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

Another use case of Big Data is in transportation and logistics. With the increasing number of connected devices, sensors, and GPS technologies, the transportation and logistics industry generate massive amounts of data on shipping routes, vehicle performance, and delivery times.

1. [2] **What are the advantages of using Hadoop and HDFS?**
   * Hadoop is written in Java and is supported on all major platforms.
   * Designed to answer the question: “How to store and process big data with reasonable cost and time?”
   * HDFS is responsible for storing large data sets on the cluster, and it can be built out of commodity hardware.
   * Streaming data access patterns
     1. Write once and read many times.
     2. Where getting the entire data faster is more important than getting a specific record
   * HDFS provides Java API for applications to use.
   * HTTP browser can be used to browse the files of the HDFS instance.
2. [2] **Explain the term block abstraction in Hadoop.**
   * Like in a filesystem for a single disk, files in HDFS are broken into block-sized chunks, which are stored as independent units.
   * A typical block size is 128 MB (64 MB in Hadoop 1.x).
   * Unlike a filesystem for a single disk, a file in HDFS that is smaller than a single block does not occupy a whole block’s worth of underlying storage. For example, a 1 MB file stored with a block size of 128 MB uses 1 MB of disk space, not 128 MB.
   * Blocks are stored on standard single-machine file systems, so HDFS lies on top of the standard OS stack (e.g., Linux).
3. [2] **What is the meaning of fault tolerance in HDFS and how is it achieved?**

Fault tolerance in HDFS refers to the ability of the system to continue operating without data loss or downtime in the event of hardware failures or other types of faults. This is achieved through the replication of data across multiple nodes in the cluster:

* + In HDFS each file is a sequence of blocks.
  + All blocks in the file except the last one are the same size.
  + Blocks are replicated for fault tolerance.
  + Block size and the number of replicas are configurable for files.
  + The replication factor is 3 by default.
  + NameNode tries to place replicas of blocks on multiple racks for improved fault tolerance.

1. [2] **Consider a 260 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 10 TB.  
   Will it be possible to store this text file in this HDFS cluster? Why or why not**?

Yes, It is possible.

The total capacity of this HDFS cluster:

(T) = 10\*100 = 1000 (TB)

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

(260\*2^40) / (128\*2^20) = 2.03125\*2^20 (blocks)

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

3 \* (2.03125\*2^20) = 6.09375\*2^20 (blocks)

In capacity, this will cost:

(C) = (6.09375\*2^20) \* 128 = 780 (TB)

The total capacity of the HDFS cluster (T) is greater than the total capacity cost (C).

**This means the file can be stored in this HDFS cluster.**