一、BASIC TASKS

a) Explain the meaning of the term "big data" and give examples

"Big data" refers to a large, complex and diverse collection of data that cannot be managed, analyzed, and processed in a reasonable amount of time by traditional data processing tools.

For example, social media platforms generate massive amounts of user-generated content every day, including text, images, videos, and more. For example, hundreds of millions of posts are published on Weibo every day, covering a variety of topics, emotional expressions, and user behavior data; The user's shopping behavior data recorded by the e-commerce platform, including browsing history, purchase history, search keywords, etc., is huge and diverse, and belongs to the category of big data.

b) How big data analytics can help increase business revenue

Big data analytics can increase business revenue in the following ways:
1. Precision marketing: By analyzing user data to understand users' interests, needs, and behavior patterns, enterprises can carry out more accurate advertising and personalized recommendations, improve marketing effectiveness, and increase sales conversion rates. For example, Amazon recommends relevant products for users based on their browsing history and purchase history to increase their purchase intent.
2. Optimize operations: By analyzing data from production, sales, supply chain and other links, enterprises can find inefficient links and optimize them to reduce costs and improve production efficiency. For example, manufacturing companies can reduce production interruptions by analyzing equipment sensor data to predict equipment failures and perform maintenance in advance.
3. Innovative products and services: By using big data to analyze market trends and user needs, companies can develop products and services that are more in line with market demand. For example, online music platforms analyze users' listening habits and preferences to launch personalized music recommendation services and exclusive playlists to improve user experience and increase user stickiness and willingness to pay.

c) Describe the difference between structured and unstructured data

Structured data has a well-defined data structure and format, is typically stored in a relational database, and can be represented and managed in tabular form. For example, a company's sales order data, including order numbers, customer information, product information, order amounts, etc., has fixed fields and formats that can be accurately queried and analyzed.

Unstructured data has no fixed structure and format, making it difficult to represent and manage with traditional database tables. For example, text files, images, audio, video, etc. This data often requires the use of special techniques and tools for processing and analysis. For example, natural language processing technology is used to analyze text data, and image recognition technology is used to process image data.

d) Create a chart that represents the big data you've contributed

Since it's unclear what you're contributing, let's take a social media user as an example. You can create a graph that shows the data generated by users on social media platforms, such as the number of posts posted, the number of likes, the number of comments, the number of users followed, the number of users being followed, etc. This data can be presented in the form of a bar chart, line chart, or pie chart.

e) Why do we need data-intensive systems? Name some of the data-intensive systems

We need data-intensive systems because the scale and complexity of big data makes traditional data processing systems inadequate. Data-intensive systems are capable of handling large-scale, high-growth data sets and provide efficient data storage, analysis, and processing capabilities.

Some data-intensive systems include:
1. Distributed file systems: such as Hadoop Distributed File System (HDFS), which is used to store large-scale datasets and provide high reliability and high availability.
2. Distributed databases, such as Cassandra, MongoDB, etc., are able to handle large-scale structured and unstructured data, and provide high scalability and high availability.
3. Data warehouse: It is used to store and manage the historical data of the enterprise, and provide data analysis and decision support. For example, Teradata, Oracle Exadata, etc.
4. Big data analysis platforms: such as Apache Spark, Hive, etc., to provide efficient big data analysis and processing capabilities.

f) Briefly describe examples of data-intensive technologies that can be used for data storage, data visualization and analysis, computation and distribution, and data warehousing

1. Data storage: distributed file systems (such as HDFS), distributed databases (such as Cassandra, MongoDB), object storage (such as Amazon S3), etc.
2. Data visualization and analysis: data visualization tools (such as Tableau, PowerBI), data analysis platforms (such as Apache Spark, Hive), machine learning algorithms (such as random forests, support vector machines), etc.
3. Compute and distribution: Distributed computing frameworks (e.g. Apache Spark, Hadoop MapReduce), stream processing platforms (e.g. Apache Flink, Storm), message queues (e.g. Kafka, RabbitMQ), etc.
4. Data warehouses: traditional data warehouses (such as Teradata, Oracle Exadata), Hadoop-based data warehouses (such as Hive, Impala), etc.

二、MEDIUM TASKS

a) "Data is known as the oil of the 21st century" for discussion

Data is known as the oil of the 21st century because it has a similar importance in today's society as oil did in the industrial age.

First and foremost, data is a valuable resource. Just as oil is an important raw material for industrial production, data is a core resource in the era of digital economy. Businesses can gain a competitive advantage by analyzing data to understand market demand, optimize operations, and innovate products and services.

Second, data has tremendous potential for value. Just as oil can be refined and processed into a variety of high value-added products, data can be analyzed and mined to generate valuable information and insights. This information can be used in decision support, precision marketing, risk assessment and other fields, bringing huge economic benefits to enterprises and society.

However, there are some differences between the data and oil. Oil is a finite natural resource, while data can be continuously generated and accumulated. At the same time, the value of data depends on its quality, availability, and analytical capabilities, not its mere quantity. In addition, there are legal, ethical, and privacy challenges to how data is owned and used.

b) Discuss the following definitions in relation to "accuracy"

1. "Accuracy is defined as uncertainty due to data inconsistencies and incompleteness, ambiguity, latency, deception, model approximation".

This definition highlights the multiple challenges to data accuracy. Data inconsistencies and incompleteness can be caused by errors in the data collection process, poor data storage and management, etc. Ambiguity can be due to ambiguity or ambiguity in the meaning of the data. Latency can affect the timeliness and availability of your data. Deception can be caused by data being deliberately tampered with or forged. Model approximation may be due to limitations in data analysis models.

2. "Other groups refer to data quality issues and include accuracy, trustworthiness, reputation, objectivity, truthfulness, consistency and unbiasedness, correctness and clarity".

This definition describes the requirements for data quality in a number of ways. Accuracy refers to how well the data matches the actual situation. Trustworthiness refers to the reliability and trustworthiness of the data. Reputation refers to the credibility and authority of a data source. Objectivity means that the data is not affected by subjective factors. Authenticity refers to the authenticity and reliability of the data. Consistency refers to the consistency of data across different times and from different sources. Unbiased means that the data is not affected by bias and discrimination. Correctness refers to the correctness and accuracy of the data. Clarity means that the meaning of the data is clear and there is no ambiguity.

三、ADVANCED TASKS

**a)

Kaggle's e-commerce dataset, which contains data on users' shopping behavior, product information, orders, and more, can be used to analyze user behavior, predict sales trends, optimize inventory management, and more, making it ideal for creating data-intensive systems.

b) Explain the dataset and explain why the data is suitable for data-intensive systems

E-Commerce Datasets:

This dataset contains a large amount of user shopping behavior data, such as the user's browsing history, purchase history, search keywords, etc., as well as product information and order data. This data has the following characteristics that make it suitable for use in data-intensive systems:

1. Large amount of data: It contains millions or even tens of millions of records, which can meet the needs of big data processing.
2. Diverse types: Includes structured data (e.g., order data, product information) and unstructured data (e.g., user reviews), which need to be processed and analyzed using a variety of technologies.
3. Timeliness: Shopping behavior data and order data are constantly updated and need to be processed and analyzed in real time to provide timely decision support.
4. High business value: By analyzing these data, you can understand user needs, optimize operations, improve sales conversion rates, and bring huge business value to enterprises.

As a result, this e-commerce dataset is ideal for creating data-intensive systems to support data analysis and decision-making in businesses.