribution for the Nutch search engine project
  • Primary use was to build indexes for search engine using Lucene.
MapReduce

• A programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster
  • Split the data into small chunks process these in parallel (Map)
  • Collate the results (Reduce)

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
Large scale Taverna workflow using ToMaR

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
Dependencies

raw2nexus/nxconvert

↓

NeXus lib/tools

↓

HDF5
Hierarchical Data Format (HDF5)

- HDF5 supports all types of data stored digitally, regardless of origin or size.
- Petabytes of remote sensing data collected by satellites
- Terabytes of computational results from nuclear testing models
- Megabytes of high-resolution MRI brain scans
- Stored in HDF5 files
- Together with metadata necessary for efficient data sharing, processing, visualization, and archiving.

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
HDFS

- Hadoop has a highly optimised filing system
  - HDFS - distributed file system
  - highly fault-tolerant and is designed to be deployed on low-cost hardware
  - high throughput access to application data and is suitable for applications that have large data sets.

- Cons
  - Have to use Hadoop API to make use of HDFS – port application
  - Third party applications with specific file system dependencies are not suitable for porting – NeXus
  - Takes time to move big data onto and off HDFS

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
ToMaR

• Tool for wrapping 3\textsuperscript{rd} party applications for execution on Hadoop platform
  • Specify the app – defines what and how to execute
  • Create a file which lists the files to be processed
    • ToMaR copies files from HDFS to a temp directory
    • Executes the 3\textsuperscript{rd} party application over the files
    • Copies outputs back to HDFS

This work was partially supported by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
Maps and Splits

- Maps have an input file which is split into multiple pieces to be processed in parallel
- Raw and NeXus files can’t be split as they can’t use standard file operations
- ToMaR takes a file with a list of operations and the file to act on
- Split is the number of lines to read in and will be processed per map
Data sets

- Small data set 1.2Gb 33 raw files ranging in size from 24Kb to 129Mb
- Large dataset 0.97Tb 20130 raw files ranging in size from 6Kb to 456Mb
- Large dataset made up from copied files from Small Dataset 1.2Tb 33000 raw files ranging in size from 24Kb to 129Mb

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
Objects per hour vs No. of maps/splits, 33 objects

Maps/Split

Objects per hour

4/50
4/1
4/4
8/1

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
Large dataset

Objects per hour vs No. of maps

- split 33 (33000 objects)
- split 1 (33000 objects)
- split 1 big file dataset (20130 objects)

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
Big files large dataset Maps 1 Split 1

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
Hadoop job_201407021714_0001 on hadoop2

User: ad43
Job Name: tomar-1.4.1-SNAPSHOT-jar-with-dependencies.jar
Job File: hdfs://hadoop2.gridpp.rl.ac.uk/dfs/tmp/mapred/staging/ad43/.staging/job_201407021714_0001/job.xml
Submit Host: hadoop8.gridpp.rl.ac.uk
Submit Host Address: 130.246.181.179
Job-ACLs: All users are allowed
Job Setup: Successful
Status: Succeeded
Started at: Wed Jul 02 17:17:31 BST 2014
Finished at: Thu Jul 03 05:03:41 BST 2014
Finished in: 11hrs, 46mins, 9sec
Job Cleanup: Successful

<table>
<thead>
<tr>
<th></th>
<th>% Complete</th>
<th>Num Tasks</th>
<th>Pending</th>
<th>Running</th>
<th>Complete</th>
<th>Killed</th>
<th>Failed/Killed Task Attempts</th>
</tr>
</thead>
<tbody>
<tr>
<td>map</td>
<td>100.00%</td>
<td>20130</td>
<td>0</td>
<td>0</td>
<td>20130</td>
<td>0</td>
<td>0 / 3</td>
</tr>
<tr>
<td>reduce</td>
<td>100.00%</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0 / 0</td>
</tr>
</tbody>
</table>
## Task Details

### Job job_201407030853_0001

<table>
<thead>
<tr>
<th>Task Attempts</th>
<th>Machine</th>
<th>Status</th>
<th>Progress</th>
<th>Start Time</th>
<th>Finish Time</th>
<th>Errors</th>
</tr>
</thead>
</table>
| attempt_201407030853_0001_m_005995_0 | taskAttempt://default://red/hadoop0.gridpp.rl.ac.uk | FAILED   | 100.00%  | 3-Jul-2014 12:02:14 | 5-Jul-2014 12:12:24 (10 mins, 14 sec) | Task attempt_201407030853_0001_m_005995_0 failed to report status for 600 seconds. Killing!
|                        |         |          |          |            |             | Cleanup Attempt: last 1 KB All                                      |
| attempt_201407030853_0001_m_005993_1 | taskAttempt://default://red/hadoop0.gridpp.rl.ac.uk | SUCCEEDED | 100.00%  | 3-Jul-2014 12:12:32 | 3-Jul-2014 12:15:00 (2 mins, 34 sec) | Last 4 KB All                                      |

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
This work was partially supported by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
### Task Details

#### Task Log for task_201407030853_0001_m_005695

<table>
<thead>
<tr>
<th>Task Id</th>
<th>Start Time</th>
<th>Finish Time</th>
<th>Host</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>attempt_201407030853_0001_m_005695_0</td>
<td>3:07 11:08:52</td>
<td>3:07 11:19:07 (10mins, 15sec)</td>
<td>hadoop7.gridpp.rl.ac.uk</td>
<td>Task attempt_201407030853_0001_m_005695_0 failed to report status for 600 seconds. Killing!</td>
</tr>
<tr>
<td>attempt_201407030853_0001_m_005695_1</td>
<td>3:07 11:19:54</td>
<td>3:07 11:22:46 (2mins, 52sec)</td>
<td>/default-rack/hadoop8.gridpp.rl.ac.uk</td>
<td></td>
</tr>
</tbody>
</table>
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
Job Status

Hadoop job_201407080914_0001 on hadoop2

User: ad43
Job Name: tomor-1.4.1-SNAPSHOT-jar-with-dependencies.jar
Job File: hdfs://hadoop2.gridpp.rl.ac.uk/dfs/tmp/mapred/staging/ad43/staging/job_201407080914_0001/job.xml
Submit Host: hadoop8.gridpp.rl.ac.uk
Submit Host Address: 130.246.181.179
Job-ACLs: All users are allowed
Job Setup: Successful
Status: Succeeded
Started at: Tue Jul 08 09:16:22 BST 2014
Finished at: Tue Jul 08 17:15:13 BST 2014
Finished in: 7hrs, 58mins, 51sec
Job Cleanup: Successful
Black-listed TaskTrackers: 2

<table>
<thead>
<tr>
<th>Kind</th>
<th>% Complete</th>
<th>Num Tasks</th>
<th>Pending</th>
<th>Running</th>
<th>Complete</th>
<th>Killed</th>
<th>Failed/Killed Task Attempts</th>
</tr>
</thead>
<tbody>
<tr>
<td>map</td>
<td>100.00%</td>
<td>20130</td>
<td>0</td>
<td>0</td>
<td>20130</td>
<td>0</td>
<td>1046 / 3</td>
</tr>
<tr>
<td>reduce</td>
<td>100.00%</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0 / 0</td>
</tr>
</tbody>
</table>

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
### Task details

**Job job_201407080914_0001**

#### All Task Attempts

<table>
<thead>
<tr>
<th>Task Attempts</th>
<th>Machine</th>
<th>Status</th>
<th>Progress</th>
<th>Start Time</th>
<th>Finish Time</th>
<th>Errors</th>
<th>Task Logs</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>attempt_201407080914_0001_m_003332_0</td>
<td>/default-rack/hadoop2/griddpo.rl.ac.uk</td>
<td>FAILED</td>
<td>100.00%</td>
<td>8-Jul-2014 09:32:03</td>
<td>8-Jul-2014 09:42:16 (10m, 12sec)</td>
<td>Task attempt_201407080914_0001_m_003332_0 failed to report status for 600 seconds. Killing!</td>
<td>Task attempt: Last 4KB Last 9KB A1 Cleanup attempt: Last 4KB Last 9KB A1</td>
<td>21</td>
</tr>
<tr>
<td>attempt_201407080914_0001_m_003332_0</td>
<td>/default-rack/hadoop2/griddpo.rl.ac.uk</td>
<td>SUCCEEDED</td>
<td>100.00%</td>
<td>8-Jul-2014 09:42:17</td>
<td>8-Jul-2014 09:44:37 (2m, 20sec)</td>
<td></td>
<td>Last 4KB Last 9KB A1</td>
<td>21</td>
</tr>
</tbody>
</table>

This work was partially supported by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
Job status

Hadoop job_201407080914_0002 on hadoop2

User: ad43
Job Name: tomar-1.4.1-SNAPSHOT-jar-with-dependencies.jar
Job File: hdfs://hadoop2.gridpp.rl.ac.uk/dfs/tmp/mapred/staging/ad43/staging/job_201407080914_0002/job.xml
Submit Host: hadoop8.gridpp.rl.ac.uk
Submit Host Address: 130.246.181.179
Job-ACLs: All users are allowed
Job Setup: Successful
Status: Failed
Failure Info: NA
Started at: Tue Jul 08 17:21:37 BST 2014
Failed at: Tue Jul 08 18:29:25 BST 2014
Failed in: 1hrs, 7mins, 48sec
Job Cleanup: Successful
Black-listed TaskTrackers: 2

<table>
<thead>
<tr>
<th>Kind</th>
<th>% Complete</th>
<th>Num Tasks</th>
<th>Pending</th>
<th>Running</th>
<th>Complete</th>
<th>Killed</th>
<th>Failed/Killed Task Attempts</th>
</tr>
</thead>
<tbody>
<tr>
<td>map</td>
<td>100.00%</td>
<td>5033</td>
<td>0</td>
<td>0</td>
<td>830</td>
<td>4203</td>
<td>192 / 63</td>
</tr>
<tr>
<td>reduce</td>
<td>100.00%</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0 / 1</td>
</tr>
</tbody>
</table>

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
Failed task details

Job job_201407080914_0002

<table>
<thead>
<tr>
<th>Task Attempts</th>
<th>Machine</th>
<th>Status</th>
<th>Progress</th>
<th>Start Time</th>
<th>Finish Time</th>
<th>Errors</th>
<th>Task Logs</th>
<th>Counters</th>
</tr>
</thead>
<tbody>
<tr>
<td>attempt_201407080914_0002_m_000833_0</td>
<td>Task attempt: /default-rack/hadoop2, gridpp.rl.ac.uk Cleanup Attempt: /default-rack/hadoop2, gridpp.rl.ac.uk</td>
<td>FAILED</td>
<td>100.00%</td>
<td>8-Jul-2014 17:35:58</td>
<td>8-Jul-2014 17:40:11 (10mins, 13sec)</td>
<td>Task attempts_201407080914_0002_m_000833_0 failed to report status for 600 seconds. Killing!</td>
<td></td>
<td>21</td>
</tr>
<tr>
<td>attempt_201407080914_0002_m_000833_1</td>
<td>Task attempt: /default-rack/hadoop2, gridpp.rl.ac.uk Cleanup Attempt: /default-rack/hadoop2, gridpp.rl.ac.uk</td>
<td>FAILED</td>
<td>100.00%</td>
<td>8-Jul-2014 17:46:15</td>
<td>8-Jul-2014 17:57:43 (11mins, 27sec)</td>
<td>Task attempts_201407080914_0002_m_000833_1 failed to report status for 600 seconds. Killing!</td>
<td></td>
<td>21</td>
</tr>
<tr>
<td>attempt_201407080914_0002_m_000833_2</td>
<td>Task attempt: /default-rack/hadoop2, gridpp.rl.ac.uk Cleanup Attempt: /default-rack/hadoop2, gridpp.rl.ac.uk</td>
<td>FAILED</td>
<td>100.00%</td>
<td>8-Jul-2014 17:57:46</td>
<td>8-Jul-2014 18:10:17 (15mins, 31sec)</td>
<td>Task attempts_201407080914_0002_m_000833_2 failed to report status for 600 seconds. Killing!</td>
<td></td>
<td>21</td>
</tr>
<tr>
<td>attempt_201407080914_0002_m_000833_3</td>
<td>/default-rack/hadoop6, gridpp.rl.ac.uk</td>
<td>KILLED</td>
<td>100.00%</td>
<td>8-Jul-2014 18:14:38</td>
<td>8-Jul-2014 18:29:19 (14mins, 41sec)</td>
<td>Last 4KB Last 8KB All Cleanup attempt: Last 4KB Last 8KB All</td>
<td></td>
<td>21</td>
</tr>
</tbody>
</table>

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
Some scalability to be had running raw2nexus on Hadoop

Hadoop not designed for use with files types that can’t be split/streamed

Significant effort required to achieve scientific file type compatibility with HDFS in order not to have to copy to local hard drive. Requires porting of NeXus tools and HDF5 tools (NetCDF may already do some of this) May not be worthwhile?

Problems will be encountered with much larger file migration. Hadoop may not be the best platform to use in this case.

Conclusions

This work was partially supported by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).