2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS) | 978-1-7281-6921-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICECOCS50124.2020.9314391

Big Data Value Chain: A Unified Approach for Integrated Data Quality and Security

Abou Zakaria Faroukhi
Laboratoire de Recherche en Sciences de l'Ingenieur
Ibn Tofail University
Kenitra, Morocco
abouzakaria.faroukhi@uit.ac.ma

Imane El Alaoui
Laboratoire des Systèmes de Télécommunications et Ingénierie de la Décision, Ibn Tofail University
Kenitra, Morocco
imane.el.alaoui@uit.ac.ma

Youssef Gahi
Laboratoire de Recherche en Sciences de l'Ingenieur
Ibn Tofail University
Kenitra, Morocco
gahi.youssef@uit.ac.ma

Aouatif Amine
Laboratoire de Recherche en Sciences de l'Ingenieur
Ibn Tofail University
Kenitra, Morocco
amine.aouatif@uit.ac.ma

***Abstract*—Big Data has grown significantly in recent years. This growth has led organizations to adopt Big Data Value Chains (BDVC) as the appropriate framework for unlocking the value to make suitable decisions. Despite its promising opportunities, Big Data raises new concerns such as data quality and security that could radically impact the effectiveness of the BDVC. These two essential aspects have become an urgent need for any Big Data project to provide meaningful datasets and reliable insights. In this contribution, we highlight the importance of considering data quality and security requirements. Then, we propose a coherent, unified framework that extends BDVC with security and quality aspects. Through quality and security reports, the model can self-evaluate and arrange tasks according to orchestration and monitoring process, allowing the BDVC to evolve at the organization pace and to align strategically with its objectives as well as to federate a sustainable ecosystem.**

Keywords—Big Data Value Chain, Data Management, Data Quality, Data Security, Process Integration, Orchestration

I. INTRODUCTION

With the rise of the Internet and the development of new information and communication technologies, such as business applications, social media, mobile devices, and sensors, many data, commonly called Big Data, are generated. These Big Data represent high-volume, high-velocity, and high-variety information assets that require scalable platforms for their processing [1]. These initial Big Data characteristics (Volume, Velocity, Variety) have been extended by additional features (7Vs) [3], namely; Veracity, Variability, Visualization & Value. This complete decomposition helps to better deal with the complex nature of Big Data to realize value, achieve a potential competitive advantage, and improve decision making.

Although Big Data characteristics reveal many opportunities, they expose business models to multiple challenges, such as adapting existing data workloads. In this regard, the traditional use of Value Chains (VC), which consists of analyzing sources, is no more suitable for managing Big Data value creation, becoming more and more digitalized. Moreover, the adoption of developed data processes such as Data Value Chain (DVC), based on data discovery, processing, and exploitation processes, remains limited for assets that are becoming intangible and more data-centric. Thus, this technological revolution has led organizations to rethink the way to discover, create, and realize value by adopting what is called a Big Data Value Chain (BDVC).

The BDVC enables organizations to deal with Big Data-driven processes successfully and to realize significant value. It consists of several phases that embody the Big Data lifecycle, namely, Data acquisition, pre-processing, storage, analysis, and visualization. It also allows organizations to align with Big Data requirements by adopting new data processing and analysis capabilities.

In addition to processing and analysis considerations, another critical aspect that should be considered to improve BDVC's performance is the data quality. Indeed, the 7Vs characteristics bring fundamental challenges of dealing with data retrieved from various and heterogeneous sources (e.g., Social Media, IoT), because these data are often incomplete, noisy, redundant, and inaccurate. Low data quality could hinder analysis and bias decision-making. Therefore, gathered data must be processed using quality measurements throughout the BDVC to extract accurate, reliable, and credible information during the entire data lifecycle.

It is also important to highlight that the security aspect should also be considered when dealing with BDVC. The wide variety of data sources exposes Big Data systems to multiple types of cyber-attacks that could affect the availability of resources and services, making the BDVC more vulnerable and less reliable. Moreover, the interaction of Big Data systems with other platforms requires security policies to define access authorization and perimeters, categorize communications and classify sensitive data, to ensure availability, confidentiality, integrity, and privacy within the BDVC. Therefore, Big Data security must be carefully considered throughout the data lifecycle to build robust, reliable, and sustainable BDVC.

The integration of both quality and security has become of utmost importance. It could be possible by rethinking managing data and adopting a unified framework to orchestrate and integrate these two aspects in the same data management process. This context requires building a novel data management approach, based on nested processes and functions to provide a robust end-to-end data-driven process.

The interest of our contribution is to propose a novel BDVC model that includes both quality and security aspects throughout its life cycle. It is worth to mention that, at best of our knowledge, no papers have treated BDVC from a nested perspective. We are confident that this contribution will provide a unified BDVC framework allowing data processes to be handled in a complete and integrated way by reducing the gap between data analytics requirements, quality, and security.

978-1-7281-6921-7/20/$31.00 ©2020 IEEE

The rest of this paper is organized as follows; the next Section presents contributions that tackle BDVC and Big Data quality and security. Section 3 describes the different domains that interfere in Big Data management scopes such as BDVC, Big Data quality, and security. In Section 4, we propose a unified framework for integrated Big Data-driven process management. Finally, the last section concludes the paper and presents some future works.

II. Related Work

The BDVC and its adaptation to new contexts have taken on increasing importance. This section goes over the most recent researches on BDVC as well as works that tackle Big Data quality and security.

The first research studies conducted on BDVCs have focused on data flow management in the context of business activities, taking into consideration interactions with external ecosystem stakeholders. In fact, in [4], authors have focused their research on achieving optimal combined value from a BDVC basing on a portfolio management approach. They have proposed a model composed of Data Discovery, Integration, and Exploitation. This BDVC enables to manage data completely from its acquisition to decision making by supporting multi-stakeholders and their technologies. In [5], Curry has proposed a BDVC model that consists of several phases, namely Data Acquisition, Analysis, Curation, Storage, and Usage. This model is designed to evolve within the European Big Data Ecosystem and to identify the stakeholders who can take part in it. From a systems engineering approach, researchers have presented in [6] a systematic framework based on a BDVC that is composed of Data generation, Acquisition, Storage, and Analysis phases. Data Visualization is considered by the authors to assist in the analysis phase. Other researchers have dealt with the BDVC by emphasizing Big Data Analytics.

Moreover, Grover et al. have provided in [7] a BDVC framework for value creation. According to the authors, organizations must ensure adequate analysis of data applications and assess the strategic role of Big Data analysis to be aligned with value creation. Mediation and alignment of analytical capacities remain essential to improve targets of strategic value. In [8], authors have designed a unified analysis framework that presents a data monetization solution basing on a set of functions to process and analyze various data on the telecom area. This BDVC allows finding like-minded communities by implementing a set of IBM tools that provide high-performance analytics for multiple data types. Ramannavar and Sidnal [9] have discussed through a survey the importance of advanced big data analytics to create valuable insights for organizations. They have also suggested a BDVC based on Big Data analysis methodologies and composed of Data Integration, Acquisition, Aggregation, Modeling, and Analysis phases, ended by Interpretation, to achieve strategic decision making.

Although the mentioned above contributions have provided attractive BDVC models to improve Big Data-driven process management and decision-making, it is essential to say that other important aspects should be integrated, such as data quality. Processing low data quality through the Value Chain could impact the process's effectiveness. This makes the data quality issue a high priority that must be addressed to achieve practical data analysis and informed decision making. Other researchers have sought to integrate data quality into BDVCs. In [10], the authors have projected Big Data quality dimensions throughout a BDVC to conceive reliable big data systems in social media. In another work [11], they have measured the impact of Big data quality on a sentiment analysis approach by giving simulation results. In [12], Serhani et al. have suggested a hybrid approach for Big Data quality management within a BDVC. Their system is based on the quality evaluation performed on Pre-Big Data, Pre-processing, Post-Big Data, Processing, and Analytics stages. The quality evaluation is based on Big Data quality specification, quality metadata, and quality of service.

On the other hand, authors in [13] have proposed a holistic quality management model for the Big Data life cycle. For this, they have addressed Big Data quality by identifying quality issues and requirements at each phase of the lifecycle. The model is mainly based on the communication of quality assessment reports.

Although the proposed solutions that integrate quality in the BDVC are impressive, what is even more interesting is to consider the security aspect in the BDVC. Effectively, the data security enables to face cyber-attacks, to protect sensitive data, and to preserve the computational infrastructures. In the following, we present some researchers that have attempted to identify data security pillars and to design secure architectures in the context of Big Data.

Authors in [14] have presented a survey on Big Data management, which considers data security an additional aspect of data processing stages. They have also provided a Big Data security management taxonomy based on four pillars: confidentiality, availability, integrity, and privacy. Researchers in [15] have enumerated different cybersecurity frameworks according to various contexts such as promoted action (cyber-strategy motivation), milieu (environment in which the framework can be deployed), and audience (intended organization type). Furthermore, they have defined five cybersecurity pillars: human, organizational, infrastructure, technology, law, and legal pillars. Authors in [16] have proposed an architecture to secure distributed information systems in a Big Data environment. This architecture allows us to detect and model incidents, predict attacks, and select countermeasures basing on the acquisition of information and events. In [17], authors have described a reference architecture (RA) for Big Data systems to meet national security needs. This RA allows addressing specific requirements by using solution models in a similar way to the NIST architecture with a certain abstraction level. It provides a knowledge capture and transfer mechanism for defining Big Data system architectures.

Despite the importance of security and quality aspects, few researchers have addressed these concerns, especially in the BDVC context. We believe that the BDVC is not limited to managing data processes such as acquisition, pre-processing, analysis, etc., but it should take into consideration data quality and security concerns.

In the next Sections, we first present a background of the BDVC, in conjunction with the data quality and security aspects. Then, we propose a unified framework for integrated data quality and security-based process management.

III. Big Data Management Scope

Big Data management is a field that embraces a set of methods, techniques, and tools that mainly allow data pre-

processing, data storage, data processing, data accessibility, and more. It will enable through a BDVC, which embodies a typical data lifecycle for the data management process, unlock and realize value, and then achieve insight. Hence, Big Data management has become a profit multiplier and considerable success for many organizations.

Next, we highlight the different processes of the Big Data Lifecycle, known as Big Data Value Chain.

A. Big Data Lifecycle - BDVC

Big Data Ecosystem may operate around several BDVCs allowing organizations to model and manage their processes to realize their goals. The BDVC proposes, in this regard, a Big Data lifecycle based on a set of steps from Data Generation to Data Visualization and Exposition, see figure 1. In what follows, we present the BDVC briefly. More details are given in [18].

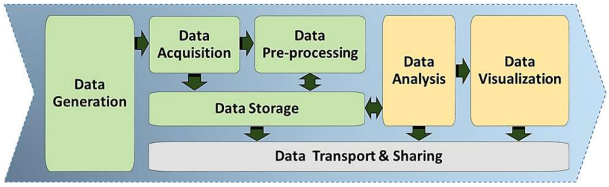

Fig. 1. Big Data Value Chain model

- Data Generation: it is the first phase of the BDVC and refers to the data generated from various sources such as operational databases, mobile device applications, social media, IoT, etc. It can be produced in structured, semi-structured, or unstructured formats.
- Data Acquisition: consists of collecting raw data from diverse sources. Data could be acquired in a different mode such as batch, micro-batch, or stream mode, and further stored in data centers.
- Data Pre-processing: Improving the collected raw data using different techniques (e.g., cleaning, transformation, reduction, integration). It allows us to provide refined and understandable content.
- Data Storage refers to store a vast amount of raw and pre-processed data in a highly distributed storage area with high availability, reliability, and fault tolerance. Different storage models are used as Storage models, Data models, Storage infrastructure, and Distributed processing infrastructure.
- Data Analysis: is a critical phase and consists of analyzing stored data using several techniques such as diagnostic, descriptive, prescriptive, and predictive models. It allows us to extract valuable insights and uncover hidden patterns from data.
- Data Visualization: refers to representing large and complex data in graphical format using dashboards, graphs, and maps. It facilitates the discovery of hidden insights and improves decision making.
- Data Transport and Sharing: concerns the exposition and transport of data, in different formats, to be shared internally or with partners. It allows monetizing data in raw form or refined format as insights.

As mentioned above, Big Data management is a relevant field that should be considered in every Big Data project. It is also important to say that several new challenges related to Big Data stood up to Big Data management, such as data quality and security. Considering these two aspects in BDVCs becomes a significant challenge that could hinder the realization of maximum data value, as discussed in the prior Sections.

In the next subsections, we discuss Big Data quality and security concerns.

B. Data Quality

The data quality has been defined in many different ways, but one of the most accepted is "fitness for use," as mentioned in [19]. It means that the data must be useful for the intended purpose. This being so, ensuring Big Data quality throughout the BDVC requires modeling rules and assessment depending on specific contexts. Notwithstanding, researchers have sought to identify the most eloquent quality measures, called Big Data quality dimensions (BDQD), to conceive trustworthy Big Data systems. These dimensions provide a way to measure and manage data quality items to validate data acceptance for use.

Based on earlier researches [10][13], [20]–[22], we provide our perspective on the BDQD by regrouping them into various aspects, as shown in figure 2.

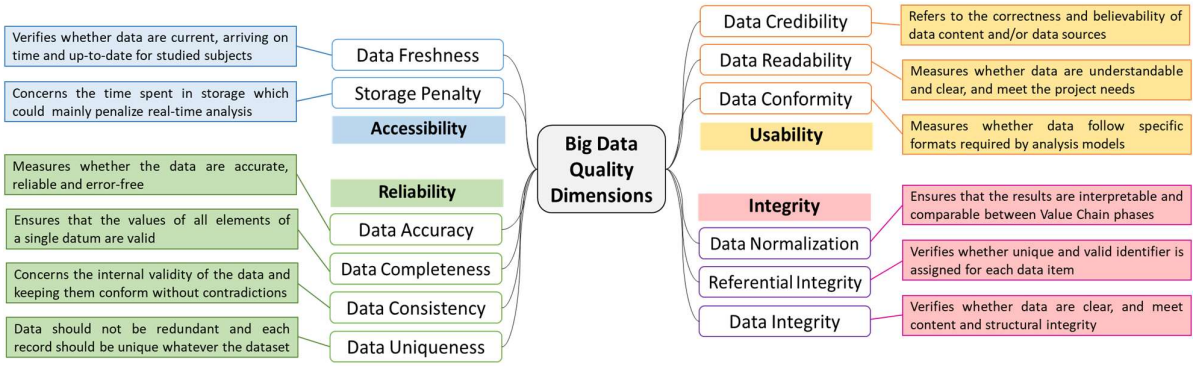

Fig. 2. Big Data Quality Dimensions

These BDQD allow us to monitor and evaluate data and facilitate implementing the data quality management policy in Big Data projects. But, what is even more interesting is to integrate BDQD all along with BDVC processing phases. It

would enable creating and delivering high-quality, reliable, and actionable information throughout every step in the BDVC.

Data quality is a required field that profoundly impacts the effectiveness of the Value Chain. It is also important to mention that quality is not a unique concern to consider in the BDVC. There is also the security, a crucial aspect that preserves data produced through the BDCV phases and ensures its protection and sustainability.

C. *Data Security*

The Big Data lifecycle is based on several stages, and each one can present threats and security risks. The Big Data ecosystem is vast. It integrates several partnerships and feeds on various data sources along with BDVCs phases. It is, therefore, essential to adopt a reference foundation that meets the security requirements for the BDVC. This means that security measurements must be sufficiently inclusive, flexible, and reflective. This leads to consider several aspects of security in Big Data management context [14], [23] that are:

- Availability: refers to the readiness of data, resource, or system, and the guarantee for a reliable access.
- Reliability: refers to the system's ability to be resilient face to various failures, incidents of cyber-attacks.
- Integrity: consists of ensuring that the data or services are not altered or modified in any way.
- Confidentiality: consists of exposing data or information to persons or processes able to handle them or know their content.
- Non-repudiation: consists of ensuring that a transaction can not be denied or disclaimed later.
- Privacy: This allows us to ensure that data handling will not lead to a privacy violation and to anonymize data.

Several security functions allow us to ensure the mentioned above aspects to acquire data asset immunity and ensure a sustainable Big Data ecosystem. These functions will be considered and integrated into the BDVC in the following Section.

To the best of our knowledge, although the importance of data security and quality in the BDVC, the coupling and orchestration of these two aspects in the same Value Chain is not yet treated in a meaningful way. In the next section, we propose a unified framework that promotes the integration and orchestration of data quality and security in the BDVC.

IV. A PROPOSED UNIFIED FRAMEWORK FOR INTEGRATED BIG DATA-DRIVEN PROCESS MANAGEMENT

In this section, we first describe a unified framework for mutualized and integrated data requirements along the BDVC, allowing the analytical aspect to interfere with quality and security.

As mentioned above, the evaluation of Big Data quality is contextual, and hence, a Big Data Quality Dimension could be interpreted differently in the different phases of the Value Chain.

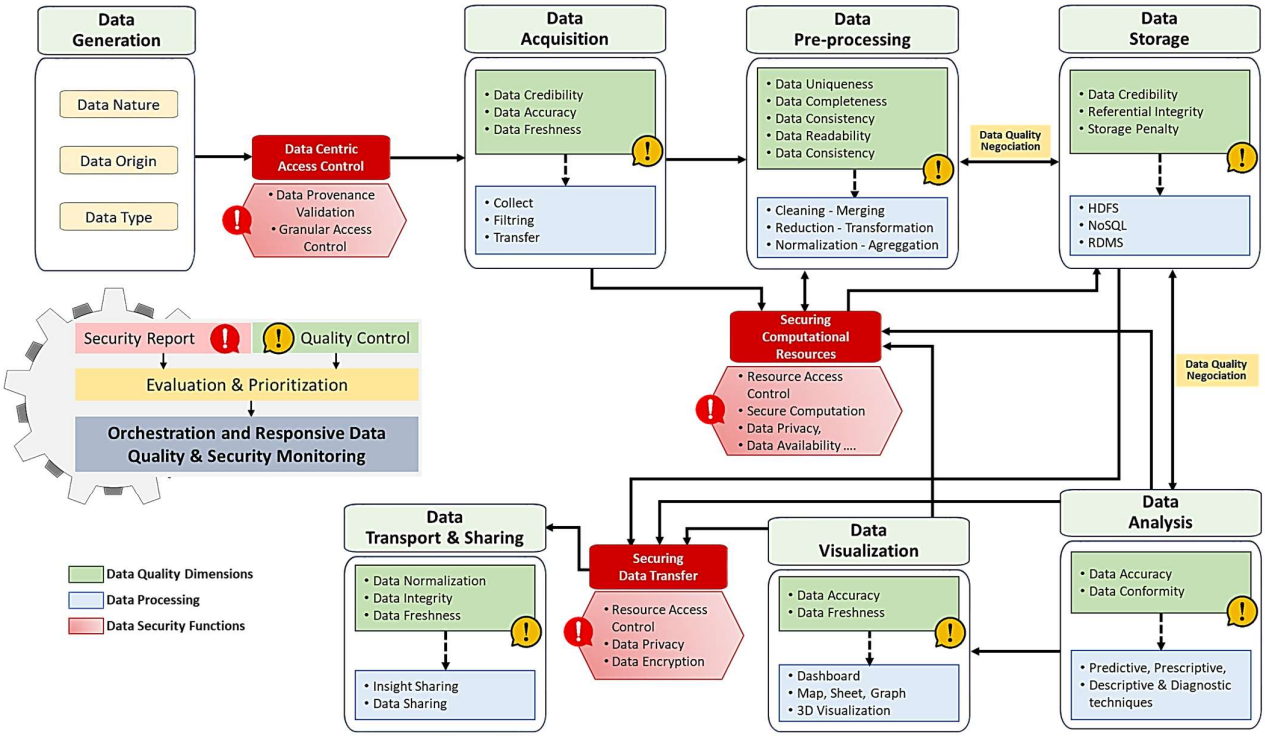

Fig. 3. A comprehensive, unified framework for integrated Big Data-driven process management

In Table 1, we explain the interpretation of BDQD in each stage of the BDVC, the security functions, the orchestration, and the agility processes of the proposed framework (shown in figure 3).

TABLE I. DESCRIPTION OF FRAMEWORK COMPONENTS

Issues	Levels	Requirements	Descriptions
	Data Acquisition	Acquisition Credibility	The acquired data come from various and heterogeneous sources in the context of big data. Data are gathered from internal and external sources such as social media, where users provide personal, subjective, and unreferenced contents. For this reason, it is necessary to measure the source and content credibility.
		Acquisition Accuracy	Organizations tend to collect a large volume of data that is not necessarily representative and relevant to the studied subject. Usually, the more accurate collect is, the more reliable is the insight we can extract. Therefore, ensuring data representativeness, reliability, correctness, and error-free is of utmost importance.
		Acquisition Freshness	As mentioned above, variability is one of the challenging characteristics of big data that refers to data whose meaning is constantly changing. One current case is social media, where users can change their posts' content regularly. However, working on outdated data will produce useless insights. Therefore, it is essential to measure the collected data currency for a given context.
Data Quality	Data Pre-processing	Uniqueness	As mentioned above, data are collected from various sources, such as social media and smartphone applications. This makes data duplication highly probable. Keeping these duplicated data will increase storage space, make data processing time-consuming, and sometimes biased. Thus, it is imperative to provide unique records.
		Completeness	The gathered data may have missing values due to network failure or from the data source itself. In health care, clinical diagnostic support systems can provide incorrect health recommendations based on missing data. It is, therefore, vital to ensure that all values of a datum must be valid and consistent for further processing.
		Consistency	In Big Data systems, data is usually aggregated from multiple sources and stored in different spaces. Sometimes, data that are supposed to be similar or correlated turn out to have changed with inconsistent meanings, leading to unreliable and non-synchronized data sets. Hence, the interest in measuring the purpose of data and checking data conflicts and contradictions.
		Readability	Collected data (e.g., text files and tweets or comments) do not often follow a schema, and their content is not always understandable. These data are difficult to handle because they are not clear and regularly do not meet the intended use. Data Readability allows us to ensure that the data are well-formatted, understandable, meet needs, and satisfy specifications.
		Conformity	As data come from various sources, they are provided with their structures. Nonetheless, non-conform data are not valid and present integration constraints. Thus, it is crucial to guarantee compliance with the defined formats and data validation to meet the analysis models requirements.
	Data Storage	Referential Integrity	The construction of datasets is based on several similar and distributed data sources where different referential integrity mechanisms are applied. Merging and integrating these data can reveal referential integrity violations, hence the need to assign a unique and valid identifier to a data item.
		Storage Penalty	As velocity is one of the fundamental characteristics of big data, the processing time is of utmost importance, especially for projects where real-time analysis is crucial. One current example is crime prediction and prevention that should provide continuous analysis with up-to-the-minute insight. Thus, the read/write access time of the data storage disks must be optimal and selective with priority management. Although storage time lost does not affect the accuracy of results, it penalizes the quality of real-time outcomes and leads to a delay in decision-making.
		Credibility	It concerns the pre-processed storage. As mentioned above, data undergo several transformations, corrections, integration in the pre-processing phase to provide smart and actionable datasets that meet the analysis and business requirements. However, these transformations techniques could impact data credibility by eliminating and modifying useful data. Therefore, it is essential to measure the believability and significance of data in a given context.
	Data Analysis	Analysis Accuracy	Although the used datasets may be reliable, imprecise analytical methods would lead to biased results. To get a valid result, it necessary to measure the degree and reliability of analysis techniques and the correctness of interpretations.
		Analysis Conformity	Once the data analysis phase is completed, the data output will be transferred to the next stage or shared with partners. The result should be presented and respect standard formats for further processing. This metric enables us to check the data compliance with the desired standard formats.
	Data Visualization	Visualization Accuracy	High-quality data visualization enables us to inform decision making. It should reflect the actual state of the analyzed data and provide a reliable value overview. Visualization Accuracy is used to check the knowledge quality and measure its meaningfulness and usefulness.
		Visualization Freshness	Data visualization, especially for real-time data monitoring purposes, often requires a dynamic response in near real-time. Feedback during the visualization process allows data to evolve as updates occur, allowing new knowledge patterns to be discovered. Wherefore, it is essential to measure data representation and visualization freshness.
	Data Transport	Integrity	Data can be corrupted during transfers, especially in distributed environments. Transferred data can be altered or distorted due to network losses, the use of inappropriate media or interface, or malicious acts. Sharing such data could mislead all BDVC processing phases. So, it is crucial to ensure that data is transferred correctly and remain intact and reliable.

		Normalization	Data are produced in different formats according to their appropriate schemas or metadata, not necessarily adapted to input interfaces ones. Standardization becomes necessary primarily to ensure exchanges with external stakeholders outside the BDVC or between completely separated internal processes. For this reason, the metadata reconciliation is required to feed the Value Chain with standardized and normalized data. Hence, the interest in using normalization techniques within data transport and sharing.
		Freshness	As mentioned above, data are shared within a Big Data ecosystem via the network in a distributed environment. This may expose data to constraints of tardiness or bottleneck, hence, forcing data to arrive with a specific latency, often penalizing. Thus, measuring and tracking transfer latency by taking into account network time lost, bottlenecks, and exchange gateways during data transfer is of utmost importance.
Data Security	Data-Centric Access Control	Data Provenance Validation	Usually, BDVCs evolve in open ecosystems that interact with other stakeholders. They receive data from various and heterogeneous, public, private, or free sources. Besides, data could sometimes be collected by non-trust platforms. For these reasons, the access gates of the BDVC must be strictly controlled to avoid any cyber-attacks during the acquisition phase. The data source validation function analyses the data source, verifies security establish metadata dependencies to ensure secure data collection. It also allows us to detect any change or variation in the data nature, type, or scalability during data generation (i.e., switching from structured to unstructured data, and from batch flow to data streaming, increasing volume).
		Granular Access Control	The Big Data system perimeter is too varied, which is very difficult to manage and control access to its resources. To limit access, to hold where, when, and how it is accessed, Granular Access Control has proved to be a security posture and well suited to Big Data contexts. It provides authorization based on data identifiers and data content. Also, some policies can be adopted that describe access control rules, depending on project requirements. Granular Access Control checks whether the environment is trusted, thereby improving the BDVC operational efficiency.
		Resource Access Control	This function is part of the Granular Access Control rules. It applies to the acquisition, pre-processing, storage, analysis, and visualization phases, which require computational capabilities to ensure data processing and analysis. Resources Access Control function implements stringent authentication/identification mechanisms to federate access to remote resources following profile management policy, adapted to have access to exclusive resources. This would enable access control, traceability, and optimal management of allocated resources.
	Securing Computational Resources	Secure Computation	Big data acquisition, pre-processing, storage, and analysis are performed in a distributed programming environment that must be secure and safeguarded. It is considered the cornerstone of the Big data system. Secure Computation is based on duplicated storage to ensure the constant availability of the system that includes thousands of nodes. In fact, in case of an event, server, node, or disk failure, the task performed by the faulty component can be immediately taken over by another element. Also, parallel computing systems allow performing several processing tasks simultaneously on the same data.
		Data Privacy	Big Data processing currently enables to extract of personal data of individuals, closely related to their private lives, such as their preferences, choices, movements, trends, purchases, future schedules, etc. Failure to regulate data processing and disclosure could lead to a violation of users' privacy. So, the data privacy function enables data processing according to specific profiles and topics and policy rules. It uses methods of differential confidentiality, anonymization, and privacy-preservation so as not to compromise privacy.
		Data Availability	Big Data storage is the backbone of BDVC. Its unavailability would inevitably paralyze the whole Value Chain. Big data storage is generally based on replicated and distributed systems to ensure data availability and fault tolerance. So, Data Availability aims to ensure that data is accessible to authorized users while providing prevention, replication, and recovery systems against human and software errors, and malicious access.
	Securing Data Transfer	Resource Access Control	This function allows us to manage the data sharing (tangible or intangible) exposed by the storage, analysis, and visualization phases and to regulate the resources access. It also depends on the Granular Access Control rules and provides profile-based access rights to manipulate and share exclusive resources with specific settings. Media sharing access rights should be regulated according to the transfer rate, capacity, time, and destination.
		Data Privacy	The BDVC can achieve insight based on its value processes or only on data assets. Information could be shared as raw data, analyzed data, or visualized data as a service with stakeholders while respecting privacy and confidentiality regulations. Thus, securing data transfer must ensure a secure sharing of sensitive data and prohibits any unauthorized changes or alterations during data transfers to stakeholders.
		Data Encryption	The data transfer in a networking environment may expose data to cyber-attacks. To ensure data confidentiality, the Data Encryption mechanism transforms precise data into indecipherable and unreadable data. It is used to secure end-to-end communication and to protect data transfer by securing the transmission itself. However, it should be moderately used to avoid the latency inherent in the compression/decompression and encryption/decryption of data.
Orchestration	Orchestration & Agility	Data Quality Control & Data Security Report	This process occurs all along the Value Chain. It is continually listening to quality and security notifications to readjust any shortcomings and respond to any incident. Data Quality Control is used to verify compliance with rules and to pass feedback between phases or to forward them centrally. Data Security Report receives notifications and provides the appropriate measures.
		Evaluation & Prioritization	Feedback, interactive and iterative negotiations between components allow providing a self-assessment and dynamic readjustment of data quality requirements. The security report evaluation also enables us to identify events and react to incidents basing on the continuous discovery of vulnerabilities and analysis of cyber-attacks. This leads to implement appropriate counter-measures and to anticipate future cyber-attacks. In case of incompatibility or non-convergence of data quality and security requirements, negotiation of priorities is triggered to look for alternative solutions (e.g.; traditional encryption is not compatible with data deduplication).

It is essential to mention that the proposed framework is perceived according to three processes, which are integrated into different phases of the BDVC, dealing mutually with end-to-end Big Data flows. These processes are explained below:

- Data Quality process: It is based on procedures, and quality rules declined in metrics that respond to a strategic objective and follow business requirements. Each phase uses assessment indicators for each data quality metric (e.g., accuracy, duplication, completeness, credibility, consistency) according to business rules based on context-awareness. For example, in the ingestion phases, gathered data are tagged to identify its type, nature, and origin to validate the quality of its source. The processing of a low-quality data source leads to a significant loss of time for Value Chain and takes up useless storage space. Besides, each data quality rule applied corresponds to an appropriate technique such as data cleaning rules (e.g., data exclusion, correction, value insertion) and data integration rules (e.g., metadata and version reconciliation). The data analysis, visualization, and transport/sharing phases contribute mainly to data quality through interactive negotiation and feedback. The reliability of results reflects the relevance of the used algorithms and especially the quality and accuracy of user data sets. Therefore, the framework allows for data quality measurement during transport and data sharing to meet business needs.

- Data Security process: Data security challenges are higher than those faced by traditional systems and need more policies and regulations to achieve the expected requirements. Therefore, this process is deployed throughout the value chain to ensure data protection, traceability, and tracking from generation to delivery. It is structured around multiple functions. The Data-Centric Access Control function provides a validation and filtering checkpoint to identity profile and data sources signature during the data acquisition phase to secure entry doors to the BDVC. Also, Securing Computational Resources provides safe access to resources through identification and authentication mechanisms. Securing computation in a distributed programming environment takes into consideration data storage and transaction logs. Mechanisms such as policy-based encryption, periodic auditing of performances, and traceability logs, ensure data availability, consistency, confidentiality, and integrity. Also, the Data Transport and Sharing in various forms require access control and compression and encryption systems. Finally, Data Privacy is maintained throughout the value chain through access control and scalable privacy-protective analytics.

- Orchestration process: It is based on information gathered throughout the value chain, especially security reports and quality controls. This agility allows a high degree of flexibility in supervising all phases and following all negotiations to pass the bridge to a self-assessment of data quality and security, and prioritization of rules and requirements. Also, the interconnectivity between components allows sharing and analysis of the quality and security log information, thus offering a high level of integration between different aspects leading to achieve a high-level of insight, see figure 4.

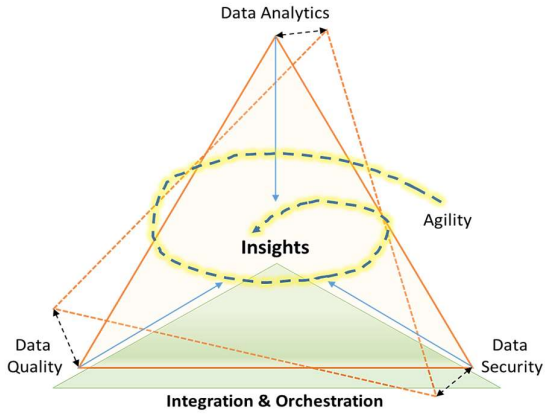

Fig. 4. Agility and orchestration of processes

Responsive data quality and security monitoring enable to design of an integrated global vision and response to needs and incidents. The proposed framework presents a unified approach that offers analytics, quality, and security processes integration under the same data umbrella.

V. CONCLUSION

The Big Data Value Chain presents a suitable framework to support significant process data-driven. However, relying solely on the Big Data Analytics side to unlock value and create insights might fail and expose the organization's ecosystem to unpredictable failures. There are limited initiatives of research that have attempted to broaden the BDVC scope to become more efficient and sustainable. Therefore, in this contribution, we have tackled a holistic and integrated approach to supporting data process management. We provide a unified framework that intrinsically incorporates the management quality and security in the BDVC. It enables end-to-end analysis of Big Data while integrating data quality and security requirements, optimized by an orchestration mechanism through responsive monitoring based on security reports and quality controls.

Implementing this integrated approach brings Big Data Analysis (BDA) to the forefront, combining efficiency, reliability, and agility with building a Big Data-driven Value Chain and enabling the organization to evolve and align with its ecosystem.

The future work will focus on the BDVC implementation, including data quality and security aspects, and using open-sources BDA tools to provide simulation results.

REFERENCE

[1] D. Laney, “3D Data Management: Controlling Data Volume, Velocity, and Variety | BibSonomy,” Feb. 06, 2001.

[2] M. K. Saggi and S. Jain, “A survey towards an integration of big data analytics to big insights for value-creation,” Information Processing & Management, vol. 54, no. 5, pp. 758–790, Sep. 2018.

[3] I. E. Alaoui, Y. Gahi, and R. Messoussi, “Full Consideration of Big Data Characteristics in Sentiment Analysis Context,” in 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, China, Apr. 2019, pp. 126–130.

[4] H. G. Miller and P. Mork, “From Data to Decisions: A Value Chain for Big Data,” IT Professional, vol. 15, no. 1, pp. 57–59, Jan. 2013.

[5] E. Curry, “The Big Data Value Chain: Definitions, Concepts, and Theoretical Approaches,” in New Horizons for a Data-Driven Economy, J. M. Cavanillas, E. Curry, and W. Wahlster, Eds. Cham: Springer International Publishing, 2016, pp. 29–37.

[6] Han Hu, Yonggang Wen, Tat-Seng Chua, and Xuelong Li, “Toward Scalable Systems for Big Data Analytics: A Technology Tutorial,” IEEE Access, vol. 2, pp. 652–687, 2014.

[7] V. Grover, R. H. L. Chiang, T.-P. Liang, and D. Zhang, “Creating Strategic Business Value from Big Data Analytics: A Research Framework,” Journal of Management Information Systems, vol. 35, no. 2, pp. 388–423, Apr. 2018.

[8] H. Cao et al., “SoLoMo analytics for telco Big Data monetization,” IBM Journal of Research and Development, vol. 58, no. 5/6, p. 9:1-9:13, Sep. 2014.

[9] M. Ramannavar and N. S. Sidnal, “Big Data and Analytics—A Journey Through Basic Concepts to Research Issues,” in Proceedings of the International Conference on Soft Computing Systems, vol. 398, L. P. Suresh and B. K. Panigrahi, Eds. New Delhi: Springer India, 2016, pp. 291–306.

[10] I. El Alaoui, Y. Gahi, and R. Messoussi, “Big Data Quality Metrics for Sentiment Analysis Approaches,” in Proceedings of the 2019 International Conference on Big Data Engineering (BDE 2019) - BDE 2019, Hong Kong, Hong Kong, 2019, pp. 36–43.

[11] I. E. Alaoui and Y. Gahi, “The Impact of Big Data Quality on Sentiment Analysis Approaches,” Procedia Computer Science, vol. 160, pp. 803–810, 2019.

[12] M. A. Serhani, H. T. El Kassabi, I. Taleb, and A. Nujum, “An Hybrid Approach to Quality Evaluation across Big Data Value Chain,” in 2016 IEEE International Congress on Big Data (BigData Congress), San Francisco, CA, Jun. 2016, pp. 418–425.

[13] I. Taleb, M. A. Serhani, and R. Dssouli, “Big Data Quality: A Survey,” in 2018 IEEE International Congress on Big Data (BigData Congress), San Francisco, CA, USA, Jul. 2018, pp. 166–173.

[14] A. Siddiqa et al., “A survey of big data management: Taxonomy and state-of-the-art,” Journal of Network and Computer Applications, vol. 71, pp. 151–166, Aug. 2016.

[15] R. Azmi, W. Tibben, and K. T. Win, “Review of cybersecurity frameworks: context and shared concepts,” Journal of Cyber Policy, vol. 3, no. 2, pp. 258–283, May 2018.

[16] R. Pavlikov and R. Beisembekova, “Architecture and security tools in distributed information systems with big data,” in 2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT), Baku, Azerbaijan, Oct. 2016, pp. 1–6.

[17] J. Klein, R. Buglak, D. Blockow, T. Wuttke, and B. Cooper, “A reference architecture for big data systems in the national security domain,” in Proceedings of the 2nd International Workshop on BIG Data Software Engineering - BIGDSE ’16, Austin, Texas, 2016, pp. 51–57.

[18] A. Z. Faroukhi, I. El Alaoui, Y. Gahi, and A. Amine, “Big data monetization throughout Big Data Value Chain: a comprehensive review,” Journal of Big Data, vol. 7, no. 1, Dec. 2020.

[19] D. M. Strong, Y. W. Lee, and R. Y. Wang, “Data quality in context,” Communications of the ACM, vol. 40, no. 5, pp. 103–110, May 1997.

[20] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, “Methodologies for data quality assessment and improvement,” ACM Computing Surveys, vol. 41, no. 3, pp. 1–52, Jul. 2009.

[21] B. Saha and D. Srivastava, “Data quality: The other face of Big Data,” in 2014 IEEE 30th International Conference on Data Engineering, Chicago, IL, USA, Mar. 2014, pp. 1294–1297.

[22] L. Cai and Y. Zhu, “The Challenges of Data Quality and Data Quality Assessment in the Big Data Era,” Data Science Journal, vol. 14, no. 0, p. 2, May 2015, doi: 10.5334/dsj-2015-002.

[23] J. Moreno, M. Serrano, and E. Fernández-Medina, “Main Issues in Big Data Security,” Future Internet, vol. 8, no. 3, p. 44, Sep. 2016.