BDA 4-1

1

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

School of Computer Science, in Beijing Institute of Technology, from this session,

we start to learn Data storing system, and in this session, we discuss about data modeling.

2

The big data computing system can be summarized into three categories: from bottom to top, they are

Data storing system, Data processing system, Data application system.

In data storing system we collect the data and store the data, then support upper level data processing,

after data has been well processed, it can support the big data application. Like data analysis, business intelligent, decision making and so on.

The data storage architecture is the foundation of big data computing. The upper-level analysis algorithms,

computing models and computing performance all depend on the performance of the data storage system.

Therefore, the data storage system is an important area of big data research.

In the data storing system, there are 4 parts to accomplish the different tasks, which are data collection and data modeling, File system, database, data warehouse and Unified data access interface.

Data collection collect data from multiple data resources, like **System logs, Web Crawler, wireless sensor network, internet of things and all kinds of data resources.**

**And after data has been collected, we need clean the data，delete dirty data like repeated data, empty data, wrong data etc.**

**Convert various types of structured, unstructured, and heterogeneous data into standard storage format data, and define data attributes and value ranges.**

**And We build the data model to organize the data.**

**Above the data collection, it is file system, which physically implement the data storing, it could be Centralized or distributed file system,**

**then database, which design the logical structure to store the data, include RDBMs and No SQL database.**

**All above file system and database is the unified data access interface, which is a data access interface, through which the data processing system can retrieve the data from data storing system.**

3

The data can be collected from CRM, ERP, financials, social media, exhaust data, logs, files.

(Data exhaust or exhaust data is the trail of data left by the activities of an Internet or other computer system users during their online activity, behavior, and transactions.

Due to the characteristics of multiple data sources, data heterogeneity, unstructured data, distributed computing environment and others, which make the design of big data storage system is more complicated than before.

The current big data storage architecture is mainly composed of data layer, distributed file system/non-relational database (NoSQL), and unified database reading interface.

Some designs will add a data mining and analysis function on top of NoSQL database. Warehouse layer

4

The data layer mainly includes the data acquisition system and provides data extraction, cleaning, and conversion ETL-Extract transform Load, data modeling functions,

1）multiple data sources include enterprise data, business data, personal social data, government statistics, Internet data, Internet of Things data, System log data, Gene sequencing data, atmospheric physics monitoring data, earth satellite observation data, etc.

2） heterogeneous data could be text, pictures, audio and video etc.

3） The unstructured data includes medical imaging data, scans of bank certificates, fragmented communication records, screenshots, etc.

All these multiple data sources, heterogeneous data, The unstructured data, all these characteristics make it difficult to store raw data directly in the database.

5

In big data collecting, there are some Problems：

the original data format cannot be recognized and processed by the data platform.

In many cases, the original data still has problems such as missing records, missing value ranges, and different level data quality problems.

This requires cleaning the original data before building a database or data warehouse, by merging or removing duplicate data items, eliminating data errors,

Data could be Extracted from multiple data sources, extract different value ranges from data items to form the data structure of the target database,

Or data items could be Extracted from one data source and decompose them into multiple structures and load them into the target database,

Transform means Transforming original data items in different formats into a unified standard target database format.

Data extraction, cleaning and conversion can be done manually or using software tools.

6

**Now let’s look at data modeling, Data modeling is an important part of the data layer work.**

**Data modeling is to establish an abstract model of entity data, including metadata, data structure, attributes, value range, association relationship, and consistency, Timeliness and other elements.**

**The data model provides a reference for further data storage structure design, database design and calculation model.**

7

**Business model normally includes process model and data model, process model describes how the business works.**

**Data model describe the data supporting the business process.**

**The data model is defined in three levels conceptual model, logic model, and physical model.**

8

In this process diagram. First based on the process models and data requirement we can do the logical modeling, then generate the logical data model as the output.

Synthesize the logical data model, technical requirement and performance requirements, we carry out the physical data modeling and generate the output physical data model. After we have the logical data model and the corresponding physical model, the generated business data can be stored by creating, updating operations in the logical and physical model.

9

**Data modeling process**

**Let’s look at the three data model one by one.**

**Based on the user's data function requirements. functions and association relationships are obtained,**

**We can find Entity Class corresponding to the business elements and functions.**

[Conceptual model schema](https://en.wikipedia.org/wiki/Conceptual_schema) describes the semantics of a domain (the scope of the model).

This consists of entity classes, representing kinds of things of significance in the domain, and relationships assertions about associations between pairs of entity classes. Simply described, a conceptual schema is the first step in organizing the data requirements.

**In logical model design, more details of data entities, including primary keys, foreign keys, attributes, indexes, relationships, constraints, and even views, with data tables, data columns, value ranges, object-oriented classes, XML tags and other can be describe.**

**The physical model describes the storage implementation of data, including data partition, data table space, and data integration.**This is concerned with partitions, CPUs, [tablespaces](https://en.wikipedia.org/wiki/Tablespace), and the like.

10

According to **American National Standards Institute,** ANSI, the three perspectives should be relatively independent of each other.

Storage technology can change without affecting either the logical or the conceptual schema.

The table/column structure can change without (necessarily) affecting the conceptual schema.

but, the structures must remain consistent across all schemas of the same data model, logical or physical.

**UML is commonly used Data modeling language, and**

**Common data modeling tools includes Power designer, ER/studio.CA Erwin, IBM Infosphere Data Architect, etc.**

11

Let’s summarize, in this session, we learned the first layer of data storing system, which includes data acquisition, extraction, transforming and modeling.

12

thank you for your attention, if you have any question, feel free to contact me.