Manufacturing big data ecosystem: A systematic literature review

Yesheng Cui[a,*], Sami Kara[a], Ka C. Chan[a,b]

[a] *Sustainable Manufacturing and Life Cycle Engineering Research Group, School of Mechanical and Manufacturing Engineering, The University of New South Wales, Sydney, NSW 2052, Australia*

[b] *School of Management and Enterprise, Faculty of Business, Education, Law and Arts, University of Southern Queensland, Springfield, QLD 4305, Australia*

Abstract

Advanced manufacturing is one of the core national strategies in the US (AMP), Germany (Industry 4.0) and China (Made-in China 2025). The emergence of the concept of Cyber Physical System (CPS) and big data imperatively enable manufacturing to become smarter and more competitive among nations. Many researchers have proposed new solutions with big data enabling tools for manufacturing applications in three directions: product, production and business. Big data has been a fast-changing research area with many new opportunities for applications in manufacturing. This paper presents a systematic literature review of the state-of-the-art of big data in manufacturing. Six key drivers of big data applications in manufacturing have been identified. The key drivers are system integration, data, prediction, sustainability, resource sharing and hardware. Based on the requirements of manufacturing, nine essential components of big data ecosystem are captured. They are data ingestion, storage, computing, analytics, visualization, management, workflow, infrastructure and security. Several research domains are identified that are driven by available capabilities of big data ecosystem. Five future directions of big data applications in manufacturing are presented from modelling and simulation to realtime big data analytics and cybersecurity.

1 Introduction

Smart manufacturing is critical to national economies by providing jobs, improving innovation and advancing sustainability [1]. Several national strategies were initiated to boost their competitiveness of manufacturing, such as: 'Industry 4.0' in Germany [2], 'Advanced Manufacturing Partnership (AMP)' program in the United States [3], 'Made in China 2025' and so on. These initiatives provide massive potential to envision the future of manufacturing: Smart manufacturing, which is defined by National Institute of Standards and Technology (NIST) as a completely integrated, collaborative manufacturing system that respond in real time to meet changing demands and conditions in the factory, in the supply network and in customer needs[4].

Manufacturing industry uses a wide range of software and automation systems to increase efficiency and productivity from shop floors to enterprise layers such as CNC machines, Programmable Logic Controllers (PLC), Supervisory Control And Data Acquisition System (SCADA) [5], Manufacturing Executive System (MES) [6], product design and development (CAx: CAD, CAPP, CAM, CAE [7]), Product Lifecycle Management (PLM) [8], Enterprise Resource Planning system (ERP) [9], Operating and Maintenance (O&M) [10], Energy Management System (EMS) [11], Supply Chain Management (SCM) [12], Customer Relationship Management (CRM) [13] etc. However, smart manufacturing cannot be realised with traditional manufacturing software and technologies due to two main challenges. First, these systems and software cannot be fully integrated and collaborative since they are developed by multiple vendors using different interfaces or protocols. Second, manufacturers cannot perceive and respond to the real-time changes on time from the factory, supply chain and market since the traditional manufacturing software lack sensory data to notice the changes inside and outside the systems.

Digital thread and digital twin are two recent concepts proposed by integrating disparate systems over the product lifecycle [14] and building up the real-time relationship between the physical space and the cyberspace in manufacturing [15] respectively. Cyber-Physical systems (CPS) are physical and engineered systems, which are monitored, controlled, coordinated and integrated with computing and communicating core [16]. Internet of Things (IoT) is data-accessing and data-processing technologies on the cyberspace to perceive the real-time changes of physical space with sensory tools [17][18]. Digital thread and digital twin can be enabled by using IoT and CPS. With the practices of these new concepts and technologies, massive data will be generated from the systems, which have to go through the processes of data collection, storage, aggregation, analysis and exchange to provide timely information to manufacturers. Being empowered with cloud computing [19], data science [20] and Artificial Intelligence (AI) [21], big data focuses on addressing the big data issues among the processes, which traditional manufacturing tools cannot. Big data could be the enabler of the concepts of digital thread and digital twin.

Smart manufacturing gains actionable knowledge in real-time with the fusion of big data and manufacturing knowledge. As big data are collected and analysed to extract timely information, manufacturing industry may still not know which approach to use, and their impacts without the domain knowledge [22]. The actionable knowledge is created when manufacturers get timely information from big data and apply manufacturing knowledge in a specific application. Some examples found in the reviewed literature are: identifying the reasons of faults from the production process by analysing real-time big processing data and manufacturing knowledge [23], predicting maintenance intervals by utilising knowledge discovery and hundreds of machine data attributes [24], making real-time scheduling and cost-effective decisions in MES system with streaming shop floor data and existing manufacturing systems [25].

Similar to big data, manufacturing industry faces the same challenges associated with 5Vs of big data (Volume, Velocity, Variety, Veracity and Value) [26]. IDC reports that manufacturing has the largest

share of data (3584 Exabyte) in 2018 and will have 30% annual growth rate of data from 2018 to 2025 [27]. Of the reported data types, structured data, such as tabular data in relational databases or spreadsheets, accounts for only 5% of all the data generated [23]; while the rest is made up of semi-structured and unstructured data with formats JSON, XML, image, video, and audio, etc. Issues of velocity, variety and veracity can be explained that the same type of data come from different devices with various sampling frequencies, formats, precisions, which leads to inconsistent data and makes challenging to extract the value-added insight to manufacturers. Amir et al. illustrate that the limitations of the traditional methods (relational database management systems (RDBMS) and on-premise software) could not handle big data [28]. As more manufacturing enterprises generate big data, the issues of big data will become pressing.

Tapping into the capabilities of big data tools presents enormous opportunities for smart manufacturing. A large number of big data tools are developed by the big players in the Internet industry such as Google, Yahoo, Facebook for their own applications in search engines, social media, and business analytics [29][30][31], such as Apache Hadoop [30], Apache Spark, Apache Flume, Apache Flink, Apache Storm, NoSQL and NewSQL databases [32], Apache Hive, Apache Pig, Apache Zookeeper etc. As these tools are enterprise-ready to use in manufacturing, many researchers reported big data based solutions, which use a set of big data tools to address problems to enable smart manufacturing. In 2012, a Hadoop-based sensor data management framework was proposed for cloud manufacturing [33]. In 2014, Tao proposed the architecture of cloud manufacturing system as well as the investigation of applying cloud computing and IoT technology in manufacturing [34]. In 2015, the Cloud-Based Design Manufacturing paradigm (CBDM) was proposed by comparing other design methods [35]. In 2017, Nagorny etc. conclude that big manufacturing data and big data analytics would provide a vast potential in smart manufacturing [26]. In 2017, Wu presented a fog computing-based framework to monitor machine health in cyber-manufacturing. In 2018, Tao proposed a data-driven conceptual framework to interoperate with ERP, MES, CRM, PLM systems in manufacturing [36]. Big data tools complement and provide additional functionalities to address manufacturing big data issues which could not have been solved by traditional approaches.

Three challenging issues have to be addressed in the research of big data in manufacturing. Firstly, big data tools from Internet industry do not consider the differences between Internet and manufacturing. Most manufacturing data is standardized which is supported by various industrial vendors and associations such as manufacturers of CNC machines, meters and sensors, controllers and software companies. The manufacturers use different hardware interfaces, communication protocols, manufacturing machine readable languages, and semantical definitions. Whereas, most data in the internet is based on natural languages and easier to be exchanged without the difficulties associated with multiple interfaces and protocols. The differences between the two industries have not been taken into account in developing big data tools. Secondly, big data tools are massive, diverse, with many overlap functions. It is challenging to design big data based solution by selecting suitable big data tools. However, designing big data solutions not only depends on big data tools but are closely associated with the specific manufacturing applications and scenarios. This paper categorizes similar types of big data tools and identifies the differences as the preparation to achieve this purpose. Thirdly, many manufacturing systems have dedicated aims, complicated and sophisticated functions, and are closely specialised with application scenarios. Social media uses big data tools to collect and store time series data from billions of customers who follow each other online; and to report trending events. For manufacturing, time series data from multiple data sources can be collected, integrated, and analysed to explain the states of manufacturing entities. For example, the sensor data collected from a CNC machine reflects the state of the machine and can be used to develop simulation models or prediction models for preventive maintenance. As many solutions are proposed, a systematic literature review is required to identify the data issues in manufacturing, capabilities of big data tools, essential components to design big data based solutions in manufacturing and the potential research directions of big data in manufacturing.

The rest of this paper is presented as follows: Section 2 presents the methodology to systematically review the state of the art of big data research in manufacturing; Section 3 presents the outcomes of the systematic review; Section 4 discusses several critical issues of big data ecosystem in

manufacturing including critical drivers, system requirements, essential components, research innovation and future directions; and finally, Section 5 presents the conclusion of this systematic review.

Nomenclature

BDA	Big Data Analytic	NoSQL	Not Only Structured Query Language
BI	Business Intelligence	NIST	National Institute of Standards and Technology, USA
CAD	Computer-Aided Design	OLAP	On-Line Analytic Processing
CAPP	Computer-Aided Process Planning	OLTP	On-Line Transaction Processing
CAM	Computer-Aided Manufacturing	O & M	Operation and Maintenance
CAE	Computer-Aided Engineering	OPC-UA	OPC Unified Architecture
CNC	Computer Numerical Control	OWL	Web Ontology Language
CRM	Customer Relationship Management	PDM	Product Data Management
DSS	Decision Support System	PLM	Product Lifecycle Management
EOL	End-Of-Life	QMS	Quality Management System
ERP	Enterprise Resource Planning	RDF	Resource Description Framework
HTML	Hypertext Markup Language	RDBMS	Relational Database Management System
IIoT	Industry Internet Of Things	SCADA	Supervisory Control And Data Acquisition
JSON	JavaScript Object Notation	SCM	Supply Chain Management
KM	Knowledge Management	STEP	Standard for Exchange of Product data
MES	Manufacturing Execution System	STEP-NC	STEP for Numerical Control
MOM	Manufacturing Operations Management	XML	Extensible Markup Language
MES	Manufacturing Executions System		

2 Methodology

This paper presents a systematic literature review (SLR) on the current state of research associated with big data technologies in manufacturing [37]. To apply big data technologies in manufacturing successfully, it is essential to systematically review the literature of big data technologies in manufacturing from the following three perspectives: manufacturing data, big data technologies and data applications in manufacturing.

Firstly, manufacturing data is the foundation to conduct data-driven manufacturing. It is impossible to propose one big data based solution to fit all manufacturing circumstances since different applications have different data issues (data types, data formats and data sources) and require specific tools to address. Therefore, systematically analysing manufacturing data could provide a useful guideline to select appropriate big data enabling technologies.

Secondly, the big data tools in the big data ecosystem have to be identified the similarities and differences. The 5Vs characters of big data are widely recognised as challenges, such as volume (TB/PB level of data size), velocity (ingesting or processing big data in streams or batches, in real time or non-real time), variety (dealing with complex big data formats, schemas, semantic models and information), value (analysing data to deliver added-value to some events), and veracity (validate data consistency and trustworthy) [38]. In general, these big data technologies are intended to address some Vs of big data. Hence, their capabilities need to be effectively classified and analysed to know which Vs are addressed.

Thirdly, gaps of data applications in manufacturing could be identified by systematically reviewing the capabilities of the traditional manufacturing systems and big data analysis. Since much traditional manufacturing software has been widely used in enterprises, big data could integrate and collaborate the software and systems as well as providing timely information. The massive amount of data generated by these applications can be fed back to the big data ecosystems for analytics and innovative applications such as prediction, optimization, monitoring, simulation and visualisation, etc. Therefore, these application gaps would be the future research directions in academia and the demands in the industry.

In summary, knowing the data requirements of manufacturing applications, understanding the capabilities of big data tools, and identifying the gaps will help define future research directions and generate new ideas for innovative applications. This systematic literature review presents a holistic overview of big data in manufacturing to study the possible use cases for manufacturing. As shown in Figure 1, the conceptual framework of this systematic literature review includes three layers: data source, big data ecosystem, and data consumers.

The first layer, at the bottom of Figure 1, is the data source, which consists of five aspects:

1. Data types refer to the meaning of data such as temperature, humidity from the physical space, and log, email, operational data from cyberspace;
2. Source devices to collect data sources, which include sensors, controllers, actuators, software systems. The data type and source devices have a close relationship with the first four characteristics of big data (Volume, Velocity, Variety, Veracity) [39]. For example, in order to know the temperature of a production line, a temperature sensor is selected with the determined sampling rate (speeds of data generation), the sizes of data accumulating by time, formats and quality of data from the sensor.
3. Data dynamics describe the states of data. Data-at-rest refers to the inactive data stored in spreadsheets, databases and data warehouses; while data-in-motion refers to the active data generated by sensors, equipment or machines, and fed into the big data ecosystem in real-time.
4. Data formats are structures of the data. Data is exchangeable among various systems with consistent data formats and languages.

The system aspect of data source refers to the system where data is originated. There is a diverse range of manufacturing systems used in different applications such as product design, manufacturing

pyramid, product lifecycle management, supply chain management, logistics. The second layer, the middle layer shown in Figure 1, is the big data ecosystem comprising all the big data software. This layer plays the role of connecting the data sources from the layer below and the big data analytics applications at the layer above.

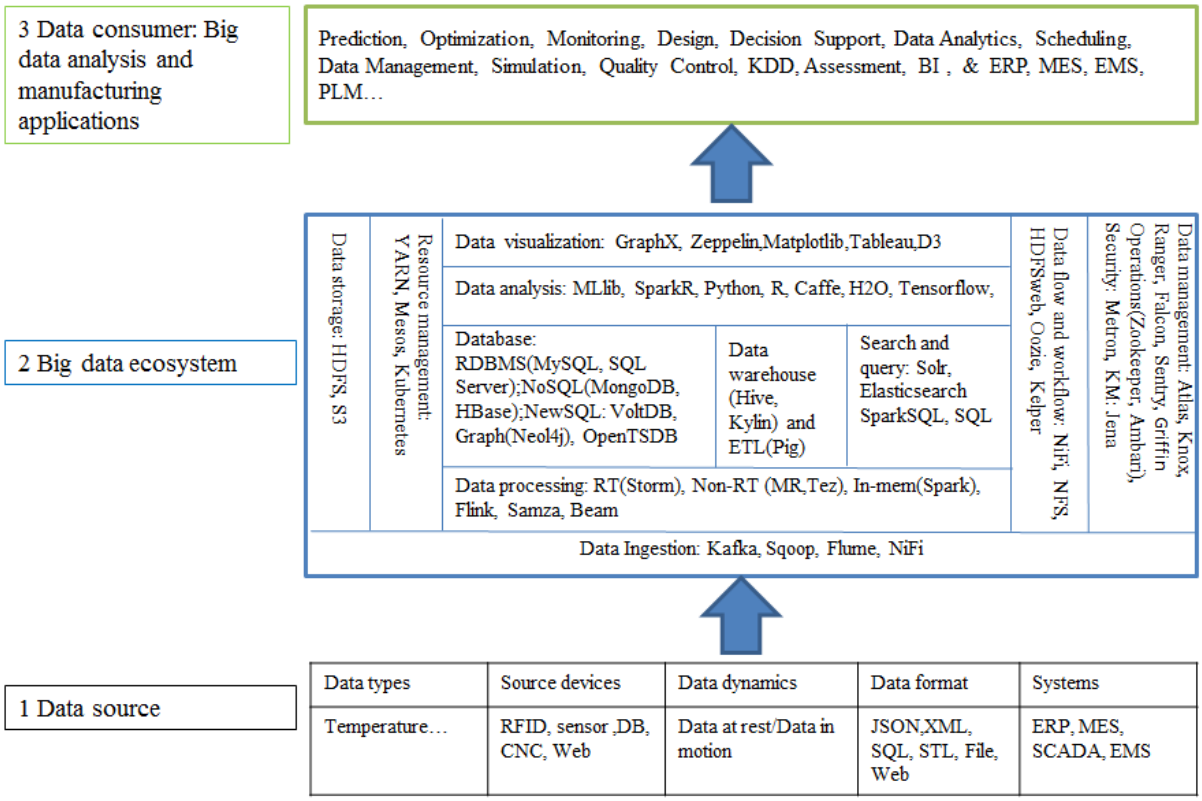

Figure 1. Conceptual framework of systematic literature review

The big data ecosystem is a set of complex and interrelated components to process and analyse big data [40]. Also, the ecosystem needs to store data from various data sources for data integration and analytics as well as other applications. Therefore, data storage layer in the ecosystem includes database and file system technologies to store big data. The ecosystem consists of the following components and tools:

- Data collection and ingestion: log data collection (Flume), bulk data collection from a relational database (Sqoop), distributed messaging system (Kafka), dataflow (NiFi);
- Computing engines: batch processing (MapReduce), iterative/near real-time processing (Spark, Flink), real-time processing/streaming (Storm, Flink) [41];
- Database: Relational database (RDBMS) has a standard schema but without scalable capabilities (MySQL, Oracle DB, SQL server, ProgresSQL); NoSQL database does not have a standard schema and has scalable capability (four types NoSQL: Column-based: HBase; Document-based: MongoDB; Key-value-based: Redis; Graph-based: Neo4j); NewSQL database is scalable relational database (VoltDB) [42][43]), search engine (Solr, Elasticsearch) ;
- Data analysis (BDA): Machine Learning (MLlib, Caffe, Tensorflow, Python), statistic (SparkR, R), OLAP,
- Data visualization (Zeppelin, Matplotlib, Tableau, D3 [44], GraphX;
- Workflow which is a scheduler of the jobs of various big data tools and dataflow which manages data transfer and data transformation among different big data tools: Oozie, Kepler, Apache NiFi;
- Data management and KM: Apache Falcon, Apache Atlas, Apache Sentry, Apache Hive, Operation (Zookeeper, Ambari), Apache Griffin, Apache Ranger, Apache Jena;

- Big data infrastructure (BDI): computing resources (general purpose computing and HPC), cluster management (YARN, Mesos) [37], network communication (Software-Defined Networks (SDN) [45], InfiniBand [46], 5G [47]) etc.;
- Big data security: Apache Metron [44], Apache Knox;

Although Hadoop is a big part of the big data ecosystem with many big data tools [48], it lacks functions such as data flow, data management and security [41].

Finally, the top layer represents the way data is used and data users. It includes the applications of big data analytics [26] and manufacturing applications [36]. The databases of traditional manufacturing software maybe the data sources of the ecosystem as well, such as SCADA or EPR.

2.1 Literature identification

Pertinent articles are identified within the scope of manufacturing data, big data technologies and big data based solution in manufacturing. Figure 2 illustrates the research method in this literature review. First, four citation databases are chosen due to their comprehensive coverage and high relevance to the scope: Scopus, IEEE Xplore, ASME Digital Collection, and ACM Digital Library. Second, several big data technologies and popular manufacturing data collection tools are selected as keywords. Hadoop is chosen since it is the earliest big data technology which is well studied and used. The underlying technologies of big data are computing and storage. There are a few big data computing engines. Three common ones are selected: Spark, Storm and Flink. For storage, because there are hundreds of available databases, it is not appropriate to limit to specific databases. All the databases could be categorized into three types: SQL, NoSQL and NewSQL. We select SQL and NoSQL because NewSQL could be recognized as SQL database with better features than traditional SQL. The features of NewSQL will be discussed in the following Sections. Time series data widely exist in manufacturing such as machines, sensors, controllers. Time-Series database (OpenTSDB) is selected since it seems more suitable for manufacturing data. Two widely adopted manufacturing data collection tools are selected: OPC-UA [49] and MTConnect[14].

Third, to further focus on the nature of this research paper, the papers were filtered by using the abstract “Manufacturing”, “Industry 4.0”, “Industrial automation”, “Smart manufacturing”, “Digital twin” and “Digital thread”. Four, manual review is implemented to select the papers about manufacturing data issues, or big data based solutions and applications in manufacturing. For example: some papers of other industrial sectors are found since they merely mentioned manufacturing in the abstract, such as Oil and Gas, Healthcare, Energy and Agriculture. Hence, the 339 articles are reviewed and 128 relevant articles are selected. The search strings of four databases are listed in Table 1 in Appendices.

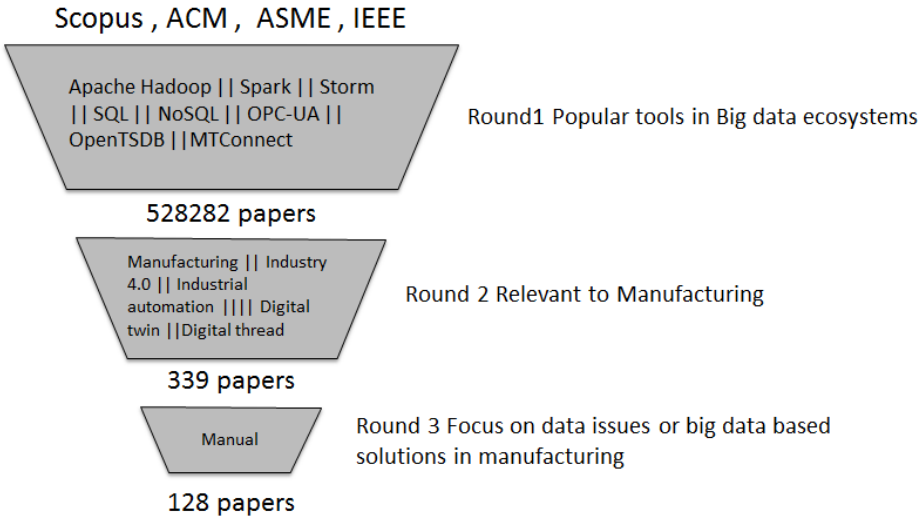

Figure 2. Process of Literature review methodology

3 Results

3.1 Manufacturing systems

In 2016, NIST reported three dimensions of concerns in smart manufacturing systems (SMS): product, production and business. Many traditional manufacturing systems and software can be categorized into one of the dimensions [49] (Figure 3). Business dimension is presented in the upper rectangle block with dash lines, which includes suppliers, customers and manufacturing enterprises (SCM, CRM, BI, asset management). Product dimension is located at the bottom rectangle block with solid lines. It includes objects and activities from product design to end-of-life of the product (CAx: CAD, CAM, CAPP and CAE, PLM). Production dimension is the triangle block, which includes an entire production system (ERP, MOM/MES, SCADA/DCS/HMI, O&M, Safety, quality management). Industry Internet of Things (IIoT) and RFID technology (grey blue) are widely used in manufacturing, logistics [50], in-use and product End-of-Life (EOL) [51].

Afterwards, all the reviewed articles are classified into four categories: Product, Production, Business and ICT (Information Communication Technology). The first three categories focus on engineering functions and business, ICT architecture underpins all three dimensions to provide the ICT infrastructure and digitalization to manufacturing, which includes several topics: CPS, Cloud manufacturing (CM), ICT, Data analytics/Data management(DM), KM. Table 2 in Appendices illustrates the distribution of reviewed articles by these four categories.

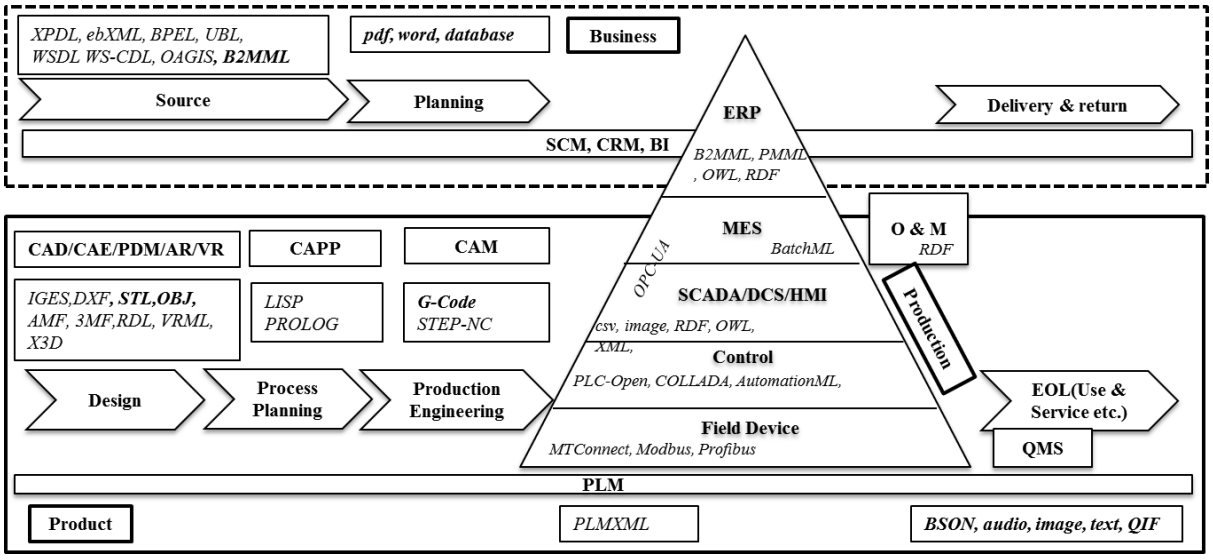

Figure 3. Smart Manufacturing Systems and various data formats

3.2 Data source

3.2.1 Data format

Based on the three dimensions of SMS in NIST report, the standards of data formats, computer languages are listed with the italic style in every manufacturing system in Figure 3. The bold black text represents some of the data formats found in the reviewed articles, whereas the black text is not found in the review but mentioned in the NIST report [49].

Table 1 demonstrates the complete data formats found from the reviewed articles and NIST report. To discuss the formats conveniently in the following chapter, all the data formats are categorized into three groups:

- Structured data: data that is presented in tables and can be stored in a relational database;

- Semi-structured data: data that has a self-described structure and is not presented in tables, such as XML, JSON, HTML [52];
- Unstructured data: data that does not have a self-described structure, such as document, image, audio, video, text and e-mail.

Figure 3 and Table 1show that one challenging issue to realize that smart manufacturing requires different data formats from various manufacturing systems. In order to make these systems collaborative and integrated, the transformation of these data formats is an essential function to the manufacturing big data solutions. It also illustrates that some data formats in the specific manufacturing systems are missing in the proposed bid data solutions. It requires solutions to fill the gap to realize data exchange among the systems. Big data tools can address the variety issue of manufacturing data such as NiFi.

Categories	Systems	Structured	Semi-structured	Unstructured
Product	CAD/CAE/CAPP/CAM	[53]	XML [54], G-code [55], STL [56], (aml, obj, UML, AutomationML) [57], **IGES, DXF, AMF, RDL**	○
	PLM	[58]	(XML, B2MML) [59], PMML[60], (RDF, SPARQL, STEP, QIF, STL) [61], PLMXML [60]	pdf [61]
Production	ERP	[62]	(XML, HTML, SCUFL) [63], **PMML**	e-mail [64]
	MOM/MES	[65][66]	JSON [67], RDF [68], AutomationML [69], **BatchML**	○
	SCADA/DCS/HMI	[70]	(RDF, OWL, XML) [71], SPDML [72]	Image [73]
	IIoT/CNC/Robot	○	(XML, UML, AutomationML) [69], OPC-UA, PLC Open, COLLADA) [74], BSON [75], JSON [76]	image [77][78],
	O&M	[79][80]	cad [80], RDF [81], XML [82]	Image [83], video [82]
	QMS	[84]	BSON [85], (XML, QIF) [86]	image [87], (audio, document) [88]
	Safety	[89]	○	○
Business	SCM//BI/AM	[90]	(XML, JSON, RDF) [91], (WSDL, EPL) [92], **XPDL, ebXML, BPEL, UBL, WS-CDL, OAGIS,**	document [93]
ICT	CPS/CM/ICT	[94][95]	(JSON, STEP, JT Open) [96], HTML [97], (AutomationML, PLCopen) [98], (XML, XSD, RDF) [99], (UML, SysML, STEP, B2MML) [5], EDDL [100], [101]	e-mail [64]
	Data analytics/DM	[102]	BSON [102], JSON [103], Parquet [104]	(image, video, document) [105]
	KM	○	(OWL, UML, RDF, SWRL) [23], (JSON, PMML, AMPL) [106], (STEP-NC, G-code, XML, DMIS, QIF) [107]	○

○: Not found

Text: Not found in reviewed articles (mentioned in the NIST report)

Table 1. Data formats in reviewed articles and NIST report

3.2.2 Data issues

Data issues are fundamental challenges to smart manufacturing, which extract actionable information from good quality of data. In order to prepare the suitable data for smart applications, amount of cost

and time is consumed to address the data issues. For example, data scientists spend over 90% of their time on data preparation before analysing data for innovative tasks such as machine learning, AI [108]. Enterprises spend billions of dollars on their data warehousing systems, which can only use well and pre-defined methods (ETL) to process product structure data and produce business reports in non-real time. Therefore, understanding and addressing data issues is critical to design big data-based solutions. From Table 2, we have identified 11 common issues of manufacturing data. Data issues are generally related to the Big data 5Vs features[38]: volumes and variety (large scale); velocity (inconsistent sampling frequencies or timestamp, batching and streaming data); veracity (missing value, imbalance, data outlier, noisy, drifting, asynchronization, data correlation); variety, veracity and value (data model, data format exchange and data integration). The researchers of these kind of literature address these issues without using big data tools. Traditional software cannot address these data issues if the data is generated in large scale systems with large number of devices. Some big data tools are discussed to provide potential solutions in Sections 4.2.7 and 4.4.2.

Issues	5Vs	Description	References
Large scale	Volume, Variety	Large volume dataset with massive features.	[109]
Inconsistent sampling frequencies or timestamp	Velocity	Sensors use various sampling frequency and timestamps; Unnecessary high sampling frequency affect the real time performance, it is related with “Smart data” topic.	[110][111][112] [113]
Batching data and streaming data	Velocity	Data modification is sensitive to time or not, some literature also uses the terms: data-at-rest and data-in-motion [13];	[114]
Missing value	Veracity	Record is empty when equipment is an anomaly; Product passes some portion of machines; Sensor is an anomaly or losing communication with sensor; Too costly to capture data by installing sensor or building models.	[115] [109] [110] [113]
Imbalance	Veracity	Small probability data in a very large dataset with most of normal data.	[109]
Data outlier	Veracity	Data is out of range of measurement device.	[110][84]
Noise or an anomaly data	Veracity	Data is out of the similar clustering dataset; Noise data is possibly generated by replacing the missing data.	[115] [109]
Drifting data	Veracity	Process drift caused by vulnerability to external environment; Sensor drift caused by modification in measuring device or calibration.	[110] [113]
Asynchronization	Veracity	Several data producers (sensor and machine) use non-central time server in manufacturing enterprise.	[113]
Data correlation	Veracity	Nature and structure of data caused by redundant sensor arrangement; Data correlation is to improve data quality with process variables. Inconsistent simulation and collection of data from CNC machine.	[110] [116] [107]
Data model, data format exchange and data integration	Variety, Veracity , Value	Merge data from multiple data sources into a single view; Exchange different data formats from various systems: OPC-UA and IoT, OPC-UA and AutomationML, MTConnect with QIF, MTConnect and IEEE 1451 wired smart transducer Integrate data models to explain the relationships of data; Data information must be available for information sharing. OWL is the enabling technology on resource description.	[115] [117],[98] [86], [118], [113] [99][118]

Table 2. Data issues in manufacturing

3.3 Big data ecosystem

Through this SLR, new big-data based research innovation could be identified by closely tracking the attention of various big data tools in manufacturing and other research domains. Figure 4 shows the distributions of big data tools over the years from the literature, including manufacturing engineering and others. Fig 4(a) shows the numbers of the term Hadoop occurred in all literature increased from 2008, reached its peak in 2016 and slightly decreased in 2017. The number of Spark articles had been higher than Hadoop since 2016. There are four factors to compare Spark and Hadoop: volume, velocity, fault-tolerant and data analysis. MapReduce and Spark Streaming are the computation engines of Hadoop and Spark, respectively. MapReduce executes batch processing by reading and writing data on disk multiple times. Spark Streaming executes micro-batch processing in memory. It results in the differences between Hadoop and Spark because disk can persistently store a larger volume of data with slower velocity than memory, which temporarily stores limited volume of data. By comparing data analysis, Spark has built-in tools (MLlib) and support third-party tools (Mahout, H2O), whereas Hadoop is only supported by the third-party tool: Mahout [41]. Spark supports iterative computation with GraphX, which is the graph processing engine. Therefore, Hadoop is suitable for the applications which need planned extraction of non-real time and critical information from a larger volume of data and guarantee without loss of data, such as ERP, Production planning. Spark can be used to provide near real-time monitoring and analytics by processing streaming data such as monitoring process and product quality, MES, SCADA, predictive maintenance. The number of Storm articles has kept increasing since 2012; however, its total number is smaller than Hadoop and Spark since it only executes streaming with limited data analysis supported by SAMOA, which is a version of Mahout [119]. Fig 4(b) presents the growing number of literatures about these three tools in the manufacturing research literature since their first inception.

From Fig 4(c) and Fig 4(d), NoSQL database, as the highest frequently mentioned database, is compared with NewSQL and time series databases (OpenTSDB) over the last eight years. Fig 4(e) also demonstrates that Kafka and OPC-UA are closely related to manufacturing as seen by their increasing patterns. Because OPC-UA protocol provides open connection to monitoring and automation systems as well as communications between MES and SCADA systems, these systems are prevalent in manufacturing [120]. Kafka is also used widely in applications at the shop floor level of manufacturing such as processing, machine and sensor data, due to it is message streaming capability [121]. Apache NiFi is entirely new to manufacturing. Only four articles were found in the industry in general and no article in manufacturing industry [122][123][124].

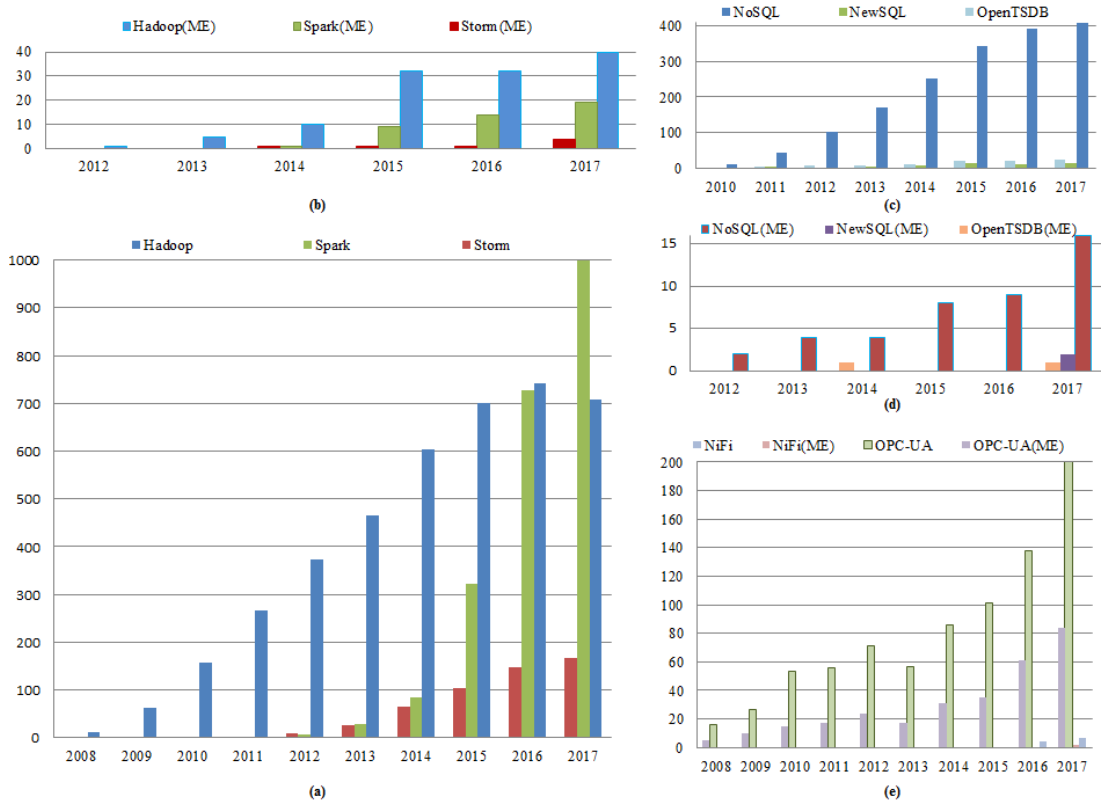

Figure 4. Chronological distribution of big data tools

3.4 Applications of big data in manufacturing

Data applications are essential to realize Smart manufacturing. Identified scopes of data applications could provide a clear guideline to design manufacturing big data platforms. Among the reviewed 128 articles, 78 articles are big-data based applications, which are categorized into17 applications. The percentages of the applications are presented in Figure 5. Monitoring (25%), prediction (23.8%), ICT framework (11.9%) and data analytics (9.5%) are the four most frequently used big-data applications in manufacturing. Through this statistic, it can be shown that the researches of big-data based solutions focus on monitoring, prediction, data analytics and propose ICT solutions in manufacturing in Table 3.

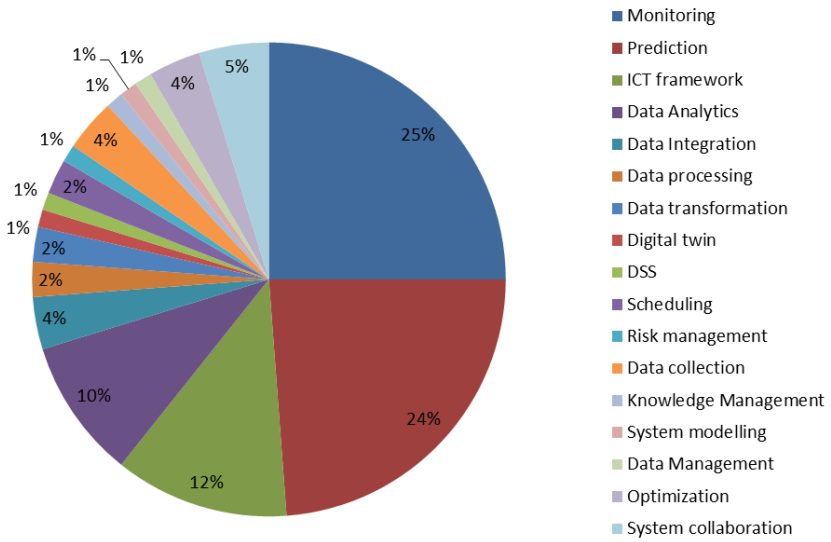

Figure 5. Percentage allocation of Big data applications (Based on our literature investigation)

Field	Systems	Big data framework	Big data computing and storage	Big data analytics
Product	**CAD/CAE/CAPP/CAM**		[57], [56], [125],	Regressions [53], ANN [53],
	PLM	[58],		
Production	**ERP**			MP [63], regression [62], K-means [126]
	MOM/MES	[69],[116],	[66],	regression [127], Distance, Regression, Self-organizing map, principal component analysis [128],
	SCADA/DCS/HMI	[129], [71],	[104],[130],[131],[104],[132],[112]	Classification [133], OPL [106], KM [134], GA [73],
	O&M	[39],[135]	[136],[81],[82],[80]	logistic regression, naïve Bayes, and a decision tree [109], regression [137], LSM [138], SVM [83], Anomaly detection [139], DTW [140], RF [141], K-means, Markov [142], KD [24],
	QMS			SPC [87], ANN, decision tree, random forest, SVM [143],[84]
	Safety		[144],	graph analytics [89],
Business	**SCM/CRM/BI/AM**	[93], [90],	[145]	
ICT architecture	**IIoT/CPS/CM/ICT**	[64],[146],[76],[147]	[68],[148],[95],[149],[150],	
	Data analytics/DM	[151],[33]	[152],[94],	Stream data analytics [103],
	KM			KM [106], Semantic data integration [23],[153],

Table 3. Allocations of proposed solutions of reviewed literatures (Based on our literature investigation)

4 Discussion

Four fundamental questions about the relationship between big data ecosystem and smart manufacturing are six drivers and requirements for big data application in manufacturing, seven essential components of the big data ecosystem, harnessing big data capabilities for research innovation in manufacturing, and future directions of big data application in manufacturing. They illustrate the driving factors of big data applications in manufacturing. These are summarized in Table 3 and Table 4 for ease of reading to interested readers.

Topics	Sub-topics	Application systems	Enabling tools	References
6 drivers for big data in smart manufacturing	System integration	Product design, AM, ERP, MES, BI, SCM, PLM,	Kepler, Hadoop, OPC-UA, RESTful API	[49][57][56][125][63][62][90][153][97][55][128][154][155]
	Data	Predictive maintenance, KM, Production planning, Safety, Anomaly detection, industrial process control, model prediction, QC, shop floor scheduling,	Cassandra, MongoDB, Blueflood, OpenTSDB, DalmatinerDB and InfluxDB, Storm, Spark, Flink, Apache Hive,	[121][62][73][156][157][82][138][158][24][89][94][153][159][75][160][88]
	Prediction	3D printing, product performance, production planning, energy consumption, MES, QC, SCM	Random Forest, Bayesian Network, statistic	[39][109][63][62][93][136][137][138][141][24][87][143][90][55][161][85][88][25][162][79]
	Sustainability	PLM, Maintenance		[163][59]

	Resource sharing and networking	SCM, ERP, Data integration	Public cloud, Private cloud, Hybrid cloud, hypervisor, container,	[19][63][62][144][95][152][97][101][164]
	Low cost hardware	SCM, Industrial automation	RFID, IoT, Robotic	[65][165][166][167]
9 essential components of big data ecosystem	Data ingestion	ERP, MES, SCADA, O&M, QC, PLM, Data management	Sqoop, Flume, Kafka,	[39][116][58][71][138][139][140][151][85][79][65]
	Storage	ERP, SCM, SCADA, OLAP, OLTP	Redis, HBase, Cassandra, MongoDB, Neo4j, HDFS, VoltDB, Clustrix, NuoDB	[30][32][43][168][169]
	Computation	ERP, SCM, PLM, MES, SCADA, O&M, QC, IoT	MapReduce, Spark, Flink, Storm,	[170][171]
	Analytics	DSS, CRM, machine vision, QC, O&M	MLlib, Scikit-Learn, CNTK, Caffe, Kylin, CaffeOnSpark, Hive, CaffeOnSpark	[109][73][82][87][25][105][172][173]
	Visualization	MES, SCADA, O&M, QC, Security,	Zeppelin, Tableau, D3.Js, Matplotlib, QlikView	[126][76]
	Workflow and dataflow	Business	Oozie, Kepler, InfoSphere, Wings/Pegasus, NiFi	[116][123][63][172]
	Data management	ICT, KM, SCADA, SCM,	Apache Falcon, Apache Atlas, Apache Sentry, Apache Griffin, Jena	[121][23][174][175][176] [177][178][179]
	Infrastructure and deployment model	ICT	HPC(AWS),	[180][181]
	Cybersecurity	SCADA, ICT		[182][183]
5 future directions	Modelling and simulation	VR/AR, PHM, PLM, CAx	Cloud computing, Quantum computing	[184][185]
	Connectivity and interoperability	File formats of PLM (PLM XML) and 3D printing (AMF, 3MF)	NiFi,	[186]
	Standardized big data platform design			[5]
	Real time big data analytics	SCADA, MES, Data warehousing	Spark, Storm, Flink, Beam, Spark R, MLlib, GraphX, SparkSQL	[26][187]
	Cybersecurity	SCM, SCADA, Security, Safety	Apache Metron, Apache Ranger, Apache Knox	[49][188]

Table 4. Summary of Discussion section

4.1 Question 1: What are the drivers and requirements for big data applications in smart manufacturing?

Identifying drivers for big data applications is essential to implement feasible smart manufacturing initiatives. Kusiak discusses the future developments in manufacturing and identified six drivers for smart manufacturing theoretically [4], which are manufacturing technology and processes, material, data, predictive engineering, sustainability, resource sharing and networking. Through reviewing the proposed big-data based solutions, six drivers are identified: data, prediction, sustainability, resource sharing, system integration, and low-cost hardware. Four common drivers from Kusiak's six drivers are verified with big-data based solutions.

4.1.1 Driver 1: System integration

System integration with big data technologies is a crucial enabler of smart manufacturing to integrate and cooperate manufacturing systems to timely adapt dynamic demands from production and supply chain [49]. Integration of production systems demonstrates the significant improvement in production efficiency and productivity since the 1980s [4]. In the context of smart manufacturing, it needs to

further expand the scope of system integration from production to product and business domains. Figure 3 shows that various manufacturing systems use different networks and protocols, which are challenging due to the necessity of implementing data and information exchange among these systems. Big data technologies can integrate these independent systems by using "cloud" as a common place to collect data, extract and exchange the required information on the cloud. With the fusion of IoT, BDA not only can integrate manufacturing systems but integrating physical and cyber worlds closer [36]. From the system integration in Table 4, many benefits of system integration are demonstrated with the proposed big-data based solutions.

Integrating product design and additive manufacturing with big data provides many benefits. With the integration of product design and additive manufacturing, product costs can be estimated by analysing more features of product model with DBA than the traditional approach [55]; using Hadoop clusters demonstrates faster velocity of converting a huge 3D model to G-code for 3D printer than traditional methods [56]. Spark and Cassandra demonstrate capabilities of computing and storing a large volume of streaming data from the application of real time monitoring 3D printing, which traditional manufacturing systems cannot offer [125].

As to production, big-data based solutions drive the integration of ERP and MES. The big data scientific workflow management system (Kepler) demonstrates the efficient scheduling capability for smart manufacturing [63]. Another paper concludes that the critical cycle time can be predicted with Hadoop for production planning [62].

In the business domain, several researchers focus on designing big data architecture to integrate BI, SCM, ERP, MES and PLM systems. Although a business intelligence architecture is proposed to advance the integration of business information from various existing systems such as ERP, CAx, SCM, PDM/PLM, the specific technical framework is not provided [90]. It demonstrates that it is necessary to have a capability to store data and information with various data formats, structures and models to integrate various systems. Another big-data based solution is proposed to integrate supply chain and production planning by retrieving and integrating SCM, ERP and MES data [97]. However, the framework focuses on designing business functionality without providing support to process the collected big data. A cloud manufacturing collaboration system is proposed to achieve better performance and functionality on production, resource planning with the Hadoop ecosystem. It is concluded that information integration from the web and other data sources is a critical issue to implement system collaboration [153]. Processing and analysing data from MES and SCADA with MapReduce and BDA can detect anomaly minutes beforehand in a large-scale production [128]. In the industry, Bosch presented a conceptual, analytic platform with a data integration method to integrate various data sources [154].

System integration with BDA requires standard data format and standard interfaces. New standards of file formats need to be developed in order to fill the deficiencies of the existing standards to transfer consistent content [49]. Additive Manufacturing File (AMF) is a new XML based standard format to replace STL by providing many new features in addictive manufacturing such as materials, material properties, colours. [49]. Because RDBMS is not suitable to store data with flexible data models (unstructured or semi-structured data), big data technologies are thus mainly used to process unstructured data such as Hadoop, NoSQL databases. The standard interface is of importance to seamlessly integrate systems in manufacturing. Because OPC-UA technology provides dedicated interfaces to production equipment such as PLC, it received increasing attention in the industry [128]. RESTful API demonstrated to be more efficient to connect web data by comparing with the traditional method SOAP [155].

4.1.2 Driver 2: Data

Timely comprehensive data with enabling big data tools is the key driver to smart manufacturing. The data row in Table 4 presents many manufacturing applications could be implemented with more comprehensive data and big data tools, which were challenging with traditional tools. Firstly, a large volume of data collecting from various sources provides sufficient data for big data analytics. Data from various industrial equipment, IoT devices, web and smartphones, is called data-in-motion [159],

which is continuously generated and ingested into the systems to provide real-time response from the physical world, such as dynamic shop floor scheduling [75], predictive maintenance [141][136][24], anomaly detection [81], diagnosis [82], prognosis [161] and systems collaboration [153]. As another significant data source for big data ecosystem, databases in manufacturing is data-at-rest, which represents static, historical data [159]. This data is mainly used to predict the long-term performance in production planning [62], global manufacturing network design [94], critical event detection in safety [89]. Both historical batching data and real-time streaming data are integrated to train models and monitor real time condition information such as anomaly detection of machines' energy consumption data [139].

Secondly, big data technologies make the management of manufacturing big data feasible. Traditionally, RDBMS is mainly designed to store structured data with limited scalability. However, NoSQL databases present better performance for handling semi-structured (JSON, XML) and unstructured data (audio, video, and email) with unlimited scalability. For example, column NoSQL database Cassandra was used to store event data of automation controller [131], document NoSQL database MongoDB was used to store machine data [85]. Time-series databases (TSDB) begin to receive increasing attention by providing dedicated applications for sensor data. A comprehensive evaluation was implemented to several TSDBs: Blueflood, OpenTSDB, DalmatinerDB and InfluxDB [82]. The collected data needs cleaning before usage in order to resolve issues such as noisy and incorrect format, as shown in Table 2. Streaming (Flink, Storm), micro-batching (Spark) and batching (MapReduce) data processing technologies provide the capabilities to clean and calculate big volume of manufacturing data. The following big-data based solutions are proposed and implemented in various manufacturing applications: complex event processing (CEP) with Storm[138], anomaly detection with Flink [121], industrial process control with Spark [73], model prediction [137] and quality control with MapReduce [88].

Lastly, past and new knowledge can emerge from the generated big data by harvesting big data technologies. Knowledge of predictive maintenance can be extracted with an Apache Hive-based platform [24]. Knowledge of intelligent applications of a smart factory is managed with Hadoop and OWL technologies [134].

4.1.3 Driver 3: Prediction

Prediction enables manufacturing to change from reaction to prevention. Because of big data and increasing applications of data analysis in manufacturing, it is feasible to predict the behaviours of various manufacturing systems accurately.
Prediction attracts many researchers' attention to manufacturing. From Figure 5, the prediction is the common BDA application in manufacturing. In the product domain, the costs of 3D printed products can be predicted with the proposed big-data based solution and three machine learning algorithms [55]. In the production domain, machine learning is used for the prediction of product performance degradation [63], cycle times for production planning [62], energy consumption and KPI values on MES system [25], and production efficiency [162]. A number of papers focus on the prediction of product quality with different machine learning algorithms such as Bayesian Network [88] and statistic analytics [87]. ANN is identified as the highest prediction accuracy by comparing with a decision tree, random forest and support vector machine [143]. Some applications need a trade-off between acquiring higher accuracy and shorter calculation time since it takes more time to calculate to get higher accuracy, such as real-time quality control. In terms of the full consideration of shorter calculation time and higher accuracy, Random forest is outperformed Naive Bayesian, Multi-Layer Perceptron and Logistic Regression [85] [109]. Many papers publish the solutions of predictive maintenance on machines by using machine learning algorithms to recommend scheduling of proactive measures before outages occurred [137][138][24][39][79][161][136][141]. In the business domain, some potential applications are found such as predicting user behaviour with using principal component analysis and Hadoop [128], proactive inventories, location and throughput times on logistics with a big data based platform [90].

4.1.4 Driver 4: Sustainability

Sustainability with big data technologies plays a vital role in smart manufacturing. Sustainable manufacturing considers the four factors: material, manufacturing processes, energy and pollutants. Big data technologies can provide a data-driven solution to analyse the big volume of data for these four factors. For example, product design could be guided by analysing the End-of-Life data of products; strategic decision making by analysing marketing, production and supply network data from CRM, ERP, MES and SCM systems; energy consumption and pollutant influence would be monitored through IoT sensors and RFID tags. Although several conceptual big-data based frameworks are proposed as shown in Table 4 [58][59], the performance of sustainability with big data technologies has not been evaluated. Further research is required to discuss the effectiveness of big data technologies on sustainability in manufacturing.

4.1.5 Driver 5: Resource sharing and networking

Manufacturing could benefit by sharing virtual and physical resources with the supply chain network. The issue of information silo results in losses of productivity and economy in manufacturing.

- **Sharing information**

Sharing information is beneficial to manufacturing systems of product, production and business. Helu et al. discuss that digital thread can improve product design and manufacturing processes by sharing data of product lifecycle systems [14]. For instance, some information on idle equipment among enterprises can be shared with big data tools in order to reduce holding costs such as machines, 3D printers, equipment, transport and warehouses [101]. Sharing information about SCM and PLM systems could meet the new business requirements such as find best supplier of a specific raw material [164]. Integrated data from various systems could identify potential production problems and improve work efficiency by selecting suitable maintenance time [152]. Collection and sharing data of supply chains with big data could be helpful to interactions among customers, manufacturers, and suppliers [97].

Addressing information silo is challenging because data cannot be easily shared among traditional manufacturing systems by different protocols. Some practical solutions are implemented, such as sharing data in various systems, which leads to the issue of data redundancy. Manufacturers also have to take massive maintenance effort on synchronizing the shared data of every system. Because the big data platform could be sat on the cloud, it has to establish only one connection to synchronize data of the platform and the system. Through big data tools, data is much easier collected from various systems to the data lake of the big data platform, synchronized, managed and shared[121].

- **Sharing BDA Infrastructure and software**

Sharing big-data infrastructure and software brings economic benefits to manufacturing. According to NIST definition, there are four deployment models of cloud computing: private cloud, community cloud, public cloud and hybrid cloud [19]. From reviewed papers, 18 solutions use private cloud while only two solutions use public cloud. The main reason for adopting a private cloud is to provide better privacy and security than public cloud [62][63][144]. However, there is a significant investment on the hardware of private cloud clusters. Wang et al. demonstrate that manufacturing can gain economic benefits from the public cloud with three aspects: pay-as-you-go service model, reducing maintenance fee on data centres and operating cost [189]. No security function was identified in the 18 proposed solutions, which means the security and privacy of the private cloud solutions do not outperform public cloud. Therefore, using the public cloud in comparison to private cloud may bring significant economic benefits to manufacturing.

Virtualization technologies such as hypervisor and container, provide faster deployments, high efficiency to share software packages in enterprises [95]. Virtualization can be used to quickly test and validate the proposed big data solutions with minimum influence on the other systems.

4.1.6 Driver 6: Low cost hardware

Low-cost hardware makes smart manufacturing more accessible. Low-cost actuator and IoT sensor reduce the wiring cost to collect data and improve automation at the factory floor. Wang et al. present a collaboration mechanism to deploy large scale robotics with big data technologies at a small factory [65]. Big-data based solution with smart sensors elevates the constraints of time and geolocation to monitor manufacturing processes [165]. Low cost RFID tags make traceability of enormous resources more feasible at the supply chain level. Industrial Internet of Things hub is proposed to realize the smart connection of various resources in a manufacturing facility with RFID tags and sensors [166]. Cost competitive NC machines and 3D printers can be flexibly deployed to demand sides such as design prototype product or serving customers. Big data technologies can secure critical data transmission between design departments and manufacturing equipment. Furthermore, low-cost data processing and storage technologies provide a cost-effective approach to manage large volumes of data from massive data sources in manufacturing [167]. Hence, cheap hardware makes it more feasible for manufacturers to monitor and respond to timely changes from production, to supply chain networks.

4.2 Question 2: What are the essential components of big data ecosystem to better serve smart manufacturing?

As big data applications enable smart manufacturing, several essential components of the big data ecosystem should be utilized to build up BDA platform for smart manufacturing, including data ingestion, storage, computing, analytics, visualization, workflow and dataflow, data management, infrastructure and security.

4.2.1 Data ingestion

Data ingestion or inception is of necessity to manufacturing in order to bring big volume data into its BDA platform. There are two types of big data: data at rest and data in motion. Apache Sqoop is used to transfer bundle of data from a relational database (MySQL, SQL server) to Hadoop in several applications such as ERP and MES [116] [65], SCADA [71], O&M [39][79] and Data Management [121][151].

Streaming data is data continuously generated from manufacturing systems and devices. Apache Flume is mainly used to collect large amounts of logs from controllers, sensors, equipment and actuators [65][71] [121][116]. Kafka is a general-purpose messaging system to collect streaming data and publish it to data consumers who subscribe to the topic of data. Kafka is applied in some cases of SCADA [140], O&M[139]. Apache Storm is a real-time data processor, which is used to collect and ingest streaming data to data consumers straightway (O&M [138][139], Quality control [85], PLM[58]). Although some data collection tools of manufacturing are widely used in production systems such as MTConnect and OPC-UA, big data ingestion tools can complement their limitations as discussed in Section 4.4.2. Therefore, data ingestion is an essential component of the big-data ecosystem to collect batching data and streaming data.

4.2.2 Storage

Data storage is critical to big data applications in smart manufacturing. As the manufacturing industry increasingly benefits from the use of big data, it is of importance to store more data [4]. Various applications require different storage technologies to provide different features, which are file system and databases. Although both technologies can store structured, semi-structured and unstructured data, they have some differences that file system is suitable to store data-at-rest or unstructured data such as files, search and compile files manually. Database is suitable to store data-in-motion, semi-structured or structured data, faster query data automatically.

There are three types of databases: RDBMS, NoSQL and NewSQL databases. RDBMS database has been used in manufacturing for general purpose applications for decades such as SQL Server, Oracle,

MySQL. Whereas, RDBMS is unable to address the challenges of big data's 3Vs (Volume, Velocity and Variety) for storage and query [32]. NoSQL and NewSQL databases can provide almost unlimited scalability and faster query capability for industrial big data. Figure 4 illustrates that the increased use of a NoSQL database is already happening in manufacturing. NoSQL is suitable for one kind of OLTP application, which does not require consistent data all the time, but simple query and frequent updates to data [168]. Hence, manufacturing could use NoSQL databases for real-time big data analytics such as quality monitoring and prediction. Moreover, manufacturing could benefit from using the four types of NoSQL data models (Key-value, Wide column, Document, Graph) to easily manage semi-structured data (XML and JSON) [43]. The widely used NoSQL databases are Redis, HBase, Cassandra, MongoDB and Neo4j. Since NoSQL databases were not designed to meet data consistency, they are not suitable for some OLTP applications, where data consistency needs to be guaranteed anytime [169]. Therefore, NoSQL should not be used in an environment of many operations and controllers that are present since the control data may be inconsistent. The NewSQL solution is used to provide relational query (SQL) and data consistency all the time for an OLTP application. Manufacturing could use NewSQL for the scenarios requiring consistent data all the time (finance data in ERP, inventory data in SCM, control signal in SCADA systems). Some examples of NewSQL are VoltDB, Clustrix and NuoDB [43]. However, NewSQL focuses on relational data, which may not fully support unstructured data for big data analytics.

Unlike database solutions, in which the data is structured as a data model for data consumers' demands, file systems do not need a data model to store unstructured data. Hadoop Distributed File System (HDFS) could store petabytes of data with a redundantly low-cost method [30]. However, querying data in HDFS is much slower than the speed in databases. HDFS could be utilized as a central data storage to retain all the data in manufacturing. Therefore, selection of the correct storage solutions is required for big data applications in manufacturing.

4.2.3 Computation

Computation is the foundation of implementing big data applications. Three types of computation engines are available to manufacturing: batching, micro-batching and streaming. MapReduce in Hadoop provides the batching method to process a petabyte level of big data by less memory usage and cannot provide real-time analytics [170]. Spark is a micro-batching processing engine, which provides near real-time computation with more memory resources than MapReduce. Flink and Storm are real-time streaming engines to process small volumes of data [171]. The differences and application scenarios of these computation engines have been discussed in Section 3.3. Based on the analysis of these computation engines and outcomes of Figure 4, there would be more Spark-based big data solutions on the factory floor from device to production planning in the near future. The proposed solutions could use Spark to replace Hadoop to get faster outcomes.

4.2.4 Analytics

Big data analytics is intended to extract information from collected big data. Big data analytics includes two types of analysis: 1) data mining and machine learning algorithms (clustering, regression, Bayesian networks, artificial neural networks (ANN), deep learning. 2) On-Line Analytic Processing (OLAP).

Big data analytics tools of the first type are identified as MLlib and Scikit-learn for machine learning [25][109], Spark R for high-level statistical analysis [73], Tensorflow and CNTK for deep learning [105]. Analytics tools have many potential use cases in manufacturing, such as machine vision for robotics[105], speech recognition for alarm and security, image processing for quality control [87]and O&M[82]. Manufacturers can save time to develop algorithms from scratch by using these tools in their big data platforms, such as image recognition for product quality control[87]. Manufacturers also benefits from new tools such as CaffeOnSpark [105], which is a deep learning framework widely used for autonomous driving in the automotive industry[105]. Because Caffe does not work on Spark clusters, it is challenging to meet the two strengths of better data analytics algorithm and faster big data computation at the same time. CaffeOnSpark could address the issue to work on Spark and Hadoop clusters for Caffe applications.

OLAP is an approach to analyse large multidimensional datasets for complex business analytics, such as BI reporting, Decision Support System (DSS), and CRM in manufacturing. Analytics tools of OLAP are Apache Hive [172] and Apache Kylin [173]. Hive is one utility of Hadoop ecosystem which is used as a data warehouse. Kylin can query a large volume of data faster than Hive.

By analysing the collected big data, data analytics tools can efficiently extract timely information to manufacturers to make decisions. It is challenging to the decision makers to apply their experience and knowledge on the new circumstances. Their experience is acquired under the previous circumstance, which may be different from the current one and their knowledge may be out of date. With the given data and big data analytics tools, manufacturers can analyse historic data, discover new knowledge, build actionable intelligence to make data-driven decisions. It would happen in some areas of different systems feed with new data such as production planning with streaming real-time IoT data.

4.2.5 Visualization

Manufacturing systems require various visualization methods to present analytics results, such as interactive dashboard, reporting, graph, document. Most of the proposed solutions did not provide a tool to construct visualization. As the Python language is widely used in big data analytics, it should be convenient to use Python plotting library (Matplotlib) to present big data analytic results by data scientists in manufacturing. However, Matplotlib works on a command line interface (CLI), which is not user-friendly to business users with less programming experience. The reviewed solutions could benefit from visualization tools such as Zeppelin, Tableau, and D3.Js[76]. Zeppelin provides interpreters with big data tools (HBase, Cassandra, HDFS, Spark, Flink etc.) and supports multi-agents. Manufacturers could use Zeppelin in their existing big data platform with less developing effort. D3.Js provides more complex visualization templates than Zeppelin. D3.Js is a JavaScript library for producing dynamic, interactive data visualizations in the web browser. It is suitable for applications, which require dynamic monitoring and control such as MES, SCADA, O&M, Quality control, safety and security systems[126]. D3.Js has less support to produce a report, which is not suitable for business analyses. Tableau and QlikView are commercial data visualization software focusing on BI and support many common databases. . Big data and visualisation can help bring data together and show the value of big data in meaningful ways. AI helps to make automatic decisions, and visualisation helps to make manual decisions. As a result, AI and human can collaboratively make data-driven decisions based on the actionable intelligence generated by big data or mined by machine learning.

4.2.6 Workflow and dataflow

Workflow and dataflow components provide efficient approaches to manage workflows for business and processes to data management. Two workflow tools are found in the review papers: Oozie[116] and Kelper [63]. Oozie is specialized in managing Hadoop jobs, which would be suitable for a manufacturer with Hadoop ready platform [174] . Kepler provides a convenient method to share workflows with other business users in cloud manufacturing [63]. However, Kepler does not support horizontal scalability to store large volume data. Another two workflow tools Wings and IBM InfoSphere have not been found through this review [175]. Wings/Pegasus provides a more intelligent method to construct and execute workflow for users by optimising the workflow automatically. For parallel executing of workflows, Wings has to work with a resource management framework (Pegasus), which may be a constraint to manufacturers with Hadoop-based platform since Hadoop uses Yarn as its cluster resource manager. IBM InfoSphere is a commercial enterprise-ready framework, which is suitable for inexperienced users. It can work with Hadoop and other IBM software together. InfoSphere is not open source, which is unable to build add-on functions. For the dataflow component, researchers had to construct their dataflow component from scratch. Usually, there is a long learning curve for beginners to go through trial-error processes. Apache NiFi is an efficient tool to construct dataflow and data integration with an interactive GUI. One paper uses NiFi to collect streaming data in process industry[123]. NiFi is initially developed and used by the National Security Agency of the United States (NSA), which has been verified in the real application

environment. Through this study, there is limited work reported in this area. Smart manufacturing needs the workflow and dataflow components to get data automatically processed among various manufacturing systems. If assisted by AI and optimization, they could significantly improve the efficiency of workflow and dataflow in manufacturing such as production planning and scheduling.

4.2.7 Data management

Data management would make big data platform more feasible for manufacturers. Data management focuses on data governance, metadata management, data modelling, data quality management, master data management, data integration and knowledge management. Although data management is highly related to the traditional IT area, it provides a holistic management method to meet the requirements of enterprise compliance, data policy, data lifecycle.

Some data management tools are available to the big data ecosystem of manufacturing:

- Data lifecycle management is based on enterprises' policy to manage data lifecycle from creation, storage, obsolescence to delete. Apache Falcon includes the functions of data retention (persistency), data replication for disaster recovery, aggregation and archive on Hadoop[176]. Falcon is beneficial to manufacturing research. For example, each acoustic experiment of aerospace engine collects hundreds of gigabytes of audio data through over a hundred sensors. The traditional method uses a single disk to store them, which is costly to manage with high risk of disk failure or data loss. Falcon could manage the data lifecycle with a friendly user interface.
- Data governance provides an enterprise policy approach to manage data availability, usability, and integrity. Apache Atlas is the Hadoop ecosystem tool for audit, lineage and service level agreement (SLA), which is used to apply agile enterprise compliance through consistent metadata management across the big data ecosystem [177].
- Data authorization manages users' privilege to access sensitive data. Apache Sentry is utilized to authorize data and metadata on Hadoop clusters based on the roles of people in manufacturing [121].
- Data quality management is crucial to the outcomes of data analyses. Table 2 presents that data quality is a challenging data issue to manufacturing. Apache Griffin maintains data quality by automating data profiling and validation on Hadoop and Spark[177]. Apache Griffin improves data quality by pre-processing data automatically from various data sources, which reduces the amount of data analyst's time to prepare data for analysing.
- Data integration has two approaches to address data silos issue [178]: Data Warehouse (DW) and Data Lake (DL). DW is the traditional approach that integrates data from various data sources into a central data store with a predefined extract-transform-load (ETL) method(schema-on-read) [176]. As data volume increases, distributed DW solution (Hive) and associated ETL tool (Apache Pig) is available to manufacturing for business applications such as BI reporting, DSS. However, the predefined ETL method is not flexible and expensive to build before this finally used by data consumers. Data Lake is a new solution of data integration, which is defined as central storage to store any data (sizes, types, rates) with the raw formats in an enterprise [190]. DL is more suitable to data consumers for an ad-hoc query of the data, which is undefined until issuing the query(schema-on-read)[179]. HDFS is a popular DL tool to store extensive unstructured data in manufacturing (video, audio, image)[105]. However, DL probably becomes a "data swamp" without practical data management tools (Falcon, Sentry, Atlas etc.)[176].
- Semantic KM includes a series of operations for including, creating, classifying, sharing, using information and knowledge in manufacturing [180]. Many technologies of the semantic web are utilized to manage knowledge in manufacturing such as application of ontology web language (OWL) in SCADA [134], Resource Description Framework (RDF), RDF query language (SPARQL) in SCM[91], RDF database(Jena) in KM[23].

Data management tools include broad areas to address the veracity and value of big data issues in manufacturing. The tools are still developing since most of the tools are Hadoop-based. As

manufacturing begins to focus on velocity of big data computation, some new data management tools based on faster computation engines would be developed in the near future such as Spark and Flink. The big data management tools are still manually operated by data stewards, which would be challenging to manufacturers when various massive data is ingested into the big data platform, and different roles of users apply to use them. New methods of big data management could address this issue by using algorithms such as rule-based, machine learning–based or hybrid of both.

4.2.8 Infrastructure and deployment model

The infrastructure and deployment model are the foundation of big data applications in manufacturing. There are two types of cloud computing infrastructures:

1. General purpose commodity computer cluster (Hadoop, Spark) to process completely parallel computing problems such as computing massive sensor data separately or responding millions of users' requests (Facebook);
2. High-Performance Computing (HPC) cluster provides faster computing speed with dedicated hardware.

Hence, HPC outperforms others in processing highly dependent data computing such as complex modelling and simulation workload in manufacturing. However, the disadvantages of HPC are that its investment is enormous, and its utilization rate is low. Some solutions address its low utilization by moving HPC to cloud [182]. Some IAAS providers offer public cloud HPC (AWS) and hybrid cloud HPC service (Microsoft Azure).

Manufacturing enterprises have different perspectives on economic, security, the privacy of big data platforms. Four deployment models (Public cloud, Private cloud, Hybrid cloud, Community cloud [19]) are available to satisfy the requirements [183]. For example: in terms of better privacy and reducing waste of computing resource, some platforms could adopt a hybrid cloud model, which puts sensitive data on private cloud and process insensitive data on public cloud. However, Hybrid and Community cloud are not found by this review. To meet diverse requirements of manufacturing enterprises, some deployment technologies, including OpenStack and Docker [191], are used to quickly deploy an agile software environment with different micro-service frameworks and programming languages.

All deployment models and infrastructure should be taken into consideration in the big data ecosystem of manufacturing.

4.2.9 Cybersecurity

Cybersecurity is an essential component of big data platform to protect data assets in manufacturing. As manufacturing becomes data-driven, the standards and tools are required to secure data in the IT architecture of manufacturers completely. Although some security standards have been provided in manufacturing systems such as SCADA [192], the standards cannot fully address the challenges of SCADA on the Internet [193]. Traditional control systems are vulnerable to unauthorized attacks from the Internet since they are designed as close systems with few capabilities of cybersecurity [194]. Because big data platform integrates the physical space and the cyber space closely, the risk of cybersecurity could quickly escalate to the physical system in manufacturing. If unauthorized people manage the critical equipment or information, it will bring irreversible disaster to manufacturing such as economic loss, personal safety.

4.3 Question 3: How can we harness the capabilities available in the big data ecosystem to drive research innovations in manufacturing?

Big data ecosystem is the comprehension of massive functional components with various enabling tools. Capabilities of the big data ecosystem are not only about computing and storing big data, but also the advantages of its systematic platform and potentials of big data analytics. Hence, according to proposed solutions of reviewed literature and big data capabilities, the maturity of big data ecosystem application is categorized into three stages:

Stage 1: proposing a big data framework and platform;

Stage 2: harvesting cloud computing capacity for big data computing and storage;

Stage 3: analysing big data with various algorithms for the applications (prediction, fault detection, optimisation etc.).

Table 3 presents the allocation of proposed solutions by these three stages. MES, SCADA and O&M have been studied within the three stages. However, big-data based solution is missing in some manufacturing systems. For example, no general big data framework has been identified in CAx (Stage 1); no practices of Stage 2 and Stage 3 are found in PLM, SCM, BI and AM. Available capabilities of big data ecosystems could drive research innovation in various big data applications (Figure 5) of the immature manufacturing areas.

Manufacturing could benefit from new development and deployment methods of the big data ecosystem. Because many tools are fast changing with highly frequent updates, it makes the design of the big data platform dramatically challenging. Most of the proposed solutions in manufacturing constructed the platform from scratch by installing and testing every tool step by step. Researchers spent much time preparing the software but not focusing on programming and big data analytics. Several popular vendors offer their Hadoop distributions to mitigate the issue, such as Cloudera (CDH), Hortonworks (HDP, HDF), MapR, IBM (Infosphere BigInsights), Microsoft (HD Insight), Pivotal HD [195]. Because the tools and versions were tested in the distributions, they are ready to be used by researchers and enterprises in manufacturing. Moreover, the packages could be easily deployed and shared with virtualization technology (VM and container) [191]. However, developers need to evaluate several conditions to select the suitable distribution such as open source, pricing, customer support, sizes of community and so on [195].

4.4 Question 4: What are the future directions of big data applications in Manufacturing?

Manufacturing could benefit from fast developing big data technology with new and matured tools. Hence, the full potential of big data has not been discovered in manufacturing. Some potential directions are proposed for the future work of interested researchers:

4.4.1 Modelling and simulation

Modelling and simulation will naturally play an important role in extracting value from data once the volume of data is available [4]. Digital twin and digital thread are the two essential methods with two important tasks: collecting real-time data from manufacturing input and output devices such as CNC machine and sensors and building up the real-time simulation model with the collected data. These two tasks require considerable computing resources and data storage to process streaming data from manufacturing devices. Researchers commonly use Matlab or some scientific software to develop the selected algorithm to simulate the models [196]. As one simulation model may be implemented on a single computer, many models are required to design, and manually implemented on computers. It is a challenging and onerous task to manufacturers to convert the Matlab simulation models to executive programs on the selected cloud computing engine. However, from the reviewed articles, simulation is rare in the big data based solutions. The ongoing practice is that of a digital twin, which provide a real-time, bi-directional management between a physical object and its digital object [36]. It is essential to simulate the digital twin model with parameters across different domains (e.g. product, process and logistics) in order to predict its dynamic performance, such as Virtual/Augmented Reality[4]. For example, simulation of the product model with predefined parameters includes FEA analysis (ANSYS). However, the parameters from other domains (processes and logistics) are not accounted for in the FEA simulation. One potential research direction is customized products by integrating End-Of-Life data from PLM into CAx software. It could improve the product's impact during its EOL phase, such as Prognostics and health management (PHM) [184]). Another direction is that simulation of digital twin models with real-time data could be used for predictive maintenance. Moreover, simulation can provide integral data to data scientists or business users to design machine

learning algorithms. Because Matlab simulation manager is available to simulate multiple models on the public cloud (Azure, AWS), more case studies of methods above could be implemented to verify the performance of big data tools for digital twin and digital thread. Since Matlab is commercial software, unlike most of the software in the big data ecosystem that is open source and free to use, users have to purchase the specific licenses to run the models on the cloud. AnyLogic is also a very popular simulation software which provides cloud-based simulation recently. Researchers could use AnyLogic Cloud to deploy and verify their digital twin models.

Simulations with these large volumes of data require enormous computing, storage and communication resources, and cloud computing is well placed to satiate these. Another disruptive technology is Quantum computing, which can provide unlimited resources to process and store the data [197]. However, the technology is still under development and still has limitations, such as fault tolerance and error correlation[185].

Future research directions could be:

- Develop big data tools to convert simulation models from scientific software to implement on public cloud or private cloud;
- Using a simulation method to generate testing and training data for machine learning at planning and decision-making processes;
- Using general-purpose cloud computing cluster to simulate FEA at product design.

4.4.2 Connectivity and interoperability

Big data ecosystem complements of existing manufacturing approaches to have connectivity and interoperability. With the aim of systems integration and collaboration for smart manufacturing, the basis is that data and information have to be timely collected, correctly formatted, analysed and exchanged among the systems. MTConnect and OPC-UA are both data collection approaches in manufacturing. MTConnect focuses on device level and control level by monitoring CNC machine tools using a predefined consistent data model, data format and definition, which matches different vendors' machine tools. OPC-UA focuses on SCADA, MES and ERP by using a generic data model which can flexibly match more industrial devices with additional configuration effort. There are three challenging issues of both approaches. One issue is the performance limitation of OPC-UA which the CPU of OPC-UA server is identified as the main bottleneck in production [198]. Since both approaches are implemented on edge devices at the factory floor, the capabilities of data processing cannot be flexibly scaled up. The servers of both approaches have to be replaced with better performing hardware, while more data sources are connecting to the systems such as IoT devices. Another issue is interoperability of the systems using both approaches since there is no common ontology on the top of both information models. The last issue is the capabilities of data analysis and information exchange between both systems and other systems such as SCM, PLM, and CRM. Although some solutions are proposed to exchange both data formats and other formats (such as MTConnect and IEEE 1451 [118], OPC-UA and IoT [117], OPC-UA and AutomationML [98]), more solutions are required for data exchange of massive data formats between these two systems and others systems as shown in Table 1.

Four existing big data ingestion tools (Sqoop, Flume, Kafka and Storm) and a new tool (NiFi) could complement the weaknesses of both approaches. All five big data tools can scale out the capability of data processing by adding new hardware to the clusters without replacing the old ones. There are some differences between the five tools. The first difference is that Sqoop and Flume are two tools of Hadoop ecosystem; while Kafka and Storm are not dedicated to Hadoop. Secondly, data consumers require to pull data with Kafka, whereas Flume pushes data to consumers. Thirdly, Kafka provides better fault-tolerance and scalable than others. Kafka provides event duplication, which means other nodes continuously make data available when one node is failed. Compared with Kafka, Storm, Flume and NiFi do not provide event duplication, which is theoretically not suitable for application of critical missions such as safety, security, and finance in manufacturing. Storm, Flume and NiFi require less developing effort to work with Hadoop.

Interoperability has two main issues: data format and data quality. With the aim of interoperability among systems on BDA platform, data needs to be correctly formatted with good quality before exchange happens. Table 1 demonstrated various data formats in manufacturing systems. In terms of data formats, there are two types of data: data with schema (XML with given XSD/DTD) and data without schema such as XML without given XSD/DTD, JSON and unstructured data (document, report, email). Since data with schema define the schemas in standard schema files (XSD/DTD: PLM XML), data is can be exchanged automatically by mapping elements of each other's schemas. It is challenging to automatically transform data without schema to an intended schema since there is no mapping between both sides. One solution may be using natural language processing (NLP) to process human readable documents. BDA platform also has to support new file formats such as AMF and 3MF[199].

Issues of data quality in Table 2 includes missing value, noise data or anomaly data, uncertainty, data outlier, data correlation, timing and synchronization. Data transformation can address these issues to extract data to correct timely information. No generic transformation tool is identified in the reviewed papers which can be used as "One size fit all" tool for all the applications in manufacturing. Because manufacturers use various manufacturing systems with different data characterises, there are many combinations of data issues which require massive data transformation tools to address. It is challenging for manufacturers to develop customized big data transformation tool for every specific scenario. Without addressing these data issues, correct information cannot be extracted and exchanged among systems. Systems integration and collaboration cannot be achieved for smart manufacturing.

Future research directions in this area could be:

- Review the availability and feasibility of data collection tools in various manufacturing scenarios;
- Develop a generic data transformation solution with big data technologies to exchange data of manufacturing systems;

4.4.3 Standardized big data platform design

Standardisation of big data platform design improves the feasibility of enterprise-ready solutions in manufacturing. Because some essential components are missing in the proposed solutions, it is likely to increase the difficulty to apply them to manufacturing. Missing components would be insufficient to apply big data applications to smart manufacturing. However, there is no standard approach to design big data platform in manufacturing. The reason may be that different profiles of manufacturing enterprises have varieties of system requirements for big data applications.

A systematic assessment method is required to analyse the limitations and strengths of the proposed big data solution in the various manufacturing systems. Data issues are not entirely assessed, such as latency of data transmission among clusters, data quality and data format exchange. Therefore, it is essential to provide a standard approach to design big data platform with related assessment method to manufacturing.

4.4.4 Real time big data analytics

Big data analytics go deeper from batch analysing to real-time streaming analysing in manufacturing. On the one hand, streaming big data analysis is considered as a high research requirement in manufacturing [187]. One the other hand, enabling technologies are changing from non-real time analytics to real time. Figure 4 illustrates that the streaming computation engine Spark is becoming a popular tool than the traditional batching engine (MapReduce of Hadoop). However, it also shows that Hadoop still receives more focus than Spark in manufacturing from Figure 4. Batch processing is not able to provide real-time analytics response such as real time monitoring, dynamic scheduling and planning on systems of workshop floor (SCADA, MES). Micro-batching and streaming engines (Spark, Storm, Flink) can provide real-time big data analytics of streaming data [26]. Spark could take advantage of batching and streaming to replace MapReduce engine.

Moreover, compared to Storm and Flink, Spark has more powerful analytics tools such as SparkSQL, Spark R, GraphX, and MLlib. However, Storm and Flink outperform Spark in real-time concerns. Hence, with the objective of streaming and analytics, it is necessary to use Storm or Flink with Spark together. The issue is that it requires more development effort to work on several computation engines. Apache Beam provides a uniform abstraction layer to run these real-time engines at the execution layer[44]. Although it has not been used in manufacturing, researchers in manufacturing could focus on analytics logics without spending time on learning various usages of engines. Another fact of big data analytics forwarding to real-time is that data warehousing tools (Pig, Hive) of BI used MapReduce to batch processing large datasets, now that their latest versions support Spark engine.

4.4.5 Cybersecurity in manufacturing

Cybersecurity will continuously challenge manufacturing since security standards are still not available in some system such as SCM [49]. Recently, NIST published a framework to improve cybersecurity on critical infrastructures [188]. It is envisioned that there would be more developments in cybersecurity tools based on the new standard. No big data tool of cybersecurity was found in all reviewed papers, which is likely a promising research direction. Manufacturing could benefit from the following new security tools. Firstly, Apache Metron is an enterprise-ready real-time big data security tool, which is used by Telstra Company [200]. Secondly, Apache Ranger provides security administration management on Hadoop clusters. Thirdly, Apache Knox provides gateway service to access Hadoop clusters. Future direction in this area is to explore the capability of big data cybersecurity tools on critical systems in manufacturing such as safety, security and SCADA. Vulnerability in both the physical and cyberspace of the manufacturing systems must be identified and protected. The potential risk and damages warrant high priority of future research in this direction.

5 Conclusions and future work

This paper systematically reviews the state of art of big data research in manufacturing to evaluate the capabilities of big data ecosystem and requirements of smart manufacturing. Six key drivers of big data ecosystem are identified for smart manufacturing, which are system integration, data, prediction, sustainability, resource sharing and hardware. Afterwards, the nine essential components of big data ecosystem are presented to design a feasible big data solution to manufacturing enterprises. These are data ingestion, storage, computing, analytics, visualization, management, workflow, infrastructure and security. The evaluation reveals that there is no enterprise-ready big data solution in the reviewed literature.

It is important to note that some research areas have received less attention from the manufacturing community such as PLM, CAx, ERP and SCM. Many big data utilities are applicable to these areas, which could drive research innovation.

Regarding future work, there are five promising directions: modelling and simulation, connectivity and interoperability, standardized big data platform design, real-time big data analytics and cybersecurity.

6 References

[1] S. de Treville, M. Ketokivi, V. Singhal, Competitive manufacturing in a high-cost environment: Introduction to the special issue, J. Oper. Manag. 49–51 (2017) 1–5. doi:10.1016/j.jom.2017.02.001.

[2] H. (Deutsche P.A. Henning, Kagermann(National Academy of Science and Engineering). Wolfgang, Wahlster (German Research Center for Artificial Intelligence). Johannes, Recommendations for implementing the strategic initiative INDUSTRIE 4.0, Final Rep. Ind. 4.0 WG. (2013) 82.

[3] Y. Liao, F. Deschamps, E. de F.R. Loures, L.F.P. Ramos, Past, present and future of Industry 4.0 - a systematic literature review and research agenda proposal, Int. J. Prod. Res. 55 (2017) 3609–3629. doi:10.1080/00207543.2017.1308576.

[4] A. Kusiak, Smart manufacturing, Int. J. Prod. Res. 56 (2018) 508–517. doi:10.1080/00207543.2017.1351644.

[5] B. (Serm) Kulvatunyou, N. Ivezic, V. Srinivasan, On Architecting and Composing Engineering Information Services to Enable Smart Manufacturing, J. Comput. Inf. Sci. Eng. 16 (2016) 031002. doi:10.1115/1.4033725.

[6] B.W. Jeon, J. Um, S.C. Yoon, S. Suk-Hwan, An architecture design for smart manufacturing execution system, Comput. Aided. Des. Appl. 4360 (2016) 1–14. doi:10.1080/16864360.2016.1257189.

[7] X. Xu, From cloud computing to cloud manufacturing, Robot. Comput. Integr. Manuf. 28 (2012) 75–86. doi:10.1016/j.rcim.2011.07.002.

[8] J. Li, F. Tao, Y. Cheng, L. Zhao, Big Data in product lifecycle management, Int. J. Adv. Manuf. Technol. 81 (2015) 667–684. doi:10.1007/s00170-015-7151-x.

[9] J. Wang, Q. Chang, G. Xiao, N. Wang, S. Li, Data driven production modeling and simulation of complex automobile general assembly plant, Comput. Ind. 62 (2011) 765–775. doi:10.1016/j.compind.2011.05.004.

[10] S. Yang, B. Bagheri, H.-A. Kao, J. Lee, A Unified Framework and Platform for Designing of Cloud-Based Machine Health Monitoring and Manufacturing Systems, J. Manuf. Sci. Eng. 137 (2015) 040914. doi:10.1115/1.4030669.

[11] H. Sequeira, P. Carreira, T. Goldschmidt, P. Vorst, Energy cloud: Real-time cloud-native energy management system to monitor and analyze energy consumption in multiple industrial sites, Proc. - 2014 IEEE/ACM 7th Int. Conf. Util. Cloud Comput. UCC 2014. (2014) 529–534. doi:10.1109/UCC.2014.79.

[12] R.Y. Zhong, S.T. Newman, G.Q. Huang, S. Lan, Big Data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives, Comput. Ind. Eng. 101 (2016) 572–591. doi:10.1016/j.cie.2016.07.013.

[13] H.M. Chen, R. Schutz, R. Kazman, F. Matthes, Amazon in the air: Innovating with big data at Lufthansa, Proc. Annu. Hawaii Int. Conf. Syst. Sci. 2016-March (2016) 5096–5105. doi:10.1109/HICSS.2016.631.

[14] M. Helu, T. Hedberg, A. Barnard Feeney, Reference architecture to integrate heterogeneous manufacturing systems for the digital thread, CIRP J. Manuf. Sci. Technol. 19 (2017) 191–195. doi:10.1016/j.cirpj.2017.04.002.

[15] F. Tao, F. Sui, A. Liu, Q. Qi, M. Zhang, B. Song, Z. Guo, S.C.Y. Lu, A.Y.C. Nee, Digital twin-driven product design framework, Int. J. Prod. Res. 7543 (2018) 1–19. doi:10.1080/00207543.2018.1443229.

[16] L. Monostori, B. Kádár, T. Bauernhansl, S. Kondoh, S. Kumara, G. Reinhart, O. Sauer, G. Schuh, W. Sihn, K. Ueda, Cyber-physical systems in manufacturing, CIRP Ann. - Manuf. Technol. 65 (2016) 621–641. doi:10.1016/j.cirp.2016.06.005.

[17] F. Bonomi, R. Milito, P. Natarajan, J. Zhu, Big Data and Internet of Things: A Roadmap for Smart Environments, 2014. doi:10.1007/978-3-319-05029-4.

[18] J. Gubbi, R. Buyya, S. Marusic, M. Palaniswami, Internet of Things (IoT): A vision, architectural elements, and future directions, Futur. Gener. Comput. Syst. 29 (2013) 1645–1660. doi:10.1016/j.future.2013.01.010.

[19] P. Mell, T. Grance, The NIST definition of cloud computing, NIST Spec. Publ. 145 (2011) 7. doi:10.1136/emj.2010.096966.

[20] B.V. Dhar, Data Science and Prediction, Commun. ACM. 56 (2013) 64–73.

[21] A.J.C. Trappey, C. V. Trappey, U. Hareesh Govindarajan, A.C. Chuang, J.J. Sun, A review of essential standards and patent landscapes for the Internet of Things: A key enabler for Industry 4.0, Adv. Eng. Informatics. 33 (2017) 208–229. doi:10.1016/j.aei.2016.11.007.

[22] R. Kosara, C. Healey, Visualization viewpoints: Data, Information and Knowledge in Visualization, Comput. Graph. …. (2003). http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1210860.

[23] S. Wang, J. Wan, D. Li, C. Liu, Knowledge reasoning with semantic data for real-time data processing in smart factory, Sensors (Switzerland). 18 (2018) 1–10. doi:10.3390/s18020471.

[24] L. Spendla, M. Kebisek, P. Tanuska, L. Hrcka, Concept of predictive maintenance of production systems in accordance with industry 4.0, SAMI 2017 - IEEE 15th Int. Symp. Appl. Mach. Intell. Informatics, Proc. (2017) 405–410. doi:10.1109/SAMI.2017.7880343.

[25] O. Morariu, C. Morariu, T. Borangiu, S. Răileanu, Manufacturing Systems at Scale with Big Data Streaming and Online Machine Learning, Stud. Comput. Intell. 762 (2018) 253–264. doi:10.1007/978-3-319-73751-5_19.

[26] K. Nagorny, P. Lima-Monteiro, J. Barata, A.W. Colombo, Big Data Analysis in Smart Manufacturing: A Review, Int. J. Commun. Netw. Syst. Sci. 10 (2017) 31–58. doi:10.4236/ijcns.2017.103003.

[27] D. Reinsel, J. Gantz, J. Rydning, Data Age 2025: The Digitization of the World From Edge to Core, 2018. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf.

[28] A. Gandomi, M. Haider, Beyond the hype: Big data concepts, methods, and analytics, Int. J. Inf. Manage. 35 (2015) 137–144. doi:10.1016/j.ijinfomgt.2014.10.007.

[29] J. Dean, S. Ghemawat, MapReduce : Simplified Data Processing on Large Clusters, (n.d.) 1–13.

[30] K. Shvachko, H. Kuang, S. Radia, R. Chansler, The Hadoop Distributed File System, in: 2010 IEEE 26th Symp. Mass Storage Syst. Technol., 2010: pp. 1–10. doi:10.1109/MSST.2010.5496972.

[31] J. Duda, Business Intelligence and NoSQL Databases, Inf. Syst. Manag. 1 (2012) 25–37.

[32] R. Cattell, Scalable SQL and NoSQL Data Stores, Acm Sigmod Rec. 39 (2010) 12–27.

[33] Y. Bao, L. Ren, L. Zhang, X. Zhang, Y. Luo, Massive sensor data management framework in cloud manufacturing based on Hadoop, IEEE Int. Conf. Ind. Informatics. (2012) 397–401. doi:10.1109/INDIN.2012.6301192.

[34] F. Tao, Y. Cheng, L. Da Xu, L. Zhang, B.H. Li, CCIoT-CMfg: Cloud computing and internet of things-based cloud manufacturing service system, IEEE Trans. Ind. Informatics. 10 (2014) 1435–1442. doi:10.1109/TII.2014.2306383.

[35] D. Wu, D.W. Rosen, L. Wang, D. Schaefer, Cloud-based design and manufacturing: A new paradigm in digital manufacturing and design innovation, CAD Comput. Aided Des. 59 (2015) 1–14. doi:10.1016/j.cad.2014.07.006.

[36] F. Tao, Q. Qi, A. Liu, A. Kusiak, Data-driven smart manufacturing, J. Manuf. Syst. (2018). doi:10.1016/j.jmsy.2018.01.006.

[37] M. Soualhia, F. Khomh, S. Tahar, Task Scheduling in Big Data Platforms: A Systematic Literature Review, J. Syst. Softw. 134 (2017) 170–189. doi:10.1016/j.jss.2017.09.001.

[38] Y. Demchenko, P. Grosso, C. De Laat, P. Membrey, Addressing big data issues in Scientific Data Infrastructure, Proc. 2013 Int. Conf. Collab. Technol. Syst. CTS 2013. (2013) 48–55. doi:10.1109/CTS.2013.6567203.

[39] J. Moyne, J. Samantaray, M. Armacost, Big Data Capabilities Applied to Semiconductor Manufacturing Advanced Process Control, IEEE Trans. Semicond. Manuf. 29 (2016) 283–291. doi:10.1109/TSM.2016.2574130.

[40] Y. Demchenko, C. De Laat, P. Membrey, Defining architecture components of the Big Data Ecosystem, 2014 Int. Conf. Collab. Technol. Syst. CTS 2014. (2014) 104–112. doi:10.1109/CTS.2014.6867550.

[41] S. Landset, T.M. Khoshgoftaar, A.N. Richter, T. Hasanin, A survey of open source tools for machine learning with big data in the Hadoop ecosystem, J. Big Data. 2 (2015) 24. doi:10.1186/s40537-015-0032-1.

[42] S. Binani, A. Gutti, S. Upadhyay, SQL vs. NoSQL vs. NewSQL-A Comparative Study, Commun. Appl. Electron. 6 (2016) 43–46. http://www.caeaccess.org/archives/volume6/number1/binani-2016-cae-652418.pdf.

[43] K. Grolinger, W.A. Higashino, A. Tiwari, M.A.M. Capretz, Data management in cloud environments: NoSQL and NewSQL data stores, J. Cloud Comput. 2 (2013). doi:10.1186/2192-113X-2-22.

[44] M. Gökalp, K. Kayabay, M. Zaki, A. Koçyiğit, Big-Data Analytics Architecture for Businesses: a comprehensive review on new open-source big-data tools, (2017) 1–35. doi:10.13140/RG.2.2.30306.84165.

[45] J. Wan, S. Tang, Z. Shu, D. Li, S. Wang, M. Imran, A. V. Vasilakos, Software-Defined Industrial Internet of Things in the Context of Industry 4.0, IEEE Sens. J. 16 (2016) 7373–7380. doi:10.1109/JSEN.2016.2565621.

[46] M.J. Koop, W. Huang, K. Gopalakrishnan, D.K. Panda, Performance analysis and evaluation of PCIe 2.0 and quad-data rate InfiniBand, Proc. - Symp. High Perform. Interconnects, Hot Interconnects. (2008) 85–92. doi:10.1109/HOTI.2008.26.

[47] J. Cheng, L. Da Xu, W. Chen, F. Tao, C.-L. Lin, Industrial IoT in 5G Environment towards Smart Manufacturing, J. Ind. Inf. Integr. (2018). doi:10.1016/j.jii.2018.04.001.

[48] P. Zikopoulos, C. Eaton, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data: Analytics for Enterprise Class Hadoop and Streaming Data, 2011. https://www.immagic.com/eLibrary/ARCHIVES/EBOOKS/I111025E.pdf.

[49] Y. Lu, K. Morris, S. Frechette, Current Standards Landscape for Smart Manufacturing Systems, Natl. Inst. Stand. Technol. NISTIR,. 8107 (2016) 39. doi:10.6028/NIST.IR.8107.

[50] S. Evdokimov, RFID and the Internet of Things: Technology, Applications, and Security Challenges, 2010. doi:10.1561/0200000020.

[51] D. Kiritsis, V.K. Nguyen, J. Stark, How closed-loop PLM improves Knowledge Management over the complete product lifecycle and enables the factory of the future, Int. J. Prod. Lifecycle Manag. 3 (2008) 54. doi:10.1504/IJPLM.2008.019970.

[52] P. Buneman, Semistructured data, Proc. 16th Symp. Princ. Database Syst. (1997) 117–121.

doi:10.1145/263661.263675.

[53] Y. Xu, G. Chen, J. Zheng, An integrated solution—KAGFM for mass customization in customer-oriented product design under cloud manufacturing environment, Int. J. Adv. Manuf. Technol. 84 (2016) 85–101. doi:10.1007/s00170-015-8074-2.

[54] S. Chen, C. Yin, X. Li, Implementation of MTConnect in Machine Monitoring System for CNCs, Proc. - 2017 5th Int. Conf. Enterp. Syst. Ind. Digit. by Enterp. Syst. ES 2017. (2017) 70–75. doi:10.1109/ES.2017.19.

[55] S.L. Chan, Y. Lu, Y. Wang, Data-driven cost estimation for additive manufacturing in cybermanufacturing, J. Manuf. Syst. 46 (2018) 115–126. doi:10.1016/j.jmsy.2017.12.001.

[56] S.K. and K.C. and H.Y. and S.O. Yang, G-code conversion from 3D model data for 3D printers on Hadoop systems, in: 2017 4th Int. Conf. Comput. Appl. Inf. Process. Technol., 2017: pp. 1–4. doi:10.1109/CAIPT.2017.8320709.

[57] S. Sierla, V. Kyrki, P. Aarnio, V. Vyatkin, Automatic assembly planning based on digital product descriptions, Comput. Ind. 97 (2018) 34–46. doi:10.1016/j.compind.2018.01.013.

[58] Y. Zhang, S. Ren, Y. Liu, T. Sakao, D. Huisingh, A framework for Big Data driven product lifecycle management, J. Clean. Prod. 159 (2017) 229–240. doi:10.1016/j.jclepro.2017.04.172.

[59] Y. Zhang, S. Ren, Y. Liu, S. Si, A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products, J. Clean. Prod. (2016). doi:10.1016/j.jclepro.2016.07.123.

[60] D. Ramanujan, W.Z. Bernstein, M.A. Totorikaguena, C.F. Ilvig, K.B. Ørskov, Generating Contextual Design for Environment Principles in Sustainable Manufacturing Using Visual Analytics, J. Manuf. Sci. Eng. 141 (2018) 021016. doi:10.1115/1.4041835.

[61] T. Hedberg, A.B. Feeney, M. Helu, J.A. Camelio, Toward a Lifecycle Information Framework and Technology in Manufacturing, J. Comput. Inf. Sci. Eng. 17 (2017) 021010. doi:10.1115/1.4034132.

[62] J.W. Wang, J.Y. Yang, J. Zhang, X.X. Wang, W. (Chris) Zhang, Big data driven cycle time parallel prediction for production planning in wafer manufacturing, Enterp. Inf. Syst. 12 (2018) 1–19. doi:10.1080/17517575.2018.1450998.

[63] X. Li, J. Song, B. Huang, A scientific workflow management system architecture and its scheduling based on cloud service platform for manufacturing big data analytics, Int. J. Adv. Manuf. Technol. 84 (2015) 119–131. doi:10.1007/s00170-015-7804-9.

[64] M.Y. Santos, J. Oliveira e Sá, C. Andrade, F. Vale Lima, E. Costa, C. Costa, B. Martinho, J. Galvão, A Big Data system supporting Bosch Braga Industry 4.0 strategy, Int. J. Inf. Manage. 37 (2017) 750–760. doi:10.1016/j.ijinfomgt.2017.07.012.

[65] S. Wang, C. Zhang, C. Liu, D. Li, H. Tang, Cloud-assisted interaction and negotiation of industrial robots for the smart factory, Comput. Electr. Eng. 63 (2017) 66–78. doi:10.1016/j.compeleceng.2017.05.025.

[66] N.M. Khushairi, N.A. Emran, M.M. Mohd Yusof, Database performance tuning methods for manufacturing execution system, World Appl. Sci. J. 30 (2014) 91–99. doi:10.5829/idosi.wasj.2014.30.icmrp.14.

[67] B.W. Jeon, J. Um, S.C. Yoon, S. Suk-hwan, An architecture design for smart manufacturing execution system, 4360 (2017). doi:10.1080/16864360.2016.1257189.

[68] M. Jirkovský, V. and Obitko, Enabling semantics within industry 4.0, in: Int. Conf. Ind. Appl. Holonic Multi-Agent Syst., 2017: pp. 39–52. doi:10.1007/978-3-642-40090-2.

[69] O. Sauer, Developments and trends in shopfloor-related ICT systems, IEEE Int. Conf. Ind. Eng. Eng. Manag. 2015-Janua (2014) 1352–1356. doi:10.1109/IEEM.2014.7058859.

[70] B.E.I. Systron, D. Division, Managing Configuration Control in an Automotive Sensor Mass Customization Manufacturing Product Line, 2006 World Autom. Congr. (2006) 1--16.

[71] J.-J. Kim, D.-W. Lee, D.-B. Ko, S.-I. Jeong, J.-M. Park, An autonomic computing based on big data platform for high-reliable smart factory, J. Eng. Appl. Sci. 12 (2017) 2662–2666. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85041023417&partnerID=40&md5=003f099909f142b4879dffa79001fe1c.

[72] P. Reboredo, M. Keinert, Integration of discrete manufacturing field devices data and services based on OPC UA, IECON Proc. (Industrial Electron. Conf. (2013) 4476–4481. doi:10.1109/IECON.2013.6699856.

[73] A.R. Khan, H. Schioler, M. Kulahci, T. Knudsen, Big data analytics for industrial process control, 2017 22nd IEEE Int. Conf. Emerg. Technol. Fact. Autom. (2017) 1–8. doi:10.1109/ETFA.2017.8247658.

[74] X. Ye, S.H. Hong, An AutomationML/OPC UA-based Industry 4.0 Solution for a Manufacturing System, IEEE Int. Conf. Emerg. Technol. Fact. Autom. ETFA. 2018-Septe (2018) 543–550. doi:10.1109/ETFA.2018.8502637.

[75] D. Mourtzis, E. Vlachou, A cloud-based cyber-physical system for adaptive shop-floor scheduling and condition-based maintenance, J. Manuf. Syst. 47 (2018) 179–198. doi:10.1016/j.jmsy.2018.05.008.

[76] C. Toro, I. Barandiaran, J. Posada, A perspective on knowledge based and intelligent systems implementation in industrie 4.0, Procedia Comput. Sci. 60 (2015) 362–370. doi:10.1016/j.procs.2015.08.143.

[77] E. Poormohammady, J.H. Reelfs, M. Stoffers, K. Wehrle, A. Papageorgiou, Dynamic algorithm selection for the logic of tasks in IoT stream processing systems, 2017 13th Int. Conf. Netw. Serv. Manag. CNSM 2017. 2018-Janua (2018) 1–5. doi:10.23919/CNSM.2017.8256009.

[78] D. Mourtzis, E. Vlachou, N. Milas, Industrial Big Data as a result of IoT adoption in Manufacturing, Procedia CIRP. 55 (2016) 290–295. doi:10.1016/j.procir.2016.07.038.

[79] J. Moyne, J. Samantaray, M. Armacost, Big data emergence in semiconductor manufacturing advanced process control, Adv. Semicond. Manuf. Conf. (ASMC), 2015 26th Annu. SEMI. (2015) 130–135. doi:10.1109/ASMC.2015.7164483.

[80] D. Mourtzis, A. Vlachou, V. Zogopoulos, Cloud-Based Augmented Reality Remote Maintenance Through Shop-Floor Monitoring: A Product-Service System Approach, J. Manuf. Sci. Eng. 139 (2017) 061011. doi:10.1115/1.4035721.

[81] Y. Busnel, N. Riveei, A. Gal, Y. Busnel, N. Riveei, A. Gal, A. Gal, FlinkMan : Anomaly Detection in Manufacturing Equipment with Apache Flink : Grand Challenge, (2017). doi:10.1145/3093742.3095099.

[82] I. Yen, S. Zhang, F. Bastani, A Framework for IoT-Based Monitoring and Diagnosis of Manufacturing Systems, in: Proc. - 11th IEEE Int. Symp. Serv. Syst. Eng. SOSE 2017, 2017: pp. 1–8. doi:10.1109/SOSE.2017.26.

[83] S. Kang, W.T.K. Chien, J.G. Yang, A study for big-data (Hadoop) application in semiconductor manufacturing, IEEE Int. Conf. Ind. Eng. Eng. Manag. 2016-Decem (2016) 1893–1897. doi:10.1109/IEEM.2016.7798207.

[84] X. Li, Z. Tu, Q. Jia, X. Man, H. Wang, X. Zhang, Deep-level quality management based on big data analytics with case study, Proc. - 2017 Chinese Autom. Congr. CAC 2017. 2017-Janua (2017) 4921–4926. doi:10.1109/CAC.2017.8243651.

[85] M. Syafrudin, N.L. Fitriyani, D. Li, G. Alfian, J. Rhee, Y.S. Kang, An open source-based real-time data processing architecture framework for manufacturing sustainability, Sustain. 9 (2017). doi:10.3390/su9112139.

[86] S. Venkatesh, Web-Enabled Real-Time Quality Feedback for Factory Systems Using MTConnect, in: ASME 2012 Int. Des. Eng. Tech. Conf. Comput. Inf. Eng. Conf., 2012: pp. 403–409.

[87] H.-M. Hou, J.-F. Kung, Y.-B. Hsu, Y. Yamazaki, K. Maruyama, Y. Toyoshima, C. Chen, Prediction of ppm level electrical failure by using physical variation analysis, Proc.SPIE. (2016) 9778. doi:10.1117/12.2229410.

[88] M.-K. Zheng, X.-G. Ming, X.-Y. Zhang, G.-M. Li, MapReduce Based Parallel Bayesian Network for Manufacturing Quality Control, Chinese J. Mech. Eng. 30 (2017) 1216–1226. doi:10.1007/s10033-017-0179-0.

[89] S.A. Jacobs, A. Dagnino, Large-Scale Industrial Alarm Reduction and Critical Events Mining Using Graph Analytics on Spark, in: Proc. - 2016 IEEE 2nd Int. Conf. Big Data Comput. Serv. Appl. BigDataService 2016, 2016: pp. 66–71. doi:10.1109/BigDataService.2016.21.

[90] H. Kemper, H. Baars, H. Lasi, An integrated business intelligence framework: Closing the gap between IT support for management and for production, in: Bus. Intell. Perform. Manag., Springer London, 2013: pp. 13–27. doi:10.1007/978-1-4471-4866-1.

[91] B.R. Ferrer, W.M. Mohammed, J.L.M. Lastra, A solution for processing supply chain events within ontology-based descriptions, (2016) 4877–4883.

[92] M.J.A.G. Izaguirre, A. Lobov, J.L.M. Lastra, OPC-UA and DPWS interoperability for factory floor monitoring using complex event processing, IEEE Int. Conf. Ind. Informatics. (2011) 205–210. doi:10.1109/INDIN.2011.6034874.

[93] J. Campos, P. Sharma, U.G. Gabiria, E. Jantunen, D. Baglee, A Big Data Analytical Architecture for the Asset Management, Procedia CIRP. 64 (2017) 369–374. doi:10.1016/j.procir.2017.03.019.

[94] P. Gölzer, L. Simon, P. Cato, M. Amberg, Designing global manufacturing networks using Big Data, Procedia CIRP. 33 (2015) 191–196. doi:10.1016/j.procir.2015.06.035.

[95] M. Zimmermann, U. Breitenbucher, M. Falkenthal, F. Leymann, K. Saatkamp, Standards-Based Function Shipping - How to Use TOSCA for Shipping and Executing Data Analytics Software in Remote Manufacturing Environments, Proc. - 2017 IEEE 21st Int. Enterp. Distrib. Object Comput. Conf. EDOC 2017. 2017-Janua (2017) 50–60. doi:10.1109/EDOC.2017.16.

[96] L.K. B, C. Gr, K. Jan, E. Hoos, C. Kiefer, C. Weber, S. Silcher, B. Mitschang, The stuttgart IT architecture for manufacturing an architecture for the data-driven factory, Int. Conf. Enterp. Inf. Syst. (2016) 53–80. doi:10.1007/978-3-319-62386-3.

[97] W.M. Mohammed, B.R. Ferrer, L. Jose, M. Lastra, D. Aleixo, C. Agostinho, Configuring and visualizing the data resources in a cloud-based data collection framework, 2017 Int. Conf. Eng. Technol. Innov. Eng. Technol. Innov. Manag. Beyond 2020 New Challenges, New Approaches, ICE/ITMC 2017 - Proc. 2018-Janua (2018) 1201–1208. doi:10.1109/ICE.2017.8280017.

[98] X. Ye, T.Y. Park, S.H. Hong, Y. Ding, A. Xu, Implementation of a Production-Control System Using Integrated Automation ML and OPC UA, 2018 Work. Metrol. Ind. 4.0 IoT, MetroInd 4.0 IoT 2018 - Proc. (2018) 242–247. doi:10.1109/METROI4.2018.8428310.

[99] Y. Cao, S. Wang, L. Kang, C. Li, L. Guo, Study on machining service modes and resource selection strategies in cloud manufacturing, Int. J. Adv. Manuf. Technol. 81 (2015) 597–613. doi:10.1007/s00170-015-7222-z.

[100] S.K. Panda, T. Schroder, L. Wisniewski, C. Diedrich, PlugProduce Integration of Components into OPC UA based data-space, IEEE Int. Conf. Emerg. Technol. Fact. Autom. ETFA. 2018-Septe (2018) 1095–1100. doi:10.1109/ETFA.2018.8502663.

[101] F. Tao, L. Zhang, Y. Liu, Y. Cheng, L. Wang, X. Xu, Manufacturing Service Management in Cloud Manufacturing: Overview and Future Research Directions, J. Manuf. Sci. Eng. 137

(2015) 040912. doi:10.1115/1.4030510.

[102] H. Yang, M. Park, M. Cho, M. Song, S. Kim, A system architecture for manufacturing process analysis based on big data and process mining techniques, Big Data (Big Data), 2014 IEEE Int. Conf. (2014) 1024–1029. doi:10.1109/BigData.2014.7004336.

[103] J.C.C. Tseng, J.Y. Gu, P.F. Wang, C.Y. Chen, C.F. Li, V.S. Tseng, A scalable complex event analytical system with incremental episode mining over data streams, 2016 IEEE Congr. Evol. Comput. CEC 2016. (2016) 648–655. doi:10.1109/CEC.2016.7743854.

[104] B. Suryajaya, C.C. Chen, M.H. Hung, Y.Y. Liu, J.X. Liu, Y.C. Lin, A fast large-size production data transformation scheme for supporting smart manufacturing in semiconductor industry, IEEE Int. Conf. Autom. Sci. Eng. 2017-Augus (2018) 275–281. doi:10.1109/COASE.2017.8256114.

[105] A. Luckow, M. Cook, N. Ashcraft, E. Weill, E. Djerekarov, B. Vorster, Deep Learning in the Automotive Industry: Applications and Tools, Big Data Int. Conf. Big Data. (2016) 3759–3768. doi:10.1109/BigData.2016.7841045.

[106] A. Brodsky, G. Shao, M. Krishnamoorthy, A. Narayanan, D. Menasce, R. Ak, Analysis and optimization in smart manufacturing based on a reusable knowledge base for process performance models, Proc. - 2015 IEEE Int. Conf. Big Data, IEEE Big Data 2015. (2015) 1418–1427. doi:10.1109/BigData.2015.7363902.

[107] S.C. Feng, W.Z. Bernstein, T. Hedberg, A. Barnard Feeney, Toward Knowledge Management for Smart Manufacturing, J. Comput. Inf. Sci. Eng. 17 (2017) 031016. doi:10.1115/1.4037178.

[108] M. Stonebraker, I.F. Ilyas, Data Integration: The Current Status and the Way Forward., IEEE Data Eng. Bull. 41 (2018) 3–9.

[109] D. Zhang, B. Xu, J. Wood, Predict failures in production lines: A two-stage approach with clustering and supervised learning, Proc. - 2016 IEEE Int. Conf. Big Data, Big Data 2016. (2016) 2070–2074. doi:10.1109/BigData.2016.7840832.

[110] V. Jirkovský, M. Obitko, P. Novák, P. Kadera, Big data analysis for sensor time-series in automation, 19th IEEE Int. Conf. Emerg. Technol. Fact. Autom. ETFA 2014. (2014). doi:10.1109/ETFA.2014.7005183.

[111] J. Bakakeu, J. Fuchs, T. Javied, M. Brossog, J. Franke, H. Klos, W. Eberlein, S. Tolksdorf, J. Peschke, L. Jahn, Multi-Objective Design Space Exploration for the Integration of Advanced Analytics in Cyber-Physical Production Systems, IEEE Int. Conf. Ind. Eng. Eng. Manag. 2019-Decem (2019) 1866–1873. doi:10.1109/IEEM.2018.8607483.

[112] R. Lynn, W. Louhichi, M. Parto, E. Wescoat, T. Kurfess, RAPIDLY DEPLOYABLE MTCONNECT-BASED MACHINE TOOL MONITORING SYSTEMS, in: Proc. ASME 12TH Int. Manuf. Sci. Eng. Conf. - 2017, VOL 3, 2017.

[113] D. Libes, S. Shin, J. Woo, Considerations and recommendations for data availability for data analytics for manufacturing, in: 3rd IEEE Int. Conf. Big Data, IEEE Big Data, 2015: pp. 68–75. doi:10.1109/BigData.2015.7363743.

[114] C. Zhao, L. Zhang, X.Z.L. Zhang, Cloud manufacturing resource management based on metadata, ASME 2015 Int. Manuf. Sci. Eng. Conf. MSEC 2015. 2 (2015) 1–8. doi:10.1115/MSEC20159388.

[115] Y. Cheng, W. Shang, L. Zhu, D. Zhang, D. Feng, Items analysis of postal supervision, 2016 IEEE/ACIS 15th Int. Conf. Comput. Inf. Sci. ICIS 2016 - Proc. (2016) 3–5. doi:10.1109/ICIS.2016.7550949.

[116] P. Gaj, K. Andrzej, P. Stera, Ontology-Based Integrated Monitoring of Hadoop Clusters in Industrial Environments with OPC UA and RESTful Web Services, Commun. Comput. Inf. Sci. 522 (2015) 162–171. doi:10.1007/978-3-319-19419-6.

[117] H. Derhamy, J. Ronnholm, J. Delsing, J. Eliasson, J. Van Deventer, Protocol interoperability of OPC UA in service oriented architectures, Proc. - 2017 IEEE 15th Int. Conf. Ind. Informatics, INDIN 2017. (2017) 44–50. doi:10.1109/INDIN.2017.8104744.

[118] K.B. Lee, E.Y. Song, P.S. Gu, Integration of MTConnect and Standard-Based Sensor Networks for Manufacturing Equipment Monitoring, in: ASME 2012 Int. Manuf. Sci. Eng. Conf. MSEC, 2012: pp. 4–8.

[119] A.N. Richter, T.M. Khoshgoftaar, S. Landset, T. Hasanin, A Multi-dimensional Comparison of Toolkits for Machine Learning with Big Data, Proc. - 2015 IEEE 16th Int. Conf. Inf. Reuse Integr. IRI 2015. (2015) 1–8. doi:10.1109/IRI.2015.12.

[120] A. Jos, Integration of Sensors , Controllers and Instruments, Sensors. (2017). doi:10.3390/s17071512.

[121] A. Luckow, K. Kennedy, F. Manhardt, E. Djerekarov, B. Vorster, A. Apon, Automotive big data: Applications, workloads and infrastructures, Proc. - 2015 IEEE Int. Conf. Big Data, IEEE Big Data 2015. (2015) 1201–1210. doi:10.1109/BigData.2015.7363874.

[122] C. Mathis, Data Lakes, Datenbank-Spektrum. (2017). doi:10.1007/s13222-017-0272-7.

[123] M. Sarnovsky, P. Bednar, M. Smatana, Data integration in scalable data analytics platform for process industries, in: 2017 IEEE 21st Int. Conf. Intell. Eng. Syst., 2017: pp. 187–192. doi:10.1109/INES.2017.8118553.

[124] M. Sarnovsky, P. Bednar, M. Smatana, Big Data Processing and Analytics Platform Architecture for Process Industry Factories, Big Data Cogn. Comput. 2 (2018) 3. doi:10.3390/bdcc2010003.

[125] G.L. Ooi, Y.-H. Wang, P.S. Tan, Z. Zhang, Y. Gao, J.K. Chow, Y. Wu, Q. Yuan, Customizable and scalable geotechnical laboratory testing and field monitoring with new sensing and big data technologies, ICSMGE 2017 - 19th Int. Conf. Soil Mech. Geotech. Eng. (2017) 471–474.

[126] H. Liang, L. Feng, Z. Chun, Application of the Big Data Technology for Massive Data of the Whole Life Cycle of EMU, in: F. Xhafa, S. Patnaik, Z. Yu (Eds.), Recent Dev. Intell. Syst. Interact. Appl., Springer International Publishing, Cham, 2017: pp. 219–224.

[127] L. Zheng, L. Tang, T. Li, B. Duan, M. Lei, P. Wang, C. Zeng, L. Li, Y. Jiang, W. Xue, J. Li, C. Shen, W. Zhou, H. Li, Applying data mining techniques to address critical process optimization needs in advanced manufacturing, Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '14. (2014) 1739–1748. doi:10.1145/2623330.2623347.

[128] S. Windmann, A. Maier, O. Niggemann, C. Frey, A. Bernardi, Y. Gu, H. Pfrommer, T. Steckel, M. Krüger, R. Kraus, Big data analysis of manufacturing processes, J. Phys. Conf. Ser. 659 (2015). doi:10.1088/1742-6596/659/1/012055.

[129] R. Bohlin, L. Lindkvist, J. Hagmar, J.S. Carlson, K. Bengtsson, DATA FLOW AND COMMUNICATION FRAMEWORK SUPPORTING DIGITAL TWIN FOR GEOMETRY ASSURANCE Robert, in: Proc. ASME 2017 Int. Mech. Eng. Congr. Expo., 2017: pp. 1–7.

[130] K.E. Harper, J. Zheng, S.A. Jacobs, A. Dagnino, A. Jansen, T. Goldschmidt, A. Marinakis, Industrial analytics pipelines, Proc. - 2015 IEEE 1st Int. Conf. Big Data Comput. Serv. Appl. BigDataService 2015. (2015) 242–248. doi:10.1109/BigDataService.2015.38.

[131] T. Goldschmidt, M.K. Murugaiah, C. Sonntag, B. Schlich, S. Biallas, P. Weber, Cloud-Based Control : A Multi-Tenant , Horizontally Scalable Soft-PLC, in: 2015 IEEE 8th Int. Conf. Cloud Comput., 2015. doi:10.1109/CLOUD.2015.124.

[132] H.-S. Park, J.-H. Kim, C.-H. Choi, B.-R. Jung, K.-H. Lee, S.-Y. Chi, W.-S. Cho, In-Memory Data Grid System for Real-Time Processing of Machine Sensor Data in a Smart Factory Environment, Proc. 2015 Int. Conf. Big Data Appl. Serv. - BigDAS '15. (2015) 92–97. doi:10.1145/2837060.2837073.

[133] N. Ramakrishnanus, R. Ghosh, Distributed Dynamic Elastic Nets : A Scalable Approach for Regularization in Dynamic Manufacturing Environments, 2015 IEEE Int. Conf. Big Data. (2015) 2752–2761.

[134] S. Wang, J. Ouyang, D. Li, C. Liu, An Integrated Industrial Ethernet Solution for the Implementation of Smart Factory, IEEE Access. (2017) 25455–25462. doi:10.1109/ACCESS.2017.2770180.

[135] S. Division, C. Hsing, N. Village, N. City, N. County, M. Availability, Developing a cloud virtual maintenance system for machine tools management, 2015 11th Int. Conf. Heterog. Netw. Qual. Reliab. Secur. Robustness. (2015) 358–364.

[136] J. Wan, S. Tang, D. Li, S. Wang, C. Liu, H. Abbas, S. Member, A. V Vasilakos, A Manufacturing Big Data Solution for Active Preventive Maintenance, in: IEEE Trans. Ind. INFORMATICS, 2017: pp. 2039–2047.

[137] J.-H. Ku, A Study on Prediction Model of Equipment Failure Through Analysis of Big Data Based on RHadoop, Wirel. Pers. Commun. 98 (2017) 3163–3176. doi:10.1007/s11277-017-4151-1.

[138] W. Lee, J. Cho, L. Lee, S. Korea, Time Series Abnormal Data Detection for Smart Factory, Int. J. Control Autom. 11 (2018) 91–98.

[139] H. Chen, X. Fei, S. Wang, X. Lu, G. Jin, W. Li, X. Wu, Energy Consumption Data Based Machine Anomaly Detection, 2014 Second Int. Conf. Adv. Cloud Big Data. (2014) 136–142. doi:10.1109/CBD.2014.24.

[140] N. Stojanovic, M. Dinic, L. Stojanovic, A data-driven approach for multivariate contextualized anomaly detection: Industry use case, Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017. 2018-Janua (2018) 1560–1569. doi:10.1109/BigData.2017.8258090.

[141] M. Canizo, E. Onieva, A. Conde, S. Charramendieta, S. Trujillo, Real-time predictive maintenance for wind turbines using Big Data frameworks, IEEE Int. Conf. Progn. Heal. Manag. (2017) 1–8. doi:10.1109/ICPHM.2017.7998308.

[142] T. Zaarour, N. Pavlopoulou, S. Hasan, U. ul Hassan, E. Curry, Grand challenge: Automatic anomaly detection over sliding windows, Proc. 11th ACM Int. Conf. Distrib. Event-Based Syst. - DEBS '17. (2017) 310–314. doi:10.1145/3093742.3095105.

[143] J.H. Lee, S. Do Noh, H.-J. Kim, Y.-S. Kang, Implementation of cyber-physical production systems for quality prediction and operation control in metal casting, Sensors (Switzerland). 18 (2018). doi:10.3390/s18051428.

[144] A. Stojadinović, Industry Paper : Dynamic Monitoring for Improving Worker Safety at the Workplace : Use Case from a Manufacturing Shop Floor, (2015) 205–216.

[145] H. Haskamp, F. Orth, J. Wermann, A.W. Colombo, Implementing an OPC UA interface for legacy PLC-based automation systems using the Azure cloud: An ICPS-architecture with a retrofitted RFID system, Proc. - 2018 IEEE Ind. Cyber-Physical Syst. ICPS 2018. (2018) 115–121. doi:10.1109/ICPHYS.2018.8387646.

[146] D. Cemernek, H. Gursch, R. Kern, Big data as a promoter of industry 4.0: Lessons of the semiconductor industry, Proc. - 2017 IEEE 15th Int. Conf. Ind. Informatics, INDIN 2017. (2017) 239–244. doi:10.1109/INDIN.2017.8104778.

[147] C. Ellwein, O. Riedel, O. Meyer, D. Schel, Rent'n'Produce: A Secure Cloud Manufacturing Platform for Small and Medium Enterprises, 2018 IEEE Int. Conf. Eng. Technol. Innov. ICE/ITMC 2018 - Proc. (2018) 1–6. doi:10.1109/ICE.2018.8436332.

[148] N. Ferry, G. Terrazas, P. Kalweit, A. Solberg, S. Ratchev, D. Weinelt, Towards a big data platform for managing machine generated data in the cloud, Proc. - 2017 IEEE 15th Int. Conf. Ind. Informatics, INDIN 2017. (2017) 263–270. doi:10.1109/INDIN.2017.8104782.

[149] A. Angrish, B. Starly, Y.S. Lee, P.H. Cohen, A flexible data schema and system architecture for the virtualization of manufacturing machines (VMM), J. Manuf. Syst. 45 (2017) 236–247. doi:10.1016/j.jmsy.2017.10.003.

[150] R.S. Peres, A.D. Rocha, A. Coelho, J. Barata Oliveira, A Highly Flexible, Distributed Data Analysis Framework for Industry 4.0 Manufacturing Systems, in: T. Borangiu, D. Trentesaux, A. Thomas, P. Leitão, J.B. Oliveira (Eds.), Serv. Orientat. Holonic Multi-Agent Manuf., Springer International Publishing, Cham, 2017: pp. 373–381.

[151] M.Y. Santos, J.O. e Sá, C. Costa, J. Galvão, C. Andrade, B. Martinho, F.V. Lima, E. Costa, A Big Data Analytics Architecture for Industry 4.0, WorldCIST 2017 Recent Adv. Inf. Syst. Technol. 0 (2017). doi:10.1007/978-3-319-56538-5.

[152] P. Tanuska, L. Spendla, M. Kebisek, Data Integration for Incidents Analysis in Manufacturing Infrastructure, in: Comput. Conf. 2017, 2017: pp. 340–345.

[153] H. Lin, J.A. Harding, C. Chen, A Hyperconnected Manufacturing Collaboration System Using the Semantic Web and Hadoop Ecosystem System, Procedia CIRP. 52 (2016) 18–23. doi:10.1016/j.procir.2016.07.075.

[154] C. Gröger, Building an Industry 4.0 Analytics Platform, Datenbank-Spektrum. (2018). doi:10.1007/s13222-018-0273-1.

[155] M.C. Domenech, L.P. Rauta, M.D. Lopes, P.H. Da Silva, R.C. Da Silva, B.W. Mezger, M.S. Wangham, Providing a smart industrial environment with the web of things and cloud computing, Proc. - 2016 IEEE Int. Conf. Serv. Comput. SCC 2016. (2016) 641–648. doi:10.1109/SCC.2016.89.

[156] A.V. Wan, J., Tang, S., Li, D., Wang, S., Liu, C., Abbas, H. and Vasilakos, A Manufacturing Big Data Solution for Active Preventive Maintenance, IEEE Trans. Ind. Informatics. 13 (2017) 2039–2047. doi:10.1201/b15906-13.

[157] N. Rivetti, Y. Busnel, A. Gal, Grand challenge: Flinkman - Anomaly detection in manufacturing equipment with apache flink, Proc. 11th ACM Int. Conf. Distrib. Event-Based Syst. - DEBS '17. (2017) 274–279. doi:10.1145/3093742.3095099.

[158] M. Canizo, E. Onieva, A. Conde, S. Charramendieta, S. Trujillo, Real-time Predictive Maintenance for Wind Turbines Using Big Data Frameworks, (2017) 1–8.

[159] M.R. Brule, Big data in EP: Real-time adaptive analytics and data-flow architecture, Soc. Pet. Eng. - SPE Digit. Energy Conf. Exhib. 2013. (2013) 305–311. doi:10.2118/163721-MS.

[160] D. Wu, J. Terpenny, L. Zhang, R. Gao, T. Kurfess, Fog-enabled architecture for data-driven cyber-manufacturing systems, ASME 2016 11th Int. Manuf. Sci. Eng. Conf. MSEC 2016. 2 (2016) V002T04A032. doi:10.1115/MSEC2016-8559.

[161] R. Gao, L. Wang, R. Teti, D. Dornfeld, S. Kumara, M. Mori, M. Helu, Cloud-enabled prognosis for manufacturing, CIRP Ann. - Manuf. Technol. 64 (2015) 749–772. doi:10.1016/j.cirp.2015.05.011.

[162] Y. Wu, S. Wang, Streaming Analytics Processing in Manufacturing Performance Monitoring and Prediction, 2017 IEEE Int. Conf. Big Data (Big Data). (2017) 3285–3289.

[163] Y. Zhang, S. Ren, Y. Liu, T. Sakao, D. Huisingh, A framework for Big Data driven product lifecycle management, J. Clean. Prod. 159 (2017) 229–240. doi:10.1016/j.jclepro.2017.04.172.

[164] M. Naeem, N. Moalla, Y. Ouzrout, A. Bouaras, An ontology based digital preservation system for enterprise collaboration, 2014 IEEE/ACS 11th Int. Conf. Comput. Syst. Appl. (2014) 691–698. doi:10.1109/AICCSA.2014.7073267.

[165] Q.P. He, J. Wang, D. Shah, N. Vahdat, Statistical Process Monitoring for IoT-Enabled Cybermanufacturing: Opportunities and Challenges, IFAC-PapersOnLine. 50 (2017) 14946–14951. doi:10.1016/j.ifacol.2017.08.2546.

[166] F. Tao, J. Cheng, Q. Qi, IIHub: An Industrial Internet-of-Things Hub Toward Smart Manufacturing Based on Cyber-Physical System, Ieee Trans. Ind. Informatics. 14 (2018) 2271–2280. doi:10.1109/TII.2017.2759178.

[167] S. Ramírez-Gallego, A. Fernández, S. García, M. Chen, F. Herrera, Big Data: Tutorial and guidelines on information and process fusion for analytics algorithms with MapReduce, Inf. Fusion. 42 (2018) 51–61. doi:10.1016/j.inffus.2017.10.001.

[168] S. Michael, SQL databases v. NoSQL databases, Commun. ACM. 53 (2010) 10–11. doi:10.1145/1721654.1721659.

[169] I. Kovačević, I. Mekterović, Alternative Business Intelligence Engines, Inf. Commun. Technol. Electron. Microelectron. (2017) 1617–1622. doi:10.23919/MIPRO.2017.7973638.

[170] P. Kannan, Beyond Hadoop MapReduce Apache Tez and Apache Spark, (2015).

[171] S. Chintapalli, D. Dagit, B. Evans, R. Farivar, T. Graves, M. Holderbaugh, Z. Liu, K. Nusbaum, K. Patil, B.J. Peng, P. Poulosky, Benchmarking streaming computation engines: Storm, flink and spark streaming, Proc. - 2016 IEEE 30th Int. Parallel Distrib. Process. Symp. IPDPS 2016. (2016) 1789–1792. doi:10.1109/IPDPSW.2016.138.

[172] J. Abadi, D., Agrawal, R., Ailamaki, A., Balazinska, M., Bernstein, P.A., Carey, M.J., Chaudhuri, S., Dean, J., Doan, A., Franklin, M.J. and Gehrke, The Beckman Report on Database Research, Commun. ACM. 59 (2016) 92–99.

[173] S.V. Ranawade, Online Analytical Processing on Hadoop using Apache Kylin, 12 (2017) 1–5.

[174] M. Islam, A.K. Huang, M. Battisha, M. Chiang, S. Srinivasan, C. Peters, A. Neumann, A. Abdelnur, Oozie: Towards a Scalable Workflow Management System for Hadoop, Proc. 1st ACM SIGMOD Work. Scalable Work. Exec. Engines Technol. - SWEET '12. (2012) 1–10. doi:10.1145/2443416.2443420.

[175] K. Sundaravarathan, P. Martin, D. Rope, M. McRoberts, C. Statchuk, MEWSE: Multi-engine Workflow Submission and Execution on Apache YARN, Proc. 26th Annu. Int. Conf. Comput. Sci. Softw. Eng. (2016) 194–200. http://dl.acm.org/citation.cfm?id=3049877.3049897.

[176] I. Suriarachchi, B. Plale, Crossing Analytics Systems : A Case for Integrated Provenance in Data Lakes, in: 2016 IEEE 12th Int. Conf. e-Science Crossing, IEEE, Baltimore, MD, USA, 2016: pp. 349–354. doi:10.1109/eScience.2016.7870919.

[177] M.O. Gökalp, K. Kayabay, M. Zaki, A. Koçyiğit, P.E. Eren, A. Neely, Big-Data Analytics Architecture for Businesses : a comprehensive review on new open-source big-data tools, (2017).

[178] B. Stein, A. Morrison, The enterprise data lake: Better integration and deeper analytics, PWC Technol. Forecast Rethink. Integr. (2014). http://www.pwc.com/us/en/technology-forecast/2014/cloud-computing/assets/pdf/pwc-technology-forecast-data-lakes.pdf.

[179] N. Miloslavskaya, A. Tolstoy, Big Data, Fast Data and Data Lake Concepts, Procedia Comput. Sci. 88 (2016) 300–305. doi:10.1016/j.procs.2016.07.439.

[180] R. Bose, V. Sugumaran, Application of Knowledge Management Technology in Customer Relationship Management, 10 (2003) 3–17. doi:10.1002/kpm.163.

[181] B.R. Ferrer, W.M. Mohammed, J.L.M. Lastra, A solution for processing supply chain events within ontology-based descriptions, IECON Proc. (Industrial Electron. Conf. (2016) 4877–4883. doi:10.1109/IECON.2016.7793020.

[182] M. Mercier, D. Glesser, Y. Georgiou, O. Richard, M. Mercier, D. Glesser, Y. Georgiou, O. Richard, B. Data, Big Data and HPC collocation : Using HPC idle resources for Big Data Analytics, in: Big Data (Big Data), 2017 IEEE Int. Conf., 2017: pp. 347--352.

[183] G. Adamson, L. Wang, M. Holm, P. Moore, Cloud manufacturing – a critical review of recent development and future trends, Int. J. Comput. Integr. Manuf. (2017) 1–34.

doi:10.1080/0951192X.2015.1031704.

[184] J. Lee, H.D. Ardakani, S. Yang, B. Bagheri, Industrial Big Data Analytics and Cyber-physical Systems for Future Maintenance & Service Innovation, Procedia CIRP. 38 (2015) 3–7. doi:10.1016/j.procir.2015.08.026.

[185] E. Knill, Quantum computing with realistically noisy devices, Nature. 434 (2005) 39–44. doi:10.1038/nature03350.

[186] Deutsche Kommission Elektrotechnik DKE; DIN e.V., German Standardization Roadmap Industry 4.0, (2016) 523.

[187] W. and S. Kagermann, H., Riemensperger, F., Hoke, D., Helbig, J., Stocksmeier, D., Wahlster, Recommendations for the Strategic Initiative Coordination and editing, 2015.

[188] NIST, Framework for Improving Critical Infrastructure Cybersecurity, 2018. doi:10.1109/JPROC.2011.2165269.

[189] P. Wang, R.X. Gao, Z. Fan, Cloud Computing for Cloud Manufacturing: Benefits and Limitations, J. Manuf. Sci. Eng. 137 (2015) 044002. doi:10.1115/1.4030209.

[190] H. Fang, Managing Data Lakes in Big Data Era, in: Cyber Technol. Autom. Control. Intell. Syst. (CYBER), 2015 IEEE Int. Conf., 2015: pp. 820–824.

[191] D. Bernstein, Containers and cloud: From LXC to docker to kubernetes, IEEE Cloud Comput. 1 (2014) 81–84. doi:10.1109/MCC.2014.51.

[192] K. Stouffer, J. Falco, K. Kent, Guide to supervisory control and data acquisition (SCADA) and industrial control systems security, NIST Spec. Publ. SP800-82. (2006) 800--82. doi:10.6028/NIST.SP.800.82.

[193] S. Nazir, S. Patel, D. Patel, Assessing and augmenting SCADA cyber security: A survey of techniques, Comput. Secur. 70 (2017) 436–454. doi:10.1016/j.cose.2017.06.010.

[194] H. Shih, W and Ludwig, The biggest challenges of data-driven manufacturing, Harv. Bus. Rev. (2016). https://hbr.org/2016/05/the-biggest-challenges-of-data-driven-manufacturing (accessed June 6, 2019).

[195] Allae Erraissi, Abdessamad Belangour, Abderrahim Tragha, Digging into Hadoop-based Big Data Architectures, Int. J. Comput. Sci. Issues IJCSI. 14 (2017) 52–59. doi:10.20943/01201706.5259.

[196] H. Fleischmann, S. Spreng, J. Kohl, D. Kisskalt, J. Franke, Distributed condition monitoring systems in electric drives manufacturing, 2016 6th Int. Electr. Drives Prod. Conf. EDPC 2016 - Proc. (2016) 52–57. doi:10.1109/EDPC.2016.7851314.

[197] J. Manyika, M. Chui, J. Bughin, R. Dobbs, P. Bisson, Marrs, Disruptive technologies: Advances that will transform life, business, and the global economy, McKinsey Glob. Insitute. (2013) 163. http://www.mckinsey.com/insights/business_technology/disruptive_technologies%5Cnhttp://www.chrysalixevc.com/pdfs/mckinsey_may2013.pdf.

[198] A. Burger, H. Koziolek, J. Rückert, M. Platenius-Mohr, G. Stomberg, Bottleneck Identification and Performance Modeling of OPC UA Communication Models, (2019) 231–242. doi:10.1145/3297663.3309670.

[199] Deutsche Kommission Elektrotechnik DKE; DIN e.V., German Standardization Roadmap Industry 4.0, 2016.

[200] D. Loshin, Data Integration, Bus. Intell. (2013) 189–210. doi:10.1016/B978-0-12-385889-4.00013-2.

[201] J.M. Gutierrez-Guerrero, J.A. Holgado-Terriza, IMMAS an industrial meta-model for automation system using OPC UA, Elektron. Ir Elektrotechnika. 23 (2017) 3–11.

doi:10.5755/j01.eie.23.3.18324.

[202] Y.H. Wu, S. De Wang, L.J. Chen, C.J. Yu, Streaming analytics processing in manufacturing performance monitoring and prediction, Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017. 2018-Janua (2018) 3285–3288. doi:10.1109/BigData.2017.8258312.

[203] A. Stojadinović, N. Stojanović, L. Stojanović, Dynamic monitoring for improving worker safety at the workplace, in: Proc. 9th ACM Int. Conf. Distrib. Event-Based Syst. - DEBS '15, 2015: pp. 205–216. doi:10.1145/2675743.2771881.

[204] S. Scholze, K. Nagorny, R. Siafaka, K. Krone, An Approach for Cloud-Based Situational Analysis for Factories Providing Real-Time Reconfiguration Services, in: 2017: pp. 118–127. doi:10.1007/978-3-319-65151-4.

[205] L. Angrisani, G. Ianniello, I. Elettrica, N. Federico, Cloud based system for measurement data management in large scale electronic production, (n.d.).

[206] G. Hesse, B. Reissaus, C. Matthies, M. Lorenz, M. Kraus, M. Uflacker, Senska – Towards an enterprise streaming benchmark, 2018. doi:10.1007/978-3-319-72401-0_3.

[207] L. Banica, A. Hagiu, 4 - Using big data analytics to improve decision-making in apparel supply chains A2 - Choi, Tsan-Ming BT - Information Systems for the Fashion and Apparel Industry, in: Woodhead Publ. Ser. Text., Woodhead Publishing, 2016: pp. 63–95. doi:https://doi.org/10.1016/B978-0-08-100571-2.00004-X.

[208] S. Saeidlou, M. Saadat, E.A. Sharifi, D. Guiovanni, An ontology-based intelligent data query system in manufacturing networks, Prod. Manuf. Res. 3277 (2017) 1–18. doi:10.1080/21693277.2017.1374887.

7 APPENDICES

	Search strings	Number of papers
Scopus	TITLE-ABS-KEY (("hadoop") OR ("spark") OR ("Storm") OR ("Flink") OR ("SQL") OR ("nosyl") OR ("time-series database") OR ("opc-usa") OR ("MTconnect")) AND (TITLE-ABS-KEY ("manufacturing") OR TITLE-ABS-KEY ("Industry 4.0") OR TITLE-ABS-KEY ("Smart manufacturing") OR TITLE-ABS-KEY ("Digital twin") OR TITLE-ABS-KEY ("Digital thread")) AND NOT (TITLE-ABS-KEY ("construction") OR TITLE-ABS-KEY ("healthcare") OR TITLE-ABS-KEY ("oil") OR TITLE-ABS-KEY ("energy") OR TITLE-ABS-KEY ("Agriculture"))	228
IEEE Xplore	(("Author Keywords":hadoop OR "Author Keywords":spark OR "Author Keywords":storm OR "Author Keywords":flink OR "Author Keywords":sql OR "Author Keywords":nosql OR "Author Keywords":"time-series database" OR "Author Keywords":"opc-ua" OR "Author Keywords":"mtconnect") AND ("Abstract":manufacturing OR "Abstract":"Industrial 4.0" OR "Abstract":"industrial automation" OR "Abstract":"smart manufacturing" OR "Abstract":"digital twin" OR "Abstract":"digital thread"))	63
ASME	(hadoop, spark, storm, flink, sql, nosql, "time series database", "opc-ua", mtconnect) AND (manufacturing, "industry 4.0", "industrial automation", "smart manufacturing", "digital twin", "digital thread")	25

ACM	((keywords.author.keyword:(hadoop, spark, storm, flink) AND content.ftsec:(apache)) OR keywords.author.keyword:(sql, nosql, "time series database", opc-ua, mtconnect)) AND recordAbstract:(manufacturing, "industrial automation", "smart manufacturing", "digital twin", "digital thread")	22

1 Table 1. Search strings of four citation databases

Categories	Systems	Reference
Product	CAD/CAE/CAPP/CAM	[55][53][56][125]
	PLM	[58][59]
Production	ERP	[126][62][62][97][94]
	MOM/MES	[25][65][127][128][69][66][116][6]
	SCADA/DCS/HMI	[137][73][71][39][134][201][133][131][130][129]
	IIoT/CNC/Robot	[166][202][78][77][140]
	O&M	[138][24][83][139][161][156][157][142][82]
	QMS	[85][88][109][87]
	Safety	[89][203]
Business	SCM/CRM/BI/AM	[93][90][181]
ICT architecture	CPS/CM/ICT	[64][146][68][160][153][99][132][101][143][75][148][95][149][204][96][76]
	Data analytics/DM	[105][121][205][102][152][206][151][207][208]
	KM	[164][23][106]

2 Table 2. Category distribution of reviewed articles