# Big Data and Distributed Computing - Q&A

## 68. What is Hadoop, and how does it handle big data?

Hadoop is an open-source framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It handles big data by breaking it into smaller chunks, distributing them across nodes in a cluster, and processing them in parallel using the MapReduce programming model.

## 69. Explain the MapReduce programming model.

MapReduce is a programming model used for processing large data sets with a distributed algorithm. It consists of two main functions: the Map function, which processes input data and produces key-value pairs, and the Reduce function, which aggregates and summarizes the results. This model enables scalable and fault-tolerant data processing.

## 70. What is Apache Spark, and why is it popular in big data processing?

Apache Spark is an open-source distributed computing system that provides an interface for programming entire clusters with implicit data parallelism and fault tolerance. It is popular because it is faster than Hadoop MapReduce due to in-memory processing, supports multiple languages, and includes libraries for SQL, streaming, machine learning, and graph processing.

## 71. Describe the concept of distributed computing.

Distributed computing is a model in which components of a software system are shared among multiple computers to improve efficiency and performance. It allows for parallel processing, fault tolerance, and scalability, making it ideal for handling large-scale computations and big data workloads.

## 72. What are the advantages and disadvantages of distributed databases?

Advantages of distributed databases include improved reliability, availability, and scalability, as well as faster data access for geographically distributed users. Disadvantages include increased complexity in database management, potential data consistency issues, and higher maintenance costs.