

Project number:	317871
Project acronym:	BIOBANKCLOUD

<p>WORK PACKAGE 2 :</p> <p>SCALABLE STORAGE</p>

Work Package Leader Name and Organisation:

Jim Dowling, KTH – Royal College of Technology (KTH)

E-mail: jdowling@kth.se

PROJECT DELIVERABLE

D2.1: Highly Available HDFS

Deliverable Due date (and month since project start): 2013-11-30, m12

Document history

Version	Date	Changes	By	Reviewed
0.1	2013-11-23	First version	Salman Niazi Kamal Hakimzadeh Alberto Lorente Mahmoud Ismail	Jim Dowling

BiobankCloud D2.1

317871

Executive Summary

This deliverable consists of a software deliverable of the highly available Hadoop Filesystem (HDFS), a userguide for the software, and a short description of the system's architecture.

Our implementation of HDFS provides a new distributed model for HDFS' metadata, based on storing the metadata in MySQL Cluster, a distributed, in-memory, highly available relational database. Our implementation strengthens the replication model of HDFS v2, which is based on eventually consistent primary-secondary replication, to one of shared atomic memory, thus simplifying some of HDFS' internal protocols and enabling support for many NameNodes (as opposed to only a primary and secondary NameNode in HDFS v2). Our implementation also maintains the consistency semantics of HDFS, and we validate this by ensuring that all 300+ unit tests for HDFS pass.

This deliverable also describes the platform-as-a-service (PaaS) support we provide for our HDFS implementation. Our HDFS implementation, along with Apache YARN, can be easily installed by unsophisticated users by just pointing and clicking from our portal website to any of the following platforms: Amazon Web Services, OpenStack or a cluster of (bare-metal) hosts. We also provide a Dashboard to administer and monitor the deployed Hadoop cluster.

The document is structured as a userguide for installing and managing a Hadoop platform containing our highly available HDFS distribution, followed by a brief description of the system architecture.

The code is available for download now, although it is still very much beta and under heavy development.

Table of Contents

1. Hop Architecture	1
Highly Available Hadoop Filesystem (HDFS)	1
Leader Election	3
Early performance measurements for Hop HDFS	5
Hop Architecture	5
Deployment model	6
Hop: Hadoop Open Platform-as-a-Service	7
2. Quickstart with Vagrant	10
Pre-requisites:	10
Launching Vagrant	10
3. Hop Web Portal	11
Requirements:	11
Installing the Hop Dashboard	12
4. Hop Dashboard	16
Change Password	16
Edit Graphs	16
Backup/Restore	20
Setup Credentials	20
Cluster Management	21
Clusters Progress	22
Monitoring	23
Hosts	24
Alerts	25
Clusters	26
5. Defining a Cluster	33
Cluster Definition Language	33
Structuring your Cluster:	34
Building your cluster:	35
Cluster in AWS	35
Cluster in OpenStack	36
Cluster on Baremetal Machines	37
Cluster Generator on Dashboard	38
Wrap up	43
6. Launching a Cluster	44
Installation on AWS	44
Pre-requisites:	44
Requirements:	44
Launching the cluster	44
Installation on OpenStack	45
Pre-requisites:	45
Requirements:	45
Launching the cluster	46
Installation on Baremetal Machines	46
Pre-requisites:	47
Requirements:	47
Launching the cluster	47
7. Configuring HDFS	49
HDFS Configuration Parameters not used	49
Additional HDFS Configuration Parameters	49

List of Figures

1.1. Hadoop v2	1
1.2. HDFS v2 NameNode Primary/Secondary Replication Model	2
1.3. Hop HDFS	3
1.4. Reduction in the DB roundtrips by snapshotting metadata at the NameNodes	5
1.5. Hop stack	6
1.6. Deployment Model	7
1.7. Hop PaaS	7
3.1. Hop Portal	11
3.2. Portal AWS	13
3.3. Portal OpenStack	14
3.4. Portal OpenStack	15
4.1. Edit Graphs	16
4.2. Graph Editor	17
4.3. Import Graphs	18
4.4. Graph Selection Detail	19
4.5. Backup/Restore	20
4.6. Setup Credentials	21
4.7. Manage Cluster	22
4.8. Clusters Progress	23
4.9. Hosts	24
4.10. Hosts Details-Services	24
4.11. Hosts Details Graphs	25
4.12. Alerts	25
4.13. Clusters	26
4.14. Cluster Detail	26
4.15. YARN Metrics	27
4.16. Resource Manager Metrics	28
4.17. Node Manager Metrics	29
4.18. Resource Manager UI	29
4.19. Node Manager UI	30
4.20. MySQL overall graphs	30
4.21. MySQL console	31
4.22. HOP console	32
5.1. Select Cluster Type:	39
5.2. Common Cluster Options:	40
5.3. Bare Metal Common Cluster Options:	40
5.4. Cluster Provider Options:	41
5.5. Cluster Group:	42
5.6. Bare Metal Groups:	42
5.7. Confirmation:	43

List of Examples

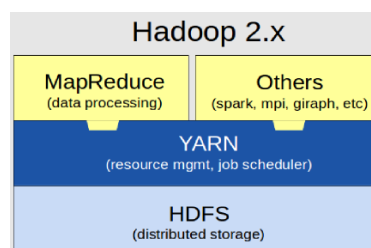
1.1. Example Cluster Definition	8
3.1. HOP Portal	11
5.1. Defining Global Properties	33
5.2. Defining Git repository	34
5.3. Defining Cloud Providers	34
5.4. Full AWS Cluster Example	36
5.5. Full OpenStack Example	37
5.6. Full Baremetal Example	38

Chapter 1. Hop Architecture

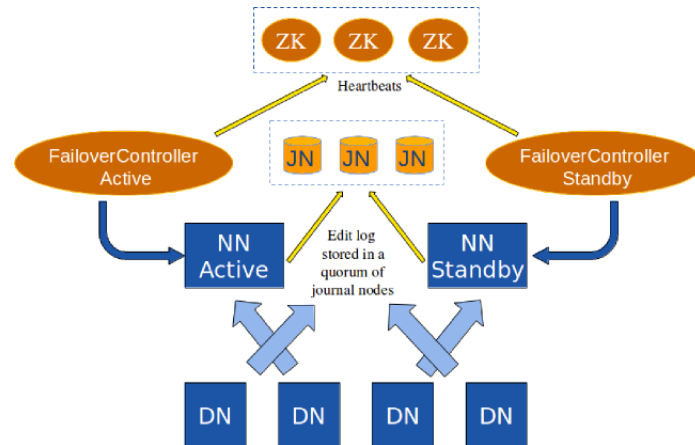
Highly Available Hadoop Filesystem (HDFS)

Due to the rapid growth of data in recent years, distributed file systems have gained widespread adoption. The new breed of distributed file systems reliably store petabytes of data, and also provide rich abstractions for massively parallel data analytics. The Hadoop Distributed File System (HDFS) is a distributed, fault-tolerant file system designed to run on low-cost commodity hardware that scales to store petabytes of data, and is the file storage component of the Hadoop platform. HDFS provides the storage layer for MapReduce, Hive, HBase, Spark and all other YARN applications, see Figure 1.1, “Hadoop v2”.

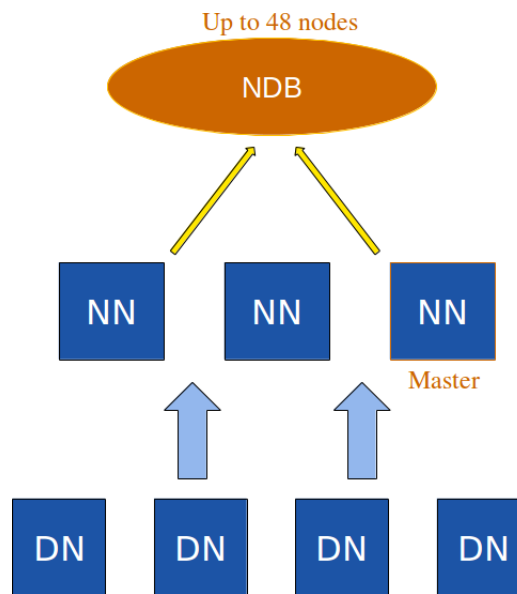
Figure 1.1. Hadoop v2



In 2013, HDFS v2 introduced a new highly available metadata architecture, where, as in HDFS v1, the entire filesystem's metadata is stored in memory on a single node, but in v2 changes to the metadata (*edit log entries*) are now replicated and persisted to a set of (at least three) Journal Nodes using a quorum-based replication algorithm. In HDFS v2, a Primary and Secondary NameNode can be configured, where the Primary NameNode is responsible for managing the metadata, and the Secondary NameNode keeps an eventually consistent copy of the metadata. The Secondary NameNode is kept in sync with the Primary by two mechanisms: firstly, by asynchronously applying all edit log entries that have been committed at the Journal Nodes, and secondly, receiving the same set of heartbeats from Data Nodes that are received by the Primary. The Primary/Secondary replication model is also known as an Active/Standby or Master/Slave replication model, and was popularized by databases in the 1990s. HDFS' implementation of this eventually consistent replication model is more limited than in the traditional relational database world, as all read and write requests are sent to the Primary. In typical Master/Slave configurations, writes are sent to the master, while reads are load-balanced across slaves. The reason all write requests are sent to the Primary to ensure a single consistent copy of the metadata. Read requests are also sent to the Primary, as reads at the Secondary could result in operations being executed on stale metadata. This is a bigger problem for a filesystem, such as HDFS, than it would be for a Web 2.0 social media application with non-critical data, using a Master/Slave database setup. Thus, reads are only sent to the Primary. If the Primary fails, however, the Secondary needs to take over as Primary. Before it can take over, it first has to catch up with the set of edit log entries applied to Primary before it failed. The period of time before all outstanding edit log entries are applied at the Secondary before it can take over may be up to tens of seconds, depending on the current load of the system and the hardware and software setup. Another limitation of the Primary/Secondary model, is that client and Data Nodes from HDFS need to have a consistent view of who the current Primary NameNode is. They do this by asking a Zookeeper coordination service that needs to run on at least 3 nodes to provide a fault tolerant reliable service. Finally, the concurrency model supported by HDFS v2 is still multiple-readers, single-writer.

Figure 1.2. HDFS v2 NameNode Primary/Secondary Replication Model

In contrast, our implementation of HDFS, called Hop HDFS, replaces the Primary-Secondary metadata model with shared, transactional memory, implemented using a distributed, in-memory, shared-nothing database, MySQL Cluster, see figure Figure 1.3, “Hop HDFS”. In our new model, the size of HDFS' metadata is no longer limited to the amount of memory that can be managed on the JVM of a single node. Our solution involves storing the metadata in a replicated, distributed, in-memory database that can scale up to several tens of nodes, all while maintaining the consistency semantics of HDFS. We maintain the consistency of the metadata, while providing high performance, all within a multiple-writer, multiple-reader concurrency model. Multiple concurrent writers are now supported for the filesystem as a whole, but single-writer concurrency is enforced at the inode level. Our solution guarantees freedom from deadlock and progress by logically organizing inodes (and their constituent blocks and replicas) into a hierarchy and having transactions defining on a global order for transactions acquiring both explicit locks and implicit locks on subtrees in the hierarchy. The use of a database, however, also has its drawbacks. As the data now resides on remote hosts on the network, an excessive number of roundtrips to the database harms system scalability and increases per-operation latencies. We ameliorate these problems by introducing a snapshotting mechanism for transactions, where, at the beginning of a transaction, all the resources it needs are acquired in the defined global order, while simulatenously taking row-level locks for those resources. On transaction commit or abort, the resources are freed. This solution enables NameNodes to perform operations on a local copy (or snapshot) of the transactions state until such time as the transaction is completed, thus reducing the number the number of roundtrips to the database, see Figure 1.4, “Reduction in the DB roundtrips by snapshotting metadata at the NameNodes” .

Figure 1.3. Hop HDFS

Leader Election

In HDFS, there are a number of background tasks that are problematic if multiple NameNodes attempt to perform them concurrently. Examples of such tasks include:

1. replication monitoring,
2. lease management,
3. block token generation,
4. and the decommissioning of datanodes.

Without any coordination between the NameNodes, and since all NameNodes have identical behaviour, then, if a block becomes under-replicated, potentially several NameNodes will identify this event, and select a DataNode to replicate that block to. This would cause multiple re-replications of the block, leading it to enter an over-replicated state, upon which, multiple NameNodes would again recognize this state and attempt to remove replicas, leading to an under-replicated block, in a circular manner. We solve this coordination problem, by implementing a Leader Election Algorithm, where only the leader is assigned the task of performing the above background tasks. Our leader election algorithm uses the shared, transactional memory abstraction provided by MySQL Cluster to coordinate the election process. *Definition: correct NameNode process* The notion of correct in this context means that a process is active and running and is able to connect and write to NDB cluster in a bounded time interval.

The bounded time interval is typically set to around 1 second. That means our MySQL Cluster should be provisioned to handle peak loads.

A leader NameNode is a NameNode from the set of NameNodes that is correct and is responsible for listening to DataNode heartbeats and assigning various tasks to them as well as responsible for managing background tasks. The properties that our leader election algorithm guarantee are:

1. *Completeness*: after a bounded time interval, all correct namenodes will detect every namenode that has crashed. *Agreement*: after a bounded time interval, all correct namenodes will recognize one among them as the leader. All will agree to the same namenode being the leader. *Stability*: If one correct namenode is the leader, all previous leaders have crashed

Leader Election Algorithm The leader election algorithm runs continuously at the namenodes as soon as it has started. Each namenode is assigned an (integer) id. At any point in time, the namenode with the lowest id is always elected as the leader. The central focus of the algorithm for detecting failures

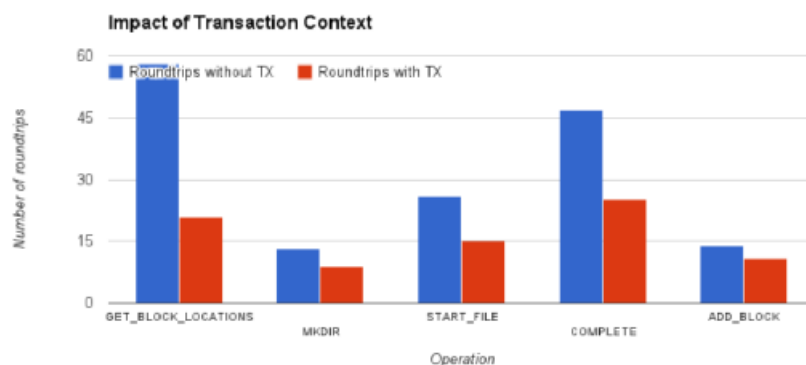
and electing a new leader is through the use of heartbeats called counters. The ids of each namenode are persisted in a table called LEADER and in this way all namenodes have a uniform view of all the namenodes in the system. The schema and sample records for these tables look like the following: Once the namenodes start, they send their heartbeats to NDB to indicate that it is currently active and running. The heartbeat involves incrementing shared counter value in the [COUNTERS] table and updating that value against its record in the [LEADER] table. For the purpose of atomic sequential update, each namenode gets a write lock on the row of [COUNTER] table, increments and updates the value in this table and finally update that value in the [LEADER] table against its row. This will eliminate the possibility of all namenodes to update to the same value during concurrent updates. *Case-I: Scenario if all goes well* If all goes well, in each round of updating the counter we should see the value of the counter as a sequence of numbers iterating in ascending order against all the namenodes. The above figure shows an example of counter values for 3 namenode in the system designated with ids 1, 2 and 3 with their corresponding counter values as 23, 24 and 25 respectively. For the purposes of simplicity in explanation, we call NN_x a namenode with an id of value x. From this example we see that NN1 has the lowest id and is therefore elected as the current leader in the system. *Case-II: Scenario when a leader crashes* However, in a distributed system it is ideal for all to go well. Namenodes can crash or experience network glitches or shutdown. This would prevent it from sending its heartbeat to NDB and thus would be unable to update its value in the [LEADER] table. So if this happens, the counter value for those namenodes would remain the same and would not progress towards incremental updates. Namenodes would experience an irregular sequence of counter values in each of these rows. For example, let's say that NN1 has crashed and the current counter value is now 25. This would allow NN2 and NN3 to progress in updating the counter with values 26 and 27 while counter value for NN1 is still at value 23. Following would be the snapshot of the view on the [LEADER] table at this round: The figure shows an irregular sequence of counter values i.e. 23, 26 and 27 when we should expect 26, 27 and 28. This means something has gone wrong. *Determining the leader* The next step is to detect this irregular sequence and decide if a new leader is to be elected. If so, the leader election will run again and a new namenode would be elected as the leader. The basic idea to determine the leader is to determine which namenodes are strongly aligned to the current counter value. In this case, the current counter after the updates from NN2 and NN3 is 27. Hence, if we have 3 namenodes in the system, we expect the counter sequence to range from 27] where NN1 would have value 25, NN2 would have value 26 and NN3 would have value 27. So we need to get all namenodes that have a counter with value greater than 24. From the list of namenodes returned, the one with the lowest id would be elected the leader. This logic is implemented in the algorithm and runs periodically during a configured time interval. So in this case, the list of namenodes returned includes NN2 and NN3. Out of these two, NN2 has the lowest and would elect itself as the leader. Accordingly NN3 would notice that it does not have the lowest id and will accept NN2 as the current leader. Ensuring single leader dominance Once a namenode determines that it is the leader, the first thing is to ensure that there are no other leaders in the system. This would mean that all previously elected leaders (or namenodes with lower id than itself) have crashed. To ensure this, it simply enforces this rule by removing all records from the [LEADER] table that have a lower id than itself. *Correctness* Following are the cases that would prove correctness of the algorithm *Case-I: Scenario when current leader crashes* Supposing we have three namenodes, each have the current counter as [NN1=6, NN2=7, NN3=8]. We assume NN1 has crashed and therefore will not update its counter. In the next round, NN2 will update the counter to 9 and later NN3 would update the counter to 10 and we would have the following state information for NNs to agree on [NN1=6, NN2=9, NN3=10]. Let's say that NN1 crashes as soon as it updated its counter to 6. In the same round, NN2 updates its counter to 7 with a counter range of [5-7] and NN3 would update its counter to 8 with a counter range of [6-8]. At this point, from both namenode perspectives, NN1 is alive as its counter lies in the range while the datanodes detect failure of NN1 via the TCP socket. This is illustrated in Figure6. Now, in the next round, NN2 would progress to updating the counter to 9 and now the counter range would be [7-9]. Thus at this round, it would detect the crash of NN1 because the counter pertaining to NN1 is still 6 and does not lie in this range. Similarly, when NN3 updates the counter to 10 it detects the crash of NN1 because the range of counter values is [8-10]. This means that at this point, both NN2 and NN3 recognizes the crash of NN1 and run the consensus to elect the leader. NN2 having the lowest id would be elected as the leader. Thus [Property#1] holds. Now, when NN2 is elected the leader, it removes all lower NN ids in comparison to its own id from the table. So now it has the lowest id and all NNs will see this view. This forces [Property#2] where all previous leaders have now crashed. This is illustrated above in Figure7. If the previous leader recovers, it notices that its id is not in the table as the lowest id and will place its record with the next highest id. DNs recognizing

the new leader All datanodes are up-to-date with the current view of namenodes in the system. This is done via requesting the namenodes for the current list of namenodes. This is done via a simple RPC call. The datanodes get the list of namenodes and assume the namenode with the lowest id is the leader. When NN1 had crashed, the DNs keep retrying for some amount of time and if they are not successful at making contact with the NN it will remove it from the list and select the next NN who potentially would be the leader. This would achieve [Property#2]. There can be two possibilities where (a) the datanode contacts the next NN and if it is actually the leader then the process flows normally. (b) The datanode contacts the next NN in the list but who may not be the leader (because it is possible that it has crashed) but this NN would then provide the updated list of namenodes and then it would recognize the new leader. Eventually all correct DNs would recognize some correct NN as the leader thereby fulfilling [Property#1]. Case-II: Scenario when current leader thinks it is alive but cannot connect to NDB cluster If a namenode process say NN1 is active and running and thinks that it has not crashed BUT cannot make contact with NDB, then such a namenode process is not considered correct as per definition. In this scenario, all datanodes would always think that NN1 is the current leader as it can make contact with that NN. But since the NN cannot make contact with NDB, it would kill itself and shutdown hoping that some other namenode would eventually be elected the leader. When NN1 will be shutdown, the datanodes will keep retrying for some amount of time after some point where they will switch to the next namenode in the list and will determine the next NN leader. This also ensures no corruption of data where the namenode thinks it is still the leader as it cannot update the metadata in the NDB since its communication with it is lost.

Early performance measurements for Hop HDFS

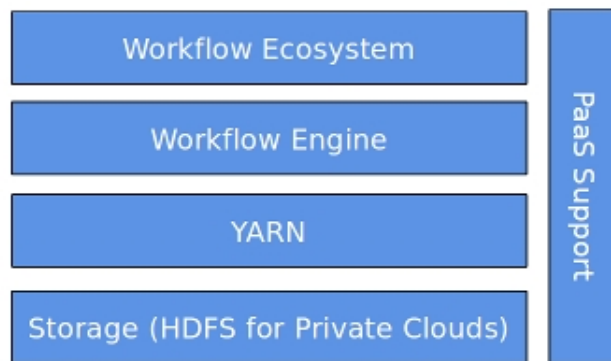
Our early performance figures show that Hop HDFS can scale to handle a similar number of read and write requests per unit time as Apache HDFS. We have introduced a number of features to enable this high level of performance, including a snapshot layer at NameNodes and row-level locking at the database level, rather than a system level lock for update operations as is done in Apache HDFS. Our snapshotting layer involves a transaction acquiring all resources it requires at the start of a primitive filesystem operation, and performing local read/write operations on the snapshot copy, and then finally committing or rolling back on transaction commit.

Figure 1.4. Reduction in the DB roundtrips by snapshotting metadata at the NameNodes



Hop Architecture

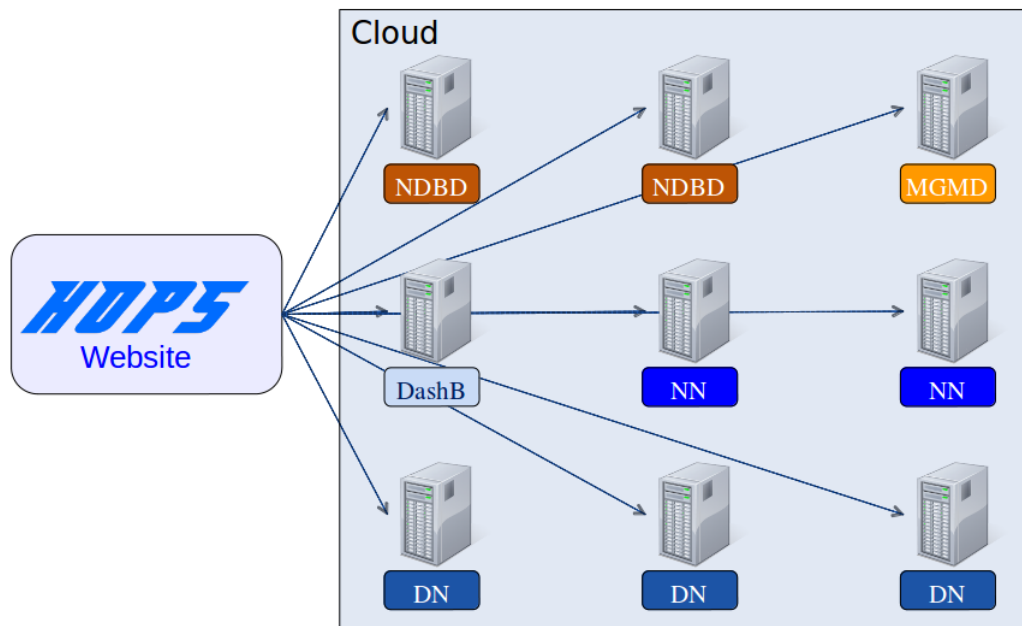
Hops, as a cloud platform for distributed processing and big data, is made up of latest Hadoop ecosystem. As you can see in Figure 1.5, “Hop stack” there are three major layers in our stack, HDFS, YARN and Workflow. Cross-layer aspects like Security and PaaS services are also included.

Figure 1.5. Hop stack

1. *Hop File System (Hop-FS)* At the bottom layer of big data stack is Hop-FS a distributed file system based on Hadoop Distributed File System (HDFS). We revisited relational representation of metadata to remove limitation of single metadata server and single point of failure. Our File System solution can scale up to several tens of nodes, while maintaining the consistency semantics for the filesystem. We store the metadata on a shared-nothing, in-memory, partitioned database by maintaining the consistency of the metadata, while providing high performance. Hop-FS also guarantees freedom from deadlock and progress.
2. *MySQL Cluster* MySQL Cluster is a highly scalable, real-time, ACID-complaint transactional database. Designed around a distributed, multi-master architecture with no single point of failure; MySQL Cluster's real-time design delivers predictable, milisecond response times with the ability to service millions of operations per second. In the case of our data platform, it is used to handle and manage the state of our multi-namenode solution of our architecture.
3. *YARN* Resource negotiator for managing high volume distributed data processing tasks against HDFS. It supports different processing models other that Map-Reduce by separating its Resource Manager from Scheduler and Application Master. Application Master gives us flexibility to accommodate heterogeneous processes by implementing a wrapper for each kind of application so it could manage any kind of processing resources that is defined for it. This allows user to process data intensive task like MapReduce jobs or in our case our future support for bioinformatic workflow tasks engine which will make use of YARN to handle and negotiate the scheduling of this type of jobs.
4. *Workflow Engine* On top of YARN, HopS workflow engine parses workflows into an execution model of arbitrary tasks. For each task, it asks YARN for a container, then for each container allocated task based on the scheduling policy it stages in data into HopSFS, launches the task and stages out the result back to HopSFS

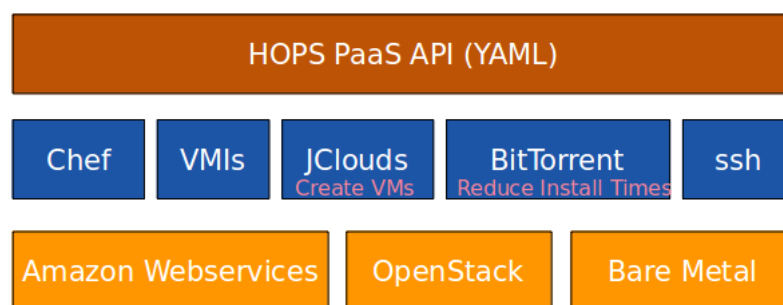
Deployment model

At the moment Hop supports Amazon Cloud, Open Stack and Bare Metal. Based on the chosen cloud provider, as it can be seen in Figure 1.6, "Deployment Model" our deployment model consist of Hop-Dashboard plus other machines either virtual in cloud or bare metal. Dashboard is the point of administration with web access through which customer could define configuration of the cluster, machines are allocated, their software stack is installed and state of the cluster is monitored. Cloud machines could be associated into security node-groups, machines inside each node-group basically have the same security credentials and could communicate with each other; however, communication between machnies from different security group is not possible. All the machins inside the cluster have the same infrastructure and basic stack of softwares, althoug based on the services each machine shoul provide, arbitrary platform softwares are installed.

Figure 1.6. Deployment Model

Hop: Hadoop Open Platform-as-a-Service

We have a prototype implementation of a Platform-as-a-Service (PaaS) framework for HDFS. As our PaaS model will cover the whole Hadoop stack, and not just HDFS, we call it Hadoop Open Platform-as-a-Service, or *Hop* for short. Our framework is *open*, as it is designed to be deployable on any cloud platform or a bare-metal environment. Currently, we support Amazon Web Services and OpenStack, as well as for bare-metal hosts. Hop consists of a set of frameworks and libraries that we use to deploy a Hadoop cluster on both cloud and bare-metal clusters, with the most significant technologies being Chef and JClouds, but also including a YAML-compatible language for defining a cluster and BitTorrent for improving download speeds of programs to hosts in a cluster.

Figure 1.7. Hop PaaS**Hop PaaS Stack**

1. **YAML** (YAML Ain't Markup Language) is markup language which takes concepts from programming languages such as C, Perl and Python, and ideas from XML. We use YAML to define the set of hosts and the services that will be installed on those hosts, making up a single cluster. This way, we can define a whole cluster in a single file, enabling easier management of clusters and even the sharing of cluster definitions. YAML syntax allows easy mappings of common data types found in high level languages like list, associative arrays and scalar. It makes it suitable for tasks where humans are likely to view or edit data structures, such as configuration files or in our case, cluster definition files. Additionally, we make use of the open source parser SnakeYAML to parse the contents of our cluster definition files. The SnakeYAML parser transforms the given cluster

definition into consecutive stages such as defining security groups, virtual machine allocation, bittorrent, installation, validation and retry. An example of a simple cluster definition is given in Example 1.1, “Example Cluster Definition”. The YAML file defines a cluster consisting of 6 nodes, with MySQL Cluster running exclusively on 2 nodes, Hadoop running exclusively on 3 nodes, and another node running both MySQL Cluster and Hadoop services. The example uses default values for the AWS image, instanceType and region. There are many other parameters that can be overridden. Our services map directly onto chef recipes for installing the services. We are developing a model for explicitly handling dependencies in chef, so that dependent services such as Java don't need to be specified as requirements in this cluster definition file.

Example 1.1. Example Cluster Definition

```
name: simpleCluster
provider:
  name: aws-ec2

nodes:
  - service:
    - ndb::ndbd
    number: 2

  - service:
    - ndb::mgm, ndb::mysqld, hadoop::namenode
    number: 1

  - service:
    - hadoop::namenode
    - hadoop::resourcemanager
    number: 1

  - service:
    - hadoop::datanode
    - hadoop::nodemanager
    number: 2
```

2. *Apache JClouds* Apache JClouds is an open source multi-cloud api interface which allows us to write reusable code for creating, destroying, and bootstrapping virtual machines (VMs) on different cloud providers. The same code can be configured to interact with Amazon, OpenStack, Azure, and Rackspace VMs, and 26 other cloud providers. Through JClouds simple Java interface, we can deploy and port Hop to different cloud environments. Hop parses cluster definition files, producing code that executes JCloud API calls to create, destroy and bootstrap VMs.
3. *Chef* Chef is a systems infrastructure and configuration framework that automates the deployment of servers and applications to any physical, virtual or cloud location. JClouds is used to bootstrap chef on VMs or physical hosts, and once chef is installed on a host, we can run *chef recipes* using the chef-client deployed on host to install software on that host. The chef-client relies in a series of abstract definitions (defined as cookbooks and recopes) which are managed in Ruby and are treated like source code. With each definition, we describe how a specific part should be built and managed, which then; the chef-client applies these definitions to deploy and configure servers and applications as specified. In most of the cases, it is simple enough to let chef-client know which cookbooks and recipes it needs to apply.
4. *BitTorrent* After machines are allocated in cloud, with the metadata information that JCloud returns, dashboard tries to open a ssh connection into every single machine and install Chef agent for installations. Before installation starts, software libraries is replicated in all machines from dashboard, though the process could overflow the bandwidth to dashboard if all machines try to download from dashboard. To handle this situation HopS run a bittorrent in which dashboard

machine is the seeder, then all machines could contribute to download process which is both faster and anti-bottleneck. After download Chef agent starts installation based on the required packages in each machine and with the order of dependencies between packages.

References

[AhoSethiUllman96] Alfred V. Aho, Ravi Sethi, and Jeffrey D. Ullman. Copyright © 1996 Bell Telephone Laboratories, Inc.. James T. DeWolf. 0-201-10088-6. Addison-Wesley Publishing Company. *Compilers, Principles, Techniques, and Tools*.

Chapter 2. Quickstart with Vagrant

This section describes the steps required to deploy a HOP cluster on a single machine using git, vagrant¹, and chef².

Pre-requisites:

You should have the following programs installed: git and vagrant. You will also need to download the vagrant virtual machine image for Ubuntu 12.04 "precise".

```
apt-get install git-core vagrant
vagrant box add "precise64" http://files.vagrantup.com/precise64.box
```

Launching Vagrant

You are ready to clone the chef recipes, and launch a vagrant instance.

```
git clone https://github.com/hopstart/hop-chef.git
cd hop-chef
vagrant up
```

Now grab a coffee, assuming you have a good network connection, it will take around 15 minutes to provision a vagrant instance. When vagrant successfully completes provisioning using chef, use the following URL and default user credentials to access the Hop Dashboard:

```
https://localhost:9191/hops-dashboard/
user: admin
password: admin
```

You can log into the VM and then get root access using:

```
vagrant ssh
sudo su
```

If needed, you can configure the glassfish webserver here:

```
https://localhost:5858
user: admin
password: admin
```

You can now jump to Chapter 4, *Hop Dashboard*

¹ Vagrant is a tool for building complete development environments. With an easy-to-use workflow and focus on automation, Vagrant lowers development environment setup time, increases development/production parity, and makes the "works on my machine" excuse a relic of the past. [http://www.vagrantup.com]

² Chef is a systems and cloud infrastructure automation framework that makes it easy to deploy servers and applications to any physical, virtual, or cloud location, no matter the size of the infrastructure. [http://docs.opscode.com/]

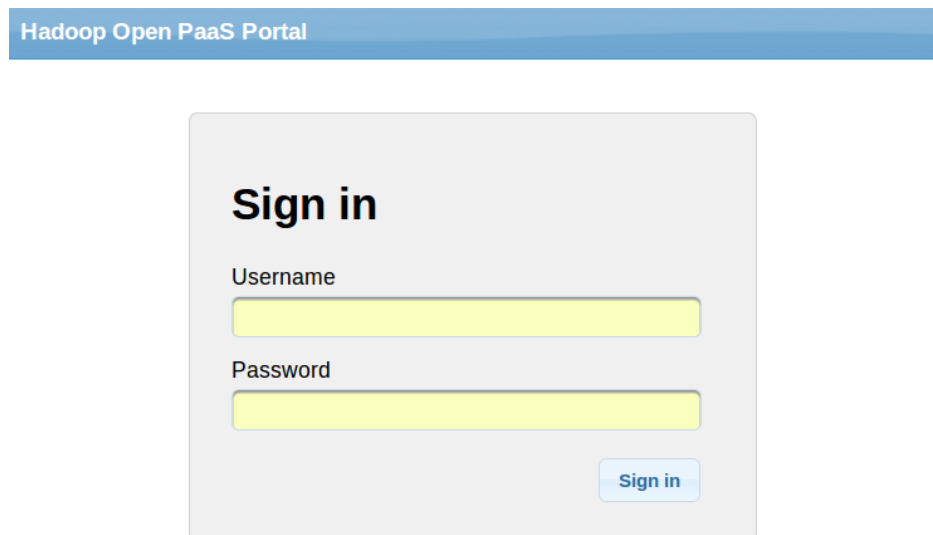
Chapter 3. Hop Web Portal

Hop is an effort to offer a high performance next generation hadoop platform focusing on high scalability, availability and reliability. This section describes the information needed to jump start Hop cluster. Hop web portal, is an entry point where users can test our platform in one of the supported environments: AWS, OpenStack and Baremetal machines. It is a simple web application that deploys the Hop Dashboard in the cloud. Using the Hop dashboard you can provision and configure the rest of the cluster. Hop Dashboard provides a centralized end-to-end management and monitoring application for Hop distribution. Following are the step by step instructions for setting up the cluster in different environments.

Example 3.1. HOP Portal

`https://snurran.sics.se:8181/hop-portal`

Figure 3.1. Hop Portal



Hadoop Open PaaS Portal

Sign in

Username

Password

Sign in

Requirements:

Hop Dashboard has to be installed on a machine in a cluster that has connectivity both to cloud infrastructure and the cluster. Using Hop Web Portal you can provision a cluster and install the dashboard to any of the following platforms: AWS, OpenStack and cluster of bare-metal hosts. The following environments are currently fully supported:

- *Amazon Web Services:* In order to use of our platform in AWS cloud Infrastructure, it is necessary to provide EC2 account credentials. When deploying our Dashboard in EC2, we recommend minimum Instance type should be of m1.large type Instance with Ubuntu.
- *OpenStack* For using Hop on OpenStack, it is necessary that you provide the credentials and configuration parameters to connect to you OpenStack end-point. The recommended instance to use in OpenStack infrastructure should be equivalent to a m1.large instance type or greater.
- *Baremetal Physical Machine:* For deployment on Baremetal machines our system requires security credentials of a user with sudo access to the machines to deploy the software through SSH .
- *Vagrant:* With our vagrant distribution, you can deploy the whole cluster locally in a VM machine. It is ideal for testing the platform before deployment in a production environment.

All the environments are tested with instances running Ubuntu 12.04.

Installing the Hop Dashboard

Intalling Hop Dashboard through the Portal is quite simple and takes a couple of minutes to deploy. Here you may find instructions on how to deploy the dashboard on AWS, OpenStack and Baremetal Machine:

1. *Amazon EC2*: For installing the dashboard in Amazon EC2 follow these simple steps:

- Login into the HOP Portal with your user name and password.
- Select from the providers option in the maintoolbar, the Amazon EC2 option. Enter the following information in the new form
 - Dashboard credentials: admin's username and password in order to access your newly created dashboard.
 - EC2 credentials which include the Access Key id from you AWS account with its related Secret key.
 - Instance configuration parameters used to deploy a virtual machine in AWS.
 - a. Security group where the machine will be deployed. If it does not exist, then a new security group will be created automatically.
 - b. The hardware ID of the instance type we want to use from Amazon EC2. For example, m1.small, t1.micro. The recommended instance type is m1.large.
 - c. Image ID which includes the region of that image and the *ami id* tag. We only support Ubuntu based images.
 - d. Location ID of the region you want to deploy the dashboard.
 - e. Selecting the option to authorize the public key will open a new option dialog box were you can insert your public key. By default we generate random key pairs for the machines through EC2 key pair service, and it is not possible to access the machines internally without this option.
 - f. Selecting the override login user, will override the default user for Ubuntu AMI images with the login user of your choice. This is necessary if you use custom Ubuntu images which are not one of the Ubuntu images that canonical offer in AWS by default.

Figure 3.2. Portal AWS

- After filling up the form, press the Launch Instance button. The whole process takes 10-15 minutes. After the deployment you will receive a notification showing the URL of the newly deployed Hop Dashboard. To login the dashboard, use the credentials you specified previously in the web portal.
2. *OpenStack*: For deployment in OpenStack cloud follow these simple steps. Note, this is in alpha state:
- Login into the Hop Portal with your user name and password.
 - Select from the providers the OpenStack option. A new form will be generated.
 - Dashboard credentials: here you specify the admin username and password for the new dashboard.
 - OpenStack credentials: the user name and password to access the OpenStack project. The username should be a concatenation of the OpenStack project name and the user for that project. For example "projectName:user". Also you should indicate the url of your OpenStack Nova end-point in order to send the requests to your OpenStack infrastructure.
 - Configuration parameters that are used to deploy a virtual machine in OpenStack:
 - a. Security group where the machine will be deployed. If it does not exist, we will automatically create a security group and open the ports needed for the application.
 - b. The hardware ID of the instance type we want to commission in OpenStack cloud. This is a number which corresponds to the type of instance you want to deploy and is supported by your OpenStack infrastructure. We recommended using a configuration similar to a m1.large in EC2.
 - c. An Image ID image located in the openstack project.

- d. Location ID identifies the dashboard in the OpenStack cluster.
- e. Selecting this option to authorize the public key based access. It will open a new dialog box where you can insert your desired public key. By default we generate random key pairs for the machine through OpenStack key pair service, and it is not possible to access the machine internally without selecting this option.
- f. Selecting the override login user: This is necessary for OpenStack if you are using a custom Ubuntu image.

Figure 3.3. Portal OpenStack

- After filling up the form, press the Launch Instance button. The whole process takes 10-15 minutes. After the deployment you will receive a notification showing the URL of the newly deployed Hop Dashboard. To login the dashboard, use the credentials you specified previously in the web portal

IP pools in OpenStack

It is necessary that you have allocated at least 1 public IP to the project. During the deployment phase the portal will query the OpenStack project and link the public ip to the VM.

3. *Baremetal Physical Machine:* For deploying the dashboard on a BareMetal hosts cluster follow these simple steps:
 - Login into the Hop Portal with your user name and password.
 - In the new page, select BareMetal from the providers list. A form will be generated where you need to fill in the following
 - Dashboard credentials: here you specify the admin username and password for the new dashboard.

- SSH credentials: includes the host address of the machine we want to connect to and the private key.
- Extra parameters that we might need:
 - a. Selecting the option to authorize the public key, it will open a new option where you can insert your desired public key to allow extra access to the machine.
 - b. Selecting the override login user will rename the sudo user to be used to deploy the dashboard on the machine.

Figure 3.4. Portal OpenStack



The screenshot shows the HOPS (Hadoop Open Platform 5) web portal interface. At the top left is the HOPS logo with the text 'HADOOP OPEN PLATFORM 5'. To its right is a small yellow cartoon elephant. Below the logo, a blue bar contains the text 'Select Provider: Baremetal' and a user profile icon with the email 'jdowling@sics'. The main content area is titled 'Create Dashboard' and contains three sections: 'Dashboard Credentials' with fields for 'Provider' (set to 'Baremetal'), 'Username *' (placeholder: 'Username for dashboard user'), and 'Password *' (placeholder: 'Password for dashboard user'); 'SSH Credentials' with a 'host *' field (placeholder: 'Machine IP or Hostname') and a large 'Private Key *' text area; and 'Instance Parameters' with two checkboxes: 'Enable Authorize Public Key: *' and 'Override Login User:'. A 'Launch Instance' button is at the bottom of the form.

- After filling up the form, press the Launch Instance button. The whole process takes 10-15 minutes. After the deployment you will receive a notification showing the URL of the newly deployed Hop Dashboard. To login the dashboard, use the credentials you specified previously in the web portal.

Chapter 4. Hop Dashboard

Hop Dashboard is a web application designed with Hadoop administrators in mind. Hop Dashboard provides a centralized end-to-end management and monitoring application for HOP distribution. It simplifies Hadoop administrator's job by having all the necessary statistics of the Hadoop cluster gathered in one place and presented in an effective manner. Also it is possible to execute maintenance commands through the different terminals available for services like MySQL, HDFS or Spark.

You will need to configure certain parameters when the Hop Dashboard is launched for the first time. In order to access the dashboard, use the credentials you specified in the Hop Portal when you were preparing to launch the dashboard Chapter 3, *Hop Web Portal*. After login, if you press your user icon you will have a menu with the following options:

- Change password
- Edit Graphs
- Backup/Restore
- Setup Credentials

We will go through these options in more detail in the following sections.

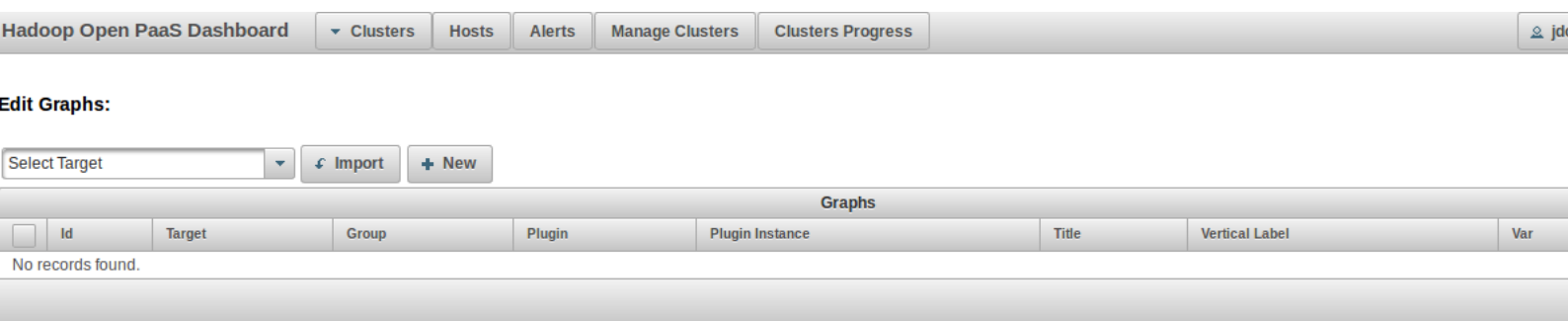
Change Password

This is currently under development.

Edit Graphs

Hop Dashboard offers customization of how the data is displayed that is retrieved from the Hop agents located in each node of the cluster. Selecting this option will enable you to define the graphs for the different services of Hop.

Figure 4.1. Edit Graphs



You can define your graphs through two different options that are visible on this new view:

- *New* Selecting this option will open a new dialog where you can define the specifications for new graph for statistics monitored by the dashboard.

Figure 4.2. Graph Editor

Hadoop Open PaaS Dashboard

Clusters Hosts Alerts Manage Clusters Clusters Progress

Edit Graphs:

Select Target

No records found.

New Graph

Graph Id:

Target:

Group:

Plugin:

Plugin-Instance:

Title:

Vertical Label:

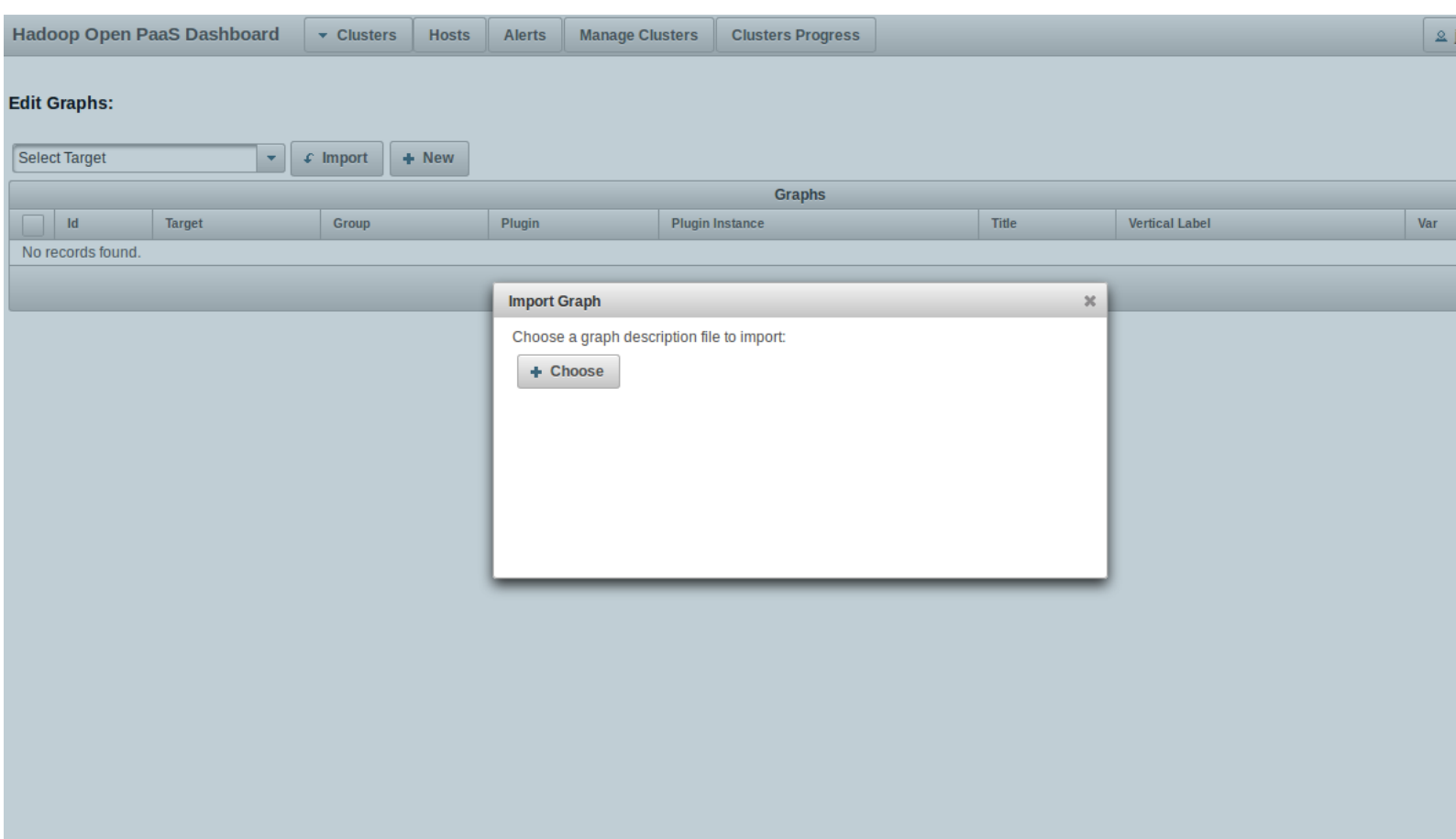
Var:

Charts

Model	Type	Type Instance	Data Set	Label	Color	Format	RRD File
No records found.							

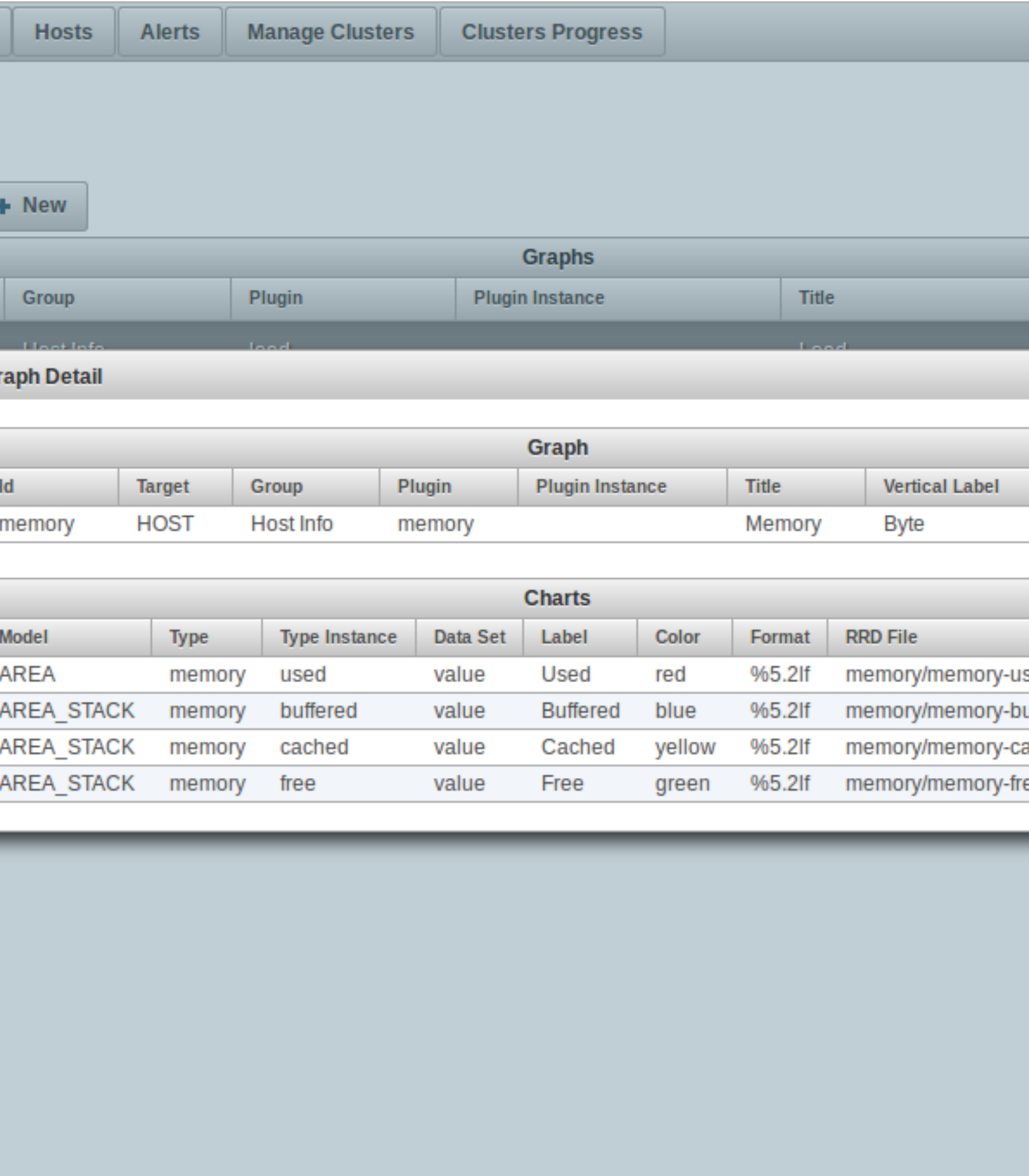
+ New Chart Save

- *Import* It is also possible to import the graphs specifications from a JSON graph file. This is the quickest way if you want to load a large number of graphs.

Figure 4.3. Import Graphs

When graphs are generated, they are stored in the database and loaded in the dashboard. In the graph table, it is possible to view all the graphs for a specific type of service or component that the dashboard is monitoring. Pressing the zoom button will generate a dialog box where you can see more details about the graph.

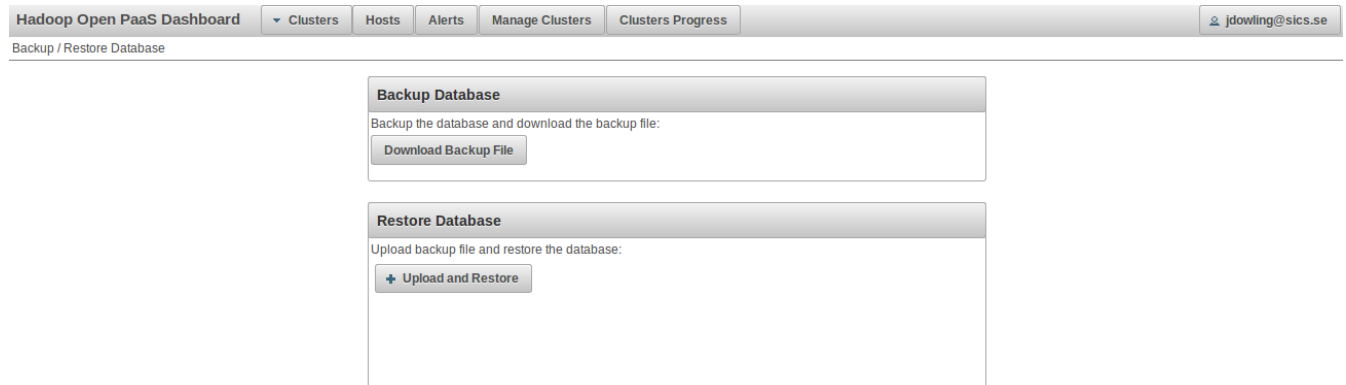
Figure 4.4. Graph Selection Detail



Backup/Restore

It takes backup of the whole dashboard state stored in the database. Incase of failure or maintenance the dashboard can restore the previous states from a backup file obtained from the dashboard.

Figure 4.5. Backup/Restore



Setup Credentials

It allows the user to setup the credentials that are used when the dashboard deploys a cluster. These are important configuration parameters. Cluster deployment will fail if these parameters are not properly set.

Figure 4.6. Setup Credentials

Setup Credentials

Setup Credentials:

Select Providers

Amazon EC2: ☒ Enabled

Openstack: ☒ Enabled

Baremetal: ☒ Enabled

Dashboard Management Parameters

Dashboard IP:

Public Key:

Private Key:

EC2 Credentials

Id:

Secret Key:

Openstack Credentials

id:

Secret Key:

Keystone url:

Cluster Management

Cluster mangement is an interesting feature in the Hop Dashboard. The dashboad keeps track of diffenent cluster applications deployed in the cloud. It allows the administrators to create, delete, edit, load and export clusters.

Figure 4.7. Manage Cluster**Cluster Management:**

You can load from a cluster definition file No file selected.

You can also select a cluster from the database

Available cluster configurations		
Cluster Name	Cluster Type	Content
test2	virtualized	prod,aws-ec2,eu-west-1
test	virtualized	dev,aws-ec2,eu-west-1
<input type="button" value="+ Create Cluster"/> <input type="button" value="- Delete clusters"/> <input type="button" value="Edit selected cluster"/> <input type="button" value="Load selected cluster"/> <input type="button" value="Export selected cluster"/>		

- *Create Cluster* Selecting this option will take you to the cluster generation wizard where a users can of generate their own clusters for Hop. In the next chapter, we will explain in detail how users can create their custom clusters using our Cluster Definition Language, see Chapter 5, *Defining a Cluster*.
- *Delete Clusters* This option will delete a selected cluster. It is possible to delete multiple clusters at once by shift clicking multiple clusters before selecting this option.
- *Edit Selected Cluster* This option will allow a user to edit an existing stored cluster. This will bring the cluster generation wizard with the values of the selected cluster. Administrators can use this wizard to modify the cluster.
- *Load Selected Cluster* This option allows the user to load one of the stored clusters into the cluster launcher. Cluster launcher is used to deploy the cluster in the cloud.
- *Export Cluster* It saves the information about the cluster in a YAML file.

Clusters Progress

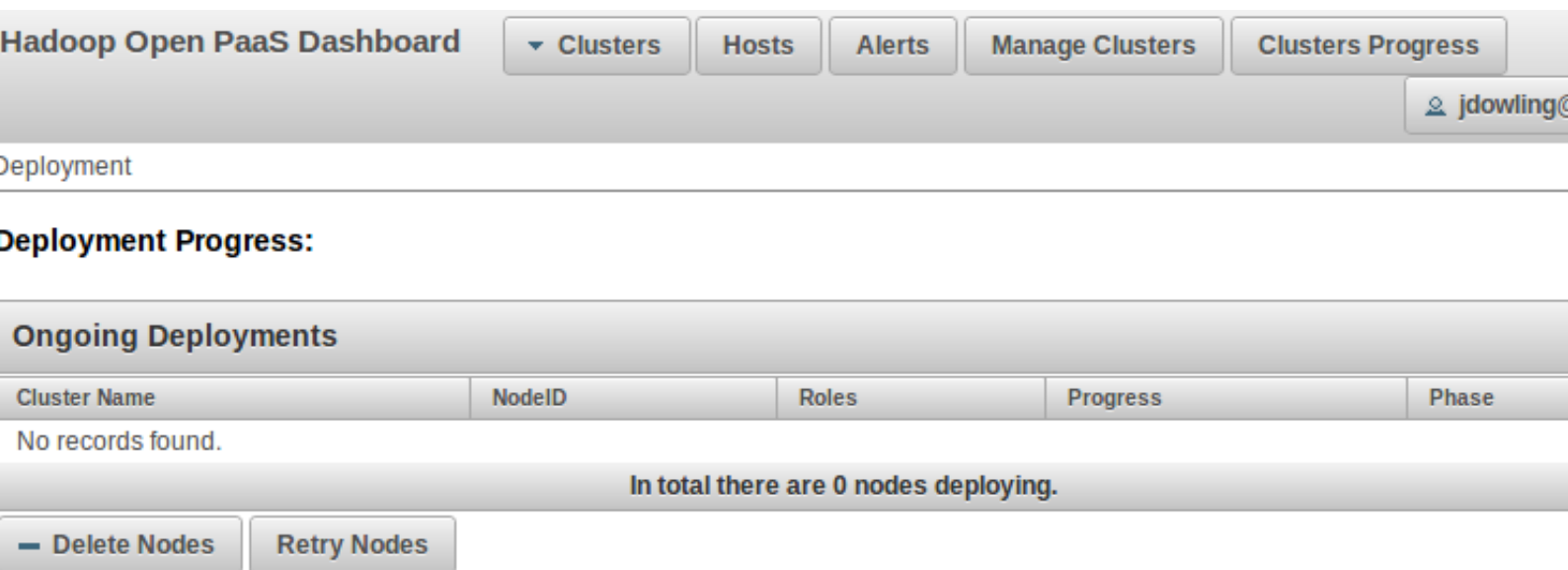
Another interesting feature for Hadoop administrators, is the option available in the main Dashboard toolbar for tracking the progress of the nodes deployed for your multiple clusters. Selecting this option will bring you a view where you will have all the history of all the nodes through multiple phases of the deployment cycle. A progress bar will appear over each of the entries in the table to see the deployment progress of the nodes and an information tag with the current phase. Here you will be notified of succesfully configured nodes or nodes that had an error during the deployment phase of your cluster which will require of special maintenance. A node entry will go through the following phases during a cluster deployment.

- *Waiting* A node in this phase means, that; the entry has been generated in the node scheduler but no node creation query has been generated to the cloud provider. It will wait until the query is submitted against the cloud provider.
- *Creation* A node in this phase means, that; the scheduler has submitted the query against the cloud provider and it is waiting for the cloud provider to finish deploying the virtual instance. After successfully getting back the information of the specific instance, an initialization script is executed to do a preliminary configuration on that node.
- *Install* A node in this phase means, that; the node is already created and we are running the install script which will execute chef install recipes for Hop services to fetch the necessary binaries from

the different repositories. This phase is optional in case of using prebuilt virtual instances which contain Hop binaries inside.

- *Configure* A node in this phase means, that; the node is receiving the configuration script which will execute chef with the selected recipes for the services defined for that node.
- *Complete* A node in this phase means, that; the node has successfully finished executing the configuration script with chef and it is now working as part of the cluster. Additionally, the nodes in this phase will appear green in the progress history.
- *Retrying* A node in this phase means, that; the deployment system has detected a problem during a previous node phase and it is triggering the retrying mechanism. It will retry submitting the previous phase script for 5 retries. This process ends when the number of retries finish or we manage to successfully execute the script.
- *Error* A node in this phase means, that; the deployment system failed to recover the node through the retrying mechanism. This will be shown as a Red entry in the progress table and further actions will be needed by the Hadoop administrator in order to recover that node, for example; SSH to the failed node in order to get more information of the type of error it got.

Figure 4.8. Clusters Progress



From the previous figure, we get an overview of the cluster progress table. In here, the user is capable of executing the following actions:

- *Delete Nodes* A user can select multiple node entries by shift click and later delete their progress history by executing this command. Note that this option will delete node entries which are registered with a node complete phase.
- *Retry Nodes* A user can do make use of this option in order to execute further retry procedures on nodes that failed to deploy correctly. This will make the deployment system to execute a recovery script on the selected nodes in an error state. For now, this script simply tries to rerun the nodes configuration script by executing only chef in that node. It is possible to retry multiple nodes at once by selecting multiple entries by shift click.

Monitoring

The Hop Dashboard offers multiple ways in how you can monitor the state of your current Hop clusters. The following options related to monitoring features can be found in the main Dashboard toolbar

- *Hosts* Information and data analytics of all the nodes this dashboard is monitoring.
- *Alerts* Information of possible alerts the Hop agents will be sending the dashboard in order to notify the current state of the nodes.
- *Clusters* A dropdown list with the available clusters been monitored by the dashboard. Selecting an entry will give further information of the state of that cluster.

Hosts

This view allows a user to get a general overview of the state of all the nodes in all the clusters, you will be able to track information of great interest like the allocated ip's for that node, its hostname, host ID, its current health in the system, and when the last heartbeat was received. Also it shows information about the nodes available resources like the number of cores that machine has, the load average in that instance in the last 1, 5 15 min, disk usage and physical memory in use.

Figure 4.9. Hosts

Hadoop Open PaaS Dashboard

▼ Clusters

Hosts

Alerts

Manage Clusters

Clusters Progress

Hosts

1 Host Under Management:

Hosts

Host Id	Hostname	Public IP	Private IP	Health	Last Heartbeat	Cores	Load Average			Disk Usage	Physical Me
10	mgmd49		10.0.2.15	Good	4.9s ago	2	0.92	1.15	0.85	<div><div>10 GB / 78.9 GB</div></div>	<div><div>2.4 C</div></div>

By clicking in one of the host ID's entries, it will show a detailed view of the hosts monitoring analytics with the graphs that are available for the Hosts components. See the previous Edit Graphs section, where we explained how users can create graphs for the dashboards monitoring components.

Figure 4.10. Hosts Details-Services

Hadoop Open PaaS Dashboard

▼ Clusters

Hosts

Alerts

Manage Clusters

Clusters Progress

Hosts »

10 (mgmd49)

Status

Host Details

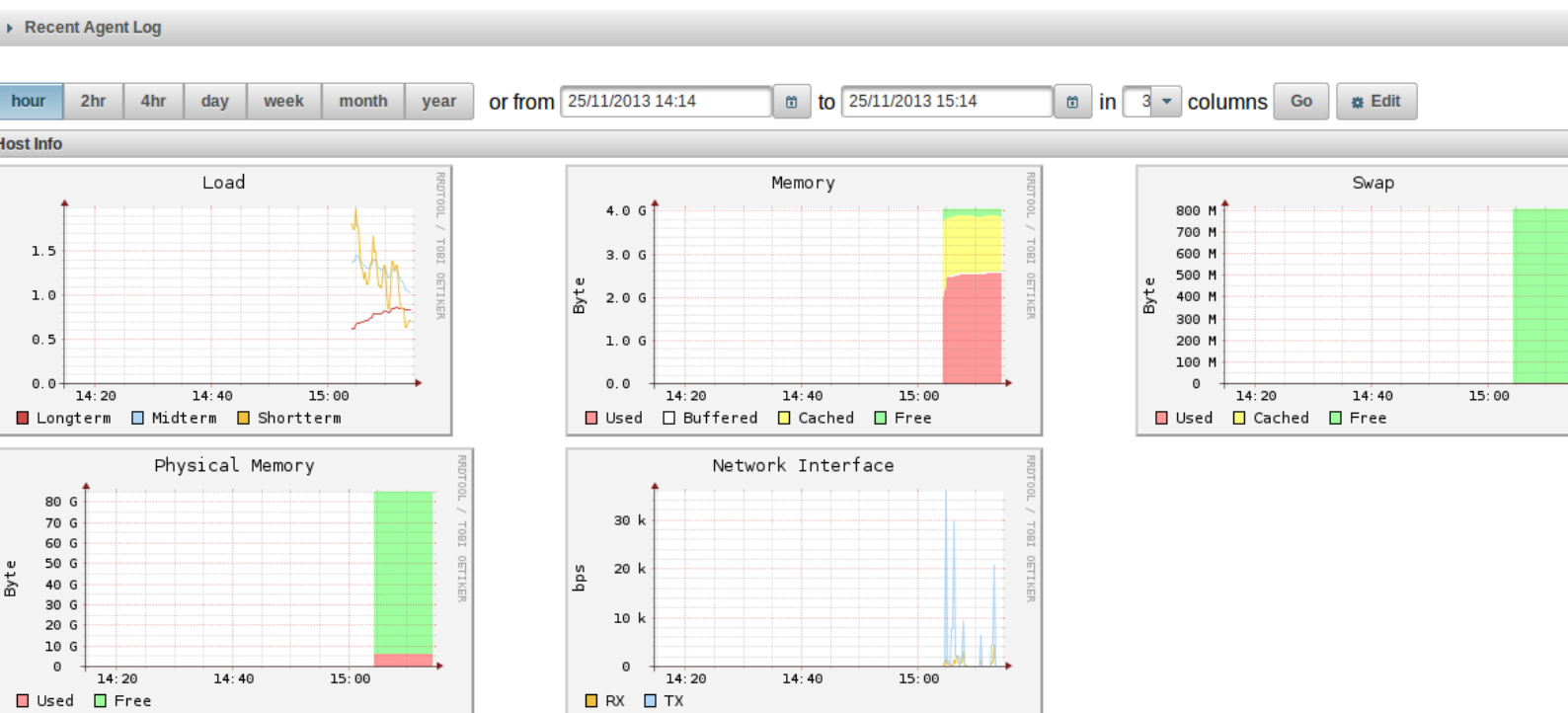
Host Id	Hostname	Public IP	Private IP	Health	Last Heartbeat	Cores	Load Average			Disk Usage	Physical M
10	mgmd49		10.0.2.15	Good	1.3s ago	2	0.94	1.07	0.85	<div>10 GB / 78.9 GB</div>	<div>2.4 C</div>

Roles

Cluster	Service	Role	Role Page
vagrant	HOPS	datanode	vagrant/HOPS/datanode @ 10
vagrant	HOPS	namenode	vagrant/HOPS/namenode @ 10
vagrant	MySqlCluster	memcached	vagrant/MySqlCluster/memcached @ 10
vagrant	MySqlCluster	mgmsrver	vagrant/MySqlCluster/mgmsrver @ 10
vagrant	MySqlCluster	mysqld	vagrant/MySqlCluster/mysqld @ 10
vagrant	MySqlCluster	ndb	vagrant/MySqlCluster/ndb @ 10
vagrant	YARN	NodeManager	vagrant/YARN/NodeManager @ 10
vagrant	YARN	ResourceManager	vagrant/YARN/ResourceManager @ 10

► Recent Agent Log

Figure 4.11. Hosts Details Graphs



Alerts

This view allows a user to keep track of the whole history of alerts submitted by the Hop agents. Here you will see for each entry an alert submitted by one of the agents containing the message provided by that alert plus the date the message was generated, the severity of the alert and the host ID that triggered the alert. Also further information is provided with the alert like the source that originated the alert and other parameters that help to track the source of the alert.

Figure 4.12. Alerts

Hadoop Open PaaS Dashboard

Clusters Hosts Alerts Manage Clusters Clusters Progress

Alerts

9 Alerts:

hour 2hr 4hr day week month year or from 25/11/2013 14:19 to 25/11/2013 15:19 by Any severity Any Go

	Date	Severity	Provider	Plugin	Type	Type Instance	Host Id	Message
<input checked="" type="checkbox"/>	Nov 25, 2013 3:04:43 PM	OKAY	Agent	Monitoring	Role	vagrant/YARN/NodeManager	mgmd49	Role is running: vagrant/YARN/NodeManager
<input type="checkbox"/>	Nov 25, 2013 3:04:39 PM	OKAY	Agent	Monitoring	Role	vagrant/YARN/ResourceManager	mgmd49	Role is running: vagrant/YARN/ResourceManager
<input type="checkbox"/>	Nov 25, 2013 3:04:35 PM	OKAY	Agent	Monitoring	Role	vagrant/HOPS/datanode	mgmd49	Role is running: vagrant/HOPS/datanode
<input type="checkbox"/>	Nov 25, 2013 3:04:29 PM	OKAY	Agent	Monitoring	Role	vagrant/HOPS/namenode	mgmd49	Role is running: vagrant/HOPS/namenode
<input type="checkbox"/>	Nov 25, 2013 3:04:19 PM	OKAY	Agent	Monitoring	Role	vagrant/MySQLCluster/memcached	mgmd49	Role is running: vagrant/MySQLCluster/memcached
<input type="checkbox"/>	Nov 25, 2013 3:04:19 PM	OKAY	Agent	Monitoring	Role	vagrant/MySQLCluster/mysqld	mgmd49	Role is running: vagrant/MySQLCluster/mysqld
<input type="checkbox"/>	Nov 25, 2013 3:04:09 PM	OKAY	Agent	Monitoring	Role	vagrant/MySQLCluster/hdb	mgmd49	Role is running: vagrant/MySQLCluster/hdb
<input type="checkbox"/>	Nov 25, 2013 3:04:09 PM	OKAY	Agent	Monitoring	Role	vagrant/MySQLCluster/mgmsrver	mgmd49	Role is running: vagrant/MySQLCluster/mgmsrver
<input type="checkbox"/>	Nov 25, 2013 3:04:09 PM	OKAY	Agent	Monitoring	Role	vagrant/MySQLCluster/mysqld	mgmd49	Role is running: vagrant/MySQLCluster/mysqld

Delete Selected Alerts Delete All

Clusters

Selecting this option from the main Dashboard toolbar, will generate a dropdown list with all the available clusters been actually monitored by the dashboard. From here you can navigate and get further detail of the cluster and the current status of the services running in that cluster. This allows a user to navigate through the different Hop services and sub roles allowing the user to grasp a deep understanding of what is currently going on each of the services.

Figure 4.13. Clusters

The screenshot shows the 'Hadoop Open PaaS Dashboard' with a navigation bar containing 'Clusters', 'Hosts', 'Alerts', 'Manage Clusters', and 'Clusters Progress'. A dropdown menu for 'Clusters' is open, showing 'All Clusters' and 'vagrant'. Below the dropdown, a table titled 'Clusters' displays the following data:

Cluster name	Services	Roles	Roles Status	Health	Hosts	Cores	Disk Capacity	Memory Capacity	Acti
vagrant	MySQLCluster HOPS YARN	1 mgmserver 1 memcached 1 namenode 1 ndb 1 NodeManager 1 datanode 1 mysqlid 1 ResourceManager	8 Started	Good	1	2	78.9 GB	3.8 GB	

From the previous image, we can see a top level overview of each cluster with general information on the status of that cluster. This contains information on the number of nodes that compose that cluster, the current health of the cluster and the number of hosts involved in the cluster. Also it keeps track of the resources allocated on that cluster like the total number of cores which compose the overall computing power of the cluster, the total disk capacity and the total physical memory capacity. Selecting a cluster entry will bring a more detailed view of that cluster.

Figure 4.14. Cluster Detail

The screenshot shows the 'Hadoop Open PaaS Dashboard' with the 'vagrant' cluster selected. The navigation bar is the same. Below the navigation bar, there is a green bar with 'vagrant', 'Status', and 'Action History'. Below this, a table titled 'Cluster Info' displays the following data:

Cluster name	Health	Hosts	Cores	Disk Capacity	Memory Capacity
vagrant	Good	1	2	78.9 GB	3.8 GB

Below the 'Cluster Info' table, there is a table titled 'Services' displaying the following data:

Service	Roles	Roles Status	Health
YARN	1 NodeManager 1 ResourceManager	2 Started	Good
MySQLCluster	1 mgmserver 1 memcached 1 ndb 1 mysqlid	4 Started	Good
HOPS	1 namenode 1 datanode	2 Started	Good

In this new view, we can see that we have gone further inside the services the cluster consists of and here the user can identify the status of each of the services that are part of the current cluster instance. Selecting one of the services will show greater detail of information of the service. We will see how it looks like each view for each of the services that form the HOP. Note that in order for the graphs to appear, a user needs to configure the specific graph for the specific component, refer to the Edit Graph section found in this chapter.

YARN monitoring

If YARN is enabled in your cluster, you can get highly detailed information of YARN like the total amount of resource managers and node managers, and metrics of interest for YARN monitoring.

Figure 4.15. YARN Metrics

Clusters » vagrant » YARN

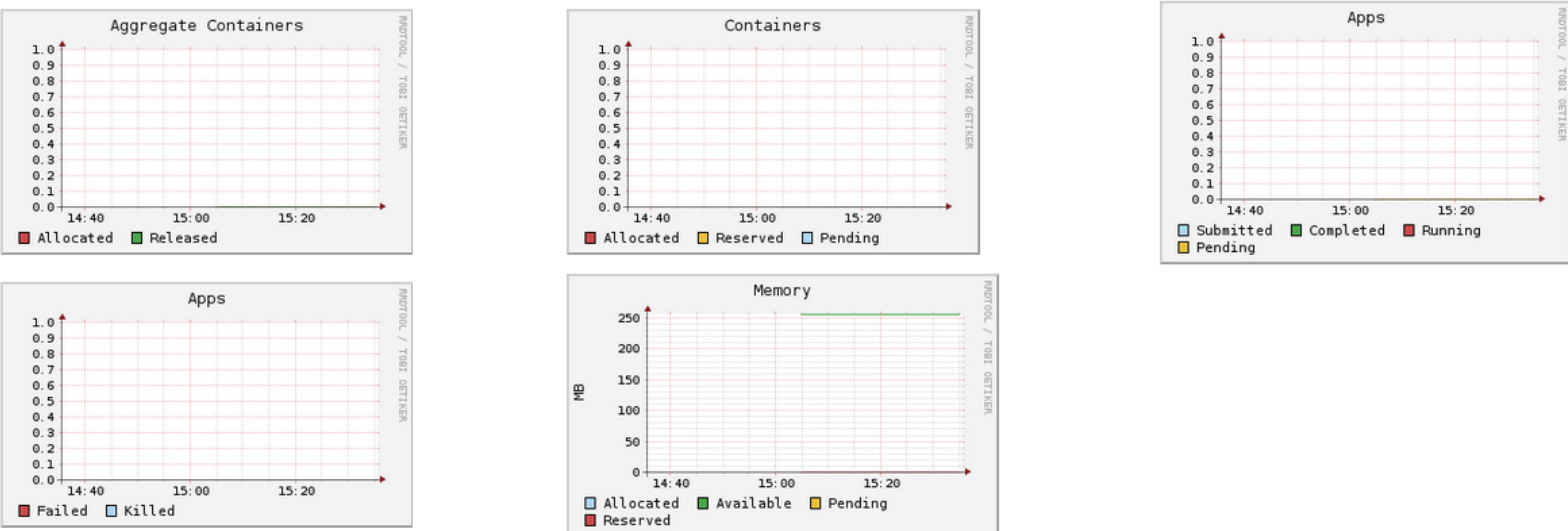
YARN [Status](#) [Instances](#) [Actions History](#)

Roles Status and Health Summary

Role name	Status	Health
Resource Manager	1 Started	1 Good
Node Manager	1 Started	1 Good

hour 2hr 4hr day week month year or from 25/11/2013 14:35 to 25/11/2013 15:35 in 3 columns Go Edit

Resource Manager Queue Metrics



If you select one of the components that form part of the YARN ecosystem, you will get more information on that specific component.

Figure 4.16. Resource Manager Metrics

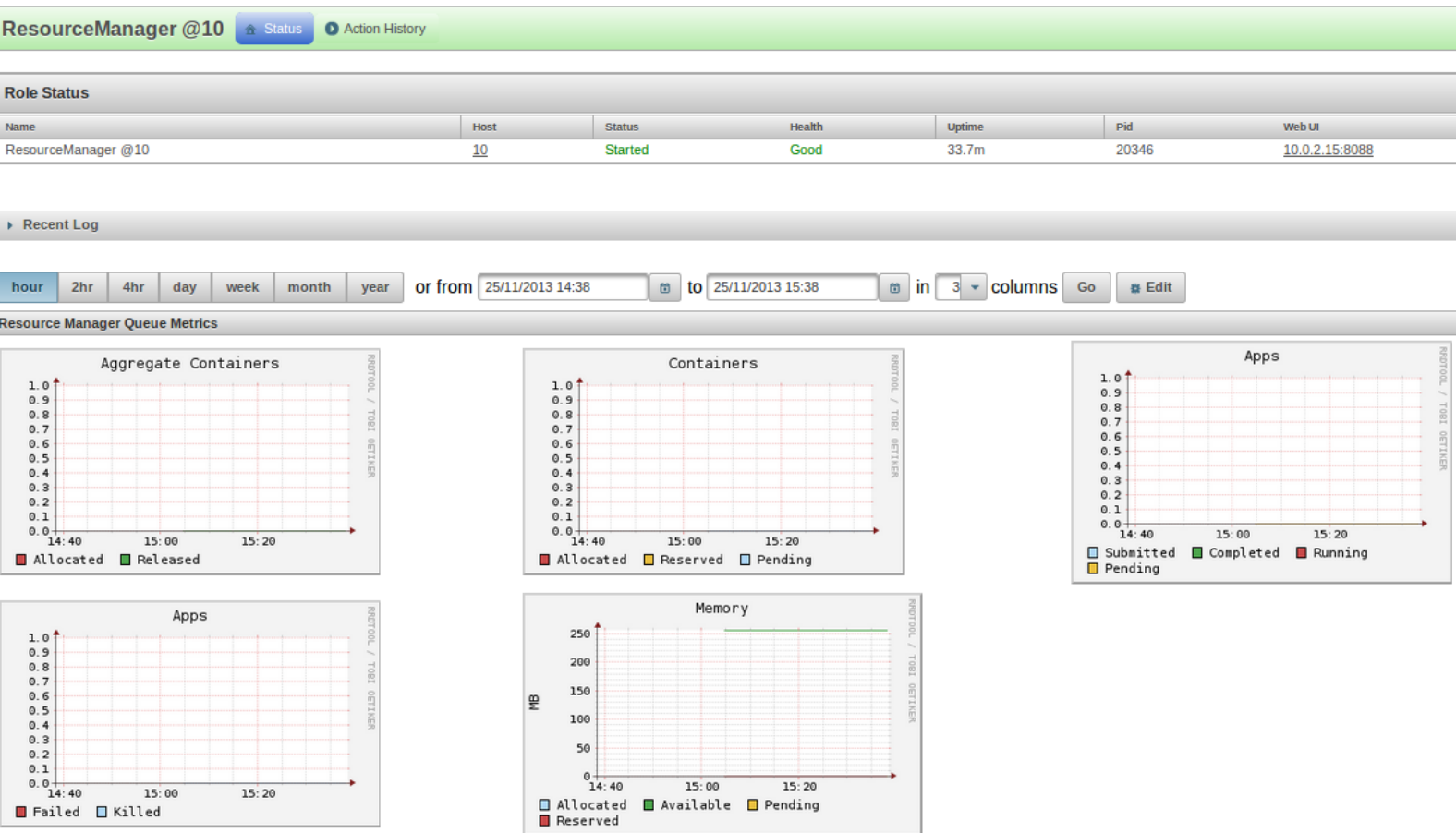


Figure 4.17. Node Manager Metrics

Clusters » vagrant » YARN » NodeManager

NodeManager @10

Status

Action History

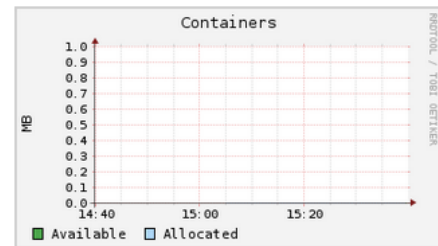
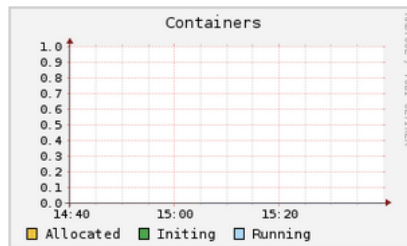
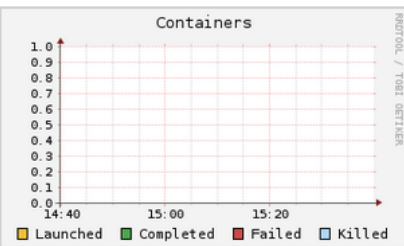
Role Status

Name	Host	Status	Health	Uptime	Pid	Web UI
NodeManager @10	10	Started	Good	35m	20617	10.0.2.15:8042

Recent Log

hour 2hr 4hr day week month year or from 25/11/2013 14:39 to 25/11/2013 15:39 in 3 columns Go Edit

Node Manager Queue Metrics



If you select the corresponding web UI link that appears in one of the YARN components, it will load the respective Hadoop information Web UI with more detailed information of that component.

Figure 4.18. Resource Manager UI



All Applications

Cluster

About
Nodes
Applications
NEW
NEW SAVING
SUBMITTED
ACCEPTED
RUNNING
REMOVING
FINISHING
FINISHED
FAILED
KILLED
Scheduler

Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes
0	0	0	0	0	0 B	256 MB	0 B	1	0	0	0

Show 20 entries Search:

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress
No data available in table									

Showing 0 to 0 of 0 entries First Previous

Figure 4.19. Node Manager UI



- ResourceManager
- NodeManager
 - Node Information
 - List of Applications
 - List of Containers
- Tools

Total Vmem allocated for Containers		537 MB
Vmem enforcement enabled	true	
Total Pmem allocated for Container	256 MB	
Pmem enforcement enabled	true	
NodeHealthyStatus	true	
LastNodeHealthTime	Mon Nov 25 15:38:43 UTC 2013	
NodeHealthReport		
Node Manager Version:	2.2.0 from 6ebdf74bac6882a5b44dda3fe506281fff01cd5f by jdowling source checksum 6afffa66f656213479c75e45d2013-11-13T09:43Z	
Hadoop Version:	2.2.0 from 6ebdf74bac6882a5b44dda3fe506281fff01cd5f by jdowling source checksum 120819887ebbbe88f4a9ce2013-11-13T09:41Z	

MySQL Cluster

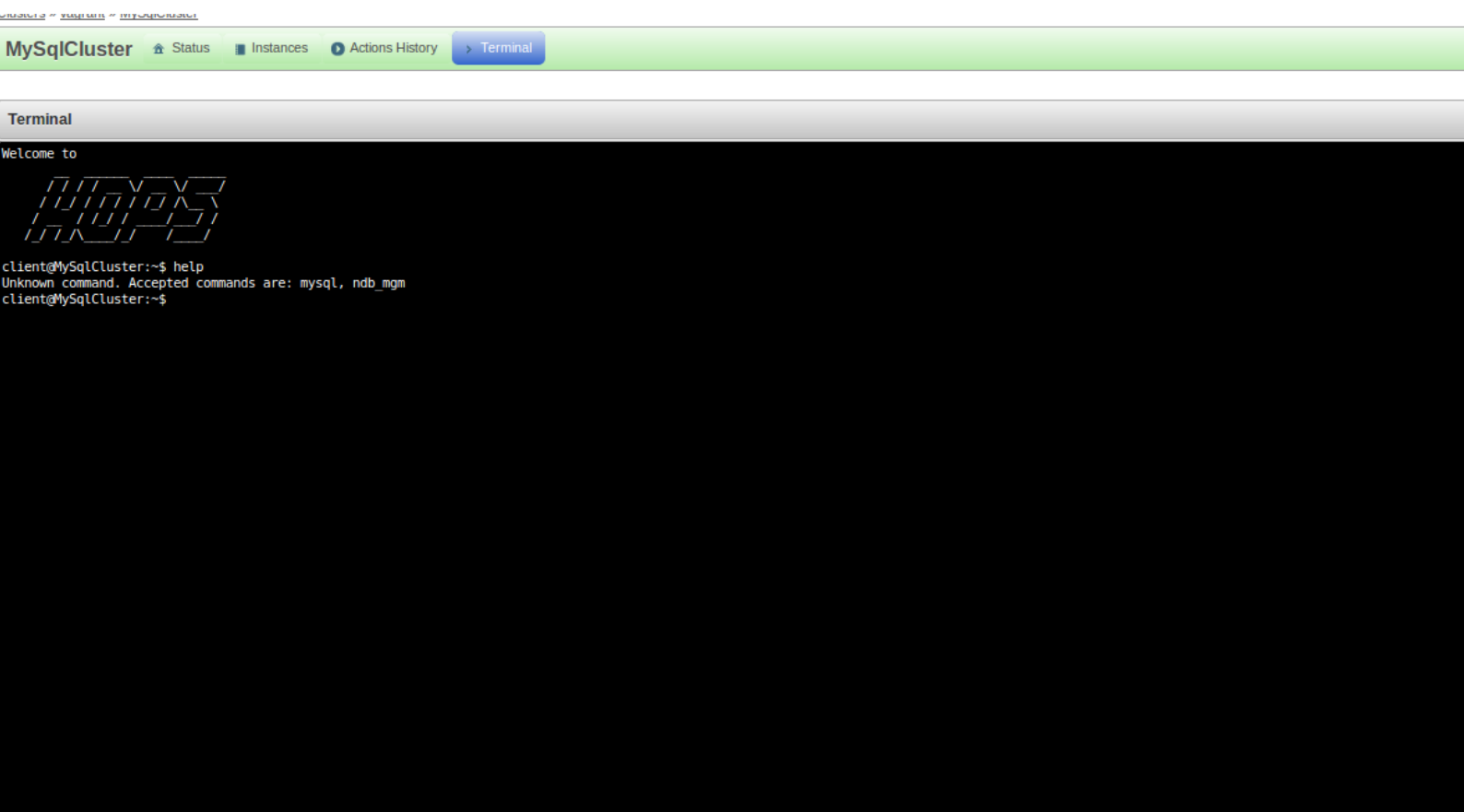
You can get a highly detailed view of what is happening in MySQL cluster after deploying a Hop cluster. You can keep track of statistics of interest in order to keep the performance of your MySQL Cluster in good shape. If you select the MySQL you can obtain the following information if the graphs are configured accordingly, see Edit Graph section in this chapter.

Figure 4.20. MySQL overall graphs



We also offer additional functionality for users to maintain and manage the status of their MySQL cluster without the need to connecting directly to the machine. We provide an online terminal where users can execute mysql commands directly to the MySQL cluster without any delay.

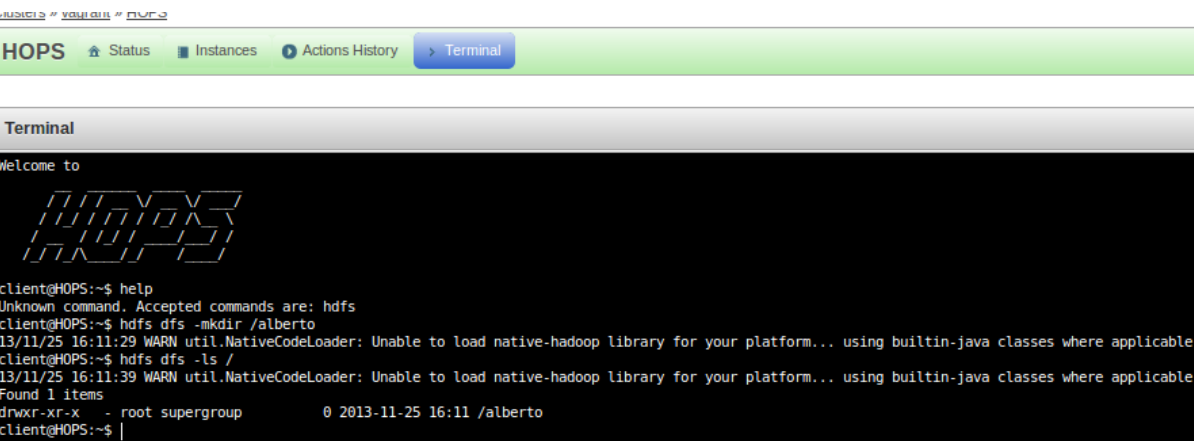
Figure 4.21. MySQL console



HOPS HDFS (Hadoop Filesystem)

You can get a highly detailed view of what is happening in the Hop file system after deploying a Hop cluster. You can keep track of statistics of interest in order to keep the performance of your Hop system in good shape. If you select the Hop option you can obtain information from the graphs previously configured in the Edit Graph section.

We also offer additional functionality for users to manage the file system without the need to connecting directly to the machine. We provide an online terminal where users can execute hdfs commands directly to the file system without any delay.

Figure 4.22. HOP console

The screenshot shows the HOP dashboard interface. At the top, there is a navigation bar with the HOP logo and links for Status, Instances, Actions History, and Terminal. The Terminal tab is active, displaying a terminal window. The terminal output shows a welcome message, a ASCII art logo, and a series of commands and their outputs. The commands include 'help', 'hdfs dfs -mkdir /alberto', and 'hdfs dfs -ls /'. The outputs show warnings about the native-hadoop library and the results of the 'ls' command.

```
client@HOPS:~$ help
Unknown command. Accepted commands are: hdfs
client@HOPS:~$ hdfs dfs -mkdir /alberto
13/11/25 16:11:29 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
client@HOPS:~$ hdfs dfs -ls /
13/11/25 16:11:39 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
drwxr-xr-x  - root supergroup          0 2013-11-25 16:11 /alberto
client@HOPS:~$
```

Chapter 5. Defining a Cluster

In this section, we describe the tools we offer in order to easily define and structure HOP clusters for their deployment through our orchestration architecture. In here, we will introduce you to our definition language to define clusters for cloud providers like Amazon EC2 and OpenStack or Baremetal clusters based on physical machines. With our cluster definition language, you will see that you will easily have a cluster deployed in a matter of minutes by making use of technologies like Chef that will be in charge of orchestrating the nodes while we provision them with Jclouds (in the case of a virtual environment) or a simple SSH client for your baremetal machines.

Cluster Definition Language

We will start first by presenting our Cluster Definition Language (CDL) with which you can defining your clusters with ease. In general, we handle the following abstractions:

- *Cluster*: A cluster is an entity that defines a whole system based on a heterogenous structure composed of multiple nodes. In most of the cases, we can classify the nodes into groups depending of the software they run. To allow further customization of your cluster, we allow interesting options like the possibility of running chef recipes globally on all the nodes and open ports that you may want to be open. In order to identify which type of cluster you are defining, it is necessary to specify the class tag of the type of cluster so our software can map with the bindings of the type of cluster you want to deploy.

Example 5.1. Defining Global Properties

```
!!se.kth.kthfsdashboard.virtualization.clusterparser.Cluster
##name of your cluster
name: test2
##enable install phase
installPhase: true
##global parameters
global:
##user defined recipes
  recipes:
    - ssh
    - chefClient
## extra ports you want to open
  authorizePorts:
    - 3306
    - 4343
    - 3321
```

Git Repositories

If you want further customization, it is possible to fork our git repository and customize our chef recipes if you want to modify some parameters of our cluster. Also you can add your own recipes if you decide to launch other services on your code. Simply add this snippet of code under global parameters.

Example 5.2. Defining Git repository

```
git:
  user: Jim Dowling
  repository: https://ghetto.sics.se/jdowling/hops-chef.git
  key: notNull
```

- *Services:* We identify multiple services, in our case related to Hop platform. You can spread this services quite easily among different nodes just indicating that information when grouping them. Also you may indicate further services to be deployed on them.

```
service:
  - datanode
  - nodemanager
  number: 2
```

- *Provider:* In the case of defining a cluster to be deployed in a virtualized environment through an Amazon EC2 infrastructure or an OpenStack environment, you can give information of the image you want to use, the type of instance to request, login credentials in case you are using custom images.

Example 5.3. Defining Cloud Providers

```
provider:
  ##name of the provider, use aws-ec2 or openstack-nova
  name: aws-ec2
  ##if EC2 use a value to one of EC2 types, in OpenStack this is an id number
  ##type of instance you want to use
  instanceType: m1.large
  ## indicate the login user of the machine with sudo access, necessary for cu
  ## or openstack image
  loginUser: ubuntu
  ## image you want in EC2 or OpenStack
  image: eu-west-1/ami-35667941
  ##region of EC2 or project name in OpenStack
  region: eu-west-1
```

We will also see that the syntax differs on whether are designing your cluster towards a virtualized environment or a physical environment. In the following sections, we will go through detailed examples for both types of clusters.

Structuring your Cluster:

Before using our tools, it is important that you have an idea of how you want to structure the services of our data platform through out the whole cluster. In our case, a fully functional cluster requires the following services deployed in different machines:

1. *MySQL Cluster:*

- *MySQL-NDB:* Your cluster should contain at least 2 instances of NDB

- *MySQL-MGM*: Your cluster should contain at least 1 instance of a Management Server.
- *MySQL-Mysqld*: Your cluster should contain at least 1 instance of a MySQL Server.

2. HOP

- *Namenode*: Your cluster should contain at least 2 namenode instances of our Hadoop Solution.
- *Datanode*: Your cluster should contain at least 2 datanode instances of our Hadoop Solution.

3. Data processing

- *ResourceManager*: Your cluster should contain at least 1 resource manager instances of YARN.
- *NodeManager*: Your cluster should contain at least 2 node manager instances of YARN.
- *Spark*: Your cluster should contain at least one instance of Spark if you want to do data processing through Spark to submit your jobs to the system.

Multiple Services per Node

The previous section gave a very simple overview of the components that are needed for a HOP cluster to work correctly. It is possible to allocate various services in one machine or group of machines as we will see in the following sections.

Now that we have a general perspective of how a cluster looks like, the next step is to identify the environment of your choice for the cluster you want to work with. In the following sections, we will describe how you can define the structure for virtualized cloud providers like Amazon EC2 and OpenStack or in a physical Baremetal environment.

Building your cluster:

In this section, we will explain through a couple of complete examples how to define your cluster for Amazon EC2, OpenStack or Baremetal. We will show you how to write your cluster from scratch using your own YAML file or you can use the available cluster wizard in order to generate your desired cluster.

Cluster in AWS

Lets imagine that we want to define a complete HOP cluster which will contain a basic minimal setup. In this case we need 2 NDBs, 1 MGM and 1 Mysqld for the MySQL cluster, 2 namenode and 2 datanode for the Hadoop File System and in order to user Spark, a Spark instance with 1 resource manager and 2 node managers. How we could map the services using only 7 machines? A very simple configuration could be as follows:

Example 5.4. Full AWS Cluster Example

```
!!se.kth.kthfsdashboard.virtualization.clusterparser.Cluster
name: test2
provider:
  name: aws-ec2
  instanceType: m1.large
  loginUser: ubuntu
  image: eu-west-1/ami-35667941
  region: eu-west-1

##lists of groups, with the roles the nodes
##will have and open ports
nodes:
- service:
  - ndb
  number: 2

- service:
  - mgm
  number: 1

- service:
  - mysqld
  - namenode
  number: 1

- service:
  - namenode
  - resourcemanager
  number: 1

- service:
  - datanode
  - nodemanager
  - spark
  number: 2
```

With this configuration file, we will create 5 security groups which will have as a name the first service defined in the list. This will also open the ports for those security groups. It will install the defined services for each of the nodes in that specific group of nodes.

Cluster in OpenStack

Taking the previous case for Amazon EC2, we can easily write the same cluster description using the same cluster definition file. In this case, the only section we need to change is related to the provider we want to use which in this case is OpenStack. The file will look as follows:

Example 5.5. Full OpenStack Example

```
!!se.kth.kthfsdashboard.virtualization.clusterparser.Cluster
name: nova
provider:
  name: openstack-nova
  instanceType: 7
  loginUser: ubuntu
  image: 0190f9c4-d64e-4412-ab88-4f9fd1d7c2e3
  region: RegionSICS

##lists of groups, with the roles the nodes
##will have and open ports
nodes:
  - service:
    - ndb
    number: 2

  - service:
    - mgm
    number: 1

  - service:
    - mysqld
    - namenode
    number: 1

  - service:
    - namenode
    - resourcemanager
    number: 1

  - service:
    - datanode
    - nodemanager
    - spark
    number: 2
```

With this configuration file, it is possible to deploy the same cluster we defined in Amazon EC2 without any major changes. You only need to change the provider specifications to match the details of your OpenStack Infrastructure.

Cluster on Baremetal Machines

How would we describe the same cluster for Amazon EC2 in a cluster of physical machines? In this case it is much simpler but you need to watch out for minor details like, for example; the class tag needs to be different for this type of clusters as we will see. Also in this case, you need to provide the IP addresses of the machines to connect to. An example is as follows:

Example 5.6. Full Baremetal Example

```
!!se.kth.kthfsdashboard.virtualization.clusterparser.Baremetal
name: baremetal
loginUser: ubuntu
totalHosts: 7
nodes:
  - service: ndb
    number: 2
    hosts:
      - 10.20.0.8
      - 10.20.0.11

  - service: mgm
    number: 1
    hosts:
      - 10.20.0.6

  - service:
    - mysqld
    - namenode
    number: 1
    hosts:
      - 10.20.0.7

  - service:
    - namenode
    - resourcemanager
    number: 1
    hosts:
      - 10.20.0.12
      - 10.20.0.14

  - service:
    - datanode
    - nodemanager
    - spark
    number: 2
    hosts:
      - 10.20.0.16
      - 10.20.0.17
```

With this configuration file, it is possible to deploy the same cluster we defined in Amazon EC2 without any major changes. You only need to change the provider specifications to match the details of your OpenStack Infrastructure.

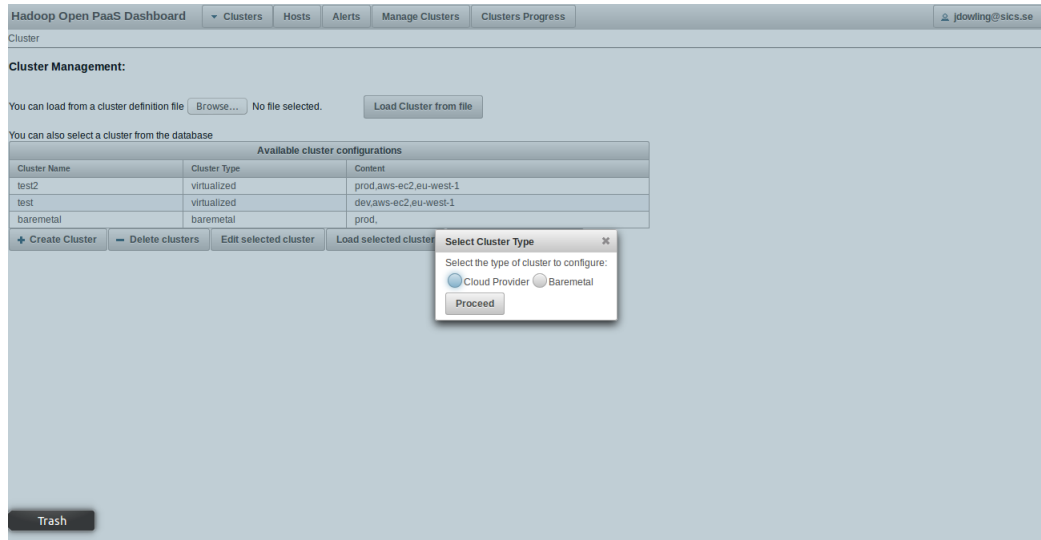
Cluster Generator on Dashboard

Apart of offering a mechanism where users can upload their clusters written in YAML to the system and later on deploy them, we also have a cluster wizard which allows the user to define a cluster step by step quite easily. To make use of this feature, follow these steps:

1. Go to the manage cluster section from the main bar in the dashboard. Select the create cluster option. Main Menu Bar → Manage Cluster → Create cluster
2. A dialog appears allowing you to select which type of cluster you want to use:

- *Virtualized*: Choose this option if you want to deploy a cluster in Amazon EC2 or OpenStack.
- *Baremetal*: Choose this option if you want to deploy a cluster in physical machines.

Figure 5.1. Select Cluster Type:



3. Selecting an option, will bring you to the cluster generator wizard. Here you can select the same options like if you were writing your own file from scratch. You will go through different phases.

Cluster Wizard → Common → Provider (not for Baremetal) → Groups → Confirmation

- *Common Section*: In this section, a form appears where you can select the following options:
 - a. *Name*: Name of the cluster
 - b. *Provider*: Select the type provider between Amazon EC2 or OpenStack, this option is available if we create a virtualized cluster.
 - c. *Git parameters*: Git repository section where you can specify as an option your own git repository based on our code. This way you can customize our recipes or even add your own.
 - d. *Global Recipes*: You can specify chef recipes that you want to execute in all the nodes
 - e. *Global Ports*: Additional Ports to open for your cluster, this option is only available for virtualized clusters.

Figure 5.2. Common Cluster Options:

Common
Provider
Groups
Confirmation

Global Attributes

Cluster name: * test2

Provider: * Amazon

Environment: * Production

Install Phase:

Git parameters (Optional)

Global Recipes

Click delete selection after selecting multiple instances to delete, use shift-click to select multiples

Recipe
Actions

No records found.

+ New Recipe
Delete Selection

Global Ports

Click delete selection after selecting multiple instances to delete, use shift-click to select multiples

Port Number
Actions

3306

4343

3321

+ New Port
Delete Selection

Next

Figure 5.3. Bare Metal Common Cluster Options:

Common
Groups
Confirmation

Global Attributes

Cluster name: * baremetal

Login user: * ubuntu

Total number of Hosts: * 8

Environment: * Production

Install Phase:

Git parameters (Optional)

Global Recipes

Click delete selection after selecting multiple instances to delete, use shift-click to select multiples

Recipe
Actions

No records found.

+ New Recipe
Delete Selection

Next

- *Provider Section:* This form enables you to define the parameters for OpenStack or Amazon EC2. Some values appear by default in the case of Amazon EC2 of defining a cluster to be used with this cloud provider.

- a. *Instance Type*: The type of instance you want to use in Amazon EC2 or in OpenStack. Note that in OpenStack we use the id number of the type of instance, not the name.
- b. *Image*: The name of the image we want to use the in Amazon EC2 or in OpenStack
- c. *Login user*: Here you include the user name with sudo access to access the instances in Amazon EC2 or OpenStack. Note that this value is necessary if you use a custom AMI in Amazon EC2 or using you use OpenStack.
- d. *Region*: Here you include the region you want to deploy in Amazon EC2 or the project to use in your OpenStack infrastructure.

Figure 5.4. Cluster Provider Options:

- *Group Section*: In this section you can specify the group of nodes for you cluster with the their services and ip addresses (if you are deploying a baremetal cluster)
 - a. *Main Service*: The main service you want to deploy in this group of nodes
 - b. *Bittorrent Support*: If you want to enable bittorrent sync of binaries from the dashboard.
 - c. *Number of nodes*: Number of nodes that will contain the same set of services.
 - d. *Extra Services*: Other services you may want to run which can be also your own services.
 - e. *ChefAttributes*: In this section, you would include a chef json which will contain the attributes you may want to override from your recipes.
 - f. *Ports*: Extra ports that you may want to enable in that group, in this case this only affect virtualized clusters.
 - g. *Hosts*: List of hosts IP addresses for the nodes that will be part of this group of nodes. In this case this option is only available for Baremetal clusters.

Figure 5.5. Cluster Group:

Common Provider **Groups** Confirmation

Cluster Nodes

Click delete selection after selecting multiple instances to delete, use shift-click to select multiples

Service	Number of Nodes	Recipes	Ports	Bittorrent	Chef Attributes
ndb	2				
mgm	1				
mysqld	1				
namenode	2				
datanode	2				

+ New Group Edit selection - Delete Selection

← Back → Next

Figure 5.6. Bare Metal Groups:

Common **Groups** Confirmation

Cluster Nodes

Click delete selection after selecting multiple instances to delete, use shift-click to select multiples

Service	Number of Nodes	Recipes	Hosts:	Bittorrent	Chef Attributes
ndb	2	[MySQLCluster-ndb]	[10.20.0.8, 10.20.0.11]		
mgm	1	[MySQLCluster-mgm]	[10.20.0.6]		
mysql	1	[MySQLCluster-mysqld]	[10.20.0.7]		
namenode	2	[KTHFS-namenode]	[10.20.0.12, 10.20.0.14]		
datanode	2	[KTHFS-datanode]	[10.20.0.16, 10.20.0.17]		

+ New Group Edit selection - Delete Selection

← Back → Next

- *Confirmation Section:* In this section you will see a summary of the details of your cluster file. When you press the submit button, your cluster file will be stored in the dashboard and it will proceed to the cluster launcher.

Figure 5.7. Confirmation:

The screenshot shows a web-based configuration wizard with four tabs: 'Common', 'Provider', 'Groups', and 'Confirmation'. The 'Confirmation' tab is active. Below the tabs is a section titled 'Confirmation' with a sub-section 'General Details:' containing the following fields:

Name:	test2
Environment:	prod
Install Phase:	false
Authorize Ports:	[3306, 4343, 3321]
Git user:	Jim Dowling
Git repository:	https://ghetto.sics.se/jdowling/kthfs-pantry.git
Git key:	notNull

Below the 'General Details' section are two expandable sections: 'Provider Details:' and 'Nodes:'. At the bottom of the form are two buttons: 'Submit' and '← Back'.

Wrap up

To summarize this section, in here we have seen the main building blocks that we need to define a cluster using our cluster domain specific language. We also explained how you can define your clusters by writing your own cluster file through multiple examples and also showed an alternative way of defining cluster through the cluster generator wizard which is accessible from the dashboard.

Chapter 6. Launching a Cluster

Installation on AWS

In this section, we will explain further steps that are required to deploy a whole functional cluster running our data platform through the dashboard. Also we refer to recommendations and aspects you should consider before deploying a cluster.

Pre-requisites:

Before starting, make sure that you have access to a functional and running Dashboard in a virtual machine in an accessible Amazon EC2 region. If you have not done so, please refer back to the Chapter 3, *Hop Web Portal*.

Requirements:

In order to install and deploy a cluster, you need to define before the structure of the cluster which includes specifying the number of machines to create in EC2 with the specific instance type with the specific software. This can be done using a cluster definition file that can be done from scratch or using the embedded wizard available on the dashboard. Further information about describing a cluster can be found on the cluster configuration section. Before continuing make sure that you have the following.

- Cluster definition for EC2 (see related section) in a file or loaded from the dashboard database.
- Amazon EC2 credentials to deploy the cluster in Amazon, configured in the dashboard. In order to do it, select the option setup credentials found in your user icon to specify the EC2 credentials to be used by the dashboard.

Additional dashboard credentials

It is possible to include other options when deploying an EC2 cluster, for example; for maintenance purposes you might want to authorize extra public keys to the virtual machines. This is possible to set in the credential section of the dashboard.

Launching the cluster

Once we have the dashboard configured with the Amazon EC2 credentials, you can proceed to launch a cluster:

1. Select the manage cluster option available in the dashboard.

Main Menu Bar → Manage Cluster → Load File

2. In this new view, you can manage available clusters that you may have defined previously. You can select a previous cluster, create a new one with the only wizard or load a cluster from a cluster definition file. For further information on managing cluster files, see Chapter 5, *Defining a Cluster*. To continue, select a cluster from the table or load a cluster from a file.
3. The file is loaded and the launcher view should appear. Here you can view the contents of the cluster to be deployed before launch.
4. Pressing the start cluster will start the deployment process. A status bar will appear giving information of the current status of the deployment. Also a progress table on the background will be generated with information of the configuration state of the nodes. The process is long and it depends on the number of nodes you deploy. On average, for 8 nodes it takes around 35 minutes.

Error Nodes

It is possible that some nodes will have issues during the deployment of our software (package configuration problem, erratic behaviour) which in this case our system will detect and will retry to relaunch the software on that specific machine automatically. The maximum number of retries specified for each node is 5, after that; the node will be tagged as an error node and it is possible to do a manual retry after the whole process has finished.

5. When the process completes, it will take you back to the progress view where you can see details of the cluster deployment. If nodes failed, you can select those nodes and try to recover them using the retry nodes option.

Retrying Nodes

Retrying node is an option that helps bringing back nodes that had minor issues when installing packages, were to slow to finish the configuration phase or the default number of retries we use were not enough. It will not bring back nodes which had a critical configuration failure, which in this case it will be necessary to log in directly through SSH to the specific machine in order to fix it.

Congratulations, if everything went okay; you have successfully deployed a complete cluster ready to use!

Installation on OpenStack

In this section, we will explain further steps that are required to deploy a whole functional cluster running our data platform through the dashboard. Also we refer to recommendations and aspects you should consider before deploying a cluster in OpenStack.

OpenStack Deployment

Note please that this option is currently in development phase and from our tests we managed to deploy functional testing clusters. Still due to issues we encountered during our tests in our personal OpenStack, we cannot guarantee the same level of performance as deploying for example a cluster in EC2. This is due to the fact that our deployment system depends greatly on how effectively OpenStack behaves with your hardware and so unexpected behaviour might take place. If you have a very good OpenStack infrastructure, we invite you to test it.

Pre-requisites:

Before starting, make sure that you have access to a functional and running Dashboard in a virtual machine accessible from your OpenStack Infrastructure. If you have not done so, please refer back to the section Chapter 3, *Hop Web Portal*.

Requirements:

In order to install and deploy a cluster, you need to define before the structure of the cluster which includes specifying the number of machines to create in OpenStack with the specific instance type with the specific software. This can be done using a cluster definition file that can be done from scratch or using the embedded wizard available on the dashboard. Further information about describing a cluster can be found on Chapter 5, *Defining a Cluster*. Before continuing make sure that you have the following.

- Cluster definition for OpenStack (see related section) in a file or loaded from the dashboard database.

- OpenStack credentials to deploy the cluster on your OpenStack infrastructure, configured in the dashboard. In order to do it, select the option setup credentials found in you user icon to specify the OpenStack credentials to be used by the dashboard.

Additional dashboard credentials

It is possible to include other options when deploying a OpenStack cluster, for example; for maintainance purposes you might want to authorize extra public keys to the virtual machines. This is possible to set in the credential section of the dashboard

Launching the cluster

Once we have the dashboard configured with the OpenStack credentials, you can proceed to launch a cluster:

1. Select the manage cluster option available in the dashboard.

Main Menu Bar → Manage Cluster → Load File

2. In this new view, you can manage available clusters that you may have defined previously. You can select a previous cluster, create a new one with the only wizard or load a cluster from a cluster definition file. For further information on managing cluster files, see the cluster configuration section. To continue, select a cluster from the table or load a cluster from a file.
3. The file is loaded and the launcher view should appear. Here you can view the contents of the cluster to be deployed before launch.
4. Pressing the start cluster will start the deployment process. A status bar will appear giving information of the current process. Also a progress table on the background will be generated with information of the configuration state of the nodes. The process is long and it depends on the number of nodes you deploy. On average, for 8 nodes it takes around 35 minutes.

Error Nodes

It is possible that some nodes will have issues during the deployment of our software (package configuration problem, erratic behaviour) which in this case our system will detect and will retry to relaunch the software on that specific machine automatically. The maximum number of retries specified for each node is 5, after that; the node will be tagged as an error node and it is possible to do a manual retry after the whole process has finished.

5. When the process completes, it will take you back to the progress view where you can see details of the cluster deployment. If nodes failed, you can select those nodes and try to recover them using the retry nodes option.

Retrying Nodes

Retrying node is an option that helps bringing back nodes that had minor issues when installing packages, were to slow to finish the configuration phase or the default number of retries we use were not enough. It will not bring back nodes which had a critical configuration failure, which in this case it will be necessary to log in directly through SSH to the specific machine in order to fix it.

Congratulations, if everything went okay; you have succesfully deployed a complete cluster ready to use!

Installation on Baremetal Machines

In this section, we explain the steps that are required through the dashboard to deploy our data platform on a cluster of hosts running the linux operating system.

Pre-requisites:

Before starting, make sure that you have access to a functional and running Dashboard in a host which you can access via a browser. If you have not done so, please refer back to Chapter 3, *Hop Web Portal*.

Requirements:

In order to install and deploy a cluster, you first need to specify the set of ip addresses for the hosts and the specific software. This can be done using a cluster definition file that can be done from scratch or using the embedded wizard available on the dashboard. Further information about describing a cluster can be found on Chapter 5, *Defining a Cluster*. Before continuing make sure that you have the following.

- Cluster definition for a baremetal cluster (see related section) in a file or loaded from the dashboard database.
- Credentials to connect to your physical machine, this means a user name with sudo access and the private key to SSH the machines.

Launching the cluster

Once we have the dashboard configured with the physical machines credentials, you can proceed to launch a cluster:

1. Select the manage cluster option available in the dashboard.

Main Menu Bar → Manage Cluster → Load File

2. In this new view, you can manage available clusters that you may have defined previously. You can select a previous cluster, create a new one with the only wizard or load a cluster from a cluster definition file. For further information on managing cluster files, see the cluster configuration section. To continue, select a cluster from the table or load a cluster from a file.
3. The file is loaded and the launcher view should appear. Here you can view the contents of the cluster to be deployed before launch.
4. Pressing the start cluster will start the deployment process. A status bar will appear giving information of the current process. Also a progress table on the background will be generated with information of the configuration state of the nodes. The process is long and it depends on the number of nodes you deploy. On average, for 8 nodes it takes around 35 minutes.

Error Nodes

It is possible that some nodes will have issues during the deployment of our software (package configuration problem, erratic behaviour) which in this case our system will detect and will retry to relaunch the software on that specific machine automatically. The maximum number of retries specified for each node is 5, after that; the node will be tagged as an error node and it is possible to do a manual retry after the whole process has finished.

5. When the process completes, it will take you back to the progress view where you can see details of the cluster deployment. If nodes failed, you can select those nodes and try to recover them using the retry nodes option.

Retrying Nodes

Retrying node is an option that helps bringing back nodes that had minor issues when installing packages, were too slow to finish the configuration phase or the default number of retries we use were not enough. It will not bring back nodes which had a critical

configuration failure, which in this case it will be necessary to log in directly through SSH to the specific machine in order to fix it.

Congratulations, if everything went okay you have successfully deployed a cluster that is ready to use!

Chapter 7. Configuring HDFS

We introduce a few new configuration parameters to HDFS, due to our support for multiple NameNodes and use of MySQL Cluster for metadata storage. These parameters are specified in *hdfs-site.xml*. The configuration parameters listed below are additional to the configuration parameters for vanilla HDFS [<http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/hdfs-default.xml>].

HDFS Configuration Parameters not used

We have replaced HDFS 2.x's Primary-Secondary Replication model with shared atomic transactional memory. This means that we no longer use the parameters in HDFS that are based on the (eventually consistent) replication of *edit log entries* from the Primary NameNode to the Secondary NameNode using a set of quorum-based replication servers. Here are the parameters that are not used in the HOP version of HDFS 2.x:

- *dfs.namenode.secondary.**: None of the secondary NameNode attributes are used.
- *dfs.namenode.checkpoint.**: None of the checkpoint attributes are used.
- *dfs.image.**: None of the FSImage attributes are used.
- *dfs.journalnode.**: None of the hadoop's journaling attributes are used.
- *dfs.ha.**: None of the hadoop high availability attributes are used.
- *dfs.namenode.num.extra.edits.**: None of the edit logs attributes are used.
- *dfs.namenode.name.dir.**: FSImage is not supported anymore.
- *dfs.namenode.edits.**: None of the edit log attributes are used.
- *dfs.namenode.shared.edits.**: None of the edit log attributes are used.

Additional HDFS Configuration Parameters

- *dfs.storage.type*: In HOP all the NameNodes in the system are stateless. All the file system metadata is stored in a relational database. We have chosen MySQL NDB Cluster for its high performance and availability for the storage of the metadata. However the metadata can be stored in any relational database. Default value is this parameter is 'clusterj'. By default HOPS uses ClusterJ libraries to connect to MySQL NDB Cluster. Later we will provide support of other DBMSs.
- *dfs.dbconnector.string*: Host name of management server of MySQL NDB Cluster.
- *dfs.dbconnector.database*: Name of the database that contains the metadata tables.
- *dfs.dbconnector.num-session-factories*: This is the number of connections that are created in the ClusterJ connection pool. If it is set to 1 then all the sessions share the same connection; all requests for a SessionFactory with the same connect string and database will share a single SessionFactory. A setting of 0 disables pooling; each request for a SessionFactory will receive its own unique SessionFactory. We set the default value of this parameter to 3.
- *dfs.storage.mysql.user*: A valid user name to access MySQL Server. For higher performance we use MySQL Server to perform aggregate queries on the file system metadata.
- *dfs.storage.mysql.user.password*: MySQL user password
- *dfs.storage.mysql.port*: MySQL Server port. If not specified then default value of 3306 is chosen.

- *dfs.quota.enabled*: Using this parameter quota can be en/disabled. By default quota is enabled.
- *dfs.namenodes.rpc.address*: HOP support multiple active NameNodes. A client can send a RPC request to any of the active NameNodes. This parameter specifies a list of active NameNodes in the system. The list has following format [ip:port, ip:port, ...]. It is not necessary that this list contain all the active NameNodes in the system. Single valid reference to an active NameNode is sufficient. At the time of startup the client will obtain the updated list of all the NameNodes in the system from the given NameNode. If this list is empty then the client will connect to 'fs.default.name'.
- *dfs.namenode.selector-policy*: For a RPC call client will choose an active NameNode based on the following policies.

1. ROUND_ROBIN

2. RANDOM

By default NameNode selection policy is set of ROUND_ROBIN

- *dfs.leader.check.interval*: One of the active NameNodes is chosen as a leader to perform housekeeping operations. All NameNodes periodically send a HeartBeat and check for changes in the membership of the NameNodes. By default the HeartBeat is sent after every second. Increasing the time interval would lead to slow failure detection.
- *dfs.leader.missed.hb*: This property specifies when a NameNode is declared dead. By default a NameNode is declared dead if it misses a HeatBeat. Higher values of this property would lead to slow failure detection.
- *dfs.block.pool.id*: Due to shared state among the NameNodes, HOP only support one block pool. Set this property to set a custom value for block pool. Default block pood id is HOP_BLOCK_POOL_123.
- *dfs.name.space.id*: Due to shared state among NameNodes, HOP only support one name space. Set this property to set a custom value for name space. Default name space id is 911 :)
- *dfs.clinet.max.retires.on.failure*: The client will retry the RPC call if the RPC fails due to the failure of the NameNode. This property specifies how many times the client would retry the RPC before throwing an exception. This property is directly related to number of expected simultaneous failures of NameNodes. Set this value to '1' in case of low failure rates such as one dead NameNode at any given time. It is recommended that this property must be set to value ≥ 1 .
- *dsf.client.max.random.wait.on.retry*: A RPC can fail because of many factors such as NameNode failure, network congestion etc. Changes in the membership of NameNodes can lead to contention on the remaining NameNodes. In order to avoid contention on the remaining NameNodes in the system the client would randomly wait between [0,MAX_VALUE] ms before retrying the RPC. This property specifies MAX_VALUE; by default it is set to 1000 ms.
- *dsf.client.refresh.namenode.list*: All clients periodically refresh their view of active NameNodes in the system. By default after every minute the client checks for changes in the membership of the NameNodes. Higher values can be chosen for scenarios where the membership does not change frequently.