**BD 2-1**

**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 this session, we will discuss**

**Some idea about Data Resources.**

**2**

**According to organization boundary, we can divide the data into internal data and external data. By the way， the reason we want to divide the big data resources into internal and external is, the methods and tools of getting internal data and external data are different.**

**1 The internal data resources refer to all the data created by the organization and stored inside the organization,**

**Including organization self operated system, business transaction system, manufacturing system, finance system, human resources system, etc. these are mainly structured data.**

**Organization internal data also include some historical legacy data, such as some documents data, emails and etc.**

**Besides there are some internal IoT devices can generate some data, for example, there are lots of sensors in the manufacturing Automated production line, these sensors can collect many production status data, which can facilitate the manufacture analysis.**

**2 External data refer to the data outside the specific organization.**

**External data includes**

**2.1 other organization business operating platform data, like data from other organization’s business transaction system, manufacturing system, finance system, human resources system, accounting system etc.**

**2.2 other organization’s IoT devices generated data**

**2.3 the government collected and published data**

**2.4 internet or mobile internet data, published by all the internet nodes, like social media, blog data, YouTube data, Wikipedia data etc.**

**All these internal and external data could be the data resources for analysis.**

**3**

**1 Government data, various departments are set up for social management purposes, such as the Public Security Bureau, the Ministry of Finance, etc., these organizations will build many business systems, for the purpose of effectively fulfilling the functions of the department.**

**The data generated by these systems is mainly stored in the corresponding data center in a specific structure. The data contains great value inside, which can provide data support for the formulation of government macro policies, national security prevention and control, and effective social management.**

**Government data has the characteristics of high credibility, good integrity, strong real-time, and clear entity description.**

**2 Self-operated data refer to the IT system, ERP, SCM, online office system, online transaction system, etc. of different organization.**

**3 Incorporate IoT data collection into the consideration of data enrichment.**

**4**

**Inside the organization, all the business data was collected from different business systems, like CRM, ERP, marketing system, finance system, HR system, Supply chain system and other data.**

**All these data were extracted and stored in the data storage, which could be the relational database like SQL Server, data warehouse like Hive, or document database like Mongo DB, or graph database like neo4j,**

**and then process the data using data processing platform like Hadoop, spark etc.**

**And analysis the data with the tools like TensorFlow, R, IBM Watson etc.**

**After the data has been processed and analyzed, we can also visualize the data using visualization tools like tableau.**

**Through the big data analysis, we can get the business improvement, benefit analysis, revenue analysis, customer profiles, suitable pricing, some business patterns and so on, all these can help organization understand the customer and market better, provide better product and services, make better decision and improve their business strategy.**

**5**

**In addition to the internal data, there are also external data available for analysis.**

**External data is mainly internet data, whose channel includes web portal, government open information, social media, E-business public data and some special forum focusing specific topic.**

**6**

**Internet has huge amount of data, but when you collect data from internet you should be aware of some issues.**

**1） there are Different IT level and structure of different website, there is no unified collection method for all the websites.**

**2）Different websites have different control policy of crawlers, Why?**

**Because Web crawlers facilitate network information collection and query, but they also introduce the following negative impacts:**

**2.1 Web crawlers always consume too much server bandwidth and increase server load as they use specific policies to browser as much information as possible.**

**2.2 Bad actors may use web crawlers to launch DoS attacks against websites. As a result, websites may fail to provide normal services due to resource exhaustion.**

**2.3 Bad actors may use web crawlers to steal mission-critical data on your websites, which will damage your economic interests.**

**((Some WAF-web application firewall provides three anti-crawler policies,**

**1）robot detection by identifying User-Agent, 2）website anti-crawler by checking browser validity, and 3）CC attack protection by limiting the access frequency, to comprehensively mitigate crawler attacks against your websites.**

**2.1 So some website will enable Robot Detection to Identify User-Agent**

**If you enable robot detection, WAF can detect and block threats such as malicious crawlers, scanners, and web shells.**

**2.2 Enabling Anti-Crawler Protection to Verify Browser Validity**

**If you enable anti-crawler protection, WAF dynamically analyzes website service models and accurately identifies crawler behavior based on data risk control and bot identification approaches.**

**2.3 Configuring CC-Challenge Collapsar挑战黑洞 Attack Protection to Limit Access Frequency**

**A CC attack- 攻击是DDoS攻击的一种类型，使用代理服务器向受害服务器发送大量貌似合法的请求(Challenge the black hole attack is a type of DDoS attack that uses a proxy server to send a large number of seemingly legitimate requests to the victim server, protection rule uses a specific IP address, cookie, or referrer to limit the access to a specific path (URL), mitigating the impact of CC attacks on web services.))**

**3） the internet data could be form of text, tables, audio, and video, the different form increase the difficulty of collection. Because you need to design different ways to collect them.**

**4） Its authenticity and data quality are inferior to other data**

**Because anyone can publish data in the internet, nobody is responsible for the data authenticity and data quality, which make Its authenticity and data quality are inferior to other data, for example the internal data.**

**7**

**Let’s summarize the data channels**

**If we want to do big data analysis, we need to have enough data, the data can be collected from inside or outside the organization, which are internal and external data.**

**Internal data includes the business system data, archived data, like documents, emails etc. and organizational IoT data.**

**External data includes government public data, other organization data, internet data, and outside IoT data.**

**8**

**In this session we learned the bigdata resources, internal and external data.**

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