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

   Sports performance analysis and betting optimization
   - By collecting and analyzing vast amounts of data from player statistics, historical game results, weather conditions, and even social media sentiment, sports analysts and betting companies can create highly accurate predictive models.
   - For teams, this data can help in strategizing game plans, optimizing player rotations, and even preventing injuries by analyzing player workload.
   - For betting companies, it enhances the accuracy of odds-setting and helps in identifying patterns that might influence betting behavior.

2. What are the advantages of using Hadoop and HDFS? Explain in your own words.

   - **Scalability**: Hadoop is built to grow across many servers. You may easily expand the Hadoop cluster by adding more nodes as your data expands, which enables it to handle growing data volumes without experiencing appreciable performance degradation.
   - **Cost-effective**: Compared to specialized, high-end servers, Hadoop uses commodity hardware, which is significantly less expensive. It is therefore an affordable option for processing and storing big datasets.
   - **Fault Tolerance**: Hardware malfunctions are accommodated for in HDFS. Since data is automatically replicated among several nodes, access to it from other nodes remains unaffected even in the event of a node failure.
   - **Flexibility**: Hadoop can handle a large range of data forms, including unstructured, semi-structured, and structured data. Because of its versatility, it may be used for a wide range of tasks, such as social media analysis and log processing.
   - **High Throughput**: Hadoop is designed to analyze big datasets quickly and efficiently in parallel. It achieves high throughput by dividing the task among several nodes, making it possible to analyze large volumes of data efficiently.
   - **Support from the Community**: Hadoop is home to a sizable and vibrant community that constantly works to enhance it. This indicates that users and developers have access to a wealth of resources, tools, and updates.

3. In Hadoop, what is the default block size and what are the advantages of block abstraction?

   Default Block Size is 128 Mb

   Advantage of Block abstraction

   - Efficient Storage Management: By splitting large files into blocks, Hadoop can efficiently manage storage across a distributed system.
   - Fault Tolerance: Blocks are replicated across multiple nodes (by default, each block is replicated three times). This replication ensures that even if one or more nodes fail, the data remains accessible from other nodes that hold copies of the blocks.

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

   Fault tolerance in HDFS refers to the system's ability to continue functioning correctly even in the presence of hardware or software failures.

   Fault tolerant can be achieved through

   - **Replication**: - If one node fails, the data is still available from the other nodes that hold the replicas, ensuring that no data is lost.
   - **Rack Awareness**: - HDFS ensures that even if an entire rack fails, data is still accessible from replicas stored in other racks.
   - **Auto Re-Replication: -** This process ensures that the replication factor is restored, maintaining the fault tolerance of the system.

5. Consider a 360 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 Data Nodes each with a capacity of 10 TB.
   Will it be possible to store this text file in this HDFS cluster? Why or why not?

   Total storage required=360TB×3=1080TB
   Total storage available=100×10TB=1000TB

   The HDFS cluster does not have enough capacity to store the 360 TB text file with a replication factor of 3. The cluster has only 1000 TB of available storage, but 1080 TB is needed. Therefore, it will not be possible to store this file in the current HDFS cluster configuration. You would need to either increase the number of Data Nodes or reduce the replication factor (which could impact fault tolerance) to store this file.