BDT- Assignment-1 (Sujiv Shrestha 610145)

1. [2] **Can you think of a use case of Big Data? Explain it briefly.**

   One of the use case of Big Data can be in following fields:

   A. **Communication System** - analysis of movement of people based on their cell-phone location can provide business intelligence for cellular network expansion etc.
   B. **Agriculture** – The analysis of climate data can suggest which geolocation will best fit for certain crops.
   C. **Stock Market** – The analysis of stock exchange data for many years can crack down the definite pattern to estimate which company will gain or lose.
   D. **Natural Disaster Prevention** – Some natural disasters like Tsunami and Earthquake can not be predicted timely by analyzing little data but when analyzed huge data of bed rock movement of earth, it may generate some timely signals to warn the people at risk.

2. [2] **What are the advantages of using Hadoop and HDFS?**

   Hadoop and HDFS uses the strategy of divide and conquer which is essential when it comes to big data. Advantages of using Hadoop and HDFS are as follows:

   i. It enables processing of huge data faster and more efficiently.
   ii. It assures reliability through replica of data file chunks.
   iii. It is scalable.
   iv. It does not have high requirements.

3. [2] **Explain the term block abstraction in Hadoop and state it's advantages.**

   In Hadoop, the huge files are divided into chunks called HDFS blocks, of size 64 MB – 128 MB before they are stored and the distributed processing is done over these blocks of files. This is called block abstraction in Hadoop. The advantage of block abstraction are as follows:

   i. It enables storage of very huge files where a file can be bigger than the individual disk in the data-node.
   ii. The block size is comparatively large so that disk seeks are made less frequent while transferring data.
   iii. High fault-tolerance due to the replication of blocks.
   iv. Efficient because if the file chunk size is less than the block size it does not occupy the whole block memory storage.
   v. Enables parallel processing on different file chunks at the same time resulting faster computation.

4. [2] **What is the meaning of fault tolerance in HDFS and how is it achieved?**

   Fault tolerance in HDFS refer to the recovery of data when the file chunk stored in one of the data node is corrupted.

   It is achieved in HDFS by replicating the file blocks with replication factor which guarantees minimum number of copies of each block of files on different nodes within the cluster even when some of the data-nodes have hardware or network failure.

5. [2] **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**?

BDT- Assignment-1 (Sujiv Shrestha 610145)

No, it is not possible to store 560TB of text file in HDFS of cluster size of 100 nodes with 15TB memory each. Because of the following fact:

Total memory capacity of data nodes in cluster: Disk capacity in one data node X cluster size

= 15TB X 100

= 1500TB

Minimum Storage Required in data-nodes : File Size X Replication Factor

= 560TB X 3

= 1680TB

So, total data-node memory capacity of cluster is less than minimum storage required and hence it can not be stored in the given cluster.