BD 1-1

1

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

School of Computer Science, in Beijing Institute of Technology, in the scope of big data, there are mainly two kinds of data which are structured data and unstructured data. in this session, We will learn what is structured data and what is unstructured data**, let’s learn about these two kinds of data**

2

According to Gartner, there is estimated 20% data of organization is structured data, the other majority is unstructured data.

Structured data is normally in the form of table with row and column, like the table at the upper-left corner, which is a table with attributes id, forename, surname and age in different column.

The structured data organized the data in predefined format, so it is easy to processed, because we know the data structure, we can design the corresponding algorithm to process it.

But for unstructured data, there is no uniform format for all the data, which makes the processing more difficult.

Let’ s talk about the characters of structured data first. structured data can be displayed in rows, columns and relational database. And the data type could be numbers, dates and strings. The structured data required less storage. and it is easier to manage and protect with legacy solutions. Because we have gained much experienced with dealing with structured data.

The Unstructured data cannot be displayed in rows, columns and relational database. They are normally images, audio, video, word processing files, e-mails, spreadsheets. They count for 80% of the enterprise data according to Gartner statistics. The Unstructured data require more storage because they are huge amount and are not well organized. And they are difficult to manage and protect with legacy solutions, in the traditional way.

3

In this diagram, data was divided into four groups, structured, semi-structured, quasi-structured and unstructured data.

From bottom to top,

1). the structured data have a defined data model, format, structure, for example the data organized in database.

2) This kind of data is somewhat structured but not completely.

This may seem to be unstructured at first and does not obey any formal structures of data models such as RDBMS. For example, NoSQL documents have keywords that are used to process the document.

semi-structured data, like Textual data files with apparent pattern, enabling analysis. For example, Spreadsheet and XML files. CSV files are also considered semi-structured data.

3) Textual data with erratic format that can be formatted with effort and software tools, for example clickstream data. After some processing, it can be formatted.

4) The unstructured data has no inherent structure and is usually stored in different type of file, like text document, PDFs, images and videos.

4

The structured is like the surfaced part of the iceberg, and the Unstructured data is like the underwater hidden part of Iceberg, which need to be explored.

The structured data like NPS（Net Promoter Score）口碑, CSAT-Client Satisfaction, and records in CRM system, sales system, Finance system and Excel.

Unstructured data is where life gets interesting. By definition, it lacks standardization and often has limited preset boundaries. It is the bits and pieces gathered from documents, social media, emails, audio/visual files, open-ended response fields, notes fields, and other forms of content that is not easily boxed and analyzed using any of our standard data analytics tools.

Its format may range from words or numbers to images and audio. Each time a customer, member, prospect, or stakeholder interacts with or mentions your organization or brand, they are generating unstructured data. Every organization has it, but many do not know where to begin when analyzing it.

5

After understand the structured data and the unstructured data, let ‘s look at the differences between traditional database and big data,

In this table, we compare the database with big data using metaphor, from aspects of scale, variety, schema, data and tool.

The traditional database is like Fishing in a pool and the big data is like fishing in the sea.

Database ‘s scale is MB, 2 to the power of 10 KB, 2 to the power of 20 Bytes,

But the scale of big data could be GB TB PB, 2 to the power of 30, 40 ,50 Bytes,

In the aspects of variety, database is like some kinds of fishes; which are different kinds of Structured data.

Big data is like thousands of kinds of fishes; Structured, semi-structured and non-structured, all of them.

Database has its own, predefined schema, so it is First schema, then data, but for big data, it is impossible to have predefined schema, so it is First data then try to find the suitable schema for the data.

In database Data is just the object to be processed, in big data we can Sense other data through some data.

as for the tool, in database we use one tool, traditional database management system can solve all the problems, in big data, no tool can solve all the problems, big data need different tools to solve the different aspects of the problem.

6

In this section we learned the bigdata concept and what is structured data and what is unstructured data, Thank you for your attention, if you have any question, feel free to connect me.