International Journal of Control and Automation
Vol. 12, No.5, (2019), pp. 546 - 552

Big Data Quality for Reliable Industrial Internet of Things Based Systems

[1]Florin Popentiu-Vlădicescu, [2]Grigore Albeanu
[1]*University "Politehnica" of Bucharest & Academy of Romanian Scientists, Romania*
[2]*"Spiru Haret" University, Romania*

Abstract

The development of modern sensors and information technologies make possible to collect and process a large amount of reliability data to predict the system health of a monitored item. This work emphasizes on conceptual issues and methodological aspects related to data registration, filtering, smoothing, analyzing in order to predict important indicators of the quality of life and describes new practical strategies to analyze reliability data in context Big Data. Industrial Internet of Things (IIoT) is considered and architectures of IIoT are discussed from the reliability computational available approaches. The following hypothesis are validated: 1) Big data is an opportunity for reliability engineers when study/analyze big networks of sensors, large grids, or very large smart cities; 2) There is at least one reference architecture supporting high connectivity when working according to the Industry 4.0 framework; 3) There are developed many platforms, frameworks, and standards to serve as main vectors in implementing large scale applications supporting integrated Big Data technologies, Industrial Data and Sensors oriented protocols.

Keywords*: Big Reliability Data, Industrial Internet of Things, Industry 4.0*

INTRODUCTION

Both industrial and social systems are increasing in complexity due to new technologies applied to information processing. Recent developments in embedding and integration had offered new opportunities to collect, filter, analyze, and interpret huge collections of data generated by fleet of smart entities (sensors, smart meters, IIoT devices etc.).

This paper contributes both to the world of Big data and the world of IIoT applications with new points of view on reliability of complex systems. The 8V+CPU model of Big Data is considered and all characteristics are detailed according to reliability practice. Moreover, methods addressing the reliability of systems based on Industrial Internet of Things world is presented related to the Smart Cities [3], Systems of Systems (SoS [7]) and Systems Operating/Environmental (SOE [19]).

The paper is organized according to the main objectives. Firstly, the world of big data applicable to reliability data collection and analysis is described in the next section. Complex systems, such as SoS, SOE etc., are considered in the third section from the reliability point of view. The fourth section addresses the statistical computing and artificial intelligence methodologies already developed and the new developments in Big Data context.

QUALITY OF BIG RELIABILITY DATA

The most recent model for Big Data is 8V+CPU [4, 8, 24], and is based on high values of the following attributes: Volume, Velocity, Variety, Variability, Veracity, Value, Visualization, Volatility, Complexity, Privacy and Usability.

The Volume is significant due to "the large size of data collections generated by a large number of actors like: people (in various societal activities), sensors, and smart devices used in applications that produce real time data streams", as given in [24]. By Velocity we understood both "the speed of data generation" and "the minimum time to produce the best decision" when working with transactional data, multidimensional data, or stream data. Variability is applied in the context of data registration, or data filtering methods used before analyzing the data. Variety addresses various types of data: structured data, semi-structured data, and unstructured data. Veracity is important to be considered when collecting and examining uncertain, imprecise, incomplete, vague, or "dirty" data. The large volumes of data collected from field have Value, being useful to make appropriate decisions and learn efficiently about the process behavior [6]. Visualization consists of various means to describe patterns, display analytics, and give three-dimensional views about the evolution of data in some interval of time [10]. Volatility is important in Big Data context due to the in-creasing size of data to be stored by digital means. It is natural to store the most recent window of data useful to make a decision, mainly for real-time decision making processes.

For real world Big Data applications, the interest is both in the size of the "sample" (volume), and the dimension of data (quantitative, categorial, univariate, multivariate). Sometimes, the second aspect is known as complexity (C). Moreover, the study of Agrawal et al [2] brings new attributes to be considered when deal with Big Data: privacy (P) and usability (U). Privacy makes sense not only for data collected by governmental agencies, but also for data collected by smart meters located in smart cities, through Internet of Things paradigm [23].

From reliability point of view any Big Reliability Data analysis procedure should include the following five major steps: (1) AR - acquisition and recording; (2) CEA - cleaning, extraction, and annotation; (3) IAR - integration, aggregation, and representation; (4) MA - modelling and analysis, and (5) VI - visualization and interpretation. This architecture is well suited for maintenance and reliability analysis of complex systems as described in the next section.

Big data AR pipeline follows some steps like: data collection (from log files, sensors, mobile items, vehicular items etc.), data transmission (depending on the ICT infrastructure, protocols, data protection assurance etc.), and data pre-processing (outlier identification, noise elimination, data fitting and smoothing etc.). Outlier identification [1] may benefit on various approaches developed by robust statistics, clustering practices, pattern identification practices, and artificial intelligence using association rules.

The CEA pipeline depends on methods for pattern identification in large collections of data and the labeling (annotated) procedure applied to such patterns for future use in learning from data and prediction oriented tasks. During cleaning, the following steps follows: error type definition, error instance identification, error correction. In this manner some incomplete and not correct data will be identified and removed. The extraction is based on mining procedures dependent on the problem to be solved. Also, redundancy elimination (by redundancy detection, data filtering, and data compression) is a required procedure to be applied in order to improve the transmission time and costs with data storage.

The final aim of IAR step is to identify an unique representation of data coming for various and heterogeneous sources. The current solution is based on RDF (Resource Description Framework [27]).

Various models should be proposed, analyzed and tested for validity during the MA step. Learning from data is a difficult task, but challenging. In order to interactively discover

patterns in data, the researcher may use adequate software tools. Depending on the type of analysis, Big Data technologies can be oriented on batch processing (analytics on “data at rest”) or stream processing (analytics on “data in motion”) [24, 28].
In the following, let enumerate some Big Data approaches useful in reliability and maintenance. The most recent solution related to monitor and optimize the systems reliability in context Big Data by maintenance optimization is IBM-PMO [34]: "IBM Predictive Maintenance and Optimization". It collects data from machines and analyzes to learn about failures and predict equipment failure". A detailed view into equipment performance is generated to be used for maintenance efforts' optimization. IBM PMO permits to the reliability engineer the risks identification and management in the case of failure or a halt in operations, through six functionalities:

1. **Health Level Evaluation**: The health scores of every source of data are calculated according to specific models and the future life is predicted;
2. **Real-Time Monitoring**: The assets and processes are monitored in real-time;
3. **Early Detection**: The asset failures and quality issues are detected earlier;
4. **Root Cause Identification**: By mining procedures the cause of failure can be identified;
5. **Recommendation**: A recommendation plan on maintenance operations is generated by an optimization module;
6. **Custom Orientation**: Custom solutions can be generated depending on specific maintenance particular procedures.

Zhang [35] proposes the ABSAD (Angle-based Subspace Anomaly Detection) approach to fault detection in high-dimensional data. Also, Cannarile et al [5] made investigations on large collections of reliability data generated by complex systems, as described in next section.

RELIABLE SYSTEMS IN INDUSTRY 4.0 CONTEXT

Industrial and Sensor Data (ISD) are a Big Data source considered in this paper. ISD contains Industrial Internet of Things (IIOT), Autonomous and Unmanned Systems, and Internet of Things used in Homes (IOTh). This practice belong to the Industry 4.0 step of modern society development [15]. However, some systems belongs to a special class, called "Systems of systems", which are more sophisticated.
Sensed information (collected from sensors or IoT devices) is transferred to a data collection point (base station) through wired or wireless networks. In this way, the sensory data is assembled in different sensor nodes of the network of sensors and sent to the base station for processing. With this assumption a fog based computing environment is necessary [12, 25, 33].
Moving from Systems of Components (SoC) to Systems of Systems (SoS), determines that engineers, following a "divide and conquer", will consider the complex engineering systems as larger cooperating systems and more similarly to natural systems. Therefore, more emphasize is necessary to understand "complex events", under imprecision and/or uncertain information. According to [18], the distinctive factor between a system with respect to SoS is "understanding the aspect of the environment or otherwise stated the differences between a system or a group of systems that constitute the SoS". This characteristic asks for high levels of safety, reliability, maintainability, and dependability for SoS.
The SoS reliability engineering depends on its nature [7], according to the following classes: virtual - based on resource sharing; collaborative - based on agreements; acknowledged - based on collaborative management through a well defined interface, and directed SoS - based on centralized management. The SoS reliability is estimated differently depending on the specific architecture and particular reliability requirements. A detailed analysis on T&E to acknowledged SoS was developed by Dahman et al [7].

The mentioned author found that a framework for T&E (Test and Evaluation) should consider evidence-based approaches, continuously assessment of SoS, and learn about SoS performance by extending T&E "to include continual feedback processes".
SOE (System Operating Environmental) concept is considered in [19] to describe data recorded by sensors or smart chips installed in a product or equipment to measure different variables like environmental parameters, the usage rate, system load etc. Two or more sensors can be used to observe the environment (collect environmental variables) and their output signals are combined by an algorithm to provide a single enhanced measurement. This process, called sensors fusion, may allow the measurement of a phenomena that would be unidentified if otherwise will be proceed.
Big SOE data are generated by various systems equipped with ISD devices: transportation engines (aircrafts, merchant or defense navies, locomotives, automobiles), solar energy devices, wind energy devices, power distribution transformers, medical systems (computed tomography scanners, pressure sensors in infusion pumps or sleep apnea machines, airflow sensors in anesthesia delivery systems etc.), and smarter meters placed in various locations situated in the new smart cities [31]. A smarter sensor sensor is a part of a larger system that comprises microprocessors, modem chips, power sources, and other related devices (a kind of small scale SoS). Any smarter sensor can sent three types of data: identification data (I-data), maintenance data (M-data) and the environmental variable (X-data).
Sometimes, alternatively to sensor fusion approach, is useful to follow a fault-tolerant approach in order to maintain the data sequence integrity over an interval of time. For instance, the sensors installed in a meteorological base point can be organized in groups of three for every parameter to be measured. At least two values should be available to report a "centroid" value to the prediction model. The faulty sensor, if is the case, should be replaced. This example illustrates that some individual systems (components) can have low operational reliability, but a mechanism to rebuild the "initial" state is activated by replacement or repairing if a failure appear.
Also, by spatial redundancy, with a minimum two communication links of any node in a sensor networks under an efficient routing protocol to find the destination, a high reliability is obtained.
The availability of the Industrial Internet Reference Architecture (IIRA) [13] makes possible the design of interoperable smart devices and the development of SoS and SOE having increased reliability and easy to be monitored for optimized maintenance in context Big Data. According to [14], the IIoT connectivity stack model has six layers supporting the exchange of various entities between participants: physical (bits), link (frames), network (packets), transport (messages), connectivity framework (data: state, events, streams) and distributed data interoperability and management (information).
The core criteria of IIRA are satisfied by the Data Distribution Service (DDS) specification. DDS proved to meet nonfunctional requirements including performance, scalability, reliability, resilience, security and safety. According to [14], "since DDS does not require servers that could fail and supports redundancy, it makes “fast” reliability and availability much easier. DDS also eliminates huge efforts with server con-figuration, startup order, or failover to backup servers." Also, the OPC-UA (Open Platform Communications Unified Architecture [21]) and oneM2M (dedicated to Machine to Machine and IoT technologies [20]) are other valuable specifications with highly conformity related to IIRA.
Due to continuously developments in the field of ISD, there are expected improved architectures, protocols and more valuable platforms to be appropriate to end-users.

RELIABILITY MODELS IN CONTEXT BIG DATA

From computational point of view, SOE data are modeled as vectors of time series [19] and should be analyzed by specific methods, and well suited tools. An important concept,

suitable to analyze SOE data, is "dynamic covariate information". With this respect, for every item is registered not only the failure time, but also its history and the current environment variables (t, f_t, H_t, X_t), t in some interval. The registration may use the fixed or variable clock model: t_0, t_1,..., t_n. If X_t describes the covariate history, then:

$$X_i(t_i) = \{X_i(s), s \text{ in } \{t_{i1}, t_{i2}, ...t_{iNi}\}\}, \quad (1)$$

where N_i is for the number of time points where the measurements were taken before time t_i. At the mentioned time (t), a binary variable (f_t) will give the information on the failure existence (1), or missing a failure (0). Analyzing data collected from IoT, with a huge history (H_t), will increase the processing time, mainly important when real time analysis is required. Depending on the application, the most relevant historical aspects will be selected (filtered) and used during the decision making process. Also, a censoring time sequence can be used and the indicator time will compare the true failing time against the censoring time in order to obtain the above binary values.

The mathematical model of the covariate process (1) is described by

$$X_i(t_i) = \square \square + Z(t_i)w_i + \square_i, \quad (2)$$

where w_i is used to model the variability (depending on the variance $\square$) in a unit's covariate process over time, and $\square_i$ is used to model the rate fluctuations at different time points. The parameters ($\square \square \square \square$) from (2) can be estimated by the Maximum Likelihood method. The distribution of remaining life of units can be estimated by Monte-Carlo algorithms as shown in [11].

An alternative approach is given by [5], for a large number of assets operating on the field. For every asset a the following data are available: the values of K control variables arranged into vectors

$$x_a = (v_a[1], v_a[2], ..., v_a[K]), \quad (3)$$

with a in {1, 2, ..., A}, and A the number of assets. Every asset a is observed over time, at moments M[0], M[1], ..., M[N_a], where the N_a observations on its behavior are registered along an information vector F[0], F[1], ..., F[N_a], telling one that M[i] is the failure time if F[i] = 1. If F[i] = 0, the asset is well functioning. If there are many types of assets under monitoring, the partitioning in a number of classes is running first. Then, the reliability analysis can be applied on every class.

Statistical analysis of reliability data can be developed using SPREDA [30], and pbdR [22]. In order to discover patterns and relations in big reliability data, various artificial intelligence methodologies can be used: data clustering [5], context-based analytics [29], statistical inference [17, 19], and deep learning [26, 32].

Based on MapReduce technique, an algorithm to estimate the reliability of a SOS system with components operating independent, but the whole system being monitored centralized, can be easily developed. In this case R(t) – the reliability function is evaluated according to a binary tree describing the formula for the reliability of the whole system. However, the reliability of some components can be evaluated in parallel, in distributed manner, according to a MAP protocol. Finally, the reliability formula is implemented by a REDUCE type protocol. This technique will apply according to the methods described by Eaton et. al [9]. This approach can be used also, for computing the reliability of a FOG distributed architecture.

CONCLUSION

The above presentation has considered both the world of big reliability data collected from a large population of devices and the world of industrial internet of things. The aim of the investigation was to validate the following hypothesis: 1) Big data is an opportunity for reliability engineers when study/analyze big networks of sensors, large grids, or very large smart cities; 2) There is at least one reference architecture supporting high connectivity when working according to the Industry 4.0 framework; 3) There are developed many platforms, frameworks, and standards to serve as main vectors in implementing large scale applications supporting Big Data technologies, Industrial Data Management, and Mission-Oriented Protocols for Sensors.

Moreover, the system reliability engineering field was revisited taking into account both data sources and the new methodologies used for reliability data. The following frameworks have been considered: Systems of Systems (SoS), Big Data, System Operating/Environmental (SOE) data, and IDS reliability.

As described above, the development of IIOT based applications will continue and the reliability of the designed system will be really increased, with a decreasing cost of maintenance.

REFERENCES

1) Aggarwal, C.C. (2017). Outlier Analysis. DOI 10.1007/978-3-319-47578-3_2, Springer International Publishing AG.
2) Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., &Widom, J. (2012). Challenges and Opportunities with Big Data: A white paper prepared for the Computing Community Consortium committee of the Computing Research Association. http://cra.org/ccc/resources/ccc-led-whitepapers, last accessed 2019/11/10.
3) Al Nuaimi, E., Al Neyadi, H., Mohamed, N., & Al-Jaroodi, J. (2015). Applications of big data to smart cities. Journal of Internet Services and Applications. DOI 10.1186/s13174-015-0041-5, 6-25.
4) Big Data Value V4.0, http://www.bdva.eu/sites/default/files/BDVA_SRIA_v4_Ed1.1.pdf, last accessed 2019/11/10.
5) Cannarile, F., Compare, M., Di Maio, F., & Zio, E. (2015). Handling reliability big data: a similarity-based approach for clustering a large fleet of assets, In: Podofillini, L., Sudret. B., Stojadinovic, B., Zio, E., Kröger, W. (eds) Safety and Reliability of Complex Engineered Systems (ESREL 2015), pp. 891-896, CRC Press.
6) Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A Survey. Mobile Netw Appl 19, 171-209.
7) Dahmann, J., Rebovich, G., Lane, J.A., Lowry, R., & Palmer, J. (2010). Systems of Systems Test and Evaluation Challenges. 5th IEEE International Conference on System of Systems Engineering, DOI: 10.1109/SYSOSE.2010.5543979.
8) DeVan, A.: The 7 V's of Big Data (2016), https://www.impactradius.com/blog/7-vs-big-data/, last accessed 2019/11/10.
9) Eaton,C., Deutsch, T., Deroos, D., Lapis, G., &Zikopoulos, P. (2012). Understanding Big Data, Ana-lytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill.
10) Garcia, I., Casado, R., &Bouchachia, A. (2016). An Incremental Approach for Real-Time Big Data Visual Analytics. IEEE International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), DOI: 10.1109/W-FiCloud.2016.46.
11) Hong, Y., & Meeker, W.O. (2011). Field-Failure Predictions Based on Failure-time Data with Dynamic Covariate Information, Statistics Preprints 72, IOWA State University.
12) IEA Wind TCP Task 33 Team - RP17 (2017). Wind Farm Reliability Data, https://community.ieawind.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=7b6fdbfe-9b58-addf-0bc6-78c2d3f4aa5d&forceDialog=0, last accessed 2019/11/10.
13) IIC/IIRA (2017). The Industrial Internet of Things, Volume G1: Reference Architecture https://www.iiconsortium.org/IIC_PUB_G1_V1.80_2017-01-31.pdf.
14) IIC//CF (2017). The Industrial Internet of Things, Volume G5: Connectivity Framework, IIC:PUB:G5:V1.0:PB:20170228.
15) Industry 4.0 (2019). https://en.wikipedia.org/wiki/Industry_4.0, last accessed 2019/11/10.
16) ISO/IEC JTC1 (2015). Information Technology. Big Data - Preliminary Report, ISO.

17) Letot, C., &Dehombreux, P. (2009). Degradation models for reliability estimation and mean residual lifetime. In Proceedings of the 8th National Congress on Theoretical and Applied Mechanics, pp. 618-625 (2009).
18) Lubas, D.G. (2017). Department of defense system of systems reliability challenges, RAMS, DOI: 10.1109/RAM.2017.7889676.
19) Meeker, W.Q., & Hong, Y. (2014). Reliability Meets Big Data: Opportunities and Challenges. Quality Engineering 26(1), DOI: 10.1080/08982112.2014.846119.
20) oneM2M (2019). Machine to Machine Communications and the Internet of Things, http://www.onem2m.org/, last accessed 2019/11/10.
21) OPC-UA (2019). Open Platform Communications - Unified Architecture, https://opcfoundation. org/about/opc-technologies/opc-ua/, last accessed 2019/11/10.
22) Ostrouchov, G., Chen, W.-C., Schmidt, D., & Patel, P. (2012). Programming with Big Data in R. In 6th Extremely Large Databases Conference (XLDB).
23) Perera, C., Ranjan, R., Wang, L., Khan, S.U., &Zomaya, A.Y. (2015). Big Data Privacy in the Internet of Things Era. IEEE IT Professional 17(3), 32-39.
24) Popentiu-Vladicescu, F., Albeanu, G., & Madsen, H, (2018). Recent Methodological and Conceptual Issues in Software Reliability Engineering. In Roceanu, I. et al. (eds) The 14th International Scientific Conference eLearning and Software for Education, Bucharest.
25) Popentiu-Vladicescu, F., &Albeanu, G.(2017). Software reliability in the Fog computing. The Inter-national Conference on Innovations in Electrical Engineering and Computational Technologies, DOI: 10.1109/ICIEECT.2017.7916578, IEEE.
26) Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data pro-cessing. EURASIP Journal on Advances in Signal Processing 67, DOI 10.1186/s13634-016-0355-x.
27) RDF (2019). Resource Description Framework, https://www.w3.org/RDF/Validator/, last accessed 2019/11/10.
28) Reeve, A. (2013). Managing Data in Motion. Elsevier.
29) Sokol, L., & Chan, S. (2013). Context-Based Analytics in a Big Data World: Better Decisions, IBM Redbooks, 1-8.
30) SPREDA (2015). Statistical Package for Reliability Data Analysis, https://cran.r-project.org/web/packages/SPREDA/SPREDA.pdf.
31) Talari, S., Shafie-khah, M., Siano, P., Loia, V., Tommasetti, A., &Catalão, P.S. (2017). A Review of Smart Cities Based on the Internet of Things Concept. Energies, 10, 421; doi:10.3390/en10040421.
32) Tamura, Y., Matsumoto, M., & Yamada, S. (2017). Software Reliability Model Selection Based on Deep Learning. Software Networking 1, 10.13052/jsn2445-9739.2017.008.
33) Volovoi, V. (2016). Big Data for Reliability Engineering: Threat and Opportunity. Reliability 2, 11-15.
34) Watson IoT - IBM Predictive Maintenance and Optimization, 2019. https://www.ibm.com/support/ knowledgecenter/SS7TH3_1.0.1/com.ibm.pmo.doc/pdf/m_pmo_master.pdf, last accessed 2019/10/10.
35) Zhang, L. (2016). Big Data Analytics for Fault Detection and its Application in Maintenance, PhD Thesis, Luleå University of Technology, Luleå.