2014 Eighth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing

Big Data for Cyber Physical Systems An Analysis of Challenges, Solutions and Opportunities

Antonio J. Jara, Dominique Genoud and Yann Bocchi
Institute of Business Information Systems
University of Applied Sciences Western Switzerland (HES-SO)
Sierre (Switzerland)
jara@ieee.org, {dominique.genoud, yann.bocchi}@hevs.ch

Abstract— **Cyber-Physical Systems (CPS) covers from M2M and Internet of Things (IoT) communications, heterogeneous data integration from multiple sources, security / privacy and its integration into the cloud computing and Big Data platforms. The integration of Big Data into CPS solutions presents several challenges and opportunities. Big Data for CPS is not suitable with conventional solutions based on offline or batch processing. The interconnection with the real-world, in industrial and critical environments, requires reaction in real-time. Therefore, real-time will be a vertical requirement from communication to Big Data analytics. Big Data for CPS requires on the one hand, real-time streams processing for real-time control, and on the other hand, batch processing for modeling and behaviors learning. This paper describes the existing solutions and the pending challenges, providing some guidelines to address the challenges.**

Keywords- Cyber Physical Systems, Internet of Things, Heterogenous Data Integration, Cloud Computing, Big Data.

I. Introduction

Cyber-Physical Systems (CPS) covers from M2M and Internet of Things (IoT) communications, heterogeneous data integration from multiple sources, security and privacy, to its integration into the cloud computing and Big Data platforms.

The integration of Big Data into CPS solutions presents several challenges and opportunities.

First, a common communication framework is required, in order to provide the proper features to make reliable, secure and sustainable the infrastructure.

Once all the relevant resources are reachable, data streams need to be integrated. Therefore, Quality of Data (QoD) and interoperability issues need to be addressed.

Big Data requires of the integration of data warehouses enabled by non-relational technologies such as Hadoop, MapReduce and Spark. These data warehouses will be usually allocated into Cloud Computing enabled platforms.

Big Data for CPS is not suitable with conventional solutions based on offline processing, since the interconnection with the real-world in industrial and critical environments requires reaction in real-time. Therefore, real-time will be a vertical requirement from communication to Big Data analytics.

Consequently, Big Data needs to be vertically integrated, and a non-classic solution is suitable, i.e., a solution based on offline or batch processing. Big Data for CPS requires on the one hand, real-time streams processing for real-time control, and on the other hand, batch processing for modeling and behaviors learning.

This paper describes briefly the state of the art for each one of the mentioned issues and offers some remarks of the existing solutions and the pending challenges that presents open opportunities.

II. M2M Communications

Machine-to-Machine communications with in-network Data Aggregation, processing, sensing and actuation for large scale Cyber-Physical Systems (CPS) [1] are enabled by Internet of Things (IoT) technologies. In particular, standardization bodies such as IETF, and the emerging oneM2M, which is composed by the most relevant worldwide standardization bodies such as ETSI and 3GPP, are working on the definition of a common IPv6-based communication layer with a RESTFul-based application protocol such as HTTP and the constrained version CoAP. Both of them are promoted by oneM2M for its integration into cellular and capillary communications.

The key features for the CPS communications as a difference with the usual communications are the requirements in terms of high reliability in emergency situations, or scenarios where privacy is extremely important, enhanced access priority in order to communicate ‘alarms’ in a variety of use cases, and treatment of unusual events (damage, equipment disappearing or location changed).

The main challenges which we intend to address in this domain revolve around these aspects:

- **Real Time Support.** The most important limitations of nowadays wireless sensors and actuators networks in CPS are coming from the fact that today’s commercial Wireless Sensor Networks (WSNs) are not tailored for the needs of the industry, since they are mostly influenced upon requirements derived from the consumer market. For example, CPS market presents serious limitations in the availability of sensors and actuators that offer real-time guarantees, when these features are of crucial importance in industrial application and the problem of safety [2].
- **Availability of sensing and communications services.** Critical applications need to be supported by a communication and data collection/aggregation infrastructure with a high degree of availability. Since some

978-1-4799-4331-9/14 $31.00 © 2014 IEEE
DOI 10.1109/IMIS.2014.139

of them are vital to save people's life, for instance, or to reestablish critical services in the shortest possible time.

Moreover, some of those applications are active on the occurrence of catastrophic events and emergencies which could make part of the sensing and communication infrastructure unavailable. So both types of infrastructure should be based on high availability solutions.

At the same time, not all applications require such a high level of availability. And the traditional approach based on over-provisioning and over-engineering sensing and communication infrastructure might turn out to be unaffordable for most applications, especially when it comes to maintenance and management costs. A challenge here is to device technologies capable of offering multiple levels of availability on the same infrastructure.

- **Maintenance.** The costs of maintaining the sensing and communication infrastructure in a large and distributed CPS system are very high, often higher than the cost of the embedded devices themselves. There is presently a lack of solutions for building high availability sensing and communication infrastructure out of cheap, commodity hardware. This calls for system level techniques for building high availability services on demand, and out of cheap hardware. The cost of reconfiguring or adapting sensing, communication and aggregation infrastructure for new services and varying service requirements are presently high and constitute a serious limit to the exploitation of the potential of large, distributed CPS.
- **Lack of open, flexible, multi-service solutions for sensing, communication and aggregation infrastructure for CPSs.** Existing CPS solutions, when it comes to sensing, communication and data aggregation, are often proprietary and closed, and ad hoc for a specific service, or a narrow set of homogeneous services, with homogeneous set of requirements. Installing new infrastructure every time a new need arises or a new service is required is often unaffordable, and it constitutes a serious barrier to innovation and to the diffusion of CPS. This calls for flexible solutions at the device level, for sensing and aggregation nodes, with the capability of being configurable and reconfigurable remotely and in response to changing service requirements.
- **Evolvability.** Another aspect which plays an important role for emergency response CPS services is related to the need for the infrastructure to discover and incorporate opportunistically elements which usually are not part of it, in order to better satisfy the needs of the application. In a scenario of a critical situation, many communication and sensing nodes might be unavailable, but a lot of other local infrastructure (mobile phones, sensors on cars, etc.) could be exploited opportunistically, possibly through participatory sensing and wireless P2P techniques, to supply in a short time the missing infrastructure elements.
- **Support of multiple different QoS requirements.** QoS at the communications and/or at the sensing level is nowadays not available. In the best cases, all systems are engineered in a very conservative ways. Or everything is offered on a "best effort basis", with very soft availability or performance guarantees [3].

Therefore, the opportunities are defined through offering high availability, flexible, multiservice sensing and communication infrastructure

In this domain, the aim should be to design a sensing, communication and aggregation infrastructure which supports multiple levels of quality of service provided, and of availability.

At the device level, a new generation of computing and communication embedded devices should be designed, based on an innovative design, capable of supporting the ever changing requirements and environmental conditions to which CPS are subject to, by means of a programmable and modular network of computing and forwarding resources. The wireless sensors nodes that have been proposed so far have too scarce memory and computing resources that prevent them to act as intelligent, reliable and secure nodes in a modern CPS infrastructure.

Security and autonomy are by far the most important features to consider and there are several issues that need to be addressed in order to deploy a safe, secure and adaptable communication over wireless sensor and actuator networks.

The innovations required for the CPS communication are:

- **Real-time cooperation.** CPS nodes interact with each other to achieve a particular application objective within a specific time frame. Users want their systems to operate in response to input on due time. This requirement is particular stringent for our different use cases which demand time responses close to milliseconds.
- **Autonomy and adaptability**. Predefined mechanisms (ranging from policy based systems, rule based systems up to neural networks) are not flexible enough in order to cope with unexpected problems and conditions in case of CPS. To avoid crash, mal-functioning and instabilities our CPS nodes need to be adaptable to unexpected conditions.
- **Secure.** A weak point of most IoT constrained systems, and consequently CPS, that have been developed so far is their security which is intrinsically linked with the constrained resources of most existing wireless sensors nodes. Authentication, integrity, availability, and non-repudiation are important security objectives for a communication infrastructure.
- **Flexibility and modularity**. Nodes should be modular enough in order to be able to stack modular multi-interface networks. be used in tandem on specific nodes requiring both interfaces.

III. HETEROGENEOUS DATA STREAMS

CPS deployments comprising by many devices (e.g., sensors, sensor networks, gateways, actuators) are nowadays widely deployed. However, it is likely that at a larger scale CPS deployments will become a growing trend and at the same time this grown will be acting as a challenge when not a bottle neck for the CPS systems in general. As the number of CPS system integration increases the multi-purpose sensors proliferates, more and more sensors will be deployed in either open or close CPS, similar what happened in IoT environments, which will create additional opportunities for CPS integration applications.

Effectively publishing information gathered from multiple sensors in heterogeneous CPS environments is a source of interoperability issues. Information must be formalized into a single, coherent format, contrasted and disambiguated with the different sources, and noise must be discarded. This overall process can be split into a set of techniques: Data Fusion or Data Aggregation techniques for merging multiple sources consistently, Quality of Data (QoD) techniques for validate the obtained data, and Linked Data techniques for publishing the information in a useful and annotated format. It must be taken into account the difficulties of using Linked Data in poorly connected environments [4], designing systems that tackle this problem.

Emerging CPS solutions will exploit those sensor/web platforms properties and deliver complex business services over a multitude of interconnected CPS environment. The overall process, with heterogeneous data streams coming from sensors or social streams, has been tackled in the recent literature. For example, context-oriented integration platforms for IoT [5], massive data management for sensor data relying on Cloud technologies [6] or approaches to tackle massive sensed information in the cloud [7] have been addressed.

For the heterogeneous data processing, since CPSs are system of systems, the data managed will be diverse and heterogeneous. Therefore, a unified picture of data will be necessary to be provided for the processing purposes. First, to provide a top-level CPS-data ontology which will be able to link and map together already-existing data specifications and formats, enabling a seamless semantic interoperability layer between already existing data format and models of different nature. And second, to create novel data fusion mechanism to integrate and aggregate the created data.

In order to work with heterogeneous data sources and user generated data, a QoD assessment mechanism needs to be carried out; using existing W3C recommendations like the PROV Data Model [8]. This recommendation describes the entities, activities and people involved in the creation of a piece of data, allowing the consumer to evaluate the reliability of the data based on their provenance information.

The QoD metrics are completeness, timeliness, ease of manipulation…) [9] to extend the available metadata. These metrics, added to the provenance information, will provide the necessary information to data-consumers to assess its usefulness.

Additionally, taking into account the multiple systems that will compose the emerging CPS, it will be necessary to implement processes that ensure the interoperability of the data at a semantic level. To do this an ontology matching mechanism will be developed. This mechanism will identify those concepts in different ontologies that are semantically similar or compatible, pairing and aligning them to create an equivalence map.

Finally, the usage of common ontologies supported by capillary and cellular networks such as the defined OMA Web Objects by oneM2M and IPSO Alliance will facilitate the interoperability and integration of heterogeneous and multi-vendor resources.

IV. Security and Privacy

Security and privacy in CPS were mainly studied in the context of smart grids [10] and M2M communications [11]. Several public key infrastructure (PKI) based solutions, including device attestation and certificate management, have been proposed in [12]. Some authentication techniques, using Diffie-Hellman key exchange, have been discussed in [13]. These approaches do not touch security in real-time sensitive data acquisition and processing environments on level that is required by this project. Privacy is also not very widely addressed issue especially with homomorphic cryptography as a tool for high level of privacy in the cloud computing of very sensitive data. Additionally in the context of synergetic approach based on the Cloud infrastructure and the real-time CPS devices no significant progress was made. There are a lot of gaps in this domain that can be made as a base for the research.

Security and Privacy in real-time sensitive architecture are required. Both security and privacy are very important for rapidly evolving and disruptive technologies. The research, development and deployment will touch all layers of our solution. We need to focus both on constrained, embedded devices that will provide the data and on the Cloud, that will collect and process information. Without holistic approach we will not be able to design any solution that could provide high security standards expected by our future users.

Opportunities are in the areas of:

- **Real-time CPS authentication**. Design authentication schemes that will provide high level of security with as minimal impact on the CPS device availability as possible.
- **Secure and efficient data transmission**. Ensure high level of security without compromising data transmission efficiency. This will be provided by combining innovative and very efficient stream-ciphers with optimized mechanisms for real-time applications.
- **Privacy-oriented data processing**. Data processing will be executed in the Cloud computing infrastructure. It is very important to provide high level of privacy in such environment. For this we will propose heterogeneous cryptography methodology. It will give the possibility to perform any required task on data that are cryptographically secured and for the processing unit completely anonymous.

V. Cloud Computing for CPS

To enable general services in the cloud, a client needs the ability to write his own applications that run on data in the cloud. A popular framework to allow a client to run distributed analytics in the cloud is Hadoop Map/Reduce, originally developed by Google [14], and now available as open source via the Apache Hadoop project [15]. Another framework that performs Map/Reduce is Apache Spark [16]. In the Map/Reduce model, data is distributed across a large number of storage servers. A "Map" operation is run in parallel on each the storage servers, on relevant data items that are stored on each of those storage servers. The results of the Map operation are then collected from the various storage servers and run through a "Reduce" process to accumulate the results from the data that was distributed among all the storage servers.

Hadoop includes a special purpose file system, HDFS (Hadoop File System), which supports the Map/Reduce programming model. The Map operation runs on data that is stored in HDFS and stores the results back in HDFS. There is typically significant overhead in setting up a Hadoop Map/Reduce program, but once it is set up, the computations on the data items take place in parallel.

It is required to expand the capabilities of performing analytics in a storage cloud, and to be able to synthesize analytics performed on an embedded system with analytics performed in the storage cloud. Nowadays, implementations of Hadoop require that the data be stored in HDFS (Hadoop File System), whereas not all clouds are built upon such a file system. In particular, Openstack is built upon Swift, not upon HDFS. Extensions are required to make a Hadoop-like operation work for Openstack Swift. Extensions are also required to allow for flows of analytics operations, with one operation feeding another..

VI. Big Data and CPS

New smart embedded systems and embedded devices are emerging and becoming connected to the Internet, which are generally defined as Internet of Things (IoT). Those devices can also act over the real world. Sensing and actuating capabilities can become ubiquitous, allowing unprecedented scenarios of interaction between the real and the virtual worlds. This army of devices will push Big Data architectures' scalability requirements to new limits. This will create a flood of real world information, considerably enriching applications, making them more aware of what happens in the real world, in real time, everywhere. In order to extract insights from such big streaming data, new data mining and machine learning techniques are required.

A main objective for Big Data in CPS is to analyze very large, fast and heterogeneous data streams from industrial environments. This can be achieved through machine learning. Machine learning is the most common technique to extract information from the data. Currently, there are two main strategies to do big data machine learning. The approaches are, on the one hand, using distributed systems: such as the described Hadoop and MapReduce. The most popular software project is Mahout [17] an open-source system to perform machine learning in batch setting. Many algorithms have been ported to run in parallel on a cluster environment [18]. Using online learning or data stream learning: Online learning [19] uses one instance at a time to update models in real-time. MOA [20] is a software framework for online machine learning running in stand-alone machines. It contains classification, clustering, and frequent graph mining.

Online algorithms make large use of approximation and statistical data structures [21] . These data structures make use of statistical properties of the data in order to reduce the computational and memory cost by forfeiting exact answers and trading off accuracy [22]. Some of these data structures have been ported to the distributed setting [23] or applied to stream mining [24, 25]. However, to date their main use has been in peer-to-peer systems and distributed information retrieval, and their application to data mining and machine learning has not been fully explored. In particular, practical issues in the application of these algorithms are just starting to be explored [26].

Finally, when dealing with large streams the issue of how to evaluate properly the mining algorithms gets even more important. The main reason is that with big data the issues of skew, class imbalance and scarcity are more pronounced [27]. Some of these issues have been studied recently [28], but their effect on the overall evaluation is not fully understood yet.

VII. Conclusions

Big Data for CPS needs to be addressed vertically, covering from communications to enable the reachability of all the data streams.

CPS needs to take care also of the heterogeneous data resources integration, in order to facilitate the aggregation and correlation of multiple data sources.

CPS needs to consider security and privacy issues such as anonymisation, data integrity and confidentiality, as a difference Big Data for CPS is involving real-world entities and people, therefore affecting more directly into our daily lives.

Big Data requires optimized data warehouses and cloud computing integration in order to make it scalable and affordable. At the same time, it needs to be local to provide real-time features, in terms of short-term reactions for events.

Big Data for CPS will require of an hybrid approach that considers, on the one hand, classic Big Data based on large amounts of data and offline data mining processes to build models and discover behaviors, and on the other hand, real-time stream data processing that adapts control, offers complex event processing, and makes systems also more autonomies.

Acknowledgment

This work has been made possible by funds from the frames of the IoT6 European project (STREP) from the 7th Framework Program (Grant 288445), and the HES-SO support for applied research.

References

[1] I. Stojmenovic, "Machine-to-Machine Communications with In-network Data Aggregation, Processing and Actuation for Large Scale Cyber-Physical Systems", IEEE Internet of Things Journal, DOI 10.1109/JIOT.2014.2311693, to be published in future issue, 2014.

[2] J. Akerberg, M. Gidlund, M. Bjorkman; "Future Research Challenges in Wireless Sensor and Actuator Networks Targeting Industrial

Automation", Proceedings of the 9th IEEE International Conference on Industrial Informatics (INDIN), 2011, pp. 410-415.

[3] Z. Pang, K. Yu, J. Akerberg, M. Gidlund, "An RTOS-based architecture for industrial wireless sensor network stacks with multi-processor support", Proceedings of the 2013 IEEE International Conference on Industrial Technology (ICIT), 2013, pp. 1216-1221.

[4] Charlaganov, Marat, Philippe Cudré-Mauroux, Cristian Dinu, Christophe Guéret, Martin Grund, and Teodor Macicas. "The Entity Registry System: Implementing 5-Star Linked Data Without the Web." arXiv preprint arXiv:1308.3357, 2013.

[5] Chen, Yeong-Sheng, and Yu-Ren Chen. "Context-oriented data acquisition and integration platform for internet of things." Technologies and Applications of Artificial Intelligence (TAAI), 2012 Conference on. IEEE, 2012.

[6] Bao, Yuan, et al. "Massive sensor data management framework in Cloud manufacturing based on Hadoop." Industrial Informatics (INDIN), 2012 10th IEEE International Conference on. IEEE, 2012.

[7] Fazio, M., et al. "Huge amount of heterogeneous sensed data needs the cloud." Systems, Signals and Devices (SSD), 2012 9th International Multi-Conference on. IEEE, 2012.

[8] L. Moreau, P. Missier, K. Belhajjame, R. B'Far, J. Cheney, S. Coppens, S. Cresswell, Y. Gil, P. Groth, G. Klyne, et al. Prov-dm: The prov data model.Candidate Recommendation, 2012.

[9] Pipino, Leo L., Yang W. Lee, and Richard Y. Wang. "Data quality assessment." Communications of the ACM 45.4 (2002): 211-218.

[10] H. Khurana, M. Hadley, N. Lu, D. A. Frincke, "Smart-Grid Security Issues", IEEE Security and Privacy, Vol. 8. Issue 1, pp. 81-85., 2010.

[11] R. Lu, X. Li, X. Liang, X. Sherman Shen, X. Lin, "GRS : The Green, Reliability and Security of Emerging Machine to Machine Communications", IEEE Communication Magazine, pp. 28-35, April 2011.

[12] A. R. Metk, R. L. Ekl R.L.. "Security Technology for Smart Grid Networks", IEEE Transactions on Smart Grids, Vol. 1, Issue 1, pp. 99-107, 2010.

[13] Z. M. Fadlullah, M. F. Fouda, N. Kato, A. Takeuchi, N. Iwasaki, Y. Nozaki. "Toward Intelligent M2M Communications in Smart Grid", IEEE Comm. Magazine, Vol. 49, Issue 4, pp. 60-65, 2011.

[14] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. Commun. ACM, 2008

[15] Apache Haddop - http://hadoop.apache.org/, 2014

[16] Apache Spark http://Spark.apache.org/, 2014

[17] Sean Owen, Robin Anil, Ted Dunning, and Ellen Friedman. Mahout in Action. Manning Publications, 2011.

[18] Ghoting, A., Krishnamurthy, R., Pednault, E., Reinwald, B., Sindhwani, V., Tatikonda, S., ... & Vaithyanathan, S. (2011, April). SystemML: Declarative machine learning on MapReduce. In Data Engineering (ICDE), 2011 IEEE 27th International Conference on (pp. 231-242). IEEE.

[19] Albert Bifet, Geoff Holmes, Bernhard Pfahringer, Philipp Kranen, Hardy Kremer, Timm Jansen, Thomas Seidl: MOA: Massive Online Analysis, a Framework for Stream Classification and Clustering. Journal of Machine Learning Research - Proceedings Track 11: 44-50 (2010)

[20] Nicolò Cesa-Bianchi and Gábor Lugosi. Prediction, learning, and games. Cambridge University Press, 2006

[21] Babcock, B., Babu, S., Datar, M., Motwani, R., & Widom, J. (2002, June). Models and issues in data stream systems. In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 1-16). ACM.

[22] Cormode, G., & Muthukrishnan, M. (2012). Approximating Data with the Count-Min Sketch. Software, IEEE, 29(1), 64-69.

[23] Ntarmos, N., Triantafillou, P., & Weikum, G. (2009). Distributed hash sketches: Scalable, efficient, and accurate cardinality estimation for distributed multisets. ACM Transactions on Computer Systems (TOCS), 27(1), 2.

[24] Chabchoub, Y., & Heébrail, G. (2010, December). Sliding hyperloglog: Estimating cardinality in a data stream over a sliding window. In Data Mining Workshops (ICDMW), 2010 IEEE International Conference on (pp. 1297-1303). IEEE.

[25] Matusevych, S., Smola, A., & Ahmed, A. (2012). Hokusai-Sketching Streams in Real Time. arXiv preprint arXiv:1210.4891.

[26] Heule, Stefan, Marc Nunkesser, and Alexander Hall. "HyperLogLog in Practice: Algorithmic Engineering of a State of The Art Cardinality Estimation Algorithm.", 2013

[27] Chawla, N. V. (2010). Data mining for imbalanced datasets: An overview. In Data Mining and Knowledge Discovery Handbook (pp. 875-886). Springer US.

[28] Gama, J., Sebastião, R., & Rodrigues, P. P. (2009, June). Issues in evaluation of stream learning algorithms. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 329-338). ACM.