**What is Ambari?**

Ambari is an open source administration tool deployed on top of Hadoop cluster and responsible for keeping track of running applications and their status. Ambari is used to install and monitor Hortonworks distribution :- HDP

* Instantaneous insight into the health of Hadoop cluster using pre-configured operational metrics
* User-friendly configuration providing an easy step-by-step guide for installation.
* Flexible and adaptive technology fitting perfectly in the enterprise environment.

## **Ambari architecture**

Ambari provides intuitive and REST APIs that automate the operations in the Hadoop cluster. Its consistent and secure interface allows it to be fairly efficient in operational control. Its easy and user-friendly interface efficiently diagnoses the health of Hadoop cluster using an interactive dashboard.

PostGres SQL

Rest Interface

Agent 3

Agent 2

Agent 1

Agent Interface

Ambari follows a master/slave architecture where the master node instructs the slave nodes to perform certain actions and report back the state of every action. The master node is responsible for keeping track of the state of the infrastructure. To do this, the master node uses a database server, which can be configured during setup time

These are the following applications in Ambari, at the core:

* Ambari server
* The Ambari agent
* Ambari web UI
* Database

**Ambari Server**

The entry point for all administrative activities on the master server is known as Ambari server. It is a shell script. Internally this script uses Python code, **ambari-server.py** and routes all the requests to it.

**Ambari Agent**

The Ambari Agent runs on all the nodes that we want to manage with Ambari.  This program periodically heartbeats to the master node. By using this agent, Ambari-server executes many of the tasks on the servers.

**Ambari web interface**

Ambari web interface is one of the powerful features of Ambari application. The web application is through the server of Ambari program which is running on the master host exposed on port 8080. One can access this application and this application is protected by authentication. Also, one can control and view all aspects of your Hadoop Cluster, once one log in to the web portal.

**Database**

Ambari supports multiple RDBMS (Relational Database Management Systems) to keep track of the state of the entire Hadoop infrastructure. Also, you can choose the database you want to use during the setup of the Ambari for the first time Few database supported are :- Postgres,Mysql,Oracle.

## **Benefits of using Ambari**

This is given with respect to Hortonworks Data Platform (HDP). Ambari eliminates the need for manual tasks used to watch over Hadoop operations. It gives a simple secure platform for provisioning, managing and monitoring HDP deployments.

1. **Installation, configuration and management is way simplified**

In present Hadoop cluster each and every service is installed manually while using Ambari will bring automation and one stop shop for entire cluster**.** Ambari also provides rolling upgrade feature where running clusters can be updated on the go with maintenance releases and feature bearing releases and therefore there is no unnecessary downtime. When there are large clusters involved then rolling updates are simply not possible in which case express updates are used

1. **Centralized security and application**

Complexity of cluster security configuration and administration is greatly reduced by Ambari which is among the components of Hadoop ecosystem. The tool also helps with automated setup of advanced security constructs like Kerboros and Ranger.

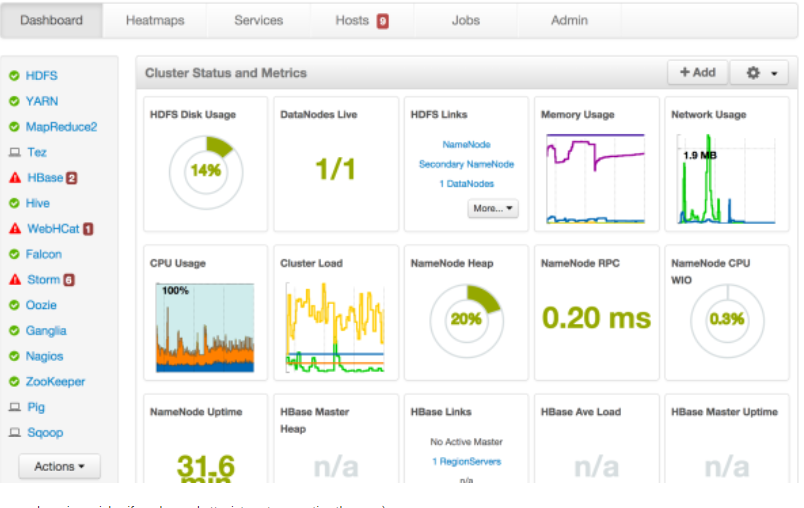
1. **Complete visibility to cluster health**

Through this tool you can monitor your cluster’s health and availability. Using the web GUI we can see the health of each and every component of cluster

1. **Metrics visualization and dashboarding**

Grafana is included with Ambari metrics with HDP which is leading graph and dashboard builder

**Below is the screenshot of Ambari dashboard.**



**HDP Capabilities**

**Various services that comes with HDP distributions are**

* Ranger
* Knox
* ZooKeeper
* HDFS
* YARN
* HBase
* Hive Metastore
* HiveServer2
* WebHCat
* Oozie
* Hue
* Storm
* Kafka
* Atlas

The restart should also follow in the same manner.

1. Ranger is a framework used to monitor,enable and manage data security across Hadoop cluster**.** With the advent of Apache YARN, the Hadoop platform can now support a true data lake architecture
2. Knox acts as a gateway which provides single point of authentication
3. Zookeeper is used to maintain consistency throughout the cluster. It stores most of the metadata about the cluster. It acts as a co-ordination service for distributed application
4. HDFS stands for Hadoop Distributed File System. It is the place where all the data is stored for Hadoop application. Here namenode and datanode comes into the picture to provide HA.
5. Yarn is used for resource management and job scheduling. It is responsible for allocating resources for Hadoop application
6. HBase is a database built on top of the HDFS. It provides fast lookups for larger tables by using key value pair as it is a nosql database
7. Hive enables easy data summarization, ad-hoc querying and analysis of large volumes of data. It is a data warehousing infrastructure built on top of apache Hadoop.
8. WebHCat is a REST API for HCatalog, a table, and storage management layer for Apache Hadoop. WebHCat is enabled by default on HDInsight clusters, and is used by various tools to submit jobs, get job status, etc. without logging in to the cluster.
9. Oozie is used to schedule Hadoop jobs.There are two basic types of Oozie jobs:
   1. Oozie Workflow jobs are Directed Acyclical Graphs (DAGs), specifying a sequence of actions to execute. The Workflow job has to wait
   2. Oozie Coordinator jobs are recurrent Oozie Workflow jobs that are triggered by time and data availability
10. Hue is used to analyze the data with Hadoop via web interface
11. Storm is used to process the data in real time. Storm is extremely fast, with the ability to process over a million records per second per node on a cluster of modest size.
12. Kafka is publish-subscribe messaging service . Here the producer publish to topic and consumer subscribe to topic. It’s equivalent can be Amazon SQS. It’s completely fault tolerant
13. Atlas provides scalable governance for Enterprise Hadoop that is driven by metadata. Atlas, at its core, is designed to easily model new business processes and data assets with agility

.