First Credit Seminar Presentation on "Privacy and Big Data: Issues and Challenges"

By,
Mr. Brijesh B. Mehta

Admission No.: D14CO002

Supervised By,
Dr. Udai Pratap Rao

Computer Engineering Department
S. V. National Institute of Technology, Surat

m.brijesh@coed.svnit.ac.in

21/11/2014

Overview

Introduction (Privacy)

- **Definition of Privacy:** Privacy is the ability of an individual or group to seclude themselves, or information about themselves, and thereby express themselves selectively[1]
- Personal information can be classified in four categories [2]
  - *Personally Identifiable Information (PII):* name and address
  - *Sensitive Information:* religion, health data
  - *Usage Data:* web usage
  - *Unique Device Identity:* MAC address, RFID tag

Introduction (Big Data)

- **Definition of Big Data:** Big data is an all-encompassing term for any collection of data sets so large and complex that it become difficult to process using traditional data processing applications [3]
- Big Data Characteristic[4]
  - Volume
  - Velocity
  - Variety
- Big Data Analytics [5]
  - Capture
  - Aggregate
  - Process

General Architecture of Big Data Analytics

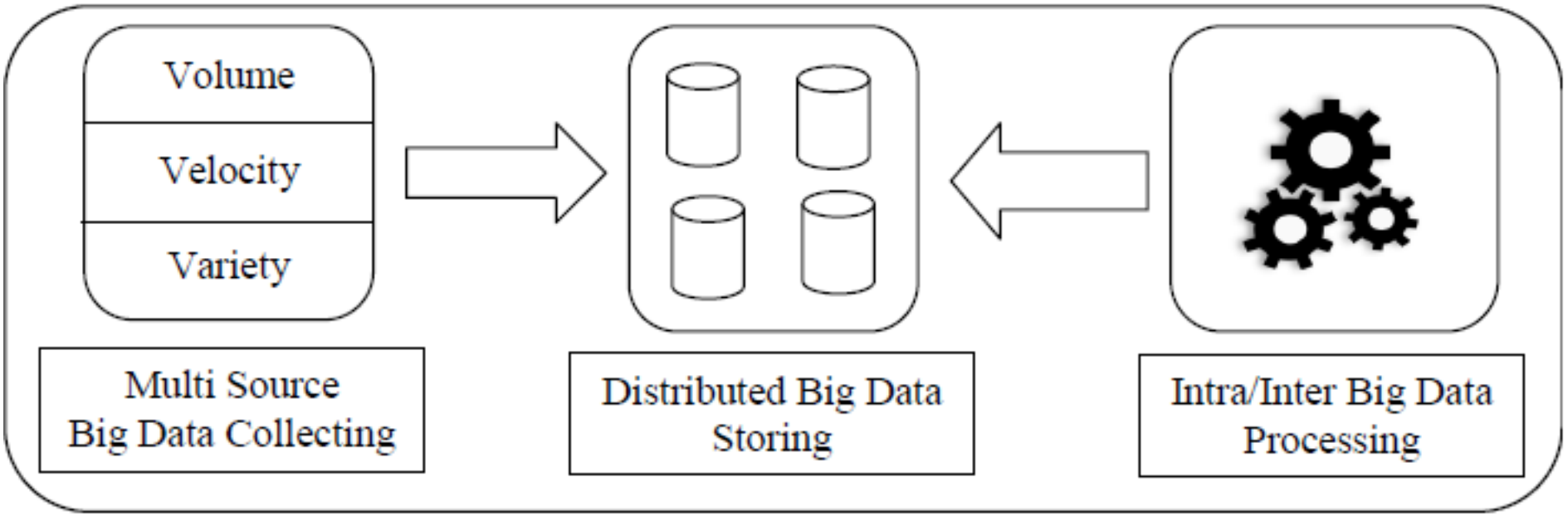

Fig 1 General Architecture of Data Analytics [4]

Multi Source Big Data Collecting

- Sources of big data [5]
  - Public Web and Social Media
  - Mobile Applications
  - Surveys
  - Traditional off-line documents scanned by optical character recognition in to electronic form
  - Sensors and Radio-Frequency Identification(RFID)chips
  - GPS chips
- Classification of online information [6]
  - **Born Analog** Created by use of some sensor or camera
  - **Born Digital** Created by use of computer
- **Data Fusion** aggregating multiple sources [6]

Intra/Inter Big Data Processing

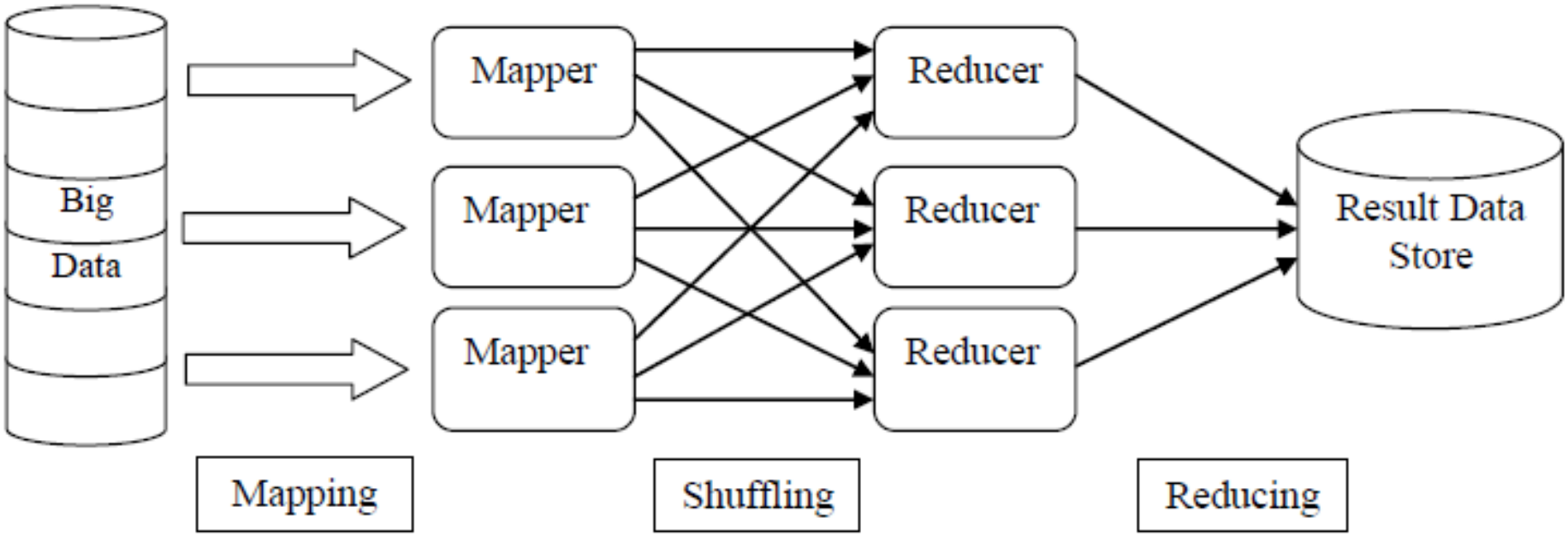

Fig 2 Map Reduce Architecture [7]

Intra/Inter Big Data Processing

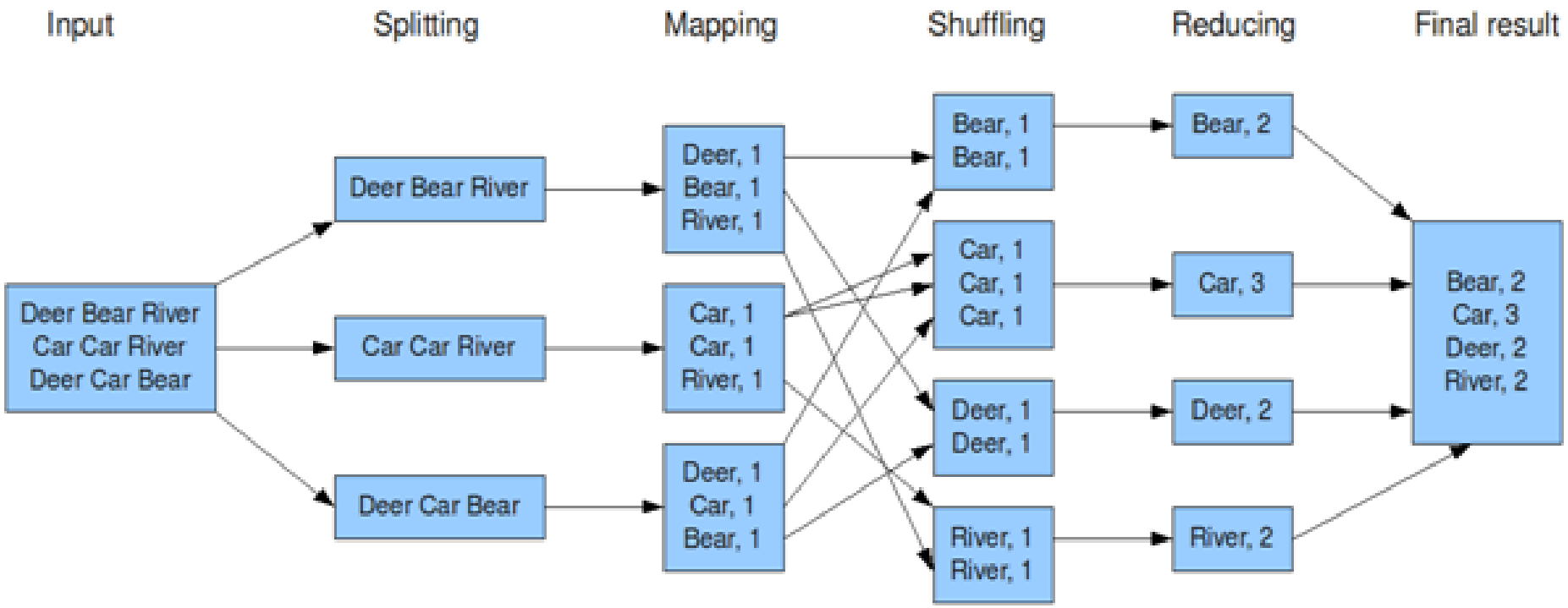

Fig 3 Word Count using Map Reduce [8]

Distributed Big Data Storing

- NoSQL database term stands for open source, distributed, and non-relational database[9]
- Most common characteristics of NoSQL databases are[10],[11]:
  - Simple and flexible non-relational data models
  - Ability to scale horizontally over many commodity clusters
  - Provide high availability
  - Most of them do not support ACID properties of Relational database, instead they support BASE properties (Basically Available, Soft state, Eventually consistence) [12]. However, couchDB[13] supports ACID properties.

NoSQL Data Models

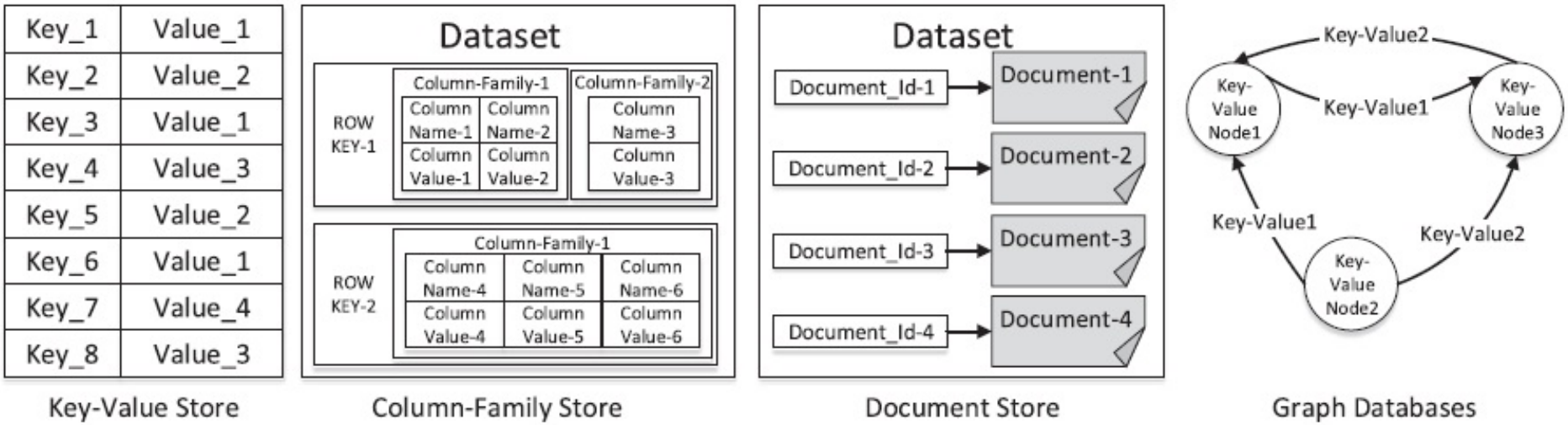

Fig 4 NoSQL Data Models [14]

- NoSQL Data Models and its members:
  - *Key-Value Stores:* Memcached, Redis, BerkeleyDB, Voldemort, Riak
  - *Column-Family Stores:* Bigtable, Hadoop HBase, SimpleDB, Cassandra
  - *Document Stores:* CouchDB, Couchbase server, MongoDB
  - *Graph Databases:* Neo4j

Privacy Issues in Big Data

- Edward Snowden Vs NSA [15]
- Surveillance program has divided data into two part:
  - Content Data
  - Context Data (Meta Data)

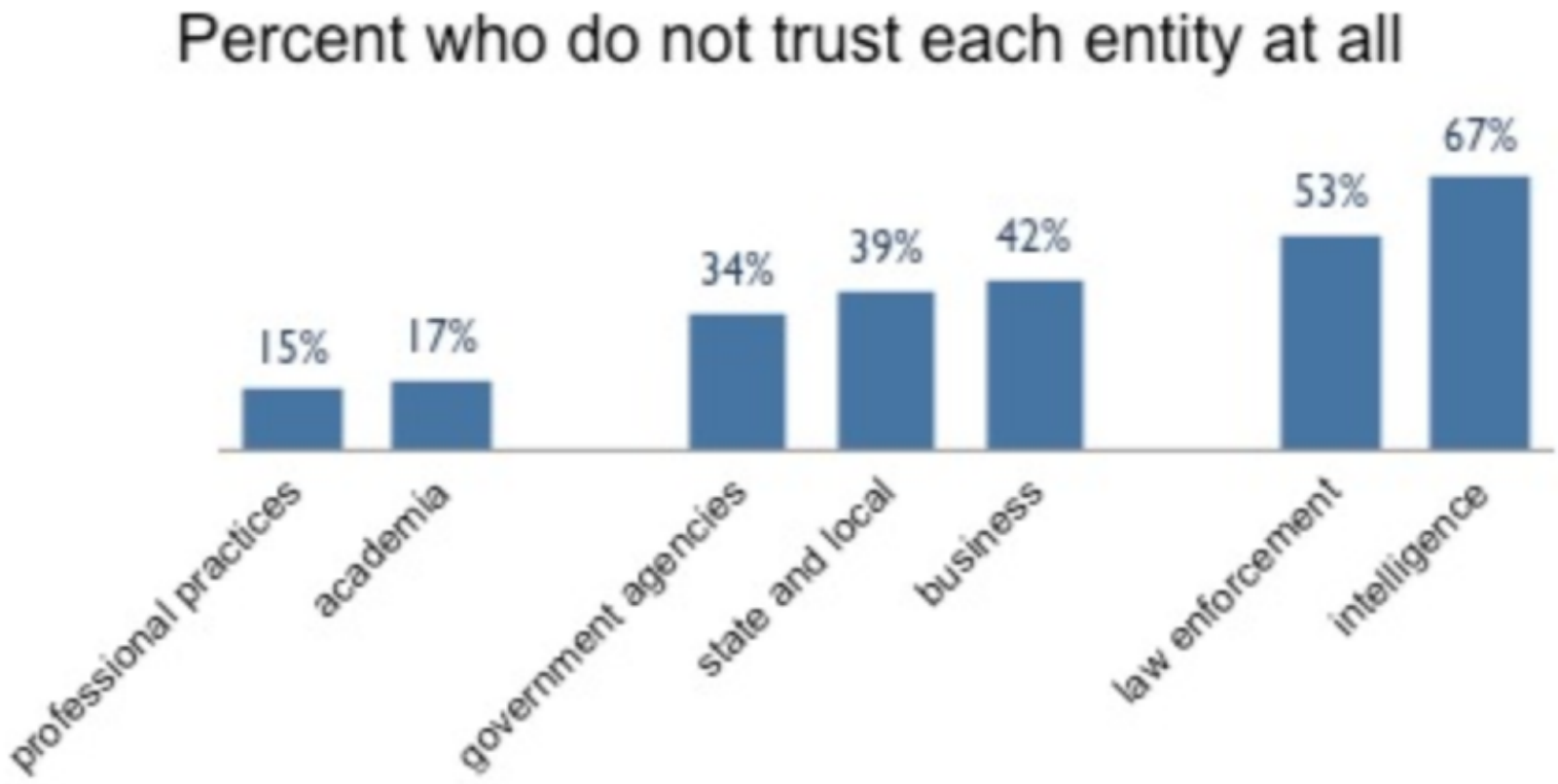

Fig 5 White House Survey Results [16]

Privacy in Big Mobile Data

- Mobile location data reveal more information about a user then any other data.
- Over-collection of mobile data is one of the major concern
  - Brightest flash light android app [17]
- Lee Garber[18] has mention that, Bit Defender, a Romanian security vendor, has analyzed 836,021 Android applications on Google Play Store and found
  - about 33% of apps could reveal location-related data
  - about 5% located and opened photos on user's phone
  - approximately 3% reveal users email
- Mirco Musolesi[19], has mentioned that In June 2013, Facebook had, on average, 819 million monthly active mobile users

Research Challenges with Privacy and Big Data (Privacy and Data Mining[21]) I

- Data Provider
  - Limit the access
  - Trade privacy for benefit
  - provide false data
- Data Collector, anonymize the data before sending it to data miner, this technique is called rivacy preserving data publishing
  - In PPDP, attributes of the table is divided as,
    - *Personal Information Identifier(PII):* Unique ID, Name, Mobile Number
    - *Quasi-identifier(QID):* Gender, Age, Zip(Postal) Code
    - *Sensitive Attribute(SA):* Disease, Salary
    - *Non-sensitive Attribute*

Research Challenges with Privacy and Big Data (Privacy and Data Mining[21]) II

- Some of the following anonymization operations[20] may apply on the data
  - Generalization
  - Suppression
  - Anatomization
  - Permutation
  - Perturbation
- Data Miner
  - Privacy preserving association rule mining
  - Privacy preserving classification
  - Privacy preserving clustering
- Decision Maker
  - Data provenance
  - Web information credibility

Research Challenges with Privacy and Big Data (Top Ten Big Data Security and Privacy Challenges[22]) I

- Big Data Working Group[22] at Cloud Security Alliance has listed top ten big data security and privacy issues as,
  1. Scalable and composable privacy preserving data analytics
  2. Cryptographically enforced data centric security
  3. Granular access control
  4. Secure computations in distributed programming frameworks
  5. Security best practices for non-relational data stores
  6. Secure data storage and transactional logs
  7. Granular audits
  8. Data provenance
  9. End-point validation and filtering
  10. Real time security monitoring

Research Challenges with Privacy and Big Data (Top Ten Big Data Security and Privacy Challenges[22]) II

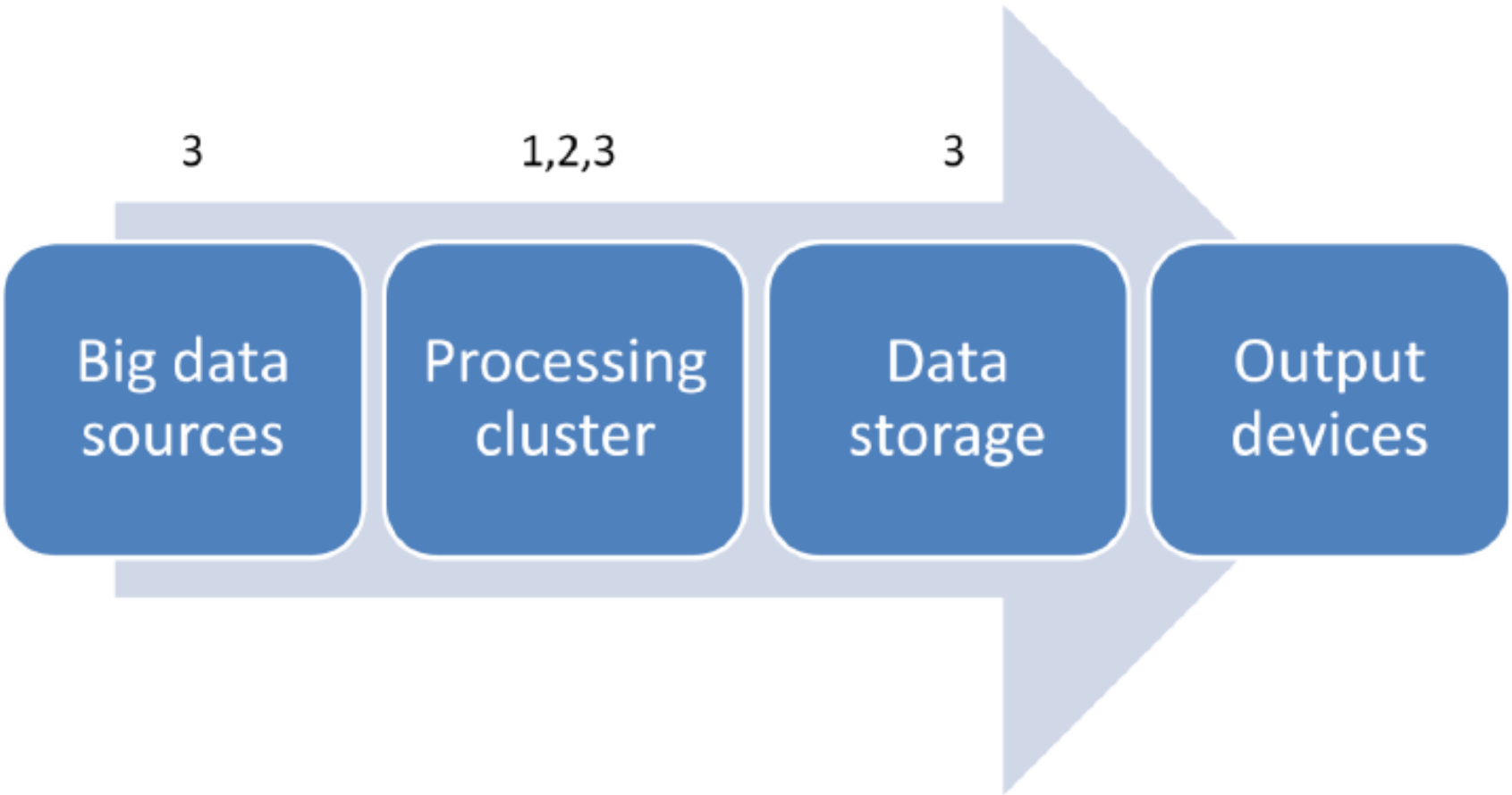

Fig 6 Big Data Eco-system [22]

Existing Privacy Preserving Techniques and their Limitations

Name of the Techniques	**Short Description**	**Limitations with respect to Big Data**
Privacy Preserving Data Publishing (Anonymization) (De-identification)	Generalization, Suppression, Anatomization, Permutation, and Perturbation are some of the techniques used in PPDP	Works for structured data only
Privacy Preserving Data Mining	Privacy Preserving Association Rule Mining, Privacy Preserving Classification, Privacy Preserving Clustering	To find and remove sensitive rule from large amount of data is not feasible
Differential Privacy	Noisy data is being added into the mined results	Risk of re-identification because same data is collected from many sources
Privacy Preserving Aggregation (Homomorphic Encryption)	Mining operation performed on encrypted data only	It is function specific so we need to write different functions for different tasks which is not feasible

Conclusion and Future Work

- Some of the existing privacy preserving techniques with little modification can be useful to privacy and big data
- As per big data working group, Homomorphic encryption and differential privacy are some of the promising technologies for preserving privacy in big data.

- In future, the detailed analysis of privacy preserving techniques like, homomorphic encryption, differential privacy, and de-identification, for privacy and big data will be presented.

References I

[1] "Definition of Privacy." [Online]. Available: http://en.wikipedia.org/wiki/Privacy

[2] S. Pearson, "Taking account of privacy when designing cloud computing services," in *Proceedings of the International Conference on Software Engineering: Workshop on Software Engineering Challenges of Cloud Computing*. IEEE, 2009, pp. 44–52. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5071532

[3] "Definition of Big Data." [Online]. Available: http://en.wikipedia.org/wiki/Big_data

[4] R. Lu, H. Zhu, X. Liu, J. Liu, and J. Shao, "Toward efficient and privacy-preserving computing in big data era," *IEEE Network*, vol. 28, no. 4, pp. 46–50, Jul. 2014. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6863131http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6863131

[5] Executive Office of the President, "Big Data: Seizing Opportunities, Preserving Values," Executive Office of the President, Washington, D.C., Tech. Rep. May, 2014. [Online]. Available: http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

References II

[6] President's Council of Advisors on Science and Technology, "Big Data and Privacy : A Technological Perspective," Executive Office of President, Washington D.C., Tech. Rep. May, 2014. [Online]. Available: http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf

[7] M. Dagli and B. Mehta, "Big Data and Hadoop: A Review," *International Journal of Advanced Research in Engineering and Science*, vol. 2, no. 2, pp. 192–196, 2014. [Online]. Available: http://arph.in/ijares/wp-content/uploads/2014/02/16.pdf

[8] "MapReduce: Word Count Example." [Online]. Available: http://xiaochongzhang.me/blog/wp-content/uploads/2013/05/MapReduce_Work_Structure.png

[9] "NoSQL Meet Up by Event Brite." [Online]. Available: http://www.eventbrite.com/e/nosql-meetup-tickets-341739151

[10] R. Hecht and S. Jablonski, "NoSQL evaluation: A use case oriented survey," in *Proceedings of the International Conference on Cloud and Service Computing*. IEEE, Dec. 2011, pp. 336–341. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6138544

[11] P. J. Sadalage and M. Fowler, *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence*. Addison-Wesley, 2012. [Online]. Available: http://books.google.co.in/books?id=AyY1a6-k3PIC

References III

[12] D. Pritchett, "BASE: AN ACID ALTERNATIVE," *Queue*, vol. 6, no. 3, pp. 48–55, May 2008. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1394127.1394128

[13] "CouchDB Database." [Online]. Available: http://couchdb.apache.org/

[14] K. Grolinger, W. a. Higashino, A. Tiwari, and M. A. Capretz, "Data management in cloud environments: NoSQL and NewSQL data stores," *Journal of Cloud Computing: Advances, Systems and Applications*, vol. 2, no. 1, pp. 1–22, 2013. [Online]. Available: http://www.journalofcloudcomputing.com/content/2/1/22

[15] "Global Surveillance Disclosures by Edward Snowden." [Online]. Available: http://en.wikipedia.org/wiki/Global_surveillance_disclosures_(2013present)

[16] "Big Data Survey." [Online]. Available: http://www.whitehouse.gov/issues/technology/big-data-review

[17] "Brightest Flash Light Android App." [Online]. Available: https://play.google.com/store/apps/details?id=goldenshorestechnologies.brightestflashlight.free&hl=en

[18] L. Garber, "Security, Privacy, Policy, and Dependability Roundup," *IEEE Security & Privacy*, vol. 12, no. 2, pp. 11–13, Mar. 2014. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6798617

References IV

[19] M. Musolesi, "Big Mobile Data Mining: Good or Evil?" *IEEE Internet Computing*, vol. 18, no. 1, pp. 78–81, Jan. 2014. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6756891

[20] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing," *ACM Computing Surveys*, vol. 42, no. 4, pp. 1–53, Jun. 2010. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1749603.1749605

[21] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, "Information Security in Big Data: Privacy and Data Mining," *IEEE Access*, vol. 2, pp. 1–28, 2014. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6919256

[22] Big Data Working Group, "Expanded Top Ten Big Data Security and Privacy Challenges," Cloud Security Alliance, Tech. Rep. April, 2013.

Thank You