International Journal of Computer Trends and Technology
ISSN: 2231 – 2803 / https://doi.org/10.14445/22312803/IJCTT-V70I1P106

Volume 70 Issue 1, 28-33, Jan, 2022

Review Article

The Importance of Data and Analytics Provenance and Governance in the Realm of Datafication

Prashant Tyagi[1], Sharada Devi P.P[2]

[1]*IEEE Senior Member, IEEE Computer Society Member, BCS CITP, BCS Professional Member Scottsdale, USA.*

[2] *Scottsdale, USA.*

Received Date: 18 December 2021
Revised Date: 20 January 2022
Accepted Date: 30 January 2022

Abstract *- This long paper discusses the importance of having a strong Data Governance program in place for organizations to trust their data so that this governed data becomes a key asset and can feed into various reports that businesses rely on heavily to drive positive results for their enterprise. In the world of digitization, where Industries, Organizations and Businesses are increasingly embracing digital-first business strategy, they are beginning to reinvent and redefine how they are conducting business fundamentally by adopting and implementing newer technologies to help them succeed, maintain their strategic position in the market and gain competitive advantage in an ever-changing business climate [1]. In the process of dramatic digitalization, globally spread-out communication networks, highly available, highly scalable computing capacity in the cloud and hybrid-cloud environment datacenters have given the capability for companies and humans to share information seamlessly to empower them with the ability they need to make faster and quality time-saving decisions. So, what is Datafication? Datafication, on the other hand, can be thought of as having a common theme with the Operational Technology (OT), which deals with more of unembedding the knowledge tied to the physical objects by detaching them for the data associated with them[2]. Having exposure to such a plethora of data, it is then very critical to trigger the discussion and the thought process around where to draw boundaries between corporate and social and governmental responsibilities and the public and private lives of people. Having said that, the provenance of the things or the data lineage matters for the data that powers the analytics and drives critical business decisions. The work in this paper discusses the importance of having an end-to-end data lineage, data provenance and data governance program in place all along the information value chain to help analyze and determine whether such information/data is "fit for purpose" as it gets reconstructed and transformed into an information product [2].*

Keywords *- Data governance, Data provenance, Data lineage, Digitization, Digitalization, Datafication, Operational Technology (OT), Information Technology (IT), Analytics, Data, Cloud, Hybrid-Cloud, Digital, Transformations, Initiatives, Security, Data quality.*

I. INTRODUCTION

A. What is Digitalization?

Digitalization can be referred to as the process of converting pieces of information into digital formats for transmission, re-use, re-construction, modification, and manipulation. For instance, conversion of text into HTML pages, images into JPEG, PNG or similar media formats, music and videos into MP3 are a few of the examples of digital transformation or digitization. "Digitalization is a process that has been active within society since the late 1950s, with the birth of the semiconductor industry" [2]. With the evolution of digitalization and the advent of disruptive technologies like cloud computing and IoT technologies, companies have been able to create new economies of scale that can span the globe. We are observing trends where companies are investing heavily in Digital Transformation initiatives to improve their productivity, increase operational efficiencies through process automation, achieve performance excellence, reduce their turnaround time to the market, thus providing an enriched customer experience. Through the implementation of Information Technology (IT) by organizations across their various business functions, like payroll, finance, operations, human resources, call centre operations, sales & marketing and other core business processes, they have been able to streamline the process of doing businesses, empowering them with the ability to share information seamlessly, link systems and disparate source systems and their data sets together to be able to gain deeper business insights and critical business information. Corporations have been able to use digital platforms to create round the clock information-technology driven businesses, thus enabling an enhanced customer experience. The evolution of digitalization and its impact on industries and organizations can be summarized in the table as shown in Figure:1.

Era	Economics on API Usage	Impact on Industries and Organizations
1960-the 1970s	API's are used for the division and distribution of information internally in companies	Different programming teams can use each other's code. Code reusability.
Late 1970's to early 2000s	APIs are used to fuel the internalization of production and globalization of industry, in particular supply chain and value networks.	Different teams located in different parts of the world can now use each other's code.
Late 2005 onwards	Competitive clash of platforms based on open API's	Underlying platforms are dominated by the total cost of ownership (TCO).
	API's create temporary monopolies on mobile devices, not just the internet.	Data analytics increases in importance the system integration and virtualization leads to massive amounts of data from both internet and mobile devices that need to be understood for organizations to be able to gain actionable business insights to empower them to improve the speed and quality of their decision-making process.
	Supply chains reconfigure, and some become virtual-'online.'	

Fig. 1 Evolution of digitalization and its impact on industries and organizations[3]

B. What is Datafication?

"Datafication, on the other hand, relates to the use of digital technology to unembed the knowledge tied to the physical objects by detaching them for the data associated with them" [2]. What this means is that datafication has made itself apparent in society and in people's lives in a variety of forms, and most often, if not always, it is associated with sensors or actuators, Internet Of Things (IoT), Industrial Internet Of Things (IIoT)[1]. In many cases, even a mobile device would be sufficient to create a siloed body of knowledge from a physical object like a piece of equipment, infrastructure, a 3D printer or even a person. Datafication is relatively newer and most of the time adopts the common theme of "Operational Technology (OT) which can be thought of as the hardware and software interaction that detects or causes a change through the direct monitoring and/or control of industrial equipment's, assets, physical processes, devices, infrastructure and events" [1].

In analogy to Operational Technology, we can think of "Datafication as the interaction between digital and physical objects and mass customization of products and services for- and by-end users rather than merely process automation or efficiency improvements although datafication techniques can, of course, be used for this as well. Datafication does not change just how we do business with one another or how we manage our companies, lives and cities – it begins to challenge some of the fundamental mechanisms upon which society has always depended" [2]. Perhaps a way to look at this is that, as a process, object or activity and their interaction between them evolves, everything around it also must evolve and adapt along; otherwise, systemic imbalances will be created. Datafication is simply the act of taking people, processes, events and indeed aspects of the world and turning it all into data that can be used on a larger scale. The word "datafication" is a new addition to the lexicon and was first used in a joint essay publication by Kenneth Neil Cukier and Viktor Mayer-Schoenberger. This essay was published on the Foreign Affairs website in 2013. In that essay, Cukier and Mayer-Schoenberger broadly discussed the role of "big data", describing it as information that is sourced from a large array of resources and which is put to use in very extraordinary ways, far beyond the regular or intended use that the data was originally collected for. (Cukier and Mayer-Schoenberger, 2013). To summarize, datafication can be thought of as a technological trend and advancement turning many aspects of our daily lives into computerized quantifiable data using processes and procedures to transform organizations into data-driven enterprises by converting this data into information that can deliver immediate business value to corporations. Datafication basically refers to the fact that daily interactions between living things can now be transformed, rendered, and presented into a data format or a consumable data structure that can be put to immediate social use. For instance, the wearable technologies like the Fitbit, which can monitor and report on the various key health metrics and suggest recommendations based on your health conditions, the Nike self-tying shoes – the 'smart-shoes' which automatically tightens once the user puts them on incorporating the 'adapt technology' which adjusts to the shape of the foot of the user. Datafication is the social platforms could be the case of Facebook or Instagram constantly monitoring and collecting data about the friendships of the users, their likes and dislikes and using this data for targeted pricing and promotion strategies to present curated marketing content to its target

audiences. Netflix, Amazon Prime, internet streaming media providers are also some excellent examples of datafication processes. Recognizing the genre of movies watched in the past, the watching patterns and trends by people and their families and also based on other similar audience preferences, these streaming media providers employ recommender systems that use the personal behaviour and choices of the audiences to recommend a new movie/series when added to their content catalogues. Another example would be Ibotta, a mobile technology company that enables users to earn cash back on their in-store/ online purchases and recommends products that are on sale in various stores and guides the users to purchase those products for availing additional savings. In this realm of digitalization, we can cite many such datafication examples.

Over the past decade, Digital Transformation initiatives in organizations and continuous automation of traditional manufacturing and industrial practices using smart modern technology have caused the blurring of boundaries and physical gaps between Datafication, Operational Technology (OT) and Information Technology (IT), and as organizations have catapulted into the fourth Industrial Revolution (Industry 4.0)[1], this has caused an explosion in their threat attack surface. This convergence of OT, IT and proliferation of datafication has increasingly stressed the importance and the need-to-know source of origin of the data, how it has been reconstructed, rebuilt, transformed and modified in transit, to make it consumable for reporting purposes. Such visibility to the 'information value chain is of paramount importance so that the data lake or data warehouse which houses such information in organizations becomes a reliable and a valuable part of the organization's ecosystem, and the data from the governed lake becomes highly useful and trusted as it becomes a key business asset that drives positive results for the enterprise.

II. WHAT IS THE RELEVANCE OF DATA PROVENANCE, DATA LINEAGE AND DATA GOVERNACE IN TODAY'S DIGITAL WORLD?

Data is everywhere, and today all our activities, from watching our favourite TV shows to doing our daily activities like taking a run, is set to leave a digital trace or a digital trail, a digital footprint. About 85 to 90% of the world's data have been produced over the past two years. According to a report released by Mckinsey, most companies are capturing only a fraction of their data's value due to the lack of completeness and trust the data citizens have on their data. The organization's capability and speed to derive value from the data is not in pace with the rate at which data is being generated. Organizations trying to harness this swamp of data are turning to self-service Business Intelligence (BI) tools to put data directly in the hands of the business users who need it.[3] "Still, according to a recent report released by Gartner, more than 60% of those initiatives will fail to deliver trustworthy data"[3]. To better curtail this data deluge and extract value from the data, organizations have to create an information value chain giving importance to Data Provenance, Data Lineage, Data Management and Governance actions.

Data Provenance and Data Lineage are not the same concepts, and yet we see many data citizens like Data Scientists and Data Engineers use both these terms interchangeably.

A. Data Lineage

Data Lineage can be thought of as a subset of Data Provenance. It can be represented visually to depict the data flow and its movement from the source to target via various transformations it has undergone throughout its life cycle. Data Lineage can be thought of as answering the 5" W's" of data as to who is using the data, what information does it contain, where does this data come from (which system is generating it) when was this data created or transformed and why does this data even exist? [3]. Data that is collated from different source systems usually land in the staging area like a data warehouse or a data lake. Before this data is delivered to the business user for consumption into various reports, this data is cleansed, massaged, curated, catalogued, aggregated, transformed, and manipulated through ETL tools, Adhoc SQL's Python Scripts, spreadsheets etc. This manipulation and transformation of data can often lead to anomalies and inaccuracies in the data and an overall lack of trust. Data Lineage graphs help address the flow of data movement from the source to the destination and provide a blueprint for 'Data Consistency', 'Completeness', 'Trustworthiness', 'Accuracy', 'Integrity' and 'Consistency'. It provides the necessary context and useful information that can help business users choose the right data for making data analytics-driven informed business decisions and taking actions.

B. Data Provenance

Data provenance, on the other hand, captures the inputs, entities, systems and processes that influence the data of interest. It is the data that is process-oriented used to reproduce the data of our interest. "It contains the historical trial of data, its origin and generates evidence that supports forensic activities such as data dependencies and analysis, recovery and auditing and compliance analysis. Data Provenance = **Data Lineage** (*what is the genealogy, history of its journey, where did it begin, how did it come into being, how did it change over time, where has it been, systems it has travelled, any loss or gain) (i.e. data-oriented, metadata*)) + **Extra** (*the inputs, entities, systems and processes that influenced the data - i.e. process-oriented, which can be used to reproduce the data)"[4]*

C. Data Governance

Data Governance is the management of practices and processes, rules and procedures that ensure the quality, integrity, availability, usability and security of the enterprise data assets (both transactional and analytical data) –on-

premises, on cloud and in a Hybrid-Cloud environment [5]. Data governance program with data quality incorporated will empower organizations to move forward with future road map projects like Digital Transformation initiatives, Cloud/Hybrid-Cloud deployment model adoption, ERP transformation and sunsetting of the legacy systems enabling corporations with usable and predictable trends to confidently develop governance policies and feel strongly about its governance decisions. This also provides data and analytics citizens and leaders an integrated governance program enabling greater self-servicing capabilities by end-users, Internet of Things [1] and decentralized analytics programs. A well thought and thoroughly planned design approach to enterprise data governance brings new avenues to organizations and presents them with more than just an opportunity to ingest vast amounts of data. Organizations at this point need to be very cautious to adopt certain best practices, which we will discuss in the following sections to keep the data lake from turning to a data flood or a data swamp. A well-managed and organized data lake provides an environment to explore and exploit data-driven actionable business insights to increase business agility, handle growing challenges in the field of rising cost pressure, market shifts and competition, leverage new growth opportunities and try out new business revenue models. Unless organizations adopt a robust Data Analytics and Business Intelligence Governance Strategy, the status-quo aspects of analytics will result in failure because of increasing vendor interest in selling their analytics capabilities which will result in fragmented (technical debt) and siloed analytics end-user initiatives driven by tactics rather than holistic optimal enterprise analytics and governance strategy. With an increase in Internet of Things (IoT) devices, corporations will have to adopt a strategy to define data governance, and data quality rules different from the traditional transactional and operational data housed in on-premises data stores such as data warehouses and datamarts. For organizations to decrease their time to value from the data, have more context-driven and focus on providing consistently high-quality, reliable data across the organization to enable business objectives, streamline processes and enable digital transformation, it is very critical for enterprises to have the right and a robust Data and Analytics Governance program in place. [6]. In fact, Analytics Governance will become increasingly important for businesses success, but analytics governance cannot be separated from data governance, and both go hand-in-hand. Having the right Data Governance and Data Management program in place in today's world of datafication will empower organizations to scale their data storage and data integration workloads, enabling them to perform advanced analytics inexpensively at scale, enhance real-time access for data on-premises or on a cloud, better usage of data and other assets making data easier to share across the enterprise and also move forward with a focus of having a modernized governed data management infrastructure in place.

III. CHALLENGES AND PAIN POINTS IN IDENTIFYING THE RIGHT DATA FOR CLOUD/HYBRID-CLOUD DATA ANALYTICS

A. Data Quality Problems

According to recent research conducted by TDWI, 67% of the organizations cite data quality issues as their biggest challenge, which inhibits them from achieving quicker turnaround time and achieving faster actionable business insights[7]. Business Users cannot trust the data they are consuming for analytics because of the absence of data provenance and data lineage to explain the origin and process of reconstructing the data for consumption purposes, which causes delays in the completion of work and sometimes also rework of the same. Hand coding of the data quality rules and point to point data loading, data integration also cannot keep up with the pace and volume with which the data is being generated and getting ingested into the data lake/data warehouse and data coming from a variety of different sources and rising workloads augment this challenge even further.

B. Technical Skill gap leading to Shadow IT projects

Piecemeal approach to multi-cloud, hybrid cloud adoption by organizations to catch up to the race of competition with breakthrough and disruptive technologies has also led to the increase in the technology gap, which cannot be substituted by training and development but would call for the need of skilled resources to detect the inconsistencies and flaws with the data ingestion and data integration process before making the data ready to consume. This is also one of the reasons organizations are leaning more towards using managed services from cloud service providers like AWS, Azure, Google Cloud to decrease their turnaround time to the market and provide self-serving capabilities to the business to empower them to take data-driven actionable biasness insights. This has led to many shadows of IT projects where it becomes increasingly difficult to track inconsistencies and anomalies under the hood.

C. Enterprise Data Catalog

With the paradigm shift of ETL (Extract, Transform, Load) to ELT (Extract, Load, Transform), organizations have allowed the data ingestion from their different sources into the data lake, repeating the same mistake of the past making it increasingly difficult to understand what information is available in the data lake and inability to track data lineage. Having an enterprise data management solution – like an enterprise data catalogue in place provides an enterprise-scale repository of all the data the business has available for data analytics, giving data citizens and end-users a single go-to data catalogue, a data marketplace to find, understand, analyze and gain quick, actionable insights from the underlying enterprise data source. Metadata makes the Enterprise Data Catalog possible.

D. Absence of Platform Agnostic Tools and Vendor Lock-Ins

According to recent research conducted by TDWI, there is an increasing expectation in organizations that the cloud service providers would have all the necessary tools and best practices embedded in their cloud platform technologies to provide corporations with the visibility to performance, data quality and data governance [7]. Also, there is a growing concern about vendor lock-ins which is one of the major driving factors for organizations in spreading their data across multi-cloud provider platforms, which further augments the problem of data quality, the ability of enterprises being able to track the source and origin of the data thus making the data inconsistent, incomplete and less reliable.

IV. KEYS TO OPTIMIZE DATA GOVERNANCE PROGRAM

A. Addressing the Data Quality Problems In Hand

By having an Artificial Intelligence(AI) infused cloud data management solution in place, smarter and automated capabilities can be applied to the data quality discovery and remediation process.[7] Common data model, Cloud Data Management[CDM], and change data capture can facilitate collaboration between stakeholders on rules and procedures agreed upon by the business, thus reducing hand-coding and creating more trust through CDM for data pipeline development. Improved data quality leads to better customer engagement, reduced manual intervention, increased automation in data ingestions and creating analytics-ready data sets made available on-demand which can be consumed by downstream applications, algorithms and other data end-users to find actionable insights in minutes, not months.[8].

B. Establishing and Socializing the Use of Enterprise Data Catalog

Having an Enterprise Data Catalog, a Data Marketplace, makes it easier and faster to protect sensitive data and apply data governance policies, procedures and constraints on the identified sensitive data. It also makes it easier and faster for data citizens and consumers to find the correct data, observe and analyze its data lineage and find the right data for their analytics, empowering the data citizens to find answers and information to the questions they have in hand. Metadata, which is used to document the data with cataloguing, classification, structuring, securing, and managing data collection, ensures organizations have a well-governed data marketplace (rather than a data flood). Then IT experts, data stewards and business users can make these data collections easier to understand and consume by creating business metadata, tags, blog fields, as well as standard and custom-defined settings and properties on their data catalogues. Having an enterprise data catalogue in place will avoid delays, inconsistent datasets for analytics, uncertainty about data definitions and help corporations avoid problems protecting sensitive data getting distributed across multi-cloud environments.

C. Having a Holistic Visibility of Data and its Workloads

Having a single platform limits the ability of organizations to gain better visibility to the data surface, performance, data availability, and workloads; having a multi-cloud architecture will again require the organizations to have the ability to have sight and manage across platforms and cloud service providers. Automations with cloud data management (CDM)/ change data capture (CDC) should be pursued to avoid manual interventions and oversight of the data ingestion and data pipeline management process.

D. Better Pipeline Management and Consolidation

By streamlining and consolidating data pipelines, efficiency, scalability, and quality of the data pipelines can be enriched. Keeping raw data pipelines separate from cleansed, transformed, curated, enriched, and governed pipeline helps in addressing advanced analytics and AI/ML (Machine Learning) workloads and use cases. Identifying redundant data pipelines and consolidating them and using a single centralized data lake to bring together divergent data warehouses and data lakes together will help in enhancing the quality of the data ingestion and transformation pipelines.

E. Providing Data Relevant to Personas Across the Enterprise

Having automated data mapping of technical metadata to business terminology enables future scaling, reduced manual intervention, reduced maintenance overheads regardless of the ever-increasing volume, speed and variety of the data mapping the data that is tailored to different personas reduced the time for deployment, which would have otherwise called for lengthy excruciating and time consuming manual process[9].

F. Streamlining of Workflows and Establishing Data Governance requirements and policies in place

By establishing data governance policies, procedures, data quality requirements like consolidating, cleansing, transforming and standardizing department databases and scheduling reports for regular updates to happen, organization's data management practices will be consistent across the enterprise giving a holistic view and better visibility to the value obtained from such trusted data. Also, identifying the data owners data stewards, who are involved in key data areas and having it documented and socialized with the enterprise community, streamlines the process of "moving data governance away from control towards collaboration" and adopting a "governance by design approach"[9].

G. Integrating Governance with Data Security

Highly sensitive data, for instance, medical records protected by HIPAA, or data that is subject to the privacy and data protection such as General Data Protection Regulation (GDPR) or the California Consumer Privacy Act(CCPA), is not just enough to ensure that the data is

correct and the integrity is maintained, but it becomes critical to define data rules for access, data encryption at rest and in transit[10], data masking, de-identification, retention and deletion as well.

H. Conducting Data Quality Reviews and Monitoring in a timely and consistent manner

Reviewing and monitoring data is pivotal when it comes to the success of the data governance and data management program for data and analytics governance [10]. Capturing the data changes that are happening daily when it comes to customers and gaining insights about the data quality metrics will help enterprises know what data governance policies and procedures need to be modified over time. It is always ideal for automating the data quality metrics and KPI (Key Performance Indicators) monitoring.

I. Make it a Consistent Practice to Cleanse and Correct your Bad Data

Once the data governance program policies, procedures, and data quality rules have been defined, it should be consistent practice within organizations to cleanse and correct any bad data and manage the trusted and governed data by incorporating data quality monitoring tools. When this practice is consistently followed, enterprises can remediate data issues and adhere to the defined data quality rules and data governance policies and procedures. This will help increase the level of trust and maintain the integrity of the data in supporting various data-driven initiatives across the entire organization.[10]

V. CONCLUSION

Information and data have become a strategic asset and a competitive advantage to the digital businesses, best able to exploit it. To manage the information generated by an organization, an effective data strategy incorporating data quality and data governance policies and procedures(a data governance program) must exist – a plan for maintaining and improving the quality, integrity, access and security of data. With data capture being widely dispersed, data mobility being challenged by network bandwidths, information security risk, data protection and data privacy restrictions, data loss, lack of data trial/data lineage, it is even more pertinent to have a robust data governance program in place, the absence of which may have serious effects such as lack of trusted data for performing advanced predictive and prescriptive analytics, automation, increased costs, reduced productivity, inaccurate and delayed decisions making, brand damage, regulatory penalties and increased cost of compliance. Continued efforts to provide Master Data Management, following a holistic approach to data governance across all data sources, information controls for various platforms that are cloud, on-premises, devices etc., capturing, storing, and partially analyzing large amounts of data to check for consistency and correctness locally, prior to transmitting and aggregating it somewhere else, identifying specific data management controls and performing audit can empower organizations to create and maintain a robust data governance program in place that can boost efficiencies and hasten workflows because it enhances the trustworthiness and reliability of the data. It also provides a shared understanding and a common data dictionary, a business glossary for all data users [9]. When this governed data can be trusted, it will improve the organization's ability to leverage data that drives business and digital transformations that not only reduce risks and costs but also increase operational efficiencies and create new business revenue models for organizations.

REFERENCES

[1] Prashant Tyagi, Convergence of IT and OT – Cybersecurity Related Challenges and Best Practices, International Journal of Computer Trends and Technology 69(2) (2021) 85-92.

[2] The impact of datafication on strategic landscapes by Ericsson

[3] Collibra.com https://www.collibra.com/wp-content/uploads/Ebook-DataLineage-20200113.pdf

[4] Topper Tips Unconevtional https://toppertips-bx67a.ondigitalocean.app/data-lineage-vs-data-provenance/

[5] Prashant Tyagi and Sharada. Devi. P.P, A Functional View of Hybrid Cloud Environment-Use Cases and Best Practices, SAP Publications http://article.sapub.org/10.5923.j.computer.20211101.02.html.

[6] Gartner Report Predicts 2020, Analytics and Business Intelligence Strategy

[7] Transforming Data With Intelligence, Avoid Mistakes of the Past on ModernCloudDataManagement, https://tdwi.org/webcasts/2020/05/adv-all-avoid-past-mistakes-use-modern-cloud-data-management-to-deliver-faster-value.aspx

[8] Datasheet from Qlik, Data Integration, Enabling Analytics with Trusted, Business Ready Data\Qlik Catalog

[9] Informatica Whitepaper, FiveKeys to Optimize Your Data Lake with DataGovernance, https://www.informatica.com/lp/five-keys-to-optimize-your-data-lake-with-data-governance_3597.html

[10] Prashant Tyagi Diagnostic, Descriptive, predictive and Prescriptive Analytics Using Geospatial 69(1) (2021) 18-22

[11] Driving Data Governance with Data Quality https://www.talend.com/resources/definitive-guide-data-governance/