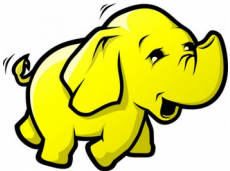


Cloudera Administrator Apache Hadoop

Parte 01-2 Planejamento e instalação do CDH



Marco Reis
<http://marcoreis.net>

Agenda



- Planejando o Cluster Hadoop
- Instalação do Hadoop e configuração inicial
- Instalação e configuração do Hive, Impala e Pig

Principais serviços do Hadoop



- Namenode
- Secondary Namenode
- ResourceManager
- Datanode
- NodeManager
- History Server

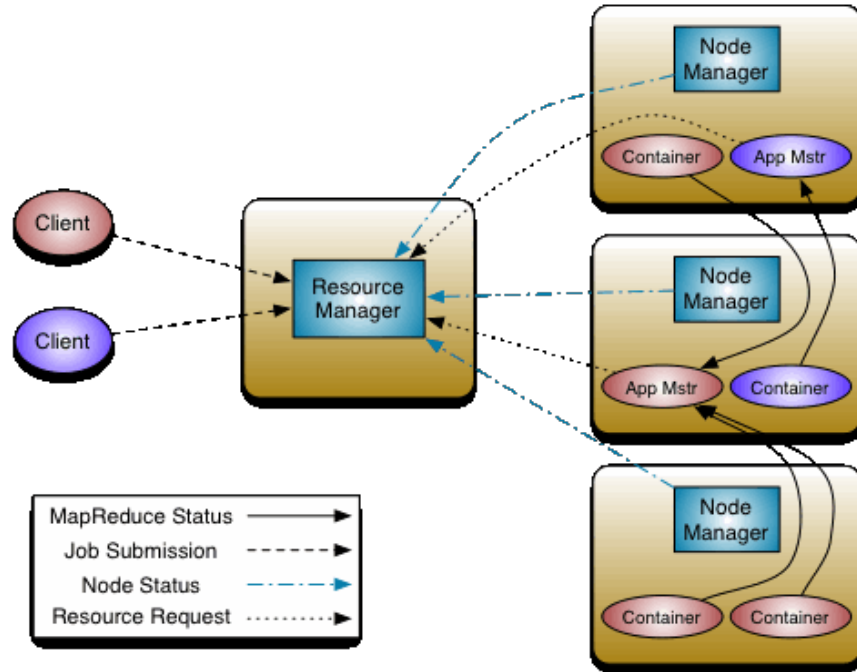
Secondary NameNode

- Responsável por duas tarefas:
 - Armazenar uma cópia do checkpoint
 - Fazer o merge do fsimage/edits, para atualizar o fsimage do Namenode

ResourceManager

- Serviço do master node (head node)
- É responsável pelo inventário dos recursos disponíveis no cluster
- Executa dois serviços:
 - Scheduler: aloca os recursos para todas as aplicações
 - ApplicationManager: cria o primeiro container (bloco de memória) de uma aplicação submetida ao cluster e gerencia o container em caso de falha

Arquitetura RM



- Centraliza e dimensiona recursos
 - NodeManager (NM)
- Nó do cluster
 - Container
- Unidade de processamento, com memória e CPU
- ApplicationMaster (AM)
 - Uma instância para cada aplicação

DataNode

- Serviço do worker node, ou slave
- É responsável por armazenar os dados no HDFS
- Os arquivos são divididos em blocos de 128 MB ou 256 MB (configurável)
- Atua como o slave do cluster, assim, envia os metadados para o namenode (master) sobre os seus arquivos e blocos

NodeManager

- Serviço que roda em cada slave (worker)
- Responsável por lançar e gerenciar os containers em um servidor, obedecendo a um ResourceManager
- Cada servidor roda o próprio NodeManager para gerenciar seus recursos (CPU, memória e disco)
- Executa as tarefas definidas pelo ApplicationMaster (map e reduce)
- Envia metadados para o master
 - Heartbeat, slots e status

History Server

- Serviço de suporte que armazena os logs de execução das aplicações finalizadas com sucesso ou com falha
- É opcional, entretanto, seu uso é recomendado para monitoramento das aplicações

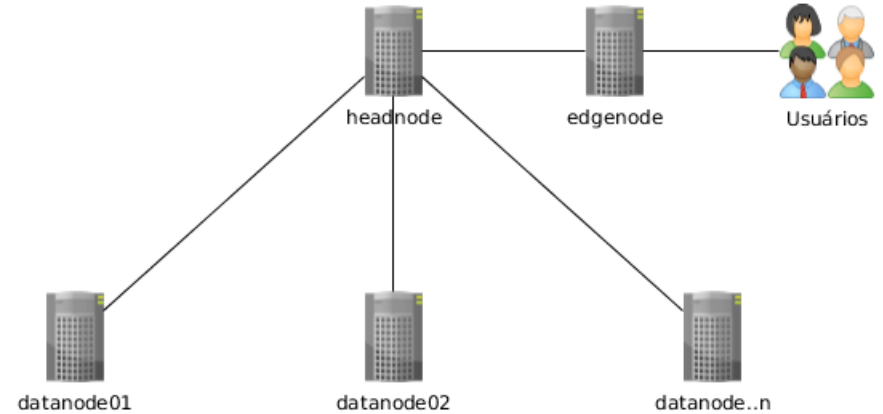


Daemons

- Programas Java que implementam os serviços
- Para visualizar use um `ps aux | grep nome-do-serviço`
- Daemons:
 - namenode – ex: `ps aux | grep namenode`
 - secondarynamenode
 - datanode
 - resourcemanager
 - nodemanager
 - historyserver

Planejamento do cluster

- Sugestão de configuração inicial
 - A rede interna do cluster deve ser isolada da rede externa corporativa
 - Uma opção é que apenas o HeadNode e o EdgeNode tenham interface externa
- 1 servidor master (HeadNode) com 16 GB
 - Principais serviços do cluster
- 1 servidor gateway (EdgeNode) com 16 GB
 - Serviços secundários e acesso dos usuários
- 3 servidores slaves com 8 GB
 - DataNode e NodeManager
- Sistema operacional
 - Linux Ubuntu Server 16
- Banco de dados
 - MariaDB/MySQL
- Hadoop
 - Cloudera 5.15
- Arquitetura de referência
 - https://www.cloudera.com/documentation/other/reference-architecture/topics/ra_private_cloud.html



Configuração dos servidores

- 1) Configurar ip estático
- 2) Verificar o /etc/hosts
- 3) Desabilitar IPv6
- 4) Configurar ssh
- 5) Copiar chaves
- 6) Configurar acesso do MariaDB

Script de preparação dos servidores

- # Todos os nós
- `sudo sed -i -e 's/127.0.1.1/#127.0.1.1/g' /etc/hosts`
- `cat /proc/sys/net/ipv6/conf/all/disable_ipv6`
- `sudo sh -c "echo 'net.ipv6.conf.all.disable_ipv6 = 1`
- `net.ipv6.conf.default.disable_ipv6 = 1`
- `net.ipv6.conf.lo.disable_ipv6 = 1' >> /etc/sysctl.conf"`
- `sudo sysctl -p`
- `cat /proc/sys/net/ipv6/conf/all/disable_ipv6`
- `sudo sed -i -e 's/PermitRootLogin prohibit-password/PermitRootLogin yes/g' /etc/ssh/sshd_config`
- `sudo service ssh restart`
- # Chaves (headserver01 como root)
- `ssh-keygen -t rsa -P " -f ~/.ssh/id_rsa`
- `cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys`
- `ssh-copy-id -i ~/.ssh/id_rsa.pub edgenode.lab`
- `ssh-copy-id -i ~/.ssh/id_rsa.pub datanode01.lab`
- `ssh-copy-id -i ~/.ssh/id_rsa.pub datanode02.lab`
- `ssh-copy-id -i ~/.ssh/id_rsa.pub datanode03.lab`

CDH

- Distribuição do Hadoop da Cloudera
 - Cloudera's Distribution Including Apache Hadoop
- Facilita a implantação de sistemas de big data
- Integra o Hadoop com dezenas de outras ferramentas
- Atualmente na Versão 5.15
- Um dos principais componentes é o Cloudera Manager
 - Responsável pela manutenção do cluster

Cloudera Manager

- Ferramenta de administração do CDH, o Cloudera Manager centraliza a operação do cluster em uma única interface web
- Permite o gerenciamento de cada um dos servidores e de seus serviços
 - CPU, memória, disco, rede, alertas, gráficos etc.
- Composto de 2 componentes:
 - Cloudera Manager Server: instalado no master, ele gerencia os serviços globais do cluster e os agentes
 - Cloudera Manager Agent: instalado em cada um dos servidores, ele gerencia os serviços locais da máquina

Preparação do ambiente

- # Repositório (todos os nós como root)
- `wget 'https://archive.cloudera.com/cm5/ubuntu/xenial/amd64/cm/cloudera.list'`
`-O /etc/apt/sources.list.d/cloudera.list`
- `wget http://archive.cloudera.com/cdh5/ubuntu/xenial/amd64/cdh/archive.key`
- `apt-key add archive.key`
- `apt-get update`
- `apt-get -y install oracle-j2sdk1.7 libmysql-java ntp`
- `echo 10 > /proc/sys/vm/swappiness`
- `sysctl -w vm.swappiness=10`

Preparação do banco de dados

- No MariaDB

- grant all privileges on *.* to 'root'@'%'
- identified by 'root'
- with grant option;
- flush privileges;
- grant all on *.* to 'scm'@'%' identified by 'scm' with grant option;
- grant all on *.* to 'amon'@'%' identified by 'amon' with grant option;
- grant all on *.* to 'hive'@'%' identified by 'hive' with grant option;
- grant all on *.* to 'hue'@'%' identified by 'hue' with grant option;
- grant all on *.* to 'rman'@'%' identified by 'rman' with grant option;
- grant all on *.* to 'oozie'@'%' identified by 'oozie' with grant option;
- #
- CREATE DATABASE scm DEFAULT CHARACTER SET utf8 DEFAULT COLLATE utf8_general_ci;
- CREATE DATABASE amon DEFAULT CHARACTER SET utf8 DEFAULT COLLATE utf8_general_ci;
- CREATE DATABASE hive DEFAULT CHARACTER SET utf8 DEFAULT COLLATE utf8_general_ci;
- CREATE DATABASE hue DEFAULT CHARACTER SET utf8 DEFAULT COLLATE utf8_general_ci;
- CREATE DATABASE rman DEFAULT CHARACTER SET utf8 DEFAULT COLLATE utf8_general_ci;
- CREATE DATABASE oozie DEFAULT CHARACTER SET utf8 DEFAULT COLLATE utf8_general_ci;

Drop do banco

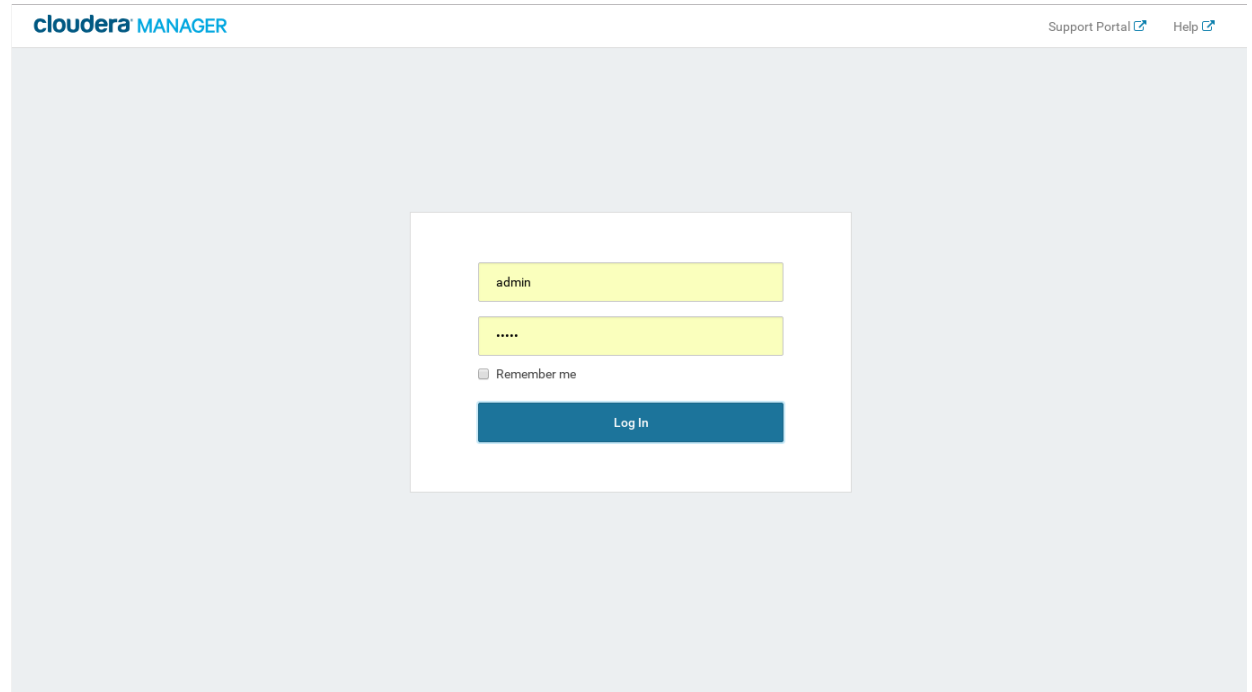
- Se precisar refazer a instalação, remova as tabelas:
 - drop database amon;
 - drop database hive;
 - drop database hue;
 - drop database rman;
 - drop database oozie;
 - drop database scm;

Instalação do Cloudera Manager Server

- Instalar no headnode:
 - `apt-get install cloudera-manager-daemons cloudera-manager-server`
- Preparar as tabelas do Cloudera Manager Server (SCM Server)
 - `/usr/share/cmfschema/scm_prepare_database.sh mysql \`
 - `-h mariadbserver \`
 - `--scm-host headnode01.lab \`
 - `scm scm scm`
- Iniciar o serviço (demora alguns minutos)
 - `service cloudera-scm-server start`
- Monitorar o log
 - `tail -f /var/log/cloudera-scm-server/cloudera-scm-server.log`
- Acesse a URL para concluir a instalação com o Wizard:
 - `http://headnode.lab:7180/`

Login

- Usuário
 - admin
- Senha
 - admin



The image shows the Cloudera Manager login interface. At the top, the header bar contains the "cloudera MANAGER" logo on the left and links for "Support Portal" and "Help" on the right. The main content area is a light gray rectangle. In the center of this area is a white rectangular box containing the login form. The form consists of two yellow input fields: the first is labeled "admin" and the second is labeled ".....". Below these fields is a checkbox labeled "Remember me". At the bottom of the white box is a blue button labeled "Log In".

Licença

- Aceite da licença do Cloudera Manager

Welcome to Cloudera Manager

End User License Terms and Conditions

Cloudera Standard License

Version 2016-05-26

END USER LICENSE TERMS AND CONDITIONS

THESE TERMS AND CONDITIONS (THESE "TERMS") APPLY TO YOUR USE OF THE PRODUCTS (AS DEFINED BELOW) PROVIDED BY CLOUDERA, INC. ("CLOUDERA").

PLEASE READ THESE TERMS CAREFULLY.

IF YOU ("YOU" OR "CUSTOMER") PLAN TO USE ANY OF THE PRODUCTS ON BEHALF OF A COMPANY OR OTHER ENTITY, YOU REPRESENT THAT YOU ARE THE EMPLOYEE OR AGENT OF SUCH COMPANY (OR OTHER ENTITY) AND YOU HAVE THE AUTHORITY TO ACCEPT ALL OF THE TERMS AND CONDITIONS SET FORTH IN AN ACCEPTED REQUEST (AS DEFINED BELOW) AND THESE TERMS (COLLECTIVELY, THE "AGREEMENT") ON BEHALF OF SUCH COMPANY (OR OTHER ENTITY).

BY USING ANY OF THE PRODUCTS, YOU ACKNOWLEDGE AND AGREE THAT:
(A) YOU HAVE READ ALL OF THE TERMS AND CONDITIONS OF THIS AGREEMENT;
(B) YOU UNDERSTAND ALL OF THE TERMS AND CONDITIONS OF THIS AGREEMENT;
(C) YOU AGREE TO BE LEGALLY BOUND BY ALL OF THE TERMS AND CONDITIONS SET FORTH IN THIS AGREEMENT

IF YOU DO NOT AGREE WITH ANY OF THE TERMS OR CONDITIONS OF THESE TERMS, YOU MAY NOT USE ANY PORTION OF THE PRODUCTS.

☒ Yes, I accept the End User License Terms and Conditions.

If your download and use of Cloudera Manager are on behalf of a company that has an existing agreement with Cloudera for the use of the software, your action does not modify that existing agreement.

Back

1 2

Continue

Edição

- Selecione a edição Cloudera Express

Welcome to Cloudera Manager

Which edition do you want to deploy?

Upgrading to **Cloudera Enterprise** provides important features that help you manage and monitor your Hadoop clusters in mission-critical environments.

Cloudera Express		Cloudera Enterprise Cloudera Enterprise Trial ✓	Cloudera Enterprise
License	Free	60 Days After the trial period, the product will continue to function as Cloudera Express . Your cluster and your data will remain unaffected.	Annual Subscription Upload License Key <div>Select License File Upload</div> Cloudera Enterprise is available in three editions: <ul style="list-style-type: none">Basic EditionFlex EditionCloudera Enterprise
Node Limit	Unlimited	Unlimited	Unlimited
CDH	✓	✓	✓
Core Cloudera Manager Features	✓	✓	✓
Advanced Cloudera Manager Features		✓	✓
Cloudera Navigator		✓	✓
Cloudera Navigator Key Trustee			✓
Cloudera Support			✓

See [full list of features available](#) in Cloudera Express and Cloudera Enterprise.

Pacotes

- Lista de pacotes disponíveis para instalação no CDH

cloudera MANAGER

Support ▾ admin ▾

Thank you for choosing Cloudera Manager and CDH.

This installer will install **Cloudera Enterprise Trial 5.15.0** and enable you to later choose packages for the services below (there may be some license implications).

- Apache Hadoop (Common, HDFS, MapReduce, YARN)
- Apache HBase
- Apache ZooKeeper
- Apache Oozie
- Apache Hive
- Hue (Apache licensed)
- Apache Flume
- Apache Impala
- Apache Sentry
- Apache Sqoop
- Cloudera Search (Apache licensed)
- Apache Spark

You are using Cloudera Manager to install and configure your system. You can learn more about Cloudera Manager by clicking on the **Support** menu above.

Before you proceed, be sure to checkout the [CDH and Cloudera Manager Requirements and Supported Versions](#)

- [Supported Operating Systems](#)
- [Supported Databases](#)
- [Supported JDK Versions](#)

Continue


Seleção de hosts

- Informe os servidores componentes do cluster
- Permite o uso de padrões
 - headnode.lab, edgenode.lab, datanode[01-03].lab

Specify hosts for your CDH cluster installation.

Hosts should be specified using the same hostname (FQDN) that they will identify themselves with.

Cloudera recommends including Cloudera Manager Server's host. This also enables health monitoring for that host.

Hint: Search for hostnames and IP addresses using [patterns](#) .

headnode.lab, edgenode.lab, datanode[01-03].lab

SSH Port:

22

Search


Hosts disponíveis

- Lista dos hosts encontrados

Specify hosts for your CDH cluster installation.

Hosts should be specified using the same hostname (FQDN) that they will identify themselves with.

Cloudera recommends including Cloudera Manager Server's host. This also enables health monitoring for that host.

Hint: Search for hostnames and IP addresses using [patterns](#) .

5 hosts scanned, 5 running SSH:

New Search

<input checked="" type="checkbox"/> Expanded Query	Hostname (FQDN)	IP Address	Currently Managed	Result
<input checked="" type="checkbox"/> datanode01.lab	datanode01.lab	192.168.25.191	No	✓ Host ready: 10 ms response time.
<input checked="" type="checkbox"/> datanode02.lab	datanode02.lab	192.168.25.192	No	✓ Host ready: 8 ms response time.
<input checked="" type="checkbox"/> datanode03.lab	datanode03.lab	192.168.25.193	No	✓ Host ready: 9 ms response time.
<input checked="" type="checkbox"/> edgenode.lab	edgenode.lab	192.168.25.194	No	✓ Host ready: 1 ms response time.
<input checked="" type="checkbox"/> headnode.lab	headnode.lab	192.168.25.190	No	✓ Host ready: 9 ms response time.

Seleção dos repositórios

- Métodos:
 - Pacotes
 - Parcels

Cluster Installation

Select Repository

Cloudera recommends the use of parcels for installation over packages, because parcels enable Cloudera Manager to easily manage the software on your cluster, automating the deployment and upgrade of service binaries. Electing not to use parcels will require you to manually upgrade packages on all hosts in your cluster when software updates are available, and will prevent you from using Cloudera Manager's rolling upgrade capabilities.

Choose Method ☐ Use Packages ?

☒ Use Parcels (Recommended) ?

[More Options](#)

[Proxy Settings](#)

CDH Version ☒ CDH-5.15.0-1.cdh5.15.0.p0.21

☐ CDH-4.7.1-1.cdh4.7.1.p0.47

Versions of CDH that are too new for this version of Cloudera Manager (5.15.0) will not be shown.

Additional Parcels ☐ ACCUMULO-1.7.2-5.5.0.ACCUMULO5.5.0.p0.8

☐ ACCUMULO-1.4.4-1.cdh4.5.0.p0.65

☒ None

☐ IMPALA-2.1.0-1.impala2.0.0.p0.1995

☒ None

☐ KAFKA-3.1.0-1.3.1.0.p0.35

☒ None

[Back](#)

1 2 3 4 5 6 7

[Continue](#)

Licença do JDK

- Aceite da licença do JDK

Cluster Installation

Accept JDK License

Oracle Binary Code License Agreement for the Java SE Platform Products and JavaFX

ORACLE AMERICA, INC. ("ORACLE"), FOR AND ON BEHALF OF ITSELF AND ITS SUBSIDIARIES AND AFFILIATES UNDER COMMON CONTROL, IS WILLING TO LICENSE THE SOFTWARE TO YOU ONLY UPON THE CONDITION THAT YOU ACCEPT ALL OF THE TERMS CONTAINED IN THIS BINARY CODE LICENSE AGREEMENT AND SUPPLEMENTAL LICENSE TERMS (COLLECTIVELY "AGREEMENT"). PLEASE READ THE AGREEMENT CAREFULLY. BY SELECTING THE "ACCEPT LICENSE AGREEMENT" (OR THE EQUIVALENT) BUTTON AND/OR BY USING THE SOFTWARE YOU ACKNOWLEDGE THAT YOU HAVE READ THE TERMS AND AGREE TO THEM. IF YOU ARE AGREEING TO THESE TERMS ON BEHALF OF A COMPANY OR OTHER LEGAL ENTITY, YOU REPRESENT THAT YOU HAVE THE LEGAL AUTHORITY TO BIND THE LEGAL ENTITY TO THESE TERMS. IF YOU DO NOT HAVE SUCH AUTHORITY, OR IF YOU DO NOT WISH TO BE BOUND BY THE TERMS, THEN SELECT THE "DECLINE LICENSE AGREEMENT" (OR THE EQUIVALENT) BUTTON AND YOU MUST NOT USE THE SOFTWARE ON THIS SITE OR ANY OTHER MEDIA ON WHICH THE SOFTWARE IS CONTAINED.

1. DEFINITIONS. "Software" means the software identified above in binary form that you selected for download, install or use (in the version You selected for download, install or use) from Oracle or its authorized licensees, any other machine readable materials (including, but not limited to, libraries, source files, header files, and data files), any updates or error corrections provided by Oracle, and any user manuals, programming guides and other documentation provided to you by Oracle under this Agreement. "General Purpose Desktop Computers and Servers" means computers, including desktop and laptop computers, or servers, used for general computing functions under end user control (such as but not specifically limited to email, general purpose Internet browsing, and office suite productivity tools). The use of Software in systems and solutions that provide dedicated functionality (other than as mentioned above) or designed for use in embedded or function-specific software applications, for example but not limited to: Software embedded in or bundled with industrial control systems, wireless mobile telephones, wireless handheld devices, kiosks, TV/STB, Blu-ray Disc devices, telematics and network control switching equipment, printers and storage management systems, and other related systems are excluded from this definition and not licensed under this Agreement. "Programs" means (a) Java technology applets and applications intended to run on the Java Platform, Standard Edition platform on Java-enabled General Purpose Desktop Computers and Servers; and (b) JavaFX technology applications intended to run on the JavaFX Runtime on JavaFX-enabled General Purpose Desktop Computers and Servers. "Commercial Features" means those features identified in

☒ Install Oracle Java SE Development Kit (JDK 7)

Check this box to accept the Oracle Binary Code License Agreement and install the JDK. Leave it unchecked to use a currently installed JDK.

☐ Install Java Unlimited Strength Encryption Policy Files

Check this checkbox if local laws permit you to deploy unlimited strength encryption and you are running a secure cluster.

Back

1 2 3 4 5 6 7

Continue

Modo de Usuário

- Distinct users
- Single user

Cluster Installation

Single User Mode

Only supported for CDH 5.2 and above.

By default, service processes run as distinct users on the system. For example, HDFS DataNodes run as user "hdfs" and HBase RegionServers run as user "hbase." Enabling "single user mode" configures Cloudera Manager to run service processes as a single user, by default "cloudera-scm", thereby prioritizing isolation between managed services and the rest of the system over isolation between the managed services.

The **major benefit** of this option is that the Agent does not run as root. However, this mode complicates installation, which is described fully in the [documentation](#). Most notably, directories which in the regular mode are created automatically by the Agent, must be created manually on every host with appropriate permissions, and sudo (or equivalent) access must be set up for the configured user.

Switching back and forth between single user mode and regular mode is not supported.

Enable Single User Mode ☐



Back

1 2 3 4 5 6 7

Continue

Credenciais de login

- Informe a senha de root ou de outro usuário com acesso ao sudo de cada um dos servidores no cluster
- Cada uma das máquinas deve ter o mesmo usuário

Cluster Installation

Enter Login Credentials

Root access to your hosts is required to install the Cloudera packages. This installer will connect to your hosts via SSH and log in either directly as root or as another user with password-less sudo/pbrun privileges to become root.

Login To All Hosts As: ☒ root
☐ Another user

You may connect via password or public-key authentication for the user selected above.

Authentication Method: ☒ All hosts accept same password
☐ All hosts accept same private key

Enter Password:

Confirm Password:

SSH Port:

Number of Simultaneous Installations:
(Running a large number of installations at once can consume large amounts of network bandwidth and other system resources)

Back

1 2 3 4 5 6 7

Continue

Instalação dos agentes

- Instalação dos agentes em cada servidor do cluster

Cluster Installation

Install Agents

Installation in progress.



0 of 5 host(s) completed successfully. [Abort Installation](#)

Hostname	IP Address	Progress	Status
datanode01.lab	192.168.25.191	<div><div></div></div>	<div><div></div> Installing cloudera-manager-agent package... Details </div>
datanode02.lab	192.168.25.192	<div><div></div></div>	<div><div></div> Installing cloudera-manager-agent package... Details </div>
datanode03.lab	192.168.25.193	<div><div></div></div>	<div><div></div> Installing cloudera-manager-agent package... Details </div>
edgenode.lab	192.168.25.194	<div><div></div></div>	<div><div></div> Installing cloudera-manager-agent package... Details </div>
headnode.lab	192.168.25.190	<div><div></div></div>	<div><div></div> Installing cloudera-manager-agent package... Details </div>

Detalhes da instalação dos agentes

- Clique na opção Details para ver o log da instalação

```
Installing cloudera-manager-agent package... View
```

```
0ubuntu0.16.04.1 [15.4 kB]
Get:23 http://179.184.208.141:80/data/0514ae68f8b8c020/us.archive.ubuntu.com/ubuntu xenial-updates/main amd64 libmysqlclient20 amd64 5.7.23-0ubuntu0.16.04.1 [812 kB]
Get:24 http://179.184.208.141:80/data/0514856811b96821/us.archive.ubuntu.com/ubuntu xenial-updates/main amd64 libpq5 amd64 9.5.13-0ubuntu0.16.04 [78.7 kB]
Get:26 http://179.184.208.141:80/data/0514a6687cbce110/us.archive.ubuntu.com/ubuntu xenial-updates/main amd64 zlib1g-dev amd64 1:1.2.8.dfsg-2ubuntu4.1 [168 kB]
Get:27 http://179.184.208.141:80/data/05140c68fcbdf71e/us.archive.ubuntu.com/ubuntu xenial-updates/main amd64 libssl-dev amd64 1.0.2g-1ubuntu4.13 [1342 kB]
Get:28 http://179.184.208.141:80/data/05141d6862be5059/us.archive.ubuntu.com/ubuntu xenial-updates/main amd64 libssl-doc all 1.0.2g-1ubuntu4.13 [1079 kB]
Get:33 http://us.archive.ubuntu.com/ubuntu xenial/main amd64 python-psycopg2 amd64 2.6.1-1build2 [131 kB]
Get:29 http://179.184.208.141:80/data/0514c568eabfca5b/us.archive.ubuntu.com/ubuntu xenial/main amd64 manpages-dev all 4.04-2 [2048 kB]
Get:30 http://179.184.208.141:80/data/0514656852c0887c/us.archive.ubuntu.com/ubuntu xenial/main amd64 python-egenix-mxtools amd64 3.2.9-1 [75.3 kB]
Get:31 http://179.184.208.141:80/data/05147b6888c1f3a3/us.archive.ubuntu.com/ubuntu xenial/main amd64 python-egenix-mxdatetime amd64 3.2.9-1 [68.3 kB]
Get:32 http://179.184.208.141:80/data/05146d688ec2a1ce/us.archive.ubuntu.com/ubuntu xenial/main amd64 python-mysqldb amd64 1.3.7-1build2 [42.4 kB]
Get:34 http://179.184.208.141:80/data/0514516896c7c890/us.archive.ubuntu.com/ubuntu xenial/main amd64 rpcbind amd64 0.2.3-0.2 [40.3 kB]
Get:35 http://179.184.208.141:80/data/0514e8689bc85bbe/us.archive.ubuntu.com/ubuntu xenial/main amd64 ssl-cert all 1.0.37 [16.9 kB]
```

Last Refreshed: Aug 7, 2018 9:51:43 AM BRT [Turn Off Auto Refresh](#)

Close

Instalação dos agentes concluída






- Conclusão da instalação com sucesso
- Em caso de erro, conserte os problemas e clique na opção Retry

Cluster Installation

Install Agents

Installation completed successfully.

5 of 5 host(s) completed successfully.

Hostname	IP Address	Progress	Status	
datanode01.lab	192.168.25.191	<div></div>	✓ Installation completed successfully.	Details 
datanode02.lab	192.168.25.192	<div></div>	✓ Installation completed successfully.	Details 
datanode03.lab	192.168.25.193	<div></div>	✓ Installation completed successfully.	Details 
edgenode.lab	192.168.25.194	<div></div>	✓ Installation completed successfully.	Details 
headnode.lab	192.168.25.190	<div></div>	✓ Installation completed successfully.	Details 

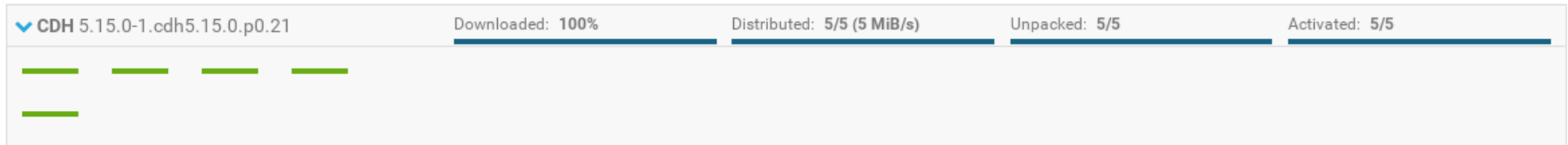
Instalação das parcels

- Aguarde o download e instalação dos componentes (parcels)

Cluster Installation

Install Parcels

The selected parcels are being downloaded and installed on all the hosts in the cluster.




Inspeção dos hosts

- Aguarde alguns instantes até terminar a inspeção

Cluster Installation

Inspect hosts for correctness

Inspecting hosts... This could take a minute. 

[Skip Host Inspector](#)

Inspeção dos hosts concluída

- A inspeção foi concluída sem erros ou alertas

Cluster Installation

Inspect hosts for correctness [Run Again](#)

Validations

✔	Inspector ran on all 5 hosts.
✔	Individual hosts resolved their own hostnames correctly.
✔	No errors were found while looking for conflicting init scripts.
✔	No errors were found while checking /etc/hosts.
✔	All hosts resolved localhost to 127.0.0.1.
✔	All hosts checked resolved each other's hostnames correctly and in a timely manner.
✔	Host clocks are approximately in sync (within ten minutes).
✔	Host time zones are consistent across the cluster.
✔	No users or groups are missing.
✔	No conflicts detected between packages and parcels.
✔	No kernel versions that are known to be bad are running.
✔	No problems were found with /proc/sys/vm/swappiness on any of the hosts.
✔	No performance concerns with Transparent Huge Pages settings.
✔	CDH 5 Hue Python version dependency is satisfied.
✔	0 hosts are running CDH 4 and 5 hosts are running CDH 5.
✔	All checked hosts in each cluster are running the same version of components.
✔	All managed hosts have consistent versions of Java.
✔	All checked Cloudera Management Daemons versions are consistent with the server.
✔	All checked Cloudera Management Agents versions are consistent with the server.

Seleção dos serviços

- Core Hadoop
 - Serviços principais
- All Services
 - Todos os serviços
- Custom Services
 - Escolha personalizada

Cluster Setup

Select Services

Choose a combination of services to install.

☒ **Core Hadoop**

HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, and Hue

☐ **Core with HBase**

HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, and HBase

☐ **Core with Impala**

HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, and Impala

☐ **Core with Search**

HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, and Solr

☐ **Core with Spark**

HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, and Spark

☐ **All Services**

HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, HBase, Impala, Solr, Spark, and Key-Value Store Indexer

☐ **Custom Services**

Choose your own services. Services required by chosen services will automatically be included. Flume can be added after your initial cluster has been set up.

This wizard will also install the **Cloudera Management Service**. These are a set of components that enable monitoring, reporting, events, and alerts; these components require databases to store information, which will be configured on the next page.

☐ Include Cloudera Navigator

Back

1 2 3 4 5 6

Continue

Atribuição dos papéis (roles)

- Distribuir entre os hosts
 - HDFS
 - Hive
 - Hue
 - CMS
 - Oozie
 - YARN
 - ZooKeeper

HDFS			
NameNode x 1 New headnode.lab	SecondaryNameNode x 1 New edgenode.lab	Balancer x 1 New edgenode.lab	HttpFS x 1 New edgenode.lab
NFS Gateway Select hosts	DataNode x 3 New datanode[01-03].lab		
Hive			
Gateway x 1 New edgenode.lab	Hive Metastore Server x 1 New edgenode.lab	WebHCat Server Select hosts	HiveServer2 x 1 New edgenode.lab
Hue			
Hue Server x 1 New edgenode.lab	Load Balancer x 1 New edgenode.lab		
Cloudera Management Service			
Service Monitor x 1 New headnode.lab	Activity Monitor x 1 New headnode.lab	Host Monitor x 1 New headnode.lab	Reports Manager x 1 New headnode.lab
Event Server x 1 New headnode.lab	Alert Publisher x 1 New headnode.lab	Telemetry Publisher Select a host	
Oozie			
Oozie Server x 1 New edgenode.lab			
YARN (MR2 Included)			
ResourceManager x 1 New headnode.lab	JobHistory Server x 1 New edgenode.lab	NodeManager x 3 New Same As DataNode	
ZooKeeper			
Server x 3 New datanode01.lab; edgenode.lab; head...			

Configurar o banco de dados

- MariaDB
 - Usuário e senha

Cluster Setup

Setup Database

Configure and test database connections. Create the databases first according to the **Installing and Configuring an External Database** section of the [Installation Guide](#).

Hive

Database Host Name: *	Database Type:	Database Name: *	Username: *	Password:
<input type="text" value="mariadbserver"/>	<input type="text" value="MySQL"/>	<input type="text" value="hive"/>	<input type="text" value="hive"/>	<input type="text" value="hive"/>

Hue

Database Host Name: *	Database Type:	Database Name: *	Username: *	Password:
<input type="text" value="mariadbserver"/>	<input type="text" value="MySQL"/>	<input type="text" value="hue"/>	<input type="text" value="hue"/>	<input type="text" value="hue"/>

Activity Monitor

Currently assigned to run on **headnode01.lab.**

Database Host Name: *	Database Type:	Database Name: *	Username: *	Password:
<input type="text" value="mariadbserver"/>	<input type="text" value="MySQL"/>	<input type="text" value="amon"/>	<input type="text" value="amon"/>	<input type="text" value="amon"/>

Reports Manager

Currently assigned to run on **headnode01.lab.**

Database Host Name: *	Database Type:	Database Name: *	Username: *	Password:
<input type="text" value="mariadbserver"/>	<input type="text" value="MySQL"/>	<input type="text" value="rman"/>	<input type="text" value="rman"/>	<input type="text" value="rman"/>

Oozie Server

Currently assigned to run on **headnode01.lab.**

Database Host Name: *	Database Type:	Database Name: *	Username: *	Password:
<input type="text" value="mariadbserver"/>	<input type="text" value="MySQL"/>	<input type="text" value="oozie"/>	<input type="text" value="oozie"/>	<input type="text" value="oozie"/>

☒ Show Password

[Test Connection](#)

Rever mudanças

- Configurações do CDH
 - Tamanho do bloco
 - Tolerância a falhas
 - Diretório de dados
 - Diretório de metadados
 - Etc.

Cluster Setup

Review Changes

HDFS Block Size
dfs.block.size, dfs.blocksize

Cluster 1 > HDFS (Service-Wide)

128

MiB

DataNode Failed Volumes
Tolerated
dfs.datanode.failed.volumes.tol
erated

Cluster 1 > DataNode Default Group

0

DataNode Data Directory
dfs.data.dir,
dfs.datanode.data.dir

Cluster 1 > DataNode Default Group

/dfs/dn

NameNode Data
Directories
dfs.name.dir,
dfs.namenode.name.dir

Cluster 1 > NameNode Default Group

/dfs/nn

HDFS Checkpoint
Directories
fs.checkpoint.dir,
dfs.namenode.checkpoint.dir

Cluster 1 > SecondaryNameNode Default Group

/dfs/snn

Hive Warehouse Directory
hive.metastore.warehouse.dir

Cluster 1 > Hive (Service-Wide)

/user/hive/warehouse

Primeira execução

- Instalação do cluster concluída

Cluster Setup











First Run Command

Status  **Finished**  Aug 7, 11:37:41 PM  6.8m

Finished First Run of the following services successfully: ZooKeeper, HDFS, YARN (MR2 Included), Hive, Oozie, Hue, Cloudera Management Service.

 Completed 8 of 8 step(s).

☒ Show All Steps ☐ Show Only Failed Steps ☐ Show Running Steps

> 	Ensuring that the expected software releases are installed on hosts.	Aug 7, 11:37:41 PM	218ms
> 	Deploying Client Configuration  Cluster 1 	Aug 7, 11:37:41 PM	16.12s
> 	Start Cloudera Management Service, ZooKeeper	Aug 7, 11:37:57 PM	27.19s
> 	Start HDFS	Aug 7, 11:38:24 PM	98.99s
> 	Start YARN (MR2 Included)	Aug 7, 11:40:03 PM	30.49s
> 	Start Hive	Aug 7, 11:40:34 PM	70.55s
> 	Start Oozie	Aug 7, 11:41:44 PM	2.4m
> 	Start Hue	Aug 7, 11:44:08 PM	23.15s

Instalação concluída

- Todos os serviços foram instalados com sucesso

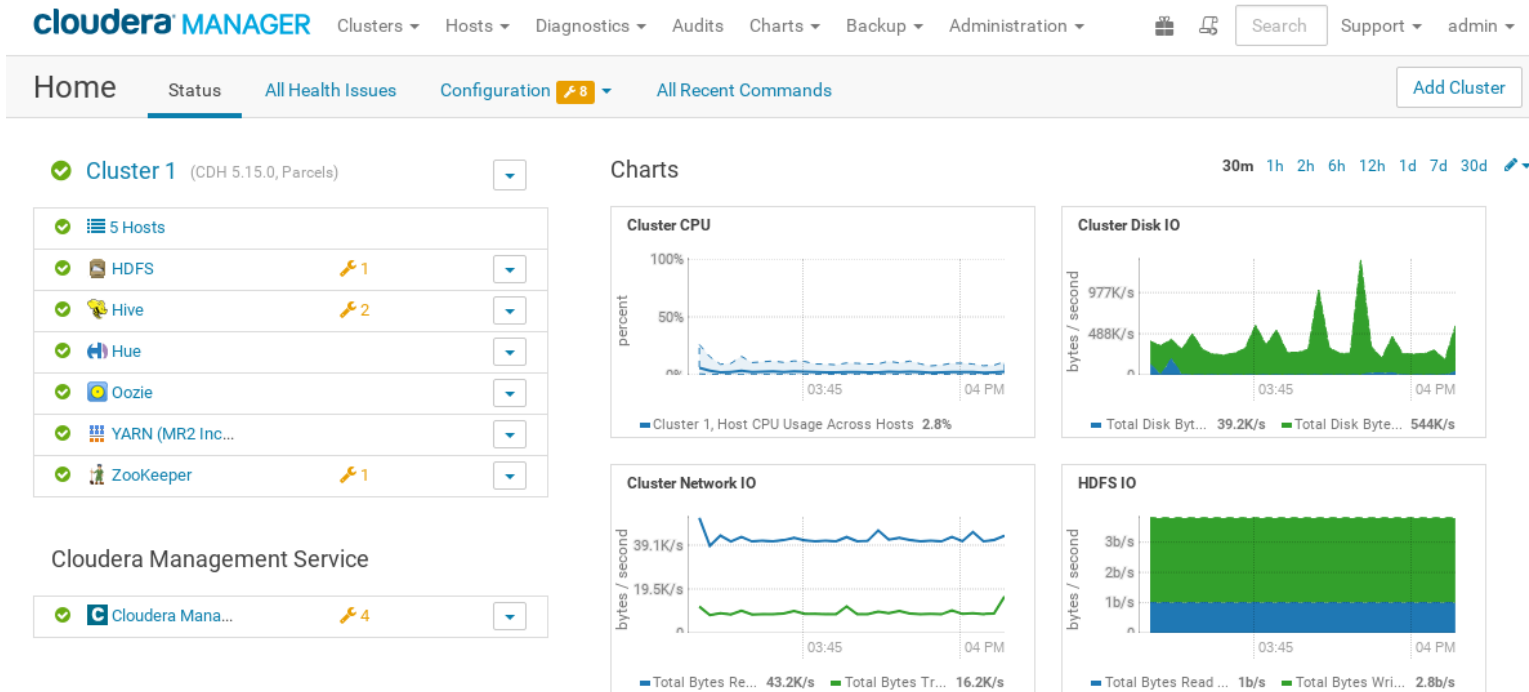
Cluster Setup

Congratulations!

✔ The services are installed, configured, and running on your cluster.

Cloudera Manager

- Bem-vindo ao Cloudera Manager!



Instalação de novos serviços

- Vamos instalar outros serviços no cluster
 - Impala
 - Flume
 - Sqoop

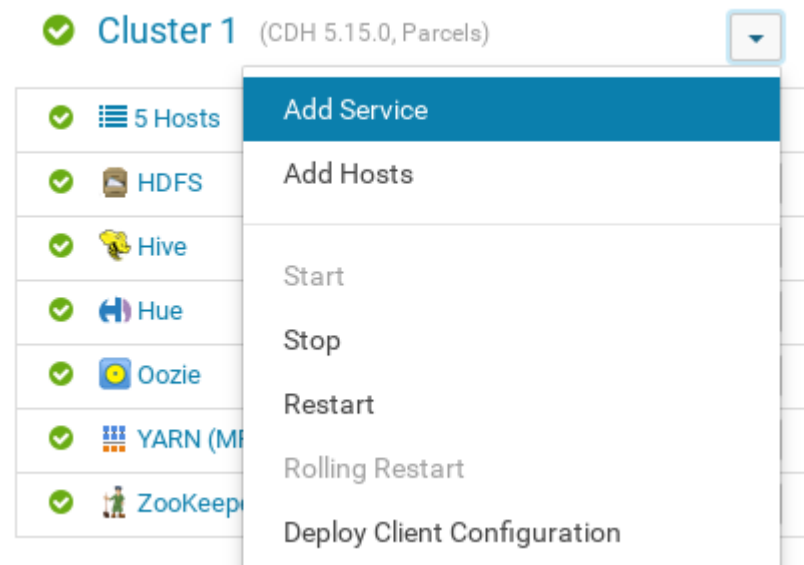
Impala

- Mecanismo de consulta para o Hadoop com suporte a SQL
- Consultas em tempo real
- Evita o MapReduce e acessa os dados com um mecanismo especializado e distribuído de consulta



Instalação do Impala

- O Impala usa a metastore do Hive
- Adicione um novo serviço:
 - Add Service



Seleção do serviço do Impala

- Selecione o serviço indicado



Impala

Impala provides a real-time SQL query interface for data stored in HDFS and HBase. Impala requires the Hive service and shares the Hive Metastore with Hue.

Seleção dos hosts do Impala

- Os novos serviços serão adicionados no EdgeNode e as daemons nos datanodes

Add Impala Service to Cluster 1

Assign Roles for Impala

You can customize the role assignments for your new service here, but note that if assignments are made incorrectly, such as assigning too many roles to a single host, performance will suffer.

You can also view the role assignments by host.

View By Host

 Impala Catalog Server x 1 New

edgenode.lab ▼

 Impala StateStore x 1 New

edgenode.lab ▼

 Impala Daemon x 3 New

datanode[01-03].lab ▼

Revisão das configurações

- Não alterar os valores

Add Impala Service to Cluster 1

Review Changes

Kudu Service

Impala (Service-Wide)



☒ none

Impala Daemon Scratch

Impala Daemon Default Group 



Directories

scratch_dirs

/impala/impalad



Primeira execução do serviço

- O serviço foi instalado com sucesso
- Provavelmente você precisará iniciar o serviço manualmente da primeira vez

Add Impala Service to Cluster 1







First Run Command

Status  **Finished**  Aug 16, 4:57:53 PM  9.84s

Finished First Run of the following services successfully: Impala.

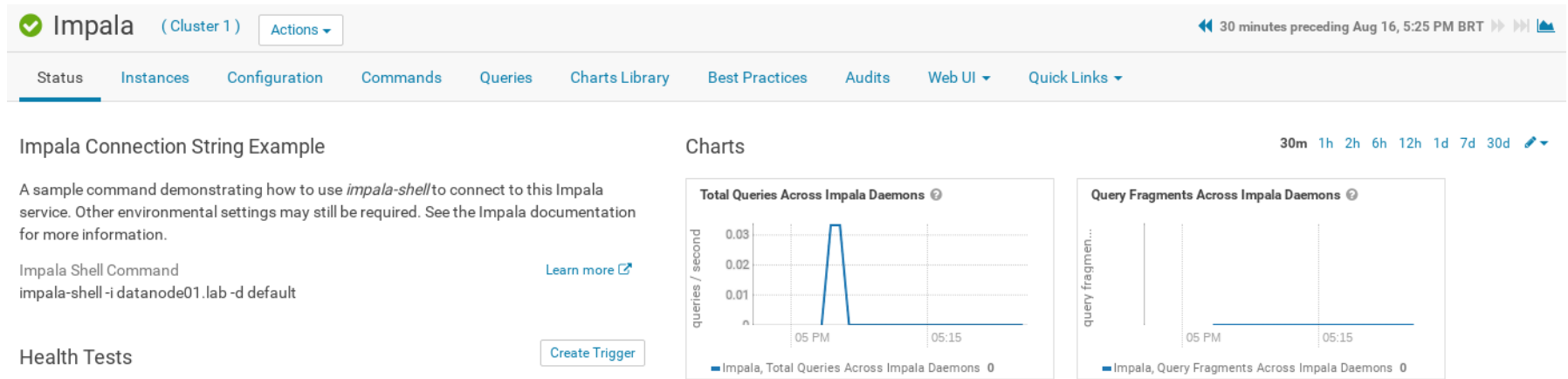
 Completed 2 of 2 step(s).

☒ Show All Steps ☐ Show Only Failed Steps ☐ Show Running Steps

  Ensuring that the expected software releases are installed on hosts.	Aug 16, 4:57:53 PM	186ms
  Creating Impala user directory	 Impala 	Aug 16, 4:57:53 PM 9.6s

Tela inicial do Impala

- Tela inicial



Ativar o Impala no Hue

- Para executar consultas com o Impala no Hue, precisamos habilitar essa opção
- Reinicie o Hue

The screenshot shows the Hue web interface. At the top, there's a header with a green checkmark, the word 'Hue', and '(Cluster 1)'. Below this is a navigation bar with tabs: 'Status', 'Instances', 'Configuration' (which is active), 'Commands', 'Charts Library', 'Audits', and 'Web UI'. A search bar contains the text 'impala'. On the left side, there's a 'Filters' section with a 'SCOPE' dropdown. The 'SCOPE' list shows: 'Hue (Service-Wide)' with a count of 1, 'Hue Server' with 2, 'Kerberos Ticket Renewer' with 0, and 'Load Balancer' with 0. The main content area is titled 'Impala Service'. It has two sections: 'Hue (Service-Wide)' with radio buttons for 'Impala' (selected) and 'none'; and 'HiveServer2 and Impala' with a 'Thrift Connection Timeout' field set to '2' and a unit dropdown set to 'minute'.

Hue (Cluster 1) Actions

Status Instances **Configuration** Commands Charts Library Audits Web UI

impala

Filters

▼ SCOPE

Hue (Service-Wide)	1
Hue Server	2
Kerberos Ticket Renewer	0
Load Balancer	0

Impala Service

Hue (Service-Wide) ↻

☒ Impala

☐ none

HiveServer2 and Impala

Thrift Connection Timeout

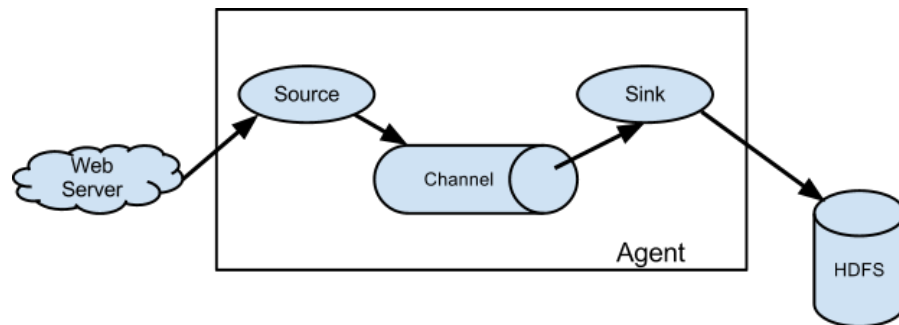
server_conn_timeout

2 minute ▼



Flume

- Apache Flume é um framework para agregação e coleta de dados em tempo real
- As operações do Flume são realizadas por um agente, que é composto por:
 - Source: fonte de dados, como o log de servidor, JMS, NetCat, Twitter ou outro agente Flume
 - Channel: mecanismo que recebe os dados da fonte e os grava no destino. O canal pode ser em memória, JDBC, em arquivo etc.
 - Sink: destino onde os dados serão gravados, que pode ser o HDFS, HBase, Hive, logger, Avro etc.



Instalação do Flume

- Adicione o novo serviço da lista



Flume collects and aggregates data from almost any source into a persistent store such as HDFS.

- Instale o agente do Flume no EdgeNode
- Provavelmente você terá de iniciar o serviço manualmente da primeira vez

Add Flume Service to Cluster 1

Assign Roles for Flume

You can customize the role assignments for your new service here, but note that if assignments are made incorrectly, such as assigning too many roles to a single host, performance will suffer.

You can also view the role assignments by host.

[View By Host](#)

Agent x 1 New

edgenode.lab ▾

Sqoop



- Ferramenta de alta performance para importação e exportação de dados entre o Hadoop e o SGBDR
- O Sqoop usa o MapReduce para as operações e é altamente configurável
- Permite importar uma tabela, uma consulta ou uma database inteira

Instalação do cliente Sqoop

- Selecione o serviço do Sqoop

 Sqoop 1 Client Configuration and connector management for Sqoop 1.


- Selecione o EdgeNode como gateway

Add Sqoop 1 Client Service to Cluster 1

Assign Roles for Sqoop 1 Client

You can customize the role assignments for your new service here, but note that if assignments are made incorrectly, such as assigning too many roles to a single host, performance will suffer.

You can also view the role assignments by host. [View By Host](#)

 Gateway x 1 New

edgenode.lab ▾

Conclusão da instalação

- Não é necessário iniciar o serviço

Add Sqoop 1 Client Service to Cluster 1







First Run Command

Status  **Finished**  Aug 16, 5:54:10 PM  16.25s

Finished First Run of the following services successfully: Sqoop 1 Client.

 Completed 2 of 2 step(s).

☒ Show All Steps ☐ Show Only Failed Steps ☐ Show Running Steps

  Ensuring that the expected software releases are installed on hosts.	Aug 16, 5:54:10 PM	53ms
  Deploying Client Configuration  Cluster 1 	Aug 16, 5:54:10 PM	15.87s

Spark



- O Spark é um mecanismo de alta performance para análise de dados
- Processamento em batch, streaming, grafos e machine learning
- Suporte para linguagem Java, Scala, Python, R e SQL
- Roda em Hadoop, Mesos, Kubernetes, standalone e nuvem
- Acessa fontes de dados diversas
 - HDFS, Cassandra, Hive, HBase

Instalação do Spark

- Adicione o serviço do Spark



Apache Spark is an open source cluster computing system. This service runs Spark as an application on YARN.

- Selecione o history server e os gateways

Add Spark Service to Cluster 1

Assign Roles for Spark

You can customize the role assignments for your new service here, but note that if assignments are made incorrectly, such as assigning too many roles to a single host, performance will suffer.

You can also view the role assignments by host.

[View By Host](#)

★ History Server x 1 New

edgenode.lab ▾

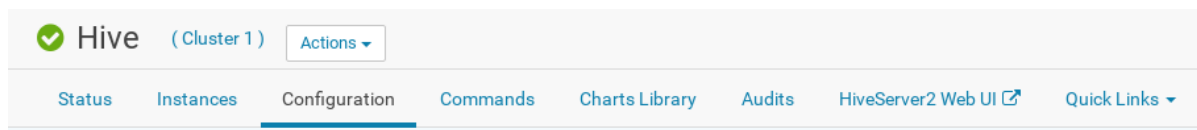
★ Gateway x 3 New

datanode[01-03].lab ▾

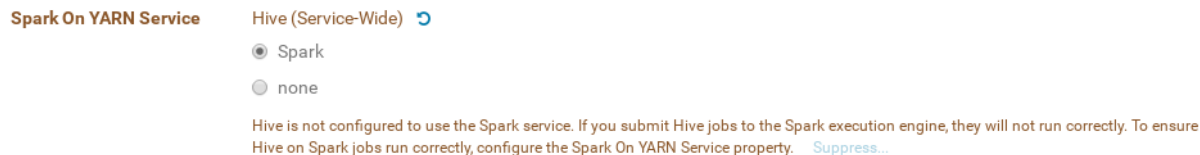
- Iniciar o serviço do Spark

Integração do Spark com o Hive

- Clique no serviço do Hive



- Configure a opção Spark On YARN Service



Serviços instalados

- Lista de serviços instalados no cluster

✓	Cluster 1	(CDH 5.15.0, Parcels)	▼
✓	5 Hosts	🔧 4	
✓	Flume		▼
✓	HDFS	🔧 1	▼
✓	Hive	🔧 2	▼
✓	Hue		▼
✓	Impala		▼
✓	Oozie		▼
✓	Spark		▼
●	Sqoop 1 Client		▼
✓	YARN (MR2 Inc...		▼
✓	ZooKeeper		▼

Dúvidas?

Marco Reis
<http://marcoreis.net>