Industrial Big Data: From Data to Information to Actions

Andreas Kirmse[1], Felix Kuschicke[2] and Max Hoffmann[1]
[1]*Institute of Information Management in Mechanical Engineering (IMA), RWTH Aachen University, Aachen, Germany*
[2]*Konica Minolta, Darmstadt, Germany*

Keywords: Industrial Big Data, Industrial Data Lakes, Information Integration, Data Acquisition, Cyber-physical Systems, Industry 4.0, Smart Manufacturing, Information Systems, RDBMS, OPC UA, MQTT.

Abstract: Technologies related to the Big Data term are increasingly focusing the industrial sector. The underlying concepts are suited to introduce disruptive changes in the various ways information is generated, integrated and used for optimization in modern production plants. Nevertheless, the adoption of these web-inspired technologies in an industrial environment is connected to multiple challenges, as the manufacturing industry has to cope with specific requirements and prerequisites that differ from common Big Data applications. Existing architectural approaches appear to be either partially incomplete or only address individual aspects of the challenges arising from industrial big data. This paper has the goal to thoroughly review existing approaches for industrial big data in manufacturing and to derive a consolidated architecture that is able to deal with all major problems of the industrial big data integration and deployment chain. Appropriate technologies to realize the presented approach are accordingly pointed out.

1 INTRODUCTION

Increasing customer demands regarding product quality and diversity as well as fast changing markets and strong competitors pose major challenges to the manufacturing industry (Brecher and Özdemir, 2017). While factory floor digitization, networking capabilities and automation significantly improve all areas of manufacturing companies, it also creates a staggering amount of available production data: the so-called Industrial Big Data (IBD) (General electric, 2012), which characterizes the increasing amount of data that is collected in industrial environments. The term has been adapted from the broader Big Data (BD) term covering all types of data and different data sources like social media, environmental or consumer data.

The digital transformation of industrial environments and a new global availability of information enabled by methods of IBD allows for new ways of realizing data-driven potentials for producing companies. The main business drivers which can be addressed by utilizing IBD are: Equipment costs reduction, product quality assurance (also warranty and after-sales management), and operational efficiency improvement (Intelligence, 2009). In order to extract the desired information from the (raw) industrial big data, various transformation steps have to be pursued: (1) data transformation (i.e. turning data into information and finally into insights), and (2) a transformation of the information user (i.e. the human being that uses insights to implement improvement measures) (Hammer et al., 2016). In order to close the feedback loop from the shop floor, followed by profound decision making through information users and finally back to the manufacturing field, new data-driven strategies, road-maps and concrete IT infrastructure planning are in need (Mourtzis et al., 2016).

In this paper, we address the entire tool chain of transforming (raw) data into useful information. The utilization of production data for insights and further improvements require several major steps to realize extraction of value from (I)BD. In principle, these major steps are implemented by a continuous tool chain and are furthermore independent of the application or use case. For this purpose, various guidelines such as reference architectures/layer structures – some of them specific for manufacturing – have been developed in the recent years. However, these guidelines are either partially incomplete or address only single aspects of BD technologies. In most BD use cases, unstructured data with low meta data footprint must be processed, while in IBD use cases, rather structured and fast generated data with known meta information and high variety of information is targeted.

Kirmse, A., Kuschicke, F. and Hoffmann, M.
Industrial Big Data: From Data to Information to Actions.
DOI: 10.5220/0007734501370146
In *Proceedings of the 4th International Conference on Internet of Things, Big Data and Security (IoTBDS 2019)*, pages 137-146
ISBN: 978-989-758-369-8

Accordingly, the specific requirements of industrial big data in contrast to (traditional) big data have to be pointed out to meet the challenges of modern factories. As mentioned above, the questions to be answered focus on how to accurately structure (known) meta-data and how to systematically extract useful information from massively generated data. Both, the format of the raw data as well as the representational form of the generated information suitable for the information user have to be clearly structured and adaptable to the targeted application or use-case. The further discussion of these requirements leads to the research question of this paper:

- How is manufacturing data currently generated and stored for industrial data solutions?
- Which steps are needed to realize the successful acquisition of information from the field level up to information management systems?
- What are the fundamental building blocks of an Industrial Big Data reference architecture?
- Which major technologies and/or methods can be used to qualify for the tasks of each step?

The authors of this research publication have been actively involved in the transformation processes of large-scale manufacturing companies and propose theoretical approaches as well as field-proven methods on how to implement these concepts into real-life applications on an industrial scale.

2 RESEARCH BACKGROUND

Similar to the traditional Big Data the Industrial Big Data term characterizes an umbrella concept for storing and dealing with huge amounts of diverse, fast incoming data. The technologies connected to IBD are increasingly used in the area of industrial production, since the hardware requirements for realizing such data-intensive environments are continuously decreasing. Driven by embedded systems (e.g. realized through edge-computing devices) and by a higher pervasion of networking technologies, IBD represent a common condition in modern factories.

The enabler technologies for IBD are strongly connected to the acquisition and networking of information across various locations of a production site. The most common realization of these technologies are characterized by terms such as Cyber-Physical Systems (CPS) and more specifically for the field of production as Cyber-Physical Production Systems (CPPS). Another enabler technology is referred to as the (Industrial) Internet of Things (IIoT), a term which characterizes the global availability of information through applications with a low footprint. After a description of these foundations, the field Industrial Big Data and its according implementation through IBD architectures is described in detail.

2.1 The Industrial Internet of Things & Cyber-physical Production Systems

One definition of Big Data characterizes its purpose quiet well by stating that “The world has always had ’big’ data. What makes ’big data’ the catch phrase [...] is not simply about the size of the data. ’Big data’ also refers to the size of available data for analysis, as well as the access methods and manipulation technologies to make sense of the data.” (Ebbert, 2018). Thus, as pointed out, BD is more about the availability of information, which due to the technological advances can be easily collected by making use by small and powerful embedded devices. The networking of the collected data finally enable its global availability for various applications.

The technological umbrella terms IIoT and CPPS represent technologies that intend to implement the described tool chain of information acquisition, collection, integration and usage. For the purpose of the present research work, we consider IIoT and CPPS as synonyms, like it is stated in (Jeschke et al., 2017) and (US Dept. of Commerce blog, 2014). Both concepts are derivatives of the non-production related Internet of Things and Cyber-Physical Systems terminology and basically describe the transfer to a production terminology (Sadiku et al., 2017). While there is a relatively sharp distinction between IoT and CPS, which is characterized by the capability of CPS to perform edge computing in the field while IoT are intended to solely provide data from distributed devices, the terms IIoT and CPPS are much closer as IIoT are equally characterized by enabling smart applications on the shop floor.

The term IIoT describes a concept, where data exchange of production devices enables embedded technologies and interconnected networks. The IIoT thereby implies the use of sensors and actuators, control systems, machine-to-machine (M2M) communication, data analytics, and security mechanisms (Mourtzis et al., 2016). According to Gartner, there will be nearly 20 billion devices connected to the IoT by 2020 – with the large majority of them coming from the industrial sector. The adoption of the IIoT results in the gradual replacement of classical network architectures like the automation pyramid (VDI/VDE-Gesellschaft, 2013) and enables the direct access to equipment data.

In an IIoT environment, data, services and functions are stored and processed where they are needed, in contrast to the traditional approach, in which the data was manipulated to fit to the systems of a grown ecosystem of information management systems, i.e. characterized by the different levels of the automation pyramid. The implementation of such data-driven approach requires new design patterns in order to comply with these business needs. The business requirements might include an application of various different use-cases to same information or a context-specific visualization of information depending on the person regarding the data. A thorough list of concerns and challenges resulting from this fact can be found in (Jeschke et al., 2017).

After a successful implementation of IIoT the acquisition of vast amounts of data can be accomplished that finally leads to the presence of IBD. Depending on the protocol, which is implemented in terms of the IIoT application, the integrated information is present in a rather structured form. In further steps this data will be processed to enable a deeper analysis to extract insights and valuable information.

2.2 Industrial Big Data

One of the most important outcomes of emerging IIoT is the generation of large data volumes centrally accumulated and stored, which grows at an unprecedented rate – this volatility of generation speed in data is one of the major characteristics about IBD (Mourtzis et al., 2016). According to (McKinsey, 2017), in 2010 manufacturing stored more data than any other sector – estimated two exabytes. To summarize the definition of IBD, the basic characteristics of IBD are the high volume, velocity, and variety of data (Laney, 2011); although new characteristics are being continuously introduced with "value" being the most important (Yin and Kaynak, 2015). In comparison with BD, IBD usually has a higher data quality and is more structured, more correlated, more orderly in time and more prepared to extract insights (for both low and advanced methods) (Lee et al., 2015). However, IBD has higher demands in terms of flexibility and application-specific utilization of data.

According to (Kuschicke et al., 2017), the term IBD stands not only for industrial data itself, but also for the techniques and methods to utilize and process the data. To underline this characterization, the definition of Wilder-James fits quiet well: "Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or does not fit the structures of your database architectures. To gain value from this data, you must choose an alternative way to process it. [...] To clarify matters, the three Vs of volume, velocity and variety are commonly used to characterize different aspects of big data. They are a helpful lens through which to view and understand the nature of the data and the software platforms available to exploit them." (Wilder-James, 2012) Thus, BD as well as IBD solutions are build around this viewpoint on Big Data, seeking for methods how to deal with the characteristics of volume, velocity and variety.

In addition to this, IBD involves further methods such as data acquisition, storage, and management techniques. To make use of the gathered and consolidated data, methods originated from the domains of data visualization, data mining, machine learning, and artificial intelligence are applied (Chen and Zhang, 2014); (Gluchowski et al., 2007) and accordingly complete the tool-set of IBD.

2.3 Industrial Big Data Architectures

The application of IBD techniques requires certain guidance, as the process of gaining valuable knowledge from industrial data involves several steps. For optimal results (little effort, short delivery time, high value) the different steps need to be performed in close coordination. Therefore, different approaches – from high level solutions up to detailed instructions with examples – were developed and presented so far. Unfortunately, the existing approaches do not comprehensively match the requirements and needs of the manufacturing industry. The following chapter contains a survey, which is not intended to be complete, but should rather provide an overview about the main streams in this research area. The examples provided offer a rather high level and functional focus or target generic solutions.

One high level architecture for BD is the so called 'big data pipeline' published in (Bertino et al., 2011). The authors describe a serial process of multiple phases which are necessary steps to enable the analysis and accordingly the exposure of the hidden potentials in the data. The pipeline consists of the phases "acquisition / recording", "extraction / cleaning / annotation", "integration / aggregation / representation", "analysis / modeling" and "interpretation". The pipeline gives readers a basic overview on how to generally extract value from BD. However, the presented approach does not describe how to link the various phases in the pipeline, nor how to implement the contents of the individual phases.

A more detailed approach has been developed by the Industrial Internet Consortium (Industrial Internet Consortium, 2017) which is referred to as the Industrial Internet Reference Architecture (IIRA). The introduced reference architecture consists of five functional domains (control, operations, information, application, and business). The last three functions represent the functionality of the big data pipeline as they contain data generation, acquisition, transformation, storage, access and analysis including a HMI.

A reference architecture model, which particularly popular in Germany, is the Reference Architectural Model Industry 4.0 (RAMI 4.0) (VDI/VDE Society Measurement and Automatic Control (GMA), 2015). This model introduces a three-dimensional view on "Industry 4.0" based on the layer structure of the Smart Grid Architecture Model (CEN-CENELEC-ETSI Smart Grid Coordination Group, 2012). Additional, a life cycle and value stream dimension plus a hierarchy level dimension complete the model. A major focus lies on the data acquisition step of Industrial Big Data. The model introduces a so-called 'Administration Shell' to collect data from shop-floor devices. This Shell contains the features device meta-data management (head) and management of data transfer (body) and so, turns devices into (smart) I4.0 objects. Lastly, the model only partially addresses the subsequent steps such as data storage, access and analysis.

Another German centered approach has been developed by the Fraunhofer society (Society, 2017). The so-called "Industrial Data Space" focuses mainly on the description of different roles within one "data ecosystem". Similar to RAMI 4.0 it introduces a five-layer structure. Each role has certain functions, depending on the layer. E.g., authorization of data usage is a task for the data owner in the functional layer. As a summary, the reference architecture is an approach to establish roles and to assign responsibilities in the data space on a higher level.

A simpler reference architecture is introduced and applied in (Pääkkönen and Pakkala, 2015). The architecture consists of the elements "data Source", "data extraction", "data loading and preprocessing", "data processing", "data analysis", "data loading and transformation", "interfacing and visualization", and "data storage". The different elements are mapped against infrastructures of tech-giants (e.g. "Facebook", "Twitter", "Netflix", etc.). In course of this, a more detailed and applicable model is generated, which covers mostly the integration, processing and analytics steps of traditional BD ecosystems.

A similarly detailed and applicable architecture is presented in (Chen et al., 2014). The key elements are "data generation", "data acquisition", "data storage" and "big data analysis". Each key element itself has sub-elements containing information that are more detailed on a lower abstraction level, e.g. including higher granularity data. Additional, available technologies are mapped against the tasks of the sub-elements, analog to (Pääkkönen and Pakkala, 2015).

Another lightweight modular based integration architecture focuses on the issue of bringing brownfield devices with proprietary protocols into a harmonized representation of information is presented in (Kirmse et al., 2018). It thereby addresses legacy devices and integrates generated data with existing network zones separating the production floor from typical office and analytic areas as well as multiple factory locations globally.

A variety of additional approaches has been published such as in Constance (Hai et al., 2016), a maintenance approach in (O'Donovan et al., 2015) or a high-level description of BD life cycle and infrastructure in (Demchenko et al., 2013). Further reference architectures for Big Data can be found in (Pääkkönen and Pakkala, 2015).

Lastly, a critical review of some reference architectures for smart manufacturing is conducted in (Moghaddam et al., 2018). Here the focus lies on individual interviews with seven experts on the two questions of the characteristics of a reference architecture for smart manufacturing and the required steps for businesses to get there. Their findings show that a unification is necessary and the diverse architectures start from different views, despite the common goal to strive towards a (macro) service-oriented architecture. All lack the definition of micro services and how humans interact in these systems and environments

3 THE BASIC ELEMENTS OF BIG DATA REFERENCE ARCHITECTURES

A thorough review of the previously mentioned reference architectures leads to the deduction of the following basic elements (and their tasks), which are needed to turn data into useful information and finally into action:

1. Industrial Big Data *Sources and Generation*
2. Industrial Big Data *Acquisition and Preparation*
3. Industrial Big Data *Storage and Access*
4. Industrial Big Data *Processing and Analysis*
5. Industrial Big Data *Information Presentation and Interfaces*

The following section describe each of these core elements with respect to realize the entire integration chain from (raw) industrial big data up to an analytics cloud and/or human decider.

3.1 Sources and Generation

The starting point of Industrial Big Data is situated at the very low level of the factory floor, where the raw data is generated by means of various devices, control units, robots, etc. Accordingly, the prior task of this element consists in the provision of generated data. This generated data hereby represents raw material of the information value chain. The quality of this data is of fundamental importance for all further steps of the information value and integration chain. However, despite the importance of raw data, the arising Industrial Big Data itself is only the result of the digitization of production processes and not its cause. The driver of the mentioned generation are mainly due to increasing equipment with more sensors that read various amounts of different information and monitor specific processes in high frequency.

3.2 Data Acquisition and Preparation

Data acquisition is the process that bridges the gap between the sole existence of distributed information and their actual collection. The capabilities of data acquisition are strongly affected by the process of machine communication and the protocols that are used for the internal data transfer between machine, automation devices and control units. This does not yet apply to the data format itself, but to the general way of accessing the source of the data generation. For the different hierarchy levels and various systems, these protocols and representational forms of data are quite diverse in nature. However, due to standards of protocols there exists a common ground for communication exchanges. On the lowest field level, where process controlling is important and thus time-sensitive, the demand is different from the upper levels, providing reporting of Key Performance Indicators (KPI).

3.3 Storage and Access

Data storage takes care of a persistence of all acquired data, but also to store possible intermediate results and (pre-)processed information. The storage system itself has to cope with the high volume of data and still deliver performance when someone, such as a data analyst, requests specific data. In combination with responsibilities connected to data protection as well as to data governance, which are both summarized under the paradigms of data security, the data storage represents the central point of handing out access to the gathered data after initial persistence.

3.4 Processing and Analysis

IBD processing and analysis represent the steps, in which the actual value creation of the entire IBD pipeline takes place. Hereby, the raw data is transformed to valuable information. For this purpose, the raw data is transformed, filtered and/or modeled with a varying degree of automation and manual efforts. The task is to filter out uninteresting data (noise) to obtain only valuable information. IBD processing and analysis can be classified according to response time into real-time and off-line analysis.

3.5 Information Presentation and Interfaces

Providing the results of the previous analysis steps to a user is an essential task, as it influences the process of transforming data to actions the most. The process of visualizing information to some higher instance can target humans but also machines. Human Machine Interfaces (HMI) includes a variety of different user interfaces (xUI) with the most common are graphical (visualization) and email/signal alerting. In comparison, machine interfacing usually involves the information transmission through digital protocols read by the receiving machine. The task of presentation and HMI is to provide the right information at the right time in the right format and quantity.

4 INDUSTRIAL BIG DATA TECHNIQUES AND METHODS

The described tasks of each step in the Industrial Big Data value chain can be fulfilled using a variety of different methods, techniques and applications. The following section, therefore, gives an overview of available methods, techniques and applications, shows basic information about them and provides decision support for the selection process.

4.1 Sources and Generation

Industrial Big Data Sources

Manufacturing data and their information backbones are usually structured according to the Purdue Model (Williams, 1994) or for German based companies more familiar the automation pyramid. This industry adopted reference model shows the interconnections and interdependencies of all main components of a manufacturing control system form the bottom to the top. Accordingly, data from manufacturing systems are available at different levels, in different density or granularity and in different quality.

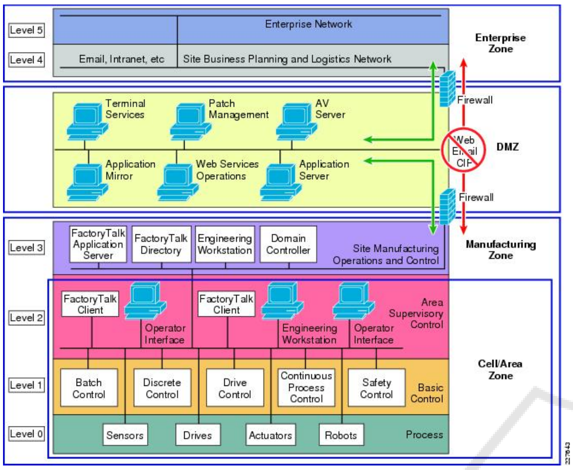

Figure 1: Purdue Model (PurdueModelCisco, 2018).

Data on the enterprise level is well structured, has a high data quality, and the number of different data storages is relatively small, as different types of systems exist only one time per plant/site. Accordingly, the data storages are very large, since they also contain a predefined history of data.

In the manufacturing zone, the number of data sources increases significantly, due to a higher variety of systems and a division of areas into subareas, each with identical but independent control systems. Shop control systems (site manufacturing operations and control), e.g. receive data from programmable logic controllers (PLC) and/or workstations as part of an Industrial PC (supervisory control & basic control), which in turn receive data from production equipment (process). As each system controls only a limited number of subordinate controllers, many identical systems are in use covering separated areas.

The break-up of the Purdue structure in the course of IoT developments results in increasing direct system connections on device level. Thus, shop floor devices become a direct data source for IBD.

The variety of different formats, unlabeled data and inconsistent naming of the data often negatively influence the data quality in the manufacturing zone. In addition, sensor data contains the risk of faulty data values being sent – e.g. assets or a sensors fail.

Manufacturing data is generated (if applicable) based on (fixed) cycle times (log-files) and sensor transmission frequencies. Therefore, data generation frequencies of milliseconds *ms* are a common phenomenon. Accordingly, the velocity of data generation, especially on the device level increases with the number of devices and processes exponentially. Additionally, data storage in production equipment itself is usually limited and often consists only of a circular buffer, so that data is stored transient according to the FIFO principal (First-In-First-Out). These points of data form a continuous flow of new and updated data points, a data stream to be handled by IBD systems.

4.2 Acquisition and Preparation

OPC UA. OPC Unified Architecture (Rinaldi, 2013) is a general machine-to-machine protocol and that also incorporates full description of data using a meta-modeling pattern for information models. Thereby, OPC UA models annotate the data in terms of its format, valid value ranges, semantics and context information. OPC UA allows modeling the complete production process and thereby preparing data for an exploration of possible correlations. The OPC UA standard defines base objects, i.e. blue prints, e.g. for sensors, devices or machines and provide these models through an OPC UA Server. Vendors are accordingly able to extend these basic information models with their domain specific object and type definition. A client accordingly connects to the server, which typically resides in each control system and contains its virtual representation. The client is able to browse and read all available nodes in that server. The client can either access data directly or register a trigger to let the server notify it on changes, either periodically or on defined value changes and other events.

MQTT

Message Queue Telemetry Transport (MQTT) is an open-source message protocol originated in the smart home automation, but is due to its simplicity also widely adopted into the industrial sector for machine communication. It is lightweight and enables publish subscribe to obtain and collect specific information through a flexible hierarchical structure, represented by topics. The hierarchical topics are created inside the queue tree of a central broker instance. The information modeling of the production process depends on these topics, describing e.g. the location of a device as well as on the message structure of the MQTT message object. The message object is represented by JavaScript Object Notation (JSON) and defines the payload of the message, which can be freely defined according to application or use-case specific requirements. Publishing clients are able define queues without former registration, thus there is no controlling instance at the server. A client that wishes to receive data has to know these specific queues beforehand to successfully receive data based on subscriptions to these topics. To organize information exchange about available topics, queues and the data representation, MQTT requires additional managing of used naming conventions and message structures.

IO-Link

Another communication standard for sensors defined under the norm IEC 61131-9 (DIN EN 61131) is IO-Link or as defined in the standard SDCI. The standard specifies communication between one master and multiple slave devices over a PLC. It bridges the link between an actual field bus of sensors and the Industrial Ethernet, where Industrial PCs reside.

(Proprietary) Message Queue

Other proprietary formats may exists which rely on a message queue system and define its own message format standard, often also called *telegrams*. Widely used are XML based telegrams, which can be also exchanged using OPC UA. Different vendors favorite this traditional mechanism to empower data exchange from their machines, but often rely on their own proprietary format, which cannot be handled directly, see 4.2.2 Industrial Big Data Transformation.

RDBMS

Another class of source systems that provide already enriched data are reporting systems of KPIs. They mostly rely on relational database management systems (RDBMS) controllable and reachable using Structured Query Language (SQL). SQL is commonly used across all database system and allows for querying specific information. Inside the database, the structured data remains in large horizontal tables containing raw values with additional new calculated or combined information including context data.

4.2.1 Industrial Big Data Extraction

The data extraction depends on the protocol and to the sort of data that is to be extracted. From the Business Intelligence (BI) point-of-view, so-called facts and dimension are differentiated in this context. Facts represent an event in time that refers entities, and a dimension is mainly static information that defines the entity itself. On the shop floor system with machine data, the majority is fact data, which contain e.g. current sensor values. One value represents one fact. To extract facts from a machine equipped with a protocol, a subscription of the client to receive the desired information is required However, to further understand and contents of the acquired data and finally extract its value, meta-data and about sensors and machines are needed. This context information is represented by dimensions. Extracting dimension data can be a complex task as context information is often not made explicit in the source system and therefore only exist implicitly due to some name or an underlying address space. To acquire this context information domain-specific and experience-based knowledge from human beings is typically needed. RDBMS systems commonly provide different mechanism to extract data using SQL. It is possible to poll for new data using a unique identifier, to define handover tables, hat only includes changed/new data points and it is possible to make use of full-fledged Change-Data-Capture systems, which completely mirror the whole database to identify changes.

4.2.2 Industrial Big Data Transformation

Data transformation touches the modification of the data format itself, thus its representation, but also the cleaning and validation of invalid information, if applicable. Especially, when the data format is proprietary and does not follow open standards for data exchange, the actual information needs to be transformed into a readable format. Furthermore, annotation and meta-data enrichment are also a part of data transformation, where all available information has to be added explicitly to the data.

4.2.3 Industrial Big Data Load/Integration

The load process partially depends on the target system, which has to retrieve or possibly store the data at hand. If the system is not able to store the information as is, which is typically the case, integration steps are required in terms of a translation of the data format into the target systems, i.e. the data storages requirements. This remodeling of data format is often associated with a manipulation of the data schema, as it is part of an Extract-Transform-Load (ETL) process. An ETL pipeline defines this behavior for a Data Warehouse system, which integrates the data into a data mart specific to the domain context of the data.

4.3 Storage and Access

4.3.1 Industrial Big Data Storage

Big Data storage systems have to be able to scale up with massive amounts of data and represent them accordingly. There are two major principles to access data. These principles also affect the storage structure as well as the complexity of the integration process: schema-on-write and schema-on-read. As the name indicates the distinction between these two concepts lies in the way data schemas are generated respectively accessed. While the schema-on-write pattern requires a fixed data schema in which the data is transferred, the schema-on-read methodology redefines a new schema every time required information is read with a specific purpose. The schema-on-read pattern explicitly shifts the homogenization of multiple different data points, with a different schema, to step of information retrieval for analytics.

One example for the schema-on-write principal is a traditional (enterprise) Data Warehouse, which uses a star-schema to represent fact and dimension information as their respective tables. In general, it is also a RDBMS applied with special notation and mechanism for larger data sets, such as hot and cold storage. These different storages inside the warehouse are represented by data marts.

The schema-on-read principle is commonly used in the Data Lake (Ignacio et al., 2015) architecture. The idea is to store data as-is and thereby with the schema, it was generated in. The user that accesses the data is required to define a schema ”on-read” that fits to all of the desired data on reading. Apache Hadoop (Shvachko et al., 2010) is an open-source distributed file storage system on commodity hardware that enables this behavior for big data. Its advantage is the utilization of existing hardware not only storage wise, but with the added Map-Reduce principle also for the computing resources.

4.3.2 Exploration

Getting around the vast amount of data points and finding exactly the relevant pieces is one major goal of data exploration. Naive approach would demand to look at each individual point, but indexes allow directly accessing a specific element, based on properties for which the index has to be created first. However, typical database indexes are only possible in a fixed data schema, such as it would exist in a Data warehouse system. In the flexible and dynamic changing data lake environment, this is not possible in the classical way, as a schema could change according to each data point, not yielding a common ground. In this case, a reading schema that fits all data points has to be determined in advance and applied to an automatic query mechanism that abstracts this application. This can be done automatically with tools like Spark, more particular SparkQL, or with Apache Hive. Both enable the use of a familiar Querying Language such as SQL to work with the data in a set schema.

4.4 Industrial Big Data Processing and Analysis

4.4.1 Visual Analytics

One of the most powerful and widely used methods for gaining knowledge from data is visual analysis. By plotting data, it is often already possible to identify anomalies in data sets. Since production data is often generated by cyclically operating systems, the data values should generally be stable along the production of identical products. Accordingly, anomalies often indicate quality or machine condition problems. In addition to pure visual analytics, thresholds can be used to automate anomaly detection.

4.4.2 Statistical Analysis

Statistical analysis of manufacturing data (the application of statistical theory) has been part of the standard repertoire in manufacturing since the advent of Six Sigma and Lean Manufacturing. These techniques allow to process and analyze larger-scaled data, as it is possible with visual analytics. Usually statistical methods condense datasets to key figures such as process stability (CPK) or equipment utilization (OEE), but they are also applied to structure and reduce complexity in the event of problems/optimization. Percentage, box-plot or distribution (average, median and mode) belong to this group.

4.4.3 Machine Learning

Machine learning (ML) is the next stage of data analysis. ML allows to discover hidden patterns in huge amounts of data independently by an algorithms. Three main categories of ML are distinguished:

Supervised Learning (SL) targets an approximation of mapping function to predict outputs based on input variables. Techniques are regression, decision trees or artificial neural networks. As in manufacturing use cases output and input variables are known, this category of algorithms appears to be applicable.

Unsupervised Learning (UL) does not require output variables. Therefore, unsupervised learning does not target to predict certain behaviors but to detect hidden patterns/clusters in unlabeled data sets. Principal component analysis (PCA) and Clustering are the most common algorithm categories of this group.

Reinforcement Learning (RL) algorithms are the most recent members of machine learning procedures. They target to automatically determine ideal behaviors in a specific situation and context. These algorithms gain knowledge about the subject behavior by doing test and memorizing the results of experiments.

4.5 Information Presentation and Interfaces

Human information interfaces have been designed in various ways and combinations of techniques. In the context of manufacturing, the major techniques to provide data to an information consumer is visualization (e.g. monitoring and reporting). Information monitoring stands for the observation of predefined metrics/figures (partially deduce from raw data by analytics) and can be seen in manufacturing control rooms and/or by making use of Supervisory Control and Data Acquisition (SCADA) systems. Information reporting provides information in a higher condensed form than high-granular information monitoring. KPI reporting for example plays a major role for the management sector of an organization. Usually, reporting is used to manage processes and to report information for operational control. For both techniques, several key aspects need to be considered: self-service/adhoc or pre-definition, frequently updated information, point of time or temporarily summarized information and distribution online or via intranet.

5 CONCLUSION

In this paper, we presented an overview of the challenges and issues that emerge when dealing with Industrial Big Data that arises from the Industrial Internet of Things. We examined different solutions and reference models and summarized the common ground as well as vital approaches of the stages towards the Industrial data ingestion goal. The major key steps of a data pipeline have not changed from the classical Big Data approach, but new Industrial specific challenges occur. There is yet still no complete universal architecture for the Industrial Internet of Things, which is capable of fully addressing all issues that arise when digitizing a factory and aim to realize a full digital transformation. The architecture references only the general concept of how to deal with data on a high level, but lack important steps, such as meta-information about the data itself or on the other hand, the proposed solution is fitted tightly for a specific use case, which cannot be easily adopted in a general manner. We see requirements for further research in the area of data governance, especially in the parts of data lineage and data protection. In particular, laws such as the General Data Protection Regulation (GDPR) of the European Commission are taken into account in industrial applications when dealing with customer-specific products. Concerning machine maintenance and the idea of predictive maintenance, the question of data value is important and needs to be strongly addressed by the manufacturing sector in the close future. Another important topic addresses business models that might arise based on discussions around data ownership. An interesting question might be, if a machine provider automatically acquires machine data of a customer free of charge in exchange for successful predictions for e.g. predictive maintenance.

ACKNOWLEDGEMENTS

The authors would like to thank the German Research Foundation DFG for the kind support within the Cluster of Excellence Internet of Production (IoP). Project-ID: 390621612

REFERENCES

Bertino, E., Bernstein, P., Agrawal, D., Davidson, S., Dayal, U., Franklin, M., Gehrke, J., Haas, L., Halevy, A., Han, J., and Jadadish, H. (2011). Challenges and opportunities with big data. *Whitepaper*.

Brecher, C. and Özdemir, D., editors (2017). *Integrative Production Technology: Theory and Applications*. Springer International Publishing.

CEN-CENELEC-ETSI Smart Grid Coordination Group (2012). Smart grid reference architecture. *Whitepaper*.

Chen, C. P. and Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, 275:314–347.

Chen, M., Mao, S., Zhang, Y., and Leung, V. C. (2014). Big data: related technologies, challenges and future prospects.

Demchenko, Y., Ngo, C., and Membrey, P. (2013). Architecture framework and components for the big data ecosystem. *Journal of System and Network Engineering*, pages 1–31.

Ebbert, J. (2018). Define it – what is big data?

General electric (2012). The rise of industrial big data. *Whitepaper*.

Gluchowski, P., Gabriel, R., and Dittmar, C. (2007). *Management Support Systeme und Business Intelligence: Computergestuetzte Informationssysteme fuer Fach- und Fuehrungskraefte*. Springer-Verlag.

Hai, R., Geisler, S., and Quix, C. (2016). Constance: An intelligent data lake system. In *Proceedings of the 2016 International Conference on Management of Data*, pages 2097–2100. ACM.

Hammer, M., Hippe, M., Schmitz, C., Sellschopp, R., and Somers, K. (2016). The dirty little secret about digitally transforming operations.

Ignacio, T., Peter, S., Mary, R., and John E., C. (2015). Data wrangling: The challenging journey from the wild to the lake. *CIDR*.

Industrial Internet Consortium (2017). The industrial internet of things volum t3: Analytics framework. *Whitepaper*.

Intelligence, M. (2009). Business intelligence in manufacturing. *Whitepaper*.

Jeschke, S., Brecher, C., Meisen, T., Özdemir, D., and Eschert, T. (2017). Industrial internet of things and cyber manufacturing systems. In *Industrial Internet of Things*, pages 3–19. Springer.

Kirmse, A., Kraus, V., Hoffmann, M., and Meisen, T. (2018). An Architecture for Efficient Integration and Harmonization of Heterogeneous, Distributed Data Sources Enabling Big Data Analytics. In *Proceedings of the 20th International Conference on Enterprise Information Systems : March 21-24, 2018, in Funchal, Madeira, Portugal. - Volume 1*, pages 175–182. 20th International Conference on Enterprise Information Systems, Funchal (Portugal), 21 Mar 2018 - 24 Mar 2018, SCITEPRESS - Science and Technology Publications.

Kuschicke, F., Thiele, T., Meisen, T., and Jeschke, S. (2017). A data-based method for industrial big data project prioritization. In *Proceedings of the International Conference on Big Data and Internet of Thing*, pages 6–10. ACM.

Laney, D. (2011). 3d data management: Controlling data volume, velocity, and variety, application delivery strategies. *Whitepaper*.

Lee, J., Ardakani, H. D., Yang, S., and Bagheri, B. (2015). Industrial big data analytics and cyber-physical systems for future maintenance & service innovation. *Procedia CIRP*, 38:3–7.

McKinsey (2017). Is manufacturing 'cool' again. *Whitepaper*.

Moghaddam, M., Cadavid, M. N., Kenley, C. R., and Deshmukh, A. V. (2018). Reference architectures for smart manufacturing: A critical review. *Journal of Manufacturing Systems*, 49:215 – 225.

Mourtzis, D., Vlachou, E., and Milas, N. (2016). Industrial big data as a result of iot adoption in manufacturing. *Procedia CIRP*, 55:290–295.

O'Donovan, P., Leahy, K., Bruton, K., and O'Sullivan, D. (2015). An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities. *Journal of Big Data*, 2(1):25.

Pääkkönen, P. and Pakkala, D. (2015). Reference architecture and classification of technologies, products and services for big data systems. *Big Data Research*, 2(4):166–186.

PurdueModelCisco (2018). Purdue Model/CPwE Logical Framework.

Rinaldi, J. (2013). *OPC UA - the basics: An OPC UA overview for those who are not networking gurus*. Amazon, Great Britain.

Sadiku, M. N., Wang, Y., Cui, S., and Musa, S. M. (2017). Industrial internet of things. *IJASRE*, 3.

Shvachko, K., Kuang, H., Radia, S., and Chansler, R. (2010). The hadoop distributed file system. In *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, pages 1–10.

Society, F. (2017). Reference architecture model for the industrial data space. *Whitepaper*.

US Dept. of Commerce blog (2014). The internet's next big idea: Connecting people information and things.

VDI/VDE-Gesellschaft (2013). Cyber-physical systems: Chancen und nutzen. *Whitepaper*.

VDI/VDE Society Measurement and Automatic Control (GMA) (2015). Reference architecture model industrie 4.0 (RAMI4.0). *Whitepaper*.

Wilder-James, E. (2012). What is big data? an introduction to the big data landscape.

Williams, T. J. (1994). The purdue enterprise reference architecture. *Computers in industry*, 24(2-3):141–158.

Yin, S. and Kaynak, O. (2015). Big data for modern industry: challenges and trends [point of view]. *Proceedings of the IEEE*, 103(2):143–146.