Assignment 1

Problem I (10 points)

Please list all characteristics of big data and use one or two sentences to explain their meanings.

Big data have five characteristics, which are **Variety, Volume, Veracity, Velocity and Value.**

Variety: There are various types of data we are using right now, including structured and unstructured. Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis applications.

Volume: It related to the data size which is enormous.

Veracity: Veracity is the quality or trustworthiness of the data. Just how accurate is about the data? For example, thinking about all the news post in Facebook with hash tags, abbreviations, typos, etc., and the reliability and accuracy of all that content.

Velocity: The flow of data is massive and continuous. This real-time data can help researchers and businesses make valuable decisions that provide strategic competitive advantages.

Value: In contemporary society, companies are continuously generating values by using big data in various industry. Nowadays, Improved customer service, better operational efficiency, Better Decision Making are few advantages of big data.

Problem II (20 points)

Please describe two scenarios where you can apply big data analytical techniques. For each application, write down the possible dataset, potential methods/algorithms, and the outcome/goal.

1. An online email service analyzes messages and user behavior to optimize ad selection and placement.

dataset : E-mails content that written by every users around the world. It is impossible that store the data in a single data warehouse, because of the enormous data size and the physical data distribution.

Potential Methods: Text mining approaches, by translating the text file into numeric vectors. We can easily implement those data mining algorithms on the vector to extract the image of the user by the numeric data. For the classification, those algorithms could be Bayesian algorithms, SVM, KNN or Random Forest. On the other hand, Clustering can be done by K-mean.

Output: the type of the users and for different class of users service providers can deliver the most suitable ads.

2.Videos - Recommendation
dataset :the sequence data of the watching history for every user, the basic information of the users and the records that who have watched the video as well as the description or categories of the video.

Potential Methods:

1.By analyzing on the users, finding the similarity behind the users. If a personal have the related background info and they have watched a collection of same type of videos. We can use the KNN or Decision tree algorithms to do so :

1.Find the K-nearest neighbors (KNN) to the user $\boldsymbol{a}$, using a similarity function $\boldsymbol{w}$ to measure the distance between each pair of users.

2.Predict the rating that user $\boldsymbol{a}$ will give to all items the $\boldsymbol{k}$ neighbors have consumed but $\boldsymbol{a}$ has not. We Look for the item $\boldsymbol{j}$ with the best predicted rating.

2.By analyzing the video itself, like rating, type, duration and description about the video .

1. Using the data, we could using clustering methods to group them into new categories.

2. Once a user watches a video $\boldsymbol{a}$, we could provide the video $\boldsymbol{j}$ in the same predicted group.

Output: A collection of video clips that the users tend to continue watch.

Problem III (10 points)

Please briefly illustrate the architecture of the Hadoop ecosystem, including layers, and components and their roles.

1. Layers : there are three layers in the Hadoop ecosystem, HDFS- Hadoop distributed files system, Mapreduce and Yarn.
2. Components:

HDFS: It is the most important component of Hadoop Ecosystem. **HDFS** is the primary storage system of Hadoop, which is consisted of two parts :

a. Name Node: It is also known as Master node. NameNode does not store **actual data or dataset**. NameNode stores Metadata i.e. number of blocks, their location, on which Rack, which Data node that the data is stored and other details. It consists of files and directories.

Role:

1. Manage file system namespace.
2. Regulates client's access to files.
3. Executes file system execution such as naming, closing, opening files and directories.

b. Data Node: It is also known as Slave. HDFS Data node is responsible for storing actual data in HDFS. Data node performs read and write operation as per the request of the clients.

Role:

1. DataNode performs operations like block replica creation, deletion, and replication according to the instruction of NameNode.

2. DataNode manages data storage of the system.

Mapreduce: Providing data processing, a software framework for easily writing applications that process the vast amount of structured and unstructured data stored in the HDFS.And improving the speed and reliability of cluster this parallel processing.

Role:

1. Map function :Taking a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs)

2. Reduce function: Taking the output from the Map as an input and combines those data tuples based on the key and accordingly modifies the value of the key.

YARN: Providing the resource management as the operating system of Hadoop. And it is responsible for managing and monitoring workloads

Role:

1. Enables other purpose-built data processing models beyond MapReduce (batch), such as interactive and streaming.

2.Runing as many applications on the same cluster, Hence, efficiency of Hadoop increases without much effect on quality of service.

3. Provides a stable, reliable, secure foundation and shared operational services across multiple workloads. Additional programming models such as graph processing and iterative modeling are now possible for data processing.

Problem IV (60 points)

For Hadoop (pseudo-distributed) installation and configuration in your first lab, please show all commands you typed, the content of files you modified (after editing), and the screenshots you obtained for all following steps.

1) Set up JAVA_HOME environment variable (5 points)

HADOOP_HOME	C:\Hadoop-2.8.0\bin	C:\Hadoop-2.8.0\bin
JAVA_HOME	C:\jdk\bin	C:\jdk\bin

2) Configuring Hadoop (modifying 4 files: hadoop-env.sh, hdfs-site.xml, core-site.xml, mapred-site.xml) (20 points)

mapred-site.xml.template | hdfs-site.xml | yarn-site.xml | hadoop-env.cmd

```
22 @rem remote nodes.
23
24 @rem The java implementation to use.  Required.
25 set JAVA_HOME=C:\jdk
```

mapred-site.xml.template | hdfs-site.xml | yarn-site.xml | hadoop-env.cmd

```
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
   <property>
       <name>dfs.replication</name>
       <value>1</value>
   </property>
   <property>
       <name>dfs.namenode.name.dir</name>
       <value>C:\hadoop-2.8.0\data\namenode</value>
   </property>
   <property>
       <name>dfs.datanode.data.dir</name>
       <value>C:\hadoop-2.8.0\data\datanode</value>
   </property>

</configuration>
```

mapred-site.xml.template | hdfs-site.xml | yarn-site.xml | hadoop-env.cmd | core-site.xml

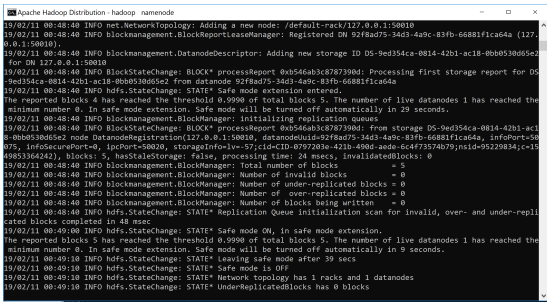
```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
   <property>
       <name>fs.defaultFS</name>
       <value>hdfs://localhost:9000</value>
   </property>

</configuration>
```

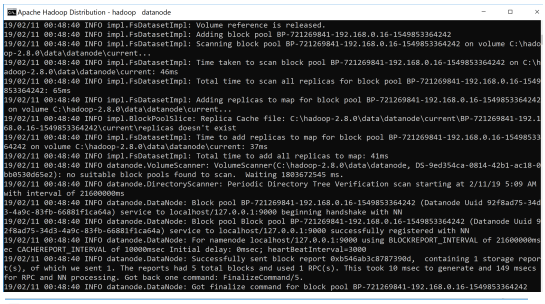
hdfs-site.xml | yarn-site.xml | hadoop-env.cmd | core-site.xml | mapred-site.xml

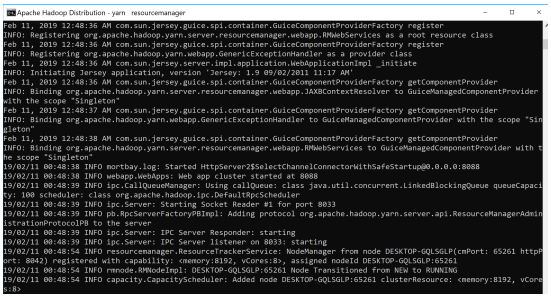
```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
    <property>
        <name>mapreduce.framework.name</name>
        <value>yarn</value>
    </property>

</configuration>
```

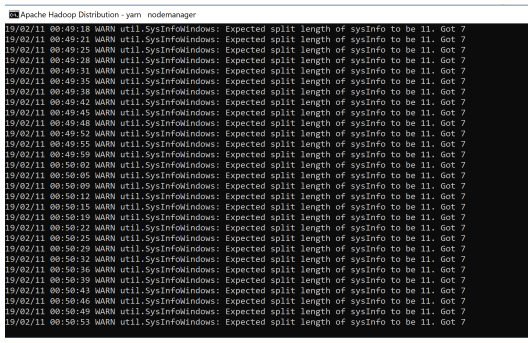
3) Hadoop startup (10 points)

```
C:\hadoop-2.8.0\sbin>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons
```

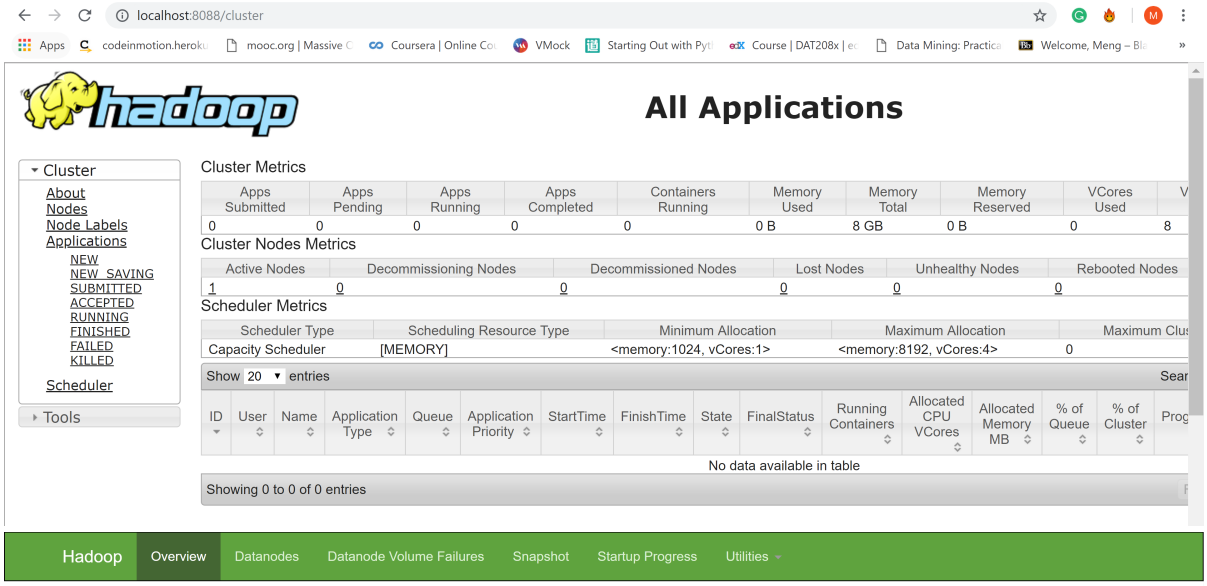








4) Hadoop monitoring using web interface (5 points)



Overview 'localhost:9000' (active)

Started:	Mon Feb 11 00:48:31 -0500 2019
Version:	2.8.0, r91f2b7a13d1e97be65db92ddabc627cc29ac0009
Compiled:	Fri Mar 17 00:12:00 -0400 2017 by jdu from branch-2.8.0
Cluster ID:	CID-0797203e-421b-490d-aede-6c4f73574b79
Block Pool ID:	BP-721269841-192.168.0.16-1549853364242

Summary

5) Running Hadoop job

a. Create a folder under Hadoop to store your data (5 points)

drwxr-xr-x	xming	supergroup	0 B	Feb 11 00:56	0	0 B	hw1file

```
C:\hadoop-2.8.0\bin>hdfs dfs -mkdir -p /hw1file/
```

b. Copy files in your local machine to remote HDFS (5 points)

```
C:\hadoop-2.8.0\bin>hdfs dfs -put C:\Users\ xming\Desktop\BigData\bigdataNotes.txt /hw1file/
```

-rw-r--r--	xming	supergroup	3.61 KB	Feb 11 01:12	1	128 MB	bigdataNotes.txt

c. Run your Hadoop job (5 points)

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	xming	supergroup	0 B	Feb 11 01:21	0	0 B	hw1result

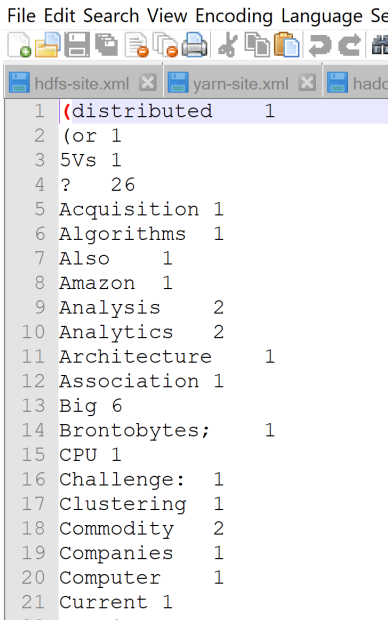
d. View your output on your local machine (5 points)

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	xming	supergroup	0 B	Feb 11 01:21	1	128 MB	_SUCCESS
-rw-r--r--	xming	supergroup	3.2 KB	Feb 11 01:21	1	128 MB	part-r-00000

```
C:\hadoop-2.8.0>hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.8.0.jar wordcount /hw1file/ /hw1file/output/hw1result
19/02/11 01:20:57 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/02/11 01:20:58 INFO input.FileInputFormat: Total input files to process : 1
19/02/11 01:20:58 INFO mapreduce.JobSubmitter: number of splits:1
19/02/11 01:20:58 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1549864109658_0001
19/02/11 01:20:59 INFO impl.YarnClientImpl: Submitted application application_1549864109658_0001
19/02/11 01:20:59 INFO mapreduce.Job: The url to track the job: http://DESKTOP-GQLSGLP:8088/proxy/application_1549864109658_0001/
19/02/11 01:20:59 INFO mapreduce.Job: Running job: job_1549864109658_0001
19/02/11 01:21:21 INFO mapreduce.Job: Job job_1549864109658_0001 running in uber mode : false
19/02/11 01:21:21 INFO mapreduce.Job:  map 0% reduce 0%
19/02/11 01:21:32 INFO mapreduce.Job:  map 100% reduce 0%
19/02/11 01:21:44 INFO mapreduce.Job:  map 100% reduce 100%
19/02/11 01:21:44 INFO mapreduce.Job: Job job_1549864109658_0001 completed successfully
19/02/11 01:21:44 INFO mapreduce.Job: Counters: 49
        File System Counters
```

```
C:\hadoop-2.8.0>hdfs dfs -cat /hw1file/output/hw1result/part-r-00000
(distributed    1
(or     1
5Vs     1
?       26
Acquisition     1
Algorithms      1
Also    1
Amazon  1
Analysis        2
Analytics       2
Architecture    1
Association     1
Big     6
Brontobytes;    1
```

```
C:\hadoop-2.8.0>hadoop fs -get /hw1file/output/hw1result/part-r-00000 C:\Users\xming\Desktop\BigData
```



Submission Notes:

- Please submit your file in a word or pdf format through BlackBoard. DO NOT email your files to the instructor. The filename must be like: *YourLastName_YourFirstName_Ass1.doc* (or *.pdf*). The file submitted without following this rule will NOT be graded.