An Architecture for Efficient Integration and Harmonization of Heterogeneous, Distributed Data Sources Enabling Big Data Analytics

Andreas Kirmse, Vadim Kraus, Max Hoffmann and Tobias Meisen
Cybernetics Lab IMA/ZLW & IfU in Aachen, RWTH Aachen University, Dennewartstr. 27, 52068 Aachen, Germany

Keywords: Big Data, Ingestion and Integration Pattern, Hadoop, OPC UA, Industry 4.0.

Abstract: We present a lightweight integration architecture as an enabler for the application of process optimization via Big Data analytics and machine learning in large scale, multi-site manufacturing companies by harmonizing heterogeneous data sources. The reference implementation of the architecture is entirely based on open-source software and makes use of message queuing techniques in combination with Big Data related storage and extraction technologies. The approach specifically targets challenges related to different network zones and security levels in enterprise information architectures and across divergent production sites.

1 INTRODUCTION

Since Internet of Things (IoT) and embedded devices have become an integral part of the manufacturing, an increasing number of data sources are part of modern production processes. This is true, especially for large companies that act on a global scale with production sites across the world. Such companies face major challenges when coping with the complexity of hundreds of different data types and sources, in particular if the application and usage of Big Data analysis promises high optimization potentials that can be exploited.

In order to make use of Big Data related technologies and to establish continuous analytics workflows that operate on heterogeneous data sources, the first major step is to make information accessible. In this paper, we present an architectural solution for a lightweight modular integration system based on message queuing concepts that synchronize heterogeneous data sources semantically in order to facilitate a generic and uniform access. This approach especially targets information and communication technology (ICT) infrastructures that are characterized by high security levels of the respective network zones, e.g., by taking into account shop floor systems and the encapsulation of different production network zones. We are able to incorporate legacy information systems such as traditional relational databases (RDBMS); network attached storages (NAS) as well as live streaming data of manufacturing equipment via interface standards such as OPC Unified Architecture (OPC UA) that are able to incorporate legacy and proprietary bus systems and insular solutions from the field level. The goal is not only to provide ease of access, but also to incorporate data security by privacy means as well as access control rights. A prototypical implementation of the architecture was evaluated in terms of large companies for multiple months.

2 PROBLEM DESCRIPTION

Current challenges within industrial production are not the lack of available information, but having the appropriate information accessible when and where needed. Separated production sites use distinct systems for data storage and transfer. The high costs for new machines and their ramp-up time impede long life cycles in a factory and thus the slow adoption of modern technologies increase the variety of (legacy-) systems, devices and communications protocols being used. Moreover, the complexity of production processes are increasing and heterogeneous types of machines, constructions, robots and sub manufacturing routines are involved. The success of each step, and therefore each machine, in the process chain is of equal significance to the whole process. For example, a typical production operation is the assembly of separate individually assembled sub-parts into a product, where a failure in the pre-build part can have an effect on the process of combining the parts together. When optimizing the entire process, it

Kirmse, A., Kraus, V., Hoffmann, M. and Meisen, T.
An Architecture for Efficient Integration and Harmonization of Heterogeneous, Distributed Data Sources Enabling Big Data Analytics.
DOI: 10.5220/0006776701750182
In *Proceedings of the 20th International Conference on Enterprise Information Systems (ICEIS 2018)*, pages 175-182
ISBN: 978-989-758-298-1

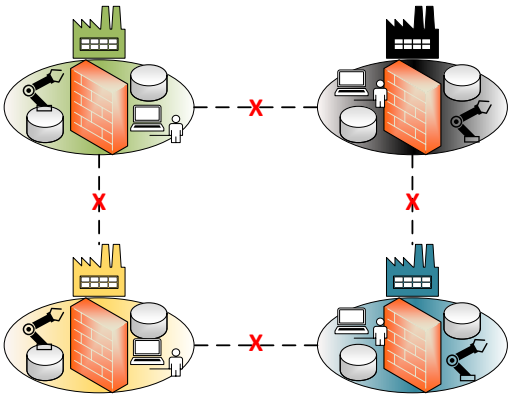

Figure 1: Scenario of four factories and their two separate network zones, production and office.

is thereby important to not limit the view to an individual part of its manufacturing routine but it is imperative to consider the complete picture. The data from every part of the chain could be of relevance. Therefore, the first challenge of optimizing production processes via Big Data Analytics lies within information consolidation and harmonization efforts across the entire process chain. In a distributed context this challenge increases significantly with the number of locations.

Each geographically distributed factory operates on its own and although they are similar and often produce the same general product, there are variances and differences in each dedicated process. Not only do they produce other parts or assembly another kind of product, the machines are naturally of different age and wear levels. Furthermore, regional factors such as working hours, legal requirements or the stability of the electric network have to be regarded. Additionally, timing optimization of the production have to deal with different time zones and the information of location is vital.

Therefore, it is necessary to be able to establish a clear distinction between several manufacturing resources and apply required harmonization and synchronization to the relevant parts, as required.

Figure 1 shows an example scenario for a company with four factories, each side has two separates network zones (e.g., production and office), which is a typical setup for an enterprise to secure and separate the production shop floor from business office computers. This separation is usually established due to the fact that productive systems should never be affected by office local area networks, e.g., in terms of networking traffic or possible attacks.

However, despite the security aspect, these restrictions oppose the endeavors of an Industry 4.0 which demands a full and seamless interconnection of machines and information systems along the manufacturing hierarchy and represent the another major challenge. Current standards, such as ISO/IEC 27002 (Technology, 2013) or (Calder, 2009), are therefore suggesting the use of technical barriers like firewalls and audited access control.

Finally, another challenge lies in the interoperability and reproducibility of analytics use cases. Currently, for each factory use cases yield insular solutions with their own specifics, not connected and not interchanging vital information about production with their similar counterparts. Across factories, these use cases are applied and similar problems often undergo the same methodology, the same steps to acquire data and to consider optimization goals to the underlying processes. The establishment of a shared data pool, available to all enterprise locations, can save tremendous efforts.

Current Situation. In terms of former modernization approaches within production and manufacturing, the goal was primarily to establish distinctive automation for most of the processes and to interconnect devices in order to perform selective analytics tasks for the optimization of dedicated processes. This traditional approach follows a hierarchical organization of the production and is manifested in terms of the automation pyramid as shown in Figure 2. Diverse systems interact on different layers with each other. The business levels on the top of the pyramid are in charge of the production organization and the bottom is characterized by the shop floor (machines, sensors and devices). This field level, represents the concrete controlling of the shop floor devices and is thus dependent on real-time information to seamlessly interact with each other and to react on events within a determined time range. Going up each level, the timeliness gets less mandatory and the velocity as well as the granularity of the incoming data reduces. One of the major goals with regard to the topics of the umbrella term Industry 4.0 is to build so-called cyber-physical system that reflect the physical processes on the shop floor and to service a digital representation of the underlying processes in terms of a digital model.

In order to achieve the desired information representation on all levels of the production, seamless data integration and information management are crucial goals to obtain. In current manufacturing systems, each level is usually characterized by specific island solutions that provide parts of the information needed to optimize not only automation itself, but give a deeper insight to the production process at the specific steps. These solutions are based on the use of data warehouses (DWH) or alike, where structured data tailored to one problem category is stored. The data integration is carried out by data processing steps, cal-

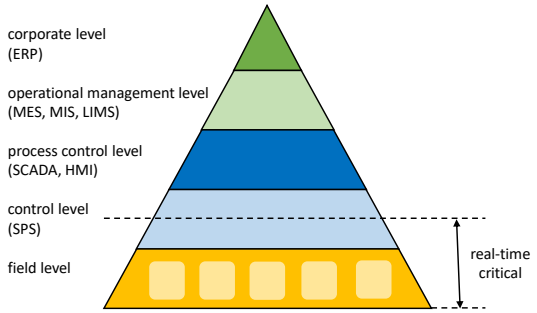

Figure 2: Automation pyramid showing the different layers of automated systems in manufacturing and the real-time critical components.

led Extract-Transform-Load (ETL) mechanisms, that apply a fixed and given star-schema to bring the information to the representation form required. Typical applications in such scenarios include the provision of reporting dashboards which generate insights about the last cycle(s) of production for a batch or for continuous monitoring, e.g., for alarm and event purposes. The described processing is useful for continuous determination of Key Performance Indicators (KPI) or daily reports. However, when it comes to gaining deeper and more holistic insights to the underlying processes, approaches such as Artificial Intelligence and Machine Learning (ML) have to be considered. These procedures, operated on chunks of data, intend to identify patterns within the data even among different sorts of information and across multiple processes.

For those, a traditional data warehousing approach to data collection and preparation is not well equipped and even lacks the required flexibility and diversity of data representation to train a (decision) model. Scaling up a data warehouse to hold years of archived data gets expensive rather quickly, as specific hardware to operate is required. Additionally, the DWH philosophy considers data that is rarely used as cold and will ultimately discard it, in order to keep access speeds high. In a production process, which is already working in a profitable range, there are almost no defect products, which need further optimizations. This leads to a data basis that is biased by design and lacks the objectivity of correct data analysis. Thus, limiting the storage to hold only faulty product process information might bias any machine learning approach of finding possible causation patterns.

Data Integration and Big Data itself is not a new topic, the concepts behind the term were already examined for decades in terms of information management (Hohpe and Woolf, 2003; Goodhue et al., 1992; Cox and Ellsworth, 1997). But, according to the described limitations, more sophisticated solutions for storing large amounts of highly heterogeneous data are needed. A first solution that fits to the needs of such storage capabilities, are Big Data based solutions, which are designed to scale very well with large amounts of fast incoming, heterogeneous data.

Having a look in the future of fully connected factories in Industry 4.0, business models like a company or even a global marketplace for optimizations or distributed knowledge generation could be realizable. In these scenarios, information would be available to stakeholders which potentially exceed the boundaries of one company. Machine manufactures could, for example, provide special services, such as predictive maintenance information in order to prevent machine malfunctions, as an incentive for manufacturing companies to share information generated on their premise.

3 RELATED WORK

Different approaches of how data aggregation that target the described challenges of heterogeneous and distributed information within production environments are suggested. We present similar, related work first before having a look into relevant technologies and protocols.

(Ball et al., 2017; Runge et al., 2016) focus on product quality assurance with the Open Manufacturing Information System (OMIS) approach. Their motivation bases on the one-off production in the small satellite manufacturing of Raytheon, where each product has its own assembly and diverse quality parameters, with the special case of Quality-in-depth (QiD). This quality method describes the over watch of manual quality inspectors, which watch the technicians. Therefore, OMIS combines existing enterprise systems with digital product assurance results (video and sensors) and claims to reduce costs by using COTS devices (commercial off-the-shelf) not particularly tested for space environment use.

(Theorin et al., 2015) proclaims the Line Information System Architecture (LISA) that uses the idea of an Enterprise Service Bus (ESB) to reduce point-to-point connections in a traditional client/server approach by making use of service mediation techniques. They claim to have made the service oriented architecture principle of ESB together with an event-driven bus system industrially applicable and scalable based on ActiveMQ. LISA uses an own message format to in-cooperate source systems and thereby solve the homogenization aspect.

(Bonci et al., 2016) show a database-centric approach based on cyber-physical production systems.

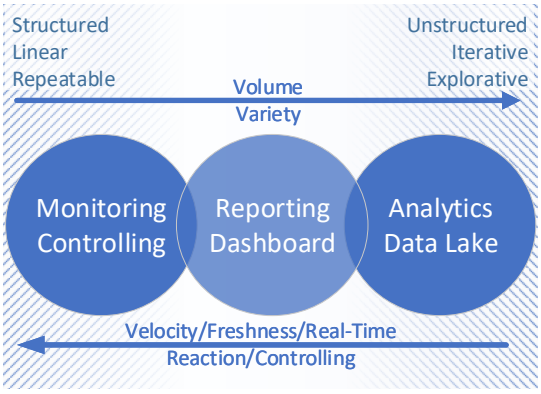

Figure 3: Placement of the Data Lake concept in an existing landscape of monitoring controllers and reporting dashboards for a production line in manufacturing.

The idea to use RDBMS along with the SQL query language is quiet established; their novel approach however focuses on lightweight database synchronization through distributed replication on every CPS device. Furthermore, they add the swarmlet concept by facilitating the publish/subscribe paradigm for IoT devices and add a plug-in structure, which extends the central database to a service-oriented architecture similar to an Enterprise Service Bus.

Related Technology. The presented related work contains architectural concepts in order to handle data load and data aggregation, but there are no technological solutions available that are directly ready to use.

Big Data storage techniques are the key aspect of managing the data load of Industry 4.0 related applications. The techniques require either specialized hardware or database instances capable of holding large amounts of data, commonly resulting in a Data Warehouse. Apache on the other hand – targeting NoSQL information storage approaches – offers a scalable data storage on commodity hardware for large data sets, called the Hadoop Distributed File System (HDFS) (Shvachko et al., 2010). Through the expandable architectural approach it is possible to provide a large-scale file storage solution across different systems at affordable costs.

Confluent provides Kafka Stream (Kleppmann and Kreps, 2015) as an idea and concept for handling real-time streaming data seamlessly and in a scalable manner. Confluent makes use of Apache Kafka, which main philosophy of log processing differs from general message queuing due to the lack of an acknowledgement (ACK) of messages and thus does not guarantee delivery. The underlying concept uses a consumer centric approach in contrast to broker or server/data centric protocols like AMQP (Vinoski, 2006) and MQTT. For server centric mechanisms, the control and thus responsibility fully lies in the master server, each client thus has to give a receipt for receiving and thereby handling the message. In contrast, the consumer centric approach of Kafka leaves all control to the receiver, where the different partitions of a broker hold numerous messages.

On the shop floor, a shift to new technologies is currently taking place, which helps to better understand and also handle data from further automated production plants. OPC Unified Architecture (OPC UA) and the Dynamic Distribution Service (DDS) are two advanced protocols that allow for easier and more holistic access to machine data. OPC UA is a machine-to-machine communication protocol specified in 2006 and advanced continuously since then that models automation hardware in the form of objects of a process work flow (Rinaldi, 2013). UA is an enhancement on the classical OPC, where the machine itself has the ability to describe its available information and enhance the raw data with meta-information. OPC UA lifts the low-level machine control on the PLC level to a higher application layer with added security and advanced capabilities to add semantics to the production data generated in the field.

More and more production machines natively run an embedded OPC UA server, which provides a base model including the process parameters and annotated information available at the specific machine.

4 CONCEPT

The concept proposed in this work focuses on data accessibility and does not change any of the incoming information. The goal is to have the raw information at hand for easy access. Therefore, our ingestion methodology differs from commonly used classical enterprise integration patterns, where semantical data transformation in terms of the data structure always takes place. The kinds of source systems retain their semantical structure and form the (Big) Data Lake (O'Leary, 2014). The general idea thus is to have a long-term storage of available data that is required to perform analysis processes such as pattern recognition and machine learning methods in a holistic manner. Artificial Intelligence methods require large sets of training data to learn. The data sets should be universally available by providing a managed data pool that allows access for users or autonomous optimization applications. The information requirements differ for each specific use case, a definitive access to all available information is thus more important than the delivery of specific information

that is already tailored to some purpose. Through the general overview, the analysis is more likely to find hidden and so far unexplored relations in the data.

The source of data covers the full range of systems located on the automation pyramid, from the real-time systems on the shop floor with streaming data up to annual reports form data warehousing systems that are not only low in frequency, but also low in volume and higher in density of information. Figure 3 shows the conceptual placement of a Data Lake component inside the existing landscape of the automation pyramid. The attributes of volume and variety, building two of the three "V"s in Big Data, are displayed on upper horizontal axis and increment from left to right. Thereby, the Data Lake concept is predestined for data of high variety, whereas e.g., a controlling instance directly at a machine heavily relies on very specific data. On the bottom axis, the velocity and actuality of data are depicted in declining order, indicating that a Data Lake by design is not reactive enough for real-time controlling.

The main goal of the described concept is to enable data analysis applications on a consolidated, long-term storage, therefore, the velocity aspect is negligible. For a live application of, e.g., trained machine learning models, however, we suggest an additional layer. This layer would follow the concept of a lambda architecture, where the data flow is split into two different streams. In this case, one stream is used to redirect the incoming data into long-term storages, whereas the other stream is processed in a timelier fashion, responding to potential requests to a machine learning model.

In the production machine-to-machine context, inconsistent network conditions, e.g. varying response times in the range of milliseconds exceed crucial time constraints for remote control commands. Therefore, another alternative model application scenario consists in the deployment of machine-learning models directly at the machine on the shop floor. The training of the models deployed on these edge computing devices, however, should be performed by analytics systems that are directly connected to the Data Lake, e.g. in the form of a server farm. This way, the model can be trained with a combination of all available data to determine optimal parameters for the processes in the scope.

As a consequence, through the described process the data analyst is able to perform model extraction of trained data from machine learning processes and to re-integrate new insights into the production process. Thus, although the Data Lake analysis on the big data aggregation is not able to directly control the machines due to the high latency, it is the enabler of improved control units in short or medium-range time cycles, which can provide real-time reactions.

4.1 Data Acquisition

As we focus on applicable data integration in the production context, we have to consider how companies in this area traditionally operate. Therefore, it is not possible to assume a factory as a green field and to assume the theoretically best-case scenario. When a factory is built, its operations are planned over decades, the processes to be performed in the production environments evolve over years. Due to the high investment costs and amortization time range an exchange of manufacturing machines is usually not an option. Consequently, the emphasis of this work focuses on a modular approach that is especially prepared to integrate legacy systems from the brown-field without the need for an adaptation of internal structures and the entire underlying legacy systems. For this purpose, each module encapsulates one category of source systems taking into account their individual challenges and individual demands. Based on a homogenization of the underlying data structure, the modular concepts is able to represent the extracted information using a generic messaging format that describes the payload accurately and without using semantic meta-data that is induced from the source system. This modular approach also allows for an extensibility in regards for future technological advances. In the following paragraphs, we will present some of the most common resources of production information that can be found in the industrial manufacturing context.

Management Systems. Various management systems are deployed at different levels of the company, a typical storage concept of these systems are databases or more advanced data warehouses for reporting purposes. (Relational) Database management systems (RDBMS), therefore, play an important role in enterprise information systems, resource planning as well as for manufacturing execution, consisting of the monitoring and reporting of the production process. The storage engine holds all relevant information of the production process ranging from different failure states or key performance indicators. This information stored in this database is characterized by its own database schema defining data types and providing additional context information to the represented values.

One key challenge is the variance in Structured Query Language (SQL) dialects, different vendors of RDBMS do not fully comply with the SQL standard. The ISO/IEC 9075 (Technology, 2008) norm defines

the bare SQL standard in its current latest version from 2016, where each vendor is free to extend it, however different database engines are even characterized by differences in the core syntax. Moreover, database structures differs per database product or system that incorporates a database itself. Often, it is not obvious that a pre-built and functional production system already comes with a defined database system of the manufacturer. Especially, in production systems equipped with a default database system, it is not possible to perform changes without risking an overall system breakdown or at least violating existing service level agreements. Therefore, the database connectors need to take into account existing legacy devices and their underlying information systems instead of relying on the requirements of a green field.

Shop Floor. Another viable information source for manufacturing enterprises consists of data from the production field. This raw machine data of sensor or actuator states is highly structured and usually characterized by high frequencies as the devices are directly based on the shop floor. The structure is based on the correlation to the actual sensor value of the physical device and thereby defined by a native register value in a state machine. Different machine vendors sometimes use proprietary native communication and control protocols, which are usable only in terms of their licensed counterparts, e.g., control units. OPC UA is one uprising protocol trying to harmonize communication protocols and their respective information flows from the shop floor. The Unified Architecture (UA) protocol extension defines an industry standard information model, which can be extended by means of vendor specific information models as needed. Hereby, the vendor of the controller fully describes the information available in the system and how they can be described for further usage. Machine manufacturers already adopt OPC UA connectors directly into their systems and controllers as machine operators demand integration of these protocols for interconnecting their future factory environments. This pseudo standardization constitutes already an improvement in comparison to directly accessing registers on a PLC without proper knowledge of the flag information. One further challenge of the described methodology consists in the high number of different vendors for control units, each providing a different information model that still has to be combined and semantically harmonized in order to describe the manufacturing process itself and tailored to the required tasks. In this sense, the heterogeneity of the shop floor is redirected to higher semantic levels, in which the harmonization takes place and the data integration is finally handled.

4.2 Data Ingestion - Message Format

In this section, we describe the implementation of a general-purpose payload format with an additional description of meta-information to make context information explicitly available after data integration. The additional meta-information includes a precise source description including the time and place in which the data first occurred, hence the time stamp from a specific shop floor device. Furthermore, the meta data describes factory localization which are needed to account for different time zones and possible synchronization timestamps as well as language differences. However, the payload of the transferred information is not adapted during transit, hence the data is simply passed on directly to the persistent storage without prior modification. Thus, in terms of the integration step, all additional semantic information from the source system is kept in their original form. In order to decouple systems and separate information security layers, all messages are buffered between gateways of two network zones in terms of a message queue. This proceeding additionally attempts to secure shop floor systems, which will not be influenced, if data is stuck on the intermediate storage and piles up.

4.3 Data Storage

Subsequently to the process of data acquisition based on the raw data of a source system, enhanced with meta-information and in a common generalized payload-driven format, the data is enabled for integration without concrete applying of a semantic transformation. Each source system of raw data provides a schema (if available and applicable), which is also used for the storage within a HDFS file. As schemata might evolve over time, different versions have to be considered in order to enable structural changes over time.

By an application of the schema-on-read principle an analyst always accesses the respective information in the desired fashion.

This principle defines the creation of a schema when reading the data and thereby enabling the combination of different schemata into one coherent description.

Drawbacks of the requirements that are connected to the creation of such schema can be neglected, because the analyst should always be able to use the schema, in which the data was written originally and thus created. Additionally, the shift from static schema descriptions to a dynamic creation of schemata allows for much easier integration of source sy-

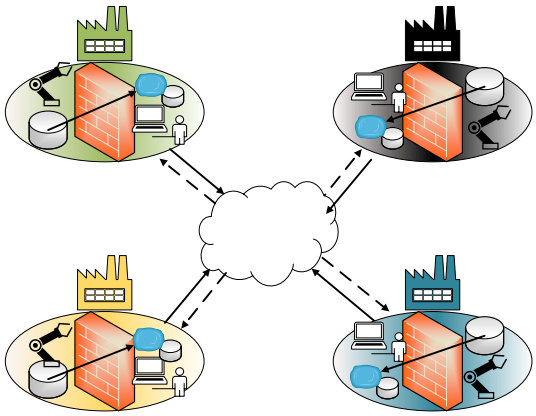

Figure 4: Aggregating and combining factories with data lakes into a (virtual) data cloud.

stem as no modeling or prior transformation to a predefined schema has to take place, e.g., as it is necessary prior integration into a Data Warehouse.

Figure 4 shows the conceptual placement of this ingestion process into Data Lakes that are located at each factory. Due to the nature of accessibility to the information by a data scientist, the data archive is not located in the production network itself, but in the neutral business network zone with restricted access. The concept furthermore enables the possibility to combine all of the production plants data lake structures into one virtual data cloud covering the entire company and exchanging insights gathered from each of these factories.

5 PROTOTYPE ARCHITECTURE

In this section, we describe the prototypical implementation of the architecture for an efficient integration and harmonization of heterogeneous distributed data sources. Hadoop (Shvachko et al., 2010) and other open-source solutions from the Apache product family provide the basic foundation for the Data Lake storage concept. These technologies enable scalable data storage based on commodity hardware and deliver the use of available computing power via the Map Reduce paradigm. HDFS combined with the scalable computing paradigm facilitates the invested hardware and allows for Big Data analytics (Dean and Ghemawat, 2008).

RabbitMQ (Videla and Williams, 2012), as the default implementation of the AMQP protocol (Vinoski, 2006), delivers a fast, robust and scalable broker system for message exchange and thereby enables the network buffering and separation of critical production systems. It is designed to not trust the underlying network infrastructure and therefore also fits the use case of the neutral business network, in which nobody is able intercept the data ingestion process. The specialized and generic message format previously mentioned is a direct adoption of AMQP messages. A message consists of a generic body for arbitrary data types without further additions coupled with a header definition comprised of key-value fields. Thereby, AMQP is a lightweight message protocol with efficient routing algorithms that assure delivery with package acknowledgements and act as the concentrator of the distributed data sources in a network zone. Spring Integration (SI) (Fisher et al., 2012) as the Java implementation for Enterprise Integration Patterns (EIP) and for message-based data integration provides a configurable pipeline approach allowing for a modular composition via reusable Spring Bean components. The entire integration process is split into different instances consisting of different pipelines for different systems. One instance alone is not required to map and process the entire integration pipeline, thus the approach proposed in this work allows for dynamic scalability of different parts of the pipeline. This dynamic scalability further demonstrates the efficiency of the proposed concept.

The Data acquisition part in terms of access to databases is achieved via the native Java Database Connectivity (JDBC) interface, that already contains the required drivers for interconnection of information resources. In a similar fashion, another modules can be implemented allowing the access to data source such as streaming data via OPC UA or legacy system protocols. Another step in the pipeline consists in the processing after completed data acquisition. The retrieved data is stored in the message queue and then enriched during the integration step into the HDFS. This step includes a determination and application of a schema onto the data along with other meta-information to be used in the message formats header fields by another Spring Integration pipeline residing connected to the Hadoop system.

The storage format of integrated information can be further facilitated by making use of Apache Avro (Apache Software Foundation, 2009), which does not only support native serialization and deserialization of arbitrary data, but also fully enables the schema-on-read principle. Data written with one schema in Avro and stored using the same format can be read with a different schema. Another advantage of the Avro format is the ability of Avro to compress information. In comparison to the JSON format, which stores the same amount of information as Avro, the Avro file format approximately consumes about 15 % of storage space.

Another member of the Apache family is Hive (Thusoo et al., 2009), which provides database like view on data in HDFS. It uses so-called SerDe modules to serialize and deserialize data for querying via SQL. SQL is a well-established query language for relational structured data in familiar tools. Due to the fact that Hive follows the SQL standard JDBC and ODBC Database drivers are available to connect with familiar frontend tools in order to analyze and work with the data. Such tools include for example Tableau for visualization and Matlab for data exploration. Using the schema-on-read principle in combination with consolidated storages a harmonized access to the data can be realized. All data can be accessed via well-defined interfaces, allowing for a fast and holistic analysis.

6 CONCLUSION

In this work, we presented an architectural framework and integration chain that enables Big Data analytics workflows for further application in the manufacturing domain. By making use of the proposed concepts, production environments with their special challenges and requirements can be appropriately mapped to an overall information management and data harmonization. One of the major inside of this work consists in the identification of lacking data accessibility in current factories. By increasing the general data accessibility through the proposed integration approach, it becomes possible to get deeper insides with regard to the production processes and to reveal patterns based on machine learning for correlations across different data sources. We shown how possible ingestion processes of these different sorts of data sets and streams into one coherent data pool can be realized by making use of established technologies. Further steps that are required for a more generic integration of shop floor devices into data lake structures can be identified especially in the field of OPC UA data and information modeling. Other topics to be further investigated are issues related to data governance and auditing. This is especially important for cases, in which privacy or customer/user data become relevant and are included into the data lake structure.

REFERENCES

Apache Software Foundation (2009). Apache Avro. https://avro.apache.org - last accessed Jan-2018.

Ball, G., Runge, C., Ramsey, R., and Barrett, N. (2017). Systems integration and verification in an advanced smart factory. In *2017 Annual IEEE International Systems Conference (SysCon)*, pages 1–5.

Bonci, A., Pirani, M., and Longhi, S. (2016). A database-centric approach for the modeling, simulation and control of cyber-physical systems in the factory of the future. *IFAC-PapersOnLine*, 49(12):249–254.

Calder, A. (2009). *Information Security Based on ISO 27001/ISO 27002: A Management Guide - Best Practice*. Van Haren Publishing.

Cox, M. and Ellsworth, D. (1997). Managing big data for scientific visualization. In *ACM Siggraph*, volume 97, pages 21–38.

Dean, J. and Ghemawat, S. (2008). Mapreduce: Simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113.

Fisher, M., Partner, J., Bogoevici, M., and Fuld, I. (2012). *Spring Integration in Action*. Manning Publications Co., Greenwich, CT, USA.

Goodhue, D. L., Wybo, M. D., and Kirsch, L. J. (1992). The impact of data integration on the costs and benefits of information systems. *MiS Quarterly*, pages 293–311.

Hohpe, G. and Woolf, B. (2003). *Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Kleppmann, M. and Kreps, J. (2015). Kafka, Samza and the Unix Philosophy of Distributed Data. *IEEE Data Eng. Bull.*, 38(4):4–14.

O'Leary, D. E. (2014). Embedding AI and Crowdsourcing in the Big Data Lake. *IEEE Intelligent Systems*, 29(5):70–73.

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

Runge, C., Lynch, K., Ramsey, R., and Pauline, T. (2016). Digital product assurance for model-based open manufacturing of small satellites. In *2016 IEEE International Conference on Emerging Technologies and Innovative Business Practices for the Transformation of Societies (EmergiTech)*, pages 206–209.

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.

Technology, I. (2008). ISO/IEC 9075 Database languages - SQL. Technical report, International Organization for Standardization.

Technology, I. (2013). ISO/IEC 27002:2013 - Information technology – Security techniques – Code of practice for information security management. Technical report, International Organization for Standardization.

Theorin, A., Bengtsson, K., Provost, J., Lieder, M., Johnsson, C., Lundholm, T., and Lennartson, B. (2015). An event-driven manufacturing information system architecture. *IFAC-PapersOnLine*, 48(3):547–554.

Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Anthony, S., Liu, H., Wyckoff, P., and Murthy, R. (2009). Hive: A warehousing solution over a map-reduce framework. *Proc. VLDB Endow.*, 2(2):1626–1629.

Videla, A. and Williams, J. J. (2012). *RabbitMQ in action: distributed messaging for everyone*. Manning.

Vinoski, S. (2006). Advanced message queuing protocol. *IEEE Internet Computing*, 10(6):87–89.