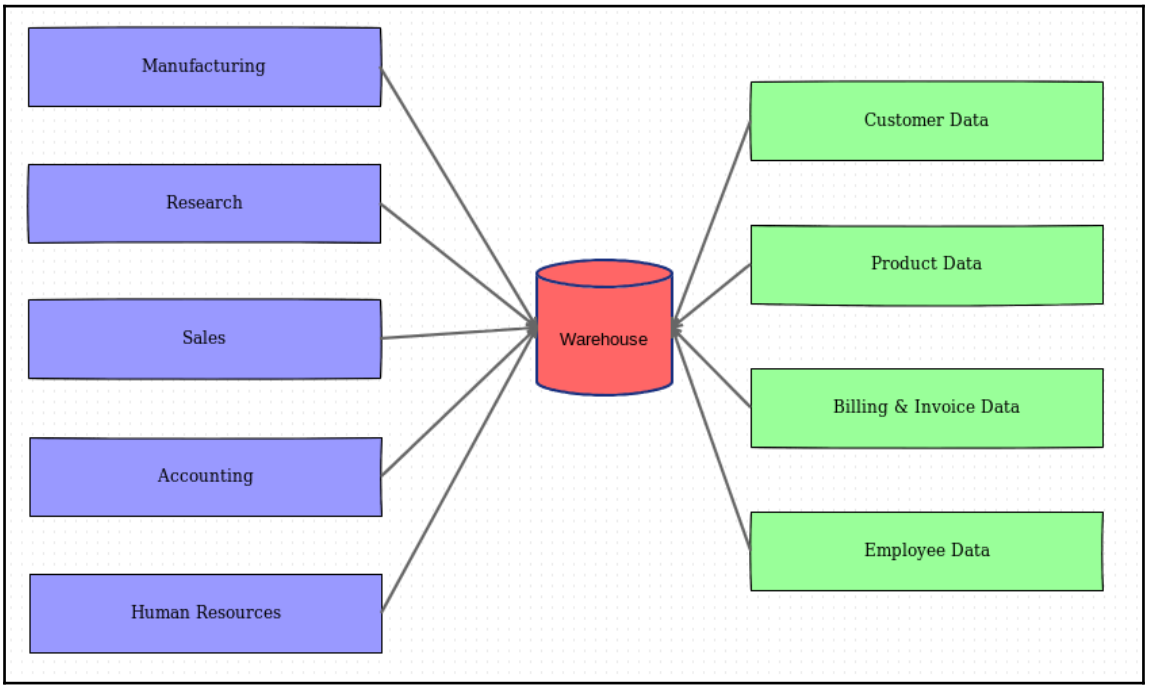
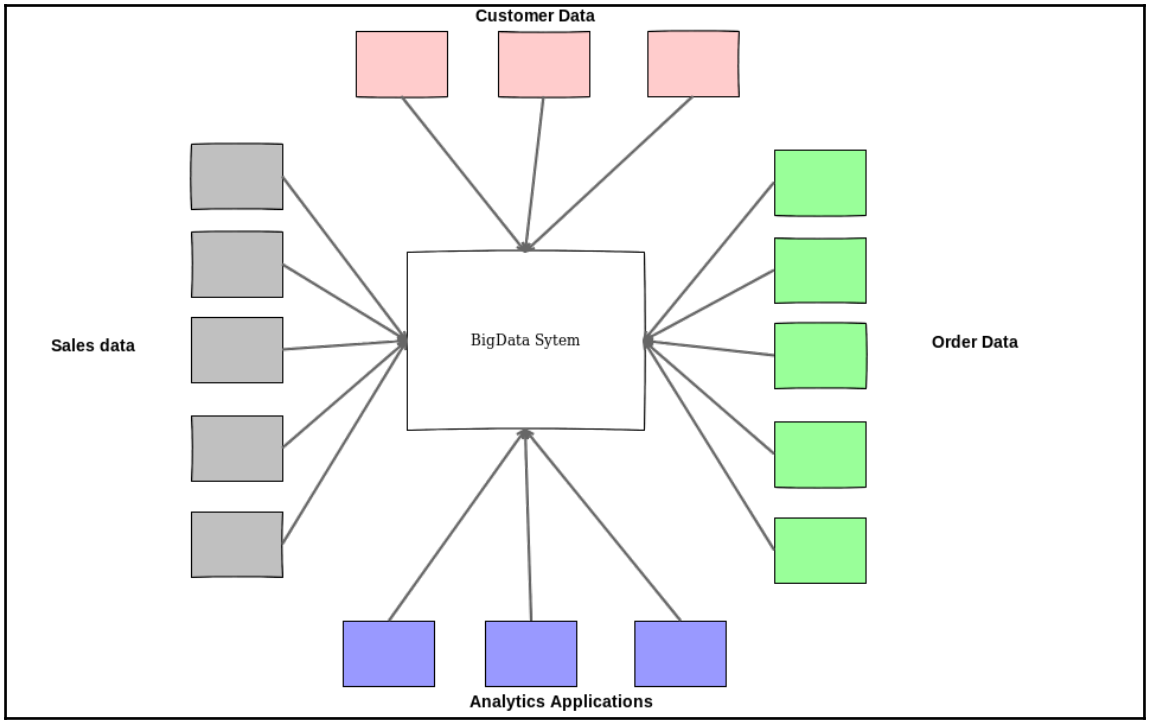
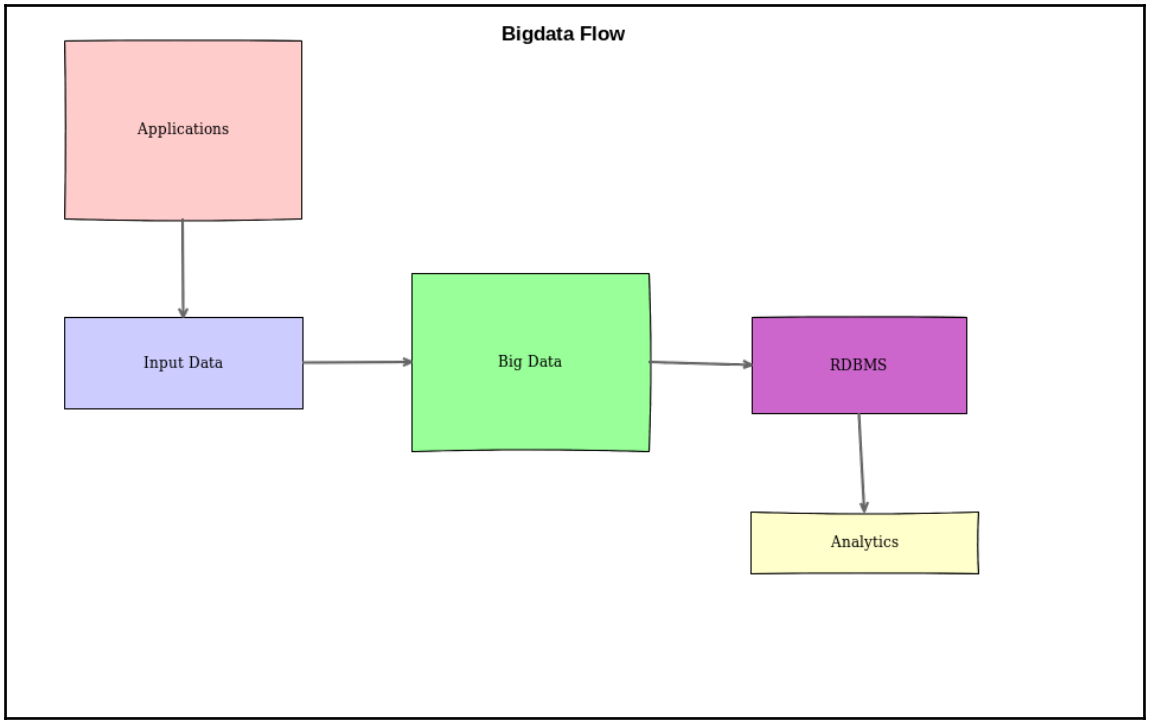
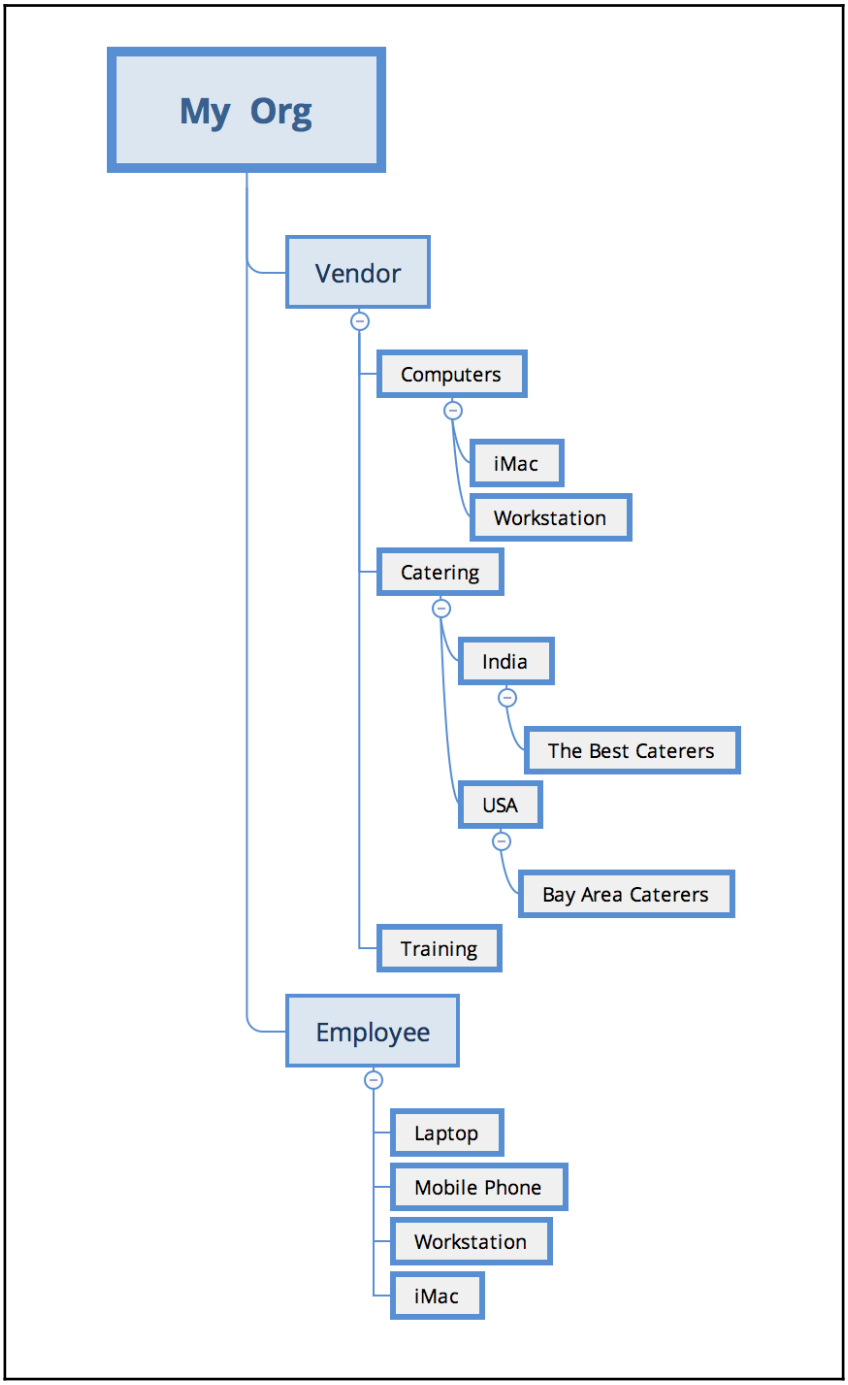
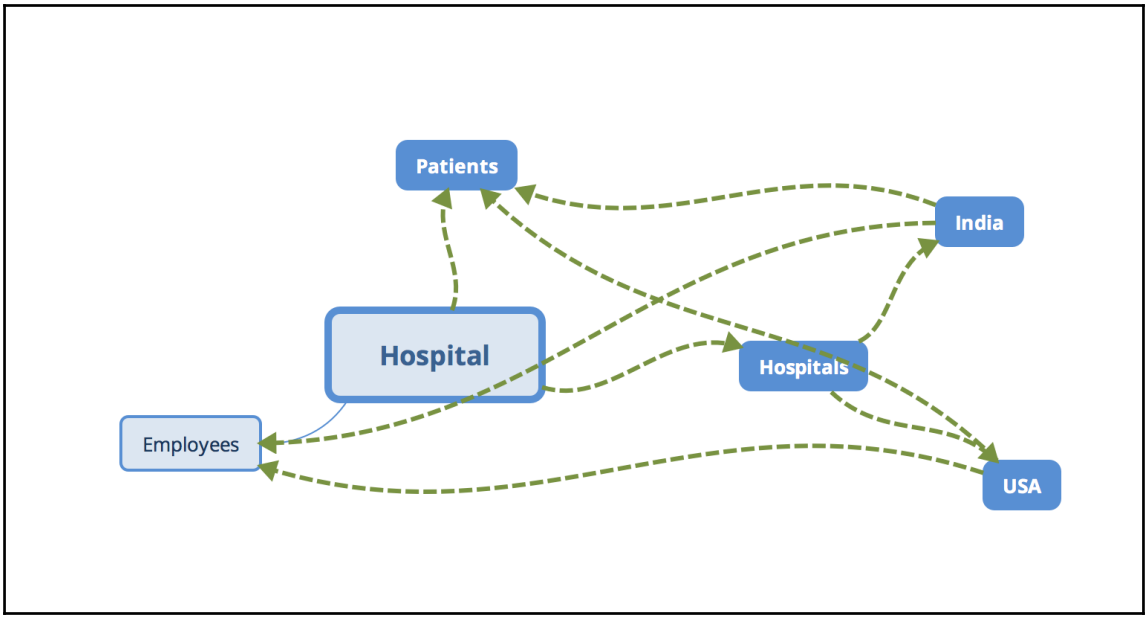
Chapter 1: Enterprise Data Architecture Principles

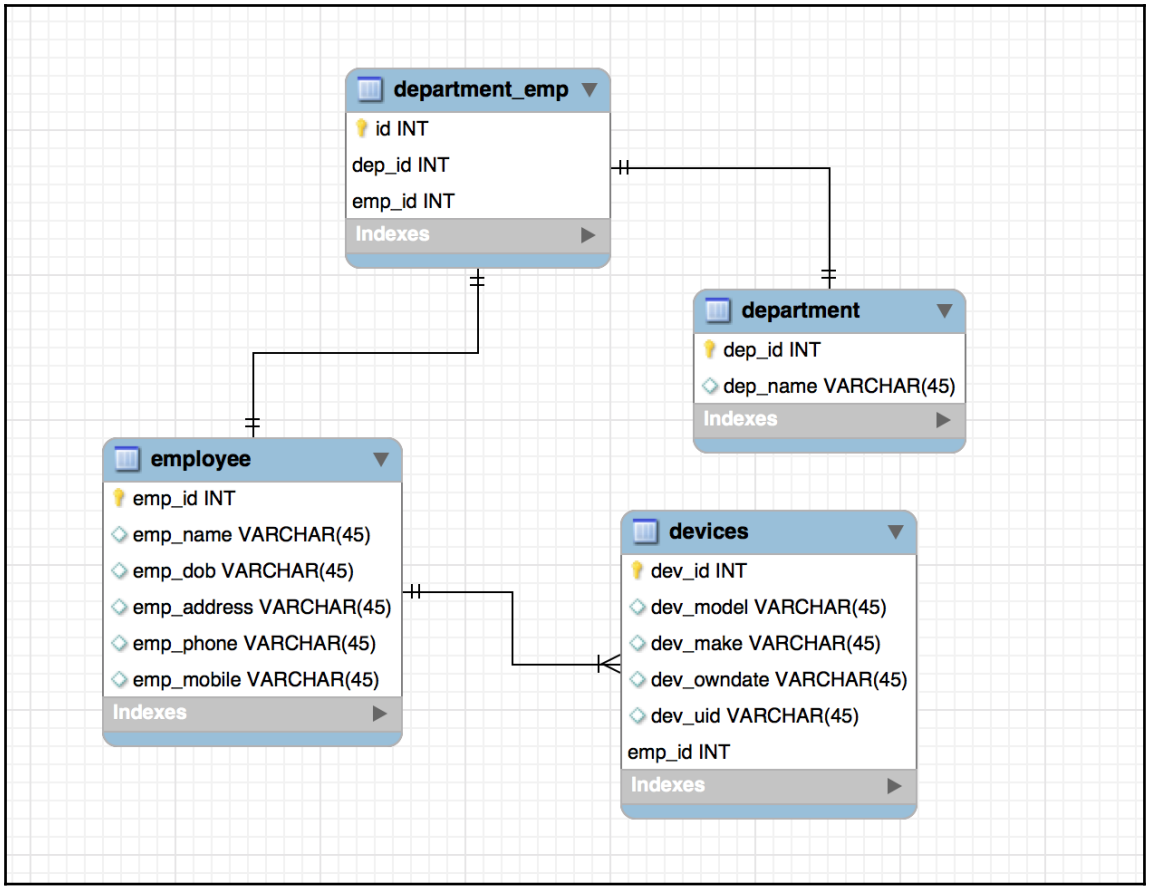


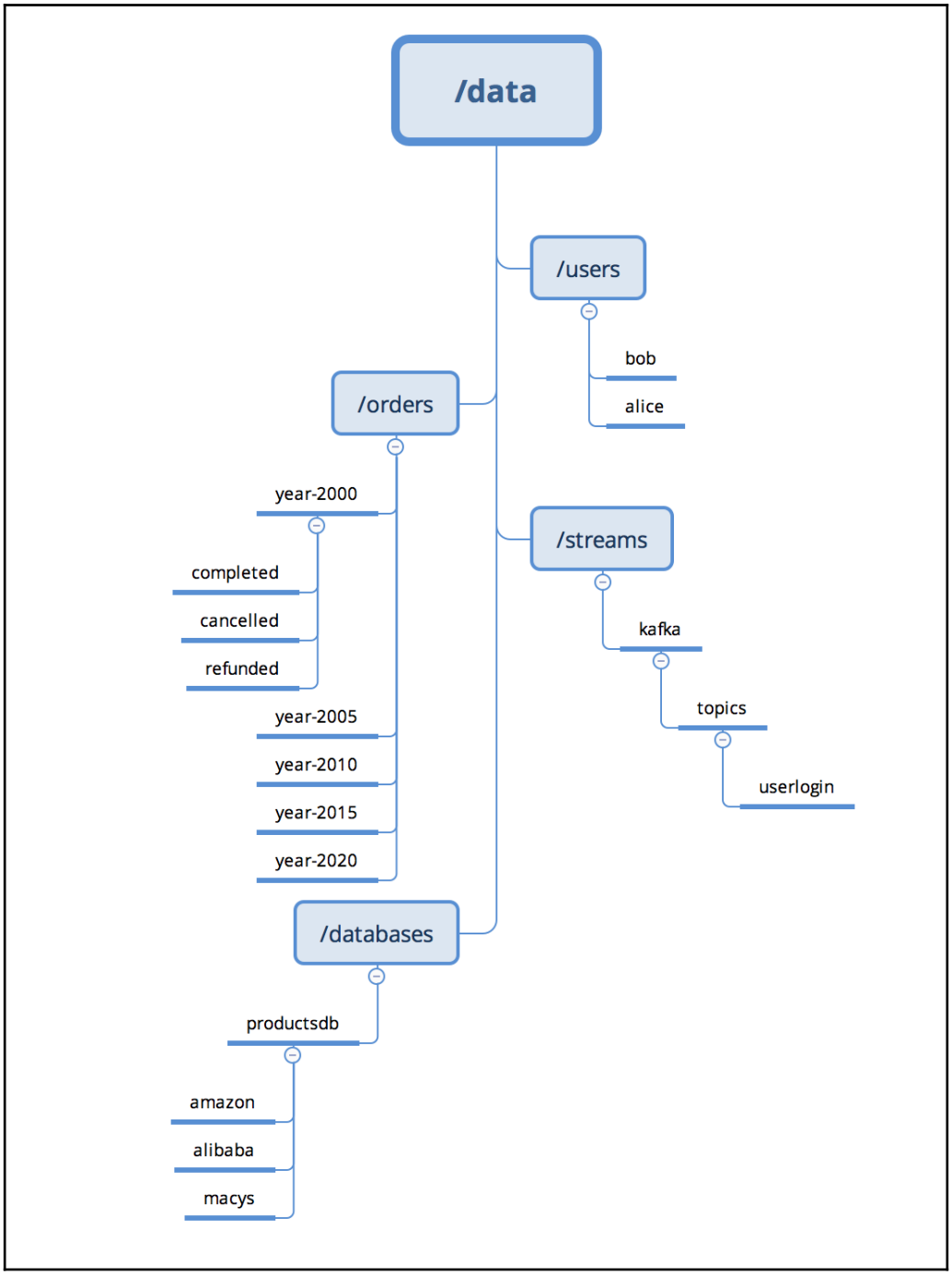

Customer Data
Sales data
BigData Sytem
Order Data
Analytics Applications


Bigdata Flow
Applications
Input Data
Big Data
RDBMS
Analytics


My Org
Vendor
Computers
iMac
Workstation
Catering
India
The Best Caterers
USA
Bay Area Caterers
Training
Employee
Laptop
Mobile Phone
Workstation
iMac

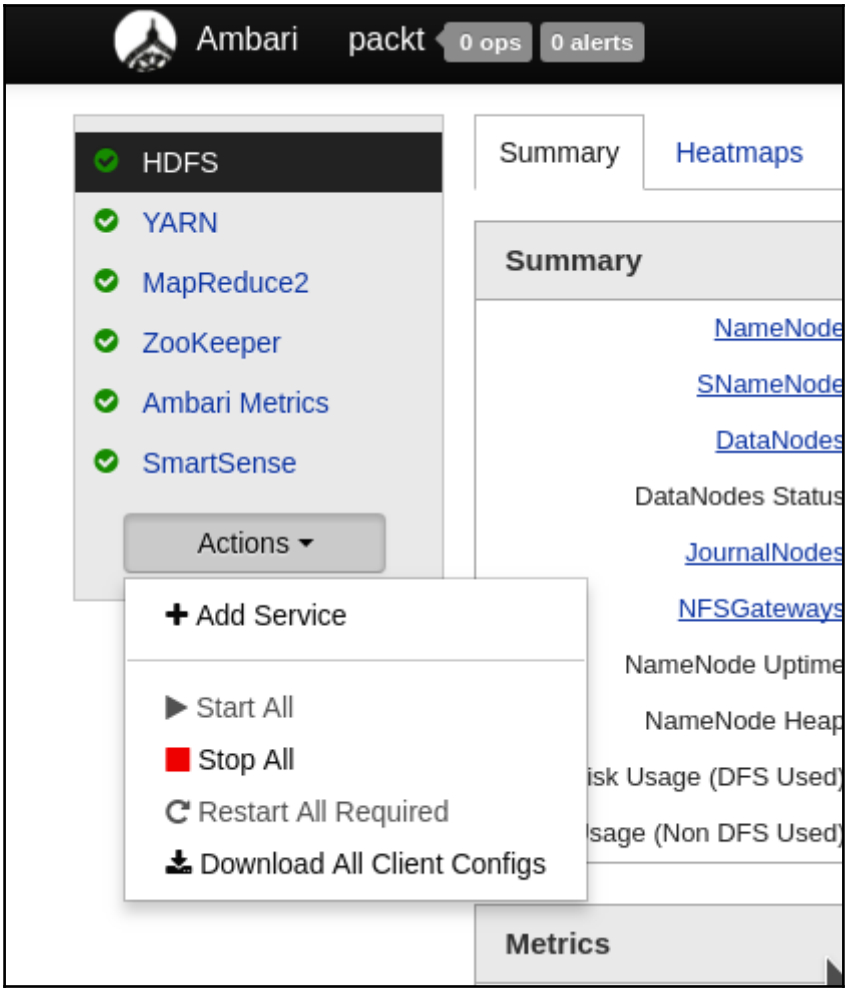

Patients
India
Hospital
Hospitals
Employees
USA


department_emp
id INT
dep_id INT
emp_id INT
Indexes
department
dep_id INT
dep_name VARCHAR(45)
Indexes
employee
emp_id INT
emp_name VARCHAR(45)
emp_dob VARCHAR(45)
emp_address VARCHAR(45)
emp_phone VARCHAR(45)
emp_mobile VARCHAR(45)
Indexes
devices
dev_id INT
dev_model VARCHAR(45)
dev_make VARCHAR(45)
dev_owndate VARCHAR(45)
dev_uid VARCHAR(45)
emp_id INT
Indexes


/data
/users
bob
alice
/orders
year-2000
completed
cancelled
refunded
year-2005
year-2010
year-2015
year-2020
/streams
kafka
topics
userlogin
/databases
productsdb
amazon
alibaba
macys

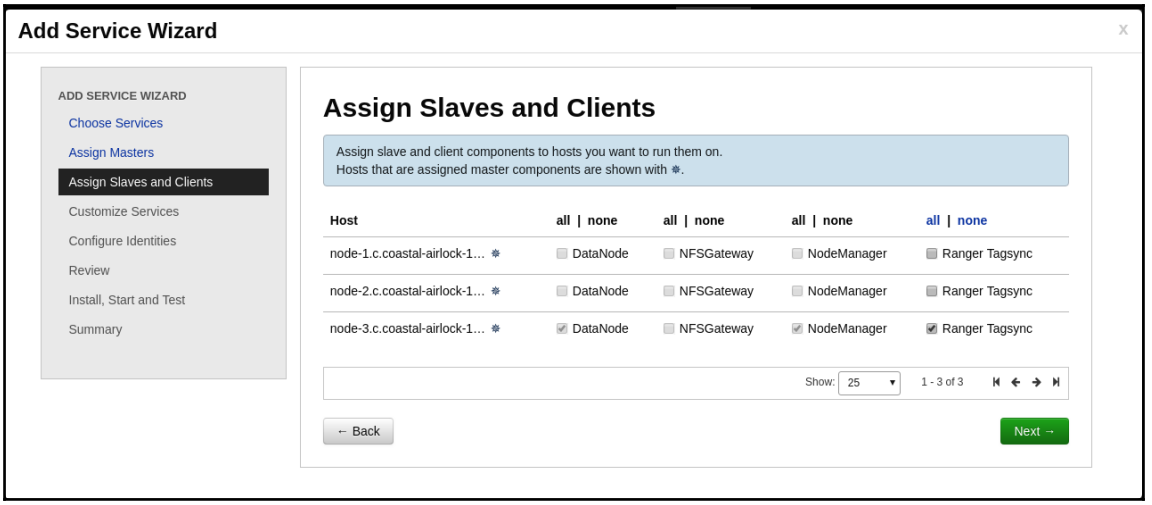
Chapter 2: Hadoop Life Cycle Management

Input Data				
First Name	**Last Name**	**Gender**	**Phone**	**Salary**
Myra	Riley	Female	550-7605-00	6502
Preston	Davis	Male	979-8283-17	4376
Kimberly	Perkins	Female	166-3607-48	7195
Jordan	Myers	Male	513-6151-62	5525
Ryan	Harper	Male	863-7471-20	2657
Shuffled Data				
First Name	**Last Name**	**Gender**	**Phone**	**Salary**
Myra	Riley	Female	550-7605-00	7195
Preston	Davis	Male	979-8283-17	4376
Kimberly	Perkins	Female	166-3607-48	5525
Jordan	Myers	Male	513-6151-62	2657
Ryan	Harper	Male	863-7471-20	65[illegible]

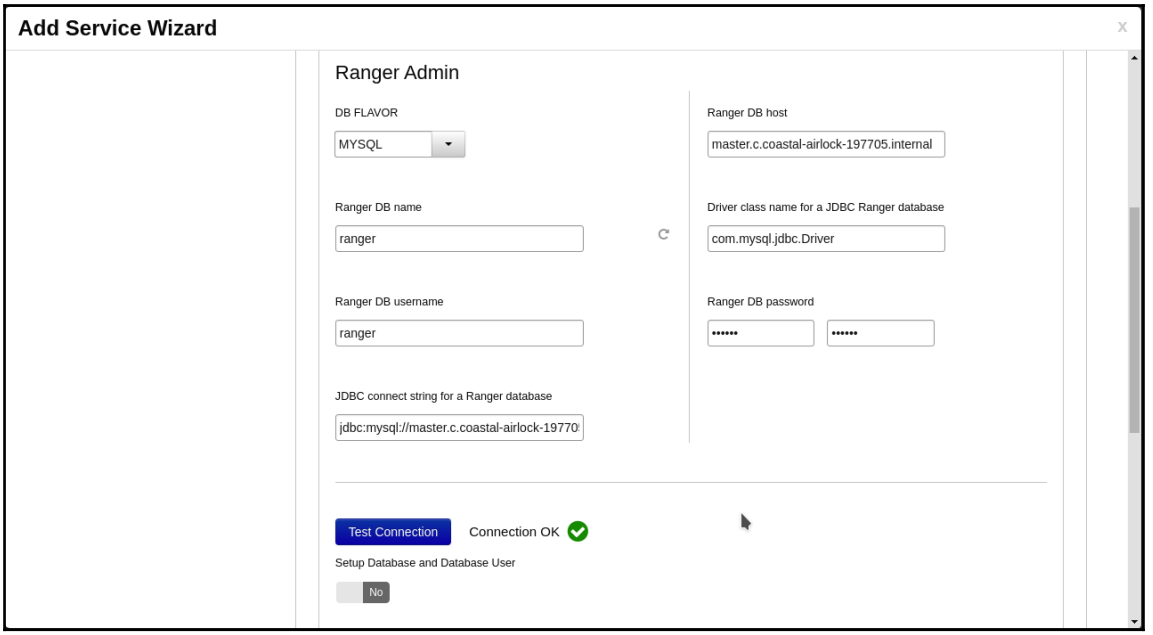

Ambari
packt
0 ops
0 alerts
HDFS
YARN
MapReduce2
ZooKeeper
Ambari Metrics
SmartSense
Actions
Add Service
Start All
Stop All
Restart All Required
Download All Client Configs
Summary
Heatmaps
Summary
NameNode
SNameNode
DataNodes
DataNodes Status
JournalNodes
NFSGateways
NameNode Uptime
NameNode Heap
Metrics

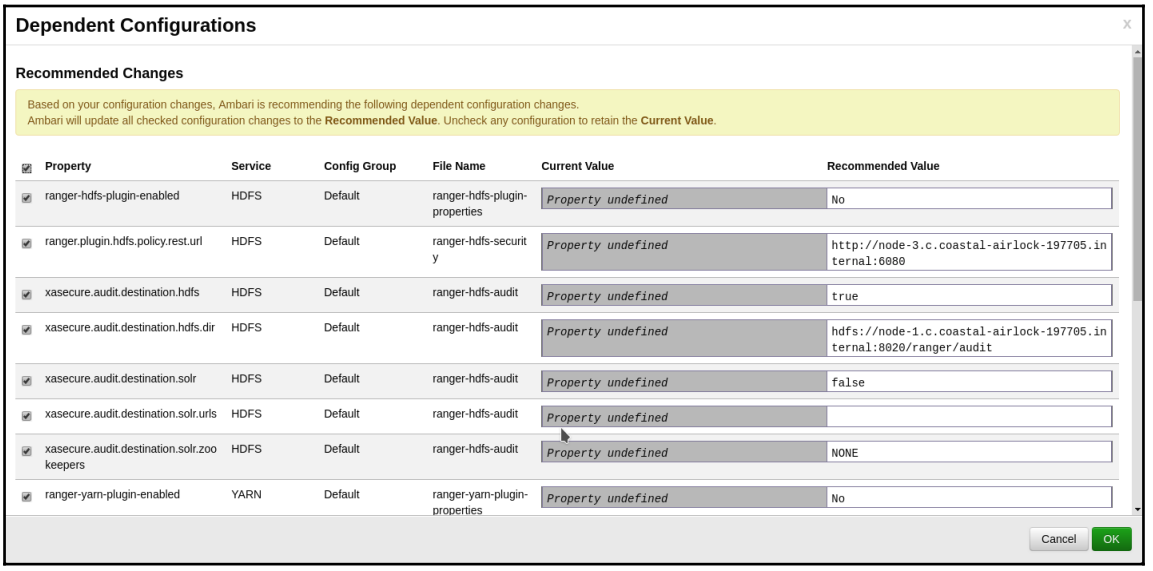

Add Service Wizard
Falcon 0.10.0 Data management and processing platform
Storm 1.1.0 Apache Hadoop Stream processing framework
Flume 1.5.2 A distributed service for collecting, aggregating, and moving large amounts of streaming data into HDFS
Accumulo 1.7.0 Robust, scalable, high performance distributed key/value store.
Ambari Infra 0.1.0 Core shared service used by Ambari managed components.
Ambari Metrics 0.1.0 A system for metrics collection that provides storage and retrieval capability for metrics collected from the cluster
Atlas 0.8.0 Atlas Metadata and Governance platform
Kafka 0.10.1 A high-throughput distributed messaging system
Knox 0.12.0 Provides a single point of authentication and access for Apache Hadoop services in a cluster
Log Search 0.5.0 Log aggregation, analysis, and visualization for Ambari managed services. This service is Technical Preview.
Ranger 0.7.0 Comprehensive security for Hadoop
Ranger KMS 0.7.0 Key Management Server
SmartSense 1.4.3.2.6.0.0-267 SmartSense - Hortonworks SmartSense Tool (HST) helps quickly gather configuration, metrics, logs from common HDP services that aids to quickly

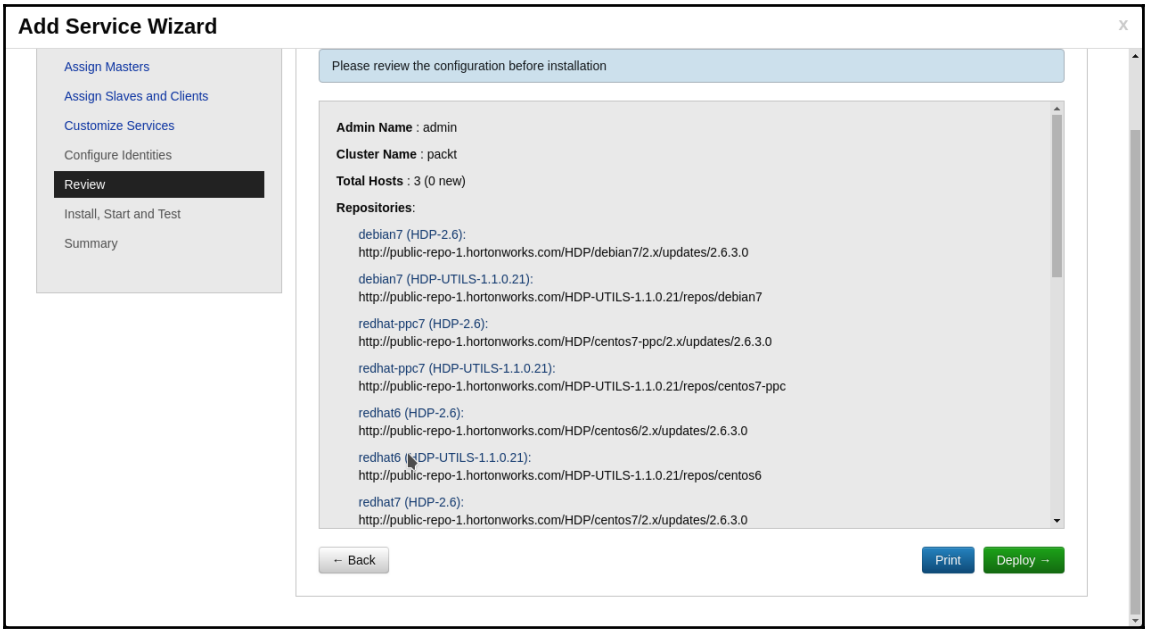

Add Service Wizard
Review
Install, Start and Test
Summary
ResourceManager: node-2.c.coastal-airlock-197705
App Timeline Server: node-2.c.coastal-airlock-197705
History Server: node-2.c.coastal-airlock-197705
ZooKeeper Server: node-2.c.coastal-airlock-197705
ZooKeeper Server: node-3.c.coastal-airlock-197705
ZooKeeper Server: node-1.c.coastal-airlock-197705
Metrics Collector: node-3.c.coastal-airlock-197705
Grafana: node-1.c.coastal-airlock-197705
Ranger Usersync: node-3.c.coastal-airlock-197705
Ranger Admin: node-3.c.coastal-airlock-197705
HST Server: node-1.c.coastal-airlock-197705
Activity Explorer: node-1.c.coastal-airlock-197705
Activity Analyzer: node-1.c.coastal-airlock-197705
HST Server
Activity Explorer
Activity Analyzer
node-2.c.coastal-airlock-197705.internal (12.6 GB, 2 cores)
SNameNode
ResourceManager
App Timeline Server
History Server
ZooKeeper Server
node-3.c.coastal-airlock-197705.internal (12.6 GB, 2 cores)
ZooKeeper Server
Metrics Collector
Ranger Usersync
Ranger Admin

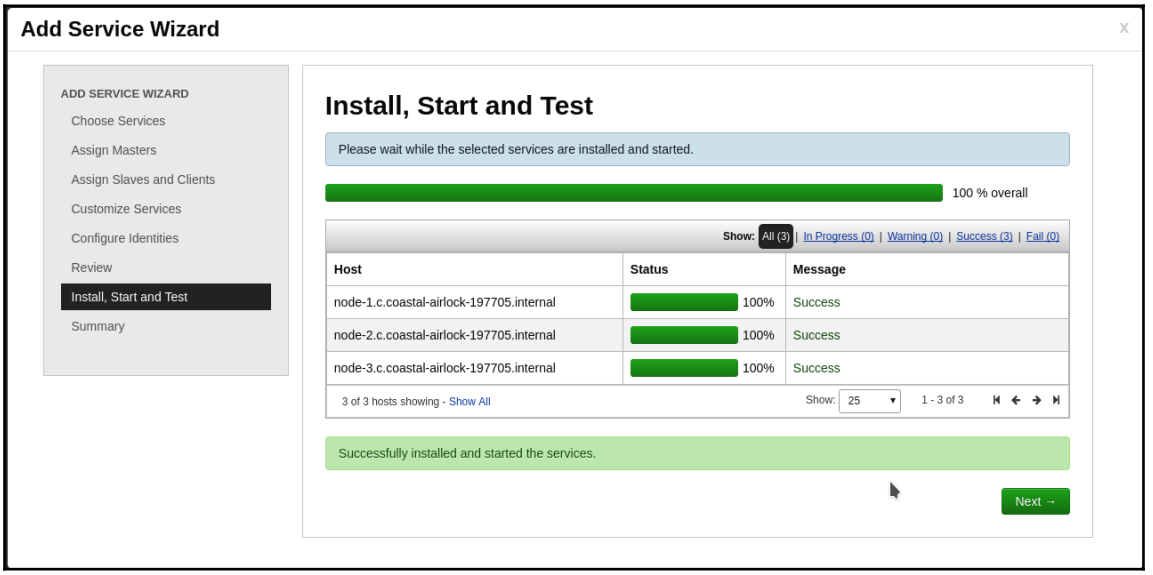


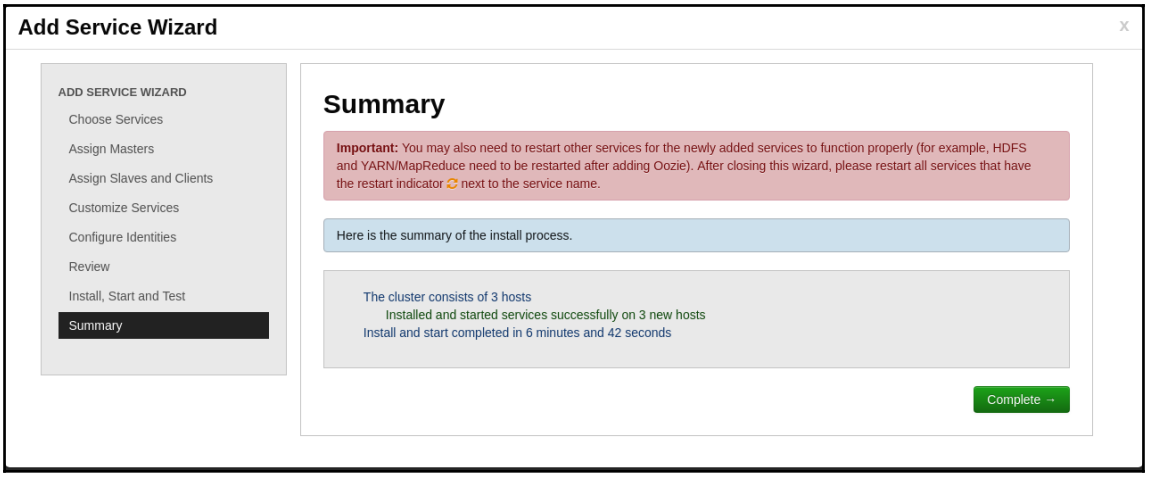

```
Welcome to the MariaDB monitor.  Commands end with ; or \g.
Your MariaDB connection id is 17
Server version: 5.5.56-MariaDB MariaDB Server

Copyright (c) 2000, 2017, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MariaDB [(none)]> create database ranger;
Query OK, 1 row affected (0.00 sec)

MariaDB [(none)]> use ranger;
Database changed
MariaDB [ranger]> grant all privileges on ranger.* to ranger@localhost identified by 'ranger';
Query OK, 0 rows affected (0.01 sec)

MariaDB [ranger]> grant all privileges on ranger.* to ranger@'%' identified by 'ranger';
Query OK, 0 rows affected (0.00 sec)

MariaDB [ranger]> flush privileges;
Query OK, 0 rows affected (0.00 sec)

MariaDB [ranger]> quit
Bye
[collabrecruit_in@master ~]$
```

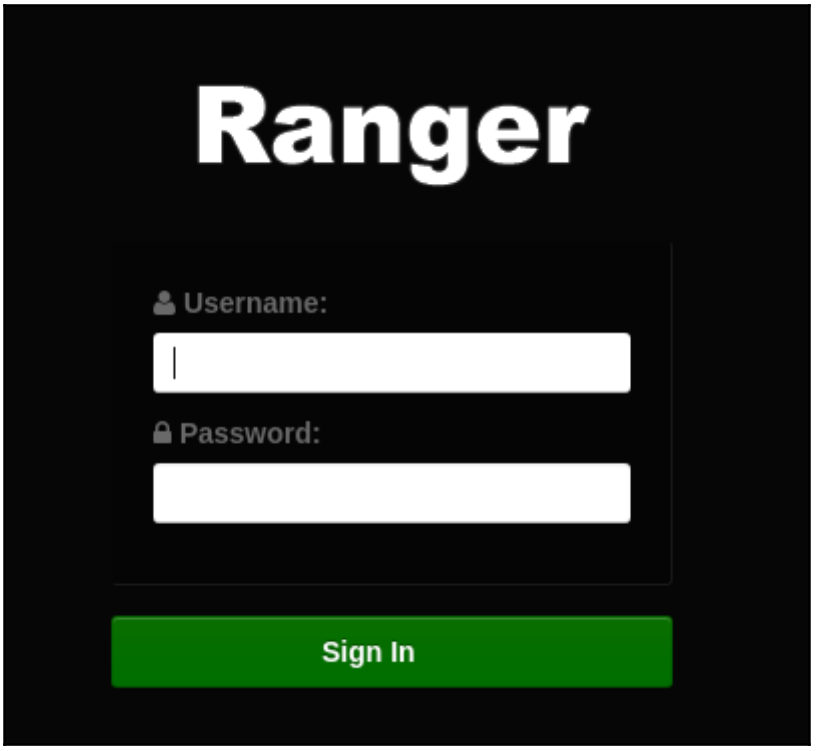

Add Service Wizard
Ranger Admin
DB FLAVOR
MYSQL
Ranger DB host
master.c.coastal-airlock-197705.internal
Ranger DB name
ranger
Driver class name for a JDBC Ranger database
com.mysql.jdbc.Driver
Ranger DB username
ranger
Ranger DB password
JDBC connect string for a Ranger database
jdbc:mysql://master.c.coastal-airlock-19770
Test Connection
Connection OK
Setup Database and Database User
No

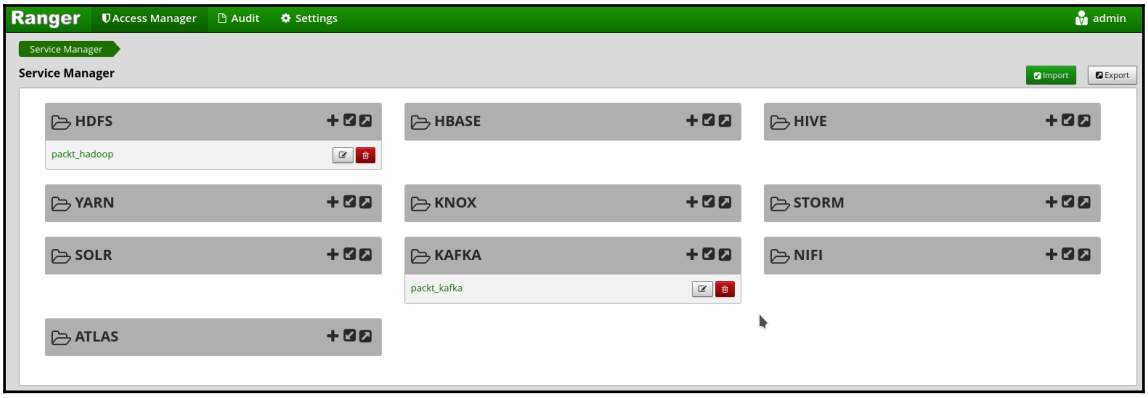

Dependent Configurations
Recommended Changes
Based on your configuration changes, Ambari is recommending the following dependent configuration changes.
Ambari will update all checked configuration changes to the Recommended Value. Uncheck any configuration to retain the Current Value.
Property
Service
Config Group
File Name
Current Value
Recommended Value
ranger-hdfs-plugin-enabled
HDFS
Default
ranger-hdfs-plugin-properties
Property undefined
No
ranger.plugin.hdfs.policy.rest.url
HDFS
Default
ranger-hdfs-security
Property undefined
http://node-3.c.coastal-airlock-197705.internal:6080
xasecure.audit.destination.hdfs
HDFS
Default
ranger-hdfs-audit
Property undefined
true
xasecure.audit.destination.hdfs.dir
HDFS
Default
ranger-hdfs-audit
Property undefined
hdfs://node-1.c.coastal-airlock-197705.internal:8020/ranger/audit
xasecure.audit.destination.solr
HDFS
Default
ranger-hdfs-audit
Property undefined
false
xasecure.audit.destination.solr.urls
HDFS
Default
ranger-hdfs-audit
Property undefined
xasecure.audit.destination.solr.zookeepers
HDFS
Default
ranger-hdfs-audit
Property undefined
NONE
ranger-yarn-plugin-enabled
YARN
Default
ranger-yarn-plugin-properties
Property undefined
No
Cancel
OK

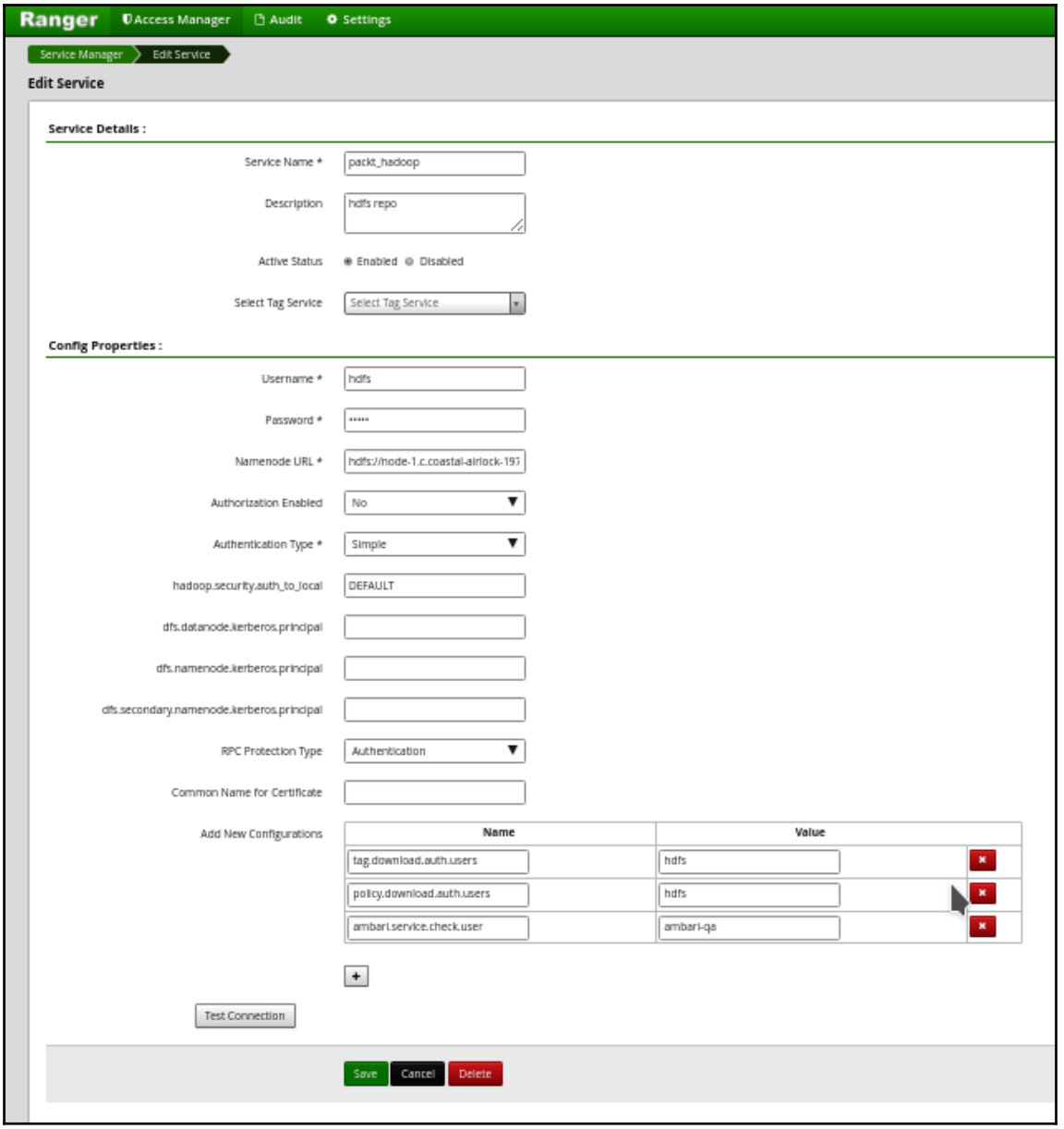

Add Service Wizard
Assign Masters
Assign Slaves and Clients
Customize Services
Configure Identities
Review
Install, Start and Test
Summary
Please review the configuration before installation
Admin Name : admin
Cluster Name : packt
Total Hosts : 3 (0 new)
Repositories:
debian7 (HDP-2.6):
http://public-repo-1.hortonworks.com/HDP/debian7/2.x/updates/2.6.3.0
debian7 (HDP-UTILS-1.1.0.21):
http://public-repo-1.hortonworks.com/HDP-UTILS-1.1.0.21/repos/debian7
redhat-ppc7 (HDP-2.6):
http://public-repo-1.hortonworks.com/HDP/centos7-ppc/2.x/updates/2.6.3.0
redhat-ppc7 (HDP-UTILS-1.1.0.21):
http://public-repo-1.hortonworks.com/HDP-UTILS-1.1.0.21/repos/centos7-ppc
redhat6 (HDP-2.6):
http://public-repo-1.hortonworks.com/HDP/centos6/2.x/updates/2.6.3.0
redhat6 (HDP-UTILS-1.1.0.21):
http://public-repo-1.hortonworks.com/HDP-UTILS-1.1.0.21/repos/centos6
redhat7 (HDP-2.6):
http://public-repo-1.hortonworks.com/HDP/centos7/2.x/updates/2.6.3.0
← Back
Print
Deploy →

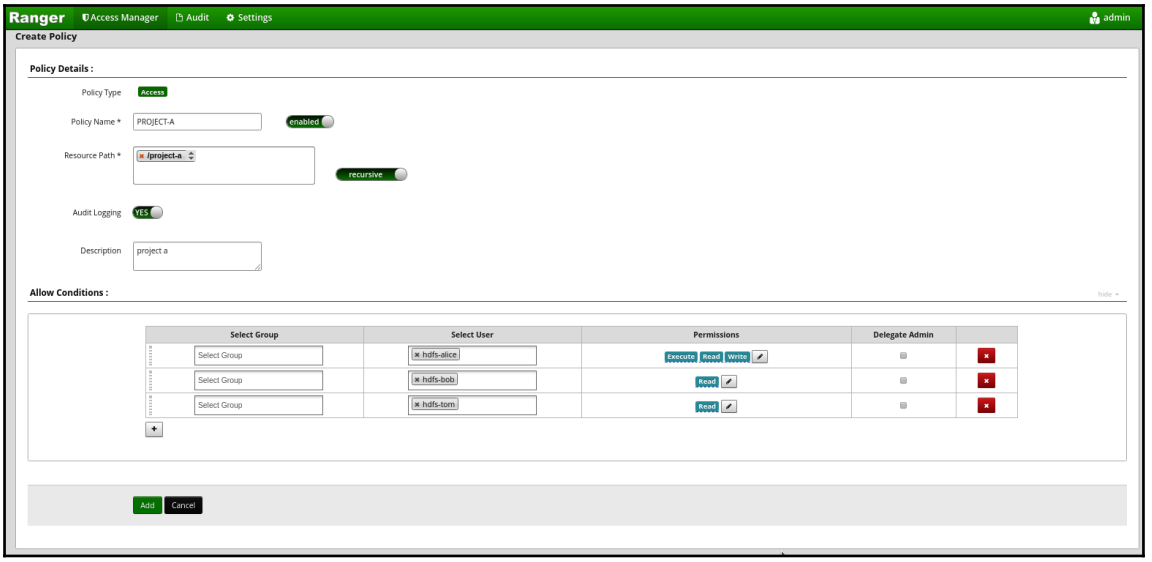

Add Service Wizard
ADD SERVICE WIZARD
Choose Services
Assign Masters
Assign Slaves and Clients
Customize Services
Configure Identities
Review
Install, Start and Test
Summary
Install, Start and Test
Please wait while the selected services are installed and started.
100 % overall
Show: All (3) | In Progress (0) | Warning (0) | Success (3) | Fail (0)
Host
Status
Message
node-1.c.coastal-airlock-197705.internal
100%
Success
node-2.c.coastal-airlock-197705.internal
100%
Success
node-3.c.coastal-airlock-197705.internal
100%
Success
3 of 3 hosts showing - Show All
Show: 25
1 - 3 of 3
Successfully installed and started the services.
Next →

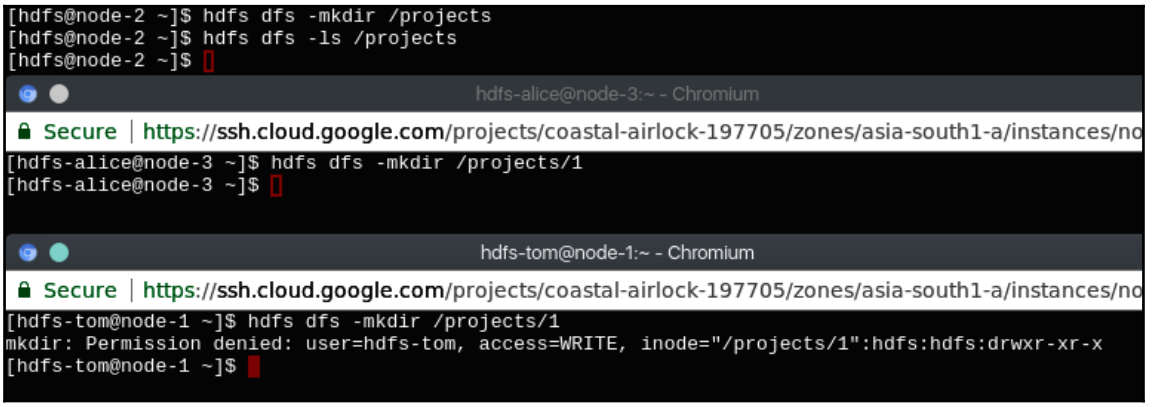

Add Service Wizard
ADD SERVICE WIZARD
Choose Services
Assign Masters
Assign Slaves and Clients
Customize Services
Configure Identities
Review
Install, Start and Test
Summary
Summary
Important: You may also need to restart other services for the newly added services to function properly (for example, HDFS and YARN/MapReduce need to be restarted after adding Oozie). After closing this wizard, please restart all services that have the restart indicator next to the service name.
Here is the summary of the install process.
The cluster consists of 3 hosts
Installed and started services successfully on 3 new hosts
Install and start completed in 6 minutes and 42 seconds
Complete →


Ranger
Username:
Password:
Sign In


Ranger
Access Manager
Audit
Settings
admin
Service Manager
Service Manager
Import
Export
HDFS
packt_hadoop
HBASE
HIVE
YARN
KNOX
STORM
SOLR
KAFKA
packt_kafka
NIFI
ATLAS


Ranger
Access Manager
Audit
Settings
Service Manager
Edit Service
Edit Service
Service Details :
Service Name *
packt_hadoop
Description
hdfs repo
Active Status
Enabled
Disabled
Select Tag Service
Select Tag Service
Config Properties :
Username *
hdfs
Password *
Namenode URL *
hdfs://node-1.c.coastal-airlock-197
Authorization Enabled
No
Authentication Type *
Simple
hadoop.security.auth_to_local
DEFAULT
dfs.datanode.kerberos.principal
dfs.namenode.kerberos.principal
dfs.secondary.namenode.kerberos.principal
RPC Protection Type
Authentication
Common Name for Certificate
Add New Configurations
Name
Value
tag.download.auth.users
hdfs
policy.download.auth.users
hdfs
ambari.service.check.user
ambari-qa
+
Test Connection
Save
Cancel
Delete

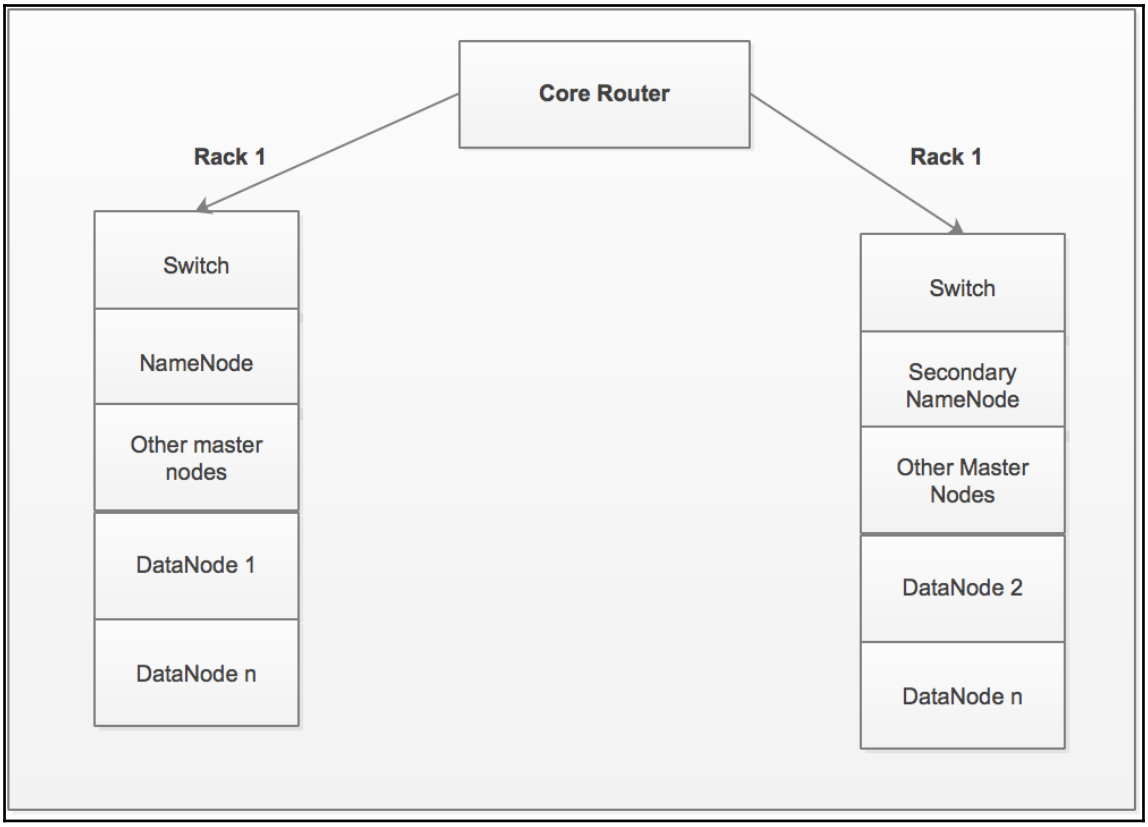

Ranger
Access Manager
Audit
Settings
admin
Create Policy
Policy Details :
Policy Type
Access
Policy Name *
PROJECT-A
enabled
Resource Path *
/project-a
recursive
Audit Logging
YES
Description
project a
Allow Conditions :
Select Group
Select User
Permissions
Delegate Admin
Select Group
hdfs-alice
Execute
Read
Write
Select Group
hdfs-bob
Read
Select Group
hdfs-tom
Read
Add
Cancel

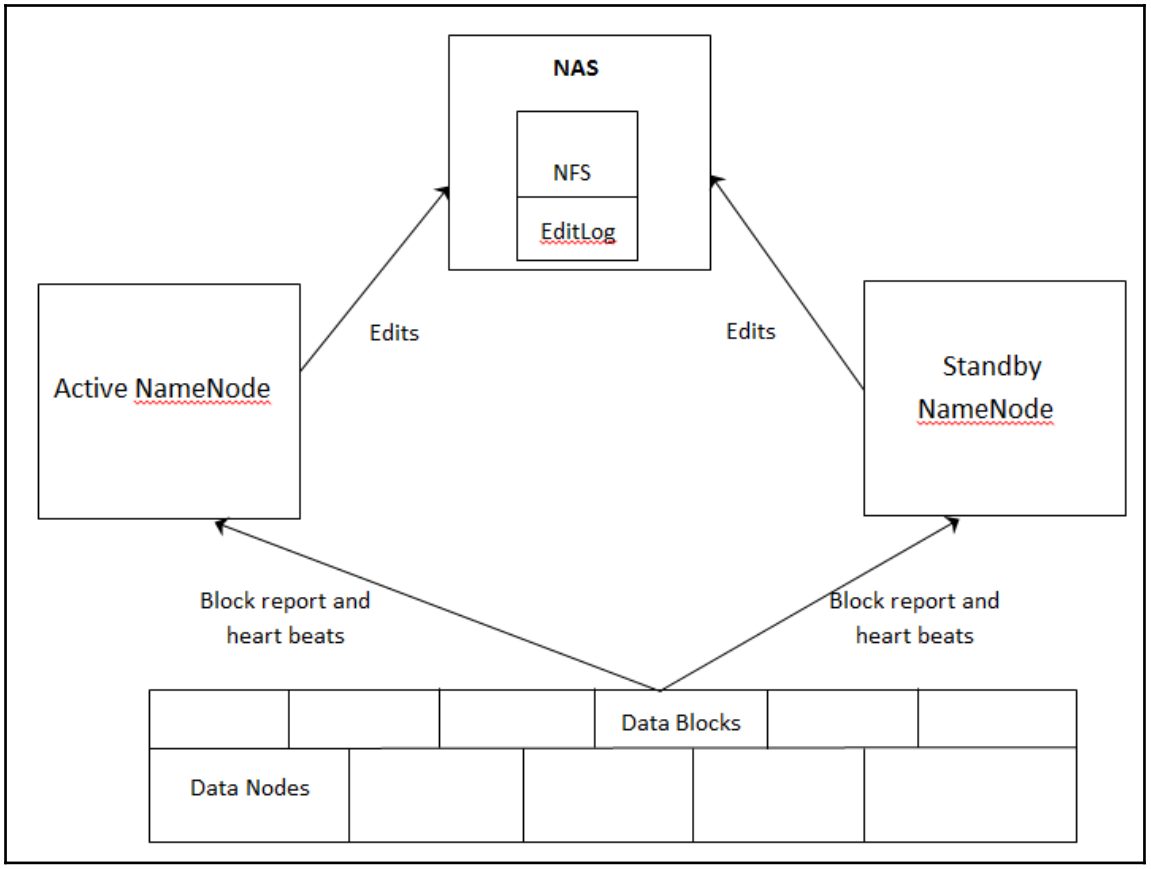

[hdfs@node-2 ~]$ hdfs dfs -mkdir /projects
[hdfs@node-2 ~]$ hdfs dfs -ls /projects
[hdfs@node-2 ~]$
hdfs-alice@node-3:~ - Chromium
Secure | https://ssh.cloud.google.com/projects/coastal-airlock-197705/zones/asia-south1-a/instances/no
[hdfs-alice@node-3 ~]$ hdfs dfs -mkdir /projects/1
[hdfs-alice@node-3 ~]$
hdfs-tom@node-1:~ - Chromium
Secure | https://ssh.cloud.google.com/projects/coastal-airlock-197705/zones/asia-south1-a/instances/no
[hdfs-tom@node-1 ~]$ hdfs dfs -mkdir /projects/1
mkdir: Permission denied: user=hdfs-tom, access=WRITE, inode="/projects/1":hdfs:hdfs:drwxr-xr-x
[hdfs-tom@node-1 ~]$

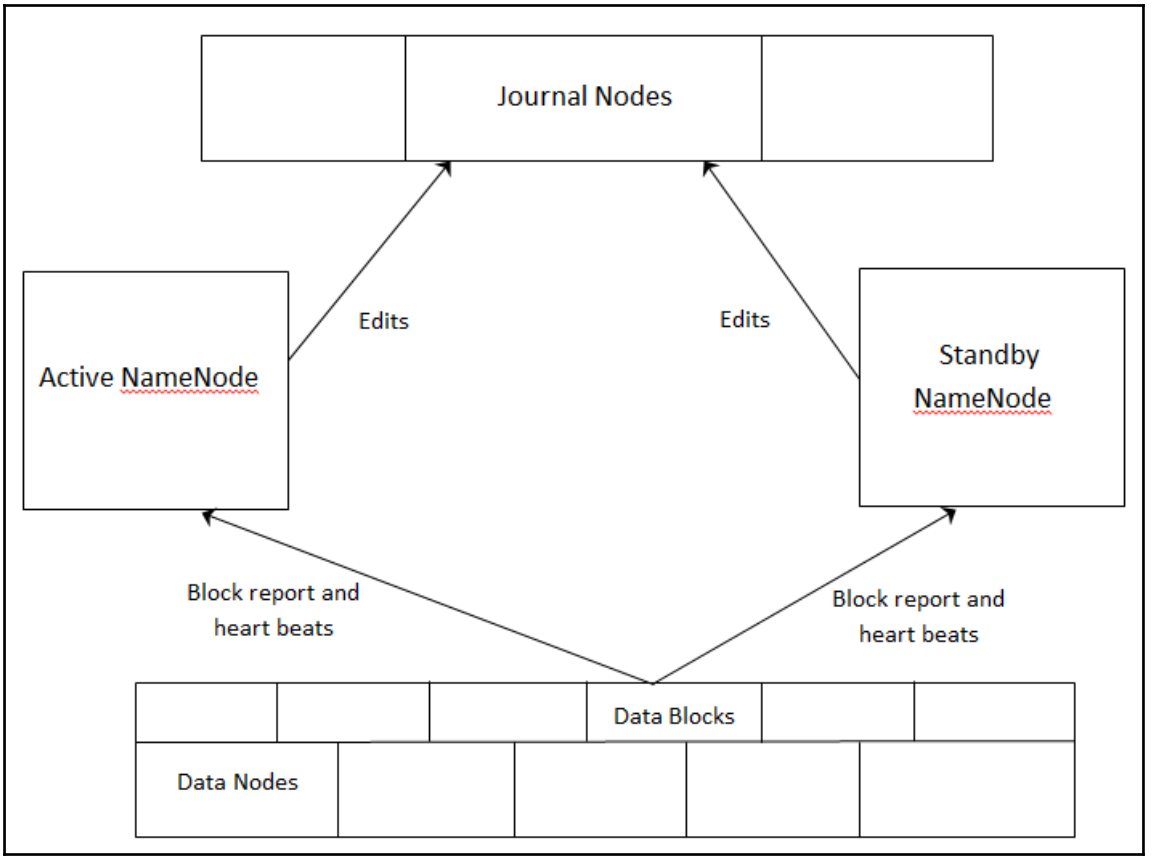

Ranger
Access Manager
Audit
Settings
admin
03/13/2018 02:39:03 PM
mapred
packt_hadoop
hdfs
/mr-history/tmp
path
READ_EXECUTE
Allowed
hadoop-acl
10.160.0.4
packt
1
9
03/13/2018 02:38:41 PM
hdfs-alice
packt_hadoop
hdfs
/projects/1
path
WRITE
Allowed
ranger-acl
10.160.0.5
packt
1
03/13/2018 02:38:17 PM
hdfs-tom
packt_hadoop
hdfs
/projects/1
path
WRITE
Denied
hadoop-acl
10.160.0.3
packt
1

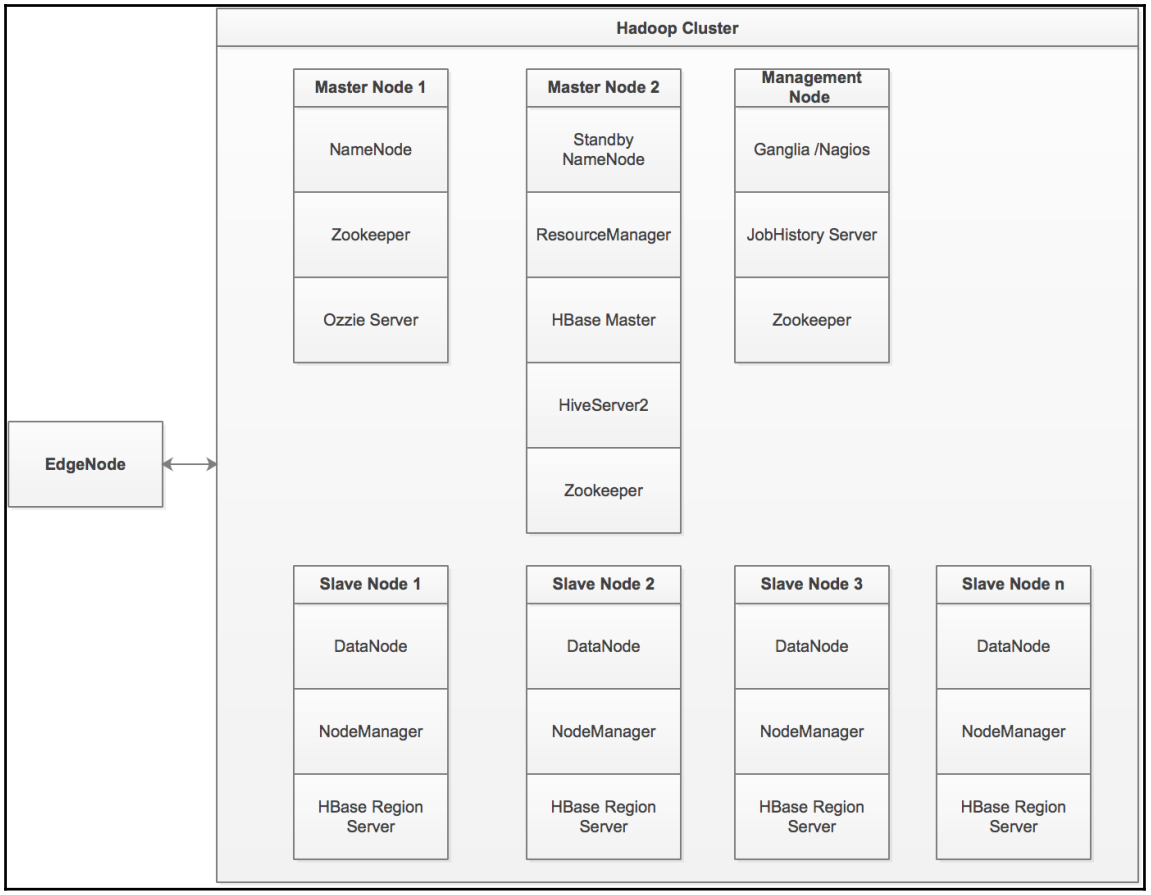
Chapter 3: Hadoop Design Consideration

Volume	Data size varies from terabytes to petabytes
Velocity	Data is generated every time tick, no limit
Variety	Data format – Structured, Semi-structured, Unstructured
Veracity	Data dependability, cleanliness
Value	Data conversion into value

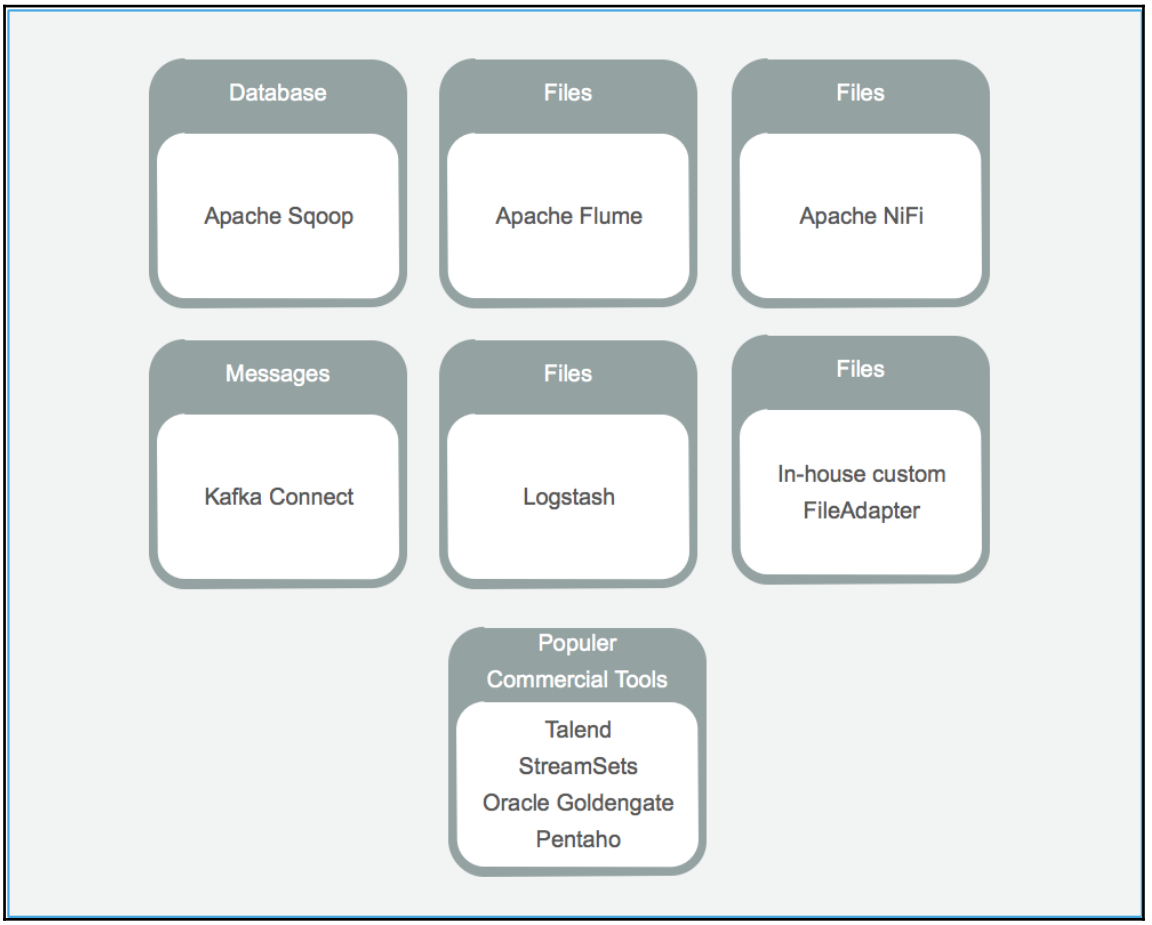

Core Router
Rack 1
Rack 1
Switch
NameNode
Other master nodes
DataNode 1
DataNode n
Switch
Secondary NameNode
Other Master Nodes
DataNode 2
DataNode n

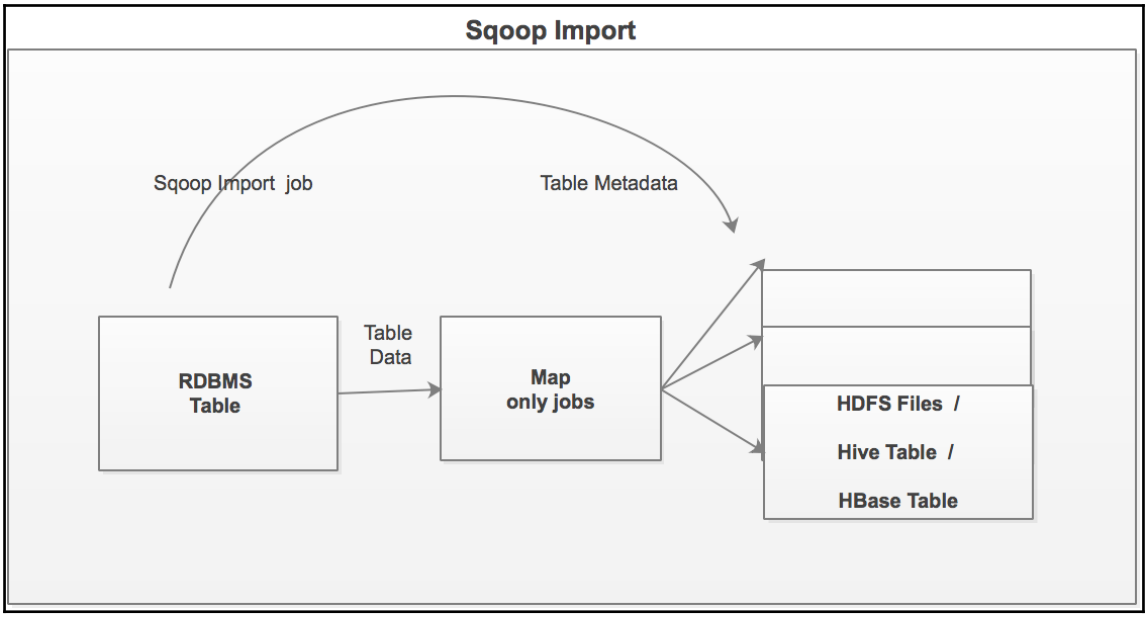

NAS
NFS
EditLog
Edits
Edits
Active NameNode
Standby NameNode
Block report and heart beats
Block report and heart beats
Data Blocks
Data Nodes

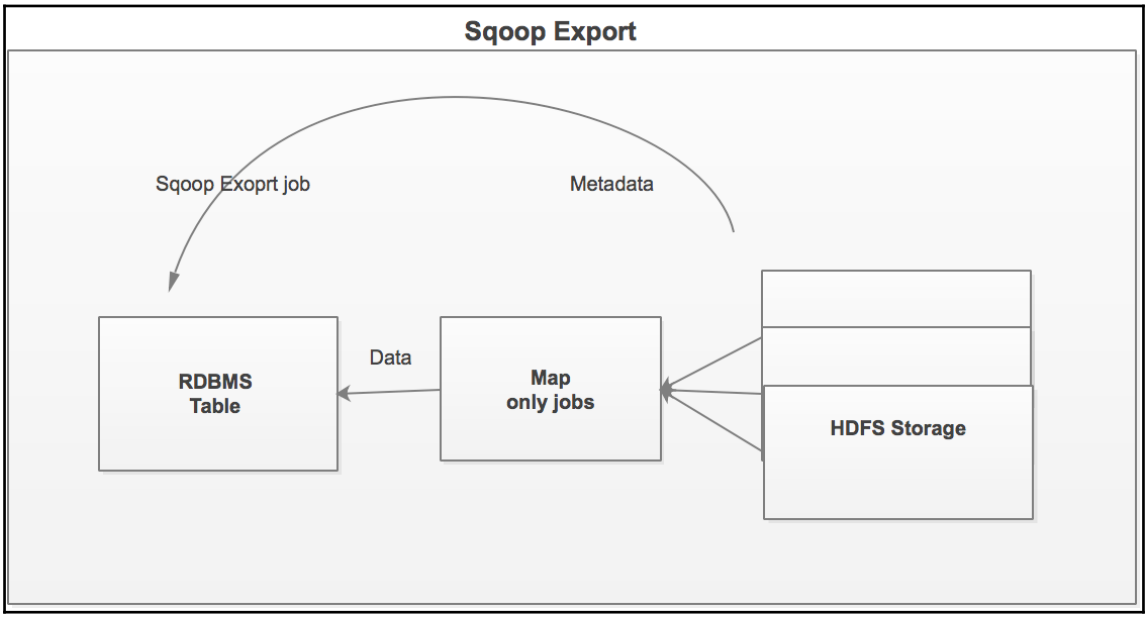

Journal Nodes
Edits
Edits
Active NameNode
Standby NameNode
Block report and heart beats
Block report and heart beats
Data Blocks
Data Nodes

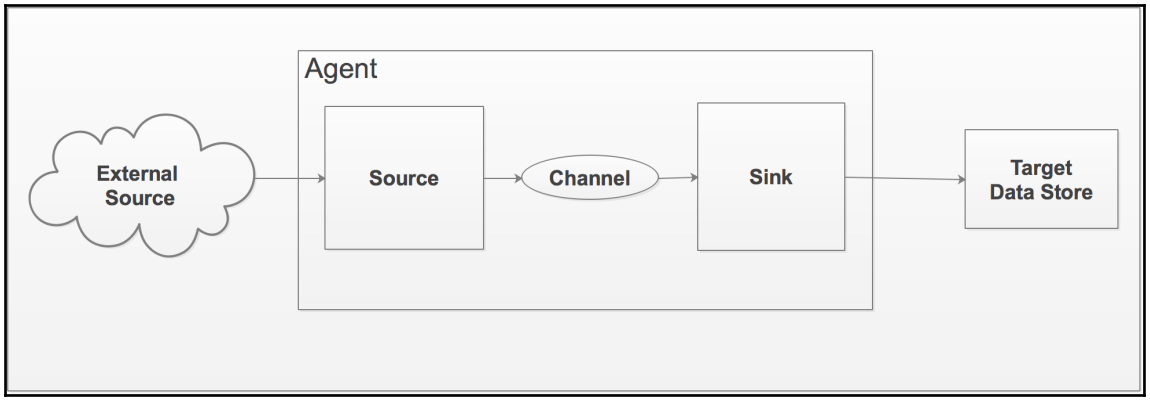

Hadoop Cluster
Master Node 1
NameNode
Zookeeper
Ozzie Server
Master Node 2
Standby NameNode
ResourceManager
HBase Master
HiveServer2
Zookeeper
Management Node
Ganglia /Nagios
JobHistory Server
Zookeeper
EdgeNode
Slave Node 1
DataNode
NodeManager
HBase Region Server
Slave Node 2
DataNode
NodeManager
HBase Region Server
Slave Node 3
DataNode
NodeManager
HBase Region Server
Slave Node n
DataNode
NodeManager
HBase Region Server

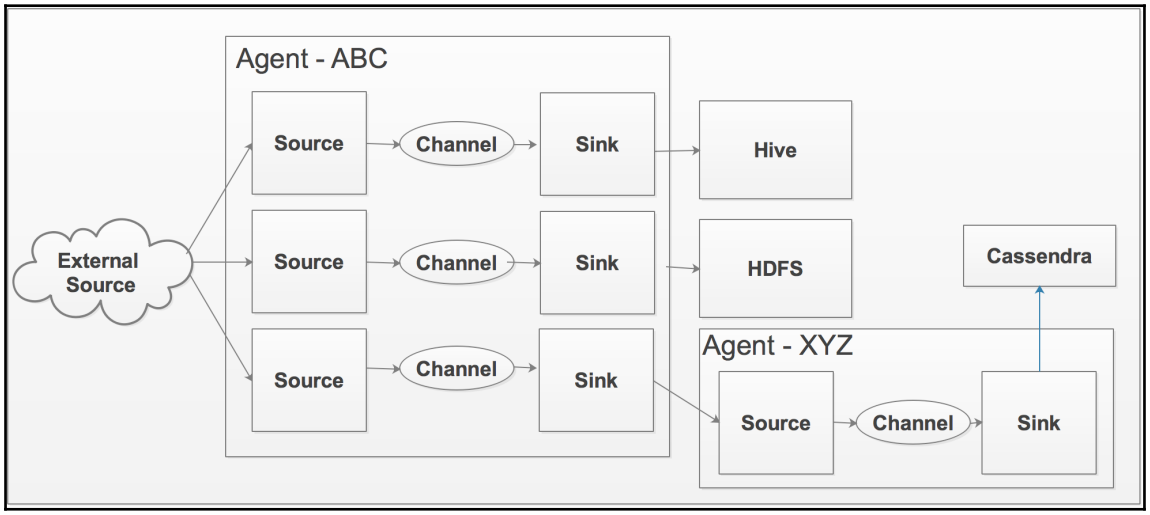
Chapter 4: Data Movement Techniques

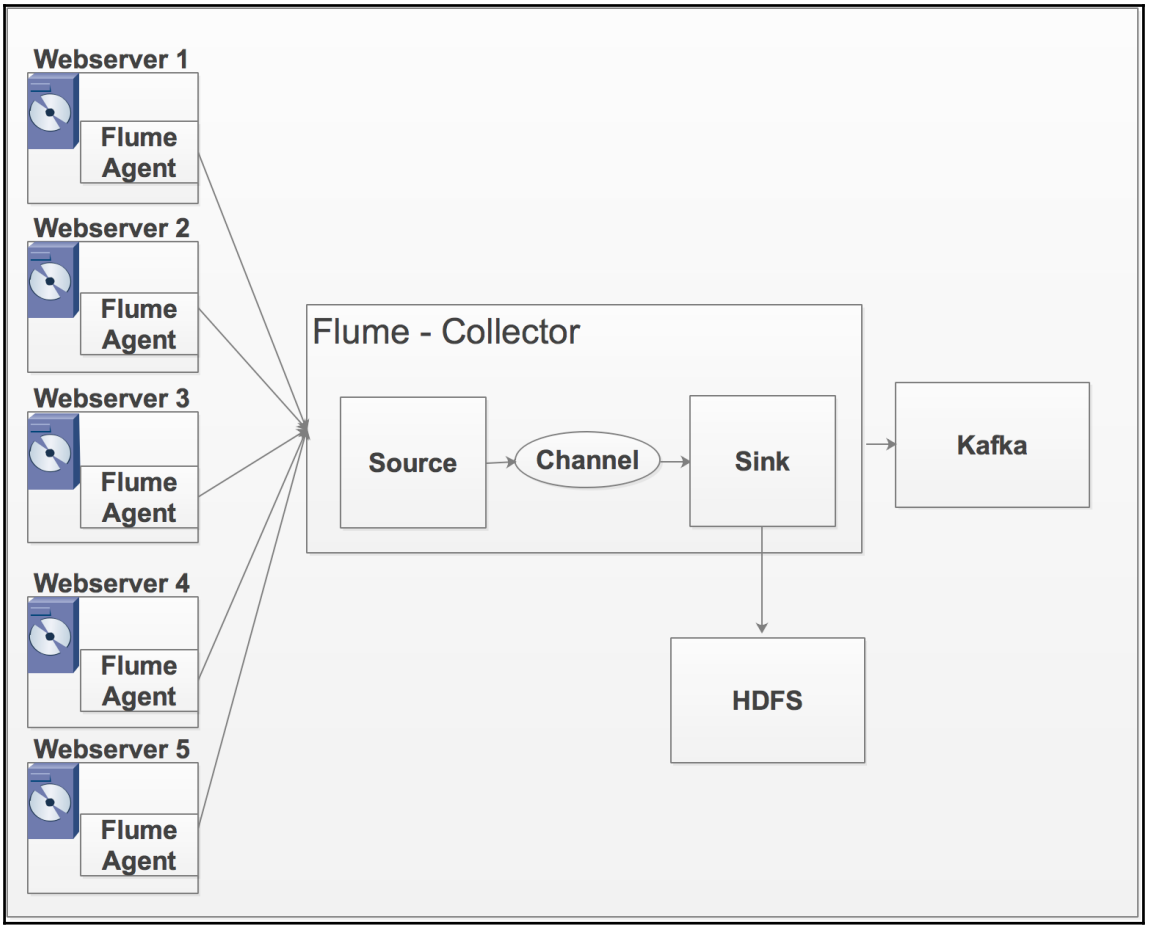


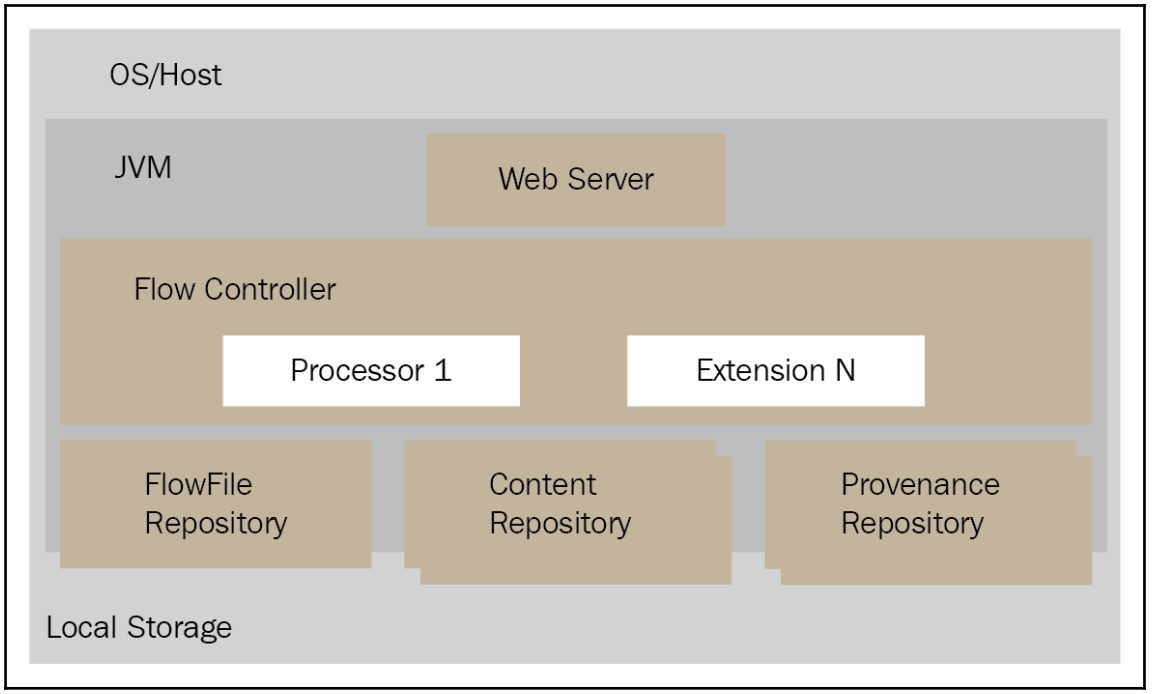

Sqoop Import
Sqoop Import job
Table Metadata
RDBMS Table
Table Data
Map only jobs
HDFS Files /
Hive Table /
HBase Table

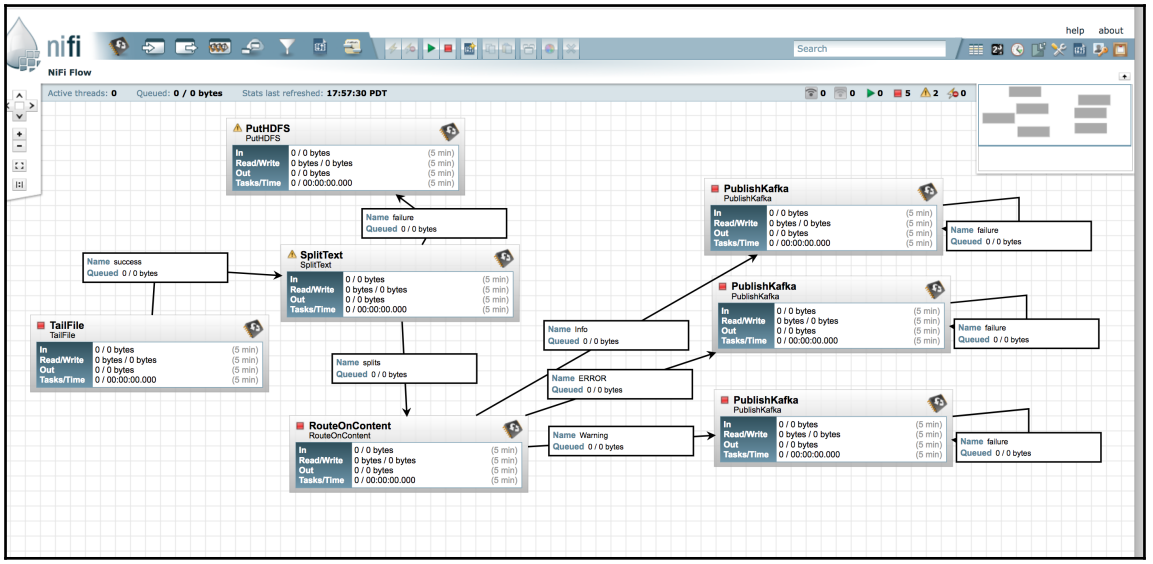

Sqoop Export
Sqoop Exoprt job
Metadata
RDBMS Table
Data
Map only jobs
HDFS Storage

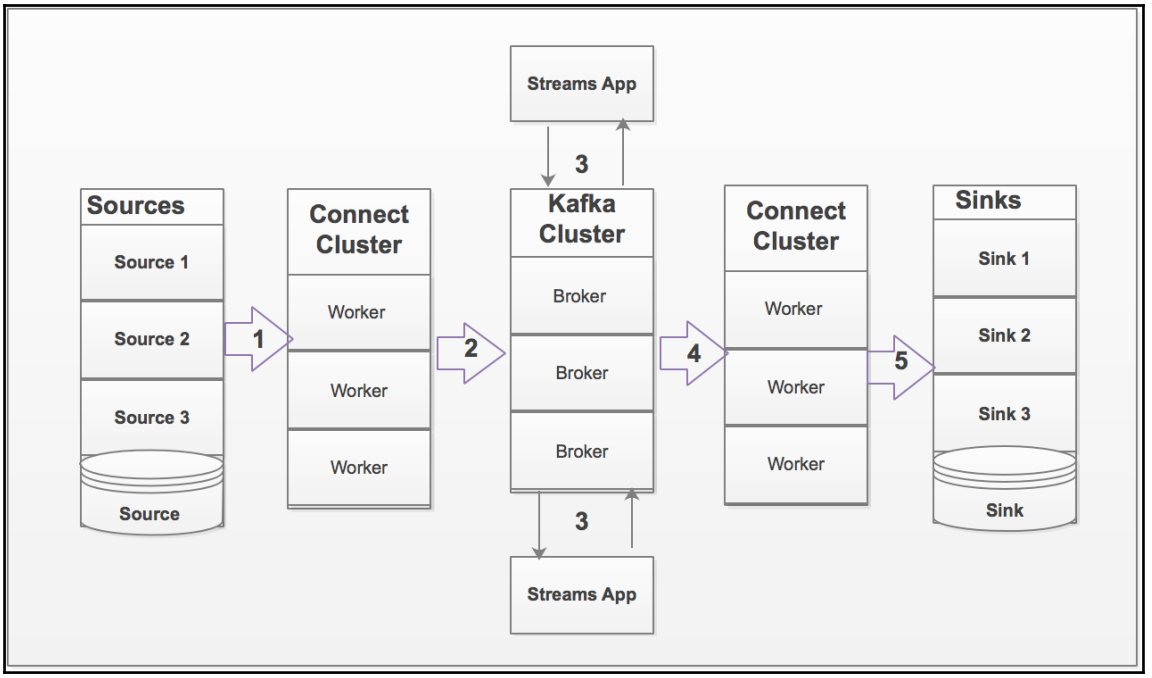

Agent
External Source
Source
Channel
Sink
Target Data Store

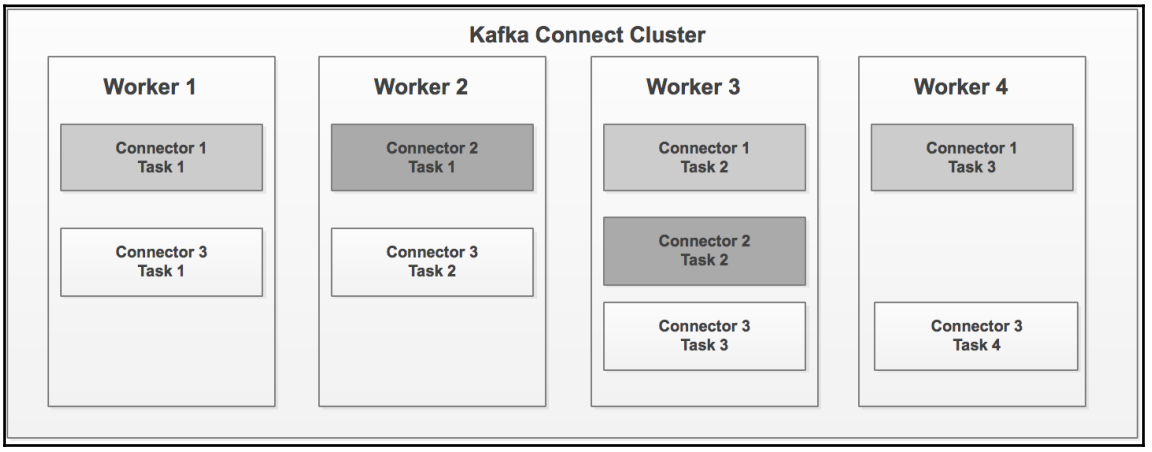

Agent - ABC
External Source
Source
Channel
Sink
Source
Channel
Sink
Source
Channel
Sink
Hive
HDFS
Cassendra
Agent - XYZ
Source
Channel
Sink

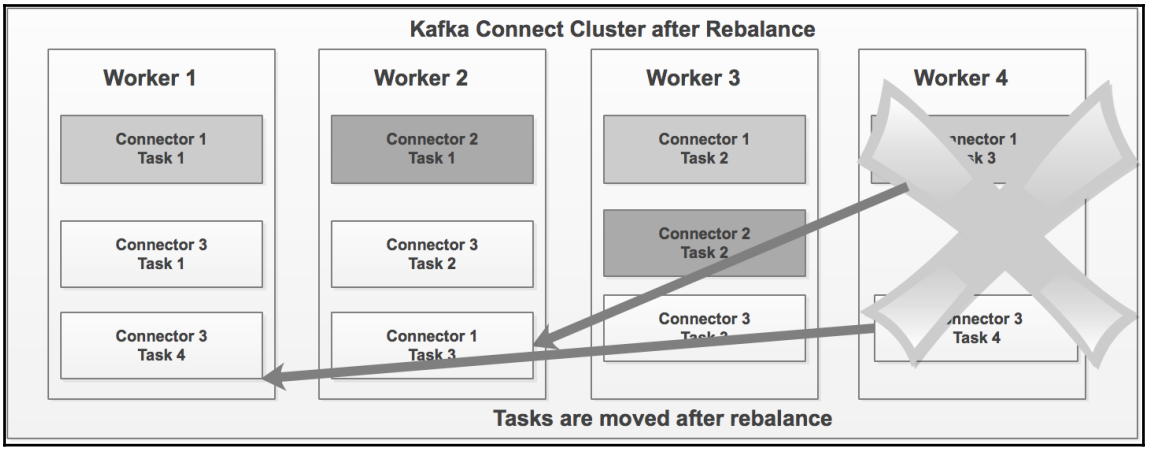

Webserver 1
Flume Agent
Webserver 2
Flume Agent
Webserver 3
Flume Agent
Webserver 4
Flume Agent
Webserver 5
Flume Agent
Flume - Collector
Source
Channel
Sink
Kafka
HDFS

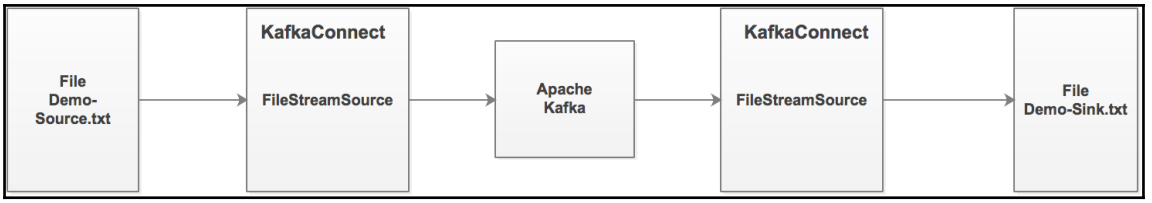

OS/Host
JVM
Web Server
Flow Controller
Processor 1
Extension N
FlowFile Repository
Content Repository
Provenance Repository
Local Storage


nifi
NiFi Flow
help
about
Search
Active threads: 0
Queued: 0 / 0 bytes
Stats last refreshed: 17:57:30 PDT
PutHDFS
PutHDFS
SplitText
SplitText
TailFile
TailFile
RouteOnContent
RouteOnContent
PublishKafka
PublishKafka
PublishKafka
PublishKafka
PublishKafka
PublishKafka
Name failure
Queued 0 / 0 bytes
Name success
Queued 0 / 0 bytes
Name splits
Queued 0 / 0 bytes
Name Info
Queued 0 / 0 bytes
Name ERROR
Queued 0 / 0 bytes
Name Warning
Queued 0 / 0 bytes
In
Read/Write
Out
Tasks/Time
0 / 0 bytes
0 bytes / 0 bytes
0 / 0 bytes
0 / 00:00:00.000
(5 min)

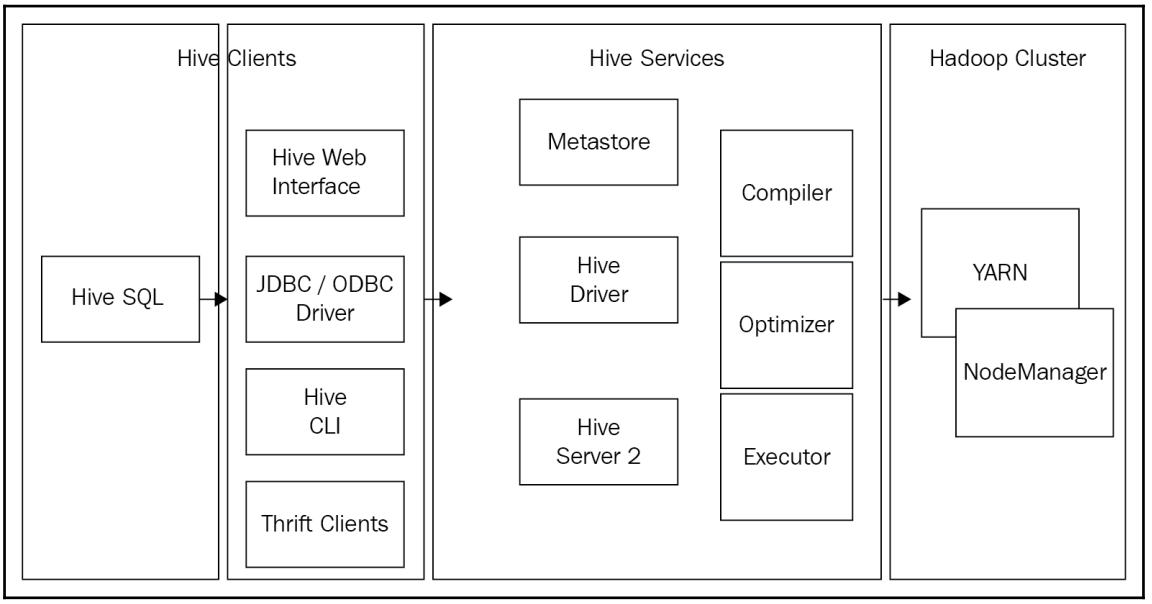

Streams App
3
Sources
Source 1
Source 2
Source 3
Source
1
Connect Cluster
Worker
Worker
Worker
2
Kafka Cluster
Broker
Broker
Broker
3
Streams App
4
Connect Cluster
Worker
Worker
Worker
5
Sinks
Sink 1
Sink 2
Sink 3
Sink

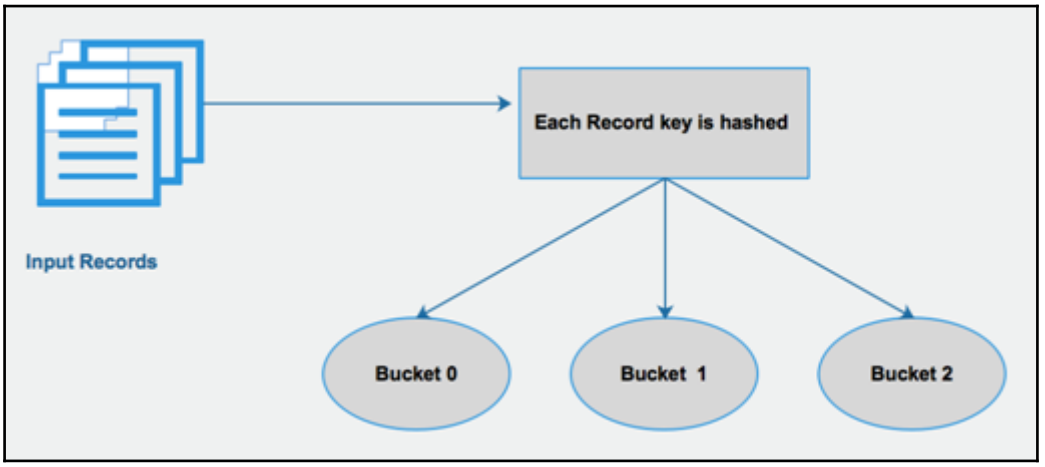

Kafka Connect Cluster
Worker 1
Connector 1 Task 1
Connector 3 Task 1
Worker 2
Connector 2 Task 1
Connector 3 Task 2
Worker 3
Connector 1 Task 2
Connector 2 Task 2
Connector 3 Task 3
Worker 4
Connector 1 Task 3
Connector 3 Task 4

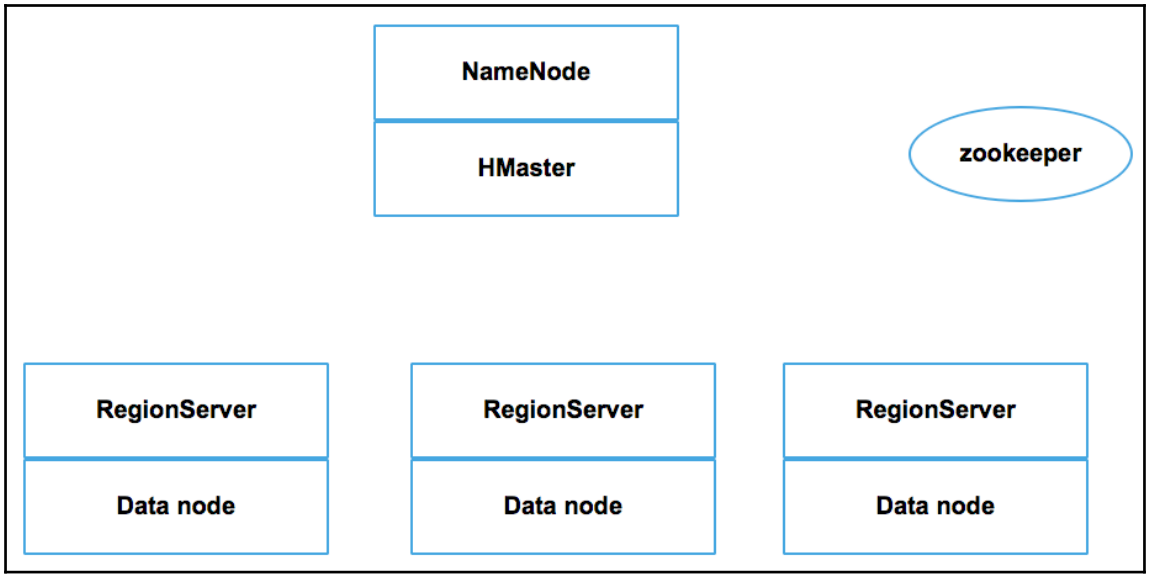

Kafka Connect Cluster after Rebalance
Worker 1
Worker 2
Worker 3
Worker 4
Connector 1 Task 1
Connector 3 Task 1
Connector 3 Task 4
Connector 2 Task 1
Connector 3 Task 2
Connector 1 Task 3
Connector 1 Task 2
Connector 2 Task 2
Connector 3
Tasks are moved after rebalance

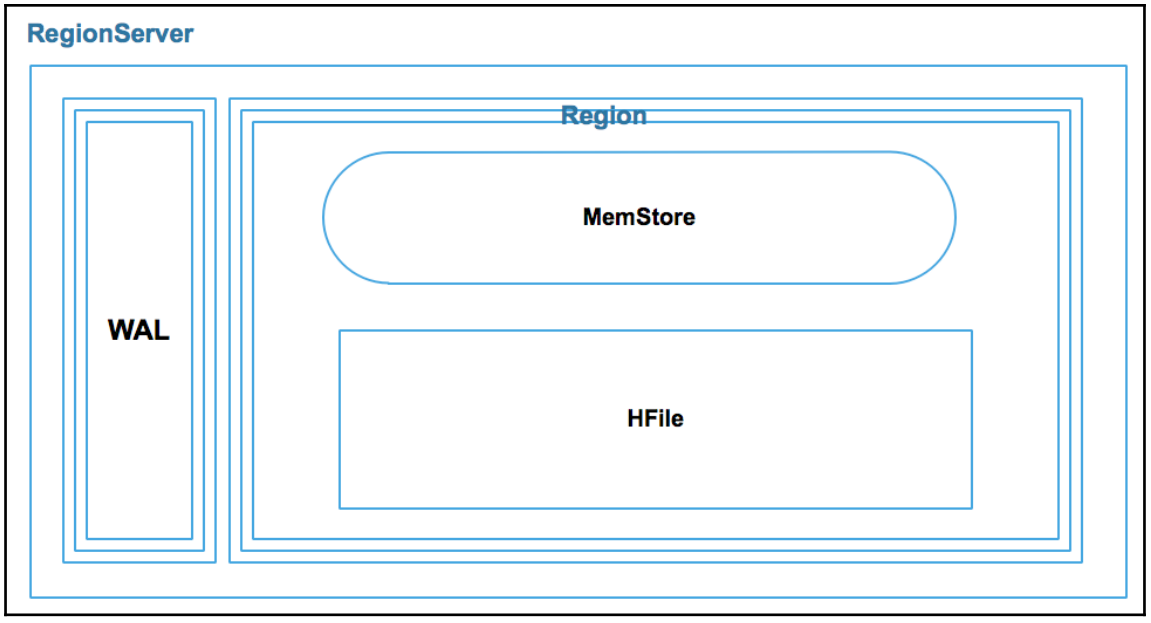

File Demo-Source.txt
KafkaConnect
FileStreamSource
Apache Kafka
KafkaConnect
FileStreamSource
File Demo-Sink.txt

Chapter 5: Data Modeling in Hadoop

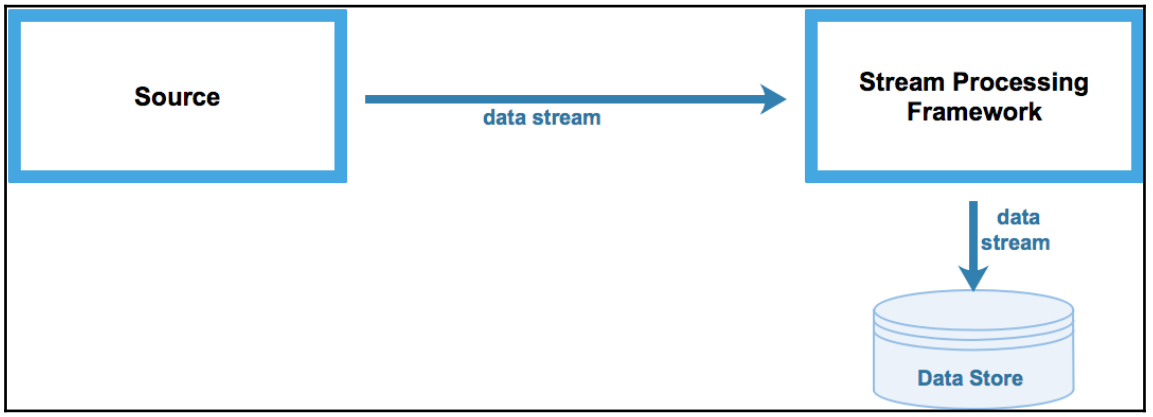


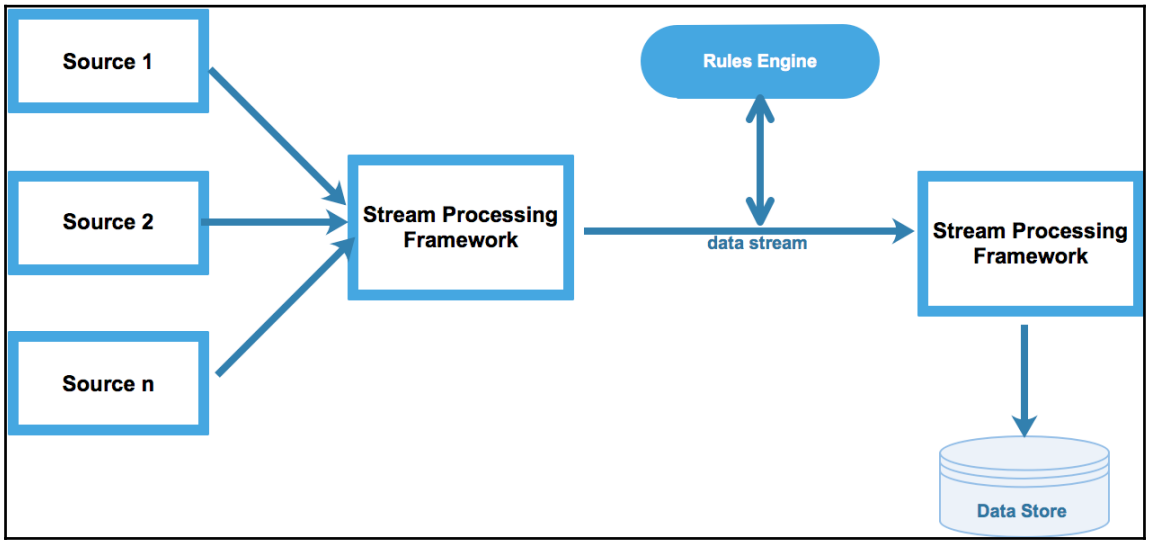


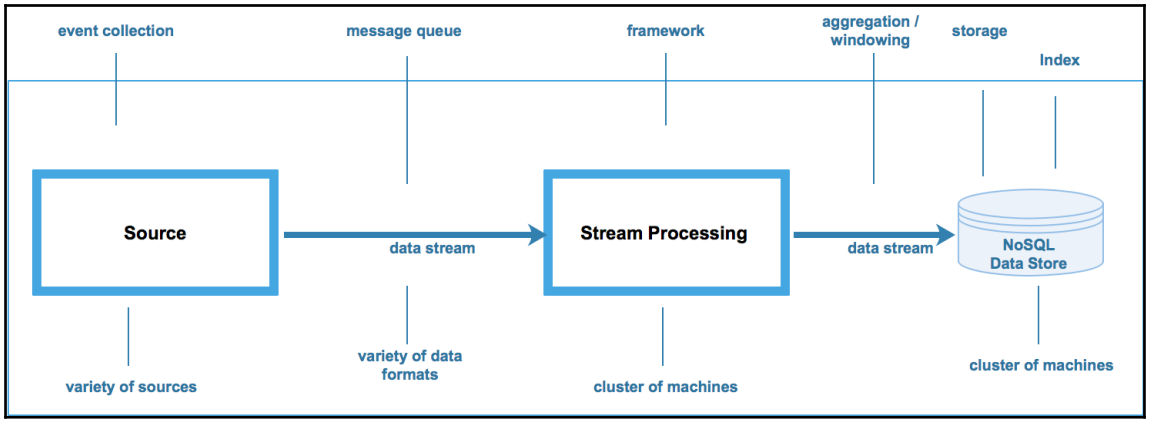

NameNode
HMaster
zookeeper
RegionServer
Data node
RegionServer
Data node
RegionServer
Data node

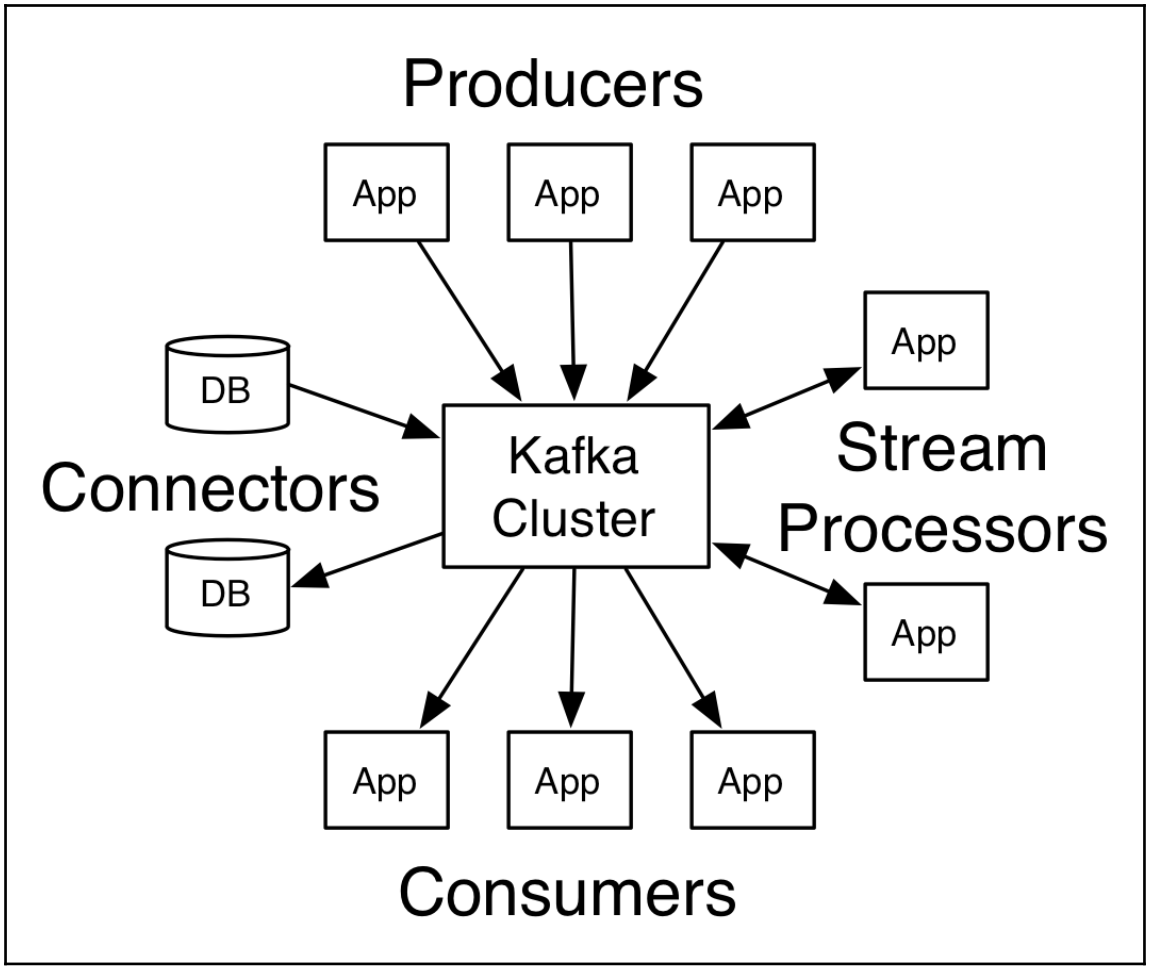

RegionServer
Region
MemStore
WAL
HFile

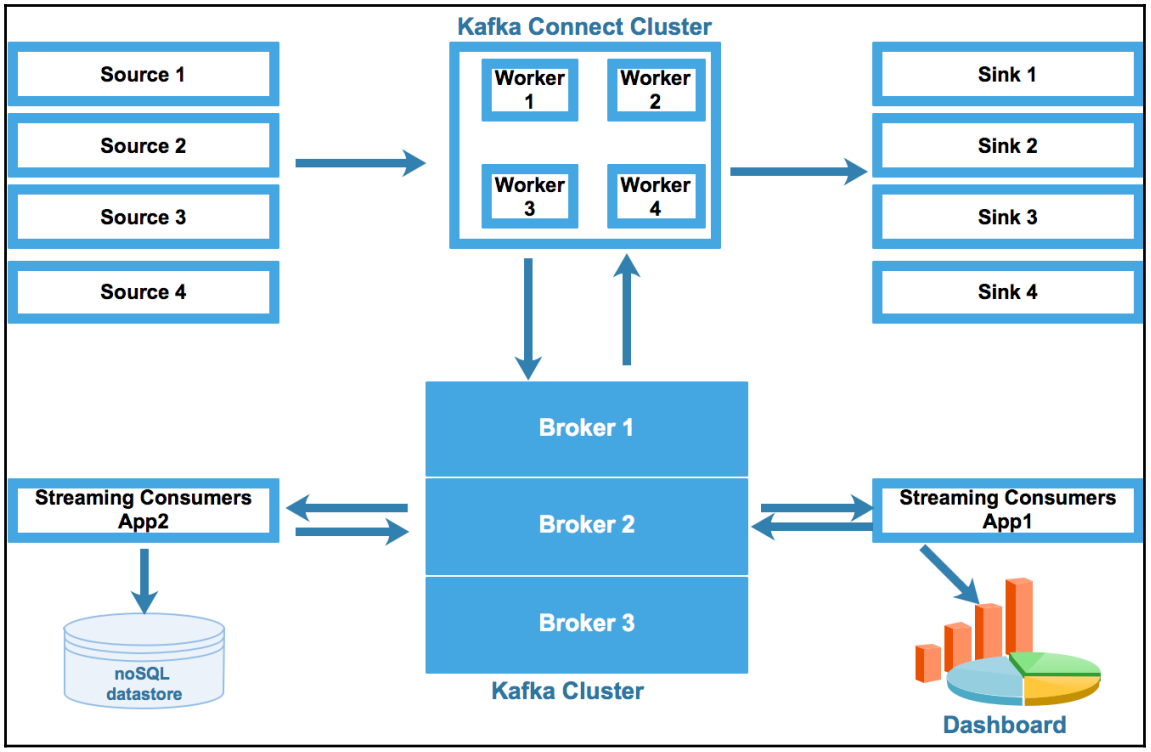
Chapter 6: Designing Real-Time Streaming Data Pipelines

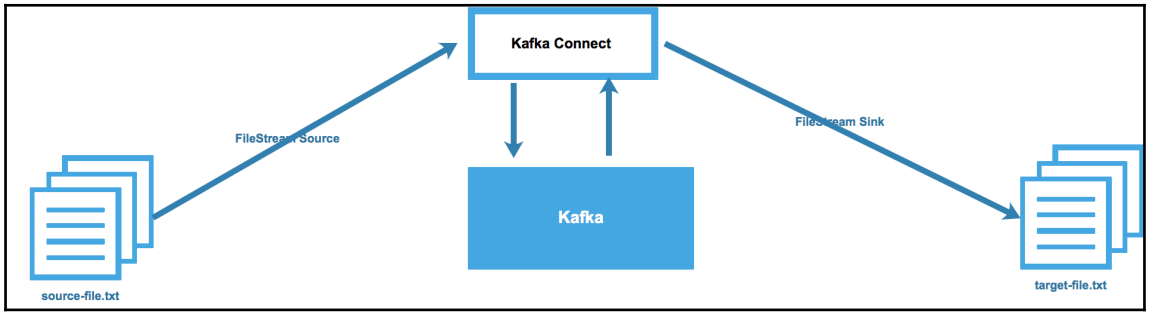


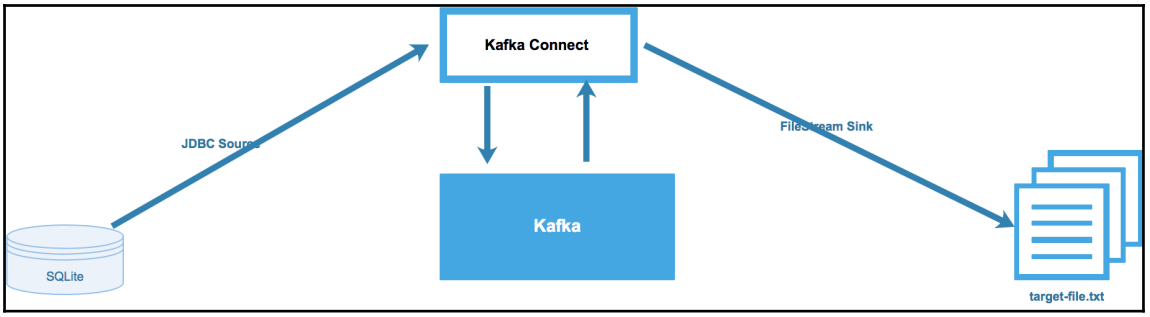


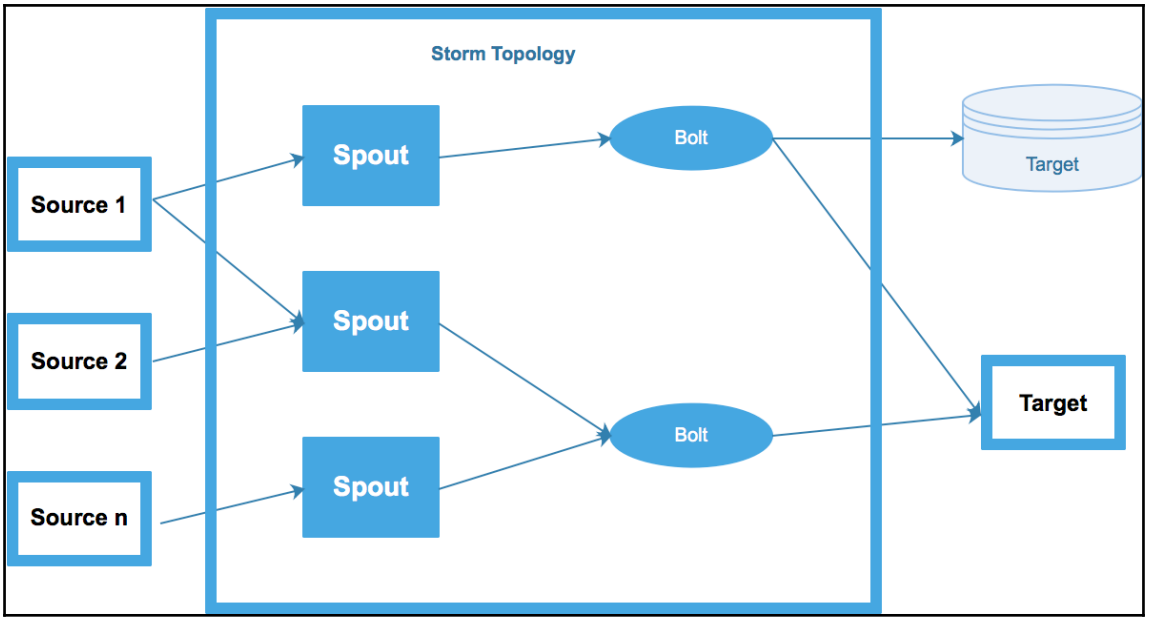

event collection
message queue
framework
aggregation / windowing
storage
Index
Source
data stream
Stream Processing
data stream
NoSQL Data Store
variety of sources
variety of data formats
cluster of machines
cluster of machines

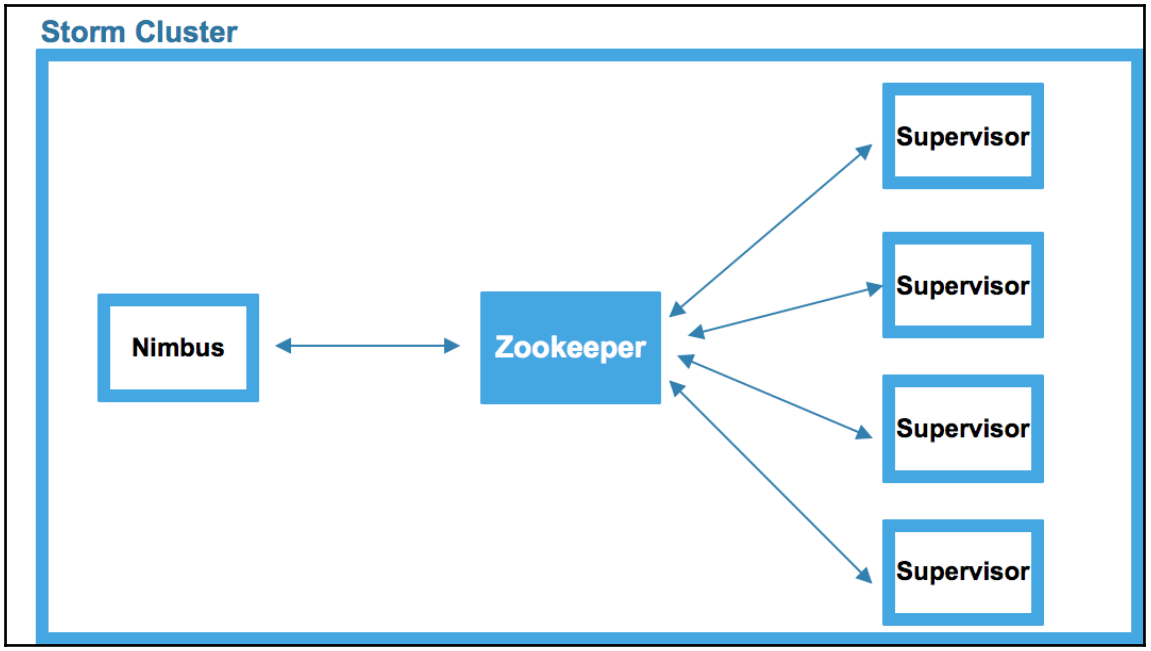

Producers
App
App
App
App
DB
Kafka
Cluster
Stream
Processors
Connectors
DB
App
App
App
App
Consumers

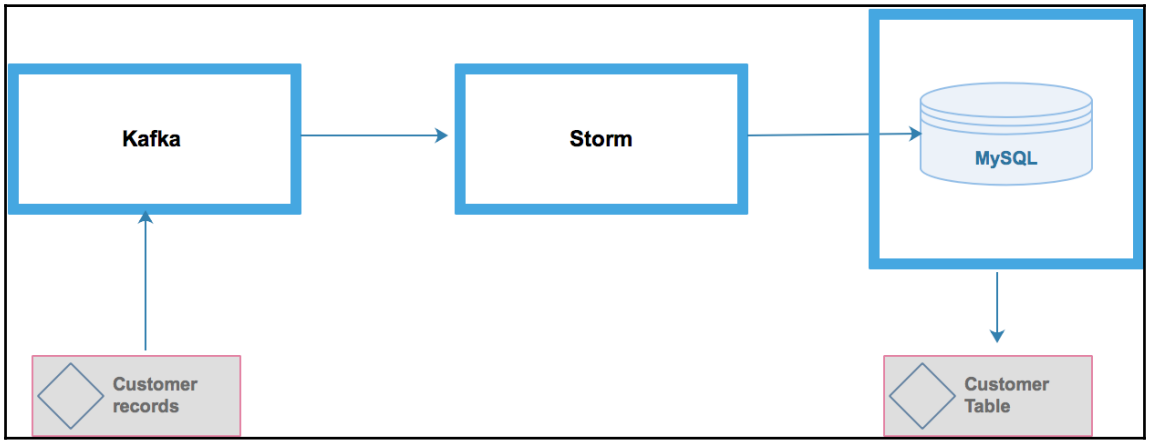

Kafka Connect Cluster
Source 1
Source 2
Source 3
Source 4
Worker 1
Worker 2
Worker 3
Worker 4
Sink 1
Sink 2
Sink 3
Sink 4
Broker 1
Broker 2
Broker 3
Kafka Cluster
Streaming Consumers App2
Streaming Consumers App1
noSQL datastore
Dashboard

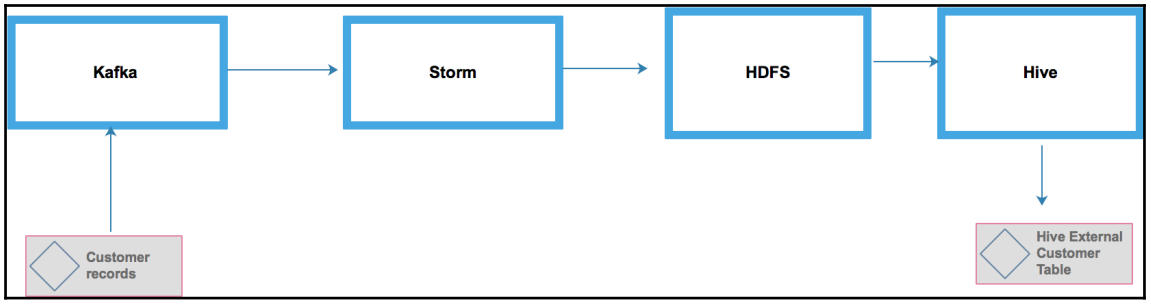

Kafka Connect
FileStream Source
FileStream Sink
Kafka
source-file.txt
target-file.txt

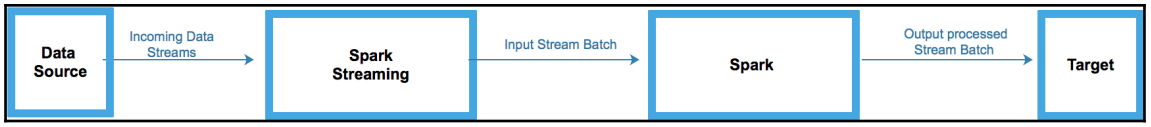

Kafka Connect
JDBC Source
FileStream Sink
Kafka
SQLite
target-file.txt

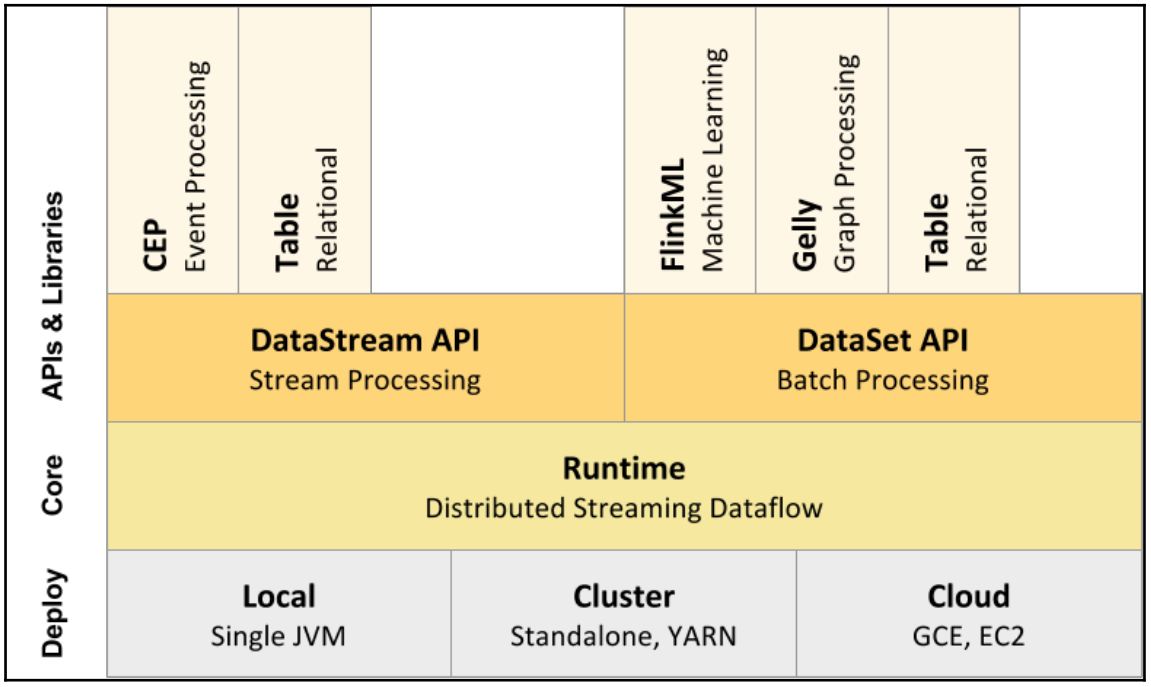

Storm Topology
Source 1
Source 2
Source n
Spout
Spout
Spout
Bolt
Bolt
Target
Target


Storm Cluster
Supervisor
Supervisor
Supervisor
Supervisor
Nimbus
Zookeeper


Kafka
Storm
MySQL
Customer records
Customer Table

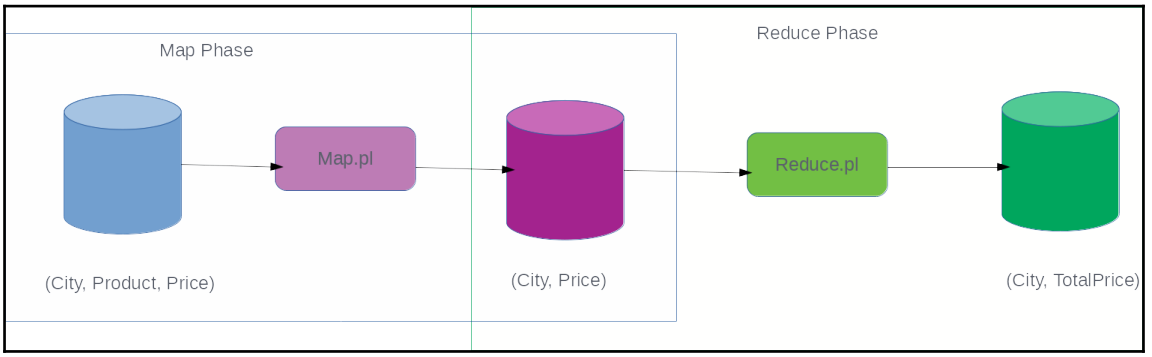

Kafka
Storm
HDFS
Hive
Customer records
Hive External Customer Table

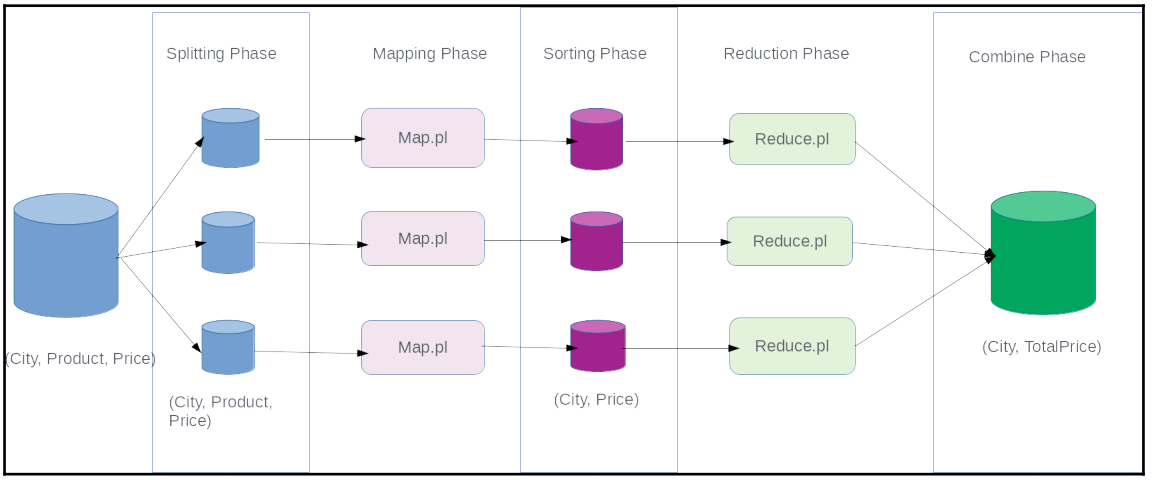

Data Source
Incoming Data Streams
Spark Streaming
Input Stream Batch
Spark
Output processed Stream Batch
Target

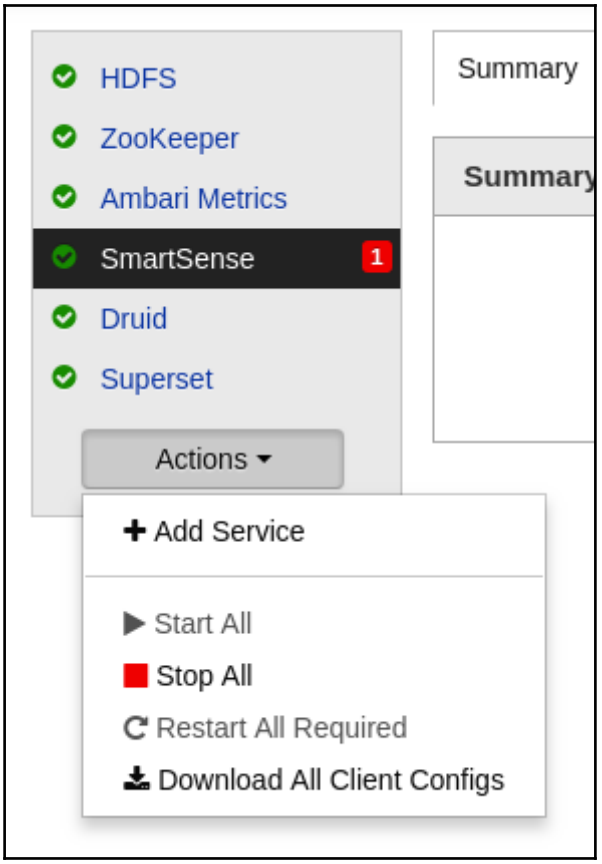

APIs & Libraries
CEP Event Processing
Table Relational
FlinkML Machine Learning
Gelly Graph Processing
Table Relational
DataStream API
Stream Processing
DataSet API
Batch Processing
Core
Runtime
Distributed Streaming Dataflow
Deploy
Local
Single JVM
Cluster
Standalone, YARN
Cloud
GCE, EC2

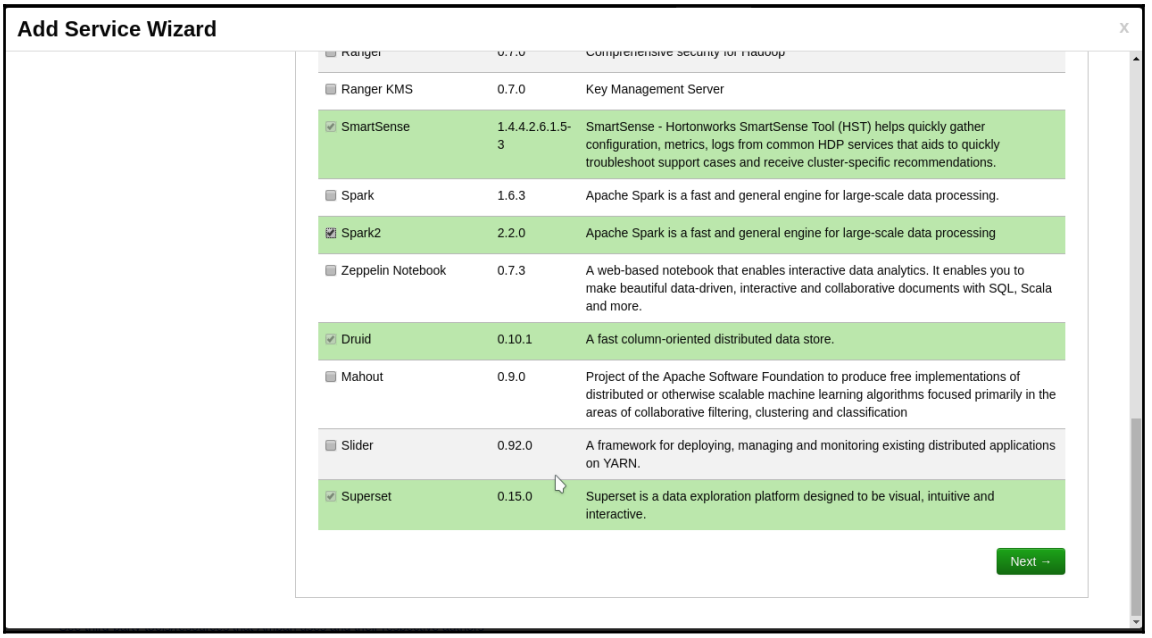

Source
Transformation
Sink

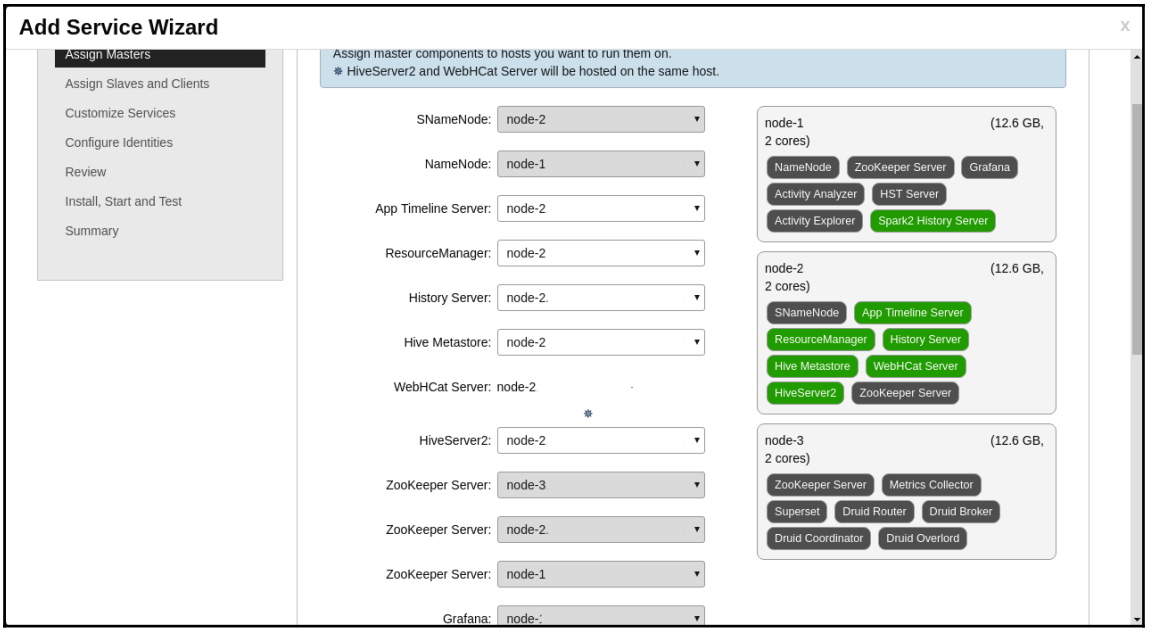
Chapter 7: Large-Scale Data Processing Frameworks

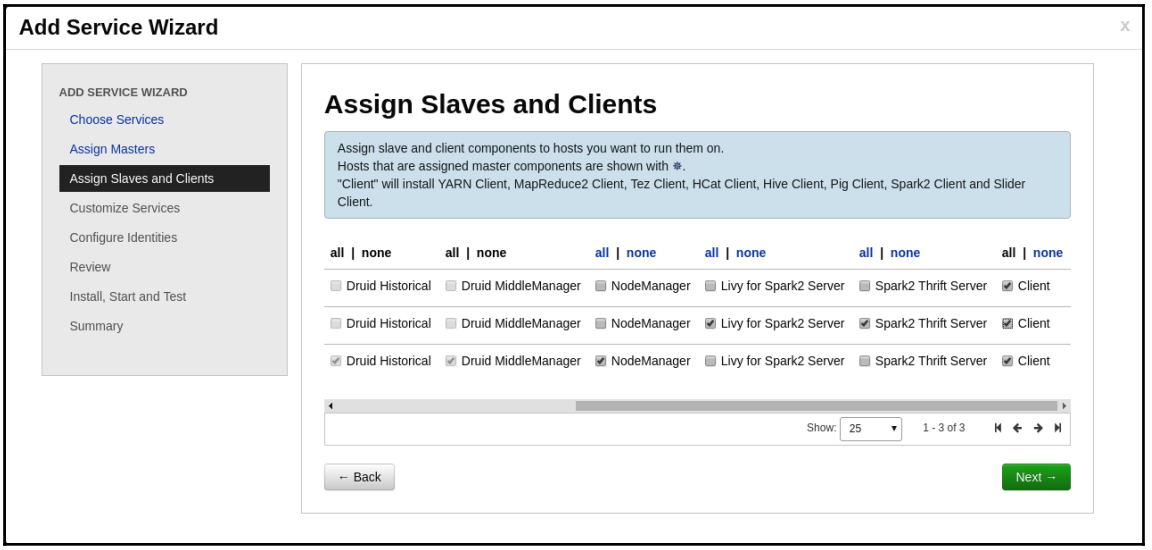


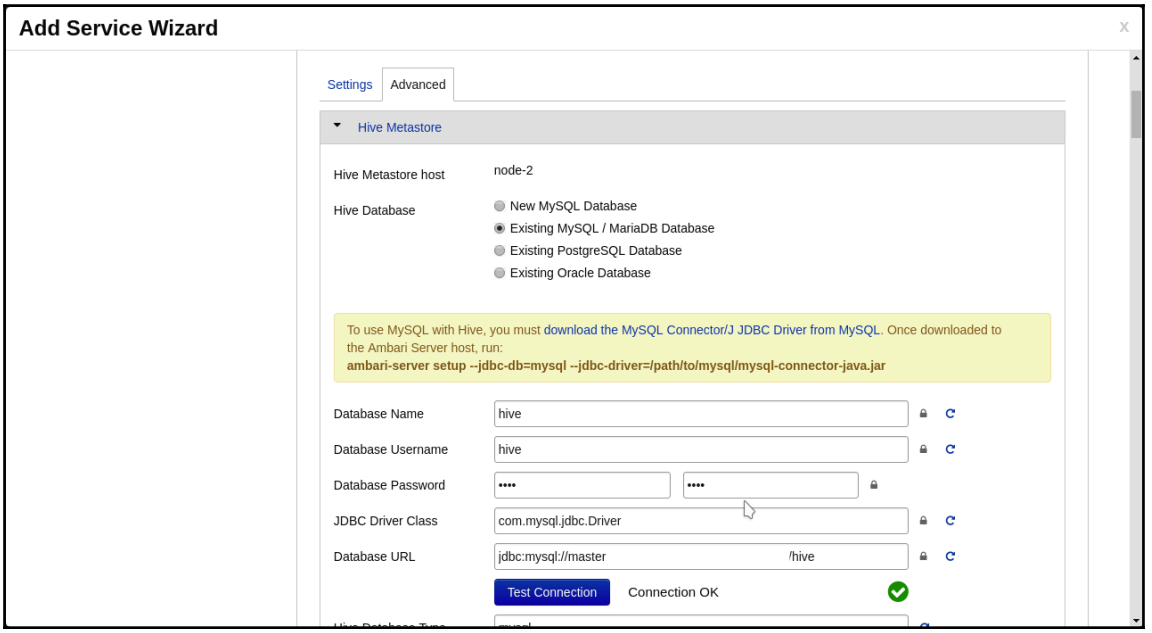


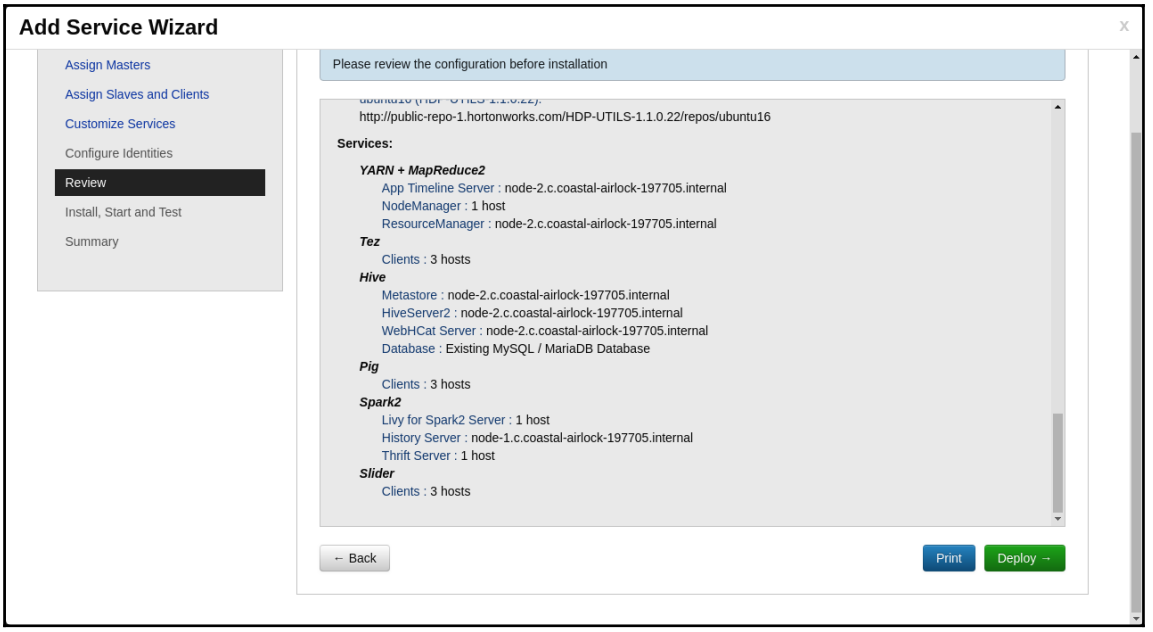

HDFS
ZooKeeper
Ambari Metrics
SmartSense
1
Druid
Superset
Actions
Add Service
Start All
Stop All
Restart All Required
Download All Client Configs
Summary
Summary

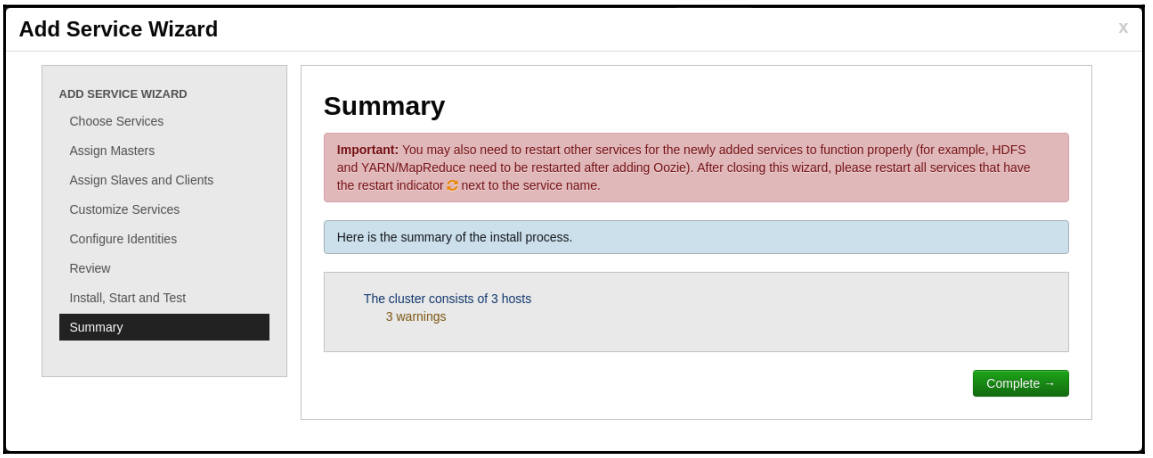

Add Service Wizard
Ranger KMS 0.7.0 Key Management Server
SmartSense 1.4.4.2.6.1.5-3 SmartSense - Hortonworks SmartSense Tool (HST) helps quickly gather configuration, metrics, logs from common HDP services that aids to quickly troubleshoot support cases and receive cluster-specific recommendations.
Spark 1.6.3 Apache Spark is a fast and general engine for large-scale data processing.
Spark2 2.2.0 Apache Spark is a fast and general engine for large-scale data processing
Zeppelin Notebook 0.7.3 A web-based notebook that enables interactive data analytics. It enables you to make beautiful data-driven, interactive and collaborative documents with SQL, Scala and more.
Druid 0.10.1 A fast column-oriented distributed data store.
Mahout 0.9.0 Project of the Apache Software Foundation to produce free implementations of distributed or otherwise scalable machine learning algorithms focused primarily in the areas of collaborative filtering, clustering and classification
Slider 0.92.0 A framework for deploying, managing and monitoring existing distributed applications on YARN.
Superset 0.15.0 Superset is a data exploration platform designed to be visual, intuitive and interactive.
Next →

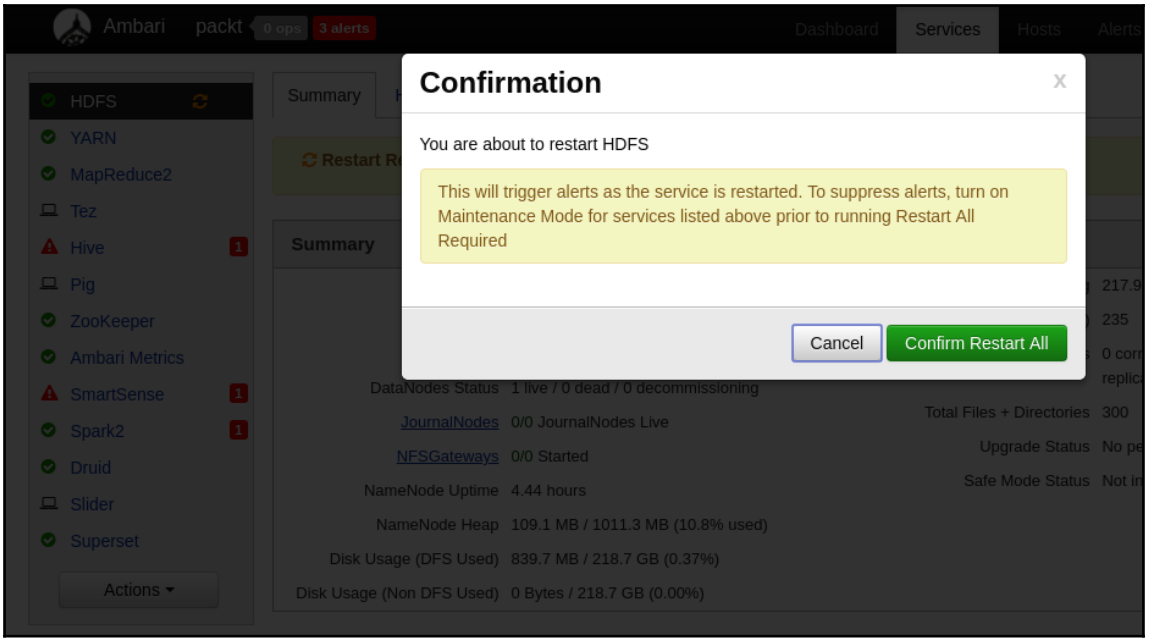

Add Service Wizard
Assign Masters
Assign Slaves and Clients
Customize Services
Configure Identities
Review
Install, Start and Test
Summary
Assign master components to hosts you want to run them on.
HiveServer2 and WebHCat Server will be hosted on the same host.
SNameNode: node-2
NameNode: node-1
App Timeline Server: node-2
ResourceManager: node-2
History Server: node-2
Hive Metastore: node-2
WebHCat Server: node-2
HiveServer2: node-2
ZooKeeper Server: node-3
ZooKeeper Server: node-2
ZooKeeper Server: node-1
Grafana: node-1
node-1 (12.6 GB, 2 cores)
NameNode ZooKeeper Server Grafana Activity Analyzer HST Server Activity Explorer Spark2 History Server
node-2 (12.6 GB, 2 cores)
SNameNode App Timeline Server ResourceManager History Server Hive Metastore WebHCat Server HiveServer2 ZooKeeper Server
node-3 (12.6 GB, 2 cores)
ZooKeeper Server Metrics Collector Superset Druid Router Druid Broker Druid Coordinator Druid Overlord

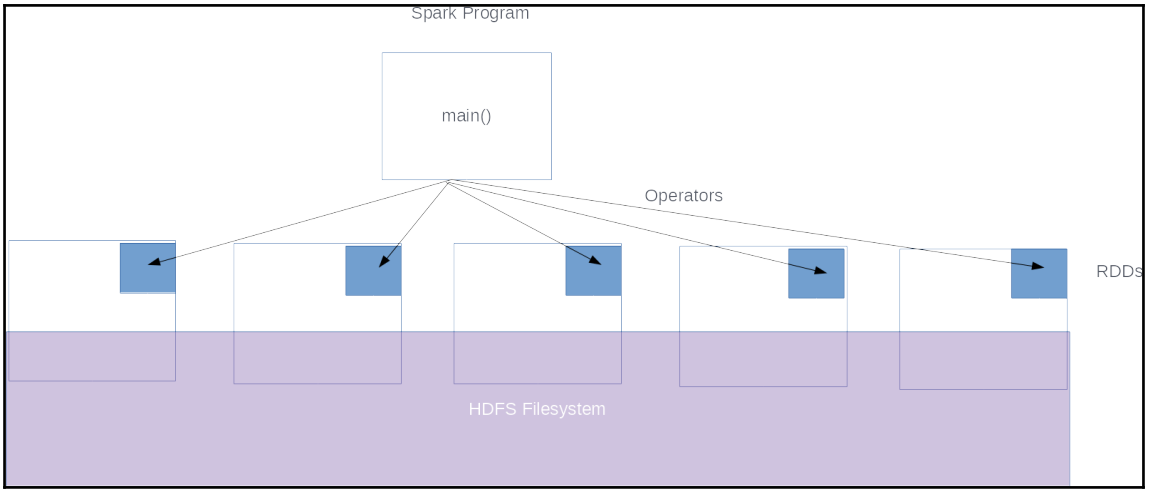

Add Service Wizard
ADD SERVICE WIZARD
Choose Services
Assign Masters
Assign Slaves and Clients
Customize Services
Configure Identities
Review
Install, Start and Test
Summary
Assign Slaves and Clients
Assign slave and client components to hosts you want to run them on.
Hosts that are assigned master components are shown with *.
"Client" will install YARN Client, MapReduce2 Client, Tez Client, HCat Client, Hive Client, Pig Client, Spark2 Client and Slider Client.
all | none
Druid Historical
Druid MiddleManager
NodeManager
Livy for Spark2 Server
Spark2 Thrift Server
Client
Show: 25
1 - 3 of 3
← Back
Next →

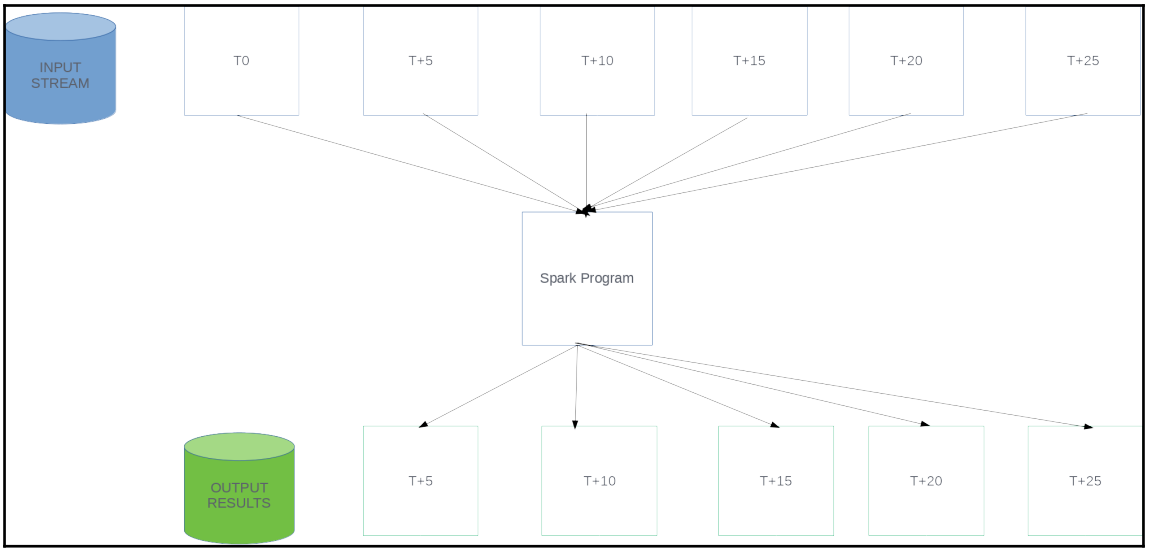

Add Service Wizard
Settings
Advanced
Hive Metastore
Hive Metastore host
node-2
Hive Database
New MySQL Database
Existing MySQL / MariaDB Database
Existing PostgreSQL Database
Existing Oracle Database
To use MySQL with Hive, you must download the MySQL Connector/J JDBC Driver from MySQL. Once downloaded to the Ambari Server host, run:
ambari-server setup --jdbc-db=mysql --jdbc-driver=/path/to/mysql/mysql-connector-java.jar
Database Name
hive
Database Username
hive
Database Password
JDBC Driver Class
com.mysql.jdbc.Driver
Database URL
jdbc:mysql://master /hive
Test Connection
Connection OK


Add Service Wizard
Assign Masters
Assign Slaves and Clients
Customize Services
Configure Identities
Review
Install, Start and Test
Summary
Please review the configuration before installation
http://public-repo-1.hortonworks.com/HDP-UTILS-1.1.0.22/repos/ubuntu16
Services:
YARN + MapReduce2
App Timeline Server : node-2.c.coastal-airlock-197705.internal
NodeManager : 1 host
ResourceManager : node-2.c.coastal-airlock-197705.internal
Tez
Clients : 3 hosts
Hive
Metastore : node-2.c.coastal-airlock-197705.internal
HiveServer2 : node-2.c.coastal-airlock-197705.internal
WebHCat Server : node-2.c.coastal-airlock-197705.internal
Database : Existing MySQL / MariaDB Database
Pig
Clients : 3 hosts
Spark2
Livy for Spark2 Server : 1 host
History Server : node-1.c.coastal-airlock-197705.internal
Thrift Server : 1 host
Slider
Clients : 3 hosts
← Back
Print
Deploy →

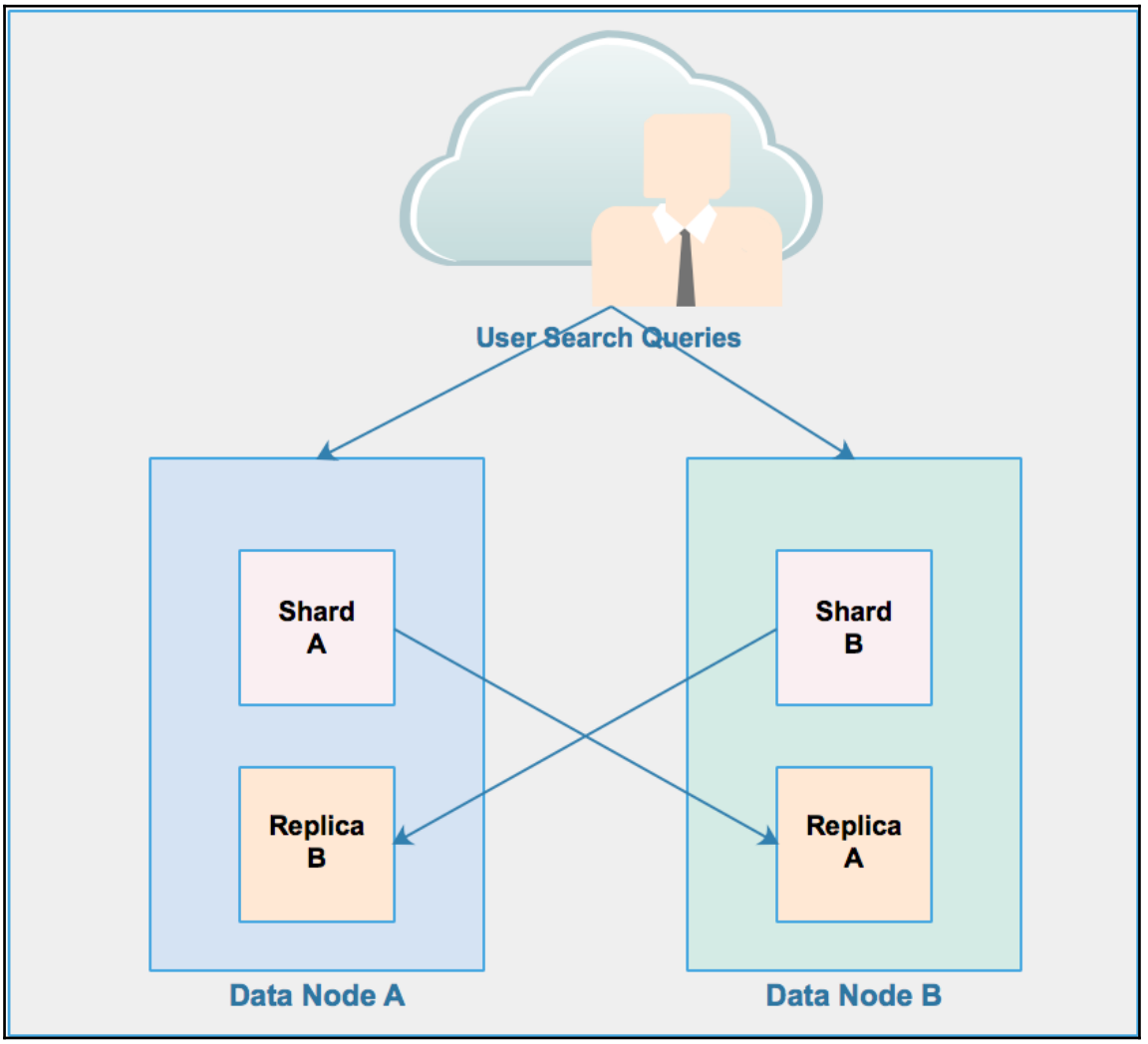

Add Service Wizard
ADD SERVICE WIZARD
Choose Services
Assign Masters
Assign Slaves and Clients
Customize Services
Configure Identities
Review
Install, Start and Test
Summary
Summary
Important: You may also need to restart other services for the newly added services to function properly (for example, HDFS and YARN/MapReduce need to be restarted after adding Oozie). After closing this wizard, please restart all services that have the restart indicator next to the service name.
Here is the summary of the install process.
The cluster consists of 3 hosts
3 warnings
Complete →

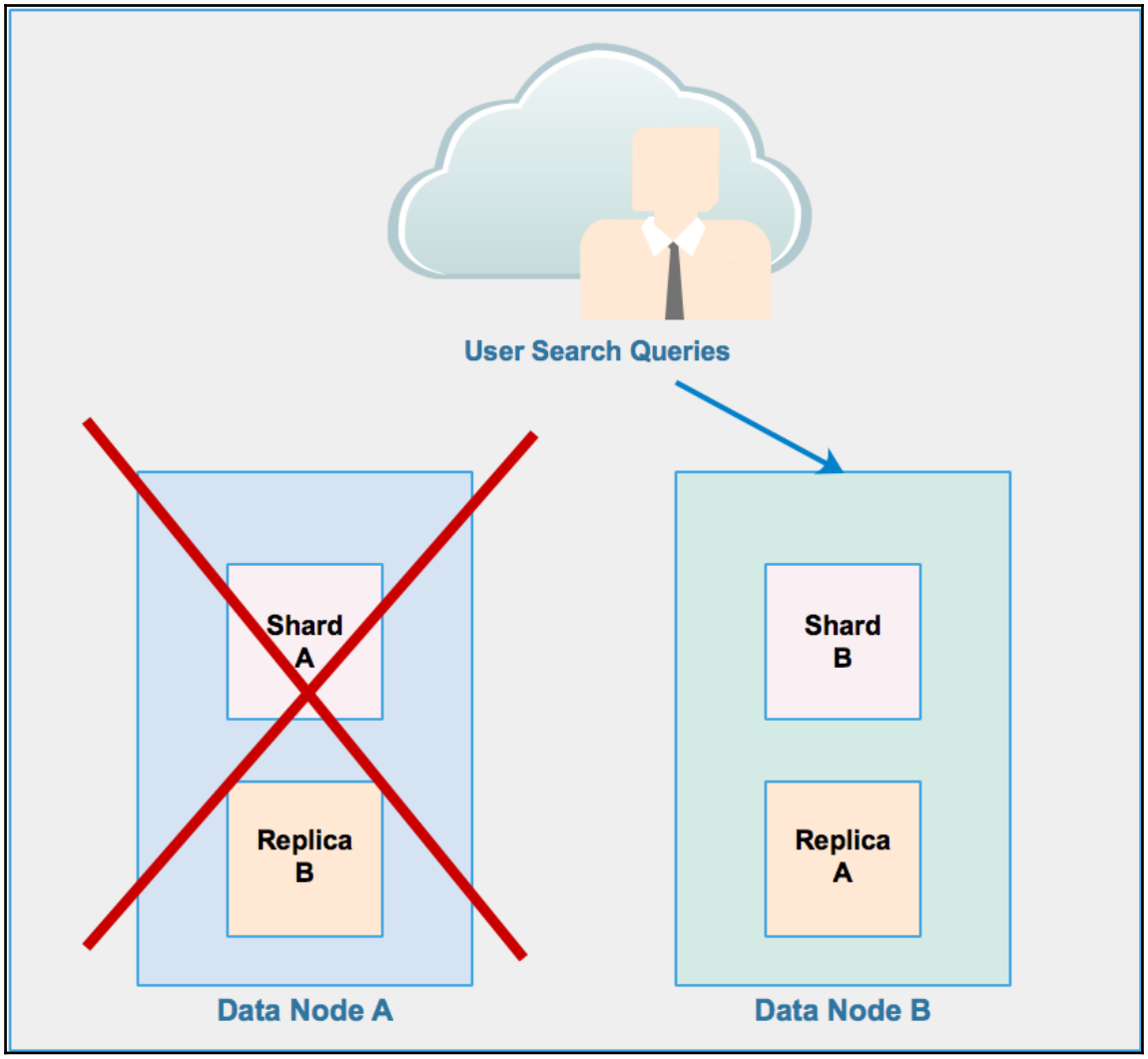

Ambari
packt
0 ops
3 alerts
Dashboard
Services
Hosts
Alerts
HDFS
YARN
MapReduce2
Tez
Hive
Pig
ZooKeeper
Ambari Metrics
SmartSense
Spark2
Druid
Slider
Superset
Actions
Summary
Restart R
Summary
Confirmation
You are about to restart HDFS
This will trigger alerts as the service is restarted. To suppress alerts, turn on Maintenance Mode for services listed above prior to running Restart All Required
Cancel
Confirm Restart All
DataNodes Status 1 live / 0 dead / 0 decommissioning
JournalNodes 0/0 JournalNodes Live
NFSGateways 0/0 Started
NameNode Uptime 4.44 hours
NameNode Heap 109.1 MB / 1011.3 MB (10.8% used)
Disk Usage (DFS Used) 839.7 MB / 218.7 GB (0.37%)
Disk Usage (Non DFS Used) 0 Bytes / 218.7 GB (0.00%)
Total Files + Directories 300
Upgrade Status
Safe Mode Status

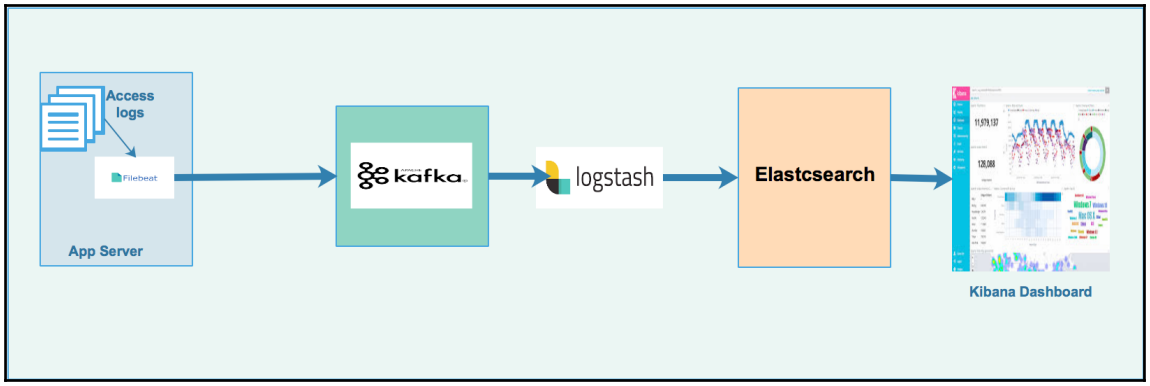

Spark Program
main()
Operators
RDDs
HDFS Filesystem


INPUT STREAM
T0
T+5
T+10
T+15
T+20
T+25
Spark Program
OUTPUT RESULTS
T+5
T+10
T+15
T+20
T+25

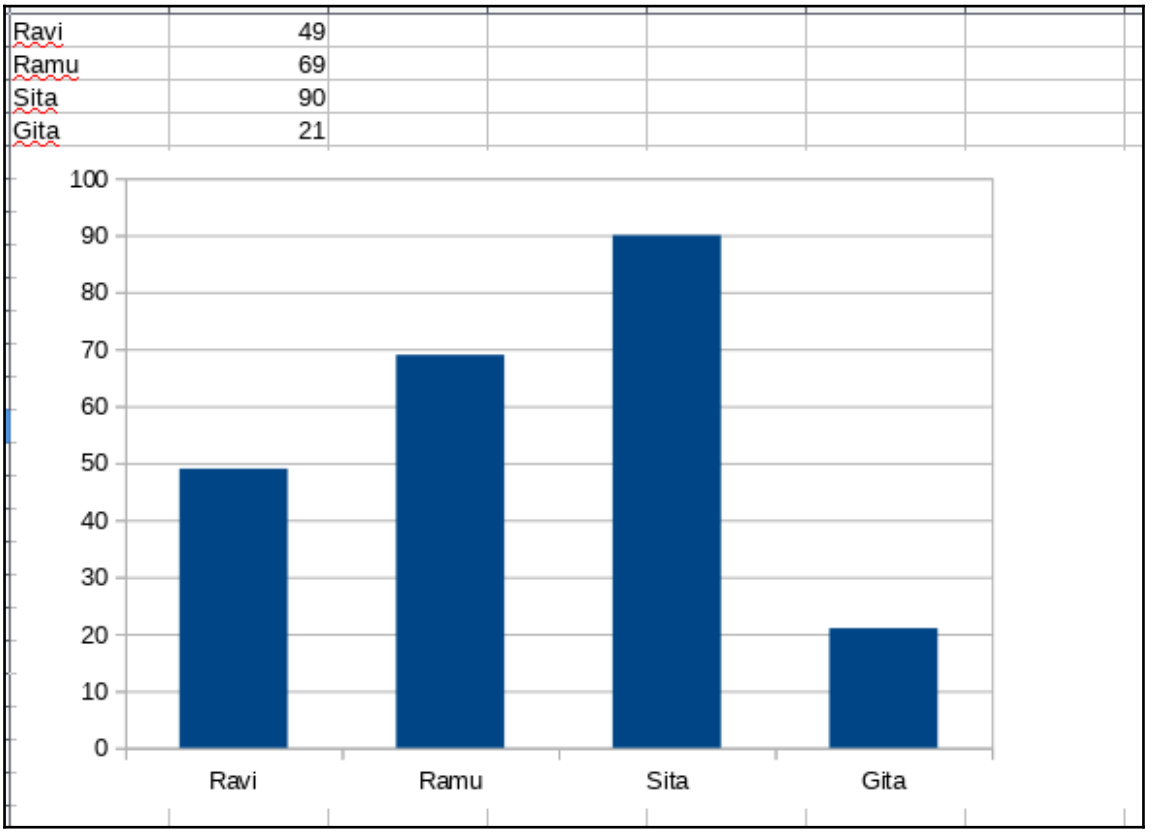
Chapter 8: Building Enterprise Search Platform



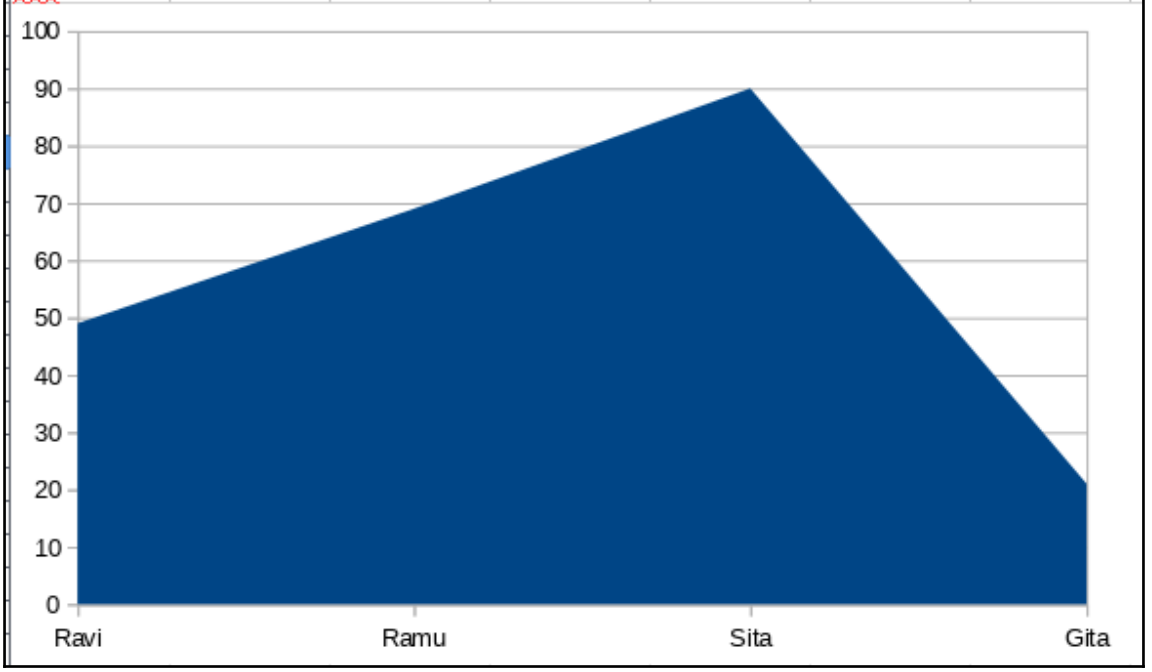

User Search Queries
Shard
A
Replica
B
Shard
B
Replica
A
Data Node A
Data Node B


Access logs
Filebeat
App Server
kafka
logstash
Elastcsearch
Kibana Dashboard

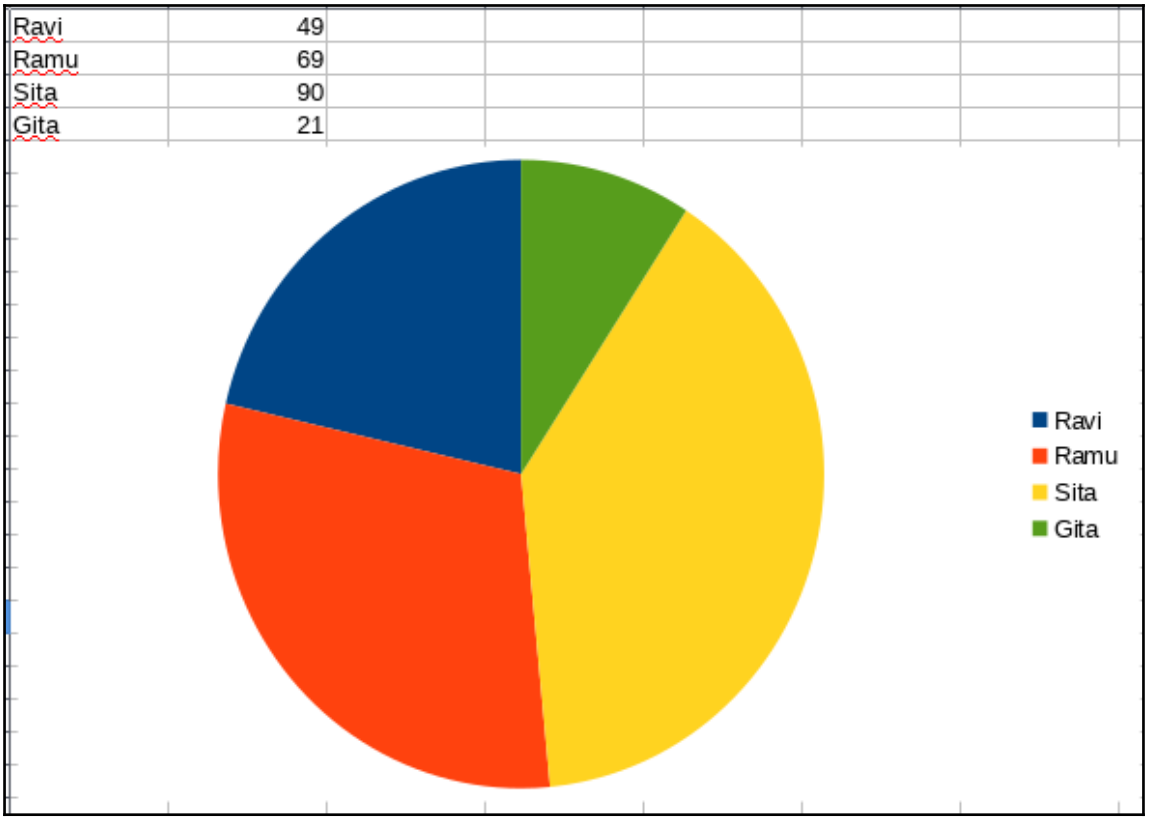
Chapter 9: Designing Data Visualization Solutions



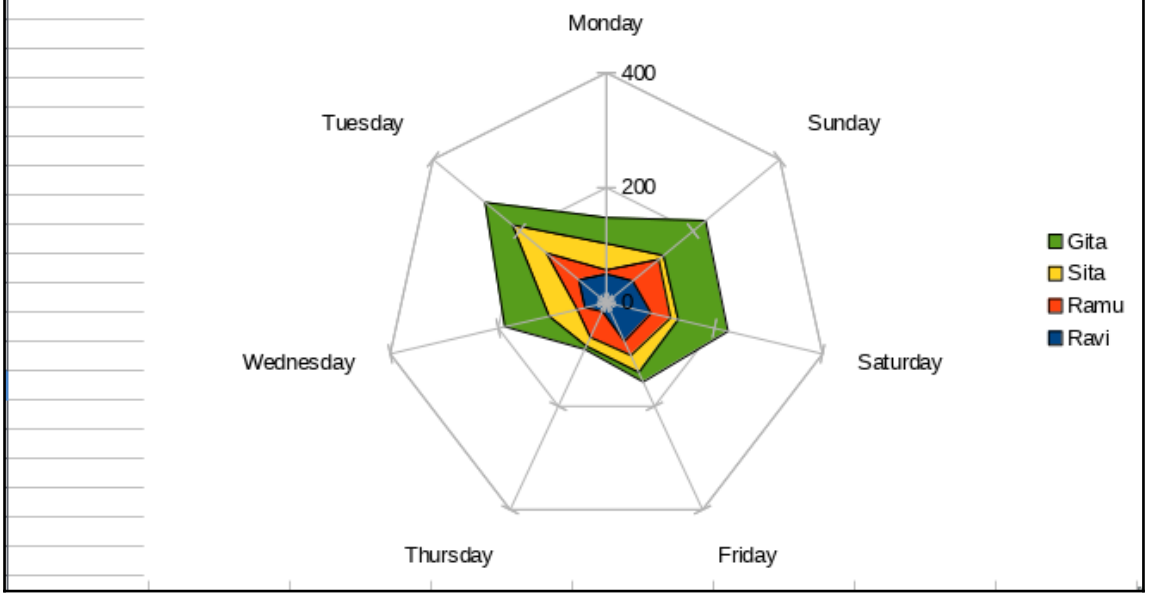
Ravi	49
Ramu	69
Sita	90
Gita	21



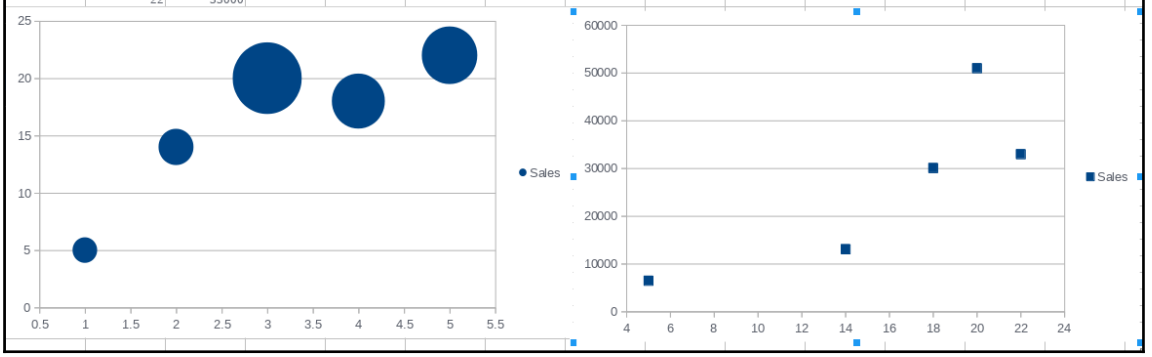
Ravi	49
Ramu	69
Sita	90
Gita	21

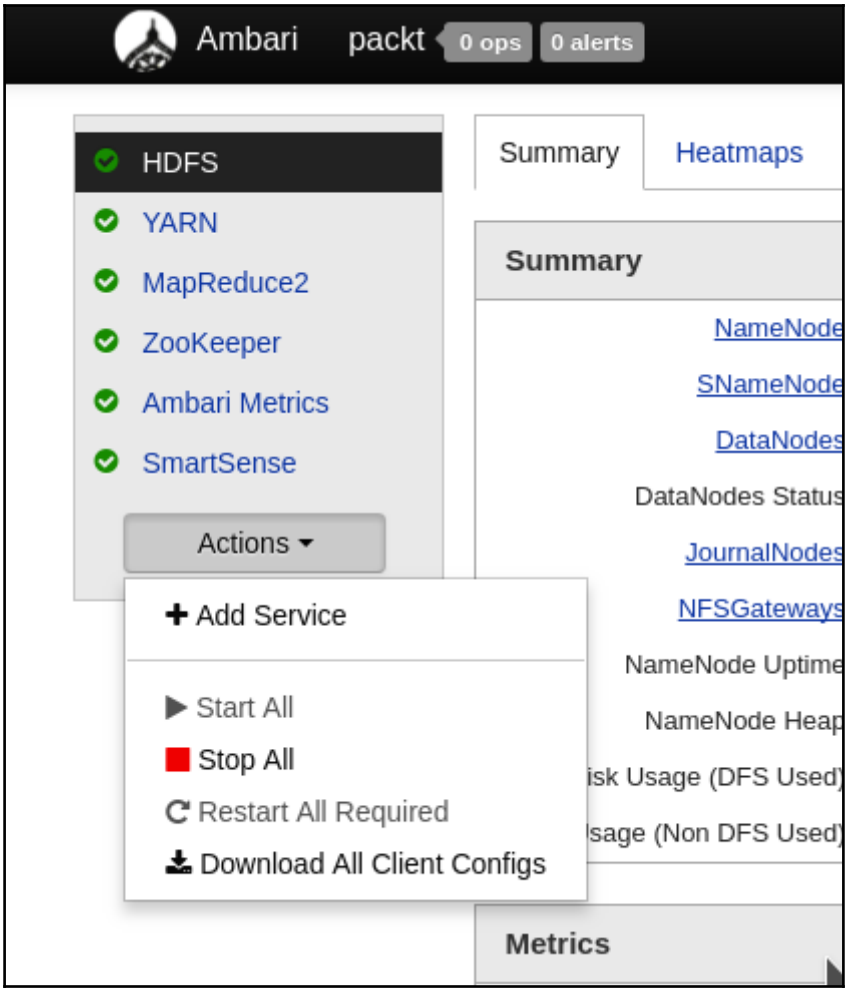


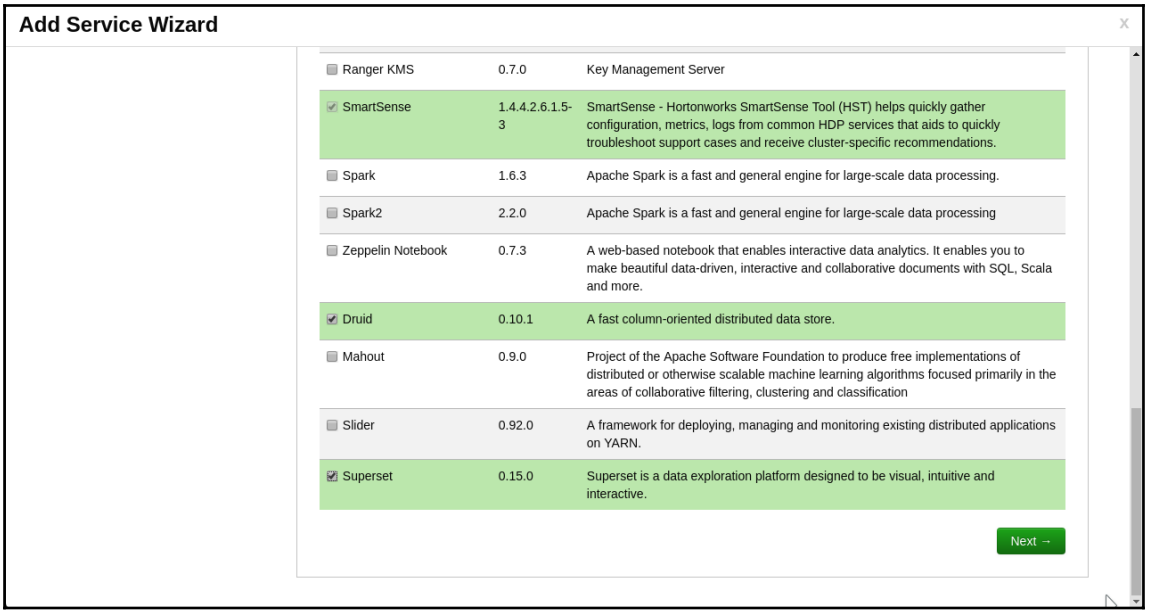
	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Ravi	50	64	40	17	75	83	61
Ramu	7	75	14	50	27	36	60
Sita	46	78	52	17	33	14	11
Gita	45	63	82	6	18	92	97

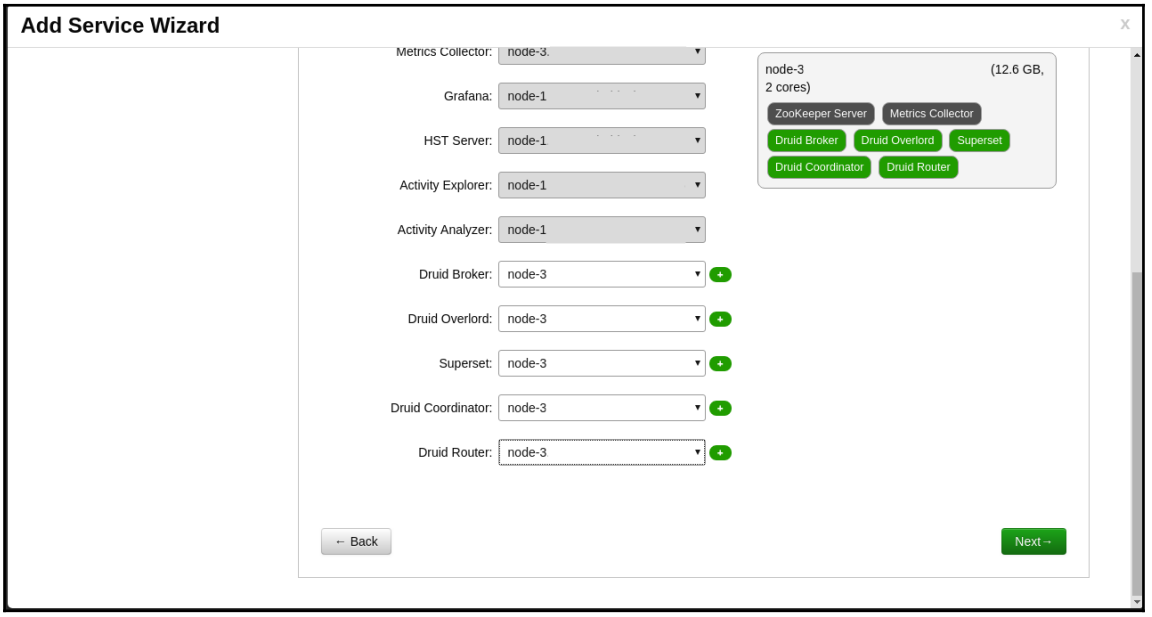


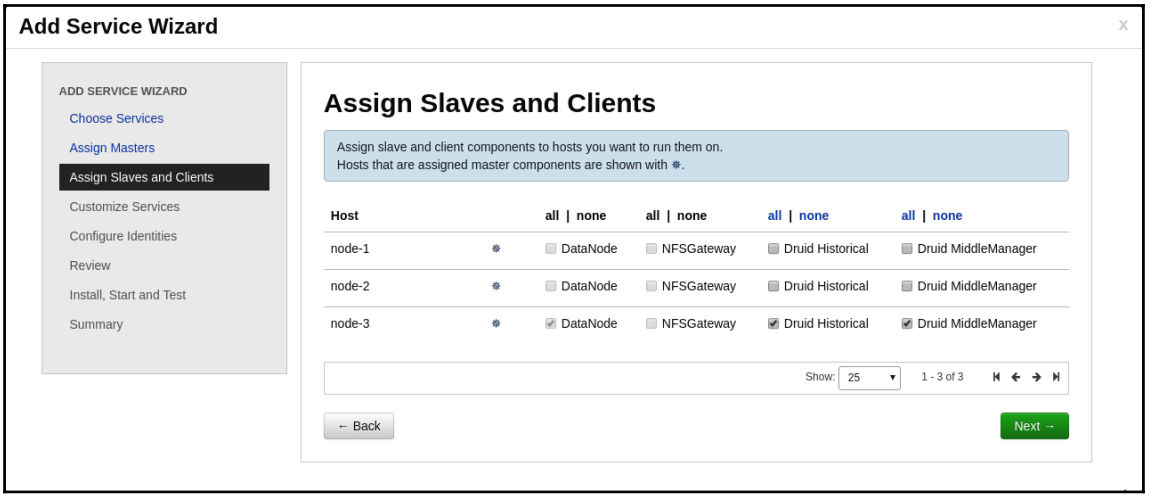
Number of Products	Sales
5	6500
14	13100
20	51000
18	30100
22	33000

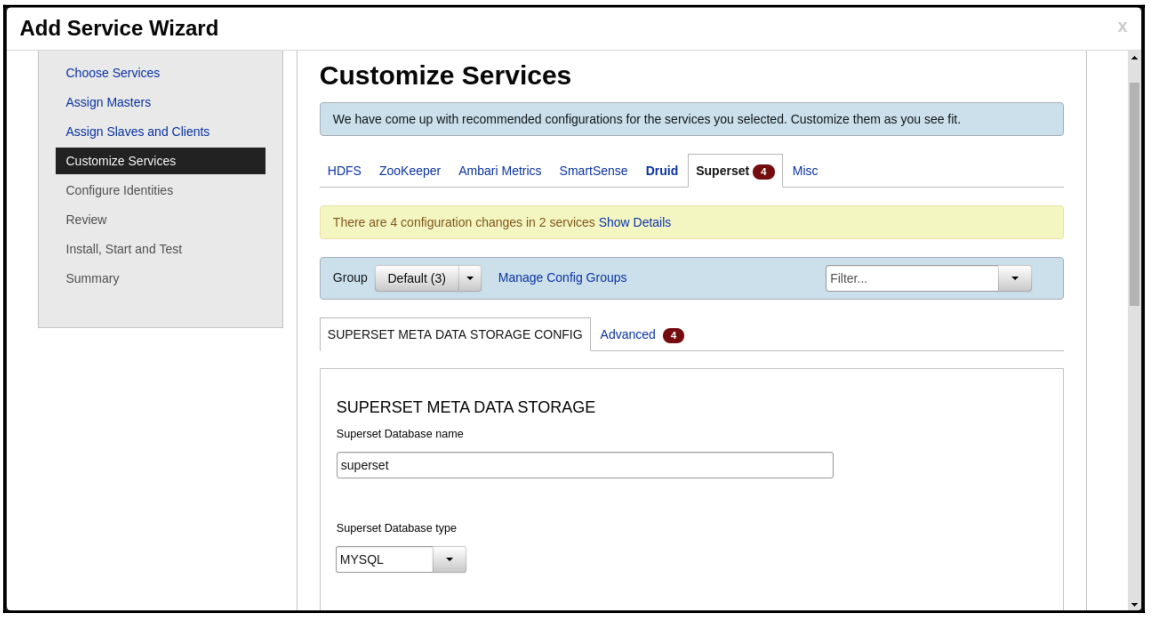


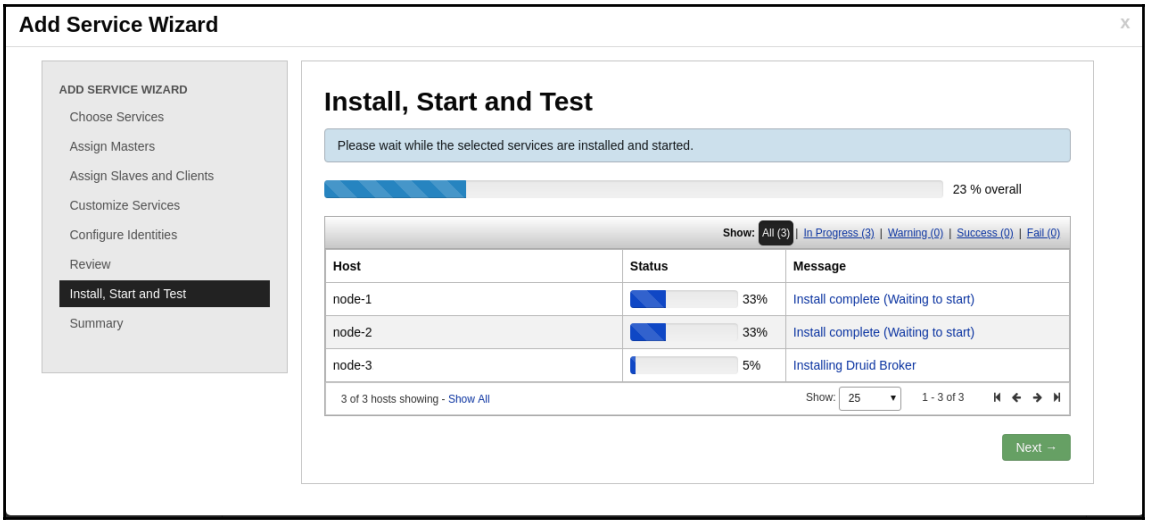

Ambari
packt
0 ops
0 alerts
HDFS
YARN
MapReduce2
ZooKeeper
Ambari Metrics
SmartSense
Actions
Add Service
Start All
Stop All
Restart All Required
Download All Client Configs
Summary
Heatmaps
Summary
NameNode
SNameNode
DataNodes
DataNodes Status
JournalNodes
NFSGateways
NameNode Uptime
NameNode Heap
Metrics

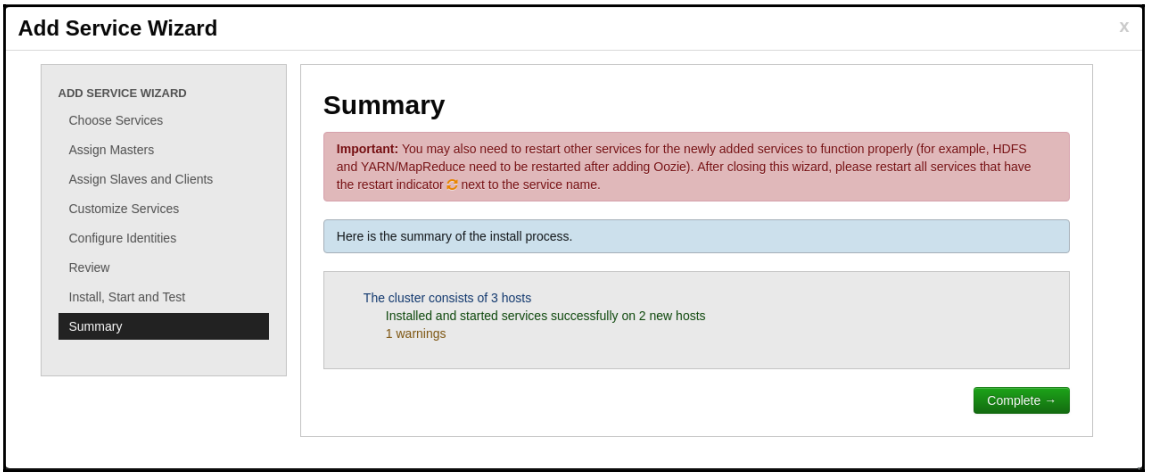

Add Service Wizard
Ranger KMS 0.7.0 Key Management Server
SmartSense 1.4.4.2.6.1.5-3 SmartSense - Hortonworks SmartSense Tool (HST) helps quickly gather configuration, metrics, logs from common HDP services that aids to quickly troubleshoot support cases and receive cluster-specific recommendations.
Spark 1.6.3 Apache Spark is a fast and general engine for large-scale data processing.
Spark2 2.2.0 Apache Spark is a fast and general engine for large-scale data processing
Zeppelin Notebook 0.7.3 A web-based notebook that enables interactive data analytics. It enables you to make beautiful data-driven, interactive and collaborative documents with SQL, Scala and more.
Druid 0.10.1 A fast column-oriented distributed data store.
Mahout 0.9.0 Project of the Apache Software Foundation to produce free implementations of distributed or otherwise scalable machine learning algorithms focused primarily in the areas of collaborative filtering, clustering and classification
Slider 0.92.0 A framework for deploying, managing and monitoring existing distributed applications on YARN.
Superset 0.15.0 Superset is a data exploration platform designed to be visual, intuitive and interactive.
Next →

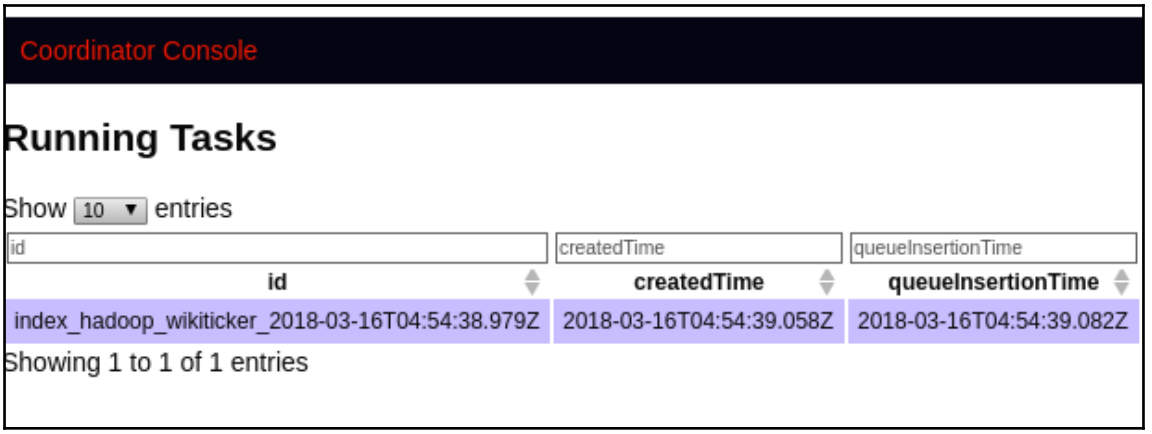

Add Service Wizard
Metrics Collector: node-3
Grafana: node-1
HST Server: node-1
Activity Explorer: node-1
Activity Analyzer: node-1
Druid Broker: node-3
Druid Overlord: node-3
Superset: node-3
Druid Coordinator: node-3
Druid Router: node-3
node-3 (12.6 GB, 2 cores)
ZooKeeper Server
Metrics Collector
Druid Broker
Druid Overlord
Superset
Druid Coordinator
Druid Router
← Back
Next→

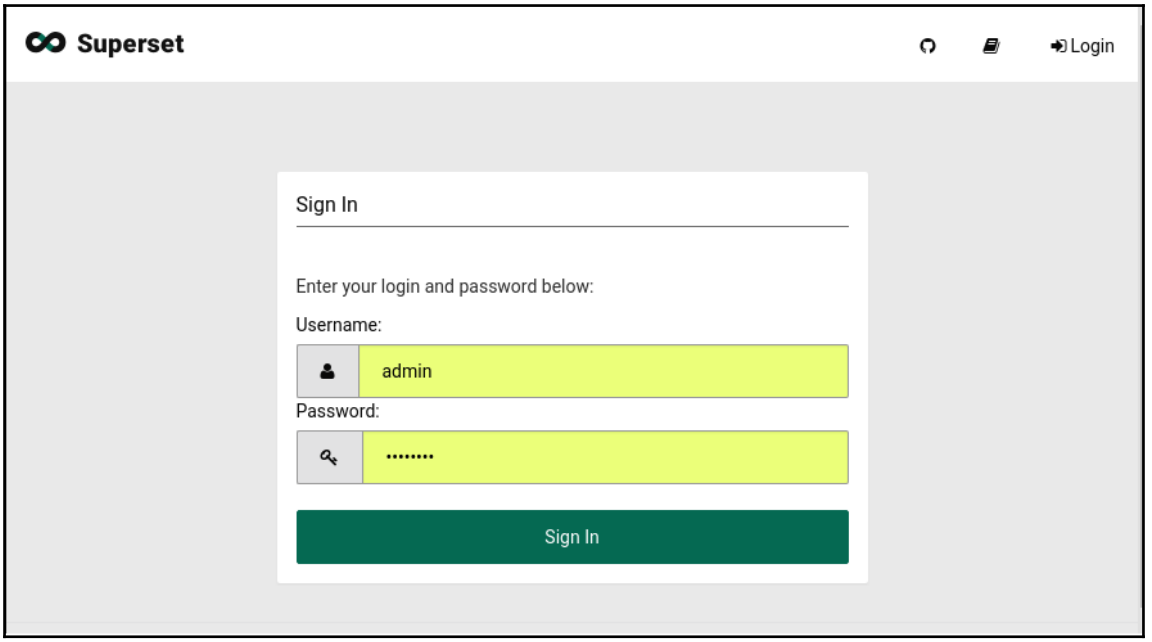

Add Service Wizard
ADD SERVICE WIZARD
Choose Services
Assign Masters
Assign Slaves and Clients
Customize Services
Configure Identities
Review
Install, Start and Test
Summary
Assign Slaves and Clients
Assign slave and client components to hosts you want to run them on.
Hosts that are assigned master components are shown with *.
Host
all | none
DataNode
NFSGateway
Druid Historical
Druid MiddleManager
node-1
node-2
node-3
Show: 25
1 - 3 of 3
← Back
Next →

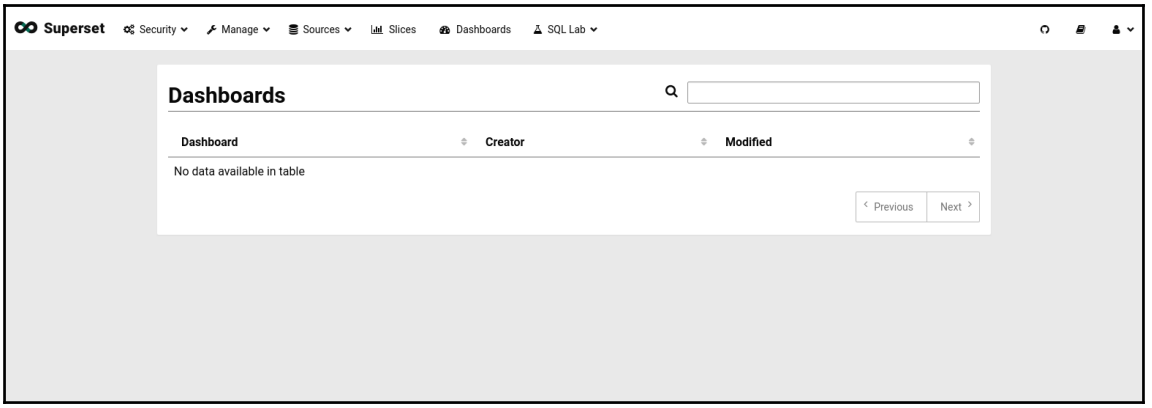

Add Service Wizard
Choose Services
Assign Masters
Assign Slaves and Clients
Customize Services
Configure Identities
Review
Install, Start and Test
Summary
Customize Services
We have come up with recommended configurations for the services you selected. Customize them as you see fit.
HDFS
ZooKeeper
Ambari Metrics
SmartSense
Druid
Superset 4
Misc
There are 4 configuration changes in 2 services Show Details
Group Default (3)
Manage Config Groups
Filter...
SUPERSET META DATA STORAGE CONFIG
Advanced 4
SUPERSET META DATA STORAGE
Superset Database name
superset
Superset Database type
MYSQL

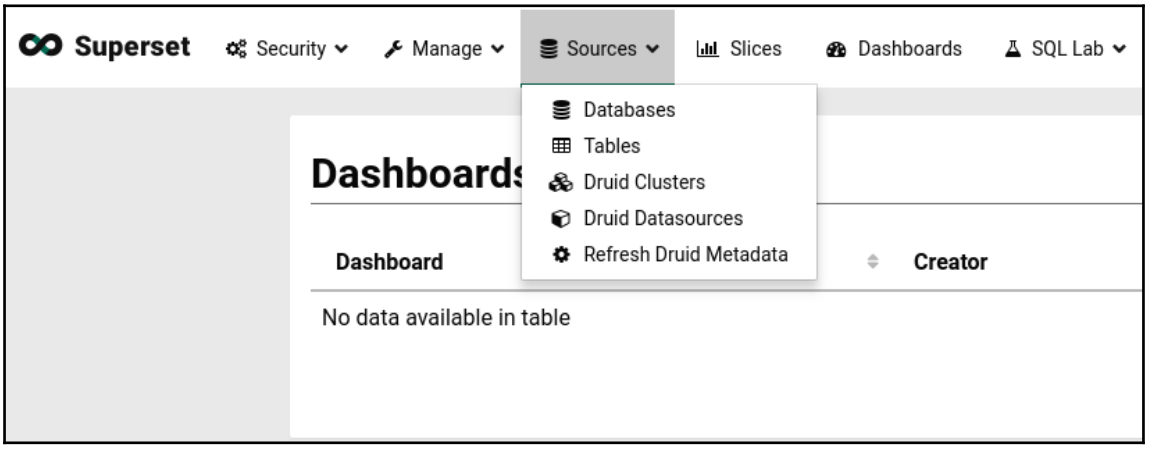

Add Service Wizard
ADD SERVICE WIZARD
Choose Services
Assign Masters
Assign Slaves and Clients
Customize Services
Configure Identities
Review
Install, Start and Test
Summary
Install, Start and Test
Please wait while the selected services are installed and started.
23 % overall
Show: All (3) | In Progress (3) | Warning (0) | Success (0) | Fail (0)
Host
Status
Message
node-1
33%
Install complete (Waiting to start)
node-2
33%
Install complete (Waiting to start)
node-3
5%
Installing Druid Broker
3 of 3 hosts showing - Show All
Show: 25
1 - 3 of 3
Next →

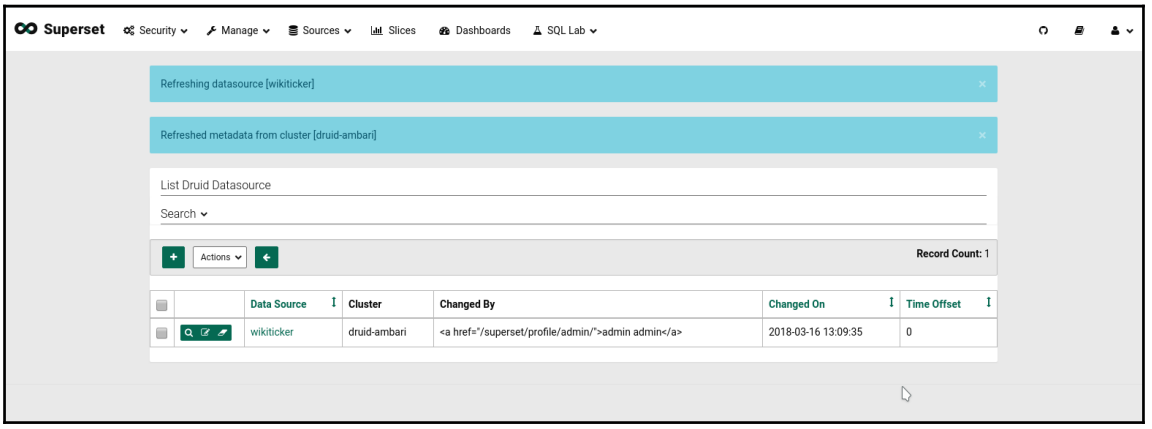

Add Service Wizard
ADD SERVICE WIZARD
Choose Services
Assign Masters
Assign Slaves and Clients
Customize Services
Configure Identities
Review
Install, Start and Test
Summary
Summary
Important: You may also need to restart other services for the newly added services to function properly (for example, HDFS and YARN/MapReduce need to be restarted after adding Oozie). After closing this wizard, please restart all services that have the restart indicator next to the service name.
Here is the summary of the install process.
The cluster consists of 3 hosts
Installed and started services successfully on 2 new hosts
1 warnings
Complete →

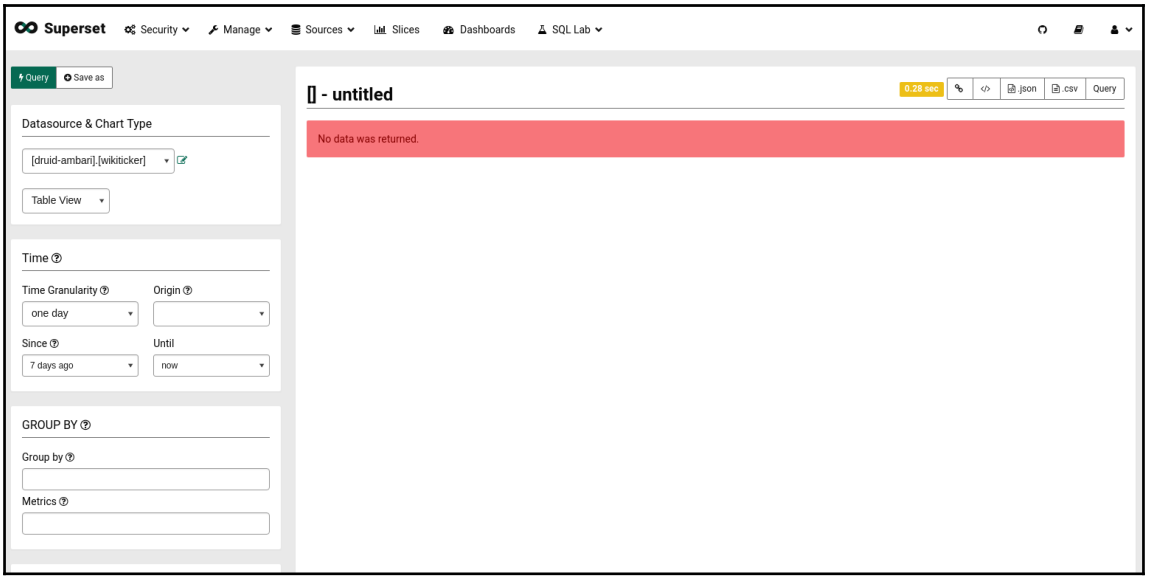

Coordinator Console
Running Tasks
Show 10 entries
id
createdTime
queueInsertionTime
id
createdTime
queueInsertionTime
index_hadoop_wikiticker_2018-03-16T04:54:38.979Z
2018-03-16T04:54:39.058Z
2018-03-16T04:54:39.082Z
Showing 1 to 1 of 1 entries

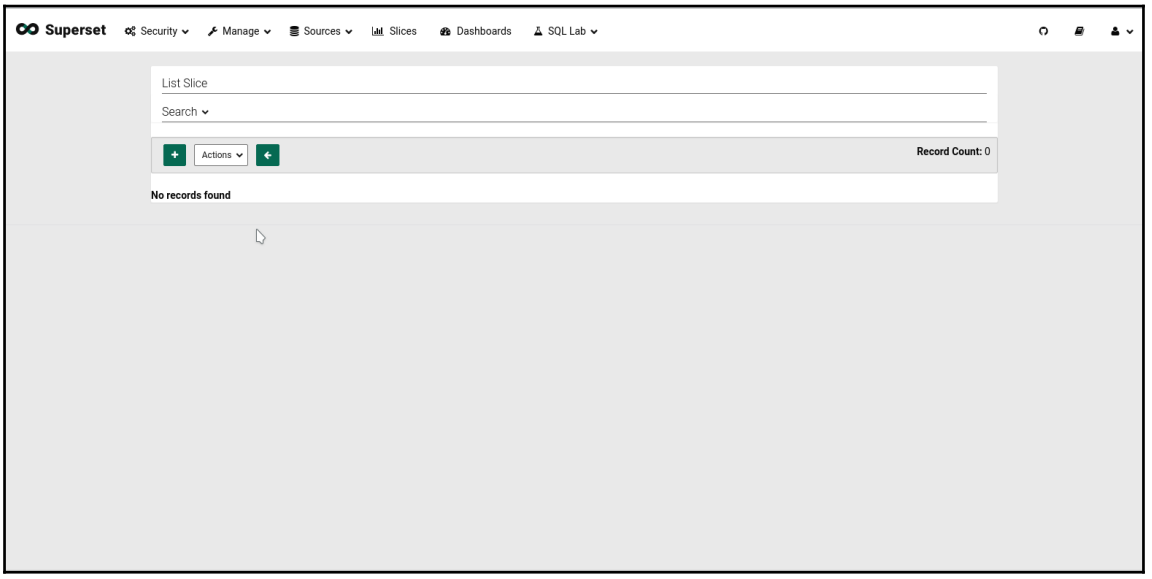

Superset
Login
Sign In
Enter your login and password below:
Username:
admin
Password:
Sign In

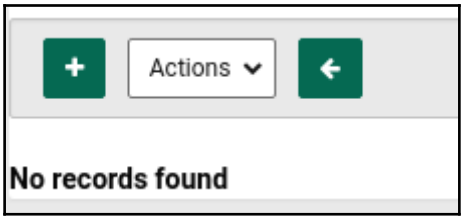

Superset
Security
Manage
Sources
Slices
Dashboards
SQL Lab
Dashboards
Dashboard
Creator
Modified
No data available in table
Previous
Next

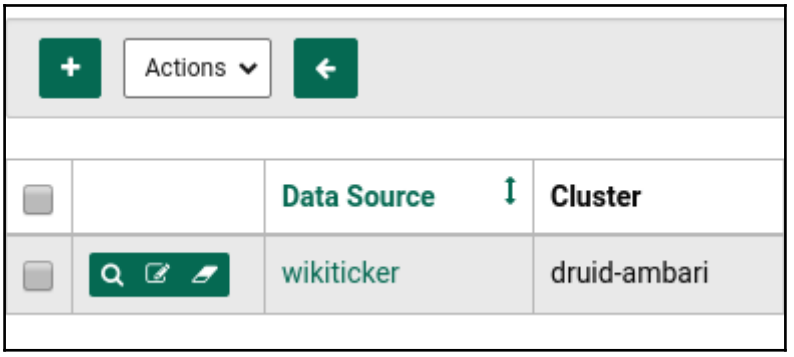

Superset
Security
Manage
Sources
Slices
Dashboards
SQL Lab
Databases
Tables
Druid Clusters
Druid Datasources
Refresh Druid Metadata
Dashboard
Creator
No data available in table

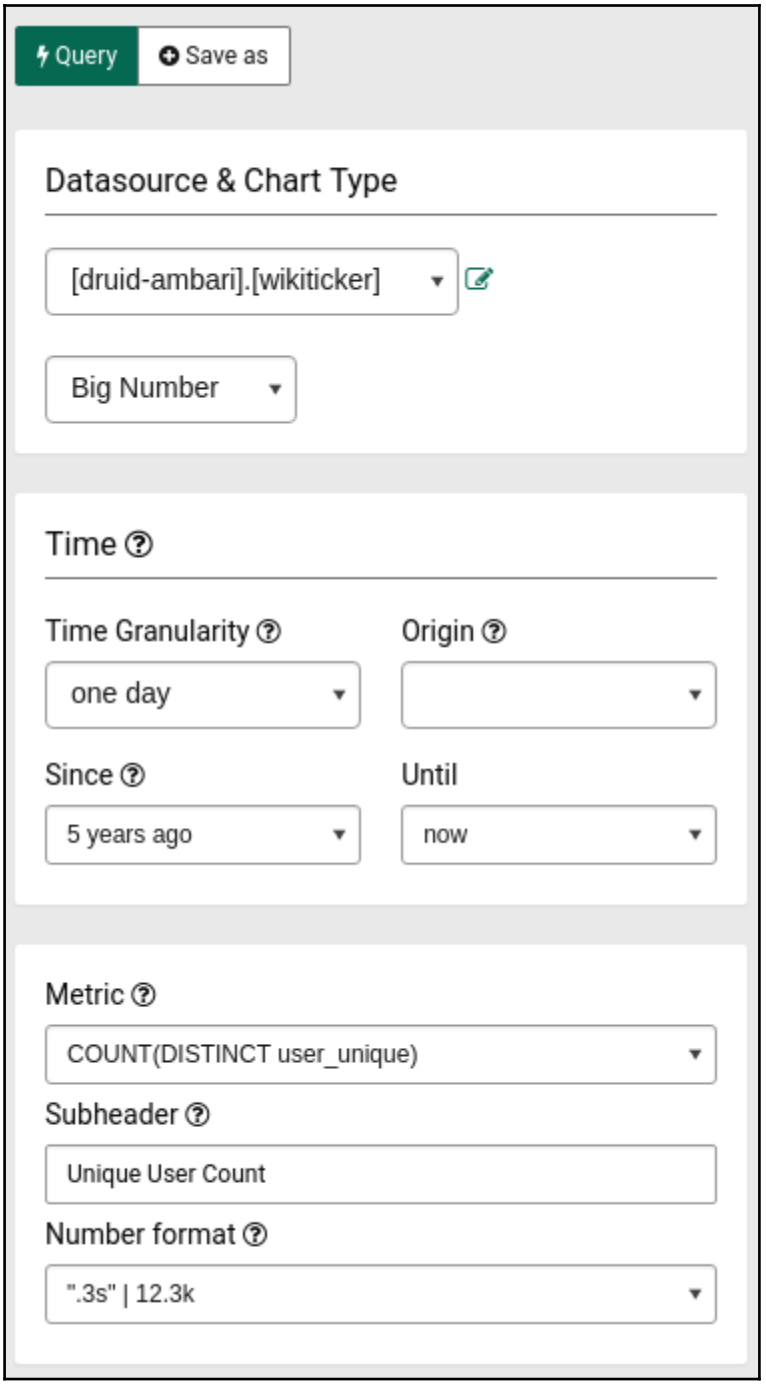

Superset
Security
Manage
Sources
Slices
Dashboards
SQL Lab
Refreshing datasource [wikiticker]
Refreshed metadata from cluster [druid-ambari]
List Druid Datasource
Search
Actions
Record Count: 1
Data Source
Cluster
Changed By
Changed On
Time Offset
wikiticker
druid-ambari
admin admin
2018-03-16 13:09:35
0

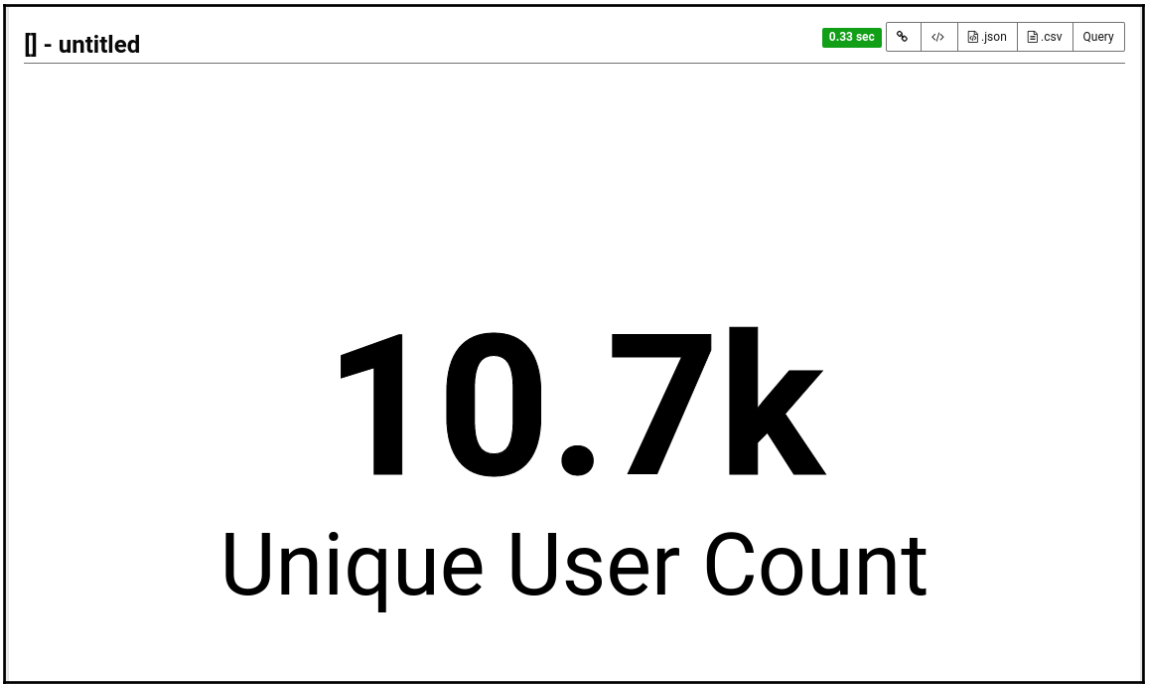

Superset
Security
Manage
Sources
Slices
Dashboards
SQL Lab
Query
Save as
Datasource & Chart Type
[druid-ambari].[wikiticker]
Table View
Time
Time Granularity
one day
Origin
Since
7 days ago
Until
now
GROUP BY
Group by
Metrics
[] - untitled
0.28 sec
.json
.csv
Query
No data was returned.

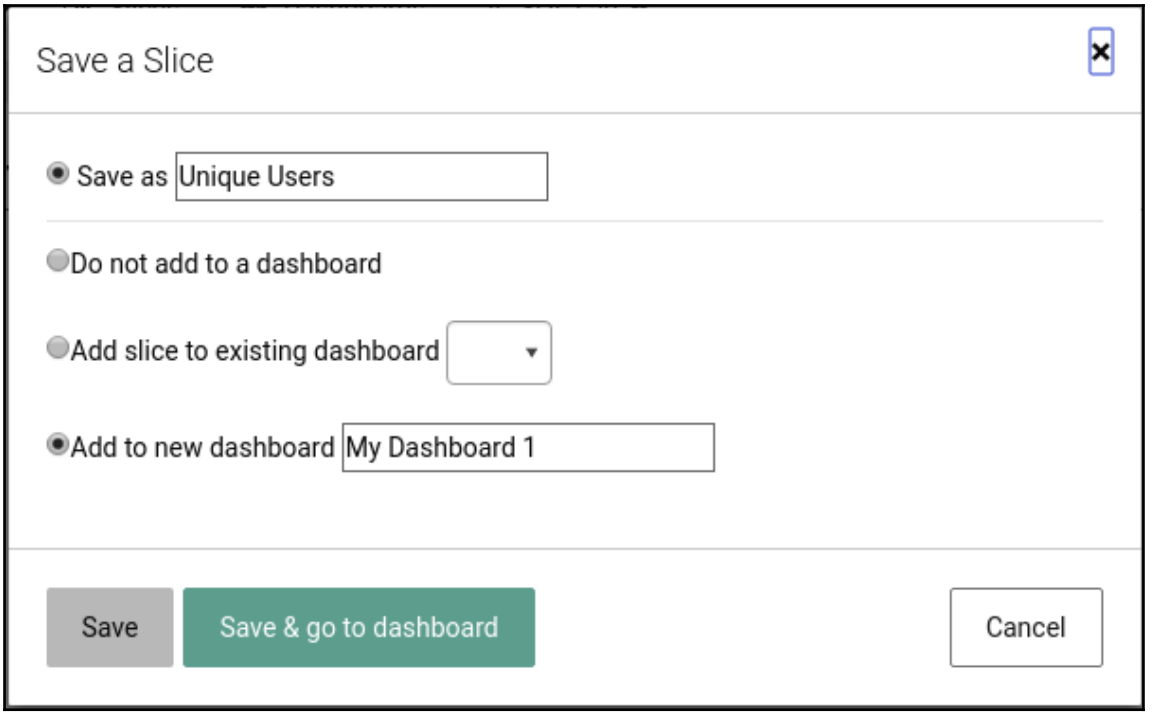

Superset
Security
Manage
Sources
Slices
Dashboards
SQL Lab
List Slice
Search
Actions
Record Count: 0
No records found

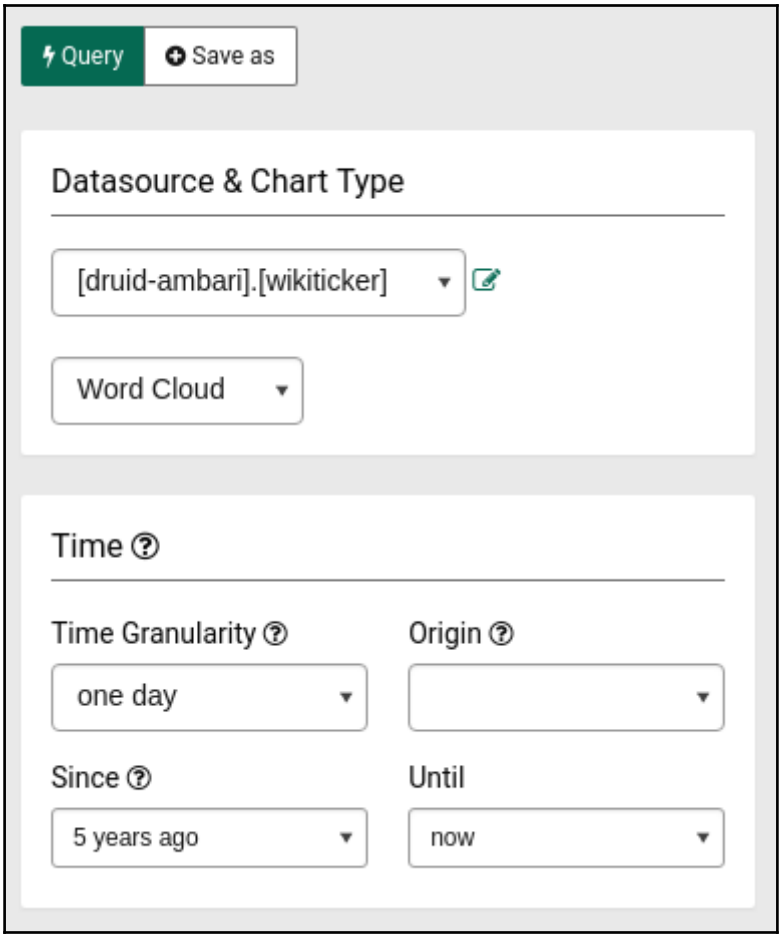

Actions
No records found


Actions
Data Source
Cluster
wikiticker
druid-ambari

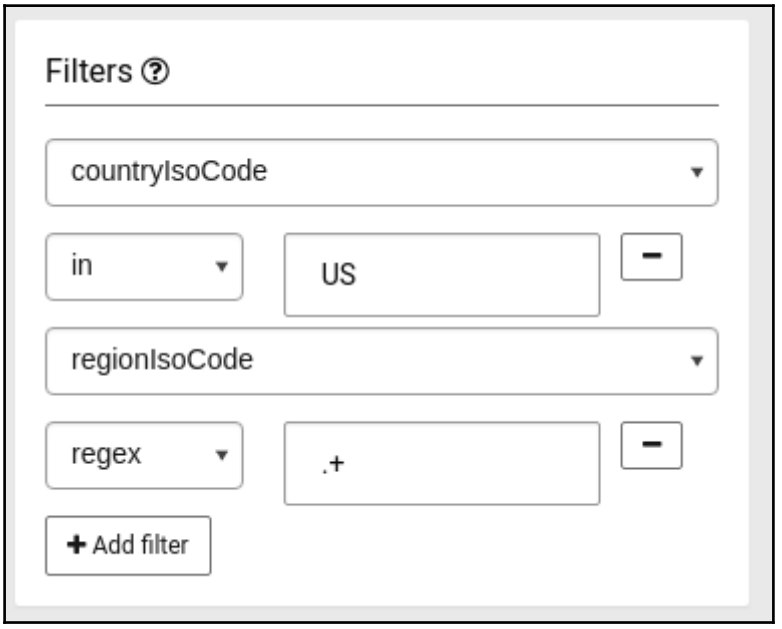

Query
Save as
Datasource & Chart Type
[druid-ambari].[wikiticker]
Big Number
Time
Time Granularity
one day
Origin
Since
5 years ago
Until
now
Metric
COUNT(DISTINCT user_unique)
Subheader
Unique User Count
Number format
".3s" | 12.3k

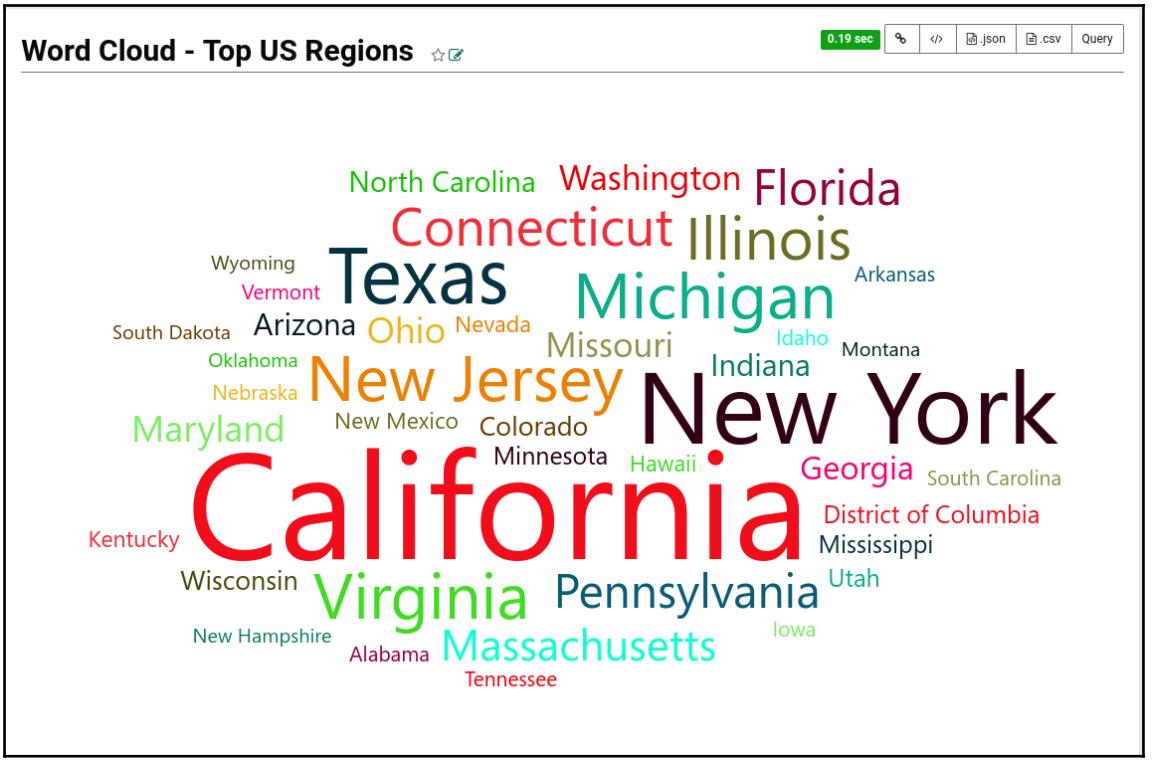

[] - untitled
0.33 sec
.json
.csv
Query
10.7k
Unique User Count

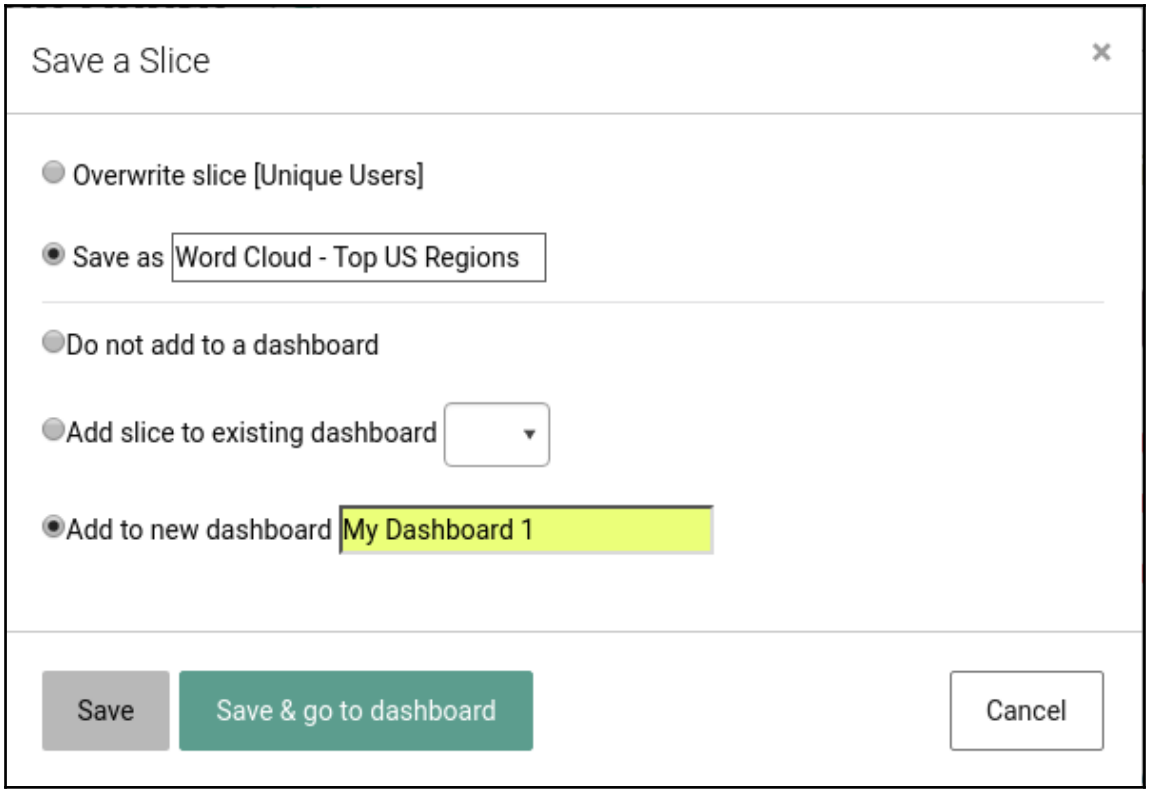

Save a Slice
Save as
Unique Users
Do not add to a dashboard
Add slice to existing dashboard
Add to new dashboard
My Dashboard 1
Save
Save & go to dashboard
Cancel

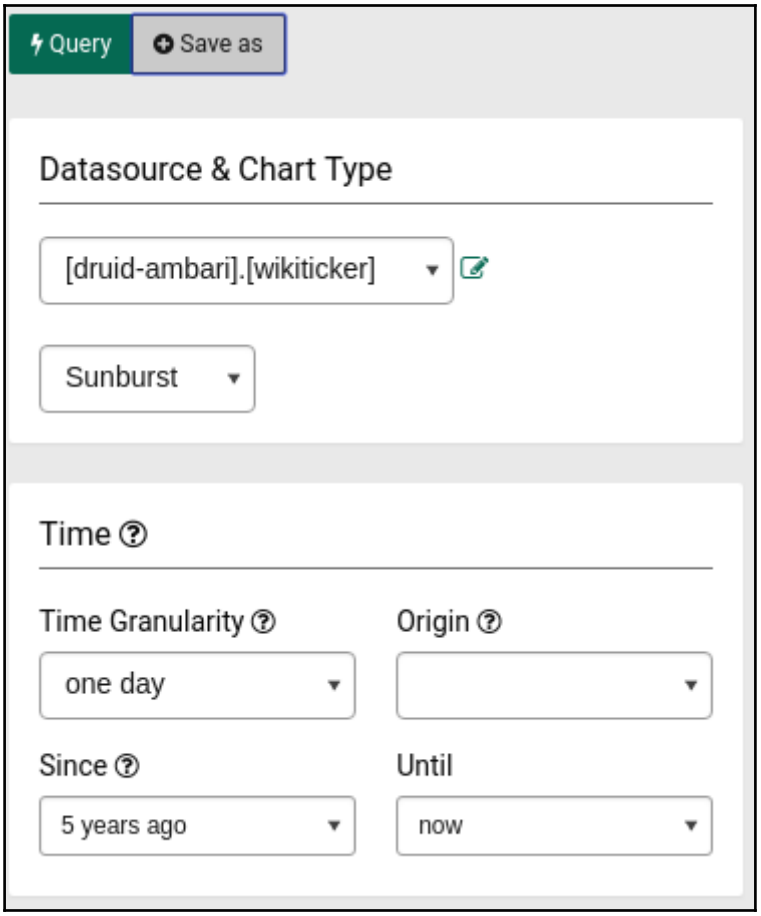

Query
Save as
Datasource & Chart Type
[druid-ambari].[wikiticker]
Word Cloud
Time
Time Granularity
one day
Origin
Since
5 years ago
Until
now

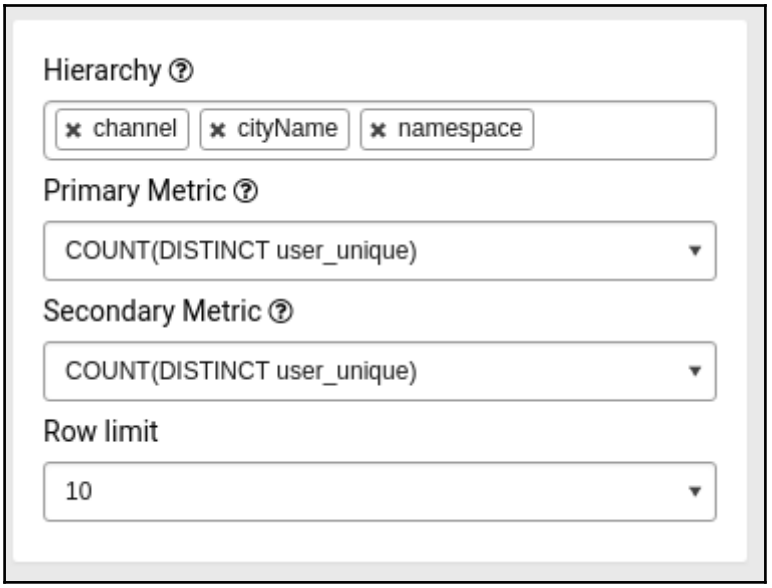

Series
regionName
Metric
COUNT(*)
Series limit
10
Font Size From
Font Size To
20
150
Rotation
flat


Filters
countryIsoCode
in
US
regionIsoCode
regex
.+
Add filter

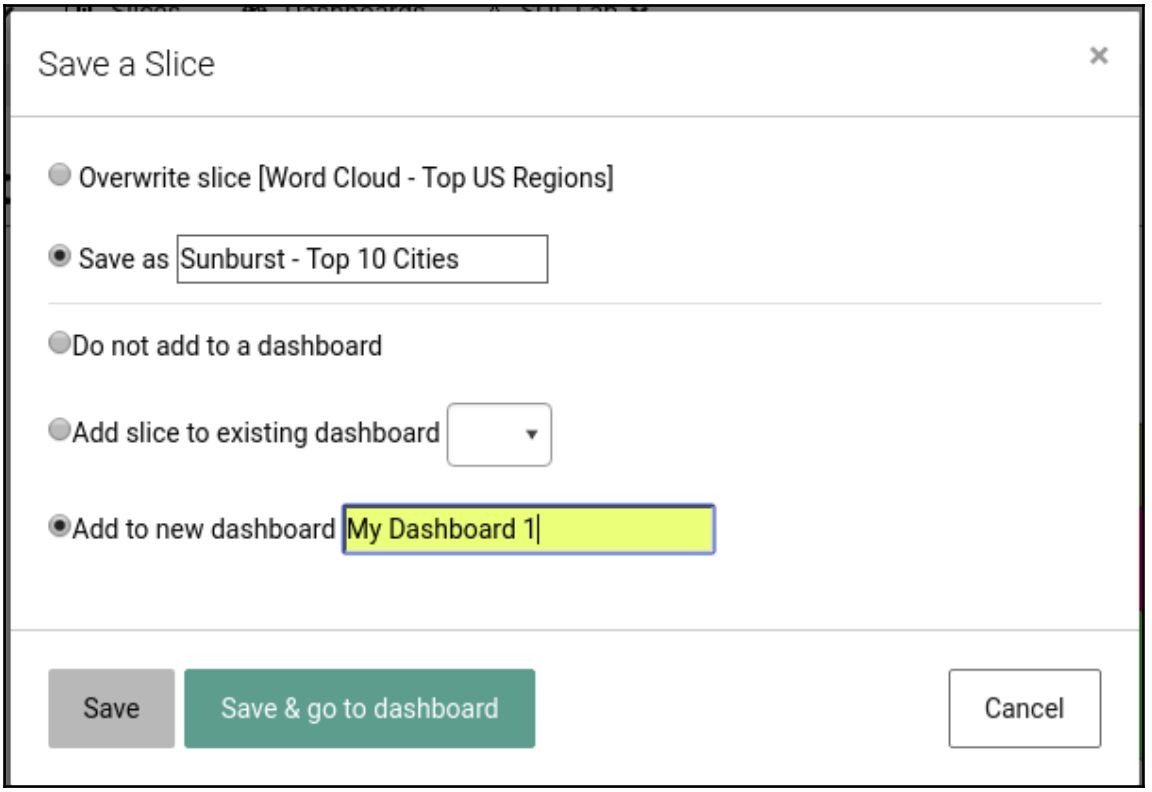

Word Cloud - Top US Regions
0.19 sec
.json
.csv
Query
North Carolina
Washington
Florida
Connecticut
Illinois
Wyoming
Texas
Arkansas
Vermont
Michigan
South Dakota
Arizona
Ohio
Nevada
Missouri
Idaho
Montana
Oklahoma
Indiana
New Jersey
New York
Nebraska
Maryland
New Mexico
Colorado
Minnesota
Hawaii
Georgia
South Carolina
California
District of Columbia
Kentucky
Mississippi
Wisconsin
Virginia
Pennsylvania
Utah
New Hampshire
Iowa
Alabama
Massachusetts
Tennessee

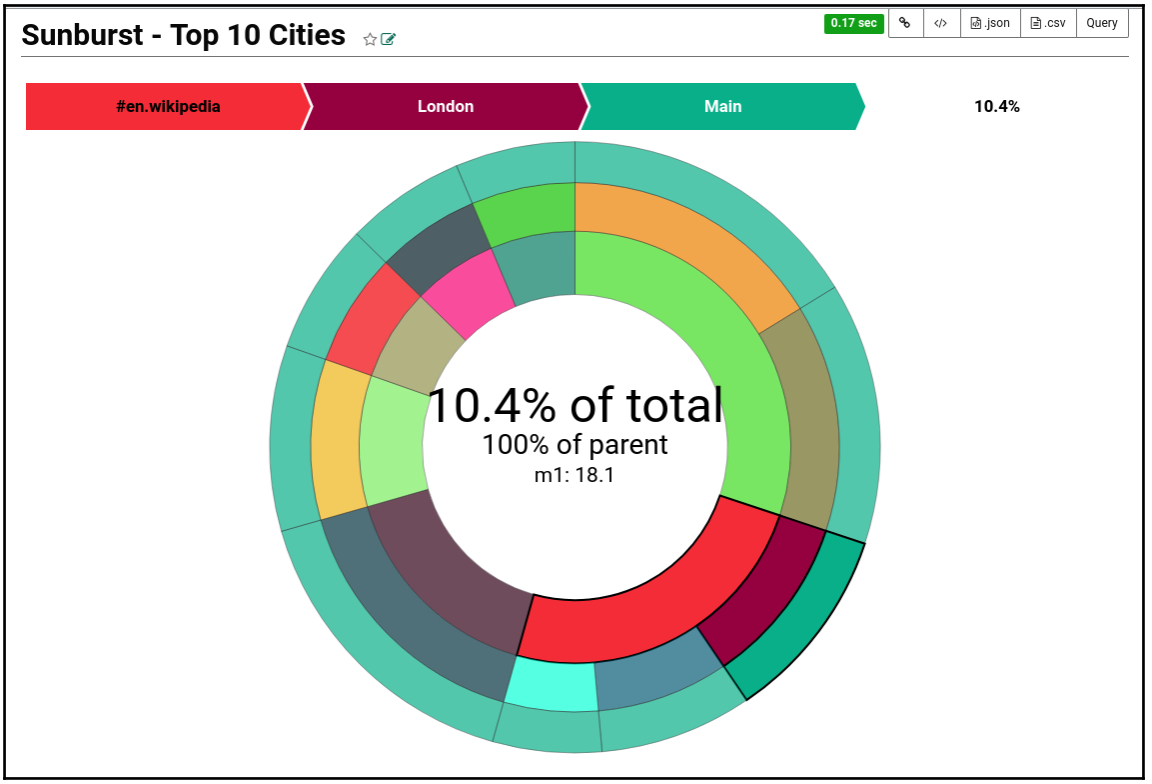

Save a Slice
Overwrite slice [Unique Users]
Save as
Word Cloud - Top US Regions
Do not add to a dashboard
Add slice to existing dashboard
Add to new dashboard
My Dashboard 1
Save
Save & go to dashboard
Cancel

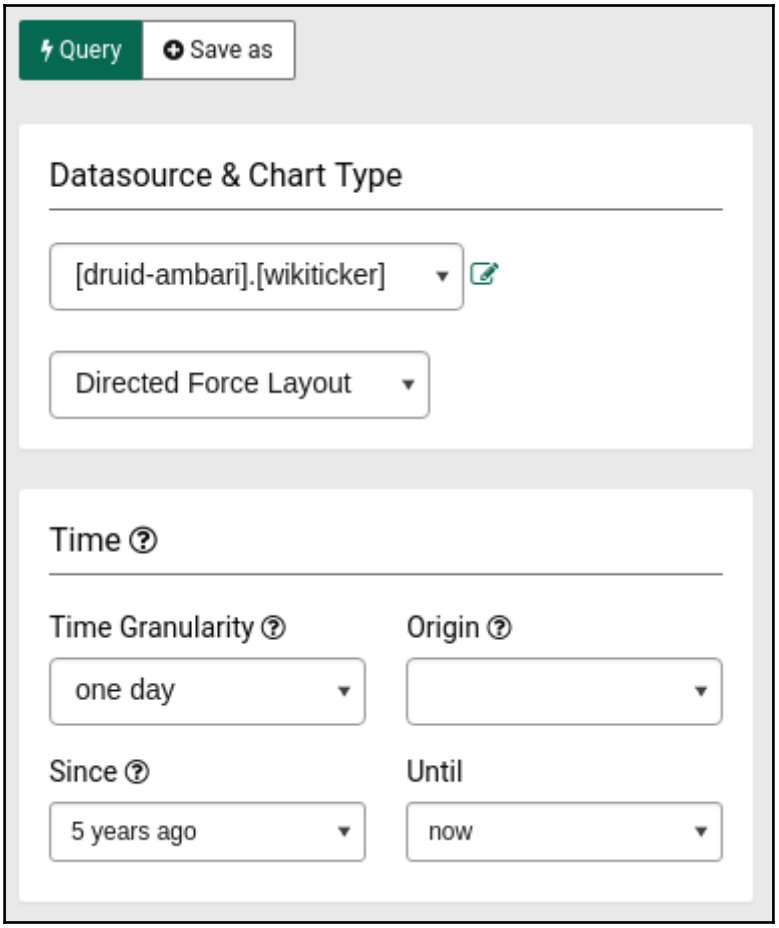

Query
Save as
Datasource & Chart Type
[druid-ambari].[wikiticker]
Sunburst
Time
Time Granularity
one day
Origin
Since
5 years ago
Until
now

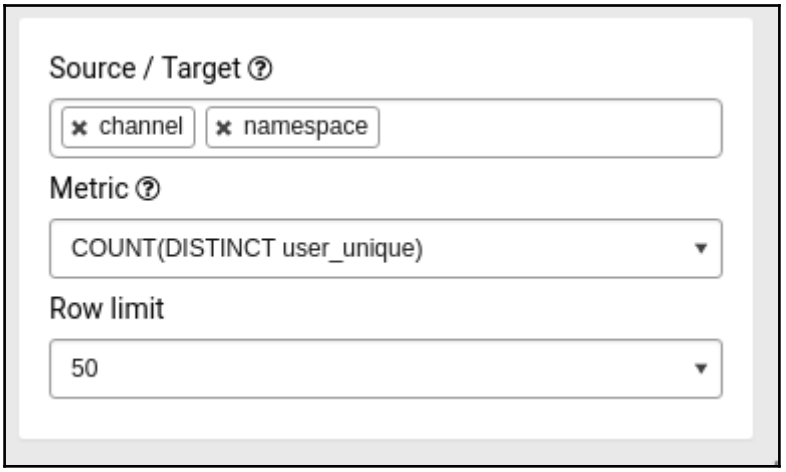

Hierarchy
channel
cityName
namespace
Primary Metric
COUNT(DISTINCT user_unique)
Secondary Metric
COUNT(DISTINCT user_unique)
Row limit
10

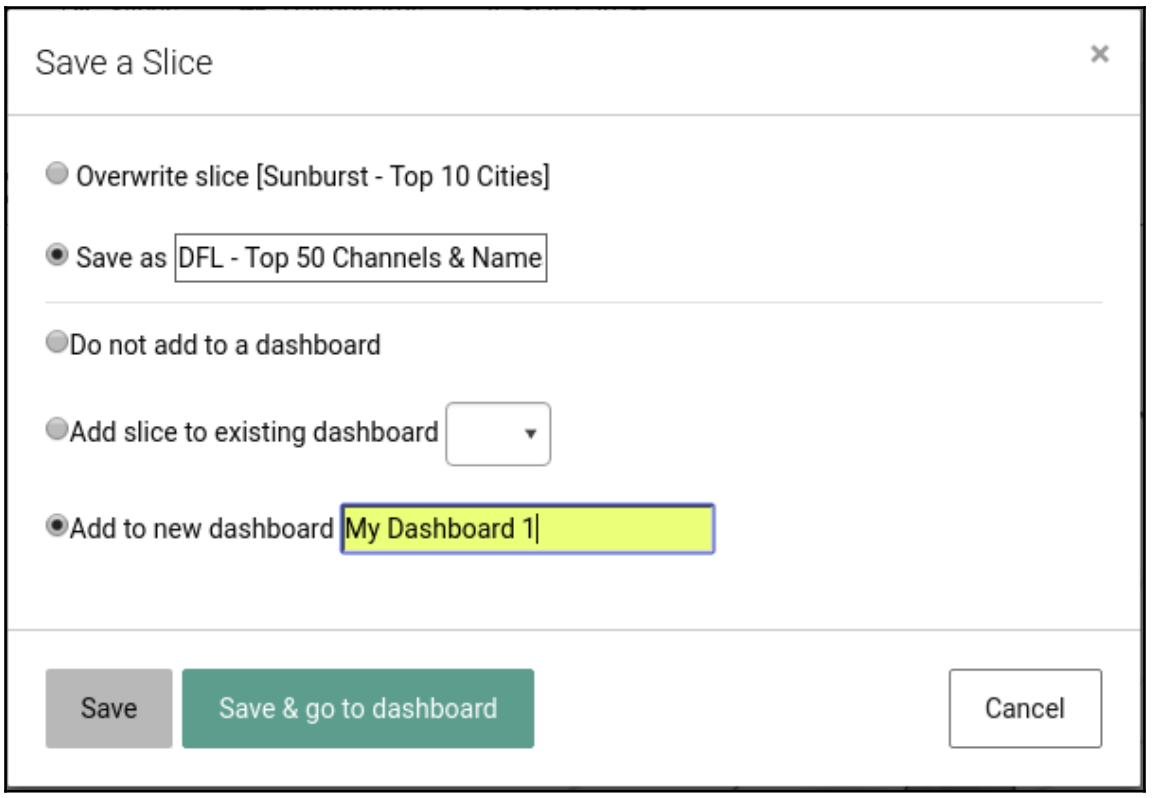

Filters
cityName
regex
.+
Add filter

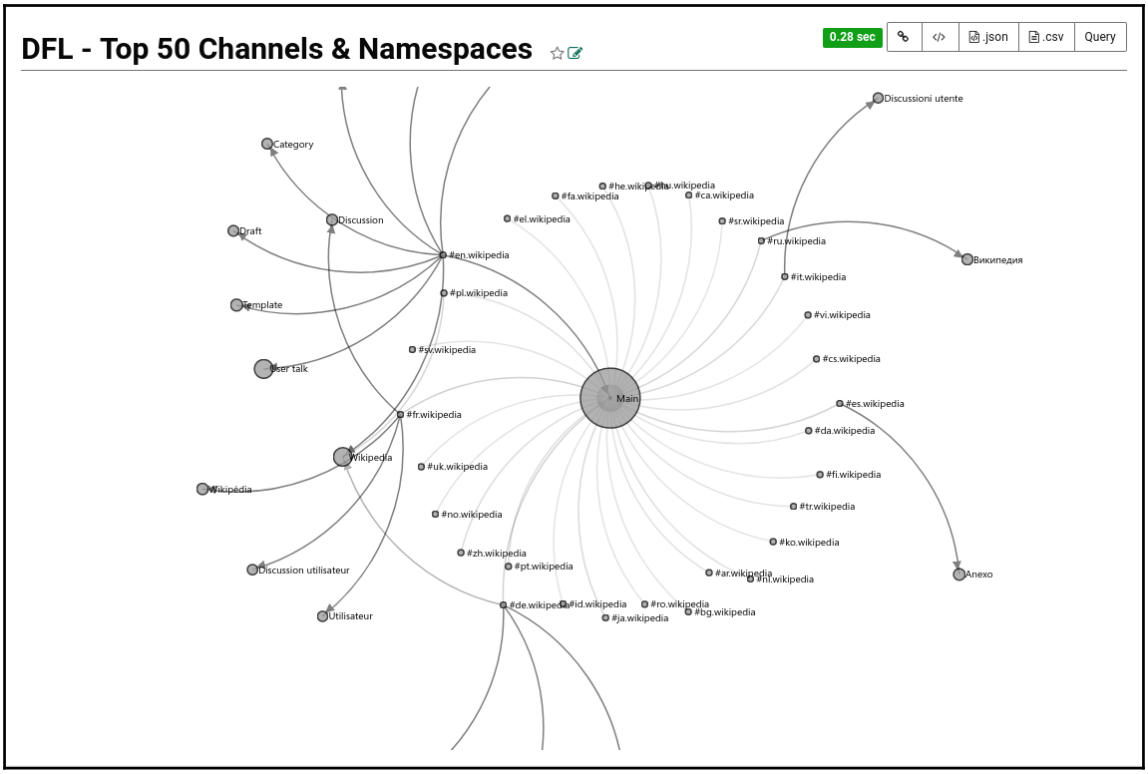

Save a Slice
Overwrite slice [Word Cloud - Top US Regions]
Save as Sunburst - Top 10 Cities
Do not add to a dashboard
Add slice to existing dashboard
Add to new dashboard My Dashboard 1
Save
Save & go to dashboard
Cancel

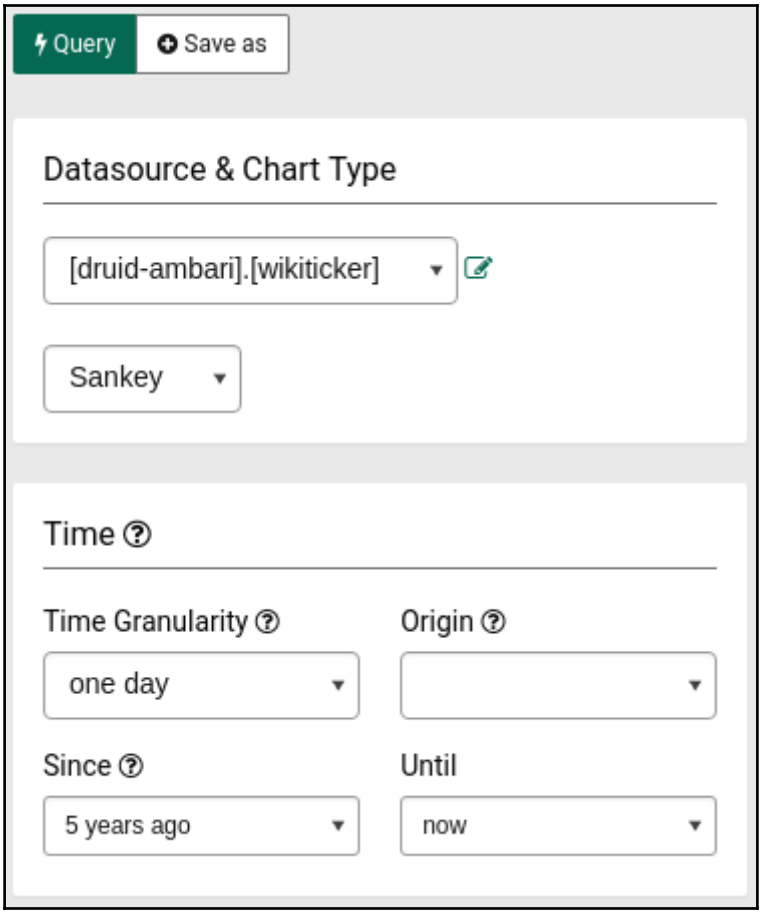

Sunburst - Top 10 Cities
0.17 sec
.json
.csv
Query
#en.wikipedia
London
Main
10.4%
10.4% of total
100% of parent
m1: 18.1

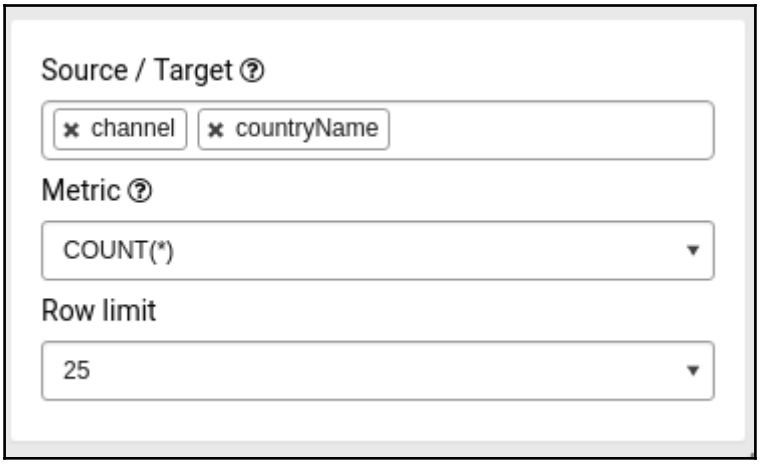

Query
Save as
Datasource & Chart Type
[druid-ambari].[wikiticker]
Directed Force Layout
Time
Time Granularity
one day
Origin
Since
5 years ago
Until
now


Source / Target
channel
namespace
Metric
COUNT(DISTINCT user_unique)
Row limit
50

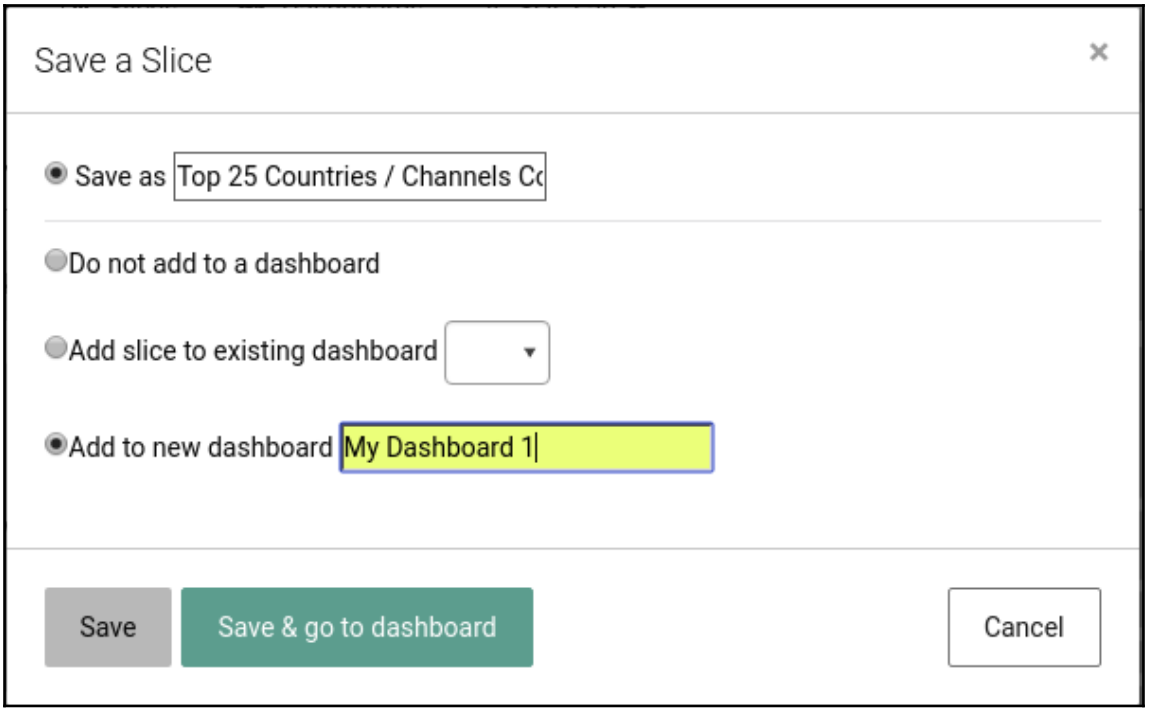

Save a Slice
Overwrite slice [Sunburst - Top 10 Cities]
Save as DFL - Top 50 Channels & Name
Do not add to a dashboard
Add slice to existing dashboard
Add to new dashboard My Dashboard 1
Save
Save & go to dashboard
Cancel

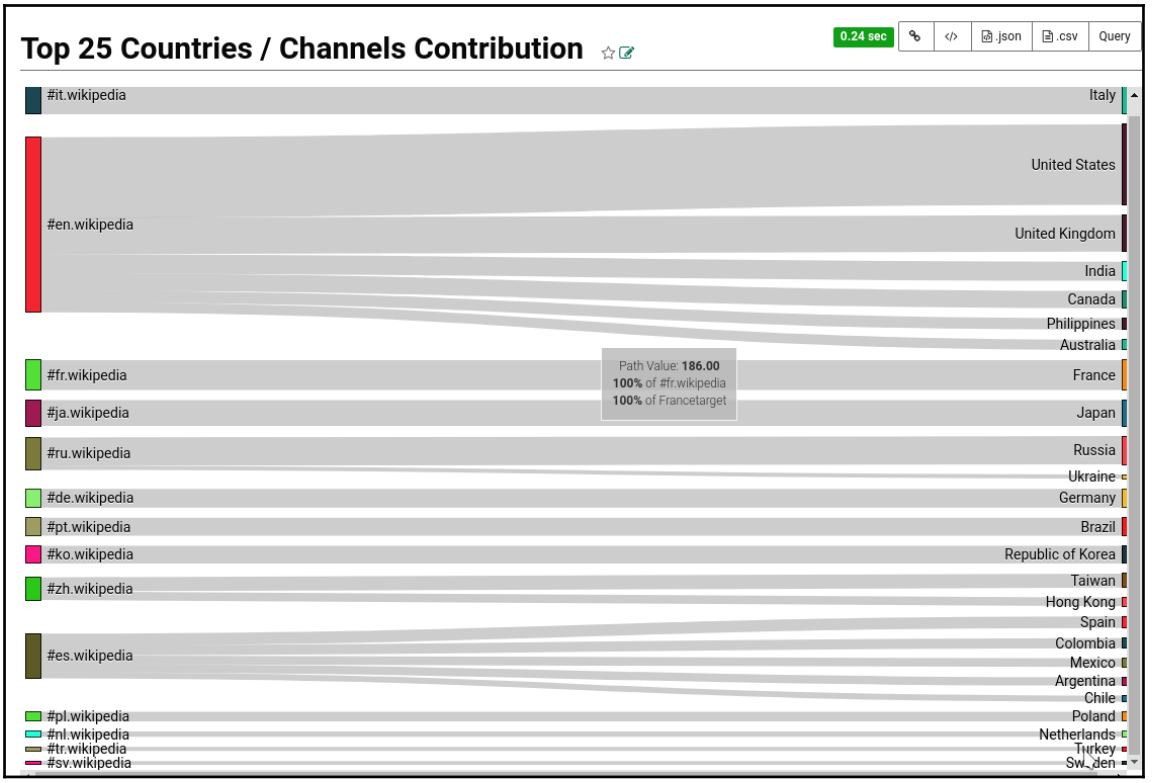

DFL - Top 50 Channels & Namespaces
0.28 sec
.json
.csv
Query
Discussioni utente
Category
Draft
Discussion
Template
User talk
Main
Wikipedia
Wikipédia
Discussion utilisateur
Utilisateur
Anexo
Википедия
#en.wikipedia
#pl.wikipedia
#sv.wikipedia
#fr.wikipedia
#uk.wikipedia
#no.wikipedia
#zh.wikipedia
#pt.wikipedia
#de.wikipedia
#id.wikipedia
#ro.wikipedia
#ja.wikipedia
#bg.wikipedia
#ar.wikipedia
#ko.wikipedia
#tr.wikipedia
#fi.wikipedia
#da.wikipedia
#es.wikipedia
#cs.wikipedia
#vi.wikipedia
#it.wikipedia
#ru.wikipedia
#sr.wikipedia
#ca.wikipedia
#fa.wikipedia
#el.wikipedia

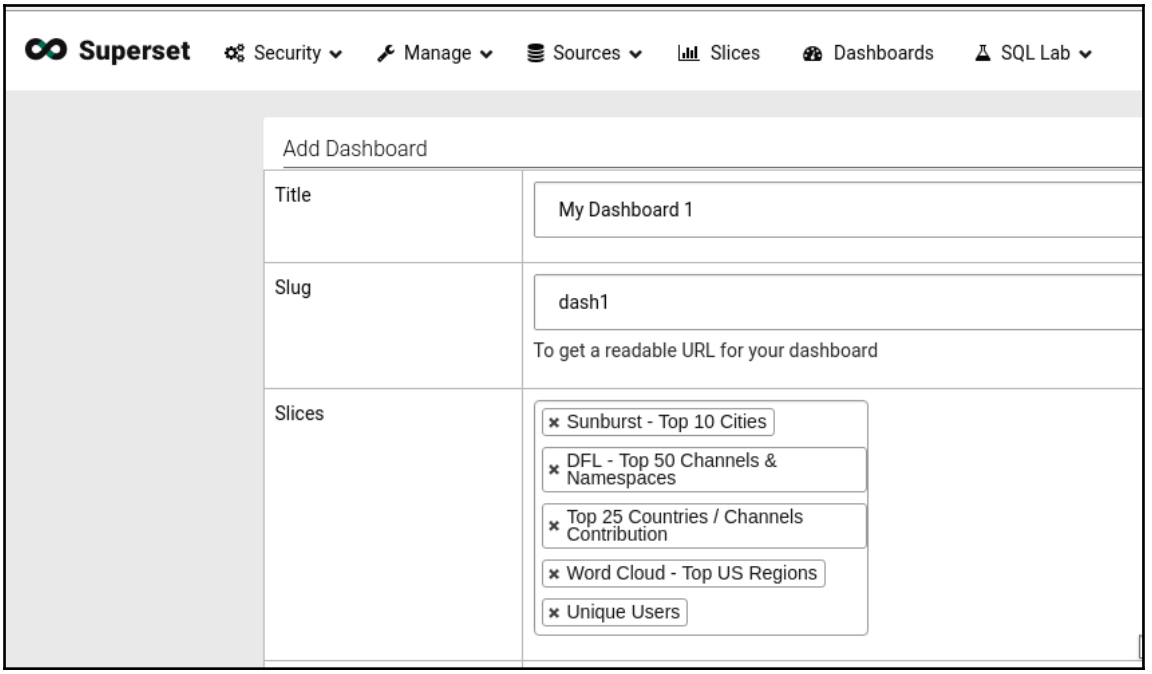

Query
Save as
Datasource & Chart Type
[druid-ambari].[wikiticker]
Sankey
Time
Time Granularity
one day
Origin
Since
5 years ago
Until
now

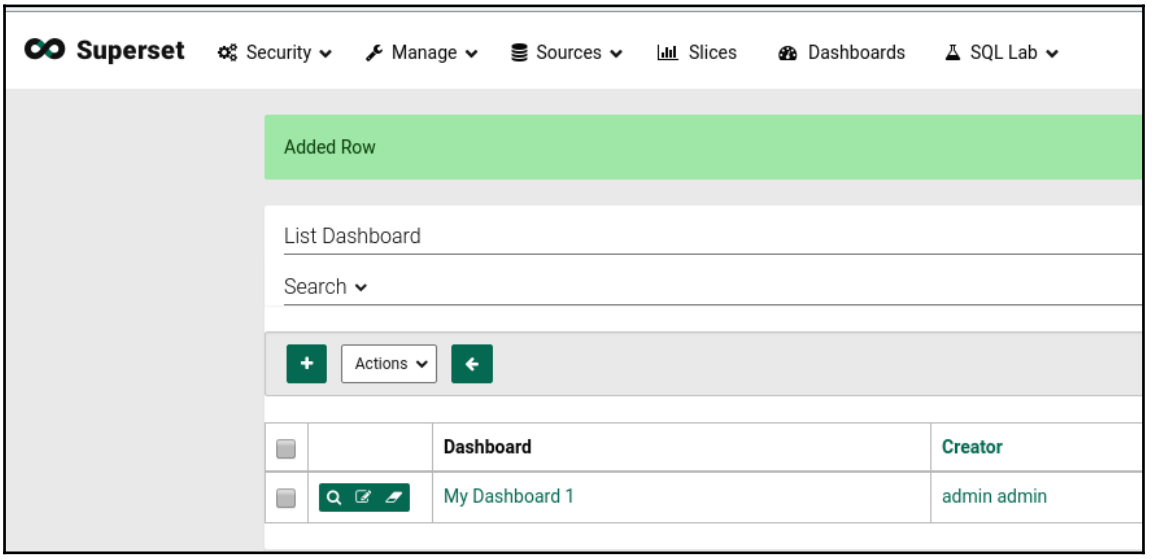

Source / Target
channel
countryName
Metric
COUNT(*)
Row limit
25

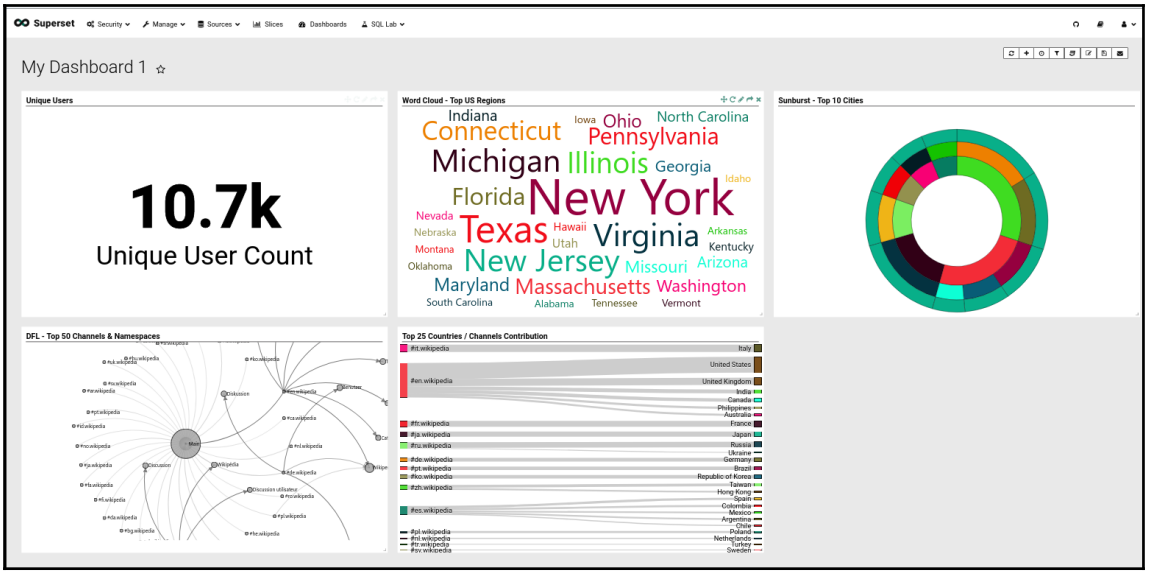

Filters
countryName
regex
.+
Add filter

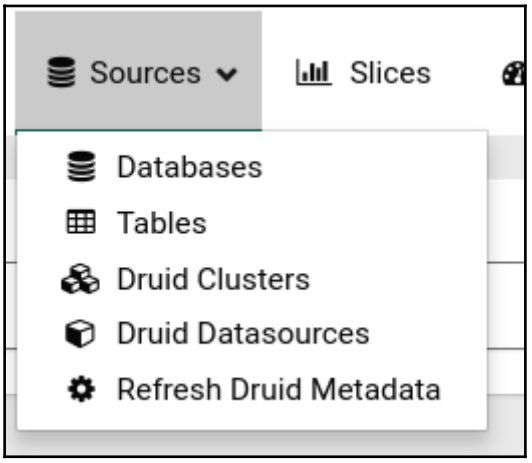

Save a Slice
Save as
Top 25 Countries / Channels Cc
Do not add to a dashboard
Add slice to existing dashboard
Add to new dashboard
My Dashboard 1
Save
Save & go to dashboard
Cancel

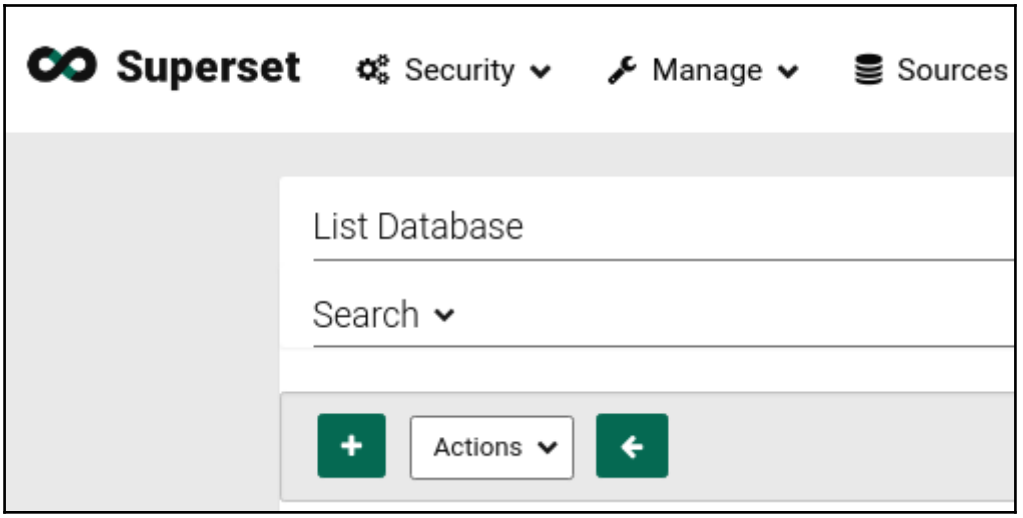

Top 25 Countries / Channels Contribution
0.24 sec
</>
.json
.csv
Query
#it.wikipedia
#en.wikipedia
#fr.wikipedia
#ja.wikipedia
#ru.wikipedia
#de.wikipedia
#pt.wikipedia
#ko.wikipedia
#zh.wikipedia
#es.wikipedia
#pl.wikipedia
#nl.wikipedia
#tr.wikipedia
#sv.wikipedia
Italy
United States
United Kingdom
India
Canada
Philippines
Australia
France
Japan
Russia
Ukraine
Germany
Brazil
Republic of Korea
Taiwan
Hong Kong
Spain
Colombia
Mexico
Argentina
Chile
Poland
Netherlands
Turkey
Sweden
Path Value: 186.00
100% of #fr.wikipedia
100% of Francetarget

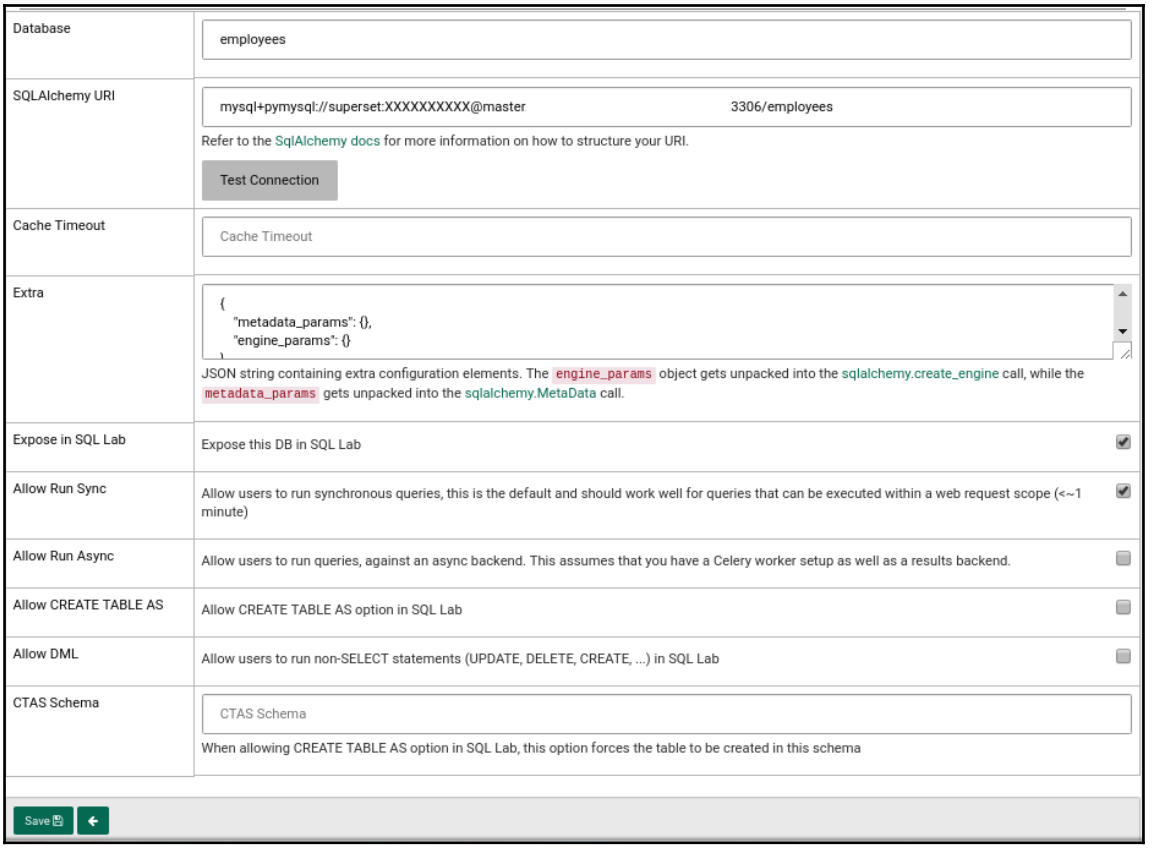

Superset
Security
Manage
Sources
Slices
Dashboards
SQL Lab
Add Dashboard
Title
My Dashboard 1
Slug
dash1
To get a readable URL for your dashboard
Slices
Sunburst - Top 10 Cities
DFL - Top 50 Channels & Namespaces
Top 25 Countries / Channels Contribution
Word Cloud - Top US Regions
Unique Users

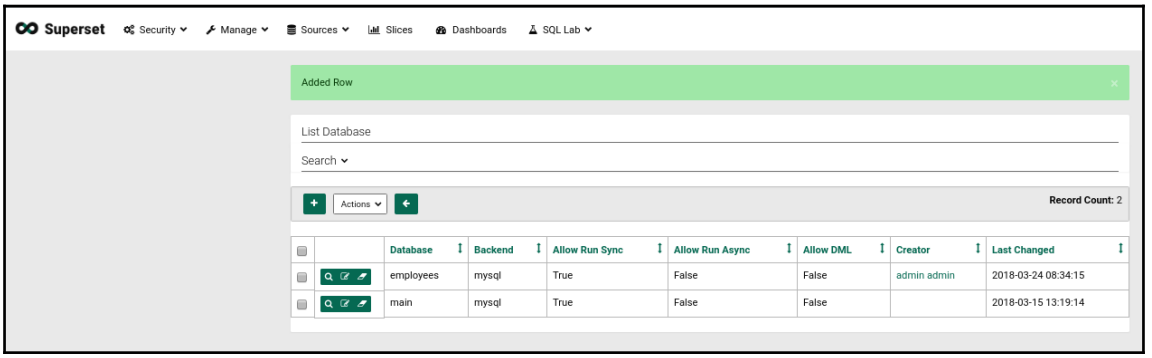

Superset
Security
Manage
Sources
Slices
Dashboards
SQL Lab
Added Row
List Dashboard
Search
Actions
Dashboard
Creator
My Dashboard 1
admin admin

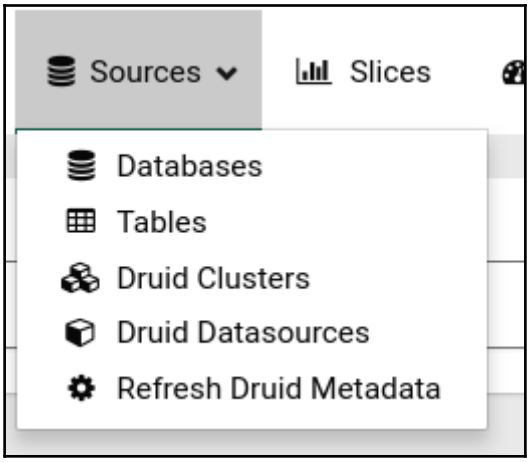

Superset
Security
Manage
Sources
Slices
Dashboards
SQL Lab
My Dashboard 1
Unique Users
10.7k
Unique User Count
Word Cloud - Top US Regions
Indiana
Iowa
Ohio
North Carolina
Connecticut
Pennsylvania
Michigan
Illinois
Georgia
Idaho
Florida
New York
Nevada
Nebraska
Texas
Hawaii
Utah
Virginia
Arkansas
Kentucky
Montana
Oklahoma
New Jersey
Missouri
Arizona
Maryland
Massachusetts
Washington
South Carolina
Alabama
Tennessee
Vermont
Sunburst - Top 10 Cities
DFL - Top 50 Channels & Namespaces
Top 25 Countries / Channels Contribution
#it.wikipedia
Italy
United States
#en.wikipedia
United Kingdom
India
Canada
Philippines
Australia
#fr.wikipedia
France
#ja.wikipedia
Japan
#ru.wikipedia
Russia
Ukraine
#de.wikipedia
Germany
#pt.wikipedia
Brazil
#ko.wikipedia
Republic of Korea
Taiwan
#zh.wikipedia
Hong Kong
Spain
#es.wikipedia
Colombia
Mexico
Argentina
Chile
#pl.wikipedia
Poland
#nl.wikipedia
Netherlands
#tr.wikipedia
Turkey
#sv.wikipedia
Sweden

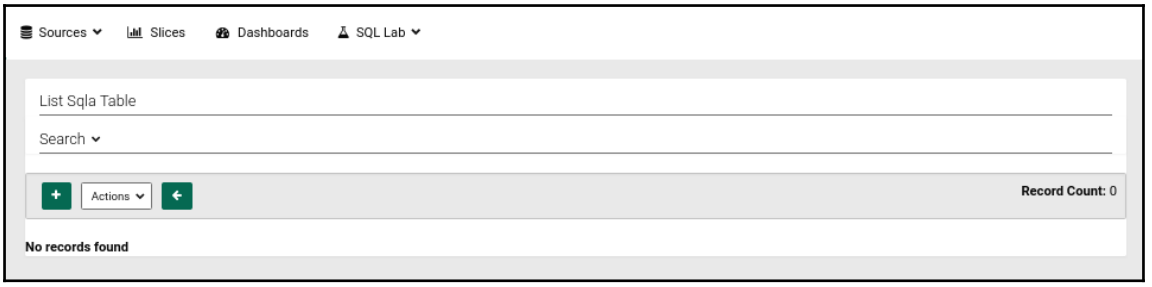

Sources
Slices
Databases
Tables
Druid Clusters
Druid Datasources
Refresh Druid Metadata

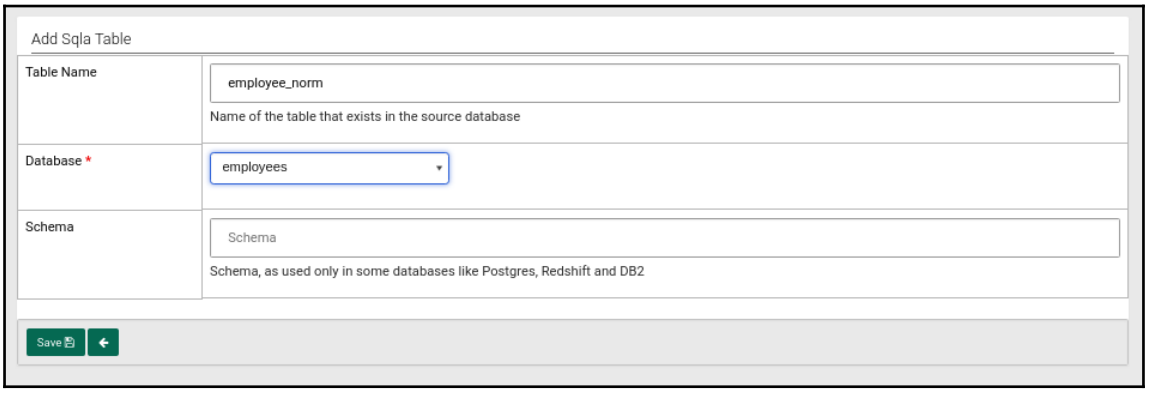

Superset
Security
Manage
Sources
List Database
Search
Actions

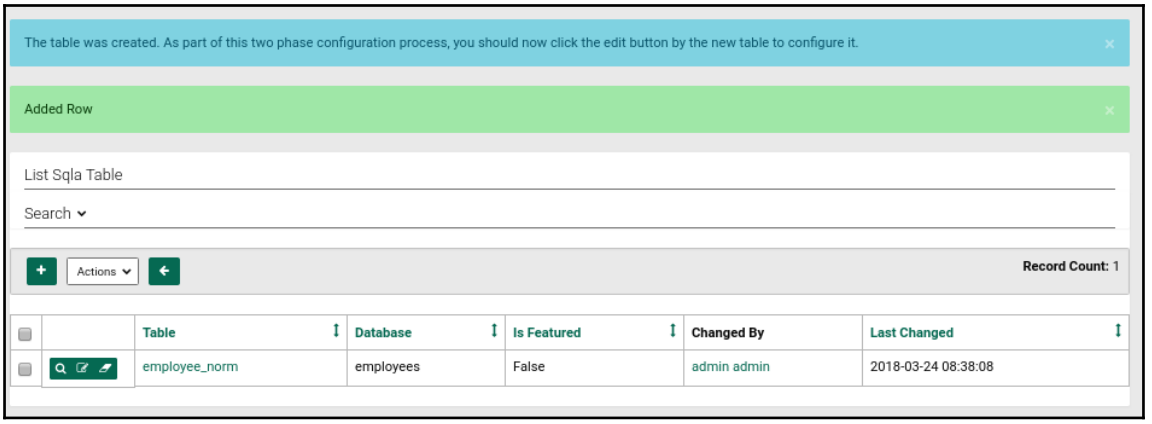

Database
employees
SQLAlchemy URI
mysql+pymysql://superset:XXXXXXXXXX@master 3306/employees
Refer to the SqlAlchemy docs for more information on how to structure your URI.
Test Connection
Cache Timeout
Cache Timeout
Extra
{
"metadata_params": {},
"engine_params": {}
}
JSON string containing extra configuration elements. The engine_params object gets unpacked into the sqlalchemy.create_engine call, while the metadata_params gets unpacked into the sqlalchemy.MetaData call.
Expose in SQL Lab
Expose this DB in SQL Lab
Allow Run Sync
Allow users to run synchronous queries, this is the default and should work well for queries that can be executed within a web request scope (<~1 minute)
Allow Run Async
Allow users to run queries, against an async backend. This assumes that you have a Celery worker setup as well as a results backend.
Allow CREATE TABLE AS
Allow CREATE TABLE AS option in SQL Lab
Allow DML
Allow users to run non-SELECT statements (UPDATE, DELETE, CREATE, ...) in SQL Lab
CTAS Schema
CTAS Schema
When allowing CREATE TABLE AS option in SQL Lab, this option forces the table to be created in this schema
Save

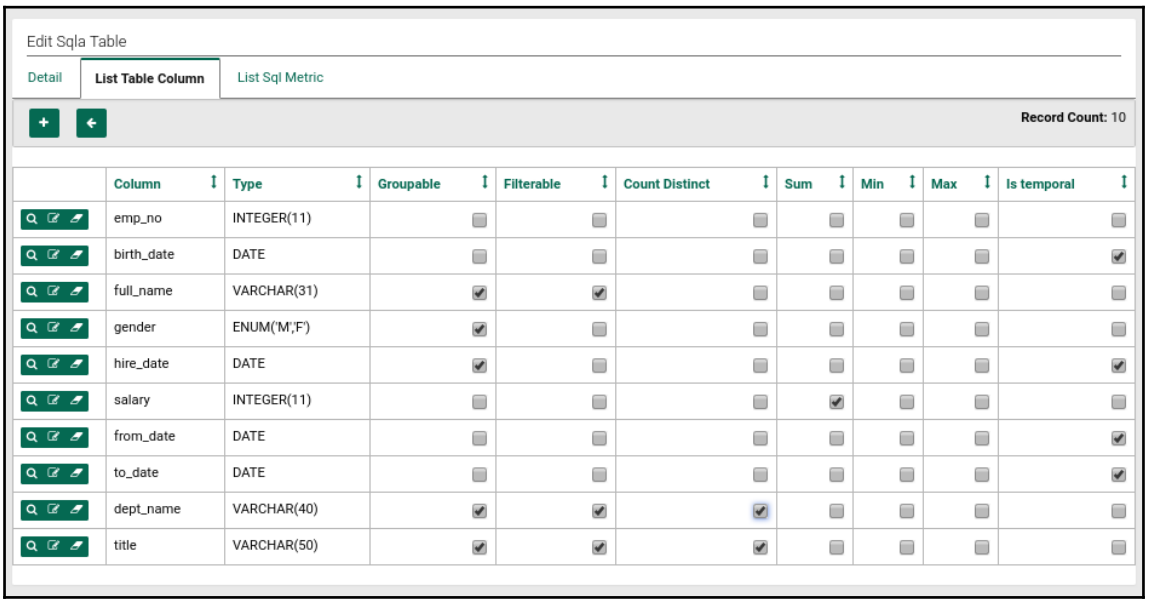

Superset
Security
Manage
Sources
Slices
Dashboards
SQL Lab
Added Row
List Database
Search
Actions
Record Count: 2
Database
Backend
Allow Run Sync
Allow Run Async
Allow DML
Creator
Last Changed
employees
mysql
True
False
False
admin admin
2018-03-24 08:34:15
main
mysql
True
False
False
2018-03-15 13:19:14

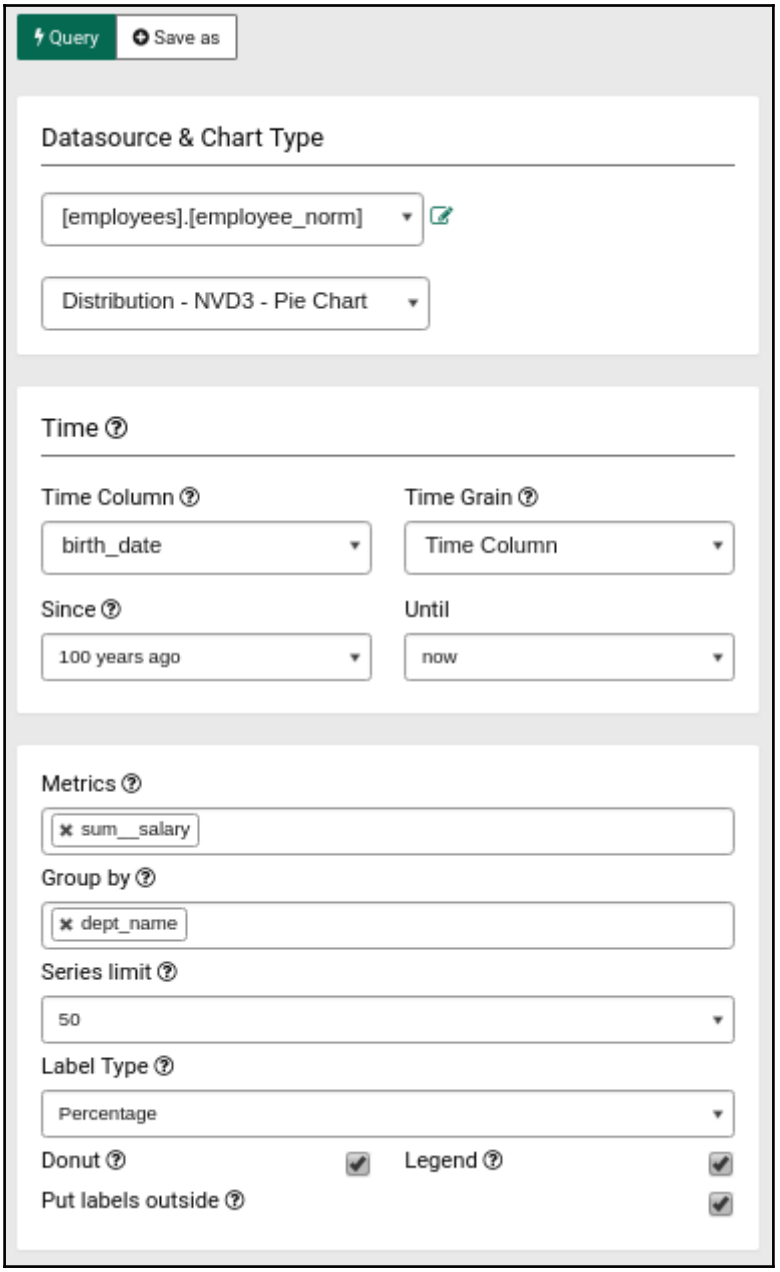

Sources
Slices
Databases
Tables
Druid Clusters
Druid Datasources
Refresh Druid Metadata

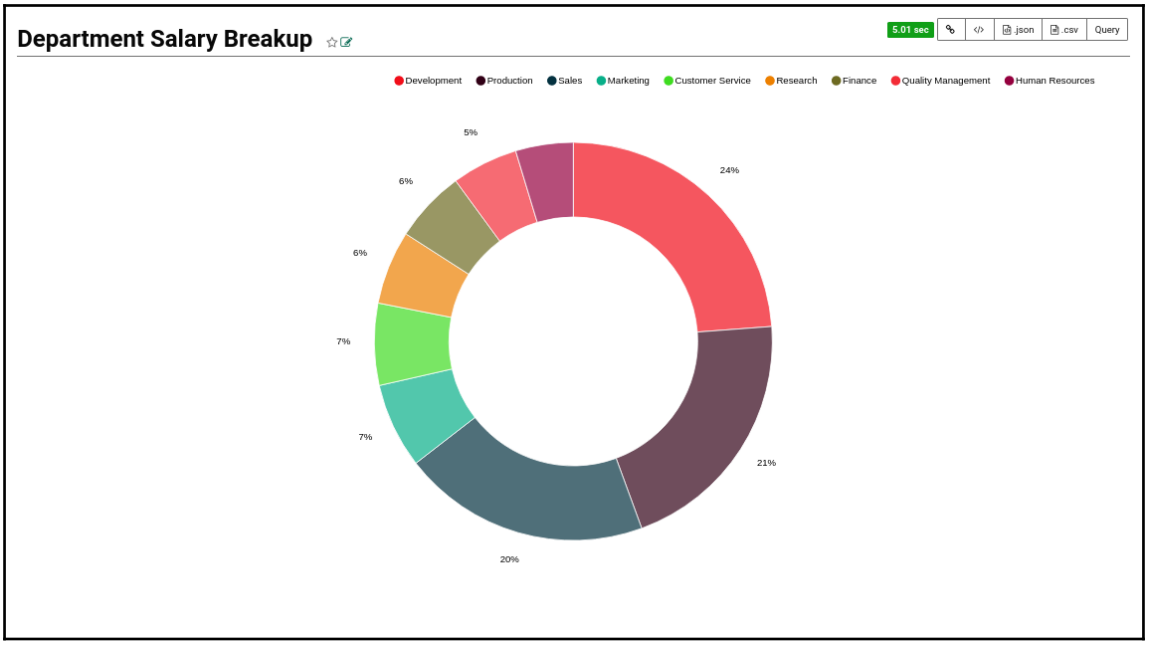

Sources
Slices
Dashboards
SQL Lab
List Sqla Table
Search
Actions
Record Count: 0
No records found

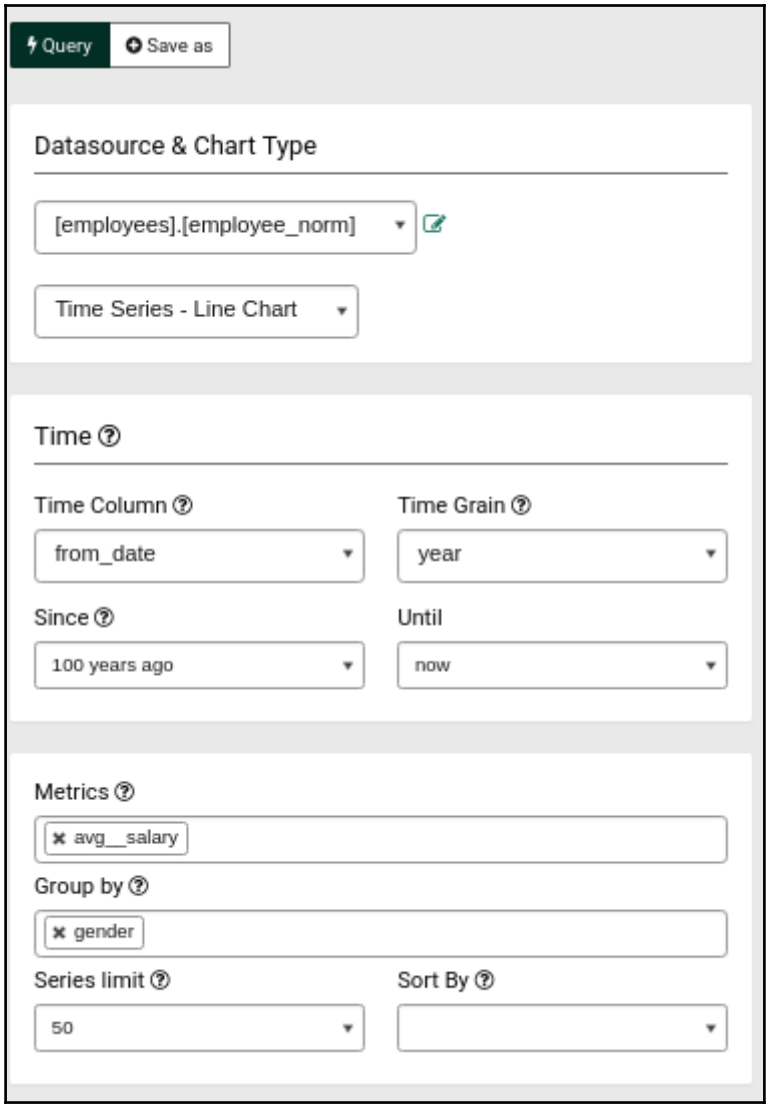

Add Sqla Table
Table Name
employee_norm
Name of the table that exists in the source database
Database *
employees
Schema
Schema
Schema, as used only in some databases like Postgres, Redshift and DB2
Save

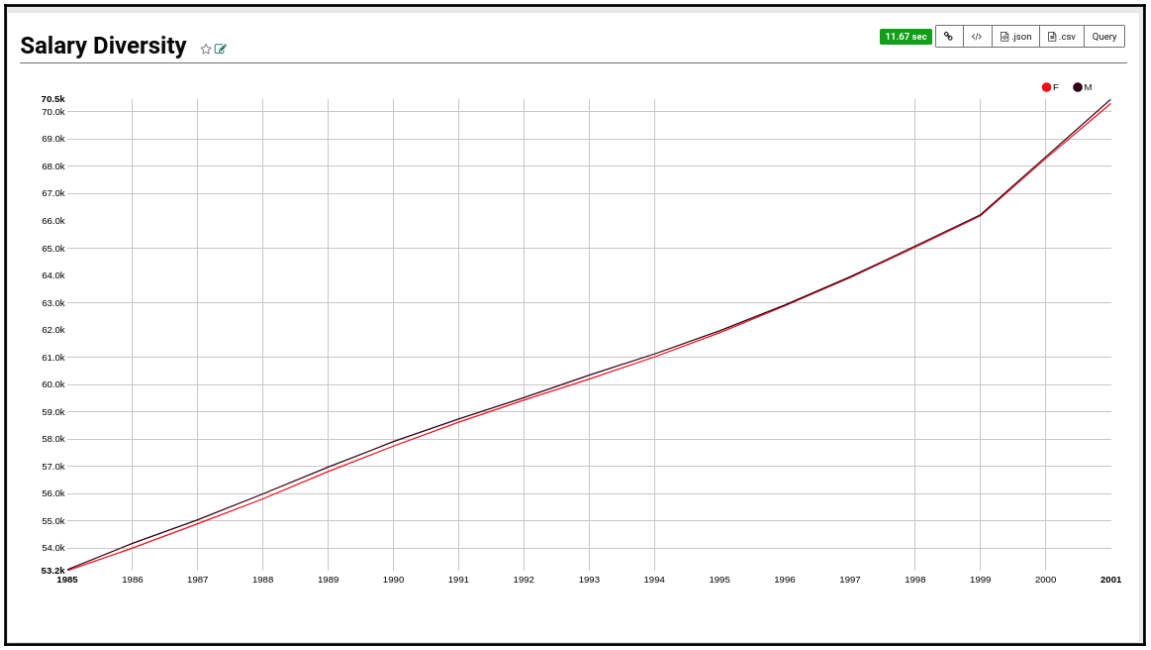

The table was created. As part of this two phase configuration process, you should now click the edit button by the new table to configure it.
Added Row
List Sqla Table
Search
Actions
Record Count: 1
Table
Database
Is Featured
Changed By
Last Changed
employee_norm
employees
False
admin admin
2018-03-24 08:38:08

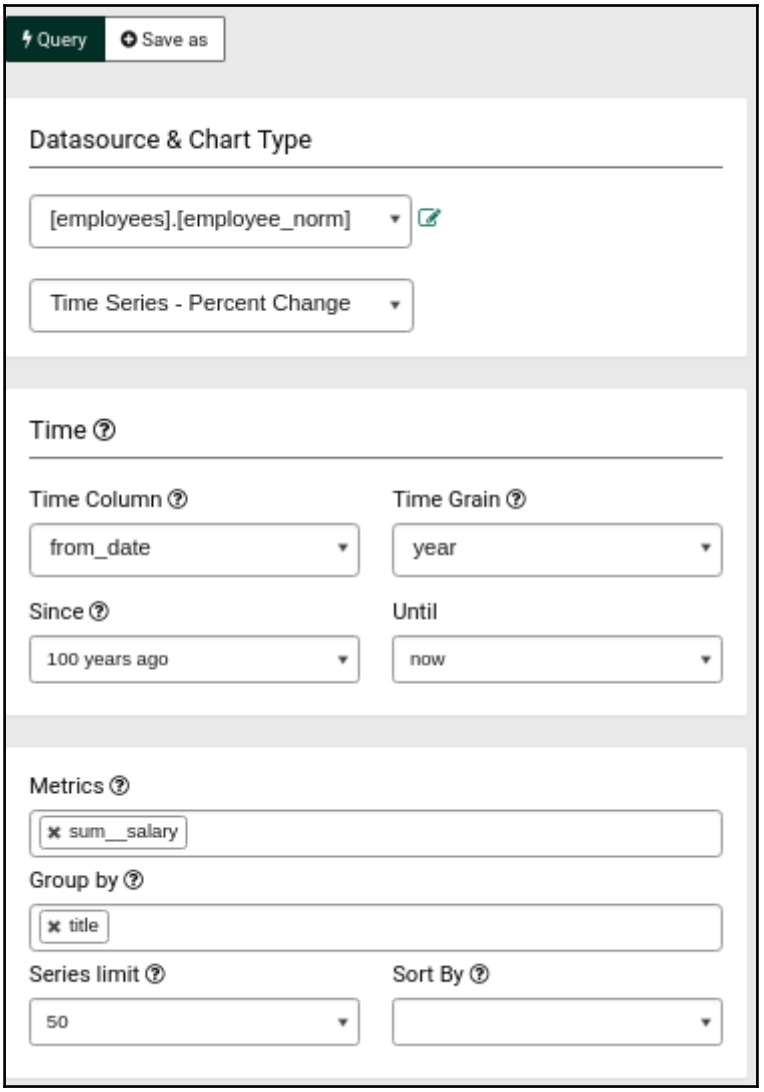

Edit Sqla Table
Detail
List Table Column
List Sql Metric
Record Count: 10
Column
Type
Groupable
Filterable
Count Distinct
Sum
Min
Max
Is temporal
emp_no INTEGER(11)
birth_date DATE
full_name VARCHAR(31)
gender ENUM('M','F')
hire_date DATE
salary INTEGER(11)
from_date DATE
to_date DATE
dept_name VARCHAR(40)
title VARCHAR(50)

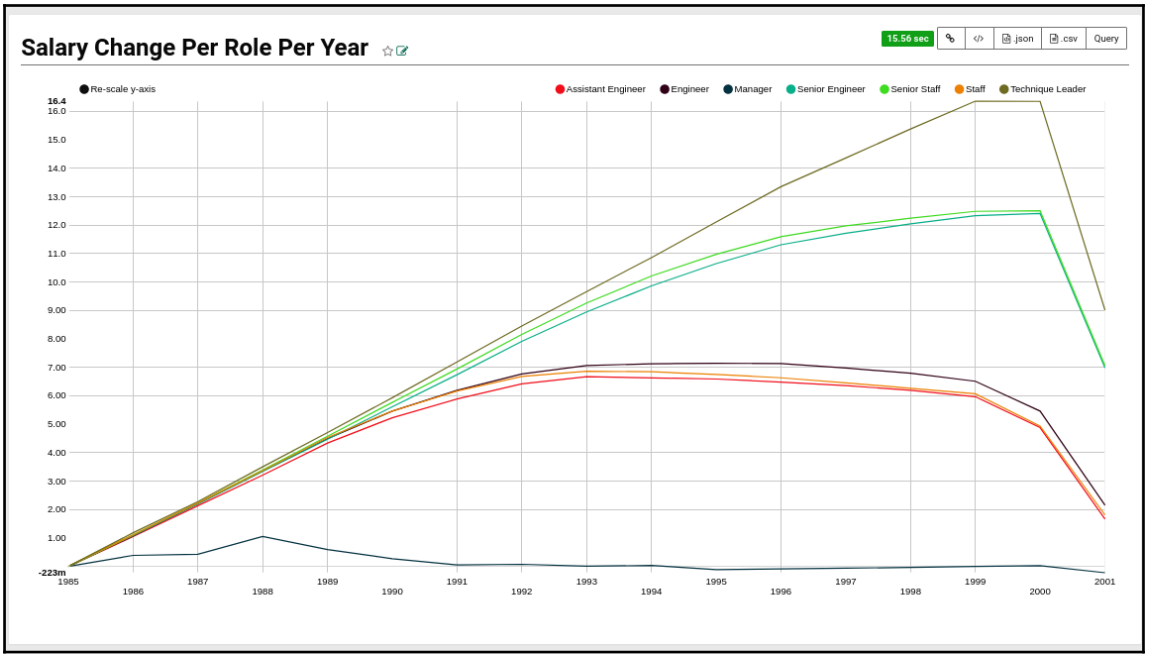

Query
Save as
Datasource & Chart Type
[employees].[employee_norm]
Distribution - NVD3 - Pie Chart
Time
Time Column
birth_date
Time Grain
Time Column
Since
100 years ago
Until
now
Metrics
sum__salary
Group by
dept_name
Series limit
50
Label Type
Percentage
Donut
Legend
Put labels outside


Department Salary Breakup
5.01 sec
.json
.csv
Query
Development
Production
Sales
Marketing
Customer Service
Research
Finance
Quality Management
Human Resources
24%
21%
20%
7%
7%
6%
6%
5%

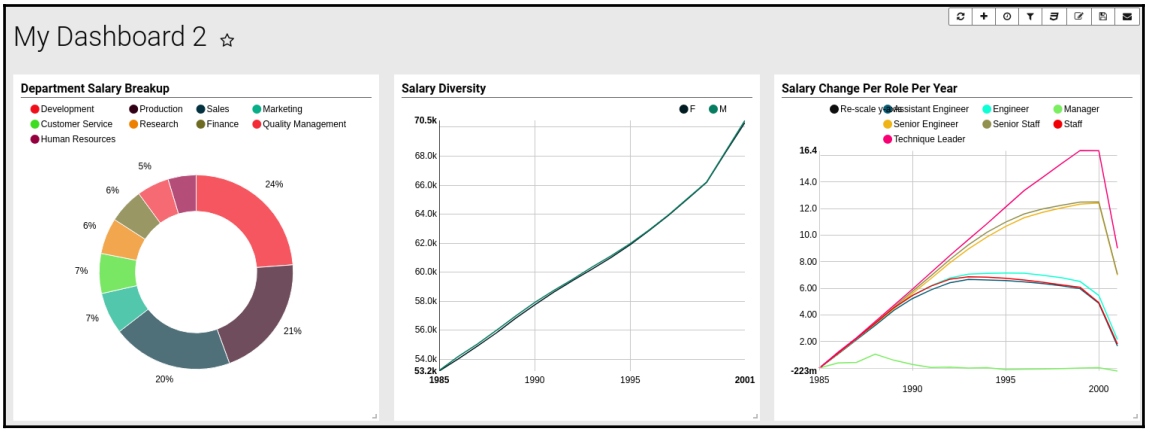

Query
Save as
Datasource & Chart Type
[employees].[employee_norm]
Time Series - Line Chart
Time
Time Column
from_date
Time Grain
year
Since
100 years ago
Until
now
Metrics
avg__salary
Group by
gender
Series limit
50
Sort By


Salary Diversity
11.67 sec
.json
.csv
Query
F
M
70.5k
70.0k
69.0k
68.0k
67.0k
66.0k
65.0k
64.0k
63.0k
62.0k
61.0k
60.0k
59.0k
58.0k
57.0k
56.0k
55.0k
54.0k
53.2k
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001

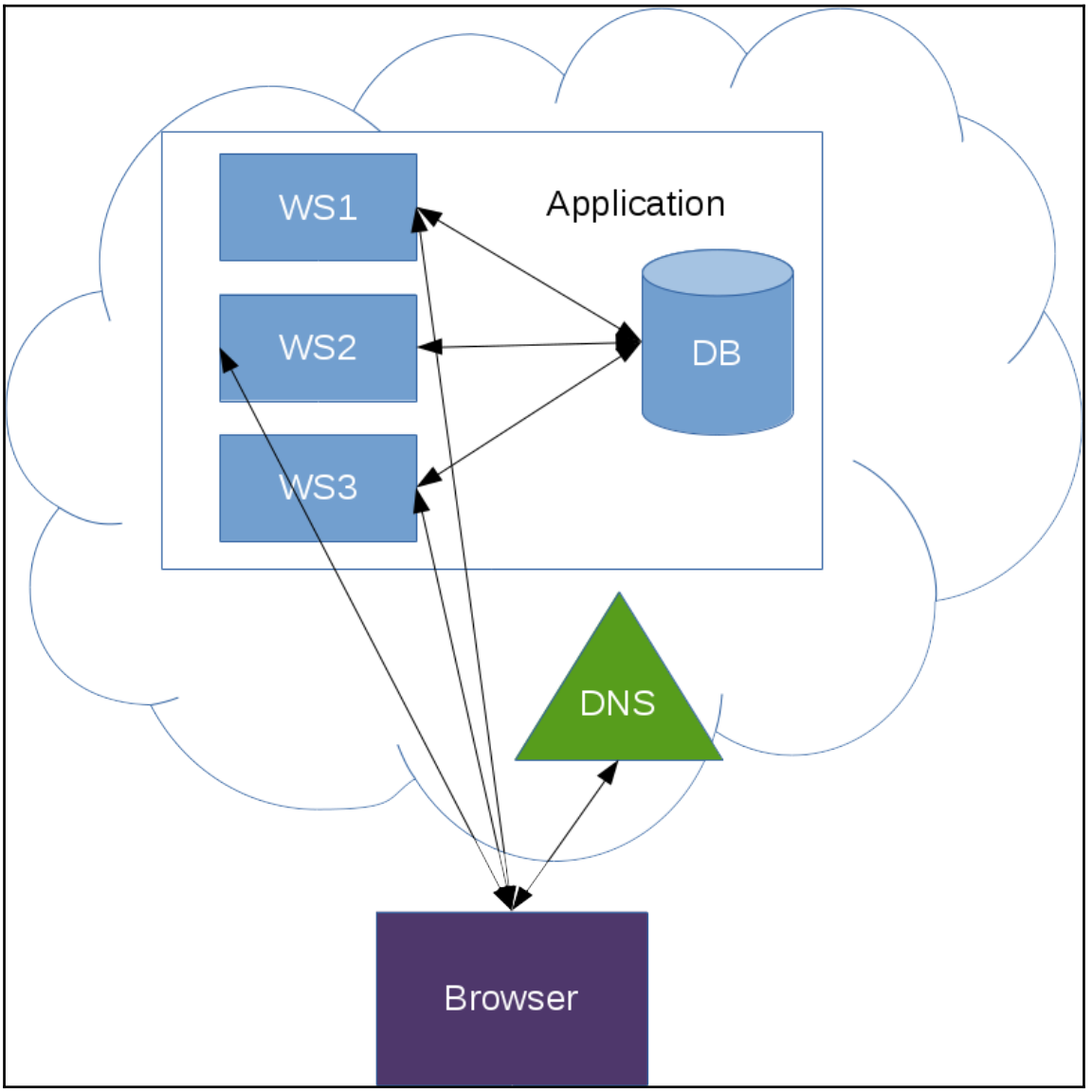

Query
Save as
Datasource & Chart Type
[employees].[employee_norm]
Time Series - Percent Change
Time
Time Column
from_date
Time Grain
year
Since
100 years ago
Until
now
Metrics
sum__salary
Group by
title
Series limit
50
Sort By

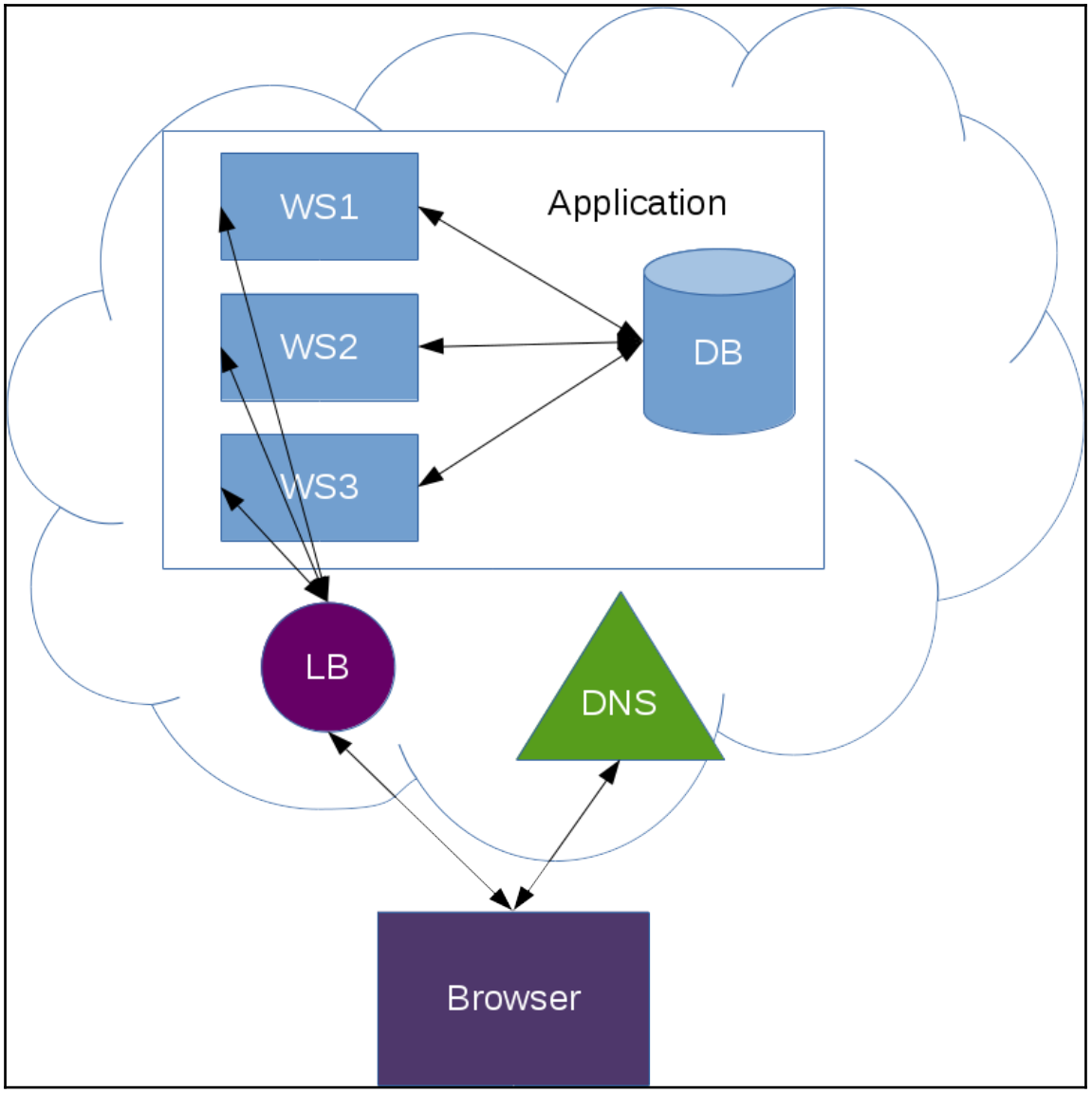


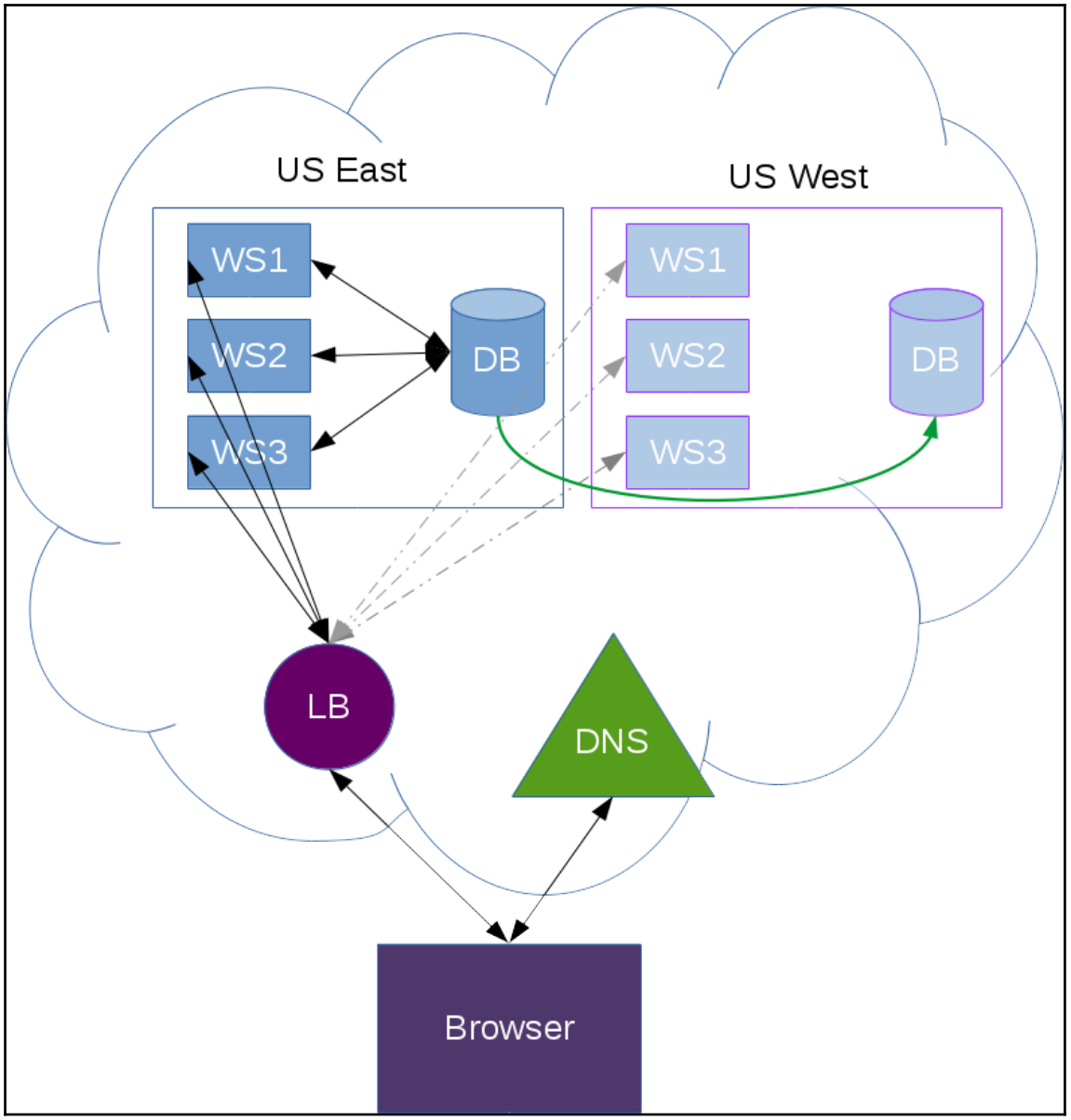
Title	My Dashboard 2
Slug	dash2 To get a readable URL for your dashboard
Slices	× Department Salary Breakup × Salary Diversity × Salary Change Per Role Per Year

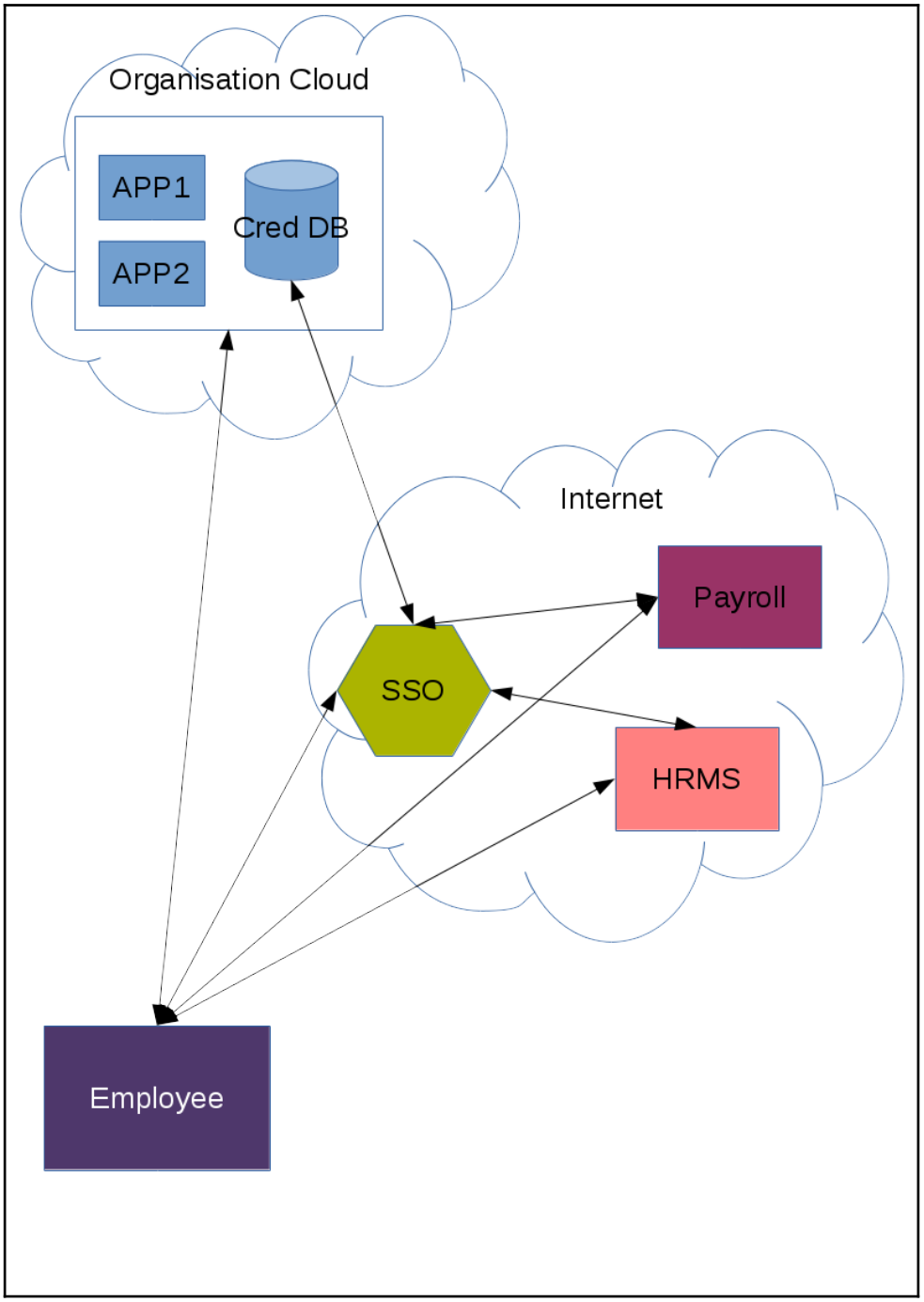

My Dashboard 2
Department Salary Breakup
Development
Production
Sales
Marketing
Customer Service
Research
Finance
Quality Management
Human Resources
24%
21%
20%
7%
7%
6%
6%
5%
Salary Diversity
F
M
70.5k
68.0k
66.0k
64.0k
62.0k
60.0k
58.0k
56.0k
54.0k
53.2k
1985
1990
1995
2001
Salary Change Per Role Per Year
Re-scale y-axis
Assistant Engineer
Engineer
Manager
Senior Engineer
Senior Staff
Staff
Technique Leader
16.4
14.0
12.0
10.0
8.00
6.00
4.00
2.00
-223m
1985
1990
1995
2000

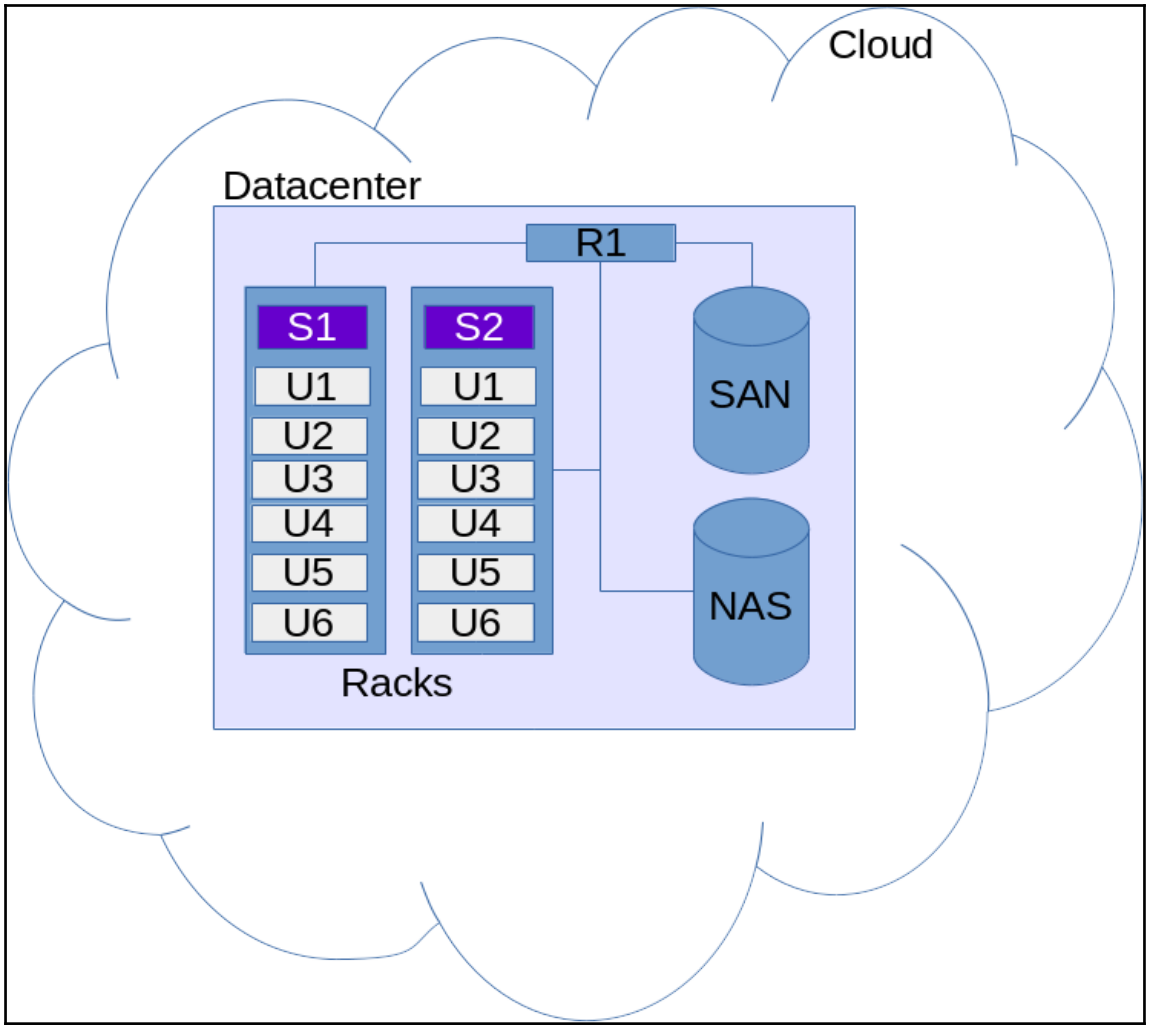
Chapter 10: Developing Applications Using the Cloud

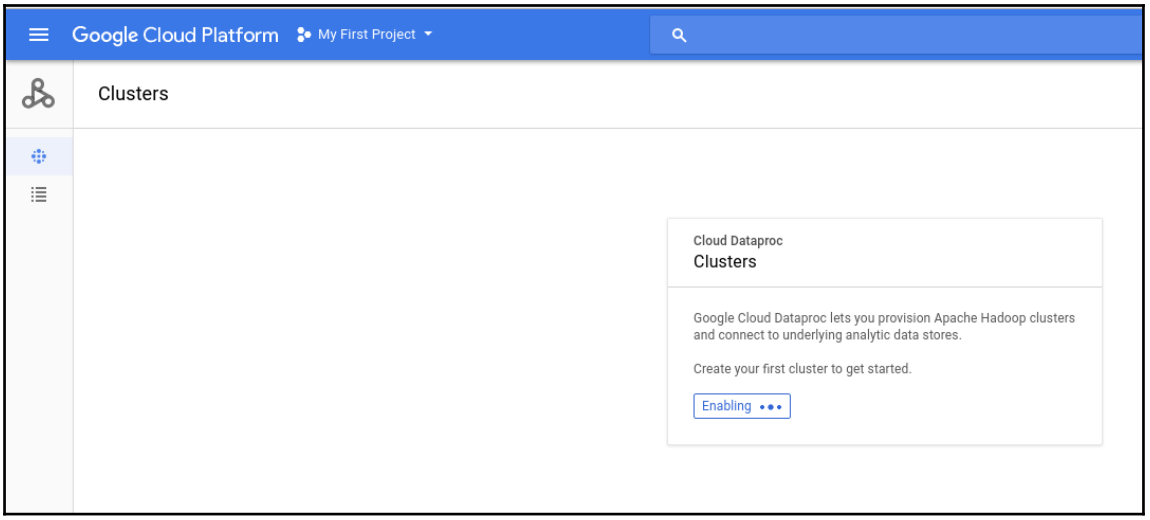


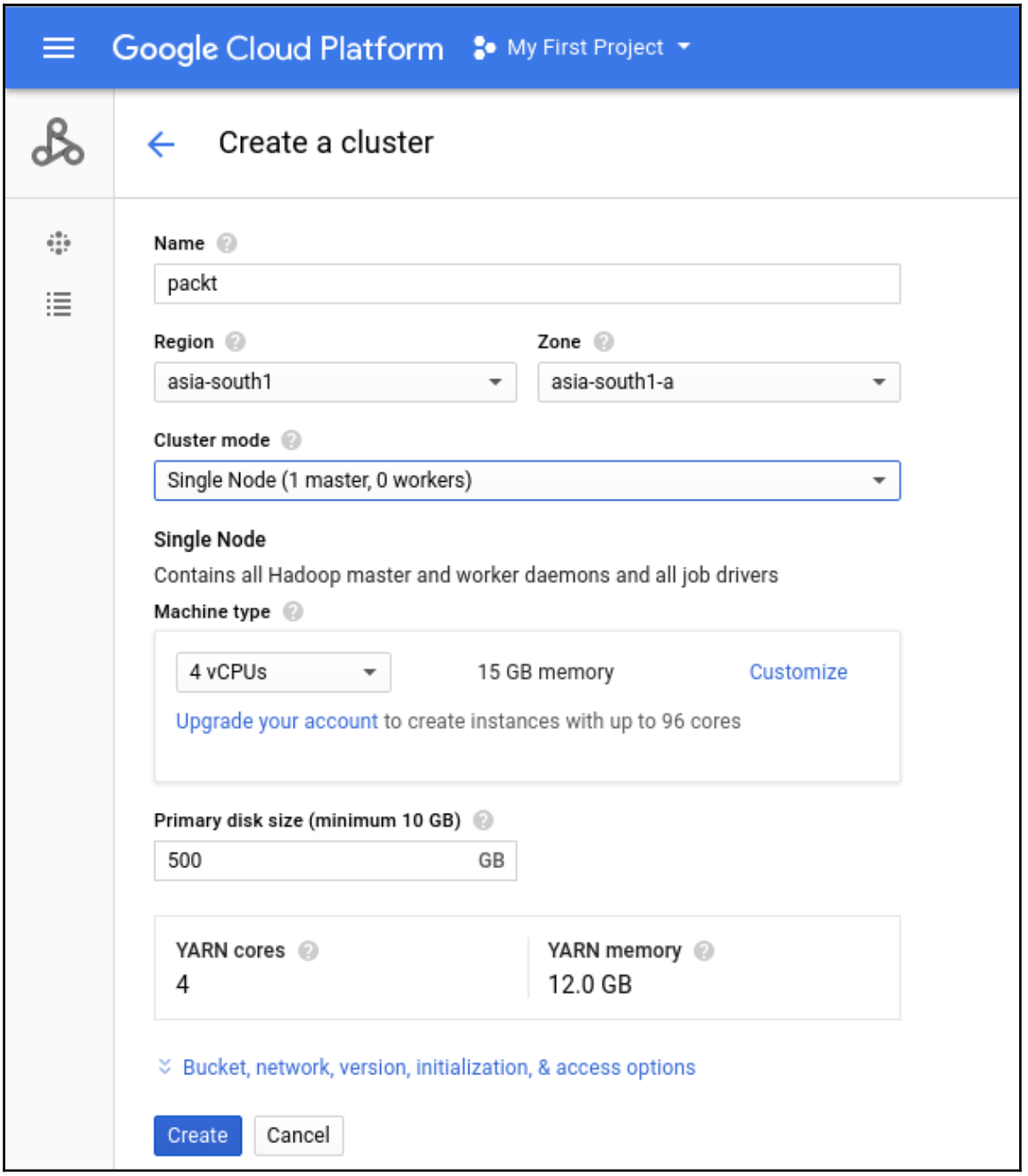

WS1
Application
WS2
DB
WS3
LB
DNS
Browser

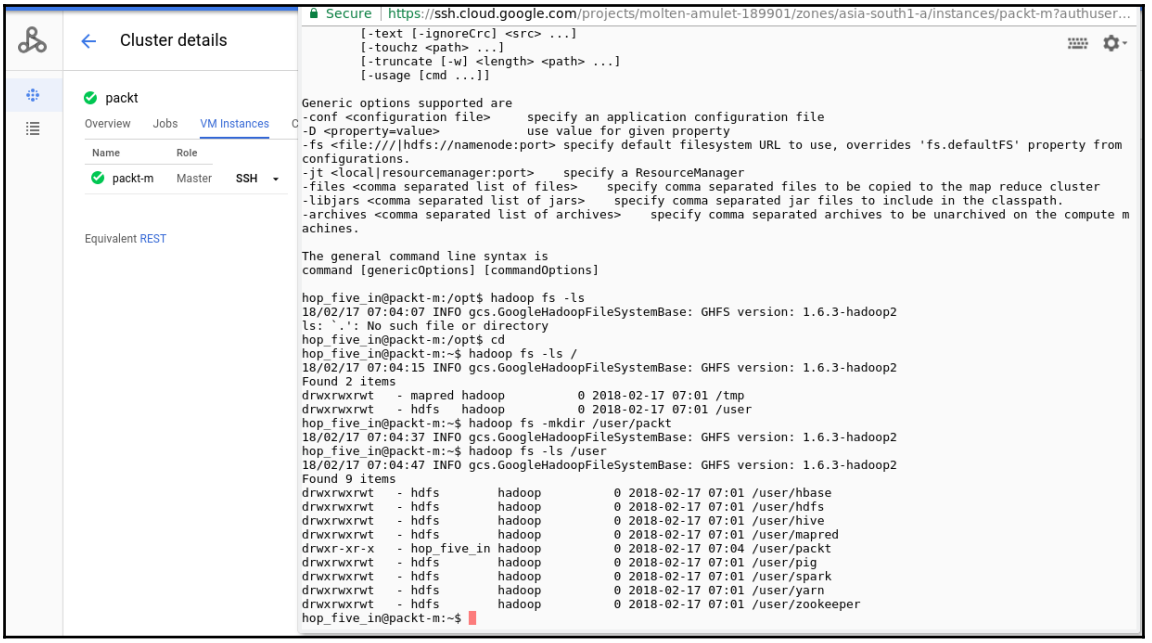

US East
US West
WS1
WS2
WS3
DB
WS1
WS2
WS3
DB
LB
DNS
Browser

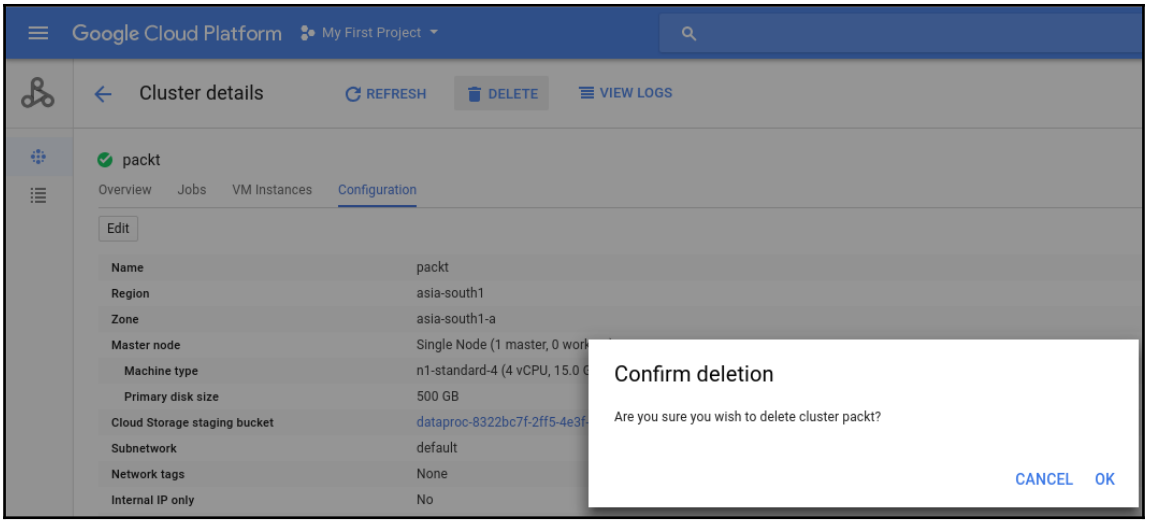

Organisation Cloud
APP1
APP2
Cred DB
Internet
Payroll
SSO
HRMS
Employee

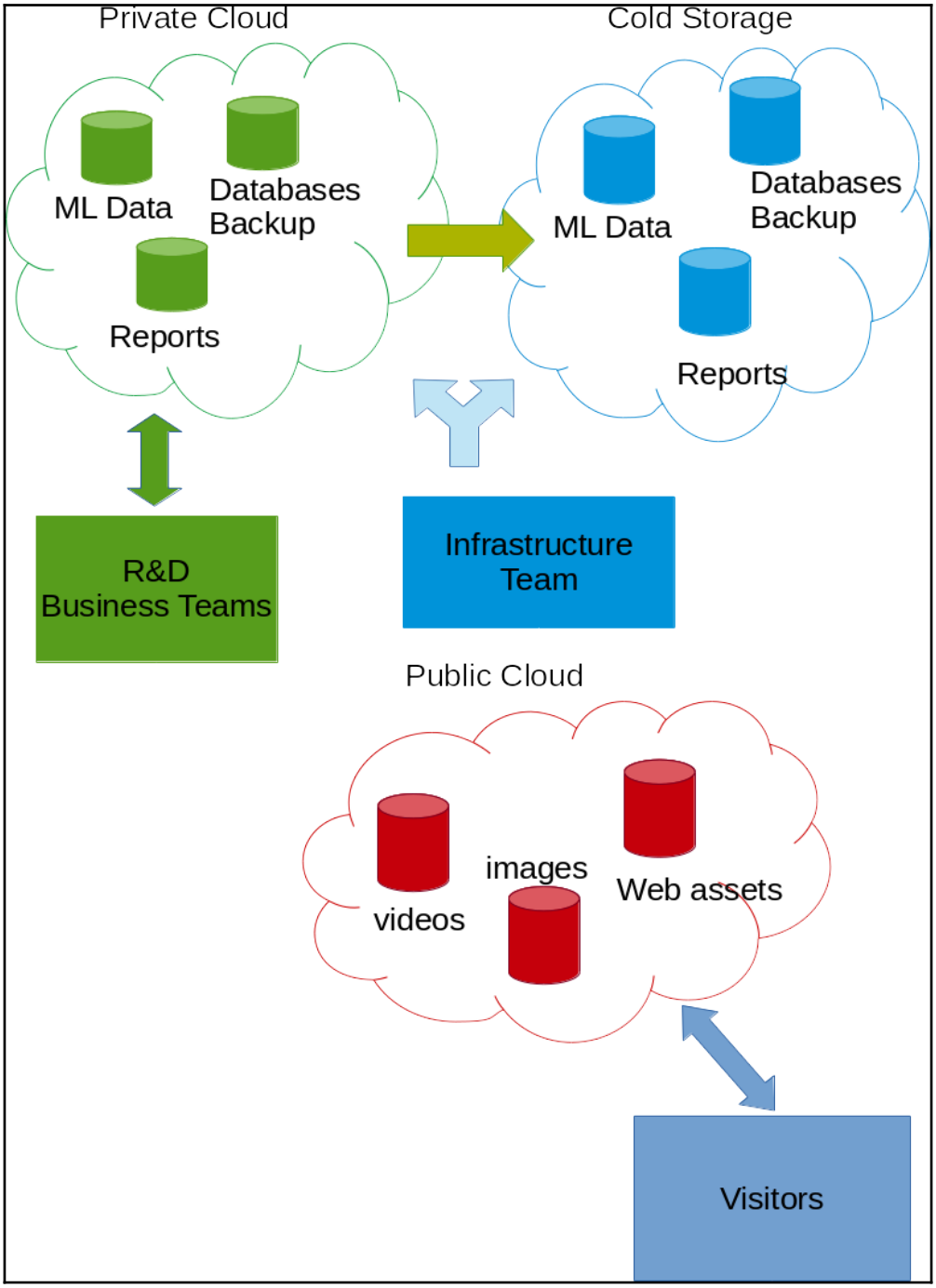

Cloud
Datacenter
R1
S1
U1
U2
U3
U4
U5
U6
S2
U1
U2
U3
U4
U5
U6
SAN
NAS
Racks

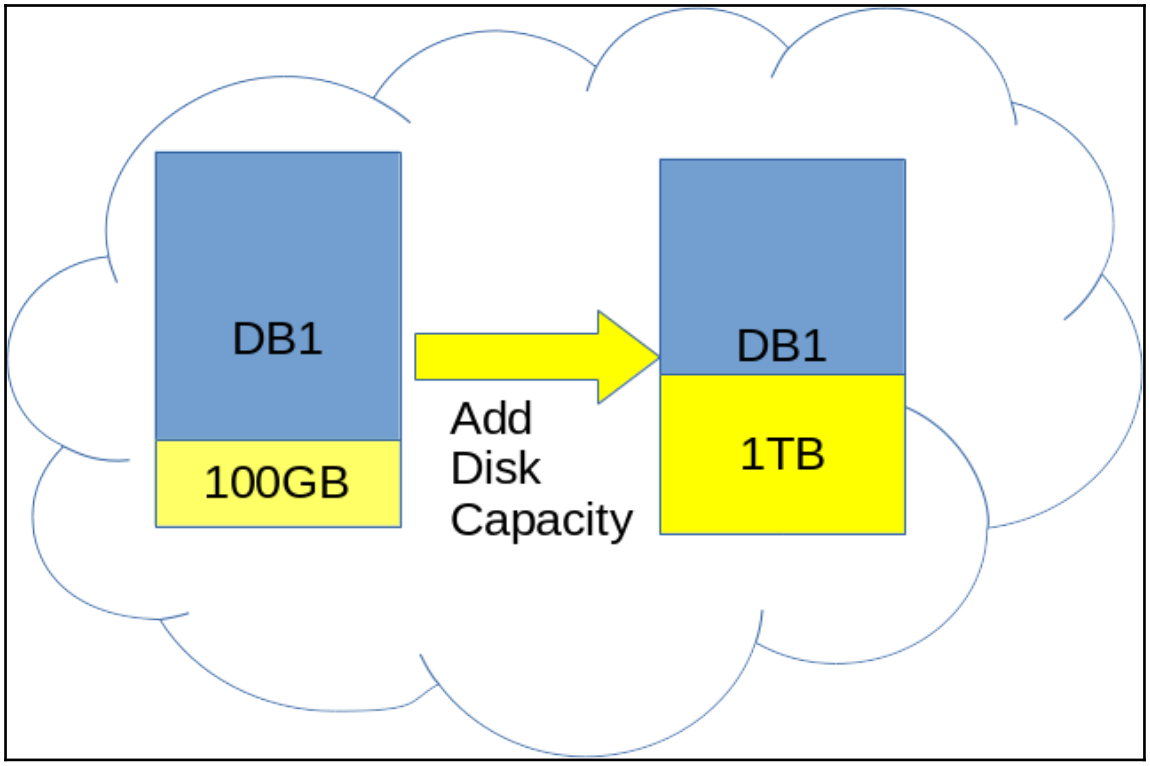

Google Cloud Platform
My First Project
Clusters
Cloud Dataproc
Clusters
Google Cloud Dataproc lets you provision Apache Hadoop clusters and connect to underlying analytic data stores.
Create your first cluster to get started.
Enabling

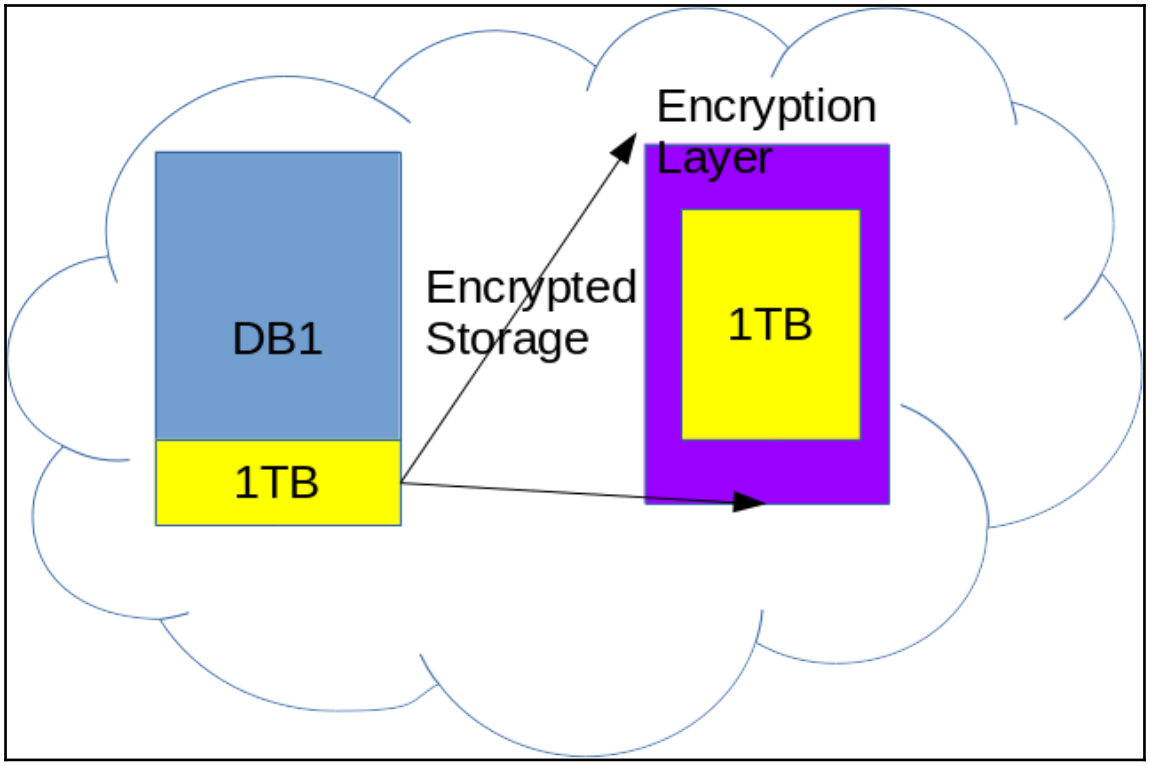

Google Cloud Platform
My First Project
Create a cluster
Name
packt
Region
asia-south1
Zone
asia-south1-a
Cluster mode
Single Node (1 master, 0 workers)
Single Node
Contains all Hadoop master and worker daemons and all job drivers
Machine type
4 vCPUs
15 GB memory
Customize
Upgrade your account to create instances with up to 96 cores
Primary disk size (minimum 10 GB)
500
GB
YARN cores
4
YARN memory
12.0 GB
Bucket, network, version, initialization, & access options
Create
Cancel


Cluster details
packt
Overview
Jobs
VM Instances
Name
Role
packt-m
Master
SSH
Equivalent REST
Secure | https://ssh.cloud.google.com/projects/molten-amulet-189901/zones/asia-south1-a/instances/packt-m?authuser...
[-text [-ignoreCrc] <src> ...]
[-touchz <path> ...]
[-truncate [-w] <length> <path> ...]
[-usage [cmd ...]]
Generic options supported are
-conf <configuration file> specify an application configuration file
-D <property=value> use value for given property
-fs <file:///|hdfs://namenode:port> specify default filesystem URL to use, overrides 'fs.defaultFS' property from configurations.
-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
command [genericOptions] [commandOptions]
hop_five_in@packt-m:/opt$ hadoop fs -ls
18/02/17 07:04:07 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.3-hadoop2
ls: `.': No such file or directory
hop_five_in@packt-m:/opt$ cd
hop_five_in@packt-m:~$ hadoop fs -ls /
18/02/17 07:04:15 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.3-hadoop2
Found 2 items
drwxrwxrwt - mapred hadoop 0 2018-02-17 07:01 /tmp
drwxrwxrwt - hdfs hadoop 0 2018-02-17 07:01 /user
hop_five_in@packt-m:~$ hadoop fs -mkdir /user/packt
18/02/17 07:04:37 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.3-hadoop2
hop_five_in@packt-m:~$ hadoop fs -ls /user
18/02/17 07:04:47 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.3-hadoop2
Found 9 items
drwxrwxrwt - hdfs hadoop 0 2018-02-17 07:01 /user/hbase
drwxrwxrwt - hdfs hadoop 0 2018-02-17 07:01 /user/hdfs
drwxrwxrwt - hdfs hadoop 0 2018-02-17 07:01 /user/hive
drwxrwxrwt - hdfs hadoop 0 2018-02-17 07:01 /user/mapred
drwxr-xr-x - hop_five_in hadoop 0 2018-02-17 07:04 /user/packt
drwxrwxrwt - hdfs hadoop 0 2018-02-17 07:01 /user/pig
drwxrwxrwt - hdfs hadoop 0 2018-02-17 07:01 /user/spark
drwxrwxrwt - hdfs hadoop 0 2018-02-17 07:01 /user/yarn
drwxrwxrwt - hdfs hadoop 0 2018-02-17 07:01 /user/zookeeper
hop_five_in@packt-m:~$

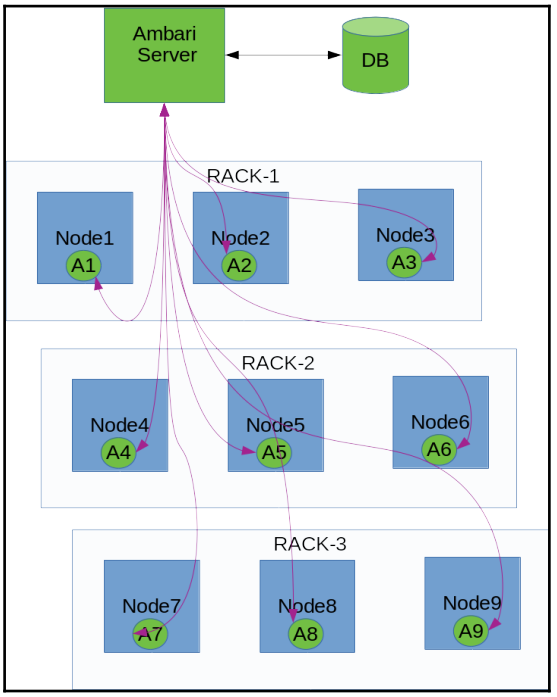

Google Cloud Platform
My First Project
Cluster details
REFRESH
DELETE
VIEW LOGS
packt
Overview
Jobs
VM Instances
Configuration
Edit
Name
packt
Region
asia-south1
Zone
asia-south1-a
Master node
Machine type
Primary disk size
500 GB
Cloud Storage staging bucket
Subnetwork
default
Network tags
None
Internal IP only
No
Confirm deletion
Are you sure you wish to delete cluster packt?
CANCEL
OK

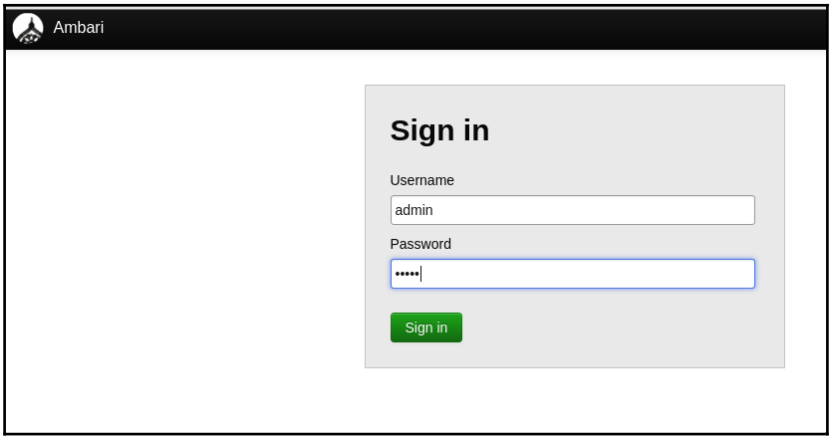

Private Cloud
ML Data
Databases Backup
Reports
Cold Storage
ML Data
Databases Backup
Reports
R&D Business Teams
Infrastructure Team
Public Cloud
videos
images
Web assets
Visitors

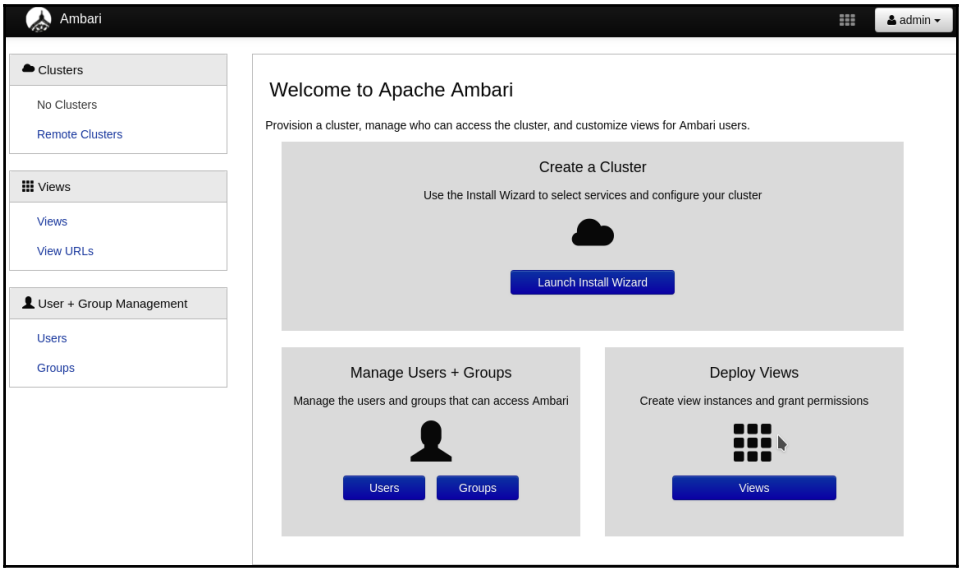

DB1
100GB
Add
Disk
Capacity
DB1
1TB

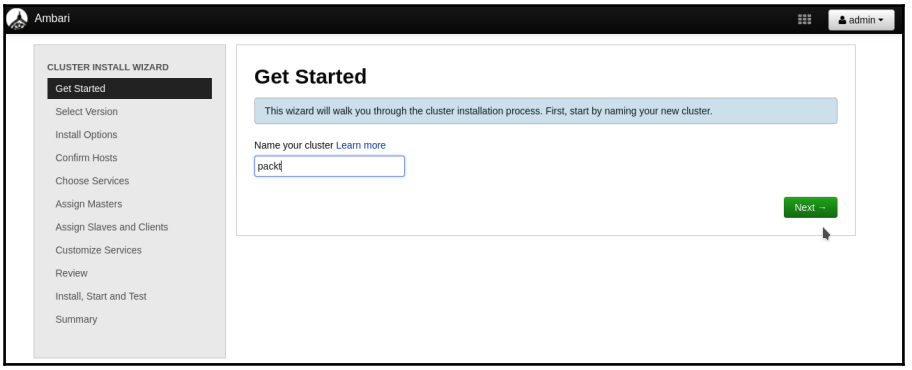

Encryption
Layer
DB1
Encrypted
Storage
1TB
1TB

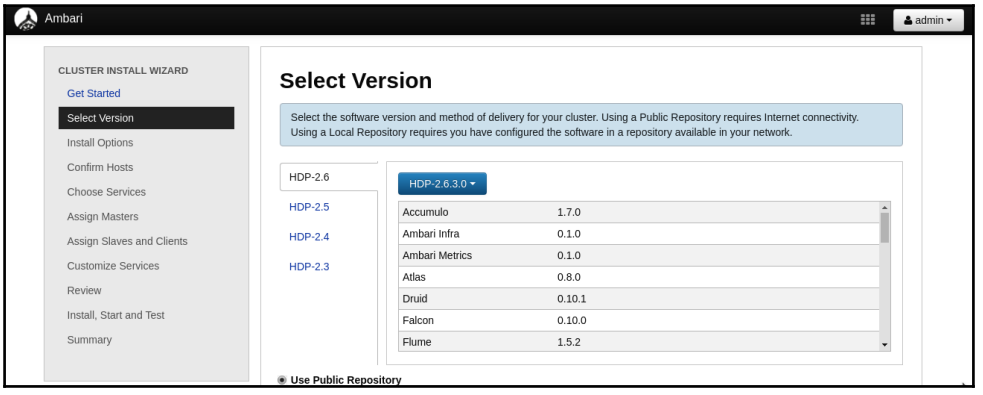
Chapter 11: Production Hadoop Cluster Deployment

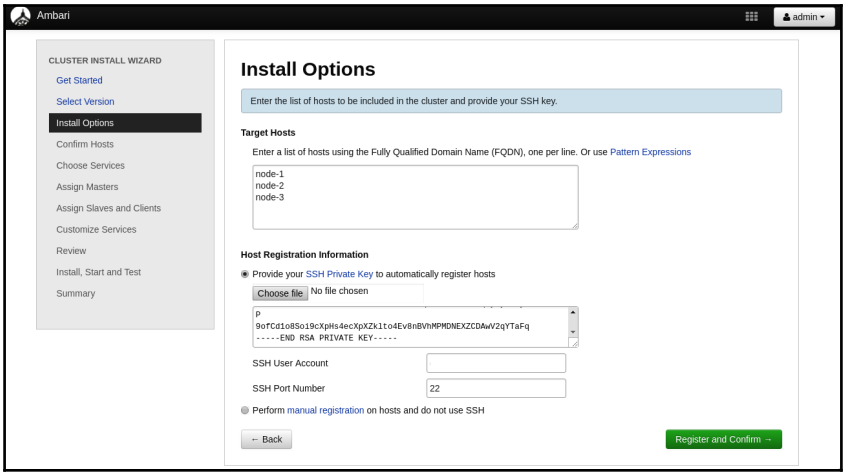


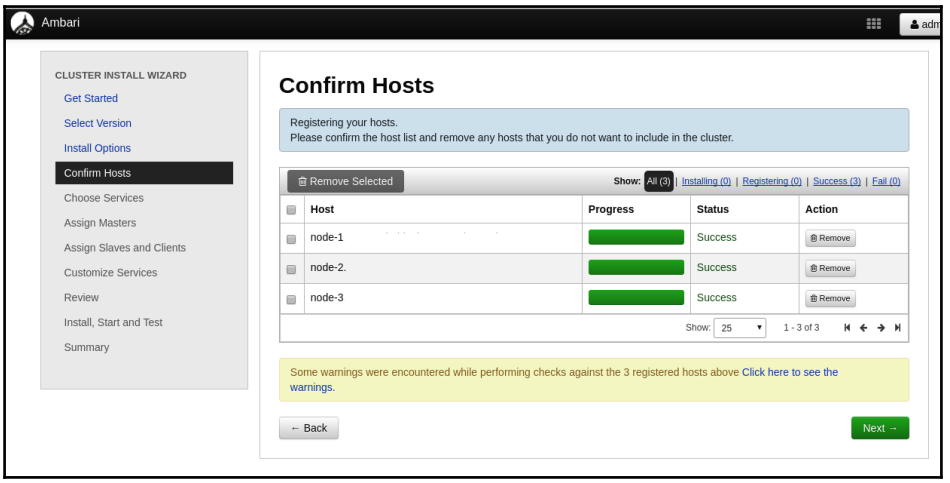


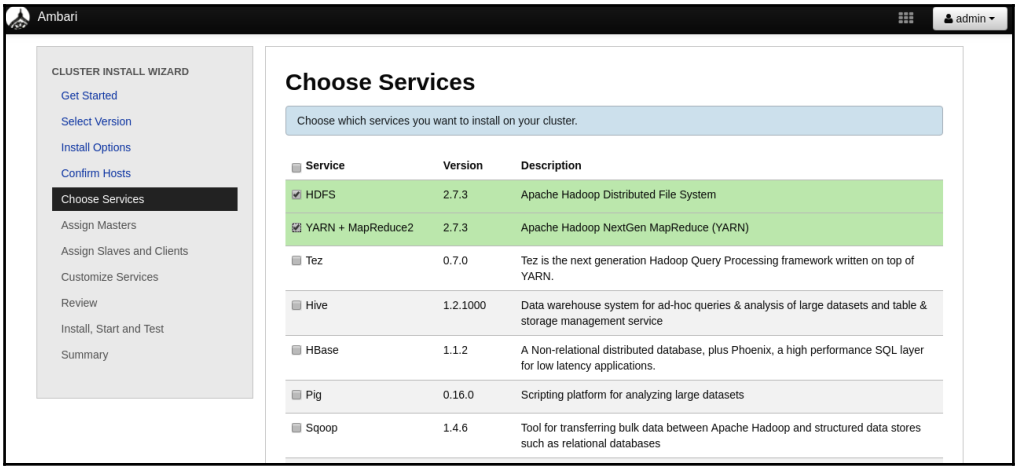

Ambari
admin
Clusters
No Clusters
Remote Clusters
Views
Views
View URLs
User + Group Management
Users
Groups
Welcome to Apache Ambari
Provision a cluster, manage who can access the cluster, and customize views for Ambari users.
Create a Cluster
Use the Install Wizard to select services and configure your cluster
Launch Install Wizard
Manage Users + Groups
Manage the users and groups that can access Ambari
Users
Groups
Deploy Views
Create view instances and grant permissions
Views

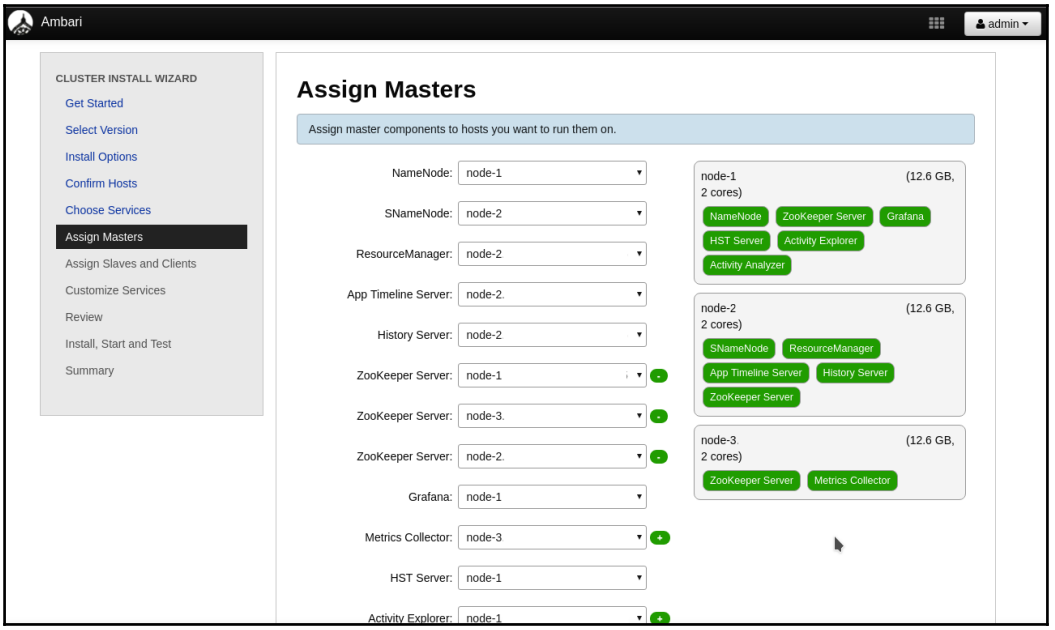

Ambari
admin
CLUSTER INSTALL WIZARD
Get Started
Select Version
Install Options
Confirm Hosts
Choose Services
Assign Masters
Assign Slaves and Clients
Customize Services
Review
Install, Start and Test
Summary
Get Started
This wizard will walk you through the cluster installation process. First, start by naming your new cluster.
Name your cluster Learn more
packt
Next →

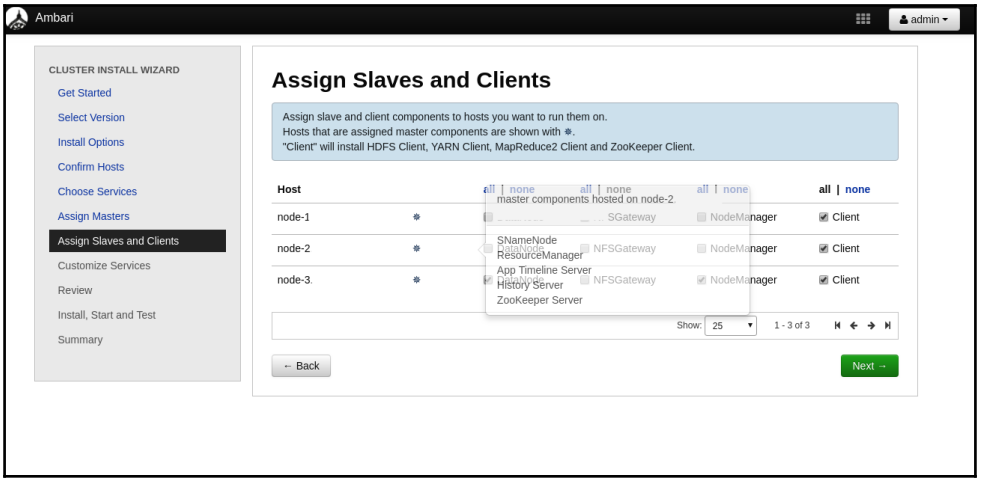

Ambari
admin
CLUSTER INSTALL WIZARD
Get Started
Select Version
Install Options
Confirm Hosts
Choose Services
Assign Masters
Assign Slaves and Clients
Customize Services
Review
Install, Start and Test
Summary
Select Version
Select the software version and method of delivery for your cluster. Using a Public Repository requires Internet connectivity.
Using a Local Repository requires you have configured the software in a repository available in your network.
HDP-2.6
HDP-2.5
HDP-2.4
HDP-2.3
HDP-2.6.3.0
Accumulo 1.7.0
Ambari Infra 0.1.0
Ambari Metrics 0.1.0
Atlas 0.8.0
Druid 0.10.1
Falcon 0.10.0
Flume 1.5.2
Use Public Repository

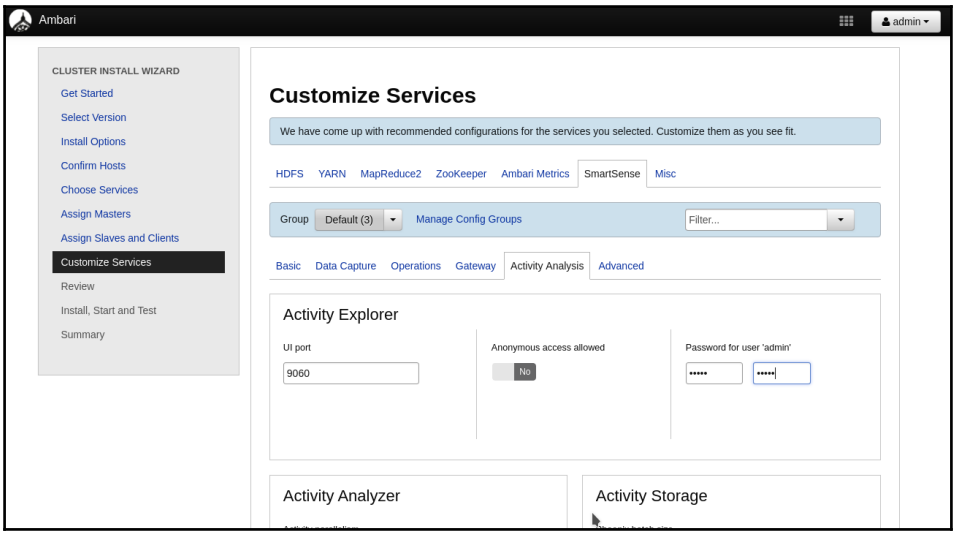

Ambari
admin
CLUSTER INSTALL WIZARD
Get Started
Select Version
Install Options
Confirm Hosts
Choose Services
Assign Masters
Assign Slaves and Clients
Customize Services
Review
Install, Start and Test
Summary
Install Options
Enter the list of hosts to be included in the cluster and provide your SSH key.
Target Hosts
Enter a list of hosts using the Fully Qualified Domain Name (FQDN), one per line. Or use Pattern Expressions
node-1
node-2
node-3
Host Registration Information
Provide your SSH Private Key to automatically register hosts
Choose file
No file chosen
P
9ofCd1o8So19cXpHs4ecXpXZk1to4Ev8nBVhMPMDNEXZCDAwV2qYTaFq
-----END RSA PRIVATE KEY-----
SSH User Account
SSH Port Number
22
Perform manual registration on hosts and do not use SSH
← Back
Register and Confirm →

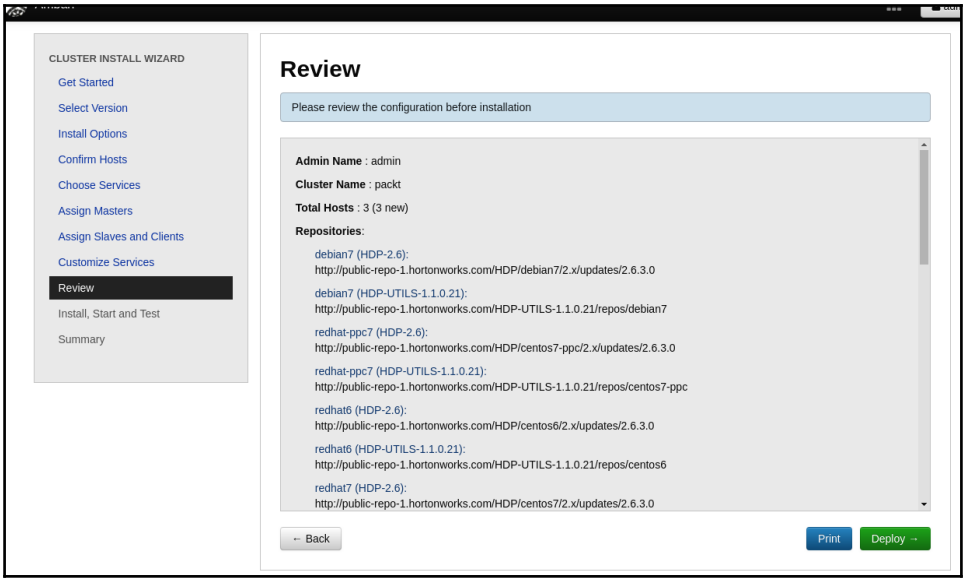

Ambari
CLUSTER INSTALL WIZARD
Get Started
Select Version
Install Options
Confirm Hosts
Choose Services
Assign Masters
Assign Slaves and Clients
Customize Services
Review
Install, Start and Test
Summary
Confirm Hosts
Registering your hosts.
Please confirm the host list and remove any hosts that you do not want to include in the cluster.
Remove Selected
Show: All (3) | Installing (0) | Registering (0) | Success (3) | Fail (0)
Host
Progress
Status
Action
node-1
Success
Remove
node-2.
Success
Remove
node-3
Success
Remove
Show: 25
1 - 3 of 3
Some warnings were encountered while performing checks against the 3 registered hosts above Click here to see the warnings.
← Back
Next →

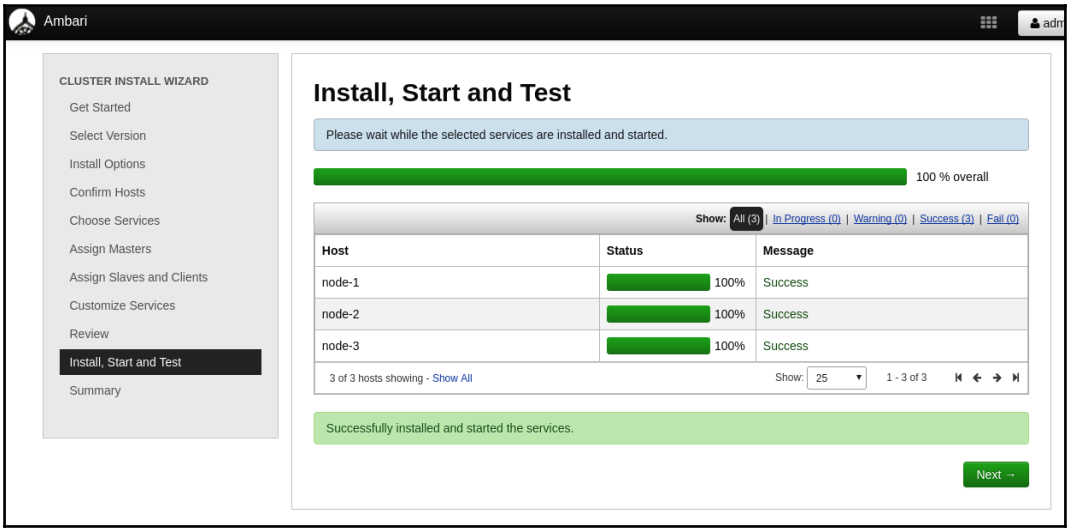

Ambari
admin
CLUSTER INSTALL WIZARD
Get Started
Select Version
Install Options
Confirm Hosts
Choose Services
Assign Masters
Assign Slaves and Clients
Customize Services
Review
Install, Start and Test
Summary
Choose Services
Choose which services you want to install on your cluster.
Service
Version
Description
HDFS
2.7.3
Apache Hadoop Distributed File System
YARN + MapReduce2
2.7.3
Apache Hadoop NextGen MapReduce (YARN)
Tez
0.7.0
Tez is the next generation Hadoop Query Processing framework written on top of YARN.
Hive
1.2.1000
Data warehouse system for ad-hoc queries & analysis of large datasets and table & storage management service
HBase
1.1.2
A Non-relational distributed database, plus Phoenix, a high performance SQL layer for low latency applications.
Pig
0.16.0
Scripting platform for analyzing large datasets
Sqoop
1.4.6
Tool for transferring bulk data between Apache Hadoop and structured data stores such as relational databases

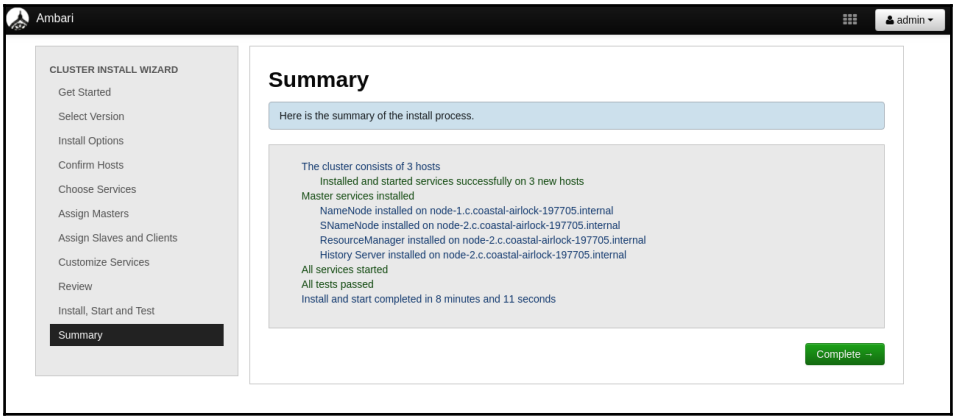

Ambari
admin
CLUSTER INSTALL WIZARD
Get Started
Select Version
Install Options
Confirm Hosts
Choose Services
Assign Masters
Assign Slaves and Clients
Customize Services
Review
Install, Start and Test
Summary
Assign Masters
Assign master components to hosts you want to run them on.
NameNode: node-1
SNameNode: node-2
ResourceManager: node-2
App Timeline Server: node-2
History Server: node-2
ZooKeeper Server: node-1
ZooKeeper Server: node-3
ZooKeeper Server: node-2
Grafana: node-1
Metrics Collector: node-3
HST Server: node-1
Activity Explorer: node-1
node-1 (12.6 GB, 2 cores)
NameNode
ZooKeeper Server
Grafana
HST Server
Activity Explorer
Activity Analyzer
node-2 (12.6 GB, 2 cores)
SNameNode
ResourceManager
App Timeline Server
History Server
ZooKeeper Server
node-3 (12.6 GB, 2 cores)
ZooKeeper Server
Metrics Collector

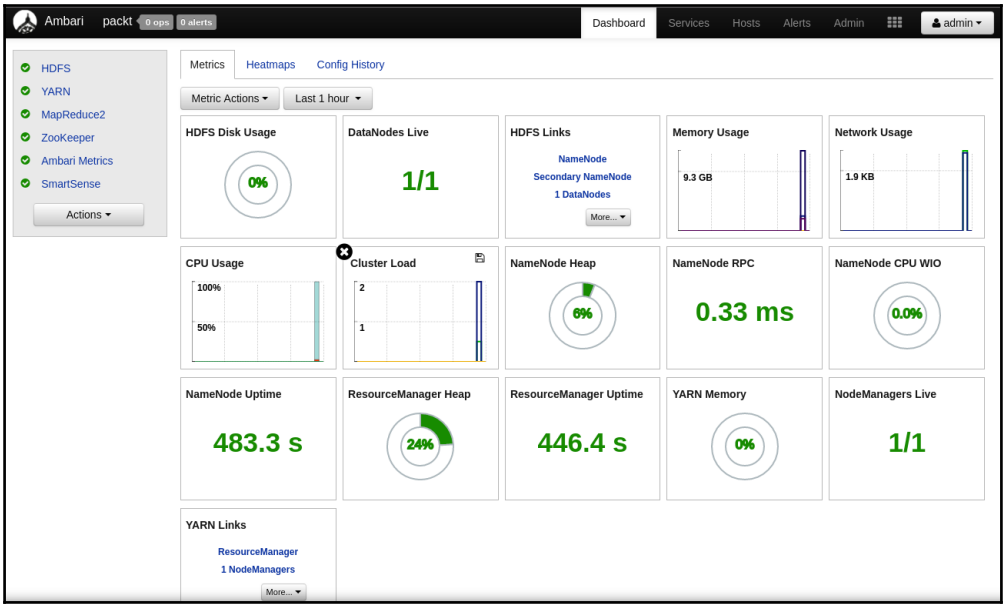

Ambari
admin
CLUSTER INSTALL WIZARD
Get Started
Select Version
Install Options
Confirm Hosts
Choose Services
Assign Masters
Assign Slaves and Clients
Customize Services
Review
Install, Start and Test
Summary
Assign Slaves and Clients
Assign slave and client components to hosts you want to run them on.
Hosts that are assigned master components are shown with ❄.
"Client" will install HDFS Client, YARN Client, MapReduce2 Client and ZooKeeper Client.
Host
all | none
all | none
all | none
all | none
master components hosted on node-2
node-1
node-2
node-3
SGateway
NodeManager
Client
SNameNode
ResourceManager
App Timeline Server
History Server
ZooKeeper Server
NFSGateway
Show:
25
1 - 3 of 3
← Back
Next →

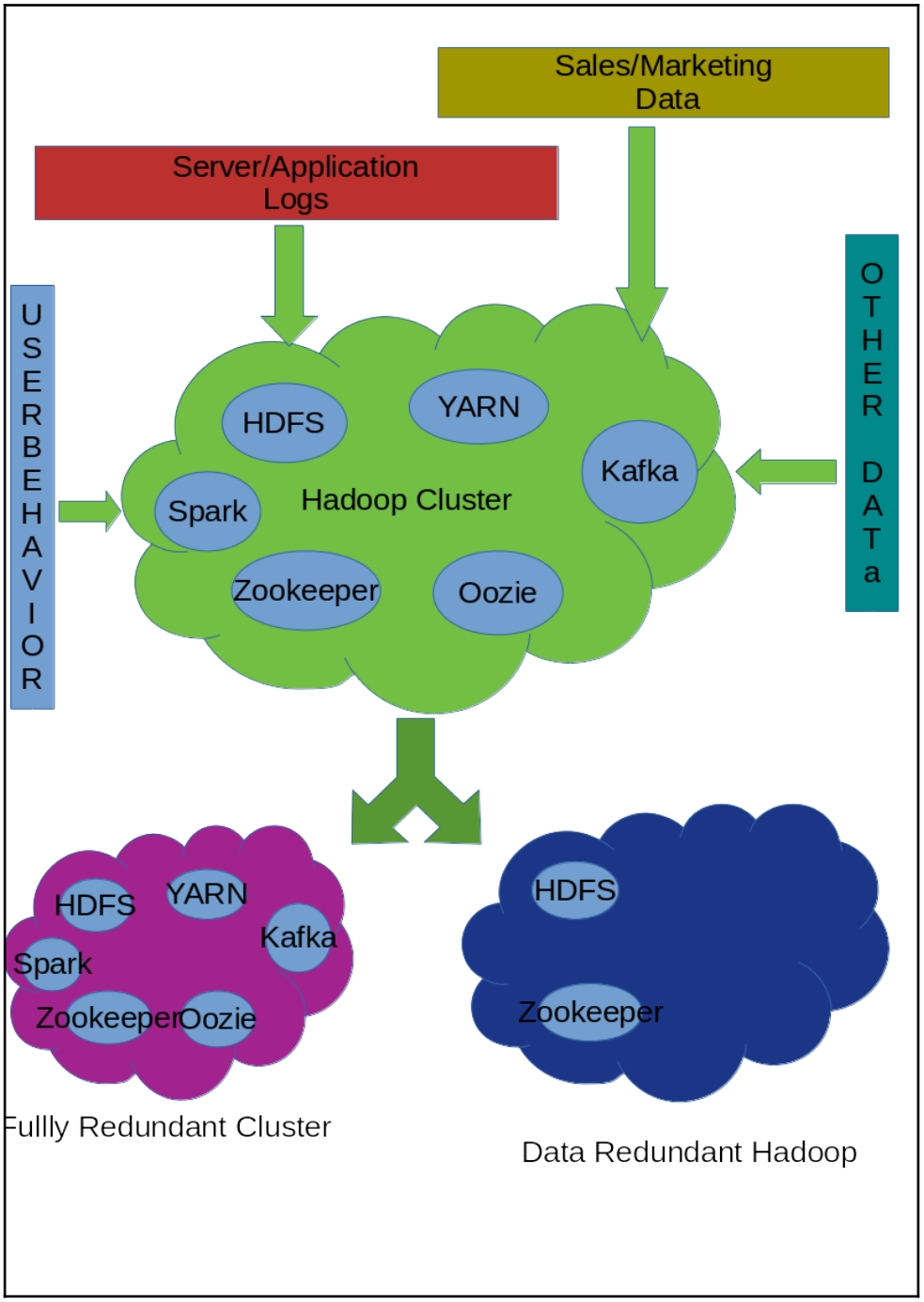

Ambari
admin
CLUSTER INSTALL WIZARD
Get Started
Select Version
Install Options
Confirm Hosts
Choose Services
Assign Masters
Assign Slaves and Clients
Customize Services
Review
Install, Start and Test
Summary
Customize Services
We have come up with recommended configurations for the services you selected. Customize them as you see fit.
HDFS
YARN
MapReduce2
ZooKeeper
Ambari Metrics
SmartSense
Misc
Group
Default (3)
Manage Config Groups
Filter...
Basic
Data Capture
Operations
Gateway
Activity Analysis
Advanced
Activity Explorer
UI port
9060
Anonymous access allowed
No
Password for user 'admin'
Activity Analyzer
Activity Storage

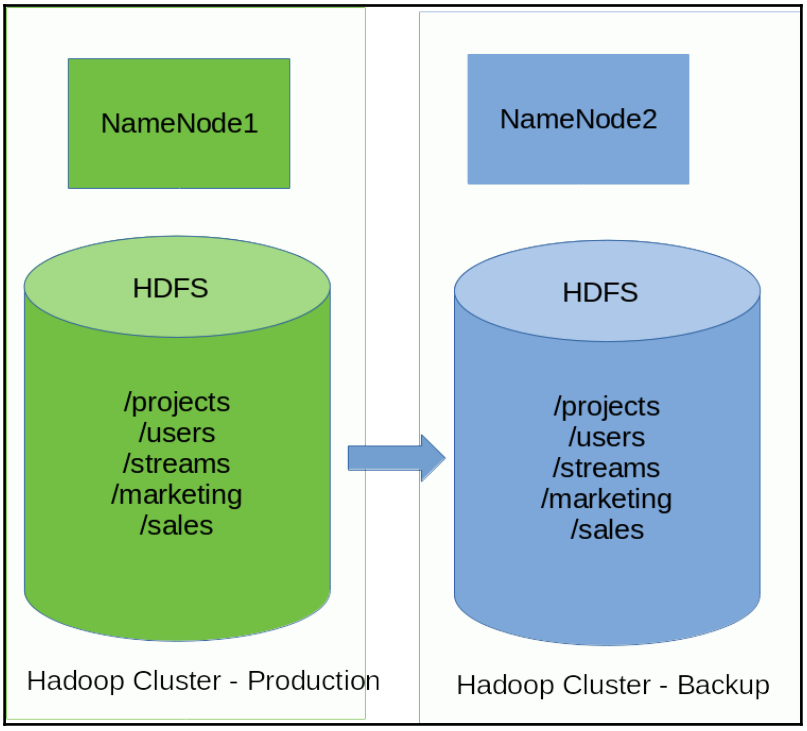

CLUSTER INSTALL WIZARD
Get Started
Select Version
Install Options
Confirm Hosts
Choose Services
Assign Masters
Assign Slaves and Clients
Customize Services
Review
Install, Start and Test
Summary
Review
Please review the configuration before installation
Admin Name : admin
Cluster Name : packt
Total Hosts : 3 (3 new)
Repositories:
debian7 (HDP-2.6):
http://public-repo-1.hortonworks.com/HDP/debian7/2.x/updates/2.6.3.0
debian7 (HDP-UTILS-1.1.0.21):
http://public-repo-1.hortonworks.com/HDP-UTILS-1.1.0.21/repos/debian7
redhat-ppc7 (HDP-2.6):
http://public-repo-1.hortonworks.com/HDP/centos7-ppc/2.x/updates/2.6.3.0
redhat-ppc7 (HDP-UTILS-1.1.0.21):
http://public-repo-1.hortonworks.com/HDP-UTILS-1.1.0.21/repos/centos7-ppc
redhat6 (HDP-2.6):
http://public-repo-1.hortonworks.com/HDP/centos6/2.x/updates/2.6.3.0
redhat6 (HDP-UTILS-1.1.0.21):
http://public-repo-1.hortonworks.com/HDP-UTILS-1.1.0.21/repos/centos6
redhat7 (HDP-2.6):
http://public-repo-1.hortonworks.com/HDP/centos7/2.x/updates/2.6.3.0
← Back
Print
Deploy →


Ambari
CLUSTER INSTALL WIZARD
Get Started
Select Version
Install Options
Confirm Hosts
Choose Services
Assign Masters
Assign Slaves and Clients
Customize Services
Review
Install, Start and Test
Summary
Install, Start and Test
Please wait while the selected services are installed and started.
100 % overall
Show: All (3) | In Progress (0) | Warning (0) | Success (3) | Fail (0)
Host
Status
Message
node-1
100%
Success
node-2
100%
Success
node-3
100%
Success
3 of 3 hosts showing - Show All
Show: 25
1 - 3 of 3
Successfully installed and started the services.
Next →


Ambari
admin
CLUSTER INSTALL WIZARD
Get Started
Select Version
Install Options
Confirm Hosts
Choose Services
Assign Masters
Assign Slaves and Clients
Customize Services
Review
Install, Start and Test
Summary
Summary
Here is the summary of the install process.
The cluster consists of 3 hosts
Installed and started services successfully on 3 new hosts
Master services installed
NameNode installed on node-1.c.coastal-airlock-197705.internal
SNameNode installed on node-2.c.coastal-airlock-197705.internal
ResourceManager installed on node-2.c.coastal-airlock-197705.internal
History Server installed on node-2.c.coastal-airlock-197705.internal
All services started
All tests passed
Install and start completed in 8 minutes and 11 seconds
Complete →


Ambari
packt
0 ops
0 alerts
Dashboard
Services
Hosts
Alerts
Admin
admin
HDFS
YARN
MapReduce2
ZooKeeper
Ambari Metrics
SmartSense
Actions
Metrics
Heatmaps
Config History
Metric Actions
Last 1 hour
HDFS Disk Usage
0%
DataNodes Live
1/1
HDFS Links
NameNode
Secondary NameNode
1 DataNodes
More...
Memory Usage
9.3 GB
Network Usage
1.9 KB
CPU Usage
100%
50%
Cluster Load
2
1
NameNode Heap
6%
NameNode RPC
0.33 ms
NameNode CPU WIO
0.0%
NameNode Uptime
483.3 s
ResourceManager Heap
24%
ResourceManager Uptime
446.4 s
YARN Memory
0%
NodeManagers Live
1/1
YARN Links
ResourceManager
1 NodeManagers
More...


Sales/Marketing
Data
Server/Application
Logs
USERBEHAVIOR
OTHER DATa
HDFS
YARN
Kafka
Spark
Hadoop Cluster
Zookeeper
Oozie
HDFS
YARN
Kafka
Spark
Zookeeper
Oozie
Fullly Redundant Cluster
HDFS
Zookeeper
Data Redundant Hadoop


NameNode1
NameNode2
HDFS
/projects
/users
/streams
/marketing
/sales
HDFS
/projects
/users
/streams
/marketing
/sales
Hadoop Cluster - Production
Hadoop Cluster - Backup