Industrial Cyberphysical Systems: Realizing Cloud-Based Big Data Infrastructures

Bo Cheng, Jingyi Zhang, Gerhard P. Hancke, Stamatis Karnouskos, and Armando Walter Colombo

***Abstract*—Future industrial systems and applications are expected to be complex constellations of cyberphysical systems (CPSs) where intelligent networked embedded devices play a pivotal role toward the realization of new sophisticated industrial scenarios. The prevalence of multifaceted devices enables new avenues for monitoring at large scale via Internet of Things (IoT) technologies, and, when coupled with the real-time analysis of massive amounts of data, it results in new insights that can enhance decision-making processes and provide a competitive business advantage. How to collect, process, analyze, and interpret big data is a challenge that affects all industries, and, if effectively addressed, it would offer numerous operational benefits. This article discusses some of the main architectural issues related to collecting and handling big data for analysis linked to IoT and cloud technologies in the industrial context. The aim is to provide a high-level introductory view of this topic, underpinned with examples from popular frameworks, and discuss open research questions and future directions.**

***Index Terms*—Cloud computing, Big Data, Industries, Companies, Protocols, Manufacturing, Real-time systems.**

I. Meeting the Challenges

With the swift penetration of IoT [1,2], technologies into a variety of industries, there is a potentially large increase in collected data that need to be processed. Effectively handling these big data is an opportunity to generate added value and provide a business advantage [3,4]. For example, the healthcare industry in the United States alone could create more than US$300 billion every year if big data are used creatively and effectively [3]. The effective management of the lifecycle of big data, including its collection, processing, and analysis for decision making and insight generation, is seen as challenging, and how effectively these aspects can be realized may impact the competitiveness and performance of companies in various industries.

IoT technology provides new opportunities to build powerful industrial systems by connecting a large number of smart networked embedded devices. Devices within these industrial IoT (IIoT) or industrial CPSs (ICPSs) can sense and control physical processes, make autonomous decisions, and communicate and cooperate, thereby collectively generating a massive amount of system data. Cloud computing [5] provides a promising solution for modern industrial systems and applications by hosting services with flexible computational and storage capacities. The integration of IoT and big data technologies into the cloud will empower industries to build a complex cloud-based CPS [6,7]. In some smart factories, for example, operators can track the entire production process, identify (often proactively) problem areas, and make informed decisions, e.g., dynamically rescheduling processes and maintaining systems on demand. The usage of cloud technologies is increasingly penetrating industrial settings, and the term industrial cloud refers to the cloud building an industrial CPS and applications with IoT and big data technologies [4,8] to distinguish it from the cloud for general purposes. To effectively handle these data, industrial companies need to consider upgrading current systems and technologies to cope with this additional data volume and, in doing so, will realize the benefits that big data handling delivers. This recommends the key capabilities and features that need to be developed to address various domainspecific requirement challenges in CPS deployment, e.g., real-time monitoring and control, flexible storage and timely generation of control and business decisions, and enhanced infrastructure management [6]–[9]. The effective handling and processing of big data is the first step in overcoming these challenges.

II. The Industrial Cloud and Big Data

The Industrial Cloud and Big Data Big data allow industrial companies and organizations to collaborate and create new value from data [3]. Interconnected industrial devices, e.g., smart meters and industrial equipment in the physical world, can sense and control processes, generate data, and communicate as part of the IoT. By capturing, processing, and analyzing significant amounts of data from these devices effectively, industrial companies and organizations can manage their enterprise resources, optimize technical processes, understand the market demand, and develop business intelligence and analytics (BI&A) [10].

Due to poor scalability and low performance, many traditional computing technologies are inadequate for handling big data, which are characterized by the volume, velocity, variety, and veracity of the data (each of these characteristics applies to ICPS data.) The volume of data will grow with the adoption of IIoT technology. The velocity, i.e., the rate at which data are generated, ingested, and processed, is crucial for decisions that feed back into the system to control real-time industrial processes. Since the ICPS consists of heterogeneous systems of systems, the variety of the data is also very high. The veracity (accuracy) of the data is also important in the cyberphysical context as incorrect decisions made from low-quality data could lead to physical disruption of industrial processes. Big data have become a critical factor of the ICPSs, and it is strongly suggested that industrial companies could create a competitive advantage and boost productivity by using cloud-based big data technologies [3]. Therefore, several

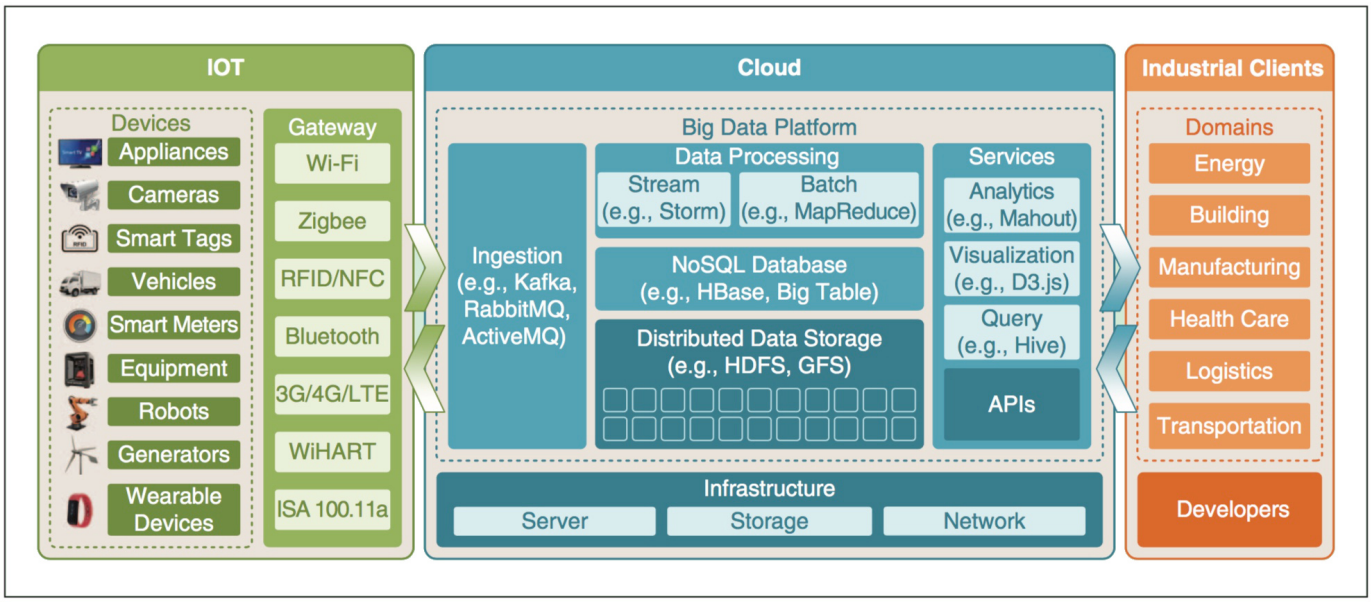


Fig. 1. The industrial CPS systems powered by big data. Smart devices could communicate with the cloud using protocols like AMQP. The big data platform hosted in the cloud enables efficient data ingestion, warehouse, processing, analytics, and visualization. Industrial clients from energy, building, manufacturing, health care, logistics and transportation domains can gain insight into the big data through the service interfaces provided by the cloud. RFID: radio-frequency identification; 3G: third generation, LTE: long-term evolution; ISA: International Society of Automation; AMQP: Advanced Message Queuing Protocol; GFS: Google File System; API: Application Programming Interface; NFC: near-field communication; WiHART: Wireless Highway Addressable Remote Transducer Protocol.

cloud-based technologies have been developed for handling big data by taking advantage of the cloud characteristics, e.g., distributed storage, dynamic computational scalability, and parallel computing [4]. Combined with IoT technology, a complex cloud-based ICPS can be built where industrial big data are collected, processed, analyzed, and stored, providing domain-specific services or software to industrial clients. The architecture of a cloud-based ICPS is shown in Figure 1, which provides a common framework that many participants can work with, by promoting innovations and accelerating the deployment of the cloud-based ICPSs and applications for future industries.

A. *Industry interest*

Many industrial companies have introduced industrial cloud platforms and services. GE is pitching an industrial cloud platform, named Predix [11], for industrial big data and analytics. Predix enables industrial-scale analytics for asset performance management and operation optimization by providing a standard way to connect machines, data, and people. IBM has developed IBM Cloud [12], which provides a platform-as-a-service (PaaS) for developers and data scientists to rapidly develop and efficiently manage their applications that take advantage of data and analytics from connected smart devices and sensors. Oracle Utilities [13] provides cloud-based software to the utility industry and their customers by analyzing smart meter readings, and uses statistical algorithms to help their customers save money through behavioral changes. Advantech, in collaboration with Microsoft, built a WISE-Cloud [14] platform that integrates IoT software and a cloud platform to provide services to industries, e.g., seamless sensor information collection, remote management of devices, and big data analytics. Several traditional cloud software companies like Google, Amazon, Microsoft, Cloudera, SAP, Oracle, and Salesforce.com have made efforts to integrate the IoT and big data into their services, which greatly promotes the development of the industrial cloud. Industrial alliances formed by companies and academia promote the extensive cooperation and standardization of the industrial cloud. The Industrial Internet Consortium (IIC) [15] was founded by AT&T, Cisco, GE, IBM, and Intel in 2014 and now has more than 200 members. The IIC was formed to accelerate the development, adoption, and widespread use of interconnected machines and devices and intelligent analytics. The Internet Protocol for Smart Objects (IPSO) Alliance [16] promotes IPv6 connected devices in energy, health care, and industrial applications. To secure industrial cloud-based implementations, the Cloud Security Alliance (CSA) IoT Working Group [17] has committed to defining actionable guidance for security practitioners.

B. *Government and Policies*

Aside from industry, government also plays an important role in speeding up the deployment of the industrial cloud by establishing active policies and providing financial support. Germany's Industry 4.0 [7] is a key strategic initiative aimed at strengthening the competitiveness of German industry. Empowered by the industrial cloud and the IoT, smart machines and production facilities can autonomously exchange information, trigger actions, and control each other independently, forming complex CPSs that pave the path to Industry 4.0. In China, the Made in China 2025 and Internet Plus strategies have made their debut, aspiring for a big leap

Preprint version for http://dx.doi.org/10.1109/MIE.2017.2788850

in innovation as well as manufacturing efficiency based on smart technology, the mobile Internet, cloud computing, big data, and the IoT. The U.S. government co-founded a national network for manufacturing innovation (NNMI) made up of several institutes for manufacturing innovation that provide a manufacturing research infrastructure where U.S. industry and academia collaborate to solve industry-relevant problems. Research areas of the NNMI include three-dimensional printing, big data, smart manufacturing, and medical devices. Although the industrial cloud is being greatly promoted by industrial companies and alliances as well as governments, it is still in its nascent stage of adoption.

III. Cloud-based Big Data Technologies

Several cloud-based technologies have been developed for handling big data by taking advantage of distributed storage and parallel computing in the cloud.

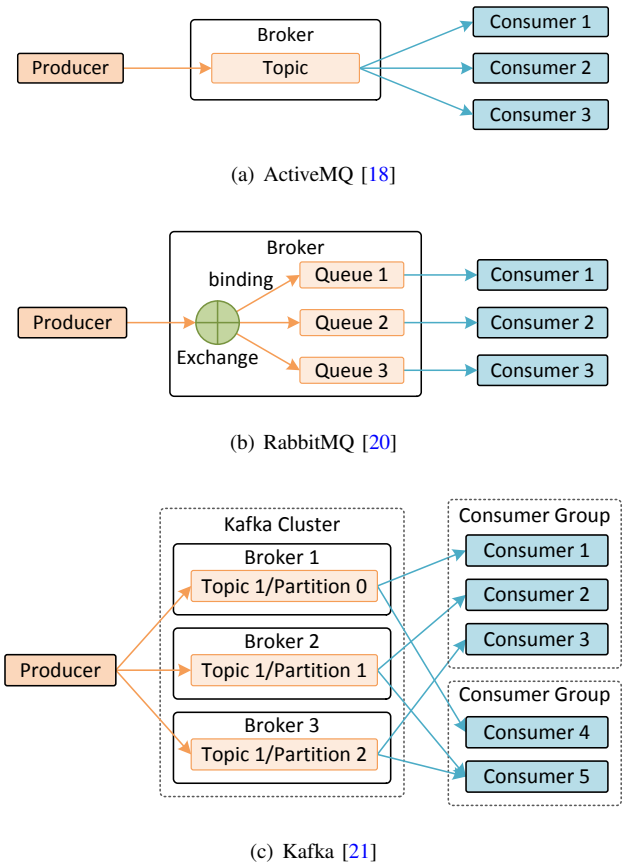
Several cloud-based technologies have been developed for handling big data by taking advantage of distributed storage and parallel computing in the cloud. In this section we discuss the core functions of big data processing – data ingestion, stream processing, scalable storage and batch processing – and discuss, as examples, how prominent existing frameworks implement these functions. To end we briefly discuss data analytics and visualization.

A. A High-Throughput and Fault Tolerance Data Ingestion

A huge amount of data from the IoT needs to be transmitted to the cloud effectively and in real time. In a conventional request/response messaging configuration, a message producer and a consumer are tightly coupled, and the performance of data transmission depends on both sides. If one side is slow, the transmission will be slow. Therefore, a request/response messaging paradigm may not be suitable for large-scale and high-throughput data ingestion. In contrast, publish/subscribe messaging provides a loosely coupled method for data transmission. In this messaging pattern, message producers, (e.g., IoT devices or gateways) publish characterized messages on one or more brokers, (i.e., servers) and message consumers, [e.g., Storm, Spark, Hadoop Distributed File System (HDFS), or a Hadoop database (HBase)] could subscribe to the messages they are interested in. Once the broker receives a message, it will deliver the message to those consumers who have subscribed to this message. Hence, a publish/subscribe messaging paradigm is potentially a good fit for scalable and distributed data ingestion in the cloud. Several publish/subscribe messaging brokers, e.g., Active Message Queuing (AMQ), RabbitMQ, and Kafka have been developed.

Apache ActiveMQ [18] is a powerful open-source messaging broker that fully supports Java Message Service [19]. In addition, it supports many protocols [such as Simple Text Oriented Messaging Protocol (STOMP), AMQP and MQ Telemetry Transport (MQTT)] that allow clients to use a variety of messaging protocols. The publish/subscribe pattern of ActiveMQ is shown in Subfigure 2(a). Producers address messages to a topic, which functions like a bulletin board. Subscribers can then receive messages on the topics they have subscribed to. Apart from a publish/subscribe messaging pattern, ActiveMQ supports a peer-to-peer (P2P) model. In the P2P model, the producer submits messages to a message queue, and recipients can browse the queue and decide which messages they wish to receive. Every single message will be received by exactly one consumer. Since the P2P model works in an asynchronous way, the sender and receiver are decoupled by this model as well.

The RabbitMQ [20] messaging broker is created by the functional language Erlang. Erlang is especially suited for distributed applications, as concurrency and availability are wellsupported. The publish/subscribe model of RabbitMQ is shown in Subfigure 2(b). Messages are published to exchanges, which are often compared to post offices or mailboxes. Exchanges then distribute message copies to queues using rules called bindings. Bindings use an optional routing key attribute (acting like a filter) to bind a queue to an exchange. RabbitMQ brokers provide four exchange types (direct, fanout, topic, and headers) where a topic exchange type is used to implement a publish/subscribe messaging pattern. A topic exchanges route messages with one or many queues based on matching between a message routing key and the pattern that was used to bind a queue to an exchange.



(a) ActiveMQ [18]

(b) RabbitMQ [20]

(c) Kafka [21]

Fig. 2. Three open source publish/subscribe messaging brokers.

Apache Kafka [21] is a distributed publish/subscribe messaging broker, rethought as a distributed, partitioned, and replicated commit log. Like ActiveMQ, Kafka maintains feeds of categorized topics and a producer can publish messages to

Preprint version for http://dx.doi.org/10.1109/MIE.2017.2788850

a topic. The messages are stored and replicated as a cluster of brokers. Since Kafka is distributive in nature, a topic can be divided into multiple partitions, and each broker maintains one or more of the partitions for the purpose of load balancing. Then, a consumer subscribes to one or more topics and acquires the data from the brokers. To achieve fault tolerance, Kafka replicates a partition to multiple brokers. Each partition has one broker acting as the leader and several brokers acting as followers. The leader is in charge of all read/ write requests for the partition, and the followers passively replicate the leader. Once the leader fails, one of the followers will be automatically selected as the new leader. Previously, there was no standard for a messaging protocol. Common protocols include AMQP, MQTT, OpenWire, Representational State Transfer (REST), STOMP, and Extensible Messaging and Presence Protocol (XMPP), and nearly all of these protocols are supported by ActiveMQ and RabbitMQ, as shown in Table 1. However, instead of adopting an existing protocol, Kafka uses a binary protocol over Transmission Control Protocol because existing messaging protocols may not work well in providing a truly distributed messaging system. The designers of Kafka must build something that works differently. Thus, Kafka is able to provide high-throughput and fault-tolerant data transmissions, which makes it a good solution for efficient large-scale data ingestion in an industrial CPS.

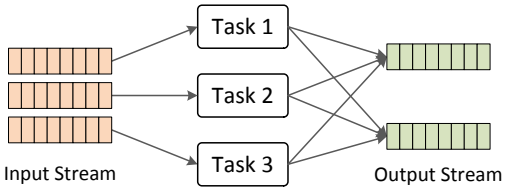
TABLE I
PROTOCOLS SUPPORTED BY ACTIVEMQ, RABBITMQ AND KAFKA

Protocol	ActiveMQ	RabbitMQ	Kafka
AMQP	1.0	0-8, 0-9-1	No
MQTT	Yes	Yes	No
OpenWire	Yes	No	No
REST	Yes	Yes	No
STOMP	Yes	Yes	No
XMPP	Yes	Over Gateway	No

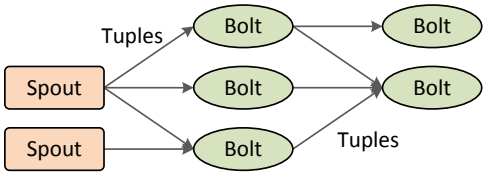
B. Real-time stream processing

Some data collected from IoT devices have to be processed in real time or near real time. This creates a demand for stream processing of big data. Essential to stream processing is the ability to continuously calculate mathematical or statistical analytics on-the-fly on the data stream. Stream processing solutions are designed to deal with high volumes at a high speed with a scalable, highly available, and fault-tolerant architecture. Many stream processing frameworks have been developed, among which Samza, Storm, and Spark Streaming have been widely adopted.

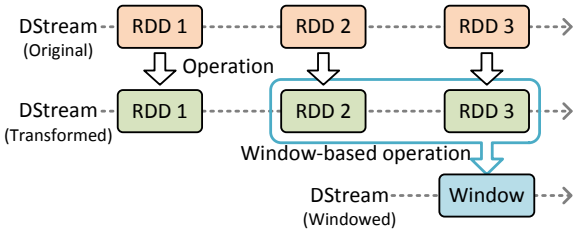
Samza [22] is an open-source distributed stream-processing framework based on Kafka. An example of Samza stream processing is shown in Subfigure 3(a). Like Kafka, a stream is divided into partitions in Samza, and each partition is a sequence of ordered messages. A job is the code that consumes and processes a set of input streams. To scale the throughput of the stream processing, a job is separated into smaller units of execution, named tasks. Each task consumes and processes messages from one or more partitions.



(a) Samza [22]



(b) Storm [23]



(c) Spark Streaming [24]

Fig. 3. Three popular stream processing frameworks.

Storm [23] is another distributed computational system for large volumes of high-velocity data. Storm users have to define topologies for how to process the data. As shown in Subfigure 3(c), the topology includes sources of streams, named spouts, and a set of bolts, which process input streams and produce output streams. In Storm, a stream is transformed into a sequence of tuples (an ordered list of elements).

Spark Streaming [25] does not process streams one at a time like Samza and Storm. Instead, it slices streams into small batches of resilient distributed datasets (RDDs) before processing, as shown in Subfigure 3(c). A continuous discretized stream (DStream) consists of a batch of RDDs, which could be operated in parallel or selectively operated over a sliding window. Therefore, Spark Streaming has the potential for combing batch and stream processing in the same framework.

Understanding the features of these frameworks will be important for every developer who wants to take full advantage of stream processing. When these technologies are used in productive systems, the industrial requirements must be considered. A comparison of Samza, Storm, and Spark Streaming is shown in Table II.

C. Scalable Data Storage

Big data generated by millions of IoT devices, such as sensors and radio-frequency identification readers and generators, need to be managed efficiently in the cloud. Since the data are always unstructured or semistructured with different

Preprint version for http://dx.doi.org/10.1109/MIE.2017.2788850

TABLE II
A COMPARISON OF SAMZA, STORM AND SPARK STREAMING [26]

Feature	Samza	Storm	Spark Streaming
Processing Model	one record at a time	one record at a time	micro-batch
Delivery Semantics	at least once	at least once	exactly once
Latency	milliseconds	milliseconds	seconds
Throughput	100k records per node per second	10k records per node per second	100k records per node per second
Language	Scala/Java	Any[1]	Scala/Java/Python

[1] Storm uses Thrift for defining topologies. Thrift can be used with any language.

volumes, the conventional relational database management system may not be appropriate. NoSQL databases provide a mechanism that does not enforce a strict schema, which allows for highly flexible data modeling. Therefore, NoSQL databases are increasingly used in big data and realtime web applications.

HBase [27] is a distributed NoSQL database and runs on top of an HDFS, which easily combines data sources that use a wide variety of different structures and schemas. In an HBase, data are stored in tables that are made of rows and columns. A table is partitioned into regions that are contiguous sorted ranges of rows of a table. A store corresponds to a column family in a region, and each store hosts a MemStore and a set of StoreFiles. The MemStore holds inmemory modifications to the store and will flush the data into an HDFS file StoreFile in HFile format when the MemStore reaches a certain size. Regions are served by HRegionServers (Figure 4), and an HMaster is used to assign regions to the HRegionServers. The assignments are then recorded in ZooKeeper. When a client wants to put data into or get data from an HBase, it has to first connect to ZooKeeper to find the HRegionServer in charge of the region. The client then talks to the HRegionServer directly.

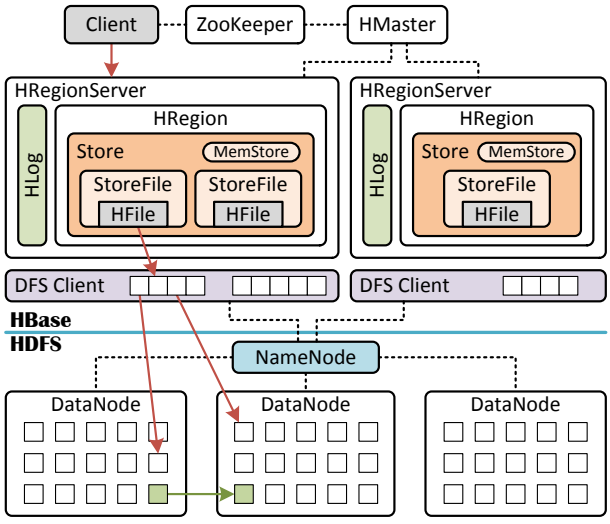


Fig. 4. Hadoop NoSQL database HBase [28].

An HBase uses an HDFS to store data, as shown in Figure 4. An HDFS has a master/slave architecture. An HDFS cluster consists of a NameNode and a number of DataNodes and exposes a file system namespace that allows the data to be stored in files. Each file is split into blocks, which are stored in a set of DataNodes. The NameNode executes file system namespace operations, e.g., opening, closing, and renaming files and directories. It also determines the mapping of blocks to DataNodes. The DataNodes are responsible for serving read and write requests and also make decisions about replication of blocks for fault tolerance. An HBase provides a wide-column data model on top of an HDFS. It can horizontally scale out to efficiently serve billions of rows and millions of columns by auto-sharding, which makes it suitable for scalable unstructured or semistructured data storage and management.

D. Large-Scale Batch Processing

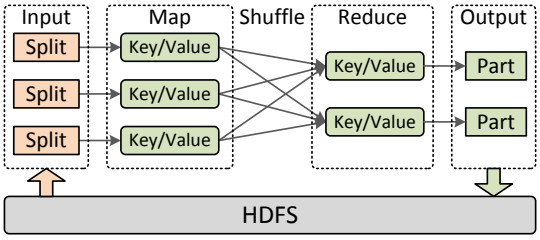
As an alternative to stream processing, data that has first been stored in file systems (e.g., HDFS) can be processed on a large scale as part of batch processing. A Hadoop MapReduce [29] is a highly scalable programming paradigm capable of processing massive volumes of data. As shown in Subfigure 5(a), MapReduce provides a programming model composed of a map procedure and a reduce procedure. The input data are divided into splits and assigned to map functions that map input key/value pairs to a set of intermediate key/value pairs. Then, the map outputs are merged during the shuffle phase. In the end, the reduce procedure carries out grouping and aggregation operations. In this way, MapReduce can process vast amounts of data in parallel on large clusters. However, since MapReduce requires considerable time to move the data in and out of the disk, it is not an appropriate solution for complex algorithms (e.g., iteration).

Spark [25] is designed to extend a Hadoop MapReduce to better support iterative algorithms (e.g., machine learning) and interactive data mining. Spark performs in-memory processing of data, which is much faster than MapReduce. As in Spark Streaming, an RDD is the basic abstraction in Spark, which is able to elegantly unify the batch and stream processing into a single framework. An RDD supports two types of operations: transformation and action [Subfigure 5(b)]. Transformation functions are lazy, and nothing actually happens when the code is evaluated. When an action is called, RDDs are processed or saved to a file. Compared to MapReduce, Spark holds immediate results in memory rather than on disk and provides more operations. For example, some of the transformation functions are map, filter, flatMap, groupByKey, reduceByKey, union, join, mapValues, and sort, while action operations include count, collect, reduce, lookup, and save. Therefore, Spark provides a faster and more flexible solution for large-scale batch processing on the cloud.

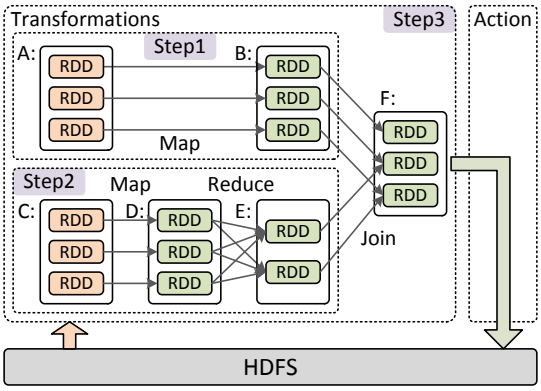
E. Big Data Analytics and Visualization

The core focus of this article is on the architecture and approaches to data collection and processing, but, for completeness, we also comment briefly on the analytics and visualization of the underlying data. Advanced analytics techniques can be applied to big data [4,30] , and visualized

Preprint version for http://dx.doi.org/10.1109/MIE.2017.2788850



(a) Hadoop MapReduce [29]



(b) Spark [25]

Fig. 5. Two example batch processing frameworks.

results represent system performance and trends. This has the potential to improve several industrial processes, e.g., energy management, supply chains, and manufacturing [31]–[35]. By using these techniques, data can be transformed into information about historical patterns, current performance, and future trends, which is critical for (BI&A) [10,36].

The most fundamental challenge for big data analytics is to explore the large volumes of existing data and extract useful information or knowledge for future application. This endeavor is referred to as data mining [37]. Data mining consists of a set of techniques, including association rule learning, cluster analysis, classification, and regression. Machine learning is another foundational technology for big data analytics [10]; it aims to design and develop algorithms that allow computers to identify behaviors based on empirical data. A major problem addressed by machine learning is to automatically recognize complex patterns and make intelligent decisions based on the data. Machine learning is highly related to data mining, and there are several similarities between them [38]. For instance, both often employ the same methods, such as classification, clustering, and regression; however, they have different emphases. Machine learning focuses on predictions based on known properties learned from the training data, while data mining focuses on the discovery of unknown properties in the data. Traditionally, data mining and machine learning are applied to small-scale data sets, so a single desktop computer is sufficient to fulfill the goals of data analytics [37]. In the era of big data, large-scale data are stored in the cloud, where a cluster of high-performance computers is deployed. Additionally, big data storage and processing are implemented in a distributed and parallel way, such as HDFS and Spark. Therefore, cloud-based data mining and machine learning solutions are needed to meet these new challenges. Examples of existing solutions are Hadoop Mahout and Spark MLlib, which provide libraries for implementing data mining and machine learning algorithms for data analytics.

Visualization can intuitively represent the results of big data analytics with images, diagrams, and animations. Effective visualization helps users analyze and reason about data and evidence. To convey ideas clearly and effectively, both aesthetic form and functionality need to be considered, with the high heterogeneity of the data making its visualization a challenge. Numerous business intelligence platforms, some with a strong IoT focus, utilize common big data platforms to deliver data analytics and visualization, e.g., Tableau, Orange, Power BI, IBM Watson Analytics, and SAP Leonardo.

IV. Research Challenges and Future Trends

The cloud, along with big data capabilities, empowers transformation to a digital, networked, intelligent, and knowledge-based industry. There is no doubt that industrial applications will benefit significantly from effective big data processing (examples of work on industrial big data and cloud are listed in Table III). However, cloud-based industrial CPSs are still in their infancy, and there are many challenges waiting to be addressed. Future steps should be taken to address challenges and upgrade conventional industrial systems, while also being mindful of new developments in cloud and IoT technologies. For the cloud-based industrial CPSs and applications to materialize and become a reality, several challenges identified in [6,9,36] need to be adequately addressed.

A. Management

It is possible that millions of smart devices could be utilized in industrial environments (e.g., a smart city), so new ways to easily manage large-scale and complex systems [9,36] must be considered. Dynamic discovery, interaction, and exchange of information, as well as lifecycle management (especially over federated systems), are challenging. The industrial cloud also allows for optimization from various perspectives, e.g., execution, communication, interaction, and management. Hence, more sustainable strategies need to be realized for managing these resources and businesses (e.g., energy-driven management). Such efforts should be seen in a greater context (e.g., smart city-wise).

B. Privacy and Security

While cloud computing brings key advantages, it also brings new and challenging security threats pertaining to outsourced data, critical infrastructure, operations, safety, and privacy [53]. For many applications, such as the smart home [36,46] and ubiquitous healthcare [54], there is much privacy-related information (e.g., patient medical history) that needs to be protected. There are also new attacks emerging, e.g., manipulating input data to misrepresent CPS state information [43,55].

Preprint version for http://dx.doi.org/10.1109/MIE.2017.2788850

TABLE III
EXAMPLE ENHANCEMENTS BASED ON THE INDUSTRIAL CLOUD

Domain	Application	Traditional On-premise System	Cloud-based Industrial CPS
Energy	Smart grid	High management costs and passive consumption	Producer & consumer participation, real-time smart grid data analytics, and sustainability [39]–[41]
	Energy management	Inflexible management and static network	Optimized and adaptive demand response, resilience, and information-driven interactions [42,43]
Building	Building automation	Information silos and inconsistent standards	Intelligent unified management (e.g., light control), and full information integration [44,45]
	Smart home	Inflexible service model	Personalized design of smart systems [46]–[48]
Manufacturing	Factory automation	Hierarchical management, inflexibility, proprietary systems	Adaptive CPS, high customization, low cost, and easy migration [49,50]
	Production management	Lack of effective operational mechanisms of resources and services	Fastest Time-to-market, highest quality, lowest cost, best service, cleanest environment, greatest flexibility, and high knowledge [51,52]

The private cloud provides a single-tenant environment for an organization in a relatively safe and efficient way compared to public clouds. A user can either implement an on-premise cloud platform or rent a virtually/physically isolated private cloud on the public cloud. However, private clouds will hinder interaction and knowledge sharing among enterprises. Private clouds also have a significant physical footprint, requiring allocation of space, hardware, and environmental control [5,56]. These assets have to be refreshed periodically, resulting in additional capital expenditure.

As an alternative solution, the public cloud provides a shared environment where clients could purchase services or computing resources based on their requirements [5]. The additional expense is virtually eliminated since the financial burden is shifted to a fee-for-service. However, security and privacy are problematic in a public cloud scenario, and the performance may be degraded due to the increased number of tenants [56]. Much work has been done to take advantage of the public cloud while protecting the privacy of clients at the same time. For instance, privacy preserving public auditing [57] enables the third-party auditing process in the public cloud without introducing vulnerabilities in user data privacy. Also, some new research focuses on searchable encryption [58],, which allows users to search encrypted data stored in the public cloud. However, there are still many privacy and security issues to be addressed, such as privacy preserving big data analytics [59] on the cloud. In addition, the current landscape of security standards for cloud computing is not yet mature [60], and the CSA is urging standardization of cloud confidentiality, integrity, and availability auditing.

C. High-Performance IoT Gateway

The IoT gateway is the bridge between the IoT and the cloud and is responsible for data collection and distribution. On one side, IoT gateways enable communication among networked devices through lightweight IoT protocols like Constrained Application Protocol and MQTT; on the other side, they exchange data with the cloud via distributed ingestion frameworks, like Kafka. It is possible for an IoT gateway to be connected to several brokers on the cloud at the same time; therefore, IoT gateways have to better manage the connections to the cloud, while handling the load balancing of the device communication [61]. IoT gateways are also the ideal protocol translation point for device networks, allowing for legacy systems to be connected to back-end big data frameworks without complete migration of the device network to new IoT technologies.

Traditional IoT gateways are mostly vendor-specific and are incompatible with other network devices. To address this problem, Intel proposed an IoT gateway solution that enables seamless and secure data flow between edge devices and the cloud by providing pre-integrated and pre-validated hardware and software building blocks. Because of its easy manageability, developers can focus on innovations for new services, big data solutions, and IoT-related applications. Others provide open-source solutions for IoT gateways, providing a wide range of Application Programming Interfaces and allow developers the flexibility to deploy customized applications. For example, Eclipse Kura runs on top of a Java virtual machine and leverages the Open Service Gateway Initiative to simplify the process of writing reusable software building blocks. However, an increased number of customized software may escalate the burden of the gateway, so resource optimization should be further considered.

D. Optimization for Data Processing

Cloud computing provides a distributed environment where data are stored in separated disks and processed by multiple servers in-parallel. Therefore, the cloud must address the challenge of minimizing the cost of data transmission [62]. Ineffective design of data processing algorithms may increase the communication burden on the cloud, and several solutions have been proposed to address this problem. For example, Map-reduce-merge [63] improves the batch processing MapReduce by adding a merge stage after the reduce phase to decrease the data to be transferred among servers. However, this type of solution may increase the design complexity since developers have to carefully design the topologies or the programming models. Rather than designing an efficient algorithm for a job, some studies [64] focus on how to allocate computing resources to match the requirements of the job. Challenges such as increasing the complexity and dynamics of jobs still need to be addressed.

Preprint version for http://dx.doi.org/10.1109/MIE.2017.2788850

E. Development and Engineering Tools

Efficient development and engineering tools will ease the creation of industrial services and applications within complex environments [6,9]. User-friendly cross-platform availability and capability are treated as the key aspects for development and engineering tools. Additionally, due to high-dynamic changes in industrial cloud-based systems, the tools should consider the robustness of the system to avoid interfering with the current operations when developing an application.

F. Migration and Integration of Legacy Systems

Transition is not expected to be instantaneous, and for existing investments in infrastructure, efforts toward migration and integration of legacy systems should not be underestimated [65]. As there is no one-size-fits-all solution, brown-field development, re-engineering, utilization of appropriate patterns, and meaningful integration into the rest of the ICPSs without sacrificing their capabilities and properties (e.g., safety and reliability) is seen as challenging.

G. Technologies for Improved Architecture

IoT and cloud technology continue to evolve and offer new options for system architecture. The rise in the number of smart devices connected to the cloud and the huge amount of data generated will significantly increase the workload of the centralized industrial cloud architecture. It may not be feasible to migrate some real-time applications that require a short reaction time to the industrial cloud because of bandwidth limitations and Internet delays. New technologies are therefore required to reduce the burden on the cloud in future.

Fog and edge computing [33,66,67] provide a promising way to enable computing services to reside within and at the edge of a network, as opposed to only on centralized servers. Fog computing uses a collaborative multitude of network devices to carry out a substantial amount of storage, communication, control, configuration, measurement, and management. Edge computing moves these functions completely to network endpoints, i.e., the edge of the network [68]. Since fog and edge computing devices are located closer to the edge of the network, they provide low latency and location awareness and improve the quality of service for streaming and real-time applications [69]. Local communication, interoperation, and decision making are supported by fog and edge computing, which will greatly improve the performance of the ICPSs, while the centralized industrial cloud provides online big data analytics and storage.

There are other technologies that will promote the development of the industrial cloud as well as future ICPSs, such as fifth-generation mobile networks [70] and deep learning [71]. These technologies will either improve the connectivity of the physical world or empower big data processing capability, which will significantly benefit a variety of industrial applications in the near future.

V. Conclusion

There are a number of technologies that could be adopted by the modern CPSs, such as the industrial cloud, the IoT, and big data methods. Combined, these technologies can contribute to the realization of a more intelligent, flexible, and cooperative industrial environment. In this article, we provided an overview of big data handling realized within the context of industrial cloud computing and CPSs. The future of industry will rely on a large ecosystem where industrial applications and enterprises are able to exchange information, share knowledge, and comprehensively collaborate in an efficient way. One of the main challenges would be to manage data flow from and to numerous operational technology devices within the IIoT. Industrial cloud services provide potentially powerful data storage, processing, and analytics capabilities for big data generated and collected from a variety of networked devices.

To gain the full benefits promised by big data and the industrial cloud, however, a number of key challenges must be met. A huge number of devices should be efficiently managed and the big data generated from these devices should be processed, stored, or visualized in real time. Comprehensive cooperation among devices, as well as enterprises in different domains, should be enabled in an efficient way, while privacy and security should be ensured at the same time. As the bridge of the IoT and the industrial cloud, IoT gateways should be well designed to support efficient protocol conversion, load balancing, and customized applications. Many other important issues like optimization in big data processing and design of development tools should be taken into account during the realization of an service-oriented architecture-based industrial cloud infrastructure. To better satisfy the new requirements of this future industry, more effort must be made to address these challenges.

References

[1] J. Höller, V. Tsiatsis, C. Mulligan, S. Karnouskos, S. Avesand, and D. Boyle, *From Machine-to-machine to the Internet of Things: Introduction to a New Age of Intelligence*. Elsevier, Apr. 2014.

[2] M. Wollschlaeger, T. Sauter, and J. Jasperneite, "The future of industrial communication: Automation networks in the era of the internet of things and industry 4.0," *IEEE Industrial Electronics Magazine*, vol. 11, no. 1, pp. 17–27, Mar. 2017.

[3] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, Tech. Rep., 2011. [Online]. Available: http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation

[4] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 2, pp. 652–687, 2014.

[5] P. M. Mell and T. Grance, "The NIST definition of cloud computing," National Institute of Standards and Technology (NIST), Tech. Rep., 2011.

[6] A. W. Colombo, S. Karnouskos, O. Kaynak, Y. Shi, and S. Yin, "Industrial cyberphysical systems: A backbone of the fourth industrial revolution," *IEEE Industrial Electronics Magazine*, vol. 11, no. 1, pp. 6–16, Mar. 2017.

[7] H. Kagermann, W. Wahlster, and J. Helbig, "Securing the future of German manufacturing industry: Recommendations for implementing the strategic initiative Industrie 4.0," German National Academy of Science and Engineering (ACATECH)., Tech. Rep., 2013. [Online]. Available: http://goo.gl/Rw4eef

[8] A. Colombo, T. Bangemann, S. Karnouskos, J. Delsing, P. Stluka, R. Harrison, F. Jammes, and J. L. Lastra, *Industrial Cloud-based Cyber-Physical Systems: The IMC-AESOP Approach*. Springer, 2014.

[9] S. Karnouskos, A. W. Colombo, and T. Bangemann, "Trends and challenges for cloud-based industrial cyber-physical systems," in *Industrial Cloud-Based Cyber-Physical Systems*. Springer, 2014, pp. 231–240.

Preprint version for http://dx.doi.org/10.1109/MIE.2017.2788850

[10] H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," *MIS quarterly*, vol. 36, no. 4, pp. 1165–1188, 2012.
[11] General Electric, "Predix: A cloud purpose-built for industrial big data and analytics." [Online]. Available: https://www.ge.com/digital/predix
[12] IBM, "Internet of Things on bluemix." [Online]. Available: http://www.ibm.com/cloud-computing/bluemix/solutions/iot/
[13] Opower, "Opower at a glance: A customer engagement platform tailor-made for utilities." [Online]. Available: https://opower.com
[14] Advantech, "Wise-paas: An integrated iot software services and cloud platform." [Online]. Available: http://www.advantech.com/embedded-boards-design-in-services/wisepaas
[15] "Industrial Internet Consortium (IIC)." [Online]. Available: http://www.iiconsortium.org
[16] "Internet Protocol for Smart Objects (IPSO) Alliance." [Online]. Available: http://www.ipso-alliance.org
[17] Cloud security alliance. [Online]. Available: https://cloudsecurityalliance.org/group/internet-of-things
[18] Apache ActiveMQ. [Online]. Available: http://activemq.apache.org/
[19] M. Richards, R. Monson-Haefel, and D. A. Chappell, *Java Message Service*. O'Reilly, 2009.
[20] RabbitMQ. [Online]. Available: https://www.rabbitmq.com/
[21] Apache Kafka. [Online]. Available: http://kafka.apache.org/
[22] Apache Samza. [Online]. Available: http://samza.apache.org/
[23] Apache Storm. [Online]. Available: http://storm.apache.org/
[24] Apache Spark Streaming. [Online]. Available: http://spark.apache.org/streaming/
[25] Apache Spark. [Online]. Available: http://spark.apache.org/
[26] G. Hesse and M. Lorenz, "Conceptual survey on data stream processing systems," in *2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS)*, Dec. 2015, pp. 797–802.
[27] N. Dimiduk, A. Khurana, M. H. Ryan, and M. Stack, *Hbase in Action*. Manning, 2013.
[28] H. Dutta, A. Kamil, M. Pooleery, S. Sethumadhavan, and J. Demme, "Distributed storage of large-scale multidimensional electroencephalogram data using hadoop and HBase," in *Grid and Cloud Database Management*. Springer, 2011, pp. 331–347.
[29] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, Jan. 2008.
[30] P. Russom, "Big data and analytics," TDWI Research, Tech. Rep., 2011.
[31] K. Wang, H. Li, Y. Feng, and G. Tian, "Big data analytics for system stability evaluation strategy in the energy internet," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 1969–1978, Aug. 2017.
[32] P. Lade, R. Ghosh, and S. Srinivasan, "Manufacturing analytics and industrial internet of things," *IEEE Intelligent Systems*, vol. 32, no. 3, pp. 74–79, May 2017.
[33] B. Tang, Z. Chen, G. Hefferman, S. Pei, T. Wei, H. He, and Q. Yang, "Incorporating intelligence in fog computing for big data analysis in smart cities," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 5, pp. 2140–2150, Oct. 2017.
[34] P. Xu, H. Mei, L. Ren, and W. Chen, "Vidx: Visual diagnostics of assembly line performance in smart factories," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 291–300, Jan. 2017.
[35] R. C. Basole, M. A. Bellamy, H. Park, and J. Putrevu, "Computational analysis and visualization of global supply network risks," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 3, pp. 1206–1213, Jun. 2016.
[36] Y. Sun, H. Song, A. J. Jara, and R. Bie, "Internet of things and big data analytics for smart and connected communities," *IEEE Access*, vol. 4, pp. 766–773, 2016.
[37] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, Jan. 2014.
[38] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Elsevier, 2011.
[39] D. Alahakoon and X. Yu, "Smart electricity meter data intelligence for future energy systems: A survey," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 1, pp. 425–436, Feb. 2016.
[40] Y. Simmhan, S. Aman, A. Kumbhare, R. Liu, S. Stevens, Q. Zhou, and V. Prasanna, "Cloud-based software platform for big data analytics in smart grids," *Computing in Science & Engineering*, vol. 15, no. 4, pp. 38–47, Jul. 2013.
[41] Z. Asad and M. A. R. Chaudhry, "A two-way street: Green big data processing for a greener smart grid," *IEEE Systems Journal*, vol. 11, no. 2, pp. 784–795, Jun. 2017.
[42] I. Hong, J. Byun, and S. Park, "Cloud computing-based building energy management system with ZigBee sensor network," in *2012 Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*. IEEE, Jul. 2012.
[43] W.-L. Chin, W. Li, and H.-H. Chen, "Energy big data security threats in IoT-based smart grid communications," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 70–75, Oct. 2017.
[44] V. E. L. Valenzuela, V. F. Lucena, P. Parvaresh, N. Jazdi, and P. Gohner, "Voice-activated system to remotely control industrial and building automation systems using cloud computing," in *2013 IEEE 18th Conference on Emerging Technologies & Factory Automation (ETFA)*, Sep. 2013.
[45] M. S. Parwez, D. B. Rawat, and M. Garuba, "Big data analytics for user-activity analysis and user-anomaly detection in mobile wireless network," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 2058–2065, Aug. 2017.
[46] M. Soliman, T. Abiodun, T. Hamouda, J. Zhou, and C.-H. Lung, "Smart home: Integrating internet of things with web services and cloud computing," in *2013 IEEE 5th International Conference on Cloud Computing Technology and Science*, Dec. 2013.
[47] Y. Cui, M. Kim, S. woo Kum, J. jin Jung, T.-B. Lim, H. Lee, and O. Choi, "Home appliance control and monitoring system model based on cloud computing technology," in *Lecture Notes in Electrical Engineering*. Springer Berlin Heidelberg, 2014, pp. 353–357.
[48] A. Yassine, S. Singh, and A. Alamri, "Mining human activity patterns from smart home big data for health care applications," *IEEE Access*, vol. 5, pp. 13 131–13 141, 2017.
[49] J. Wan, S. Tang, D. Li, S. Wang, C. Liu, H. Abbas, and A. V. Vasilakos, "A manufacturing big data solution for active preventive maintenance," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 2039–2047, Aug. 2017.
[50] S. Karnouskos, A. W. Colombo, T. Bangemann, K. Manninen, R. Camp, M. Tilly, M. Sikora, F. Jammes, J. Delsing, J. Eliasson, P. Nappey, J. Hu, and M. Graf, "The IMC-AESOP architecture for cloud-based industrial cyber-physical systems," in *Industrial Cloud-Based Cyber-Physical Systems*. Springer, 2014, pp. 49–88.
[51] J. Zhu, Z. Ge, and Z. Song, "Distributed parallel pca for modeling and monitoring of large-scale plant-wide processes with big data," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 1877–1885, Aug. 2017.
[52] F. Tao, Y. Cheng, L. D. Xu, L. Zhang, and B. H. Li, "CCIoT-CMfg: Cloud computing and internet of things-based cloud manufacturing service system," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 1435–1442, May 2014.
[53] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, "Information security in big data: Privacy and data mining," *IEEE Access*, vol. 2, pp. 1149–1176, 2014.
[54] J. Wan, C. Zou, S. Ullah, C.-F. Lai, M. Zhou, and X. Wang, "Cloud-enabled wireless body area networks for pervasive healthcare," *IEEE Network*, vol. 27, no. 5, pp. 56–61, Sep. 2013.
[55] S. Karnouskos, "Stuxnet worm impact on industrial cyber-physical system security," in *IECON 2011 - 37th Annual Conference of the IEEE Industrial Electronics Society*, Nov. 2011, pp. 4490–4494.
[56] T. W. Wlodarczyk, C. Rong, and K. A. H. Thorsen, "Industrial cloud: Toward inter-enterprise integration," in *Lecture Notes in Computer Science*. Springer, 2009, pp. 460–471.
[57] C. Wang, Q. Wang, K. Ren, and W. Lou, "Privacy-preserving public auditing for data storage security in cloud computing," in *2010 Proceedings IEEE INFOCOM*, Mar. 2010.
[58] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-preserving multi-keyword ranked search over encrypted cloud data," in *2011 Proceedings IEEE INFOCOM*. IEEE, Apr. 2011.
[59] R. Lu, H. Zhu, X. Liu, J. Liu, and J. Shao, "Toward efficient and privacy-preserving computing in big data era," *IEEE Network*, vol. 28, no. 4, pp. 46–50, Jul. 2014.
[60] J. Ryoo, S. Rizvi, W. Aiken, and J. Kissell, "Cloud security auditing: Challenges and emerging approaches," *IEEE Security & Privacy*, vol. 12, no. 6, pp. 68–74, Nov. 2014.
[61] S. K. Datta, C. Bonnet, and N. Nikaein, "An IoT gateway centric architecture to provide novel m2m services," in *2014 IEEE World Forum on Internet of Things (WF-IoT)*. IEEE, Mar. 2014.
[62] C. Ji, Y. Li, W. Qiu, U. Awada, and K. Li, "Big data processing in cloud computing environments," in *2012 12th International Symposium on Pervasive Systems, Algorithms and Networks*. IEEE, Dec. 2012.
[63] H. chih Yang, A. Dasdan, R.-L. Hsiao, and D. S. Parker, "Map-reduce-merge," in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data - SIGMOD '07*. ACM Press, 2007.

Preprint version for http://dx.doi.org/10.1109/MIE.2017.2788850

[64] D. Warneke and O. Kao, "Exploiting dynamic resource allocation for efficient parallel data processing in the cloud," *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, no. 6, pp. 985–997, Jun. 2011.
[65] J. Delsing, O. Carlsson, F. Arrigucci, T. Bangemann, C. Hübner, A. W. Colombo, P. Nappey, B. Bony, S. Karnouskos, J. Nessaether, and R. Kyusakov, "Migration of SCADA/DCS systems to the SOA cloud," in *Industrial Cloud-Based Cyber-Physical Systems*. Springer International Publishing, 2014, pp. 111–135.
[66] B. Tang, Z. Chen, G. Hefferman, S. Pei, T. Wei, H. He, and Q. Yang, "Incorporating intelligence in fog computing for big data analysis in smart cities," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 5, pp. 2140–2150, Oct. 2017.
[67] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in *Proceedings of the first edition of the MCC workshop on Mobile cloud computing - MCC '12*. ACM Press, 2012.
[68] E. Ahmed, A. Ahmed, I. Yaqoob, J. Shuja, A. Gani, M. Imran, and M. Shoaib, "Bringing computation closer toward the user network: Is edge computing the solution?" *IEEE Communications Magazine*, vol. 55, no. 11, pp. 138–144, Nov. 2017.
[69] I. Stojmenovic and S. Wen, "The fog computing paradigm: Scenarios and security issues," in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*. IEEE, Sep. 2014.
[70] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. A. Uusitalo, B. Timus, and M. Fallgren, "Scenarios for 5g mobile and wireless communications: the vision of the METIS project," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 26–35, May 2014.
[71] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, Jan. 2015.

Preprint version for http://dx.doi.org/10.1109/MIE.2017.2788850