



Co-funded by the
Erasmus+ Programme
of the European Union

SHYFTE 4.0

Building Skills 4.0 through University and Enterprise Collaboration

Principle and Application of BigData Technology- Introduction of BigData

Pilot of SE&BD

2020-06-20

Xi- YU
yuxi@cdu.edu.cn



Principle and Application of BigData Technology

-Introduction of Big Data



Co-funded by the
Erasmus+ Programme
of the European Union

- Examples Of Big Data
- Types Of Big Data
- Characteristics Of Big Data
- Key technologies of Big Data
- Benefits Of Big Data Processing

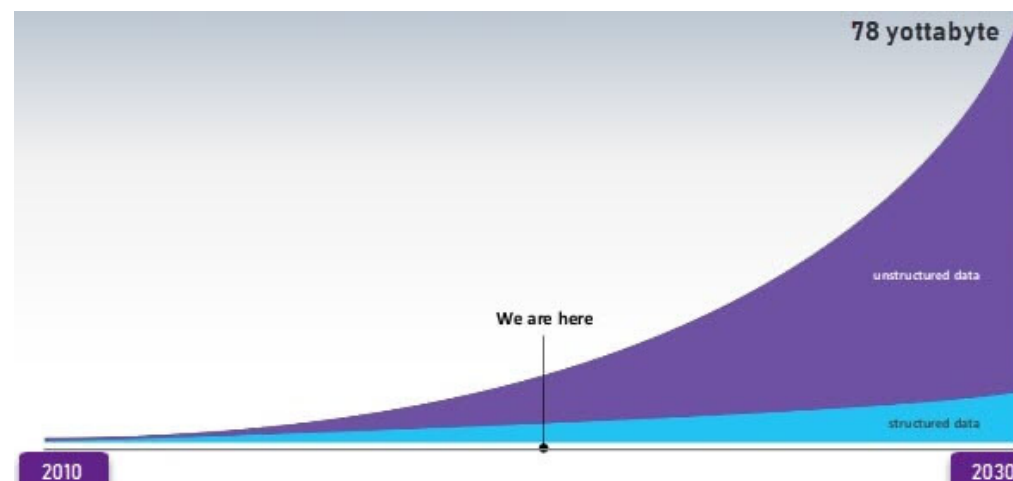
- DATA
 - The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.
- BIG DATA
 - Big Data is also **data** but with a **huge size**. Big Data is a term used to describe a collection of data that is huge in volume and yet growing exponentially with time. In short such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.

Principle and Application of BigData Technology

-Introduction of Big Data



Co-funded by the
Erasmus+ Programme
of the European Union



Principle and Application of BigData Technology

-Introduction of Big Data



Co-funded by the
Erasmus+ Programme
of the European Union



In the book “Big Data”, Viktor Mayer-Schonberger makes it clear that the biggest shift in the ara of big data is to abandon the desire for **causality** and focus instead on **correlations**: not knowing **why** but only **What**. This overturns centuries of established practices and challenges our most basic understanding of how to make decisions and comprehend reality. The book argues that the core of big data is prediction. Big data will create unprecedented quantifiable dimensions of human life.

- According to Lou Gerstner, former CHIEF executive of IBM, the IT world is experiencing a major change every 15 years

Three waves of informationization

waves of informationization	Time	Flag	Solved Problem	Representative enterprise
First Wave	Around 1980	Personal computer	Information processing	Intel, AMD, IBM, Apple, MS , Lenovo, Dell, HP etc.
Second Wave	Around 1995	Internet	Information transmission	Yahoo, Google, Alibaba, Baidu, Tencent etc
Third Wave	Around 2010	IoT, cloud computing and big data	Information explosion	

Principle and Application of BigData Technology

-Introduction of Big Data

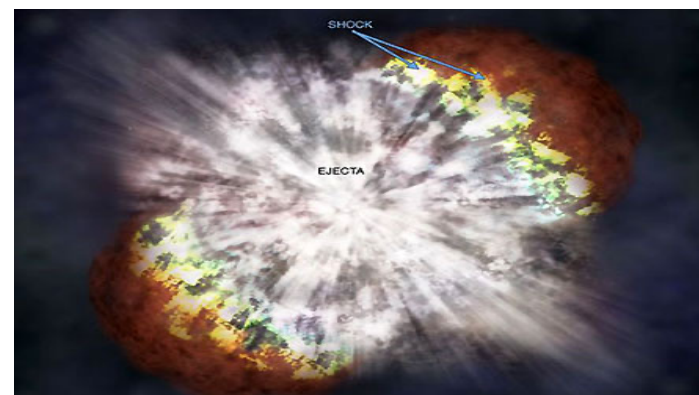


Co-funded by the
Erasmus+ Programme
of the European Union

Era of digital

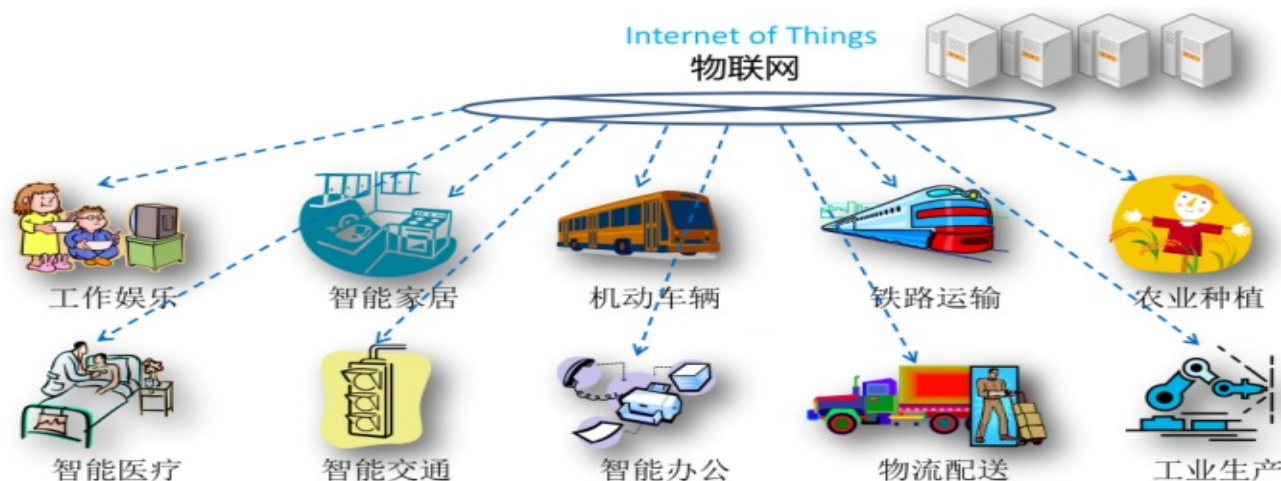


Era of data explosion



Unit	B	KB	MB	GB	TB	PB	EB	ZB	YB
Base	2	2	2	2	2	2	2	10	10
Power	0	10	20	30	40	50	60	21	24

- Ubiquitous data



“物” 皆在数据化的路上

Examples Of Big Data



Co-funded by the
Erasmus+ Programme
of the European Union

The **New York Stock** Exchange generates about **one terabyte** of new trade data per day.



The statistic shows that **500+terabytes** of new data get ingested into the databases of social media site **Facebook**, every day. This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments etc.



A single **Jet engine** can generate **10+terabytes** of data in **30 minutes** of flight time. With many thousand flights per day, generation of data reaches up to many **Petabytes**.



1. Structured Data

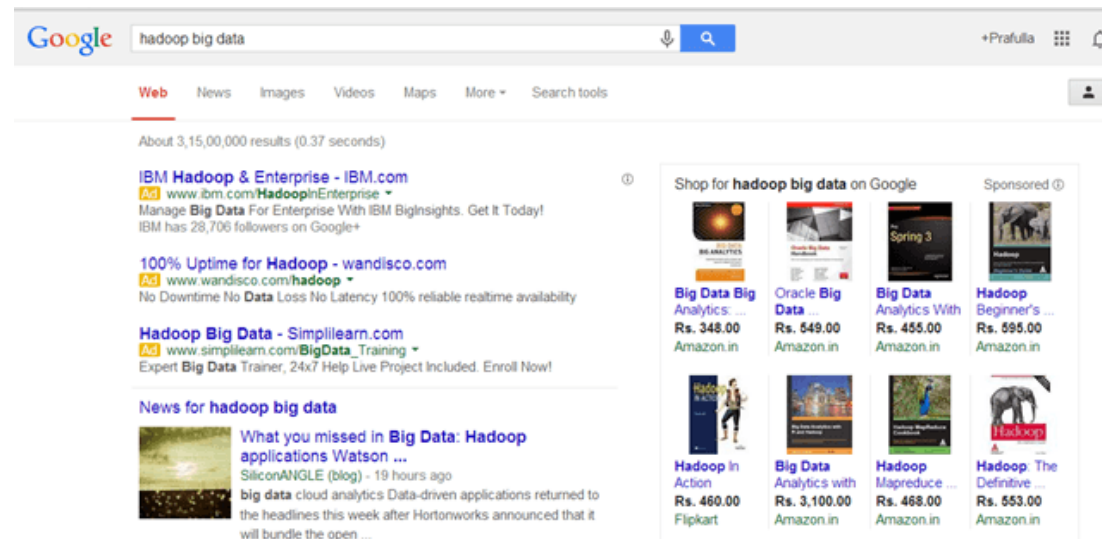
Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data. Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it. However, nowadays, we are foreseeing issues when a size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabytes.

An 'Employee' table in a database

Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

2. Unstructured

Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc. Now day organizations have wealth of data available with them but unfortunately, they don't know how to derive value out of it since this data is in its raw form or unstructured format.



Google search results for "hadoop big data". The search bar shows "hadoop big data" and the results indicate "About 3,15,00,000 results (0.37 seconds)".

Web News Images Videos Maps More Search tools

IBM Hadoop & Enterprise - IBM.com
www.ibm.com/HadoopinEnterprise
Manage Big Data For Enterprise With IBM Bigsights. Get It Today!
IBM has 28,706 followers on Google+

100% Uptime for Hadoop - wandisco.com
www.wandisco.com/hadoop
No Downtime No Data Loss No Latency 100% reliable realtime availability

Hadoop Big Data - Simplilearn.com
www.simplilearn.com/BigData_Training
Expert Big Data Trainer, 24x7 Help Live Project Included. Enroll Now!

News for hadoop big data

What you missed in Big Data: Hadoop applications Watson ...
SiliconANGLE (blog) - 19 hours ago
big data cloud analytics Data-driven applications returned to the headlines this week after Hortonworks announced that it will bundle the open ...

Shop for hadoop big data on Google Sponsored

Product	Price	Platform
Big Data Analytics ...	Rs. 348.00	Amazon.in
Oracle Big Data ...	Rs. 549.00	Amazon.in
Big Data Analytics With ...	Rs. 455.00	Amazon.in
Hadoop Beginner's ...	Rs. 595.00	Amazon.in
Hadoop in Action	Rs. 460.00	Flipkart
Big Data Analytics with ...	Rs. 3,100.00	Amazon.in
Hadoop Mapreduce ...	Rs. 468.00	Amazon.in
Hadoop: The Definitive ...	Rs. 553.00	Amazon.in

3. Semi-structured

Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in an XML file.

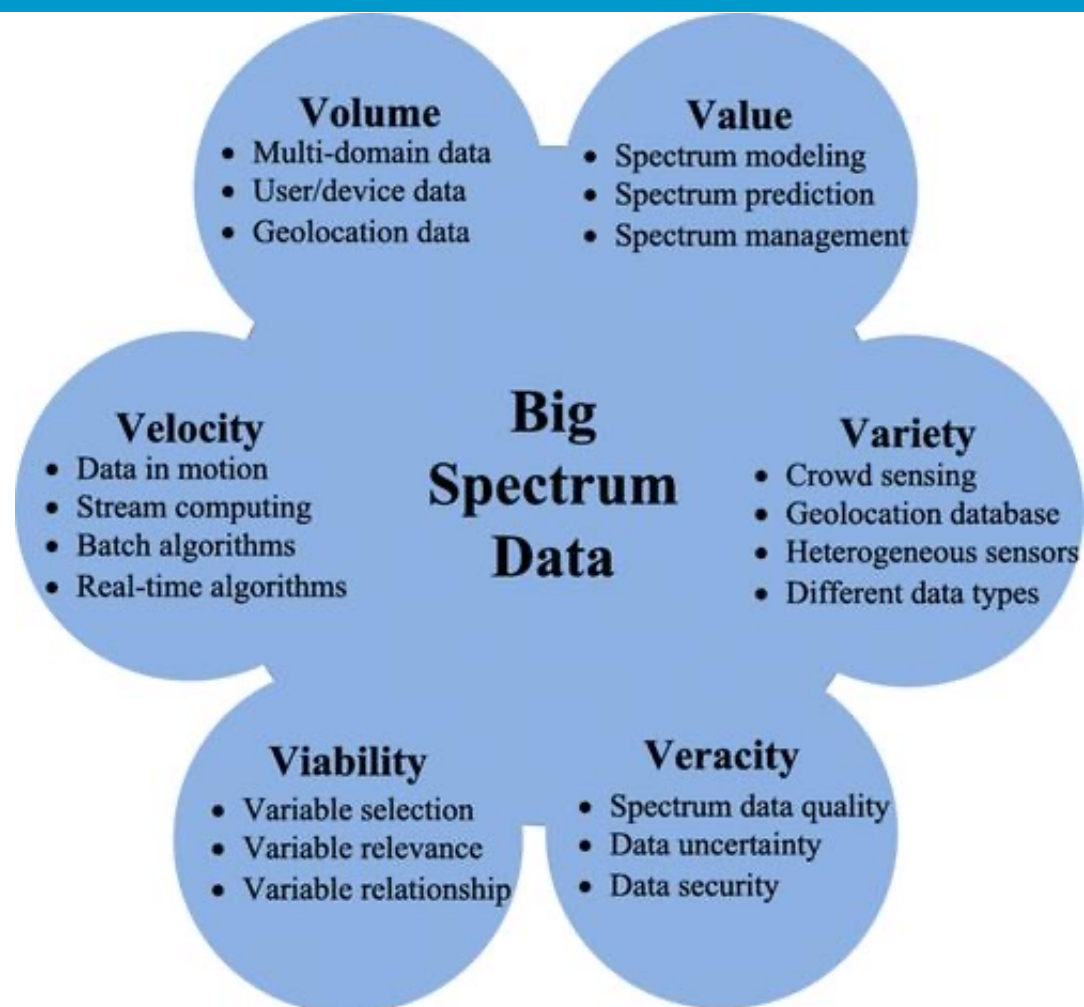
Personal data stored in an XML file-

```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>  
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>  
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>  
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>  
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```

Characteristics Of Big Data



Co-funded by the
Erasmus+ Programme
of the European Union

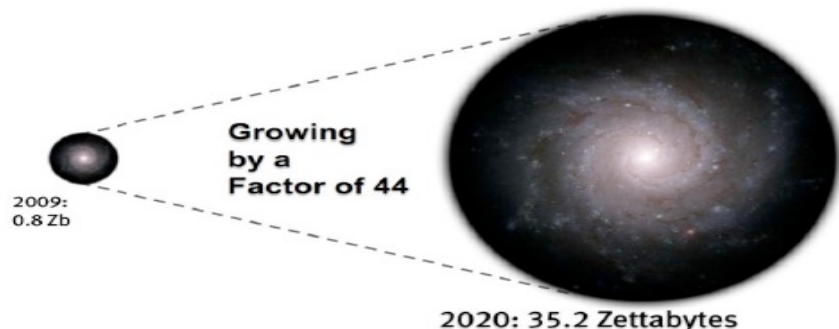


Characteristics Of Big Data-Volume



Co-funded by the
Erasmus+ Programme
of the European Union

- According to estimates by IDC, data has been growing at an annual rate of 50%, which means it is doubling every two years (Moore's Law of Big Data)
- The amount of data that humans have produced in the last two years is equal to the amount of data they have produced in the past.
- By 2020, the world will have a total of 35 ZBS of data, an increase of nearly 30 times compared with 2010



TERABYTE	10 的 12 次方	一块 1TB 硬盘		200,000 照片或 mp3 歌曲
PETABYTE	10 的 15 次方	两个数据中心机柜		16 个 Blackblaze pod 存储单元
EXABYTE	10 的 18 次方	2,000 个机柜		占据一个街区的 4 层数据中心
ZETTABYTE	10 的 21 次方	1000 个数据中心		纽约曼哈顿的 1/5 区域
YOTTABYTE	10 的 24 次方	一百万个数据中心		特拉华州和罗德岛州

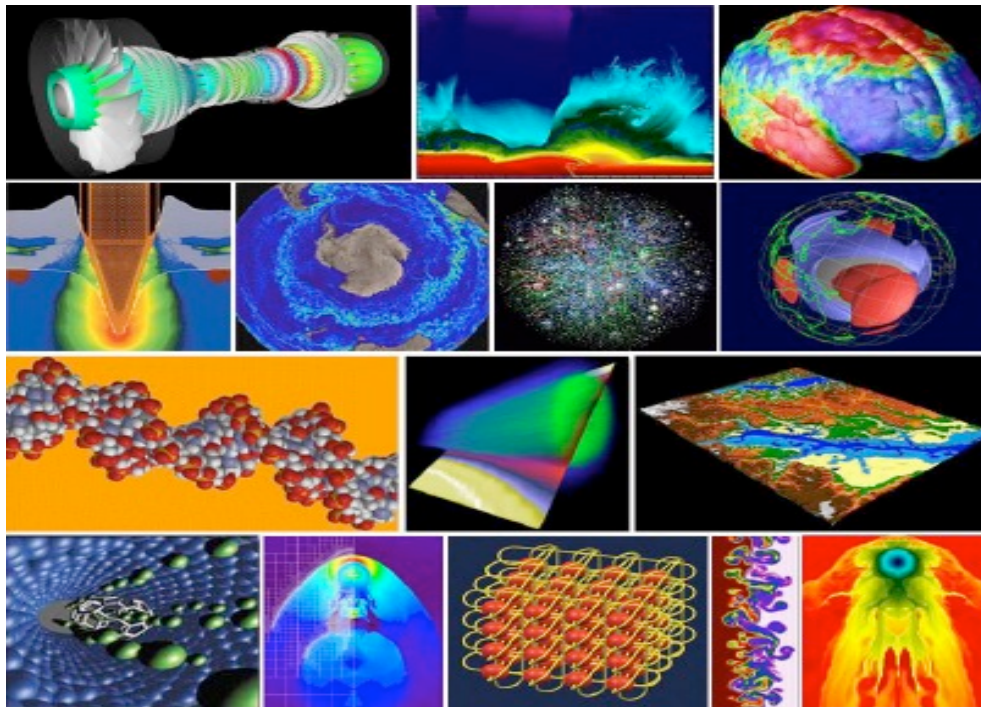
Characteristics Of Big Data- Variety



Co-funded by the
Erasmus+ Programme
of the European Union

Big data is composed by structured and unstructured data.

- 10% of the structured data is stored in the database.
- 90% of unstructured data, which is closely related to human information.



❑ Scientific research

- genome
- the LHC accelerator
- Earth and space exploration

❑ Enterprise application

- Email, documents, files
- Application log
- Transaction records

❑ Web 1.0 data

- Text
- Image
- video

❑ Web 2.0 data

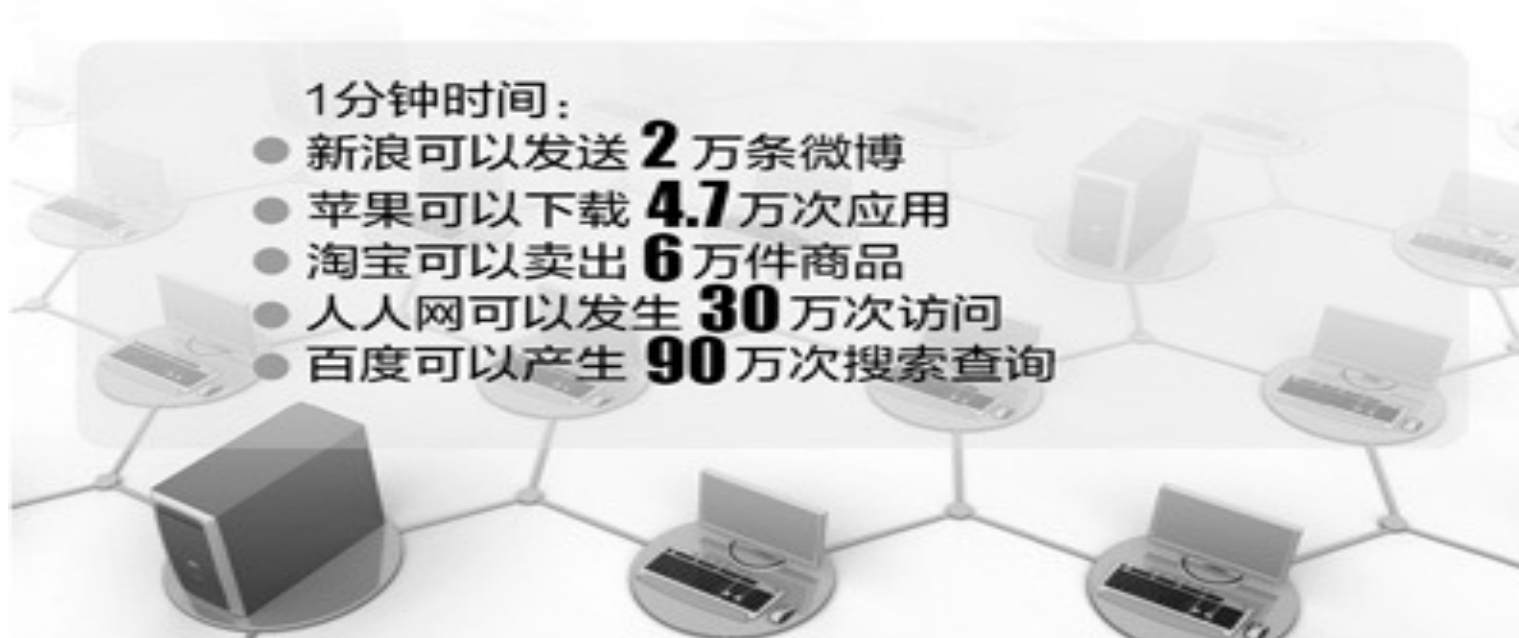
- Query log/click stream
- Twitter/Blog/SNS
- Wiki

Characteristics Of Big Data-Velocity



Co-funded by the
Erasmus+ Programme
of the European Union

- ❑ From data generation to consumption, the time window is very small, and the time available to make decisions is very small
- ❑ 1 second law: this point is also different from the traditional data mining technology



Characteristics Of Big Data-Value



Co-funded by the
Erasmus+ Programme
of the European Union

Low value density, high commercial value

In the case of video, for example, there may be only one or two seconds of useful data in the continuous monitoring process, but it has a high commercial value.



Key technologies of Big Data



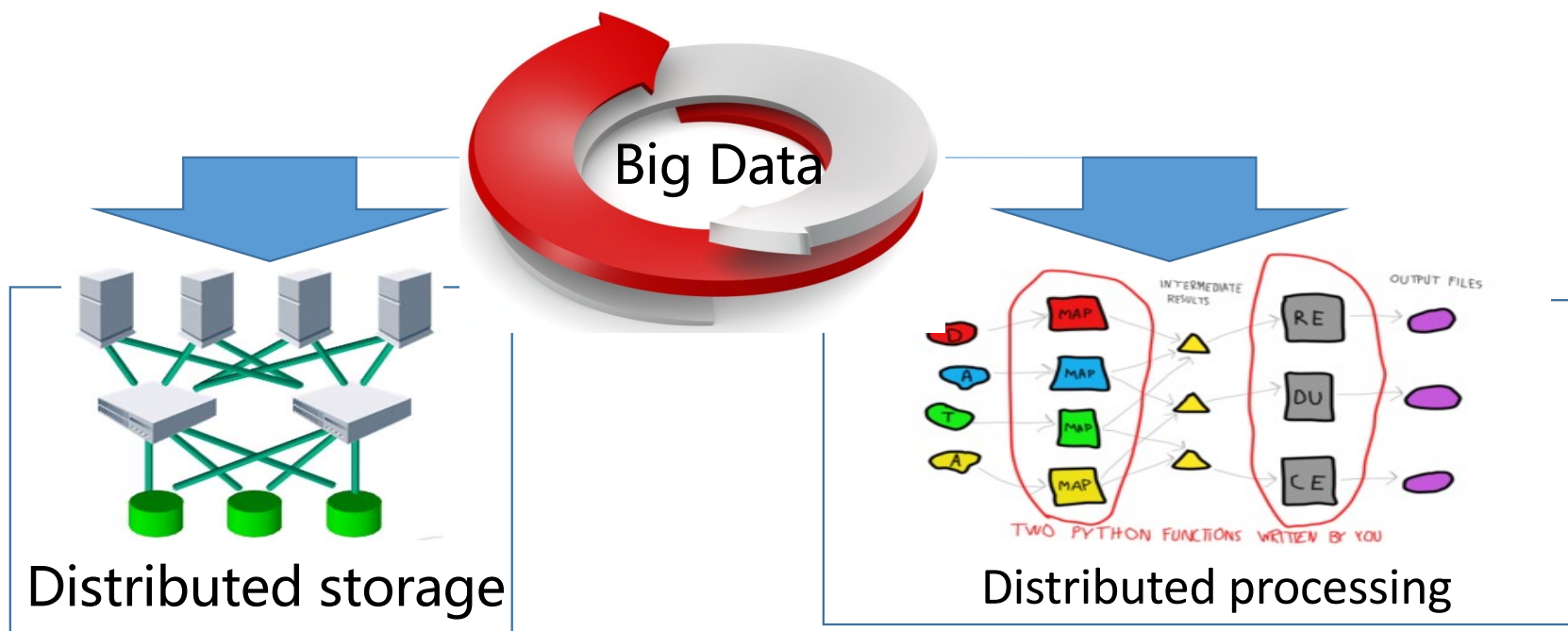
Different aspects and functions of big data technology

Technology aspect	Function
Data collection	ETL tool is used to extract the data such as relational data and flat data files from distributed and heterogeneous data sources to the temporary middle layer for cleaning, transformation and integration, and finally loading into the data warehouse or data mart, which becomes the basis of online analysis and processing and data mining. Alternatively, the data collected in real time can be used as input to the flow computing system for real time processing and analysis.
Data storage and management	Using distributed file system, data warehouse, relational database, NoSQL database, cloud database, etc., to achieve the storage and management of structured, semi-structured and unstructured massive data.
Data processing and analysis	The distributed parallel programming model and computing framework, combined with machine learning and data mining algorithm, can realize the processing and analysis of massive data. The visual presentation of the analysis results can help people better understand and analyze the data.
Data privacy and security	In addition to mining huge potential commercial value and academic value from big data, the privacy data protection system and data security system should be built to effectively protect personal privacy and data security.

Key technologies of Big Data



Co-funded by the
Erasmus+ Programme
of the European Union



GFS\HDFS
BigTable\HBase
NoSQL
NewSQL

MapReduce

- **Businesses can utilize outside intelligence while taking decisions**

Access to social data from search engines and sites like facebook, twitter are enabling organizations to fine tune their business strategies.

- **Improved customer service**

Traditional customer feedback systems are getting replaced by new systems designed with Big Data technologies. In these new systems, Big Data and natural language processing technologies are being used to read and evaluate consumer responses.

- Early identification of risk to the product/services, if any
- Better operational efficiency

Big Data technologies can be used for creating a staging area or landing zone for new data before identifying what data should be moved to the data warehouse. In addition, such integration of Big Data technologies and data warehouse helps an organization to offload infrequently accessed data.

- **Big Data is defined as data that is huge in size. Bigdata is a term used to describe a collection of data that is huge in size and yet growing exponentially with time.**
- **Examples of Big Data generation includes stock exchanges, social media sites, jet engines, etc.**
- **Big Data could be 1) Structured, 2) Unstructured, 3) Semi-structured**
- **Volume, Variety, Velocity, and Variability are few Characteristics of Bigdata**
- **Improved customer service, better operational efficiency, Better Decision Making are few advantages of Bigdata**



Co-funded by the
Erasmus+ Programme
of the European Union

SHYFTE 4.0

Thank you !

Xi, YU

Pilot of SE&BD

ChengDu University

yuxi@cdu.edu.cn