Ismail *et al. J Big Data* (2019) 6:1
https://doi.org/10.1186/s40537-018-0162-3

Journal of Big Data

SURVEY PAPER Open Access

Manufacturing process data analysis pipelines: a requirements analysis and survey

Ahmed Ismail*, Hong-Linh Truong and Wolfgang Kastner

*Correspondence:
aismail@auto.tuwien.ac.at
Faculty of Informatics,
Technische Universität Wien,
Karlsplatz 13, 1040 Vienna,
Austria

Abstract

Smart manufacturing is strongly correlated with the digitization of all manufacturing activities. This increases the amount of data available to drive productivity and profit through data-driven decision making programs. The goal of this article is to assist data engineers in designing big data analysis pipelines for manufacturing process data. Thus, this paper characterizes the requirements for process data analysis pipelines and surveys existing platforms from academic literature. The results demonstrate a stronger focus on the storage and analysis phases of pipelines than on the ingestion, communication, and visualization stages. Results also show a tendency towards custom tools for ingestion and visualization, and relational data tools for storage and analysis. Tools for handling heterogeneous data are generally well-represented throughout the pipeline. Finally, batch processing tools are more widely adopted than real-time stream processing frameworks, and most pipelines opt for a common script-based data processing approach. Based on these results, recommendations are offered for each phase of the pipeline.

Keywords: Big data, Smart manufacturing, Industry 4.0, Analysis pipelines, Industrial Internet of Things, Data-driven decision making, High performance computing

Introduction

Smart manufacturing is a manufacturing strategy that is principally based on the digitization of manufacturing related activities and the rapid conversion of data into information. Innovations in big data analysis can be used to support the quick data-driven decision making processes needed for today's turbulent markets [1–3].

Big data refers to the large volumes of structured, semi-structured, and unstructured data, acquired from a variety of heterogeneous sources [4]. This data is typically assumed to have the valuable information hidden in it because substantial efforts and resources are needed to uncover it [5, 6]. According to the U.S. National Institute of Science and Technology (NIST) Big Data Public Working Group (Reference Architecture Subgroup) [7], big data does not refer to the increasingly large datasets or the requirement for improved performance and efficiency. Instead, it refers to the fundamental reforms in the architecture needed to manage this data [1–3].

Big data analytics are currently used for many industrial applications. This includes product lifecycle management [8], process re-design [9], supply chain management [10], and production systems data analysis [11]. Of these, production systems analysis has

received a considerable amount of attention from academia and industry. According to Vodenčarević and Fett in [11], this is because "production systems are big sources of raw data that are often hard to model manually". A number of applications have thus emerged to investigate process data for process monitoring, anomaly detection, root cause analysis, and knowledge extraction [11].

Recent publications have proposed platforms, frameworks, and architectures to address the different functions needed for process data analysis [12]. However, manufacturing systems present unique requirements for big data analysis platforms. This includes the ability to acquire and process billions of values [11], apply stringent constraints for low and/or bounded latency processing, and uphold high requirements for data quality. These demands translate to imposed minimum standards for the tools and technologies used in big data analysis pipelines and processing frameworks.

The goal of this article is to assist data engineers in designing big data analysis pipelines for manufacturing process data. This is achieved by investigating two previously unaddressed research questions.

1. RQ1: What are the requirements for a big data analysis pipeline for manufacturing process data?
2. RQ2: What are the available big data analysis pipelines for process data in academic literature?

Thus, the first section of this article presents an overview on the smart manufacturing strategy. Next, the infrastructure, data sources, and characteristics of manufacturing systems, and the challenges facing the adoption of big data analysis solutions in enterprises are detailed. This is followed by an analysis on the requirements for big data platforms (RQ1) and on the recent big data platforms for process data analysis (RQ2). The results are discussed to explain the decisions and choices of the surveyed pipelines. Finally, this article also includes recommendations and conclusions based on the results of the analysis as well as a discussion on possible future work.

Background—Smart manufacturing

A manufacturing strategy, according to [3], is a framework for the design, organization, management, and development of a manufacturing enterprise's resources. It is used to focus the decisions of a company towards achieving a select number of characteristics that would continuously improve the company's competitive advantage.

According to [3], the need for a manufacturing strategy can be understood from the following five characteristics.

1. Manufacturing involves the majority of the enterprise's resources.
2. Manufacturing decisions often take a long time to have an effect. Thus, a long term perspective is needed to support them.
3. Reverting a manufacturing decision normally requires a substantial investment in both time and resources.
4. Manufacturing decisions impact the manufacturing characteristics and performance of a company. Thus, it has a direct effect on its ability to compete in the market place.

5. A strategic outlook is needed for manufacturing decisions to ensure that it supports the business strategy of the enterprise.

Smart manufacturing is focused on the optimized application of resources and the workforce to achieve the on-time production of high quality goods while maintaining the enterprise characteristics necessary for the company to control and quickly respond to internal and external stimuli. Smart manufacturing places special emphasis on the role of emerging technologies. Thus, it calls for the digitization of all manufacturing-relevant activities as well as the adoption of technologies such as big data analysis [1].

The digitization of manufacturing services increases the amount of internal and external data available to the enterprise. Internal data is generated by sources inside the manufacturing enterprise. This includes data from manufacturing equipment, automation systems, work pieces, and enterprise information management, typically hosted in a Manufacturing Execution System (MES) and Enterprise Resource Planning System (ERP). External data refers to environmental data accumulated from sources external to the enterprise. This includes data from the supply network, the government (e.g., legislature and incentive programs), strategic partners, distribution channels, and customers. Analyzing this data for data-driven decision making programs may lead to increases in productivity and profits. These programs may include the systematic analysis of data for yield management, product re-engineering, and predictive maintenance. The remainder of this section will discuss these three use cases.

Yield management

Yield is highlighted as a key performance indicator that impacts the product's price, profit margin, quality, and customer satisfaction level. As opposed to the strategy of yield models, low yields can be combated by detecting anomalous behavior, preferably during the product ramp up phase at low batch sizes. Previously, abnormal behavior could be detected by domain experts. However, the increased digitization and complexity of manufacturing functions has led to an enormous rise in the amount of data available. This has made the process of monitoring and analyzing process data for quality control an increasingly difficult task. In more detail, manufacturing plants typically implement alarm systems based on simple logic for the detection of abnormalities in the production process. These systems use predefined thresholds to detect and alert an operator on violation events. However, high rates of false alarms and alarm flooding are endemic problems in industry. For example, Vodenčarević et al. [11] report on alarm bursts of 200 alarms per second. Such rates overload plant operators and make it difficult to detect genuine root cause alarms. To combat these alarm floods, big data analysis techniques are proposed to mine production data and assist operators by suppressing redundant alarms and narrowing down the number of variables relevant to root cause detection [13, 14].

Product re-engineering

The increasing complexity of produced goods have led to a rise in demands by customers for maintenance, repair, and overhaul (MRO) services from product manufacturers. These MRO services are major cost factors that require concise engineering to ensure

a favorable whole-life cost to the manufacturer and customer. Product re-engineering for the effective and constant improvement of product design and maintenance requires closing the loop on product development, manufacturing operations, and customer relationship management from an engineering, economic, and social perspective. This depends on the overall ecosystem's ability to process large data sets including customer surveys, inspection reports, maintenance operations plans, production sensors and actuators, wireless devices, software logs, photographic systems, audio-capturing devices, and other sources of metrology [15, 16].

Predictive maintenance

Maintenance in manufacturing has evolved over time from reactive and preventative maintenance to predictive maintenance. Reactive or corrective maintenance refers to the act of fixing or replacing components once they break down. The cost associated with the damage and down time caused by component failures has led to the development of preventative maintenance. This involves the prevention of failures through the regular inspection and servicing of assets based on pre-defined intervals (time-based maintenance, TBM) or the condition of assets (condition-based maintenance, CBM). CBM is also known as predictive maintenance and most commonly involves the continuous collection and analysis of raw sensor data to detect faults in production equipment before it manifests as a failure. Predictive maintenance avoids unnecessary inspections by prompting interventions only when necessary. However, the data characteristics and possibility of alarm floods require data mining techniques for alarm masking and root cause detection. The latter may be used to isolate and fix the source of failures rather than the symptoms, thereby improving asset lifetimes [17].

Altogether, the hypothesis of smart manufacturing is that technological transformations would allow for the sought after leaner and more agile innovation cycles. However, a number of hindrances stand in the way of smart manufacturing. These are discussed in the next section.

Challenges to data analysis systems in manufacturing

The manufacturing ecosystem may be viewed as a multi-dimensional grouping of systems designed to support the various business, operations, engineering, maintenance, and training functions involved in the manufacturing process. Figure 1 illustrates an example network of technologies, life cycles, systems, users, and applications.

The lowest layer of Fig. 1 represents the various tools and equipment that produce data. The data sources range from cross-domain design and diagnostics applications to physical sensors and cyber-physical devices. These provide large amounts of data that can include technical, social, environmental, and other types of data that are spread across the entire manufacturing life cycle (layer 2). The tools and equipment compose functionally specific systems (layer 3) such that the data is distributed across diverse databases, tools, and systems for access and use by different users (layer 4) and for different purposes (layer 5).

The majority of this ecosystem is governed by numerous standards, a complex technical landscape, safety and security considerations, and regulations and legislation. Each of

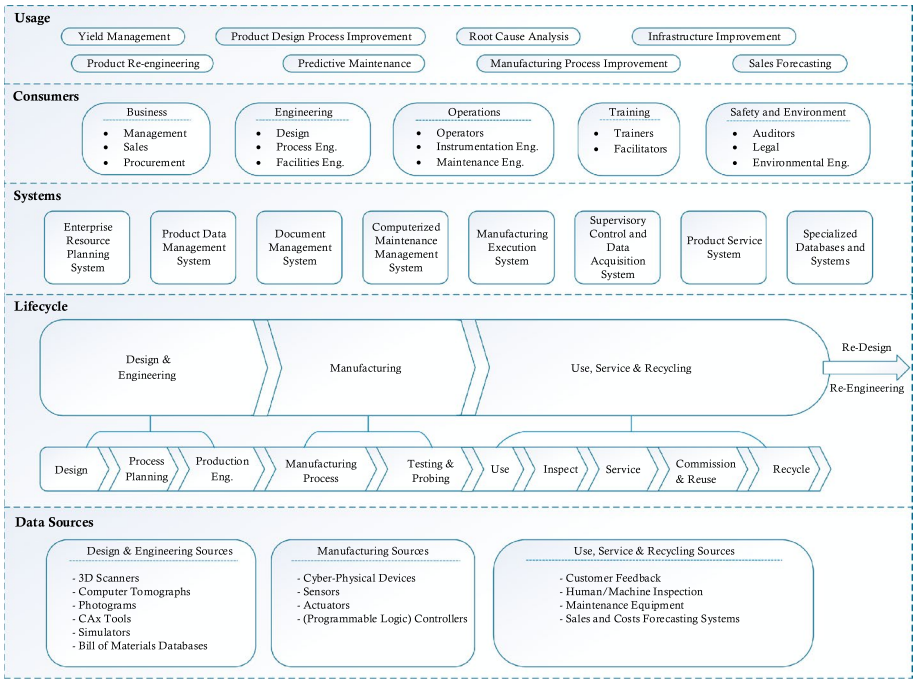

Fig. 1 A superimposed sample of different elements found in a manufacturing ecosystem

these factors pose a significant challenge to the design, implementation, and deployment of a digital platform for data analysis.

Complex standards landscape

Challenge 1: Integrating data of heterogeneous characteristics

Information integration is a challenging issue born of the multi-disciplinary nature of manufacturing. Specifically, a more accurate representation of layer 2 from Fig. 1 is the composition of three life cycles shown in Fig. 2. These are the product, production system, and supply chain life cycles.

The product development life cycle is concerned with creating the entity to be marketed and is associated with a collection of information flows and controls that span through design, process planning, production engineering, manufacturing, use and service, and end-of-life and recycling phases [1, 18].

The production systems are the means through which products are realized. They typically include systems of machines, equipment, and human labor that coordinate to convert resources into manufactured goods and services. The production system life cycle is generally much longer than product life cycles and involves design, build, commission, operation and maintenance, and decommission and recycling phases [1, 19].

Finally, the supply chain life cycle is focused on the flow of interactions and functions between customers, manufacturers, suppliers, and other entities impacting the business factors of the manufacturing ecosystem [1].

As shown in Fig. 2, each of these life cycles encompasses a large number of standards designed to meet the key needs of their activities, functions, and components. These standards encompass different protocols for communication and therefore represent

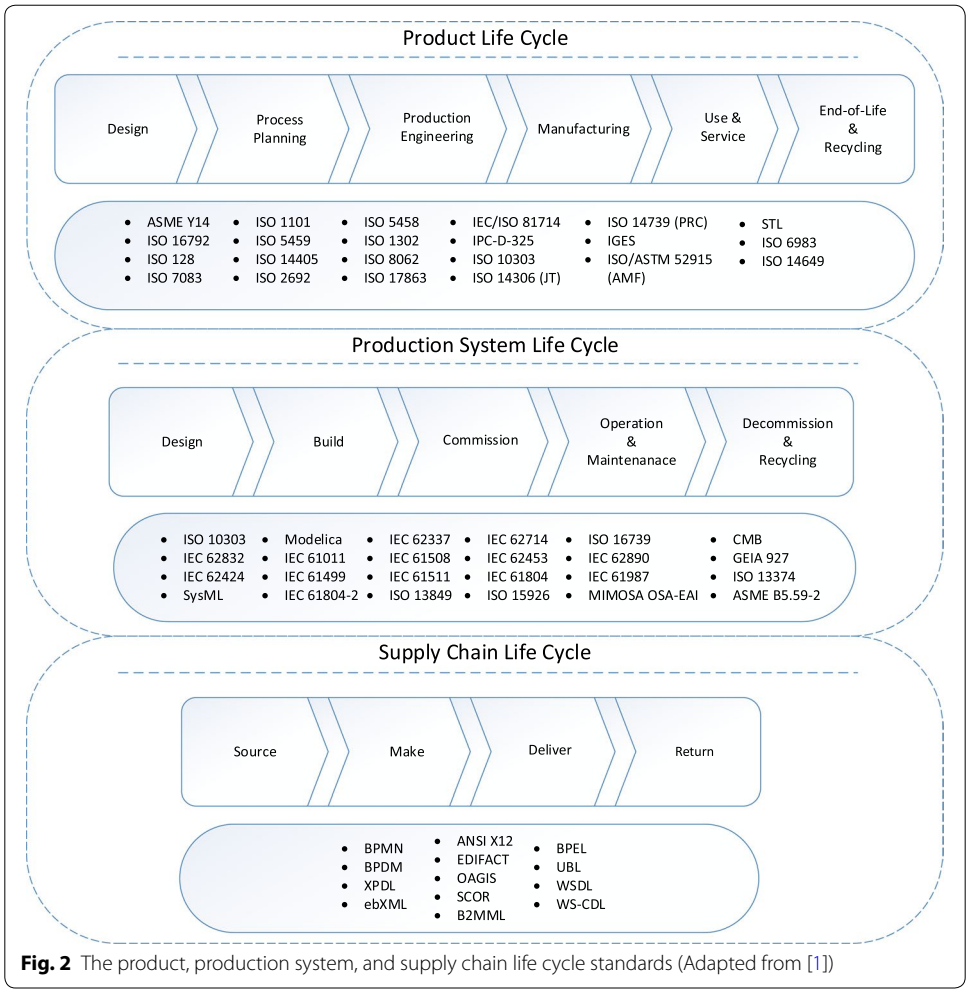

Fig. 2 The product, production system, and supply chain life cycle standards (Adapted from [1])

and exchange information in ways that are not strictly compatible with each other. Unfortunately, each of the life cycles has been treated as isolated dimensions and, consequently, work on the integration of information flows has primarily focused on establishing mechanisms for the exchange of information along each of these dimensions. Thus, the ability to integrate data of different formats and from widely diverse sources is a considerable challenge for data engineers.

Complex technical architecture

Challenge 2: Secure integration of the data analysis system in a "defense in depth" architecture

A manufacturing enterprise normally employs a layered architecture that segregates its infrastructure into multiple zones of operation, as shown in Fig. 3, based on the 5 layers of the automation pyramid (see Fig. 4). The infrastructure can be seen as a composition of two networks, an enterprise IT network (levels 4–5) and a factory automation network (levels 0–3) that are separated by a de-militarized zone (DMZ) for access control. The enterprise levels are where the majority of business processes are located. These use applications and protocols technologically similar to what is found in most other enterprises; i.e., using standard IT infrastructure. The latter, may incorporate

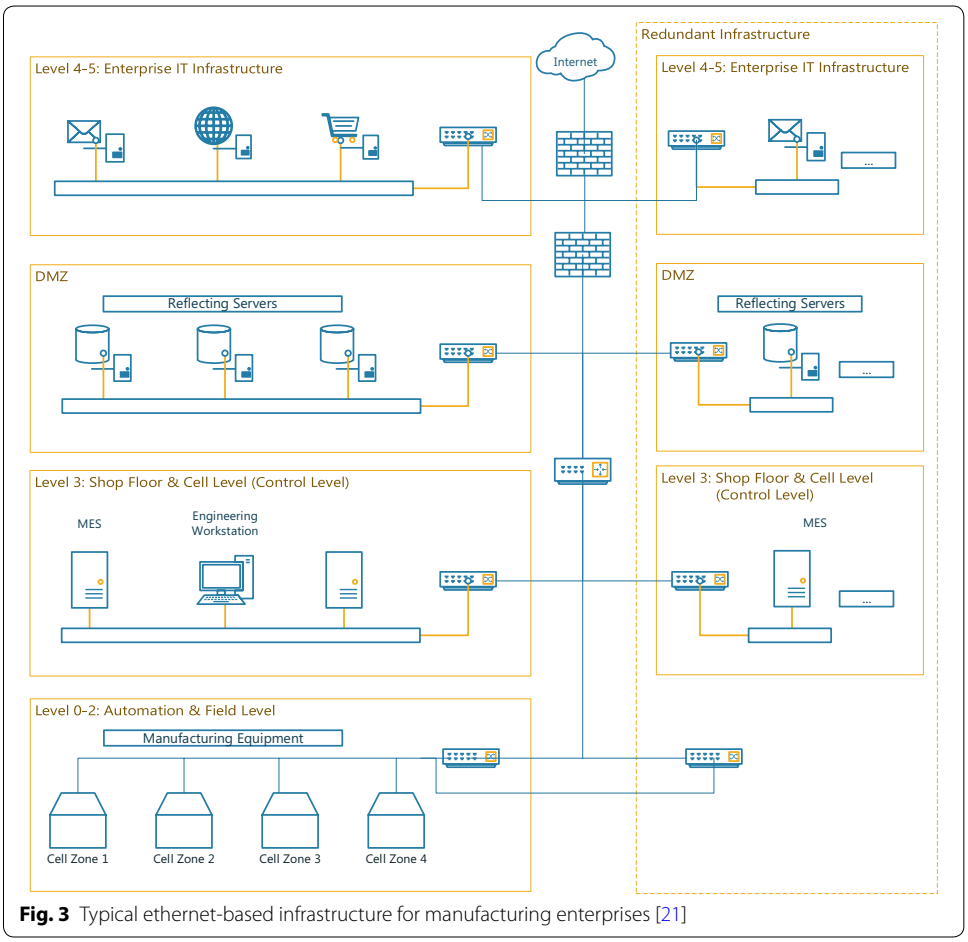

Fig. 3 Typical ethernet-based infrastructure for manufacturing enterprises [21]

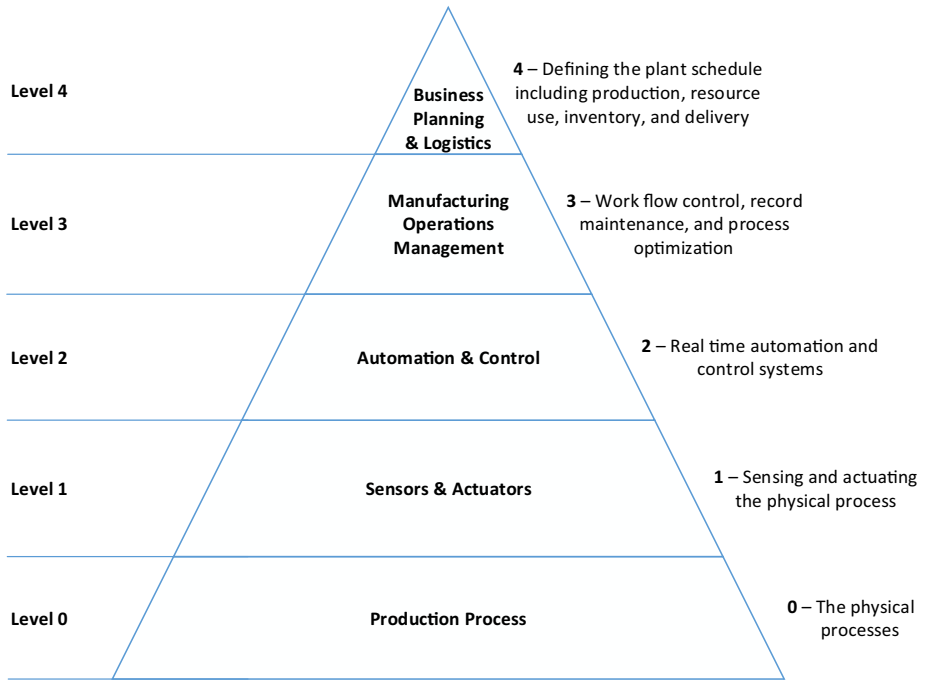

Fig. 4 The five layers of the automation pyramid (Reproduced with permission from [22])

domain-specific technologies to meet the unique requirements of manufacturing. This can include strict real-time and deterministic behavior, domain-specific physical infrastructure and topologies, and different expectations for the system's availability, reliability, safety, and security levels than what exists in the enterprise IT network [20].

This type of architecture is based on the concept of "defense in depth". This is a security approach that protects against the failure of one or more components in a level from cascading to other levels by installing security and other measures at the communication boundaries between levels. Thus, network traffic can freely flow between devices within the same layer while traffic that crosses between layers is strictly controlled. Any digital system deployed as part of the smart manufacturing strategy must therefore comply with the architecture and mirror the controls in place [20, 23].

Safety

Challenge 3: Integration of safety-critical data sources

Challenge 4: Using the data platform for safety-critical functions

There is a strong requirement for safety in manufacturing as malfunctions may result in serious accidents. The Union Carbide India Limited chemical plant disaster, which caused the death of 2000 people and the injury of over 50,000, stands out as a prominent example of the hazards of compromised safety and maintenance practices. Standards, such as S84, IEC 61508, and IEC 61511 exist to minimize the dangers of hardware, operator, and information errors [24].

A common industry practice is to employ a high degree of redundancy in the plant to increase the availability of the overall system. This is demonstrated by the redundant infrastructure shown in Fig. 3. In concrete terms, manufacturers typically deploy a separate control system, the Safety Instrumented System (SIS), in conjunction with a basic process control system (BPCS). The BPCS is responsible for the normal operation of the plant, yet, if the BPCS fails, and manual operator intervention also fails, the SIS then becomes responsible for safely returning the process to normal operating levels. An effective SIS requires that the number of components shared between the SIS and BPCS be kept at a minimum to avoid the cascading of failures between systems. Although it increases the reliability of the acquired data and imposed control over the plant, it also results in increased expenses [25].

The process components (e.g., transmitters and valves) that form the safety control loops, often referred to as Safety Instrumented Functions (SIF), are rated using a Safety Integrity Level (SIL) metric. This four level system ranges from 1 to 4. SIL 1 represents the worst possible level of safety and has a Required Safety Availability (RSA) of 90–99%. Each subsequent SIL level introduces an additional 9 to the RSA such that SIL 4, being the best level, requires an RSA of 99.99%–99.999%. As expected, every increment in the number of nines is more difficult to achieve. The safe and economic integration of data acquired from the SIS is a challenge that needs to be addressed convincingly for current technologies. This includes having strong guarantees throughout the data analysis pipeline for fault isolation, system availability, and data integrity [25, 26].

Regulations and legislation

Challenge 5: Data governance, data life cycle management, and liability

Manufacturing is regulated from the local to the international level for product and technical compliance, safety, health, and environmental protection [20]. From the information-centric perspective of digital systems, legislation for long-term data preservation is already in place for specific sectors. The US and Germany also employ cyber- and data breach notification laws [27]. Thus, strict data governance is considered an imposed requirement on current digital systems that necessitates data management policies for the entire information lifecycle. This lifecycle spans from "creation or receipt, [through] storage, distribution and transmittal, access and use, maintenance, disposition and destruction" [28]. However, data management in the manufacturing industry is further complicated by data sharing practices (between suppliers, manufacturers, and customers), privacy laws, liability, and IP protection. These are complex issues that are difficult to resolve and are associated with measures that are also hard to enforce [29]. Thus, there is a strong need for versatile and well-integrated controls for platform-wide data governance and policy enforcement.

This section presented some of the most prominent challenges to big data analysis platforms in manufacturing. While others exist (see [30]) the points discussed here are sufficient to highlight the difference between designing and employing a digital system for manufacturing as opposed to other candidate domains. In light of these challenges, the next section addresses the necessary requirements for a big data pipeline for manufacturing process data analysis.

Process data analysis

The manufacturing industry is characterized by highly sensitive and closely monitored production processes with extensive opportunities for big data analytics applications. Process operations and controls produce a variety of massive and complex data. This includes field level, control level, and quality control data that is collected using both direct and indirect measurement technologies. The data originating from the field and process levels are generally born of hierarchical manufacturing systems that produce structured time series data acquired through regular polling, with sampling possibly occurring in the range of milliseconds [16]. The data acquired from this level is typically expected to be structured, well-selected, and well-defined in both syntactic and semantic means. This is evident from the survey of [1], which demonstrates that standards for structuring information are abundant and apply across industries. In contrast, quality control data can either be structured data from property and quality measurement tests, or unstructured data from social media networks, customer feedback forms, interviews, surveys, and other communication channels [16].

As such, the product manufacturing process produces a large variety and volume of data from complex process operations. Typically, the flow of data follows the hierarchy (shown in Fig. 3) from the automation and field layers, to the control layer, and finally the enterprise layer. In contrast, control data flows in the reverse direction, as demonstrated in Fig. 5. Since the platform is focused on process data analysis, it is logical to assume that it will predominantly consume structured automation and field level data. However, quality control data may also be required from both the control and enterprise

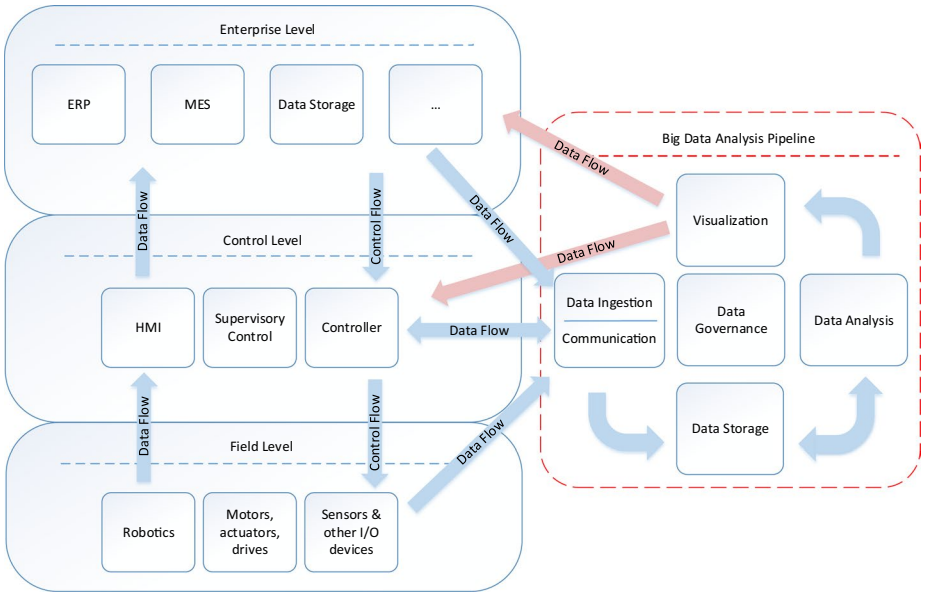

Fig. 5 Control and data flow between the manufacturing layers and a big data analysis pipeline

layers. These may be used to isolate component faults or to supplement analysis with information that is only available from these layers. For example, this may include controller debug logs, material properties, and customer feedback forms. The heterogeneous nature of the data will need to be ingested to formalize its structure and to filter, de-duplicate, and synchronize the different records. While it is typical for ingested data to be then directly sent for long term storage and analysis, it may also be returned to elements of the manufacturing infrastructure to be used or displayed in its pre-processed form. Next, the analysis of ingested data may include mining for knowledge in real time streams, processing for predictive analytics, and searching for historical patterns. It is typical for the analysis results to be stored while also being delivered through reporting and visualisation tools for real-time monitoring, alerting, and decision-making purposes. Hence, five basic components are expected of process data analysis platforms. These are data ingestion, communication, storage, analysis, and visualization [4, 31, 32].

The next subsection will address the first research question (RQ1) by characterizing the non-functional and functional requirements for each of these components and for process data analysis platforms as a whole. The requirements are summarized in Tables 1 and 2.

RQ1: What are the requirements for a big data analysis pipeline for manufacturing process data?

Data ingestion

The data ingestion component is the main entry point for data into the big data analysis platform. Thus, it is responsible for tasks such as the identification, validation, transformation, filtering, compression, noise reduction, and integration of incoming data [33]. Data cleaning is considered to be an extremely resource-intensive task that may consume 50% of the effort and 80% of the time in a data mining project [34]. It requires a concise understanding of the data sources to allow for the selection, cleaning, construction, and formatting of data.

Table 1 Non-functional requirements for the platform based on the FURPS+ model

	Parameter	Requirement
Functionality	Security	Compliant with legislative and regulatory requirements[a]
		Compliant with enterprise security policies
	Extensibility	Capable of integrating new interfaces, data types, connectors, and components
	Reusability	System functions should, at minimum, handle structured time-series data. The system should also have sufficient connectors to allow its reuse for new compositions of data and functions
Usability	Aesthetics	Generic and intuitive interfaces providing interpretable data. The various data types should be queried via a single interface. Visualisation interfaces should present data in multiple common formats (trend charts, bar charts, etc.) [11, 42]
	Documentation	Well-documented to assist in the reduction of system ambiguity and entropy, and to allow for system extensibility, component replacement, user training, etc
	Responsiveness	Limits on stream analysis response times depend on the use case and can range from milliseconds to seconds (alarms and eventing), to daily and weekly reports (process optimization)
Reliability	Accuracy	Intolerant of data and event loss
	Availability	Data acquisition, storage systems, and event processing and reporting should have the highest guarantees for availability
	Recoverability	Recovery of persisted data (raw and processed) is necessary. Speed of system recovery from faults and the resumption of functions is important
Performance	Throughput	Typically, this is in the order of 10s of Gigabytes (GB) per day
	Scalability	The system should scale to accommodate geographically dispersed sources/sinks
Supportability		The components should be well-maintained, stable, active, well-documented, and with a strong, supportive, and responsive user and developer community. They should also be compatible with well-established monitoring solutions

Robert Grady's FURPS + model. This acronym stands for functionality, usability, reliability, performance, and supportability (FURPS), and the + represents additional needs [43]

[a] Dependent on regulatory concerns imposed by local to international laws and agreement. Can impact storage and application locations, communication paths, security configurations, and other system features

Several challenges exist for manufacturing data ingestion. These include the following [5, 16].

1. Manufacturing ecosystems can have a large number of diverse protocols for communication. Protocol integration normally incurs high engineering costs. Therefore, a system that natively supports a large number of relevant connectors is very useful (I1, I2).
2. The ubiquity of proprietary protocols in manufacturing requires that the system be extensible with custom processors and connectors (I3).
3. The number of data sources and sinks may be in the thousands or hundreds of thousands implying the need for a scalable distributed solution (I4).
4. The presence of redundancy at the field level results in multiple readings of the same parameter, which may not be identical in values or sampling rates, thus requiring data synchronization (I5).
5. While certain sources are expected to provide real-time streams of data (I6), the scheduled transfer of data in bulk from back-end servers, repositories, and applications may also occur in a manufacturing setting (I7). In fact, the cleaned data may

Table 2 Platform requirements by function

Pipeline stage	Requirement
Ingestion	I1: Native support of a large number of technology connectors
	I2: Can ingest a large variety of formats
	I3: Supports custom processors and connectors
	I4: Scales to support a large number of sources and sinks (1000s to 100,000s)
	I5: Native processors for data validation, transformation, filtration, compression, noise reduction, identification, and integration
	I6: Supports active (real-time) ingestion
	I7: Supports passive (batch) ingestion
Communication	C1: Scalable It should be able to support a large number of sources (ms poll rate) and sinks. The combined number can range from 1000s to 100,000s
	C2: Secures data in transit
	C3: Exactly-once message delivery semantics
	C4: Publish-subscribe communication
	C5: Efficient bandwidth utilization
	C6: Supports both real-time data streams and bulk data transfer
	C7: Pull-based data consumption
Storage	S1: Scalable up to 10s GB/day
	S2: Read/Write speed independent of volume of stored data.
	S3: Large variety of formats and types (structured, semi-structured, and unstructured)
	S4: Compression features for cost-efficient long-term storage (years)
	S5: Intolerant of data loss
	S6: Secures stored data
	S7: Exports data to relational databases
Analysis [16]	A1: Scalable up to 100,000 variables
	A2: Heterogeneous data types
	A3: Imperfect data
	A4: Real-time and batch processing required
	A5: Supports time-series analysis & data mining and machine learning
Visualization	V1: Scalable
	V2: Visualization methods for large data volumes, variety, and velocity
	V3: Dynamic and static visualization
	V4: Interactive
	V5: Extensible interfaces

also need to be transferred back to some of these same databases for transactional operations.

Communication

While communication channels are necessary to transfer data between the different tools in the analysis pipeline, the need for a dedicated communication middleware is not as clear. In some cases, ingestion and communication tasks are handled by the same tool. Figure 5 acknowledges this possibility by displaying both functions in the same box. However, data may flow to an increasingly large number of destinations. Certain tools also split up or parallelize tasks across a large number of distributed workers. Having middleware decouple the ingestion component from other processes supports "asynchronous operations and [promotes] scalability, robustness and performance" [30].

Thus, using dedicated middleware for communication tasks is encouraged for big data systems.

The network infrastructure of a manufacturing plant is divided into layers based on function and requirements with strict controls on communication between them. The top layers are IP-based networks and the lowest layers use one or more field-level technologies [35]. Field-level technologies interconnect manufacturing equipment that can consist of thousands of data sources and sinks that produce enormous and continuous streams of data. For example, NXP Semiconductor's Assembly Plant in GuangDong is recorded to "[produce] millions of products per day, on a few thousand pieces of equipment, [collecting] over 26 Gigabytes of data per day" [36].

Furthermore, the sensory, actuation, and control devices embedded in manufacturing equipment are typically resource-constrained systems that are sensitive to any variations in the characteristics of network inputs. For example, two recorded incidents published by NIST in [37] demonstrate that unsolicited traffic (ping sweeps) in online control systems caused a robotic arm to switch from standby to active mode and swing around 180 degrees. The second incident, also involving a ping sweep, caused the control system in charge of fabricating integrated circuits to hang, and resulted in the loss of $50,000 of wafers.

In addition, the communication protocols used by control systems usually have limited or no security features[1] [39]. This means that access by an unauthorized individual may lead to unlimited access to the production equipment. Thus, manufacturing plants typically employ a defense in depth approach using firewalls, virtual private networks, intrusion detection systems, and other technologies to secure the control network [39].

Therefore, the communication system should be able to support thousands to hundreds of thousands of endpoints (C1). It should have a proven track record of operating safely in defense in depth architectures. It should also incorporate its own security measures to limit access to its data and functions and to secure data in transit (C2). This may include access control mechanisms and cryptography. Furthermore, since data loss may reduce the operating safety level of the plant, result in litigation, or degrade the quality of analysis, the middleware should provide strong guarantees for fault tolerance and availability. While this may imply at-least-once message delivery semantics, message duplicates should be avoided to prevent unnecessary load on resource-constrained devices. Thus, the ideal tool should support exactly-once message delivery semantics (C3). For efficiency, the tool should also support the publish/subscribe communication model and message compression and be capable of both real-time and bulk data transfer (C4, C5, C6). Finally, given the sensitivity of endpoints to unsolicited traffic, the tool should offer pull-based communication (C7).

Storage

As far back as 1997, the production of a single manufactured item could produce megabytes of sensor data in a single step of a single phase of production. As the number of units and production steps increase, the amount of data produced also increases. The

[1] See Section 22.5.2 of [38] for more details.

tools used throughout the process may vary, and therefore, may produce a variety of different data types. Fractions of this data may be highly correlated due to a multi-step production process or because of the wide deployment of redundant sensors. For reasons related to liability determination and regulations, this data may also need to be stored securely and for many years [33, 40].

Thus, there is a need for a database that is capable of scaling in the order of gigabytes per day (S1). It should be able to do so while maintaining fairly consistent per-second read and write speeds regardless of the volume of stored data (S2). The tool should also be able to handle highly correlated data of a large variety (S3). Due to the costs associated with long-term storage, it should be able to hold this data in an efficiently compressed state and in a secure manner (S4, S5, S6). Finally, it should be able to export to relational databases since manufacturing applications still depend on this type of databases (S7) [33, 40].

Analysis

The requirements for manufacturing data analysis described in this section are based on the perspective paper by Qin in [16].

According to Qin [16], process data analytics has heavily been in favor of modeling data using multivariate statistical methods such as the latent variable methods of principle components analysis (PCA) and projection to latent structures (PLS). Data reconciliation, neural networks, and time-series trend analysis are also explored and applied in industry for inferential modeling and performance assessment. Yet, these methods have been disconnected from recent developments in big data analysis, machine learning, and data mining. The benefit in progressing beyond the status quo lies in being able to use imperfect data in analysis, improving performance through deep learning techniques, exploiting underutilized time-series trend analysis methods for data mining, and extracting valuable information from largely unused unstructured data [16].

To achieve these improvements, data analysis should strive for scalability, versatility, simplicity, and timeliness. The requirements for analysis outlined by Qin include scalability up to a 100,000 variables (A1), the ability to analyze heterogeneous data types from diverse data sources (A2), and supporting analytical techniques for the real-time and batch processing of imperfect data (A3, A4). Finally, since the manufacturing process mainly generates time-series data, analysis should also favor tools that are designed to handle this type of data (A5) [16].

Visualization

Visualization involves the systematic representation of data. In [11], the results of data analytics were found to be more readily accepted by engineers if the models were interpretable and easy to visualize. According to [41], big data visualization requires different tools than traditional methods because of their differences in properties such as velocity, variety, and volume. The largeness of datasets implies the need for parallel visualization algorithms that can divide the workload into separate tasks that can be processed concurrently (V1). Visualization methods are also needed to meaningfully display structured, semi-structured, and unstructured data, as well as data of high complexity and

dimensionality (V2). Dynamic visualization is also needed to report on real-time processing streams (V3). According to [41], interactive tools are also stated as being more useful and leading to more discoveries than static ones (V4). Finally, the applied tools should have extensible interfaces to allow for cross-platform access (e.g., browsers, mobile devices) (V5).

This section discussed the main requirements for the five phases of manufacturing process data analysis pipelines. These non-functional and functional requirements are summarized in Tables 1 and 2. The next section addresses the second research question by discussing the existing data analysis pipelines in academic literature.

RQ2: What are the available big data analysis pipelines for process data in academic literature?

This subsection describes the big data analysis pipelines for manufacturing process data. Thus, it is divided in three parts. First, the search methodology used to source relevant literature is described. Then the inclusion and exclusion criteria applied are explained. Next, the main tools used in the different platforms are described.

Search methodology The primary search string ("Manufacturing" AND "Big Data") was used in the IEEE Xplore, ACM Digital Library, and Scopus digital databases. When possible, the search was limited to peer-reviewed English journal articles and conference papers in the computer science and engineering fields between the years 2014 and 2018. This yielded 939 unique papers.

Inclusion/exclusion criteria Four criteria are used to limit the number of surveyed papers. These are as follows.

1. *Recent* The paper should have been published within the last 5 years, including this year (i.e., 2014, 2015, 2016, 2017, 2018). Technologies change often, thus, the properties and capabilities of tools that justified their inclusion may change drastically between the time that a tool was included in a platform and when this survey was carried out. Therefore, it is not useful for this survey to aggregate platforms developed based on the requirements of systems that existed 10 years ago with those developed for today's needs.
2. *Manufacturing process data analysis* The reviewed platforms must be designed for manufacturing and specifically for process data analysis.
3. *Platform* The purpose of this survey is to review platforms. We base our definition of a platform on [12]. Thus, a paper qualifies if it is "research that provides a system with hardware and software components, which enables applications to execute" [12]. For this survey, we relax the definition to only software components and look specifically for platforms that address the full analysis pipeline from ingestion to visualization.
4. *Big data* The platforms should be designed specifically for big data use cases.

The full text of all 939 papers were manually inspected to ensure that they met the specified inclusion and exclusion criteria. Only 38 papers qualified. The 38 papers, their publication year, type (conference paper or journal article), industry, and use case are summarized in Table 3.

Table 3 Papers that satisfy the review criteria

Paper	Year	Type	Industry	Use case
[44]	2014	C	Agnostic	Model discovery and analysis
[45]	2014	C	Agnostic	Knowledge management
[46]	2014	C	Agnostic	Cloud manufacturing
[47]	2014	C	Agnostic	Anomaly detection
[48]	2014	C	SCM	Predictive maintenance
[49]	2015	C	Agnostic	Air quality
[50]	2015	A	Polymer	Yield optimization
[51]	2015	A	Cement	Performance monitoring
[52]	2015	A	Chemical agricultural recycling	Anomaly detection
[53]	2015	C	SCM	APC
[54]	2016	A	Agnostic	Agnostic
[55]	2016	C	Agnostic	Risk management
[56]	2016	C	Agnostic	Agnostic
[57]	2016	C	SCM	Yield improvement
[58]	2016	C	Agriculture	Quality control
[59]	2016	C	Printing	Anomaly detection
[60]	2016	C	Tire	Quality control
[61]	2017	A	Polymer	Quality control
[62]	2017	A	Die casting	Quality control
[63]	2017	A	Agnostic	Quality control
[64]	2017	A	SCM	APC
[65]	2017	C	SCM	Process monitoring
[66]	2017	C	Oil and gas	Predictive maintenance
[67]	2017	C	Agnostic	Prognostics
[68]	2017	C	Hydroelectric	Semantic integration
[69]	2017	C	Weichai Power Co., Ltd.	Quality management
[70]	2017	C	Machining	Energy use tool use and wear
[71]	2017	C	Agnostic	Energy use
[72]	2017	C	Polymer	Prognostics
[73]	2017	C	Automotive	Quality management
[74]	2018	A	SCM	Production planning
[75]	2018	A	Metal casting	Quality management
[76]	2018	A	Machining	Kanban
[77]	2018	A	Food	Event processing
[78]	2018	A	Agnostic	Supply chain management
[79]	2018	A	Agnostic	Agnostic
[80]	2018	C	Agnostic	OEE
[81]	2018	C	Agnostic	Agnostic

Type 'C' represents a conference paper and 'A' a journal article

SCM: semi-conductor manufacturing, APC: advanced process control, OEE: overall equipment effectiveness

Classification The papers are analyzed and the tools used in their respective pipelines are defined. These tools are classified by function into one of five stages: ingestion, communication, storage, analysis, and visualization. Table 4 shows the results of this classification.

Table 4 The tools used in the respective data analysis pipelines of each paper

Paper	Ingestion	Communication	Storage	Analysis	Visualization
[44]	Custom	–	HDFS, HBase, MongoDB Infinispan,	Hadoop, Hive, Pig, Elasticsearch	Custom
[45]	Custom	–	HDFS, MySQL	Hadoop	– (~)
[46]	WSO2 BAM	WSO2 ESB	HDFS, RDB (~), Cassandra (~)	Hadoop, WSO2 CEP	Custom (WSO2 UES)
[47]	–	Kafka	HDFS	Hadoop, Storm	–
[48]	–	–	HDFS, HBase, MongoDB Cassandra,	Hadoop, Hive	–
[49]	–	–	HDFS	Hadoop, Mahout, Jena Elephas	–
[50]	–	–	MySQL	Matlab, QuickCog	–
[51]	Custom	–	Microsoft SQL 2012	Custom	Custom
[52]	Custom	Kafka	HDFS, HBase	Hadoop, Storm, Hive, Radoop, Rapidminer	– (~)
[53]	Sqoop	–	HDFS, HBase	Hadoop, Hive, Impala	–
[54]	Sqoop	Flume	HDFS, HBase, MySQL	Hadoop, Hive	Custom
[55]	Custom	Custom	MongoDB	Custom	Custom
[56]	Custom	–	MongoDB, PostgreSQL	RStudio, Watson Analytics, Qliksense	Custom
[57]	Flume (~), Sqoop (~)	Custom	HDFS, HBase	Hadoop, Hive, Impala, Spark, Pig	Custom
[58]	Custom	Custom	Cassandra	Spark	Zeppelin (~)
[59]	Kafka	Kafka	Cassandra, OntoQUAD	Spark	Custom, Jupyter, Ontos Eiger
[60]	Custom	–	HDFS	Hadoop, Hive, Spark	Custom
[61]	Storm	Kafka	MongoDB	Storm	Custom
[62]	Pig, Hive	Custom	HDFS	Hadoop, Hive, Pig	Flamingo, Custom
[63]	ODI, Talend, Sqoop	Kafka	HDFS, HBase	Hadoop, Spark, IPython	Tableau, Microsoft BI
[64]	Sqoop, Custom	Custom	HDFS, RDB ~	Hadoop, Hive, Impala, Spark, Matlab	–
[65]	Custom	–	–	Custom (~)	Custom
[66]	Custom	Kafka, RabbitMQ	HDFS, HBase, Cassandra, PostgreSQL	Hadoop, Spark, Storm	Custom
[67]	Custom	Custom	Microsoft SQL 2008R2	Custom	Custom
[68]	Custom	–	Cassandra	Spark	–
[69]	Sqoop	–	HDFS	Spark	Custom
[70]	Custom, Storm	Kafka	CouchDB	–	–
[71]	Sqoop	–	HDFS, HBase	Hadoop, Hive, Pig	Custom
[72]	Custom	–	MongoDB	Custom	–
[73]	WSO2 ESB	WSO2 ESB	Alfresco CMS, Neo4j	Apache UIMA, WEKA	Custom
[74]	– (~)	–	HDFS, HBase	Hadoop, Hive	–
[75]	Spark	–	HDFS, HBase	R, Drools	Custom
[76]	Custom	–	MySQL	Custom	Custom
[77]	Custom	–	Microsoft SQL 2008R2	Custom	Custom
[78]	Flume, Sqoop	–	HDFS	Hadoop, Hive, Solr, RServe, Mahout	Custom
[79]	Flume	Kafka	HDFS, HBase, MySQL	Hadoop, Hive, Storm	Custom

Table 4 (continued)

Paper	Ingestion	Communication	Storage	Analysis	Visualization
[80]	–	Kafka	Cassandra	Spark	Custom
[81]	Custom	RabbitMQ	HDFS	Hadoop	–

~, implies that it is uncertain if the tool was used for this stage of the pipeline. –, implies that no tool was used for this stage of the pipeline

ODI: Oracle Data Integrator, BAM: Business Activity Monitor, ESB: Enterprise Service Bus, CEP: Complex Event Processor, UES: User Engagement Server

Results

A number of anecdotal observations collected during this survey include the following.

1. Papers often do not state the requirements of their systems prior to the design of the pipeline.
2. Papers often do not justify their design choices and decisions in tools.
3. Papers often do not describe how the tools were applied and used. This is especially true for custom tools that are also not openly accessible.

This restricts the information available to explain the observed trends in this study.

Yet, the results of RQ1 define the requirements that should be met by a data analysis pipeline for manufacturing process data. They are agnostic to industry and use case. Thus, process data analysis pipeline should at the least meet these requirements and those that are specific to the industry and use case. Therefore, the results of RQ1 are used to establish the context needed to explain the pipelines in RQ2.

The discussion is split into two sections that give an overview of the results and a more in-depth discussion on the tools used and design choices made for the analysis pipelines, respectively.

Results overview

Figure 6 shows an overview on the tools used in the different pipelines and demonstrates the following. The analysis and storage stages of the pipeline are well-addressed by most paper with respectively only 1 and 3 papers out of 38 not using a tool (custom or otherwise) for the task. In contrast, 6 pipelines do not have a tool set for ingestion (15.8%), 19 for communication (50.0%), and 13 for visualization (34.2%). Thus, a considerable number of papers focused predominantly on the analysis and storage phase and, in the process, neglected the ingestion, communication, and visualization stages.

Results by analysis stage

This section does not compare the tools and pipelines used against the requirements defined in RQ1 for two reasons.

1. Figure 6 shows the frequent use of custom and ad-hoc tools in the different pipelines. These tools are not openly available for review.
2. The descriptions of how the tools are applied vary drastically in quality and detail between papers. Therefore, while the tools may be evaluated independent of how

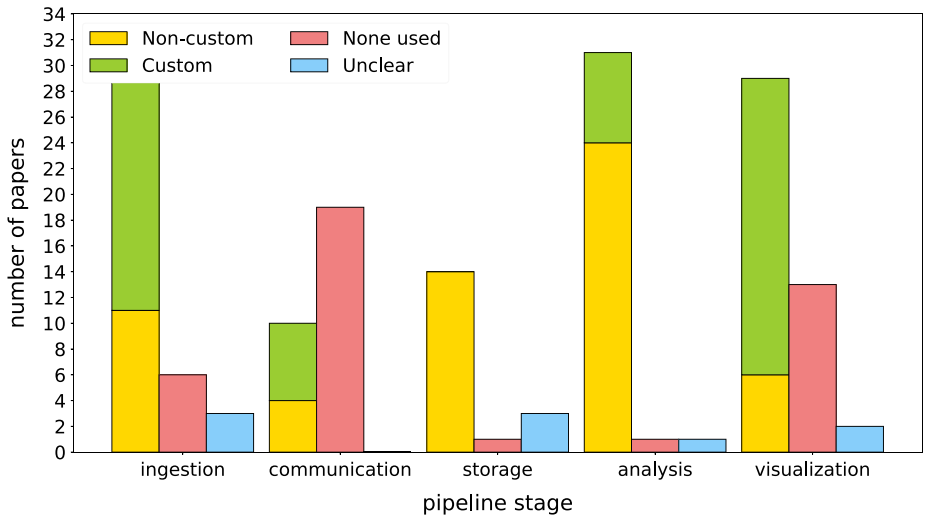

Fig. 6 An overview of the tools used. The yellow bar is the number of unique and non-custom made tools (e.g., NiFi, Kafka). The green bar is the number of custom tools (e.g., ad hoc scripts/applications). The red bar shows the number of pipelines that did not use any tool for that specific stage. The blue bar counts the number of cases of uncertainty in the results (e.g., a paper includes a tool in its figure of the data analysis pipeline but then does not mention it in-text)

they are actually used, this would not give a realistic and fair assessment of the pipelines themselves.

The rest of this section will instead describe the trends in Fig. 7 in light of the requirements from RQ1.

Ingestion (Fig. 7a) 18 of the 38 pipelines (47.3%) use custom tools. Custom tools are therefore the largest type of tool used for data ingestion. This design choice may be due to the nature of manufacturing ecosystems. To explain, a manufacturing ecosystem may employ a number of protocols and standards for communication and data representation, some of which are proprietary. The protocols, standards, and the combination in which they are used can be unique to the manufacturing industry. Since these properties would directly impact the characteristics of the ingested data, the response has been to develop custom and ad-hoc connectors for data ingestion.

Apache Sqoop is the second most used tool for data ingestion. It is used in 8 out of 38 pipelines (21.0%). Apache Sqoop is used to integrate Hadoop directly with existing relational databases and manufacturing operations management systems (e.g., ERP, MES) that commonly depend on relational databases. Since, Apache Sqoop is a big data tool for the transfer of data between relational databases and Hadoop, explaining this trend is straight forward.

The third most common design choice is to not use any tool and to leave the ingestion phase unaddressed. This is done in 6 of the 38 pipelines (15.8%).

The remainder of the tools include message queuing middleware (e.g., Flume, Kafka), processing frameworks (e.g., Storm, Spark), query engines (e.g., Pig, Hive), and others. These are each applied in 3 pipelines or less (7.9%).

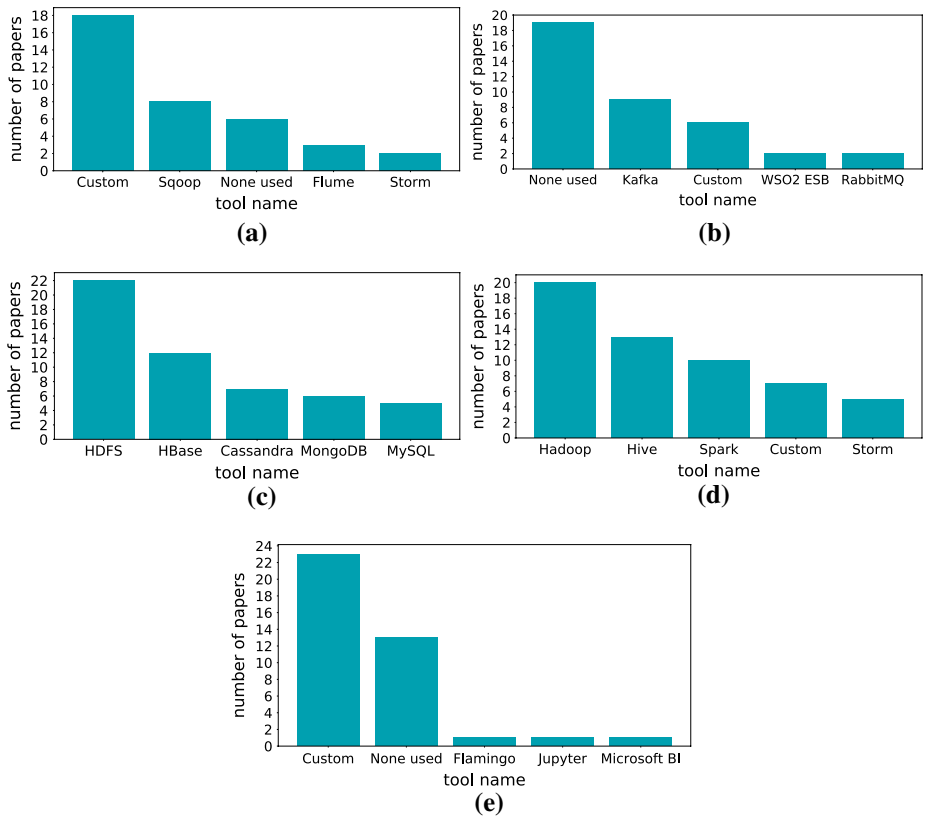

Fig. 7 The top 5 tools used for **a** ingestion, **b** communication, **c** storage, **d** analysis, and **e** visualization in the pipelines

Communication (Fig. 7b) 18 out of 38 pipelines (47.3%) did not use any tool for communication. Instead, the pipelines would chain the different tools using their connectors or respective interfaces.

Apache Kafka is the second most used tool and is applied in 9 out of 38 pipelines (23.7%). Kafka is a message queuing solution. Thus, it can be used to decouple the pipeline components and promote scalability, robustness, and performance [30]. Kafka is a good choice based on the requirements defined in RQ1 for communication middleware. However, Kafka may be difficult to integrate in systems where data flows back to the process and control levels from the analysis pipeline. To explain, Apache Kafka publishes messages from producers to queues ('topics'), that are then processed by consumers. Instances of Kafka consumers operate in consumer groups. If a consumer group is subscribed to a topic, each published record is delivered to a single consumer instance in that group. Yet, manufacturing systems often have redundant equipment that needs to be in sync and up to date. If the redundant components operate as a single consumer group, then only one component out of the redundant set will receive each message from the Kafka topic. Additional mechanisms will therefore be necessary to ensure that all of the redundant components are in sync. Alternatively, each component will have to operate as an independent consumer group. This is one example of a caveat that requires extra care to engineer a correct system for message queuing in manufacturing.

Custom tools are used in 6 of 38 pipelines (15.8%). This is normally seen as integrated communication interfaces in middleware and other components of the pipeline. However, several papers also construct pipelines that predominantly use custom tools. Thus, having a custom tool for the communication stage is unsurprising as it conforms with their all-encompassing design choice.

The remainder of the tools used (e.g., WSO2 ESB, and RabbitMQ) are each applied in 2 pipelines or less (5.3%).

Storage (Fig. 7c) HDFS is used in 22 out of 38 pipelines (57.9%), thus making it the most popular choice for storage in the surveyed pipelines. HDFS is bundled with big data analysis tools (e.g., Hadoop) that depend on HDFS to function. It is therefore unsurprising that it is so well-represented in the analysis platforms.

NoSQL (Not Only SQL) databases are used in 27 out of 38 pipelines (71.1%). The databases used are HBase, Cassandra, MongoDB, Infinispan, and CouchDB. NoSQL databases are normally described as optimal solutions for storing heterogeneous data. Some papers, however, have cited scalability as their main reason for using a NoSQL database instead of a relational one.

Relational databases are used in 12 out of 38 pipelines (31.6%). The tools used include MySQL, Microsoft SQL Server, and PostgreSQL. Two papers did not specify the exact type and referred only to a relational database in the pipeline. Relational databases are traditionally common in manufacturing systems. Thus, this choice in database systems may be driven by historical reasons and context.

One pipeline, described in [65], did not address the storage stage and did not explain this decision.

Analysis (Fig. 7d) Hadoop is used in 20 of 38 pipelines (52.6%). This number is inflated because Hadoop is often used to execute jobs on behest of other tools. For example, Hive is used in 13 of 38 pipelines (34.2%) and runs on top of Hadoop, submitting SQL queries to Hadoop for processing. Similarly, Pig submits SQL-like jobs to Hadoop and is used in 4 separate pipelines (10.5%). However, several pipelines do use Hadoop as intended for batch processing tasks.

Spark and Storm are used for stream processing in 10 pipelines (26.3%) and 5 pipelines (13.2%) out of 38, respectively. Using this type of tools conforms with the well-known fact that the majority of data from field level equipment is structured time series data produced through regular polling.

Custom tools are used in 7 of 38 pipelines (18.4%). This number represents applications and ad hoc scripts written in a variety of languages.

It is worth noting that this stage shows the highest diversity of tools. In total, 24 different tools (excluding custom tools) are used in the 38 pipelines. Yet, 19 of the 24 tools are each used in 3 pipelines or less.

Finally, 1 of the 38 pipelines (2.6%) does not use a tool for this stage. This pipeline is described in [70]. This paper focuses on establishing a pipeline for ingestion, communication, and storage. Plans exist to hand off the stored and cleaned data to an analysis component in the future. However, the analysis setup is not specified as of yet.

Visualization (Fig. 7e) A custom tool is used in 23 of the 38 pipelines (60.5%). This includes web frameworks developed in a number of languages to display web pages on diverse terminals.

The second most common choice is to have no tool assigned for the visualization stage of the pipeline. This is done in 13 of the 38 pipelines (34.2%). Instead, they depend on the outputs of the analysis tools used or leave it unaddressed.

The commercial off-the-shelf (COTS) tools used are Flamingo, Jupyter, Microsoft BI, Ontos Eiger, Tableau, and Zeppelin. These are each used in 1 pipeline.

Recommendations

This section presents the main recommendations for each stage in the pipeline based on the study's results.

Ingestion Custom tools are the largest type of tool used for data ingestion. This may highlight the lack of readily available industry-wide capable technology connectors for ingestion or the need for specialized tools in general. For example, the latter may include tools that have direct support for proficiently handling highly correlated and redundant process level data. Developing a standard and openly available tool with the required features can remove the redundancy of redeveloping fundamental components, such as protocol connectors.

Communication 19 out of the 38 pipelines (50.0%) neglect the communication phase. Using an enterprise-level message queuing service would introduce middleware that can decouple the components in the pipelines, and thus promote scalability, robustness and performance [30]. Also, 6 of the 38 pipelines (15.8%) use custom tools for communication. Using a COTS service would relieve data engineers from the cost of having to redevelop communication logic for custom tools.

Storage 27 out of the 38 pipelines (71.1%) use NoSQL databases. This is an understandable design choice since the context dictates support for storing and analyzing heterogeneous data. However, NoSQL databases should provide a familiar interface similar to relational databases, e.g., the ability to strictly enforce data schema, since it is more likely that in-house expertise in the manufacturing sector are more aligned with relational systems.

Analysis Real-time processing frameworks are under-represented in this stage of the pipeline. This limits the capabilities of the overall system and may, in the future, require the re-engineering of systems. Incorporating a suitable real-time processing tool in the initial design is normally justified since it is highly relevant to a number of current manufacturing use cases.

Visualization In [11], data analysis results were found to be more readily accepted by engineers if the models were interpretable and easy to visualize. Having no tool assigned for this stage or depending on the output of analysis tools instead of using standardized interfaces for the systematic representation of data antagonizes these findings. Thus, this aspect should be addressed in accordance to the requirements defined in RQ1.

Conclusion

This survey identifies and addresses two research questions with the goal of supporting data engineers in the development of big data analysis pipelines for manufacturing process data. The first research question addresses the requirements for big data analysis pipelines for manufacturing process data. The second research question surveys the available pipelines in academic literature.

Most pipelines focus on the analysis and storage phases, and neglect the ingestion, communication, and visualization stages. Furthermore, custom tools are frequently used for ingestion and visualization. While these trends may currently be justified in the manufacturing context, it highlights an opportunity for the development of standardized and openly accessible tools. Moreover, tools capable of handling heterogeneous data are well-represented. Storage and analysis tools for relational data are also well-represented. Finally, batch processing tools are more widely adopted than real-time stream processing frameworks, and most pipelines tackle the analysis phase using a common script-based data processing approach.

The derived recommendations are as follows.

1. Data ingestion is in need of a suitable tool with the standard technology connectors and common features necessary for manufacturing.
2. A COTS enterprise-level message queuing solution should be used for communication to free-up developers from having to re-implement message queuing logic between pipelines. It also ensures that the overall system can decouple the components in the pipelines, thereby promoting features such as scalability, robustness, and performance.
3. For storage, NoSQL databases with a familiar interface (e.g., similar to relational databases) should be favored over others. This would allow companies to capitalize on existing in-house expertise that are typically in relational systems for historical reasons.
4. The analysis stage should strive to include a stream processing tool in the pipeline since it is relevant to most use cases on manufacturing process data.
5. The visualization stage should not be left unaddressed so that the data analysis results are more accessible to engineers.

Future work can include a complete comparison of the tools identified in this survey against the requirements of RQ1. This future comparison may help further determine a set of tools that are best-suited for the big data analysis of manufacturing process data. They may then serve as a good basis for future development and standardization efforts.

Abbreviations

ERP: enterprise resource planning; MES: manufacturing execution system; MRO: maintenance, repair and overhaul; TBM: time-based maintenance; CBM: condition-based maintenance; DMZ: de-militarized zone; SIS: safety instrumented system; BPCS: basic process control system; SIF: safety instrumented functions; SIL: safety integrity level; RSA: required safety availability; NIST: national institute of science and technology; PCA: principal components analysis; PLS: projection to latent structures; FURPS: functionality, usability, reliability, performance, and supportability; GB: gigabytes; SCM: semi-conductor manufacturing; APC: advanced process control; OEE: overall equipment effectiveness; ODI: oracle data integrator; BAM: business activity monitor; ESB: enterprise service bus; CEP: complex event processor; UES: user engagement server; NoSQL: not only SQL; COTS: commercial off-the-shelf.

Authors' contributions

AI identified the background, challenges, and research questions, designed and conducted suitable research methodologies, collected, inspected, and classified the papers and tools, formulated the discussion, recommendations, and conclusion, and wrote the manuscript. HLT identified topics, edited the manuscript, and supervised the research. WK edited the manuscript and supervised the research. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the reviewers. AI thanks G. Gridling and R. Trubko for their helpful discussions.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
Not applicable.

Funding
The paper has been partially funded by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie Grant Agreement No. 764785 (Fog Computing for Robotics and Industrial Automation). The authors acknowledge the TU Wien University Library for financial support through its Open Access Funding Program.

Publisher's Note

Received: 24 September 2018 Accepted: 7 December 2018
Published online: 07 January 2019

References

1. Lu Y, Morris K, et al. Current standards landscape for smart manufacturing systems. In: Tech. rep. NIST IR 8107. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8107.
2. Choudri A. The Agile enterprise. In: ReVelle J, editor. Manufacturing handbook of best practices: an innovation, productivity, and quality focus. New York: CRC Press; 2001. p. 3–23.
3. Corrêa H. Agile manufacturing as the 21st century strategy for improving manufacturing competitiveness. In: Gunasekaran A, editor. Agile manufacturing: the 21st century competitive strategy. Oxford: Elsevier Science Ltd; 2001. p. 3–23.
4. Rehman M.H.u, Chang V, et al. Big Data reduction framework for value creation in sustainable enterprises. In: International journal of information management. 2016. p. 917–28. https://doi.org/10.1016/j.ijinfomgt.2016.05.013.
5. Jirkovsky V, Obitko M, et al. Big Data analysis for sensor time-series in automation. In: International conference on emerging technology and factory automation (ETFA). IEEE; 2014. p. 1–8.
6. Han J, Kamber M. Data mining: concepts and techniques. 3rd ed. Burlington: Elsevier; 2011 ISBN: 978-0-12-381479-1.
7. NIST Big Data Interoperability Framework. Volume 1, definitions. In: Tech. rep. NIST SP 1500-1. National Institute of Standards and Technology; 2015. https://doi.org/10.6028/NIST.SP.1500-1.
8. Li J, Tao F, et al. Big Data in product lifecycle management. Int J Adv Manuf Technol. 2015;81:667–84. https://doi.org/10.1007/s00170-015-7151-x.
9. Palma-Mendoza JA, Neailey K. A business process re-design methodology to support supply chain integration: application in an airline MRO supply chain. Int J Inform Manag. 2015;35:620–31. https://doi.org/10.1016/j.ijinfomgt.2015.03.002.
10. Hazen BT, Boone CA, et al. Data quality for data science, predictive analytics, and big data in supply chain management: an introduction to the problem and suggestions for research and applications. Int J Prod Econ. 2014;154:72–80. https://doi.org/10.1016/j.ijpe.2014.04.018.
11. Vodenčarević A, Fett T. Data analytics for manufacturing systems. In: International conference on emerging technology and factory automation (ETFA). IEEE; 2015. p. 1–4.
12. O'Donovan P, Leahy K, et al. Big Data in manufacturing: a systematic mapping study. J Big Data. 2015;1:1. https://doi.org/10.1186/s40537-015-0028-x.
13. Chien C-F, Liu CW, et al. Analysing semiconductor manufacturing Big Data for root cause detection of excursion for yield enhancement. Int J Prod Res. 2017;55:5095–107. https://doi.org/10.1080/00207543.2015.1109153.
14. Mannila H, Toivonen H, et al. Discovery of frequent episodes in event sequences. Data Mining Knowl Dis. 1997;1(3):259–89.
15. Stark R, Grosser H, et al. Advanced technologies in life cycle engineering. Procedia CIRP. 2014;22:3–14. https://doi.org/10.1016/j.procir.2014.07.118.
16. Qin SJ. Process data analytics in the era of big data. AIChE J. 2014;60:3092–100. https://doi.org/10.1002/aic.14523.
17. Puig Ramírez J. Asset optimization and predictive maintenance in discrete manufacturing industry. MA thesis. Universitat Politècnica de Catalunya; 2010.
18. Biffl S, Gerhard D, et al. Introduction to the multi-disciplinary engineering of cyber-physical production systems. In: Multi-disciplinary engineering for cyber-physical production systems. Oxford: Springer; 2017.
19. Colombo A, Bangemann T, et al. Industrial cloud-based cyber-physical systems. Cham: Springer; 2014.
20. Ismail A, Kastner W. Vertical integration in industrial enterprises and distributed middleware. Int J Internet Protocol Technol. 2016;9(2/3):79–89. https://doi.org/10.1504/IJIPT.2016.079547.
21. Ismail A, Kastner W. Discovery in SOA-Governed Industrial Middleware with mDNS and DNS-SD. In: International conference on emerging technology and factory automation (ETFA). IEEE. 2016.
22. Ismail A. Service oriented manufacturing infrastructure. Dissertation. Vienna: TU Wien; 2018.
23. Zerbst J, Schaefer M, et al. Zone principles as cyber security architecture element for smart grids. In: 2010 IEEE PES innovative smart grid technologies conference Europe (ISGT Europe). 2010. https://doi.org/10.1109/ISGTEUROPE.2010.5638900.
24. Bowonder B. An analysis of the Bhopal accident. Project Appraisal. 1987;2(3):157–68.
25. Spooner M, MacDougall T. Safety Safety instrumented systems. Can they be integrated but separate?" In: ABB White Paper. 2011.

26. Liptak BG, Venczel K, et al. Instrument and Automation Engineers' Handbook. Process measurement and analysis. 5th ed. Boston: CRC Press; 2016.
27. Raul AC. The privacy, data protection and cybersecurity law review. English. 2014. ISBN: 978-1-909830-28-8.
28. Bernard R. Information lifecycle security risk assessment: a tool for closing security gaps. Comput Secur. 2007;26:26–30.
29. Appt S, Fietz E, et al. Smart manufacturing. The legal and regulatory challenges. Pinsent: Pinsent Masons LLP; 2015.
30. O'Donovan P, Leahy K, et al. An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities. J Big Data. 2015;1:1. https://doi.org/10.1186/s40537-015-0034-z.
31. Didier P, Macias F, et al. Converged plantwide ethernet (CPwE) design and implementation guide. Milwaukee: Cisco Systems, San Jose, Californiaand Rockwell Automation; 2011.
32. Krumeich J, Werth D, et al. Advanced planning and control of manufacturing processes in steel industry through big data analytics: case study and architecture proposal. In: 2014 IEEE international conference on Big Data (Big Data). 2014. p. 16–24.
33. Sawant N, Shah H. Big Data application architecture Q & A: a problem-solution approach. Expert's voice in big data. New York: Apress. 2013. ISBN: 978-1-4302-6292-3.
34. Gupta G. Introduction to Data Mining with case studies. English. 2015. ISBN: 978-81-203-5002-1.
35. Sauter T, Soucek S, et al. Vertical integration. In: Wilamowski B, Irwin J, editors. Industrial communication systems. 2nd ed. London: CRC Press; 2011. p. 1–12.
36. Wilschut T, Adan I.J, et al. Big Data in daily manufacturing operations. In: IEEE simulation conference (WSC). 2014. p. 2364–75.
37. Stouffer K, Pillitteri V, et al. NIST special publication 800-82 revision 2: guide to industrial control systems (ICS) security. Tech. rep. National Institute of Standards and Technology. 2015.
38. Granzer W, Treytl A. Industrial Communication Systems. In: Irwin J, editor. Security in industrial communication systems. New York: CRC Press; 2011. https://doi.org/10.1201/b10603-24.
39. Kuipers D, Fabro M. Control systems cyber security: defense in depth strategies. Department of Homeland Security: Tech. rep. Prepared by Idaho National Laboratory. U.S; 2006.
40. Staff NRC. Massive data sets: proceedings of a workshop. Washington: National Academies Press; 1900.
41. Wang L, Wang G, et al. Big Data and visualization: methods, challenges and technology progress. Dig Technol. 2015;1:33–8. https://doi.org/10.12691/dt-1-1-7.
42. Lee F, Smith S. Yield Analysis and Data Management Using Yield ManagerTM. In: IEEE/SEMI 1998 IEEE/SEMI Advanced semiconductor manufacturing conference and workshop (Cat. No.98CH36168). 1998. pp. 19–30. https://doi.org/10.1109/ASMC.1998.731377.
43. Grady R, Caswell D. Software metrics: establishing a company-wide program. New Yrok: Prentice Hall; 1987 ISBN: 0138218447.
44. Yang H, Park M, et al. A system architecture for manufacturing process analysis based on big data and process mining techniques. In: International conference on Big Data. 2014. p. 1024–9. https://doi.org/10.1109/BigData.2014.7004336.
45. Wang C, Zhao C, et al. A framework for management of massive knowledge in cloud environment. In: International conference on BioMedical engineering and informatics (BMEI). 2014. p. 843–7. https://doi.org/10.1109/BMEI.2014.7002889.
46. Qanbari S, Zadeh S, et al. CloudMan: a platform for portable cloud manufacturing services. In: International conference on Big Data (Big Data). IEEE; 2015. p. 1006–14. https://doi.org/10.1109/BigData.2014.7004334.
47. Chen H, Fei X, et al. Energy consumption data based machine anomaly detection. In: International conference on advanced cloud and big data (CBD). IEEE. 2015. p. 136–42. https://doi.org/10.1109/CBD.2014.24.
48. Munirathinam S, Ramadoss B. Big data predictive analtyics for proactive semiconductor equipment maintenance. In: International conference on Big Data (Big Data). IEEE. 2015. p. 893–902. https://doi.org/10.1109/BigData.2014.7004320.
49. Obitko M, Jirkovský V. Big data semantics in industry 4.0. In: Lecture Notes in Computer Science 9266 2015. p. 217–29. https://doi.org/10.1007/978-3-319-22867-9_19.
50. Kohlert M, König A. Large, high-dimensional, heterogeneous multi-sensor data analysis approach for process yield optimization in polymer film industry. Neural Comput Appl. 2015;26(3):581–8. https://doi.org/10.1007/s00521-014-1654-5.
51. Dutta D, Bose I. Managing a big data project: the case of Ramco cements limited. Int J Prod Econ. 2015;165:293–306. https://doi.org/10.1016/j.ijpe.2014.12.032.
52. Windmann S, Maier A, et al. Big data analysis of manufacturing processes. J Phys. 2015;1:1. https://doi.org/10.1088/1742-6596/659/1/012055.
53. Moyne J, Samantaray J, et al. Big data emergence in semiconductor manufacturing advanced process control. In: Annual SEMI advanced semiconductor manufacturing conference (ASMC). IEEE. 2015. p. 130–5. https://doi.org/10.1109/ASMC.2015.7164483.
54. Wan J, Tang S. Software-defined industrial internet of things in the context of industry 4.0. IEEE Sens J. 2016;16(20):7373–80. https://doi.org/10.1109/JSEN.2016.2565621.
55. Niesen T, Houy C, et al. Towards an integrative big data analysis framework for data-driven risk management in industry 4.0. In: International conference on system sciences, Vol. 2016. 2016. p. 5065–74. https://doi.org/10.1109/HICSS.2016.627.
56. Gerrikagoitia J, Unamuno G, et al. Making sense of manufacturing data. In: International conference on informatics in control, automation and robotics, vol. 2. SciTePress. 2016. p. 590–4.
57. Chen C-C, Hung M-H, et al. Development of a cyber-physical-style continuous yield improvement system for manufacturing industry. In: International conference on automation science and engineering. IEEE. 2016. p. 1307–12. https://doi.org/10.1109/COASE.2016.7743559.
58. De Silva PP, De Silva PA. Ipanera: an industry 4.0 based architecture for distributed soil-less food production systems. In: Manufacturing & industrial engineering symposium (MIES). IEEE. 2016. p. 1–5.

59. Huber M, Voigt M, et al. Big data architecture for the semantic analysis of complex events in manufacturing. In: Lecture Notes in Informatics (LNI), Proceedings—series of the Gesellschaft fur Informatik (GI). p. 353–60.
60. Shi Y, Chen Y, et al. A data services-based quality analysis system for the life cycle of tire production". In: Lecture notes in computer science 9936 2016. p. 715–29. https://doi.org/10.1007/978-3-319-46295-0_51.
61. Syafrudin M, Fitriyani N, et al. An open source-based real-time data processing architecture framework for manufacturing sustainability. In: Sustainability. 2017. p. 2139. https://doi.org/10.3390/su9112139.
62. Lee JY, Yoon JS, et al. A big data analytics platform for smart factories in small and medium-sized manufacturing enterprises: an empirical case study of a die casting factory. Int J Precis Eng Manuf. 2017;18:1353–61. https://doi.org/10.1007/s12541-017-0161-x.
63. Lade P, Ghosh R. Manufacturing analytics and industrial Internet of Things. IEEE Intell Syst. 2017;32(3):74–9.
64. Lin Y-C, Hung M-H. Development of advanced manufacturing cloud of things (AMCoT)—a smart manufacturing platform. IEEE Robot Automat Lett. 2017;2(3):1809–16. https://doi.org/10.1109/LRA.2017.2706859.
65. Fan X, Zhu X, et al. Big data analytics to improve photomask manufacturing productivity. In: International conference on industrial engineering and engineering management, Vol. 2017. IEEE, 2018. p. 2341–5. https://doi.org/10.1109/IEEM.2017.8290310.
66. Stojanovic L, Stojanovic N. PREMIuM: Big data platform for enabling self-healing manufacturing. In: International conference on engineering, technology and innovation. IEEE, 2018. p. 1501–8. https://doi.org/10.1109/ICE.2017.8280060.
67. Yan J, Meng Y, et al. Big-data-driven based intelligent prognostics scheme in industry 4.0 environment". In: Prognostics and system health management conference (PHM-Harbin), 2017. IEEE. 2017. p. 1–5.
68. Jirkovský V, Obitko M. Enabling Semantics within Industry 4.0. In: Mařík V, Wahlster W, editors. Industrial applications of holonic and multi-agent systems, vol. 10444. Cham: Springer; 2017. p. 39–52. https://doi.org/10.1007/978-3-319-64635-0_4.
69. Li X, Tu Z, et al. Deep-level quality management based on big data analytics with case study. In: 2017 Chinese Automation Congress (CAC). 2017. p. 4921–6.
70. Ferry N, Terrazas G, et al. Towards a big data platform for managing machine generated data in the cloud. In: International Conference on Industrial Informatics (INDIN). IEEE. 2017. https://doi.org/10.1109/INDIN.2017.8104782.
71. Xu W, Liu Q, et al. Energy condition perception and Big Data analysis for industrial cloud robotics. In: Procedia CIRP 61 2017. p. 370–5. https://doi.org/10.1016/j.procir.2016.11.164.
72. Kozjek D, Vrabič R, et al. A data-driven holistic approach to fault prognostics in a cyclic manufacturing process. In: Procedia CIRP 63: 2017. p. 664–9. https://doi.org/10.1016/j.procir.2017.03.109.
73. Kassner L, Gröger C, et al. The Stuttgart IT Architecture for Manufacturing. In: Hammoudi S, Maciaszek LA, editors. Enterprise Information Systems, vol. 291. Cham: Springer; 2017. p. 53–80. https://doi.org/10.1007/978-3-319-62386-3_3.
74. Wang J, Yang J, et al. Big data driven cycle time parallel prediction for production planning in wafer manufacturing. In: Enterprise information systems. 2018. p. 714–32. https://doi.org/10.1080/17517575.2018.1450998.
75. Lee J, Noh S, et al. Implementation of cyber-physical production systems for quality prediction and operation control in metal casting. In: Sensors. 2018. p. 1428. https://doi.org/10.3390/s18051428.
76. Ding K, Jiang P. RFID-based production data analysis in an IoT-enabled smart job-shop. IEEE/CAA J Autom Sinica. 2018;5(1):128–38. https://doi.org/10.1109/JAS.2017.7510418.
77. Li S, Chen W, et al. ASPIE: a framework for active sensing and processing of complex events in the internet of manufacturing things. In: Sustainability. 2018. p. 692. https://doi.org/10.3390/su10030692.
78. Noh K-S. Model of knowledge-based process management system using big data in the wireless communication environment. Wireless Personal Commun. 2018;98:3147–62. https://doi.org/10.1007/s11277-017-4769-z.
79. Bai Y. Industrial Internet of things over tactile Internet in the context of intelligent manufacturing. Cluster Computing. 2018;21:869–77. https://doi.org/10.1007/s10586-017-0925-1.
80. Arantes M, Bonnard R, et al. General architecture for data analysis in industry 4.0 using SysML and model based system engineering. In: International systems conference (SysCon). IEEE. 2018. p. 1–6.
81. Kirmse A, Kraus V, et al. An architecture for efficient integration and harmonization of heterogeneous, distributed data sources enabling big data analytics. In: International conference on enterprise information systems. INSTICC. SciTePress. 2018. p. 175–82. ISBN: 978-989-758-298-1. https://doi.org/10.5220/0006776701750182.