BD 1-3 **Big Data Characters**

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 will discuss about big data Characters, 5Vs, which are volume, velocity, variety, veracity and value.

2.

Originally the Characters of Big data is 3 Vs, which are volume, velocity, variety

As the information technology developed, the data recorded by the system or devices expanded dramatically.

Shown in the diagram.

At the very beginning, the ERP (Enterprise resource planning) system record the purchase detail, purchase record and payment record, which is mainly transactions, data scale is MB.

and then they expanded the content of customer, in CRM the segmentation, offer details, customer touches, supports contacts information have been recorded in system, which is transactions and interactions with the boundary of organization, data scale is GB.

Further with the help of Web, the web logs, offer history, A/B testing, dynamic Pricing, affiliated networks, search marketing, behavioral targeting, dynamic Funnel that useful information can also be recoded and stored. Which are transactions and interactions with the inside and outside of the organization, data scale is TB.

In the Big data scope, the sensors/RFID/Devices, user click stream, mobile web, sentiment analysis, user generated content, social interaction & feeds, spatial & GPS coordinates, external demographics, business data feeds, HD video, audio, images, speech to text, product/service logs, SMS/MMS and etc. the data scope is very comprehensive, which are transactions, interactions and observations, data scale is PB,

So, from this development, we can see, the data volume and variety, also the velocity is increased.

Let’ s goes through these characters one by one.

3 volume

1. Volume

Volume refers to the huge amounts of data that is collected and generated every second in large organizations.

data scale has increased 44 times from 2009 to 2020，From 0.8 zettabytes to 35 zettabytes， **The amount of data is growing exponentially**

This data is generated from different sources such as IoT devices, social media, videos, financial transactions, and customer logs.

Data volume can vary. For example, a text file is a few kilobytes whereas a video file is a few megabytes.

**IDC's "Data Era 2025" report, by 2025, the global data will reach 175 ZB (218.75 times in 2009)**

4 **The project Earth scope**

Earth scope is the largest scientific project in the world. The observatory aims to track the geological evolution of North America, recording more than 3.8 million square miles of data and accumulating 67 TB of data. It analyzes the seismic slip in the San Andreas fault, and of course, the magma plume below Yellowstone Park, and so on.

5 **Variety**

Another one of the most important Big Data characteristics is its variety. It refers to the different sources of data and their nature.

The sources of data have changed over the years. Earlier, it was only available in spreadsheets and databases.

Nowadays, data is present in photos, audio files, videos, text files, and PDFs.

Which includes

* Relational Data (Tables/Transaction/Legacy Data)
* Text Data (Web)
* Semi-structured Data (XML)
* Graph Data

Social Network, Semantic Web (RDF), …

* Streaming Data

That you can only scan the data once

* A single application can be generating/collecting many types of data
* Big Public Data (online, weather, finance, etc.)

In order to extract knowledge, all these types of data need to be linked together

The variety of data is crucial for its storage and analysis.

6

For example, if you want to analyze your customer, you want to know all different aspects of him.

Including his purchase preference, our known history in transaction system, banking, finance history, social media, gaming and entertainment and so on, with all these data you will know comprehensive information about your customer and provide better service.

7

This term refers to the speed at which the data is created or generated.

**Data is generated fast and need to be processed fast**

This speed of data producing is also related to how fast this data is going to be processed. This is because only after analysis and processing, the data can meet the demands of the clients/users.

Massive amounts of data are produced from sensors, social media sites, and application logs – and all of it is continuous. If the data flow is not continuous, there is no point in investing time or effort on it.

**Late decisions will lead to miss opportunities**

**For Example,**

**E-Promotions: Based on your current location, your purchase history to analyze what would you like and send promotions right now for store next to you, if the Promotion was sent when you already far from the store, it is no use.**

**And Healthcare monitoring, sensors monitoring your activities and body, when any abnormal measurements happens, it requires immediate reaction**

8 **Real-time/Fast Data**

A lot of fast data is real time.

**In Social media and networks** all of us are generating data.

**Scientific instruments are** collecting all sorts of data.

**Mobile devices are** tracking all objects all the time.

**Sensor technology and networks are** measuring all kinds of data.

People and devices are all generating data.

The progress and innovation is no longer hindered by the ability to **collect data**

But, by the ability to **manage, analyze, summarize, visualize, and discover knowledge** from the collected data in a timely manner and in a scalable fashion.

9 **Real-Time Analytics/Decision Requirement**

Real-Time Analytics is important, which can enable Realtime better decision, act when it is happening.

For example:

Product Recommendations that are Relevant & Compelling.

Learning why Customers Switch to competitors and their offers in time to Counter

Improving the Marketing Effectiveness of a Promotion while it is still in Play.

Preventing Fraud as it is Occurring & preventing more proactively.

10 **5V’s**

Let’s look at the 5Vs,

1 volume, data at rest, terabytes to exabytes of existing data to process,

2 velocity, data in motion, streaming data, milliseconds to seconds to respond.

3 variety, data in many forms, structured, unstructured, text, multimedia and so on

The other two are veracity and value.

4 **Veracity: data in Doubt**

**Uncertainty due to data inconsistency and incompleteness, ambiguities, latency, deception, model approximations.**

This feature of Big Data is connected to the previous one. It defines the degree of trustworthiness of the data.

As most of the data you encounter is unstructured, it is important to filter out the unnecessary information and use the rest for processing.

5 value:

Among the characteristics of Big Data, value is perhaps the most important. No matter how fast the data is produced or its amount, it has to be reliable and useful. Otherwise, the data is not good enough for processing or analysis. Research says that poor quality data can lead to almost a 20% loss in a company’s revenue. Data scientists first convert raw data into information. Then this data set is cleaned to retrieve the most useful data. Analysis and pattern identification is done on this data set. If the process is a success, the data can be considered to be valuable.

Even the valuable data, it is low value density, like panning for gold in the sand. A few seconds among a video of several house may be useful.

11 conclusion

Big Data is the driving force behind major sectors such as business, marketing, sales, analytics, and research. It has changed the business strategies of customer-based and product-based companies worldwide. Thus, all the Big Data characteristics have to be given equal importance when it comes to analysis and decision making. In this session we learned the Big Data characteristics, volume, velocity, variety, veracity and value, which are key factors you need to consider when you do big data analysis.

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