**Assignment 1**

**Question 1: Can you think of a use case of Big Data?  Explain it briefly.**

Personal Health Record.

It is able that people can record each personal sickness symptoms, allergy elements and history, patient record… from birthdate and keep it up-to-date. So that, the doctor can easily figure out what happened when one person got ill or predict what possible sickness.

**Question 2: What are the advantages of using Hadoop and HDFS?**

* Storing the large files (TB, PB+)
* Streaming data
* Using cheap commodity hardware

**Question 3: Explain the term block abstraction in Hadoop and state it's advantages.**

* Hadoop block abstraction is the term that describe HDFS block size is of 128MB (64MB for Hadoop 1.x) and unlike other filesystems, a file which size is smaller than the block size does not occupy the fully block size’s worth of memory.
* Hadoop block abstraction’s advantage:
  + The block size is kept so large so that less time is made doing disk seeks as compared to the data transfer rate.
  + Files can be bigger than individual HDD.
  + Filesystem metadata does not need to be associated with each and every block.
  + Simplifies storage management - Easy to figure out the number of blocks which can be stored on each disk.
  + Fault tolerance and storage replication can be easily done on a per-block basis (storage/HA policies can be run on individual blocks).

**Question 4: What is the meaning of fault tolerance in HDFS and how is it achieved?**

The fault tolerance in HDFS refers to the process of HDFS system to handle the “failed” cases of data- missing or DataNode(s)/Cluster(s)/Rack(s) is down…

HDFS is highly fault tolerant. It handles faults by the process of data replication. The replica of data block is created and stored on different DataNode/Rack in the HDFS cluster. So whenever if any machine in the cluster goes down, then user still can accessed that data block (replica) which stored in other machines.

**Question 5: Consider a 560 TB of text file which needs to be stored in HDFS. The block size has been set to be 128 MB with a replication factor of 3. The cluster has 100 DataNodes each with a capacity of 15 TB. Will it be possible to store this text file in this HDFS cluster? Why or why not**?

No, this 560 TB file can’t be stored in this HDSF cluster.

The total capacity of this HDFS cluster:

15\*100 = 1500 (TB)

Due to block size is 128 MB, this file is split into:

(560\*2^40) / (128\*2^20) = 4.375\*2^20 (blocks)

Due to this HDFS system’s **replication factor is 3**, so total block number**:**

3 \* (4.375\*2^20) = 13.125\*2^20 (blocks)

In capacity, this will cost:

(13.125\*2^20) \* 128 = 1680 (TB)

It is clear that the total capacity cost is greater than the total capacity of HDFS cluster. Hence, this file can’t be stored in this HDFS cluster.