


Article

Leveraging the Organisational Legacy: Understanding How Businesses Integrate Legacy Data into Their Big Data Plans

Sanjay Jha ^{1,*}, Meena Jha ¹, Liam O'Brien ², Michael Cowling ^{3,*}  and Marilyn Wells ⁴¹ School of Engineering and Technology, Central Queensland University, Sydney 2000, Australia; m.jha@cqu.edu.au² Department of Home Affairs, Government of Australia, Belconnen 2617, Australia; liamob99@hotmail.com³ School of Engineering and Technology, Central Queensland University, Brisbane 4000, Australia⁴ School of Engineering and Technology, Central Queensland University, Rockhampton 4702, Australia; m.wells@cqu.edu.au

* Correspondence: s.jha@cqu.edu.au (S.J.); m.cowling@cqu.edu.au (M.C.)

Received: 19 May 2020; Accepted: 17 June 2020; Published: 23 June 2020



Abstract: Big Data can help users attain a competitive advantage, and evidence suggests that by utilising Big Data, organisations can generate insight that can help strengthen their decision-making capabilities. However, a key issue remains that much data is trapped in legacy systems, and is hence not being appropriately retrieved and utilised. This paper builds on the existing literature base to investigate the challenges and issues organisations face in utilising Big Data. Through results of a survey with 97 respondents, this work shows that these issues can be categorised into six areas, including issues of format and structure of the data, as well as identification of the key need for a framework and architecture for organising Big Data.

Keywords: Big Data approaches; Big Data analytics; legacy systems; survey; decision making

1. Introduction

By 2025, the International Data Corporation (IDC) predicts there will be 163 zettabytes of data [1]. By collecting and analysing this data, businesses can understand their customers better and make better decisions. In fact, organisations adopting Big Data solutions have been shown to have gained a significant competitive advantage [2]. The potential of Big Data is great; however, there remain several challenges to overcome [3]. Organisations are investing more in data manipulation, however, they are still not processing and using stored data to its fullest. In particular, data stored in their data warehouses and data marts are not being retrieved and utilized in a proper way [4]. Data marts are a subset of data warehouse used to retrieve user specific data. Data in data marts and data warehouses are organised for ready retrieval, however, these do not support retrieval of unstructured data.

Zou et al. [5] conducted a survey of Big Data analytics for smart forestry and established the need for Big Data technology in forestry. In their work, they noted that Big Data will bring greater opportunities for forestry development as the speed and accuracy of forestry data acquisition have been greatly improved with the development of technology. However, the authors pointed out the challenges of reasonably and effectively organizing the massive amount of data and analysing it quickly. This relates to the technological problems that organisations are facing in accessing and implementing Big Data.

Similarly, Inoubli et al. [6] conducted an experimental survey of existing Big Data frameworks and provided an experimental evaluation with several representative batch and iterative workloads. They compared the Big Data frameworks Hadoop, Spark, Storm, Flink, and Samza based on data

formats, processing mode data sources, programming modes, supported programming languages, cluster management, comments, iterative computation, interactive mode machine learning capability, and fault tolerance, and provided a list of best practices for batch and stream processing. The authors found that it is difficult to organize the Big Data landscape.

Despite these attempts to utilize data in new ways, evidence suggests that organisations traditionally only use in-house data for decision making [7]. Further, in-house data is generated from legacy systems that cannot evolve with the changing requirements of the organisation. Business Intelligence and Data Analytics (BI & DA) based only on internal data from legacy systems does not provide complete insight into business problems, even though analysing large data sets and extracting meaning from them can help the organisation in building informed decisions and gaining competitive advantage [8].

Organisations use a variety of systems to collect data, including data from enterprise resource planning systems, in-house information systems, decision support systems, and many others to support their day-to-day activities and operations. Once analysed, this data can provide more insight into how businesses perform and has the potential to support more accurate and effective business decisions. Specifically, an organisation's operational capability is based on existing information systems. These existing information systems are called legacy systems [9]. Legacy systems contain significant and invaluable business logic of the organisation. These legacy systems are old processes, technology, computer systems, or application programs that continue to be used, typically because they still function for the users' needs, even though newer technology or more efficient processes of performing a task are now available [10].

What would appear to be needed, therefore, is a model of how this legacy data can be understood in the context of Big Data systems, a sort of guidance plan for industry to use to understand the existing data they have and how it might fit into a Big Data solution. A survey of the literature would indicate that this model is still being sought, with different researchers using different classification schemes to understand the components of a Big Data system and how they interact with organisations.

For instance, Liu et al. [11] conducted a survey of real-time processing systems for Big Data based on open source technologies. Their survey focused on system architectures and system platforms used for real-time/near real-time processing. They identified that "due to the nature of Big Data, it has become a challenge to achieve the real-time capability using the traditional technologies", such as a Relational Database Management System and legacy systems where data is stored in a row and column format.

Achariyya et al. [12] conducted a survey of Big Data analytics to identify the potential impact of Big Data challenges, open research issues, and the various tools associated with Big Data. The authors categorized Big Data challenges into four broad categories, namely: data storage and analysis; knowledge discovery and computational complexities; scalability and visualization of data; and information security. They also discussed Big Data tools such as Apache Hadoop and MapReduce, Apache Mahout, Apache Spark, Dryad, Storm, Apache Drill, JasperSoft, and Splunk.

Oussous et al. [13] conducted a survey of Big Data technologies to facilitate the adoption and the right combination of different Big Data technologies according to their technological needs and specific applications' requirements. The authors categorized the tools according to different layers, such as the data storage layer, data processing layer, data querying layer, data access layer, and management layer. They identified the tools that can be used in each layer. This brings an understanding of the tools and where they can be used in a Big Data solution.

Khan et al. [14] conducted a survey on technologies, opportunities, and challenges associated with Big Data. The authors noted that difficulties lie in data capture, storage, searching, sharing, analysis, and visualization. According to the authors, the architecture of Big Data must be synchronized with the support infrastructure of the organisation. The authors claim that, to date, all data used by organisations are stagnant. However, data are increasingly sourced from various fields that are

disorganized and messy, such as information from machines or sensors and large sources of public and private data. It is pertinent to have these data for decision making.

Tsai et al. [15] conducted a survey of Big Data analytics to develop a high-performance platform to efficiently analyse Big Data. The authors work is more aligned to selecting data mining algorithms to handle Big Data analytics and relates to selection, pre-processing, transformation, data mining, and interpretation/evaluation. Their work is based on knowledge discovery in databases (KDD) and its operations. The authors discussed three main processes, namely, input, data analytics, and output. These processes work on seven operators: gathering, selection, pre-processing, transformation, data mining, evaluation, and interpretation. Big Data analytics systems are designed to work on parallel computing, with other systems which can be hosted on the cloud, or on another search engine. The communication between the Big Data analytics and other systems will strongly impact the performance of the whole process of input, data analytics and output. In practice, assuming there are infinite computing resources for Big Data analytics, which is a thoroughly unrealistic assumption, the input and output ratio (e.g., return on investment) will need to be considered before an organisation constructs their Big Data analytics centre.

Praveena-Anto and Bharathi [16] conducted a survey and provided an overview of Big Data analytics and the issues, challenges, and various technologies related to Big Data. The authors provided a Big Data architecture with different layers, including the technologies to be used in different layers. The Big Data challenges reported related to storage; data representation; data life cycle management; analysis; reporting; redundancy reduction and data compression; energy management; data confidentiality; expendability; scalability; cooperation; and Big Data dimensional reduction.

Overall, the literature [5,6,11–16] shows that there are existing surveys which discuss Big Data challenges and issues. However, authors have not identified surveys that list how organisations are integrating Big Data solutions with their legacy systems, and the challenges and issues they are facing in doing so. In essence, rather than identifying a guidance plan or model based on the data requirements, why not instead identify these aspects based on discussions with the organisations themselves, using the existing models as a guide? With this in mind, in this paper, the research questions addressed are:

- RQ 1: What are the current issues and challenges with the integration of Big Data solutions into legacy systems?
- RQ 2: What are the practices related to the use of Big Data solutions with legacy systems?
- RQ 3: What strategies does an organisation need to adapt before implementing Big Data solutions within their organisations?

The remainder of the paper is organised as follows. In Section 2, we present materials and methods highlighting the research methodology and survey motivation of our work. In Section 3, results from the survey we conducted are discussed. In Section 4, we present discussion and analysis of our survey, and concluding points are given in Section 5.

2. Materials and Methods

We used a combination of qualitative and quantitative methodologies to answer our research questions. This approach was well suited to the study as it allowed identification of both the strengths and weaknesses of integrating Big Data solutions into legacy systems and data while addressing our research questions. We designed our survey to provide a “snapshot of how things are at a specific time” [17]. More importantly, the data collected allowed us to answer the research questions. In this study, we filtered results by cross-tabulating subgroups, interrogating the data, and analysing the results, and drew conclusions using Excel because of the nature of the data.

Our survey of existing approaches for Big Data integration with legacy systems and data was conducted between 2018 and 2019. The survey had 97 ($n = 97$) respondents from different organisations with responses from industries such as financial services, healthcare, aviation, higher education,

energy sector, and insurance. Table 1 shows the total number of survey respondents. The survey questions were developed in discussion with Big Data experts and people in industries where there is a need for a Big Data solution, and based on literature [5,6,11–16] that indicates integrating a Big Data solution is a challenge for organisations. The table shows that the energy sector and aviation were represented by only 4 organizations each.

Table 1. Survey respondents in percentage terms.

Organisations	Total Numbers	Percentage
Research Organisations	6	6.18%
Service Providers	25	25.77%
Higher Education	22	22.68%
Financial Organisations	15	15.46%
Energy Sector	4	4.12%
Supermarket Organisations	5	5.14%
Insurance	9	9.27%
Aviation	4	4.12%
Others	7	7.21%

The survey was divided into three subsections with 31 questions in total:

- Section 1: Big Data Organisational and General Questions (1–7). This section captures answers to find out if Big Data projects are running within the organisation or not. This section also captures answers on the educational level, role, and experience level of the respondents.
- Section 2: Legacy Systems Issues and Concerns Questions (8–13): This section captures answers about legacy systems and data, advantages, disadvantages, and factors influencing legacy systems and data in the business intelligence community.
- Section 3: Big Data Initiatives and Implementation Questions (14–31). This section captures answers about finding the proper fit of a Big Data solution and technologies for an organisation based on characteristics of Big Data. This section also gathered information regarding Big Data activities, in particular, the integration of a Big Data solution with legacy systems. The survey was targeted specifically where Big Data could be of use.

3. Results

3.1. Big Data Organisational Questions

To understand how our respondents perceived Big Data and its implementation, we needed to know the respondent's role in the organisation. The statistics show that we had a good cohort of respondents having important roles in the surveyed organisations around areas such as data analytics, decision making, and business intelligence. Figure 1 shows the respondents' roles in percentage terms and their educational qualifications. The survey results show that the years of experience of the respondents ranges from 10 to 20 years. Furthermore, the experience of the respondents indicates a more mature population that is working in the area of legacy systems and Big Data analytics and is relatively well educated. The mature and experienced population comprising the respondent group means that, due to its breadth of coverage of understanding how businesses integrate legacy data into their Big Data plans, the group constitutes a representative sample, and can therefore be generalizable to a population.

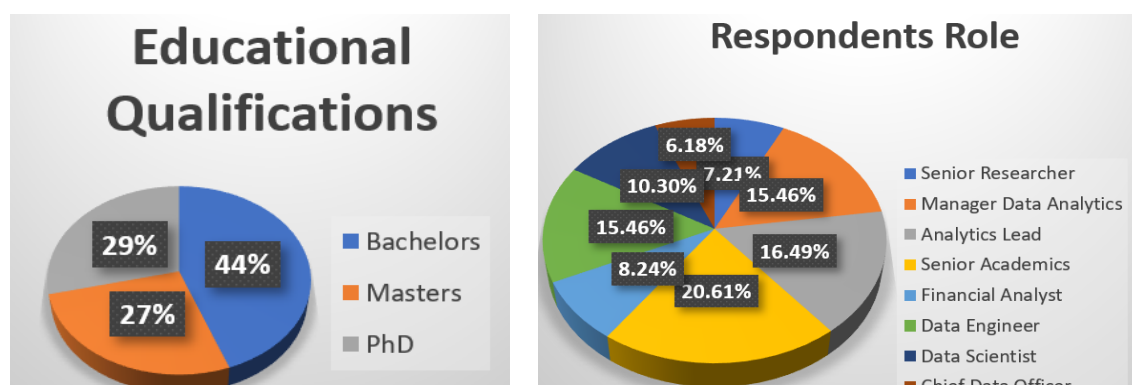


Figure 1. Respondents' roles in organisations.

A share of 27.83% of the respondents believed that Big Data projects were running in their organisation; the remaining 72.16% of the respondents believed that they were not running any Big Data projects in their organisation. Constituting the 27.83% of organisations running Big Data projects were financial organisations, research organisations, insurance companies, and some of the service providers. Figure 2 shows the organisations running Big Data projects were in the sectors: Research Organisations; Service Providers; Financial Organisations; Insurance and Aviation.

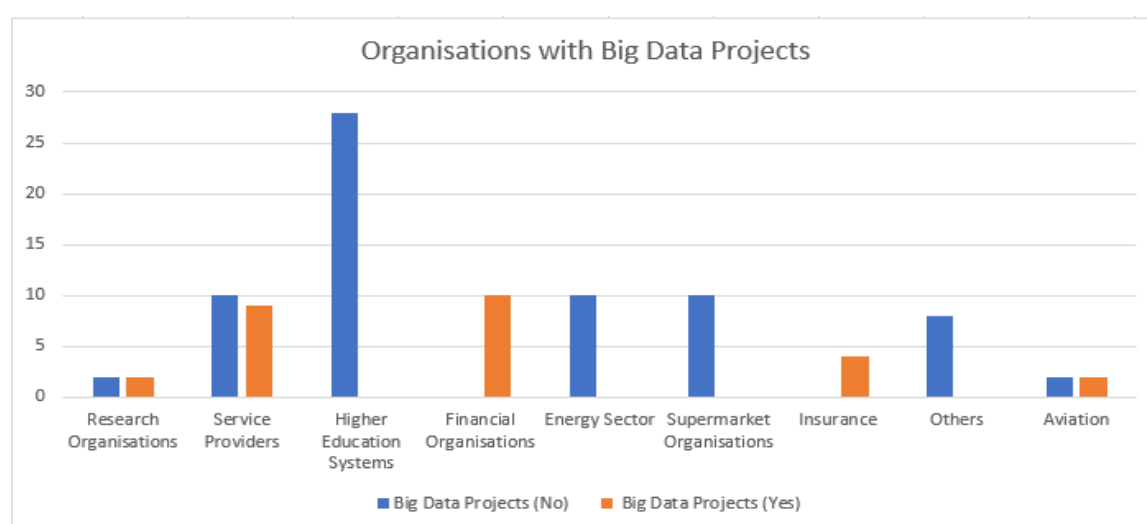


Figure 2. Organisations with Big Data projects.

3.2. Legacy Systems Issues and Concerns Questions

This section captures answers about legacy systems and data, and the advantages, disadvantages, and factors influencing legacy systems and data in the business intelligence community. The following provides a description of the survey results.

What kind of data does your organisation have?

Most organisations we surveyed were working on organisational transactional and historical data. All of the respondents from financial organisations believed that they had good practices in place to manage data, processes, and infrastructure for detecting fraud. Similarly, all of the respondents believed that data was the biggest asset of their organisation. A share of 48% of the respondents believed that the data generated within the organisation fitted into the characteristics of Big Data and if analysed effectively can provide a competitive edge to the organisation. The drawback was that the existing technology inhibited the organisation from exploiting this volume of data. Our survey identified that the kinds of data used include structured data, unstructured data, sensor data, log files

data, big data, time stamped data, machine data, spatiotemporal data, open data, real-time data, operational data, and unverified outdated data.

Figure 3 shows the responses relating to the various kinds of data used by our respondents. Our survey identified that all organisations had structured, unstructured, sensor, log file, time stamped, machine, and operational data. However, only 5.21% had social media (Facebook, Twitter, and LinkedIn) data. The share of unverified outdated data sitting within organisations was 22.30%. This data was collected and stored within organisations, and no employees had knowledge about the data, its relevancy, or whether it could be put to use. A share of 63.91% of respondents believed that data silos are one of the biggest identified challenges to be addressed and diminish the power of Big Data. These respondents believed that if data was stored in different data sources, different data systems, and different organisational units, without some link between them, then no complete insights can be generated because the available data was not integrated at the back end.

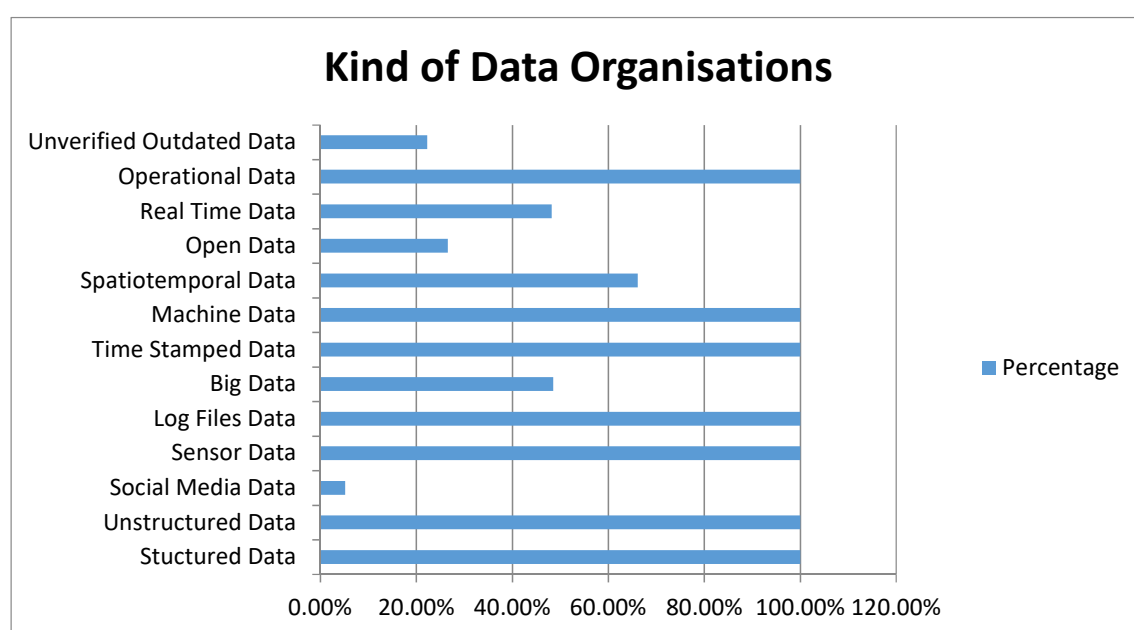


Figure 3. Kinds of data held by organisations.

What do you think are the benefits of legacy systems and data in your organisation?

Most respondents believed that their organisations used legacy systems for day-to-day operations. People were very comfortable using these legacy systems as they were familiar with them. All respondents believed that business continuity was important, and a legacy system that worked and kept everyone on the same page was good for the organisation. A share of 64.94% of the respondents believed that using a legacy system was easy to manage and within the control of the organisation. All respondents believed that using existing systems was less complex as over time people gained confidence in using it. Globally, data is being retrieved from legacy systems for business intelligence and decision making. However, the insights available from existing data have become more meaningful with the advancement in tools and technologies. Legacy systems are considered as the life-blood of organisations, such as Supply Chain Management (SCM), Human Resources Systems (HRS), Customer Relationship Management (CRM), and Learning Management Systems (LMS). Respondents admitted that legacy systems were very important to the running of their organisations. All respondents believed that the legacy systems running in their organisations contained significant and invaluable business logic of the organisation. These respondents further believed that the business logic embedded in the legacy systems was very important for the running of the organisations and organisations could not afford to throw them away. All respondents believed that the “legacy systems cannot be replaced

because of obvious reasons and hence Big Data solutions need to be integrated to legacy systems". All respondents believed that "there is a need for integrating Big Data with their legacy systems".

What do you think are the main disadvantages with reporting when using legacy system and data?

A share of 56% of respondents believed that existing systems and data could not fulfil the requirements of changing technology, which should be treated as being of prime importance if organisations wished to gain and maintain a competitive edge in today's world. Business requirements change continuously. The disadvantages of legacy systems were identified as: costly, outdated, limited flexibility, cannot generate reports in real time, store only organisational data in a structured format, immobile systems, and data integrity problems. Data and processes are not scalable and cannot be cross referenced. Data redundancy exists, together with the inability of systems to share data with each other. Vital information is lost as a result. Legacy systems support certain types of reporting and can supplement legacy system data with other data sources, thereby improving insights. A share of 57.73% of the respondents believed that their customer contact information was incorrect. One of the respondents mentioned that "if you've got a database full of inaccurate customer data, you might as well have no data at all". A share of 57.73% of the respondents believed that data silos can be eliminated by integrating data. They also believed that, in their organisation, accessing legacy system data takes a long time. Business decisions become totally dependent on accessing legacy data. Business users cannot access legacy systems directly, so the request to extract data from legacy systems has to go through the IT department for processing. This takes a long time as too many processes are involved. Organisations need to reduce these unwanted processes by integrating and eliminating data silos in today's world of BI.

Where and when do you use legacy systems and data for decision making?

Legacy data are used for historical analysis and operational analysis to understand how to improve the future based on historical evidence. Data is analysed for descriptive and predictive analytics. Respondents want to have prescriptive analytics; however, they do not have the skill sets, tools, and technologies to run prescriptive analytics. Our respondents required the use of legacy systems for five key themes; information access; insight; foresight; business agility; and strategic alignment. All of the respondents and their organisations were using data generated from systems such as enterprise resource planning systems, attendance tracking systems, and e-commerce systems. The data generated from these systems were stored in data warehouses, data marts, and database management systems. These technologies have existed in various forms for many years. These are large amounts of data and our respondents believed that these data fit into the meaning of Big Data. However, Big Data is not only about storing and retrieving semi-structured and unstructured data.

3.3. Big Data Initiatives and Implementation Questions

This section captures answers about finding the proper fit of a Big Data solution and technologies for an organisation. The section discusses the results we found from the survey.

In your opinion what is the range (high; moderate and low) of data processing in your organisation in relation to volume, velocity, variety, and veracity?

Most of the respondents believed that their organisations were contributing towards one or other of the stated characteristics of Big Data (volume, velocity, variety, and veracity) in the range of high, moderate, and low. Figure 4 shows the responses from our respondents about data volume, velocity, variety, and veracity in their organisation.

Does your organisation have information management big data and analytics capabilities?

A share of 72.16% of the respondents believed that their organisational data was measured in gigabytes and that applying new tools for business intelligence would help their organisation. However, they also believed applying new tools should not disrupt their existing running systems. Organisations require different types of analytics for different purposes, so new technology should help them with these changing requirements. Organisations having gigabytes of data do not have any Big Data project in their organisation. However, data analytics is enhanced using machine learning

to provide more insight into the data and make use of the data that the organisations are already collecting. A share of 27.83% of the respondents believed that their organisations had petabytes of data and Big Data projects were underway to identify real time fraud and detection. These included finance, insurance, and service provider organisations. The aviation industry is establishing a framework to use Big Data. Higher education systems are using Learning Analytics on already collected data. This fits into Big Data, as the analysis of discussion forums, emails, etc., uses unstructured data. A share of 70.1% of the respondents stated that Big Data had primarily been used to drive profits. All of the respondents believed that Big Data analytics can provide deep insights into customer behaviour and help in gaining a 360° view of their customers, by analysing and integrating existing data. One of the respondents stated that “Big Data analytics is all about understanding the customer, and that means harnessing all resources not just analysing all contacts with the organisation, but also linking to external sources such as social media and commercially available data. For the digital supply chain, it is about collecting, analysing and interpreting the data from the myriad of connected devices”.

Organisation Sector	Volume	Velocity	Variety	Veracity
Research Organisations	Low	Very Low	High	Moderate
Service Providers	High	Moderate	High	Moderate
Higher Education	Low	Low	High	Moderate
Financial Organisations	High	Low	High	High
Energy Sector	Moderate	Low	Low	Low
Supermarket Organisations	High	Moderate	Low	Moderate
Insurance	High	Moderate	Moderate	High
Aviation	Moderate	Low	Low	High
Others	Moderate	Low	Moderate	Moderate

Figure 4. Range of data processing in organisations.

Do you think that integrating a big data solution will benefit your organisation?

All of the respondents believed that integrating Big Data solutions would bring benefits to their organisation in different forms. This question had interesting answers, in which higher education respondents highlighted that contract cheating, which is a form of fraud, can be detected using Big Data analytics. Fraud detection is very common in financial, service provider, and insurance organisations. One of the respondents stated that “if you can obtain all the relevant data, analyse it quickly, surface actionable insights, and drive them back into operational systems, then you can affect events as they’re still unfolding”.

Has your organisation developed a big data strategy?

Organisations with gigabytes of data had not developed any Big Data strategy within their organisation. They were using advanced analytics, pattern recognition, and deep learning techniques to identify any irregular patterns. Organisations with petabytes of data had a Big Data strategy in place as these organisations were already working on a Big Data framework. However, their reporting systems were not integrated to their legacy systems. According to our survey, 86.62% of the respondents using Big Data (that is, 27.83% of all respondents) believed that Big Data framework implementation requires organisational efforts. These respondents believed that one of the disruptive facets of Big Data is the use of a wide range of Big Data tools and technologies for innovative data management to support different analytics.

What kind of analytics is used in your organisation?

All of the respondents believed that they were using descriptive analytics. A share of 53% believed that they were using predictive analytics and 27% believed that they were using prescriptive analytics. Organisations on the forefront of money management were using prescriptive analytics. The ecosystem of Big Data is daunting and confusing. A share of 10.3% of the respondents believed that there should be guidelines with requirements on how to use Big Data tools, technology, and architectures. One of the respondents stated that “the most practical use cases for Big Data involve the availability of data, augmenting existing storage of data, as well as allowing access to end-users employing business intelligence tools for the purpose of the discovery of data”.

How do you generate reporting for business intelligence?

All of the respondents were using legacy systems, such as CRM, SCM, LMS, etc., to generate reports for business intelligence. A share of 27% of the respondents believed that they were using a Big Data framework to generate reports. However, the frameworks did not include data from legacy systems. All of the respondents believed that they used legacy systems and data for business intelligence and decision making.

What approaches does your organisation use to integrate legacy systems and data? What problems are associated with it?

All of the respondents believed that there was no framework implemented to integrate Big Data solutions with legacy systems. However, they were generating reports from legacy systems, generating reports from Big Data frameworks, and combining the reports for final analysis. Respondents cited a lack of experience slowing project progress (48%), struggling to keep up with new data sources (58%), and issues with constantly changing business requirements (44%) as their top challenges.

When making a business decision in your organisation what do you mostly rely on?

Our survey identified that business decisions are made at different levels, which can be classified as functional level, business unit level, and corporate level. At all three levels, people rely on data and reports generated by existing organisational systems. As identified by our respondents, 100% of the respondents believed that information is the key success factor influencing the decision-making process. A share of 19.58% of the respondents believed that Big Data analytics was integrated into the decision-making process. The remaining 80.42% respondents believed that Big Data analytics was not used in the decision-making process.

Do you see value in integrating big data solutions into legacy systems and data in your organisation?

All of the respondents believed that integrating a Big Data solution with legacy systems and data in their organisation would bring benefit to their organisation. However, 87.62% of the respondents also stated that their organisation must have a strategic plan for integrating data from multiple data sources. This was required for Big Data integration for receiving holistic information from different data sources, including legacy systems and data. Integrating new datasets into existing pipelines (72%) was cited as a primary obstacle to Big Data projects. This was shown as the biggest concern that would hinder an organisation's progress towards a Big Data solution. These respondents believed that their organisation had hundreds of systems. This means that to receive all relevant data, the data must be extracted from many different sources, and the volumes could be overwhelming. In addition to the volume, the variety of sources also needs to be considered for integration purposes.

Does your organisation use any big data technology? Please specify the technology in the text box below.

A share of 12.38% of the respondents said that their organisation was using some kind of Big Data technology and tools. The most common tools used were Hadoop, Apache Spark, Apache storm, Cassandra, RapidMiner, MongoDB, R programming, and Neo4j. A share of 3.09% of the respondents believed that Hadoop was not suitable for social networking. For large volumes and graph-related issues, such as social networking or demographic patterns, Neo4j, a graph database, may be a better choice. Neo4j is one of the Big Data tools that is widely used as a graph database in the Big Data industry.

What is the biggest challenge in your organisation for collecting, accessing, storing, processing, and analysing data?

While Big Data offers many benefits, implementation of Big Data also has many challenges. The Big Data landscape is vast, making it even more challenging and complex for organisations to implement Big Data. Business users do not have enough understanding and knowledge of how Big Data can be utilised within organisation. Some of the commonly identified issues include inadequate knowledge about the technologies involved, data privacy, and inadequate analytical capabilities of organisations. A share of 85.56% of the respondents believed that their organisation lacked the skills of Big Data implementation in the workforce. Employees are not trained enough to handle Big Data technologies with confidence. Not many people are actually trained to work with Big Data, which then becomes an even bigger problem. Volumes, velocities, and varieties of data are growing continuously, giving organisations a large number of opportunities to gain insights that might otherwise be hidden in their available raw data. A share of 87.62% of the respondents were looking to increase their data team headcount to support Big Data solutions, but 85.56% also said that it was difficult to find professionals with the right skills and experiences within Big Data. Organisations were struggling to satisfy the requirements of implementing Big Data.

What are your goals of adopting big data projects?

Our respondents believed that there were several goals for adopting Big Data projects within their organisations. Following are the listed goals for adopting Big Data projects according to our identified nine organisational categories.

- Research Organisation: Goals are to process and analyse a high variety of data to generate more insight from the data.
- Service Providers: Goals are to collect, process, and analyse high volume and variety of data so that consumers' insights can be utilised and recommendations can be built.
- Higher Education: Goals are to collect, process, and analyse a high variety of data so that it can be used to enhance good practices in learning and teaching. Educators and learners should be able to take control of informed decisions. Teachers' performances can be fine-tuned and measured against student numbers, subject matter, student demographics, student aspirations, and behavioural classification.
- Financial Organisations: Goals are to collect, process, and analyse a high volume and high variety of data for early fraud detection and mitigation, and anti-money laundering.
- Energy Sectors: Goals are to collect, process, and analyse data from smart meters so that energy consumption can be analysed for improved customer feedback and better control of utilities use. Big Data analytics in the energy sector plays a crucial role in reducing energy consumption and improving energy efficiency. Through Big Data analytics, energy utilities can optimize power generation and planning.
- Supermarket Organisations: Goals are to collect, process, and analyse data to optimise staffing through data from shopping patterns and local events.
- Insurance: Goals are to collect, process, and analyse data derived from social media, GPS-enabled devices, and CCTV footage for fraudulent claims.
- Aviation: Goals are to collect, process, and analyse data to strengthen customer value, relationships, and loyalty.
- Others: Goals are to collect, process, and analyse data for optimising resources within the organisation.

If you do not have big data framework in your organisation, how beneficial will it be for your organisation? Do you see any value in having big data framework in your organisation?

A share of 12.38% of the respondents believed that they had a Big Data Framework within their organisation; 87.62% of the respondents believed that they did not have any Big Data Framework within their organisation. However, they did believe that a Big Data Framework would be beneficial to their organisation as a framework provides structure to achieve long term success. A Big Data framework concerns structure, technology, capabilities, and skilled people. The respondents believed

that a Big Data framework can provide a structure for organisations that want to start with Big Data or aim to develop their Big Data capabilities further. Respondents also suggested that the Big Data framework should be vendor independent and applied to any organisation regardless of choice of technology, specialisation, or tools. It should be able to provide a common reference model that can be used by any organisation depending on their requirements of Big Data analytics and solutions. All of the respondents believed that having a Big Data framework would add value where organisations were struggling to embed a successful Big Data solution in their organisation. The respondents also believed that a Big Data framework would be useful in integrating a Big Data solution with legacy systems, as it will dictate the architecture of Big Data and help in developing a Big Data strategy.

4. Discussion

The analysis of our survey answered our research questions and uncovered the current issues and challenges with the integration of Big Data solutions into legacy systems; practices related to the use of Big Data solutions with legacy systems; and strategies an organisation needs to adapt before implementing Big Data solutions within their organisations. Based on this, we can begin to develop some guidance plans for organisations to integrate legacy data. These conclusions are summarized in answer to our research questions across six categories answering our three research questions.

RQ 1, “What are the current issues and challenges with the integration of Big Data solutions into legacy systems?” is answered by the following two categories:

- While data is well structured, silos between legacy system (which are caused by organisation silos) can present issues: Organisations have different kinds of data generated from different systems on which decisions are built. Our survey respondents believed that data captured in different data sources cannot provide a complete view on which an informed decision can be made. These different data sources need to be integrated to fulfil the requirements of organisations in today’s digital age. This result correlates with conclusions reached by [5] in relation to the challenges faced by organisations in managing data reasonably, effectively, and quickly. Respondents believed that data within silos should be integrated with other data sources so that an organisation can gather insights from it to make data-driven decisions. A share of 57.73% of the respondents believed that their customer contact information was incorrect or not up to date. Data silos were identified as one of the reasons for having different data stored in different systems giving rise to inaccurate data. All of these respondents believed that data silos could be eliminated by integrating data. Data integration would produce a single, unified view of an organisation’s data. Business users for BI & DA applications can access this unified view to develop an actionable plan based on the organisation’s data assets.
- Legacy Data has rigid format issues that do not suit Big Data Applications: All of the respondents believed that the data in legacy systems are typical relational data from enterprise applications such as CRM and SCM, and are very structured. Respondents also believed that the legacy data tended to be on-premises, behind firewalls in a bounded and constrained infrastructure, so external security and data management were not an issue and never considered while developing legacy systems. This result correlates with conclusions reached by [16], where Big Data challenges reported relate to: storage; data representation; data life cycle management; analysis; reporting; redundancy reduction and data compression; energy management; data confidentiality; expendability; scalability; cooperation; and Big Data dimensional reduction. All of the respondents believed that legacy systems were not built to accommodate today’s different variety of data, as opposed to Big Data. All of the respondents believed that Big Data sources were identified to have different characteristics, such as frequency, volume, velocity, type, and veracity of the data. All respondents believed that processing, accessing, visualising, and storing Big Data is very complex. Organisations need to consider many dimensions, such as where the data is coming from, who is the owner of data, and how data can be shared. Policies, structure, procedures, and governance need to be in place before Big Data can be processed. All of the respondents

were of the opinion that building an appropriate architecture for Big Data solutions is challenging because so many factors are required to be considered. They also believed that it becomes even more challenging when Big Data solutions require reports from legacy systems for decision making. Legacy systems have different data formats. All of the respondents believed that a review of legacy applications and data was needed to get a complete picture of integrating Big Data solutions with legacy systems.

RQ 2, “What are the practices related to the use of Big Data solutions with legacy systems?” is answered by the following two categories:

- Skill shortages in Big Data and in integration of Big Data with legacy systems: All of the respondents believed that Big Data solutions were beneficial to their organisation. However, 85.56% of the respondents believed that there were skill shortages within organisations to implement Big Data solutions. This result correlates with conclusions reached by [15], where Big Data processes require working on seven operators: gathering, selection, pre-processing, transformation, data mining, evaluation, and interpretation. These seven processes require technical skills such as engineering for parallel computing. Big Data analytics systems are designed to work on parallel computing, with other systems which can be hosted on the cloud, or on another search engine. Employees are not trained to work with such Big Data analytics systems. Employees are not aware of the ethical concerns of processing Big Data. Volumes and varieties of data are growing continuously. A share of 87.62% of the respondents were looking to increase their data team headcount to support their Big Data solution, but the respondents also said it was difficult to find data professionals with the right skills and experience. The skillset required to integrate a Big Data solution with legacy systems is totally dependent on Big Data skills, and understanding and handling of legacy systems and data issues. Integration of different data sets requires understanding of different data formats and how they can be integrated to provide actionable insights.
- Lack of a framework and architecture for legacy and Big Data integration: Frameworks provide structure and the core objective of the Big Data framework is to provide a structure for enterprise organisations that aim to benefit from the potential of Big Data. All of the respondents believed that integrating a Big Data solution with legacy systems and data in their organisation would be beneficial to their organisation. The deep understanding derived from integrating Big Data could help organisations to improve their business processes, optimisation of resources, fraud detection, and improve customer relationships and satisfaction. Organisations can use Big Data analytics as their primary source for reporting and analytics after integrating Big Data solutions with legacy systems. Most organisations do not have a strategic plan to execute Big Data integration. A share of 87.62% of the respondents also stated that their organisation needed to develop a Big Data strategy concerning Big Data integration. This strategy would help plan receiving information from multiple data sources. Integrating new datasets into existing pipelines (72%) was cited as the primary obstacle to Big Data integration with legacy systems. This was shown as the biggest concern that would hinder an organisation’s progress towards a Big Data solution. All of the respondents believed that there was no framework implemented to integrate Big Data solutions to legacy systems. However, they were generating reports from legacy systems, generating reports from Big Data frameworks, and combining the reports for final analysis.

RQ 3, “What strategies does an organisation need to adapt before implementing Big Data solutions within their organisations?” is answered by the following two categories:

- Framework to integrate Big Data solution with legacy systems: A share of 87.62% of the respondents believed that they do not have any Big Data framework within their organisation. However, all of them believed that a Big Data framework would help those organisations who are struggling with implementing Big Data solutions. Legacy systems cannot be replaced, as they have many other challenges associated with them. Legacy systems represent many years of changes of

organisational processes. These legacy systems are assets of the organisation. Redevelopment of these systems would be unaffordable in terms of time, cost, and the required human resources. Big Data tools, such as Apache Hadoop and MapReduce, Apache Mahout, Apache Spark, Dryad, Storm, Apache Drill, JasperSoft, and Splunk [12], can be used to develop a framework to integrate Big Data solutions with legacy systems. The framework will show the organisation the best possible approach and required capabilities to proceed with a Big Data project. This will provide the organisation the required support and structure to start their Big Data projects.

- Legacy systems and the data are assets: All of the respondents said that their organisations were using legacy systems and accrued benefits from doing so. The difficulties in different types of data capture, storage, searching, sharing, analysis, and visualization [14], can be removed by using Big Data solutions. According to the authors, the architecture of Big Data must be synchronized with the support infrastructure of the organisation. Legacy systems and data are assets to the organisation which are used for building informed decisions. The applications used in BI & DA require access to legacy systems and data. An organisation's payment strategies, sales strategies, marketing strategies, optimisation of human resources, and satisfaction of customer needs contribute to an organisation's competitive advantage. Legacy systems have in-built business processes and data, ranging from mainframes, to organisationally developed custom applications and proprietary applications. Data is being retrieved from legacy systems for business intelligence and decision making. Legacy data is a crucial resource for BI & DA, and its retrieval is also very inefficient and expensive. Organisations are struggling to find ways in which legacy systems can be operated efficiently and in a cost-effective manner. There are many issues and challenges associated with legacy systems and data residing in legacy systems. Some of the issues and challenges are the cost and complexity of migrating to newer platforms; challenges in accessing data in legacy systems; slowness of data refreshes from legacy systems for BI & DA purposes; and outdated, obsolete technology. All of the respondents believed that legacy systems cannot be replaced as they are beneficial and in use. However, 100% of the respondents believed that their legacy systems must be integrated with Big Data technology to leverage a competitive advantage for the organisation.

5. Conclusions

The Big Data movement is fuelling business transformation within organisations. Furthermore, an organisation that uses Big Data is presented with more business opportunities. Our survey demonstrated that many organisations are implementing Big Data solutions and integrating their legacy systems and data with these solutions. However, no systematic framework is being used. Big Data integration with legacy systems should be able to provide solutions for integrating data from a variety of data sources requiring a variety of heterogeneous data formats. The integration should be able to maintain data accuracy and integrity, and should be addressed by the Big Data framework and Big Data architecture that organisations use.

Having a Big Data integration solution in place is as important as having good analytics tools for creating insights. However, from our survey we identified numerous challenges to integrating Big Data solutions with existing legacy systems. These include skill shortages in Big Data, together with technical challenges, such as the lack of Big Data frameworks, the lack of an organisational Big Data strategy, the use of cloud computing, the lack of a Big Data architecture or inability to define one, and the lack of knowledge of software systems to support Big Data. All of these challenges require further investigation.

Author Contributions: Formal analysis, S.J.; Investigation, S.J.; Methodology, S.J.; Resources, L.O.; Software, M.J.; Writing—original draft, S.J.; Writing—review & editing, M.J., L.O., M.C. and M.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors do not have any competing interests in this study.

References

1. Reinsel, D.; Gantz, J.; Rydning, J. *Data Age 2025: The Evolution of Data to Life-Critical*; International Data Corporation: Framingham, MA, USA, 2017.
2. Pearson, T.; Wegener, R. Big Data: The Organisational Challenge. Available online: https://www.bain.com/contentassets/25c167a5149c42168994338f9dc99ffe/bain_brief_big_data_the_organizational_challenge.pdf (accessed on 1 February 2019).
3. Bhadani, A.K.; Jothimani, D. Edited book chapter Big Data: Challenges and Opportunities, and Realities. In *Effective Big Data Management and Opportunities for Implementation*; IGI Global: Hershey, PA, USA, 2016; pp. 1–24.
4. Arputhamary, B.; Arockiam, L. Data Integration in Big Data Environment. *Bonfring Int. J. Data Min.* **2015**, *5*, 1–5. [CrossRef]
5. Zou, W.; Jing, W.; Chen, G.; Lu, G.; Houbing, S. A Survey of Big Data Analytics for Smart Forestry. *IEEE Access. Spec. Sect. Urban Comput. Intell.* **2019**, *7*, 46621–46636. [CrossRef]
6. Inoubli, W.; Aridhi, S.; Mezni, H.; Maddouri, M. An Experimental Survey on Big Data Frameworks. *Future Gener. Comput. Syst.* **2018**, *86*, 546–564. [CrossRef]
7. Jha, M.; Jha, S.; O'Brien, L. Combining Big Data Analytics with Business Process Using Reengineering. In Proceedings of the 2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS), Grenoble, France, 1–3 June 2016.
8. Raghupathi, W.; Raghupathi, V. Big Data Analytics in Healthcare: Promise and Potential. *Health Inf. Sci. Syst.* **2014**, *2*, 3. [CrossRef] [PubMed]
9. Bisbal, J.; Lawless, D.; Wu, B.; Grimson, J. Legacy Information Systems: Issues and Directions. *IEEE Softw.* **1999**, *16*, 103–111. [CrossRef]
10. Jha, S.; Jha MO'Brien, L.; Wells, M. Integrating Legacy Systems into Big Data Solutions: Time to make the Change. In Proceedings of the IEEE Conference on Asia-Pacific World Congress on Computer Science and Engineering, Nadi, Fiji, 4–5 November 2014.
11. Liu, X.; Iftikhar, N.; Xie, X. Survey of Real Time Processing Systems for Big Data. In Proceedings of the 18th International Database Engineering & Application Symposium, Porto, Portugal, 7–9 July 2014; Association for Computing Machinery: New York, NY, USA, 2014; pp. 356–361. [CrossRef]
12. Acharjya, D.P.; Ahmed, P.K. A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools. *Int. J. Adv. Computer Sci. Appl. (IJACSA)* **2016**, *7*, 511–518.
13. Oussousa, A.; Benjellouna, F.Z.; Lahcen, A.A.; Belfkih, S. Big Data technologies: A survey. *Comput. Inf. Sci.* **2018**, *30*, 431–448. [CrossRef]
14. Khan, N.; Yaqoob, I.; Hashem, I.A.T.; Inayat, Z.; Ali, W.K.M.; Alam, M.; Shiraz, M.; Gani, A. Big Data: Survey, Technologies, Opportunities, and Challenges. *Sci. World J.* **2014**. [CrossRef] [PubMed]
15. Tsai, C.W.; Lai, C.F.; Chao, H.C.; Vasilakos, A.V. Big Data Analytics: A Survey. *J. Big Data* **2015**, *2*, 21. [CrossRef]
16. Praveena-Anto, M.D.; Bharathi, B. A Survey Paper on Big Data Analytics. In Proceedings of the International Conference of Information, Communication & Embedded Systems (ICICES), Chennai, India, 23–24 February 2017; ISBN 978-1-5090-6135-8.
17. Denscombe, M. *The Good Research Guide: For Small-scale Social Research Projects*; Open University Press: Buckingham, UK, 1998.



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