BDA5-1

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 processing system, and in this session, we discuss Data processing Architecture.

2

The data processing system provides big data computing and processing capabilities and an application development platform.

From the perspective of computing architecture, the data processing system is divided into data algorithm layer, computing model layer, computing platform layer, computing engine layer, etc.

Big data related computing algorithms include machine learning algorithms and data mining algorithms.

Computing models are the way that different kinds of big data is processed in different scenarios, which include batch processing, stream computing, **Large-scale concurrent processing (MPP) model** for structured data, In-memory Computing model, and Data Flow Graph models.

As far as the computing platform and engine, normally the typical representative big data processing platforms are Hadoop, Spark , storm, Pregel, etc,

3

Big data processing algorithm basically include data mining algorithms and Machine learning algorithms.

Data mining algorithms can be categorized into 4 groups, classification, clustering, correlation Analysis and anomaly detection.

Machine learning algorithms include supervised learning, unsupervised learning semi-supervised learning and reinforcement learning.

And in the scope of supervised learning there are regression, classification, and deep learning.

And in the scope of unsupervised learning there are clustering, dimensionality reduction ,and also anomaly detection.

For semi-supervised learning , it include self-training and low density separation models.

and reinforcement learning include dynamic programming and Monte Carlo methods.

4

Basic big data processing start from Batch processing, Represented by MapReduce.

Batch processing is the processing of a large volume of data all at once.

The data easily consists of millions of records for a day and can be stored in a variety of ways (file, record, etc).

The jobs are typically completed simultaneously in non-stop, sequential order.

A go-to example of a batch processing job is all of the transactions a financial firm might submit over the course of a week. Batching can also be used in:

* Payroll processes
* Line item invoices
* Supply chain and fulfillment

Batch data processing is an extremely efficient way to process large amounts of data that is collected over a period. It also helps to reduce the operational costs that businesses might spend on labor as it doesn’t require specialized data entry clerks to support its functioning. It can be used offline and gives managers complete control as to when to start the processing, whether it be overnight or at the end of a week or pay period.

Stream processing is the process of being able to almost instantaneously analyze data that is streaming from one device to another.

This method of continuous computation happens as data flows through the system with no compulsory time limitations on the output.

With the almost instant flow, systems do not require large amounts of data to be stored.

Stream processing is highly beneficial if the events you wish to track are happening frequently and close together in time.

It is also best to utilize if the event needs to be detected right away and responded to quickly Stream processing, then, is useful for tasks like fraud detection and [cybersecurity](https://www.bmc.com/blogs/cybersecurity/).

If transaction data is stream-processed, fraudulent transactions can be identified and stopped before they are even complete.

5

Massively parallel processing (MPP) is a storage structure designed to handle the coordinated processing of program operations by multiple processors. This coordinated processing can work on different parts of a program, with each processor using its own operating system and memory. This allows MPP databases to handle massive amounts of data and provide much faster analytics based on large datasets.

MPP processors can have up to 200 or more processors working on application and most commonly communicate using a messaging interface. MPP works by allowing messages to be sent between processes through an “interconnect” arrangement of data paths.

6

In-memory computation (or in-memory computing) is the technique of running computer calculations entirely in computer memory (e.g., in RAM). This term typically implies large-scale, complex calculations which require specialized systems software to run the calculations on computers working together in a cluster. As a cluster, the computers pool together their RAM so the calculation is essentially run across computers and leverages the collective RAM space of all the computers together.

In-memory computation works by eliminating all slow data accesses and relying exclusively on data stored in RAM.

Overall computation performance is greatly improved by removing the latency commonly seen when accessing hard disk drives or SSDs.

Software running on one or more computers manages the computation as well as the data in memory, and in the case of multiple computers, the software divides the computation into smaller tasks which are distributed out to each computer to run in parallel.

7

Massively Parallel Graph Computation

Graphs are useful theoretical representations of the connections between groups of entities, and have been used for a variety of purposes in data science, from ranking web pages by popularity and mapping out social networks, to assisting with navigation.

In many cases, such applications require the processing of graphs containing hundreds of billions of edges, which is far too large to be processed on a single consumer-grade machine.

A typical approach to scaling graph algorithms is to run in a distributed setting, i.e., to partition the data (and the algorithm) among multiple computers to perform the computation in parallel.

While this approach allows one to process graphs with trillions of edges, it also introduces new challenges. Namely, because each computer only sees a small piece of the input graph at a time, one needs to handle inter-machine communication and design algorithms that can be split across multiple computers.

8

Applying the Data Algorithm Layer algorithms and combined with the Batch, stream , Massive Parallel Processing, In Memory Computing

Or Graph Computing model, we can process big data problems, but it is hard for us to implement all the algorithm and computing model by ourselves,

The Computing Platform & Computing Engine Layer can provide us the platform and engine which includes the needed tools, libraries to facilitate the implementing the complex Algorithms and Computing models.

9

**Computing platform and engine refer to a development integrated environment that provides Technical Standards, Computing Architecture, and a series of Development Technologies And Tools for big data computing and analysis.**

**The current representative computing platforms are: Hadoop, Cloudera, Spark, Storm, and Google's commercial platforms based on a series of big data computing technologies.**

10

In this session we learned the general architecture of the data processing system,

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