Big Data in the construction industry: A review of present status, opportunities, and future trends

Bilal, M, Oyedele, LO, Qadir, J, Munir, K, Ajayi, SO, Akinade, OO, Owolabi, HA, Alaka, HA & Pasha, M

Original citation & hyperlink:
Bilal, M, Oyedele, LO, Qadir, J, Munir, K, Ajayi, SO, Akinade, OO, Owolabi, HA, Alaka, HA & Pasha, M 2016, 'Big Data in the construction industry: A review of present status, opportunities, and future trends' *Advanced Engineering Informatics*, vol 30, no. 3, pp. 500-521
https://dx.doi.org/10.1016/j.aei.2016.07.001

DOI 10.1016/j.aei.2016.07.001
ESSN 2352-7102

Publisher: Elsevier

Big Data in the Construction Industry: A Review of Present Status, Opportunities, and Future Trends

***Abstract*—The ability to process large amounts of data and to extract useful insights from data has revolutionised society. This phenomenon—dubbed as Big Data—has applications for a wide assortment of industries, including the construction industry. The construction industry already deals with large volumes of heterogeneous data; which is expected to increase exponentially as technologies such as sensor networks and the Internet of Things are commoditized. In this paper, we present a detailed survey of the literature, investigating the application of Big Data techniques in the construction industry. We reviewed related works published in the databases of American Association of Civil Engineers (ASCE), Institute of Electrical and Electronics Engineers (IEEE), Association of Computing Machinery (ACM), and Elsevier Science Direct Digital Library. While the application of data analytics in the construction industry is not new, the adoption of Big Data technologies in this industry remains at a nascent stage and lags the broad uptake of these technologies in other fields. To the best of our knowledge, there is currently no comprehensive survey of Big Data techniques in the context of the construction industry. This paper fills the void and presents a wide-ranging interdisciplinary review of literature of fields such as statistics, data mining and warehousing, machine learning, and Big Data Analytics in the context of the construction industry. We discuss the current state of adoption of Big Data in the construction industry and discuss the future potential of such technologies across the multiple domain-specific sub-areas of the construction industry. We also propose open issues and directions for future work along with potential pitfalls associated with Big Data adoption in the industry.**

I. Introduction

The world is currently inundated with data, with fast advancing technology leading to its steady increase. Today, companies deal with petabytes (10^{15} bytes) of data. Google processes above 24 petabytes of data per day [1], while Facebook gets more than 10 million photos *per hour* [1]. The glut of data increased in 2012 is approximately 2.5 quintillion (10^{18}) bytes per day [2]. This data growth brings significant opportunities to scientists for identifying useful insights and knowledge. Arguably, the accessibility of data can improve the status quo in various fields by strengthening existing statistical and algorithmic methods [3], or by even making them redundant [4].

The construction industry is not an exception to the pervasive digital revolution. The industry is dealing with significant data arising from diverse disciplines throughout the life cycle of a facility. Building Information Modelling (BIM) is envisioned to capture multi-dimensional CAD information systematically for supporting multidisciplinary collaboration among the stakeholders [5]. BIM data is typically 3D geometric encoded, compute intensive (graphics and Boolean computing), compressed, in diverse proprietary formats, and intertwined [6]. Accordingly, this diverse data is collated in federated BIM models, which are enriched gradually and persisted beyond the *end-of-life* of facilities. BIM files can quickly get voluminous, with the design data of a 3-story building model easily reaching 50GB in size [7]. Noticeably, this data in any form and shape has intrinsic value to the performance of the industry. With the advent of embedded devices and sensors, facilities have even started to generate massive data during the operations and maintenance stage, eventually leading to more rich sources of Big BIM Data. This vast accumulation of BIM data has pushed the construction industry to enter the Big Data era.

Big Data has three defining attributes (a.k.a. 3V‘s), namely (i) volume (terabytes, petabytes of data and beyond); (ii) variety (heterogeneous formats like text, sensors, audio, video, graphs and more); and (iii) velocity (continuous streams of the data). The 3V‘s of Big Data are clearly evident in construction data. Construction data is typically large, heterogeneous, and dynamic [8]. Construction data is voluminous due to large volumes of design data, schedules, Enterprise Resource Planning (ERP) systems, financial data, etc. The diversity of construction data can be observed by noting the various formats supported in construction applications including DWG (short for drawing), DXF (drawing exchange format), DGN (short for design), RVT (short for Revit), ifcXML (Industry Foundation Classes XML), ifcOWL (Industry Foundation Classes OWL), DOC/XLS/PPT (Microsoft format), RM/MPG (video format), and JPEG (image format). The dynamic nature of construction data follows from the streaming nature of data sources such as Sensors, RFIDs, and BMS (Building Management System). Utilising this data to optimise construction operations is the next frontier of innovation in the industry.

[Fig. 1 about here.]

To understand the subtleties of Big Data, we need to disambiguate between two of its complementary aspects: Big Data Engineering (BDE) and Big Data Analytics (BDA). The domain of BDE is primarily concerned with supporting the relevant data storage and processing activities, needed for analytics [9]. BDE encompasses technology stacks such as Hadoop and Berkeley Data Analytics Stack (BDAS). Big Data Analytics (BDA), the second integral aspect, relates to the tasks responsible for extracting the knowledge to drive decision-making [9]. BDA is mostly concerned with the principles, processes, and techniques to understand the Big Data. The essence of BDA is to discover the latent patterns buried inside Big Data and derive useful insights therefrom [10]. These insights have the capability to transform the future of many industries through data-driven decision-making. This ability to identify, understanding and reacting to the latent trends promptly is indeed a competitive edge in this hyper-competitive era.

Contributions of this paper: While some data-driven solutions have been proposed for the fields of the construction industry, there is currently no comprehensive survey of the literature, targeting the application of Big Data in the context of the construction industry. This paper fills the void and presents a wide-ranging interdisciplinary study of fields such as Statistics, Data Mining and Warehousing, Machine Learning, Big Data and their applications in the construction industry.

Organization of this paper: The discussion in this paper follow the review structure shown in Fig. 1. We start with a thorough review of extant literature on BDE and BDA in the construction industry in Section II and III, respectively. After which, opportunities of Big Data in the construction industry sub-domains are presented in Section IV. Discussions about open research issues and future work, and pitfalls of Big Data in the construction industry are then presented in Section V and VI, respectively.

II. Big Data Engineering (BDE)

Big Data Engineering (BDE) provide infrastructure to support Big Data Analytics (BDA). Some discussions about the Big Data platforms worth consideration to understand the BDE adequately. Various Big Data platforms are developed so far with varied characteristics, which can be divided into two groups: (i) *horizontal scaling platforms (HSPs)*, the ones that distribute processing across multiple servers and scale out by adding new machines to the cluster. (ii) And *vertical scaling platforms (VSPs)*, in which scaling is achieved by upgrading hardware (processor or memory or disk) of the underlying server since it is a single server-based configuration. In the interest of brevity of this paper, the discussion here is confined to HSPs, notably Hadoop and BDAS only. We refer interested readers to Singh et al. [11] for a detailed explanation on their comparison and selection criterion.

Due to clear performance gains of BDAS over Hadoop, it is getting more attention recently. However, BDAS is in its infancy with limited support and supporting tools. Whereas, Hadoop is still widely adopted and has become the de-facto framework for Big Data applications. These platforms offer tools to store and process Big Data. Some of the most prominent tools are discussed in the subsequent sections.

A. Big Data Processing

[Fig. 2 about here.]

Parallel and distributed computation is at the core of BDE. A large number of processing models are developed for this purpose, which includes but not limited to:

*1) **MapReduce (MR)**:* MR is the distributed processing model to handle Big Data [13]. The entire analytical tasks in MR are written as two functions, i.e., *map* and *reduce* (see Fig. 2), which are submitted to separate processes called Mappers and Reducers. Mapper read data, process it, and generate intermediate results. Reducers work on mappers' output and produce final results which are stored back to the file system. Hadoop—a popular Big Data platform—introduced MR initially to the wider public and provided an ecosystem to execute MR programs successfully. In a typical Hadoop cluster, several mappers and reducers simultaneously run MR programs. MR is a powerful model for batch-processing tasks. However, it is struggling with applications that require real-time, graph, or iterative processing. Latest versions of Hadoop have encountered this issue to some extent where processing aspect of MR is detached *from* rest of the ecosystem. To this end, *Yet Another Resource Negotiator (YARN)* is introduced that has taken Hadoop to an actually computationally-agnostic Big Data platform. MR runs as a service over YARN, while YARN handles scheduling and resource management related functionalities. This separation has made Hadoop suitable for implementing innovative applications.

*2) **Directed Acyclic Graphs (DAG)**:* DAG is an alternative processing model for Big Data platforms. In contrary to MR, DAG relaxes the rigid *map-then-reduce* style of MR to a more generic notion. BDAS—an emerging Big Data platform—supports this kind of data processing through its resilient component called Spark [14]. Spark holds supremacy over MR in many aspects. Particularly, in-memory computation and high expressiveness are keys to wider adoption of Spark. These capabilities heralded the Spark a natural choice to support iterative as well as reactive applications [14]. Spark is reported to have ten times faster than MR on disk-resident tasks, whereas hundred times faster for memory-resident tasks [11]. Fig. 3 shows components of Spark. These technologies are designed to support functions that are vital to the development of enterprise applications.

[Fig. 3 about here.]

Examples of Construction Research using Big Data Processing: MR and Spark have use cases across myriad information systems (IS) of the construction industry. Despite significance, these tools are rarely used to process BIM data in construction industry applications.

Chang et al. [16] customised MR for BIM data (MR4B) to optimise the retrieval of partial BIM models. They found legacy data distribution logic of Hadoop MR inadequate, since BIM data is intertwined as well as highly relative, and merely placing it randomly might sparsely distribute BIM elements across different blocks on Hadoop cluster nodes. Such placement degrades querying performance due to increased disk I/O required to bring sparsely distributed data together for analysis using MR. To overcome this, a data pre-partition and processing step is devised to parse, analyse and partition logically relevant parts of BIM data (by floor number or material family) and store it in the adjacent spaces on the Hadoop cluster. Node multi-threading is introduced to utilise the CPU maximally during analysis [16]. This way Hadoop is customised for BIM data and querying components are implemented as YARN applications. A BIM system for clash detection and quantity estimation is developed to exploit the proposed YARN applications. It is reported that the system has improved the performance manifold, and the required tasks are executed at real time with reasonable response time.

Lin et al. [7] presented the development of a specialised big BIM data storage and retrieval system for experts and naive BIM users. The intentions are to develop a highly interactive user interface for querying BIM data through mobile devices to maximise its utility and usability. User queries in plain English are re-formulated using the proposed natural language processing approach to retrieve highly complex BIM data, which are mapped onto a variety of visualisations. To optimise query execution, an MR join pre-processing is demonstrated to merge two BIM collections before query evaluation. The response time is reported to have enhanced by more than 40% compared to the same join pre-processing written in traditional technologies.

B. *Big Data Storage*

Another aspect of BDE is the Big Data storage, which is provided either by the distributed file systems or emerging NoSQL databases. These technologies are briefly discussed in the following subsections.

*1) **Distributed File Systems**:* In this subsection, we are discussing two competing distributed file systems, namely HDFS and Tachyon.

- ***Hadoop Distributed File System (HDFS)***—HDFS is suitably designed for managing the larger datasets [17]. It is designed specifically to work with a cluster of commodity servers. Since the chances of hardware failure are higher in such settings, it provides greater fault tolerance for hardware failures. Data distribution and replications are the key traits of HDFS to achieve the fault tolerance and high availability. There are however situations when the usage of HDFS degrades performance, particularly in applications requiring low-latency data access. Similarly, it is also not ideal for storing a large number of small files due to the associated overhead for managing their metadata. Lastly, HDFS is not the choice of technology if applications require a significant number of concurrent modifications at random places in data.
- ***Tachyon*** is the BDAS flagship distributed file system that extends HDFS and provides access to the distributed data at memory speed across the cluster. Some of the features where Tachyon has outsmarted HDFS include: (i) in-memory data caching for enhanced performance and (ii) backwards compatibility to work seamlessly with Spark as well as MR tasks without any code changes required to the programs.

*2) **NoSQL Databases**:* Relational databases served IT industry for the past couple of decades as de facto data management standard. However, recently applications emerged that demanded more scalability, performance, and flexibility. Relational databases are found unsuitable for these applications due to their specialised storage and processing needs. Consequently, new systems came into being—called "Not only SQL" (NoSQL) systems—to fill this technological gap. NoSQL systems improved traditional data management in numerous ways. More importantly, NoSQL systems eschew the rigid schema-oriented storage in favour of schema-less storage to achieve flexibility [18]. Today these systems are prevalent in myriad data-intensive applications in many industries. Pointedly, the architecture of NoSQL systems is well suited to fragmented nature of construction industry's data.

NoSQL systems store schema-less data in a non-relational data model. Presumably, there are four data models for these systems.

1) ***Key-Value:*** This is the simplest data model to store unstructured data. However, the underlying data is not self-describing.
2) ***Document:*** This data model is suitable for storing self-describing entities. However, the storage of this model can be inefficient.
3) ***Columnar:*** This data model favours the storage of sparse datasets, grouped sub-columns, and aggregated columns.
4) ***Graph:*** This is a relatively new data model that supports relationship traversal over a huge dataset of property-graphs. Graph databases are getting popular than other data models (see Fig. 4, where the *x-axis* represents the period of popularity and *y-axis* shows a change in popularity). Table VIII describes features of 12 prominent databases.

[TABLE 1 about here.]

[Fig. 4 about here.]

Examples of Construction Research using Big Data Storage: Despite significance for massive BIM data storage, existing applications are still lacking their successful implementation. Das et al. [20] proposed Social-BIM to capture social interactions of users along with the building models. A distributed BIM framework, called BIMCloud, is developed to store this data through IFC. Apache Cassandra, hosted on Amazon EC2, is used. Jeong et al. [21] proposed a hybrid data management infrastructure comprised two tiers. The *client tier* that utilises MongoDB for storing the structured data temporarily for efficiently completing analytical tasks, whereas, the *central tier* employs Apache Cassandra to store permanently the streams of sensor data generated over time. Cheng et al. [22] have also employed the Apache Cassandra for presenting their query language to extract partial BIM models. Similarly, Lin et al. [7] exploited MongoDB to store BIM data of building models for distributed processing through MapReduce. MongoDB is tailored for IFC, with minor alterations to IFC hierarchy for supporting MR-efficient query execution.

III. Big Data Analytics

Big Data Analytics has a rich intellectual tradition and borrows from a wide variety of fields. There have been traditionally many related disciplines that have essentially the same core focus: finding useful patterns in data (but with a different emphasis). These related fields are Statistics (1830[1])

[1]While it can be difficult to pin down the exact time of genesis of a technology, the year in which the domain's seminal work was proposed is provided to approximately sequence the various Big Data Analytics enabling technologies chronologically.

[23], Data Mining (1980), Predictive Analytics (1989 [24]), Business Analytics (1997), Knowledge Discovery from Data (KDD) (2002), Data Analytics (2010), Data Science (2010) and now Big Data (2012). Fig. 5 shows the relevance of these multidisciplinary fields to Big Data. So, Big Data Analytics is a broadening of the field of data analytics and incorporates many of the techniques that have already been performed. This is the key reason that most of the existing work, presented in subsequent subsections, has focused on data analytics rather than Big Data is that the Big Data revolution—i.e., the ability to process large amounts of diverse data on a large scale—has only recently happened. Existing approaches can be possibly extended to the environments, dealing with large, diverse datasets.

[TABLE 2 about here.]

Some ML-based tools have been developed for Big Data analytics. Table IX highlight some of the important ones. To showcase the implementation of BDA, we use MLlib (MLbase) code in the subsequent subsections.

[Fig. 5 about here.]

*1) **Statistics**:* In scientific studies, rigorous and efficient techniques are used to answer research questions. Careful observations (data) comprise the backbone of underpinning investigations. Statistics is the study of collecting, analysing, and drawing conclusions from the data, with the primary focus on selecting the right tools and techniques at every data analysis stage [29]. Right from the data collections, to efficiently analysing it, and then inferring or formulating conclusions out of it, all of these steps comes under the scope of statistics [30]. Various fields of analytics are borrowing techniques from statistics [29].

Examples of Construction Research using Statistics: The industry is employing statistical methods in a variety of application areas, such as identifying causes of construction delays [31], learning from post-project reviews (PPRs) [32], decision support for construction litigation [33], detecting structural damages of buildings [34], identifying actions of workers and heavy machinery [35], [36], etc., are to name a few.

[TABLE 3 about here.]

*2) **Data Mining**:* ***Data Mining*** is concerned with the automatic or semi-automatic exploration and analysis, of large volumes of data, to discover meaningful patterns or rules. Data Mining has the broader scope than other traditional data analysis fields (such as statistics) since it tends to answer non-trivial questions [37], [38]. For patterns discovery and extraction, Data Mining is primarily based on the technique(s) from statistics, machine learning, and pattern recognition [39], [40]. Several models are created and tested to assess the suitability of particular technique(s) for solving the given business problem. Models with the highest accuracy and tolerance are chosen and applied to the actual data for generating predictive results (including predictions, rules, probability, and predictive confidence).

Databases are crucial to empowering various aspects of data mining, in particular by taking care of the activities of efficient data access, group and ordering of operations and optimising the queries to scale up data mining algorithms. Databases provide native support for analytics in the form of *data warehousing*. In data warehousing, the copy of the transactional data is stored specifically structured for querying and the analysis [37], [41]. The transactional data is collated from the operational databases using a process usually known as Extract, Transform, and Load (ETL) [42]. Data in the warehouse is typically analysed through the Online Analytical Processing (OLAP). OLAP outperforms SQL in computing the summaries (roll-up) and breakdown (roll-down) of the data.

Examples of Construction Research using Data Mining: Kim et al. [31] employed data mining techniques to identify the key factors that cause delays in construction projects. They presented knowledge discovery in databases (KDD) framework to analyse massive construction datasets. Limitations of ML algorithms (such as incorrect prediction) are discussed and overcome through statistical methods. Buchheit et al. [43] also presented KDD process for the construction industry. Data preprocessing is found to be the most challenging and time-consuming step. Also, Soibelman et al. [44] illustrated the applicability of KDD to construction datasets for identifying causes of construction delays, cost overrun, and quality controls.

Carrilli et al. [32] used data mining to learn from past projects and improve future project delivery. Approaches such as text analysis, link analysis and dimensional matrix analysis are performed on data from multiple projects. Liao et al. [45] employed association rule mining to proactively prevent occupational injuries. In another similar study [46], data mining is used to explore the causes and distribution of occupational injuries and revealed that falls and collapses are the primary reasons of occupational fatalities. While Oh et al. [47] employed DW in construction productivity data, which is utilised using a multi-layer analysis through OLAP in the proposed system. SQL is quite prevalent in the industry for querying partial BIM models query languages such as Express Query Language (EQL) and Building Information Modelling Query Language (BIMQL) are developed in the various construction industry sub-domain applications [48], [49].

These datasets underlying the identification of causes of delays, learning from PPRs, BIM-based knowledge discovery, preventing occupational injury, among others, evidently present the 3V's of Big Data, and these applications can easily be extended to this emerging revolution of Big Data Analytics for features like efficiently processing querying partial BIM models.

[TABLE 4 about here.]

*3) **Machine Learning Techniques**:* Machine learning (ML), a sub-field of Artificial Intelligence (AI), focuses on the task of enabling computational systems to learn from data about specific task automatically. ML tasks can be categorized into: *i)* classification (or supervised learning); *ii)* clustering (or unsupervised learning); *iii)* association; *iv)* numeric prediction

[51].

ML has many applications across the construction applications, such as the modelling of judicial reasoning and predicting the outcomes of litigation is thoroughly studied using rule-based learning approaches [52], artificial neural networks methods [53], [54], [55], case-based reasoning techniques [56], [57], and hybrid methodologies [58], [59]. Such applications are discussed by ML techniques in the subsequent sections.

A. *Regression Techniques*

Regression is the supervised ML method, which is concerned about predicting the numerical value of a target variable based on input variables. For instance, estimating the cost of the design based on design specifications. Regression can be of the following types. The simple linear regression that is used for modelling the relationship between a dependent variable *y* and one explanatory variable *x*. Multiple linear regression that is used for modelling the relationship between one dependent variable (continuous) and two or more explanatory variables. This is commonly used regression approach. The logistic regression that is used for modelling the relationship between on categorical dependent variable and one or more explanatory variables. Listing 1 shows the MLlib code to demonstrate loading data, customising regression algorithm, developing the model, and finally using it to predict data point.

```
val df=sqlContext.createDataFrame(data).
toDF("label", "features")

val reg = new LogisticRegression().setMaxIter(15)
val model = reg.fit(df)
val weights = model.weights
model.transform(df).show()


```

Listing 1. A Snapshot of MLlib Code for Regression Analysis

Examples of Construction Research using Regression: Siu et al. [60] employed regression for predicting the cycle times of construction operations using least-square-error and least-mean-square. The approach is evaluated on a project installing Viaduct Bridge and is reported to have higher accuracy of predictions. Aibinu et al. [61] employed linear regression for identifying the delays on construction projects. Their findings reveal that cost and time overruns are frequently occurring delay factors. Similarly, Sambasivan et al. [62] studied relationship between the cause and effect of delays in the Malaysian construction industry using regression models.

Trost et al. [63] used multivariate regression analysis for predicting the accuracy of estimate during the early stages of construction projects. Estimates are given scores for gaining prediction accuracy. The results reveal that estimate score model is predicting the accuracy with very high significance. Chan et al. [64] employed multiple regression analysis for predicting the partnering success of contracting parties.

Fang et al. [65] applied logistic regression analysis to explore the relationship between safety climate and individual behaviour. The results demonstrate the vivid relationship of safety climate and personal behaviour such as gender, marital status, education level, number of family members to support, safety knowledge, drinking habits, direct employer, and individual safety behaviour.

B. *Classification Techniques*

Classification is the supervised learning technique in which programs emulate decisions automatically based on the previously made correct decisions. The input to classification algorithms is a particular set of features, and the output is to make a single selection from a short list of choices (categorical or mutually exclusive). It suits situations where single but more focused decisions are involved. Since these algorithms learn by examples, carefully crafted examples of correct decisions aside with input data are vital for algorithms to learn precisely. These algorithms learn to mimic the examples of right decisions contrary to clustering in which algorithms decide on their own without prior guidance. Classification intends to choose a single choice from the limited set of possible choices. Prominent classification algorithms include Logistic Regression, Naive Bayes, Decision Trees, and Support Vector Machine (SVM). These algorithms are slightly discussed in the subsequent sections.

*1) **Naive Bayes Classifier**:* Naive Bayes is very simple but the popular algorithm to create a broad class of ML classifiers for diverse industrial applications. It is used to calculate the joint probabilities of values with their attributes (features) within the given set of cases. The attributes are considered independent of each other, and this consideration is known as naive assumption of conditional independence. The classifier makes this assumption while evaluating cases. The classification is made by taking into account the prior information and likelihood of incoming information to constitute posteriori probability model, which can be denoted by the following expression.

$$Posterior = (Prior * Likelihood)/Evidence \quad (1)$$

Listing 2 shows MLlib code for Naive Bayes classifier, where data is split into training (60%) and test (40%), and model is built and used for making predictions.

```
val splits = parsedData.randomSplit(Array(0.6, 0.4),
    seed = 11L)
val training = splits(0)
val test = splits(1)
val model = NaiveBayes.train(training, lambda = 1.0,
    modelType = "multinomial")
val predictionAndLabel = test.map(p => (model.
    predict(p.features), p.label))
val accuracy = 1.0 * predictionAndLabel.filter(x =>
    x._1 == x._2).count() / test.count()
```

Listing 2. A Snippet of MLlib Code for Naive Bayes

Examples of Construction Research using Naive Bayes Classifiers: Jiang et al. [34] presented a Bayesian probabilistic methodology for detecting the structural damages. Bayes factor evaluation metric is computed from Bayes theorem and Gaussian distribution assumption for accurate damage identification. The effectiveness of the proposed techniques is reported for assessing damage confidence of structures over five damaged scenarios of four-story buildings benchmark. Gong et al. [35] presented a framework for automated classification of actions of workers and heavy machinery in complex construction scenarios. They employed Bag-of-Video-Features-Model alongside the Bayesian probability for evaluating and tuning action discovery. It is revealed that the proposed approach is capable of identifying several actions in highly complex situations and is faster than the traditional methods. Huang et al. [36] studied the effect of severe loading events, namely earthquakes or long environmental degradation, on civil structures. A Bayesian probabilistic framework is proposed to compute the stiffness reduction. Using simulated data, the proposed approach is found to measure the stiffness accurately. The approaches as mentioned earlier are reportedly revealed as compute-intensive; hence require contemporary Big Data technologies for enhanced accuracy and response.

[TABLE 5 about here.]

*2) **Decision Trees**:* Decision trees (DTs) *is the modern ML method* for predicting about qualitative and quantitative target features. The process of building DT begins with identifying decision node and then recursively split nodes until no further *divisions* are possible. The robustness of DT depends on the logic for splitting nodes, which is assessed by using concepts such as information gain (IG) or entropy reduction. Listing 3 shows MLlib code to show DTs implementation; the data is split into training and testing sets, initialized parameters, created DT, and evaluated the model using data.

```
val splits = data.randomSplit(Array(0.7, 0.3))
val (trainData, testData) = (splits(0), splits(1))
val numClasses = 2
val categoricalFeaturesInfo = Map[Int, Int]()
val impurity = "gini"
val maxDepth = 5
val maxBins = 32
val model = DecisionTree.trainClassifier(trainData,
    numClasses, categoricalFeaturesInfo, impurity,
    maxDepth, maxBins)
val labelAndPreds = testData.map { point => val
    prediction = model.predict(point.features) (
    point.label, prediction)
```

Listing 3. A Snippet of MLlib Code for Decision Trees

Examples of Construction Research using Decision Trees: Pietrzyk et al. [66] studied the issue of mould germination in building structures using fault tree analysis. Structure related deficiencies that are introduced during the construction process are identified and classified. A probabilistic quantification model is generated to compare building structures based on their tendency for mould germination. Desai et al. [67] have employed decision trees to analyse and assess the labour productivity in the construction industry. The traditional decision tree algorithm is slightly customised to *suit* construction data, which is reported to have improved the accuracy of proposed methodology, with more realistic results are obtained.

*3) **Support Vector Machines (SVM)**:* SVM is a widely used technique that is remarkable for being practical and theoretically sound, simultaneously. SVM is rooted in the field of statistical learning theory, and is systematic: e.g., training *an SVM* has a unique solution (since it involves *optimisation* of a concave function). SVM uses kernel methods to map data from input/parametric space to higher level dimensional feature space. Listing 4 shows MLlib code to illustrate SVM, where algorithm builds *a model*, compute accuracy on test data, and evaluate *the model*.

```
val splits = data.randomSplit(Array(0.6, 0.4),
seed = 11L)
val train = splits(0).cache()
val test = splits(1)

val numIterations = 100
val model = SVMWithSGD.train(train, numIterations)

val scoreAndLabels = test.map { point =>
val score = model.predict(point.features) (score,
    point.label)}

val metrics = new BinaryClassificationMetrics(
    scoreAndLabels)
```

Listing 4. A Snippet of MLlib Code for SVM

Examples of Construction Research using SVM: *To* identify the damages in bridges, Liu et al. [68] employed SVM and genetic algorithms (GA). The selection, crossover, and mutation in GA *are* used for selecting best kernel parameters which are used in SVM as model parameters. A numerical simulation is presented to see the feasibility of the proposed approach. Comparative analysis of *GA-RBF (radical basis function)* and GA-BP (back propagation networks) is conducted, which reveals that the proposed technique has outsmarted these previously used approaches significantly for damage identification in bridges.

Mahfouz et al. [69] studied automated construction document classification using models, based on SVM and Latent Semantic Analysis (LSA). The classification accuracy of these models is compared and contrasted against the Gold standard of human agreement measures. Relatively better results are attained (with accuracy between 71% to 91%) than the previously used models. In another study [70], a construction legal decision support system is developed using SVM. SVM models extract legal factors from earlier cases to help the judges *to check the basis for their verdicts*. Results of first, second, and third-degree polynomial kernel SVM models are compared and contrasted. Highest accuracy is revealed for the first and second-degree polynomial SVM, of 76% and 85% respectively, implemented using TF-IDF. Similarly, SVM is used in fault detection system for HVAC under real working conditions [71]. The SVM classifiers for fault detection and isolation (FDI) are developed. The proposed approach can efficiently detect and isolate many typical HVAC faults.

*4) **Artificial Neural Networks (ANN)**:* Artificial Neural Networks (ANNs) algorithms are well suited to problems of classification or function estimation. Since their advent, these algorithms are widely used in solving complex industrial problems. Multi-layer perceptron (MLP) is the most commonly used type of ANN. ANNs are typically made up of three layers including an input layer, hidden (intermediate) layer, and output layer.

Data samples in MLP neural network are normalised and are fed into the input layer. This data moves from the input layer to one or two hidden layers and is finally passed onto the output layer, producing an output of the given ANN algorithm. Typically, x:y:z is used to describe the ANN topology in which x, y, z corresponds to the number of nodes in input, hidden, and output layers, respectively. During the training phase, the values of connections between nodes (a.k.a weights) are adjusted. Back propagation, simulated annealing and genetic algorithms are commonly used for training ANNs. Listing 5 shows MLlib code to explain lifecycle stages of ANN model development and evaluation.

```
val splits = data.randomSplit(Array(0.6, 0.4), seed = 1234L)
val train = splits(0)
val test = splits(1)

val layers = Array[Int](4, 5, 4, 3)

val trainer = new MultilayerPerceptronClassifier()
.setLayers(layers).setBlockSize(128)
.setSeed(1234L).setMaxIter(100)

val model = trainer.fit(train)
val result = model.transform(test)
val predictionAndLabels = result
.select("prediction", "label")
val evaluator = new
    MulticlassClassificationEvaluator().
    setMetricName("precision")
```

Listing 5. A Snippet of MLlib Code for ANN

Examples of Construction Research using ANN: Chen et al. [72] tailored ANN for fault detection of engineering structures, caused due to vibration and fatigue. The approach is reportedly revealed to yield better results in structural fault diagnosis. Fang et al. [74] employed ANN for structural damage detection. Back propagation algorithm, empowered by heuristics-based tunable steepest descent method, is used for training the neural network. Frequency response functions (FRF) are used for structural damage detection. A case study of cantilevered beam is analysed for unseen, single, and multiple damage types. Similarly, ANN is employed alongside GA in [73] for fault classification, in which ANN and GA complemented each other in reconstructing the missing input data. Moselhi et al. [75] deliberated the usefulness of ANN over the conventional expert-based systems, employed in developing various applications for the construction industry. A generic neural network based architecture is described, which is validated by implementing an application for optimal markup estimation. It is argued that ANN-based intelligent systems guarantee ideal performance over the systems, developed using conventional expert systems based approaches.

ANN algorithms have recently brought revolution in machine learning through deep learning. New algorithms of ANN are designed to learn from high dimensionality data (i.e., Big Data), which seek special attention in all the construction industry applications where ANN is employed.

*5) **Genetic Algorithms (GA)**:* Genetic Algorithms (GA) are evolutionary ML algorithms that are inspired by the natural evolution process. It computes better solutions to optimisation problems using the concepts such as inheritance, mutation, selection, and crossover. Typically GA algorithms involve creating two integral components, including (i) genetic representation (array of bits) of the problem, and (ii) a fitness function to evaluate solution domain. The process starts with initiating a solution randomly and then keeps improving it through iterative application of mutation, crossover, inversion and selection unless an optimal solution is found.

Examples of Construction Research using GA: Chen et al. [76] used GA to develop cost/schedule integrated planning system (CSIPS) which is focused on assigning crew optimally under complex set of constraints pertaining to resources and workforce. GA couple with BIM and object sequencing matrix is used to achieve feasible crew assignment in CSIPS system. Similarly, Moon et al. [77] developed an active BIM system for assessing the risks imposed by schedule and workspace conflicts that typically happens during the construction phase of a project. This active BIM system used fuzzy and GA algorithms for efficiently generating the optimal plan for workspace conflicts.

*6) **Latent Document Analysis (LDA)/ Latent Semantic Analysis (LSA)**:* LSA determines the meaning of words over a large corpus of documents using statistical techniques. It uses singular value decomposition method as its entire basis for computation. It is widely used in text analytics where it is used for vocabulary recognition, word categorization, sentence word priming, discourse comprehension, and essay quality assessment. LSA is based on the following measures.

```
val corpus = parsedData.zipWithIndex.map(_.swap).
    cache()
val ldaModel = new LDA().setK(3).run(corpus)
val topics = ldaModel.topicsMatrix
```

Listing 6. A Snippet of MLlib Code for Latent Semantic Analysis

1) *Precision*—is the fraction of retrieved documents, which are relevant. It is useful to assess the quality of LSA approaches.
2) *Recall*—is the fraction of the relevant documents, which are retrieved. Recall mostly informs about the completeness of LSA approaches.
3) *F-Measure*—is often used to combine precision and recall for assessing the accuracy of tests.

Listing 6 shows MLlib code to demonstrate the implementation of LDA, where a corpus is created, and documents are clustered based on word distribution.

Examples of Construction Research using LDA & LSA: Kandil et al. [79] employed LSA for automated construction document classification. The proposed technique classified two sets of documents: (1) documents with low word variations (claims and legal documents), and (2) documents with high word variations (correspondence and meeting minutes). The evaluation of proposed technique provided satisfactory classification results. Mahfouz et al. [69] employed a hybrid ML-based construction document classification methodology built on top of SVM and LSA. The presented results are relatively better than approaches based on a single ML technique. Salama et al. [78] employed LSA-based classifiers for this purpose where the documents clauses are automatically classified into predefined categories such as environmental, health, etc., before rule extraction. The developed method is reported to achieve 100% and 96% recall and precision respectively.

*7) **More Construction Industry Research using Classification:*** Classification algorithms have been used in construction for many tasks. In this subsection, we will discuss some of the important applications of classification for the construction industry. In particular, we will review document classification, document analysis, image-based classification, the classification for predicting project overrun, and finally, the classification for safety analysis. Pointedly, these applications need to be revamped with the Big Data technologies, since they present similar challenges of high dimensionality, velocity, and variety. Besides, these applications also involve classy computation while performing domain-specific tasks.

Document Classification: Different techniques are devised to classify automatically documents based on various classification systems such as CSI MasterFormat, CSI UniFormat, and UniClass. Caldas et al. [80] used SVM to organise construction documents based on the CSI MasterFormat classes. The relevance of documents with terms is calculated by Boolean weighting, absolute frequency, TF/IDF, and IFC weighting. The prototype system is evaluated and found very relevant. Rehman et al. [81] classified construction documents into two distinct groups of good and bad information-containing documents. Three layered ML approach is employed. Decision Trees (DT), Naive Bayes, SVM, and KNN algorithms are used to check the accuracy of classification. Except for the DT, the rest of algorithms have significantly improved the classification accuracy. Similarly, Liu et al.[82] presented the process for structured document retrieval for engineering based document management.

Document Analysis: Soibelman et al. [83] proposed a comprehensive platform to store and analyse unstructured documents used within a construction project. The system captures the essential attributes of these document types containing diverse data about text, web, image, and linking and stores it in an analytic-friendly format. These documents are then automatically linked to the appropriate binary files (building models) using different ML classifiers, which dramatically improved the information retrieval and significantly reduced overall searching time of project managers.

Image-Based Classification: Construction site photography logs comprise a significant chunk of construction documentation. A novel ML-based classification system is proposed in [84] which uses Whitening Transform (WT), SVM, and Biased Discriminant Transform (BDT) algorithms to classify and index construction site images. The proposed approach has significantly boosted the results of traditional search engines.

Predicting Overrun Potential: Williams et al. [85] analysed highway project bidding data for interested trends informing about project overruns. Data exploration revealed that bids with higher ratios tend to have significant cost overruns. Based on these ratios (as independent variables), an automated ML-based algorithm (Ripple Down Rules) is employed to classify the overrun potential of construction projects into following discrete values of *Near*, *Overrun*, *BigOverrun*. This exploration has revealed interesting rules for assessing the dilemma of project cost overruns. Similarly, Elfaki et al. [86] explored the whole breadth of intelligent systems developed using different ML algorithms for construction project cost estimation.

Safety Analysis: Han et al. [87] presented an approach that uses site videos to measure the workers' behaviour towards safety. The proposed approach analyse the 3D skeleton motion model of the workers to identify their actions. Since safe and unsafe actions are known, so the training data is correctly labelled for safe and unsafe actions, which is exploited by the classifier for learning. As a case study, the motion of worker while climbing the ladder is analysed. It is revealed that classifier can successfully identify the moves that can potentially lead to site injuries.

[TABLE 6 about here.]

C. Clustering Techniques

Clustering is used to find groups that have similarity in their characteristics. Intuitively, clustering is akin to unsupervised classification: while classification in supervised learning assumed the availability of a correctly labelled training set, the unsupervised task of clustering seeks to identify the structure of input data directly. Items in one cluster are similar to each other whereas different from the items of other clusters. Some of the examples of clustering algorithms include K-means, O-means, fuzzy K-means, and canopy. Listing 7 shows MLlib code for clustering data using K-Means and evaluating the model using Within *Set-Sum-of-Squared-Errors*.

```
val numClusters = 2
val numIterations = 20
val clusters = KMeans.train(parsedData, numClusters,
    numIterations)
val WSSSE = clusters.computeCost(parsedData)
```

Listing 7. A Snapshot of MLlib Code for K-Means

Examples of Construction Research using Clustering: Ng et al. [88] used clustering to group the facilities based on the deficiency descriptions stored in the facility condition assessment database. The results have shown that facility deficiencies are unique and always a function of location and type of the facility. Fan et al. [89] employed clustering for developing construction case retrieval system to identifying accidents occurred in the past. The goal is to resolve the disputes before provoking litigation and work interruptions. It is noticed that the NLP based approaches performed far better than case-based reasoning techniques, while measuring the similarity of case documents.

A hybrid approach is adopted in [90] to group construction project documents automatically. The approach initially uses clustering to generate classes for these documents based on textual similarity measures. Later on text classifier is used to classify relevant documents from the construction document information system. This hybrid approach has drastically improved the recall and F-measure. Clustering becomes non-trivial with massive datasets comprising millions of dimensions.

D. Natural Language Processing (NLP)

The NLP is concerned with creating computational models that resemble the linguistic abilities (reading, writing, listening, and speaking) of human beings. It provides basic concepts and methods for text processing and analysis, such as part of speech (POS) tagging, tokenization, sentence splitting, named entity recognition, and semantic role labelling, etc. This field brings together diverse techniques from computational linguistics, speech recognition, and speech synthesis to process human languages.

Examples of Construction Research using NLP: The NLP has a broad range of applications for knowledge acquisition and retrieval in the construction industry. Al-Qady et al. [91] used NLP to develop ontologies from construction contractual documents. They employed NLP-based Concept Relation Identification using Shallow Parsing (CRISP) for automatically extracting the concepts and concept relationships from the text of contract documents. The Kappa score and F-measure have significantly improved knowledge acquisition, while constructing legal ontology. The works in [92], [93], [94] proposed an NLP-based information extraction system for automated compliance checking from construction regulatory documents. A set of pattern-matching and conflict resolution rules has been developed that employ syntactic (syntax/grammar-related) and semantic (meaning/context-related) text features during NLP processing. A technique for tagging, separation, and sequencing of regulatory document elements is proposed to generate high-quality ontology. The proposed algorithm is tested on the regulatory documents, retrieved from the International Building Code and the results are promising with higher precision and recall.

E. Information Retrieval (IR)

Web search engines are the most common examples of IR systems, where information is typically organised as a collection of documents. IR systems deal mainly with unstructured textual data (that have no defined schemas). Besides, these systems can also handle complex, unstructured data such as images. Approximation and ranking are the vital attributes of the IR query languages. Queries are specified as search terms encapsulated in keywords and logical (AND & OR) connectives. These queries are evaluated with approximation based relevance ranking, where documents are identified and returned based on their relevance to a query.

Examples of Construction Research using IR: Demian et al. [95] developed CoMem-XML system to augment searching through granularity and context. The system is enhanced for contextual similarity, which is revealed to be of greater usefulness and usability to construction professionals. Tserng et al. [96] developed IR system called Knowledge Map Model System (KMMS) to facilitate construction professionals for managing and reusing construction knowledge from a variety of unstructured documents. Fan et al. [97] proposed a framework for managing unstructured construction project documents where terms dictionaries and dependency textual documents are used. The framework is evaluated, and its usefulness is revealed.

Hsu et al. [98] employed context-based text mining for 3D CAD documents exploration. Traditional systems depend on textual naming and require designers to memorise and embed these descriptions within the design documents. To this end, a context-based CAD document retrieval system (CCRS) is developed for extracting the context from CAD documents into the characteristic document (CD), which is exploited by query planner to select the documents. Lin et al. [99] studied the retrieval of technical documents like journal papers, patents, technical reports, or domain handbooks. A concept-based IR system is developed to illustrate the effectiveness of proposed partitioning approach. It is shown that the proposed approach is quite useful for concept-based IR of technical documents. Al-Qasy et al. [100] introduced an electronic document management system (EDMS) to manage construction project documents. At the crux of this system lies the proposed idea of *document discourse*, which determines the semantic similarity of documents. A classification algorithm, using document discourse, is implemented for classifying project documents. The system is evaluated by a group of experts.

IV. Opportunities

A. Resource and Waste Optimization

Rapid urbanisation has escalated construction activities globally, which triggered construction industry to consume the bulk of natural resources and produce massive construction and demolition (C&D) waste [101]. The adverse impact of construction activities on the environment has serious implications worldwide [102]. Existing waste management approaches are based on *Waste Intelligence (WI)*, which suggests remedial measures to manage waste only after it happens [103]. These systems mostly answer close-ended questions such as project/site wise waste generated, progress towards defined waste targets, and understanding how a particular design strategy produces waste [104]. The end users are provided hindsight with limited insight on waste minimisation.

However, data-driven decision-making at the design stage is revealed to bring a revolution for preventing a significant proportion of construction waste [105], [104]. This compels a paradigmatic shift from the static notion of WI to a more progressive idea of *Waste Analytics (WA)* [106]. Waste minimisation through design is the future of waste management research [101]. WA advocates proactive analyses of disaggregated and massive datasets to uncover non-obvious correlations related to design, procurement, materials, and supply-chain, which could lead to waste during the actual construction stage. It explores waste data in a forward-looking way [104], [106]. Advanced analytical approaches could be employed to forecast waste and prescribe the best course of actions to pre-emptively minimise waste.

However, WA depends increasingly on the high-performance computation and large-scale data storage. It requires a significant number of diverse data of building design, material properties, and construction strategies to successfully carry out the process. Storing these datasets, using traditional technologies, is not only insurmountable, but the real-time processing for underpinning high-dimensional analytical models is highly challenging. This calls for the application of Big Data technologies for effective construction waste management. Particularly, robust waste generation estimation models, BIM-based optimal materials selection during design specification, and holistic waste minimisation framework are key research areas which call for the applications of these Big Data technologies to be employed. Table XIV summarises the state of the art and potential opportunities for resource and waste optimisation. Some of these opportunities are further explained in Section V.

[TABLE 7 about here.]

B. Value Added Services

This section discusses a broad range of non-core services, which can be benefited from the emerging trend of Big Data in the construction industry.

*1) **Generative Design**:* Generative design (GD) is another paradigm shift in the construction industry. The idea is to generate many designs automatically based on the specified design objectives, such as functional requirements, material type, manufacturing method, performance criteria, and cost restriction, among others. The intended GD tools employ sophisticated algorithms to synthesise design space and generate a wide assortment of design solutions that meet the given design requirements. These designs are presented to designers for evaluation based on their performance. This evaluation enables the designers to reiterate designs by adjusting design goals and constraints unless a design is produced to their satisfaction. Advancements in this field can bring lots of benefits, particularly for resource optimisation and waste reduction through design.

Attempts are made to verify the adequacy of this idea. To this end, Autodesk has come up with the Dreamcatcher tool, to facilitate designers, for generating designs based on abstract design requirements. However, Dreamcatcher is still in its infancy and is far from being a promising tool to be used for professional purposes. Many challenges are underlying to achieve GD realistically. Particularly, the generation and exploration of design space is time-consuming and is massive. The tool has to generate and compare a permutation of models for single client requirement. This field requires more R&D for getting mature to be usable in the enterprise-grade applications. These challenges of GD tools are expressly the jurisdiction of using Big Data technologies. These technologies can undoubtedly bring new levels of usability, accessibility, and democratisation in the design exploration and optimisation in next generation GD tools. Table XIV summarises the state of the art and potential opportunities for this subdomain.

*2) **Clash Detection and Resolution**:* The identification of design clashes is an integral part of the building model. Ideally, this phase should be carried out before the start of construction stage for effective project management. Traditional paper-based approaches are widely substituted by BIM-enabled automated approaches, which are found relatively inefficient as well as less accurate to identify the majority of design conflicts. However, existing BIM-enabled conflict resolution solutions are still tedious and time-consuming for efficient process automation. There are two aspects of these systems. Firstly, adequate knowledge management is at the crux of these systems to achieve accuracy. Wang et al. [38] proposed a knowledge-based system for acquiring, formulating, and deploying knowledge in BIM-enabled MEP design coordination. However, much is required in this direction. Additionally, for the later, design conflicts identification requires non-trivial algorithms for design exploration, which are time-consuming. These aspects are the subject of Big Data technologies, which can augment knowledge representation as well as computation through its well-known distributed and parallel computational capabilities. Table XIV summarises the state of the art and potential opportunities for this subdomain.

*3) **Performance Prediction**:* Performance prediction models have been wide applicability in various domains of the construction industry. Particularly, these models are instrumental for pavement management systems, where system engineers are facilitated to take right decisions while constructing, maintaining, and rehabilitating the pavement structures. These models use a large number of variables and their great combinations, in which they influence each other as well as overall model performance, and are developed using simple statistical approach (like linear regression) to computational intelligence techniques (as ANN). Karagah et al. [109] evaluated various prediction models for predicting their accuracy for pavement deterioration trends. Their evaluation shows that these system involve computation-savvy analysis, which is time-consuming and hard for traditional technologies to process at a real time. Moreover, it is highlighted that high dimensionality is inherent to the dataset produced for these applications, where the extremely large number of variables contribute to the model development. To this end, performance prediction field offers opportunities to utilise Big Data technologies. Consequently, Big Data technologies are of immense relevance and can aid in the area regarding real-time computation, reliable model development, and enhanced visualisation. Table XIV summarises the state of the art and potential opportunities for this subdomain.

*4) **Visual Analytics**:* Analytical problems are of two kinds: (1) the problems that have clearly defined and logical solutions; and (2) the problems that have approximate heuristic solutions (and no logic-based straightforward solution applies). The former category is handled through automated approaches, whereas the later ones are tackled through visualisation. Human knowledge, creativity, and intuition are pivotal for effective visualisation. Human knowledge works perfectly with smaller datasets, but its application in involving high dimensional larger datasets becomes impractical. The field of Visual Analytics (VA) came into existence to combine automated reasoning and visualisation to solve complex analytical problems. Such systems are phenomenal to empower analytical abilities of users while perceiving, understanding, and reasoning about complex and uncertain situations. VA is one of the key domains that require Big Data technologies to execute data visualisation to provide personal views and interactive exploration of data.

One of the key reasons behind the widespread adoption of BIM lies in its versatile visualisation capabilities. Existing software are quite competitive to visualise all dimensions (nD) of the design using the right set of tools and techniques. In this context, Castronov et al. [110] studied the role of visualisation in 4D construction management. Shortcomings of existing BIM visualisation are identified, and general guidelines/ protocols are prescribed for developing 4D visualisation in BIM authoring tools. To enable participation of technically unskilled BIM users, Zhadanovsky et al. [154] studied the issue of generating master plan visualisation. Similarly to promote sustainable energy use, Goodwin et al. [111] employed VA for classifying energy users. The data of household energy consumption along with geo-demographic data is used for deeper insights. Classification is reported to enable clusters and trends for understanding energy usage. However, state-of-the-art approaches of visualisation are needed during clustering process and decision making to enable overall comprehension. Chuang et al. [112] studied the development of a cloud-enabled web-based system for BIM visualisation and manipulation. The system improved communication and distribution of relevant information among the stakeholders.

The scope of BIM is widening with more applications from construction as well FM stage has started utilising and extending it. As BIM data grows, these models get highly dimensional, so the visualisation of high-dimensional BIM models is challenging. VA is essential to both BIM and Big Data and provides sophisticated techniques to improve BIM and Big Data visualisation for better comprehension and interpretation. Table XIV summarises the state of the art and potential opportunities for this subdomain.

*5) **Social Networking Services/ Analytics**:* Majority of construction industry problems are communication-related [113]. Social media is another interesting trend that can help the industry to improve communication among the project team. This trend is slowly penetrating the industry. Social networking services to share updated project information along with wider practices for communicating the best practices of sustainability could be the next application areas.

Some studies have been carried out in these directions. Jiao et al. [113] studied the usage of social media to communicate project management data, including schedules, progress monitoring data, and work assignments. The proposed approach facilitates the integration of useful project data with BIM. Meadati et al. [114] studied the integration of RFID, BIM, and social media to support facility managers in locating data from multiple documents. Jiao et al. [115] brought the web3D-based AR environment for integration of BIM and business social networking services (BSNS) over the cloud-enabled platform. The goal is to enhance the overall comprehension of BIM models.

However, a robust framework is required to capture every useful social interaction into the BIM right from the design to end-of-life of the building. Since data of social interactions are likely to be in variety, velocity, and volume, Big Data technologies could be harnessed to develop interesting domain applications for enhancing the productivity of stakeholders. Table XIV summarises the state of the art and potential opportunities for this subdomain.

*6) **Personalized Services**:* In personalised services, the primary emphasis lies on an adaptation of the given facilities based on the user' choice. The users are empowered to control the overall usage of services the way they desire. These systems adapt based on various parameters such as user behaviour. The input to such services could be manual as well as automatic.

Gao et al. [116] developed SPOT+ system to enable office workers to personalise the indoor thermal comfort. SPOT+ used Predictive Personal Vote (PPV) to automatically adjust indoor thermal comfort that mainly involve heating. The system turns on the heating before the arrival of occupants whereas turns the heating off immediately after their departure. Rabbani et al. [117] proposed an enhanced personalised thermal comfort system called SPOT* that enables users to adjust lower and upper bounds of indoor temperature as desired, which is automatically regulated accordingly. SPOT* supports heating as well cooling of indoor spaces. The system has significant potential for energy reduction while maintaining the overall comfort at desired level. Panagopoulos et al. [118] proposed the AdaHeat system that uses intelligent agents to regulate the heating for domestic consumption. A novel aspect of this system is that it requires minimal user input. Chen et al. [119] studied the correlation of human behaviour and energy consumption in smart homes. Computational models to predict energy consumption based on user behaviour are developed. These models are used to develop a web-based system that provides user with insights based on behaviour for optimal energy consumption.

The applications to enable personalised services always require scanning the surrounding environment for the events of interest using sensing technologies, generating large volumes of data. Accumulating such streams of data and then processing it to generate actionable insights at real time for point-in-time adaptation is non-trivial and is the subject of interest for Big Data technologies. To this end, robust Big Data enabled platform is required that provides a unified interface to support the needs of diverse personalisation services, employed in modern buildings. Table XIV summarises the state of the art and potential opportunities for this subdomain.

C. Facility Management

Facilities management (FM) integrate organisational processes to maintain the agreed services that support and improve the effectiveness of its primary activities. Operations and management are the central parts of FM and are the longest stage in whole building lifecycle. Mostly FM activities (such as assets management, preventive maintenance, etc.) are laborious, and the efficiency of such tasks can improve by incorporating suitable supporting technology. Localization information is of great importance to these technology solutions. Today these facilities utilise advanced automation and integration to measure, monitor, control, and optimise building operations and maintenance. They provide adaptive, real-time control over an ever-expanding array of building activities in response to a wide range of internal and external data streams. As investment ramps up and more intelligent systems are brought online, more data will enter the energy management platform at faster speeds.

Taneha et al. [120] proposed an approach to determine the FM personal location information using localization technologies to support FM related activities. The system employs three technologies like RFID, Wireless LAN, and Inertial measurement units (IMUs) for this localization. To reduce the FM cost, Ng et al. [121] applied knowledge discovery and data mining over the facilities maintenance databases. Liu et al. [122] evaluated the capabilities of BIM to support the FM operations. A detailed needs of FM professionals are identified to harness BIM to support relevant tasks. The factors affecting the maintainability of facilities are mainly considered.

Motamedi et al. [155] highlighted three challenges faced by the majority of FM systems. These include (i) inefficient & time-consuming searching interfaces, (ii) no unified interface for FM system to exchange information, and (iii) inability to store and process large volumes of data generated by these systems. These challenges evidently call for the applications of Big Data technologies in the development of FM systems. Particularly, in the case of predictive maintenance, BDA can inform FM managers whenever equipment is likely to break or require an upgrade. Consequently, FM organisations could benefit from lowered operating expenses, higher profit margins and enhanced service availability. Table XIV summarises the state of the art and potential opportunities for this subdomain.

D. *Energy Management & Analytics*

Two type of energy software are prevalent. Firstly, building energy simulation software to model the energy consumption of buildings. Their accuracy depends on the accuracy of provided parameters that are fine-tuned by experts. This fine tweaking is laborious and time-consuming. Automatic fine tweaking involves lots of computations. Sanyal et al. [123] studied the automatic generation of accurate input model with proposed Autotune workflow for the EnergyPlus energy simulation software. Pointedly, it is informed that the software operates on raw data of about 270 terabytes, and condenses that to approximately 80 terabytes of useful data. Data storage, transfer, and processing such datasets is inevitably the subject of Big Data technologies.

Secondly, Building Energy Management Systems (BEMSs) are vital for buildings. And as part of their architecture, hundreds to thousands of sensors are installed to capture data. Linda et al. [124] used computational intelligence based anomaly detection to fuse data from multiple heterogeneous data sources and to process it for generating actionable insights. Despite BEMSs use the state-of-the-art multi-processor infrastructures, the issue of data management and processing is reported to have taxed the boundaries of these systems. Hong et al. [125] proposed a cloud-based storage system to store and process energy data generated from a network of thousands of Zigbee sensors. To persist this data, Singh et al. [126] proposed cloud-based storage and processing architecture. Berges et al. [127], [128] proposed novel approach for identifying appliances and their events (on/off or low/high) to measure their electric consumption precisely from the electric influx. It is reportedly revealed that proposed approach requires emerging data management and processing capabilities for real life deployment. Similarly, Goodwin et al. [111] employed visual analytics for energy users classification. It is highlighted that state-of-the-art approaches of visualisation are at the core of clustering process, decision making, and enhanced overall comprehension of energy consumption. Wei et al. [128] proposed an IOT-based framework to monitor and analyse the energy consumption of Smart Buildings.

The software as mentioned above perfectly presents the opportunities for Big Data analytics to advance the field. Pointedly, energy-related data is of immense importance for various analytics, which is usually discarded by building owners and utility companies at a time interval. To present this data nicely for advanced analytics is the next frontier of innovation in this field. Table XIV summarises the state of the art and potential opportunities for this subdomain.

E. *Other Emerging Trends that Triggered Big Data*

This section presents a few technologies that amplified the advent of Big Data in the construction industry. Their successful deployment to advance the industry is indeed the function of Big Data analytics.

*1) **Big Data with BIM**:* Building Information Modelling (BIM) is conceived to revolutionise construction industry in many aspects [156], [131]. BIM is empowered with an extra layer of data, captured throughout the whole building lifecycle [131], [132]. This data can be unleashed to develop useful applications for improving the overall building delivery process. Theoretically, BIM is declared as the de facto standard for managing building data, its applications, in practice, across every lifecycle stages of building are yet to develop, however. Preconstruction stages are well-known for widely adopting the BIM, whereas, it is progressively used lesser in the later stages of building lifecycle [5]. Substantial research is made to extend BIM for encapsulating different types of related data.

Goedert et al. [133] extended BIM for construction process documentation. Chiang et al. [157] integrated power consumption data with BIM models. Isikdag et al. [134] integrated geographic information systems (GIS) data with BIM for developing a fire response system. Yeh et al. [158] employed BIM for onsite building information retrieval using augmented reality. Wang et al. [38] extended BIM for spatial conflict data for MEP models. Yu et al. [135] integrated BIMserver with OpenStudion (a platform for assessing the energy efficiency of building designs). Das et al. [20] tailored BIM for social interactions taking place while reviewing and commenting on different aspects of the design. Zheng et al. [136] integrated BIM with diverse project data sources. Chaung et al. [112] exploited BIM for cloud-enabled design exploration and manipulation. Jiao et al. [113] sorted out the issues of integrating BIM with project schedules, progress monitoring data, and work assignments. Meadit et al. [114] integrated RFID data into BIM to locate project documents. Volk et al. [129] illustrated the automated creation of BIM models for existing buildings.

These illustrate the gradual increase in the size and scope of the contents of BIM models, which eventually restricts the capabilities of traditional BIM-based storage and processing systems. To tackle this, Jiao et al. [6] tailored MapReduce for storage and processing BIM. However, there are still many use cases which may require sophisticated customizations to the way BIM is stored and processed. So in future, we are expecting BIM specialised Big Data storage and processing platforms. Up until recently, BIM is envisaged to contain data of construction industry only; however the emergence of linked building data has changed this perception. Despite linking BIM data to inter-industry applications, many interesting applications can be developed by enabling the integration of BIM with Linked Open Data (LOD) datasets, such as weather, flooding, population densities, road congestions, and so on [147]. Such integration of BIM is undoubtedly resulting in Big BIM data, which justifies the emergence of Big Data in the specialised area of BIM. Table XIV summarises the state of the art and potential opportunities for this subdomain.

*2) **Big Data with Cloud Computing**:* Cloud computing is Internet computing paradigm in which on-demand access to a shared pool of configurable resources is provided [159]. The idea is to outsource data storage and computation to third-party datacentres. Multiple users can simultaneously access the cloud services without having to purchase individual licenses. Cloud computing offers three service models. (i) *Infrastructure-as-a-service (IaaS)*: In IaaS, the user is provided with an abstraction to manage virtual/physical computers and cloud network services. (ii) *Platform-as-a-service (PaaS):* In PaaS, a user is provided with services pertaining to development environments such as operating systems, programming languages, or databases, among others; (iii) *Software-as-a-service (SaaS):* In SaaS, the user is provided access to enterprise applications via the internet such as Revit 360.

Cloud computing is widely adopted in the construction industry since it supports the integration of tasks in BIM-based applications. Hong et al. [125] utilised cloud computing for building energy management systems using Zigbee sensors. Das et al. [20] proposed a cloud-based BIM framework for integrating stakeholders interactions with BIM. Zhang et al. [136] utilised private clouds to offer BIM services across the whole building lifecycle. Klinc et al. [139] proposed SaaS platform for the structural analysis applications. Kumar et al. [140] employed cloud for SMEs design and construction firms. Chuang et al. [112] used cloud computing for BIM design exploration and manipulation. Redmond et al. [160] employed cloud for interoperability between BIM applications. Amarnath et al. [161] deployed Revit Server on the cloud for collaboration and coordination of architectural and structural models. Rawai et al. [162] explored cloud computing for green and sustainable developments. Fathi et al. [142] used the cloud for BIM-based context-aware computing. Beach et al. [143] discussed the issues of enabling Google SketchUp over the Amazon EC2 cloud. Chong et al. [144] evaluated existing cloud computing applications and highlighted Google Apps, Autodesk BIM 360, and Viewpoint, among others, support majority of designers features on the cloud. Grilo et al. [145] used the cloud for creating e-procurement platform—Cloud Marketplaces. Jiao et al. [113] exploited cloud framework for integrating project management data with building models. Jiao et al. [115] integrated cloud computing with latest technologies such as AR and business social networking services to create virtual environment to visualise better and understand BIM models. Wong et al. [137] highlighted the legal issues related to cloud-based BIM models, including security, responsibility, liability, and design ownership.

Cloud computing has already accelerated the uptake of IT in the construction industry by transforming many domain specific applications as discussed above. And the role of Big Data in this transformation is overwhelming. Table XIV summarises the state of the art and potential opportunities for this subdomain.

*3) **Big Data with Internet of Things (IOT)**:* An exciting fact about Internet is it keeps evolving since its perception. It started with *Internet-of-Computers* and had evolved into *Internet-of-People*, and is recently facing new paradigm shift. With fast emerging technologies, the devices are getting smaller and powerful, and the broadband connectivity is getting cheaper and ubiquitous. This has led to the proliferation of connected devices on the Internet, eventually resulted in an exciting trend coined as the Internet-of-Things (IOT) [150]. The primary vision behind IOT is to bring together the smart devices and objects the vital parts of Internet. Fusing these exciting physical and digital worlds are creating fascinating opportunities of growth. Some of the popular areas where IOT applications are successfully demonstrated across the industries include logistics, transport, assets tracking, smart homes, smart buildings, to energy, defence and agriculture.

Elghamrawy et al. [146] demonstrated RFID usage for construction monitoring and quality control. Meadati et al. [114] integrated RFID with 3D BIM documents of assets for searching and locating objects quickly. Wei et al. [128] proposed an IOT-based framework for building energy monitoring. Zanella et al. [148] presented specifications of urban IOT to envision the idea of Smart Cities. Kortuem et al. [163] discussed the technical specifications of the smart object for petrochemical and road construction industries. Curry et al. [147] examined the storage and processing of energy sensors data using cloud-based data management framework.

The applications of IOT are non-trivial and often deploy hundreds or even thousands of sensor devices for data collection. Since construction industry presents unlimited use cases for IOT, Big Data is inherently the subject of interest. IOT and Big Data are complementary trends, with former to generate large volumes of data and the later to store and analyse these data at the real time in construction specific domain applications. Table XIV summarises the state of the art and potential opportunities for this subdomain.

*4) **Big Data for Smart Buildings**:* Buildings evolved considerably over time. While providing comfort and security, buildings cause adverse environmental impact by consuming energy and producing lots of greenhouse gases [159]. Smart building technology is a paradigm shift to embrace the integration of contemporary technologies with the prevailing building systems for striking the trade-off between the comfort maximisation and energy minimisation [149]. Building systems such as building automation, life safety, telecommunications, user systems, facility management systems, among others, provide actionable insights about different aspects of building and allows the users to control their interactions with building services better. The smart building incorporates technologies into building systems through a unified view. Often, these systems generate vast amounts of data and majority of this data remain untapped and often discarded. To truly realise smart buildings, this data of unprecedented size need to be analysed—a task that presents significant data management and processing issues. To this end, Big Data analytics is of immense importance to optimise total building performance via predictive analytics.

McKinsey [159] highlighted smart buildings amongst the top ten emerging technology businesses. Azam et al. [149] implemented a prototype software Project Dasher to illustrate Smart Buildings. Data of sensors related to motion, CO_2, temperature, airflow, lighting, and other acoustics properties are gathered and analysed. It is reportedly revealed that more than 2 billion data entries are accumulated in 3 months that reached the limits of legacy relational databases. Stankovic et al. [150] developed sensor based fire-fighting systems for skyscraper office building with the authorities to detect fires, alter fire situations, and aid in evacuation. Bonino et al. [151] studied complex event processing in smart buildings. *spChain* framework is proposed to support the real-time processing of sensor data. Miller et al. [152] analysed significant energy data through proposed DayFilter approach to precisely identify the diurnal patterns from the data.

Despite the fact that sophisticated IT systems are currently being used for controlling various building operations via sensors with enhanced data collection and analysis capabilities. However, these systems are still a long way off the actual vision of smart building apps that empower the end user in understanding and controlling their interactions with the building systems and spaces [164]. This discrepancy is due to the following reasons: (i) the services and functionalities currently being offered are quite rigid; (ii) the services are isolated and robust solutions for vertical and horizontal integration are not as yet available; and (iii) the supporting apps and APIs are often proprietary and lack standardization in many cases. For these reasons, these APIs can only be exploited by the BMS software itself, and are not amenable to the third-party development of applications, which restricts innovation at scale. In the future, Big Data Analytics based standard buildings APIs can bridge this technology gap and enable integration of sensors, users, control systems, machinery for providing innovative smart building services that promise comfort, safety, and energy. Table XIV summarises the potential opportunities for research on the applications of Big Data in Smart Buildings.

*5) **Big Data with Augmented Reality (AR)**:* Augmented reality (AR), which is an offshoot of virtual reality, is the field in which computer-generated virtual objects are superimposed over real-world scenes to produce *mix worlds*. It enables a semi-immersive environment that accurately aligns real scenes with corresponding virtual world imagery. This mixed overlay enables the users to obtain additional information about the real world. It is an emerging technology for enhancing human perception.

Rankohi et al. [165] argued that visualisation and simulation aspects of the construction industry apps can be revamped with AR to enhance their usability. Some of the exciting AR application areas are highlighted such as virtual site visits, proactive schedule dispute identification and resolution, and as-planned vs. as-built comparison. Chi et al. [166] pointed out the following four pillars for wider AR adoption in the construction industry. (i) *Localization*, the ability to accurately impose virtual object on the real-life scene. (ii) *A natural user interface*, which provides easy and intuitive user experiences to increase the usability of AP software. (iii) *Cloud computing*, which enables apps to store and retrieve information seamlessly everywhere, and (iv) *mobile devices*, which are getting smaller, cheaper, and powerful and play a vital role in AR environment. William et al. [153] went ahead by bringing BIM, mobile technology and AR together. The BIM aspects of geometry translation, indoor localization, attribute assignment, and registration are explored for integration with mobile AR. The study proposed BIM2MAR, which provides general guidelines for integrating BIM with mobile AR. It is emphasised robust BIM integration requires new approaches for BIM geometry conversion and indoor localisation of BIM using geo-coordinates. Jiao et al. [115] developed a web3D-based AR environment to integrate BIM, business social networking services (BSNS), and cloud services.

AR and Big Data inevitably converge. The complexity associated with Big Data in construction is enormous, which can only be surmounted by advanced methods of visualisation, particularly Augment and Virtual reality technologies. This requires new interactive platforms and methodologies to visualise construction related datasets. The aim is to comprehend better and interpret the complicated structures and interconnection buried inside the Big BIM Data for design exploration and optimisation. Table XIV summarises of progress and potential opportunities for AR in the construction industry.

V. Open Research Issues and Future Work

There are many interesting open research issues within the construction industry for Big Data. Some of these include (but are not limited to) the following:

A. Construction Waste Simulation Tool:

Construction waste minimisation is the perennial issue of the construction industry. Estimating construction waste accurately, at the early stages of design or as the project proceeds, is core to so many exciting project activities. Particularly, waste estimation is preliminary to waste minimisation at the early stages of design, where it provides insights about how the design is generating waste. These insights enable the designers to explore further and carry out corrective measures proactively, for waste efficiency at the early stages of design. So, construction waste estimation has become the key research question in construction waste management research. This estimation requires thorough design exploration and optimisation from a myriad of dimensions. Existing waste estimation models are based on very limited, and static project attributes such as GFA, project contract sum, etc. [107], [108], [167], [168]. However, these attributes are incapable of informing about the true size of construction waste, hence unable to generate a reliable waste estimate, regardless of how much data is used during their model development. A comprehensive waste estimation model that considers dynamic project attributes of deconstruction, standardisation and dimension coordination, reuse and recycling, and procurement, among together, needs to be developed. The model is also required to consider many attributes of construction materials, which heralds the development of a comprehensive materials database using open and linked data standards. The waste estimation model and construction materials database will be bundled into a standard and handy simulation tool, where waste estimates are visualised onto design elements through analytical dashboard alongside necessary prescriptions to minimise it through alternative materials or better design strategies. This tool presents a rich application of BDA in construction waste minimisation to backstage its storage and computation related workloads.

B. BDA enabled Linked Building Data Platform:

Existing interoperability efforts in the *construction* industry are mainly concerned about exchanging the building data between domain-specific applications (architectural, structural, MEP, energy simulation, etc.) pertaining to the *construction* industry. However, many interesting use cases can be achieved from greater integration of BIM data with external data sources such as materials, GIS, sensors, geodata, etc. This interoperability, at a wider scale, enables the *construction* industry to achieve automation of its business processes, which can improve the overall efficiency of the project participants. Linked data coupled with the Web of data technologies are found phenomenal for this integration. Substantial progress is made to develop various enabling artefacts for this integration such as ifcOWL ontology [169], [170]. However, much has yet to be done. To this end, the development of a robust BDA enabled platform that supports the storage and processing of these diverse linked data sets pertaining to the building as well as other data, is required. This platform can provide the basis for the development of interesting applications, particularly for energy analytics and smart buildings.

C. Big Data driven BIM System for Construction Progress Monitoring:

Currently, BIM is prevalent in the design world, with very limited utilisation across the construction and FM stages of the building. The real intent of BIM could never be achieved until it is employed in every stage of the building lifecycle. At present, no such mechanism can facilitate the tracking of progress of various construction sites using automated tools. It is indeed labour-intensive as well impractical (to some extent) to update the BIM model with such minute details pertaining to the daily construction progress. As a result, real-time construction progress monitoring is not an easy task, because managers are required to visit their sites regularly and assess the progress subjectively with the intended schedule, which is less effective and error-prone. Employing Big Data and sensing technologies could move the state of the art in domain of construction progress monitoring to the next level. Using latest imaging technology, the progress of the on-going construction is captured at the real time. Big Data analytics will process the real-time streams of these images to measure the daily change and updated the BIM models and construction schedule accordingly. The project managers are presented with an update to date progress on the schedule, which will, in turn, enable them to see whether they are lagging behind on the project or still follow the schedule. Accordingly, the project managers can proactively respond in case of any delay is reported. This will save them a lot of money due to penalty whenever the deadline is missed, and improve the overall project monitoring and control. This is also aligned with the vision of BIM adoption. In this way, Big Data can help the industry to deliver the projects on time.

D. Big Data for Design with Data:

Currently, designs are produced solely based on the client requirements and the designers experience. Thus, such designs that suit wider needs of the users, as well as the surrounding environment, are rare. For example, designers rarely consider the data collected by manufacturers on hundred or thousands of their product lines during design specification, which might be quite valuable. Similarly, many other sources of data can be relevant to designs such as users' sentiments while interacting with facilities, weather, flooding, energy consumption, commute pattern in that vicinity, and population densities, to name a few. These datasets could be harnessed to support for example the generation of an optimal construction schedule. And the good thing is that these data are captured using technologies such as the web, sensors, smart meters, mobile phones, etc., and are made available through open data initiative (in most cases). However, the design world is still detached from harnessing these data sources for their purpose. Currently, there is no such tool that can facilitate the designers to leverage these data during their design activities. If this is achieved, this can result in the paradigm shift of Design with Data, where these diverse data sources are integrated within the BIM authoring tools and made available to architects, engineers, contractors, and facility managers at early design stages. Big Data Analytics is indeed the key to this frontier of innovation. This symbiotic integration of diverse data sources with BIM will ultimately lead to the generation of next generation designs that can meet the wider requirements of sustainability, users, environment, and even broader infrastructures of emerging concept of smart cities.

VI. Pitfalls of Big Data in Construction Industry

Despite the opportunities and benefits accruable from Big Data in this industry, some challenging issues remain of concern. This section discusses some of these challenges and provides suggestions to deal with them for the successful implementation and dissemination of Big Data technologies across various domain applications of the construction industry.

A. Data Security, Privacy and Protection:

Prominent among these concerns is the issue of data security, data ownership, and management issues. To scale the hurdles posed by these challenges, several research studies have proposed and implemented security measures such as access control, intrusion prevention, Denial of Service (DoS) prevention, etc. [171], [172], [173]. These issues also require more study in the context of BIM-related construction data, and the appropriate solutions also need to be adopted in the underlying analytics workflows.

B. Data Quality of Construction Industry Datasets:

The construction industry is well-known for fragmented data management practices. Despite the aggressive promotion of BIM, companies using BIM are rare. Null values, misleading values, outliers, non-standardised values, among others, are some of the essential traits of industry data. And producing high-valued analytics is challenging due to poor data management practices. High-quality data is preliminary for successful Big Data projects. It is observed that analytics projects usually require approximately 80% of time cleaning noisy datasets before embarking on analytics. So, Big Data projects in construction industry shall also be specially taken care of, for data quality related issues. Otherwise, the resulting insights are likely to mislead, which in turn will result in unpleasant and pessimistic feeling in the industry. Consequently, the industry will be reluctant towards adopting such fascinating trends like Big Data.

C. Cost Implications for Big Data in Construction Industry:

Every technology incurs cost so introducing Big Data in construction is not for free of charge. Companies are required to set up data centres and purchase software licenses, which can be an attractive investment. Also, skilled IT personnel to keep the entire ecosystem running is another overhead. So Big Data has inevitably substantial cost implication. The construction business is considered amongst the low-profit-margin businesses, and introducing such costly add-ons to projects are more likely to be opposed and difficult to be defended. However, Big Data has the potential to enhance the overall project delivery by optimising processes and reducing risks that companies usually bear due to myriad inefficiencies such as delays, litigations, etc. It is highly optimistic that construction industry can gain huge revenue from this investment as experienced by other industries, provided the right methodology is used to employ Big Data. The exact cost implication of Big Data is, however, difficult to quantify. More studies on cost-benefit analysis of using Big Data technologies in construction projects are required.

D. Internet Connectivity for Big Data Applications:

To monitor project site activities at real-time, instant data transmission between project sites (dams, highways, etc.) and centralised Big Data repository should be supported. However, project sites usually have low bandwidth; due to unavailability of sophisticated networking infrastructure in rural, underdeveloped areas. Advanced wireless sensor networks need to be extended to tackle internet connectivity issues in these types of Big Data applications; otherwise, the decisions on stale offline data will not be useful for effective monitoring.

E. Exploiting Big Data to its Full Potential:

The effectiveness of Big Data cannot be measured just by accumulating large volumes of data; it is more of the use cases or industrial problems that dictate the usefulness of these technologies. It is feared that the construction industry might not extract the full value of accessible Big BIM Data if the conceived use cases are vague. To this end, researchers or domain experts are required to highlight domain-specific problems that are the subject of Big Data. This way Big Data as a technology will not be the driving force rather the industry itself will lead the innovation by applying contemporary tools to solve its topical issues. Additionally, Big Data is not the silver bullet, it merely sets the stage. Skilled professionals and domain experts, empowered with sophisticated analytical workflows, are equally necessary to reap the overall benefits. Without whom, the applications are likely to get into the pitfall of producing too much information that should not be delivering significant insights for the purpose.

VII. Conclusions

Although the construction industry generates massive amounts of data throughout the life cycle of a building, the adoption of Big Data technology in this sector lags the progress made in other fields. With the commoditization of the technology necessary for storing, computing, processing, analysing, and visualising Big Data, there is immense interest in leveraging such technologies for improving the efficiency of construction processes. In this exploratory study, we have analysed the extent to which the industry has employed Big Data technologies. Towards this end, we have reviewed not only the latest research but also relevant research articles that have been published over the last few decades in which the precursor to modern Big Data Analytics techniques have been deployed in various domain-specific construction applications. Principal Big Data technology streams are explained to help readers to understand the complicated subject. Concepts of Big Data Engineering and Big Data Analytics are demarcated; the works utilising these technologies across various subdomains of the construction industry are deliberated.

Through our research, we conclude that while data-driven analytics have long been used in the construction industry due to the broad applicability of such techniques in many construction subdomains, the adoption of the recent, much agiler and powerful, Big Data technology has been relatively slow. Although Big Data trend is gradually creeping in the industry; its applicability is amplified further by many other emerging trends such as BIM, IOT, cloud computing, smart buildings, and augmented reality, which are also slightly elaborated. We presented some of the prominent future works along with potential pitfalls associated with Big Data while adopting it in the industry. To the best of our knowledge, this is the first in-depth review of the applications of Big Data related techniques in the construction industry. In our work, we have identified many potential application areas in which Big Data techniques can significantly advance the state-of-the-art in the construction industry. This work is of utility and relevance to all the construction researchers and practitioners who will like to harness the power of Big Data in the construction industry for developing exciting business applications.

References

[1] V. Mayer-Schönberger and K. Cukier, *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Eamon Dolan/Houghton Mifflin Harcourt, 2013.

[2] E. Siegel, *Predictive Analytics: The Power to Predict who Will Click, Buy, Lie, Or Die*. John Wiley & Sons, 2013.

[3] "Blog post by Anand Rajaraman on how 'More Data Usually Beats Better Algorithms' (2013) [Online]." http://anand.typepad.com/datawocky/2008/03/more-data-usual.html. Accessed: 2013-09-12.

[4] C. Anderson, "The end of theory," *Wired Magazine*, vol. 16, 2008.

[5] R. Eadie, M. Browne, H. Odeyinka, C. McKeown, and S. McNiff, "BIM implementation throughout the UK construction project lifecycle: An analysis," *Automation in Construction*, vol. 36, pp. 145–151, 2013.

[6] Y. Jiao, S. Zhang, Y. Li, Y. Wang, B. Yang, and L. Wang, "An augmented Mapreduce framework for building information modeling applications," in *Computer Supported Cooperative Work in Design (CSCWD), Proceedings of the 2014 IEEE 18th International Conference on*, pp. 283–288, IEEE, 2014.

[7] J.-R. Lin, Z.-Z. Hu, J.-P. Zhang, and F.-Q. Yu, "A natural-language-based approach to intelligent data retrieval and representation for cloud BIM," *Computer-Aided Civil and Infrastructure Engineering*, 2015.

[8] G. Aouad, M. Kagioglou, R. Cooper, J. Hinks, and M. Sexton, "Technology management of it in construction: a driver or an enabler?," *Logistics Information Management*, vol. 12, no. 1/2, pp. 130–137, 1999.

[9] F. Provost and T. Fawcett, *Data Science for Business: What you need to know about data mining and data-analytic thinking*. " O'Reilly Media, Inc.", 2013.

[10] F. T. Matsunaga, J. D. Brancher, and R. M. Busto, "Data mining techniques and tasks for multidisciplinary applications: a systematic review," *Revista Eletrônica Argentina-Brasil de Tecnologias da Informação e da Comunicação*, vol. 1, no. 2, 2015.

[11] D. Singh and C. K. Reddy, "A survey on platforms for big data analytics," *Journal of Big Data*, vol. 29, no. 9, 2015.

[12] J. Qadir, N. Ahad, E. Mushtaq, and M. Bilal, "SDNs, clouds, and big data: New opportunities," in *Frontiers of Information Technology (FIT), 2014 12th International Conference on*, pp. 28–33, IEEE, 2014.

[13] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[14] V. S. Agneeswaran, *Big Data Analytics Beyond Hadoop: Real-Time Applications with Storm, Spark, and More Hadoop Alternatives*. FT Press, 2014.

[15] H. Karau, A. Konwinski, P. Wendell, and M. Zaharia, *Learning Spark: Lightning-Fast Big Data Analysis*. " O'Reilly Media, Inc.", 2015.

[16] C.-Y. Chang and M.-D. Tsai, "Knowledge-based navigation system for building health diagnosis," *Advanced Engineering Informatics*, vol. 27, no. 2, pp. 246–260, 2013.

[17] T. White, *Hadoop: The definitive guide*. " O'Reilly Media, Inc.", 2012.

[18] P. Helland, "If you have too much data, then'good enough'is good enough," *Communications of the ACM*, vol. 54, no. 6, pp. 40–47, 2011.

[19] M. Gelbmann, "Graph DBMSs are gaining in popularity faster than any other database category," 2014.

[20] M. Das, J. C. Cheng, and S. S. Kumar, "BIMCloud: A distributed cloud-based social BIM framework for project collaboration," in *the 15th International Conference on Computing in Civil and Building Engineering (ICCCBE 2014), Florida, United States*, 2014.

[21] S. Jeong, J. Byun, D. Kim, H. Sohn, I. H. Bae, and K. H. Law, "A data management infrastructure for bridge monitoring," in *SPIE Smart Structures and Materials+ Nondestructive Evaluation and Health Monitoring*, pp. 94350P–94350P, International Society for Optics and Photonics, 2015.

[22] J. C. Cheng and M. Das, "A cloud computing approach to partial exchange of BIM models," in *Proc. 30th CIB W78 International Conference*, pp. 9–12, 2013.

[23] L. Goldman, "The origins of British social science: Political economy, natural science and statistics, 1830–1835," *The Historical Journal*, vol. 26, no. 03, pp. 587–616, 1983.

[24] R. O. Duda, P. E. Hart, *et al.*, *Pattern classification and scene analysis*, vol. 3. Wiley New York, 1973.

[25] S. Owen, R. Anil, T. Dunning, and E. Friedman, *Mahout in action*. Manning Shelter Island, 2011.

[26] R. Ihaka and R. Gentleman, "R: a language for data analysis and graphics," *Journal of computational and graphical statistics*, vol. 5, no. 3, pp. 299–314, 1996.

[27] R. Yadav, *Spark Cookbook*. Packt Publishing Ltd, 2015.

[28] J. Dean, *Big data, data mining, and machine learning: value creation for business leaders and practitioners*. John Wiley and Sons, 2014.

[29] L. Wasserman, *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[30] D. S. Moore and G. P. McCabe, *Introduction to the Practice of Statistics*. WH Freeman/Times Books/Henry Holt & Co, 1989.

[31] H. Kim, L. Soibelman, and F. Grobler, "Factor selection for delay analysis using knowledge discovery in databases," *Automation in Construction*, vol. 17, no. 5, pp. 550–560, 2008.

[32] P. Carrillo, J. Harding, and A. Choudhary, "Knowledge discovery from post-project reviews," *Construction Management and Economics*, vol. 29, no. 7, pp. 713–723, 2011.

[33] T. S. Mahfouz, *Construction legal support for differing site conditions (DSC) through statistical modeling and machine learning (ML)*. PhD thesis, Iowa State University, 2009.

[34] X. Jiang and S. Mahadevan, "Bayesian probabilistic inference for nonparametric damage detection of structures," *Journal of engineering mechanics*, vol. 134, no. 10, pp. 820–831, 2008.

[35] J. Gong, C. H. Caldas, and C. Gordon, "Learning and classifying actions of construction workers and equipment using bag-of-video-feature-words and Bayesian network models," *Advanced Engineering Informatics*, vol. 25, no. 4, pp. 771–782, 2011.

[36] Y. Huang and J. L. Beck, "Novel sparse Bayesian learning for structural health monitoring using incomplete modal data," in *Proceedings of the 2013 ASCE International Workshop on Computing in Civil Engineering, Los Angeles*, 2013.

[37] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to

knowledge discovery in databases," *AI magazine*, vol. 17, no. 3, p. 37, 1996.

[38] L. Wang and F. Leite, "Knowledge discovery of spatial conflict resolution philosophies in BIM-enabled mep design coordination using data mining techniques: a proof-of-concept," in *Proceedings of the ASCE International Workshop on Computing in Civil Engineering (WCCE13)*, pp. 419–426, 2013.

[39] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons, 2011.

[40] M. J. Berry and G. S. Linoff, *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons, 2004.

[41] T. Rujirayanyong and J. J. Shi, "A project-oriented data warehouse for construction," *Automation in Construction*, vol. 15, no. 6, pp. 800–807, 2006.

[42] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 26, no. 1, pp. 97–107, 2014.

[43] R. Buchheit, J. Garrett Jr, S. Lee, and R. Brahme, "A knowledge discovery case study for the intelligent workplace," *Proc., Computing in Civil Engineering*, pp. 914–921, 2000.

[44] L. Soibelman and H. Kim, "Data preparation process for construction knowledge generation through knowledge discovery in databases," *Journal of Computing in Civil Engineering*, 2002.

[45] C.-W. Liao and Y.-H. Perng, "Data mining for occupational injuries in the taiwan construction industry," *Safety Science*, vol. 46, no. 7, pp. 1091–1102, 2008.

[46] C.-W. Cheng, S.-S. Leu, Y.-M. Cheng, T.-C. Wu, and C.-C. Lin, "Applying data mining techniques to explore factors contributing to occupational injuries in taiwan's construction industry," *Accident Analysis & Prevention*, vol. 48, pp. 214–222, 2012.

[47] S.-W. Oh, M.-H. Kim, and Y.-S. Kim, "The application of data warehouse for developing construction productivity management system," *Korean Journal of Construction Engineering and Management*, vol. 7, no. 2, pp. 127–137, 2006.

[48] D. Koonce, L. Huang, and R. Judd, "EQL an express query language," *Computers & industrial engineering*, vol. 35, no. 1, pp. 271–274, 1998.

[49] W. Mazairac and J. Beetz, "BIMQL–an open query language for building information models," *Advanced Engineering Informatics*, vol. 27, no. 4, pp. 444–456, 2013.

[50] L. Soibelman, L. Y. Liu, and J. Wu, "Data fusion and modeling for construction management knowledge discovery," in *International Conference on Computing in Civil and Building Engineering, Weimar, Germany*, 2004.

[51] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[52] J. E. Diekmann and T. A. Kruppenbacher, "Claims analysis and computer reasoning," *Journal of construction engineering and management*, vol. 110, no. 4, pp. 391–408, 1984.

[53] D. Arditi, F. E. Oksay, and O. B. Tokdemir, "Predicting the outcome of construction litigation using neural networks," *Computer-Aided Civil and Infrastructure Engineering*, vol. 13, no. 2, pp. 75–81, 1998.

[54] M. P. Kim, "Us army corps engineers construction contract claims guidance system," in *Excellence in the Constructed Project*, pp. 203–209, ASCE, 1989.

[55] K. Chau, "Predicting construction litigation outcome using particle swarm optimization," in *Innovations in Applied Artificial Intelligence*, pp. 571–578, Springer, 2005.

[56] O. B. Tokdemir, "Using case-based reasoning to predict the outcome of construction litigation," *Computer-Aided Civil and Infrastructure Engineering*, vol. 14, no. 6, pp. 385–393, 1999.

[57] K. Chau, "Prediction of construction litigation outcome–a case-based reasoning approach," in *Advances in Applied Artificial Intelligence*, pp. 548–553, Springer, 2006.

[58] D. Arditi and T. Pulket, "Predicting the outcome of construction litigation using boosted decision trees," *Journal of Computing in Civil Engineering*, vol. 19, no. 4, pp. 387–393, 2005.

[59] J.-H. Chen and S. Hsu, "Hybrid ann-cbr model for disputed change orders in construction projects," *Automation in Construction*, vol. 17, no. 1, pp. 56–64, 2007.

[60] M. Siu, R. Ekyalimpa, M. Lu, and S. Abourizk, "Applying regression analysis to predict and classify construction cycle time," *Proc., Computing in Civil Engineering (2013)*.

[61] A. Aibinu and G. Jagboro, "The effects of construction delays on project delivery in nigerian construction industry," *International journal of project management*, vol. 20, no. 8, pp. 593–599, 2002.

[62] M. Sambasivan and Y. W. Soon, "Causes and effects of delays in malaysian construction industry," *International Journal of project management*, vol. 25, no. 5, pp. 517–526, 2007.

[63] S. M. Trost and G. D. Oberlender, "Predicting accuracy of early cost estimates using factor analysis and multivariate regression," *Journal of Construction Engineering and Management*, vol. 129, no. 2, pp. 198–204, 2003.

[64] A. P. Chan, D. W. Chan, Y. Chiang, B. Tang, E. H. Chan, and K. S. Ho, "Exploring critical success factors for partnering in construction projects," *Journal of Construction Engineering and Management*, vol. 130, no. 2, pp. 188–198, 2004.

[65] D. Fang, Y. Chen, and L. Wong, "Safety climate in construction industry: A case study in hong kong," *Journal of construction engineering and management*, 2006.

[66] K. Pietrzyk, "A systemic approach to moisture problems in buildings for mould safety modelling," *Building and Environment*, vol. 86, pp. 50–60, 2015.

[67] V. S. Desai and S. Joshi, "Application of decision tree technique to analyze construction project data," in *Information Systems, Technology and Management*, pp. 304–313, Springer, 2010.

[68] H.-B. Liu and Y.-B. Jiao, "Application of genetic algorithm-support vector machine (ga-svm) for damage identification of bridge," *International Journal of Computational Intelligence and Applications*, vol. 10, no. 04, pp. 383–397, 2011.

[69] T. Mahfouz, J. Jones, and A. Kandil, "A machine learning approach for automated document classification: A comparison between svm and lsa performances," *International Journal of Engineering Research & Innovation*, p. 53, 2010.

[70] T. Mahfouz and A. Kandil, "Construction legal decision support using support vector machine (svm)," *Proc. of the CRC 2010: Innovation for Reshaping Construction*, 2010.

[71] D. Dehestani, F. Eftekhari, Y. Guo, S. Ling, S. Su, and H. Nguyen, "Online support vector machine application for model based fault detection and isolation of hvac system," *International Journal of Machine Learning and Computing*, vol. 1, no. 1, p. 66, 2011.

[72] Q. Chen, Y. Chan, and K. Worden, "Structural fault diagnosis and isolation using neural networks based on response-only data," *Computers & structures*, vol. 81, no. 22, pp. 2165–2172, 2003.

[73] T. Marwala and S. Chakraverty, "Fault classification in structures with incomplete measured data using autoassociative neural networks and genetic algorithm," *Bangalore-Current Science*, vol. 90, no. 4, p. 542, 2006.

[74] X. Fang, H. Luo, and J. Tang, "Structural damage detection using neural network with learning rate improvement," *Computers & structures*, vol. 83, no. 25, pp. 2150–2161, 2005.

[75] O. Moselhi, T. Hegazy, and P. Fazio, "Neural networks as tools in construction," *Journal of Construction Engineering and Management*, 1991.

[76] Y.-J. Chen, C.-W. Feng, Y.-R. Wang, H.-M. Wu, *et al.*, "Using BIM model and genetic algorithms to optimize the crew assignment for construction project planning," *International Journal of Technology*, no. 3, pp. 179–187, 2011.

[77] H. Moon, H. Kim, L. Kang, and C. Kim, "BIM functions for optimized construction management in civil engineering," *Gerontechnology*, vol. 11, no. 2, p. 67, 2012.

[78] D. M. Salama and N. M. El-Gohary, "Semantic text classification for supporting automated compliance checking in construction," *Journal of Computing in Civil Engineering*, p. 04014106, 2013.

[79] T. Mahfouz, "Unstructured construction document classification model through support vector machine (SVM)," pp. 126–133, 2011.

[80] C. H. Caldas and L. Soibelman, "Automating hierarchical document classification for construction management information systems," *Automation in Construction*, vol. 12, no. 4, pp. 395–406, 2003.

[81] N. Ur-Rahman and J. A. Harding, "Textual data mining for industrial knowledge management and text classification: A business oriented approach," *Expert Systems with Applications*, vol. 39, no. 5, pp. 4729–4739, 2012.

[82] S. Liu, C. A. McMahon, and S. J. Culley, "A review of structured document retrieval (SDR) technology to improve information access performance in engineering document management," *Computers in Industry*, vol. 59, no. 1, pp. 3–16, 2008.

[83] L. Soibelman, J. Wu, C. Caldas, I. Brilakis, and K.-Y. Lin, "Management and analysis of unstructured construction data types," *Advanced Engineering Informatics*, vol. 22, no. 1, pp. 15–27, 2008.

[84] I. Brilakis, *Content based integration of construction site images in AEC/FM model based systems*. PhD thesis, 2005.

[85] T. P. Williams, "Using classification rules to develop a predictive indicator of project cost overrun potential from bidding data," *Computing in Civil Engineering (2007)*, p. 35, 2007.

[86] A. O. Elfaki, S. Alatawi, and E. Abushandi, "Using intelligent techniques in construction project cost estimation: 10-year survey," *Advances in Civil Engineering*, vol. 2014, 2014.

[87] S. Han, S. Lee, and F. Peña-Mora, "A machine-learning classification approach to automatic detection of workers actions for behavior-based safety analysis," in *Proceedings of ASCE International Workshop on Computing in Civil Engineering*, 2012.

[88] H. Ng, A. Toukourou, and L. Soibelman, "Knowledge discovery in a facility condition assessment database using text clustering," *Journal of infrastructure systems*, vol. 12, no. 1, pp. 50–59, 2006.

[89] H. Fan and H. Li, "Retrieving similar cases for alternative dispute resolution in construction accidents using text mining techniques," *Automation in Construction*, vol. 34, pp. 85–91, 2013.

[90] M. Al Qady and A. Kandil, "Automatic clustering of construction project documents based on textual similarity," *Automation in Construction*, vol. 42, pp. 36–49, 2014.

[91] M. Al Qady and A. Kandil, "Concept relation extraction from construction documents using natural language processing," *Journal of Construction Engineering and Management*, vol. 136, no. 3, pp. 294–302, 2009.

[92] J. Zhang and N. M. El-Gohary, "Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking," *Journal of Computing in Civil Engineering*, 2013.

[93] J. Zhang and N. El-Gohary, "Extraction of construction regulatory requirements from textual documents using natural language processing techniques," *Journal of Computing in Civil Engineering*, pp. 453–460, 2012.

[94] J. Zhang and N. El-Gohary, "Automated regulatory information extraction from building codes leveraging syntactic and semantic information," in *Proc., 2012 ASCE Construction Research Congress (CRC)*, 2012.

[95] P. Demian and P. Balatsoukas, "Information retrieval from civil engineering repositories: Importance of context and granularity," *Journal of Computing in Civil Engineering*, vol. 26, no. 6, pp. 727–740, 2012.

[96] H. P. Tserng, S. Y.-L. Yin, and M.-H. Lee, "The use of knowledge map model in construction industry," *Journal of Civil Engineering and Management*, vol. 16, no. 3, pp. 332–344, 2010.

[97] H. Fan, F. Xue, and H. Li, "Project-based as-needed information retrieval from unstructured AEC documents," *Journal of Management in Engineering*, vol. 31, no. 1, p. A4014012, 2014.

[98] J.-y. Hsu *et al.*, "Content-based text mining technique for retrieval of cad documents," *Automation in Construction*, vol. 31, pp. 65–74, 2013.

[99] H.-T. Lin, N.-W. Chi, and S.-H. Hsieh, "A concept-based information retrieval approach for engineering domain-specific technical documents," *Advanced Engineering Informatics*, vol. 26, no. 2, pp. 349–360, 2012.

[100] M. Al Qady and A. Kandil, "Document discourse for managing construction project documents," *Journal of Computing in Civil Engineering*, vol. 27, no. 5, pp. 466–475, 2012.

[101] M. Osmani, J. Glass, and A. D. Price, "Architect and contractor attitudes to waste minimisation," 2006.

[102] L. O. Oyedele, M. Regan, J. von Meding, A. Ahmed, O. J. Ebohon, and A. Elnokaly, "Reducing waste to landfill in the UK: identifying impediments and critical solutions," *World Journal of Science, Technology and Sustainable Development*, vol. 10, no. 2, pp. 131–142, 2013.

[103] Z. Wu, T. Ann, L. Shen, and G. Liu, "Quantifying construction and demolition waste: An analytical review," *Waste Management*, vol. 34, no. 9, pp. 1683–1692, 2014.

[104] W. Lu, H. Yuan, J. Li, J. J. Hao, X. Mi, and Z. Ding, "An empirical investigation of construction and demolition waste generation rates in shenzhen city, south china," *Waste management*, vol. 31, no. 4, pp. 680–687, 2011.

[105] L. L. Ekanayake and G. Ofori, "Building waste assessment score: design-based tool," *Building and Environment*, vol. 39, no. 7, pp. 851–861, 2004.

[106] M. Bilal, L. O. Oyedele, O. O. Akinade, S. O. Ajayi, H. A. Alaka, H. A. Owolabi, J. Qadir, M. Pasha, and S. A. Bello, "Big data architecture for construction waste analytics (cwa): A conceptual framework," *Journal of Building Engineering*, 2016.

[107] W. Lu, X. Chen, Y. Peng, and L. Shen, "Benchmarking construction waste management performance using big data," *Resources, Conservation and Recycling*, vol. 105, pp. 49–58, 2015.

[108] W. Lu, X. Chen, D. C. Ho, and H. Wang, "Analysis of the construction waste management performance in hong kong: the public and private sectors compared using big data," *Journal of Cleaner Production*, vol. 112, pp. 521–531, 2016.

[109] N. Kargah-Ostadi, "Comparison of machine learning techniques for developing performance prediction models," in *Computing in Civil and Building Engineering (2014)*, pp. 1222–1229, ASCE.

[110] F. Castronovo, S. Lee, D. Nikolic, and J. I. Messner, "Visualization in 4d construction management software: A review of standards and guidelines," in *In: Proceedings of the International Conference on Computing in Civil and Building Engineering. Orlando USA.*, pp. 315–322, 2014.

[111] S. Goodwin and J. Dykes, "Visualising variations in household energy consumption," in *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pp. 217–218, IEEE, 2012.

[112] T.-H. Chuang, B.-C. Lee, and I.-C. Wu, "Applying cloud computing technology to BIM visualization and manipulation," in *28th International Symposium on Automation and Robotics in Construction*, vol. 201, pp. 144–149, 2011.

[113] Y. Jiao, Y. Wang, S. Zhang, Y. Li, B. Yang, and L. Yuan, "A cloud approach to unified lifecycle data management in architecture, engineering, construction and facilities management: Integrating BIMs and SNS," *Advanced Engineering Informatics*, vol. 27, no. 2, pp. 173–188, 2013.

[114] P. Meadati, J. Irizarry, and A. K. Akhnoukh, "BIM and RFID integration: a pilot study," *Advancing and Integrating Construction Education, Research and Practice*, pp. 570–78, 2010.

[115] Y. Jiao, S. Zhang, Y. Li, Y. Wang, and B. Yang, "Towards cloud

augmented reality for construction application by BIM and SNS integration," *Automation in Construction*, vol. 33, pp. 37–47, 2013.

[116] P. X. Gao and S. Keshav, "Optimal personal comfort management using spot+," in *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, pp. 1–8, ACM, 2013.

[117] A. Rabbani and S. Keshav, "The spot* system for flexible personal heating and cooling," in *Proceedings of the 2015 ACM Sixth International Conference on Future Energy Systems*, pp. 209–210, ACM, 2015.

[118] A. A. Panagopoulos, M. Alam, A. Rogers, and N. R. Jennings, "Adaheat: A general adaptive intelligent agent for domestic heating control," in *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pp. 1295–1303, International Foundation for Autonomous Agents and Multiagent Systems, 2015.

[119] C. Chen and D. J. Cook, "Behavior-based home energy prediction," in *Intelligent Environments (IE), 2012 8th International Conference on*, pp. 57–63, IEEE, 2012.

[120] S. Taneja, A. Akcamete, B. Akinci, J. Garrett, L. Soibelman, and E. W. East, "Analysis of three indoor localization technologies to support facility management field activities," in *Proceedings of the International Conference on Computing in Civil and Building Engineering, Nottingham, UK*, 2010.

[121] H. S. Ng and L. Soibelman, "Knowledge discovery in maintenance databases: Enhancing the maintainability in higher education facilities," in *Proc., 2003 Construction Research Congress*, ASCE, Reston, Va, 2003.

[122] R. Liu and R. Issa, "Issues in BIM for facility management from industry practitioners perspectives," *Computing in Civil Engineering (2013)*, pp. 411–418, 2013.

[123] J. Sanyal and J. New, "Simulation and big data challenges in tuning building energy models," in *Modeling and Simulation of Cyber-Physical Energy Systems (MSCPES), 2013 Workshop on*, pp. 1–6, IEEE, 2013.

[124] O. Linda, D. Wijayasekara, M. Manic, and C. Rieger, "Computational intelligence based anomaly detection for building energy management systems," in *Resilient Control Systems (ISRCS), 2012 5th International Symposium on*, pp. 77–82, IEEE, 2012.

[125] I. Hong, J. Byun, and S. Park, "Cloud computing-based building energy management system with zigbee sensor network," in *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2012 Sixth International Conference on*, pp. 547–551, IEEE, 2012.

[126] R. P. Singh, S. Keshav, and T. Brecht, "A cloud-based consumer-centric architecture for energy data analytics," in *Proceedings of the fourth international conference on Future energy systems*, pp. 63–74, ACM, 2013.

[127] M. Berges, E. Goldman, H. S. Matthews, and L. Soibelman, "Learning systems for electric consumption of buildings," in *ASCI international workshop on computing in civil engineering*, vol. 38, 2009.

[128] C. Wei and Y. Li, "Design of energy consumption monitoring and energy-saving management system of intelligent building based on the internet of things," in *Electronics, Communications and Control (ICECC), 2011 International Conference on*, pp. 3650–3652, IEEE, 2011.

[129] R. Volk, J. Stengel, and F. Schultmann, "Building information modeling (BIM) for existing buildingsliterature review and future needs," *Automation in Construction*, vol. 38, pp. 109–127, 2014.

[130] X. Wang, "BIM handbook: A guide to building information modeling for owners, managers, designers, engineers and contractors," *Construction Economics and Building*, vol. 12, no. 3, pp. 101–102, 2012.

[131] S. Azhar, "Building information modeling (BIM): Trends, benefits, risks, and challenges for the AEC industry," *Leadership and Management in Engineering*, 2011.

[132] K. Barlish and K. Sullivan, "How to measure the benefits of BIMa case study approach," *Automation in construction*, vol. 24, pp. 149–159, 2012.

[133] J. D. Goedert and P. Meadati, "Integrating construction process documentation into building information modeling," *Journal of construction engineering and management*, vol. 134, no. 7, pp. 509–516, 2008.

[134] U. Isikdag, J. Underwood, G. Aouad, and N. Trodd, "Investigating the role of building information models as a part of an integrated data layer: a fire response management case," *Architectural Engineering and Design Management*, vol. 3, no. 2, pp. 124–142, 2007.

[135] N. Yu, Y. Jiang, L. Luo, S. Lee, A. Jallow, D. Wu, J. I. Messner, R. M. Leicht, and J. Yen, "Integrating BIMserver and openstudio for energy efficient building," in *Proceedings of 2013 ASCE International Workshop on Computing in Civil Engineering*, 2013.

[136] J. Zhang, Q. Liu, F. Yu, Z. Hu, and W. Zhao, "A framework of cloud-computing-based BIM service for building lifecycle," *Computing in Civil and Building Engineering*, pp. 1514–1521, 2014.

[137] J. Wong, X. Wang, H. Li, G. Chan, and H. Li, "A review of cloud-based BIM technology in the construction sector," *The Journal of Information Technology in Construction*, vol. 19, pp. 281–291, 2014.

[138] Y. Hsieh, "Cloud-computing based parameter identification system–with applications in geotechnical engineering," in *Computing in Civil and Building Engineering (2014)*, pp. 1384–1392, ASCE.

[139] R. Klinc, I. Peruš, and M. Dolenc, "Re-using engineering tools: Engineering SaaS web application framework," in *Proceedings of the 2014 International Conference on Computing in Civil and Building Engineering*, pp. 23–25, 2014.

[140] B. Kumar, J. C. Cheng, and L. McGibbney, "Cloud computing and its implications for construction it," in *Computing in Civil and Building Engineering, Proceedings of the International Conference*, vol. 30, p. 315, 2010.

[141] A. Redmond, A. V. Hore, and R. West, "Building support for cloud computing in the Irish construction industry," 2010.

[142] F. Fathi, M. Abedi, S. Rambat, S. Rawai, M. Z. Zakiyudin, *et al.*, "Context-aware cloud computing for construction collaboration," *Journal of Cloud Computing*, vol. 2012, pp. 1–11, 2012.

[143] T. H. Beach, O. F. Rana, Y. Rezgui, and M. Parashar, "Cloud computing for the architecture, engineering & construction sector: requirements, prototype & experience," *Journal of Cloud Computing*, vol. 2, no. 1, pp. 1–16, 2013.

[144] H.-Y. Chong, J. S. Wong, and X. Wang, "An explanatory case study on cloud computing applications in the built environment," *Automation in Construction*, vol. 44, pp. 152–162, 2014.

[145] A. Grilo and R. Jardim-Goncalves, "Cloud-marketplaces: Distributed e-procurement for the AEC sector," *Advanced Engineering Informatics*, vol. 27, no. 2, pp. 160–172, 2013.

[146] T. Elghamrawy and F. Boukamp, "Managing construction information using rfid-based semantic contexts," *Automation in construction*, vol. 19, no. 8, pp. 1056–1066, 2010.

[147] E. Curry, J. ODonnell, E. Corry, S. Hasan, M. Keane, and S. ORiain, "Linking building data in the cloud: Integrating cross-domain building data using linked data," *Advanced Engineering Informatics*, vol. 27, no. 2, pp. 206–219, 2013.

[148] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of things for smart cities," *Internet of Things Journal, IEEE*, vol. 1, no. 1, pp. 22–32, 2014.

[149] A. Khan and K. Hornbæk, "Big data from the built environment," in *Proceedings of the 2nd international workshop on Research in the large*, pp. 29–32, ACM, 2011.

[150] J. Stankovic *et al.*, "Research directions for the internet of things," *Internet of Things Journal, IEEE*, vol. 1, no. 1, pp. 3–9, 2014.

[151] D. Bonino, F. Corno, and L. De Russis, "Real-time big data processing for domain experts, an application to smart buildings," *Big Data Computing/Rajendra Akerkar; Taylor & Francis Press: London, UK*, vol. 1, pp. 415–447, 2013.

[152] C. Miller, Z. Nagy, and A. Schlueter, "Automated daily pattern filtering of measured building performance data," *Automation in Construction*, vol. 49, pp. 1–17, 2015.

[153] G. Williams, M. Gheisari, P.-J. Chen, and J. Irizarry, "BIM2MAR: An efficient BIM translation to mobile augmented reality applications," *Journal of Management in Engineering*, 2014.

[154] B. Zhadanovsky and S. Sinenko, "Visualization of design, organization of construction and technological solutions," in *Computing in Civil and Building Engineering (2014)*, pp. 137–142, ASCE.

[155] A. Motamedi, *Improving Facilities Lifecycle Management Using RFID Localization And BIM-Based Visual Analytics*. PhD thesis, Concordia University, 2013.

[156] N. Gu and K. London, "Understanding and facilitating BIM adoption in the AEC industry," *Automation in construction*, vol. 19, no. 8, pp. 988–999, 2010.

[157] C. Chiang, T. Ho, and C. Chou, "A BIM-enabled platform for power consumption data collection and analysis," *Computing in Civil Engineering*, pp. 90–98, 2012.

[158] K.-C. Yeh, M.-H. Tsai, and S.-C. Kang, "On-site building information retrieval by using projection-based augmented reality," *Journal of Computing in Civil Engineering*, 2012.

[159] J. Bughin, M. Chui, and J. Manyika, "Clouds, big data, and smart assets: Ten tech-enabled business trends to watch," *McKinsey Quarterly*, vol. 56, no. 1, pp. 75–86, 2010.

[160] A. Redmond, A. Hore, M. Alshawi, and R. West, "Exploring how information exchanges can be enhanced through cloud BIM," *Automation in Construction*, vol. 24, pp. 175–183, 2012.

[161] C. Amarnath, A. Sawhney, and J. U. Maheswari, "Cloud computing to enhance collaboration, coordination and communication in the construction industry," in *Information and Communication Technologies (WICT), 2011 World Congress on*, pp. 1235–1240, IEEE, 2011.

[162] N. M. Rawai, M. S. Fathi, M. Abedi, and S. Rambat, "Cloud computing for green construction management," in *Intelligent System Design and Engineering Applications (ISDEA), 2013 Third International Conference on*, pp. 432–435, IEEE, 2013.

[163] G. Kortuem, F. Kawsar, D. Fitton, and V. Sundramoorthy, "Smart objects as building blocks for the internet of things," *Internet Computing, IEEE*, vol. 14, no. 1, pp. 44–51, 2010.

[164] H.-A. Jacobsen, R. H. Katz, H. Schmeck, and C. Goebel, "Smart buildings and smart grids (dagstuhl seminar 15091)," *Dagstuhl Reports*, vol. 5, no. 2, 2015.

[165] S. Rankohi and L. Waugh, "Review and analysis of augmented reality literature for construction industry," *Visualization in Engineering*, vol. 1, no. 1, pp. 1–18, 2013.

[166] H.-L. Chi, S.-C. Kang, and X. Wang, "Research trends and opportunities of augmented reality applications in architecture, engineering, and construction," *Automation in construction*, vol. 33, pp. 116–122, 2013.

[167] B. Bossink and H. Brouwers, "Construction waste: quantification and source evaluation," *Journal of construction engineering and management*, vol. 122, no. 1, pp. 55–60, 1996.

[168] V. W. Tam, C. M. Tam, S. Zeng, and W. C. Ng, "Towards adoption of prefabrication in construction," *Building and environment*, vol. 42, no. 10, pp. 3642–3654, 2007.

[169] P. Pauwels and W. Terkaj, "Express to owl for construction industry: towards a recommendable and usable ifcowl ontology," *Automation in Construction*, vol. 63, pp. 100–133, 2016.

[170] J. Beetz, J. Van Leeuwen, and B. De Vries, "Ifcowl: A case of transforming express schemas into ontologies," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, vol. 23, no. 01, pp. 89–101, 2009.

[171] R. Wong, "Big data privacy," *Journal of Information Technology and Software Engineering*, vol. 2, p. e114, 2012.

[172] A. Cavoukian and J. Jonas, *Privacy by design in the age of big data*. Information and Privacy Commissioner of Ontario, Canada, 2012.

[173] O. Tene and J. Polonetsky, "Big data for all: Privacy and user control in the age of analytics," *Nw. J. Tech. & Intell. Prop.*, vol. 11, p. xxvii, 2012.

LIST OF FIGURES

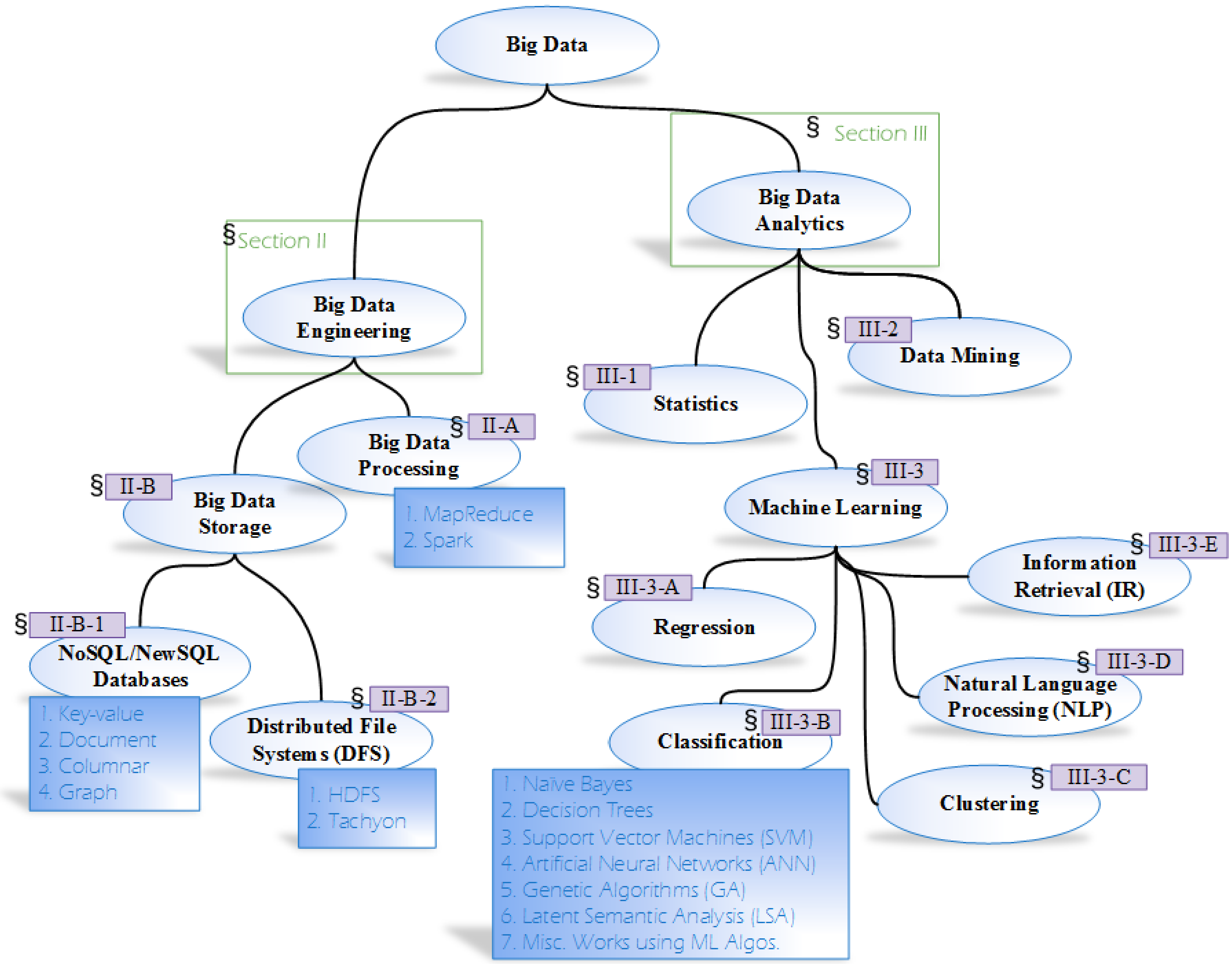

Fig. 1. Proposed Review Structure of the Paper

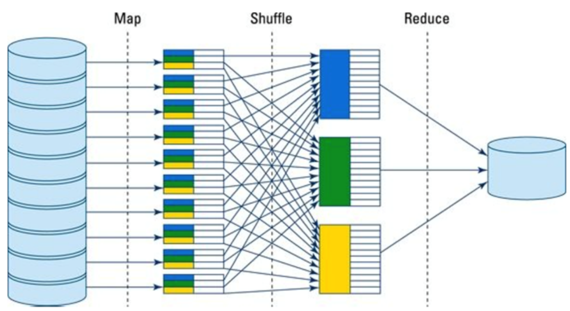

Fig. 2. An overview of MapReduce processing [12]

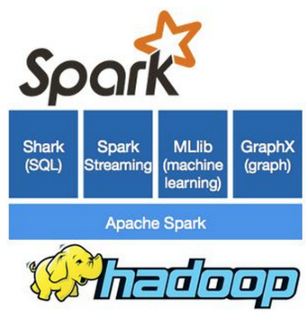

Fig. 3. Apache Spark and Related Technology Stack [15]

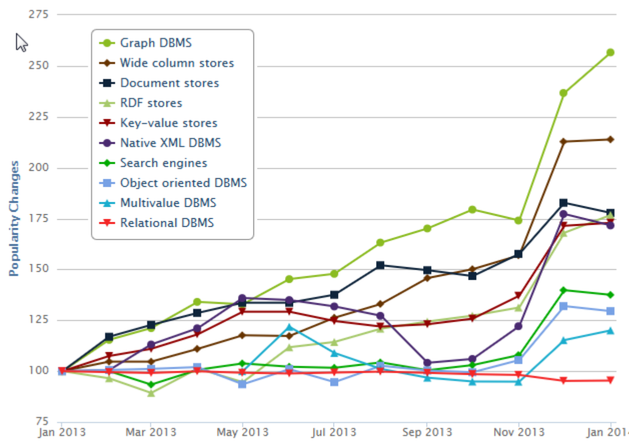

Fig. 4. Databases Popularity Trend [19]

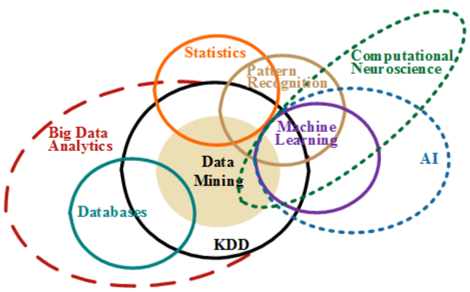

Fig. 5. Multidisciplinary Nature of Big Data Analytics (Adapted from [28])

LIST OF TABLES

TABLE I. PROMINENT NOSQL SYSTEMS AND THEIR CRITICAL FEATURES.

Product Name	**Product Description**	**Data Model(s)**	**Language**	**Concurrency**	**Storage**	**Key Features**
Cassandra	Apache Cassandra is scalable database that provides proven fault-tolerance and tunable consistency on cluster of commodity servers.	Columnar Key-Value	Java, Python	MVCC	Disk, Hadoop, Plugin	High Availability, Partition Tolerance
HBase	HBase is distributed data store that extended Google Bigtable to scale on HDFS. Its novelty lies in storing and accessing data with random access. It doesn't restrict the kind of data being stored.	Columnar Key-Value	Java	Locks	Hadoop	Consistent, Partition Tolerance
HyperTable	Hypertable supports data distribution for scalable data management. It offers maximum efficiency and superior performance. However, it lacks data management features such as transaction and join processing.	Columnar	C++	MVCC	Disk Hadoop GlusterFS, Kosmos File System	Consistent, Partition Tolerance
MongoDB	MongoDB is a document-oriented database. It facilitates storage of documents with variable schemas and is suitable for applications, storing complex types.	Document Key-Value	C++	Locks	Disk,GFS, Plugin	Consistent, Partition Tolerance
CouchDB	CouchDB is suitable for large scale web and mobile applications. It facilitate data storage that are queried through web browsers, via HTTP. JavaScript is used to index, integrate, and transform the database.	Document Key-Value	Erlang, C	MVCC	Disk	High Availability, Partition Tolerance
MarkLogic	MarkLogic facilitates storing documents efficiently for easy and intuitive search. It is suitable for applications that derive revenue, streamline operations, risk management, and security.	Document	C, Java, Python	ACID	GFS Hadoop S3, RDF	Consistent, High Availability, Partition Tolerance
Redis	Redis is in-memory system that can be used as a database, cache, and message broker. When configured on cluster, it becomes scalable and highly available. It also supports transaction processing.	Key-Value	ANSI C	Locks	RAM	Consistent, Partition Tolerance
Riak	Riak is a distributed database that provides scalability and high availability. It achieves performance and fault tolerance through built-in distribution and replications.	Key-Value	Erlang	ACID	Disk, Plugin	High Availability, Partition Tolerance
BarkeleyDB	Berkeley DB is embedded database for key-value dataset. It is written in C but supports application development for C++, PHP, Java, Perl, among others.	Key-Value	Java	ACID	RDF	Consistency, Availability, Partition-Tolerance
Neo4J	Neo4J is a semantic store for creating, updating, deleting, and retrieving graph data. It captures relationships natively and processes queries as paths through language called Cypher. Neo4J is good option for applications, dealing with connected data.	Graph	Java	Locks	Disk	High Availability, Partition Tolerance
OrientDB	OrientDB is a system for large-scale and distributed graph management. The core features include multi-master replication and sharding.	Graph	Java	ACID	Disk, Plug-in RAM, SSD	Consistent, High Availability, Partition Tolerance
Oracle NoSQL	Oracle NoSQL is designed specifically to provide highly reliability, scalability, and maximum availability across the cluster of storage nodes. Data is replicated to survive rapid failure and load balancing for distributed query processing.	Columnar Document Key-Value Graph	Java	ACID	Berkeley DB Architecture, RDF	Consistency, Availability, Limited Partition-Tolerance

TABLE II. Big Data Analytics (BDA) ML Tools

Tool Name	Description	Supported Languages	ML at Scale	Supported Algorithms
Apache Mahout [25]	Mahout is an open-source machine learning framework for quickly writing scalable and high performance ML applications.	-Java -Scala.	Yes	-Collaborative Filtering -Classification -Clustering -Regression
R [26]	R is an open-source programming language for statistical analysis. R is extremely extensible. With huge developer base, thousands of R packages are available to provide variety of functionalities. The graphics supported by R are highly polished and very powerful.	Many languages	Yes	-Collaborative Filtering -Classification -Clustering -Regression
MLbase [27]	Spark has constituted a novel ML platform called MLbase, which has brought together highly robust ML components, such as *ML optimizer*, *MLI*, and *MLlib*, to support the full lifecycle activities, required to implement as well as use ML algorithms. ML optimizer automates the tasks of ML pipeline construction to efficiently search algorithms of MLI and MLlib. MLI is the API to develop ML algorithms using high-level constructs. MLlib is the Spark distributed ML library.	-Java -Scala -Python	Yes	-Collaborative Filtering -Classification -Clustering -Regression
Oryx [14]	Oryx is an open-source ML library that has evolved over time out of the libraries and toolkits developed by Cloudera. Based on the distributed input from HDFS, it builds predictive models that are written to output in predictive model markup language (PMML). An interesting feature of Oryx is its ability to keep the model updated under emerging streams of data from Hadoop.	-Java	Yes	-Collaborative Filtering -Classification -Clustering -Regression

TABLE III. SUMMARY OF WORKS WITH STATISTICAL METHODS

Purpose of use	**Technique(s) employed**	**References**
Identifying the causes of construction delays	-Frequency charts -Correlation matrix -Factor analysis -Bayesian networks	[31]
Learning from post project reviews (PPRs)	-Link analysis -Dimensional matrix analysis	[32]
Decision support systems for construction litigation	-Naïve Bayes -Decision trees -Rule inductive	[33]
Structural damage detection in buildings	-Gaussian distribution -Monte Carlo simulation	[34]
Identifying workers and heavy machinery actions towards site safety	-Gaussian distribution -Naïve Bayes -Bags of video feature	[35], [36]

TABLE IV. SUMMARY OF WORKS ON DATA MINING

Purpose of use	**Technique(s) employed**	**References**
Causes of construction project delays	-KDD -Statistics	[31]
Cost overruns and quality control in construction projects	-KDD -Data function	[43], [50], [44]
Learning from past Projects (PPRs)	-Text mining -Link analysis	[32]
Identifying and coordinating spatial conflicts in MEP design	-KDD	[38]
Presenting occupational injuries	-Association rule mining -Classification and regression tree (CART)	[45], [46]
Construction data integration for enhanced productivity	-Data warehousing -OLAP	[41], [47]
Querying partial BIM models in information systems	-SQL -EQL -BIMQL	[48] [49]

TABLE V. CLASSIFICATION-BASED WORKS FOR THE CONSTRUCTION INDUSTRY, CATEGORIZED PER CLASSIFICATION TECHNIQUE

Purpose of use	Technique(s) employed	References
Naïve Bayes		
Detecting structural damages of buildings	-Gaussian distribution -Probability density function	[34]
Complex actions classification of workers and heavy machinery	-Bags-of-video-features -Bayesian probability	[35]
Stiffness reduction of structures, caused by earthquakes	-Bayesian probability	[36]
Decision Trees (DTs)		
Assessment of mould germination in building structures	-Fault tree analysis	[66]
Construction labour productivity assessment	-Augmented decision tree	[67]
Support Vector Machine (SVM)		
Damage identification in bridges	-SVM -GA-RDF	[68]
Automated construction document classification	-SVM -LSA	[69]
Legal decision support system	-SVM -TF -TF/IDF -LTF	[70]
Semi-supervised fault detection and isolation system for HVAC	-SVM	[71]
Artificial Neural Networks (ANN)		
Structural fault detection, caused by vibration and fatigue	-Transmissibility Functions -ANN	[72]
Fault classification system	-GA -ANN	[73]
Structural damage detection	-Tuneable steepest descent method -Frequency response function	[74]
Expert system for optimal markup estimation	-ANN	[75]
Genetic Algorithms (GA)		
Cost/schedule integrated planning system for optimal crew assignment	-GA -Object sequencing matrix	[76]
Risks imposed by schedule and workspace conflicts	-GA -Fuzzy logic	[77]
Latent Semantic Analysis (LSA)		
Automated construction document classification	-LSA	[69]
Automated regulatory and contractual compliance system	-LSA	[78]

TABLE VI. CLASSIFICATION-BASED APPLICATIONS IN THE CONSTRUCTION INDUSTRY

Purpose of use	**Technique(s) employed**	**References**
Document classification		
Document classification based on CSI MasterFormat	-Boolean weighting -Absolute frequency -TF/IDF -IFC weighting	[80]
Classifying post project review documents	-SVM -KNN -DT -Naïve Bayes	[81]
Structured document retrieval system	-SGML -XML	[82]
Document analysis		
Unstructured document analysis system	-ML classifiers	[83]
Image-based classification		
Indexing construction site imagery	-Whitening Transform (WT) -SVM -Biased Discriminant Transform (BDT)	[84]
Predicting overrun potential		
Highway project bidding system for overrun prediction	- Ripple Down Rules	[85]
Construction project estimation	-ML algorithms	[86]
Safety analysis		
Worker behaviour modelling to predict site injury from construction site videos	-Bayesian classifier	[87]

TABLE VII. SUMMARY OF OPPORTUNITIES WITHIN THE SUB-DOMAIN OF CONSTRUCTION INDUSTRY

Construction Industry Sub-domains	**State of the Art**	**Potential Opportunities**
Resource and Waste Optimization	-Construction waste generation estimation [105], [107] -Waste generation benchmarking [107] Comparative analysis of waste management performance [108]	-BIM tools to actualize circular economy for sustainability, green supply chains, and closed-loop supply chains -BIM tools for optimal & auto design specification -BIM integrated materials database using open standards -BIM integrated linked data for waste data management -BIM based waste estimation using predictive analytics -BIM based waste minimisation through deconstruction -BIM based waste minimisation through resource optimisation -BIM based waste minimisation through interactive visualisation
Generative Designs	-Autodesk Dreamcatcher—a prototype system to showcase the feasibility of this idea of generating design from abstract requirements	-Framework to exploit Big Data analytics to parallelize algorithms for real time GD computation -Big Data algorithms to accurately reduce the design space -Big Data enabled GD tool
Clash Detection and Resolution	-BIM enabled approaches are developed to resolve conflicts in MEP design, however these approaches are time consuming [38]	-Big Data analytics based MEP design checker that uses prescriptive analytics not only to identify conflicts but also describe the best action to resolve it.
Performance Prediction	-Pavement management system using pavement deterioration prediction [109]	-Big Data driven BIM system for pavement deterioration prediction
Visual Analytics	-4D BIM Visualisation [110], energy user classification [111], using VA -Cloud-based BIM system for design visualisation and exploration [112]	-Visual analytics driven Big Data framework for BIM model visualisation -Visual analytics driven design optimiser for energy reduction and comfort maximisation
Social Networking Services/ Analytics	-Integration of project management data using social networking [113] -BIM, RFID, and social data integration [114] -AR based Business social networking services (BSNS) [115]	-BIM framework for social network information modelling using Big Data
Personalized Services	-SPOT+ indoor air personalisation [116], SPOT* for heating/cooling [117] -AdaHeat domestic heat regulator [118], Behavioural energy adaptation [119]	-Personalisation energy monitor that requires less user input to regulate optimal energy consumption
Facility Management	-BIM based indoor localisation [120], FM cost reduction through massive data exploration [121], FM data modelling through BIM [122]	-Big Data Analytics based BIM system for FM activities
Energy Management and Analytics	-Energy simulation software (EnergyPlus) [123], energy management systems [124], Cloud based energy data storage and processing [125], [126] -Appliance event identification using NILM and Wire Spy [127], [128], Energy user classification [111], IOT framework for energy analytics [128]	-Big Data framework for BIM based open energy data persistence -Big Data analytics platform to simulate and optimise energy usage of buildings
Big Data with BIM	-BIM models for building designs [129], [130], [131], [132], BIM models for construction process documentation [133], BIM models for GIS data [134], BIM for MEP conflict resolution [38], BIM open platform [135], BIM via cloud [20], [136], [112], BIM and RFID [114], BIM models for project management data [113], MapReduce savvy BIM data storage and processing [6]	-Big Data enabled IFC-compliant BIM storage system -BIM platform for IOT applications -Open data platform for linking BIM models with external sources -Big Data enabled BIM processing platform to developing applications
Big Data with Cloud Computing	-Cloud based energy data management [125] -Cloud based BIM data management [20] -Cloud enabled BIM design data storage & exploration [112], [137], [136], [138] -SaaS platform for structural MEP analysis [139] -Cloud based BIM system for SMEs [140], [141] -BIM based context-aware computing [142] -Amazon EC2 enabled Google SketchUp [143], [144] -Cloud based e-procurement platform [145]	-A BDA platform to store and process BIM models on cloud for developing domain specific applications.
Big Data with Internet of Things (IOT)	-RFID based construction document retrieval & assets management system [146], [114] -IOT based energy monitoring and analysis system [128], [147] -Urban IOT, a framework for smart cities [148]	-Big Data driven IOT platform for Smart Buildings
Big Data for Smart Buildings	-Project dasher for measuring and visualising CO_2 emission of buildings [149] -A robust firefighting systems for buildings [150] -Complex event processing for smart buildings [151] -DayFilter, for pattern recognition in energy data [152]	-Building Personalisation Services using Big Data -Mobiles apps to exploit building personalisation services
Big Data with Augmented Reality (AR)	-BIM2MAR, a platform to integrate BIM, mobile and AR [153] -Web3D-based AR system for BIM and social networking services (SNS) [115]	-Big BIM Data Visual Exploration System -Big Data and AR based virtual site exploration -Big Data and AR enabled Proactive Dispute Identification and resolution System -Big Data and AR enabled As-planned vs. As-built Comparison System