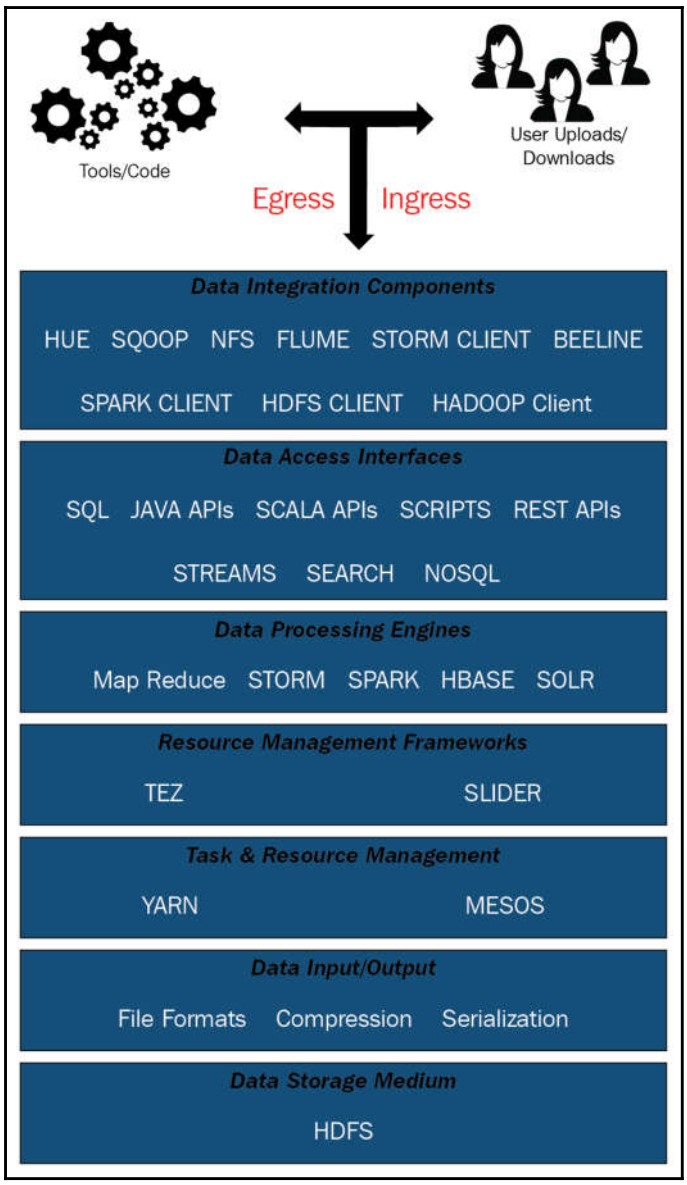
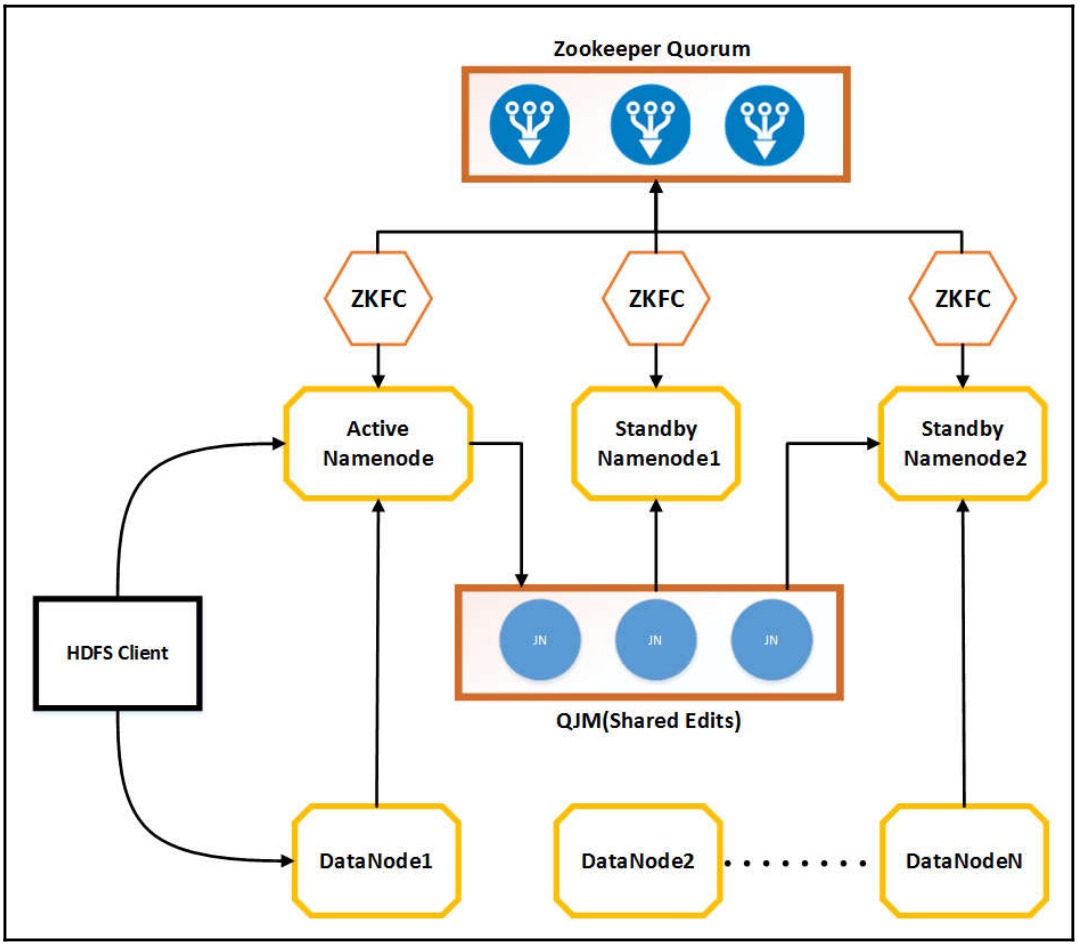
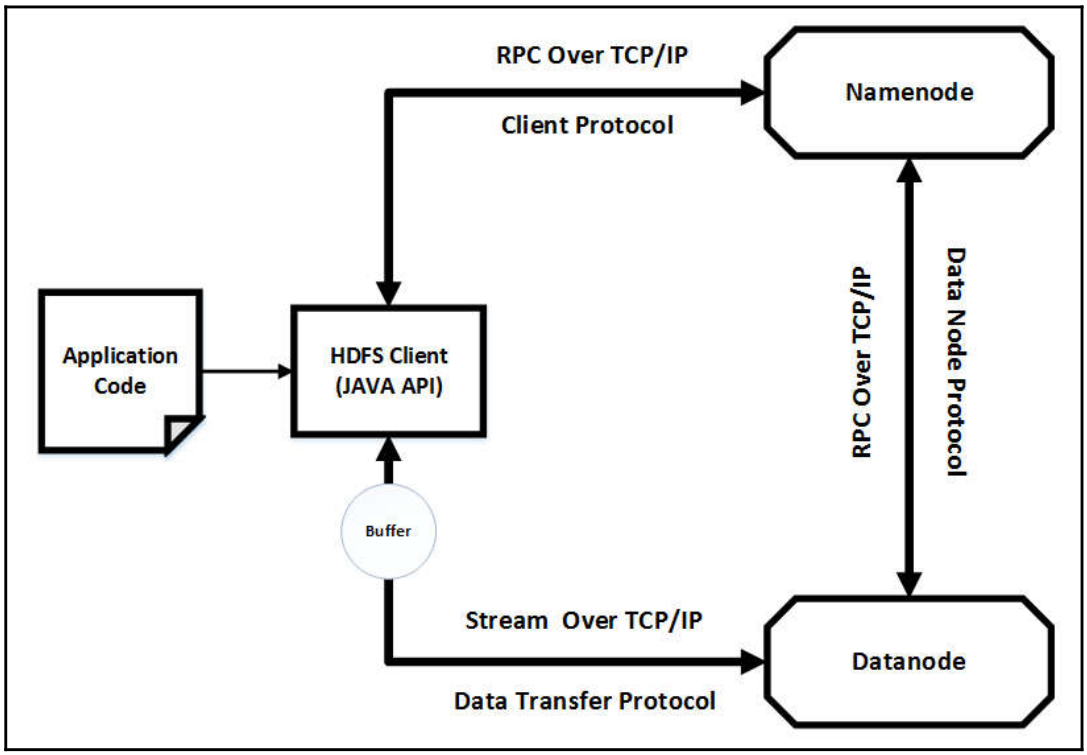
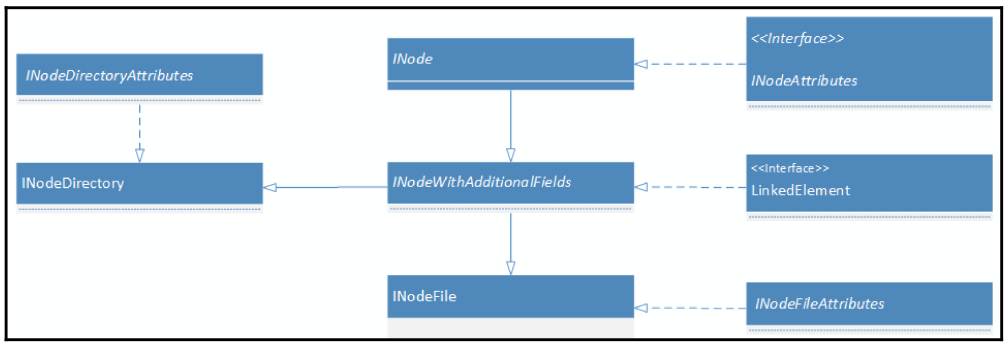
Chapter 1: Journey to Hadoop 3

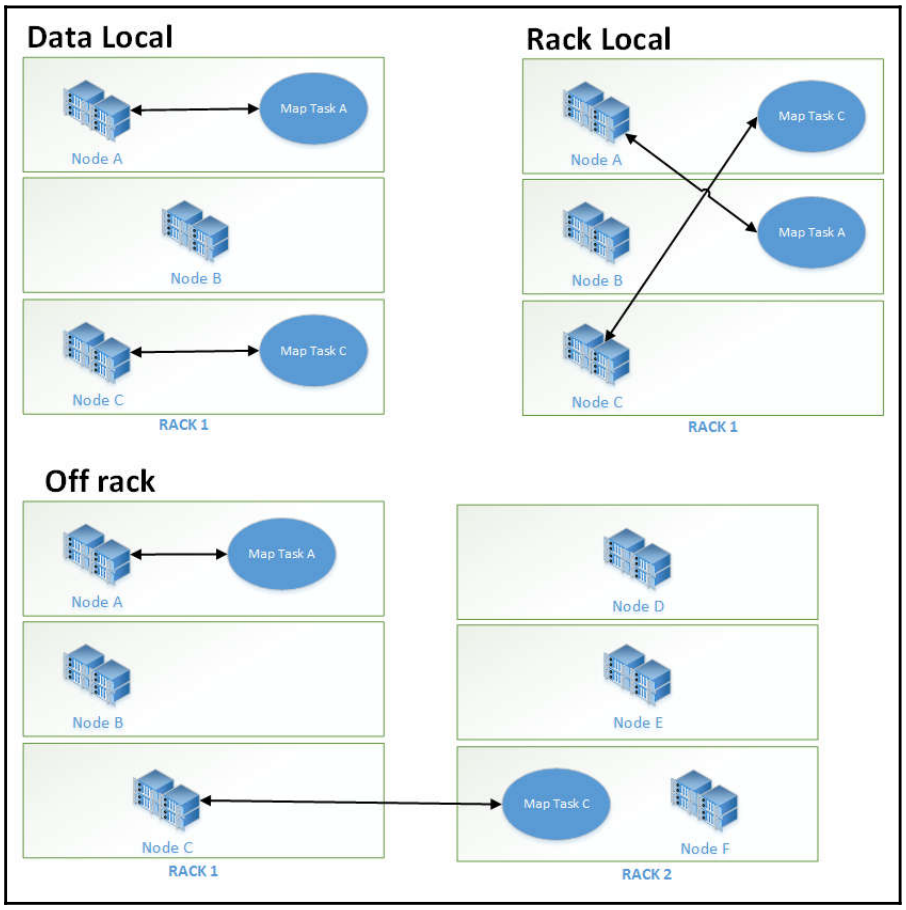


Chapter 2:
Deep Dive into the Hadoop Distributed File System




RPC Over TCP/IP
Client Protocol
Namenode
Application Code
HDFS Client (JAVA API)
Buffer
RPC Over TCP/IP
Data Node Protocol
Stream Over TCP/IP
Data Transfer Protocol
Datanode

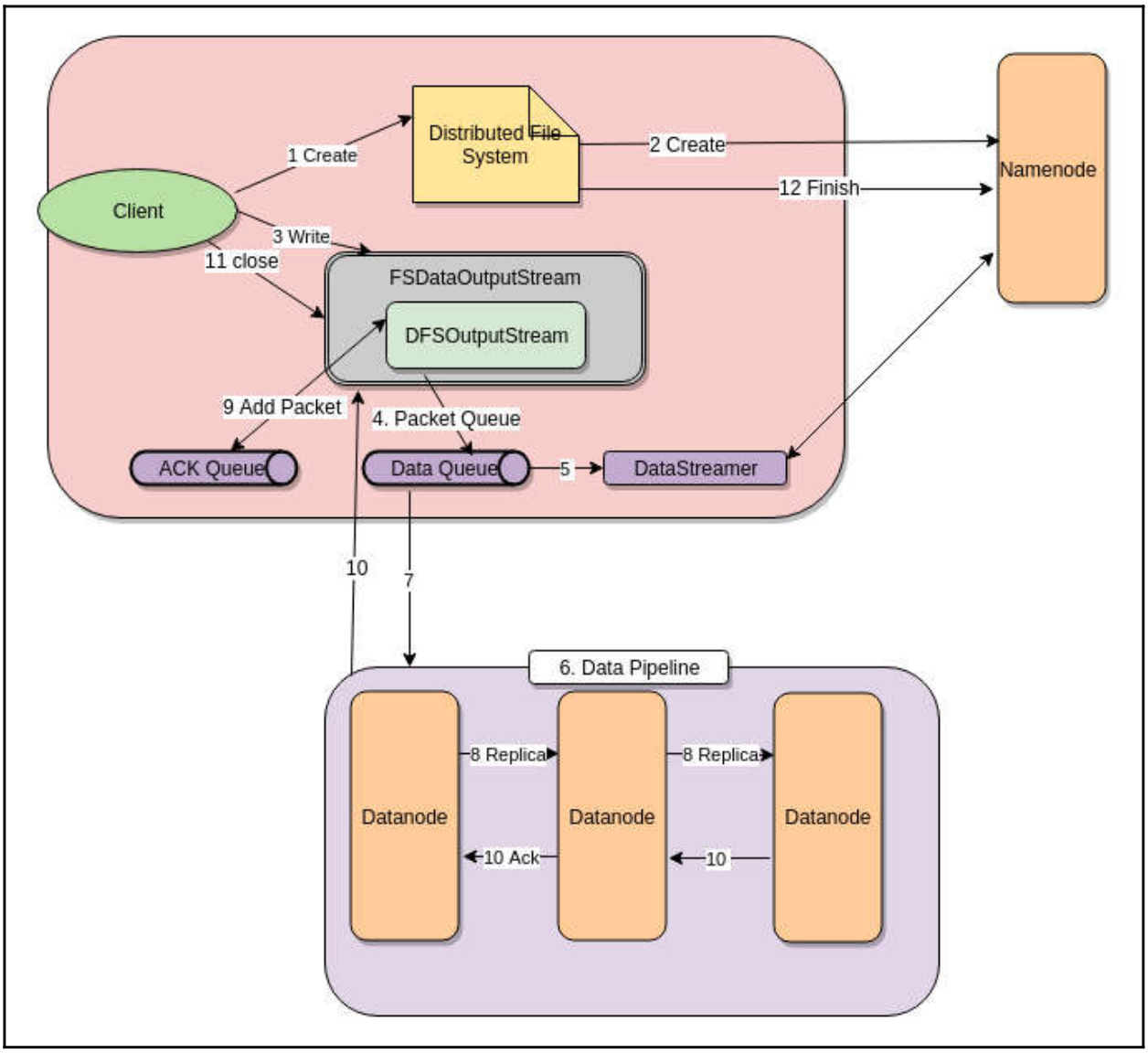

INodeDirectoryAttributes
INode
<<Interface>>
INodeAttributes
INodeDirectory
INodeWithAdditionalFields
<<Interface>>
LinkedElement
INodeFile
INodeFileAttributes

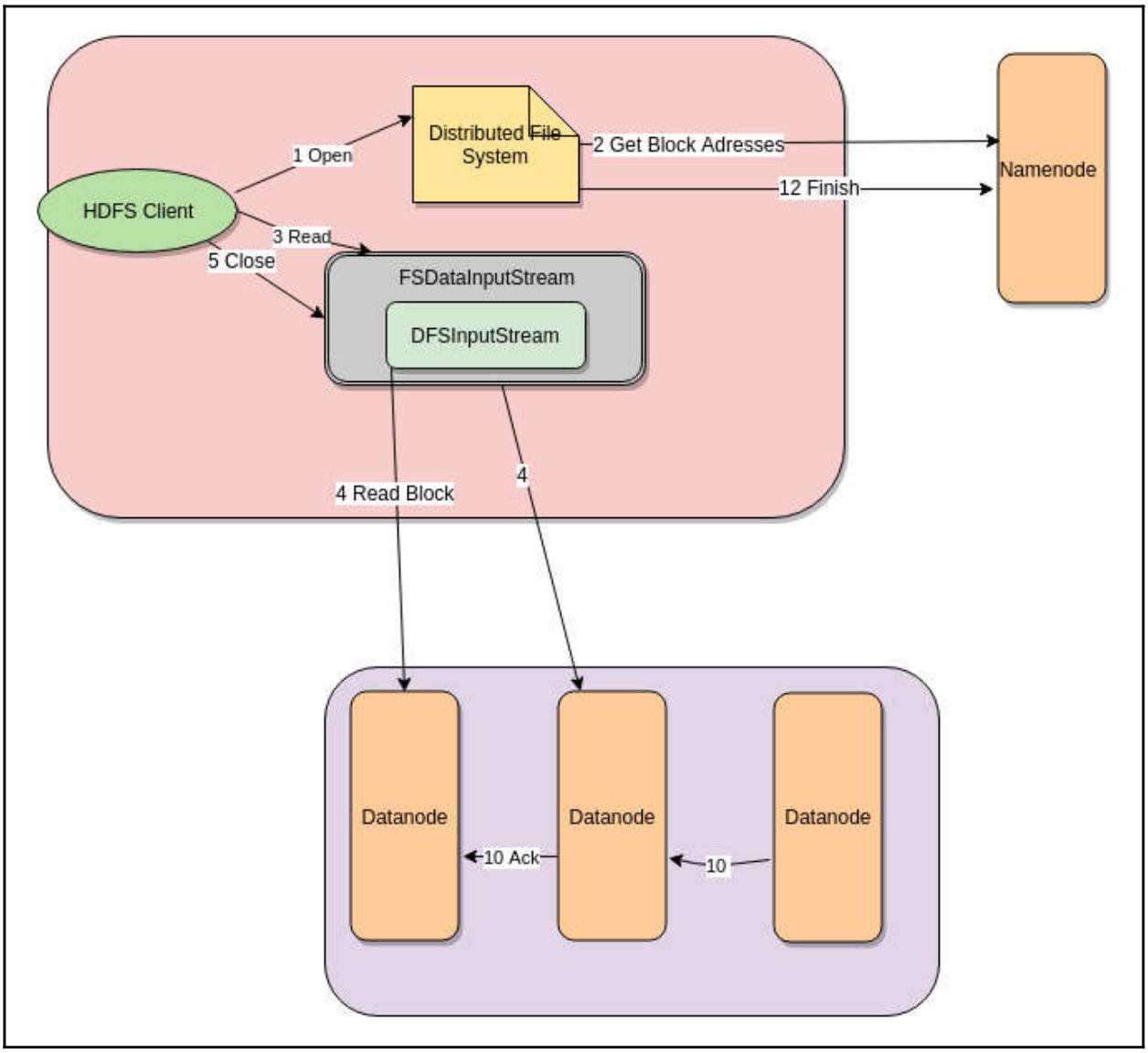

Data Local
Map Task A
Node A
Node B
Map Task C
Node C
RACK 1
Rack Local
Map Task C
Node A
Map Task A
Node B
Node C
RACK 1
Off rack
Map Task A
Node A
Node B
Node C
RACK 1
Node D
Node E
Map Task C
Node F
RACK 2

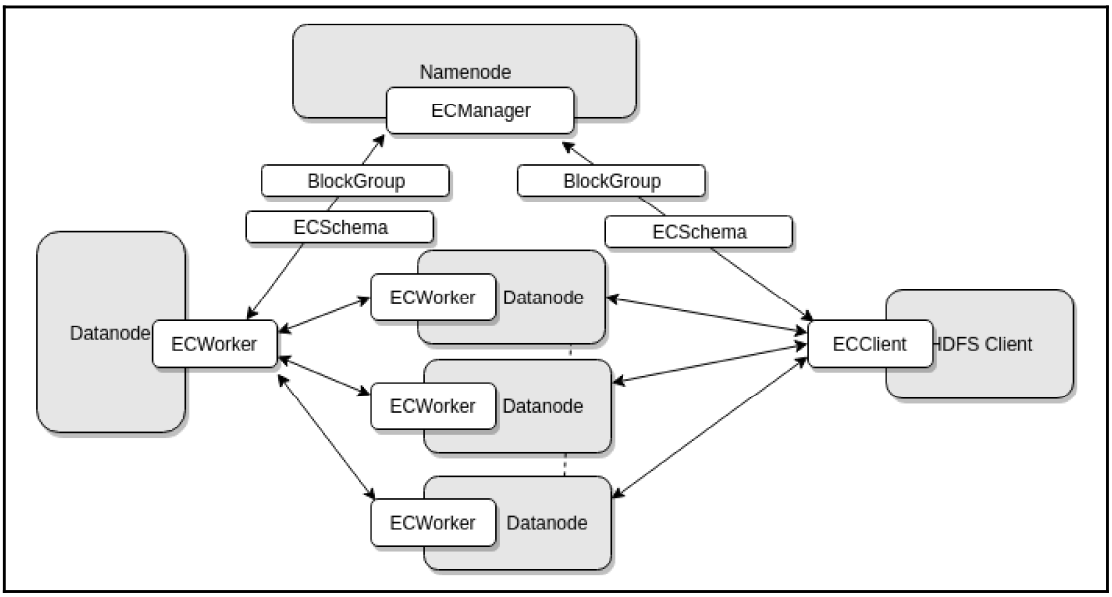
```
chanchal@chanchal-Lenovo-ideapad-510-15IKB:~$ hdfs oiv --help
Error parsing command-line options:
Usage: bin/hdfs oiv [OPTIONS] -i INPUTFILE -o OUTPUTFILE
Offline Image Viewer
View a Hadoop fsimage INPUTFILE using the specified PROCESSOR,
saving the results in OUTPUTFILE.

The oiv utility will attempt to parse correctly formed image files
and will abort fail with mal-formed image files.

The tool works offline and does not require a running cluster in
order to process an image file.

The following image processors are available:
  * XML: This processor creates an XML document with all elements of
    the fsimage enumerated, suitable for further analysis by XML
    tools.
  * FileDistribution: This processor analyzes the file size
    distribution in the image.
    -maxSize specifies the range [0, maxSize] of file sizes to be
     analyzed (128GB by default).
    -step defines the granularity of the distribution. (2MB by default)
  * Web: Run a viewer to expose read-only WebHDFS API.
    -addr specifies the address to listen. (localhost:5978 by default)
  * Delimited (experimental): Generate a text file with all of the elements common
    to both inodes and inodes-under-construction, separated by a
    delimiter. The default delimiter is \t, though this may be
    changed via the -delimiter argument.

Required command line arguments:
-i,--inputFile <arg>   FSImage file to process.

Optional command line arguments:
-o,--outputFile <arg>  Name of output file. If the specified
                       file exists, it will be overwritten.
                       (output to stdout by default)
-p,--processor <arg>   Select which type of processor to apply
                       against image file. (XML|FileDistribution|Web|Delimited)
                       (Web by default)
-delimiter <arg>       Delimiting string to use with Delimited processor.
-t,--temp <arg>        Use temporary dir to cache intermediate result to generate
                       Delimited outputs. If not set, Delimited processor constructs
                       the namespace in memory before outputting text.
```

```xml
<?xml version="1.0" encoding="UTF-8"?>
<EDITS>
  <EDITS_VERSION>-63</EDITS_VERSION>
  <RECORD>
    <OPCODE>OP_START_LOG_SEGMENT</OPCODE>
    <DATA>
      <TXID>488053</TXID>
    </DATA>
  </RECORD>
  <RECORD>
    <OPCODE>OP_MKDIR</OPCODE>
    <DATA>
      <TXID>488054</TXID>
      <LENGTH>0</LENGTH>
      <INODEID>190335</INODEID>
      <PATH>/tmp/hive/hive/124dd7e2-d4d3-413e-838e-3dbbbd185a69</PATH>
      <TIMESTAMP>1509663411129</TIMESTAMP>
      <PERMISSION_STATUS>
        <USERNAME>hive</USERNAME>
        <GROUPNAME>hdfs</GROUPNAME>
        <MODE>448</MODE>
      </PERMISSION_STATUS>
    </DATA>
  </RECORD>
  <RECORD>
    <OPCODE>OP_ADD</OPCODE>
    <DATA>
      <TXID>488055</TXID>
      <LENGTH>0</LENGTH>
      <INODEID>190336</INODEID>
      <PATH>/tmp/hive/hive/124dd7e2-d4d3-413e-838e-3dbbbd185a69/inuse.info</PATH>
      <REPLICATION>3</REPLICATION>
      <MTIME>1509663411169</MTIME>
      <ATIME>1509663411169</ATIME>
      <BLOCKSIZE>134217728</BLOCKSIZE>
      <CLIENT_NAME>DFSClient_NONMAPREDUCE_1006023362_1</CLIENT_NAME>
      <CLIENT_MACHINE>10.1.2.26</CLIENT_MACHINE>
      <OVERWRITE>true</OVERWRITE>
      <PERMISSION_STATUS>
        <USERNAME>hive</USERNAME>
        <GROUPNAME>hdfs</GROUPNAME>
        <MODE>420</MODE>
```

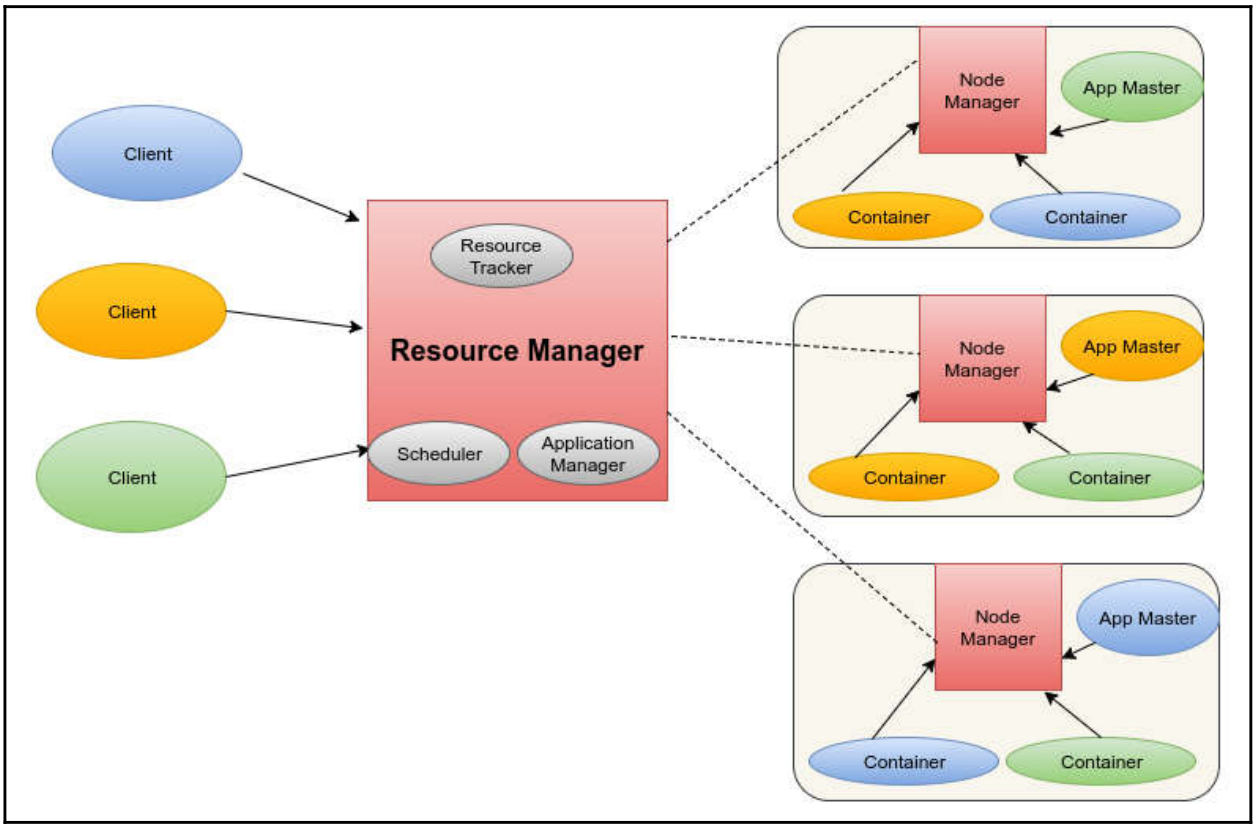
```
sshuser@hn0-apdc-s:~$ hdfs balancer --help
Usage: hdfs balancer
        [-policy <policy>]      the balancing policy: datanode or blockpool
        [-threshold <threshold>]        Percentage of disk capacity
        [-exclude [-f <hosts-file> | <comma-separated list of hosts>]]  Excludes the specified datanodes.
        [-include [-f <hosts-file> | <comma-separated list of hosts>]]  Includes only the specified datanodes.
        [-source [-f <hosts-file> | <comma-separated list of hosts>]]   Pick only the specified datanodes as source nodes.
        [-idleiterations <idleiterations>]      Number of consecutive idle iterations (-1 for Infinite) before exit.
        [-runDuringUpgrade]     Whether to run the balancer during an ongoing HDFS upgrade.This is usually not desired since it will not
 used space on over-utilized machines.

Generic options supported are
-conf <configuration file>     specify an application configuration file
-D <property=value>            use value for given property
-fs <local|namenode:port>      specify a namenode
-jt <local|resourcemanager:port>    specify a ResourceManager
-files <comma separated list of files>    specify comma separated files to be copied to the map reduce cluster
-libjars <comma separated list of jars>    specify comma separated jar files to include in the classpath.
-archives <comma separated list of archives>    specify comma separated archives to be unarchived on the compute machines.

The general command line syntax is
bin/hadoop command [genericOptions] [commandOptions]
```

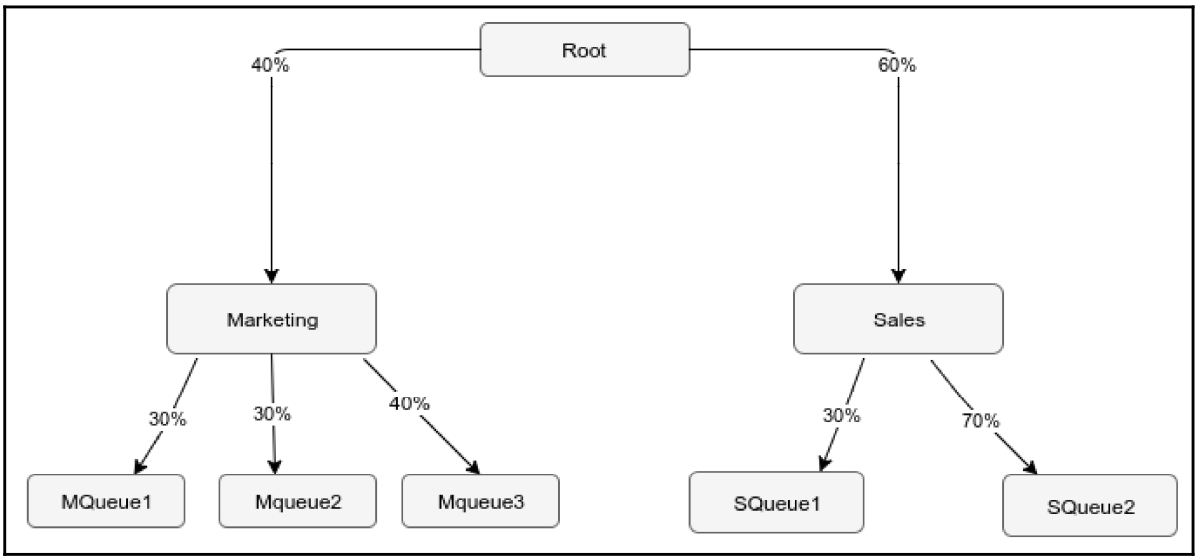

Client
1 Create
Distributed File System
2 Create
Namenode
12 Finish
3 Write
11 close
FSDataOutputStream
DFSOutputStream
9 Add Packet
4. Packet Queue
ACK Queue
Data Queue
5
DataStreamer
10
7
6. Data Pipeline
Datanode
8 Replica
Datanode
8 Replica
Datanode
10 Ack
10

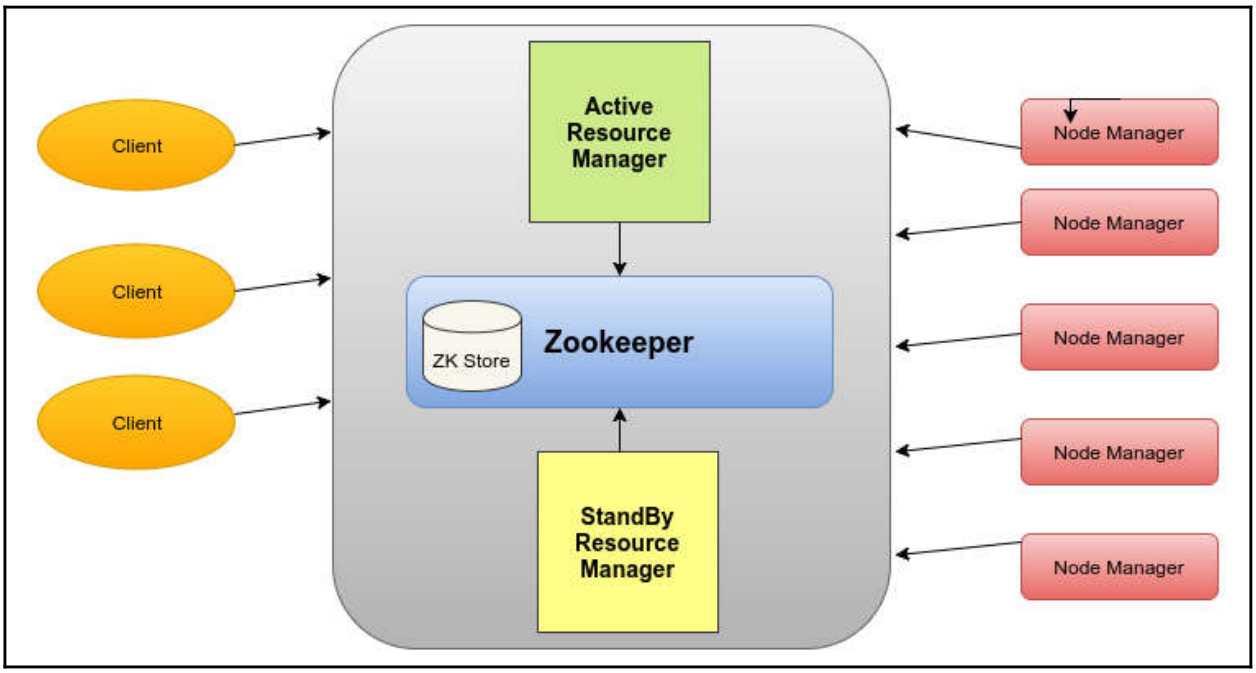

Distributed File System
1 Open
2 Get Block Adresses
12 Finish
Namenode
HDFS Client
3 Read
5 Close
FSDataInputStream
DFSInputStream
4 Read Block
4
Datanode
Datanode
Datanode
10 Ack
10


Namenode
ECManager
BlockGroup
BlockGroup
ECSchema
ECSchema
Datanode
ECWorker
ECWorker
Datanode
ECClient
HDFS Client
ECWorker
Datanode
ECWorker
Datanode

Chapter 3:
YARN Resource Management in Hadoop



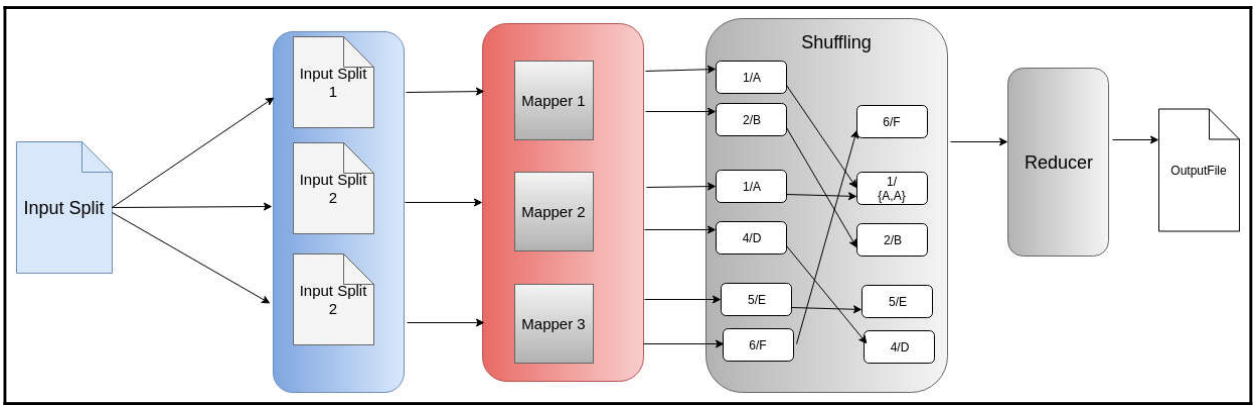

Root
40%
60%
Marketing
Sales
30%
30%
40%
30%
70%
MQueue1
Mqueue2
Mqueue3
SQueue1
SQueue2

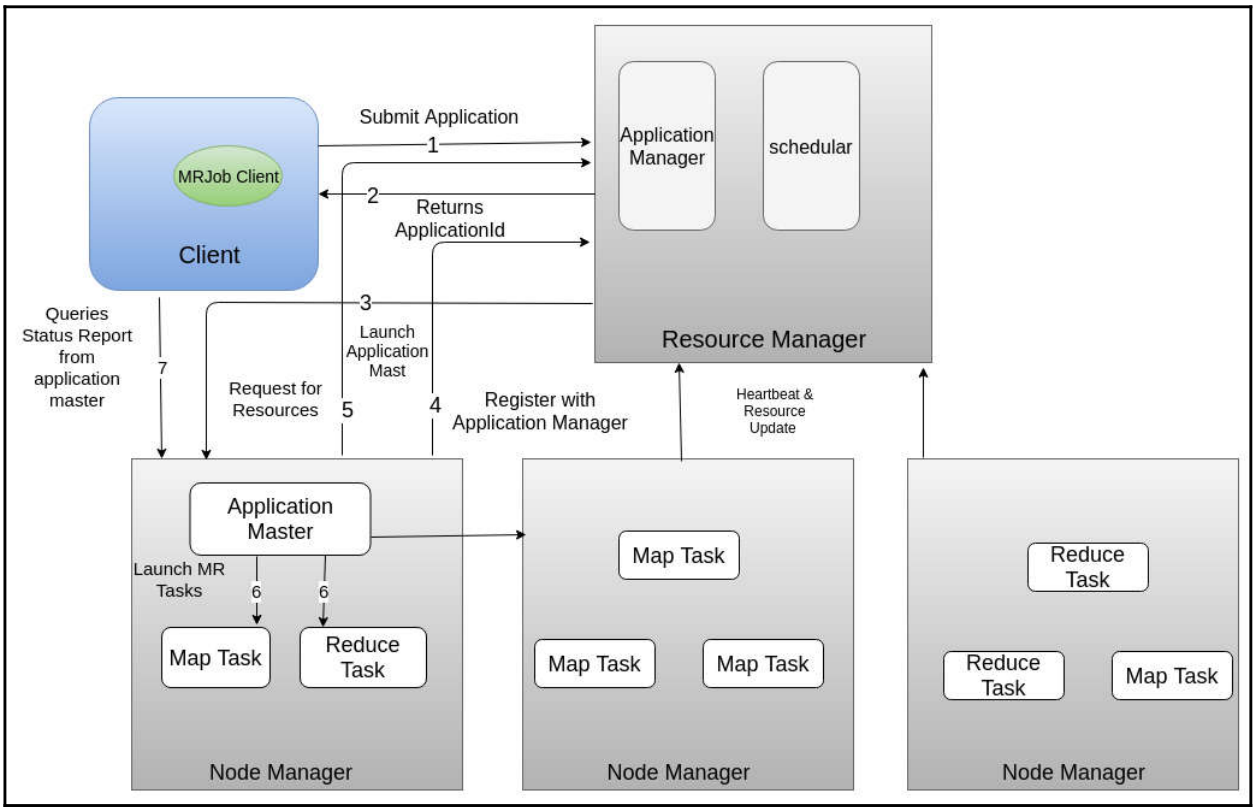

Client
Client
Client
Active Resource Manager
ZK Store
Zookeeper
StandBy Resource Manager
Node Manager
Node Manager
Node Manager
Node Manager
Node Manager

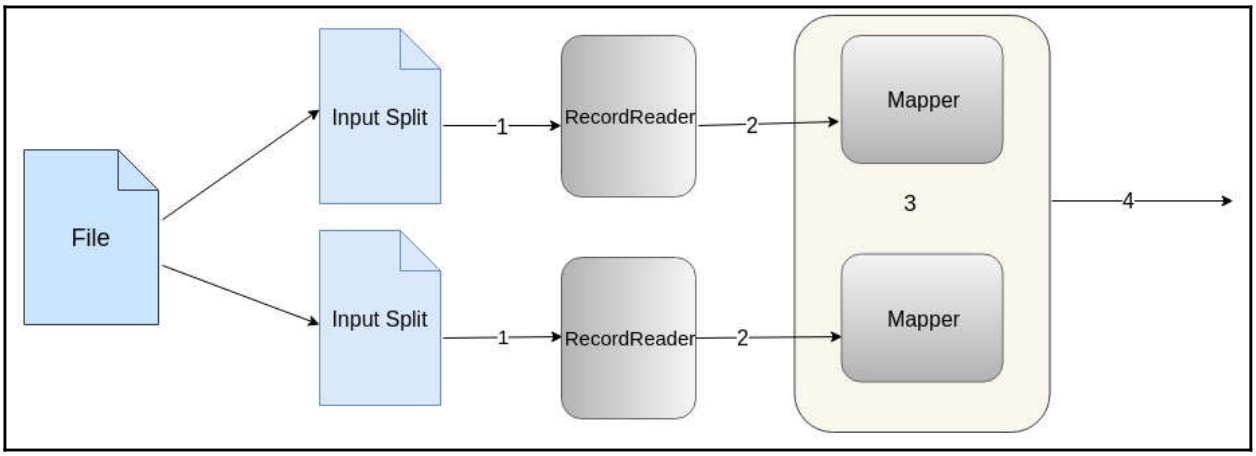
```
[root@ip-10-254-0-45 hadoop]# yarn application -list -appStates failed
18/01/18 08:08:03 INFO impl.TimelineClientImpl: Timeline service address: http://ip-10-254-0-45.ap-south-1.compute.internal:8188/ws/v1/timelin
18/01/18 08:08:03 INFO client.RMProxy: Connecting to ResourceManager at ip-10-254-0-45.ap-south-1.compute.internal/10.254.0.45:8032
Total number of applications (application-types: [] and states: [FAILED]):5
                Application-Id      Application-Name        Application-Type          User           Queue                   State
inal-State            Progress                        Tracking-URL
application_1513582536692_0011  HIVE-dcbc8001-3339-470e-a303-b367d17fa83c                       TEZ        hadoop         default
    FAILED                  FAILED                  0% http://ip-10-254-0-45.ap-south-1.compute.internal:8088/cluster/app/application_1513582
6692_0011
application_1513582536692_0009  select count(*) from tests3...service_region(Stage-1)             MAPREDUCE        hadoop         default
          FAILED                    FAILED                   0% http://ip-10-254-0-45.ap-south-1.compute.internal:8088/cluster/app/application
513582536692_0009
application_1513582536692_0010  HIVE-9a9368ef-6e76-4ddb-8d72-e8b0dccaa534                       TEZ        hadoop         default
    FAILED                  FAILED                  0% http://ip-10-254-0-45.ap-south-1.compute.internal:8088/cluster/app/application_1513582
6692_0010
application_1513582536692_0007  select sum(transaction_amou...service_region(Stage-1)             MAPREDUCE        hadoop         default
          FAILED                    FAILED                   0% http://ip-10-254-0-45.ap-south-1.compute.internal:8088/cluster/app/application
513582536692_0007
application_1513582536692_0008  select sum(transaction_amount) from tests3(Stage-1)               MAPREDUCE        hadoop         default
          FAILED                    FAILED                   0% http://ip-10-254-0-45.ap-south-1.compute.internal:8088/cluster/app/application
513582536692_0008
[root@ip-10-254-0-45 hadoop]#
```

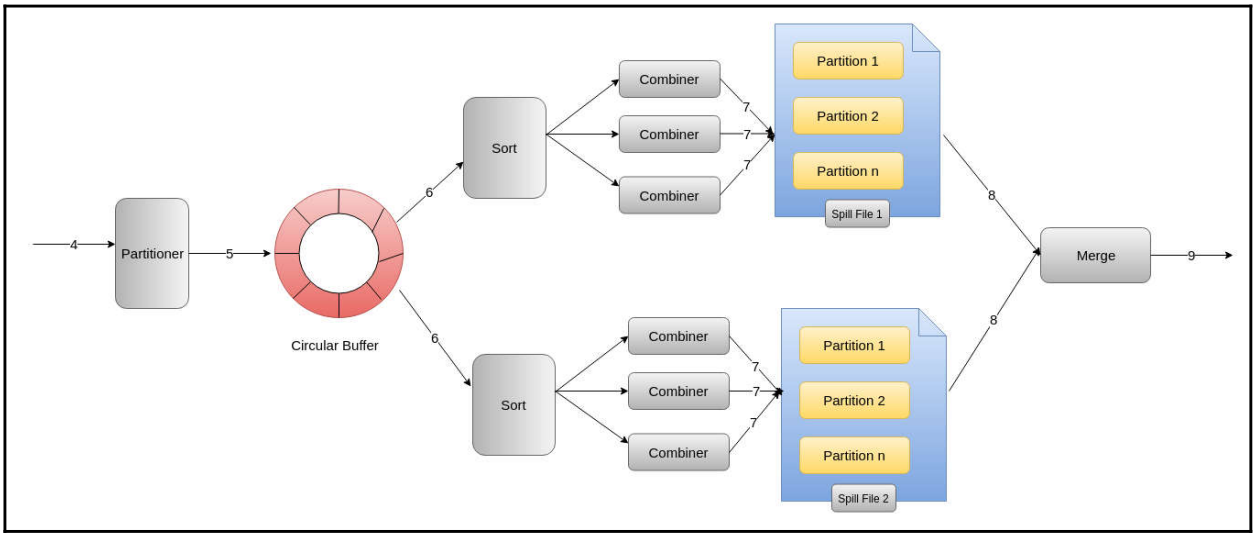
```
[root@ip-10-254-0-45 hadoop]# yarn application -status application_1513582536692_0119
18/01/18 08:20:26 INFO impl.TimelineClientImpl: Timeline service address: http://ip-10-254-0-45.ap-south-1.compute.internal:8188/ws/v1/timeline/
18/01/18 08:20:26 INFO client.RMProxy: Connecting to ResourceManager at ip-10-254-0-45.ap-south-1.compute.internal/10.254.0.45:8032
Application Report :
        Application-Id : application_1513582536692_0119
        Application-Name : HIVE-a34d1844-cdb8-4691-a289-8a53a55f9dec
        Application-Type : TEZ
        User : root
        Queue : default
        Start-Time : 1515995486809
        Finish-Time : 1515995488561
        Progress : 100%
        State : KILLED
        Final-State : KILLED
        Tracking-URL : http://ip-10-254-0-45.ap-south-1.compute.internal:8088/cluster/app/application_1513582536692_0119
        RPC Port : -1
        AM Host : N/A
        Aggregate Resource Allocation : 1337 MB-seconds, 1 vcore-seconds
        Diagnostics : Application killed by user.
[root@ip-10-254-0-45 hadoop]#
```

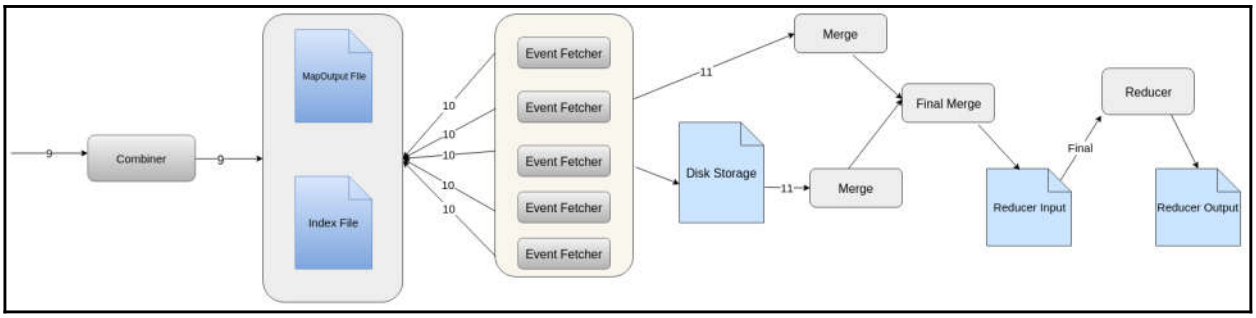
Chapter 4: Internals of MapReduce



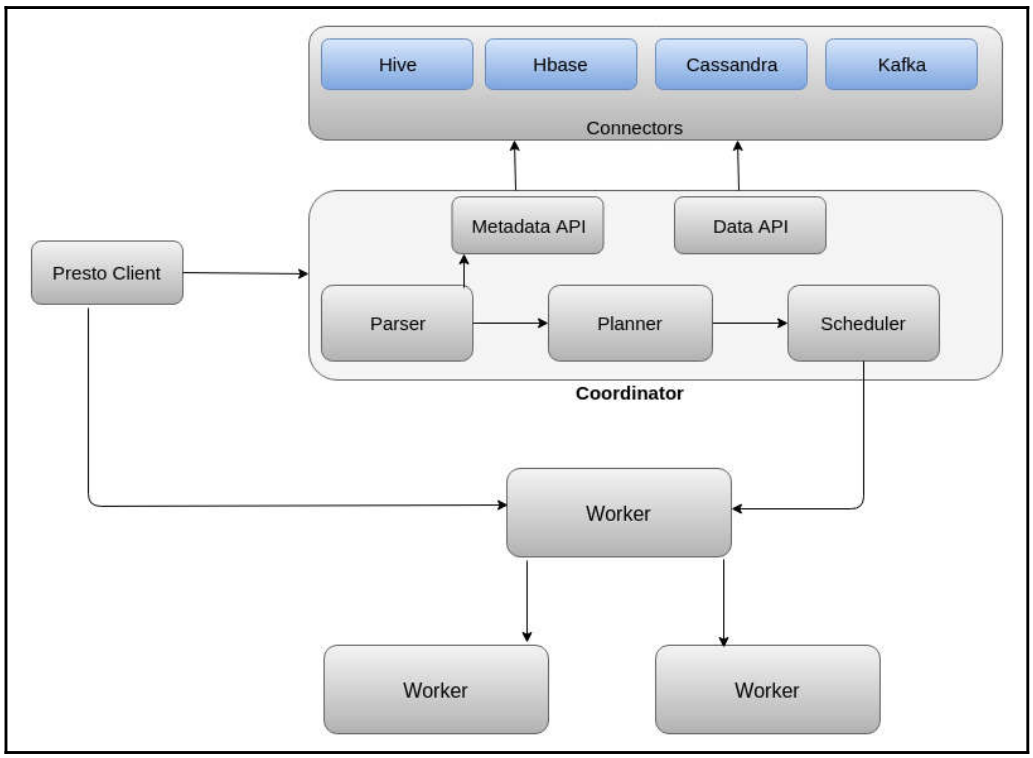


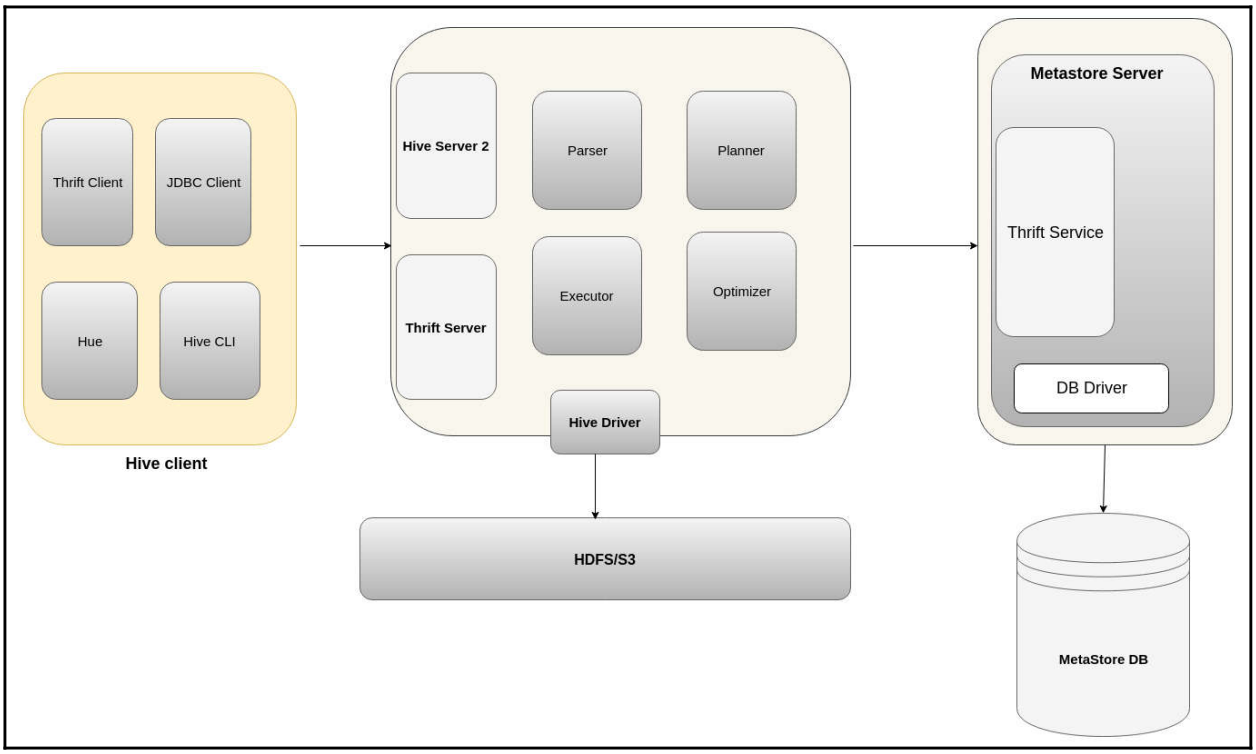

File
Input Split
Input Split
1
1
RecordReader
RecordReader
2
2
Mapper
3
Mapper
4


4
Partitioner
5
Circular Buffer
6
6
Sort
Sort
Combiner
Combiner
Combiner
Combiner
Combiner
Combiner
7
7
7
7
7
7
Partition 1
Partition 2
Partition n
Spill File 1
Partition 1
Partition 2
Partition n
Spill File 2
8
8
Merge
9


9
Combiner
9
MapOutput File
Index File
10
10
10
10
10
Event Fetcher
Event Fetcher
Event Fetcher
Event Fetcher
Event Fetcher
11
Merge
Disk Storage
11
Merge
Final Merge
Reducer
Final
Reducer Input
Reducer Output

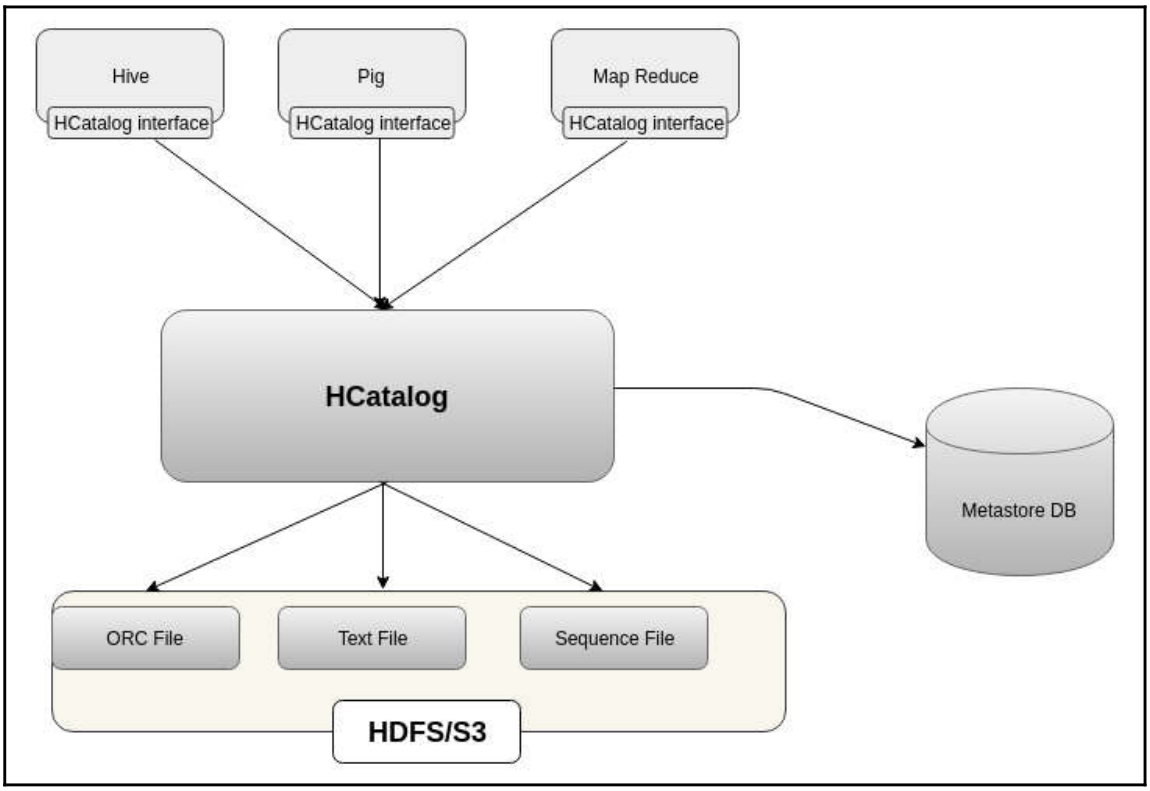
Chapter 5: SQL on Hadoop




Thrift Client
JDBC Client
Hue
Hive CLI
Hive client
Hive Server 2
Parser
Planner
Thrift Server
Executor
Optimizer
Hive Driver
HDFS/S3
Metastore Server
Thrift Service
DB Driver
MetaStore DB

```
hive> describe formatted product;
OK
# col_name              data_type               comment

product_id              int
product_name            string
product_price           double
manufacturer            string

# Detailed Table Information
Database:               default
Owner:                  hadoop
CreateTime:             Mon May 21 08:40:53 UTC 2018
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://ip-10-254-0-45.ap-south-1.compute.internal:8020/user/hive/warehouse/product
Table Type:             EXTERNAL_TABLE
Table Parameters:
        EXTERNAL                TRUE
        numFiles                1
        numRows                 1
        rawDataSize             22
        totalSize               27
        transient_lastDdlTime   1526894412

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        field.delim             \t
        line.delim              \n
        serialization.format    \t
```

```
hive> explain  Select count(distinct product_id) , manufacturer from product group by manufacturer;
OK
Plan optimized by CBO.

Vertex dependency in root stage
Reducer 2 <- Map 1 (SIMPLE_EDGE)

Stage-0
  Fetch Operator
    limit:-1
    Stage-1
      Reducer 2
      File Output Operator [FS_11]
        Select Operator [SEL_10] (rows=1 width=22)
          Output:["_col0","_col1"]
          Group By Operator [GBY_9] (rows=1 width=22)
            Output:["_col0","_col1"],aggregations:["count(_col0)"],keys:_col1
            Select Operator [SEL_5] (rows=1 width=22)
              Output:["_col0","_col1"]
              Group By Operator [GBY_4] (rows=1 width=22)
                Output:["_col0","_col1"],keys:KEY._col0, KEY._col1
              <-Map 1 [SIMPLE_EDGE]
                SHUFFLE [RS_3]
                  PartitionCols:_col0
                  Group By Operator [GBY_2] (rows=1 width=22)
                    Output:["_col0","_col1"],keys:manufacturer, product_id
                    Select Operator [SEL_1] (rows=1 width=22)
                      Output:["manufacturer","product_id"]
                      TableScan [TS_0] (rows=1 width=22)
                        default@product,product,Tbl:COMPLETE,Col:NONE,Output:["product_id","manufacturer"]

Time taken: 0.108 seconds, Fetched: 29 row(s)
```




```
ravishankarnair — -bash — 80×7
ravion:~ ravishankarnair$ hive --version
Hive 2.3.2
Git git://stakiar-MBP.local/Users/stakiar/Desktop/scratch-space/apache-hive -r 8
57a9fd8ad725a53bd95c1b2d6612f9b1155f44d
Compiled by stakiar on Thu Nov 9 09:11:39 PST 2017
From source with checksum dc38920061a4eb32c4d15ebd5429ac8a
ravion:~ ravishankarnair$
```

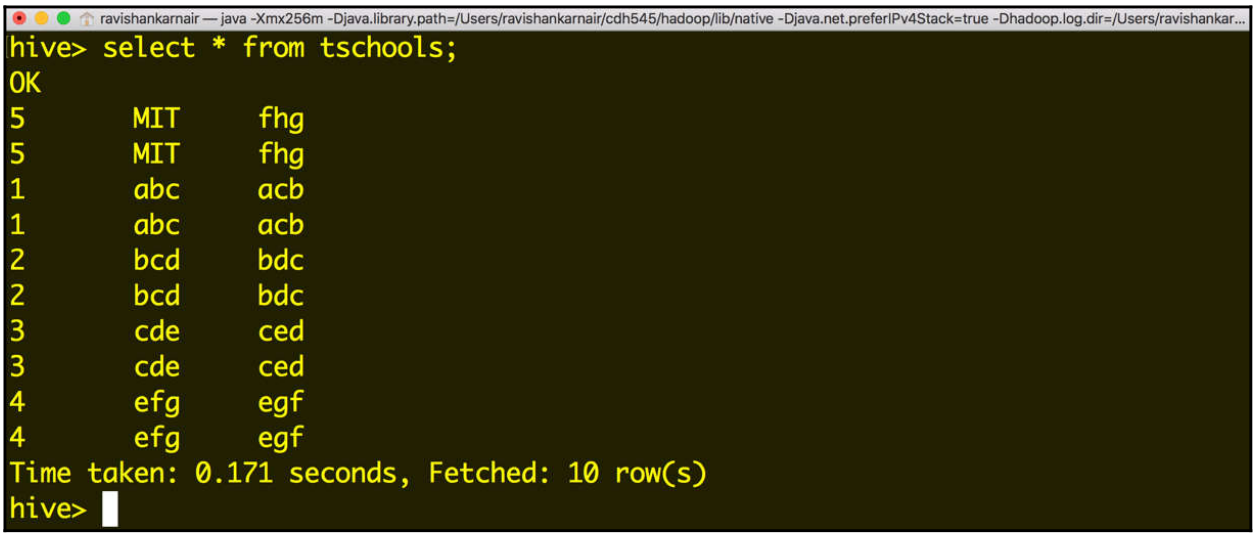
```
ravishankarnair — java -Xmx256m -Djava.library.path=/Users/ravishankarnair/cdh545/hadoop/lib/native -Djava.net.preferIPv4Stack=true -Dhadoop.log.dir=/Users/ravishankar...

hive> set hive.support.concurrency = true;
hive> set hive.enforce.bucketing = true;
hive> set hive.exec.dynamic.partition.mode = nonstrict;
hive> set hive.txn.manager = org.apache.hadoop.hive.ql.lockmgr.DbTxnManager;
hive> set hive.compactor.initiator.on = true;
hive> set hive.compactor.worker.threads = 1;
hive>
```

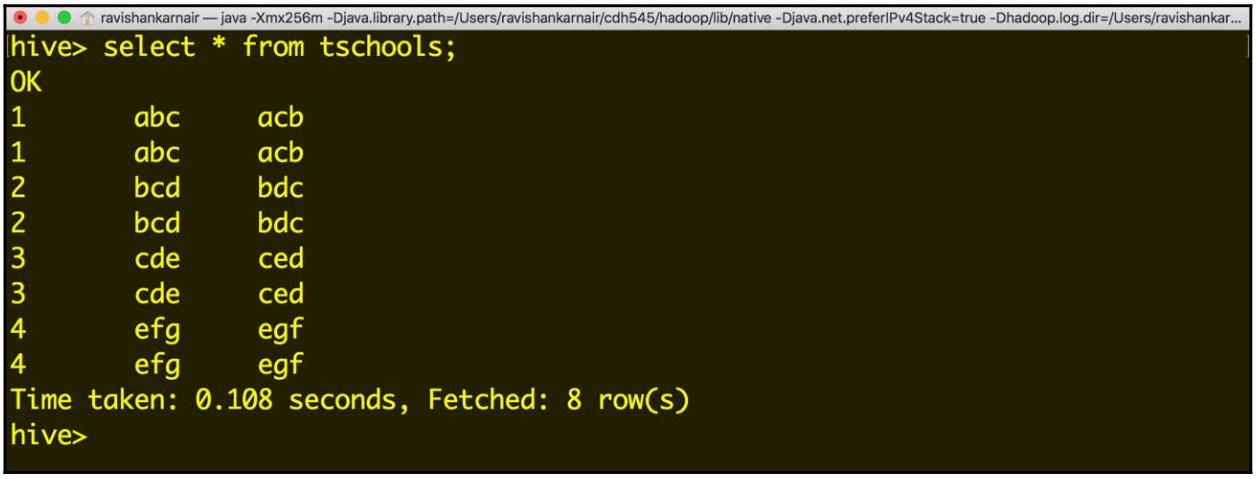
```
ravishankarnair — java -Xmx256m -Djava.library.path=/Users/ravishankarnair/cdh545/hadoop/lib/native -Djava.net.preferIPv4Stack=true -Dhadoop.log.dir=/Users/ravishankar...

hive> set hive.enforce.bucketing;
hive.enforce.bucketing=true
hive>
```

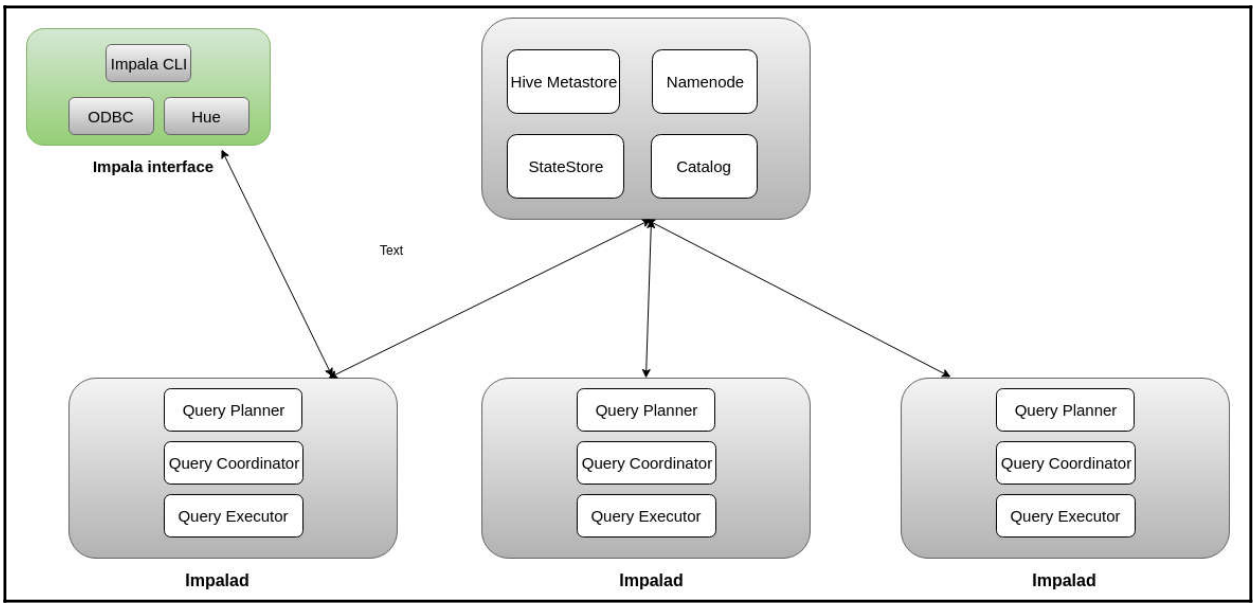
```
ravishankarnair — java -Xmx256m -Djava.library.path=/Users/ravishankarnair/cdh545/hadoop/lib/native -Djava.net.preferIPv4Stack=true -Dhadoop.log.dir=/Users/ravishankar...
hive> create table tschools(school_id int, school_name string, school_loc string
) clustered by (school_id) into 5 buckets stored as ORC TBLPROPERTIES ('transact
ional' = 'true');
OK
Time taken: 0.231 seconds
hive>
```

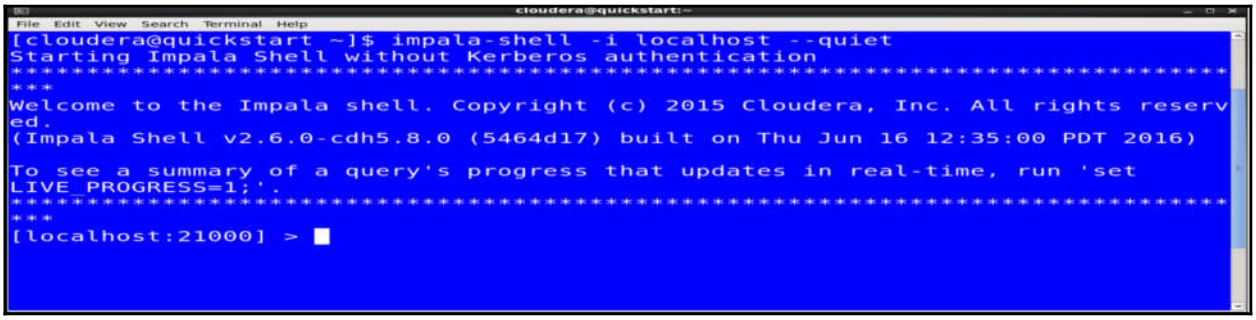
```
ravishankarnair — java -Xmx256m -Djava.library.path=/Users/ravishankarnair/cdh545/hadoop/lib/native -Djava.net.preferIPv4Stack=true -Dhadoop.log.dir=/Users/ravishankar...
hive> select * from tschools;
OK
5       fgh     fhg
5       fgh     fhg
1       abc     acb
1       abc     acb
2       bcd     bdc
2       bcd     bdc
3       cde     ced
3       cde     ced
4       efg     egf
4       efg     egf
Time taken: 0.298 seconds, Fetched: 10 row(s)
hive>
```

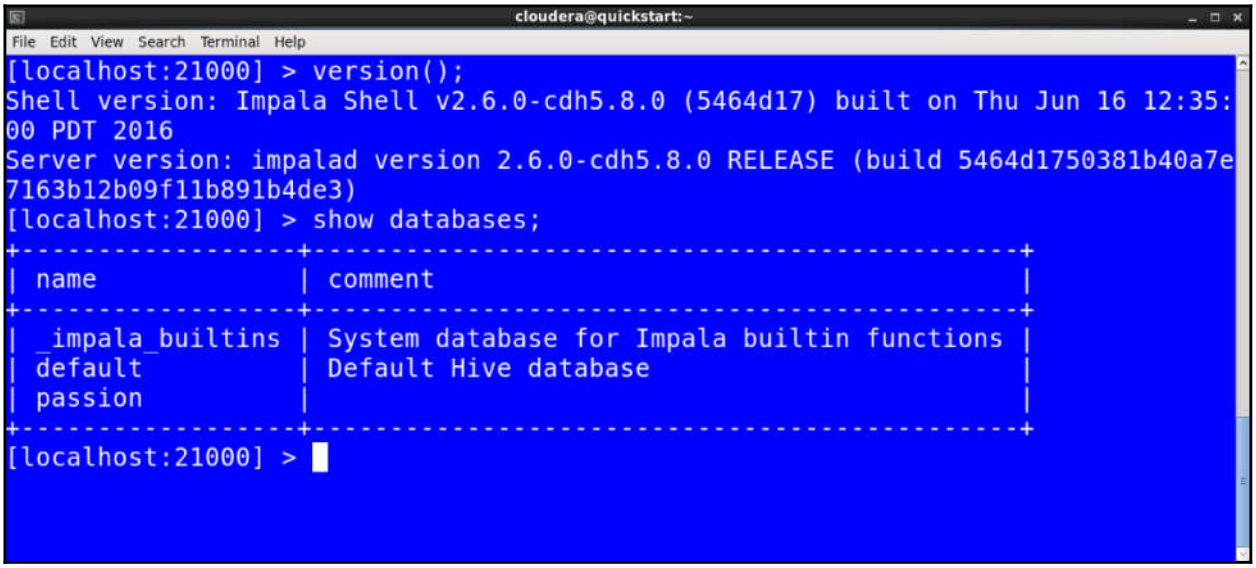
```
ravishankarnair — java -Xmx256m -Djava.library.path=/Users/ravishankarnair/cdh545/hadoop/lib/native -Djava.net.preferIPv4Stack=true -Dhadoop.log.dir=/Users/ravishankar...

hive> update tschools set school_id = 10 where school_id = 5;
FAILED: SemanticException [Error 10302]: Updating values of bucketing columns is
 not supported.  Column school_id.
hive>
```

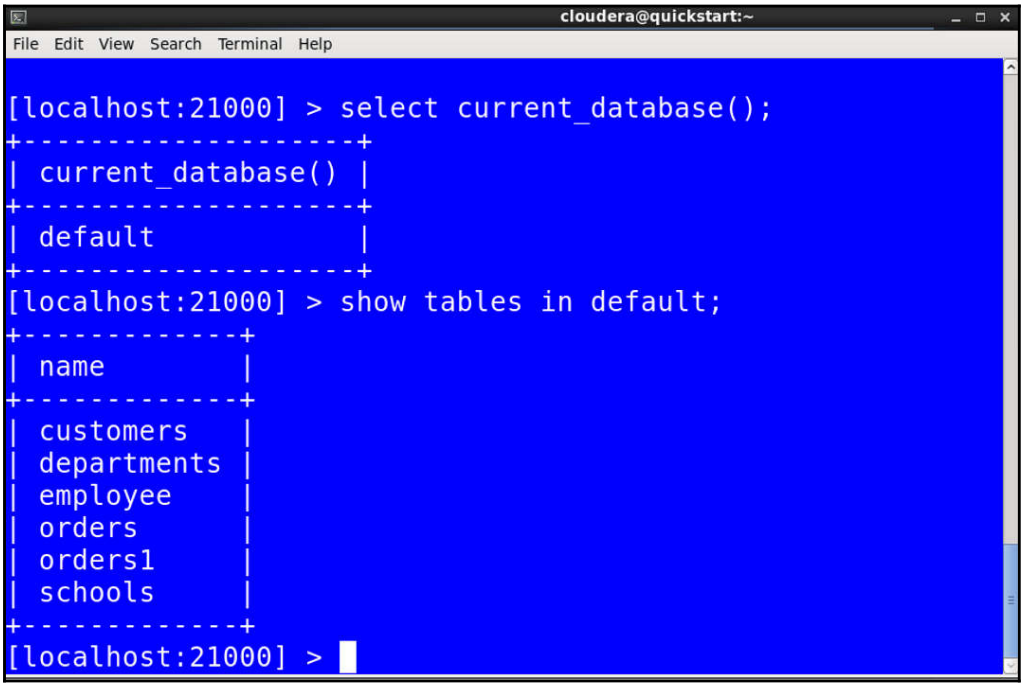

ravishankarnair — java -Xmx256m -Djava.library.path=/Users/ravishankarnair/cdh545/hadoop/lib/native -Djava.net.preferIPv4Stack=true -Dhadoop.log.dir=/Users/ravishankar...
hive> select * from tschools;
OK
5 MIT fhg
5 MIT fhg
1 abc acb
1 abc acb
2 bcd bdc
2 bcd bdc
3 cde ced
3 cde ced
4 efg egf
4 efg egf
Time taken: 0.171 seconds, Fetched: 10 row(s)
hive>

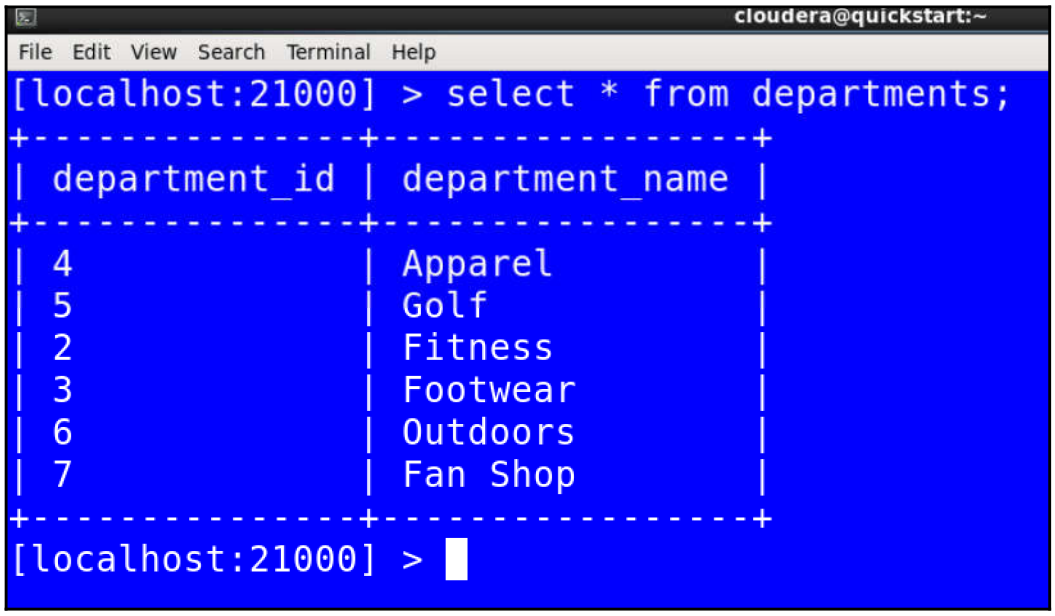

ravishankarnair — java -Xmx256m -Djava.library.path=/Users/ravishankarnair/cdh545/hadoop/lib/native -Djava.net.preferIPv4Stack=true -Dhadoop.log.dir=/Users/ravishankar...
hive> select * from tschools;
OK
1 abc acb
1 abc acb
2 bcd bdc
2 bcd bdc
3 cde ced
3 cde ced
4 efg egf
4 efg egf
Time taken: 0.108 seconds, Fetched: 8 row(s)
hive>

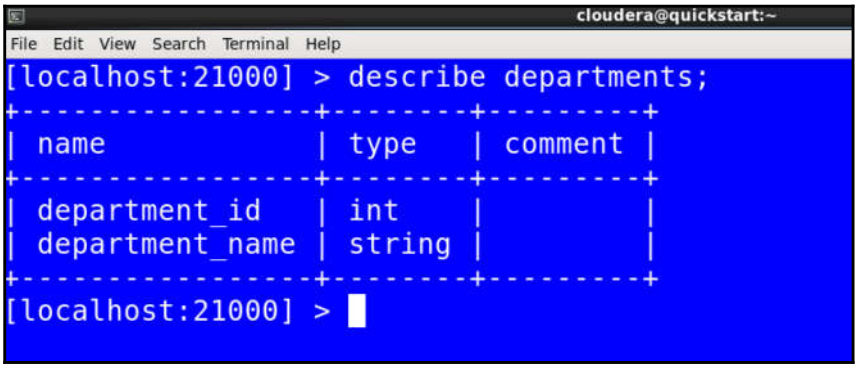

Impala CLI
ODBC
Hue
Impala interface
Hive Metastore
Namenode
StateStore
Catalog
Text
Query Planner
Query Coordinator
Query Executor
Impalad
Query Planner
Query Coordinator
Query Executor
Impalad
Query Planner
Query Coordinator
Query Executor
Impalad

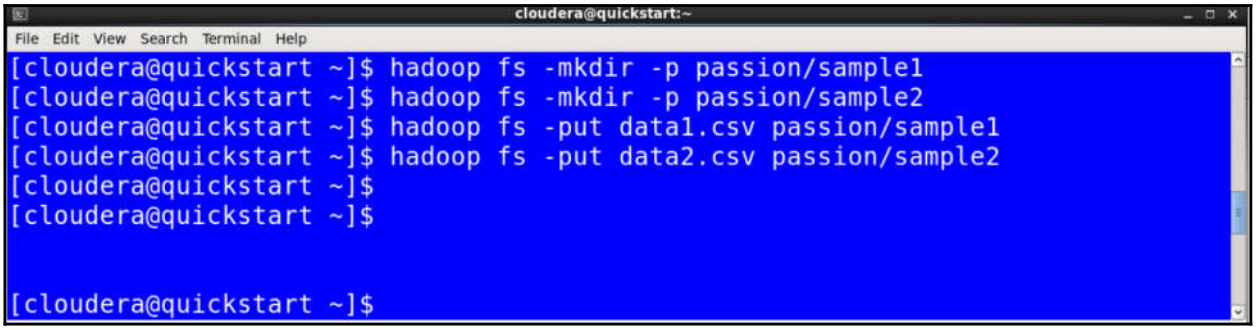

cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ impala-shell -i localhost --quiet
Starting Impala Shell without Kerberos authentication
Welcome to the Impala shell. Copyright (c) 2015 Cloudera, Inc. All rights reserved.
(Impala Shell v2.6.0-cdh5.8.0 (5464d17) built on Thu Jun 16 12:35:00 PDT 2016)
To see a summary of a query's progress that updates in real-time, run 'set LIVE_PROGRESS=1;'.
[localhost:21000] >

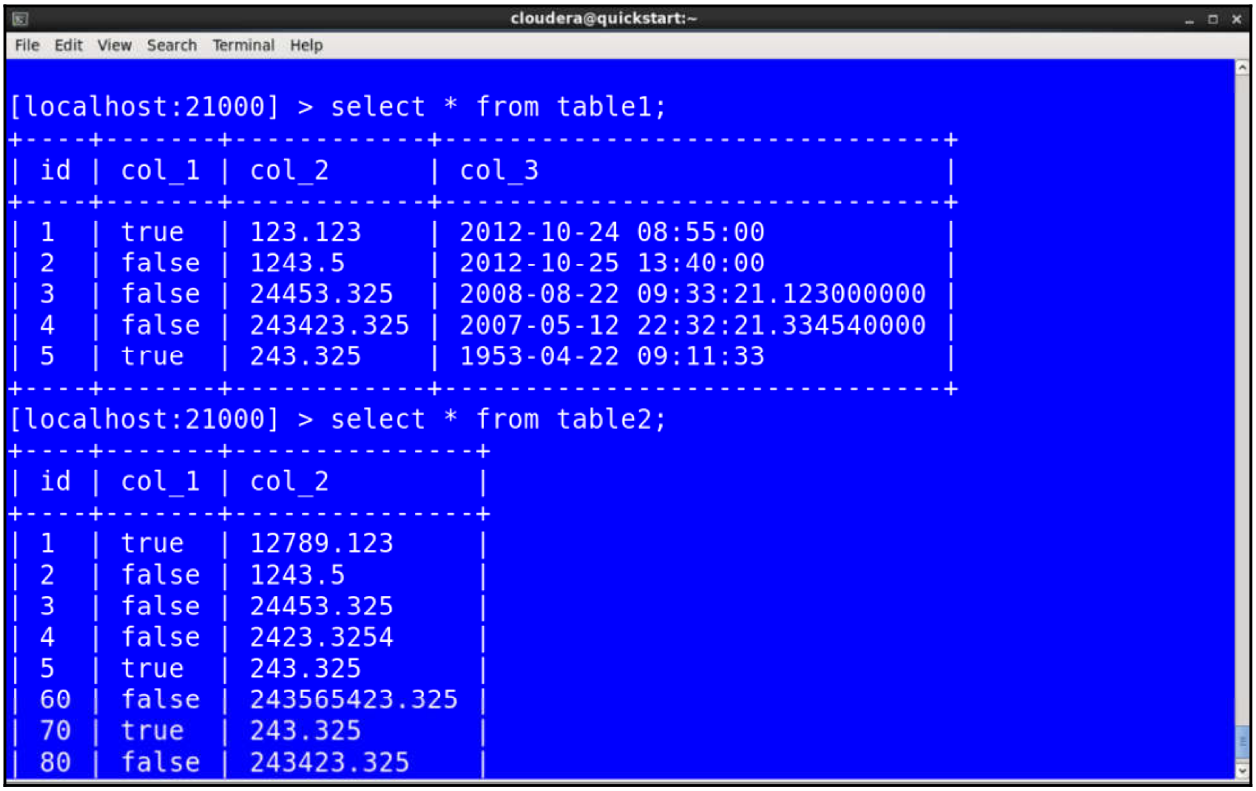

cloudera@quickstart:~
File Edit View Search Terminal Help
[localhost:21000] > version();
Shell version: Impala Shell v2.6.0-cdh5.8.0 (5464d17) built on Thu Jun 16 12:35:00 PDT 2016
Server version: impalad version 2.6.0-cdh5.8.0 RELEASE (build 5464d1750381b40a7e7163b12b09f11b891b4de3)
[localhost:21000] > show databases;
name	comment
_impala_builtins	System database for Impala builtin functions
default	Default Hive database
passion
[localhost:21000] >


cloudera@quickstart:~
File Edit View Search Terminal Help
[localhost:21000] > select current_database();
| current_database() |
| default |
[localhost:21000] > show tables in default;
| name |
| customers |
| departments |
| employee |
| orders |
| orders1 |
| schools |
[localhost:21000] >

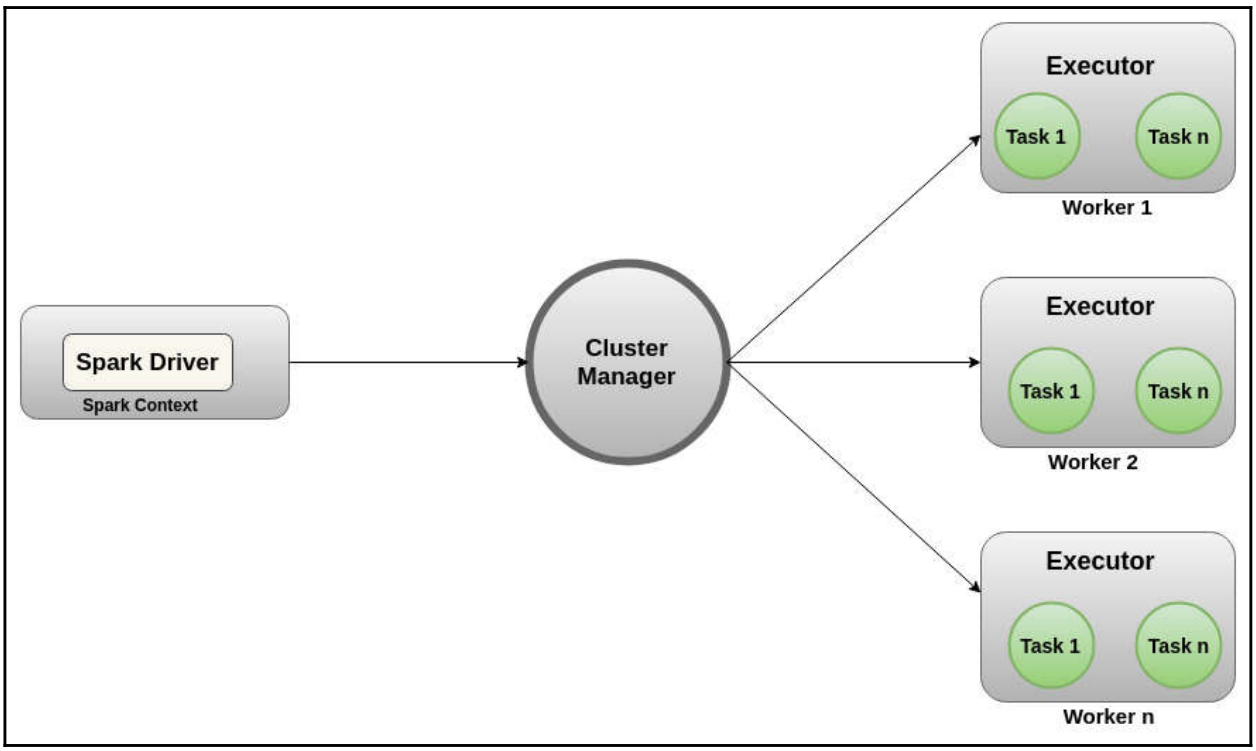

cloudera@quickstart:~
File Edit View Search Terminal Help
[localhost:21000] > select * from departments;
+---------------+-----------------+
| department_id | department_name |
+---------------+-----------------+
4	Apparel
5	Golf
2	Fitness
3	Footwear
6	Outdoors
7	Fan Shop
+---------------+-----------------+
[localhost:21000] >


cloudera@quickstart:~
File Edit View Search Terminal Help
[localhost:21000] > describe departments;
+-----------------+--------+---------+
| name | type | comment |
+-----------------+--------+---------+
| department_id | int | |
| department_name | string | |
+-----------------+--------+---------+
[localhost:21000] >

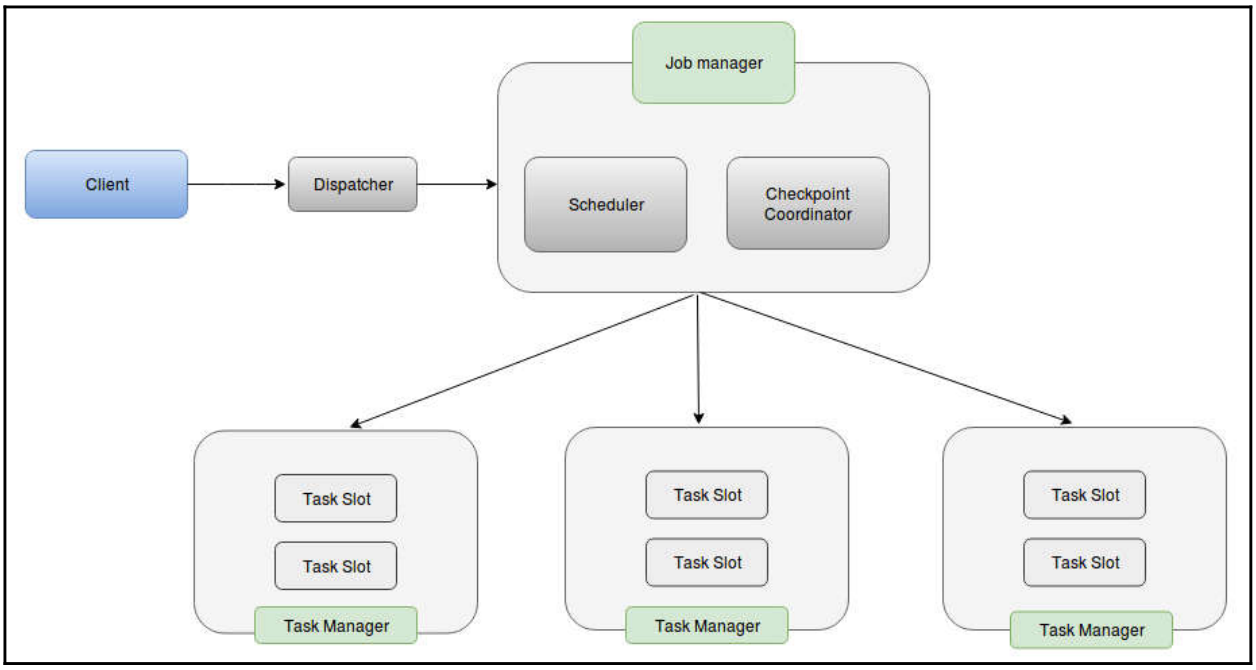

cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ hadoop fs -mkdir -p passion/sample1
[cloudera@quickstart ~]$ hadoop fs -mkdir -p passion/sample2
[cloudera@quickstart ~]$ hadoop fs -put data1.csv passion/sample1
[cloudera@quickstart ~]$ hadoop fs -put data2.csv passion/sample2
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$

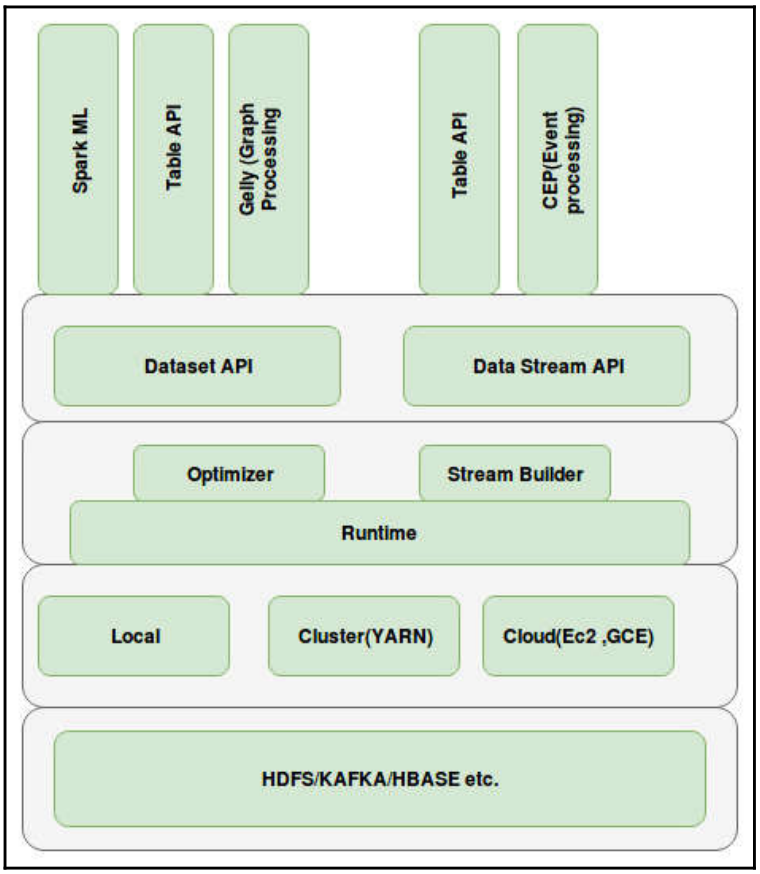

cloudera@quickstart:~
File Edit View Search Terminal Help
[localhost:21000] > select * from table1;
id	col_1	col_2	col_3
1	true	123.123	2012-10-24 08:55:00
2	false	1243.5	2012-10-25 13:40:00
3	false	24453.325	2008-08-22 09:33:21.123000000
4	false	243423.325	2007-05-12 22:32:21.334540000
5	true	243.325	1953-04-22 09:11:33
[localhost:21000] > select * from table2;
id	col_1	col_2
1	true	12789.123
2	false	1243.5
3	false	24453.325
4	false	2423.3254
5	true	243.325
60	false	243565423.325
70	true	243.325
80	false	243423.325

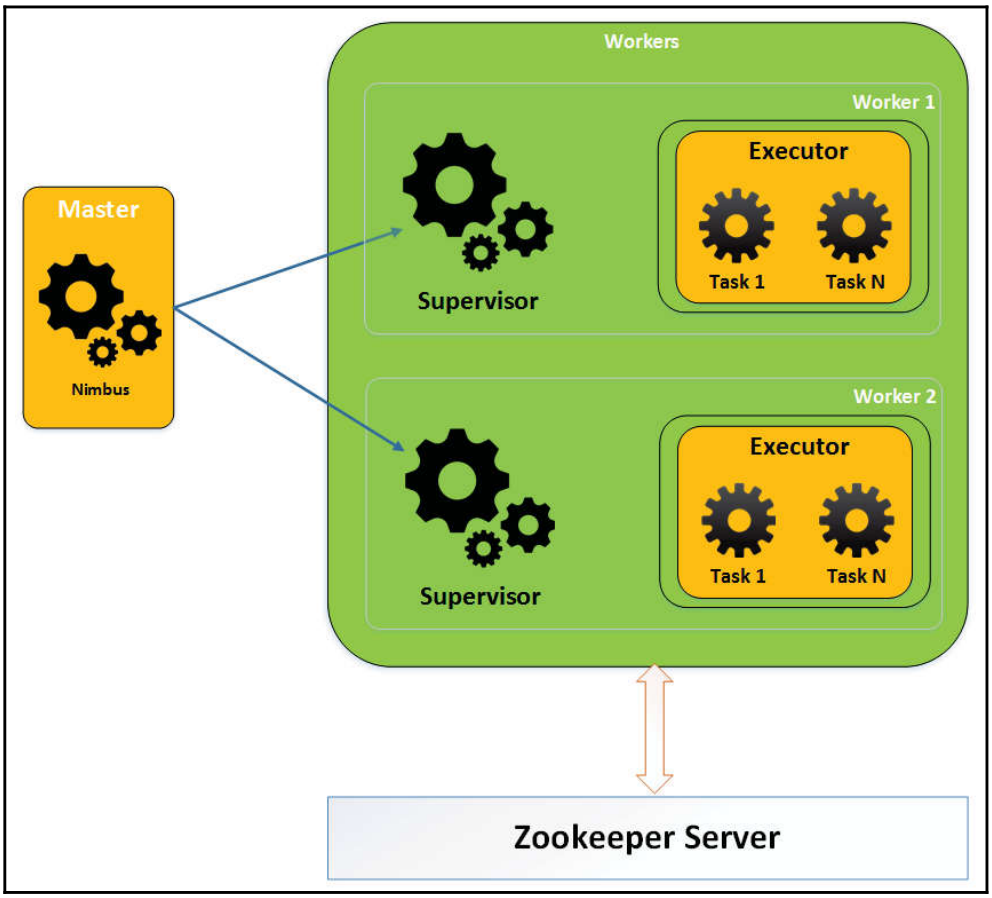
Chapter 6: Real-Time Processing Engines

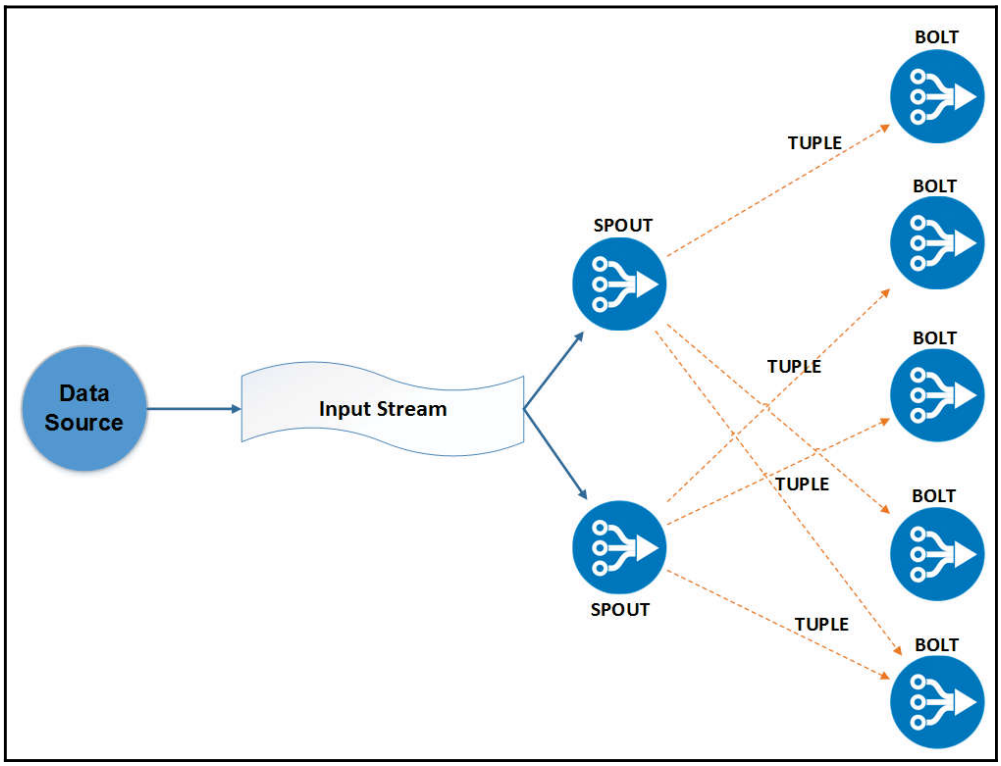


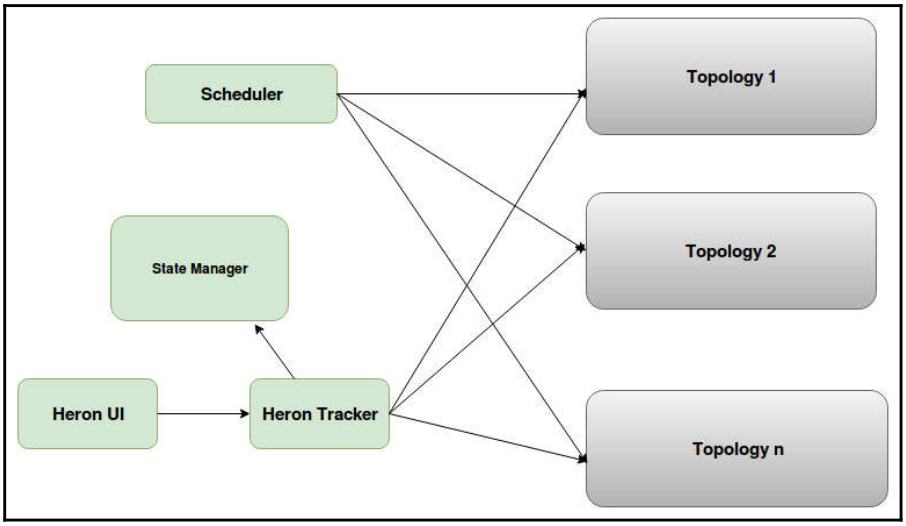

```
exa00077@exa00071:~/Softwares/spark/bin$ ./spark-shell
2018-09-15 20:30:19 WARN  Utils:66 - Your hostname, exa00071 resolves to a loopback address: 127.0.1.1; using 192.168.0.103 instead
ce wlp5s0)
2018-09-15 20:30:19 WARN  Utils:66 - Set SPARK_LOCAL_IP if you need to bind to another address
2018-09-15 20:30:20 WARN  NativeCodeLoader:62 - Unable to load native-hadoop library for your platform... using builtin-java classes
icable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://192.168.0.103:4040
Spark context available as 'sc' (master = local[*], app id = local-1537023626092).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.3.1
      /_/

Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_181)
Type in expressions to have them evaluated.
Type :help for more information.
```

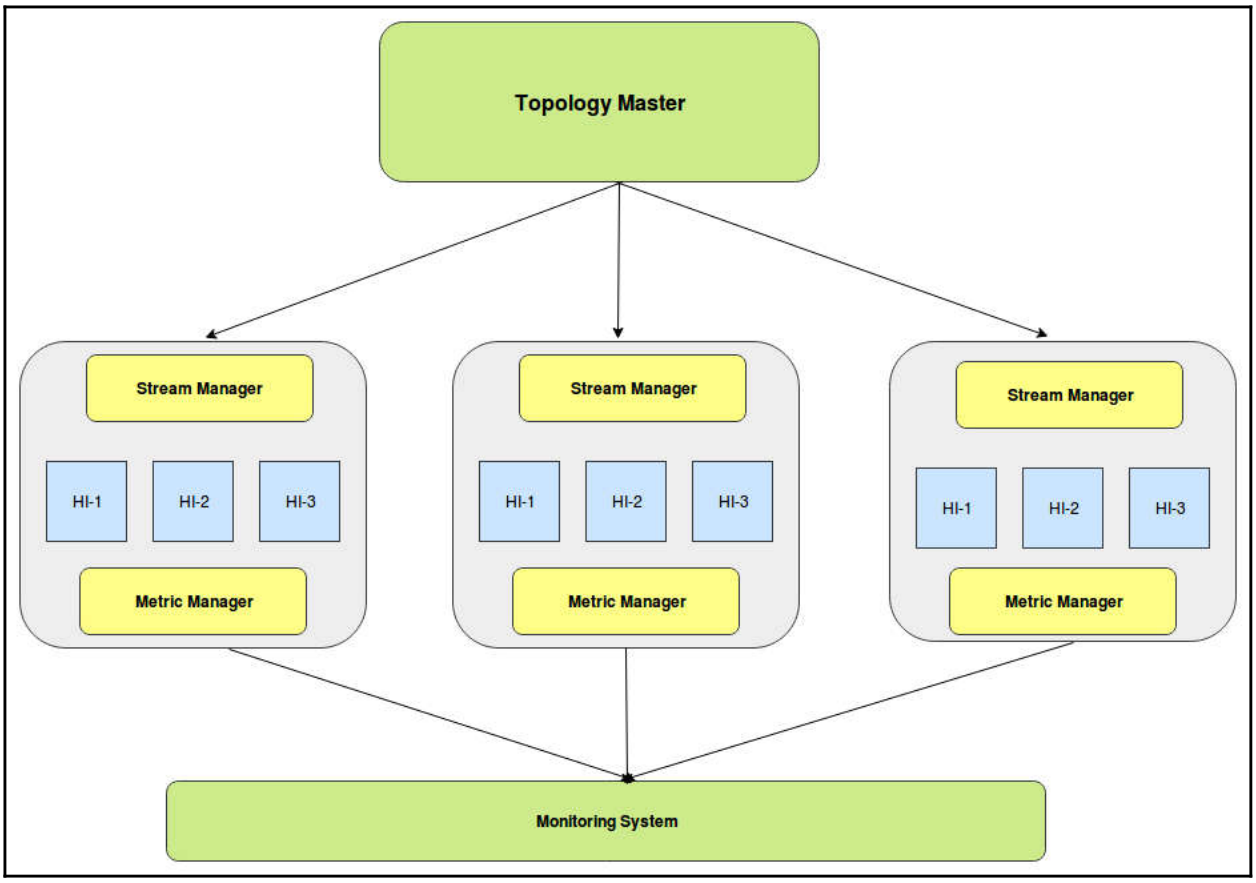

Job manager
Client
Dispatcher
Scheduler
Checkpoint
Coordinator
Task Slot
Task Slot
Task Manager
Task Slot
Task Slot
Task Manager
Task Slot
Task Slot
Task Manager


Spark ML
Table API
Gelly (Graph Processing
Table API
CEP(Event processing)
Dataset API
Data Stream API
Optimizer
Stream Builder
Runtime
Local
Cluster(YARN)
Cloud(Ec2 ,GCE)
HDFS/KAFKA/HBASE etc.

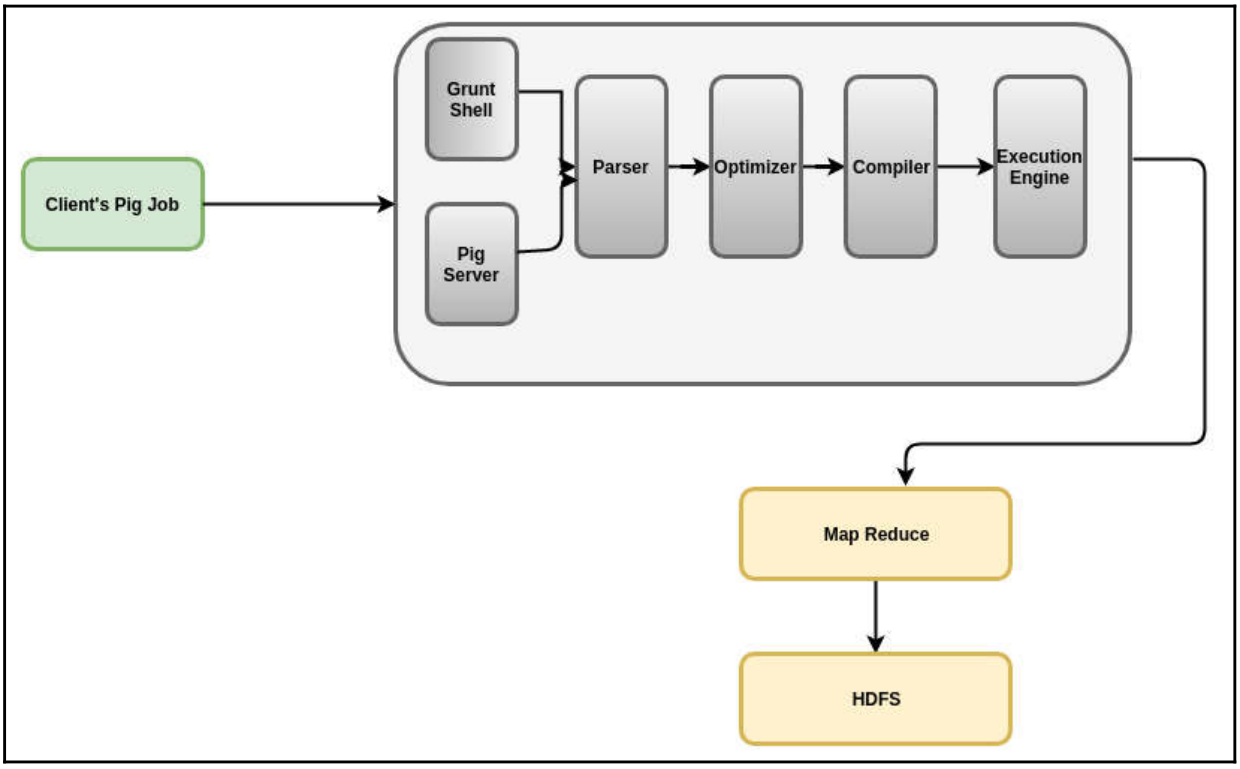

Workers
Worker 1
Executor
Task 1
Task N
Supervisor
Master
Nimbus
Worker 2
Executor
Task 1
Task N
Supervisor
Zookeeper Server

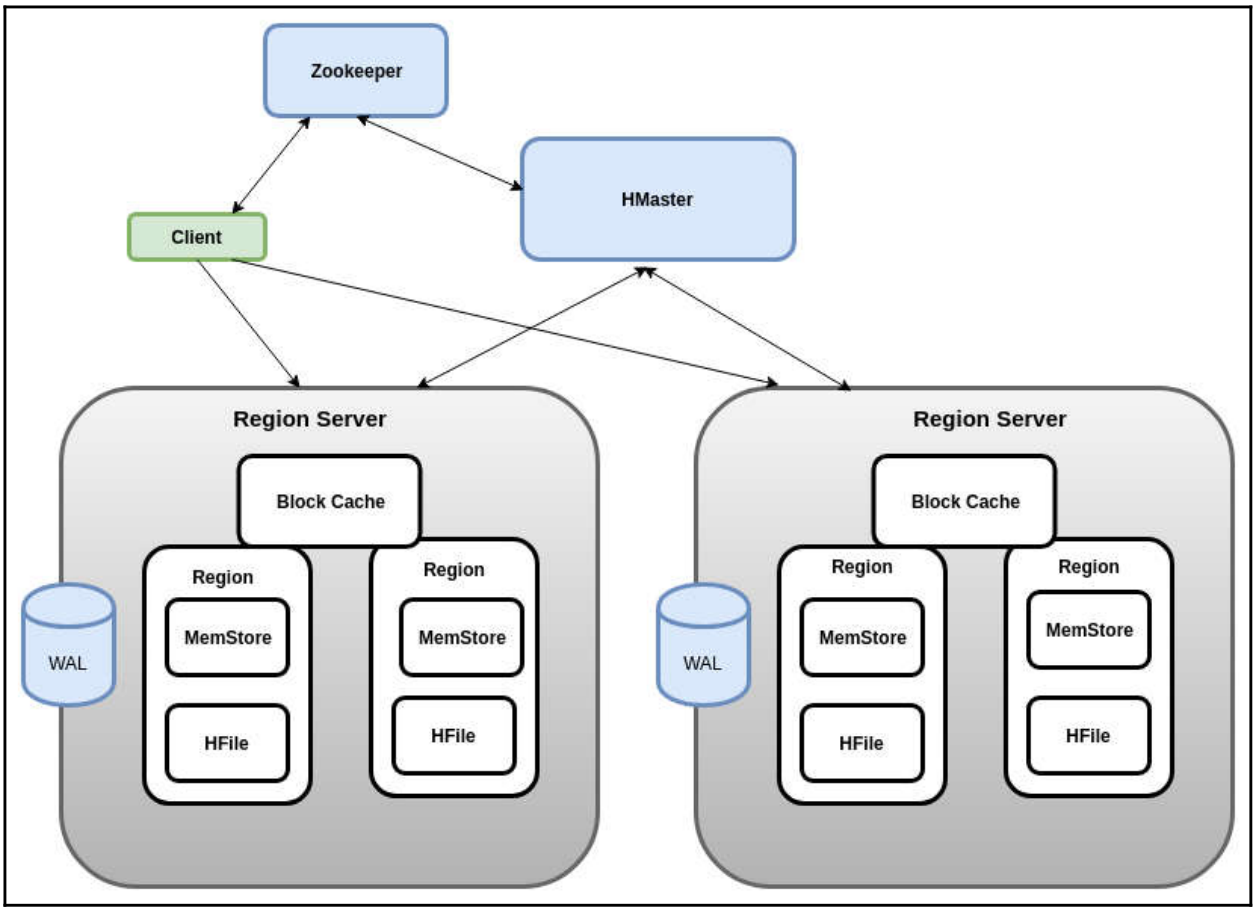

Data Source
Input Stream
SPOUT
SPOUT
TUPLE
TUPLE
TUPLE
TUPLE
BOLT
BOLT
BOLT
BOLT
BOLT

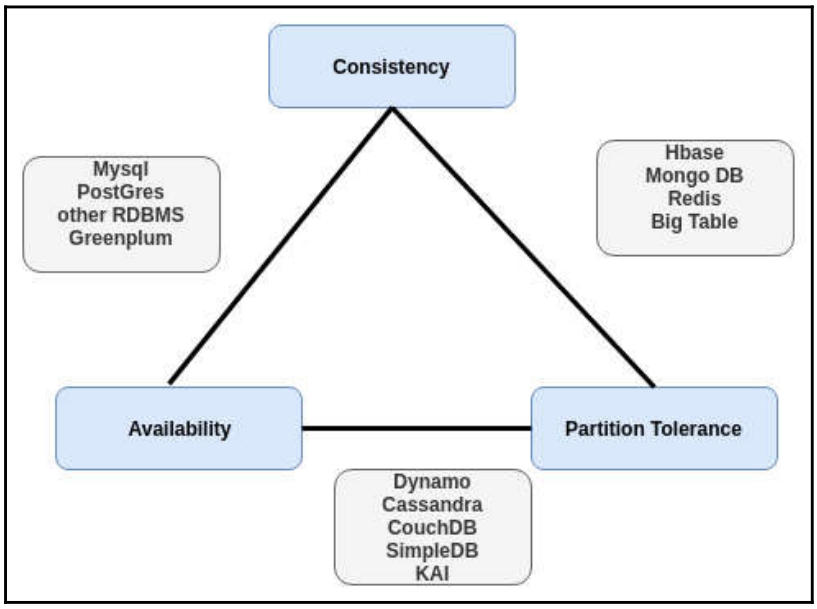

Scheduler
State Manager
Heron UI
Heron Tracker
Topology 1
Topology 2
Topology n


Topology Master
Stream Manager
HI-1
HI-2
HI-3
Metric Manager
Stream Manager
HI-1
HI-2
HI-3
Metric Manager
Stream Manager
HI-1
HI-2
HI-3
Metric Manager
Monitoring System

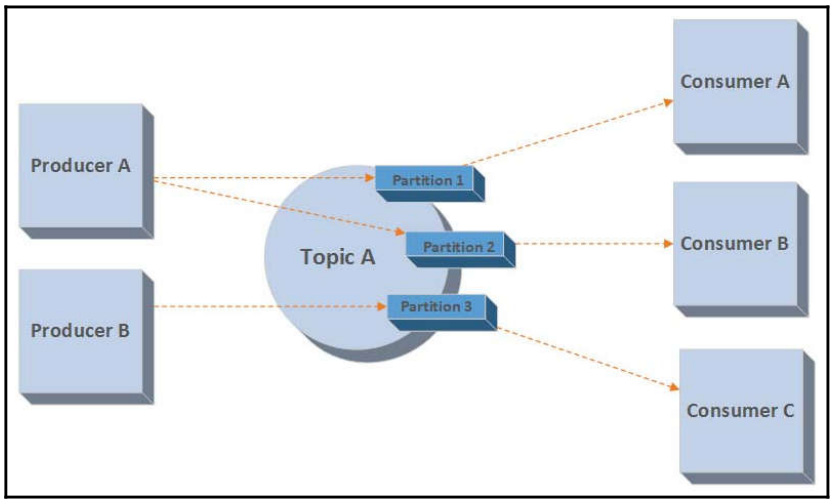
Chapter 7:
Widely Used Hadoop Ecosystem Components

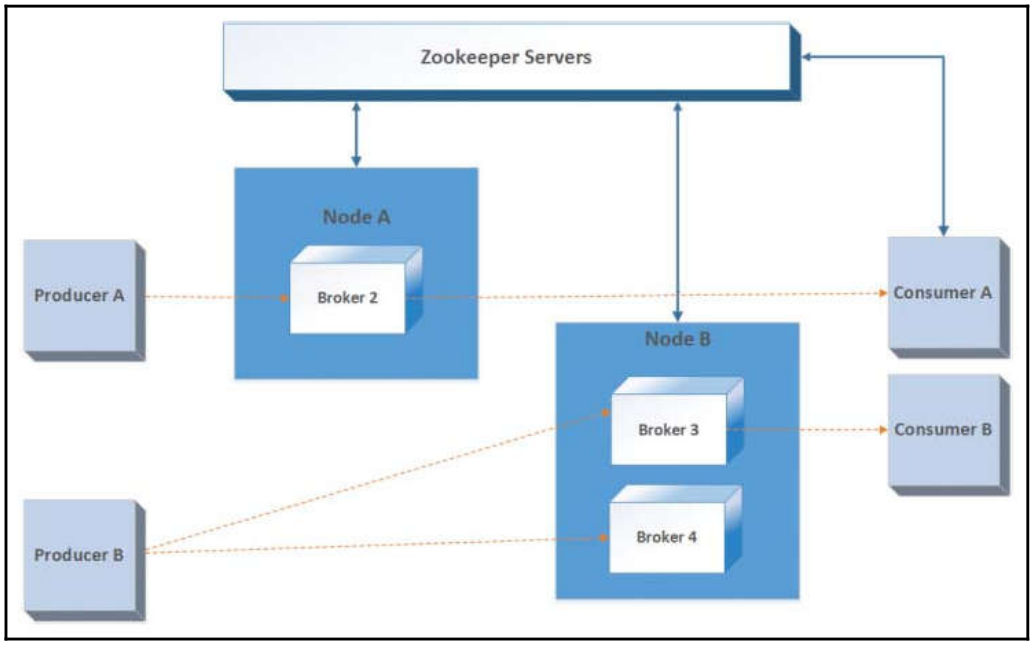


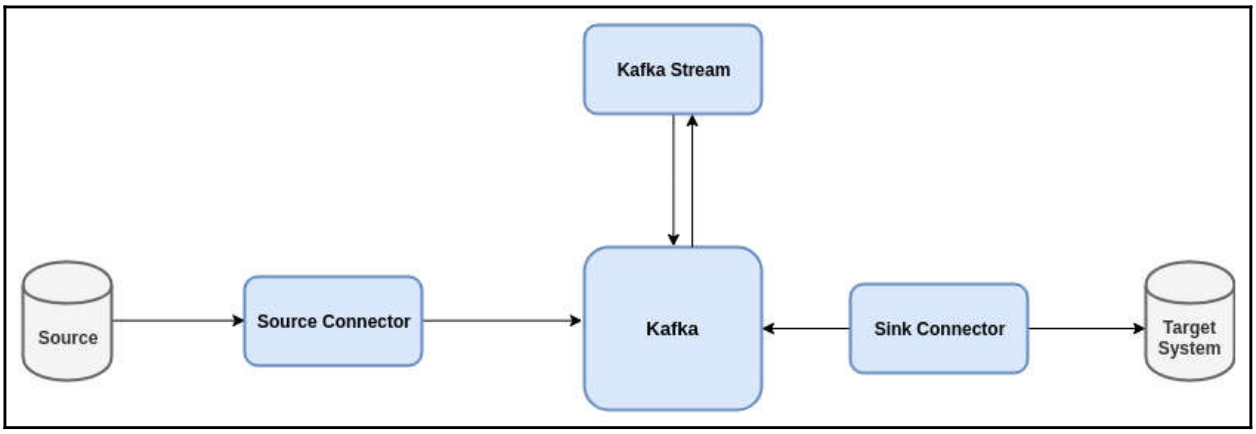

Zookeeper
HMaster
Client
Region Server
Block Cache
Region
MemStore
HFile
Region
MemStore
HFile
WAL
Region Server
Block Cache
Region
MemStore
HFile
Region
MemStore
HFile
WAL




```
# Set environment variables here.

# This script sets variables multiple times over the course of starting an hbase process,
# so try to keep things idempotent unless you want to take an even deeper look
# into the startup scripts (bin/hbase, etc.)

# The java implementation to use.  Java 1.8+ required.
 export JAVA_HOME=/opt/java/jdk1.8.0/

# Extra Java CLASSPATH elements.  Optional.
# export HBASE_CLASSPATH=

# The maximum amount of heap to use. Default is left to JVM default.
# export HBASE_HEAPSIZE=1G

# Uncomment below if you intend to use off heap cache. For example, to allocate 8G of
# offheap, set the value to "8G".
# export HBASE_OFFHEAPSIZE=1G

# Extra Java runtime options.
# Below are what we set by default.  May only work with SUN JVM.
# For more on why as well as other possible settings,
# see http://hbase.apache.org/book.html#performance
export HBASE_OPTS="$HBASE_OPTS -XX:+UseConcMarkSweepGC"

# Uncomment one of the below three options to enable java garbage collection logging for the server-side processes.

# This enables basic gc logging to the .out file.
# export SERVER_GC_OPTS="-verbose:gc -XX:+PrintGCDetails -XX:+PrintGCDateStamps"

# This enables basic gc logging to its own file.
# If FILE-PATH is not replaced, the log file(.gc) would still be generated in the HBASE_LOG_DIR .
# export SERVER_GC_OPTS="-verbose:gc -XX:+PrintGCDetails -XX:+PrintGCDateStamps -Xloggc:<FILE-PATH>"

# This enables basic GC logging to its own file with automatic log rolling. Only applies to jdk 1.6.0_34+ and 1.7.0_2+.
# If FILE-PATH is not replaced, the log file(.gc) would still be generated in the HBASE_LOG_DIR .
# export SERVER_GC_OPTS="-verbose:gc -XX:+PrintGCDetails -XX:+PrintGCDateStamps -Xloggc:<FILE-PATH> -XX:+UseGCLogFileRotation -XX:NumberO
```

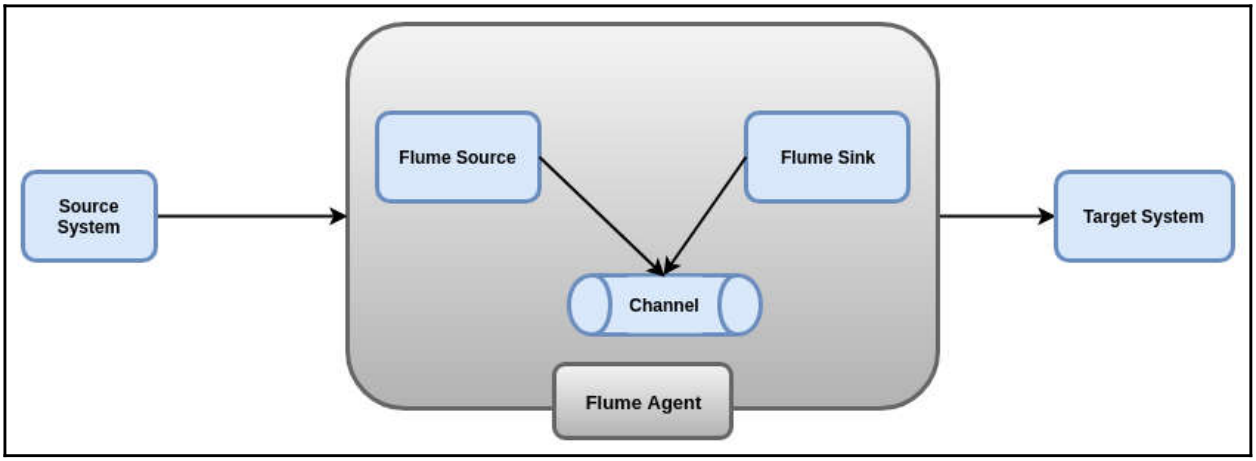

Producer A
Producer B
Topic A
Partition 1
Partition 2
Partition 3
Consumer A
Consumer B
Consumer C

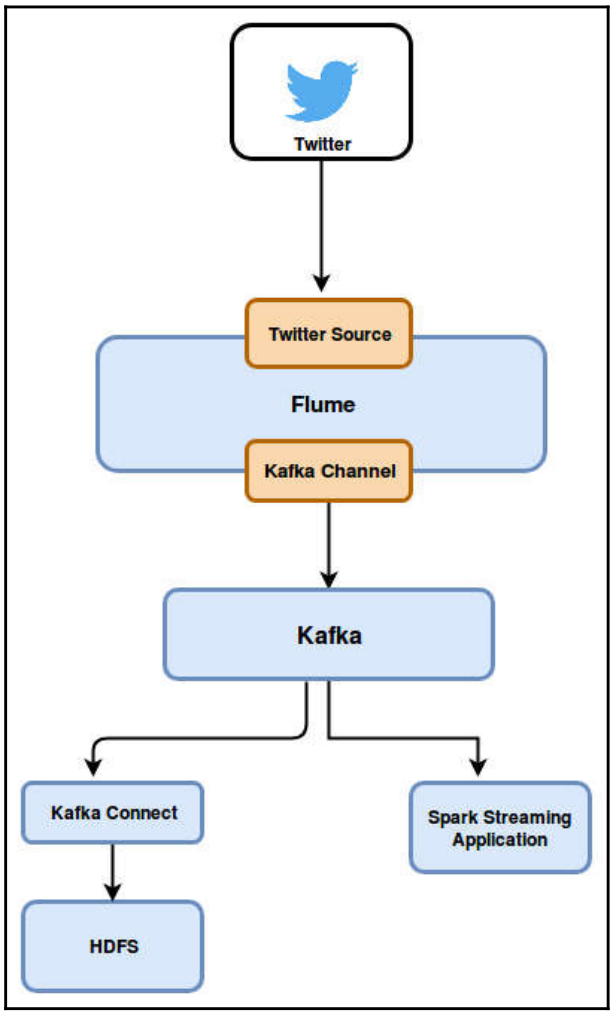

Zookeeper Servers
Node A
Broker 2
Node B
Broker 3
Broker 4
Producer A
Producer B
Consumer A
Consumer B



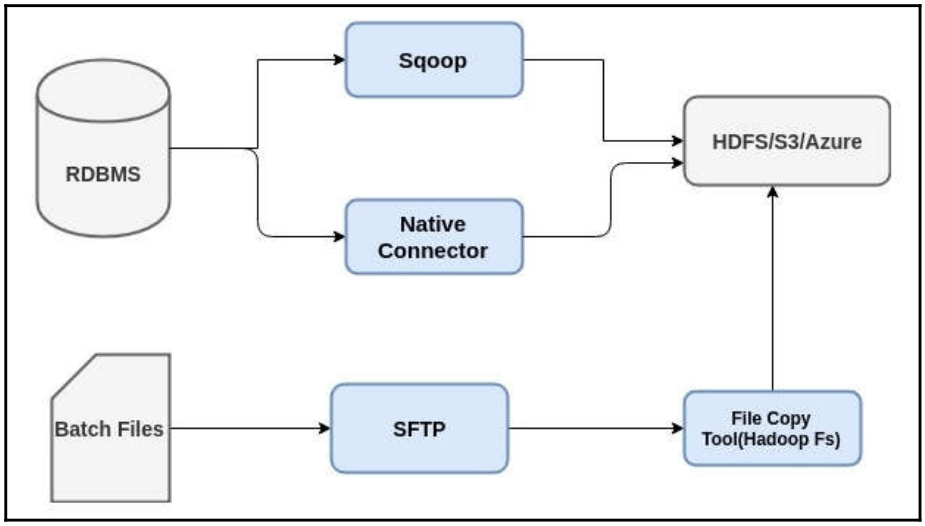

```
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/chanchal/projects/confluent-3.2.2/share/
StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/chanchal/projects/confluent-3.2.2/share/
aticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
{"id":1,"name":{"string":"Manish"}}
{"id":2,"name":{"string":"Chanchal"}}
```

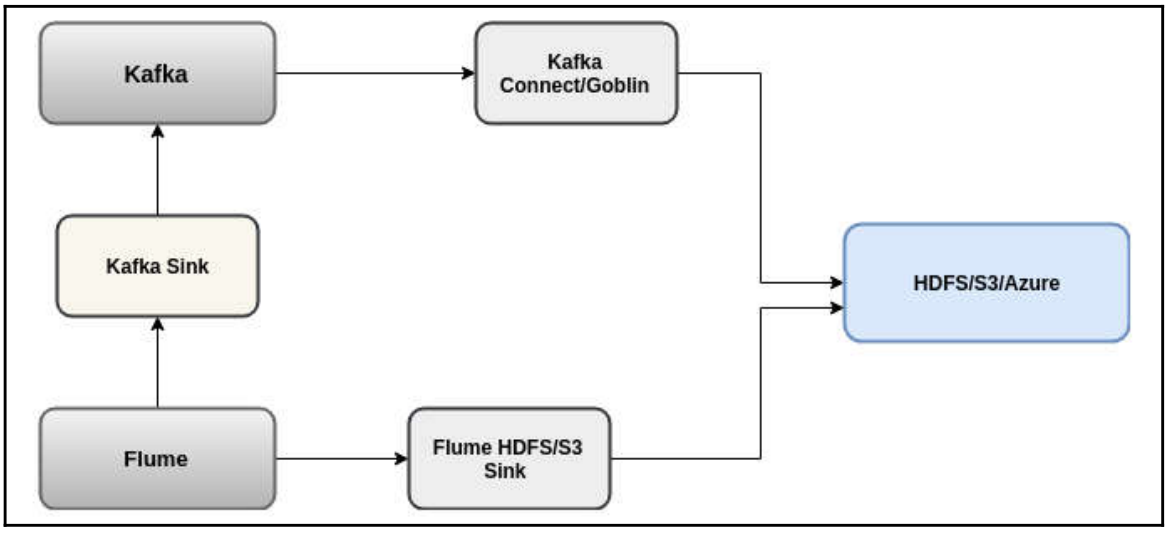
```
sqlite> select * from authors_sink;
Chanchal|60.0|1|26
Manish|80.0|2|32
sqlite>
```

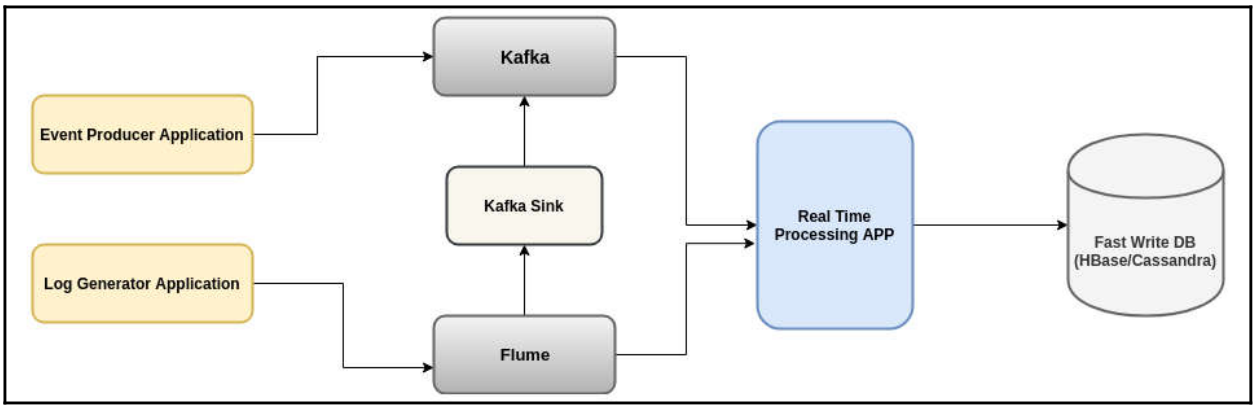


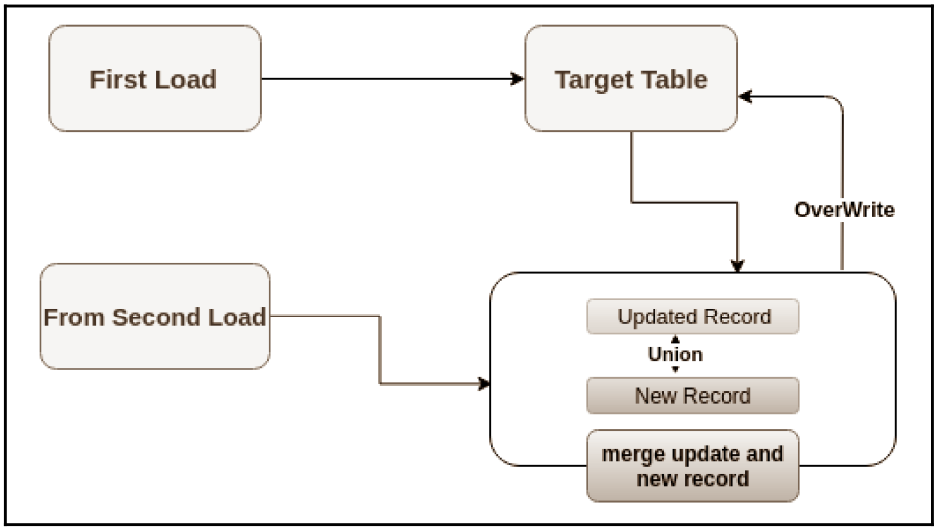

Twitter
Twitter Source
Flume
Kafka Channel
Kafka
Kafka Connect
HDFS
Spark Streaming Application

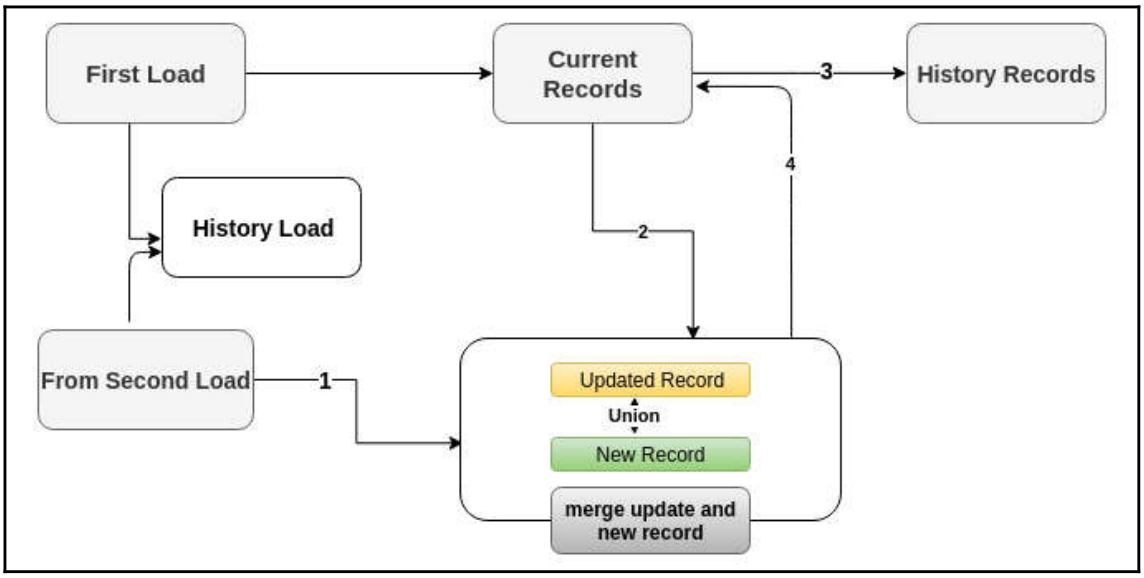
Chapter 8:
Designing Applications in Hadoop

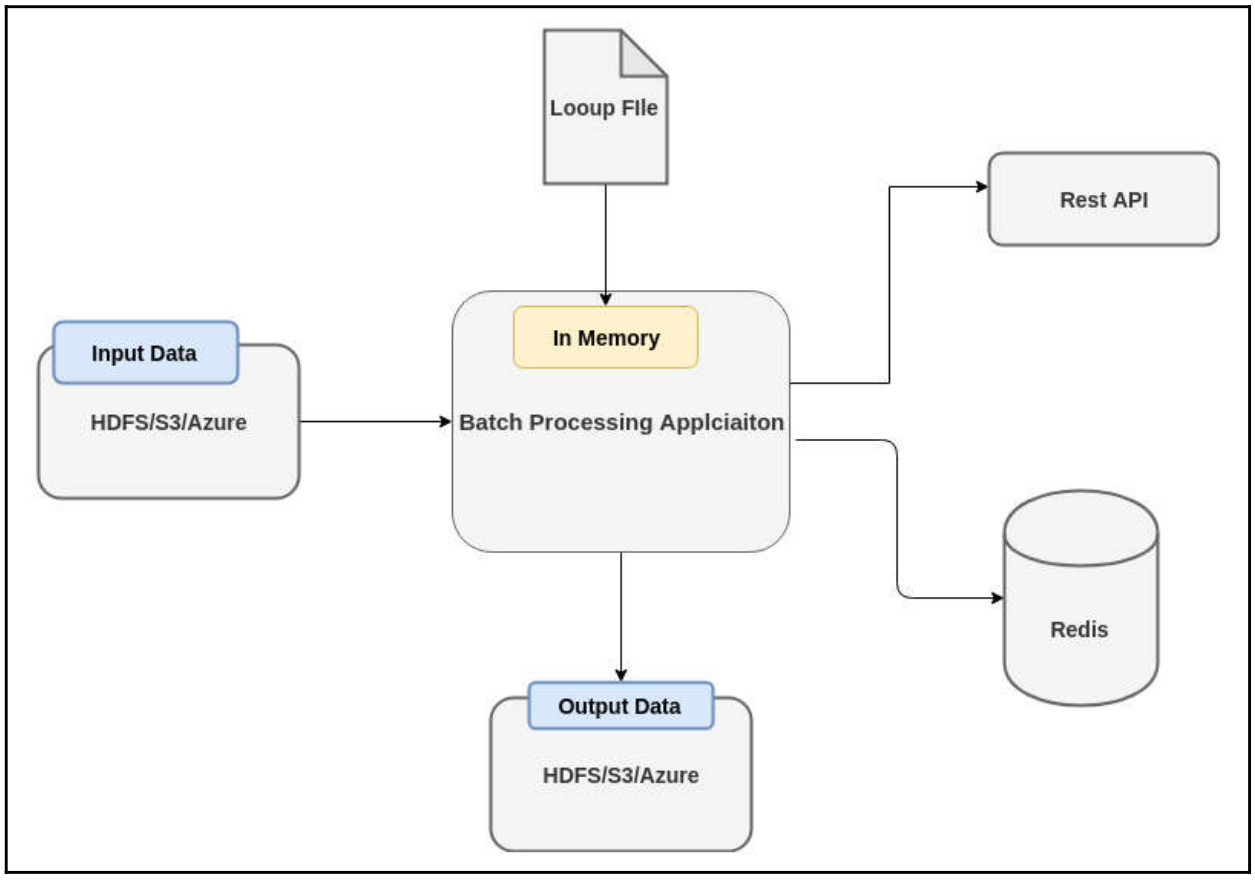


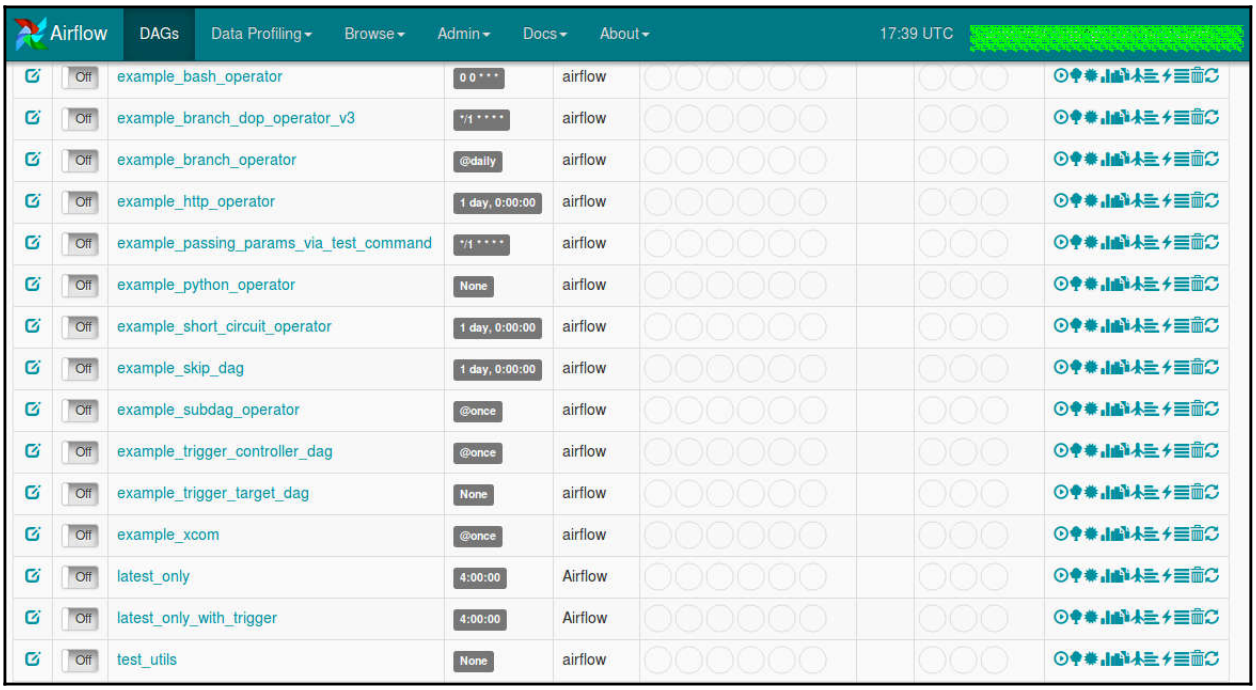


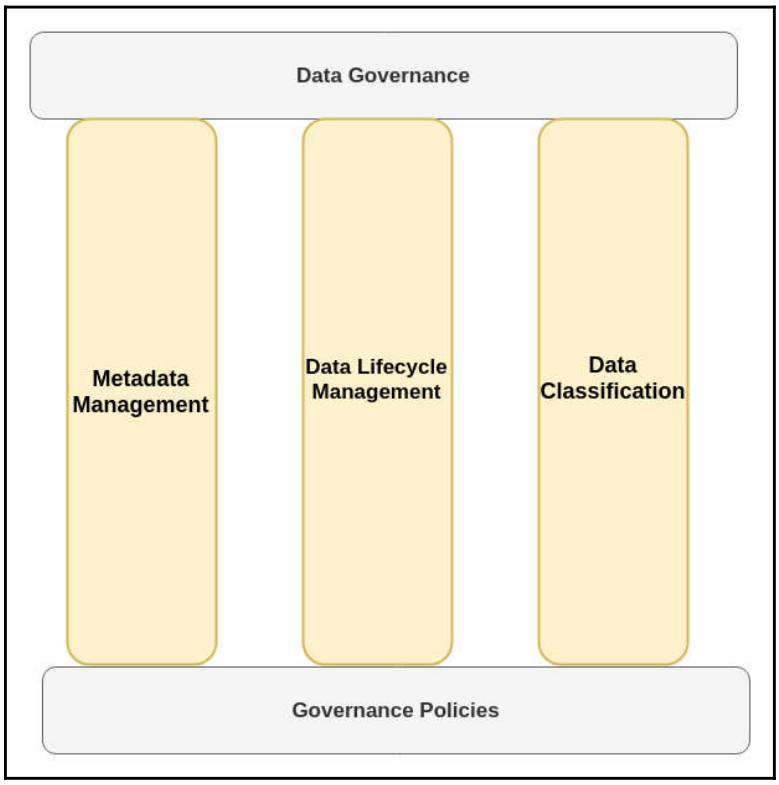

Kafka
Event Producer Application
Kafka Sink
Log Generator Application
Flume
Real Time
Processing APP
Fast Write DB
(HBase/Cassandra)


First Load
Target Table
OverWrite
From Second Load
Updated Record
Union
New Record
merge update and
new record

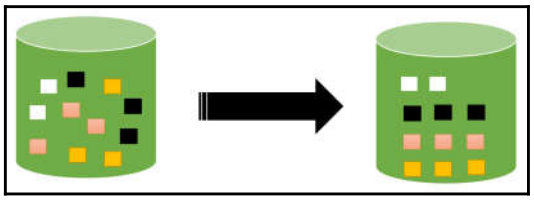

First Load
Current Records
3
History Records
4
History Load
2
From Second Load
1
Updated Record
Union
New Record
merge update and new record

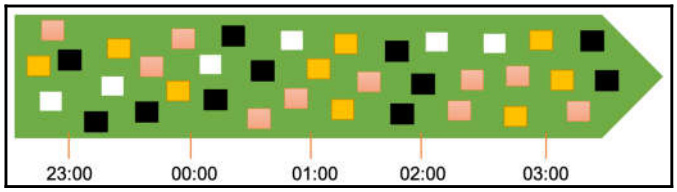

Looup FIle
Rest API
Input Data
HDFS/S3/Azure
In Memory
Batch Processing Applciaiton
Redis
Output Data
HDFS/S3/Azure

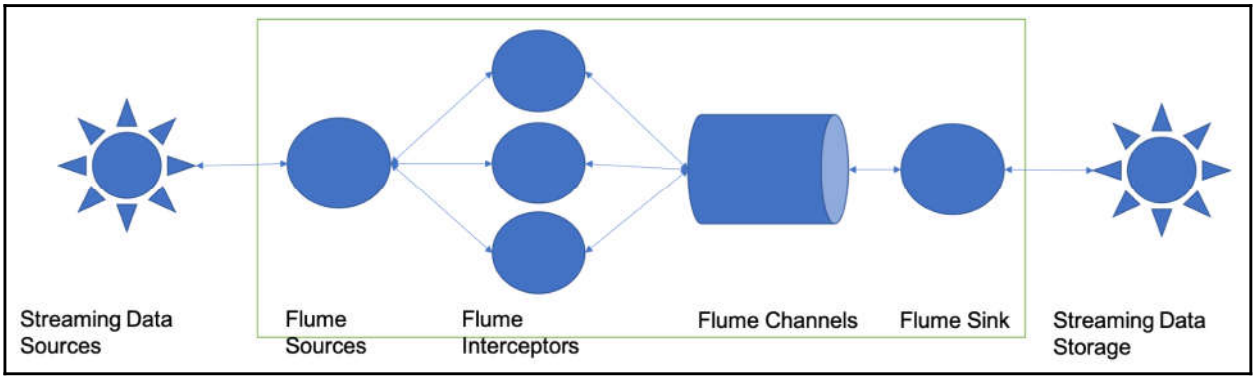

Airflow
DAGs
Data Profiling
Browse
Admin
Docs
About
17:39 UTC
Off example_bash_operator 0 0 * * * airflow
Off example_branch_dop_operator_v3 */1 * * * * airflow
Off example_branch_operator @daily airflow
Off example_http_operator 1 day, 0:00:00 airflow
Off example_passing_params_via_test_command */1 * * * * airflow
Off example_python_operator None airflow
Off example_short_circuit_operator 1 day, 0:00:00 airflow
Off example_skip_dag 1 day, 0:00:00 airflow
Off example_subdag_operator @once airflow
Off example_trigger_controller_dag @once airflow
Off example_trigger_target_dag None airflow
Off example_xcom @once airflow
Off latest_only 4:00:00 Airflow
Off latest_only_with_trigger 4:00:00 Airflow
Off test_utils None airflow

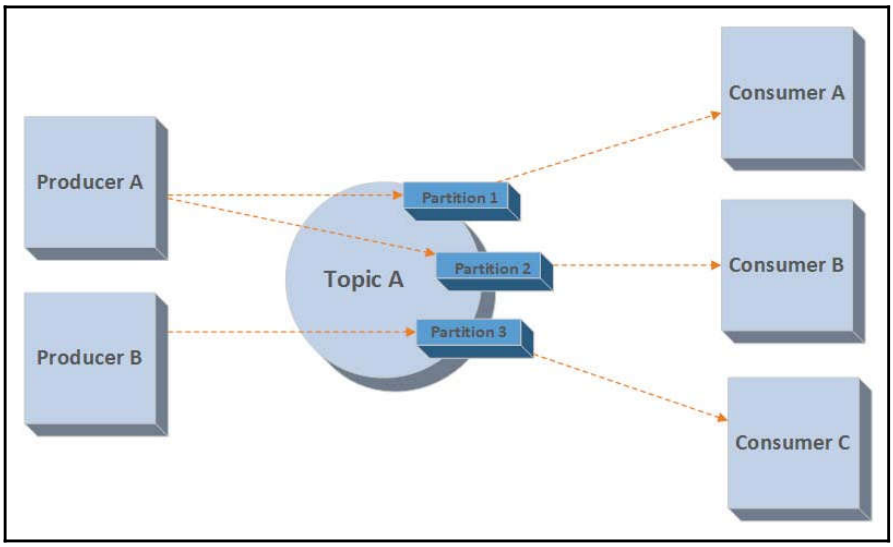

Data Governance
Metadata Management
Data Lifecycle Management
Data Classification
Governance Policies

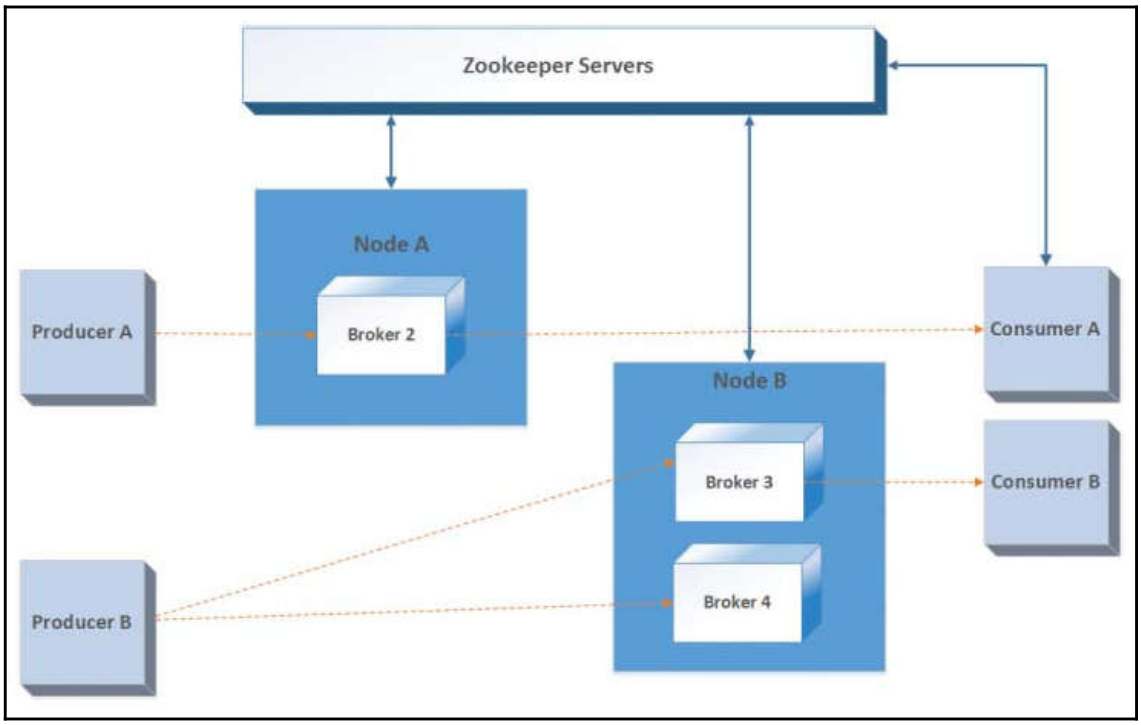
Chapter 9: Real-Time Stream Processing in Hadoop

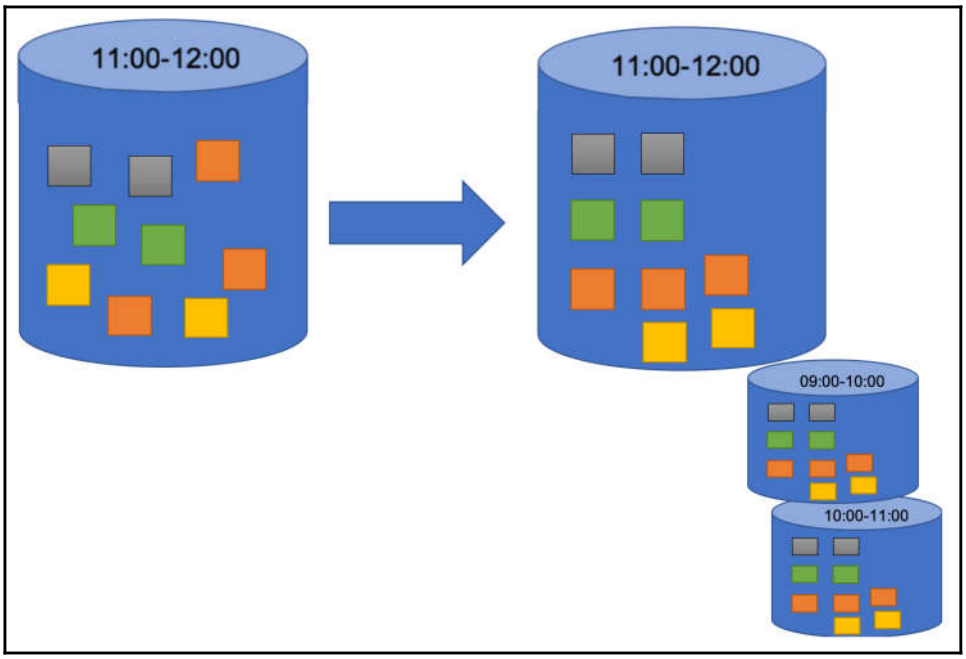


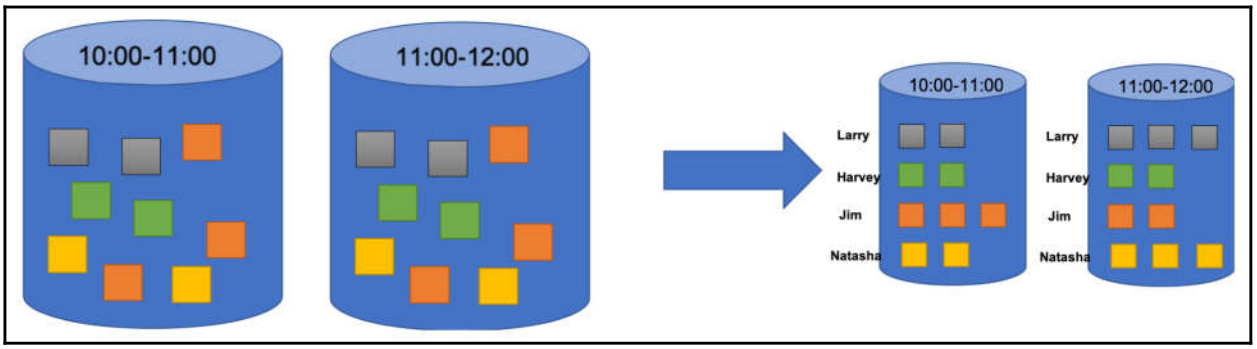




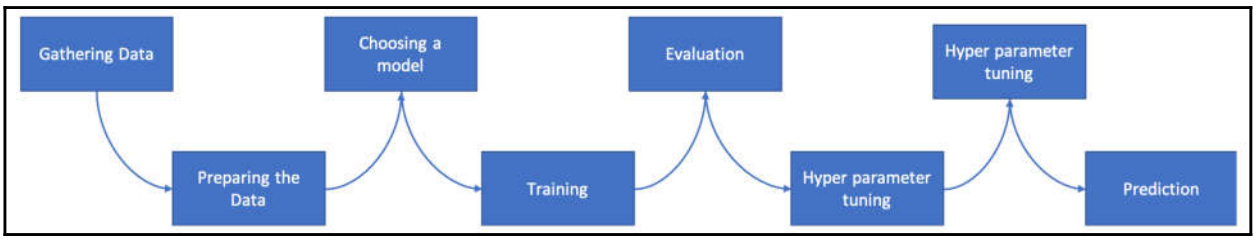

Producer A
Producer B
Topic A
Partition 1
Partition 2
Partition 3
Consumer A
Consumer B
Consumer C

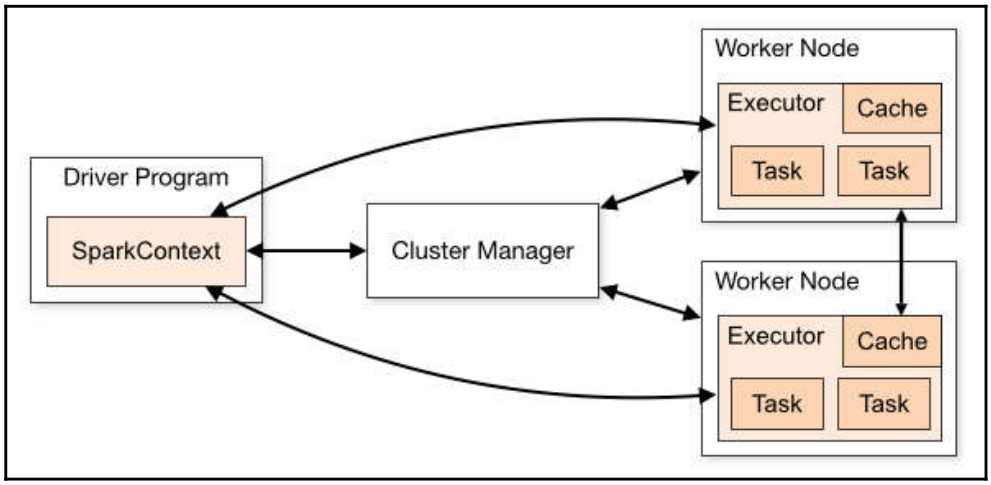

Zookeeper Servers
Node A
Broker 2
Node B
Broker 3
Broker 4
Producer A
Producer B
Consumer A
Consumer B

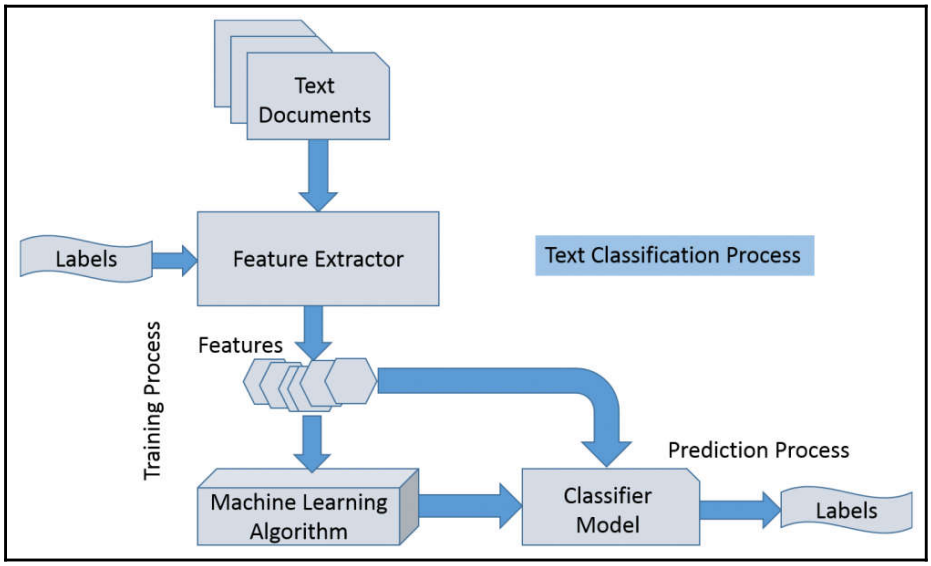

11:00-12:00
11:00-12:00
09:00-10:00
10:00-11:00


10:00-11:00
11:00-12:00
10:00-11:00
Larry
Harvey
Jim
Natasha
11:00-12:00
Larry
Harvey
Jim
Natasha

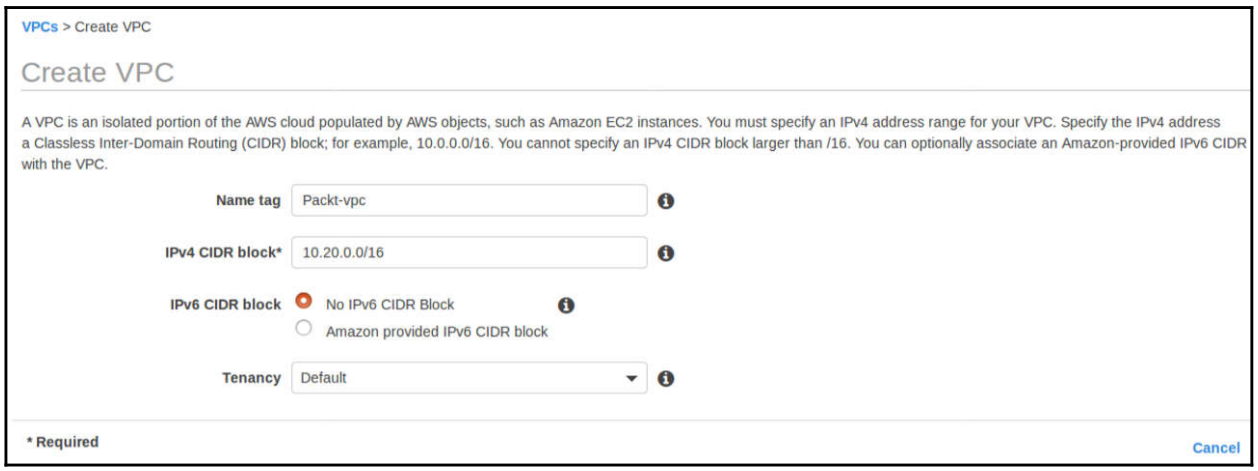
Chapter 10: Machine Learning in Hadoop

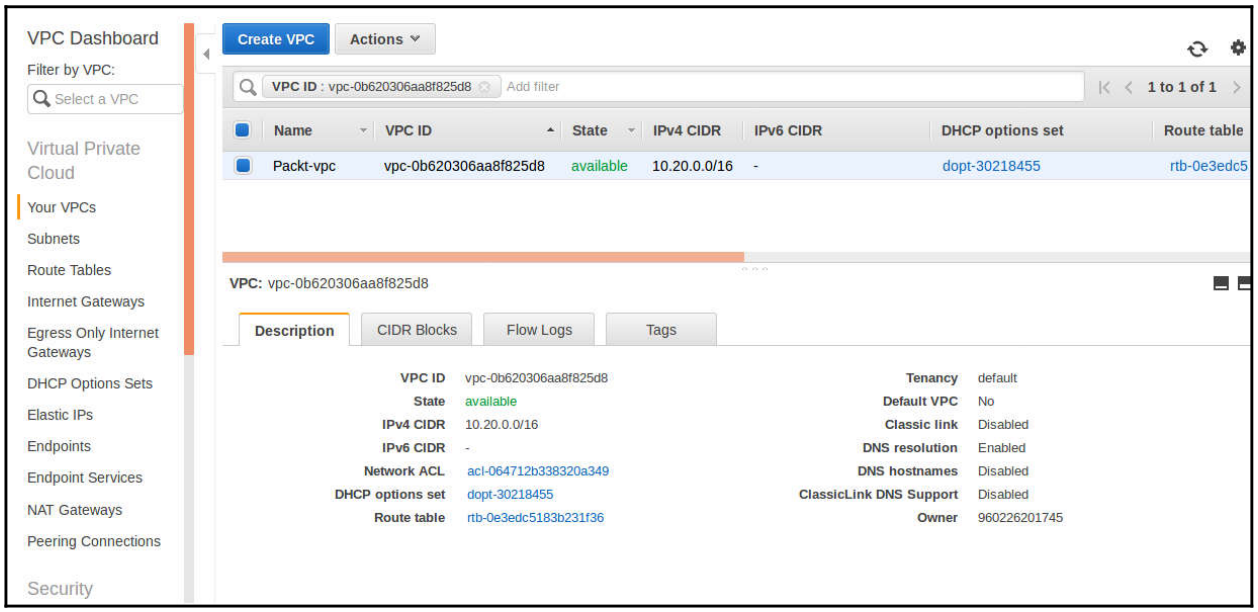


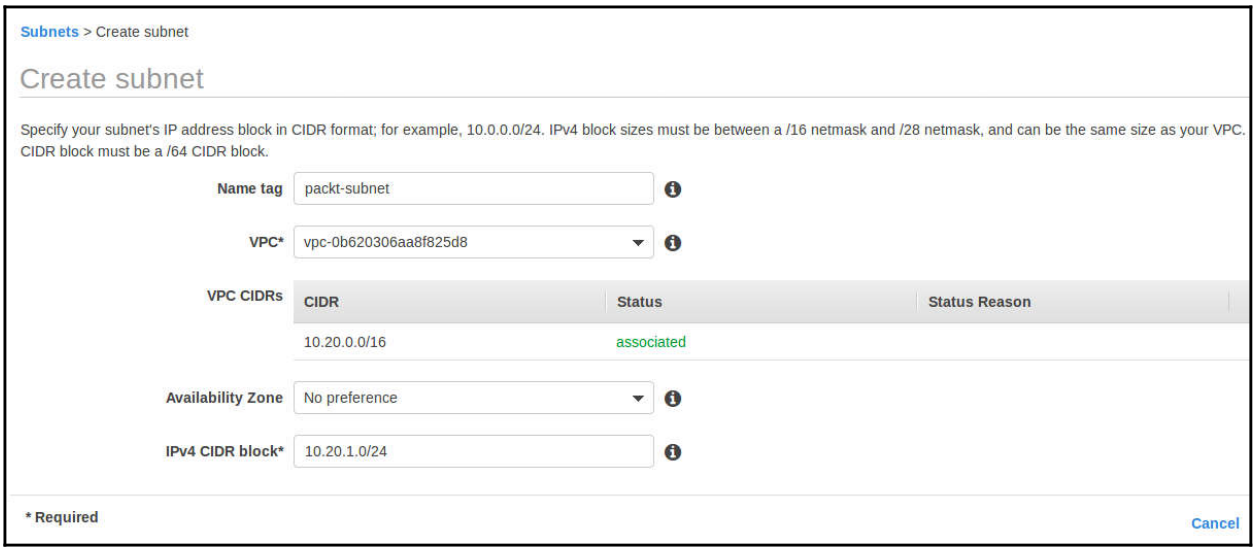


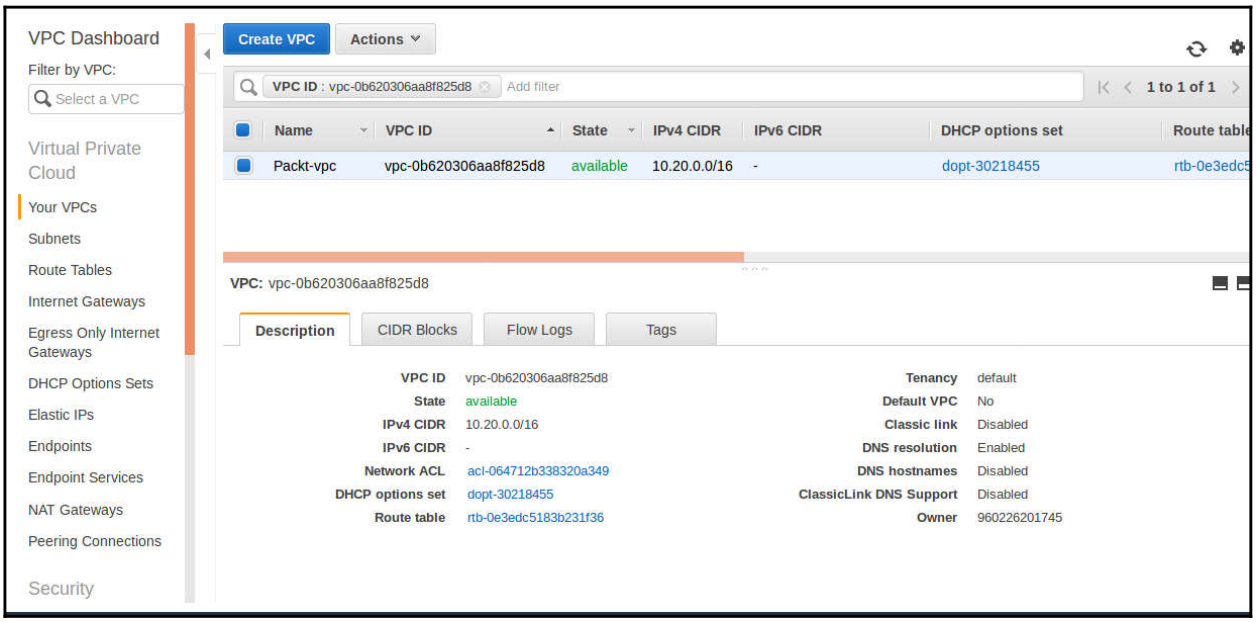


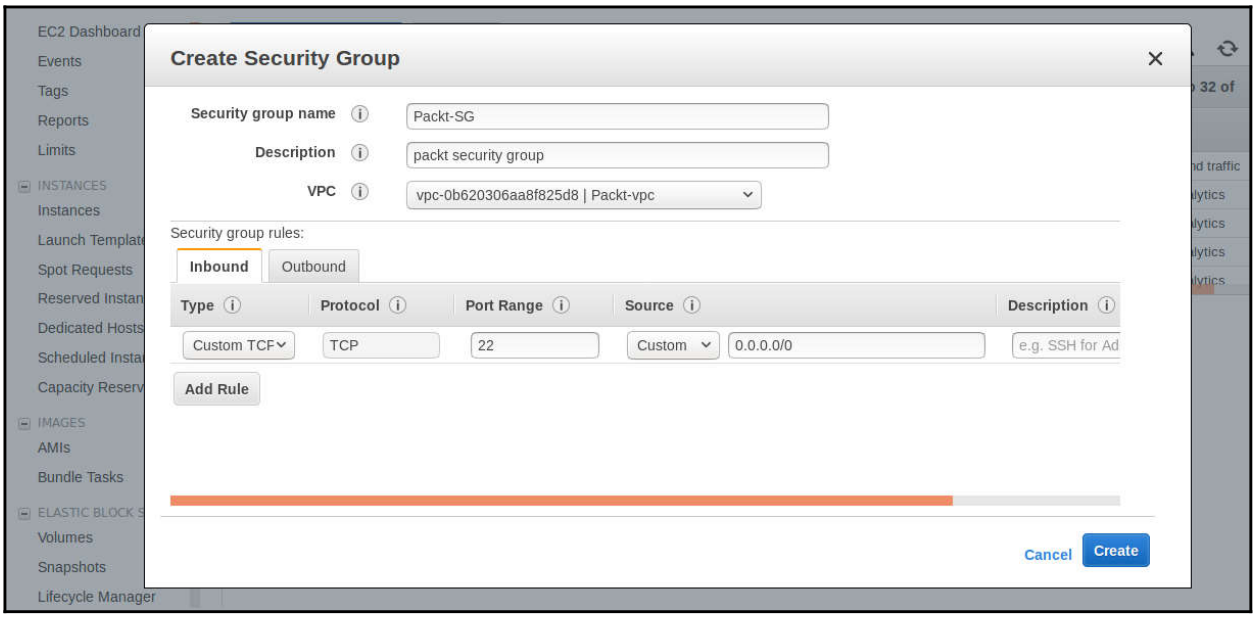
Chapter 11: Hadoop in the Cloud

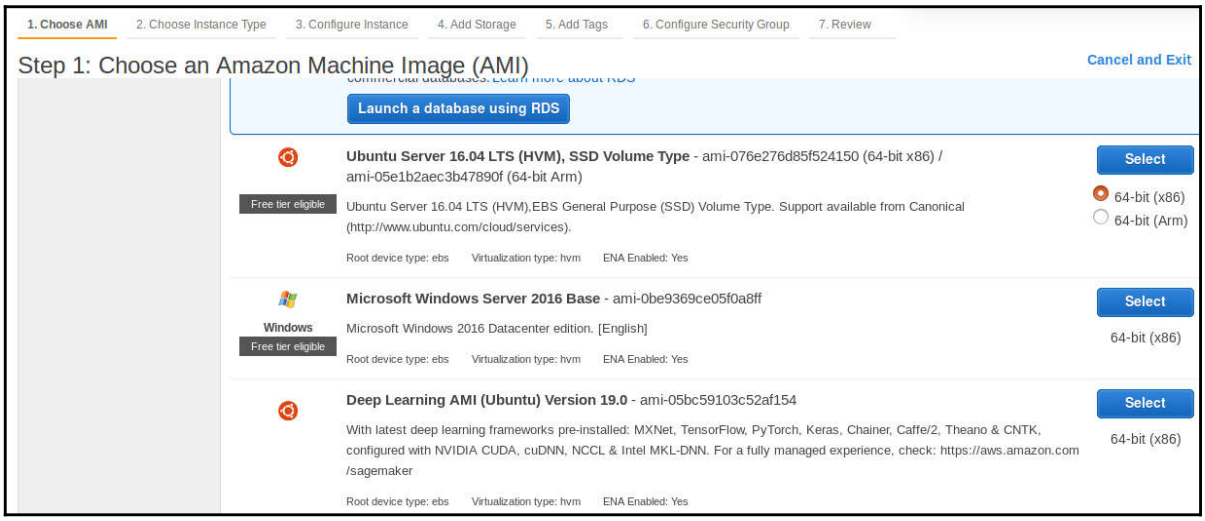


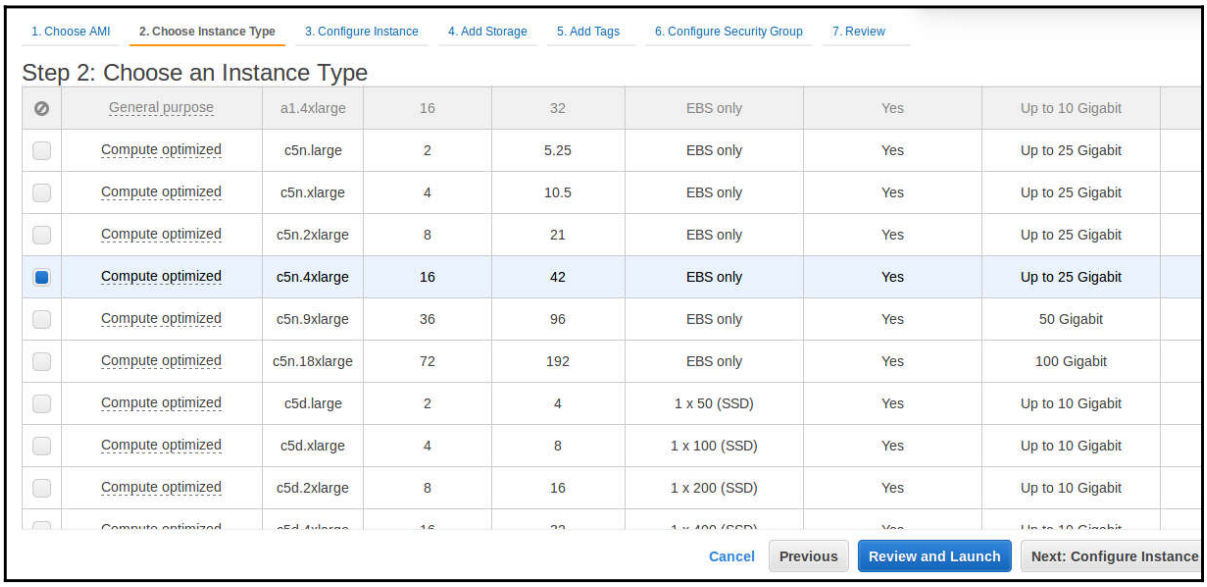


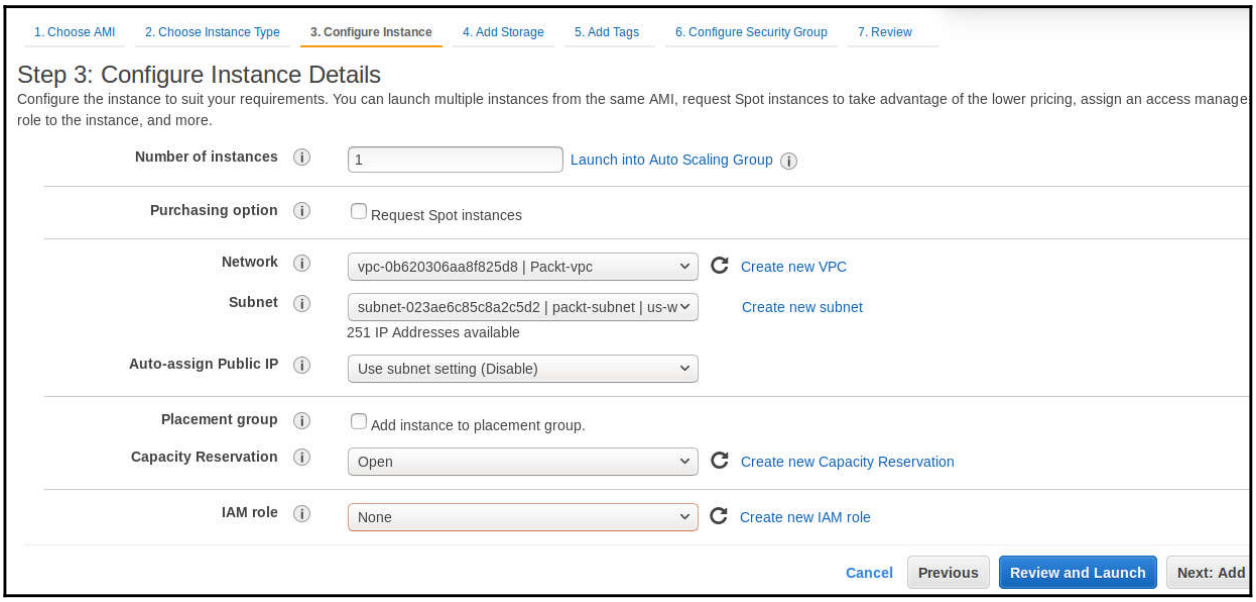

Subnets > Create subnet
Create subnet
Specify your subnet's IP address block in CIDR format; for example, 10.0.0.0/24. IPv4 block sizes must be between a /16 netmask and /28 netmask, and can be the same size as your VPC. CIDR block must be a /64 CIDR block.
Name tag
packt-subnet
VPC*
vpc-0b620306aa8f825d8
VPC CIDRs
CIDR
Status
Status Reason
10.20.0.0/16
associated
Availability Zone
No preference
IPv4 CIDR block*
10.20.1.0/24
* Required
Cancel

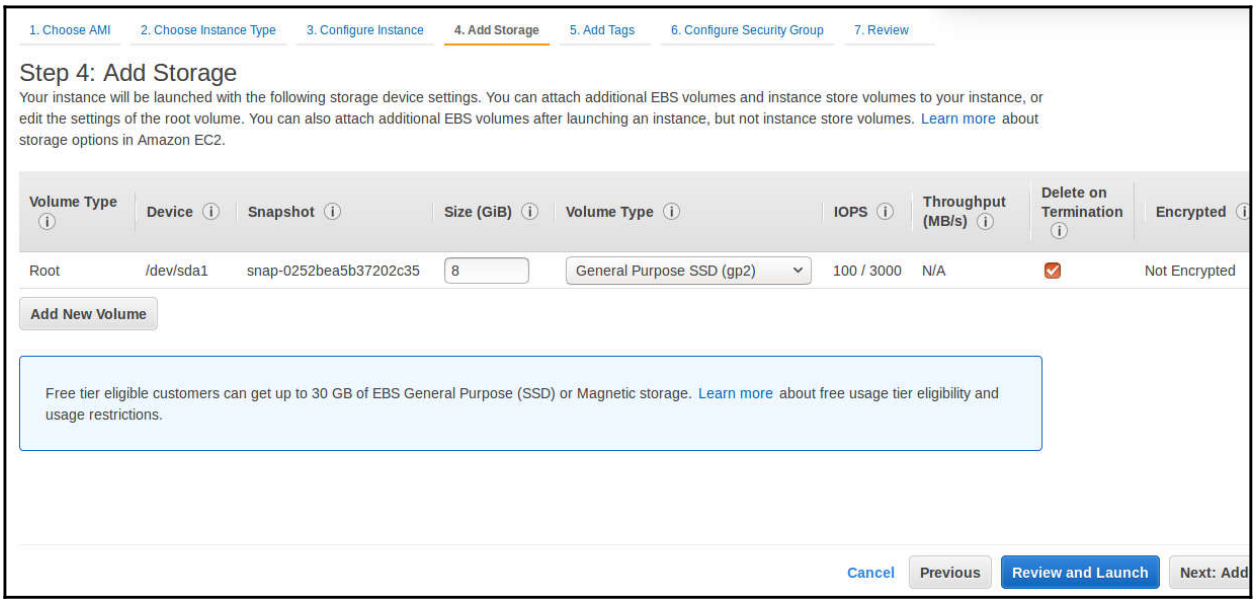

VPC Dashboard
Filter by VPC:
Select a VPC
Virtual Private Cloud
Your VPCs
Subnets
Route Tables
Internet Gateways
Egress Only Internet Gateways
DHCP Options Sets
Elastic IPs
Endpoints
Endpoint Services
NAT Gateways
Peering Connections
Security
Create VPC
Actions
VPC ID : vpc-0b620306aa8f825d8
Add filter
1 to 1 of 1
Name
VPC ID
State
IPv4 CIDR
IPv6 CIDR
DHCP options set
Packt-vpc
vpc-0b620306aa8f825d8
available
10.20.0.0/16
-
dopt-30218455
VPC: vpc-0b620306aa8f825d8
Description
CIDR Blocks
Flow Logs
Tags
VPC ID vpc-0b620306aa8f825d8
State available
IPv4 CIDR 10.20.0.0/16
IPv6 CIDR -
Network ACL acl-064712b338320a349
DHCP options set dopt-30218455
Route table rtb-0e3edc5183b231f36
Tenancy default
Default VPC No
Classic link Disabled
DNS resolution Enabled
DNS hostnames Disabled
ClassicLink DNS Support Disabled
Owner 960226201745

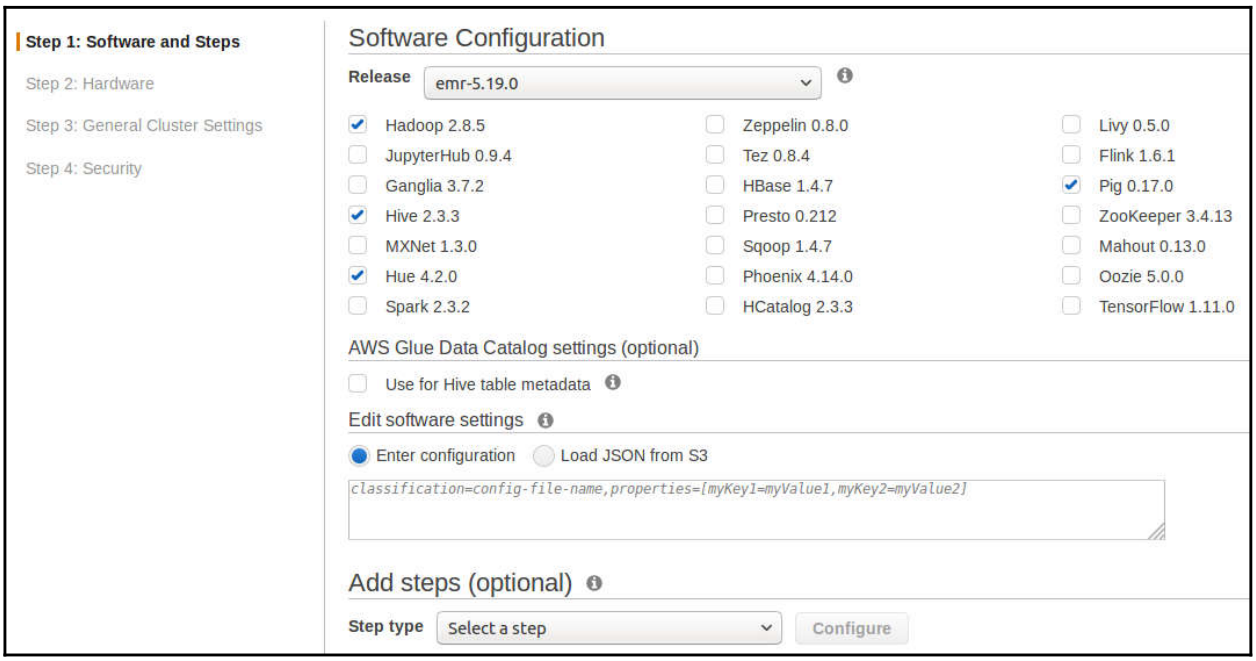

EC2 Dashboard
Events
Tags
Reports
Limits
INSTANCES
Instances
Launch Templat
Spot Requests
Reserved Instan
Dedicated Hosts
Scheduled Insta
Capacity Reserv
IMAGES
AMIs
Bundle Tasks
ELASTIC BLOCK S
Volumes
Snapshots
Lifecycle Manager
Create Security Group
Security group name
Packt-SG
Description
packt security group
VPC
vpc-0b620306aa8f825d8 | Packt-vpc
Security group rules:
Inbound
Outbound
Type
Protocol
Port Range
Source
Description
Custom TCP
TCP
22
Custom
0.0.0.0/0
e.g. SSH for Ad
Add Rule
Cancel
Create

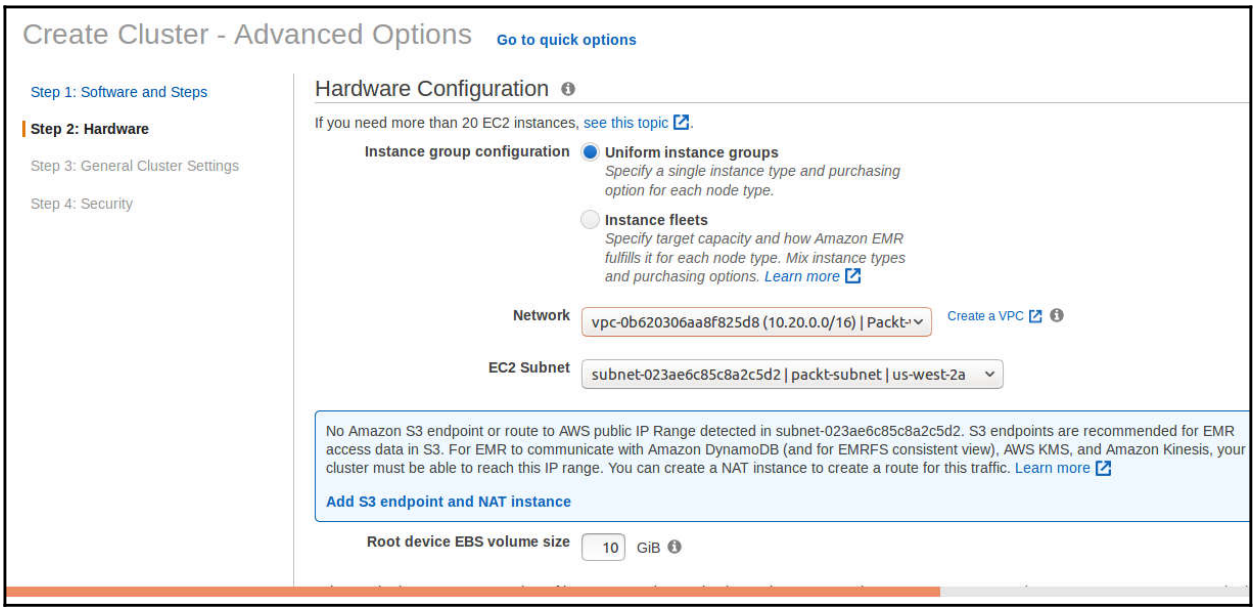

1. Choose AMI
2. Choose Instance Type
3. Configure Instance
4. Add Storage
5. Add Tags
6. Configure Security Group
7. Review
Step 1: Choose an Amazon Machine Image (AMI)
Cancel and Exit
Launch a database using RDS
Ubuntu Server 16.04 LTS (HVM), SSD Volume Type - ami-076e276d85f524150 (64-bit x86) / ami-05e1b2aec3b47890f (64-bit Arm)
Free tier eligible
Ubuntu Server 16.04 LTS (HVM),EBS General Purpose (SSD) Volume Type. Support available from Canonical (http://www.ubuntu.com/cloud/services).
Root device type: ebs
Virtualization type: hvm
ENA Enabled: Yes
Select
64-bit (x86)
64-bit (Arm)
Microsoft Windows Server 2016 Base - ami-0be9369ce05f0a8ff
Windows
Free tier eligible
Microsoft Windows 2016 Datacenter edition. [English]
Root device type: ebs
Virtualization type: hvm
ENA Enabled: Yes
Select
64-bit (x86)
Deep Learning AMI (Ubuntu) Version 19.0 - ami-05bc59103c52af154
With latest deep learning frameworks pre-installed: MXNet, TensorFlow, PyTorch, Keras, Chainer, Caffe/2, Theano & CNTK, configured with NVIDIA CUDA, cuDNN, NCCL & Intel MKL-DNN. For a fully managed experience, check: https://aws.amazon.com/sagemaker
Root device type: ebs
Virtualization type: hvm
ENA Enabled: Yes
Select
64-bit (x86)

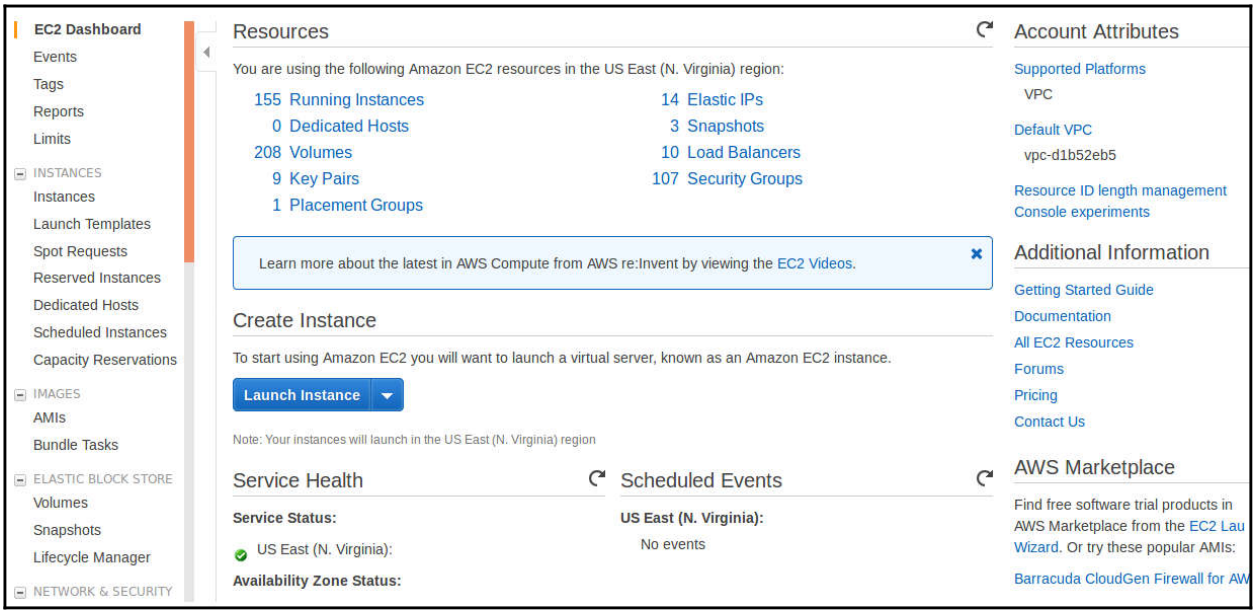

1. Choose AMI
2. Choose Instance Type
3. Configure Instance
4. Add Storage
5. Add Tags
6. Configure Security Group
7. Review
Step 2: Choose an Instance Type
General purpose a1.4xlarge 16 32 EBS only Yes Up to 10 Gigabit
Compute optimized c5n.large 2 5.25 EBS only Yes Up to 25 Gigabit
Compute optimized c5n.xlarge 4 10.5 EBS only Yes Up to 25 Gigabit
Compute optimized c5n.2xlarge 8 21 EBS only Yes Up to 25 Gigabit
Compute optimized c5n.4xlarge 16 42 EBS only Yes Up to 25 Gigabit
Compute optimized c5n.9xlarge 36 96 EBS only Yes 50 Gigabit
Compute optimized c5n.18xlarge 72 192 EBS only Yes 100 Gigabit
Compute optimized c5d.large 2 4 1 x 50 (SSD) Yes Up to 10 Gigabit
Compute optimized c5d.xlarge 4 8 1 x 100 (SSD) Yes Up to 10 Gigabit
Compute optimized c5d.2xlarge 8 16 1 x 200 (SSD) Yes Up to 10 Gigabit
Cancel
Previous
Review and Launch
Next: Configure Instance

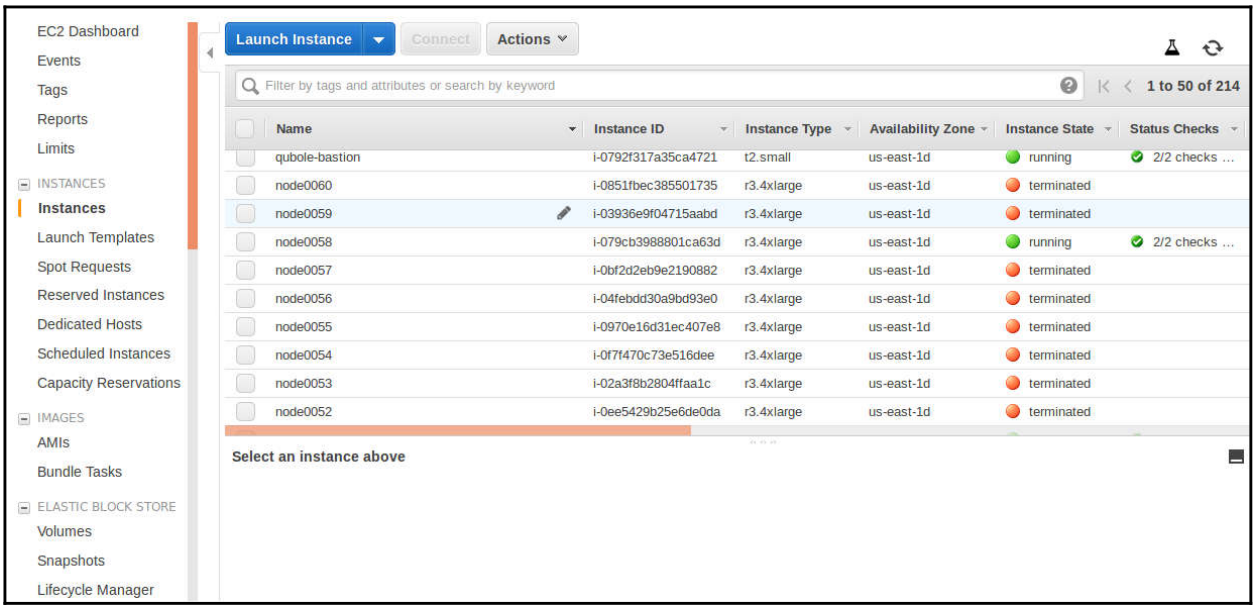

1. Choose AMI
2. Choose Instance Type
3. Configure Instance
4. Add Storage
5. Add Tags
6. Configure Security Group
7. Review
Step 3: Configure Instance Details
Configure the instance to suit your requirements. You can launch multiple instances from the same AMI, request Spot instances to take advantage of the lower pricing, assign an access manage
role to the instance, and more.
Number of instances
1
Launch into Auto Scaling Group
Purchasing option
Request Spot instances
Network
vpc-0b620306aa8f825d8 | Packt-vpc
Create new VPC
Subnet
subnet-023ae6c85c8a2c5d2 | packt-subnet | us-w
Create new subnet
251 IP Addresses available
Auto-assign Public IP
Use subnet setting (Disable)
Placement group
Add instance to placement group.
Capacity Reservation
Open
Create new Capacity Reservation
IAM role
None
Create new IAM role
Cancel
Previous
Review and Launch
Next: Add

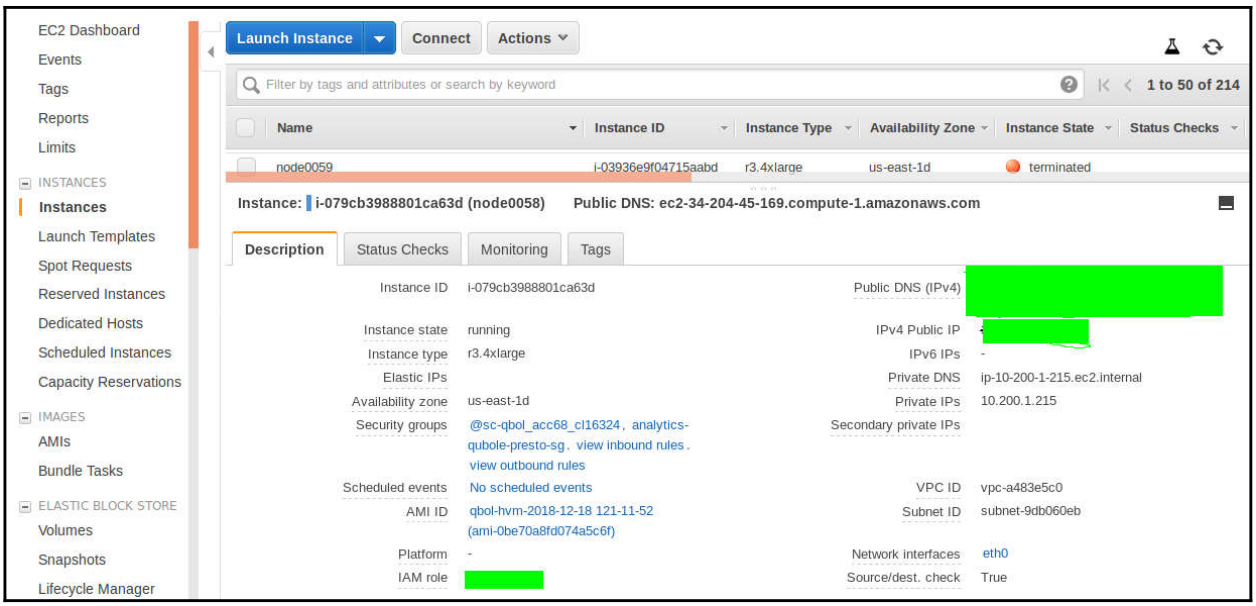

1. Choose AMI
2. Choose Instance Type
3. Configure Instance
4. Add Storage
5. Add Tags
6. Configure Security Group
7. Review
Step 4: Add Storage
Your instance will be launched with the following storage device settings. You can attach additional EBS volumes and instance store volumes to your instance, or edit the settings of the root volume. You can also attach additional EBS volumes after launching an instance, but not instance store volumes. Learn more about storage options in Amazon EC2.
Volume Type
Device
Snapshot
Size (GiB)
Volume Type
IOPS
Throughput (MB/s)
Delete on Termination
Encrypted
Root
/dev/sda1
snap-0252bea5b37202c35
8
General Purpose SSD (gp2)
100 / 3000
N/A
Not Encrypted
Add New Volume
Free tier eligible customers can get up to 30 GB of EBS General Purpose (SSD) or Magnetic storage. Learn more about free usage tier eligibility and usage restrictions.
Cancel
Previous
Review and Launch
Next: Add

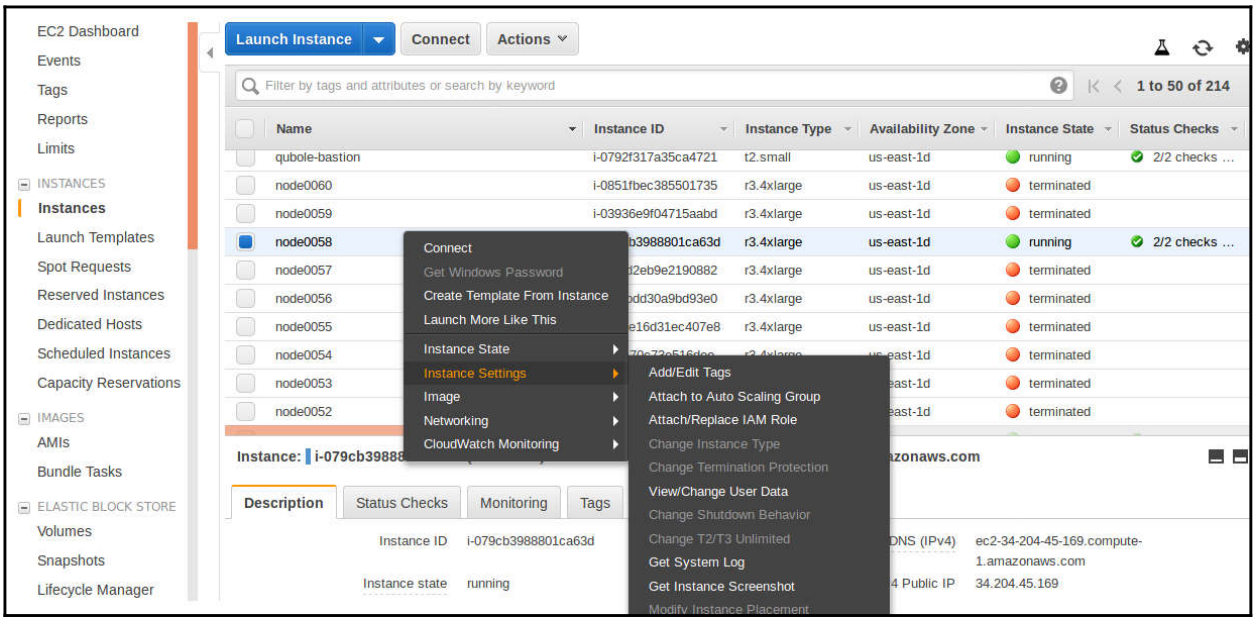

Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security
Software Configuration
Release
emr-5.19.0
Hadoop 2.8.5
JupyterHub 0.9.4
Ganglia 3.7.2
Hive 2.3.3
MXNet 1.3.0
Hue 4.2.0
Spark 2.3.2
Zeppelin 0.8.0
Tez 0.8.4
HBase 1.4.7
Presto 0.212
Sqoop 1.4.7
Phoenix 4.14.0
HCatalog 2.3.3
Livy 0.5.0
Flink 1.6.1
Pig 0.17.0
ZooKeeper 3.4.13
Mahout 0.13.0
Oozie 5.0.0
TensorFlow 1.11.0
AWS Glue Data Catalog settings (optional)
Use for Hive table metadata
Edit software settings
Enter configuration
Load JSON from S3
classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]
Add steps (optional)
Step type
Select a step
Configure

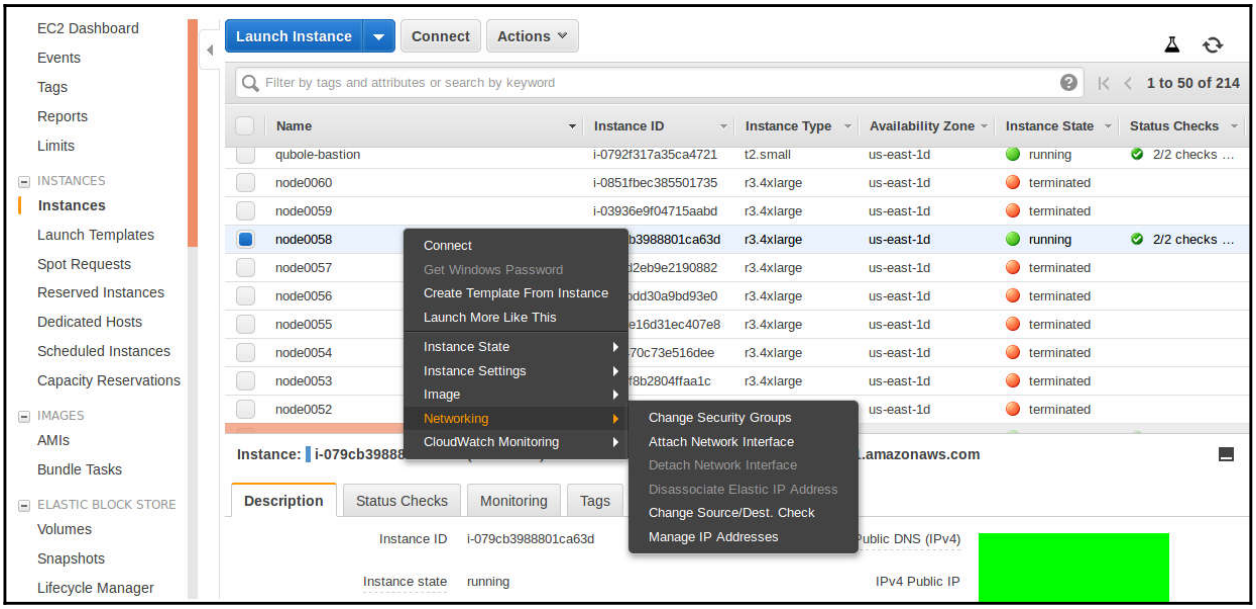

Create Cluster - Advanced Options
Go to quick options
Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security
Hardware Configuration
If you need more than 20 EC2 instances, see this topic.
Instance group configuration
Uniform instance groups
Specify a single instance type and purchasing option for each node type.
Instance fleets
Specify target capacity and how Amazon EMR fulfills it for each node type. Mix instance types and purchasing options. Learn more
Network
vpc-0b620306aa8f825d8 (10.20.0.0/16) | Packt-
Create a VPC
EC2 Subnet
subnet-023ae6c85c8a2c5d2 | packt-subnet | us-west-2a
No Amazon S3 endpoint or route to AWS public IP Range detected in subnet-023ae6c85c8a2c5d2. S3 endpoints are recommended for EMR access data in S3. For EMR to communicate with Amazon DynamoDB (and for EMRFS consistent view), AWS KMS, and Amazon Kinesis, your cluster must be able to reach this IP range. You can create a NAT instance to create a route for this traffic. Learn more
Add S3 endpoint and NAT instance
Root device EBS volume size
10
GiB

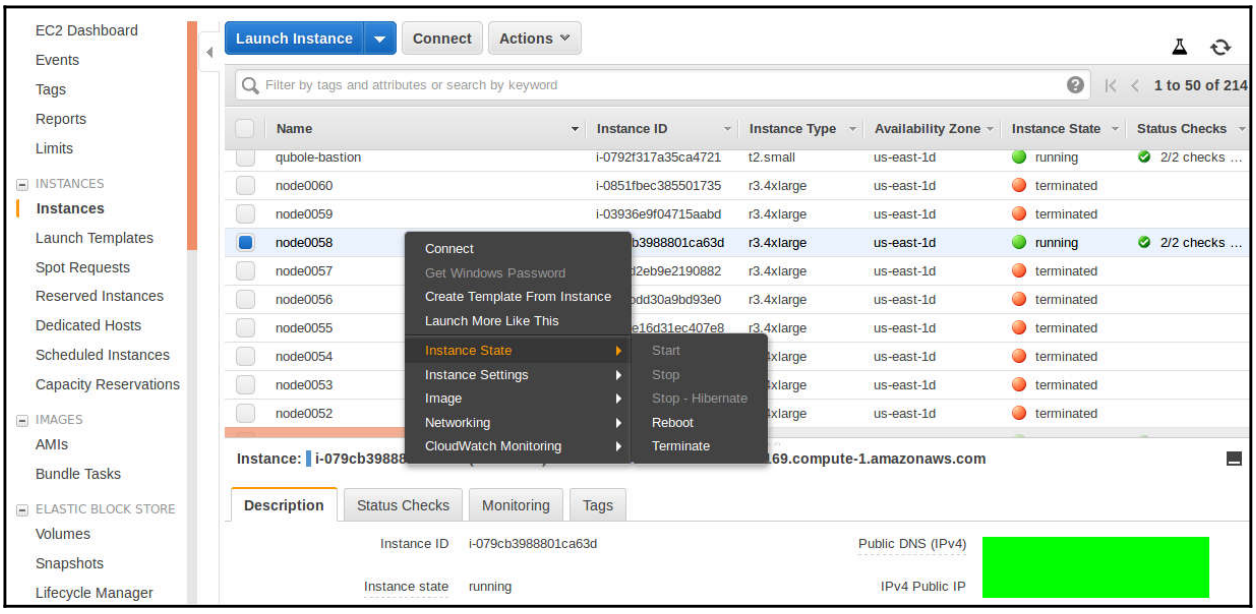

EC2 Dashboard
Events
Tags
Reports
Limits
INSTANCES
Instances
Launch Templates
Spot Requests
Reserved Instances
Dedicated Hosts
Scheduled Instances
Capacity Reservations
IMAGES
AMIs
Bundle Tasks
ELASTIC BLOCK STORE
Volumes
Snapshots
Lifecycle Manager
NETWORK & SECURITY
Resources
You are using the following Amazon EC2 resources in the US East (N. Virginia) region:
155 Running Instances
0 Dedicated Hosts
208 Volumes
9 Key Pairs
1 Placement Groups
14 Elastic IPs
3 Snapshots
10 Load Balancers
107 Security Groups
Learn more about the latest in AWS Compute from AWS re:Invent by viewing the EC2 Videos.
Create Instance
To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance.
Launch Instance
Note: Your instances will launch in the US East (N. Virginia) region
Service Health
Service Status:
US East (N. Virginia):
Availability Zone Status:
Scheduled Events
US East (N. Virginia):
No events
Account Attributes
Supported Platforms
VPC
Default VPC
vpc-d1b52eb5
Resource ID length management
Console experiments
Additional Information
Getting Started Guide
Documentation
All EC2 Resources
Forums
Pricing
Contact Us
AWS Marketplace
Find free software trial products in AWS Marketplace from the EC2 Lau
Wizard. Or try these popular AMIs:
Barracuda CloudGen Firewall for AW

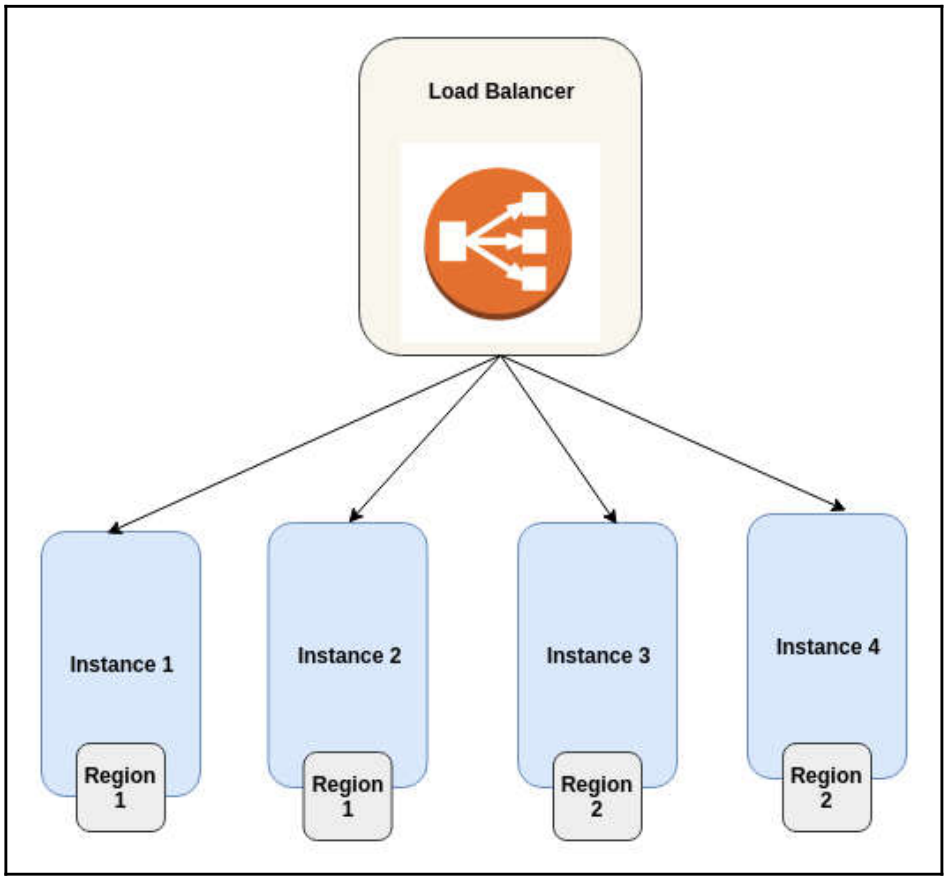

EC2 Dashboard
Events
Tags
Reports
Limits
INSTANCES
Instances
Launch Templates
Spot Requests
Reserved Instances
Dedicated Hosts
Scheduled Instances
Capacity Reservations
IMAGES
AMIs
Bundle Tasks
ELASTIC BLOCK STORE
Volumes
Snapshots
Lifecycle Manager
Launch Instance
Connect
Actions
Filter by tags and attributes or search by keyword
1 to 50 of 214
Name
Instance ID
Instance Type
Availability Zone
Instance State
Status Checks
qubole-bastion i-0792f317a35ca4721 t2.small us-east-1d running 2/2 checks ...
node0060 i-0851fbec385501735 r3.4xlarge us-east-1d terminated
node0059 i-03936e9f04715aabd r3.4xlarge us-east-1d terminated
node0058 i-079cb3988801ca63d r3.4xlarge us-east-1d running 2/2 checks ...
node0057 i-0bf2d2eb9e2190882 r3.4xlarge us-east-1d terminated
node0056 i-04febdd30a9bd93e0 r3.4xlarge us-east-1d terminated
node0055 i-0970e16d31ec407e8 r3.4xlarge us-east-1d terminated
node0054 i-0f7f470c73e516dee r3.4xlarge us-east-1d terminated
node0053 i-02a3f8b2804ffaa1c r3.4xlarge us-east-1d terminated
node0052 i-0ee5429b25e6de0da r3.4xlarge us-east-1d terminated
Select an instance above

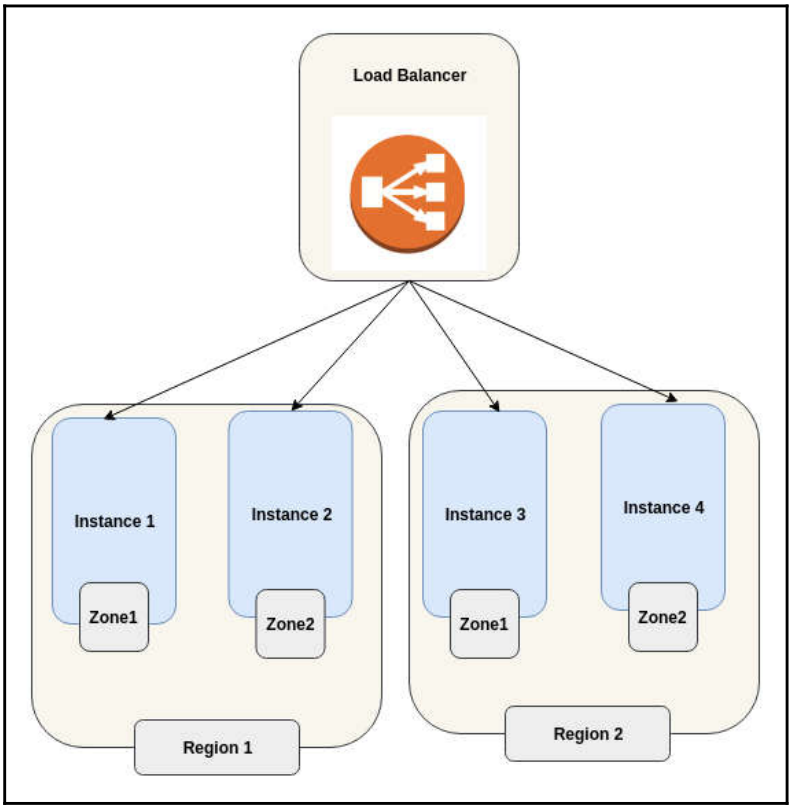

EC2 Dashboard
Events
Tags
Reports
Limits
INSTANCES
Instances
Launch Templates
Spot Requests
Reserved Instances
Dedicated Hosts
Scheduled Instances
Capacity Reservations
IMAGES
AMIs
Bundle Tasks
ELASTIC BLOCK STORE
Volumes
Snapshots
Lifecycle Manager
Launch Instance
Connect
Actions
Filter by tags and attributes or search by keyword
1 to 50 of 214
Name
Instance ID
Instance Type
Availability Zone
Instance State
Status Checks
node0059 i-03936e9f04715aabd r3.4xlarge us-east-1d terminated
Instance: i-079cb3988801ca63d (node0058) Public DNS: ec2-34-204-45-169.compute-1.amazonaws.com
Description
Status Checks
Monitoring
Tags
Instance ID i-079cb3988801ca63d
Instance state running
Instance type r3.4xlarge
Elastic IPs
Availability zone us-east-1d
Security groups @sc-qbol_acc68_cl16324, analytics-qubole-presto-sg. view inbound rules. view outbound rules
Scheduled events No scheduled events
AMI ID qbol-hvm-2018-12-18 121-11-52 (ami-0be70a8fd074a5c6f)
Platform -
IAM role
Public DNS (IPv4)
IPv4 Public IP
IPv6 IPs -
Private DNS ip-10-200-1-215.ec2.internal
Private IPs 10.200.1.215
Secondary private IPs
VPC ID vpc-a483e5c0
Subnet ID subnet-9db060eb
Network interfaces eth0
Source/dest. check True


EC2 Dashboard
Events
Tags
Reports
Limits
INSTANCES
Instances
Launch Templates
Spot Requests
Reserved Instances
Dedicated Hosts
Scheduled Instances
Capacity Reservations
IMAGES
AMIs
Bundle Tasks
ELASTIC BLOCK STORE
Volumes
Snapshots
Lifecycle Manager
Launch Instance
Connect
Actions
Filter by tags and attributes or search by keyword
1 to 50 of 214
Name
Instance ID
Instance Type
Availability Zone
Instance State
Status Checks
qubole-bastion i-0792f317a35ca4721 t2.small us-east-1d running 2/2 checks ...
node0060 i-0851fbec385501735 r3.4xlarge us-east-1d terminated
node0059 i-03936e9f04715aabd r3.4xlarge us-east-1d terminated
node0058 b3988801ca63d r3.4xlarge us-east-1d running 2/2 checks ...
node0057 d2eb9e2190882 r3.4xlarge us-east-1d terminated
node0056 dd30a9bd93e0 r3.4xlarge us-east-1d terminated
node0055 e16d31ec407e8 r3.4xlarge us-east-1d terminated
node0054 r3.4xlarge terminated
node0053 terminated
node0052 terminated
Connect
Get Windows Password
Create Template From Instance
Launch More Like This
Instance State
Instance Settings
Image
Networking
CloudWatch Monitoring
Add/Edit Tags
Attach to Auto Scaling Group
Attach/Replace IAM Role
Change Instance Type
Change Termination Protection
View/Change User Data
Change Shutdown Behavior
Change T2/T3 Unlimited
Get System Log
Get Instance Screenshot
Modify Instance Placement
Instance: i-079cb39888
azonaws.com
Description
Status Checks
Monitoring
Tags
Instance ID i-079cb3988801ca63d
Instance state running
DNS (IPv4) ec2-34-204-45-169.compute-1.amazonaws.com
4 Public IP 34.204.45.169


EC2 Dashboard
Events
Tags
Reports
Limits
INSTANCES
Instances
Launch Templates
Spot Requests
Reserved Instances
Dedicated Hosts
Scheduled Instances
Capacity Reservations
IMAGES
AMIs
Bundle Tasks
ELASTIC BLOCK STORE
Volumes
Snapshots
Lifecycle Manager
Launch Instance
Connect
Actions
Filter by tags and attributes or search by keyword
1 to 50 of 214
Name
Instance ID
Instance Type
Availability Zone
Instance State
Status Checks
qubole-bastion i-0792f317a35ca4721 t2.small us-east-1d running 2/2 checks ...
node0060 i-0851fbec385501735 r3.4xlarge us-east-1d terminated
node0059 i-03936e9f04715aabd r3.4xlarge us-east-1d terminated
node0058 b3988801ca63d r3.4xlarge us-east-1d running 2/2 checks ...
node0057 d2eb9e2190882 r3.4xlarge us-east-1d terminated
node0056 dd30a9bd93e0 r3.4xlarge us-east-1d terminated
node0055 e16d31ec407e8 r3.4xlarge us-east-1d terminated
node0054 70c73e516dee r3.4xlarge us-east-1d terminated
node0053 f8b2804ffaa1c r3.4xlarge us-east-1d terminated
node0052 us-east-1d terminated
Connect
Get Windows Password
Create Template From Instance
Launch More Like This
Instance State
Instance Settings
Image
Networking
CloudWatch Monitoring
Change Security Groups
Attach Network Interface
Detach Network Interface
Disassociate Elastic IP Address
Change Source/Dest. Check
Manage IP Addresses
Instance: i-079cb39888
.amazonaws.com
Description
Status Checks
Monitoring
Tags
Instance ID i-079cb3988801ca63d
Instance state running
Public DNS (IPv4)
IPv4 Public IP

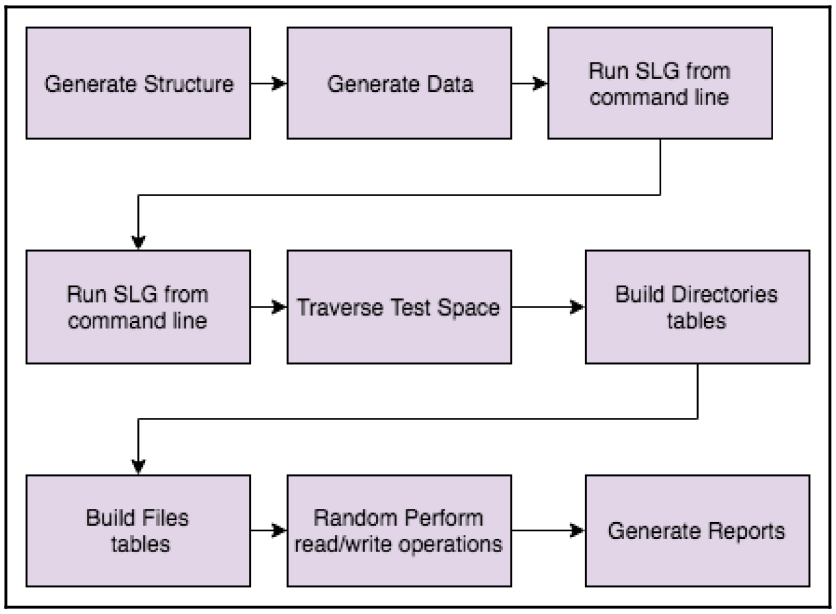

EC2 Dashboard
Events
Tags
Reports
Limits
INSTANCES
Instances
Launch Templates
Spot Requests
Reserved Instances
Dedicated Hosts
Scheduled Instances
Capacity Reservations
IMAGES
AMIs
Bundle Tasks
ELASTIC BLOCK STORE
Volumes
Snapshots
Lifecycle Manager
Launch Instance
Connect
Actions
Filter by tags and attributes or search by keyword
1 to 50 of 214
Name
Instance ID
Instance Type
Availability Zone
Instance State
Status Checks
qubole-bastion
i-0792f317a35ca4721
t2.small
us-east-1d
running
2/2 checks ...
node0060
i-0851fbec385501735
r3.4xlarge
us-east-1d
terminated
node0059
i-03936e9f04715aabd
r3.4xlarge
us-east-1d
terminated
node0058
r3.4xlarge
us-east-1d
running
2/2 checks ...
node0057
r3.4xlarge
us-east-1d
terminated
node0056
r3.4xlarge
us-east-1d
terminated
node0055
r3.4xlarge
us-east-1d
terminated
node0054
us-east-1d
terminated
node0053
us-east-1d
terminated
node0052
us-east-1d
terminated
Connect
Get Windows Password
Create Template From Instance
Launch More Like This
Instance State
Instance Settings
Image
Networking
CloudWatch Monitoring
Start
Stop
Stop - Hibernate
Reboot
Terminate
Instance:
Description
Status Checks
Monitoring
Tags
Instance ID
i-079cb3988801ca63d
Instance state
running
Public DNS (IPv4)
IPv4 Public IP

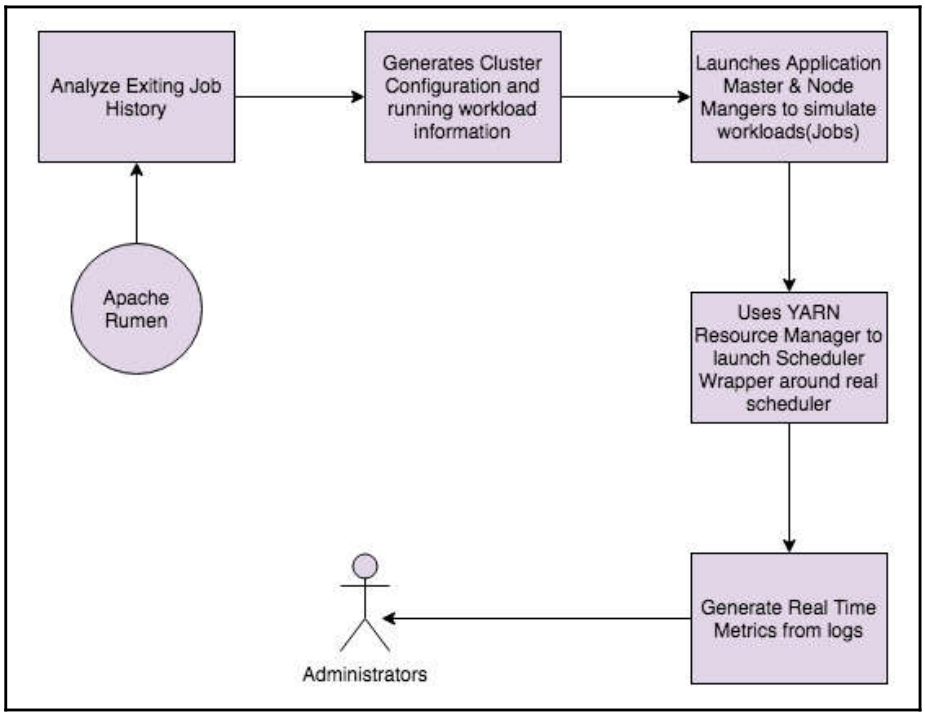

Load Balancer
Instance 1
Region 1
Instance 2
Region 1
Instance 3
Region 2
Instance 4
Region 2


Load Balancer
Instance 1
Zone1
Instance 2
Zone2
Region 1
Instance 3
Zone1
Instance 4
Zone2
Region 2

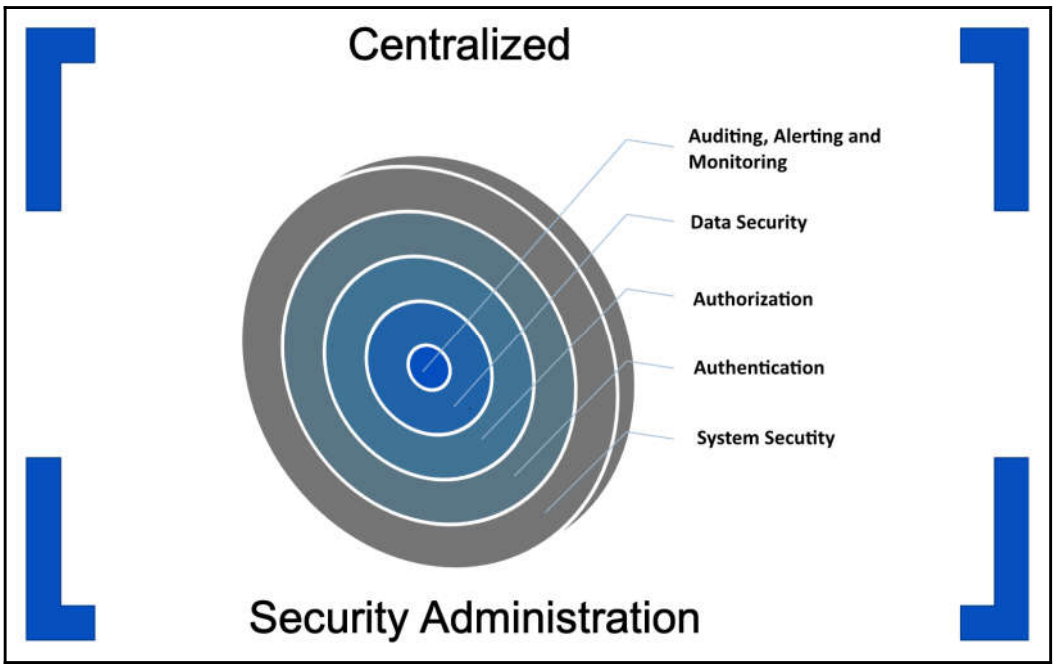
Chapter 12: Hadoop Cluster Profiling

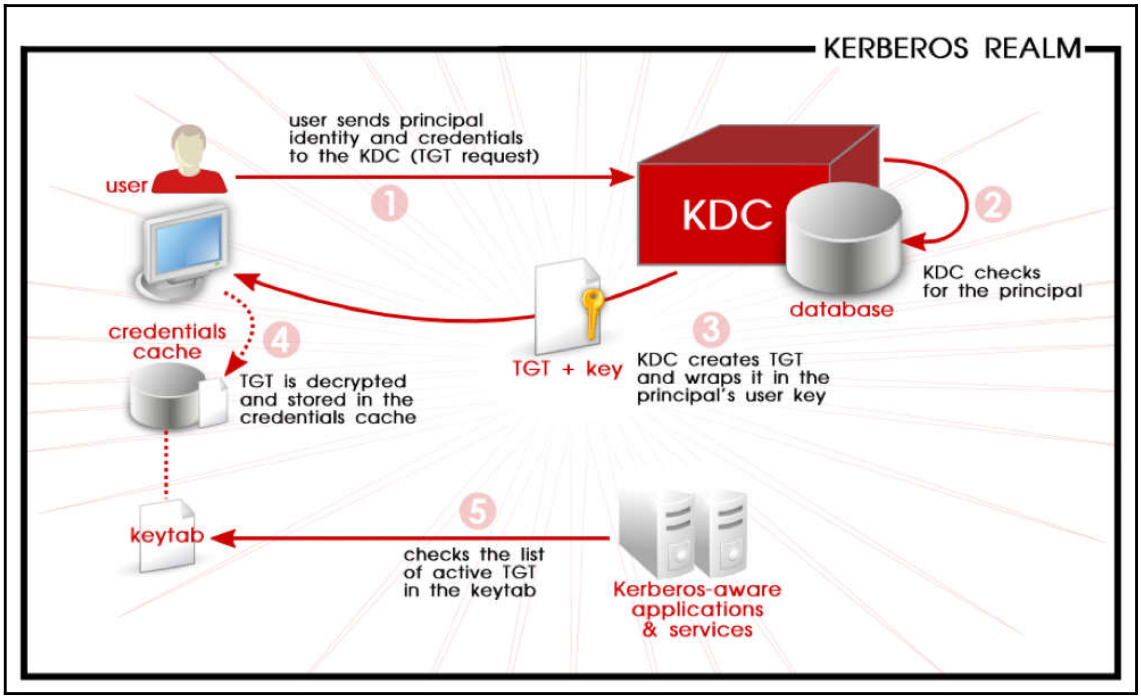
Benchmarking Hadoop Cluster		
HDFS	TestDFSIO	
Namenode	NNBench	NNThroughputBenchmark
	Synthetic Load Generator (SLG)	
YARN	YARN Scheduler Load Simulator (SLS)	
HIVE	TPC-DS	TPC-H
MIX-WORKLOADS	GRIDMIX	RUMEN

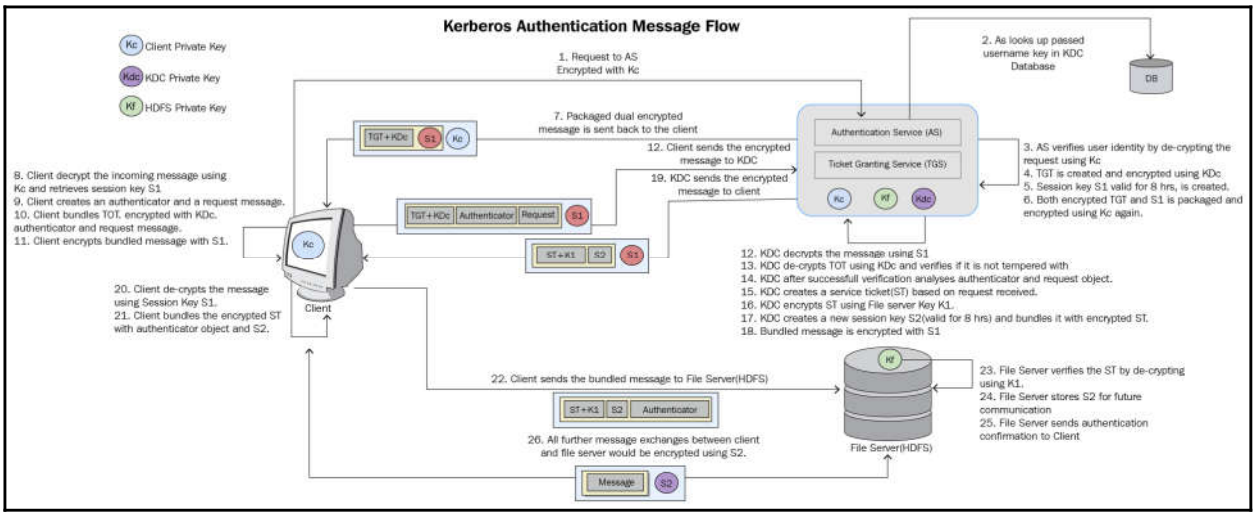


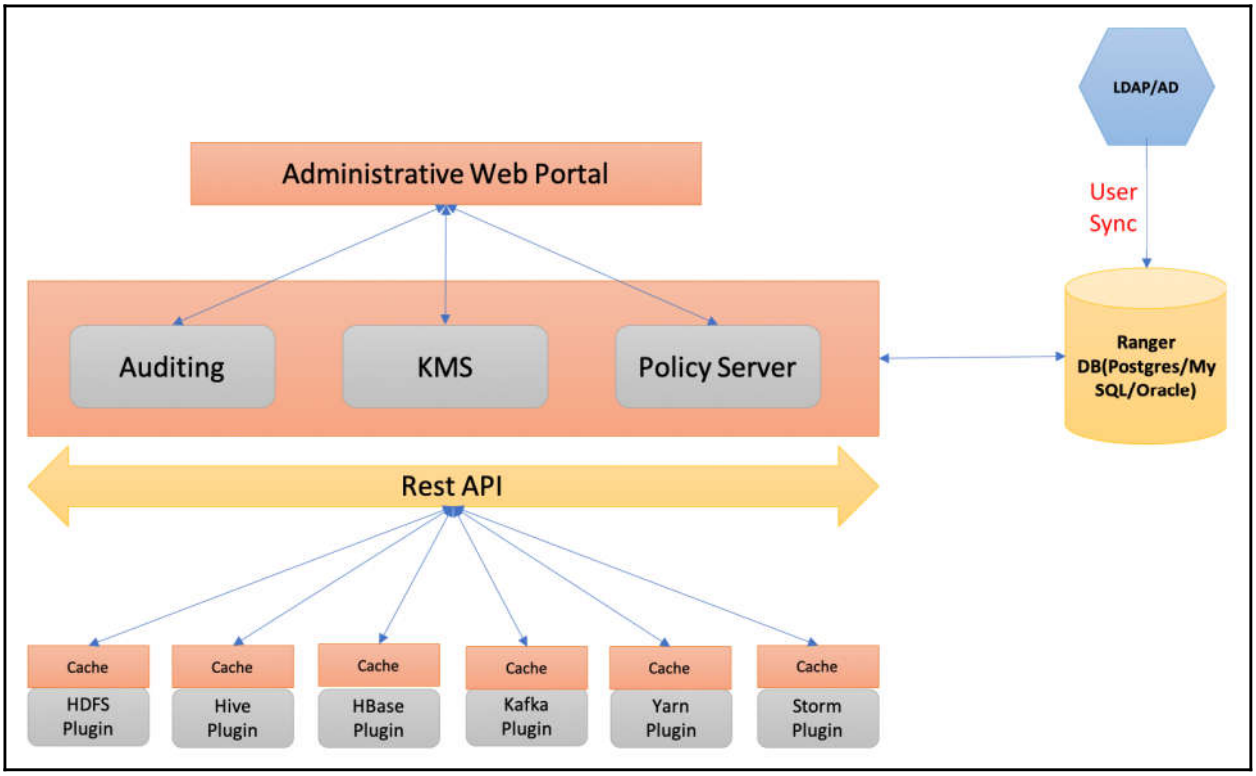

Analyze Exiting Job History
Generates Cluster Configuration and running workload information
Launches Application Master & Node Mangers to simulate workloads(Jobs)
Apache Rumen
Uses YARN Resource Manager to launch Scheduler Wrapper around real scheduler
Generate Real Time Metrics from logs
Administrators

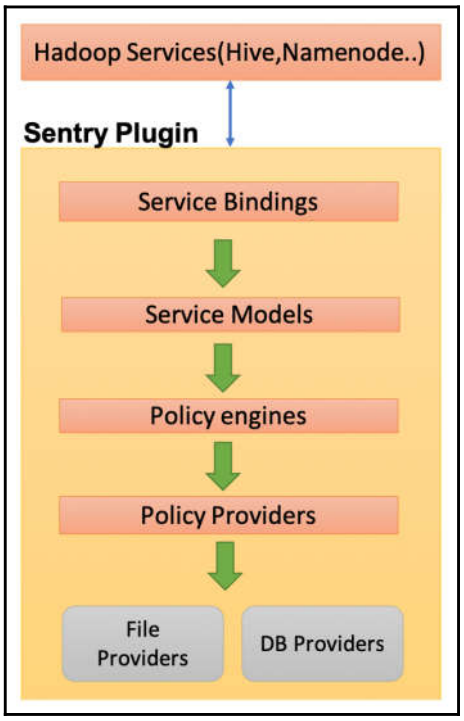
Chapter 13: Who Can Do What in Hadoop

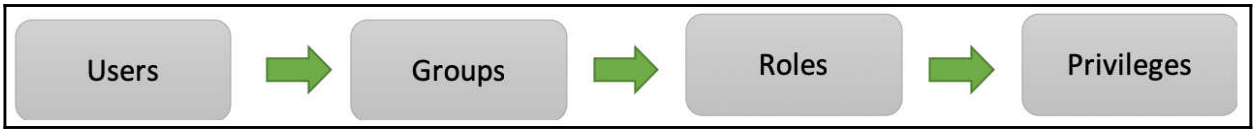




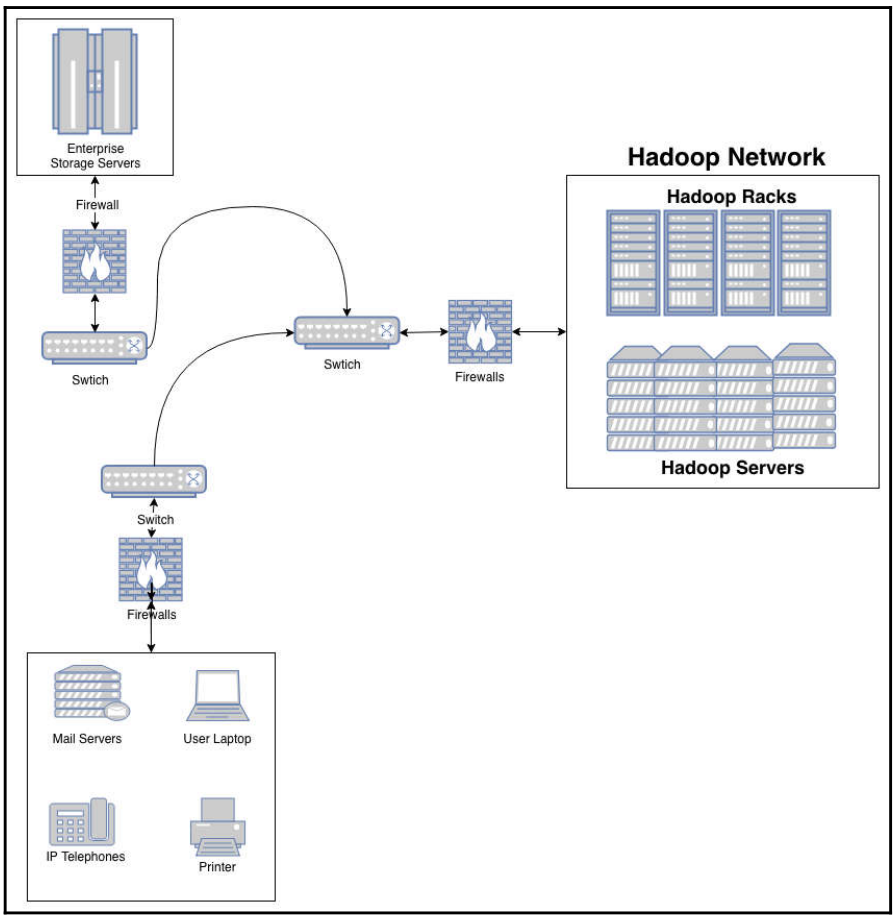

Kerberos Authentication Message Flow
Kc Client Private Key
Kdc KDC Private Key
Kf HDFS Private Key
1. Request to AS
Encrypted with Kc
2. As looks up passed
username key in KDC
Database
DB
7. Packaged dual encrypted
message is sent back to the client
TGT+KDc
S1
Kc
Authentication Service (AS)
Ticket Granting Service (TGS)
Kc
Kf
Kdc
12. Client sends the encrypted
message to KDC
19. KDC sends the encrypted
message to client
3. AS verifies user identity by de-crypting the
request using Kc
4. TGT is created and encrypted using KDc
5. Session key S1 valid for 8 hrs, is created.
6. Both encrypted TGT and S1 is packaged and
encrypted using Kc again.
8. Client decrypt the incoming message using
Kc and retrieves session key S1
9. Client creates an authenticator and a request message.
10. Client bundles TOT. encrypted with KDc.
authenticator and request message.
11. Client encrypts bundled message with S1.
TGT+KDc
Authenticator
Request
S1
ST+K1
S2
S3
Kc
Client
12. KDC decrypts the message using S1
13. KDC de-crypts TOT using KDc and verifies if it is not tempered with
14. KDC after successfull verification analyses authenticator and request object.
15. KDC creates a service ticket(ST) based on request received.
16. KDC encrypts ST using File server Key K1.
17. KDC creates a new session key S2(valid for 8 hrs) and bundles it with encrypted ST.
18. Bundled message is encrypted with S1
20. Client de-crypts the message
using Session Key S1.
21. Client bundles the encrypted ST
with authenticator object and S2.
Kf
22. Client sends the bundled message to File Server(HDFS)
ST+K1
S2
Authenticator
23. File Server verifies the ST by de-crypting
using K1.
24. File Server stores S2 for future
communication
25. File Server sends authentication
confirmation to Client
File Server(HDFS)
26. All further message exchanges between client
and file server would be encrypted using S2.
Message
S2

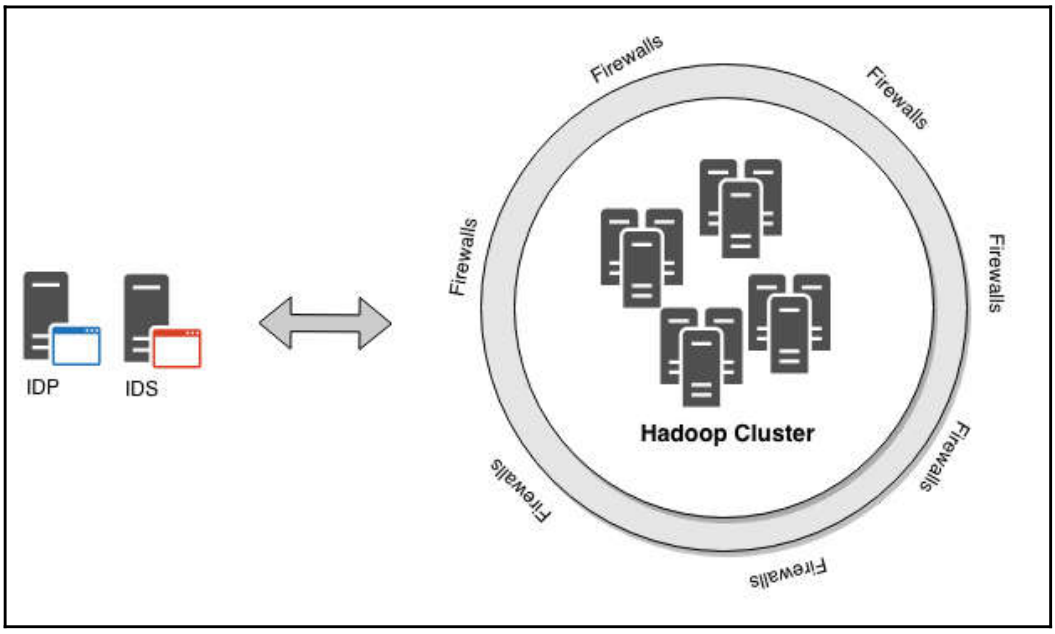

LDAP/AD
User
Sync
Administrative Web Portal
Auditing
KMS
Policy Server
Ranger
DB(Postgres/My
SQL/Oracle)
Rest API
Cache
HDFS
Plugin
Cache
Hive
Plugin
Cache
HBase
Plugin
Cache
Kafka
Plugin
Cache
Yarn
Plugin
Cache
Storm
Plugin

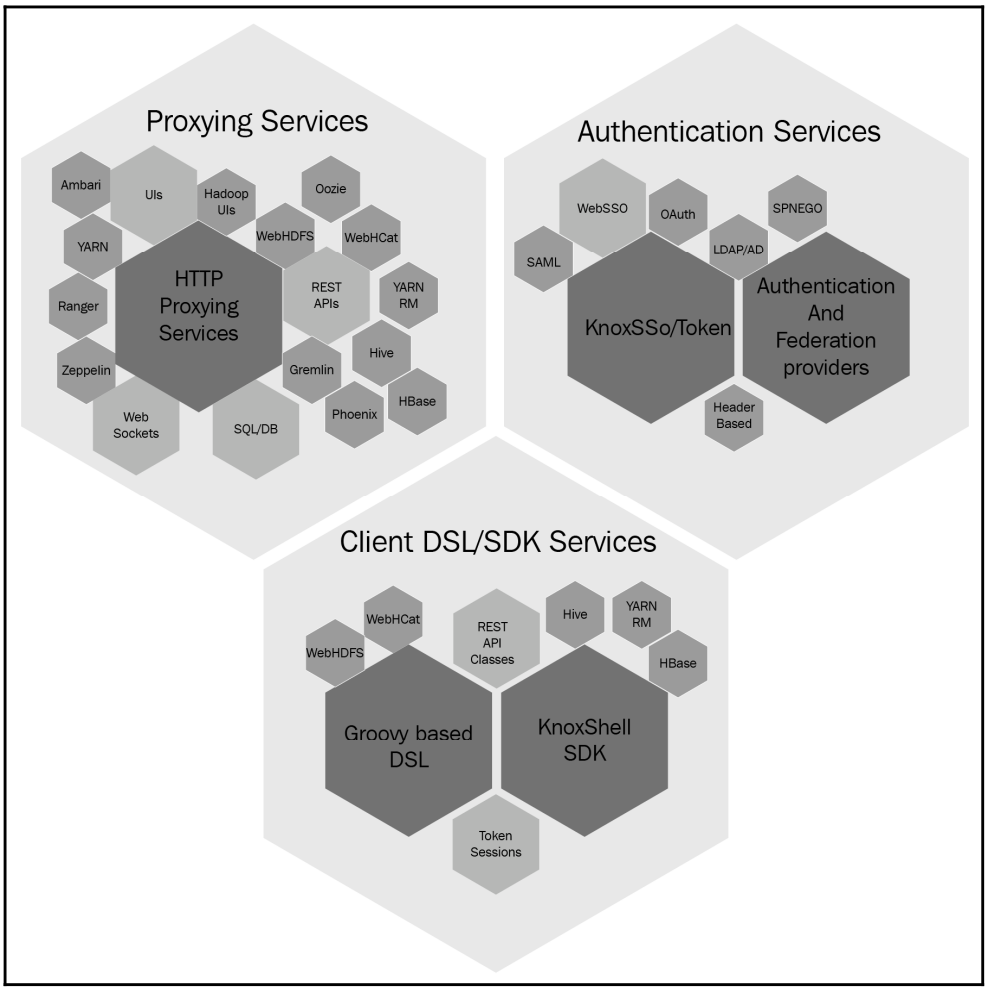

Hadoop Services(Hive,Namenode..)
Sentry Plugin
Service Bindings
Service Models
Policy engines
Policy Providers
File Providers
DB Providers


Users
Groups
Roles
Privileges

Chapter 14: Network and Data Security

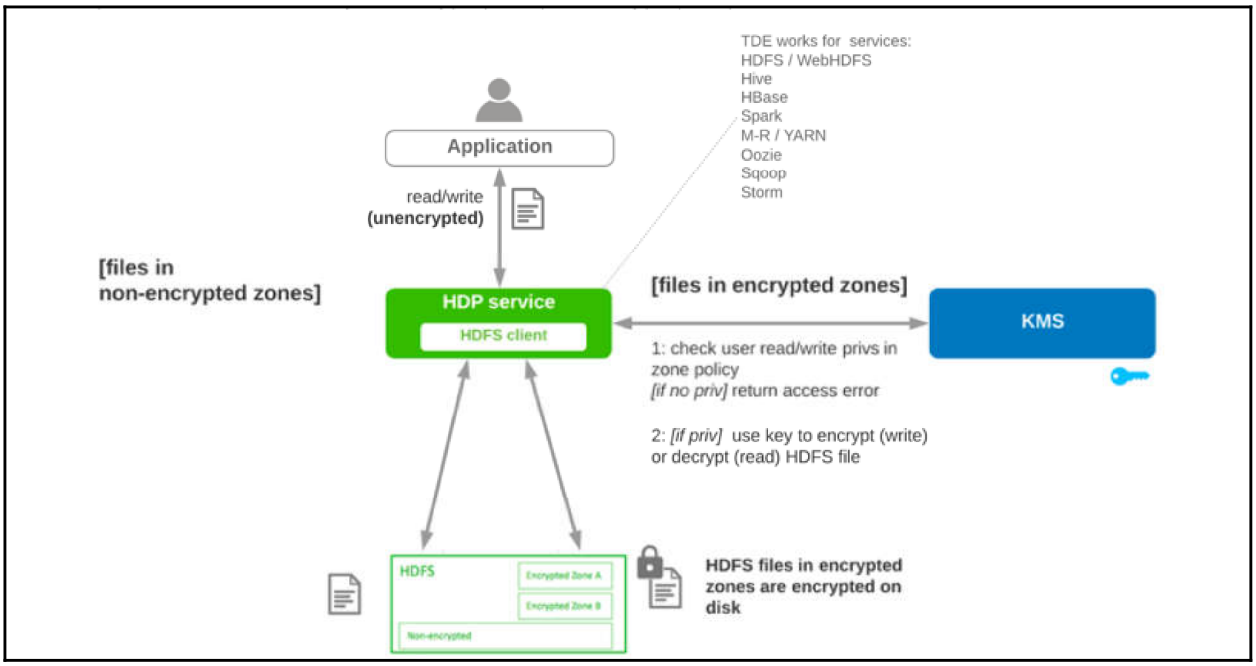



IDP
IDS
Firewalls
Firewalls
Firewalls
Firewalls
Firewalls
Firewalls
Hadoop Cluster



Hadoop Data In Transit Security

	Transit Protocols	
RPC	**TCP/IP**	**HTTP**
Mapreduce	**HDFS CLIENTS**	**Mapreduce Shuffles**
JobTracker		
TaskTracker		**Web Interfaces**
NameNode		



Data Masking

ID	Employee Name	Employee Salary
1	John	10,000
2	Tim	20,000

ID	Employee Name	Employee Salary
1	John	XXXXX
2	Tim	XXXXX

Data Masking By Random Substitution

ID	Employee Name	Employee Salary
1	John	10,000
2	Tim	20,000

ID	Employee Name	Employee Salary
1	John	5000
2	Tim	1000

Data Masking By Encryption

ID	Employee Name	Employee Salary
1	John	10,000
2	Tim	20,000

ID	Employee Name	Employee Salary
1	John	AB2H345EDNE98TYUO
2	Tim	SDF2096FT32UO7I9OP

Row Level Filtering

ID	Employee Name	Employee Salary
1	John	10,000
2	Tim	20,000

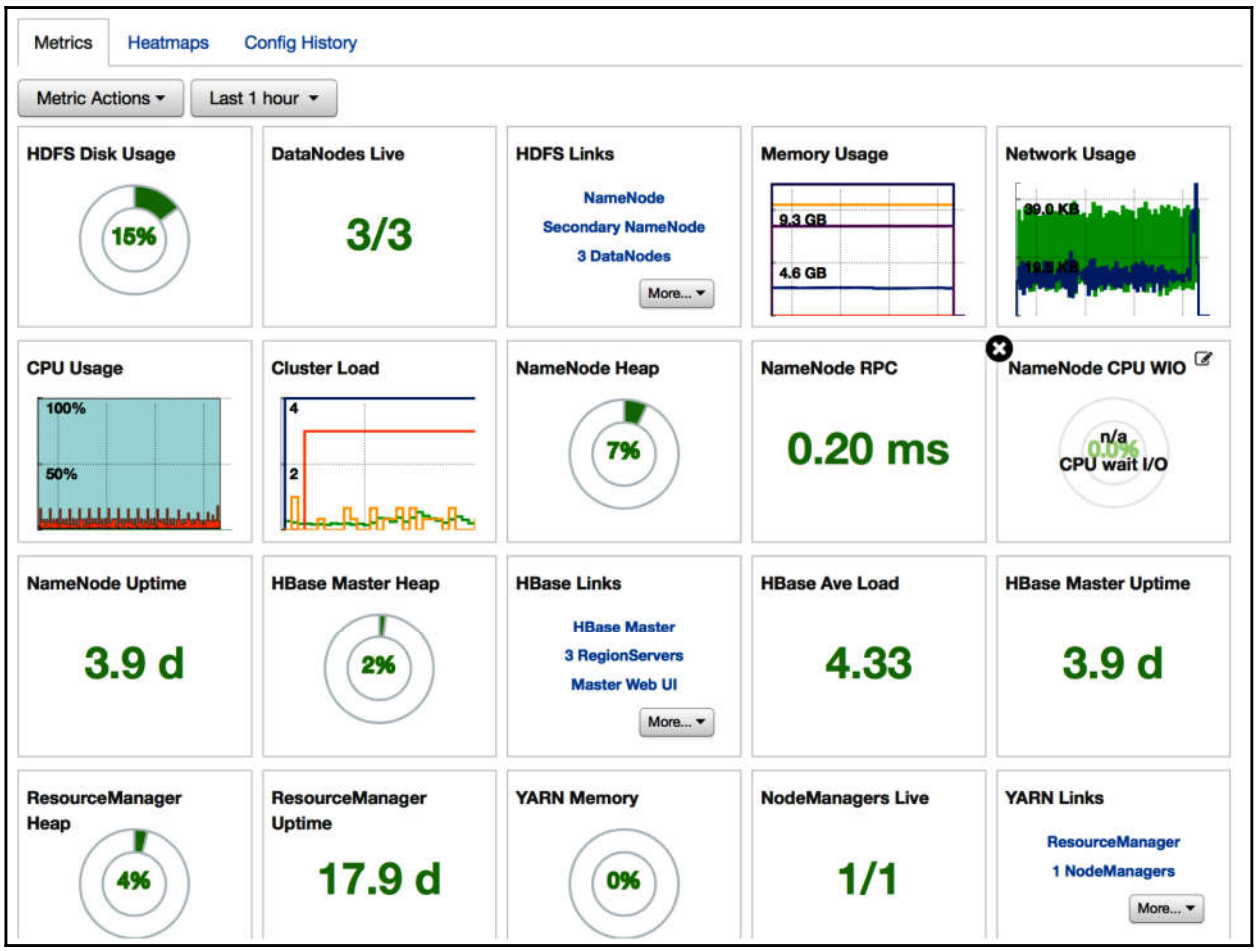
ID	Employee Name	Employee Salary
1	John	10,000

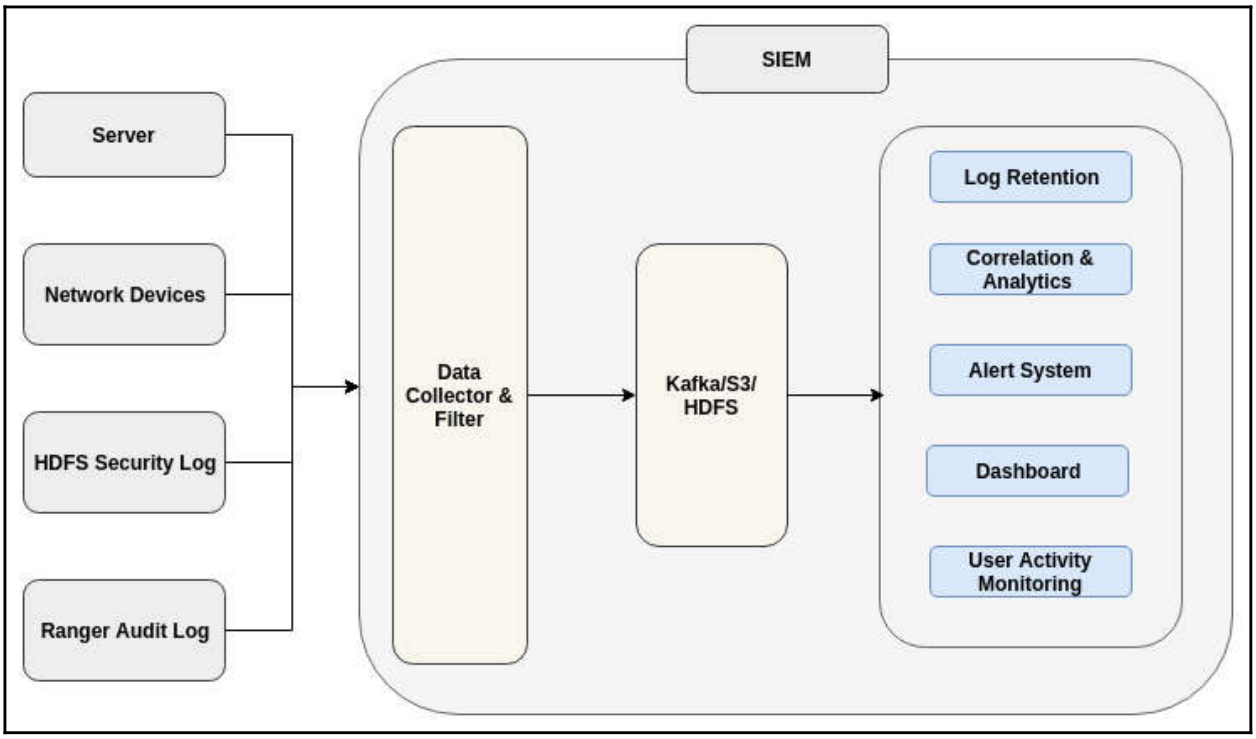
Data Masking By Random Substitution

ID	Employee Name	Employee Salary
1	John	10,000
2	Tim	20,000

ID	Employee Name
1	John
2	Tim

Chapter 15: Monitoring Hadoop




SIEM
Server
Network Devices
HDFS Security Log
Ranger Audit Log
Data Collector & Filter
Kafka/S3/ HDFS
Log Retention
Correlation & Analytics
Alert System
Dashboard
User Activity Monitoring