Universitat Politècnica de Catalunya - BarcelonaTECH
Barcelona School of Informatics

IT4BI Master Thesis

Entrepreneurial decision support engine in forecasting business opportunities

Candidate: Burghelea, Madalina

Advisor: Manso, Andres

Official Supervisors: Eguiguren, Marcos and Romero,Oscar

Planned period: February 2014 to July 2014

DatoSphera is a platform that helps entrepreneurs validate their business ideas and investors to identify the best opportunities in the market, with the use of Big Data.

Table of Contents

1.Introduction

1.1.Business Context

The master thesis aims at creating a decision support system for new entrepreneurs, which studies the online data on web domains and could recommend to the new entrepreneur if it is a good idea to join that particular business niche.

DatoSphera is a data startup for entrepreneurs, inside the Big Data Incubator Incubio. The project followed different stages of development from being part of the Incubio Research department to being incubated after only one month and transformed into a real company.

Given that the product is following the Lean Startup Model into a real company, frequent changes and flexibility were the keys to the product development. The user feedback was directly and immediately reflected in the product. The team had also an important role in the definition of the product and its construction. Contact with entrepreneurs and investor's trough participation to dedicated events and personalized interviews was the main contribution factor to test the viability of DatoSphera. Currently the project is fully funded.

The project was based on the web domains data available within the project Trakty, in Incubio and aimed at creating a spin-off with a big part of this data. For all that, given the errors made during the data aggregation strategy in the Trakty project, the data strategy had to change completely, as well as the product idea.

1.2.Motivation

The studies organized inside Incubio show a high amount of failure among the technical entrepreneurs, due to the fact that market research is neglected, and a high importance is given to the product development. A normal consequence is that a product technically complete will not have place on the current market, becoming one of the copies of the companies that have already proved to be successful long time ago and that occupy all the market and have their clients.

This situation is not common only among technical entrepreneurs but also occurs among the others, given that the market situation is changing daily and new competitors appear day by day.

DatoSphera aims at helping in particular the technical entrepreneurs to deal with market research, even if this is to be done automatically and as a web service. This would allow an entire movement into creating original applications that are currently needed on the market and not replicas of the existing ones.

Even though the project was designed for entrepreneurs, soon it appeared a strong interest from the investors, who wanted to evaluate all the business ideas they use to receive. The investors were considering reviewing the project a time-consuming task and therefore were willing to use an application that could study the market viability of the ideas.

The main advantage of the tool is saving time with the market research and allowing real-time analysis of your competitors. The data is updated daily and therefore allows the fastest access on the market to this sort of data.

1.3.Objectives

The goal of the project is to build a decision support engine that allows technical entrepreneurs to see what are the main existing companies in the field and see a complete analyses of them (Security level, geographical users, trends and evolution, financial information etc.)

The module will check how many competitors are in the market for that particular business will see how well or bad they perform and advise the new entrepreneur if there is a good idea to open a business in this sector.

More concrete, different tasks need to be accomplished and for them, different goals have been set:

- Refine the business plan *[Business bibliography- 2]*
- Define the business algorithm for recommendation *[Business bibliography- 3]*
- Get public funding with the business plan, for the project
- Conduct complex market research *[Business bibliography- 4]*
- Design a marketing and promotional strategy
- Establish partnerships with market researchers
- Ensure scalability of the business model *[Business bibliography- 1]*
- Consider extension and working in virtual teams, by involving specialists worldwide
- Search for private investors

Business Model goals

- Prepare data in the project Trakty for the purpose of the project
- Conduct data cleaning and preparation *[Technical bibliography- 1 and 3]*
- Decide on the database to be used *[Technical bibliography- 1 and 2]*
- Plan and supervise the construction of the front and back end
- Plan strategies for user experience
- Define strategies for optimisation and data management *[Technical bibliography- 4]*
- Design an extension plan for historical data and other modules *[Technical bibliography- 5 and 6]*
- Introduce historical evidence

Technical goals

- Recruit and form a team suitable for the format of the project
- Implement AGILE methodologies in the development of the project
- Evaluate the amount of work and distribute the tasks according to the person's profile and the time constraints *[Project Management bibliography- 1]*
- Make a growth plan for the team *[Project Management bibliography- 2]*
- Document all the findings and results *[Project Management bibliography- 3]*
- Analyse the results, with the metodologies

Project Management goals

The expected result would be a decision support engine that aggregates the web domains characteristics, based on the keywords that define them and then display to the new entrepreneur what is the situation in the specific niche.

1.4.Initial planning

The initial plan was to reuse the data from the Trakty project, to create a recommendation system. For all that various errors in defining the data strategy, made this data difficult to be used for the purpose of accuracy.

Initially the approach was based on reusing the data sources that have been already extracted in Incubio.

The data available from Incubio in the Trakty project was aggregated by domain and includes various web sources, such as : **Alexa** (Geographical distribution of users of a domain, Traffic data - bounce rate, time on site, daily page views) Top related links (information about possible competitors extracted from their webpages), **Home** (technologies used in a website, that could be detected from the front code - for instance HTML, PHP, Apache etc.), **Phishtank** (a database of phishing sites), **sharedCount** (social media popularity given by LinkedIn likes, Facebook likes and shares and also Twitter followers), **SimilarSites** (information about possible competitors from AdWords), **Topsy** (all the tweets that have been associated with the name of a domain), **WOT**(valuable source of reputation scores such as confidentiality and reputation of vendor or privacy or even child safety on the web domain), **Semrush** (stores the most important organic keywords associated with a domain and their position),**YahooBoss**(that stores all the information about the links associated with the keywords in SemRush) and **Seomoz** (that extract for an URL all the possible versions in which an URL appears). All these data sources have been aggregated by domain. As a starting point the Alexa domains were considered and then enriched with information from the other sources.

This data from the Trakty project proved to be not sufficient accurate for the purpose of recommendations and therefore a new strategy was defined, involving less data sources but more carefully selected and merged.
The new approach uses less data sources, known to have a better data quality, and it maximizes the potential that these data sources have.
The impact on the user was considered to be more significant, in case of using less data sources, but more accurate than merging diverse data sources, but with very low scores of accuracy and low coverage between them.

The project followed different stages, from a research project to a real company, fully funded with public support.
The project is part of a business incubator, which means that the development of the business model has to fit into the requirements of the incubator. Some of the roles inside a start-up will be realized with the support of the incubator, as shown below:

Master thesis contributions

Business plan construction

- A draft is developed as output of the VBP course
- Master thesis contribution:
  - Enrich the plan with practical experience
  - Add complexity and plan details

Business assistance from Incubio

- There are specialists in marketing, finance ,legal, systems and IT, that could refine the business plan
- Master thesis contribution:
  - Synthetize the feedback received
  - Adapt the business plan to the suggestions received

Register the business

- Incubio provides legal support but a significant part of registering the company belongs to the entrepreneur
- Master thesis contribution:
  - Register the business legally
  - Register the contracts of the people hired

Ownership

- As the business is part of an incubator, the ownership will be decided in agreement with Incubio
- Master thesis contribution:
  - Analyse different legal structures of ownership
  - Determine which form of ownership suits the best the business model, the entrepreneur and the incubator

Public funding and office location

- The files to apply for public funds will be realised in the first week of February
- Master thesis contribution:
  - prepare the files for public funds
  - provide all the necessary documentation required

Recruiting

- Incubio provides support in recruitment but for all that, a big part of the month of February will be dedicated to recruit the right freelancers/ developers.
- Master thesis contribution:
  - conduct the entire process
  - decide the funding allocated to personnel expenses

Technical implementation

- Personal Contribution
  - Data strategy
  - Database operations
  - Semantics User experience + Programmers management

Establish the brand and advertise it

- After the product was developed, support is required. With the help of the Marketing department in Incubio we could develop a marketing and sales strategy
- Master thesis contribution
  - Actively participate in promotion
  - Online advertising- prepare strategies

The project is conceptually developed within the Service Oriented Business Intelligence Course and also in *the Viability of the Business Projects*. The amount of work realized was done on a small batch of data, on only 4000 domains, in order to define a methodology to do it on the terabytes of data available. A business plan is developed during the Viability of Business Projects course.

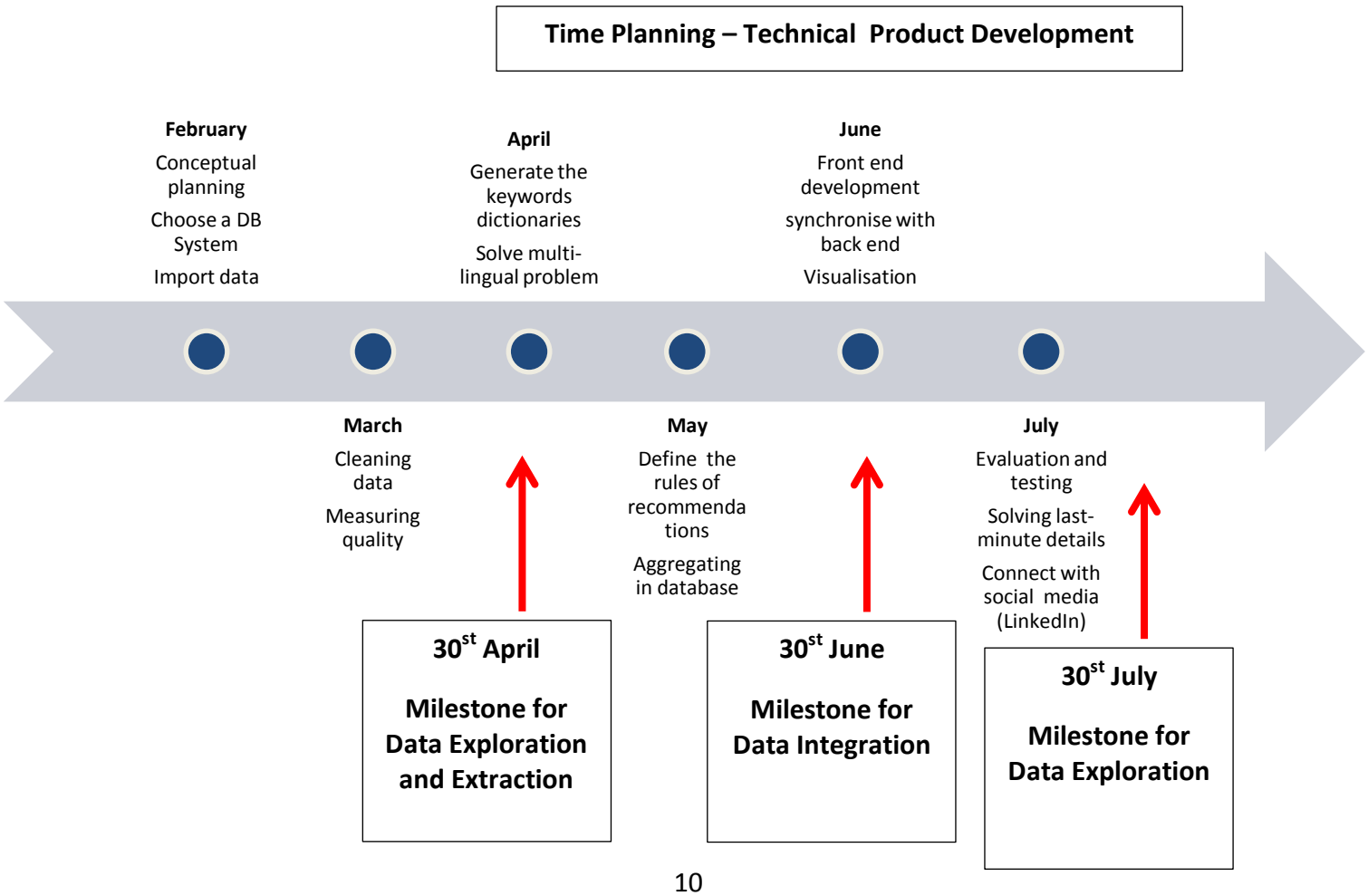
The master thesis was planning to replicate the entire process on a much bigger scale and also to increase the complexity of the rules for recommendation. This was not achieved because while evaluating and testing the possibilities of building a large scale system with these data, the accuracy of the data was considered too low, as well as the scores of missing data.

Instead of continuing to work with these data, a new strategy was defined and the data was extracted all over again, but much more carefully measured. The result in this case was significantly better and leading to a complete product.

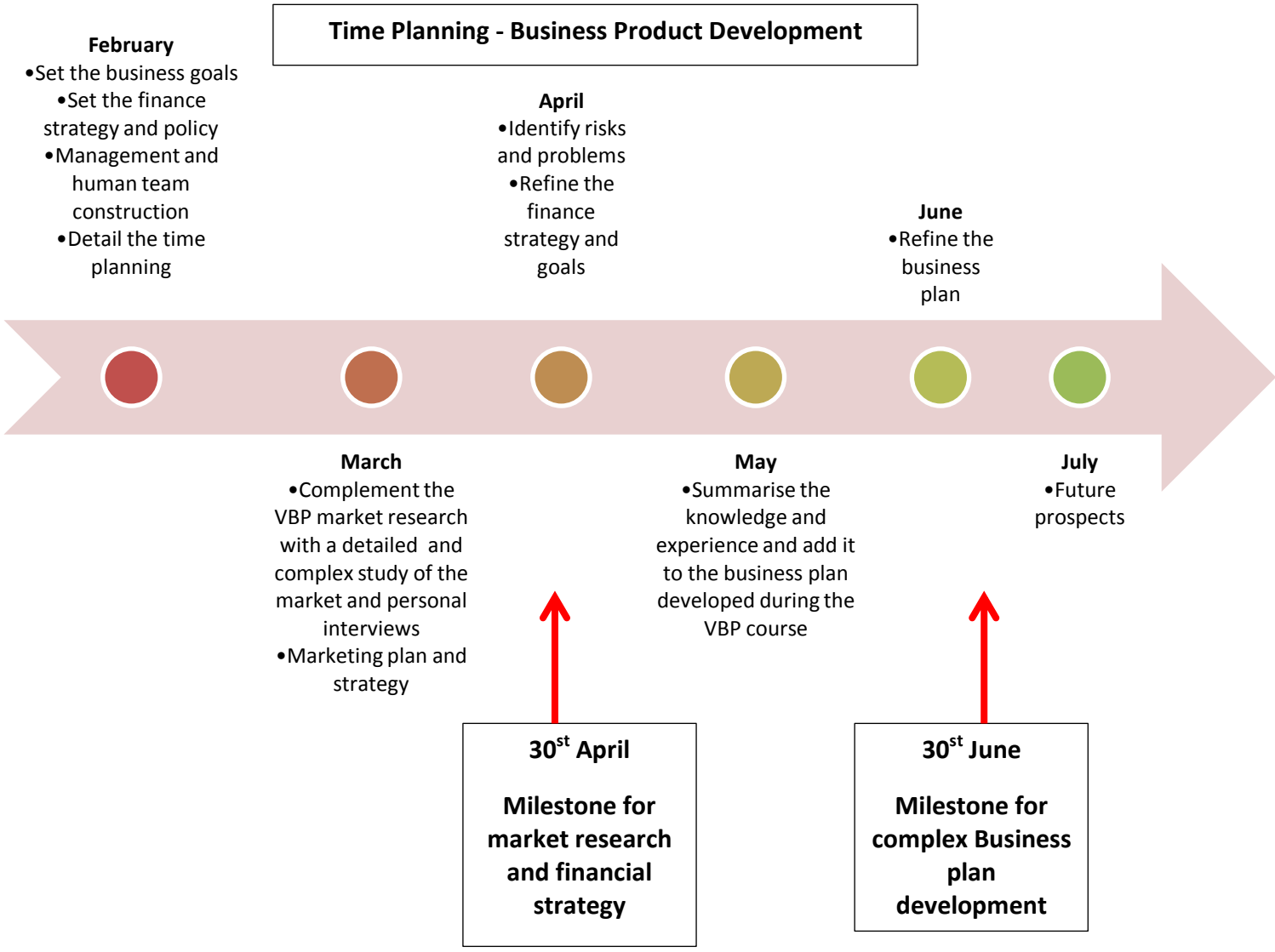
The business plan will be developed and enriched with practical experience. In the same time the project aims at creating a company out of the final product.

The methodological plan has been developed during the SOBI project, by testing on a small batch of data of 4000 domains.

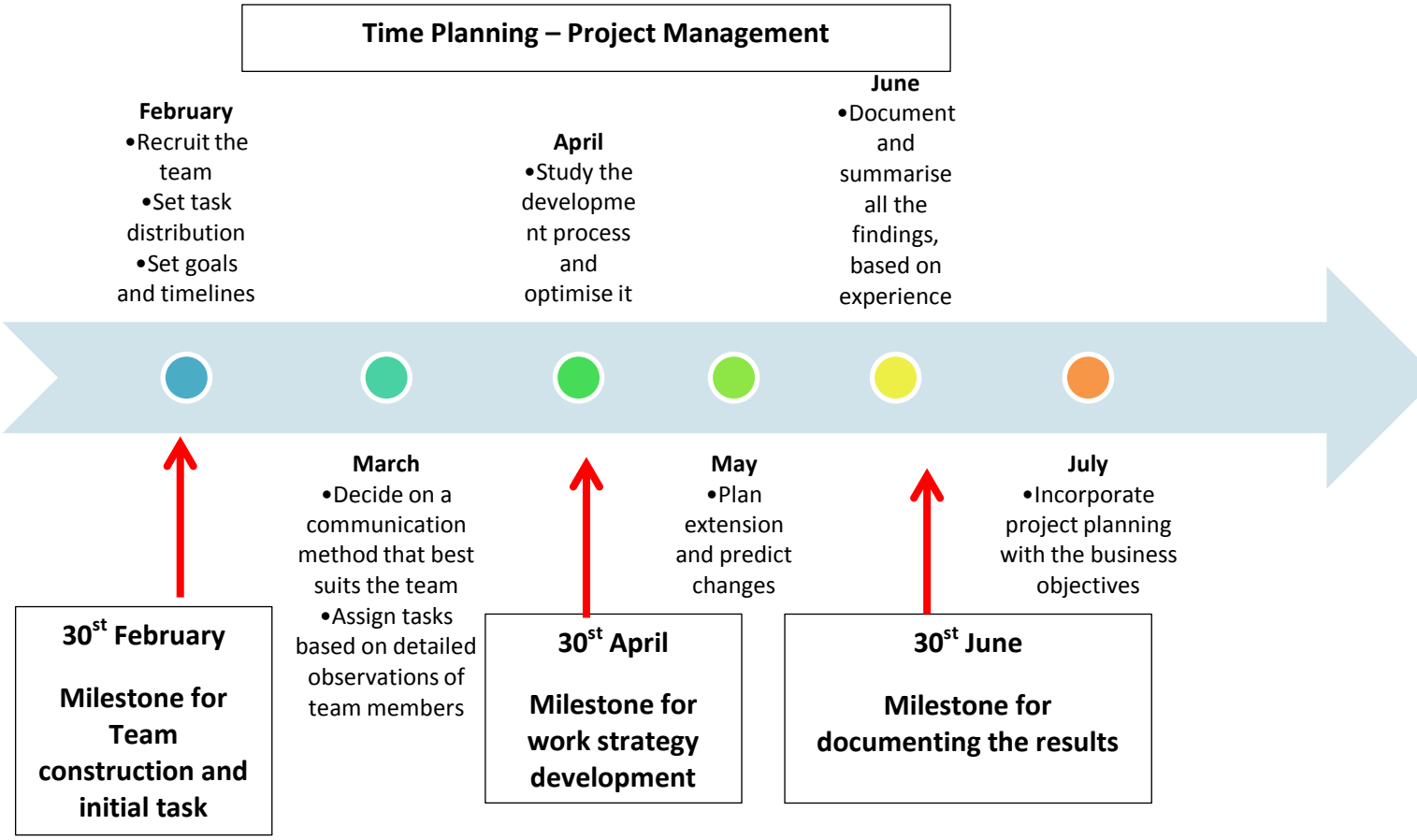
The time planning for the technical part is the following:



The time planning for the business product development is presented below:



Given that the project has also a managerial component, in terms of project management, the time structure to be followed is the one below:



1.5.Structure of the document

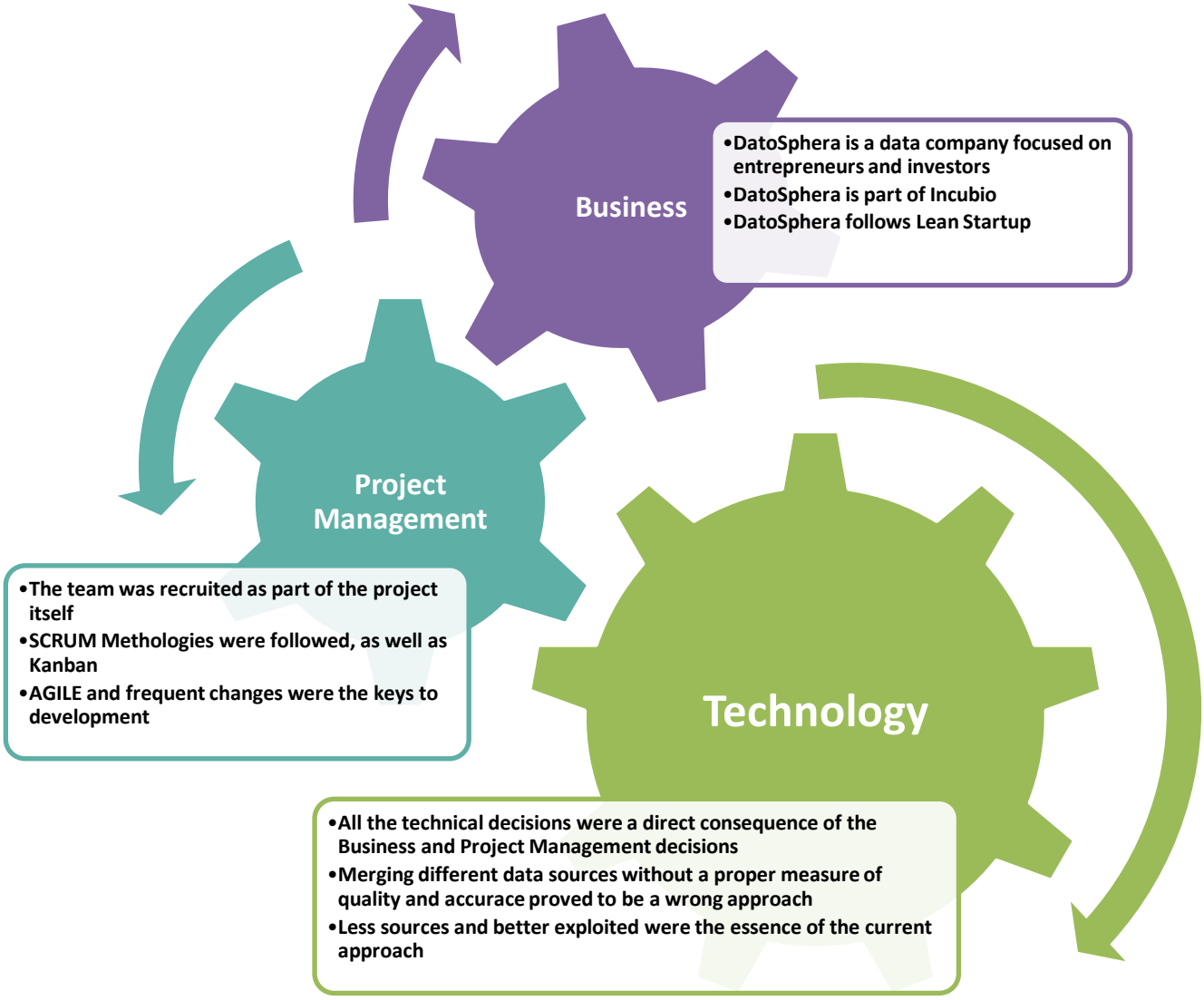
The structure of the document involves diverse aspects from business to project management and technical.

Therefore each of the aspect has been developed in a dedicated chapter, for the purpose of understanding each of the different types of tasks that have been conducted.

Is important to understand that as a real project, these tasks cannot be separated and therefore one technical decision would always have a business motivation or a project management or team foundation.

For these particular dependencies, the business part will be treated before, to motivate the context of the company and its strategy, followed by the project management part, where all the team constraints were explained and finally the technical part that is derived from the business and project management part.

It is mandatory to understand the business context and the project management constraints, to fully understand why some technical decisions were taken. As a real company, these constrains have direct impact on technical development.



Given the diversity of topics to be discussed, each of these chapters will be divided into chapters personalized for the type of content.

It is important to emphasize that the technological decisions are a direct consequence of the business and project management constraints and therefore a full understanding of the context motivates the technical decisions.

AGILE and Lean Startup methodologies were followed all over the process and therefore frequent changes were a constant of the project, until its validation with user input.

2.Product Development and Business Decisions

The product development is directly related to the technical development and the decisions are motivated in correlation with the data coverage statistics. Diverse methodologies are used to conduct market research and to test the viability of the project.

2.1.Motivation

Most of the technical entrepreneurs are likely to fail, because they never check the market viability of their projects. Very few of them have support or can afford ping a market researcher to conduct market research, prior to application development.

This is how the technology market becomes full of replicas of exactly the same application, instead of coming with original ideas.

Information on competitor, market evolution, funding and acquisition can be very valuable to any entrepreneur at the beginning.

2.2.Market research and competitive advantage

2.2.1. Interviews

2.2.1.1. Entrepreneurs

We have conducted face to face interviews with the entrepreneurs from Incubio, in the first phase, followed by interviews with technical students and graduates, willing to start a company. A total number of 30 people have been interviewed, 5 of them running their own company in Incubio.

The questions that the participants have been asked are the following:

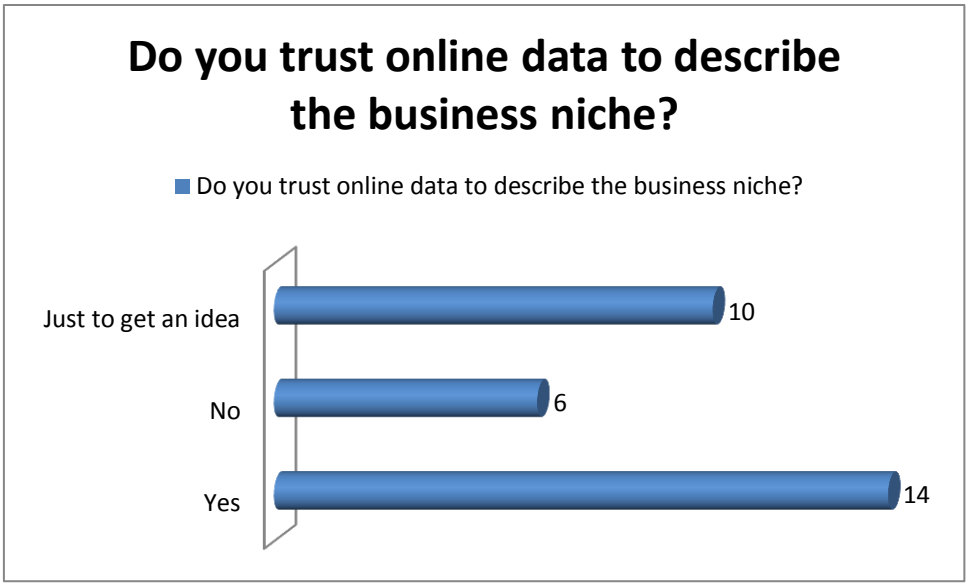
Quantitative (based on closed questions)

- Would you trust online data to describe the business niche?

Qualitative (based on open questions)

- How have you/How would you conduct market research?
- How would you/have you got a business idea?
- Do you have any suggestions for DatoSphera?

The results are synthetized below:



To the qualitative questions, the answers vary, but the most significant and complete ones are synthetized below:

How have you/How would you conduct market research?

- I haven't ; A friends has done it for me ; I have no idea how to do it ; Google ; With the help of a specialist ; Online companies

How would you/have you got a business idea?

- Google ; From last work place ; From my employees ; From my professors ; From family ; Online statistics ; Problems I faced ; Different websites ; Traffic data

Do you have any suggestions for DatoSphera?

- Use the funding rounds; Use tags, not keywords; Introduce fields of activity; Substitute Ad Words, by telling me on which sites I can promote my brand. Search for websites using particular tags that accept websites. Or search for online communities. ; Watch out on promotion strategy ; I don't believe that such an engine could be developed- maybe just some analytics ; I want to see figures and numbers; I would use it to find partners and see necessity in other countries

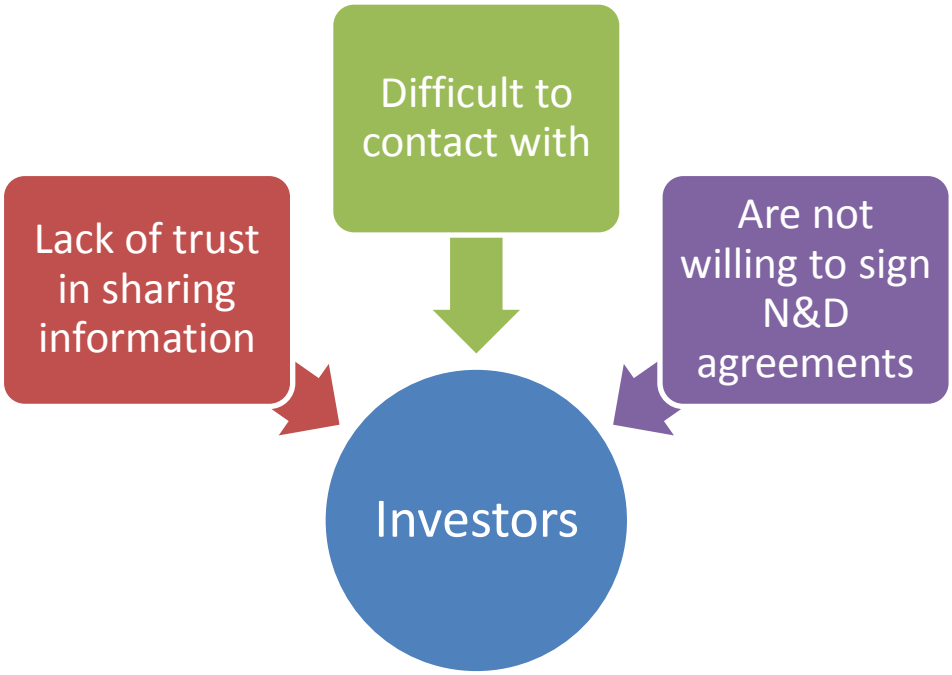
Conclusions

Entrepreneurs are a target that is willing to test the product, as they have a real necessity. One issue might be regarding the confidentiality of the business idea but this could be easily talked by introducing the industries of activity. We can therefore conclude that the interviews reveal a real necessity of the tool for startups and entrepreneurs.

2.2.1.2. Investors

Special personalized interviews were conducted with investors, even though the complexity of these interviews was much higher than with the entrepreneurs. DatoSphera interviewed 5 investors, having different profiles, from Seed funding to Venture Capital.

This category was raising significant problems due to the following aspects:



Therefore only 5 investors were interviewed and carefully selected to the meetings. The investors were coming from different backgrounds, mostly investing in technical projects and having small seed investments.

Given that most of the investors were very skeptical to share information on their process of screening and identifying the most successful projects, the research approach was different, following some questions but in a form of a conversation. Another aspect is that all the investors refused to sign an N&D agreement. Therefore the questions were asked in more general lines.

- How you do evaluate the business ideas you receive?
- Do you have any suggestions for DatoSphera?

The answers to these questions are synthetized below:

How you do evaluate the business ideas you receive?

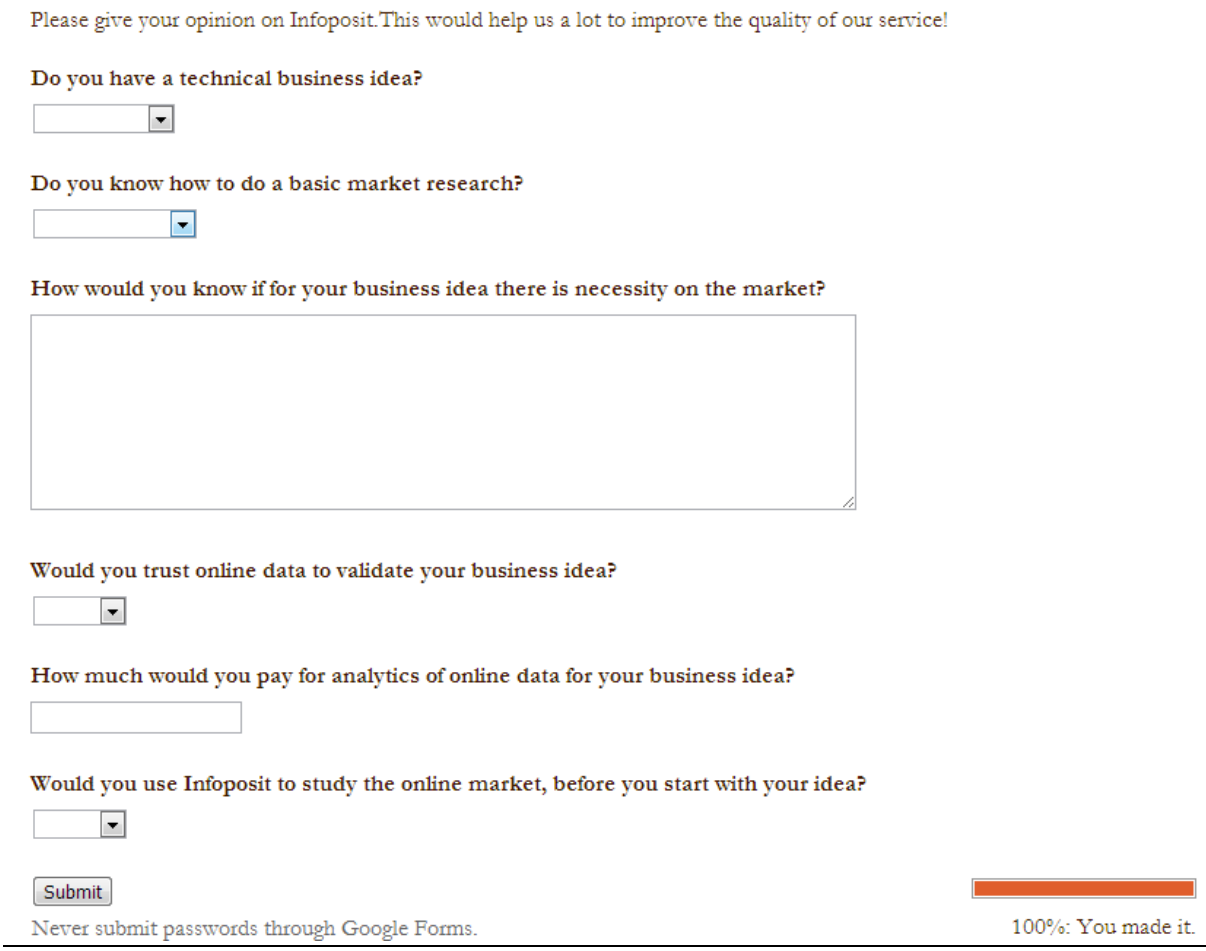
- Searching through online data sources of information
- Use the LinkedIn profile to see recommendations
- Study the trends from the USA
- Study TechCrunch
- Use advisors opinions and experts in technology

Do you have any suggestions for DatoSphera?

- Involve more team aspects
- Involve LinkedIn profiles or sources like People.com
- Offer reports to be downloaded

2.2.2. Surveys

As the interviews could be complemented with even more data, we have developed a simple survey, with few questions, that could be easily filled in by the respondents.


Please give your opinion on Infoposit.This would help us a lot to improve the quality of our service!

Do you have a technical business idea?

Do you know how to do a basic market research?

How would you know if for your business idea there is necessity on the market?

Would you trust online data to validate your business idea?

How much would you pay for analytics of online data for your business idea?

Would you use Infoposit to study the online market, before you start with your idea?

Submit

Never submit passwords through Google Forms.

100%: You made it.

The form was promoted by email and also in social networks of technical students and graduates.

To ensure that the respondents will fill it in, just the most representative questions have been selected.

Also to select exactly our target, just the ones having a technical idea have been selected, as the rest of the backgrounds are not our target group.

The number of respondents was 46, but as some of the answers were not of a good quality, the total number that could be used was 40.

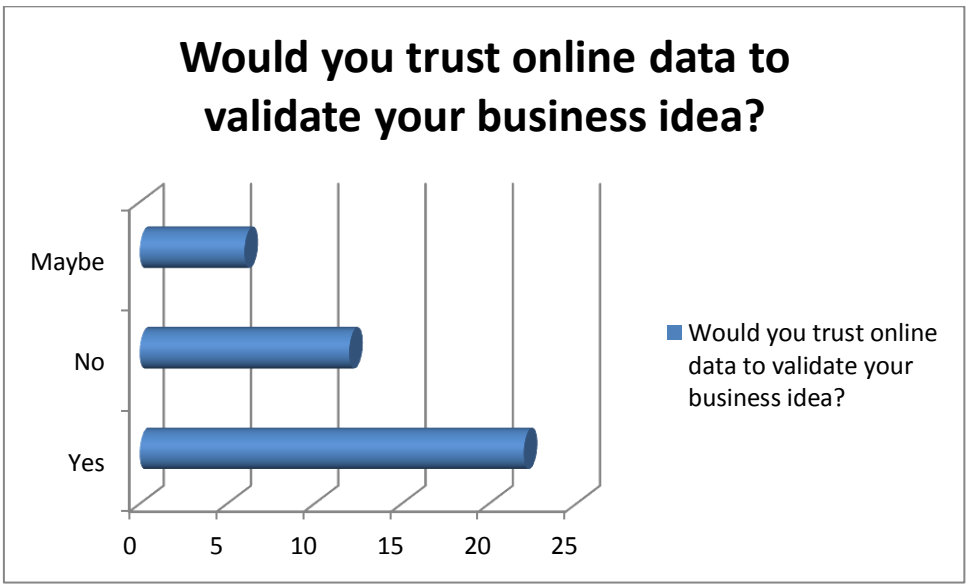
The results summarized and cleaned are shown below:

How would you know if for your business idea there is necessity on the market?

The results for these questions varied from people not answering it to complex answers:

- Study the competition
- Ask specialists
- Re-engineer the processes
- Nothing similar exists so far
- It could be realized easily and could be tested after to see the traffic data
- Asking possible customers

For all that, most of the respondents (62%) left this question empty.



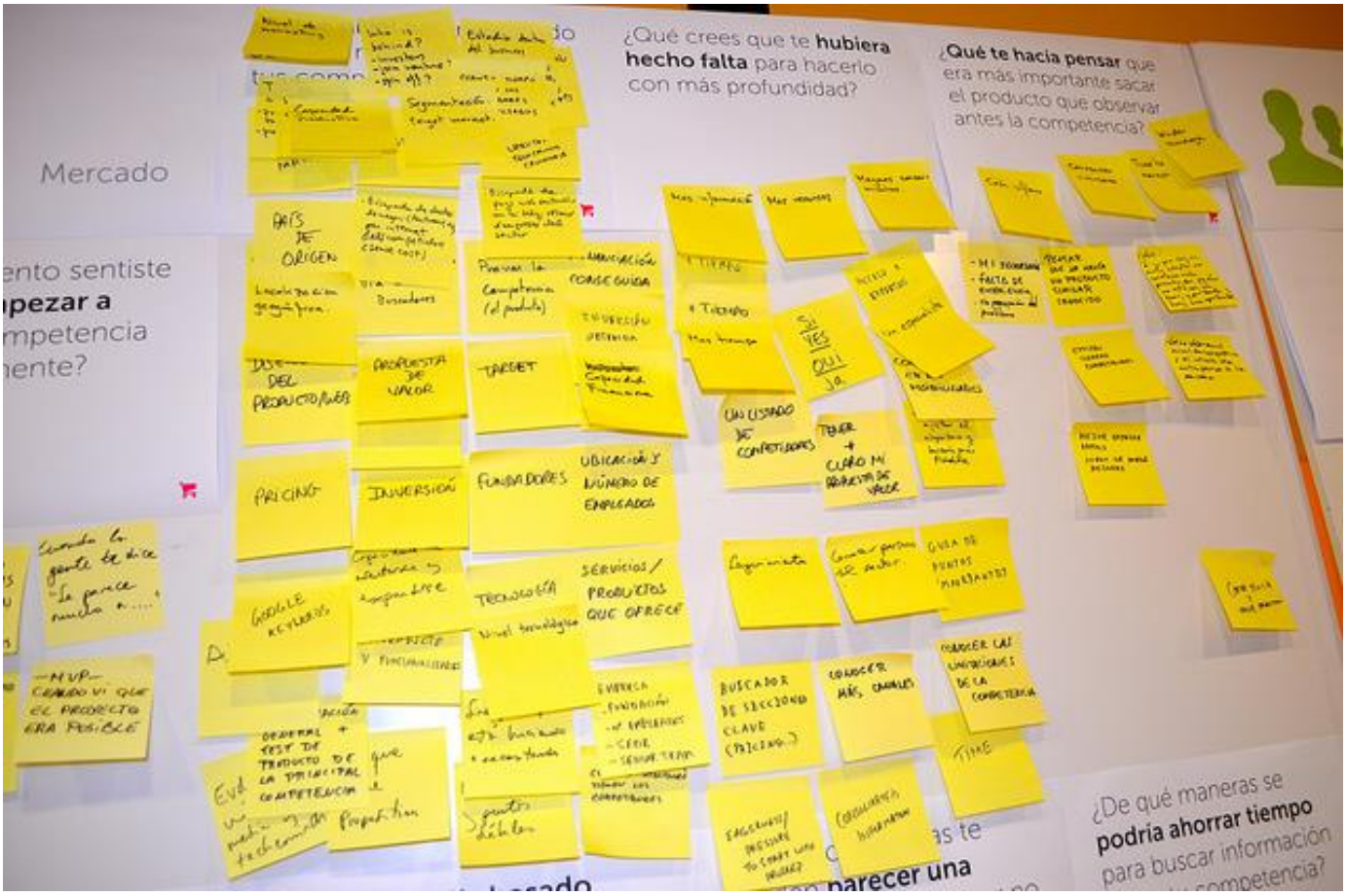
2.2.3. Co-creation Workshop: focus groups

We have gained 15 people in Incubio, to do a Co-creation workshop with our focus group. During an hour, each of them contributed with ideas in the brain storming.

The structure of the workshop was the following:

- Open questions were raised ,regarding entrepreneurship, ideas and market research
- Each participant had to write down the answers, that were shaping the application
- The answers were publically discussed
- The discussions followed the brainstorming

The results out of the co-creation workshop were different ideas for applications and necessities. These ideas will be considered in respect to the data that we have and we can extract. The output of the workshop was the following:



The discussion was moderated by Marc Segarra, expert in Co-creation workshops. The questions were selected carefully and the participants were encouraged to share more information.

A big part of the taller was dedicated to technical innovation, due to the fact that the backgrounds of the participants were ideal for getting technical ideas.

The conclusions of the workshop were:

- The participants would need the application, as it is right now
- The price will be a decision factor for using the application
- Extra modules are needed: where and how to promote, recruit and find business partners
- Technical innovation is possible and extra data sources could be added
- The recommendation of most popular keywords or technologies seems to be very popular
- They would want to have momentum data

2.2.4. Interview market researchers

In Incubio there are currently 4 market researchers and marketing specialists, which we could ask regarding the idea itself.

The questions designed for them were the following:

- Do you think that the service could provide a very first step in market research and idea validation?
- Would you be able to use the results to enrich the quality of the reports?
- Would you collaborate with DatoSphera, as a partnership?
- Do you have any suggestions?

The answers we received were the following:

Do you think that the service could provide a very first step in market research and idea validation?

- 3 of them mentioned that it would be a very fast response and therefore a very first step

- 1 of them mentioned that it could be used for something else than market research, such as promotion, recruitment and finding business partners

Would you be able to use the results to enrich the quality of the reports?

- 2 of them would use such the engine to enrich the quality of the final report and to get a general overview
- The other 2 would use it just to get an idea and to study the market, find the competitors and their possible problems

Would you collaborate with DatoSphera, as a partnership?

- 3 of them would do it, if the price is low and if there is a reasonable number of entrepreneurs reaching them
- The other one would only collaborate after some years of studying the application

Do you have any suggestions?

- Use it as a replacement of AdWords, to find websites/groups where entrepreneurs could promote their brand
- Recruitment facility could be developed and it can turn into an interesting feature
- Extend it for marketing campaigns

The conclusion from talking to the marketing specialists would be that they would be interested in collaborating with DatoSphera, as the second step in market research, but they would first want to see the performance and the usage of the application in the very first years. This means that the strategy should be based in finding and maintaining entrepreneurs, to have a batch of contacts for the marketing agencies and agents.

2.3.Competitor analysis

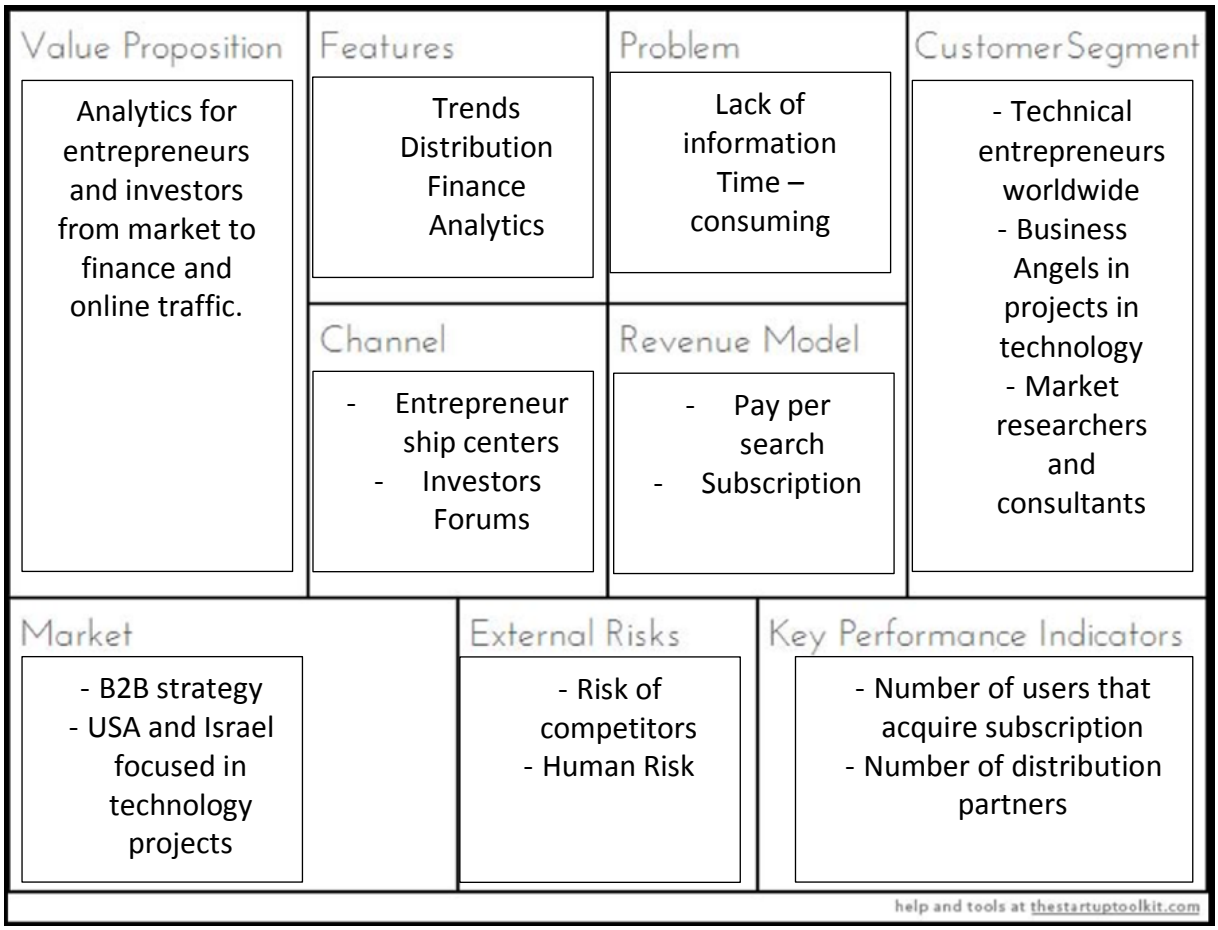
So far the competitive analysis of DatoSphera reveals very few direct competitors, most of them having connected activities and that could easily transform into partners.

Competitor	Activity	Competitive advantage of DatoSphera
Quick sprout	Is studying all the traffic data that exists on a certain web domain.	DatoSphera involves financial, traffic and market data, compared with Quick sprout. DatoSphera focuses on entrepreneurs, while Quick sprout targets companies.
Sisense	SiSense is using Crunchbase to offer generic financial information's and statistics.	DatoSphera offers personalized analysis for strictly entrepreneurs and investors and not general ones.

Kapowsoftware	The software helps the companies to merge different data sources and to generate analytics.	There is no evidence that this company is focusing on entrepreneurs or is using any of our data sources. DatoSphera already has the data merged and doesn't depend as much on the user input.

2.4.Business Model

The business model is the result of the market research conducted and it can summarize in this form:

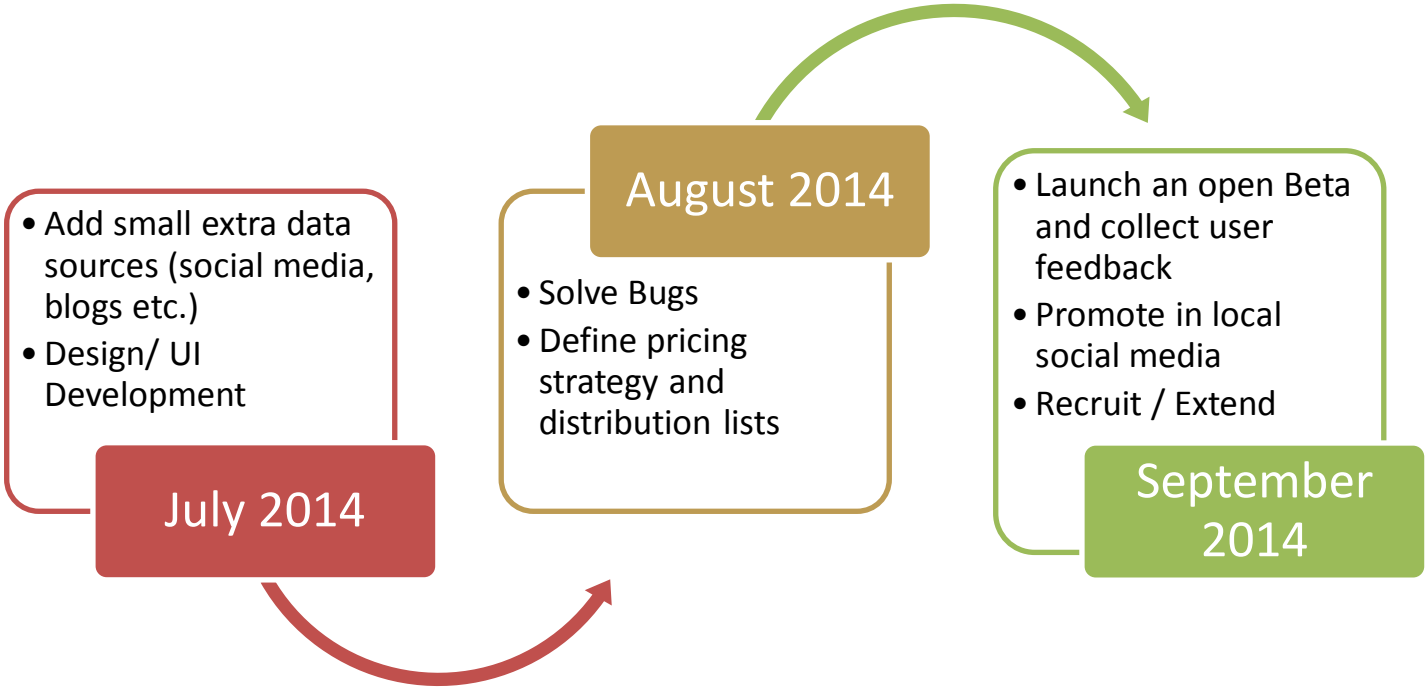


As we can observe, DatoSphera implements a business model that allows the users to pay per search or to acquire a subscription, given that the needs of its users are very diverse.

2.5.Road Map

The roadmap reflects the product development, once the team is completed and the company was fully incorporated and funded.

As we work with Lean Startup technologies, the Road Map is realized only for 3 months and adapted to the changes.



2.6.Marketing and distribution strategy

The distribution scheme is based on the following channels:

- Entrepreneurship centers
  - They will distribute the tool to the entrepreneurs and to the start-ups that want to evaluate themselves on the market
- Investors forums
  - They will distribute the tool to their investors and evaluate the projects they receive

Certain geographic market evaluations have been realized and by checking the top cities where the entrepreneurs and investors are present, we decided to focus promotion strategies in the USA and Israel market. On Israel we already have a distribution partner, Tel Aviv Centre of Entrepreneurship. On USA were still evaluating the possibilities to enter.

2.7.Branding

The name of the company has changed from Info Posit to DatoSphera, given that the principal objective of the company is to offer Big Data analytics. For all that the current name generates confusion given the objective of the company and its link with entrepreneurs and investors. The very first version of the logo was the following one:

We considered that this version of the logo was too complex in terms of the small data cubes and therefore was changed to the following and current one:

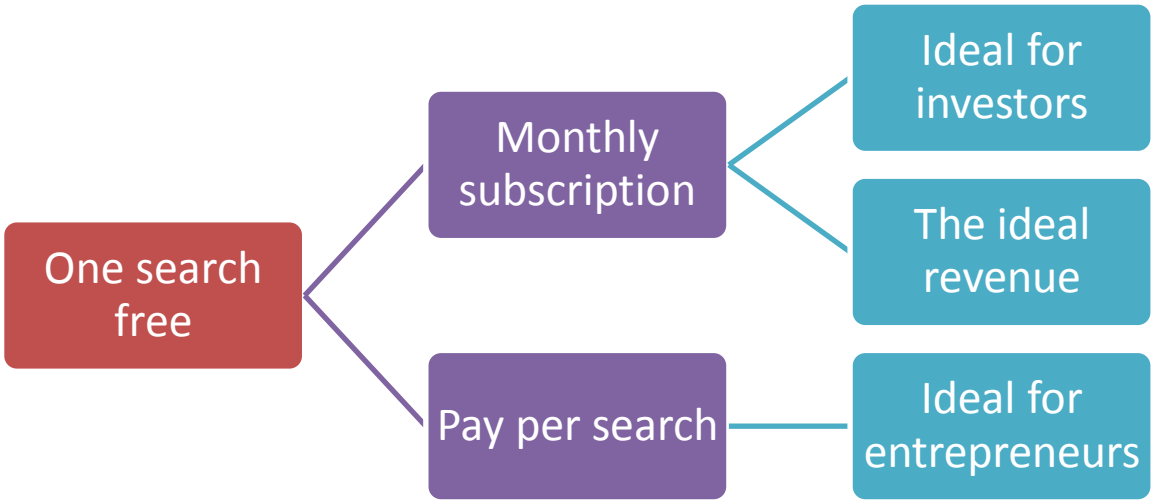
We simplified the original version, combining colors associated with technology. The claim was also changed to reflect better that the product is evaluating business ideas.

2.8.Pricing strategy and associated costs

During the market research several issues about pricing have been raised and therefore the only strategy to monetize the application had to deal with the following challenges:

- There was a high mistrust from the users regarding the accuracy of data
- There were different sorts of users with different needs
- There appeared to be a huge number of possible users that would use it for various reasons

Therefore the strategy defined was the following:



The user has the option to choose the best option for his/her needs. We understand that the tool will be used for various reasons and purposes and this proved to be one of the most transparent strategies.

As DatoSphera is part of an incubator, all the services are offered except for the costs of the infrastructure dedicated to the project. These costs include:

- Instances EC2
- Auto-scaling
- Elastic Load balancer
- RDS (Relational Database Storage)

- S3 (Simple Storage Service)
- Elastic Cache
- Cloud Front
- Route 53
- Amazon Beanstalk

All these costs and the personnel costs are fully covered from European grants and Incubio. The finance documents are to be delivered at the particular terms when the projects can apply for funding.

2.9.Product and features

DatoSphera aims at being a complete analytics service focused on entrepreneurs and investors. Therefore the product build is designed to create a competitor analysis tool, focused on giving tips for entrepreneurs and investors.

2.9.1. Value proposition

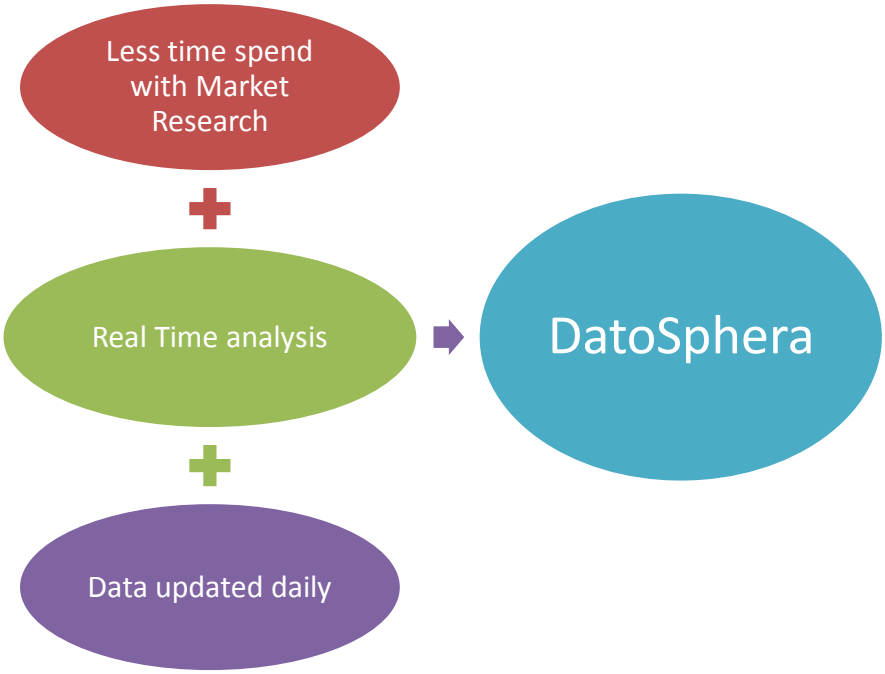
What DatoSphera offers to entrepreneurs is:

- Access to worldwide competitors
- Real-time analytics – data updated daily
- Estimations of funding achieved per sector
- Online data aggregated into one single point view
- A simple and intuitive interface
- Results offered in real time

What DatoSphera offers to investors is:

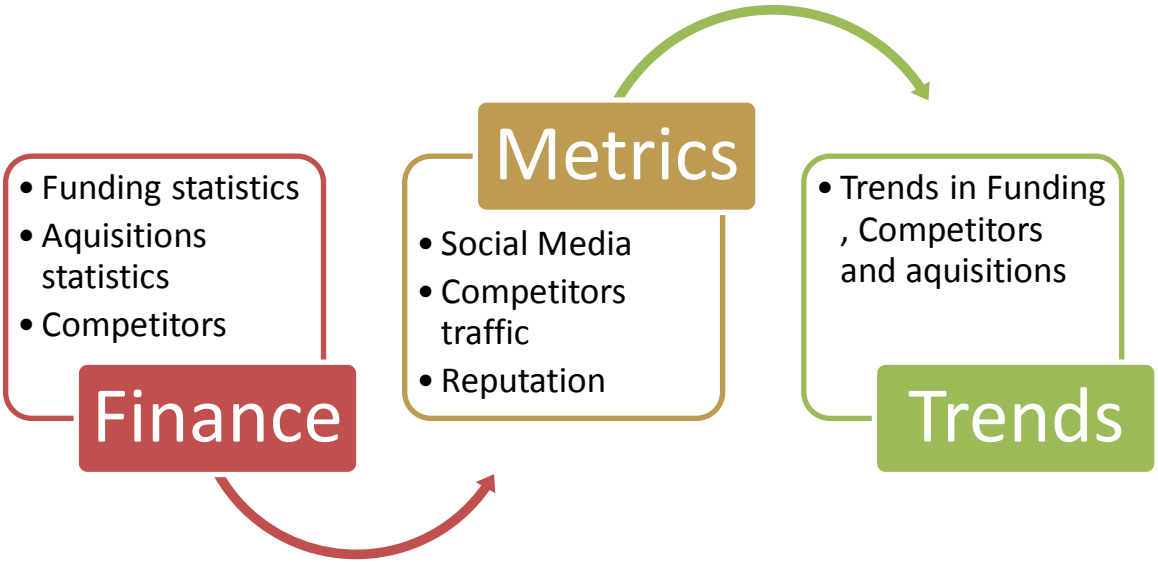
- Saved time in evaluating projects
- See the market potential of a product
- Access to the funding trends worldwide
- Compare two possible business ideas to be funded
- Study trends of acquisitions

Overall the value proposition of DatoSphera is summarized in the following schema:



2.9.2. Features

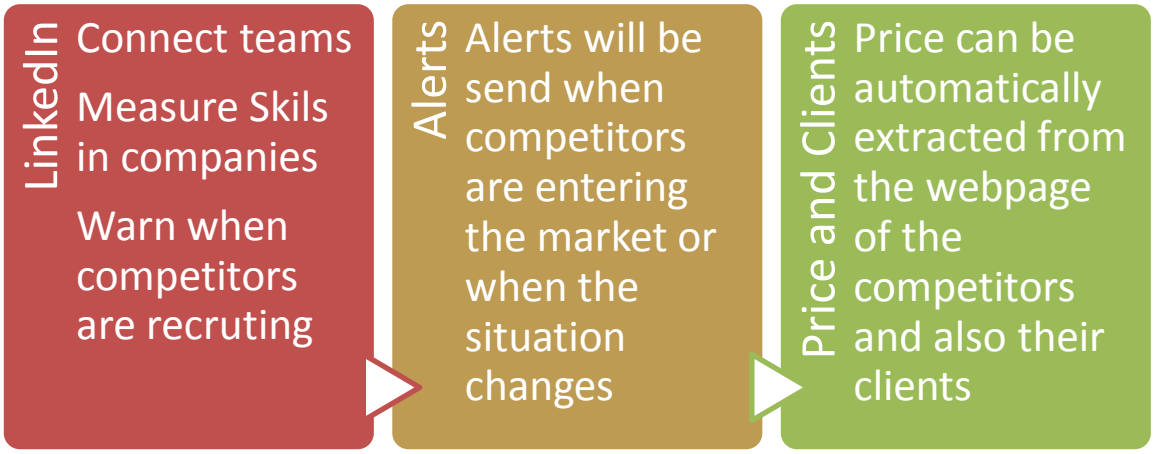
The product is offering a great variety of information, making it a complete solution for entrepreneurs and investors. The variety of data is given by the data sources and is represented in the product features:



The diversity of information available makes DatoSphera a complete option for entrepreneurs that want to study their market, before joining it.

2.9.2.1. Future features to be considered

Given the user feedback, the following features will be introduced, after their importance to the user will be defined:



All these modules have been raised during the interviews or the sessions of brainstorming with the users and the future step is to prioritize them according their complexity and their relevance to the user.

2.10. User retention strategy

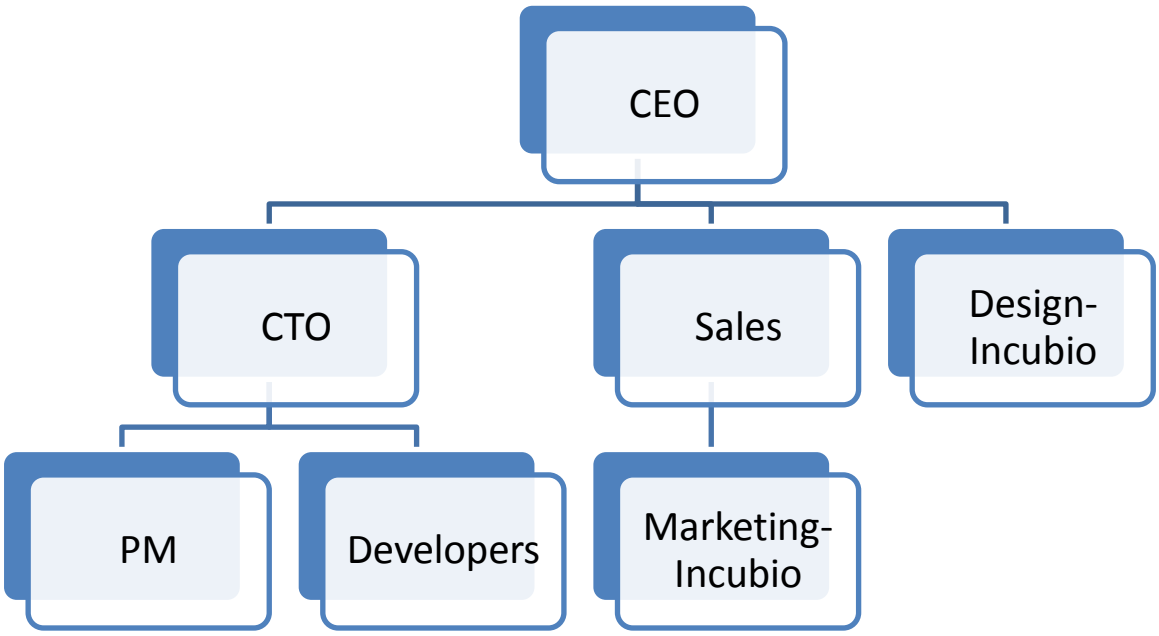
Once a user has done a search on the platform, there will be stored and automatic updates will be given on email, at certain times. This will ensure that the platform is not a one-time use but it has a continuous use.

The model is inspired by the Signl.com model of retention and stores the searches that a person has done and after that sends updates by email whenever the scores for the business idea change or a new competitor enters in the market.

The goal of this strategy is to create a need and to convert users from simple searches to monthly subscriptions, even if they are entrepreneurs.

2.11. Organizational Structure and Human Resources

The team follows the structure above:



The following persons are forming the team:

Madalina Burghelea, CEO & Founder
She began working as an R&D Engineer at Incubio Research, the center of innovation from the incubator Incubio.
Being surrounded by entrepreneurs who struggled to validate their ideas, Madalina was inspired to create DatoSphera.

Jordi Masramon, CTO
A native from Barcelona, Jordi has a MBA specialized in Entrepreneurship, Finance and Negotiation Techniques from University of Minnesota, Carlson School of Management, in addition to studies in Information Technology and Telecommunications Engineering.

Xavier Ruiz Project Manager
Xavier studied technology at Polytechnic University of Catalunya and is passionate about entrepreneurship and after several years of dealing with cloud architectures as the Systems Architect at ADmira, he moved to Incubio as a Project Manager.

Denis di Paolo Data Mining Specialist and Developer
Denis received his Masters in Computer Science Engineering through a 6 months' work experience in Istanbul, at Boğaziçi University, working with a team composed of people from all over the world.

Jose Luis Padilla Sales Specialist
Jose Luis is an Executive MBA with ten years of experience in consumer goods in sales, marketing, financial and executive processes.

Incubio

DatoSphera is part of the Incubio community: a start-up incubator specialized in the creation of projects that offer business processes as a service, with the use of Big Data technologies. Incubio provides all the other services necessary such as Design and Marketing, through the specialized departments.

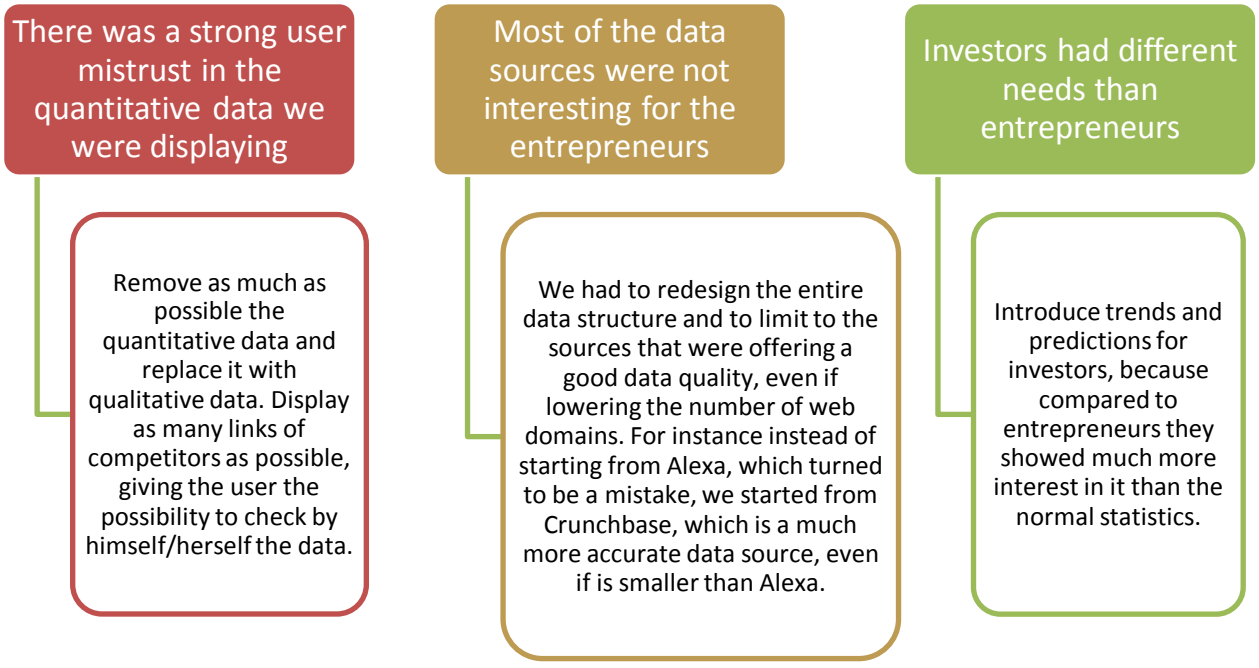
2.12. Company legal information

We company was officially registered in Barcelona on the 28th of April as an S.L. having as founder Madalina Burghelea and as partner, Incubio Partners S.L.. Previously the project was inside of the Research Department of Incubio, as the Trakty Project. DatoSphera is a spin-off of the Trakty Project, sharing the database with the Trakty project and Incubio Research.

2.13. Strategic business decisions

During the market research there were several issues that were critical for the product development and these issues are strongly correlated with the technical development.

The critical issues were the following:



The project is following the Lean Start-up Methodology and therefore it tests the hypothesis with clients, causing the entire project to change the direction. The AGILE methodology proved to be one of the key success factors.

3.Project Management

The Project Management part has two major components: The Business Development methodologies and the Software development methodologies. The choice for the methodologies was motivated by the team structure and the business road map. Two compatible methodologies are Lean Startup and AGILE, which will be detailed below.

3.1.Business Methodology- Lean Startup

Given that DatoSphera is a real project, dealing with real issues and time constraints, Lean Startup was the methodology chosen to develop the product.It is based on the validated learning concept, which states that previous experiences, mistakes and failures are contributing to the success of the next product release.

The Lean-Startup methodology states that iterative product releases and involving the customer in the releases are key success factors.

This is why DatoSphera proceeded with Co-creation workshops, customer interviews and focus groups. All the conclusions were then refined to a Minimum Viable Product schema. Once developed this prototype, users were testing it and giving feedback.

3.2.Software Development Methodology- AGILE

The methodology imposed by Incubio is the one that proved to work very good in data projects and this is an AGILE methodology. Furthermore Lean Startup contributed also to the AGILE methodology, because the end user was always part of the project. The product owner had the role to evaluate, collect and plan extensions, all according to the user feedback.

Being AGILE also translated into choosing the modules technically easy to implement and quickly test the user reactions to them. Therefore we ensured that there is no module that is not going to be ignored by the users.

3.2.1. SCRUM

SCRUM methods are basically a reflection of AGILE into the way that code is designed. Key concepts of SCRUM were:

- Daily meetings
- The existence of a SCRUM master, preferably not from the team
- Frequent changes coming from the product owner
- Let team self-organize

The main asset of SCRUM is the ability to predict the time for each task.

3.2.2. Kanban

Kanban as a model comes from industry, where a bottleneck can have huge impacts on the work of the team. This means that the workflow should always be equilibrated and also the time distribution.

It is based on a classical board, where the tasks are counted, to see where is possible to occur a bottle-neck. An example is shown below:

	Requirement analysis	Development	Test
Limit of tasks:	3	5	3

Basically the new issues are related to limiting the number of tasks, based on the team potential.

For instance there can be situations when instead of 5 tasks in development there will be 7 and a bottleneck could be detected immediately:

	Requirement analysis	Development	Test
Limit of tasks:	3	5	3
Current tasks:	3	7	1

This means that the development team is overloaded, while the testers are being delayed because the features were not implemented.

The main advantage of the Kanban model is time-saving and avoiding overload of work.

3.2.3. Critical evaluations: SCRUM vs. Kanban in DatoSphera

The project was developed in different stages and the necessities at the particular stages were therefore different, motivating the usage of both methodologies.

In the Inception sprint (the very first iteration, dedicated to forming the team and exploring the business opportunities), Kanban was chosen as the best option for the following reasons:

- The team was very small (consisting of 2 persons) and the as the team was at the beginning, flexibility was needed.
- The work was based on brainstorming on the most simple technical solution
- There were no work estimations, to choose SCRUM as the best option

When the team was growing to 5 persons, working together on certain modules, Kanban was not sufficient any more, being too flexible with the workload.

Another issue of Kanban was the inability to predict the workload.

Whenever there were deadlines to accomplish for different events, the product was never ready in time, because of wrong estimations done with Kanban.

This made the team realize that was time to change to SCRUM, and to be able to predict how much time the team will need to develop the modules. Furthermore, as the team was growing two times, the necessity for a more strict methodology occurred.

A comparative analysis is presented below:

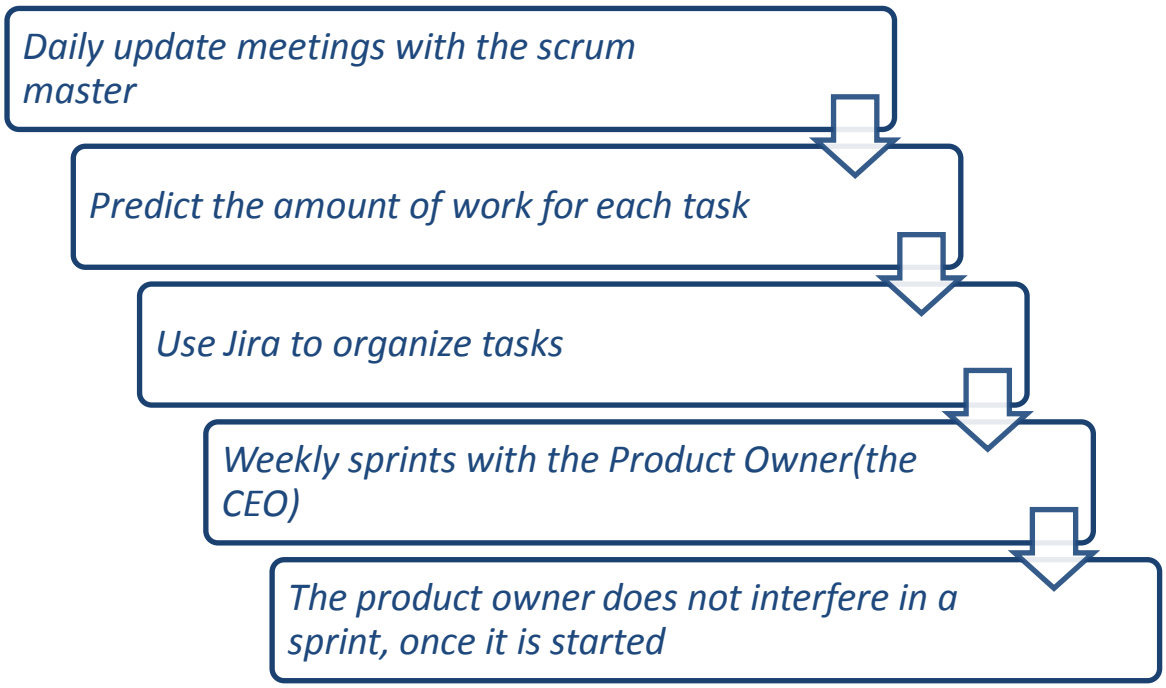
SCRUM	Kanban
What is the best time approximation?	**What is the best time-distribution ?**
Strict with timing	Focused on avoiding bottlenecks
Focused on time-estimations	Flexible with timing issues
Ideal for big teams	

Kanban was successfully implemented in the team of DatoSphera, due to its flexibility and focus on predictions. Due to the fact that the rhythm of the team members is different and many unpredicted problems might occur, flexibility was the key factor.

3.2.4. Current Solution

3.2.4.1. SCRUM Methodology

We have decided to use SCRUM, because the team structure was suitable for a more rigid model. The SCRUM implementation was following the pattern below:



3.3.Tools for project management – JIRA

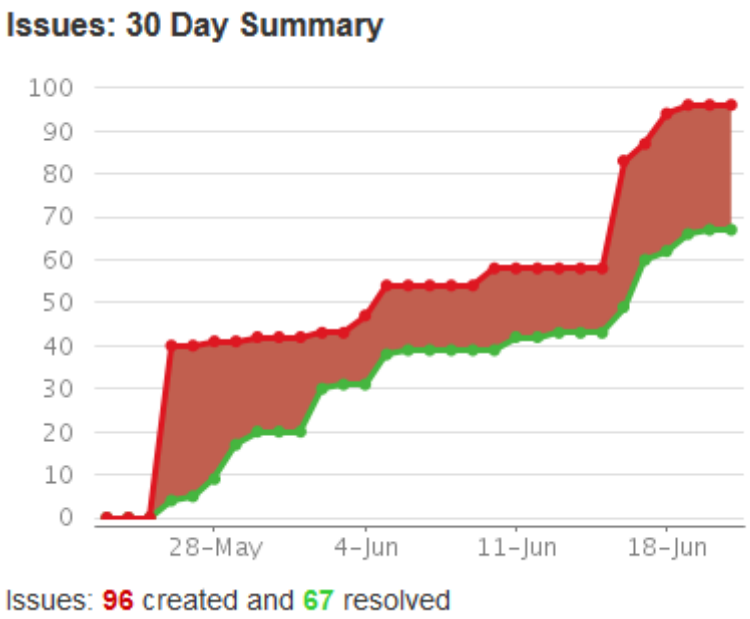
JIRA is an instrument for tracking bugs, user stories and tasks within a team, while developing a product.

JIRA was chosen in DatoSphera because the experience with JIRA in Incubio confirmed that for the other similar teams worked perfectly.

Other criteria for choosing JIRA is the fact that incorporates all the possible tools for project management, from boards, tracking and communication.

3.3.1. Optimizing JIRA with strictly defined priorities

Another decision taken regarding the project management aspect and the detailed tasks was to start adding priorities to tasks because it generated a situation in which many issues were created in JIRA, but without priorities and therefore these issues were paused for a long time. The situation is revealed bellow:



Therefore we decided to add priorities also in Jira, to make the pausing of issues more difficult and to discourage it.

3.4.TDD Method

TDD is a methodology for software development that states the importance of very short tests of minimal code to pass and then refactors the new code. It explores the potential of short development cycles.

Given that we follow the AGILE methodologies, tests of code that work are very important for the product itself.

To encourage working tests we have added in the definition of done for each of the stories that the tests should work perfectly, because otherwise was leading to quality issues.

One major issue when working with TDD is the scalability, because the test suite takes a lot of time to run and therefore affects the current development. One the other side, regarding the human resources aspects, getting the new members in the team to use and properly understand TDD was a long process.

We have performed unit testing with AngularJS as a framework.

3.5.Team Roles

During the different phases of the project the team roles were slightly modified but the usual work distribution was the following:

Product Owner	SCRUM Master	Technical Leader	Data Processing	Front-end
Madalina	**Xavier**	**Jordi**	**Pablo**	**Denis**

During the daily meetings, the SCRUM Master along with the development team stands by the board and discusses the evolution of tasks. Each task is written on a post-it and is assigned to a particular person and moved in one of the three categories. An example is presented below:

During the daily meeting, the team and the SCRUM master were standing by the board and discussing the progress done and also the problems. The daily meetings were for the technical team only, as the Product Owner is not allowed to change the stories during a sprint.

The weekly meetings with the Product Owner were assigned to the following tasks:

- Demo of the progress
- Discuss challenges
- Make the sprint review
- Decide the stories that are going to be tackled for the next sprint

An example of a demo meeting is shown below:

The team roles definition helped DatoSphera to better organize and focus on the task distribution.

3.6.Product Backlog

The product backlog is the result of the market analysis, summarized in Chapter 2- Section 2 and it is derived from the customer interviews and the co-creation workshop. The priorities have been assigned depending on the quantitative evaluation realized with the Google forms.

User Story	Business Priority	Technical Complexity
As a company representative I have to see all the information online about my competitors	1	Medium
As an entrepreneur I want to see in which countries are my competitors based.	2	Low
As an investor I want to see charts about market evolution	3	Medium
As an entrepreneur I want to see from where and how much I can get funding	4	Low
As an investor I want to see how easy is to sell my company	5	High
As an investor I want to see the number of competitors evolution	6	Medium
As an entrepreneur I want to see if somebody else tried that before and failed	7	High
As an entrepreneur I want to check my direct and my indirect competitors	8	Medium
As an entrepreneur I want to retrieve all the competitors, doesn't matter if they exist from very short term on the market	9	Medium
As an investor I want to access the most updated data that	10	High

exists		
As an European Investor I want to narrow my search to the European market	11	Low
As a market researcher I want to see why companies failed in this entire sector	12	High
As an entrepreneur I need as many contacts as possible, from investors, to get funded by them	13	Low
As an investor I want to see the traffic associated with the certain business.	14	Low
As a market researcher I want to download a pdf with all the information.	15	Medium
As a business developer I want to see geographically where can I extend	16	Medium
As an entrepreneur I want to log in with my LinkedIn account and to have all the traffic data about my company, compared with my competitors	17	Low
As a market researcher I want to see how much the users trust my competitors	18	Low

The order of execution was following the Lean startup methodology and was based on the following schema:

	Low technical complexity	High technical complexity
Low Business Priority	III	IV
High Business Priority	I	II

As you see, regardless the technical complexity, the business priorities and the user requirements will be considered first, even if the time dedicated to them is very high. The following chapter will explain how exactly the user stories were tackled.

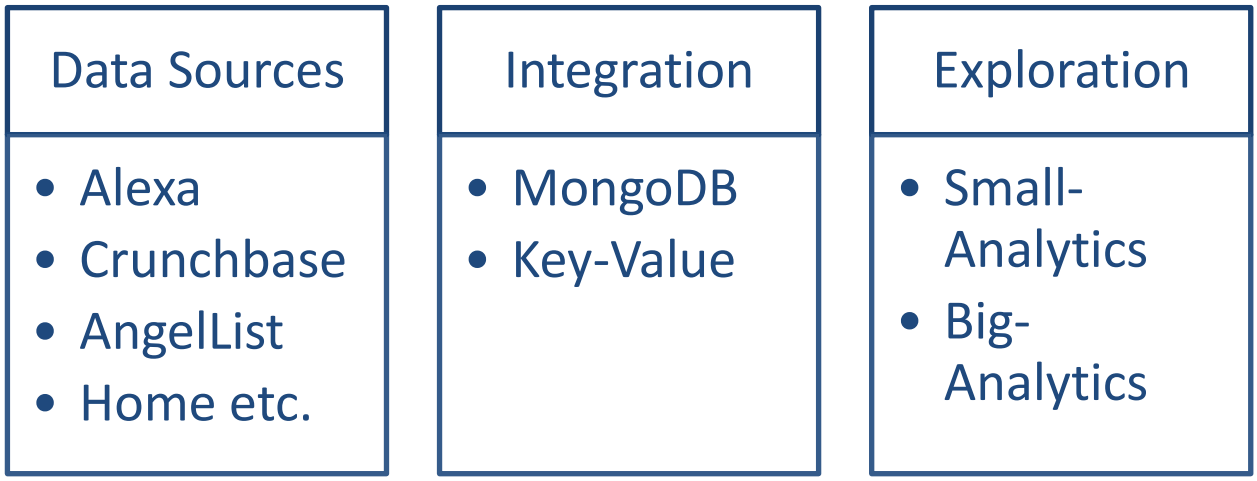
4.Development of the data architecture

The implementation of the current data architecture is a result of the business and project management constraints presented in the previous chapters.

DatoSphera is a spin-off of the project Trakty, which gathered different data sources during a period of 6 months. Trakty is a project belonging to Incubio Research, which provides the projects incubated an additional help with the Big Data technologies. These data sources were aggregated for a different purpose and their quality was not measured accordingly. This translates into a strategic decision on using the Trakty data or following the same principle but with different data sources. The decisions will be motivated below.

A full analysis of the data sources, of possible schemas of integration layers and exploration techniques has been done.

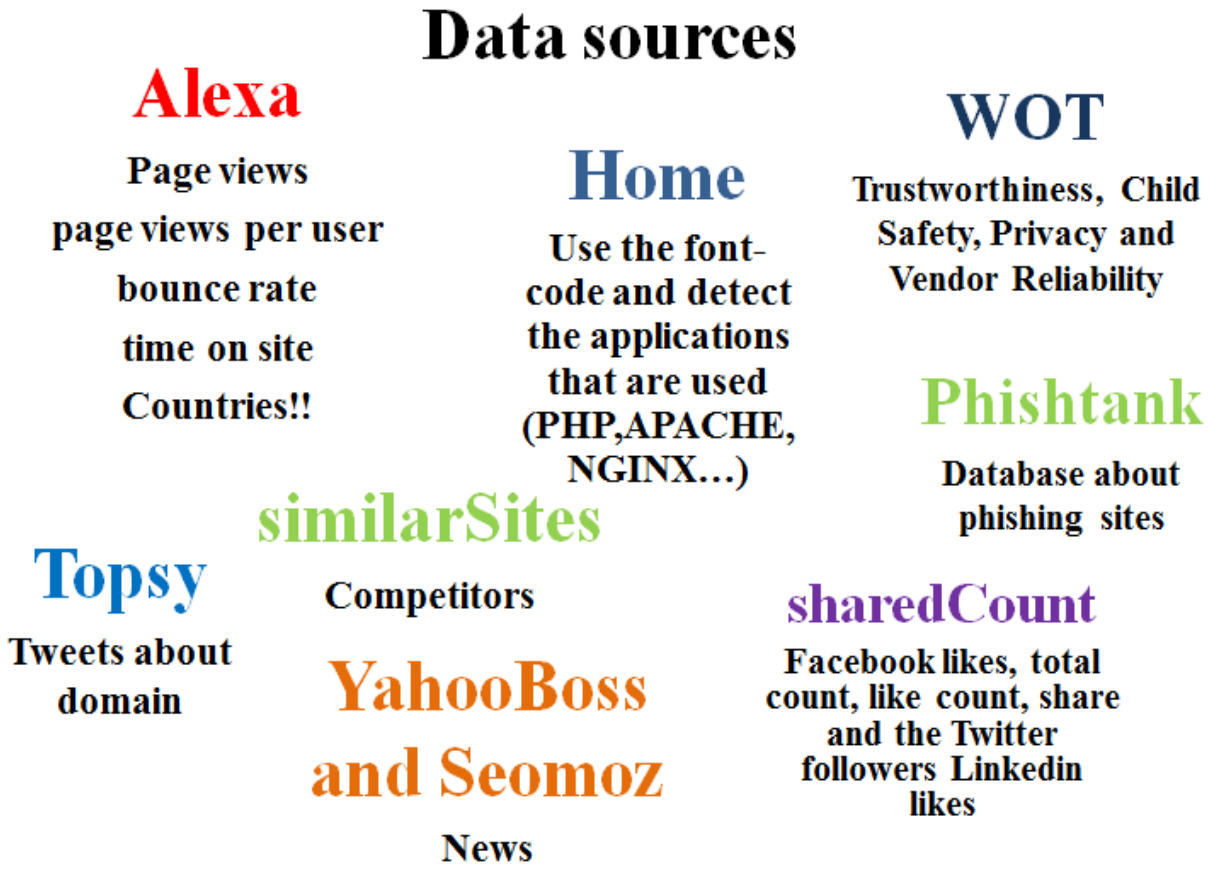
The chapter is organized into 3 layers:



Each of the different layers will be explained into a different section. Each of the sections has correspondents into sprint that are annexed and reflect the historical evolution of the project.

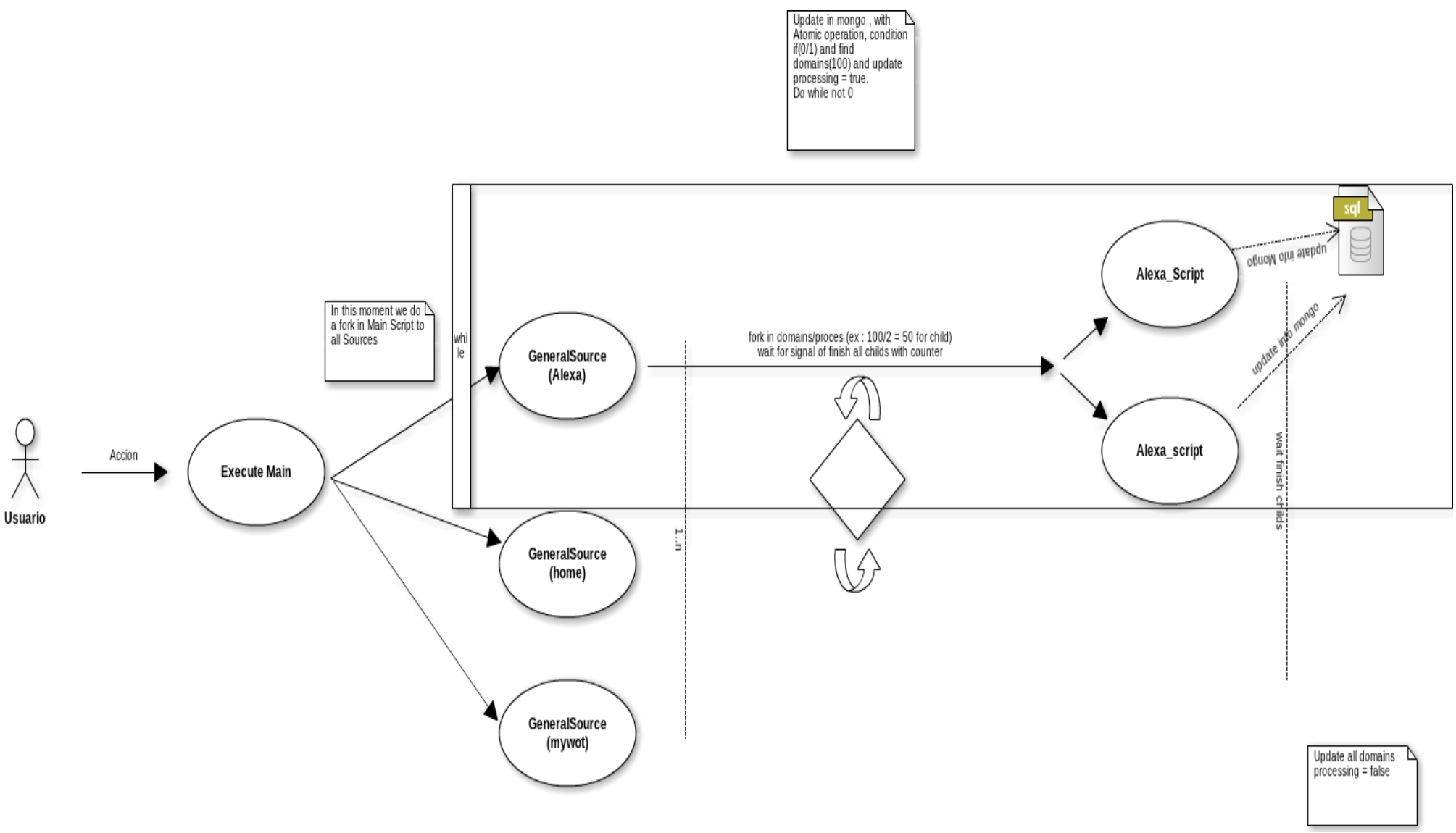
The data sources have been selected strictly on being as diverse as possible and on measuring data relevant for web domains.

The following data sources were analyzed firstly for being aggregated. Afterwards Crunchbase and Angel List were added to this list. The problems come from the heterogeneity and the qualities of these different data sources.



These data sources were extracted with the APIS provided and through scrapping and merged based on the web domains they were referring to.

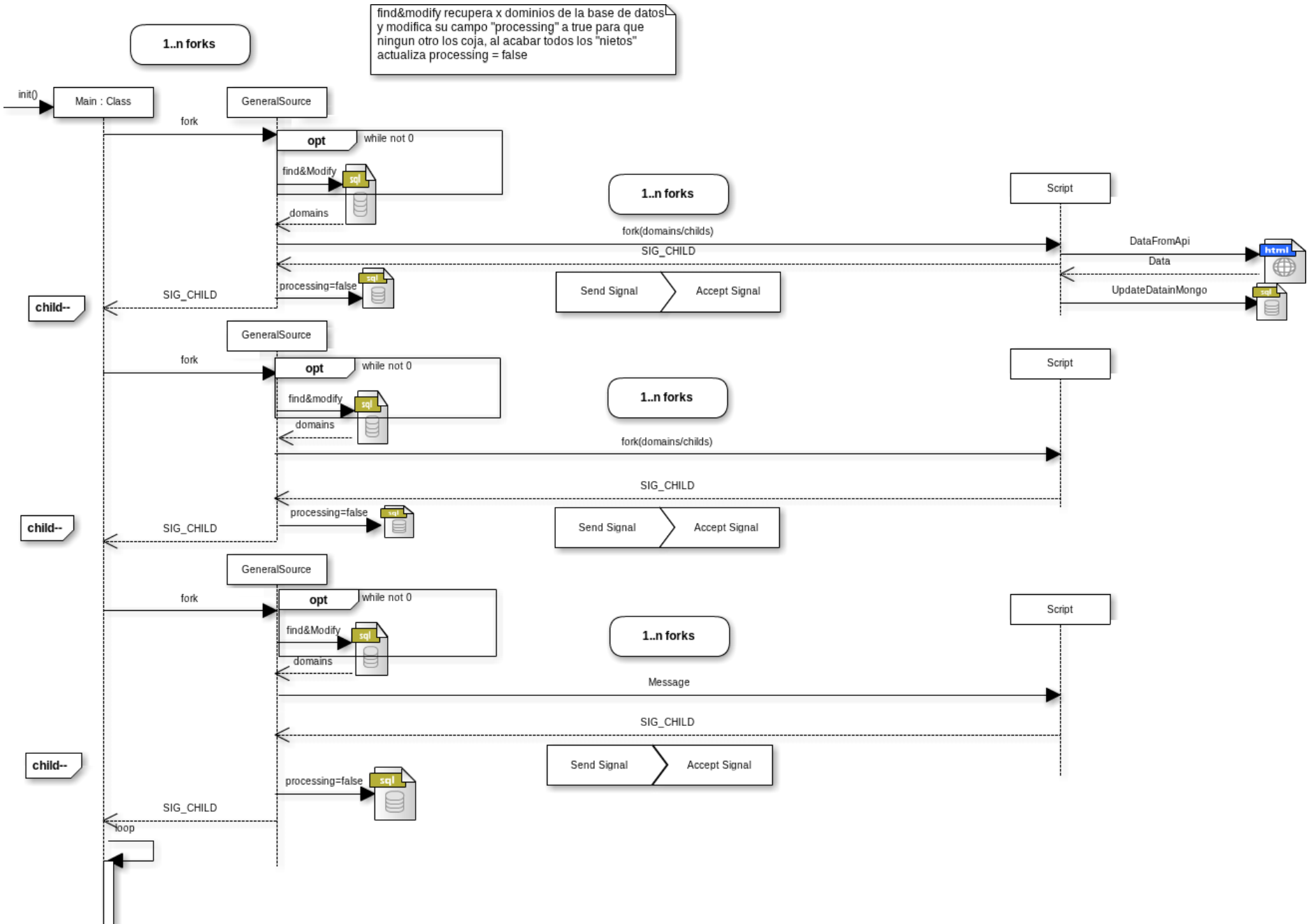
The following schema reflects the general strategy to extract data:


Update in mongo , with Atomic operation, condition if(0/1) and find domains(100) and update processing = true.
Do while not 0
sql
Alexa_Script
update into Mongo
update into mongo
Usuario
Accion
Execute Main
In this moment we do a fork in Main Script to all Sources
whi le
GeneralSource (Alexa)
fork in domains/proces (ex : 100/2 = 50 for child)
wait for signal of finish all childs with counter
Alexa_script
wait finish childs
GeneralSource (home)
GeneralSource (mywot)
1...n
Update all domains processing = false

The schema reflects a perfect parallel application that proved to be one of the best strategies to work with this amount of data coming from different sources. The distribution between children of the same script was an efficient way of working with chunks of data.

There are certain discussions regarding this parallel approach such as timing of operations and evaluation of how much these processes last. Working with children inside the same data source and defining forks for them can raise synchronization problems .

The following schema details even more the extraction process and this schema proved to work good, until introduction of versioning:


1..n forks
find&modify recupera x dominios de la base de datos y modifica su campo "processing" a true para que ningun otro los coja, al acabar todos los "nietos" actualiza processing = false
init()
Main : Class
GeneralSource
fork
opt
while not 0
find&Modify
domains
1..n forks
Script
fork(domains/childs)
SIG_CHILD
DataFromApi
Data
UpdateDatainMongo
processing=false
Send Signal
Accept Signal
child--
SIG_CHILD
find&modify
Message
loop

As we can see all the web domains are equally distributed by the main script between the children, each one of them writing the results in the integration layer and signaling to the main script that the operation is finished.

A parallel application was defined, given the different web sources available and their volume. The application involved using Forks in PHP and defining processes that generates a certain number of forks, depending directly on the number of sources. Each source will be extracted automatically using the particular API or even scrapping the data through CURL. The data is afterwards extracted as a JSON document and send to the integration layer.

For example while trying to execute the following line of code: nohup php execSource.php alexa sharedCount home , the following process is defined:

1. General Source creates a fork that generates afterwards all the other forks to process the other data sources and also controls memory and the loop through the domains
2. In integration layer General Source calculates the necessities for each of the data sources: for instance Alexa for 1000 domains will use 4 sub-scripts and then each of the processing sub-scripts will take 250 web domains.
3. When the domains are finished to process by each of the sub-script, General Source will delete the fork and create another one to process these web domains.
4. In principle the General source stops after 1 million domains, considering also the elapsed time to avoid the infinite loops.

It reflects the same distribution with sub-scripts for each of the data sources and then writing directly to the integration layer, once the task has finished.

The script is adapting for each of the data sources and distributing the web domain data to be found among them. Once a task is completed and the processing variable takes the value 1, the script will be reused for the next task.

There are two main groups of data sources and are summarized below:

Based on companies	Based on web domain
Use as key the name of the company	Use as key the web domain
User inputs the information	The information is measured on the web
Crunchbase and AngelList	Alexa, SemRush, SharedCount, MyWOT

The links between a company and a web domain was not always available and therefore a full analysis of coverage was performed, to check the missing data.

In this very moment there were problems extracting the name of the company, which was extracted from the URL. For all that a big ambiguity was left because this strategy was not correct, because one web domain could have been associated with more than one company, causing a problematic 1-TO- MANY links in the database.

One problematic application requirement was the lack of good quality tags, because the organic keywords from Alexa were not sufficient to retrieve companies. The quality of the Alexa data was one of the lowest ones, causing a real problem when retrieving accurate data.

In that moment, we thought about adding Crunchbase as a source, because it was one of the only one online data sources that were containing tags that were of a good quality. The problem of adding Crunchbase was that the Crunchbase web domains were matching very few of the Alexa domains. This means that if the application would have retrieved data by the tags, would have miss many of the Alexa domains, uncovered in Crunchbase.

The table shows the coverage and missing proportions between Alexa and Crunchbase:

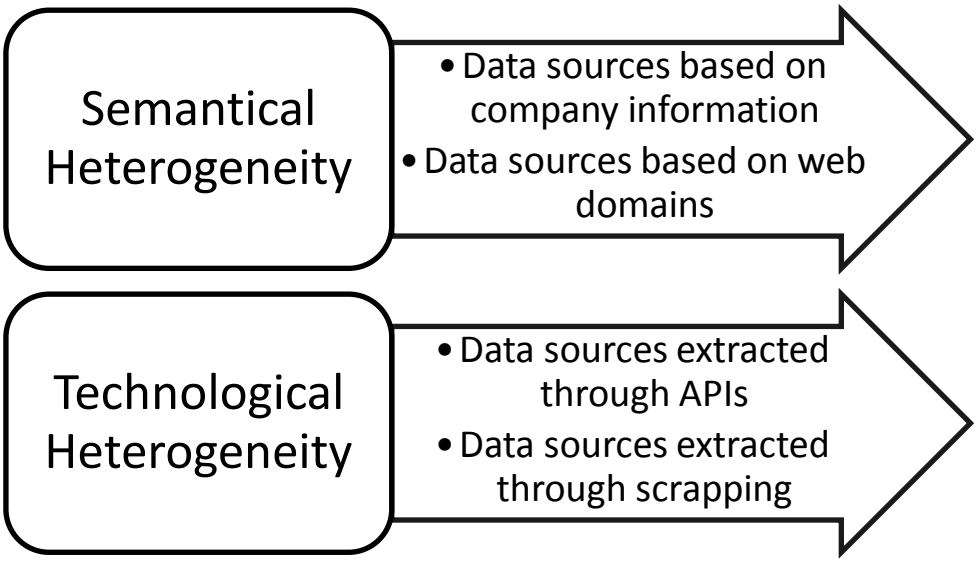
TYPE DATA / SOURCE	**PERCENTAGE MISSING**			**Total**		
	Domains TOP (ALEXA)	(201413) Companies in CrunchBase				
		Companies in Top 1M Alexa	Companies out of Top 1M Alexa	**(Alexa Top 1M + Companies Crunch)**		
	1000000	27852	173561			
rank global / Alexa	0%	0%	100%	14,79%	1000000	173561
bounce rate / Alexa	0%	0%	100%	14,79%	1000000	173561
daily page views / Alexa	0%	0%	100%	14,79%	1000000	173561
daily time site / Alexa	0%	0%	100%	14,79%	1000000	173561
Company info / CrunchBase			-	14,79%	1000000	173561
FundingRounds /			-	14,79%	1000000	173561

CrunchBase						
Investments / CrunchBase			-	14,79%	1000000	173561
Acquisitions / CrunchBase			-	14,79%	1000000	173561
Trustworthiness / Wot	80%		100%	82,96%	1000000	173561
Vendor Reliability / Wot	25%		100%	36,09%	1000000	173561
Privacy / Wot	34%		100%	43,76%	1000000	173561
Child Safety / Wot	75%		100%	78,70%	1000000	173561
StumbleUpon / SharedCount			100%	14,79%	1000000	173561
Facebook / SharedCount			100%	14,79%	1000000	173561
Twitter / SharedCount			100%	14,79%	1000000	173561
LinkedIn / SharedCount			100%	14,79%	1000000	173561
Competitors / SimilarSites			100%	14,79%	1000000	173561
Home info / Home			100%	14,79%	1000000	173561
Keywords / Home			100%	14,79%	1000000	173561
Applications / Home			100%	14,79%	1000000	173561
Influencers / Topsy			100%	14,79%	1000000	173561

Looking at these coefficients we realized that Alexa, Crunchbase and the other sources we added were covering very few common web domains, **less than 15%,** which determined us to analyses the viability of the project.

This meant that we had to choose between Crunchbase and Alexa and from then on, work only with the remaining compatible sources. A full explanation of the compatibilities of the sources (based on user input- like Crunchbase and based on studying the web domain- like Alexa) is presented below.

The data heterogeneity could be categorized into the following aspects:



The data acquisition process is very similar for the sources, by categorizing all of them into 2 groups:

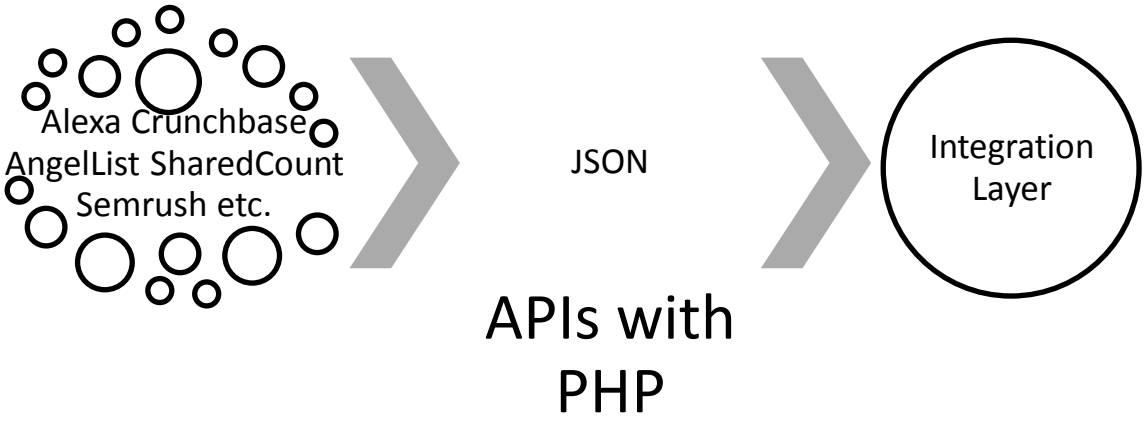
	Data Sources where users inputs information	Data Sources that study the web domains automatically
Examples	Crunchbase, Angel List	Alexa, Topsy, Semrush, SharedCount etc.
Risks	Accuracy Dependent of user Subjectivity in describing tags Updates of data whenever the user decides to	Automatized process need to be verified
Benefits	Data that normally wouldn't be available in any source	Very accurate data (automatic) Easy to work with the APIs Real time updates

The data is after extracted with the API from the sources. So the quality of the data is not under DatoSphera control, but under the sources control.

The data sources, although offering different information, were uniform in terms of formats. Each one of the sources (except for Alexa) has an API that made possible to extract all or partial information. In case of Alexa, an API was not available so we had to extract the front-end code and start processing from this point. In case of Angel List the data was extracted only partially, due to legal reasons.

Therefore the data was heterogeneous in the kind of information they offer, but pretty much homogenous, as an API to extract JSONs documents was available for all of them.

Each of the tools was offering an API to extract the data. The processing to extract the data sources was facilitated by their APIs. A schema is shown below:



As you can see that data source initially were extracted with PHP , all of them as JSON documents and therefore stored in the integration lawyer.

Details on the each one of the data sources will be presented in the following section.

4.1.Data Sources

4.1.1. Alexa

4.1.1.1. Description

Alexa is one of the most famous data sources, containing more than 1 million web domains. The data available on the web domains is relatively clean, lacking only keywords of activities. For instance the information available about google on Alexa is retrieved with the following link: http://www.alexa.com/siteinfo/google.com and contains web traffic, geographical distribution and general statistics.

4.1.1.2. Data Model

We started by extracting the data that **Alexa** was offering, having a big inconvenient that at that time there was not an API to extract all the data in structured format, but we extracted an HTML code of each website and starting from this code, using DOM structures, we managed to extract all the information and place it in a structured form.

We have stored the front code and then extract the entities. An example of the results for Bing.com is shown below:

```
[extracted data] => Array
        ( [pageviews] => Array
            (
              [yesterday] => 0.1916%
              [7 day] => 0.2055%
```

```
        [1 month] => 0.2103%
        [3 month] => 0.22306%
    )

[bounce rate] => Array
    (
        [yesterday] => 36.5%
        [7 day] => 36.0%
        [1 month] => 36.0%
        [3 month] => 40.9%
    )

[time on site] => Array
    (
        [yesterday] => 3:36
        [7 day] => 3:42
        [1 month] => 3:39
        [3 month] => 3:23
```

The advances of this way of processing was that was cost-efficient but the html codes were significantly variable, making the extraction very difficult. The only thing that happens in the structure is that there are new fields being added, but the process is more or less automatized, not generating any problems and continuously updating data.

4.1.1.3. Data Quality

Alexa				
Accuracy	**Completeness**	**Consistency**	**Freshness**	**Mentions**
The accuracy of Alexa is based on internal algorithms that Alexa uses to extract data, therefore dependent on the source.	More than 30% of the records have missing data.	Various tests have been performed with Traffic Estimate.com and the web traffic is consistent in more than 90% of the cases, with very small variations.	The data is up to date as is based on the websites. The modifications of websites are visible within few days in Alexa.	Alexa studies the websites of the companies but it doesn't provide some keywords to understand what a web domain is about. The main handicap of Alexa is that other than the organic keywords, doesn't provide any industry or good quality keywords.

4.1.2. Crunchbase

4.1.2.1. Description

Crunchbase is a data collection of information on companies, mostly popular in the USA and updated once there is funding given to companies.

4.1.2.2. Data Model

Crunchbase architecture is based on storing the company's information as JSON and providing an API to have it extracted. The Crunchbase API documentation is more than sufficient and comprehensive and the usage of the API is encouraged by Crunchbase.

An example of the data that could be extracted from Crunchbase is shown below:

Type of information and usage	JSON correspondent

General Company Information that will be displayed individually	`"name": "Facebook",`
	`"number_of_employees": 5299,`
	`"founded_year": 2004,`
	`"deadpooled_year": null,`

Tag List to retrieve the companies in the search results
 	
The description and the category will be used as parameters in retrieving the competitors.
	`"tag_list": "facebook, college, students, profiles, network, online-communities, social-networking",`
	`"description": "Social network",`
	`"category_code": "social",`

The people that work for the company are specified in the section relationships and these persons will be displayed as contacts to the entrepreneurs and investors	`"relationships": [{`
	`"is_past": false,`
	`"title": "Founder and CEO, Board Of Directors",`
	`"person": {`
	`"first_name": "Mark",`
	`"last_name": "Zuckerberg",`

The financial information is one of the most relevant one, given that it is very detailed and it includes also the funding rounds. This is particularly an interesting feature to be studies aggregated, for all the competitors.
 	
Knowing who funded similar startups before is important for the entrepreneurs that are searching for contacts to be funded.
	`"total_money_raised": "$2.43B",`
	`"funding_rounds": [{`
	`"id": 2,`
	`"round_code": "angel",`
	`"raised_amount": 500000.0,`
	`"raised_currency_code": "USD",`
	`"funded_year": 2004,`
	`"investments": [{`
	`"person": {`
	`"first_name": "Peter",`
	`"last_name": "Thiel",`

Most of the investors show interest in finding to whom they can sell after the startups that they buy or invest in and therefore this information will be aggregated and displayed as a contact list.


```
acquisition": null,
    "acquisitions": [{
        "price_amount": 100000,
        "acquired_year": 2007,
        "company": {
            "name": "Parakey",
            "permalink":
"parakey"
```


The Crunchbase data is crawled and the different types of data are organized into different tabs.

4.1.2.3. Data Quality

Crunchbase				
Accuracy	**Completeness**	**Consistency**	**Freshness**	**Mentions**
The accuracy is given by the user input and in most than 95% is defining accurate the industry of activity. Most of the time the keywords are either subjective or too wide.	Around 20% have missing funding data.	The tags in Crunchbase are different than the ones in AngelList in approximately 10% of the cases. Small semantic variations occur.	The data is not very fresh, given that a normal user would update a Crunchbase profile only when there are funding or team changes.	Crunchbase is one of the best data quality sources, easy to extract, but the only issues is the small size (300k companies) and the freshness of the data.

4.1.3. Angel List

4.1.3.1. Description

Angel List is one of the biggest databases of start-ups, in Europe. It is known for having more than 1 million start-ups and being one of the most updated data set. Angel List has a complex system of encouraging startups to advertise what they recruit or where they get funding.

The benefits and the problems of using Angel List are summarized below:

Angel List	
Pros	Cons
Biggest data source for start-ups	Legal issues with scrapping their data
High popularity in Europe	
The data is the most frequent updated one	

4.1.3.2. Legal Implications

Angel List imposes certain restrictions that do not allow the use of their data in other applications. A summary of their legal implications is shown below:

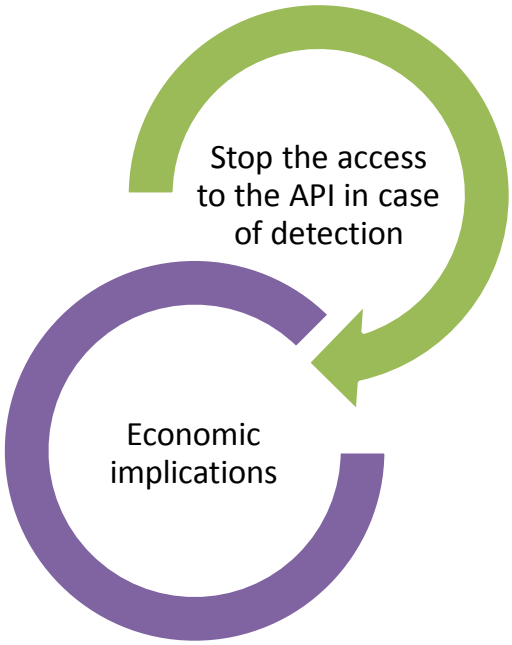
Plain English Terms of Service

Don't store any raw data returned via the API for more than 24 hours.
Don't use the API to scrape data.
Don't use any undocumented endpoints without our explicit permission.
Don't store any of our users' login credentials.
Don't publicly discuss an ongoing private financing. There are federal laws which govern these announcements, commonly referred to as General Solicitation. You can read more about it here.
Do credit AngelList wherever our data is used.
Don't use the AngelList logo or the word "AngelList" as the logo or name for the product you build upon our API.

These terms may change at any time. The last update occurred on January 5, 2012.

We have used the help of a consultancy firm of lawyers, **Avatic Abogados** , and they have evaluated the risks that using Angel Data will imply.

The risks were the following, depending on the amount of norms that were not respected:



The technical decision taken after studying the lawyers report is that we are going to use the open information and find other data sources to complement the missing information.

4.1.3.3. Data Model

One important mention is that the access is not forbidden to all the information they offer and Angel.Co was very cautious with this aspect because the data extracted has the following characteristics:

Information provided	Data
The data that is retrieved in the JSON format is the data on the right, minimal company information and location identification. This data is public and could be used in any way.	`{ id: 19, hidden: false, name: 'Syntyche, Inc.', product_desc: 'myRepLunch.com\r\n1) increases medical staff efficiency so they can help more patients\r\n2) increases sales rep productivity high_concept: ' Vendor-client meeting coordinator', follower_count: 5, company_url: 'http://www.myRepLunch.com', twitter_url: 'http://twitter.com/myRepLunch', markets: [], locations: [{ id: 1694, tag_type: 'LocationTag', name: 'palo alto', display_name: 'Palo Alto',`
This financial data is hidden inside an HTML code that makes the extraction very difficult, as it doesn't follow any patterns. This is most probably a privacy policy of Angel List to prevent this data from being used.	Financial Data
The employee data is as well hidden inside an HTML code that makes it difficult to extract it automatically.	Employee data

The Angel.Co data was used therefore partially, by extracting the JSON with the company information.

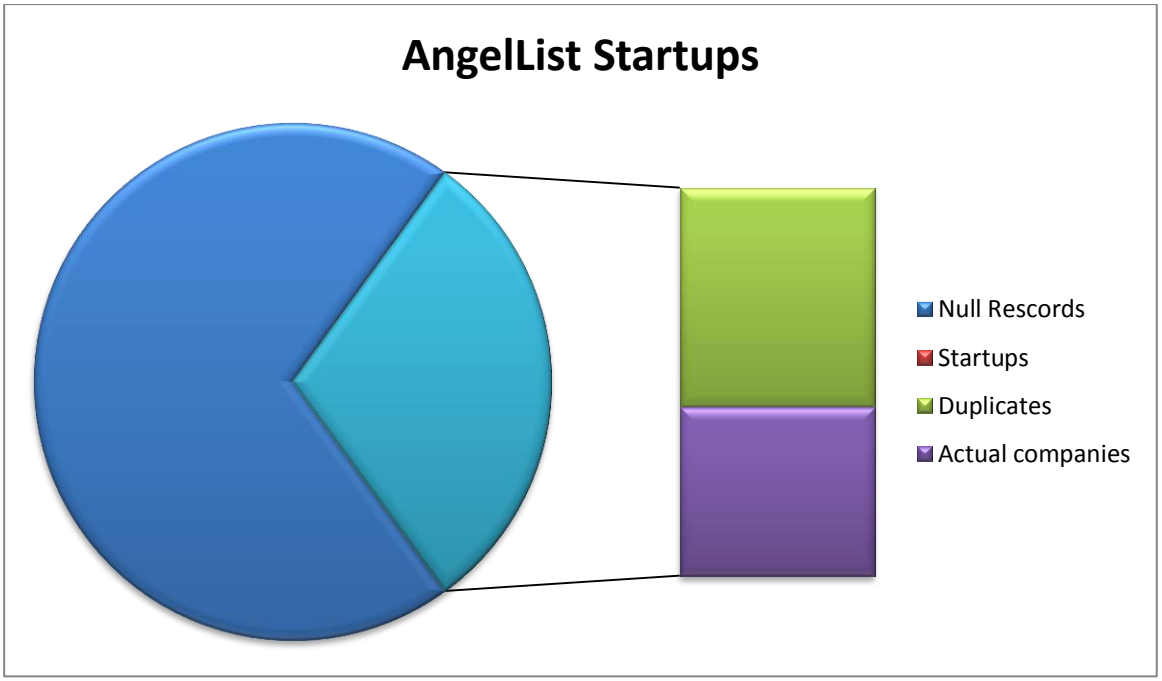
The issues with Angel List are related to the number of records they claim to have. They claim to have 1 million startups but when we extracted the data, the crawler stopped after around than 200k, starting retrieving empty records. The process was started all over again but with no success, because exactly the same thing happened.

What we noticed is that Angel List really has 1 million ids of companies, but out of these 300k start-ups have any sort of other information. Our guess was that Angel List generated all these ids to claim that it has 1 million startups but in reality their number is much lower.

From this point the around 300k of startups were studied in detailed and what we noticed is that many of them were duplicates were exactly the same company, under different ids. Out of them, what we detected was at least 130k , out of 300k that was duplicates. For sure a deeper analysis would reveal more.

Randomly, out of there 130k that we are left with, we check if these are real companies or not and we discover that out of 10 tests, 1 company was a fake one, having absolute no information available in any other source. For this analysis a deeper study needs to be conducted and automatized.

The distribution is shown below:



Angel List was one of the sources that disappointed the most, because they were associated with one of the best data sources for startups.

4.1.3.4. Data Quality

Angel List				
Accuracy	**Completeness**	**Consistency**	**Freshness**	**Mentions**
Angel List is one of the most famous and famous sources. The accuracy of the data is given by the company representative that introduces the data.	The funding information is not available for more than 30% of their startups and the people information is missing for more	There are many fake companies and generated ids, added probably to claim the 1 million	Angel List data is one of the freshest ones, being updated anytime a company	Although Angel List doesn't have 1 million startups as it claims, the freshness and accuracy of the real startups is impressive. The

In more than 95% of companies the tags are accurate.	than 40%.	startups. A good separation needs to be done, between the real and the fake companies.	gets funding or recruits.	data available is also the most updated one existing.

4.1.4. Home

4.1.4.1. Description

Home is a source that extracted the technologies used within a web page and it offers an API to share these technologies. The source is relatively unique in this sector and the quality of the data is very good. For all that, the number of websites they offer is not impressive.

4.1.4.2. Data Model

Home provides an API to extract directly the technologies used on the home page of a website. What Home does is extracting the front code and then detecting all the applications that are used within the web page, for instance: PHP , APACHE etc. Also from the front end code we can extract security seals, emails or even links to their social media.

One of the most significant problems we had was to include the API of **Home**, another web source that offers the technologies used. This source is based also on Alexa and was retrieving all the technologies that were used on a web page, such as PHP, Apache, jQuery and as well the social media profiles and organic keywords.
The problem was caused because Curl (library containing protocols for transferring data) was mostly timed out or was blocked on certain domains that were most probably to contain java script code that was not loading or because of flash programs. This turned to be a time consuming process and the only solution to this was to use **Phantomjs**, a library of Java Script that was able to inject a jquery in a web and to extract content out of it, like for instance an image of the home page, all the technologies used inside of the web page. Another factor that helped was using **Wappalyzer,** a browser extension for detecting technologies used on a website.
Even after involving these technologies we continued having the same problem, that sometimes for some web domains the time was huge and blocked the process. A solution to that was to check if the time to process one domain was too high and then if there appears to be too much, to send a SIGTERM. This proved to be the best strategy to extract the data for all the domains, without blocking the entire process for only some problematic web domains. The cause of the technologies failing to work for some web domains remains a question.

4.1.4.3. Data Quality

Home				
Accuracy	**Completeness**	**Consistency**	**Freshness**	**Mentions**
Given the known problems of this API, many technologies were not extracted. A random test showed that the technologies were retrieved in a percentage of 95%.	All the web domains extracted have at least one technology.	There are around 20 technologies mentioned that do not appear to exist or to be updated.	Due to the issues associated with this data source, the data is not updated on a weekly basis. This is also shown by the names of the technologies that do not exist.	Home is one of the most complicated APIs and therefore the quality of the data might have been distorted during the process. For all that, the quality is enough to be used.

4.1.5. sharedCount

4.1.5.1. Description

SharedCount is one of the sources that are focused on tracking social media, starting from a web domain. They extract the social media links from the webpage and then connect with Twitter, Facebook and LinkedIn and display the number of shares and comments associated with this web domain. Is one of the most famous sources in social media, among many others.

4.1.5.2. Data Model

SharedCount has an API that allows us to extract in a JSON all the information available, starting from a web domain.

An example of the social media data is available below:

```
[StumbleUpon] => 0
        [Reddit] => 0
        [Facebook] => Array
            (
                [commentsbox_count] => 6
                [click_count] => 0
                [total_count] => 1993
                [comment_count] => 82
                [like_count] => 205
                [share_count] => 1706
            )

        [Delicious] => 470
        [GooglePlusOne] => 10161
        [Buzz] => 0
        [Twitter] => 65
        [Diggs] => 0
        [Pinterest] => 154
        [LinkedIn] => 2457
```

As you can observe there are absolutely all the social media channels and their activity is synthetized in this format.

It basically received as input a URL of a company and then extracts all the mentions in social media, associated with the URL, or a very close form of it. It was one of the most simple APIs we worked with, receiving as input a URL and outputting a JSON with all the information.

4.1.5.3. Data Quality

SharedCount				
Accuracy	**Completeness**	**Consistency**	**Freshness**	**Mentions**
The test conducted show that the social media scores are accurate.	More than 20% have at least one social media score missing, showing no social media presence in all networks.	There are no consistency doubts.	The data is updated in real-time.	SharedCount is simply automatized and real time social media monitoring and therefore there are no big issues.

4.1.6. MyWOT

4.1.6.1. Description

WOT (Web of Trust) is one of the most well-known data sources in terms of preventing scams and evaluating reputations of web domains. Based on crowd evaluations, MyWOT managed to collect user ratings on different domains, as well as reputation and confidentiality.

4.1.6.2. Data Model

MyWOT offers an API retrieving scores of trustworthiness, privacy and child reputation, having two coefficients: reputation and confidence for each of the parameters. The extraction of domains was in chunks of 100 domains and the usage of the API was smooth and easy to automatize. The source webpage is : www.mywot.com

An example in the case of Bing.com offers the following result:

```
[trustworthiness] => Array
                (
                    [reputation] => 94
                    [confidence] => 80
                )
            [vendor_reliability] => Array
                (
                    [reputation] => 95
                    [confidence] => 81
                )
            [privacy] => Array
                (
                    [reputation] => 94
                    [confidence] => 80
                )
            [child_safety] => Array
                (
                    [reputation] => 94
                    [confidence] => 80
                )
```

As we can see these 4 coefficients are measured each in terms of reputation (what the user has heard) and in terms of confidence (what the user trusts).

4.1.6.3. Data Quality

MyWOT				
Accuracy	**Completeness**	**Consistency**	**Freshness**	**Mentions**
Accurate, given that the results were validated with the community support	5% of them have only one coefficient	Less than 3% had very big variations in the confidence levels.	The data is updated in real-time.	MyWOT is one of the sources that validate the data, before providing it.

4.1.7. Other data sources

4.1.7.1. LinkedIn

LinkedIn offers an API through which we can extract information on their companies, teams and also recruitment. For commercial use, the API is not free and only some modules can be used free of charge. Other than that, the People Search with LinkedIn was disabling in the last versions of the API, making impossible to access peoples information. LinkedIn is associated with terms of use that are very strict and therefore a preliminary study will be done on their terms and conditions and on their pricing policy.

4.1.7.2. Google Search

Google Search is a paid module of google that automatizes searches of companies on google and automatically storing their data. Google Search will be used to detect the main blog posts about failure and to automatize decision tracking. It is indeed a very expensive source of data.

4.1.7.3. Freebase

Freebase is a source similar to Crunchbase, having all the information linked. For all that, Freebase is not specialized in startups and companies and it also has around 200k of companies, mostly not updated.

4.1.7.4. Similar Sites

Similar sites is an interesting combination between the competitors extracted from Alexa and Google AdWords. It is currently one of the good sources to retrieve the exact competitors. For all that the data set is very small, counting no more than 100k of companies. This is due to the fact that accessing AdWords is very expensive and they haven't invested much in buying data.

4.1.7.5. Phishtank

This is a free data source about the entire phishing website. It returns a CSV file with the entire phishing website. In this way we could aggregate it by domain and see which one of them are phishing websites.

4.1.7.6. Topsy

Topsy is a data source that has extracted for each of the webs domains in Alexa all the tweets available. So instead of using the Twitter API they offer a much comfortable way to use the tweets already extracted. Their data is updated in real time.

4.1.7.7. Semrush

This source contains all the most frequent keywords associated with one domain. Given the poor quality of the data we might assume that these are organic keywords, without a proper treatment. There are problems regarding multi-language keywords and very poor quality of them.

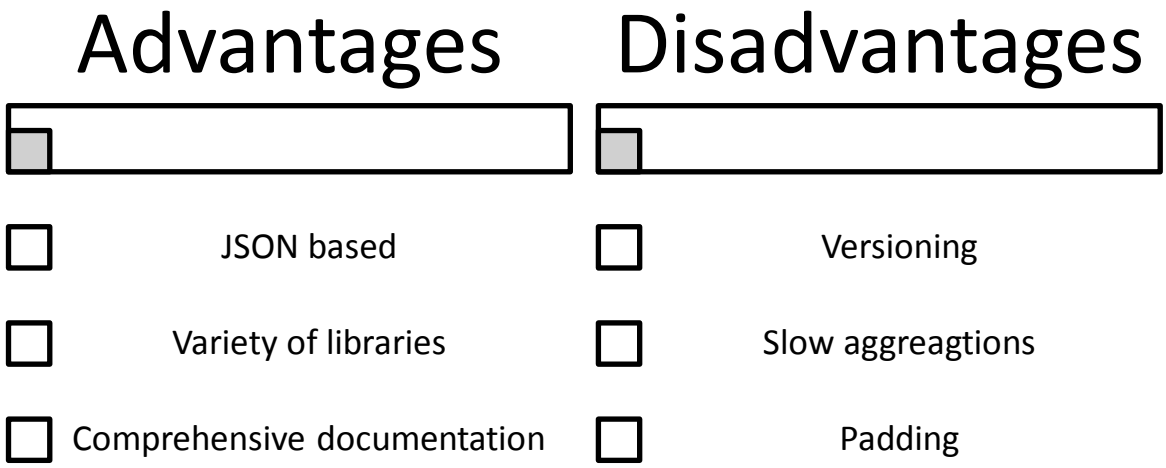
4.2.Integration layer

The integration layer is based on the data sources that have been described before and it considers their semantic heterogeneity. As discussed before, there are possibilities of aggregating the sources based on the company or the sources based on the web domains. The current solutions are being evaluated, as well as the data source to be integrated.

4.2.1. Critical Evaluation of current existing solutions

Each of the data storage methods was studied individually, depending on the constraints at that point in time.
Mongo DB was a natural choice at the beginning of the project , due to the JSON format of all the data extracted.

An analysis of Mongo DB will be presented below:



The main inconvenients of MongoDB that we have seen were the problems with storing different versions, performing very slow aggreagations and also padding scores.

Once these problems of MongoDB were significantly affecting the quality of the results, new options had to be found and therefore Cassandra and Hbase were studied.

At the time of studying the tool, Hbase was recommending in the documentation usage of 3 columns only, although now there is evidence of possibility to use more than 3 columns.

When studying the writing speed to disk, Cassandra proved to be the best option. Even more, Cassandra offered the possibility to add as many columns as possible and given the current architecture, Cassandra was chosen as main database.

Details on each of step of the evaluation process of the tools are given below.

4.2.2. Mongo DB

4.2.2.1. Description

Mongo DB is a document oriented database, considered a NoSQL database, due to the usage of JSON documents. The integration with any application and Mongo DB is natural and there is plenty of documentation its usage.

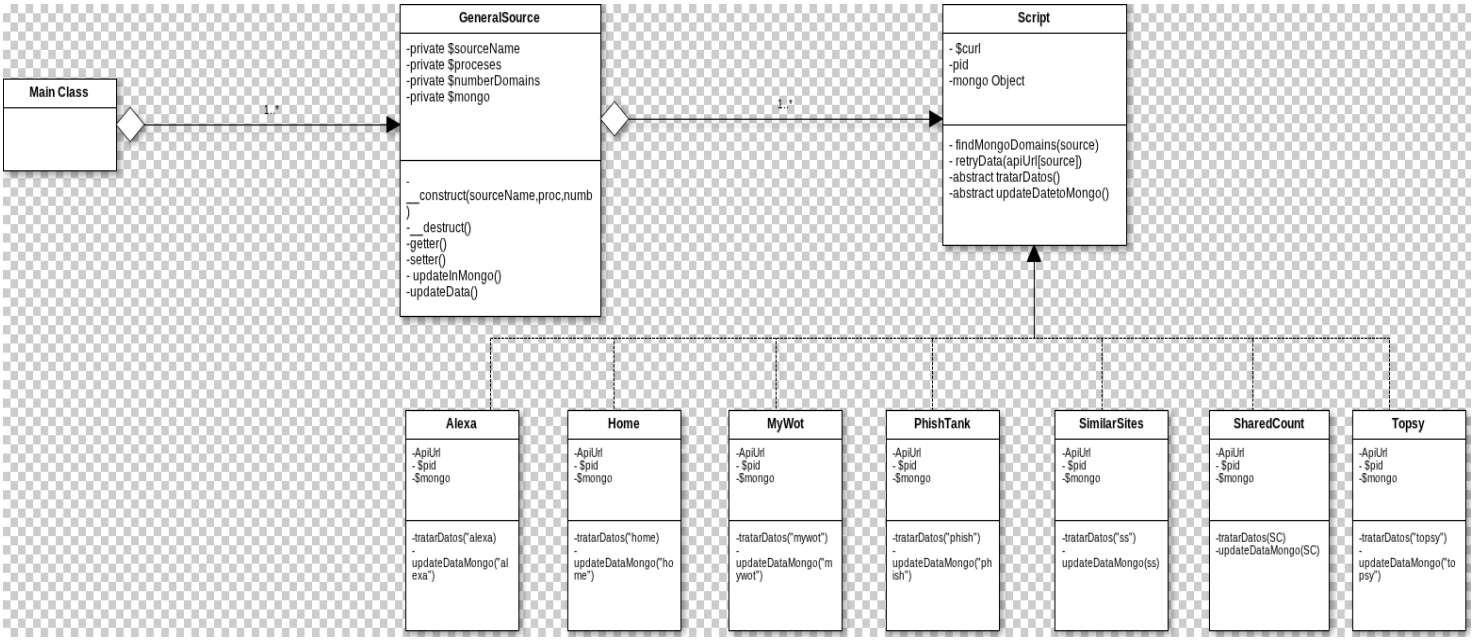
4.2.2.2. Feasibility in DatoSphera

We started using Mongo DB, being at that time one of the most common software to be used for this sort of projects and semi-structured data. Furthermore all the data structures were extracted as JSON documents, making the choice for Mongo DB obvious.

At some point in time what we were writing was far below what we were reading, causing Mongo DB to fail in various occasions, which lead to the decision of changing the storage method. At the beginning of the project, the necessity was basically to read but then it turned into a necessity to write, which caused problems.

Another major problem of Mongo DB was basically the versioning of historical data, because adding a new document with another date was a very simple thing to do but it caused in time a significant slower process. This lead to the necessity to use aggregation frameworks, just to make some very simple queries, only because the historical data for the text format we were storing turned into an impressive value, growing exponential. We decided that given all these problems, Mongo DB was not a very good option. The class diagram of Mongo DB is presented below:

Class Diagram Mongo DB



The best strategy to process so many data sources is the parallel approach instead of one by one approach. We define the main scripts to define the data and we need to define an order for the data source processing. For instance updates in myWot could be processed each week; Phishtank could be updated daily and so on.

This main script is in charge of managing all the data sources extraction, creating some sub scripts that extract from each of the sources, following the pre-defined strategy.

After this being done, each instance of this class will search in Mongo DB data that are not processed (processing=false) and will order it by the most recent data or by null.
Having the domains now will split the number of domains between a numbers of child instances.

This children scripts are in charge of searching information on the given web domains and then store it in Mongo. At the end there are signals send once the tasks of the children is completed and the father will change the value of the variable processing to false, will change the date when the information was updated and will prepare for the next task.

4.2.3. Key-Value stores

Out of the Key-Value stores, Cassandra and HBase were studied in depth for DatoSphera.

Both of them are second generations of the Key-Value stores, with great scalability. Other key-value stores were also considered but not studied in detail, as there was no previous experience in the team working with them. A special attention will be given to the critical comparison between them.

4.2.3.1. Cassandra

Cassandra is an Apache storage based on column indexes, that fully uses the potential of materialized views and caching. It is known for powerful replication techniques across nodes. It is scalable and fast in retrieving results.

Cassandra was used after there were several Big issues with Mongo DB. Versioning and also benchmarking shown that Mongo DB was no longer an ideal option.

4.2.3.2. Hbase

Modelled after Google Big Table, HBase is an Apache project that is developed on the top of the Hadoop Distributed File System.

At the time when HBase was studied the documentation encouraged usage of only 3 columns. Given that there were many data source, this 3 column architecture would have been challenging to implement. Currently there are demonstrations of usage with even more than 3 columns.

4.2.3.3. Critical Evaluation: Cassandra versus HBase

A critical evaluation was performed to decide on a storage method, between Cassandra and HBase. To support the evaluation, technical literature has been used.

The benchmark was done, by studying the evidence on Cassandra and HBase. The starting point was the Technical Bibliography 7 : ***Solving big data challenges for enterprise application performance management - Tilmann Rabl, Sergio Gómez-Villamor, Mohammad Sadoghi, Victor Muntés-Mulero, Hans-Arno Jacobsen.***

The results are shown below. The workload to be studied can be summaries in the following table:

Table 1: Workload specifications

Workload	% Read	% Scans	% Inserts
R	95	0	5
RW	50	0	50
W	1	0	99
RS	47	47	6
RSW	25	25	50

(Extracted from : ***Solving big data challenges for enterprise application performance management - Tilmann Rabl, Sergio Gómez-Villamor, Mohammad Sadoghi, Victor Muntés-Mulero, Hans-Arno Jacobsen***)

Out of these cases, the relevant ones for DatoSphera were R, RW and W. A particular attention was given to the W tasks that had a particular importance.

Given then relevant cases, the paper reveals the following results for the workload R:

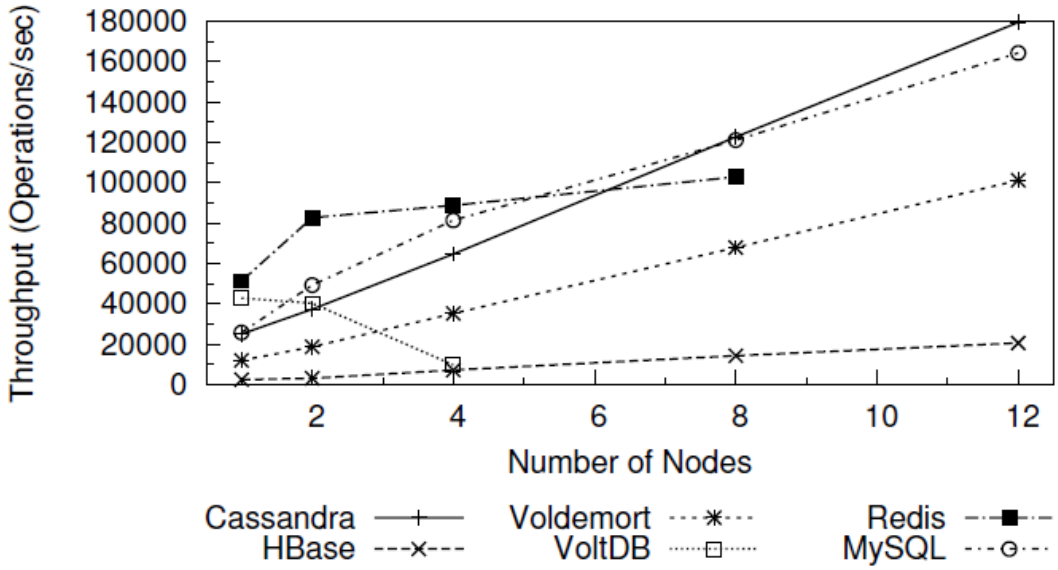


Figure 3: Throughput for Workload R

(Extracted from : ***Solving big data challenges for enterprise application performance management - Tilmann Rabl, Sergio Gómez-Villamor, Mohammad Sadoghi, Victor Muntés-Mulero, Hans-Arno Jacobsen***)

Is very interesting to see that HBase and Cassandra follow a linear exponential model, but for all that the performance of Cassandra is significantly better.

Even for one node Cassandra performs better and the difference increases with the growing number of nodes. For another perspective we observe that Redis performs better than Cassandra, up to 6 nodes, when its performance remains constant.

We now study the latency of the two options, in the R workload:

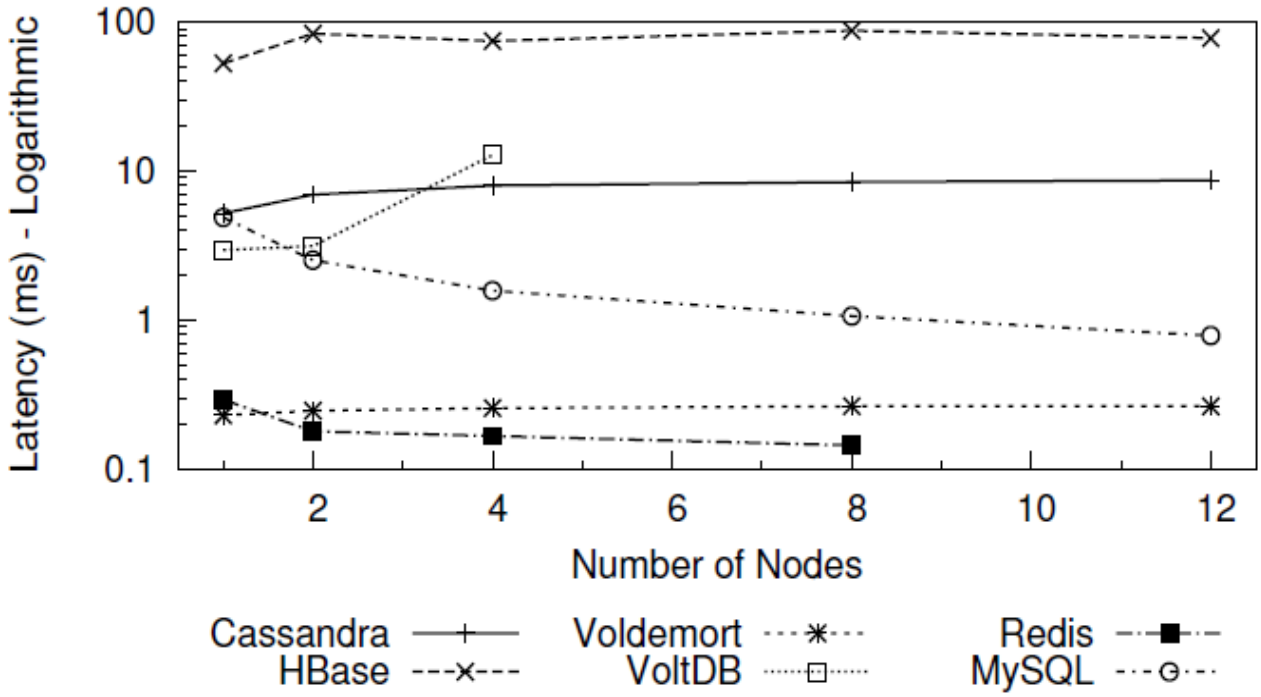


Figure 4: Read latency for Workload R

(Extracted from : ***Solving big data challenges for enterprise application performance management - Tilmann Rabl, Sergio Gómez-Villamor, Mohammad Sadoghi, Victor Muntés-Mulero, Hans-Arno Jacobsen***)

What we notice is that except for the very first node, the latency is constant for the two options. The latency of HBase is significantly higher than the one of Cassandra.

We will now study the scalability in case of a workflow RW and the results are summarized below:

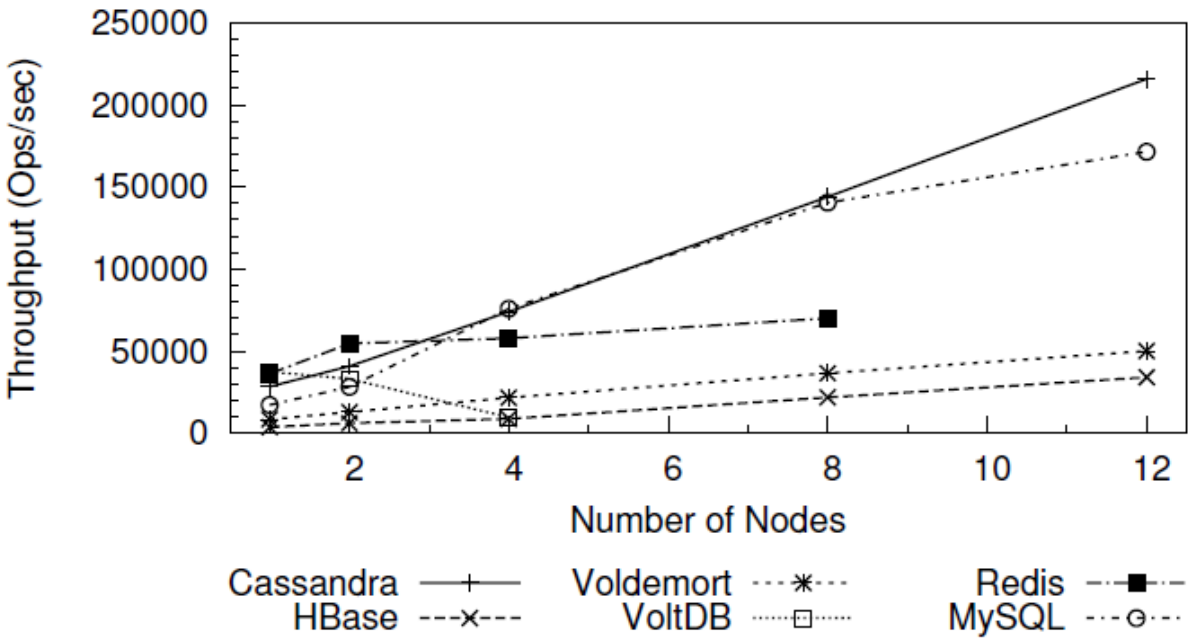


Figure 6: Throughput for Workload RW

(Extracted from : ***Solving big data challenges for enterprise application performance management - Tilmann Rabl, Sergio Gómez-Villamor, Mohammad Sadoghi, Victor Muntés-Mulero, Hans-Arno Jacobsen***)

This states a clear difference between Cassandra and HBase, Cassandra increasing exponentially the number of operations, while Hbase increases with small variations.

The only option that for up to 2 nodes performs better is Redis, but for more than 2 nodes Cassandra proves to be way more scalable.

When we check the latency, the results are the following:

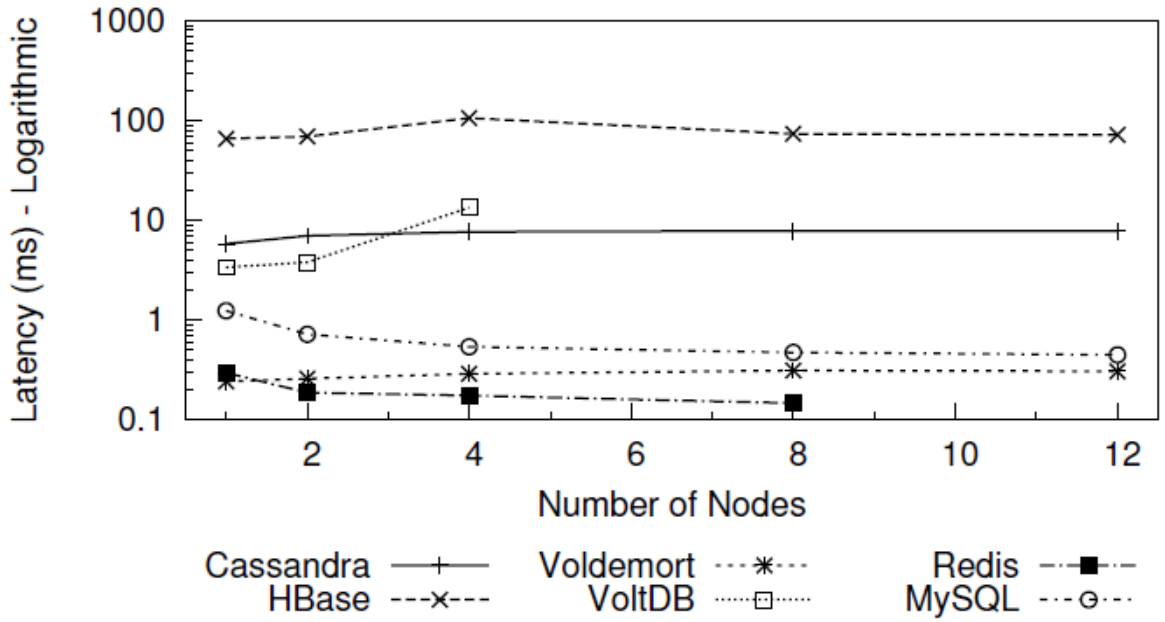


Figure 7: Read latency for Workload RW

(Extracted from : ***Solving big data challenges for enterprise application performance management - Tilmann Rabl, Sergio Gómez-Villamor, Mohammad Sadoghi, Victor Muntés-Mulero, Hans-Arno Jacobsen***)

We observe again that HBase has the worst performance in terms of latency, but Cassandra is not very far from this performance.

They both follow a linear evolution, constant regardless the number of nodes. For all that, the other options performs much better than Cassandra and HBase.

The remaining workflow to check is the write one, which is summarized below:

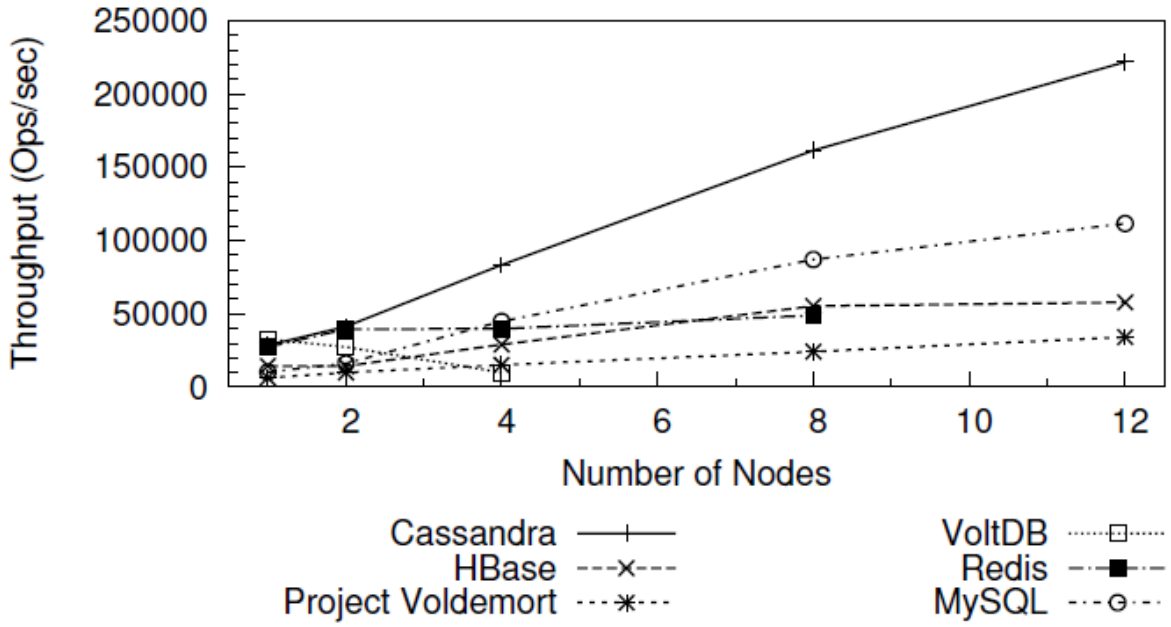


Figure 9: Throughput for Workload W

(Extracted from : ***Solving big data challenges for enterprise application performance management - Tilmann Rabl, Sergio Gómez-Villamor, Mohammad Sadoghi, Victor Muntés-Mulero, Hans-Arno Jacobsen***)

We observe that even in this case Cassandra proves to be the best option in terms of latency, even when the number of nodes grows.

Up to 2 nodes, HBase has a comparable performance with Cassandra, but after that HBase performance is linear, while Cassandra grows again exponentially, differentiating from all the other options.

We study now the latency of the workload W:

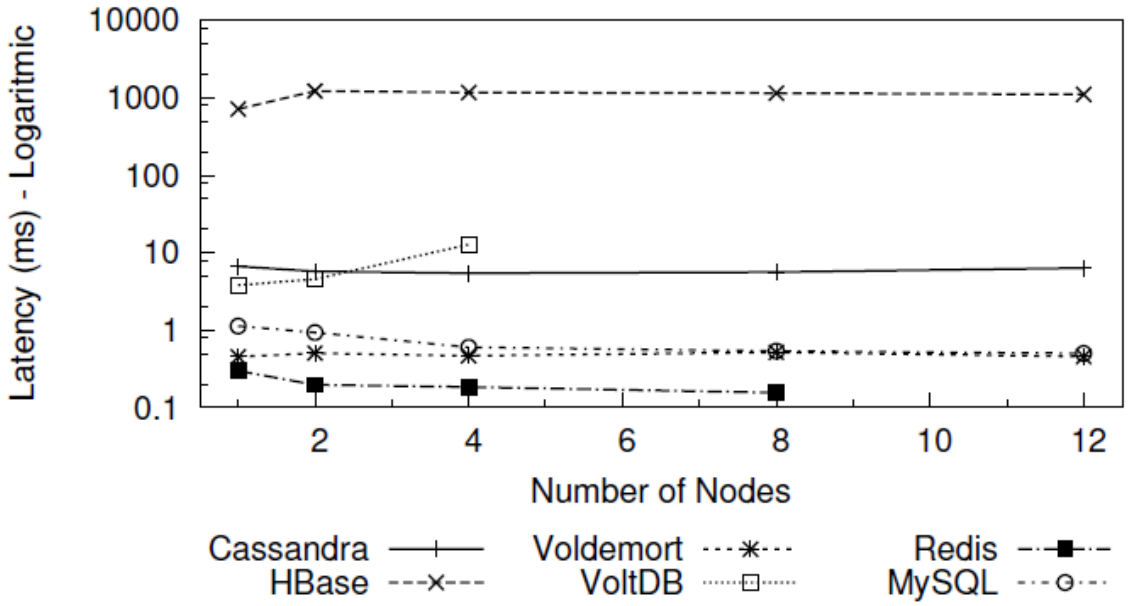


Figure 10: Read latency for Workload W

(Extracted from : ***Solving big data challenges for enterprise application performance management - Tilmann Rabl, Sergio Gómez-Villamor, Mohammad Sadoghi, Victor Muntés-Mulero, Hans-Arno Jacobsen***)

Exactly the same latency as in the RW workflow is represented. HBase has a big latency, compared to Cassandra, even if the number of nodes is low. An interesting mention is that the latency in case of HBase starts growing from 2 nodes and then remains constant, but the opposite situation happens for Cassandra, that after the first node starts slowly decreasing the latency, but finally maintains constant, regardless of the nodes.

Now considering the benchmark analysis, Cassandra was chosen as optimal solution. We can observe that Cassandra allows replicas across multiple machines. The literarily increase in performance once new machines are added was also a criteria to be considered.

The fact that adding different machines was not causing downtime or interruptions in access and also synchronous and asynchronous replication, were important factors to be analyzed.

A part of the decision is based on the fact that at the point in time when the decision was made, HBase was recommending in its documentation to use only 3 columns, which would have been impossible for DatoSphera. Current studies show that the limitation to the 3 columns is no longer valid. For all that the benchmark states a difference between Cassandra and HBase performance.

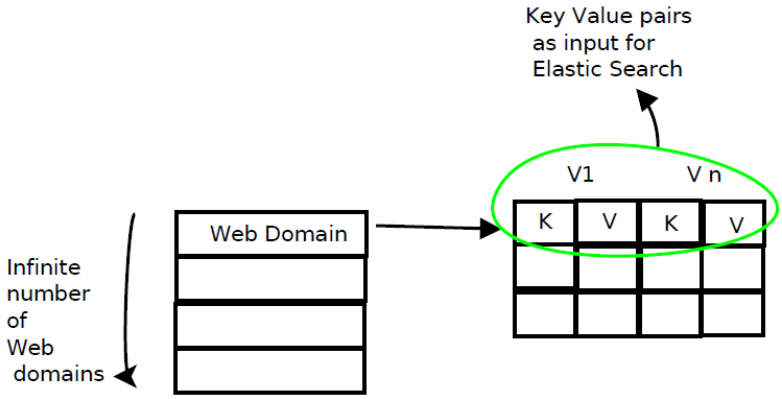
4.2.3.4. Description of the solution

A combination of Cassandra with Elastic Search has been chosen as optimal. Cassandra was chosen as a result of all the discussion presented in the previous section and Elastic Search (Search Server based on Lucene) was chosen to complement the speed of Cassandra.

4.2.3.4.1. Cassandra

The current solution is based on Cassandra that proved to be the best option in terms of workload, but also regarding versioning.

A summary of the Cassandra column representation is shown below:



Each domain is stored as a record and after that the key value structure is followed, for each version of the domain.

The key value pairs are defined as follows:

$$\begin{cases} k \rightarrow X \\ Value \rightarrow JSON\ Document \end{cases}$$

The column schema for Cassandra is the following:

SIMPLE	COMPLEX	COMPLEX	COMPLEX	SIMPLE	SIMPLE	SIMPLE	SIMPLE
CRUNCH	**CRUNCH**	**CRUNCH**	**CRUNCH**	**ALEXA**	**WOT**	**SHARED COUNT**	**HOME**
company_info	**Fundinground**	**Investment**	**acquisition**	**seo_info**	**wot_info**	**socialNetwork_info**	**home_info**
Name	source_description	company_permalink	company_permalink	rank	trustworthinessReputation	Facebook_likes	detectedApp_appName[0..X]
Permalink	round_code	company_name	company_name	countryRank	trustworthinessConfidence	Facebook_shares	detectedApp_networkId_[0..X]
crunchbase_url	source_url	round_code	price_amount	stadisticsBounceRate	vendor_reliabilityReputation	Twitter	detectedApp_domain_[0..X]
homepage_url	raised_amount	source_url	price_currency_code	stadisticsDailyPageViews	vendor_reliabilityConfidence	Delicious	detectedApp_type_[0..X]
blog_url	raised_currency_code	source_description	term_code	stadisticsDailyTimeSites	privacyReputation	Pinterest	keyword_[0..N]
blog_feed_url	funded_date	raised_amount	source_url	stadisticsSearchVisits	privacyConfidence	etc	html
twitter_user	investments_compa	raised_curr	source_de	demographics[0..	child_safetyRe		title

name	ny[0..N]_name	ency_code	scription	X]Country	putation		
category_code	investments_company[0..N]_permalink	funded_date	acquired_date	demographics[0..X]PercentVisitors	child_safetyConfidence		description
number_of_employees	investments_financial_org[0..M]_name			demographics[0..X]Rank			mail

The key value pair is the web domain, considering all the information existent. The columns are represented above and measured for each of the key values. The families of columns are based on the simple and complex types, being grouped in these categories.

4.2.3.4.2. Elastic Search

4.2.3.4.2.1. Description

Elastic Search is a module based on Lucene, which provides distributed full text search, very fast and scalable. It can be used through a RESTful web interface and it can be personalized for the needs of the project.

4.2.3.4.2.2. Feasibility in DatoSphera

Combining Cassandra with Elastic Search is motivated by the fact that storing directly in Elastic Search might give problems of accessing the data, in case one of the Elastic Search indexes is deleted. Therefore it offers stability.

A major issue of Elastic Search is the scores that it creates, based on word frequency. In DatoSphera the scores were modified because term frequency was not relevant for the results.

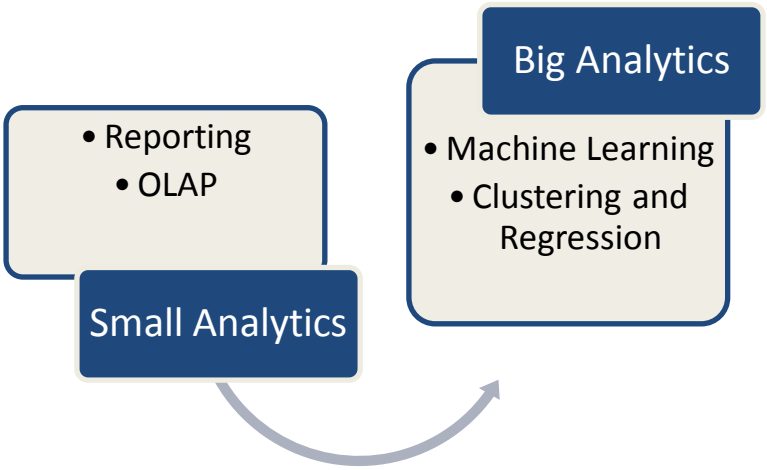
The input for Elastic Search was the Version (Key, Value) pair and therefore the searches were very fast.

4.3.Exploration

Given that DatoSphera is at the beginning, in the phase of exploration of data source, the time dedicated to analytics was limited to basic aggregations and illustrations of graphics.

For a better understanding of the exploration level the following source has been used: Michael Stonebraker, professor at the Massachusetts Institute of Technology- http://cacm.acm.org/ - “What does Big Data mean”.

The source explains that currently there are analytics that can be grouped into Small and Big Analytics, applied on Big Data. The interesting concept introduced is that every company has to pass from Small Analytics to Big Analytics, process during which the data is being explored and prepared for the Big Analytics.

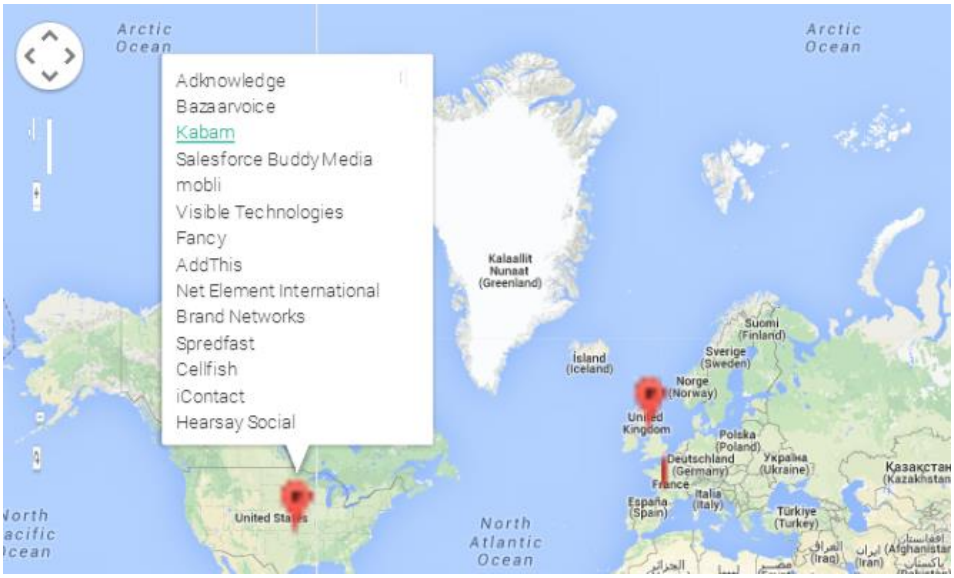


Small analytics were necessary to better understand the data and then based on the small analytics, generate big analytics, such as pattern or sequence mining.

4.3.1. Small analytics in DatoSphera

- Reporting

Basic reporting tools have been used to generate aggregated graphs and to display them in a pdf report. Some samples of these reports are shown below:



These aggregations are performed by countries and from groups of countries; the user can go into details on each of the records. Another performed aggregation is displayed for competitors' number, funding or acquisitions:

Evolutions with minimum and maximum are displayed per years, as well as aggregations of funding and acquisitions.

- OLAP

OLAP cubes will be used to display general statistics about funding's per sector of activity and also numbers of competitors. Currently no OLAP instrument is being used.

4.3.2. Big Analytics in DatoSphera

- Machine Learning

Big Analytics will be implemented to evaluate the real competitors, with the user input. Collective inputs about competitors will reveal the real competitors for each company. The user will have the possibility to select from a list the real competitors and to see data aggregated only for them. This will allow DatoSphera to significantly improve the keywords and to retrieve the real competitors. Once the platform is being used and the volume of business ideas tested is high, Big data techniques will be applied to predict the possibility of failure of startups.

Pattern and sequence mining will be combined with time series to predict the probability that a startup will fail, based on historical evidence. The algorithm will be defined of the past data of the startups and their closing dates.

- Clustering and Regression

The potential in text mining will be explored with clusters of similar tags and associations between tags and industries. Recommendations will be performed inside of the clusters and regression models will be defined.

4.4.Project challenges and solutions

As a result of the extraction, integration and exploration of the data source the following problems were identified:

The Source Home to extract the web technologies was raising significant problems because even using an alternative solution with Wappalyzer and Phantlyzer was very time consuming and the probability of failure was very big.

- **Solution**: The probability of failure could be minimized , by using 10 proxies per source and only 10 domains per iteration and this would mean 1 domain per proxy . Even though the probability of failure of the source Home was a great risk.

Alexa, Topsy and Shared Count have the same execution time of approximately 1 week each, extracting 1 million domains. This is directly correlated with the number of processes that are at a given time executed by the source.

- **Solution**: We have upgraded the machine and execute 10 processes per source and therefore the time was reduced with 150%. The maximum achieved was 4 proxies per source.

FlipTop and Crunchbase were used to extract the email addresses of the main persons in the companies. There were several problems in generating all the possibilities of email addresses.

- **Solution**: The time was depending on the Crunchbase API and how fast could call, insert and update in Mongo and then send to Fliptop to be processed.

SimilarSites has a high dependency on Alexa, because is basically a source that extracts the competitors from Alexa. There were certain errors done while this source is extracting the similar sites from Alexa and therefore a double check were needed.

- **Solution**: check the HTML code extracted from Alexa with the information provided by Similar Sites. The only issue is that Similar Sites API limits us to 2500 calls per day, turning into a time-consuming task.

Phishtank was a very simple tool to use, providing a CSV with around 10000 phishing websites that was directly updated in Mongo DB.

- **Solution**: Given that the number of phishing sites is very less, search for alternative sources.

WOT was one of the fastest data source but it covers very few domains.

- **Solution**: Their API allowed up to 1000 web domains per chunk of data in only one process and this turned to be a thousand times faster than all the other API. that were extracting only one domain.

Crunchbase and Alexa shared only 15% of records

- **Solution:**Given the good data quality of Crunchbase, start aggregating from Crunchbase instead of Alexa

Alexa was based on the web domain, while Crunchbase on companies

- **Solution:**Find web sources that mention what is the website for a particular company and doduble check

AngelList is protected legally against use

- **Solution:** Evaluate risks and store only the free information

Beside these technical issues with the data, there were issues with the latency as well.

The latency started becoming an issue just in the moment when we started working with Elastic search, because of the types of queries that were posed and were not efficiently optimized. Once Elastic Search was fully understood and the locations were introduced, the process become smooth and fast.

On the other side, testing the prototype with users revealed an interesting situation: they tend not to believe the results, if the results come too fast. This determined us to introduce a spinner, to create an impression of latency, although this doesn't properly exist. The user feedback was much better. Therefore we let 9 seconds for the spinner and then we display the results.

Other than latency the following types of technical risks are the most common ones to occur, and the planned solutions are explained below:

Risk of data dependency

- This is probably the most common to occur, given that the data comes from external data sources
- This risk could be prevented by having partnership with Crunchbase or AngelList, or even storing the last version of their data

Risk of data quality

- Most users are worried about how we measure the quality of the results and are worried that the tags are not correct
- This risk was already partially adressed, by strategically displaying the links of the competitors and letting the user the possibility to chekc it himself

Legal Data Risks

- As explained before, the data from AngelList comes with several restrictions, that can modify anytime and that can affect DatoSphera legally and can have economic implications, if the legal terms are not studied regularily.
- A solution to the legal risks is having a partnership with AngelList, but this implies huge resources that currently DatoSphera canot afford.

4.4.1. Future technical improvements

The following data sources could be involved in the future:

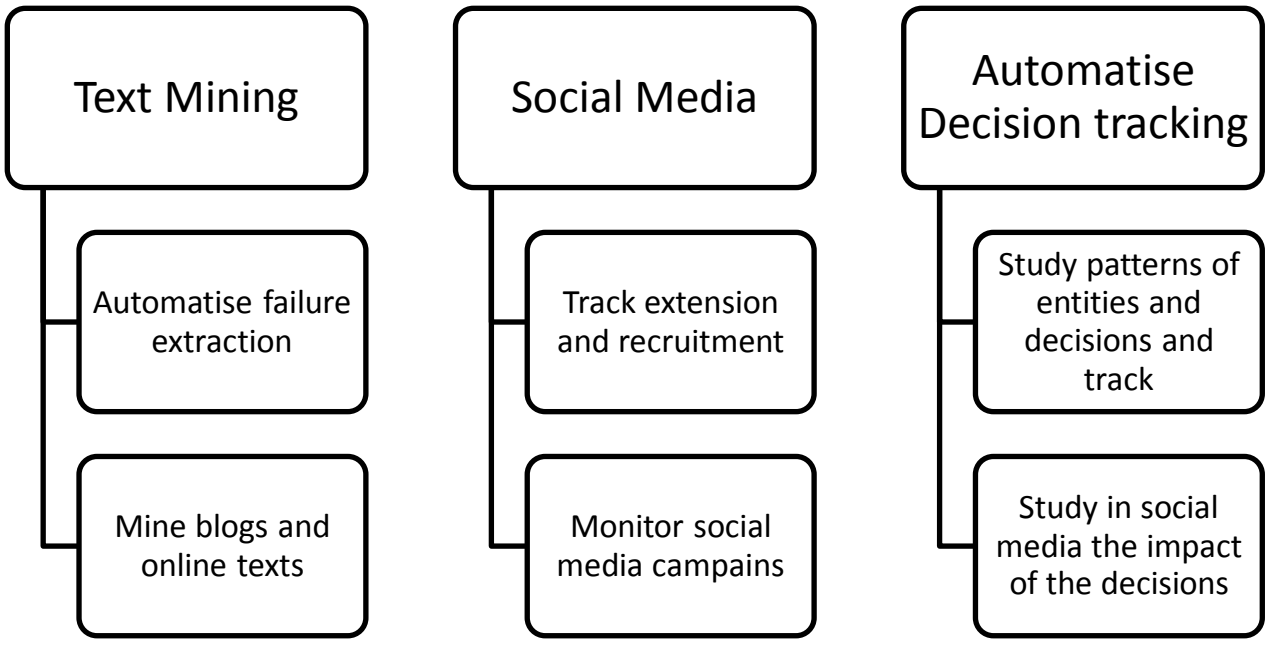
Sources:
MyWOT
SharedCount
Freebase

Having all the experience from the project Trakty that started from a poor quality data source and aggregated all the other data sources, aggregating future data source will studied in detail, considering the future prospects.

An interesting application is the full integration with LinkedIn that will be explained below:

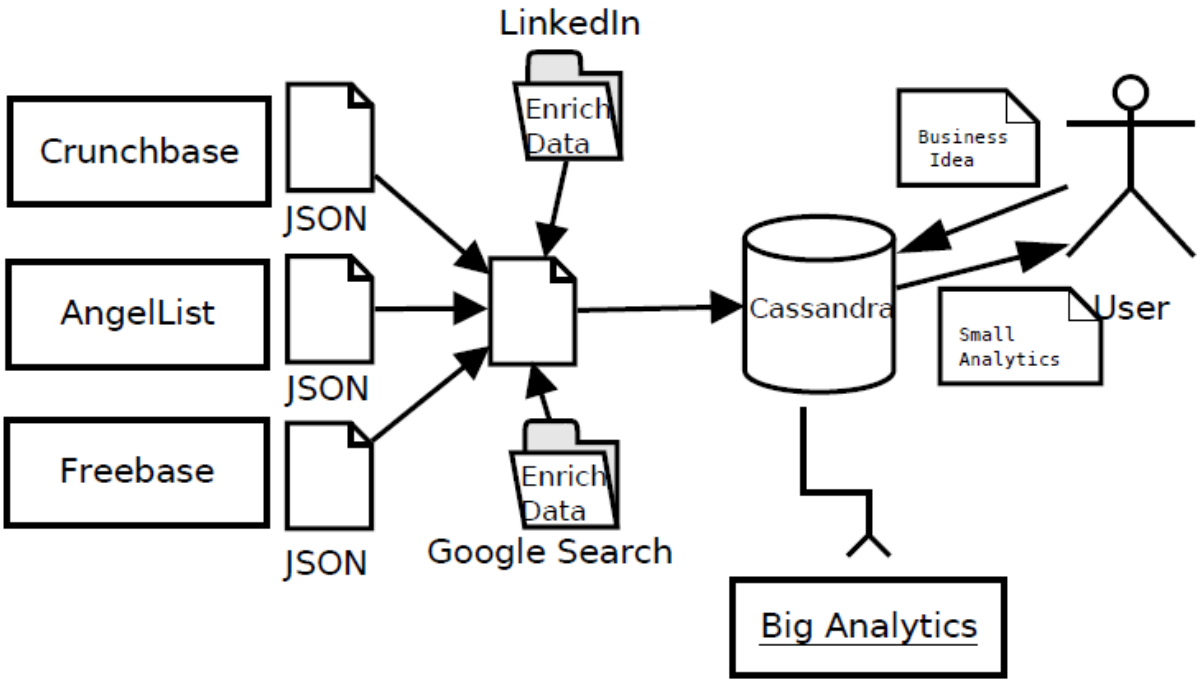
LinkedIn
Detect Extension of companies
Check connections between users and competitors
Aggregate the Skills of competitors
Study what competitors recruit

Extension plans might include the following modules:



4.4.2. Desired data architecture

The desired data architecture is the following one:



Good quality data sources will be aggregated, all of them based on the companies and not on the web domains. This data will be enriched with paid data sources such as LinkedIn and Google Search and the merged data will be uploaded in Cassandra. From this step, Big analytics are performed and small analytics are send as result to the queries of the user.

The stages of extraction, integration and exploitation are all illustrated in the figure.

4.4.3. Conclusions

The exploration part of the data sources was very important and revealed issues associated with the data sources. Although the data sources were not aggregated correctly from the beginning, the experiences gained are very important for future extensions.

Sources like Angel List or Crunchbase, based on the companies, have proved to be the best options to accomplish the business goals. Modules involving LinkedIn are also considered, to complement the data.

Although for moment are implemented just small analytics, big analytics and machine learning will be key parts of DatoSphera.

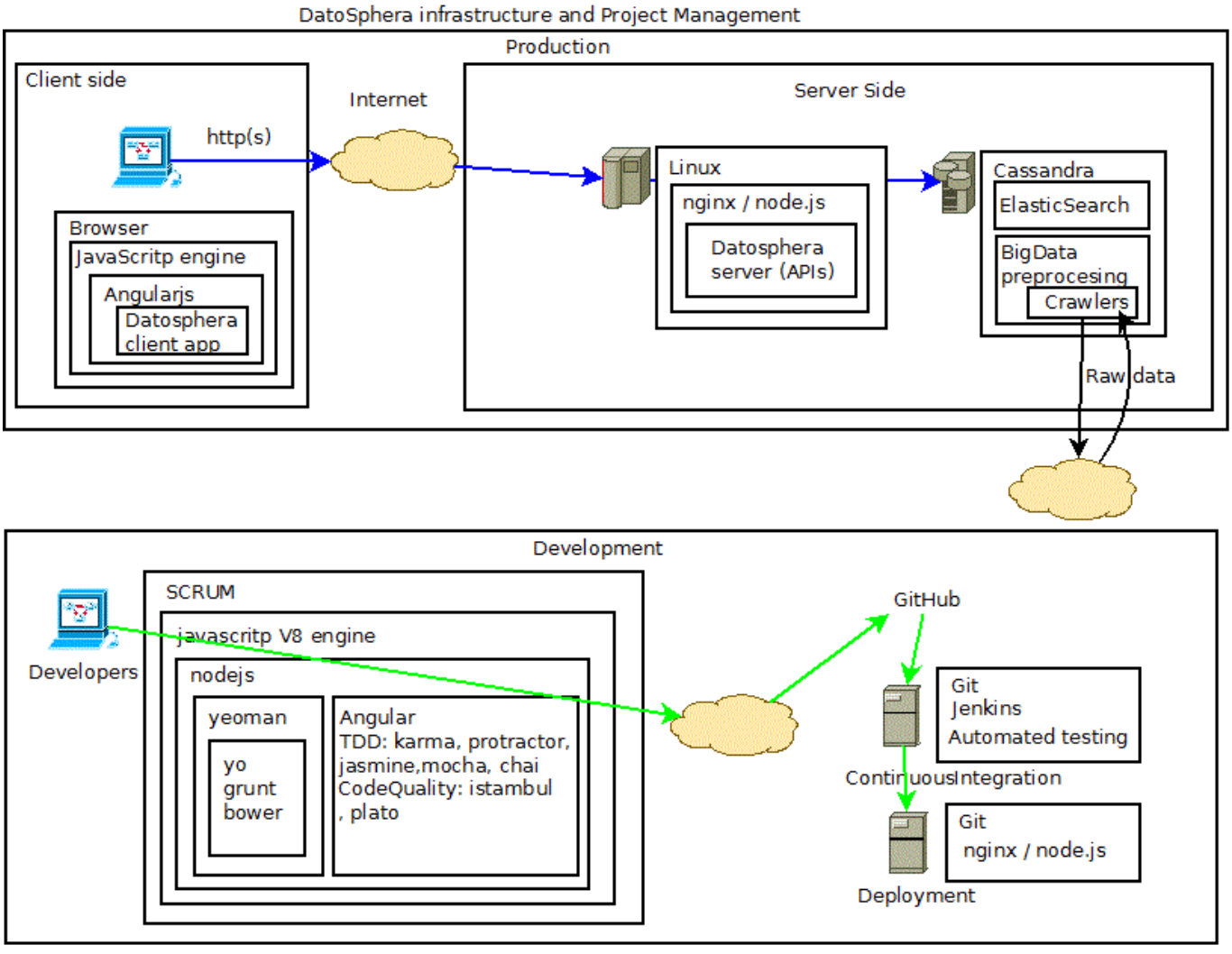
5.Current solution

DatoSphera is a technical research project, which has business and project management constraints. Therefore the current solution is an illustration of all these constrains on the technical side.

DatoSphera is currently using Crunchbase and is aggregating Angel List, Freebase, LinkedIn and Google Search modules.

5.1.Functional Architecture

The following schema shows the current functional architecture.



As you can see the client sides' core is Angular and Java Script and the backend is not very complex for the moment. Minimum processing is done in this very moment but for future extensions, a separate module to pre-process the data will be necessary.

In the development process TDD methods have been followed, and the entire process from the developer's side to deployment is being explained.

5.2.Tools used

The most important decision in this phase was to decide on the language to develop the application. The alternatives along with the pros and cons are presented below:

Language	Pros	Cons
Java Script	- **Previous experience inside the team** - **Fast and optimal for web development** - **The developing costs of personnel are reasonable in Barcelona**	- **Not the best technology to work in data manipulation**
Python	- Ideal in working with data sources and manipulating them	- Expensive in Barcelona and difficult to scale the team
PHP	- Currently the cheapest language to program in Barcelona	- Very low performance in working with data and in speed

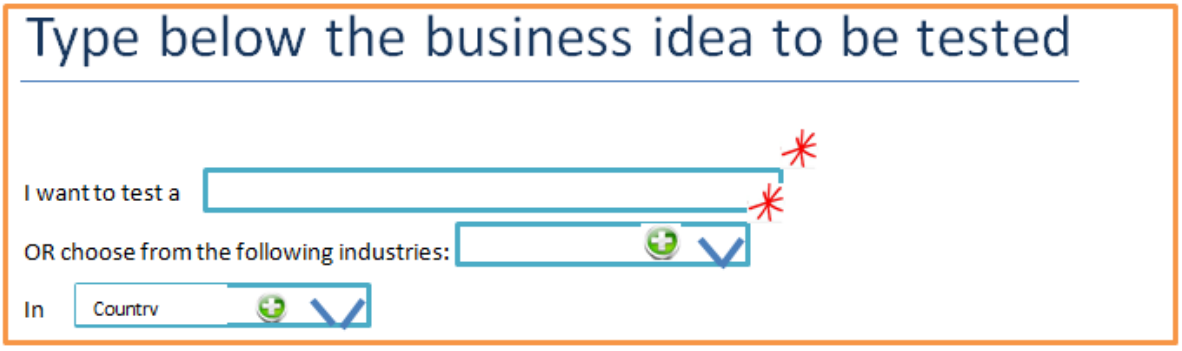
Considering all the options, Java Script was chosen as the best alternative (costs of developing in Barcelona), although will have to change to a combination with Python, for data manipulation. Due to budget restrictions and to the skills of the current team, Java Script was chosen as the best option.

The application is realized in Java Script, using Angular JS as a basis. These technologies proved to be some of the fastest and optimal ones. At the beginning all the application was done just with a client side, followed by introducing the server side very late, when was unavoidable.

5.3.Use Case

The interface developed was only supporting the business decisions taken and representing the data from Crunchbase.

The initial schema to display the Crunchbase data aggregated from the user query is the following:



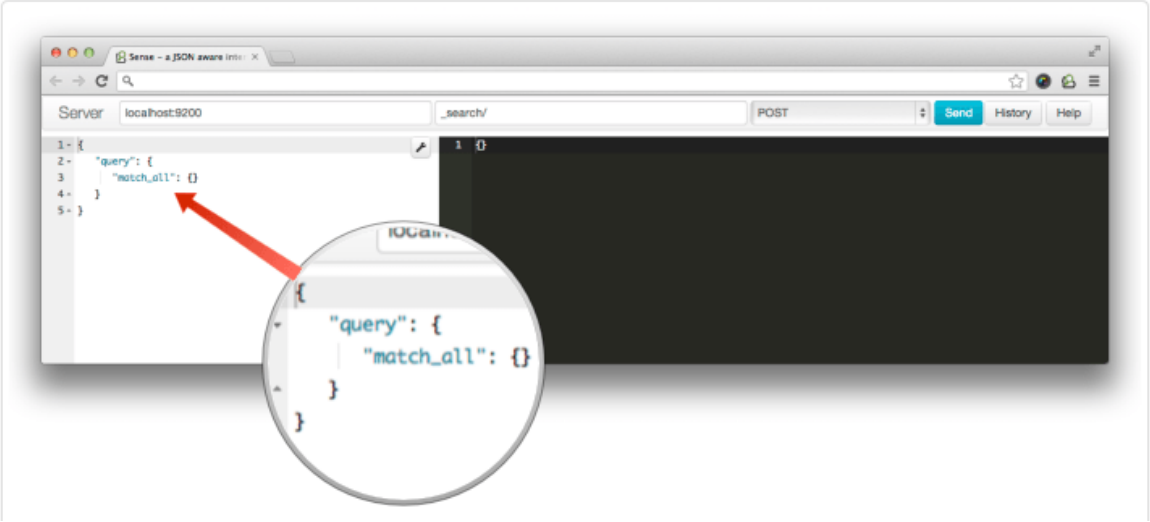
The user could type:

I want to start a business in market niche identification for investments IN Spain

We take each of these words and we see if it appears in the tag list.

In the Crunchbase data set we have the tags, and we try to match each of the words in the sentence.

These are now the inputs for the elastic search (“market”, “niche”, ”investment”, “Spain”)



Now when the Elastic Search score is 1 , this is a direct competitor. It means it has all the characteristics (Market AND niche AND investment). [The country serves just to show personalized per country. If we narrow by country might not retrieve any direct competitors at all.]

What we have so far is a combinatorial problem of all the tags.

Possible alternatives:

Indirect Competitors

Direct Competitor

Market	Company 1, Company 2, Company 3
Niche	Company 1, Company 2
Investment	Company 2
Market Niche	Company 1
Niche Investment	Company 2
Market Investment	Company 2
Market Niche Investment	Company 2 [Elastic score = 1]

Now we have:

Direct Competitors	Company 2
Indirect competitors	Company 1 and Company 3

From this point the aggregated data will be displayed in the following form:

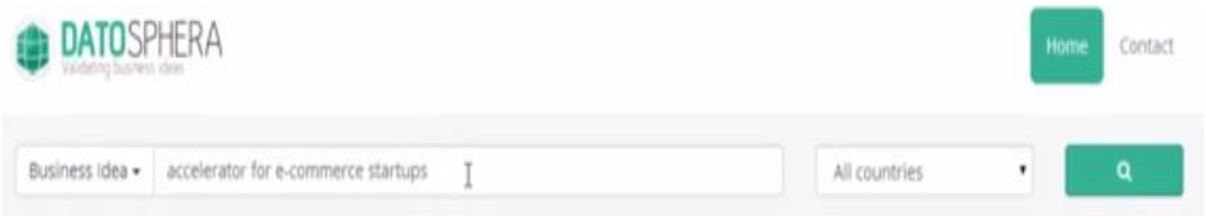
Companies doing exactly the business idea (Direct Competitors)

Companies doing something similar (Indirect Competitors)

What	Source
How many companies already failed in this and when they closed?	**Crunchbase – Company- Status =Closed** **Display the name and year they got closed**
Where are your competitors based?	**Crunchbase – Company-Country** 

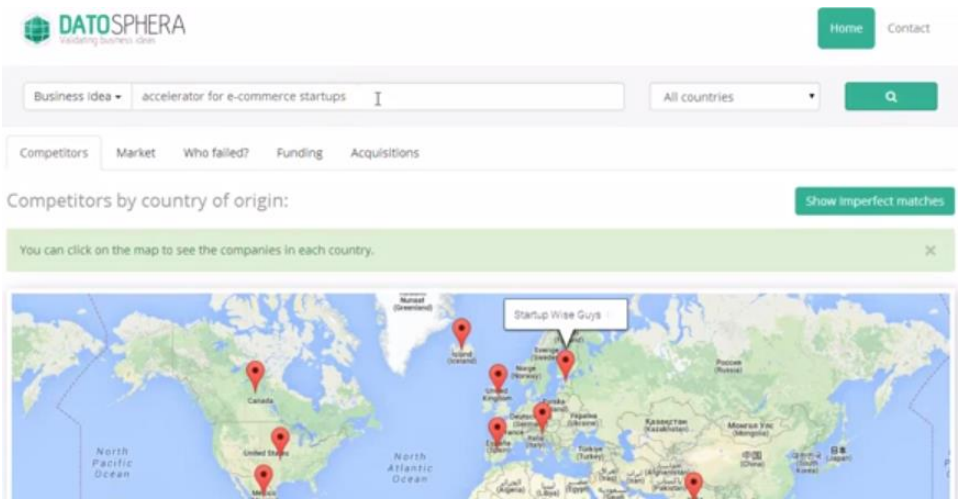
Age of competitors	**Crunchbase – Company**-founded_year
How much funding they usually get?	**Crunchbase – Company-funding_total_usd**
Who invests in them?	**Crunchbase- Investments-**investor_name Display them on the map
Who buys them, and for how much?	**Crunchbase-Acquisitions-A**cquirer_name Display on map

Due to the fact that users do the market research using Google, which has a common interface, we are going to replicate the interface of google, specialized for an entrepreneurs and investors product.



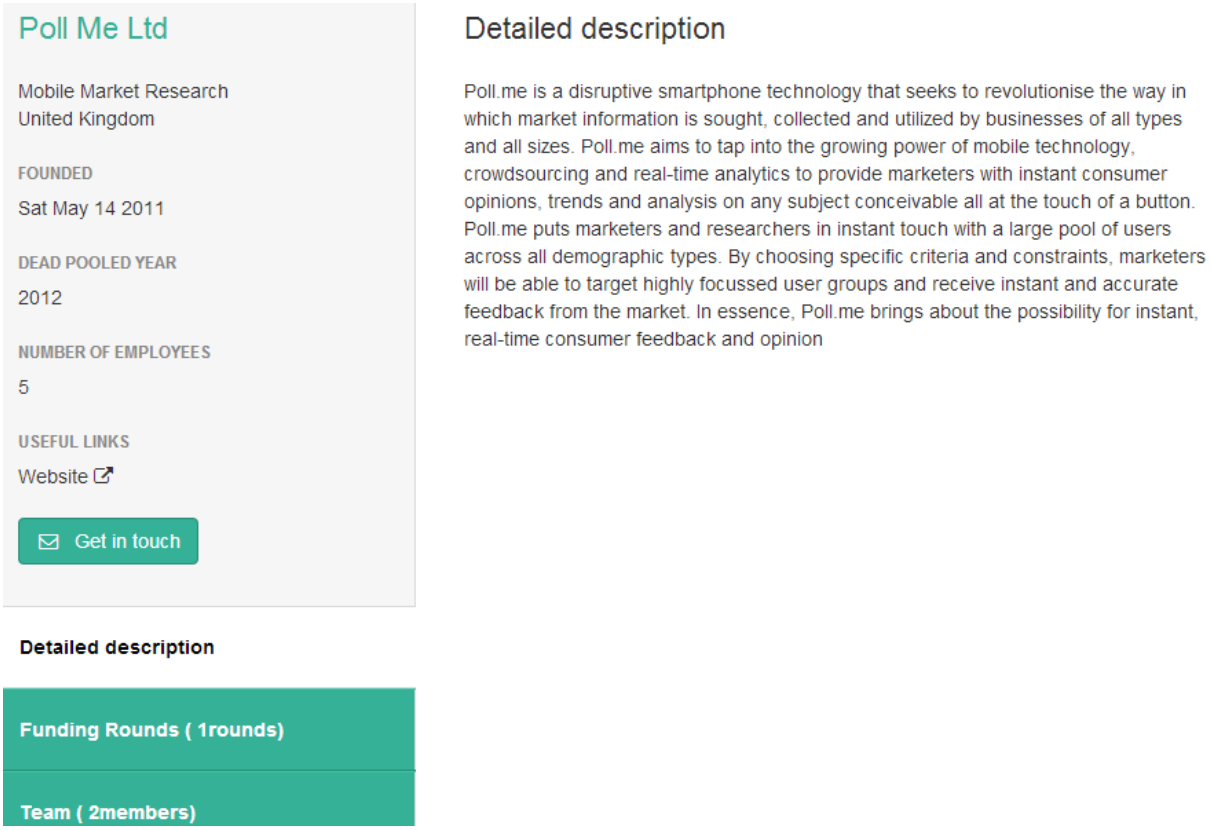
The only module that was incorporated is an extension of Elastic Search for natural language processing and lemma extraction. Name Entity Recognition is another external module that was incorporated, when we noticed that the users tend to type locations in the search box, instead of separating them.

After the user types the idea, just like the Google interface we maintain the idea on the top of the page, displaying the results. This is done in case the user wants to modify his/her search and see what the results are. This will encourage the users to use more than once the platform.



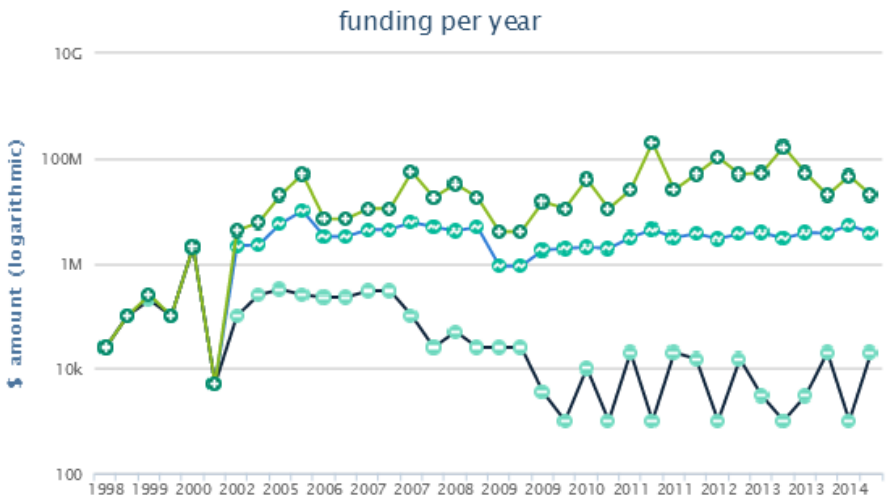
Given that the users want to click on the map and to have everything linked, a way to accomplish this was to display on the maps the links to the company information and jump from one part of the webpage to other.

The user can now click on each of the competitors and find out more information on them:

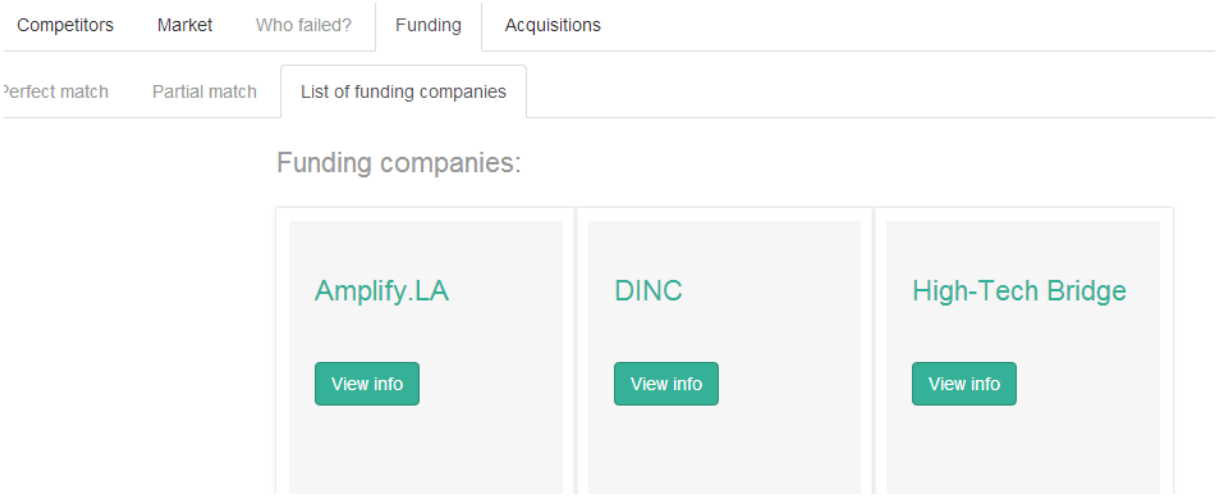


The user can now select the description, the team, the funding, the news on the company and also contacts of the company. This will help the entrepreneur study the competitors, aggregated but also one by one.

From this step users need to see the aggregated data of their competitors and therefore this data will be displayed with charts, such as:



We also display the companies that funded similar projects :



We could also check the market potential, by displaying the trend of competitor's number, per years:

We display the average, minimum and maximum, depending on the corresponding years. The user can click on the graphs on the different points and the corresponding companies will be displayed.

As the user stories show, the linkable content was vital for the users and therefore links were placed in all the places.

5.3.1. Strategic decisions

The technical decision were directly related to the product development are directly related the feedback received from the customer, presented in the product development part and to the evaluation of work amount necessary, estimated in the project management part.

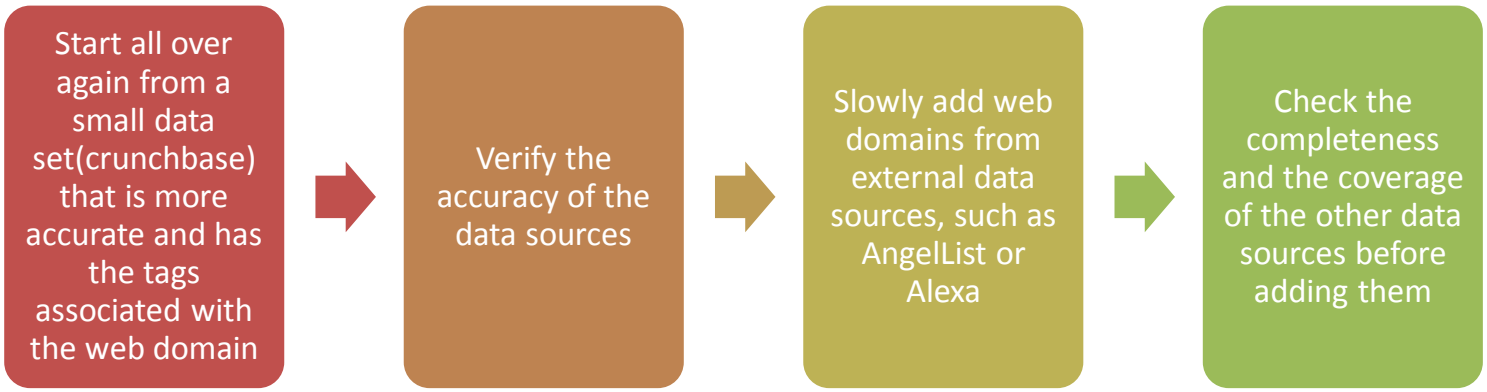
The constrains were the following:

- The web domains will be retrieved by tags , which were good quality in Crunchbase and of bad quality in Alexa
- The Crunchbase data was restricted to 15% of the total Alexa number of web domains

The evaluation of the problem turned into the following conclusions:

- Starting with Alexa as a data source was a problem because even though it contains information on more than 1 million web domains, these domains are unable to be retrieved by tags of a very poor quality
- The usage of Crunchbase would have been mandatory, being the only data source with a very good data quality and tags associated
- A previous analysis of the coverage and missing data would have saved much time
- The different sources of data added are not covering much of the web domains available in Alexa

The decisions taken were the following:



6.Conclusion and future work

Given that DatoSphera is a real startup dealing with project management and business constraints, which affect the technical development we have learned that all the three parts should be treated as a whole system and not as isolated ones. Tending to ignore either business or project management or technical part tends to seriously affect the project.

One of the most consuming tasks in DatoSphera were forming the team and evaluating the best way to work, adapted to the team structure and skills. Incorporating the company and legal aspects of public funding were also given special attention.

Lean-Startup and AGILE methodologies proved to work perfectly for DatoSphera as the project was changing one week to the other, depending on the user feedback. One of the most challenging parts was to focus on the minimum product, instead of developing and testing complex architectures.

The technical temptation of trying and evaluating all the tools available was almost causing the project to lose focus, at the beginning. A good balanced team was what saved the project from losing focus, because our team involved people who execute and also who think in research.

Regarding the data, the mistakes done were huge but helped us to better explore the data and reuse the experience. Although we started with Alexa, adding data sources that were not needed, we should have measured the data quality beforehand and ensure that the data sources that we are adding are relevant for the user's needs. This mistake was leading to months wasted merging data that in reality could be very difficult retrieved.

Starting with a huge data source, but of poor quality was a bad decision, soon replaced by a smaller dataset, Crunchbase, with fewer web domains but more accurate one. From Crunchbase we will step by step measure data quality and coverage of the next data sources to be added. We will also be caution regarding the legal implications and the risks of using online data sources, such as Angel List.

We have learned that adding fewer but carefully selected data sources might be a better strategy than adding all the data sources and not measuring the coverage of one with the other or their data quality.

Overall DatoSphera as a real-life project was dealing with all sorts of problems, from losing the focus on minimal product to business constraints, as funding or legal. What helped DatoSphera to overcome these challenges was a very well balanced team, a combination of administration, execution and research.

7.Bibliography

7.1.Business bibliography

1. *Venu Vasudevan. 2006. Global Software Enterpreneurship. In* Proceedings of the 30th Annual International Computer Software and Applications Conference - Volume 01 *(COMPSAC '06), Vol. 1. IEEE Computer Society, Washington, DC, USA, 55-56. DOI=10.1109/COMPSAC.2006.50 http://dx.doi.org/10.1109/COMPSAC.2006.50*

2. *Joanne Eglash. 2000.* How to Write a .COM Business Plan: The Internet Entrepreneur's Guide to Everything You Need to Know about Business Plans and Financing Options*. McGraw-Hill Professional.*

3. *Shan Wang, Ji-Ye Mao, and Norm Archer. 2012. On the performance of B2B e-markets: An analysis of organizational capabilities and market opportunities.* Electron. Commer. Rec. Appl.*11, 1 (January 2012), 59-74. DOI=10.1016/j.elerap.2011.07.001 http://dx.doi.org/10.1016/j.elerap.2011.07.001*

4. *C. Lee Giles, Yves Petinot, Pradeep B. Teregowda, Hui Han, Steve Lawrence, Arvind Rangaswamy, and Nirmal Pal. 2003. eBizSearch: a niche search engine for e-business. In*Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval *(SIGIR '03). ACM, New York, NY, USA, 413-414. DOI=10.1145/860435.860527 http://doi.acm.org/10.1145/860435.860527*

7.2.Technical bibliography

[1] Shifeng Zhang and Steve Goddard. 2007. A software architecture and framework for Web-based distributed Decision Support Systems. Decis. Support Syst. *43, 4 (August 2007), 1133-1150. DOI=10.1016/j.dss.2005.06.001 http://dx.doi.org/10.1016/j.dss.2005.06.001*

[2] Thomas C. Redman. 1997. Data Quality for the Information Age *(1st ed.). Artech House, Inc., Norwood, MA, USA.*

[3] Stuart E. Madnick, Richard Y. Wang, Yang W. Lee, and Hongwei Zhu. 2009. Overview and Framework for Data and Information Quality Research. J. Data and Information Quality *1, 1, Article 2 (June 2009), 22 pages. DOI=10.1145/1515693.1516680 http://doi.acm.org/10.1145/1515693.1516680*

[4] Heiko Müller, Johann-Christoph Freytag, and Ulf Leser. 2012. Improving data quality by source analysis. J. Data and Information Quality *2, 4, Article 15 (March 2012), 38 pages. DOI=10.1145/2107536.2107538 http://doi.acm.org/10.1145/2107536.2107538*

[5] Alexander Alexandrov, Christoph Brücke, and Volker Markl. 2013. Issues in big data testing and benchmarking. In Proceedings of the Sixth International Workshop on Testing Database Systems*(DBTest '13). ACM, New York, NY, USA, , Article 1 , 5 pages. DOI=10.1145/2479440.2482677 http://doi.acm.org/10.1145/2479440.2482677*

[6] *Vladimír Olej, Jana Filipová, and Petr Hájek. 2010. Time series prediction of web domain visits by IF-inference system. In* Proceedings of the 14th WSEAS international conference on Computers: part of the 14th WSEAS CSCC multiconference - Volume I *(ICCOMP'10), Nikos E. Mastorakis, Valeri Mladenov, and Zoran Bojkovic (Eds.), Vol. I. World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, 156-161.*

[7] *Tilmann Rabl, Sergio Gómez-Villamor, Mohammad Sadoghi, Victor Muntés-Mulero, Hans-Arno Jacobsen, and Serge Mankovskii. 2012. Solving big data challenges for enterprise application performance management.* Proc. VLDB Endow. *5, 12 (August 2012), 1724-1735. DOI=10.14778/2367502.2367512 http://dx.doi.org/10.14778/2367502.2367512*

7.3.Project management bibliography

1) *Shai Rozenes. 2011. The Impact of Project Management Methodologies on Project Performance.*Int. J. Inf. Technol. Proj. Manag. *2, 2 (April 2011), 64-73. DOI=10.4018/jitpm.2011040105 http://dx.doi.org/10.4018/jitpm.2011040105*

2) *San Murugesan and Behnaz Gholami. 2011. Global IT Project Management Using Web 2.0.* Int. J. Inf. Technol. Proj. Manag. *2, 3 (July 2011), 30-52. DOI=10.4018/jitpm.2011070103 http://dx.doi.org/10.4018/jitpm.2011070103*

3) *Alan R. Peslak. 2012. Information Technology Project Management and Project Success.* Int. J. Inf. Technol. Proj. Manag. *3, 3 (July 2012), 31-44. DOI=10.4018/jitpm.2012070103 http://dx.doi.org/10.4018/jitpm.201207010*

8.Annex

8.1.Annex - Sprints- Historical evolution

8.1.1. Initial context – Inception Sprint

8.1.1.1. Context

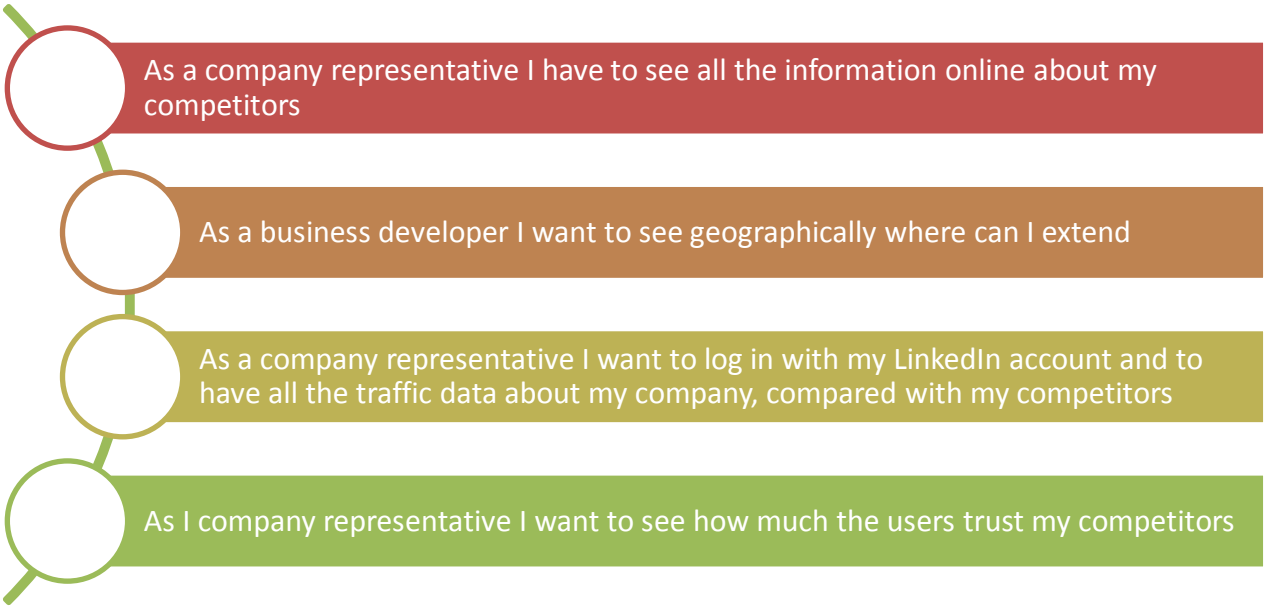
DatoSphera is a spin-off of the project Trakty, which gathered different data sources during a period of 6 months. Trakty is a project belonging to Incubio Research, that provides the projects incubated an additional help with the Big Data technologies. These data sources were aggregated for a different purpose and their quality was not measured accordingly. This translates into a strategic decision on using the Trakty data or following the same principle but with different data sources.

This sprint was used also to form the current team of DatoSphera and to incorporate the company itself.

8.1.1.2. User Stories

Trakty was a project designed to study the competitors for web domains. Therefore the user stories were based on the competitors' analysis and on their traffic data, more particularly.

Some samples of the most relevant user stories are the following:



8.1.1.3. Team Roles

The team roles followed inside the project Trakty was the following:

- Frederic Montes- Technical Coordinator
- Madalina Burghelea – Business Intelligence and Big Data Development
- David Berruezo- Developer
- Cristian Vitales- Developer

AGILE methods and SCRUM were used to distribute the amount of work inside the team. JIRA was used as a tool to distribute the user stories.

8.1.1.4. Estimated versus Realized Analysis

As the user stories were vague and therefore leaving much space to interpret, different data sources have been aggregated. For all that, the only constraint was the time constraint, that was not accomplished. The deadline for finishing the project was 6 months, although the aggregation of data sources finished in 9 months. All the other tasks were accomplished properly.

A data analysis of data quality was missing and therefore the data sources were aggregated without a proper analysis. The analysis was done just after finishing aggregating them.

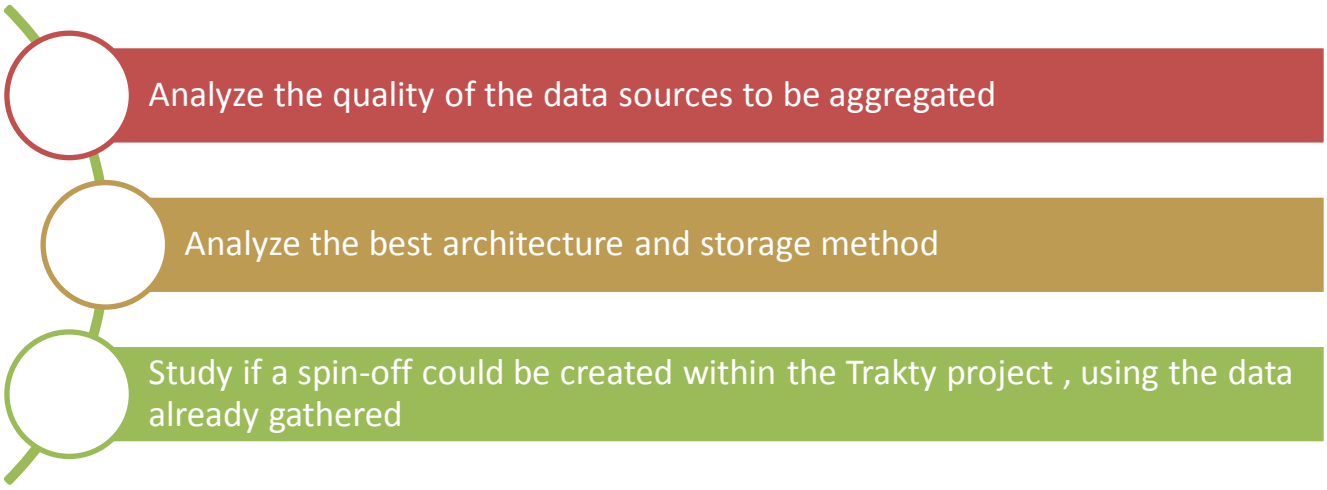
8.1.2. Sprints 1 -4 – Analyze the current architecture

8.1.2.1. Context

The data from the Trakty project was to be used inside DatoSphera, if the quality was enough to create a spin-off within this architecture. The measures were necessary in order to continue with the development or decide on an new strategy of aggregating the data.

8.1.2.2. User Stories

The stories to be solved in this sprint were the following:



8.1.2.3. Team Roles

The team roles were the following from the Trakty project, now combined with the DatoSphera team.

- Frederic Montes- Technical Coordinator Trakty
- Madalina Burghelea – Business Intelligence and Big Data Development
- Jordi Masramon – Technical Coordinator DatoSphera
- David Berruezo- Developer
- Cristian Vitales- Developer

8.1.2.4. Estimated versus Realized Analysis

The time estimation for the data analysis was 2 weeks but for all that, this process took 1 month, due to the technical difficulties in evaluating all the data sourced and in detecting all the mistakes that were done when this aggregation was performed.

8.1.3. Sprints 5-10- Crunchbase based data set

8.1.3.1. Context

As decided in the previous sprint, starting from Alexa and merging data sources was an error, due to the quality problems that Alexa faces. Therefore we started again, this time with Crunchbase as the main high quality data source and then planning to aggregate other data sources that have a similar good quality.

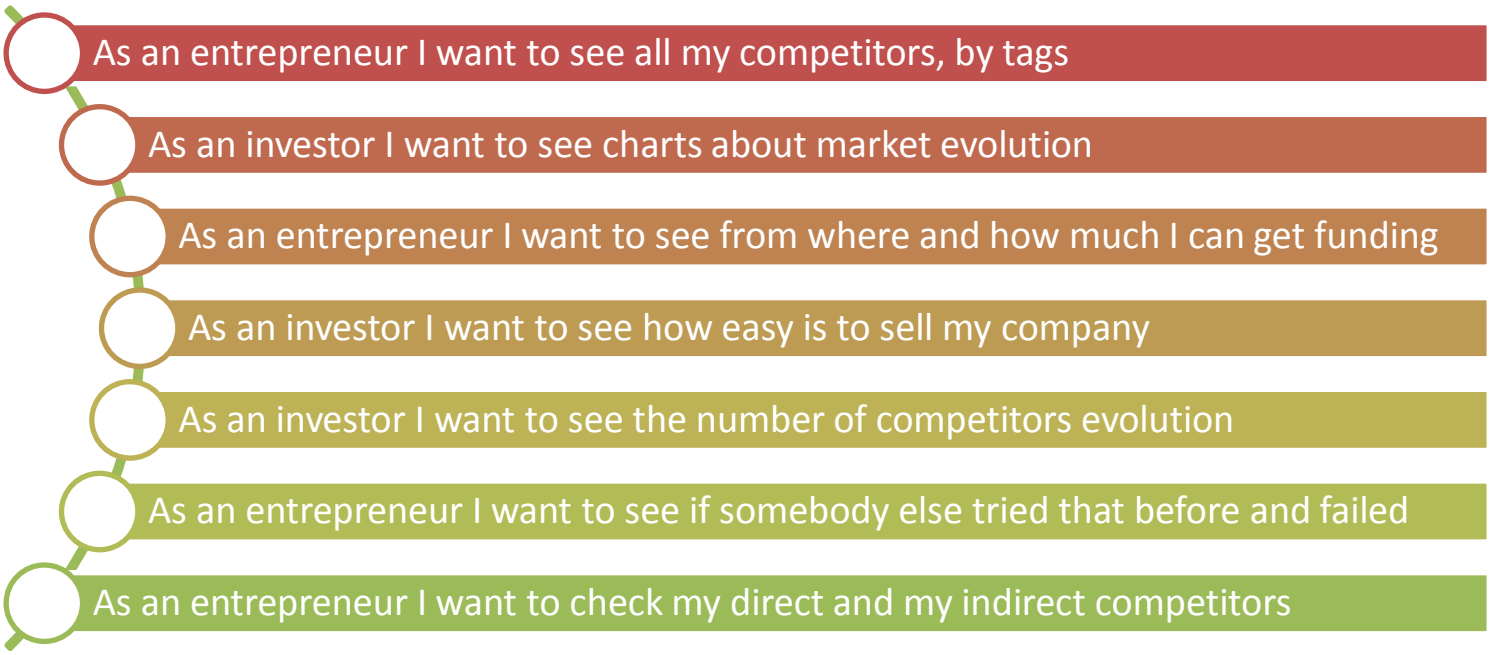
At this point in time the DatoSphera team was complete and although we had the opportunity to reuse the data in the project Trakty, we decided to start from scratch and measure intensively the data quality coefficients, to retrieve the best results.

We also had to decide if we follow SCRUM model or Kanban , depending on the structure of the team and its skills.

The company in this point was incorporated from Incubio Research as a funded company inside Incubio.

8.1.3.2. User Stories

The user stories treated were the following:



8.1.3.3. Team Roles

At this point in time the DatoSphera team was complete and the team roles during this sprint were the following:

- Madalina Burghelea – Product Owner
- Jordi Masramon – CTO
- Denis di Paolo – Developer
- Xavier Ruiz- SCRUM Master and Project Manager
- Pablo Casado – Data Architect

The AGILE methodologies were followed completely starting from this sprint and therefore SCRUM was better defined in this sprint.

8.1.3.4. Estimated vs Realized

The time allocated to finding the team was 1 month and was accomplished. The time dedicated to extracting Crunchbase data was one week and was accomplished, while the one for developing the prototype was one month and was extended to two months, due to the difficulties in understanding the Crunchbase data.

8.1.4. Sprint 10-13 – Aggregating Angel List

8.1.4.1. Context

We measured how many companies were in Crunchbase and although the quality was good there were the following problems with it:

- Crunchbase had only 300k companies and most of them big companies
- Crunchbase was mainly populated by companies in the United States

Therefore an external data source had to be found and the most suitable one looked to be Angel List, one of the biggest databases of start-ups, in Europe.

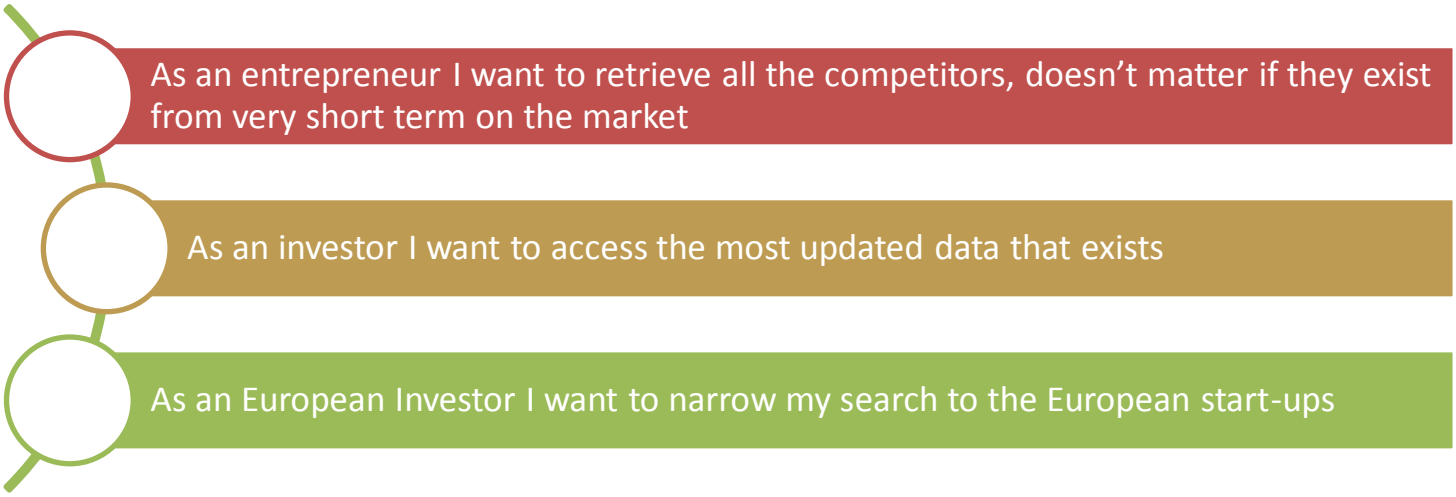
The benefits and the problems of using Angel List are summarized below:

Angel List	
Pros	Cons
Biggest data source for start-ups	Legal issues with scrapping their data
High popularity in Europe	
The data is the most frequent updated one	

The decisions taken will be motivated in the following sections.

8.1.4.2. User Stories

The stories to be treated in this sprint were the following:



8.1.4.3. Team Roles

The team roles during this sprint were the following:

- Madalina Burghelea – Product Owner
- Jordi Masramon – CTO
- Denis di Paolo – Developer
- Xavier Ruiz- SCRUM Master and Project Manager
- Pablo Casado- Data Architect

8.1.4.4. Estimated vs Realized

The problems faced during this sprint were related to the legal issues that interrupt the normal process of extracting the companies. Due to legal issues, extracting the companies from Angel List was creating significant delays. The initial time estimation was one week, but due to lawyers and the legal implications, the process took 3 weeks.

8.1.4.5. Legal Implications

Angel List imposes certain restrictions that do not allow the use of their data in other applications. A summary of their legal implications is shown below:

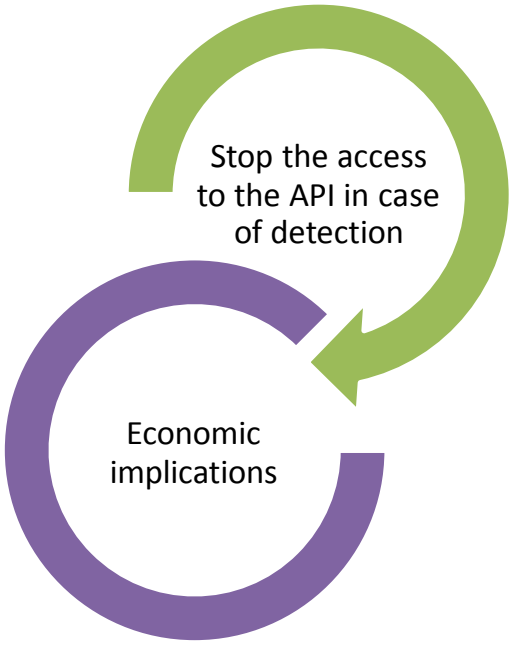
Plain English Terms of Service

Don't store any raw data returned via the API for more than 24 hours.
Don't use the API to scrape data.
Don't use any undocumented endpoints without our explicit permission.
Don't store any of our users' login credentials.
Don't publicly discuss an ongoing private financing. There are federal laws which govern these announcements, commonly referred to as General Solicitation. You can read more about it here.
Do credit AngelList wherever our data is used.
Don't use the AngelList logo or the word "AngelList" as the logo or name for the product you build upon our API.

These terms may change at any time. The last update occurred on January 5, 2012.

We have used the help of a consultancy firm of lawyers, **Avatic Abogados** , and they have evaluated the risks that using Angel Data will imply.

The risks were the following, depending on the amount of norms that were not respected:



The technical decision taken after studying the lawyers report is that we are going to use the open information and find other data sources to complement the missing information.

8.1.5. Sprint 13-18 – Improve usability and user experience

8.1.5.1. Context

The product was supposed to be presented at events and ready to be tested with online users. The only problem is that the interface was far from being adapted to the user needs. The users were typing tags, instead of writing directly their ideas.

We have worked on a very simple, Google-like interface, to make the user familiar with the way of introducing the business ideas.

8.1.5.2. User Stories

The stories treated in this sprint were the following:

8.1.5.3. Team Roles

The team roles during this sprint were the following:

- Madalina Burghelea – Product Owner
- David Mundo- Designer and UI

8.1.5.4. Estimated vs Realized

This process was planned for 2 weeks but due to little details that affected the implementation, the design was realized in 3 weeks. During this time it was continuously tested with users. The technical team meanwhile was supporting the natural language processing implementation.