Big Data Analytics Project Proposal

A Fog Computing Prototype

Ali Alizadeh Mansouri
40102969

Marco Sassano
26658245

Abstract

Internet of Things (IoT) aims to bring every object (e.g. smart cameras, wearables, environmental sensors, home appliances, and vehicles) online, hence generating massive volumes of data that can overwhelm storage systems and data analytics applications. Cloud computing offers services at the infrastructure level that can scale to IoT storage and processing requirements. However, there are two major downsides to this two-layer infrastructure: The bottleneck is the network bandwidth, meaning that transmission of large volumes of raw data will oversaturate the network bandwidth on the way to the cloud, and there will be a high latency and response time for the results to come back from the cloud to the IoT devices. To overcome this limitation, Fog computing paradigm has been proposed, where cloud services are extended to the edge of the network to decrease the latency and network congestion. In our project, we implemented a Fog Computing infrastructure for early detection of epilepsy seizures using EEG timeseries data consisting of three layers: IoT devices and sensors, a Fog layer, and the Cloud. We performed the analytics on the 2nd and the 3rd layers. We report the methods, the results, and discuss the possible challenges, limitations, and future work.

1 Introduction

The number of Internet of Things (IoT) devices has increased to a great extent in recent years. It is estimated that 50 billion devices will be connected to the Internet by 2020 [4]. On the other end side of the infrastructure, Cloud computing as a paradigm delivers computing services over the Internet — the Cloud — to offer flexible resources to deal with a wide range of scalable computational demands. This includes analysis, aggregation, and storage of large volumes of data (Big Data) from the IoT devices. The total Internet bandwidth crossing international borders in 2013 was 100 Tbps. Furthermore, while application demands are growing from 100s of terabytes towards petabytes per day, network capacity growth has been decelerating [7]. In other words, the bottleneck of such infrastructure lies on the network bandwidth between the IoT devices and the Cloud. This issue arises from the fact that most Cloud computing datacenters are geographically centralized, and situated far from the proximity of the end devices.

Fog Computing (FC) was proposed by Cisco in 2012 to address the needs of the applications which demand low latency and high response times [2]. As a distributed computing paradigm, FC acts as an intermediate layer in between Cloud services and IoT devices (or end users/devices in general) [5]. In this manner, the concept of FC is analogous to *data locality*, in which computational tasks are moved towards the data, instead of the other way. Figure 1 shows a typical FC environment.

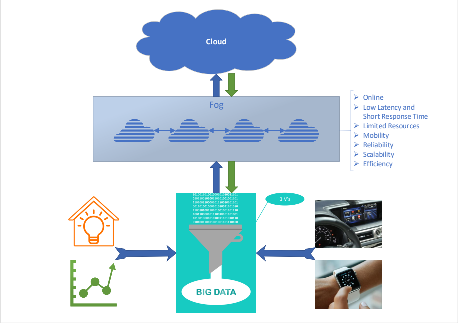

Figure 1: A Fog Computing environment.

1.1 Problem Specification

We consider the healthcare application of early detection of epilepsy seizures using EEG timeseries data, because such healthcare applications require close to real-time response times.

Our goal was to achieve less network congestion between the Fog and Cloud layers as well as higher response times for the EEG sensors. More specifically, we aim for a balanced tradeoff between fast and light-weight — even though less accurate — computations on the IoT side, and more accurate classifications but with higher latency on the cloud side.[1]

1.2 Related Work

FC in general as well as different applications requiring real-time response times have been considered before. For example, Tang *et al.* [6] implemented a hierarchical FC architecture for anomaly detection of pipe leakage in smart cities. We were inspired by the work of Diab Abdulgalil *et al.* [3], where they propose a FC architecture with SVM classification at the edge, and deep convolutional networks in the cloud. To the best of our knowledge, and to the date of this writing, this has been the only work which considered FC for efficient epilepsy seizures detection.

2 Materials and Methods

2.1 Dataset

Andrzejak *et al.* [1][2] provide a dataset of 500 individuals, each with 4097 data points for 23.5 seconds. This dataset is further reshaped and shuffled to 23 × 500 = 11500 records, where each record contains 178 data points (features) for 1 second by Qiuyi and Fokoue on UCI Machine Learning Repository [8]. Each record of this dataset is labeled with one of 5 classes, with one class being the seizure state. Since our goal was predicting the seizures, we further restructured this dataset into `(time,value)` tuples, and considered the seizure class of consecutive 178 tuples as positive, and all the groups tuples belonging to other classes

[1] *A note on the revision of the first two sections*: These sections have been mildly revised to 1) more accurately reflect the actual implementation of the project, and the materials — including the dataset used — and, 2) incorporate the professor's helpful comments from the project proposal.

[2] http://epileptologie-bonn.de/cms/front_content.php?idcat=193&lang=3&changelang=3

Table 1: The breakdown of the EEG timeseries dataset used.

	Total	Positive	Negative
Training	7 000	1 392	5 608
Test	4 500	908	3 592
Total	11 500	2 300	9 200

as negative. This resulted in 11500 × 178 = 2 047 000 tuples, which would be simulated as a stream of data generated by an EEG sensor. The final breakdown of the dataset is shown in table 2.1.

2.2 Project Structure

We will consider a three-tier infrastructure for our FC implementation, as described below:

1. **The IoT devices/sensors**: This layer consists of the IoT data generators/producers/sensors. Such devices generate large volumes of data in a small time-frame; however, they usually lack the required processing power, storage, and energy to process this data. They may also expect a response from the higher layers depending on the event inferred from their produced data, for example, in case of a gas/pipe leak sensor.

   Several IoT hubs and data stream processing frameworks exist, such as Apache Kafka[3], IBM Watson Internet of Things (IoT)[4], Microsoft Azure IoT Platform[5], etc.

2. **The Fog layer**: The Fog layer is the intermediate layer between the end devices and the Cloud. Its job is to perform any preliminary analytics on the data, in order to both prevent the whole raw data travel to the Cloud, and to infer any fast responses required by the IoT devices in the lower layer based on the mentioned analysis. The Fog layer may also serve as a resource provisioning coordinator in some applications.

   At the time of this writing, the only existing framework dedicated to Fog/Edge Computing is Apache Edgent[6], which is currently in a beta state.

3. **The Cloud**: The Fog sends the results of its analyses to the Cloud for further analysis and aggregation. The Cloud may also serve as a coordinator for the bottom two layers, and/or it may send required responses or information to the IoT applications.

   As the Cloud services have been offered since the 2000s, there is a large number of options, both commercial and open-source. Examples include IBM Bluemix[7], Microsoft Azure[8], Amazon Web Services (AWS)[9], and Apache Cloudstack[10].

[3] https://kafka.apache.org/

[4] https://www.ibm.com/internet-of-things

[5] https://azure.microsoft.com/en-us/overview/iot/

[6] http://edgent.apache.org/

[7] https://www.ibm.com/cloud/

[8] https://azure.microsoft.com/

[9] https://aws.amazon.com/

[10] https://cloudstack.apache.org/

Our top preference is the open-source solution — Apache Cloudstack, while we may consider educational services of the other Cloud providers depending on the risks faced in the course of the development of the project.

2.3 Project Phases

We aim to proceed through the following phases for our project. Details are subject to change due to our inexperience with the FC infrastructure. Nevertheless, we believe that we will achieve the following by the end of the project deadline:

- We plan to develop a functional implementation of the layers described in the previous section in separation first. Since our ultimate goal is to connect each layer functionally, we will choose from the available technologies and frameworks those that we deem will best suit to this end. This phase will also include a good deal of research and experimentation with the available tools and frameworks. Meanwhile, we will also research and define an application use-case to test with our framework. The choice of the application must be approved by the supervisor of the course.

- For the next step, we will connect the three layers with their available APIs, and set up a fully-functioning prototype. This phase will also include part of the debugging of the infrastructure.

- We will implement our application use-case on the developed infrastructure. We plan to analyze the data generated by the IoT devices in the first layer, both on the Fog layer and the Cloud layer. Preferably, the two analyses will be such that the one in the Fog layer is advantageous to the Cloud. For example, we may implement data pre-processing in this layer. The certain types of analyses performed will be both determined by the specific use-case that we come up with, as will as be one mentioned during the lectures (e.g. Random Forests) as required by the project description.

- Finally, we plan to analyze the process, aggregate the results, discuss the challenges and future works, and compile the project report containing all said deliverables.

References

[1] Ralph G. Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, and Christian E. Elger. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Phys. Rev. E*, 64:061907, Nov 2001.

[2] Flavio Bonomi, Rodolfo Milito, Jiang Zhu, and Sateesh Addepalli. Fog Computing and Its Role in the Internet of Things. In *Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing*, MCC '12, pages 13–16, New York, NY, USA, 2012. ACM.

[3] Huda Diab Abdulgalil. *A Multi-Tier Distributed fog-based Architecture for Early Prediction of Epileptic Seizures.* PhD thesis, University of Waterloo, 2018.

[4] D Evans. The internet of things: How the next evolution of the internet is changing everything. Technical report, 2011.

[5] Redowan Mahmud, Ramamohanarao Kotagiri, and Rajkumar Buyya. Fog Computing: A Taxonomy, Survey and Future Directions. In Beniamino Di Martino, Kuan-Ching Li, Laurence T Yang, and Antonio Esposito, editors, *Internet of Everything: Algorithms, Methodologies, Technologies and Perspectives*, pages 103–130. Springer Singapore, Singapore, 2018.

[6] Bo Tang, Zhen Chen, Gerald Hefferman, Tao Wei, Haibo He, and Qing Yang. A Hierarchical Distributed Fog Computing Architecture for Big Data Analysis in Smart Cities. In *Proceedings of the ASE BigData &#38; SocialInformatics 2015*, ASE BD&#38;SI '15, pages 28:1—-28:6, New York, NY, USA, 2015. ACM.

[7] Ashish Vulimiri, Carlo Curino, P Brighten Godfrey, Thomas Jungblut, Jitu Padhye, and George Varghese. Global Analytics in the Face of Bandwidth and Regulatory Constraints. In *12th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 15)*, pages 323–336, Oakland, CA, 2015. {USENIX} Association.

[8] Qiuyi Wu and Ernest Fokoue. Epileptic seizure recognition data set. https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition. Accessed: 2019-04-15.