Optimized Privacy Preserving Big Data Publishing Approach

Salheddine Kabou, Imad eddine Kimi, Arbia Boudaouad

Abstract

In the context of big data, balancing individual privacy with the need for data utility remains a critical challenge. This research presents the Bottom-Up k-Concealment (BU-KC) framework, a pioneering solution for Privacy-Preserving Big Data Publishing (PPBDP). At its core, the k-concealment model addresses the limitations of traditional k-anonymity by minimizing excessive generalizations, ensuring stronger privacy protection while preserving data utility. Combined with the Bottom-Up Generalization (BUG) strategy, BU-KC demonstrates superior performance compared to Top-Down Specialization (TDS) in terms of computational efficiency, scalability, and privacy preservation, while concurrently sustaining data utility. Furthermore, the integration of Apache Spark's distributed computing paradigm enables us to effectively mitigate scalability constraints and processing bottlenecks commonly observed in the anonymization of large-scale datasets.

Keywords: Data anonymization, Big data, Apache Spark, k-concealment model, Bottom-Up.

MSC 2020: 68P27, 68M14, 62R07.

1 Introduction

In recent years, organizations such as hospitals and universities require the sharing of vast amounts of data on the web with the research institutes to undergo analysis and have their intricate patterns identified. A major challenge in data publishing is ensuring individual privacy by anonymizing data that could reveal personal identities [24]. The key

doi:10.56415/csjm.v34.07

lies in developing systems and methods that strike the perfect balance between preserving privacy and maintaining data utility. The area of study that incorporates all these techniques is referred to as Privacy Preserving Data Publishing (PPDP).

As data volumes have increased and the big data paradigm, defined by the 42 Vs, has emerged, many privacy models have become highly influential in this field. Among these approaches, k-anonymity and l-diversity are extensively used. K-anonymity, proposed by [2], guarantees that each record cannot be distinguished from at least $k-1$ other records based on specific identifying attributes, thus safeguarding individual privacy. Although this model produces an acceptable level of data privacy, it generally fails to retain data utility, particularly in high-dimensional datasets, by applying excessive generalization, resulting in significant information loss for anonymized data. On the other hand, the k-concealment model provides a more sophisticated approach. It provides greater data utility by reducing less generalization while assuring a similar degree of privacy [5].

Generalization operations are widely applied in data anonymization, where original attribute values are replaced with generalized ones using a taxonomy tree to ensure semantic consistency between the old and new values. There are typically two main approaches to navigating the taxonomy tree: Top-Down Specialization (TDS) begins at the root of the taxonomy tree and progresses downward, while Bottom-Up Generalization (BUG) starts at the leaf nodes and works its way upward. Unlike the top-down approach, we utilize Bottom-Up Generalization (BUG), a more optimized method that maintains the quality of data [6].

As data quantities reach zettabyte levels, optimizing processing time becomes critical for effective data exchange and publication. Traditional centralized techniques fail to handle massive volumes of data, resulting in inefficiencies in processing speed, scalability, and overall performance [1]. To overcome these challenges, distributed computing frameworks like MapReduce and Apache Spark have been created. Among these, Apache Spark stands out as an open-source, scalable platform capable of handling large datasets efficiently. Its capacity to handle vast amounts of data fast and efficiently makes it an ideal tool

for modern data processing tasks. Programs running on the Spark platform can achieve significantly faster processing speeds, up to 100 times faster in memory and 10 times faster on disk compared to MapReduce. Consequently, Spark has emerged as a highly efficient ecosystem for performing the anonymization process on large-scale data. By leveraging Spark's capabilities, the anonymization process can be executed more efficiently, enabling faster processing and improved scalability.

Most of the previous approaches have focused on the application of k-anonymity [2] and l-diversity [3] models using MapReduce and Spark. To the best of our knowledge, the k-concealment model has not yet been integrated into distributed big data publishing frameworks based on Apache Spark. In addition, although Bottom-Up Generalization (BUG) [21] has been explored in centralized and MapReduce environments, it has not been combined with the k-concealment model within a Spark-based framework.

In this work, we explicitly address this gap by proposing a unified Bottom-Up k-Concealment (BU-KC) framework on Apache Spark. While the underlying components have been individually studied, our contribution lies in their systematic integration and optimization for scalable privacy-preserving data publishing, with a particular emphasis on minimizing unnecessary generalization to preserve data utility. Experimental results demonstrate that the proposed approach significantly reduces information loss compared to traditional top-down methods, while maintaining strong privacy guarantees and scalability.

The remainder of this paper is organized as follows: Section 2 reviews relevant literature and prior work in the field. Section 3 delves into the key concepts and foundational components of the proposed approach. Section 4 presents and analyzes the experimental results, evaluating the performance of the method. Finally, Section 5 concludes the paper and outlines potential directions for future research.

2 Related works

Privacy-Preserving Big Data Publishing (PPBDP) is a crucial area of research that focuses on balancing privacy and utility when releasing large-scale datasets. In the context of big data, various privacy models,

including k-anonymity, are widely applied. Section 2.1 addresses different centralized approaches using the k-anonymity model, while Section 2.2 presents some of distributed big data methods to accomplish the anonymization using whether top-down or Bottom-up generalization operators centralized approaches.

2.1 Centralized approaches

LeFevre et al. [9] presented the Mondrian algorithm, which uses multidimensional generalization in a top-down fashion for partitioning the data. Initially, all data instances are grouped into a single partition. The algorithm then recursively splits each partition into pairs of subpartitions until the k-anonymity conditions are satisfied. At each step, the algorithm selects the widest dimension as the cut dimension and uses the median value of that dimension as the split point.

Yaseen et al. [10] introduced three centralized approaches for generalization hierarchies: conventional generalization hierarchies, divisors generalization hierarchies, and cardinality-based generalization hierarchies. Each method prioritizes data utility and partitions the attribute domain space in a distinct way. Their performance is evaluated based on the discernibility penalty, distortion ratio, and the number of nodes at each level.

In the paper [11], a new data anonymization algorithm that enhances users' community privacy is presented. The algorithm assesses the vulnerability of each attribute in a user's dataset to effectively protect community privacy. It conducts adaptive data generalization while simultaneously assessing attribute susceptibility and entropy.

Torra and Navarro [13] proposed a new data publishing method for minimizing the attribute disclosure risk in the k-anonymity model. To avoid internal and external attacks, the solution is to apply well-known codes to identify sensitive cells in a tabular dataset by using p-sensitivity and p-diversity models.

In order to avoid the danger of skewness and similarity attacks, Su et al. [12] presented a k-anonymity mechanism based on clustering for multi-dimensional data, combined with the t-closeness privacy model. Based on an enhanced African vultures optimization, the sug-

gested approach adheres to extremely accurate clustering of the multidimensional dataset and can offer the ideal solution with multiple dimensions.

A new anonymization model known as z-anonymity is proposed in [16] and deals with publishing data streams. The key idea behind this mechanism is the capability of realising such data if at least $(z-1)$ other users have shared it within a certain time period. The suggested approach creates a flexible framework for evaluating privacy in data streams by combining the z-anonymity and k-anonymity properties using a probabilistic model.

Even though these methods significantly enhance privacy, they are still reliant on the excessive generalization caused by the traditional k-anonymity paradigm. Our method, on the other hand, is based on the k-concealment model, which, by minimizing generalization, greatly reduces information loss during anonymization.

2.2 Decentralized approaches

Sopaoglu and Abul [14] developed a Top-Down Specialization (TDS) anonymization solution designed specifically for the Apache Spark platform. This method assesses both numerical and category attributes by iteratively generalizing the original data, beginning with the most general node and using the k-anonymity model.

Zakerzadeh et al. [15] enhanced the centralized Top-Down Mondrian technique for massive data applications by using the MapReduce framework. Reducing the algorithm's processing time across the Mapper, Combiner, and Reducer computational nodes is the aim of this work.

In a related study, Ashkouti et al. [8] enhanced earlier work on the Apache Spark Framework by applying a multidimensional Top-Down Mondrian approach. The primary difference between the two approaches resides in selecting the axis dimension criterion to which the partitioning is carried out. While [15] utilized the domain width as the primary criterion for axis dimension selection, the improved study diverges by prioritizing the dispersion of attribute values over the domain. This research aims to enhance time efficiency through the

implementation of RDD programming, while simultaneously improving data privacy by adopting the l-diversity model in lieu of k-anonymity.

More recently, the same authors [17] developed a partition clustering model using the Apache Spark ecosystem that improves data privacy by using l-diversity and t-closeness. The model more effectively partitions data using city block and Pearson distance functions to improve scalability, minimize runtime, and reduce information loss.

In the presence of big data, the centralized traditional BUG techniques are not suitable for adhering to privacy needs. This motivated researchers to propose parallel Bottom-Up approaches to address limitations brought about by centralized BUG. In [18] [19], a scalable, advanced Bottom-up generalization method is proposed to meet the k-anonymity model using the MapReduce platform on Cloud. The process involves two main steps: first, the original dataset is divided into multiple tables, and each partition is anonymized in parallel to generate intermediate results. In the second step, these intermediate results are grouped, and an additional anonymization operation is performed to ensure compliance with the privacy model. In terms of scalability, the authors demonstrate that the parallel BUG approach is more efficient than the top-down specialization method.

The key distinction between our research and the studies mentioned above lies in our focus on data utility as the primary consideration in the big data paradigm. While previous approaches have primarily concentrated on developing top-down k-anonymization techniques within MapReduce or Apache Spark environments, Bottom-Up methods have largely been confined to MapReduce frameworks. In this paper, we introduce a Bottom-Up k-concealment approach implemented within Apache Spark, designed to enhance data utility in a big data context. This contribution is particularly noteworthy, as studies [7] and [24] have shown that Bottom-up approaches tend to outperform top-down methods in terms of both efficiency and privacy preservation.

3 Methodology

3.1 Background and key concepts

In the context of privacy-preserving big data publishing, data is typically structured in a table with different types of attributes, classified into three main categories: (i) Explicit Identifiers (EI): These are attributes that immediately reveal an individual's identifier, such as a name. (ii) Quasi-Identifiers (QI): While these attributes are not individually revealing, when paired with other data, they can be utilized to identify individuals, such as zip code. (iii) Sensitive Identifiers (SI): These attributes include confidential information, such as medical conditions [20]. It's important to note that EI and QI traits might be numerical or categorical, whereas SI attributes are always categorical.

Adhering to anonymity requires converting the data into generalized groups, known as Equivalence Classes (EC) or QI-groups. These groups serve as the basis for numerous privacy-preserving models. The k-anonymity model ensures that each group in the dataset has at least k entries with identical QI values, with a maximum group size of $2k-1$. Adding to this, the l-diversity model improves k-anonymity by ensuring that each group also has at least l distinct SI values [3]. In contrast, the k-concealment model guarantees that each record is computationally indistinguishable from at least $k-1$ other records [4]. This model enhances privacy and data utility by minimizing unnecessary generalization, surpassing the effectiveness of k-anonymity.

Definition 1. *Generalized Anonymization Levels: Let* $T = \{T_1, \ldots, T_n\}$ *represent a dataset (or table), and let* $g(T) = \{T'_1, \ldots, T'_n\}$ *denote the corresponding generalized version of* T*. The relationships between* T *and* $g(T)$ *define different levels of anonymization as follows:*

- *The generalised table g(T) is a (1, k)-anonymization of T if each original record in T corresponds to at least k records in the generalised dataset g(T). This ensures that no record in T can be distinguished from the other k records in g(T).*
- *(k, 1)-Anonymization: A generalised table (T) is (k, 1)-anonymized if each generalised record in g(T) corresponds to at least k*

original entries in T. This ensures that each anonymised group in g(T) represents at least k individuals in T.

- *(k, k)-Anonymization: A table g(T) is a (k, k)-anonymization of T if it satisfies both (1, k)- and (k, 1)-anonymization conditions. This means that every record in T is interchangeable with k records in g(T), and every group in g(T) corresponds to at least k records in T.*

Definition 2. *k-Concealment Based on Bipartite Graph Matching: Let T represent an original table, g(T) its generalized version, and* $B_{T,g(T)}$ *the bipartite graph formed between T and g(T). The concept of k-concealment can be formalized as follows:*

- *Matching Criterion: A record* $R \in g(T)$ *is considered a match for a record* $R' \in T$ *if an edge* (R', R) *exists in the bipartite graph* $B_{T,g(T)}$ *and this edge is part of at least one perfect matching in* $B_{T,g(T)}$.

k-Concealment: The generalized table $g(T)$ is said to achieve *k-concealment* of T if each record $R' \in T$ has at least k distinct matches in $g(T)$. This ensures that R' becomes computationally indistinguishable among k generalized records, effectively enhancing privacy.

3.2 Bottom-up generalization model

Figure 1 provides a visual representation of the sequential phases involved in the implementation of our distributed k-concealment model. The process is executed using a series of RDDs, which are manipulated through transformations and actions in the Apache Spark library.

3.2.1 Data initiation

Data preparation: The dataset obtained from the data holder often includes missing or duplicate items. This stage attempts to enhance the original data, making it suitable for the anonymisation operation. By removing incorrect items, the cleaned data is efficiently distributed over several worker nodes for further processing.

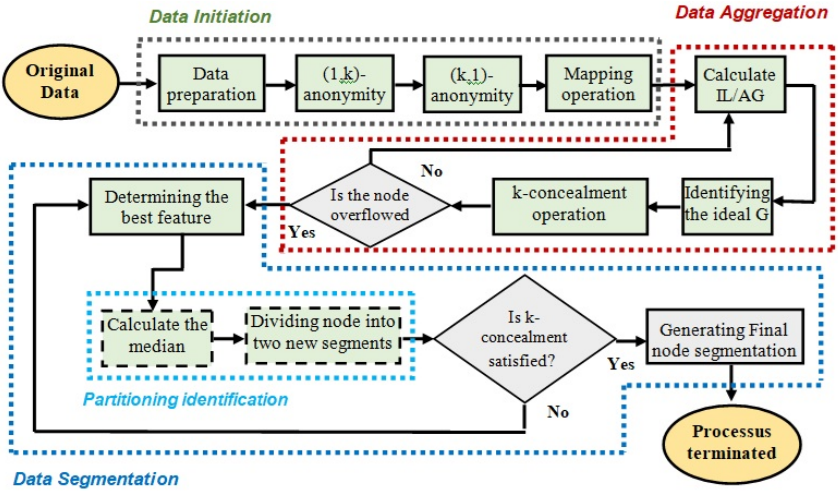

Figure 1. Key phases of the proposed method

(1,k)-anonymity operation: After removing erroneous and unnecessary items, the dataset is partitioned across worker nodes, establishing a foundation for further processing. The emphasis is therefore on record aggregation to ensure that each item has a connection with at least k counterparts inside its generalised representation. This procedural step is critical for the construction of a thorough anonymisation strategy that uses the cleaned and structured dataset from the previous phase.

(k,1)-anonymity operation: This phase converts the anonymisation process from a (1,k)-anonymity configuration to a (k,1)-anonymity configuration by building on the structural foundation established in the previous step. After determining that each record in the original dataset is indistinguishable from at least k other records in the generalised table g(T), the (k,1)-anonymity operation requires that each generalised record g(T) correspond to at least k original records in the dataset. This interaction between the two processes ensures consistency, improves privacy protections, and preserves the accuracy and integrity of the anonymised data.

Mapping operation: In this step, we perform a mapping operation

to link each data record to a $< key, value >$ pair, as described in Algorithm 1. Since the previous step produces a set of equivalence classes following the (k,k)-anonymity model, the key represents the label of each data record, which corresponds to the distinct sensitive values within its equivalence class. The value corresponds to the actual data in each attribute of the record.

Algorithm 1 Mapping operation

Input: RDD3
Output: (key, value) pairs as RDD4

```
Key = [ ] // list of SI of each EC
Values = [ ]
for each data records in EC do
    Value = Value.append (QI)
    Key = All_SI (value)
end for
```

3.2.2 Data aggregation

Calculate Score (G): The Bottom-Up generalization method entails a progressive assessment of possible generalizations, starting with the most exact data and progressing to larger categories. At each iteration, the system determines the optimal generalization level, effectively balancing privacy and data usefulness. Every alternative generalization is assessed to see how well it fits the k-concealment model while causing the least amount of important data loss. The ideal choice, as determined by the highest score provided by the balance of privacy and data usefulness, is then chosen and used. The approach includes iteratively developing candidate generalizations, assessing their different scores, selecting the best candidate, and repeating these stages until the predetermined anonymity conditions are met. A generalization is considered optimal when it has the highest score as determined by the Information Loss/Anonymity Gain (IL/AG) trade-off criterion. This metric tries to assess how well a generalization preserves data value for analytical tasks such as classification while also safeguarding privacy via robust anonymization techniques. The score assigned to a generalization G, marked Score(G), is obtained by using Equation 1, as shown

below:

$$\text{Score}(G) = 1 + \frac{IL(G)}{AG(G) \cdot IL(G)}, \quad AG \neq 0. \tag{1}$$

The Information Loss, denoted as $IL(G)$, is a metric that quantifies the loss of data granularity caused by the application of generalization G. Specifically, it articulates the degree to which information detail is attenuated during the anonymization procedure, as formally defined in Equation (2):

$$IL(G) = \text{Entropy}(R_g) - \sum_{d_i} \frac{|R_{d_i}|}{|R_g|} \cdot \text{Entropy}(R_{d_i}). \tag{2}$$

Here, R_g represents the set of records that contain the generalized value g, while R_{d_i} refers to the set of records that contain the original, more specific value d_i. The entropy $\text{Entropy}(R_x)$, where $x \in \{g, d_i\}$, represents the entropy of the set R_x. It is computed as follows:

$$\text{Entropy}(R_x) = -\sum_{\text{cls}} \frac{\text{freq}(R_x, \text{cls})}{|R_x|} \cdot \log_2 \left(\frac{\text{freq}(R_x, \text{cls})}{|R_x|} \right), \tag{3}$$

where $\text{freq}(R_x, \text{cls})$ represents the percentage of individuals in R_x belonging to the class labeled *cls*.

The *Anonymity Gain*, denoted $AG(G)$, corresponds to the increase in anonymity resulting from the application of generalization G. Intuitively, this measure is computed by comparing the level of anonymity in the table before and after applying generalization G. Formally, it is defined as:

$$AG(G) = \text{Anonymity}(T, \text{after } G) - \text{Anonymity}(T, \text{before } G). \tag{4}$$

Anonymity (T, after G) refers to the size of the smallest equivalence class in T after applying generalization G, where the equivalence class contains the fewest individuals sharing the same QI. On the other hand, *Anonymity (T, before G)* represents the size of the smallest equivalence class in T before applying G.

Identifying the Ideal Generalization (G_{id}): In this phase, the algorithm evaluates candidate generalizations to identify the one that best balances data utility and privacy enhancement. Each generalization is assessed by computing its score, which reflects this trade-off. The algorithm selects the generalization with the highest score, referred to as G_{id}, and applies it to the dataset. This process ensures that the applied generalization achieves the desired level of anonymization while preserving maximum data utility.

k-Concealment Operation: After applying the ideal generalization G_{id}, each equivalence class (EC) within the generalized dataset undergoes the k-concealment operation. This operation ensures that each EC satisfies the k-concealment model, thereby enhancing privacy protection. The k-concealment operation modifies the ECs to meet the privacy requirements, ensuring that each EC contains at least k original records, thus preventing re-identification. The outcome is stored in `RDD5` as $< Key, ECc >$ pairs, where the `Key` represents the set of distinct sensitive identifiers within the equivalence class, and the `ECc` corresponds to the EC that satisfies the k-concealment model based on the ideal generalization G_{id}.

Algorithm 2 Data Aggregation

Input: RDD4, ECs
Output: (key, value) pairs as RDD5

while Node is not overloaded **do**
  for all generalizations G **do**
    compute Score$_G$
  end for
  determine the ideal generalization G_{id}
  generalize ECs using G_{id}
  for each EC$_{G_{\text{id}}} \in g(T)$ satisfying (k, k)-anonymity **do**
    EC$_c \leftarrow$ `k_concealed`(EC$_{G_{\text{id}}}$)
  end for
end while

3.2.3 Data segmentation

Determining the Best Feature: In this phase, the focus is on identifying the most appropriate dimension or feature for splitting the nodes. Since data records are distributed across multiple worker nodes, this selection process must be carried out in parallel, with each node independently determining its segmentation criterion. The selection is guided by criteria from the Mondrian algorithm [9], which aims to optimize data segmentation while respecting privacy constraints.

The feature chosen for segmentation is typically the one associated with the attribute that has the widest range, as this broader scope increases the likelihood of identifying an optimal segmentation point, ultimately improving the efficiency of the data segmentation process. As a result, this stage culminates in the persistence of the resultant data within `RDD6`, structured as $< Key, Feat >$, where `Key` denotes the set of unique Sensitive Identifiers (SIs) within the equivalence class, and `Feat` signifies the quasi-identifier (QI) used for segmentation.

Partitioning Identification: Once the best feature has been selected, the algorithm establishes two isolated segments within that dimension. It separates overloaded nodes that contain equivalence classes previously adhering to the k-concealment requirements. The main objective is to eliminate segment overlap in order to improve data utility.

To accomplish this, the algorithm employs a median-based segmentation strategy, partitioning segments around the median value of the selected feature. For example, when using the attribute "Age," the median value serves as the segmentation point, resulting in two distinct segments: one containing data records (DRs) below the median, and the other containing records above it, as illustrated in Figure 2.

Verifying Privacy Compliance: After the segment identification process, it is crucial to verify that data handling complies with privacy regulations and protection requirements. In this step, the data is partitioned as extensively as possible while ensuring adherence to the k-concealment model, with the aim of minimizing information loss.

The primary criterion is to ensure that each segment contains at least k records, and that each equivalence class (EC) includes at least l distinct sensitive values. This segmentation process is repeated it-

eratively until both segments satisfy the k-concealment requirements, thereby ensuring that sensitive information is properly protected and privacy standards are upheld.

Algorithm 3 Data Segmentation

Input: RDD6
Output: Segm1 and Segm2
Step 1: Partitioning Identification
for each best attribute Att **do**
    compute median m of Att
    select m as splitting point
    divide the overloaded node at the segmentation point into two segments
    Segm1 ← records where DR ¡ m
    Segm2 ← records where DR ¿ m
end for
Step 2: Verifying Privacy Compliance
if $k < \text{count}_{\text{Segm-i}}$ **and** $l \leq$ labels **then**
    generate final node segmentation
else
    repeat Step 2
end if

4 Results and discussion

For the experimental evaluation, we used the widely studied Adult dataset from the UCI Machine Learning Repository, which has been extensively adopted in the literature on privacy-preserving data publishing and in several related works used for comparison in this study. This choice ensures a fair and consistent evaluation within a common experimental context. The dataset comprises 48,842 records with 14 attributes. To identify quasi-identifiers (QIs), we selected the attributes age, occupation, relationship, education, and workclass, while disease was designated as the sensitive attribute (SI). The selected QIs and the sensitive attribute were organized using taxonomy structures consistent with those adopted in previous studies [24].

Since our approach is applied in a big data context, expanding the

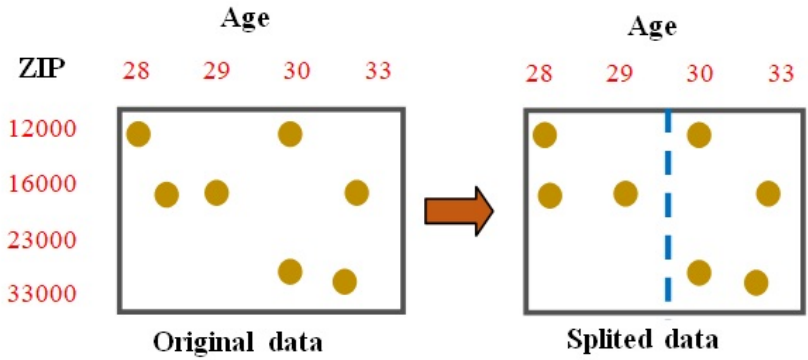

Figure 2. Partitions identification

dataset was necessary to ensure its suitability for large-scale analysis. To accomplish this, we increased the size of the Adult dataset by randomly generating additional records based on the selected quasi-identifiers. The details of the final dataset are presented in Table 1.

Table 1. Recommended datasets and their sizes

Dataset	**Number of data records**	**Size**
Adult 48k	48,842	1.5 MB
Adult 500K	500 Thousand	15 MB
Adult 1M	1 Million	31 MB
Adult 10M	10 Million	295 MB

4.1 Data utility evaluation

A data publisher's main goal is not only to ensure data privacy but also to maintain its utility. One crucial metric for assessing the impact of generalization on data quality is information loss, which helps quantify the extent of data degradation. While numerous methodologies can be employed to assess this parameter, the Normalized Cardinality Penalty (NCP) has emerged, particularly within the realm of big data, as the most suitable metric for quantifying information loss [24] [22].

Equation 5 defines the NCP calculation, where the numerator captures the difference between the upper and lower bounds of attribute m within data record n following generalization. The denominator represents the difference between the maximum and minimum values of attribute m across the entire dataset. The variables x and y denote the number of records and attributes, respectively. To evaluate the performance of different anonymization methods concerning information loss, we utilized the synthetic Adult dataset. The methods compared include Top-Down K-Anonymization (TD-KA) [14], Bottom-Up K-Anonymization (BU-KA), and our proposed Bottom-Up K-Concealment (BU-KC). Figure 3 illustrates the results of these evaluations. The assessments spanned a range of parameter values for k and l, with k varying between 30 and 160 and l ranging from 3 to 6. These evaluations demonstrate the relative effectiveness of the methods in preserving data utility while adhering to privacy requirements.

$$\mathrm{NCP} = \sum_{n=1}^{x} \sum_{m=1}^{y} \frac{|\mathrm{upper}_{nm} - \mathrm{lower}_{nm}|}{x \cdot y \cdot |\mathrm{Max}_m - \mathrm{Min}_m|}. \tag{5}$$

As shown in Figure 3, the NCP metric increases with higher values of k and l for all methods. This increase reflects the growing number of data records grouped within equivalence classes due to higher generalization requirements. However, BU-KC consistently outperforms TD-KA in terms of information loss. The key advantage of BU-KC lies in its Bottom-Up approach, which starts with the most granular data and only generalizes when necessary, minimizing unnecessary information loss. In contrast, TD-KA uses more generalizations right from the start, which leads to more data loss and a discernible drop in usefulness. This targeted approach makes BU-KC more effective by enabling it to better retain data utility. The BU-KC approach exhibits enhanced performance relative to BU-KA, a result attributable to its utilization of the k-concealment model. This model, while achieving k-anonymity, distinguishes itself by its capacity to minimize information loss through refined generalization strategies. In contrast to k-anonymity, which frequently resorts to extensive and imprecise generalization, k-concealment facilitates a more discerning and efficacious data transformation. By strategically limiting superfluous generaliza-

tion, our approach effectively retains a greater proportion of salient information, thereby yielding improved data utility.

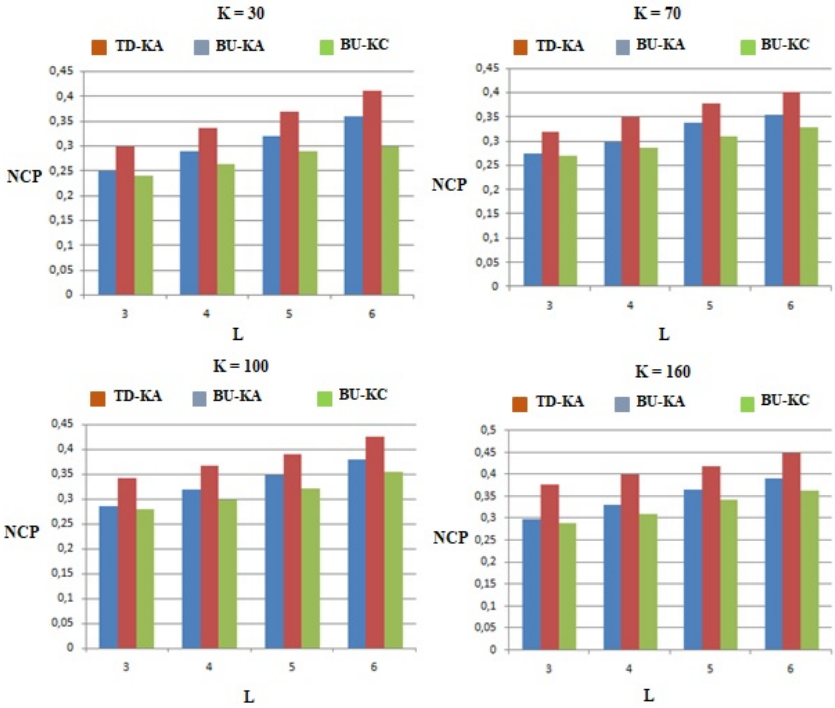

Figure 3. Data utility evaluation

4.2 Impact on classification models

To assess the effectiveness of the proposed BU-KC method, both original and anonymized datasets were evaluated using classification models. The classification outputs were then compared to determine the extent of performance degradation caused by the anonymization process (Figure 4). Accuracy served as the primary metric for evaluating the performance of the classifiers. The supervised learning models employed in this analysis included Naïve Bayes, Decision Tree, and Random Forest algorithms [23].

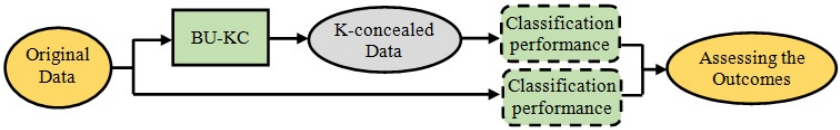

Figure 4. Evaluating classification model results

Figure 5 illustrates the evaluation metrics obtained for the synthetic Adult dataset. Both the BU-KA and BU-KC models were applied across varying k values (ranging from 20 to 160) to evaluate the effectiveness of the anonymization techniques. Balancing privacy protection and data utility remains a central objective of this study. As illustrated in Figure 5, the classification accuracy is highest when the k value is set to 20. For the Naïve Bayes classifier, accuracy levels reached 70.83% for BU-KC and 68.01% for BU-KA, reflecting reductions of only 0.21% and 3.03%, respectively, compared to pre-anonymization accuracy (B.Anony). The Random Forest classifier demonstrated accuracy rates of 71.86% for BU-KC and 69.77% for BU-KA, corresponding to decreases of 0.39% and 2.48%. Similarly, the Decision Tree classifier yielded accuracy levels of 71.47% for BU-KC and 68.77% for BU-KA, showing reductions of 0.41% and 3.11%, respectively. These results underscore the superior performance of the BU-KC method in preserving classification accuracy. The minimal reduction in accuracy observed with BU-KC highlights its ability to better retain essential data attributes during the anonymization process compared to the traditional BU-KA model. This highlights the effectiveness of the k concealment model in striking a compromise between privacy requirements and data value.

4.3 Performance metrics: Time efficiency

Apache Spark is a fast environment for big data operating. One of the key strengths of this framework is its ability to perform in-memory computation and store data as Resilient Distributed Datasets (RDDs). This design allows for extremely fast distributed data processing. When compared to MapReduce, it can run applications up to 100 times faster

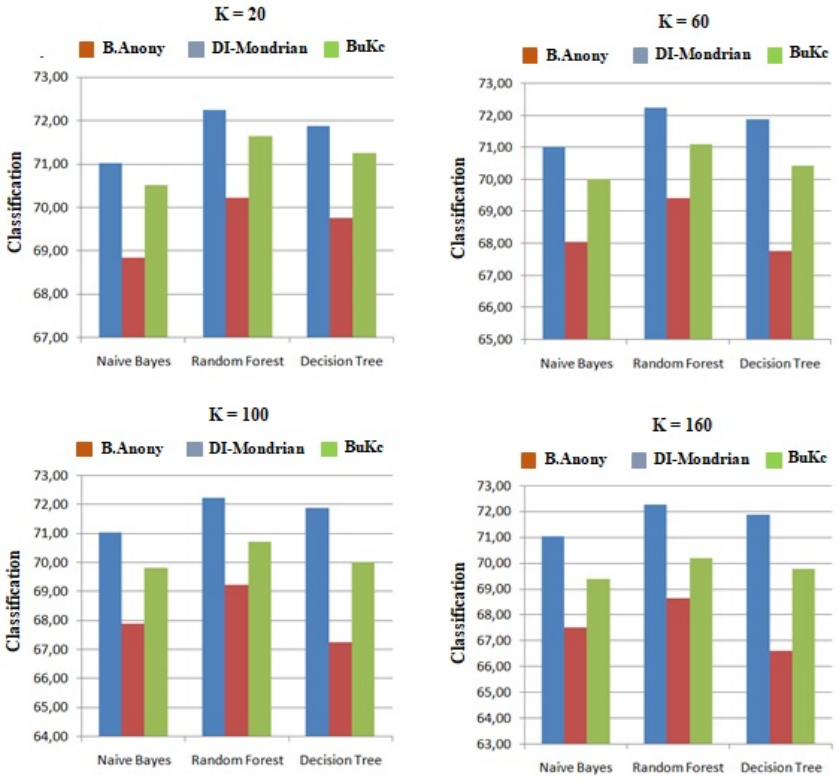

Figure 5. Impact of classification models

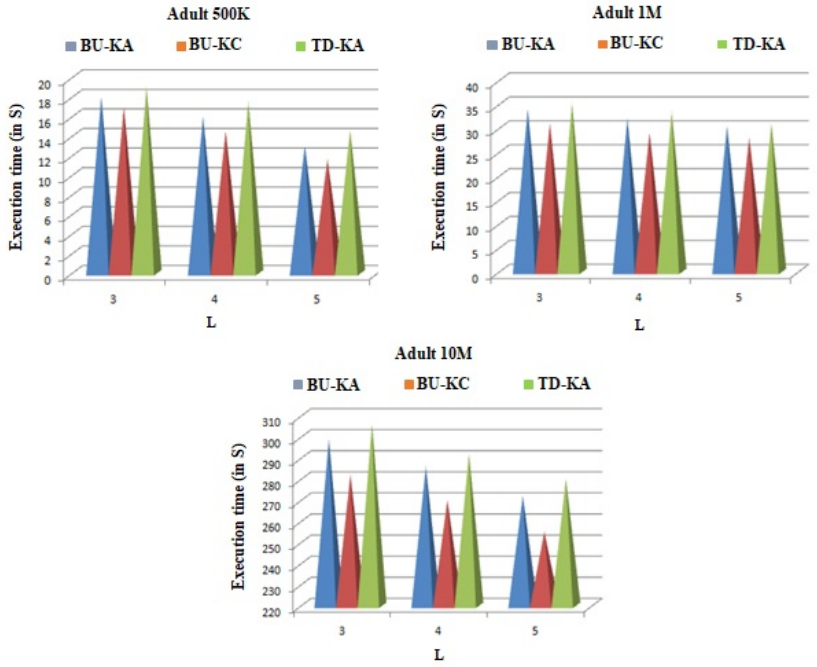

Figure 6. Time efficiency

in memory and 10 times faster when using disk storage. In this section, we evaluate the time efficiency of our proposed BU-KC, Top-Down K-Anonymization (TD-KA), and Bottom-Up K-Anonymization (BU-KA) over a synthetic Adult dataset. The running time was measured in seconds, and the anonymization setups have been carried out with l values ranging from 3 to 5. To further enhance the scalability, all the methods were tested over Adult of varied sizes: Adult 500K, Adult 1M, and Adult 10M. Figure 6 demonstrates that the Bottom-Up K-Concealment (BU-KC) approach outperforms both Top-Down K-Anonymization (TD-KA) and Bottom-Up K-Anonymisation (BU-KA) approaches regarding the time efficiency. The BU-KC methodology is consistently faster than other methods since it generalizes data step by step, only when required. Focusing on data records that require anonymization avoids redundant processing, making it significantly more efficient. This contrasts significantly with the TD-KA method, which starts with very broad generalizations, forcing the framework to handle a larger amount of data and consequently leading to longer execution times. Similarly, despite outperforming TD-KA, the Bottom-Up K-Anonymization (BU-KA) approach is less computationally efficient than BU-KC. The increased processing time observed in BU-KA is directly attributable to the necessity of excessive generalization to achieve k-anonymity.

5 Conclusion

In this study, we presented a novel approach known as Bottom-Up k-Concealment (BU-KC), meant to increase data privacy while keeping its utility in the Big Data publication paradigm. This technique is constructed using Apache Spark, a platform that facilitates the quick processing of large-scale data. Our results reveal that BU-KC outperforms classic k-anonymity approaches by decreasing information loss and greatly boosting processing performance. These gains are driven by two major elements: (i) the k-concealment model, which eliminates superfluous generalization while guaranteeing robust privacy protection, and (ii) the Bottom-Up Generalization (BUG) technique, which simplifies the anonymization process for increased efficiency. Future

work will focus on extending the proposed framework by integrating stronger privacy models, such as t-closeness or differential privacy, in order to enhance privacy guarantees from a security perspective. In addition, further optimization of the BU-KC framework to reduce computational resource consumption and its adaptation to other distributed computing platforms will be investigated.

References

[1] S. Kabou, L. Gasmi, and A. Kabou, "Privacy Preserving Continuous Big Data Publishing," in *Proc. 4th Int. Conf. Embedded & Distributed Systems (EDiS)*, 2024, pp. 109–114. DOI: 10.1109/EDiS63605.2024.10783344.

[2] L. Sweeney, "k-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.

[3] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, Article ID: Pages 3–es, 2007. DOI: https://doi.org/10.1145/1217299.1217302.

[4] T. Tassa, A. Mazza, and A. Gionis, "k-concealment: An alternative model of k-type anonymity," *Trans. Data Privacy*, vol. 5, pp. 189–222, 2012.

[5] S. Kabou and S. M. Benslimane, "A new distributed anonymization protocol with minimal loss of information," *Int. J. Organizational and Collective Intelligence*, vol. 7, no. 1, pp. 1–19, 2017. DOI: https://doi.org/10.4018/IJOCI.2017010101.

[6] S. Kabou, S. M. Benslimane, and A. Kabou, "Toward a new way of minimizing the loss of information quality in the dynamic anonymization," in *Proc. 2nd Int. Conf. Mathematics and Information Technology (ICMIT)*, IEEE, 2020, pp. 186–189. DOI: 10.1109/ICMIT47780.2020.9046981.

[7] B. C. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.*, vol. 42, no. 4, Article No.: 14, pp. 1–53, 2010. DOI: https://doi.org/10.1145/1749603.1749605.

[8] F. Ashkouti, K. Khamforoosh, and A. Sheikhahmadi, "DI-Mondrian—Distributed improved Mondrian for satisfaction of the L-diversity privacy model using Apache Spark," *Inf. Sci.*, vol. 546, pp. 1–24, 2021. DOI: https://doi.org/10.1016/j.ins.2020.07.066.

[9] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *Proc. 22nd Int. Conf. Data Eng. (ICDE)*, (Atlanta, GA, USA, 2006), Apr. 2006, pp. 25–25. DOI: 10.1109/ICDE.2006.101.

[10] S. Yaseen, S. Abbas, A. Anjum, T. Saba, A. Khan, and S. Malik, "Improved generalization for secure data publishing," *IEEE Access*, vol. 6, pp. 27156–27165, 2018. DOI: 10.1109/ACCESS.2018.2828398.

[11] A. Majeed and S. Lee, "Attribute susceptibility and entropy based data anonymization to improve users community privacy and utility in publishing data," *Appl. Intell.*, vol. 50, pp. 2555–2574, 2020. DOI: https://doi.org/10.1007/s10489-020-01656-w.

[12] B. Su, J. Huang, K. Miao, Z. Wang, X. Zhang, and Y. Chen, "K-Anonymity privacy protection algorithm for multi-dimensional data against skewness and similarity attacks," *Sensors*, vol. 23, no. 3, Article No.: 1554, 2023. DOI: https://doi.org/10.3390/s23031554.

[13] V. Torra and G. Navarro-Arribas, "Attribute disclosure risk for k-anonymity: The case of numerical data," *Int. J. Inf. Secur.*, vol. 22, pp. 2015–2024, 2023. DOI: https://doi.org/10.1007/s10207-023-00730-x.

[14] U. Sopaoglu and O. Abul, "A top-down k-anonymization implementation for Apache Spark," in *Proc. IEEE Int. Conf. Big*

Data, (Boston, MA, USA), Dec. 2017, pp. 4513–4521. DOI: 10.1109/BigData.2017.8258492.

[15] H. Zakerzadeh, C. C. Aggarwal, and K. Barker, "Privacy-preserving big data publishing," in *Proc. 27th Int. Conf. Scientific and Statistical Database Management*, Jun. 2015, Article No.: 26, pp. 1–11. DOI: https://doi.org/10.1145/2791347.2791380.

[16] N. Jha, L. Vassio, M. Trevisan, E. Leonardi, and M. Mellia, "Practical anonymization for data streams: z-anonymity and relation with k-anonymity," *Perform. Eval.*, vol. 159, Article No.: 102329, 2023. DOI: https://doi.org/10.1016/j.peva.2022.102329.

[17] F. Ashkouti and K. Khamforoosh, "A distributed computing model for big data anonymization in the networks," *PLoS ONE*, vol. 18, no. 4, Article ID: e0285212, 2023. DOI: 10.1371/journal.pone.0285212.

[18] K. R. Pandilakshmi and G. R. Banu, "An advanced bottom up generalization approach for big data on cloud," *International Journal of Communication and Networking System*, vol. 3, no. 1, pp. 12–15, June, 2014. DOI: 10.20894/IJCNES.103.003.001.003.

[19] A. Irudayasamy and L. Arockiam, "Parallel bottom-up generalization approach for data anonymization using MapReduce for security of data in public cloud," *Indian J. Sci. Technol.*, vol. 8, no. 22, pp. 1–9, 2015. DOI: 10.17485/ijst/2015/v8i22/79095.

[20] S. Kabou, S. M. Benslimane, and M. Mosteghanemi, "A survey on privacy preserving dynamic data publishing," in *Res. Anthol. Privatizing and Securing Data*, IGI Global, 2021, ch.79, pp. 1635–1657.

[21] K. Wang, P. S. Yu, and S. Chakraborty, "Bottom-up generalization: A data mining solution to privacy protection," in *Proc. 4th IEEE Int. Conf. Data Mining (ICDM)*, 2004, pp. 249–256. DOI: 10.1109/ICDM.2004.10110.

[22] B. B. Mehta and U. P. Rao, "Improved l-diversity: Scalable anonymization approach for privacy preserving big data publishing," *J. King Saud Univ.—Comput. Inf. Sci.*, vol. 34, no. 4, pp. 1423–1430, 2022. DOI: https://doi.org/10.1016/j.jksuci.2019.08.006.

[23] L. Gasmi, S. Kabou, N. Laiche, and R. Nichani, "Time series forecasting using deep learning hybrid model (ARIMA-LSTM)," *Stud. Eng. Exact Sci.*, vol. 5, no. 2, Article No.: e6976, 2024. DOI: https://doi.org/10.54021/seesv5n2-125.

[24] S. Kabou, L. Gasmi, A. Kabou, and S. M. Benslimane, "ImDMI: Improved distributed M-Invariance model to achieve privacy continuous big data publishing using Apache Spark," *Big Data Research*, vol. 40, Article ID: 100519, 2025. DOI: https://doi.org/10.1016/j.bdr.2025.100519

Salheddine Kabou,
Imad eddine Kimi, Arbia Boudaouad

Received June 04, 2025
Revised 1: December 30, 2025
Revised 2: January 26, 2026
Accepted February 12, 2026

Salheddine Kabou
ORCID: https://orcid.org/0000-0002-1423-7215
Higher Normal School of Bechar
Bechar city, Algeria
E–mail: kabou.salheddine@ensbechar.dz

Imad eddine Kimi
ORCID: https://orcid.org/0009-0005-6997-9194
Higher Normal School of Bechar
Bechar city, Algeria
E–mail: kimi.imad@ensbechar.dz

Arbia Boudaouad
ORCID: https://orcid.org/0009-0007-3735-4477
Ahmed Draia University
Adrar city, Algeria
E–mail: bouda.arbia@univ-adrar.edu.dz