Yue Ma, Yuheng LI and Conger Yang

BASIC TASKS

(a)Many company exploit short video application, which are based on big data. For example, network technology means, such as crawler, etc., to collect effective user information, after this pre-processing and data cleaning, stored in the distributed file system and database, the amount of these data is huge. In the above example, the first meaning of big data is explained: the ability to mine and store large amounts of data.
Tiktok also recommends what users want based on their preferences and habits. This is the second meaning of big data, which can be applied to the processed data.
In summary, big data is a technology that can mine, store and apply data.

(b) Based on the above explanation of big data, we can know that big data has the function of mining data. Enterprises use big data analysis and obtain useful information to make up for the gaps in management mode, market positioning, etc., so as to reduce the consumption of enterprises. Second, big data can be applied to data, here is an example, e-commerce platform: Amazon. Amazon analyzes data through big data, understands customers' preferences, and pushes relevant products to customers, thus increasing the revenue of enterprises. There are even technology companies that analyze huge amounts of data as their income. Therefore, big data analysis itself can bring revenue to enterprises, not only that, but also help other types of enterprises to provide effective advice and help in sales and strategy.

(c) The primary distinction between structured and unstructured data lies in their organization and storage methods. Structured data is typically stored in relational databases, such as SQL databases, with a fixed format and schema, where each record contains the same data fields like names, addresses, and phone numbers, making them easy to query and analyze. In contrast, unstructured data lacks a fixed format or schema, encompassing text files, emails, social media posts, images, and videos, which are generally more challenging to process and analyze due to their lack of a uniform structure.

(d)

[Data Sources]
↓
[Data Collection]
↓
[Data Storage]
↓
[Data Processing]
↓
[Data Analysis]
↓
[Data Visualization]
↓
[Data Application]

(e) Due to the rapid increase in the amount of modern data, people have been given the opportunity to discover patterns in data intensive systems through this vast

amount of data. This allows for the analysis of user behavior and the customization of personalized services, as well as the prediction of user behavior to a certain extent. Ultimately, data intensive systems can better protect user privacy. Examples include Tiktok's push videos and Taobao's recommended products. They can predict your next behavior under your behavior, so as to provide accurate services.

(f) 1、 Data storage

HDFS (Hadoop Distributed File System) can store massive amounts of data in large-scale clusters. It disperses data across multiple nodes, improving storage capacity and reliability. There is redundant storage, so even if some nodes are broken, data cannot be lost. Suitable for storing large-scale unstructured and semi-structured data, such as log files, images, videos, etc.

2、 Data visualization and analysis

Tableau is a powerful data visualization tool that can connect various data sources such as databases, spreadsheets, and cloud storage. Users can create interactive dashboards and reports through operations, visually observing the trends, distributions, and relationships of data. It supports multiple types of charts, such as bar charts, line charts, pie charts, etc., which can help users quickly understand data.

3、 Calculation and distribution

Apache Spark is a fast and versatile big data processing framework. Capable of processing large-scale data in memory, faster than traditional Hadoop MapReduce. Supports multiple computing modes such as batch processing, stream processing, and machine learning. It can be integrated with various data sources and storage systems for easy reading and writing of data.

4、 Data Warehouse

Amazon Redshift is a fully hosted cloud data warehouse service. Capable of quickly processing large-scale data, supporting SQL queries and analysis. It has high scalability and performance, and users can adjust storage and computing resources according to their own needs. Well integrated with other AWS services, convenient for importing and exporting data.

MEDIUM TASKS

(a) It makes sense that data is referred to as the 'oil of the 21st century'. On the one hand, data is as valuable as oil. By analyzing data, enterprises can understand consumer needs and market trends, formulate precise marketing strategies, and optimize product services. For example, e-commerce platforms recommend personalized products based on user data. On the other hand, data, like oil, needs to be refined and processed. The raw data is messy and needs to be cleaned, integrated, analyzed, etc. to become useful information. Just like how oil needs to be processed into products such as fuel. In addition, data acquisition and storage have similarities with oil and need to be collected, stored in a database, and managed. At the same time, we should attach importance to data security and privacy protection. But data is also different from oil, as oil is limited and data can continuously accumulate. And the value of data changes over time, and ownership and usage rights are also complex.

(b)

1.The uncertainty in data veracity primarily refers to the instability of analysis results

due to factors such as data inconsistency, incompleteness, ambiguity, latency, deception, and model approximation. This uncertainty may stem from errors in data collection, damage during data storage, biases in data processing, or the use of inaccurate models or algorithms. It poses a challenge to the accuracy and reliability of data analysis because even minor data errors can be magnified in large-scale data analysis, thereby affecting the quality of decision-making. Therefore, reducing data uncertainty and improving data veracity is crucial for ensuring the effectiveness and credibility of data analysis results.

2. Accuracy: Data accuracy means a high degree of conformity with the actual situation. Accurate data is the foundation for accurate evaluation and decision-making.

Reliability: The reliability of data is determined by reliable sources and rigorous processing. Trustworthy data is easier to accept and use.

Reputation: Data reputation is linked to provider reputation. A data source with a good reputation usually has high data quality.

Objectivity: Data objectivity requires not being influenced by subjectivity. Objective data can provide a fair basis for decision-making.

Authenticity: The authenticity of data reflects the actual situation without any falsehood. Real data is the foundation of effective analysis.

Consistency: Data consistency refers to maintaining consistency in different areas. Consistent data avoids confusion and misunderstanding.

Without bias: Data without bias ensures fair decision-making. Unbiased data ensures fair decision-making.

Correctness: Correctness is similar to accuracy, emphasizing correctness and accuracy. Correct data to avoid incorrect judgments.

Clarity: Data clarity should be clear and easy to understand. Clarify data to reduce ambiguity and improve efficiency.

ADVANCED TASKS

dataset: Online Retail

This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.

The data set contains:

InvoiceNo,StockCode,Description,Quantity,InvoiceDate,CustomerID,UnitPrice and Country. All of that were not missing value.

This data set can be used to build data-intensive systems about the transactions of goods between merchants and buyers. Of course, the object of this data-intensive system adds some additional data on the basis of the data set, such as: data analysis of buyers' purchasing preferences, the relationship between climate and commodity sales, etc., to better facilitate the delivery of goods.

Through this data-intensive system, merchants can determine the quantity of goods sold and adjust the inventory in the warehouse, reducing the loss of goods due to overstocking and other problems. Second, the data-intensive system can be used to estimate the number of foreign visitors in the area, giving merchants control over what they buy.