IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 18, 2019, accepted January 13, 2020, date of publication January 17, 2020, date of current version January 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2967436

Requirements for Big Data Adoption for Railway Asset Management

P. MCMAHON, (Member, IEEE), T. ZHANG, (Member, IEEE), AND R. DWIGHT, (Member, IEEE)
Faculty of Engineering and Information Sciences, University of Wollongong, Wollongong, NSW 2522, Australia

Corresponding author: T. Zhang (tieling@uow.edu.au)

ABSTRACT Nowadays, huge amounts of data have been captured along with the day-to-day operation of assets including railway systems. Hence, we have come to the era of big data. The utilization of big data technologies for asset condition information management is becoming indispensable for improving asset management decision making. The vital information such as precursor information collected on failure modes and knowledge that may be available for analysis is hidden within the large extent of data. There are analysis tools incorporated with techniques such as multiple regression analysis and machine learning that are facilitated by the availability of big data. Therefore, the utilization of big data technologies for asset condition information management is becoming indispensable for improving asset management decision making. This paper provides a review of the requirements and challenges for big data analytics applications to railway asset management. The review focuses on railway asset data collection, data management, data applications with the implementation of Blockchain technology as well as big data analytics technologies. The need for, and the importance of big data analytics in railway asset management; and the requirement for the asset condition data collection in the railway industry are highlighted. Research challenges in railway asset management via application of big data analytics are identified and the future research directions are presented.

INDEX TERMS Big data analytics, asset management, railway, data management, blockchain.

I. INTRODUCTION

The term of 'Big Data' was coined by a few pioneer researchers and scientists in later 1990s ([1]–[3]) since the first proposal of World Wide Web (WWW) was invented by Berners-Lee in 1989 [4] and its wide application started in 1993 [5]. Since then, the global data volume has been growing extremely fast year by year. 'Big Data' refers to "the explosion in the quantity (and sometimes, quality) of available and potentially relevant data, largely due to the result of recent and unprecedented advancements in data recording and storage technology" [6]. With the advancement of technologies, increasing numbers of sensors have been implemented in various industry fields. Therefore, huge amounts of data are collected in day to day operations in different industry sectors including railways. Those large volumes of data bring us great value while with big challenges as well. The traditional way to handle the data including data storage, processing and management is obviously incapable of satisfying the current social and industrial activity needs. The technologies being capable of dealing with large volumes of data are therefore required thereby the appearance of a new term of 'Big Data Analytics'. 'Big Data Analytics' refers to a process from data collection, data management to real application by applying available techniques/technologies applicable to dealing with large volumes of data, i.e., big data. The discussion and investigation of related technologies and their applications have been soon rapidly expanded through almost every industrial field.

Big data analytics has become a hot research topic in recent ten years. The number of published articles has been exponentially increasing especially in the last five years. These include many review papers discussing the related technologies and open issues for application of big data and big data analytics, e.g., [7]–[11] and many others by focusing on application of big data analytics in specific industrial fields such as [12] covering agricultural production, [13] in healthcare, [14] in supply chain management, [15] in smart

The associate editor coordinating the review of this manuscript and approving it for publication was Jiajie Fan.

manufacturing, and also a very recent review article on application of big data analytics in railway transportation systems (RTS) [16]. The key contribution of this review paper given by [16] is the taxonomy of the selected papers in review with analysis and discussions on big data sources and by considerations of applications of big data analytics on maintenance, operation, and safety of RTS. In view of asset management life cycle activities and to the best of authors' knowledge based on the literature survey, it lacks a comprehensive review with discussions in detail about the application advances of big data analytics in railway network area for railway asset management. From railway system operation point of view, however, there are huge amounts of data collected every day. As a result, it has been identified that this is a typically important area that big data analytics will show and demonstrate its power. In addition, the migration towards the digital railway by a number of rail operators worldwide will increase the uptake of big data analytics into rail infrastructure asset management [17]. For these reasons, we think an overall review on big data analytics and its application status in railway industry is required and it is also timely needed in order to provide an insight into the technological development, challenges and gaps for the railway system operators and researchers.

Railway networks have been one of the largest assets in most countries and to manage it well has always been a concern of the railway asset owners and operators. This has led to the initial idea for asset management within a railway infrastructure environment which has evolved from a number of sources including the concept of Total System Support [18]. As industry support for asset management systems has developed, standardization processes have resulted in the development of asset management standards such as ISO 55000 ∼ ISO 55002 [19]. An asset management system is concerned with the planning and control of all asset-related activities and their relationships to ensure that asset performance meets the intended competitive strategy of the organization. All aspects related to asset life cycle activities from concept design to disposal are crucial to the success of an organization. Asset lifecycle activities can be considered to be both interdisciplinary and interrelated [20].

Rail assets are capital intensive. They drive a significant proportion of the organization's service delivery costs. Even a small improvement in asset management can bring a large benefit. The rail industry is concerned with the condition of its assets and is actively involved in developing advanced strategies and techniques to maintain its infrastructure. The motivation for managing the asset condition and the movement towards condition-based maintenance has provided an impetus to the adoption of big data analytics for rail asset management. The challenge that asset management presents in the rail industry, however, extends beyond infrastructure and into assets of all classes, including transportation systems, operations, rolling stock, services, safety, and security. There is a mix within the infrastructure of complex linear assets such as track and electrical overhead wires with discrete assets such as bridges, switches, and stations [21]. This amalgam of infrastructure assets needs to be correlated with rolling stock assets to provide an overview of asset condition performance including the wheel/rail interface [22], [23].

Asset condition monitoring plays a primary role in asset management because it provides asset condition information that enables subsequent activities to be decided efficiently and effectively. As highly sophisticated condition monitoring systems produce huge amounts of data every day, it is obviously impractical or arguably impossible to handle the data in an alphanumerical form. Instead, most of the data gathered must be properly visualized in order for users to gain insight into the actual behavior of the objects in question. Hence high-quality visualization of the analyzed data is needed when it comes to infrastructure management. This has led to the utilization of evidence-based decision making [24].

In order to provide high-quality decision-support, a railway asset management system requires a large amount of data to be available for analysis. This data, however, is usually contained in various databases and storage systems. It requires careful data-selection and transfers to a railway asset management database for infrastructure condition analysis and work-planning purposes. The data may be in a structured format (which can include sensor data, real-time monitoring data, failure codes) or in semi-structured or unstructured data such as maintenance reports and service logs. In traditional asset management systems, the maintenance reports and equipment service logs are normally stored separately. However, for analysis purposes, the maintenance reports and equipment service logs may be reviewed in a combined format. In terms of big data classification, structured, semi-structured and unstructured data are often classified under "Content Format" in terms of their characteristics [8]. For traditional asset management systems, data selection is known to be very sensitive (and important) as all the future analyses and subsequent planning are to be performed based on the data stored in the asset management database [25]. Hence, the quality and reliability of the transferred data represent one of the crucial issues and keys to the success of any railway asset management system in use. However, data characteristics are not normally considered in conventional asset management data models, whereas big data utilizes the data characteristics to identify heterogeneous formats for handling [11]. The other issue is the Rule-creation, i.e., the transfer of the user's knowledge, standards and regulations comprising, in fact, the overall maintenance policy into asset management decision-rules [26]. In very recent years, the idea of big data analytics within the railway industry has been introduced [27]. In terms of building transport network sustainability, the measurement of performance indicators to show progress towards sustainability can be challenging in the presence of big data where managing the amounts of big data can be a significant challenge [28]. Other issues have also been identified in using big data for predicting the behavior of assets in use.

The purpose of this paper is to conduct a thorough review of big data analytics and its applications in railway industry in order to identify the gaps in the areas that are particularly important for asset management and condition monitoring. More importantly, this paper seeks to identify cross-disciplinary concepts and opportunities for both rail asset management practitioners who have traditionally used small data and are now facing big data challenges, and big data researchers who are engaged in technological development for handling big data. It is our view that the recent progress made in use of big data analytics has the potential to drive fundamental advances in research in rail asset performance monitoring and, at the same time, the knowledge accumulated in transportation research in the past many decades can guide big data studies to answer questions that matter to the rail transportation systems.

With the purposes as described above, this paper is organized as follows: Section II gives a brief introduction to the methodology utilized in this review; Section III provides background knowledge about big data and big data analytics in railway asset management; Section IV gives an overview of railway asset data collection requirements in big data environments; Section V discusses the big data analytics technologies available for implementation where the challenges and technical gaps are also highlighted; Section VI presents a brief overview of big data analytics applications to railway infrastructure where challenges in the applications are indicated, and this paper is concluded as given in Section VII.

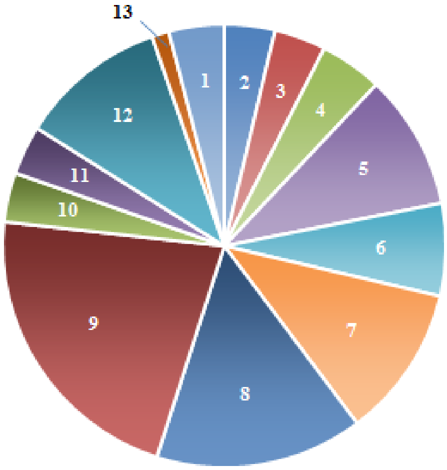
II. METHODOLOGY

Methodologies utilized in literature review papers in the related fields have been studied. For simplicity, the methodological approach followed in this paper is adapted from [29] and [30] where there were three phases identified, namely, research question statement or focus, research methodology and research scope. In this paper, six dimensions or aspects of big data analytics were considered in the selection of keywords for literature search as follows:

i. The areas of railway infrastructure asset management in which big data analytics could be applied;
ii. the level of big data analytics in rail network management;
iii. the trend towards condition-based maintenance;
iv. prognostic health management and smart monitoring of rail assets;
v. types of big data models, and big data tools and techniques used for applying these models; and
vi. Blockchain technology for railways and its current application status.

The selection of these six aspects is based on detailed discussions among the authors and the determined scope of this work. The research scope is a literature review on applications of big data analytics in the railway industry from the published research articles and reports published between 1992 to 2019. The reason to choose this time period is due to that the research area is relatively recent.

To evaluate the available literature, multiple Scopus searches were carried out initially based on the following query format. As an initial phase, keywords used included: (TITLE-ABS-KEY(''asset management'' OR ''condition monitoring'' OR ''data quality'' OR ''missing data'' OR ''machine learning'' OR ''big data'') AND TITLE-ABS-KEY (decision∗) AND TITLE-ABS-KEY (rail∗)). Then the literature search was progressively adapted and driven by considerations of the industry issues. Research was undertaken to further identify the relevant literature and discussions to focus on the issues, challenges, and opportunities being faced by industry including potential application domains. Some of the key challenges identified within rail asset management included both the trend for increasing volume and concern about the veracity of data available for analysis within an asset management system. Other challenges identified included the need for selection of appropriate big data analysis tools to match the data being collected and the trend towards condition-based maintenance and prognostic health monitoring systems within rail infrastructure organizations. The volume of data being collected for import and analysis within asset management has exponentially increased with the increasing number of condition monitoring systems implemented onto the railway network. However, the veracity of the data is possibly more critical in cases where rail safety models are required as part of the analysis. Data quality is inherently impacted by uncertainty and unreliability of data sources to some extent [31]. As traditional corporate IT networks may not be compatible with the introduction of big data analytics, these aspects were considered to be a part of the focus of the paper and included in the methodological framework for this paper. The trend towards condition-based maintenance has also focused attention on the requirement for intelligent asset management systems with machine learning capabilities built in. Additional search terms were identified during this process for the methodology. As other industries have also faced challenges in big data analytics, the search terms were then widened to include other industry sectors such as the health and manufacturing industries to include case studies for analysis. For evaluation of available literature on big data technologies and big data analytics, search engines through Google Scholar, Web of Science, and IEEE Xplore were utilized. Based on the article search process as described above, over 2000 papers were identified for review and analysis. The selection process was adapted from Ngai *et al.* [29] and involved a three-stage approach where papers were judged primarily on their title, abstracts and keywords. Two co-authors then debated the extent to which the paper addressed the six aspects of literature search criteria identified above. The selected articles were then grouped in different categories as outlined in Figure 1 based on their primary contributions to each of the category. It is understood that one paper may cover several categories, but it is only allocated to the category to which it contributed the most



1. Blockchain (Other) (45); 2. Decision Support (41); 3. Missing Data (42); 4. Condition Monitoring (51); 5. Big Data Analytics (110); 6. Big Data (75); 7. Asset Management (124); 8. Rail Infrastructure (169); 9. Infrastructure Maintenance (243); 10. PHM (Structural Health Monitoring) (40); 11. Asset Management (Oil & Gas, Utilities) (40); 12. Fault diagnostics and prediction (120); 13. Blockchain (Rail) (14).

FIGURE 1. Grouping of the searched articles in different categories.

and counted only once to avoid any duplicated counting. As condition-based maintenance for railway transportation systems with application of big data was discussed in [16] and it was identified that the fault diagnostics and prognostic health management (PHM) for railway systems is well suitable for a separate paper, the detailed discussion on these two aspects was therefore not included in this paper due to the paper length limit requirement. Finally, 200 articles (including 9 published through online sources) were cited in this paper to support discussion and analysis. Some others were excluded as not being relevant to the focus of this paper or similar contents have been covered by the ones cited in this paper.

Instead of providing a taxonomy analysis of the selected articles, the analysis and discussions in the following sections are more focused on the technical perspective.

III. UNDERSTANDING OF BIG DATA ANALYTICS FOR RAILWAY ASSET MANAGEMENT

A. BIG DATA IN RAILWAY NETWORK MANAGEMENT

Big data is often referred to as the data as having five characteristics, i.e., large volume (Volume), fast processing speed (Velocity), multiple domains (Variety), low density of the value distributed (Value) and complexity of data formats (Veracity) [32]. Veracity was initially a term developed by IBM in 2012 [31] to convey the idea of uncertainty and unreliability of the data. However, some researchers would like to highlight the 4 V's as 'Value' is the results extracted from the data through analysis and modeling. The value of big data in the decision sciences has been highlighted by Wang *et al.* [33] where the different stages of data capture, curation, analysis, visualization, and decision-making were discussed. The availability of big data is regarded as one of the enablers for intelligent decision making with complex systems [34]. Big data analysis technologies are required to process and analyze the data having these five characteristics. In addition, big data has facilitated the use of techniques such as machine learning and expert systems for knowledge discovery [35]. Examples of application of these techniques to big data analysis in other industries include the use of statistical process control with big data analytics in smart manufacturing processes [36]. A generic system for machine learning using big data has been explored by [37]. Another case of using big data analytics has been discussed by [38] where the existing systems using business intelligence and data warehouse are limited to handling and relating unstructured data to structured data. Structured data, once uploaded from source and analyzed, can be used as variables in a statistical/machine learning model. Unstructured data, however, requires further analysis and decomposition into a set of structured data elements [39]. The significance of big data analytics for asset management has been described in [40]. In this particular case, the whole of the asset management system can be supported by the use of big data analytics and the Internet of Things (IoT).

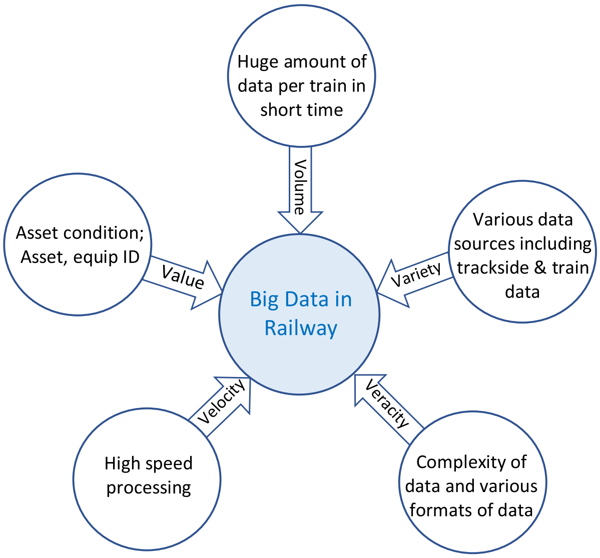

Huge amount of data per train in short time
Volume
Asset condition; Asset, equip ID
Value
Big Data in Railway
Variety
Various data sources including trackside & train data
Velocity
Veracity
High speed processing
Complexity of data and various formats of data

FIGURE 2. Five "V's".

Figure 2 provides an overview of some of the characteristics of big data pertaining to railway network management. Big data dimensions have been proposed in different forms to meet challenges in implementation for various disciplines [41]. Numerous people have increased the number of "V's"; however, we focus our discussion on the five V's in Figure 2. For railway asset management, the big data problem can be described in terms of a) aggregating multiple databases which are individually manageable and come from different sources, b) individual datasets that by themselves are too large to be processed by standard algorithms on legacy hardware [42] and c) relational databases have been designed to deal mainly with structured data, and little support is provided to semi-structured or unstructured data [43]. While linking the asset and equipment identification (ID) with asset condition information is seen as a key aspect for adding value to railway asset management systems, this information may be from different sources and requires aggregation and processing to adding value. As data is aggregated from multiple sources, the data may sometimes exhibit heavy tail behavior and non-trivial tail behavior [44]. There may also be an imbalance in the datasets where the most important relationships to be discovered in the datasets are presented by a small number of examples in the dataset [45]. In a similar manner to trackside systems, rolling stock on-board systems may perform the analysis or processing prior to transmission to trackside-based storage, in a similar approach to the Internet of Things (IoT) applications. The challenge here for the asset condition data analysis is to understand where the raw data analysis is being performed.

B. UNDERSTANDING OF BIG DATA ANALYTICS

Big data analytics is a process of collecting, organizing and processing large amounts of data to discover useful information and extract patterns for the purpose of asset condition prediction, process optimization and decision making in management to drive better business decisions. A taxonomy of different processes within big data is provided by Khan *et al.* [46]. The focus on big data analysis is to typically discover the insight into the knowledge that comes from analyzing the data [47].

Asset condition data on its own without analysis will not create value for the asset owners who are collecting and own the data. Once the data is stored and analyzed, it can create tremendous value [48]. The data analysis or process can consist of a number of technologies and approaches such as in-memory analytics, in-database analytics, and appliances to examine large and varied data sets [49]. There are six analytical techniques [50] which can be grouped into four different kinds of analytics as follows [32]:

a) descriptive analytics, which includes reporting/online analytical processing (OLAP), dashboards/scorecards, and data visualization. These applications have been widely used for some time, and are the core applications of traditional asset management systems;
b) diagnostic analytics, which is used for discovery or to determine why something happened;
c) predictive analytics, which can suggest what will occur in the future. The methods and algorithms for predictive analytics include regression analysis, machine learning, and neural networks that have existed for some time; and
d) prescriptive analytics, which can identify optimal solutions, often for the allocation of scarce resources. Prescriptive analytics are seen as the future of big data [51].

To select one of the four kinds of analytical approaches, an awareness of the concepts of design data and organic data is required. Hence, in addition to the evolution of analytics, the ideas of ‘‘design data and organic data’’ have also been discussed and developed in parallel [52].

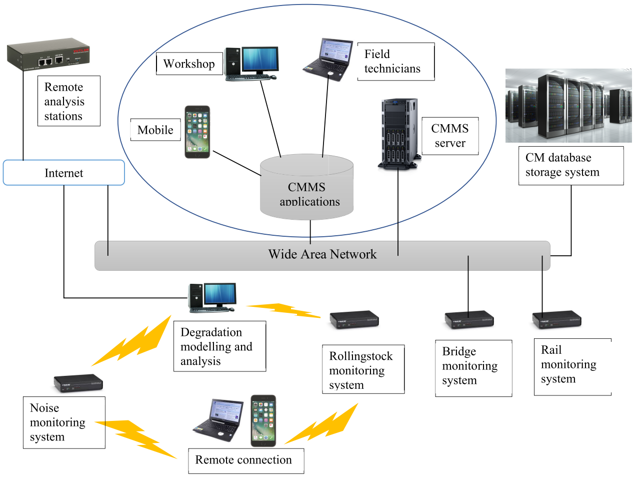


FIGURE 3. Integrated rolling stock and infrastructure field data collection.

IV. DATA COLLECTION AND MANAGEMENT IN RAILWAY ASSET MANAGEMENT

A. UNDERSTANDING OF BIG DATA ANALYTICS

A typical railway infrastructure condition monitoring data collection is shown in Figure 3.

Each of the monitoring systems in Figure 3 is collecting structured data where the format of the data is dependent on the type of system being monitored as well as unstructured data including log files. For example, the noise monitoring station may be collecting acoustic audio files. Storage requirements may also be different for different data types. Field Technician will collect unstructured data and log files that need to be linked to the monitoring system structured data. Linking the work order generation from the CMMS with the correct asset and accessing log files provides information for later analysis. Within traditional CMMS systems, the unstructured data including log files may be stored separately from the structured data within the CMMS.

Within typical asset condition monitoring systems, sensors are located either on trackside or within the railway vehicle undertaking measurements. Typical asset condition data may include a mixture of related data from high to low sample rates [53]. The asset condition data sample rate may be affected by the requirement for the data collection when equipment passes near a sensor, e.g., a rolling stock vehicle passes over a track sensor or by a trigger when measured data passes over a threshold. In this particular example, the amount of data sent may increase significantly to allow the evaluation of the trends by an expert system for the particular threshold that has been reached [54].

Condition monitoring sensor systems also provide data that must be interpreted to turn into alarms [55] and stored separately in a database for later verification. This highlights the development of maintenance recording features of asset condition data which are the precursors for asset management systems today. Other key considerations in the introduction of condition monitoring to railway asset management are ensuring that the technology being used is as reliable as the asset being monitored and being prepared for analysis and storage of large volumes of data. Analysis of the data may have challenges due to the volume of the data being captured in real-time [56].

Railways across Australia including ARTC and Sydney Trains have implemented a number of trackside condition monitoring systems within their network for monitoring of infrastructure and train assets including the wheel/rail interface. The Wheel Impact Load Detector (WILD) system [55] has been utilized for providing notification when wheel impacts are detected. Condition assessment is based on visual inspection and some other measures, e.g., track machines run over the track and provide a record of the condition of the rail [57]. Integration of data from condition monitoring systems with track machine recording can identify the locations where defects are present and assist in the decision making process [54]. In recent years, Sydney Trains has implemented pantograph condition monitoring using laser and computer vision technology [58]. A range of applications using LIDAR and video recording technology has also been utilized in rail corridors to provide track degradation measurement data for prediction of track deflection.

B. DATA MANAGEMENT REQUIREMENTS

There are a number of key requirements within condition data management including a) data retention requirements that may be linked to organizational requirements, b) data accuracy and quality requirements c) data volume requirements and d) identification of key data required for decision making and verification purposes. Each of the requirements can be traced back to the asset management data model required to meet the asset management objectives of the organization. Within the advent of intelligent railway networks, data management has also been identified as a key challenge in their implementation.

Condition-based maintenance and the requirement for management of the increasing volume of data have also required organizations to focus on the development of new requirements to meet these needs as railway organizations migrate away from a preventative maintenance approach [16]. These requirements may include real-time analysis of streaming data, identification of owner and labeling of source data in a real-time environment, sharing of data in real-time to provide for safety hazard notifications as well as the storage and identification of models to be applied to the condition-based maintenance data. The condition-based maintenance data must be converted into information including details about the quality of the data collected, any uncertainties, maintenance or operational options [59]. A case study was undertaken by the Swedish Railways with the support of Bombardier systems for the collection of condition-based maintenance data of railway vehicles [60]. A key requirement is to identify at an early stage what needs to be measured and how to measure, collect and store the data. The real-time identification of alarms and trend prediction is identified as requirements [61]. Further development is occurring within the American Association of Railways (AAR) to develop interface standards for condition-based maintenance [62]. A ten-year initiative has been set up to provide nationwide monitoring and repair information of freight cars operating in North America. Data modeling prior to integration may be seen as key to identifying relationships between the data for rail maintenance purposes [63].

While standards including IEEE 1451 [64] have been proposed for regulating the condition monitoring sensor interface, the asset condition data may be collected and stored in a proprietary format [65]. Innovative approaches to converting the data to an open standards format may be required to extract useful information. Integration of the condition-based maintenance data using different condition-monitoring tools used within railway infrastructure is a critical requirement to ensure the usefulness of the data [66]. The requirement for integration of data from different data sources has been extended to other infrastructure areas where complex decision-making is required to manage a hierarchy of assets to support maintenance decisions [67].

C. DATA MANAGEMENT SYSTEM ARCHITECTURES

A data warehouse model incorporating condition-based maintenance using the open systems framework (OSA) is proposed for an asset management system [68], in which it is identified that the use of standard terminologies and tools can assist in more effective use of asset management information in data warehousing scenarios, with all of the data being used. This model consists of seven (7) layers as shown in Figure 4 which is modified for the railway environment. With the advent of the internet and IoT, the presentation layer in some cases has evolved to provide a web-based user interface to match the internet technologies or via a dashboard. However, the lowest three layers may still be based on proprietary or bespoke solutions from the manufacturer or supplier of the condition monitoring solution. Migrating the top three layers of the OSA-CBM model to a big data analytics framework can be achieved if the business process rules for the asset being monitored is known and understood. In the UK rail industry, an ontology-based data management approach has been proposed by the Rail Safety and Standards Board [69]. An agent-based approach is suggested to relate the identities and locations of asset equipment.

To utilize the data in an operational environment, dashboards have been implemented to display real-time asset condition data [71] to allow for real-time decision making on asset operational management using descriptive and diagnostic analytics. Figure 5 provides a technical architecture overview of data collection, integration, storage, application and presentation. The dashboard allows for real-time monitoring of agreed key performance indicators to allow early indication and notification of divergence from baseline

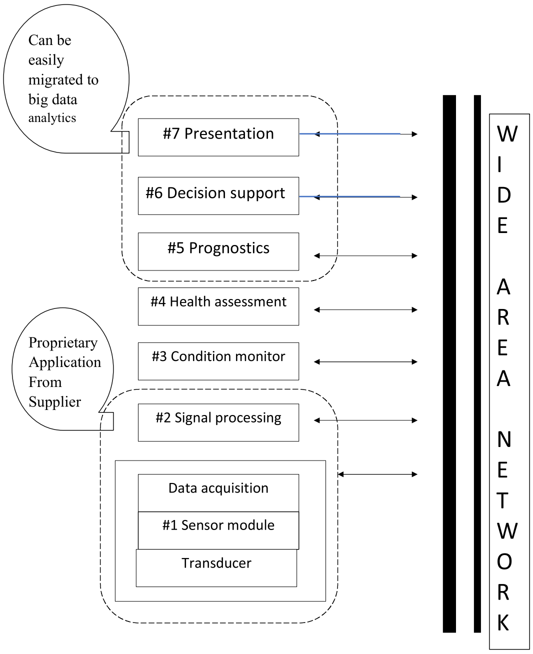
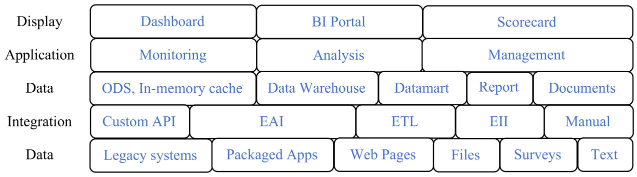


FIGURE 4. OSA - CBM Architecture [70].



Display	Dashboard	BI Portal	Scorecard			
Application	Monitoring	Analysis	Management			
Data	ODS, In-memory cache	Data Warehouse	Datamart	Report	Documents	
Integration	Custom API	EAI	ETL	EII	Manual	
Data	Legacy systems	Packaged Apps	Web Pages	Files	Surveys	Text

FIGURE 5. Overview of technical architecture for data integration and application [71].

performance. From those blocks shown in Figure 5, it identifies how to develop appropriate customer application programming interfaces (API) and other technologies including enterprise information integration (EII), enterprise application integration (EAI) as well as data extraction, transformation and loading (ETL) tools to pull data from source systems.

The type of bespoke applications or services and the analytical requirements are shown in Figure 6 which is modified to suit the rail asset data model. Asset condition information is outputting from each of the applications and exported to the asset management database. In terms of big data analytics, analytics engines are required to be tailored to be compatible with the data formats and calculations performed at the source data collection points. For the MongoDB shown in Figure 6, a document store is used "that does not have any schema restrictions and supports multi-attribute lookups on records" [72].

In terms of a migration strategy between the existing traditional relational database systems storing asset condition data and the NoSQL storage provided by MongoDB, a hybrid approach can be utilized to minimize the risk of losing data during the migration [73].

In each of the rail applications identified in Figure 6, there are separate calculations, in parallel, on the data with

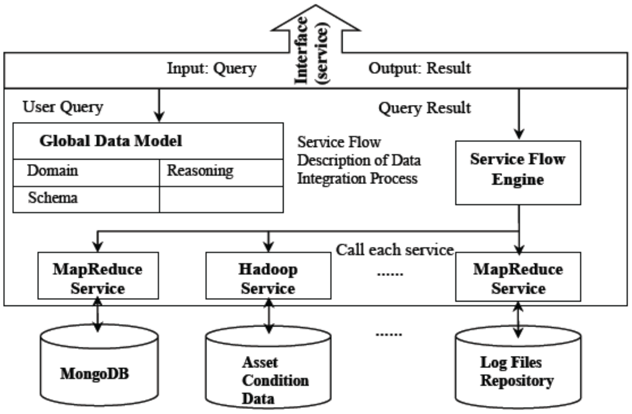


FIGURE 6. Data integration platform [81].

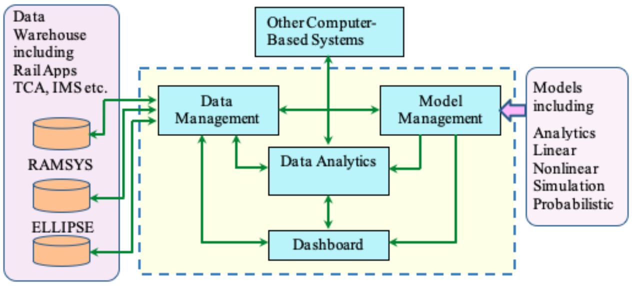


FIGURE 7. Example of data source integration for asset management systems.

common access to the asset register. Asset condition information is output from each of the applications and exported to the asset management database. In terms of big data analytics, analytics engines are required to be tailored to be compatible with the data formats and calculations performed at the source data collection points. Data records from train describer systems are a valuable source of information for analyzing railway operations performance and assessing railway timetable quality [74]. The asset condition database in Figure 6 may be installed by the condition monitoring applications provider. The database choice may be subject to Corporate IT policies within an organization and utilize databases such as Microsoft SQL, Oracle and other commercially available or open-source databases. To ensure compatibility across data sources, a common data management framework [75] has been adopted by different railway infrastructure organizations including CrossRail [76] and Deutsche Bahn [77] to meet asset master data register requirements [78]. This practice has been adopted by rail infrastructure organizations within Australia [79], [80].

Figure 7 provides an idea of the integration of various data sources within the asset management system. Data analytics tools using linear, nonlinear and simulation models such as FlexSim can provide predictive trends for maintenance management. Data integration is another critical task that needs to be handled well in asset management. It requires a common platform to process and manage diverse sources' data.

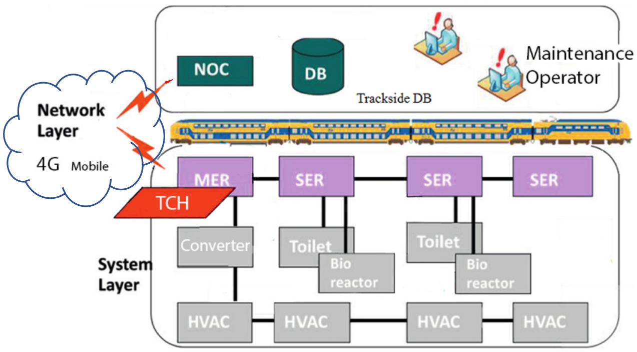


FIGURE 8. Condition monitoring data integration for rolling stock modified from [88].

The data can be divided at component level, system level, and operation level data. It then requires efficient database management using techniques such as data warehouse for data store and data mining for classification, etc.

D. CHALLENGES OF DATA VOLUME MANAGEMENT

Collection and classification of condition monitoring data have become an issue in terms of management of large amounts of data where duplication of resources and data can occur within the asset management discipline [82]. These challenges and complexities include the volume of data in collection, tracking of changes to the assets, tracking and/or recording of maintenance performed on these assets, and willingness to share data within the organization [83]. For rail assets that may consist of 100's of kilometers of rail track including ballast and undertrack infrastructure, the necessity of spatial information to identify exact locations in terms of asset condition is an immense challenge. Converting asset location information to GPS coordinates may utilize time and resources away from maintenance tasks. This challenge is not unique to railway organizations and synergies can be gained from reviewing how other organizations with spatial datasets are meeting the challenge [84]. For complex projects, delivery of the asset management information can be a significant challenge from a big data perspective [85]. The unstructured nature of design delivery with information capture can provide a large amount of unstructured data that needs to be included within an asset management system for configuration change management. In terms of the volume of data, the information technology (IT) infrastructure may require re-engineering to support the volume of data available for collection with each asset operation. The impact of using information and communication technology to achieve the benefits of big data analytics for rail infrastructure assets have been highlighted by Takikawa [86]. This may mean a new platform dedicated to big data analytics [87]. An example of the complexities in providing access to real-time asset condition information for rolling stock is shown in Figure 8 provided from NedTrain [88]. The complexities can include the interrelationship among multiple suppliers for an asset. For legacy assets such as older rolling stock, retrofitting of condition monitoring equipment may be desirable to meet organizational safety and performance requirements. In these particular cases, where there is no train management system (TMS) installed, other modules are required to be implemented such as the train communication handler (TCH) connected to the main equipment room (MER) and the secondary equipment room (SER) modules to pull data from the diagnostic modules and sensors. The TCH connects to the network layer which can be either a cellular 3G, 4G network or Wi-Fi to connect to the trackside layer which includes network operations control center (NOC) and databases (DB) for storage of the condition monitoring data. A further key consideration of the complexities is the utilization of radio frequency identification device (RFID) technologies to accurately identify the location of the asset (GPS plus track location information) with trackside information using real-time data analytics [89]. In terms of real-time data collection and continuous analytics, a paradigm shift may be required from traditional database methods. This could include "keeping analytics results in small-sized tables" [90]. An alternative approach may be to utilize the many task computing paradigm (MTC) for ensemble-based prediction methods [91]. This can be extended to utilizing artificial intelligence or other stochastic methods such as Markov chains for degradation modeling using asset condition data.

Other railway manufacturers including Bombardier and IBM have developed big data analytics for train condition monitoring [92]. These approaches have generally been focused on the collection of data for fault prediction and diagnostic explanation. Transportation systems in large cities such as London are integrating sensor data streams for prediction purposes and transformation [93]. However, data volume management is still challenging. The Public Transport Services Division of the Department of Planning, Transport and Infrastructure in South Australia has utilized an IBM Maximo system for asset management and is progressively rolling out asset condition monitoring systems as part of transportation expansions [94]. The Swedish Railways have implemented *Maintenance 4.0* with the requirement for support of large databases [40]. Scalable data structures are recommended as a requirement to cope with the increasing volumes of data being collected within a railway environment.

E. SUMMARY OF DATA MANAGEMENT AND OTHER ASPECTS

With the volume of data now available, tools and technologies have been developed to extract useful information and patterns from datasets, particularly for spatial data sets where high dimensional data is stored [95]. The complexity of spatial data types, spatial relationships, and spatial autocorrelation has meant that organizations need to develop Spatial Data Analysis (SDA) to handle this complexity [96]. Therefore, intelligent data analysis methods with application of advanced technologies need to be developed. This is the focus of research in the past about 20 years and a number of techniques for condition monitoring data interpretation have been developed (e.g., [97]). None of the techniques, however, is perfect and each is only applicable to certain circumstances. To assist in application of these intelligent data analysis methods, technologies such as streaming data analysis have been developed to cater to the large amounts of data [98].

The idea of missing or incomplete data within spatial data sets undergoing data mining processes has been discussed in Brown and Kros [99]. The impact of missing data on the prediction models can be significant and for this reason, owners of incomplete data sets may be reluctant to share the data for asset prediction purposes [100].

In summary, the key challenges facing organizations managing data on railway infrastructure asset condition include a) handling of different data types based on systems being monitored; b) presence of structured and unstructured data; c) identification of missing or incomplete data sets; d) handling of volume of data collected with asset condition systems to sift through for information content and e) asset data model maturity in measuring conflicting key performance indicators (KPIs). Display of data sets that are missing or incomplete which are required for KPIs may require flags to label the status of these data sets for potential users of the data sets. A further challenge for management of the asset data are bespoke applications provided with the legacy condition monitoring systems that may not be easily migrated to support a big data analytics framework. However, rail organizations will need to consider the future-proofing of asset information using data analytics technologies [101]. Only a small number of railway organizations have implemented big data analytics within their organizations due to the challenges in the management of asset data [16]. However, there is a ''potential for the application of big data techniques to manage railway safety'' [102]. Big data analytics for railways include the utilization of sensor technology for data collection and real-time monitoring of track geometry, waves, joints and sun curves of rail as well as relative height, position, wear and defects of catenary wires. Big data analytics combined with artificial intelligence (AI) has allowed for real-time prediction of alarms and trends within the Swedish railways. This can allow an organization to identify what is critical and what the real problems are. Further discussions on the application of big data are also extended to other aspects such as selection of suitable technologies and tools for some specific applications when facing challenges.

V. BIG DATA ANALYTICS TECHNOLOGIES AND CHALLENGES

Advanced analytic techniques have become available with the advent of big data. At data collection stage, the data to be collected may include unstructured data from heterogeneous sources managed through big data analytics engines such as HADOOP combined with a NoSQL database. At data analytic stage, the techniques utilized may include regression analysis, machine learning, deep learning, Bayesian inference and data mining. The technique selection is closely related to the data characteristics as well as platforms used for data collection, management and processing.

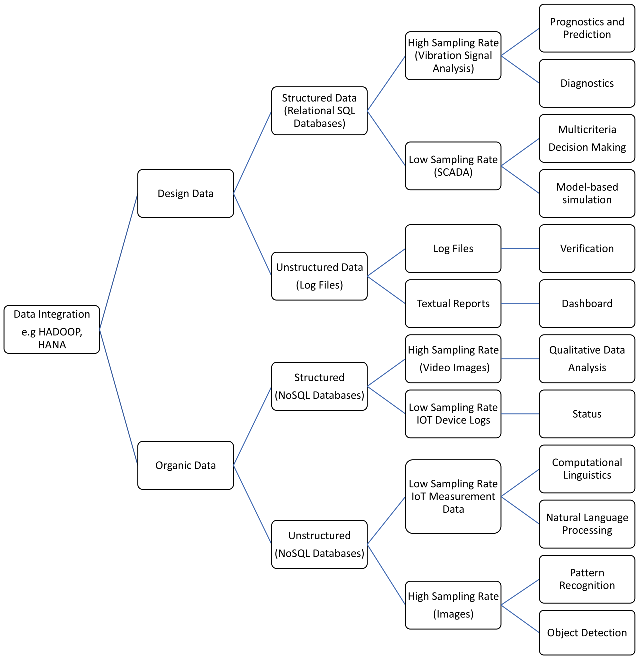

Data Integration e.g HADOOP, HANA
Design Data
Organic Data
Structured Data (Relational SQL Databases)
Unstructured Data (Log Files)
Structured (NoSQL Databases)
Unstructured (NoSQL Databases)
High Sampling Rate (Vibration Signal Analysis)
Low Sampling Rate (SCADA)
Log Files
Textual Reports
High Sampling Rate (Video Images)
Low Sampling Rate IOT Device Logs
Low Sampling Rate IoT Measurement Data
High Sampling Rate (Images)
Prognostics and Prediction
Diagnostics
Multicriteria Decision Making
Model-based simulation
Verification
Dashboard
Qualitative Data Analysis
Status
Computational Linguistics
Natural Language Processing
Pattern Recognition
Object Detection

FIGURE 9. Considerations of asset data types and context.

A. CONSIDERATION OF BIG DATA ANALYTICS APPLICATIONS FOR RAILWAY INFRASTRUCTURE

Before we start to consider big data analytics applications, it would be good to have an overall view of data in integration such as what is shown in Figure 9. The diagram gives an overall picture of asset data divided into different types. The right-hand side of the diagram presents a number of key applications depending upon the types and context of the data. That means data availability determines the potential applications. Identification of the context of data being collected is critical to the usefulness of the data.

The considerations on data collection that impact on the selection of the big data analytics techniques in railway industry may include: a) which part of the track or rolling stock is to be monitored, b) which type of sensor is to be placed and what kind of data (structured, semi-structured or unstructured) is expected from the sensor systems, c) sparsity of the data to be collected and whether the collected data reached the SQL database on time or not, and d) how to deal with bad data (or missing information such as missing tag reader identification). Each of the considerations above can be addressed initially with the concept of ''design data'' where the data requirements are included in the design of the system to be monitored. However, as the system is monitored in a real-time environment, the concept of ''organic data'' is introduced which can include sensor data, sentiment data and various types of machine data that can be characterized by context.

The data context is important in providing the links or relationships between each of the nodes of big data [103]. Big data analytics techniques such as machine learning and Bayesian inference could be used to read monitoring data and also learn data context and correlation (e.g., rolling stock identity vs wheel temperature) for efficient maintenance planning and decision-making.

There are new questions on data analytics in terms of accuracy such as: a) which technique accurately predicts the condition of the asset; b) how to deal with missing data in the modeling and how it will affect the trends produced from the modeling and c) how to manage the data flow where the data may be intermittent based on equipment availability and timeliness requirements of the data. Veracity of the data is of key importance within railway asset management due to the requirement for safety as a priority where uncertainty and reliability of data sources can arise. While large data sets are assumed to be better, they may contain systematic biases or have large amounts of missing information, and even missing key variables which can be magnified with the size of the dataset [104]. Big data sets may be collected under some complex and unknown measurement process [105]. This process needs to be included in the process of design data. Alternatively, the big data set may have been collected for a different purpose and an analysis is being performed to see if any inference can be drawn from this data set [39].

In terms of diversity of data, the key challenges are in matching the diverse data sources to a data management model. Alternate approaches using heterogeneous data integration are provided by [106] where multiple kernel estimators have been proposed to develop an understanding of the underlying data structure for integration into a data model. Data sources such as equipment maintenance logs may not match the structure of the other data sources being imported. Analytics approaches using sentiment analysis and natural language processing may be required for information extraction to associate the unstructured data with a failure event [107]. Further development is required for utilization of these techniques within engineering asset management. Computational analytical solutions, particularly using unstructured data, may not yet be sufficiently developed for safety-critical infrastructure systems [108]. While some failure modes can be identified during the design phase for a system, failure modes may be identified after operational trials [109] which mean additional separate sensors to be used for collecting failure information that are not part of the asset system model. A way of achieving data integration with diverse data sources is to utilize scenario-driven data modeling (SDDM) [110]. This method also allows for the identification of gaps within a data management model to allow adaptation for missing data elements. This approach can also allow for event-driven data collection where the collected data can change depending on an event [111]. If failure modes can be identified after the event, the use of event-driven data analysis and scenario data modeling can help identify gaps within the existing data management model to identify patterns in failure data. The required changes to the data management model can then be modeled using universal modeling language tools (UML) to assess impacts on the data collection requirements. Where failure modes are not readily identifiable, then maintenance staff may need to collect all of the data and search for patterns within the data. An example of an ontology method for data integration with big data is provided by Eine *et al.* [112] where the relationships can be mapped using equivalence rules to automatically deduce the relationships between different data sources. This method can also be prepared with the approach of using ontology for data management discussed earlier.

In the railway asset condition monitoring example, the trackside data collection site can be designated as the primary site while a large amount of data collection process can be designated as the secondary site where the data integration and mapping begin. The other part of the integration process is in identifying an authoritative source uniting overlapping datasets from separate primary data sources, which may supply redundant or conflicting values [113]. For each of the databases (DB) in Figure 6, a separate service is required to pull the data from the DB to the primary source (Global Model). Each of the separate services may utilize proprietary technology based on the origin of the primary source. For example, if the trackside sensor produces sound files with text data, a 'Service' that is suited to importing these types of files may be utilized for import facility to the Global Model. The RailBAM^{TM}application, for example, produces wave files with text files based on the train consistently for each train pass.

Another example of the type of bespoke applications or services and the analytical requirements is given in Oweis [21] where, in each of the bespoke applications identified, there are separate calculations in parallel on the data, with common access to the asset register.

An illustration of a modeling framework for track maintenance is shown in Figure 10 [114]. From this figure, a model is selected for the particular asset being assessed and the live asset condition data is applied to the model with an evidence-based track working model as its output which is stored in the asset management system.

The collection of rail inspection data has increased in volume and quality in recent years and has been presented as a big data analytics challenge [115]. Rail inspection processes have developed from two approaches: a) on-board sensors mounted on rolling stock and b) track-mounted sensors. In these approaches, further discussion is required to ensure data format standards and requirements can be met by the different manufacturers within the rail industry.

The considerations outlined above need to be evaluated for each of the asset classes and types included within the railway infrastructure asset management model.

B. BIG DATA ANALYTICS TECHNOLOGIES

A short review of big data technologies can be found from [116]–[119]. These technologies are related to the

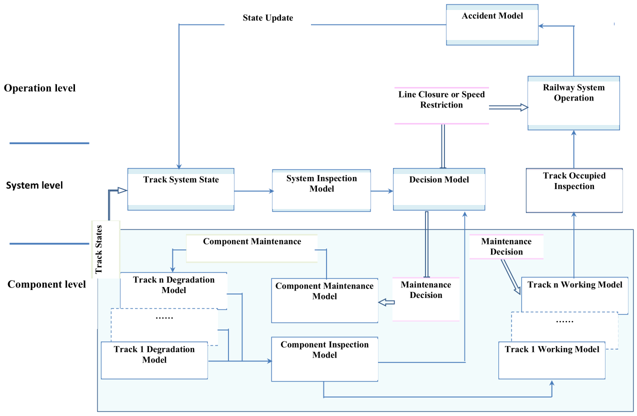


FIGURE 10. Modeling framework for track maintenance modified from [114].

challenges being faced by using big data, namely, volume, complexity and high-speed processing. Advanced analytics technologies are developed to support predictions using scenario developments rather than probabilities as in more traditional analytics technologies. Analytics is about the discovery of new or existing relationships within the data sets and making sense of the data for modeling purposes. Multiple analytics technologies may be utilized within big data applications to support the variety of heterogeneous data being collected. In-memory analysis can be performed during the data collection stage to determine the value and relationships of the data prior to storage [48]. These technologies can be grouped into two key categories of a) storage and b) querying/analysis [120]. Storage relates to the "persistently storing and managing of large-scale datasets" [121]. The technologies associated with data storage and management are also closely related to the frameworks and platforms utilized, e.g., HADOOP frameworks, HDFS and non-relational databases such as NoSQL. Due to the volume and speed, incoming data may be initially split and stored across several machines. Big data storage technologies may be required to split the incoming data across several machines simultaneously for storage and analysis. These technologies and tools are discussed further in Section V.D. Querying and analysis relate to the use of analytical technologies or tools to inspect, transform, and model data to extract value (information). At this stage, big data analytics may be divided as descriptive, diagnostic, predictive and prescriptive analytics with the purpose of application from information learning to insight and further to foresight (optimization) of the process. Technologies utilized for the querying and analysis include data mining, clustering, knowledge discovery, machine learning, MapReduce, Massively Parallel Processing (MPP), Multi-Dimensional On-Line Analytical Processing (MOLAP), visual analytics and statistical models [122]–[125]. Table 1 provides an exemplar of big data analytics technologies for reference.

In each of the examples provided in Table 1, aspects of the analytics technologies may be used in more than one example. Machine learning has been utilized across multiple analytics technologies to increase efficiency. A key challenge being faced by railway organizations in introducing asset condition monitoring sensors is issues with scalability, security, economics, and engineering. Currently, connectivity between railway condition monitoring sensors utilizes a number of mediums for connectivity (see Figure 8). This means that the IoT structure can be regarded as a centralized structure. This can impede the flow of communications between sensor devices and the analysis processing units. To address the limitations of a centralized structure, Blockchain technology has been proposed to assist in decentralized computation and asset condition monitoring by providing a distributed ledger that can record transactions between two parties efficiently and in an authentic way [126]. Further discussions on Blockchain and its applications to the railway distributed processing are provided below.

TABLE 1. Examples of big data analytics technologies.

Big Data Analytics Technologies	Examples
Descriptive Technologies	Examples include OLAP, EAI, EII, ETL with dashboards for display and real-time reports. Data visualization allows for condensing big data into smaller, more useful nuggets of information. Mining historical data to look for the reasons behind past success or failure.
Diagnostic Technologies	Variety of technologies such as drill-down, clustering, data discovery, data mining and correlations. It is to examine data or content to answer the question "Why did it happen?".
Predictive Technologies	Variety of statistical, modeling, data mining, machine learning, and sentiment analysis technologies (natural language processing, text analysis, computational linguistics) to study recent and historical data. Predictive analytics can be probabilistic in nature.
Prescriptive Technologies	Variety of technologies such as graph analysis, simulation, complex event processing, neural networks, recommendation engines, heuristics, and machine learning. Prescriptive technologies can be considered to be predictive technologies with additional components for actionable data and feedback that tracks the outcome produced by the action taken to provide optimization and then suggest decision options to take advantage of the predictions.

C. BLOCKCHAIN TECHNOLOGIES

1) CURRENT APPLICATION STATUS OF BLOCKCHAIN TECHNOLOGY IN RAILWAY INDUSTRY

Blockchain can be described as a shared distributed ledger [127] with encryption to provide authenticity [128]. For distributed parallel processing of data coming from multiple sensors, there is a problem arisen that is the trust between sensors and processing components and the generation of digital signatures for each of the sensor devices. Blockchain can be utilized to build a trust relationship using a smart contract concept [129]. An alternative decentralized system to register and assign IoT devices to an owner based on the Blockchain technology has been proposed by Ghuli *et al*. [130]. In this approach, initial ownership is provided by the manufacturer of the device and then being transferred to an owner based on Blockchain technology. However, it is not clear with this approach if replacement of the failed devices would have occurred with updated registration of devices.

Blockchain can be utilized to provide parallel processing and communications architecture where the data flow is decentralized and security is improved [131]. This can also decrease response times where data flows can be directed to the closest node where the data may be utilized. Data on track status such as faults or obstacles including location information can be shared between the track devices and the train (rolling stock) without first sending the information to a central device (Rail Control Centre). The same data can be sent to the central device in parallel using Blockchain technology to verify the receipt of the data [132]. The types of applications currently being reviewed by Deutsche Bahn with Blockchain include safety applications where track obstacle reporting is provided directly to trains, and track condition monitoring and maintenance systems [133]. Enterprise asset management systems such as the IBM Maximo product currently support Blockchain for asset management applications [134]. Pacific National within the Australian rail freight network has been utilizing Blockchain for supply chain management of perishable goods [135]. This involves the tracking of the perishable goods across the network with multiple partners. The approach adopted within the rail industry has utilized a similar approach to that outlined for the energy sector in [136]. Other examples representing the current application status of Blockchain technology in railways are summarized as shown in Table 2.

TABLE 2. Summary of current application examples of Blockchain in railway industry.

Name of Company or Organization	Examples
Swiss Federal Railways	Tested a worker identity management system ([137], [138]).
Blockchain in Transport Alliance (BiTA)	Making efforts to develop new framework and standards for transportation companies ([139], [140]).
Go-Ahead Group Plc in the UK	The company is reportedly partnering with Blockchain start-up DOVU to launch a tokenized, Blockchain-based rewards system for its customers ([141]).
Russian Railway (RZD)	Launched the Freight Transport, an electronic trading platform underpinned by Emercoin and more than 5000 freight consignments ordered via the platform over a nine-month period [142]. The company is going to implement Blockchain applications for ticket sales and smart contracts all in crypto ([143], [139]).
Bourque Logistics in the US	Unveiled the RAILChain™ platform to explore the application of Blockchain technology for its rail shipper clients. The platform is designed to enable shippers to securely exchange bill of lading information, as well as settle freight, repair and lease costs using smart contract technology [144].
The State Railway of Thailand	Is investing Blockchain to manage signalling, passenger information systems, ticketing and goods delivery [139].
Shenzhen Metro in China	Launched the first-ever Blockchain-based digital invoicing system in March 2019 [145].

2) CONSIDERATION OF THE FUTURE APPLICATIONS OF BLOCKCHAIN TECHNOLOGY IN RAILWAYS

The data flows in a railway network can be expressed using an IoT architecture as shown in Figure 11, in which the Rail IoT Connections represent parallel distributed connections between track, trains and station staff; and the Rail Network Connections represent the centralized communications path. Thus, the asset condition data collected from trackside sensors may be received twice and arbitration may be required to distinguish between the two paths for the same data. However, the data available in a single public distributed ledger shared among several parties can be more reliable than multiple centralized databases. A time-stamped version of the 'data' can be available on the distributed ledger to show the arrival of the data at various devices for audit purposes.

The data flows can be enabled more effectively with Blockchain technology. For example, the verification of the asset condition status by the asset management system after maintenance intervention by trackside staff can be achieved using Blockchain technology and IoT network in Figure 11. Each of the devices can be arranged into a peer-to-peer (P2P) connection where the Blockchain is a decentralized, distributed ledger (public or private) of different kinds of transactions arranged into each device. The data cannot be altered without the consensus of the whole network. Track location

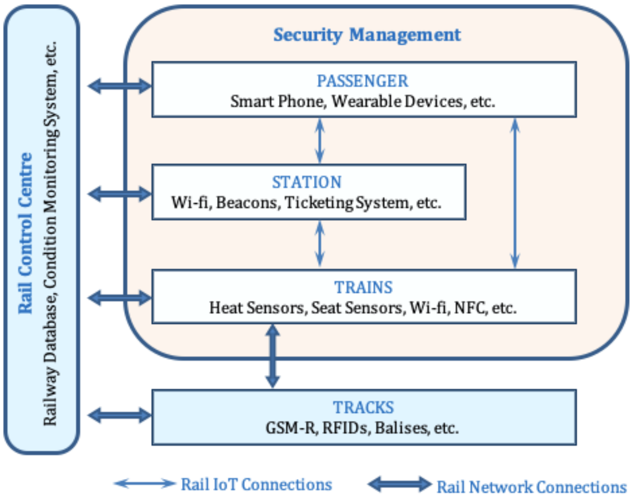

Rail Control Centre
Railway Database, Condition Monitoring System, etc.
Security Management
PASSENGER
Smart Phone, Wearable Devices, etc.
STATION
Wi-fi, Beacons, Ticketing System, etc.
TRAINS
Heat Sensors, Seat Sensors, Wi-fi, NFC, etc.
TRACKS
GSM-R, RFIDs, Balises, etc.
Rail IoT Connections
Rail Network Connections

FIGURE 11. Rail IoT Architecture [146].

information with vehicle RFID information can be shared between trackside staff and the Rail Operations Centre for decision-making with the assurance that the information has been validated without manual verification processes being utilized. Work orders can be sent directly to maintenance staff for allocation of maintenance resources without human intervention. The utilization of Blockchain technology allows the verification of the source of the information as a trusted source in real-time with distributed processing and storage of the data by big data analytics. Different communication mediums such as WIFI, 4G mobile and trackside communication systems can be utilized directly as available without the concern about the security of the communication mediums.

Given the description above, it can be foreseen that the applications of Blockchain technology in railway network management will be promising in the near future. These include smart maintenance with better data tracking for improving safety and reducing service costs, transit payment and ticketing for improving service and cutting customer costs. To sum up, Blockchain technology is at the start-up stage [147]. "The opportunities are exciting but companies will need time to familiarize themselves with the technology and to identify the best prospective application areas which are supported by a universally-accepted set of standards" [142]. These, however, are still in the developing process.

Blockchain has its strengths and advantages in application but the speed and scalability would be remaining as a big concern in developing Blockchains for a busy network.

D. BIG DATA ANALYTICS TOOLS AND OTHER ASPECTS

Tools have been developed to manage the five characteristics of big data described above and the five-stage process which forms two main sub-processes of data management and analytics [148]. These tools can allow for parallel processing and loading of chunks of data to assist in key activities for the handling of the large volume and diverse types of data. Examples of these tools include MapReduce which allows, for example, splitting of the input data-set into independent chunks which can then be processed in a completely parallel manner [69], [107], [149]. Python is commonly used when data analysis tasks need to be integrated with web apps or if statistics code needs to be incorporated into a production database [150]. The utilization of Python for big data analytics is generally via a framework using Python libraries and tools [151]. Each of these Python frameworks has characteristics that are designed for specific models. Other tools such as Regular Expressions (RegEx) can be used to detect and repair errors for data input as part of the pre-processing stage [152]. The requirement for expert systems or artificial intelligence was identified to pre-process the large amounts of data [70]. This approach has been further developed with the concept of machine learning [153]. The Machine Learning Library (MLIB) is utilized to provide a number of machine learning algorithms that can be implemented within big data infrastructure [154]. However, other machine learning tools such as backpropagation can be parallelized with big data analytics such as MapReduce [155]. In this example, the original data information is maintained in the data subset, which can be useful for verification purposes. For the types of data collected by condition monitoring systems, different tools may be selected to process the data from the raw to the processed data stage (e.g., semantic analysis for the raw text files). Semantic analysis for message analysis can be done with big data analytics [156] and semantic analysis can be performed using support vector machines (SVM) where the accuracy of sentiment classifications can be improved [157]. The semantic analysis process developed could be applied to unstructured data such as log files or status messages provided as part of the asset condition monitoring process.

E. BIG DATA ANALYTICS TOOL SELECTION AND CHALLENGES

The challenges for the use of big data within an asset management environment relate to a) storage of the diverse and also heterogeneous types of data; b) management of the data collection process to ensure the accurate and reliable collection of data; c) managing the changing relationships between the data and the assets that the data describes and d) inconsistency and incompleteness of the data [158]. These challenges include, for example, handling of heterogeneity, dealing with inconsistency and incompleteness, merging data, timely process and analysis, and ensuring privacy and data ownership [159]. Track geometry measurements (see Figure 10) in particular are described as requiring a high sample rate to capture the trends for prediction and analysis purposes. Analyzing large volumes of data in real-time may require the combination of in-memory processing with analysis by advanced machine learning techniques. Companies have had to deal with profound changes in the use of technologies for big data analytics [160]. The selection of different unconventional tools such as Python and MongoDB to meet big data challenges has meant that companies have had to adapt to changes in the enterprise data model and data analytics [161]. The use of JavaScript Object Notation (JSON) data format with NoSQL databases such as MongoDB allows for efficient analytics functionality [162].

The selection of tools for managing big data is often described as a collection of related techniques and tool types. Those are usually utilized for predictive analytics, data mining, statistical analysis, and database management that support analytics [32] such as MapReduce with in-database analytics, in-memory databases, and columnar data stores. As most traditional tools and algorithms are regarded as inefficient, the development of new algorithms is required. The design of an asset management system at this stage may appear to be different from that expected for a conventional asset management system using off the shelf applications and tools. Python and R programming languages can be used to implement machine learning (ML) techniques to support complex degradation models requiring sufficient data. This can help improve the outcomes of complex maintenance decisions [163]. The benefits of migrating to new tools, particularly in a data-rich environment can be large [164].

While big data has been applied to transit forecasting and ridership forecasting within rail systems [27], the use of big data analytics for predictive maintenance is in its infancy. Network Rail has utilized Deloitte's suite of cloud-based analytics tools to provide real-time timetable information for timetable management [165]. The development of advanced analytics for train delay prediction using exogenous data is provided by Oneto *et al.* [166] where multivariate statistical concepts are implemented using big data analytics tools, and improvements in train delay prediction using advanced data analytics including multivariate statistics over traditional delay prediction methods are discussed. Further work is suggested for inclusion of railway asset condition data to improve prediction accuracy. Within rail infrastructure environments, energy efficiency has been given increasing priority to improve sustainability. An example of predictive analytics using artificial intelligence (AI) for rolling stock power consumption on the railway network has been given by Furutani *et al.* [167]. While AI has been regarded as being in its infancy within rail infrastructure networks, impetus is being provided to roll out AI to the rail sector. The key challenge identified here is the real-time processing of the stream of data from sensors. This is also the case for other industries such as health care and infrastructure organizations [10]. A key item is identified as the maturity of the available big data platforms and analytics software which are still regarded as being in their infancy [125]. This includes concerns about scalability to meet increasing data volumes and substantial real-time requirements for asset condition monitoring. The use of big data analytics for transit forecasting analysis of human mobility has been given by [168]. This has applications for transportation passenger utilization prediction using passenger mobile phone location data. Trials of this approach are currently being conducted in the US and Europe to assist in passenger load forecasting and congestion reduction. Access to passenger information can be difficult when evaluating the value of rail infrastructure investments. There is currently a digital transformation focus on railways in terms of utilizing big data for better management of resources including passenger utilization of assets [169]. The recent development of European standard (EN12896) for a reference data model for public transport information has provided for sharing of timetabling, fares, operational management, real-time data, and journey planning across railways [170]. In addition, the AAR has published a set of standards for data handling to meet railway operational requirements [171].

In terms of rail maintenance applications, an example of using machine to machine (M2M) procedures with a customized condition-based maintenance platform to handle big data challenges has been proposed by Palem [172]. For this application, real-time data collection and analysis have been emphasized for fleet maintenance prediction purposes. A reference architecture (platform) is proposed where Python and R have been proposed for the data analytics engines. In particular, R-frameworks can be utilized where the analyst does not need to know the underlying intricacies of the big data architecture or data model [173]. A decision support system using smart data for a railway metro system is provided by He *et al.* [174]. In this case, the use of data from different sources for the decision support system is described and compared with the innovation requirements to support a large-scale railway metro system. This has supported a change in focus from equipment-centric to asset-operation centered. The concept of smart maintenance decision support systems is being trialed in the US where the data analytics is used to extend the linkage between the analysis of condition monitoring data and statistical trending with prediction and simulation-based scenarios [175]. The availability of large volumes of data to build accurate simulations of complex systems can extend the predictive maintenance concept within large infrastructure organizations.

Challenges have also been identified in capturing changes in assets particularly when delivering a major piece of infrastructure such as a new railway in the era of big data [85]. Related IoT technologies are being deployed within industrial asset management environments to provide the connectivity for big data [176] through the improvement of the data acquisition process for condition monitoring sensors. The choice of big data analytics tool is then based on the IT & technology policies of the particular organization undertaking the data analytics development. In addition, the security and privacy of the stored data have become a challenge as the volume of data has increased. Automation and centralisation of the processing of the data with the added security and privacy requirements may require a re-think on IT strategy [177]. Where datasets are heavily skewed, changes in the clustering algorithms such as DBSCAN for parallel processing may be required [178]. In addition, a reference architecture may need to be developed for implementation [179]. There must be very clear requirements for implementation of big data analytics [180]. Identifying the underlying structure and key information is important in achieving the value of big data analytics.

In summary, rail organizations are focusing on the development of big data analytics framework for railways while technology choices are being made by individual organizations based on requirements [181].

VI. OVERVIEW OF BIG DATA ANALYTICS APPLICATION TO RAILWAY INFRASTRUCTURE

Big data analytics within the railway environment is still in its infancy as highlighted earlier. Key applications of big data analytics within railway infrastructure include decision making for infrastructure maintenance based on available condition monitoring data as well as operational performance and train timetabling data. The data processing requirements for the collection of decision-making data include integration of data from heterogeneous sources where low latency requirements for decision making need to be provided through parallel architectures. These include track circuit status monitoring where data flows may be triggered on train movements through track sections in short time periods of seconds or more. Other approaches include the collection of rolling stock condition information for condition analysis to provide a plan for maintenance when the rolling stock reaches a destination.

A. BRIEF DISCUSSION ON APPLICATIONS OF BIG DATA ANALYTICS FOR PROGNOSTICS (PHM)

As organizations move towards predictive maintenance programs away from planned and reactive maintenance, the importance of prognostics and health management (PHM) has increased. Detection of anomaly patterns within PHM can be used to detect the existence of a fault before a failure happens [182]. Key aspects of the prognostics models available have been identified by Peng *et al.*[183]. Hybrid approaches using machine learning using a combination of two or more algorithms to model the system were suggested. The four dimensions of prognostics, i.e., a) sensing, b) prognosis, c) diagnosis and d) management have been described by Kwon *et al.* [184] where prognosis requires additional data including maintenance history, operational and performance parameters that were not previously available with sensors or diagnosis but are available with data integration from heterogeneous sources. PHM has been described as a recent advance that can allow for a better understanding of the degradation process of a component and a system so as to extend and manage efficiently the life duration of industrial systems [185]. The PHM methodology allows for the utilization of remote sensing data, condition monitoring data and, the interpretation of environmental, operational and performance parameters to indicate systems' health status [186]. The availability of data from multiple sources such as condition monitoring systems and reliability analysis systems has allowed for the use of regression, degradation model, support vector machines (SVM), neural networks, and other Artificial Intelligence (AI) techniques particularly with decision support and decision-making models for PHM [164], [187]. A discussion of the model-driven and data-driven approaches for PHM using artificial intelligence has been conducted by Schwabacher and Goebel [188], where verification and validation of the diagnostic models are difficult to achieve before deployment. A key reason for this is that the failure modes can occur which are not identified in the model-driven approaches prior to deployment as well as the lack of historical state (condition) data that can impact on the data-driven approaches. Choice of algorithms to be used for PHM will be based on the level of complexity required of the model and the amount of noise presented in the data [189], [190]. Another key challenge faced in using PHM methods is in ''taking uncertainty into account'' [191]. However, a key challenge with machine learning in prognostics health monitoring based on analysis of asset condition data is the handling of high dimensional data sets. Using domain knowledge that faults and their implications correlate to a type of information contained in an asset's life cycle data and are translatable to a type of domain knowledge representation with an entropy measure can be utilized to provide a dimension reduction framework [192]. Alternative approaches for dimension reduction using principal components with machine learning have been identified by Gorban and Zinovyev [193]. Table 3 below provides a summary of the application of big data analytics with prognostics health monitoring.

B. DATA MINING

Traditional techniques such as data mining have also been utilized with big data analytics to provide improved outcomes for condition-based maintenance and rail infrastructure asset management. Predictive analytics using data mining with big data has been identified as a future need where real-time analysis will be the key to meet scenario-based analysis techniques [194]. The utilization of technology for onboard train equipment can assist in the collection of condition data for analysis using data mining techniques. In the example provided by Sammouri *et al.* [195], the association of Spatio-temporal data is utilized to determine if a significant asset condition related event has occurred to trigger the mining of the data associated with this event for analysis. Condition-based maintenance may require a continual monitoring of an asset leading to volumes of data requiring real-time analysis. A challenge for rail operators is in being able to predict a defect beforehand so that the failure does not occur in operation and does not cause a train delay [196]. To meet this challenge, a spatio-temporal-nodal model has been proposed to analyze the interdependencies of various rail systems and identify root causes for failures and system behaviors.

An approach using big data streaming analysis has been provided by Fumeo *et al.* [98]. In this approach, the continual monitoring is akin to a prognostic approach where triggering of maintenance actions occurs when degradation is detected. However, older data in some cases may no longer be available

TABLE 3. Summary of application of big data analytics with prognostics health monitoring.

Application	Type of condition monitoring sensors	Analytics approach
Rail condition monitoring	Vehicle-mounted sensors	Real-time analytics comparing with a model for diagnostics of deviations from track geometries. Environmental data can be added to predict when track geometries may require attention
Train performance monitoring	Trackside and vehicle-mounted sensors; wheel flats measured by impact sensors mounted on track	Real-time analytics comparing with different models for threshold alarms. Hybrid data approaches used to enhance existing models. Alarm threshold information can be provided to train operators in real-time for planning maintenance intervention
Traction Overhead Wire Sag measurement	Trackside and vehicle-mounted sensors; sag measured by deviation from known height using CCTV images	Real-time analytics of sensor and CCTV images to measure traction wire sag while train is passing under overhead wire. Environmental data can be added to predict when overhead wire sag could reach a threshold.
Railway Points monitoring	Trackside audio measurement sensors collecting audio data to efficiently detect and diagnose faults in railway condition monitoring systems.	Real-time analytics of sensor and audio data to detect anomalies while a train is crossing a set of points using support vector machines (SVMs).
Foreign object detection on track	Vehicle-mounted sensors using CCTV images to detect objects in front on track	Real-time analytics using convolutional neural network (CNN) of CCTV images to detect objects while train is travelling on track Radar and location data can be added to predict when objects may be present.

to revise earlier suboptimal models. Further strategies including data mining may be required to assist in the data analysis process. An example provided by Cannarile *et al.* [197] outlines how data mining for a large number of assets can be used to optimize the maintenance strategies for a particular group of railway infrastructure assets. Population-based strategies were compared with cluster-based strategies for a group of 30,000 switch point machines. While improvements in analysis were reported when using cluster-based strategies, further work in the selection of the starting point or decision variables for the cluster analysis is required. This is consistent with the traditional usage of data mining and clustering algorithms. It may be difficult to choose an appropriate clustering algorithm for use in specific big data sets without detailed knowledge of the characteristics of the big data set [198].

C. SUMMARY

Big data analytics with the increased availability of data has provided tools for the analysis of more complex systems, while also improving the outcomes for analysis of existing systems [187]. However, the increased availability of selection of tools has complicated the task of selection of the right tools for big data analytics with rail infrastructure assets. The selection of methods for a practical application will be based on the availability and quality of the data for the analytics process. However, there is a tendency towards the hybrid approach due to the limitations with the model-driven approach, in which the model-driven approach is used as a starting point for the hybrid approach. While the volume of big data is increasing, the challenge is in being able to select the right data to describe the asset condition information and also in undertaking degradation modeling analysis. Selection of machine learning techniques may impact on the scalability of the applications for increasing data volumes and high dimensionality. Each selected algorithm may have a point where performance starts to drop with increasing volumes of data. Hybrid approaches to analysis techniques may be required. Uncertainty is built into the data and needs to be accommodated in the collection and analysis of the data. Big data analytics algorithms need to be developed further in order to make full use of information hidden in big data. Techniques are crucial to reducing uncertainty in big data analytic process for improving the performance of prediction. Further research has been proposed for the utilization of deep learning techniques to recognize model limitations and improve the prediction accuracy when data quality is poor, or missing data patterns are evident within the data collected where the concept of platform analytics including machine learning and expert systems is introduced for optimization of timing and types of maintenance to be performed for different rail infrastructure assets [86].

VII. CONCLUSION

Big data analytics for rail infrastructure management is not fully mature, while condition monitoring systems are in general use for asset condition classification purposes. However, there is no agreed common method of application or use of algorithms for asset condition assessment. There is also no agreed interface for integration of the types of information utilized for measurement of the asset condition with asset management systems for big data analytics. Issues, challenges, and future research directions are summarized below.

A. GAPS AND ISSUES IN CURRENT RAILWAY ASSET MANAGEMENT WITH APPLICATIONS OF BIG DATA ANALYTICS

The collection and updating of asset condition data at organizational levels of control are extremely time-consuming. Collection and classification of condition monitoring data is an issue in terms of management of large amounts of data and duplication of resources within the asset management discipline. This may involve the collection of historical data with current component failure data to uncover new patterns for prognostics applications. A change in approach from model-driven to data-driven algorithms may be required to analyze data/information collected from the system(s) and the operational environment. The addition of unstructured data into the mix and the requirement to identify relationships for decision-making purposes can increase the complexity of the challenge. Timelines of the analysis may be difficult with the volume of the asset condition data being collected. Smart frameworks with machine learning combined with, for example, fuzzy systems models can be used to optimize the decision-making based on the data-driven inputs [199]. However, the proprietary nature of the condition monitoring systems used to collect the data can restrict the ability of end-users to integrate condition monitoring data from different sources to provide aggregate views of asset condition. This has led to the development of open systems architecture for condition-based maintenance. Condition monitoring systems data collection can provide timely data for maintenance planning at scales appropriate to a variety of maintenance activities. The problems of quality control and quality monitoring from the perspective of condition monitoring data classification by non-experts require development of new approaches. Disparate and large registers of all asset types, most of which are remote, are all subject to annual verification primarily motivated by financial accounting concerns. The volume of data collected is so huge that it has become humanly impossible to do any intelligent data analysis. Hence, there is an increased reliance on automated big data analysis tools which requires an increased level of testing and verification. A lot of managerial tasks including research works need to be done in a large railway organization before a condition-monitoring database is established and can be used in assisting in asset management decisions.

Evidence-based asset management models are currently being developed to meet the requirement for justification of investment decisions in assets. Applying big data analytics design methods can provide potential values in the transportation field in terms of cost, risk reduction, delay reduction and optimization. A well-developed evidence-based database is very helpful to transportation executives and researchers as it is more strategic by taking advantage of proven and useful information. Utilization of evidence-based approaches with risk management requires a great use of condition assessment data. This can help in the identification and measurement of benefits achieved with the optimization of the asset under management. One of the key tasks confronting the asset management custodian is the selection of the approach for big data analytics techniques for classification of condition data. Approaches utilizing big data analytics with design data combined with organic data from measurements during the asset lifecycle may be required including a selection of big data technologies discussed earlier. Asset integrity reports utilizing asset condition data are required for compliance with the relevant standards or regulators' requirements within Australia and overseas. Current approaches do not provide the required accuracy levels for classification of condition monitoring data where there are several asset condition classes. Hence further development of autonomous approaches as outlined using artificial intelligence may assist in meeting the required accuracy levels required by organizations for safety-critical assets.

B. THE FUTURE OF BIG DATA IN INFRASTRUCTURE ASSET MANAGEMENT

A key challenge facing railway asset owners and operators is in being able to view the interdependencies of various systems that are a part of the railway network and identify root causes for failures and system behaviors before delays occur [200]. This may also involve big data mining with spatial and Spatio-temporal properties to provide visualization of trends and relationships between assets. The spatial dimension would involve the real-time location of key assets and their respective systems and sub-systems in each key asset. Technologies exist to provide the real-time location information, but generally are stored separately to measurement and log data. The temporal dimension would involve the display of historical and current asset condition measurements with performance criteria as a sub-category. The temporal dimension would require the integration of design and organic data to a common asset data model with machine learning techniques utilized to identify the relationships. While there would be value in displaying asset performance in the two dimensions, namely spatial and temporal, the further value could be derived if a third dimension could be added to show the interdependencies and hierarchy of various systems that make up the network. The third dimension would also involve artificial intelligence with expert systems and machine learning utilized to identify the interdependencies and hierarchies. While building information management (BIM) systems exist for fixed assets, a key challenge will be in integrating the fixed asset information with mobile asset information using heterogeneous data sources. The development of scenario-based modeling with PHM would also become a necessary adjunct to big data analytics in railway asset management.

C. RECOMMENDED RESEARCH DIRECTIONS IN USING BIG DATA ANALYTICS

Research efforts are required to investigate and develop commonly agreed processes and methods for application of big data analytics as well as well accepted tools for integration of asset data including condition monitoring data in railway asset management systems.

Big data analytics technologies are still in their infancy within the IT pillars of rail organizations. It is recommended that the research efforts could be directed to define potential big data analytics frameworks, and to integrate different big data approaches for condition and failure monitoring.

The introduction of big data analytics to the railway sector will require further development of tools and models for investigation of asset condition status such as using multistate degradation modeling and correlation analysis of management activities in asset lifecycle.

Application of Blockchain technology in the railway industry is at the start-up stage but the research direction is promising. The future research efforts are directed at developing new frameworks and standards for the industry to use and easily implement their Blockchains. New technologies will be explored to handle the issues related to the required speed and scalability when developing Blockchains for a complicated and busy network.

On the basis of the work presented in this paper, the next step would be directed at providing a review of application of big data analytics to PHM for railway communication networks, infrastructure and transportation systems.

REFERENCES

[1] M. Cox and D. Ellsworth, "Managing big data for scientific visualization," *ACM Siggraph*, vol. 97, pp. 21–38, Aug. 1997.

[2] J. R. Mashey, "Big data and the next wave of infrastress," Ph.D. dissertation, Dept. Comput. Sci. Division Seminar, Univ. California, Berkeley, Berkeley, CA, USA, Oct. 1997.

[3] S. Bryson, D. Kenwright, M. Cox, D. Ellsworth, and R. Haimes, "Visually exploring gigabyte data sets in real time," *Commun. ACM*, vol. 42, no. 8, pp. 82–90, Aug. 1999, doi: 10.1145/310930.310977.

[4] T. J. Berners-Lee. *Information Management: A Proposal*. Accessed: Apr. 2019. [Online]. Available: http://cds.cern.ch/record/369245/files/dd-89-001.pdf

[5] *CERN Website*. Accessed: Apr. 2019. [Online]. Available: https://home.cern/science/computing/birth-web/short-history-web

[6] F. Diebold. (2012). *On the Origins and Development of Big Data: The Phenomenon, the Term, and the Discipline*. Accessed: Mar. 16, 2016. [Online]. Available: https://economics.sas.upenn.edu/sites/economics.sas.upenn.edu/files/12-037.pdf

[7] C. Philip Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Inf. Sci.*, vol. 275, pp. 314–347, Aug. 2014, doi: 10.1016/j.ins.2014.01.015.

[8] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, "The rise of 'big data' on cloud computing: Review and open research issues," *Inf. Syst.*, vol. 47, pp. 98–115, Jul. 2015, doi: 10.1016/j.is.2014.07.006.

[9] A. Siddiqa, A. Karim, and A. Gani, "Big data storage technologies: A survey," *Frontiers Inf. Technol. Electron. Eng.*, vol. 18, no. 8, pp. 1040–1070, 2017, doi: 10.1631/fitee.1500441.

[10] P. Grover and A. K. Kar, "Big data analytics: A review on theoretical contributions and tools used in literature," *Global J. Flexible Syst. Manage.*, vol. 18, no. 3, pp. 203–229, Sep. 2017, doi: 10.1007/s40171-017-0159-3.

[11] A. Oussous, F.-Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "Big data technologies: A survey," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 30, no. 4, pp. 431–448, Oct. 2017, doi: 10.1016/j.jksuci.2017.06.001.

[12] A. Kamilaris, A. Kartakoullis, and F. X. Prenafeta-Boldú, "A review on the practice of big data analysis in agriculture," *Comput. Electron. Agricult.*, vol. 143, pp. 23–37, Dec. 2017, doi: 10.1016/j.compag.2017.09.037.

[13] J. Luo, M. Wu, D. Gopukumar, and Y. Zhao, "Big data application in biomedical research and health care: A literature review," *Biomed Inf. Insights*, vol. 8, pp. 1–10, Jan. 2016, doi: 10.4137/BII.S31559.

[14] S. Tiwari, H. Wee, and Y. Daryanto, "Big data analytics in supply chain management between 2010 and 2016: Insights to industries," *Comput. Ind. Eng.*, vol. 115, pp. 319–330, Jan. 2018, doi: 10.1016/j.cie.2017.11.017.

[15] K. Nagorny, P. Lima-Monteiro, J. Barata, and A. W. Colombo, "Big data analysis in smart manufacturing: A review," *Int. J. Commun., Netw. Syst. Sci.*, vol. 10, no. 3, pp. 31–58, 2017, doi: 10.4236/ijcns.2017.103003.

[16] F. Ghofrani, Q. He, R. M. Goverde, and X. Liu, "Recent applications of big data analytics in railway transportation systems: A survey," *Transp. Res. C, Emerg. Technol.*, vol. 90, pp. 226–246, May 2018, doi: 10.1016/j.trc.2018.03.010.

[17] H. J. Parkinson and G. Bamford, "The potential for using big data analytics to predict safety risks by analyzing rail accidents," presented at the 3rd Int. Conf. Railway Technol., Res., Develop. Maintenance, Cagliari, Italy, Apr. 2016.

[18] J. R. Kennedy and J. D. Parrish, "Total system support concept for railway signalling systems," presented at the Railway Eng. Conf., Adelaide, SA, Australia, Sep. 23–25, 1991.

[19] *Asset Management—Management Systems—Guidelines for the Application of ISO 55001*, Standard ISO 55002:2018, International Standards Organization, 2014.

[20] P. Scarf, R. Dwight, A. Mccusker, and A. Chan, "Asset replacement for an urban railway using a modified two-cycle replacement model," *J. Oper. Res. Soc.*, vol. 58, no. 9, pp. 1123–1137, Sep. 2007, doi: 10.1057/palgrave.jors.2602288.

[21] R. Oweis, "Aligning asset management policy, strategies and proceses to deliver effective and sustainable results," presented at the ICOMS Asset Manage. Conf., Gold Coast, QLD, Australia, May 16–21, 2011.

[22] K. J. Epp and P. J. Mutton, "Wheel/rail interaction: Current'state of the art'in the Australasian railway industry." presented at the CORE Cost Efficient Railways Through Eng., Wollongong, NSW, Australia Nov. 13, 2002.

[23] M. Palo, D. Galar, T. Nordmark, M. Asplund, and D. Larsson, "Condition monitoring at the wheel/rail interface for decision-making support," *Proc. Inst. Mech. Engineers, F, J. Rail Rapid Transit*, vol. 228, no. 6, pp. 705–715, Aug. 2014, doi: 10.1177/0954409714526164.

[24] C. A. Lengnick-Hall and R. J. Griffith, "Evidence-based versus tinkerable knowledge as strategic assets: A new perspective on the interplay between innovation and application," *J. Eng. Technol. Manage.*, vol. 28, no. 3, pp. 147–167, Jul. 2011, doi: 10.1016/j.jengtecman.2011.03.003.

[25] D. S. Michele and L. Daniela, "Decision-support tools for municipal infrastructure maintenance management," *Procedia Comput. Sci.*, vol. 3, pp. 36–41, 2011, doi: 10.1016/j.procs.2010.12.007.

[26] S. Jovanovic, "Modern railway infrastructure asset management," presented at the 24th Southern African Transp. Conf. (SATC), Pretoria, South Africa, Jul. 11–13, 2005.

[27] A. Thaduri, D. Galar, and U. Kumar, "Railway assets: A potential domain for big data analytics," *Procedia Comput. Sci.*, vol. 53, pp. 457–467, 2015, doi: 10.1016/j.procs.2015.07.323.

[28] C. D. Cottrill and S. Derrible, "Leveraging big data for the development of transport sustainability indicators," *J. Urban Technol.*, vol. 22, no. 1, pp. 45–64, Jan. 2015, doi: 10.1080/10630732.2014.942094.

[29] E. Ngai, Y. Hu, Y. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decis. Support Syst.*, vol. 50, no. 3, pp. 559–569, Feb. 2011, doi: 10.1016/j.dss.2010.08.006.

[30] W. A. Günther, M. H. Rezazade Mehrizi, M. Huysman, and F. Feldberg, "Debating big data: A literature review on realizing value from big data," *J. Strategic Inf. Syst.*, vol. 26, no. 3, pp. 191–209, Sep. 2017, doi: 10.1016/j.jsis.2017.07.003.

[31] *The Four V's of Big Data*. Accessed: Jun. 2019. [Online]. Available: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

[32] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 137–144, Apr. 2015, doi: 10.1016/j.ijinfomgt.2014.10.007.

[33] H. Wang, Z. Xu, H. Fujita, and S. Liu, "Towards felicitous decision making: An overview on challenges and trends of Big Data," *Inf. Sci.*, vols. 367–368, pp. 747–765, Nov. 2016, doi: 10.1016/j.ins.2016.07.007.

[34] L. Osuszek, S. Stanek, and Z. Twardowski, "Leverage big data analytics for dynamic informed decisions with advanced case management," *J. Decis. Syst.*, vol. 25, no. sup1, pp. 436–449, Jun. 2016, doi: 10.1080/12460125.2016.1187401.

[35] A. L'heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. 5, pp. 7776–7797, 2017, doi: 10.1109/access.2017.2696365.

[36] Q. P. He and J. Wang, "Statistical process monitoring as a big data analytics tool for smart manufacturing," *J. Process Control*, vol. 67, pp. 35–43, Jul. 2018, doi: 10.1016/j.jprocont.2017.06.012.

[37] E. P. Xing, Q. Ho, P. Xie, and D. Wei, "Strategies and principles of distributed machine learning on big data," *Engineering*, vol. 2, no. 2, pp. 179–195, Jun. 2016, doi: 10.1016/j.eng.2016.02.008.

[38] F. Portela, L. Lima, and M. F. Santos, "Why big data? Towards a project assessment framework," *Procedia Comput. Sci.*, vol. 98, pp. 604–609, 2016, doi: 10.1016/j.procs.2016.09.094.

[39] B. Franke, J.-F. Plante, R. Roscher, A. Lee, C. Smyth, A. Hatefi, F. Chen, E. Gil, A. Schwing, A. Selvitella, M. M. Hoffman, R. Grosse, D. Hendricks, and N. Reid, "Statistical inference, learning and models in big data," *Int. Stat. Rev.*, vol. 84, no. 3, pp. 371–389, Dec. 2016.

[40] M. Kans and D. Galar, "The impact of maintenance 4.0 and big data analytics within strategic asset management," presented at the 6th Int. Conf. Maintenance Perform. Meas. Manage., Luleå, Sweden, Nov. 28, 2016.

[41] I. Lee, "Big data: Dimensions, evolution, impacts, and challenges," *Bus. Horizons*, vol. 60, no. 3, pp. 293–303, May 2017, doi: 10.1016/j.bushor.2017.01.004.

[42] S. Landset, T. M. Khoshgoftaar, A. N. Richter, and T. Hasanin, "A survey of open source tools for machine learning with big data in the Hadoop ecosystem," *J. Big Data*, vol. 2, no. 1, p. 24, 2015, doi: 10.1186/s40537-015-0032-1.

[43] A. Bechini, F. Marcelloni, and A. Segatori, "A MapReduce solution for associative classification of big data," *Inf. Sci.*, vol. 332, pp. 33–55, Mar. 2016, doi: 10.1016/j.ins.2015.10.041.

[44] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *Nature Sci. Rev.*, vol. 1, no. 2, pp. 293–314, Jun. 2014, doi: 10.1093/nsr/nwt032.

[45] V. López, S. D. Río, J. M. Benítez, and F. Herrera, "Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data," *Fuzzy Sets Syst.*, vol. 258, pp. 5–38, Jan. 2015, doi: 10.1016/j.fss.2014.01.015.

[46] N. Khan, M. Alsaqer, H. Shah, G. Badsha, A. A. Abbasi, and S. Salehian, "The 10 Vs, issues and challenges of big data," presented at the Int. Conf. Big Data Educ. (ICBDE), Honolulu, HI, USA, Mar. 9–11, 2018.

[47] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of big data challenges and analytical methods," *J. Bus. Res.*, vol. 70, pp. 263–286, Jan. 2017, doi: 10.1016/j.jbusres.2016.08.001.

[48] H. Watson, "Tutorial: Big data analytics: Concepts, technologies, and applications," *Commun. Assoc. Inf. Syst.*, vol. 34, no. 6, p. 65, 2014.

[49] P. Raj, A. Raman, D. Nagaraj, and S. Duggirala, "In-database processing and in-memory analytics," in *High-Performance Big-Data Analytics* (Computer Communications and Networks). Cham, Switzerland: Springer, 2015, ch. 8, pp. 207–231.

[50] H. Shu, "Big data analytics: Six techniques," *Geo-Spatial Inf. Sci.*, vol. 19, no. 2, pp. 119–128, Apr. 2016, doi: 10.1080/10095020.2016.1182307.

[51] D. den Hertog and K. Postek, "Bridging the gap between predictive and prescriptive analytics-new optimization methodology needed," Tilburg Univ., Tilburg, The Netherlands, 2016. [Online]. Available: http://www.optimization-online.org/DB_HTML/2016/12/5779.html

[52] R. M. Groves, "Three eras of survey research," *Public Opinion Quart.*, vol. 75, no. 5, pp. 861–871, Dec. 2011, doi: 10.1093/poq/nfr057.

[53] P. Pace, L. Kontokostas, L. Wessels, and G. Sockalingam, "Condition-based maintenance using an integrated track maintenance management system," presented at the Conf. Railway Eng., Perth, WA, Australia, Sep. 7–10, 2008.

[54] D. Jagan, C. Southern, A. Kane, and K. Bladon, "Managing the Australian defined interstate rail network wayside condition monitoring data," presented at the CORE Rail Transp. Vital Economy, Adelaide, SA, Australia, 2014.

[55] K. Bladon, D. Rennison, R. Tracy, and T. Bladon, "Predictive condition monitoring of railway rolling stock," presented at the Conf. Railway Eng., Darwin, NT, Australia, Jun. 20–23, 2004.

[56] P. R. Cohen, N. M. Adams, and M. Berthold, "Advances in intelligent data analysis IX," in *Proc. 9th Int. Symp. (IDA)*, Tucson, AZ, USA, May 2010.

[57] G. Hardie, G. Tew, G. Crew, and M. Courtney, "Track condition assessment and monitoring in heavy haul railroads," presented at the Int. Heavy Haul Assoc. Conf., Calgary, AB, Canada, Jun. 19–22, 2011.

[58] S. Doyle, C. Bastucescu, and T. Vale, "Pantograph condition monitoring system for automated maintenance inspections and prevention of overhead wiring tear downs," presented at the CORE Maintaining Momentum, Melbourne, VIC, Australia, May 16–18, 2016.

[59] K. Laakso, T. Rosqvist, and J. L. Paulsen, "The use of condition monitoring information for maintenance planning and decision-making," Nordisk Kernesikkerhedsforskning, Finland, Tech. Rep. NKS–80, 2002.

[60] M. Bengtsson, "Condition based maintenance on rail vehicles-possibilities for a more effective maintenance strategy," Mälardalen Univ., Västerås, Sweden, Tech. Rep. 1756, 2003. [Online]. Available: http://www.ipr.mdh.se/pdf_publications/1756.pdf

[61] K. Lee, J. Lee, and I. Kim, "A study on strategy of condition based maintenance for Korea metro rolling stocks," presented at the 7th IET Conf. Railway Condition Monit. (RCM), Birmingham, U.K., Sep. 27–28, 2016.

[62] S. Kalay and M. Witte, "Recent advances in automated wagon health monitoring and its effects on safety in North America," presented at the IHHA, Perth, WA, Australia, Jun. 21–24, 2015.

[63] K. Yano, T. Suzuki, K. Okada, W. Wang, and T. Takayanagi, "Data modeling technology in railway operation and maintenance," *Hitachi Rev.*, vol. 67, no. 7, pp. 876–877, 2018.

[64] D. Wobschall, "IEEE 1451—A universal transducer protocol standard," presented at the IEEE Autotestcon, Baltimore, MD, USA, Sep. 17–20, 2007.

[65] R. Nappi, "Integrated maintenance: Analysis and perspective of innovation in railway sector," 2014, *arXiv:1404.7560*. [Online]. Available: https://arxiv.org/abs/1404.7560

[66] M. Bengtsson, "Standardization issues in condition based maintenance," presented at the Condition Monit. Diagnostic Eng. Manage. 16th Int. Congr., Växjö Univ., Sweden, Aug. 27–29, 2003.

[67] S. Burnett and P. Vlok, "A simplified numerical decision-making methodology for physical asset management decisions," *South African J. Ind. Eng.*, vol. 25, no. 1, pp. 162–175, Jul. 2014.

[68] A. D. Mathew, S. Zhang, L. Ma, and D. Hargreaves, "A water utility industry conceptual asset management data warehouse model," presented at the 36th Int. Conf. Comput. Ind. Eng., Howard Hotel, Taipei, Taiwan, 2006.

[69] J. Tutcher, J. Easton, C. Roberts, R. Myall, M. Hargreaves, and C. Tiller, "Ontology-based data management for the GB rail industry feasibility study," Rail Saf. Standards Board, London, U.K., Tech. Rep., 2013.

[70] S. Liu, A. H. B. Duffy, R. I. Whitfield, and I. M. Boyle, "Integration of decision support systems to improve decision support performance," *Knowl. Inf. Syst.*, vol. 22, no. 3, pp. 261–286, Mar. 2010, doi: 10.1007/s10115-009-0192-4.

[71] W. W. Eckerson, "Deploying dashboards and scorecards," TDWI Best Practices Rep., Chicago, IL, USA, Jul. 2006.

[72] A. Makris, K. Tserpes, V. Andronikou, and D. Anagnostopoulos, "A classification of NoSQL data stores based on key design characteristics," *Procedia Comput. Sci.*, vol. 97, pp. 94–103, 2016, doi: 10.1016/j.procs.2016.08.284.

[73] B. E. James and P. O. Asagba "Hybrid database system for big data storage and management," *Int. J. Comput. Sci., Eng. Appl. Int. J. Comput. Sci., Eng. Appl.*, vol. 7, nos. 3–4, pp. 15–27, Aug. 2017, doi: 10.5121/ijcsea.2017.7402.

[74] P. Kecman and R. M. P. Goverde, "Process mining of train describer event data and automatic conflict identification," presented at the Comput. Railways XIII, Southampton, U.K., 2012.

[75] *Specification for Information Management for The Capital/Delivery Phase of Construction Projects Using Building Information Modeling*, Standard PAS 1192:2013, BSI, 2013.

[76] P. Fraga-Lamas, T. M. Fernandez-Carames, and L. Castedo, "Towards the Internet of smart trains: A review on industrial IoT-connected railways," *Sensors*, vol. 17, no. 6, pp. 1457–1501, Jun. 2017, doi: 10.3390/s17061457.

[77] A. Borrmann, M. Hochmuth, M. König, T. Liebich, and D. Singer, "Germany's governmental BIM initiative—Assessing the performance of the BIM pilot projects," presented at the 16th Int. Conf. Comput. Civil Building Eng., Osaka, Japan, Jul. 6–8, 2016.

[78] G. A. Boyes, C. Ellul, and D. Irwin, "Exploring bim for operational integrated asset management—A preliminary study utilising real-world infrastructure data," *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. 4/W5, pp. 49–56, Oct. 2017, doi: 10.5194/isprs-annals-iv-4-w5-49-2017.

[79] A. Bradley, H. Li, R. Lark, and S. Dunn, "BIM for infrastructure: An overall review and constructor perspective," *Autom. Construct.*, vol. 71, pp. 139–152, Nov. 2016, doi: 10.1016/j.autcon.2016.08.019.

[80] A. Sanchez, J. A. Kraatz, K. D. Hampson, and S. Loganathan, "BIM for sustainable whole-of-life transport infrastructure asset management," presented at the Sustainability Public Works Conf., Tweed Heads, Australia, Jul. 7–29, 2014.

[81] G. Hao, S. Ma, J. Lv, and Y. Sui, "A service-oriented data integration architecture and the integrating tree," presented at the 5th Int. Conf. Grid and Cooperat. Comput. (GCC), Hunan, China, Oct. 21–23, 2006.

[82] J. Gao, A. P. Koronios, and C. S. Lin, "A data quality model for asset management in engineering organizations," presented at the Int. Conf. Inf. Quality (ICIQ), Cambridge, MA, USA, Nov. 4–6, 2005.

[83] A. Zoeteman, "Asset maintenance management: State of the art in the European railways," *Int. J. Crit. Infrastruct.*, vol. 2, nos. 2–3, pp. 171–186, 2006.

[84] V. Saquicela, L. M. Vilches-Blázquez, and A. Tello, "Challenges and trends about smart big geospatial data: A position paper," presented at the IEEE Int. Conf. Big Data (BIGDATA), Boston, MA, USA, Dec. 11–14, 2017.

[85] J. Whyte, A. Stasis, and C. Lindkvist, "Managing change in the delivery of complex projects: Configuration management, asset information and big data," *Int. J. Project Manage.*, vol. 34, no. 2, pp. 339–351, Feb. 2016, doi: 10.1016/j.ijproman.2015.02.006.

[86] M. Takikawa, "Innovation in railway maintenance utilizing information and communication technology (smart maintenance initiative)," *Jpn. Railway Transp. Rev.*, no. 67, pp. 22–35, Mar. 2016. [Online]. Available: http://www.ejrcf.or.jp/jrtr/jrtr67/pdf/22-35.pdf

[87] A. Alharthi, V. Krotov, and M. Bowman, "Addressing barriers to big data," *Bus. Horizons*, vol. 60, no. 3, pp. 285–292, May 2017, doi: 10.1016/j.bushor.2017.01.002.

[88] F. P. J. H. M. C. Mimeche and L. A. M. van Dongen, "Application of remote condition monitoring in different rolling stock life cycle phases," *Procedia CIRP*, vol. 11, pp. 135–138, Jan. 2013, doi: 10.1016/j.procir.2013.07.050.

[89] V. Rajaraman, "Big data analytics," *Resonance*, vol. 21, pp. 695–716, Sep. 2016, doi: 10.1007/s12045-016-0376-7.

[90] Q. Chen, M. Hsu, and H. Zeller, "Experience in continuous analytics as a service (CaaaS)," presented at the 14th Int. Conf. Extending Database Technol., Uppsala, Sweden, Mar. 21–24, 2011.

[91] C. Evangelinos, P. F. J. Lermusiaux, J. Xu, P. J. Haley, and C. N. Hill, "Many task computing for real-time uncertainty prediction and data assimilation in the ocean," *IEEE Trans. Parallel Distrib. Syst.*, vol. 22, no. 6, pp. 1012–1024, Jun. 2011, doi: 10.1109/tpds.2011.64.

[92] G. Manco, E. Ritacco, P. Rullo, L. Gallucci, W. Astill, D. Kimber, and M. Antonelli, "Fault detection and explanation through big data analysis on sensor streams," *Expert Syst. Appl.*, vol. 87, pp. 141–156, Nov. 2017.

[93] D. E. Boyle, D. C. Yates, and E. M. Yeatman, "Urban sensor data streams," *IEEE Internet Comput.*, vol. 17, no. 6, pp. 12–20, Nov. 2013.

[94] R. Taverner and V. Lammerse, "Implementing maximo asset management system for Adelaide's public transport rail assets," presented at the CORE Rail Transp. Vital Economy, Adelaide, SA, Australia, May 5–7, 2014.

[95] S. Shekhar, P. Zhang, and Y. Huang, "Spatial data mining," in *Data Mining and Knowledge Discovery Handbook*, O. Z. Mainmon and L. Rokach, Eds. New York, NY, USA: Springer, 2005, pp. 833–851.

[96] Chen, Chiang, and Storey, "Business intelligence and analytics: From big data to big impact," *MIS Quart.*, vol. 36, no. 4, p. 1165, 2012.

[97] E. E. Mangina, S. D. J. McArthur, and J. R. McDonald, "Autonomous agents for distributed problem solving in condition monitoring," in *Intelligent Problem Solving. Methodologies and Approaches* (Lecture Notes in Computer Science), vol. 1821. Berlin, Germany: Springer, 2000, pp. 683–692.

[98] E. Fumeo, L. Oneto, and D. Anguita, "Condition based maintenance in railway transportation systems based on big data streaming analysis," *Procedia Comput. Sci.*, vol. 53, pp. 437–446, 2015, doi: 10.1016/j.procs.2015.07.321.

[99] M. L. Brown and J. F. Kros, "Data mining and the impact of missing data," *Ind. Manage. Data Syst.*, vol. 103, no. 8, pp. 611–621, Nov. 2003.

[100] P. Leonard, "Data and the Internet of Things," presented at the AFR Innov. Summit, Sydney, NSW, Australia, Aug. 16–18, 2016.

[101] P. E. Love, J. Zhou, J. Matthews, M. Lavender, and T. Morse, "Managing rail infrastructure for a digital future: Future-proofing of asset information," *Transp. Res. A, Policy Pract.*, vol. 110, pp. 161–176, Apr. 2018, doi: 10.1016/j.tra.2018.02.014.

[102] A. D'agostino, "Big data in railways common occurrence reporting programme," Eur. Union Agency Railways, Valenciennes, France, Tech. Rep. ERA-PRG-004-TD-003 V 1.0, 2016.

[103] D. Karamshuk, F. Shaw, J. Brownlie, and N. Sastry, "Bridging big data and qualitative methods in the social sciences: A case study of Twitter responses to high profile deaths by suicide," *Online Social Netw. Media*, vol. 1, pp. 33–43, Jun. 2017, doi: 10.1016/j.osnem.2017.01.002.

[104] R. M. Kaplan, D. A. Chambers, and R. E. Glasgow, "Big data and large sample size: A cautionary note on the potential for bias," *Clin. Transl. Sci.*, vol. 7, no. 4, pp. 342–346, Aug. 2014, doi: 10.1111/cts.12178.

[105] D. B. Dunson, "Statistics in the big data era: Failures of the machine," *Statist. Probab. Lett.*, vol. 136, pp. 4–9, May 2018, doi: 10.1016/j.spl.2018.02.028.

[106] J. Thomas and L. Sael, "Overview of integrative analysis methods for heterogeneous data," presented at the Int. Conf. Big Data Smart Comput. (BigComp), Jeju Island, South Korea, Feb. 9–12, 2015.

[107] V. Subramaniyaswamy, V. Vijayakumar, R. Logesh, and V. Indragandhi, "Unstructured data analysis on big data using map reduce," *Procedia Comput. Sci.*, vol. 50, pp. 456–465, 2015, doi: 10.1016/j.procs.2015.04.015.

[108] S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, "Big data: Issues and challenges moving forward," presented at the 46th Hawaii Int. Conf. Syst. Sci., Maui, HI, USA, Jan. 7–10, 2013.

[109] National Research Council, *Reliability Growth: Enhancing Defense System Reliability*. Washington, DC, USA: National Academies Press, 2015.

[110] S. D. Griffith, D. J. Quest, T. S. Brettin, and R. W. Cottingham, "Scenario driven data modeling: A method for integrating diverse sources of data and data streams," *BMC Bioinf.*, vol. 12, no. 10, p. S17, Oct. 2011, doi: 10.1186/1471-2105-12-S10-S17.

[111] F. Gao, S. Bhiri, and Z. Zhou, "User-centric modeling and processing for ubiquitous events using semantic capability models," *Commun. Mobile Comput.*, vol. 1, no. 1, pp. 1–7, 2012.

[112] B. Eine, M. Jurisch, and W. Quint, "Ontology-based big data management," *Systems*, vol. 5, no. 3, p. 45, Jul. 2017, doi: 10.3390/systems5030045.

[113] O. Brazhnik and J. F. Jones, "Anatomy of data integration," *J. Biomed. Informat.*, vol. 40, no. 3, pp. 252–269, 2007. Accessed: May 2019, doi: 10.1016/j.jbi.2006.09.001.

[114] H. Shang and C. Berenguer, "A colored petri net model for railway track maintenance with two-level inspection," presented at the Eur. Saf. Rel. Conf. (ESREL), Wrocław, Poland, Sep. 14–18, 2014.

[115] Q. Li, Z. Zhong, Z. Liang, and Y. Liang, "Rail inspection meets big data: Methods and trends," presented at the 18th Int. Conf. Netw.-Based Inf. Syst. (NBiS), Taipei, Taiwan, Sep. 2–4, 2015.

[116] A. A. da Silva, G. L. Bellotti, and H. H. Chaya, "Utilizing asset data for predictive asset management," *Ind. Comput.*, vol. 58, no. 6, pp. 30–33, Nov./Dec. 2011.

[117] S. G. Alonso, I. de la Torre Díez, J. J. P. C. Rodrigues, S. Hamrioui, and M. López-Coronado, "A systematic review of techniques and sources of big data in the healthcare sector," *J. Med. Syst.*, vol. 41, no. 11, p. 183, 2017, doi: 10.1007/s10916-017-0832-2.

[118] C. Kacfah Emani, N. Cullot, and C. Nicolle, "Understandable big data: A survey," *Comput. Sci. Rev.*, vol. 17, pp. 70–81, Aug. 2015, doi: 10.1016/j.cosrev.2015.05.002.

[119] H. Shuijing, "Big data analytics: Key technologies and challenges," presented at the Int. Conf. Robots Intell. Syst. (ICRIS), Zhangjiajie, China, Aug. 27–28, 2016.

[120] X. Zhou, A. Su, G. Li, W. Gao, C. Lin, S. Zhu, and Z. Zhou, "Big data storage and parallel analysis of grid equipment monitoring system," *Int. J. Performability Eng.*, vol. 14, no. 2, pp. 202–209, 2018, doi: 10.23940/ijpe.18.02.p2.202209.

[121] C. Lakshmi and V. V. Nagendra Kumar, "Survey paper on big data," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 6, no. 8, pp. 368–381, 2016.

[122] N. Honest, "A survey of big data analytics," *Int. J. Inf. Sci. Techn.*, vol. 6, nos. 1–2, pp. 35–43, 2016, doi: 10.5121/ijist.2016.6204.

[123] M. Praveena and M. K. Rao, "Survey on big data analytics in healthcare domain," *Int. J. Eng. Technol.*, vol. 7, nos. 2–7, p. 919, May 2018.

[124] C.-W. Tsai, C.-F. Lai, H.-C. Chao, and A. V. Vasilakos, "Big data analytics: A survey," *J. Big Data*, vol. 2, no. 1, p. 21, 2015, doi: 10.1186/s40537-015-0030-3.

[125] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 2, pp. 652–687, 2014, doi: 10.1109/access.2014.2332453.

[126] A. Bahga and V. K. Madisetti, "Blockchain platform for industrial Internet of Things," *J. Softw. Eng. Appl.*, vol. 09, no. 10, pp. 533–546, 2016, doi: 10.4236/jsea.2016.910036.

[127] B. Hub. *Blockchains and Distributed Ledger Technologies*. Accessed: May 2019. [Online]. Available: https://www.worldbank.org/en/topic/financialsector/brief/blockchain-dlt

[128] A. Gausdal, K. Czachorowski, and M. Solesvik, "Applying blockchain technology: Evidence from norwegian companies," *Sustainability*, vol. 10, no. 6, p. 1985, Jun. 2018, doi: 10.3390/su10061985.

[129] N. Teslya and I. Ryabchikov, "Blockchain-based platforrm architecture for industrial IoT," presented at the 21st Conf. Open Innov. Assoc. (FRUCT), Helsinki, Finland, Nov. 6–10, 2017.

[130] P. Ghuli, U. Priyam Kumar, and R. Shettar, "A review on blockchain application for decentralized decision of ownership of IoT devices," *Adv. Comput. Sci. Technol.*, vol. 10, no. 8, pp. 2449–2456, 2017.

[131] A. Dorri, S. S. Kanhere, and R. Jurdak, "Blockchain in Internet of Things: Challenges and solutions," 2016, *arXiv:1608.05187*. [Online]. Available: https://arxiv.org/abs/1608.05187

[132] A. Ouaddah, A. A. El Kalam, and A. A. Ouahman, "Harnessing the power of blockchain technology to solve IoT security & privacy issues," presented at the ICC, 2017, Cambridge City, U.K., Mar. 22–23, 2017.

[133] M. Kuperberg, D. Kindler, and S. Jeschke, "Are smart contracts and blockchains suitable for decentralized railway control?" 2019, *arXiv:1901.06236*. [Online]. Available: https://arxiv.org/abs/1901.06236

[134] L. X. Downey, F. Bauchot, and J. Roling, "Blockchain for business value: A contract and work flow management to reduce disputes pilot project," *IEEE Eng. Manag. Rev.*, vol. 46, no. 4, pp. 86–93, Dec. 2018.

[135] O. Probert, "Pacific national takes part in experimental blockchain delivery," Rail Express, Jul. 31, 2018.

[136] M. Andoni, V. Robu, D. Flynn, S. Abram, D. Geach, D. Jenkins, P. McCallum, and A. Peacock, "Blockchain technology in the energy sector: A systematic review of challenges and opportunities," *Renew. Sustain. Energy Rev.*, vol. 100, pp. 143–174, Feb. 2019, doi: 10.1016/j.rser.2018.10.014.

[137] A. Alexandre. *National Swiss Railway Operator Completes Pilot of Blockchain ID Management System*. Accessed: May 2019. [Online]. Available: https://dollardestruction.com/18847/

[138] M. Rodríguez. *Swiss Federal Railways Develop a Digital Identity Pilot on Ethereum*. Accessed: May 2019. [Online]. Available: https://www.cryptoworldjournal.com/swiss-federal-railways-develop-a-digital-identity-pilot-on-ethereum/

[139] N. Newman. *How Can Blockchain Open Up New Opportunities for Rail Freight*. Accessed: May 2019. [Online]. Available: https://www.nicnewmanoxford.com/how-can-blockchain-open-up-new-opportunities-for-rail-freight/

[140] BiTA. *BiTA Appoints IEEE-ISTO for Standards*. Accessed: May 2019. [Online]. Available: https://www.freightwaves.com/news/blockchain/bita-to-use-ieee-standards

[141] F. Memoria, "Leading UK public transport provider to launch Blockchain-based loyalty program," Accessed: May 2019. [Online]. Available: https://www.cryptoglobe.com/latest/2019/02/leading-uk-public-transport-provider-to-launch-blockchain-based-loyalty-program/

[142] S. Morant, "How can Blockchain open up new opportunities for rail freight?" *Int. Railway J.*, vol. 1, pp. 1–3, Jun. 2018. [Online]. Available: https://www.railjournal.com/in_depth/how-can-blockchain-open-up-new-opportunities-for-rail-freight

[143] S. Bhatanagar. *Russian Railways Are Eyeing Crypto for Tickets, Blockchain for Cargo*. Accessed: May 2019. [Online]. Available: http://blockchainmagnets.com/russian-railways-are-eyeing-crypto-for-tickets/

[144] M. Popke, "Technology disruptor: Blockchain is a potential game-changer in the rail realm," Rail News, Internet-Digital, Apr. 2018. [Online]. Available: https://www.progressiverailroading.com/internet-digital/article/Technology-disruptor-Blockchain-is-a-potential-game-changer-in-the-rail-realm-industry-adopters-say-54395

[145] O. Emmanuel, "Shenzhen adopts blockchain technology for transportation," Bitcoin, Blockchain & Cryptocurrency News, Mar. 20, 2019. [Online]. Available: https://btcmanager.com/shenzhen-adopts-blockchain-technology-transportation/?q=/shenzhen-adopts-blockchain-technology-transportation/

[146] M. H. Eiza, M. Randles, P. Johnson, N. Shone, J. Pang, and A. Bhih, "Rail Internet of Things: An architectural platform and assured requirements model," presented at the IEEE Int. Conf. Comput. Inf. Technol. Ubiquitous Comput. Commun. Depen., Autonomic Secure Comput., Pervasive Intell. Comput., Liverpool, U.K., Oct. 26–28, 2015.

[147] J. M. Easton, "Blockchains: A distributed data ledger for the rail industry," in *Innovative Applications of Big Data in the Railway Industry*, S. Kohli, A. Kumar, J. M. Easton, and C. Roberts, Eds. Hershey, PA, USA: IGI Global, 2018, pp. 27–39.

[148] A. Labrinidis and H. V. Jagadish, "Challenges and opportunities with big data," in *Proc. VLDB Endowment*, vol. 5, 2012, pp. 2032–2033.

[149] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[150] P. Louridas and C. Ebert, "Embedded analytics and statistics for big data," *IEEE Softw.*, vol. 30, no. 6, pp. 33–39, Nov. 2013.

[151] M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani, "Deep learning for IoT big data and streaming analytics: A survey," 2017, *arXiv:1712.04301*. [Online]. Available: https://arxiv.org/abs/1712.04301

[152] Z. Li, H. Wang, W. Shao, J. Li, and H. Gao, "Repairing data through regular expressions," *Proc. VLDB Endow.*, vol. 9, no. 5, pp. 432–443, Jan. 2016.

[153] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. Sydney, NSW, Australia: Morgan Kaufmann, 2016.

[154] M. U. Ali, S. Ahmad, and J. Ferzund, "Harnessing the potential of machine learning for bioinformatics using big data tools," *Int. J. Comput. Sci. Inf. Secur.*, vol. 14, no. 10, pp. 668–675, Oct. 2016.

[155] Y. Liu, L. Xu, and M. Li, "The parallelization of back propagation neural network in MapReduce and spark," *Int. J. Parallel Program.*, vol. 45, no. 4, pp. 760–779, 2016, doi: 10.1007/s10766-016-0401-1.

[156] B. Liu, E. Blasch, Y. L. Chen, D. Shen, and G. Chen, "Scalable sentiment classification for big data analysis using naive Bayes classifier," presented at the IEEE Int. Conf. Big Data, Silicon Valley, CA, USA, Oct. 6–9, 2013.

[157] J. Khairnar and M. Kinikar, "Sentiment analysis based mining and summarizing using SVM-MapReduce," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 3, pp. 4081–4085, 2014.

[158] N. Attoh-Okine, "Big data challenges in railway engineering," presented at the IEEE Int. Conf. Big Data (Big Data), Washington, DC, USA, Oct. 27–30, 2014.

[159] A. M. Zarembski, "Some examples of big data in railroad engineering," presented at the IEEE Int. Conf. Big Data, Washington, DC, USA, Oct. 27–30, 2014.

[160] S. F. Wamba, A. Gunasekaran, S. Akter, S. J.-F. Ren, R. Dubey, and S. J. Childe, "Big data analytics and firm performance: Effects of dynamic capabilities," *J. Bus. Res.*, vol. 70, pp. 356–365, Jan. 2017, doi: 10.1016/j.jbusres.2016.08.009.

[161] E. Raguseo, "Big data technologies: An empirical investigation on their adoption, benefits and risks for companies," *Int. J. Inf. Manage.*, vol. 38, no. 1, pp. 187–195, Feb. 2018, doi: 10.1016/j.ijinfomgt.2017.07.008.

[162] S. Venkatraman, K. F. S. Kaspi, and R. Venkatraman, "SQL versus NoSQL movement with big data analytics," *Int. J. Inf. Technol. Comput. Sci.*, vol. 8, no. 12, pp. 59–66, Dec. 2016, doi: 10.5815/ijitcs.2016.12.07.

[163] A. Malviya, A. Udhani, and S. Soni, "R-tool: Data analytic framework for big data," presented at the Symp. Colossal Data Anal. Netw. (CDAN), Indore, India, Mar. 8–19, 2016.

[164] I. Durazo-Cardenas, A. Starr, C. J. Turner, A. Tiwari, L. Kirkwood, M. Bevilacqua, A. Tsourdos, E. Shehab, P. Baguley, Y. Xu, and C. Emmanouilidis, "An autonomous system for maintenance scheduling data-rich complex infrastructure: Fusing the railways' condition, planning and cost," *Transp. Res. C, Emerg. Technol.*, vol. 89, pp. 234–253, Apr. 2018, doi: 10.1016/j.trc.2018.02.010.

[165] Deloitte. *Cloud-Based Data Analytics Enables Network Rail to Improve the Performance of its Network*. Accessed: Apr. 4, 2019. [Online]. Available: https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/consultancy/deloitte-uk-network-rail.pdf

[166] L. Oneto, E. Fumeo, G. Clerico, R. Canepa, F. Papa, C. Dambra, N. Mazzino, D. Anguita, "Advanced analytics for train delay prediction systems by including exogenous weather data," presented at the IEEE Int. Conf. Data Sci. Adv. Anal. (DSAA), Montreal, QC, Canada, Oct. 17–19, 2016.

[167] R. Furutani, F. Kudo, and N. Moriwaki, "Utilization of AI in the railway sector: Case study of energy efficiency in railway operations," *Hitachi Rev.*, vol. 65, no. 6, pp. 128–133, 2016.

[168] C. Chen, J. Ma, Y. Susilo, Y. Liu, and M. Wang, "The promises of big data and small data for travel behavior (AKA human mobility) analysis," *Transp. Res. C, Emerg. Technol.*, vol. 68, pp. 285–299, Jul. 2016, doi: 10.1016/j.trc.2016.04.005.

[169] A. Ø. Sørensen, N. O. E. Olsson, M. M. Akhtar, and H. Bull-Berg, "Approaches, technologies and importance of analysis of the number of train travellers," *Urban, Planning Transp. Res.*, vol. 7, no. 1, pp. 1–18, Jan. 2019, doi: 10.1080/21650020.2019.1566022.

[170] A. Tibaut, B. Kaučič, and D. Rebolj, "A standardised approach for sustainable interoperability between public transport passenger information systems," *Comput. Ind.*, vol. 63, no. 8, pp. 788–798, Oct. 2012, doi: 10.1016/j.compind.2012.08.002.

[171] *Manual of Standards and Recommended Practices Section A, Part I*, AAR, Wood Dale, IL, USA, 2019.

[172] G. Palem, "Designing condition-based maintenance management systems for high-speed fleet," *Int. J. Comput. Sci. Bus. Inform.*, vol. 17, no. 1, pp. 28–40, 2017.

[173] I. Patel, A. Rau-Chaplin, and B. Varghese, "A platform for parallel R-based analytics on cloud infrastructure," presented at the 41st Int. Conf. Parallel Process. Workshops, Pittsburgh, PA, USA, Sep. 10–13, 2012.

[174] L. He, Q. Liang, and S. Fang, "Challenges and innovative solutions in urban rail transit network operations and management: China's Guangzhou metro experience," *Urban Rail Transit*, vol. 2, no. 1, pp. 33–45, 2016, doi: 10.1007/s40864-016-0036-y.

[175] D. Bumblauskas, D. Gemmill, A. Igou, and J. Anzengruber, "Smart maintenance decision support systems (SMDSS) based on corporate big data analytics," *Expert Syst. Appl.*, vol. 90, pp. 303–317, Dec. 2017, doi: 10.1016/j.eswa.2017.08.025.

[176] S. K. Kinnunen, A. Yla-Kujala, S. Marttonen-Arola, T. KArri, and D. Baglee, "Internet of things technologies to rationalize the data acquisition in industrial asset management," presented at the World Congr. Comput. Sci., Comput. Eng., Appl. Comput., Las Vegas, NV, USA, Jul. 25–28, 2016.

[177] J. Lee, H. D. Ardakani, S. Yang, and B. Bagheri, "Industrial big data analytics and cyber-physical systems for future maintenance & service innovation," *Procedia CIRP*, vol. 38, pp. 3–7, 2015, doi: 10.1016/j.procir.2015.08.026.

[178] Y. He, H. Tan, W. Luo, S. Feng, and J. Fan, "MR-DBSCAN: A scalable MapReduce-based DBSCAN algorithm for heavily skewed data," *Front. Comput. Sci.*, vol. 8, no. 1, pp. 83–99, Feb. 2014, doi: 10.1007/s11704-013-3158-3.

[179] P. Pääkkönen and D. Pakkala, "Reference architecture and classification of technologies, products and services for big data systems," *Big Data Res.*, vol. 2, no. 4, pp. 166–186, Dec. 2015, doi: 10.1016/j.bdr.2015.01.001.

[180] X. Jin, B. W. Wah, X. Cheng, and Y. Wang, "Significance and challenges of big data research," *Big Data Res.*, vol. 2, no. 2, pp. 59–64, Jun. 2015, doi: 10.1016/j.bdr.2015.01.006.

[181] A. Nunez, J. Hendriks, Z. Li, B. De Schutter, and R. Dollevoet, "Facilitating maintenance decisions on the Dutch railways using big data: The ABA case study," presented at the IEEE Int. Conf. Big Data, Washington, DC, USA, Oct. 27–30, 2014.

[182] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection," *ACM Computing Surv.*, vol. 41, no. 3, pp. 1–58, 2009, doi: 10.1145/1541880.1541882.

[183] Y. Peng, M. Dong, and M. J. Zuo, "Current status of machine prognostics in condition-based maintenance: A review," *Int. J. Adv. Manuf. Technol.*, vol. 50, nos. 1–4, pp. 297–313, Sep. 2010, doi: 10.1007/s00170-009-2482-0.

[184] D. Kwon, M. R. Hodkiewicz, J. Fan, T. Shibutani, and M. G. Pecht, "IoT-based prognostics and systems health management for industrial applications," *IEEE Access*, vol. 4, pp. 3659–3670, 2016, doi: 10.1109/access.2016.2587754.

[185] D. Zhang, C. Cadet, N. Yousfi-Steiner, F. Druart, and C. Bérenguer, "PHM-oriented degradation indicators for batteries and fuel cells," *Fuel Cells*, vol. 17, no. 2, pp. 268–276, Apr. 2017, doi: 10.1002/fuce.201600075.

[186] C. Kulkarni, J. R. Ceyla, G. Biswas, and K. Goebel, *Physics Based Modeling and Prognostics of Electrolytic Capacitors*. Reston, VA, USA: American Institute of Aeronautics and Astronautics, 2012.

[187] E. Zio, "Some challenges and opportunities in reliability engineering," *IEEE Trans. Rel.*, vol. 65, no. 4, pp. 1769–1782, Dec. 2016, doi: 10.1109/tr.2016.2591504.

[188] M. Schwabacher and K. Goebel, "A survey of artificial intelligence for prognostics," presented at the AAAI Fall Symp., 2007.

[189] G. W. Vogl, B. A. Weiss, and M. Helu, "A review of diagnostic and prognostic capabilities and best practices for manufacturing," *J. Intell. Manuf.*, vol. 30, no. 1, pp. 79–95, Jan. 2019, doi: 10.1007/s10845-016-1228-8.

[190] D. An, J. H. Choi, and N. H. Kim, "Options for prognostics methods: A review of data-driven and physics-based prognostics," presented at the 54th AIAA/ASME/ASCE/AHS/ASC Struct., Struct. Dyn., Mater. Conf., Boston, MA, USA, Apr. 8–11, 2013.

[191] P. Dersin, "Keynote speech: PHM in railways: Big data or smart data?" presented at the Prognostics Syst. Health Manage. Conf. (PHM-Harbin), Harbin, China, Jul. 9–12, 2017.

[192] O. O. Aremu, A. S. Palau, A. K. Parlikad, D. Hyland-Wood, and P. R. Mcaree, "Structuring data for intelligent predictive maintenance in asset management," presented at the 16th IFAC Symp. Inf. Control Problems Manuf., Bergamo, Italy, Jun. 11–13, 2018.

[193] A. N. Gorban and A. Y. Zinovyev, "Principal graphs and manifolds," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques*. Hershey, PA, USA: IGI Global, 2010, pp. 28–59.

[194] M. Vignesh, "Predictive analytics in data mining with big data: A literature survey," *Int. J. Res. Eng. Appl. Sci.*, vol. 6, no. 11, pp. 10–22, 2016.

[195] W. Sammouri, E. Come, L. Oukhellou, P. Aknin, and C. E. Fonlladosa, "Floating train data systems for preventive maintenance: A data mining approach," presented at the Int. Conf. Ind. Eng. Syst. Manage. (IESM), Rabat, Morocco, Oct. 28–30, 2013.

[196] B. Rapolu. *Bringing Artificial Intelligence to the Rail Industry*. Data Economy. Accessed: Jan. 2019. [Online]. Available: http://www.techx365.com/document.asp?doc_id=731414

[197] F. Cannarile, M. Compare, F. Di Maio, and E. Zio, "A clustering approach for mining reliability big data for asset management," *Proc. Inst. Mech. Eng., O, J. Risk Rel.*, vol. 232, no. 2, pp. 140–150, Apr. 2018, doi: 10.1177/1748006x17716344.

[198] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, and A. Bouras, "A survey of clustering algorithms for big data: taxonomy and empirical analysis," *IEEE Trans. Emerg. Topics Comput.*, vol. 2, no. 3, pp. 267–279, Sep. 2014.

[199] M. Menéndez, C. Martínez, G. Sanz, and J. M. Benitez, "Development of a smart framework based on knowledge to support infrastructure maintenance decisions in railway corridors," *Transp. Res. Procedia*, vol. 14, pp. 1987–1995, 2016, doi: 10.1016/j.trpro.2016.05.166.

[200] R. Veit-Egerer, M. Widmann, G. Achs, J. Rajek, A. Weninger-Wycudil, C. Stefan, and J. Litzka, "EINSTEIN risk-based decision model for the determination of optimized maintenance intervention schedules for infrastructure," presented at the IABSE Conf.-Eng. Past, Meet Needs Future, Copenhagen, Denmark, Jun. 25-27, 2018.

• • •