BD1-5

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 about Big Data Processing flow both from the analytical perspective and technical perspective.

2

All the big data analysis and processing is about to get value from data.

That is the ultimate goal of big data analysis. In order to achieve it, we need to do several steps, this diagram shows the bigdata process flow in analytical perspective.

1According to the analysis requirement to find the suitable data sources, which could be structured data, semi structured data or unstructured data.

2 Then collect data from the selected data sources, which could use the ETL(Extract Transform load) to extract data from different database or data warehouse or use data crawlers to crawl the data.

3 After get data from different data sources, sometime the data format is different, it need transform to facilitate the further data store or processing.

we need plan the suitable way to store the data, due to the nature of big data, data volume is big, which may need distributed data storage solution.

4 Then the data need to be processed to support analysis. In data Processing,

data scientist maybe first clean the data, delete the wrong data, redundant data, and other dirty data.

After that, data scientist need to do the feature engineering, to extract or build the representative key features which can enable the further model or algorithm.

Then data scientist could apply statistic, mathematic models or machine learning algorithms to find the pattern, correlation, classification and so on.

5 the pattern, correlations, classification and so on will be visualized, which can make the results easy to understand. For example, the pandemic data of the different provinces could explain the distribution better in the form of map than table.

6 also some reports could be generated to do the further analysis. Or the results could be used to monitor the business.

These are the Big data process flow in analytical perspective.

3

Now let’ s look at Big data process flow-technical perspective.

At the most left side, are the data assets, including transactions, OLTP, online transaction processing and OLAP online analyzing processing,

different kinds of documents, social media content and the IoT data generated by machine devices.

All these data need to be integrated into cluster or big data storage to support the further analysis.

1 transaction data in database or data warehouse need to be offload & ETL processed to Hadoop, which is mostly the first step of journey.

The transaction data is mostly structured data.

2 Some document data can also be used for data analysis, these data also need to be offloaded to Hadoop cluster for the further processing and analysis.

3 Social media data can help to understand customer’s opinion and preference, which enable better customized service .

4 With the help of the IoT device , we can collect IoT data from different kind of sensors, video cameras and so on , those data can help us

learn the status of the devices and the its content.

All these raw data can be integrated into cluster, and the metadata changed also need to update to cluster,

besides the real-time data need to be collected timely to support the further real-time analysis.

After the data collected in cluster, the data scientist will try to discover and profile the data according to the analysis purpose,

and use the defined meta data to govern and enrich the data set. Then parse and prepare data for next analysis.

When data scientist analysis the data, they should mask the sensitive data like name , id etc. to protect the privacy.

Based on the data in the cluster, data scientist can generate the business intelligence report or other visualization graphs.

Sometimes, in order to do the effective analysis, data scientist would like to build an Enterprise Data Warehouse and move the curated data to the

Enterprise Data warehouse, which just store the relevant data for the business intelligence. And on the basis of Enterprise Data Warehouse,

analytics can generate the business intelligence report to support decision making.

4

Let’s make an example, JD big data analysis.

JD has lots of transactions data from the online shopping history.

And they also have lots of document data, like customer email, industry reports, user agreements and so on.

And the customer’ s review about their feeling after shopping.

And they also have lots of IoT data, like the logistic trucks GPS, product package ‘s RFID etc.

With all these data integrated into the Hadoop cluster, JD data scientist can discover and profile about the customer, product, express and

with all the predefined model, metadata, they can parse the original raw data, and generate the JD sales trend BI report, express efficiency BI report, product category BI report and etc.

But of cause, customer sensitive data should be masked to protect customer privacy.

5

In this session, we discussed the Big Data Processing flow both from the analytical perspective and technical perspective. We understand the general steps of big data analysis, including selecting the data sources, data collecting, cleaning the data before storing the data.

Based on the data stored in the big data distributed storage, data scientist can process the data for further analysis, after all the data has been processed, data scientist can use algorithm or designed model to analyze the data, generate the business intelligence report or do the visualization analysis to dig the insight of the data and support decision making.

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