DERIVING INTELLIGENT DATA ANALYTICS USING ANOMALY DETECTION FRAMEWORK FOR IOT NETWORK AND SMART ENVIRONMENTS

Anil Lamba[1], Satinderjeet Singh[2], Balvinder Singh[3], Natasha Dutta[4], Sivakumar Sai Rela Muni[5]
Department of Computer Science, Charisma University, Turks and Caicos Islands

Abstract: Data Analytics is by far the component with more added value in Internet of Things (IoT) networks. One aspect of data analytics is anomaly detection within data points received in some cases in real time that help to conduct predictive maintenance, weather monitoring or cyber security forensics for instance. Although there exists a number of web dashboards that allow IoT users to visualize data in time domain and perform statistical analysis, anomaly detection is often absent else if present not that straightforward, reliable and accurate. The development and implementation of Anomaly Detection Engine (ADE) poses a number of challenges that are in fact addressed in this paper. The research work exposes the multifaceted aspect of IoT networks and applications based on real life use cases and the difficulties engendered in mounting an ADE from both software system engineering and network convergence perspectives. Moreover a comparative description of diverse time series models adopted in anomaly detection is undertaken. It was noticed that there is neither one size fit all solution nor a plug n play alternative and that the unsupervised mode in machine learning as a model for time series analysis is the most versatile and efficient technique for IoT analytics developers.

Keywords: Time Series; Data Analytics; Machine Learning; IoT Networks; Sensor Fusion; Smart Cities; Anomalies; Anomaly based algorithm; Classification algorithms; Data communication; Denial of service attack; Intrusion detection; Cyber Security; Cloud Security; Network ; Cyber; Cyber Threats; Threat Analysis ; Information Security; Data security.

Citation: Anil Lamba, 2017."DERIVING INTELLIGENT DATA ANALYTICS USING ANOMALY DETECTION FRAMEWORK FOR IOT NETWORK AND SMART ENVIRONMENTS", International Journal for Technological Research in Engineering, Volume 4 Issue 6, pp.5682-5686, 2347-4718.

I. INTRODUCTION

The Internet has evolved from its original aim of providing access to web resources globally to what is commonly called today Internet of Things, where it is expected that objects will internetwork and have a presence on the Internet just with an IPv6 address for example. The objects market is estimated in billions and trillions, very far away from the globe human population.

This has led to new business models with development of dedicated IoT networks like SigFox, LoRa, Symphony Link and NB-IoT and production of IoT compliant devices from microcontrollers' manufacturers such as Microchip, Intel, and Raspberry PI. Software companies have come up with virtual machines and statistical tools for big data analytics whereas network devices constructors like Cisco and Juniper for instance have come up with network gateways and routers

to accommodate devices connection, routing and IoT data transit. The myriad of technologies involved within the IoT ecosystem should empower smart environments as it happens likewise in smart cities. Section depicts the multiple use cases of IoT in smart environments whereby anomalies arouse and need to be identified. Moreover the importance ADE within the data analytics umbrella is emphasized within the IoT value chain. In section III the challenges pertaining to the network infrastructure and convergence are illustrated. The categories of anomalies and a framework to mount an ADE are presented in Section IV including the intricacies in dealing with the data points in the time domain, different time series models for ADE are discussed with specific attention to the unsupervised mode in machine leaning. Finally the paper provides a conclusion on the key points surrounding the challenges and proposes future work.

II. SMART ENVIRONMENT USE CASES

2.1 IoT Value Chain

As per the IoT value chain depicted in Fig. 1, it is understood that there need to be chips and devices that can be integrated onto IoT networks per se, from there data can be channeled to Web of Things (WoT) and Machine to Machine applications. According to Cisco's annual Visual Networking Index, machine-to-machine (M2M) connections that support IoT applications will account for more than half of the world's 27.1 billion devices and connections by 2021.

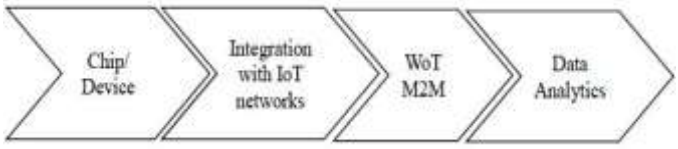

Fig. 1. IoT Value Chain.

However, data analytics provide real value added services to IoT as a business key differentiator, more than key technology enablers such as SDN, IPv6 & 5G. Beyond process automation analytics as a service is current on the cloud. One aspect of data analytics happens to be anomaly detection that is discussed in this paper in the context of

smart environments IoT applications.

2.2 Anomalies in Smart Environments

Smart cities are currently present in economic development plans across the globe. With different IoT use cases in smart cities, there is a market for anomaly detection APIs. Table I shows a few IoT use cases with their anomalies descriptions and benefits in tracking those anomalies.

Smart/	Anomalies in Smart City IoT Use Cases	
	Anomalies Description	*Benefits*
Water	Water leakages	To prevent water waste
Lighting	Broken Bulbs	Save time and fuel for maintenance
Home	Gas leakage	Alert home users on the incident
Building	Electricity peak and Pipe leakage	Energy monitoring
Farm	Anomalies in farm data and weather	Monitor growth
Goods	Traffic congestion spots	Optimize route and delivery

Table 1. Anomalies in Smart Applications.

Streaming data from the different utilities expressed in Table 1 requires an appropriate IoT network architecture. IoT networks forward packets of around 3 bytes and a number of limited messages per device per day. Less than 10% of the 30 billion nodes by 2020 across the world will be connected to the Internet via cellular networks, thus the dire need for IoT networks. Fig. 2 illustrates the physical architecture for an IoT network from data analytics perspective.

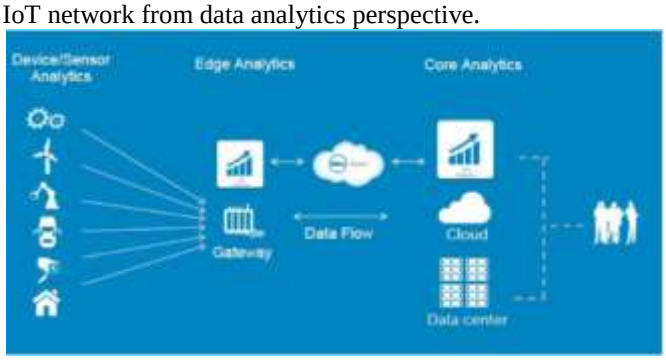

Fig. 2. IoT Architecture

2.2.1 Where to run Analytics?

In any development of an ADE, it is important to understand and identify where to place the engine within an IoT network, bearing in mind that many IoT network operators present their network as a global network. Elements to be considered are computational resources, the need for cleansing and filtering of data before running analytics, the choice of moving application to data in case of big data analytics, the choice of running analytics at the Business Intelligence (BI) layer in an IoT network [1]. Examples of analytics' platforms on the market are IBM Watson Analytics, Python, and R language. Academics usually use Mathematica, Matlab, and other Time Series Analysis (TSA) tools. Machine Learning Deep Learning slowly but surely making their way too.

Device/Object Computing

The physical and computational limitations of IoT devices (sensors with microprocessors or smartphones) make it hard to execute complex analytics at the device layer. Besides memory and battery life cycle, Device Identification (DI) is another challenge to IoT networks and device computing. The tendency is to adopt the MAC address concept coupled with IPv6.

Mobile Code Offloading (MCO)

With the rapid depletion of batteries on the smartphones and sensors, mobile code offloading is proposed in situations where it takes lesser power to transmit executable code at run time on a cloud or a cloudlet, the latter being a decentralised cloud. In the context of IoT, MCO is relevant only when the ultimate device is a smartphone, as sensors will normally transmit a few bytes. Another technique consists of cloning the VMs to the cloud instead of moving only the executable files [2].

In-memory Analytics

This is technology initiated by SAP whereby queries happen on unidimensional data in the memory bypassing round trips to the hard disk. Again it makes sense only if the end device has a considerable primary storage.

Edge Computing

The rationale behind Edge Computing is to filter unwanted data at the level of an edge router closer to the end devices before actually running the analytics on a cloud. Cisco views the network as a platform andimagine the analytics can thus take place at different points on a network with the resources on the network with a difference between creating analytics model and executing analytics model [3].

2.3 Protocols, Gateways and APIs

One major consideration in the design of an ADE is the choice of protocols for the gateways and APIs given the heterogeneous nature of the wireless communications environment. If we compare the IoT protocols with the TCP/IP suite, the Physical layer is almost the same with standards such as IEEE 802.11/ 802.15.4/ 802.3, GSM and others. At the network layer over and above IPv4 and IPv6, the gateways should be compliant with 6LoWPAN. The transport layer for IoT is UDP or TCP whereas the application layer comprises lightweight protocols such as CoAP, MQTT, XMPP, more are described in the coming section.

2.3.1 Constrained Applications Protocol (CoAP)

Devices and sensors operate with constrained resources, thus CoAP a constrained application protocol that runs on UDP therefore allow broadcasting and multicasting depending on IPv4 or IPv6 addresses. It is in fact a client/server model where a sensor is the server and the user a client. Similar to HTTP or HTTPS it can modify states and is a one to one protocol. CoAP is not meant for TCP communications [4].

2.3.2 Message Queue Telemetry Transport (MQTT)

This is a lightweight protocol that supports TCP connection. As an open source protocol MQTT works with multiple clients via a broker. There are multiple versions of MQTT, for instance with Raspberry PI, Github [5] has specific libraries that facilitate interconnectivity with message exchanging. MQTT is not efficient with small packets protocols such as IEEE 802.15.4 [4].

2.3.3 Extensible Messaging and Presence Protocol (XMPP)
Based on XML this is a message oriented protocol. XMPP works without broker in a decentralized client/server model. It is highly scalable, supports thousands of nodes (Sensors & Users) concurrently. As communications are through messaging XMPP can assure security for different applications separately, namely through preference settings [4].

2.3.4 Advanced Message Queuing Protocol (AMQP)
Contrary to XMPP, Advanced Message Queuing Protocol is a brokered protocol with queuing capability. The broker provides all the discovery services. Both are message oriented protocols with features like security and reliability, however AMQP is not scalable as XMPP. AMQP was actually developed to overcome interoperability issues with proprietary protocols. There is no need for APIs, AMQP easily understands and converts the message format from another protocol as AMQP operates on the wire-protocol. As a messaging protocol AMQP supportspoint to point as well publish/subscribe paradigms, efficient for real time applications [4].

2.3.5 Java Message Service (JMS)
Like AMQP, Java Message Service works on messaging and publish/subscribe model, inefficient for more than one thousand nodes. JMS provides a platform for middleware and APIs. However JMS requires Java APIs built in Java Enterprise Edition. One drawback is that JMS APIs under different implementations may not interoperate [4].

2.3.6 Representational State Transfer (REST) APIs
REST is a famous protocol in web programming with HTTP. It is a stateless, point to point and publish/subscribe protocol. REST is applicable to IoT due to its simplicity and light weight nature as a request/reply protocol. Poorly scalable but with its caching capability, REST is still relevant for optimizing client/server communications in IoT and 5G cellular networks [4].

2.3.7 Data Distribution Service (DDS)
Data Distribution Service protocol is meant for programming large data centric applications. DDS is efficient for real time applications and it is highly scalable with all security features pertaining to data. It should be noted that DSS requires the wire-protocol to interoperate and data interchange [4].

III. ANOMALY DETECTION IN DATA ANALYTICS

In this section, we explore a proposed ADE framework, categorise the different types of anomalies in the time domain.

3.1 Anomaly Detection Engine (ADE) framework
From a software engineering perspective, a framework for the development of an ADE can be outlined as follows: Raw data → Time Series representation → Time Series Transformation → Model the time series→ Evaluation → Labelling → Alert

Raw Data normally comes in Binaries, CSV, TTL or JSON format [6] then a window captures the time series representation followed by a specific algorithm that will lead to evaluation, anomalies detection and labelling (mainly in machine learning supervised mode). In some situations namely for real time applications, detection is followed by alerts. In case of massive data, techniques like Hadoop and MapReduce are appropriate. As far as the transformation of the Time Series Java, Python, C, R [7] are common languages. Prior to any ADE development, it is important to understand the types of anomalies. These are presented in the next section.

3.2 Different Types of Anomalies
It is crucial to define clearly the criteria for anomalies given a particular IoT application as they can be subjective or contextual. The anomalies are classified into four classes from time series representations. As a matter of fact, visualization seems an easy way, nevertheless the more challenging anomalies are the ones unpredicted or without historical data. This is where AI algorithms are highly solicited.

3.2.1 Static vs Dynamic Anomalies
As depicted in Fig. 3, whenever there are some data points not following same patterns as the rest, dynamic anomalies occur. Such anomalies are relatively easy to detect based on existing pattern.

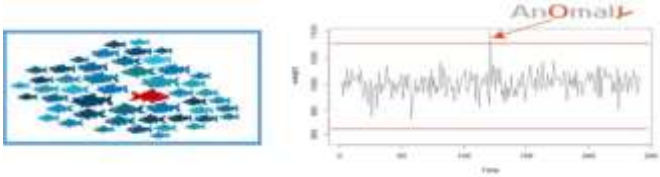

Fig. 3. Dynamic Anomaly, Fig. 4. Outlier anomaly [8]

3.2.2 Point/Outlier Anomaly
An outlier anomaly occurs normally if there is a surge or dip in the amplitude as illustrated in Fig. 4 time seriesshowing sugar bags weight on a conveyor belt with respect to time. In this case sugar bags < 920 g or >1080 g are considered to be abnormal [8]. It is important to not the anomalies are defined based on specific criteria. If the current criteria is modified the one in Fig. 4 may not be an anomaly. Thus an outlier is not necessarily de facto an anomaly.

3.2.3 Contextual Anomaly
Such anomalies are contextual, in other words what may be an anomaly in a particular situation is not necessarily true in another situation. For instance, as per Fig. 5 [9] the temperature t2 for June is an anomaly if we are looking at the northern hemisphere, Europe, however t1would be an anomaly if we are looking at the southern hemisphere, for example South Africa.

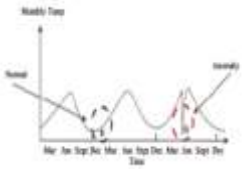

Fig. 5. Contextual Anomaly [9]

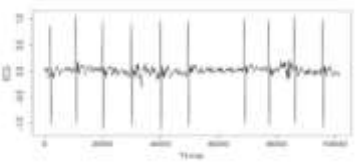

Fig. 6. ECG signal time series [9]

3.2.4 Collective Anomaly

Collective anomalies occur when there is continuous discrete anomalies in time. As it can be seen from Fig. 6 representing ECG signals in time. Another example is the case of jitter in network transmission, short delays tend to last on a longer period.

IV. TIME SERIES MODELS

It is always an advantage to have some clues about the type of anomalies expected before applying the right time series models that would yield an optimal result.

4.1 Time Series models:

There is actually no one size fit all solution for the development of an ADE as well as no de facto time series model that suit the ADE. Below are some of the popular time series models adopted for ADE in IoT.

- Autoregressive models: An autoregressive model specifies that the output variable depends linearly on its ownprevious values. It's based on an approach that several points from the past generate a forecast of the next point with the addition of some random variable, which is usually white noise. The Autoregressive Integrated Moving Average (ARIMA) is applicable to stationary time series only.
- Symbolic TSA [10]: Data points are converted to bits and bytes 10100111001, then we can apply InformationTheory; Shannon, Fast Fourier Transform (FFT) [11], DFT, DWT.
- Seasonal Trend Loss (STL) Decomposition: Data points together with the noise or multiple data sets over aperiod are decomposed and analysed to detect eventual anomalies [12].
- Machine learning: There are two main branches of machine learning namely supervised learning whereby thepattern for the anomaly is learnt and known, and unsupervised learning where detection is done by inference or featuring. The latter is more challenging as the anomaly pattern is unknown and the algorithm learn fromthe data points to be analysed. The supervised mode comprises the following methods: Artificial Neural Networks (ANN), Decision Table, Random Forest, K-nearest Neighbour, Support Vector Machine (SVM), Deep Learning, and Naive Bayes. The popular "unsupervised" algorithms are K-means clustering, Density-based spatial clustering of applications with noise (DBSCAN), N-SVM, Stream Clustering, and Latent Dirichlet Allocation (LDA).

4.2 Problematics:

Below are listed the ten main issues, some inherent to the IoT network others to the time series properties.

- Missing data points / holes: Missing data can happen due to device malfunctioning for instance or issuesrelated to device identification. It could also happen due to human failures or omissions, for example "Potent, climate warming gases are being emitted into the atmosphere but are not being recorded in official inventories" [13].
- Data corruption: For instance data can be corrupted due to external factors or device malfunctioning, thus itis important to ensure that the data points analysed are accurate and come really from the system under investigation.
- Encrypted data: In most IoT networks data is encrypted during transmission and normally decrypted forcustomer usage. If detection is to be performed on encrypted data, anomaly detection might not be straightforward.
- Sensor Fusion: Data points from different sensors can be aggregated for a specific function. For example,different parameters like temperature, carbon footprint, and wind speed can be captured from different sensors and merged for modelling on a server for environmental impact study. In such cases the TSA needs to deal with multiple datasets. Sensor fusion deals also with evolving sources.
- Real time detection: This is probably more inherent the network itself, but the processing and programmingaspects of the TSA is also determinant. Real time anomalies involve streaming data whereby historical data is matched with real time data [14].
- Seasonality: Also called as periodic time series, arrives when the time series is influenced by the seasonalfactors like day, night, month, and so on.
- Heteroscedasticity: It involves frequent changes in variances that can render the transform of the time seriesmore complex.
- Noisy data: Data points with very low amplitude can be drowned into the intrinsic transmission electronicnoise. Network equipment vendors are proposing edge computing routers that would actually clean the IoT device data in a closer location prior to run the complete analytics on the cloud.
- Traffic surge: At time there could be excessive throughput like number SMS on the eve of New Year thatcould bring an overload on the ADE.
- Non-linearity: Non stationary data points that are changing with time would require multivariate analysis.

V. CONCLUSION

This paper highlights the challenges relevant to core

elements involved in the development of an Anomaly Detection Engine (ADE). It was found that an accurate and reliable ADE relies on three main selection factors namely, the quality of the data points, the time series transformation, and where analytics are executed. Moreover due to theheterogeneous nature of networking environments, the convergence of communication and data protocols in IoT requires special attention when it comes to anomaly detection software development. For instance, raw data points from a smart water application are entirely different from that of a health care IoT application, hence the domain of application is another determinant factor in the construction of an efficient ADE. Machine learning in the unsupervised mode is indeed very efficient in situations where datasets are unpredictable. Moreover, cases where data points show non-liner time series require multivariate analysis that make the process more computing intensive. This property is not favourable to real time anomaly detection as more computation at the ADE level will affect the accuracy of the ADE. From a software development perspective the trend is similar to data mining tools embedded in popular database servers. Once the dataset is compiled the user can choose the most appropriate statistical tool. However, ERP solution providers will probably propose the ADE as a customizable module that would best fit the customers' requirements. Future work will investigate into the challenges from empirical experimentations and how anomaly detection can be translated as a service in cloud computing.

REFERENCES

[1] M. Marjani, F. Nasaruddin, A. Gani1, A. Karim, Ibrahim A. Targio Hashem, A. Siddiqa, and I. Yaqoob," Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges", Digital Object Identifier 10.1109/ACCESS.2017.2689040, , Vol 5, pp 5247-5261, May 17, 2017
[2] Khadija Akherfi a,*, MichealGerndt a, Hamid Harroud Mobile cloud computing for computation offloading: Issues and challenges
[3] http://www.kdnuggets.com/2016/09/evolution-iot-edge-analytics.html accessed on 8/1/2018
[4] O. Vermesan, P. Friess, Internet of Things-From Research and Innovation to Market Deployment, River Publishers Series in Communication, pp109-113
[5] https://github.com/dsmi-lab-ntust/AnomalyDetectionToolboxaccessedon8/1/2018
[6] http://iot.ee.surrey.ac.uk:8080/datasets.html accessed on 8/1/2018
[7] https://stats.stackexchange.com/questions/153498/how-do-i-obtain-the-anomaly-series-of-a-time-series accessed on 8/1/2018
[8] https://anomaly.io/anomaly-detection-normal-distribution/ accessed on 8/1/2018
[9] Varun Chandola, Arindam Banerjee, Vipin Kumar Anomaly Detection: A Survey, ACM Computing Surveys, September 2009.
[10] http://www.sciencedirect.com/science/article/pii/S0165168405001039 accessed on 8/1/2018
[11] Phil Winters, Iris Adae, Rosaria Silipo Anomaly Detection in Predictive Maintenance Anomaly Detection with Time Series Analysis, KNIME,2014 pp 3-9
[12] https://blog.statsbot.co/time-series-anomaly-detection-algorithms-1cef5519aef2accessed 8/1/2018
[13] http://www.bbc.com/news/science-environment-40669449 accessed on 8/1/2018
[14] Subutai Ahmad, Scott Purdy, Real-Time Anomaly Detection for Streaming Analytics, ar Xiv:1607.02480v1 [cs.AI] 8 Jul 2016

