Barcelona School of Informatics

Bachelor Degree Thesis

Author:
Miquel Sabaté Solà

Director:
Jordi Garcia
Almiñana

April 23, 2014

Contents

1 Introduction

1.1 Context

This is the final report of the GEP course. The GEP course is part of my Bachelor Degree Thesis, and consists of a set of deliveries and instructions that has helped me in the preparation of my Thesis.

My Bachelor Degree Thesis is being presented at the Barcelona School of Informatics and is directed by Mr. Jordi García Almiñana. The project was originally envisioned by the director, and the author must develop a solution for the given problem.

1.2 Brief project description

My Bachelor's Degree Thesis is about building an infrastructure that is capable of providing a set of services from the data that has been collected and processed. This data will come from initiatives like OpenData BCN and iCity. This infrastructure will fetch and process all this data in realtime, by using a set of new technologies that allow us to do so.

1.3 Brief state of the art

Since the appearence of the MapReduce algorithm, a lot of different technologies have evolved so we can now fetch and process huge amounts of data. This is important for the computer industry, but also for governments and other associations.

We have a huge list of technologies that deal with Big data, but maybe the more important ones are Hadoop and Storm. Storm is the base technology that I will be using in this project. It's a modern and mature framework that will allow me to build a realtime pipeline that is capable of fetching and processing huge amounts of data.

1.4 Purpose

The purpose of this project is to provide a base infrastructure capable of fetching and processing large amounts of data from iniciatives like OpenData BCN and iCity to provide a set of services improving the actual situation.

2 Project scope

2.1 The context: Big data and streaming

Once upon a time, Google released a paper called: "MapReduce: Simplified Data Processing on Large Clusters"[2]. This was the beginning of a revolution in the data processing front. After that, lots of efforts have been put in this direction, being Hadoop one of the most important accomplishments. Hadoop (and related technologies) has made it possible to store and process data at scales previously unthinkable.

However, these technologies have one major drawback: they are not realtime systems, nor are they meant to be. In order to solve this issue one has to manually implement a network of queues and workers. These workers would eventually send messages, update databases, and send new messages to other queues for further processing. This, of course, has some serious limitations: it's tedious, it doesn't scale, and it has little fault-tolerance.

In order to fix the previously cited issues from the core, some projects like Yahoo! S4 and Twitter Storm emerged. Both projects have some differences but they have a clear focus: to ease the writing of parallel realtime computation. Sadly, nowadays the S4 project is dead. This is why in this thesis I'm picking Storm[4] as the base technology (plus, the Storm project has a very active community).

2.2 My proposal

My Bachelor Degree Thesis is about building an infrastructure that is capable of providing a set of services from the data that has been collected and processed. The raw data for this infrastructure will come from OpenData BCN[10] iniciative. That is, the infrastructure that I'm going to build for my thesis will provide additional features to the OpenData BCN iniciative that are only possible by building a cluster that will process data continiously.

This infrastructure will fetch and process all this data in realtime, using Storm as the base technology in the software front. This is important because the users of all the services build upon these infrastructure will be able to fetch all the processed data also in realtime.

This means that I am not just going to do research only in the Big data and streaming front, but I'm also going to research more deeply in the whole concept of "Smart City". This is because this infrastructure will be built in a way so other services can fetch our processed data to improve, promote and have a better understanding of the city of Barcelona.

I am very motivated for this project for the following reasons:

- **Research**. Because of this thesis, I will be doing a lot of research. My main focus will be: Big data, streaming processing and smart cities. I will be investigating, for each of them, their state of the art, what should we know about and what can we do with them.
- **Learn** some technologies as a side effect. My base technology will be the Storm framework. This means that I will be learning and improving my skills in programming languages like Java and Scala. I will also be improving my skills with concurrent programming. I will further also my knowledge on hardware by analyzing what kind of machines we need in order to provide and scale all the services.
- **Provide** a useful technology base. The main goal of this Thesis is to create a base technology that will be useful to other people. This way we will be able to provide a set of services that will potentially improve the city of Barcelona on multiple ways. I am really optimistic about this project, and I can see other thesis from other students about improving or extending the infrastructure that I'll be building in my thesis.
- **Practice**. The Bachelor Degree Thesis is the final stop of this journey that I've been doing in the Barcelona School of Informatics. I've learnt a lot and I am ready to apply some of the concepts learnt in this university to this Thesis.

2.3 Goals

In the previous section I have clarified my proposal and my motivations. Now it is time to specify the specific goals of this project:

- Design a cluster that can match our requirements.
- Design a base system that will fetch and process all the data.
- Provide a couple of useful services on top of the base system.

2.4 Limitations and risks

There are more things to evaluate before starting the project. In this section I will describe the limitations and the risks of this project.

I will start by specifying the limitations. This project does not intend to cover as much as possible in regards to the technologies that I will be using. This is not a project about Storm or any other technology that I will be using. Instead, this is a project focused on solving a specific problem.

Another limitation comes with the hardware. First of all, the sensors. An ideal situation would be that I design the sensors and I establish a way for the

sensors to communicate with the cluster. However, this is unlikely to happen, so that I am relying in an already existing solution such as OpenData BCN. Another limitation is the cluster itself. Ideally I would design a base cluster to handle all the software, but it is extremely unlikely that this cluster gets built anytime soon. For this reason, I will do my experiments in my local machine and, hopefully, also in the clusters provided by the Computer Architecture Department.

There are some risks to evaluate too. First of all, I want to talk about time related problems. If I do not have enough time to build the software, I intend to cut on services. That is, I won't be building more services. The core software is a must, though, so there won't be any time related issues here.

Another possible risk comes with the possibility that maybe the OpenData BCN and the iCity initiatives don't meet my expectations. In this case I will be developing an "alternate" solution, by specifying how ideally the sensors or the interface to the sensor would be sending requests to my cluster.

2.5 About the development

Lastly, I want to clarify how I want to develop this project. First of all, the code that will make all this to happen is open source. This means that all of the base technology and the code that I will be developing is all open source. In fact, all my code is hosted on Github[1]. This is really important for me for the following reasons:

- I am a strong advocate of Open Source.
- It could help future thesis based on this infrastructure.
- It could help anyone in the world that is developing something similar.
- It is fair.

My workflow will consist in discussing any doubts that I would have with my teacher and keep on developing the whole platform in an open way. Finally, I want to point out my resources, that I quite humble. I will spend most of my time doing this project, with all the resources that I have.

[1] https://github.com/mssola/thesis

3 Planning

3.1 Project planning and feasability

The first period of the project is dedicated to studying the feasability of the project and to do the initial planning of the project. This is done through a course called GEP, that is currently running. This course includes the following stages:

- Scope of the project.
- Project planning.
- Budget and sustainability.
- Preliminary presentation.
- Bibliography.
- List of conditions.
- Oral presentation and delivery of the final document.

3.2 Project analysis and design

The main goal of this stage is to draw a clear picture of the project and analyze all the goals of the project. After doing this, I will move my efforts towards the design of the application.

Therefore, this stage is made out of two sub stages. The first one, the analysis of the project. In this sub stage I will be defining all the goals,lose requirements, features and use cases of my application.

The other sub stage consists in design the application. This will be done by creating diagrams, drawing flow charts, etc.

3.3 Project iterations

1. **Development of the core software infrastructure**

   In this iteration I am going to focus on the core infrastructure that has to hold the whole application. This will be executed only in the software front. I will consider that this stage has ended when there is production-ready base software that can hold the infrastructure and that is well documented and thoroughly tested.

2. **Providing services**

   The next iteration consists of building services on top of the base infrastructure that has been created in the previous stage. Similarly to the

previous stage, all the services build at this point will be documented and tested consciously.

3. **Designing the cluster**

   At this point, we have the software ready to be put in production. Now we need to design and implement a cluster that can hold the software. In this stage I will be using lots of concepts learnt at the university, specially from the CPD course.

4. **Concluding the development**

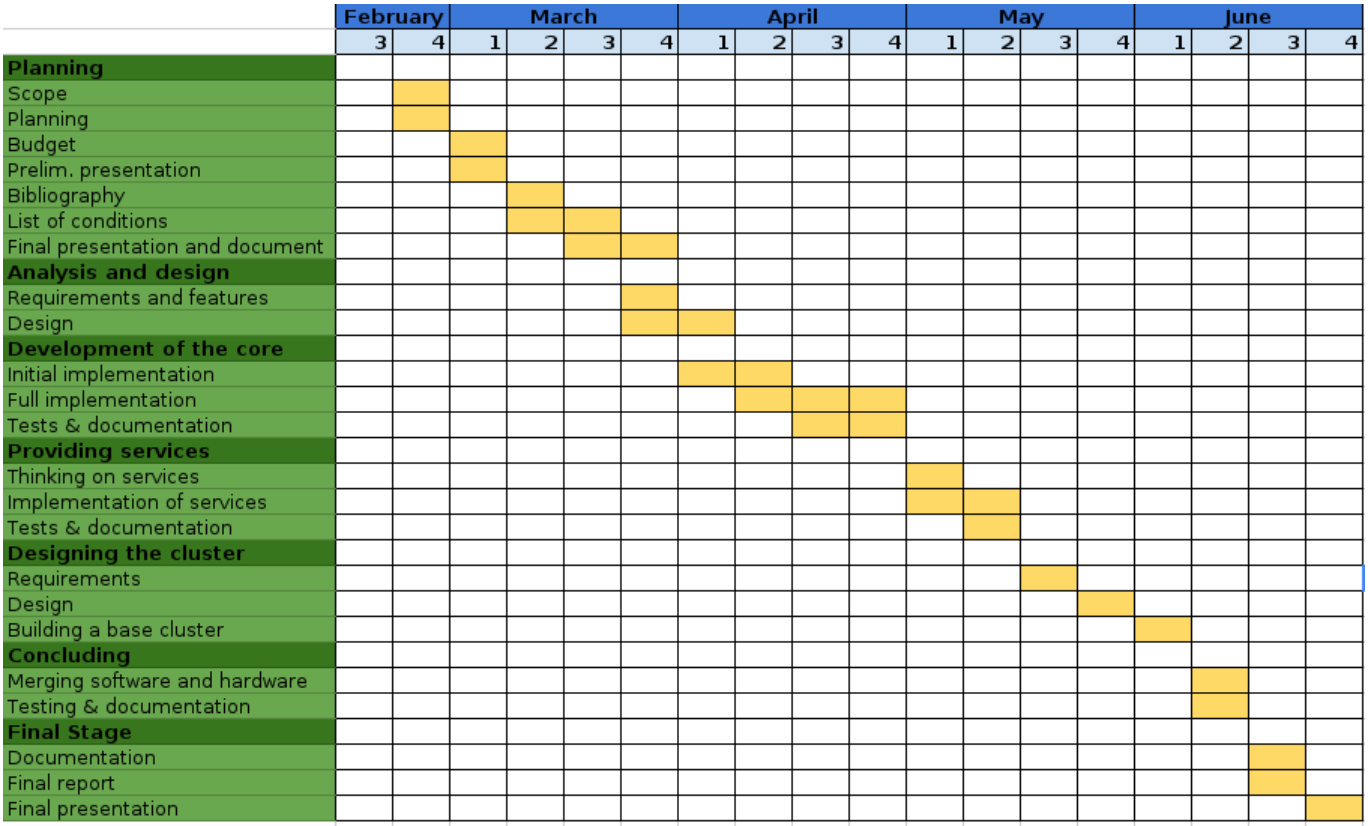
   In the last iteration I will be concluding the development of this project. It consists of doing final tests on both the software and the hardware. This last iteration will be used to make sure that everything runs smoothly.

3.4 Final stage

The final stage consists on closing the project. Since all the code and the cluster design has already been set up to this point, I am just going to focus on the following topics:

- Documentation.
- Final report.
- Final presentation.

Gantt chart



I don't like to compute the work to be done in hours. I prefer to do it in weeks. In the Gantt chart we can see the scheduling that I've made for the project divided by weeks. Some tasks are overlapping in some weeks. This is quite normal.

The total amount of hours is not fixed, it depends on the work that is needed to be done for each task. The important thing here is that we make the deadlines with no problems.

3.5 Action plan

At this point, it is clear how the stages are scheduled. Now, what happens if any stage has a different duration than expected ? In my case, I will just start the next stage. That is, if I will start each stage as soon as possible.

In the end of each stage I will do an evaluation of the work done so far. This will allow me to rearrange the schedule so I don't lose the track of the development of the project.

Moreover, at the end of each iteration I will meet with my director to analyze how is the project going. This will force me to work hard to meet the deadlines and it will also give me the oportunity to have some feedback from the director.

4 Budget

4.1 Introduction

In this delivery I am focusing on the budget and on the impact of my project. First of all, I am going to focus on the budget. This includes doing some math in the following topics: human resources, hardware and software.

Lastly, I will be explaining the impact that I expect that this project will have. My only considerations for the impact will be the social impact and the environmental impact.

4.2 Budget

In this section I will be describing the estimated costs that this project will have. After specifying each of the costs, I will make up a total amount of estimated costs. This will lead me to conclude the needed budget for this project.

4.3 Human resources

What I want to do in this section is to describe all the costs related to employing people to develop this project.

Luckily enough, it is just me in this project, so I just have to compute my own salary. After doing some research, I have found that the average salary for a Big data analyst is around 99,000 USD a year. This boils down to 51 USD per hour. I did my scheduling based on weeks, instead of hours, and I think that I will spend 20 hours in average per week. This means that in total, my salary during the project will be:

$$20\ hours \cdot 16\ weeks \cdot 51\ USD = 16,320\ USD$$

4.4 Hardware

Ideally in this section I would explain all the costs regarding the hardware. However, I can't compute this because the only hardware that I will specify will be the cluster which, in turn, is just a proposal. Other things like my laptop won't be counted in this section They are not a cost for this project since I use them for personal purposes too.

4.5 Software

In this section I will describe all the costs associated with software. In short, this is my budget for software:

0 $

All the software that I will be using for this project is free: both as in "free beer" as in "free speech". There are no fees for any license, I won't be using any paid service, etc. Nothing.

4.6 Total

The main conclusion is that the only cost associated to this project is just my salary. Therefore, my estimated costs are a total of $16,320.

4.7 Social impact

This section is really important for this project. It is about the impact that this platform will make to society. This is not a direct impact, but an indirect one. The social impact of this platform comes in two ways:

- The fact that a number of businesses will take advantage that this platform exists.
- Indirectly, all the impact that all the services built on top of this platform will produce.

This means that, ideally, the social impact of this platform can be huge.

4.8 Environmental impact

In the same way that this platform can bring a lot of goodness in the social front, it certaintly comes with a cost. In my case the cost is an environmental impact that cannot be understated.

In this project I will propose an ideal cluster that would be able to run the software that I will design. The problem here is that maintaining a cluster means the following:

- Power supply.
- Maintaining a cooling system.
- The implied environmental costs of building the cluster.

All of these can be reduced by using the minimum amount of cluster time as possible. This means to run the software in "batches" or with a very low latency. However, this is not possible at all if the cluster has a lot of requests, and that is to be expected.

Therefore, in the development of the cluster I will focus most of my efforts into keeping the cluster as environmental friendly as possible.

5 Bibliography

5.1 Context

This section aims to describe the context within which the project will be developed. We differntiate between: developers, the tutor, library developers and the users.

5.2 Developers

I am the only developer of this project. I will make use of my knowledge on the topic, the guidance of my tutor and the previous work of the library developers to accomplish my goals.

5.3 Tutor

Jordi García Almiñana is my tutor. His role is to make sure that I have both my focus and my goals straight. He will be guiding me through the development of this project and giving me some tips of advice.

5.4 Library developers

Library developers are all the developers that have contributed to the libraries that I will be using to build this project. There are a lot of developers, coming from different backgrounds and goals.

5.5 Users

The users are everyone that will make use of this project. We can distinguish different users:

- End users that will make use of the data generated from our platform.
- Developers that will make use of our services.
- Developers that will read the code of this project, so they can get a benefit from it.

5.6 State of the art

In this section I will be describing the state of the art. That is, I will describe what is the current situation, the problems arising, and solutions from other developers.

5.7 Big data

Big data[1] is a term that has been coined lately. Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers". The definition of Big data varies depending on the organization.

5.8 Map Reduce

Since we cannot perform operations on Big data on a traditional way, we have to build tools that will tackle this problem accordingly. The first step was made by Google with its "MapReduce" algorithm[2]. The MapReduce algorithm solves this problem by mapping a specified problem (e.g. a file with data inside of it) and finally reducing the results back. This is an easy way to distribute the work on different machines, so they are as efficient as possible.

Multiple implementations of MapReduce arised after the announcement of MapReduce. The most famous implementation being Hadoop[3]. Hadoop implements the algorithm of MapReduce with fault tolerance. It also provides the idea of HDFS (Hadoop File System), a distributed file system.

5.9 Streaming

At this point, it is clear that Hadoop and the MapReduce algorithm brings a lot of advantages and opportunities. This is nice but Hadoop does not do one thing: realtime operations. Hadoop and similar technologies are not realtime systems, nor are they meant to be. Hadoop is not designed to be a realtime system, and thus any effort to hack Hadoop in any realtime direction is a waste of time. This is because realtime data processing has a fundamentally different set of requirements than batch processing.

This is really a problem, because realtime data processing at massive scale is becoming more and more of a requirement these days for businesses. To fill this hole, technologies such as Yahoo! S4 and Twitter Storm[4] were developed. Right now Yahoo! S4 is more or less abandoned and Twitter Storm is really alive, but under the umbrella of the Apache foundation[5] This is why my project will be based on Storm.

Before Storm, you would typically have to manually build a network of queues and workers to do realtime processing. Workers would process messages off a queue, update databases, and send new messages to other queues for further processing. Unfortunately, this approach has serious limitations:

- **Tedious**: You spend most of your development time configuring where to send messages, deploying workers, and deploying intermediate queues. The realtime processing logic that you care about corresponds to a relatively small percentage of your codebase.
- **Painful** to scale: When the message throughput get too high for a single worker or queue, you need to partition how the data is spread around. You need to reconfigure the other workers to know the new locations to send messages. This introduces moving parts and new pieces that can fail.

Apart from Storm, I will make use of other technologies such as Summingbird[6] and other technologies from Twitter[8] and Yahoo[9]. These two companies have open sourced a lot of libraries for Storm that will be useful in the development of this project.

5.10 Future

The future of Big data will be tied in the future to the streaming technologies. They solve increasingly important issues and they perform efficiently when dealing with lots of data.

It's clear, then, that more companies will follow this lead. This implies more support and more libraries available out there.

5.11 Smart cities

Finally, I would like to talk about the concept of Smart Cities[7]. Smart cities is a new concept that embodies all the efforts to make cities more efficient and more reliable.

That key component are IT technologies. We use data centers, sensors, etc. to compute key elements. With all the computed data, we can then analyze which parts of the cities can be improved in order to make citizens happier and services more efficient.

Bibliography

[1] Dan Kusnetzky, *W*hat is Big Data?. http://www.zdnet.com/blog/virtualization/what-is-big-data/1708", 2010.

[2] Jeffrey Dean and Sanjay Ghemawat, *M*apReduce: Simplified Data Processing on Large Clusters 2004.

[3] Hadoop. Wikipedia. http://en.wikipedia.org/wiki/Hadoop

[4] Storm. http://storm.incubator.apache.org/

[5] Apache Software Foundation. http://www.apache.org/

[6] Github page for Summingbird. https://github.com/twitter/summingbird

[7] Smart Cities. http://en.wikipedia.org/wiki/Smart_cities

[8] Github page of Twitter. https://github.com/twitter

[9] Github page of Yahoo!. https://github.com/yahoo

[10] OpenData BCN webpage. http://opendata.bcn.cat/opendata/en