IIT KHARAGPUR

NPTEL ONLINE CERTIFICATION COURSES

Industry 4.0: Big Data and Advanced Analysis

Dr. Sudip Misra

Professor

Department of Computer Science and Engineering

Indian Institute of Technology Kharagpur

Email: smisra@sit.iitkgp.ernet.in

Website: http://cse.iitkgp.ac.in/~smisra/

Research Lab: cse.iitkgp.ac.in/~smisra/swan/

What is Big Data?

- Big data means
  - data which is “too big” to be handled by
    - processing tools, and
    - conventional databases.
- Big data consists of
  - structured and
  - non-structured data

  such as web blogs, FB chats, images, news, tweets, comments, etc.

Source: cs.kent.edu: Big data

Big Data: Definition

- *"Big data will represent the data of which acquisition speed, data volume or data characterization restricts the capacity of using conventional associated methods to manage successful analysis or the data which can be successfully operated with important horizontal zoom technologies."*

  [NIST(National Institute of Standards and Technology)]

Source: cs.kent.edu: Big data

Data Types

- Structured data
  - Data that can be easily organized.
  - It is stored in relational databases.
  - It is managed by Structured Query Language (SQL) in databases.
  - It accounts for only 20% of the total available data today in the world.

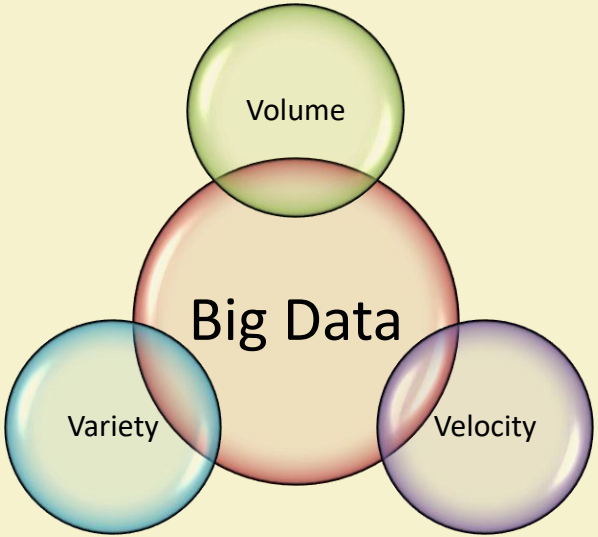
Source: Big data analytics : Srinivasa

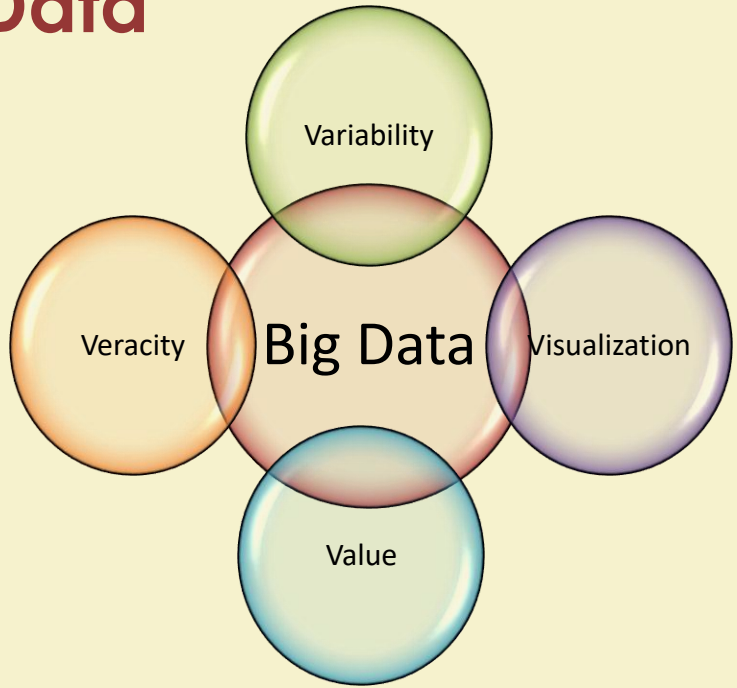
Data Types(Contd.)

- Unstructured data
  - Data that do not possess any pre-defined model.
  - Traditional RDBMSs are unable to process unstructured data.
  - Enhances the ability to provide better insight to huge datasets.
  - It accounts for 80% of the total data available today in the world.

Source: Big data analytics : Srinivasa

Characteristics of Big Data





- There are mainly 3 Vs in Big Data
- Some authors also include another 4 Vs

Source: Big data analytics : Srinivasa

Characteristics of Big Data (Contd.)

- Volume
  - Quantity of created data.
  - Sources of data are added continuously.
  - Example of *volume* -
    - More than 32TB of pictures will be created each night from the Large Synoptic Survey Telescope (LSST).
    - In every minute, 70 hours of video is uploaded to Youtube.

Source: Big data analytics : Srinivasa

Characteristics of Big Data (Contd.)

- Velocity
  - Speed of generation of data.
  - Data processing time is decreasing day by day to provide real-time services.
  - Older processing technologies can not help to handle high velocity of data.
  - Example of *velocity* –
    - 140 million tweets per day on average (according to a survey conducted in 2011)
    - NYSE(New York Stock Exchange) measures 1TB of exchange data during every exchanging session.

Source: Big data analytics : Srinivasa

Characteristics of Big Data (Contd.)

- Variety
  - Category of the data.
  - No restriction over the input data formats.
  - Mostly data are not structured.
  - Example of *variety* –
    - Pure text, images, audio, video, web, GPS data, sensor data, SMS, documents, PDFs, flash etc.

Source: Big data analytics : Srinivasa

Characteristics of Big Data (Contd.)

- Variability
  - Variability is different from variety.
  - Data whose meaning is constantly changing.
  - Such data appear as an indecipherable mass without structure.
  - Example:
    - Language processing, Hashtags, Geo-spatial data, Multimedia, Sensor events.

Source: Big data analytics : Srinivasa

Characteristics of Big Data (Contd.)

- Veracity
  - Veracity indicates biasness in the data, unusualness and noise in data.
  - It is important in programs which involve automated decision-making.
  - It is also important for feeding the data into an unsupervised machine learning algorithm.
- Veracity deals about the data understandability, not just the data quality.

Source: Big data analytics : Srinivasa

Characteristics of Big Data (Contd.)

- Visualization
  - Data can be in form of pictures or in form of a graphical format.
  - Visualization provides the power to decision makers to see visually.
  - It is helpful to identify new patterns.
- Value
  - It means extracting useful business information from scattered data.
  - Simple to access and provides quality investigation that empowers informed decisions.

Source: Big data analytics : Srinivasa

Data Sources

Enterprise data

- Online trading & data analysis
- Production and inventory data
- Sales and other financial data

IoT data

- Industrial data
- Healthcare data
- Agricultural data

Source: The Making of ENCODE: Lessons for Big-Data Projects : Birney

Data Sources

Biomedical data

- Data generated from gene sequencing
- Data from medical clinics

Others

- Computational biology
- Astronomy
- Nuclear research

Source: The Making of ENCODE: Lessons for Big-Data Projects : Birney

Data Acquisition

- Data collection
  - Data sources automatically generate log files or record files to record activities for further analysis.
  - Complex and variety of data collection through mobile devices. E.g. – geographical location, 2D barcodes, pictures, videos etc.
- Data transmission
  - Categorized as – Inter-DCN transmission and Intra-DCN transmission.
  - Collect data and transfer to storage system for further processing and analysis of the data.

Source: The Making of ENCODE: Lessons for Big-Data Projects : Birney

Data Acquisition (Contd.)

- Data pre-processing
  - Pre-processing of data is necessary as collected datasets suffer from noise, redundancy etc.
  - Pre-processing of relational data mainly follows-

Source: The Making of ENCODE: Lessons for Big-Data Projects : Birney

Data Acquisition (Contd.)

- Integration:
  - combine data from various sources and
  - delivers the users a constant data view.
- Clearing:
  - spot incorrect, insufficient, or uncooperative data, and
  - correct or remove such data.
- Redundancy mitigation:
  - eliminate data repetition through detection, filter and compression of data to avoid unnecessary transmission.

Source: The Making of ENCODE: Lessons for Big-Data Projects : Birney

Data Storage

- File system
  - Distributed file systems that store massive data and ensure – consistency, accessibility, and fault tolerance of data.
    - GFS is a distributed file system that supports large-scale file system.
    - HDFS(Hadoop Distributed File System) is a notable file systems, derived from the open source codes of GFS.
- Databases
  - Emergence of non-traditional relational databases (NoSQL) in order to deal with the characteristics that big data possess.

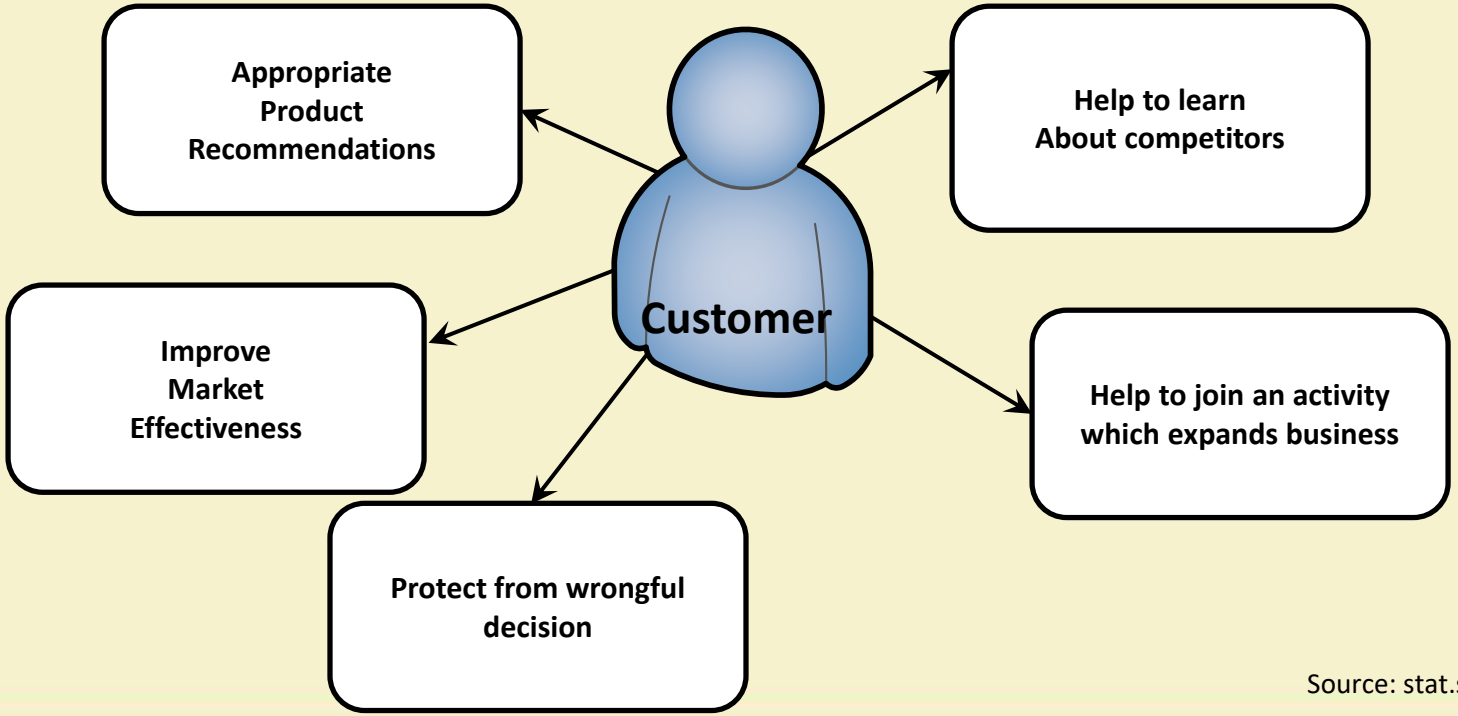
Source: The Making of ENCODE: Lessons for Big-Data Projects : Birney

Why Data Analytics?

Sensors are very small in sizes. They can be placed anywhere and transfer the data over wireless technology, because of this explosion of data moving to systems from sensors. Some data are irrelevant for systems. How can one know which data are relevant, this requires analysis of the data.

Source: Industry 4.0:The Industrial Internet of Things: Gilchrist

Why Data Analytics?(Contd.)



Source: stat.si: Big data tutorial

Big Data Analytics

- Big data is different from conventional Data Warehouse (DW) approaches.
- Big data apps cannot be fit in traditional DW architectures (e.g. Exadata, Teradata).
- Distributed nothing, mighty parallel performing, scale out frameworks are convenient for big data apps.

Source: Industry 4.0:The Industrial Internet of Things: Gilchrist

Big Data Analytics for Industry 4.0

- Industrial Internet require an approach to manage and process data coming from thousand of sensors for precious perceptions .
- To manage and handle the huge data in health services and manufacturing etc. is not new. For example-
  - An event is detected by a sensor and sent to the operational recorder. An operational recorder is a database which stores data. After that this data is optimized by querying such as, what about this hour's production from the norm.

Source: Industry 4.0:The Industrial Internet of Things: Gilchrist

Big Data Analytics for Industry 4.0 (Contd.)

- IIoT can be recognized as a big benefactor of Big Data.
- It needs new technologies to manage vast data.
- Cloud services are accessible to handle Big Data with no-limit of storage on demand.
- In IIoT, Hadoop (open source cloud based distributed data storage) is also available for managing the data.

Source: Industry 4.0:The Industrial Internet of Things: Gilchrist

Cloud-Based Method for Analytics

- Essential features (according to NIST)
  - On-demand self service
  - Wide network access
  - Method grouping
  - Fast flexibility
  - Measured service

Source: Industry 4.0:The Industrial Internet of Things: Gilchrist

Types of Analytics

Prescriptive Analytics

-> Best action?

-> Should we try this?

Predictive Analytics

->What next?

->Pattern?

Descriptive Analytics

->When, where?

->What happened?

Source: Industry 4.0:The Industrial Internet of Things: Gilchrist

References

[1] A. Machanavajjhala and J.P. Reiter, “Big Privacy: Protecting Confidentiality in Big Data,” ACM Crossroads, vol. 19, no. 1, pp. 20-23, 2012.

[2] E. Birney, “The Making of ENCODE: Lessons for Big-Data Projects,” Nature, vol. 489, pp. 49-51, 2012.

[3] J. Bughin, M. Chui, and J. Manyika, Clouds, Big Data, and Smart Assets: Ten Tech-Enabled Business Trends to Watch. McKinSey Quarterly, 2010.

[4] S. Banerjee and N. Agarwal, “Analyzing Collective Behavior from Blogs Using Swarm Intelligence,” Knowledge and Information Systems, vol. 33, no. 3, pp. 523-547, Dec. 2012.

[5] Marko Grobelnik (2012).Big-Data Tutorial.Online .URL https://www.stat.si/dokument/8682/BigDataIntro-MarkoGrobelnik.pdf.

[6] Ruoming Jin.Introduction to Big Data.Online.URL https://www.cs.kent.edu/~jin/BigData/.

[7] S. Aral and D. Walker, “Identifying Influential and Susceptible Members of Social Networks,” Science, vol. 337, pp. 337-341, 2012.

[8] Srinivasa S.,& Bhatnagar, V.(2012), Big data analytics, Springer.

[9] Gilchrist A.(2016).Industry 4.0:The Industrial Internet of Things.Apress.

Thank You!!