Review

Smart Supervision of Public Expenditure: A Review on Data Capture, Storage, Processing, and Interoperability with a Case Study from Colombia

Jaime A. Restrepo-Carmona [1] , **Juan C. Zuluaga** [2] , **Manuela Velásquez** [3] , **Carolina Zuluaga** [3] , **Rosse M. Villamil** [2] , **Olguer Morales** [2] , **Ángela M. Hurtado** [2] , **Carlos A. Escobar** [1] , **Julián Sierra-Pérez** [1,3] **and Rafael E. Vásquez** [1,3,*]

1 Corporación Rotorr, Universidad Nacional de Colombia, Cr. 45 26-85, Bogotá 111311, Colombia; jarestrepoca@unal.edu.co (J.A.R.-C.); pmo@rotorr.co (C.A.E.); julian.sierra@upb.edu.co (J.S.-P.)
2 Dirección de Información, Análisis y Reacción Inmediata, Contraloría General de la República, Cr. 69 No 44-35, Bogotá 111071, Colombia; juan.zuluaga@contraloria.gov.co (J.C.Z.); rosse.villamil@contraloria.gov.co (R.M.V.); olguer.morales@contraloria.gov.co (O.M.); angela.hurtado@contraloria.gov.co (Á.M.H.)
3 School of Engineering, Universidad Pontificia Bolivariana, Medellín 050031, Colombia; manuela.velasquezj@upb.edu.co (M.V.); carolina.zuluagaa@upb.edu.co (C.Z.)
* Correspondence: rafael.vasquez@upb.edu.co; Tel.: +57-4-448-8388

Abstract: Effective fiscal control and monitoring of public management are critical for preventing and mitigating corruption, which in turn, enhances government performance and benefits citizens. Given the vast amounts of data involved in government operations, applying advanced data analysis methods is essential for strengthening fiscal oversight. This paper explores data management strategies aimed at enhancing fiscal control, beginning with a bibliometric study to underscore the relevance of this research. The study reviews existing data capture techniques that facilitate fiscal oversight, addresses the challenges of data storage in terms of its nature and the potential for contributing to this goal, and discusses data processing methods that yield actionable insights for analysis and decision-making. Additionally, the paper deals with data interoperability, emphasizing the importance of these practices in ensuring accurate and reliable analysis, especially given the diversity and volume of data within government operations. Data visualization is highlighted as a crucial component, enabling the detection of anomalies and promoting informed decision-making through clear and effective visual representations. The research concludes with a case study on the modernization of fiscal control in Colombia, focusing on the identification of user requirements for various data-related processes. This study provides valuable insights for modern audit and fiscal control entities, emphasizing that data capture, storage, processing, interoperability, and visualization are integral to the effective supervision of public expenditure. By ensuring that public funds are managed with transparency, accountability, and efficiency, the research advances the literature by addressing both the technological aspects of data management and the essential process improvements and human factors required for successful implementation.

Keywords: digital transformation; smart government; public expenditure; fiscal control; governmental accounting; Industry 5.0

1. Introduction

Today, the use and management of data in various fields has increased significantly, facilitating business and industry analysis for decision making [1]. The amount of data generated is growing exponentially [2]. It is estimated that around 2.5 quintillion bytes of data are created every day [3]. In other words, the data are fully available to those willing to mine them. Data are a valuable asset for businesses in the 21st century [4]. Clive Humby, a British mathematician, stated that 'data is the new oil' in 2006. Like oil, data are not

valuable in their raw state; their value comes when they are collected quickly, completely and accurately, and linked to other relevant data [5].

The public sector and tax analytics are also seizing the opportunity to use data to combat tax evasion [6]; public resources management is a critical government function that helps ensure fiscal responsibility and accountability. The public budget, usually set for one year, is a fundamental plan of action that reflects the government's priorities and objectives through the amounts allocated for revenue and expenditure [7]. Public expenditure management ensures that resources are used efficiently and effectively and supports the delivery of essential public services. It is, therefore, a key element of good governance.

In addition, the management of public expenditures is essential to achieve macroeconomic stability and can help control inflation and reduce deficits [8]. It also promotes transparency and accountability of public finances [9,10]. By providing a clear and understandable picture of spending, citizens can better assess economic factors and their impact on society. This allows them to support or challenge spending decisions in a more informed way, while governments can demonstrate progress against their commitments [11].

Today, problems such as corruption, lack of transparency and inefficiency are among the main difficulties facing the public sector [12]. These problems impede the proper functioning and management of government, which has a negative impact on the development of a nation and the well-being of its citizens. The institutional design of local governments has a significant impact on poverty: the increased risk of corruption and inefficiency in the use of transfers for education and health spending increase poverty. On the contrary, transparent governance can reduce it; in fact, a one percentage point increase in transparency and relative efficiency indices reduces the poverty rate by 0.6 percentage points [13].

The risk of corruption tends to increase with the size of the local state and decrease with improvements in fiscal performance, tax collection, and average years of education [14]. It is essential to implement strategies to mitigate this situation, one of the most promising is the use of analytics and data management in the public sector [15]. Technology in public administration (e-government) has enabled the delivery of public services over the Internet, promoting efficiency in data collection, processing, and reporting, and improving decision-making [16].

Advances in smart technologies, better informed and connected citizens, and globally interconnected economies have created new opportunities [17]. The concept of e-government is being enhanced by governments by recognizing the power of data and heuristic processing through artificial intelligence (AI) to improve services, interact with citizens and society, propose new policies, and implement solutions for the well-being of the community, therefore, becoming a smart government [18].

However, the diverse sources and formats of public sector documents make the collection, processing and organization of these documents challenging from a data analytics perspective [19]. It is important to develop approaches to managing these data, as their analysis allows for greater citizen involvement by giving them more access to public decisions and spending [20], increasing transparency in the public sector, and giving citizens a greater sense of accountability by providing different views on government performance in meeting its public policy objectives. As an example, Valle-Cruz et al. [16] presented the use of AI as a promising tool for intelligent monitoring of the public sector and combating tax evasion. The availability of this information is argued to help stakeholders make better-informed decisions. The use of AI would work like any other decision support system, providing several alternatives and a wealth of information to enable the final decision to be made.

Some of the advantages of AI-based techniques relate to their ability to analyze any data, regardless of organization, size, or format [21]. However, some limitations of AI-based methods are related to the computational capacity available at the time. For example, Yahyaoui and Tkiouat [22] showed how reinforcement learning based on Markov models and partially observable computations of the behavior of a contributing agent can refine the analysis of an audit policy. They showed that by synchronizing procedures and dynamically

updating intelligent behaviors, a hybrid model can be built that combines bottom-up agent-based execution with various partially observable intelligent behaviors. This can serve as a platform for testing the effectiveness of an audit policy and for training an intelligent auditor. Rukanova et al. [23] developed a framework for the value of data analysis in government oversight that is intended to serve as a tool for analysis and understanding. In their study, they found that collective capacity building processes in data analytics and their link to capacity processes in an individual organization are very interesting but not yet well understood.

This paper contributes to the field of information by describing how effective data management and processing can significantly enhance the quality of information used for financial control and auditing practices, in the digital era where several industries use digital technologies to support their operations [24]. To do so, we defined the following research questions (RQ): RQ1. What are the research trends regarding data capture, storage, processing, interoperability, and visualization in the context of fiscal control? RQ2. Which countries have contributed the most to these topics in the last years? RQ3. What groups of co-authors are more representative of collaborative research on these topics? The study emphasizes the critical role of accurate data collection and secure storage as foundational elements for reliable fiscal oversight. Provides information on advanced data processing techniques that enable proactive decision-making in public expenditure management. In addition, the paper underscores the importance of interoperable systems that facilitate seamless data exchange between government entities, promoting efficiency and transparency. Finally, the article advocates for the adoption of appropriate data visualization strategies to improve communication of financial insights and support evidence-based policy-making. The paper contributes by providing a stakeholder-driven framework for designing data management architectures that enhance real-time supervision and interoperability in public expenditure oversight.

The organization of the paper is as follows. Section 2 contains the bibliometric analysis on the role of data collection, storage, processing, interoperability, and visualization in improving the smart monitoring of public expenditure. Section 3 provides an overview of data capture techniques, data capture tools, data capture types, and a data capture architecture proposed for the CGR. Section 4 describes the data storage options, on-premises and in the cloud, and the data storage architecture proposed for the CGR. Section 5 contains the three stages in data processing: data cleaning, data transformation, data analysis, and a data processing architecture proposed for the CGR. Section 6 refers to the architecture that facilitates effective interoperability. Section 7 provides an overview of the basic principles and advanced technologies related to data visualization in the modern era. Section 8 contains a case report for fiscal control in Colombia and provides user requirement lists, generated from workshops that involved multiple stakeholders. Then, Section 9 contains the discussion, and finally, conclusions are presented in Section 10.

2. Literature Review

The study employed a systematic literature review methodology comprising three main stages: defining the research question, searching for relevant studies, and selecting them. The first stage involved defining the research question to delineate the review's aim and scope, guiding the subsequent search and selection of pertinent studies. The literature review focused on analyzing the impact and role of data collection, storage, processing, interoperability, and visualization in enhancing the smart monitoring of public expenditure.

This review involved a comprehensive search for articles using the following predefined search equation: (("DATA CAPTURE" OR "DATA STORAGE" OR "DATA PROCESSING" OR "DATA VISUALIZATION") AND "PUBLIC EXPENDITURE" AND "SUPERVISION"). This equation was used in Scopus, a prominent engineering database. This process yielded 547 papers, that covered the period from 2014 to 2024, ensuring a ten-year search window. Figure 1 illustrates the publication trends over a 10-year search period. It is evident that the number of publications focusing on data management and fiscal control

techniques has increased over time, especially since 2019, indicating a trend toward digital strategies to enhance control and reduce fraud. Hence, to focus the scope and select the most recent literature, the review included articles from journals and conference papers that have been published in English after 2019; 403 documents met these criteria for detailed analysis, Figure 2.
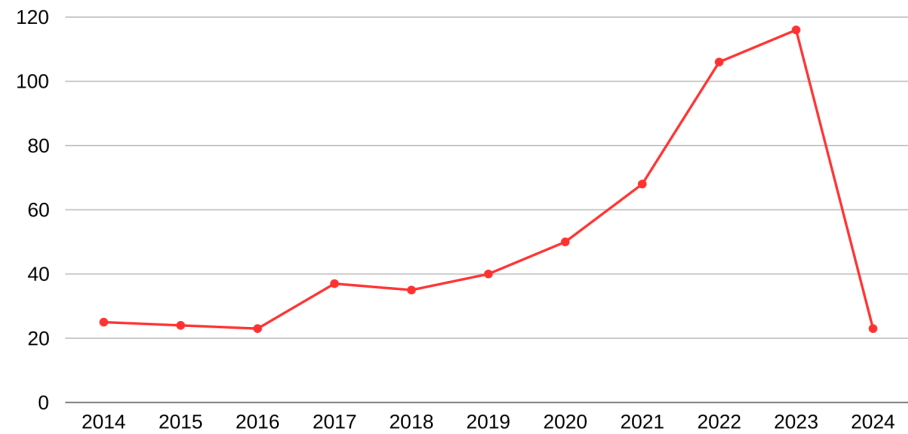

Figure 1. Publications over a 10-year search window.
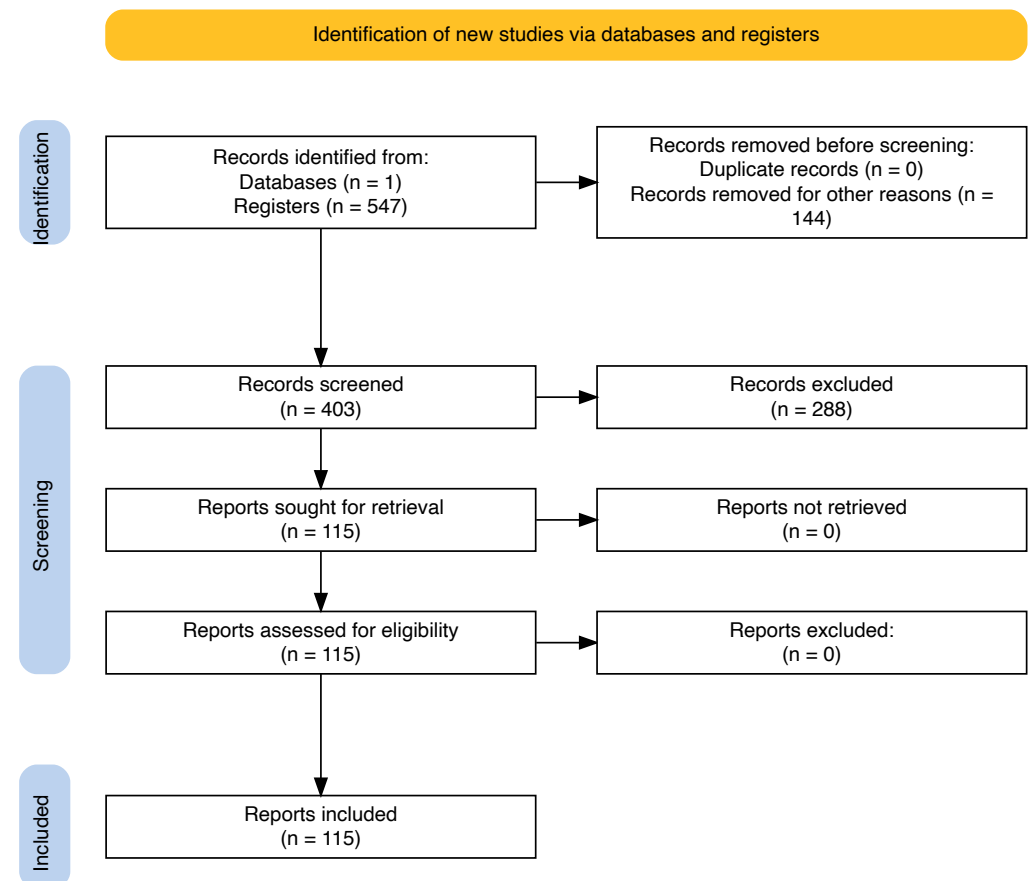
Figure 2. Methodology for the systematic literature review, using the PRISMA framework [25].

The selection of articles was based on their relevance to the research question and the number of citations. Priority was given to the most recent articles that directly addressed the research question. However, older articles with more than five citations were also included. Following this process, 115 articles were chosen for the study, as illustrated in Figure 2.

Bibliometric Analysis

The bibliometric analysis enables the extraction of pertinent information regarding trends and quantitative research data using mathematical tools. For metadata analysis, the Scopus tool was employed to visualize publication behaviors, while VOSviewer (version 1.6.20) was used to analyze keywords, co-authors, and their relationships. The selected tool is noted for its user-friendly interface in developing thematic maps and conducting analyses such as clustering, keywords, and sources, among others. All findings discussed in this section are derived from primary literature sources.

The global expansion of data handling and management is impacting not only sectors where data analysis is critical, such as financial analysis, power generation, and industrial plants, but also those that historically managed data differently. With the emergence of machine learning (ML) and related tools, the tax sector has significant opportunities to continually monitor data and mitigate tax fraud. Through the collection, processing, manipulation, and visualization of data, trends, patterns, and outliers indicative of potential fraud can be identified, enabling more effective actions to be taken.

Figure 3 presents a collection of common keywords extracted from the primary bibliography used in the bibliometric analysis conducted with VOSViewer. Three primary clusters were identified, interconnected in a way that reflects thematic coherence within the database. Prominent keywords are represented by larger circles, highlighting terms such as 'deep learning', 'machine learning', 'fraud', 'public sector', 'information processing', 'crime', 'human', and 'machine learning'. Notably, the term 'crime' serves as a bridge between the 'human and information processing' cluster and the 'public sector and machine learning' clusters. This connection underscores the growing importance of data management and analytics in the public sector for proactively addressing and mitigating fraud, reflecting key trends in current research, answering RQ1.

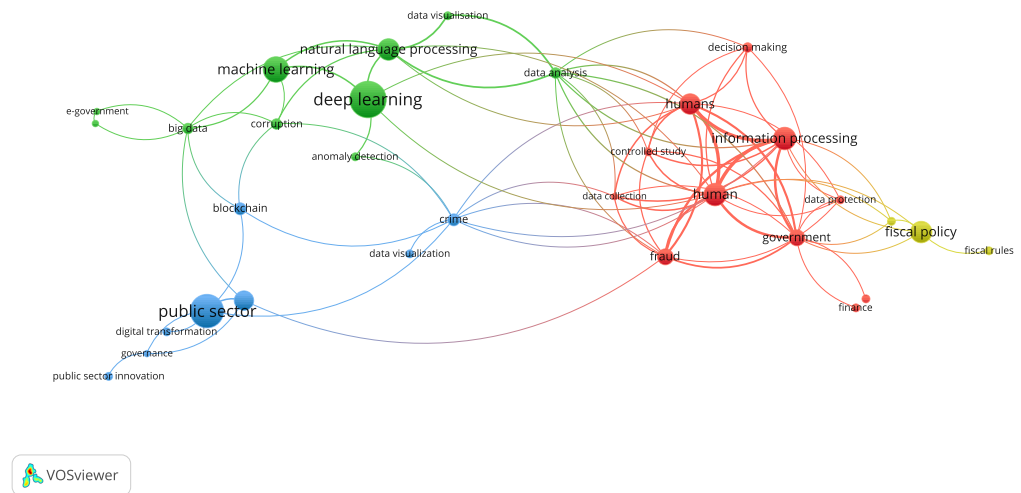**Figure 3.** Co-occurrence network of keywords. Performed with VOSviewer 1.6.20.

Figure 4 highlights the countries with the highest number of publications on these topics, with the United States leading, followed by China and the United Kingdom, answering RQ2. This suggests a stronger research focus on data management and the prevention, detection, and prosecution of corruption and fraud in these countries.
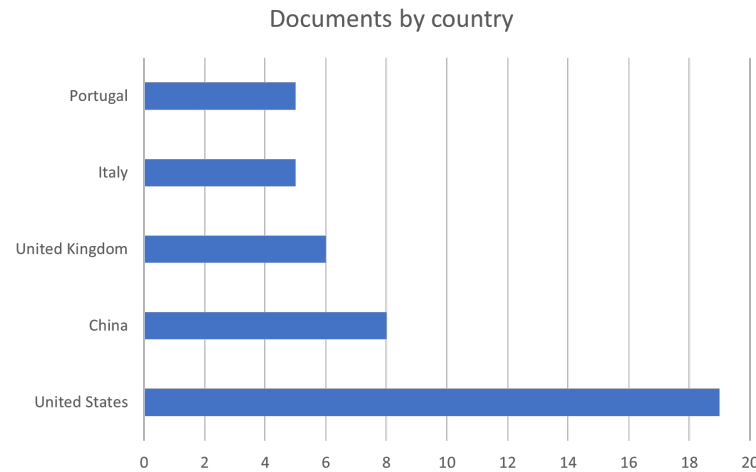
Figure 4. Documents by country.

Finally, Figure 5 illustrates a co-authorship network that shows collaborations among researchers, with distinct clusters representing different groups, answering RQ3. The nodes vary in size based on publication activity, and the colors reflect the timeline from 2021 to 2024. Most collaborations occurred between 2021 and 2023, with a few groups, such as those involving "sun, yuan" and "basile, pierpaolo" showing more recent activity in 2024. This pattern highlights an increasing momentum in the field, with both ongoing and newly emerging collaborations.
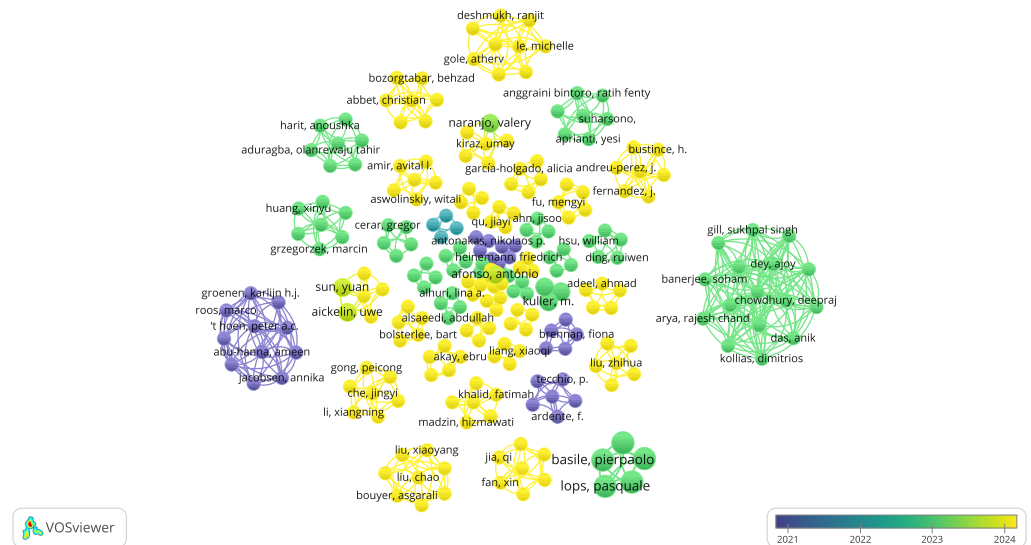
Figure 5. Co-author network and publication years. Performed with VOSviewer 1.6.20.

3. Data Capture

This section provides an overview of data capture techniques, tools, types, and hardware architectures used for data collection in organizations. Historically, data extraction was performed manually, which, while offering control and flexibility, was prone to errors, lacked scalability, and often led to inconsistencies. Today, various technologies enable data extraction with differing levels of accuracy.

The primary goal of data capture is to facilitate processing, integration with other datasets, and transfer to databases, helping organizations identify valuable information for decision-making. However, challenges such as standardizing and formatting unstructured data arise [26]. Furthermore, data from diverse sources often require an ETL (Extract, Transform, Load) process to address compatibility issues [27,28].

3.1. Types of Data for Public Expenditure Monitoring

For effective monitoring of public spending, it is crucial to have access to specific data, which can often be challenging due to the dispersed nature of the data. Most of these data can be found in different forms, listed as follows.

- Public budgets. Financial documents detail government revenues and expenditures for a given period, including allocations for different departments, programs, and projects. They are crucial for evaluating resource allocation efficiency and ensuring funds align with government priorities [29].
- Government contracts. These legal agreements between the government and suppliers outline conditions, prices, delivery terms, and penalties for non-compliance. They are essential for ensuring transparency and integrity in procurement by verifying fair contract awards and compliance, and helping to prevent corruption and misappropriation of funds [30].
- Public bidding. Competitive processes where the government solicits and evaluates bids from suppliers for goods or services. This includes solicitations, received bids, evaluation criteria, and award results. Public bidding promotes competition and transparency, ensuring the government achieves the best value for money [31].
- Public payments and disbursements. These records encompass all financial transactions made by the government, including payments to suppliers, employee salaries, transfers to other entities, and grants. They are essential for monitoring public fund flow, ensuring payments align with approved budgets, and detecting irregularities or fraud [32].
- Audits and public expenditure control. Independent evaluations by control agencies, such as comptrollers and external auditors, assess the legality, efficiency, and effectiveness of public resource use. These evaluations offer an objective review of the government's financial performance, identify areas for improvement, and help prevent fraud and corruption [33].
- Performance indicators. These quantitative and qualitative metrics assess the performance of programs and projects funded by public resources, including indicators of efficiency, impact, and beneficiary satisfaction. They help evaluate whether public resources achieve expected outcomes and enable policy and program adjustments to enhance effectiveness [34].
- Beneficiary data. Information on individuals, communities, or organizations benefiting from publicly funded programs includes demographic, socioeconomic, and geographic data. These data are crucial for assessing equity in public resource distribution, ensuring support reaches those most in need [35].
- Transparency and accountability data. These include public reports, performance indicators, and independent evaluations that enable citizens and oversight bodies to monitor the use of public resources. They promotes citizen participation and accountability by providing accessible information on public fund utilization [36].

3.2. Data Capture Techniques

To support the new fiscal control model, advanced data capture techniques—such as full, incremental, and wrap extraction—are essential for efficiently gathering information from various sources [37].

Full extraction is an ETL technique that transfers the entire dataset from the source system to the target database, ensuring complete data integrity. However, it can be resource-intensive and time-consuming, potentially impacting system performance, especially with large datasets.

Incremental extraction extracts only the data that have changed since the last extraction, usually triggered by events like successful prior extractions or updates. This technique minimizes the volume of data transferred, reduces resource consumption, and accelerates the process, making it particularly efficient for large datasets by ensuring that only new or modified data are processed.

Wrap extraction is designed to manage and optimize the extraction of large data volumes. It involves breaking down data into smaller parts or batches to enhance efficiency and prevent overloading network or system resources. This method allows for effective handling of substantial datasets while maintaining performance and minimizing the risk of errors during data transfer.

3.3. Types of Data Capture

Regarding types of data capture, several methods can be used in fiscal control, including batch data ingestion, data extraction from documents, web data extraction, data extraction from enterprise applications, extraction from logs and records, streaming data extraction, IoT data extraction, social network data extraction, real-time data extraction, and email data extraction.

3.3.1. Batch Data Ingestion

Batch data ingestion captures datasets from various sources at regular intervals, making it ideal for handling large volumes of data that do not require real-time processing. In this process, data can be grouped and sent to the system based on predefined criteria. One advantage of batch data ingestion is its cost-effectiveness, as it helps prevent network overload by allowing continuous data transmission in controlled intervals [38].

3.3.2. Data Extraction from Documents

Data extraction from documents involves retrieving information from printed or digital documents. The goal is to identify and extract specific unstructured data, transforming it into structured data for storage and analysis. Automated document data extraction employs a combination of techniques, tools, and algorithms to extract necessary data from complex documents. The key steps in this process include [39]:

- Scanning high-quality documents or images improves OCR (Optical Character Recognition) output and extraction accuracy.
- ML models should be updated and trained frequently using diverse and representative datasets, adapting them to new document layouts and formats, which improves extraction performance over time.
- Use a hybrid approach using rules for fields in structured data with predictable patterns in conjunction with ML algorithms to handle unstructured or complex data.
- Implement robust data validation to ensure data accuracy and integrity.
- Design the data extraction process to handle large volumes of documents without failure.

The automated document data extraction process enables organizations to efficiently process and retrieve data from multiple document types with minimal effort. It offers numerous advantages, including improved process efficiency, accuracy, accessibility, scalability, flexibility, and adaptability.

3.3.3. Web Data Extraction

Web data extraction, also known as web harvesting or web scraping, involves collecting data from websites using software that simulates HTTP or web browser interactions [40]. Search engines widely use web scraping to convert unstructured web data into structured data for storage and analysis in databases [40]. Often confused with web scraping, web crawling focuses on locating and storing various data from the internet, organizing it into a database based on relevant words or search keywords. Each keyword is associated with a hyperlink identifier, allowing traceability to its source [41]. This technique employs data mining to analyze multiple statistical properties of the collected information, facilitating data monitoring services that generate user alerts. Essentially, web data extraction aims to obtain new or updated information [42].

3.3.4. Data Extraction from Enterprise Applications

The use of data mining and AI has become essential for process analysis and generating smart recommendations within organizations. These tools enable accurate information extraction from processes and facilitate the collection and evaluation of behavioral patterns and preferences among specific groups [43]. In business applications, data mining involves searching, collecting, and analyzing data to uncover specific attributes, outcomes, or hidden patterns. For instance, customer information is collected through data mining, transforming raw data into actionable insights for both the organization and its customers [44].

3.3.5. Data Extraction from Logs and Records

Log files are plain text files typically analyzed during emergencies to investigate program or system errors, offering insights into system and user behaviors. Log data extraction tools allow an organization's IT teams to access this information from a centralized location, eliminating the need to manage multiple software tools and simplifying troubleshooting for swift issue detection and resolution. These tools help determine what data to record, the format for storage, the retention period for data, and the optimal deletion methods when data are no longer needed [45].

3.3.6. Data Extraction from Streaming Sources

Data extraction from streaming sources involves tools that transmit data to generate basic reports and perform simple actions, such as issuing event alerts. These tools utilize complex algorithms for data extraction and stream processing, focusing on extracting data from smaller time intervals [46]. Key features of streaming data extraction include continuous querying and processing of data, logging of individual data or micro-batches with a few records, low transmission latency in seconds or milliseconds, straightforward response functions, and incremental metrics.

3.3.7. IoT Data Extraction

The Internet of Things (IoT) has transformed how organizations interact with technology, necessitating efficient tools to manage various sources and types of data for automated collection from different IoT devices [47]. Since IoT data are generated and transmitted in real-time, there is a critical need for tools that can automatically ingest, process, integrate, and transmit these data.

3.3.8. Data Extraction from Social Networks

Tools like web scraping and data mining are used to extract data from social networks based on predefined criteria [48]. Web scraping employs bots to access social media platforms and extract information via an unblocking API acting as a proxy server. Captured data may include profile information, job details, URLs, images, videos, hashtags, geolocation, timestamps, comments, marketplace posts, engagement rates, and emerging trends. Data mining utilizes statistical, mathematical, and machine learning methods to analyze the extracted information. Following extraction, techniques such as classification, association, pattern tracking, predictive analysis, keyword extraction, sentiment analysis, and trend analysis are applied to derive insights.

3.3.9. Real-Time Data Extraction

Real-time data are generated by various technologies, including IoT devices, which must connect to multiple data sources for effective data capture [49]. Real-time data capture tools have three primary characteristics: a single data source, a single channel, and minimal resource usage. Additionally, these tools must provide:

- High-speed response to a large number of events
- Management of large volumes of data
- Data capture from multiple sources and formats
- Data filtering and aggregation

- Management of a constant, asynchronous, persistent, and connection loss-resistant data transmission channel

3.3.10. Data Extraction from Emails

Unstructured data from the header, body, and attachments of emails are extracted and stored in appropriate databases [50]. Structured data are extracted using bots that validate the data's characteristics. Additionally, some tools analyze email activity to extract metrics such as the number of emails sent per day, emails opened, top senders and recipients, and average response time, generating statistics to measure productivity.

3.4. Challenges in Data Capture

Data capture is a critical process that can face various challenges to ensure the reliability of the data extracted from the source, as follows:

- Data Quality: Refers to the accuracy, consistency, and timeliness of captured data. Manual data entry can introduce errors that affect information accuracy and consistency, and data may be incompletely captured due to a lack of standardized systems or clear procedures [51].
- Data Integrity: Refers to the accuracy and reliability of data, ensuring it remains unaltered and trustworthy. Data can be intentionally manipulated to conceal fraud or errors, and capture processes can be vulnerable to cyber-attacks, compromising information integrity [52].
- Data Accessibility: Refers to the ease with which data can be retrieved and used by authorized stakeholders. Restricted access policies may limit data availability to authorized users [53].
- Data Security and Privacy: Crucial for protecting sensitive information and ensuring public confidence in the governance system [54].

4. Data Storage

To store large volumes of data, organizations can choose between on-premise and cloud storage. Each option has its pros and cons based on the organization's needs, but cloud storage is increasingly favored for its flexibility and scalability [55]. A summary of the advantages and disadvantages of both storage types is provided in Table 1.

The selection of a storage solution can be guided by the following aspects [56]:

- Redundancy: Having backups of data is essential to prevent loss from failures. Data should be copied and stored in different systems to ensure its persistence.
- Persistence and Preservation: Data can become inaccessible over time, so regular reviews are needed to identify and address any damage.
- Transformation: As storage technologies evolve, it is crucial to migrate data to new systems without altering its semantic content.

Table 1. Advantages and disadvantages of on-premise and cloud storage.

Storage Type	Advantages	Disadvantages	Ref.
On-premise	Full control over the hardware and the actions to be performed with the software. No access restrictions. They are reliable as they do not depend on internet connections.	Dependence on providers for hardware and software purchases. Security risk due to internet connection. Low flexibility and scalability. High cost of installation and maintenance.	[57]

Table 1. *Cont.*

Storage Type	Advantages	Disadvantages	Ref.
Cloud	Fast Implementation in the Start-up. Accelerated time to productive use of applications. Lower initial and operating costs. No need for additional infrastructure for servers, networking, among others. No additional IT resources are needed to support both your infrastructure and your applications. Cloud providers offer a best-class enterprise infrastructure with the appropriate servers, networks, and storage systems and are responsible for frequent upgrades of your application with each new release; regular backups, and necessary restores, and compliance with the latest security and legal compliance requirements. Accessibility from anywhere in the world Backup, which avoids data loss.	Requires internet access and a good, stable, and fast connection. Risk in the security of the data, since they can be hacked. Privacy can also be lost. Dependence on third-party hardware and software to operate.	[58–61]

4.1. On-Premise Storage

The on-premise solution is the name given to the storage systems installed in the organization that will use them. They contain all the servers and software necessary to provide the different services. The size and economic activity of the organization determine the number of servers and the dimensions of the installations [57]. Studies have shown the drawbacks of storing data at the desktop level. Problems increase with the increased cost of hardware and maintenance that involve high labor costs. In addition, there are increased possibilities of data loss due to the low robustness of these systems [59]. However, because of the characteristics mentioned above, this type of storage is no longer being used as much. Gradually, it is migrating to cloud storage or mixed versions.

4.2. Cloud Storage

Cloud storage systems consist of a network of connected devices that facilitate storage virtualization, abstracting physical storage from applications. This network enables access to stored information regardless of location or mode. It can be categorized into file, block, and object storage based on client interaction and access methods [62,63].

4.2.1. File Storage

Data are organized hierarchically in files, with the storage system maintaining all file information as metadata accessible by specifying the path to a file [63]. Some storage systems include the Hadoop Distributed File System (HDFS) and Google File System (GFS).

- HDFS [64] efficiently stores large volumes of unstructured data on commodity hardware, enabling rapid data ingestion and bulk processing [65]. It segments large files into 128 MB blocks, replicating them across nodes with a replication factor of 3. A standard configuration features an active name node and multiple data nodes, adhering to a master-slave architecture [66].
- The GFS is tailored for organizing and managing large files according to user needs, emphasizing scalability, reliability, and availability [67]. It consists of three primary components: Masters, which handle file system metadata; Chunkservers, which manage storage units and Clients, which engage with the master for metadata tasks and directly communicate with chunkservers for data operations [68].

4.2.2. Block Storage

In block storage, files are divided into blocks, each assigned an address for easy access and combination by applications, resulting in good performance. However, this setup does

not guarantee secure data transmissions [62]. The storage remains raw, with data organized as an array of unrelated blocks [69].

4.2.3. Object Storage

Object storage encapsulates data, attributes, metadata, and object identifiers within virtual containers [70]. Objects can be of any type and are geographically distributed, enabling direct and secure data access through metadata. This architecture offers excellent scalability, ideal for supporting Big Data applications [62].

4.3. Database for Storing Government Expenditure Data

Storing public expenditure data efficiently, securely, and accessibly is crucial. Below are several database options for managing and storing such data.

- Relational Databases (SQL): relational databases are database management systems that use a table-based model to organize data. They are ideal for structured data and complex transactions. Table 2 shows some examples of relational databases.
- NoSQL Databases: NoSQL databases are database management systems designed to handle unstructured or semi-structured data, with a focus on scalability and flexibility. Table 3 shows some examples of NoSQL databases.
- Data Warehouses: data warehouses are storage systems specifically designed for queries and analysis of large volumes of data [71].

Table 2. Relational Databases.

Name	Advantages	Application	Ref.
ySQL	Ease of use. Reliability. Scalability. Performance. High availability	Web applications and medium to large data management systems	[72,73]
PostgreSQL	Open-source. Highly extensible. Robust support for complex transactions and queries.	Applications requiring a high level of integrity and advanced data operations.	[74,75]
Microsoft SQL Server	Integration with other Microsoft products. Enterprise support. Advanced analysis and reporting tools.	Common in enterprise environments using Microsoft products	[76]

Table 3. NoSQL databases.

Name	Advantages	Application	Ref.
MongoDB	Open-source, flexible, dynamic scheme, high scalability	Applications that handle large volumes of unstructured data.	[77]
Cassandra	Open-source, horizontal scalability, high availability, fault tolerance, performance	Distributed applications requiring high availability and management of large volumes of data.	[75,78]

Ensuring data security is crucial for the effective and responsible management of public spending. In this regard, robust security measures are required to protect sensitive information. Encryption practices, stringent access controls, continuous monitoring, regular backups, timely updates, and comprehensive staff training are essential steps for government entities to mitigate risks and protect their data from potential threats.

5. Data Processing

Data processing involves collecting, cleaning, transforming, and analyzing valuable information from diverse sources. Today, data processing has evolved into a specialized discipline, particularly in managing large-scale information. The process typically includes three stages: data cleaning, data transformation, and data analysis [79].

5.1. Data Cleansing

Data cleansing, or preprocessing, is vital for producing high-quality datasets that support accurate and meaningful analysis [80]. It reduces the risk of extracting less valuable or erroneous information during subsequent analysis processes [80,81]. Key techniques in data cleansing include the following:

- Integration: Also known as data fusion, this process consolidates data from various sources into a unified, coherent structure.
- Selection: Also referred to as filtering, it focuses on retaining data relevant to the specific analysis or application.
- Reduction: Involves reducing dimensionality or data quantity while maintaining its core characteristics.
- Conversion: Techniques like normalization and discretization transform data to enhance analysis efficiency or result interpretability.
- Imputation: Addresses missing data through methods such as averaging, regression, Bayesian inference, or decision trees to estimate values.
- Outlier Cleaning: Identifies and handles outliers, correcting them if possible or treating them as missing values to ensure data integrity.

Each of these techniques plays a vital role in preparing data for analysis, ensuring that the dataset is accurate, complete, and suitable for deriving meaningful insights.

5.2. Data Transformation

Data transformation involves converting data into a specific format to meet your requirements, enabling efficient management. This is achieved through applying rules or merging data. The process includes various operations, such as moving, splitting, translating, merging, sorting, and classifying data [82].

5.3. Data Analysis

At this stage, data are transformed into useful information that supports decision-making. The data collected are explored to identify trends and patterns that offer valuable insights. Several recognized techniques can be applied [80]:

- Descriptive statistical analysis: These involve estimates or values derived from a data sample to describe its key characteristics, such as mean, median, or standard deviation.
- Supervised models: These models estimate a function or relationship from training data to predict outcomes for new data. Training sets contain input-output pairs, with the outputs being numerical values (regression) or class labels (classification).
- Unsupervised models: These models learn patterns from data without predefined outcomes. Clustering is a common unsupervised technique, which groups similar data points without prior knowledge.

Processing large volumes of information demands dedicated computing resources, influenced by factors such as CPU speed, network capabilities, and storage capacity. Traditional processing systems are outpaced when faced with substantial data loads. Hence, cloud computing has emerged as an excellent solution to overcome these limitations [83]. Taking into account the information above, Table 4 shows some of the most widely used tools and algorithms in the cloud for data cleansing, Table 5 for data transformation, and Table 6 for data analysis.

Table 4. Tools and algorithms for data cleansing or preprocesing.

Name	Description	Ref.
Apache Spark	Open-source software that enables efficient large-scale data cleansing and transformation, including modules for data management such as: Spark SQL and Data Frames	[84]
Databricks	Cloud-based platform, widely used for large-scale data cleansing and preparation tasks	[85]
Talend	Open-source software that offers a wide range of data cleansing functions such as: validate, standardize, data deduplication, and delimited file cleansing	[86]

Table 5. Tools and algorithms for data transformation.

Name	Description	Ref.
Apache Hadoop	Open-source framework for distributed storage and processing. It has three main elements: Hadoop Distributed File System (HDFS), Hadoop MapReduce, and Hadoop Common System. Through the first two components, it facilitates data transformation. The framework is scalable and fault-tolerant.	[87,88]
Apache Hive	It operates as a data storage application, presenting information in tabular format. Tailored for low-latency operations, it facilitates script creation using the Hive Query Language (HQL). Recognized for its scalability and extensibility, this system is specifically crafted for Online Analytical Processing (OLAP) operations.	[89]
Apache Spark	Open source framework widely used in industries for data processing and transformation. It is recognized because data processing is conducted in memory. Therefore, it supports interactive queries and online streaming.	[89]

Table 6. Tools and algorithms for data analysis.

Name	Description	Ref.
MapReduce	It is a distributed programming model that enables large-scale parallel data processing in a timely, faultless, and scalable manner. It has two stages: a mapping stage in which data are taken and key-value pairs are created and a reduce stage that seeks to reduce the data output by joining data according to its key and value.	[87,90,91]
Machine learning algorithms	They are a set of techniques and methods used to learn patterns and perform specific tasks without explicit human intervention. Naive Bayes: estimate the probability that a piece of data belongs to a category. Decision Trees: is a tree-like structure that has leaves, denoting groupings and divisions, which thus establish the conjunctions of salient aspects that motivate those classifications. Random Forest: generates multiple decision trees from random subsets of data. It may have less error compared to other algorithms. Support Vector Machines (SVM): The model labels classes based on certain inputs and the machine is trained to construct a predictive model to determine the class of newly introduced data. These algorithms undergo automatic improvement based on experience.	[92–94]
Linear regression	A simple regression is the calculation of the equation corresponding to the line that best describes the relationship between the response and the variable that explains it.	[95]
Clustering Algorithms	These are classification algorithms. KNN (K-Nearest Neighbors): Recognized for its nearest neighbor approach, it operates on the premise that similar data points tend to cluster in space. Agglomerative Clustering: This algorithm organizes data into a tree or dendrogram structure. DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Known for its ability to identify groups of varying densities within datasets.	[96,97]

Some of the most common applications in the area of fiscal control using the above techniques are as follows:

- Pattern identification: This involves detecting recurring relationships and structures within the data. Correlation analysis can reveal relationships between different expenditure variables [98].

- Anomaly detection: Identifies data that deviate from expected patterns, possibly indicating errors or suspicious activities. Algorithms like K-means clustering, Random Forest, PCA, and Neural Networks can be used for this purpose [99].
- Predictive Analytics: Uses statistical and ML models, such as linear regression, Random Forest, and Neural Networks, to predict future spending based on historical data [100].
- Trend Analysis: Examines data over time to identify trends and changes [101].

  Various tools and platforms can be used to carry out these analyses:
- Statistical Software: R, SAS, SPSS.
- BI (Business Intelligence) tools: Tableau, Power BI, QlikView.
- Programming Languages: Python (pandas, NumPy, scikit-learn, statsmodels), R.

6. Architecture and Interoperability

In large data environments, building an architecture that enables effective interoperability is crucial for accurate business insights and informed decision-making [102]. Interoperability is especially challenging in the public sector, where data collection involves diverse formats—images, text, audio, and video—across fragmented databases, incompatible systems, and proprietary software, limiting efficient data exchange, analysis, and interpretation [102].

Recognizing the challenge of data interoperability, various industries have adopted solutions to improve access, analysis, and communication between systems, devices, and applications. Application Programming Interfaces (APIs) have emerged as a valuable tool for managing data flow within internal applications, enabling secure and efficient data exchange and functionality sharing [102].

Government agencies generate vast amounts of structured and unstructured data, which can enhance services and policy-making processes. However, leveraging these data for informed decision-making necessitates robust tools and architectures [103].

Interoperability is vital for future advancements, especially in the public sector. Emerging technologies like IoT, SaaS, and cloud computing rely on APIs to improve performance in security, revenue generation, and customer insights [102,104]. APIs standardize data access and transfer between systems, facilitating integration with legacy applications and optimizing data utilization [102]. By abstracting technologies and data formats, APIs ensure seamless integration, making them essential for achieving interoperability goals [105,106]. APIs solve many of the challenges related to data format and data capture, among others. There are a number of additional challenges that exist when talking about interoperability, which are outlined below [107].

- Managing data at scale: Implementing interoperable systems necessitates consolidating data from multiple sources, often stored in incompatible formats. This process requires specialized expertise and substantial computing resources for extraction, cleansing, transformation, and loading [104,107].
- Addressing privacy concerns: Interoperability involves security measures to protect user information, especially when multiple systems exchange data through complex channels. This necessitates reliable security technologies and policies, leading to additional costs [107].
- Enforcing interoperability standards: Organizations often use customized systems that lack a common standard, which is necessary for effective communication. Introducing interoperability standards requires upgrades to equipment, software, and data infrastructure [107].
- Cross-platform incompatibility: Each system has its own data format, leading to inefficiencies during file exchanges. Data often require time-consuming and error-prone conversion processes [107].

- Data loss during transfer: Critical information can be lost when converting or transferring files, affecting project integrity and necessitating rework or manual adjustments [107].

When structuring systems to capture data from various sources, addressing challenges is crucial to prevent future issues, inconsistencies, or data loss that could impact business operations or investigations. In the public sector, promoting interoperability in data management is essential for fiscal control and auditing, ensuring that valuable information is not overlooked and aiding in fraud prevention efforts.

7. Data Visualization

This section outlines the fundamental principles and advanced technologies of modern data visualization. Technological advances in information technology have transformed data presentation compared to traditional print media like books and newspapers [108].

Data visualization is increasingly important for understanding business processes and facilitating informed decision-making [109]. Various theories and methodologies have emerged to establish best practices for creating effective data visualization dashboards that enhance communication and user experience [110].

To engage in data visualization, it is essential to understand foundational theories, including the human visual system, appropriate visual metaphors, Gestalt psychology, and diverse data types. Proficiency in economics, statistics, color theory, graphic design, visual storytelling, and emotional intelligence skills is also crucial [111,112].

7.1. Basic Concepts for Data Visualization

User Experience (UX) focuses on the interaction between users and products or services, particularly in relation to websites and applications. Multiple factors influence user perception, and if a system's design does not meet user needs, acceptance will decline. UX and usability are essential for ensuring that systems effectively fulfill users' requirements [113].

User Interface (UI) encompasses the various screens, pages, and visual components, such as buttons and forms, through which users interact with a digital product or service. UI can be categorized into three levels [114], as shown in Table 7.

Color theory plays a vital role in UI design by influencing user perception and associations with visual elements. The choice of a color palette significantly impacts user decisions [115]. For instance, vibrant colors enhance the visibility of elements, as depicted in Figure 6, where birds stand out against grayscale backgrounds.

This influence extends to alarm systems, where a neutral color palette is used until an alarm is activated, prompting a shift to a more prominent hue. This change enhances visibility and serves as a clear indicator of significant changes, with green typically representing positive changes and red signaling alarms or potential danger [116,117].

Table 7. User interface levels.

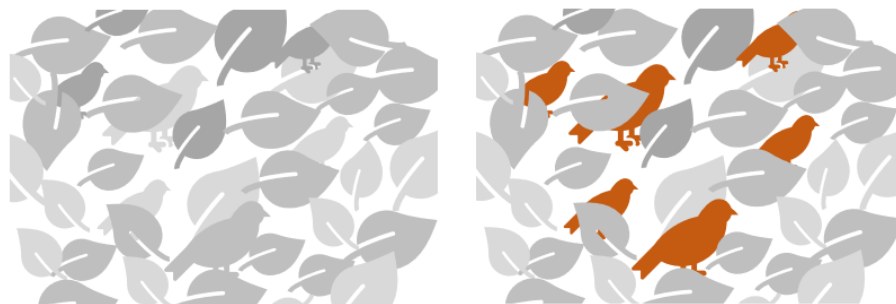| Item	Description
Conceptual	Consists of the general idea of application use and operation.
Functional	It focuses on the input language, including what operations are available and how the user gets to use them.
Visual	Determines how the application and data are presented to the user.

Figure 6. Ease of display and viewing the birds.

7.2. Principles of Data Visualization

According to [118,119], there are a number of general principles that apply to a complete presentation or analysis and serve as a guide for constructing new visualizations, as follows.

7.2.1. Structure Planning

Despite being non-technical, this aspect is fundamentally important. Before creating any visual elements, prioritizing the information to be communicated is essential. Although this may seem obvious, focusing first on the information and message is crucial before using software tools that could limit or distort visual options [118,119].

7.2.2. Choosing the Right Tool

To achieve effective visual effects, mastering specific software programs is crucial. It is unrealistic to expect complex and effective results from tools like spreadsheets or other software not designed for creating advanced graphics [118,119].

7.2.3. Use Effective Geometries to Represent Data

Geometries refer to the shapes and characteristics associated with specific graph types; for instance, bar geometries are used for bar graphs. While it may be tempting to move directly from a dataset to a familiar geometry, it is important to remember that geometries are visual representations of data. Often, considering multiple geometries is necessary to achieve a comprehensive and effective representation of the data [118–120].

7.2.4. Color Always Conveys Meaning

Incorporating color into visualizations can be very powerful, as there is typically no reason to avoid it. A comprehensive study found that color visualizations are often easier to remember, highlighting their effectiveness in creating memorable representations [117–119].

7.2.5. Incorporate Uncertainty

Uncertainty is an inherent aspect of understanding most systems, and omitting it from visualizations can lead to misinterpretation. Including uncertainty in visual representations is crucial to prevent misunderstandings and misrepresentations. Often overlooked, uncertainty can omit critical statistical messages and impact other aspects of the analysis, such as inferences about the mean [118,119].

7.2.6. Distinction Between Data and Models

Information can be presented in various forms, such as raw data (e.g., scatter plots), summary data (e.g., box plots), or inferential statistics (e.g., fitted regression lines). While raw and summary data are generally straightforward, graphical models require detailed explanations for reproducibility [118,119]. Comprehensive presentations of models should include visual representations with explanatory legends or references to relevant sections for necessary details.

7.2.7. Seek External Opinions

While established principles and theories inform effective data visualization, the most impactful figures resonate with readers. Authors should seek external reviews of their figures, as many are created quickly during study development and often lack objective feedback, even if crafted carefully [118,119]. Following these principles aims to convey messages effectively, making improved design and presentation essential for enhancing communication and reducing misinterpretations.

7.2.8. Graphic Quality Indicators

Table 8 contains some of the most prominent and relevant indicators when designing graphs for data visualization [118,119]. These indicators can be used as a guide to assess whether good data visualization practices are being followed [121].

Table 8. Relevant indicators to data visualization.

Indicators	Description	Ref.
Graphical integrity	This indicator refers to the possibility of distortion or deception in the graphical representation of data. For example, the use of three-dimensional effects in a graph can create the illusion that one piece of data is larger than another when it is not. This indicator addresses any element that may lead to misinterpretation of data.	[118,120]
Excessive chart elements	Chart junk refers to the presence of visual elements that do not correspond to variations in the data or that make the chart difficult to interpret. For example, the use of textures in bar or line charts, or the inclusion of unnecessary grids in some charts, can create unwanted visual effects.	[118]
Data Ink	This indicator measures the amount of ink used to represent the data compared to the total amount of ink used in the chart. For example, charts with very sharp borders can make it difficult to extract information efficiently.	[118]
Data Density	Data density refers to the ratio of the area occupied by the chart to the amount of data it contains. If a chart displays only four pieces of data, it does not need to take up a lot of space; space can be optimized by making the chart smaller. This makes it easier to understand the information and increases the efficiency of visualization by making better use of the available space.	[118]

7.3. Public Sector Data Visualization

Governments worldwide have diligently worked to establish a culture of open government, embracing principles like transparency and citizen participation in decision-making and policy formulation. Open data signifies a shift toward transparency and accountability, fostering public engagement through accessible data and promoting a culture of information sharing and collaboration [103]. These cultural shifts not only change how governments operate but also drive economic innovation, enhance performance, and increase accountability among elected officials with active citizen involvement [103]. A crucial aspect of this transformation is the presentation of data visualization, as it significantly impacts understanding and decision-making; therefore, the visual presentation of government data must be clear and accurate, adhering to visualization theory principles to promote transparency [103,122].

In technical terms, it is essential to have the appropriate tools to perform an adequate visualization of data for fiscal control, some of which are shown in Table 9.

Table 9. Tools for data visualization.

Tool	Advantages	Disadvantages	Ref.
Tableau	Easy to use interface, with extensive community support and resources. It also has a variety of customization options.	Expensive, especially for features at the enterprise level.	[123]
Power BI	It has an easy-to-use interface. Provides drag-and-drop functionality that makes it easy to create compelling visualizations. Includes advanced dash-boarding and reporting features that allow users to effectively monitor and analyze data.	Advanced functionality requires knowledge of Power Query and DAX.	[124]
Google Data Studio	It is free and user friendly.	Lack of sophisticated analytics and limited customization.	[125]
Plotly	Interactive, customizable, open source libraries with active communities.	Fewer predefined chart types compared to other visualization platforms.	[126]
Grafana	Open source and compatible with a wide range of databases, it provides real-time monitoring.	Some dashboards can be complex to customize.	[127]
IBM Cognos Analytics	It has enterprise-level features and powerful protection and management tools.	Complex reporting requires learning, and large deployments require resources.	[128]

It is important to understand some data visualization tools and recognize their importance in the fields of financial reporting, auditing, and fiscal control. The choice of appropriate data visualization is essential, as it can facilitate the identification of irregularities or suspicious patterns that may indicate fraudulent activity. This approach not only strengthens detection capabilities, but also promotes transparency in the public sector.

8. Results: A Case Study for Fiscal Control in Colombia

The Colombian Comptroller General's Office (CGR, or Contraloría General de la República) is the nation's highest fiscal oversight authority, which is responsible for ensuring the efficient and lawful use of public resources. Established as a constitutional entity, the CGR has evolved significantly since the 1945 constitutional reform that transformed it from a technical accounting department into a comprehensive supervisory body. Subsequent reforms in 1968 and 1976 refined its role [129], emphasizing financial and legal oversight, and introducing prior and perceptual control mechanisms. The 1991 constitutional reform shifted to posterior and selective control [130], reducing interference in audited entities' decisions and promoting citizen participation in fiscal monitoring. In 2019 further modifications introduced concurrent and preventive control measures [131]. Over time, the CGR has integrated advanced technologies like big data analytics and AI to enhance its real-time oversight capabilities while continuing to foster transparency, accountability, and citizen involvement in safeguarding public resources.

Recognizing the need to process large volumes of data to ensure the proper use of public resources, the CGR's strategic plan for 2018–2022 focused on enabling digital transformation through enterprise architecture [132]. The plan included the creation of the Integrated Information Center, a hub designed to enhance data monitoring, analysis, and interoperability between the organization and its strategic partners. The Center leverages cutting-edge technologies, such as AI, ML, and blockchain, to support fiscal control actions, strategic decision-making, and the optimization of institutional processes, while also fostering inter-institutional collaboration and enhancing participatory fiscal control. In this regard, the CGR established the Directorate of Information, Analysis, and Immediate Reaction (DIARI) as part of the country's broader efforts to integrate advanced technologies into the supervision of public resources.

The establishment of DIARI was formalized by the Legislative Act 04 of 2019 and was regulated by Decree Law 2037 of 2019 [133]. The DIARI is structured into three units:

Information, Analysis, and Immediate Reaction, each playing a crucial role in accessing and analyzing data to issue preventive alerts and assess risks to public resources [134]. The Information Unit connects to relevant data sources for fiscal supervision, ensuring data quality before passing it to the Analysis Unit. The Analysis Unit develops and validates analytical models to address critical business questions, producing data reports, alerts, and dashboards that highlight potential fiscal risks. The Immediate Reaction Unit is responsible for implementing timely and effective real-time supervision actions to protect public resources at risk of imminent loss. Together, these units ensure comprehensive monitoring, analysis, and rapid response in safeguarding public resources [134].

This new fiscal control model, which combines preventive, concomitant, and subsequent oversight, aims to improve the efficiency and timeliness of fiscal control in Colombia. To support this model, there is a critical need for advanced data capture systems that can efficiently gather information from diverse sources, robust data storage solutions that handle large volumes of information securely, and powerful data processing capabilities to analyze and extract actionable insights in real-time. Furthermore, achieving interoperability between various data systems is essential to ensure seamless information exchange across platforms. Effective data visualization tools are also necessary to present complex data in an accessible and intuitive way, allowing decision-makers to quickly identify potential risks and respond accordingly. The DIARI's multidisciplinary team, composed of professionals from over 15 fields, is tasked with integrating these needs by connecting information sources, structuring data, and coordinating continuous monitoring efforts, ensuring that the CGR can respond quickly to potential threats to public resources.

8.1. Co-Creation Workshops

A series of co-creation workshops were organized in Bogotá at the end of 2023, bringing together participants from the Directorate of Information, Analysis, and Immediate Reaction (DIARI), academia, and various companies in the technology sector in the process of defining a strategic platform for smart supervision of public expenditure in the context of Society 5.0 [135]. These workshops focused on defining the user requirements and engineering characteristics necessary for the DIARI's digital transformation. Through a participatory approach involving open innovation and design thinking techniques, participants participated in in-depth discussions and collaborative activities to address the challenges facing DIARI, Figure 7. The workshops provided a platform for analyzing the current processes of the three DIARI operational units, identifying areas for improvement, and proposing innovative solutions to optimize these processes.

Figure 7. Workshops: CGR's DIARI + academia + industry.

A key outcome of the workshops was the collaborative development of proposed architectures for data capture, storage, processing, interoperability, and visualization. These architectures were designed to address the specific needs identified during the workshops, ensuring that they aligned with DIARI's strategic objectives and technological capabilities. The interdisciplinary nature of the workshops, with contributions from DIARI officials, industry experts, and academic representatives, enriched the process, resulting in comprehensive and practical solutions. The finalized proposals included detailed implementation action plans, along with mechanisms for monitoring and evaluation to ensure the success of the digital transformation initiatives. Table 10 contains the activities that were carried out during the workshops.

Table 10. Activities conducted in the workshops with DIARI officials

Workshop	Activity 1	Activity 2	Activity 3
(1) Data Acquisition and Capture	• Select a process or activity involving data acquisition and capture. • Draw a flow chart or representative diagram.	Conduct a SWOT analysis, including Weaknesses, Opportunities, Strengths, and Threats related to the selected process.	Answer a questionnaire from the user's perspective, focusing on data acquisition and capture.
(2) Data Storage	• In pairs, select a process or activity involving data storage. • Draw a flow chart or representative diagram.	Conduct a SWOT analysis, including Weaknesses, Opportunities, Strengths, and Threats related to the selected process.	Answer a questionnaire from the user's perspective, focusing on data storage.
(3) Data Processing and Representation	• In groups, draw a model representing the components of data processing and presentation for DIARI.	Answer a questionnaire from the user's perspective, focusing on data processing and representation.	No activity

8.2. User Requirements

Based on the workshops focused on data acquisition and collection, data storage, and data processing, 15 flowcharts were produced, representing day-to-day processes in DIARI that involve data at various stages; 12 SWOT matrices corresponding to the analysis of the processes depicted in the flowcharts and 11 responses to two surveys with a total of 60 questions. Using this collected information and a study of general information about DIARI, analyses and filtering processes were conducted, leading to consolidated representations reflecting the perspectives of the workshop participants. The first of these outcomes is a general SWOT matrix presented in Table 11. It shows the general categories corresponding to weaknesses, opportunities, strengths, and threats, and also includes suggestions for improvement strategies. This matrix summarizes the analysis performed by DIARI officials on the flowcharts of processes they carry out daily.

Table 11. Consolidated SWOT Analysis.

Category	Description	Improvement Strategies
Strengths	- Experienced team - Experience in managing successful projects - Robust technological infrastructure - Diverse information sources	- Leveraging the team and experience - Exploiting existing infrastructure - Migration to cloud solutions
Weaknesses	- Dependency on personnel - Manual and non-automated processes - Non-permanent monitoring - Outdated and duplicated information - Traceability issues	- Process automation - Continuous updating and validation of data - Traceability protocols - Platform integration
Opportunities	- Construction or acquisition of traceability tools - Automation of alert assignment and monitoring - Application of cutting-edge technologies - Advanced data processing techniques	- Development of an automated dashboard - Redefinition of technological architecture - Technical collaboration tables - Automation of data validation and management
Threats	- Risks of information loss or deterioration - Potential cyberattacks - Dependency on government policies	- Backup and disaster recovery solutions - Strengthening cybersecurity - Risk management strategy

8.2.1. Data Capture

The analysis of the responses to the surveys conducted has resulted in a list of requirements that serve as a guide to define the most appropriate architecture for data acquisition and capture, as follows.

a. Support for multiple data sources and formats. The data acquisition solution must be capable of connecting to and collecting data from various sources, such as application systems, databases, sensors, IoT devices, and social networks, while supporting a variety of data formats, including CSV, JSON, XML, and more.

b. Real-time and scheduled data capture. It must support both real-time data capture and the scheduling of automated data capture tasks at specific intervals to meet the entity's needs.

c. Secure data acquisition. It must ensure secure and encrypted connections with data sources to protect the confidentiality and integrity of the information, particularly when handling sensitive data.

d. Data validation and error management. The solution must include robust data validation mechanisms to ensure accuracy and quality, as well as error logging and management to provide detailed information on issues and exceptions.

e. Duplicate detection and management. It must have the capability to detect, manage, and prevent the capture of duplicate data, maintaining the integrity of the database.

f. Adaptability to data structure changes. It must be able to adapt to changes in data structures from various sources without significant disruptions to the data capture process.

g. High-speed data flow and performance optimization. It must handle high-speed data flows efficiently, optimizing performance to minimize latency and ensure the timely processing of data.

h. Regulatory compliance and documentation. The solution must comply with applicable regulations and standards regarding data privacy and security, while providing clear documentation of data capture processes, including sources, frequency, procedures, and metadata management.

i. Historical record and metadata management. It must maintain a historical record of captured data and manage metadata that describes the data source, structure, and transformation.
j. Integration with external sources. It must support integration with external data sources through APIs or other connection methods, ensuring seamless data flow and interoperability.
k. Testing and validation. The solution must allow for thorough testing and validation of data capture processes before they are implemented in a production environment.

8.2.2. Data Storage

The analysis of the responses to the surveys conducted has resulted in a list of requirements that serve as a guide for selecting the most appropriate architecture for data storage, as follows:

a. Scalability and format compatibility. The data storage system must be scalable to handle large volumes of data as the entity grows, and compatible with various data formats, such as CSV, JSON, XML, among others.
b. Performance, response times, and efficient access. It must provide high performance, fast response times, and allow efficient indexing and search capabilities to quickly and accurately retrieve information.
c. Data security and integrity. It must have strong security measures, including encryption of data at rest and in transit, ensure the integrity of stored data to prevent corruption or loss, and log and audit access to the data.
d. Backup, recovery, and redundancy. It must have a regular backup system and recovery capabilities, incorporate redundancy and fault tolerance, and allow data compression to optimize storage space usage.
e. Historical data maintenance and metadata management. It must be able to maintain and manage historical data for retrospective analysis, comply with applicable regulations, and maintain metadata records that describe the structure and content of the data.
f. Regulatory compliance and Retention Policies. It must comply with applicable regulations and standards, allow for the implementation of data retention policies to manage the deletion or archiving of data, and clearly document the data storage structure.
g. Ease of Integration and Space Recovery. It must be compatible with analysis and visualization tools, and have mechanisms for managing and recovering space when data are deleted or no longer needed.

8.2.3. Data Processing

The analysis of the responses to the surveys conducted has resulted in a list of requirements for data processing with Industry 4.0 tools.

a. Process automation. Develop an automated process management system that allows for the assignment, tracking, and traceability of requirements and alerts. Implement an automated workflow for receiving and processing alerts, eliminating reliance on manual processes.
b. Data integration. Establish a system that integrates data from multiple sources to ensure a comprehensive view of the information. Ensure data consistency and quality through integration with data quality tools.
c. Visualization and dashboards. Develop automated dashboards that display the real-time status of projects, alerts, and requirements, providing a comprehensive and detailed view of relevant information. Implement advanced data visualization tools to facilitate decision-making.
d. Natural Language Processing (NLP). Utilize NLP algorithms to extract relevant information and context from documents and texts, enhancing analytical and search capabilities.
e. Cloud infrastructure. Migrate to a cloud or a hybrid infrastructure to leverage the scalability, flexibility, and security it offers. This includes cloud storage and data processing.

f.   Data security. Implement robust data security measures to protect against cyber threats, including firewalls, intrusion detection systems, and encryption of sensitive data.

g.   Data governance. Establish clear policies and procedures for data governance to ensure the quality, integrity, and availability of information. Designate specific roles and responsibilities regarding data governance.

h.   Artificial Intelligence and Machine Learning. Utilize these techniques to identify patterns, trends, and risks in the data, enabling more informed decision-making.

i.   Automated data updates. Implement automatic data update processes that track and record the update dates of information sources.

j.   Staff training. Provide training to staff so they can effectively use the new tools and technologies implemented.

k.   Continuous monitoring. Establish continuous monitoring systems that alert about potential disruptions in the process, ensuring the continuity of operations.

l.   Continuity and sustainability. Design a continuity and sustainability strategy for the process, ensuring that new administrations continue to support these improvements.

m.   Risk assessment and mitigation. Conduct periodic risk assessments and establish mitigation plans to address identified threats, including political and security risks.

n.   Documentation and reporting. Generate customized reports and detailed documentation on data processing to ensure transparency and regulatory compliance.

8.2.4. Interoperability Needs

The key to establishing interoperability is determining the corporate operating model. According to The TOGAF Standard [136], the operating model is the necessary level of integration and standardization of business processes to provide goods and services to customers. The analysis of the responses to the surveys conducted has resulted in a list of requirements for interoperability.

a.   Extend data governance to more domains, integrating technical roles and unifying the criteria of the working units.

b.   Design and implement a comprehensive semantic model to improve internal interoperability and standardize information sources and data models, including contracts.

c.   Improve internal and external interoperability, including the integration of government data sources and interconnection of units to enhance the timeliness of information, track all alerts, and eliminate duplication.

d.   Strengthen the chain of custody for information exchanged and consider implementing a multi-cloud solution to mitigate the risk of low-quality data.

e.   Implement a robust data quality management model with a validator or supporting tool to enhance data cleansing processes and properly manage raw external data sources and data warehouse setup.

f.   Address reprocessing issues in databases caused by differences in data structures by standardizing and harmonizing data between units.

g.   Design and implement a complete information flow for interoperability between units, with a contingency plan to prioritize needs.

h.   Introduce real-time evaluation processes, integrate technologies and methods that support this capability, and leverage artificial intelligence models to effectively exploit the vast amount of existing information.

i.   Continue to develop enterprise architecture by supplementing the existing analytical and IT architectures with new data architectures, ensuring interoperability across the organization.

8.2.5. Data Visualization

To identify the purpose for which the visualizations will be used, it was necessary to first identify the people who could explain the characteristics of the case. Usually, this is not the only person who will use the solution. Throughout a problem or case, there are

several types of users, and it is important that each one is identified to ensure success in the design and solution of the case or problem. The following questions can be considered as input to determine user requirements regarding data visualization.

a. Who in the organization is investing in the visual representation of the data, and why? What are the project's objectives, and how do they align with the organization's needs?
b. Who will be using the solution, and who will act on the findings? What actions or decisions will they take based on the analysis?
c. Who will be presented with the results or findings, and what level of interaction and collaboration with others is expected? Will the data analysis be shared?
d. What questions can be answered with these data, and what more is expected to be discovered?
e. What do they want to achieve with the data that are not currently possible, and what more would they want to achieve if it were possible?
f. What is their current workflow, what tools do they use, and what difficulties do they encounter when analyzing the data?
g. Are there system limitations, and how could they be overcome? Do they understand what the system does with the data and the algorithms applied, or is it a black box (an unknown process)?
h. What is the end user's level of experience in analytics, and what tools are they currently using? How capable is this tool in visual data exploration, and do they know how to select appropriate charts based on the questions being asked?
i. Can they read visual patterns in the data, make notes, and act on insights?

9. Discussion

The literature review examines essential components of effective data management crucial for enhancing public expenditure oversight. It begins by highlighting the importance of understanding diverse data types and acquisition methodologies, focusing on recent research trends in data capture, storage, processing, interoperability, and visualization within the context of fiscal control. A significant trend is the increasing adoption of cloud-based solutions, which provide scalability and accessibility advantages over traditional on-premise data storage. This shift reflects broader global movements toward digital transformation in the public sector.

A key finding is the critical role of interoperability in facilitating information exchange among various systems and organizations within the public sector. Recent studies emphasize the integration of smart government processes—such as financial reporting and auditing—strengthened by technologies like the Internet of Things (IoT), Software as a Service (SaaS), cloud computing, and APIs. Countries such as the United States, Canada, and members of the European Union have been at the forefront of research in these areas, contributing significantly to the literature and best practices for effective fiscal oversight.

The role of public managers at various levels, from central government to municipal offices, presents both challenges and opportunities in the realm of data management and fiscal control. Public managers often face challenges such as resource constraints, resistance to change, and the need for training in advanced technologies. At the central government level, managers are tasked with implementing comprehensive data management strategies across multiple agencies, requiring coordination and integration of disparate systems. Conversely, this complexity also presents an opportunity for central managers to champion cross-agency collaboration, promoting a unified approach to fiscal oversight that can enhance overall transparency and efficiency.

At the municipal level, public managers deal with unique challenges, including limited budgets and a narrower range of technological resources. However, municipalities also have the advantage of being closer to citizens, which allows for more direct engagement and feedback. This proximity enables managers to tailor data management initiatives to the specific needs of their communities, fostering innovation in public service delivery. By leveraging local partnerships and community resources, municipal managers can imple-

ment agile solutions that enhance data capture and visualization, improving oversight of public expenditures.

Furthermore, the discussion must include the evolving role of citizens in Society 5.0, where technology enhances human-centric approaches to governance. Citizens are not only users of public services but also active participants in monitoring public spending. In this context, open data initiatives empower citizens to engage with government data, enhancing transparency and accountability. By providing tools and platforms for data visualization, governments can encourage citizen participation in fiscal oversight, allowing communities to identify potential risks and hold officials accountable.

This participatory model shifts the perception of citizens from passive recipients of services to active co-producers of governance, aligning with the principles of Society 5.0, which emphasizes a symbiotic relationship between humans and technology. Citizens equipped with data literacy skills can analyze public spending patterns and advocate for responsible resource allocation, thereby reinforcing democratic engagement and civic responsibility.

Ongoing advancements in technologies supporting public spending supervision are essential. These developments aim to produce more accurate reviews, leveraging automation to reduce constraints on fiscal supervision capacity and optimize resource allocation. Notably, collaborative research efforts have emerged as a vital strategy in this domain. Analysis of co-authorship networks indicates that research groups from academia, governmental organizations, and industry are increasingly collaborating on projects related to fiscal control. Identifying leading co-authors in this field reveals a network of prominent scholars and practitioners who are shaping research trends and fostering interdisciplinary approaches.

The SWOT analysis conducted with diverse stakeholders proved to be a vital tool for understanding user requirements in the design of respective architectures. By involving DIARI officials, industry experts, and academic representatives, the SWOT analysis offered a comprehensive view of the strengths, weaknesses, opportunities, and threats in current data management processes. This collaborative approach highlighted specific needs and challenges, facilitating the identification of strategic improvement areas. Insights from this analysis were instrumental in ensuring that the developed architectures are both technically robust and aligned with practical realities and user expectations.

User requirement lists, generated from multi-stakeholder workshops, are critical for designing a robust data management architecture tailored to organizational needs. These lists ensure that the architecture addresses diverse functional requirements, aligns with strategic goals, and prioritizes key features like real-time data capture, secure storage, and interoperability. Reflecting stakeholder input enhances system flexibility, collaboration, and ownership while also mitigating risks such as integration challenges and data quality issues. This comprehensive approach ensures effective support for smart oversight of public expenditure.

9.1. Limitations

It is essential to acknowledge the inherent limitations of this review and the case study presentation. While synthesizing existing literature and identifying current trends in data capture, storage, processing, interoperability, and visualization, the findings depend on the quality and scope of the included studies. The depth of analysis may vary and may not encompass emerging developments or specific contextual nuances. Additionally, the review relied on published sources, potentially excluding recent unpublished or proprietary advances from technology companies. This limitation may result in an incomplete view of the latest innovations or practical applications. Furthermore, the variability in methodologies and frameworks used in the analyzed studies could affect the generalizability of the conclusions drawn.

9.2. Proposed Research Agenda

The literature review and case study on Colombia highlight several critical areas for future research in data management for public expenditure oversight. A proposed research

agenda focuses on key topics, existing gaps, and methodologies to enhance fiscal control through effective data handling.

First, data capture methodologies need further exploration, particularly in acquiring data from real-time sources such as IoT devices and social media. Research should investigate the role of emerging technologies like blockchain in improving data capture efficiency and security, utilizing comparative case studies across countries.

Second, studies on cloud-based storage solutions versus traditional systems are essential, especially concerning sustainability, cost-effectiveness, and security in developing countries like Colombia. Longitudinal studies could map performance and user satisfaction over time, informing better decision-making.

Research on interoperability frameworks is also crucial, focusing on integrating data systems across public institutions. A design science research approach could develop frameworks that address the specific data-sharing needs of different government levels and citizens.

The contribution of data processing and analytics to fiscal oversight should be examined, particularly the application of machine learning for predictive analytics in risk assessment. Experimental designs can identify effective algorithms for detecting financial anomalies.

Data visualization for public engagement warrants focused research to enhance citizen oversight of public funds. Investigating how different demographic groups interact with visualization formats will provide insights into engagement strategies, utilizing qualitative methods like surveys and focus groups.

The impact of citizen participation in Society 5.0 is significant, emphasizing citizens as key players in holding public spending accountable. Research should explore participatory methods that enable citizens to co-design oversight tools, testing their effectiveness in promoting transparency.

Lastly, examining collaborative research networks will reveal trends in co-authorship and collaborative efforts in public expenditure oversight. Bibliometric analysis can identify major research networks and their contributions, while qualitative studies can explore collaboration dynamics.

10. Conclusions

The findings of this review, regarding data capture, storage, processing, interoperability, and visualization for smart public expenditure control, underscore the importance of adopting an integrated approach that combines several technologies. This approach can significantly improve public spending oversight by fostering more efficient, transparent, and accountable management of public resources.

In the public sector, analytics and data management are increasingly recognized as essential for continuous monitoring. Persistent challenges such as corruption, fund mismanagement, and inefficiency have highlighted the need for robust data analysis strategies. Implementing techniques that facilitate intelligent monitoring in the public sector is, therefore, recommended. Such techniques refer to advanced technologies such as machine learning, artificial intelligence, and blockchain.

The integration of advanced technologies presents a transformative opportunity to improve financial reporting and auditing in public organizations in the transition from Industry 4.0 to Society 5.0. Such technologies offer unprecedented capabilities for detecting irregularities, optimizing resource management, and improving operational efficiency. However, their effective implementation requires training and cultural change within public sector institutions.

The identification of user requirements in this study advances the state of the art by addressing not only the technological aspects of implementing data capture and processing solutions but also by considering the critical process improvement and human factors necessary for successful deployment. By thoroughly analyzing the needs of various users, including their interaction with data visualization tools and their ability to engage with

complex data structures, this study provides valuable insight into integration issues and the broader implications for corporate data management. Furthermore, the focus on adapting technology to meet specific user needs aligns with the growing emphasis on developing new competencies and skilling within organizations, offering practical guidance for professionals aiming to optimize the deployment of these technologies in real-world scenarios. This approach not only improves the efficacy of technology adoption, but also contributes to a more nuanced understanding of the intersection between technology, process, and human capital in modern data-driven environments.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available on request from the authors.

Conflicts of Interest: Authors Jaime A. Restrepo-Carmona, Carlos A. Escobar, Julián Sierra-Pérez and Rafael E. Vásquez were employed by the company Corporación Rotorr, Universidad Nacional de Colombia. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
API	Application Programming Interface
CGR	Contraloría General de la República
	(Office of the Comptroller General of the Republic)
CPU	Central Processing Unit
DAMA	Data Management International
DIARI	Directorate of Information, Analysis, and Immediate Reaction of the CGR
DT	Digital Transformation
ETL	Extract, Transform, and Load
GFS	Google File System
HDFS	Hadoop Distributed File System
IoT	Internet of things (IoT)
OCR	Optical Character Recognition
PCA	Principal Component Analysis
RQ	Research Question
SaaS	Software as a Service
SQL	Structured Query Language
TOGAF	The Open Group Architecture Framework
UI	User Interface
UX	User Experience

References

1. DAMA-International. *The Data Management Body of Knowledge (DAMA DMBOK)*; Technics Publications: Sedona, AZ, USA, 2017.
2. Lone, K.; Sofi, S.A. Effect of Accountability, Transparency and Supervision on Budget Performance. *Utopía Y Prax. Latinoam.* **2023**, *25*, 130–143.

3. Bharany, S.; Sharma, S.; Khalaf, O.I.; Abdulsahib, G.M.; Al Humaimeedy, A.S.; Aldhyani, T.H.H.; Maashi, M.; Alkahtani, H. A Systematic Survey on Energy-Efficient Techniques in Sustainable Cloud Computing. *Sustainability* **2022**, *14*, 6256. [CrossRef]

4. Kubina, M.; Varmus, M.; Kubinova, I. Use of Big Data for Competitive Advantage of Company. *Procedia Econ. Financ.* **2015**, *26*, 561–565. [CrossRef]

5. Spiekermann, S.; Novotny, A. A vision for global privacy bridges: Technical and legal measures for international data markets. *Comput. Law Secur. Rev.* **2015**, *31*, 181–200. [CrossRef]

6. Fang, C. Taxation with information: Impacts of customs data exchange on tax evasion in Pakistan. *Econ. Syst.* **2024**, 101243. [CrossRef]

7. Sisto, R.; Garcia, J.; Quintanilla, A.; deJuanes, A.; Mendoza, D.; Lumbreras, J.; Mataix, C. Quantitative Analysis of the Impact of Public Policies on the Sustainable Development Goals through Budget Allocation and Indicators. *Sustainability* **2020**, *12*, 10583. [CrossRef]

8. Alsaadi, A. Financial-tax reporting conformity, tax avoidance and corporate social responsibility. *J. Financ. Report. Account.* **2020**, *18*, 639–659. [CrossRef]

9. Petropoulos, T.; Thalassinos, Y.; Liapis, K. Greek Public Sector's Efficient Resource Allocation: Key Findings and Policy Management. *J. Risk Financ. Manag.* **2024**, *17*, 60. [CrossRef]

10. Gao, S. An Exogenous Risk in Fiscal-Financial Sustainability: Dynamic Stochastic General Equilibrium Analysis of Climate Physical Risk and Adaptation Cost. *J. Risk Financ. Manag.* **2024**, *17*, 244. [CrossRef]

11. Dammak, S.; Jmal Ep Derbel, M. Social responsibility and tax evasion: Organised hypocrisy of Tunisian professionals. *J. Appl. Account. Res.* **2023**, *25*, 325–354. [CrossRef]

12. Adam, I.; Fazekas, M. Are emerging technologies helping win the fight against corruption? A review of the state of evidence. *Inf. Econ. Policy* **2021**, *57*, 100950. [CrossRef]

13. Gidigbi, M.O. Assessing the impact of poverty alleviation programs on poverty reduction in Nigeria: Selected programs. *Poverty Public Policy* **2023**, *15*, 76–97. [CrossRef]

14. Valle-Cruz, D.; García-Contreras, R. Towards AI-driven transformation and smart data management: Emerging technological change in the public sector value chain. *Public Policy Adm.* **2023**, 09520767231188401. [CrossRef]

15. Thomas, M.A.; Cipolla, J.; Lambert, B.; Carter, L. Data management maturity assessment of public sector agencies. *Gov. Inf. Q.* **2019**, *36*, 101401. [CrossRef]

16. Valle-Cruz, D.; Fernandez-Cortez, V.; Gil-Garcia, J.R. From E-budgeting to smart budgeting: Exploring the potential of artificial intelligence in government decision-making for resource allocation. *Gov. Inf. Q.* **2022**, *39*, 101644. [CrossRef]

17. Oliveira, T.A.; Oliver, M.; Ramalhinho, H. Challenges for Connecting Citizens and Smart Cities: ICT, E-Governance and Blockchain. *Sustainability* **2020**, *12*, 2926. [CrossRef]

18. Kankanhalli, A.; Charalabidis, Y.; Mellouli, S. IoT and AI for Smart Government: A Research Agenda. *Gov. Inf. Q.* **2019**, *36*, 304–309. [CrossRef]

19. Bendre, M.R.; Thool, V.R. Analytics, challenges and applications in big data environment: A survey. *J. Manag. Anal.* **2016**, *3*, 206–239. [CrossRef]

20. Aftabi, S.Z.; Ahmadi, A.; Farzi, S. Fraud detection in financial statements using data mining and GAN models. *Expert Syst. Appl.* **2023**, *227*, 120144. [CrossRef]

21. Parycek, P.; Schmid, V.; Novak, A.S. Artificial Intelligence (AI) and Automation in Administrative Procedures: Potentials, Limitations, and Framework Conditions. *J. Knowl. Econ.* **2023**, *15*, 8390–8415. [CrossRef]

22. Yahyaoui, F.; Tkiouat, M. Partially observable Markov methods in an agent-based simulation: A tax evasion case study. *Procedia Comput. Sci.* **2018**, *127*, 256–263. [CrossRef]

23. Rukanova, B.; Tan, Y.H.; Slegt, M.; Molenhuis, M.; van Rijnsoever, B.; Migeotte, J.; Labare, M.L.; Plecko, K.; Caglayan, B.; Shorten, G.; et al. Identifying the value of data analytics in the context of government supervision: Insights from the customs domain. *Gov. Inf. Q.* **2021**, *38*, 101496. [CrossRef]

24. Wynn, M.; Jones, P. Corporate Responsibility in the Digital Era. *Information* **2023**, *14*, 324. [CrossRef]

25. Haddaway, N.R.; Page, M.J.; Pritchard, C.C.; McGuinness, L.A. PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. *Campbell Syst. Rev.* **2022**, *18*, e1230. [CrossRef] [PubMed]

26. Maté Jiménez, C. Big data. Un nuevo paradigma de análisis de datos. *An. De Mecánica Y Electr.* **2014**, *91*, 10–16.

27. Muñoz, L.; Mazon, J.N.; Trujillo, J. ETL Process Modeling Conceptual for Data Warehouses: A Systematic Mapping Study. *IEEE Lat. Am. Trans.* **2011**, *9*, 358–363. [CrossRef]

28. Mositsa, R.J.; Van der Poll, J.A.; Dongmo, C. Towards a Conceptual Framework for Data Management in Business Intelligence. *Information* **2023**, *14*, 547. [CrossRef]

29. Cretu, C.; Gheonea, V.; Talaghir, L.; Manolache, G.; Iconomesu, T. Budget—Performance Tool in Public Sector. In Proceedings of the 5th WSEAS International Conference on Economy and Management Transformation, Timisoara, Romania, 24–26 October 2010.

30. Dawar, K.; Oh, S.C. *The Role of Public Procurement Policy in Drivingindustrial Development*; Technical Report; United Nations Industrial Development Organization (UNIDO): Vienna, Austria, 2017.

31. Adam, I.; Hernandez-Sanchez, A.; Fazeka, M. *Global Public Procurement OpenCompetition Index*; Technical report; Government Transparency Institute: Budapest, Hungary, 2021.

32. CABRI. *Value for Money in Public Spending*; Technical Report; CABRI: Kabri, Israel, 2015.

33. Popov, M.P.; Prykhodchenko, L.L.; Lesyk, O.V.; Dulina, O.V.; Holynska, O.V. Audit as an Element of Public Governance. *Stud. Appl. Econ.* **2021**, *39*, 1–9. [CrossRef]

34. Abdullah, A.S.B.; Zulkifli, N.F.B.; Zamri, N.F.B.; Harun, N.W.B.; Abidin, N.Z.Z. The Benefits of Having Key Performance Indicators (KPI) in Public Sector. *Int. J. Acad. Res. Account. Financ. Manag. Sci.* **2022**, *12*, 719–726. [CrossRef]

35. Leite, P.; George, T.; Sun, C.; Jones, T.; Lindert, K. *Social Registries for Social Assistance and Beyond: A Guidance Note & Assessment Tool*; Technical report; World Bank: Washington, DC, USA, 2017.

36. Han, Y. The impact of accountability deficit on agency performance: Performance-accountability regime. *Public Manag. Rev.* **2020**, *22*, 927–948. [CrossRef]

37. Vassiliadis, P. A Survey of Extract-Transform-Load Technology. *Int. J. Data Warehous. Min. (IJDWM)* **2009**, *5*, 1–27. [CrossRef]

38. Yaqoob, I.; Hashem, I.A.T.; Gani, A.; Mokhtar, S.; Ahmed, E.; Anuar, N.B.; Vasilakos, A.V. Big data: From beginning to future. *Int. J. Inf. Manag.* **2016**, *36*, 1231–1247. [CrossRef]

39. Eito-Brun, R. *Gestión de Contenidos*; Editorial UOC: Barcelona, Spain, 2014.

40. Hernandez, A.T.; Vazquez, E.G.; Rincon, C.A.B.; Montero-García, J.; Calderon-Maldonado, A.; Ibarra-Orozco, R. Metodologías para analisispolitico utilizando web scraping. *Res. Comput. Sci.* **2015**, *95*, 113–121. [CrossRef]

41. Kumar, N.; Gupta, M.; Sharma, D.; Ofori, I. Technical Job Recommendation System Using APIs and Web Crawling. *Comput. Intell. Neurosci.* **2022**, *2022*, 7797548. [CrossRef] [PubMed]

42. Puñales, E.M.; Salgueiro, A.P. Aplicación de minería de datos a lainformación recuperada de la intranet para agrupar los resultados relevantes. In *Jornada Científica ICIMAF-2015*; Instituto de Cibernética, Matemática y Física: Havana, Cuba, 2015.

43. Güemes, V.L. Business Intelligence Para la Toma de Decisiones Estratégicas: Un Casode Aplicación de Minería de Datos Dentro del Sector Bancario. Master's Thesis, Universidad de Cantabria, Santander, Spain, 2019.

44. Olson, D.L.; Lauhoff, G. *Descriptive Data Mining*; Springer: Singapore, 2019. [CrossRef]

45. He, S.; Zhu, J.; He, P.; Lyu, M.R. Experience Report: System Log Analysis for Anomaly Detection. In Proceedings of the 2016 IEEE 27th International Symposium on Software Reliability Engineering (ISSRE), Ottawa, ON, Canada, 23–27 October 2016. [CrossRef]

46. Bahri, M.; Bifet, A.; Gama, J.; Gomes, H.M.; Maniu, S. Data stream analysis: Foundations, major tasks and tools. *WIREs Data Min. Knowl. Discov.* **2021**, *11*, e1405. [CrossRef]

47. Krishnamurthi, R.; Kumar, A.; Gopinathan, D.; Nayyar, A.; Qureshi, B. An Overview of IoT Sensor Data Processing, Fusion, and Analysis Techniques. *Sensors* **2020**, *20*, 6076. [CrossRef] [PubMed]

48. Bazzaz Abkenar, S.; Haghi Kashani, M.; Mahdipour, E.; Jameii, S.M. Big data analytics meets social media: A systematic review of techniques, open issues, and future directions. *Telemat. Inform.* **2021**, *57*, 101517. [CrossRef] [PubMed]

49. Patnaik, S.K.; Babu, C.N.; Bhave, M. Intelligent and adaptive web data extraction system using convolutional and long short-term memory deep learning networks. *Big Data Min. Anal.* **2021**, *4*, 279–297. [CrossRef]

50. Gangavarapu, T.; Jaidhar, C.D.; Chanduka, B. Applicability of machine learning in spam and phishing email filtering: Review and approaches. *Artif. Intell. Rev.* **2020**, *53*, 5019–5081. [CrossRef]

51. Taleb, I.; Serhani, M.A.; Dssouli, R. Big Data Quality: A Survey. In Proceedings of the 2018 IEEE International Congress on Big Data (BigData Congress), San Francisco, CA, USA, 2–7 July 2018. [CrossRef]

52. World Economic Forum. *Data Integrity*; Technical report; World Economic Forum: Cologne, Switzerland, 2020.

53. OECD. *Data Accessibility: Open, Free and Accessible Formats*; OECD: Paris, France, 2019. [CrossRef]

54. Nikiforova, A. Data Security as a Top Priority in the Digital World: Preserve Data Value by Being Proactive and Thinking Security First. In *Springer Proceedings in Complexity*; Springer International Publishing: Cham, Switzerland, 2023; pp. 3–15. [CrossRef]

55. Gupta, I.; Singh, A.K.; Lee, C.N.; Buyya, R. Secure Data Storage and Sharing Techniques for Data Protection in Cloud Environments: A Systematic Review, Analysis, and Future Directions. *IEEE Access* **2022**, *10*, 71247–71277. [CrossRef]

56. Blumzon, C.F.I.; Pănescu, A.T. Data Storage. In *Good Research Practice in Non-Clinical Pharmacology and Biomedicine*; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; pp. 277–297. [CrossRef]

57. Mishra, S.P.; Sahoo, S.K.; Jena, B.; Tirthankar. Migrating on-premise application workloads to a hybrid cloud architecture. *J. Inf. Optim. Sci.* **2022**, *43*, 1099–1108. [CrossRef]

58. Sriramoju, S. A Comprehensive Review on Data Storage. *Int. J. Sci. Res. Sci. Technol.* **2019**, *6*, 236–241. [CrossRef]

59. Sen, R.; Sharma, A. Optimization of Cost: Storage over Cloud Versus on Premises Storage. In Proceedings of the 2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT), Gwalior, India, 10–12 April 2020; pp. 179–181. [CrossRef]

60. Yang, P.; Xiong, N.; Ren, J. Data Security and Privacy Protection for Cloud Storage: A Survey. *IEEE Access* **2020**, *8*, 131723–131740. [CrossRef]

61. Syed, A.; Purushotham, K.; Shidaganti, G. Cloud Storage Security Risks, Practices and Measures: A Review. In Proceedings of the 2020 IEEE International Conference for Innovation in Technology (INOCON), Bangluru, India, 6–8 November 2020. [CrossRef]

62. Nachiappan, R.; Javadi, B.; Calheiros, R.N.; Matawie, K.M. Cloud storage reliability for Big Data applications: A state of the art survey. *J. Netw. Comput. Appl.* **2017**, *97*, 35–47. [CrossRef]

63. Saadoon, M.; Ab. Hamid, S.H.; Sofian, H.; Altarturi, H.H.; Azizul, Z.H.; Nasuha, N. Fault tolerance in big data storage and processing systems: A review on challenges and solutions. *Ain Shams Eng. J.* **2022**, *13*, 101538. [CrossRef]

64. Shvachko, K.; Kuang, H.; Radia, S.; Chansler, R. The Hadoop Distributed File System. In Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), Incline Village, NV, USA, 3–7 May 2010. [CrossRef]

65. Strohbach, M.; Daubert, J.; Ravkin, H.; Lischka, M. Big Data Storage. In *New Horizons for a Data-Driven Economy*; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 119–141. [CrossRef]

66. Shin, H.; Lee, K.; Kwon, H.Y. A comparative experimental study of distributed storage engines for big spatial data processing using GeoSpark. *J. Supercomput.* **2021**, *78*, 2556–2579. [CrossRef]

67. Verma, C.; Pandey, R. Comparative Analysis of GFS and HDFS: Technology and Architectural landscape. In Proceedings of the 2018 10th International Conference on Computational Intelligence and Communication Networks (CICN), Esbjerg, Denmark, 17–19 August 2018. [CrossRef]

68. Wang, M.; Li, B.; Zhao, Y.; Pu, G. Formalizing Google File System. In Proceedings of the 2014 IEEE 20th Pacific Rim International Symposium on Dependable Computing, Singapore, 18–21 November 2014. [CrossRef]

69. Rao, M.V. 16-Data duplication using Amazon Web Services cloud storage. In *Data Deduplication Approaches*; Thwel, T.T., Sinha, G., Eds.; Academic Press: Cambridge, MA, USA, 2021; pp. 319–334. [CrossRef]

70. Mondal, A.S.; Sanyal, M.; Barua, H.B.; Chattopadhyay, S.; Mondal, K.C. Comparative Analysis of Object-Based Big Data Storage Systems on Architectures and Services: A Recent Survey. *J. Inst. Eng. (India) Ser. B* **2024**, *105*, 685–700. [CrossRef]

71. Nambiar, A.; Mundra, D. An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management. *Big Data Cogn. Comput.* **2022**, *6*, 132. [CrossRef]

72. Oracle. What is MySQL? 2024. Available online: https://www.oracle.com/mysql/what-is-mysql/ (accessed on 15 August 2024).

73. Oracle. MySQL Documentation. 2024. Available online: https://dev.mysql.com/doc/ (accessed on 15 August 2024).

74. IBM. What is PostgreSQL? 2024. Available online: https://www.ibm.com/topics/postgresql (accessed on 15 August 2024).

75. The PostgreSQL Global Development Group. PostgreSQL 16.3. 2024. Available online: https://www.postgresql.org/docs/release/16.3/ (accessed on 15 August 2024).

76. Microsoft. SQL Server Technical Documentation. 2024. Available online: https://learn.microsoft.com/en-us/sql (accessed on 15 August 2024).

77. MongoDB. MongoDB Documentation. 2024. Available online: https://www.mongodb.com/docs/ (accessed on 15 August 2024).

78. The Apache Software Foundation. Cassandra Documentation. 2024. Available online: https://cassandra.apache.org/doc/latest/ (accessed on 15 August 2024).

79. Maharana, K.; Mondal, S.; Nemade, B. A review: Data pre-processing and data augmentation techniques. *Glob. Transit. Proc.* **2022**, *3*, 91–99. [CrossRef]

80. Luengo, J.; García-Gil, D.; Ramírez-Gallego, S.; García, S.; Herrera, F. *Big Data Preprocessing: Enabling Smart Data*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020. [CrossRef]

81. Liang, W.; Tadesse, G.A.; Ho, D.; Fei-Fei, L.; Zaharia, M.; Zhang, C.; Zou, J. Advances, challenges and opportunities in creating data for trustworthy AI. *Nat. Mach. Intell.* **2022**, *4*, 669–677. [CrossRef]

82. Shehab, N.; Badawy, M.; Arafat, H. Big Data Analytics and Preprocessing. In *Machine Learning and Big Data Analytics Paradigms: Analysis, Applications and Challenges*; Springer International Publishing: Cham, Switzerland, 2021; pp. 25–43. [CrossRef]

83. Yang, C.; Huang, Q.; Li, Z.; Liu, K.; Hu, F. Big Data and cloud computing: Innovation opportunities and challenges. *Int. J. Digit. Earth* **2016**, *10*, 13–53. [CrossRef]

84. Ahmed, N.; Barczak, A.L.C.; Susnjak, T.; Rashid, M.A. A comprehensive performance analysis of Apache Hadoop and Apache Spark for large scale datasets using HiBench. *J. Big Data* **2020**, *7*, 110. [CrossRef]

85. L'Esteve, R. Databricks. In *The Azure Data Lakehouse Toolkit*; Apress: Berkeley, CA, USA, 2022; pp. 83–139. [CrossRef]

86. Sreemathy, J.; Joseph, V.I.; Nisha, S.; Prabha, I.C.; Priya, R.M.G. Data Integration in ETL Using TALEND. In Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 6–7 March 2020. [CrossRef]

87. Dolev, S.; Florissi, P.; Gudes, E.; Sharma, S.; Singer, I. A Survey on Geographically Distributed Big-Data Processing Using MapReduce. *IEEE Trans. Big Data* **2019**, *5*, 60–80. [CrossRef]

88. Zhang, J.; Lin, M. A comprehensive bibliometric analysis of Apache Hadoop from 2008 to 2020. *Int. J. Intell. Comput. Cybern.* **2022**, *16*, 99–120. [CrossRef]

89. Sharma, M.; Kaur, J. A comparative study of big data processing: Hadoop vs. spark. In Proceedings of the 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 13–15 March 2019; pp. 690–701.

90. Ibtisum, S.; Bazgir, E.; Rahman, S.M.A.; Hossain, S.M.S. A comparative analysis of big data processing paradigms: Mapreduce vs. apache spark. *World J. Adv. Res. Rev.* **2023**, *20*, 1089–1098. [CrossRef]

91. Bawankule, K.L.; Dewang, R.K.; Singh, A.K. Historical data based approach to mitigate stragglers from the Reduce phase of MapReduce in a heterogeneous Hadoop cluster. *Clust. Comput.* **2022**, *25*, 3193–3211. [CrossRef]

92. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef]

93. Ariyaluran-Habeeb, R.A.; Nasaruddin, F.; Gani, A.; Hashem, I.A.T.; Ahmed, E.; Imran, M. Real-time big data processing for anomaly detection: A survey. *Int. J. Inf. Manag.* **2019**, *45*, 289–307. [CrossRef]

94. Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci.* **2021**, *2*, 160. [CrossRef]

95. Maulud, D.; Abdulazeez, A.M. A Review on Linear Regression Comprehensive in Machine Learning. *J. Appl. Sci. Technol. Trends* **2020**, *1*, 140–147. [CrossRef]

96. Bansal, M.; Goyal, A.; Choudhary, A. A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning. *Decis. Anal. J.* **2022**, *3*, 100071. [CrossRef]

97. Ezugwu, A.E.; Ikotun, A.M.; Oyelade, O.O.; Abualigah, L.; Agushaka, J.O.; Eke, C.I.; Akinyelu, A.A. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Eng. Appl. Artif. Intell.* **2022**, *110*, 104743. [CrossRef]

98. Psycharis, Y. Public Spending Patterns. In *Contributions to Economics*; Physica-Verlag HD: Heidelberg, Germany, 2008; pp. 41–71. [CrossRef]

99. Aliyu, G.; Umar, I.E.; Aghiomesi, I.E.; Onawola, H.J.; Rakshit, S. Anomaly Detection of Budgetary Allocations Using Machine-Learning-Based Techniques. In *Engineering Innovation for Addressing Societal Challenges*; Trans Tech Publications Ltd.: Lausanne, Switzerland, 2021. [CrossRef]

100. Wolniak, R.; Grebski, W. Functioning of predictie analytics in bussiness. *Sci. Pap. Silesian Univ. Technol. Organ. Manag. Ser.* **2023**, *2023*, 631–649. [CrossRef]

101. Bhikaji, V.; Abdul, S. Trends of public expenditure in India: An empirical analysis. *Int. J. Soc. Sci. Econ. Res.* **2019**, *4*, 3307–3318.

102. Mishra, R.; Kaur, I.; Sahu, S.; Saxena, S.; Malsa, N.; Narwaria, M. Establishing three layer architecture to improve interoperability in Medicare using smart and strategic API led integration. *SoftwareX* **2023**, *22*, 101376. [CrossRef]

103. Hagen, L.; Keller, T.E.; Yerden, X.; Luna-Reyes, L.F. Open data visualizations and analytics as tools for policy-making. *Gov. Inf. Q.* **2019**, *36*, 101387. [CrossRef]

104. Ramadhan, A.N.; Pane, K.N.; Wardhana, K.R.; Suharjito. Blockchain and API Development to Improve Relational Database Integrity and System Interoperability. *Procedia Comput. Sci.* **2023**, *216*, 151–160. [CrossRef]

105. Borgogno, O.; Colangelo, G. Data sharing and interoperability: Fostering innovation and competition through APIs. *Comput. Law Secur. Rev.* **2019**, *35*, 105314. [CrossRef]

106. Platenius-Mohr, M.; Malakuti, S.; Grüner, S.; Schmitt, J.; Goldschmidt, T. File- and API-based interoperability of digital twins by model transformation: An IIoT case study using asset administration shell. *Future Gener. Comput. Syst.* **2020**, *113*, 94–105. [CrossRef]

107. Amazon Web Services. What Is Interoperability? 2024. Available online: https://aws.amazon.com/what-is/interoperability/ (accessed on 15 August 2024).

108. Chen, J.X. Data Visualization and Virtual Reality. In *Handbook of Statistics*; Elsevier: Amsterdam, The Netherlands, 2005; pp. 539–563. [CrossRef]

109. Chandra, T.B.; Dwivedi, A.K. Data visualization: Existing tools and techniques. In *Advanced Data Mining Tools and Methods for Social Computing*; Elsevier: Amsterdam, The Netherlands, 2022; pp. 177–217. [CrossRef]

110. Prokofieva, M. Using dashboards and data visualizations in teaching accounting. *Educ. Inf. Technol.* **2021**, *26*, 5667–5683. [CrossRef]

111. Bina, S.; Kaskela, T.; Jones, D.R.; Walden, E.; Graue, W.B. Incorporating evolutionary adaptions into the cognitive fit model for data visualization. *Decis. Support Syst.* **2023**, *171*, 113979. [CrossRef]

112. Ryan, L. Data visualization as a core competency. In *The Visual Imperative*; Elsevier: Amsterdam, The Netherlands, 2016; pp. 221–242. [CrossRef]

113. Zaki, T.; Islam, M.N. Neurological and physiological measures to evaluate the usability and user-experience (UX) of information systems: A systematic literature review. *Comput. Sci. Rev.* **2021**, *40*, 100375. [CrossRef]

114. Lindholm, M.; Sarjakoski, T. Designing a Visualization User Interface. In *Visualization in Modern Cartography*; Elsevier: Amsterdam, The Netherlands, 1994; pp. 167–184. [CrossRef]

115. Ryan, L. The importance of visual design. In *The Visual Imperative*; Elsevier: Amsterdam, The Netherlands, 2016; pp. 153–175. [CrossRef]

116. Ware, C. Color. In *Information Visualization*; Elsevier: Amsterdam, The Netherlands, 2021; pp. 95–141. [CrossRef]

117. Ware, C. Foundations for an Applied Science of Data Visualization. In *Information Visualization*; Elsevier: Amsterdam, The Netherlands, 2021; pp. 1–29. [CrossRef]

118. Tufte, E.R. *Beautiful Evidence*; Graphics Press LLC: Cheshire, Connecticut, 2006.

119. Midway, S.R. Principles of Effective Data Visualization. *Patterns* **2020**, *1*, 100141. [CrossRef] [PubMed]

120. Ware, C. Images, Narrative, and Gestures for Explanation. In *Information Visualization*; Elsevier: Amsterdam, The Netherlands, 2021; pp. 331–358. [CrossRef]

121. Zhang, Y.; Cao, T.; Li, S.; Tian, X.; Yuan, L.; Jia, H.; Vasilakos, A.V. Parallel processing systems for big data: A survey. *Proc. IEEE* **2016**, *104*, 2114–2136. [CrossRef]

122. Pazzi, S.; Svetlova, E. NGOs, public accountability, and critical accounting education: Making data speak. *Crit. Perspect. Account.* **2023**, *92*, 102362. [CrossRef]

123. Tableau. Tableau. 2024. Available online: https://www.tableau.com (accessed on 15 August 2024).

124. Microsoft. Power BI. 2024. Available online: https://www.microsoft.com/en-us/power-platform (accessed on 15 August 2024).

125. Google. Data Studio: Make Interactive Data Visualizations. 2024. Available online: https://newsinitiative.withgoogle.com/resources/trainings/data-studio-make-interactive-data-visualizations/ (accessed on 15 August 2024).

126. Plotly. Plotly. 2024. Available online: https://plotly.com/ (accessed on 15 August 2024).
127. Grafana. Grafana. 2024. Available online: https://grafana.com/ (accessed on 15 August 2024).
128. IBM. IBM Cognos Analytics. 2024. Available online: https://www.ibm.com/products/cognos-analytics (accessed on 15 August 2024).
129. Congreso de la República de Colombia. Ley 20 de 1975. 1975. Available online: https://www.funcionpublica.gov.co/eva/gestornormativo/norma_pdf.php?i=79924 (accessed on 15 June 2024).
130. Congreso de la República de Colombia. Ley 42 de 1993. 1993. Available online: https://www.funcionpublica.gov.co/eva/gestornormativo/norma_pdf.php?i=289 (accessed on 15 June 2024).
131. Congreso de la República de Colombia. Acto Legislativo 04 de 2019. 2019. Available online: https://www.funcionpublica.gov.co/eva/gestornormativo/norma_pdf.php?i=100251 (accessed on 15 June 2024).
132. Contraloría General de la República. Informe de Gestión al Congreso y al Presidente de la República 2021–2022. 2022. Available online: https://www.contraloria.gov.co/en/resultados/informes/informes-constitucionales/historico-informes-constitucionales (accessed on 15 June 2024).
133. Presidencia de la República de Colombia. Decreto Ley 2037 de 2019. 2020. Available online: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=102213 (accessed on 15 April 2024).
134. Economía Colombiana. DIARI, la Herramienta Que Revolucionó en Tiempo y Resultados las Inspecciones Fiscales. 2024. Available online: https://www.economiacolombiana.co/desarrollo-futuro/plataforma-diari-herramienta-aliada-de-cgr-4055 (accessed on 15 June 2024).
135. Restrepo-Carmona, J.A.; Zuluaga, J.C.; Flórez, D.A.; Gómez, M.S.; Londoño, L.; Gómez, G.; Villamil, R.M.; Morales, O.; Hurtado, A.M.; Escobar, C.A.; et al. The Design of a Strategic Platform for the Smart Supervision of Public Expenditure for Colombia in the Context of Society 5.0. *Urban Sci.* **2024**, *8*, 117. [CrossRef]
136. The Open Group. *The TOGAF Standard*, 10th ed.; 2024. Available online: https://www.opengroup.org/togaf (accessed on 15 August 2024).