Lanzhou University

School of Information Sciences and Engineering

Cloud Computing and Big Data

Assignment 1

Module Code: 2043623

Name: Yuming Chen

Student Number: 320180939611

Set date: Thursday, March 4, 2021 Due Date: 11:59pm, Friday, March 12, 2021

A. Requirements

1. Provide full and detailed answer to each question.
2. Work individually. The Lanzhou University Academic Honesty Rules and Procedures (as stated in the student handbook) will be adhered to strictly.
3. There are 6 questions. The total marks are 100. This assignment contributes 15% to the overall assessment for this module.

B. Questions

Question 1 [15 marks]. Discuss the characteristics of Big Data. Explain the Big Data Problem faced by current enterprises. Describe how cloud computing framework attempts to solve the Big Data Problem.

1. The characteristics of Big Data are:

Characteristics	*Explanation*
Volume	Data volume refers to the tremendous scale of big data. Nowadays, data volume is increasing exponentially, because data is generated from various sources in different formats. From 2009 to 2020, data volume has increased from 0.8 zettabytes to 35zb. ***Example***: ① Walmart handles more than 1 million customer transactions every hour; ② YouTube users upload 48 hours of new video every minute of the day. ③ 294 billion emails are sent every day
Variety	Data variety refers to big data has various formats, types, and structures which makes the process of data become complexity. In other words, big data is not always structured data, it also can be semi-structured and unstructured data. In fact, almost 80 percent of data produced globally including photos, videos and social media content, is unstructured in nature. In order to extract knowledge from data, all these types of data need to linked together. ***Example***: The 230+ millions of Tweets which created every day contain data which has various formats, types, and structures data. Most of Tweets contain structured data such as personal information, which can be stored into relational database, such as twitter account, name and age. They also contain unstructured data such as emoji, video and photos.
Velocity	Data velocity refers to big data begin generated fast and need to be processed fast. In Big Data velocity data flows in from sources like machines, networks, social media, mobile phones etc. There is a massive and continuous flow of data. This determines the potential of data that how fast the data is generated

	and processed to meet the demands. ***Example***: ① E-Promotions should send promotions based on customer current location and customer purchase history to customer on time before customer finishing shopping. ② In order to achieve healthcare monitoring, any abnormal measurements require immediate reaction when sensors monitoring your activities and body. ③ There are more than 3.5 billion searches per day are made on Google.
Veracity	Data veracity refers to the big data in doubt or uncertainty of data available due to data inconsistency and incompleteness. Big Data is available can sometimes get messy and quality and accuracy are difficult to control. It is not possible that all of the big data is going to be 100% correct there will be dirty data. Nowadays, dirty data cost $600 billion to the companies every year in the United States. ***Example***: ① **Duplicate data**: It may occur due to repeated submissions, improper data joining or user error. ② **Incomplete data**: Data with missing values is the main type of incomplete data. ③ **Inconsistent data:** Due to unchecked data redundancy, the format and type of the same data are different.
Value	Data value refers to how useful the data is in decision making, which means the people can extract the value of the Big Data, such as trends, patterns and associations. Data value can be analyzed computationally and it is related to human behaviors or interests. ***Example***: Shopping websites can extract useful information from a large amount of user data, such as shopping history, browsing record, interests and hobbies. Therefore, they can understand the products that users are more interested in and provide corresponding recommendations for users, in order to increase the sales volume of products and gain value from them.

Other characteristics	
Validity	Data validity refers to the correctness of data is important.
Variability	Data variability refers to the dynamic behavior of data.
Volatility	Data volatility refers to the tendency of data is easy to change in time.
Vulnerability	Data vulnerability refers to data is vulnerable to breach or attacks.

2. The Big Data Problem faced by current enterprises:

Challenges	*Explanation*
Data Quality	The problem is related to Data Veracity. Today, data here is very messy, inconsistent and incomplete. If the accuracy of data is low, it will make data analysis cost increase greatly. In order to deal with the problem, companies should pay a lot to clean the dirty data. Nowadays, it costs many companies $600 billion dollars every year in the United States to clean the dirty data.
Data Discovery	It is imperative for business organizations to gain important insights from Big Data analytics, and also it is important that only the relevant department has access to this information. However, It is a very difficult challenge to analyze petabytes of data using extremely powerful algorithms to find patterns and insights. Companies need to mend this wide gap in an effective manner.
Data Storage	The more data an organization has, the more complex the problems of managing it can become. The storage of this massive amount of data is becoming a real challenge for every companies. Companies have to develop a storage system which can easily scale up or down on-demand.
Data Analytics	The variety of Big Data which is extremely complex makes companies difficult to analyze the data.
Data Security	Once business enterprises discover how to use Big Data, it brings them a wide range of possibilities and opportunities. However, it also involves the potential risks associated with big data when it comes to the privacy and the security of the data. The Big Data tools used for analysis and storage utilizes the data disparate sources. This eventually leads to a high risk of exposure of the data, making it vulnerable. Thus, keeping it secure is an important and enormous challenge for companies. It includes user authentication, restricting access based on a user, recording data access histories, proper use of data encryption etc.
Lack of Talent/Technical skills	The analysis of data is important to make this voluminous amount of data being produced in every minute, useful. With the exponential rise of data, a huge demand for big data scientists and Big Data analysts has been created in the market. It is important for business organizations to hire a data scientist having skills that are varied as the job of a data scientist is multidisciplinary. Nowadays, most of companies faced the challenge that the shortage of professionals who understand Big Data analysis. There is a sharp shortage of data scientists in comparison to the massive amount of data being produced.

Insufficient budget	Developing Big Data technology is expensive for companies. It not only cost a lot in software, but also the hardware, such as GPU and cloud servers.

3. Solutions generated by cloud computing framework are:

Challenges	*Solution*
Data Quality	In a cloud-based system, all documents are stored in one place and in a single format. With everyone accessing the same information, companies can maintain consistency in data, avoid human error, and have a clear record of any revisions or updates. Conversely, managing information in silos can lead to employees accidentally saving different versions of documents, which leads to confusion and diluted data.
Data Storage	Cloud computing is based on the distribution file system which offers almost limitless storage capacity. At any time, companies can quickly expand their storage capacity with very nominal monthly fees.
Data Analytics	Cloud computing offer integrated cloud analytics for a bird's-eye view of data. With the information stored in the cloud, companies can easily implement tracking mechanisms and build customized reports to analyze information organization wide.
Data Security	Cloud computing providers implement baseline protections for their platforms and the data they process, such as authentication, access control, and encryption. In this protection, it can actually be much safer to keep sensitive information offsite. RapidScale claims that 94% of businesses saw an improvement in security after switching to the cloud, and 91% said the cloud makes it easier to meet government compliance requirements. Moreover, cloud computing provide disaster recovery which will keep data in a much safer zone.
Lack of Talent/Technical skills	Cloud computing provide companies an easy way to deal with big data. The cloud computing platform will serve as a talent data scientist to solve companies' problem. Moreover, cloud computing makes collaboration a simple process and reduce the cost of collaboration. Team members can view and share information easily and securely across a cloud-based platform.
Insufficient budget	Cost saving is one of the biggest Cloud Computing benefits. By using a cloud infrastructure, today's companies do not have to spend a lot of money to buy and maintain expensive physical hardware investments for cloud computing. Also, you do not need trained personnel to maintain the hardware. Cloud computing has significantly reduced capital expenditure costs.

Question 2 [15 marks]. Discuss and explain the main characteristics of cloud computing in terms of data partition, data replication, parallel computing, cluster scalability, and fault tolerance.

1. Definition

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

Essential characteristics	*Explanation*
On-demand self-service	A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.
Broad network access	Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms.
Resource pooling	The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction
Rapid elasticity	Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.
Measured service	Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service. Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

2. data partition

Data partition refers to divide data into serval pieces regularly and move each piece into different rank according the rules are set earlier. Data partition is a basic operation of distributed file system

which makes it easier for users to access shared files distributed across the network. Data partition reflects the resource pooling which is one of the cloud computing's characteristics. It means that the provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand.

3. data replication

Data replication is the process of making multiple copies of data and storing them at different locations to improve their overall accessibility across a network. Similar to data mirroring, data replication can be applied to both individual computers and servers. The data replicates can be stored within the same system, on-site and off-site hosts, and cloud-based hosts. In cloud computing, data replication can increase the data scalability which means the capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. This is the basic of rapid elasticity which is one of the cloud computing's characteristics. Moreover, data replication facilitates the recovery of data which is lost or corrupted by maintaining accurate backups at well-monitored locations, thereby contributing to enhanced the fault tolerance of cloud computing. Data replication also can increase the speed of data access by providing high concurrency service.

4. parallel computing

Parallel computing refers to a type of computation where many calculations or the execution of processes are carried out simultaneously. Large problems can often be divided into smaller ones, which can then be solved at the same time. Parallel computing is the source of cloud computing which can solves different demands in the same time and increase the speed of data process.

5. cluster scalability

Scalability is the property of a system to handle a growing amount of work by adding resources to the system. Cluster scalability refers to the data process unit can be increased or decreased with demand. This is the basic of rapid elasticity which is one of the cloud computing's characteristics. Moreover, the cluster scalability can reduce computing cost and enable data parallelism for cloud computing.

6. fault tolerance

Fault tolerance refers to the ability of a system (computer, network, cloud cluster, etc.) to continue operating without interruption when one or more of its components fail. The objective of creating a fault-tolerant system is to prevent disruptions arising from a single point of failure, ensuring the high availability and business continuity of mission-critical applications or systems. Fault tolerance enhances the reliability of cloud computing.

Question 3 [15 marks]. Describe and explain the following service models: IaaS, PaaS, and SaaS. Do some research on the Internet; for each type of service model, find some real-world examples and describe them.

1. IaaS

1.1 Describe and explain

Infrastructure as a service (IaaS) is a form of cloud computing that delivers fundamental compute, network, and storage resources to consumers on-demand, over the internet, and on a pay-as-you-go basis. It provides basic storage and computing capabilities as standardized services over the network. Therefore, customers would typically deploy their own software on the infrastructure.

Example

① IBM Cloud

IBM Cloud is another classic example of how top Cloud provider cover the entire spectrum. Its complete product includes a comprehensive IaaS segment as well. This covers compute elements, network resources, storage, and more. Most unique about IBM Cloud is their Bare Metal as a Service (BMaaS) offering. This allows their IaaS users to get unprecedented access to the hardware that lies beneath their Cloud service. Another notable product under their IaaS range is Cloud Object Storage.

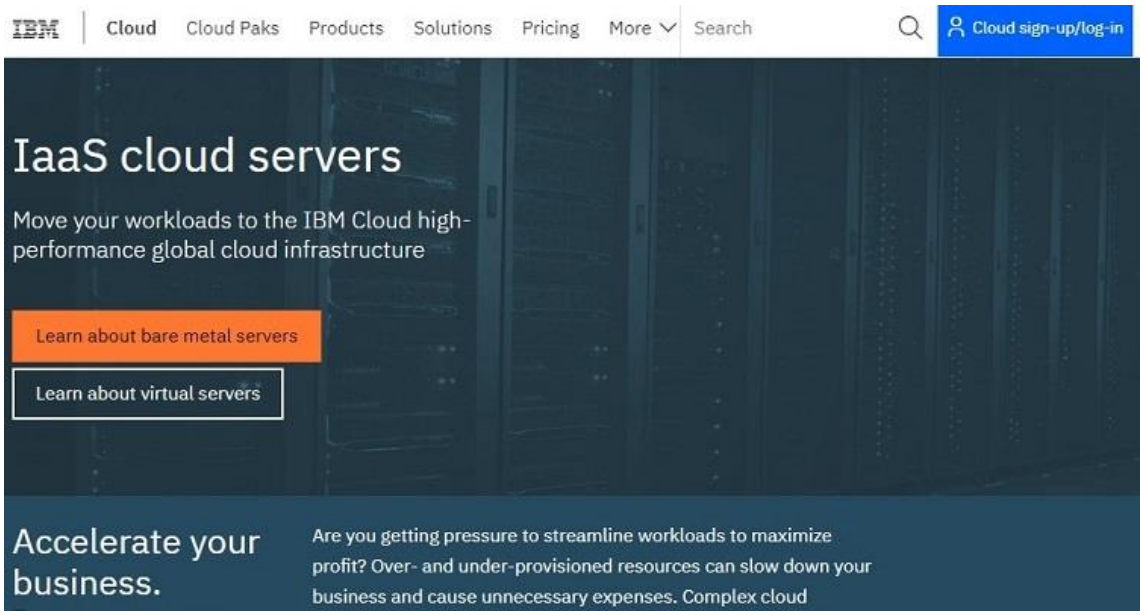


Image 1. IBM Cloud website: https://www.ibm.com/cloud

③ Digital Ocean

Many of those in the web hosting industry may be familiar with providers like Digital Ocean. Although Digital Ocean (DO) focuses on the areas of web hosting and web application deployment, it is nonetheless a good example of a niche IaaS provider. DO offers users piecemeal allocation of various infrastructure resources that they can combine to customize a Cloud to their unique

requirements and enables developers to quickly build and deploy powerful sites and applications. These can be run indefinitely or sold quickly for profit.

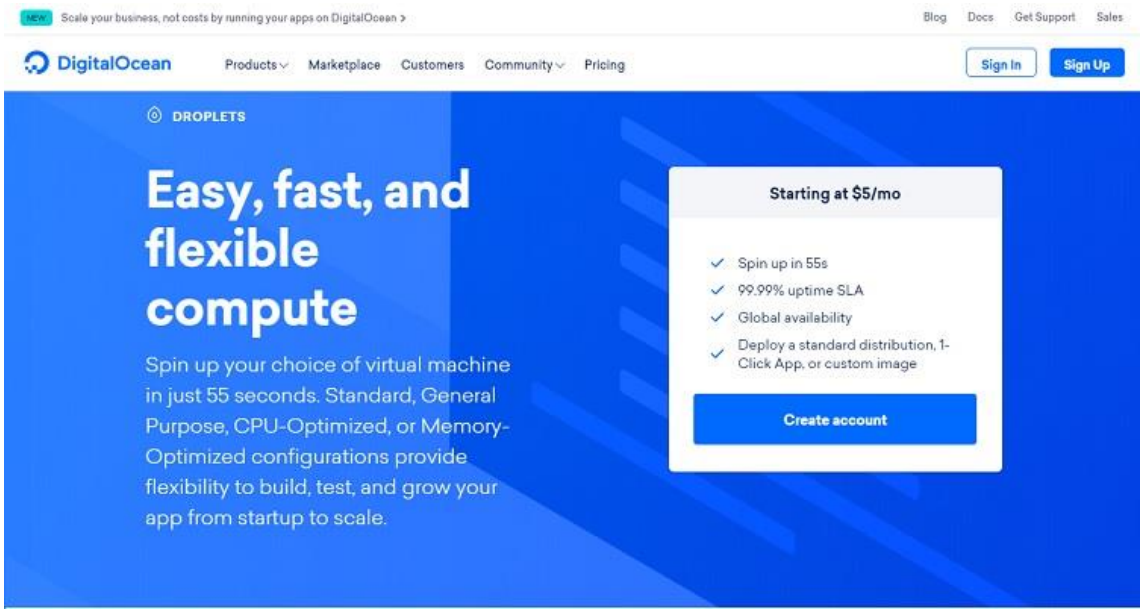


Image 2. Digital Ocean website: https://www.digitalocean.com/

2. PaaS

2.1 Describe and explain

Platform-as-a Service (PaaS) refers to a layer of software, or development environment is encapsulated and offered as a service. It has the freedom to build his own applications, which run on the provider's infrastructure. It offers a predefined combination of OS and application servers, such as LAMP platform, restricted J2EE, Ruby etc. Therefore, customer has the freedom to build his own applications, which run on the provider's infrastructure.

Example

① SAP Cloud

SAP is a really big company, so much so that its offerings span multiple service models. Among them is their Cloud PaaS which is an open business platform. It was designed to help developers build applications more easily, offering both breadth and depth of service. The platform also opens up the possibility of integrating Cloud and on-premise apps and provides many supporting services. Part of this is thanks to SAP's immense partner ecosystem which delivers a stunning library of over 1,300 apps built on the same platform.

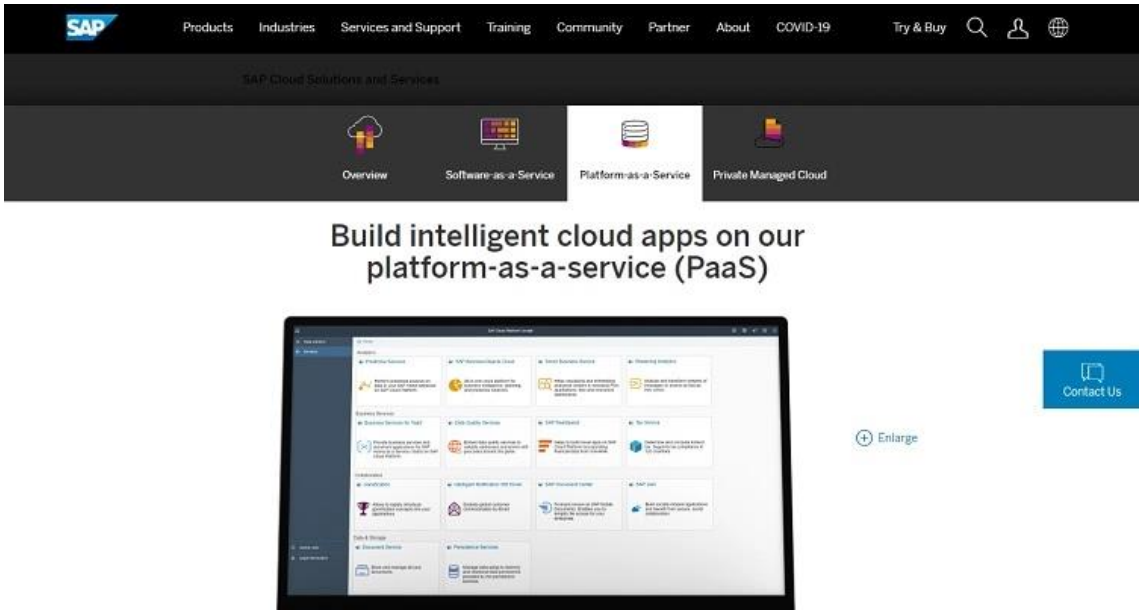


Image 3. SAP Cloud website: https://www.sap.com/index.html

② **Heroku**

Heroku now belongs to Salesforce and is an example of PaaS based on the managed container concept. As with many PaaS environments, it is highly self-contained and integrates data services as well as a complete ecosystem of its own.

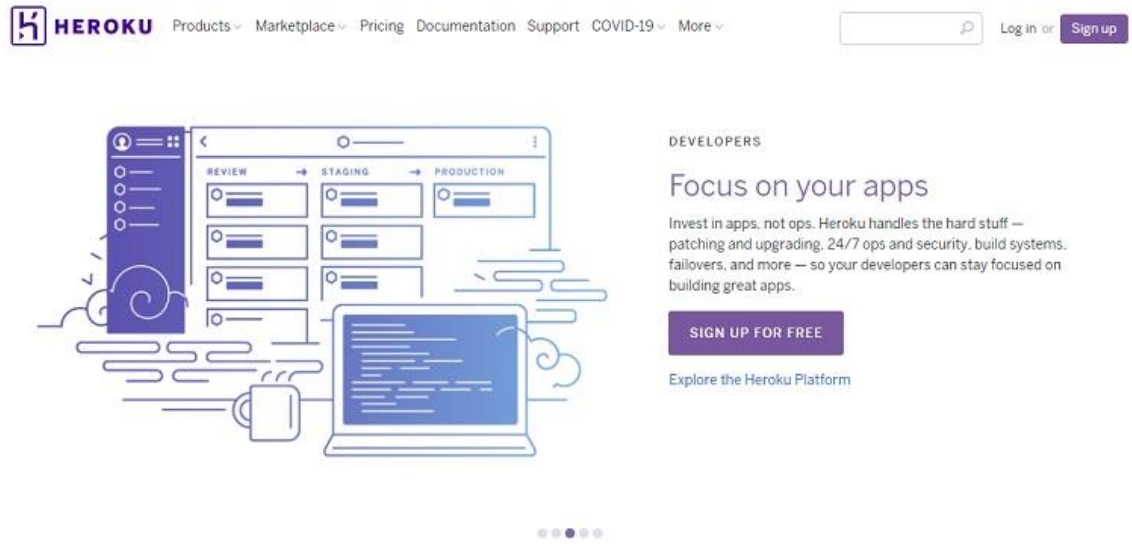


Image 4. Heroku website: https://www.sap.com/index.html

3. SaaS

3.1 Describe and explain

Software-as-a-Service (SaaS) refers to a complete application is a single instance of the service which runs on the cloud and multiple end users are serviced. The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications

are accessible from various client devices through either a thin client interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited userspecific application configuration settings.

Example

① Box

Box is an online workspace enables professionals to collaborate with anyone, anywhere. Users can securely share large files via traditional link or custom URL, safeguarding data and documents via permissions and password protection. Box also automates tasks such as employee onboarding and contract approvals, reducing repetition and abbreviating review cycles. These functions are provided by servers which run on the cloud.

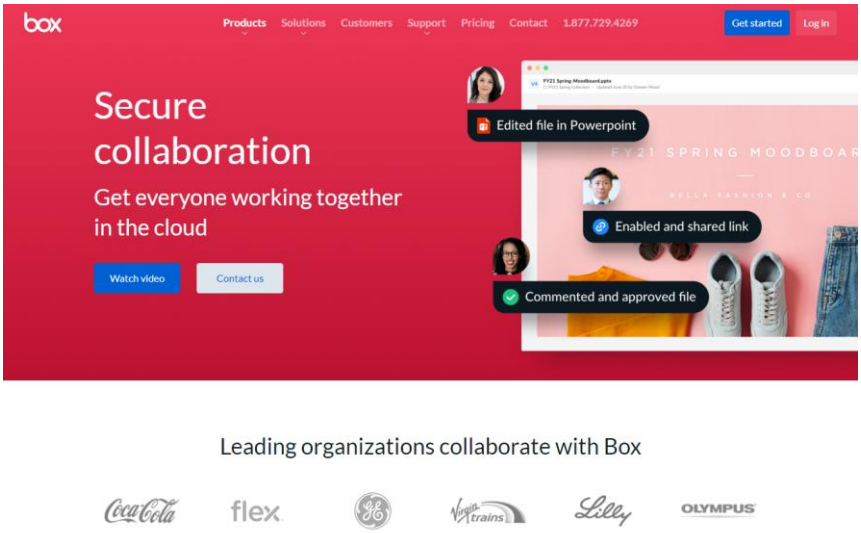


Image 5. Box website: https://www.box.com/collaboration

② Lumen5

Lumen5 is a leading video creator SaaS app that lets businesses create amazing videos with its drag-and-drop interface. It creates video automatically form text or any URL. The text is converted into a video that can be personalized by positioning text, adding images from the library, highlighting keywords, adding brand colors, tweaking font style, and changing video resolution.

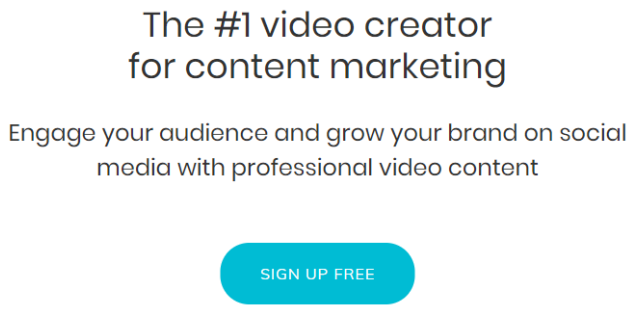


Image 6. Lumen5 website: https://lumen5.com/

Question 4 [15 marks]. Describe the core components in Hadoop architecture.

Describe the steps of Hadoop process handling a distributed computing task across a commodity cluster.

1. The core components in Hadoop architecture are

1.1 Hadoop Distributed file system (HDFS) - **storage**

HDFS is a Java based file system for providing scalable and reliable storage of large datasets. It serves as the foundation of Hadoop. Moreover, HDFS stores the Data in the form of blocks (the smallest continuous location on hard drive where data is stored) and operates it on the Master-Slave Architecture. It can split input files into blocks across nodes for parallel processing. In order to achieve fault tolerance, it will replicate data 3 times (default).

1.2 MapReduce framework - **processing**

MapReduce framework is a Java based programming paradigm of the Hadoop framework that provides scalability across various Hadoop clusters. MapReduce executes user jobs specified as "map" and "reduce" functions

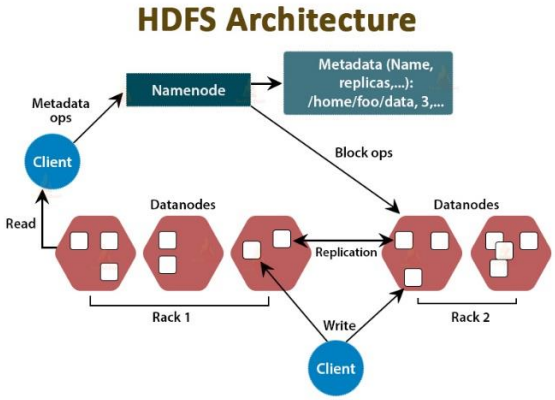
2. The steps of Hadoop process handling a distributed computing task across a commodity cluster are:

2.1 Store input data in Hadoop HDFS

① The client interacts with the NameNode firstly. NameNode will check for the client privileges, and if the client has sufficient privileges, then the NameNode will provide the address of the DataNodes from where the client can write input data.

② The input data will be divided into blocks. The block size is 128 MB by default. DataNode will store the blocks of files.

③ To provide fault-tolerance, HDFS will create replicas of blocks depending on the replication factor. By default, replication factor is 3, which means 3 copies of a block are stored in HDFS. HDFS stores replicas of the block on different DataNodes by following the Rack Awareness algorithm.



2.2 Use Hadoop MapReduce to process data

① Writes the MapReduce program for processing the Data.

② Hadoop divides the input job into tasks of two types, that is, map tasks and reduce tasks.

③ YARN scheduled these tasks which run on different DataNodes.

④ The input data to the MapReduce job is divided into fixed-size pieces which is called input splits.

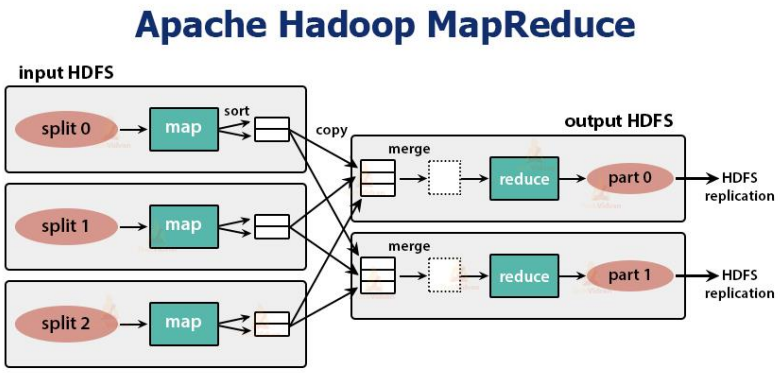
⑤ One map task which runs a user-defined map function for each record in the input split is created for each input split. These map tasks run on the DataNodes where the input data resides.

⑥ After task is finished, the output of the map task is intermediate output and is written to the local disk.

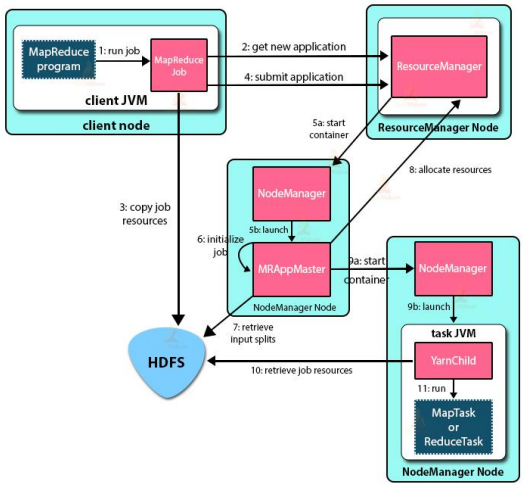
⑦ The intermediate outputs of the map tasks are shuffled and sorted and are then passed to the reducer.

⑧ For a single reduce task, the sorted intermediate output of mapper is passed to the node where the reducer task is running. These outputs are then merged and then passed to the user-defined reduce function. For multiple reduce functions, the user specifies the number of reducers. When there are multiple reduce tasks, the map tasks partition their output, creating one partition for each reduce task.

⑨ The reduce function summarizes the output of the mapper and generates the output.



2.3 The result of distributed computing task is stored on HDFS.



Question 5 [15 marks]. Do some research on the Internet. Describe the types of applications for which Hadoop is not a good choice to be applied.

1. Applications which use many small files

Reason

1.1 Issue with Small Files

Hadoop does not suit for the situation with too many small data. It is a major problem of Hadoop. The reason is that Hadoop distributed file system (HDFS) lacks the ability to efficiently support the random reading of small files because of its high capacity which is designed for working properly with a small number of large files for storing large data sets. If we are storing huge numbers of small files which is significantly smaller than the HDFS block size (default 128MB), the NameNode will get overload since it stores the namespace of HDFS. Therefore, HDFS can't handle too much of small files.

2. Applications which need real-time data processing

Reason

2.1 Processing Overhead

In Hadoop, the data is read from the disk and written to the disk which makes read/write operations very expensive when we are dealing with tera and petabytes of data. Hadoop cannot do in-memory calculations hence it incurs processing overhead. Therefore, the speed of processing data in Hadoop is slow which is not suitable to the real-time data processing.

2.2 Support for Batch Processing only

At the core, Hadoop has a batch processing engine which is not efficient in stream processing. It cannot produce output in real-time with low latency. It only works on data which we collect and store in a file in advance before processing.

3. Applications which store Sensitive Data

Reason

3.1 Vulnerable by Nature

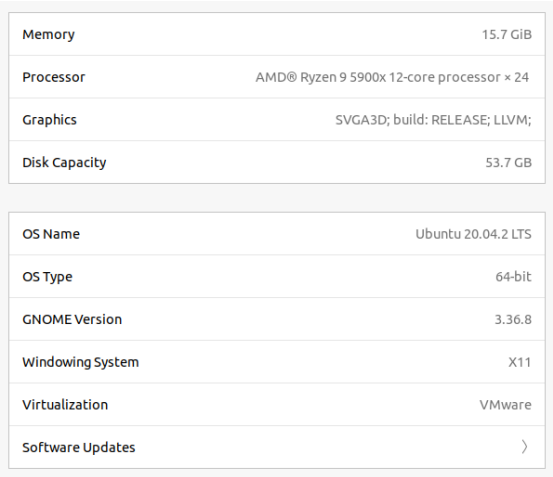
Hadoop is entirely written in Java, a language most widely used, hence java been most heavily exploited by cyber criminals and as a result, implicated in numerous security breaches.

3.2 Security

Hadoop is challenging in managing the complex application. If the user doesn't know how to enable a platform who is managing the platform, your data can be a huge risk. At storage and network levels, Hadoop is missing encryption, which is a major point of concern. Hadoop supports Kerberos authentication, which is hard to manage.

Question 6 [25 marks]. Hands-on Exercises. Do some research on the Internet. You need to complete this question in an Apache Hadoop environment installed in a Linux box. Suppose that you have installed one Virtual Machine (VM) (VirtualBox or VMware) in your machine.

1. Show the Linux distribution name and version. Capture a screenshot and insert here.




```
fishandwasabi@hadoop1:~$ lsb_release -a
No LSB modules are available.
Distributor ID: Ubuntu
Description:    Ubuntu 20.04.2 LTS
Release:        20.04
Codename:       focal
```

2. Open a terminal in the Linux. What is the command to list the content in your home directory? Show a screenshot with the result.

```
fishandwasabi@hadoop1:~$ ls -lsa
total 80
4 drwxr-xr-x 16 fishandwasabi fishandwasabi 4096 Mar  4 11:04 .
4 drwxr-xr-x  5 root          root          4096 Mar  5 17:04 ..
4 -rw-------  1 fishandwasabi fishandwasabi  267 Mar 10 15:07 .bash_history
4 -rw-r--r--  1 fishandwasabi fishandwasabi  220 Mar  4 00:33 .bash_logout
4 -rw-r--r--  1 fishandwasabi fishandwasabi 3771 Mar  4 00:33 .bashrc
4 drwxrwxr-x 13 fishandwasabi fishandwasabi 4096 Mar  9 16:55 .cache
4 drwx------ 11 fishandwasabi fishandwasabi 4096 Mar  4 11:04 .config
4 drwxr-xr-x  2 fishandwasabi fishandwasabi 4096 Mar  4 13:59 Desktop
4 drwxr-xr-x  2 fishandwasabi fishandwasabi 4096 Mar  4 00:54 Documents
4 drwxr-xr-x  2 fishandwasabi fishandwasabi 4096 Mar  4 10:34 Downloads
4 drwx------  3 fishandwasabi fishandwasabi 4096 Mar  4 00:54 .gnupg
4 drwxr-xr-x  3 fishandwasabi fishandwasabi 4096 Mar  4 00:54 .local
4 drwx------  5 fishandwasabi fishandwasabi 4096 Mar  4 01:24 .mozilla
4 drwxr-xr-x  2 fishandwasabi fishandwasabi 4096 Mar  4 00:54 Music
4 drwxr-xr-x  2 fishandwasabi fishandwasabi 4096 Mar  4 00:54 Pictures
4 -rw-r--r--  1 fishandwasabi fishandwasabi  807 Mar  4 00:33 .profile
4 drwxr-xr-x  2 fishandwasabi fishandwasabi 4096 Mar  4 00:54 Public
4 drwxr-xr-x  3 fishandwasabi fishandwasabi 4096 Mar  4 01:24 snap
0 -rw-r--r--  1 fishandwasabi fishandwasabi    0 Mar  4 01:13 .sudo_as_admin_su
ccessful
4 drwxr-xr-x  2 fishandwasabi fishandwasabi 4096 Mar  4 00:54 Templates
4 drwxr-xr-x  2 fishandwasabi fishandwasabi 4096 Mar  4 00:54 Videos
```

3. Create a folder named after your university user account name such as "abc123" in your home directory. Show the command and screenshot.

```
fishandwasabi@hadoop1:~$ mkdir 320180939611
fishandwasabi@hadoop1:~$ ls -lsa
total 84
4 drwxr-xr-x 17 fishandwasabi fishandwasabi 4096 Mar 12 15:15 .
4 drwxr-xr-x  5 root          root          4096 Mar  5 17:04 ..
4 drwxrwxr-x  2 fishandwasabi fishandwasabi 4096 Mar 12 15:15 320180939611
4 -rw-------  1 fishandwasabi fishandwasabi  267 Mar 10 15:07 .bash_history
4 -rw-r--r--  1 fishandwasabi fishandwasabi  220 Mar  4 00:33 .bash_logout
4 -rw-r--r--  1 fishandwasabi fishandwasabi 3771 Mar  4 00:33 .bashrc
4 drwxrwxr-x 13 fishandwasabi fishandwasabi 4096 Mar  9 16:55 .cache
4 drwx------ 11 fishandwasabi fishandwasabi 4096 Mar  4 11:04 .config
4 drwxr-xr-x  2 fishandwasabi fishandwasabi 4096 Mar  4 13:59 Desktop
4 drwxr-xr-x  2 fishandwasabi fishandwasabi 4096 Mar  4 00:54 Documents
4 drwxr-xr-x  2 fishandwasabi fishandwasabi 4096 Mar  4 10:34 Downloads
4 drwx------  3 fishandwasabi fishandwasabi 4096 Mar  4 00:54 .gnupg
4 drwxr-xr-x  3 fishandwasabi fishandwasabi 4096 Mar  4 00:54 .local
4 drwx------  5 fishandwasabi fishandwasabi 4096 Mar  4 01:24 .mozilla
4 drwxr-xr-x  2 fishandwasabi fishandwasabi 4096 Mar  4 00:54 Music
4 drwxr-xr-x  2 fishandwasabi fishandwasabi 4096 Mar  4 00:54 Pictures
4 -rw-r--r--  1 fishandwasabi fishandwasabi  807 Mar  4 00:33 .profile
4 drwxr-xr-x  2 fishandwasabi fishandwasabi 4096 Mar  4 00:54 Public
4 drwxr-xr-x  3 fishandwasabi fishandwasabi 4096 Mar  4 01:24 snap
0 -rw-r--r--  1 fishandwasabi fishandwasabi    0 Mar  4 01:13 .sudo_as_admin_su
ccessful
4 drwxr-xr-x  2 fishandwasabi fishandwasabi 4096 Mar  4 00:54 Templates
4 drwxr-xr-x  2 fishandwasabi fishandwasabi 4096 Mar  4 00:54 Videos
```

4. Go to the folder created at previous step. Create a text file named "test.txt" with at least three lines. Show the commands and screenshots.

```
fishandwasabi@hadoop1:~$ nano 320180939611/test.txt
fishandwasabi@hadoop1:~$ cat 320180939611/test.txt
test
test
test
test
```

5. What command you use to show the Hadoop version? Show the screenshot.

The Hadoop was installed in the "hadoop" user's environment. Therefore, I need to switch to "hadoop" user firstly.

```
fishandwasabi@hadoop1:~$ su hadoop
Password:
hadoop@hadoop1:/home/fishandwasabi$ cd ~
hadoop@hadoop1:~$ hadoop version
Hadoop 3.2.2
Source code repository Unknown -r 7a3bc90b05f257c8ace2f76d74264906f0f7a932
Compiled by hexiaoqiao on 2021-01-03T09:26Z
Compiled with protoc 2.5.0
From source with checksum 5a8f564f46624254b27f6a33126ff4
This command was run using /home/hadoop/hadoop-3.2.2/share/hadoop/common/hadoop
-common-3.2.2.jar
```

6. List the content of the HDFS in your machine. Show the commands and screenshots.

6.1 Start the HDFS server

```
hadoop@hadoop1:~$ start-dfs.sh
Starting namenodes on [hadoop1]
Starting datanodes
Starting secondary namenodes [hadoop1]
```

6.2 List the content

```
hadoop@hadoop1:~$ hdfs dfs -ls -R /
```

Because the pervious demo, there are many files in HDFS.

7. Delete all content of the HDFS. Show the commands and screenshots.

```
hadoop@hadoop1:~$ hdfs dfs -ls /
Found 2 items
drwx------   - hadoop supergroup          0 2021-03-12 14:46 /tmp
drwxr-xr-x   - hadoop supergroup          0 2021-03-12 14:46 /user
hadoop@hadoop1:~$ hdfs dfs -rm -r /tmp
Deleted /tmp
hadoop@hadoop1:~$ hdfs dfs -rm -r /user
Deleted /user
```

8. Copy the “test.txt” file to HDFS. Show the commands and screenshots.

8.1 Create “/usr/hadoop ”

```
hadoop@hadoop1:~$ hdfs dfs -mkdir /user
hadoop@hadoop1:~$ hdfs dfs -mkdir /user/hadoop
```

8.2 Copy the “test.txt” file to HDFS

```
hadoop@hadoop1:~$ hdfs dfs -put /home/fishandwasabi/320180939611/test.txt test.txt
hadoop@hadoop1:~$ hdfs dfs -cat test.txt
test
test
test
test
```

9. Copy the “test.txt” file in HDFS to your home directory as “test-from-HDFS.txt”. Show the commands and screenshots.

```
hadoop@hadoop1:~$ hdfs dfs -get test.txt ~/test-from-HDFS.txt
hadoop@hadoop1:~$ cat test-from-HDFS.txt
test
test
test
test
```

Reference:

[1] https://www.webhostingsecretrevealed.net/blog/web-business-ideas/paas-examples/

[2] https://www.webhostingsecretrevealed.net/blog/web-business-ideas/iaas-examples/

[3] https://joshfechter.com/software-service-examples/

[4] https://getnerdio.com/academy/10-popular-software-service-examples/

[5] https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

[6] https://core.ac.uk/download/pdf/38126642.pdf

[7] https://techvidvan.com/tutorials/how-hadoop-works-internally/

[8] https://www.salesforce.com/products/platform/best-practices/benefits-of-cloud-computing/

[9] https://www.manageengine.com/device-control/data-replication.html

[10] https://hadoop.apache.org/docs/current/

[11] https://data-flair.training/blogs/13-limitations-of-hadoop/

[12] https://www.datanami.com/2014/01/27/when_to_hadoop_and_when_not_to/

[13] http://www.bigdatacompanies.com/5-big-disadvantages-of-hadoop-for-big-data/

[14] https://data-flair.training/blogs/advantages-and-disadvantages-of-hadoop/

[15] https://data-flair.training/blogs/13-limitations-of-hadoop/

[16] Wang L, Von Laszewski G, Younge A, et al. Cloud computing: a perspective study[J]. New generation computing, 2010, 28(2): 137-146.

[17] Dillon T, Wu C, Chang E. Cloud computing: issues and challenges[C]//2010 24th IEEE international conference on advanced information networking and applications. Ieee, 2010: 27-33.

[18] Qian L, Luo Z, Du Y, et al. Cloud computing: An overview[C]//IEEE International Conference on Cloud Computing. Springer, Berlin, Heidelberg, 2009: 626-631.

[19] Labrinidis A, Jagadish H V. Challenges and opportunities with big data[J]. Proceedings of the VLDB Endowment, 2012, 5(12): 2032-2033.

[20] Shvachko K, Kuang H, Radia S, et al. The hadoop distributed file system[C]//2010 IEEE 26th symposium on mass storage systems and technologies (MSST). Ieee, 2010: 1-10.

[21] Alam A, Ahmed J. Hadoop architecture and its issues[C]//2014 International Conference on Computational Science and Computational Intelligence. IEEE, 2014, 2: 288-291.