

CSEE5590 Big Data Programming

In Class Programming –1 Report
(Jongkook Son)

Project Overview:

Installation of Cloudera and visualize data with Hue

Requirements/Task(s):

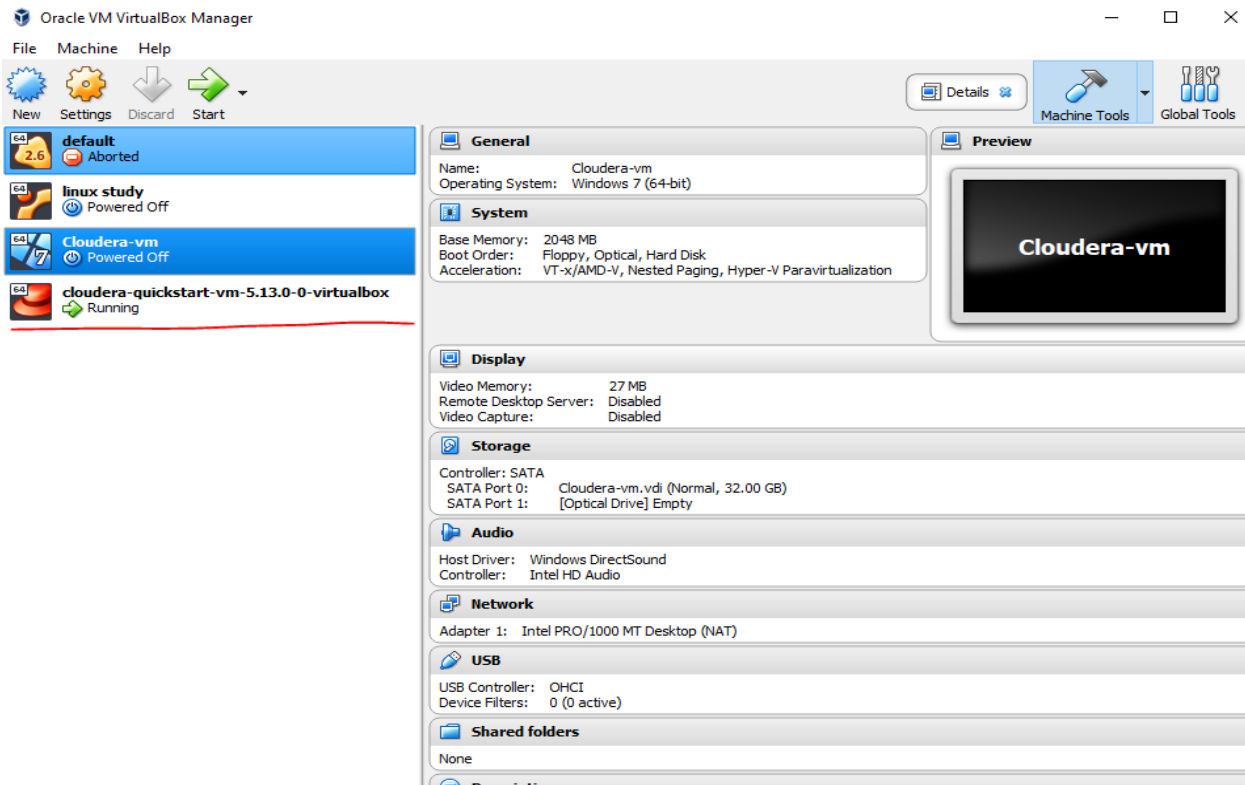
You are required to follow the steps below to complete your ICP today

- Use the given dataset
- Load it in hadoop hdfs
- Use the second file
- Append it to the first file
- Visualize file with Hue
- View the first and last lines (approximately 5) of merged dataset using appropriate hdfs commands
- Create a new text file and load it into hdfs and try to append all three datasets.

What I learned in ICP:

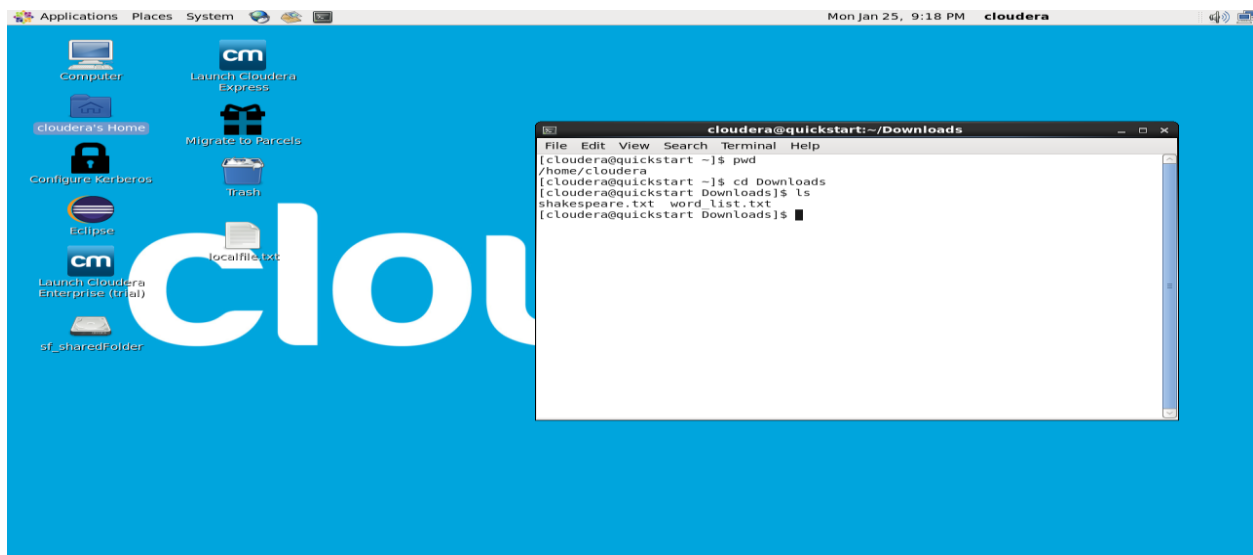
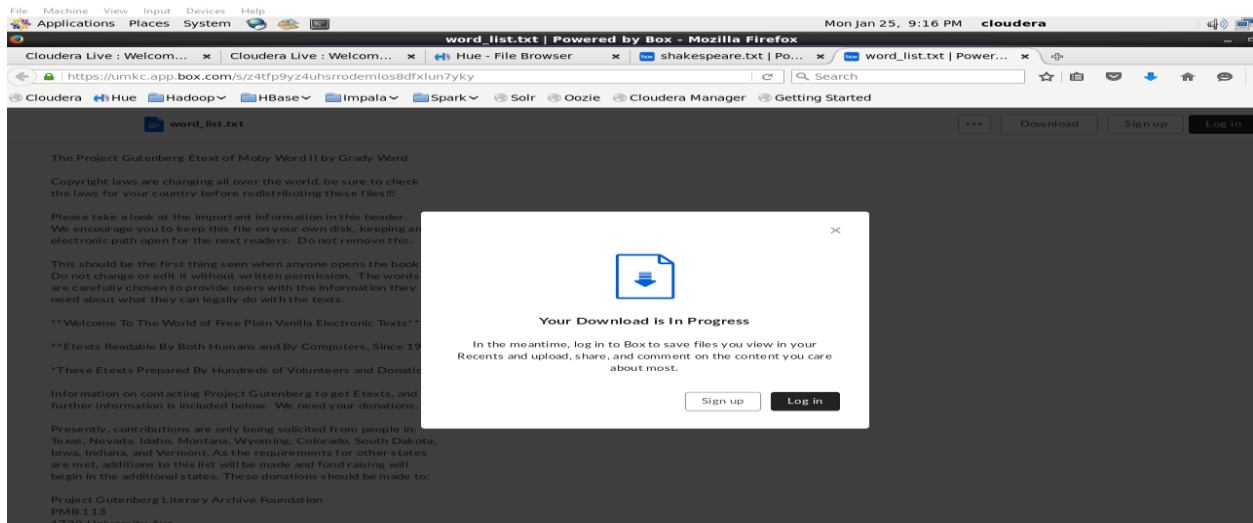
How to install cloudera using oracle virtual box. How to load data from local file system to the hdfs using cloudera virtual machine. Also I learned how to visualize hdfs file by using Hue. Finally I learned some basic hdfs commands to control files in the hdfs.

ICP description what was the task you were performing and Screen shots that shows the successful execution of each required step of your code



1. I download and install oracle virtual box download and install the cloudera virtualbox Extract cloudera file to a directory. In VirtualBox, Go to file and select import Appliance to select the VM image from my hard drive.

Select the virtual machine to start loading the VM image in the virtual box.
(cloudera-quickstart-vm-5.13.0-0-virtualbox)



2. Downloaded given datasets in cloudera virtual machine. One can find out that there are two given datasets in local virtual machine download folder

3. File Manipulation with command line

- Move to Downloads directory

```
[cloudera@quickstart ~]$ cd Downloads
```

```
[cloudera@quickstart Downloads]$ ls
```

```
shakespeare.txt word_list.txt
```

- Putting a file (Shakespeare.txt) from local system to hdfs

```
[cloudera@quickstart Downloads]$ hdfs dfs -put shakespeare.txt /user/hadoop
```

	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	↑		hdfs	supergroup	drwxr-xr-x	January 25, 2021 09:28 PM
<input type="checkbox"/>	.		cloudera	supergroup	drwxr-xr-x	January 25, 2021 09:37 PM
<input type="checkbox"/>	hadoopfile	0 bytes	cloudera	supergroup	-rw-r--r--	January 25, 2021 08:08 PM
<input type="checkbox"/>	shakespeare.txt	5.3 MB	cloudera	supergroup	-rw-r--r--	January 25, 2021 09:37 PM

Show 45 of 2 items Page 1 of 1

- Append a second file in local system to first file in the hdfs

```
[cloudera@quickstart Downloads]$ hdfs dfs -appendToFile word_list.txt /user/hadoop/shakespeare.txt
```

Mon Jan 25, 9:45 PM cloudera

Hue - File Browser - Mozilla Firefox

10.0.2.15:8888/hue/filebrowser/view=/user/hadoop/shakespeare.txt

Search data and saved documents...

File Browser

View as binary
Download
View file location
Refresh

Last modified: 01/26/2021 5:43 AM
User: cloudera
Group: supergroup
Size: 5.35 MB
Mode: 100644

Home / user / hadoop / shakespeare.txt

The Project Gutenberg EBook of The Complete Works of William Shakespeare, by William Shakespeare

This eBook is for the use of anyone anywhere at no cost and with almost no restrictions whatsoever. You may copy it, give it away or re-use it under the terms of the Project Gutenberg License included with this eBook or online at www.gutenberg.org

** This is a COPYRIGHTED Project Gutenberg eBook, Details Below **
** Please follow the copyright guidelines in this file. **

Title: The Complete Works of William Shakespeare
Author: William Shakespeare
Posting Date: September 1, 2011 [EBook #100]
Release Date: January, 1994
Language: English

*** START OF THIS PROJECT GUTENBERG EBOOK COMPLETE WORKS--WILLIAM SHAKESPEARE ***

Produced by World Library, Inc., from their Library of the Future

- View the first 5 line of merged dataset

```
[cloudera@quickstart Downloads]$ hdfs dfs -cat /user/hadoop/shakespeare.txt | head -n 5
```

The Project Gutenberg EBook of The Complete Works of William Shakespeare, by William Shakespeare

This eBook is for the use of anyone anywhere at no cost and with almost no restrictions whatsoever. You may copy it, give it away or cat: Unable to write to output stream.

- View the last 5 line of merged dataset

```
[cloudera@quickstart Downloads]$ hdfs dfs -cat /user/hadoop/shakespeare.txt | tail -n 5
```

using any compatible zip file extraction utility.

- 4) Delete the original zip file from your disk to save space. (optional)

End of this Project Gutenberg etext of Moby Word II by Grady Ward.

- Store the first 5 line of merged dataset

```
[cloudera@quickstart Downloads]$ hdfs dfs -cat /user/hadoop/shakespeare.txt | head -n 5 |hadoop fs -put - /user/hadoop/head.txt
```

- Store the last 5 line of merged dataset

```
[cloudera@quickstart Downloads]$ hdfs dfs -cat /user/hadoop/shakespeare.txt | tail -n 5 |hadoop fs -put - /user/hadoop/tail.txt
```

File Browser

Actions Move to trash Upload New

Home / user / hadoop Trash







<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	.		hdfs	supergroup	drwxr-xr-x	January 25, 2021 09:28 PM
<input type="checkbox"/>	..		cloudera	supergroup	drwxr-xr-x	January 27, 2021 06:52 PM
<input type="checkbox"/>	head.txt	238 bytes	cloudera	supergroup	-rw-r--r--	January 27, 2021 06:47 PM
<input type="checkbox"/>	shakespeare.txt	5.3 MB	cloudera	supergroup	-rw-r--r--	January 25, 2021 09:43 PM
<input type="checkbox"/>	tail.txt	201 bytes	cloudera	supergroup	-rw-r--r--	January 27, 2021 06:52 PM

Show of 5 items Page 1 of 1 Navigation icons

- Append all three datasets(shakespeare, word_list, first last 5 sentence) to the new text file

```
[cloudera@quickstart Downloads]$ hdfs dfs -text /user/hadoop/*.txt | hdfs dfs -put - /user/hadoop/Merged.txt
```

Home / user / hadoop Trash

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	 .		hdfs	supergroup	drwxr-xr-x	January 25, 2021 09:28 PM
<input type="checkbox"/>	 .		cloudera	supergroup	drwxr-xr-x	January 27, 2021 06:57 PM
<input type="checkbox"/>	 Merged.txt	5.3 MB	cloudera	supergroup	-rw-r--r--	January 27, 2021 06:57 PM
<input type="checkbox"/>	 head.txt	238 bytes	cloudera	supergroup	-rw-r--r--	January 27, 2021 06:47 PM
<input type="checkbox"/>	 shakespeare.txt	5.3 MB	cloudera	supergroup	-rw-r--r--	January 25, 2021 09:43 PM
<input type="checkbox"/>	 tail.txt	201 bytes	cloudera	supergroup	-rw-r--r--	January 27, 2021 06:52 PM

Show 45 of 4 items Page 1 of 1

HUE Query Jobs

Search data and saved documents...

File Browser

View as binary

Download

View file location

Refresh

Last modified: 01/28/2021 2:57 AM

User: cloudera

Group: supergroup

Size: 5.35 MB

Mode: 100644

Home / user / hadoop / **Merged.txt** Page 1 to 50 of 1369

The Project Gutenberg eBook of The Complete Works of William Shakespeare, by William Shakespeare

This eBook is for the use of anyone anywhere at no cost and with almost no restrictions whatsoever. You may copy it, give it away or re-use it under the terms of the Project Gutenberg license included with this eBook or online at www.gutenberg.org

** This is a COPYRIGHTED Project Gutenberg eBook, Details Below **
 ** Please follow the copyright guidelines in this file. **

Title: The Complete Works of William Shakespeare

Author: William Shakespeare

Posting Date: September 1, 2011 [Ebook #100]

Release Date: January, 1994

Language: English

*** START OF THIS PROJECT GUTENBERG EBOOK COMPLETE WORKS--WILLIAM SHAKESPEARE ***