



Article

An Adaptable Big Data Value Chain Framework for End-to-End Big Data Monetization

Abou Zakaria Faroukhi ¹, Imane El Alaoui ², Youssef Gahi ^{1,*}  and Aouatif Amine ¹

¹ Laboratory of Engineering Sciences, Ibn Tofail University, Kenitra 14000, Morocco; abouzakaria.faroukhi@uit.ac.ma (A.Z.F.); aouatif.amine@uit.ac.ma (A.A.)

² Telecommunications Systems and Decision Engineering Laboratory, Ibn Tofail University, Kenitra 14000, Morocco; imane.el.alaoui@uit.ac.ma

* Correspondence: gahi.youssef@uit.ac.ma

Received: 23 October 2020; Accepted: 20 November 2020; Published: 23 November 2020



Abstract: Today, almost all active organizations manage a large amount of data from their business operations with partners, customers, and even competitors. They rely on Data Value Chain (DVC) models to handle data processes and extract hidden values to obtain reliable insights. With the advent of Big Data, operations have become increasingly more data-driven, facing new challenges related to volume, variety, and velocity, and giving birth to another type of value chain called Big Data Value Chain (BDVC). Organizations have become increasingly interested in this kind of value chain to extract confined knowledge and monetize their data assets efficiently. However, few contributions to this field have addressed the BDVC in a synoptic way by considering Big Data monetization. This paper aims to provide an exhaustive and expanded BDVC framework. This end-to-end framework allows us to handle Big Data monetization to make organizations' processes entirely data-driven, support decision-making, and facilitate value co-creation. For this, we present a comprehensive review of existing BDVC models relying on some definitions and theoretical foundations of data monetization. Next, we expose research carried out on data monetization strategies and business models. Then, we offer a global and generic BDVC framework that supports most of the required phases to achieve data valorization. Furthermore, we present both a reduced and full monetization model to support many co-creation contexts along the BDVC.

Keywords: big data; big data value chain; big data monetization; big data management; value co-creation

1. Introduction

A wide variety of data sources such as social media, mobile devices, and the Internet of Things (IoT) generate a considerable amount of data evolving rapidly in a highly connected society. By 2020, the International Data Corporation estimates that accumulated data should increase to 44 zeta-octets [1], and ABI Research predicts that 30 billion devices will be connected [2]. These data are expected to grow more, marking a new step in the era of Big Data.

Big Data was initially defined as large volume, high velocity, and data asset diversity, which require new processing capabilities to find hidden knowledge and improve decision-making [3]. These three essential characteristics (volume, velocity, and variety) have been extended to become more refined. This new extended definition is called 10Vs [4,5], grouping ten characteristics that should be considered in every big data use case. These 10Vs are:

Volume: Refers to a large amount of generated data that is no more possible to master using traditional processing and storage systems.

Variety: Refers to the nature of data that may follow different forms as structured (e.g., database records), semi-structured (e.g., XML and JSON documents), or unstructured data (e.g., IoT and Social Media).

Velocity: Refers to the high speed at which data are both gathered and processed.

Veracity: Refers to the characteristics of Big Data quality. Collected data are often uncertain and inaccurate and need to be preprocessed before going further.

Variability: Refers to the meaning of continuously changing data, which may impact their consistency and homogeneity.

Virality: Refers to the spread rate of data and sharing speed of insight over the network.

Viscosity: Refers to the time lag between when the event occurred and when it is described. It is mainly observed when handling massive datasets.

Validity: Refers to the state of data and its reliability for the intended use.

Visualization: Refers to comprehensive visual data representation, allowing the identification of correlations and hidden patterns.

Value: Corresponds to the insights and knowledge that are extracted from hidden, varied, and sophisticated big data.

The consideration of these Vs exposes business model implementations to several challenges, particularly those relating to computational capabilities and data sharing regulations [6,7]. Therefore, organizations are forced to redesign their data processes and propose well-adapted Data Value Chains (DVC) that are more suitable for Big Data needs. This rethinking of data management has led companies to adopt what we call Big Data Value Chain (BDVC) to master every step the data passes through [3,8–10].

BDVC allows the management of the flow of data based on several phases to generate value and follow data evolution. These phases are data generation, data acquisition, data preprocessing, data storage, data analysis, data visualization, and data exposition [11].

It is worth noting that processes involved in a BDVC act on a large amount of data to achieve a final objective, extracting reliable information. By contrast, the intermediate states before we can attain knowledge are also of utmost importance for either the company or its partners. Therefore, each form of data must be carefully considered. In this regard, BDVC is a suitable pattern. It is a flexible and adaptable set of stages, not only allowing data to be made useful at each state but also monetizing and sharing valuable content and insights with contributors partners [12,13].

Big Data monetization, according to Prakash, refers to the organizations' capabilities to generate new kinds of revenue from their internal and external data sources, based on data processing, to create useful information, insight, and observations [14].

The modular and complete structure of the BDVC makes it very promising to monetize Big Data as each phase allows us to valorize the processed records at its level. Furthermore, Big Data monetization could be an integral part of the BDVC to support data-driven management based on interconnected processes with partners [11,12]. It thus reduces the gap between native monetization and the corresponding business model. However, few contributions to this field have addressed the BDVC in a general way by emphasizing Big Data monetization aspects.

In this contribution, we propose a marriage between data monetization and BDVC. For this, we present a comprehensive review of the existing BDVCs. After, we discuss data strategies and business models in the context of data monetization. Next, we propose an end-to-end BDVC framework that forms the core platform to take charge of data-driven processes in an agile way. This extended BDVC will support end-to-end data analysis to ensure efficient decision-making. Furthermore, we propose and discuss two possible sharing models, a reduced and a full one, enabling monetization of both data and insights along that BDVC.

The following research questions will guide this contribution:

- Can the BDVC support data monetization?
- How does one achieve end-to-end Big Data monetization?

- What are the different models to ensure data monetization throughout the BDVC?

The rest of this article is structured as follows. In Section 2, we present the adopted research methodology. Section 3 surveys existing work on both Big Data Value Chain and Data Monetization. Then, we highlight the monetization strategies and Big Data business models in Section 4. Section 5 provides a synoptic Big Data Value Chain framework. In Section 6, we propose suitable models to monetize data throughout BDVC. Then, we present and discuss the simulation results in Section 7. The last section concludes the paper and points to some research outlooks.

2. Research Methodology

We have adopted a systematic literature review methodology to identify, select, and synthesize relevant research contributions narratively to address the above questions. For this, we have followed Tranfield et al.'s [15] guidelines to conduct our systematic review.

Several steps were identified in these guidelines, such as planning the study, driving a review, and extracting reports. This literature review helped us to develop a concrete Big Data monetization model adapted to several contexts. To conduct the research, we followed three steps:

- Research, identification, and selection of articles of primary study.
- Filter and evaluate the selected papers.
- Validate papers for data synthesis and analysis of finding.

The selection process for primary studies was conducted according to inclusion and exclusion criteria applied to the examined articles, the scope of the study, and according to verifications carried out by subject and by title and reading the abstract and results. Articles reporting theoretical, empirical, qualitative, or quantitative case studies were also considered.

This research process allows the focusing of the study on articles that are published in prestigious academic journals recognized in Business and Big Data Management and Analytics with a focus on BDVC and Data Monetization. The collected primary studies were carried out from the following sources:

- Google Scholar,
- ScienceDirect,
- Wiley,
- IEEE Xplore digital library,
- ACM (Association for Computing Machinery) digital library,
- Springer,
- Other diverse sources (Taylor and Francis, Emerald, Oxford, and Books).

The research strategy aimed at papers and magazines to cover the period from 2000 to 2020. This allowed us to draw the concepts' evolution. Thus, to ensure a better approach in literature analysis, we used specific keywords such as "Big Data," "Big Data Adoption," "Big Data Value Chain," "Value Chain," "Big Data Management," "Big Data Analytics," "Data Monetization," "Data Sharing," "Business Value," and "Business Management."

In this regard, we have conducted an iterative deep examination to identify the most relevant contributions (Figure 1).

The selection of primary studies identified a large number of articles of the order of 1269. This initial package was refined by a relevant and careful analysis, using cross-referenced and overlapping keyword combinations. A total of 223 articles were selected, followed by a more in-depth analysis. They were then carefully read and reviewed to be relevant to the scope of our study. It is essential to mention that the final datasets are issued only from peer-reviewed and indexed journals. Finally, the descriptive details of the articles were checked and then filed in the Zotero database. The final panel consisted of 97 articles.

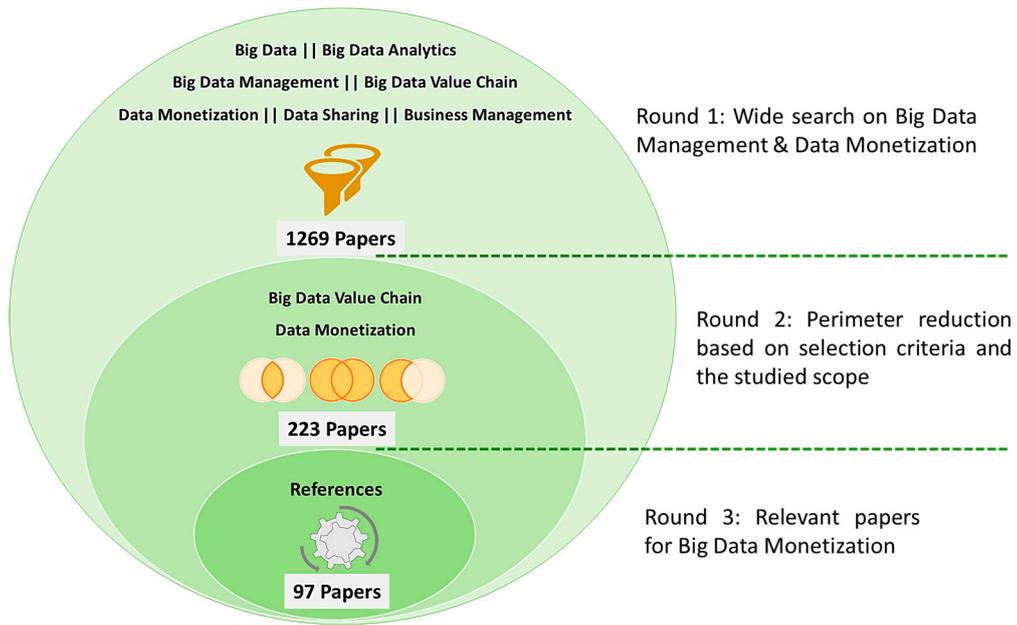


Figure 1. Literature review methodology process.

Figure 2 shows the selected references grouped by the publishing year.

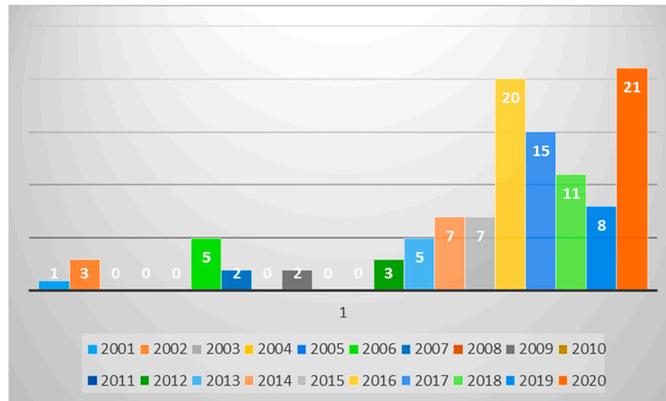


Figure 2. Distribution of references by publishing year.

Data synthesis has shown an essential growth of research in Big Data management over the last few years, which motivates us to study further the relative contributions of the BDVC and Data monetization.

3. Related Work and Background

It is necessary to review the literature to integrate data monetization into BDVC systematically. The literature review has covered the evolution of BDVCs, various existing monetization models, and application fields.

In this section, we present the examined contributions and discuss their definitions and theoretical foundations.

3.1. Big Data Value Chain to Support Data-Driven Operations

The Value Chain (VC), initiated by Porter in the 1980s [16,17], consists of decomposing commercial activities and identifying the interaction between them. It allows the creation of added value by going through the chain activities.

The literature has provided several types of Value Chains such as Knowledge Value Chain [18–20], Value Grid [21,22], Linked Data Value Chain [23], Data Value Chain network [24,25], and Virtual Value Chain [26].

In the 2000s, data digitization led to the emerging concept of Data Value Chains, which replaced the traditional idea of Value Chains for tangible products [25].

The Data Value Chain model is based on a set of activities that consider data as raw materials and a vital resource, enabling the organization to generate value and useful insights. It allows the movement from data to knowledge by going through several steps, such as discovery, processing, and exploitation, to facilitate decision-making [10]. In this regard, the authors in [27] have proposed a framework for Data Value Chain as a Service (DVCaaS) that allows acquisition, integration, management, and data sharing related to business flow.

With the advent of Big Data and the large-scale digitization of information systems, organizations have shown a growing interest in adopting BDVC concepts to define models that engross and support Big Data characteristics. Researchers have shown keen attention to this field and have provided several definitions and models that have enriched the literature. According to Curry, “Big Data Value Chains can describe the information flow within a big data system as a series of steps needed to generate value and useful insights from data” [3]. In addition, the authors in [11] have listed an extended list of BDVC models in various application fields.

In [28], the authors have conceived the BDVC as four successive stages: Generation, acquisition, storage, and data analysis. The authors have discussed this compressed chain of managing data as a dedicated process focusing on data analysis.

In the business management context, Miller and Mork [10] have proposed a BDVC model that allows data discovery, data integration, and data exploitation leading to decision-making. The proposed framework is based on collaborative partnership strategies through data sharing. For the same field, the authors in [3] have made a model that enables value generation and extracted meaningful insights through managing information flow from high-level activities within Big Data ecosystems. This contribution has been adopted to create a new value and bring retail, transport, health, and manufacturing sectors opportunities. Likewise, the authors in [29] have suggested a framework for Big Data reduction at a customer level to achieve value creation. This framework provides a business model aiming at the new data reduction at the customer end, preserving data confidentiality, the cost reduction of using services, and the delegation of data sharing control to customers. Always in the same context, the authors in [30] have presented a seven-step model that uses an interactive Big Data analytical process to help avoid failed business implementations. For this, they have defined a framework for business managers to understand and evaluate the effectiveness of Big Data processes before seeking their applications.

In the education context, the authors in [31] have presented a study that follows a BDVC relying on acquisition, aggregation, linking, processing, analysis, and delivery. The conducted research aims to identify schools in which students have learning disabilities and a risk to leave the education system early. The proposed BDVC allows the retained educational institutions to have more knowledge and a global vision of their students.

In the smart grid context, the authors in [32] have proposed a new approach to segmenting Big Data systems near real-time for energy visualization scenarios. It introduces a framework able to manage massive smart grid data, with a visual data analysis aspect. It also provides a guideline to conduct future research in this field. In the same context, the authors in [33] have proposed a multi-step Big Data grid architecture to manage smart grid requirements, including processing, storage, and visualization. The proposed architecture supervises different flows from the power grid and various energy producers, users, and communicating devices. It provides valuable contributions to the consumption of energy systems and the monitoring of their sustainability.

It is worth noting that digitalization has led organizations improving the employed technologies in supply chain management. According to Korpela et al. [34], digital supply chain integration has

become more dynamic and scalable. The supply chain allows shared access and visibility in tracking product delivery through intelligent logistics in collaboration with partners. It has also permitted end-to-end data product integration through digital enablers, following integrated, standardized, and interoperable processes.

By going over the literature, we noted that BDVCs are becoming increasingly attractive to master data management processes. However, BDVCs form only the skeleton of big data management strategies, and it should be coupled with other equally important aspects such as Big Data monetization. BDVC and Big Data monetization together form extensible and unlimited value chains. But unfortunately, this area suffers from a lack of contributions. Next, we present a few works that have tackled data monetization.

3.2. Data Monetization: Definitions and Theoretical Foundations

Organizations are having increasingly more trends to analyze the various data generated from different sources, look for correlations, and derive market trends. They rely on customer relationship management techniques to exploit customer data and ensure customer loyalty by providing an excellent basis to deduce behaviors. Thus, it allows increased marketing effectiveness, allowing the monetization of data and insights [35].

According to the Gartner glossary, “Data monetization refers to using data for a quantifiable economic benefit. This may include indirect methods, measurable improvements in performance, advantageous terms for trading partners, information about the information (for example, new information-based offers), product information (an existing offer), or outright sale of data (via a data broker or independently)” [36]. The Gartner Institute has already predicted that from 2016, a significant increase in organizations will start monetizing information assets directly or indirectly [37]. They explain that generating a monetary value from data could be from either directly selling or sharing or indirectly creating new products and data-based services.

Likewise, Wixom and Ross consider data monetization as “the act of exchanging information-based products and services for legal tender or something of perceived equivalent value” [38]. In this regard, organizations can adopt three different approaches to monetize their data assets: (i) Improving internal decisions making and operational processes, (ii) guiding and packaging information to frame core products and services, and (iii) selling information that can provide new and existing markets and generate new revenues for the organization. To adopt many of these approaches simultaneously, the organization requires support for specific organizational changes and the alignment of technology and data management to identify the promising approach to ensure integration of other techniques [39].

This would lead to implementing more focused strategies and sustainable business benefits [40]. For this purpose, computational capabilities have enabled enterprises to build an extensive customer database, inform business decisions, and provide data-driven products and services or sell data at different maturity levels [41]. Besides, the IoT explosion has led firms into the Analytics 3.0 era, allowing them to make all decisions based on data and unlock value through data monetization [42].

Several contributions have tackled data monetization in several contexts. In the retailing context, Najjar and Kettinger [43] have given lessons on the monetization of data, deduced from an American retailer’s journey. According to the authors, data monetization can be achieved by improving technical capabilities, enhancing analytical rigor, or simultaneously.

In the healthcare context, Visconti et al. [12] have presented monetization as an output of the BDVC, such as creation, storage, processing, and consumption as the sharing model. This approach is part of a global strategy that includes different actors and several heterogeneous data sources. The processing of these collected data enables us to achieve the simultaneous creation of value that could be monetizable.

In the online social networks context, the authors in [44] have proposed six new routes to monetize value, including affiliate programs, content aggregation, and integrated mobile platforms. They have also focused on improving customer interaction and creating knowledge within an organization to

develop unconventional revenue sources. To monetize personal and user data, the authors in [45] have identified the client's activity or publications to monetize and eliminate off-topic content in these publications via relevant keywords and feedback to infer users' intentions. For the same field, the contribution presented in [46] has proposed a personal data monetization platform via a broker that manages the data as a tangible economic asset and sharing activities as a transaction with suppliers and data consumers.

In [42], the authors have reported that the data economy as an asset has become emerging and could be monetized. The IoT has generated vast data resources leading to business transformation and improved human experience. Thus, the supply chain will be transformed into meshed, complex ecosystems based on interactions of suppliers, retailers, manufacturers, and distributors. The authors have noted that consumers and partners facilitate the creation of interconnected supply chains that allow high collaboration, upstream or downstream, based on effective orchestration.

In [47], researchers have proposed a theory-driven Big Data analysis guidance process to reach data monetization. This theory is based on guidance that relies on a sequence of phases: Data acquisition, preprocessing, analytics, and interpretation, to protect the analysis process from pitfalls.

The authors in [48] have conducted a systematic review to investigate and clarify data monetization models' configuration. They have built the model configuration using a thematic analysis based on the inductive approach. They have also identified four global themes that constitute the main aspects of data, namely, monetization layer, data refinement process layer, base layer, and accessing and processing restrictions layer. They have provided organizations with a practical approach to ensure a data monetization process.

Finally, several data monetization use cases in the industry have been presented in [41]. The authors have projected monetization to telecommunication, banking, tourism, and transportation sectors considered by Franzetti as data monetization industries [49].

Despite the importance of data monetization, the state of the art of BDVC does not consider it an integrated process. Most existing contributions tackle data monetization in old-fashioned ways, completely ignoring the Big Data characteristics and capabilities offered by Big Data analytics. Moreover, they marginalize the need for a data monetization configuration model to be aligned with data strategy. In the following section, we discuss data strategies that present the foundations of data monetization and the business model to be implemented.

4. Big Data Monetization Strategies and Business Models

Big Data monetization strategies allow the realization of economic insights and achievement of more business goals. This section provides various data strategies associated with specific monetization models that are promising pathways to follow.

The idea herein is that each enterprise should consider different kinds of data and make it available by creating processes to monetize these business assets [50]. That way, other businesses could exploit it to reorient their plan and perform transformations based on various approaches.

Data monetization is becoming a core component in business and digital transformations. According to [51], transformation strategies should follow three essential transitions: (i) Process-directed transformation, which consists of standardizing processes; (ii) the IT-driven transformation that uses emerging technologies, such as analytics tools and cloud computing; and (iii) the data-driven transformation that monetizes global data assets by better decision-making. The convergence of these transformations would result in better-targeted strategies and sustainable business benefits [35].

In [50], the authors have listed four data strategies to follow up on business emergence. The data monetization remains a central component in what is projected as strategies, namely:

- The strategy of retaining proprietary data, which can lead to various specific monetization strategies, such as data licensing.
- The trading data strategy with partners for shared benefits considers two monetization strategies: Trading data with business suppliers or downstream business partners.

- The strategy of selling data to many customers under several big data monetization strategies depends on the customer priority premium.
- The strategy to make data open and available to everyone would lead to monetization strategies such as “if you are not paying for it, you are the product.”

It is worth mentioning that several monetization models can be adopted in the same data strategy. Therefore, organizations should rethink the way of dealing with data to extract the maximum benefits.

According to [52], organizations must observe four input components: (i) Business levers, (ii) decision analysis and agile analytics, (iii) competitive and market information, and (iv) monetization guiding principles. Indeed, developing a data monetization strategy starts with building business levers to determine the organization’s strategy impact. The decision analysis will determine monetization needs and requirements. The study of competitive and market information will also reframe the solution, which will define the guiding principles necessary for a monetization strategy development. Thus, Big Data monetization strategies rely on a data strategy. They represent a roadmap to achieve business goals and realize the value and monetization of big data.

Business models for data monetization are deployed in multiple ways to cover and serve the different actors and components, internal and external. An exciting contribution to framing a mining data monetization strategy presented in [53] relates four business models for monetizing data:

- Return of Advantage: Uses its internal performance data triangulated with external data to create a business advantage. Customer targeting, risk mitigation, and fraud detection are examples of this model.
- Premium Service: The data are processed, presented, and delivered to end-user consumption via an access interface offering data products.
- Differentiator: Consists of service or value to the customer with zero or negligible costs to build brand or customer loyalty or develop other services.
- Syndication: Refers to data delivered in a nonraw format to third-party entities.

By correlating the data strategy to its monetization, a generic approach to Big Data business models can be built around four pillars [54,55] (Figure 3).

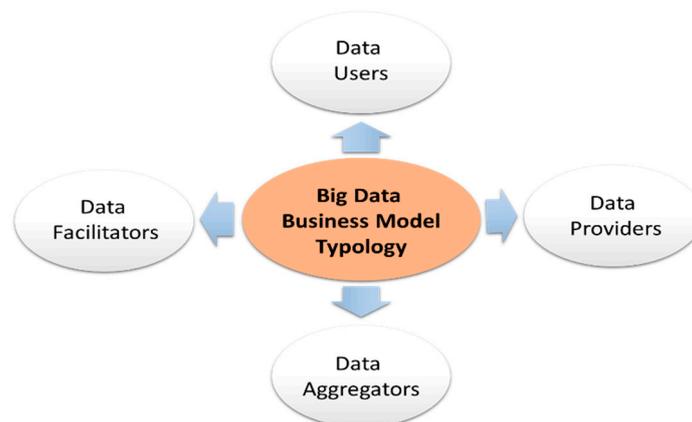


Figure 3. Big Data business models typology.

- Data Users: Refer to businesses that use retrieved data from customers’ internet activities, product usage, behaviors, and preferences to develop strategies and enable data-driven decision-making.
- Data Providers: Refer to data brokerage firms engaged in primary and secondary data collection and sales activities.
- Data Aggregators: Provide customers with aggregated services and data, enabling them to produce a targeted advertising business model.

- Data Facilitators: Correspond to a technical platform based on tools collection, processing, storage, analysis, and data visualization. They will enable businesses to ensure more informed decisions.

Designing a business model for a monetization strategy is based on understanding the organization’s knowledge, mastering its environment and processes, and adopting a data strategy. However, Big Data characteristics are constraining. For business models, they appear in Big Data users that often generate scrambled data with low quality. In addition, data sources are diverse and multiple, making the understanding of data context challenging, especially without homogeneous metadata. Finally, the violation of data privacy remains an ethical and legal issue to solve. Big Data Value Chain can manage data-driven processes and achieve data monetization by building computational and analysis capabilities, sharing information with suppliers, and optimizing costs to overcome these issues.

The BDVC allows, through its various phases, to collect and manage complex and heterogeneous data and subject them to extensive processing following a well-defined data strategy. These data can then be analyzed for visualization to assist decision-makers or to share and expose them to other stakeholders. Thus, conceiving an efficient BDVC offers different forms of value. In the next section, we describe this BDVC framework before going through monetization models.

5. Big Data Value Chain Framework

As explained above, a BDVC is a succession of steps aiming to manage organizations’ data processes. Our generic BDVC framework retains the same basic principle. As illustrated in Figure 4, it relies on several successive phases, namely, data generation, data acquisition, data preprocessing, data storage, data analysis, data visualization, and data exposition.

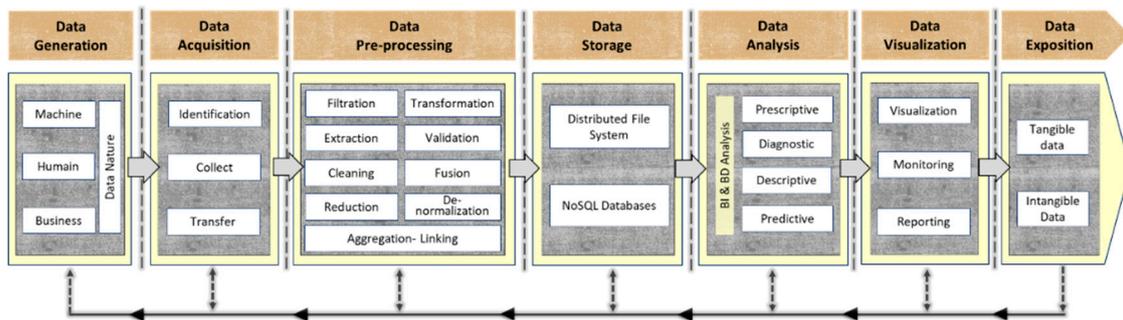


Figure 4. Big Data Value Chain framework.

The proposed BDVC is characterized by a more profound and more refined segmentation of the phases in a Big Data context, with an iterative approach. Each stage consists of various methods and techniques, either grouped or used separately, depending on its strategy. They allow particular processing of the data flow to transform data from a raw state to another form, which is more perceptible, significant, and mature. The transition from one phase to another is ensured by specific interfaces defined at each level, supporting an iterative way if necessary.

This framework remains adaptable to several contexts and supports data generated from different sources as it integrates all phases to address data-driven processes, including data exposition (monetization). Furthermore, our contribution focuses on process-generated data, which will be enhanced and processed along the BDVC to make it entirely data-driven and achieve end-to-end value realization.

In what follows, we detail all phases forming this framework.

- Data Generation: Refers to data that are generated from various sources. It is classified by either the nature of data (structured, semi-structured, and unstructured data) or the source of data (IoT, social media, operational, and commercial data) [56–58]. Moreover, data might be generated by business, machines, or human processes [13]:

- Machine-generated data, coming from equipment and connected objects.
- Human-generated data, coming directly from people through forms, e-mails, research engines, and social media platforms.
- Business-generated data, coming from giant business data aggregators, internal platforms such as data warehouses, government agencies, and public institutions.

This phase's challenge is to correctly identify data generation platforms and the nature of data suitable for the managed use cases.

- Data Acquisition: Refers to the way data can be received and collected. This phase consists of identifying the data flow mode when connecting to generation platforms. This data flow could be:
 - Batch loading mode can be performed on large datasets, grouped in a defined time interval. It is often used for data sources from legacy systems with proof processes or when data streams cannot be technically delivered.
 - Stream loading mode is continuous data inputs. It should perform in real-time or near real-time and has a faster loading rate than the incoming data rate.
 - Micro-batch loading mode allows the dividing of input flows into micro-lots. As a result, the data are obtained in near-real-time.

Once the acquisition mode is identified, and all kinds of connections are established, the acquisition phase should go through three sub-phases:

- Data identification: Refers to determining the content that should be considered.
- Data collected: Refers to identifying preliminary data structures that should be followed to suit the data management strategy.
- Data transfer: Refers to transferring the collected raw data to a specific data storage infrastructure. Most of the time, it is a Data Lake [59].

These sub-phases mainly consist of identifying and collecting data, then transferring them to a storage infrastructure without considering the preprocessing phase's quality aspects.

- Data Pre-processing: The collected data from several heterogeneous sources contain a lot of noise, redundancy, and anomalies, which increases storage space by retaining unneeded data that could affect the data management workflow. Besides, analytical methods require a certain level of data quality [3,60]. For this, data preprocessing is a crucial step to ensure efficient data processing. This phase includes the following sub-phases:
 - Filtration: Refers to eliminating data considered as corrupt according to the organization's data strategy requirements. Several techniques could be applied (e.g., filtration of URLs from web data, low-memory pre-filtration of data streams).
 - Extraction: Refers to reworking incompatible data, often specifically grouped or compressed. This sub-phase allows transforming disparate data into supported formats [61].
 - Transformation: Refers to modifying, adapting, and packaging data into appropriate forms and the scaling standardization of attributes to improve data analytics processes [13].
 - Validation: Refers to establishing validation and deletion rules to manage the syntactic and semantic structures of data and remove invalid and unknown data [62].
 - Cleaning: Refers to identifying and processing incomplete, inaccurate, and unreasonable data to remove or complete it.
 - Fusion: Refers to sensitive data mining processes that consist of merging data from various multimodal sources continuously. It enables us to produce coherent datasets [29,63].

- Reduction: Refers to techniques used to reduce the dimension of essential data extracted with redundancy, noise, and inconsistency, to produce highly relevant datasets for Big Data analysis [64–66].
- Linking: Refers to converting and connecting data using appropriate Linked Data vocabularies and semantic web technologies to provide uniform representation. It reduces Big Data variety for some less critical dimensions, such as using the RDF (Resource Description Framework) representation [65,67].
- Aggregation: Refers to treating together datasets' content belonging to the same field. This aggregation enables us to deal with voluminous data by combining similar and correlated data and eliminating redundancy to produce a sizeable unified view [62].
- Denormalization: Refers to the data modeling process that involves collecting information from multiple tables to form a larger one. It allows optimizing queries' performance and making data-oriented applications [68].

Once the data are preprocessed, they are then transmitted to data centers for storage.

- Data Storage: refers to storing a massive amount of collected and preprocessed data. Storage systems strategies have a significant impact on the scalability and performance of BDVC in terms of data access and exposition. It is based on several aspects, namely:
 - Storage models: Developed mainly around three storage models: Block, File, and Object.
 - Data models: Often follow NoSQL topologies such as key-value, column-oriented, graph-oriented, or document-oriented. This NoSQL view is reasonable for efficient storage, leading to effective processing and, above all, native exposition capabilities.
 - Distributed storage systems: Operate as CA systems (consistent and highly available), CP systems (compatible and partition-tolerant), or AP systems (highly available and partition-tolerant).
- Data Analysis: Refers to manipulating massive data to identify patterns, find correlations, and discover new emerging knowledge models. This phase mainly relies on dedicated Big Data analytics capabilities categorized as descriptive, diagnostic, predictive, or prescriptive [4,57,62,69].
 - Descriptive analysis refers to the description and synthesis of knowledge models using statistical methods that describe a situation, such as standard reports, dashboards, and detailed analysis.
 - The diagnostic analysis identifies causes leading to better performance by reviewing past performances.
 - Predictive analysis refers to the prediction probabilities employed to define future trends. It uses supervised, unsupervised, and semi-supervised learning models to provide predictive analytical models.
 - The prescriptive analysis is applied to predict future events and drive proactive decisions outside human interaction bounds.

These analysis processes combine both statistical and computational approaches, such as:

- Machine learning (ML) belongs to the scope of artificial intelligence (AI). It relies on analytical methods to create predictive models. ML models could be either supervised, semi-supervised, or unsupervised. The most common ML techniques are classification, association analysis, regression, graph analysis, clustering, and decision tree [70–72].
- Deep learning is a set of methods allowing us to create computational models based on nonlinear processing and hierarchical representations. It is used to build classification patterns and learn feature representations from multiple layers of abstraction.

Data analysis output is either a piece of final information that could be represented visually with concrete meaning or another projection of data that requires other data analysis rounds.

- **Data Visualization:** Refers to illustrating data relationships with an artistic visual representation such as graphs, maps, data grids, and alerts, which help rapid and efficient decision-making [73–75]. Big Data Visualization uses suitable tools with extended capabilities that allow business users to find new trends or discover answers to questions not formulated [62].
- **Data Exposition:** Refers to making data available for consumption. This exposition consists of setting up many APIs (application programming interfaces), respecting security and confidentiality policies, and allowing the access to data in different states: Analyzed, preprocessed, transformed, or even as raw as collected. The data exposition generally serves many internal applications, such as CRM (Customer Relationship Management), to promote specific products, but it could be extended to serve partners as well.

The proposed BDVC framework covers most of the required steps needed to accompany and fully exploit data from data generation to data exposition. It is worth mentioning that conceiving the BDVC as a sequential chain is not a realistic approach; we should always have flexibility in going through BDVC phases. Therefore, we propose in our framework a connection between all stages, allowing for example, to re-ingest the output of data exposition to serve as input (data acquisition) to other use cases. This flexibility in the chain of value allows for the reworking of the data as necessary to make it meaningful. Furthermore, each phase in the proposed BDVC has an interface that exposes data at its level, leading to a full Big Data monetization strategy.

6. Big Data Monetization through the BDVC Framework

We propose in this section two big data monetization models integrated into the proposed BDVC. Whereas the monetization at every phase is sometimes very expensive to set-up, the first model is a reduced strategy only to monetize data at some specific stages. The second model is more exhaustive and consists of monetizing every step throughout the BDVC.

6.1. Reduced Big Data Monetization Model

The described model below allows monetizing data at two different states. The first stage consists of those already generated, acquired, preprocessed, and ready for analysis. This first way of monetization enables data sharing (data as goods), allowing other entities or partners to use it for their specific use cases. The second way of monetization refers to sharing insights, which is the output of the whole BDVC. It relies on data that represents readily added values and strategic orientations.

In the proposed BDVC framework, presented in the previous section, data exposition is placed at the value chain's end. However, to support data and insights sharing under the reduced monetization model, the BDVC has to be slightly adapted. Data exposition should be coupled with storage and visualization phases (see Figure 5). This data exposition is ensured through interfaces called "data monetization nodes," which implement all required security rules.

In more details, data exposition under the reduced monetization model respects the following:

- Data could be either raw or specially prepared to serve for extensive analysis processes at the storage phase. This kind of sharing is often useful for the data scientist to run their models or entities with their analysis platforms. The challenge here is to find a balance between confidentiality and exposure. Therefore, an efficient sharing strategy must be implemented by employing suitable tools to control this exposure up to the most exquisite detail.
- At the visualization phase, insights are shared as final components. This kind of sharing, which consists of dashboards, maps, or just some text reflecting some strategic information, is often useful for entities that do not prefer to deal with BDVC phases and prefer to rely on ready-for-use insights.

Many security and access rules are established at both exposition stages to allow authorized entities and partners to access exposed data and insights.

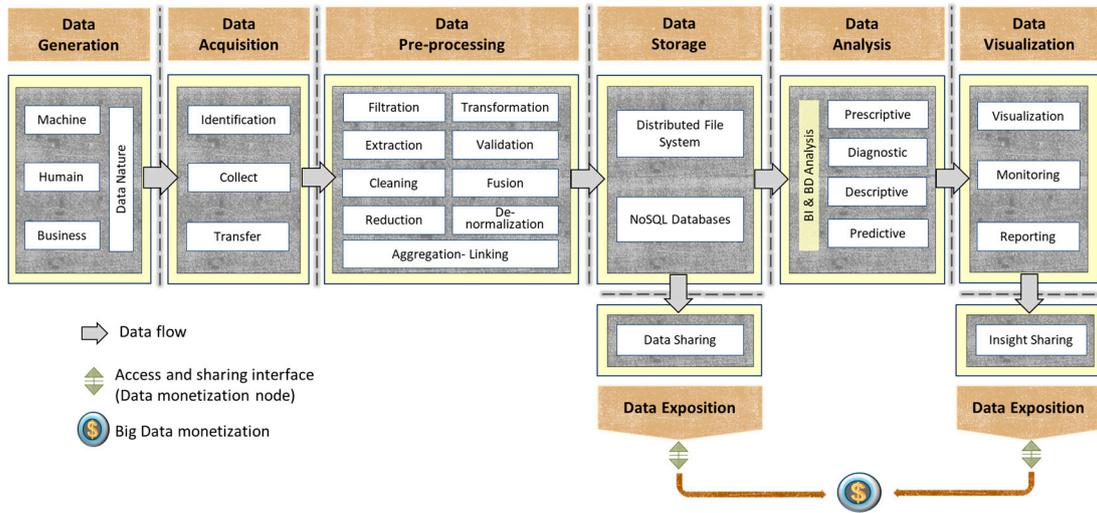


Figure 5. A reduced model to monetizing data via Big Data Value Chain (BDVC).

6.2. Full Big Data Monetization Model

The whole Big Data monetization model is a generalization of the reduced one. As the reduced model consists of monetizing data at storage and visualization phases, the entire model is conceived to support monetization at every part crossed by the data. For this, the proposed BDVC has been slightly adapted to support this way of monetization. Namely, data storage becomes transversal to the acquisition, preprocessing, and analysis (see Figure 6). On the other hand, the analysis output in data storage is considered to come from the fact that it is sometimes necessary to store the analysis output to serve for further analysis or as a monetizable asset.

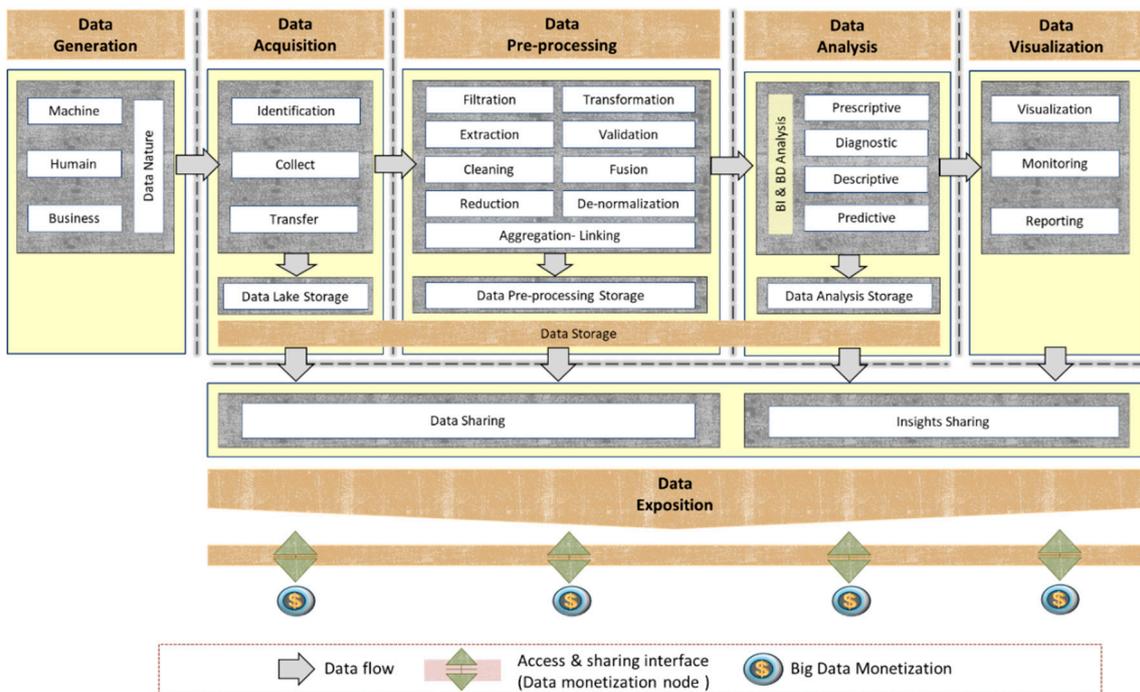


Figure 6. A full model for monetizing data via a BDVC.

The data exposition through this model supports two kinds of monetization, data sharing and insights sharing (see Figure 4). The data sharing is possible via the data storage layer. The insights sharing is possible via either data analysis or data visualization phases. It is worth mentioning that data analysis could serve both data and insights, depending on the data maturity.

To support these different layers of monetization, dedicated storages have been associated with acquisition, preprocessing, and analysis, which is described as follows:

- **Data lake storage:** Contains the data as it is collected. As many use cases require one to rely on the original data to go over their BDVC, this storage allows the sharing of the raw data gathered from several operating systems.
- **Data preprocessing storage:** Contains the filtered, transformed, extracted, validated, cleaned, merged, and reduced data. This kind of sharing allows many entities to gain more time by relying on data respecting some maturity. This phase enables their processes to go faster in implementing their Big Data use cases.
- **Data analysis storage:** Contains the outputs of the analysis models and programs. This kind of sharing is more useful for either use cases that aim at building their visualization components or as input to other BDVCs seeking extensive analysis targets.

We note that there is no storage associated with visualization, as it is continuously updated with volatile content. Thus, the monetization of this layer is actually in real-time. Moreover, data exposition and data storage layers should be efficiently protected as they are the entry points for the core of each organization's business. Data monetization interfaces are also significant bridges connecting many data management strategies.

In term of data value, BDVC can traditionally be segmented into: (i) Value discovery, achieved by Big Data sources and processing; (ii) value creation, reached by Big Data analytics and capabilities; and (iii) value realization, by Big Data analytics and value insights [13]. However, the models studied can reach value realization prematurely compared to the classical BDVC approaches by providing a new way to monetize end-to-end data.

The gradual richness of data storage levels in the BDVC allows for sudden value realization, which demolishes the boundaries of the already fragile DIKW (data, information, knowledge, wisdom) hierarchical model [76].

On the other hand, the proposed models offer alternative and scalable implementations, according to the organization's data strategies and analytical capabilities, which enable customized integration into a co-creative value system. Indeed, the first model allows monetizing data on two main outputs, storage and visualization. This reduces constraints related to data governance, sharing metadata, and data exchange negotiation standards. Moreover, the second model contains several nodes for data production and monetization. In this case, the setting up of a data quality policy and the holding of adequate analytical capacity are required to share processed data, smart and actionable datasets, and analysis patterns. Hence, the second model's chain becomes strongly connected to a value co-creation system through various data monetization nodes.

It is worth mentioning that Big Data monetization is very contextual and mainly depends on the addressed use cases and functional perimeter.

6.3. Value Co-Creation via the BDVC and Cloud Computing

Intrinsic qualities of the Value Chain and the modularity of proposed models allow the BDVC to focus on business modeling and planning. However, it remains tributary on the definition of strategic goals, the framing of the functional perimeter, and the business ecosystem's technological alignment, ensuring suitable conditions for data monetization and value co-creation.

Wixom has defined three business models, such as Data Selling, Bartering, and Wrapping, while Moore has proposed direct and indirect monetization types for customer-related data. Hence, the direct monetization can be mapped to data sales processes, and the indirect monetization can be

assigned to data barter and wrapping information around products. In Figure 7, we show how to project these concepts on the proposed BDVC, according to Moore’s and Waxom’s models [37,38,48].

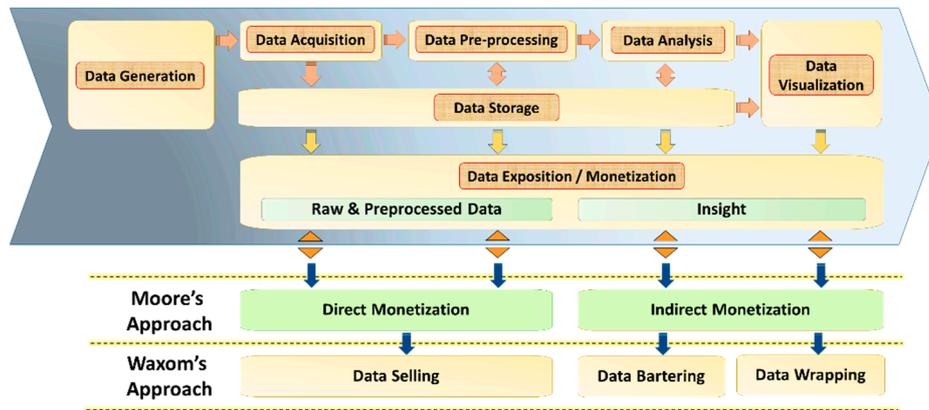


Figure 7. Correspondence of BDVC and data monetization trading types.

The proposed models monetize data in raw or insight form in different ways through various data monetization nodes (shared interfaces and services). Organizations and intermediary players such as aggregators, facilitators, or providers can sell raw goods or provide data barter or wrapper as product- or information-based services.

It is worth noting that the value co-creation results from a mix of data sharing between exchange platforms, data providers, data users, and stakeholders [12], through which it combines and integrates each other’s resources. This extensive collaboration brings out several technological trends, such as cloud computing, to ensure an efficient and coherent mix.

Cloud computing allows BDVCs to handle and process data in a distributed way and share data and insights in a secure mode. In the Big Data monetization context, the cloud computing topology could be implemented as a public, private, hybrid, or community model, see Figure 8.

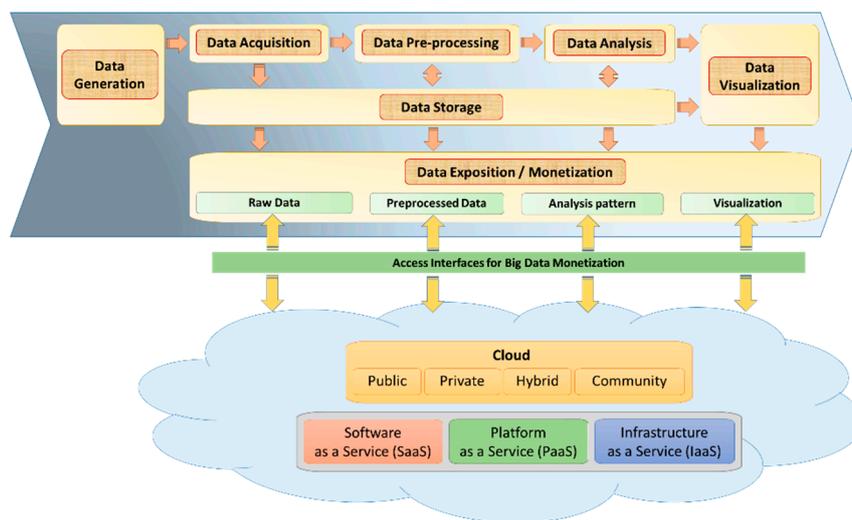


Figure 8. Big Data monetization using cloud computing.

The proposed BDVC provides multiple inputs and outputs interfaces to receive and transmit data according to a certain maturity. This exchange can be provided and automated as deployed services leading to monetization as a service topology at different levels. Thus, the cloud represents a suitable platform to go beyond conventional hosting and classical sharing and propose “X as a Service” (XaaS) for different phases. Namely, Raw data-as-a-Service, Preprocessing data-as-a-Service,

Storage-as-a-Service, Analytics-as-a-Service, Insight-as-a-Service, and Visualization-as-a-Service [77] (see Figure 8).

Monetizing data in such an environment would enable enterprises to quickly develop appropriate professional networks to exchange and share data and resources. Companies can co-create new values within a Big Data ecosystem that no single enterprise can achieve alone [78]. BDVCs can share and exchange data in a cloud network environment via interfaces of monetization using metadata standards and communication protocols such as APIs (application programming interfaces) [79] to provide timely services (see Figure 9).

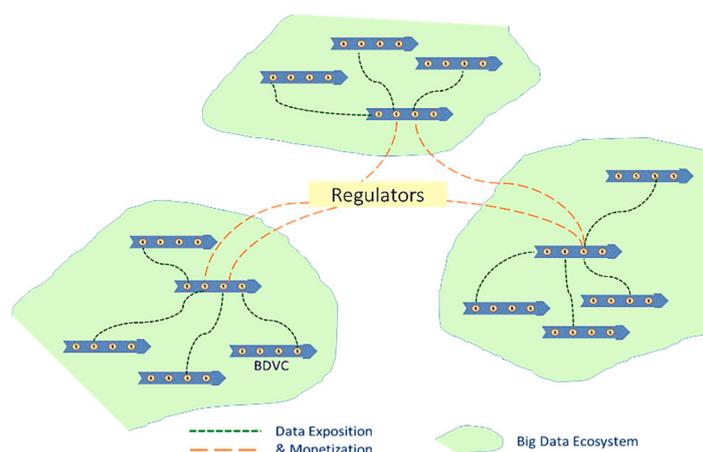


Figure 9. The value co-creation through BDVCs.

However, affiliated organizations need to be aware of this kind of open platform's challenges, such as data discovery, type of storage, linking, business modeling, and synchronization. Regulators deal with legal and data privacy issues, standards, and whether organizations ensure suitable standards [28].

Blockchain is a possible solution that can guarantee reliable transactions by executing a decentralized consensus protocol. It consists of a blocklist containing a cryptographic hash of the previous block, timestamps, and transaction data conducted without any third party. The blockchain is an exchange and decentralization technology, following a Public, Private, and/or Consortium mode, characterized by decentralized, immutable, fraud-free, secure, user anonymity, and auditability aspects. It also uses digital networks in which different stakeholders interact and share data. It could reduce the risk of privacy leakage, mitigate the failure point, and avoid transaction disputes [34,80–82].

However, value co-creation needs to work differently in blockchain business models than in digital business models. The latter rely on pivoting platforms to federate the co-creation of value while blockchains operate in a decentralized way, without dominant players. Thus, it is difficult to evaluate the blockchain because it presents a shareable good by its participants instead of a quantifiable firm asset. This is why the collateral value of the blockchain often depends on its related applications. Thus, it can be valued as an asset owned by a firm or as a value brought to an external user [83].

Thus, Big Data has increasingly brought data to the mainstream of enterprise strategies, determining the business model approach to use, leading to efficient VC implementation. As a core part of BDVC, the Big Data analysis allows data manipulation mainly in data preprocessing and analysis phases, based on artificial intelligence technologies, potentially machine learning, and deep understanding. This leads to innovation in decision-making processes, product delivery, and digital marketing intangibles, which mark a strong discontinuity in business models and evolving extra-returns [84]. This approach significantly impacts the organization's digital know-how, fed by the wealth of knowledge in business data models.

7. Simulation and Evaluation

To validate our approach, we built a prototype for a specific use case on a Big Data platform. This simulation has been conducted to extract several metrics over two scenarios, with and without BDVC. The next subsections describe the guided simulation by showing the evaluation part for different BDVC phases. Furthermore, we provide a detailed discussion regarding the extracted results.

7.1. Use Case Description and Scenario

In our simulation, we rely on a geolocation dataset of a trucking company [85]. It essentially consists of geolocation data collected from sensors that transmit the location, event, speed, and mileage of trucks. It also contains data extracted from a relational database that describes trucks and their characteristics.

It is essential to mention that IoT-based trucking is beneficial for companies. It allows them to interact in real-time with drivers and to anticipate through predicting specific incidents and events. In our case study, the objective is to understand the risks of trucking like driver’s behavior, tiredness, speed, abnormal occurrences, and knowing which cities have drivers with a high-risk factor. We consider the risk factor for a driver as the number of bizarre events recorded over a total of miles.

Due to the absence of a dataset with low data quality in our case, the dataset was intentionally scrambled to stress the work performed by different phases. This dataset will be loaded and injected into a BDVC pipeline to be handled by various steps. Likewise, the same dataset will be processed by a conventional process involving data discovery, processing, and exploitation phases to compare the outputs.

7.2. Adopted Platform and Tools

In this section, we discuss platforms, tools, and used technologies to build our use case. When writing these lines, the Big Data tools’ landscape is highly diverse and diversified. The choice of tools is not exhaustive but selective to fit the case.

Considering the wide variety of Big Data tools and their associated tasks and roles, we propose a set of components based on the Cloudera distribution based on the open-source Apache Hadoop. It covers all functions for building BDVC phases, such as processing, analysis, and monetization, as shown in Figure 10 and explained below.

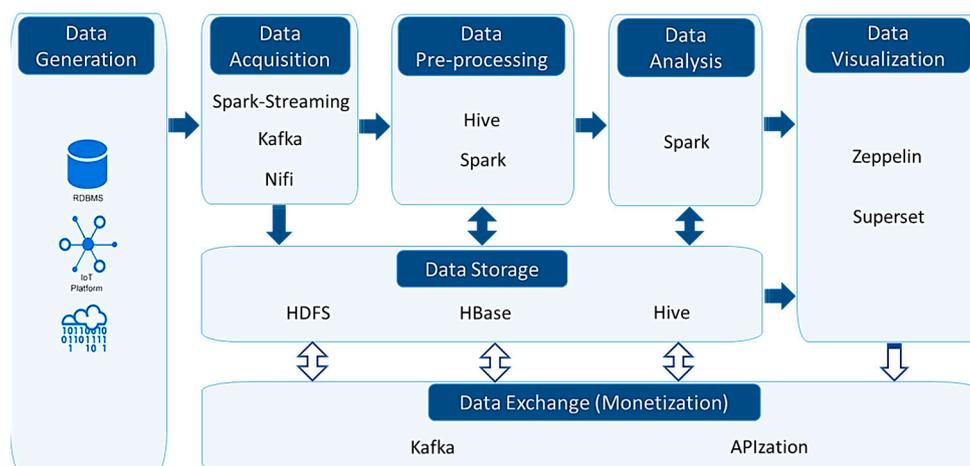


Figure 10. Selected tools for BDVC.

1. Data Acquisition

Several tools are needed to collect data depending on their nature (real-time data streams and data flows).

- Spark-Streaming: A component of the Spark core. It enables high-speed and scalable processing of real-time data from a variety of sources with fault tolerance. Once processed, data are delivered in real-time to databases, file systems, or dashboards. Spark Streaming works internally by dividing input data streams into batches and processes them through the powerful Spark engine [86].
- Kafka: An open-source distributed event messaging platform. It offers high-performance data pipelines, stream analysis, and integration of complex data and applications. It provides three main capabilities to publish and distribute event streams to and from other systems, store event streams sustainably and reliably, and process event streams on-the-fly or retrospectively. Kafka also provides high scalability and message consistency. Kafka is also a message broker that provides a data flow pipeline to Spark-Streaming to be divided into micro-batches for processing [87,88].
- NiFi: An open-source dataflow management system that integrates data streaming and simple event processing. It allows us to automatically inject data streams between different source systems and other systems. Fault-tolerant and scalable, NiFi ensures the entire data stream and authentication and access authorization security via Kerberos [89,90].

2. Data Pre-processing

During this phase, we mainly use Hive and Spark to improve the quality of collected data.

- Hive: A data platform that exploits Hadoop's capacities to offer reasonable possibilities to handle massive preprocessed and post-processed datasets. It is based on its HQL (Hive query language) similar to standard SQL statements for data query and analysis. It is used to summarize Big Data and facilitate querying and data aggregation. It is also considered a language for real-time queries and row-level updates [91].
- Spark: A parallel and unified analysis engine for large-scale data processing, known for its speed, ease of use, and versatility. It provides high-level APIs through several languages such as Java, Scala, and Python. It consists of several components such as Spark SQL to preprocess and process structured data, Structured Streaming [92].

3. Data Storage

This is based on Hadoop/HDFS linked to Hbase and Hive to ensure various processing kinds of storage.

- HDFS (Hadoop distributed file system): A distributed file system with a high level of fault tolerance that stores files as a replicated series of blocks. It is one of the core components of the Apache Hadoop framework. It provides high-speed data access and is suitable for Big Data applications based on distributed processing [93].
- Hbase: Based on the concept and features of Google BigTable as a nonrelational structure (NoSQL). It relies on a family column-oriented concept with key-value-pair data stores [94].
- Hive (warehouse): Besides its processing capability, it is a data warehouse that allows reading, writing, and managing large dataset files stored in HDFS. Hive tables are similar to those in a relational database, are organized from largest to most granular, and are queried using HQL. Simultaneously, storage is more scalable than a relational database, and schema reading is faster. It supports many forms, such as Avro, orc, and parquet [91].

4. Data Analysis

Concerning analysis, we mainly rely on the Spark engine and its rich API.

- Spark: In addition to what was presented, Spark allows the application of machine learning analysis through Spark MLlib and GraphX. The API MLlib provides several functionalities for learning, underlying statistics, optimization, and linear algebra. It supports multiple languages and harnesses the rich Spark ecosystem, and feeds the ML pipeline from end to end. In addition, the GraphX component for Spark is dedicated to model and graph processing [92].

5. Data Visualization

The following tools will carry it out:

- Zeppelin: A tool that provides a web interface as a notebook form to analyze and display visually and interactively a large volume of processed data. It is coupled with various software components, such as Spark. It is based on a set of plugins that make it more flexible [95].
- Superset is a fast and intuitive tool. It allows for simple exploitation and ready-to-use data visualizations by creating shared dashboards. It supports linear graphs and very detailed geospatial maps. It also ensures integration with most SQL-speaking RDBMS (Relational DataBase Management System) [96].

6. Exchange/Data Monetization

It allows data exchange using standardized communication protocols based on APIs.

- Kafka: Previously presented above, it remains the centralized standard platform for data exchange pipelines for data consumption and data production.
- APIzation: Refers to using third-party services and data access interfaces to allow external/internal applications to connect to a resource application to exchange data and outsource services [97].

For stability and integration of these tools, the choice was made on Hortonworks Data Platform version 3.0 [98]. It is a data platform based on the Hadoop ecosystem, which includes HDFS, Pig, Hive, HBase, Spark, Zeppelin, ZooKeeper, and Data Analytics Suite (DAS). This platform is used to manipulate, store, and analyze large amounts of data.

The simulation is carried out over HP Z4 G4 Workstation, having 32 GB of memory and an Intel Xeon 3.60 GHz processor. It runs on a Oracle VM VirtualBox 6.1 virtual machine, on which Hortonworks Data Platform 3.0 is pre-installed.

7.3. Simulation and Results

This subsection aims to experimentally prove the positive impact of adopting BDVC on the performance results. For this, we compare data processing with and without BDVC, using the following set of criteria to assess different aspects:

- Data quality dimensions: To assess the impact on data handling and to measure data relevance and maturity, i.e., credibility, consistency, time penalty, accuracy, and reliability.
- Computational capabilities: To calculate the speed of processing and responses to different tasks.
- Generated insights: To verify the two processes' ability to deliver value, knowledge, and wisdom.
- Data monetization and exchange: To check the ability of both processes to monetize valuable data.

The following table shows our simulation's comparative study between with and without BDVC applied to the geolocation trucking use case.

In addition to the impact of the two scenarios on quality dimensions, insights and monetization, the two scenarios' final results display a value shift, as shown in Figure 11. Catches (a) and (b) reflect values recorded by the BDVC, respectively, to the risk factor per driver and city. They show as a maximum value the driver A39 with a risk factor of 27.4 M and Santa Rosa city with a risk factor of 46.9 M. Catches (c) and (d) correspond to similar values recorded by the classical process to the 2nd scenario. They display as a maximum value the driver A 80 with a risk factor of 0.68 M and Santa Rosa city with a risk factor of 465 M.



Figure 11. Output results of both scenarios with and without BDVC.

The value shift reflects the alteration that may occur in a data processing cycle. The data manipulation following the BDVC life-cycle, potentially the passage through preprocessing and analysis phases, confers high reliability and accuracy of the results and output values.

As shown in Table 1 and Figure 11, the simulation results revealed the positive impact of the BDVC on Big Data processing and monetization. The drawn conclusions converge on the BDVC models' opportunity to implement a Big Data monetization policy at the organization's value chain level.

Indeed, evolving in a digitized environment, the BDVC takes charge of data processes from the beginning, starting with the data sources validation and ending with the generation of strategic indicators to display or share. As these processes evolve according to business rules, they are aligned with data management and quality rules, allowing Value Chains to align these data processes with those of business. These data, which evolve at the same rate as processes, present an opportunity for organizations to monetize them, especially as the Big Data ecosystem architecture promises an integrated deployment for organizations seeking transformation and becoming entirely data-driven.

The theoretical study and simulation have shown that the BDVC allows us to produce and generate data from several points and make them evolve from acquisition to visualization within a Big Data analytics pipeline, allowing the enhancement and valorization of data assets, and achievement of maximum insights. These analytical capacities associated with Big Data ecosystems' computational capabilities allow the BDVC to monetize data from end to end.

Finally, it is prescribed that the deployment of the two proposed monetization models in Section 6 should be used at their accurate scale. The first model has two monetization stations, which are storage and visualization. This model remains moderately simple to implement, while the second model is more complicated. This latter is dedicated to organizations having advanced analytical capabilities deployed in the Cloud. These two models, potentially the second one, offer the possibility to share and monetize the BDVC insight via APIzation interfaces, allowing organizations to evolve within big data ecosystems.

Table 1. Comparative results of the simulation with and without BDVC.

Criteria	Processes	with BDVC						without BDVC				Impact	Conclusions
		Acquisition	Pre-Processing	Storage	Analysis	Visualization	Monetization	Discover	Storage	Processing	Visualization		
Data Quality Dimensions	Credibility	-	91%	-	89%	-	-	-	-	55%	-	Medium	In the first scenario, the dataset was preprocessed by filtering, cleaning, and detecting the missing and outlier values, thus generating more credible insights. Meanwhile, in the second scenario, collected raw data only underwent Null values treatment, leaving the information still impure and sometimes biased. The generated data throughout the BDVC were much more credible than the second case.
	Consistency	-	81%	-	-	-	-	-	-	42%	-	Medium	In the BDVC, we applied several techniques to obtain consistent data such as transformation, reduction, and aggregation operations on collected data using Spark and Hive SQL. These operations allow us to build smart datasets that are preprocessed and directly exploitable, potentially by the analysis phase or monetization. By contrast, in the second scenario, this process was limited to simple SQL operations, with unreliable data.
	Time Penalty	-	-	90%	90%	-	-	-	51%	31%	-	Strong	The access to stored data (i.e., in HDFS, HBase, Hive in our case) in the BDVC is variable. However, as a data warehouse, Hive offers high-speed and granular access to data, far exceeding the performance. In addition, the use of Spark RDD during data analysis in the BDVC allows a fast display of data compared to the second scenario, which uses standard JDBC access.
	Accuracy and Reliability	-	91%	-	89%	-	90%	-	-	45%	35%	Strong	In the BDVC, we prepared and processed the data using different techniques implemented on Spark and Hive, which gave us a reliable and smart dataset. Then, we analyzed them using Spark SQL, which made them more accurate. The processing pipeline by which the data flows into the BDVC allows the ensuring of efficient data processing, while in the second case, raw data are not fully preprocessed, which raises serious questions about the process outputs' accuracy and reliability. In our case, it reduced drastically the accuracy of the results, which made queries on the database inaccurate.

Table 1. Cont.

Criteria	Processes	with BDVC					without BDVC					Impact	Conclusions
		Acquisition	Pre-Processing	Storage	Analysis	Visualization	Monetization	Discover	Storage	Processing	Visualization		
	Computation	80%	80%	90%	90%	90%	90%	25%	50%	40%	45%	Strong	<p>Hortonworks, as a big data platform, offers high computation capabilities when combined with dedicated hardware. Despite the modest test platform, the queries were executed in the milliseconds, and the dataset loading did not exceed one second.</p> <p>Meanwhile, without the BDVC, data loading and the execution of nested or complex analysis are penalized. If not prepared in boxes or views, they can have a polynomial behavior in terms of execution time in the Big Data context.</p>
	Insights	35%	70%	80%	90%	90%	80%	25%	50%	40%	45%	Strong	<p>The simulated BDVC provides multiple valuable outputs. Indeed, the storage contains raw data (in HDFS), which are then filtered well and updated, resulting in preprocessed data (in HBase). Then, they are transformed and aggregated to become intelligent datasets (in Hive), ready for analysis and visualization. It is essential to highlight that this configuration allows the generation of insights into several points through the BDVC. By contrast, without BDVC, we only have access to stored, unreliable data and their visualization by a query. Thus, the use of the BDVC allows the generation of more valuable insight in a different format.</p>
	Exchange/ Monetization	90%	90%	90%	90%	90%	99%	-	40%	-	40%	Strong	<p>As mentioned above, the BDVC generates end-to-end insight under different formats (e.g., raw, preprocessed, smart data, analysis model, web page). These data can be exchanged or shared as needed by other systems using exchange, APIzation, and exchange protocols. The low level of insight without BDVC makes the process limited only to its output and eventually to the storage level.</p>

8. Conclusions and Research Outlooks

Big Data analysis has allowed us to discover hidden insights and define new trends by analyzing and predicting behaviors. With data becoming more valuable than ever, organizations have to adapt all processes dealing with big data specificities to make them more profitable. For that, researchers have assigned the utmost attention to conceive well-defined Big Data management strategies with detailed processes, called Big Data Value Chains (BDVC), aiming to master data in its different states. Another aspect that widely attracted organizations, but a little less the research community, is Big Data monetization. This latter consists of exposing data to entities and partners to form vast and homogenous interconnected data value chains. We believe that coupling BDVC and Big Data monetization would attract much interest in inefficient data management systems. This contribution aims to propose a marriage between these two concepts. For this, above all, we went through a literature review of BDVC models. Then, we proposed a global and generic BDVC framework that supports most of the needed phases, namely: Data generation, data acquisition, data preprocessing, data storage, data analysis, data visualization, and data exposition.

Furthermore, we discussed how this generic BDVC framework could be slightly adapted to host two possible monetization models that expose data and insights. The first model is a reduced form aiming to monetize data only through storage and visualization phases. The second one, which is more generic but expensive to set-up, supports monetization data and insights along the whole BDVC. Both models cover most of the data monetization needs, contribute to co-creating value, and help data monetization models. In future work, we plan to detail other monetization aspects such as security and privacy, projecting both BDVC and its sharing models to real-use cases.

Author Contributions: All the authors have participated in the elaboration of this paper. A.Z.F. carried out the preliminary research and the draft. I.E.A. made the framing of the manuscript. The validation of the content and models and the revisions were performed by I.E.A., Y.G. and A.A. The research orientations were dictated by Y.G. and A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. IDC's 2016 Global IoT Decision Maker Survey Finds Organizations Moving Past Pilot Projects and Toward Scalable Deployments. Available online: <https://www.businesswire.com/news/home/20160921005122/en/IDCs-2016-Global-IoT-Decision-Maker-Survey> (accessed on 14 September 2020).
2. More Than 30 Billion Devices Will Wirelessly Connect to the Internet of Everything in 2020. Available online: <https://www.abiresearch.com/press/more-than-30-billion-devices-will-wirelessly-conne/> (accessed on 28 April 2019).
3. Curry, E. The Big Data Value Chain: Definitions, Concepts, and Theoretical Approaches. In *New Horizons for a Data-Driven Economy*; Cavanillas, J.M., Curry, E., Wahlster, W., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 29–37. ISBN 978-3-319-21568-6.
4. Alaoui, I.E.; Gahi, Y.; Messoussi, R. Full Consideration of Big Data Characteristics in Sentiment Analysis Context. In Proceedings of the 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, China, 12–15 April 2019; pp. 126–130.
5. Moro Visconti, R.; Morea, D. Big Data for the Sustainability of Healthcare Project Financing. *Sustainability* **2019**, *11*, 3748. [[CrossRef](#)]
6. Forgó, N.; Hånold, S.; Schütze, B. The Principle of Purpose Limitation and Big Data. In *New Technology, Big Data and the Law*; Corrales, M., Fenwick, M., Forgó, N., Eds.; Perspectives in Law, Business and Innovation; Springer: Singapore, 2017; pp. 17–42. ISBN 978-981-10-5037-4.
7. Yang, C.; Huang, Q.; Li, Z.; Liu, K.; Hu, F. Big Data and cloud computing: Innovation opportunities and challenges. *Int. J. Digit. Earth* **2017**, *10*, 13–53. [[CrossRef](#)]
8. Liu, J.; Li, J.; Li, W.; Wu, J. Rethinking big data: A review on the data quality and usage issues. *ISPRS J. Photogramm. Remote Sens.* **2016**, *115*, 134–142. [[CrossRef](#)]

9. Vahi, K.; Rynge, M.; Juve, G.; Mayani, R.; Deelman, E. Rethinking data management for big data scientific workflows. In Proceedings of the 2013 IEEE International Conference on Big Data, Silicon Valley, CA, USA, 6–9 October 2013; pp. 27–35.
10. Miller, H.G.; Mork, P. From Data to Decisions: A Value Chain for Big Data. *IT Prof.* **2013**, *15*, 57–59. [CrossRef]
11. Faroukhi, A.Z.; El Alaoui, I.; Gahi, Y.; Amine, A. Big data monetization throughout Big Data Value Chain: A comprehensive review. *J. Big Data* **2020**, *7*. [CrossRef]
12. Moro Visconti, R.; Larocca, A.; Marconi, M. Big Data-Driven Value Chains and Digital Platforms: From Value Co-Creation to Monetization. *SSRN Electron. J.* **2017**. [CrossRef]
13. Saggi, M.K.; Jain, S. A survey towards an integration of big data analytics to big insights for value-creation. *Inf. Process. Manag.* **2018**, *54*, 758–790. [CrossRef]
14. Big Data Led Big Monetization—ProQuest. Available online: <https://search.proquest.com/openview/7cc2f1e5ca16b5f000da83e0e96eeb2d/1?pq-origsite=gscholar&cbl=936333> (accessed on 1 September 2020).
15. Tranfield, D.; Denyer, D.; Smart, P. Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. *Br. J. Manag.* **2003**, *14*, 207–222. [CrossRef]
16. Porter, M.E. Clusters and the new economics of competition. *Harv. Bus. Rev.* **1998**, *76*, 77–90.
17. Micek, G. Competition, Competitive Advantage and Clusters: The Ideas of Michael Porter—Edited by Robert Huggins & Hiro Izushi: BOOK REVIEWS. *Tijdschr. Voor Econ. En Soc. Geogr.* **2012**, *103*, 250–252. [CrossRef]
18. Holsapple, C.W.; Singh, M. The knowledge chain model: Activities for competitiveness. *Expert Syst. Appl.* **2001**, *20*, 77–98. [CrossRef]
19. Carlucci, D.; Schiuma, G. Knowledge asset value spiral: Linking knowledge assets to company's performance. *Knowl. Process Manag.* **2006**, *13*, 35–46. [CrossRef]
20. Chyi Lee, C.; Yang, J. Knowledge value chain. *J. Manag. Dev.* **2000**, *19*, 783–794. [CrossRef]
21. Pil, F.K.; Holweg, M. Evolving from Value Chain to Value Grid. Available online: https://www.researchgate.net/publication/285703652_Evolving_from_value_chain_to_value_grid (accessed on 12 February 2019).
22. Hahn, I.; Kodó, K. Literature Review of the Value Grid Model. *Open Access DiVA* **2017**, *11*. Available online: <http://urn.kb.se/resolve?urn=urn:nbn:se:hh:diva-33421> (accessed on 22 November 2020).
23. Latif, A.; Saeed, A.U.; Hoefler, P.; Stocker, A.; Wagner, C. The Linked Data Value Chain: A Lightweight Model for Business Engineers. In Proceedings of the I-KNOW '09 and I-SEMANTICS '09, Graz, Austria, 2–4 September 2009; pp. 568–575.
24. Peppard, J.; Rylander, A. From Value Chain to Value Network. *Eur. Manag. J.* **2006**, *24*, 128–141. [CrossRef]
25. Attard, J.; Orlandi, F.; Auer, S. Data Value Networks: Enabling a New Data Ecosystem. In Proceedings of the 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Omaha, NE, USA, 13–16 October 2016; pp. 453–456.
26. Bhatt, G.D.; Emdad, A.F. An analysis of the virtual value chain in electronic commerce. *Logist. Inf. Manag.* **2001**, *14*, 78–85. [CrossRef]
27. Kasim, H.; Hung, T.; Li, X. Data Value Chain as a Service Framework: For Enabling Data Handling, Data Security and Data Analysis in the Cloud. In Proceedings of the 2012 IEEE 18th International Conference on Parallel and Distributed Systems, Singapore, 17–19 December 2012; pp. 804–809.
28. Han, H.; Yonggang, W.; Tat-Seng, C.; Xuelong, L. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. *IEEE Access* **2014**, *2*, 652–687. [CrossRef]
29. ur Rehman, M.H.; Chang, V.; Batool, A.; Wah, T.Y. Big data reduction framework for value creation in sustainable enterprises. *Int. J. Inf. Manag.* **2016**, *36*, 917–928. [CrossRef]
30. Rajpurohit, A. Big data for business managers—Bridging the gap between potential and value. In Proceedings of the 2013 IEEE International Conference on Big Data, Silicon Valley, CA, USA, 6–9 October 2013; pp. 29–31.
31. Petrova-Antonova, D.; Georgieva, O.; Ilieva, S. Modelling of Educational Data Following Big Data Value Chain. In Proceedings of the 18th International Conference on Computer Systems and Technologies—CompSysTech'17, Ruse, Bulgaria, 23–24 June 2017; pp. 88–95.
32. Munshi, A.A.; Mohamed, Y.A.-R.I. Big data framework for analytics in smart grids. *Electr. Power Syst. Res.* **2017**, *151*, 369–380. [CrossRef]
33. Daki, H.; El Hannani, A.; Aqqal, A.; Haidine, A.; Dahbi, A. Big Data management in smart grid: Concepts, requirements and implementation. *J. Big Data* **2017**, *4*. [CrossRef]

34. Korpela, K.; Hallikas, J.; Dahlberg, T. Digital Supply Chain Transformation toward Blockchain Integration. In Proceedings of the 50th Hawaii International Conference on System Sciences, Hilton Waikoloa Village, HI, USA, 4–7 January 2017.
35. News Room, TM Forum Gartner: Companies Missing Out on OPPORTUNITIES to Monetize Data. Available online: <https://inform.tmforum.org/news/2015/10/gartner-companies-missing-out-on-opportunities-to-monetize-data/> (accessed on 19 November 2020).
36. Data Monetization—Gartner IT Glossary. Available online: <https://www.gartner.com/it-glossary/data-monetization> (accessed on 28 April 2019).
37. Moore, S. How to Monetize Your Customer Data. Available online: <http://www.gartner.com/smarterwithgartner/how-to-monetize-your-customer-data/> (accessed on 7 September 2020).
38. Cashing In on Your Data. Available online: https://cisr.mit.edu/publication/2014_0801_DataMonetization_Wixom (accessed on 31 August 2020).
39. Wixom, B.H.; Ross, J.W. How to Monetize Your Data. MIT Sloan Management Review. Available online: <https://sloanreview.mit.edu/article/how-to-monetize-your-data/> (accessed on 19 November 2020).
40. Data Monetization Strategies. How to Make Money or Save Money With Data and Analytics. Available online: <https://www.irmconnects.com/white-papers/data-monetization-strategies-how-to-make-money-and-save-money-with-data-and-analytics/> (accessed on 19 November 2020).
41. Liu, C.-H.; Chen, C.-L. A review of data monetization: Strategic use of big data. In Proceedings of the Fifteenth International Conference on Electronic Business (ICEB 2015), Hong Kong, China, 6–10 December 2015; p. 7.
42. Opher, A.; Chou, A.; Onda, A. The Rise of the Data Economy: Driving Value through Internet of Things Data Monetization. *IBM Glob. Serv.* **2016**. Available online: <https://assets.toolbox.com/research/the-rise-of-the-data-economy-driving-value-through-internet-of-things-data-monetization-42098> (accessed on 22 November 2020).
43. Najjar, M.S.; Kettinger, W.J. Data Monetization: Lessons from a Retailer’s Journey. *MIS Q. Exec.* **2013**, *12*, 14.
44. Gomez-Arias, J.T.; Genin, L. Beyond monetization: Creating value through online social networks. *Int. J. Electron. Bus. Manag.* **2009**, *7*, 79–85.
45. Nagarajan, M.; Baid, K.; Sheth, A.; Wang, S. Monetizing User Activity on Social Networks - Challenges and Experiences. In Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, Milan, Italy, 15–18 September 2009; pp. 92–99.
46. Bataineh, A.S.; Mizouni, R.; Barachi, M.E.; Bentahar, J. Monetizing Personal Data: A Two-Sided Market Approach. *Procedia Comput. Sci.* **2016**, *83*, 472–479. [[CrossRef](#)]
47. Elragal, A.; Klischewski, R. Theory-driven or process-driven prediction? Epistemological challenges of big data analytics. *J. Big Data* **2017**, *4*. [[CrossRef](#)]
48. Hanafizadeh, P.; Harati Nik, M.R. Configuration of Data Monetization: A Review of Literature with Thematic Analysis. *Glob. J. Flex. Syst. Manag.* **2020**, *21*, 17–34. [[CrossRef](#)]
49. Franzetti, A. Data Monetization in the Big Data Era: Evidence from the Italian Market. Available online: https://www.academia.edu/34951960/Data_monetization_in_the_big_data_era_evidence_from_the_Italian_market/ (accessed on 19 November 2020).
50. Walker, R. *From Big Data to big Profits: Success with Data and Analytics*; Oxford University Press: New York, NY, USA, 2015; ISBN 978-0-19-937832-6.
51. Berman, S.J. Digital transformation: Opportunities to create new business models. *Strategy Leadersh.* **2012**, *40*, 16–24. [[CrossRef](#)]
52. Wells, A.R.; Chiang, K. *Monetizing Your Data: A Guide to Turning Data into Profit-Driving Strategies and Solutions*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2017; ISBN 978-1-119-35627-1.
53. KPMG Framing a Winning Data Monetization Strategy. Available online: <https://home.kpmg/mu/en/home/insights/2015/10/raming-a-winning-data.html/> (accessed on 19 November 2020).
54. Schroeder, R. Big data business models: Challenges and opportunities. *Cogent Soc. Sci.* **2016**, *2*. [[CrossRef](#)]
55. The Data Monetization | Big Data Business Models. Available online: <https://www.feedough.com/the-data-monetization-big-data-business-models/> (accessed on 13 February 2019).
56. Rathore, M.M.; Paul, A.; Hong, W.-H.; Seo, H.; Awan, I.; Saeed, S. Exploiting IoT and big data analytics: Defining Smart Digital City using real-time urban data. *Sustain. Cities Soc.* **2018**, *40*, 600–610. [[CrossRef](#)]

57. Kibria, M.G.; Nguyen, K.; Villardi, G.P.; Zhao, O.; Ishizu, K.; Kojima, F. Big Data Analytics, Machine Learning, and Artificial Intelligence in Next-Generation Wireless Networks. *IEEE Access* **2018**, *6*, 32328–32338. [[CrossRef](#)]
58. Ge, M.; Bangui, H.; Buhnova, B. Big Data for Internet of Things: A Survey. *Future Gener. Comput. Syst.* **2018**, *87*, 601–614. [[CrossRef](#)]
59. Mehmood, H.; Gilman, E.; Cortes, M.; Kostakos, P.; Byrne, A.; Valta, K.; Tekes, S.; Riekkki, J. Implementing Big Data Lake for Heterogeneous Data Sources. In Proceedings of the 2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW), Macao, China, 8–12 April 2019; pp. 37–44.
60. Azeroual, O. Treatment of Bad Big Data in Research Data Management (RDM) Systems. *Big Data Cogn. Comput.* **2020**, *4*, 29. [[CrossRef](#)]
61. Grzegorowski, M.; Stawicki, S. Window-Based Feature Extraction Framework for Multi-Sensor Data: A Posture Recognition Case Study. In Proceedings of the 2015 Federated Conference on Computer Science and Information Systems (FedCSIS), Lodz, Poland, 13–16 September 2015; pp. 397–405.
62. Erl, T.; Khattak, W.; Buhler, P. Big data fundamentals: Concepts, drivers & techniques. In *The Prentice Hall Service Technology Series from Thomas Erl*, 1st ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2016; ISBN 978-0-13-429121-5.
63. Yan, Z.; Liu, J.; Yang, L.T.; Chawla, N. Big data fusion in Internet of Things. *Inf. Fusion* **2018**, *40*, 32–33. [[CrossRef](#)]
64. Chen, M.; Mao, S.; Liu, Y. Big Data: A Survey. *Mob. Netw. Appl.* **2014**, *19*, 171–209. [[CrossRef](#)]
65. Siddiqua, A.; Hashem, I.A.T.; Yaqoob, I.; Marjani, M.; Shamshirband, S.; Gani, A.; Nasaruddin, F. A survey of big data management: Taxonomy and state-of-the-art. *J. Netw. Comput. Appl.* **2016**, *71*, 151–166. [[CrossRef](#)]
66. Trovati, M.; Bessis, N. An influence assessment method based on co-occurrence for topologically reduced big data sets. *Soft Comput.* **2016**, *20*, 2021–2030. [[CrossRef](#)]
67. Zhang, Y.; Cheung, Y.-M. Discretizing Numerical Attributes in Decision Tree for Big Data Analysis. In Proceedings of the 2014 IEEE International Conference on Data Mining Workshop, Shenzhen, China, 14 December 2014; pp. 1150–1157.
68. Rodrigues, R.A.; Filho, L.A.L.; Gonçalves, G.S.; Mialaret, L.F.S.; da Cunha, A.M.; Dias, L.A.V. Integrating NoSQL, Relational Database, and the Hadoop Ecosystem in an Interdisciplinary Project involving Big Data and Credit Card Transactions. In *Information Technology—New Generations*; Latifi, S., Ed.; Springer International Publishing: Cham, Switzerland, 2018; Volume 558, pp. 443–451. ISBN 978-3-319-54977-4.
69. Hassani, H.; Beneki, C.; Unger, S.; Mazinani, M.T.; Yeganegi, M.R. Text Mining in Big Data Analytics. *Big Data Cogn. Comput.* **2020**, *4*, 1. [[CrossRef](#)]
70. Qiu, J.; Wu, Q.; Ding, G.; Xu, Y.; Feng, S. A survey of machine learning for big data processing. *EURASIP J. Adv. Signal Process.* **2016**, *2016*. [[CrossRef](#)]
71. Zhou, L.; Pan, S.; Wang, J.; Vasilakos, A.V. Machine learning on big data: Opportunities and challenges. *Neurocomputing* **2017**, *237*, 350–361. [[CrossRef](#)]
72. Alfrjani, R.; Osman, T.; Cosma, G. A Hybrid Semantic Knowledgebase-Machine Learning Approach for Opinion Mining. *Data Knowl. Eng.* **2019**, *121*, 88–108. [[CrossRef](#)]
73. Becker, T. Big Data Usage. In *New Horizons for a Data-Driven Economy*; Cavanillas, J.M., Curry, E., Wahlster, W., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 143–165. ISBN 978-3-319-21568-6.
74. Hirve, S.; Pradeep Reddy, C.H. A Survey on Visualization Techniques Used for Big Data Analytics. In *Advances in Computer Communication and Computational Sciences*; Bhatia, S.K., Tiwari, S., Mishra, K.K., Trivedi, M.C., Eds.; Springer: Singapore, 2019; Volume 924, pp. 447–459. ISBN 9789811368608.
75. Miu, M.; Zhang, X.; Dewan, M.; Wang, J. Development of Framework for Aggregation and Visualization of Three-Dimensional (3D) Spatial Data. *Big Data Cogn. Comput.* **2018**, *2*, 9. [[CrossRef](#)]
76. Rowley, J. The wisdom hierarchy: Representations of the DIKW hierarchy. *J. Inf. Sci.* **2007**, *33*, 163–180. [[CrossRef](#)]
77. Sharma, S. Expanded cloud plumes hiding Big Data ecosystem. *Future Gener. Comput. Syst.* **2016**, *59*, 63–92. [[CrossRef](#)]
78. Adner, R. Match your innovation strategy to your innovation ecosystem. *Harv. Bus. Rev.* **2006**, *84*, 98. [[PubMed](#)]
79. Asaithambi, S.P.R.; Venkatraman, R.; Venkatraman, S. MOBDA: Microservice-Oriented Big Data Architecture for Smart City Transport Systems. *Big Data Cogn. Comput.* **2020**, *4*, 17. [[CrossRef](#)]

80. Xie, J.; Yu, F.R.; Huang, T.; Xie, R.; Liu, J.; Liu, Y. A Survey on the Scalability of Blockchain Systems. *IEEE Netw.* **2019**, *33*, 166–173. [CrossRef]
81. Atlam, H.F.; Azad, M.A.; Alzahrani, A.G.; Wills, G. A Review of Blockchain in Internet of Things and AI. *Big Data Cogn. Comput.* **2020**, *4*, 28. [CrossRef]
82. Xie, S.; Zheng, Z.; Chen, W.; Wu, J.; Dai, H.-N.; Imran, M. Blockchain for cloud exchange: A survey. *Comput. Electr. Eng.* **2020**, *81*, 106526. [CrossRef]
83. Moro Visconti, R. Blockchain Valuation: Internet of Value and Smart Transactions. In *The Valuation of Digital Intangibles*; Springer International Publishing: Cham, Switzerland, 2020; pp. 401–422. ISBN 978-3-030-36917-0.
84. Moro Visconti, R. Big Data Valuation. In *The Valuation of Digital Intangibles*; Springer International Publishing: Cham, Switzerland, 2020; pp. 345–360. ISBN 978-3-030-36917-0.
85. Cloudera. Getting Started with HDP Sandbox: Loading Sensor Data into HDFS. Available online: <https://www.cloudera.com/content/dam/www/marketing/tutorials/getting-started-with-hdp-sandbox/assets/datasets/Geolocation.zip> (accessed on 12 November 2020).
86. Spark Streaming—Spark 3.0.1 Documentation. Available online: <https://spark.apache.org/docs/latest/streaming-programming-guide.html> (accessed on 12 November 2020).
87. Apache Kafka. Available online: <https://kafka.apache.org/> (accessed on 12 November 2020).
88. Gurcan, F.; Berigel, M. Real-Time Processing of Big Data Streams: Lifecycle, Tools, Tasks, and Challenges. In Proceedings of the 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, 19–21 October 2018; pp. 1–6.
89. Apache NiFi Overview. Available online: <https://nifi.apache.org/docs/nifi-docs/html/overview.html> (accessed on 12 November 2020).
90. Cloudera. ApacheNiFi: A real-time integrated data logistics and simple event processing platform. Available online: <https://www.cloudera.com/content/www/en-us/products/open-source/apache-hadoop/apache-nifi.html> (accessed on 12 November 2020).
91. What is Apache Hive? | IBM. Available online: <https://www.ibm.com/analytics/hadoop/hive> (accessed on 12 November 2020).
92. Salloum, S.; Dautov, R.; Chen, X.; Peng, P.X.; Huang, J.Z. Big data analytics on Apache Spark. *Int. J. Data Sci. Anal.* **2016**, *1*, 145–164. [CrossRef]
93. HDFS Architecture Guide. Available online: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html (accessed on 12 November 2020).
94. Apache HBase—Apache HBase™ Home. Available online: <https://hbase.apache.org/> (accessed on 12 November 2020).
95. Zeppelin. Available online: <https://zeppelin.apache.org/> (accessed on 12 November 2020).
96. Welcome | Superset. Available online: <https://superset.apache.org/> (accessed on 12 November 2020).
97. Schöne, P. APIzation in the B2B Space: Integration & Infrastructure. API Friends 2017. Available online: <https://apifriends.com/api-management/apization/> (accessed on 19 November 2020).
98. Cloudera. Hortonworks Data Platform (HDP) on Sandbox. Available online: <https://www.cloudera.com/content/www/en-us/downloads/hortonworks-sandbox/hdp.html> (accessed on 12 November 2020).

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).