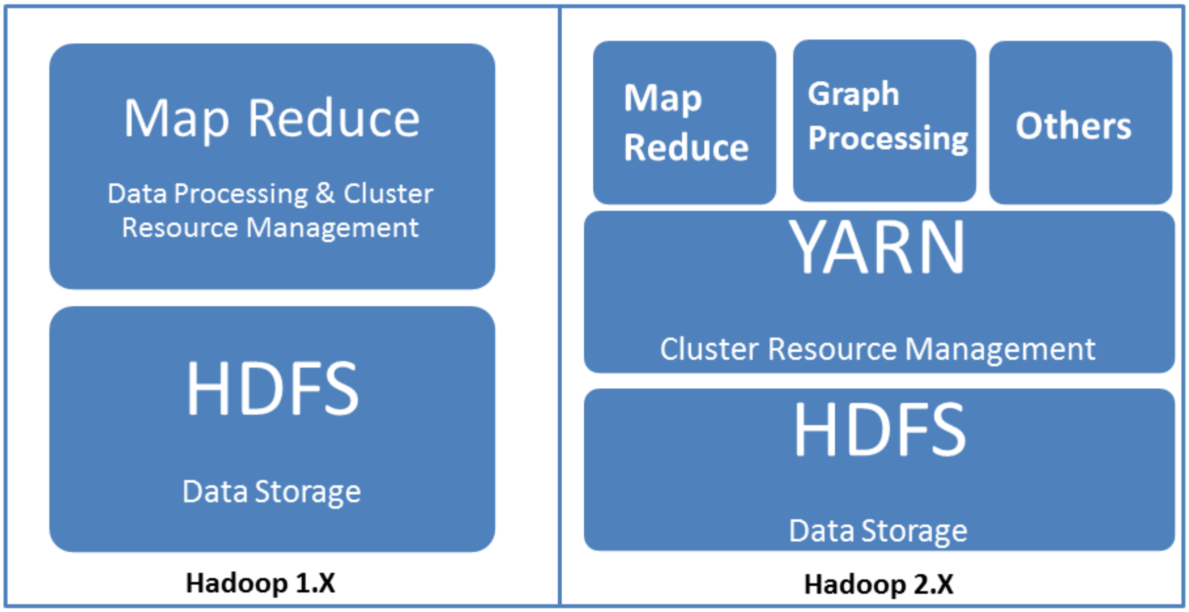
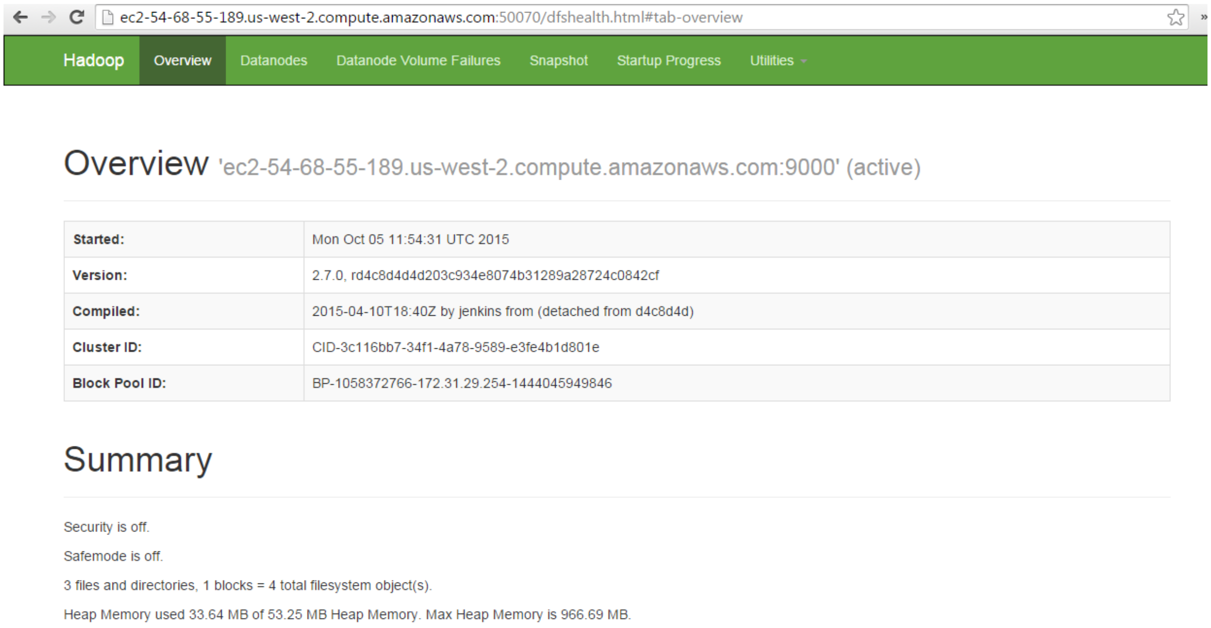
Chapter 1 : Getting Started with Hadoop




ec2-54-68-55-189.us-west-2.compute.amazonaws.com:50070/dfshealth.html#tab-overview
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities

Overview 'ec2-54-68-55-189.us-west-2.compute.amazonaws.com:9000' (active)

Started:	Mon Oct 05 11:54:31 UTC 2015
Version:	2.7.0, rd4c8d4d4d203c934e8074b31289a28724c0842cf
Compiled:	2015-04-10T18:40Z by jenkins from (detached from d4c8d4d)
Cluster ID:	CID-3c116bb7-34f1-4a78-9589-e3fe4b1d801e
Block Pool ID:	BP-1058372766-172.31.29.254-1444045949846

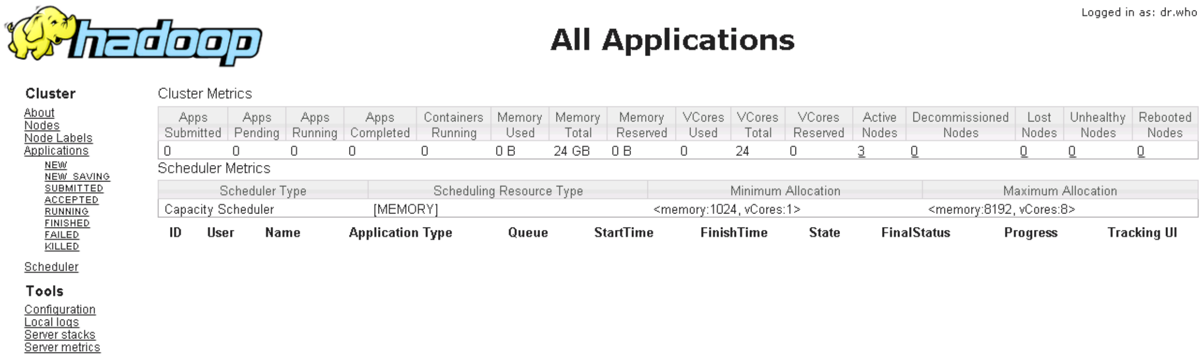
Summary

Security is off.

Safemode is off.

3 files and directories, 1 blocks = 4 total filesystem object(s).

Heap Memory used 33.64 MB of 53.25 MB Heap Memory. Max Heap Memory is 966.69 MB.



Logged in as: dr.who

All Applications

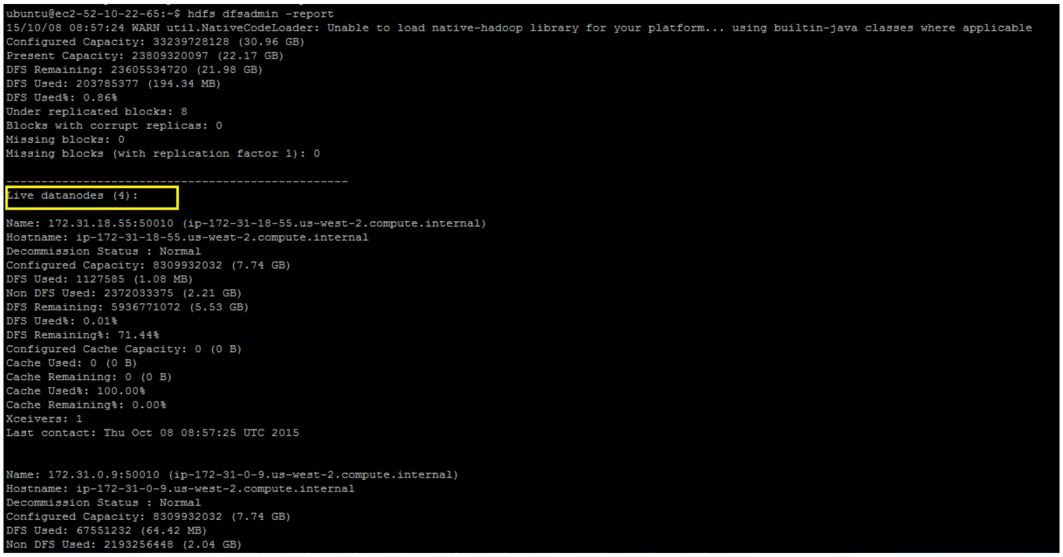
Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
0	0	0	0	0	0 B	24 GB	0 B	0	24	0	3	0	0	0	0

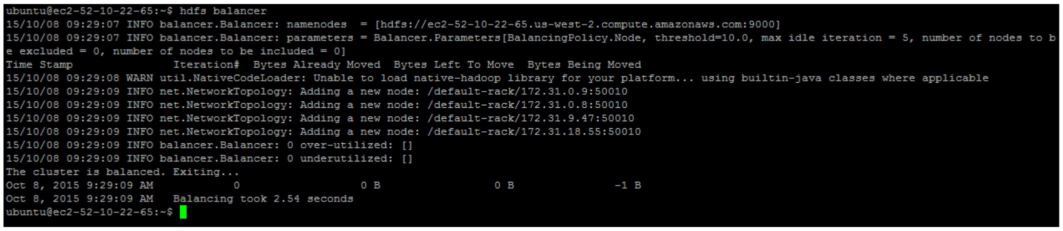
Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[MEMORY]	<memory:1024, vCores:1>	<memory:8192, vCores:8>

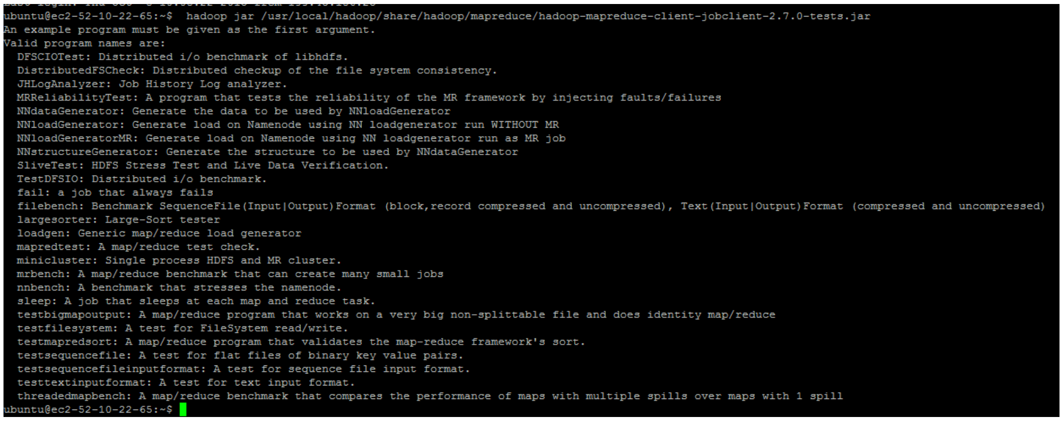
ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI



```
ubuntu@ec2-52-10-22-65:~$ hdfs dfsadmin -report
15/10/08 08:57:24 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Configured Capacity: 33239728128 (30.96 GB)
Present Capacity: 23809320097 (22.17 GB)
DFS Remaining: 23605534720 (21.98 GB)
DFS Used: 203785377 (194.34 MB)
DFS Used%: 0.86%
Under replicated blocks: 8
Blocks with corrupt replicas: 0
Missing blocks: 0
Missing blocks (with replication factor 1): 0

-------------------------------------------------
Live datanodes (4):

Name: 172.31.18.55:50010 (ip-172-31-18-55.us-west-2.compute.internal)
Hostname: ip-172-31-18-55.us-west-2.compute.internal
Decommission Status : Normal
Configured Capacity: 8309932032 (7.74 GB)
DFS Used: 1127585 (1.08 MB)
Non DFS Used: 2372033375 (2.21 GB)
DFS Remaining: 5936771072 (5.53 GB)
DFS Used%: 0.01%
DFS Remaining%: 71.44%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Thu Oct 08 08:57:25 UTC 2015


Name: 172.31.0.9:50010 (ip-172-31-0-9.us-west-2.compute.internal)
Hostname: ip-172-31-0-9.us-west-2.compute.internal
Decommission Status : Normal
Configured Capacity: 8309932032 (7.74 GB)
DFS Used: 67551232 (64.42 MB)
Non DFS Used: 2193256448 (2.04 GB)
```

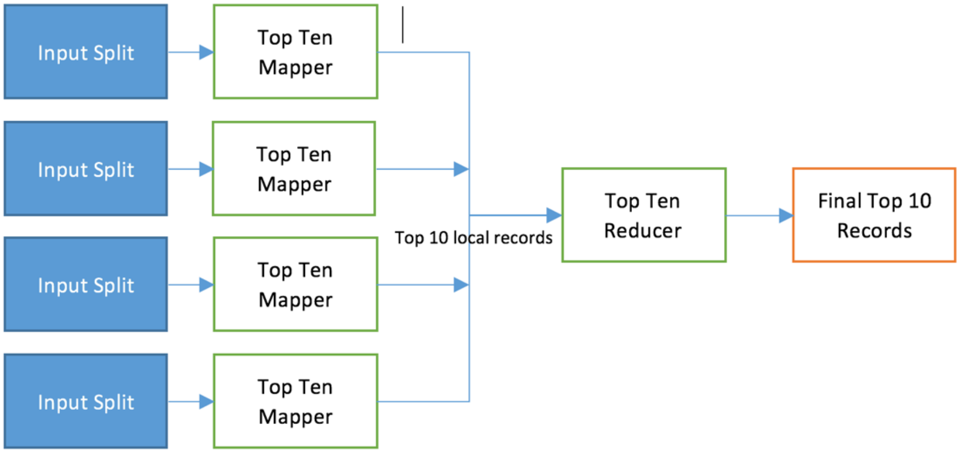


```
ubuntu@ec2-52-10-22-65:~$ hdfs balancer
15/10/08 09:29:07 INFO balancer.Balancer: namenodes  = [hdfs://ec2-52-10-22-65.us-west-2.compute.amazonaws.com:9000]
15/10/08 09:29:07 INFO balancer.Balancer: parameters = Balancer.Parameters[BalancingPolicy.Node, threshold=10.0, max idle iteration = 5, number of nodes to b
e excluded = 0, number of nodes to be included = 0]
Time Stamp               Iteration#  Bytes Already Moved  Bytes Left To Move  Bytes Being Moved
15/10/08 09:29:08 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
15/10/08 09:29:09 INFO net.NetworkTopology: Adding a new node: /default-rack/172.31.0.9:50010
15/10/08 09:29:09 INFO net.NetworkTopology: Adding a new node: /default-rack/172.31.0.8:50010
15/10/08 09:29:09 INFO net.NetworkTopology: Adding a new node: /default-rack/172.31.9.47:50010
15/10/08 09:29:09 INFO net.NetworkTopology: Adding a new node: /default-rack/172.31.18.55:50010
15/10/08 09:29:09 INFO balancer.Balancer: 0 over-utilized: []
15/10/08 09:29:09 INFO balancer.Balancer: 0 underutilized: []
The cluster is balanced. Exiting...
Oct 8, 2015 9:29:09 AM            0                  0 B                 0 B               -1 B
Oct 8, 2015 9:29:09 AM   Balancing took 2.54 seconds
ubuntu@ec2-52-10-22-65:~$
```

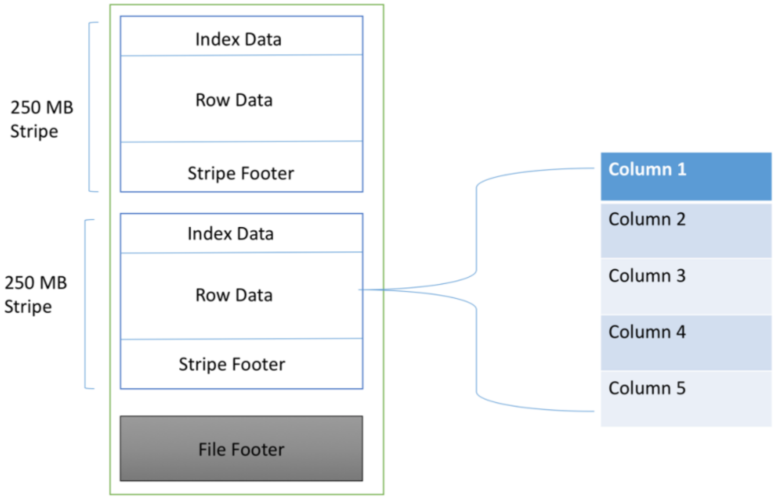


```
ubuntu@ec2-52-10-22-65:~$  hadoop jar /usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-client-jobclient-2.7.0-tests.jar
An example program must be given as the first argument.
Valid program names are:
  DFSCIOTest: Distributed i/o benchmark of libhdfs.
  DistributedFSCheck: Distributed checkup of the file system consistency.
  JHLogAnalyzer: Job History Log analyzer.
  MRReliabilityTest: A program that tests the reliability of the MR framework by injecting faults/failures
  NNdataGenerator: Generate the data to be used by NNloadGenerator
  NNloadGenerator: Generate load on Namenode using NN loadgenerator run WITHOUT MR
  NNloadGeneratorMR: Generate load on Namenode using NN loadgenerator run as MR job
  NNstructureGenerator: Generate the structure to be used by NNdataGenerator
  SliveTest: HDFS Stress Test and Live Data Verification.
  TestDFSIO: Distributed i/o benchmark.
  fail: a job that always fails
  filebench: Benchmark SequenceFile(Input|Output)Format (block,record compressed and uncompressed), Text(Input|Output)Format (compressed and uncompressed)
  largesorter: Large-Sort tester
  loadgen: Generic map/reduce load generator
  mapredtest: A map/reduce test check.
  minicluster: Single process HDFS and MR cluster.
  mrbench: A map/reduce benchmark that can create many small jobs
  nnbench: A benchmark that stresses the namenode.
  sleep: A job that sleeps at each map and reduce task.
  testbigmapoutput: A map/reduce program that works on a very big non-splittable file and does identity map/reduce
  testfilesystem: A test for FileSystem read/write.
  testmapredsort: A map/reduce program that validates the map-reduce framework's sort.
  testsequencefile: A test for flat files of binary key value pairs.
  testsequencefileinputformat: A test for sequence file input format.
  testtextinputformat: A test for text input format.
  threadedmapbench: A map/reduce benchmark that compares the performance of maps with multiple spills over maps with 1 spill
ubuntu@ec2-52-10-22-65:~$
```

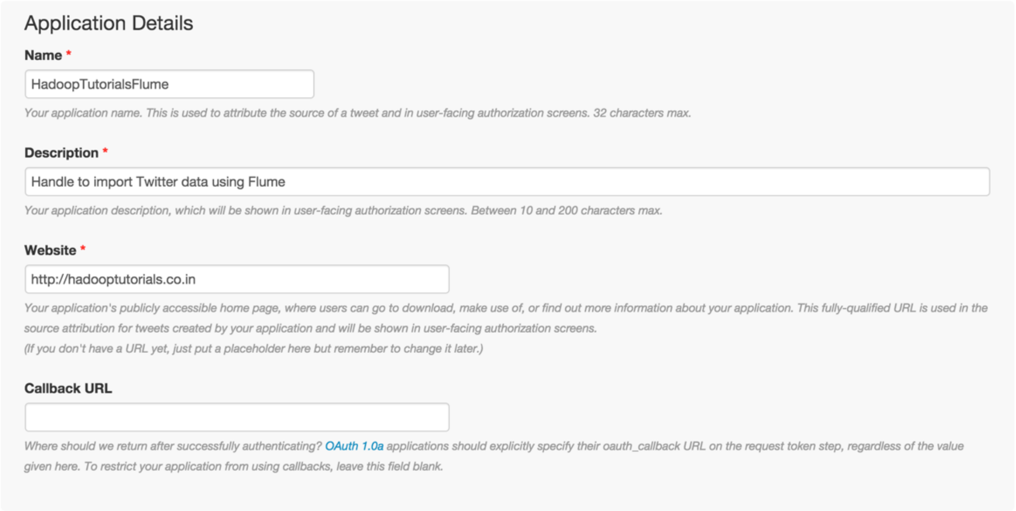
Chapter 3 : Mastering Map Reduce Programs



Chapter 4 : Performing Common tasks using Hive, Pig and Hbase



Chapter 6: Data import/Export using Sqoop and Flume



Application Details

Name *

HadoopTutorialsFlume

Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.

Description *

Handle to import Twitter data using Flume

Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

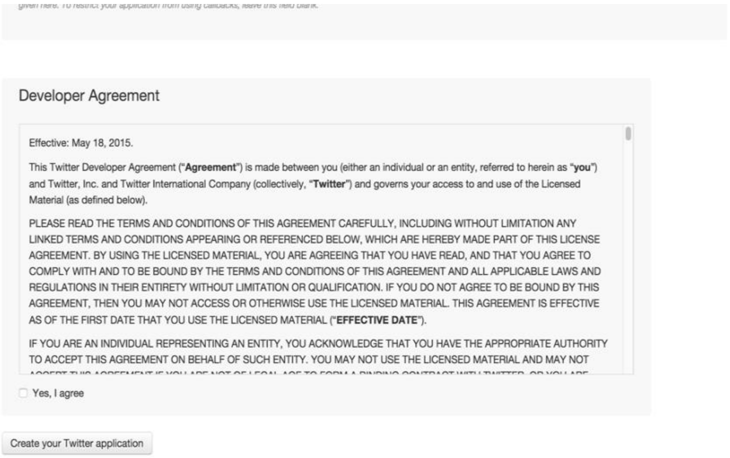
Website *

http://hadooptutorials.co.in

Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens.
(If you don't have a URL yet, just put a placeholder here but remember to change it later.)

Callback URL

Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth_callback URL on the request token step, regardless of the value given here. To restrict your application from using callbacks, leave this field blank.



given here. To restrict your application from using callbacks, leave this field blank.

Developer Agreement

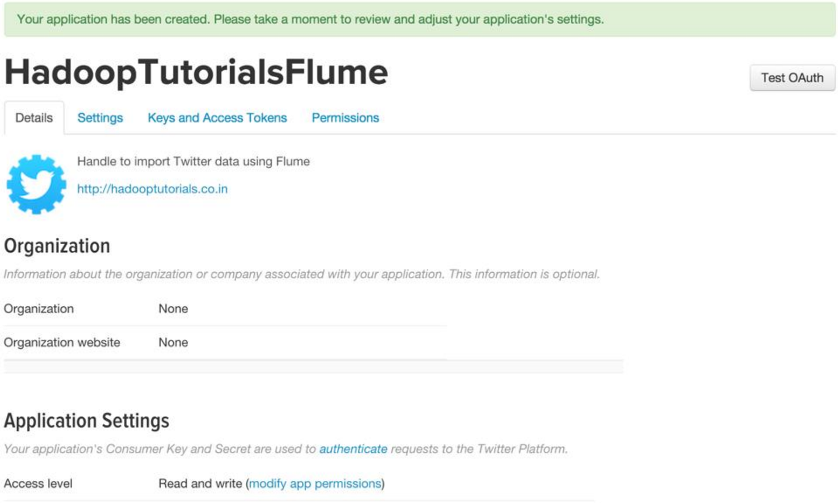
Effective: May 18, 2015.

This Twitter Developer Agreement ("**Agreement**") is made between you (either an individual or an entity, referred to herein as "**you**") and Twitter, Inc. and Twitter International Company (collectively, "**Twitter**") and governs your access to and use of the Licensed Material (as defined below).

PLEASE READ THE TERMS AND CONDITIONS OF THIS AGREEMENT CAREFULLY, INCLUDING WITHOUT LIMITATION ANY LINKED TERMS AND CONDITIONS APPEARING OR REFERENCED BELOW, WHICH ARE HEREBY MADE PART OF THIS LICENSE AGREEMENT. BY USING THE LICENSED MATERIAL, YOU ARE AGREEING THAT YOU HAVE READ, AND THAT YOU AGREE TO COMPLY WITH AND TO BE BOUND BY THE TERMS AND CONDITIONS OF THIS AGREEMENT AND ALL APPLICABLE LAWS AND REGULATIONS IN THEIR ENTIRETY WITHOUT LIMITATION OR QUALIFICATION. IF YOU DO NOT AGREE TO BE BOUND BY THIS AGREEMENT, THEN YOU MAY NOT ACCESS OR OTHERWISE USE THE LICENSED MATERIAL. THIS AGREEMENT IS EFFECTIVE AS OF THE FIRST DATE THAT YOU USE THE LICENSED MATERIAL ("**EFFECTIVE DATE**").

IF YOU ARE AN INDIVIDUAL REPRESENTING AN ENTITY, YOU ACKNOWLEDGE THAT YOU HAVE THE APPROPRIATE AUTHORITY TO ACCEPT THIS AGREEMENT ON BEHALF OF SUCH ENTITY. YOU MAY NOT USE THE LICENSED MATERIAL AND MAY NOT

☐ Yes, I agree



Your application has been created. Please take a moment to review and adjust your application's settings.

HadoopTutorialsFlume

Handle to import Twitter data using Flume

http://hadooptutorials.co.in

Organization

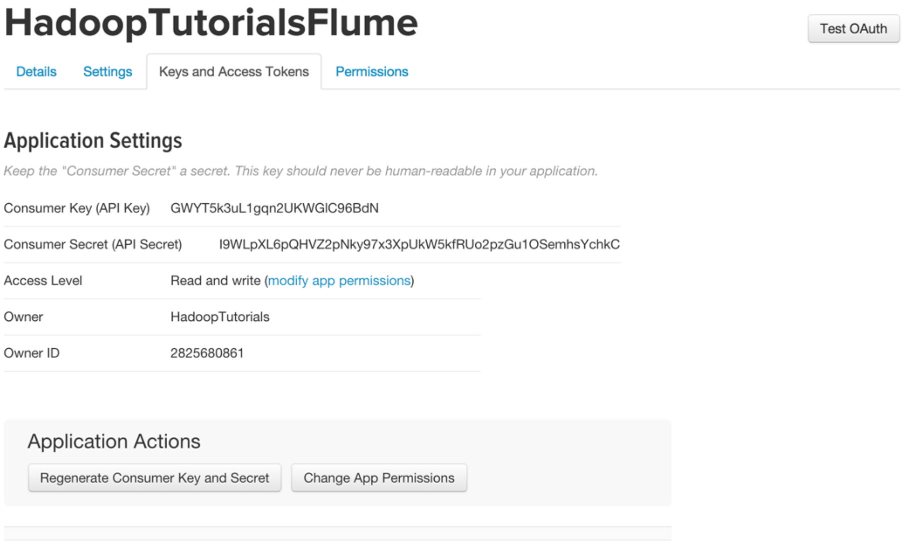
Information about the organization or company associated with your application. This information is optional.

Organization	None
Organization website	None

Application Settings

Your application's Consumer Key and Secret are used to authenticate requests to the Twitter Platform.

Access level	Read and write (modify app permissions)



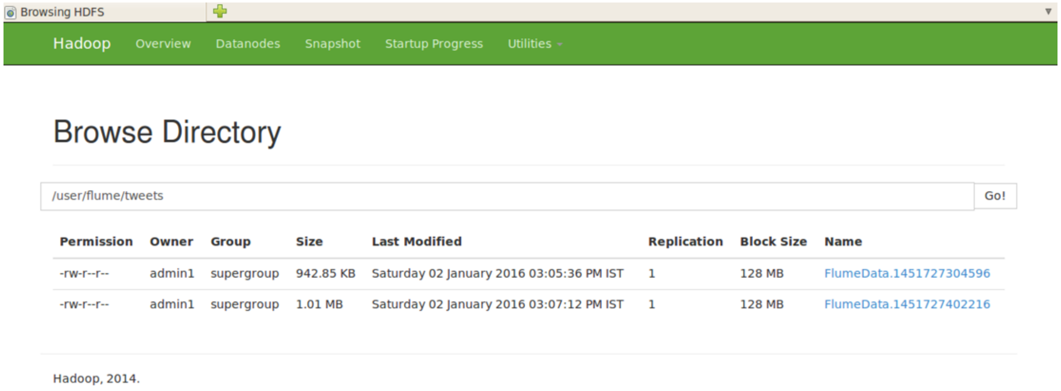
HadoopTutorialsFlume

Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	GWYT5k3uL1gqn2UKWGlC96BdN
Consumer Secret (API Secret)	I9WLpXL6pQHVZ2pNky97x3XpUkW5kfRUo2pzGu1OSemhsYchkC
Access Level	Read and write (modify app permissions)
Owner	HadoopTutorials
Owner ID	2825680861

Application Actions

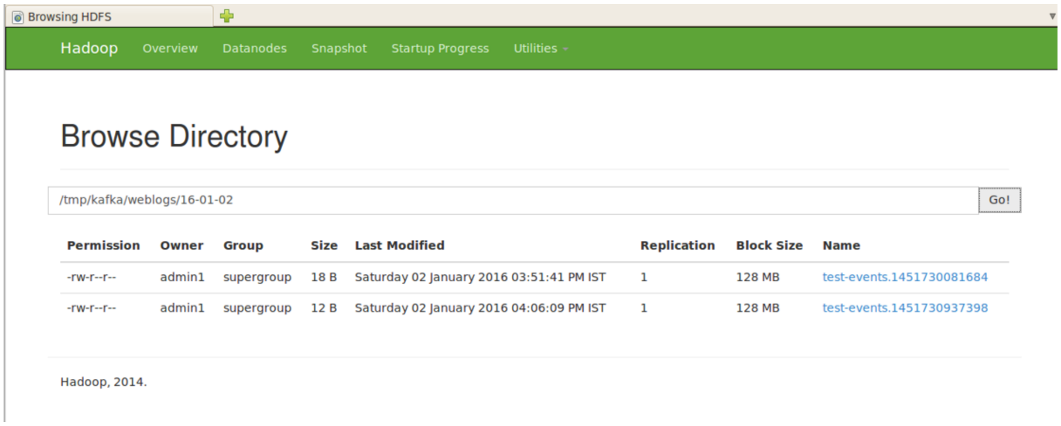


Browse Directory

/user/flume/tweets

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	admin1	supergroup	942.85 KB	Saturday 02 January 2016 03:05:36 PM IST	1	128 MB	FlumeData.1451727304596
-rw-r--r--	admin1	supergroup	1.01 MB	Saturday 02 January 2016 03:07:12 PM IST	1	128 MB	FlumeData.1451727402216

Hadoop, 2014.

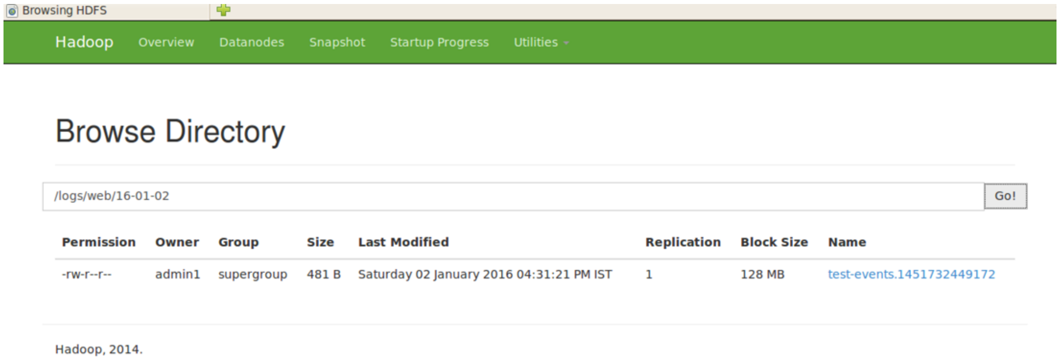


Browse Directory

/tmp/kafka/weblogs/16-01-02

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	admin1	supergroup	18 B	Saturday 02 January 2016 03:51:41 PM IST	1	128 MB	test-events.1451730081684
-rw-r--r--	admin1	supergroup	12 B	Saturday 02 January 2016 04:06:09 PM IST	1	128 MB	test-events.1451730937398

Hadoop, 2014.



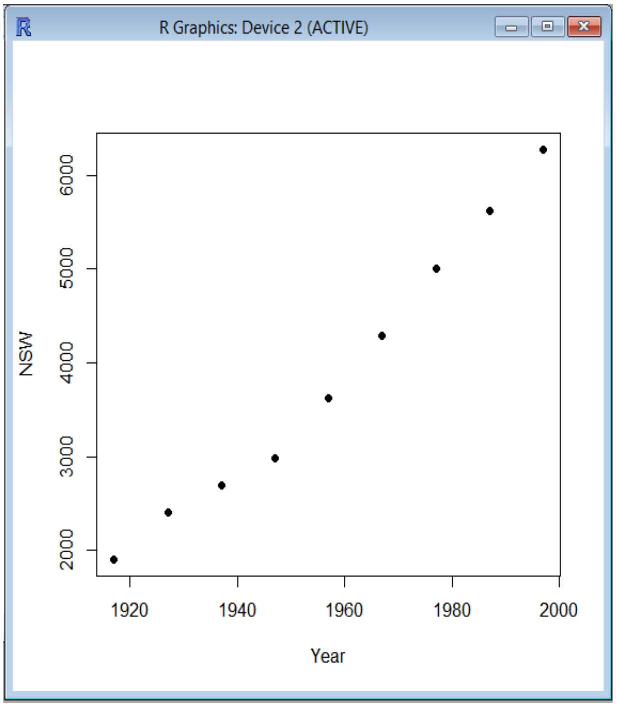
Browse Directory

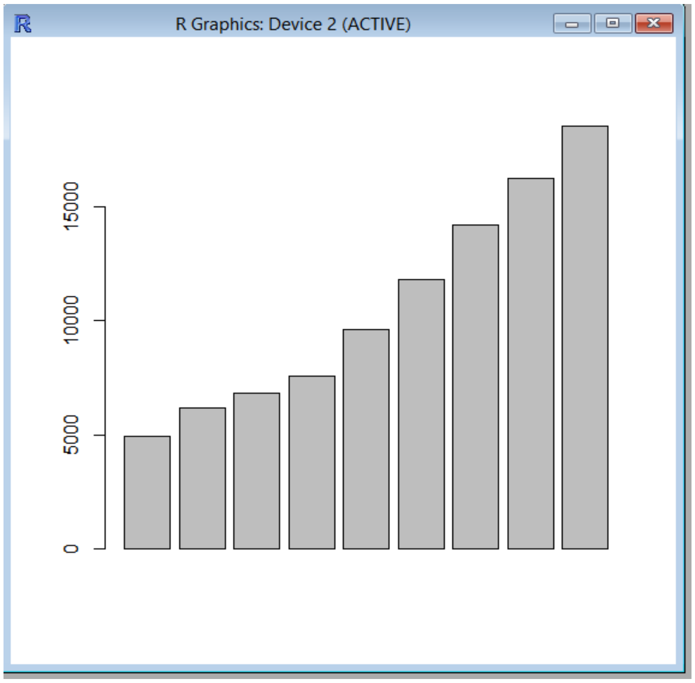
/logs/web/16-01-02

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	admin1	supergroup	481 B	Saturday 02 January 2016 04:31:21 PM IST	1	128 MB	test-events.1451732449172

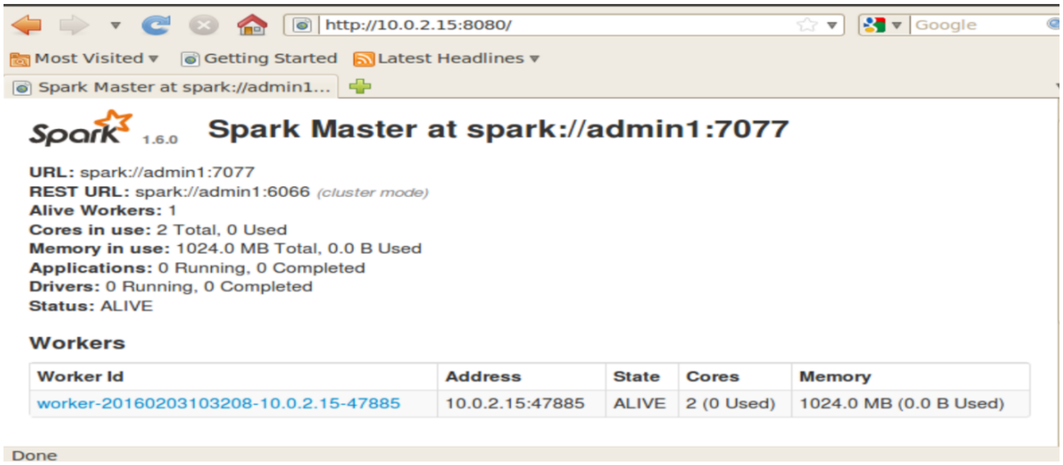
Hadoop, 2014.

Chapter 8: Machine Learning and Predictive Analytics using Mahout and R




R Graphics: Device 2 (ACTIVE)
0
5000
10000
15000

Chapter 9 : Integration with Apache Spark

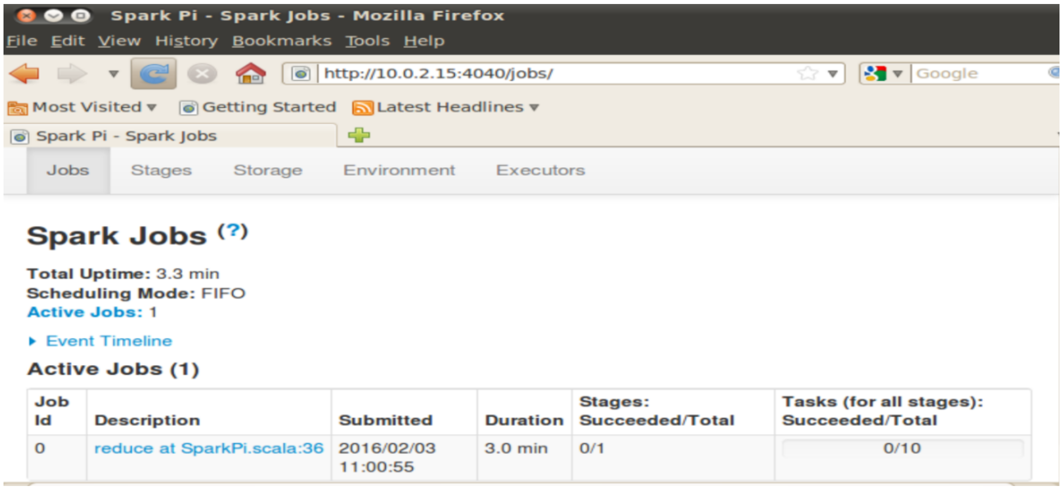


Spark Master at spark://admin1:7077

URL: spark://admin1:7077
REST URL: spark://admin1:6066 *(cluster mode)*
Alive Workers: 1
Cores in use: 2 Total, 0 Used
Memory in use: 1024.0 MB Total, 0.0 B Used
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers

Worker Id	Address	State	Cores	Memory
worker-20160203103208-10.0.2.15-47885	10.0.2.15:47885	ALIVE	2 (0 Used)	1024.0 MB (0.0 B Used)


Spark Pi - Spark Jobs - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://10.0.2.15:4040/jobs/
Google
Most Visited
Getting Started
Latest Headlines
Spark Pi - Spark Jobs
Jobs
Stages
Storage
Environment
Executors

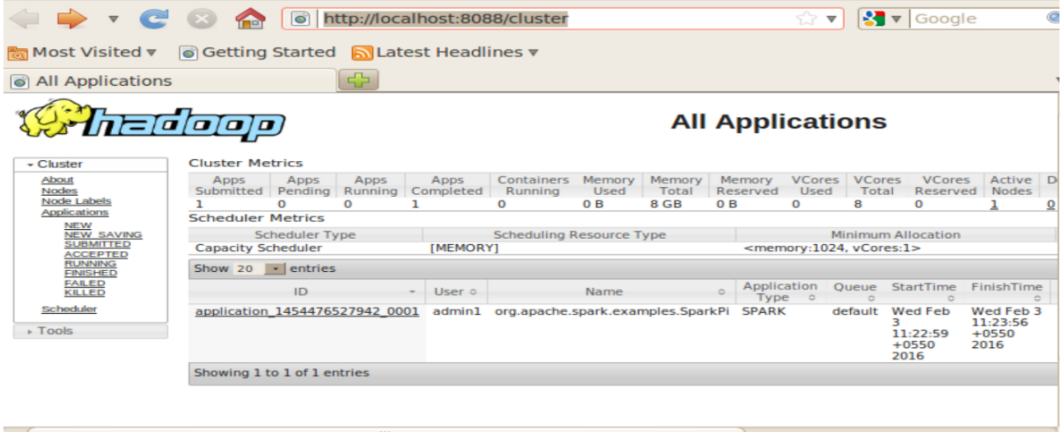
Spark Jobs (?)

Total Uptime: 3.3 min
Scheduling Mode: FIFO
Active Jobs: 1

▸ Event Timeline

Active Jobs (1)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
0	reduce at SparkPi.scala:36	2016/02/03 11:00:55	3.0 min	0/1	0/10



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	D
1	0	0	1	0	0 B	8 GB	0 B	0	8	0	1	0

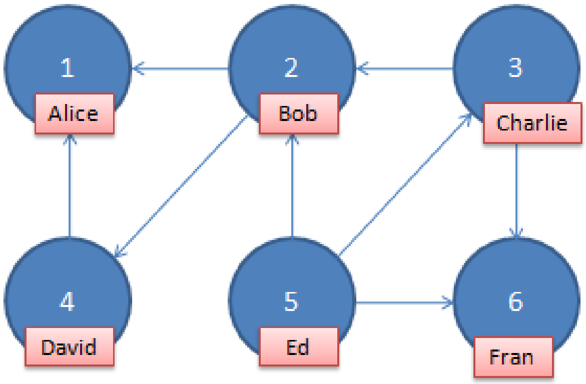
Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation
Capacity Scheduler	[MEMORY]	<memory:1024, vCores:1>

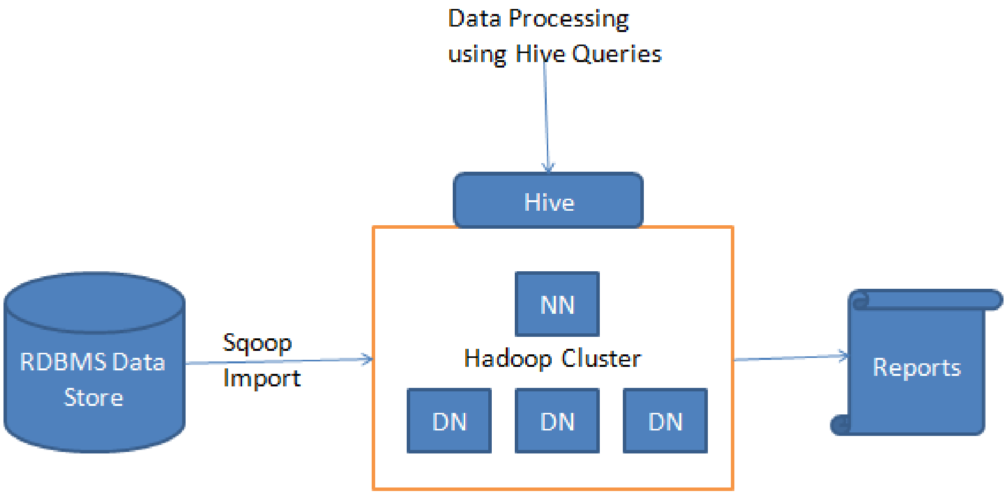
Show 20 entries

ID	User	Name	Application Type	Queue	StartTime	FinishTime
application_1454476527942_0001	admin1	org.apache.spark.examples.SparkPi	SPARK	default	Wed Feb 3 11:22:59 +0550 2016	Wed Feb 3 11:23:56 +0550 2016

Showing 1 to 1 of 1 entries


1
Alice
2
Bob
3
Charlie
4
David
5
Ed
6
Fran

Chapter 10 : Hadoop Use Cases



Call Detail Record Analytics using Hadoop