BIG DATA

TOO BIG TO IGNORE

SÜMEYYE KAYNAK

OUTLINE

Big Data Modeling and Data Management

Hadoop

Cloudera

Hue

Apache Hive

INTRODUCTION TO BIG DATA MODELING

- We need to have an idea of how the data looks.
- The goal of data modeling is to formally explore the nature of data so that you can find out what kind of storage you need, and what kind of processing you can do on it.
- Data modeling is a technique that helps to give meaningful insight into data by defining and categorizing it and establishing official definitions and descriptors so that the data can be utilized by all information systems in a company.

INTRODUCTION TO BIG DATA MODELING

- A high-level data model illustrates the core concepts and principles of any company in a very simplistic way, employing short descriptions.
- One of the biggest advantages of developing the high-level model is that it helps us to arrive at common terminology and definitions of the ideas and principles.

INTRODUCTION TO BIG DATA MODELING

- A high-level data model utilizes simplistic graphical images to illustrate the core concepts and principles of an organization and what they mean.
- For example, a database model shows the logical structure of a database, including the relationships and constraints that determine how data can be stored and accessed.

INTRODUCTION TO BIG DATA MODELING

Scenario:

- A student has a **First name**, a **Last name**, and a unique **identifier**.
- Each student is associated with an **institution**.
- Each student has a **Start date** and other data associated with them.

INTRODUCTION TO BIG DATA MODELING

Student ID	First name	Last name
52-743965	Charles	Peters
48-209689	Anthony	Sondrup
14-204968	Rebecca	Philips

Student ID	Institution	Start data	Type of plan
52-743965	156-983	04/01/2016	HSA
48-209689	146-823	12/01/2015	HMO
14-204968	447-784	03/14/2016	HSA

Institution	Provider name
156-983	UnitedHealth
146-823	Blue Shield
447-784	Carefirst

INTRODUCTION TO BIG DATA MODELING

Student ID	First name	Last name
52-743965	Charles	Peters
48-209689	Anthony	Sondrup
14-204968	Rebecca	Philips

Student ID	Institution	Start data	Type of plan
52-743965	156-983	04/01/2016	HSA
48-209689	146-823	12/01/2015	HMO
14-204968	447-784	03/14/2016	HSA

Institution	Provider name
156-983	UnitedHealth
146-823	Blue Shield
447-784	Carefirst

INTRODUCTION TO BIG DATA MODELING

Student ID	First name	Last name
52-743965	Charles	Peters
48-209689	Anthony	Sondrup
14-204968	Rebecca	Philips

Student ID	Institution	Start data	Type of plan
52-743965	156-983	04/01/2016	HSA
48-209689	146-823	12/01/2015	HMO
14-204968	447-784	03/14/2016	HSA

Institution	Provider name
156-983	UnitedHealth
146-823	Blue Shield
447-784	Carefirst

TO START WITH DATA MODELING

It is important to know the following:

- Understanding how the business works in order to understand data flow inside the organization.
- Understanding what type of data is gathered and stored in the organization.
- Understanding business processes and relationships. This knowledge guides us in building data and relationships in a data model.

BENEFITS OF MODELING

- **Gaining insight**: A detailed model shows the process from various angles.
- **Discussion**: The detailed data model can be used for discussions with the stakeholders.
- **Knowledge transfer:** A data model can be used as a source of documentation for instructing people or developers. Data modelling is a sort of documentation.

BENEFITS OF MODELING

- **Verification**: The process models are analyzed to find errors in systems or procedures.
- **Performance analysis:** A detailed model made from the data can be used to analyze the performance of the system by employing several available techniques, such as simulations.

BENEFITS OF MODELING

- **Specification**: A relevant model generated from an organization's data can be utilized to create a Software Requirement Specification (SRS) document.
- **Configuration**: The models constructed from data can be applied to configure a system. A detailed model constructed with precision shows the relationship between modules and how a module can communicate with another module.

TO MANAGING BIG DATA

- The intent of big data management is to figure out what kind of infrastructure support you would require for the data.

TO MANAGING BIG DATA

The big data management answers the following questions:

- How do we ingest or consume the data?
- Where and how do we store it?
- How can we ensure as well as enforce data quality?
- What operations do we perform on the data?
- How can these operations be efficient?
- How do we manage data scalability, variety, velocity, ...

BENEFITS OF BIG DATA MANAGEMENT

- Accelerates revenue
- Improved customer service
- Improves marketing

HADOOP SETUP MODES

- Hadoop can be run in 3 different modes.

- Standalone mode
  - It uses to test and debug
  - It is installed on a single machine.
  - The HDFS system is not used.
  - Doesn’t need configuration.

- Single node cluster
  - Hadoop cluster is built on a single machine.
  - HDFS replication factor value is 1.
  - Need configuration.

- Multiple node cluster
  - Hadoop cluster is built on multiple machine.
  - These machines are connected to each other in the network as clusters.
  - HDFS replication factor value is greater than 1.

SETTING UP BIG DATA MODELING PLATFORMS

- We are going to set up Cloudera VM on Windows.

SETTING UP BIG DATA MODELING PLATFORMS

VirtualBox

Download VirtualBox

Here you will find links to VirtualBox binaries and its source code.

VirtualBox binaries

By downloading, you agree to the terms and conditions of the respective license.

If you're looking for the latest VirtualBox 6.0 packages, see VirtualBox 6.0 builds. has been discontinued in 6.1. Version 6.0 will remain supported until July 2020.

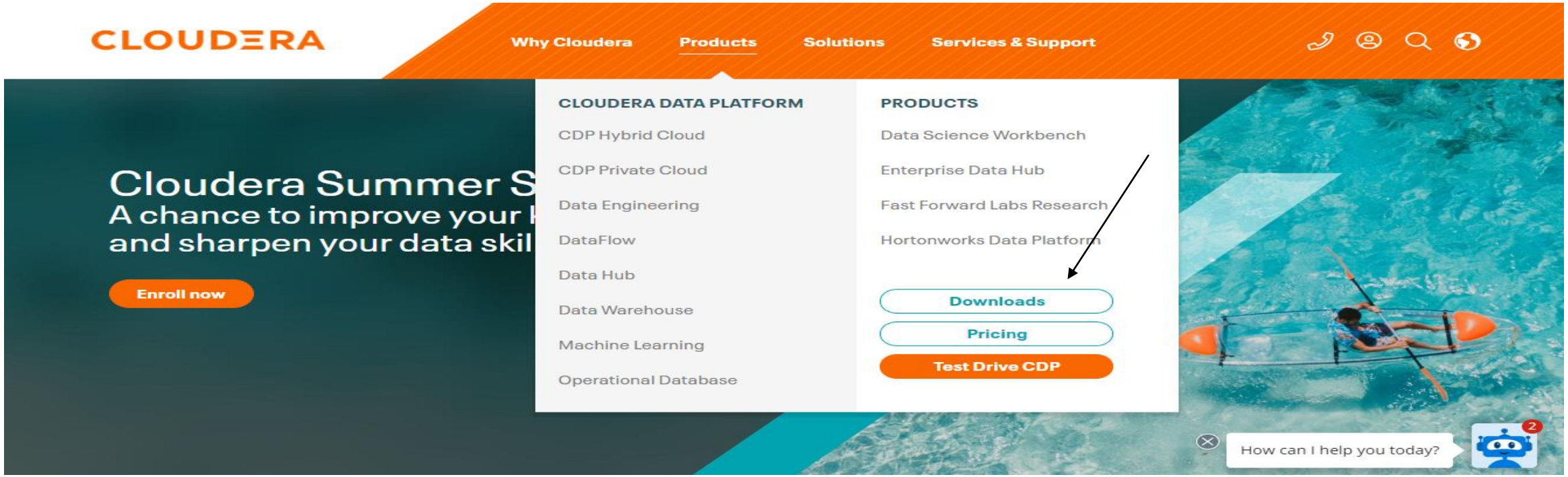
If you're looking for the latest VirtualBox 5.2 packages, see VirtualBox 5.2 builds. discontinued in 6.0. Version 5.2 will remain supported until July 2020.

VirtualBox 6.1.26 platform packages

- Windows hosts
- OS X hosts
- Linux distributions
- Solaris hosts
- Solaris 11 IPS hosts

The binaries are released under the terms of the GPL version 2.

CLOUDERA


CLOUDERA
Why Cloudera
Products
Solutions
Services & Support
Cloudera Summer S
A chance to improve your
and sharpen your data skil
Enroll now
CLOUDERA DATA PLATFORM
CDP Hybrid Cloud
CDP Private Cloud
Data Engineering
DataFlow
Data Hub
Data Warehouse
Machine Learning
Operational Database
PRODUCTS
Data Science Workbench
Enterprise Data Hub
Fast Forward Labs Research
Hortonworks Data Platform
Downloads
Pricing
Test Drive CDP
How can I help you today?

CLOUDERA INSTALLATION

- **Title**: Virtual Box
- **Value**: vb
- **Download Location:** https://downloads.cloudera.com/demo_vm/virtualbox/cloudera-quickstart-vm-5.13.0-0-virtualbox.zip

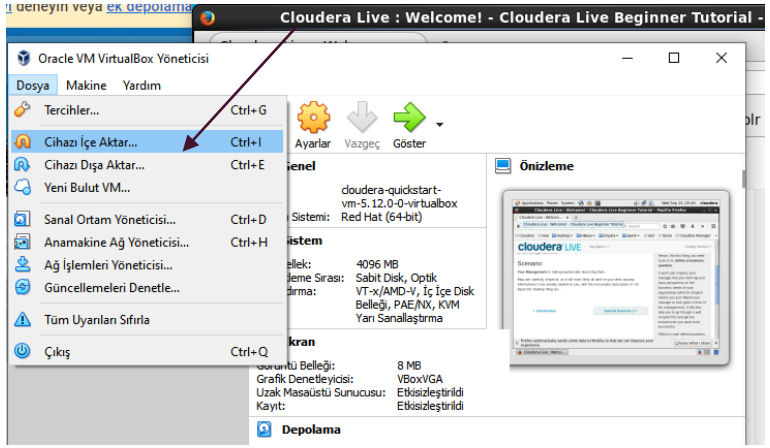
- **Title**: VMWare
- **Value**: vmw
- **Download Location:** https://downloads.cloudera.com/demo_vm/vmware/cloudera-quickstart-vm-5.13.0-0-vmware.zip

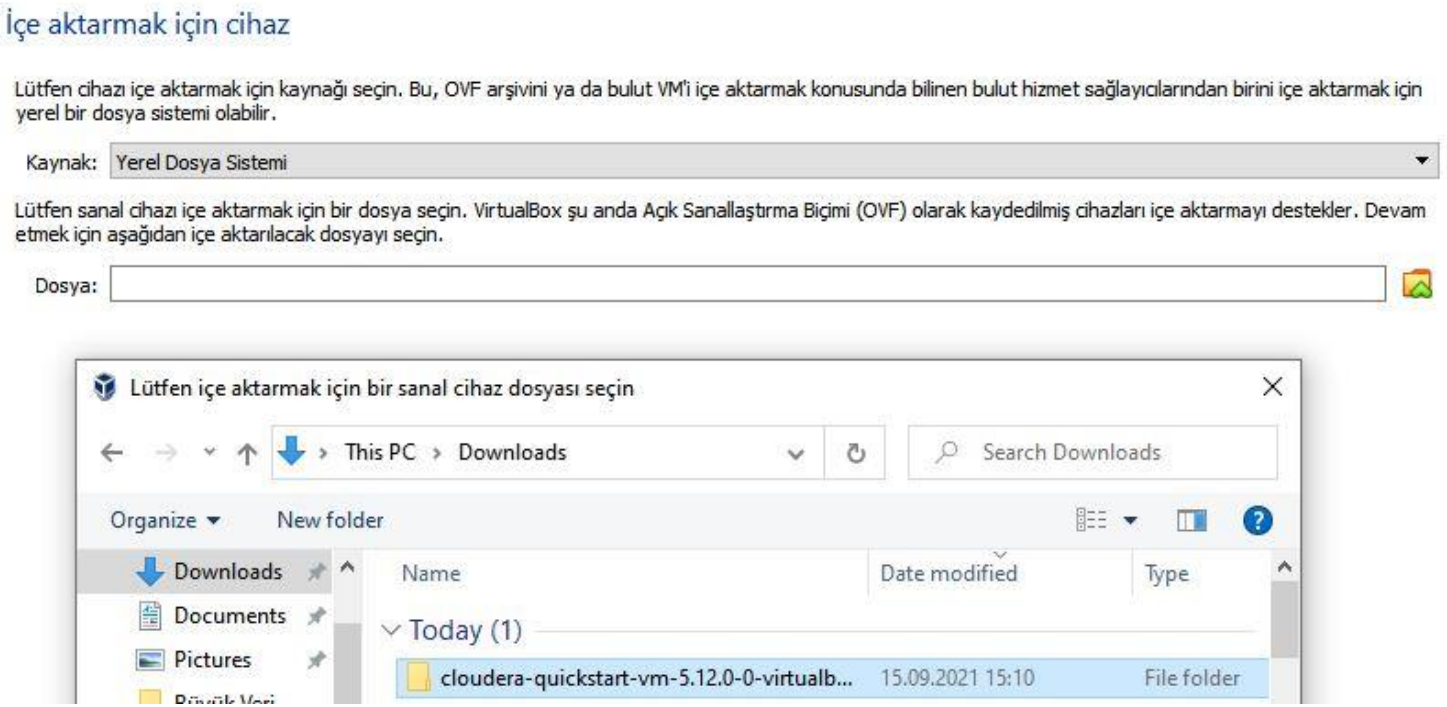
CLOUDERA INSTALLATION

- **Title**: KVM
- **Value**: kvm
- **Download Location**: https://downloads.cloudera.com/demo_vm/kvm/cloudera-quickstart-vm-5.13.0-0-kvm.zip

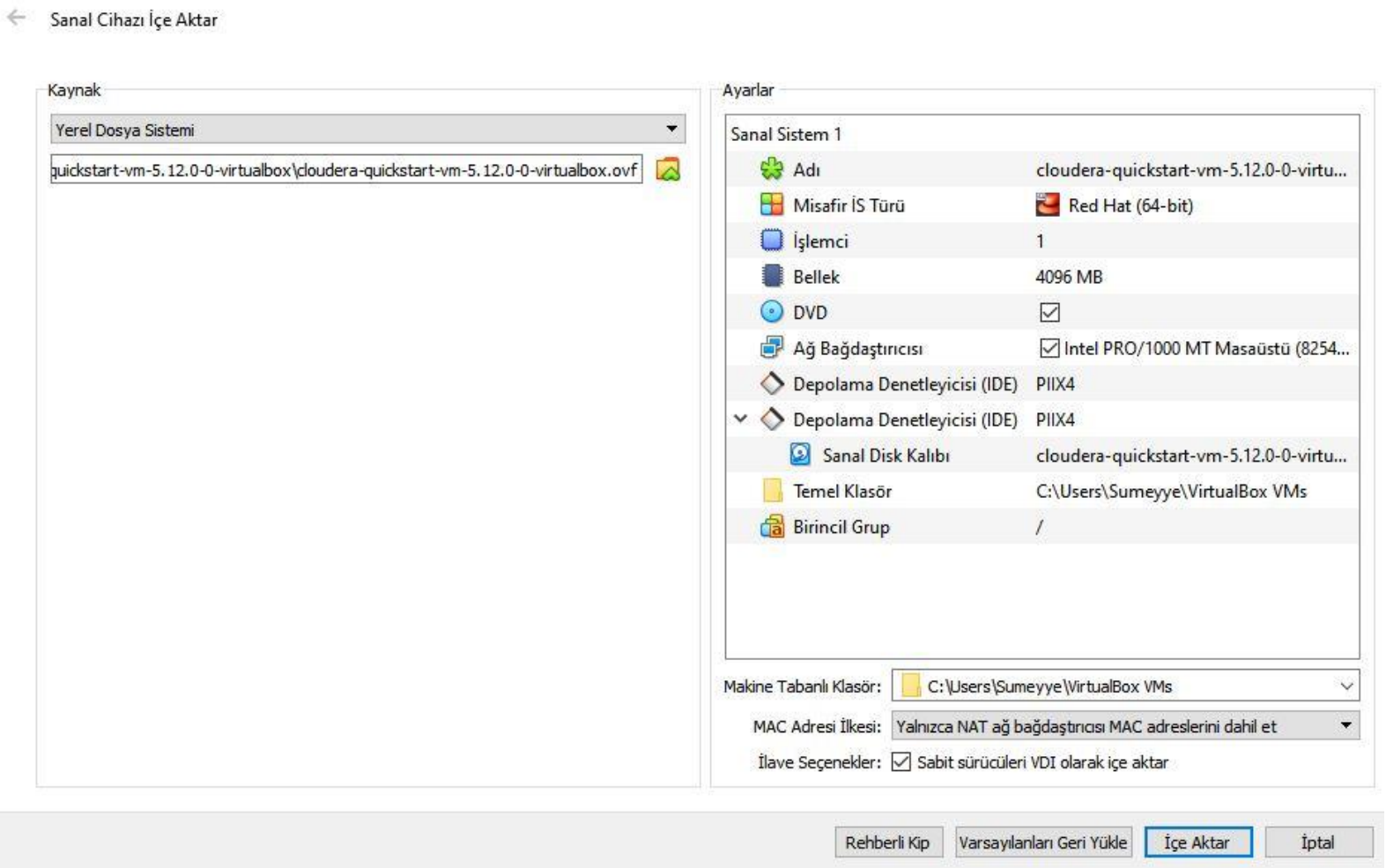
- **Title**: Docker Image
- **Value**: docker
- **Download Location**: https://downloads.cloudera.com/demo_vm/docker/cloudera-quickstart-vm-5.13.0-0-beta-docker.tar.gz

VM-CLOUDERA

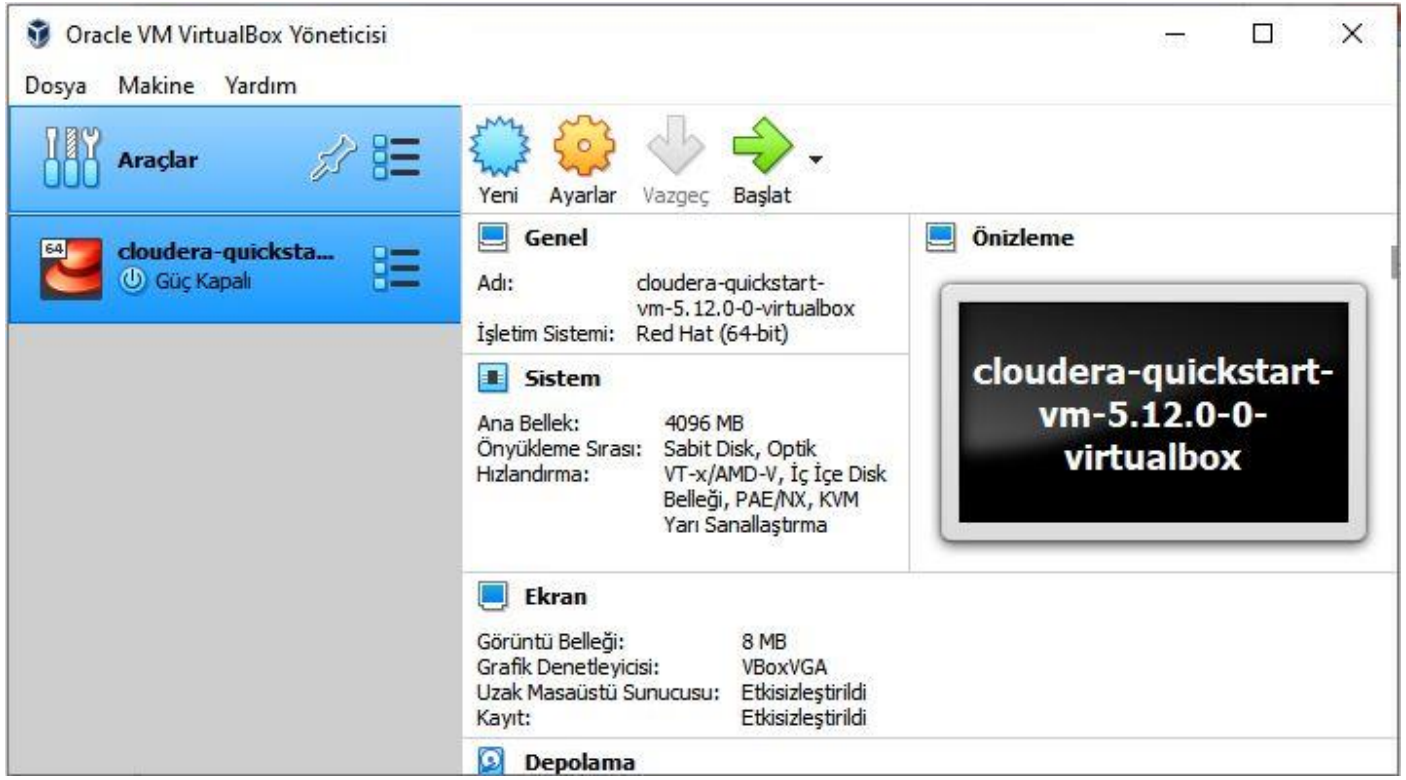

Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial -
Oracle VM VirtualBox Yöneticisi
Dosya
Makine
Yardım
Tercihler... Ctrl+G
Cihazı İçe Aktar... Ctrl+I
Cihazı Dışa Aktar... Ctrl+E
Yeni Bulut VM...
Sanal Ortam Yöneticisi... Ctrl+D
Anamakine Ağ Yöneticisi... Ctrl+H
Ağ İşlemleri Yöneticisi...
Güncellemeleri Denetle...
Tüm Uyarıları Sıfırla
Çıkış Ctrl+Q
Ayarlar
Vazgeç
Göster
Önizleme
Depolama


İçe aktarmak için cihaz
Lütfen cihazı içe aktarmak için kaynağı seçin. Bu, OVF arşivini ya da bulut VM'i içe aktarmak konusunda bilinen bulut hizmet sağlayıcılarından birini içe aktarmak için yerel bir dosya sistemi olabilir.
Kaynak: Yerel Dosya Sistemi
Lütfen sanal cihazı içe aktarmak için bir dosya seçin. VirtualBox şu anda Açık Sanallaştırma Biçimi (OVF) olarak kaydedilmiş cihazları içe aktarmayı destekler. Devam etmek için aşağıdan içe aktarılacak dosyayı seçin.
Dosya:
Lütfen içe aktarmak için bir sanal cihaz dosyası seçin
This PC › Downloads
Search Downloads
Organize
New folder
Downloads
Documents
Pictures
Name
Date modified
Type
Today (1)
cloudera-quickstart-vm-5.12.0-0-virtualb...
15.09.2021 15:10
File folder

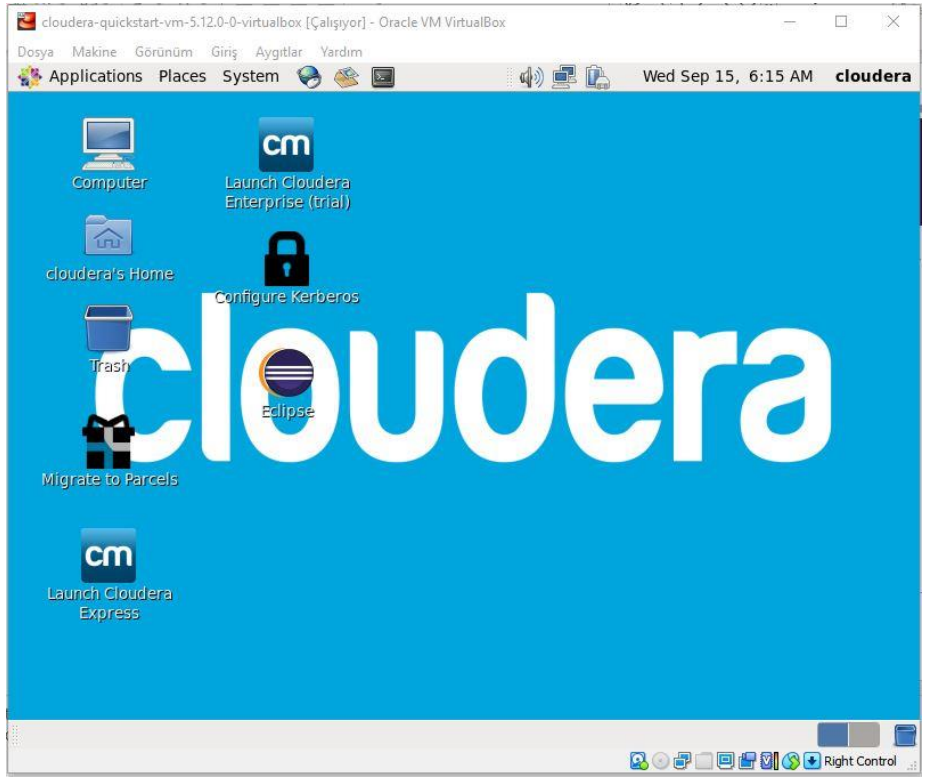
VM-CLOUDERA


Sanal Cihazı İçe Aktar
Kaynak
Yerel Dosya Sistemi
quickstart-vm-5.12.0-0-virtualbox\cloudera-quickstart-vm-5.12.0-0-virtualbox.ovf
Ayarlar
Sanal Sistem 1
Adı
cloudera-quickstart-vm-5.12.0-0-virtu...
Misafir İS Türü
Red Hat (64-bit)
İşlemci
1
Bellek
4096 MB
DVD
Ağ Bağdaştırıcısı
Intel PRO/1000 MT Masaüstü (8254...
Depolama Denetleyicisi (IDE)
PIIX4
Depolama Denetleyicisi (IDE)
PIIX4
Sanal Disk Kalıbı
cloudera-quickstart-vm-5.12.0-0-virtu...
Temel Klasör
C:\Users\Sumeyye\VirtualBox VMs
Birincil Grup
/
Makine Tabanlı Klasör:
C:\Users\Sumeyye\VirtualBox VMs
MAC Adresi İlkesi:
Yalnızca NAT ağ bağdaştırıcısı MAC adreslerini dahil et
İlave Seçenekler:
Sabit sürücüleri VDI olarak içe aktar
Rehberli Kip
Varsayılanları Geri Yükle
İçe Aktar
İptal

VM-CLOUDERA


Oracle VM VirtualBox Yöneticisi
Dosya
Makine
Yardım
Araçlar
cloudera-quicksta...
Güç Kapalı
Yeni
Ayarlar
Vazgeç
Başlat
Genel
Adı:
cloudera-quickstart-vm-5.12.0-0-virtualbox
İşletim Sistemi:
Red Hat (64-bit)
Sistem
Ana Bellek:
4096 MB
Önyükleme Sırası:
Sabit Disk, Optik
Hızlandırma:
VT-x/AMD-V, İç İçe Disk Belleği, PAE/NX, KVM Yarı Sanallaştırma
Önizleme
cloudera-quickstart-vm-5.12.0-0-virtualbox
Ekran
Görüntü Belleği:
8 MB
Grafik Denetleyicisi:
VBoxVGA
Uzak Masaüstü Sunucusu:
Etkisizleştirildi
Kayıt:
Etkisizleştirildi
Depolama

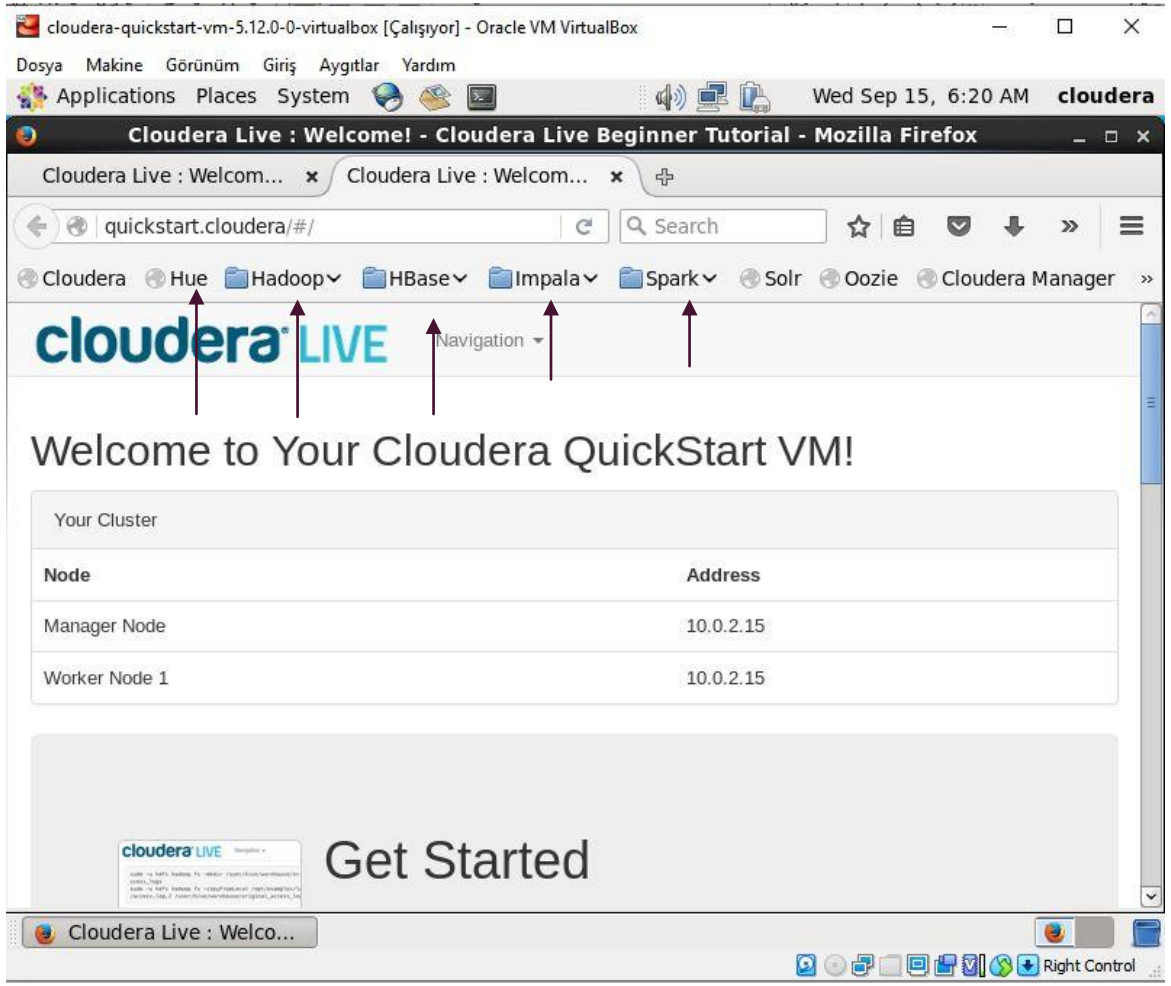
VM-CLOUDERA DESKTOP SCREEN



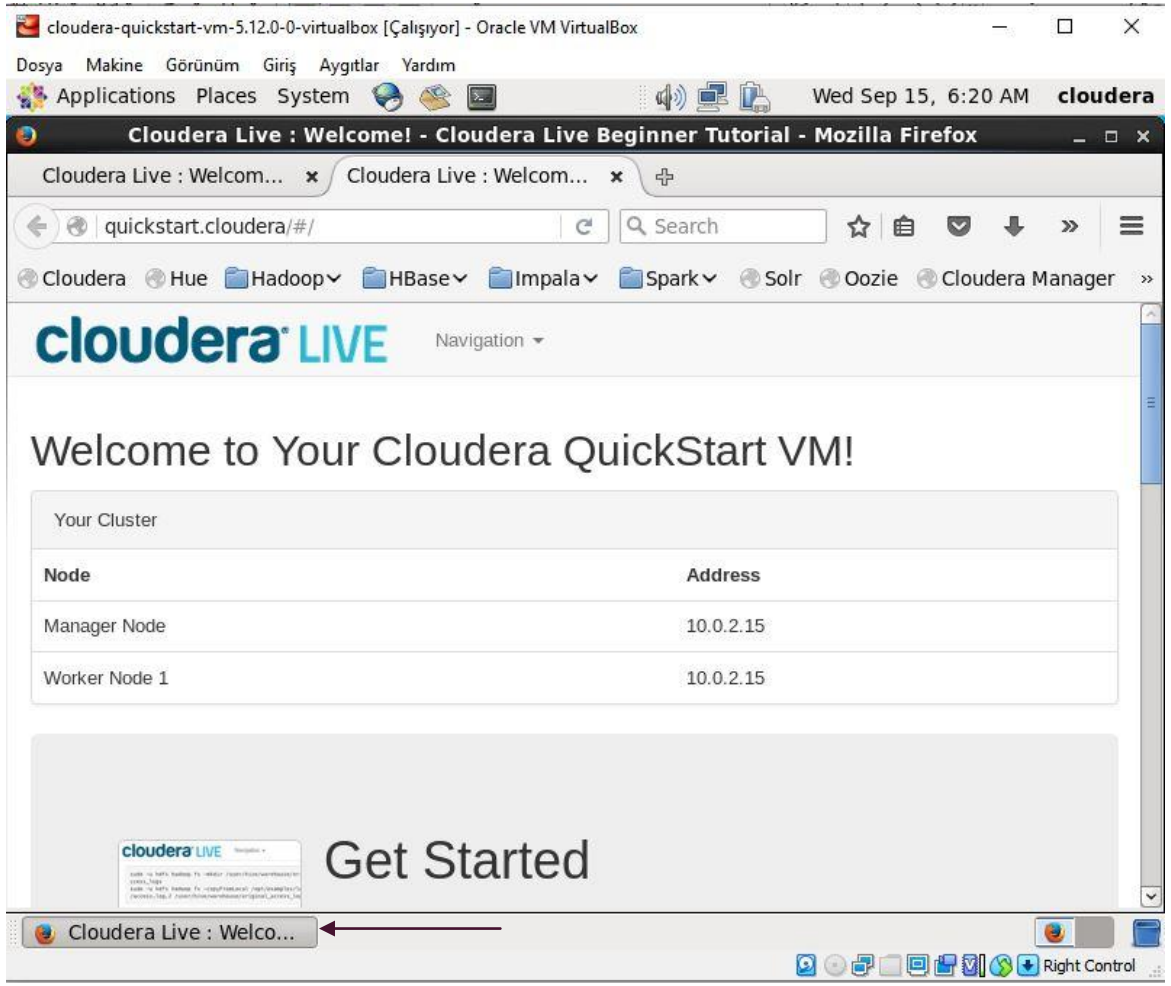
VM SETTINGS


Oracle VM VirtualBox Yöneticisi
Dosya
Makine
Yardım
Araçlar
Yeni
Ayarlar
Vazgeç
cloudera-quicksta...
Çalışıyor
Genel
Adı:
İşletim Sistemi:
Sistem
Ana Bellek:
Önyükleme Sırası:
Hızlandırma:
Ekran
Görüntü Belleği:
Grafik Denetleyicisi:
Uzak Masaüstü Sunucusu:
Kayıt:
Depolama
VirtualBox - Tercihler
Genel
Giriş
Güncelleme
Dil
Ekran
Ağ
Uzantılar
Vekil
Giriş
VirtualBox Yöneticisi
Sanal Makine
Adı
Kısayol
ACPI Kapama
Anlık Görüntü Bölmesi: Al...
Ctrl+Shift+T
Anamakine Ağ Yöneticisi
Anamakine Ağ Yöneticisi...
Ctrl+H
Dosya Yöneticisi: Anamakineye kopyala
Klavyeyi Otomatik Yakala
TAMAM
İptal

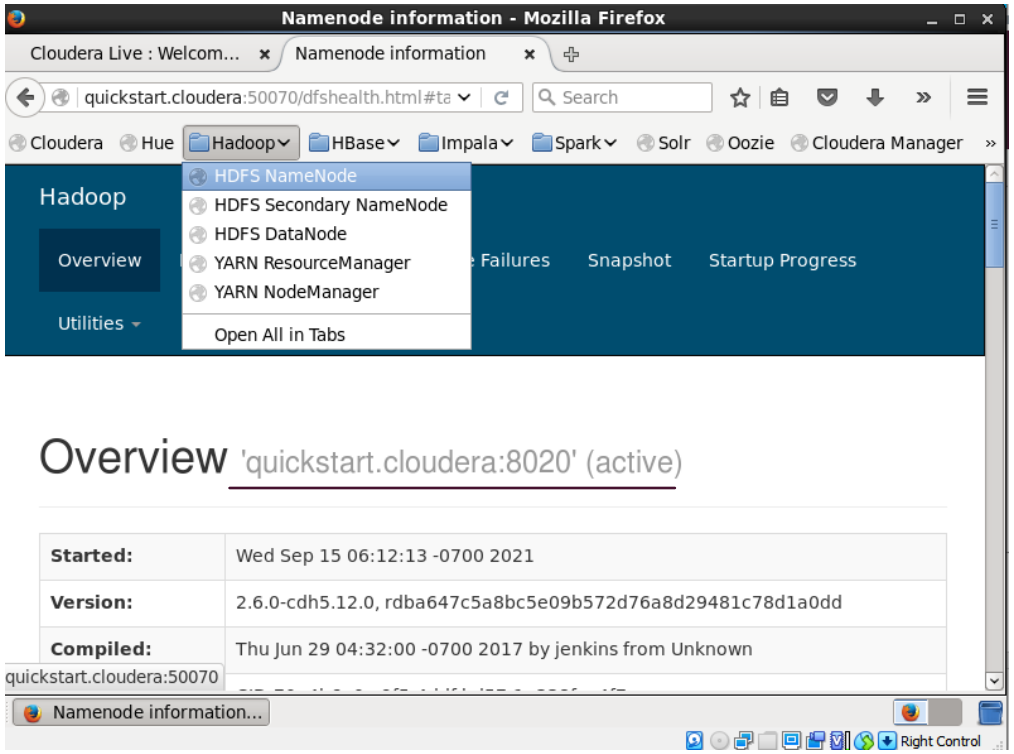
VM-CLOUDERA


cloudera-quickstart-vm-5.12.0-0-virtualbox [Çalışıyor] - Oracle VM VirtualBox
Dosya Makine Görünüm Giriş Aygıtlar Yardım
Applications Places System
Wed Sep 15, 6:20 AM cloudera
Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox
Cloudera Live : Welcom...
Cloudera Live : Welcom...
quickstart.cloudera/#/
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager
cloudera LIVE
Navigation
Welcome to Your Cloudera QuickStart VM!
Your Cluster
Node
Address
Manager Node
10.0.2.15
Worker Node 1
10.0.2.15
Get Started
Cloudera Live : Welco...
Right Control

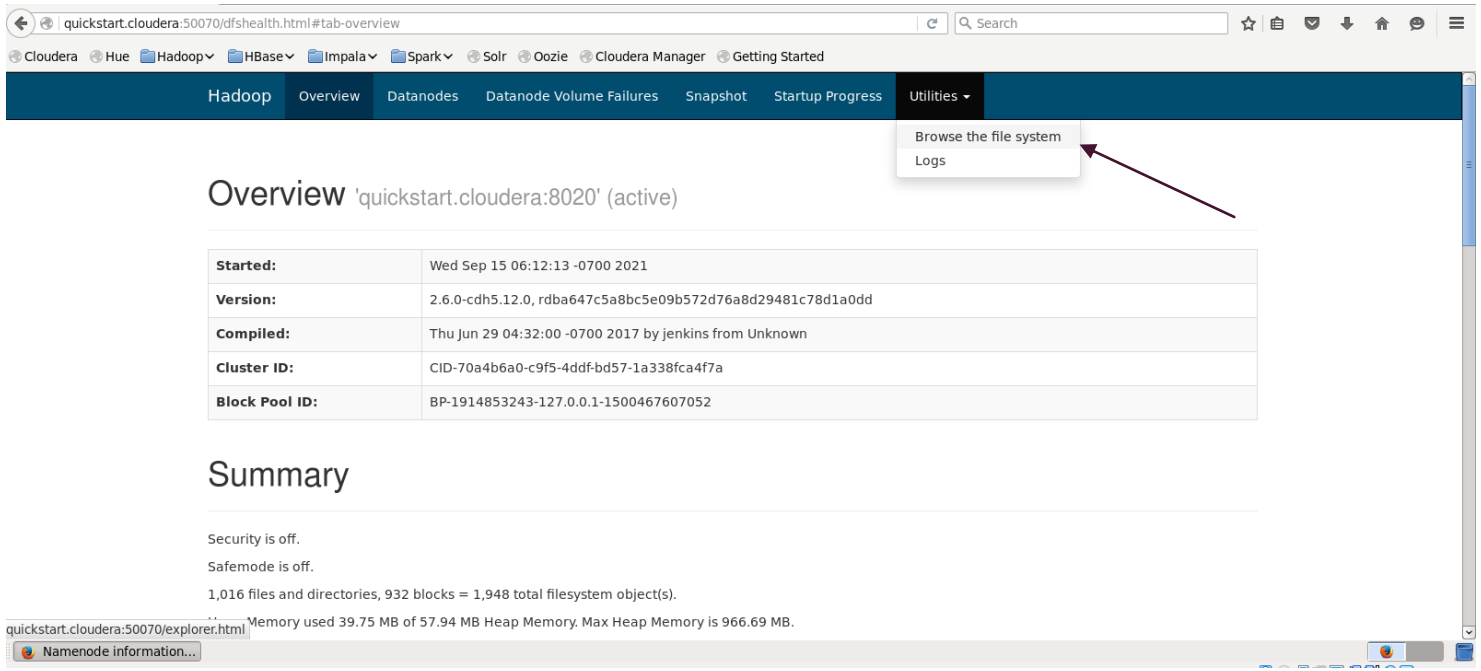
VM-CLOUDERA


cloudera-quickstart-vm-5.12.0-0-virtualbox [Çalışıyor] - Oracle VM VirtualBox
Dosya Makine Görünüm Giriş Aygıtlar Yardım
Applications Places System
Wed Sep 15, 6:20 AM cloudera
Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox
Cloudera Live : Welcom...
Cloudera Live : Welcom...
quickstart.cloudera/#/
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager
cloudera LIVE
Navigation
Welcome to Your Cloudera QuickStart VM!
Your Cluster
Node
Address
Manager Node
10.0.2.15
Worker Node 1
10.0.2.15
Get Started
Cloudera Live : Welco...
Right Control

HADOOP


Namenode information - Mozilla Firefox
Cloudera Live : Welcom...
Namenode information
quickstart.cloudera:50070/dfshealth.html#ta
Search
Cloudera
Hue
Hadoop
HBase
Impala
Spark
Solr
Oozie
Cloudera Manager
HDFS NameNode
HDFS Secondary NameNode
HDFS DataNode
YARN ResourceManager
YARN NodeManager
Open All in Tabs
Hadoop
Overview
Utilities
Failures
Snapshot
Startup Progress
Overview 'quickstart.cloudera:8020' (active)
Started: Wed Sep 15 06:12:13 -0700 2021
Version: 2.6.0-cdh5.12.0, rdba647c5a8bc5e09b572d76a8d29481c78d1a0dd
Compiled: Thu Jun 29 04:32:00 -0700 2017 by jenkins from Unknown
quickstart.cloudera:50070
Namenode information...
Right Control

HADOOP


quickstart.cloudera:50070/dfshealth.html#tab-overview
Search
Cloudera
Hue
Hadoop
HBase
Impala
Spark
Solr
Oozie
Cloudera Manager
Getting Started
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse the file system
Logs
Overview 'quickstart.cloudera:8020' (active)
Started: Wed Sep 15 06:12:13 -0700 2021
Version: 2.6.0-cdh5.12.0, rdba647c5a8bc5e09b572d76a8d29481c78d1a0dd
Compiled: Thu Jun 29 04:32:00 -0700 2017 by jenkins from Unknown
Cluster ID: CID-70a4b6a0-c9f5-4ddf-bd57-1a338fca4f7a
Block Pool ID: BP-1914853243-127.0.0.1-1500467607052
Summary
Security is off.
Safemode is off.
1,016 files and directories, 932 blocks = 1,948 total filesystem object(s).
Memory used 39.75 MB of 57.94 MB Heap Memory. Max Heap Memory is 966.69 MB.
quickstart.cloudera:50070/explorer.html
Namenode information...

HADOOP

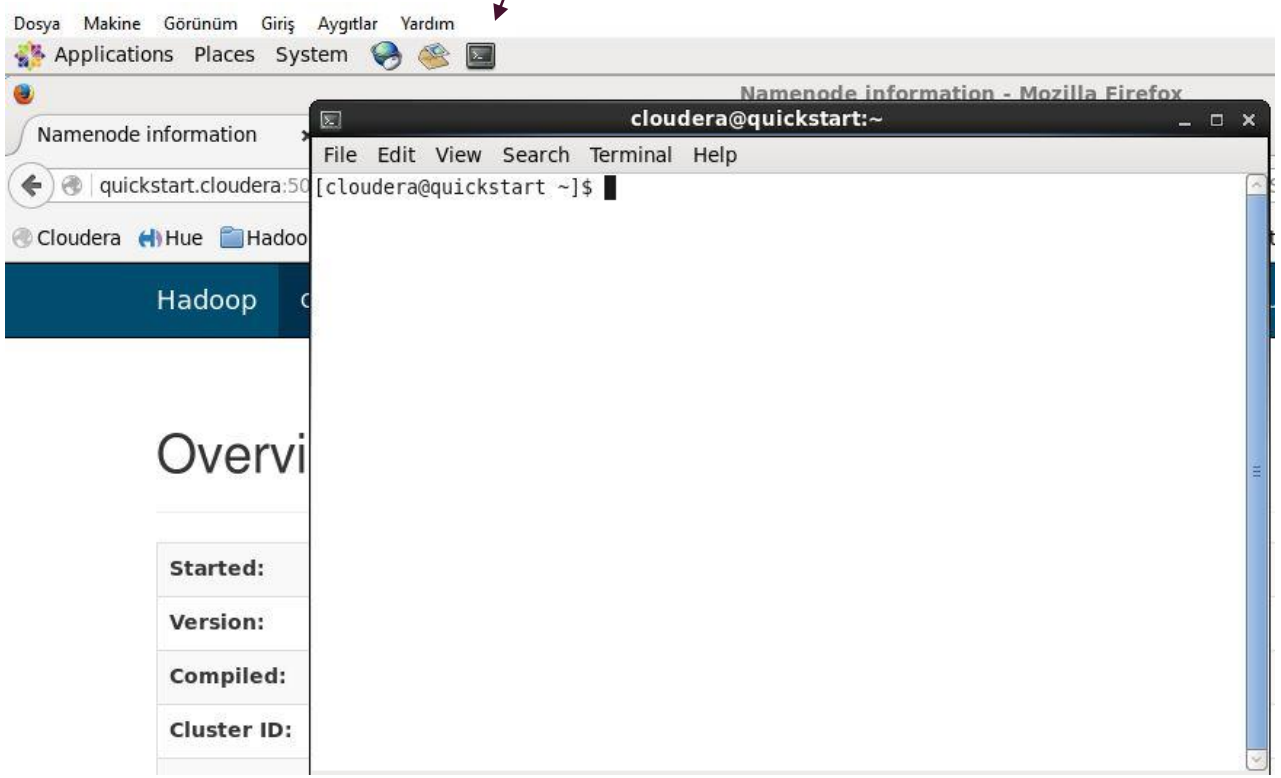
Browse Directory

/ Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxrwxrwx	hdfs	supergroup	0 B	Wed Jul 19 05:34:46 -0700 2017	0	0 B	benchmarks
drwxr-xr-x	hbase	supergroup	0 B	Wed Sep 15 06:15:56 -0700 2021	0	0 B	hbase
drwxr-xr-x	solr	solr	0 B	Wed Jul 19 05:37:04 -0700 2017	0	0 B	solr
drwxrwxrwt	hdfs	supergroup	0 B	Wed Sep 15 05:20:01 -0700 2021	0	0 B	tmp
drwxr-xr-x	hdfs	supergroup	0 B	Wed Jul 19 05:36:36 -0700 2017	0	0 B	user
drwxr-xr-x	hdfs	supergroup	0 B	Wed Jul 19 05:36:28 -0700 2017	0	0 B	var

Hadoop, 2017.

HADOOP COMMANDS



HADOOP COMMANDS

HADOOP COMMANDS

Browse Directory

/ Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxrwxrwx	hdfs	supergroup	0 B	Wed Jul 19 05:34:46 -0700 2017	0	0 B	benchmarks
							example
drwxr-xr-x	hbase	supergroup	0 B	Wed Sep 15 06:15:56 -0700 2021	0	0 B	hbase
drwxr-xr-x	solr	solr	0 B	Wed Jul 19 05:37:04 -0700 2017	0	0 B	solr
drwxrwxrwt	hdfs	supergroup	0 B	Wed Sep 15 05:20:01 -0700 2021	0	0 B	tmp
drwxr-xr-x	hdfs	supergroup	0 B	Wed Jul 19 05:36:36 -0700 2017	0	0 B	user
drwxr-xr-x	hdfs	supergroup	0 B	Wed Jul 19 05:36:28 -0700 2017	0	0 B	var

```
[cloudera@quickstart ~]$ hdfs dfs -mkdir /example
```

HADOOP COMMANDS

userId,movieId,rating,timestamp
1,31,2.5,1260759144
1,1029,3.0,1260759179
1,1061,3.0,1260759182
1,1129,2.0,1260759185
1,1172,4.0,1260759205
1,1263,2.0,1260759151
1,1287,2.0,1260759187
1,1293,2.0,1260759148

Copying Data from Local Machine to HDFS

- hdfs dfs -copyFromLocal / local source folder / destination hdfs folder

```
[cloudera@quickstart ~]$ hdfs dfs -copyFromLocal /home/cloudera/Downloads/ratings.csv /example
```

/example Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	cloudera	supergroup	2.33 MB	Fri Sep 17 05:40:20 -0700 2021	1	128 MB	ratings.csv

HADOOP COMMANDS

- Finding the Number of Files in the Folder

```
[cloudera@quickstart ~]$ hdfs dfs -count /example
           1            1            2438266 /example
```

- Printing the Contents of the File to the Console

```
[cloudera@quickstart ~]$ hdfs dfs -cat /example/ratings.csv
```

- Copying Files Between Folders in HDFS

```
[cloudera@quickstart ~]$ hdfs dfs -cp /example/ratings.csv /var
```

HADOOP COMMANDS

- Move files between folders in HDFS

```
[cloudera@quickstart ~]$ hdfs dfs -mv /example /var
```

```
[cloudera@quickstart ~]$ hdfs dfs -mv /var/example /example
```

```
[cloudera@quickstart ~]$ hdfs dfs -mv /example/ratings.csv /var
```

- Delete files or folders

```
[cloudera@quickstart ~]$ hdfs dfs -rmr /var/ratings.c
rmr: DEPRECATED: Please use 'rm -r' instead.
```

- List files in folder

```
[cloudera@quickstart ~]$ hdfs dfs -ls /var
Found 2 items
drwxr-xr-x   - hdfs supergroup          0 2017-07-19 05:36 /var/lib
drwxrwxr-t   - yarn mapred              0 2017-07-19 05:34 /var/log
```

HADOOP COMMANDS

drwxr-xr-x	cloudera	supergroup	0 B	Fri Sep 17 05:40:20 -0700 2021	0	0 B	example

Permission	Files
r	Can read the file
w	Can write the file
x	Can execute the file

Permission	Files	Operator	Access class
r	Can read the file	+ (add access)	u (user)
w	Can write the file	- (remove access)	g (group)
x	Can execute the file	= (set exact access)	a (all)

HADOOP COMMANDS

For example, to add permission for everyone to read a file in the current directory named myfile, at the Unix prompt, enter:

```
chmod a+r myfile
```

The a stands for "all", the + for "add", and the r for "read".

Note:

This assumes that everyone already has access to the directory where myfile is located and its parent directories; that is, you must set the directory permissions separately.

HADOOP COMMANDS

- To remove read and write permission for group on a file named myfile.

chmod g-rw myfile

- To remove write permission and add execute for all users on myfile.

chmod a-w+x myfile

HADOOP COMMANDS

```
[cloudera@quickstart ~]$ hdfs dfs -chmod -x /example/ratings.csv
[cloudera@quickstart ~]$ hdfs dfs -chmod +r /example/ratings.csv
[cloudera@quickstart ~]$ hdfs dfs -chmod +xr /example/ratings.csv
[cloudera@quickstart ~]$ hdfs dfs -chmod +wxr /example/ratings.csv
```

- You can set replication factor.

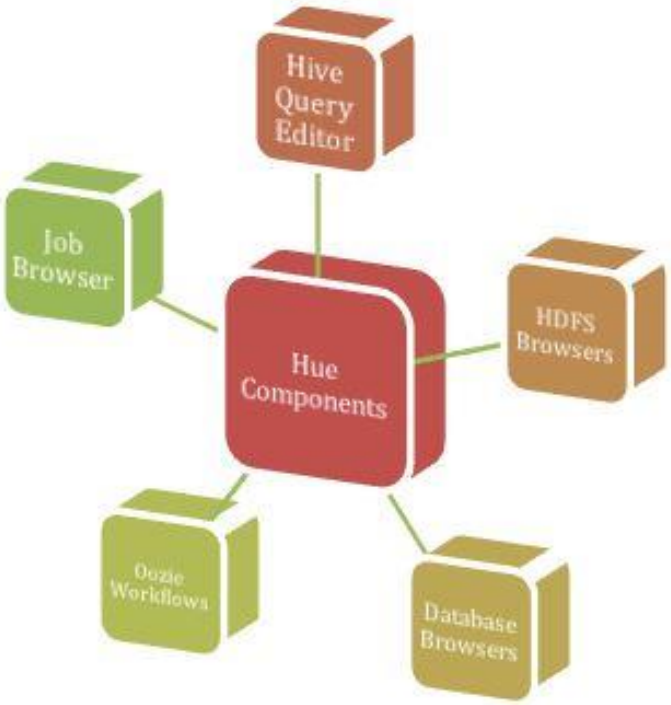
hdfs dfs -setrep RepDegeri -R /hdfs-folders

```
[cloudera@quickstart ~]$ hdfs dfs -setrep 4 -R /example/ratings.csv
setrep: `-R': No such file or directory
Replication 4 set: /example/ratings.csv
```

HUE

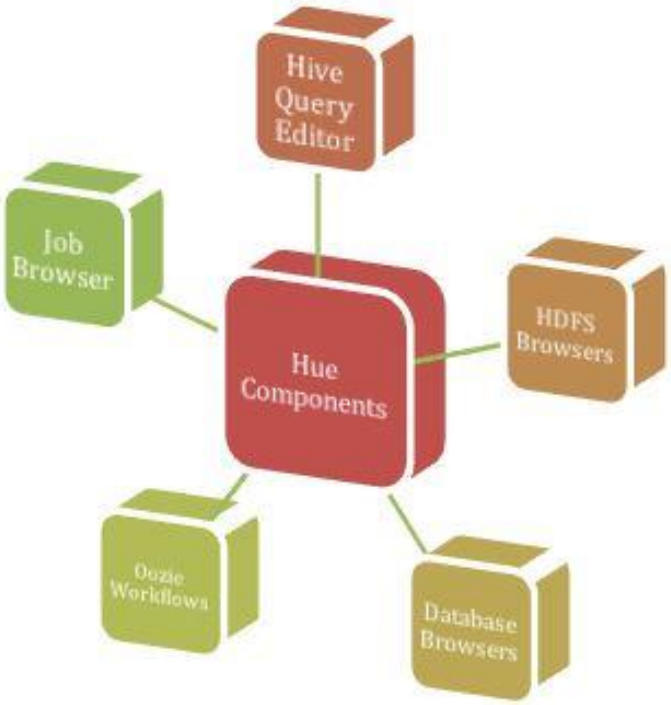
- Hue is an open source SQL Assistant for Databases & Data Warehouses.
- Hue can connect to all the databases such as Apache Hive, Apache Impala, SparkSQL, Elastic Search..

HUE COMPONENT



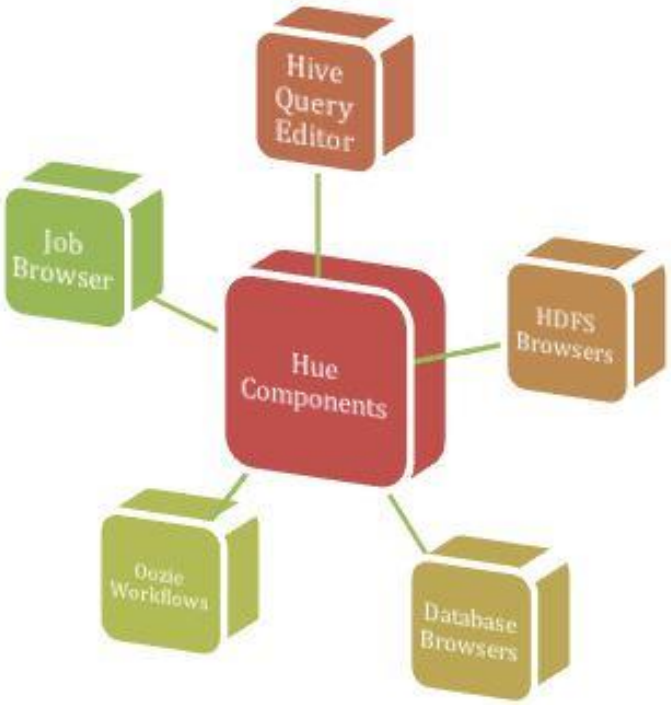
- **HDFS Browser:** While working with Hadoop Ecosystem, one of the most important factors is the ability to access the HDFS Browser.
- User can interact with the HDFS files.

HUE COMPONENT



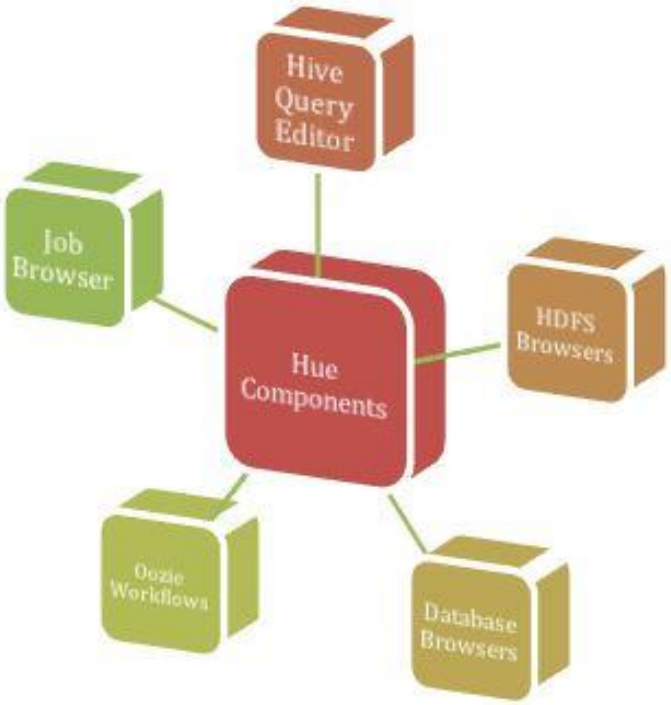
- **Job Browser:** Hadoop ecosystems consist of many jobs and sometimes developers may need to know that which job is currently running on the Hadoop cluster and which job has been successfully completed and which has errors.

HUE COMPONENT



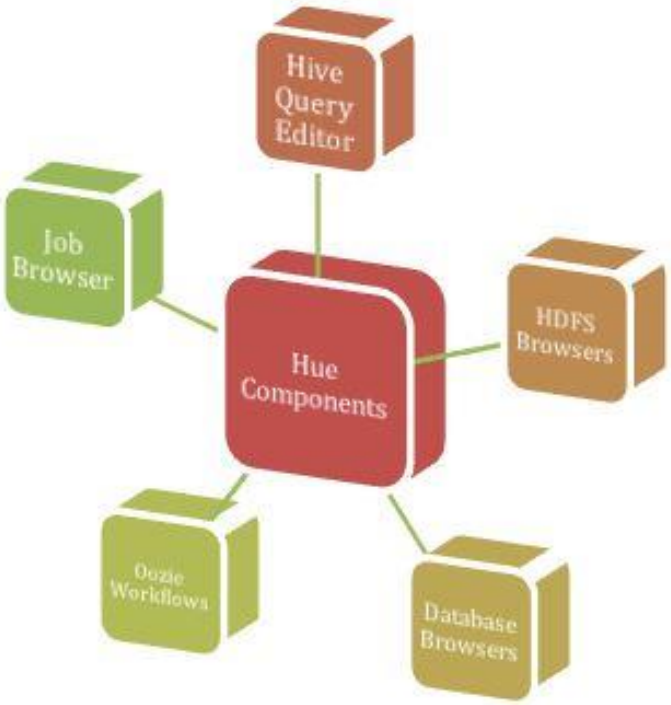
- **Hive Query Editor:** Hive query editor allows us to write SQL Hive queries and the result can also be shown in the editor.

HUE COMPONENT



- **Database Browser:** All of the available datastore tables can be displayed, exported and imported through Database browser.

HUE COMPONENT

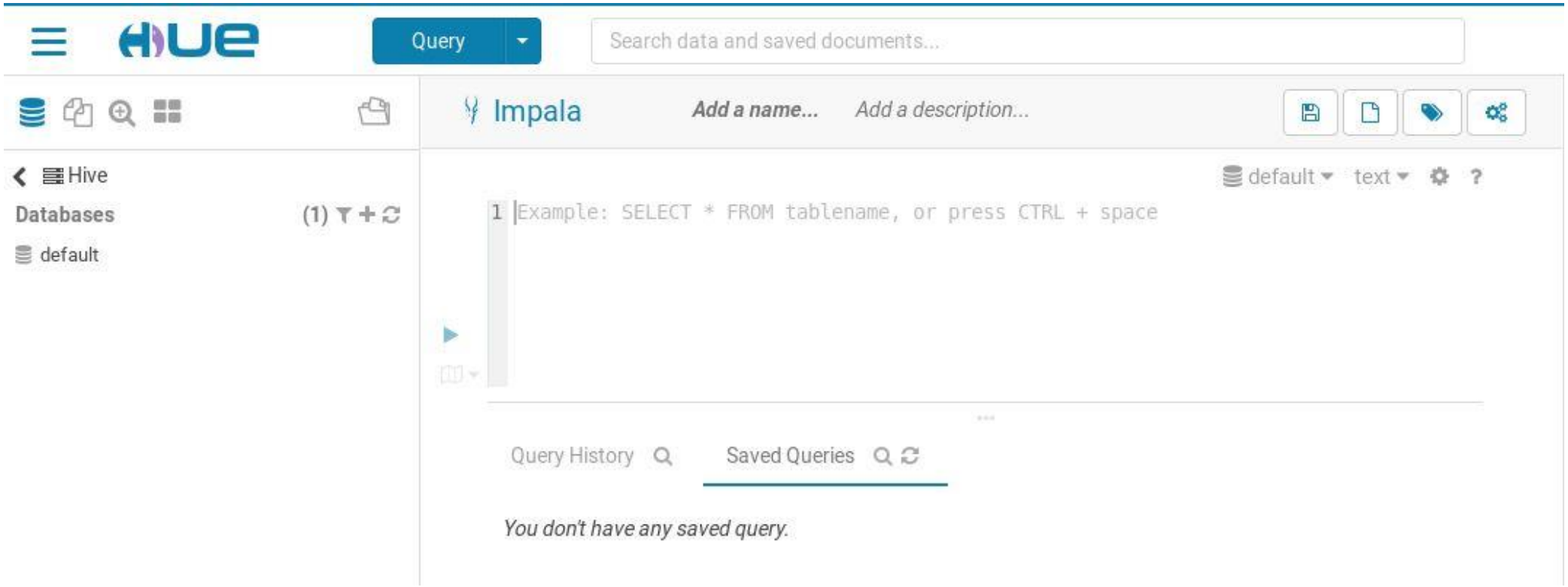


- **Oozie Workflows:** All of the past and previous workflows of Hadoop cluster can be checked through this workflow interface.

APACHE HIVE

- Apache Hive is developed by Facebook.
- It is a map-reduce development method used to process big data on Hadoop.
- It is an open-source library.

APACHE HIVE


Query
Search data and saved documents...
Impala
Add a name...
Add a description...
Hive
Databases
(1)
default
default
text
1 Example: SELECT * FROM tablename, or press CTRL + space
Query History
Saved Queries
You don't have any saved query.

APACHE HIVE

- Apache Hive is a data warehouse infrastructure build over Hadoop platform for performing data intensive tasks such as querying, analysis, processing and visualization.
- Apache Hive is versatile in its usage as it supports the analysis of large datasets stored in Hadoop's HDFS and other compatible file systems.
- Apache Hive uses an SQL – like language (HiveQL) and transparently converts queries to MapReduce and Spark jobs.

APACHE HIVE

- Apache Hive is a data warehouse infrastructure build over Hadoop platform for performing data intensive tasks such as querying, analysis, processing and visualization.
- Apache Hive is versatile in its usage as it supports the analysis of large datasets stored in Hadoop's HDFS and other compatible file systems.
- Apache Hive uses an SQL – like language (HiveQL) and transparently converts queries to MapReduce and Spark jobs.

APACHE IMPALA

- Apache Impala has an open source massively parallel processing (MPP) SQL engine.

APACHE IMPALA

- Cloudera Impala is an excellent choice for programmers for running queries on HDFS and Apache HBase as it doesn’t require data to be moved or transformed prior to processing.
- Cloudera Impala easily integrates with Hadoop ecosystem, as its file and data formats, metadata, security and resource management frameworks are same as those used by MapReduce, Apache Hive, Apache Pig and other Hadoop software.

HIVE & IMPALA

- Hive is written in Java.
- Impala is written in C++ and Java.

HIVE & IMPALA

- Hive is built over MapReduce and hence is slower than Impala for less complex queries due to many I/O operations that have to run, for single query execution.
- Hive is better able to handle longer-running, more complex queries on much larger datasets.
- Since Impala is not built over the MapReduce algorithms, the latency is reduced allowing Impala to run faster than Hive. Impala supports in-memory data processing, which means that it accessed data that is stored on the Hadoop data nodes without movement of data.

HIVE & IMPALA

- Hive supports complex types, but Impala does not.
- Apache Hive might not be ideal for interactive computing whereas Impala is meant for interactive computing.
- Apache Impala was primarily designed for speed. It is written in C++, a CPU efficient language that allows fast query execution and metadata caching.

HIVE & IMPALA

- Apache Hive was primarily built to handle sophistication. This makes it a little more difficult for beginners to get comfortable with it. However, Hive is very effective when it comes to running complex queries, possibly requiring heavy transformations and/or multiple joins. The latency in Hive is higher since queries take longer to execute since queries go through planning and ramp-up prior to execution.

HIVE & IMPALA

- Hive is also very fault tolerant. If a part of a long-running query fails, Hive will ensure that this part of the query is reassigned and tried again.
- Impala does not support fault tolerance.

	Apache Hive	Cloudera Impala
Latency	Hive is built on Hadoop's MapReduce and hence has a higher latency when it comes to processing queries. Hive was built mainly for sophistication and not speed.	Impala was built primarily for speed and hence has a very low latency.
Throughput	Hive has the lower throughput than impala.	Impala has the higher throughput than Hive.
Use cases	Hive is ideal for situations where multiuser support is required and complex queries.	Impala is best suited for business interactive workloads where a low latency is required, and queries have to be interactive
Fault tolerance	Hive can recover from mid-query faults.	Impala is not fault-tolerant. If query execution fails mid-query, then the query has to be executed again.
Query complexity	Hive is built to handle long running queries which require multiple transformations and joins.	Impala is built to handle shorter queries on large data sets, but due to its low latency, it is ideal for interactive computing.

APACHE HIVE CODING STEPS

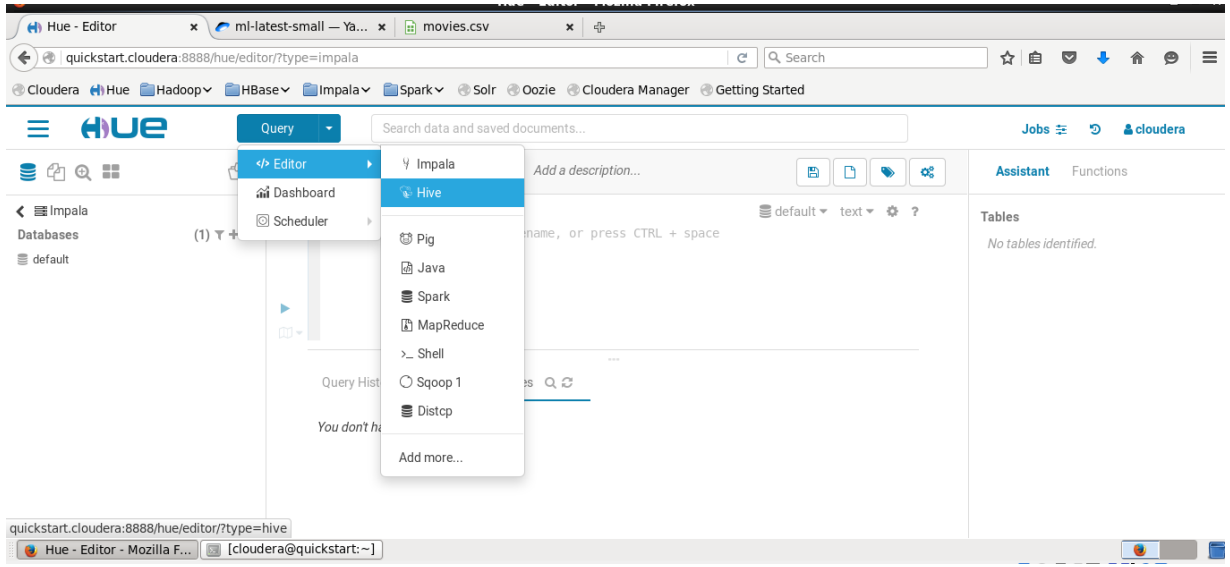
1. Export logs to HDFS
2. Create database.
3. Create table.
4. Write SQL query.

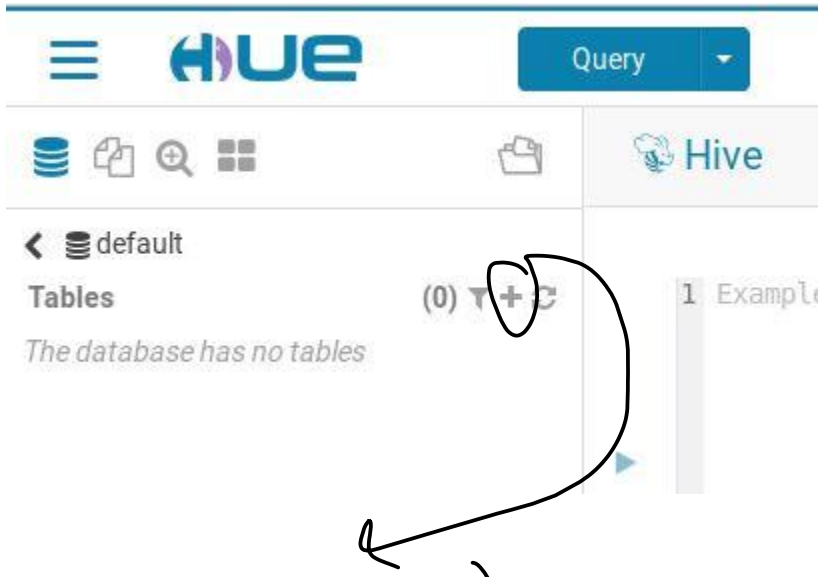
APACHE HIVE- APPLICATION

```
[cloudera@quickstart ~]$ hdfs dfs -mkdir /hive_example
```

```
[cloudera@quickstart ~]$ hdfs dfs -copyFromLocal /home/cloudera/Downloads/movies
.csv /hive example
```

APACHE HIVE- APPLICATION





create databases

APACHE HIVE- APPLICATION

Pick data from file /user/hive/warehouse/movie/movies.csv

Move it to table movies.m

Source

Type: File

Path: /user/hive/warehouse/movie/movies.csv

Format

Field Separator: Comma (,)

Record Separator: New line

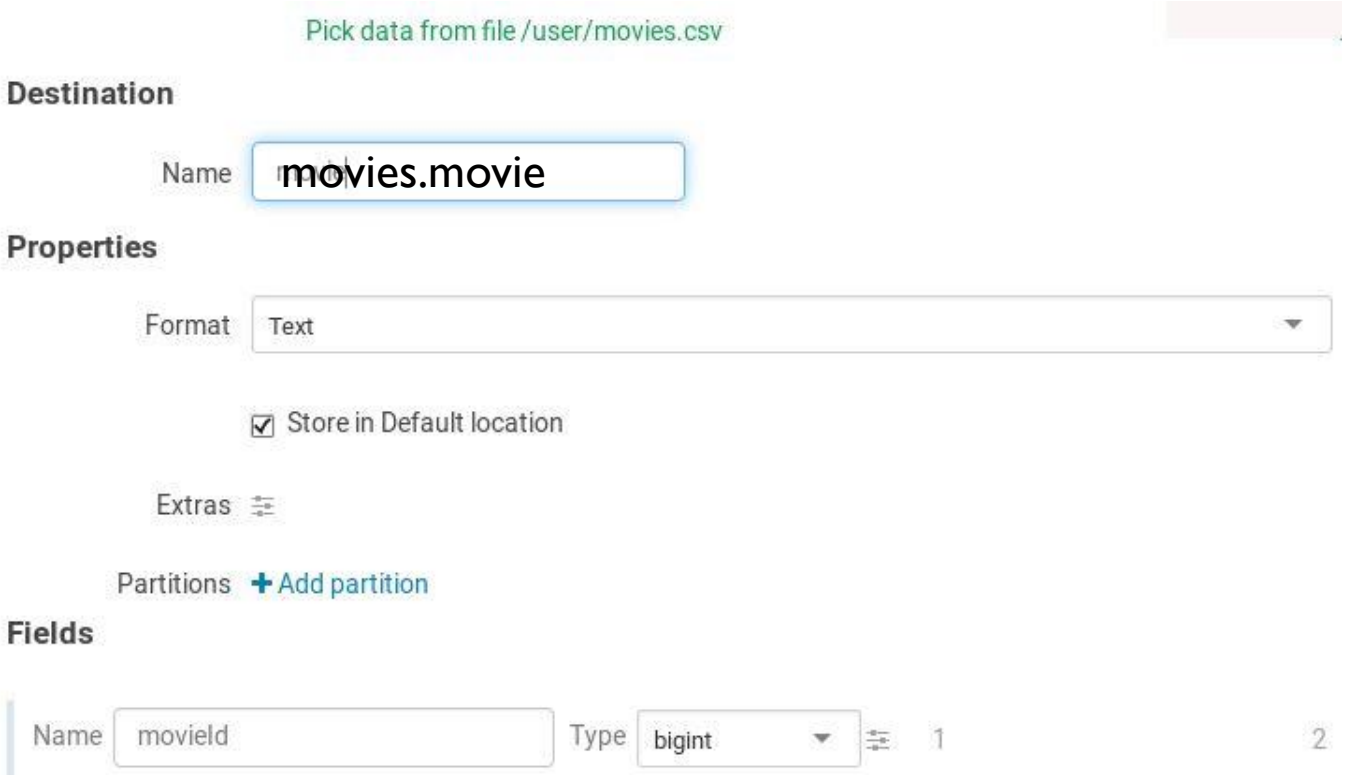
Quote Character: Double Quote

☑ Has Header

Preview

movieId	title	genres
1	Toy Story (1995)	Adventure|Animation|...
2	Jumanji (1995)	Adventure|Children|F...

APACHE HIVE-APPLICATION



APACHE HIVE-APPLICATION

Databases > default > movie

No description available

Overview | Columns (3) | Sample | Details

PROPERTIES

Table
cloudera
09/20/2021 11:43 AM
text Not compressed

STATS

Location
1 files
447.65 KB

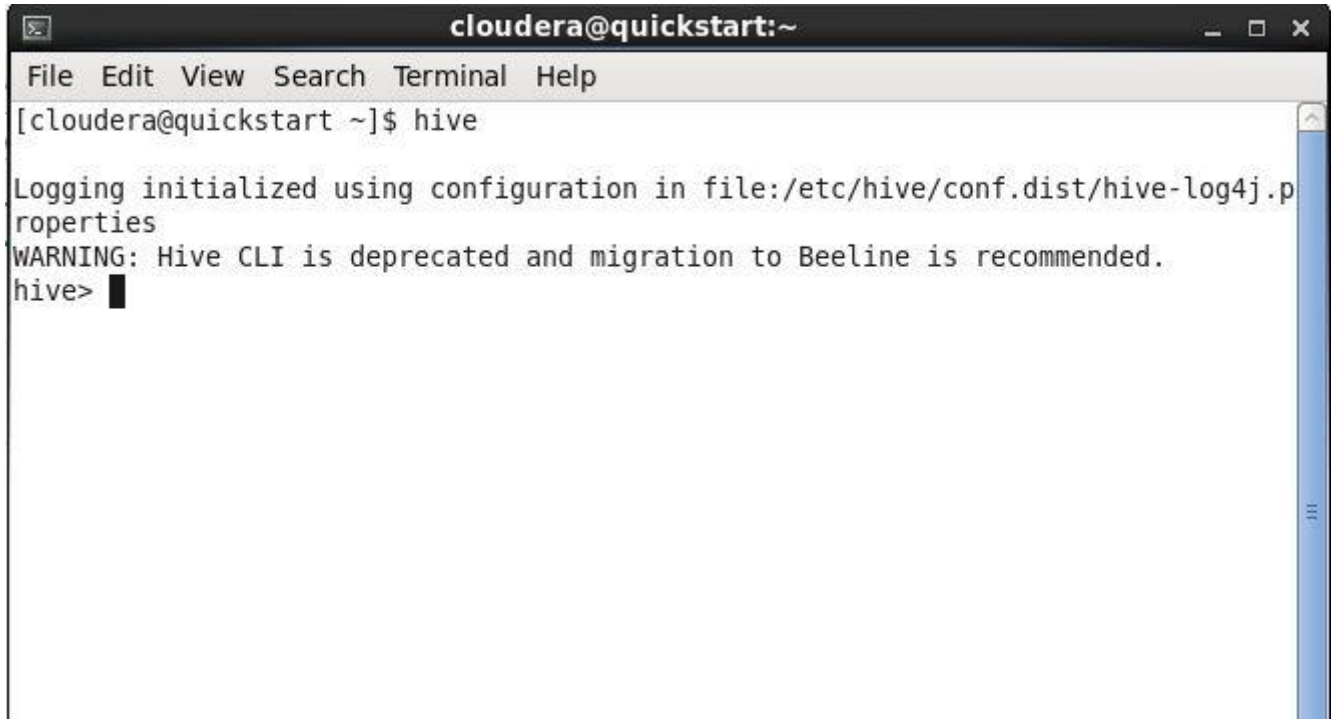
COLUMNS (3)

		Name	Type	Comment
1	i	movieid	bigint	*Add a comment...*

APACHE HIVE-APPLICATION

1. Select title from movie where genres=‘Comedy’
2. Select title from movie where genres like ‘%Comedy%’
3. Select count(title) from movie where genres like ‘%Comedy%’
4. Select title, genres from movie where title like ‘%1995%’
5. Select title, genres from movie where title like ‘%1995%’ and genres like ‘%Comedy%’

APACHE HIVE-APPLICATION



ASSIGNMENT

- *http://www.bit.ly/nyt_march2018*