Implementing Data Strategy: Design Considerations and Reference Architecture for Data-Enabled Value Creation

Radhakrishnan Balakrishnan
Indian Institute of Management
Raipur, India
radhakrishnan.efpm2016@iimraipur.ac.in

Satyasiba Das
Indian Institute of Management
Raipur, India

Manojit Chattopadhyay
Indian Institute of Management
Raipur, India

Abstract

With the arrival of Big Data, organizations have started building data-enabled customer value propositions to increase monetizing and cost-saving opportunities. Organizations have to implement a set of guidelines, procedures, and processes to manage, process and transform data that could be leveraged for value creation. This study has approached the journey of an organization towards data-enabled value creation through four levels of data processing, such as data extraction, data transformation, value creation, and value delivery. This study has critical inferences on using data management solutions such as RDBMS, NoSQL, NewSQL, Big Data and real-time reporting tools to support transactional data in internal systems, and other types of data in external systems such as Social Media. The outcome of this study is a methodological technology independent data management framework an organization could use when building a strategy around data. This study provides guidelines for defining an enterprise-wide data management solution, helping both the academicians and practitioners.

Keywords: Data Management; Data Management Reference Architecture; Data Strategy; Design considerations for Big Data; Governance for Big Data; Data-enabled value creation;

1 Introduction

Organizations across industries consider data as an essential factor of production similar to other critical assets such as labour, capital, and land. In the past, organizations were primarily focusing on enterprise-specific structured data, and of late, organizations have started collecting data regardless of the size, structure, and the speed at which it is created (Manyika et al., 2011). Big Data is no longer a buzz word used by a select number of technology companies and consultants. Today, most organizations have to deal with large volumes of data as their interactions with stakeholders are recorded with increasing level of granularity. The initial definition of Big Data confining only to Volume has expanded to include Velocity at which the data gets created, Variety such as structured, semi-structured and unstructured formats, Veracity or accuracy of data, and the Value that gets generated through data insights (T. H. Davenport, 2014). There is value hidden within Big Data collected by organizations, and the effective extraction and application of that value have become a key differentiator in the battle for supremacy in the marketplace.

Traditionally, organizations have been collecting and analysing data from various transactional systems to identify patterns and arrive at inferences with which they have enhanced operational efficiency and reduced cost. Now there is a plethora of additional data, such as extensive customer and supplier information, detailed product catalogues, more information surrounding online transactions, and an increase in the volume of communication with third parties, including email and voice recording systems. This is further supplemented with data available on social media (SM) that exist outside of the organization. With this barrage of data, organizations have to identify the right means to store, retrieve and process data to arrive at actionable inferences.

The study on data management has received extensive attention in the recent past, and some of the topics covered in data management are summarized in Table 1. In the literature related to data management, 'How do organizations build a robust data architecture to support different types of data management systems, and what are the critical elements of the proposed architecture that enable data-driven value generation?' has been identified as a less researched topic. To this end, the proposed study attempts to arrive at a reference architecture, including different types of data management solutions, and how these systems can combine to drive value-creation for customers. There are two objectives of this study: (i) to identify how organizations support a variety of data management solutions from traditional RDBMS to NewSQL, and the recommended data architecture for supporting data-enabled value creation, and (ii) to understand the design considerations in implementing the proposed architecture. This study is organized as follows: In Section-2, the study presents a literature review on the need for data management solutions to create data-enabled value creation. In Section-3, the study covers the research methodology, research setup, and data analysis associated with the investigation. The findings are presented in Section-4, followed by a discussion on the findings in Section-5, and the conclusion in Section-6. The study has identified a methodological technology independent framework an organization could use when building a strategy around data-enabled value creation and delivery.

Aspects of Data Management	Notable Contributions
Fundamentals on Data Analytics	(T. H. Davenport, Harris, Long, & Jacobson, 2001)
Data Analysis on Web 3.0	(Hendler, 2009; Spivack, 2011)
Data management – Value Creation Framework	(Lim et al., 2018)
Data Management – Big Data based value creation	(Mazzei & Noble, 2017)
Data Management – Building blocks for a data-driven culture	(Anderson & Li, 2017)
Data Strategy – Guidelines for development	(Bowen & Smith, 2014)
Data Management – usage of relational database management systems (RDBMS) and NoSQL	(Link & Prade, 2019; Pokorny, 2013)
Data Management – Migration from RDBMS to NoSQL	(Rocha, Vale, Cirilo, Barbosa, & Mourao, 2015)
Data Management – Features of NoSQL systems	(Mohan, 2013; Pavlo & Aslett, 2016)
Data Management – Use of NewSQL systems	(Kaur & Sachdeva, 2017; Simsek, 2019)

Table 1: Aspects of Data Management and Notable Research Contributions

2 Literature Review

The first section of the literature review discusses how data strategy is critical for an organization, and the design considerations for data-enabled value creation. The second section discusses the use of traditional relational database management systems (RDBMS) to store structured data and provides a brief on the issues related to RDBMS. The third section provides an overview of how NoSQL systems could be used to overcome some of the problems related to RDBMS. The fourth section discusses NewSQL systems and how they could be leveraged to build data management solutions. The last section discusses the need for data governance and critical elements of data governance.

2.1 Data Strategy – A critical component of Web 3.0

John Markoff of the New York Times had coined the term Web 3.0 to refer to the intelligent web such as the "*applications using semantic web, natural language search, data mining, machine learning (ML), and artificial intelligence (AI) technologies that emphasize machine-facilitated understanding of information to provide a more productive and intuitive user experience*" (Spivack, 2011). Web 3.0 technologies enable firms to respond to external changes quickly by integrating data and applications and providing the ability to "*infer relationships between data available in different applications or different parts of the same application*" (Hendler, 2009). Web 3.0 comprises of technologies such as ubiquitous connectivity, network computing, open technologies, and the intelligent internet (Spivack, 2011). The intelligent web consists of a variety of technology solutions associated with data stores and data processing using natural language processing (NLP), ML and AI.

In the existing literature on data management, the terms *data* and *information* have often been used interchangeably. This study differentiates data and information based on the widely popular *data-information-knowledge* hierarchy (Braganza, 2004; T. H. Davenport et al., 2001). Data are facts and observations, and in a given context, processed data becomes information (Zack, 1999). The information that leads to results becomes part of the knowledge. Data analytics is defined as the transformation of quasi-finished data into useful insights and knowledge (T. H. Davenport et al., 2001). The quasi-finished data goes through a process of cleansing, amending, and augmenting before going through further processing by statistical techniques, computer programs, and manual analysis.

Organizations consider knowledge as a strategic asset, and organizations have to create, capture, and share knowledge to remain competitive (Zack, 1999). Organizations implement knowledge management mechanisms by building repositories of explicit knowledge, defining roles to execute and manage the knowledge accumulation process, and developing solutions using information technologies to support knowledge repositories and processes (Zack, 1999). An organization achieves data-enabled value creation by focusing on four areas, such as data collection, information creation, value creation, and distribution through provider networks (Lim et al., 2018).

Organizations use data in a variety of ways, such as for exploring new opportunities, for arriving at insightful analysis on various internal and external factors, and for improving the organization's data eco-system by compiling all possible data from end customers (Mazzei & Noble, 2017). Organizations could build innovative business models by building data-enabled assets and resources. Organizations, by developing analytical capabilities, could improve product development, customer service, and other value-chain activities (Mazzei & Noble,

2017). Organizations look at data and analytics as tools for enhancing innovation and operational efficiency, thus leading to additional revenue streams. Organizations could reap huge rewards by implementing effective means of data management solutions. Data-driven organizations have higher output and productivity than their less data-driven counterparts. Organizations could facilitate data-driven culture by implementing various measures such as building a repository that provides a single version of truth, building a repository containing data dictionary to understand what the data means, giving access to information for all relevant roles in the organization, and enabling managers to support data-driven decision making (Anderson & Li, 2017). Technology is the critical enabler for big-data aided value generation. However, the quality and integrity of data used for arriving at inferences depend on business processes and rules rather than on technology (Bowen & Smith, 2014).

Firms generate value by analysing data available at their disposal, and the value derived from analysis helps on both internal operations front, and external market services front. The data-enabled value generation options range from managerial actions such as implementing data-driven process management to managing operations through data-driven workforce management and creating new market offerings based on data analytics, ML and AI techniques. For an organization to build data-enabled value creation, it has to deploy the right set of data management processes, tools, and methods for facilitating the building blocks necessary to create a data-driven culture. This study is an attempt in that direction to provide insights on a reference data architecture for handling all types of data in a corporate environment such as data in traditional transactional systems, data warehouses, Big Data repositories, schema-less data stores, and a newer variant of the scalable transactional system such as NewSql. The next few sections deal with various types of data, an orientation on the usage, shortcomings, and alternative options.

2.2 Relational Database Management Systems (RDBMS)

RDBMS store and represent data in two-dimensional tables. This is a simple and effective way of automated storage and retrieval of structured data. While using RDBMS, programmers employ query-based scripting languages to perform data management tasks. This continues to be useful when the data is in a structured format, and the scale of the data is manageable. RDBMS are suitable for online transaction processing (OLTP) applications and characterized by atomicity (A), consistency (C), isolation (I) and durability (D). The ACID properties of RDBMS ensure that the related transactions get completed as one unit – *all or nothing*, transactions happen independently, transactions that are written to the storage result in consistent data stores, and data stays durable irrespective of system failures (Pokorny, 2013).

RDBMS have been used for decades and are mature systems, yet, in the face of ever-changing types of data and the complicated relationships among them, they become difficult to use. In RDBMS, especially in OLTP systems, data gets stored in normalized forms following Codd's rules, eliminating redundancy and making sure that all changes in the data are consistent across the models (Link & Prade, 2019). The more complex the collection of data is, the more levels of hierarchy and cross relationships, the less possible it is to represent the collected data within the simple tabular structures of an OLTP relational database system.

The issues associated with normalized data resulting in inefficient retrievals and lack of support for complex structures such as hierarchies are eliminated in data-warehouses. Data-warehouses are relational in nature and store aggregate information that can readily be used for analysis and reporting. RDBMS supporting data warehouses feature non-normalized or

de-normalized data stores to facilitate easy retrieval for analysis and reporting. Data warehouses are built with star-schemas or snowflake schemas. Data warehouses comprise of dimensional stores through which hierarchies can be supported. Also, data warehouses enable storage of business metrics associated with various dimensional stores.

When it comes to performance, RDBMS could scale up very well, but always with an associated cost. The initial investment needed for RDBMS is high as one has to pay for both software systems such as Oracle, and high-end servers with excellent processing capabilities. If one uses RDBMS for storing some reasonably common forms of data like storing the content from channels such as SM, articles, blogs, forums, to identify prospects in various domains, it will result in processing heavy-duty data. While using RDBMS for this purpose, one would run into issues on cost, performance, scalability, and even on building complex programming logic for text processing. The applications using a massive amount of unstructured data streaming on a real-time basis would need a scalable system with extensive storage and high availability (Rocha et al., 2015). RDBMS support scaling-up vertically by moving the database to a machine with better hardware, resulting in additional cost. The sources of data such as real-time feeds and SM warrant a new type of system called NoSQL that facilitates scaling-up horizontally (or scaling-out) by adding more low-cost machines or commodity servers (Pavlo & Aslett, 2016).

2.3 'Not Only SQL or NoSQL' – A Schema-less database solution

With the arrival of Web 2.0 and Web 3.0, different types of data formats and various data channels emerged, and this new paradigm warranted for a different approach for data storage and retrieval. The modified approach to storing data in NoSQL is called CAP. CAP deals with data consistency - data being correct all the time, data availability - reading and writing data all the time, and partition tolerance - no complete failure of the system, and the system becomes consistent when it comes online. The type of systems based on CAP is not ACID and called BASE systems (Pokorny, 2013). BASE reflects the features of NoSQL systems that include *BAsic-availability*, *Soft-state*, and *Eventual-consistency*. The critical feature of NoSQL system is that it sacrifices robust transactional guarantees and data models of RDBMS in favour of eventual consistency and more flexible data models such as key-value pairs, document-stores, columnar stores, graph databases and likes (Pavlo & Aslett, 2016). Performance and Flexibility are the two critical reasons for firms to move to NoSQL systems (Lourenço, Cabral, Carreiro, Vieira, & Bernardino, 2015). Performance is focused on sharing and managing distributed data, and flexibility is the ability to handle different types of data such as unstructured and semi-structured data that may arise on the web.

In the BASE model, the main emphasis is on data availability despite multiple system failures. NoSQL databases spread data across many storage systems with a high degree of replication. NoSQL solutions operate similarly as other distributed systems, providing critical features such as scalability, concurrency, data fail-over and recovery, and data security. The storage failures do not necessarily result in a complete outage of the system. In RDBMS, data is organized in the form of database tables, whereas NoSQL databases use different approaches for storing data (Sareen & Kumar, 2015). For example, a key-value NoSQL database such as MongoDB stores values for each key and distributes those values across the database to allow for easy and efficient retrieval. The NoSQL database does not have a standard schema or physical model that stores the meta-data for the data elements. In other words, there is no schema available for NoSQL databases, and they could be categorized as *schema-less* databases.

NoSQL systems are highly optimized for storing high volumes of data and efficient data retrieval. NoSQL system could be a better fit for storing blogs, articles, SM content and any unstructured data that does not fit into a relational database table. For NoSQL systems, developers do not create logical data models, and hence, data is unconstrained (Stantic & Pokorny, 2014). There are query tools such as XQuery and JSONiq developed to support data retrieval from NoSQL systems. However, they are not widely used. In cases where NoSQL query tools are not used, the NoSQL query options are built into the application layer. NoSQL systems are inherently huge in nature because of the type of data that gets saved. The usage of unconstrained data in NoSQL systems is generally supported by metadata such as data catalogues for facilitating efficient and faster data retrievals.

NoSQL systems complement RDBMS, and hence these systems co-exist. The issues associated with RDBMS, such as data model change management, and data integrity enforcement, are eliminated in NoSQL systems. NoSQL systems are useful when the data's nature does not require a relational data model (Sareen & Kumar, 2015). NoSQL systems present a different set of issues such as lack of designer tools to manage NoSQL systems, and lack of routines or options to migrate from one NoSQL system to another NoSQL system (Mohan, 2013). NoSQL systems do not have any standardized way of storing and retrieving data like RDBMS and SQL based systems, and hence, facilitating change management on NoSQL based systems could be frightening (Mohan, 2013). NoSQL systems are not recommended for applications requiring functionality, such as data integrity and security. However, they could be used for analysing high volume and real-time data such as web-log data, and click-stream data (Stantic & Pokorny, 2014). NoSQL cannot be used as the standard data store for all types of data, and organizations have to use NoSQL systems for the right use cases to facilitate efficient data storage, retrieval, and change management. This study attempts to understand use cases relevant for NoSQL systems and how these systems could work with existing traditional RDBMS supporting *online transaction processing* (OLTP) solutions.

2.4 NewSQL systems

NewSQL systems represent a newer version of the already existing RDBMS technologies. NewSQL systems are the ones that combine the critical features of both RDBMS and NoSQL systems (Pavlo & Aslett, 2016). NewSQL systems are based on distributed architecture operating on shared-nothing resources and contain components to support multi-node concurrency control, fault tolerance through replication, and distributed query processing (Pavlo & Aslett, 2016). This includes cloud computing providers that offer NewSQL database-as-a-service (DBaaS) products. NewSQL DBaaS products are hosted on cloud infrastructure, and the customers are provided with a universal-resource-locator (URL) to connect to the system. DBaaS enables the pay-per-use of computing resources.

NewSQL systems support horizontal scaling by adding new servers, and support partitioning, sharding, and clustering. NewSQL systems, by combining features from both NoSQL and RDBMS, could support both ACID transactions and enable high-availability systems. NewSQL systems meet many of the requirements for data management in cloud environments, combined with the benefits offered by traditional RDBMS (Grolinger, Higashino, Tiwari, & Capretz, 2013). NewSQL could be considered as an alternative to NoSQL systems for new OLTP applications (Kaur & Sachdeva, 2017). Some of the NewSQL systems, such as VoltDB, support in-memory processing of complex ACID transactions in a distributed manner. NewSQL systems are developed from scratch using shared-nothing architecture, or

built by augmenting the existing RDBMS, enabling them to scale-out. Through augmentation, the features of NewSQL could be extended on top of existing RDBMS solutions, and this feature would be useful for organizations that are not willing to migrate to new database solutions. The traditional one-size-fits-all approach of RDBMS does not hold good anymore, and NewSQL empowers organizations to move towards specialized database designs catering to their specific data storage and processing needs (Simsek, 2019).

2.5 Data Governance

Information is a critical resource for organizations, and it forms the basis for business intelligence and data analytics (Newman & Logan, 2009). Information takes on new importance in IT-related strategies such as service-oriented architecture. Data governance is defined as an organizational approach to data and information management through a set of policies and procedures, to manage the life cycle of data from acquisition to use to disposal (Korhonen, Melleri, Hiekkanen, & Helenius, 2013). With data governance, organizations could define guidelines and standards for data quality management and assure compliance with laws governing data (Weber, Otto, & Osterle, 2009). Data or information governance is a mechanism through which data that is considered a liability, get converted into a trusted strategic asset (Kooper, Maes, & Lindgreen, 2011)

The organization's use of Information Technology (IT) is primarily focused on operational efficiency; however, the evolving complex IT landscape has resulted in redundant and inconsistent data in enterprise environments (Korhonen et al., 2013). Incorrect and erroneous data add additional costs and risks and lead to poor managerial decisions. Ever since the inception of Big Data and streaming data, organizations have started considering data they collect from various internal systems and external sources, as assets. Information assets are facts that have value or potential value that are documented (Khatri & Brown, 2010).

The critical business benefit of information governance is the realization of data as a valuable and manageable organizational asset, and other business benefits include improved business decisions due to accurate data, and increased user trust within the organizations' data sources (Marco, 2006). Data governance strategy results in improved reliability, traceability, and authenticity of data, thus providing the backbone communication platform across all stakeholders.

Some of the critical elements of data governance include data quality, meta-data management, and data life-cycle management. Having high-quality information is a pre-requisite to building data-driven customer value propositions (Weber et al., 2009). Hence, it is critical for organizations to focus on data quality so that they could make effective managerial and strategic decisions. Data quality management helps an organization to plan, provision, organize, use, and retire high-quality data (Weber et al., 2009). For a firm to implement the right governance program, it has to manage various aspects of data management such as information security, risk management, data privacy, data security, data loss, and business continuity. Organizations in the digital era have to adopt a holistic approach to manage the three critical challenges associated with data management such as data security, data privacy and compliance obligations (Salido, 2010). These problem areas of data management have to be adeptly supported by three capabilities of firms such as people, process and technology. Organizations design three kinds of governance mechanisms such as decision-making structures, alignment processes, and formal communications (Weill & Ross, 2005).

Successful information governance warrants accountability, and accountability is executed through a formalized process called *stewardship* (Newman & Logan, 2009). The people aspect of data management deals with identifying *data-stewards* - people collectively responsible for data quality in the organization by defining policies and procedures for the classification, protection, use, and management of data (Salido, 2010). The data stewards engage in information governance by implementing managerial activities such as setting standards, enforcing policies around data, resolving intra-team issues and implementing day-to-day operational procedures (Newman & Logan, 2009).

The process aspect of data management deals with defining a set of processes to manage elements such as information security needs and compliance needs of the organization. The process aspect of information governance deals with monitoring and reconciling data at critical points along data's path through the organization's IT systems, and this includes initial data entry points and data aggregation points (Griffin, 2005). At the data-entry points, the IT systems could check for valid values or required values. At data aggregation points, the process steps could include eliminating duplicate data entries and checking for valid cross-references to strengthen the authenticity of data. Also, the process options include features for manual corrections to exceptions and errors that cannot be fixed automatically (Griffin, 2005).

The technology aspects of information governance and data management deal with activities such as providing secure infrastructure, implementing identity and access control, strengthening information protection features, and implementing robust auditing and reporting functionalities (Salido, 2010). The software solution considered for big data-related data management solutions should scale to accommodate large datasets, leverage the underlying hardware platforms efficiently, and bridge the increasing gap between growing data and computing power (Kambatla, Kollias, Kumar, & Grama, 2014).

The terms *data governance* and *data quality* are often used interchangeably, but both are different in theory and practice. Data quality happens at both the front-end through data validations and back-end through databases. The process of data quality is defined, structured, and implemented through a well-designed *data governance* framework. In the *data governance* framework, rules and policies related to data ownership, data processes, and data technologies are clearly defined. In other words, *data governance* provides a framework for managing *data quality* (Ghosh, 2019).

The existing literature on data management covers a variety of topics, as mentioned in Table 1. However, in the current research of data management, queries such as 'How do organizations support variety of data management solutions ranging from traditional RDBMS to NewSQL, and how do organizations integrate them for facilitating data collection, aggregation, and analysis?, 'What is the recommended data architecture for supporting data-enabled value creation?', and 'What are the design considerations in implementing the proposed architecture, in terms of data governance, data security, disaster management, and business continuity?', have been identified as less explored, and recommended for research. This study attempts to delve into various aspects of data management, to arrive at inferences on reference architecture and design considerations to enable data strategy for customer value creation.

3 Research Methodology

3.1 Research setting

This study is based on an explorative approach to understand data management solutions used by organizations, and the inferences are generic in nature. The study is executed through interviews with three data architects focusing on building solutions using the latest data management tools and technologies. All three data architects have significant industry experience, worked with multi-national corporations, and architected complex data management solutions for global clients. One of the architects – A1, has been working on building solutions for clients in North America. Another interviewee – A2, has been working with clients in Western Europe and the Americas, and the third one – A3 has experience in building solutions for e-commerce, banking and financial services clients in North America.

This research is conceptual in nature, and there are no control variables, moderating, and mediating factors considered for this study. This study includes interviews that have enabled the researchers to understand the data architecture used for creating a robust data management solution covering all design aspects, end-user aspects, and governance aspects. The interviewees have provided documents on processes used while building data warehouses and presented architecture diagrams that accommodate various data management solutions.

The study involves interviewing three candidates, and for insightful qualitative research, the number of interviewees may not be considered sufficient, as interviewing three people does not satisfy the condition of having a representative sample. However, the three senior people found for this study A1, A2 and A3 are playing leadership roles in their respective teams. A1, as a team lead, manages fifteen data architects, A2, manages twenty-five data architects, and A3 manages fifteen data architects. All interviewees have exposure to variety of business challenges in data management domain, and they have provided solutions to complex business problems, with help from team members. They have a collective knowledge base on various data management techniques and worked on all tools and technologies related to data management. This criterion qualifies them as subject matter experts (SME) who could provide insights on all aspects of data management. The SME has profound knowledge in the subject, and their experience was evident as they could comment and voice their opinion on all prominent and latest methodologies and technologies used for data management. All the chosen interviewees are from different organizations, and hence their comments and opinions are independent, and not influenced by other members. In addition to these discussions, the researchers have met with the team members to have informal discussions on finer aspects of data management. During those meetings, the team members had shared artifacts covering the data management solutioning and implementation details for various clients. The researchers have used the transcripts, documentation, and artifacts to arrive at the inferences and contributions for this paper.

3.2 Research Design and Data Collection

The scope of the research is limited to four dimensions of data-enabled value creation, as suggested by Lim et al. (2018), and the dimensions include data collection, information creation, value creation, and distribution or information delivery. The scope is technical in nature, and covers governance aspects of data management such as information security, data privacy, and data quality. The queries covering these four dimensions are created in

consultation with faculties teaching courses on data management for under-graduate engineering students. The study does not focus on covering the aspects related to hardware or infrastructure components associated with data management.

The study comprises of interviews with subject matter experts, and the research is exploratory in nature. The researchers have avoided any leading queries to interviewees and included probing queries to offer more clarity on the base queries when needed. The questions are open-ended, and some of the queries used for this study are, *"How are various data management systems integrated? From your area of expertise, could you provide insights on why the integration is needed, and how it was implemented?"*, *"How would you handle exceptions while loading different types of data?" "What is the new customer value proposition that comes with the data exploration tools?" "How does your new value proposition get monetized? Could you give specific scenarios on how you helped your client monetize on data offerings?"*, and *"What kind of data governance mechanisms you have implemented? What are the design considerations for business continuity?"*. The study is conducted without any underlying assumptions, and the researchers have personally met all the three data architects, conducted interviews and transcribed the audio files. The transcripts are reviewed with the interviewees for approval, thus enhancing the accuracy of collected data (Huber & Power, 1985). The study comprises of 150 minutes of audio recordings, and 16000 words in transcripts. Also, the authors have referred to various process documentation, and discussions with critical team members through informal *question-and-answer* sessions to answer queries, for a period of another 140 minutes;

3.3 Data Analysis

The researchers have transcribed the recorded conversations and validated the transcribed content through the *maker-checker process*, a dual-approval process typically used in the financial services industry (Commercial Bank, 2018). In this *dual-approval* process, a *maker* enters the transactions into a system, and a *checker* reviews these transactions and approves the same. For this study, an author analysing the transcript files, gets his analysis reviewed and approved by the other author. The researchers have analysed the data, verified each other's work, and frequently referred to both the data and existing literature when developing the research report. This study has identified findings across four aspects of data-enabled value creation, and these inferences are summarized in the next section.

4 Research Findings

The study proposes a methodological technology independent framework an organization could implement to support data-enabled value creation. All the critical elements of the structure supporting data-enabled value creation are portrayed in Fig.1. These components are discussed at length in subsequent sections. The first section provides details on different types of data analytics solutions implemented by organizations. The second section contains the findings on data integration methods adopted by firms, explicitly covering all types of data such as structured data, semi-structured data, and unstructured data. The third section summarizes the findings on data transformation methods used by firms to create *information* that can be part of the *knowledge*. The fourth section lists the findings on the value creation resulting from such an architecture, and the last section summarizes the findings on other business critical considerations such as data quality, data security, and business continuity.

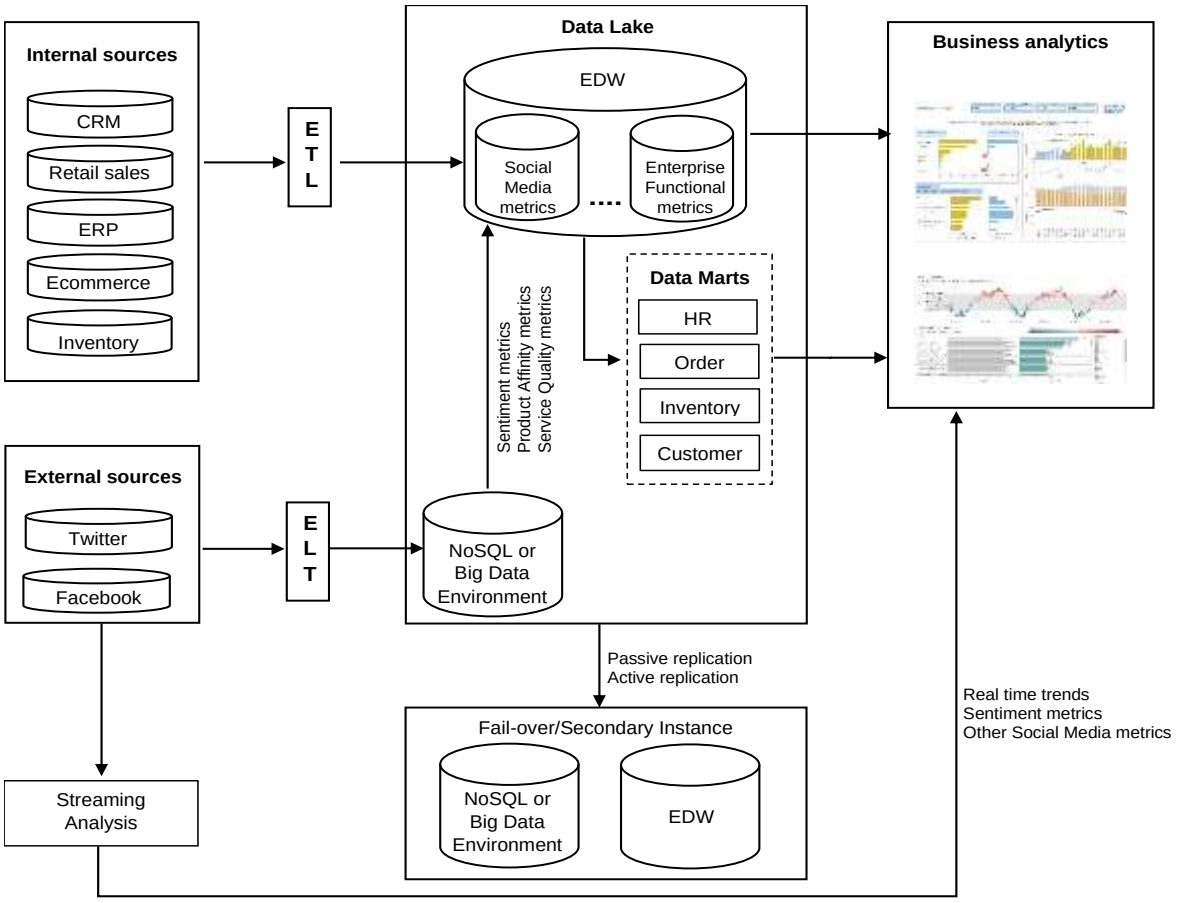

Fig.1 Reference Architecture for creating Big Data enabled value creation

4.1 Use Cases for various data management technologies

One of the data architects has spelled out the use cases of these data management technologies, and opined,

> *"anything that is transactional in nature – where you have structured schema and transactions …things like ERP and stuff – RDBMS is the preferred choice; … when you have aspects related to data integrity and things, that is where RDBMS comes into play; …. the challenge comes as (when) the data volume increases; the only way you can handle volumes is to grow vertically; like you have to increase the processing capability of the server. ….we had influence of social media data coming; influence of semi-structured data coming; semi-structured in the form of XML, JSON, and other files have come; these things resulting in change in data structures as well; that is how we got NoSQL; …lots of vital data that needs to be stored and does not need to be in an integrated manner (data integrity need not be there) then, we have to go for NoSQL systems … a new generation data management solution – you could handle big volumes and scaling like NoSQL – that is where NewSQL comes into (the) picture; low latency data retrievals – huge storage, scaling up - horizontally, with very low latency;"* (A2, Data and Solution Architect)

Data management solutions are implemented for descriptive, predictive and prescriptive analytics solutions. The following table lists the suggestions as indicated by the interviewees, on the type of data management solutions possible with different types of data.

Type of Data Management Technology	Descriptive Analytics Solutions	Predictive and Prescriptive Analytics Solutions
RDBMS (Supports Structured Data and ACID transactions)	Various reporting solutions based on Customer Relationship Management (CRM) and Enterprise Resource Planning (ERP) applications; Management Information Systems (MIS) generating daily reports;	Retail: Recommendation solutions Banking: Fraud monitoring solutions Health-care: Identifying the right solution based on given health conditions (Prescriptive solution)
NoSQL (Supports unstructured and semi-structured data)	Statistics information such as *tag-clouds*, sentiments on social-media content, and visual reporting on hierarchical information	Arriving at the sentiment trend based on the tone of sentiments collected over a period of time; Predicting the most-preferred features in a new product or service based on the comments collected from semi-structured and unstructured data sources
NewSQL (Structured data)	Similar to RDBMS; However, NewSQL is mostly used for real-time analytics and predictive analytics	Streaming analytics – on customer sentiments; Identifying trends in user perceptions; Real-time recommendations on product purchases; Streaming analytics on data from IoT; Market basket analysis in Retail industry;
Solutions supporting Big Data **(HDFS based systems)**	All types of data; Mainly used as 'Data staging' area and data archival area; Helpful to deal with 'Save now and analyze later' kind of data;	Staging area for Extract-Load-Transform (ELT) kind of data extraction; Storage for archived-data wherever needed; System can support all type of data to support descriptive, predictive and prescriptive analytics solutions

Table 2: Data Management Technologies and Use Cases supported

4.2 Data Integration

Data integration involves combining data from disparate and heterogeneous source systems to arrive at a unified view of the data. Data integration consolidates applications within one firm to provide a unified view of the firm's data assets ("Data Integration," 2018). Data integration enables businesses to provide users with a real-time view of business performance by combining data residing in different sources, and inclusion is the first step towards transforming data into meaningful insights such as business metrics and key-performance-indicators (KPI) (Doyle, 2017).

Organizations use various means to extract data from source systems, integrate them, arrive at different business metrics, and store those KPI in a relational data warehouse that could drive further analysis. Organizations could use either ETL (*Extract-Transform-Load*) or ELT (*Extract-Load-Transform*) to extract data from the source systems. When using ETL, the data is extracted from source systems, transformational rules are applied to the data, all necessary business metrics are calculated, and the transformed information is loaded into the target database. In the case of ELT, the extracted data is loaded into a temporary work area called 'staging area', transformation rules are applied to arrive at business metrics, and the

transformed information or business metrics is loaded into the target data warehouse. To answer a specific query on the right use cases for using either of ETL or ELT, A3 claims,

> *"If the systems you extract data from (source systems) are all local (within your internet/intranet), then ETL would be good; there won't be much of latency issues here as the data is internal; the data you have on external systems – especially social media, etc – you extract data, you load them as-is into some sort of temporary store (data staging area) – and then transform and load/persist into data warehouse; (Use ELT) "* (A3, Data Architect)

ETL is the standard way of data extraction when sourcing from internal transaction processing systems. However, data from SM such as Twitter and Facebook get loaded first into a data store like NoSQL, necessary metrics such as sentiment scores are calculated, and the results are loaded into a relational data warehouse. SM data is usually extracted through scripts, software programs, and crawler programs.

An ETL or ELT process might throw exceptions or errors for data that cannot be processed completely. These exceptions are sent back to source systems, and alert messages are raised to business users. Business users view error records through user interfaces or applications wherein they would have options to fix the data, and resubmit for processing. The data extraction and load procedures maintain data integrity when loading data into target systems, and reconcile target data with source systems to load all relevant information into target systems. The data extraction process, depending on the business rules, could be set as an *incremental-refresh* or *full-refresh* procedures. In the case of *incremental-refresh*, new transactions since the last *incremental-refresh*, will get extracted from the source system, and loaded into the target system. The incremental transactions are usually identified through a timestamp field in transactions. In the case of *full-refresh*, the entire target data gets overwritten by the source data. The following quote summarizes how *incremental-refresh* and *full-refresh* data extraction procedures work:

> *"When you do data extraction, and if you know that the data (source data) is changing, then, you will have to set that as a full refresh; if the data doesn't change much, it has to be incremental refresh; if I have to quote example from the credit card expense transactions, the transactions happen on a daily basis, and no need to do full refresh (on those daily transactions); likewise, your banking transactions – no need to do full refresh as all the transactions like deposits, withdrawals, etc – are incremental in nature; and also, once you do a transaction you hardly change anything; maybe there is a credit onto your account, but that is a new transaction – which can again be incremental; full refreshes happen when you set up the target (data) for the first time; also – things like billing for one of your corporate clients: there, may be billing has to be done as a full refresh on a periodic basis – say 3 months or 6 months or even a year;"* (A3, Data Architect)

Data transformation is a process through which the data extracted from source systems is transformed to reflect the precise business metrics used for end-user analysis. The data transformation happens either at the source system or at the target system after the load to the staging area is completed. For data extraction and transformation, one could choose to transform at the source, provided the source systems are internal to the firm or available within the firm's intranet. Data extracted from SM and external sites is brought inside the firm's intranet, loaded into the staging area, transformed to arrive at metrics such as sentiment scores and product affinity scores, and loaded into a target data warehouse for further analysis.

Organizations could implement data transformation using in-memory techniques that are part of new data management solutions called NewSQL systems. NewSQL technologies offer in-memory computing features that support real-time analysis of data. In this case, the data that comes in real-time such as Internet-of-Things (IoT) sensor data and real-time Twitter feeds, could be loaded into the NewSQL database, and analysed on a real-time basis. The new generation analytics tools such as Spotfire and Tableau provide in-memory-computing features through which the entire data used for reporting could be cached into the reporting server, and cached data in memory would expedite information delivery and visualization. One of the subject matter experts has mentioned the following while discussing in-memory computing:

> *"It (In-memory analysis) is something that we typically use it for visualization; when you want to write a query and load lots of data; it is supposed to be available for visualization very quickly – then you can use In-memory; …In-memory computing will be beneficial for sporting, casino kind of businesses where the update should be in seconds and decisions should be taken in seconds."* (A1, Senior Data Architect)

Organizations store business metrics and KPI data in a data warehouse that is usually represented by a star-schema wherein the business-relevant KPI are stored in one or more *fact* tables (or a data store), and all reporting entities are stored as *dimensions*. For an organization, there would be an enterprise-wide data warehouse (EDW) that hosts business metrics relevant for all functional teams within the organization. In addition to this EDW, there would be many subject-area data warehouses one for each functional unit within the organization. These subject-area data warehouses are otherwise known as *data-marts*. The data-marts are built by extracting subject-specific data from EDW, or in some cases, an organization could build EDW from the underlying subject-specific data-marts. A1 has suggested,

> *"The other important thing is for all data warehouses to have subject areas (on their own); it would help in semantic analysis; you put customer data into 'customers' store, sales into a separate store; you can put that into (create) order, inventory; likewise, you can create subject areas, and keep the data into different subject areas; anybody who want to do customer analysis can easily do it from the subject-area data-mart (data base); it is always better to persist in subject-matter data store and do further analysis; "* (A1, Senior Data Architect)

Data transformation process is preceded with the pre-processing phase during which data from disparate systems get cleansed, validated and modelled to support transforming and loading the data into target database. The processes of extracting data from source systems, transforming data, and loading necessary business-relevant metrics into EDW are portrayed in Fig.2.

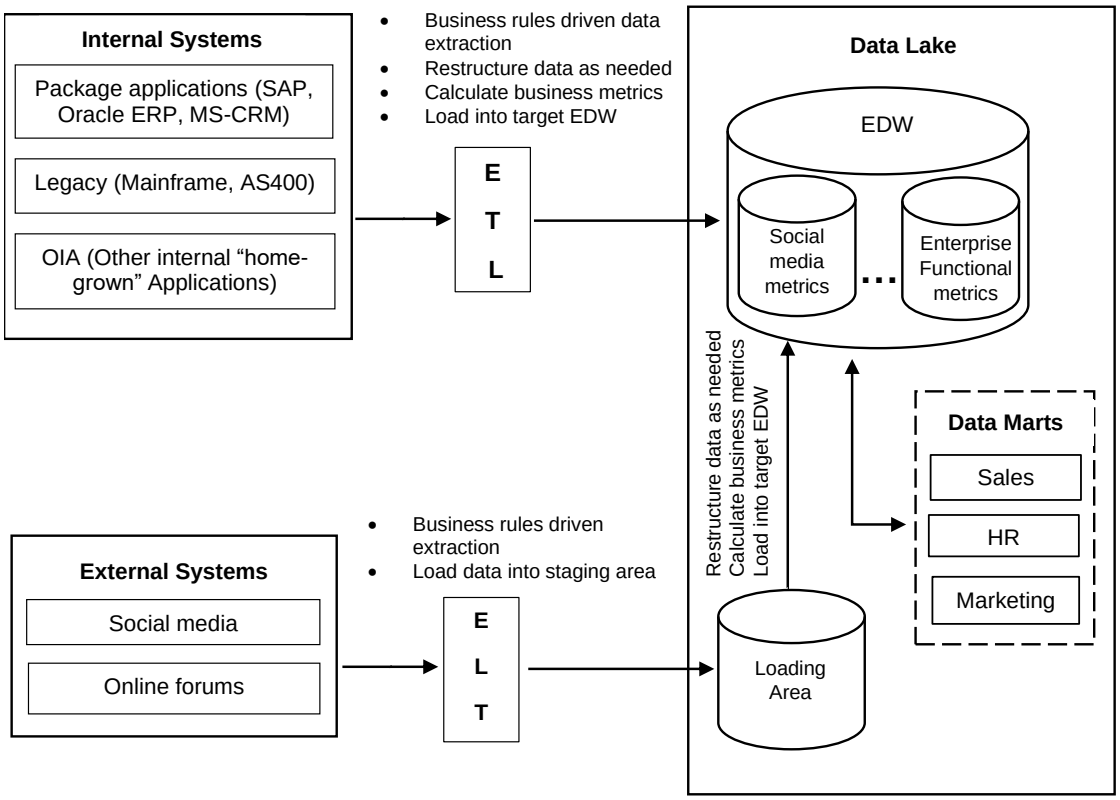

Fig.2 EDW built using ETL and ELT processes

The findings from this study are in line with inferences on big data extraction methods identified by Paakkonen and Pakkala (2015). In addition, this study provides inferences on extraction options for both internal and external systems, and outlines the need for data-staging to handle data from external systems.

4.3 Data Lakes and Dataspaces

Data lakes enable large-scale data storage at a relatively low cost. It is cheap and persistent storage that can support huge volumes of data (T. Davenport & Verma, 2018). The data stewards have to identify all possible sources from where the data has to be brought into the enterprise data lake. Once the causes are identified, the architects have to identify the right means to get the incremental and full dump of data from the source, and these mechanisms would include techniques such as writing a crawler job, developing a program to call API function to get data, configuring an ELT tool to extract data, and using vendor (source) provided adapters or scripts. The architects have to define an integrated semantic layer as mentioned by A2,

> *"to give a specific example – assume you are getting social media comments on your product, service quality and product ideation; these are three different things – or you could call it 'subject areas or domains'. When you bring it over (the data over to staging), you put it (the extracted data) on a staging (temporary) data store; a semantic layer integrating all these three domains have to be modelled – to be precise, integrating your products through product-id and relating those products to the comments you get on service quality (related to those products), and user feedback on new features that could be added to the product through product ideation.*

> *Building this integrated model is a complex process as the relationships are not defined in a concrete way; you will have to do lot of trial and error things (modelling); one need to create ontologies, tag-clouds for often-used words, classifications on the type of comments used, etc – to arrive at this integrated (semantic) layer; data analytics and your predictive stuff (analytics) can be driven off of this (integrated layer); all these things constitute your data lake"* (A2, Data and Solution Architect)

Hadoop based Big Data environments provide flexibility to organizations when dealing with high volumes of data from different source systems. Since systems supporting big data are hosted on low-cost commodity hardware, organizations could choose to save data from less popular social forums, newspapers and online magazines into a big-data environment, without worrying about how it could be leveraged for customer value creation. In the words of an architect,

> *"We always tell them (clients) – 'forget about what you want to do with data; Have a data-lake in Hadoop; any data, you come across – make sure that you store in Hadoop (Big Data environment)'; We are not saying that high-volume data only should go to Hadoop; All other data – streaming, etc -more like a 'Staging' area; Dump everything into one store; we then figure out a way to use it, build analysis and store the inferences;"* (A2, Data and Solution Architect)

All the subject matter experts handling data management solutions feel that data lakes could be used to archive all types of data that may or may not be used later. This would result in a data store that is humongous in nature, and hence retrieving data for further processing might be time-consuming and cumbersome in nature. Dataspaces would help overcome these issues. Dataspaces are not part of a data integration layer, and they provide base functionality over all data sources, regardless of how integrated they are (Franklin, Halevy, & Maier, 2005). Dataspace comprises of data models with a collection of relationships between data repositories. One of the basic dataspace services is cataloguing data elements from all participating data sources (Franklin et al., 2005). Dataspace platform should provide means for searching and querying the metadata. The metadata forms the basis for all retrieval operations, and the metadata in dataspace could be configured at various levels such as time-bound data, source-specific data, and certainty of data. Dataspace is considered as a hybrid of a database management system storing data from variety of sources, a search engine, a data-sharing system and an information integration system (Mirza, Chen, & Chen, 2010).

In the words of a team member from A2,

> *"Any data that you get into data lake, one has to create a model or a catalogue supporting all sources of data, where the data is retrieved from, time-stamp associated with data, all the meta-data associated with data like which user (if possible to identify) has created it, geo (geography) associated with the data, relationship between data elements from various sources – if any, etc. we could build more intelligence in the meta data by providing filtering elements such as the key tags that are being referred, the impact the data have – in terms of decision making, etc. all this could be part of the meta data ; when we retrieve data, usually you retrieve it through an app(application) layer; there, you query the meta-data to identify the right source we should query from, and initiate our retrieval; by having an effective cataloguing, you would be able to isolate which sources need to be queried, and also – how impactful the inference could be. If you don't have catalogue information or namespace kind of setup, querying from a data lake might*

> *be impossible as no one knows where to get it (your search results) from"* (Team-A2, Data and Solution Architect)

The inferences presented in this study that are related to data spaces is similar to the concepts suggested by Franklin et al. (2005) and Mirza et al. (2010). Additionally, this study emphasizes the need for having subject-area specific data stores along with catalogue and meta-data for each subject-area. Business Intelligence solutions built on this premise enable combining data across various subject-areas to arrive at valuable inferences that could drive business decisions.

4.4 Customer Value Creation

The organizations of the new digital era use data as a means by which they could create customer value proposition and leverage the same for monetization. A2 has opined,

> *"Data monetization is an area where every organization wants to get into; Data is very valuable and not that easy to get; now organizations look at – we have data now, and how they can make money out of it? Something we are looking into – of late is 'benchmarking'; benchmarking is very relevant for people in health-care industry; benchmarking gives an idea on where an organization stands with respect to various parameters that are standardized for the industry – and gives an indication of where the company should focus on; there are organizations in health-care industry, who collect data (from various firms), these people collect non-sensitive data, gender data, demographic data – and start creating benchmarks with data; the data can be federated and third-party people can bank on the data for benchmarking; that is one way of monetizing; this probably would fall into 'data-as-a-service' model as well.."* (A2, Data and Solution Architect)

One of the architects has opined on the data-enabled value creation saying,

> *"(by listening to data in SM and other channels), businesses get more informed in terms of what products have to be developed, and how they have to be positioned … if you provided targeted discounts, people tend to buy more; ..people tend to buy more as there is a discount which in turn increases sales. if you look at a data warehouse, it helps – and gives directions to salespeople on what to do (which product to develop), when and where (Positioning in different markets)?"* (A1, Senior Data Architect)

Firms have deployed self-service business intelligence tools for doing explorative data analysis. The information visualization tools support a variety of analysis such as *what-if analysis, scenario analysis, correlation analysis*, and enable business users to understand business KPI in a better way. The new generation data explorative tools are more tuned for business use than for IT use, and engage customers more efficiently than the previous generation reporting tools. The inferences on value creation and value delivery are aptly summarized by one of the interviewees:

> *"you provide value-add like self-service reports (to business users), real-time analysis, help customers in their product launch, product positioning, enhancing your target market reach - those are your monetization opportunities; internally, you use data-driven decisions to improvise your operational metrics; that is your cost saving opportunities; any data-driven initiatives (value propositions) help you on two fronts: open (enables) new avenues for revenue (additional revenue generation), and cost-saving opportunities by process optimization; and by doing your data management the right way – you get to benefit from both ends (revenue generation and cost cutting);"* (A3, Data Architect)

The findings from this study resonate with the ideas of using data-driven decisions for process optimization and value creation that are presented in various other studies. Furthermore, this study highlights options for data-enabled new value creation and revenue generation opportunities such as building *data-as-a-service* model.

4.5 Other Critical Considerations: Data Quality, Data Security and Business Continuity

The organization's ability to create a data-driven value proposition is severely impacted by low-quality data. Organizations strive hard to strengthen the quality of data used for building EDW. A slang expression '*Garbage in, Garbage out* (GIGO)' (Techopedia, 2011) refers to inaccurate results due to erroneous input data. Data governance mechanisms make sure that the data entered into both transactional systems and reporting systems are valid, valuable, safe and could be used effectively for decision making. One of the subject matter experts has to say this, when queried on the aspects of data governance:

> *"implementing master data management, cleansing rules for data, business rules for data validation, implementing maker-checker for enforcing transactional integrity, data backups and archival based on business rules - at stipulated time periods – all are part of data governance mechanisms; an organization has to look at these things holistically – to set up their governance mechanisms for ensuring 'clean data';"* (A3, Data Architect)

Organizations have to comply with stringent data security mandates, and one of the rules imposed by Sarbanes-Oxley (SOX) considers encryption as a critical security control, and one of the best practices, for both *data-at-rest*, and *data-in-transit* (McAfee, 2019). This implies that all types of data residing in enterprise systems within the organization, have to be encrypted. Data encryption happens at the database level or the hardware level. Organizations could choose to encrypt entire databases such as MongoDB or Oracle hosting transactional applications. The encryption at the database level is preferred over incorporating security at the code level because database-level encryption provides uniformity across all functional elements of the application. Data centres hosting servers running enterprise applications could be protected by the use of firewall software and access control mechanisms. For encrypting *data-in-transit*, firms could use a secured protocol such as *https* instead of *http*. For data-transfers from the source system to the target system, firms could use *secured-ftp* instead of regular *FTP*.

Organizations implement data replication to support recovery from data failures and business continuity. Data replication happens through two modes such as *active-replication* and *passive-replication*. The databases are replicated on a real-time basis during active-replication, and during passive-replication, databases are replicated when they are offline. Organizations usually keep primary server hosting data and a secondary, failover server in case of system failures. In the case of system failures or data loss, the fail-over server takes over automatically. Organizations define business rules for configuring data replication frequencies between primary and fail-over sites. For all critical transactional systems, replication happens frequently, or on a real-time basis. For all non-critical transactions, replication could occur during off-hours. The non-critical data such as facts and dimensions in an EDW are replicated less-frequently, usually once a day. In case of data loss, this incremental data in EDW could be built from the data in transactional systems. The objective of replication is to keep the data loss at a very minimal level.

5 Discussion

This study on data architecture related to using diverse data management solutions has inferences on using *data-access* concepts such as data virtualization, implementing alternate options for real-time analytics, and defining general guidelines for building a robust data architecture for enterprise data lake and EDW. These inferences are discussed at length in subsequent sections. The last section discusses the implications for both practitioners and academicians and the limitations of this study.

5.1 Options available for Data Extraction

While extracting data from source systems, organizations could choose either ETL or ELT as an option to transform and load data into target systems. For both ETL and ELT data extraction, organizations create necessary scripts or use *commercial-off-the-shelf* tools to complete the data extraction and load jobs. Creating these jobs is time-consuming, and one needs to go through rigorous testing to cover all possible data scenarios. One of the options available for an organization to extract data from source systems instead of setting up ETL or ELT jobs is to build data virtualization on top of underlying databases. Data virtualization integrates data from disparate source systems without replicating the data, to create a single *virtual data layer* that delivers unified data to multiple applications and end-users (Patrizio, 2019). When answering a query on specific scenarios for using data virtualization over ETL or ELT, A1 has opined,

> *"Data virtualization is a case where you want to do ad-hoc-queries (for reporting); when your business wants to do adhoc-querying, and you want to run something which is not already existing. They (business users) can look at the data, define some queries on it, and explore the data further; so, data virtualization helps with ad-hoc queries which are not existing already; and it adds value without adding a long ETL process…. I would rather say – use data virtualization for all PoCs (proof of concepts) that you want to run; find the value of it; and for better performance (if it needs to be repeated), you make it into ETL or ELT (for further extraction, processing, and analysis);"*

The suggestion is to use the data virtualization option to query directly from underlying systems when working on *proof-of-concept* or demo-version of applications. Organizations could choose to go with ETL or ELT jobs when they want to implement the proof-of-concept version as a full-fledged application. Data virtualization helps a firm to increase productivity when building EDW, brings down the cost as the data does not get replicated through ETL, and lessens the data governance needs (Patrizio, 2019).

5.2 Real-time analytics

Organizations could use streaming analytics or real-time analytics to query continuous data streams and detect conditions instantly, within a small period from the time the data is received (Perera, 2018). Real-time analysis is the right choice in scenarios where the data insights are more valuable only when the data is processed shortly after it is received, and the value diminishes fast with time (Perera, 2018). Some of the typical use cases for real-time analytics are applications such as providing real-time personalization options, detecting anomalies and frauds in real-time, providing real-time emergency health-care services, implementing smart device applications, and building algorithmic trading applications (Lawton, 2019). Options such as an e-commerce portal recommending a product to buy, or offering a discount based on the transactions one has made in the last few hours, are made

possible through *real-time analytics*. Organizations could use technologies such as NewSQL to drive real-time analytics. One of the architects has commented on the popularity and relevance of NewSQL systems,

> *"I have my apprehensions about NewSQL; if you look at the way it has come up, it has been there for almost 8 years now; 2011 is when it came first; – not much of adoption happened in the industry; that is what I have seen; even though it (NewSQL) happened, it happened in a small way; In places like ours (e-commerce portals), it is difficult to adopt (a tool or technology) which is not previously adopted or proven (by others); … the maturity of NewSQL is still a question;* (A1, Senior Data Architect)

The architects feel that NewSQL systems handle real-time analysis at a higher cost compared to other streaming analytics tools that are open-source or offered at a significantly lower cost. One of the subject matter experts used for this study, feel that tools such as Kafka, provided by Apache Foundation, could listen to messages streaming from SM, and synchronously do analysis such as trends and sentiments. Information visualization tools could directly connect to streaming analytics tools and plot real-time reports on inbound comments. The messages coming through Kafka are cached in Kafka for a duration that could be configured. Developers could write scripts to move the streaming messages coming through Kafka, and persist them in a data-store if needed (Yang, 2016). A1 has opined,

> *"you will have a lot more channels Kafka listens at (to); you could push messages to various channels – the product could be one, and so on. … If they (firms) want to do that (real-time analysis), we will have (a tool like) Tableau listening to Kafka; then, you can do real-time analytics; So, Kafka would be your data source for Tableau; It is like messaging queue; the data gets persisted for a period of time (in Kafka); then, you could refresh that also; if you do not want to keep that data (in Kafka) for a long time, you could set the configuration to 'delete the messages', or you could move the messages (to your data warehouse or transactional systems) also"* (A1, Senior Data Architect)

The big data and data management reference architecture depicted in Fig.1 covers all the components of EDW and data lake. The *streaming-analytics* part in Fig.1 is explained in Fig.3, which contains high-level architecture and process flow associated with Kafka enabled stream-processing.

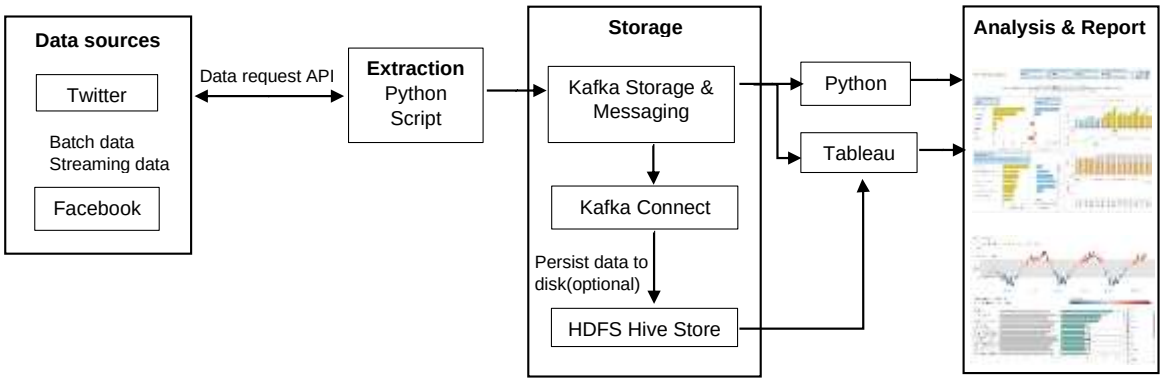

Fig.3 Stream Processing using Kafka – High-level Architecture and Process Flow

5.3 Guidelines for building a robust data architecture

Organizations have predicaments on the validity of the data management solutions discussed in this paper, and the chances of these solutions becoming obsolete. The solutions mentioned in this study such as RDBMS, NoSQL, and NewSQL or real-time analysis tools, complement each other and will continue to co-exist in an enterprise. The choice of using an appropriate data management solution is dependent on the type of data that needs to be managed, the way the data has to be integrated and transformed, and the method the data has to be analysed. In the interest of time, effort and cost, organizations tend to standardize the usage of data management solutions. The standardization of solutions is considered a good initiative, but it may not be the right choice. As one could notice, data keeps on evolving. As the data evolves, the data management solutions have to evolve as well. With that, the adage of *one-size-fits-all* would not work anymore, when discussing data management solutions.

It is suggested that organizations should extract all types of data available on external systems, load them into a low-cost Big Data setup, regardless of whether it could be leveraged or not. When organizations arrive at a need to use the saved data from external sources, they could implement processes to transform the *staged data* into meaningful business metrics and persist into an EDW. Organizations have to be mindful of data quality, as the data quality determines the effectiveness of data analytics. Organizations have to implement a business-rule driven data cleansing, data profiling, and data validation procedure to make sure that the data is *analysis-ready* before it is transformed and loaded into EDW. For transactional systems, to strengthen the data quality, organizations could implement *maker-checker*, a widely implemented, successful process implemented in the financial services industry (Commercial Bank, 2018). This process mandates a transaction entered by a *maker* to be verified and approved by a *checker*, to make sure that the data is valid and useable.

The data extracted from external source systems such as SM, forums, blogs, and newsletters get inserted into data lakes, and data governance enables the data to be tagged. In other words, data governance mechanisms warrant architects to catalogue the data that gets into data lakes by building a meta-data repository. The data has to be governed in terms of ownership, accountability, sharing, and usage, to avoid accumulating unwanted data in the form of data swamps (Ghosh, 2019).

The organization's success in building a robust data management system depends on implementing solutions such as master data management (MDM) solution, implementing data privacy, and maintaining data portability to other environments when needed (Addagada, 2019). Master data is defined as an organization's core data, containing the necessary information that is critical to run the business, and the data is relatively stable in nature (Foote, 2019a). Some of the master data domains include customers, products, and locations. In the words of A1,

> *"the centralized master data – used by all systems in the enterprise; generally, one (the firm) stores only one instance of master data – that gets replicated or federated to other systems; the data in master data should reflect 'the single version of your most important entities in the enterprise – like customer and part/product details'. A well-defined master data of customers and products can enable the firm to get a unified view of customer-wise product sales across various channels – in an effective way in very short span of time; MDM increases data analytics efficiency"* (A1, Senior Data Architect)

As the type of data evolves, data modelling has become even more important (Foote, 2019b). The benefits associated with modelling NoSQL data include improved data quality, lot more identifiable information, and enhanced business intelligence. Architects could start with simple models, and make the models agile, keeping them evolve as new data sources and types of data get added into data lakes (Foote, 2019b). In the words of A2 and his team,

> *"Data models are very important regardless of the data you use; all of us are familiar with OLTP and OLAP related data models (normalization and relational databases); now, with all the evolving database types, we need to constantly add meta-data repositories and models to define and hold that data; this (modelling NoSQL data) will play a critical role when you do analysis; to give an example, if you want to monitor twitter feeds for a given trending topic, the data could be stored in a relational model that store, topic-id, topic-name, comment or tweet, and some meta-data on the person posted the tweet including the date-time stamp; once, you have it in this form, assuming that it is saved in BigData Hadoop cluster, querying this data using tools like Hive would be much easier; you could query, do your aggregates and analytics on top of this; as and when new types and sources evolve, one should keep adding flexible data models in the form of your meta-data, and also tables to host your data.. this is very important to have a control over what comes into your system"* (A2, Data and Solution Architect)

Good data governance mechanisms, coupled with data management best practices contribute to the successful implementation of data-driven value creation. The data governance issues related to data security, data privacy, and data quality could be managed effectively through data-stewardship, the right set of processes for extraction, transformation and loading, and the right tools and technologies. In this section, as mentioned earlier, the aspects to related to technology governance are included, and mechanisms related to people and process aspects of governance are not dealt with.

5.4 Implications and Limitations of this study

This study has implications for both practitioners and academicians. For the practitioners, this study provides an overview of building an EDW by encompassing data from disparate sources and external systems. This study provides insights to managers on the choices of data management solutions available, and the ideal use cases for the same. With this information, managers would be able to choose the right data management tool for a specific business purpose. This study provides inferences to academicians on the methods used for creating value propositions from data available in internal systems and external systems. Academicians would be able to understand the components involved in building an EDW, and the use of various data management solutions. This study provides extensive coverage of the processes needed for creating an EDW. The subject matter experts used for the study have offered insights on generic processes used in data analysis. However, the insights offered on data architecture are oriented more towards practitioners. This study proposes research to be conducted in academic settings, to understand the concepts and theory behind implementing data management solutions. This suggested research would be able to bring perspective on the nature of transactions, properties of data conducive for certain analysis, extraction and transformation methods, and options available for analytics.

6 Conclusion

Many of the findings from this study are already available as part of technical white papers, research articles, and conference proceedings. However, in this paper, the researchers have

attempted to collate all those findings to arrive at a reference architecture for all types of data management solutions. The critical contribution from this paper is to provide a consolidated reference material that would provide input to practitioners on: a) using the right technology and methods for data analysis, b) implementing the right use cases for various data management technologies, c) recommending a reference architecture for managing various data management solutions and design considerations for implementing the solution, d) providing a literature review and overview of all types of data management solutions, their pros and cons, and how they are part of the data management eco-system, and e) providing a consolidated view on various technologies used for data management such as data lakes, dataspaces, virtualization, and real-time monitoring tools.

This study has approached the journey of an organization towards data-enabled value creation through four levels of data processing, such as data extraction, data transformation, data exploration or information visualization, and value creation. The study has critical observations with respect to using RDBMS, NoSQL, and in-memory data management solutions. The query on organizations supporting a variety of data management solutions, and the recommended data architecture for supporting the same, is aptly answered by one of the subject matter experts. The entire process of building an EDW that enables further data analysis is succinctly quoted here:

> *"The traditional data in various transactional systems, you store in relational databases; you have data from social media, and whatever other external sources – you want to bring it over to a NoSQL kind of environment which is hosted on a Big Data environment, or, you could use any other form of Big Data systems (systems supporting big data - provided by various third parties), and then you use some level of processing and transformation to take all metrics that are relevant for your business needs, and then bring it (those metrics and KPI) into a structured kind of data warehouse, and you drive your information visualization (reporting and analysis) on top of that"* (A1, Senior Data Architect)

Organizations could create value by providing real-time analysis on data retrieved from SM, and an architect, when discussing the technology used for real-time analytics, has mentioned,

> *"Any data that does not have to be persisted, and you need real-time analysis, then go for NewSQL; here, what you do is – you stream the data into NewSQL in-memory database, enable all your analysis on the same; … NewSQL is not proven, and a bit expensive, and you have other options available as well – for doing real-time analysis (Like Kafka)"* (A3, Data Architect)

Organizations manage the data-enabled value creation through governance mechanisms implemented in a holistic way to address challenges associated with data security, data privacy, data loss and recovery, and business continuity. The design considerations should include elements such as encryption at the database and hardware-level for data-at-rest, security through firewall at the data-centre level, data transfer over secured protocols for data-in-transit, access controls at various levels, scheduled back-ups for data, data storage through both primary and fail-over instances, and data replication through active and passive modes. Organizations have to implement robust information-governance mechanisms across data handling processes to achieve higher level of quality. One of the architects has commented on the aspect of data governance,

> *"' clean and good data' enables good metrics, and good metrics enable great inferences – and hence, these things (data governance mechanisms and best practices) are very critical for*

> *implementing quality data management solutions that would drive your business;"* (A3, Data Architect)

In summary, data is considered as the gold of the 21st century (Aviza, 2017), and using it the right way would yield economic returns at an epic scale. Organizations could harvest the value embedded in data by implementing the right data management technologies, coupled with appropriate information governance mechanisms.

Acknowledgements

We thank our interviewees Rajendran Subramanian, Narayanan Kandanchatha, Mukundakumar KG, Brijesh Madhavan, Shiju Jalaludin, Murali Avanamuthu, and their team members for sharing their pearls of wisdom during the course of this research. We are immensely grateful to Dr. S. Varadhaganapathy for helping us in framing our research queries. We extend our sincere thanks to Latha Venkitachalam and Baiju Ambadan for providing comments that have greatly improved the manuscript. Please note that any errors in the manuscript are our own, and the errors should not tarnish the reputations of these esteemed persons.

References

Addagada, T. (2019). *Customer Data Protection: Deriving Value and Ownership*. Retrieved May 24, 2019, from https://www.dataversity.net/customer-data-protection-deriving-value-and-ownership/

Anderson, C., & Li, M. (2017). *Five building blocks of a data-driven culture*. Retrieved May 27, 2019, from https://techcrunch.com/2017/06/23/five-building-blocks-of-a-data-driven-culture/

Aviza, E. (2017). *Data is the Gold of the 21st Century*. Retrieved May 20, 2019, from https://www.cloudbakers.com/blog/data-is-the-gold-of-the-21st-century

Bowen, R., & Smith, A. R. (2014). Developing an enterprisewide data strategy. *Healthcare Financial Management*, *68*(4).

Braganza, A. (2004). Rethinking the data – information – knowledge hierarchy : towards a case-based model. *International Journal of Information Management*, *24*, 347–356. https://doi.org/10.1016/j.ijinfomgt.2004.04.007

Commercial Bank, C. G. (2018). *Why Dual Approval Matters*. Retrieved May 6, 2019, from https://businessaccess.citibank.citigroup.com/basprod/citiiwt/images/Why_Dual_Approval_Matters.pdf

Data Integration. (2018). Retrieved May 20, 2019, from https://www.dataintegration.info/data-integration

Davenport, T. H. (2014). *Big Data at work: Dispelling the myths and uncovering the opportunities*. *Harvard Business Review Press*.

Davenport, T. H., Harris, J. G., Long, D. W. De, & Jacobson, A. L. (2001). Data to Knowledge to Results: Building an Analytic Capability. *California Management Review*, *43*(2).

Davenport, T., & Verma, A. (2018). *It's time to modernize your big data management techniques.* Retrieved May 18, 2019, from

https://www2.deloitte.com/us/en/insights/topics/analytics/data-management-techniques-approaches-tools.html

Doyle, M. (2017). *The Importance of a "Data Integration First" Strategy*. Retrieved May 15, 2019, from https://www.dqglobal.com/2017/07/25/importance-data-integration/

Foote, K. D. (2019a). *A Brief History of Master Data*. Retrieved May 12, 2019, from https://www.dataversity.net/a-brief-history-of-master-data/

Foote, K. D. (2019b). *Data Modeling in an Agile World*. Retrieved May 15, 2019, from https://www.dataversity.net/data-modeling-in-an-agile-world/

Franklin, M., Halevy, A., & Maier, D. (2005). From databases to dataspaces: A new abstraction for information management. *SIGMOD Record*, *34*(4), 27–33. https://doi.org/10.1145/1107499.1107502

Ghosh, P. (2019). *Data Governance and Data Quality Use Cases*. Retrieved May 12, 2019, from https://www.dataversity.net/data-governance-and-data-quality-use-cases/

Griffin, J. (2005). Data governance: a strategy for success. *DM Review*, *15*(8), 15, 70.

Grolinger, K., Higashino, W. A., Tiwari, A., & Capretz, M. A. M. (2013). Data management in cloud environments : NoSQL and NewSQL data stores. *Journal of Cloud Computing*. https://doi.org/https://doi.org/10.1186/2192-113X-2-22

Hendler, J. (2009). Web 3.0 Emerging. *IEEE Computer Society*, *42*(January), 111–113.

Huber, G. P. ., & Power, D. J. . (1985). Retrospective Reports of Strategic-Level Managers : Guidelines for Increasing Their Accuracy. *Strategic Management Journal*, *6*(2), 171–180.

Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. *Journal of Parallel and Distributed Computing*, *74*(7), 2561–2573. https://doi.org/10.1016/j.jpdc.2014.01.003

Kaur, K., & Sachdeva, M. (2017). Performance Evaluation of NewSQL Databases. In *International Conference on Inventive Systems and Control* (pp. 1–5). https://doi.org/10.1109/ICISC.2017.8068585

Khatri, V., & Brown, C. V. (2010). Designing Data Governance. *Communications of the ACM*, *53*(1). https://doi.org/10.1145/1629175.1629210

Kooper, M., Maes, R., & Lindgreen, R. E. (2011). Information Governance as a Holistic Approach to Managing and Leveraging Information Prepared for IBM Corporation. *International Journal of Information Management*, *31*.

Korhonen, J. J., Melleri, I., Hiekkanen, K., & Helenius, M. (2013). Designing Data Governance Structure : An Organizational Perspective. *GSTF Journal On Computing*, *2*(4), 11–17. https://doi.org/10.5176/2251-3043

Lawton, G. (2019). *7 enterprise use cases for real-time streaming analytics*. Retrieved May 20, 2019, from https://searchbusinessanalytics.techtarget.com/feature/7-enterprise-use-cases-for-real-time-streaming-analytics

Lim, C., Kim, K., Kim, M., Heo, J., Kim, K., & Maglio, P. P. (2018). From data to value : A nine-factor framework for data-based value creation in information-intensive services.

International Journal of Information Management, *39*(January 2017), 121–135. https://doi.org/10.1016/j.ijinfomgt.2017.12.007

Link, S., & Prade, H. (2019). Relational database schema design for uncertain data Relational Database Schema Design for Uncertain Data $. *Information Systems*. https://doi.org/10.1016/j.is.2019.04.003

Lourenço, J. R., Cabral, B., Carreiro, P., Vieira, M., & Bernardino, J. (2015). Choosing the right NoSQL database for the job: a quality attribute evaluation. *Journal of Big Data*, *2*(1), 1–26. https://doi.org/10.1186/s40537-015-0025-0

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation , competition , and productivity. *McKinsey Global Institute*, (May).

Marco, D. (2006). Understanding data governance and stewardship, Part 1. *DM Review*, *16*(9), 28.

Mazzei, M. J., & Noble, D. (2017). Big data dreams : A framework for corporate strategy. *Business Horizons*, (60), 405–414.

McAfee. (2019). *Overview of Serbanes-Oxley*. Retrieved May 20, 2019, from https://www.skyhighnetworks.com/cloud-compliance/sarbanes-oxley-encryption-compliance-requirements/

Mirza, H. T., Chen, L., & Chen, G. (2010). Practicability of dataspace systems. *International Journal of Digital Content Technology and Its Applications*, *4*(3), 233–243. https://doi.org/10.4156/jdcta.vol4.issue3.23

Mohan, C. (2013). History Repeats Itself : Sensible and NonsenSQL Aspects of the NoSQL Hoopla. *IBM Alamaden Research Center*, 11–16.

Newman, D., & Logan, D. (2009). Governance Is an Essential Building Block for Enterprise Information Management. *Gartner Research*, (May 2006).

Pääkkönen, P., & Pakkala, D. (2015). Big Data Research Reference Architecture and Classification of Technologies , Products and Services for Big Data Systems. *Big Data Research*, *2*(4), 166–186. https://doi.org/10.1016/j.bdr.2015.01.001

Patrizio, A. (2019). W*hat is Data Virtualization?* Retrieved May 20, 2019, from https://www.datamation.com/big-data/what-is-data-virtualization.html

Pavlo, A., & Aslett, M. (2016). What is really new with NewSQL ? *SIGMOD Record*, *45*(2), 45–55.

Perera, S. (2018). *A Gentle Introduction to Stream Processing.* Retrieved May 20, 2019, from https://medium.com/stream-processing/what-is-stream-processing-1eadfca11b97

Pokorny, J. (2013). NoSQL databases: A step to database scalability in web environment. *International Journal of Web Information Systems*, *9*(1), 69–82. https://doi.org/10.1108/17440081311316398

Rocha, L., Vale, F., Cirilo, E., Barbosa, D., & Mourao, F. (2015). A Framework for Migrating Relational Datasets to NoSQL ∗. *Proceedia Computer Science*, *51*, 2593–2602. https://doi.org/10.1016/j.procs.2015.05.367

Salido, J. (2010). Data Governance for Privacy, Confidentiality and Compliance: A Holistic Approach. *ISACA Journal*, *6*, 1–7.

Sareen, P., & Kumar, P. (2015). NoSQL Database and its comparison with SQL Database. *International Journal of Computer Science & Communication Networks*, *5*(5), 293–298.

Simsek, G. (2019). *What is new about NewSQL?* Retrieved June 7, 2019, from https://softwareengineeringdaily.com/2019/02/24/what-is-new-about-newsql/

Spivack, N. (2011). *Web 3.0: The Third Generation Web is Coming*. Retrieved May 18, 2019, from https://lifeboat.com/ex/web.3.0

Stantic, B., & Pokorny, J. (2014). Opportunities in Big Data Management and Processing. *Databases and Information Systems VIII*. https://doi.org/10.3233/978-1-61499-458-9-15

Techopedia. (2011). *Garbage In, Garbage Out (GIGO).* Retrieved May 15, 2019, from https://www.techopedia.com/definition/3801/garbage-in-garbage-out-gigo

Weber, K., Otto, B., & Osterle, H. (2009). One Size Does Not Fit All — A Contingency Approach to Data Governance. *ACM Journal of Information Quality*, *1*(1). https://doi.org/10.1145/1515693.1515696.http

Weill, P., & Ross, J. (2005). A matrixed approach to designing IT governance. *MIT Sloan Management Review*, *46*(2), 26–34.

Yang, F. (2016). *Building a Streaming Analytics Stack with Apache Kafka and Druid*. Retrieved May 20, 2019, from https://www.confluent.io/blog/building-a-streaming-analytics-stack-with-apache-kafka-and-druid/

Zack, M. H. (1999). Managing Codified Knowledge. *Sloan Management Review*, *40*(4), 45–58.

doi: https://doi.org/10.3127/ajis.v24i0.2541