Máster en Ingeniería de Sistemas y Servicios para la Sociedad de la Información

Trabajo Fin de Máster

Título	Dataflow Specification of a K-Means Clustering Algorithm	
Autor	Marta Rodríguez Ramos	
Tutor	Eduardo Juárez	VºBº
Director		
Tribunal		
Presidente	César Sanz Álvaro	
Secretario	Antonio Carpeño Ruiz	
Vocal	Juana María Gutiérrez Arriola	
Fecha de lectura		
Calificación		

El Secretario:

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA Y SISTEMAS DE TELECOMUNICACIÓN

UNIVERSIDAD POLITÉCNICA
DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENERÍA

Y SYSTEMAS DE TELECOMUNICACIÓN

MSc in Systems and Services Engineering for the Information Society

Dataflow Specification of a K-Means Clustering Algorithm

Master Thesis

Marta Rodríguez Ramos

Madrid, October 2019

Contents

List of Figures

List of Tables

List of Acronyms

6LoWPAN	IPv6 over Low-Power Wireless Personal Area Networks
ACK	Acknowledge
ACL	Access Control List
API	Application Programming Interface
CoaP	Constrained Application Protocol
CoaPS	DTLS-Secured Constrained Application Protocol
DTLS	Datagram Transport Layer Security
DDoS	Distributed Denial of Service
EUI	Extended Unique Identifier
FS	File System
GPU	Graphics Processing Unit
Hops	Hadoop Open Platform-as-a-Service
HDFS	Hadoop Distributed File System
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure
IMEI	International Mobile Equipment Identity
IP	Internet Protocol
IPSO	Internet Protocol for Smart Objects
IoT	Internet of Things
JDBC	Java Database Connectivity
JSON	JavaScript Object Notation

JVM Java Virtual Machine

JWT JSON Web Token

MAC Media Access Control

ML Machine Learning

MVC model-view-controller

MVVM model-view-viewmodel

NAT Network Address Translation

OMA LwM2M Open Mobile Alliance Lightweight Machine-to-Machine

PKI Public Key Infrastructure

PSK Pre-Shared Key

REST Representational State Transfer

RPK Raw Public Key

SQL Structured Query Language

TLS Transport Layer Security

TSDB Time-Series Database

UI User Interface

URN Uniform Resource Name

UUID Universally Unique Identifier

VM Virtual Machine

Summary

The number of internet connected devices has already by far surpassed the number of human beings. The pace of growth is still so big that in the next five years that number will double. The ecosystem of these devices, collectively called Internet of Things (IoT), is a source of a tremendous amount of data and creates several unheard challenges for researchers and companies. New, unconventional ways of storing, analyzing, and processing of the data had to be proposed. One such a solution is Hadoop Open Platform-as-a-Service (Hops), a result of years-long research between KTH Royal Institute of Technology in Stockholm and RISE SICS AB. It is a platform enabling an analysis of extremely large volumes of data with cutting-edge, open-source technologies for Big Data and Machine Learning (ML). This master thesis provides support for connecting these two environments. It provides instruments for secure and reliable ingestion of IoT data into Hops platform. Moreover, it provides tools for ensuring the level of security by supporting the execution of mitigating measures, such as automated exclusion of misbehaving devices and dropping traffic from sources of Distributed Denial of Service (DDoS) attacks. To allow the data ingestion a new element was introduced to the ecosystem - IoT Gateway. It is a platform, where the authenticated IoT devices can stream data to. Furthermore, Hopsworks, one of the Hops' main component, was extended with REST API that allowed the gateways to securely connect to the Hops ecosystem. A testbed, including IoT software simulator and a real IoT device with dedicated hardware, was built and comprehensively tested and benchmarked. The architecture is based on the publicly open and very popular security protocols - Raw Public Key (RPK) and Hypertext Transfer Protocol Secure (HTTPS). It is shown that the proposed solution is performant, scalable, and provides high reliability in a real-life case scenario. Up to our knowledge, the work done in this thesis makes Hopsworks the world's first open source Big Data platform with secure IoT data ingestion.

Resumen

Hyperspectral imaging collect and process information from across the electromagnetic spectrum.While the human eye sees color of visible light in mainly three bands, however, hyperspectral images cover a wide range of wavelengths.This imaging modality was developed, as the first approach, for remote sensing and earth observations, the multiple applications and the wide information provided do

The accurate delimitation of brain cancer is an important task having a surgery. Several techniques are used in order to guide neurosurgeons in the removal of the tumor. Hyperspectral imaging (HSI) is a promising non-invasive and non-ionizing optical imaging modality that has the ability to speed-up medical imaging research and clinical practice.

Acknowledgments

I would like to express my deepest gratitude to my supervisor Theofilos Kakantousis for the continuous support during the time of the internship. His experience in the fields of the project, a head full of ideas and accurate questions, and willingness to advice were extremely helpful and made the internship an invaluable experience.

I would also like to thank my university supervisor Professor Mariano Ruiz for helping me with the project, making the thesis procedure smooth, and for agreeing to do all of it fully remotely.

Chapter 1

Introduction

1.1 Motivation

The present project has been developed within the context of HELICoiD (HypErspectraL Imaging Cancer Detection), which is an European project funded by the Seventh Framework Programme of the European Union and, especially, by the FET - Open (Future & Emerging Technologies) initiative. Several institutions from different countries such as France, United Kingdom and The Netherlands have been involved in this European project as well. Furthermore, it includes two hospitals, three companies and four universities (Spain included). This research, particularly, is a collaborative work with UPM (Spain) within research design group of Electronic and Microelectronic.

The main aim of the HELICoiD European project is to provide to the surgeon a technique which informs accurately about healthy tissue and tumours in real time. This is all thanks to Hyperspectral images since traditional methods have low level in terms of sensitivity and the boundaries of the image are not clearly defined. In other words, HELICoiD aims at distinguishing between healthy tissue and tumours by extracting the spectral information of each pixel.It can be assumed that the spectral information is correlated with the chemical composition of a particular material. Therefore, each hyperspectral pixel has a spectral signature of a specific substance.

With regard to this line of research, the present work implements an unsupervised clustering method called K-Means on a parallel architecture in order to supply information in real time to surgeons.

In this regard, a dataflow language called πSDF is used, in order to perform the parallelization of this algorithm. The πSDF is a generalization of SDF MoC, is a syncronous dataflow model of computation. An application is modelled by directed graph of computational entities, called actors, that exchange data packets called data tokens, through a network of First-In First-Out queues (FIFOs)[**lee1987synchronous**]

The procedure is as follows; hyperspectral (HS) sensors attain hyperspectral cubes, and HS cubes are pre-processed in order to reduce dimensionality and noise. Afterwards, they are clustered employing K-Means, which defines different areas properly. After using this algorithm, an unsupervised segmentation map is generated. Meanwhile in parallel, the system executes a number of algorithms belonging to supervised classification. These algorithms are PCA (Principal Components Analysis), SVM (Support Vector Machine) and KNN (K-Nearest Neighbour). After performing these algorithms, tissues are displayed using different colours in order to represent the associated classes. Applying the majority voting, the unsupervised segmentation map obtained from K-Means clustering algorithm as well as the classification map obtained from supervised classification, are merged.

The implementation of this algorithm is carried out using a dataflow specification tool called PREESM. This tool is widely used for manycore architectures and signal processing applications. The objective of parallelizing this algorithm is to speed up computations for data clustering to target real time response.

1.2 Objectives

As was mentioned earlier, the main objective of this project is to analyse a clustering algorithm called K-means in order to find possible parallelization methods and approach real time . The following points have been developed to achieve the global aim:

- Research how to efficiently parallelize an algorithm by considering the bottlenecks that generate delays on the execution of the algorithm.
- Study in depth the unsupervised clustering algorithm, especially, an optimized model for hyper-spectral images provided by Universidad de Las Palmas De Gran Canaria.
- Learn how to parallelize the chosen algorithm using a dataflow specification tool called PREESM. Ones of the great advantages of this dataflow used in PREESM is the flexibility, predictability and expressivity provided,since this semantic is based

on interfaces that fix the number of tokens consumed/produced by a hierarchical vertex.

- Test the reached speedup by comparing it with the sequential implementation one.

Chapter 2

Background

2.1 Project context

2.2 Hyperspectral Images

2.2.1 Description

2.2.2 Applications

Remote Sensing

Cancer Detection

2.3 Classifiers

2.3.1 Concept

2.3.2 Unsupervised classification

KMeans

2.3.3 Supervised classification

2.3.4 Processing chain for hyperspectral image

PCA (Principal Component Analysis)

SVN (Support Vector Machine)

KNN (K-Nearest Neighbor)

2.4 Parameterized and Interfaced Synchronous DataFlow MoC

2.4.1 PREESM

Hyperspectral images are defined as an image that has high spectral as well as spatial resolution, which means that a pixel does not have just the 3 colors of reflectance's values characteristic of usual images.

Hyperspectral Images (HSI) collects high resolution spectral information gathering hundreds of bands from the ultraviolet to the infra-red range of the electromagnetic spectrum. This information is used to distinguish among the different materials composing the captured scene [**lee1987synchronous**].

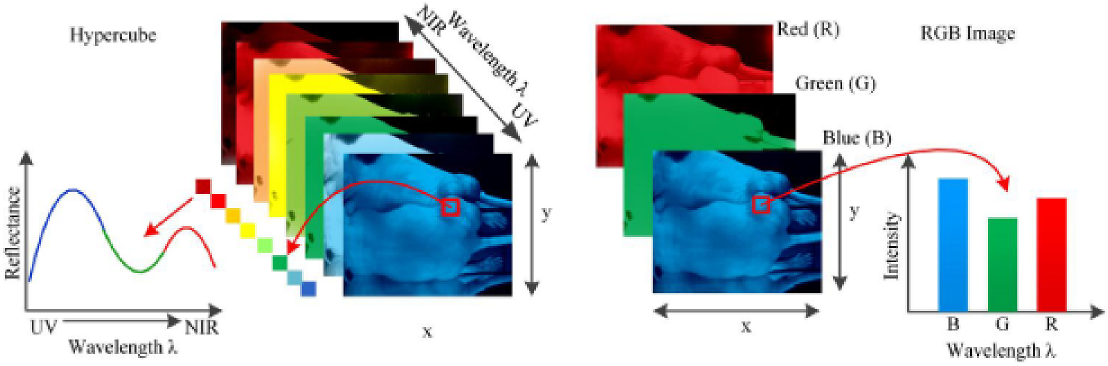

Figure 2.1: Comparison between hyperspectral image(left) and RGB image(right)

This electromagnetic radiation is not only present among the visible range of wavelengths, as it covers the most part of the electromagnetic spectrum. This kinds of images provide us with useful information as said earlier, as it includes electromagnetic spectrum invisibly to the human eye. However, due to the huge amount of data, processing of HSI images is very computationally expensive.

Chapter 3

Implementation

This section describes how the design challenges were solved in this project.

3.1 K-Means algorithm for hyperspectral images

3.1.1 Serial code profiling

3.2 Parallel K-Means implementation

3.2.1 Methods

3.2.2 Bottlenecks

3.3 PREESM implementation

Chapter 4

Results

The evaluation of the project was performed in three steps - verification, validation, and benchmarking. The following subsections describe each of the steps in details.

4.1 Accuracy analysis

4.2 Timing analysis

Chapter 5

Conclusion

This chapter summarizes the work done in this project. It reviews if the goals were achieved, presents the main areas of future work, and provides final reflections.

5.1 Goals Achieved

The project met all the set goals. It was empirically proven that the project is feasible at its scope. Support for IoT data ingestion was provided through IoT Gateway and extending Hopsworks. Security was ensured by the use of the HTTPS and RPK protocols. On top of that, the gateways were authenticated using JSON Web Token (JWT). Besides, the `hops-util` library was extended to provide tools for the exclusion of misbehaving devices and blocking traffic from sources of DDoS attack. Sample streaming jobs were provided to test the added functionality. Moreover, multiple tests were run to prove the reliability of the system and the ability to recover from potentially harmful situations like a power outage, unexpected reboot of the machines, and others. The system demonstrated its resilience and capability to return after a collapse of any of the elements. In addition, the IoT Gateway was tested against bigger traffic on a scale that the test machines were able to simulate. It was shown that the gateway can deliver data fast enough and in a reliable manner. The gateway generally performed very well, however, some parts, like the DatabaseService, can be optimized thus making the gateway work faster under heavy traffic. Lastly, examples of streaming analytics jobs were presented to visualize the measurements. The data was correctly retrieved from storage, processed and shown in a graphical form.

5.2 Future Work

The scope of the project was limited because of time constraints. To meet both the project requirements and deadlines some simplification were introduced. The following elements are expected to be further developed to make sure the system is production-ready:

- The Open Mobile Alliance Lightweight Machine-to-Machine (OMA LwM2M) protocol was implemented only in terms of two types of messages - temperature and presence. It is advised to implement the rest of the Internet Protocol for Smart Objects (IPSO) objects to make the IoT Gateway fully compliant with the protocol.
- Currently, the IoT Nodes are provided with the hostname and port of the IoT Gateway. To make the system truly scalable, a bootstrap server needs to be introduced. It would contain the list of active IoT Gateways and would redirect the nodes to the optimal one. In other words, the bootstrap server would server the role of a load balancer. This would also ease the usage of hostnames instead of IP addresses which would make the system much more flexible. In this case, the gateways would perform a DNS lookup.
- Extracting gateways as a separate resource not bounded to a single project would highly extend flexibility and ease the analysis of the data. Currently, the gateways are a subresource of a project and only the stored datasets can be shared between projects.
- The Hops Kafka Authorizer currently supports access based on the IP address. In the case of a Network Address Translation (NAT), it creates a conflict between the gateways. Blocking one gateway would potentially block a whole range of gateways. Adding authorization based on the port would mitigate the problem.
- The work done in this project provides tools for the automatic exclusion of the devices and/or gateways. The next step would be to develop a real ML model that could protect the Hops platform against DDoS attacks.
- Another approach to data ingestion would be to make the IoT Nodes push the data directly to a Kafka broker. It would require a complete redesign of the system but could potentially enable end-to-end Public Key Infrastructure (PKI) security. This design would also require the deployment of Kafka brokers not only in the main data center but also in the field introducing new challenges.

5.3 Reflections

It was shown that the IoT Gateway and Hopsworks IoT extension work as expected. We were able to connect real IoT devices and stream the data to the cloud in a secure, performant, and reliable manner. The gateway was designed with a flexible architecture so, by replacing `LeshanService`, the system can be extended to other IoT protocols, such as MQTT. The code developed in this thesis is fully open-source and free to use and distribute under the GNU v3.0 license. It was not, however, tested in a production environment. The system would be required to go through an exhaustive quality assurance phase before being installed with a real-life IoT network. We hope that the work conducted in this thesis will be the subject of further research and development in a production environment and that the extended Hops platform will open new possibilities of data analysis to researchers, companies, and organizations.