# Understanding Privacy Violations in Big Data Systems

# Abstract

- Big data systems have been instrumental in solving computational problems for business intelligence and predictive analysis.
- Despite this, they exhibit serious concerns for user privacy. The authors provide an overview of privacy in the context of big data, categorizing four types of existing privacy violations in big data systems and assessing the strengths and weaknesses of their protection techniques.
- They also provide measures that can be taken to strengthen users’ privacy.

# Existing

- Big data applications have helped analyze and solve many data-science problems for businesses and governments alike. Governments have used big data applications to identify criminals, detect terrorist activities, and enhance citizen services.
- For example, in a smart city, vehicle movement can be tracked through sensors to determine volumes and patterns of traffic.
- This information can then be linked with vehicle owner information to determine the relationships between age groups and their travel times and locations. This analysis can then be used for improved city planning.

# Contd..

- Similarly, corporate organizations use big data to improve the customer experience, generate revenue, and provide cost-effective solutions.
- For example, a department store can keep track of customer spending and determine relevancy between types of products purchased at the store and their relationship to customers’ age groups. The store can then focus on popular items that will increase sales.

# Disadvantage

- Despite these benefits, there are increasing concerns that the information collected by government agencies and corporate organizations can lead to leakage of private and confidential information.

# Proposed

- From social networks to financial transactions and shopping records, a large amount of data is consistently being collected, integrated, and analyzed.
- Data analysis is extremely useful for fore-casting and predictions, but it has also led to increased concerns about and violations of privacy.
- Broadly speaking, a privacy violation is an undesired leakage, exposure, or inference of private or confidential information.
- Big data systems are susceptible to privacy violations primarily be-cause of their large and continually growing datasets. As more data becomes available,

# Contd..

- a user’s confidential information can be collected directly from a single source or be gathered indirectly through meticulous linking of data from multiple sources.
- An email provider might automatically scan emails from users to infer confidential information, and a user might agree to the provider’s privacy terms without realizing that her emails are being analyzed. This is an example of a direct violation.
- In contrast, if data from the user’s emails is linked with the user’s data from another source, such as web search, then more information about the same user can be assessed. This is an example of an indirect violation.

# Advantage

- It is imperative that methods be adopted to preserve and protect confidential information in big data systems

# HARDWARE REQUIREMENTS

| | | |
|---|---|---|
| Processor | – | Pentium –III |
| Speed | – | 1.1 Ghz |
| RAM | – | 256 MB(min) |
| Hard Disk | – | 20 GB |
| Floppy Drive | – | 1.44 MB |
| Key Board | – | Standard Windows Keyboard |
| Mouse | – | Two or Three Button Mouse |
| Monitor | – | SVGA |

# SOFTWARE REQUIREMENTS

- Operating System : Windows 8
- Front End : Java /DOTNET
- Database : Mysql/HEIDISQL

# Conclusion

- Although the importance of big data systems has been established for analytics and prediction, it is imperative that methods be adopted to preserve and protect confidential information in big data systems.
- Substantial and cohesive efforts are needed to achieve this important goal.

# Reference

1. P.A. Laplante, “Who's Afraid of Big Data?,” IT Professional, vol. 15, no. 5, 2013, pp. 6–7.
2. A. Narayanan and V. Shmatikov, “Robust De-Anonymization of Large Sparse Datasets,” IEEE Symp. Security and Privacy(SP), 2008, pp. 111–125.
3. M. De Goede, “The Politics of Privacy in the Age of Preemptive Security,” Int'l Political Sociology, vol. 8, no. 1, 2014, pp. 100–104.
4. D. Barth-Jones, “The ‘Re-identification’ of Governor William Weld’s Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now,” SSRN, July 2012; https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2076397.
5. M. Lindh and J. Nolin, “Information We Collect: Surveillance and Privacy in the Implementation of Google Apps for Education,” European Educational Research J., vol. 15, no. 6, 2016, pp. 644–663.