Micro Data Center
&
Hadoop Big Data WareHouse

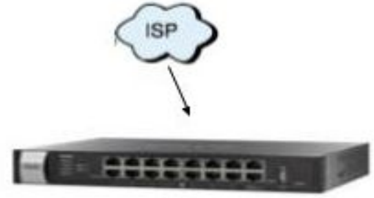
V0.01, r 2018

Open Source Platform

- Micro Data Center 25 TB, Small Business Solution (Plug & Play)
- Hadoop Open Source Technology
- Hive Data Warehouse
- Hadoop Testing Data model
- Software & Tools Library
- Business Intelligence report

Infrastructure V0.07

Micro Data Center



Cisco Small Business Router

Wifi Cisco Small Business Router

Cisco Small Business Smart Switch

WD Sentinel Server - Xeon E3

Seagate Personal Cloud NAS Server
(Backup - Set 1)

Seagate Backup Plus External Hard Drive
(Backup - Set 2)

cloudera

ubuntu Linux

amazon web services™

HIVE

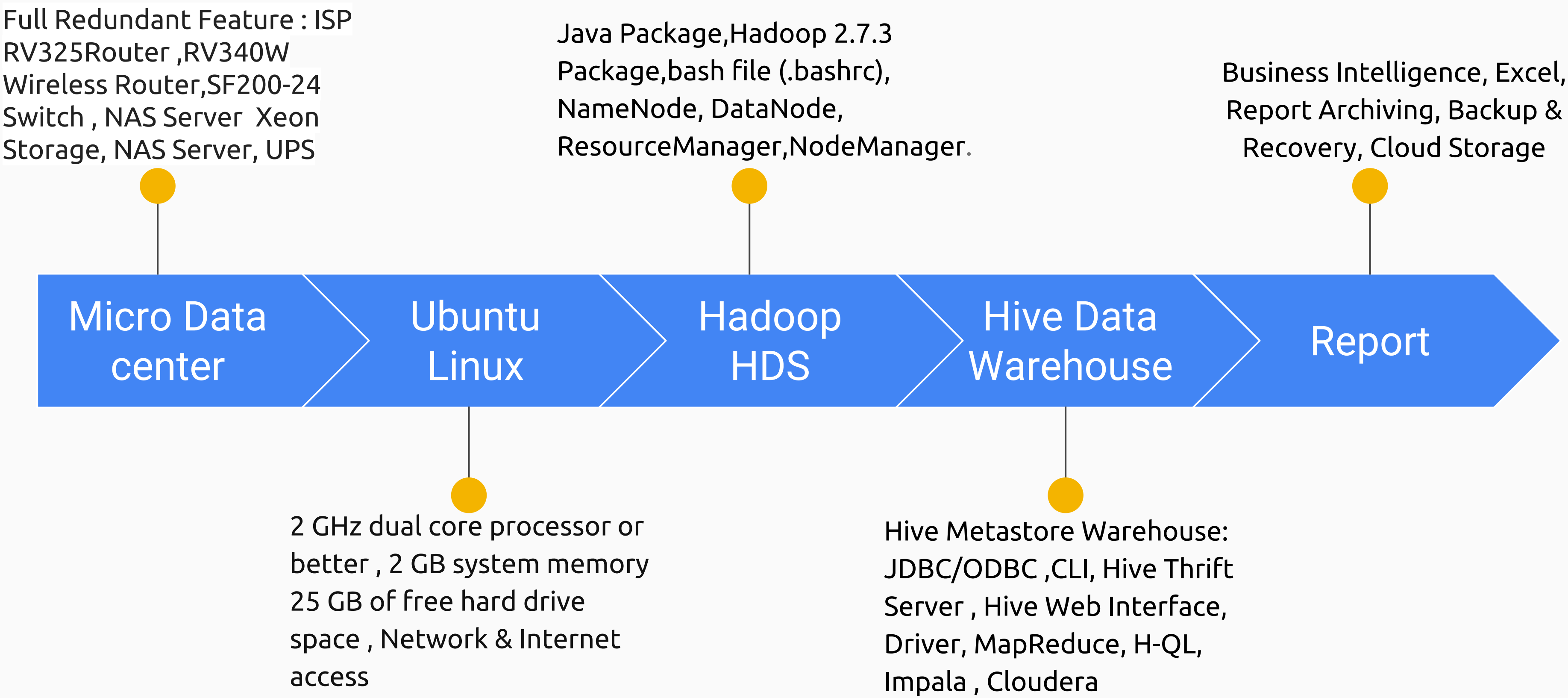
Impala

hadoop

Excel

The Data Warehouse

SQL



Micro Data center

Full Redundant Feature : ISP RV325Router ,RV340W Wireless Router,SF200-24 Switch , NAS Server Xeon Storage, NAS Server, UPS

Ubuntu Linux

2 GHz dual core processor or better , 2 GB system memory 25 GB of free hard drive space , Network & Internet access

Hadoop HDS

Java Package,Hadoop 2.7.3 Package,bash file (.bashrc), NameNode, DataNode, ResourceManager,NodeManager.

Hive Data Warehouse

Hive Metastore Warehouse: JDBC/ODBC ,CLI, Hive Thrift Server , Hive Web Interface, Driver, MapReduce, H-QL, Impala , Cloudera

Report

Business Intelligence, Excel, Report Archiving, Backup & Recovery, Cloud Storage

Micro Data Center (Hardware Specification)

Networking :

1. Cisco Small Business RV325 Router - 14-port - Gigabit Ethernet
2. Cisco Small Business RV340W Wireless Router - 2.4 GHz / 5 GHz
3. Cisco Small Business Smart SF200-24 Switch - 24 Ethernet Ports

Central Storage :

1. WD Sentinel DS5100 WDBYVE0080KBK Server Xeon - 15 TB
2. Seagate Personal Cloud STCR3000101 NAS Server -
3. Seagate 5TB Backup Plus External Hard Drive - 5TB

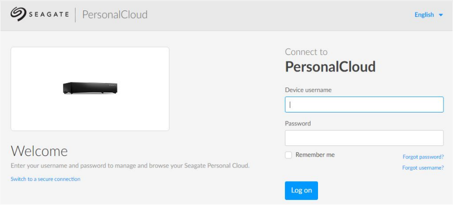
Redundant Power UPS :

1. OL1000RTXL2U, Runtime @ 450 W: 20 min

Network Connectivity

Micro Data Center storage v0.07

& Linux Ubuntu Workstation v18.04 LTS



Welcome

Enter your username and password to manage and browse your Seagate Personal Cloud.

Switch to a secure connection

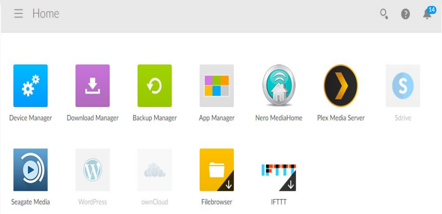
Connect to
PersonalCloud

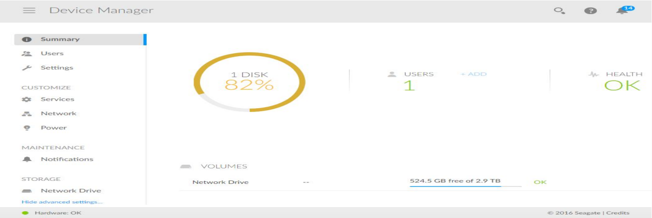
Device username

Password

□ Remember me

Forgot password?
Forgot username?





1 DISK 82%

USERS 1

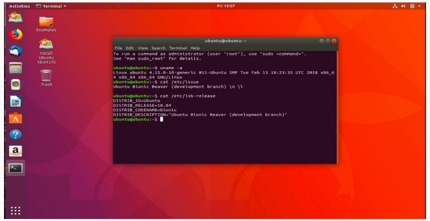
HEALTH OK

VOLUMES

Network Drive -- 524.5 GB free of 2.9 TB OK

Hardware: OK

© 2016 Seagate | Credits



```
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

ubuntu@ubuntu:~$ uname -a
Linux ubuntu 4.15.0-10-generic #11-Ubuntu SMP Tue Feb 13 18:23:35 UTC 2018 x86_6
4 x86_64 x86_64 GNU/Linux
ubuntu@ubuntu:~$ cat /etc/issue
Ubuntu Bionic Beaver (development branch) \n \l

ubuntu@ubuntu:~$ cat /etc/lsb-release
DISTRIB_ID=Ubuntu
DISTRIB_RELEASE=18.04
DISTRIB_CODENAME=bionic
DISTRIB_DESCRIPTION="Ubuntu Bionic Beaver (development branch)"
ubuntu@ubuntu:~$
```

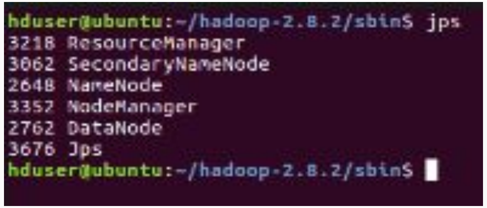
Connect to Micro Data Center Storage:

1. Connect Network/Wifi router
2. PersonalCloude : http://192.168.1.82/
   Device user’s name & PW
3. Network Configuration for Micro Datacenter Storage

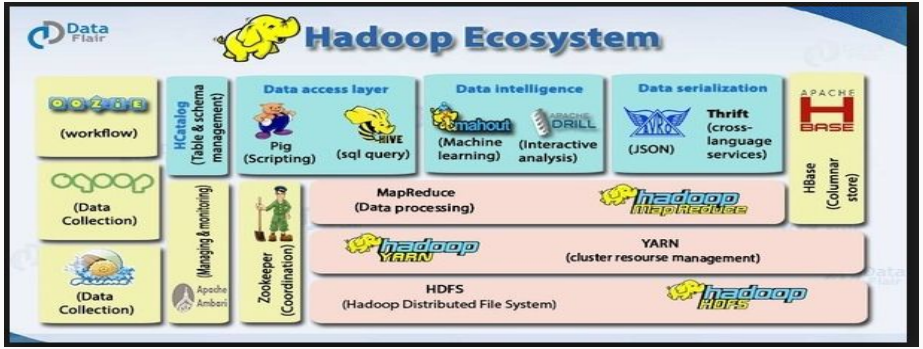
Linux Ubuntu (Workstation) to Micro Data Center:

1. Boot from USB/DVD
2. Prepare to install Ubuntu
3. Allocate drive space
4. Begin installation
5. Login as User Admin for Storage

Hadoop Installation v2.7.3



```
hduser@ubuntu:~/hadoop-2.8.2/sbin$ jps
3218 ResourceManager
3062 SecondaryNameNode
2648 NameNode
3352 NodeManager
2762 DataNode
3676 Jps
hduser@ubuntu:~/hadoop-2.8.2/sbin$
```



Install Hadoop

1. Download & Installation :

a. Linux Ubuntu

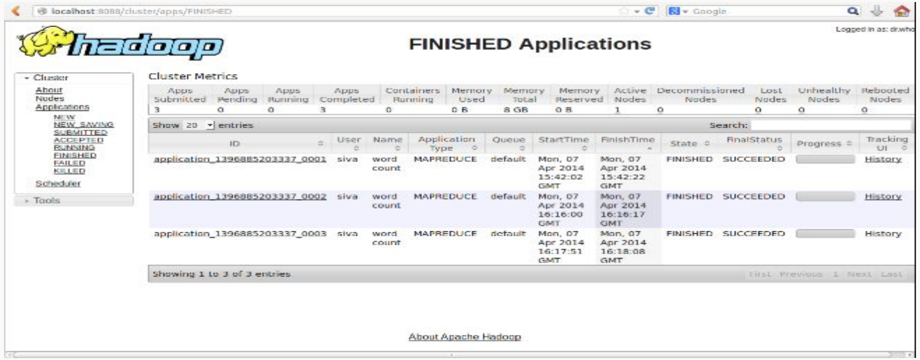
b. Java JDK c. Vim CLI (command line interface) d.hadoop-2.6.5.tar.gz

2. Group & Admin for Hadoop User

3. Configuration : a. sysctl.conf (ipv6) b. Generating public/private rsa key pair c. ssh localhost d..bashrc (Hadoop Variables), e. Hadoop Core conf files (hadoop-env.sh, core-site.xml,mapred-site.xml & hdfs-site.xml) f. Namenode, Datanode & hadoop_store g. Namenode format h. Start-all.sh jps

4. Hadoop Daemons JPS (ResorceManager, SecondaryNamenode, NodeManager, & Datanode)

Hadoop Web Interface



FINISHED Applications

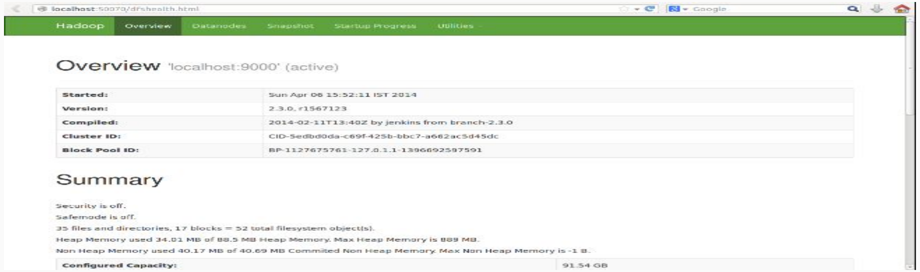
Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
3	0	0	3	0	0 B	8 GB	0 B	1	0	0	0	0

Show 20 entries Search:

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI
application_1396885203337_0001	siva	word count	MAPREDUCE	default	Mon, 07 Apr 2014 15:42:02 GMT	Mon, 07 Apr 2014 15:42:22 GMT	FINISHED	SUCCEEDED		History
application_1396885203337_0002	siva	word count	MAPREDUCE	default	Mon, 07 Apr 2014 16:16:00 GMT	Mon, 07 Apr 2014 16:16:17 GMT	FINISHED	SUCCEEDED		History
application_1396885203337_0003	siva	word count	MAPREDUCE	default	Mon, 07 Apr 2014 16:17:51 GMT	Mon, 07 Apr 2014 16:18:08 GMT	FINISHED	SUCCEEDED		History

Showing 1 to 3 of 3 entries



Overview 'localhost:9000' (active)

Started:	Sun Apr 06 15:52:11 IST 2014
Version:	2.3.0, r1567123
Compiled:	2014-02-11T13:40Z by jenkins from branch-2.3.0
Cluster ID:	CID-5edbd0da-c69f-425b-bbc7-a662ac5d45dc
Block Pool ID:	BP-1127675761-127.0.1.1-1396692597591

Summary

Security is off.

Safemode is off.

35 files and directories, 17 blocks = 52 total filesystem object(s).

Heap Memory used 34.01 MB of 88.5 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 40.17 MB of 40.69 MB Commited Non Heap Memory. Max Non Heap Memory is -1 B.

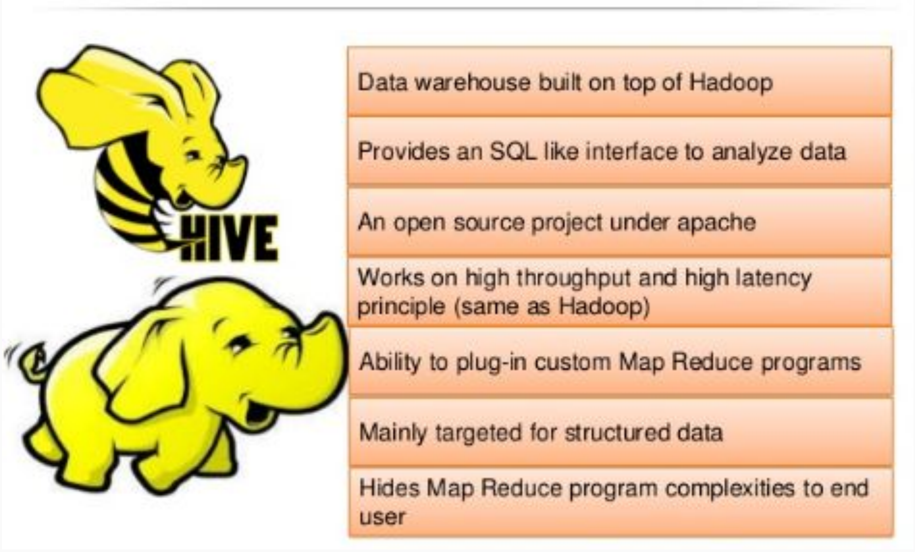
Configured Capacity:	91.54 GB

- **http://localhost:50070/ of the NameNode daemon :**
  Namenode Summary report,Security , Safemode status, DFS Used%, DFS Remaining%, Block Pool Used, DataNodes usages%, Live Nodes, Dead Nodes, Decommissioning Nodes, Number of Under-Replicated Blocks, NameNode Journal Status, Journal Manager, NameNode Storage

- **Datanode Information :**
  Node, Admin State, Capacity, Used, Non DFS Used, Remaining, Block pool used, Failed Volumes.

- **Browsing HDFS :**
  Browse Directory, Permission, Owner, Group, Size, Block Size, Folder Name.

Hive
Data Warehouse Implementation v2.0



Data warehouse built on top of Hadoop

Provides an SQL like interface to analyze data

An open source project under apache

Works on high throughput and high latency principle (same as Hadoop)

Ability to plug-in custom Map Reduce programs

Mainly targeted for structured data

Hides Map Reduce program complexities to end user

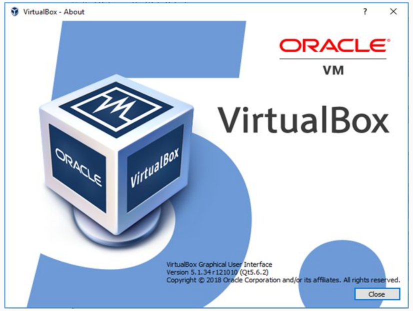
Step1. **Hive Installation**
- Download the Hive
- Configure ~/.bashrc and set the environment variables

Step2. **Hive Warehouse Directory Creation**
- Hive is based on Hadoop platform in Hadoop in PATH
- HDFS create the Hive Warehouse Directory

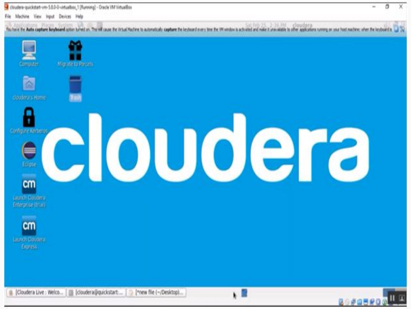
Step3. **Hive Configuration**
- Configure Hive with Hadoop
- Congigure “hive-env.sh” file
- Configure to external database to configure Metastore

Step4. **Hive Data Warehouse Files Location**
- $hadoop fs-ls /user/hive/warehouse

VirtualBox Installation v5.1.34r

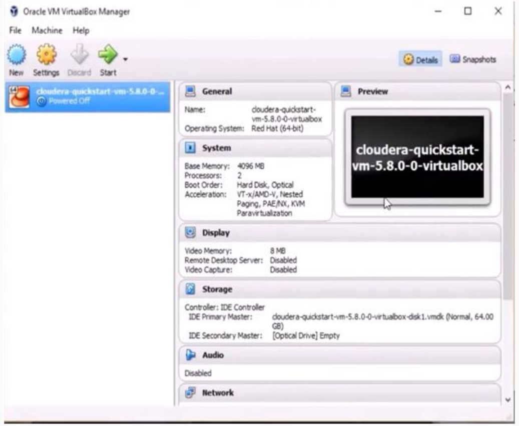


VirtualBox Graphical User Interface
Version 5.1.34 r121010 (Qt5.6.2)
Copyright © 2018 Oracle Corporation and/or its affiliates. All rights reserved.



cloudera-quickstart-vm-5.8.0-0-virtualbox_1 [Running] - Oracle VM VirtualBox

You have the **Auto capture keyboard** option turned on. This will cause the Virtual Machine to automatically **capture** the keyboard every time the VM window is activated and make it unavailable to other applications running on your host machine: when the keyboard is


Oracle VM VirtualBox Manager
File
Machine
Help
New
Settings
Discard
Start
Details
Snapshots
cloudera-quickstart-vm-5.8.0-0-...
Powered Off
Preview
cloudera-quickstart-vm-5.8.0-0-virtualbox

General
Name: cloudera-quickstart-vm-5.8.0-0-virtualbox
Operating System: Red Hat (64-bit)

System
Base Memory: 4096 MB
Processors: 2
Boot Order: Hard Disk, Optical
Acceleration: VT-x/AMD-V, Nested Paging, PAE/NX, KVM Paravirtualization

Display
Video Memory: 8 MB
Remote Desktop Server: Disabled
Video Capture: Disabled

Storage
Controller: IDE Controller
IDE Primary Master: cloudera-quickstart-vm-5.8.0-0-virtualbox-disk1.vmdk (Normal, 64.00 GB)
IDE Secondary Master: [Optical Drive] Empty

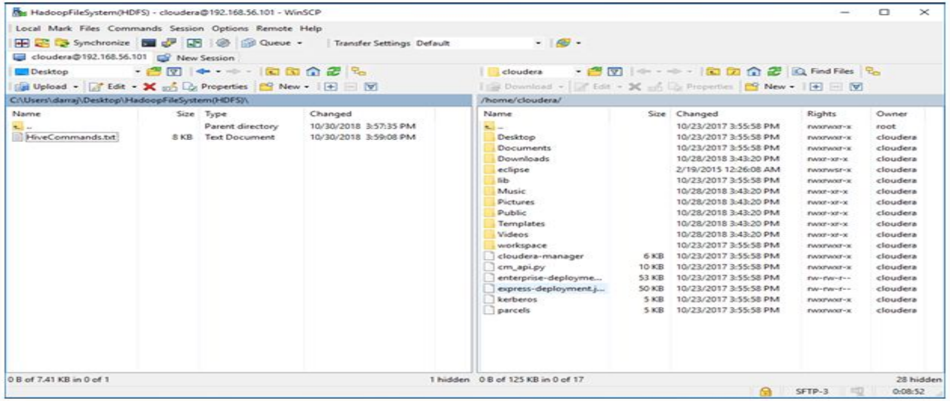
Audio
Disabled

Network

R&D platform (Cloudera QuickStart)

- Download from Oracle Virtual Box.org
- Configure Network Interfaces
- Open VM virtualBox Manager
- Appliance to import
- Appliance Settings : Name, Guest OS Type, CPU 2, RAM 10GB, DVD, Network Adapter
- Enable Network :
- Adapter 1: Inter PRO/1000 MT Desktop (NAT)
- Adapter 2: Configure Host only Adapter (VirtualBox Host-only Ethernet Adapter 2)
- System : Motherboard, Base Memory :10GB ,Processor 2 CPU
- VirtualBox running : Booting CentOS 6 (2.6.32-573.e16.x86_64)

Tools & Software Library



HadoopFileSystem(HDFS) - cloudera@192.168.56.101 - WinSCP

cloudera@192.168.56.101

C:\Users\darraj\Desktop\HadoopFileSystem(HDFS)\

Name	Size	Type	Changed
..		Parent directory	10/30/2018 3:57:35 PM
HiveCommands.txt	8 KB	Text Document	10/30/2018 3:59:08 PM

0 B of 7.41 KB in 0 of 1 — 1 hidden

/home/cloudera/

Name	Size	Changed	Rights	Owner
..		10/23/2017 3:55:58 PM	rwxrwxr-x	root
Desktop		10/23/2017 3:55:58 PM	rwxrwxr-x	cloudera
Documents		10/23/2017 3:55:58 PM	rwxrwxr-x	cloudera
Downloads		10/28/2018 3:43:20 PM	rwxr-xr-x	cloudera
eclipse		2/19/2015 12:26:08 AM	rwxrwsr-x	cloudera
lib		10/23/2017 3:55:58 PM	rwxrwxr-x	cloudera
Music		10/28/2018 3:43:20 PM	rwxr-xr-x	cloudera
Pictures		10/28/2018 3:43:20 PM	rwxr-xr-x	cloudera
Public		10/28/2018 3:43:20 PM	rwxr-xr-x	cloudera
Templates		10/28/2018 3:43:20 PM	rwxr-xr-x	cloudera
Videos		10/28/2018 3:43:20 PM	rwxr-xr-x	cloudera
workspace		10/23/2017 3:55:58 PM	rwxrwxr-x	cloudera
cloudera-manager	6 KB	10/23/2017 3:55:58 PM	rwxrwxr-x	cloudera
cm_api.py	10 KB	10/23/2017 3:55:58 PM	rwxrwxr-x	cloudera
enterprise-deployme...	53 KB	10/23/2017 3:55:58 PM	rw-rw-r--	cloudera
express-deployment.j...	50 KB	10/23/2017 3:55:58 PM	rw-rw-r--	cloudera
kerberos	5 KB	10/23/2017 3:55:58 PM	rwxrwxr-x	cloudera
parcels	5 KB	10/23/2017 3:55:58 PM	rwxrwxr-x	cloudera

0 B of 125 KB in 0 of 17 — 28 hidden

SFTP-3 0:08:52

ETL ELT

Run in the ETL grid
Run in the database
Run in Hadoop

WINSCP Ftp infteracе between Window & Linux

- SSH and SPC code based on Putty
- Login : New Site > File protocol :FSTP, Host name IP : 192.168.56.101, Port number:22 and Username/PW.
- Open Two different OS windows with Window OS:c:\Users\Document and Linux OS: /home/cloudera/
- Upload & Download file : File Upload to Linux/Window OS and File Download to Windows/Linux OS

PuTTY Key Generator :

- Private key file for authentication
- Public key for pasting into OpenSSH authorized File,
- Type of Key Parameters RSA, Save public key

PuTTY release 0.70 :

- Host Name IP : 3.17.0.143 & Port 22
- SSH authentication : Private key file for authentication
- Controlling session logging : Open remote terminal 3.17.0.143-Putty

ETL (extract, transform, load) ELT (extract, load, transform) : SQOOP

Cloudera CDH5.3.

Business Data Testing & Analysis 12k+ Customers

Hadoop, Hive & Impala (SQL) , Source Cloudera@quickstart

MySQL (retail_db)

```
mysql> show tables;
```

Table	Records
categories	58
customers	12,435
Departments	6
order_items	68,883
orders	1,72,198
products	1,345

BigData/Hive/Impala

```
hive> show tables;
```

Table	Records
categories	58
customers	12,435
Departments	6
order_items	68,883
orders	1,72,198
products	1,345

MySQL (retail_db) :

```
mysql> show databases;
mysql> use retail_db;
mysql> select count(*) from customers;
```

Sqoop :

```
[cloudera@quickstart ~]$ sqoop import-all-tables \
```

Hive (retail_db) :

```
hive> show databases;
hive> use default;
hive> show tables;
hive> select count(*) from customers;
```

Hive Data Warehouse :

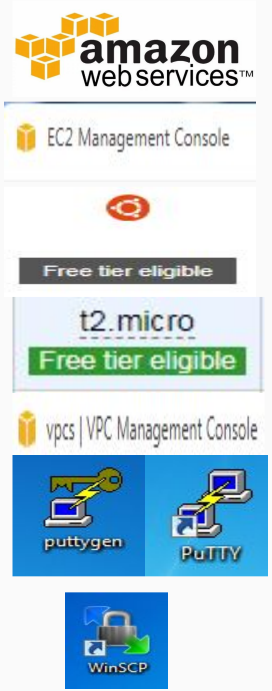
```
[cloudera@quickstart ~]$  hadoop fs -ls /user/hive/warehouse/
```

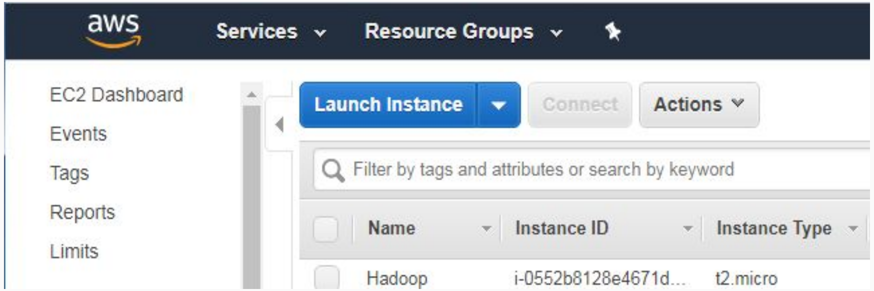
Business Intelligence Report :

Most popular product categories

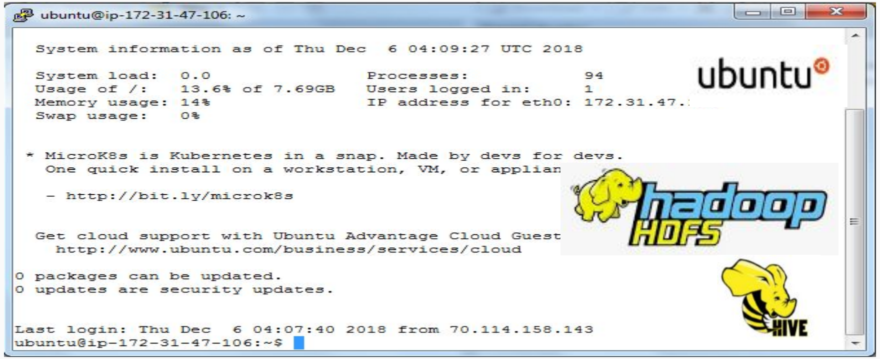
Top 10 revenue generating products

AWS Cloud Services Ubuntu Server 18.04 LTS





Name	Instance ID	Instance Type
Hadoop	i-0552b8128e4671d...	t2.micro



```
ubuntu@ip-172-31-47-106: ~

  System information as of Thu Dec  6 04:09:27 UTC 2018

  System load:  0.0               Processes:            94
  Usage of /:   13.6% of 7.69GB   Users logged in:      1
  Memory usage: 14%               IP address for eth0: 172.31.47.
  Swap usage:   0%


 * MicroK8s is Kubernetes in a snap. Made by devs for devs.
   One quick install on a workstation, VM, or applian

   - http://bit.ly/microk8s


  Get cloud support with Ubuntu Advantage Cloud Guest
    http://www.ubuntu.com/business/services/cloud

0 packages can be updated.
0 updates are security updates.


Last login: Thu Dec  6 04:07:40 2018 from 70.114.158.143
ubuntu@ip-172-31-47-106:~$
```

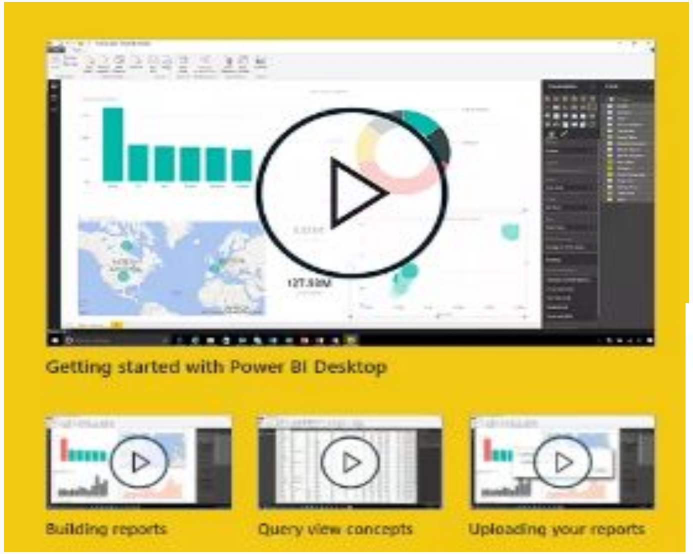
AWS Management Console:

- Step 1: Amazon Machine Image Ubuntu Server 18.04 LTS
- Step 2: Build an Instance
- Step 3: Configure Instance Details
- Step 4: Add Storage
- Step 5: Add Tags
- Step 6: Configure Security Group
- Step 7: Review Instance Launch

Connect AWS Management Console:

- Connect ubuntu@ip-172.47.106:~$
- Generate Private Key by PPuttygen
- Connect AWS from Putty FTP
- File Transfer by WinSCP SFTP

Business Intelligence Reporting Tools v2.65



Report Generation

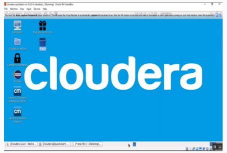
- Opening BI window
- Run Power BI Desktop
- Import Data from different source
- Connecting Dataset
- Load Data into BI
- Management Data as per Query
- Export as BI / Export to PDF

Prototype Demo Cloudera (Remote Login)

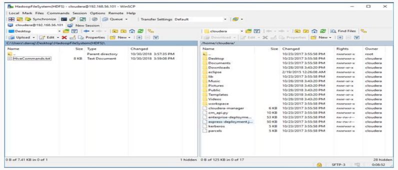
TeamViewer

Windows 10

ubuntu Linux



amazon web services™


HadoopFileSystem(HDFS) - cloudera@192.168.56.101 - WinSCP
Local Mark Files Commands Session Options Remote Help
Synchronize
Queue
Transfer Settings Default
cloudera@192.168.56.101
New Session
Desktop
Upload
Edit
Properties
New
cloudera
Find Files
Download

C:\Users\darraj\Desktop\HadoopFileSystem(HDFS)\

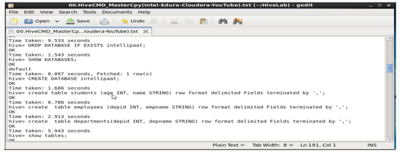
Name	Size	Type	Changed
..		Parent directory	10/30/2018 3:57:35 PM
HiveCommands.txt	8 KB	Text Document	10/30/2018 3:59:08 PM

/home/cloudera/

Name	Size	Changed	Rights	Owner
..		10/23/2017 3:55:58 PM	rwxrwxr-x	root
Desktop		10/23/2017 3:55:58 PM	rwxrwxr-x	cloudera
Documents		10/23/2017 3:55:58 PM	rwxrwxr-x	cloudera
Downloads		10/28/2018 3:43:20 PM	rwxr-xr-x	cloudera
eclipse		2/19/2015 12:26:08 AM	rwxrwsr-x	cloudera
lib		10/23/2017 3:55:58 PM	rwxrwxr-x	cloudera
Music		10/28/2018 3:43:20 PM	rwxr-xr-x	cloudera
Pictures		10/28/2018 3:43:20 PM	rwxr-xr-x	cloudera
Public		10/28/2018 3:43:20 PM	rwxr-xr-x	cloudera
Templates		10/28/2018 3:43:20 PM	rwxr-xr-x	cloudera
Videos		10/28/2018 3:43:20 PM	rwxr-xr-x	cloudera
workspace		10/23/2017 3:55:58 PM	rwxrwxr-x	cloudera
cloudera-manager	6 KB	10/23/2017 3:55:58 PM	rwxrwxr-x	cloudera
cm_api.py	10 KB	10/23/2017 3:55:58 PM	rwxrwxr-x	cloudera
enterprise-deployme...	53 KB	10/23/2017 3:55:58 PM	rw-rw-r--	cloudera
express-deployment.j...	50 KB	10/23/2017 3:55:58 PM	rw-rw-r--	cloudera
kerberos	5 KB	10/23/2017 3:55:58 PM	rwxrwxr-x	cloudera
parcels	5 KB	10/23/2017 3:55:58 PM	rwxrwxr-x	cloudera

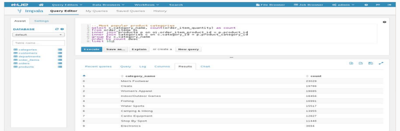
0 B of 7.41 KB in 0 of 1 1 hidden 0 B of 125 KB in 0 of 17 28 hidden

SFTP-3 0:08:52


00.HiveCMD_MasterCpy(Intel-Edura-Cloudera-YouTube).txt (~/HiveLab) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
00.HiveCMD_MasterCp...loudera-YouTube).txt

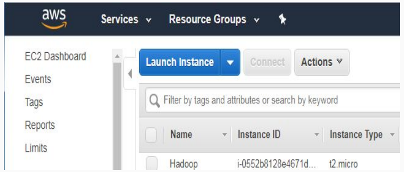
```
Time taken: 9.533 seconds
hive> DROP DATABASE IF EXISTS intellipaat;
OK
Time taken: 1.543 seconds
hive> SHOW DATABASES;
OK
default
Time taken: 0.097 seconds, Fetched: 1 row(s)
hive> CREATE DATABASE intellipaat;
OK
Time taken: 1.686 seconds
hive> create table students (age INT, name STRING) row format delimited Fields terminated by ',';
OK
Time taken: 0.786 seconds
hive> create  table employees (depid INT, empname STRING) row format delimited Fields terminated by ',';
OK
Time taken: 2.913 seconds
hive> create  table departments(depid INT, depname STRING) row format delimited Fields terminated by ',';
OK
Time taken: 5.943 seconds
hive> show tables;
OK
```

Plain Text Tab Width: 8 Ln 191, Col 1 INS



```
-- Most popular product categories
select c.category_name, count(order_item_quantity) as count
from order_items oi
inner join products p on oi.order_item_product_id = p.product_id
inner join categories c on c.category_id = p.product_category_id
group by c.category_name
order by count desc
limit 10;
```

	category_name	count
0	Men's Footwear	23029
1	Cleats	19786
2	Women's Apparel	19685
3	Indoor/Outdoor Games	18456
4	Fishing	16991
5	Water Sports	15517
6	Camping & Hiking	13955
7	Cardio Equipment	12827
8	Shop By Sport	11446
9	Electronics	3694



Name	Instance ID	Instance Type
Hadoop	i-0552b8128e4671d...	t2.micro

Connected over Cloud Through :

- Team Viewer
- Windows10
- VirualBox
- Start Cloudera Desktop
- Cloudera CLI Terminal
- Run Mysql database
- Run HIVE open source database
- Cloudera QuickStart Hive/Impala SQL terminal
- SQL Data Analysis

FTP (File Transfer Protocol)

- Run WinSCP
- Connect Window Desktop to Linux Desktop

Business Intelligence Report

- Visual Analytics at your fingertips and creating interactive data visualizations and reports.

Meetup

Registration :

- Free Orientation
- Prototype demo
- Consultancy

Info & Registration :
Micro DataCenter & Data Warehouse
MDCDWH@gmail.com
https://goo.gl/forms/SuCTolEeZNNIL35V2