A data pipeline for PHM data-driven analytics in large-scale smart manufacturing facilities

P. O'Donovan[12], K. Leahy[12], D. Og Cusack[2], K. Bruton[12], D. T. J. O'Sullivan[12]

[1] *IERG, University College Cork, Ireland*
[2] *Department of Civil and Environmental Engineering, University College*

peter_odonovan@umail.ucc.ie

ABSTRACT

The term smart manufacturing refers to a future-state of manufacturing, where the real-time transmission and processing of information across the factory will be used to produce advanced manufacturing intelligence that can optimize every aspect of its operation. In recent years, initiatives and groups such as the Smart Manufacturing Leadership Coalition (SMLC), Industry 4.0, and the Industrial Internet Consortium (IIC), have led the way by bringing together industry, academia and government to establish policies, roadmaps and platforms to support smart manufacturing. Although there are many characteristics that can be associated with smart manufacturing across these initiatives, a common theme is the emphasis on transitioning operations from reactive and responsive, to predictive and preventative. The research presented in this paper focuses on the development of a data pipeline that supports the development of data-driven Prognostics and Health Management (PHM) applications. In the context of smart manufacturing, PHM enables facilities to transition from preventative and reactive maintenance strategies, to predictive, preventative and condition-based strategies. The benefits that can be derived by PHM are aligned with those of smart manufacturing, which include the opportunity to decrease costs, increase machine availability, reduce energy consumption, and improve production yield.

However, the process of ingesting, cleaning and transforming real-time data streams for data-driven PHM is a difficult, complex and time-consuming task, with estimates from business intelligence projects ranging from 80% to 90% of total project effort. This effort may only be exacerbated further in manufacturing environments due to additional technology challenges, such as low levels of standardization, disparate protocols and interfaces, and ad hoc data management. While emerging technologies such as Cyber Physical Systems (CPS) and Internet of Things (IoT) can overcome many of these challenges and provide an open platform for transmitting data, existing large-scale manufacturing facilities that are subject to compliance, regulation, and stringent quality assurance policies may not be able to adopt these technologies in the short-term due to the associated cost, risk and effort. Therefore, PHM applications that need to access data streams in large-scale manufacturing facilities must do so using transparent data integration that does not discriminate between emerging and legacy technologies in the factory. To this end, this research presents a real-time, scalable, robust, and fault tolerant data pipeline for ingesting, cleaning, transforming, processing and contextualizing time-series data from a wide-range of sources in the factory.

1. INTRODUCTION

The term smart manufacturing refers to a paradigm that describes the transmission and sharing of real-time information across pervasive networks with the aim of creating manufacturing intelligence in every aspect of the factory (Davis, Edgar, Porter, Bernaden, & Sarli, 2012a; Lee, Lapira, Bagheri, & Kao, 2013; Lee, 2014; Wright, 2014). Experts predict that smart manufacturing may become a reality in the next 10 to 20 years. The objective of smart manufacturing is similar to manufacturing intelligence insofar as it focuses on the transformation of raw data to knowledge, which can improve decision-making and have a positive impact on operations. However, smart manufacturing supersedes manufacturing intelligence in its emphasis on real-time data collection and aggregation, which facilitates knowledge sharing across physical and computational processes that can result in seamless operating intelligence (Manufacturing et al., 2011). In general terms, smart manufacturing can be considered an intensified application of manufacturing intelligence, where every aspect of the factory is monitored, optimized and visualized (Davis et al., 2012a).

While smart technologies facilitate the creation of knowledge, workers must apply this knowledge in some way before it can have a positive impact on operations. Therefore, while technology transformation is arguably the most publicized aspect of smart manufacturing, the transformation and education of workers should not be ignored. The demands placed on workers in smart manufacturing facilities are not be entirely limited to vertical operations, and will therefore require a multi-disciplinary perspective. Many of the technologies and systems associated with smart manufacturing discuss high data visibility across the factory, where the potential impact of a decision can be evaluated in the context of the entire facility rather than being isolated to a particular department. Without adopting this type of holistic decision-making, it is difficult to envisage how smart manufacturing objectives such as demand-driven and intelligent production, real-time data management, system interoperability, and cyber security, can be realized (Manufacturing et al., 2011). Therefore, decision-makers embedded in smart manufacturing operations will need a basic understanding of multiple disciplines, including engineering, computing, analytics, design, planning, automation, and production (Meziane, Vadera, Kobbacy, & Proudlove, 2000; Sharma & Sharma, 2014).

1.1. Groups and initiatives focused on smart manufacturing

There are a number of government, academic and industry groups promoting an awareness of smart manufacturing. These initiatives include the Smart Leadership Coalition (SMLC) (Manufacturing et al., 2011), Technology Initiative SmartFactory (Zuehlke, 2010), Industry 4.0 (Lee, Kao, & Yang, 2014), and The Industrial Internet Consortium (IIC). These initiatives formed from the realization that challenges facing smart manufacturing adoption are too big for any single organization to address, and while terminology used by initiatives may differ, they share an overarching vision of smart manufacturing where real-time data streams are used to realize operational efficiencies. The two most prominent smart manufacturing initiatives are the SMLC and Industry 4.0, with each loosely related to their geographical origin – the US and EU respectively.

The SMLC working group differs from other initiatives in a couple of ways. The SMLC is comprised of numerous academic institutions, government agencies and industry partners. This blend enables the SMLC to identify real problems by consensus, which may mitigate from bias recommendations that do not serve the wider manufacturing community. Furthermore, the SMLC have not only developed theoretical artifacts relating to smart manufacturing, such as roadmaps, recommendations and guidelines, they have also undertaken the development of a smart manufacturing platform that implements many of these ideas. Industry 4.0 is a high-tech strategy that was created by the German government to promote an awareness of smart manufacturing and its potential economic benefits. The term Industry 4.0 is a simple naming convention that serves to partition each industrial revolution, with 4.0 referring to an anticipated fourth revolution. Expert opinions differ regarding a realistic timeline for Industry 4.0, with general estimates ranging from 10 to 20 years. Exploring the Industry 4.0 naming convention further, previous industrial revolutions are predictably labelled 1.0, 2.0 and 3.0. Industry 1.0 was brought about by the introduction of mechanical production using water and steam power, with the first mechanical loom used in 1784. Industry 2.0 was brought about by the division of labor and the realization of mass production, which were largely facilitated by electrical energy, with the first assembly line introduced in the Cincinnati slaughter house in 1870. Finally, Industry 3.0 was brought about by advances in electronics and IT systems, which enabled automation of production using control networks, with the first programmable logic controller (PLC) Modicon 084 introduced in 1969.

1.2. Benefits of smart manufacturing

Smart manufacturing focuses on pervasive networking and intelligent data-driven analytics that are highly integrated, intelligent, and flexible. The combined application of these technologies can be used to facilitate highly customized and optimized demand-driven supply chains that can dynamically respond to the needs of the customer. Furthermore, smart manufacturing addresses many common business and operating challenges, such as increasing global competition and rising energy costs, while also facilitating shorter production cycles that respond quickly to customer demand (Manufacturing et al., 2011; Sharma & Sharma, 2014). In addition to these high-level efficiencies, more quantifiable benefits have also been cited. For example, the SMLC identified realistic performance targets for different aspects of smart manufacturing, including (1) a 30% reduction in capital intensity, (2) up to a 40% reduction in product cycle times, as well as (3) an overarching positive impact across energy, emissions, throughput, yield, waste, and productivity. Furthermore, smart manufacturing can also provide benefits to the wider economy. A recent report from Fraunhofer Institute and Bitkom highlights the potential economic benefit of Industry 4.0 to the German economy. The report states the transformation of traditional factories to Industry 4.0 could be worth 267 billion euros cumulatively to the German economy by 2025 (Heng, 2014).

1.3. Impediments to smart manufacturing adoption

While the potential benefits of smart manufacturing are apparent, there are numerous challenges and issues that must be overcome before they can be realized. In particular, facilities must develop the infrastructure and network-intensive real-time technologies needed to support smart manufacturing, as well as cultivating multidisciplinary workforces and next-generation IT departments that are capable of working with smart technologies(Manufacturing et al., 2011). The degree to which these challenges exist in each facility will vary. For example, there are obvious differences between implementation challenges in greenfield and brownfield sites (Davis, Edgar, Porter, Bernaden, & Sarli, 2012b). Excluding fundamental challenges, such as budgetary constraints, technology availability and the presence of a skilled workforce, greenfield sites are better positioned to adopt emerging smart technologies when compared with brownfield sites. Brownfield sites may be restricted by legacy devices, information systems, and protocols, which can also include proprietary and ad hoc technologies. These technologies are from a time when low latency distributed real-time networks and large-scale data storage and processing were simply not a concern. In some instances legacy technologies may be replaced with smarter equivalents, but there are numerous reasons why substitution may not be an option;

- *Historical investment in IT and automation.* Many facilities invested in information systems and automation networks over the last 40 years. Therefore, facilities may be reluctant to replace technologies that received significant investment and continue to operate at an appropriate level.
- *Regulatory and quality constraints*. In certain industries, such as pharmaceuticals and medical devices, internal or external constraints may exist in the form of regulatory and/or quality standards. In these instances, the existence of exhaustive processes and procedures may negate the enthusiasm for legacy technology replacement.
- *Dependency on proprietary systems or protocols*. While numerous open standards exist for manufacturing information systems and automation networks, such as ISA95 for system interoperability and OPC for device-level communication, their adoption is sporadic. Therefore, where proprietary and closed technologies are used in place of open standards, technology adoption (i.e. smart technologies) is limited by the proprietary vendors offerings.
- *Weak vision and insufficient commitment*. The transition to smart manufacturing is a significant undertaking that requires strong leadership and a shared vision of the short and long-term benefits for the facility. Facilities that do not have a clear vision of how smart manufacturing can improve their operations may be less likely to have an appetite for technology replacement.
- *High risk and disruption.* The implementation of new and emerging technologies and systems are considered high-risk projects, which can negatively impact operations while technical competency is being achieved. Therefore, the appetite to undertake large-scale IT projects may be weak until such time lost opportunities effect the facilities competitiveness.
- *Skills and technology awareness.* IT and automation departments are entrenched in mature computing, automation and networking methods that have been in existence for decades. However, technologies synonymous with smart manufacturing (e.g. IoT, CPS, Big Data, Cloud Computing) require a shift from these approaches. Therefore, if the relevant departments do not embrace these technologies and contribute to the organizations smart manufacturing roadmap, their lack of knowledge may impede technology replacement.

There are numerous impediments surrounding the introduction of technologies for smart manufacturing, but the majority of these relate to brownfield sites where technology replacement can be problematic. The main challenge facing brownfield sites is the encapsulation and integration of legacy technologies with emerging smart technologies, methodologies and roadmaps. Facilities that do not address these issues may be restricted in their adoption of smart manufacturing, and the realization of its associated performance enhancements and benefits.

This paper focuses on Prognostics and Health Management (PHM) applications for equipment maintenance in the context of smart manufacturing. PHM comprises methods for detecting and predicting equipment faults to optimize equipment uptime and availability (Bruton et al., 2014; Bruton, Coakley, O'Donovan, Keane, & O'Sullivan, 2013; Lee, Bagheri, & Kao, 2015; Lee et al., 2013; O'Donovan, Leahy, Bruton, & O'Sullivan, 2015; Wright, 2014). The main contributions of this paper are high-level requirements for data-driven smart manufacturing systems in highly regulated and quality controlled brownfield sites, where legacy integration may impede smart manufacturing adoption, and a system architecture that satisfies these requirements and enables real-time data ingestion and big data processing in the cloud.

2. RESEARCH METHODOLOGY

This research employed an embedded study, which was undertaken in DePuy Ireland - a large-scale manufacturing facility which is part of the Johnson & Johnson family of companies. The aim of this research was to identify the main requirements and associated system architecture, to support the development of PHM applications by reducing expensive and time-consuming activities, such as ad hoc data integration, and improving overall data accessibility and reusability.

2.1. Establishing scope

As there were numerous potential applications of PHM and industrial analytics in the context of smart manufacturing, the first priority was to establish research boundaries. After an initial discussion between research team members, and automation personnel in DePuy Ireland, the focus of the research was narrowed using the following specificities.

2.1.1. Type of PHM applications

It was agreed that research efforts would focus on data-driven applications that deal with predictive and intelligent equipment maintenance. Equipment uptime and availability was considered a critical aspect of operations given the potential impact downtime can have on production. Therefore, the development of a solution that can stream data directly to PHM applications focused on promoting machine uptime and availability was deemed a worthwhile pursuit.

2.1.2. Regulation and compliance

Given this research was undertaken in a highly regulated and quality-focused environment, and an empirical research methodology was employed, there is an implied narrowing of the research scope insofar as observations may only apply to facilities with the same characteristics. As legacy technology replacement is not easily achieved in these environments (e.g. smart technologies), it was agreed an emphasis would be placed on legacy technology integration, with the aim of amalgamating legacy and smart technologies in a single framework. This was considered a significant real-world challenge for brownfield sites, which could only improve the value of this research. Furthermore, it was agreed that the final solution requirements and prerequisites should be minimal (e.g. it should not require a facility to use a particular brand of controller).

2.1.3. Time-series data

This research focuses on data-driven PHM applications for equipment maintenance in the context of smart manufacturing. Therefore, it was agreed data ingestion, processing and accessibility aspects of the research could be limited to time-series data measurements. Based on experiences of research team members and feedback from automation personnel in DePuy Ireland, time-series data was the format most relevant to equipment maintenance monitoring, analysis and decision-making. By limiting the pipeline to a particular class of data the number of permutations for extraction, transformation and loading operations were reduced, given the predictable and low-dimensional structure of the data (e.g. time/value pairs).

2.1.4. Data flow and direction

The overarching theme of this research is the investigation of real-time data integration and transmission from large-scale industrial facilities. Therefore, it was agreed that data flows in the pipeline would only move one-way (i.e. factory to cloud) and this data would be immutable (i.e. read only). While this research may not consider a two-way communication channel for PHM applications to send instructions back to the factory, these applications can extend the framework and implement their own protocol to initiate actions in the factory if required.

2.1.5. Industrial data integration

Legacy integration was deemed an important aspect of this research given the prominence of proprietary systems and diverse communication protocols that can exist in industrial environments. However, given the broad and ill-defined nature of this problem it was agreed initial legacy integration would be limited to log files produced by Programmable Logic Controllers (PLC) and Manufacturing/Building Systems, OLE Process Control (OPC), Modbus, and BACnet.

2.1.6. Industry collaboration

To better understand the manufacturing systems, processes and technologies in DePuy Ireland, and to gain a greater appreciation for manufacturing operations in general, we engaged with internal teams across automation, energy, big data and smart manufacturing. Discussions with these teams assisted in the identification of data sources, processes and industrial protocols that were relevant to PHM applications in the factory.

- **Automation** - the automation team consisted of eight staff with skills covering control and automation, production, energy and information technology. The automation team informed the research teams understanding of infrastructure supporting production in the factory, as well as scheduling and maintenance strategies for machinery.
- **Energy** – the energy team comprised of five staff with skills in engineering and energy. The energy team informed the research teams understanding of energy consumption monitoring for equipment, as well as highlighting how malfunctioning equipment can produce energy fluctuations.
- **Big Data and Smart Manufacturing** - there are no dedicated teams currently responsible for big data and smart manufacturing. Therefore, the research team interacted with multiple teams and personnel to form a better understanding of how emerging technologies, such as Internet of Things (IoT) and big data

technologies, were being considered for use in the factory.

2.2. Research questions

Two research questions were identified to guide research efforts. The purpose of the first question (RQ1) was to establish the real-world data integration requirements for large-scale industrial environments, with an emphasis on those that are not supported by traditional data integration tools (e.g. ETL tools). The purpose of the second question (RQ2) was to create an open and accessible architecture for data ingestion, processing and management that could satisfy requirements identified by RQ1.

2.2.1. RQ1 – What requirements and characteristics are important to large-scale manufacturing facilities when it comes to data integration methods?

This question focuses on establishing requirements and characteristics that may support the development of data-driven PHM applications in real-world large-scale manufacturing environments, with a particular emphasis on facilitating transparent data flows across the entire factory, which is aligned with the vision of smart manufacturing.

2.2.2. RQ2 – How can a data pipeline serve data-driven PHM applications using legacy and emerging technologies in an indiscriminate manner?

This question considers the design of a data pipeline architecture that can provide a framework for PHM applications focused on equipment maintenance, while satisfying the requirements from RQ1. Recommendations from smart manufacturing are combined with those of RQ1 to further inform the pipelines design, incorporating the need for real-time capabilities, open standards, and seamless data access.

3. RESULTS AND DISCUSSION

3.1. RQ1 – Requirements and characteristics

The following requirements and characteristics were identified in response to RQ1 during the study. Although these findings were derived from discussions relating to PHM applications focused on equipment maintenance, they should also be considered representative of industrial data integration challenges facing facilities transitioning to smart manufacturing.

3.1.1. Legacy integration

Some facilities will not be in a position to adopt emerging and smart technologies to realize intelligent systems associated with smart manufacturing. Based on observations of the research team, many large-scale manufacturing facilities may have invested too much time and resources in control and automation networks to consider replacing legacy devices with smarter equivalents. Similarly, while many facilities may be aware of the potential benefits associated with emerging technologies, such as big data analytics, they may not know how they can integrate with existing operations, or fully appreciate the multi-disciplinary and technical skills needed to implement them in the facility. Therefore, facilities may want to leverage and maximize existing investments, skills, knowledge, vocabulary and systems, while incrementally transitioning to smart manufacturing rather than completely overhauling technologies and operations. To achieve this transparent data integration will be an important requirement, whereby legacy and smart technologies are abstracted to deliver indiscriminant data access.

3.1.2. Cross-network communication

Real-time data transmission across pervasive networks is a fundamental aspect of smart manufacturing. However, networks in modern manufacturing facilities were not designed with these characteristics in mind. The research team encountered several instances during the study where equipment maintenance data was restricted by firewalls and other security measures. Furthermore, limited access to equipment data was also encountered due to external maintenance and support agreements with vendors (e.g. wind turbines). While these measures may make sense in the context of traditional manufacturing operations, they represent a challenge to data-driven smart manufacturing. Therefore, to provide data visibility across facilities (and/or multiple sites) it may be necessary to communicate across secure networks.

3.1.3. Fault tolerance

Information systems and technologies that play a role in production, automation and maintenance may have high demands placed on them given their ability to directly impact facilities production yield and operational efficiency. Based on observations of the research team, information systems deployed in industrial environments must be highly available and fault tolerant. Therefore, these characteristics will be expected of new systems and tools operating in similar environments.

3.1.4. Extensibility

Proprietary and/or ad hoc technologies and systems in large-scale manufacturing facilities are common. Based on observations of the research team, it appears that facilities have become more aware of technology integration and consolidation, but duplication and disparity across systems is still evident. Some of these inefficiencies may be due to the inextensibility of existing information systems, which can result in ad hoc implementations. Therefore, an important requirement for systems operating in industrials environments is extensibility, where new data types, methods and protocols can be supported as requirements emerge.

3.1.5. Scalability

As the digitization of factories accelerate, the ability to dynamically scale based on demand is becoming a desirable characteristic for industrial information systems. This is especially relevant when considering emerging technologies (e.g. IoT) in smart manufacturing, and the unknown load they will place on these systems. While modern large-scale manufacturing facilities are entrenched in technology, the real-time and data-rich nature of smart manufacturing may expose unforeseen limitations due to their inability to scale. For example, the normal resolution for data measurements observed during the study was 15 minutes. Without considering the addition of new sensors and measurements (i.e. IoT), the data production rate in the facility would increase by 900% if measurement intervals were reduced to 1 second. Therefore, the ability to scale based on demand is an important requirement.

3.1.6. Data accessibility

Modern large-scale manufacturing facilities produce a lot of data. However, based on observations of the research team, access is inhibited by diverse protocols, formats and structures. These issues are currently overcome using expensive data discovery and integration procedures, which focus on proprietary and ad hoc data integration routines that address the peculiarities of a particular project. However, these approaches typically result in poor reuse, which results in tasks being repeated and duplicated across projects. Therefore, it is important to abstract and generalize low-level data integration routines to provide a consistent data interface for data sources and devices in the factory.

3.2. RQ2 – Data pipeline architecture

A high-level data pipeline architecture for data-driven PHM applications focused on equipment maintenance was produced using the requirements from RQ1. Figure 1 illustrates the data pipeline architecture, with each stage of the factory-to-cloud workflow numbered and highlighted. The aim of the data pipeline is to deliver a low cost turnkey solution for industrial data integration, which is built on a real-time, open, scalable and fault tolerant infrastructure. The purpose and function of each component and stage in the data pipeline is described in the proceeding sections.

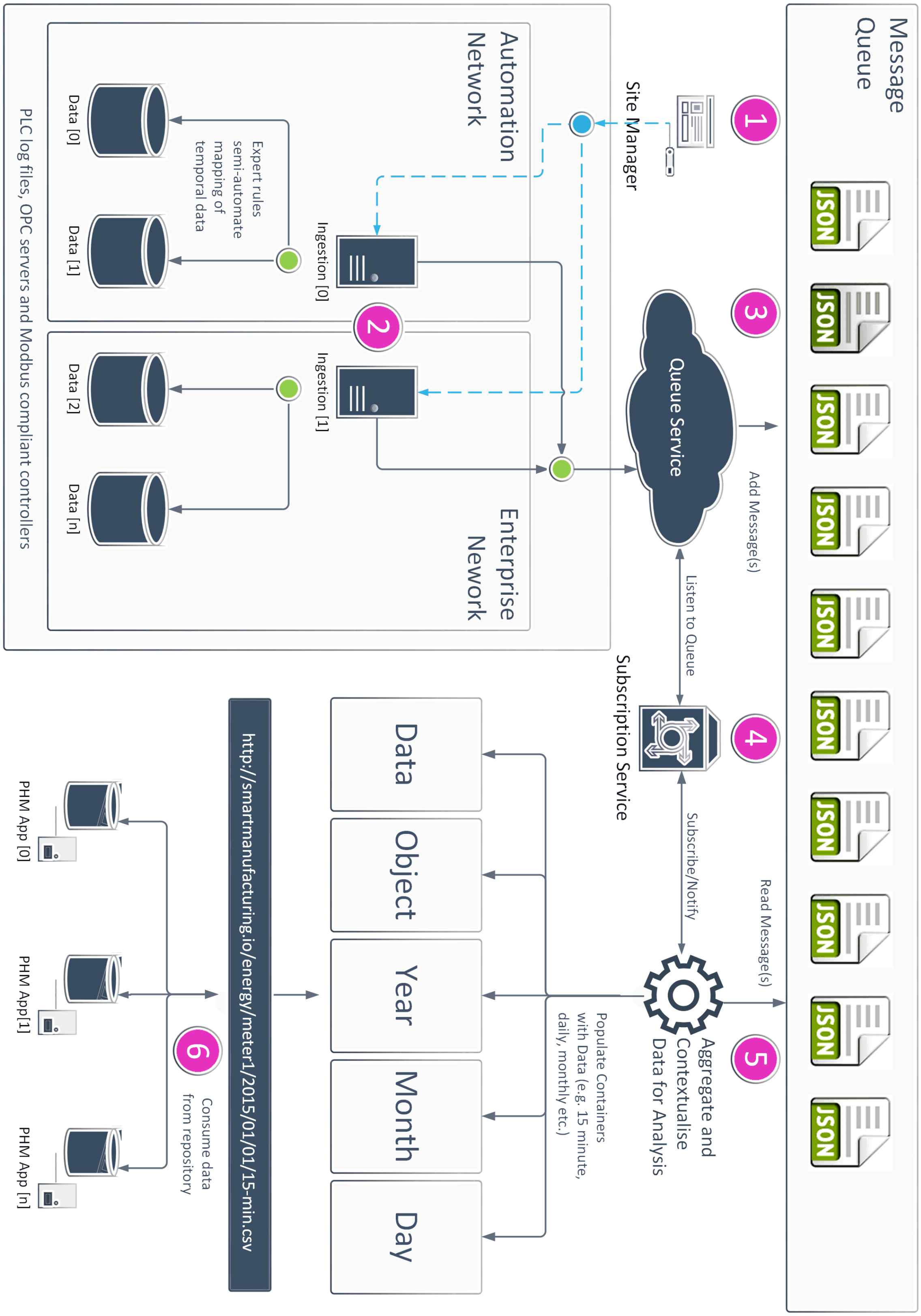
Message Queue
1
3
4
5
JSON
Site Manager
Add Message(s)
Queue Service
Listen to Queue
Subscription Service
Subscribe/Notify
Read Message(s)
Aggregate and Contextualise Data for Analysis
Populate Containers with Data (e.g. 15 minute, daily, monthly etc.)
Automation Network
Enterprise Nework
2
Ingestion [0]
Ingestion [1]
Expert rules semi-automate mapping of temporal data
Data [0]
Data [1]
Data [2]
Data [n]
PLC log files, OPC servers and Modbus compliant controllers
Data
Object
Year
Month
Day
http://smartmanufacturing.io/energy/meter1/2015/01/01/15-min.csv
6
Consume data from repository
PHM App [0]
PHM App[1]
PHM App [n]

Figure 1. Data pipeline architecture and workflow

3.2.1. Stage 1 – Site manager

Purpose: The site manager resides on a cloud server and stores meta-data regarding each facility and associated data sources. Its purpose in the architecture is to persist essential site information, such as the location of each data point and the type of protocol that needs to be used for integration.

Functions: The site manager has multiple functions that are related to the factory – (1) store details relating to the site, such as the type and location of local data sources that shall be ingested, (2) schedule and assign jobs to ingestion engines in the factory based on the availability and location of each node, and (3) derive a suitable amount of data to ingest for each engine based on its current location, bandwidth, CPU and bandwidth availability.

3.2.2. Stage 2 – Ingestion process

Purpose: Ingestion engines are distributed software agents that are deployed across networks in the factory to collect and integrate time-series data for different data-driven PHM applications (e.g. HVAC, Chillers, Boilers). They execute as background workers on a server and continually send their status to the site manager (Stage 1), and when instructed, collect and transmit data from local data sources to the cloud. As illustrated in Figure 1, the distributed and autonomous nature of an ingestion engine enables them to be deployed across different networks that are separated by firewalls and/or geographical boundaries. Furthermore, these characteristics also allow the ingestion process to scale by deploying more ingestion engines, which increases the throughout capacity of the pipeline from the factory.

Functions: The ingestion engine has multiple functions – (1) communicate location, bandwidth, CPU and memory availability to the site manager so an appropriate ingestion task can be assigned, (2) interpret ingestion tasks sent by the site manager and automatically extract time-series data from the relevant sources in accordance with the task parameters (e.g. particular date range), and (3) transmit the collected time-series data to the cloud message queue. A novel aspect of the ingestion process is an expert ruleset that can automatically map and extract time-series data to limit expensive and manual data mapping tasks.

3.2.3. Stage 3 – Message queue

Purpose: The highly available and distributed message queue service in the cloud accepts time-series data from ingestion engines in the factory. Its main purpose is to provide intermediary storage between the factory and processing components in the pipeline. This decouples data ingestion components from data processing components, which instills resilience by facilitating asynchronous communication and parallel operations when the pipeline is at peak demand.

Functions: The message queue has two main functions – (1) notify subscription service when new data has been ingested, and (2) add received data to a queue so data processing components can access it further down the pipeline.

3.2.4. Stage 4 – Subscription service

Purpose: The subscription service provides an endpoint for the data ingestion process and functions as a notification mechanism for data processing components when new data is received. The notification of new data results in one or more data preparation and/or analysis tasks being undertaken. The number of data processing actions executed can be increased or decreased by subscribing or unsubscribing from the subscription service.

Functions: The functions of the subscription service are limited, but essential in the orchestration of events in the pipeline – (1) listen to the message queue for new data and (2) notify subscribers when new data are available for processing.

3.2.5. Stage 5 – Data processing

Purpose: Data processing components are responsible for transforming raw time-series data to a format suitable for analysis. The aim of data processing is to remove the onus on each PHM application to undertake expensive and time-consuming operations. At the most basic level the pipeline aggregates time-series data at different levels of granularity, such as hourly, daily, monthly and annual averages. Examples of more sophisticated processing may include the execution of expert rules to identify faults, or the semantic encoding of time-series data (e.g. Project Haystack) to promote interoperability with other applications. Each data processing component in the architecture is responsible for executing a single processing operation to promote modularity. This enables new data processing components to be easily added to the pipeline as new requirements emerge.

Functions: The functions that may be associated with data processing are truly diverse. Therefore, data processing components in the pipeline cannot be strictly prescribed given processing requirements will vary from application-to-application, and factory-to-factory. However, common use cases could be built over time to form a library of default processing components. The current default scenario illustrated in the data pipeline is time-series aggregation – (1) daily average, (2) monthly average, and (3) annual average.

3.2.6. Stage 6 – Data access

Purpose: The data access stage exposes a consistent and open method for PHM applications to consume data that originated from equipment in the factory. A naming convention is used to promote consistency in data access and contextualize data requests. The convention uses an encoded URL to request data for an object (e.g. HVAC) and date. Figure 1 illustrates the naming convention between stage 5 and 6 in the pipeline, and Table 1 describes each parameter of the naming convention in more detail;

Table 1. URL convention for data requests

Parameter	**Description**
Data	Refers to a particular data set (e.g. energy).
Object	Identifying name or code that exists within the data set (e.g. machine number).
Year	Year relevant to data request/query.
Month	Month relevant to data request/query.
Day	Day relevant to data request/query.

Functions: Functions relating to data access include – (1) ensuring data are stored in the appropriate location as per the naming convention and (2) return the appropriate data for requests that utilize this convention.

3.3. Alignment of architecture with requirements

This section discusses how the data pipeline architecture from RQ2 satisfies requirements from RQ1. Table 2 supports this discussion by describing how different stages of the pipeline address different requirements.

Table 2. Relationship between requirements and architecture

Requirement	**Data pipeline stages**
Legacy integration	**Stages 1 and 2** in the architecture are responsible for legacy integration. The site manager creates meta-data for each data source in the factory, which is then used by the ingestion engine to extract data independent of the underlying source (i.e. either legacy or smart).
Cross-network communication	**Stages 2 and 3** in the architecture facilitate communication across networks. Ingestion engines are remote autonomous agents that are network agnostic. Therefore, given an outbound connection to the queue service in the cloud, ingestion engines can be deployed across multiple networks to unify data in the pipeline.
Extensibility	**Stages 2, 5 and 6** promote extensibility in the data pipeline architecture. First, data ingestion instructions are dynamically disseminated from the site manager, which means these instructions can be extended to support new types of data sources etc. Second, data processing components are modular, which enables processing capabilities of the pipeline to be extended through the addition of new processing components. Finally, the type of data served to PHM applications via the data interface can be extended to include additional formats.
Scalability	**Stages 1-6** illustrate how scalability is embedded in the pipeline. First, the distributed design of the ingestion process is realized using autonomous agents, which enables integration routines to run in parallel and scale based on the number of data points being measured. Second, the message queue, notification and storage services are inherently scalable given the selection of a cloud provider that supports auto-scaling. Finally, data processing components (i.e. workers) can benefit from load balancing and auto scaling features of cloud computing, which enables these components to dynamically distribute and scale based on the quantity of data to be processed.
Data accessibility	**Stage 6** provides a common interface for PHM applications to consume data from the pipeline. The architecture employs a cloud-based repository to serve low-latency precompiled views of time-series data to geographically distributed end-users and PHM applications. Furthermore, interoperability with 3rd party applications is supported by the use of open standards and protocols (e.g. HTTP, JSON etc.).

4. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a set of challenges and characteristics associated with collecting and integrating industrial data for data-driven manufacturing, and a system architecture that addresses these challenges. In particular, the system architecture supports real-time data ingestion from a range of legacy and smart devices throughout the factory, while using a mix of novel and conventional technologies to promote fault tolerance, scalability and accessibility. The automated data pipeline architecture can provide facilities with a robust, flexible and adaptable managed framework to enable data-driven manufacturing (i.e. smart manufacturing) while mitigating low-level technical details, such as legacy and smart technology integration. While emerging smart sensors and technologies (e.g. IoT) will eventually eliminate the need for legacy integration, given the fact 20 year old PLC's are still in operation, it is advisable that researchers and innovators should be conservative when estimating timelines for when large-scale industrial facilities will be operating using smart technologies exclusively. Therefore, to transition to smart manufacturing and develop the insightful data-driven applications that can deliver predictive and efficient operations, facilities must be capable of addressing the fundamental issue of transparent data integration.

ACKNOWLEDGEMENTS

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).

REFERENCES

Bruton, K., Coakley, D., O'Donovan, P., Keane, M. M., & O'Sullivan, D. (2013). Development of an Online Expert Rule based Automated Fault Detection and Diagnostic (AFDD) tool for Air Handling Units: Beta Test Results. In *ICEBO - International Conference for Enhanced Building Operations*. Montréal, Canada.

Bruton, K., Raftery, P., O'Donovan, P., Aughney, N., Keane, M. M., & O'Sullivan, D. T. J. (2014). Development and alpha testing of a cloud based automated fault detection and diagnosis tool for Air Handling Units. *Automation in Construction*, *39*, 70–83. doi:10.1016/j.autcon.2013.12.006

Davis, J., Edgar, T., Porter, J., Bernaden, J., & Sarli, M. (2012a). Smart manufacturing, manufacturing intelligence and demand-dynamic performance. *Computers & Chemical Engineering*, *47*, 145–156. doi:10.1016/j.compchemeng.2012.06.037

Davis, J., Edgar, T., Porter, J., Bernaden, J., & Sarli, M. (2012b). Smart manufacturing, manufacturing intelligence and demand-dynamic performance. *Computers and Chemical Engineering*, *47*, 145–156. doi:10.1016/j.compchemeng.2012.06.037

Heng, S. (2014). Industry 4.0: Huge potential for value creation waiting to be tapped. Deutsche Bank Research. Retrieved from http://www.dbresearch.com/servlet/reweb2.ReWEB?rwsite=DBR_INTERNET_EN-PROD&rwobj=ReDisplay.Start.class&document=PROD0000000000335628

Lee, J. (2014). Recent Advances and Transformation Direction of PHM, 1–31.

Lee, J., Bagheri, B., & Kao, H. (2015). A Cyber-Physical Systems architecture for Industry 4 . 0-based manufacturing systems. *MANUFACTURING LETTERS*, *3*, 18–23. doi:10.1016/j.mfglet.2014.12.001

Lee, J., Kao, H.-A., & Yang, S. (2014). Service Innovation and Smart Analytics for Industry 4.0 and Big Data Environment. *Procedia CIRP*, *16*, 3–8. doi:10.1016/j.procir.2014.02.001

Lee, J., Lapira, E., Bagheri, B., & Kao, H. (2013). Recent advances and trends in predictive manufacturing systems in big data environment. *Manufacturing Letters*, *1*(1), 38–41. doi:10.1016/j.mfglet.2013.09.005

Manufacturing, S., Manufacturing, C. S., Coalition, L., Smart, T., Leadership, M., Incorporated, E., … Any, D. (2011). *About this Report About the Smart Manufacturing Leadership Coalition*.

Meziane, F., Vadera, S., Kobbacy, K., & Proudlove, N. (2000). Intelligent systems in manufacturing: current developments and future prospects. *Integrated Manufacturing Systems*, *11*(4), 218–238. doi:10.1108/09576060010326221

O'Donovan, P., Leahy, K., Bruton, K., & O'Sullivan, D. T. J. (2015). Big data in manufacturing: a systematic mapping study. *Journal of Big Data*, *2*(1), 20. doi:10.1186/s40537-015-0028-x

Sharma, P., & Sharma, M. (2014). Artificial Intelligence in Advance Manufacturing Technology-A Review Paper on Current Application, (1), 4–7.

Wright, P. (2014). Cyber-physical product manufacturing. *Manufacturing Letters*, *2*(2), 49–53. doi:10.1016/j.mfglet.2013.10.001

Zuehlke, D. (2010). SmartFactory—Towards a factory-of-things. *Annual Reviews in Control*, *34*(1), 129–138. doi:10.1016/j.arcontrol.2010.02.008