Yahoo! Grid Services

Where Grid Computing at Yahoo! is Today

Marco Nicosia
Grid Services Operations
marco@yahoo-inc.com

What is Apache Hadoop?

- Distributed File System and Map-Reduce programming platform combined
  - Designed to scale out
    - Must satisfy requirements of a full-scale WWW content system
    - SAN and NAS devices don't support enough storage or IO bandwidth
  - Combine the storage and compute power of any set of computers
  - Highly portable: Framework in Java, user code in user's preferred language
- Apache Software Foundation Open Source
  - Originally part of Lucene, now a sub-project
  - Yahoo! does not maintain a separate source code repository
  - You can download and use what I'm using today

Why Open Source?

- Infrastructures that scale are not common
  - If our infrastructure becomes popular, we'll be able to hire people who are already familiar with our technology
- Execution should not be a competitive advantage
  - The game is killer search engine results and monetization
  - My job is providing hosted infrastructure to Yahoo!'s researchers, scientists, architects and developers
- Yahoo! Search Technology was comprised of several search companies: Inktomi, Altavista, Fast, Overture
  - We are behind in infrastructure technology
    - Companies in difficult situations don't invest in infrastructure
  - By releasing our work, we all will catch up more quickly

A Quick Timeline

- 2004 - Initial versions of what is now Hadoop Distributed File System and Map-Reduce implemented by Doug Cutting & Mike Cafarella
- December 2005 - Nutch ported to the new framework. Hadoop runs reliably on 20 nodes.
- January 2006 - Doug Cutting joins Yahoo!
- February 2006 - Apache Hadoop project official started to support the standalone development of Map-Reduce and HDFS.
- March 2006 - Formation of the Yahoo! Hadoop team
- May 2006 - Yahoo sets up a Hadoop research cluster - 300 nodes
- April 2006 - Sort benchmark run on 188 nodes in 47.9 hours
- May 2006 - Sort benchmark run on 500 nodes in 42 hours (better hardware than April benchmark)
- October 2006 - Research cluster reaches 600 Nodes
- December 2006 - Sort times 20 nodes in 1.8 hrs, 100 nodes in 3.3 hrs, 500 nodes in 5.2 hrs, 900 nodes in 7.8
- January 2006 - Research cluster reaches 900 node
- April 2007 - Research clusters - 2 clusters of 1000 nodes

Interacting with the HDFS

- Simple commands: hadoop dfs -ls, -du, -rm, -rmr
- Uploading files
  - hadoop dfs -put foo mydata/foo
  - cat ReallyBigFile | hadoop dfs -put - mydata/ReallyBigFile
- Downloading files
  - hadoop dfs -get mydata/foo foo
  - hadoop dfs -get - mydata/ReallyBigFile | grep “the answer is”
  - hadoop dfs -cat mydata/foo
- File Types
  - Text files
  - SequenceFiles
    - Key/Value pairs formatted for the framework to consume
    - Per-file type information (key class, value class)

Browse the HDFS in a Browser

NameNode 'kry-nn1.[redacted].com:8020'

Started:	Mon Aug 13 07:25:08 UTC 2007
Version:	0.13.1, r558872
Compiled:	Mon Jul 23 22:07:51 UTC 2007 by hadoopqa

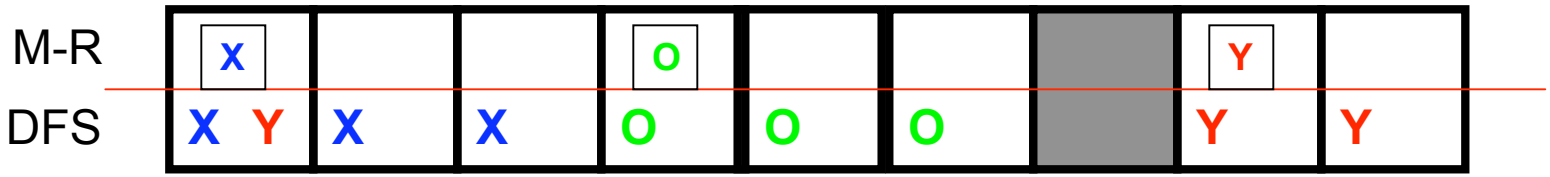
Browse the filesystem

Cluster Summary

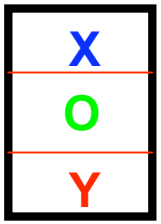
Capacity	:	1.46 PB
Remaining	:	63.49 TB
Used	:	95.77 %
Live Nodes	:	1003
Dead Nodes	:	158

Hadoop: Two Services in One

Cluster Nodes run both DFS and M-R



Input File
(128MB blocks)



Map-Reduce Demo

- hod -m 10
- run jar hadoop-examples.jar wordcount -r 4 /data/vespanews/20070218 /user/marco/wc.out

```
60x12
% hod -m 10
HDFS UI on kry-nn1.[illegible].com:50070
Mapred UI on kry1787:56208

Hadoop config file in: hod/tmp/hod-17613-tmp/hadoop-site.xml

allocation information:

     1 job tracker node
     9 task tracker nodes
    10 nodes in total
%
```

```
84x12
[hod] (marco) >>
[hod] (marco) >> run jar hadoop-examples.jar wordcount -r 4 /data/vespanews/20070218
 /user/marco/wc.out
07/09/04 21:26:19 INFO mapred.FileInputFormat: Total input paths to process: 4
07/09/04 21:26:25 INFO mapred.JobClient: Running job: job_0002
07/09/04 21:26:26 INFO mapred.JobClient:  map 0% reduce 0%
07/09/04 21:26:35 INFO mapred.JobClient:  map 1% reduce 0%
07/09/04 21:26:37 INFO mapred.JobClient:  map 3% reduce 0%
07/09/04 21:26:38 INFO mapred.JobClient:  map 4% reduce 0%
07/09/04 21:26:40 INFO mapred.JobClient:  map 5% reduce 0%
07/09/04 21:26:32 INFO mapred.JobClient:  map 6% reduce 0%
%
```

Running job: One failed task automatically retried

Hadoop job_0002 on kry1787

User: marco
Job Name: wordcount
Job File: /mapredsystem/kry1787/submit_k2sh7k/job.xml
Started at: Tue Sep 04 21:26:25 UTC 2007
Status: Running

Kind	% Complete	Num Tasks	Pending	Running	Complete	Killed	Failed/Killed Task Attempts
map	100.00%	41	0	0	41	0	0 / 0
reduce	93.45%	4	0	1	3	0	1 / 0

	Counter	Map	Reduce	Total
org.apache.hadoop.examples.WordCount$Counter	WORDS	579,909,134	0	579,909,134
	VALUES	579,909,134	53,428,324	633,337,458
Map-Reduce Framework	Map input records	60,490,175	0	60,490,175
	Map output records	579,909,134	0	579,909,134
	Map input bytes	5,447,617,164	0	5,447,617,164
	Map output bytes	7,695,121,884	0	7,695,121,884
	Combine input records	579,909,134	0	579,909,134
	Combine output records	66,750,520	0	66,750,520
	Reduce input groups	0	12,621,009	12,621,009
	Reduce input records	0	53,428,325	53,428,325
	Reduce output records	0	12,621,008	12,621,008

File: /user/marco/wc.out/part-00000

Go back to dir listing
Advanced view/download options

View Next chunk View Prev chunk

```
Moraga  88
Moraga, 24
Moraga/Fotos    1
Moragahakanda   9
Moragahakanda,  3
Moragne 4
Moragne,        1
Moragues.       4
Moraine-Pewaukee        4
Moraine.        3
Morais  147
Morais's        2
Morais, 80
Moraka  5
Moraka, 15
Morakul,        1
Moral"  1
Moral," 1
Moral-Luc??a    3
Moral-Spritze   2
Moral.  18
Moral:  7
Moral?]]></body>        1
Morala]]></title>       2
Moralas 1
```

Download this file
TAIL this file
Chunk Size to view (in bytes, upto file's DFS blocksize): 1048576 Refresh

Total number of blocks: 2

blk_5989985298053736360: 72.30.126.73:50010 72.30.63.44:50010 72.30.63.13:50010
blk_7159560097026934642: 72.30.63.234:50010 72.30.63.24:50010 72.30.63.38:50010

Streaming Map-Reduce Demo

```
run jar hadoop-streaming.jar -numReduceTasks 4
    -input /data/vespanews/20070218 -output wc-streaming.out
    -mapper "perl -ane 'print join(\"\n\", @F), \"\n\"'"
    -reducer "uniq -c"
```

File: /user/marco/wc-streaming.out/part-00000

Goto : /user/marco/wc-streamin go

Go back to dir listing
Advanced view/download options

View Next chunk View Prev chunk

```
     88 Moraga
     24 Moraga,
      1 Moraga/Fotos
      9 Moragahakanda
      3 Moragahakanda,
      4 Moragne
      1 Moragne,
      4 Moragues.
      4 Moraine-Pewaukee
      3 Moraine.
    147 Morais
      2 Morais's
     80 Morais,
      5 Moraka
     15 Moraka,
      1 Morakul,
      1 Moral"
      1 Moral,"
      3 Moral-Luc??a
      2 Moral-Spritze
     18 Moral.
      7 Moral:
      1 Moral?]]></body>
      2 Morala]]></title>
      1 Moralas
```

Download this file
TAIL this file
Chunk Size to view (in bytes, upto file's DFS blocksize): 2097152 Refresh

Total number of blocks: 2

org.apache.hadoop.dfs.LocatedBlock@f23851:	74.6.130.87:50010	72.30.62.167:50010	72.30.62.152:50010
org.apache.hadoop.dfs.LocatedBlock@4f88cb:	72.30.62.152:50010	72.30.62.213:50010	72.30.62.149:50010

PIG

- A high-level declarative language (similarity to SQL)
  - Pig Latin is a simple query algebra that lets you express data transformations such as merging data sets, filtering them, and applying functions to records or groups of records. Users can create their own functions to do special-purpose processing.
- A SW layer above MR, implementing a grouping syntax

- Wordcount example in PigLatin:

```
input   = LOAD 'documents' USING StorageText();
words   = FOREACH input GENERATE FLATTEN(Tokenize(*));
grouped = GROUP words BY $0;
counts  = FOREACH grouped GENERATE group, COUNT(words);
```

Questions?

http://lucene.apache.org/hadoop/

http://developer.yahoo.com/blogs/hadoop/

Marco Nicosia
Grid Services Operations
Yahoo! Inc.
marco@yahoo-inc.com

I Am Hiring!