Big Data Platform for Health and Safety Accident Prediction

Abstract

Purpose
This paper highlights the use of the Big Data technologies for health and safety risks analytics in the power infrastructure domain with large datasets of health and safety risks, which are usually sparse and noisy.

Design/methodology/approach
The study focuses on using Big Data frameworks for designing a robust architecture for handling and analysing (exploratory and predictive analytics) accidents in power infrastructure. The designed architecture is based on a well coherent health risk analytics lifecycle. A prototype of the architecture interfaced various technology artefacts was implemented in the Java language to predict the likelihoods of health hazards occurrence. A preliminary evaluation of the proposed architecture was carried out with a subset of an objective data, obtained from a leading UK power infrastructure company offering a broad range of power infrastructure services.

Findings
The proposed architecture was able to identify relevant variables and improve preliminary prediction accuracies and explanatory capacities. It has also enabled conclusions to be drawn regarding the causes of health risks. The results represent a significant improvement in terms of managing information on construction accidents, particularly in power infrastructure domain.

Originality/value
This study carries out a comprehensive literature review to advance the health and safety risk management in construction. It also highlights the inability of the conventional technologies in handling unstructured and incomplete dataset for real-time analytics processing. The study proposes a technique in Big Data technology for finding complex patterns and establishing the statistical cohesion of hidden patterns for optimal future decision-making.

Keywords: Big Data analytics, Machine learning, Health hazards analytics, Health and Safety

1. Introduction

Occupational accidents are things of worry in modern society, especially in construction sites where a high number of construction activities take place (Zhu et al. 2016). The power infrastructure delivery sector, for instance, has high incidences of nonfatal occupational injuries as workers using heavy machinery are confronted with health risks such as radiation, dust, temperature extremes, and chemicals amongst others (McDermott & Hayes 2016). According to the UK Health and Safety Executive, a total cost of £4.8 billion was expended in 2014/15 for workplace injury (HSE 2016). Similarly, repair costs of buried communication lines are significant when disrupted during excavations (McDermott & Hayes 2016).

Several machine-learning techniques have been used for health and safety risks prediction in construction. For instance, decision trees (Cheng et al. 2011), the generalised linear model (Esmaeili et al. 2015), and fuzzy-neural method (Debnath et al. 2016) have all been used to analyse incident data to reduce accident rates. Techniques such as the Bayesian network was used to quantify occupational accident rates (Papazoglou et al. 2015), and fuzzy Bayesian networks for damaged equipment analysis (Zhang et al. 2016). Others are the bow tie representation for occupational risks assessment (Jacinto & Silva 2010), and Poisson models for occupational injury impacts modelling (Yorio et al. 2014).

However, a significant problem associated with these existing models is their limited ability to process large-scale raw data since considerable effort is needed to transform them into an appropriate internal form to achieve high prediction accuracy (Esmaeili et al. 2015). Construction accident data are typically large, heterogeneous and dynamic (Fenrick & Getachew 2012), nonlinear relationships among accident causation variables (Gholizadeh & Esmaeili 2016), imbalance data, and appreciable missing values (Bohle et al. 2015). Besides, these techniques simplify some key factors and pay little attention to analysing relationships between a safety phenomenon and the safety data (Landset et al. 2015).

Based on the preceding, the Big Data technology due to its parallel processing feature and ability to efficiently handle high dimensional, noisy data with nonlinear relationships, will be beneficial for health and safety risks analytics in the power infrastructure domain. Also, the technology will uncover potential factors contributing to accidents in this domain. The objectives of this study are, therefore, to chart lifecycle stages of occupational hazards analytics and develop a Big Data architecture for managing health and safety risks.

1.1. Big data for health and safety risk analytics

Big Data is an emerging technology, which refers to data sets that are many orders of magnitude larger than the standard files transmitted via the Internet (Suthakar et al. 2016). There is tremendous interest in utilising information in Big Data for various analytics (exploratory, descriptive, predictive and prescriptive) to determine future occurrences. Most importantly, Big Data technologies support analytical techniques for occupational health and safety risk analytics; thus, a system being proposed in this study, named Big Data Accident Prediction Platform (B-DAPP) offers unparalleled opportunities to minimise occupational hazards at construction sites. The seamless combination of the following technologies: Big Data, Health and Safety, and Machine-learning is an outcome of a robust health and safety risk management tool to help stakeholders in making appropriate decisions to minimise occupational accidents in Power Infrastructure projects.

Health and safety risk analytics is dependent on a high-performance computation and large-scale data storage requiring a large number of diverse datasets of health and safety risks, and machine-learning knowledge to successfully provide the needed analytical responsibilities. The datasets, however, are unreliable, unstructured, incomplete, and imbalanced (Chen et al. 2017). Hence, storing the datasets using conventional technologies and subjecting them to real-time processing for advanced analytics is highly challenging. A robust technique for finding complex patterns and establishing the statistical cohesion of hidden patterns in such datasets for optimal future decision-making is inevitable. Thus, motivating the use of Big Data technologies to address these challenges.

1.2. Research Justification

There exists an apparent technological gap in existing literature regarding health and safety risk management. In particular, there is limited research on the application of Big Data techniques for managing health and safety risk in Power Infrastructure. The development of a robust B-DAPP for health and safety risk is the objective of the ongoing R&D effort. The proposed tool will provide stakeholders with well-informed and data-driven insights to reduce accidents and incidents at construction sites. Therefore, a Big Data architecture is proposed for managing health and safety risks. Also, a presentation of components and relevant technologies of the proposed architecture necessary for storing and analysing health and safety risk datasets for real-time exploration and prediction is made. The term 'Architecture' as used in this text refers to high-level structures of a software system. Similarly in the context of this study, 'Accident' is an unplanned, unpremeditated event caused by unsafe acts or conditions resulting in injury while 'Incident' is an event causing actual damage to property (including plant or equipment) or other loss with potential to cause injury.

The remainder of the paper is structured as follows: Section 2 discusses on the research methodology, Big Data analytics, and Big Data ecosystem. Section 3 deliberates on the health hazards analytics lifecycle. Section 4 presents the proposed Big Data architecture for health and safety risk management while Section 5 presents the preliminary outcomes. Conclusions and future work are given in Section 6.

2. Methodology

In this section, a discussion on the methodology employed in this research is made. Foremost, a comprehensive literature review is performed to advance the health and safety risk management with respect to the system architecture and system analytics lifecycle. Then the proposed architecture and occupational hazard analytics lifecycle are validated in a preliminary analysis of the health and safety risk related data. To be able to offer a holistic Big Data architecture and occupational hazard analytics lifecycle, a careful review of existing literature on health and safety risk prediction models, Big Data, and machine learning have been carried out. In this regard, online databases such as Journal of Big Data, Big Data Research, Safety Science, Journal of construction engineering, Journal of Decision Systems, Journal of Safety Research, Journal of Construction Engineering and Management, Reliability Engineering and System Safety are searched for research articles between 2005 and 2017. Recent reviews of research and books on Big Data Analytics are also considered (Camann et al. 2011; Gandomi & Haider 2015; Guo et al. 2016).

Examples of search words used include: “managing health and safety risks”, “design strategies for occupational hazards in construction”, “Prediction models for occupational health risks”, “Big Data in Construction”, “Big Data based Application Architecture”, and “Big Data Analytics”. In general, 94 publications were selected even though literature search was in-exhaustive as a result of a vast amount of published articles. However, it is believed that the literature search has captured a representative balanced sample of the related research. Studies in which Big Data is used to develop enterprise applications were included, and those focusing on road traffic related hazards and health hazards in domains not related to construction (e.g., mining and fishing) were excluded. This elimination procedure further reduced the selected articles to 66. These articles are furthermore scrutinised for relevancy by reading abstracts, introductions, and conclusions. Ultimately, the articles are reduced to 50. Table 1 depicts how these selected articles are relevant and contributing to the development of the proposed architecture, which is essentially based on three concepts, namely Big Data, Health and safety risk, and Machine learning. In this study, we introduce the proposed B-DAPP architecture and the occupational hazards analytics lifecycle stages for managing incidents and accidents.

2.1. Big data analytics

Big data consists of large and complex datasets often difficult to manipulate using the conventional processing methods. It has six defining attributes (Gandomi & Haider 2015), which are volume, variety, velocity, veracity, variability and complexity, and value. The term 'volume' represents the magnitude of the data (measured in terabytes, petabytes and beyond). 'Variety' is the structural heterogeneity in a dataset while the 'Velocity' is the rate of generating data. 'Veracity' is the unreliability inherent in data sources while 'Variability' (complexity) represents the variation in data flow rates. Finally, 'Value' measures the information extracted from historical incident datasets for optimal control decision to mitigate incidents and reduce their impact.

These attributes are evident in a typical power infrastructure health and safety dataset, which is typically large, heterogeneous and dynamic (Fenrick & Getachew 2012). Big data analytics is a concept that inspects, cleans, transforms, and models the big data to discover useful information to support decision-making(Power 2014). The Big data analytics have rich intellectual traditions and borrow from a wide variety of related fields such as statistics, data mining, business analytics, knowledge discovery from data (KDD), and data science. The forms of big data analytics are descriptive (Schryver et al. 2012), predictive (Esmaeili et al. 2015), prescriptive (Delen & Demirkan 2013) and causal (Schryver et al. 2012).

2.2. Big Data for safety risk management

A wide variety of technologies and heterogeneous architectures are available to implement Big data applications. Since this paper intends to develop a robust Big Data architecture for health hazards analytics, A brief discussion of tools and Big Data platforms to facilitate the creation of a compact architecture and increase the understanding of the concept is made. Primarily, focusing on the Hadoop ecosystem, a system designed for solving Big Data problems.

Table 1:Summary of articles reviewed

#	Article	Contribution to health and safety risk analytics architecture		
		Health and safety risk	Machine learning	Big data
1	Liu & Tsai (2012)	×	×	
2	Zhou et al. (2015)	×		
3	García-Herrero et al. (2012)	×	×	
4	Groves et al. (2007)	×		
5	Li et al. (2016)	×	×	
6	Soltanzadeh et al. (2016)		×	
7	Power (2014)			×
8	Yi et al. (2016)	×	×	
9	Cheng et al. (2011)	×	×	
10	Silva et al. (2016)	×		
11	Raviv et al. (2017)	×		
12	Liao & Perng (2008)		×	
13	Li & Bai (2008)			
14	Törner & Pousette (2009)	×		
15	Pinto et al. (2011)	×		
16	Tixier et al. (2016)	×	×	
17	Hallowell & Gambatese (2009)	×		
18	Pääkkönen & Pakkala (2015)			×
19	Venturini et al. (2017)			×
20	Suthakar et al. (2016)			×
21	Najafabadi et al. (2015)		×	×
22	Landset et al. (2015)			×
23	Tsai et al. (2015)			×
24	Zang et al. (2014)		×	×
25	Jin et al. (2015)			×
26	Rahman & Esmailpour (2016)			×
27	Al-Jarrah et al. (2015)			×
28	Zhang et al. (2016)	×	×	
29	Love & Teo (2017)	×	×	
30	Rivas et al. (2011)	×	×	
31	Guo et al. (2016)	×		×
32	Zou et al. (2007)	×		
33	Wu et al. (2010)	×		
34	Carbonari et al. (2011)	×		
35	Weng et al. (2013)	×	×	
36	Naderpour et al. (2016)	×	×	
37	Yoon et al. (2016)	×		
38	Favarò & Saleh (2016)	×	×	
39	Jocelyn et al. (2017)	×	×	
40	Papazoglou et al. (2017)	×	×	
41	Papazoglou et al. (2015)	×	×	
42	Fragiadakis et al. (2014)	×	×	
43	Ciarapica & Giacchetta (2009)	×	×	
44	Khakzad et al. (2015)	×	×	
45	Galizzi & Tempesti (2015)	×		
46	Gürcanli & Müngena (2009)	×	×	
47	Debnath et al. (2016)	×	×	
48	Nanda et al. (2016)	×	×	
49	Zeng et al. (2008)	×		
50	Guo et al. (2016)	×	×	

2.2.1. Hadoop ecosystem

Hadoop is a MapReduce processing engine with distributed file systems (White 2012). However, it has evolved into a vast web of projects (Hadoop ecosystem) related to every step of a Big Data workflow. The concept now is being referred to as the Hadoop ecosystem, which encompasses related projects and products developed to either complement or replace original components. Further examination of the two concepts for ease of understanding follows.

The Hadoop project consists of four modules (White 2012):

a) Hadoop distributed file system (HDFS) is a fault-tolerant file system designed to store massive data across multiple nodes of commodity hardware. It has a master-slave architecture that is made up of data nodes and name nodes. Data nodes store blocks of the data, retrieve data on request and report to the name node with inventory. The name node keeps records of the inventory and directs traffic to the data nodes upon client requests.
b) MapReduce Data processing engine. A MapReduce job consists of a map phase and a reduce phase. A map phase organises raw data into key/value pairs, while the reduce phase processes data in parallel.
c) YARN (“Yet Another Resource Negotiator”) is a resource manager of the Hadoop project introduced to address the limitations of the MapReduce. It separates infrastructures from program representations.
d) Common is a set of utilities required by the other Hadoop modules. These include compression codecs, I/O utilities, error detection, proxy users authorisation, authentication, and data confidentiality.

The Hadoop ecosystem consists of several tools built on top of the core Hadoop modules described above to support researchers and practitioners in all aspects of data analyses. The ecosystem structure has the following layers: storage, processing, and management. Figure 1. depicts examples of standard tools used in Big Data applications. The right selection requires in-depth knowledge of critical features of these platforms and the characteristics of the problem to be solved. In the case of health hazards analytics the platforms to adapt as a result of increased workload, outweighs the rest of the selection criteria. In the real sense, Hadoop ecosystem is made up of well over 100 projects, and readers are referred to (White 2012) or the Hadoop website for more information.

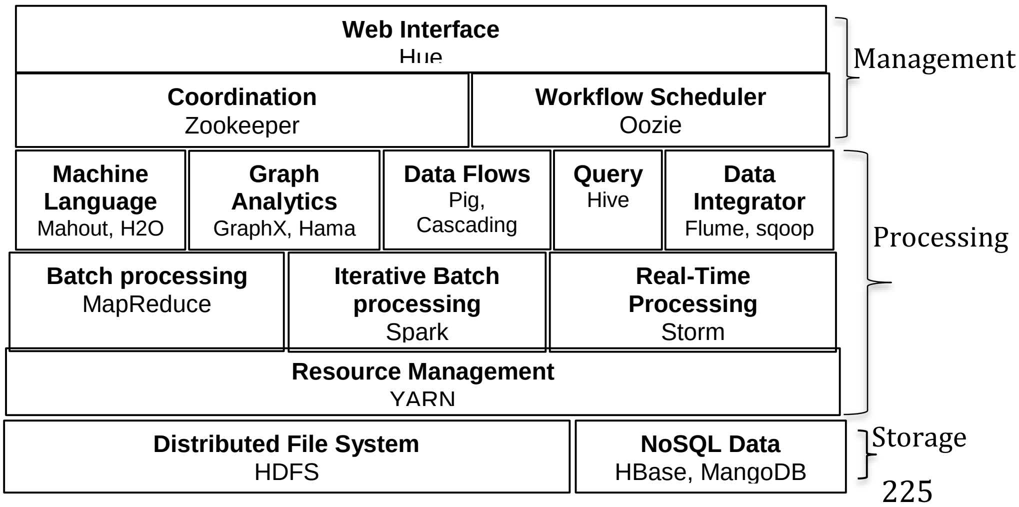

Figure 1. Hadoop ecosystem

a) Storage layer- This layer includes the HDFS described earlier and Non-relational databases (NoSQL). Non-relational databases are nested, semi-structured, and unstructured data that support machine-learning tasks. These databases use the following data representation models: Key-value stores (i.e. Redis), Document stores (i.e. MongoDB), Column-oriented Data (i.e. HBase), and Graph-based models (Neo4J). The graph model is regarded as more flexible than other models.

b) Processing layer - This layer carries out the actual analysis using YARN, which allows one or more processing engines to run on a Hadoop cluster. Additionally, a layer has frameworks for data transfer, aggregation, and interaction. Examples include Flume, Sqoop, Hive, Spark, and Pig. Flume collects, aggregates, and moves data log in HDFS. Kafka is a distributed messaging system on HDFS, and Sqoop transports bulk data between the HDFS and relational databases. Hive is a query engine for querying data stored in the HDFS and NoSQL databases. Spark supports iterative computation, and it improves on speed and resource issues by utilising in-memory computation. Finally, Pig offers an execution framework and data flow language to support user-defined functions written in Python, Java, JavaScript, etc. Machine learning frameworks are used to perform machine-learning tasks in Hadoop. Examples are Mahout, H2O, etc. Mahout is one of the more well-known machine-learning tools. It is known for having a wide selection of robust algorithms, but with inefficient runtimes due to the slow MapReduce engine. H2O provides a parallel processing engine, analytics, math, and machine learning libraries for data pre-processing and evaluation.

c) Management layer - This layer has tools for user interaction and high-level organisation. It carries out functions such as scheduling, monitoring, coordination, amongst others. Examples of tools available in this layer are Oozie, Zookeeper, and Hue. Oozie is a workflow scheduler, which manages jobs for many of the tools in the processing layer. Zookeeper provides tools to handle the coordination of data and protocols and can handle partial network failures. It includes APIs for Java and C and also has bindings for Python and REST clients. Hue is a web interface for Hadoop projects with support for widely used Hadoop ecosystem components.

3. **Proposed health hazards analytics stages**

Developing a health hazards analytics tool for health and safety risk data is a challenging task since the data are typically dynamic (Fenrick & Getachew 2012), and unbalanced with significant missing values (Bohle et al. 2015). Besides, the traditional accident-causing modelling may ignore or simplify some key factors as well as assume the same format for the input data. Thus, an efficient methodology to address these challenges requires a well-articulated process to break the task into smaller manageable stages to ensure adequate preparation of various analytical approaches. In this section, a discussion on the lifecycle of the proposed Big Data architecture for the health hazards analytics tool is made. The lifecycle has six stages (see Figure 2) that are iteratively executed to suit the requirements of the proposed tool.

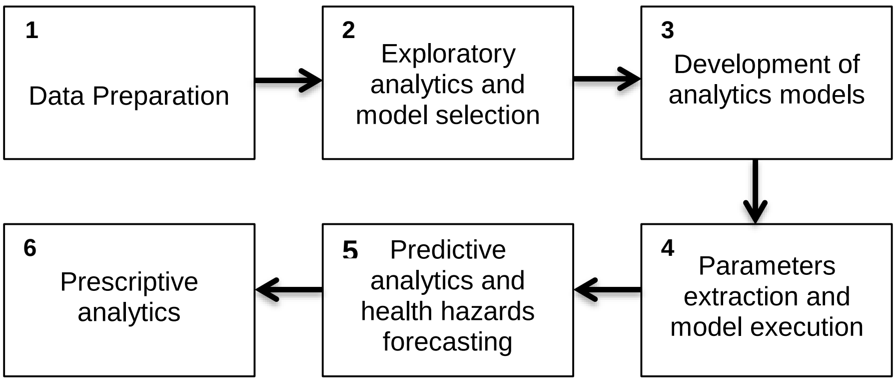

Figure 2. Stages of the health hazards analytics

3.1. *Data preparation*

Data preparation is a procedure to detect and repair errors in the dataset. For the health hazards analytics, sufficient data quality is necessary for high-quality analytics. Thus, data from various sources are obtained, transformed, and loaded into the centralised data store. Before this, outliers are inadvertently eliminated using techniques such as mean/mode imputation, transformation, and binning. Missing data issues should also be solved using appropriate technology. The k-nearest neighbour (kNN) imputation and mean/mode imputation are few examples to eliminate the missing data problem. Apparently, machine-learning techniques can also be applied to quickly filter through hundreds of thousands of narratives (texts) to accurately and consistently retrieve and track high-magnitude, high-risk and emerging causes of injury. The retrieved information is then utilised to guide the development of interventions to prevent future incidents.

In the event of having large data, methods for parallel data movement may be required, which may necessitate using the appropriate component of the Hadoop ecosystem. Data is often analysed to get familiar with the health and safety risk as it pertains to the construction domain. For the sake of preliminary analysis presented here, the health and safety data are provided as .csv files that are stacked on the Hadoop cluster. The respective files are queried to retrieve specific details on health and safety hazards such as injured body parts, loss type injury, and damaged equipment amongst others. For this purpose, tools like Apache Flume are of immense relevance to capture current versions of datasets.

3.2. *Exploratory analytics and model selection*

For the health hazards management, the analysis starts with exploratory analytics and then to the predictive analytics. For each activity in the proposed tool, a clear objective is essential for the right selection of analytical approaches (prescription, exploratory, predictive, etc.) to execute. The data exploration of health and safety records is performed to understand the relationship between different explanatory variables. This exploratory data analysis informs the selection of relevant variables to build a robust health hazards prediction model. In this study, a visualisation technique is used for exploratory data analysis. At this phase, the purpose of the analysis is to capture essential predictors and independent variables while eliminating the least relevant ones for building the model. Variable selection methods include All Possible regression, Stepwise Forward regression, Best Subset regression, etc. These selection methods are often iterative and require a series of steps to identify the most useful variables for the given model. Tools such as R Studio could be exploited to build these models.

3.3. *Development of analytics models*

In this stage, analytics models are created for health and safety risk prediction using robust Big Data analytics techniques. The data are divided first into the training and test sets. The analytics models are then fitted to the training data and evaluated using the test data. Models with optimal accuracy or higher predictive power are selected. Often, this step may involve dealing with certain optimisation issues such as multicollinearity. The best model is selected and deployed to predict health and safety risk from a large volume of data. Many times the production environment may require adjusting and redeploying models to support more practical situations (Camann et al. 2011).

3.4. *Parameters extraction and model execution*

Here, vital parameters are extracted to execute the predictive models. Parameters such as task, equipment type, project complexity, etc. are extracted and the relationship between a safety phenomenon and safety data explored to uncover potential factors that contribute to the likelihood of accidents. These relationships bring those potential trends into the focus that could be utilised to predict the health and safety risk of an infrastructure project under execution. A series of transformations are applied to make the application user-friendly. Specifically, by standardising contents using the ifcOWL ontology (Chaudhuri & Dayal 1997). The data are then stored as graph-annotated formats to support broader computations required from the proposed tool.

3.5. *Predictive analytics and health hazards forecasting*

Health hazards prediction provides the necessary foundation for understanding causes and types of health and safety risk arising from a construction project in execution. Thus, this stage employs predictive models generated through the big data analytics approaches to analyse health and safety risk database and give notice of a possible health hazard occurrence. Indeed, the critical thing about this evaluation is the accuracy of the health and safety risk prediction models that are employed.

The traditional accident-causing modelling has the following limitations: may ignore or simplify some key factors, uses qualitative analysis, and focuses on causality analysis and explanations of an accident (Landset et al. 2015). Hence, these methods pay little attention to the analysis of relationships between a safety phenomenon and safety data. They are also unable to uncover potential factors that contribute to the likelihood of accidents, such as frequency, relevance, locale, and timeliness.

The development of robust health hazards prediction models is the ultimate goal of this lifecycle, and using the prediction models, comprehensive accident and equipment damage forecasts are generated to organisations implement strategies and techniques to improve the safety of their construction sites.

3.6. *Prescriptive analytics*

This phase optimises various safety strategies based on myriad factors (the interaction between deficiencies in work teams, workplace, equipment and materials, weather, etc.) to recommend the best course of action for a given situation. It uses simulation and optimisation to offer the best strategy to employ for different health and safety risks. Consequently, a large number of alternative optimisation plans are generated and converted into user-friendly prescriptions for stakeholders to aid in data-driven decision-making for minimising accidents.

3.7. Analysis and preliminary results

The proposed architecture is further assured and validated with the objective data, obtained from a leading UK construction company, offering a broad range of power infrastructure services, including building and refurbishing overhead lines, substations, underground cabling, fibre optics, etc. The company uses a relational database to store the health and safety risks data, which consist of a large number of power infrastructure projects constructed over 13 years (2004 to 2016) across five UK regions. Each time an incident (or hazard) occurs, a digital record is created in the database. Details of some of the relevant explanatory variables in the database are shown in Table 2.

A subset of 5000 randomly selected projects from 20000 projects in total was used for a preliminary evaluation and analysis presented in this study. The criteria for this selection include project types (i.e. overhead lines, cabling, and substations) and construction mode (i.e. new built, refurbishment). The distribution of data across the UK regions will help to generate advanced visualisations such as geographic heat map. Data from the relational database is accessed via the front-end application and exported to comma-separated files (.csv). Plainly, occupational hazards data of 5000 projects will not be labelled as Big Data to justify the use of data-intensive platforms for its analysis. However, the approach adopted in this study can be used to analyse larger sets of health and safety risk data. Exploratory data analytics is applied to understand the underlying trends in the data using geographical and chronological dimensions. Thus, a variety of visualisations such as bar plot, box plot, and geographic heat map are used for data investigation.

Table 2:Explanatory variables in the database

Variable	Meaning
Incident reference	Identification of a given incident
Project type	The specific project (overhead line, cabling, offshore, etc.)
Project contract	The nature construction project being built (i.e. new built, maintenance, refurbishment)
Region	The specific region of the construction site (Scotland, North, South East, Midlands, etc.)
Sub region	The sub-region where the site is located i.e. Yorkshire East, Midlands North, East England, Tyrone, etc.
City	UK cities where the construction site is located.
Location	A specific area or location of the site
Client	An organisation using the services of the power infrastructure company.
Equipment type	Specifies the machinery (e.g. drill, hammer, haulage, etc.) used for a task.
Age	The age of the victim at the time of the accident.
Year	The year when the health hazard occurred.
Season	External factor such as the weather
Month	The month (1-12) when the incident occurred
Time	The period incident happened (0-6- early morning, 6-12- morning, 12-18 afternoon, 18-23 -evening).
Day of the week	Day (1-31) when the accident occurred.
Weekday	The weekday i.e. Monday, Tuesday, Wednesday, etc.
Task	Specific task or operation to be carried out (excavating, lifting, cutting, etc.
Accident type	The type of accident, for instance, fall, trip, struck by, Inhalation, Caught in/between, etc.
Injury type	The physical consequence for a victim, i.e. first aid, fatal, no injury, etc.
Severity cost	Financial cost incurred as a result of the accident
Hazard type	Forms of health hazards, for example, illness, injury, loss or damage, etc.,
Injured body part	The part of the body that is injured, i.e. Fingers, shoulder, head, back, etc.
Total cost	The cost of the project
Equipment	Part of the equipment damaged during operation.

4. **Proposed big data architecture for health hazards analytics**

This section discusses the proposed Big Data architecture for health hazards analytics (see Figure 3). Components of the architecture are the Application layer, Analytics and Functional Model layer, Semantic layer, and Data Storage layer which are discussed in subsequent subsections.

4.1. *Data storage*

This layer is the data source (finance and health and safety risks), which are needed for efficient functioning of B-DAPP and analytics models (predictive and prescriptive) development. The finance data includes information such as project cost, margin, labour cost, material cost, etc. The health and safety data contains historical occupational risk data while multimedia data consists of images and videos depicting accidents scenes.

As a result of the diverse nature of data to be stored in this layer, a NoSQL database (i.e. MongoDB, Neo4J, Oracle NoSQL) is used for the implementation due to its robust storage mechanisms and efficient handling of structured, semi-structured and unstructured data (Leavitt 2010).

4.2. *Semantic layer*

This layer provides the data exchange formatting and data provisioning to the application layer. The data exchange formatting allows the sharing of a common data format in the entire system. The DDAXML is used to share data among different modules in the system since it is an industrially supported schema for sharing information. The data provisioning functionality provides the application layer of the architecture with seamless access to databases through the Representation State Transfer (REST) web service. This database access approach is considered the most appropriate due to the different nature of health and safety risk data.

4.3. *Analytics and functional model layer*

The significance of health and safety risk management tool lies in its ability to analyse and promptly act upon complex and high volume data. The layer has one functional model (Health and Safety visualisation) and three analytics models (discussed earlier), which are exploratory analytics, predictive analytics, and prescriptive analytics. As discussed earlier, predicting and managing health hazards is data-driven and highly intensive. Consequently, the Apache Spark engine was chosen over the MapReduce to build the analytics (predictive and prescriptive), due to its efficient in-memory storage and computation (Ryza et al. 2015). The analytical pipelines for health hazards management are actualised using SparkR, H2O, and GraphX.

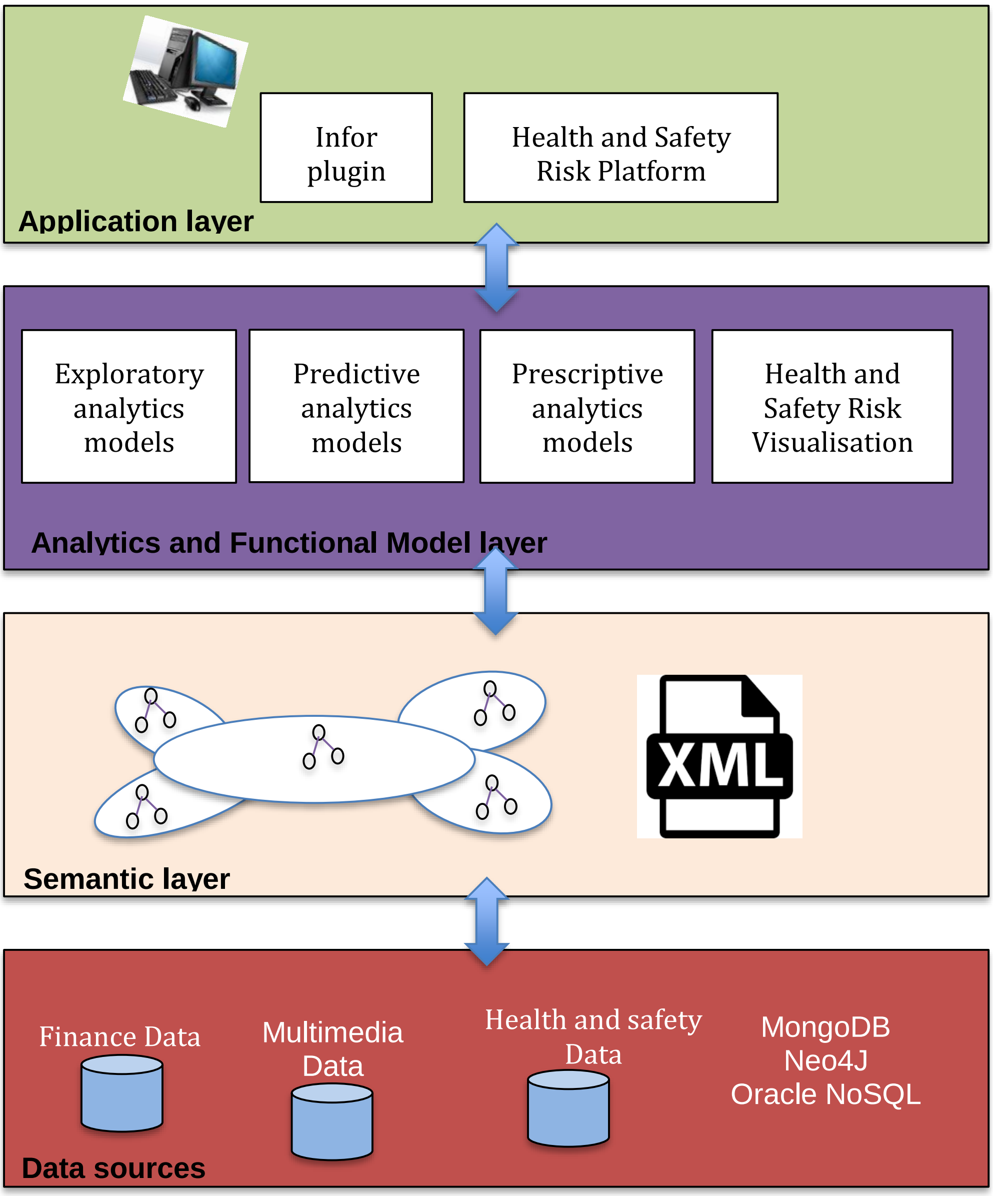

Figure 3. B-DAPP architecture

During each iteration in the analytical pipeline, different predictive models for health hazards are explored and optimised for optimum accuracy.

The H2O framework is selected because of its rich graphical user interface (GUI) and numerous tools for developing deep neural networks models. Additionally, it offers a comprehensive open source machine learning toolkit that is suitable for big data (Landset et al. 2015). It also provides tools for varied machine learning tasks, optimisation tools, data preprocessing and deep neural networks. Additionally, it offers coherent integration with Java, Python, R and R Studio, as well as Sparkling Water for integration with Spark and MLlib. Prior to or during an infrastructure project construction, health hazards are predicted and disseminated to stakeholders to help in mitigating the impact of hazards.

4.4. *Application layer*

This layer is built by exploiting its powerful API programs. The end users of the tool are stakeholders (Engineers, Health and Safety officers, Site managers, Top level directors, etc.). The explanatory variables for infrastructure projects under B-DAPP are captured through appropriate the user interface and loaded to the HDFS and then to the Triplestore. Spark Streaming triggers the analytics pipeline to predict health hazards and suggests actionable insights to minimise health hazards. The predictions and prescriptions are communicated as the Predictive Model Markup Language (PMML). Stakeholders are provided with information to manage health hazards effectively.

5. **Results and discussions**

The prototype of the B-DAPP architecture is implemented by considering and interfacing the various technology artefacts. A sample screenshot produced by simulating the B-DAPP system is as shown in Figure 4, where the system predicts probable and number of injuries to body parts after the specification of input parameters (i.e. “Project type”, “Region”, “Operation”, etc). It informs stakeholders of probable risks and allowing them adequate attention to risk factors when managing occupational hazards to achieve a safer environment.

The B-DAPP architecture is evaluated using exploratory data analysis and some preliminary results are provided. The purpose of this evaluation is to test the appropriateness of the B-DAPP architectural components and present some of these initial results. Interestingly, results obtained support findings in the literature. The future goal is to conduct a more rigorous evaluation through predictive analytics, by exploiting the preliminary analysis results presented in this paper.

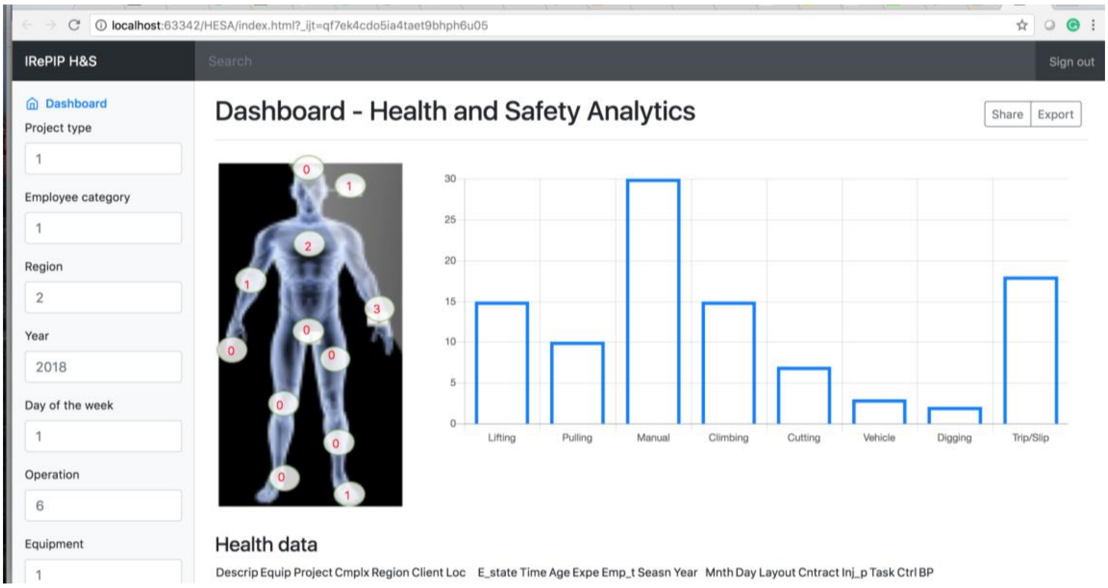

Figure 4. Screenshot of sub-module

5.1. *Injury distribution by body parts*

Since, Health and Safety dataset include the operation type variable, which describes the type of operation (lifting, pulling, cutting, etc.) with the specific tool (equipment) for the given task. Understanding the distribution of injury by body parts can highlight the top-k operations, for instance, that result in accidents to body parts. A graphical statistical tool (Pie chart) to explore this information is as depicted in Figure 5, where it is observed that certain body parts are prone to injuries during the power infrastructure project construction. The injury distribution of the top-5 body parts as specified in the database is as follows: Fingers (23%), Hand (13%), Back/Buttocks (12%), and Ankle (8%). The top five operations resulting in these injuries are pulling (stringing), lifting, loading/offloading, manual handling, and cutting because these parts are essential for carrying out these operations (Chi & Han 2013). The observation from this is probably that most of the accidents are as a result of carelessness, distractions, and disregard for safety procedures. The exploratory analysis results are in agreement with Fan et al. (2014). This fine-grained knowledge is not only integral to the development of robust construction health and safety risk management but also critical for stakeholders to enforce best safety practices to minimise accidents.

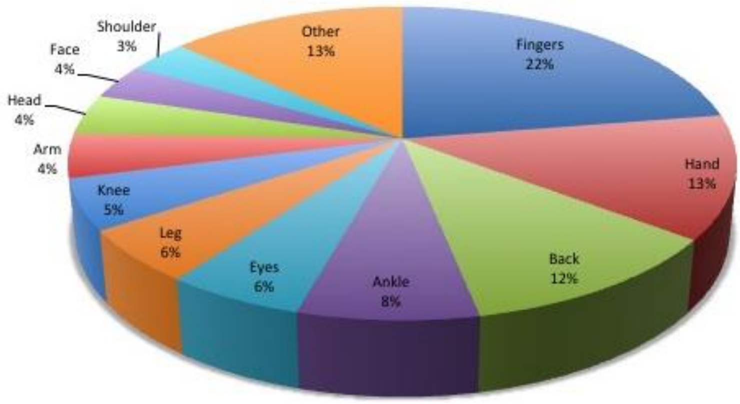

Figure 5. Injury distribution by body parts

5.2. *Incident distribution by season*

Constructing power infrastructure (i.e. overhead lines) is mostly an outdoor activity, and certain types of accidents are more likely due to the changing seasonal conditions (summer, winter, autumn, and spring). Figure 6 shows that winter has the highest percentage of incidents (29%), followed by autumn (25%), spring (24%) and summer (23%). Scotland has a temperate and oceanic climate that is very cold in winter, due to frequent and heavy hail and snow showers. Wales likewise, has a temperate climate and tends to be wetter than England.

Trips, slips, and falls are among the most common incidents in these regions due to the reduced visibility. Temperatures near or below freezing and strong winds can also result in severe illness and injury. Additionally, vehicle accidents occur due to the effects of ice and snow on muddy roads.

The use of Big Data analytics for automatic extraction and dissemination of climatic conditions of a region in real-time will go a long way at mitigating injuries that are synonymous to that region (location).

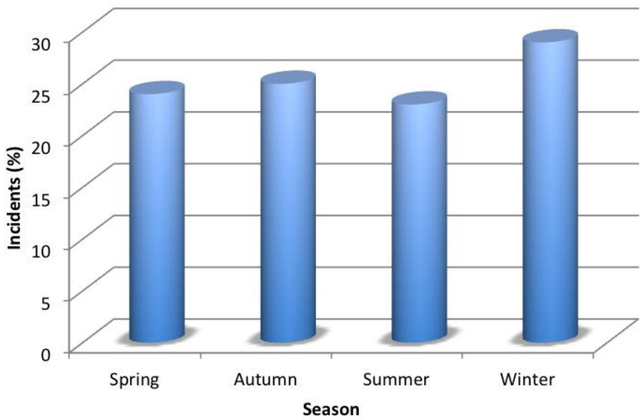

Figure 6. Incidents distribution by season

5.3. *Accident distribution by spatial analysis*

Often, the top management of a construction company may be interested in regions with high incident rates. Offering this service will equip managers with adequate information to proactively react to health and safety challenges in such regions. Thus, spatial analysis is of immense importance in such situations in that it enables the analysis of incidents over the topological and geographical spread. In the health and safety dataset, the location information is captured in the 'site' column. For the spatial analysis, the dataset is pre-processed to extract the UK postcode of each incident record and linked with the corresponding latitude and longitude data from Doogal (http://www.doogal.co.uk/UKPostcodes.php). The geographical heat map is employed to visualise the resulting data. Figure 7 shows the summary of this distribution, where the size of spheres represents the proportion of accidents (computed as percentages) in each region. Scotland has the highest (30%), followed by Wales and South West (25%), North (16%), South East (14%), and Midlands (2%). The frequency of severe weather is observed to be the leading cause of accidents in Scotland as well as Wales and South West regions. Strong wind, for instance, may lead to shattering of vehicle windscreens and a collapse of a fence or unit. Icy weather may result in trips and slips. Also, heavy-duty machinery operation (i.e. excavation and road cutting) is often the cause of utility service damage (i.e. gas pipelines, water supply). Even though geological conditions in different cities are complex, existing health and safety risk management approaches do not consider making this information available for proper health and safety risk prevention. To efficiently bring health and safety risk in the site under control, incorporating a module to automatically compute the geology and hydrology condition of construction sites in real-time will improve the optimal control of occupational hazards.

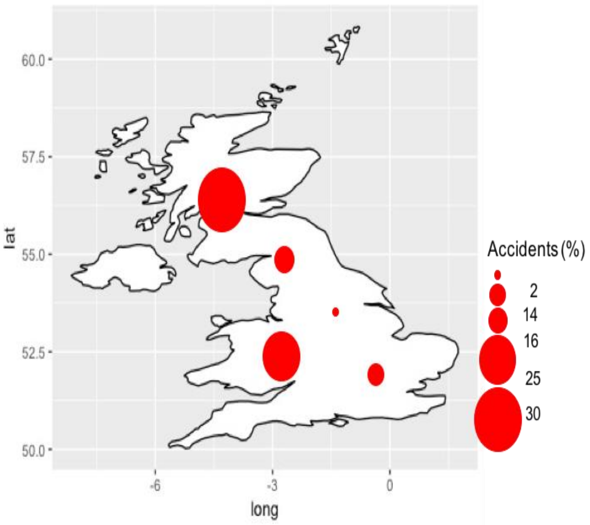

Figure 7. Spatial analysis of accidents

Additionally, the result of viewing the regions with respect to incident (or accident) rate can further be narrowed to cities and a specific location. The impact of location on incidents is worth further exploration. This investigation is the focus of future research on the proposed architecture.

5.4. *Modelling the relationship between variables*

Tremendous R&D efforts have been carried out to reduce the impacts of occupational health hazards. One such attempt is in modelling and analysing several variables (i.e. determining the relationships between the predictors (independent variables) and the dependent variable. Robust and efficient machine learning techniques such as deep learning, gradient boosting machines, and linear multivariate regression are employed in modelling relationships among variables. In this paper, a demonstration of the linear regression technique is made due to its simplicity.

Linear multivariate regression, in this regard, advocates methods for analysing health hazards with respect to the project cost. This concept not only enables the exploratory analysis of injury but also allows predictive accident analytics. The principle of the linear multivariate regression is to predict Y as a linear combination of the input variables $(x_1, x_2, \cdots, x_p)$ plus an error term ϵ_i.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots\cdots + \beta_p x_{ip} + \epsilon_i, i \in [1, n]$$

n is the number of sample data, p the number of variables and β_0 a bias. This model can conveniently be written as $y = X\beta + \epsilon$, where

$$y = (y_1, \cdots, y_n)^T,\ \epsilon = (\epsilon_1, \cdots, \epsilon_n)^T,\ \beta = (\beta_1, \cdots, \beta_n)^T, \text{ and } X = \begin{pmatrix} 1 & x_{11} & \vdots & x_{1p} \\ 1 & x_{21} & \vdots & x_{2p} \\ 1 & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \vdots & x_{xp} \end{pmatrix}$$

The predicted or fitted value is thus, $\hat{y} = X\hat{\beta}$, where $\hat{\beta}$ is the least squares estimate of β.

The model can be used for example to predict the body part injured given a set of inputs such as the type of operation (task), equipment being used, kind of power infrastructure project, the project complexity, project contract type, etc. A practical but straightforward illustration is to determine the relationship between the project cost and occupational hazards (linear regression with one predictor) is depicted using a line plot (Figure 8). The x-axis of the plot represents the project cost while the y-axis represents the health hazards risk (incidents and accidents). The line plot shows a significant increase in the number of health hazards (accident and incidents) as the project cost increases. Consequently, the number of occupational health risk is proportional to the project cost. This result is expected since the project cost is a crucial factor in determining the complexity of a project. Thus, the more complex a project is, the more are incidents associated with it.

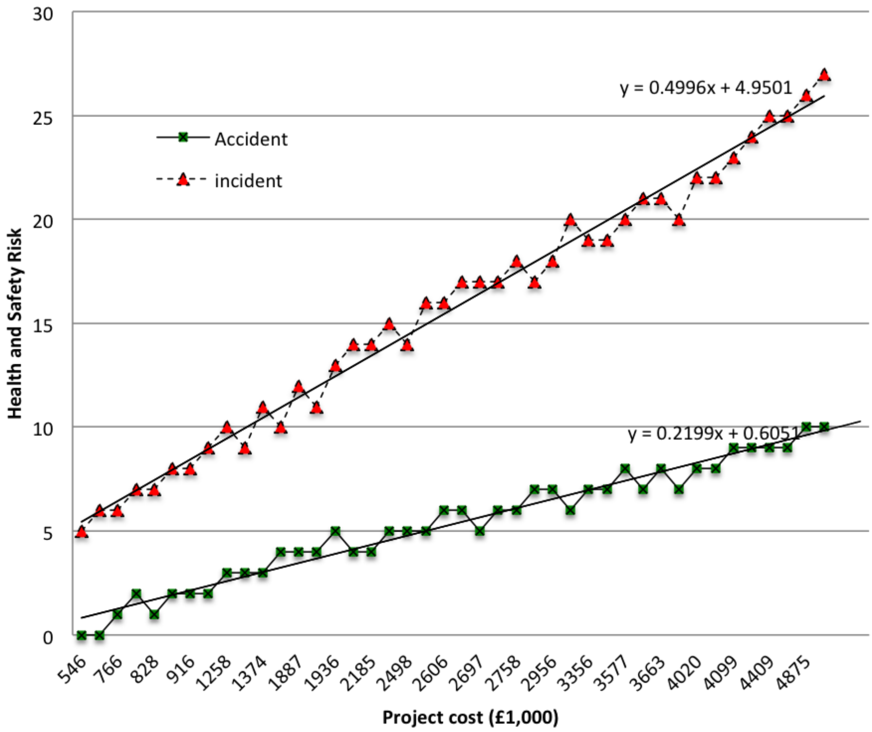

Figure 8. Relationship among variables

6. **Conclusions**

Construction safety risk analyses are currently limited because existing techniques overlook the complex and dynamic nature of construction sites. Besides, they ignore or simplify some key factors and pay little attention to analysing the relationship between a safety phenomenon and safety data. Today, large and dynamic data with various data types are to be analysed. In implementing the health hazards management tool, the Big Data architecture that is based on a well coherent health risk analytics lifecycle is proposed. The Big Data technology was selected due to its support for massive, high dimensional, heterogeneous, complex, unstructured, incomplete, and noisy data.

The preliminary results obtained in this study using the various Big Data frameworks have enabled us to design a robust architecture to handle and analyse power infrastructure accident data. The proposed architecture can identify relevant variables and improve preliminary prediction accuracies and explanatory capacities. It has also enabled conclusions to be drawn regarding the causes of health hazards. The results obtained in this study represent a significant improvement in terms of managing information on construction accidents, particularly for power infrastructure companies. The satisfactory results of the B-DAPP tool have indicated the reliability and appropriateness of the selected Big Data components for studies of construction health risks and their causes.

Future research is aimed at rigorously evaluating accuracies of both the prediction and prescription of the software deployed in real-time. Additionally, other researchers should look in the area of designing and planning a more ambitious, larger scale models to gain a deeper understanding of accident causes in various industrial sectors.

References

Al-Jarrah, O.Y. et al., 2015. Efficient Machine Learning for Big Data: A Review. *Big Data Research*, 2(3), pp.87–93.

Bohle, P. et al., 2015. Health and well-being of older workers: Comparing their associations with effort–reward imbalance and pressure, disorganisation and regulatory failure. *Work & Stress*, 8373, pp.1–14.

Camann, D.E. et al., 2011. *Data Science and Big Data Analytics*, New York: EMC.

Carbonari, A., Giretti, A. & Naticchia, B., 2011. A proactive system for real-time safety management in construction sites. *Automation in Construction*, 20(6), pp.686–698. Available at: http://dx.doi.org/10.1016/j.autcon.2011.04.019.

Chaudhuri, S. & Dayal, U., 1997. An overview of data warehousing and OLAP technology. *ACM SIGMOD Rec*, 26, pp.65–74.

Chen, J., Qiu, J. & Ahn, C., 2017. Construction worker’s awkward posture recognition through supervised motion tensor decomposition. *Automation in Construction*, 77, pp.67–81.

Cheng, C. et al., 2011. Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan’s construction industry. *Accident Analysis and Prevention*, 48, pp.214–222.

Chi, S. & Han, S., 2013. Analyses of systems theory for construction accident prevention with specific reference to OSHA accident reports. *International Journal of Project Management*, 31(7), pp.1027–1041.

Ciarapica, F.E. & Giacchetta, G., 2009. Classification and prediction of occupational injury risk using soft computing techniques : An Italian study. *Safety Science*, 47(1), pp.36–49. Available at: http://dx.doi.org/10.1016/j.ssci.2008.01.006.

Debnath, J. et al., 2016. Fuzzy inference model for assessing occupational risks in construction sites. *International Journal of Industrial Ergonomics*, 55, pp.114–128.

Delen, D. & Demirkan, H., 2013. Data, information and analytics as services. *Decision Support Systems*, 55(1), pp.359–363.

Esmaeili, B., Hallowell, M.R. & Rajagopalan, B., 2015. Attribute-based safety risk assessment. II: Predicting safety outcomes using generalized linear models. *Journal of Construction Engineering and Management*, 141(8), pp.1–11.

Fan, Z.J. et al., 2014. The association between combination of hand force and forearm posture and incidence of lateral epicondylitis in a working population. *Human factors*, 56, pp.151–165.

Favarò, F.M. & Saleh, J.H., 2016. Toward risk assessment 2.0: Safety supervisory control and model-based hazard monitoring for risk-informed safety interventions. *Reliability Engineering and System Safety*, 152, pp.316–330. Available at: http://dx.doi.org/10.1016/j.ress.2016.03.022.

Fenrick, L. & Getachew, S., 2012. Cost and reliability comparisons of underground and overhead power lines. *Utilities Policy*, 20(1), pp.31–37.

Fragiadakis, N., Tsoukalas, V. & Papazoglou, V., 2014. An adaptive neuro-fuzzy inference system (anfis) model for assessing occupational risk in the shipbuilding industry. *Safety Science*, 63, pp.226–235.

Galizzi, M. & Tempesti, T., 2015. Workers' risk tolerance and occupational injuries. *Risk Analysis*, 35(10), pp.1858–1875.

Gandomi, M. & Haider, A., 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), pp.137–144.

García-Herrero, S. et al., 2012. Working conditions, psychological/physical symptoms and occupational accidents Bayesian network models. *Safety Science*, 50(9), pp.1760–1774.

Gholizadeh, P. & Esmaeili, B., 2016. Applying classification trees to analyze electrical contractors' accidents. In *Construction Research Congress*. San Juan, Puerto Rico, pp. 2699–2708.

Groves, W., Kecejovic, V. & Komljenovic, D., 2007. Analysis of fatalities and injuries involving mining equipment. *Journal of Safety Research*, 38(4), pp.461–470.

Guo, B., Yiu, T. & González, V., 2016. Predicting safety behavior in the construction industry: Development and test of an integrative model. *Safety Science*, 84, pp.1–11. Available at: http://dx.doi.org/10.1016/j.ssci.2015.11.020.

Guo, S. et al., 2016. A Big-Data-based platform of workers' behavior: Observations from the field. *Accident Analysis and Prevention*, 93, pp.299–309. Available at: http://dx.doi.org/10.1016/j.aap.2015.09.024.

Gürcanli, G. & Müngena, U., 2009. An occupational safety risk analysis method at construction sites using fuzzy sets. *International Journal of Industrial Ergonomics*, 39(2), pp.371–387.

Hallowell, M.R. & Gambatese, J.A., 2009. Construction Safety Risk Mitigation. *Journal of Construction Engineering and Management*, 135(12), pp.1316–1323. Available at: http://ascelibrary.org/doi/10.1061/%28ASCE%29CO.1943-7862.0000107.

Haslam, R.A. et al., 2005. Contributing factors in construction accidents. *Applied Ergonomics*, 36(4), p.401–415.

HSE, 2016. Health and safety at work Summary statistics for Great Britain. Available at: http://www.hse.gov.uk/statistics/overall/hssh1516.pdf?pdf=hssh1516 [Accessed April 14, 2017].

Jacinto, C. & Silva, C., 2010. A semi-quantitative assessment of occupational risks using bow-tie representation. *Safety Science*, 48, pp.973–979.

Jin, X. et al., 2015. Significance and Challenges of Big Data Research. *Big Data Research*, 2(2), pp.59–64.

Jocelyn, S. et al., 2017. Application of logical analysis of data to machinery-related accident prevention based on scarce data. *Reliability Engineering and System Safety*, 159(May 2016), pp.223–236. Available at: http://dx.doi.org/10.1016/j.ress.2016.11.015.

Khakzad, N., Khan, F. & Amyotte, P., 2015. Major Accidents (Gray Swans) Likelihood Modeling Using Accident Precursors and Approximate Reasoning. *Risk Analysis*, 35(7), pp.1336–1347.

Landset, S. et al., 2015. A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *Journal of Big Data*, 2(1), pp.1–36.

Leavitt, N., 2010. Will NoSQL Databases Live Up to Their Promise? *IEE Computer Journal*, 43(2), pp.12–14.

Li, H. et al., 2016. Stochastic state sequence model to predict construction site safety states through real-time location systems. *Safety Science*, 84, pp.78–87.

Li, Y. & Bai, Y., 2008. Comparison of characteristics between fatal and injury accidents in the highway construction zones. *Safety Science*, 46(4), pp.646–660.

Liao, C.-W. & Perng, Y.-H., 2008. Data mining for occupational injuries in the Taiwan construction industry. *Safety Science*, 46(7), pp.1091–1102.

Liu, H. & Tsai, Y., 2012. A fuzzy risk assessment approach for occupational hazards in the construction industry. *Safety Science*, 50(4), pp.1067–1078. Available at: http://dx.doi.org/10.1016/j.ssci.2011.11.021.

Love, P.E.D. & Teo, P., 2017. Statistical Analysis of Injury and Nonconformance Frequencies in Construction : Negative Binomial Regression Model. *Journal of Construction Engineering and Management*, 143(8), pp.1–9.

McDermott, V. & Hayes, J., 2016. “We’re still hitting things”: The effectiveness of third party processes for pipeline strike prevention. In *Proceedings of the eleventh international pipeline conference (IPC 2016)*. Calgary, Alberta, Canada, pp. 1–10.

Naderpour, M., Lu, J. & Zhang, G., 2016. A safety-critical decision support system evaluation using situation awareness and workload measures. *Reliability Engineering and System Safety*, 150, pp.147–159. Available at: http://dx.doi.org/10.1016/j.ress.2016.01.024.

Najafabadi, M.M. et al., 2015. Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1), pp.1–21.

Nanda, G. et al., 2016. Bayesian decision support for coding occupational injury data. *Journal of Safety Research*, 57, pp.71–82. Available at: http://dx.doi.org/10.1016/j.jsr.2016.03.001.

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

Papazoglou, I. et al., 2017. Quantitative occupational risk model: Single hazard. *Reliability Engineering & System Safety*, 160, pp.162–173.

Papazoglou, I. et al., 2015. Uncertainty Assessment in the Quantification of Risk Rates of Occupational Accidents. *Risk Analysis*, 35(8), pp.1536–1561.

Pinto, A., Nunes, I. & Ribeiro, R., 2011. Occupational risk assessment in construction industry – Overview and reflection. *Safety Science*, 49, pp.616–624.

Power, D., 2014. Using “Big Data” for analytics and decision support. *Journal of Decision Systems*, 23(2), pp.222–228.

Rahman, M.N. & Esmailpour, A., 2016. A Hybrid Data Center Architecture for Big Data. *Big Data Research*, 3, pp.29–40.

Raviv, G., Shapira, A. & Fishbain, B., 2017. AHP-based analysis of the risk potential of safety incidents: Case study of cranes in the construction industry. *Safety Science*, 91, pp.298–309. Available at: http://dx.doi.org/10.1016/j.ssci.2016.08.027.

Rivas, T. et al., 2011. Explaining and predicting workplace accidents using data-mining techniques. *Reliability Engineering & System Safety*, 96(7), pp.739–747.

Ryza, 0] S. et al., 2015. *Advanced Analytics with Spark*, Cambridge: O’Reilly,.

Schryver, J., Shankar, M. & Xu, S., 2012. Moving from descriptive to causal analytics: Case study of discovering knowledge from US health indicators warehouse. In *ACM SIGKDD Workshop on Health Informatics*. Beijing, China, pp. 1–8.

Silva, S.A. et al., 2016. Organizational practices for learning with work accidents throughout their information cycle. *Safety Science*, In Press.

Soltanzadeh, A. et al., 2016. Analysis of occupational accidents induced human injuries: A case study in construction industries and sites. *Journal of Civil Engineering and Construction Technology*, 7(1), pp.1–7. Available at: http://academicjournals.org/journal/JCECT/article-abstract/15EEFC357741.

Suthakar, U. et al., 2016. An efficient strategy for the collection and storage of large volumes of data for computation. *Journal of Big Data*, 3(1), pp.1–17.

Tixier, A.. et al., 2016. Application of machine learning to construction injury prediction. *Automation in Construction*, 69, pp.102–114.

Törner, M. & Pousette, A., 2009. Safety in construction - a comprehensive description of the characteristics of high safety standards in construction work, from the combined perspective of supervisors and experienced workers. *Journal of Safety Research*, 40(6), pp.399–409.

Tsai, C.W. et al., 2015. Big data analytics: a survey. *Journal of Big Data*, 2(1), pp.1–32.

Venturini, L., Baralis, E. & Garza, P., 2017. Scaling associative classification for very large datasets. *Journal of Big Data*, 4(1). Available at: https://doi.org/10.1186/s40537-017-0107-2.

Weng, J., Meng, Q. & Wang, D.Z.W., 2013. Tree-based logistic regression approach for work zone casualty risk assessment. *Risk Analysis*, 33(3), pp.493–504.

White, T., 2012. *Hadoop: The Definitive Guide*, Sebastopol, CA: O’Reilly Media, Inc.

Wu, W. et al., 2010. Towards an autonomous real-time tracking system of near-miss accidents on construction sites. *Automation in Construction*, 19(2), pp.134–141. Available at: http://dx.doi.org/10.1016/j.autcon.2009.11.017.

Yi, W. et al., 2016. Development of an early-warning system for site work in hot and humid environments: A case study. *Automation in Construction*, 62, pp.101–113. Available at: http://dx.doi.org/10.1016/j.autcon.2015.11.003.

Yoon, Y.S., Ham, D.H. & Yoon, W.C., 2016. Application of activity theory to analysis of human-related accidents: Method and case studies. *Reliability Engineering and System Safety*, 150, pp.22–34. Available at: http://dx.doi.org/10.1016/j.ress.2016.01.013.

Yorio, P.L., Willmer, D.R. & Haight, J.M., 2014. Interpreting MSHA citations through the lens of occupational health and safety management systems: Investigating their impact on mine injuries and illnesses 2003-2010. *Risk Analysis*, 34(8), pp.1538–1553.

Zang, W. et al., 2014. Comparative study between incremental and ensemble learning on data streams: Case study. *Journal Of Big Data*, pp.1–16. Available at: http://www.journalofbigdata.com/content/1/1/5/abstract.

Zeng, S.X., Tam, V.W.Y. & Tam, C.M., 2008. Towards occupational health and safety systems in the construction industry of China. *Safety Science*, 46, pp.1155–1168.

Zhang, L. et al., 2016. Towards a Fuzzy Bayesian Network Based Approach for Safety Risk Analysis of Tunnel-Induced Pipeline Damage. *Risk Analysis*, 36(2), pp.278–301.

Zhou, Z., Goh, Y. & Li, Q., 2015. Overview and analysis of safety management studies in the construction industry. *Safety Science*, 72, pp.337–350.

Zhu, Z. et al., 2016. Predicting movements of onsite workers and mobile equipment for enhancing construction site safety. *Automation in Construction*, 68, pp.95–101.

Zou, P.X.W., Zhang, G. & Wang, J., 2007. Understanding the key risks in construction projects in China. *International Journal of Project Management*, 25(6), pp.601–614.