Student: Peng simin 2023020207

Chen yinuo2023020101

BASIC TASKS

a) Using examples, explain what is meant by the term big data.

The application of Taobao Data Rubik's Cube. Taobao collects and analyzes massive data on users' shopping behaviors, including browsing records, purchase history, search keywords, etc., to optimize the product recommendation system and advertising strategy. This not only improves the user's shopping experience, but also increases the platform's sales and advertising revenue

b) How is big data analysis helpful in increasing business revenue?

Innovation and competitiveness

Big data has become a key competitive foundation

Drive productivity growth, innovation, and consumer surplus

Customer insights and decision support

Gain customer insights through big data analytics

Decision support system based on big data

Market value extraction

The impact of the enterprise data explosion on an organization's computing technology and architecture

Big data technology helps organizations make decisions more effectively

Business growth and innovation

Leverage big data to support business growth and innovation

Enhance traditional analytics frameworks with big data analytics

Implementation Strategies vs. Implementation Challenges

IT works directly with marketing to ensure data access

The challenges of processing and analyzing large data sets

c) Describe the difference between structured and unstructured data.

definition

Structured data: There are predefined data models

Unstructured data: There's no predefined data model and isn't suitable for relational databases

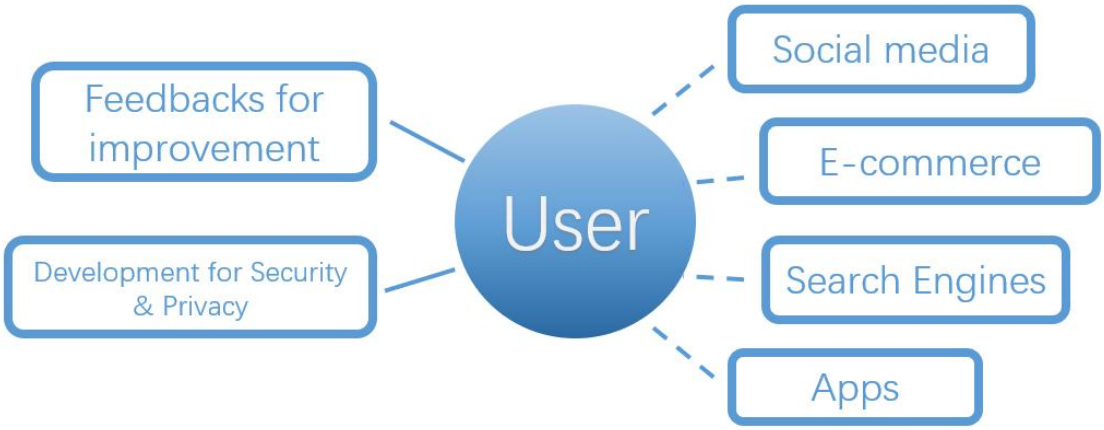
peculiarity

Structured data: Easy to store and query. It is commonly used in traditional RDBMS

Unstructured data: Lots of sources like social media, sensor data, etc. Special technical handling is required for effective analysis.

When it comes to processing and analysis, structured data can be efficiently queried and analyzed using traditional database query languages such as SQL. Unstructured data requires specific technologies and tools to be processed, such as natural language processing (NLP), image recognition, and video analytics.

d) Create a diagram that represents the big data that you contribute too



e) Why do we need data-intensive systems? Give a list of data-intensive systems.

People's life is closer to modern information technology, data has become an important driving force for enterprise and social development, and data intensive systems have emerged in this context, becoming an important component of modern information technology.

Relation: MYSQL ,Oracle
NoSQL: MongoDB ,Redis ,Cassardra

f) Briefly describe examples of data-intensive technology that can be used for data storage.
data visualisation and analysis, compute and distributions and data warehouses.

Storage: it usually involves distributed file systems (such as Hadoop Distributed File System, HDFS)which distributes data storage across multiple physical nodes in the network, while providing a logically unified file system view to users.

Visualization: (applications such as Tableau, Power BI, and D3.js) convert processed data into charts, dashboards, and interactive reports, helping users understand data analysis results more intuitively.

Compute and distributions: including data mining algorithms such as clustering analysis, classification analysis, association rule mining, and temporal analysis, as well as machine learning techniques such as classification, clustering, regression, etc., are used to extract valuable information and insights from large amounts of data

MEDIUM TASKS

a) Data has been coined "The oil of the 21st century". Discuss

This metaphor emphasizes the core role and value of data in the modern

economy. Data, like oil, is a fundamental resource that can drive economic growth, innovation, and decision-making

b) Discuss the following definitions related 'veracity':

1. Veracity is defined as "uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations
2. Other groups refer to issues of the quality of the data and include "accuracy, believability, reputation, objectivity, factuality, consistency and freedom from bias, correctness, and unambiguousness."

Regarding the authenticity of the data, it usually involves the accuracy and completeness of the data. The authenticity of the data refers to the consistency of the data with the actual situation, that is, the data reflects the true state of the object it describes. For example, the authenticity of real-world data is a basic requirement for data quality, which involves the characteristics of data from a wide range of sources, large amounts of data, and rich content. In addition, the authenticity of information on the Internet is also an important issue, as inconsistencies and incompleteness of information on the Internet can lead to poor decision-making. When it comes to data quality, it includes multiple dimensions such as accuracy, trustworthiness, reputation, objectivity, factfulness, consistency, and unbiasedness. Together, these dimensions constitute a comprehensive evaluation system of data quality. For example, the quality of statistics includes not only their truthfulness, but also their accuracy, completeness, etc.

ADVANCED TASKS

a) Find a big data set that might be useful for creating a data-intensive system. You might look at finding a dataset that could be used for your interest.

Based on a high-performance unified data management base, the Xinghuo cognitive model released by iFLYTEK effectively improves the efficiency of gene sequencing and meets the capacity requirements of data-intensive gene sequencing scenarios

b) Explain the data set. Indicate why you feel that this data is suitable for data-intensive systems.

Quantity of a type of plant in arboretum is massive. Besides we have four types of attributes from three types of plants here to compare. The data is very likely to be difficult to be handled by normal computing tools.