Graduation Project

# Data Mining Framework for Internet of Things

**Hamdy Michael Ayas**

**UNIVERSITY OF CYPRUS**

**DEPARTMENT OF COMPUTER SCIENCE**

**May 2017**

# UNIVERSITY OF CYPRUS

## DEPARTMENT OF COMPUTER SCIENCE


# Data Mining Framework for Internet of Things


**Hamdy Michael Ayas**


Supervisor

Dr. George Pallis

This Graduation Project is submitted for the partial fulfillment of the requirements for the acquisition of the degree of Computer Science of the Department of Computer Science of University of Cyprus.

May 2017

# Acknowledgments

I would like to express my gratitude to my beloved family: my parents Mohammad Ayas and Eleni Michael Ayas, for believing always in me and for their continuous support and encouragement throughout this year. My dearest sister Sandy and dearest brother Malek and of course my grandparents: Vasiliki, Kiriakos, Khadija, Ahmad. Thank you all for the unconditional love and support.

I would also like to thank my supervisor Dr. George Pallis. In addition, I would like to thank the Phd candidate Demetris Trihinas, for his knowledge and insightful comments.

# Summary

The concept of Internet of Things (IoT) has been introduced and it is considered as the next big step of the technological evolution. Smart and internet-enabled devices, or inter-connected objects, percept their surrounding environments and come with features of intelligence and decision making. These devices and objects, are blended in any environment and they interact and cooperate with each other in order to reach common goals. In order to percept and understand their surroundings, and consequently, create intelligence and features of decision making, the objects sense and gather data from their environments. Data that need to be transformed into sensible information and knowledge. However the generated data are so big and valuable, that it is needed the development of technologies to deal with knowledge extraction from this enormous amount of data.

Consequently, we believe that a good addition to the available toolkit is a framework that will enhance the transmission of Data Mining and data processing, to the edge of IoT Applications – a data analytics framework specialized in Edge Mining. It is intended the development of such a framework, that will enable IoT with early application of data mining to limit the transmission of unnecessary data from IoT devices to the cloud infrastructure and reduce any unnecessary overwhelming of the cloud with unnecessary processing and analysis of data. Thus, it enhances the move of a significant part of the data mining and processing from the cloud infrastructure on the IoT devices themselves.

Therefore, due to the variety of different types of hardware and Operating Systems (OS), the framework needs to operate on a platform and OS independent manner, as well as to provide a low cost on complexity and processing analysis, in order to reach lower energy consumption which is a point of vast importance for the energy constrained devices. The framework contribute to transform a part of the data mining to the devices on the edge of IoT, enabling them with a variety of data mining and edge mining techniques and algorithms. The results of the data mining, can be extracted in various data types in order to be disseminated for more extensive and accurate analysis on the cloud. The framework is not enabling the devices with the complex and sophisticated analytical capabilities of the cloud but with the early application of data mining, supports the cloud and offers a mean to reduce the processing and handling of unnecessary data.

# Table of Contents


Chapter 1 ..... - 1 -
Introduction ..... - 1 -
General Introduction ..... - 1 -
Objective & Contribution ..... - 4 -
Structure ..... - 5 -
Chapter 2 ..... - 7 -
State of The Art research and description of Challenges ..... - 7 -
The Internet of Things ..... - 7 -
Development of IoT Applications ..... - 10 -
Data mining solutions ..... - 11 -
Edge Mining solutions for Internet of Things ..... - 15 -
Devices and objects on the Edge of Internet of Things ..... - 17 -
Internet of Things Platforms ..... - 18 -
Challenges & Motivations ..... - 22 -
Chapter 3 ..... - 23 -
Proposition of an IoT specialized data mining framework ..... - 23 -
Architectural Design ..... - 23 -
Methodology for defining Requirements ..... - 26 -
Specifications ..... - 27 -
Required knowledge and technologies ..... - 28 -
Chapter 4 ..... - 32 -
Implementation of the IoT specialized data mining framework ..... - 32 -
Design and Interfaces ..... - 32 -
Data Interface & Data Dissemination ..... - 33 -
Data Processing ..... - 35 -
Data Mining Techniques & Algorithms ..... - 37 -
Chapter 5 ..... - 40 -
Results & Use Case ..... - 40 -
Use Case ..... - 40 -
Results ..... - 42 -
Chapter 6 ..... - 45 -
Conclusion and Open Challenges ..... - 45 -
Conclusion ..... - 45 -
Discussion ..... - 46 -

Bibliography ..... - 48 -

# Chapter 1

## Introduction

---

General Introduction
Objective & Contribution
Structure

---

### General Introduction

We are at the point of time, in humanity's history, where Computers are evolving from simple calculating machines to ubiquitous systems with their existence trying to be transferred from our individual devices to every single object in our surrounding environments (1). And by "surrounding environments", we mean every possible aspect of our every-day life comprising machines, products, devices, wearables, means of transport and people that need to interact and communicate with each other. Hence, we are heading to a new age where our societies and our everyday lives are strongly related and based on information and computerization (2) (1). With the technological evolution and development we are able to enhance all our environments and our activities with intelligence and embed in everything the characteristic of smartness. Enabled by the concept of Internet of Things, the environments in which we exist are becoming smarter and that is due to the newly created abilities of every single object that consists an environment to connect and interact in a network of objects, exchange data, create knowledge from that data and actuate accordingly in an autonomous manner (2). Therefore, the Information Age in which we are living in is formed and a Digital Revolution in every aspect of our lives is now a fact.

This transfer and evolution of the computing systems into ubiquitous systems predisposes the ability of smart internet-enabled devices, often referred to as things/objects blended in any environment, to interact and cooperate with each other in order to reach common goals. And this is basically the main idea and definition of the disruptive concept of

Internet of Things (IoT) (1) or Internet of Everything as some prefer to call it (3). In very simple words, an ability is created to take any object, connect it in a network of objects, enabling them to exchange data and create knowledge from that data (1). That equals to the creation of a new type of intelligence that offers features of automization in environments as well as features of "decision" making from the connected objects themselves. Therefore, every asset can have its own, as Gartner calls it, "digital voice" (4) that enables the exchange of information through a network of distributed intelligence - a network consisted by smart objects or devices that sense and exchange data. So, by applying "digital voices" to devices and objects, a lot of data can be gathered and analysed creating a seamless generation of information and knowledge (2).

Undeniably, the successful analysis and conversion of the data into useful and valuable information can offer a tremendous insight knowledge about the state and use of devices and environments, creating a new kind of distributed intelligence and communication in assets, disrupting entire fields by introducing them the characteristic of "smartness" and intelligence in objects and environments. An interesting example is that of the smart factory. Enabled by the concept of Internet of Things, is taking industrial automation and manufacturing to the next level creating Industry 4.0 (5). Another related example, is that of business intelligence that is evolving radically with data-driven business models being the ultimate goal of every modern business in order to success. It is very common in the enterprise world, to create intelligence from the development of smart products and services that are generating data and are using these data to make them more valuable. A good real world example is the data driven approach developed by Thyssen Krupp elevator with MAX that is "bringing the elevator industry into the digital age by integrating machine learning IoT technology to deliver a breakthrough service solution", thereby delivering maximum possible uptime (6). By using what IoT technology offers, prediction of maintenance issues is achieved before they even occur and relevant functionalities and procedures are optimised. The embedded intelligence in elevators achieved the reduction of downtime by up to 50% and saved approximately half of the 16.6 years that New York City office workers were spending for waiting elevators (6). Furthermore, it cut elevator electricity consumption in Grand Avenue Courtyard in El Segundo, California by 58.5% (7). It is getting clearer that fields related with products and services development (Service Composition, Service Management etc.) are disrupted

from the added potentials with the newly introduced knowledge and information that can be enabled from the data generated.

However, it turns out that most of the times it is extremely challenging to successfully implement feasible and viable IoT Applications with the required (data) analytical capabilities. The smart and inter-connected objects/products of any IoT Application, are sensing from their different environments and they are generating data constantly. Connecting a lot of objects together and letting them to generate data, may lead to the creation of enormous amounts of data that have to be stored and processed. Specifically, it is estimated that by the year of 2020 there will be 50,000,000,000 devices creating and sharing 40,000,000,000,000 GB of data that will need to be stored, analysed and managed across the Internet (8). Hence, the successful development of Internet of Things applications can be proved very challenging, with several aspects in need of consideration. Undeniably, Big Data Analytics and Data Mining techniques are momentous in order to convert the information successfully into knowledge (9) (10) and developers of Internet of Things applications need guidance for best practices in designing and implementing applications in order to overcome the challenges and obstacles that come with Internet of Things.

Moreover, the amount of data that can be generated from a network of sensing objects and the information included in them is "enormous and considered highly useful and valuable" (10). Machine Learning and other data mining techniques may offer important insight knowledge, giving a new dimension to the potential intelligence. For example, not all the data are equally important and this can be identified using data mining. Some are very vital at the point of time that are generated, some can only add value in later stages of analysis and some are not important at all. If a machine is going to fail, it is vital to take some actions immediately before it fails but if a user made a mistake using a machine is not that vital. However, if a lot of users are making the same mistake over time it means that the machine needs an improvement in its User Interface and User Experience. This kind of knowledge can only come over time and analysis of historical data. In addition, the data that are not important at all, they are not required to be considered and expend resources on them at all. Moreover, the general objective of the concept of Internet of Things (1) is to make objects to actuate in an autonomous manner, so consequently, it is

very essential that data mining technologies like clustering, classification, frequent patterns and so on, have high impact on objects and devices in order to achieve fast decision making (2). And it is very critical, for example, in fast decision making to transform data into valuable knowledge and information as fast as possible and in the most efficient and effective way.

Considering the estimated (huge) amount of data that is going to be created (10), the analytics' requirements of IoT and the capabilities of the available tools for analysing data (9) (2), we can come to the conclusion that "Data analysis tools available are simply not powerful enough to handle and analyse big data of IoT" (10) and it is necessary to move a part of the data processing to the devices sitting on the edge in order to manage and process the data successfully as well as in order to enhance their capabilities of actuation and analysis (11) (12). So it is necessary to discover ways and methodologies that will significantly reduce the generation and exchange of the data that are sensed from devices and objects from the time they are created (12). In addition, the uptime of devices and objects needs to be maximized making very crucial the factor of energy consumption. As a result, it is of vast importance and there is a big necessity of tools that will help to tame data volume and velocity, as well as energy efficiency (11) in order to reduce energy consumption and network traffic with unnecessary data and increase the uptime of devices as much as possible. So it will always be needed a trade-off between the quality of the analytics and processing capabilities offered by the device and the energy consumption of the device or object.

**Objective & Contribution**

The developed work, is about the development of a framework that enhances the transmission of Data Mining and data processing, to the edge of Internet of Things (IoT) applications. The particular framework, aims to limit unnecessary data transmission from IoT devices to the cloud infrastructure and unnecessary analysis for knowledge extraction. Thus, it enhances the move of a significant part of the data mining and processing from the cloud infrastructure on the IoT device itself, trying to face the challenges of dealing with the enormous amount of data generated by IoT applications and mining knowledge from those data from the time of their creation. Due to the variety

and the different types of devices and operating systems (Android, Linux etc.) the Data Mining (Edge Mining) requires to be on a platform and OS independent manner, as well as to provide a low cost on complexity and processing analysis, in order to help lower energy consumption which is a point of vast importance for the energy constrained devices.

To achieve that, the framework is firstly receiving data from different types of data sources (e.g. CSV files) with the ability to easily be extended and giving the option for adding other data sources types. Those data once they are read, they are transformed in particular data types that are suitable for analysis. Then, after studying data mining techniques and algorithms, a variety of them that are suitable for edge mining and are dominant in attributes of low cost data processing, complexity and use of resources are developed in extensible algorithm models. The results of the data mining, have the ability to be extracted in various types of files and data types in order to be disseminated for further and more accurate analysis, because the low cost attribute may need some conventions (e.g. in accuracy). The methodology followed, is using Object Oriented design and development in Java programming language, aiming to be fully extensible and to give options of future additions on the developed framework (e.g. of algorithms, data sources etc.). On the development part, a test-driven approach is taken for each part of the framework.

**Structure**

This Graduation Project presents the great impact of IoT in people's lives and the importance of low cost data mining and edge mining techniques in the development of IoT applications with a proposed framework for that use. This framework's objective is to increase the feasibility and viability of IoT Applications by giving the ability to their devices to manage data, information and knowledge they generate in the most energy efficient and low in cost ways. Specifically, in Chapter 2 we present an overview of the available data mining technologies and frameworks as well as different IoT solutions and an indicative development process for IoT Applications along with some examples of real world applications. In Chapter 3, we propose and describe the architecture and a high level overview of the framework for software developers that will enhance and improve

the data mining capabilities of IoT applications by enabling processing and data mining capabilities on devices that sit in the edge of IoT and in Chapter 4 we present how this framework is implemented along with its detailed specifications and how it can be used. Finally, Chapter 5 will be concluding with summarizing this work and with discussions on open challenges. In addition, there is an Appendix for a reference on the implementation of the framework.

# Chapter 2

## State of The Art research and description of Challenges

---

The Internet of Things
Development of IoT Applications
Data mining solutions
Edge Mining solutions for Internet of Things
Devices and objects on the Edge of Internet of Things
Internet of Things Platforms
Challenges & Motivations

---

### The Internet of Things

The concept of Internet of Things is applicable in all aspects of our lives and with the introduction of the characteristic of smartness and intelligence in our environments, entire fields are disrupted with very big influences in societies, industries and every-day life (13). Our residences and working places are transformed into connected/smart homes and offices, giving us capabilities of customizing our living environment, controlling appliances and monitoring the energy we consume as well as monitoring of living behaviours and habits. Smart wearables and smart gyms can take outdoor and sporting activities of an average person to a totally new level offering detailed analysis and visualization of related data. For example, a wearable can indicate in detail everything relevant to the sport/activity committed by the athlete, starting from how many steps he/she run or walked during a day until how many calories are burned every 5 minutes of a day. Autonomous cars and smart means of transportation are shaping new routines and habits in the way we live and our cities, are becoming smart cities. Urban design and planning is vastly improved after leveraging data that are generated from people who use technology within an urban setting. In healthcare, the automation of several activities and constant monitoring of behaviours can decrease the ratio of human error and of course can result in a better quality on taking care of patients, elderly and kids. Therefore, there

is a huge number of opportunities in the development of IoT applications and "Developers are on a technological cusp with opportunities for innovation riper than ever" (11).

Hence, the development of Internet of Things applications is emerging in the industrial and academic world in all fields of action, as it is also indicated by Gartner's IT Hype Cycle (see Figure 2.1). It is quite remarkable and impressive, in Gartner's IT Hype Cycle, the fact that approximately half of the technologies on 'Innovation Trigger' and on 'Peak of Inflated Expectations' are technologies directly related to Internet of Things or technologies strongly associated with the management and use of the information and data that are generated from Internet of Things.


**Figure 2.1** Gartner 2016 Hype Cycle of emerging technologies

Source: Gartner Inc. (4)

Consequently from the emerging needs of IoT development, it is very crucial to clearly identify and define the perspectives/vision of Internet of Things as well as the technological aspects of an Internet of Things application. Having in mind the example where we can design a smart product with a "digital voice", connect it in a network of

distributed intelligence to communicate and make it to exchange data that will be converted into knowledge, we can get a good overview for the technological perspectives (see Figure 2.2) that can guide to the creation of solid IoT applications. Through the technological vision of Internet of Things, there is the perspective of "Things" (the objects with "digital voices" in the example), the perspective of the "Internet" (the network that establishes communication) and the perspective of the "Semantics" (the brain and soul of an IoT solution that converts Data into knowledge) (1). In simple words, the perspective of "Things" is referred to the devices and objects that are enabled with sensing and computing capabilities. The perspective of the "Internet" is referring to the network of inter-connected objects, appliances and people resulting the evolution of the Internet to the web3 (ubiquitous computing web). Finally, the perspective of the "Semantics" is referring to the understanding of the information generated and to the process where we try to make sense of the data and information sensed and to transform them into knowledge (1).

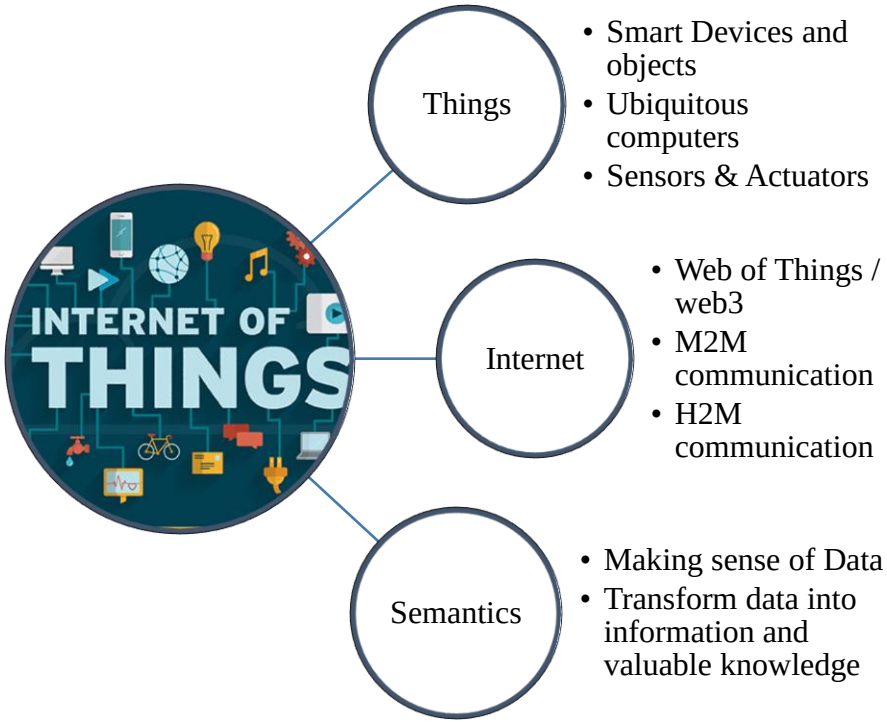


**Figure 2.2** Perspectives of Internet of Things

Hence, the technological and technical aspects that need to be developed for a successful Internet of Things application are related to these perspectives. In a five layered high level architecture (2), it is firstly required, on the lower layer, the development of the devices and objects that are sitting on the Edge of IoT and have sensing capabilities. These devices

need to have established communication and connection in the network of other interconnected objects and devices through an access gateway. Then, is required the development of a Middleware (2) that will contain the logical grouping of the data and information that is sensed and their preparation and transformation in order to become sensible and valuable knowledge. The Middleware, will then compose and offer knowledge and decisions as services to the Application. Finally, in the Application it is required to visualize and develop the presentation of the information and knowledge, depending on the context of the application, in an understandable and sensible manner (2) – according to the semantics.

**Development of IoT Applications**

For example, in an IoT application for a smart transportation system, all cars are connected and send data of their locations in order to be analysed and to calculate estimations and predictions of the estimated traffic of a city centre. So Machine to Machine (M2M) communication is established. After analysing the data sent and transforming them into information, this system will visualize them to inform drivers about the situated traffic in the direction they are heading. The system will analyse the data and make some predictions and calculations, transforming them into, meaningful for the drivers, knowledge. Additionally, in order to prevent a possible traffic jam in a certain road, according to predictions from the data mining process, they may suggest alternative roots for reaching their destinations. As a result, the cars will communicate with their drivers and actuate with informing them about traffic. With the mined knowledge, the cars can also take the decision to follow an alternative root and will suggest possible alternatives to their drivers. Then, will maintain communication and wait for the drivers' interference to change the predefined roots.

Data visualization is very important in IoT for the required communication with humans, and fields like Human-Computer Interaction and interaction design are thriving due to IoT requirements. Users need to interact with their environments and need to have access in attractive and easy to understand data representations. The visualizations of knowledge and information must be representable of the application domain and useful for the purpose of use (2) and this is better achieved by using more than 2 Dimensional

representations, resulting in 3D screens and technologies to support these needs. With technologies like Virtual Reality and especially Augmented Reality, Human to Machine communication is vastly improved. Specifically, is evolving and transforming simple data representation to tangible representation of knowledge with simulations of real world examples and situations helping users to make better sense out of the information presented to them. IoT devices and objects are embedded in people's environments and very often interact with users. As a result they need to be designed and developed to offer the best possible usability and discretion in order to be useful and usable.

**Data mining solutions**

A big technological evolution is made to enable devices with capabilities to sense from their environments, create data and share these data but this is not enough to enrich devices and objects with intelligence. The inter-connected objects and devices are naturally resulting in a seamless generation of data. This will eventually result in the creation of an enormous amount of data and information that need to be stored, processed, and most importantly data that need to be converted into knowledge. The data can be either meta-data that describe the state of the smart objects or data produced from the objects or devices (parameters sensed from the environment) (10). However, the intelligence and smartness of devices is introduced only when decision making is enabled in environments and it can be measured by the quality of the decisions that objects and devices are able to make as well as from the value of the knowledge and information they offer depending on the context of the implemented solution. Receiving the data in high frequency is important to be able to achieve the intelligence needed in Internet of Things, but just receiving the data is not enough.

The successful analysis of the data is very important in order to mine useful information from that data and to ultimately convert them into valuable knowledge. This conversion can offer a tremendous insight knowledge about the state and use of devices and environments, creating a new kind of distributed intelligence and communication in assets. As a result, data-driven decision making will be introduced in environments of smart objects and these objects will be enhanced with capabilities to actuate and interact autonomously. In addition, the insight knowledge about the state and use of devices, may

result in the composition and creation of new services to offer. Services that will empower any application, product or service with additional value. But what this knowledge really is? And how it is possible to manage and handle the enormous amount of data created in order to mine it (see Figure 2.3)?

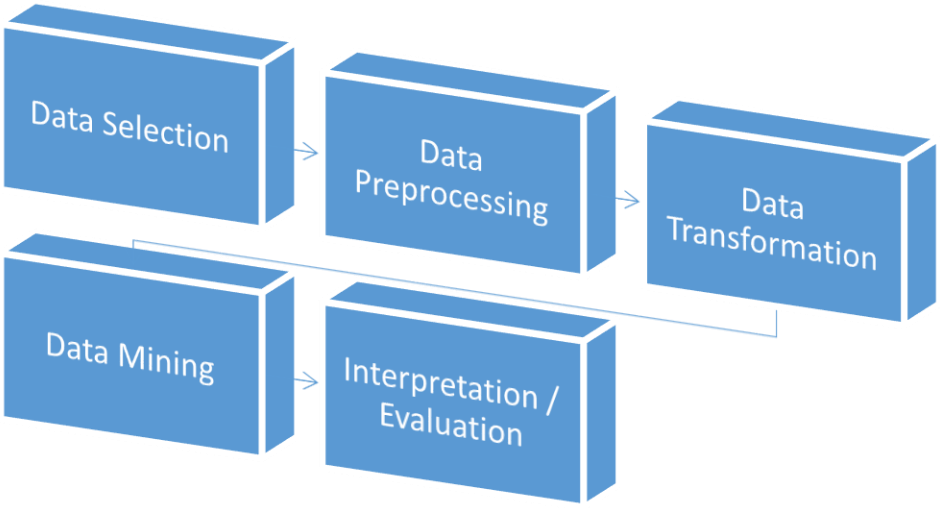


**Figure 2.3** The Data Mining Process to convert data into knowledge

Source: (1)

First of all, it is very important to define the information requirements and the knowledge that needs to be mined from the data. Then it is necessary to see the form of the available data and decide which is the most suitable data mining system or technique. For example, if we want to create a recommendation system that will introduce additional services to customers, trying to predict what kind of services they may also be interested in and recommend those services to them, we have to define and use data related to customers' profiles (profile of what services they use and how they use them) and maybe their location. Of course, we must also give indications about the knowledge we want. In our case, we will give indications about what a correct recommendation is in terms of customers' profiles, history and location. And here comes the touch of real intelligence of Data Mining and Machine Learning. Specifically, Machine Learning gives the ability to create a Model capable to improve and adapt itself - capable to learn and build experience. Such models are eventually being used in Data Mining in order to take raw data as input and offer in return real world knowledge like recommendations, predictions,

frequent patterns, outlier detection, classification and clustering of data. However, the successful creation of an effective data mining model that will be valuable predisposes a lot of research in Machine Learning algorithms and a lot of implementation, testing, benchmarking and evaluation in order to end up in the most optimized solution that may offer the greater value. Hence, it can be from very useful to a necessity for the viability of a data mining solution, the use of tools that group data mining techniques and machine learning algorithms.

Developing the best Data Mining model that will be able to derive the most valuable knowledge out of the data, is a challenging process and an iterative one rather than a linear one. Every time a Machine Learning algorithm is tested and evaluated on a dataset, more information are revealed about the dataset, leading to a better understanding of the problem and ultimately to the right algorithm and data mining model. Consequently, it is very useful the use of tools, like Weka which is a "workbench" that offers the software incorporation of several standard Machine Learning (ML) techniques. Weka is written in Java language and by using it can any data scientist extract useful knowledge from data, of a database, that are too large to let their manual analysis be feasible. In addition, with Weka it is possible to test and explore several Machine Learning algorithms on samples of the available data in one place without needing to switch environments, programming languages and write "endless" configurations and scripts to adapt the available datasets. Weka's ease of use, both with its Graphical User Interface as well as with its Command Line Interface, makes it one of the ideal tools to experiment with Machine Learning and conclude to the best application of Machine Learning.

With the Weka explorer, a data specialist has features to overview the actions that are possible to take place on the available datasets. First of all, data pre-processing techniques may be explored in order to apply the best choice on the dataset and achieve the best possible manipulation of the data to transform them into a desired form, suitable for the following analysis. Then it is possible to experiment with different algorithms and techniques on the dataset. For example, it is possible to select and try classification and regression algorithms to classify data, in the case that datasets are suitable for supervised machine learning. If they are not, there are available several clustering algorithms that may be selected and run to result in grouping the data in several clusters. Moreover, there

are available association algorithms that may reveal great insights from the dataset and different attribute selection algorithms and techniques that may be applied to process and analyse only relevant data for extracting the desired knowledge. Finally, in order to create a complete data mining solution the knowledge needs to be visualized and Weka has a feature for choosing the best possible visualization of knowledge. After exploring the different available techniques and algorithms it is important to run experiments and this can be done through the Weka Experimenter. The Experimenter offer features to design experiments with the several selected algorithms and datasets ending up to results that can be analysed and compared.

Another excellent example of a software solution with implemented data mining software is ELKI (14). ELKI is also an open source software implemented in Java and focuses in research of algorithms with an emphasis on unsupervised methods in cluster analysis and outlier detection. One of the main challenges that ELKI is overcoming is that due to the existence of many algorithms that can deal with tasks of similar characteristics from research over time in data mining, the development and implementation of data mining solutions may vary making practical evaluation of data analytics difficult and strongly dependant on the programming skills of the scientist implementing them. As a result, ELKI proposes a software framework with data mining algorithms and data management tasks to help in the development and objective evaluation of advanced data mining algorithms. This is achieved with the modular architecture that is followed and the distinct separation of concepts in a Data Mining solution. For example, algorithms, data types, distance functions and every other aspect of a solution consist a model that can be extended and further developed. Each of the implemented extensions of a model can be used as a component for reuse and combination in any other Data Mining solution.

Hence, after generating datasets from an IoT application, such tools and frameworks can be used for the creation of the proper analytical model that will successfully transform the data generated into useful knowledge and information. A framework like ELKI is ideal in order to benchmark a data mining algorithm and compare it with others and a ‘workbench’ like Weka are ideal in order to experiment with different data mining techniques and evaluate them in order to conclude on the most suitable algorithms and techniques. In addition, they are very good in giving the opportunity to modify and create

new techniques and algorithms based on the existing ones and with the proper iterative process of developing a data mining model, they can constitute a significant contribution on the transformation of data into valuable knowledge. Undeniably they are very strong tools to shape an excellent data mining solution.

However, Internet of Things requires some additional characteristics and are needed techniques and algorithms more specialized for these characteristics. For example, not all the data are equally important. Some are very vital at the point of time that are generated and some can only add value in later stages of analysis. If a machine is going to fail, it is vital to take some action immediately before it fails, but if a user makes a mistake using a machine, it is not that vital. However, if a lot of users are making the same mistake over time it means that the machine needs improvement in its Human Computer Interaction. This kind of knowledge can only come over time and analysis of historical data and tools like ELKI and Weka are suitable. However, if we desire the development of devices enabled with the required intelligence to predict, for example, a failure and with or without the aid of human intervention to actuate accordingly and interact in good time, then more specialized techniques and algorithms are needed than the ones offered by Data Mining Frameworks of general purpose, like these. They are needed techniques and algorithms that will decentralize the processing and analysis in order to successfully deal with the big data of Internet of Things (15).

**Edge Mining solutions for Internet of Things**

In a network of inter-connected objects that sense data and cooperate to reach common goals, it is generated a very big amount of data that come from multiple data sources and with the general purpose (traditional) data mining technologies, it is very usual to have a big congestion of data. As Baraniuk observed (16), a very crucial bottleneck of data processing will be shifted from sensor to the data processing, communication, and storage capability of sensor. Consequently, semantics definition and Data Analysis start to take place even on devices that sit at the edge points of the IoT in order to cope efficiently with the huge amount of data generated (9). Therefore, we need devices smart enough in order to contribute as much as possible in the data mining and data analytical requirements of transforming as fast as possible raw data into information but we need to

always keep in consideration that IoT applications use resources constrained devices and every processing action that the devices need to do, comes with a cost in the energy that powers the device (11).

With Edge Mining, it is intended to transfer a part of the processing of sensory data near or at the point of the battery powered devices and inter-connected objects that sit on the edge of Internet of Things, in order to convert the raw sensed data into "contextually relevant information" (2). In the edge mining process (see Figure 2.4) that takes place at an inter-connected object, the sensed data are not transmitted directly in the host of the application (cloud) for analysis and data mining. Instead, every piece of sensed data pass from some stages that eventually determine if the particular piece of data needs to be transmitted or not (12). First, a smart object, can be triggered and awakened to sense from its environment some data. The smart object will not directly transmit the data but will pass it from a state estimator. The state estimator is basically the stage where the raw data will be pre-processed and translated into a contextually meaningful information according to the IoT application. After the translation into a meaningful information, the event detector may trigger in case of the information generated is useful and only if the event detector is triggered the information will be transmitted. Otherwise, the information is considered irrelevant and the device will return on its sleeping mode (12).

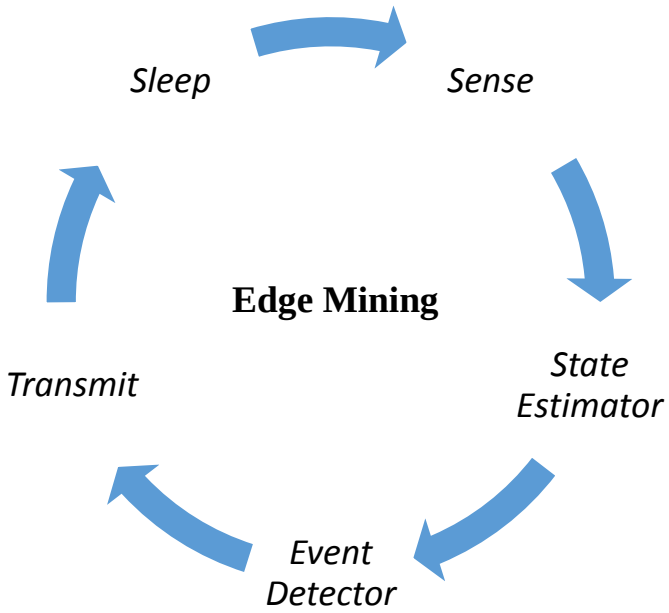


**Figure 2.4** The Edge Mining process at an inter-connected object

Source: (12)

An example of an edge mining algorithm that fulfils these attributes is G-SIP (General Spanish Inquisition Protocol) (12). This algorithm, creates a mechanism on the interconnected object that, with its connection with the middleware of the receiver (cloud), it creates a model of the expected data to sense that the receiver does not want to receive. The event detector decides to send the sensed data only if they are not similar to the data expected. In that way, it is defined a pattern of the data and the data are transmitted for analysis only if they are different from the expected pattern (12).

Another example, is the framework ADAM (ADAptive Monitoring) which consists techniques for IoT (11). Those techniques, regulates the amount of data transmitted and disseminated through the network to the cloud with taking in consideration the current evolution and viability of the metric stream. Specifically, with ADAM's algorithms (a. Adaptive Sampling, b. Adaptive Filtering) one-step ahead estimations are enabled and are seriously getting considered and effectively detected abrupt changes in the stream. As a result, it can possibly reach a significant reduction of the data transmitted (by at least 74%), of the energy consumption (by at least 71%) and maintain an 89% of accuracy (11).

**Devices and objects on the Edge of Internet of Things**

Undeniably, smart devices and objects are one of the most important elements of IoT applications. According to Cluster of European research projects on the Internet of Things (1) – "'Things' are active participants in business, information and social processes where they are enabled to interact and communicate among themselves and with the environment by exchanging data and information sensed about the environment, while reacting autonomously to the real/physical world events and influencing it by running processes that trigger actions and create services with or without direct human intervention". Consequently, devices and objects are enabled with capabilities for sensing their environments and creating data and information from what they sensed. These sensed data and information need to be shared with their surroundings and as a result the devices need to be enabled with communication capabilities that will allow the exchange of data and information. Devices need to have established communication with other devices in order to cooperate and reach common goals (1) (Machine to Machine communication) as well as with humans in the case of an action that needs to be received

from a human or in the case that it is needed the human intervention to proceed with a certain action.

Therefore, devices and objects in IoT have sensors in order to fulfil the required sensing capabilities and are enriched with technologies that enable them with communication and data transfer. Such a technology is RFID (Radio Frequency IDentification) (2) technology that enables the design of super small chips for wireless data communication. Another example is WSN (Wireless Sensor Networks) technologies, that integrate sensing and processing capabilities in hardware level, establish communication with data transformation through the network and serves sensor resources through the WSN Middleware. Moreover, the constant data generation may challenge any network and any asset that is going to process and analyse these data. Another crucial aspect that needs to be considered it that devices are mostly battery-powered and it is very crucial to optimize as much as possible energy efficiency, or even ensure independency from battery consumption (e.g. RFIDs are not powered by batteries). On the other hand, WSN technologies rely on battery powered systems (see table 2.3) and it is critical to ensure the use of energy efficient technologies but to also regulate the generation and transmission of data. (13) (2) (16)

| | *Standard* | *Range* | *Power* | *Lifetime* | *Communication* | *Processing & Sensing* |
|---|---|---|---|---|---|---|
| *RFID* | ISO 18000 | 10 | Harvested | Indefinite | Asymmetric | No |
| *WSN* | IEEE 802.15.4 | 100 | Battery | < 3years | Peer-to-peer | Yes |

**Table 2.2** Gartner 2016 Hype Cycle of emerging technologies
Source: [6]

**Internet of Things Platforms**

During the last years, there has been a significant shape-change in the software industry strongly influenced from the development and evolution of Cloud Computing (17). Various and very attractive service models are developed, from large public cloud providers like Microsoft (see Figure 2.5), Amazon, Google and ThingWorx that offer

important reduces of capital expenses and costs to interested parties. An example of such a service model offers pay-as-you-go services making the development and deployment of web services very easy and simple. In addition, these service models offer high elasticity for dynamic load adaption and very simplified resource management. Finally, the developed cloud based platforms are transforming the traditional, expensive and time consuming process of developing complex software solutions. Very sophisticated solutions with software components are offered as a service in the cloud, making the development of complex solutions a matter of assembling and configuring components.

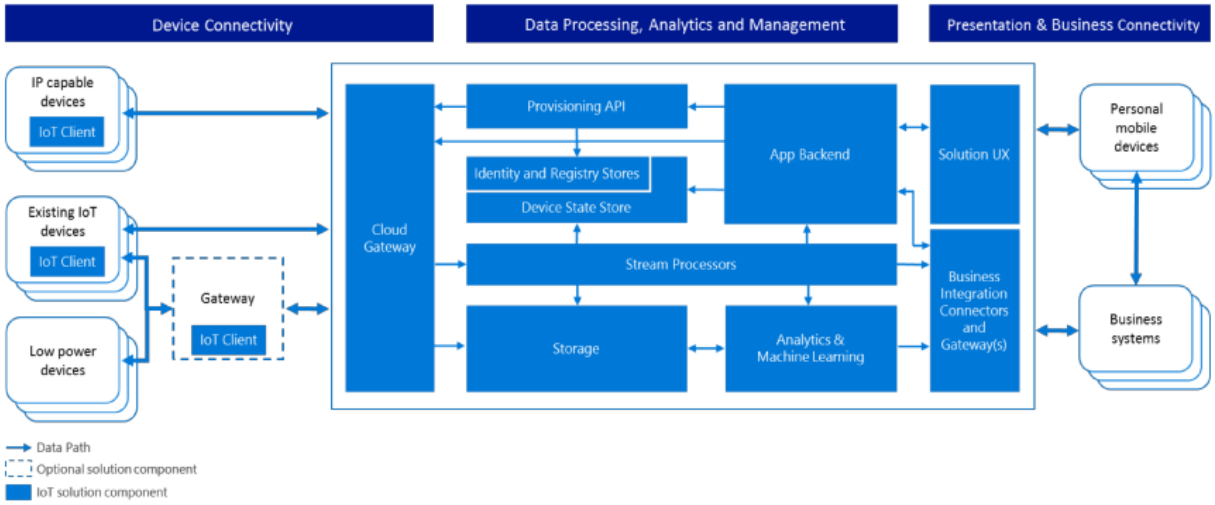


**Figure 2.5** Microsoft Azure IoT solution

Source: (18)

Hence, using the same architecture and selling model, the biggest cloud providers have Internet of Things platforms and along with other providers of IoT platforms, are offering components of data analytics, solutions for IoT and configurable software solutions, suitable for the development of end to end IoT solutions and applications. For example, we have a case where it is required the development of a smart transportation application for a city with the Amazon Web Services for IoT (AWS IoT) in order to improve driver safety with connected cars. The connected cars, are developed by the specialized AWS IoT Device SDK that enables devices to connect, authenticate, and exchange messages with AWS IoT using different protocols like the MQTT, HTTP, or WebSockets protocols. The AWS IoT Device SDK supports C, JavaScript, and Arduino, and includes the client libraries, the developer guide, and the porting guide for manufacturers.

Therefore, the cars sense and gather data from the road regarding its condition (see Figure 2.6) and using the AWS IoT Device Gateway, establish one-to-one and one-to-many communications with AWS IoT. The gathered data are transmitted to the AWS IoT where are being processed (filtered and transformed) and analysed on the fly, based on predefined rules. Then it is identified how dangerous the road possibly is with the rules that the developers of the particular application set and the analytics solutions that are used as a service from the application. If the road is categorized and identified to be dangerous, then alerts are sent to the nearby cars that inform their drivers. In addition, all the data are sent and stored to the accordingly configured space of Amazon's S3 for future analysis and retrieval of relevant information. Then, the great amount of data that is stored, is used to train an Amazon's Machine Learning model, configured in the needs of the developed application in order to increase accuracy and improve the quality of the knowledge and information that may be offered. Finally, the analytics and knowledge extracted from those data, are visualized on dashboard-like mobile and web applications which is again developed with AWS IoT Device SDK.

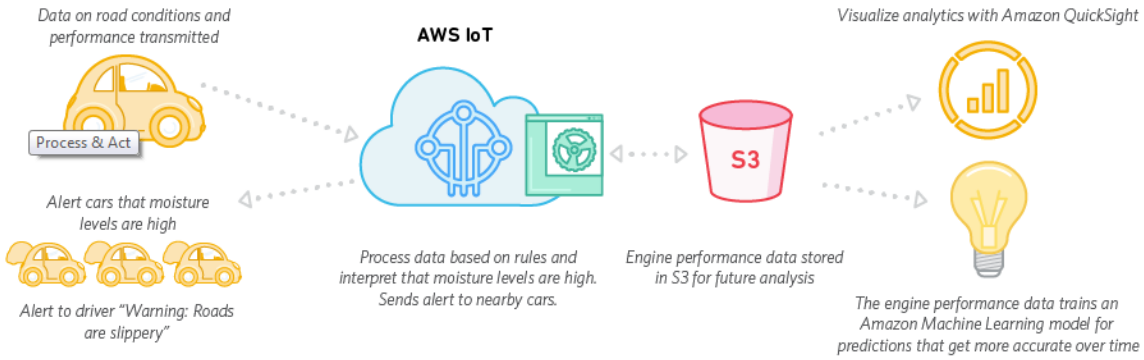


**Figure 2.6** AWS IoT makes it easy to use AWS services like AWS Lambda, Amazon Kinesis, Amazon S3, Amazon Machine Learning, Amazon DynamoDB, Amazon CloudWatch, and Amazon Elasticsearch Service for even more powerful IoT applications.

Source: (19)

Using very similar concepts, there are a lot of IoT platforms can be used to develop Internet of Things solutions and applications with their architectures having many

indispensable elements in common. In a high level perspective (see Figure 2.7) of these an IoT solution, we have at the lower level the IoT Infrastructure which are the inter-connected objects that sense from there environment. These devices communicate with the cloud and through a Data Gateway or API, are transmitting to it the sensed data. Abstractly, the cloud will then receive these data and it will attempt to manage, process, store and analyze them to ultimately transform them into knowledge. In data processing, the data are manipulated and brought in a suitable form in order to analyze them and in order to bring them in good form, suitable for analysis. In order to be managed, the data need to be stored and easily retrieved with simple queries. In addition, they are created Analytical and Machine Learning models that are trained with the data and transform them into information and knowledge. And of course, it is needed to pass this knowledge so it is visualized or presented in the best possible way to explore from it as much as possible.

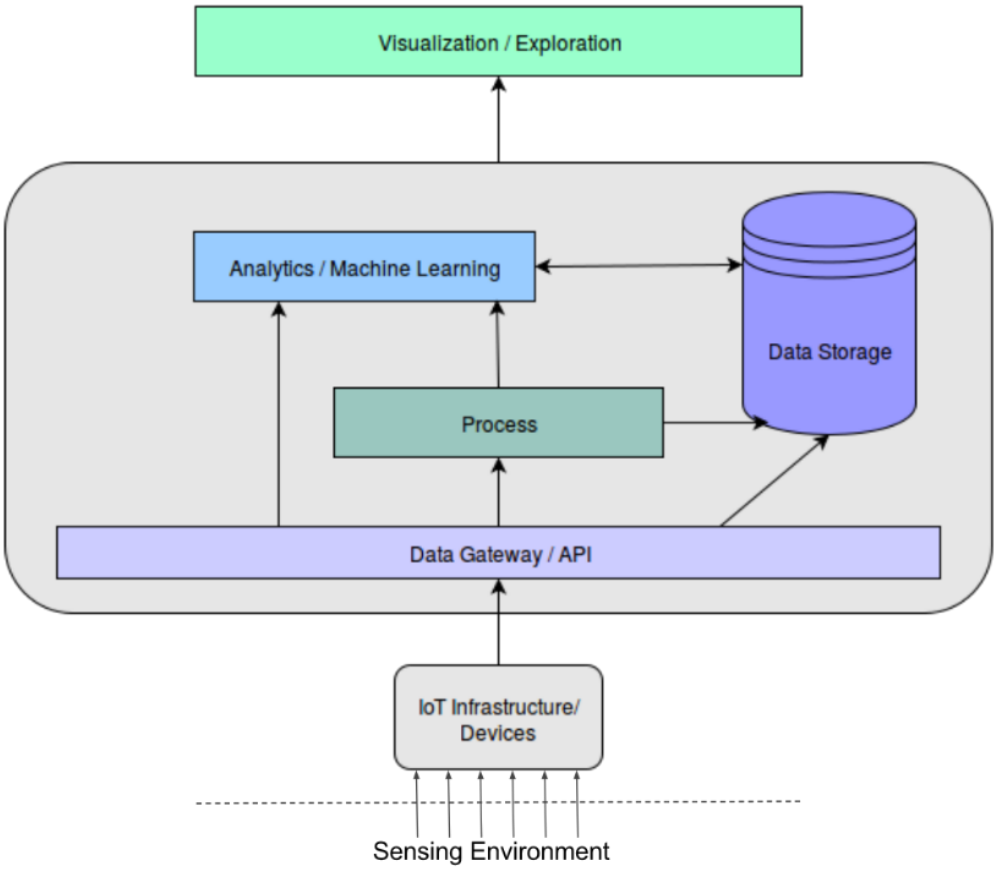


**Figure 2.7** Abstract architecture of IoT solutions

**Challenges & Motivations**

However, with this architecture, there are several pitfalls caused by the incompatible nature of IoT with the most commonly used cloud infrastructure (17). Devices are remote on edge and they create huge amounts of data, in high frequencies, that need to be sent through the network to reach the cloud, creating all kinds of bottlenecks in all the levels of analysis and communication. Hence it is required a particular background knowledge of how to deal with and analyse the big data generated in order to eventually transform them into knowledge. In addition, a pay-as-you-go model, it may be proved extremely expensive which is okay if it is created valuable knowledge, but a big percentage of the data sent are not considerably useful (18).

Hence, data need to start being processed and analysed from their birth and with significant operations of pre-processing in order to exclude and eliminate unnecessary data from the time they are created and reduce the dimensions of the data generated that need to be exchanged through the network for analysis (18). However, not all developers acquire the needed background knowledge related to data science and data engineering. Consequently, except of the enormous opportunities for development of IoT applications, there is a need of tools and frameworks to guide and support developers in order to successfully deal with the challenges of developing IoT applications and "it is of vast importance the deployment of large-scale, platform independent, wireless sensor network infrastructure that includes data management and processing, actuation and analytics" (2).

However, most of the devices that consist this sensor network infrastructure are battery powered and resources constrained. As a result, except from capabilities to manage, process, analyse data and actuate it is very critical to ensure energy efficiency in devices and always consider the possible costs. Beyond that, it is of vast importance to enable IoT applications with the ability to consider only the useful data and information and consequently, to use less resources for transmitting and receiving data and less effort to handle data bottlenecks that may possibly be created. Moreover, due to the big number of available IoT solutions, the devices and objects may run in a big variety of operating systems creating the need that development of data mining and edge mining solutions is in a platform and operating system independent manner.

# Chapter 3

## Proposition of an IoT specialized data mining framework

---

Architectural Design

Methodology for defining Requirements Specifications

Required knowledge and technologies

---

**Architectural Design**

One of the possible solutions is the proposed framework for low cost data mining for IoT. The particular framework, tries to assist developers of IoT Applications to face the challenges that are created with IoT and to successfully and efficiently cope with the huge amount of data generated (10). In addition, it attempts to enhance the development of IoT applications, with the establishment of early data mining techniques and algorithms that take place on the devices that sit on the edge of IoT. Except from attacking on the challenges that are resulted from the generation of big data of IoT, these techniques and algorithms, aim to improve the energy efficiency of devices. For example, it is intended to help distinguish between useful and un-useful data and consequently to regulate the ratio of transmitting data. Some of the benefits of applying early data mining is that it is not necessary for inter-connected objects to constantly communicate wirelessly and send unnecessary packages of data using the battery powered wireless transmitters in constantly on-mode. It enhances the try to keep only necessary elements of a device in a working state and only when it is really necessary - when it will have a valuable outcome.

In order to successfully do that, edge mining is essential and we need devices smart enough that contribute as much as possible in the data mining and data analytical requirements of transforming as fast as possible raw data into information. So, with the framework is intended to empower the development of edge mining and the transition of some part of the data mining from the cloud to the devices and interconnected objects that

sit on the edge of IoT (see Figure 3.1). However, it is very important to have a realistic and clear picture. At the moment, it is impossible for the devices with the offered edge mining to prevail over the data mining that happens on the cloud in terms of the quality of the knowledge extracted from data. The data processing and data analytics that take place on the cloud, use lots of resources in memory and processing power and consider multiple parameters. In addition, with the possible advanced programming techniques (e.g. distributed computing), the analytical capabilities are much higher than the possible capabilities of the low cost, in processing and memory, hardware that devices and inter-connected objects use. Finally, it is very critical for devices to execute every action timely and effectively. Therefore, the solutions offered by the framework may possibly do some compromises (e.g. in the analytical accuracy) in favour of these particularities of the devices and inter-connected objects.

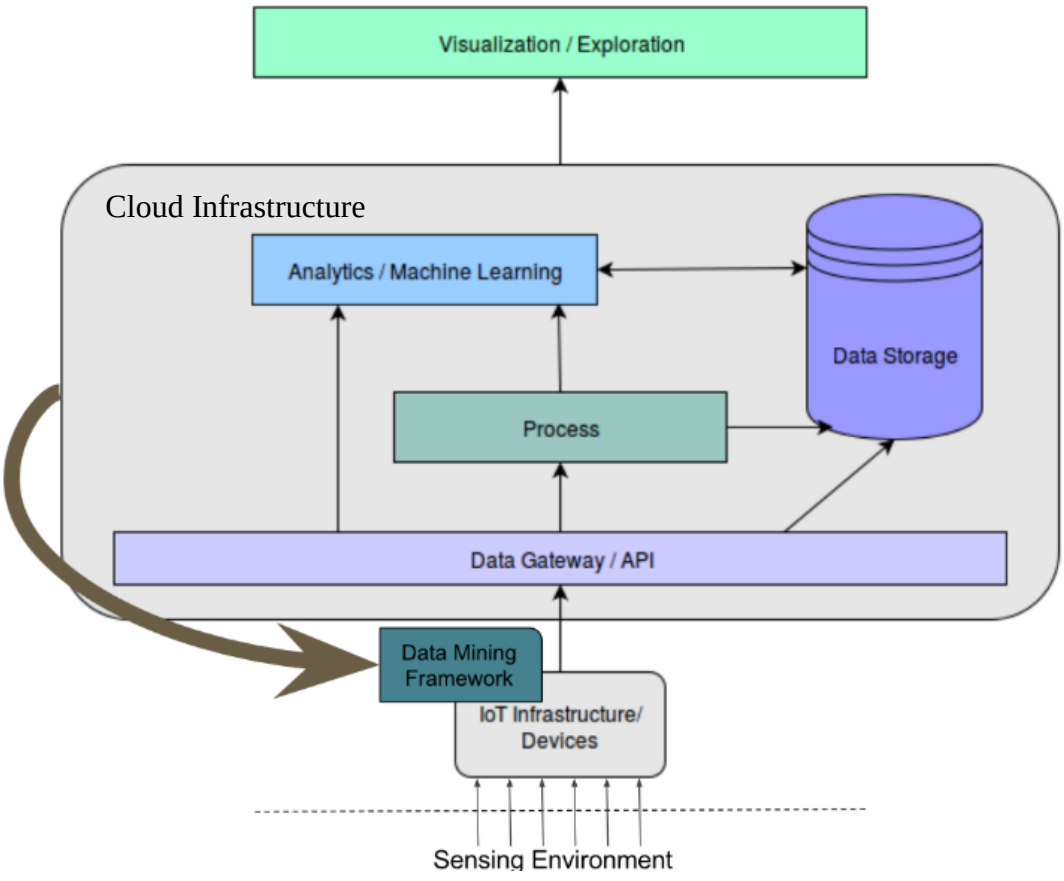


**Figure 3.1** Abstract architecture of where the proposed framework for low cost data mining for IoT takes place on the available IoT solutions

As a result, the framework does not try to replace the data mining work that happens in the cloud but contrariwise, tries to assist it with executing pre-processing on the data and with enabling an early analysis of data with available data mining and machine learning techniques that can operate in a low cost, in terms of processing and energy consumption, manner. Influenced by the already existing models and frameworks that exist on the cloud, the particular framework adopts a similar architecture, but instead having beneath it the cloud infrastructure, it has the infrastructure of the remote devices of IoT (see Figure 3.2). After receiving data that are sensed from the environment by the devices, through a data interface, the data are transformed into data types suitable for analysis. Then, the data can be processed and manipulated and eventually be stored in the data structures that are suitable for the data mining techniques and algorithms that will take place. These data structures will pass through models that execute data mining algorithms and techniques on them in order to mine the useful information from them. Finally, a model for data dissemination prepares the resulted information in a form that can be sent to the cloud for further analysis. In addition, the data dissemination is possible to take place even without executing any kind of processing or analysis on the data and transmit them even in their raw form.

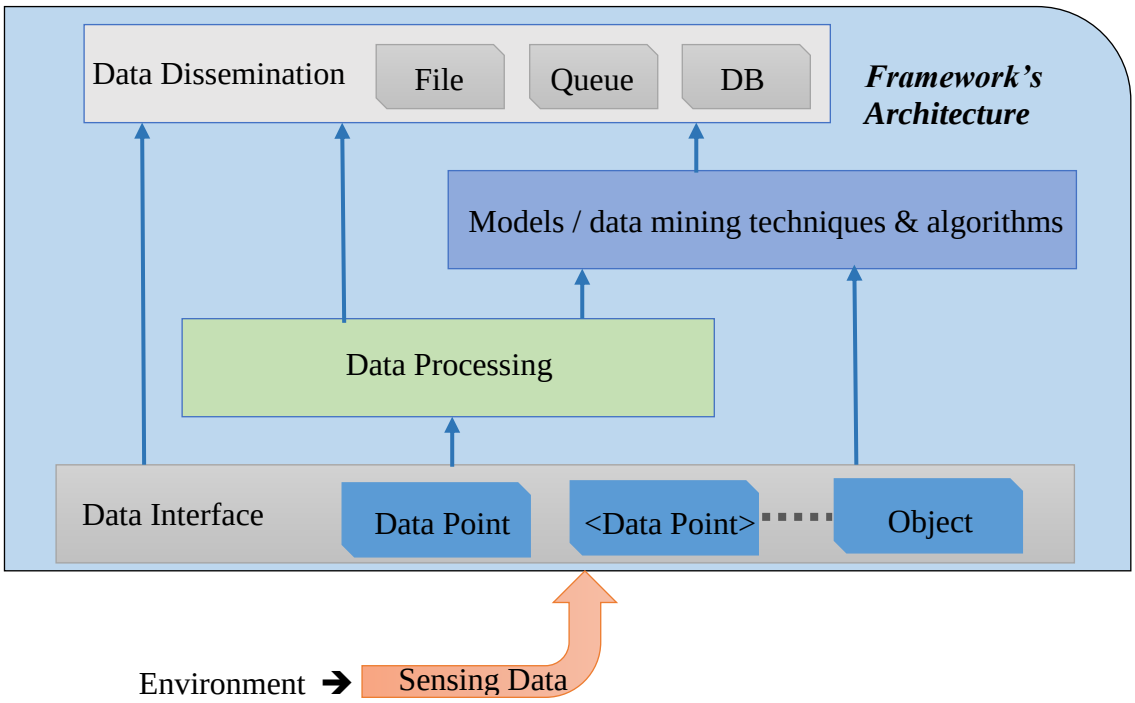


**Figure 3.2** Abstract architecture of the Data Mining framework for devices that sit on the edge of Internet of Things

**Methodology for defining Requirements**

As it was indicated, there are available solutions for developing IoT applications and robust solutions suitable for IoT that are able to successfully transform data into knowledge. However, even though it is possible the development of data mining models, which can significantly contribute on the transformation of data into valuable knowledge, Internet of Things requires some additional characteristics and are needed techniques and algorithms more specialized for these characteristics. Thus, the development of IoT applications with the specialized requirements of smart devices that with or without the aid of human intervention actuate accordingly and interact in good time, the existing technologies and solutions face a lot of challenges to overcome.

Hence, the requirements of the framework are strongly related to these challenges and with the particularities that devices and inter-connected objects on the edge of IoT have, as well as with the characteristics of Edge Mining. The objective is to contribute on the process of transforming raw data into knowledge and improve the energy efficiency and power consumption of the devices sitting on the edge of IoT, using smart ways to reduce the volume of the data that are transmitted for analysis. In addition, it is important the ease of use and extendibility of the framework in order to be used by developers and to easily create their specialized data mining models.

Furthermore, the basic elements and the functional requirements of the framework are strongly relevant with the phases of the data mining process. Specifically, in an abstract data mining process, we first need to gather data. Then, the data need to be processed and analysed in order to transform them into sensible information. Finally, the generated information is required to be disseminated or visualized. Thus, it is required to have a model that will receive data in various formats and a model for manipulating and transforming these data in the proper data types and data structures. In addition, it is needed an element in which intelligence is developed for mining information from the received data. This element, needs to give the ability to run already implemented data mining techniques and algorithms, to modify already implemented data mining models and to develop new models based on the available functionalities and techniques of already implemented models. Then, it is needed an element that will receive the resulted information and it will format them with a diversity of types and forms for dissemination.

**Specifications**

First of all, the framework is destined to be used by developers that will be able to use it as is and to extend and develop on it according to their specific requirements as well as their data mining objectives. Hence, the framework needs to be highly extensible and easy to use and further develop on it. In addition, it is required to distinguish every phase of the data mining process and keep independent their development and use. As a result, it is necessary that the framework is designed and implemented in a modular way that will clearly have separated the data mining phases.

In addition, the inter-connected objects and devices are remote on the network and they are generating constantly data, leading to the creation of huge amounts of data. This creates the requirement to enable the development and implementation of data mining and edge mining techniques and algorithms on devices and inter-connected objects. Hence, through the framework it is necessary to give abilities of implementing data mining and data analytics models to developers of IoT applications. Consequently, it is required that the framework will enhance the development of analytical models that will receive raw data as input and will give information as output for dissemination. This way, it will give more capabilities for handling and dealing with the big amount of data that the inter-connected objects and devices create.

However, the interconnected objects and devices are mostly powered by batteries and they are resources constraint. As a result, it is required to consider energy efficiency and try to improve the amount of the resources that the device consume. Hence, in favour of the energy consumption, it is very possible to undertake some assumptions and compromises that may be needed to take. For example, to make a prediction with confidence of accuracy close to 90% may be very demanding and costly in processing, time and consequently in energy consumption. But, it is possible to find techniques and algorithms that their developed knowledge is on approximation but they are very efficient in processing, time and energy consumption costs. For example, it can run on a device, to reach and make a prediction with confidence of accuracy close to 60% instead of 90% but it will not drain its battery. As a result, it could be the case that there is a compromise

in the accuracy of the results after applying edge mining. Hence, the data are being processed and analysed in order to be transformed into a premature type of knowledge.

Finally, due to the diversity of IoT platforms and IoT solutions that exist in the market, it is required that the framework can run in most of them with few dependencies. IoT devices may be Arduinos, Rasberry Pis, Microsoft devices or any other company specific devices. In addition, the devices it is possible to run in various Operating Systems like Android, iOS, Windows, Linux etc. As a result, it is required that the framework is implemented, run and built in a platform and OS independent way. The framework needs to have abilities to be used with very few or even no prerequisites and dependencies of other systems. Considering all these facts and the challenges that consequently come out of their combination, it is critical that developers find in the framework the best possible compromises and trade-offs that will bring results of good quality. The framework's purpose is to help the development of viable and feasible applications, which means it will always try to make the best possible processing and mining on the data with the available resources and energy consumption. Thus, it is created a trade-off between accuracy on the quality of analysis and the cost of this analysis in terms of energy consumption, processing and memory. In addition, the usefulness of the framework is relied on the offering of a great degree of specializing it according to the requirements of the developed IoT application. All models can be easily extended and specifically developed and it can run in any hardware that can run Java.

**Required knowledge and technologies**

As a result, the programming language that was chosen for the development of the framework is Java. Some of the very primary characteristics of Java is simplicity, dynamic, Object Oriented, high performance, portable, architecture neutral and secure. All these characteristics are essential for such a framework as the developed one. The Data Mining framework for devices that sit on the edge of Internet of Things, needs to be modular with all its elements and phases of the data mining distinguished and separated. The Object Oriented structure and design of Java, is the mean to reach the required modularity due to the introduced abilities to logically group and package the different software classes and objects that are created to offer the various functionalities and

attributes of entities. In addition, the logically grouped and packaged software classes and objects, may be called in many other places of the software and easily reused as independently implemented components or even be extended in order to provide new and more specialized behavior. Hence, the framework is built with Maven and all the dependencies and needed regulations are handled automatically.

One of the most important requirements, is that the framework has the ability to run and be used in a platform and Operating System independent manner. The way java works in combination with the way that Maven builds the framework, allow to overcome this challenge and create software with abilities to work in any device and operating system. Specifically, with the java software that any java program can run, comes the Java Runtime Environment which consists of the Java Virtual Machine, some standard core classes and some standard supporting libraries (19) (20). Any java software can run on any operating system and any processing hardware because Java's Runtime Environment creates a virtual machine (JVM (19)) that sits on top of the used hardware and OS and can execute any compiled java code. A software's source code written in Java, is stored in a .java file and this file is passed in the java compiler for compilation, resulting in the creation of the binary .class file. The compiled .class file does not contain code native to any processor, but it consists of bytecodes which is the machine language of the Java Virtual Machine (19) (see Figure 3.3).

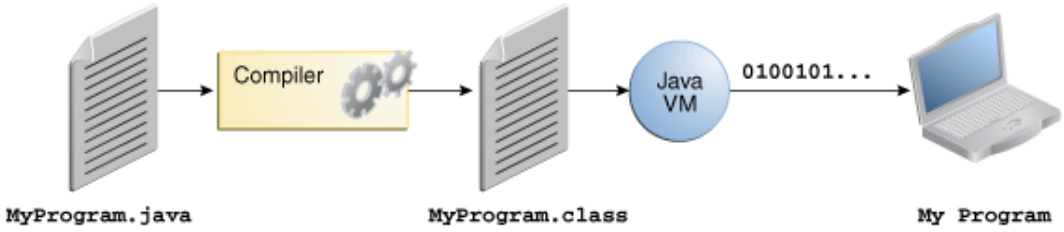


**Figure 3.3** Overview of the software development process in java

Source:

Hence, because the execution of software is handled by the JVM which is available in all the operating systems that run in all available processing hardware, it does not make any difference the kind of the Operating System that runs (see Figure 3.4).

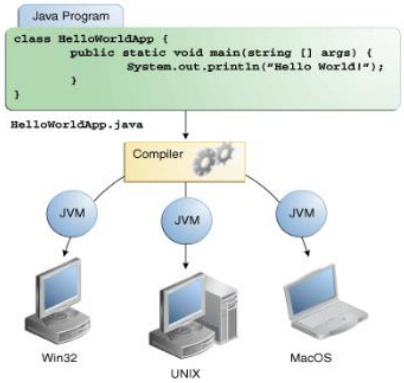


**Figure 3.4** JVM enables the execution of the same software in different platforms

Source:

This is achieved with the assembly of the java platform. The java platform is basically a software-only platform that runs on top of all the other available hardware-based platforms such us Microsoft Windows, Linux and Mac OS. The java platform of course contains the JVM along with a java Application Programming Interface (API) (see Figure 3.5), which is a large collection of existing software components that may offer a lot of useful capabilities. The software components are basically packages that are the grouped libraries of related classes and interfaces. As a result, the framework is developed in a way that its components can independently run in any operating system a device may have as long as it has installed the Java Runtime Environment. In addition, because of the features offered by Maven of handling dependencies and the big community of Maven repositories, the framework will have the ability later on to be used in the same exactly way as the built in components.

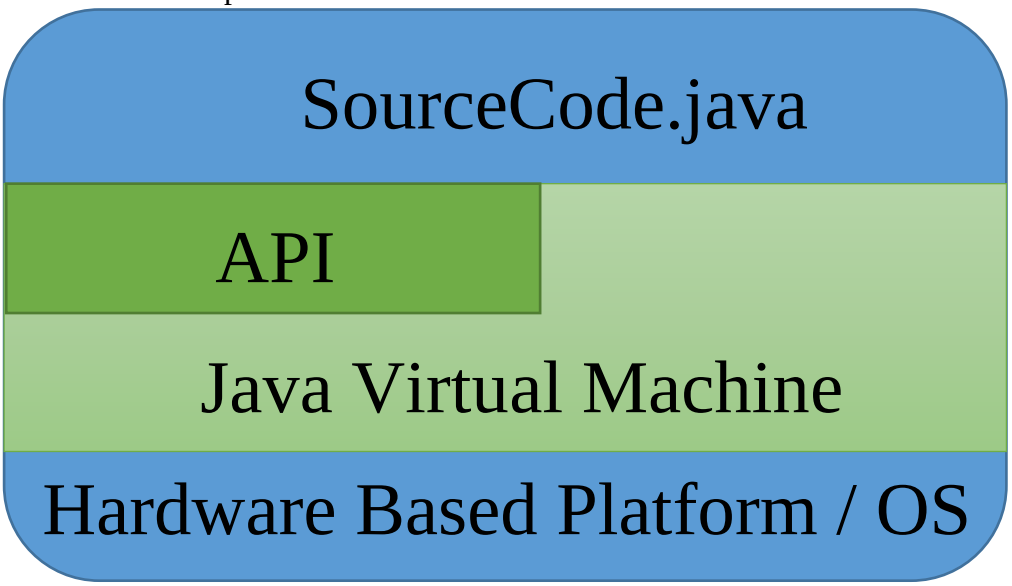


**Figure 3.5** JVM and Java API

Another requirement of vast importance and one of the objectives of the framework is to contribute on the transformation of data into knowledge timely, effectively and efficiently. That means that the raw data received need to be processed and analyzed in order to make some sense out of them and create information and some kind of knowledge from them. Thus, Big Data Analytics and Data Mining techniques and algorithms are momentous in order to successfully do that.

As a result, it is essential to conceptually understand perfectly the possible data mining models that can be implemented. There are many available data mining techniques and algorithms suitable for different needs and requirements (see Figure 3.6), like Classification, Clustering, Divide and Conquer, Incremental Learning, Sampling, Filtering, Data Condensation, Frequent Patterns, Outlier detectors etc..

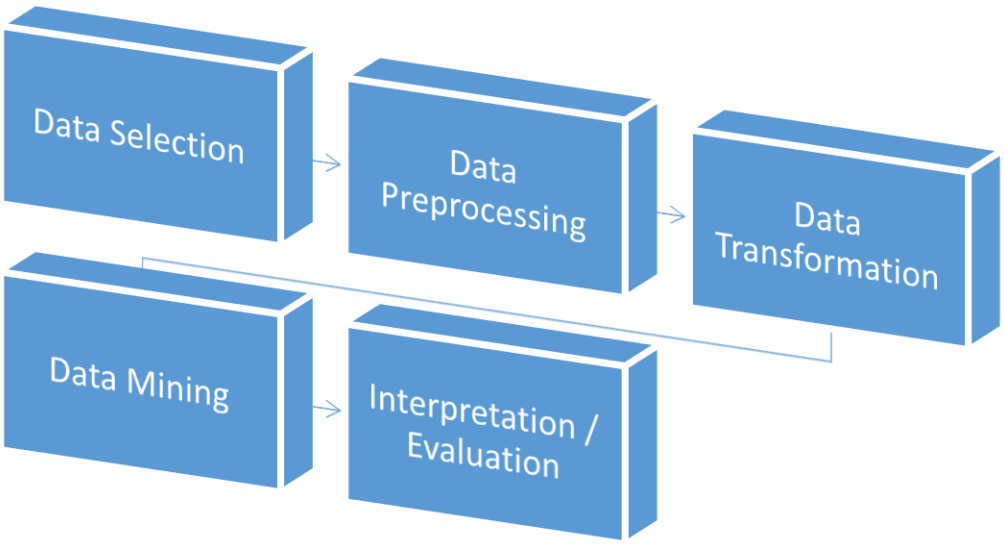


**Figure 3.6** Data Mining Process

Source: (1)

# Chapter 4

## Implementation of the IoT specialized data mining framework

---

Design and Interfaces
Data Interface
Data Processing
Data Mining Techniques & Algorithms
Data Dissemination
Results

---

**Design and Interfaces**

For the developed data mining algorithms and techniques that consist the framework, it is followed a modular design that can realize the use, development and implementation of models for specific data mining tasks and activities. The modular design, give to the developer, that tries to implement a data mining solution for IoT, abilities that allow him/her to extend an already implemented model of a data mining activity and develop a more specific and specialized model aligned to the data mining requirements of the particular problem. Additionally, it is possible for anyone to implement the offered interfaces and further develop models of algorithms and techniques that are not already implemented, offering multiple options for a solution to a data mining activity.

Furthermore, for a more comprehensive analysis and mining of information, it is possible the chaining of models of data mining activities. All the aspects implemented are independent from each other and there is a distinct separation of concepts in the framework. For example, algorithms, data types and data dissemination models as well as every other aspect of a solution, is an independent model that can be extended and further developed. Hence, these independent models, can be configured in a sequential order where the output of the preceding models can be passed to the following ones.

In the fundamental use of the framework (see Figure 4.1), it is required to develop a model for receiving data from a data source, a model for a first processing of these data and the creation of the proper data types and data structures and of course, a model with the different data mining algorithms and techniques that are going to be used in order to transform the data of a given IoT application in more sensible information. Finally, on the last part of the models' chaining that consist the fundamental use of the framework, there is a model that forms the resulted information in order to be disseminated. Thus, each of the implemented models and extensions of a model can be used as a component for reuse and can be differently combined in any other data mining solution.

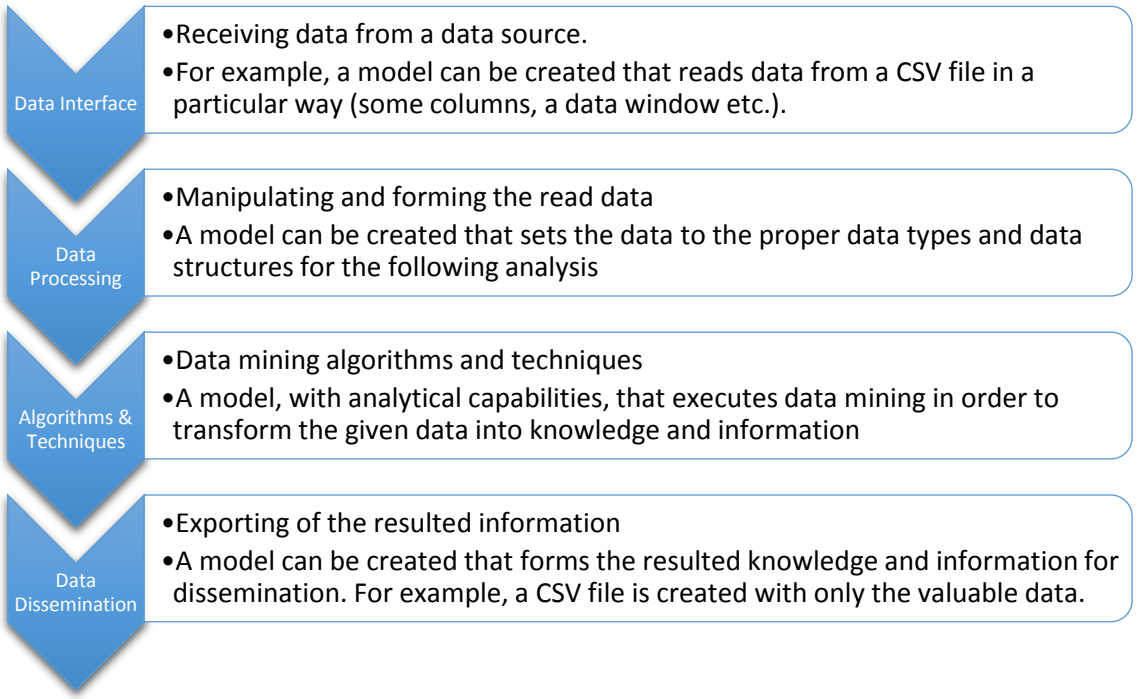


**Figure 4.1** The design of the fundamental use of the framework. One model is created for each one of the four activities.

### Data Interface & Data Dissemination

In IoT, the inter-connected devices and sensors, constantly sense and generate raw data. Thus, the framework first of all needs to give abilities to receive these raw data and it offers the development of data interfaces for the in development IoT application. Specifically, it is possible to implement the baseline of reading raw data and set the detailed specifications of the required attributes that represent a data entity. For example,

an inter-connected object of an IoT application is a module of a smart wearable (e.g. fitbit). This, smart object, senses and generates various types of data. For example, it may sense the heart rate of the person wearing it and consecutively (every five minutes) provide metrics for it. It may also count how many steps and the distance walked and provide these counts in a basis of five minutes. Hence, with all the enabled sensors, are generated data that represent many attributes that may be relevant or irrelevant with the required knowledge to be extracted.

With the framework, it is possible to develop a data interface that can be used to define which of those attributes are valuable and need to be considered. A developed data interface, can take a certain form of a resource that contains data and read from it. It is already implemented a data interface for reading from a CSV file and it is possible to be extended in order to work with other types of resources that contain data (see Figure 4.2).

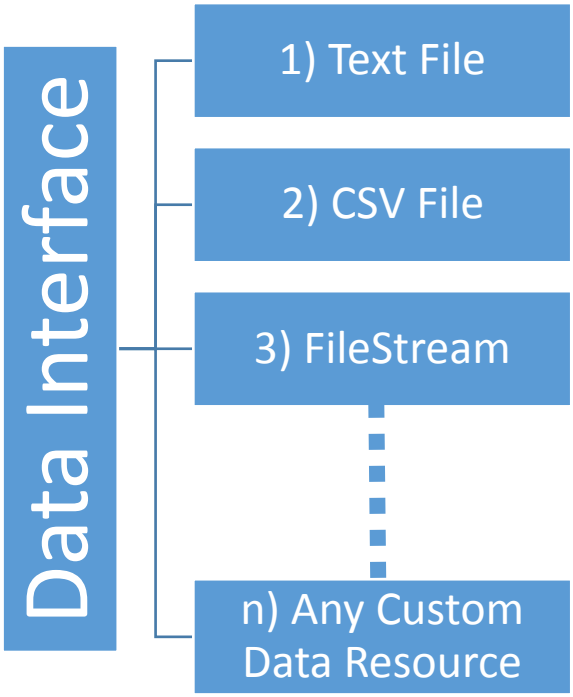


**Figure 4.2** It is possible to develop on the Data Interface a variety of models for receiving data.

From the time that the data are read, it is possible to execute a selection on the attributes that are most relevant and ignore the rest. It can be predefined, the required attributes or columns that need to be analysed. However, a developed data interface (see Table 4.1),

enables the receiving of raw data that cannot stand by themselves and therefore enable their mining.

| Modifier and Type | Method and Description |
|---|---|
| void | closeFile() |
| void | getLine(String filename, String delimiter, int num) |
| void | openFile(String filename) |
| void | readAll(String filename, String delimiter) |
| ArrayList of String | readData(ArrayList of DataPoint data, String delimiter, int[] columns, int lblCol, DataPointType type) |
| ArrayList of DataPoint | readData(String delimiter, int column, DataPointType type) |
| void | readSome(String filename, String delimiter, int limit) |

**Table 4.1** API for Data Interface. *Interface DataResource*

Finally, after the data have been received as well as processed and data mining techniques and algorithms have been executed on them transforming them into knowledge, the resulted sensible information need to be disseminated in order to be transmitted to the cloud. The resulted information and knowledge, are the data that after the analysis are considered valuable. Hence, with the framework, it is possible to develop a data dissemination model that can be used to define the form that the resulted information are going to be disseminated. A developed data dissemination model, can take the resulted valuable information and write them in a certain form for transmitting them to the cloud for further analysis, because the data mining executed on them is just for preparing them and cleaning them from non-useful ones. The data dissemination model follows a similar modular design and it is already implemented for disseminating in CSV form and it is possible to be extended in order to work with other data forms (e.g. queue).

**Data Processing**

Hence, with the framework it is possible to develop models of data processing that manipulate the raw data and create the proper data types and data structures. This manipulation of the data into specific data types and data structures, will enable the analysis and the execution of data mining techniques and algorithms on the data. After the proper data interface model is created, it is required to form the data in types and structures that will contribute as much as possible in their analysis in order to extract from

them sensible information. Each data point creates an object with the name that represents it, a timestamp that represents the time of its occurrence, a unique id and of course, its value (see Table 4.2).

| Modifier and Type | Method and Description |
|---|---|
| abstract Label | getLabel() Assigned label of training data. |
| String | getName() |
| int | getSequenceID() |
| java.sql.Timestamp | getTimestamp() |
| DataPointType | getType() |
| java.lang.Object | getValue() |
| abstract java.lang.Object | getValue(java.lang.String column) Get sample data value from specified column. |
| boolean | has(Feature feature) |
| void | setName(java.lang.String name) |
| void | setSequenceID(int seq) |
| void | setTimestamp(java.sql.Timestamp timestamp) |
| void | setType(DataPointType type) |

**Table 4.2** API for Data Processing. *Abstract Class DataPoint*

The type of the value may vary for each case and consequently, there are capabilities for the generic creation of data points according to any type that the given data are. For example, a metric of the heart rate for the average adult is between 60 and 100 beats per minute that means that the values of a heart rate's data point is an integer. Other metrics may require data points of double types of value and others of String types of value and so on. In addition, there are data points that represent more than one dimensions, for example location or directions, creating requirements for data points with vectors of data types as values. Hence, there are capabilities, for example, to create a data point that represents location and it is a vector of doubles (a double for Longitude and a double for Latitude). Thus, it is created a model of data points that may be easily extended and develop on it any custom type of data and keeps the data mining capabilities intact.

Most importantly, it is maintained a significant ease of use and simplicity in the creation of these data points in all the aspects of the framework and their creation and use is clearly separated and totally independent with any other aspect, either is the models for data

interface, or the models for the data mining techniques and algorithms. This, implements and realizes the modular requirements of the framework (see Figure 4.3). The connection of the data point's model and the data mining model is irrelevant with the type of the data that are passed. The data mining models of techniques and algorithms, understand and can work with any given object that has and implements the attributes and characteristics of a data point.

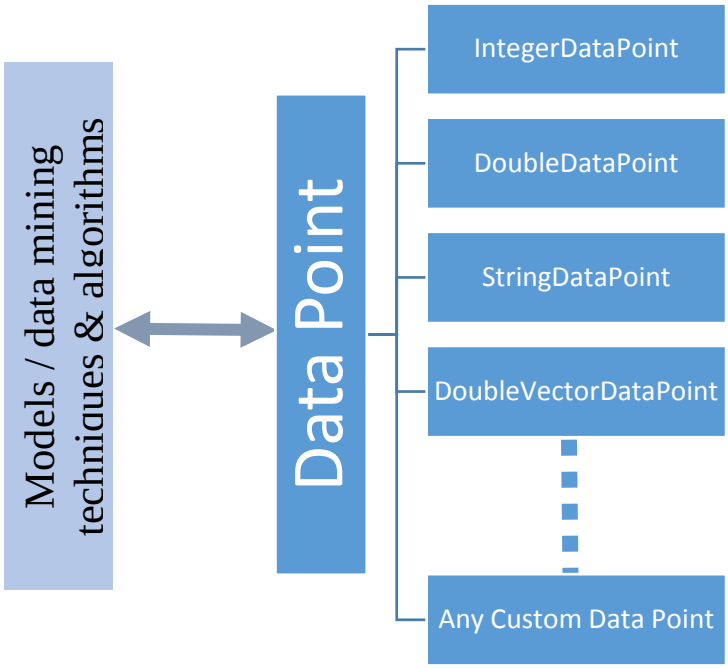


**Figure 4.3** It is possible to develop on the Data Point a variety of models for forming and structuring data.

**Data Mining Techniques & Algorithms**

Moreover, once the data processing models are created and the received data are set into the proper data types, it is possible the creation of models with capabilities of data mining techniques and algorithms. The algorithms and techniques are distinguished and grouped according to their types and they are designed and implemented in a modular way in order to be clearly separated (see Figure 4.4). For example, all the clustering algorithms are grouped together and they have their common interface that is used to call them, pass data and execute them. This design is adopted for all the data mining algorithms and techniques and resulted to the creation of a model for classification algorithms, a model for data mining techniques that work with time-series and so on.

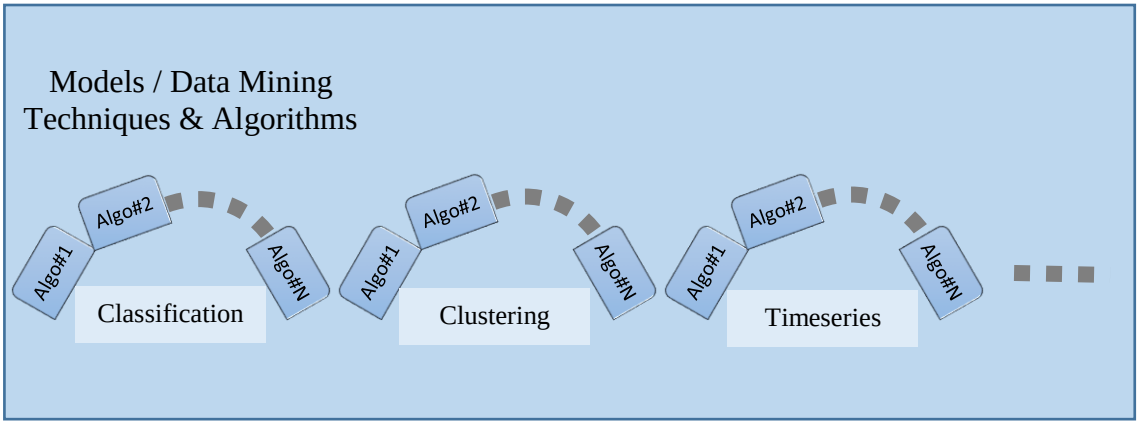


**Figure 4.4** It is possible to develop on the Data Mining Model a variety of models with different algorithms and techniques.

With classification techniques, data are classified / separated into various classes according to a model trained with classified data. With $X$ a set of feature vectors, $C$ a set of classes and $c: X \rightarrow C$ the ideal classifier for $X$, there is a set of examples $D = \{(x_1, c(x_1)), ..., (x_n, c(x_n))\} \subseteq X \times C$. Given the set of examples $D$, a model function can be formulated that can get any vector of features and determine in which class from $C$ belongs to, with some approximation. An example of such a model is to use supervised machine learning to create and train a neural network[20] that separates data into various classes. Another example of a classification technique is that of Decision Tree. Given a set of possible features $X$, and a set of classified examples $D$, the creation of a decision tree executes a splitting of $X$, into subsets $X_1,..., X_n$ and a partitioning of $D$ into subsets $D_1,..., D_n$. Then, a tree node is created for $D_j$, where j=1,…, s and we have $\{(x, c(x)) \in D \mid x \in Xj\}$. In simple words, for a possible feature from the set, is created one tree node, splitting that way the imported training data into j sub-lists. Same thing is repeated until the tree is created (21).

A decision tree T, for X and C, is a tree with finite number of nodes. It contains a distinguished root node and each non-leaf node t of T has a set $X(t) \subseteq X$ assigned, a splitting of X(t) and a one-to-one mapping of the subsets of the splitting to its successors. Each leaf node of T has assigned a class from C. The classification of an $x \in X$ determines a unique path to a leaf node of T beginning from the root node of T. At each non-leaf node a particular feature of x is evaluated in order to find the next node along with a possible next feature to be analysed. Hence, with each possible path from the root node

to a leaf node, the features that correspond to the nodes of the path are creating a sequence of values that are successively tested and formulate decision rules (21).

Abstract Class DecisionTree

Package: Framework.algorithm.classification.decisiontree.DecisionTree

| Modifier and Type | Method and Description |
|---|---|
| Label | classify(DataPoint dataSample) Classify dataSample. |
| Node | getRoot() Get root. |
| void | printSubtree(Node node) |
| void | printTree() |
| void | train(List of DataPoint trainingData, List of Feature features) Trains tree on training data for provided |

Interface ImpurityCalculationMethod

Package: Framework.algorithm.classification.decisiontree.impurity

| Modifier and Type | Method and Description |
|---|---|
| double | calculateImpurity(List of DataPoint splitData) Calculates impurity value. |
| default | double getEmpiricalProbability(List of DataPoint splitData, Label positive, Label negative) Calculate and return empirical probability of positive class. |

Abstract Class Label

Package: Framework.algorithm.classification.decisiontree.label

| Modifier and Type | Method and Description |
|---|---|
| abstract boolean | equals(Object o) Force overriding equals. |
| abstract String | getName() |
| abstract String | getPrintValue() Label value used to print to predictions output. |
| abstract int | hashCode() Force overriding hashCode. |

**Table 4.2** API for the Data Mining Model of Decision Tree.

Except from the classification data mining model that implements a Decision Tree, the framework also implements other techniques and algorithms. A Clustering model is developed implementing kMeans along with various other techniques that analyse time-series. Specifically, is implemented an extensible model for calculating the simple moving average, the Cumulative Moving average, EWMA and PEWMA. The combination and the use of those techniques and algorithms, give the ability to develop many other models. For example, data mining models can be developed for prediction and anomaly detection.

# Chapter 5

## Results & Use Case

---

Use Case
Results

---

**Use Case**

With the framework, it is possible the development of models that represent data mining and edge mining activities. Hence, a significant part of the required data mining is transformed to the devices on the edge of IoT and Edge Mining is enabled. A use case of the framework is, for example, the development of an IoT application where data are sensed from a wearable and is required to define how intensive are the activities and the exercises of the person/athlete wearing it (22). In the simplest form of this application the wearable, is enabled with sensors that sense the heart rate of the user constantly. With the most naïve approach, we can say that the device sends the heart rate of the person wearing it every a short period of time. For example, every minute the device records and transmits to the cloud one package of data with the heart rate of the person. But, if we consider that all or even half of the professional football players of the world, which are 265 million footballers (23), wear such a device then we have a very big congestion of data to be transmitted and analysed from the cloud, making impossible the viability of the application.

Following a different and smarter approach, we can say that the heart rate of the person wearing it is transmitted to the cloud only when necessary and when it changes from its previous time sent. Specifically, we can create a data interface to receive the person's heart rate and develop an edge mining model that implements the PEWMA. In our developed model, we can predict each time the expected value for the following heart rate with a one step ahead prediction technique. Then, using the standard deviation, we can see how much differentiates the actual value of the heart rate with the expected value that

our developed edge mining model predicted. And of course, with the Standard Deviation our model can define if the actual heart rate that it received is an abrupt change on the heart rate's time-series. Hence, our wearable can decide whether it is worthy to transmit a heart rate metric or not and the transmitted data are reduced significantly.

However, there are still a lot of data for the Cloud to analyse and the battery powered wearables struggle to complete a good period of time in one charge, mostly due to the amount of time that they need to be connected wirelessly to the internet and send them data. In addition, anyone can make the observation that an athlete, even if he/she wears the wearable 24 hours a day, it is impossible to need metrics of all the day. For example, most of the people dedicate a big part (the one third – 1/3) of their days to sleep and almost all of them, do not require their heart rate while sleeping. As a result, extending our already improved solution, we may create another edge mining model that classifies the activity of the person wearing it according to the metrics of his/her heart rate. A good example of this classification could be to distinguish with each heart rate, whether the person wearing the wearable is a) sleeping, b) awake but not working out, c) working out and d) intensively working out.

However, such a classification needs a lot of training and consequently, it is costly. For example, it could possibly consume a lot of energy and require a lot of processing power. Thus, it would be more preferable, that our second edge mining model, compromise in the accuracy and preciseness of the classification in order to lower the cost of processing and energy consumption. Consequently, our model classifies and distinguishes only if the athlete wearing the wearable is sleeping or not.

As a result, with the use of the framework in the specific use case, we can reach a significant reduction of the data transmitted to the cloud, by only sending the suddenly changed heart rates that are classified to a specific type of activity. Finally, depending from the requirements of the application, we can create the specific model and specify the type and form on which we will disseminate the resulted information.

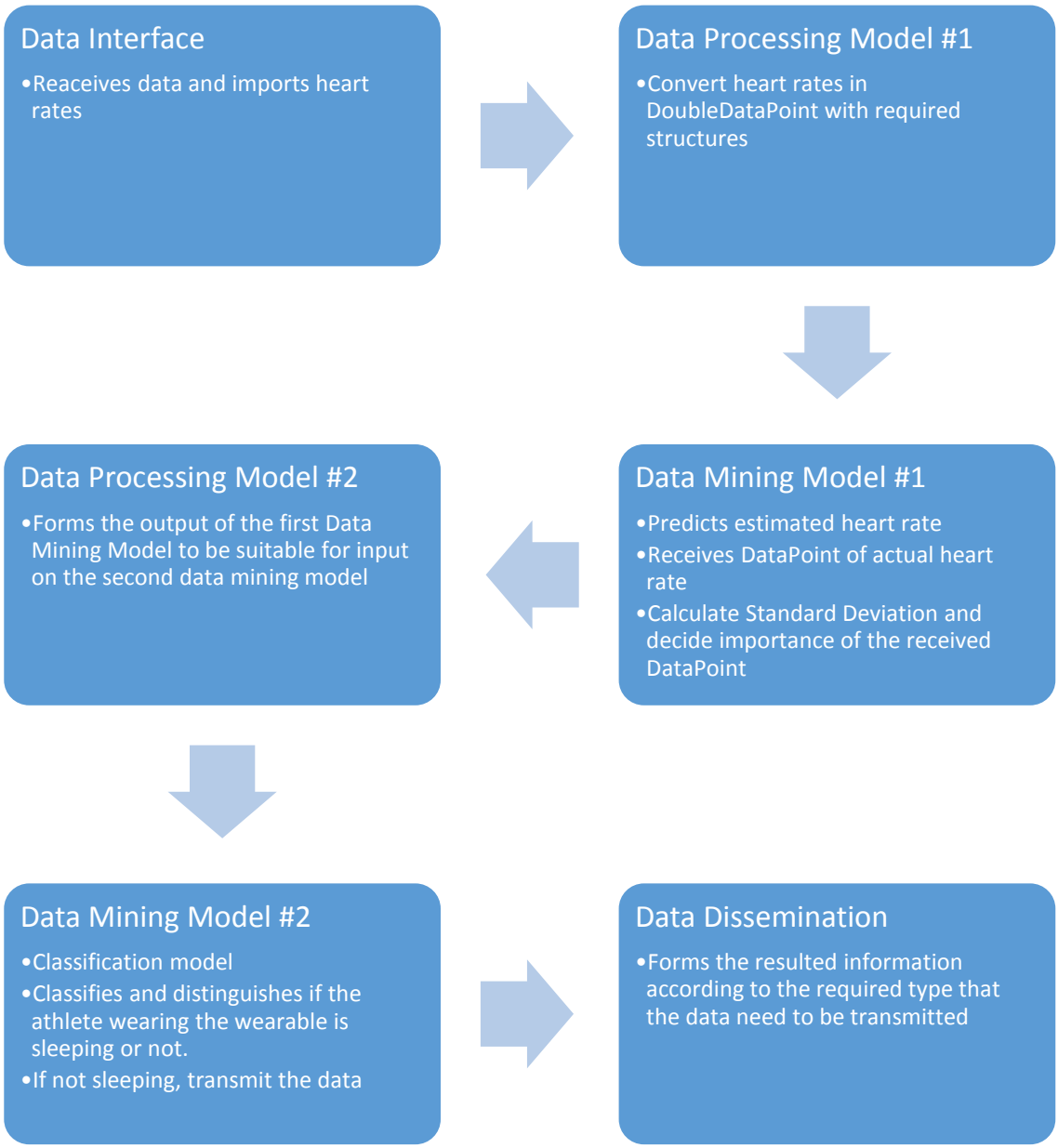


**Figure 5.1** Representation of a use case of the framework where wearables sense data from athletes and identify if they are useful to send or not

**Results**

After running the example, it is very interesting to see the improvement on the frequency that data are disseminated to the cloud. With the first, naïve approach where the data points are disseminated to the cloud, independently from their value, there is a congestion of data points to be sent (see Figure 5.2). With only applying a selection when an abrupt value appears in the data set, the data frequency and data ratio is dropped to approximately

1/3 (see Figure 5.3). In addition, it is very clear in the graph, that the data are disseminated mostly when there is an abrupt change and a pattern change on the exercise rate of the athlete.

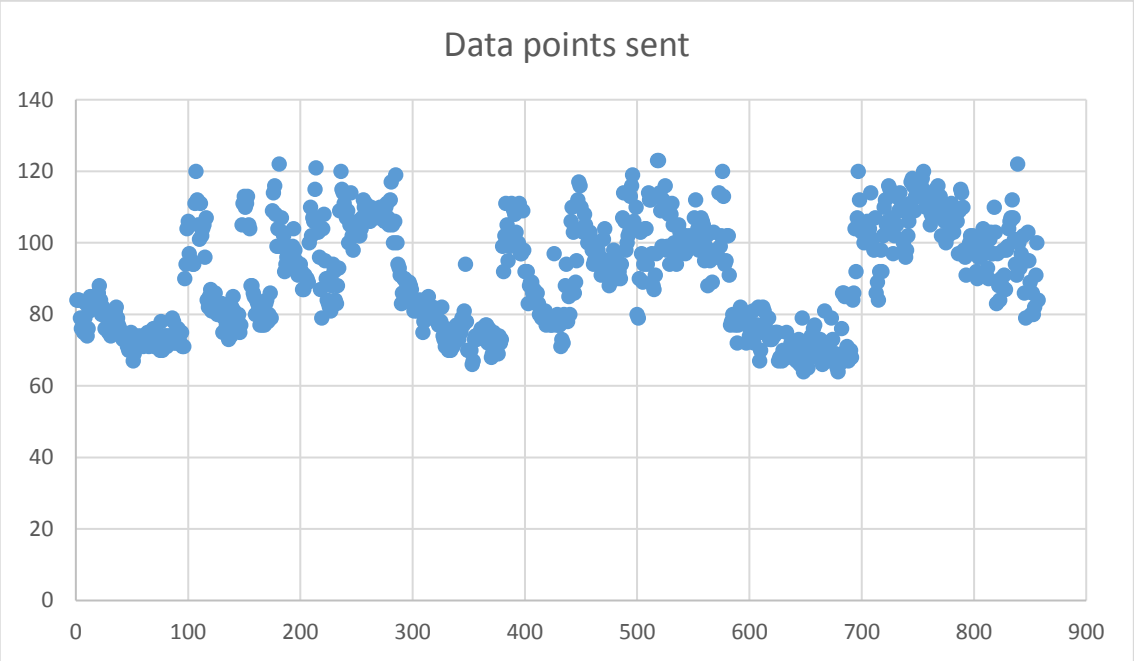


**Figure 5.2** Data points transmitted to the cloud for further analysis with the Naïve approach

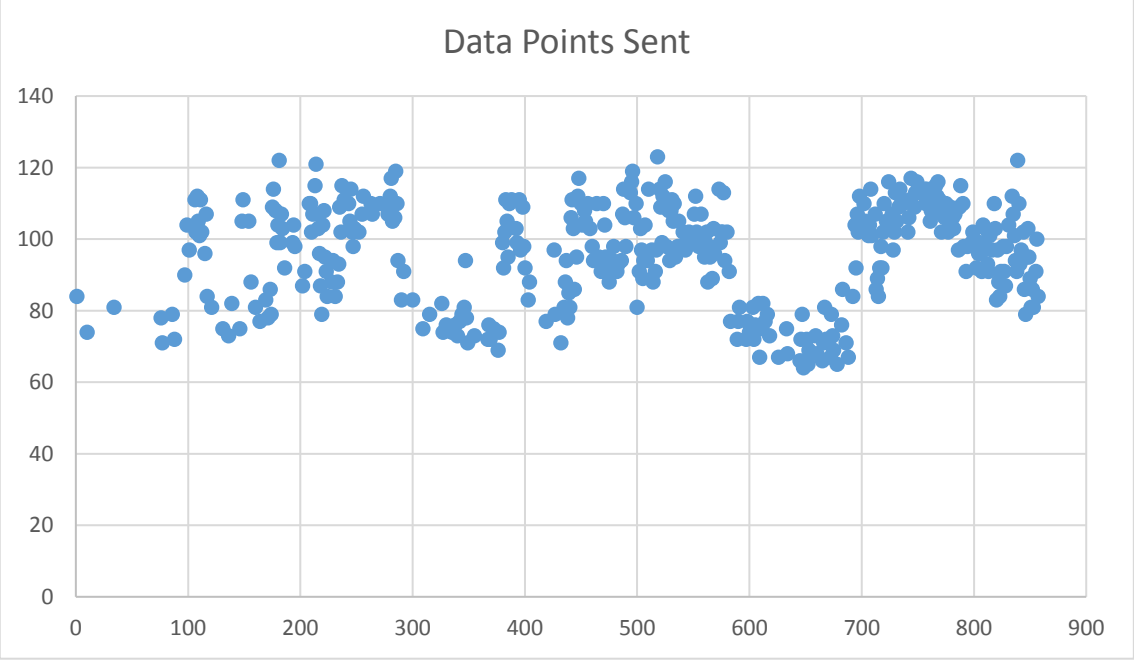


**Figure 5.3** Data points transmitted to the cloud after filtered with a timeseries technique based on moving average technique (PEWMA)

Finally, after applying a classification technique on the data, it can be seen (figure 5.4), that the volume and velocity of the data are even further reduced. It is very interesting to note that the pattern of the data reduction, it indicates the night hours most probably, where the person wearing the wearable was sleeping.

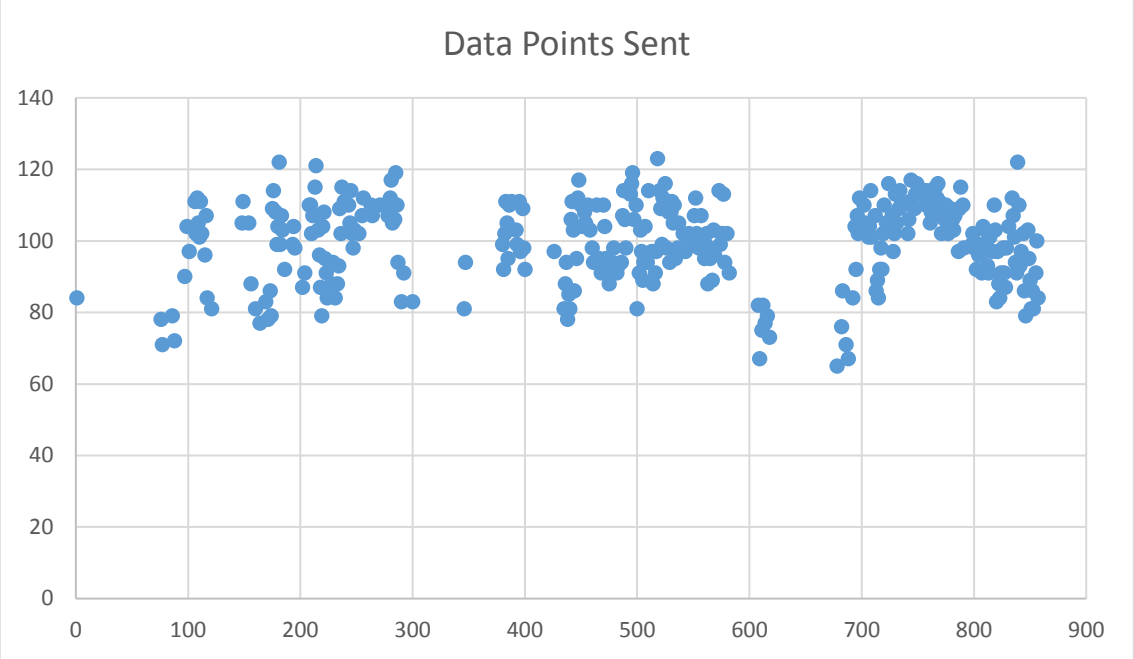


**Figure 5.4** Data points transmitted to the cloud for further analysis with the Naïve approach

# Chapter 6

## Conclusion and Open Challenges

---

Conclusion
Discussion

---

**Conclusion**

The devices and objects of IoT, are blended in any environment and they interact and cooperate with each other in order to reach common goals. In order to percept and understand their surroundings, and consequently, create intelligence and features of decision making, the objects sense and gather data from their environments. Data that need to be transformed into sensible information and knowledge. However, the generated data are so big and valuable, that it is needed the development of technologies to deal with knowledge extraction from this enormous amount of data.

Hence, there is a big variety of tools and frameworks for the creation of the proper analytical models that will successfully transform the data generated into useful knowledge and information and a lot of IoT solutions. Many data mining solutions and frameworks are very good in order to benchmark and compare data mining algorithms and other solutions are ideal in order to experiment with different data mining techniques and evaluate them in order to conclude on the most suitable algorithms and techniques. In addition, they are very good in giving the opportunity to modify and create new techniques and algorithms based on the existing ones and with the development of data mining models, they can constitute a significant contribution on the transformation of data into valuable knowledge.

However, Internet of Things require some additional characteristics and are needed techniques and algorithms more specialized for these characteristics. For example, the

development of devices enabled with the required intelligence to analyse data and with or without the aid of human intervention to actuate accordingly and interact in good time, require more specialized techniques and algorithms. Techniques and algorithms that will also contribute to decentralize the processing and analysis in order to successfully deal with the big data of Internet of Things (15). And that is where the developed Data Mining Framework contributes.

The developed Data Mining Framework for IoT, enhances the transmission of Data Mining and data processing, to the edge of IoT applications. IoT applications may generate such a huge amount of data that it is extremely difficult for the Cloud to analyse them and convert them into knowledge. In addition, a big percentage of those data are useless for the required conversion into information and knowledge and their analysis is unnecessarily expensive (in terms of how valuable they are). The particular framework, limits the unnecessary data transmission from IoT devices to the cloud infrastructure and it enhances the move of a significant part of the data mining and processing from the cloud infrastructure on the IoT device itself, trying to face the challenges of dealing with the enormous amount of data generated by IoT applications and mining knowledge from those data from the time of their creation.

It operates on a platform and OS independent manner, and has low cost on complexity and processing analysis features and characteristics. Hence, the viability and feasibility of IoT applications can be vastly improved. In addition, the extensible and universal architecture, design and implementation can result in the creation of a community to further develop edge mining techniques and architectures on the framework.

**Discussion**

A very good addition on the implemented framework, would definitely be the development of more data mining algorithms and techniques. In addition, it could be very useful the development of a benchmarking tool with some standard use cases, to compare the efficiency and effectiveness of data mining techniques and algorithms. Consequently, it is suggested the creation of a repository with different use cases and possible scenarios of applications that could benefit from the framework.

In addition, the framework covers a part of the overall Data Mining process of an IoT Application. However, the development and applicability of IoT is relevant to also other aspects that the current framework doesn't reach. For example, it is possible the development and implementation of a complementary middleware, that manages many devices that carry the Data Mining framework for IoT and connects them directly to devices that communicate with users. Hence, these end-users' devices, may have directly visualized information, not perfectly accurate, that are derived from the devices themselves. Hence, it is possible the further expansion in other aspects and phases of Internet of Things.

# Bibliography

1. *The Internet of Things: A survey.* **Atzori, Luigi, Iera, Antonio and Morabito, Giacomo.** 2010, Computer Networks.

2. *Internet of Things (IoT): A vision, architectural elements, and future directions.* **Gubbi, Jayavardhana , et al.** 2013.

3. **CISCO.** How does Cisco define the Internet of Everything, and how is it different from the "Internet of Things"? *The Internet of Everything Global Public Sector Economic Analysis.* [Online] 2013. http://internetofeverything.cisco.com/sites/default/files/docs/en/ioe_value_at_stake_public_sector%20_analysis_faq_121913final.pdf.

4. **Gartner.** Gartner says 4.9 Billion Connected "Things" Will Be in Use in 2015. [Online] Gartner, 11 November 2014. www.gartner.com/newsroom/id/2905717.

5. *A Complex View of Industry 4.0.* **Vasja Roblek, Maja Meško, and Alojz Krapež.** s.l. : SAGE, 2016.

6. **thyssenkrupp.** MAX The game-changing predictive maintenance for elevators. *Elevator Technology Max.* [Online] https://max.thyssenkrupp-elevator.com/assets/pdf/TK-Elevator-MAX-Brochure_EN.pdf.

7. —. El Segundo, California. *Grand Avenue Courtyard.* [Online] 2012. https://www.thyssenkruppelevator.com/downloads2/Grand%20Avenue%20Courtyard%20Case%20Study_One-Page%20Version_Ver3A_071212.pdf.

8. **Evans, Dave.** How the Next Evolution of the Internet. *The Internet of Things.* [Online] 2011. http://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf.

9. *Data Mining for the Internet of Things: Literature Review and Challenges.* **Chun-Wei Tsai, Chin-Feng Lai, Ming-Chao Chiang, and Laurence T. Yang.** 2015.

10. *Data Mining for Internet of Things: A Survey.* **Chun-Wei Tsai, Chin-Feng Lai, Ming-Chao Chiang, and Laurence T. Yang.** 2014.

11. *Low-Cost Adaptive Monitoring Techniques for the Internet of Things.* **Demetris Trihinas, George Pallis, Marios D. Dikaiakos.** 2016.

12. *Edge Mining the Internet of Things.* **Elena I. Gaura, James Brusey, Michael Allen, Ross Wilkins, Dan Goldsmith, and Ramona Rednic.** 2013.

13. *Towards the Implementation of IoT for Environmental Condition Monitoring in Homes.* **Sean Dieter Tebje Kelly, Nagender Kumar Suryadevara, and Subhas Chandra Mukhopadhyay.** 2013.
14. **Team, The ELKI.** *ELKI: Environment for Developing KDD-Applications Supported by Index-Structures.* [Online] https://elki-project.github.io/.
15. *Compressing Historical Information in Sensor Networks.* **Antonios Deligiannakis, Yannis Kotidis, Nick Roussopoulos.**
16. *More is Less: Signal processing and the data deluge.* **R.G.Baraniuk.** 2011.
17. *The Cloud is Not Enough: Saving IoT from the Cloud.* **Ben Zhang, Nitesh Mor, John Kolb, Douglas S. Chan, Nikhil Goyal, Ken Lutz, Eric Allman, John Wawrzynek, Edward Lee, John Kubiatowicz.** s.l. : University of California, Berkeley.
18. *RainMon: An Integrated Approach to Mining Bursty Timeseries Monitoring Data.* **Ilari Shafer, Kai Ren, Vishnu Naresh Boddeti, Yoshihisa Abe, Gregory R. Ganger, Christos Faloutsos.**
19. **Tim Lindholm, Frank Yellin, Gilad Bracha, and Alex Buckley.** The Java Virtual Machine Specification, Java SE 8 Edition. *Java Language and Virtual Machine Specifications.* [Online] 2015. http://docs.oracle.com/javase/specs/jvms/se8/jvms8.pdf.
20. **James Gosling, Bill Joy, Guy Steele, Gilad Bracha, and Alex Buckley.** The Java Language Specification, Java SE 8 Edition. *Java Language and Virtual Machine Specifications.* [Online] http://docs.oracle.com/javase/specs/jls/se8/jls8.pdf.
21. **Tom Mitchell, McGraw Hill.** *Machine Learning.* 1997.
22. *Energy-Harvesting Wearables for Activity-Aware Services* . **Sara Khalifa, Mahbub Hassan, and Aruna Senerivatne.** 2015.
23. **KUNZ, FIFA - MATTHIAS.** fifa.com. *fifa.com.* [Online] 2007. https://www.fifa.com/mm/document/fifafacts/bcoffsurv/emaga_9384_10704.pdf.