

**Rochester Institute of Technology  
B. Thomas Golisano College  
of  
Computing and Information Sciences**

**Master of Science in Networking and Systems Administration  
- NSSA**



**~ Project Proposal Approval Form ~**

Student Name: QUOC TRUNG, KHUONG

Project Title: IoT Statistic and Analytics of Networking Traffic Data  
using AWS IoT Cloud Core

Project Area(s): ☐ Application Dev.   ☐ Database   ☐ Website Dev.  
(√ primary area)   ☐ Game Design   ☐ HCI   ☐ eLearning  
                                 ☒ X Networking   ☐ Project Mngt.   ☐ Software Dev.  
                                 ☐ Multimedia   ☐ System Admin.   ☐ Informatics  
                                 ☐ Geospatial   ☐ Other \_\_\_\_\_

**~ MS Project Committee ~**

Name	Signature	Date
Dr. Tae Oh		1/11/2021
<hr/>		
Chair		
Bruce Hartpence		
<hr/>		
Committee Member		

Approved: \_\_\_\_\_ Date: \_\_\_\_/\_\_\_\_/\_\_\_\_

☐ electronic copy received

# **IoT Statistic and Analytics of Networking Traffic Data using AWS IoT Cloud Core**

**by**

**Quoc Trung, Khuong**

Project submitted in partial fulfillment of the requirements for the  
degree of Master of Science in **Networking and Systems**  
**Administration MS**

**Rochester Institute of Technology**

**B. Thomas Golisano College  
of  
Computing and Information Sciences**

**Department of Information Sciences and Technologies**

**Jan 6<sup>th</sup> 2021**

## 1. Introduction

Analytics, in general, has always been in development for improvement, however, analytics tools are not as developed as its applications. Specifically, in the scalability of analytics tools, despite being extremely effective as technologies progress, most of them are not compatible with one another for requiring an installation on the premises. Networking analytics, in particular, should be concentrated more due to its impacts, by using AWS IoT Cloud Core and Jupyter Notebook for data analytics, this project will create a much more flexible, and simple platform for Networking analytics while keeping the depth of programming analytics. On the same note, IoT analytics is affected by these issues, the enormous amount of data requires a fast but scalable platform for processing and expanding. With our current infrastructure, it is difficult to achieve both of these aspects while reducing cost, and in this stage, Cloud platforms could be the ideal solution for IoT analytics because of its features and flexibility. As stated by Kang-Pyo, Spencer, Henry, Benjamin, and Daniel in July 2018 “by adopting a cloud-based solution, we compressed our development timeline because we did not have to install and maintain the necessary hardware and software. We were able to begin the prototype with minimal specifications of servers and databases, and most of the services were scalable enough to handle bigger data and heavier computation.”, this and many supported reasons, which we are going to discuss in other sections, are the foundation of this project’s decision to integrate Cloud services and IoT Analytics.

IoT analytics has proven itself to be highly useful in gaining a competitive advantage, however, to narrow down our project, the details regarding advanced IoT analytics might not be concentrated due to its complexity but rather the general and intermediate level of analytics, and data visualization with Networking data. For instance, with a suitable data set, we could perform IoT analytics on a network to predict future events such as node failure due to high workload, or Botnet behaviors. “Botnets are responsible for many types of attacks these days, including but not limited to spam spreading, distributed denial of service (DDoS) attacks, distribution of malicious software, information harvesting, and identity theft.” (Duc, A. Nur, Malcolm, 2016, p.1), this project is going to try to achieve this idea along with other results depending on the process. Furthermore, a software-defined network – SDN, can benefit from the data analytic of networking data as well. In 2018, Yan, Nejabati, and Simeonidou has suggested an NCMdB - network-scale centralized network configuration and monitoring database, model, which utilizes the data analytic of networking data to manipulate the network infrastructure based on demands and predictions. Although this application of SDN is highly impactful, it still has a considerable amount of overhead and scalability. My project might not “compatible” with the suggested infrastructure but it still can be a great ally in data analytic of networking data in the future, and there is valuable information from their proposal that will aid my progress.

## **2. Problem**

### **2.1 Problem statement**

The rising in amount and types of IoT data as a whole and Networking data in individuals reveals obstacles in the scalability of IoT platforms, as well as increasing complexity in performing analytics.

### **2.2 Motivation to study, significance of problem, and potential benefits**

Preventing a problem from occurring by predicting is much better than fixing an issue when it happened, the same saying can be applied in Networking's data, there are thousands of packets with different types of data and functions going through the network daily. Moreover, the use of IoT for collecting Network data or traffic information will greatly improve the first step of performing analytics, which is data collection. There is no best method to perform anything, however, there are always more optimal ways to execute it, the Cloud is one of the best ways to "scale" your project and because there are already features on the Cloud to support IoT analytics, it is now the most "convenience" method to implementing it. Despite being supported by many providers such as AWS, Azure, IBM, Google... IoT analytics is far from being limited, the features that these cloud providers introduce to the users are programmable to your specific demands, and there will be a steep learning curve if they want to use the features to an advance level.

As the traffic of networks increased, including IoT's traffic, previous methods for monitoring and analytics network data become less effective and ultimately, obsoleted. By utilizing cloud services such as AWS IoT Core, Microsoft Azure IoT ... the process is much more flexible in terms of scalability and performance. Additionally, the features that Cloud providers offer are highly relevant, straight forward setup, and directly impact the function of IoT's Network data analytic and statistic due to the ability of filtering data and customizing filters for specific requirements. This would lower the technical aspect of IoT analytic for IT technicians while raising the accuracy and visibility of the analytic process, which will provide long term benefits for businesses.

By implementing IoT analytics for Network data on Cloud services, we will be able understand IoT Analytic better as well as the comparison between cost, scalability and performance of Cloud services and on-site infrastructure, improving the topic in general. Furthermore, different data sets or requirements can be considered in the decision of choosing a suitable platform.

### **2.3 Project goals**

The goals of this project are evaluating traffic patterns including malicious behavior and potential network issues at IoT devices using AWS IoT Cloud Core features. The features include traffic analytics, QuickSight, Jupyter Notebook and the statistic evaluation especially at the hosts in their traffic patterns. An overall process of IoT analytics with networking data using AWS will be illustrated.

### 3. Prior works

In the beginning, the project was headed toward an “on-premises” database rather than the cloud for more control over the data. However, in 2016, Wentao et al. [1], have published a paper which indicates why the current TCP/IP architecture is not suitable for IoT Networking because of multiple constraints. To get started, the MTU – Maximum transmission unit, of IoT devices for networking is quite insignificant (around 127 bytes) compared to the standard 1500 bytes in an IP network, which is why the authors imply that the act of having a fixed MTU and reduce router’s overall workload “are intended for performance optimization in the current Internet, without the consideration of constrained IoT environment with small MTU sizes” [1, p. 2]. Additionally, IP networks rely on “multicast” and “broadcast” to function, which is not ideal for IoT devices since IoT devices have extremely limited power consumption. On a related note, due to the power limitation’s IoT devices, most routing protocols that are used on the IP networks are not as effective when implemented on IoT networks, once again, because they could not utilize broadcast and multicast to maintain the connection. Lastly, the security aspect of IoT, there is a suggestion of using TLS - Transport Layer Security, to provide encryption for IoT devices’ communication, however, the main issue would be the devices have to “maintain the states of the channel until it is closed” [1, p. 5] which IoT devices are not capable of due to battery consumption and frequent “sleep” mode to conserve resources.

Continuing on the same note, the Cloud, or any services that specialize in IoT analytics is commonly considered a superior decision than the existed IP/TCP infrastructure. This is due to the fact that IoT’s data and devices are extremely vary, depending on the requirements, their data can be quite large in terms of variety and size. This is the reason for Ahab [2] was created in 2016 by Michael, Johannes, Christian, and Schahram. Although we would not be using Ahab in this project, the concept of Cloud IoT is rather similar, especially in API, “Ahab, a distributed, cloud-based stream processing framework that offers a unified way for operators to better understand and optimize the managed infrastructure in reaction to data from the underlying infrastructure resources, that is, connected IoT devices” [2, p. 3]. Furthermore, AWS is also using API for most of its services, and features despite it might not be “purely” for IoT Analytic, it can provide the required flexibility in data adaptation.

With all the data traffic that IoT generated, despite implementing the Cloud, our existed infrastructure might not be able to handle it as we have mentioned and clarified above. New technology is definitely the solution, however, it is not something that would happen overnight, and that is why utilizing the existed devices and architecture will be our “silver bullet” for the near future. Specifically, a type of networking architecture that allows every device in the network to be a part of the network’s processing, computing, and storage, is identified as “fog computing”, and one of its weaknesses is the “unbalance” between devices, especially in storage. This weakness is eliminated by having the cloud as the storage service as well as handling “heavy task”. As indicated by Abdulsalam et al [3] in 2018, fog computing will enable near “real-time analytics” for IoT devices while Cloud computing is going to be responsible for “the abundance of computing and storage resources... in intensive applications” [3, p. 2]. In term of the benefits of this combination of technology, it is mainly to maximize the advantages of IoT analytics itself,

for instances, a car factory can use IoT sensors and this technology to predict upcoming maintaining and possible faulty parts in real-time, or a power company can observe the behavior of their customers' power usage and offer a much more suitable power plan for the consumers with the ultimate goal to save cost for both sides.

Complexity is one of the main issues when developing or implementing IoT analytics on the clouds, what kind of routing protocol and infrastructure is the best suit for the enormous number of IoT devices. One of the answers is “Iowa Quantified (IQ)” which was suggested by Kang-Pyo, et al. [4]. A system, or rather an architecture that is based on AWS, which provides high flexibility in terms of portable between environments such as from IoT devices to AWS cloud is a transition between two environments. This system “sensor hardware details are hidden behind standard, widely used protocols such as MQTT (Message Queueing Telemetry Transport) or HTTP, allowing diverse data to be processed and visualized in real time, and then transferred to storage for detailed, off-line analysis.” [4, p. 1]. Similar to our project development, AWS IoT Core will be used to “Processing and streaming” incoming IoT message, however, IQ system is much more in-depth because of the utilization of AWS S3 for back up service, AWS ESS for database storage, AWS Kibana for Dashboard display, and AWS EC2 for Analytics.

In 2018, Ravulavaru Arvind published a book titled “Enterprise Internet of Things Handbook: Build end-to-end IoT solutions using popular IoT platforms” [5]. In his book, the challenges as well as the suggestion of developing IoT Analytics on the Cloud is heavily focused. One of the ideas that is quite innovative is the implementation of the “Shadow device”, it “is a replica of the device present on the cloud, with the latest states and attributes that were persisted by the device.” [5, p. 74]. Due to the instability of IoT devices such as data errors, blank data, and devices’ connection... A “shadow device” has to be used to remain “persistent” of data. Personally, this idea will be considered in my project, however, the details and implementing behind it can be challenging, therefore, it is not guaranteed.

IoT’s applications are enormous, from a simple sensor to gather the temperature to image capturing. However, Suhas et al. have suggested AgrOne [6] - using drones to gather data of the farm and send it to the cloud for analysis, in 2018. The interesting part about this project is the compatibility of the drone and the IoT devices themselves. The drone itself is highly mobile but it is not able to transfer the data over the Internet without an assistant of a Raspberry Pi GSM Module. In general, the ground devices will gather data regarding the temperature, moisture, humidity... of the field while the drone is going to capture images and then analyzing the “green pixel” as a variable for analytic on the cloud. As mentioned by the authors, the project is working as required although there is room for improvements such as, “the framework can further be improved by providing synchronization between On-Ground nodes and Quad-copter. By providing synchronization, power can be saved and On-ground Sensor nodes can use Quadcopter’s GSM to upload data.” [6, p. 6]. Despite not having the ideal results, this technology is an embodiment of IoT integration among various devices, and it is also a practical application of IoT analytic.

One of the most useful material that would help my project tremendously is from AWS themselves, AWS IoT Developer Guide [7], although this book was created in 2015, it is still relevant in today's AWS IoT Core. This book not only guides the users through AWS IoT Core but also give them the information and clarification for the features and services they need. Within the book, the section regarding MQTT and the demonstration of how it works is quite interesting for me, because MQTT is a crucial part of IoT connections, being able to configure and observing the packets is highly informative.

In the introduction, we have mentioned Botnet prediction as one of the expectations of this project, and it is going to be elaborated. In 2016, Duc, A. Nur, and Malcolm have published an article in which they describe their “neural network technique” called SOM - Kohonen's Self Organizing Map [8]. Although the technical aspect of this idea is not feasible to implement in our project, the logic behind it can be a valuable asset for IoT analytics. According to the authors, botnets are extremely dangerous and responsible for many threads over the Internet, one of the main reasons is because of its intelligence. Unlike malware, virus... Botnets' behavior can be modified in real-time thanks to CC – communication channel, which can be deployed over HTTP, HTTPS, P2P (Peer-to-peer)... Generally, the CC provides the control of “bots” to the “master”, and it is also easy to setup. In addition, there are two main architectures for botnets, centralized and decentralized, both of which have their strengths and weaknesses, the former allows for better synchronization while having a single point of failure, the latter is more flexible and "robustness" but difficult to act as one. The reason that we need to comprehend how botnets operate is because of SOM's ability to learn these patterns, the more data SOM learned, the more accurate it will become.

As suggested by Shuangyi, Reza, Dimitra in 2018 [9], the use of data analytic in networking can be used as a valuable asset for a network-scale centralized network configuration and monitoring database (NCMdB). This implementation would allow NCMdB to alter devices' configuration and the SDN network itself, based on the result from the data analytic part. “Data-driven network analytics” is the term used to describe their project, it will make the decision based on traffic prediction, dynamic network abstraction, payload aware planning algorithm, and so much more. Despite everything is done inside of NCMdB instead of in the AWS Cloud such as my project, nonetheless, there are similarities, especially in the data analytic methods, which will be highly useful and related to my Jupyter notebook analytics.

Although data analytics has been developed for quite some time, the progress of analytics itself is quite unclear or demand's based rather. In 2018, Prasant et al. [10], the editors for a conference proceeding titled “Progress in Computing, Analytics and Networking” have discussed and outlined the progress of developing an analytics network. The book covered everything from sensors to analytics network, how the limitation in power can affect certain aspects of analytics such as security, and flow aggregator module for analyzing network traffic due to the large size and speed of today's networks. In general, the authors suggested many possible solutions for analytics, especially in security and performance, which seems to be the main issues in this topic.

## **4. Approach**

### **4.1 Plan**

The first step would be to search for a third party IoT that provides Networking's data for our project, however, generating our own data is also viable although it would need more structure. The scope of creating personalized IoT devices and data are not included in this project due to its complexity, therefore, finding a suitable data source is vital. Some services provide IoT data online such as "thingful.net" or "Arrayofthings.github.io", however, it is not the only reference for Data source but rather just an example, and there is also a probability of creating or generating our own data set for this project if required by using Wireshark captures. All in all, the will be directed to Cloud services, which is our next step.

Continuing, various Cloud services provide IoT analytics such as AWS IoT Core, IBM Watson, Google IoT Core, Microsoft Azure IoT. However, AWS will be our only and unchangeable decision for the Cloud provider because of its popularity and features. Moreover, these services on AWS IoT Core are subjected to change, therefore, making an early decision on them might not be optimal since they might change in their functionality or new features could be introduce.

Our last stage would be to customize and optimize the analytic and statistical progress of the data through QuickSight and Jupyter Notebook. Depending on the kind of data sources, the results might vary with a goal of trying to predict future issues such as heavy load on devices, bot net behaviors... In a nutshell, this step is going to reflect the goals of the project as visible as possible.

### **4.2 Feasibility**

#### **4.2.1 As a capstone project**

##### **4.2.1.1 Challenges/barriers**

Once again, the data is vital for the project, due to the fact that IoT for Networking data is quite vague, and there is not a general structure for it at the moment, therefore, it can be quite a challenge to find a suitable data source, and extremely difficult to customize an external data source. Another obstacle would be the Cloud services themselves, the features and configuration of the cloud services have a learning curve which will take time to fully comprehend. As we have mentioned previously, things are subjected to changes and if there are any changes in the duration of this project, it could affect the process directly.

##### **4.2.1.2 Project scope**

In order to make this project achievable with my skill sets, the creation of IoT devices are not considered, which leads to our external and modified data source. Specifically, only the analytics and statistical progress of IoT Networking data are focused and elaborated. Additionally, there are multiples cloud services online such as AWS, IBM, Google, and Microsoft but the project will only include AWS.



## 4.2.2 As a project intended to be field deployable

### 4.2.2.1 Challenges/barriers

In a practical environment, the cost of deploying IoT analytic and statistic for Networking data on the Cloud can be unpredictable. Repeatedly, the data source is vital to how much processing and filtering the Cloud service has to perform. Personally, the biggest barrier in my project would be the adaptability in the future where the pattern of Networking data changes since the project relies just on a single data set.

### 4.2.2.2 Functional limitations

Although the topic of this project is concentrating on the analytics and statistic part of IoT Networking data, it is only on the operation and average complexity analytic. Advance analytic is out of my capability as well as some of the features of statistics, such as full statistics using algorithm.

## 4.3 Deliverables

In the final stage of the project, there will be a functional script of AWS Jupyter Notebook for various Networking analytics such as malicious behavior detection, traffic prediction of nodes... Additionally, a demonstration of the process will be illustrated and explained, however, the focus is still on AWS Jupyter Notebook script.

## 5. System description

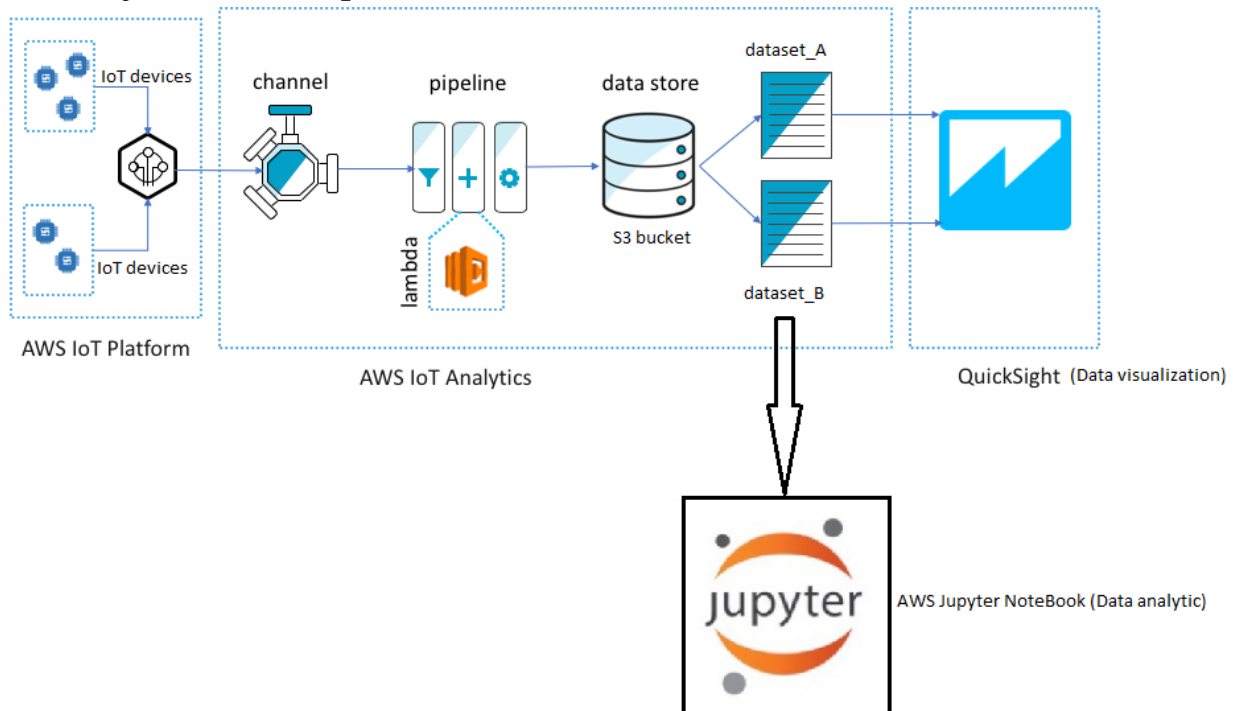


Figure 1. AWS IoT Analytic general description

The original version of the image can be found on AWS site:

<https://aws.amazon.com/blogs/iot/using-aws-iot-analytics-to-prepare-data-for-quicksight-time-series-visualizations/>

## 6. References

- [1] W. Shang, Y. Yu, R. Droms, and L. Zhang, “Challenges in IoT Networking via TCP/IP Architecture,” *Named Data Networking (NDN)*, 10-Feb-2016. [Online]. Available: <https://named-data.net/publications/techreports/ndn-0038-1-challenges-iot/>. [Accessed: 18-Dec-2020].
- [2] Vögler, M., Schleicher, J., Inzinger, C. and Dustdar, S., “Ahab: A Cloud-Based Distributed Big Data Analytics Framework For The Internet Of Things,” *Software: Practice and Experience*, 2016. Available at: <https://doi.org/10.1002/spe.2424> [Accessed 18 December 2020].
- [3] A. Yassine, S. Singh, M. S. Hossain, and G. Muhammad, “IoT big data analytics for smart homes with fog and cloud computing,” *Future Generation Computer Systems*, 15-Sep-2018. [Online]. Available: <https://doi.org/10.1016/j.future.2018.08.040>. [Accessed: 18-Dec-2020].
- [4] K.-P. Lee, S. J. Kuhl, H. J. Bockholt, B. P. Rogers, and D. A. Reed, “A Cloud-Based Scientific Gateway for Internet of Things Data Analytics,” *A Cloud-Based Scientific Gateway for Internet of Things Data Analytics | Proceedings of the Practice and Experience on Advanced Research Computing*, 01-Jul-2018. [Online]. Available: <https://doi.org/10.1145/3219104.3219123>. [Accessed: 18-Dec-2020].
- [5] A. Ravulavaru, “Enterprise Internet of Things Handbook: Build End-To-end IoT Solutions Using Popular IoT Platforms,” *Packt Publishing*, 30-Apr-2018. [Online]. Available: [https://books.google.com/books/about/Enterprise\\_Internet\\_of\\_Things\\_Handbook.html?id=8aeItQEACAAJ](https://books.google.com/books/about/Enterprise_Internet_of_Things_Handbook.html?id=8aeItQEACAAJ). [Accessed: 18-Dec-2020].
- [6] M. V. Suhas, S. Tejas, Snigdha, S. Yaji, and S. Salvi, “AgrOne: An Agricultural Drone using Internet of Things, Data Analytics and Cloud Computing Features - IEEE Conference Publication,” *4th International Conference for Convergence in Technology (I2CT)*, 27-Oct-2018. [Online]. Available: <https://doi.org/10.1109/i2ct42659.2018.9057995>. [Accessed: 18-Dec-2020].
- [7] AWS. (2015, September). AWS IoT CoreDeveloper Guide. Retrieved November 12, 2020, Available: <https://docs.aws.amazon.com/iot/latest/developerguide/iot-dg.pdf>. [Accessed: 18-Dec-2020].
- [8] D. C. Le, A. N. Zincir-Heywood, and M. I. Heywood, “Data analytics on network traffic flows for botnet behaviour detection,” *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, 06-Dec-2016. [Online]. Available: <https://doi.org/10.1109/ssci.2016.7850078>. [Accessed: 18-Dec-2020].

- [9] S. Yan, R. Nejabati, and D. Simeonidou, "Data-driven network analytics and network optimisation in SDN-based programmable optical networks," *2018 International Conference on Optical Network Design and Modeling (ONDM)*. [Online]. Available: <https://doi.org/10.23919/ondm.2018.8396137>. [Accessed: 18-Dec-2020].
- [10] P. K. Pattnaik, S. S. Rautaray, H. Das, and J. Nayak, Eds., "Progress in Computing, Analytics and Networking," - *Proceedings of ICCAN 2017, Part of the Advances in Intelligent Systems and Computing book series (AISC, volume 710)*, 2018. [Online]. Available: <https://link.springer.com/book/10.1007/978-981-10-7871-2>. [Accessed: 18-Dec-2020]