Optimizing Condition Monitoring of Big Data Systems

David Baglee[1], Unai Gorostegu[2], Erkki Jantunen[3], Jaime Campos[4], Pankaj Sharma[5]

[1] Faculty of Engineering and Advanced Manufacturing, University of Sunderland, UK
[2]Mondragon University, Loramendi 4, 20500 Mondragon, Spain
[3]VTT Research Centre. Helsinki. Finland.
[4]Linnaeus University, Faculty of Technology, Department of Informatics, Sweden
[5]Department of Mechanical Engineering, IIT Delhi, New Delhi, India

***Abstract*— Industrial communication networks are common in a number of manufacturing organisations. The high availability of these networks is crucial for smooth plant operations. Therefore local and remote diagnostics of these networks is of primary importance in determining issues relating to plant reliability and availability. Condition Monitoring (CM) techniques when connected to a network provide a diagnostic system for remote monitoring of manufacturing equipment. The system monitors the health of the network and the equipment and is therefore able to predict performance. However, this leads to the collection, storage and analyses of large amounts of data, which must provide value. These large data sets are commonly referred to as Big Data. This paper presents a general concept of the use of condition monitoring and big data systems to show how they complement each other to provide valuable data to enhance manufacturing competiveness.**

I. INTRODUCTION

In the big data era, we are dealing with the exponential increase of global data and big datasets. Unlike seemingly similar terms such as "massive data" or "very big data," big data refers to the datasets that could not be perceived, acquired, managed, and processed by traditional Information Technology (IT) and software/hardware tools. Big data if often characterized by the 'four dimensions, namely **Volume** (scale), **Variety** (heterogeneity), **Velocity** (rapidly streamed data) and **Veracity** (uncertainty). Very large amounts, rapidly streamed, heterogeneous and/or uncertain – maybe called big data. It is important to understand which of the former mentioned aspects of the four Vs' that a maintenance strategy will involve, in this case, CBM.

In certain domains, such as the use of data for manufacturing intelligence it is the combination of the 4 V's that need to be considered. It is therefore crucial to have an understanding of the data, and the use of data mining algorithms, which are to be employed in the domain of interest. The main reason is to ensure scalability and the big data technology required such as Hadoop, MapReduce, and hive systems. However, in the area of maintenance management the data can be divided into traditional and non-traditional data based upon the four V's.

Traditional data, when expressed as 'volume' is retrieved from condition mmonitoring systems, maintenance planning and a large set of detailed work orders (Zhang and Karim, 2011). The non-traditional data is often described as data, which is indirectly related to the manufacturing process, but just as important. This includes purchase contract, production, scheduling, asset depreciation value etc.

Variety includes semi-structured data, such as spreadsheets and data stored in relational databases, emails, XML files, and log files. Data classed as unstructured consists of pictures, audios, videos, web pages, Word and PDF documents.

Velocity and traditional maintenance data are transaction data and multidimensional data. Whereas, the data considered outside the traditional maintenance data and part of the velocity are the real-time condition monitoring data collected by sensors and instruments.

Veracity generally refers to the storage and data mining strategy to ensure the 'abnormalities' in the data are identified and removed. The data mining techniques employed to ensure usefulness is the biggest challenge when compared to volume and velocity. The data mining strategy must ensure (1) the data is 'clean' and (2) any 'dirty data' does not accumulate in the systems, therefore providing false and inaccurate information from which to develop your condition monitoring strategies.

The classification of big data is important to be able to understand its characteristics (Hashem et al. 2015). A Big data classification made by the authors is illustrated in Figure 1.

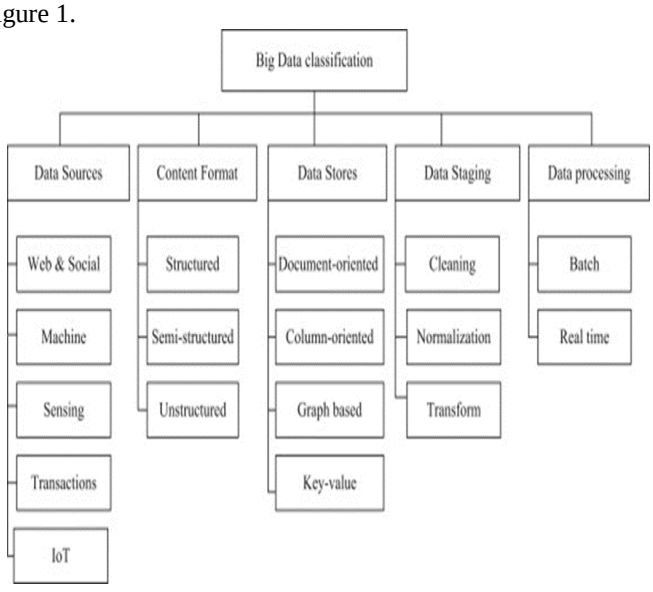

Figure 1. Big data classification (Hashem et al. 2015).

Figure 1 highlights several types of big data. The arrangements of the big data are made based on five features, data sources, content format, data stores, data staging, and data processing. To support the use of big data, based upon the sources of data identified within figure 1, aanalytical methods are required to derive or infer patterns of behaviour.

With regard to condition monitoring systems and data mining techniques to support machine learning, it is necessary to utilise descriptive and inferential statistics that are required to find patterns and help construct models. Data mining methods help to find regular patterns and similar items in data. Condition monitoring systems utilising machine-learning techniques make it possible for the system to learn and create the best model that describes the relation between a number of known and often unknown variables.

Consequently, an example of these is the classifiers, which are one of the frequent algorithms used in data mining (Wu et al. 2008). Examples of classifiers are the decision trees and ruleset classifiers. Another approach is based on the K-means algorithm. It is a simple iterative data mining technique to divide a given dataset into a user specified number of clusters, K. In addition, a machine-learning algorithm is the support vector machine which is a robust and precise method amid all recognised algorithms. Moreover, the Apriori algorithm is a popular algorithm to find frequent item sets from for instance a transaction and elicit association rules.

II. CONDITION MONITORING

The Condition Based Maintenance (CBM) is a predictive maintenance strategy where the main objective is to know the current health state of an asset and predict its future behaviour to carry out maintenance in the most appropriate time to decrease the downtime and consequently increase its availability.

To implement a CBM system, it is necessary to do an initial study on the asset that is going to be monitored. It is possible that not every asset could benefit from this strategy. For example, if a component fails but the machine can continue operating until the next scheduled maintenance, it can be assumed that this is not a critical component and therefore does not require monitoring. The use of available historical data could prove useful in this instance to establish the type of failures or the signs of an impending failure (Raheja et al. 2006).

A CBM system is usually composed of the following phases:

Data Acquisition: It is necessary to collect and store asset data to be able to process the data hen required. The data is typically collected using transducers, sensors such as, an accelerometer, thermometer and pressure. These sensors are used to transform a physical phenomenon into an electrical signal that can later be processed in a computer. The physical properties measured can vary and typically include vibration, temperature, pressure, force, flow and light. The output of the signal usually needs to be cleaned from the noise through different filters. Afterwards, the data is digitized using an analogue to digital converter when necessary and sent to the processing phase.

Data Manipulation: Once the raw data is digitized, it is stored in a local or remote database. Signal analysis is performed to compute meaningful descriptors and create virtual readings from the raw data (ISO-13374-1). The information extracted from the raw readings is then available for use.

State Detection: The State Detection block helps create baseline profiles and compares them to the new data to search for abnormalities to determine if there are any problems and, if so, what kind of problems exist. The use of mathematical calculations and algorithms such as kurtosis, envelope analysis and Fast Fourier Transform are required to ensure accuracy.

Health Assessment, A diagnosis is made using the information from the previous steps, to create a diagnosis of the current state of health of the asset, component or sub-component.

Prognostics Assessment: Once the diagnosis undertaken it is possible to predict how the asset will operate or fail. Different technical approaches exist to support this stage, (1) prognostics and (2) data-driven prognostics, which can be used in isolation or as a combined system.

Advisory Generation: This block provides the information about the actions that need to be taken regarding the maintenance of the asset to optimize life expectancy. With the information obtained from the prognostics, an action report is created and the system should be able to take consequent actions such as ordering a new component if necessary, scheduling a maintenance service or doing nothing if the system reports that the component is in good health condition. It is worth noting that the Advisory Generation then can be linked to a Computerized Maintenance Management System (CMMS) to create a fully automatic maintenance system, where the new component orders are taken care of by the system and the technicians work only when there is a real need for a replacement or maintenance.

As it has been mentioned before, to implement a CBM system it is necessary to do an initial analysis. This is the first phase where the big data systems can provide useful intelligence. From gathering big quantities of data, it is possible to draw valuable information. This information could lead to making decisions such as if it is necessary to monitor the component but can also help realize what kind of failures the system suffers, when it is necessary to replace the component and the general relevance of the component in the system. Different analytical tools such as Fault Tree Analysis (FTA), Root Cause Analysis (RCA) or Failure Mode, Effect and Criticality Analysis (FMECA) can also be helpful when determining the potential failures of a system and the most critical components.

III. The Use of Large Data Sets to Support Condition Monitoring

Regarding the big data analysis, it is worth mentioning that there are different types of data processing, suitable for various applications. Data can be processed in real-time for applications such as finance, navigation or intelligent transportation (Philip Chen et al. 2014). However, this type of processing requires very high computational power, and usually for CBM systems, a batch data processing system may be implemented. It is efficient in terms of processing a wide quantity of data collected over a period of time (Walker 2013). Batch processing is fundamentally high-latency. Therefore, if analysis is required on a terabyte of data all at once, it will not be possible to do that computation in less than a second with batch processing. Stream processing looks at smaller amounts of data as they arrive. Intense computations, like parallel search, and merge queries on the fly are possible by this method (Barlow, 2013). Critical systems like nuclear power plants will still require real time data stream analysis because batch data processing takes relatively longer time for diagnosis. This delay may have catastrophic results.

Real Time Big Data Analytics (RTBDA) is a complex analytical process that requires bringing data in from various sources. A data loss in a real time data analytic scenario can bias the diagnosis and may result in incorrect decision-making. New tools like *Flume* and *SQOOP* are helpful in transmitting data into Hadoop from different networks. Another tool called *Impala* enables real time ad-hoc querying. *Spark* is another open source cluster computing system that can be programmed quickly and runs fast. *Spark* relies on "resilient distributed datasets" (RDDs) and can be used to interactively query 1 to 2 terabytes of data in less than a second. In scenarios involving machine learning algorithms and other multi-pass analytics algorithms, *Spark* can run 10x to 100x faster than Hadoop MapReduce can. *Storm* is an open source low latency processing stream processing system designed to integrate with existing queuing and bandwidth systems. Normally, for a search query, there is a need to create search indexes, which can be a slow process on one machine. With *Storm*, the process can be streamed across many machines to get much quicker results (Barlow, 2013).

The data will be analysed using mathematical algorithms to determine the current health level of the asset. Available mathematical tools are the Fast Fourier Transform (FFT), Envelop analysis or Kurtosis. Once the current state is established, the system will do a prognostic and predict the future health levels and the Remaining Useful Life (RUL) of the asset by analysing the degradation level compared to the normal "profiles".

Stated earlier, various approaches exist to perform a prognostic, however, the two main methods are the data-driven and the model-based approaches. The data-driven method consists on a statistical approach where the historical data is required. Algorithms such as neural networks or pattern recognition with stochastic models are used to perform the required data analysis. A disadvantage of this approach is the large amount of time required to obtain meaningful.

The model-based procedure is creates a physical model of the system using mathematical tools such as dynamic or differential equations that describe the physical phenomena of the system. The model-based approach could become complex depending on the degree of detail of the system. There is usually a trade-off between the coverage and the accuracy of the model. However, usually it is possible to combine these two methods to get the advantages of both of them, and not to rely only on one. However, accurate data and a long lead-time to obtain results can be a disadvantage; therefore, care needs to be taken when deciding upon which methodology to employ.

To be successful, technology such as sensors are hugely beneficial. Intelligent devices can monitor manufacturing systems to schedule agile service intervals, by utilising Big Data algorithms, which help you, predict future service issues to prevent customer downtime. The aim is to create an expanded network of condition monitoring sensors, which when combined with digital technologies such as advanced analytics and connectivity, will enable manufacturing plants to collect and aggregate data from disparate, geographically distributed plants to produce, merge, store and analyse information on specific equipment to detect anomalies and trends.

IV. Big Data in Maintenance

Big Data in Maintenance has started finding numerous applications. These are in the fields of maintenance of railway tracks (Thaduri et al., 2015), Oil and gas sector (Perrons and Jensen, 2015), maintenance of electric grid (Jaradat et al., 2015), etc. There are a number of maintenance applications utilising big data available within the current literatures. However, in this section, two new

application areas of big data in asset management have been discussed.

Asset Information. Complex assets are associated with large amounts of manufacturing and maintenance data that need to be interpreted and presented in the form of information. This information becomes a foundation to achieve sustainable and safe performance of these complex systems through the life cycle (Whyte et al., 2016). The asset information details include the provenance, part types and serial numbers, design life, maintenance schedule, and design rationale for sub-systems or components. As data is reused across the life cycle, sets of data and information become combined and can be mined, interpreted and used in new ways. Big data and new analytical techniques can have a great impact on how the asset information is stored, analysed and presented in order to achieve positive impact on the maintenance. Baselines are often used to provide an agreed description of one, or a number of, assets at a point in time (Whyte et al., 2016). Such configurations can be stored in the big data storages and can be used to supervise any modifications that may occur later.

V. Discussion

The recent advancements in big data techniques and technologies have enabled many organisations to collect and analyse big data efficiently to support a range of manufacturing initiatives. There needs to be caution exercised in our endeavours to chase the 'big data' hype that has been created in the recent times; mostly by those who have something to sell. The transformational effects and dividends of big data can sometimes be overestimated and overstated. The data has always been increasing even before the term 'Big Data' came into use. So the term Big Data is a rhetorical nod to the reality that "big" is a fast-moving target when it comes to data (Lohr, 2012). How 'big' 'big data' is varies with time, industry, type of data and numerous other factors. The focus of 'Big data' must shift from the data to uses of the data and insights that it can provide. Most of the large-scale Computerised Maintenance Management Systems (CMMS) and CBM initiatives have failed not because of any lack of data. The failure emanates from incorrect analysis being conducted on wrong data. This leads to incorrect diagnosis and hence wasteful expenditure in maintenance actions. This has a negative effect on the motivation of the workforce as well as the management to pursue such initiatives.

According to Harford (2014), big data is being sold today with four claims:

1) It provides accurate results
2) All data points can be captured
3) Old statistical sampling techniques are now obsolete
4) Statistical correlation can substitute causation

These claims have been made because of the high dimensionality of big data that also enables in understanding heterogeneity better. The data is gathered from many sources and can take into account those populations as well that were earlier treated as 'outliers'. Nevertheless, we must be cautious. *High dimensionality* also brings in spurious correlation, referring to the fact that many uncorrelated random variables may have high sample correlations in high dimensions. *Spurious correlation* may cause false scientific discoveries and wrong statistical inferences (Fan et al., 2014). Large number of data sets and testing many parameters can also lead to accumulation of individual estimation errors; also called as *noise accumulation.* It is easier and cheaper to establish statistical correlation between events than to find the cause of the event. This can often be misleading because "If you have no idea what is behind a correlation, you have no idea what might cause that correlation to break down" (Harford, 2014).

Big Data Analytics in asset maintenance has a bright future. The application of big data can lead to revolutionizing the way we do maintenance. However, there is a need to be careful in what we actually require. The focus must remain on effective insights that data can provide and not on how much data can be collected. As the amount of data being gathered today is increasing, the mistakes of statistical inferences will also become bigger.

Computational approaches that will allow manufacturing organisations to collect store, rapidly integrate, and compare a number of different datasets are required. This will require the design of automated tools that are intuitive and allow organisation to reuse large amounts of data, often readily available but unused. The peculiar features of Big Data make traditional statistical methods invalid. The need to develop new robust methods of data analysis will become the most important field in the future.

Future technologies will create smart factories operated by smart production systems with the capacity to remember large amounts of past production data. This will create a system where organisations move from fixed maintenance intervals, towards specific tailored predictive maintenance based upon accurate past and current data support by intuitive data mining systems.

Big data and analytics combined with condition monitoring systems can support maintenance planning

References

[1] Raheja, D, Llinas, J, Nagi, R, & Romanowski, C 2006, 'Data fusion/data mining-based architecture for condition-based maintenance', International Journal Of Production Research, 44, 14, pp. 2869-2887, Business Source Premier, EBSCOhost, viewed 9 February 2017.

[2] ISO 13373-1 Condition monitoring and diagnostics of machines -- Vibration condition monitoring -- Part 1: General procedures

[3] C.L. Philip Chen, Chun-Yang Zhang, Data-intensive applications, challenges, techniques and technologies: A

SURVEY ON BIG DATA, INFORMATION SCIENCES, VOLUME 275, 10 AUGUST 2014, PAGES 314-347

[4] MICHAEL WALKER, BATCH VS. REAL TIME DATA PROCESSING (2013). RETRIEVED FROM HTTP://WWW.DATASCIENCECENTRAL.COM/PROFILES/BLOGS/BATCH-VS-REAL-TIME-DATA-PROCESSING

[5] ZHANG, L AND KARIM, R. BIG DATA MINING IN EMAINTENANCE: AN OVERVIEW, PROCEEDINGS OF THE 3RD INTERNATIONAL WORKSHOP AND CONGRESS ON EMAINTENANCE, 2014, 17-18, LULEÅ, SWEDEN

[6] HASHEM, I. A. T., YAQOOB, I., ANUAR, N. B., MOKHTAR, S., GANI, A., & KHAN, S. U. (2015). THE RISE OF "BIG DATA" ON CLOUD COMPUTING: REVIEW AND OPEN RESEARCH ISSUES. INFORMATION SYSTEMS, 47, 98-115. DOI:10.1016/J.IS.2014.07.006.

[7] WU, X., KUMAR, V., QUINLAN, J. R., GHOSH, J., YANG, Q., MOTODA, H.,& ZHOU, Z. H. (2008). TOP 10 ALGORITHMS IN DATA MINING. KNOWLEDGE AND INFORMATION SYSTEMS, 14(1), 1-37.

[8] BARLOW, M., (2013), REAL-TIME BIG DATA ANALYTICS: EMERGING ARCHITECTURE, O'REILLY MEDIA, INC., 1005 GRAVENSTEIN HIGHWAY NORTH, SEBASTOPOL, CA, 95472.

[9] LOHR, S., 2012, HOW BIG DATA BECAME SO BIG. NEW YORK TIMES, AUGUST 11. ⟨HTTP://WWW.NYTIMES.COM/2012/08/12/BUSINESS).

[10] HARFORD, T., 2014, BIG DATA: ARE WE MAKING A BIG MISTAKE? FINANCIAL TIMES, MARCH 28. (HTTP://WWW.FT.COM/INTL/CMS).

[11] FAN, J., HAN, F. AND LIU, H., (2014), "CHALLENGES OF BIG DATA ANALYSIS", NATIONAL SCIENCE REVIEW, VOL. 1, 2014, PP. 293 – 314.

[12] PERRONS, R.K. AND JENSEN, J.W., (2015), "DATA AS AN ASSET: WHAT THE OIL AND GAS SECTOR CAN LEARN FROM OTHER INDUSTRIES ABOUT BIG DATA", ENERGY POLICY, VOL. 81, PP. 117–121.

[13] JARADAT, M., JARRAH, M., BOUSSELHAM, A., JARARWEH, Y. AND AL-AYYOUB, M., (2015), "THE INTERNET OF ENERGY: SMART SENSOR NETWORKS AND BIG DATA MANAGEMENT FOR SMART GRID", PROCEDIA COMPUTER SCIENCE, VOL. 56, PP. 592 – 597.

[14] THADURI, A., GALAR, D. AND KUMAR, U., (2015), "RAILWAY ASSETS: A POTENTIAL DOMAIN FOR BIG DATA ANALYTICS", PROCEDIA COMPUTER SCIENCE, VOLUME 53, PAGES 457–467.

[15] TURBAN, E., SHARDA, R., & DENLEN, D. (2011). DECISION SUPPORT AND BUSINESS INTELLIGENCE SYSTEMS (9TH ED.). UPPER SADDLE RIVER, NJ: PEARSON PRENTICE HALL

[16] WHYTE, J., STASIS, A. AND LINDKVIST, C., (2016), "MANAGING CHANGE IN THE DELIVERY OF COMPLEX PROJECTS: CONFIGURATION MANAGEMENT, ASSET INFORMATION AND BIG DATA", INTERNATIONAL JOURNAL OF PROJECT MANAGEMENT VOL. 34, PP. 339–351

[17] LEE, J, RAMJI, A, ANDREWS K. , DARNING L, DRAGAN, B., 2004. AN INTEGRATED PLAT-FORM FOR DIAGNOSTICS, PROGNOSTICS AND MAINTENANCE OPTIMIZATION, IN E-PROCEEDINGS OF INTELLIGENT MAINTENANCE SYSTEM, ARLES, FRANCE, 15-17 JULY