BDA5-4

1

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

School of Computer Science, in Beijing Institute of Technology,

in this session, we discuss Stream computing model.

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.

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.

Now let’s look at the stream computing model represented by Storm

3

Stream computing is a computing model that processes real-time dynamic data.

The traditional enterprise database stores historical data, that is, static data, that is, all data must be entered into the database before calculation and processing.

Technicians can query and update the database, and use data mining and OLAP analysis tools to extract static data from the database. Find valuable information to support business decision analysis.

However, in Internet applications (user web click tracking, online real-time recommendation systems, etc.), intelligent transportation systems, wireless sensor network monitoring and other fields, its data generation methods and data characteristics have the following characteristics and calculation requirements

1）Data is no longer arriving in batches but continuously arriving dynamically

2）Computational analysis requires real-time, fast response, and low latency

3）The amount of data is large, but the storage of the data is not valued, but the immediate processing and analysis of the data are emphasized

4）Pay attention to the calculation and analysis results of the data as a whole, but not to pay attention to the individual data

5）The order and timing of the arrival of data elements cannot be predicted or controlled, and the calculation program must be able to respond

MapReduce performs offline batch calculations for static data that has entered the database, and the calculation results are also stored in the static database;

while stream computing is real-time analysis and calculation for dynamic continuous data streams.

After the calculation results are obtained, the data is either imported into the static database, either discard, that is, one-time use.

To support this data flow calculation mode, the flow calculation framework generally includes three steps: real-time data collection, real-time data calculation, real-time data query service

4

Directed Acyclic Graph (DAG, Directed Acyclic Graph) is commonly used in distributed systems to characterize the calculation process or calculation model.

The figure shows the combination of chained tasks in a distributed system.

The nodes in different colors in the figure represent computing tasks (or computing objects) at different stages.

The one-way arrow indicates the order of the calculation steps and the dependencies.

5

Storm is a Native Stream Processing System, that is, the processing of stream data is based on each piece of data, and its parallel calculation is implemented based on a directed topology graph.

Topology composed of Spout (data source) and Bolt (processing unit).

Topology Defines the logical model (or abstract model) of parallel computing, that is, designs the calculation steps and processes from the perspective of function and architecture.

The out put of a spout is a series of tuples stream, and the input and output of Bolt is also is a series of tuples stream.

6

Storm's computing system also adopts a master-slave (Master/Slave) architecture. There are mainly two types of nodes: master node and slave node.

A Nimbus daemon runs on the master node, like Hadoop's JobTracker, responsible for task distribution and fault monitoring of the cluster.

Nimbus manages many worker nodes through a group of Zookeeper

Each worker node runs a Supervisor daemon, monitors the status of the local node, and starts and shuts down the worker process of the node when necessary according to Nimbus instructions.

7

**There are two ways of stream computing,** Native Stream Processing System represented by storm and Micro-batch Stream Processing System represented by spark

**Micro-batch processing** is the practice of collecting data in small groups (“batches”) for the purposes of taking action on (processing) that data.

Compared with Native Stream Processing System, **Micro-batch processing** is not really Realtime processing , but nearly Realtime processing , it can save the computing cost because it wait until the micro batch composed.

Contrast this to traditional “batch processing,” which often implies taking action on a large group of data.

Micro-batch processing is a variant of traditional batch processing in that the data processing occurs more frequently so that smaller groups of new data are processed. In both micro-batch processing and traditional batch processing, data is collected based on a predetermined threshold or frequency before any processing occurs.

8

During the real application, sometimes the Batch computing and Stream computing combined to finish the history records analysis and Realtime analysis like showed in the diagram.

Kafka broker take the data form the data producer, and deliver them into Stream computing storm cluster and the Batch computing Hadoop cluster, the processing output of storm can be stored in NoSQL Cassandra DB The output of Batch processing can be stored in Hbase.

After the data has been processed, the data analytics like virtualization, decision making, Prediction, OLAP and recommendation can be done based on the result of Batch processing and Stream processing.

9

In this session we learned stream computing represented by Storm.

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