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

Human Behavioural Analytics

Most of the organizations and to be specific almost 50% organizations use big data to understand meaningful insights from customer behaviour data.

The beauty of big data lies in understanding the customer behaviour. Organizations are bridling the force of huge information through conduct examination to convey huge worth to organizations. Organizations that utilization conduct examination to foresee client conduct have quite recently gone up ten times in increasing the value of their business. Amazon has dominated the proposal of items a long while back dependent on clients interest and different organizations like Spotify, Pinterest and Netflix are following a similar suit.

Bank of America uses big data for behavioural analytics

Bank of America rewards program named BankAmeriDeals rewards its customer with different cashback offers by analysing their previous credit and debit card purchase histories.

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

Hadoop is highly scalable storage platform and cost-effective storage solution for businesses.

The **DataNodes** that store the data rely on inexpensive off-the-shelf hardware, which cuts storage costs.

HDFS is open source, there's no licensing fee.

Large data set storage. HDFS stores variety of data in different size and in any format including structured and unstructured data.

Fast recovery from hardware failure. HDFS is designed to detect faults and automatically recover on its own.

HDFS is **portable** across all hardware platforms, and it is compatible with most of the operating systems, including Windows, Linux and Mac OS/X.

Streaming data access. HDFS is built for high data throughput, which is best for access to streaming data.

3. Explain the term block abstraction in Hadoop.

Normal file system and Hadoop architecture are little different. In normal file system if we create a blank file, then it holds the 4k size, as it is stored on the block. In HDFS it won't happen, for 500MB file it hold 500MB memory, not 1GB.

The block abstraction in HDFS is just logically built over the physical blocks of file system. The file system is not physically divided into blocks. 64MB or 128MB or whatever may be the block size. It's just an abstraction to store the metadata in the NameNode. Since the NameNode has to load the entire metadata in memory therefore there is a limit to number of metadata entries thus explaining the need for a large block size.

Therefore, three 200MB files stored on HDFS logically occupies 3 blocks but physically occupies 200*3= 600MB space in the file system.

The large block size is to account for proper usage of storage space while considering the limit on the memory of NameNode.

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

Fault tolerance can be defined as, When the system functions properly without any data loss even if some hardware components of the system has failed. HDFS provide high throughput to access data application and suitable to have large data sets as their input. The main purpose of this fault tolerance is to remove frequently taking place failures, which occurs commonly and disturbs the ordinary functioning of the system.

Single point failure nodes occur when a single node failure causes the entire system to crashes. The primary duty of fault tolerance is to remove such node which disturbs the entire normal functioning of the system. Fault tolerance is one of the major advantages of using Hadoop. The three main solutions which are used to produce fault tolerance are data replication, heartbeat messages and checkpoint and recovery.

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?

Let's calculate to see the answer.

Capacity = 15 TB
DataNodes = 10 (DataNodes each with capacity)

Block Size = 128 MB
Replication factor = 3
Text file size = 560 TB = 5.6x10^8 MB

Total Number of Blocks = 5600000000/128 x 3 = 131250000 (Approximately 13 million)

To store 13 million metadata, NameNode need to have 131250000 x 150 Bytes = 1950 MB or 2 GB Approximately.

So, it's possible.