Not Small - Not Big Data: the Missing Size in the Data Spectrum

Schmidt Peter[1], Krčová Ingrid[2], Zelinová Silvia[2], Simonka Zsolt[2]

[1] Department of Applied Informatics, Faculty of Economic Informatics

[2] Department of Mathematics and Actuarial Science, Faculty of Economic Informatics

University of Economics in Bratislava, Dolnozemská cesta 1, 852 35 Bratislava, Slovakia

e-mail: peter.schmidt@euba.sk, ingrid.krcova@euba.sk, silvia.zelinova@euba.sk, zsolt.simonka@euba.sk

Abstract: This paper addresses the classification of data sets that do not fit into the traditional categories of Small Data or Big Data, introducing an intermediate category called Not-Small-Not-Big Data (NoS-NoB Data). Understanding and effectively handling NoS-NoB Data is crucial for various fields of data science, as it encompasses structured, semi-structured, and unstructured information that challenges conventional computing environments due to hardware and software limitations. NoS-NoB Data is primarily defined by its volume, which surpasses the capacity of standard office tools yet does not necessitate the complexity of distributed big data technologies. Typically ranging from tens of gigabytes to several terabytes, such data require specialized processing approaches, including advanced database management systems and programming languages with dedicated data analysis libraries. Processing NoS-NoB Data effectively demands a combination of relational databases, NoSQL solutions, and computational tools such as Python, R, and, in some cases, Hadoop or Spark, depending on specific analytical requirements. The application domains of NoS-NoB Data span various industries, including small and medium enterprises (SMEs), healthcare, education, retail, financial services, logistics, and public administration. This study summarises the fundamental distinctions between data categories based on volume, structure, processing methods, and technological requirements. Additionally, a visual representation illustrates the relationships between Small Data, NoS-NoB Data, and Big Data, highlighting the overlap between categories and the proportional representation of structured, semi-structured, and unstructured data.

Keywords: Small Data, Big Data, Not-Small-Not-Big Data, intermediate data, data science, data processing, NoSQL, Python, Hadoop

1 Introduction

The rapid growth of data has made it an essential element of modern technology, leading to the need for precise categorization based on volume and processing requirements. Traditionally, data has been classified as either small data, which can be quickly processed using conventional tools, or big data, which requires extensive computational resources and specialized infrastructure (Davenport & Harris, 2017). However, the increasing size and complexity of datasets that do not fit neatly into either category necessitate a more refined classification (Duarte, 2023).

To address this gap, we introduce the concept of Not-Small-Not-Big Data (NoS-NoB Data), which encompasses datasets that exceed the capabilities of standard office tools but do not require full-scale big data infrastructure (Serrant, 2023). These datasets present unique processing challenges, such as handling a 3 GB text file, which is problematic for most text editors, or working with a 10 GB CSV file, which cannot be loaded into a typical spreadsheet. Processing such data demands alternative software solutions or customized scripts in languages like Python.

The necessity of NoS-NoB Data is further highlighted by the high failure rate of big data technology implementations, with Gartner reporting a failure rate of 85-90% and IDC also acknowledging significant implementation challenges (Walker, 2017). It is believed that many organizations that have attempted to implement big data solutions have found that their datasets do not qualify as big data, making tools like Hadoop and Spark unnecessarily complex and inefficient. This suggests that a significant number of failed implementations may stem from a mismatch between the actual size of the dataset and the chosen technologies.

Furthermore, academic surveys of dataset classifications have revealed a lack of consensus on the precise boundaries between small, medium (Inmon, 2019), and big data. Students in Big Data courses often ask for precise definitions of these categories, but the existing literature does not provide definitive answers. This ambiguity motivated us to explore why these boundaries remain undefined and whether they can be determined empirically. For example, students asked how to process structured datasets exceeding 1.5 million rows without relying on traditional databases, highlighting the need for affordable and efficient solutions for medium-sized datasets.

Other practical challenges emerged from computational experiments. For example, students using brute-force methods to solve the Traveling Salesman Problem (TSP) encountered severe performance bottlenecks, even for small datasets (e.g., 12 nodes and 13 edges). The resulting computation generated a 38 GB file that took more than 24 hours to process, illustrating how datasets not traditionally considered “large” can still overwhelm conventional tools and workflows. Moreover, the increasing role of datafication, i.e., the transformation of real-world entities and business processes into digital data, further underscores the need for a classification that bridges the gap between small and big data. Misut and Jurik (2021) emphasize that data-driven decision making relies on effectively structuring and processing

data, particularly in logistics and enterprise decision support systems, where classification models and analytical tools are essential for managing large but not necessarily big datasets.

Building on these observations, we formally define and delineate the concept of NoS-NoB Data, distinguishing it from both small and big data. This study explores the practical challenges associated with processing NoS-NoB Data and proposes suitable tools and methodologies to address them. Additionally, we examine its applications across various domains, including small and medium-sized enterprises (SMEs) (Tawil et al., 2023), healthcare (Zhang et al., 2019), and public administration (Dossouza & Jacob, 2017).

This paper does not aim to provide an exhaustive literature review or experimental validation. However, it focuses on clearly defining the concept of NoS-NoB Data, identifying its challenges, and exploring its practical applications. Doing so contributes to a more structured and practical understanding of NoS-NoB Data, bridging the gap between traditional small and big data classifications.

2 Definition of Not Small-Not Big Data

"Not Small - Not Big Data" (NoS-NoB Data) refers to data that surpasses the processing capabilities of office tools designed for structured data or the capacity of traditional database systems running on low-end servers. However, it is not large enough to necessitate specialized big data processing tools and techniques, such as cloud-based or distributed computing frameworks (e.g., Hadoop technologies) (Mishra, 2019). Typically, NoS-NoB Data ranges from 10 to 100 GB, although the exact threshold varies depending on the application domain. While such data can be managed using advanced database systems, it does not inherently require large-scale distributed processing or complex machine learning algorithms. Moreover, NoS-NoB Data may include relatively small files (in the MB range) whose structure renders them incompatible with conventional desktop tools. For example, a table containing 15 million rows may exceed the processing limits of Excel, necessitating alternative solutions such as Python or other programming languages equipped with appropriate libraries."

2.1 Characteristics of NoS-NoB Data

Before we delve into the specific characteristics of NoS-NoB Data, it is important to understand the context in which this data operates. NoS-NoB Data represents a path between small and big data, requiring a unique approach to processing and analysis. Below are the key characteristics of NoS-NoB data:

- *Size*: NoS-NoB Data is larger than traditional data but smaller than typical big data. It typically ranges from tens of gigabytes to several terabytes.

- *Complexity*: NoS-NoB Data can contain structured, semi-structured, and unstructured data, but generally to a lesser extent than big data.
- *Generation Speed*: NoS-NoB Data is generated faster than traditional data but not as fast as big data.
- *Processing*: NoS-NoB Data can be processed by relational database systems, NoSQL databases, or more advanced data analytics tools, but it does not require extensive parallel processing.
- *Analytics*: NoS-NoB Data requires advanced analytics, but not at the level of big data.

2.1.1 Size

A defining characteristic of NoS-NoB Data is its size, which significantly influences the selection of appropriate data processing and storage solutions. Several key factors contribute to its classification, including the size of individual files, the overall dataset composition within a project, and the demarcation of NoS-NoB Data relative to Small and Big Data. The following sections provide a detailed examination of these aspects, elucidating their implications for data management, computational efficiency, and system compatibility.

File size: The size of individual files is crucial, as it can significantly influence the choice of data processing and storage technology. Large files may require systems capable of efficiently handling large volumes of data simultaneously, such as NoSQL databases or distributed file systems.

Project size: In projects where the raw database (dataset) comprises numerous small files that must be processed collectively, significant challenges arise in organizing and coordinating access to these files. The efficiency of file system operations and the capacity for parallel data processing are critical for addressing such issues. Technologies like Hadoop and Spark are designed to handle massive datasets exceeding terabytes, leveraging distributed computing for efficient processing (Abu-Salih et al., 2021). By enabling distributed processing and optimizing the handling of large datasets, these tools facilitate the efficient management of numerous smaller files. Notably, even when individual files are relatively small (measured in hundreds of kilobytes), their combined volume can amount to tens of gigabytes. While these files might individually fall into the category of Small Data, the need to process them collectively to identify patterns or trends renders manual processing extremely inefficient.

Upper bound: However, Big Data, which frequently surpasses terabytes in size, NoS-NoB Data typically ranges from tens to hundreds of gigabytes. This range defines its unique position in the data spectrum, bridging the gap between manageable Small Data and the extensive requirements of Big Data.

Lower limit: The lower limit of NoS-NoB Data does not necessarily exceed the capabilities of traditional computing systems. In practice, datasets in the megabyte range are often encountered, but their structure may render them inefficient for

processing with standard office applications such as Excel or Power BI. In such cases, software compatibility or performance becomes the limiting factor.

Traditional relational databases can efficiently process gigabytes of data when deployed on adequately configured servers, provided the database structure is well designed and optimized. However, many organizations that have not transitioned to cloud-based solutions often rely on outdated technologies. Consequently, we categorize NoS-NoB Data as any dataset that cannot be processed efficiently on typical office computers (e.g., systems with a CPU i7, 16GB RAM, and a 1TB SSD). Based on current capabilities, this category includes data in the range of 10 to 20GB.

The efficiency of traditional relational database systems (RDBMS) decreases as data volumes increase. The specific point at which this occurs depends on multiple factors, including hardware configuration, database structure, query complexity, and database optimization. For instance, a server equipped with a 4-8 core CPU, running a Linux or Windows Server operating system, coupled with database software such as Oracle Database or MS SQL Server, a 1-2TB SSD, and 32GB of RAM, is generally sufficient for processing NoS-NoB Data. Such a configuration can handle data volumes ranging from 10 to 100GB without significant performance degradation, assuming optimal indexing, well-designed queries, and effective resource management. The key factors for effective NoS-NoB data processing are database optimization, indexing, properly designed queries, and adequate resource management. These factors significantly affect how quickly and efficiently the system processes data.

2.1.2 Data Complexity

The classification of NoS-NoB Data is not solely determined by its size but also by its structural characteristics, which influence how it is processed and analyzed. Depending on its degree of organization, NoS-NoB Data can be categorized into structured, semi-structured, and unstructured, each requiring distinct processing approaches and computational tools.

Structured Data: Structured data refers to traditional database formats, such as tables and records, with fixed formats that are relatively straightforward to process. In the context of "Not Small-Not Big Data" (NoS-NoB Data), structured data often includes extensive records from systems like customer relationship management (CRM) platforms, financial transactions, or inventory lists. Although these datasets do not require the extensive infrastructure of Big Data systems, they necessitate more advanced tools and techniques than ordinary office data to ensure efficient processing. Effectively managing structured NoS-NoB Data enables organizations to extract deeper insights and make more informed business decisions.

Semi-Structured Data: Semi-structured data, such as XML or JSON files, tends to grow rapidly in volume while remaining manageable for less powerful systems. Examples include email archives, website logs, or data obtained via APIs. Unlike

traditional structured data, semi-structured data contain some organizational framework but lack the rigid definition of database records. This complexity demands advanced methods for extracting and analyzing meaningful information, often involving tools that handle hybrid data formats.

Unstructured Data: Unstructured data encompasses formats such as text documents, images, and videos. While this type of data can be relatively large, its volume in the context of NoS-NoB Data typically does not reach the scale that necessitates massive Big Data infrastructure. Nevertheless, processing unstructured NoS-NoB Data often benefits from Big Data technologies tailored to mid-sized datasets, allowing for effective management and analysis within these constraints.

2.1.3 Generation Speed

The generation of Not-Small-Not-Big Data (NoS-NoB Data) occurs faster than traditional data but slower than Big Data. This category typically includes data streams originating from online transactions, medical sensors, or records of customer behavior. Although NoS-NoB Data does not necessitate massively distributed processing systems, its speed, volume, and diversity demand more efficient and faster database systems or data warehouses.

Small-scale distributed systems with several interconnected nodes are often employed to handle these requirements. These systems provide the computational power and scalability to manage the increased speed and complexity of NoS-NoB Data streams, ensuring efficient storage, processing, and analysis.

2.1.4 Processing

The processing of NoS-NoB Data encompasses several critical aspects that facilitate its efficient utilization without necessitating the full-scale deployment of Big Data infrastructures. Key stages in this process include data collection, storage, management, and analysis.

Given their distinctive size and complexity, the processing of NoS-NoB Data requires a balanced approach. While these datasets do not demand the extensive infrastructures typically associated with Big Data, they nonetheless necessitate more advanced technologies and methodologies than those employed for traditional small-scale datasets. Organizations can effectively manage NoS-NoB Data and extract meaningful insights by leveraging appropriate tools and techniques.

The following sections outline the fundamental aspects of NoS-NoB Data processing:

Data collection

- Integration of different sources: NoS-NoB Data can come from different sources, such as internal databases, online transactions, social media, or sensors. Integrating these sources is important to ensure the uniformity and completeness of the data sets.

- *Automatic collection:* Using automated tools to collect data from various sources simplifies the process and increases efficiency.

Data storage
- *Advanced database systems:* Unlike more straightforward and open-source databases such as Access, JavaDB, or MariaDB, NoS-NoB Data require more advanced database systems to handle greater data size and complexity. These systems may include high-performance relational or NoSQL databases (Taylor, 2023).
- *Optimization of storage:* Efficient storage of this data may require data compression, proper indexing, or optimized database design to minimize data redundancy.

Data management
- *Data quality:* Maintaining high data quality through data cleaning and inconsistency removal processes is important.
- *Security and privacy:* If NoS-NoB Data contain personal or sensitive information, it is important to ensure its adequate protection and compliance with legal regulations regarding data privacy.

Data analysis
- Advanced analytical tools: To analyze NoS-NoB Data, it is necessary to use tools that can handle its diversity and volume. These may include data analysis software, BI (Business Intelligence), and data visualization tools.
- Machine learning and AI: For more complex analyses, machine learning and artificial intelligence algorithms can reveal patterns and trends in data.

2.2 Analytics - Tools and Techniques

The following tools and platforms are widely employed in data processing and analysis, offering extensive capabilities for handling NoS-NoB Data. Their selection should be tailored to the organization's specific requirements and objectives.

Data Preprocessing and Cleaning
- *Pandas:* A Python library for data manipulation, cleaning, transformations, and normalization.
- *Apache NiFi:* An open-source data flow management system that enables automation and control of data flow between different systems.
- *Talend:* A data integration tool that supports ETL processes, allowing users to transform and integrate data from various sources efficiently.

Database Analysis
- *MySQL/Microsoft SQL Server:* Popular relational database systems suitable for structured data in the NoS-NoB range.
- *MongoDB/Cassandra:* NoSQL databases suitable for working with semi-structured and unstructured data.

Business Intelligence (BI) and Visualization
- *Tableau:* Enables advanced data visualization and is suitable for interactive BI reports.
- *Microsoft Power BI:* Provides comprehensive BI for data analysis and visualization.
- *Qlik:* Offers tools for data analysis and visualization using self-service BI.

Machine Learning and AI
- *Scikit-learn:* A Python library for machine learning, suitable for various analytical models.
- *TensorFlow/Keras:* Deep learning libraries used for more complex analyses, such as image and text analysis.
- *RapidMiner:* A platform that provides a visual programming language for machine learning and advanced analytics.

Predictive and Prescriptive Analysis
- *SAS/SPSS:* Traditional statistical tools suitable for predictive analysis.
- *R*: A programming language and environment for statistical calculations and graphics.
- *Python with libraries like NumPy and SciPy:* Provides a robust background for scientific computing and can be used for predictive analytics.

Challenges and solutions in analysis
Although NoS-NoB Data is not as extensive as Big Data, its volume can still present challenges for specific analytics tools. The heterogeneity of data formats and the need to analyze structured, semi-structured, and unstructured data necessitate distinct methodologies and techniques. Data quality is a fundamental factor that directly influences the reliability of analytical outputs. Inaccurate or incomplete data can lead to misleading conclusions, undermining the effectiveness of data-driven decision making.
To address these challenges, various tools have been developed to facilitate the analysis and processing of NoS-NoB Data:
- *DataRobot:* An automated machine learning platform that helps solve predictive analytics challenges.
- *Alteryx:* A comprehensive platform offering tools for data integration, analysis, and visualization, simplifying the processing of NoS-NoB Data and improving analytical efficiency.

2.2.1 Key Steps for Managing and Analysing NoS-NoB Data

The management and analysis of NoS-NoB Data involves several essential steps, each requiring careful consideration to ensure accuracy, efficiency, and security.

Identifying Objectives: The first step is establishing clear objectives, such as defining target analytics, enhancing customer satisfaction, or optimizing internal processes. These objectives determine the relevant data sources and the appropriate analytical approaches.

Data Collection and Storage: Implementing suitable data acquisition and storage mechanisms is crucial. Depending on the data structure and volume, organizations may use relational database management systems (e.g., MySQL, MS SQL Server, PostgreSQL), NoSQL databases (e.g., MongoDB, Cassandra), or cloud-based solutions (e.g., AWS S3, Google Cloud Storage). Distributed frameworks such as Apache Hadoop facilitate the management of large datasets.

Data Cleaning and Transformation: Ensuring data quality is fundamental. Tools such as Pandas in Python or Apache NiFi assist in removing duplicates, correcting errors, and transforming raw data into an analyzable format, thereby improving data integrity.

Data Analysis: Various analytical techniques can be applied, ranging from traditional statistical methods (e.g., SAS, SPSS, R) to machine learning approaches (e.g., Scikit-learn, TensorFlow, Keras) to uncover patterns, detect anomalies, or predict future trends (Chollet, 2021).

Data Visualisation and Communication: Effective presentation of analytical insights is essential for decision making. Tools such as Tableau, Microsoft Power BI, and Qlik facilitate the visualization, interpretation, and communication of results, ensuring clarity for stakeholders.

Data Security: Protecting data from unauthorized access and potential loss is critical. Security measures include encryption protocols, regular backups, and access control mechanisms to maintain data integrity and confidentiality.

Managing and analyzing NoS-NoB Data requires selecting appropriate tools and technologies tailored to the project's specific needs and objectives. For example, relational databases may be adequate for straightforward data processing tasks, whereas distributed frameworks like Apache Hadoop or analytical libraries in Python may be necessary for comprehensive data analysis. While this wide range of available technologies poses challenges in selecting the optimal solution, it also offers flexibility in addressing diverse analytical requirements.

3 Application Domains and Comparative Analysis of Small Data, NoS-NoB Data, and Big Data

Data's application in various domains depends on its size, structure, and processing requirements. While Small Data and Big Data have been well-defined in prior literature, the increasing number of datasets that do not fit neatly into either category necessitates a more refined classification (Banafa, 2018). However, NoS-NoB Data represents an intermediate category, requiring more sophisticated approaches than Small Data but not the full capabilities of Big Data. This section provides a comparative overview of data application in key domains, highlighting differences in data characteristics, processing needs, and practical implications.

3.1 Key Application Domains and Data Characteristics

Small and Medium Enterprises (SMEs): SMEs often generate NoS-NoB Data through customer transactions, CRM systems, and web interactions. Processing and analyzing this data can help better understand customer behavior and optimize marketing strategies.

Healthcare: Smaller hospitals and clinics generate significant medical records and data from medical facilities. By analyzing this data, healthcare facilities can improve patient care and the efficiency of their services.

Insurance and reinsurance: Insurance companies generate vast data through policy administration, claims handling, and customer interactions. This data encompasses structured formats (e.g., policy details), semi-structured formats (e.g., claims descriptions), and unstructured formats (e.g., customer communications). The analysis of NoS-NoB Data enables insurers to enhance risk assessment, customize insurance products, optimize claims processing efficiency, and improve customer service. Furthermore, NoS-NoB Data plays a critical role in fraud detection, actuarial modeling, and accurate premium calculation, ensuring that pricing structures appropriately reflect the risk profiles of policyholders. Leveraging advanced data analytics, insurance and reinsurance companies can develop more precise underwriting models and strengthen their overall risk management strategies.

Education: Colleges and universities use NoS-NoB Data to track student performance, manage courses, and oversee research projects. By analyzing student records, attendance data, and online learning interactions, institutions can refine curricula, enhance adaptive learning strategies, and improve student retention rates. Additionally, processing mid-scale academic datasets enables universities to conduct trend analyses on student success factors.

Retail and E-commerce: Retailers and e-commerce platforms generate high-frequency transactional data, customer preferences, and insights into online shopping behavior. NoS-NoB Data is critical in dynamic pricing models, demand forecasting, and targeted marketing strategies. Analyzing these datasets enables businesses to automate inventory management, personalize promotions, and optimize supply chain logistics based on historical sales patterns and market trends.

Financial services: Small and medium-sized banks and financial institutions use NoS-NoB Data for portfolio risk assessment, compliance monitoring, and fraud detection. Transactional data at this scale allows institutions to implement real-time anomaly detection systems, improve loan approval processes, and develop more accurate credit risk models. Additionally, structured and semi-structured datasets help enhance anti-money laundering (AML) frameworks and regulatory reporting efficiency.

Logistics and transport: Transport and logistics companies rely on NoS-NoB Data for route optimization, vehicle tracking, and fleet management. These datasets,

often derived from GPS tracking, supply chain monitoring, and fuel consumption analytics, allow companies to enhance operational efficiency, predict maintenance needs, and reduce delivery delays. Furthermore, mid-scale data processing enables businesses to adjust logistics networks dynamically based on seasonal demand fluctuations.

Local government and public administration: Cities and local governments leverage NoS-NoB Data for urban planning, real-time traffic analysis, and intelligent infrastructure management. Municipalities can optimize public transport networks, improve emergency response coordination, and support sustainable urban development initiatives by integrating data from IoT sensors, public service usage statistics, and environmental monitoring systems. Government institutions use data analytics to improve administrative processes and public service delivery, leveraging structured and semi-structured datasets for better decision making (Desouza & Jacob, 2017).

3.2 Comparative Analysis of Small Data, NoS-NoB Data, and Big Data

To better illustrate the distinctions between Small Data, NoS-NoB Data, and Big Data, Table 1 presents a structured comparison based on key processing characteristics, highlighting their unique attributes and practical implications.

Table 1
Comparison of Small Data, NoS-NoB Data, and Big Data
(Source: own processing)

Aspect	***Small Data***	***NoS-NoB Data***	***Big Data***
Volume	MBs to small GBs	GBs to tens of GBs	Hundreds of GBs to PBs
Diversity	Mostly structured	Structured and semi-structured	Highly diverse, including unstructured data
Processing Speed	Quick processing with simple tools	Requires optimized batch processing	Real-time or near-real-time processing
Purpose	Specific tasks, business operations	Business intelligence, risk assessment	Predictive analytics, deep learning
Data Preparation	Minimal processing, direct analysis	Requires cleaning, transformation, and advanced storage	Extensive pre-processing, distributed systems
Storage & Archiving	Simple file storage, relational databases	Distributed storage, optimized query handling	Cloud computing, big data platforms
Risk of Error	Minimal due to small-scale	Moderate, dependent on processing techniques	High requires sophisticated error-handling mechanisms

3.3 Practical Implications and Data Processing Challenges

Efficient management and utilization of Small, NoS-NoB, and Big Data require tailored approaches to address their distinct processing needs. This section examines the key challenges and practical considerations involved.

Storage and Infrastructure: While Small Data can be stored locally and managed with essential tools, NoS-NoB Data often require distributed databases or optimized storage solutions to ensure efficient retrieval and analysis. In contrast, Big Data relies on cloud-based, high-performance computing systems.

Processing Techniques: Small Data can be analyzed with conventional statistical tools, while NoS-NoB Data requires more advanced data processing techniques, such as data warehousing and machine learning pipelines. Big Data, on the other hand, is processed using large-scale distributed computing frameworks like Hadoop and Spark.

Scalability & Performance: NoS-NoB Data represent a critical transition stage, where traditional small-scale methods become inefficient, but full-scale big data solutions remain unnecessary or cost-prohibitive.

Reproducibility & Data Integrity: Errors in NoS-NoB Data processing may have significant business impacts, necessitating rigorous validation techniques. Big Data introduces even greater complexity, requiring frequent backups and automated error-handling mechanisms.

3.4 Summary and Future Directions

This section provides a structured comparison of Small Data, NoS-NoB Data, and Big Data, offering a clearer perspective on how each category is applied across different industries. By distinguishing the challenges and requirements associated with NoS-NoB Data, this study highlights the need for tailored data processing strategies beyond traditional small-data techniques while avoiding the complexity of full-scale big-data infrastructure.

Future research should focus on:

- *Developing specialized tools tailored for NoS-NoB Data processing, bridging the gap between small and big data methodologies.*
- *Exploring industry-specific best practices to optimize data storage, processing, and analysis for mid-scale datasets.*
- *Investigating the scalability of existing data processing frameworks to support NoS-NoB Data efficiently.*

4 Graphical Representation of Data Categories

Figure 1 visually represents the relationship between small, NoS-NoB, and big data, which are classified based on data volume and structure.

The *x*-axis represents the range of data sizes, spanning from 1 GB to beyond 1 PB. A logarithmic scale is employed to accommodate the vast differences in data magnitude, allowing for a more precise depiction of the gradual transitions between these categories. Using a linear scale would render it difficult to effectively display all three categories within a single graph, as Big Data (typically measured in petabytes or more) vastly exceeds the scale of Small Data (ranging from megabytes to gigabytes).

The *y*-axis represents the distribution of structured, semi-structured, and unstructured data across each category, highlighting how data format complexity evolves with increasing volume.

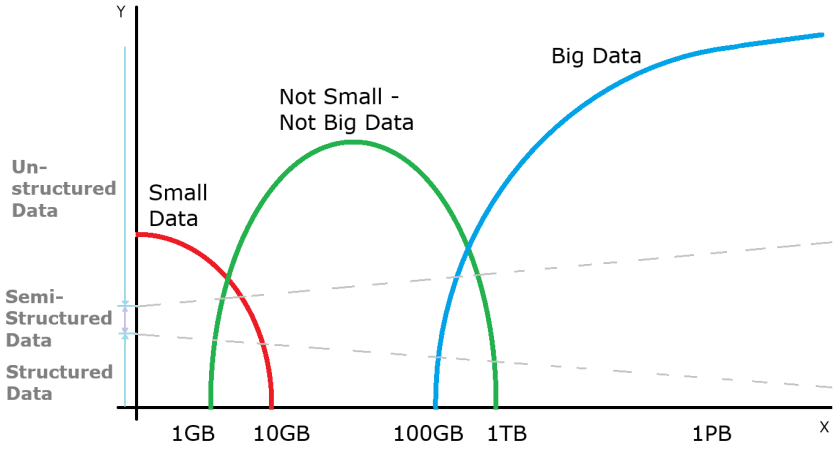

Figure 1
Relationship between data types and their structural characteristics
(Source: own processing)

Interpretation of the Graph

The visualization presents a schematic data classification based on volume (x-axis) and degree of structuring (y-axis). The categorization into "Small Data," "Not Small – Not Big Data," and "Big Data" is intended as an illustrative heuristic rather than a strictly defined or universally accepted taxonomy.

In practical applications, no absolute thresholds definitively delineate "small" from "big" data. Consequently, indicative benchmarks (1 GB, 10 GB, 100 GB, 1 TB, 1 PB) are provided, though these values vary significantly across industries. For instance, in some domains, data volumes of tens of gigabytes may be classified as

"big data," In others, only datasets at the scale of hundreds of terabytes or petabytes warrant such classification.

The vertical axis suggests that smaller datasets exhibit a higher degree of structure, whereas larger datasets are frequently semi-structured or unstructured (e.g., text data, log files, multimedia content). Increasing data volume generally correlates with greater heterogeneity in format and structure. However, this is not an absolute principle—large enterprises may store terabytes of highly structured relational database tables, while a small start-up may primarily generate unstructured log data.

Small Data (Red Curve): The area under the red curve represents "Small Data," predominantly structured and typically ranges from 1 GB to 10 GB. Such datasets can be efficiently processed using standard office tools and are commonly stored on personal devices or within single-server databases.

NoS-NoB Data (Green Curve): The green curve represents datasets that do not fall neatly into the small or big data categories. This category encompasses structured, semi-structured, and unstructured data, with an upper boundary typically around 100 GB. Depending on the context, larger datasets may also be classified as NoS-NoB Data, reflecting the diversity and variability in how organizations define and handle data. This category acknowledges the prevalence of datasets that are too large for traditional office tools, but do not require the complete infrastructure of big data technologies (e.g., Hadoop or Spark).

Big Data (Blue Curve): Big Data is characterized by a predominance of unstructured formats, including video streams, IoT-generated data, and social media logs. These datasets typically exceed 1 TB and necessitate distributed processing frameworks such as Hadoop or Spark to ensure efficient storage and computational performance.

The intersections between Small Data and NoS-NoB Data and between NoS-NoB Data and Big Data are of particular significance. These transitional zones represent areas where datasets exhibit characteristics of both categories. For example, datasets in the 1GB to 10GB range may possess structural complexity or semi-structured formats that exceed the capabilities of traditional tools, necessitating the use of programming languages such as Python or advanced relational databases. A clear example of a dataset that falls into the overlapping region between Small Data and NoS-NoB Data is a 4GB log file, which cannot be fully processed using standard Microsoft Office tools such as Excel yet can be efficiently handled using a simple Python script. This illustrates the necessity of the NoS-NoB Data category. While such datasets do not require full-scale Big Data technologies, they also exceed the processing limitations of conventional small-scale tools.

Similarly, datasets ranging between 100GB and 1TB can push the limits of traditional storage systems, necessitating partial adoption of Big Data frameworks. One illustrative example is the output generated by a brute-force solution to the Traveling Salesman Problem (TSP) for a 14-node, 25-edge graph. The calculation was executed on a laptop with an RTX 3070 graphics card to optimize computation time, leveraging CUDA cores for parallel processing. Despite this optimization, the

computation required 27 hours and generated a 478GB output file. Processing large datasets, particularly when extracting all optimal solutions, necessitates streaming data processing techniques implemented in Python rather than a full Hadoop-based infrastructure, despite the dataset's substantial size. This example highlights the nuanced distinction between NoS-NoB Data and valid Big Data, demonstrating that not all large datasets inherently require Big Data technologies.

These transitional zones highlight the importance of understanding a dataset's specific characteristics when selecting processing and storage technologies. The decision on which technological approach to use ultimately rests with data architects and engineers, who must assess the volume, structure, and processing requirements to ensure optimal performance.

It is important to note that this graph serves primarily a didactic purpose, offering a conceptual framework to differentiate data typologies and their corresponding technological infrastructures. While the overlaps between categories underscore the complexity of data management in real-world applications, the exploration of specific datasets and their classification will be addressed in a future study. A key insight derived from this visualization is that unstructured components typically become more prevalent as data volume increases, potentially exacerbating scalability challenges for conventional data management systems, particularly in the intermediate data range.

Trends in Data Structure

The graph illustrates how the proportion of structured, semi-structured, and unstructured data evolve with increasing data volume:

- *Structured data* predominantly characterizes smaller datasets, as observed in Small Data, which includes tables, relational databases, and well-organized records;
- *Semi-structured data* become increasingly prevalent in the NoS-NoB data range, reflecting the growing complexity of formats such as XML, JSON, and log files;
- *Unstructured data* dominate datasets exceeding 1 TB, a defining characteristic of Big Data, which frequently comprises text, images, audio, and video content.

This visualization (Figure 1) illustrates how the proportional representation of structured, semi-structured, and unstructured data evolve with increasing data volume, highlighting a "middle" zone where traditional tools begin to encounter limitations.

Conclusions

This article introduces the concept of "Not-Small-Not-Big Data" (NoS-NoB Data), a category that bridges the gap between traditional classifications of "Small Data" and "Big Data." NoS-NoB Data is characterized by its size, which exceeds the processing capabilities of standard office tools but does not necessitate the infrastructure required for Big Data technologies. The NoS-NoB Data category encompasses a wide range of formats, including structured, semi-structured, and

unstructured data, each present unique processing challenges for conventional computing systems.

Typically, NoS-NoB Data spans from 1 gigabyte to hundreds of gigabytes. Processing such data efficiently demands specialized tools and methodologies. Advanced database systems, programming languages such as Python or R, and technologies like Hadoop, Spark, and NoSQL databases are necessary for effectively handling and processing these datasets. The article underscores the importance of selecting appropriate tools and methodologies based on the specific needs of each project, which can significantly enhance processing efficiency and improve overall project outcomes.

Moreover, the article highlights a variety of application domains for NoS-NoB Data, including small and medium-sized enterprises (SMEs), healthcare, education, retail, financial services, logistics, and public administration. These findings demonstrate the extensive applicability of NoS-NoB Data in scenarios where traditional tools are insufficient, yet Big Data technologies are not necessary. The versatility of NoS-NoB Data allows for their use across multiple industries, offering practical solutions in many real-world contexts.

A comparative analysis, supported by comprehensive figures and tables, illustrates the distinctions and overlaps between Small Data, NoS-NoB Data, and Big Data. The visualizations provide a detailed view of the evolving distribution of structured, semi-structured, and unstructured data types across varying data volumes. One key insight is that the boundaries between different data size categories are not static; instead, they are dynamic and influenced by the continual advancements in computational technology. As processing power and storage capabilities improve, these boundaries gradually shift upward, which explains the presence of overlapping areas in the graph. These transitional zones reflect the fluid nature of data classification and the need for flexible approaches to managing datasets at the edges of these categories.

While this article presents a wide overview of NoS-NoB Data, it is important to note that this topic remains a dynamic area of research. The findings suggest that precise planning and problem identification are essential for organizations to estimate the expected input and output data volumes. Furthermore, incorporating the principles of the "5Vs" (Volume, Variety, Velocity, Veracity, and Value) into decision making processes can assist organizations in selecting the most suitable technologies, ultimately enhancing operational efficiency and profitability.

Acknowledgment

This work has been supported by the project VEGA 1/0096/23 - Selected methods of risk management in the implementation of partial internal models for determining the solvency capital requirement.

References

[1] Abu-Salih, B., Wongthongtham, P., Zhu, D., Chan, K. Y., & Rudra, A. (2021) *Introduction to Big Data Technology*. arXiv. https://doi.org/10.48550/arXiv.2104.08062

[2] Banafa, A. (2018, August 1) Small data vs. Big Data: Back to the basics. OpenMind. https://www.bbvaopenmind.com/en/technology/digital-world/small-data-vs-big-data-back-to-the-basics/

[3] Chollet, F. (2021) Deep learning with Python. Manning Publications.

[4] Davenport, T. H., & Harris, J. G. (2017) Competing on analytics: The new science of winning. Harvard Business Review Press.

[5] Desouza, K. C., & Jacob, B. (2017) Big data in the public sector: Lessons for practitioners and scholars. *Administration & Society, 49*(7), 1043–1064. https://doi.org/10.1177/0095399714555751

[6] Duarte, F. (2023, December 13) Amount of data created daily (2024). Exploding Topics. https://explodingtopics.com/blog/data-generated-per-day

[7] EDUCBA. (2023, June 15) Small Data vs big data: Top 10 useful comparisons you need to know. https://www.educba.com/small-data-vs-big-data/

[8] Gartner_Inc. (n.d.) Definition of big data - Gartner information technology glossary. Gartner. https://www.gartner.com/en/information-technology/glossary/big-data

[9] Inmon, W. H., Linstedt, D., & Levins, M. (2019) Parallel processing. Data Architecture, 81–87. https://doi.org/10.1016/b978-0-12-816916-2.00012-7

[10] Mishra, D. (2019, June 24) What is Big Data, and why do we need Hadoop for Big Data? SAP Blogs. https://blogs.sap.com/2019/06/24/what-is-big-data-and-why-do-we-need-hadoop-for-big-data/

[11] Misut, M., & Jurik, P. (2021) Datafication as a necessary step in the processing of Big Data in decision-making tasks of business. *International Conference on Innovations in Science and Education (Natural Sciences and ICT), Prague, Czech Republic*

[12] Serrant, D. (2023, August 19) How can one tell if their dataset is small, medium, or big? Quora. https://www.quora.com/profile/Daryle-Serrant

[13] Taylor, P. (2023, September 14) *Most popular database management systems 2023*. Statista. https://www.statista.com/statistics/809750/worldwide-popularity-ranking-database-management-systems/

[14] Tawil, A.-R., Mohamed, M., Schmoor, X., Vlachos, K., & Haidar, D. (2023) *Trends and challenges towards an effective data-driven decision making in UK SMEs: Case studies and lessons learnt from the analysis of 85 SMEs*. arXiv preprint arXiv:2305.15454. https://arxiv.org/abs/2305.15454

[15] Walker, B. (2017, November 23) *Big Data Strategies Disappoint with an 85 percent failure rate*. Digital Journal. https://www.digitaljournal.com/tech-science/big-data-strategies-disappoint-with-85-percent-failure-rate/article/508325

[16] Zhang, Y., Li, H., Liu, X., Chen, H., & Chang, V. (2019) A survey of Big Data architectures and machine learning algorithms in healthcare. *Journal of King Saud University - Computer and Information Sciences, 31*(4), 552-567. https://doi.org/10.1016/j.jksuci.2018.09.016