O'Donovan *et al. Journal of Big Data* (2015) 2:25
DOI 10.1186/s40537-015-0034-z

Journal of Big Data
a SpringerOpen Journal

RESEARCH Open Access

An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities

P. O'Donovan*, K. Leahy, K. Bruton and D. T. J. O'Sullivan

*Correspondence:
peter_odonovan@umail.ucc.ie
IERG, University College Cork, Cork, Ireland

Abstract

The term smart manufacturing refers to a future-state of manufacturing, where the real-time transmission and analysis of data from across the factory creates manufacturing intelligence, which can be used to have a positive impact across all aspects of operations. In recent years, many initiatives and groups have been formed to advance smart manufacturing, with the most prominent being the Smart Manufacturing Leadership Coalition (SMLC), Industry 4.0, and the Industrial Internet Consortium. These initiatives comprise industry, academic and government partners, and contribute to the development of strategic policies, guidelines, and roadmaps relating to smart manufacturing adoption. In turn, many of these recommendations may be implemented using data-centric technologies, such as Big Data, Machine Learning, Simulation, Internet of Things and Cyber Physical Systems, to realise smart operations in the factory. Given the importance of machine uptime and availability in smart manufacturing, this research centres on the application of data-driven analytics to industrial equipment maintenance. The main contributions of this research are a set of data and system requirements for implementing equipment maintenance applications in industrial environments, and an information system model that provides a scalable and fault tolerant big data pipeline for integrating, processing and analysing industrial equipment data. These contributions are considered in the context of highly regulated large-scale manufacturing environments, where legacy (e.g. automation controllers) and emerging instrumentation (e.g. internet-aware smart sensors) must be supported to facilitate initial smart manufacturing efforts.

Introduction

A 2011 report on big data authored by McKinsey Global Institute, an economic and business research arm of McKinsey and Company, highlighted big data analytics as a key driver in the next wave of economic innovation [1]. However, the report suggests that this innovation may be impeded by a shortage of personnel with the skills needed to derive insights from big data, with demand in the US predicted to double between 2008 and 2018. This prediction seems credible when current data growth estimates are considered, with one estimate suggesting that the worlds data is doubling approximately every 1.5 years [2], and another estimate proposing that 2.5 quintillion bytes of

data are being produced each day [3]. This exponential growth in data can be attributed to a number of technological and economic factors, including the emergence of cloud computing, increased mobile and electronic communication, as well as the overall decreased costs relating to compute and data resources. In addition, emerging technology paradigms, such as the internet of things (IoT), which focus on embedding intelligent sensors in real-world environments and processes, will result in further exponential data growth. In 2011 it was estimated that more than 7 billion interconnected devices were in operation, which was greater than the world's population at that time. However, given the potential applications of IoT across numerous sectors and industries, including manufacturing, engineering, finance, medicine, and health, the number of interconnected devices in circulation is expected to rise to 24 billion by 2020 [4]. Therefore, given the anticipated shortage of personnel that are capable of managing this exponential data growth, there is a need for tools and frameworks that can simplify the process.

As big data analytics permeates different sectors, the tools and frameworks that are needed to address domain-specific challenges will emerge. For example, modern large-scale manufacturing facilities utilise sophisticated sensors and networks to record numerous measurements in the factory, such as energy consumption, environmental impact and production yield. Given the existence of such data repositories, these facilities should be in a position to leverage big data analytics. However, a number of domain-specific challenges exist, including diverse communication standards, proprietary information and automation systems, heterogeneous data structures and interfaces, as well as inflexible governance policies regarding big data and cloud integration. These challenges coupled with the lack of inherent support for industrial devices, makes it difficult for mainstream big data tools and methods (e.g. Apache Hadoop, Spark, etc.) to be directly applied to large-scale manufacturing facilities. Although some of the aforementioned challenges are addressed by different commercial tools, their scope is typically limited to data (e.g. energy and environmental) that is needed to feed a particular application, rather than facilitate open access to data from across the factory. To address these constraints, as well as many more, a new interdisciplinary field known as smart manufacturing has emerged. In simple terms, smart manufacturing can be considered the pursuit of data-driven manufacturing, where real-time data from sensors in the factory can be analysed to inform decision-making. More generally, smart manufacturing can be considered a specialisation of big data, whereby big data technologies and methods are extended to meet the needs of manufacturing. Other prominent technology themes in smart manufacturing include machine learning, simulation, internet of things (IoT) and cyber physical systems (CPS).

The application of big data has been demonstrated in different areas of manufacturing, including production, supply chain, maintenance and diagnosis, quality management, and energy [5]. This paper focuses on maintenance and diagnosis because of the role it plays in promoting machine uptime, as well as the potential impact it can have on operating costs, with some estimates claiming equipment maintenance can exceed 30 % of total operating costs, or between 60 and 75 % of equipment lifecycle cost [6]. The role of equipment maintenance is an important component in smart manufacturing. Firstly, smart manufacturing revolves around a demand-driven, customer-focused and highly-optimised supply chain. Given the dynamic and optimised nature of such a

supply chain, there is an implied dependency on machine uptime and availability. Secondly, smart manufacturing promotes energy and environmentally efficient production. The amount of energy used by equipment can increase if it is operating in an inefficient state (e.g. increased range of motion). Thirdly, smart manufacturing aims to maximise production yield. Machinery that is not functioning as per its design specification may negatively impact production yield (e.g. scrapped product). Finally, equipment maintenance can have an overall positive impact on capital costs. The lifetime of machinery may be enhanced by limiting the the number of times it enters a state of disrepair, while on-going costs may be reduced by using predictive and preventative maintenance strategies to optimise the scheduling of maintenance activities.

This paper presents an industrial big data pipeline architecture, which is designed to meet the needs of data-driven industrial analytics applications focused on equipment maintenance in large-scale manufacturing. It differs from traditional data pipelines and workflows in its ability to seamlessly ingest data from industrial sources (e.g. sensors and controllers), co-ordinate data ingestion across networks using remote agents, automate the mapping and cleaning process for industrial sources of time-series data, and expose a consistent data interface on which data-driven industrial analytics applications can be built. The main contributions of this research are the identification of information and data engineering requirements which are pertinent to industrial analytics applications in large-scale manufacturing, and the design of a big data pipeline architecture that addresses these requirements—illustrating the full data lifecycle for industrial analytics applications in large-scale manufacturing environments, from industrial data integration in the factory, to data-driven analytics in the cloud. Furthermore, this research provides big data researchers with an understanding of the challenges facing big data analytics in industrial environments, and informs interdisciplinary research in areas such as engineering informatics, control and automation, and smart manufacturing.

Related work

Smart manufacturing

The term smart manufacturing refers to a data-driven paradigm that promotes the transmission and sharing of real-time information across pervasive networks with the aim of creating manufacturing intelligence throughout every aspect of the factory [7–10]. Experts predict that smart manufacturing may become a reality over the next 10–20 years. The objective of smart manufacturing is similar to that of traditional manufacturing and business intelligence, which focuses on the transformation of raw data to knowledge. In turn, this knowledge can have a positive effect on operations by promoting better decision-making. However, smart manufacturing can be delineated from traditional manufacturing intelligence given its extreme focus on real-time collection, aggregation and sharing of knowledge across physical and computational processes, to produce a seamless stream of operating intelligence [11]. In simple terms, smart manufacturing can be considered an intensified application of manufacturing intelligence, where every aspect of the factory is monitored, optimised and visualised [7]. More expansive descriptions of smart manufacturing and its constituent parts can be found here [11, 12]. The level of digitalisation derived through smart manufacturing can facilitate radical transformations, such as;

- *Knowledge-embedded facilities*—embedding traditional facilities with knowledge and intelligence with the aim of enabling 'smart' operations across the factory.
- *Predictive and preventive operations*—transforming operations from reactive and responsive, to those that are predictive and preventative.
- *Performance-based operations*—promoting performance over compliance, with an emphasis on minimising energy and material usage, while maximising sustainability, health and safety, and economic competitiveness.
- *Distributed intelligence*—replacing vertical and localised decision-making with collective intelligence, where decisions are made to benefit the entire organisation.
- *Multi-disciplinary workforce*—removing the boundaries of vertical factory operations and cultivating a culture of an interdisciplinary workforce.
- *Next-generation IT departments*—retraining personnel to handle data-intensive, real-time and internet-aware infrastructures and technologies, such as real-time IoT sensors and big data technologies.

While these high-level transformations may seem achievable at first glance, it is generally accepted that the practicalities of smart manufacturing adoption are simply too complex for any single organisation to address [7]. Therefore, a number of groups and initiatives emerged in recent years to address the challenges and support the adoption of smart manufacturing.

Groups and initiatives

There are currently a number of government, academic and industry groups focused on of smart manufacturing. The most prominent of which include the Smart Leadership Coalition (SMLC) [11], Technology Initiative SmartFactory [13], Industry 4.0 [14], and The Industrial Internet Consortium (IIC). The emergence of these initiatives stemmed from the realisation that smart manufacturing is simply too large for any single organisation to address [7], and while the terminology used across these initiatives may differ, they share an overarching vision of real-time, digitalised and data-driven smart factories, which rely on sophisticated simulation and analytics to optimise operations.

The most prominent initiatives found in research are the SMLC and Industry 4.0, with each initiative loosely connected to their geographical origins (i.e. US and EU). The SMLC working group is comprised of academic institutions, government agencies and industry partners. These diverse perspectives are an important characteristic of the SMLC as it ensures challenges relevant to the wider community are addressed. While technology roadmaps, recommendations and guidelines have been central to the SMLC's activities to date, they have also been involved in the development of a technology platform that implements these recommendations. Industry 4.0 is a high-tech strategy developed by the German government to promote smart manufacturing and the benefits that can be derived by the greater economy. The term Industry 4.0 can be considered a naming convention that serves to partition each industrial revolution, with 4.0 referring to an anticipated fourth revolution (i.e. smart manufacturing) that experts anticipate will come to pass in the next 10–20 years. Expanding on this naming convention further, previous industrial revolutions can be labelled 1.0, 2.0 and 3.0. The first revolution (Industry 1.0) was brought about by the use of water and steam power to enable

mechanical production, with the first mechanical loom employed in 1784. The second revolution (Industry 2.0) was brought about by the use of electricity to realise mass production, which in turn promoted the division of labour in production processes. Finally, the third revolution (Industry 3.0) was brought about by advances in electronics and information systems that enabled control networks to automate the production process, with the first programmable logic controller (PLC) introduced in 1969.

Impact and benefits

Real-time and internet-aware pervasive networks, as well as highly integrated and intelligent data-driven analytics applications, are central to smart manufacturing. This highly measured and digitalised environment enables facilities to derive the knowledge needed to realise highly efficient and customised demand-driven supply chains, from the acquisition of raw materials, through to the delivery of the final product to the customer. In addition, smart manufacturing addresses many business and operating challenges that exist today, such as increasing global competition and rising energy costs, while also creating shorter production cycles that can quickly respond to customer demand [11, 15]. The SMLC outline several performance targets that relate to the aforementioned benefits, including (1) 30 % reduction in capital intensity, (2) up to 40 % reduction in product cycle times, and (3) overarching positive impact across energy, emissions, throughput, yield, waste, and productivity. Furthermore, smart manufacturing adoption can also benefit the greater economy. Recent research produced by the Fraunhofer Institute and Bitkom highlighted the potential benefits of Industry 4.0 to the German economy, where they estimated the transformation of factories to Industry 4.0 could be worth up to 267 billion Euros to the German economy by 2025 [16].

Roadmap for smart manufacturing

The progression to smart manufacturing can be decomposed into three distinct phases, with each phase deriving benefits that are exponentially greater than the last [17]. These sequential phases provide a high-level view of the journey to smart manufacturing adoption;

- *Phase 1—data integration and contextualisation.* Initially, facilities evaluate what data is available in different areas of the factory (e.g. sensors, controllers, databases, etc.) to form a global and contextualised view of data in the facility. Similarly to business intelligence and data warehouse projects, the process of integrating data can be a time-consuming and complex task. However, in manufacturing environments this complexity can be exacerbated due to the wide-range of industrial devices and protocols that can exist. The benefits of data integration and contextualisation alone may have a positive impact on operating costs, health and safety, and environmental factors.
- *Phase 2—simulation, modelling, and analytics.* Once data has been integrated and contextualised it can be processed and synthesised to create manufacturing intelligence that can inform decision-making. While data integration and contextualisation increases data visibility throughout the factory, data processing can be used to formulate actions that positively affect operations, which eclipse the benefits of fundamental data management alone. The potential benefits derived from advanced data

processing include flexible manufacturing, optimal production rates and enhanced product customisation.

- *Phase 3—process and product innovation.* As manufacturing intelligence is accumulated from data processing, new insights will begin to emerge from repositories of collective intelligence. In turn, these insights can inspire major innovations in processes and production. The benefits derived from such insights will be capable of causing significant market disruption that result in game-changing economics (e.g. 90 % reduction in retail price of laptop).

To surpass the benefits of current manufacturing intelligence systems, facilities must navigate each phase in the smart manufacturing roadmap. However, the early stages of adoption can be particularly challenging due to the complexity of industrial data integration, which may result in the effort-to-benefit ratio being perceived as low. Generally, the potential benefits from each phase (1–3) move from low to high. This is in contrast with the effort required at each phase, with significant effort required to integrate data, and less effort needed during process innovation. Therefore, in the first phase facilities should expect significant effort and cost, with modest gains in operational intelligence. However, the effort in each subsequent phase is reduced given that residual technologies, knowledge and skills are carried forward from the previous phase.

Impediments and challenges

There are numerous challenges facing smart manufacturing adoption. These challenges include the development of infrastructures to support real-time smart communication, as well as the cultivation of multidisciplinary workforces and next-generation IT departments that are capable of working with these technologies [11]. The extent to which these challenges exist will vary from facility-to-facility. For example, there is a distinct difference in the challenges facing greenfield and brownfield sites [7]. Apart from budgetary constraints, technology availability, and the presence of a skilled workforce, greenfield sites (i.e. new manufacturing facilities) can choose to adopt smart technologies without significant impediments. This is in contrast with brownfield sites that are encumbered by the existence of legacy devices, information systems, and protocols, some of which may be proprietary. Furthermore, many of these legacy technologies were not designed to operate across low-latency distributed real-time networks that are synonymous with smart manufacturing. Although legacy technologies could be replaced with smarter equivalents in other business domains, there are numerous reasons why substitution may not be a viable option in industrial facilities. A summary of these impediments are provided below;

- *Historical investment in IT and automation* Over the last 40 years many facilities invested in IT systems and automation networks. Therefore, facilities may be reluctant to replace technologies that received significant investment, and continue to function at an appropriate level.
- *Regulatory and quality constraints* In certain industries, such as pharmaceutical and medical devices, internal or external regulatory and/or quality standards may restrict the adoption of new technology. While in some scenarios this impediment can be

circumvented through policy alteration, the exhaustive procedures associated with such a process may negate the enthusiasm for technology replacement.

- *Dependency on proprietary systems or protocols* Although many manufacturing information systems and automation network standards exist, such as ISA95 for system interoperability and OPC for device communication, they are not always adopted. Therefore, when facilities are locked-in to proprietary and closed technologies, rather than those built on open standards, technology adoption (i.e. smart technologies) may be restricted to the offerings of their vendor.
- *Weak vision and commitment* Transitioning to smart manufacturing requires strong leadership and a shared vision of its short and long-term benefits. Facilities that do not have a clear understanding of how smart manufacturing can impact operations, may not have an appetite to replace equipment and technologies that are currently operating as intended.
- *High risk and disruption* There is a high-risk associated with new technology and system implementation. These type of projects may negatively impact operations while personnel are achieving competency or fail to operate as originally intended. Therefore, the desire to undertake large-scale IT projects for smart manufacturing adoption may remain weak until such time lost opportunities affect competitiveness.
- *Skills and technology awareness* IT, automation and facilities departments are entrenched in legacy computing, automation and networking methods that have been in existence for decades. However, smart manufacturing adoption requires a significant shift from these approaches. Therefore, unless these new technologies and methods are embraced, smart manufacturing adoption may be severely impeded.
- *Multi-disciplinary workforce* A strong multi-disciplinary workforce is an important aspect of smart manufacturing, where key decision-makers may need knowledge from multiple disciplines, such as engineering, computing, analytics, design, planning, automation, and production [15, 18]. Multi-disciplinary personnel may be particularly important for demand-driven supply chains, large-scale data analysis, system interoperability, and cyber physical systems [11].

These impediments address many of the high-level challenges that can be expected when transitioning to smart manufacturing. Inevitably, additional challenges are likely to emerge as different areas in the factory are explored (e.g. energy, production, maintenance, etc.). Recent research focused on big data technologies in manufacturing, which is closely related to smart manufacturing efforts, suggests that there are eight areas in manufacturing where data-driven methods are being explored [5]—these are (1) process and planning, (2) business and enterprise, (3) maintenance and diagnosis, (4) supply chain, (5) transport and logistics, (6) environmental, health and safety, (7) product design, and (8) quality management.

Industrial equipment maintenance

Maintaining equipment in a proper working state is an important aspect of manufacturing. However, industrial equipment maintenance is an expensive activity that can account for over 30 % of a facilities annual operating costs, and between 60 and 75 % of a machines lifecycle cost [6]. These figures will vary from site-to-site depending on the

type of equipment being maintained, and maintenance strategies being employed. Maintenance strategies range from those that focus on reacting to issues when they arise, to those that focus on preventing issues from occurring. Strategies that embrace a predictive and preventative approach to maintenance are well suited to smart manufacturing, given their affinity to optimising machine uptime and availability. Most maintenance strategies are supported by information systems that monitor particular measurements (e.g. temperature, revolutions per minute, etc.) from equipment in the factory. However, a criticism of existing real-time maintenance systems is their inability to describe, predict or prescribe specific maintenance actions [19].

Maintenance strategies

There are numerous strategies that can be used for industrial equipment maintenance. Each strategy possesses its own strengths and weaknesses, which will suit different scenarios depending on the type of equipment being maintained, and its role in the facility and/or manufacturing process. Table 1 provides a comparison matrix of common maintenance strategies that describe the trade-offs between each, as well as providing guidelines relating to their use.

Each maintenance strategy has an obvious maintenance goal. For example, in the case of predictive maintenance, the goal is simply to identify a potential issue before it occurs while optimising resources and reducing costs. These maintenance goals are achieved using issue identification techniques. Issue identification techniques are typically implemented in standalone or integrated information systems in the factory, and may embody one or more maintenance strategies.

Issue identification techniques

Similar to maintenance strategies, there are several common issue identification techniques. Apart from reactive maintenance strategies, where issue identification is the result of an observed equipment failure, maintenance strategies are realised using information systems that implement particular issue identification techniques. Table 2 provides a comparison matrix of that describe the most common issue identification techniques, and the maintenance strategies they can support.

While there are many issue identification techniques to choose from, smart manufacturing places a strong emphasis on developing predictive capabilities throughout the factory. As previously discussed, being able to accurately predict issues can deliver machine uptime and availability for demand-driven supply chains, while also optimising energy consumption and equipment lifecycle costs [7]. Therefore, PM and PHM issue identification techniques appear to be the most relevant to smart manufacturing. These techniques can be found in contemporary research, where emerging technologies and methods have been applied to equipment maintenance [8, 10, 14, 20]. The Center for Intelligent Maintenance Systems (IMS) is a noteworthy contributor to research in this space. IMS is a leading research initiative comprised of university and industry partners, which employs industrial big data analytics, real-time smart systems and continuous health monitoring, to create intelligent, self-aware and self-learning systems that can achieve near-zero breakdown [21].

Table 1 Common industrial maintenance strategies

Strategy	Intent	Benefits	Weaknesses	Suitability
Reactive	Only undertake maintenance when a complete equipment failure occurs	No upfront planning or scheduling	Unpredictable equipment availability, shorter equipment lifetime, increased energy costs, and potentially lower production yield due to partial malfunctions	Suitable for non-essential equipment, or in situations where the cost of maintaining equipment is greater than the cost of failure
Corrective	Identify and address individual/minor faults when they occur to avoid a complete equipment failure	Manages risk of complete failure, provides visibility of equipment health, and can increase lifetime of equipment through timely maintenance activities	Investment in diagnostic technologies, as well as the labour cost associated with monitoring and managing faults	Suitable for broad classes of equipment maintenance, but may not be suitable for mission critical equipment, where a complete failure must be mitigated at all costs
Preventative	Perform regular maintenance to avoid either partial or complete equipment failures. Preventative maintenance can be undertaken at time intervals (e.g. change component X every 4 weeks regardless of its state), or when a particular condition has been met (e.g. heating element begins to take X minutes to reach its target temperature)	Promotes confidence in machine availability by mitigating equipment failure using pre-determined maintenance intervals/conditions	Prematurely replacing components and carrying out maintenance activities may come at the expense of high maintenance costs, or at least costs that are sub-optimal	Suitable for scenarios where every attempt must be made to ensure that mission critical equipment is available and operating correctly at all times, but this is typically done at the expense of resource efficiency and cost
Predictive	Predict an issue before it occurs and be capable of estimating the remaining useful life (RUL) of the equipment and/or its internal components	Optimises resources and reduces costs by predicting the lifetime of components to avoid premature replacement and circumvent redundant maintenance activities	Given prediction is probabilistic rather than deterministic, there is potential for false positives that could lead to unnecessary maintenance actions	Suitable for scenarios where the operation, cost and output derived from equipment must be optimised, but occurrences of machine availability can be tolerated

Table 2 Common issue identification techniques

Technique	Strategies	Description	Implementation	Weakness
Basic intelligence and reporting	Corrective	Reporting is used to manually assess if a particular parameter is outside an expected operating boundary. If so, further investigation of the potential issue can be undertaken	Existing industrial information systems, such as Manufacturing Execution Systems (MES) or Building Management Systems (BMS), which are used in day-to-day operations, can be used to generate reports	Largely manual process, with static and ad hoc issue identification. Also dependent on the ability of the expert analysing the report to observe the anomaly, which may be somewhat subjective and easy to overlook due to human error
Fault detection and diagnosis (FDD)	Corrective	FDD consists of a set of encoded fault logic (e.g. IF/THEN rules), to identify potential issues based on a set of input data	FDD capabilities are embedded in some industrial information systems, but also exist as standalone systems and tools that can be used to monitor specific types of equipment	Logic employed is typically specific to equipment, and detection means that the issue is already present and may be impacting operations in some way
Condition-based monitoring (CM)	Corrective Preventative	CM focuses on monitoring a particular measurement, or set of measurements, to determine if an issue has, or is likely to occur. The condition is fired when the monitored parameter(s) are outside a predefined range	CM is available in many modern industrial information systems, and can be viewed as an extension to reporting and monitoring modules, with a condition/trigger that automatically highlights issues	The condition is specific to equipment and/or its components. Therefore, performance and accuracy is dependent on the appropriate parameter(s) being chosen, and condition values set
Predictive maintenance (PM)	Preventative Predictive	PM employs statistical learning techniques to anticipate the occurrence of an issue, and/or estimate the RUL of equipment and components	Predictive methods in current industrial information systems are limited, and therefore, PM it is common to see implementations as standalone systems or tools	To develop an accurate tool, an appropriate amount of high-quality data must be available to inform the statistical learning model
Prognostics and health management (PHM)	Corrective Preventative Predictive	PHM uses a holistic approach to issue identification, and comprises FDD, CM, and PM, to highlight issues at different stages so that optimal equipment health is maintained	Given multiple techniques are used in PHM; it is typically implemented as a dedicated system. In some cases, where interoperability exists, PHM systems may leverage the FDD, CM or PM capability of an existing system	Implementing multiple techniques represents challenges – e.g. when should a particular technique be used. This arguably makes PHM more complex than any single technique

Industrial information and data systems

Figure 1 illustrates a topology of an industrial network with common components, such as devices, systems and databases, which facilitate data flow in the factory. This topology depicts a special type of network that is used in manufacturing to orchestrate the production process, which is known as an automation network. As the adoption of standards is a major issue for these types of industrial networks, the topology presented is abstracted to a level that enables the hierarchical data flow to be depicted without being burdened by low-level details. The transmission of data begins with instruments at the bottom level recording measurements (e.g. temperature), and culminates in end-users accessing information at the top level.

Instruments (i.e. sensors) stream continuous measurements (e.g. room temperature) [22–24] to programmable logic controllers (PLC). PLC's are digital computers that are programmed with logic to automate the production process. This logic is programmed by automation engineers to evaluate each instruments measurement and initiate appropriate actions based on their state. Measurements transmitted to PLC's are persisted in memory at set intervals (e.g. every 15 min) as tuples of timestamp/value. This format is common in industrial information systems and is referred to as time-series, measured, or temporal data. Subsequently, time-series data persisted in PLC memory is transferred to an archive in batch (e.g. every 24 h), with the archive typically taking the form of a relational database or flat log file. Once data is persisted in the archive, information systems are able to consume data from the repository and generate reports for end-users. Examples of such systems include building management systems (BMS), manufacturing execution systems (MES), and monitoring and targeting systems (M&T).

Accessing data from archives can be achieved using standard database and I/O interfaces. However, when underlying data models are proprietary, some cleaning and transformation may be required to unify measurements. While accessing data from archives

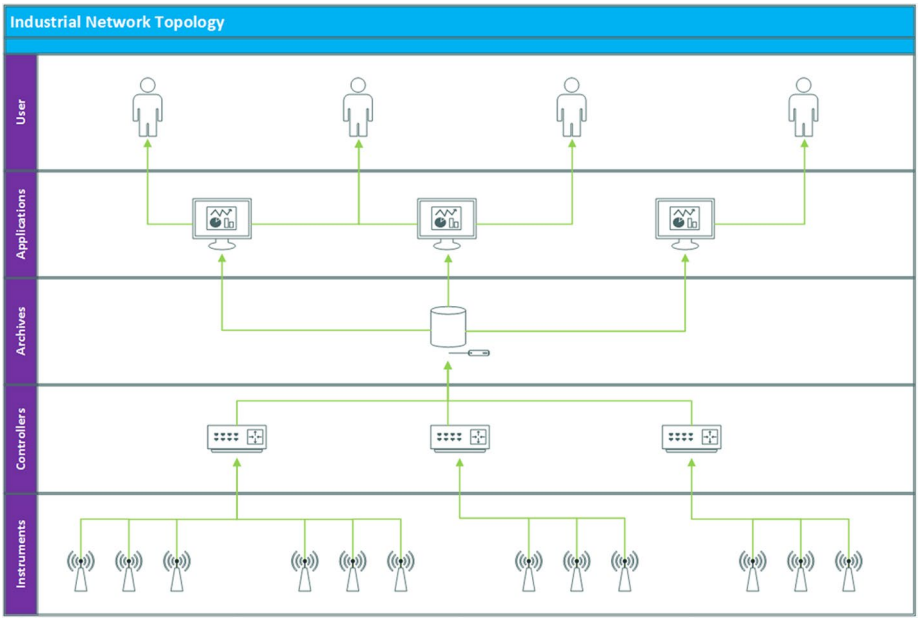

Fig. 1 Topology of an industrial environment

can be used in some industrial applications, the high latency characteristics of archiving across automation networks (e.g. 24 h dumps) means that it is not inherently suited to real-time industrial applications. Where real-time data access is a requirement, direct communication with field devices (e.g. PLC) on the lower levels of an automation network must be undertaken. This can be achieved using industrial protocols and interfaces, such as Modbus, LonWorks, BACnet, OLE Process Control (OPC), and MT Connect [25–28], or where smart sensors (e.g. IoT devices) exist, using MQTT, COAP and HTTP [4]. However, as previously stated, the adoption of smart sensors and emerging technologies in large-scale manufacturing brownfield sites may be restricted by regulation and quality control, as well as the high risks associated with new technology adoption [7]. Therefore, initial data management and infrastructure requirements for smart manufacturing may need to consider how legacy and emerging technologies can operate transparently. Table 3 summarises potential sources of equipment maintenance data in industrial environments.

Unifying data in industrial environments is a significant challenge due to the disparity of data models in legacy information systems, and diverse number of industrial protocols. Many existing brownfield sites transitioning to smart manufacturing must also address device interoperability, real-time capabilities, big data infrastructures, and open standards.

This paper presents a big data pipeline for industrial analytics applications focused on equipment maintenance. The design of the data pipeline is based on real-world requirements obtained from an embedded study in a large-scale manufacturing facility, as well as adhering to relevant smart manufacturing recommendations. The proposed big data pipeline employs an open and extendable system architecture that is capable of indiscriminately interacting with legacy and smart devices on automation networks, as well as exposing a common data interface for industrial analytics applications to consume

Table 3 Potential sources of equipment data

Classification	Data source	Data type	Common interfaces	End user	Latency
Database	Building management system (BMS)	Energy and environmental	ODBC OLEDB JDBC etc.	Energy and facilities	Batch
Database	Monitoring and targeting (M&T)	Energy	ODBC OLEDB JDBC etc.	Energy and facilities	Batch
Database	Manufacturing execution system (MES)	Production and automation	ODBC OLEDB JDBC etc.	Automation, production and quality	Real-time and batch
Device	Programmable logic controller (PLC)	Production, energy and environmental	OPC MT Connect BACnet Modbus LonWorks	Automation and building services	Real-time
Device	Gateways	Multiple	HTTP OPC Modbus I/O (i.e. CSV)	Multiple	Real-time and batch
Device	Smart devices (i.e. IoT)	Multiple	MQTT COAP HTTP	Multiple	Real-time

measured data. This research should be of interest to big data researchers and practitioners that wish to extend their work to industrial environments, engineering informatics researchers operating at the crossroads of manufacturing and technology, and large-scale manufacturing facilities moving towards smart manufacturing adoption.

Research methodology

This research employs an embedded study. The study was undertaken in DePuy Ireland, which is a large-scale manufacturing facility that is part of the Johnson & Johnson group. An empirical approach was chosen to observe proprietary and ad hoc technologies, architectures and configurations in a real-world industrial environment, while also utilising industry collaboration to identify relevant requirements for industrial big data analytics. Although a theoretical solution could be derived without empirical methods, it is unlikely to address all of the peculiarities associated with industrial environments. Therefore, this study was necessary to inform the future implementation of an industrial big data pipeline for smart manufacturing.

Research scope

The first priority was to establish boundaries that could limit efforts to a subset of data flows and operations in the factory. While many smart manufacturing scenarios employ data from across the factory, it was necessary to define a scope for this research given available resources. Therefore, after discussions between members of the research team and automation personnel in DePuy Ireland, the following boundaries were agreed.

Type of applications

The industrial big data pipeline focuses on supporting data-driven analytics applications for predictive and intelligent equipment maintenance. These types of applications were chosen given their alignment with the goals of smart manufacturing, such as enabling predictive capabilities, promoting machine availability, and optimising energy consumption. The main facets of these applications were identified as industrial data integration, real-time processing, and data-driven analysis.

Regulation and compliance

Given the highly regulated and quality-focused environment of the study, the research considers smart manufacturing in the context of such facilities. This limitation implied that smart manufacturing cannot be achieved using technology replacement alone. Therefore, an additional emphasis was placed on supporting legacy technologies in the data pipeline. However, it was agreed this should not be done at the expense of supporting emerging smart technologies.

Time-series data

As the industrial big data pipeline focuses on equipment maintenance applications, data flows through the pipeline will be limited to time-series data. Based on the experiences of research team members and automation personnel in DePuy Ireland, no other class of data was highlighted as being relevant to equipment maintenance and monitoring.

Limiting the data pipeline to a single class of data meant the complexity of operations (e.g. cleaning and transformation) could be reduced given its predefined structure of timestamp/value tuples.

Immutable data

A data pipeline comprises sequential processes for data management and preparation. In the context of this research, the industrial big data pipeline was concerned with funnelling data from the factory to data-driven equipment maintenance applications. Therefore, it was only necessary for data in the pipeline to flow outwards to serve industrial analytics applications (i.e. one-way). This meant the pipeline would not be a closed loop system given the lack of an inbound communication channel to the factory. However, this would not prevent third party applications from implementing their own methods to close the loop. Furthermore, time-series data was considered immutable given it is a record of a measured state at a particular point in time. This meant the data pipeline would only need to service read requests from consuming applications.

Requirements analysis

The area of smart manufacturing is characterised by high-level heuristics, guidelines and roadmaps that are too abstract to inform system design or implementation. Therefore, this research combines requirements that were observed and elicited from a real-world large-scale manufacturing facility, with the ethos of smart manufacturing, to derive a set of functional requirements for the industrial big data pipeline. This meant requirements would be limited to characteristics of the study environment and assertions of industry collaborators.

Industry partner

This research was undertaken with an industry partner, DePuy Ireland. The aim of this partnership was to familiarise members of the research team with advanced manufacturing operations in highly regulated and quality-controlled environments. The research team were also given access to group meetings across automation, energy, big data and manufacturing. These meetings were beneficial for eliciting high-level requirements, determining attitudes towards smart manufacturing, and highlighting information systems and data sources pertinent to equipment maintenance. The following section broadly summarises the contributions from each team;

- *Automation*—the automation team comprised personnel with knowledge of control, production, energy and information technology. Liaising with the automation team provided a better understanding of the technologies, infrastructure and procedures relating to production, scheduling and maintenance.
- *Energy*—the energy team comprised personnel with specialist knowledge of engineering, energy and environment. The energy team demonstrated current systems for measuring the energy consumption of machinery, and illustrated how malfunctioning equipment can negatively impact operating cost.
- *Big data and smart manufacturing*—no single team was responsible for big data and smart manufacturing, which is understandable given the contemporary and inter-

disciplinary of both fields. However, several teams were considering applications of big data in manufacturing, while other teams were investigating strategies for smart manufacturing. These teams showed a positive attitude towards smart manufacturing approaches, with plans to integrate IoT, big data systems and data-driven analytics being a recurring theme.

Research questions

RQ1: How can industrial big data analytics applications be enabled and integrated in existing large-scale manufacturing facilities?

The purpose of this question was to establish the real-world requirements for the industrial big data pipeline, understand how data is currently being transmitted across the factory, and identify overlaps or similarities with smart manufacturing.

RQ2: What type of data architecture is needed to support the implementation of industrial big data analytics focused on equipment maintenance?

The purpose of this question was to prescribe industrial big data pipeline architecture based on results from RQ1, while using smart manufacturing philosophies, such as open standards, real-time monitoring, and advanced analytics, to inform architectural decisions.

Results

RQ1: Requirements and challenges

The following requirements and challenges were identified in response to RQ1. These requirements were produced through elicitation with industry personnel, observations relating to industrial information systems in the facility, and adherence to prominent smart manufacturing guidelines and recommendations. Synthesising and collating requirements proved to be a significant undertaking. Therefore, to communicate results effectively in this research, requirements were limited to those deemed essential to the design and implementation of an industrial big data pipeline.

Legacy integration

Integrating legacy and smart technologies is needed for smart manufacturing adoption given practicalities regarding technology replacement. Facilities that invested heavily in multiple iterations of sophisticated automation networks may find it difficult to justify the cost of technology replacement in the immediate-term. Those facilities that can justify the expense must overcome regulatory, quality, infrastructure and technical challenges that are needed to support and integrate smart technologies in the factory. Therefore, integration with legacy automation networks is desirable given historical investments regarding infrastructure, skills and knowledge can be leveraged.

Cross-network communication

Pervasive real-time networks are a key aspect of smart manufacturing. However, current network topologies in industry can be complicated by the existence of multiple Virtual Private Networks (VPN) and Local Area Networks (LAN). These configurations exist to secure operations across departments and processes, but they can also restrict data

access for consuming applications (e.g. equipment maintenance analytics software). Data accessibility can be further exacerbated where resources are under the control of external vendors. While these measures make sense from a security or management perspective, they represent a challenge to the adoption of smart manufacturing. Therefore, legacy technology configurations must be overcome to facilitate communication across networks without being encumbered by rigid governance policies and procedures. However, this requirement must be considered on a site-by-site basis to ensure security protocols are not violated.

Fault tolerance

Information systems and technologies that contribute to production, automation and maintenance activities must exhibit high availability. Given these systems can affect operating metrics, such as production yield and energy consumption, it is imperative they are resilient to failure and provide an appropriate level of redundancy. Therefore, it is logical that smart manufacturing initiatives, such as the industrial big data pipeline for equipment maintenance proposed in this research, must exhibit fault tolerance and operational resilience.

Extensibility

The presence of proprietary technologies and systems in large-scale manufacturing facilities is common. In recent years facilities have become more aware of technology integration and consolidation, but after discussions with industry personnel it appears that duplication and disparity across systems persists. Much of this is due to the plethora of data types and protocols that exist in industrial environments. Therefore, systems dealing with data integration must be extensible to allow new data formats and protocols to be added when they arise.

Scalability

While the digitisation of the factory accelerates, the number of sensors that will be operating in smart manufacturing facilities is yet unknown. Therefore, dynamic scalability is a highly desirable attribute for systems focused on smart manufacturing. Although current manufacturing technology may be considered advanced when compared with other sectors, the highly disruptive and intensive nature of smart manufacturing methodologies may place unpredictable and unforeseen demands on existing information systems. Therefore, technologies and systems focused on smart manufacturing must exhibit the ability to scale seamlessly based on demand.

Openness and accessibility

Large-scale manufacturing facilities produce large quantities of data. However, problems persist regarding its proprietary and inaccessible nature. Such characteristics make data integration complicated and time-consuming, which can result in data-driven projects being overlooked due to the associated costs. Those that are commissioned typically exhibit high levels of duplicate and redundant data integration, with poor reuse due to hardcoded routines and proprietary requirements. Therefore, systems and tools should be built using open standards to promote data accessibility. Furthermore, data access

should be realised through a common interface to enable reuse and interoperability between systems and applications in the factory.

RQ2: Data pipeline architecture

Given the requirements identified by RQ1, a big data pipeline architecture for industrial analytics applications focused on equipment maintenance was created. Figure 2 presents the big data pipeline architecture with each stage of the workflow numbered and highlighted. The proposed data pipeline provides a turnkey solution for industrial data integration, processing and analysis. It uses a distributed architecture that is highly scalable and fault tolerant, which is suitable for operation in a smart manufacturing facility. The purpose and function of each stage in the data pipeline is described in the following sections.

Stage 1: Site manager

Purpose The site manager resides on a cloud server and acts as a central repository of metadata relating to facilities and equipment being monitored. Its purpose within the architecture is to persist and communicate essential site information to other components in the architecture.

Functions The site manager has multiple functions that are directly related to the factory—(1) store details relating to the site, such as the type and location of local data sources to be integrated, (2) schedule and assign jobs to ingestion engines based on their availability and location, and (3) decide how much data each node should ingest based on current CPU and bandwidth availability.

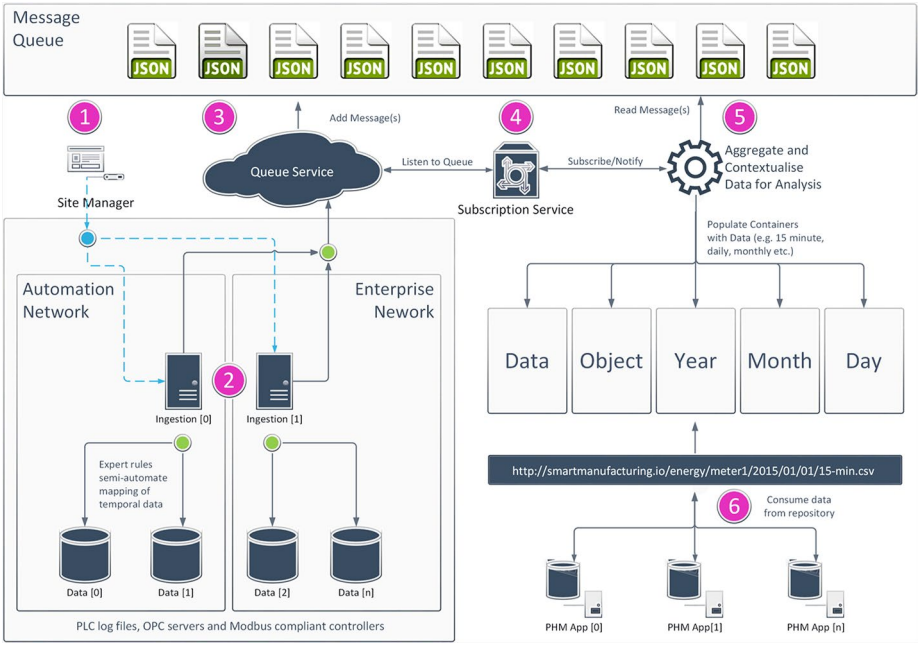

Fig. 2 Big data pipeline architecture and workflow

Stage 2: Ingestion process

Purpose The ingestion engines are distributed software agents deployed autonomously across multiple automation networks in the factory. They ingest data from time-series data sources relevant to industrial equipment maintenance applications (e.g. HVAC, Chillers, Boilers). Ingestion engines run as background applications on local servers and communicate their status to the site manager (Stage 1), and transmit time-series data to the cloud when instructed to do so. Figure 2 illustrates the ingestion engines distributed and autonomous nature, which makes it easy to deploy them across automation networks that are separated by firewalls and/or geographical boundaries. These characteristics can also be used to increase ingestion work capacity or improve latency by adding more engines.

Functions The ingestion engine has multiple functions—(1) communicate location, bandwidth, CPU and memory availability to the site manager, (2) interpret data collection tasks sent from the site manager and automatically extract time-series data in accordance with task parameters (e.g. particular date range), and (3) transmit the acquired time-series data to the cloud. Furthermore, the extraction of data from sources in the factory can be automated using the expert ruleset embedded in the ingestion engine to automatically identify the appropriate data mapping for each source.

Stage 3: Message queue

Purpose The message queue is a highly available and distributed service that stores JSON encoded time-series data transmitted from the factory. It acts as an intermediary data store between the factory and data processing components in the pipeline. The message queue decouples the ingestion process from data processing components to facilitate asynchronous operations and promote scalability, robustness and performance.

Functions The message queue has two functions—(1) notify the subscription service when new data has been received from the factory, and (2) persist the received data in a queue so it may be read by data processing components in the data pipeline.

Stage 4: Subscription service

Purpose The subscription service provides a notification service between the endpoint for data ingestion (i.e. message queue) and data processing components responsible for transforming raw data to a state suitable for industrial analytics applications. It decouples the message queue from data processing components, which enables both to scale and operate independently.

Functions The subscription service functions are important to the chain of events in the data pipeline—(1) listen to the message queue for new data, and (2) notify subscribers when new data is available for processing.

Stage 5: Data processing

Purpose The processing components are responsible for transforming time-series data to a form that is useful for analysis. The data processing components in the pipeline aim to remove the onus on ad hoc processing and aggregation routines on raw data. The basic processing illustrated for time-series data is the transformation of high residual data to different levels of granularity, such as hourly, daily, monthly and annual averages. More

sophisticated data processing may include the execution of expert rules to identify early fault signals, or encoding of time-series data in a semantic format (e.g. Project Haystack) to support interoperability with a particular application. Each processing component in the architecture is responsible for a single use case, such as those previously mentioned. Therefore, new requirements that cannot be met by existing components can be facilitated through the creation and deployment of a new component.

Functions The potential requirements relating to data processing are diverse and will vary from application-to-application and site-to-site. Therefore, the data processing aspect of the pipeline has been designed with customisation and extensibility in mind. It is envisaged that a library of default data processing components will eventually be included in the final architecture. Currently, the data pipeline architecture incorporates simple aggregation functions for time-series data—(1) daily average, (2) monthly average, and (3) annual average.

Stage 6: Data access

Purpose The purpose of the data access stage is to provide a consistent and open interface for industrial analytics applications focused on equipment maintenance to consume data. The data access interface serves files output from data processing components using a cloud repository. Data queries utilise a URL naming convention to describe the time-series data being requested (e.g. equipment identifier and date range), adhering to this naming convention can promote consistency and standardisation for industrial analytics applications in the facility. The naming convention is illustrated in Fig. 2 and described in further detail below;

- *Data*—refers to a particular data set that was processed (e.g. energy).
- *Object*—an identifier within the data set (e.g. equipment identifier).
- *Year*—the year that is relevant to the query.
- *Month*—the month that is relevant to the query.
- *Day*—the day that is relevant to the query.

Functions The functions of the data access component are—(1) ensure data is stored in the appropriate location/context, and (2) respond to requests for data that adhere to the aforementioned convention.

Discussion

The qualitative results in this research include a set of empirically derived requirements (RQ1) elicited through industry collaboration, and a big data pipeline architecture (RQ2) for industrial analytics applications focused on equipment maintenance. Given the theoretical aspects of this research, the following section simulates a real-world scenario to provide additional context and derive points for discussion. The simulation decomposes the data pipeline into three parts—(1) data ingestion in the factory, (2) data processing in the cloud, and (3) feeding industrial analytics applications. Each part is discussed in terms of its ability to satisfy the aforementioned requirements, as well as highlighting potential implementation challenges.

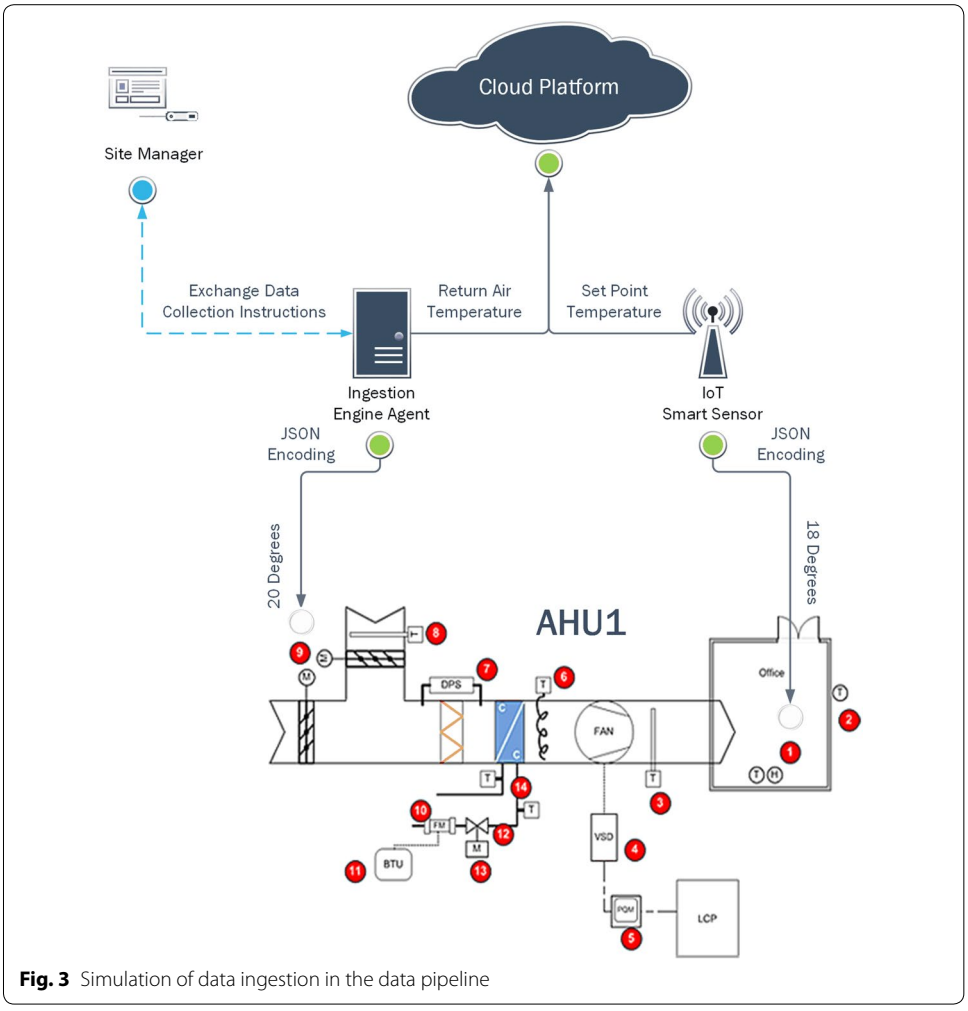

Fig. 3 Simulation of data ingestion in the data pipeline

Part 1 of simulation: data ingestion

Figure 3 depicts the ingestion of measurements from an Air Handling Unit (AHU) using the industrial big data pipeline. AHU's are mainly used in manufacturing to control air quality and ensure thermal comfort in the facility. There are two entities in the simulation that are used to transmit data from the factory to the cloud, namely the ingestion engine and smart sensor. While the ingestion engine is an internal component that is directly controlled by the data pipeline, the smart sensor can be considered a third party component that is programmed and managed externally. Given its tighter integration with the pipeline, the ingestion engine receives data collection instructions from the site manager, which returns an instruction to read the Return Air Temperature (RAT) for AHU1. This measurement is obtained by communicating with the PLC associated with the RAT instrument using an industrial communication protocol (e.g. BACnet). Similarly, the smart sensor is programmed to read the Set Point Temperature (SPT) for AHU1. However, its operation is programmed by a third party and due to on-board instrumentation there is no need for industrial protocols. Both measurements are read at regular intervals using their respective methods, encoded in a common JSON format and pushed to the cloud.

Table 4 Data ingestion discussion

Requirement	Discussion
Legacy integration	The simulation illustrates how legacy and smart devices can coexist in the industrial big data pipeline, with legacy integration realised using the ingestion engine to abstract and encapsulate legacy devices and instrumentation relating to the RAT measurement. Both measurements are pushed to the cloud for industrial analytics applications to consume without being aware of their origin (i.e. legacy or smart device)
Cross-network communication	Ingestion engines can operate across different networks due to lack of local dependencies and cloud data processing. The simulation shows that an ingestion engine can be deployed on a network with an active internet connection, access to the site manager, and reachable data sources. Considering the flow of data depicted in the simulation, the network location of the ingestion engine and smart sensor are irrelevant as both measurements reach the same endpoint (i.e. cloud platform)
Fault tolerance	Reliability and resilience can be instilled by adding more ingestion engines to the process and monitoring the status of each. As illustrated by the simulation, the site manager controls the ingestion process by sending data collection instructions to the ingestion engine. In scenarios where an ingestion engine fails to complete its assigned task, the instructions can be reassigned to another ingestion engine. This promotes fault tolerance in the ingestion process by removing a single point of failure, while allowing for different levels of resilience
Extensibility	It is feasible that the data pipeline may need to be extended to support additional data sources. As ingestion engines encapsulate data integration logic they are the logical point of extension. Given the existence of an abstract function that expects data collection instructions from the site manager as input, and outputs JSON encoded measurements, additional data sources could be facilitated by implementing new instances of the function
Scalability	The ingestion process can be scaled horizontally by deploying additional ingestion engines across machines and networks. This provides the data pipeline with a greater work capacity given that more ingestion jobs can run in parallel. The ingestion engine characteristics that make this possible have already been addressed in discussions regarding cross-network communication and fault tolerance
Openness and accessibility	The simulation shows how open standards, such as HTTP and JSON, can be used to facilitate communication and data exchange amongst distributed components in the data pipeline. This is exemplified by communication between the site manager and ingestion engine to relay data collection instructions, as well as the transmission of measurements to the cloud from the ingestion engine and smart sensor

Table 4 discusses the simulation of data ingestion in the data pipeline in the context of the requirements identified in RQ1 of this study.

Part 2 of simulation: data processing

Figure 4 illustrates data processing of RAT and SPT measurements transmitted from the factory. Both measurements are added to the message queue for processing. The subscription service is notified when the new measurements arrive in the queue. In turn, the data processing component for aggregation and contextualisation is notified of the new measurements by the subscription service. The data processing component reads both messages from the message queue and executes its routine, which results in aggregated time-series data (i.e. 15 min, hourly, daily and monthly intervals) being pushed to a contextualised repository. These repositories are encoded to convey the dataset, object, year, month and day the data relates to. By the end of this part of

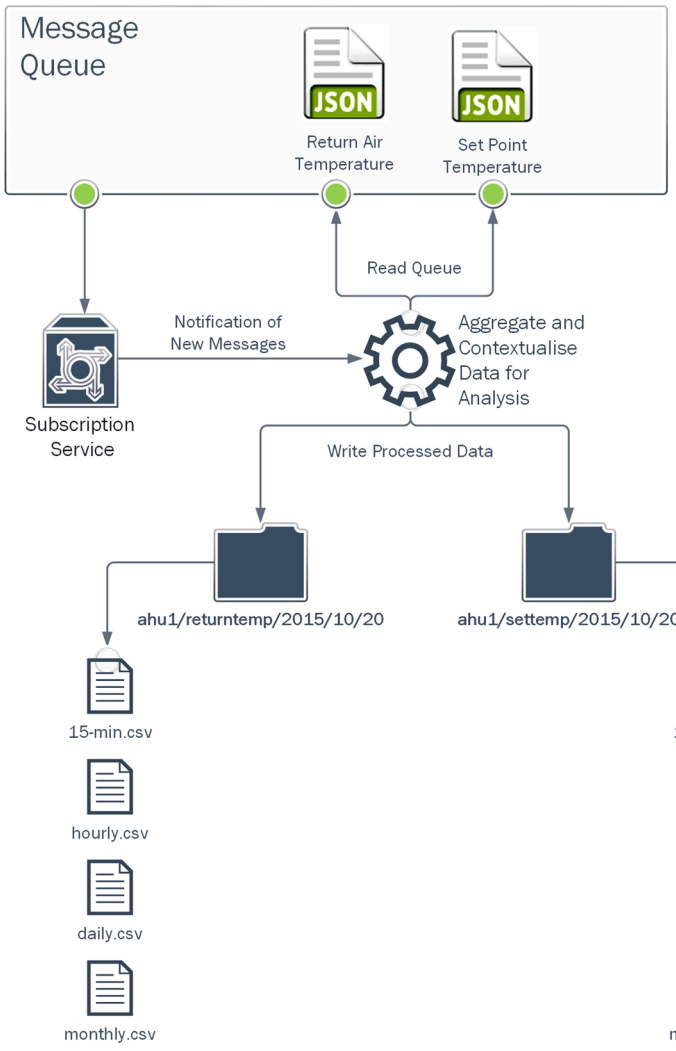

Fig. 4 Simulation of data processing in the data pipeline

the simulation data is located in a directory that is accessible to industrial analytics applications.

Table 5 discusses the simulation of data processing in the data pipeline in the context of the requirements identified in RQ1 of this study.

Part 3 of simulation: industrial analytics

Figure 5 illustrates how industrial analytics applications consume data from the pipeline. The processed time-series data resides in a directory structure that gives context to the data being accessed (i.e. returntemp and settemp for AHU1). There are two applications in the simulation. The dashboard is a business intelligence application that implements basic descriptive analytics. It uses the data pipeline to access precompiled aggregates of time-series data to eliminate the overhead of running this routine dynamically. The simulation shows that the dashboard accesses the hourly, daily and monthly data for both measurements. The other data consumer is a predictive maintenance model that is used

Table 5 Data processing discussion

Requirement	Discussion
Legacy integration	The simulation shows JSON encoded RAT and SPT measurements in the message queue. It is apparent that legacy integration was successful given the RAT measurement originated from a legacy source, but now resides in the same format as SPT
Cross-network communication	The unified message queue shown in the simulation illustrates how messages from across devices and networks are assimilated after a successful ingestion process
Fault tolerance	Cloud computing provides a fault tolerant environment for data processing due to its ability to scale resources based on demand. The data pipeline simulation depicts a high-throughput cloud infrastructure with data processing and message queue components. It is implied that these components reside in a highly distributed cloud service that provides fault tolerance across multiple compute nodes. Furthermore, the simulation clearly shows how the message queue decouples the data ingestion from data processing. This promotes additional fault tolerance by protecting the ingestion process from faults that originate from data processing components
Extensibility	The simulation illustrates the aggregation and contextualisation component that is responsible for preparing data for analysis. This is one example of processing that may be required for time-series data. As new processing needs arise additional components can subscribe to the message queue subscription service and begin processing in parallel with other components. This extension is facilitated by the decoupling of processing components and message queue using the subscription service
Scalability	The data processing simulation inherits its scalability from its cloud infrastructure. Many of the benefits of cloud computing have already been discussed with regard to fault tolerance. These same load balancing features facilitate scalable data processing in the pipeline by scaling compute resources based on demand (i.e. amount of processing required)
Openness and accessibility	The simulation shows data stored in a contextualised cloud repository after completion of the aggregation process. This repository uses a naming convention to identify a dataset (i.e. AHU1), object (i.e. RAT), and chronological association for accessing time-series data. To promote openness and accessibility this data is accessed using standard HTTP requests. Furthermore, the ability of the data pipeline to support additional data formats, standards and representations has been addressed in discussions regarding extensibility

to identify issues in AHU's. The predictive model requests 15-min data for both measurements given its need for granular data. Both applications in the simulation were able to access data using a common interface without having to engage low-level industrial protocols.

Table 6 discusses the simulation of industrial analytics in the data pipeline in the context of the requirements identified in RQ1 of this study.

Conclusions and future work

In this paper, we addressed the main challenges and desirable characteristics associated with large-scale data integration and processing in industry, such as automating and simplifying data ingestion, embedding fault tolerant behaviour in systems, promoting scalability to manage large quantities of data, supporting the extension and adaption of systems based on emerging requirements, and harmonising data access for industrial analytics applications. The contributions and findings of this research are important for facilitating big data analytics research in large-scale industrial environments, where the requirements and demands of data management are significantly different to traditional information systems. While emerging technologies (e.g. IoT) may eventually eliminate

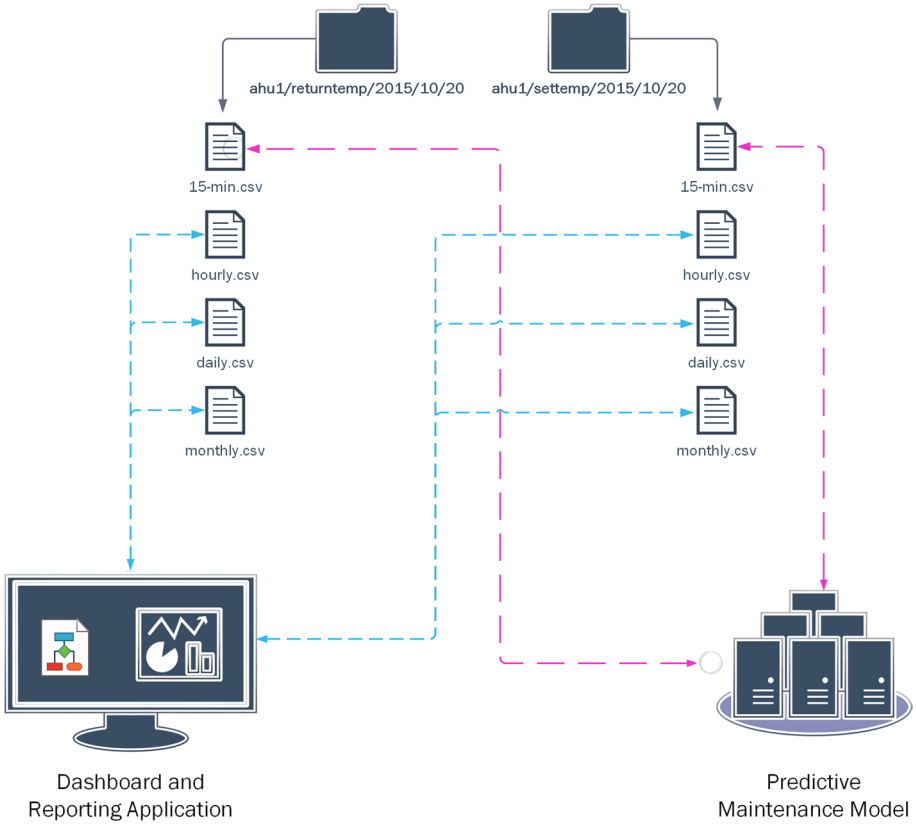

Fig. 5 Simulation of industrial analytics in the data pipeline

the need for legacy support in the factory, given the fact 20 year old devices are still in operation, we feel the big data research community should be cognisant of a potential lag in smart technology adoption across large-scale manufacturing facilities. Therefore, the big data pipeline presented in this research facilitates transparent data integration that enables facilities to begin their smart manufacturing journey without committing to extensive technology replacement.

Future work will focus on the implementation and deployment of the big data pipeline in DePuy Ireland. Our aim is to validate the big data pipeline architecture, reassess and extend the requirements presented, quantify the percentage of data sources that can be accessed in the factory using the ingestion process, and estimate the throughput capacity of the pipeline using load testing. Finally, there are two research projects in DePuy Ireland where we plan to use the data pipeline to feed predictive maintenance applications for Wind Turbine and Air Handling Units in the facility.

Authors' contributions

POD, KL, KB and DOS all contributed to the review of the literature, eliciting system requirements in DePuy Ireland, and refining and prioritising those requirements in the context of large-scale manufacturing facilities. POD designed and modelled the main technical architecture to support the requirements, and was responsible for the industrial big data pipeline concept as a means of managing and enabling 'big data' in industrial environments. KL created the naming convention and methodology for accessing measured data from the data pipeline, and aligned the convention with the needs of PHM for wind turbines. KB identified the relevant data repositories for maintenance and production, and created a mapping protocol that the data pipeline uses for data ingestion. DOS identified and aligned the cloud services for each component in the data pipeline architecture, while also fulfilling the role of principal investigator, which involved supporting and guiding contributions made by all authors. All authors read and approved the final manuscript.

Table 6 Industrial analytics discussion

Requirement	Discussion
Legacy integration	The previous parts of the simulation ensured that legacy data integration was achieved. Data access illustrated in this simulation shows both applications are able to access measurements regardless of whether they originated from legacy or smart devices
Cross-network communication	Similarly, previous parts of the simulation have realised cross-network communication. Data ingestion undertaken across all networks is unified by message queues before being exposed to applications for reading
Fault tolerance	The inherent qualities of cloud computing are used to provide fault tolerance in the data access part of the simulation. For example, file storage and delivery services in the cloud (e.g. Amazon S3) can provide a distributed, low latency and fault tolerant platform for serving time-series data
Extensibility	Extensible data access in the pipeline is necessary to service industrial analytics applications. For example, an application may require a certain data format or standard to be presented. In this scenario the new format can be generated by a processing component (part 2 of simulation) and pushed to the cloud repository for industrial applications to access
Scalability	Similarly to fault tolerance, the scalability of the data access part of the simulation is dependent on the cloud service on which it resides. The ability of cloud-based file delivery services to scale horizontally across multiple compute nodes and data centres provides a highly scalable infrastructure for serving time-series data
Openness and accessibility	The simulation illustrates how data access can be achieved from the data pipeline with context encoded URL's over HTTP. Furthermore, no proprietary or commercial technologies or drivers are required to consume the time-series data from the cloud. Therefore, there are no obvious technology barriers preventing users, applications and systems from accessing the data

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

Funding
The authors would like to thank the Irish Research Council, DePuy Ireland and Amazon Web Services for their funding of this research, which is being undertaken as part of the Enterprise Partnership Scheme (EPSPG/2013/578).

Received: 7 September 2015 Accepted: 3 November 2015
Published online: 16 November 2015

References

1. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH. Big data: the next frontier for innovation, competition, and productivity; 2011.
2. Siddiqui S. Big data process analytics : a survey. Int J Emerg Res Manag Technol. 2014;3(7):117–23.
3. Dobre C, Xhafa F. Intelligent services for big data science. Futur Gener Comput Syst. 2014;37:267–81.
4. Gubbi J, Buyya R, Marusic S, Palaniswami M. Internet of things (IoT): a vision, architectural elements, and future directions. Futur Gener Comput Syst. 2013;29(7):1645–60.
5. O'Donovan P, Leahy K, Bruton K, O'Sullivan DTJ. Big data in manufacturing: a systematic mapping study. J Big Data. 2015;2(1):20.
6. Dhilon BS. Maintainability, maintenance, and reliability for engineers. CRC Press; 2006.
7. Davis J, Edgar T, Porter J, Bernaden J, Sarli M. Smart manufacturing, manufacturing intelligence and demand-dynamic performance. Comput Chem Eng. 2012;47:145–56.
8. Lee J, Lapira E, Bagheri B, Kao H. Recent advances and trends in predictive manufacturing systems in big data environment. Manuf Lett. 2013;1(1):38–41.
9. Lee J. Recent advances and transformation direction of PHM. NIST; 2014. http://www.nist.gov/el/isd/upload/Keynote_Lee_IMS-University_of_Cincinnati_updated.pdf. Accessed 03 Aug 2015.
10. Wright P. Cyber-physical product manufacturing. Manuf Lett. 2014;2(2):49–53.
11. Smart Manufacturing Leadership Coalition. Implementing 21st century smart manufacturing. In Workshop Summary Report; 2011.
12. Smart Process Manufacturing Engineering Virtual Organization Steering Committee. Smart process manufacturing: an operations and technology roadmap. Full Report; 2009.
13. Zuehlke D. SmartFactory—towards a factory-of-things. Annu Rev Control. 2010;34(1):129–38.

14. Lee J, Kao H-A, Yang S. Service innovation and smart analytics for industry 4.0 and big data environment. Procedia CIRP. 2014;16:3–8.
15. Sharma P, Sharma M. Artificial intelligence in advance manufacturing technology-a review paper on current application. Int J Eng Manag Sci. 2014;1(1):4–7.
16. Heng S. Industry 4.0: huge potential for value creation waiting to be tapped. Deutsche Bank Research. 2014. http://www.dbresearch.com/servlet/reweb2.ReWEB?rwsite=DBR_INTERNET_EN-PROD&rwobj=ReDisplay.Start.class&document=PROD0000000000335628.
17. Chand S, Davis JF. What is smart manufacturing? Time Magazine Wrapper; 2010. p. 28–33.
18. Meziane F, Vadera S, Kobbacy K, Proudlove N. Intelligent systems in manufacturing: current developments and future prospects. Integr Manuf Syst. 2000;11(4):218–38.
19. Efthymiou K, Papakostas N, Mourtzis D, Chryssolouris G. On a predictive maintenance platform for production systems. Procedia CIRP. 2012;3(1):221–6.
20. Lee J, Bagheri B, Kao H. A cyber-physical systems architecture for industry 4. 0-based manufacturing systems. Manuf Lett. 2015;3:18–23.
21. Center for Intelligent Maintenance Systems. http://www.imscenter.net/. Accessed 03 Aug 2015.
22. Samad T, Frank B. Leveraging the web: a universal framework for building automation. In: Proceedings of the 2007 American control conference, 2007.
23. Nagorny K, Colombo AW, Schmidtmann U. A service- and multi-agent-oriented manufacturing automation architecture. Comput Ind. 2012;63(8):813–23.
24. Kastner W, Neugschwandtner G, Soucek S, Newman HM. Communication systems for building automation and control. 2005;93(6).
25. Xu X. From cloud computing to cloud manufacturing. Robot Comput Integr Manuf. 2012;28(1):75–86.
26. Wang XV, Xu XW. An interoperable solution for cloud manufacturing. Robot Comput Integr Manuf. 2013;29(4):232–47.
27. Alves Santos R, Normey-Rico JE, Merino Gómez A, Acebes Arconada LF, de Prada Moraga C. OPC based distributed real time simulation of complex continuous processes. Simul Model Pract Theory. 2005;13(7):525–49.
28. Hong X, Jianhua W. Using standard components in automation industry: a study on OPC Specification. Comput Stand Interfaces. 2006;28(4):386–95.