Q). Why Spark uses HDFS to process billions of records?

Apache Spark is a distributed processing engine that is designed to handle large-scale data processing workloads efficiently. One of the key benefits of Spark is its ability to process large volumes of data in parallel across a cluster of machines.

Hadoop Distributed File System (HDFS) is a distributed file system that is designed to store and manage large volumes of data across a cluster of machines. HDFS is highly fault-tolerant and can scale to handle petabytes of data.

Spark leverages HDFS to process large volumes of data because HDFS provides several advantages that make it a good fit for big data processing with Spark. Some of the benefits of using HDFS with Spark include:

1. Scalability: HDFS is designed to scale to handle large volumes of data, making it an ideal storage solution for Spark workloads that process billions of records.
2. Fault-tolerance: HDFS is highly fault-tolerant, meaning that data stored in HDFS is replicated across multiple nodes in the cluster. This ensures that even if a node fails, the data remains available and processing can continue without interruption.
3. Data locality: HDFS stores data in a distributed manner across the cluster, and Spark can take advantage of this data locality to process the data where it resides. This can reduce network overhead and improve processing performance.

Overall, the combination of Spark and HDFS provides a powerful platform for processing big data workloads efficiently and effectively.