Big Data Security and Privacy

Presented
by
Sayali Vaidya and Priyanka N. Murthy
Spring 2017
University of Massachusetts, Lowell
220 Pawtucket St,Lowell,MA-01854

Outline

- Introduction & Background
- Architecture of Big Data
- Framework of Big Data
- Challenges of Big Data
- Security Techniques
- Privacy in Big Data
- Risk of Big Data and impact on IT
- Conclusion

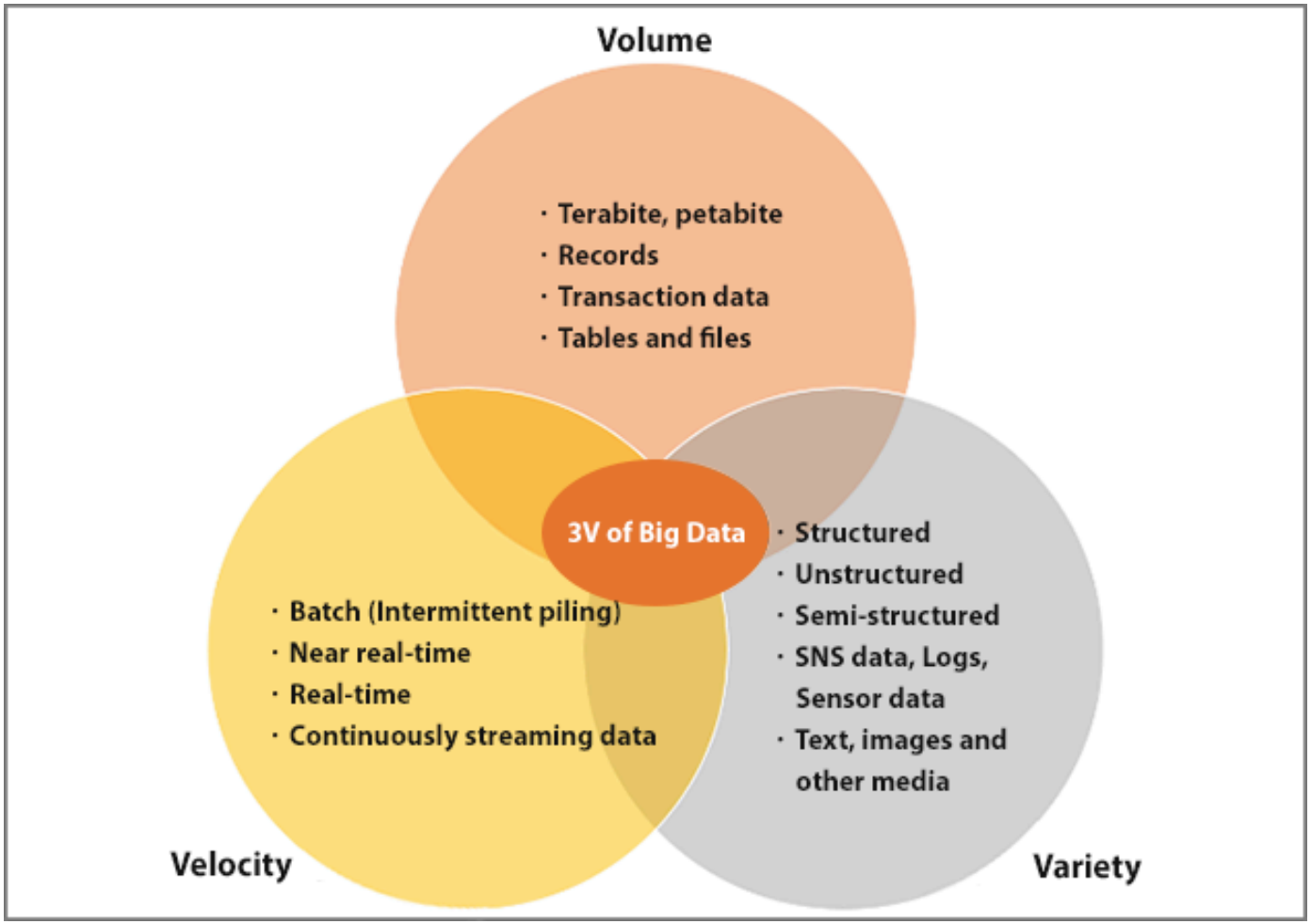
Introduction[1]

- Dramatic increase of data.
- From 2005 to 2020: increase 300 times, from 130 Exabytes to 40,000 Exabytes.(10^18 bytes)
- Generated from different Data Sources

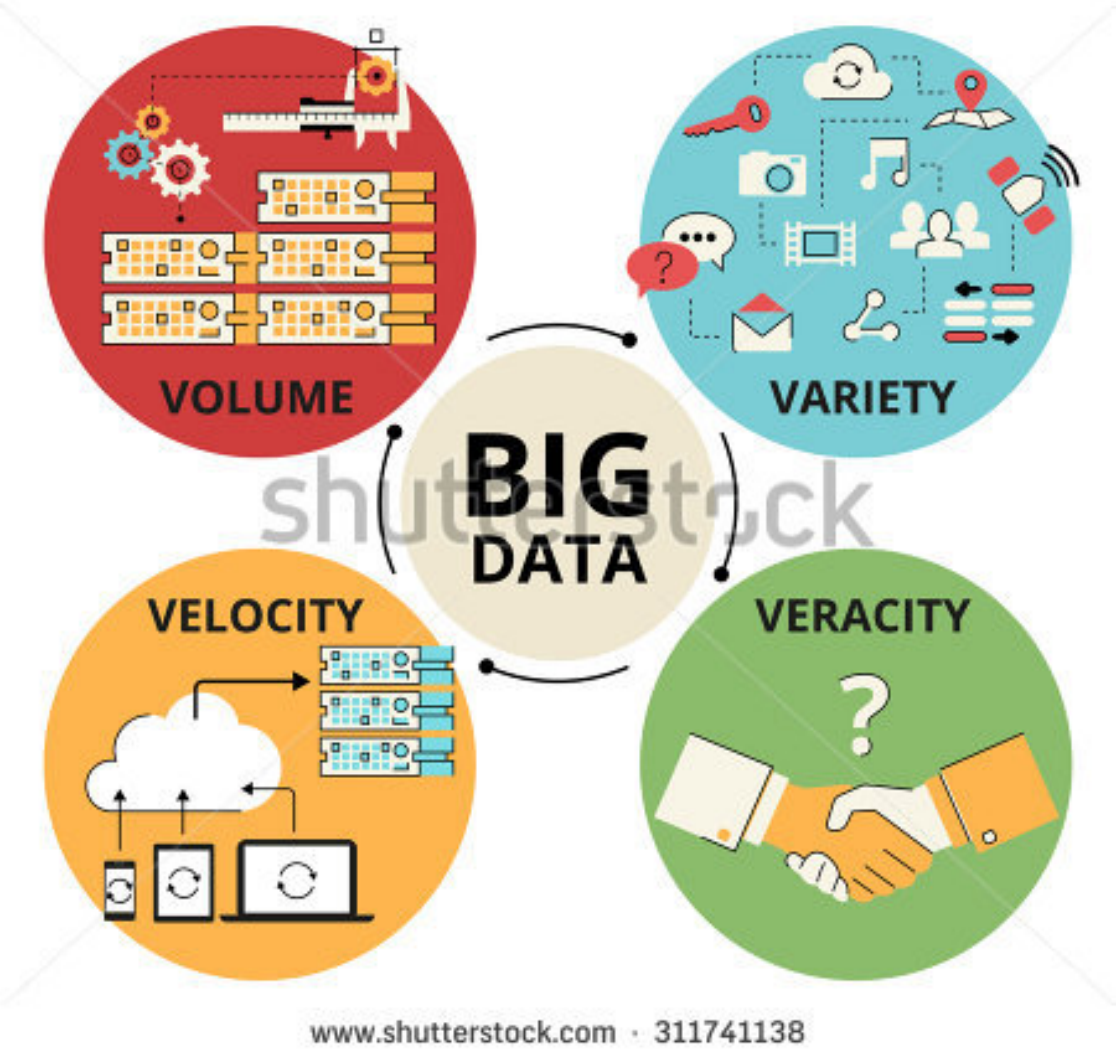
Definition-Big Data[1]

- The term Big Data remains vague.
- **Wikipedia Definition :** “Big Data is an all- encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications. ”
- Criteria for Big data

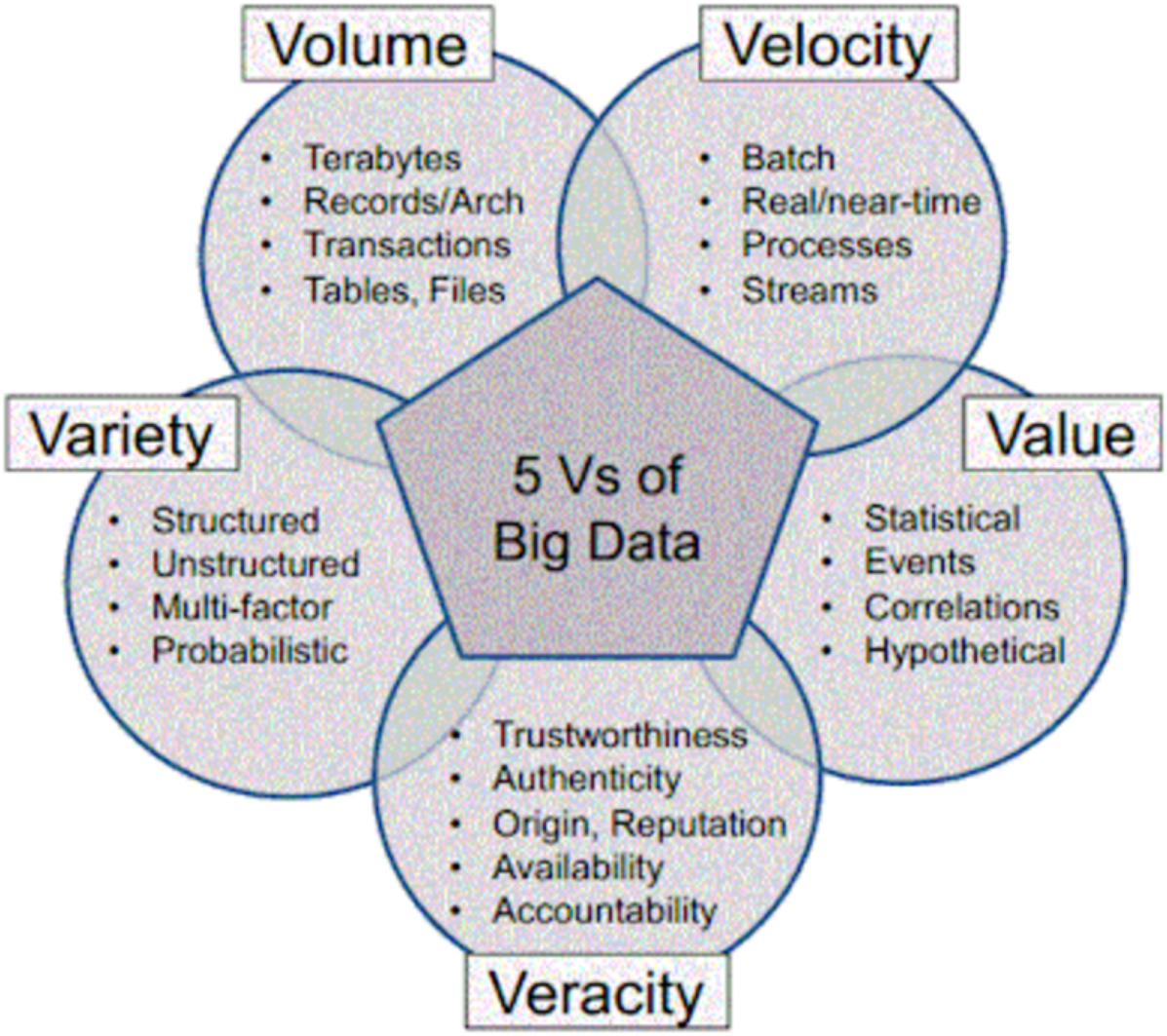
3V's Model


Volume
· Terabite, petabite
· Records
· Transaction data
· Tables and files
3V of Big Data
· Batch (Intermittent piling)
· Near real-time
· Real-time
· Continuously streaming data
· Structured
· Unstructured
· Semi-structured
· SNS data, Logs, Sensor data
· Text, images and other media
Velocity
Variety

4V's Model


VOLUME
VARIETY
BIG
DATA
VELOCITY
VERACITY
?
shutterstock
www.shutterstock.com · 311741138

5 V's Model


Volume
Terabytes
Records/Arch
Transactions
Tables, Files
Velocity
Batch
Real/near-time
Processes
Streams
Variety
Structured
Unstructured
Multi-factor
Probabilistic
5 Vs of Big Data
Value
Statistical
Events
Correlations
Hypothetical
Trustworthiness
Authenticity
Origin, Reputation
Availability
Accountability
Veracity

Motivation [1][2]

- Giga-Terabytes of new data generated daily.
- Increase of storage capacity
- Increase of processing power
- Availability of data

Applications of Big Data[1]

Example-1: Google

- Prediction of the outbreak of H1N1 disease

Example-2: Microsoft - Farecast

- Airline Ticket Price

Applications of Big Data[1]

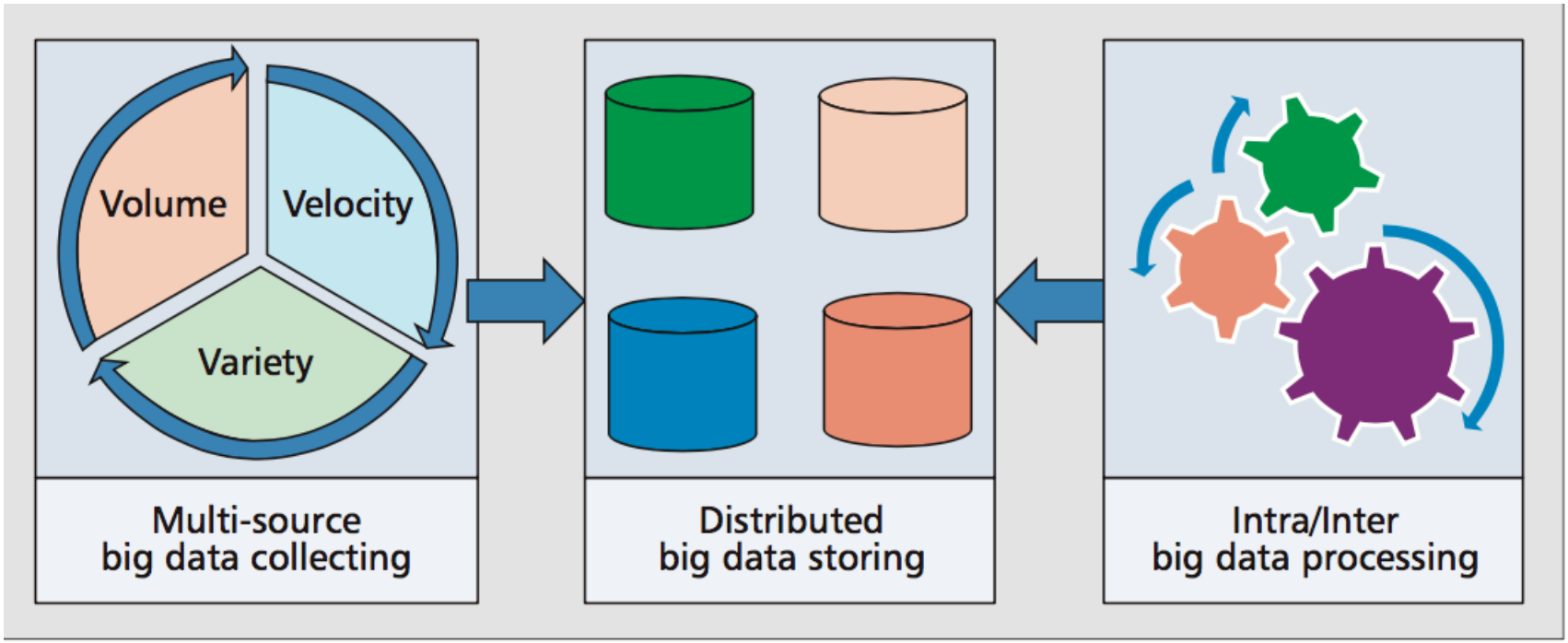
Example-3

❖Data Analysis in Politics

❖improved decision-making

❖Health care, employment, economic productivity, crime, security, natural disaster

Architecture of Big Data Analytics[4]



Multi-Source Bit Data Collecting[4]

- High volume, high velocity, and high variety.
- IBM - 2.5 quintillion bytes of created every day.
- Challenges
  - Efficiently store
  - Organize high-volume data
  - Quickly process streaming
  - Accurately analyze structured
  - Unstructured data

Distributed Big Data Storing & Intra/Inter Big Data Processing[4]

Distributed Big Data Storing

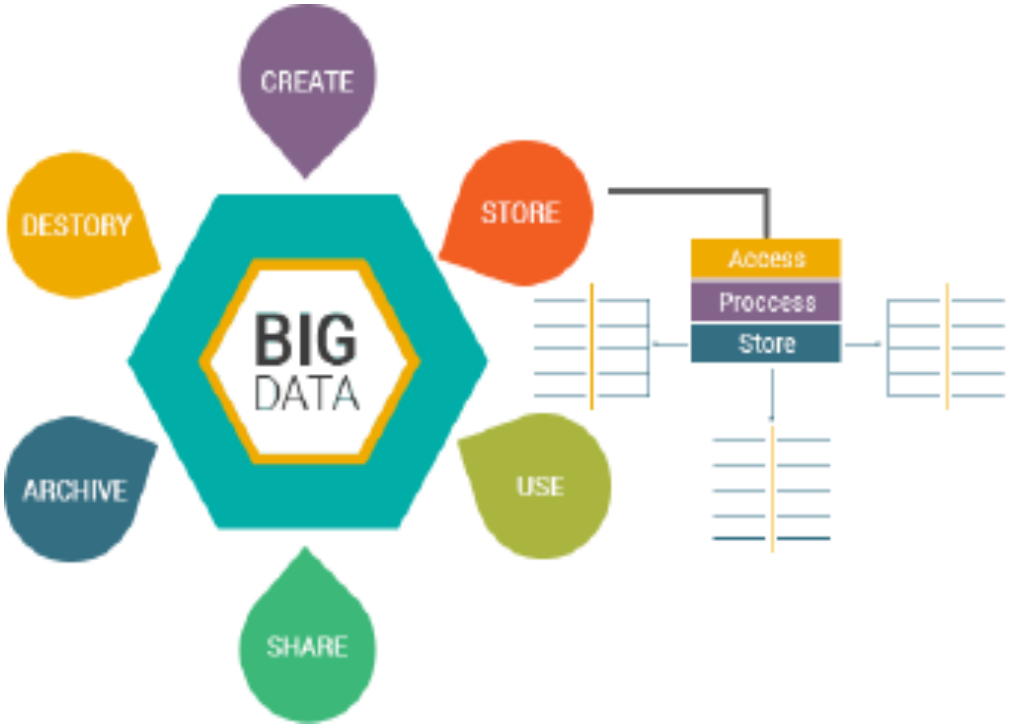
- ❖Limitation: Centralized storage
- ❖Solution : Distributed big data storage is suggested

Two types of Big Data Processing

- ❖Intra Big Data Processing
- ❖Inter Big Data Processing

Life Cycle of Big Data[6]

- Data Generation
- Data Collection
- Data Storage
- Data Management
- Data Processing
- Data Transmission

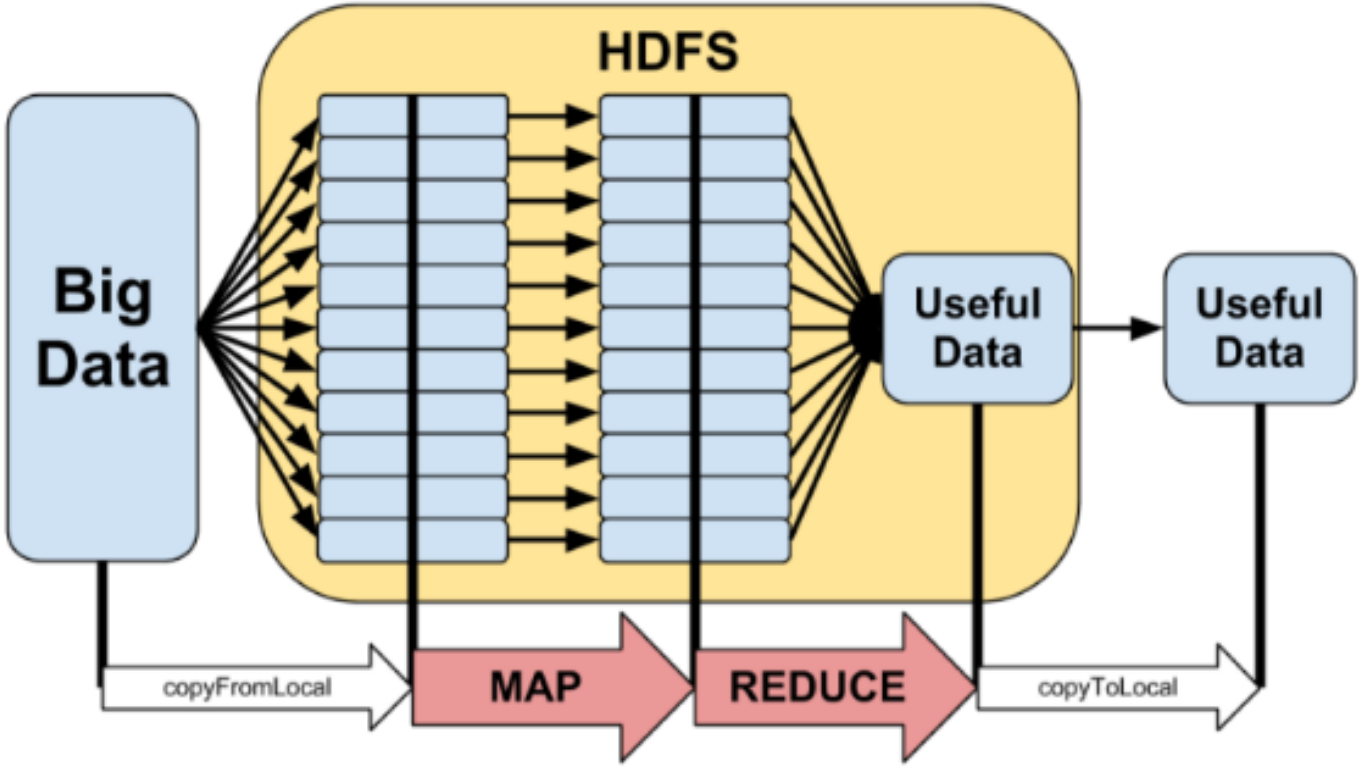


Framework of Big Data[1]

❖ **Framework for-Big data Analysis : Hadoop**

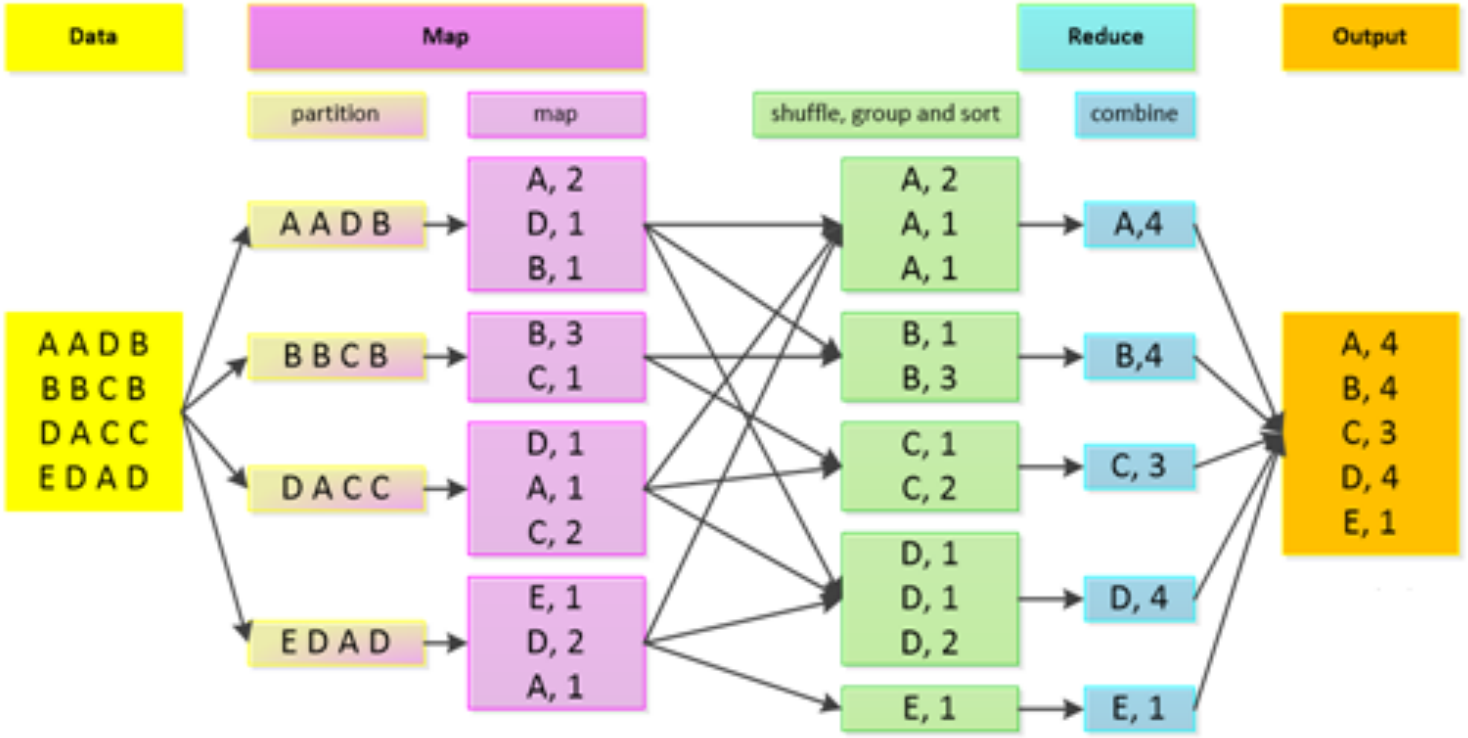
- Scalable
- Fault tolerant and reliable
- Easy to use

Map Reduce Framework



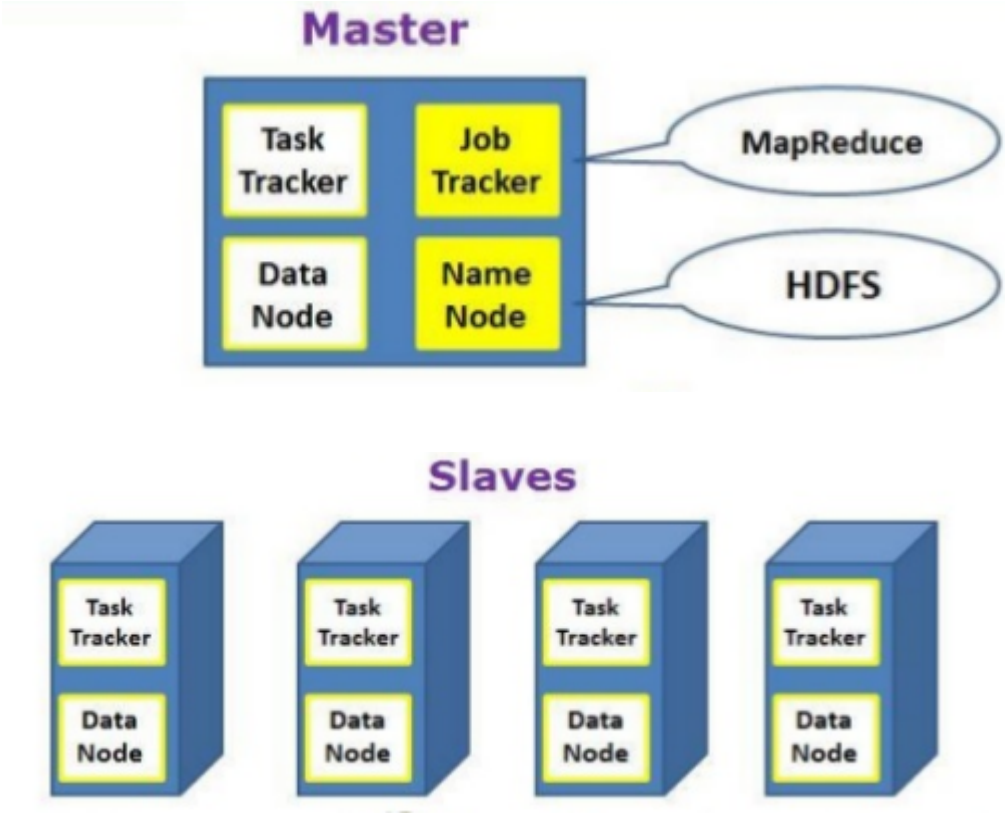
Map Reduce Example

Count the number of occurrences of each word in given input file



High Level Architecture of Hadoop

Master – Slave Architecture



Challenges of Big Data[1]

- Big Data is so large and complex .

- Security and privacy

Other challenges of Big Data[1]

The other challenges are listed as following

- Data preparation
- Efficient distributed storage and search
- Effective online data analysis
- Effective machine learning techniques
- Efficient handling of Big Data streams
- Semantic lifting techniques
- Programming models
- Social analytics

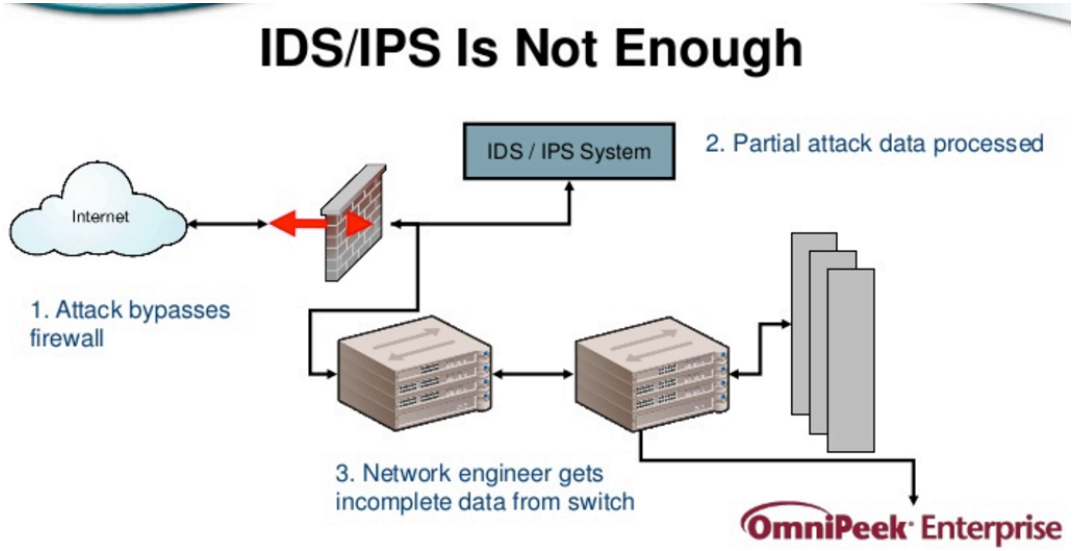
Big Data Security[3]

- Any comprehensive data security solution must meet three requirements

Cont'd..

- The security vulnerabilities inherited from -virtualization, IP, APIs

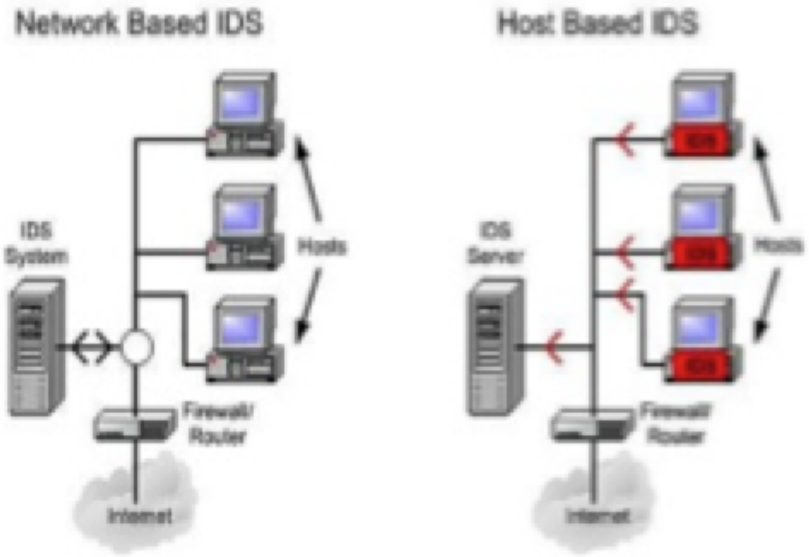
- Various schemes address these security issues
  - Encryption
  - Authentication
  - Access Control
  - Firewalls
  - IDS
  - DLPS



- Schemes should be integrated

Intrusion Detection Systems(IDS)[2]

- IDS
  - –Host based and Network based
  - –Signature Based and Anomaly Based
- CIDS
  - –Share alerts
  - –Share data,logs
  - –Share knowledge



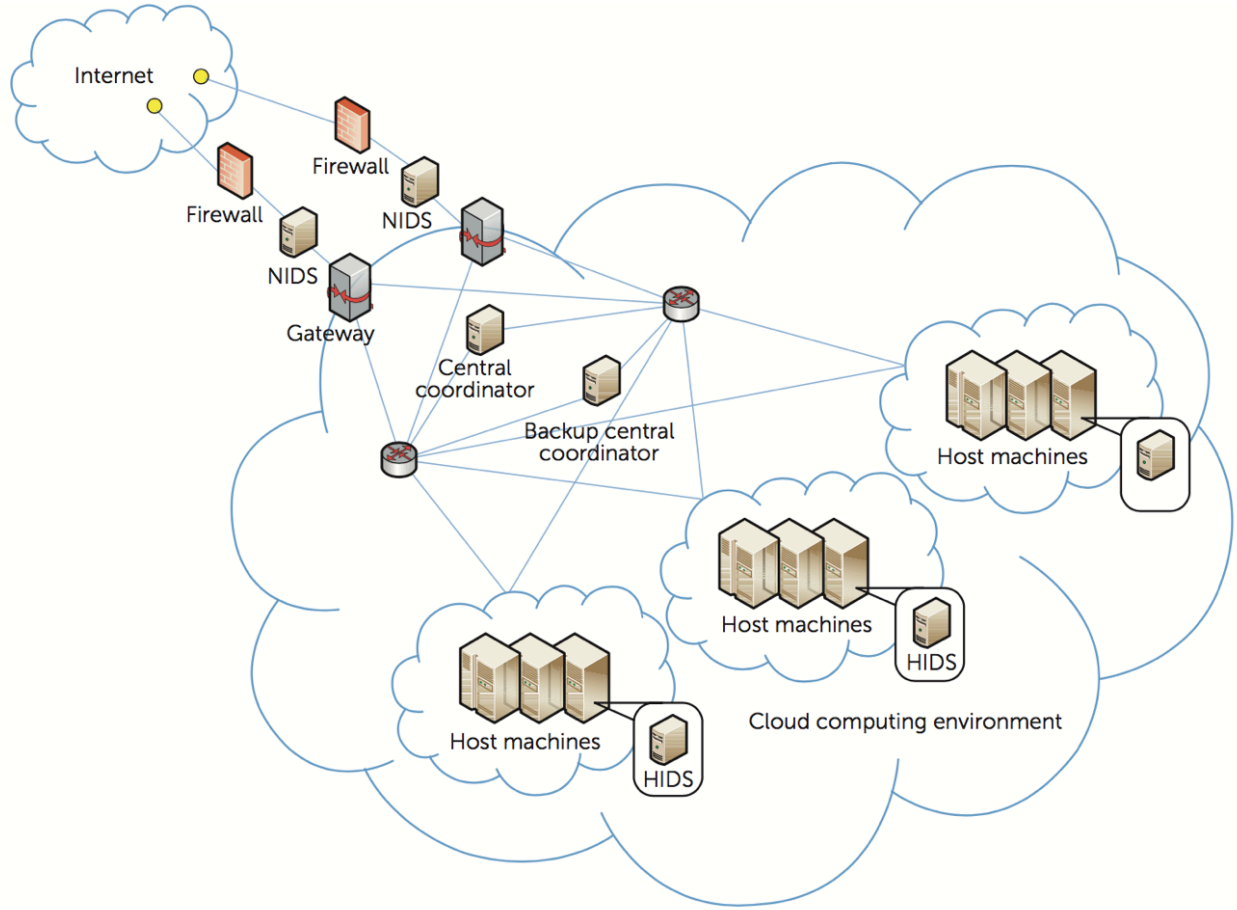
Enhanced Collaborative Intrusion Detection[2]

- Designed by Integrating Conventional IDSs.

- Collaborative intrusion detection systems (CIDSs)

  - Shares traffic information
  - organize IDSs within a CIDS in a decentralized manner.
  - IDSs communicate with each other or with Central Coordinator.
  - Central coordinator generates a complete attack diagram of the network.

Collaborative Intrusion Detection Framework Diagram[2]



Collaborative Intrusion Detection Framework [2]

- Co-operative Agents

  - Detect malicious behavior on networks.
  - Located on host machines are a new type of HIDS.
  - Comply with service-level agreements (SLAs) and legal restrictions.
  - Report intrusive behavior or activity.

Collaborative Intrusion Detection Framework [2]

- Central Co-ordinator

  - Network traffic aggregation
  - Captures sophisticated cooperative intrusions
  - Hybrid detection mechanisms
  - Enhance the detection accuracy of attacks.

Benefits Collaborative Intrusion Detection

Limitations of Collaborative Intrusion Detection[2]

- **Accuracy**
  - provides a minimal information gain

- **Efficiency**
  - denser traffic require additional computation to process summarization

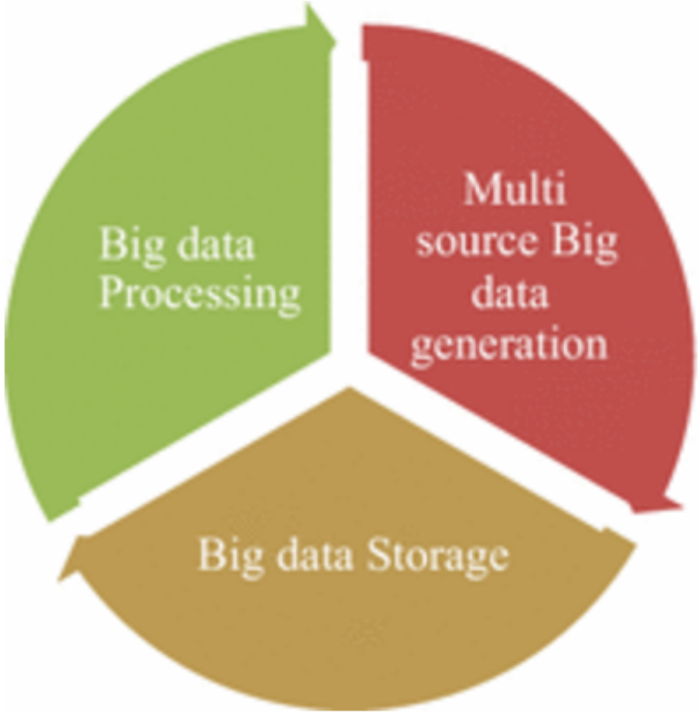
Big Data Privacy[1]

- ❖Amazon monitors shopping preferences.
- ❖Google learns our browsing habits.
- ❖Twitter knows what's on our minds.
- ❖Facebook seems to catch all that information too.

Privacy Requirements[4]

- Study of privacy in big data is still in its early stage

- Privacy requirements of big data analytics as follows



Privacy in Data Generation Phase[6]

- Active data generation and passive data generation.
  - Active data
  - Passive data
- Minimize the risk of privacy violation during data generation by
  - Access restriction
  - Falsifying data

Measures to Control Privacy Violation

- Access Control:
  - Passive Data - ensures privacy
    - Anti-tracking extensions
    - Advertisement/script blockers
    - Encryption tools.

  - In addition to these tools
    - anti-malware
    - anti-virus software

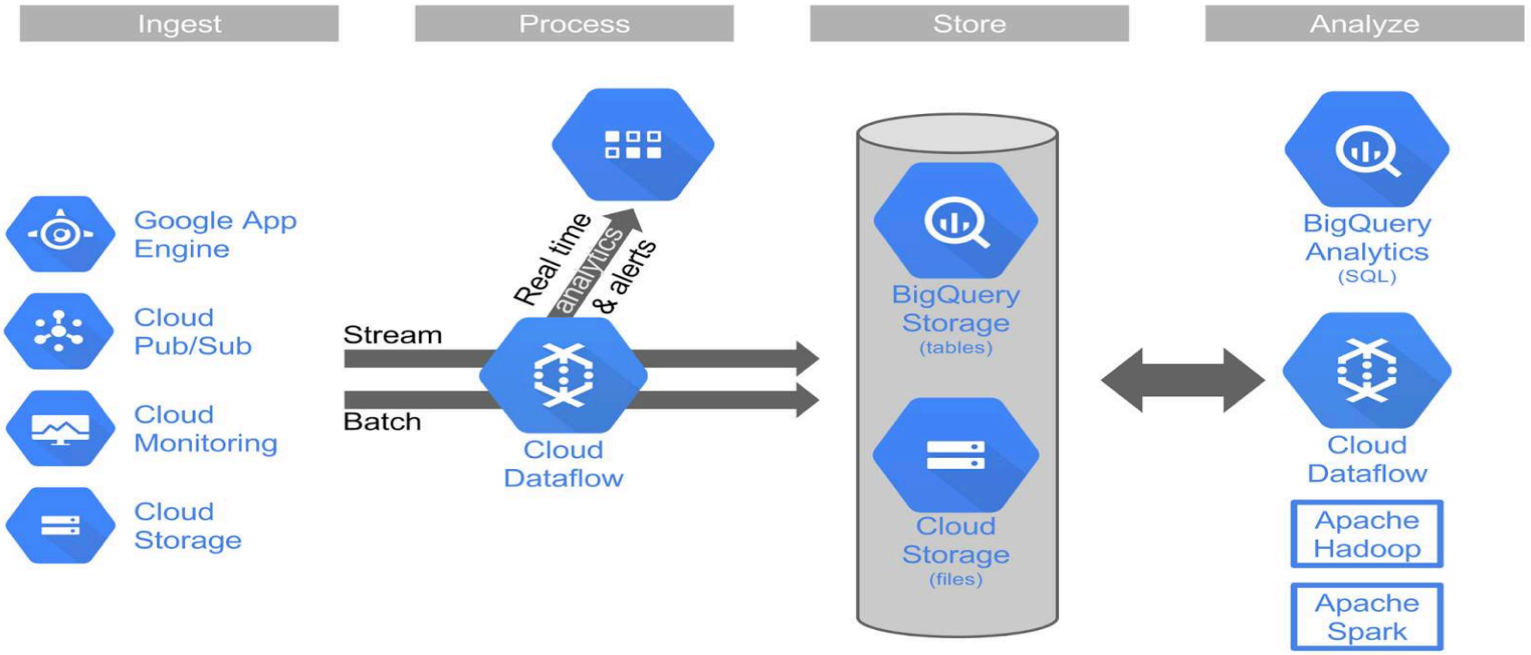
- Falsifying Data:

Distort data before it reaches the third party.

- **Socket Puppet:** hide online identity of individual by deception.
- **Mask Me:** mask individual's identity .

Privacy in Data Storage phase[6]

- Storage infrastructure should be scalable
- Configurable dynamically
- Solution: Virtual storage – empowered by cloud
- This could affect the privacy of the data.



Approaches to Privacy Preservation Storage on Cloud[6]

- The approaches to preserve the privacy are as follows
  - Attribute based encryption
  - Identity based encryption
  - Homomorphic encryption
  - Storage path encryption
  - Usage of hybrid clouds

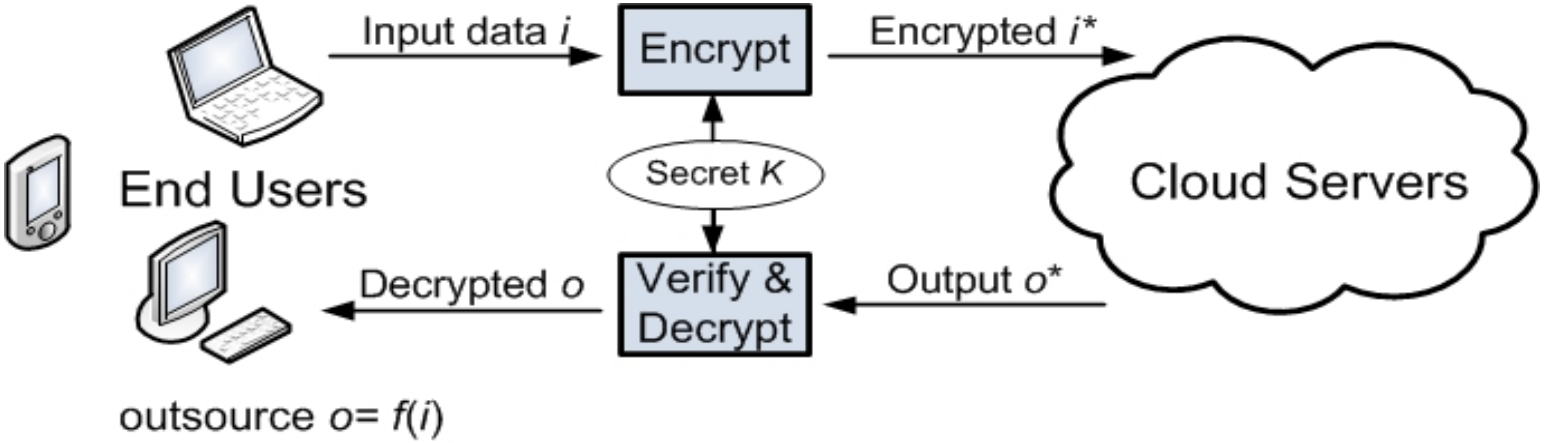


TABLE 1. Comparison of encryption schemes.

Encryption scheme	Features	Limitations
Identity based encryption	• Access control is based on the identity of a user • Complete access over all resources	• Time consuming in large environment • Granular access control is hard to implement • Changing ciphertext receiver is not possible • Data to be processed must be downloaded and decrypted
Attribute based encryption	• Access control is based on user's attribute • More secure and flexible as granular access control is possible	• Computational overhead in handling different user categories • Updating ciphertext receiver is not possible • Data to be processed must be downloaded and decrypted
Proxy re-encryption	• Can be deployed in IBE or ABE scheme settings • Updating Ciphertext receiver is possible	• Computational overhead • Data to be processed must be downloaded and decrypted
Homomorphic encryption	• Computations are performed on the encrypted data • Very secure	• Computational overhead is very high

Privacy preserving in Data Processing[6]

First phase: PPDP

Goal: Protect information from unwanted user viewing it

❖Identifier (ID)

❖Quasi-identifier (QID)

❖Sensitive attribute (SA)

❖Non-sensitive attribute (NSA)

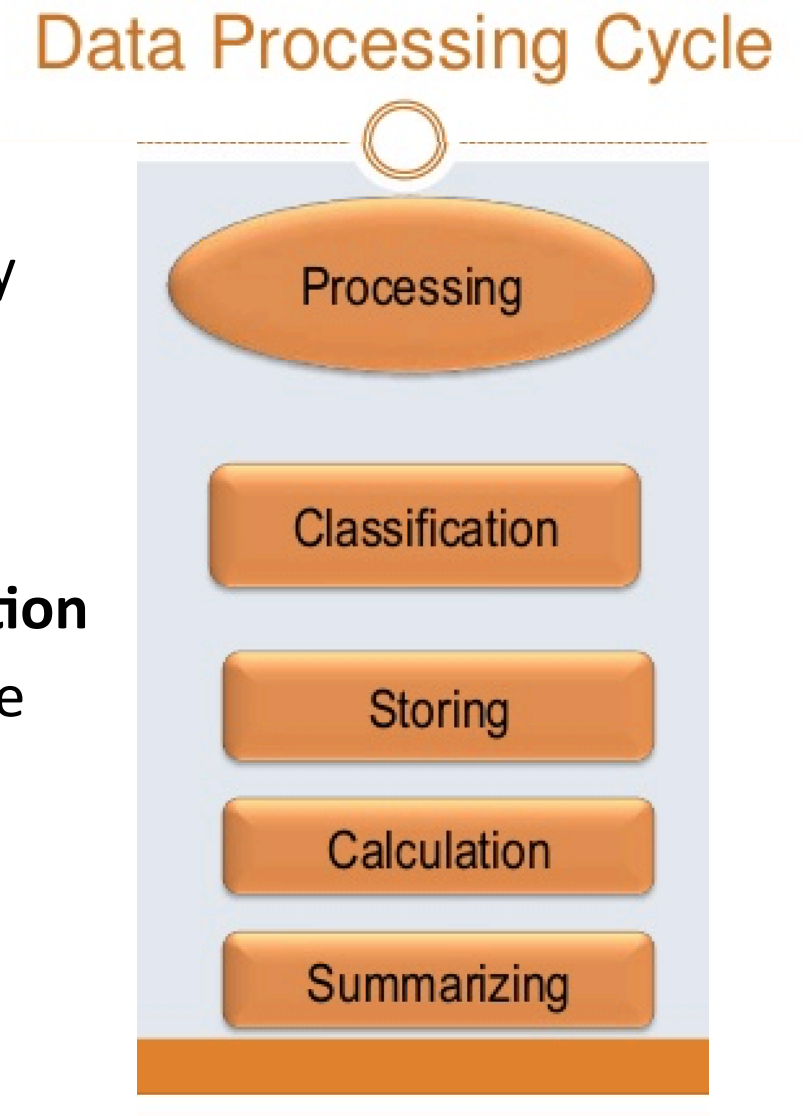
Privacy preserving in Data Processing[6]

Second phase
Goal: Extract the meaningful
Information without violating privacy

❖**Privacy Preserving Clustering**

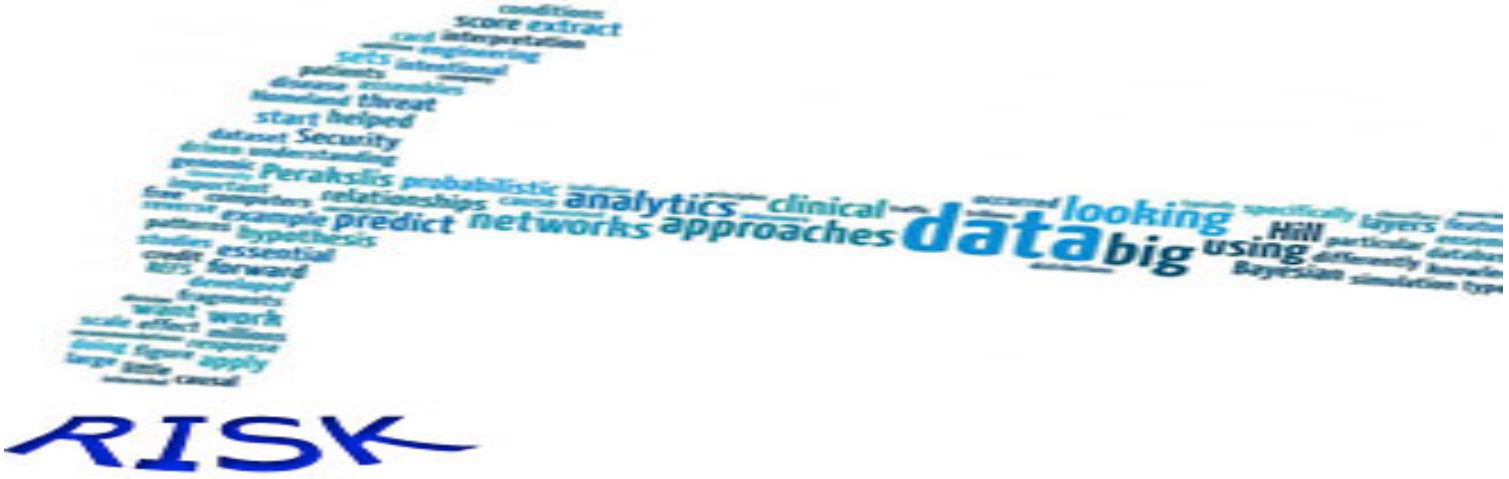
❖**Privacy Preserving Data Classification**

- Classification as a technique
- Classification algorithms



Risks of Big Data and Impact on IT

- Need the right people to solve problems
- Costs escalates too fast
- Challenges to IT organizations
- Creating billions of jobs in Big Data
- Require many data scientists , data analysts,data managers



Conclusion

- A review on Big Data Security and Privacy.
- Different approaches / framework /techniques for implementing Big Data security and privacy .
- Misuse of user's privacy if it is not properly handled.

References

[1] B. Matturdi, X. Zhou, S. Li and F. Lin, "Big Data security and privacy: A review," in *China Communications*, vol. 11, no. 14, pp. 135-145, Supplement 2014.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7085614&isnumber=7085375

[2] Z. Tan *et al*., "Enhancing Big Data Security with Collaborative Intrusion Detection," in *IEEE Cloud Computing*, vol. 1, no. 3, pp. 27-33, Sept. 2014.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7036256&isnumber=7036253

[3] E. Bertino, "Big Data - Security and Privacy," *2015 IEEE International Congress on Big Data*, New York, NY, 2015, pp. 757-761.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7207310&isnumber=7207183

[4] R. Lu, H. Zhu, X. Liu, J. K. Liu and J. Shao, "Toward efficient and privacy-preserving computing in big data era," in *IEEE Network*, vol. 28, no. 4, pp. 46-50, July-August 2014.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6863131&isnumber=6863119

[5] A. K. Tiwari, H. Chaudhary and S. Yadav, "A review on Big Data and its security," *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Coimbatore, 2015, pp. 1-5.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7193110&isnumber=7192777

[6] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua and S. Guo, "Protection of Big Data Privacy," in *IEEE Access*, vol. 4, no. , pp. 1821-1834, 2016.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7460114&isnumber=7419931

[7] http://www.ey.com/gl/en/services/advisory/ey-big-data-big-opportunities-big-challenges

Contd...

[8] A survey of security and privacy in big data .Haina Ye; Xinzhou Cheng; Mingqiang Yuan; Lexi Xu; Jie Gao; Chen Cheng
2016 16th International Symposium on Communications and Information Technologies (ISCIT)

[9] Challenges and solutions of information security issues in the age of big data Yang Mengke; Zhou Xiaoguang; Zeng
Xu Jianjian China Communications Year: 2016, Volume: 13, Issue: 3

[10] SmartGrids: MapReduce framework using Hadoop. Vaibhav Fanibhare; Vijay Dahake
2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN)

Thank you!