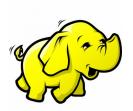
#### Cloudera Administrator Apache Hadoop

Parte 02-1 Clientes Hadoop



Marco Reis http://marcoreis.net

# Agenda



- Clientes Hadoop
  - Máquina não gerenciada
  - Máquina gateway (EdgeNode)
  - Hue

## Clientes Hadoop

- O Hadoop é operado por meio dos clientes (máquina não gerenciada, máquina de gateway e Hue)
- Um cliente é usado para que o usuário não acesse diretamente um HeadNode
- É uma medida de segurança e performance:
  - Evita acessos indevidos ao HeadNode que possam danificar o cluster
  - Evita que uma operação de I/O degrade o HeadNode, por exemplo, uma cópia de arquivos que exige muito do sistema operacional e rede

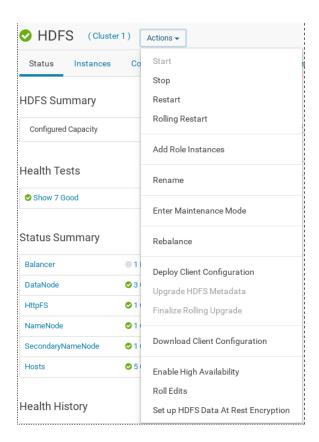
# Máquina não gerenciada



- Um computador que n\u00e3o faz parte do cluster pode ser usado para opera\u00f3\u00f3es no CDH
- A Cloudera disponibiliza um pacote com os programas do Hadoop (hadoop-client)
- A configuração é feita manualmente e as opções são bastante limitadas
  - Suporte apenas para operações do HDFS e YARN
  - Quando a configuração do cluster é alterada, devem ser atualizadas manualmente nos clientes
- Script para instalação:
  - \$ sudo su -
  - \$ wget 'https://archive.cloudera.com/cdh5/ubuntu/xenial/amd64/cdh/cloudera.list' \
  - O /etc/apt/sources.list.d/cloudera.list
  - \$ apt-get update
  - \$ apt-get install hadoop-client hive impala-shell
- Após a instalação são criados novos usuários, como hdfs, mapred, yarn e zookeeper

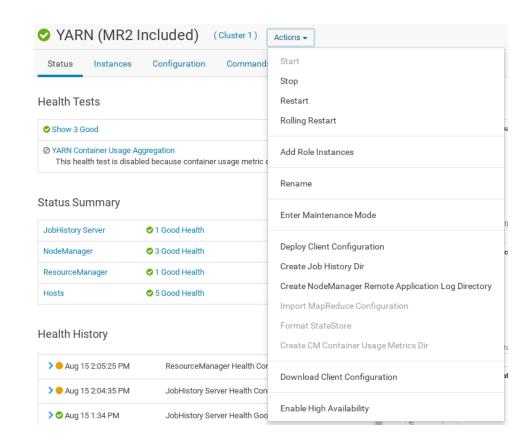
# Configuração do cliente HDFS

- Arquivos de configuração:
  - Download Client Configuration (hdfs-clientconfig.zip)
- Descompactar os arquivos e copiar para o computador cliente no diretório /etc/hadoop/conf
  - O computador cliente é o <u>hadoopclient</u>
  - Antes, faça uma cópia do diretório /etc/hadoop/conf, só por precaução
  - Exemplo: scp \* root@hadoopclient:/etc/hadoop/conf/
- Certificar que os hosts (/etc/hosts) estão configurados no cliente (hadoopclient) e no cluster



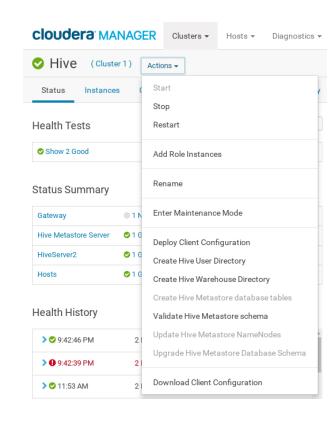
## Configuração do cliente YARN

- Arquivos de configuração:
  - Download Client Configuration (yarnclientconfig.zip)
- Copiar os arquivos mapred-site.xml e yarn-site.xml para o hadoopclient com os comandos:
  - \$ scp mapred-site.xml root@hadoopclient:/etc/hadoop/conf/
  - \$ scp yarn-site.xml root@hadoopclient:/etc/hadoop/conf/



# Configuração do cliente Hive

- Arquivos de configuração
  - Download Client Configuration (hiveclientconfig.zip)
- Copiar os arquivos hive-site.xml e hive-env.sh para o hadoopclient:
  - \$ scp hive-site.xml root@hadoopclient:/etc/hive/conf
  - \$ scp Downloads/hive/hive-conf/hive-site.xml root@hadoopclient:/etc/hive/conf



#### Conexão com o HDFS

Listar os arquivos do cluster

```
Found 2 items

drwxrwxrwt - hdfs supergroup 0 2018-08-14 16:40 /tmp

drwxr-xr-x - hdfs supergroup 0 2018-08-14 16:41 /user
```

- sudo -u hdfs hdfs dfs -ls /
- Listar os datanodes disponíveis
  - sudo -u hdfs hdfs dfsadmin -printTopology

```
Rack: /default
192.168.25.191:50010 (datanode01.lab)
192.168.25.192:50010 (datanode02.lab)
192.168.25.193:50010 (datanode03.lab)
```

#### Conexão com o YARN

- Comando no hadoopclient para listar os nós disponíveis
  - sudo -u yarn yarn node -list

```
Total Nodes:3

Node-Id Node-State Node-Http-Address Number-of-Running-Containers

datanode01.lab:8041 RUNNING datanode01.lab:8042 0

datanode02.lab:8041 RUNNING datanode02.lab:8042 0

datanode03.lab:8041 __RUNNING datanode03.lab:8042 0
```

#### Conexão com o Hive

- Para conectar no Hive usamos o programa beeline
- Exemplo:
  - \$ beeline -n dataengineer -p dataengineer -u jdbc:hive2://edgenode.lab:10000
- Para mostrar as bases disponíveis:
  - > show databases;

```
dataengineer@hadoopclient:~$ beeline -n dataengineer -p dataengineer -u jdbc:hive2://edgenode.lab:10000
OpenJDK 64-Bit Server VM warning: ignoring option MaxPermSize=512M; support was removed in 8.0
OpenJDK 64-Bit Server VM warning: ignoring option MaxPermSize=512M; support was removed in 8.0
Connecting to jdbc:hive2://edgenode.lab:10000
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
Connected to: Apache Hive (version 1.1.0-cdh5.15.0)
Driver: Hive JDBC (version 1.1.0-cdh5.15.1)
Transaction isolation: TRANSACTION REPEATABLE READ
Beeline version 1.1.0-cdh5.15.1 by Apache Hive
  jdbc:hive2://edgenode.lab:10000> show databases
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:database name, type:string, comment:from deserial
izer)], properties:null)
INFO : Completed compiling command(queryId=hive 20180824220202 e60231a0-0a8e-47be-a87a-c76607ef3ff6); Time taken: 0.0
INFO : Executing command(queryId=hive_20180824220202_e60231a0-0a8e-47be-a87a-c76607ef3ff6): show databases
      : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive 20180824220202 e60231a0-0a8e-47be-a87a-c76607ef3ff6); Time taken: 0.1
INFO : OK
 database name
  row selected (0.34 seconds)
   jdbc:hive2://edgenode.lab:10000>
```

## Conexão com o Impala

- Para conectar no Impala usamos o programa impala-shell
- Exemplo:
  - impala-shell -i datanode01.lab
- Para mostrar as bases disponíveis:
  - > show databases;

## API de serviços

- O CDH tem uma API de serviços que permite verificar a configuração do servidor por meio de web services
- Listar os clusters disponíveis
  - http://headnode.lab:7180/api/v18/clusters/
- Listar os dados do Cluster 1
  - http://headnode.lab:7180/api/v18/clusters/Cluster%201
- Listar os serviços do Cluster 1
  - http://headnode.lab:7180/api/v19/clusters/Cluster%201/ser vices
- Na imagem vemos o trecho do serviço HDFS

# API de serviços (HDFS)

- Para listar as funcionalidades do HDFS
  - http://headnode.lab:7180/api/v19/clusters/Cluster%201/services/hdfs/roles
- Selecione a propriedade <u>name</u>

```
"name": "hdfs-NAMENODE-ab8735db1bc6548069202e72e8a31842",
"type": "NAMENODE",

"serviceRef": {
    "clusterName": "cluster",
    "serviceName": "hdfs"
},

"hostRef": {
    "hostId": "f10fb064-3af4-435d-9aab-6a5839be83d2"
},
    "roleUrl": "http://headnode.lab:7180/cmf/roleRedirect/hdfs-NAMENODE-ab8735db1bc6548069202e72e8a31842",
"roleState": "STARTED",
"healthSummary": "CONCERNING",
```

## API de serviços (HDFS)

- Visualizar as configurações do namenode
  - http://headnode.lab:7180/api/v19/clusters/Cluster%201/s ervices/hdfs/roles/hdfs-NAMENODE-ab8735db1bc6548069202e 72e8a31842/process/

```
▼ "configFiles": [
     "core-site.xml",
     "dfs all hosts.txt",
     "cloudera manager agent fencer.py",
     "cloudera-stack-monitor.properties",
     "navigator.client.properties",
     "cloudera-monitor.properties".
     "hdfs-site.xml".
     "log4j.properties",
     "redaction-rules.ison".
     "ssl-server.xml".
     "cloudera manager agent fencer secret key.txt",
     "topology.py",
     "hdfs.keytab",
     "http-auth-signature-secret",
     "event-filter-rules.json",
     "topology.map",
     "hadoop-policy.xml",
     "ssl-client.xml",
     "hadoop-metrics2.properties"
```

## API de serviços (HDFS)

- URL Para baixar um dos arquivos, por exemplo o core-site.xml:
  - http://headnode.lab:7180/api/v19/clusters/Cluster
     %201/services/hdfs/roles/hdfs-NAMENODEab8735db1bc6548069202e72e8a31842/process/configFiles/coresite.xml

# Máquina de gateway

- É gerenciado pelo cluster
- Garante a atualização das configurações do CDH automaticamente, não precisa da configuração manual
- Pode ser usado de diversas formas:
  - Cliente para todos os serviços do CDH, como HDFS, YARN, HBase, Hive etc.
  - Área de staging (temporária) para os arquivos que serão copiados no HDFS

## Acesso ao gateway

- Acesso normal via ssh
  - ssh root@edgenode
- Todos os serviços do CDH já estão instalados no gateway
  - Ele é um servidor gerenciado pelo CDH
- Comandos para testar
  - sudo -u hdfs hdfs dfs -ls /
  - sudo -u hdfs hdfs dfsadmin -printTopology

#### Hue

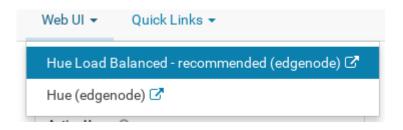


- Originalmente, a operação do Hadoop era descentralizada e complexa, utilizando diversas ferramentas diferentes e sem integração
- O Hue surgiu como uma interface web para operar o Hadoop e realizar análises de dados de forma simples e centralizada
- O foco é principalmente SQL, com suporte a autocompletar
- Acesso por interface gráfica para os serviços do Hadoop:
  - Pig
  - HDFS
  - Hive
  - Impala
  - Oozie



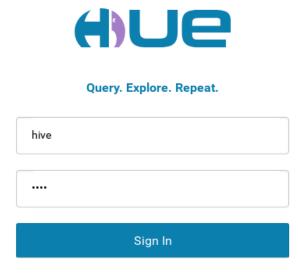
## Acesso pela Web UI

 Recomendado que o acesso seja feito por um EdgeNode para evitar carga excessiva no HeadNode



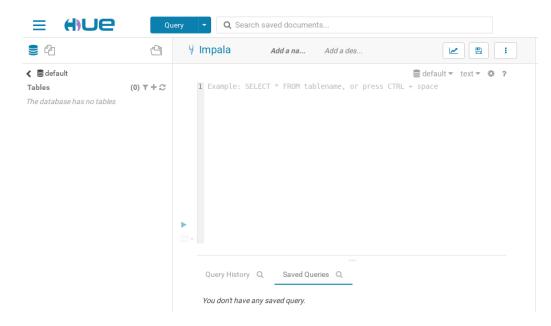
#### Acesso ao Hue

- No primeiro acesso será definido o usuário e a senha do administrador
  - Sugestão para usuário e senha: hive/hive



#### Tela inicial do Hue

- A tela inicial de consulta do Hue permite a operação dos diversos serviços habilitados
- O Hue é principalmente utilizado para consultas Impala e Hive

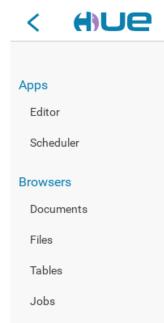


# Tipos de queries

- Menu Query → Editor
  - Impala
  - Hive
  - Pig
  - Java
  - Spark
  - MapReduce
  - Shell
  - Sqoop 1
  - Distcp

#### Menu Lateral

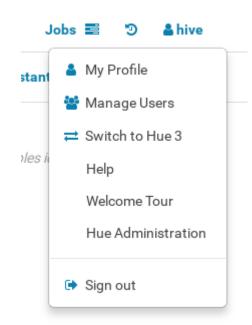
- Aplicações
  - Editor: editores de consulta
  - Scheduler: workflows do Oozie
- Browser
  - Documents: scripts gravados
  - Files: navegação no HDFS
  - Tables: administração das tabelas
  - Jobs: execuções do MapReduce, consultas e workflows



#### Gerenciar usuários

- Para listar os usuários vá no menu do usuário (hive) → Manage Users
- Para adicionar um novo usuário clique no botão Add user





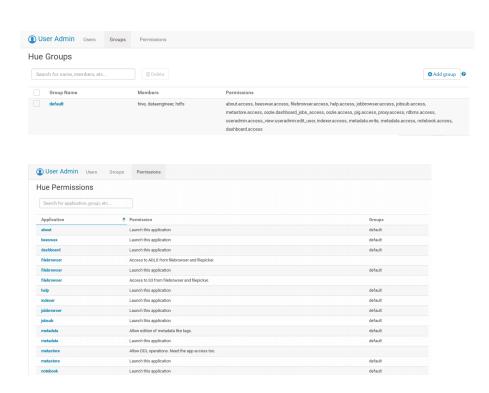
#### Credenciais

- Nos exemplos o usuário será dataengineer
- Cada aluno deve criar um usuário diferente
  - Sugestão: nome e sobrenome, como em marcoreis
- No Step 3: Advanced
  - Checar a opção Superuser status
  - Permite a administração do ambiente
- Clique no botão Add user
- Faça o Sign out do Hue e acesse novamente com o novo usuário

# Step 1: Credentials (required) Username dataengineer New Password Password confirmation Create home directory Create h

## Grupos e Permissões

- O Hue permite a criação de grupos e a definição de permissões para acesso aos recursos da ferramenta
- As permissões são autoexplicativas e indicam o nome da aplicação e o tipo de ação permitida



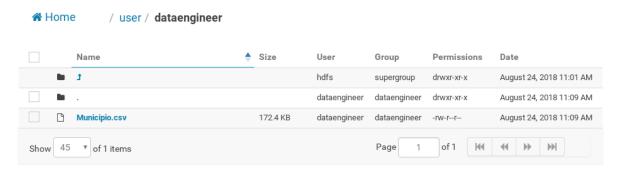
# Inserir dados de exemplo



- Usando o Hue com o usuário individual
  - Copiar um arquivo CSV para o HDFS
  - Criar uma base de dados
  - Criar uma tabela
  - Carregar dados do CSV na tabela

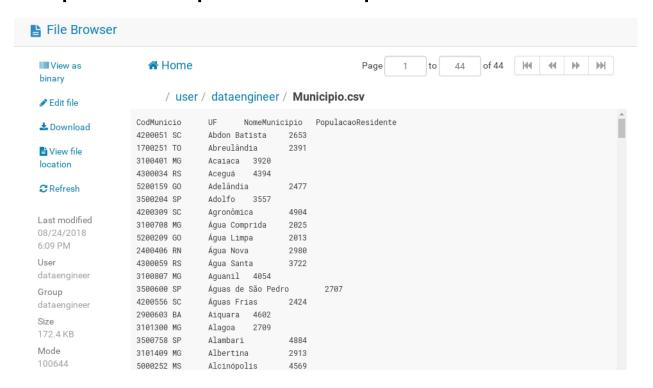
# Copiar para o HDFS

- Menu lateral → Files
  - Mostra o diretório raiz do usuário
- Clique no botão Upload e selecione o arquivo Municipio.csv



## Visualizar arquivo

Clique no arquivo Municipio.csv



#### Nova base de dados

- Menu lateral → Tables
- Clicar em Databases
- Clicar no botão +



#### Nova base de dados

- Database
  - Name: datalake
  - Cada usuário terá seu próprio ambiente, assim, a sugestão é usar o identificador datalake + nome do usuário. Ex: datalakemarcoreis



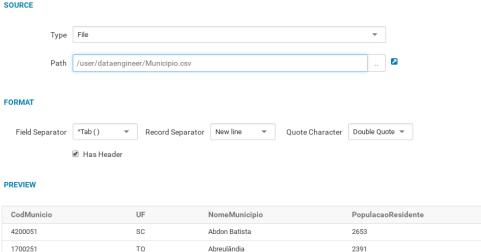
# Nova tabela (municipio)

3100401

MG

GO

- Naveguar até a nova base de dados
- Clicar no botão +
- Clicar na caixa de texto Path para selecionar o arquivo dos dados
- Selecione o Municipio.csv
- O Hue tenta identificar os separadores e mostra um preview dos dados do arquivo



Acaiaca

Aceguá

Adelândia

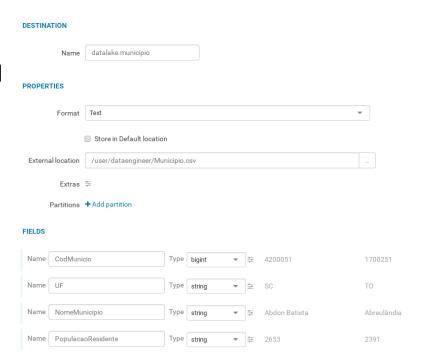
3920

4394

2477

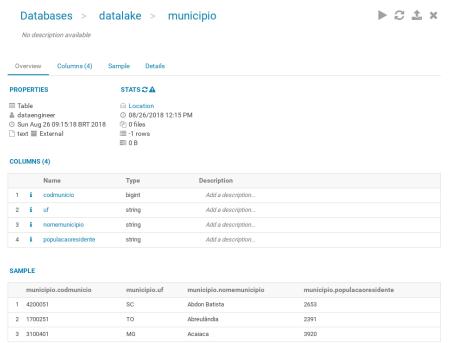
#### Create table

- O Hue tenta identificar os tipos de dados dos campos com base em seu conteúdo
- Não é muito confiável...
- Desmascar a opção Store in Default location



#### Tabela criada

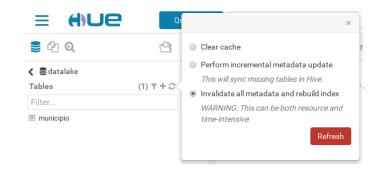
- Verifique as estatísticas e informações da tabela criada
- Para consultar os dados clique no botão

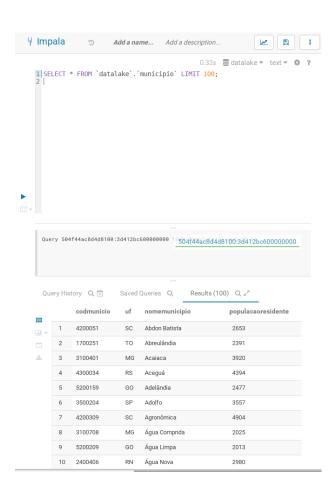


View more..

## Consultando no Impala

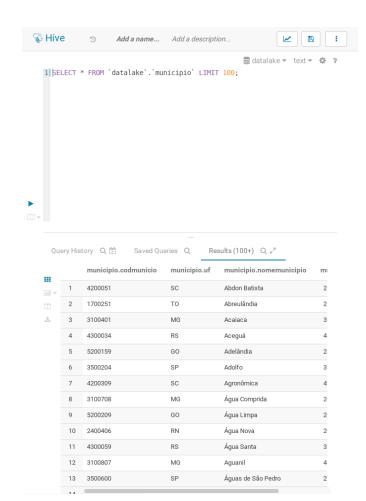
- A consulta pode ser executada no Impala ou Hive
- Tabelas criadas no Hive podem não aparecer no Impala. Neste caso, atualize os metadados
  - Invalidate all metadata





#### Consultando no Hive

- A consulta pelo Hive deve retornar exatamente os mesmos registros
- O editor padrão é o do Impala
- Para alterá-lo, clique na estrela ao lado no nome do Hive:
  - Set default application



### Dúvidas?

Marco Reis http://marcoreis.net