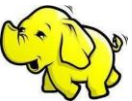

Reducers
File System
Cassandra
Flume
Combiners
Distributed
MapReduce
HBase
HiveQL
API
Sqoop
MRUnit
Hive
Big Data
Hadoop
Petabytes
Oozie
Zookeeper
Architecture
HDFS
YARN
PIG
Mappers
Partitioners
TaskTracker
Localization
JobTracker
MongoDB

Course Content

- **Module1**
  - BigData Introduction
- **Module2**
  - HDFS Installation and Commands
- **Module 3**
  - Deployment Modes
- **Module 4**
  - Mapreduce
- **Module 5**
  - Advance Mapreduce - Part1
  - Join and Counters
- **Module 6**
  - Advance MapReduce - Part 2
  - Custom Input Formats and MRUnit
- **Module 7**
  - Pig and Pig Latin
- **Module 8**
  - Hive, Hive QL and HIve Architecture
- **Module 9**
  - Advance Hive QL, Rollups and Custom Functions

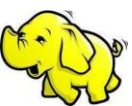
- **Module 10**
  - Flume
- **Module 11**
  - Sqoop
- **Module 12**
  - Oozie
- **Module 13**
  - NoSQL Databases
- **Module 14**
  - MongoDB and Cassandra
- **Module 15**
  - Hbase and Advanced Hbase
- **Module 16**
  - Zookeeper
- **Module 17**
  - Hadoop 2.0
  - HA and YARN and MRV2
- **Module 18**
  - Project



DeZyre

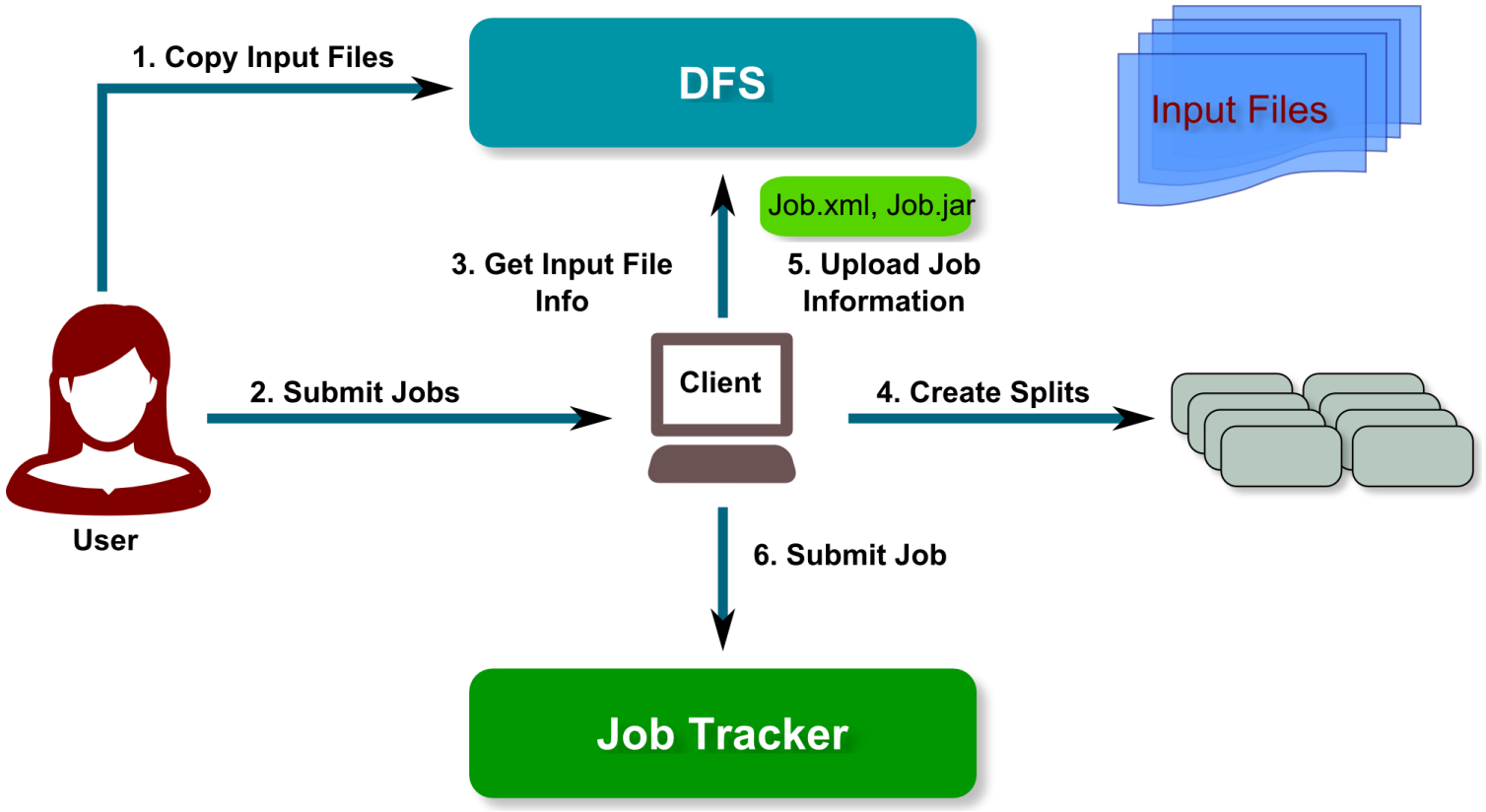
Agenda for the day

1. Understanding Job Tracker
2. Hadoop Modes
3. Hadoop Terminal Commands
4. Production Hadoop Clusters
5. Cluster Configuration
6. Hadoop Configuration Files
7. Recovery
8. MapReduce in Action

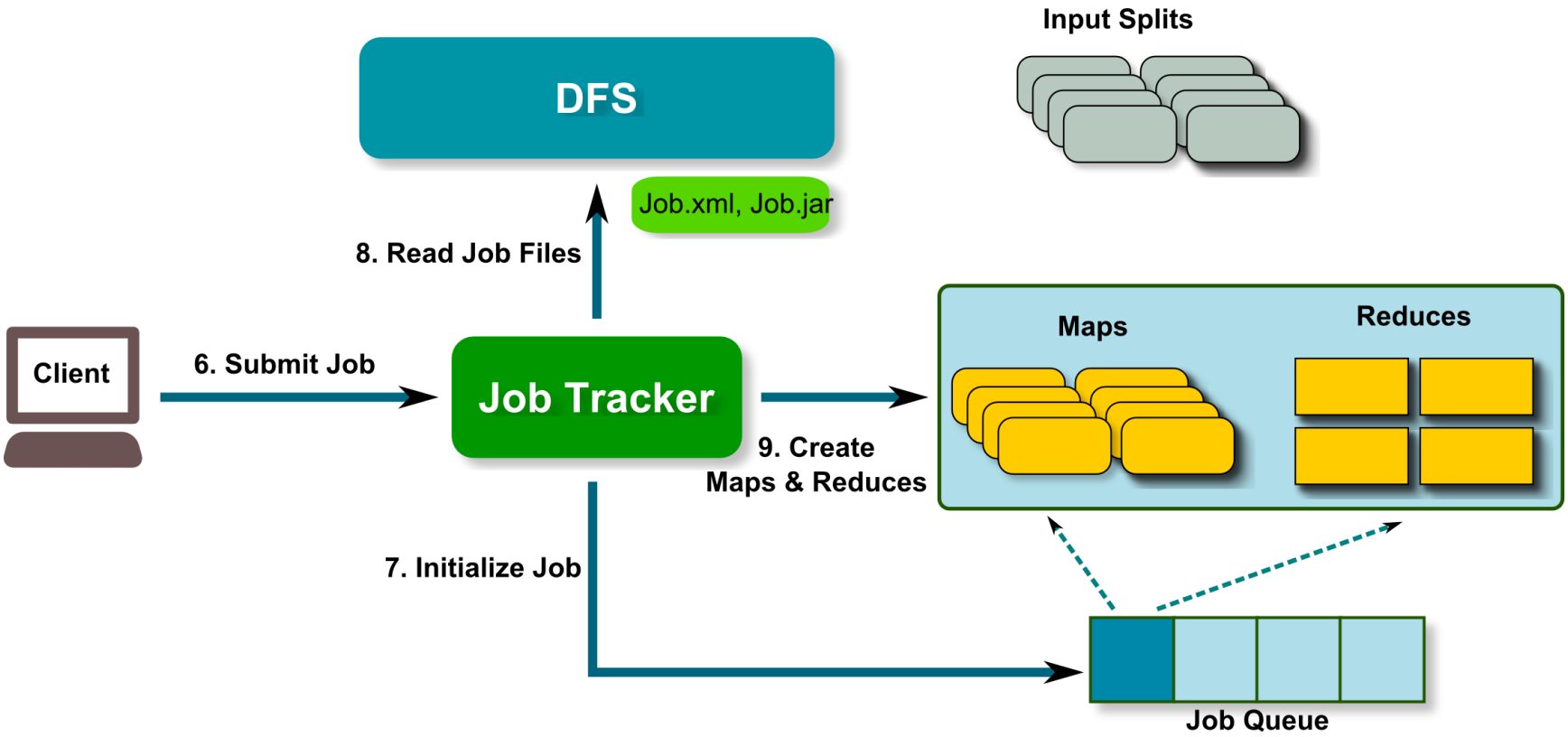


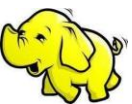
DeZyre

Job Tracker



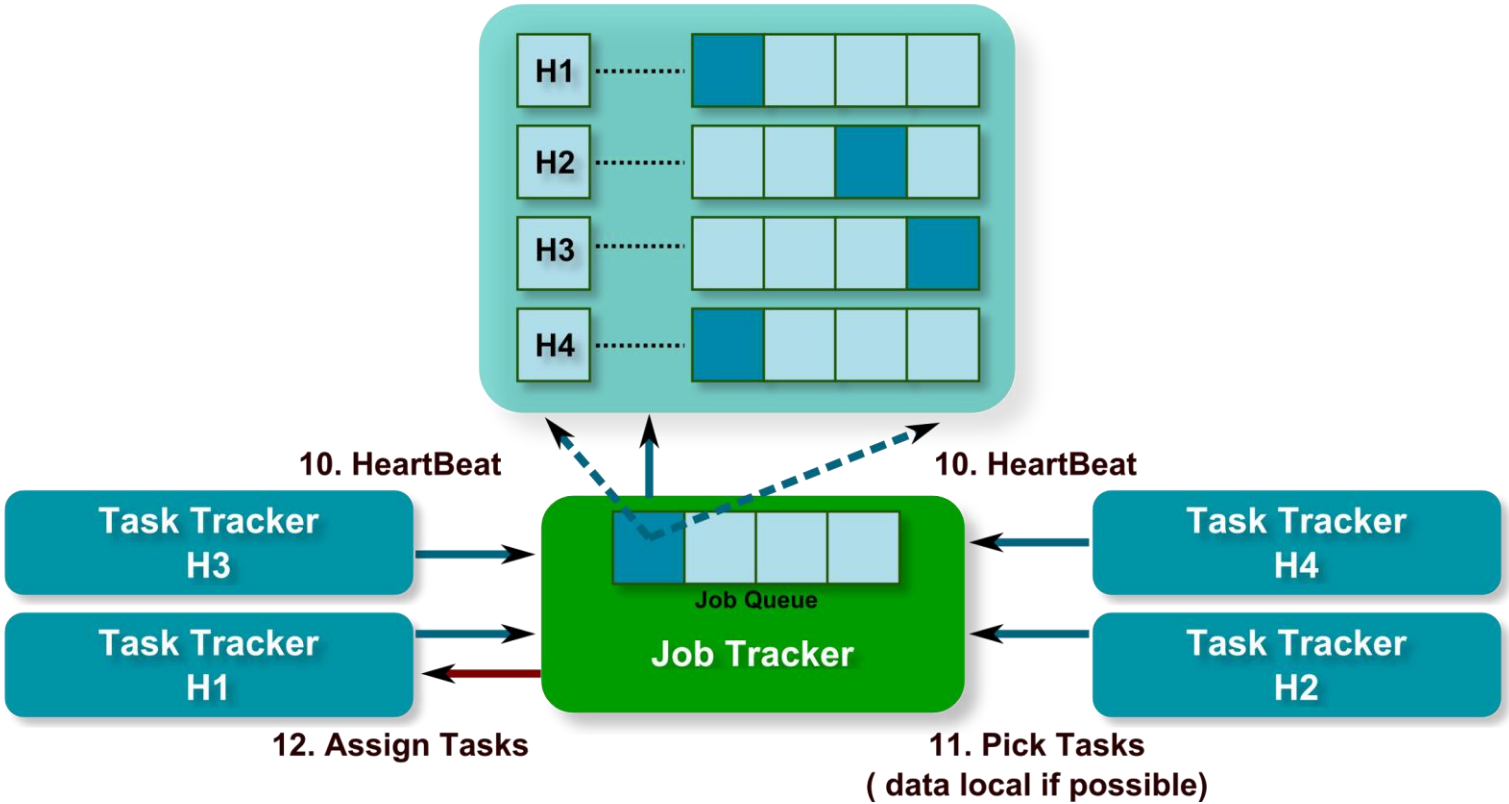
Job Tracker

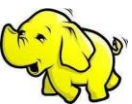




DeZyre

Job Tracker

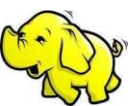




DeZyre

Job Tracker

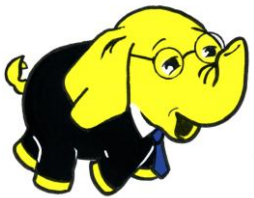
- MapReduce Master : delegating jobs to task tracker
  - Client submits jobs to job tracker , jobs are kept in queue
  - FIFO Scheduler
- Capacity Scheduler
- Job Tracker Determines the location of data through Name Node
- Job Tracker determines available task tracker (prefers the slots near to the data)
- Job Tracker submits the work to Task Tracker
- Task Tracker monitors it and send update to Job Tracker
- After completion Job Tracker Updates its Status
- Job Tracker is a single point of failure



Hadoop Wants to Know?

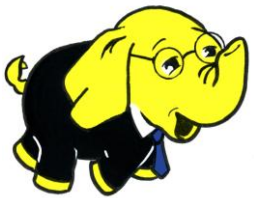
Which of these is responsible to assign a task to TaskTracker ?

a) Namenode
b) Jobtracker
c) Secondary NameNode
d) Data Node

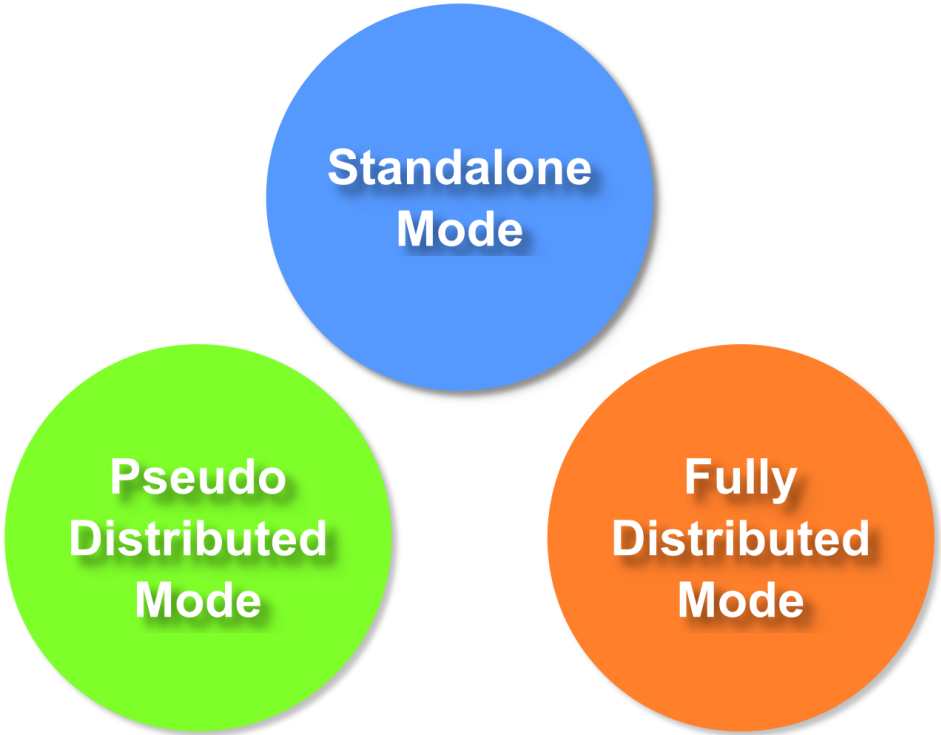


Hadoop Wants to Know?

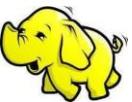
Which of these is responsible for executing a task ?
a) Namenode
b) Jobtracker
c) TaskTracker
d) Data Node



Hadoop Modes



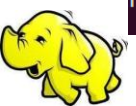
- Standalone (or Local) Mode
  - No daemons, everything runs in a single JVM.
  - Suitable for running MapReduce programs during development.
  - Has no DFS.
- Pseudo-Distributed Mode
  - Hadoop daemons run on the local machine.
- Fully Distributed Mode
  - Hadoop daemons run on a cluster of machines.



DeZyre

Terminal Commands

```
hadoopjob@ubuntu:~$ hadoop
Warning: $HADOOP_HOME is deprecated.

Usage: hadoop [--config confdir] COMMAND
where COMMAND is one of:
  namenode -format     format the DFS filesystem
  secondarynamenode    run the DFS secondary namenode
  namenode             run the DFS namenode
  datanode             run a DFS datanode
  dfsadmin             run a DFS admin client
  mradmin              run a Map-Reduce admin client
  fsck                 run a DFS filesystem checking utility
  fs                   run a generic filesystem user client
  balancer             run a cluster balancing utility
  oiv                  apply the offline fsimage viewer to an fsimage
  fetchdt              fetch a delegation token from the NameNode
  jobtracker           run the MapReduce job Tracker node
  pipes                run a Pipes job
  tasktracker          run a MapReduce task Tracker node
  historyserver        run job history servers as a standalone daemon
  job                  manipulate MapReduce jobs
  queue                get information regarding JobQueues
  version              print the version
  jar <jar>            run a jar file
  distcp <srcurl> <desturl> copy file or directories recursively
  distcp2 <srcurl> <desturl> DistCp version 2
  archive -archiveName NAME -p <parent path> <src>* <dest> create a hadoop archi
ve
  classpath            prints the class path needed to get the
                       Hadoop jar and the required libraries
  daemonlog            get/set the log level for each daemon
 or
  CLASSNAME            run the class named CLASSNAME
Most commands print help when invoked w/o parameters.
hadoopjob@ubuntu:~$
```



DeZyre

Terminal Commands

Listing of Files present on HDFS

```
hadoopjob@ubuntu:~$ hadoop fs -ls /
Warning: $HADOOP_HOME is deprecated.

Found 4 items
drwxr-xr-x   - stratapps supergroup          0 2014-02-12 09:56 /data
drwxr-xr-x   - stratapps supergroup          0 2014-02-12 10:19 /datasets
drwxr-xr-x   - stratapps supergroup          0 2014-02-12 16:47 /tmp
drwxr-xr-x   - stratapps supergroup          0 2014-02-12 15:53 /user
```

Listing of files present in bin Directory

```
hadoopjob@ubuntu:~$ cd /usr/local/hadoop/bin
hadoopjob@ubuntu:/usr/local/hadoop/bin$ ls
hadoop            hadoop-daemons.sh  start-all.sh       start-jobhistoryserver.sh  stop-balancer.sh          stop-mapred.sh
hadoop-config.sh  rcc                start-balancer.sh  start-mapred.sh            stop-dfs.sh               task-controller
hadoop-daemon.sh  slaves.sh          start-dfs.sh       stop-all.sh                stop-jobhistoryserver.sh
hadoopjob@ubuntu:/usr/local/hadoop/bin$
```

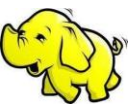
Web UI URL's

Name Node Status : http://localhost:50070/dfshealth.jsp

Job Tracker Status : http://localhost:50030/jobtracker.jsp

Task Tracker Status : http://localhost:50060/tasktracker.jsp

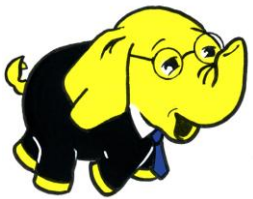
Data Block Scanner Report : http://localhost:50075/blockScannerReport



DeZyre

Hadoop Wants to Know?

Which of these is used for uploading data from local to HDFS?
a) Hadoop dfs -ls /input
b) Hadoop dfs -get /input/data.json
c) hadoop dfs -mkdir /input/data
d) Hadoop dfs -copyFromLocal /input/data/data1.json /input/data/



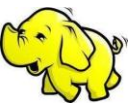
DeZyre

Real Life Hadoop Implementations

- **EBay**
  - 532 nodes cluster (8 * 532 cores, 5.3PB).
  - Heavy usage of Java MapReduce, Apache Pig, Apache Hive, Apache HBase
  - Using it for Search optimization and Research.

- **Facebook**
  - Uses Apache Hadoop to store copies of internal log and dimension data sources and use it as source for reporting/analytics and machine learning.
  - Currently FB have 2 major clusters:
  - A 1100-machine cluster with 8800 cores and about 12 PB raw storage.
  - A 300-machine cluster with 2400 cores and about 3 PB raw storage.
  - Each (commodity) node has 8 cores and 12 TB of storage.
  - FB are heavy users of both streaming as well as the Java APIs. Have built a higher level data warehousing framework using these features called Hive

(see the http://hadoop.apache.org/hive/).Have also developed a FUSE implementation over HDFS.



DeZyre

Real Life Hadoop Implementations

- **Spotify**

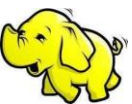
  - Uses Apache Hadoop for content generation, data aggregation, reporting and analysis
  - 690 node cluster = 8280 physical cores, 38TB RAM, 28 PB storage (read more about Hadoop issues while growing fast: Hadoop Adventures At Spotify)
  - 7,500 daily Hadoop jobs (scheduled by Luigi, our home-grown and recently open-sourced job scheduler - https://vimeo.com/63435580

http://www.slideshare.net/AdamKawa/hadoop-adventures-at-spotify-strata-conference-hadoop-world-2013

http://files.meetup.com/5139282/SHUG%201%20-%20Hadoop%20at%20Spotify.pdf

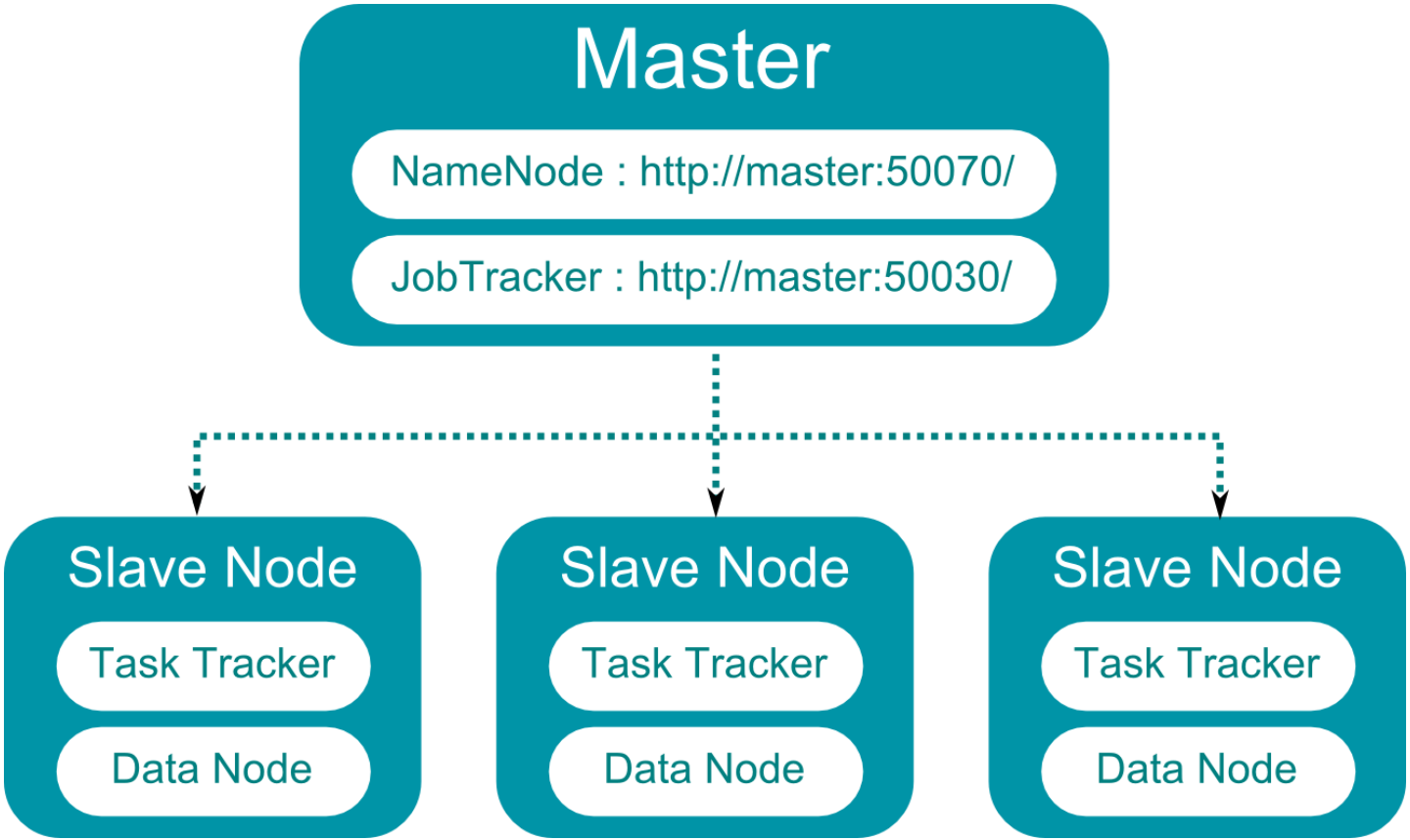
- **Last.fm**

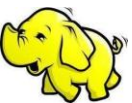
  - 100 nodes
  - Dual quad-core Xeon L5520 @ 2.27GHz & L5630 @ 2.13GHz , 24GB RAM, 8TB(4x2TB)/node
  - Used for charts calculation, royalty reporting, log analysis, A/B testing, dataset merging
  - Also used for large scale audio feature analysis over millions of tracks



DeZyre

Sample Cluster Configuration

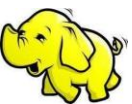




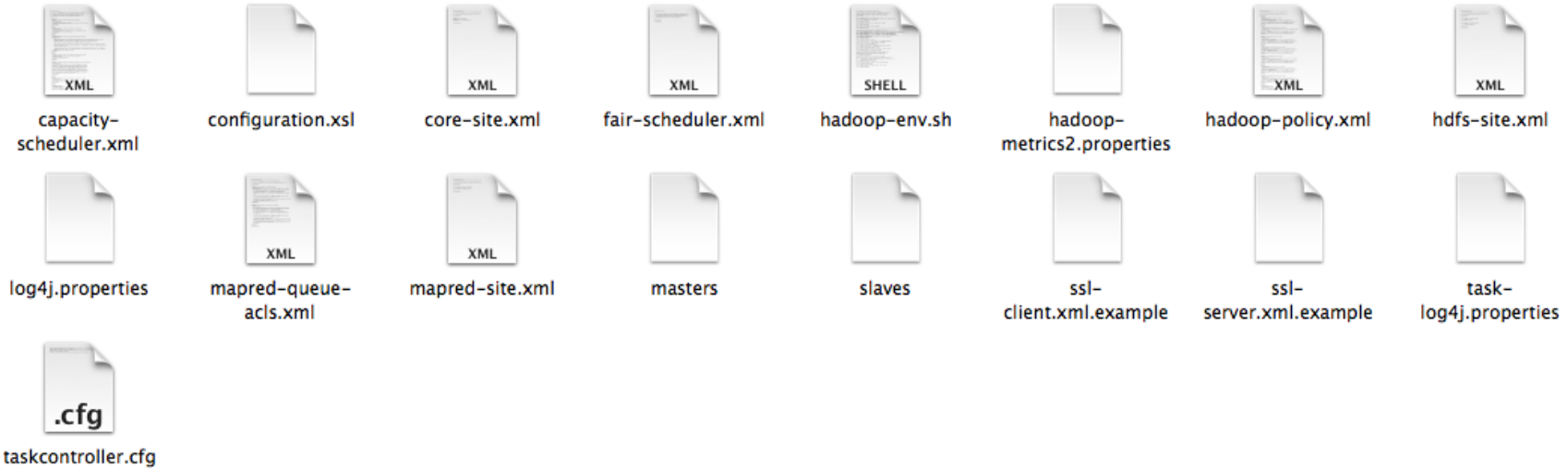
DeZyre

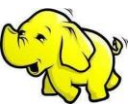
Hadoop Configuration Files

Configuration Files	Format	Description
hadoop-env.sh	Base Script	Environment variables that are used in the scripts to run Hadoop.
core-site.xml	Hadoop Configuration XML	Configuration settings for Hadoop Core such as I/O settings that are common to HDFS and MapReduce.
hdfs-site.xml	Hadoop Configuration XML	Configuration settings for HDFS daemons, the namenode, the secondary namenode and the data nodes.
mapred-site.xml	Hadoop Configuration XML	Configuration settings for MapReduce daemons : the job-tracker and the task-trackers.
masters	Plain Text	A list of machines (one per line) that each run a secondary namenode.
slave	Plain Text	A list of machines (one per line) that each run a datanode and a task-tracker.
hadoop- metric. properties	Java Properties	Properties for controlling how metrics are published in Hadoop.
log4j.properties	Java Properties	Properties for system log files, the namenode audit log and the task log for the task-tracker child process.



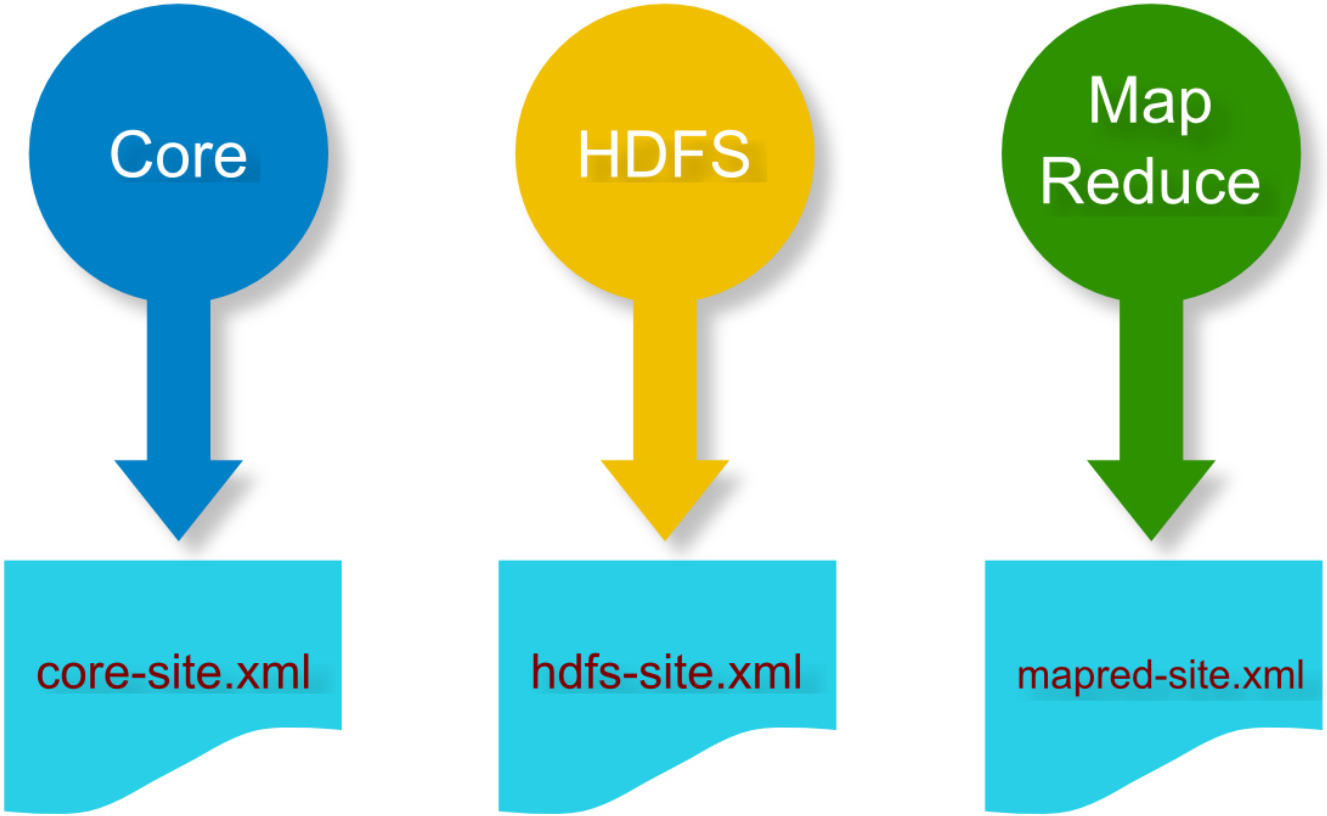
Hadoop Configuration Files

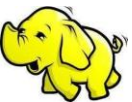




DeZyre

DD for each Component

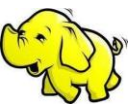




DeZyre

Core-site.xml and hdfs-site.xml

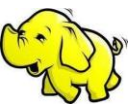
hdfs-site.xml	core-site.xml
<?xml version -"1.0"?>	<?xml version -"1.0"?>
<!--hdfs-site.xml-->	<!--core-site.xml-->
<configuration>	<configuration>
<property>	<property>
<name>dfs.replication</name>	<name>fs.default.name</name>
<value>1</value>	<value>http://localhost:8020</value>
</property>	</property>
</configuration>	</configuration>



DeZyre

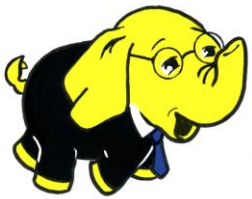
Defining HDFS details in hdfs-site.xml

Property	Value	Description
dfs.data.dir	<value> /disk1/hdfs/data, /disk2/hdfs/data </value>	A list of directories where the datanode stores blocks. Each block is stored in only one of these directories. ${hadoop.tmp.dir}/dfs/data
fs.checkpoint.dir	<value> /disk1/hdfs/namesecondary, /disk2/hdfs/namesecondary </value>	A list of directories where the secondary namenode stores checkpoints. It stores a copy of the checkpoint in each directory in the list ${hadoop.tmp.dir}/dfs/name secondary



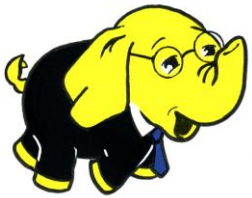
Hadoop Wants to Know?

Can you provide two different data directory locations for a single data node?
a) True
b) False



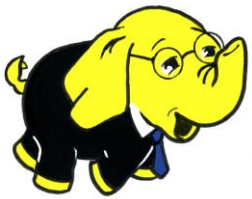
Hadoop Wants to Know?

Default Replication Factor in HDFS ?
a) 2
b) 3
c) 4
d) None



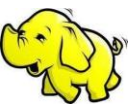
Hadoop Wants to Know?

Which of following daemons would be running on Slave Machine ?
a) DataNode
b) Secondary NameNode
c) NameNode
d) TaskTracker
e) JobTracker



mapred-site.xml

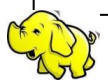
mapred-site.xml
<?xml version -"1.0"?>
<!—mapred-site.xml-->
<configuration>
<property>
<name>mapred.job.tracker</name>
<value>localhost:8021</value>
</property>
<configuration>



DeZyre

Defining mapred-site.xml

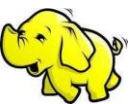
Property	Value	Description
mapred.job.tracker	<value> localhost:8021 </value>	The hostname and the port that the jobtracker RPC server runs on. If set to the default value of local, then the jobtracker runs in-process on demand when you run a MapReduce job.
mapred.local.dir	${hadoop.tmp.dir}/mapred/local	A list of directories where MapReduce stores intermediate data for jobs. The data is cleared out when the job ends.
mapred.system.dir	${hadoop.tmp.dir}/mapred/system	The hostname and the port that the jobtracker RPC server runs on. If set to the default value of local, then the jobtracker runs in-process on demand when you run a MapReduce job.
mapred.tasktracker. map/reducer .tasks.maximum	2	The number of map/reducer tasks that may be run on a tasktracker at any one time



All Properties

http://hadoop.apache.org/docs/r1.1.2/core-default.html

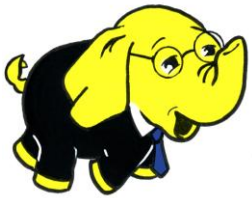
http://hadoop.apache.org/docs/r1.1.2/mapred-default.htm
l
http://hadoop.apache.org/docs/r1.1.2/hdfs-default.html



DeZyre

Hadoop Wants to Know?

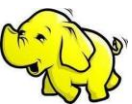
Which configuration file contains information about task trackers?
a) /conf/masters
b) /conf/slaves
c) /conf/mapred-site.xml
d) /conf/hdfs-site.xml



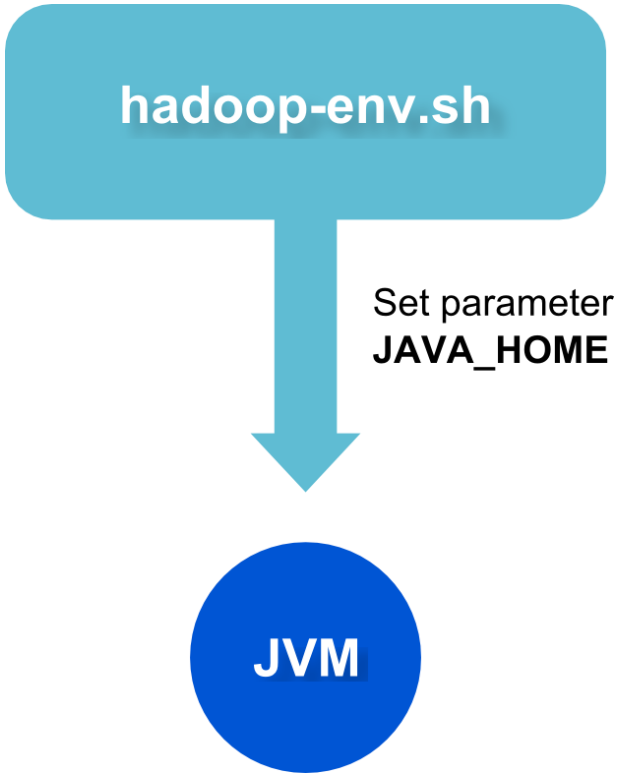
Slaves and Masters

Two files are used by the startup and shutdown commands:

- **Slaves**
  - Contains a list of hosts, one per line, that are to host DataNode and TaskTracker servers.
- **Masters**
  - Contains a list of hosts, one per line, that are to host Secondary NameNode servers.



Per Process Run-time Environment



- This file also offers a way to provide custom parameters for each of the servers.
- Hadoop-env.sh is sourced by all of the Hadoop Core scripts provided in the conf/directory of the installation.

Examples of environment variables that you can specify:

Export:

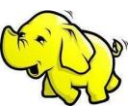
HADOOP_DATANODE_HEAPSIZE="128"

Export :

HADOOP_TASKTRACKER_HEAPSIZE="512"

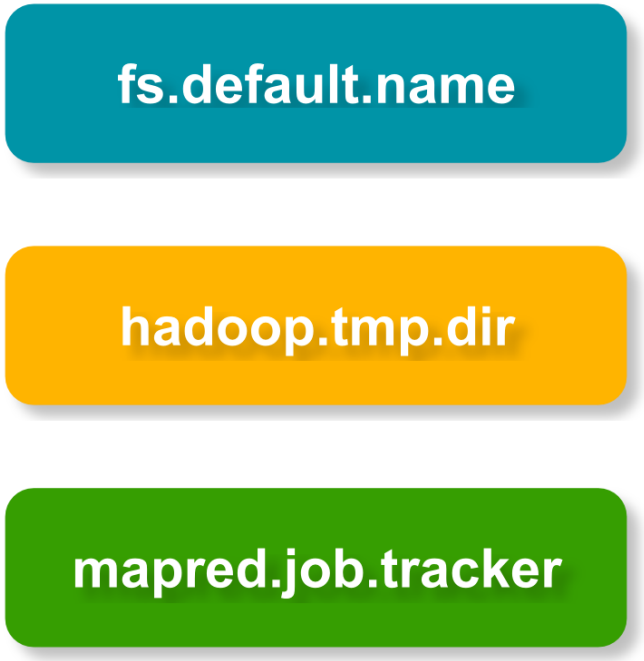
hadoop.env.sh - Sample

```
# Set Hadoop-specific environment variables here.
# The only required environment variable is JAVA_HOME. All others are
# optional. When running a distributed configuration it is best to
# set JAVA_HOME in this file, so that it is correctly defined on
# remote nodes.

# The java implementation to use. Required.
export JAVA_HOME=/usr/lib/jvm/java-6-sun-1.6.0.45

# Extra Java runtime options. Empty by default.
export HADOOP_OPTS="-Djava.net.preferIPv4Stack=true ${HADOOP_OPTS}"
…..
…..
# A string representing this instance of hadoop. $USER by default.
export HADOOP_IDENT_STRING=$USER
```

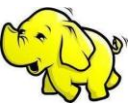


DeZyre

Critical Properties

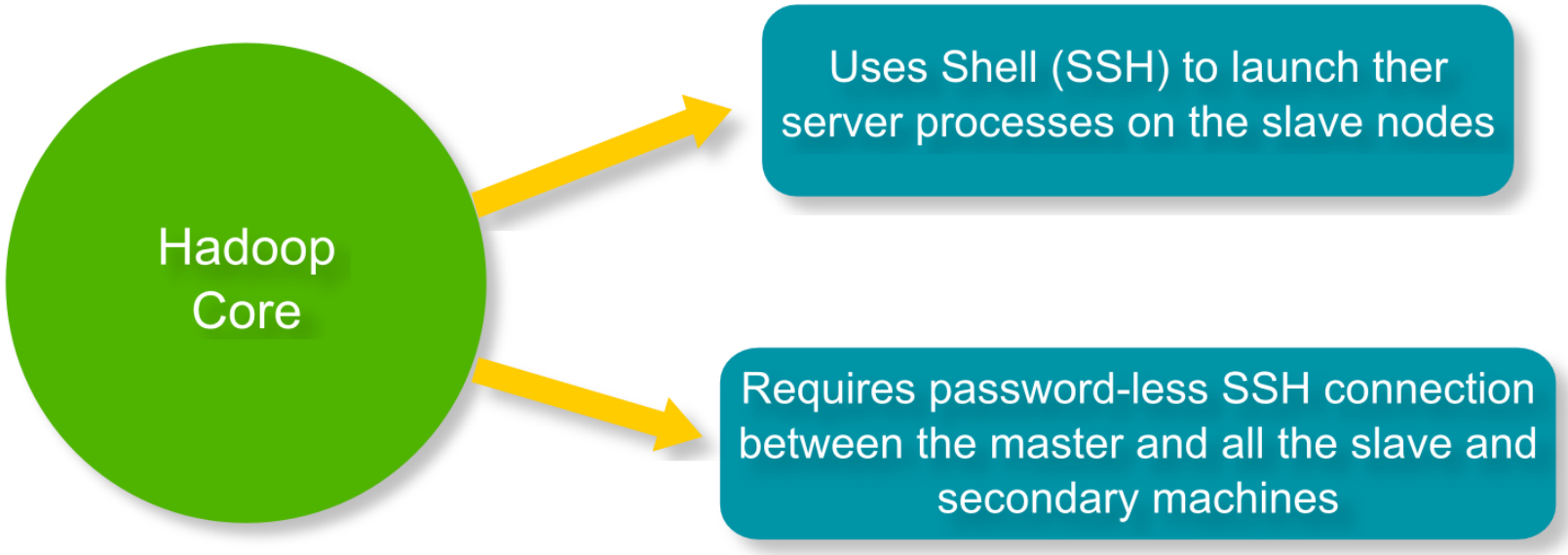


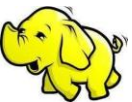
- **fs.default.name:**
  - It points to the default URI for all file system requests in Hadoop.
- **Hadoop.tmp.dir**
  - hadoop.tmp.dir is used as the base for temporary directories locally, and also in HDFS
- **Mapred.job.tracker**
  - The host and port of the MapReduce job tracker where it runs. If "local", then jobs are run in-process as a single map and reduce task



DeZyre

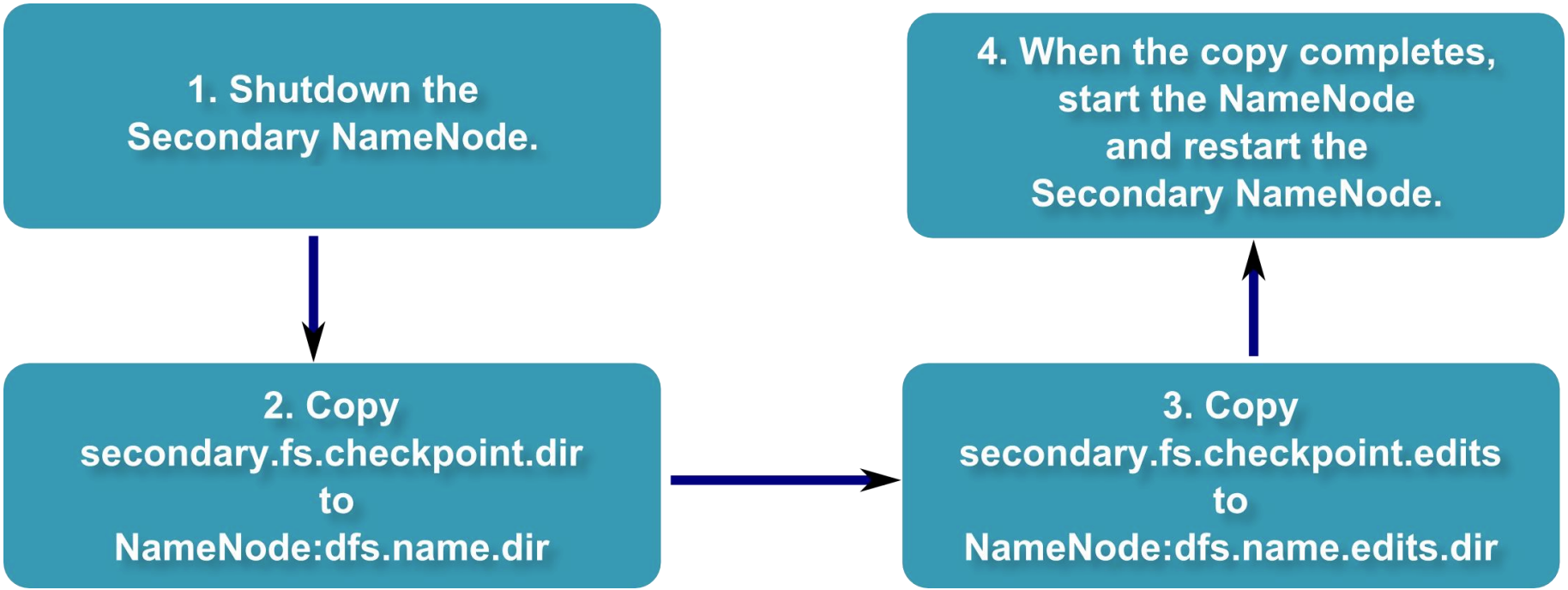
Network Requirements

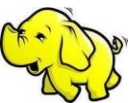




DeZyre

Name Node Recovery

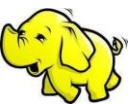




DeZyre

Sample Examples List

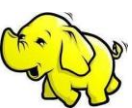
```
hadoopjob@ubuntu:/usr/local/hadoop$ ls
bin             hadoop-ant-1.2.1.jar          ivy          README.txt
build.xml       hadoop-client-1.2.1.jar       ivy.xml      sbin
c++             hadoop-core-1.2.1.jar         lib          share
CHANGES.txt     hadoop-examples-1.2.1.jar     libexec      src
conf            hadoop-minicluster-1.2.1.jar  LICENSE.txt  webapps
contrib         hadoop-test-1.2.1.jar         logs
docs            hadoop-tools-1.2.1.jar        NOTICE.txt
```



DeZyre

Running the Example

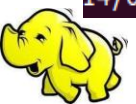
```
hadoopjob@ubuntu:/usr/local/hadoop$ hadoop jar hadoop-examples-1.2.1.jar teragen 100000 /user/teragen/teragen-inputTest
Warning: $HADOOP_HOME is deprecated.

Generating 100000 using 2 maps with step of 50000
14/02/14 14:31:35 INFO mapred.JobClient: Running job: job_201402141140_0001
14/02/14 14:31:36 INFO mapred.JobClient:  map 0% reduce 0%
```



DeZyre

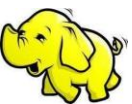
Running the Example

```
14/02/14 14:32:04 INFO mapred.JobClient:  map 100% reduce 0%
14/02/14 14:32:07 INFO mapred.JobClient: Job complete: job_201402141140_0001
14/02/14 14:32:07 INFO mapred.JobClient: Counters: 19
14/02/14 14:32:07 INFO mapred.JobClient:   Job Counters
14/02/14 14:32:07 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=43652
14/02/14 14:32:07 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0
14/02/14 14:32:07 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
14/02/14 14:32:07 INFO mapred.JobClient:     Launched map tasks=2
14/02/14 14:32:07 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=0
14/02/14 14:32:07 INFO mapred.JobClient:   File Input Format Counters
14/02/14 14:32:07 INFO mapred.JobClient:     Bytes Read=0
14/02/14 14:32:07 INFO mapred.JobClient:   File Output Format Counters
14/02/14 14:32:07 INFO mapred.JobClient:     Bytes Written=10000000
14/02/14 14:32:07 INFO mapred.JobClient:   FileSystemCounters
14/02/14 14:32:07 INFO mapred.JobClient:     HDFS_BYTES_READ=164
14/02/14 14:32:07 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=107108
14/02/14 14:32:07 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=10000000
14/02/14 14:32:07 INFO mapred.JobClient:   Map-Reduce Framework
14/02/14 14:32:07 INFO mapred.JobClient:     Map input records=100000
14/02/14 14:32:07 INFO mapred.JobClient:     Physical memory (bytes) snapshot=149934080
14/02/14 14:32:07 INFO mapred.JobClient:     Spilled Records=0
14/02/14 14:32:07 INFO mapred.JobClient:     CPU time spent (ms)=2370
14/02/14 14:32:07 INFO mapred.JobClient:     Total committed heap usage (bytes)=63307776
14/02/14 14:32:07 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=1946124288
14/02/14 14:32:07 INFO mapred.JobClient:     Map input bytes=100000
14/02/14 14:32:07 INFO mapred.JobClient:     Map output records=100000
14/02/14 14:32:07 INFO mapred.JobClient:     SPLIT_RAW_BYTES=164
```



Running the Example

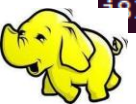
```
hadoopjob@ubuntu:/usr/local/hadoop$ hadoop fs -ls /user/teragen/teragen-inputTest
Warning: $HADOOP_HOME is deprecated.

Found 4 items
-rw-r--r--   1 stratapps supergroup          0 2014-02-14 14:32 /user/teragen/teragen-inputTest/_SUCCESS
drwxr-xr-x   - stratapps supergroup          0 2014-02-14 14:31 /user/teragen/teragen-inputTest/_logs
-rw-r--r--   1 stratapps supergroup    5000000 2014-02-14 14:31 /user/teragen/teragen-inputTest/part-00000
-rw-r--r--   1 stratapps supergroup    5000000 2014-02-14 14:31 /user/teragen/teragen-inputTest/part-00001
hadoopjob@ubuntu:/usr/local/hadoop$ hadoop fs -cat /user/teragen/teragen-inputTest/part-00000
```



DeZyre

Running the Example

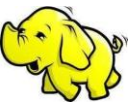
```
I#wql^@Woe      49972AAAAAAAAAABBBBBBBBBBCCCCCCCCCCDDDDDDDDDDEEEEEEEEEEFFFFFFFFFFGGGGGGGGGGHHHHHHHH
wyGUMkZHG^      49973IIIIIIIIIIJJJJJJJJJJKKKKKKKKKKLLLLLLLLLLMMMMMMMMMMNNNNNNNNNNOOOOOOOOOOPPPPPPPP
=&z_SmO,bL      49974QQQQQQQQQQRRRRRRRRRRSSSSSSSSSSTTTTTTTTTTUUUUUUUUUUVVVVVVVVVVWWWWWWWWWWXXXXXXXX
HqY:M y%A=      49975YYYYYYYYYYZZZZZZZZZZAAAAAAAAAABBBBBBBBBBCCCCCCCCCCDDDDDDDDDDEEEEEEEEEEFFFFFFFF
DNm&R>f$be      49976GGGGGGGGGGHHHHHHHHHHIIIIIIIIIIJJJJJJJJJJKKKKKKKKKKLLLLLLLLLLMMMMMMMMMMNNNNNNNN
(4L|p/"sBM      49977OOOOOOOOOOPPPPPPPPPPQQQQQQQQQQRRRRRRRRRRSSSSSSSSSSTTTTTTTTTTUUUUUUUUUUVVVVVVVV
<AIQFo]_k;      49978WWWWWWWWWWXXXXXXXXXXYYYYYYYYYYZZZZZZZZZZAAAAAAAAAABBBBBBBBBBCCCCCCCCCCDDDDDDDD
=;WN,D$xmp      49979EEEEEEEEEEFFFFFFFFFFGGGGGGGGGGHHHHHHHHHHIIIIIIIIIIJJJJJJJJJJKKKKKKKKKKLLLLLLLL
KD$-4a.-]g      49980MMMMMMMMMMNNNNNNNNNNOOOOOOOOOOPPPPPPPPPPQQQQQQQQQQRRRRRRRRRRSSSSSSSSSSTTTTTTTT
_A=w*qK\`4      49981UUUUUUUUUUVVVVVVVVVVWWWWWWWWWWXXXXXXXXXXYYYYYYYYYYZZZZZZZZZZAAAAAAAAAABBBBBBBB
08a)"V\<p`      49982CCCCCCCCCCDDDDDDDDDDEEEEEEEEEEFFFFFFFFFFGGGGGGGGGGHHHHHHHHHHIIIIIIIIIIJJJJJJJJ
xQAr$"TP&Q      49983KKKKKKKKKKLLLLLLLLLLMMMMMMMMMMNNNNNNNNNNOOOOOOOOOOPPPPPPPPPPQQQQQQQQQQRRRRRRRR
pK0~WIn5k:      49984SSSSSSSSSSTTTTTTTTTTUUUUUUUUUUVVVVVVVVVVWWWWWWWWWWXXXXXXXXXXYYYYYYYYYYZZZZZZZZ
55E(X $+aq      49985AAAAAAAAAABBBBBBBBBBCCCCCCCCCCDDDDDDDDDDEEEEEEEEEEFFFFFFFFFFGGGGGGGGGGHHHHHHHH
J[OUDpFN]U      49986IIIIIIIIIIJJJJJJJJJJKKKKKKKKKKLLLLLLLLLLMMMMMMMMMMNNNNNNNNNNOOOOOOOOOOPPPPPPPP
hSf3v9T1v_      49987QQQQQQQQQQRRRRRRRRRRSSSSSSSSSSTTTTTTTTTTUUUUUUUUUUVVVVVVVVVVWWWWWWWWWWXXXXXXXX
q|0HybIHYM      49988YYYYYYYYYYZZZZZZZZZZAAAAAAAAAABBBBBBBBBBCCCCCCCCCCDDDDDDDDDDEEEEEEEEEEFFFFFFFF
~eA6?(=>>R      49989GGGGGGGGGGHHHHHHHHHHIIIIIIIIIIJJJJJJJJJJKKKKKKKKKKLLLLLLLLLLMMMMMMMMMMNNNNNNNN
-ZsQxF9FCn      49990OOOOOOOOOOPPPPPPPPPPQQQQQQQQQQRRRRRRRRRRSSSSSSSSSSTTTTTTTTTTUUUUUUUUUUVVVVVVVV
.J@O^WDE/d      49991WWWWWWWWWWXXXXXXXXXXYYYYYYYYYYZZZZZZZZZZAAAAAAAAAABBBBBBBBBBCCCCCCCCCCDDDDDDDD
Bg`+BA8=k      49992EEEEEEEEEEFFFFFFFFFFGGGGGGGGGGHHHHHHHHHHIIIIIIIIIIJJJJJJJJJJKKKKKKKKKKLLLLLLLL
4{y*(O\%bL      49993MMMMMMMMMMNNNNNNNNNNOOOOOOOOOOPPPPPPPPPPQQQQQQQQQQRRRRRRRRRRSSSSSSSSSSTTTTTTTT
w`KOq5xJ/I      49994UUUUUUUUUUVVVVVVVVVVWWWWWWWWWWXXXXXXXXXXYYYYYYYYYYZZZZZZZZZZAAAAAAAAAABBBBBBBB
K!E*n*3\!l      49995CCCCCCCCCCDDDDDDDDDDEEEEEEEEEEFFFFFFFFFFGGGGGGGGGGHHHHHHHHHHIIIIIIIIIIJJJJJJJJ
Q@6$O\\?my      49996KKKKKKKKKKLLLLLLLLLLMMMMMMMMMMNNNNNNNNNNOOOOOOOOOOPPPPPPPPPPQQQQQQQQQQRRRRRRRR
f'+H CzV)H      49997SSSSSSSSSSTTTTTTTTTTUUUUUUUUUUVVVVVVVVVVWWWWWWWWWWXXXXXXXXXXYYYYYYYYYYZZZZZZZZ
;olHJ9Z\i<      49998AAAAAAAAAABBBBBBBBBBCCCCCCCCCCDDDDDDDDDDEEEEEEEEEEFFFFFFFFFFGGGGGGGGGGHHHHHHHH
i8IQ$H^yv,      49999IIIIIIIIIIJJJJJJJJJJKKKKKKKKKKLLLLLLLLLLMMMMMMMMMMNNNNNNNNNNOOOOOOOOOOPPPPPPPP
```



DeZyre

MapReduce

```
13/08/03 00:58:40 WARN mapred.JobClient: Use GenericOptionsParser for parsing th
e arguments. Applications should implement Tool for the same.
13/08/03 00:58:40 INFO mapred.FileInputFormat: Total input paths to process : 1
13/08/03 00:58:40 INFO mapred.JobClient: Running job: job_201308022025_0003
13/08/03 00:58:41 INFO mapred.JobClient:  map 0% reduce 0%
13/08/03 00:58:44 INFO mapred.JobClient:  map 100% reduce 0%
13/08/03 00:58:51 INFO mapred.JobClient:  map 100% reduce 11%
13/08/03 00:58:52 INFO mapred.JobClient:  map 100% reduce 66%
13/08/03 00:58:59 INFO mapred.JobClient:  map 100% reduce 100%
13/08/03 00:58:59 INFO mapred.JobClient: Job complete: job_201308022025_0003
13/08/03 00:58:59 INFO mapred.JobClient: Counters: 23
13/08/03 00:58:59 INFO mapred.JobClient:   Job Counters
13/08/03 00:58:59 INFO mapred.JobClient:     Launched reduce tasks=3
13/08/03 00:58:59 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=4053
13/08/03 00:58:59 INFO mapred.JobClient:     Total time spent by all reduces wai
ting after reserving slots (ms)=0
13/08/03 00:58:59 INFO mapred.JobClient:     Total time spent by all maps waitin
g after reserving slots (ms)=0
13/08/03 00:58:59 INFO mapred.JobClient:     Launched map tasks=2
13/08/03 00:58:59 INFO mapred.JobClient:     Data-local map tasks=2
13/08/03 00:58:59 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=23684
13/08/03 00:58:59 INFO mapred.JobClient:   FileSystemCounters
13/08/03 00:58:59 INFO mapred.JobClient:     FILE_BYTES_READ=81770
13/08/03 00:58:59 INFO mapred.JobClient:     HDFS_BYTES_READ=136111
13/08/03 00:58:59 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=429317
13/08/03 00:58:59 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=61194
13/08/03 00:58:59 INFO mapred.JobClient:   Map-Reduce Framework
13/08/03 00:58:59 INFO mapred.JobClient:     Reduce input groups=3586
13/08/03 00:58:59 INFO mapred.JobClient:     Combine output records=4027
13/08/03 00:58:59 INFO mapred.JobClient:     Map input records=2403
13/08/03 00:58:59 INFO mapred.JobClient:     Reduce shuffle bytes=81788
13/08/03 00:58:59 INFO mapred.JobClient:     Reduce output records=3586
13/08/03 00:58:59 INFO mapred.JobClient:     Spilled Records=8054
13/08/03 00:58:59 INFO mapred.JobClient:     Map output bytes=151013
13/08/03 00:58:59 INFO mapred.JobClient:     Map input bytes=132663
13/08/03 00:58:59 INFO mapred.JobClient:     Combine input records=11037
13/08/03 00:58:59 INFO mapred.JobClient:     Map output records=11037
13/08/03 00:58:59 INFO mapred.JobClient:     SPLIT_RAW_BYTES=146
13/08/03 00:58:59 INFO mapred.JobClient:     Reduce input records=4027
```



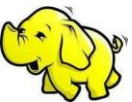
DeZyre

MapReduce

```
13/08/03 00:58:40 WARN mapred.JobClient: Use GenericOptionsParser for parsing th
e arguments. Applications should implement Tool for the same.
13/08/03 00:58:40 INFO mapred.FileInputFormat: Total input paths to process : 1
13/08/03 00:58:40 INFO mapred.JobClient: Running job: job_201308022025_0003
13/08/03 00:58:41 INFO mapred.JobClient:  map 0% reduce 0%
13/08/03 00:58:44 INFO mapred.JobClient:  map 100% reduce 0%
13/08/03 00:58:51 INFO mapred.JobClient:  map 100% reduce 11%
13/08/03 00:58:52 INFO mapred.JobClient:  map 100% reduce 66%
13/08/03 00:58:59 INFO mapred.JobClient:  map 100% reduce 100%
13/08/03 00:58:59 INFO mapred.JobClient: Job complete: job_201308022025_0003
13/08/03 00:58:59 INFO mapred.JobClient: Counters: 23
13/08/03 00:58:59 INFO mapred.JobClient:   Job Counters
13/08/03 00:58:59 INFO mapred.JobClient:     Launched reduce tasks=3
13/08/03 00:58:59 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=4053
13/08/03 00:58:59 INFO mapred.JobClient:     Total time spent by all reduces wai
ting after reserving slots (ms)=0
13/08/03 00:58:59 INFO mapred.JobClient:     Total time spent by all maps waitin
g after reserving slots (ms)=0
13/08/03 00:58:59 INFO mapred.JobClient:     Launched map tasks=2
13/08/03 00:58:59 INFO mapred.JobClient:     Data-local map tasks=2
13/08/03 00:58:59 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=23684
13/08/03 00:58:59 INFO mapred.JobClient:   FileSystemCounters
13/08/03 00:58:59 INFO mapred.JobClient:     FILE_BYTES_READ=81770
13/08/03 00:58:59 INFO mapred.JobClient:     HDFS_BYTES_READ=136111
13/08/03 00:58:59 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=429317
13/08/03 00:58:59 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=61194
13/08/03 00:58:59 INFO mapred.JobClient:   Map-Reduce Framework
13/08/03 00:58:59 INFO mapred.JobClient:     Reduce input groups=3586
13/08/03 00:58:59 INFO mapred.JobClient:     Combine output records=4027
13/08/03 00:58:59 INFO mapred.JobClient:     Map input records=2403
13/08/03 00:58:59 INFO mapred.JobClient:     Reduce shuffle bytes=81788
13/08/03 00:58:59 INFO mapred.JobClient:     Reduce output records=3586
13/08/03 00:58:59 INFO mapred.JobClient:     Spilled Records=8054
13/08/03 00:58:59 INFO mapred.JobClient:     Map output bytes=151013
13/08/03 00:58:59 INFO mapred.JobClient:     Map input bytes=132663
13/08/03 00:58:59 INFO mapred.JobClient:     Combine input records=11037
13/08/03 00:58:59 INFO mapred.JobClient:     Map output records=11037
13/08/03 00:58:59 INFO mapred.JobClient:     SPLIT_RAW_BYTES=146
13/08/03 00:58:59 INFO mapred.JobClient:     Reduce input records=4027
```

1

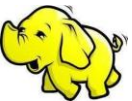
2



DeZyre

MapReduce

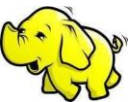
```
13/08/03 00:58:40 WARN mapred.JobClient: Use GenericOptionsParser for parsing th
e arguments. Applications should implement Tool for the same.
13/08/03 00:58:40 INFO mapred.FileInputFormat: Total input paths to process : 1
13/08/03 00:58:40 INFO mapred.JobClient: Running job: job_201308022025_0003
13/08/03 00:58:41 INFO mapred.JobClient:  map 0% reduce 0%
13/08/03 00:58:44 INFO mapred.JobClient:  map 100% reduce 0%
13/08/03 00:58:51 INFO mapred.JobClient:  map 100% reduce 11%
13/08/03 00:58:52 INFO mapred.JobClient:  map 100% reduce 66%
13/08/03 00:58:59 INFO mapred.JobClient:  map 100% reduce 100%
13/08/03 00:58:59 INFO mapred.JobClient: Job complete: job_201308022025_0003
13/08/03 00:58:59 INFO mapred.JobClient: Counters: 23
13/08/03 00:58:59 INFO mapred.JobClient:   Job Counters
13/08/03 00:58:59 INFO mapred.JobClient:     Launched reduce tasks=3
13/08/03 00:58:59 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=4053
13/08/03 00:58:59 INFO mapred.JobClient:     Total time spent by all reduces wai
ting after reserving slots (ms)=0
13/08/03 00:58:59 INFO mapred.JobClient:     Total time spent by all maps waitin
g after reserving slots (ms)=0
13/08/03 00:58:59 INFO mapred.JobClient:     Launched map tasks=2
13/08/03 00:58:59 INFO mapred.JobClient:     Data-local map tasks=2
13/08/03 00:58:59 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=23684
13/08/03 00:58:59 INFO mapred.JobClient:   FileSystemCounters
13/08/03 00:58:59 INFO mapred.JobClient:     FILE_BYTES_READ=81770
13/08/03 00:58:59 INFO mapred.JobClient:     HDFS_BYTES_READ=136111
13/08/03 00:58:59 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=429317
13/08/03 00:58:59 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=61194
13/08/03 00:58:59 INFO mapred.JobClient:   Map-Reduce Framework
13/08/03 00:58:59 INFO mapred.JobClient:     Reduce input groups=3586
13/08/03 00:58:59 INFO mapred.JobClient:     Combine output records=4027
13/08/03 00:58:59 INFO mapred.JobClient:     Map input records=2403
13/08/03 00:58:59 INFO mapred.JobClient:     Reduce shuffle bytes=81788
13/08/03 00:58:59 INFO mapred.JobClient:     Reduce output records=3586
13/08/03 00:58:59 INFO mapred.JobClient:     Spilled Records=8054
13/08/03 00:58:59 INFO mapred.JobClient:     Map output bytes=151013
13/08/03 00:58:59 INFO mapred.JobClient:     Map input bytes=132663
13/08/03 00:58:59 INFO mapred.JobClient:     Combine input records=11037
13/08/03 00:58:59 INFO mapred.JobClient:     Map output records=11037
13/08/03 00:58:59 INFO mapred.JobClient:     SPLIT_RAW_BYTES=146
13/08/03 00:58:59 INFO mapred.JobClient:     Reduce input records=4027
```



DeZyre

MapReduce

```
13/08/03 00:58:40 WARN mapred.JobClient: Use GenericOptionsParser for parsing th
e arguments. Applications should implement Tool for the same.
13/08/03 00:58:40 INFO mapred.FileInputFormat: Total input paths to process : 1
13/08/03 00:58:40 INFO mapred.JobClient: Running job: job_201308022025_0003
13/08/03 00:58:41 INFO mapred.JobClient:  map 0% reduce 0%
13/08/03 00:58:44 INFO mapred.JobClient:  map 100% reduce 0%
13/08/03 00:58:51 INFO mapred.JobClient:  map 100% reduce 11%
13/08/03 00:58:52 INFO mapred.JobClient:  map 100% reduce 66%
13/08/03 00:58:59 INFO mapred.JobClient:  map 100% reduce 100%
13/08/03 00:58:59 INFO mapred.JobClient: Job complete: job_201308022025_0003
13/08/03 00:58:59 INFO mapred.JobClient: Counters: 23
13/08/03 00:58:59 INFO mapred.JobClient:   Job Counters
13/08/03 00:58:59 INFO mapred.JobClient:     Launched reduce tasks=3
13/08/03 00:58:59 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=4053
13/08/03 00:58:59 INFO mapred.JobClient:     Total time spent by all reduces wai
ting after reserving slots (ms)=0
13/08/03 00:58:59 INFO mapred.JobClient:     Total time spent by all maps waitin
g after reserving slots (ms)=0
13/08/03 00:58:59 INFO mapred.JobClient:     Launched map tasks=2
13/08/03 00:58:59 INFO mapred.JobClient:     Data-local map tasks=2
13/08/03 00:58:59 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=23684
13/08/03 00:58:59 INFO mapred.JobClient:   FileSystemCounters
13/08/03 00:58:59 INFO mapred.JobClient:     FILE_BYTES_READ=81770
13/08/03 00:58:59 INFO mapred.JobClient:     HDFS_BYTES_READ=136111
13/08/03 00:58:59 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=429317
13/08/03 00:58:59 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=61194
13/08/03 00:58:59 INFO mapred.JobClient:   Map-Reduce Framework
13/08/03 00:58:59 INFO mapred.JobClient:     Reduce input groups=3586
13/08/03 00:58:59 INFO mapred.JobClient:     Combine output records=4027
13/08/03 00:58:59 INFO mapred.JobClient:     Map input records=2403
13/08/03 00:58:59 INFO mapred.JobClient:     Reduce shuffle bytes=81788
13/08/03 00:58:59 INFO mapred.JobClient:     Reduce output records=3586
13/08/03 00:58:59 INFO mapred.JobClient:     Spilled Records=8054
13/08/03 00:58:59 INFO mapred.JobClient:     Map output bytes=151013
13/08/03 00:58:59 INFO mapred.JobClient:     Map input bytes=132663
13/08/03 00:58:59 INFO mapred.JobClient:     Combine input records=11037
13/08/03 00:58:59 INFO mapred.JobClient:     Map output records=11037
13/08/03 00:58:59 INFO mapred.JobClient:     SPLIT_RAW_BYTES=146
13/08/03 00:58:59 INFO mapred.JobClient:     Reduce input records=4027
```



DeZyre

Further reading and reference...

Companies powered by Hadoop

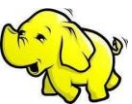
https://wiki.apache.org/hadoop/PoweredBy

MapReduce Paper by Google

http://research.google.com/archive/mapreduce.html

MapReduce Tutorial

https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html



DeZyre

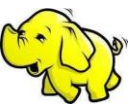
Further reading and reference...

Single Node Cluster Installation on Linux(Ubuntu)

http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/

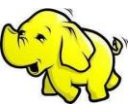
Multi-Node Cluster Installation on Linux(Ubuntu)

http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-multi-node-cluster/



DeZyre

Thank You



DeZyre