EAI Endorsed Transactions
on Scalable Information Systems

Research Article EAI.EU

Proposing a streaming Big Data analytics (SBDA) platform for condition based maintenance (CBM) and monitoring transportation systems

Jamal Maktoubian[1],*

[1]International School of Information Management(ISIM), University of Mysore, India

Abstract

Statistics demonstrate that public transportation plays a significant role in people's movement in metropolises. However, transit systems are aging and are facing rising maintenance costs. Technologies such as Condition-Based Maintenance (CBM) could be used in order to monitor performance conditions of transportation and industrial assets in real-time to detect when and what maintenance is required. CBMs could help to identify risk scenarios in real-time, enhance reliability, reduce call out costs, increase productivity, and better asset functioning visibility. Since the high volume of maintenance data is generated from the different source, managing assets conditions with traditional inspection system such as planned maintenance (PM) is impossible. Therefore, providing a comprehensive performance management program is essential. My research is motivated by interesting challenges increasing from the growing size, variety, and complexity of maintenance data in CBM systems. This paper presents a knowledge-based approach of CBM using streaming big data analysis (SBDA) in order to solve real-time big data management, storage and computation challenges and predictive data analytics in CBM systems. This platform could detect changes in asset's behaviour before they stop.

Keywords: Condition-Based Maintenance(CBM), Apache Kafka, Spark Streaming, Real-Time Data, Streaming Big Data Analysis (SBDA), Maintenance Management.

Received on 26 March 2017, accepted on 23 April 2017, published on DD 28 June 2017

doi: 10.4108/_______________

1. Introduction

Chicago's transit system has been operated by Chicago transit authority (CTA) since the early 1900s. CTA's bus system has about 1,880 buses that support 130 routes and over 1,300 route miles of track. Totally, Buses make approximately 18,840 journeys a day and cover more than 10,800 bus stops[1]. Recently, encouraging people to use public transportation become critical for authorities as cities become more and more crowded, and greenhouse gas emissions (GHGs) is becoming a major problem all over the world. Officials have to find some solutions to improve transportation services in order to reduce the commuting and promote reliability and quality of public transportation. Recently, governments are spending a huge amount of money on research fields such as intelligent transportation system (ITS), smart city, traffic engineering, transportation engineering, and to name but a few in order to improve this sector[2]. Since the number of complexes, fast-flowing, unstructured, heterogeneous and high-dimensional sources such as sensors, devices, applications and infrastructures which are generated in the

[*]Corresponding author. Email:jamal.maktoubian@gmail.com

transportation sector have been increased dramatically; utilizing data analytics would be important to vibrant and thriving cities and improve user satisfaction[3].

Every day around 1.6 million trips are recorded by CTA on the transit system and collect a vast amount of valuable data on travel routes. Train and buses are tracked and control in real-time which is called Automatic Vehicle Location technology (AVL). When this trend mixed with the Internet and smart devices, it could lead to better performance in transit accessibility, produces customer alerts, allow users to track a bus in real-time, traffic reduction, reducing vehicle trips and traffic congestion. However, this is not the only fact that could help cities have high-quality public transportation; Transportation asset management (TAM) is the other factor should be considered, as assets require a systematic process of operating, maintaining, and upgrading in their life cycle. In the past, different techniques such as Reliability-Centered maintenance (RCM), Time-based maintenance, and Condition-Based maintenance are utilized for monitoring assets condition. Recent Progress in producing advance devices and sensor technology, we are able to detect and predict failure more accurately than before. However, the major problem in this type of predictive maintenance system is making a massive amount of real-time data[4] which should be collected, processed, stored, and analyse with special infrastructure and techniques. Combination of CBM systems and big data analytics might be the best solution to tackle the problem for extracting valuable information such as detection of Components and equipment wear, breakage and the estimation of remaining Components life.

This paper proposes a framework in order to solve real-time big data management, storage, computation challenges and predictive data analytics in CBM systems in order to predict and monitor changes in components' behaviour before they stop. To deal with real-time data, Apache Kafka[5] is used as a distributed messaging system to collect unstructured and semi-structured data. Collected data are delivered to Spark Streaming which is a distributed stream processing engine. In this stage, spark streaming breaks up the input data stream in small batches namely Resilient Distributed Datasets (RDD)[6]. Continuous sequence of RDDs is called DStream (discretized stream) which pass through spark engine in order to be processed and could be used in machine learning libraries such as MLlib[7] for data analysis. Visualization is critical part of this system as users should be able to monitor the environment and tracking assets [8].

2. Condition-Based Maintenance (CBM)

The continuous massively parallel data generated by various sources are convincing different sectors to think about huge opportunities in this area. The CTA has been utilizing Global Positioning System (GPS) devices in order to track buses and provide useful information for Chicago's citizen. Implementing predictive maintenance management program on transit systems such as CTA require powerful big data analytics platform as data which are produced by embedded sensors and other equipment are extremely vast and complex. The program is not simply stated for some operating condition such as vibration monitoring, thermal imaging, and lubricating oil analysis. While, efforts to find wear, imbalance, misalignment, and breakage components and equipment before they become critical. Clearly, preventive maintenance program is more advantageous than unscheduled destruction.

Intelligent CBMs are systems which could understand and making decisions without operator's intervention. It is a set of actions based on monitoring real-time and near-real-time maintenance data which is collected from embedded sensors and devices. Sensors and data collectors should be connected to the analytic platforms using wireless technologies. Smart sensors play a vital role in this type of system as they are reprogrammable and designed to analyse recovered sensory data. The sensors collect maintenance data from different part of assets such as control and communication system which most likely will be monitored. Intelligent system should be design in order to perform analysis of the measured data. Therefore, Artificial Intelligence techniques and algorithms such as Neural Networks, Fuzzy Logics, Case Based Reasoning, and Expert Systems need to be taken into account.

3. CBMs and Big-Data Analytics

Since the last decade, there has been an exponential growth in the volume of streaming data which are produced from several sources to measure some characteristics of the environment [9]. These data ought to be collected, processed, stored, and analysed[10]. Gradually, a vast amount of data in short periods of time is produced in CBM systems which is called Big data streaming (BDS) [1]. Analysing these data has provided great opportunities for various sectors in the different fields to minimize maintenance cost and avoid unexpected failures long before they happen.

Although there are plenty of advantages in Processing and analysing BSD in CBMs system, this could make series of problems as using traditional techniques, infrastructures, and databases are costly. So, it requires special frameworks, algorithms, programming languages, and databases which provide the data scalability and availability. To deal with real-time events, distributed messaging system such as Kafka and flume is used to collect the generated data from various sources. The other important factor which needs to be taken into account in this type of system is BSD that should be processed as soon as they received in CBMs' system. Apache Spark presents distributed processing framework that offer a data structure for in-memory computations on large clusters. In the next stage, processed data should be store in an appropriate database; Databases should have the ability to distribute data over many nodes automatically to provide data accessibility attribute to the system. No-SQL databases are much more desirable for big data analytics[11]. They are mostly open source and could be divided into different categories, for instance, Key-Value Stores, Document Stores, Wide-Column Stores, and In-memory databases. Finally, data need to be served for detecting valuable information such as assets wear, breakage and the estimation of remaining Components life.

4. Related Work

In this section we survey the recent work in analytics of maintenance data and related research in condition-based maintenance in the area of smart transportation for failure and wear detection. Maintenance is introduced as technical actions taken to refit the required functionality of component in a system. It could be divided into unscheduled and preventive maintenance [12]. There are numbers of tools, which use various techniques to monitor where attributes are replicated within clusters using multicast methods and aggregated via a tree structure[13]. Snort[14], Bro[15] and Tstat[16] are some of the tools to monitor network data. According to the explosive advance of various devices and sensors, the research attention given to IOT, big data and analytics are also increased. Therefore, recent tools are unable to scale the sizes required in data centre systems. As a result, devices could collect various data and send it through the net [17]. Researchers bring up big data technologies like Apache Spark as a computing platform for analytics. They represent a survey of several techniques that could be considered in the data analytics process. Particularly, various big data processing techniques have been proposed in order to use for smart grid technology. Emanuele Fumeo[1] presented a Big Data platform for Condition Based Maintenance (CBM) in the context of Rail Transportation Systems (RTS) based on real-time, high velocity data. The current platform, tools, and techniques are inflexible and point only some particular issues. Moreover, finding assets failure and their reduction performance will evaluate the future status of the monitored assets in order to reduce risks related to failures and to avoid service disruptions.

4. Proposed Framework

There have been research and studies on a variety of methodologies for the storage, processing and analysing of large amounts of sensor data. As relational database model, which is most frequently used in storing data, suffers from the reduction of their performance and needs more processor speed and memory capacity as the volume of data to be processed grows, the costs of additional devices like servers are also increased. To overcome this issues, studies were done to decrease the problems with using distributed processing like Apache Spark, which use Hadoop distributed file system, a large distributed file system. When we talk about big data, it should have some characteristics such as volume, velocity and variety[18] which could not be collected, stored, computed with normal systems. Data complexity is a characteristic was added to Vs of Big data when CBMs'data is considered. Since captured data in maintenance systems are high dimension, nonlinear, and poor, This could cause many critical issues such as security problems [19]. In this section, we present system architecture for collecting, processing, and analysing the sensor data in order to monitor CBM systems. Figure 1 illustrates the overall architecture of system for real-time processing of machine sensor data proposed in this study. The propose architecture should have capability to consider these properties and find appropriate solution for analysing real-time signals in transportation systems. The architecture consists of four main integrated sections which should work together simultaneously:

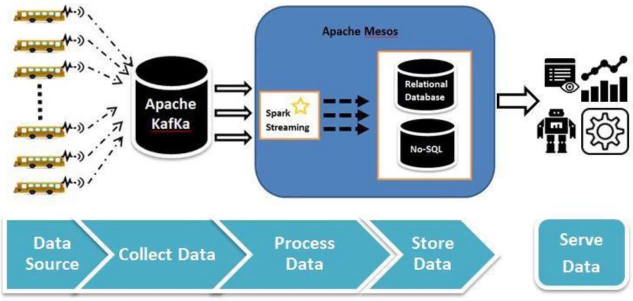

Figure 1. Architecture

5.1. Data Source

The wide range of data are produced by the different type of sources every second such as real-time condition data created by sensors or equipment, Pictures, audios, videos, web pages and other documents (Word, PDF, etc.), and manufacturing component information from different parts of system, to name but a few.

Structured data are mainly stored in relational databases (such as Oracle, DB2, SQL Server, etc.) which could be

seen in figure-1. Unstructured CBM data consist of technical documentation (TD) of tools and component, photos taken by Thermographic cameras[20], frequency spectrums collected by vibration spectrum[21], etc. Semi-structured data are primarily text-based, tags or other types of markers are utilized to depict certain elements. In this category, data are deal with emails, XML files, and log files, etc.
Big Data can be described by the following characteristics:

- Volume: First quality after hearing the term “Big data” come to our mind is the size of data. In systems that deal with maintenance data, all prior and afterward (life cycle) maintenance data ought to be considered as invaluable data[22].
- Velocity: Velocity[23] reflects the rate of data which is created in high speed. In CBM system, sensors and other equipment, transfer high-speed real-time data to the system which could to be collected and processed as soon as they arrived.
- Variety: It refers to maintenance data in different types and forms, for example, wireless sensor networks(WSN), image, media, and log files and so on. These data could be structured, Semi-structured and unstructured data[24].
- Complexity: Generally, maintenance data which are created by CBMs systems are complex. It deals with quality, accuracy, certainty and many other factors of data which need to be tackled with different techniques.

5.2. Collect Data

Real-time data could be ingested from data sources like HDFS directories, Kafka, Flume, Twitter, ZeroMQ, Kinesis, TCP sockets, [25] etc. LinkedIn developed Kafka as a messaging system to work as the data source for their real-time data and support data processing pipeline. It is built to be fault-tolerant, high-throughput, horizontally scalable, distributing data streams and processing[5]. Real-time data are converted into key-value data with timestamp and pass through spark streaming. The architecture of Kafka is illustrated in Figure-2. Kafka has three main components namely message producer, message consumer and message broker that all could be applied on a cluster of machines that work as a logical group. Apache Kafka in some cases is considered as a core for data stream processing.

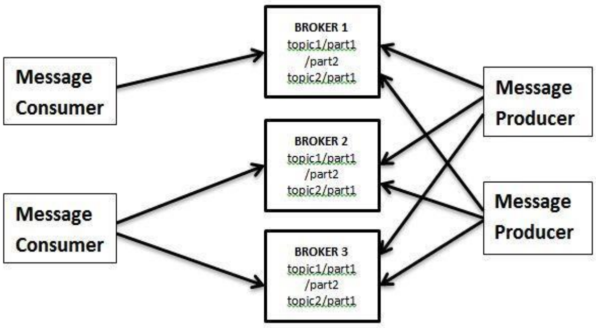

Figure 2. Kafka Architecture [2]

5.3. Processing Data

In this stage, Spark Streaming is considered as a part of the core Spark API. It is scalable, high-throughput, fault-tolerant stream processing of real-time data. The data is ingested into Spark streaming from multiple sources like Kafka, Flume, Kinesis, or TCP sockets. Spark Streaming is able to execute some complex functions such as map, reduce, join and window on received data. It can process real-time data using GraphX, Spark SQL, DataFrames, and utilizing machine learning techniques from MLlib.
When collected data are received by spark streaming, they convert into effective storage called resilient distributed datasets (RDDs)[26]. Arrays of RDDs could make discretized stream or D-Stream groups. DStreams are extremely suitable for different complex operations such as Map, Reduce, Join, etc.

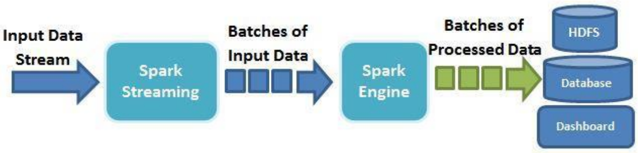

Figure 3. Apache Spark[12]

5.4. Store Data

Due to the recent advances in CBM, the numbers of devices and the volumes of data have been expanding with remarkable rates. Immense computational and storage costs require to collect, process, analyze and store with a relational database. As the received data are varied and huge, traditional databases cannot provide high availability, performance, and scalability for big data. Therefore, No-SQL databases are considered which could store, retrieve and distribute data over many servers[27].
Processed data (RDDs) can push out to external data storage system like Hadoop distributed file system (HDFS), or No-SQL databases such as Cassandra, MongoDB, and HBase, etc. As the volumes of data is extremely large, we cannot store all them in our storages, so they should be scored based on entry time. Obviously, invaluable data are real-time and near real-time data and based on CBM system we should eliminate useless data.

However, some information like components information, installation date, age, manufacturer, and grade, etc. are stored in relational databases and use in analytics stages.

5.5. Cluster Manager

Apache Mesos is an open source, centralized, fault-tolerant cluster manager, provided efficient resource isolation for distributed computing environments. Spark standalone, Apache Mesos[28] and Hadoop YARN are three cluster managers supported by Apache Spark. When SparkContext needs resources for executing jobs, it could communicate with one of these cluster managers. Apache Mesos could combine physical resources and design schedule cluster frameworks, it consists of two major framework components namely executor, and scheduler. Using the scheduler, system would be able to know what to do with resources and the executor is designed to runs framework tasks.

5.6. Serve Data

Machine learning technics could facilitate decision-making automatically without human interaction. Serve data combines many techniques from various disciplines, including machine-learning, statistical analysis, mathematical modelling, and Data mining which could be applied to evaluate the operation condition of the components and tools. Traditional machine learning frameworks such as WEKA[29] and RapidMiner[30] could not scale to big data as their memory constraints. Apache mahout is a distributed machine learning framework which is licensed under the Apache Software Foundation[31]. Since in real-time processing specially in CBM systems quick action is required, Apache Mahout is not really best choice as data need to be written to the disk and read from it.

Spark MLlib is an in-memory-based distributed machine-learning library, consists of various machine learning algorithms in order to do Filtering, prediction, Clustering, and Classification. The MLlib is much more desirable for applying real-time machine learning algorithms especially when Spark streaming is used for processing data. Processed data also could be pushed out to live dashboard for monitoring reason.

6. Conclusion

In this short paper, we propose a CBM platform sensing technologies to monitor big data stream which are generated by several devices like sensors, camera, and Microphones. This platform is capable of making decision automatically without human control. This technique enables us to mix structured, Simi-structure and unstructured data in order to monitor all components of real-time systems. Using appropriate technologies like Apache Kafka, Mesos, Spark streaming and HDFS could give us this ability to manage and respond to the system failures in exact time. In future, we will present next versions of the framework with use of Spark Streaming + Kafka Integration which is new combination of these technologies.

References

[1] Fumeo, E., L. Oneto, and D. Anguita, Condition based maintenance in railway transportation systems based on big data streaming analysis. Procedia Computer Science, 2015. 53: p. 437-446.
[2] Kitchin, R., The real-time city? Big data and smart urbanism. GeoJournal, 2014. 79(1): p. 1-14.
[3] Zhang, L., Big data analytics for emaintenance: Modeling of high-dimensional data streams. 2015, Luleå tekniska universitet.
[4] Rusitschka, S. and E. Curry, Big Data in the Energy and Transport Sectors, in New Horizons for a Data-Driven Economy. 2016, Springer. p. 225-244.
[5] Garg, N., Apache Kafka. 2013: Packt Publishing Ltd.
[6] Zaharia, M., et al., Discretized Streams: An Efficient and Fault-Tolerant Model for Stream Processing on Large Clusters. HotCloud, 2012. 12: p. 10-10.
[7] Meng, X., et al., Mllib: Machine learning in apache spark. Journal of Machine Learning Research, 2016. 17(34): p. 1-7.
[8] Gubbi, J., et al., Internet of Things (IoT): A vision, architectural elements, and future directions. Future generation computer systems, 2013. 29(7): p. 1645-1660.
[9] Akyildiz, I.F., et al., Wireless sensor networks: a survey. Computer networks, 2002. 38(4): p. 393-422.
[10] Dumbill, E., A Revolution That Will Transform How We Live, Work, and Think: An Interview with the Authors of Big Data. Big Data, 2013. 1(2): p. 73-77.
[11] Berger, M.L. and V. Doban, Big data, advanced analytics and the future of comparative effectiveness research. 2014.
[12] Yardley, R.J., Impacts of the fleet response plan on surface combatant maintenance. Vol. 358. 2006: Rand Corporation.
[13] Massie, M.L., B.N. Chun, and D.E. Culler, The ganglia distributed monitoring system: design, implementation, and experience. Parallel Computing, 2004. 30(7): p. 817-840.
[14] Roesch, M. Snort: Lightweight intrusion detection for networks. in Lisa. 1999.
[15] Bro. Nov 2014; Available from: http://www.bro-ids.org.
[16] Finamore, A., et al. Live traffic monitoring with tstat: Capabilities and experiences. in International Conference on Wired/Wireless Internet Communications. 2010. Springer.
[17] Kawamoto, Y., et al., Internet of things (IoT): Present state and future prospects. IEICE TRANSACTIONS on Information and Systems, 2014. 97(10): p. 2568-2575.
[18] Montgomery, D.C., Big data and the quality profession. Quality and Reliability Engineering International, 2014. 30(4): p. 447-447.
[19] Beyer, M.A., et al., Big Data's only the beginning of extreme information management. Gartner report G, 2011. 211490.
[20] Wong, W.K., et al. An effective surveillance system using thermal camera. in Signal Acquisition and Processing, 2009. ICSAP 2009. International Conference on. 2009. IEEE.
[21] Khwaja, H.A., S. Gupta, and V. Kumar, A statistical approach for fault diagnosis in electrical machines. IETE Journal of Research, 2010. 56(3): p. 146-155.

[22] Levrat, E., B. Iung, and A. Crespo Marquez, E-maintenance: review and conceptual framework. Production Planning & Control, 2008. 19(4): p. 408-429.

[23] Chen, C.P. and C.-Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences, 2014. 275: p. 314-347.

[24] Warren, P. and N. Davies, Managing the risks from information—through semantic information management. BT Technology Journal, 2007. 25(1): p. 178-191.

[25] Namiot, D., On big data stream processing. International Journal of Open Information Technologies, 2015. 3(8): p. 48-51.

[26] Zaharia, M., et al. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. in Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation. 2012. USENIX Association.

[27] Han, J., et al. Survey on NoSQL database. in Pervasive computing and applications (ICPCA), 2011 6th international conference on. 2011. IEEE.

[28] Apache Mesos. Available from: http://mesos.apache.org/.

[29] Holmes, G., A. Donkin, and I.H. Witten. Weka: A machine learning workbench. in Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on. 1994. IEEE.

[30] Klinkenberg, R., RapidMiner: Data mining use cases and business analytics applications. 2013: Chapman and Hall/CRC.

[31] Owen, S., et al., Mahout in action. Greenwich, CT. 2011, USA: Manning Publications Co.