BD 1-4 Big Data lifecycle

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 Lifecycle and big data processing flow.

2

The effective use of Big Data with exponential growth in data types and data volumes has the potential to transform economies，useful business and marketing information and customer surplus. Big Data has become a key success mantra for current competitive markets for existing companies, and a game changer for new companies in the competition. This all can be proven true if VALUE FROM DATA is leveraged

Let’s look at the stages of digging the value in bigdata lifecycle .

3

As this figure explains, the Big Data life cycle can be divided into four stages. Let's study them in detail.

1. Collecting data is key in a Big Data life cycle; it defines which type of data is captured at the source. Some examples are gathering logs from the server, fetching user profiles, crawling reviews of organizations for sentiment analysis, and order information.
2. Examples that we have mentioned might involve dealing with local language, text, unstructured data, and images, which will be taken care of as we move forward in the Big Data life cycle.

With an increased level of automating data collection streams, organizations that have been classically spending a lot of effort on gathering structured data to analyze and estimate key success data points for business are changing. Mature organizations now use data that was generally ignored because of either its size or format, which, in Big Data terminology, is often referred to as unstructured data. These organizations always try to use the maximum amount of information whether it is structured or unstructured, as for them, data is value.

You can use data to be transferred and consolidated into Big Data platform like HDFS (Hadoop Distributed File System). Once data is processed with the help of tools like Apache Spark, you can load it back to the MySQL database, which can help you populate relevant data to show which MySQL consists.

1. Storing data that has been collected from various sources. Let's consider an example of crawling reviews of organizations for sentiment analysis, where in each gathers data from different sites with each of them having data uniquely displayed.

Traditionally, data was processed using the ETL (Extract, Transform, and Load) procedure, which used to gather data from various sources, modify it according to the requirements, and upload it to the store for further processing or display. Tools that were every so often used for such scenarios were spreadsheets, relational databases, business intelligence tools, and so on, and sometimes manual effort was also a part of it.

The most common storage used in Big Data platform is HDFS. HDFS also provides HQL (Hive Query Language), which helps us do many analytical tasks that are traditionally done in business intelligence tools. A few other storage options that can be considered are Apache Spark, Redis, and MongoDB. Each storage option has their own way of working in the backend; however, most storage providers exposes SQL APIs which can be used to do further data analysis.

There might be a case where we need to gather real-time data and show case in real time, which practically doesn't need the data to be stored for future purposes and can run real-time analytics to produce results based on the requests.

1. Analyze

how these various data types are being analyzed with a common question starting with what if...?

The way organizations have evolved with data also has impacted new metadata standards, organizing it for initial detection and reprocessing for structural approaches to be matured on the value of data being created.

Most mature organizations reliably provide accessibility, superiority, and value across business units with a constant automated process of structuring metadata and outcomes to be processed for analysis.

A mature data-driven organization's analyzing engine generally works on multiple sources of data and data types, which also includes real-time data.

During the analysis phase, raw data is processed, for which MySQL has Map/Reduce jobs in Hadoop, to analyze and give the output. With MySQL data lying in HDFS, it can be accessed by the rest of the ecosystem of Big Data platform-related tools for further analysis.

4)Governance

Value for data cannot be expected for a business without an established governance policy in practice. In the absence of a mature data governance policy, businesses can experience misinterpreted information, which could ultimately cause unpredictable damages to the business. With the help of Big Data governance, an organization can achieve consistent, precise, and actionable awareness of data.

Data governance is all about managing data to meet compliance, privacy, regulatory, legal, and anything that is specifically obligatory as per business requirements.

For data governance, continuous monitoring, studying, revising, and optimizing the quality of the process should also respect data security needs. So far, data governance has been taken with ease where Big Data is concerned; however, with data growing rapidly and being used in various places, this has drawn attention to data governance. It is gradually becoming a must-considerable factor for any Big Data project.

4

The purpose of big data analysis is to extract knowledge and wisdom from the data, but all the beginning is raw data,

There are four layers in the DIKW pyramid model data，which are Information ，Knowledge ，Wisdom.

Data is in the bottom, which is individual facts, figures, signals, measurements,

And in the special context, we can draw information from the data, then reach the second level;

Information is organized, structured, categorized, useful, condensed, calculated data.

For example , we have the Taobao online shopping data, and form one customer’s shopping history, we can summarized the categories this customer preferred, which are his shopping interests.

Above the information, it is knowledge, which is idea, learning, notion, concept, synthesized, compared, thought-out, discussed. In the example, from the shopping interest, we can learn this customers characters, knowledge is the meaningful information.

According to the knowledge, we can get the insight, which is understanding, integration, applied, reflected upon, actionable, accumulated, principles, patterns, decision-making progress.

And from bottom to top, the decision risk is decreased, because from processed data we got the information, knowledge and wisdom, the uncertainty is reduced, we know more and more about the decision to make.

5

Harnessing Big data to get more business intelligence.

Stage1, at the very beginning, the hierarchical database invented in 1968 and the relational database invented in 1970, with these kinds of operational database, based on the data at rest, reporting and human analysis can be made on historical data. In this stage, it was mainly OLTP: Online Transaction Processing

Stage 2, with the help of data warehousing, people can analyze the current data to improve business transaction, because data warehousing can integrate all the different aspects of the organization comprehensive information, which can help understand the business and customer better, and provide better services. In this stage, it was mainly **OLAP:** Online Analytical Processing

Stage 3

With the rapid growth of data, the data is coming faster and need to be processed faster, to solve this in 2000, streaming computing was created, which can process the data fast when it comes, and support **RTAP:** Real-Time Analytics Processing to make the Realtime decision and improve Realtime business response.

6

With the data stored in the organization system, we can do the business intelligence, which is Ad-hoc querying and reporting

Using Data mining techniques, based on Structured data, typical sources in Small to mid-size datasets.

But with the advantage of new technology, more and more data was collected, we want to know more and predict something when it is not happen, in that case the complexity will increase and the business value will expand.

Based on All types of data from many sources, integrate them into very large datasets, Complex statistical analysis can be done, Optimizations and predictive analytics can be made, And More of a real-time manner, and vice versa, which promote people to gather more data. And all this will drive the big data progress.

For example, with the sensors equipped in all the parts of the machine, we can collect the status data and by using data mining, we can predict when the specific part will be expired or broken, and what part of the machine need to be repaired or replaced by analyzing the data perceived from the machine part sensors.

7

Let’s look at the evolution of business intelligence.

At 1990’s Business Intelligence Reporting OLAP & Data warehouse can help organize to generate the useful information to summarize the organization situation with the tools like Business Objects, SAS, Informatics, Cognos other SQL Reporting Tools.

In 2000, When the dataset gets big, the analysis takes longer time, we need speed the analysis, with the help of Interactive Business Intelligence & In-memory RDBMS like Tableau, HANA, analysis can be faster,

When the scale gets huge, the big data processing technology is needed.

Big Data: Batch Processing & Distributed Data Store Like Hadoop/Spark, HBase/Cassandra can handle these problems.

In 2010’s both the scale and speed are demanded, Big Data Real Time & Single View of all comprehensive data are needed, then call for more advanced technology to address these challenge.

8

In this session we learned about big data lifecycle, which includes Collect, Store, Analyze and Governance.

Among all these, analyzing is the purpose, we want to dig value from the raw material. We reviewed the traditional analysis and Big data analysis,

and nowadays we need fast analyze the huge amount of big data.

That is all for this session, **thank you for your attention, if you have any question, feel free to connect me.**