Big Data Hadoop & Spark

Session 1 Assignment

Introduction

Big Data

According to Gartner, Big Data is a high-volume, high-velocity, and high-variety information asset that demands cost-effective, innovative forms of information processing for enhanced insight and decision making

3 Vs of Big Data

Volume	**Variety**	**Velocity**
Data size	Data Sources	Speed of Change
• Terabyte	Structured	Batch
• Records	Unstructured	Near-Time
• Transactions	Semi Structured	Real time
• Tables/Files	All of the above	Streams

Exploding of the Problem

Big Data constitutes large data sets in petabytes & zettabyte which cannot be processed by a single machine within an expected timeframe.

Big Data Challenges

- Big Data Storage
- Big Data Processsing
- Manual Distributed Computing

Possible Solutions

- Scale up
- Scale out

Scale up

- Increase the configuration of a single system, like disk capacity, RAM, data transfer speed, etc.
- Complex, costly, and a time consuming process

Scale Out

- Use multiple commodity (economical) machines and distribute the load of storage/processing among them.
- Economical and quick to implement as it focuses on distribution of load
- Instead of having a single system with 10 TB of storage and 80 GB of RAM, use 40 machines with 256 GB of storage and 2 GB of RAM.

Solution For Big Data Explosion-Hadoop

Apache Hadoop

Apache Hadoop is an open source framework which provides an automated distributed computing environment that supports storage of big data sets. It does that storage using a cluster of commodity machines. It then analyses this stored big data using a very simple programming model.

The storage mechanism is known as **HDFS** (Hadoop Distributed File System). It is based on google GFS(Google File System) white paper.

The analytical mechanism is known as **Map Reduce** and is based on google map reduce white paper.

Apache Hadoop Philosophies

There are 4 basic philosophies on which hadoop works.

a) All the basic software that helps start a hadoop cluster is a software daemons.
b) All the above daemons are based on master and slave architecture.
c) The entire hadoop framework is divided into 2 broad parts - storage (HDFS) and processing (Map Reduce).

- Single Node Hadoop Installation (dev and testing)
- Mulinode Installation (Production)

Hadoop 2.x

- **HDFS (Storage Part)**
- **Master Daemon** - **Namenode** (High End Admin Machine) (1 in number)
- Back up Master Daemon - Secondary Namenode (High End Admin Machine) (1 in number)
- **Slave Daemons** - **Datanode** (Commodity Machines) (**Many in number)**
- **Map Reduce (Processing Part)** - YARN (Yet Another Resource Negotiater)
- **Master Daemon** - **ResourceManager** (High End Admin Machine) (1 in number)
- **Slave Daemon** - **NodeManager** (Commodity Machines) (Many in number)

d) Hadoop is a batch oriented system which can never be plugged behind and online transaction processing system. Moreover, it a write once, ready many times data storage mechanism. This means, you can never update the data. If you really want to update, you need to delete the previous version and upload a new copy.
Data Node is a slave daemon software for the storage part of hadoop.
Likewise, Node manager is a slave daemon software for the processing part of hadoop.
When you refer to hardware in hadoop, you always refer to it as either a commodity machines or a slave node.
Important point to remember, there is a difference between “Slave daemon” and “slave node” (which btw is also called as commodity hardware.)
Both the datanode slave daemon and node manager slave daemon run on a commodity machine or a slave machine (which is a hardware)

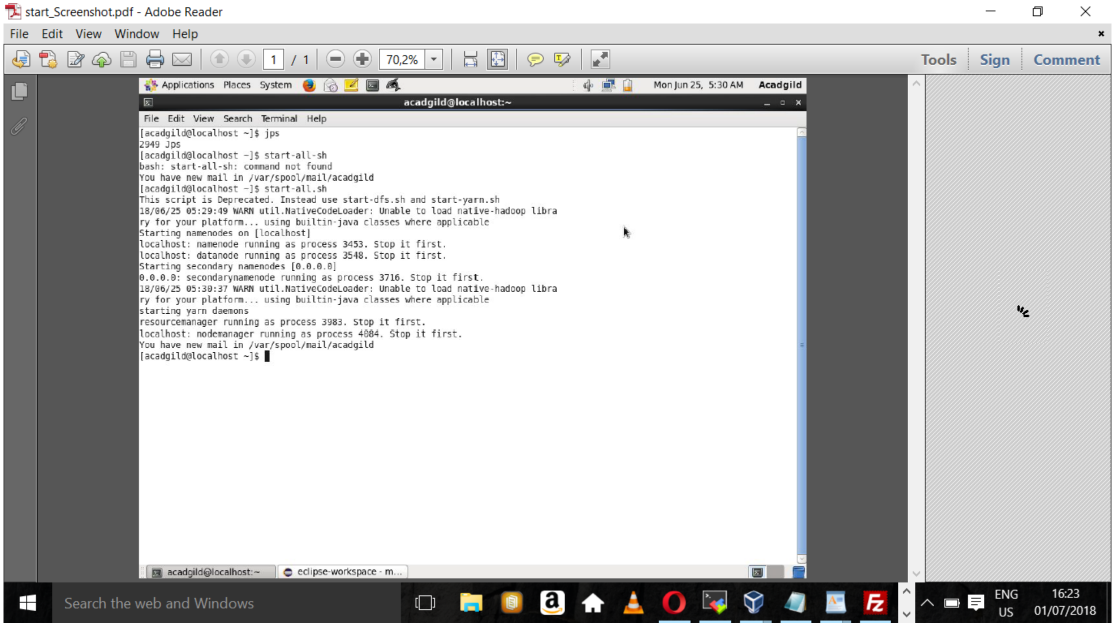
- Structured Data (Big and Small)
- Unstructured Data (Big and Small)
- Semi - Structure Data (Email, XML)

10 GB - 1 Machine -1 hr to process the data
10 machines - 1 GB of Data in each machine (60/10=6 minutes)

Cluster - a combination of machines which collabrates to give you a combined processing power

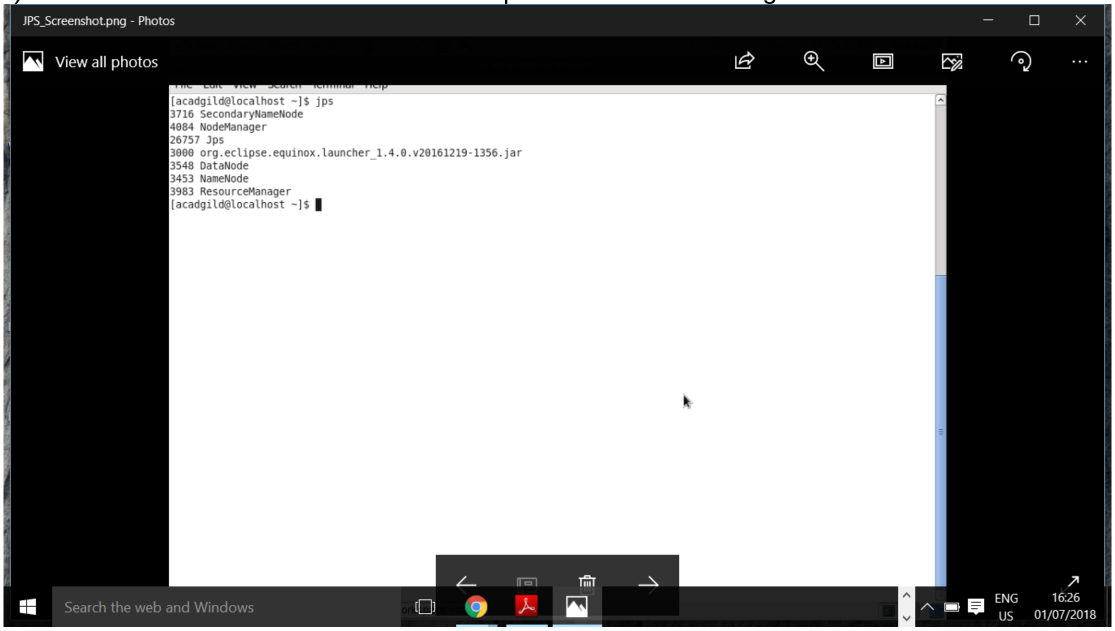
TASK 2

1) Start Hadoop single node on AcadGild VM. The command is start-all.sh.


start_Screenshot.pdf - Adobe Reader
File Edit View Window Help
70,2%
Tools Sign Comment
Applications Places System
Mon Jun 25, 5:30 AM Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
acadgild@localhost:~
eclipse-workspace - m...
Search the web and Windows
ENG US 16:23 01/07/2018

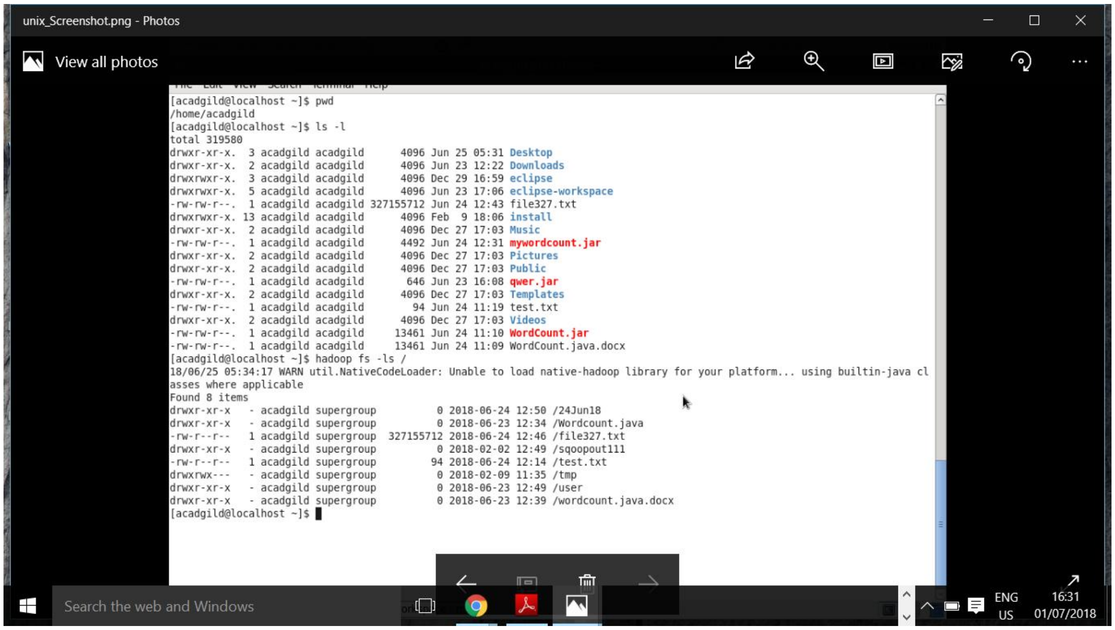
```
[acadgild@localhost ~]$ jps
2949 Jps
[acadgild@localhost -]$ start-all-sh
bash: start-all-sh: command not found
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost -]$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
18/06/25 05:29:49 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: namenode running as process 3453. Stop it first.
localhost: datanode running as process 3548. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 3716. Stop it first.
18/06/25 05:30:37 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
starting yarn daemons
resourcemanager running as process 3983. Stop it first.
localhost: nodemanager running as process 4084. Stop it first.
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ 
```

2) Run a JPS command to see if all Hadoop daemons are running.



```
[acadgild@localhost ~]$ jps
3716 SecondaryNameNode
4084 NodeManager
26757 Jps
3000 org.eclipse.equinox.launcher_1.4.0.v20161219-1356.jar
3548 DataNode
3453 NameNode
3983 ResourceManager
[acadgild@localhost ~]$ 
```

3) Run few Unix commands like pwd, ls -ls, etc.

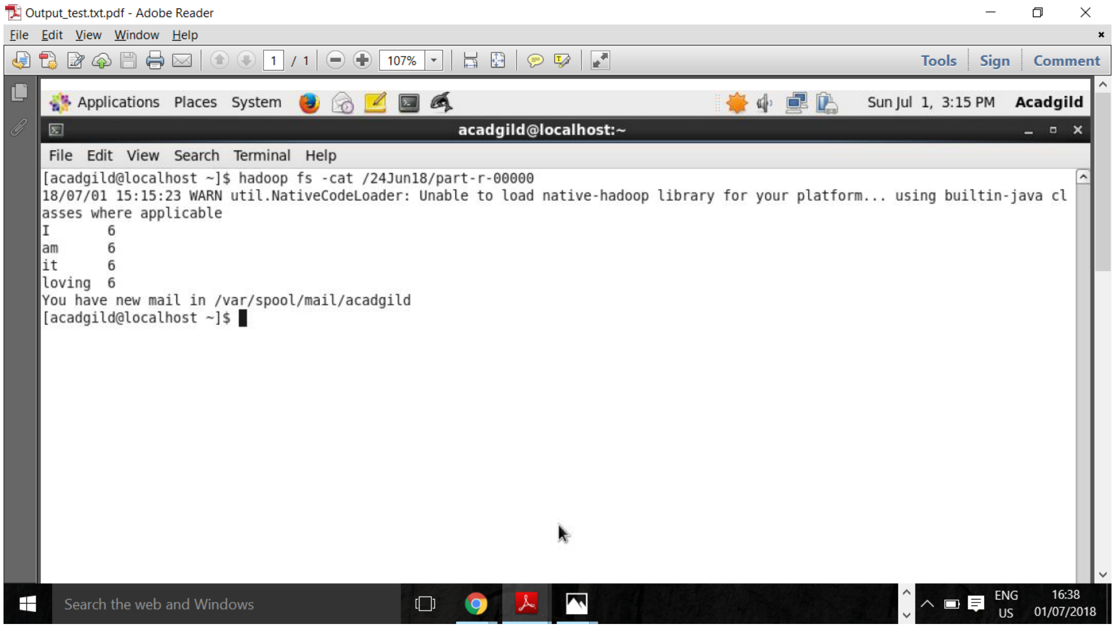


```
[acadgild@localhost ~]$ pwd
/home/acadgild
[acadgild@localhost ~]$ ls -l
total 319580
drwxr-xr-x.  3 acadgild acadgild      4096 Jun 25 05:31 Desktop
drwxr-xr-x.  2 acadgild acadgild      4096 Jun 23 12:22 Downloads
drwxrwxr-x.  3 acadgild acadgild      4096 Dec 29 16:59 eclipse
drwxrwxr-x.  5 acadgild acadgild      4096 Jun 23 17:06 eclipse-workspace
-rw-rw-r--.  1 acadgild acadgild 327155712 Jun 24 12:43 file327.txt
drwxrwxr-x. 13 acadgild acadgild      4096 Feb  9 18:06 install
drwxr-xr-x.  2 acadgild acadgild      4096 Dec 27 17:03 Music
-rw-rw-r--.  1 acadgild acadgild      4492 Jun 24 12:31 mywordcount.jar
drwxr-xr-x.  2 acadgild acadgild      4096 Dec 27 17:03 Pictures
drwxr-xr-x.  2 acadgild acadgild      4096 Dec 27 17:03 Public
-rw-rw-r--.  1 acadgild acadgild       646 Jun 23 16:08 qwer.jar
drwxr-xr-x.  2 acadgild acadgild      4096 Dec 27 17:03 Templates
-rw-rw-r--.  1 acadgild acadgild        94 Jun 24 11:19 test.txt
drwxr-xr-x.  2 acadgild acadgild      4096 Dec 27 17:03 Videos
-rw-rw-r--.  1 acadgild acadgild     13461 Jun 24 11:10 WordCount.jar
-rw-rw-r--.  1 acadgild acadgild     13461 Jun 24 11:09 WordCount.java.docx
[acadgild@localhost ~]$ hadoop fs -ls /
18/06/25 05:34:17 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 8 items
drwxr-xr-x   - acadgild supergroup          0 2018-06-24 12:50 /24Jun18
drwxr-xr-x   - acadgild supergroup          0 2018-06-23 12:34 /Wordcount.java
-rw-r--r--   1 acadgild supergroup  327155712 2018-06-24 12:46 /file327.txt
drwxr-xr-x   - acadgild supergroup          0 2018-02-02 12:49 /sqoopout111
-rw-r--r--   1 acadgild supergroup         94 2018-06-24 12:14 /test.txt
drwxrwx---   - acadgild supergroup          0 2018-02-09 11:35 /tmp
drwxr-xr-x   - acadgild supergroup          0 2018-06-23 12:49 /user
drwxr-xr-x   - acadgild supergroup          0 2018-06-23 12:39 /wordcount.java.docx
[acadgild@localhost ~]$
```

Explanation of Commands

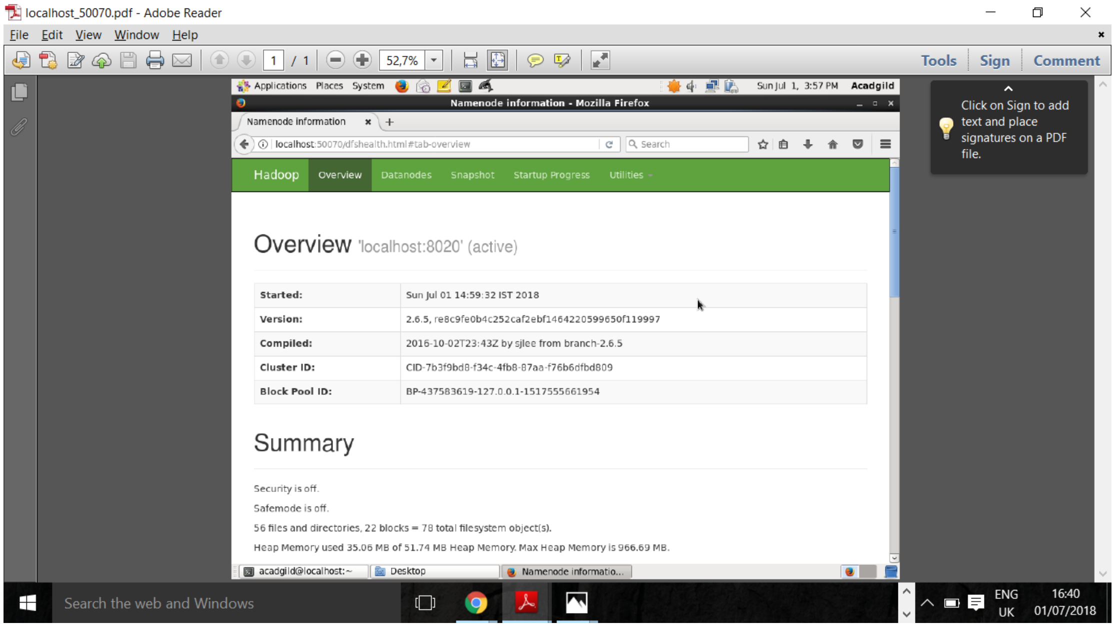
- hadoop fs -ls / - Another useful command to display the list of files and directory in HDFS File path
- hadoop fs – mkdir /direcory_name – creates the directory in HDFS
- hadoop fs -touchz /directory-name/file_name – creates the file HDFS with a size of 0 bytes
- hadoop fs –cat /directory-name/file_name – displays file content to stdout
- hadoop fs –count /directory-name/ – displays the count of directories, files, bytes for path provided
- hadoop fs –put XYZ.txt /directory-name/XYZ.txt – Copies the local file to HDFS on the specified path
- hadoop fs –get /directory-name/XYZ.txt ./XYZ.txt – reverse of put
- hadoop fs -rm -r to remove the files and hadoop fs –rm to remove directories

4) Create a file from the terminal using nano editor (example: nano test.txt), and add some content in it. Cat it to see if the content is saved.


Output_test.txt.pdf - Adobe Reader
File Edit View Window Help
1 / 1
107%
Tools Sign Comment
Applications Places System
Sun Jul 1, 3:15 PM Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
Search the web and Windows
ENG
US
16:38
01/07/2018

```
[acadgild@localhost ~]$ hadoop fs -cat /24Jun18/part-r-00000
18/07/01 15:15:23 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
I       6
am      6
it      6
loving  6
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

5)Open the hdfs web page by typing localhost:50070 in the browser. Check all the details of the HDFS.



localhost_50070.pdf - Adobe Reader
File Edit View Window Help
1 / 1
52,7%
Tools Sign Comment
Click on Sign to add text and place signatures on a PDF file.
Applications Places System
Sun Jul 1, 3:57 PM Acadgild
Namenode information - Mozilla Firefox
Namenode information
localhost:50070/dfshealth.html#tab-overview
Search
Hadoop
Overview
Datanodes
Snapshot
Startup Progress
Utilities
acadgild@localhost:~
Desktop
Namenode informatio...
Search the web and Windows
ENG UK
16:40 01/07/2018

Overview 'localhost:8020' (active)

Started:	Sun Jul 01 14:59:32 IST 2018
Version:	2.6.5, re8c9fe0b4c252caf2ebf1464220599650f119997
Compiled:	2016-10-02T23:43Z by sjlee from branch-2.6.5
Cluster ID:	CID-7b3f9bd8-f34c-4fb8-87aa-f76b6dfbd809
Block Pool ID:	BP-437583619-127.0.0.1-1517555661954

Summary

Security is off.

Safemode is off.

56 files and directories, 22 blocks = 78 total filesystem object(s).

Heap Memory used 35.06 MB of 51.74 MB Heap Memory. Max Heap Memory is 966.69 MB.