**Task 1**

Basic Tasks 1

Li Weixuan

Zhang chiyun

Chen ziang

a

Big data refers to a large - scale, diverse, and fast - processing data set. For example, in the field of e - commerce, companies like Amazon have to process a huge amount of user transaction data every day, including purchase records, browsing history, search keywords, etc. These data are not only huge in quantity (there may be hundreds of millions of transactions every day), but also diverse in data types, including structured order information (such as commodity number, price, purchase time, etc.) and unstructured user evaluation text. The rapid generation and accumulation of such data require special technologies and tools for processing and analysis.

b)

- Customer segmentation: By analyzing big data such as customers' purchasing behavior, preferences, and demographic information, enterprises can segment customer groups more accurately. For example, an online clothing retailer can divide customers into different groups such as fashion seekers and cost - effective seekers according to their browsing and purchasing history, and then launch personalized marketing activities for different groups to improve marketing effects and sales.

- Product optimization: By using big data analysis of product usage data, user feedback and other information, enterprises can timely find problems and improvement directions of products. For example, mobile phone manufacturers determine the key points of function optimization of the next - generation products by analyzing the frequency of users' use of mobile phone functions and feedback opinions, so as to improve the competitiveness and market share of products and increase revenue.

- Supply chain optimization: By analyzing big data such as sales data, inventory data and supplier data, enterprises can optimize supply chain management. For example, predict the peak and trough of demand, arrange inventory reasonably, reduce inventory cost and out - of - stock loss, improve the efficiency and response speed of the supply chain, and then increase business revenue.

c)

- Structured data: It has a clear format and a predefined structure, usually stored in a relational database in the form of a table. For example, the basic information table of employees, which contains fields such as name, age, employee number, and entry date, and each field has a clear data type and length limit. This kind of data is easy to store, query and analyze because the data format is fixed and conforms to specific rules.

- Unstructured data: There is no fixed structure and various forms, such as text files, images, audio, video, etc. For example, the text of a news report, a painting, a song or a video. Unstructured data is difficult to be directly stored and processed by traditional relational databases, and special technologies such as text mining, image recognition and audio processing are required to extract useful information.

d) (Since the specific big data field and content are not clear, the following is a simple example framework of a big data graph)

Suppose taking a social media company as an example, the following is a chart representing big data:

Data Types Description Examples

User - generated text Text data created by users, such as posts, comments Tweets on Twitter, comments on Facebook

User profiles Structured data about users Age, gender, location in a user's profile on a social media platform

Media content Images, videos uploaded by users Photos on Instagram, videos on YouTube

e) Why we need data - intensive systems and list data - intensive systems:

We need data - intensive systems because in the era of big data, traditional data processing systems cannot effectively handle the huge amount of data with high complexity. Data - intensive systems are designed to process, store and analyze large - scale data. Examples of data - intensive systems include distributed file systems (such as Hadoop Distributed File System - HDFS), distributed databases (such as Cassandra), and data processing frameworks (such as Apache Spark).

f) Briefly describe examples of data - intensive technologies for data storage, data visualization and analysis, computing and distribution, and data warehousing:

- Data storage: Technologies such as HDFS provide a distributed and scalable way to store large amounts of data. It divides data into blocks and distributes them across multiple nodes in a cluster for reliable storage.

- Data visualization and analysis: Tools like Tableau and PowerBI enable users to create visualizations and perform in - depth analysis on large datasets. They offer various chart types and interactive features to help users understand data patterns and trends.

- Computing and distribution: Apache Spark is a powerful data processing framework that can perform in - memory computing and distributed processing. It can handle large - scale data processing tasks efficiently by distributing the workload across multiple nodes.

- Data warehousing: Technologies like Amazon Redshift and Snowflake provide cloud - based data warehousing solutions. They are designed to store and manage large amounts of data and support complex querying and analysis.

Medium Tasks

a) Data is called "the oil of the 21st century" because just as oil was the driving force behind the industrial revolution, data has become a crucial resource in the digital age. It powers businesses, drives innovation, and creates value in various fields.

b) Discussion on the following definitions related to "accuracy":

1. Accuracy is defined as the uncertainty caused by data inconsistency, incompleteness, ambiguity, delay, deception, and model approximation. This definition emphasizes the factors that can lead to inaccuracies in data. For example, if data is collected from multiple sources and there are inconsistencies between them, or if there are missing values in the data, it can affect the accuracy of the analysis results.

2. Other categories involve data quality issues including "accuracy, credibility, reputation, objectivity, authenticity, consistency, unbiasedness, correctness, and clarity". This broader definition encompasses multiple aspects of data quality. Accuracy refers to how close the data is to the true value. Credibility and reputation refer to the trustworthiness of the data source. Objectivity means the data is not influenced by personal biases. Authenticity ensures the data is genuine. Consistency means the data is consistent across different sources or over time. Unbiasedness means the data is not skewed in a particular direction. Correctness and clarity ensure the data is correct and easy to understand.

Advanced Tasks

a) Find big data sets that may help create data - intensive systems. You can consider looking for data sets that can be used in your area of interest. Possible data sources include:

https://archive.ics.uci.edu/ml/datasets.php, https://www.kaggle.com/, https://toolbox.google.com/datasetsearch, https://datarepository.wolframcloud.com/category/MachineLearning, http://data.europa.eu/euodp/en/data/, https://www.data.gov/

b) Explanation of the data set and indication of why you think this data is suitable for a data - intensive system:

For example, on Kaggle, you might find a data set related to weather patterns. This data set could be very useful for a data - intensive system in the field of meteorology. The reasons are as follows: Firstly, the data set is likely to be large in size as it may contain historical weather data from multiple locations and over a long period of time. Secondly, it requires complex processing to extract meaningful information such as weather trends, seasonal variations, and extreme weather events. Thirdly, the data can be used for various applications such as weather forecasting, climate research, and agricultural planning, which require a data - intensive system to handle the large - scale data and perform in - depth analysis.