BD1-7

1

Hello everyone, I am Haiying Che, from Institute of Data Science and knowledge Engineering，

School of Computer Science, Beijing Institute of Technology,

in this session, we will discuss big data general architecture.

2

When the volume of data get bigger and bigger, any single traditional high performance server can not satisfy the requirement, more servers are needed. But the traditional high-performance server is very expensive, normally Hundreds of millions of yuan, because it is very reliable, like IBM Z series server, which can reach the reliability of 99.9999%, Known as zero downtime, so it is called Z series. High performance and high reliability means high price. But no company can afford such expensive server to deal with the big data issues. So, the distributed computing was used to address this issue.

The logic of distributed computing is use a group of cheap servers working as a team to replace the expensive high performance servers, which could be distributed processing, distributed storage. And use high redundancy to achieve high performance.

Key hardware and software breakthroughs revolutionized the data management industry.

First, innovation and demand increased the power and decreased the price of hardware.

New software emerged that understood how to take advantage of this hardware by automating processes like load balancing and optimization across a huge cluster of nodes.

The software included built-in rules that understood that certain workloads required a certain performance level.

The software treated all the nodes as though they were simply one big pool of computing, storage, and networking assets, and moved processes to another node without interruption if a node failed, using the technology of virtualization.

3

In the scope of big data, there are 2 main components, distributed storage and distributed processing.

The data need to be stored, then processed, and the intermediate and results also need to be stored.

These two components complement each other and closely work together.

4

Generally the big data architecture includes 3 layers, from bottom to top,

1 data storage system includes distributed database/Datawarehouse, which store the structured data in relational database, distributed file system, which store the unstructured data, and also the data collection and modeling, which responsible for collecting data according to the predefined model.

2 Data processing system includes calculation engine, Computing platform, calculation model and algorithm. Based on the stored data, we need design the calculation model, algorithms, also we need big data distributed calculation engine like MapReduce, and computing platform, like Hadoop, spark, Storm etc. to facilitate the data processing.

3 Based on the data storage and data processing results, the customized big data application can be built, which includes big data applications, data products and services, and also the data visualization.

With the help of processing system, all the needed data has already been calculated and modeled, the upper level application can present the data visualized results or give suggestions, or make smart decisions.

5

In the storage structure: the database provides the logical storage structure of the data;

the distributed file system provides the physical storage structure of the data.

Data storage system could include several parts.

1. **Data collection layer:**

**System logs, Web Crawler, wireless sensor network, internet of things and all kinds of data resources, from which we can collect data.**

**2）Data cleaning, extraction and MODELING**

**Then we can convert various types of structured, unstructured, and heterogeneous data from different sources into standard storage format data, and define data attributes and value ranges to prepare for the data analysis.**

**3）Data storage architecture;**

**Centralized or distributed file system, relational database or distributed database, row based storage data structure or column based storage data structure, key-value pair structure, hash table retrieval, etc. data scientist can choose the suitable data storage architecture to store the data, making the data storing and data retrieval more convenient.**

**4）****Unified Data Access Interface, etc.**

The application 's access to the database and data exchange is an important issue in distributed computing systems.

The industry earlier used the database application programming access interface provided by Microsoft, ODBC-Open DataBase Connectivity, which uses the X/open and ISO/IEC call interface (CLI call-level interface) standards as the basis, and uses a structured query language-SQL as the database access language.

ODBC is essentially a set of database access APIs, consisting of a set of function calls, and the core is SQL statements. When an ODBC-based application operates on the database, the user directly transmits SQL statements to ODBC. At the same time, ODBC does not rely on any DBMS for database operations, and does not directly deal with DBMS， All database operations are handled by the corresponding ODBC driver.

6

Computing models for different types of data,

such as the **MapReduce Batch Processing Model** for massive data,

the **Stream Computing model** for dynamic data streams,

the **Large-scale concurrent processing (MPP-Massively Parallel Processing) model** for structured data,

and the **large-scale physical memory In-memory Computing model**;

**Data Flow Graph model** for machine learning algorithms;

Implementation of various analysis algorithms, and computing platforms that provide various development kits and operating environments, such as **Hadoop, Spark , Storm**, etc.

**The computing engine is** a server-side program designed and encapsulated for a specific computing model based on a computing platform.

It is used to support the back-end big data processing, computing and tasks analysis in a specific computing mode.

For example, the MapReduce computing engine provides big data division and node allocation, job scheduling and calculation result integration, directly supporting the development of upper-level applications.

Google’s interactive computing engine uses Dremel and PowerDrill technologies to provide rapid calculation and analysis of large-scale data sets;

The open source **Apache drill** project is based on column storage structure, data localization, In-memory storage and other technologies to achieve large-scale data Quick query access.

The graph parallel computing engine provides efficient computing and processing of network graph data (social networks, telecommunications networks, and brain function connection networks are often represented by weighted directed graphs)

(20% of the data processed by the Google search engine is enabled by a graph computing engine). This technology includes Google’s Pregel, open source technology Hama, GraphLab, etc.

S4 (Simple, Scalable Streaming System) is Yahoo! A streaming computing engine is provided. The initial goal is to improve the click-through rate of cost-per-click advertisements, and predict the user's possible click behavior on advertisements through real-time data calculations

7

Based on the above-mentioned computing architecture and processing platform, big data application technology solutions in various industries and fields were provided.

At present, the Internet, e-commerce, e-government, finance, telecommunications, medical and health industries are the most popular areas for big data applications,

while manufacturing, education, energy, environmental protection, and smart transportation are the areas that big data technology will or have begun to expand.

8

After go through the general architecture of the big data, let’s look at the big data layer in another logical perspective.

In the bottom layer, the network data was collected, probe data was collected, service data was collected, business data was collected, terminal data was collected.

In the middle layer, user behavior was analyzed, Key indicators was monitored and other data was integrated and analyzed using the models and algorithms. And the processed data was provided to the upper level data services through unified data platform.

In the top layer, based on the analyzed data provided through the unified data platform, user can do different kinds of analysis, business process analysis, user analysis, data analysis and data output visualization.

9

In this session, we discussed the big data architecture 3 layers, **Data storage system;** **Data processing system;** **Data application system;** **Thank you for your attention, if you have any question, feel free to contact me.**