Chapter 3 Big Data Generation and Acquisition

Big Data Generation and Acquisition

Key technologies related to big data (already discussed) -

- Cloud computing
- IoT
- Data Center
- Hadoop

The value chain of big data, which can be generally divided into four phases:

- **Data generation**
- **Data acquisition**
- Data storage
- Data analysis

Big Data Generation and Acquisition

Big Data Generation

- Enterprise Data
- IoT Data
- Internet Data
- Bio-medical Data
- Data Generation from Other Fields: Computational biology-GenBank, Astronomy-Sloan Digital Sky Survey, High-energy Physics-Nuclear Research

Big Data Acquisition

- Data Collection
- Data Transportation
- Data Pre-processing - Integration, Cleaning, Redundancy Elimination

Big Data Generation

Big Data Generation

Data generation is the first step of big data.

Specifically, it is large-scale, highly diverse, and complex datasets generated through **longitudinal and distributed data sources**. Such data sources include sensors, videos, click streams, and all other available data sources.

At present, the main sources of big data are the operation and trading information in enterprises, logistic and sensing information in the IoT, human interaction information and position information in the Internet world, and data generated in scientific research, etc.

Enterprise Data

In 2013, IBM issued a report titled "Analytics: The Real-world Use of Big Data," which indicates that the internal data of enterprises are the main sources of big data.

The internal data of enterprises mainly consists of online trading data and online analysis data, most of which are historically static data and are managed by RDBMSs in a structured manner.

In addition, production data, inventory data, sales data, and financial data, etc., also constitute enterprise internal data.

IoT Data

IoT is an important source of big data. Among smart cities constructed based on IoT, big data may come from industry, agriculture, traffic and transportation, medical care, public departments, and households, etc.

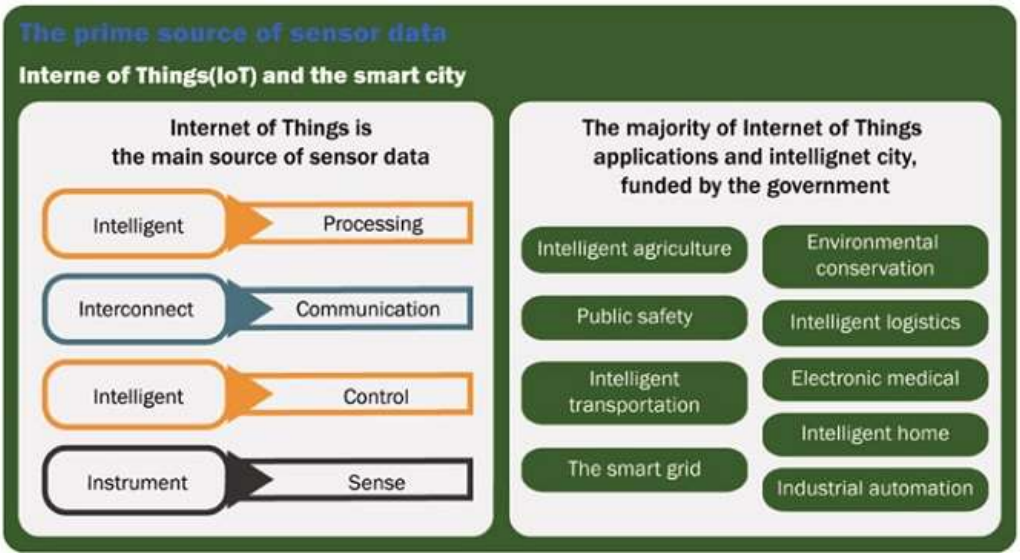

Figure: Illustration of the prime source of sensory data

Big Data Generation and Acquisition

According to the processes of data acquisition and transmission in IoT, its network architecture may be divided into three layers:

Sensing layer - is responsible for data acquisition and mainly consists of sensor networks.

Network layer- is responsible for information transmission and processing, where close transmission may rely on sensor networks, and remote transmission shall depend on the Internet.

Application layer- support specific applications of IoT.

According to the characteristics of IoT, the data generated from IoT has the following features: **Large-Scale Data, Heterogeneity, Strong Time and Space Correlation, and Effective Data Accounts for Only a Small Portion of the Big Data**.

Big Data Generation and Acquisition

Internet Data

Internet data consists of **searching entries, Internet forum posts, chatting records, and microblog messages** among others, which have similar features, such as high value and low density.

Such Internet data may be valueless individually, but through the exploitation of accumulated big data, useful information such as the habits and hobbies of users can be identified, and it is even possible to forecast users' behavior and emotional moods.

Big Data Generation and Acquisition

Bio-medical Data

As a series of high-throughput bio-measurement technologies are innovatively developed in the beginning of the twenty-first century, the frontier research in the bio-medicine field also enters the era of big data.

By constructing smart efficient, and accurate analytical models and theoretical systems for bio-medicine applications, the essential governing mechanism behind complex biological phenomena may be revealed.

Not only the future development of bio-medicine can be determined, but also the leading roles can be assumed in the development of a series of important strategic industries related to the national economy, people's livelihood, and national security, with important applications such as medical care, new drug R&D, and grain production.

Big Data Generation and Acquisition

One sequencing of human genes may generate 100–600GB of raw data.

In the China National Genebank in Shenzhen, there is 1.3 million samples including 1.15 million human samples and 150,000 animal, plant, and microorganism samples. By the end of 2013, 10 million traceable biological samples will be stored, and by the end of 2015, this figure will reach 30 million. (The info was updated in 2014 by Min Chen and others.)

Big Data Generation and Acquisition

Data Generation from Other Fields

As scientific applications are increasing, the scale of datasets is gradually expanding, and the development of some disciplines greatly relies on the analysis of masses of data.

The **first example** is related to computational biology. GenBank is a nucleotide sequence database maintained by the U.S. National Bio-Technology Innovation Center.

The **second example** is related to astronomy. Sloan Digital Sky Survey (SDSS), the biggest sky survey project in astronomy, has recorded 25TB of data from 1998 to 2008.

The **last application** is related to high-energy physics. At the beginning of 2008, the Atlas experiment of the Large Hadron Collider (LHC) of the European Organization for Nuclear Research generated raw data at 2PBs and stored about 10TB of processed data per year.

Big Data Generation and Acquisition

Big Data Acquisition

Big data acquisition includes data collection, data transmission, and data preprocessing.

During big data acquisition, once the raw data is collected, an efficient transmission mechanism should be used to send it to a proper storage management system to support different analytical applications.

The collected datasets may sometimes include much redundant or useless data, which unnecessarily increases storage space and affects the subsequent data analysis.

Data Collection

Data collection is to utilize special data collection techniques to acquire raw data from a specific data generation environment. **Four common data collection methods** are shown as follows.

Log Files: As one widely used data collection method, log files are record files automatically generated by the data source system, so as to record activities in designated file formats for subsequent analysis.

Log files are typically used in nearly all digital devices. For example, web servers record in log files the number of clicks, click rates, visits, and other property records of web users.

Sensors: Sensors are common in daily life to measure physical quantities and transform physical quantities into readable digital signals for subsequent processing (and storage). Sensory data may be classified as sound waves, voice, vibration, automobile, chemical, current, weather, pressure, temperature, etc. The sensed information is transferred to a data collection point through wired or wireless networks. For example, see page 24.

Methods for Acquiring Network Data: At present, network data acquisition is accomplished using a combination of web crawler, word segmentation system, task system, and index system, etc. Web crawler is a program used by search engines for downloading and storing web pages. Data acquisition through a web crawler is widely applied in applications based on web pages, such as search engines or web caching.

The current network data acquisition technologies mainly include traditional Libpcap-based packet capture technology, zero-copy packet capture technology, as well as some specialized network monitoring software such as Wireshark, SmartSniff, and WinNetCap.

Data Transportation

Upon the completion of raw data collection, data will be transferred to a data storage infrastructure for processing and analysis. Big data is mainly stored in a data center. The data layout should be adjusted to improve computing efficiency or facilitate hardware maintenance. In other words, internal data transmission may occur in the data center.

Therefore, data transmission consists of two phases:

Inter-DCN transmissions - are from data source to data center, which is generally achieved with the existing physical network infrastructure.

Intra-DCN transmissions - are the data communication flows within data centers. Intra-DCN transmissions depend on the communication mechanism within the data center .

Data Pre-processing

Because of the wide variety of data sources, the collected datasets vary with respect to noise, redundancy, and consistency, etc., and it is undoubtedly a waste to store meaningless data.

In addition, some analytical methods have stringent requirements on data quality. Therefore, data should be pre-processed under many circumstances to integrate the data from different sources, so as to enable effective data analysis.

Some relational data pre-processing techniques are - Integration, Cleaning, and Redundancy Elimination.

Big Data Acquisition

Big Data Acquisition

As the second phase of the big data system, big data acquisition includes data collection, data transmission, and data pre-processing. During big data acquisition, once the raw data is collected, an efficient transmission mechanism should be used to send it to a proper storage management system to support different analytical applications.

Big Data Acquisition-Data Collection

Data Collection

Data collection is to utilize special data collection techniques to acquire raw data from a specific data generation environment. Four common data collection methods are shown as follows.

- Log Files
- Sensors
- Methods for Acquiring Network Data:
  - Libpcap-based packet capture technology
  - Zero-copy packet capture technology
  - as well as some specialized network monitoring software such as Wireshark, SmartSniff, and WinNetCap.
- Mobile Equipments

In addition to the aforementioned three data acquisition methods of main data sources, there are many other data collect methods or systems. For example, in scientific experiments, many special tools can be used to collect experimental data, such as magnetic spectrometers and radio telescopes.

Big Data Acquisition

Data Transportation

Upon the completion of raw data collection, data will be transferred to a data storage infrastructure for processing and analysis. As discussed in the previous Section, big data is mainly stored in a data center. The data layout should be adjusted to improve computing efficiency or facilitate hardware maintenance.

Data transmission consists of two phases:

- Inter-DCN transmissions: Inter-DCN transmissions are from data source to data center, which is generally achieved with the existing physical network infrastructure. Because of the rapid growth of traffic demands, the physical network infrastructure in most regions around the world are constituted by high-volume, high-rate, and cost-effective optic fiber transmission systems.
- Intra-DCN transmissions: Intra-DCN transmissions are the data communication flows within data centers. Intra-DCN transmissions depend on the communication mechanism within the data center

Big Data Acquisition

Data Pre-processing

Because of the wide variety of data sources, the collected datasets vary with respect to noise, redundancy, and consistency, etc., and it is undoubtedly a waste to store meaningless data. In addition, some analytical methods have stringent requirements on data quality. Therefore, data should be pre-processed under many circumstances to integrate the data from different sources.

Some relational data pre-processing techniques are discussed in the following:

- Integration
- Cleaning
- Redundancy Elimination