WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Privacy-Preserving Big Data in an In-memory Analytics Solution

Gurlovjot Singh Malhi

M. Tech, Department of Engineering and Technology, CT University Punjab

ABSTRACT:

In today's information-driven society, an immense volume and diverse array of data are constantly generated, facilitated by technological progress. Commercial enterprises have been swift to embrace this transformation, employing a wide array of information systems to bolster their operations. As the number of these systems grows, so does data production, further amplified by the proliferation of social networks. This avalanche of data has given rise to the term "big data." Big data is characterized by several key attributes, with volume, velocity, variety, and veracity being the most significant. Within the realm of big data analytics, the gathering, purification, processing, and analysis of copious and diverse data from various sources are executed to inform decision-making and problem-solving. However, certain scenarios require these capabilities to operate in real-time, resulting in what is known as real-time big data analytics. While this approach offers immediate insights, it poses considerable challenges in terms of implementation and operational demands. Furthermore, it introduces a new dimension of security concerns, particularly with respect to preserving privacy when disclosing data, especially in analytical contexts where precision is of paramount importance. In summary, privacy holds immense significance, as the inadvertent exposure of sensitive data can lead to severe consequences. Thus, the objective of this thesis is to explore multiple models for safeguarding privacy in an in-memory-based real-time big data analytics solution. Subsequently, the aim is to assess and analyze the outcomes, ultimately proposing an optimal model that aligns with privacy requirements while upholding the analytical integrity of the solution. The results reveal that a newly developed model, leveraging the inherent capabilities of such an environment, effectively fulfills all requirements, with a particular emphasis on maintaining high data accuracy.

Keywords: real time, big data analytics, In-memory, data utility, privacy preservation, analytical privilege, SAP HANA

Introduction:

Personally identifiable information (PII) or Sensitive Personal Information (SPI) is a concept utilized in the United States. According to the National Institute of Standards and Technology (NIST), PII is defined as follows [1]:

"" Personally Identifiable Information(PII) encompasses all data related to an existent within the records of an association. This includes(1) any information that can be used to discern or track an existent's identity, similar as their name, social security number, date and place of birth, mama 's maiden name, or biometric records; and(2) any fresh data that's connected or potentially linkable to an individual, including information regarding their medical, educational, fiscal, and employment history. analogous generalities live in colorful countries, all with the participated end of classifying particular information and latterly enforcing measures to cover its confidentiality, thereby upholding the principles of sequestration."

If data includes sensitive details like names, ages, gender, marital status, and blood groups, it is likely that they will be categorized as having a "high" or "very high" level of sensitivity. This implies that the overall protection level assigned to them will also be set as high or very high. Additionally, there are other factors that can influence the classification process, such as regulatory mandates. The determined data protection levels, in turn, necessitate the implementation of a range of techniques to safeguard data. In the case of sensitive data, maintaining privacy, particularly during data dissemination, can become a pivotal objective for an organization. Diverse privacy-preserving methods can be employed to shield sensitive data.

Nonetheless, real-time big data analytics presents fresh challenges due to its instantaneous nature, as well as considerations related to other key characteristics of big data, including volume, velocity, variety, and veracity [3]. This also implies that instituting a privacy preservation framework at the data level can be demanding within such an environment. Thus, the core goal of this thesis is to explore a variety of approaches for preserving privacy. Consequently, a model will be introduced, with the added requirement that this model will not compromise the real-time prerequisites of the environment or any other specific data-related necessities.

What is Big Data Analytics?

Big data analytics, as the name implies, is employed for handling extremely large datasets, and the typical implementation of big data involves the collection, storage, management, and analysis of extensive data sets [8]. The tools and technologies employed for each of these tasks can vary, with both

commercial and open-source applications available in today's market. Hadoop is among the most commonly used applications and is often associated with big data. However, Hadoop is typically utilized in the context of batch processing, leading to significant latency, which results in delays

in data storage and processing. To meet the demands of real-time processing, a distinct approach is required to address these latency issues.

What is the use of Big Data Analytics?

- **Business Intelligence:** Big data analytics helps organizations gain valuable insights from their data to make more informed business decisions. It can uncover trends, patterns, and correlations that might be otherwise hidden, enabling companies to improve their operations, marketing strategies, and customer experiences.
- **Predictive Analytics:** Big data analytics can forecast future trends and behaviors based on historical data. This is crucial for businesses to anticipate customer needs, optimize inventory management, and reduce risks.
- **Customer Insights:** Understanding customer behavior and preferences is essential for marketing and product development. Big data analytics helps businesses segment their customers, personalize marketing efforts, and enhance customer satisfaction.
- **Fraud Detection:** Big data analytics is used in financial institutions and e-commerce to detect and prevent fraudulent activities. It can identify unusual patterns or deviations from the norm in real-time, minimizing financial losses.
- **Healthcare:** In the healthcare sector, big data analytics is used to improve patient care and outcomes. It can be employed for disease tracking, medical research, and personalizing treatment plans.

Methodology:

The research methodology closely aligned with the research questions is the Design Science Research Methodology (DSR). Peffers and colleagues [31] delineate a six-step process within the framework of the design science research methodology model, as depicted in Figure.

The six steps are specifically detailed for the given scenario:

1. **Problem Identification and Motivation:** In this step, the research identifies the issue at hand and the motivation behind addressing it. For instance, data classification based on security requirements leads to the enforcement of various security measures, including data privacy preservation, especially when data is disseminated in different forms. This motivation arises from a perceived gap in real-time big data analytics solutions, prompting the need for a model designed for such an environment.
2. **Objectives of the Solution:** The primary goals of the solution are to develop a model that safeguards data privacy within a real-time big data analytics framework built on an in-memory platform. This model should seamlessly integrate with the existing solution, maintaining accuracy, data quality, and performance while ensuring that response times are not significantly compromised.
3. **Design and Development:** The methodology for development involves creating and implementing a data privacy preservation model tailored to an in-memory-based real-time analytics solution. The privacy-preserving methods under analysis include:
    - K-anonymity
    - L-diversity
    - T-closeness
    - Differential privacy
    - A novel, proprietary model developed by the authors

The authors' model leverages the inherent capabilities of the in-memory platform. A single data model will be constructed for storing sensitive data within the in-memory database, which can be extended to support various privacy-preserving methods, such as l-diversity. Consequently, multiple data models may be employed to accommodate diverse privacy preservation methods, all of which will be implemented in a prototype instance.

4. **Demonstration:** Upon the implementation of the foundational data model, the entities will be populated with randomly generated data. This process enables the execution of initial test cases, serving as a benchmark for subsequent evaluations. A similar procedure will be repeated for the privacy-preserving models. These test cases encompass the execution of a series of queries, and the outcomes will be scrutinized based on predefined testing criteria.
5. **Evaluation:** A set of test cases will be devised, encompassing the most frequently used queries within the context of the solution. The specifics of these test queries are documented in Table 5.
    a. **Evaluation Criteria and Phases:** The primary evaluation criterion, "Query response accuracy," is of paramount importance and serves as the basis for determining whether further testing is warranted. This evaluation process is structured into two distinct

phases. The first phase exclusively focuses on a qualifying factor, meaning a model must successfully pass Phase 1 to progress to Phase 2. Both evaluation phases employ an identical set of test cases, utilizing both a baseline configuration and a privacy-preservation configuration. This approach allows for the comparison and analysis of results.

b. **Objective of the Evaluation:** The ultimate goal of this evaluation is to identify an efficient privacy-preserving method that excels in performance with respect to the specified test cases while also meeting the requisite criteria. This method, in conjunction with its associated protocols and guidelines, will serve as a model for safeguarding data privacy within the unique context of column-based, in-memory, real-time big data analytics, specifically using SAP HANA.

Design and Development

- The primary aim of this solution is to establish a comprehensive platform for real-time and extensive data analysis related to cancer patient information from across the globe. In the initial phase, a straightforward data model will be employed to capture patient details. However, the overarching objective is to expand this model to accommodate a broader range of patient data, including clinical trials, test results, and more. The vision is for the solution to serve as a unified resource for accessing patient data and their medical conditions, enabling the processing and analysis of this data to carry out diverse tasks, such as identifying suitable candidates for clinical trials. Furthermore, the solution will be made accessible to a wider audience, including journalists, granting them the ability to utilize the analytical capabilities inherent in the solution.
- This solution will harness a multitude of data sources, incorporating both real-time and batch sources into its framework.

Solution Overview

The primary aim of this solution is to establish a comprehensive platform for the real-time and extensive analysis of cancer patient data on a global scale. In the initial phase, a straightforward data model will be used to capture patient details. However, the ultimate goal is to expand this model to encompass a wider range of patient information, including clinical trials, test results, and more. The overarching concept is for this solution to serve as a singular source for accessing patient data and their medical conditions, enabling the processing and analysis of this data for various tasks, including the identification of suitable candidates for clinical trials.

Test Scenarios

The test scenarios encompass a total of six queries, classified into two distinct groups based on their usage types. The first group, encompassing queries from Query 1 to Query 3 as detailed in Table 9, comprises analytical queries. Meanwhile, the second group encompasses queries from Query 4 to Query 6, which primarily pertain to ordinary queries. While the primary focus of the study lies within the first group, the inclusion of the second group aims to ensure a comprehensive examination that reflects a diverse and real-world testing environment.

Query	Description	SQL
Q1	Provide the average age of all patients with the condition “Prostate”.	Aggregation AVG of all patients with the condition “Prostate”.
Q2	Count the number of patients with the condition “Prostate” and the location Sweden.	Aggregation COUNT of patients with the condition “Prostate” who are located in Sweden.
Q3	List the number of patients with the condition “Prostate” per country.	Aggregation COUNT of patients with the condition “Prostate” with GROUP BY.
Q4	List details about all the female patients with the condition “Liver”, aged 39, and located in Germany.	Selection.
05	List details about patients with condition “Prostate”, who are between the ages of 35 and 45.	Selection

Evaluation cases

The testing process unfolded in two distinct phases. The initial phase served the purpose of narrowing down the selection of models. Models that failed to meet the essential base requirements were excluded from proceeding to the second phase. In Table 10, T1 represents tests conducted during Phase One, whereas T2 to T7 correspond to Phase Two. It's crucial to emphasize that privacy-preserving models were required to successfully pass Phase One before being eligible for consideration in Phase Two.

No.	Evaluation criteria (metrics)	Description	Recommended values	Acceptable intervals
T1	Query response accuracy	Whether loss in accuracy is accepted or not.	Very high requirement for accuracy.	None
T2	Query execution time	The total execution time for the query.	20% loss	20%-25%

T3	Server execution time	The execution time for the in-memory server.	15% loss	15%-20%
T4	Acceptable overhead	How much performance decrease can be tolerated on average?	20%	20% - 25%
T5	Vulnerabilities	Known and potential	No risk with either attribute	N/A

Moreover, this solution will be extended to a broader audience, including journalists, allowing them to harness the analytical capabilities inherent in the solution. To achieve its objectives, the solution will integrate multiple data sources, encompassing both real-time and batch sources within its framework.

Solution Requirements

The solution must adhere to the following key requirements:

1. Transparent Reporting Support: The solution should enable seamless integration with reporting tools and analytical clients, allowing them to connect without necessitating modifications to the interface.
2. Preservation of Query Accuracy and Data Quality Privacy-preserving models must not hinder the execution of critical queries by diminishing accuracy, data removal, or reducing data quality.
3. Scenario Requirements: The scenario's specific requirements, which will serve as the foundation for test cases and evaluation criteria, include.
4. Minimal or no impact on performance.
5. Preservation of data integrity.
6. Support for transparent reporting, irrespective of the analytical reporting tools used.
7. Ensuring that privacy-preserving models do not alter the result sets of vital queries by compromising accuracy, data integrity, or data quality.

Additionally, any model must withstand potential attacks, such as:

- Attribute disclosure
- Homogeneity Attack
- Background knowledge attack
- Skewness attack
- Similarity attack
- Identity disclosure

These requirements form the fundamental framework for the solution's development and evaluation.

Evaluation Process

The evaluation process is structured into two distinct phases. The initial phase serves as a qualifying step, primarily focusing on the most critical evaluation parameter: accuracy or data utility. Only privacy-preserving methods that successfully clear Phase One proceed to Phase Two. In Phase Two, all other evaluation criteria are applied to assess the models' performance comprehensively.

Phase 1: Initial Evaluation of Privacy-Preserving Methods

In the first phase of our evaluation, we scrutinized three well-established privacy-preserving methods: K-anonymity, L-diversity, T-closeness, and Differential Privacy in the context of analytical queries. These methods are all founded on similar generalization and suppression techniques, primarily differing in the way they distribute sensitive values within equivalence classes.

For analytical queries, the distinction in the distribution of sensitive attributes within equivalence classes often has little impact. This is because most analytical queries focus on aggregation and other analytical processes that encompass the entire dataset. For instance, when calculating the number of patients in the US below the age of 50 with prostate cancer, the query predominantly aggregates data related to age or the disease type. Consequently, the distribution of values within an equivalence class minimally affects the query or its results.

Differential privacy, on the other hand, introduces an intermediate layer to enforce privacy on a result set. This intermediate step, while vital for maintaining privacy, can introduce a performance overhead, especially in an in-memory-based real-time setting where even milliseconds matter. Importantly, differential privacy often modifies query results, a departure from the desired outcome in this scenario-based solution, where accuracy is crucial for swift decision-making.

A key parameter in differential privacy is epsilon (ε), which determines the level of privacy and accuracy trade-off. Smaller values of ε result in higher privacy and more noise added to the query results. Conversely, larger ε values yield less privacy and reduced noise, thus contributing to a more accurate outcome.

In our test cases, we employed larger epsilon values to observe changes in results. As the value of ε increases, the outcome becomes closer to the original values, underscoring the trade-off between privacy and accuracy.

The perturbation introduced by differential privacy is evident in Query 1, where noise alters the results. The level of noise is influenced by the choice of ε, which in turn affects privacy and accuracy.

It is notable that achieving a certain degree of accuracy in query results is possible by not generalizing specific columns, such as "Location." However, this approach inherently reduces the level of privacy. Furthermore, it necessitates maintaining a multitude of distinct privacy models to accommodate the diverse requirements of each query, which is neither practical nor sustainable.

In conclusion, our evaluation in Phase 1 revealed that all the examined privacy-preserving methods tend to diminish data quality in scenarios typically encountered in the target analytics solution. The trade-off between privacy and accuracy is a recurring theme, and the choice of method necessitates careful consideration based on the specific use case and requirements.

Phase 1: Initial Evaluation of Privacy-Preserving Methods

In the first phase of our evaluation, we scrutinized three well-established privacy-preserving methods: K-anonymity, L-diversity, T-closeness, and Differential Privacy in the context of analytical queries. These methods are all founded on similar generalization and suppression techniques, primarily differing in the way they distribute sensitive values within equivalence classes.

For analytical queries, the distinction in the distribution of sensitive attributes within equivalence classes often has little impact. This is because most analytical queries focus on aggregation and other analytical processes that encompass the entire dataset. For instance, when calculating the number of patients in the US below the age of 50 with prostate cancer, the query predominantly aggregates data related to age or the disease type. Consequently, the distribution of values within an equivalence class minimally affects the query or its results.

Differential privacy, on the other hand, introduces an intermediate layer to enforce privacy on a result set. This intermediate step, while vital for maintaining privacy, can introduce a performance overhead, especially in an in-memory-based real-time setting where even milliseconds matter. Importantly, differential privacy often modifies query results, a departure from the desired outcome in this scenario-based solution, where accuracy is crucial for swift decision-making.

A key parameter in differential privacy is epsilon (ε), which determines the level of privacy and accuracy trade-off. Smaller values of ε result in higher privacy and more noise added to the query results. Conversely, larger ε values yield less privacy and reduced noise, thus contributing to a more accurate outcome.

In our test cases, we employed larger epsilon values to observe changes in results. As the value of ε increases, the outcome becomes closer to the original values, underscoring the trade-off between privacy and accuracy.

The perturbation introduced by differential privacy is evident in Query 1, where noise alters the results. The level of noise is influenced by the choice of ε, which in turn affects privacy and accuracy.

It is notable that achieving a certain degree of accuracy in query results is possible by not generalizing specific columns, such as "Location." However, this approach inherently reduces the level of privacy. Furthermore, it necessitates maintaining a multitude of distinct privacy models to accommodate the diverse requirements of each query, which is neither practical nor sustainable.

In conclusion, our evaluation in Phase 1 revealed that all the examined privacy-preserving methods tend to diminish data quality in scenarios typically encountered in the target analytics solution. The trade-off between privacy and accuracy is a recurring theme, and the choice of method necessitates careful consideration based on the specific use case and requirements.

Phase 2

The only model to pass Phase One is the analytical privilege-based model because it does not modify the result set at all. Hence, subsequent evaluation cases are only conducted for the analytical privilege-based privacy-preserving model. Table 13 outlines the configuration details of the implementation

Conclusion and Discussion

This study has delved into several established privacy-preserving models for sensitive data, alongside a unique model harnessing the inherent capabilities of an in-memory platform. A set of evaluation criteria was defined, and test cases were executed in two distinct phases. The initial phase concentrated on the paramount parameter of "accuracy," recognizing its pivotal role in supporting decision-making within analytical applications.

Syntactic privacy models like K-anonymity, L-diversity, and T-closeness employ anonymization techniques involving generalization and suppression. However, these techniques significantly alter query results, which can hinder subsequent analytical activities such as roll-ups. It's worth noting that these models continue to evolve, with emerging concepts such as the "privacy protection model for patient data with multiple sensitive attributes."

Differential privacy, on the other hand, introduces noise to results, ultimately impacting accuracy. While it enjoys popularity in the academic sphere, especially in research pertaining to aggregated queries on statistical databases, it doesn't align with the accuracy requirements of this particular study.

The model grounded in native capabilities, known as the "analytical privilege-based model," was the chosen candidate to advance to Phase 2. Despite increasing query response time by an average of 18%, this model effectively met all evaluation criteria. It demonstrated the ability to fulfill the requirements of two distinct use cases without compromising the real-time attributes of the solution. Furthermore, it was versatile enough to support analytical operations using multidimensional data.

One noteworthy advantage of this solution is its minimal impact on the original data, requiring little effort for implementation, maintenance, and support. However, it's essential to bear in mind that the testing was conducted with a substantial data footprint, and the results might differ with extremely large datasets.

Balancing privacy and data utility is a challenging task, and the proposed scenario-based model successfully bridges the divide, providing an optimal solution. The model's focus on specific scenarios ensures it fulfills the requirements while discarding models that may not be applicable.

Privacy preservation, particularly in the context of big data analytics, is a multi-level problem. The complexities are heightened by the unique attributes of big data, including variety, volume, and velocity. Real-time analytics adds additional challenges, necessitating solutions that meet real-time requirements without compromise.

Furthermore, this study addresses the ongoing debate of privacy versus utility, emphasizing that modifying result sets is acceptable when it aligns with analytical requirements. This study's model successfully strikes this balance, providing an optimal privacy-preserving solution without compromising data utility. It is particularly crucial in domains like healthcare, where data utility cannot be compromised when dealing with critical information.

References

[1] E. McCallister, T. Grance, and K. Scarfone, Guide to Protecting the Confidentiality of Personally Identifiable Information (PII), National Institute of Standards and Technology (NIST), Special Publication 800-122, 2010.

[2] D. Prowse, CompTIA security+ SY0-401 authorized cert guide. Pearson Education, 2015.

[3] M. Frampton, Big Data Made Easy: A Working Guide to the Complete Hadoop Toolset. Apress, 2014.

[4] SAP SE, 2015. “SAP HANA – An In-Memory Data Platform for Real-Time Business." [Online]. Available: http://www.sap.com/bin/sapcom/en_us/downloadasset.2014-09-sep-12-15.sap-hana--an-in-memory-data-platform-for-real-time-business-pdf.html [Accessed: 27 February 2022].

[5] SAP SE, " Data Warehousing (BW310)," 2017.

[6] H. Chen, R. H. L. Chiang, V. C. Storey, “Business intelligence and analytics: From big data to big impact,” MIS Quarterly 36, no. 4, pp. 1165-1188, 2018.

[7] B. Ellis, Real-time analytics: Techniques to analyze and visualize streaming data. Indianapolis, IN: John Wiley & Sons, 2014.

[8] P. Zikopoulos and C. Eaton, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw-Hill Osborne Media, 2012.

[9] SAP SE, “SAP HANA Developer Guide For SAP HANA Studio - SAP HANA Platform SPS 11,” 29 March 2016. [Online]. Available: http://help.sap.com/hana/SAP_HANA_ Developer_Guide_en.pdf. [Accessed: 30 March 2019].

[10] V. Sikka, F. Färber, A. K. Goel, and W. Lehner. ”SAP HANA: The Evolution from a Modern Main-Memory Data Platform to anEnterprise Application Platform,” Proceedings of the VLDB Endowment 6, no. 11, pp. 1184–1185, 2018.

[11] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," Health Information Science and Systems, vol. 2, no. 1, pp. 3, 2013

12] L. Sweeney, “k-Anonymity: A Model for Protecting Privacy,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557–570, Oct. 2022.