cs523 - BDT

Ans Assignment 1 – Day 1 (Monday)

DO NOT COPY, SHARE OR REPRODUCE THIS MATERIAL AS IT CAN GET YOU EXPELLED FROM M.U.M.

1. **Can you think of a use case of Big Data? Explain it briefly.**
   (Do not repeat the ones from the slides!)

2. **What are the advantages of using Hadoop and HDFS?**
   ANS:
   Following are the various advantages of using Hadoop and HDFS.
   1) The main advantage of Hadoop is that it is highly reliable.
   2) It is simple and robust.
   3) It is used to store large data set without any limit on storage.*
   4) It is highly scalable storage platform.
   5) Cost effective as it is 100% open source software.
   6) Quick recovery from system failures.
   7) Ability to rapidly process large amounts of data in parallel.
   8) Once data written in HDFS can be read several times.
   9) Provides Faults tolerance by detecting faults and provide mechanism for overcoming them.

3. **Explain the term block abstraction in Hadoop.**
   ANS:
   This term can be explained in 2 different ways.
   1) DNs don't know anything about the disk blocks that they are storing is called as block abstraction.
   2) Internally in Hadoop, a file is split into one or more blocks and these blocks are stored in a set of DataNodes. DataNodes also perform block creation, deletion, and replication upon instruction from the NameNode. A typical block size used by HDFS is 64 MB. (128 MB in Hadoop 2) The block size is kept so large so that less time is required for doing disk seeks as compared to the data transfer rate.
   Advantages of block abstraction:
   1. Files can be bigger than individual disks
   2. Filesystem metadata does not need to be associated with each and every block.
   3. Simplifies storage management - Easy to figure out the number of blocks which can be stored on each disk.
   4. Fault tolerance and storage replication can be easily done on a per-block basis.

4. **What is the meaning of fault tolerance in HDFS and how is it achieved?**
   ANS:
   Fault tolerance means, system continues to function correctly even after some components fail to work properly. Fault tolerance is mainly achieved using data duplication and making copies of same data sets in two or more data nodes.

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?

    Not possible; it needs at least (560*3) = 1680 TB of storage space and 100 DNs together has only 1500 TB space.

DO NOT COPY, SHARE OR REPRODUCE THIS MATERIAL AS IT CAN GET YOU EXPELLED FROM M.U.M.