Big Data Learning Report

Lab1

Big data is crucial in this day and age. "Big data" is a collection of data that is large and complex. It has the characteristics of "4V", i.e., massive, varied, high-speed, and questionable authenticity.

For example, the content generated by social media reflects the abundance and diversity of big data, including text, images, videos, etc. Structured data, such as enterprise sales orders, has a fixed format for easy query and analysis; Unstructured data, such as images, requires special technical processing.

Big data analytics can increase revenue, precision marketing, optimize operations, and innovate products and services. We need data-intensive systems to address big data challenges, such as distributed file systems, databases, and data warehouses. Data has been called the "oil of the 21st century" because it is a valuable resource with great value potential. In short, big data and related systems are of great significance to the development of modern society.

Lab2# Big Data Database Type Analysis Report

This report provides an in-depth analysis of the types of big data databases, clarifies their application characteristics through the study of real cases, and compares the differences between relational and non-relational databases, so as to provide a reference for the selection of databases for related projects.

In the real-world database example, in the online video streaming scenario, the video content data (semi-structured) is suitable for management in a document-based database (such as MongoDB), and the user viewing record (structured) relies on a relational database (such as MySQL) to ensure consistency and accurate analysis. In e-commerce scenarios, relational databases (such as PostgreSQL) are more suitable for structured product, customer, and order information due to their complex relationships and high integrity requirements.

Relational databases are based on the relational model and tablestore, which have the advantages of strong consistency, complex query capabilities and good specification, and are suitable for enterprise-level applications (ERP, CRM), etc., but the scalability is

limited, the processing of unstructured data is inflexible, and complex transactions affect the performance. The non-relational database data model is flexible, diverse, highly scalable, and has good performance in specific scenarios, such as big data processing (log analysis, data mining) and social networking (user relationship management), but the data consistency is weak, and the query and transaction support is not mature enough.

In summary, when selecting a database, it is necessary to comprehensively consider factors such as data characteristics, business requirements, performance requirements, and costs, weigh the advantages and disadvantages of different types of databases, and make the most appropriate decision.

Lab3

Summary report on the application of big data in the campus security system

This report comprehensively summarizes the application of big data technology in campus security systems.

In terms of middleware functions, it undertakes data collection and pre-processing tasks, collects data from multiple types of security equipment, cleanses, converts formats and extracts features; At the same time, it ensures equipment communication and collaboration, ensures real-time and reliable information interaction, and supports collaboration between robots and with the monitoring center; It also provides intelligent decision-making assistance, which provides a basis for security strategy formulation based on data and algorithms.

In terms of security scenario and database selection, in the personnel intrusion detection scenario, the document database is used for video surveillance data, and the relational database is selected for access control and geographic information data. In the fire alarm scenario, the smoke sensor data is suitable for the time series database, and the fire protection equipment and building structure data are used

for the relational database. In the early warning scenario of campus violence, the social media data is a text-based database, the user behavior data is a graphical database, and the student file data is a relational database, all of which are stored and processed efficiently according to the data characteristics and scenario requirements.

In the application of communication and artificial intelligence technology, Starlink satellite communication ensures the stability of communication across the campus, especially in remote or weak network areas to ensure secure data transmission. Azure machine learning is used for abnormal behavior identification and fire risk prediction, and analyzes data through machine learning algorithms to achieve intelligent early warning and risk assessment, and improve the intelligence level of security systems.

In summary, big data technology is deeply integrated in all aspects of the campus security system, effectively enhancing the campus security and security capabilities.

Big Data Experiment IV Report

This report focuses on the experimental tasks in the fourth week of the big data course.

First, the basic tasks

1. File Replication Decisions: The advantages are improved availability and read performance; The disadvantages are that storage resources are consumed and data consistency maintenance is complicated.
2. Remote file processing: The disadvantage is that the network transmission overhead is large, and the improvement method is to transfer the results after processing by the remote node, which also involves the concept of data locality.
3. Matrix multiplication: Multi-core systems can be divided by rows or columns, and distributed systems can be distributed with storage nodes for collaborative computing.
4. Parallel execution of mathematical formulas: Some sub-expressions can be parallel, but they are limited by data dependence and resource competition.
5. Distributed Maximum Find: Split files to find local maximum numbers for multiple nodes, and then summarize them. Parallel

processing with mathematical formulas differs in data characteristics and parallelism.

2. Medium tasks

1. In HDFS, NameNode manages metadata, and DataNode stores data, and the two work together.

2. 180MB 文件按 64MB 块分割为 3 块,HDFS 默认 3 副本,通过 NameNode 分配存储,节点故障时 NameNode 会重新复制副本。

3. Advanced tasks

1. MapReduce 中 JobTracker 管理作业，TaskTracker 执行任务并汇报。

2. The Map function preprocesses the data as key-value pairs, and the Reduce function merges the values of the same key. Logic can be applied in color counts, device order statistics, and streaming billing.

Through this experiment, we have a deep understanding of the principles and application scenarios of big data distributed technology, which provides a foundation for subsequent learning and practice.

Lab5、

Research report on MongoDB and big data

I. MongoDB handles document relationships and characteristics

1. **Document Relationship Handling**

- MongoDB handles document relationships through embedding and referencing. Embedding is suitable for scenarios where child documents are closely associated with parent documents, which can reduce the number of queries. References are suitable for complex relationships or large data volumes, avoiding data redundancy, but may require multiple queries.

2. **Comparison with relational databases and reasons for "no mode" and "dynamic mode"**

- Relational databases need to create tables with a clearly defined structure that must be followed for inserting data. MongoDB, on the other hand, does not need to define a strict structure in advance when creating a collection, and has the flexibility to insert documents, which makes it considered "schemaless". The documents in its collection can have different fields, and fields can be added or removed at any time according to needs, reflecting the characteristics of "dynamic mode".

二、SQL-MongoDB 语句转换及查询

1. **Statement Transformation Example**

- For example, the SQL statement that creates an employee table and inserts employee data is converted to the insertMany statement of MongoDB. Convert various SQL statements that query employee information into MongoDB 'find' and related operation statements, including conditional query, sorting, and counting.

2. **Query statement function implementation**

- Implemented the query function of multiple conditions, such as querying the publication year by range, filtering books by specific criteria (such as author, ISBN, and title containing specific strings, etc.), counting the quantity, sorting, and updating fields.

3. Big data related concepts and MongoDB processing methods

1. **Nature and Diversity of Big Data**

- Big data has characteristics such as volume, velocity, variety, value, and veracity. The diversity is reflected in the abundance of data types, including structured, semi-structured, and unstructured data.

2. **MongoDB Handles Data Diversity**

- MongoDB can store structured data, similar to traditional databases. Convenient storage of semi-structured data in JSON format; Store unstructured data (e.g., binary data) in BSON format. With a flexible document model and embedded referencing, multiple types of data can be processed in a single database, simplifying the data architecture.

Lab6

Data visualization learning report

I learned a lot from this data visualization study.

For Microsoft Power BI, I learned through a learning path to learn more about its capabilities, successfully connect to data sources, and create multiple reports. When analyzing sales data, I use its interactive function to dynamically see sales changes over time, which greatly improves my insight into business data and makes me feel its strong ability to support business decisions.

When using Anaconda Spyder, I was able to follow the installation guide and get familiar with the interface. After the Stroke dataset was imported, a variety of visualization operations were performed. Create histograms to analyze the distribution of numeric fields, such as the number of people in a specific interval of average blood glucose levels; Histograms are used to show the differences in categorical data; Although the scatter plot did not find a strong correlation between age and BMI, exploring the relationship between other variables gave me a deeper understanding of the data. The pie chart is optimized to clearly show the proportion of work types.

Learning the basics of data visualization helped me understand how to choose the right chart type. Anaconda Spyder is flexible and suitable for in-depth analysis, and Microsoft Power BI is easy to use and can quickly create reports.

In the Spotify dataset visualization, I draw a line chart to show the changes in the number of songs played. The Financial Data Analysis Workshop has enhanced my data processing skills in the financial field.

This study not only allowed me to master visualization skills, but also cultivated data analysis thinking, and I will continue to work hard to improve my ability in this field in the future and contribute more to data processing and decision-making.