Name: Pong Kien Yiep
Student ID: 0341541
Module code: ITS66904

Question 1

The big data refer to how the company can manage a large amount of data and ensure the data processed can produce a good result, while the data mining means that how the business can search and retrieve the important information in the large amount of data by identify the pattern in large amount of data.

Question 2

The Hadoop HDFS is a great software to store and manage the big data. To manage the file, a huge file will be separated into few smaller files and then stored inside different machines. When the huge file breaks into smaller files, a copy of smaller files will also be created, and goes into different nodes, which can allow the user to store the big data in a distributed way. This way, even though one machine has failed, the data is still safe on another machine, since there is a copy of data in other machine. Hence, it can defend against the single point failure and ensure the availability of the data. Besides that, HDRS also use a technique called MapReduce which can allow the parallel processing, which can make the process easier and speed up the process.

Question 3

MapReduce is a technique used by the HDRS. MapReduce used to process big data. For example, this technique will separate a big task into few smaller tasks, and then different machines will take up each task, and then each machine will process the data, and complete the processing in parallel, and then combine the result at the end. Hence, this technique can provide parallel processing, which make the processing become faster and easier.

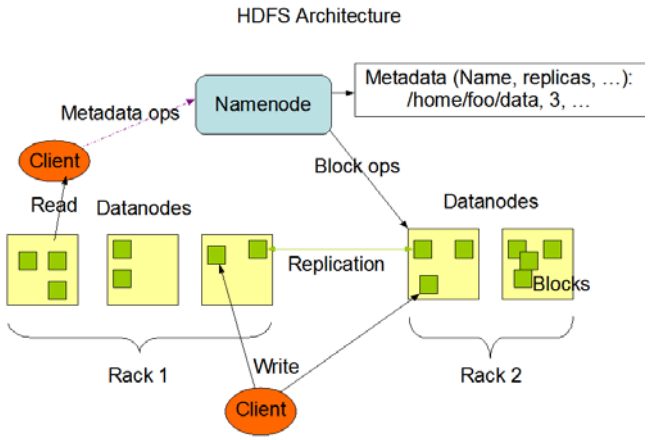
Question 4

One of the benefits of using Hadoop is scalability because Hadoop can store and distribute the data across hundreds of low-cost servers and allow the server to add node to scale the cluster. Hence, it can benefit the business to run the application on thousands of nodes and manage high amount storage of data. Besides that, another benefit of using Hadoop is speed, since Hadoop is using the map reduce which can allow the parallel processing, which each data can be stored and processed inside the different node in multiple devices. Hence, multiple device can process the data at the same time and then produce an output, and then the output will be combined to achieve the final result, which can increase the processing speed due to parallel processing.

Question 5

In Hadoop, there are many different types of operations are running, and each operation will need some resource to complete the task. The resource will need to be managed, and the yarn will be used to manage the resources. Hence, yarn will ensure each role or operation will retrieve its right resource, so that the operation can be continue processed without any small failure.

Question 6



HDFS contain a master server called Namenode, which is responsible in managing the file system, and allow the files to be accessed by the clients. Besides of Namenode, there are also number of Datanodes, which one Datanode will manage one node in the cluster. The Datanode will be responsible in manage the storage in the node. For example, the HDFS will store the user data in the files, and then the files will be split into few blocks, and these blocks will be stored in the Datanodes. The Namenode will process the file operation like opening, closing, and renaming the files. Besides that, the Namenode will also map the block to the Datanode, and also responsible in creating, deleting and replicating the block. Instead of managing the block, the Datanode also will serve the client by performing the read and write request.