SE ZG522: Big Data Systems

01 Logistics and Introduction

Dr. Kamlesh Tiwari
Assistant Professor, Department of CSIS,
BITS Pilani, Pilani Campus, Rajasthan-333031 INDIA

Jan 28, 2021 ONLINE (WILP @ BITS-Pilani Jan-May 2022)

http://ktiwari.in/bds

Introduction: Data, Information and Knowledge

What is data?

Fact or values

What is Information?

Processed output of date

What is Knowledge?

Understanding of information.

X1	X2	X3		Y
2	3	4		1
8	7	5		10
9	6	1		14
7	4	9		2
1	5	5		1
5	3	6		2
7	4	4		7
6	3	5		4
3	2	3		

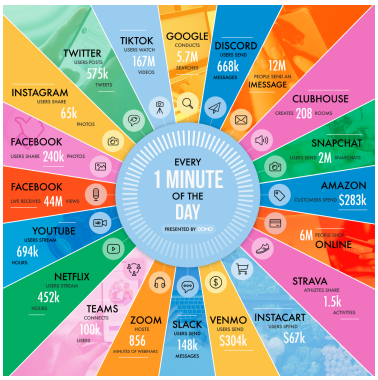
Type of Data

- Structured - RDBMS (Relational Databases)
- Semi-Structured - Ecel/XML (Object Stores)
- Quasi-Structured - Search result on google
- Unstructured - youTube (File systems)

Big Data

What is big data?[1]

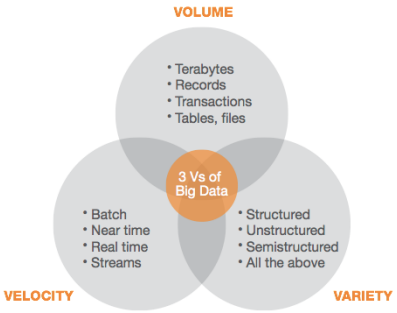
- 2.5 quintillion (1×10^{18}) bytes of data/day
- 90% generated in last 2 year. 80% of it is unstructured.
- Facebook, 4 million posts/min
- Aircraft: each engine generates 10 TB data in 30 min. UK to New York would generate 640 TB of data



[1] https://www.visualcapitalist.com/from-amazon-to-zoom-what-happens-in-an-internet-minute-in-2021/

Characteristics of Big Data

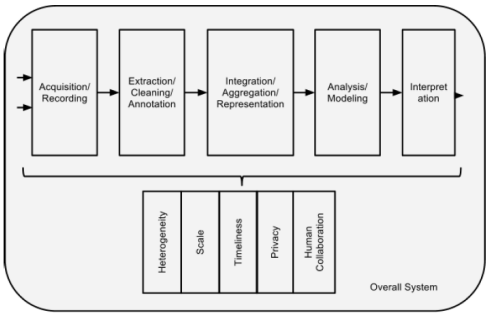
3V's: The "BIG" in big data is not only for volume



Few More: Veracity, Value

Big Data

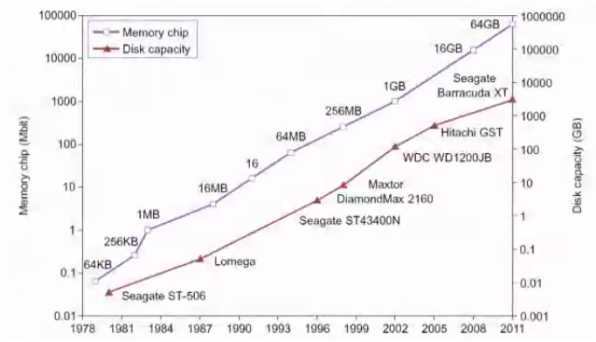
Major steps in analysis of big data: [2]



Solution is **Distributed computing**

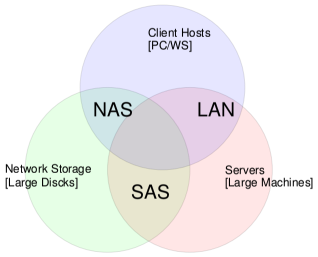
[2] "Challenges and Opportunities with Big Data" A community white paper developed by leading researchers across the United States

Storage at Hardware Level



10x increase in every eight year.

System Perspective



Enabeling technologies for distributed data systems

Network Attached Storage, Storage Area Network, Local Area Network

Locality of Reference

- Memory Hirarchy
- Processing: In-memory, Secondary, network
- Locality of Reference
  - Temporal locality - same instr
  - Spatial locality - nearby instr
- Latency

Perspective: Knowledge Discovery in Databases

What is **data-mining**?

Computation to facilitate Knowledge Discovery in Databases (**KDD**)

Goal of Data Mining (motivation of doing the same?)

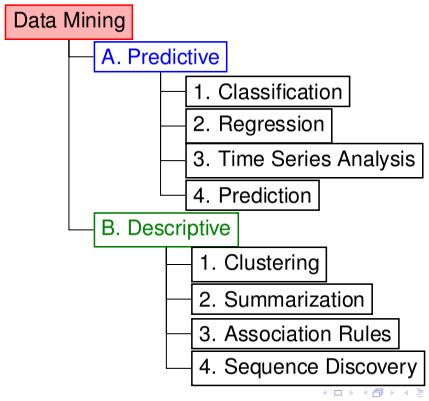
To provide efficient tools and techniques for KDD

Knowledge Discovering in Databases (KDD) involves

1. **Selection:** collection of data
2. **Preprocessing:** deal with incorrect/missing data
3. **Transformation:** common format and preprocessing
4. **Data Science:** algorithmic tools
5. **Interpretation/Evaluation:** presentation and visualization

Data Mining: Tasks

Two broad categories of data mining models are *Predictive* and *Descriptive*. Some of the related tasks are



KDD Issues

- Human interaction
- Overfitting, Outliers
- Large dataset
- High dimension
- Multimedia data
- Missing data
- Irrelevant data
- Noisy data
- Changing data

And much more...

Course and Syllabus

Introduction and basics Storage for Data Big Data
Locality of Reference Distributed Processing Cloud Computing
Big Data Analytics, CAP Theorem, DynamoDB, Parallel Processing
Distributed Computing, Hadoop, NoSQL, Stream Processing

Evaluation Scheme (July-Nov 2019)

- A1, A2, Q1: 10%, 15%, 5% Feb/March/Apr 14-24, 2022
- Mid-Semester Test: 30% (2H, Open Book) 12 Mar Sept 2022 (FN)
- Comprehensive Exam: 40% (2H, Open Book) 21 May 2021 (FN)

NOTE: All evaluation components are to be attempted individually. Plagiarism of any form is not accepted.

Thank You!

Thank you very much for your attention!

Queries ?