ISSN: 2583-9888
DOI: doi.org/10.51219/JAIMLD/santosh-deshmukh/57

URF PUBLISHERS
connect with research world

Journal of Artificial Intelligence, Machine Learning and Data Science

https://urfpublishers.com/journal/artificial-intelligence

Vol: 1 & Iss: 4

Research Article

Improve Data Security Safeguards through Data Aggregation Purge Process

Santosh S Deshmukh*

Senior Member, IEEE, Consulting Project Manager, USA

Citation: Deshmukh SS. Improve Data Security Safeguards through Data Aggregation Purge Process. *J Artif Intell Mach Learn & Data Sci 2023*, 1(4), 214-213. DOI: doi.org/10.51219/JAIMLD/santosh-deshmukh/57

Received: 01 November, 2023; **Accepted:** 25 November, 2023; **Published:** 27 November, 2023

***Corresponding author:** Santosh S Deshmukh, Consulting Project Manager, USA Email: san_desh3@yahoo.com

ABSTRACT

"Data is money", in today's digital economy, data is often equated to money due to its immense value in informing decisions, driving strategies, and personalizing customer experiences. Organizations harness data to gain competitive advantages, improve operations, and create new revenue streams. Thus, managing, analyzing, and securing data have become pivotal to success in virtually all industries. Expecting data left on the shelf to continually yield more valuable insights after every reprocessing risk organizational resources unnecessarily. The costs and risks associated with safeguarding dormant data often outweigh the potential benefits, suggesting a need for a more strategic approach to data management and analysis. This article suggests a mutually beneficial solution for both data consumers and providers. It proposes that after data is analyzed and insights are generated, the original source data can be deleted to ensure privacy and security. This allows for the retention of analytical results for future reference or analysis, without the need to store the original data, addressing concerns about data security and privacy while maintaining the utility of the analysis.

Keywords: Data retention; Data analysis; Analytics; Data Aggregation; Data Security; Data Protection; Machine Learning

1. Introduction

Data refers to raw, unprocessed facts and figures that, in isolation, may not carry any specific meaning. It consists of numbers, characters, symbols, or images that are collected through observations, measurements, research, or analysis[1]. Data is the primary input that needs to be processed and analyzed to derive meaning. **Information**, on the other hand, is data that has been processed, organized, or structured in a way that adds context and relevance, making it useful and meaningful to the recipient. Information arises from interpreting or analyzing data, thereby providing insight, and conclusions, or aiding in decision-making. In summary, data is the raw material that, when processed and interpreted, becomes information - the meaningful output that informs, clarifies, and aids understanding and decision- making processes[2].

As we navigate through the vast span of the internet and engage with various digital platforms, every click, every search, and every interaction generate data. This data, when curated and analyzed, transforms into valuable information, offering insights that can drive decision-making, influence consumer behavior, and shape the future of businesses.

In our digital age, the scale of data we generate, and share is immense and multifaceted, encompassing personal health records, financial behavior, mundane shopping lists, sensitive insurance particulars, identifiers like driving license and social security numbers, biometrics such as face IDs, intimate medical histories including medications and allergies, earnings and losses, and all manner of tax-related. documentation. The transmission of such data, often necessitated by legal, medical, or commercial requirements, has become a commonplace and legally sanctioned aspect of everyday life. Frequently, this data serves as the bedrock for analytical endeavors, underpinning targeted marketing campaigns, informing trend analyses, and guiding business promotions among other uses. However, with the exchange of such sensitive information comes a profound

responsibility. The responsibility of safeguarding this data falls squarely upon the shoulders of its recipients.

Certain sectors operate under strict regulatory frameworks like (GDRP or HIPAA) with stringent data compliance policies, including defined data retention periods. Conversely, industries without such mandated policies may retain data indefinitely. This disparity can introduce significant risks to data security, highlighting the need for a balanced approach to data retention that safeguards privacy and compliance across all sectors. All business who captures the data from the general public, should remind themselves of the chapter of ethics all the time and ask the questions "Why should I keep this data?" If content has outlived its usefulness to your business, you should delete it[3].

Indeed, the implementation of a data aggregation purge solution is not universally prescriptive but rather contingent upon the strategic priorities and operational exigencies of individual organizations. It is largely a reflection of an organization's disposition towards data retention and its governance policies. Certain institutions, such as research organizations, often have a compelling need to preserve data comprehensively. The mandate to retain data indefinitely in such contexts supersedes the impetus to purge. Conversely, there exist entities for whom data serves a less critical role, where the risks associated with data retention - such as theft or breach - do not carry the same weight of consequence. Such organizations may deal with data that is non-sensitive or oblique to human life and privacy, rendering stringent data preservation less crucial.

Life Classification of Data

In human life, data can be categorized into distinct lifecycle types, reflecting the frequency and circumstances under which they change. These types range from static, lifelong information to data that undergoes continuous updates. While these categories can apply to any living thing, let's focus on the data woven into the fabric of our daily lives[4].

1. Immutable Data (One-time Life): This data forms the bedrock of our identity. The constancy of this data simplifies its management, providing a stable foundation for various systems and processes to build upon. Elements like name, date of birth, and biological makeup are persistent, remaining constant throughout our lives. This permanence simplifies storage, but a breach can have lasting consequences. As once compromised, such information remains perpetually vulnerable, its integrity irrevocably breached.

2. The Evolving Landscape (Infrequent Life): This category encompasses information that undergoes occasional shifts often measured in decades rather than years or months, marking pivotal chapters in an individual's life narrative. Changes in residence, marital status, or legal documents like driver's licenses fall into this group. These updates, while less frequent, require vigilance to ensure accuracy.

3. The Shifting Sands (Frequent Changes): More ephemeral in nature, dynamic data captures the evolving facets of our personal and professional lives. This data reflects the dynamic nature of our lives. Academic qualifications, professional certifications, or car ownership are examples. These details update periodically, over the years, demanding regular checks and updates.

4. Ephemeral (Continuous Changes): This category captures the constant flux of our existence. Daily activities like grocery lists, location data, food we consume, or health metrics fall under this umbrella. This data normally represents out habits and lifestyle patterns. This data offers a real-time snapshot, constantly generated and requiring ongoing management.

Together, these four categories form a comprehensive framework for understanding the lifecycle of data in human life. Each category not only highlights the diversity and complexity of the data we generate but also underscores the varying implications for privacy, security, and data management strategies. Understanding these categories helps us navigate the ever-growing data landscape. It allows us to prioritize information security based on its criticality and develop strategies for efficient storage and management. Ultimately, recognizing the spectrum of human data empowers us to harness its potential while safeguarding our privacy.

The stability of data enhances its vulnerability, necessitating stronger protection measures, especially for information that underpins an entity's identity over extended periods, marking a distinct individuality without predictable patterns due to minimal changes. The greater the immunity of data to change, the more it beckons for fortified protections. Conversely, swiftly changing data exhibits discernible patterns, offering valuable insights for informed decision- making, contrasting the static nature of persistent data with the dynamic and actionable intelligence of rapidly evolving information.

The classification of data by its lifetime aims to establish its relevance duration, guiding the determination of its appropriate retention period. This approach helps in managing data efficiently by recognizing its varying degrees of importance and utility over time. By understanding the distinct temporal characteristics of data-from static and enduring to ephemeral and fleeting-organizations and individuals can make informed decisions about data management practices. This includes how to store, protect, and eventually retire data in a manner that aligns with its inherent longevity and significance (**Figure 1**).

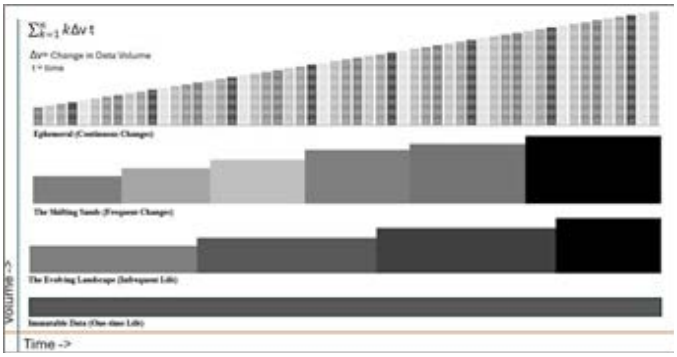

Figure 1. Figure represents the four data lifetime classifications. Source: Author

Δv represents the change in data volume over the period t. The subsequent classification type represents the level of aggregation of data over the period. **Data aggregation** is the process of collecting and summarizing information from various sources, aiming to gather related data to provide comprehensive insights, analyze trends, or make decisions. This process involves compiling detailed data into a more digestible and useful format, often used in statistical analysis, business intelligence, and within various technological frameworks to enhance decision-making processes.

In the given example 'immutable Data' does not require further data aggregation. However, the other data types can be aggregated if they are statically in nature.

Further, we understand the importance of aggregation as it will address some of the solutions towards making an attempt to reduce data breaches and keeping the data secure enough without indulging in the security of data.

Traditional Data Analytical Process

Data's journey begins with its varied forms, sizes, and sources, transitioning through capture to secure storage locations like expansive data warehouses, versatile data lakes, or the increasingly prevalent cloud storage solutions. This initial phase is crucial, as it lays the foundational stone for data integrity and accessibility. Originating from a diverse spectrum of sources, this data embarks on a complex journey before unveiling its inherent value. This data undergoes refinement and organization, readying it for analysis, often referred to as data wrangling or massaging- where it is cleaned, normalized, and structured. This step is vital for mitigating inconsistencies and ensuring the data is in a usable state. Subsequently, the refined data is moved to a curated zone, a specially designated area where it is primed for consumption. With the data now in its most pristine and organized form, it is fed into a stack of analytical tools and platforms, each selected for its capability to handle specific data types and volumes. These tools, powered by sophisticated algorithms and computing processes, dive deep into the data, unraveling patterns, trends, and anomalies that were previously obscured. Analytical tools dive into this prepared data, tailored to its specifics, to extract actionable insights. These insights should not only illuminate understanding but also compel action-whether it's optimizing operations, enhancing customer experiences, or identifying new market opportunities.

The entire cycle-from sourcing to insight generation- relies on the retention of original and processed data on durable media, safeguarded until needed again, emphasizing a cycle of use, analysis, and protection integral to leveraging data effectively. Ensuring the longevity and reusability of this data, the source material and processed information are meticulously preserved on persistent media.

This iterative process underscores the critical nature of data management and security throughout its lifecycle, highlighting the value of insights derived from well-curated and analyzed data. This safeguarding is pivotal, for the value of data is not solely in its initial use but in its potential for re- analysis, comparison, and historical reference. Through this intricate and nuanced process, data transcends its raw state, becoming the lifeblood of informed decision-making and strategic foresight in the modern era.

In an ideal scenario, once insights have been extracted and actions formulated, the original (sourced or raw) and processed data remain stored, ready for future re-analysis. This necessitates that, until new data supplants it, considerable effort be dedicated to maintaining and securing this stored data. This dormant period persists until the arrival of new data necessitates their reactivation. During this interim, the data, both in its emerging and refined forms, is meticulously preserved, awaiting its next moment in the spotlight. It is within this waiting period that a significant commitment of resources is necessary to ensure the ongoing

security and maintenance of this valuable digital asset. This preservation phase is not merely a matter of storage but a critical endeavor to safeguard the integrity and confidentiality of the data against evolving threats. This diligent stewardship ensures that when the time comes, the data remains a pristine and compelling resource for driving continued innovation and informed decision-making.

The question is, is this necessary? In the era of advanced technology, superior data modeling and analytical capabilities, and frequent posing of security threats, is this worth retaining the source information that has been already processed?

This article proposes an innovative approach by using the outcomes of previous analyses to fuel the next wave of insight discovery, rather than revisiting already scrutinized data. It advocates for the deletion of original and processed data post-analysis to mitigate risks of unintended exposure and enhance security, thereby streamlining the insight generation cycle and reinforcing data protection protocols.

Data Aggregation Purge Process

To grasp the essence of this strategy from a broader perspective, the primary goal is to declutter the data repository by discarding any source data deemed redundant after its analysis and the subsequent extraction of insights. This action is taken because, post-analysis, such data becomes superfluous, eliminating the need for its further storage or management. This proactive measure not only simplifies data handling processes but also significantly mitigates risks associated with data security and potential breaches. By ensuring that only pertinent, current data is retained, we enhance the repository's efficiency and security, streamlining operations and safeguarding valuable information against unauthorized access or exposure (**Figure 2**).

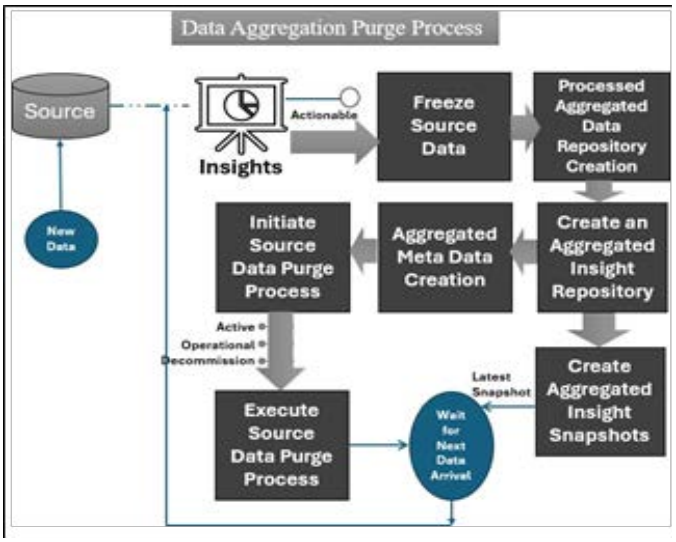

Figure 2. Simplified flowchart of Data Aggregation Purge Process. Source: Author

This process will not intervene in the existing data analytical process as described in the previous section. Rather this process will trigger after the insights are prepared and analytical results are frozen.

Freeze Source Data: The snapshot of source data on which the analytics was performed will be frozen and no more changes will be allowed to ensure the integrity of the results drawn as part of the analysis. This can be hard freeze or soft freeze by tagging the candidate data elements. The frozen data can be kept aside as it won't be required anymore.

Processed Aggregated Data Repository Creation: In the event of this whole process, the data insights are already created. This processed data with additional attributes will bring more meaning to the processed data. The additional attributes should

suffice the need in case of insights repository (next step) needs any reference to source data. For example, adding a timeline/ time series to aggregated data. Adding geolocations, universal IDs, identifying additional statistical attributes, adding P -value, or any attributes that might be useful for descriptive, diagnostic, predictive, and prescriptive analytics. This process also demands that the required analytics model be rerun to bring more attributes and insights to this repository. This step is a prerequisite for the next Analytical repository building.

Create an Aggregated Insight Repository: This involves crafting a specialized data storage facility where comprehensive analytical outcomes are housed. Formed from the processed data collated in prior stages, this repository is a refined collection of significant statistical analyses and valuable analytics. Despite its extensive nature and the potential inclusion of repetitive aggregated attributes, its value is unequivocal. The insight repository stands as a vault of refined data, ready to substitute the raw source material for any subsequent analytical ventures. In scenarios where re-analysis is crucial, this repository offers a wealth of insights, precluding the necessity to revisit the original, unprocessed source. By integrating exhaustive analytical information, it not only facilitates swift re-examination of data but also fortifies the overall data handling process against repetitive accumulation. This design is tailored for the efficient recycling of insights, ensuring that the fountain of knowledge within the repository can be tapped into repeatedly, rendering the raw source data redundant for future analysis.

Aggregated Metadata Creation: Metadata is data about data. Metadata is imperative in all data management and data processing activities. Aggregated Metadata will help set the context of the Aggregated Insight Repository. It will define the relevance, use, and relationship and describe the data that will be retrieved, reused, and reprocessed.

Create Aggregated Insight Snapshots: The Aggregated Insight will be accessed and used repetitively every time the new data arrives. After every successful cycle of regeneration of Aggregated Insight, the old will be achieved and the new will become active. This specific arrangement is to ensure the old versions are not lost and can be pulled in at any time when they are required. Collectively all snapshots represent cumulative Insights.

Initiate and execute the Data Purge Process: This will logically end the purge workflow. This step activates upon the culmination of the data aggregation and insight assimilation. After every data arrival, an aggregated data repository, aggregated insight repository, and metadata will be created and updated, snapshots will be reconstructed. This whole process will be efficiently streamlined. Once the completion of the whole cycle is ensured, the data purge process will be triggered. This process will identify all data points and actual data which has been flown through the cycle and are no longer required further. The data will be tagged and processed for further purge. The purging act is both logical and systematic, designed to excise data without disrupting the repository's structure or the availability of vital insights. By doing so, it maintains a lean and relevant data environment, ready for the subsequent wave of incoming data and the ensuing analytical process.

The purge can use 3 strategies, Data that remains potentially valuable for upcoming analytical cycles is put on hold, thus enabling the data to be **active** through several iterations. This is particularly applicable for data that feeds into models requiring longitudinal analyses or those that may benefit from a more extensive temporal context. In the **operational** queue, data is not immediately decommissioned but is instead removed from the primary operational workflow. It resides in a dormant state within the repository, akin to being on standby. This data is retrievable in the short term, should it be deemed necessary for unexpected analytical needs or for validation of recent insights. For data assessed as non-critical to ongoing operations or future analysis, an immediate **decommission** strategy is adopted. This data is expunged from the system promptly to free up resources and maintain a streamlined database.

By applying these approaches, the system ensures that data management is not only responsive to current needs but is also strategically aligned with the anticipated future requirements of the analytical process.

New Event / New Data Arrival: The duration of new data arrival can be in seconds, minutes, days, or years depending on the data classification type. The New data arrival is the starting point to trigger the cycle/workflow. The next data will not push in the cycle unless first finished or can be effectively managed in parallel with proper segregation of data sets.

In essence, the objective of each cycle here is to systematically refine the repository's contents, reassemble the snapshots to reflect the most current analytical insights and execute a data purge upon cycle completion. The purging component is pivotal; it ensures that data, once it has outlived its utility, is removed from the system. This act of data sanitization serves a dual purpose: it maintains the repository's relevance and efficiency by eliminating redundancy, and it significantly reduces security risks. By disposing of data that is no longer necessary, we preemptively mitigate the potential threat of data theft or breach, strengthening the integrity and trustworthiness of our data environment.

"Data is money", much like currency, the more data we accumulate and process, the more its marginal value can decrease unless it's directly contributing to the generation of insights. In this continuous process of data consumption and analysis, it's the extraction of actionable intelligence that represents the peak of data's worth. Once we distill the actionable insights, the original, raw source data's value depreciates significantly, often to the point where it becomes expendable. However, to be successful here, other areas need to be supportive of the purge process. Firstly, the ingrained habit of indefinitely storing raw data 'just in case' it may prove useful needs to evolve. Instead, a more strategic, purpose-driven approach to data retention should be adopted, prioritizing data that has a clear utility for future analytics. Secondly, the core of this system, the Aggregated Insight Repository, must possess a robust data modeling capability. This is vital to extract the maximum number of insights and to ensure that the processed data serves a strategic function. Thirdly, as data analytics is inherently time-intensive, maintaining momentum in the cycle is critical. The cadence of new data introductions and the agility of processing will dictate the pace and effectiveness of the entire operation.

By eliminating redundant data, we reduce the 'attack surface' that could potentially be exploited in breaches, thereby protecting the integrity of the system and, in broader applications, safeguarding human lives by preventing sensitive information from falling into the wrong hands.

It maintains the repository's relevance and efficiency by eliminating redundancy, and it significantly reduces security risks. By disposing of data that is no longer necessary, we preemptively mitigate the potential threat of data theft or breach, strengthening the integrity and trustworthiness of our data environment.

Conclusion

"Data is money", much like currency, the more data we accumulate and process, the more its marginal value can decrease unless it's directly contributing to the generation of insights. In this continuous process of data consumption and analysis, it's the extraction of actionable intelligence that represents the peak of data's worth. Once we distill the actionable insights, the original, raw source data's value depreciates significantly, often to the point where it becomes expendable. However, to be successful here, other areas need to be supportive of the purge process. Firstly, The ingrained habit of indefinitely storing raw data 'just in case' it may prove useful needs to evolve.

References

1. https://www.techtarget.com/searchdatamanagement/definition/data-management
2. Dhudasia M, Grundmeier R, Mukhopadhyay S. Essentials of data management: An overview. Pediatric Research, 2021; 93: 2-3.
3. https://blog.box.com/what-is-a-data-retention-policy
4. Mandala, Vishwanadham, Mahindra Sai Mandala. Anatomy of big data lake houses. Neuroquantology, 2022; 20: 6413.