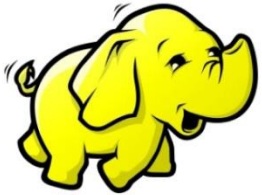
edureka!

Big Data & Hadoop



Course Topics

edureka!

- ✓ Module 1
  - ✓ Understanding Big Data
  - ✓ Hadoop Architecture
- ✓ Module 2
  - ✓ **Hadoop Cluster Configuration**
  - ✓ **Data loading Techniques**
  - ✓ **Hadoop Project Environment**
- ✓ Module 3
  - ✓ Hadoop MapReduce framework
  - ✓ Programming in Map Reduce
- ✓ Module 4
  - ✓ Advance MapReduce
  - ✓ MRUnit testing framework
- ✓ Module 5
  - ✓ Analytics using Pig
  - ✓ Understanding Pig Latin
- ✓ Module 6
  - ✓ Analytics using Hive
  - ✓ Understanding HIVE QL
- ✓ Module 7
  - ✓ Advance Hive
  - ✓ NoSQL Databases and HBASE
- ✓ Module 8
  - ✓ Advance HBASE
  - ✓ Zookeeper Service
- ✓ Module 9
  - ✓ Hadoop 2.0 – New Features
  - ✓ Programming in MRv2
- ✓ Module 10
  - ✓ Apache Oozie
  - ✓ Real world Datasets and Analysis
  - ✓ Project Discussion

Topics for Today

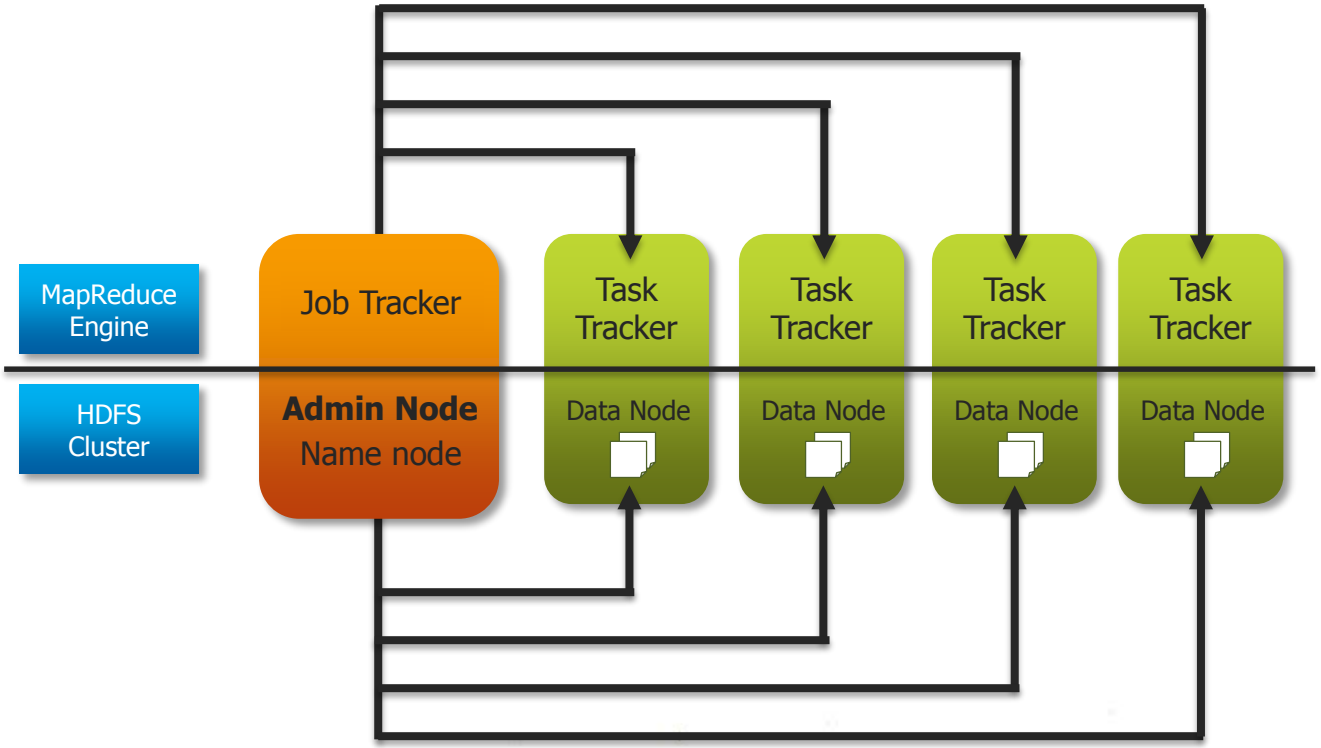
edureka!

- ✓ Revision
- ✓ A Typical Hadoop Cluster
- ✓ Hadoop Cluster Configuration
- ✓ Hadoop Configuration Files
- ✓ Hadoop Cluster Modes
- ✓ Sample example list in Hadoop
- ✓ Running Teragen Example
- ✓ Dump of MR job
- ✓ Data Loading Techniques
  - ✓ Using Hadoop Copy Commands
  - ✓ FLUME
  - ✓ SQOOP
- ✓ Course Project Problem Statement

Lets's Revise

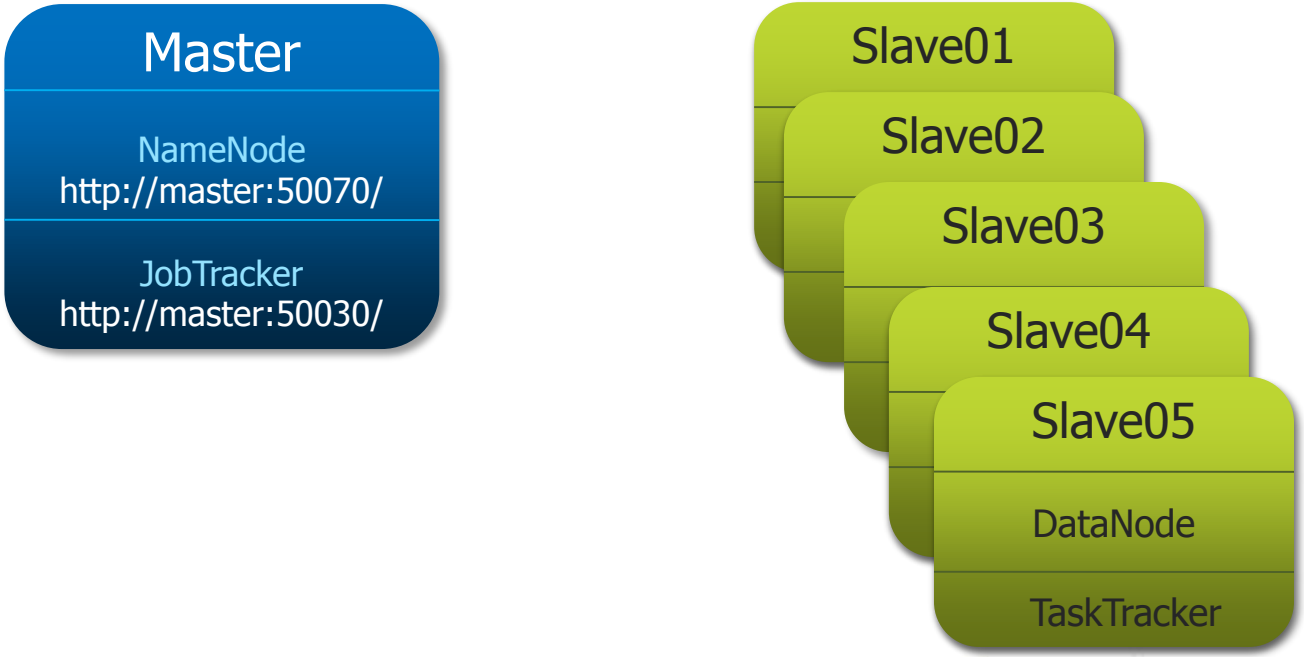
edureka!

1. Hadoop Core Components
2. HDFS Architecture
3. What is HDFS?
4. Hadoop Vs. traditional systems
5. Namenode and Secondary Namenode



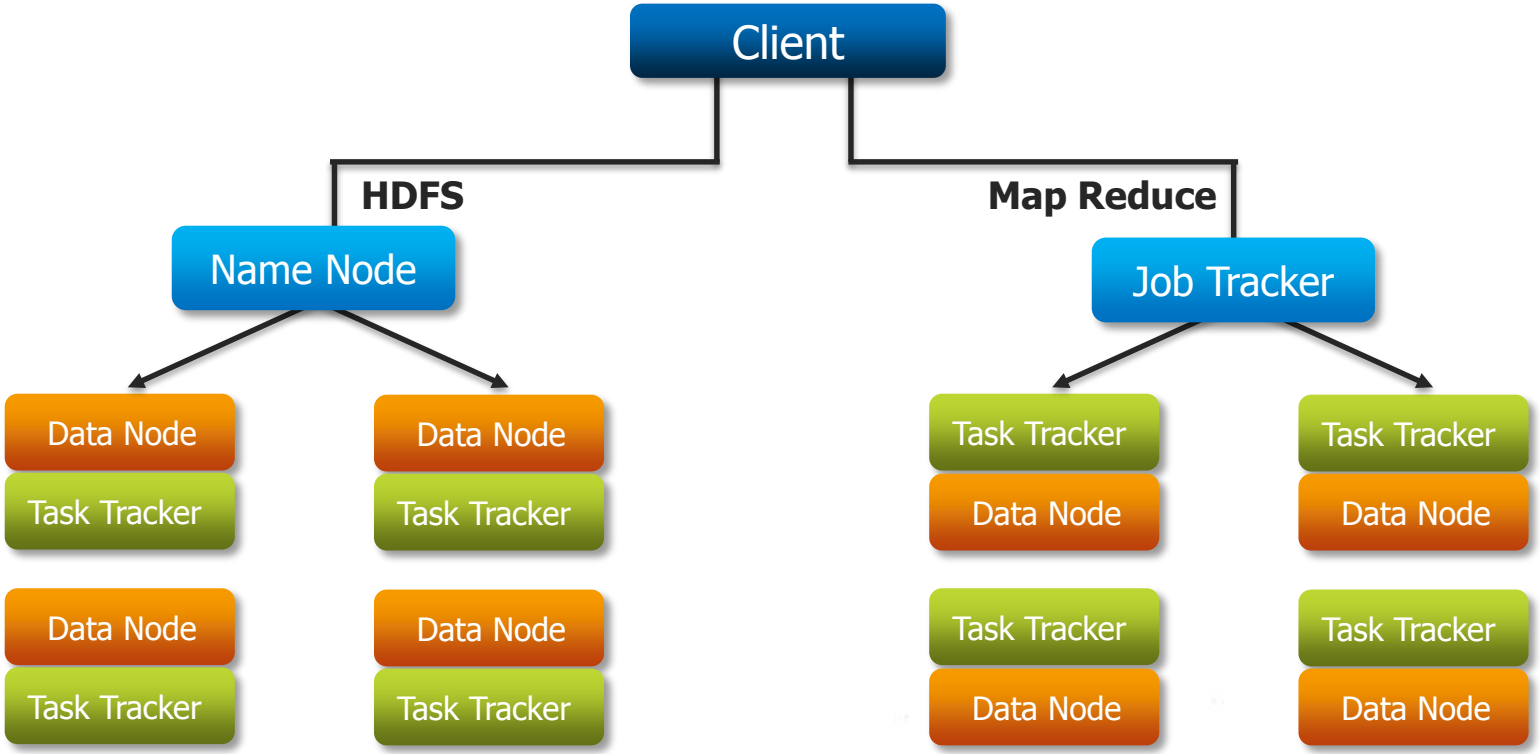
Hadoop Cluster Architecture

edureka!



Hadoop Cluster Architecture (Contd.)

edureka!



Pre-Class Questions

Annie's Question

edureka!

The default replication factor is:

- 2
- 4
- 5
- 3

Annie's Answer

edureka!

The answer is 3. It means if you move a file to hdfs then by default 3 copies of the file will be stored on different datanodes.

Annie's Question

edureka!

Every Slave node has two daemons running on them that is DataNode & TaskTracker in a MultiNode Cluster.

- TRUE
- FALSE

Annie's Answer

edureka!

TRUE. DataNode service for HDFS and TaskTracker for processing (MapReduce) tasks.

Annie's Question

edureka!

A block is replicated in 4 nodes K,L,M, and N. If M, K and N fails. A client can still read the data.

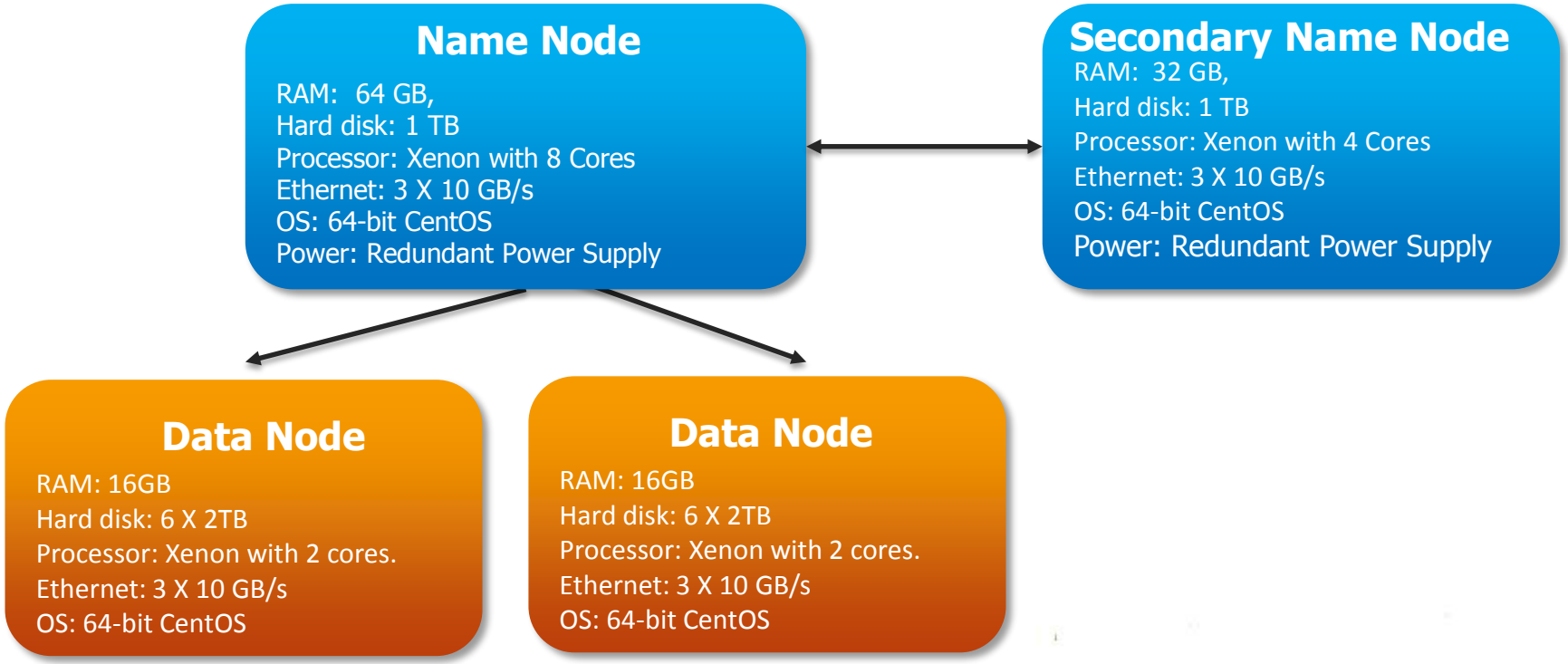
- TRUE
- FALSE

Annie’s Answer

TRUE. As the remaining node ‘L’ will contain the block in question.

Hadoop Cluster: A Typical Use Case

edureka!



Hadoop Cluster: Facebook

edureka!

- *Facebook*
  - *We use Hadoop to store copies of internal log and dimension data sources and use it as a source for reporting/analytics and machine learning.*
  - *Currently we 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.*
    - *We are heavy users of both streaming as well as the Java APIs. We have built a higher level data warehousing framework using these features called Hive (see the http://hadoop.apache.org/hive/). We have also developed a FUSE implementation over HDFS.*

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

Hadoop Cluster Modes

edureka!

Hadoop can run in any of the following three 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

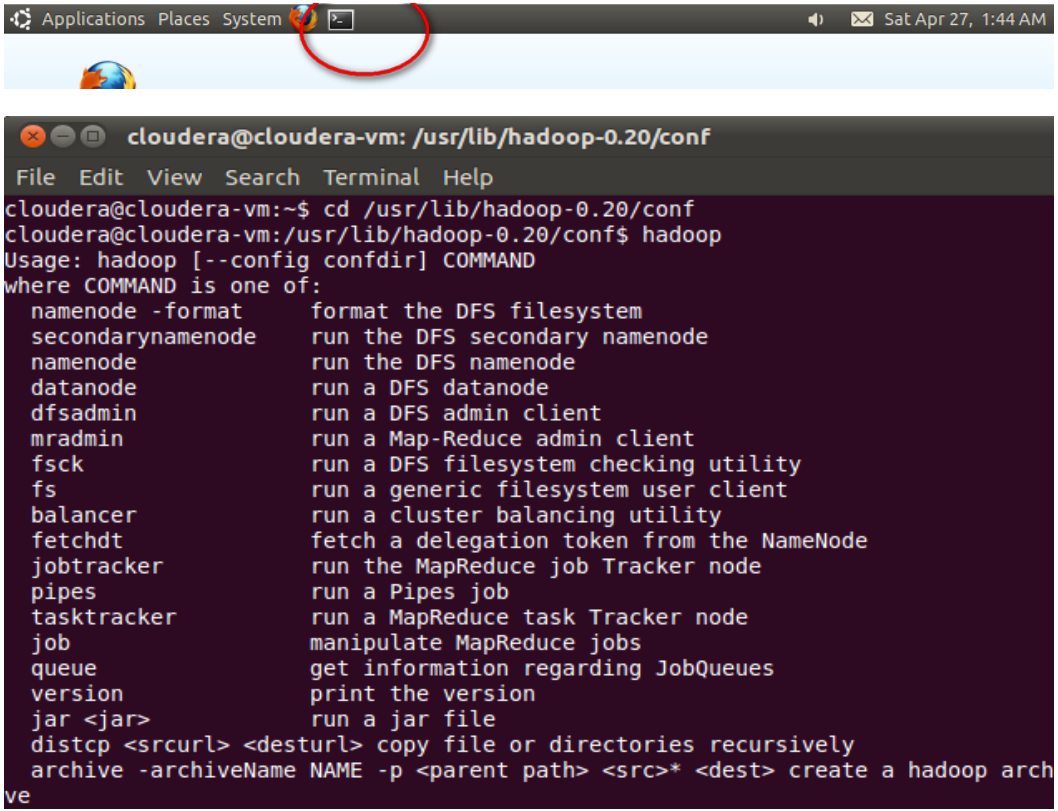
- ✓ Hadoop daemons run on the local machine.

Fully-Distributed Mode

- ✓ Hadoop daemons run on a cluster of machines.

Terminal Commands

edureka!



Terminal Commands

edureka!

Listing of files present on HDFS

```
cloudera@cloudera-vm:/usr/lib/hadoop-0.20/conf$ hadoop dfs -ls /
Found 3 items
drwxrwxrwx   - hue    supergroup          0 2013-04-20 03:01 /tmp
drwxr-xr-x   - hue    supergroup          0 2013-04-20 03:22 /user
drwxr-xr-x   - mapred supergroup          0 2011-04-12 06:20 /var
cloudera@cloudera-vm:/usr/lib/hadoop-0.20/conf$
```

Listing of files present in bin Directory

```
cloudera@cloudera-vm:~$ cd /usr/lib/hadoop-0.20/bin/
cloudera@cloudera-vm:/usr/lib/hadoop-0.20/bin$ ls
fuse_dfs_wrapper.sh  hadoop-daemons.sh  start-all.sh       stop-all.sh
hadoop               rcc                start-balancer.sh  stop-balancer.sh
hadoop-config.sh     README             start-dfs.sh       stop-dfs.sh
hadoop-daemon.sh     slaves.sh          start-mapred.sh    stop-mapred.sh
cloudera@cloudera-vm:/usr/lib/hadoop-0.20/bin$
```

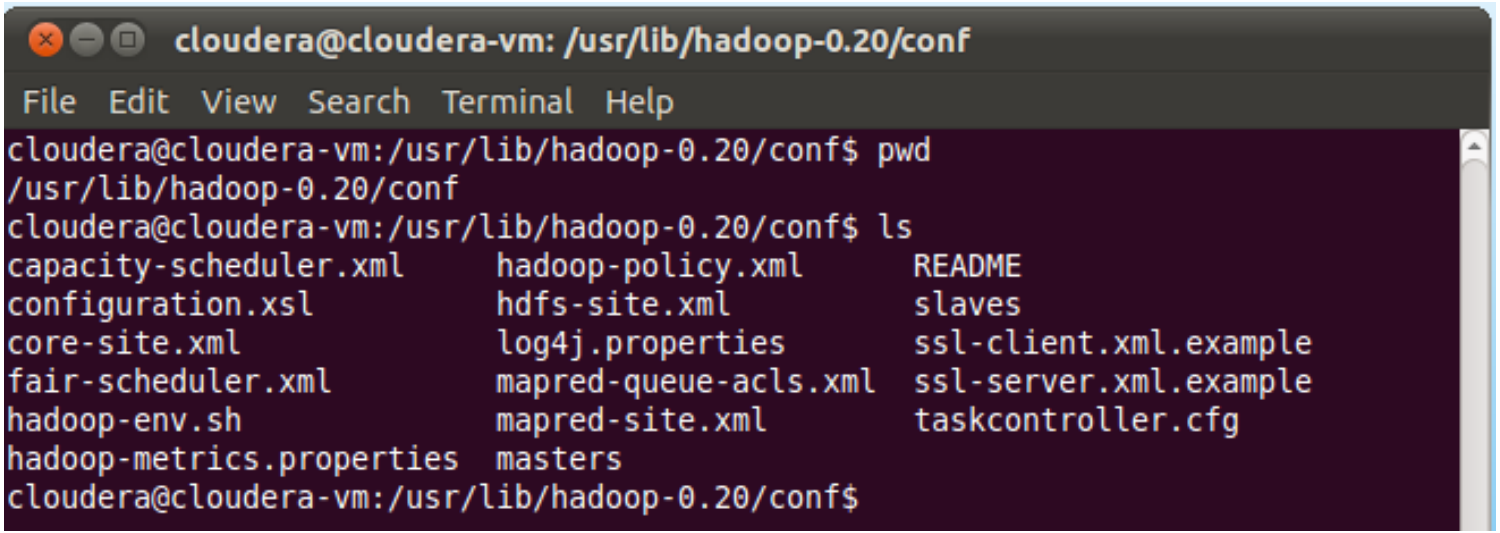
Hadoop Configuration Files

edureka!

Configuration Filenames	Description of Log Files
hadoop-env.sh	Environment variables that are used in the scripts to run Hadoop.
core-site.xml	Configuration settings for Hadoop Core such as I/O settings that are common to HDFS and MapReduce.
hdfs-site.xml	Configuration settings for HDFS daemons, the namenode, the secondary namenode and the data nodes.
mapred-site.xml	Configuration settings for MapReduce daemons : the job-tracker and the task-trackers.
masters	A list of machines (one per line) that each run a secondary namenode.
slaves	A list of machines (one per line) that each run a datanode and a task-tracker.

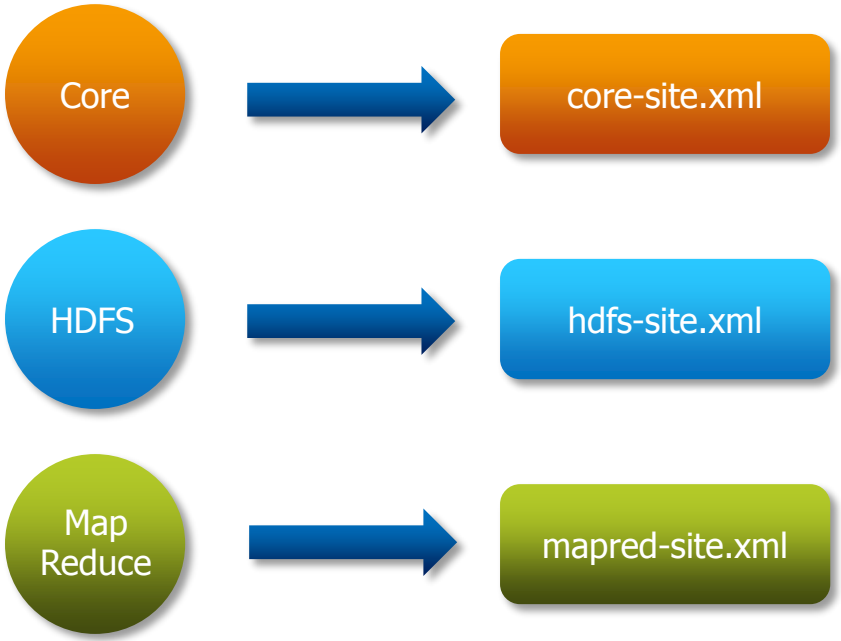
Hadoop 1.x Configuration Files – CDH3

edureka!



Hadoop 1.x: Core Configuration Files

edureka!



core-site.xml and hdfs-site.xml

edureka!

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>**hdfs://localhost:8020/** </value>
</property>	</property>
</configuration>	</configuration>

Defining HDFS Details In hdfs-site.xml

edureka!

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/namesecondary

mapred-site.xml

edureka!

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

Defining mapred-site.xml

edureka!

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 directory relative to fs.default.name where shared files are stored, during a job run.
mapred.tasktracker.map.tasks.maximum	2	The number of map tasks that may be run on a tasktracker at any one time
mapred.tasktracker.reduce.tasks.maximum	2	The number of reduce tasks tat may be run on a tasktracker at any one time.

All Properties

edureka!

1. http://hadoop.apache.org/docs/r1.1.2/core-default.html
2. http://hadoop.apache.org/docs/r1.1.2/mapred-default.html
3. http://hadoop.apache.org/docs/r1.1.2/hdfs-default.html

Slaves and Masters

edureka!

Two files are used by the startup and shutdown commands:

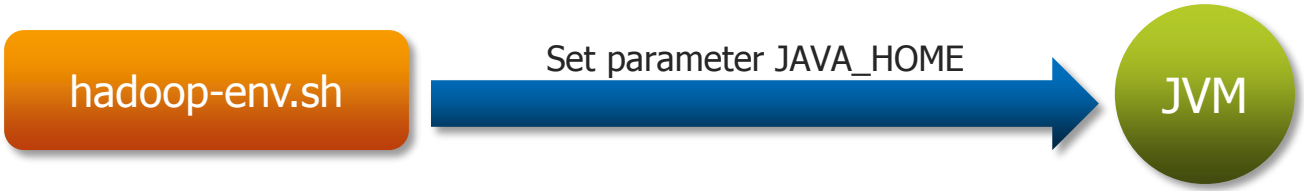
Slaves

- ✓ 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 Runtime 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"
```

Web UI URLs

edureka!

- ✓ NameNode status: http://localhost:50070/dfshealth.jsp
- ✓ JobTracker status: http://localhost:50030/jobtracker.jsp
- ✓ TaskTracker status: http://localhost:50060/tasktracker.jsp
- ✓ DataBlock Scanner Report: http://localhost:50075/blockScannerReport

Annie's Question

edureka!

Which of the following file is used to specify the NameNode's heap size?

- .bashrc
- hadoop-env.sh
- hdfs-site.sh
- core-site.xml

Annie's Answer

hadoop-env.sh. This file specifies environment variables that affect the JDK used by Hadoop Daemon (bin/hadoop).

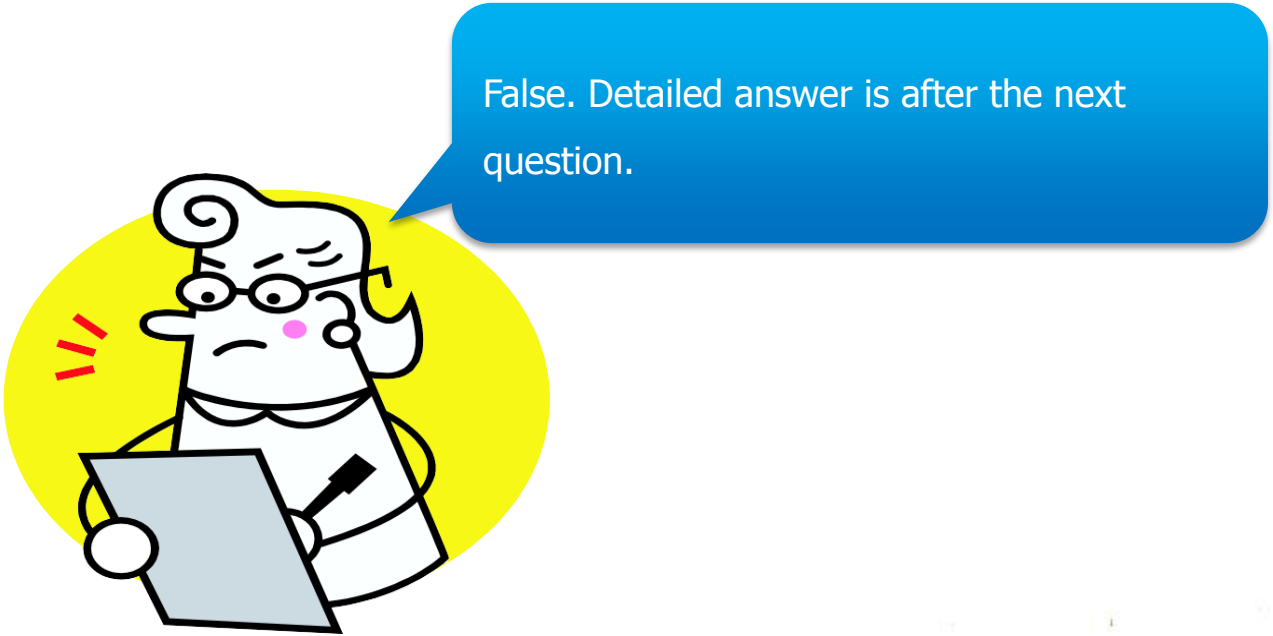
Annie's Question

edureka!

It is necessary to define all the properties in core-site.xml, hdfs-site.xml & mapred-site.xml.

- TRUE
- FALSE

Annie’s Answer



Annie's Question

Standalone Mode uses default configuration?

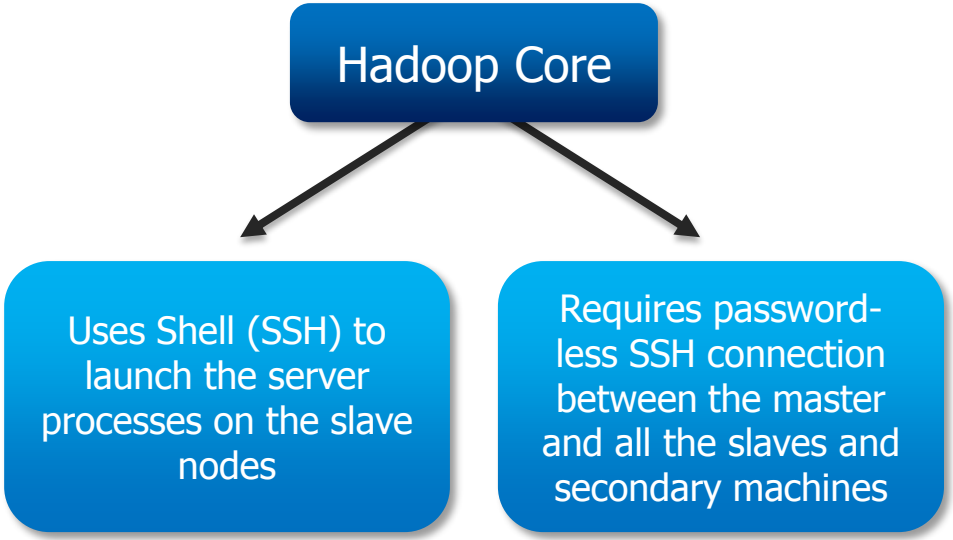
- TRUE
- FALSE

Annie’s Answer

edureka!

True. In Standalone mode Hadoop runs with default configuration (Empty configuration files i.e. no configuration settings in core-site.xml, hdfs-site.xml, and mapred-site.xml). If properties are not defined in the configuration files, hadoop runs with default values for the corresponding properties.

Hadoop Cluster: Password-Less SSH



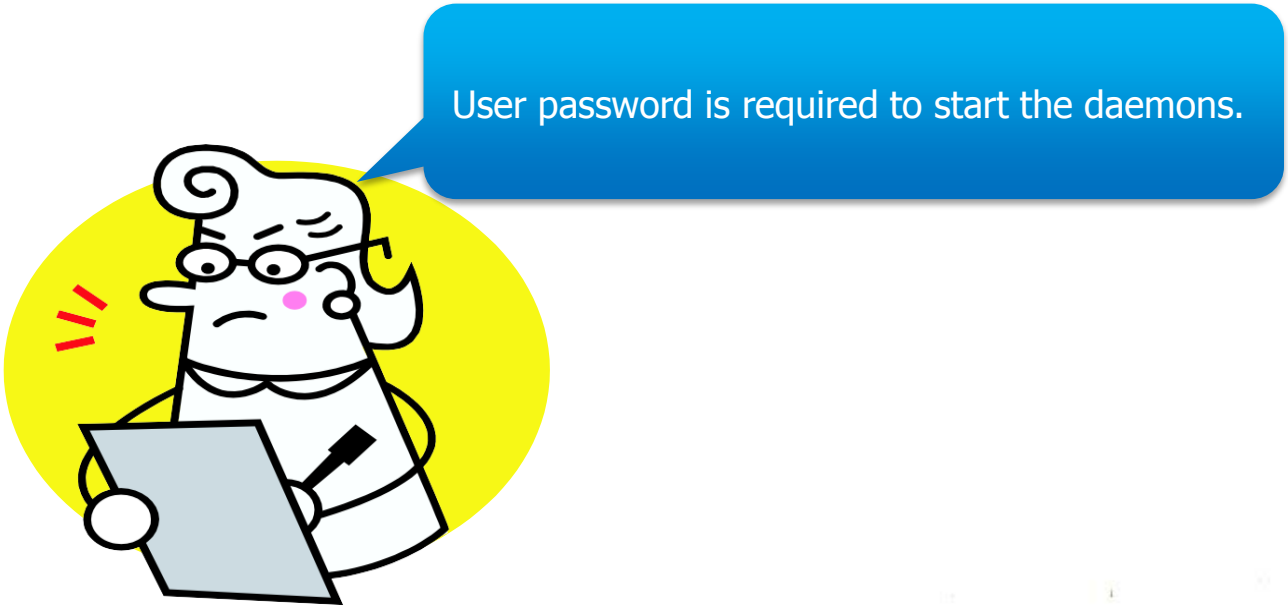
Annie’s Question

edureka!

If password-less SSH Login is not set-up,

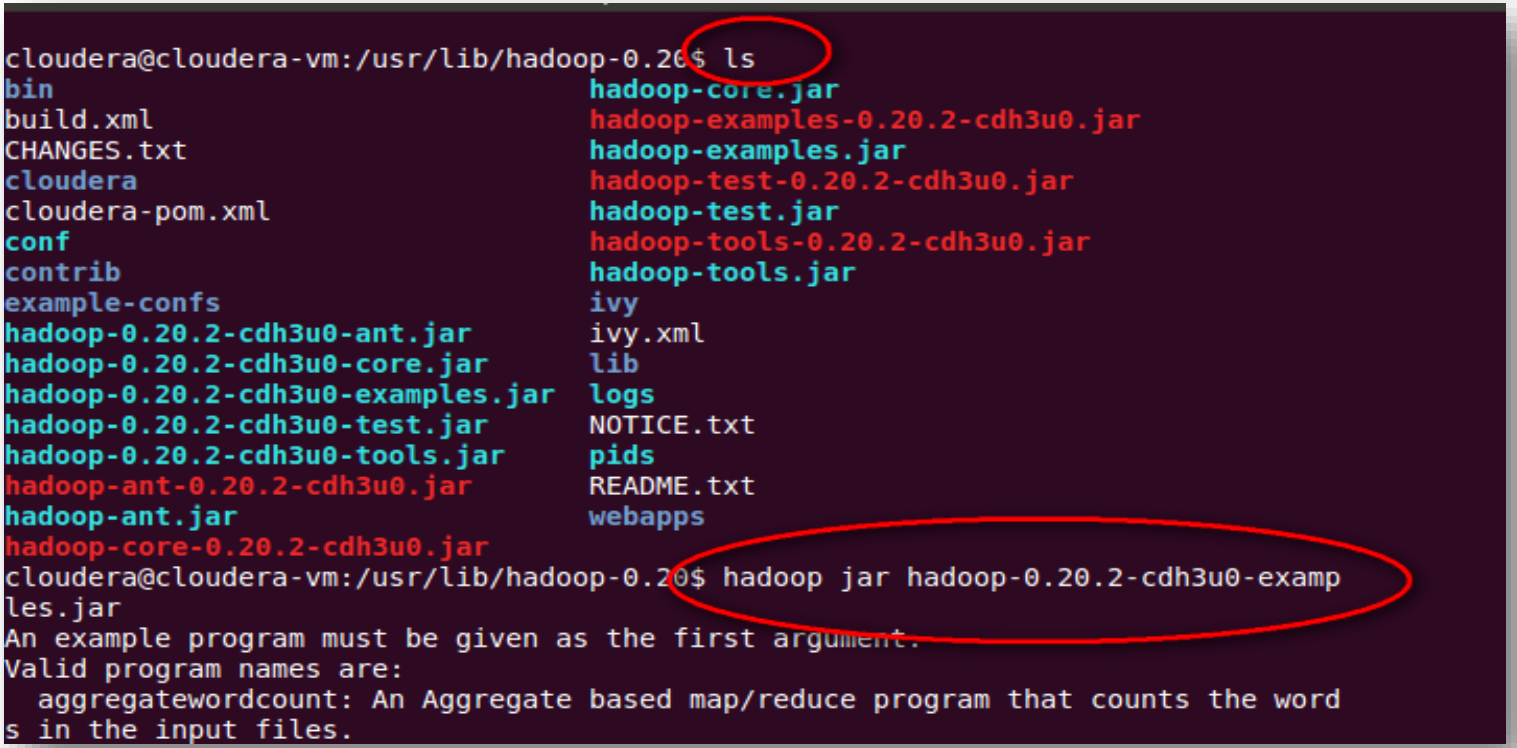
- hadoop deamons won't start
- only NameNode will start in Master
- user's password has to be entered to start every daemon
- None of these

Annie’s Answer



Sample Examples List

edureka!



Running the Teragen Example

edureka!

```
cloudera@cloudera-vm:/usr/lib/hadoop-0.20$ hadoop jar hadoop-0.20.2-cdh3u0-examples.jar teragen 1000000 /user/teragen-inputTes
t
Generating 1000000 using 2 maps with step of 500000
13/04/27 02:20:58 INFO mapred.JobClient: Running job: job_201304270136_0001
13/04/27 02:20:59 INFO mapred.JobClient:  map 0% reduce 0%
```

```
Generating 1000000 using 2 maps with step of 500000
13/04/27 02:20:58 INFO mapred.JobClient: Running job: job_201304270136_0001
13/04/27 02:20:59 INFO mapred.JobClient:  map 0% reduce 0%
13/04/27 02:21:15 INFO mapred.JobClient:  map 49% reduce 0%
13/04/27 02:21:18 INFO mapred.JobClient:  map 81% reduce 0%
13/04/27 02:21:21 INFO mapred.JobClient:  map 100% reduce 0%
13/04/27 02:21:23 INFO mapred.JobClient: Job complete: job_201304270136_0001
13/04/27 02:21:23 INFO mapred.JobClient: Counters: 13
13/04/27 02:21:23 INFO mapred.JobClient:   Job Counters
13/04/27 02:21:23 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=37341
13/04/27 02:21:23 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0
13/04/27 02:21:23 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
13/04/27 02:21:23 INFO mapred.JobClient:     Launched map tasks=2
13/04/27 02:21:23 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=0
13/04/27 02:21:23 INFO mapred.JobClient:   FileSystemCounters
13/04/27 02:21:23 INFO mapred.JobClient:     HDFS_BYTES_READ=167
13/04/27 02:21:23 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=105170
13/04/27 02:21:23 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=100000000
13/04/27 02:21:23 INFO mapred.JobClient:   Map-Reduce Framework
13/04/27 02:21:23 INFO mapred.JobClient:     Map input records=1000000
13/04/27 02:21:23 INFO mapred.JobClient:     Spilled Records=0
13/04/27 02:21:23 INFO mapred.JobClient:     Map input bytes=1000000
13/04/27 02:21:23 INFO mapred.JobClient:     Map output records=1000000
13/04/27 02:21:23 INFO mapred.JobClient:     SPLIT_RAW_BYTES=167
```

Checking the Output

edureka!

```
cloudera@cloudera-vm:/usr/lib/hadoop-0.20/conf$ hadoop dfs -ls /
Found 3 items
drwxrwxrwx   - hue    supergroup          0 2013-04-27 01:59 /tmp
drwxr-xr-x   - hue    supergroup          0 2013-04-27 02:20 /user
drwxr-xr-x   - mapred supergroup          0 2011-04-12 06:20 /var
cloudera@cloudera-vm:/usr/lib/hadoop-0.20/conf$ hadoop dfs -ls /user
Found 5 items
drwxr-xr-x   - cloudera supergroup          0 2011-04-08 16:48 /user/cloudera
drwxr-xr-x   - hue      supergroup          0 2011-04-01 17:25 /user/hive
drwxr-xr-x   - cloudera supergroup          0 2013-04-20 03:19 /user/teragen-input14
drwxr-xr-x   - cloudera supergroup          0 2013-04-20 03:22 /user/teragen-input15
drwxr-xr-x   - cloudera supergroup          0 2013-04-27 02:21 /user/teragen-inputTest
```

Checking the Output

edureka!

```
cloudera@cloudera-vm:/usr/lib/hadoop-0.20$ hadoop fs -get /user/teragen-inputTest/part-00001 .
get: Permission denied
cloudera@cloudera-vm:/usr/lib/hadoop-0.20$ sudo su
[sudo] password for cloudera:
root@cloudera-vm:/usr/lib/hadoop-0.20# hadoop fs -get /user/teragen-inputTest/part-00001 .
root@cloudera-vm:/usr/lib/hadoop-0.20# ls
bin                hadoop-0.20.2-cdh3u0-ant.jar       hadoop-core.jar                    ivy.xml
build.xml          hadoop-0.20.2-cdh3u0-core.jar      hadoop-examples-0.20.2-cdh3u0.jar  lib
CHANGES.txt        hadoop-0.20.2-cdh3u0-examples.jar  hadoop-examples.jar                logs
cloudera           hadoop-0.20.2-cdh3u0-test.jar      hadoop-test-0.20.2-cdh3u0.jar      NOTICE.txt
cloudera-pom.xml   hadoop-0.20.2-cdh3u0-tools.jar     hadoop-test.jar                    part-00001
conf               hadoop-ant-0.20.2-cdh3u0.jar       hadoop-tools-0.20.2-cdh3u0.jar     pids
contrib            hadoop-ant.jar                     hadoop-tools.jar                   README.txt
example-confs      hadoop-core-0.20.2-cdh3u0.jar      ivy                                webapps
root@cloudera-vm:/usr/lib/hadoop-0.20# mkdir test
root@cloudera-vm:/usr/lib/hadoop-0.20# cd test
root@cloudera-vm:/usr/lib/hadoop-0.20/test# hadoop fs -get /user/teragen-inputTest/part-00001 .
root@cloudera-vm:/usr/lib/hadoop-0.20/test# ls
part-00001
root@cloudera-vm:/usr/lib/hadoop-0.20/test# vi part-00001
root@cloudera-vm:/usr/lib/hadoop-0.20/test# vi part-00001
root@cloudera-vm:/usr/lib/hadoop-0.20/test#
```

Dump of a MR Job

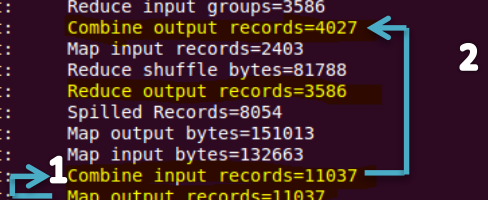
edureka!

```
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
```

Dump of a MR Job

edureka!

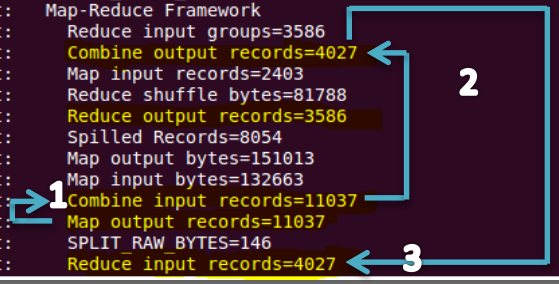
```
13/08/03 00:58:40 WARN mapred.JobClient: Use GenericOptionsParser for parsing th
e arguments. Applications should implement Tool for the same.
13/08/03 00:58:40 INFO mapred.FileInputFormat: Total input paths to process : 1
13/08/03 00:58:40 INFO mapred.JobClient: Running job: job_201308022025_0003
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
```



Dump of a MR Job

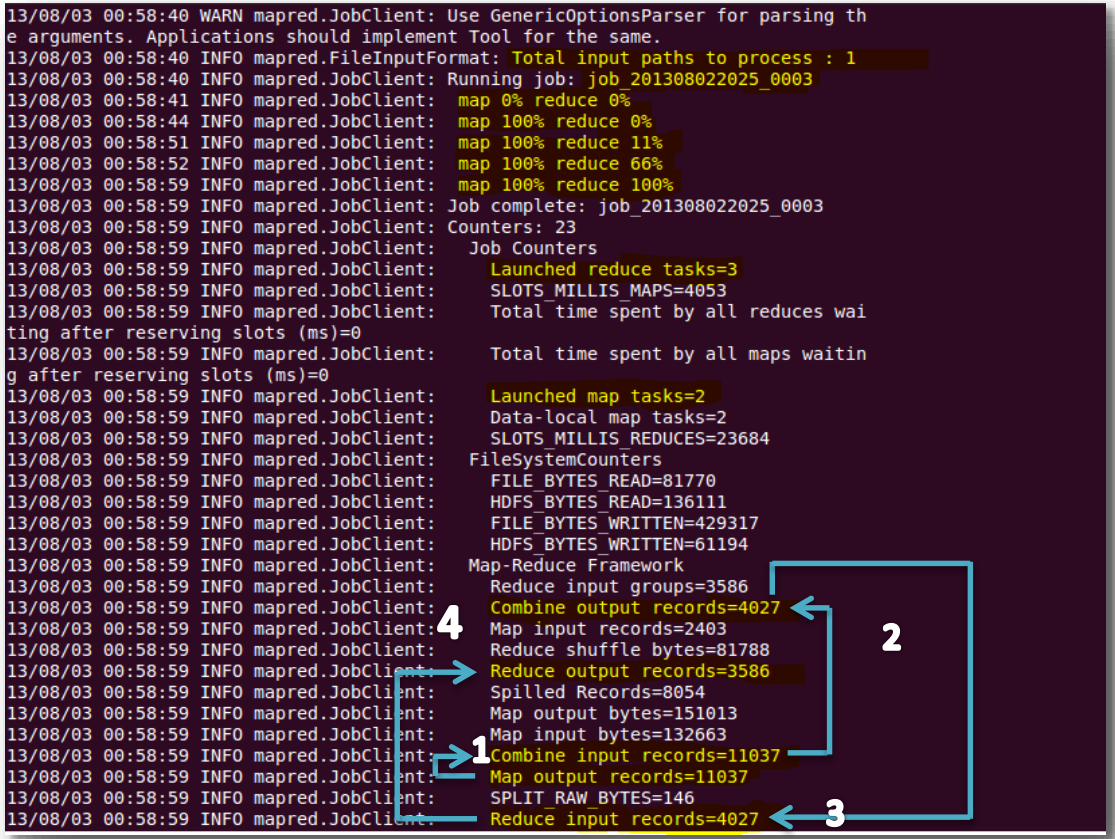
edureka!

```
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
```



Dump of a MR Job

edureka!



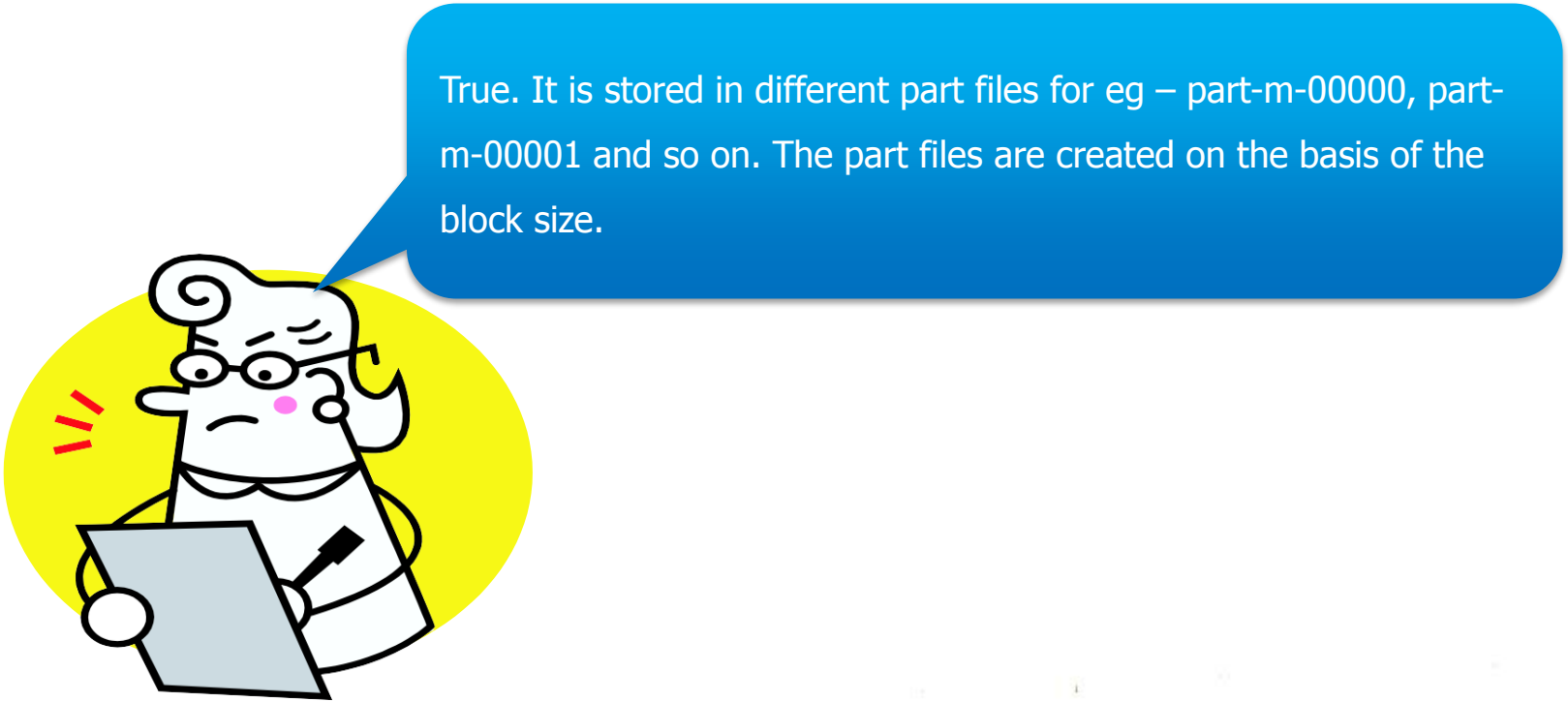
Annie's Question

edureka!

The output of a MR job will be stored on HDFS:

- TRUE
- FALSE

Annie's Answer



Annie's Answer

edureka!

To run MR job data should be present on HDFS:

- TRUE
- FALSE

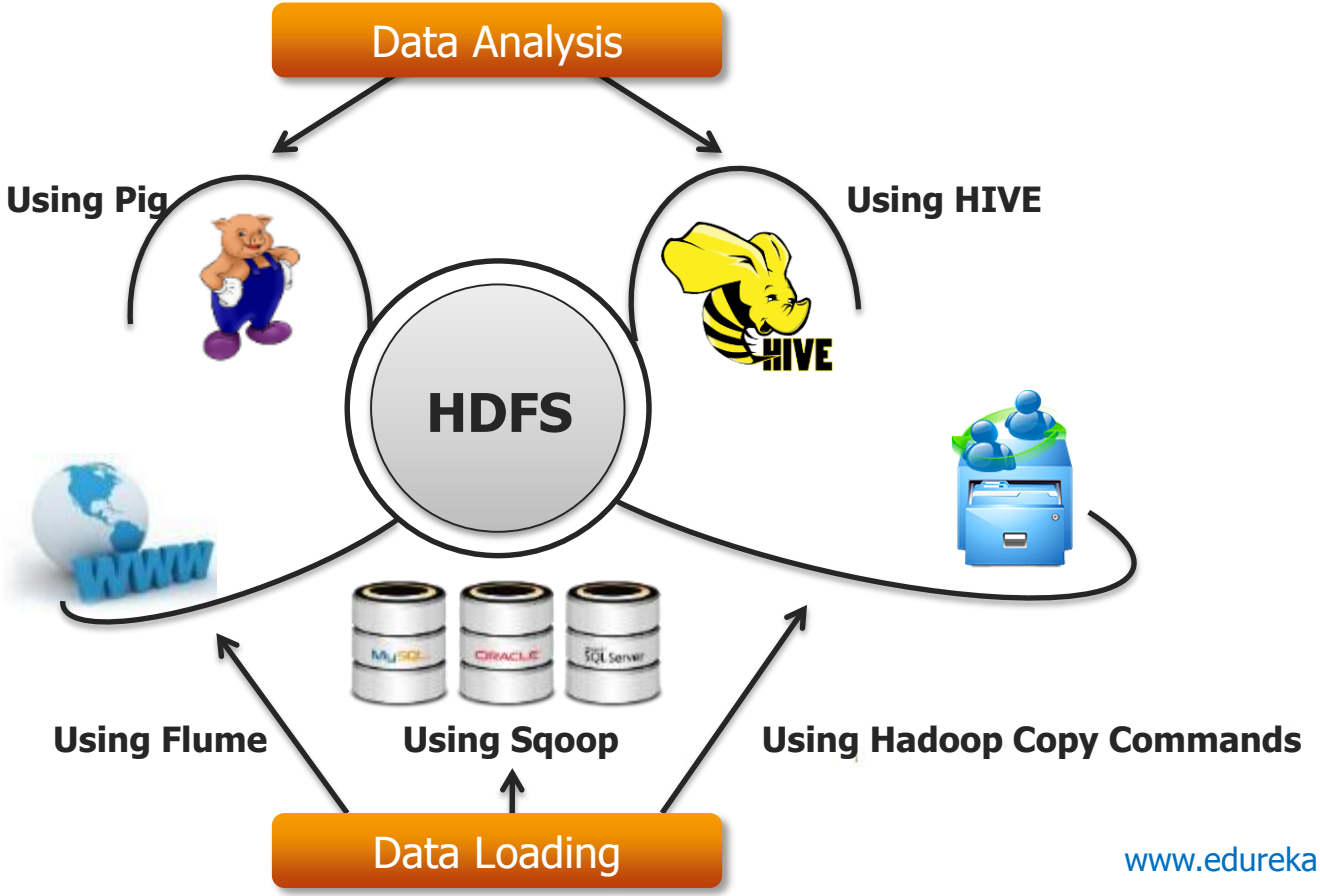
Annie's Answer

edureka!

True. In order to process data in parallel it is necessary that it is present on HDFS so that MR can work on chunks of data in parallel.

Data Loading Techniques and Data Analysis

edureka!



Hadoop Copy Commands

edureka!

put: Copy file(s) from local file system to destination file system. It can also read from "stdin" and writes to destination file system.

```
hadoop dfs –put weather.txt hdfs://<target Namenode>
```

copyFromLocal: Similar to "put" command, except that the source is restricted to a local file reference.

```
hadoop dfs –copyFromLocal weather.txt hdfs://<target Namenode>
```

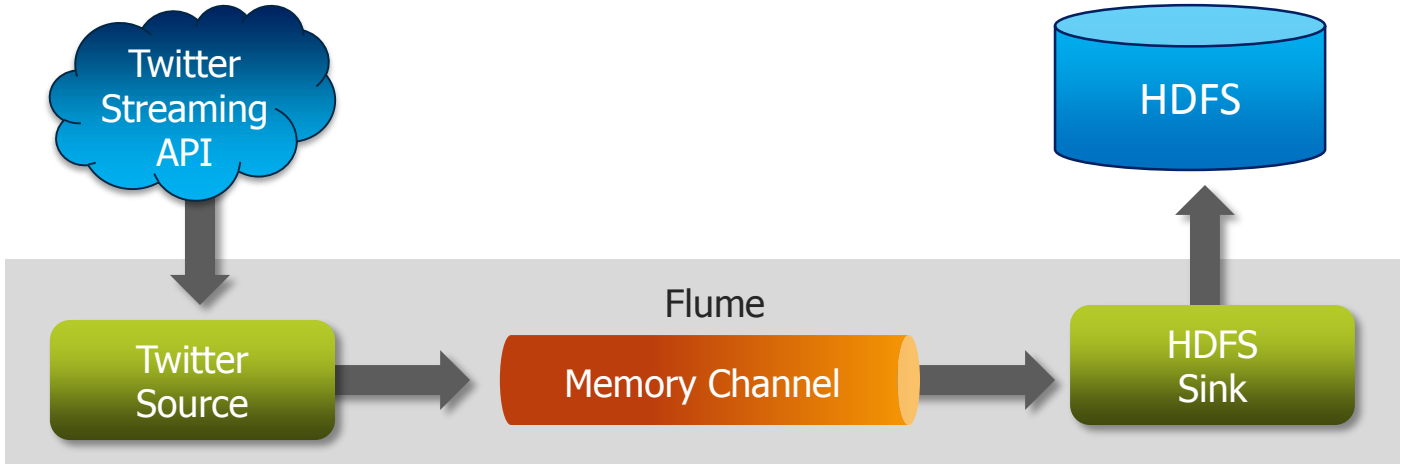
distcp: Distributed Copy to move data between clusters, used for backup and recovery

```
hadoop distcp hdfs://<source NN> hdfs://<target NN>
```

Data Loading Using Flume

edureka!

Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming event data.



Data Loading Using Sqoop

edureka!

Apache Sqoop (TM) is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured data stores such as relational databases.

- ✓ Imports individual tables or entire databases to HDFS.
- ✓ Generates Java classes to allow you to interact with your imported data.
- ✓ Provides the ability to import from SQL databases straight into your Hive data warehouse.

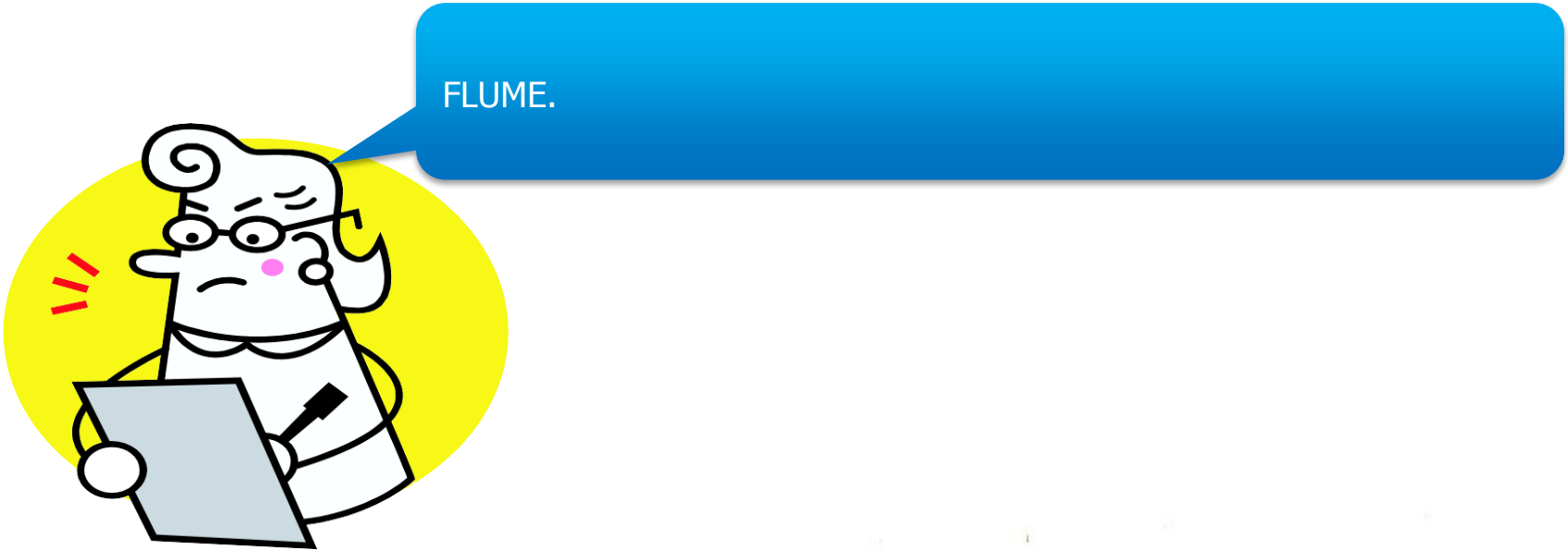
Annie's Question

edureka!

Your website is hosting a group of more than 300 sub-websites. You want to have an analytics on the shopping patterns of different visitors? What is the best way to collect those information from the weblogs?

- SQOOP
- FLUME

Annie’s Answer



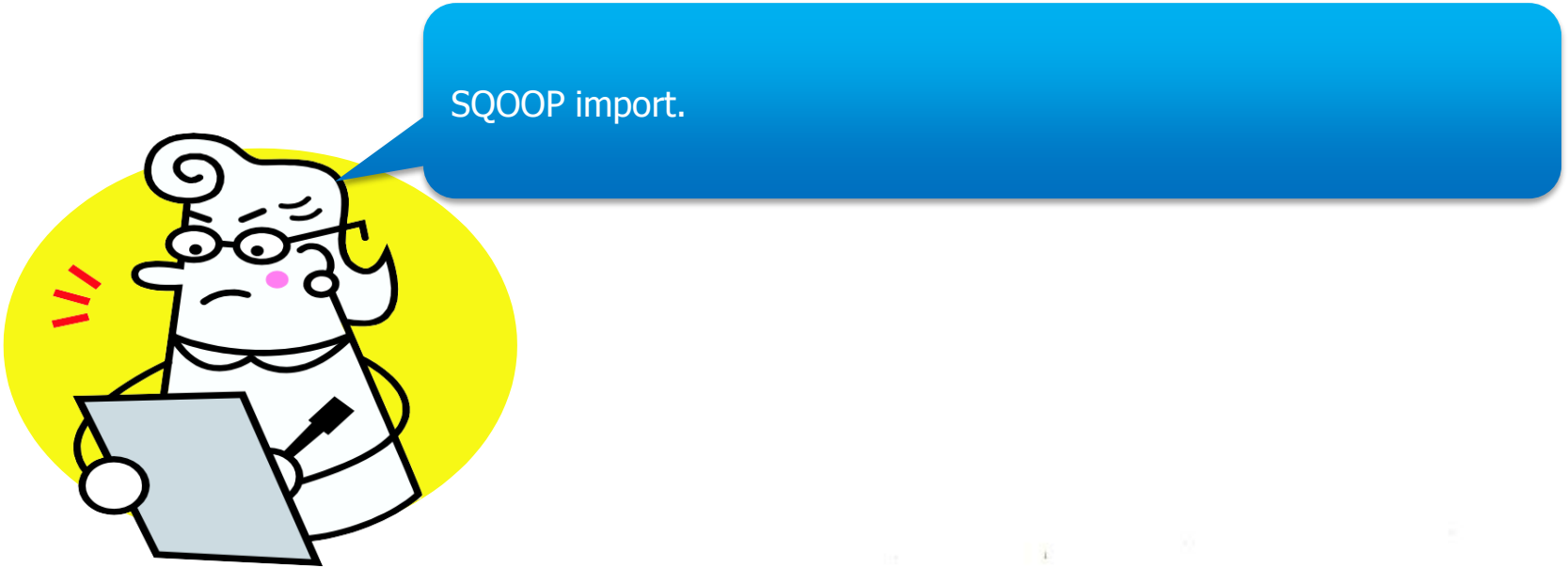
Annie's Question

edureka!

You want to join data collected from two sources. One source of data collected from a big database of call records is already available in HDFS. The another source of data is available in a database table. The best way to move that data in HDFS is:

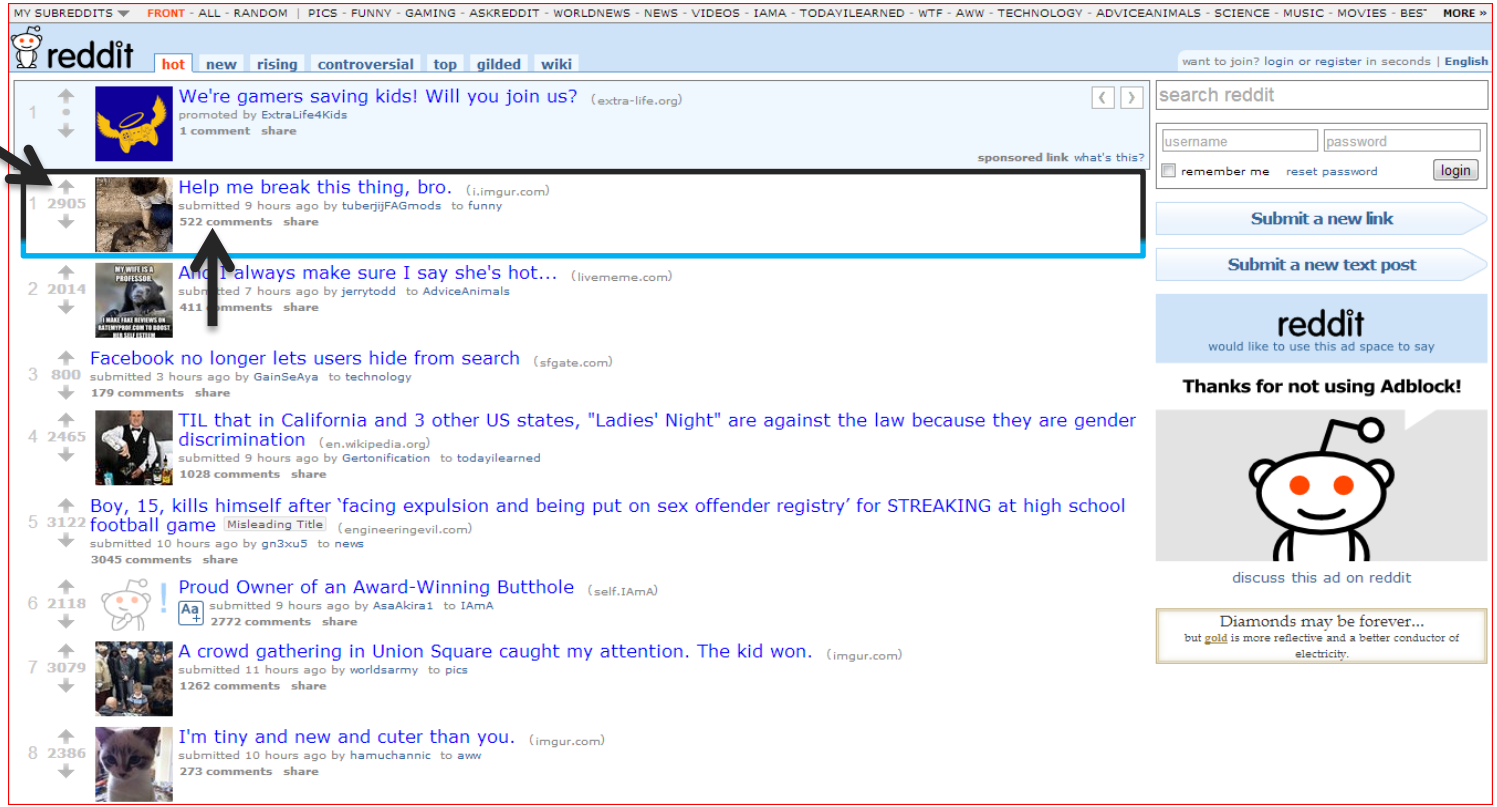
- SQOOP import
- PIG script
- Hive Query

Annie’s Answer



Use Case – How do I find out the best?

edureka!



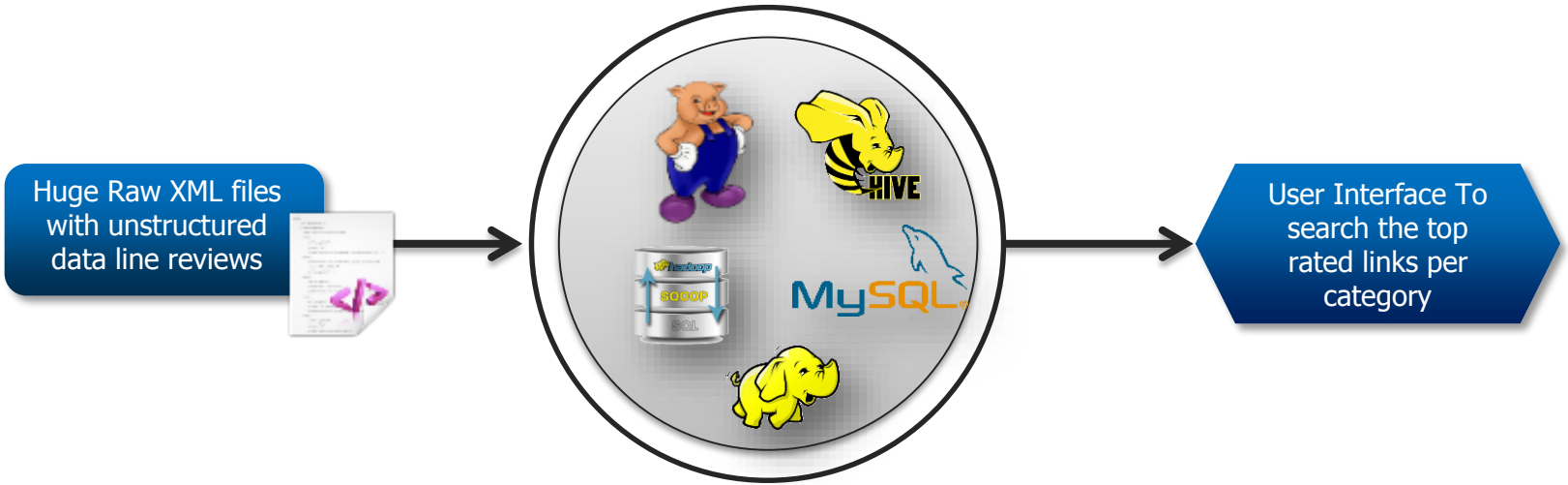
Use Case – Data to Analyse

edureka!

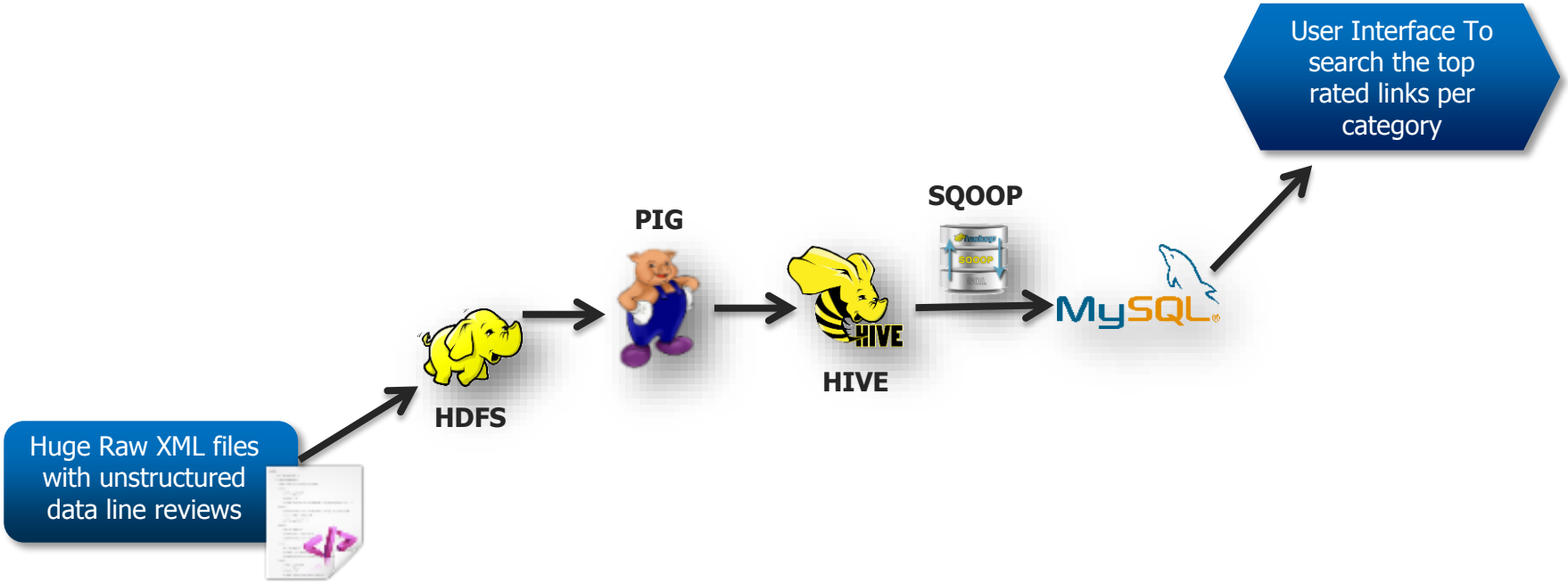
```
<?xml version="1.0" encoding="UTF-8"?>
<documents>
  <document>
    <hash>0c539040d03f4ba7edb386b3d5b7c8c8</hash>
    <url>http://www.bemboszoo.com/</url>
    <category>Top/Reference/Education/Early_Childhood/Subjects/Language_Arts/Alphabet</category>
    <usercount>724</usercount>
        <tag>
          <name><![CDATA[fun]]></name>
          <count>114</count>
        </tag>
        <tag>
          <name><![CDATA[kids]]></name>
          <count>69</count>
        </tag>
        <tag>
          <name><![CDATA[alphabet]]></name>
          <count>66</count>
        </tag>
        <tag>
          <name><![CDATA[children]]></name>
          <count>55</count>
        </tag>
    </tags>
    <reviews>
        <review><![CDATA["This high concept abcdedary, the picture book debut for deVicq Cumptich, should delight collectors c
Publishers Weekly]]></review>
        <review><![CDATA[Really neat flash site making animal art out of the letters of the alphabet.]]></review>
        <review><![CDATA[I love this site!]]></review>
        <review><![CDATA[Since I'm forver young LOL I LOVE this and it's even better in the flash version!!!!!! :)]]></review>
        <review><![CDATA[Cool once you get past the illegible splash page.]]></review>
        <review><![CDATA[Silly, yet clever and funny.  Probably helps if you have some interest in typography, and know that E
        <review><![CDATA[Clever!

]]></review>
```

Abstract Flow Diagram



Flow Diagram



Demo for Data Load

Further Reading

edureka!

- ✓ Hadoop Cluster Setup
  http://hadoop.apache.org/docs/r0.19.1/cluster_setup.html
- ✓ Hadoop on Amazon AWS ec2
  http://www.edureka.in/blog/install-apache-hadoop-cluster/
- ✓ Hadoop Hardware Selection
  http://blog.cloudera.com/blog/2013/08/how-to-select-the-right-hardware-for-your-new-hadoop-cluster/
  http://docs.hortonworks.com/HDPDocuments/HDP1/HDP-Win-1.3.0/bk_cluster-planning-guide/content/ch_hardware-recommendations.html
- ✓ Hadoop Cluster Configuration
  http://www.edureka.in/blog/hadoop-cluster-configuration-files/

Further Reading

edureka!

- ✓ MapReduce Job execution
  http://www.edureka.in/blog/anatomy-of-a-mapredcue-job-in-apache-hadoop/
- ✓ Add/Remove Nodes in a Cluster
  http://www.edureka.in/blog/commissioning-and-decommissioning-nodes-in-a-hadoop-cluster/
- ✓ Secondary Namenode
  https://hadoop.apache.org/docs/r1.2.1/hdfs_user_guide.html#Secondary+NameNode

Module-3 Pre-work

edureka!

- ✓ Setup the Hadoop development environment using the documents present in the LMS.
  - ✓ Flume Set-up on Cloudera
  - ✓ SQOOP Set-up on Cloudera
  - ✓ Refresh your Java Skills using Java Essential for Hadoop Tutorial
- ✓ Attempt the Module-2 Assignments present in the LMS.
- ✓ Review the Interview Questions for setting up hadoop cluster
  http://www.edureka.in/blog/hadoop-interview-questions-hadoop-cluster/

edureka!

Thank You

See You in Class Next Week