



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)

IMPROVING SMART CITY SERVICES BY MEANS OF BIG DATA TECHNIQUES

Autor: Ignacio Rosario Carrera

Director: Eugenio F. Sánchez Úbeda, Ana Salgado Ortega

Madrid
Julio 2016

Ignacio
Rosario
Carrera

“IMPROVING SMART CITY SERVICES BY MEANS OF BIG DATA TECHNIQUES”



AUTORIZACIÓN PARA LA DIGITALIZACIÓN, DEPÓSITO Y DIVULGACIÓN EN RED DE PROYECTOS FIN DE GRADO, FIN DE MÁSTER, TESINAS O MEMORIAS DE BACHILLERATO

1º. Declaración de la autoría y acreditación de la misma.

El autor D. IGNACIO ROSARIO CARRERA

DECLARA ser el titular de los derechos de propiedad intelectual de la obra: “IMPROVING SMART CITY SERVICES BY MEANS OF BIG DATA TECHNIQUES” que ésta es una obra original, y que ostenta la condición de autor en el sentido que otorga la Ley de Propiedad Intelectual.

2º. Objeto y fines de la cesión.

Con el fin de dar la máxima difusión a la obra citada a través del Repositorio institucional de la Universidad, el autor **CEDE** a la Universidad Pontificia Comillas, de forma gratuita y no exclusiva, por el máximo plazo legal y con ámbito universal, los derechos de digitalización, de archivo, de reproducción, de distribución y de comunicación pública, incluido el derecho de puesta a disposición electrónica, tal y como se describen en la Ley de Propiedad Intelectual. El derecho de transformación se cede a los únicos efectos de lo dispuesto en la letra a) del apartado siguiente.

3º. Condiciones de la cesión y acceso

Sin perjuicio de la titularidad de la obra, que sigue correspondiendo a su autor, la cesión de derechos contemplada en esta licencia habilita para:

- a) Transformarla con el fin de adaptarla a cualquier tecnología que permita incorporarla a internet y hacerla accesible; incorporar metadatos para realizar el registro de la obra e incorporar “marcas de agua” o cualquier otro sistema de seguridad o de protección.
- b) Reproducirla en un soporte digital para su incorporación a una base de datos electrónica, incluyendo el derecho de reproducir y almacenar la obra en servidores, a los efectos de garantizar su seguridad, conservación y preservar el formato.
- c) Comunicarla, por defecto, a través de un archivo institucional abierto, accesible de modo libre y gratuito a través de internet.
- d) Cualquier otra forma de acceso (restringido, embargado, cerrado) deberá solicitarse expresamente y obedecer a causas justificadas.
- e) Asignar por defecto a estos trabajos una licencia Creative Commons.
- f) Asignar por defecto a estos trabajos un HANDLE (URL *persistente*).

4º. Derechos del autor.

El autor, en tanto que titular de una obra tiene derecho a:

- a) Que la Universidad identifique claramente su nombre como autor de la misma
- b) Comunicar y dar publicidad a la obra en la versión que ceda y en otras posteriores a través de cualquier medio.
- c) Solicitar la retirada de la obra del repositorio por causa justificada.
- d) Recibir notificación fehaciente de cualquier reclamación que puedan formular terceras personas en relación con la obra y, en particular, de reclamaciones relativas a los derechos de propiedad intelectual sobre ella.

5º. Deberes del autor.

El autor se compromete a:

- a) Garantizar que el compromiso que adquiere mediante el presente escrito no infringe ningún derecho de terceros, ya sean de propiedad industrial, intelectual o cualquier otro.
- b) Garantizar que el contenido de las obras no atenta contra los derechos al honor, a la intimidad y a la imagen de terceros.
- c) Asumir toda reclamación o responsabilidad, incluyendo las indemnizaciones por daños, que pudieran ejercitarse contra la Universidad por terceros que vieran infringidos sus derechos e intereses a causa de la cesión.

- d) Asumir la responsabilidad en el caso de que las instituciones fueran condenadas por infracción de derechos derivada de las obras objeto de la cesión.

6º. Fines y funcionamiento del Repositorio Institucional.

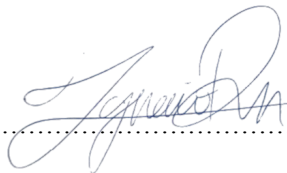
La obra se pondrá a disposición de los usuarios para que hagan de ella un uso justo y respetuoso con los derechos del autor, según lo permitido por la legislación aplicable, y con fines de estudio, investigación, o cualquier otro fin lícito. Con dicha finalidad, la Universidad asume los siguientes deberes y se reserva las siguientes facultades:

- La Universidad informará a los usuarios del archivo sobre los usos permitidos, y no garantiza ni asume responsabilidad alguna por otras formas en que los usuarios hagan un uso posterior de las obras no conforme con la legislación vigente. El uso posterior, más allá de la copia privada, requerirá que se cite la fuente y se reconozca la autoría, que no se obtenga beneficio comercial, y que no se realicen obras derivadas.
- La Universidad no revisará el contenido de las obras, que en todo caso permanecerá bajo la responsabilidad exclusiva del autor y no estará obligada a ejercitar acciones legales en nombre del autor en el supuesto de infracciones a derechos de propiedad intelectual derivados del depósito y archivo de las obras. El autor renuncia a cualquier reclamación frente a la Universidad por las formas no ajustadas a la legislación vigente en que los usuarios hagan uso de las obras.
- La Universidad adoptará las medidas necesarias para la preservación de la obra en un futuro.
- La Universidad se reserva la facultad de retirar la obra, previa notificación al autor, en supuestos suficientemente justificados, o en caso de reclamaciones de terceros.

Madrid, a DIECISIETE de JULIO de 2016

ACEPTA

Fdo.....



Motivos para solicitar el acceso restringido, cerrado o embargado del trabajo en el Repositorio Institucional:

--

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título
“IMPROVING SMART CITY SERVICES BY MEANS OF BIG DATA
TECHNIQUES”

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el
curso académico 2015/2016 es de mi autoría, original e inédito y
no ha sido presentado con anterioridad a otros efectos. El Proyecto no es
plagio de otro, ni total ni parcialmente y la información que ha sido tomada
de otros documentos está debidamente referenciada.

Fdo.: Ignacio Rosario Carrera

Fecha: 17/ 07/ 2016

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: Eugenio F. Sánchez Úbeda

Fecha://

Vº Bº del Coordinador de Proyectos

Fdo.: Álvaro Sánchez Miralles

Fecha://

“IMPROVING SMART CITY SERVICES BY MEANS OF BIG DATA TECHNIQUES”

Autor: Rosario, Ignacio

Dirigido por:

Ana Salgado Ortega, Eugenio F. Sánchez Úbeda

ICAI – Universidad Pontificia Comillas

RESUMEN DEL PROYECTO

1. Introduction

“Data really powers everything that we do”, the quote corresponds to Jeff Weiner, chief executive officer of LinkedIn. Nowadays, data has emerged as one of the most important assets of a company, enabling data driven decisions and actions in order to maintain its power inside the market and not let the competition overtake them.

The main problem these days is the vast amount of devices generating the data: GPS, cars, RFID tags, social networks, smart meters, customer data, mobiles ... Two of the main problems encountered

are: the amount of data being generated (more than 90% of today’s data was generated in the last two years) and the different natures of the created data. Not only it can come from a huge variety of sources, but also it doesn’t have the same structure inside each of the data sources (see Figure 1). Moreover, 88% of today’s data is estimated to be unstructured data, the new way to transmit data (XML documents stored in NoSQL databases for example) and the fastest and largest growing sector.

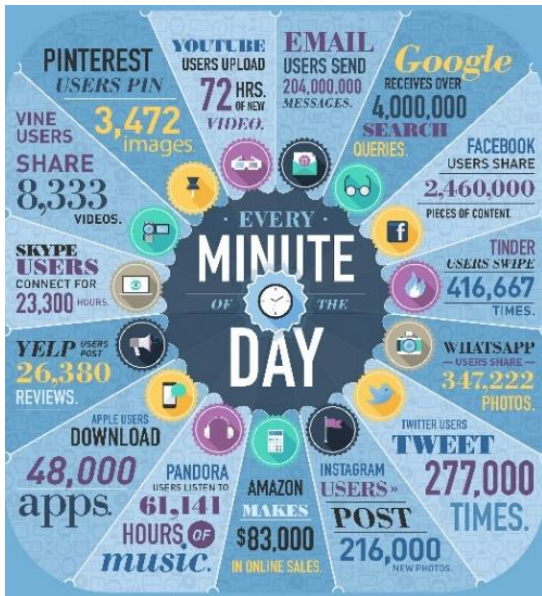


Figure 1: What happens in 1 minute in the internet? [1]

Old technologies can't keep up with the speed at which the data grows and so, Big Data offers a new set of tools that enable organizations to gather, store and manipulate vast amounts of data at the right speed and time.

It is not possible to talk about a single technology that takes care of the Big Data challenges. It is a set of tools and methods to be used together in the different stages of the data acquisition and processing in order to analyze the huge, differently structured pieces of data that will bring crucial insight to the company

Nowadays, many sectors benefit from the new tools brought by Big Data

technologies such as banking, health, construction and Internet of Things (IoT). IoT can be defined as a network of devices that exchange and collect data between them. Big Data technologies really empowers this kind of networks as sensors are one of the most data-prolific entities out there and now, we can not only store all the information measured by this sensor networks, but also make deep analytics in real time on the data just generated by the sensor network.

The combination between IoT and Big Data technologies can help the companies who own the data created by those networks really step ahead of the competition as they get the ability to gain very meaningful insight about their products or services in the area where the IoT network is implemented.

Each year, Gartner, one of the leaders in the Information Technologies market, publishes a "hype cycle" for emerging technologies, where it introduces the latest technologies created and their corresponding stage in the hype cycle: Technology trigger, Peak of inflated expectations, Trough of disillusionment, Slope of enlightenment and Plateau of productivity.

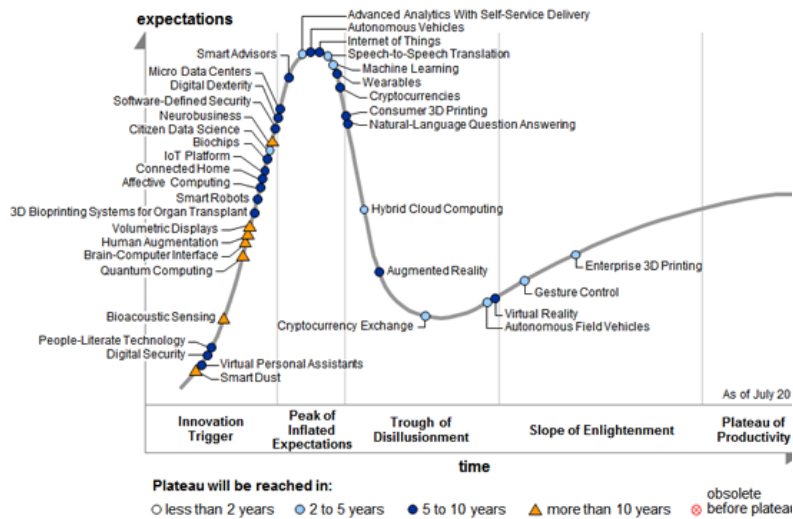


Figure 2: Gartner's hype cycle for 2015 [2]

If we analyze the hype cycles for 2014 and 2015 (Figure 3) we can see that in both of them, IoT appears as the technology with maximum expectations and that Big Data disappears in the 2015 cycle whereas in the 2014 it was presented as a technology with a high amount of expectations that would reach the plateau in 5 to 10 years. Big Data is present in the 2015 cycle as it is present in most of the technologies presented by Gartner: IoT, Machine Learning, Neuro-business, etc. Gartner presents IoT and Big Data as one of the most promising technologies in the market nowadays.

One application of the combination between the IoT and Big Data technologies is the creation of Smart Cities in order to make cities smarter, more efficient, and environmentally friendly, as

well as to understand the city in order to provide better services to the citizens. This also means new services and efficiency actions for the companies that offer services to the cities as Ferrovial, in order words, new revenues from the new services and reduction of costs from the efficiency actions based on the use of data.

2. Methodology

As the project can be categorized as a “Software Development Project”, it has been carried out using the *Agile* methodology in order to have the final client’s feedback during the development, instead of at its completion. Thus, having its feedback on the product as soon as possible ensures it achieves what the client really wants.

Being able to create discussions with the business based on data results as soon as possible and make the business participate in the definition of the final platform is a strategic goal behind the agile methodology.

The Agile methodology is very common in the Software Project Management as it enables the team to provide value as soon as possible. Instead of committing to yearlong projects, the agile methodology suggests to work in “Sprints”. Each Sprint having a duration about three weeks and no more than four, has unchangeable objectives defined at the beginning of the Sprint and the development has to commit to them, nothing more and nothing less. Once the Sprint is finished, the “product” is released for the customer to see the evolution and the Sprint is evaluated in order to think about possible improvements to add in the next Sprint.

In this project the first sprints are organized by cities and analytics layers, the first one dedicated to the city of Santander creating the base for the description layer, being followed by Murcia, Barcelona and Madrid and getting also the predictive and prescriptive layers in following sprints.

The first tasks correspond to the first layer of analytics: Retrieve the data for the corresponding city, cleaning it and storing data in the database with the correct formatting. As a result the platform is gathering real time and historic data for the specific city to which the sprint is dedicated, enabling realistic discussion about first view trends and relationships between sensor information and service operation.

In the second iteration for each city, the data visualization results from the first iteration were improved upon the conclusions obtained from the previous iteration. Deeper analytics were also performed in order obtain more insight about the city’s behavior and make data driven decisions about the services offered in those cities.

3. Results

The first city in the iteration process defined by the Agile methodology in the scope of the project is Santander. In this case, Ferrovial has a partnership with the Cantabria’s college to implement a number of sensors throughout the whole city. The two approaches to the implemented platform are the following:

- **Deep Analytics platform** to visualize historical data and run predictive analytics on it to understand how the citizens behave and thus, the pace of the city.
- **Online Analytics platform** to monitor the city in real time in order to improve the reactive maintenance of services.

The sensor's data is accessed via the FIWARE API and the sensors update their measurements every 10 minutes theoretically. The stored dataset from more than 1200 sensors going from August 2015 to February 2016 consists of more than 50 variables and 150 millions of rows.

The main problem resides in the API maintenance, as the data from the sensors was never used until this new initiative from Ferrovial. The sensors' maintenance is non-existent and thus the measures are very volatile. In particular, 20% of the sensors recorded their last measurement in 2013 or 2014 and, the ones that are currently measuring data, have problems in their ability to acquire accurate measurements as they have periods of normal measurements followed by months of inactivity with no measurements at all.

Due to the data quality and the golden rule applied to predictive analytics of "garbage in – garbage out" the data from the FIWARE API of Santander wasn't used for any predictive task but was used to monitor the city and do exploratory analysis of the city's behavior. The next city in the roadmap is Barcelona, where Ferrovial is responsible for waste collection. Among all the containers in the city of Barcelona, 303 of them were selected in Sarriá to have sensors implemented to monitor the container's interior temperature and fill level.

As for Santander, the data from Barcelona was included in the developed Smart Cities' platform but without the problem of data quality as the majority of containers have correct measurements.

The predictive models implemented belong to two main categories:

- **Classification:** The algorithm predicts, for each container, at each hour of the day if it will be full or not with a 75% accuracy.
- **Regression:** The implemented algorithm predicts, for each container, the incremental fill level for each moment of the day (the day is divided into segments of equal length), with

an average Root Mean Squared Error of 2.3% in the test set.

4. Conclusions

A platform for smart city services analytics has been developed and tested in a real case study for the waste containers use. This platform has two main modules. The Deep Analytics module allows analyzing historical data as well as running predictive analytics, whereas the Online Analytics one allows the user to monitor the city in real time.

In the classification approach, only one model was implemented, using the container type as an input for the model, whereas in the regression approach each container has a particular model implemented on its own. These approaches have been tested on the Barcelona case study, with classification errors of 25% and a prediction accuracy of 2.3% RMSE for the regression approach.

The regression approach, based on using one model per container, results in more accurate predictions, but in a problem of scalability when it is generalized to the whole city or cities. Another drawback of the regression approach is the

impossibility of adding a new container unseen previously as no model would have been trained to fit its data.

One approach for the future work could be to run a cluster algorithm to group containers behaving similarly and adjust a predictive model for each cluster, solving the problem of scalability and unseen containers.

References

- [1] https://web-assets.domo.com/blog/wp-content/uploads/2014/04/DataNeverSleeps_2.0_v2.jpg
- [2] <http://digitaltechdiary.com/gartners-2015-hype-cycle-for-digital-marketing/2241/>

“IMPROVING SMART CITY SERVICES BY MEANS OF BIG DATA TECHNIQUES”

Autor: Rosario, Ignacio

Dirigido por:

Ana Salgado Ortega, Eugenio F. Sánchez Úbeda

ICAI – Universidad Pontificia Comillas

RESUMEN DEL PROYECTO

1. Introducción

“Data really powers everything that we do”, la frase corresponde a Jeff Weiner, consejero delegado de LinkedIn. Hoy en día los datos han emergido como uno de los activos más importantes de las empresas, permitiendo decisiones basadas en datos y acciones con el fin de mantener y crear las ventajas competitivas sobre el mercado y seguir por encima de la competencia.

El problema principal encontrado es el gran número de dispositivos generando datos: GPS, coches, etiquetas RFID, redes sociales, smart meters, datos de

consumidores, móviles... Los dos problemas principales encontrados son: la cantidad de datos generada (más del 90% de los datos de hoy fueron generados a lo largo de los dos últimos años) y los distintos tipos de datos generados.

No sólo la gran cantidad de fuentes de datos sino las distintas estructuras de cada una de las fuentes de datos (ver Figura 1). Más en detalle, se estima que el 88% de los datos de hoy se trata de datos no estructurados, la nueva manera de transmitir datos (XML, documentos almacenados en bases de datos NoSQL...)

y el sector que crece más rápidamente hoy en día.

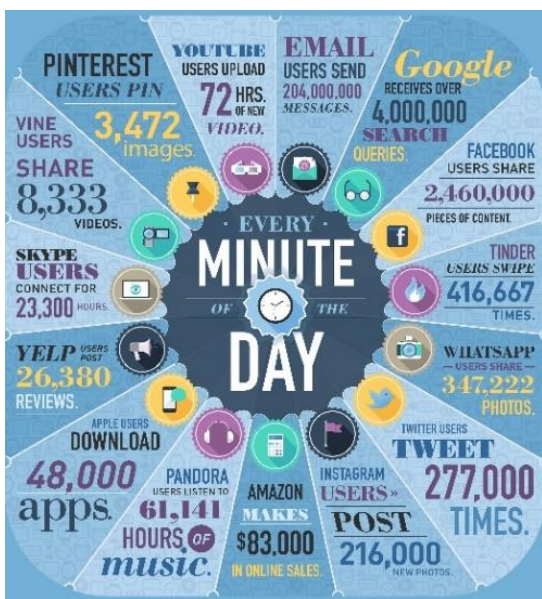


Figura 1: What happens in 1 minute in the internet? [1]

Las antiguas tecnologías no pueden mantener el ritmo impuesto por el crecimiento de los datos y, por lo tanto, las tecnologías Big Data ofrecen un nuevo grupo de herramientas para que las empresas puedan reunir, almacenar y manipular grandes cantidades de datos a unas velocidades y tiempos de procesamiento aceptables.

Es importante recalcar que no hay una tecnología que resuelva los retos que plantea el Big Data. Se trata de una conjunción de herramientas y métodos que, usados en las distintas etapas de la adquisición y procesamiento de datos,

permiten analizar las enormes cantidades y diferentes estructuras de datos que traerán nuevos puntos de vista a los análisis de la compañía.

Muchos sectores se benefician hoy en día de las ventajas aportadas por las tecnologías Big Data: banca, sanidad, construcción o el Internet de las Cosas (IoT por sus siglas en inglés). IoT puede definirse como una red de artefactos que intercambian y almacenan datos entre ellos. Las tecnologías Big Data pueden aportar mucho valor a dichas redes de sensores ya que se trata de una de las fuentes más prolíficas de datos que puedan existir y no sólo permiten el almacenamiento de dicha información sino que también permiten la realización de analítica avanzada sobre dichos datos en tiempo real.

La combinación de IoT y tecnologías Big Data puede permitir a las compañías propietarias de los datos generados por dichas redes de sensores diferenciarse respecto a la competencia ya que consiguen aportar una visión nueva sobre los productos o servicios del área en el que la red de sensores fue implementada.

Cada año, Gartner, uno de los líderes en el sector de las Tecnologías de la

Información, publica un “hype cycle” en el que expone las últimas tecnologías y su nivel de madurez.

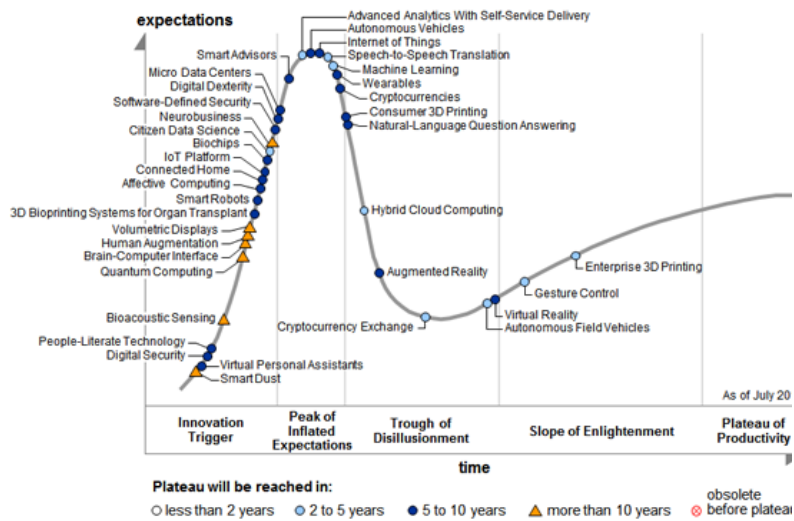


Figura 2: Gartner's hype cycle for 2015 [2]

Si se analiza el hype cycle para 2015, se puede observar que el IoT aparece como la tecnología con las máximas expectativas y que el Big Data desaparece con respecto el hype cycle de 2014.

El Big Data está presente en el ciclo de 2015 ya que forma parte de la mayoría de las tecnologías presentadas por Gartner: IoT, Machine Learning, Neuro-business, etc. Gartner presenta la combinación de Big Data e IoT como una de las tecnologías más prometedoras del presente y el futuro.

Una aplicación de la combinación de IoT y Big Data es la creación de Smart Cities

tanto como para crear ciudades “inteligentes”, más eficientes y que respeten el medio ambiente como para entender el comportamiento de los ciudadanos y así poder proveer mejores servicios en dichas ciudades.

Esto representa a su vez nuevas oportunidades para las empresas responsables de dichos servicios, como Ferrovial, ya sea en forma de nuevos servicios o mejora de la eficiencia de servicios ya implementados.

2. Metodología

El proyecto puede ser categorizado como un proyecto de desarrollo de software, se

ha desarrollado usando metodologías ágiles, obteniendo feedback continuo del cliente final durante la fase de desarrollo en lugar de esperar a la conclusión del proyecto. Dicha manera de proceder permite asegurar que al final del proyecto se obtiene lo que de verdad el cliente pretendía.

El ser capaces de generar discusiones con el negocio lo antes posible y hacer partícipe al negocio en la definición de la plataforma final es un objetivo estratégico planteado por las metodologías ágiles.

Las metodologías ágiles son muy comunes en los proyectos de desarrollo de software ya que permite al equipo aportar valor lo antes posible. En lugar de comprometerse en proyectos de larga duración, las metodologías ágiles sugieren el trabajo en “sprints”. Cada sprint tiene una duración de entre tres y cuatro semanas y objetivos incambiables definidos al principio del mismo. Los desarrolladores tienen que comprometerse a ellos, nada más y nada menos. Una vez el sprint finaliza, se muestra el producto al cliente para que vea la evolución y dé feedback. Se evalúa cada sprint para identificar áreas de mejora aplicables en sprints próximos.

En el cuadro del presente proyecto los sprints se organizan por ciudades y por capas de analítica.

El primero es dedicada a la ciudad de Santander, creando la base para la primera capa de analítica descriptiva, siendo continuado por Murcia, Barcelona y Madrid, iterando después para añadir analítica predictiva y prescriptiva.

Las primeras tareas se corresponden con la primera capa de analítica: Recolección de datos de la ciudad en cuestión, limpieza y almacenamiento en la base de datos con el formato correcto. El resultado es una plataforma recoge datos en tiempo real y datos históricos para la ciudad en cuestión, permitiendo discusiones acerca de patrones y relaciones entre sensores de información y servicios de la ciudad.

A partir de la segunda iteración de cada ciudad, las visualizaciones resultantes de la primera iteración se mejoran a raíz de las conclusiones obtenidas. Analítica más avanzada es ejecutada para obtener información más detallada del comportamiento de la ciudad y poder tomar decisiones basadas en datos para los servicios ofrecidos en dichas ciudad.

3. Resultados

La primera ciudad en el proceso iterativo definido por la metodología ágil en el cuadro de este proyecto es Santander. En este caso Ferrovial tiene un acuerdo con la universidad de Cantabria para la sensorización de la ciudad y el acceso a los datos de dichos sensores. Se divide la plataforma creada en dos partes:

- **Plataforma de analítica avanzada** para visualizar datos históricos y poder ejecutar analítica avanzada sobre ellos para entender más profundamente cómo funcionan la ciudad y sus habitantes.
- **Plataforma de analítica Online** para monitorizar la ciudad en tiempo real con el fin de mejorar el mantenimiento reactivo de los servicios.

Lo datos de los sensores son accedidos a través de una API de FIWARE y los sensores actualizan sus mediciones cada diez minutos. El dataset recopilado de más de 1200 sensores entre Agosto de 2015 y Febrero de 2016 consiste de más de cincuenta variables y 150 millones de registros.

El problema principal reside en el mantenimiento de dicha API y de los sensores, ya que los datos de dichos sensores no se habían usado para nada

hasta la llegada de la iniciativa de Ferrovial.

El mantenimiento de los sensores es prácticamente inexistente por lo que las mediciones son muy volátiles. En particular, 20% de los sensores midieron por última vez en los años 2013 o 2014 y, de los que siguen midiendo hoy en día, presentan largos períodos de inactividad antes de volver a medir correctamente.

Debido a la calidad de los datos y a la norma de analítica predictiva “garbage in – garbage out” los datos de la API de FIWARE no se usaron para el desarrollo de analítica predictiva sino que se usó para análisis descriptivo del comportamiento de la ciudad.

La siguiente ciudad del proyecto es Barcelona, donde Ferrovial es responsable de la recogida de residuos. Se seleccionaron 303 contenedores para montar un proyecto piloto de sensorización para monitorizar tanto el nivel de llenado del contenedor como la temperatura interior.

Al igual que para la ciudad de Santander, los datos de Barcelona se incluyeron en la plataforma de Smart Cities.

Los modelos predictivos implementados para la ciudad de Barcelona pertenecen a dos categorías principales:

- Clasificación: El algoritmo predice, para cada contenedor, para cada hora del día si el contenedor estará lleno o no con una precisión del 75%.
- Regresión: El algoritmo implementado predice, para cada contenedor, el incremental de llenado para cada momento del día (se dividen las 24h en segmentos de igual longitud) con un RMSE medio de 2.3% en el conjunto de test.

4. Conclusiones

Se ha desarrollado una plataforma para la analítica de Ferrovial en Smart cities en un caso real para los contenedores de basura. La plataforma consiste de dos módulos principales. El módulo de analítica avanzada permite el análisis de datos históricos y la creación de modelos predictivos mientras que el módulo de analítica Online permite al usuario monitorizar la ciudad en tiempo real.

En el enfoque de clasificación, se desarrolla un solo modelo usando el tipo de contenedor como una de las entradas del modelo mientras que en el enfoque de

regresión cada contenedor tiene un modelo particular.

El modelo de regresión, basado en un modelo para cada contenedor, resulta en predicciones más precisas y que aportan más información pero presenta un problema de escalabilidad al generalizar dicho enfoque a una ciudad completa. Otro problema de realizar un modelo por contenedor es la imposibilidad de realizar predicciones para contenedores no vistos previamente ya que ningún modelo lo recogería.

UN enfoque para el futuro sería el de clusterizar los contenedores para así llegar a un punto medio escalable de complejidad entre los dos opuestos presentados en el proyecto.

Referencias

- [1] https://web-assets.domo.com/blog/wp-content/uploads/2014/04/DataNeverSleeps_2.0_v2.jpg
- [2] <http://digitaltechdiary.com/gartners-2015-hype-cycle-for-digital-marketing/2241/>



Index

Part I	Memoria.....	9
Chapter 1	Introduction	11
	State of the art.....	14
	Project's motivation	18
	Objectives.....	18
	Methodology.....	19
	Resources.....	21
Chapter 2	Smart cities platform	23
	Introduction	23
	Platform design.....	23
	Infrastructure.....	23
	Back-end	23
	Front-end.....	24
	Analysis.....	24
	Online analytics platform.....	25
	Deep analytics platform	26
	Results.....	27
Chapter 3	Classification	31
	Introduction	31
	Data preparation / wrangling	31
	Modeling.....	33
	Support Vector Machine	33



Results.....	36
Chapter 4 Regression.....	37
Introduction	37
Data preparation / wrangling	37
Modeling.....	38
KNN Regressor	39
Decision trees	40
Gradient Boosting Regressor	42
Results.....	43
Chapter 5 Route optimization.....	45
Introduction	45
Route optimization	45
Results.....	46
Chapter 6 Conclusions.....	47
Chapter 7 Future developments	49
Algorithms.....	49
Platform.....	50
Part II Annexes.....	51
Annex 1 kNN algorithm	53
Annex 2 Decision trees.....	56
Annex 3 Gradient Boosting Regression	61
Annex 4 Support Vector Machines.....	63
Annex 5 Amazon Web Services.....	67
Amazon EC ₂	70
Amazon S3.....	70
Amazon EMR	71



Annex 6	Hadoop and MapReduce	73
	Hadoop Distributed File System	74
	MapReduce	74
	Yet Another Resource Negotiator	76
	Bibliography	81



UNIVERSIDAD PONTIFICIA COMILLAS
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)
INGENIERO INDUSTRIAL

ÍNDICE DE LA MEMORIA



Figure index

Figure 1: What happens in 1 minute in the internet? [1].....	11
Figure 2: Big Data landscape [2].....	12
Figure 3: Gartner's hype cycle for 2015 [3].....	13
Figure 4: Big Data/ Smart Cities ecosystem [4].....	15
Figure 5: Example of a custom open data dashboard for the city of Glasgow [5]	17
Figure 6: The Agile cycle.....	20
Figure 7: Illustration of online platform for the city of Santander.....	26
Figure 8: Illustration of the deep analytics platform.....	27
Figure 9: Sensors showing inactivity.....	28
Figure 10: Kernel mapping.....	34
Figure 11: ROC curve example.....	35
Figure 12: KNN Regression example.....	39
Figure 13: Decision tree regression example.....	41
Figure 14: GBM Regression example.....	42
Figure 15: Comparison of optimized vs. normal route.....	45
Figure 16: Illustration of kNN prediction [7].....	53
Figure 17: kNN bias-variance tradeoff [7].....	54
Figure 18: Decision tree example.....	57
Figure 19: Classification based on decision trees [7].....	58
Figure 20: Ensemble methods flowchart [7].....	61



Figure 21: Linear separation in 2D and 3D [7]	63
Figure 22: Support vectors [7].....	64
Figure 23: Kernel trick [7].....	65
Figure 24: AWS availability zones	68
Figure 25: Amazon Web Services	69
Figure 26: Amazon EC2.....	70
Figure 27: Amazon S3.....	71
Figure 28: Amazon EMR	71
Figure 29: Hadoop ecosystem [6]	73
Figure 30: MapReduce paradigm [6]	75
Figure 31: Hadoop 2.0 architecture [6]	76
Figure 32: Apache Hive [6].....	77



Table index

Table 1: Classification features	33
Table 2: Regression features	38
Table 3: KNN Regression results	40
Table 4: Decision tree regression results.....	41
Table 5: GBM Regression results	42



UNIVERSIDAD PONTIFICIA COMILLAS
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)
INGENIERO INDUSTRIAL



UNIVERSIDAD PONTIFICIA COMILLAS
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)
INGENIERO INDUSTRIAL

PART I MEMORIA



UNIVERSIDAD PONTIFICIA COMILLAS
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)
INGENIERO INDUSTRIAL



CHAPTER 1 INTRODUCTION

“Data really powers everything that we do”, the quote corresponds to Jeff Weiner, chief executive officer of LinkedIn. Nowadays, data has emerged as one of the most important assets of a company, enabling data driven decisions and actions in order to maintain its power inside the market and not let the competition overtake them.

The main problem these days is the vast amount of devices generating the data: GPS, cars, RFID tags, social networks, smart meters, customer data, mobiles ... Two of the main problems encountered are: the amount of data being generated (more than 90% of today’s data was generated in the last two years) and the different natures of the created data. Not only it can come from a huge variety of sources, but also it doesn’t have the same structure inside each of the data sources (see Figure 1). Moreover, 88% of today’s data is estimated to be unstructured data, the new way to transmit data (XML documents stored in NoSQL databases for example) and the fastest and largest growing sector.

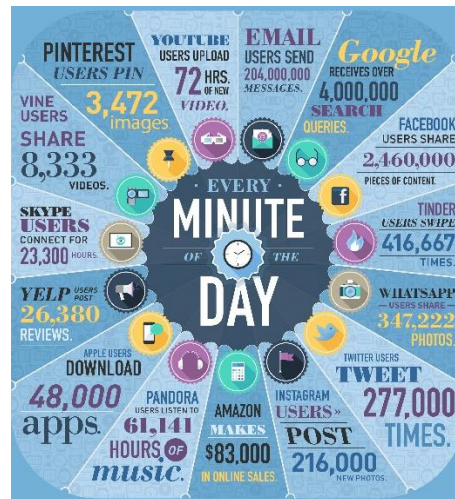


Figure 1: What happens in 1 minute in the internet? [1]

Old technologies can't keep up with the speed at which the data grows and so, Big Data offers a new set of tools that enable organizations to gather, store and manipulate vast amounts of data at the right speed and time.

It is not possible to talk about a single technology that takes care of the Big Data challenges. It is a set of tools and methods to be used together in the different stages of the data acquisition and processing in order to analyze the huge, differently structured pieces of data that will bring crucial insight to the company (see Figure 2).

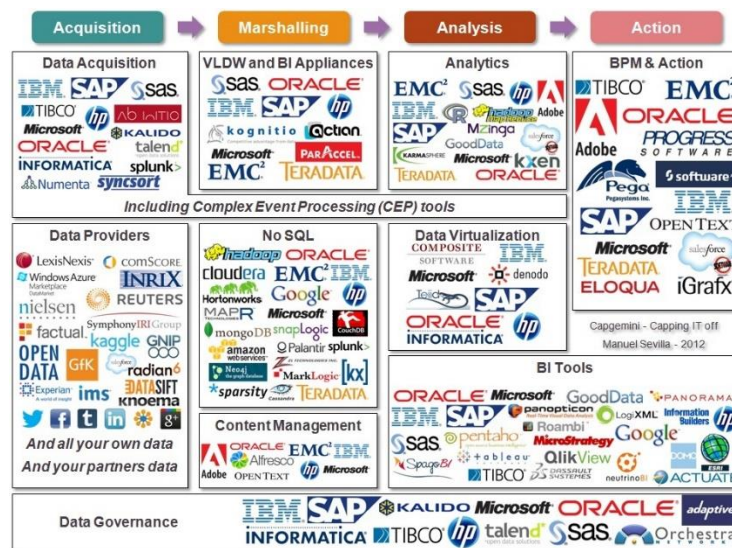


Figure 2: Big Data landscape [2]

Nowadays, many sectors benefit from the new tools brought by Big Data technologies such as banking, health, construction and internet of things. The Internet of Things (IoT) can be defined as a network of devices that exchange and collect data between them. Big Data technologies really empowers this kind of networks as sensors are one of the most data-prolific entities out there and now, we can not only store all the information measured by this sensor networks, but also make deep analytics in real time on the data just generated by the sensor network. The combination between the IoT and Big Data technologies can help the companies who own the data created by those networks really step ahead of the



competition as they get the ability to gain very meaningful insight about their products or services in the area where the IoT network is implemented.

Each year, Gartner, one of the leaders in the Information Technologies market, publishes a “hype cycle” for emerging technologies, where it introduces the latest technologies created and their corresponding stage in the hype cycle: Technology trigger, Peak of inflated expectations, Trough of disillusionment, Slope of enlightenment and Plateau of productivity.

If we analyze the hype cycles for 2014 and 2015 (Figure 3) we can see that in both of them, the IoT appears as the technology with maximum expectations and that Big Data disappears in the 2015 cycle whereas in the 2014 it was presented as a technology with a high amount of expectations that would reach the plateau in 5 to 10 years. Big Data is present in the 2015 cycle as it is present in most of the technologies presented by Gartner: IoT, Machine Learning, Neurobusiness, etc. Gartner presents IoT and Big Data as one of the most promising technologies in the market nowadays.

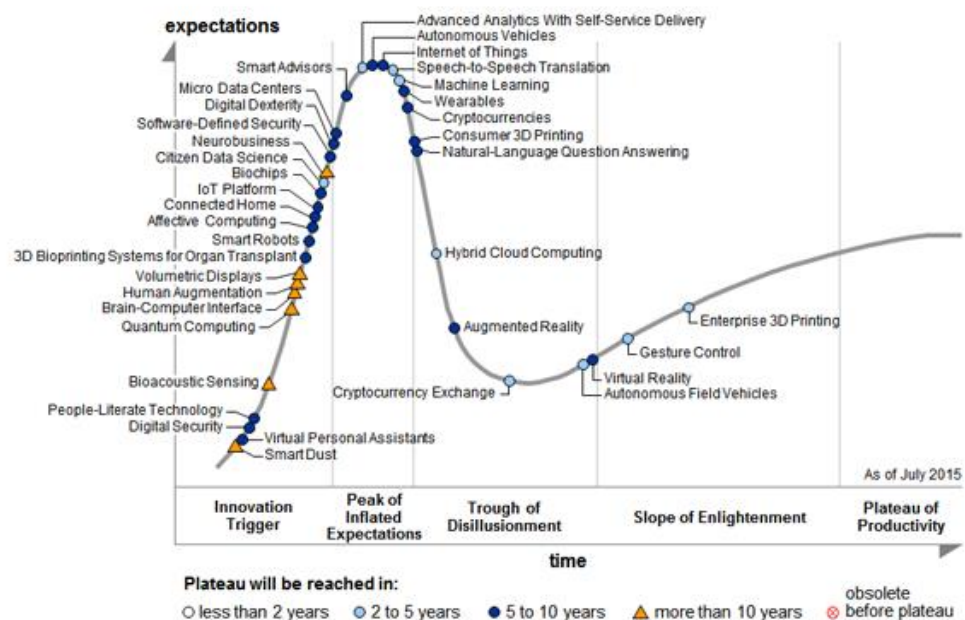


Figure 3: Gartner's hype cycle for 2015 [3]



One application of the combination between the IoT and Big Data technologies is the creation of Smart Cities in order to make cities smarter, more efficient, environment friendly and understand the city in order to provide better services to the citizens. This also means new services and efficiency actions for the companies that offer services to the cities as Ferrovial, in order words, new revenues from the new services and reduction of costs from the efficiency actions based on the use of data.

STATE OF THE ART

The rise of the Internet technologies and data collection are really transforming the way we live in the cities and will change it even more in the years to come. Using the Internet of Things, the environments we live in are becoming “alive”, making them interact between them or with the citizens living in it. The transport inside the cities, public services such as waste collection, street lightening and energy utilization are being transformed into data-driven decisions and actions (see Figure 4).

The combination between Internet of Things, for data creation, and Big Data technologies, for data processing and storage, can help the companies in charge of the smart city’s services use its resources more efficiently such as waste collection route optimization or lightening the streets according to the people present in it. Other examples of smart city services may include energy mapping throughout the city to better understand electrical demand or studying people’s movement across the city to maximize the use of bicycle and foot paths.

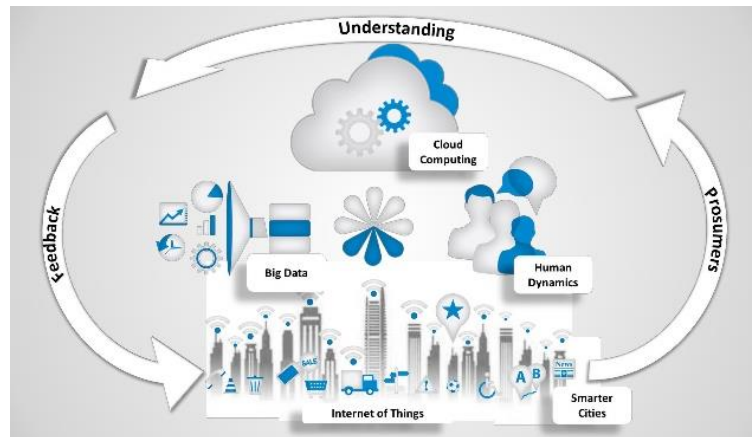


Figure 4: Big Data/ Smart Cities ecosystem [4]

The city of Songdo (South Korea) has been built entirely with embedded technology since its foundations in order to create a truly smart and connected city. The city is currently being developed by a consortium consisting of Cisco, 3M, Posco E&C and United Technology and started in the year 2000.

By giving to almost any device, building or road a microchip and wireless sensors, the city will have intelligent public services such as street lights that automatically adjust to the number of people in the street or a terminal inside each home, connected to the systems monitoring the public infrastructure and transport. The amount of traffic inside the city is tracked using RFID tags, sending geo-location data to the servers powering the city and giving feedback to the citizens about congested areas and optimized routes taking real-time traffic data into account. Every activity inside the city of Songdo generates data, garbage disposal will be done using smart cards and children will wear bracelets in order to be easily found if they get lost.

The smart energy grid can measure the presence of people in a particular area in a particular moment and can accordingly adjust the street lights. For example, the smart grid will ensure that areas that are scantily populated will automatically have some of the street lights turned off. This results in energy savings.



Sensors fitted in the roads will measure the total traffic at different times of a day and the total emissions. The data is sent to a central unit which will coordinate with the traffic lights. Traffic can be managed or diverted along other less congested areas to reduce carbon emissions in a particular area.

Smart traffic lights and smart grids represent two of the examples of big data applications serving smart cities. (Eiman Al Nuaimi).

The creation of the Korean city of Songdo rose the social concerns and mistrust about monitoring the citizen's daily activities and processing the data in the cloud. The people interact with the city's devices through their own smartphones and wearable devices and will reveal some details about their personal lives and daily habits such as their destinations throughout the city, people they get in touch with, daily routines and habits... The society, although it loves the new era of technology, may not be fully ready to give more details about their personal lives in order to receive better services inside the cities. (Michelle Selinger)

The analytics and real-time monitoring of the city of Songdo are performed using Cisco's tools for machine learning and cognitive computing.

Another example of a fully working smart city is the city of Glasgow, Scotland. The city hall spent more than twenty million pounds over the past two years in order to become the smartest city on the planet.

The city has become an example as far as data transparency inside the city is concerned, having more than 400 datasets on their open data portal. Intelligent street lights that adjust their brightness taking into account the city's activity, reporting their own faults and measuring air pollution inside the city. The traffic lights inside the city are regulated using real-time data from sensors under the roads in order to reduce bottlenecks. The city collects health records, figures for footfall in each street, demographics, air pollution measurements... all of which the city uses to plan ahead, for business developments, schools and health services. There are no social complaints about Glasgow's technological development as all of the



data is completely free for everyone who wants to take a look at it and no personal records are stored.

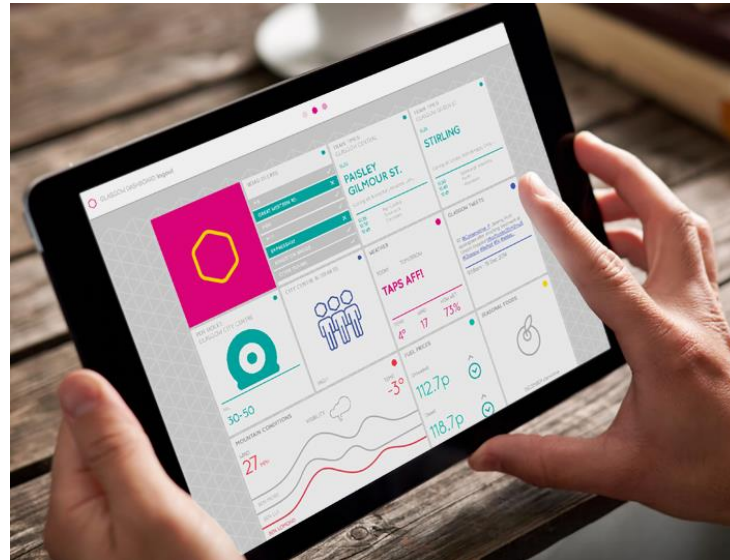


Figure 5: Example of a custom open data dashboard for the city of Glasgow [5]

The IoT will be, in a very near future, a reality and is transforming the way cities can be managed. It helps to better understand the behavior of the city and the citizens living in it. By feeding huge amounts of data, collected throughout the whole area, to Big Data tools and analytics it enable the city's services to be data driven in order to gain efficiency and an improved experience for the citizens. One of the most important parts of the union between IoT and Big Data for smart cities is the cooperation of people inside the city, the platforms deployed across the smart cities mentioned include smartphone applications to interact with the city and without a good use of those tools by the users, the insight brought by analytics wouldn't be as rich as it is in Glasgow for example, where the society response was incredibly positive.



PROJECT'S MOTIVATION

The combination of the Internet of Things and Big Data technologies is starting to redesign cities across the globe nowadays and seems to be the path to be followed by many cities in the years to come.

In order to keep up with technology and the competition, Ferrovial has decided to take their city management services to a new level, making use of the mentioned combination between internet of Things and Big Data. In order to do so, the company has deployed through some of the major cities of Spain, sensors related to the services it offers inside those cities (for example sensor inside the waste containers in Barcelona).

Ferrovial acknowledges the opportunity to gain insight about the cities it has contracts with and, in order to improve those services and create new ones, has created a new DataLAB department to carry out projects related with Big Data, Internet of Things and advanced analytics. The company doesn't want to externalize these services yet because the company wants to learn about the new technologies on offer in the market nowadays and have a better control over the insight brought by the data created inside the company thought the services that provides as in the cities, roads, constructions and airports.

OBJECTIVES

The main objective of the project is to develop a platform to collect, process and visualize the city's real time and historic data. The data concerning the city will be accessed through APIs in the case of IoT networks and cloud stored CSV files in the case of Ferrovial's services data. The platform should also gather data from other sources other than the city's sensors (i.e. social networks, open data portals, etc.) in order to enhance the model and the quality of the information provided.

The platform will analyze at three different levels:



Descriptive analytics: After a data ingestion process (ETL process to get, transform/clean and load the data), the user will be presented with the data evolution in order to detect trends and/or cyclic behaviors across the data. This type of analytics summarizes what happened – more than 80% of current business analytics are descriptive ones. This allows us to have discussions with the business to plan next iterations as first conclusions appear easily.

Predictive analytics: The next step. In this case, by using a variety of statistical modeling, data mining, and machine learning techniques, the platform should be able to forecast the city's behavior in order to improve the services' efficiency and be able to have new data based in the one we have to have the chance to analyze new variables. Big data technologies represent a breakthrough regarding the amount and the speed of data that can be introduced into the algorithms. The project will seek to predict waste levels in Barcelona's containers in the next three days.

Recommend data driven decisions to the different business areas using the information acquired in the 2 previous analytics layers and be able to prescribe an action that will be the third layer – prescriptive analytics. The platform developed in the scope of this project will make use of the predictions made in the second layer of analytics to optimize garbage trucks' routes and thus, reducing the amount of fuel used and increasing the efficiency.

The platform should perform the three layers for different cities where Ferrovial offers its services enabling the user to compare information about different cities in a unified manner.

METHODOLOGY

As the project can be categorized as a “Software Development Project” it will be carried out using the agile methodology in order to have the final client review the product multiple times before its completion and thus, having its feedback on the product as soon as possible to ensure the product is developing what the client really wants.



Being able to create discussions with the business based on data results as soon as possible and make the business participate in the definition of the final platform is a strategic goal behind the agile methodology.

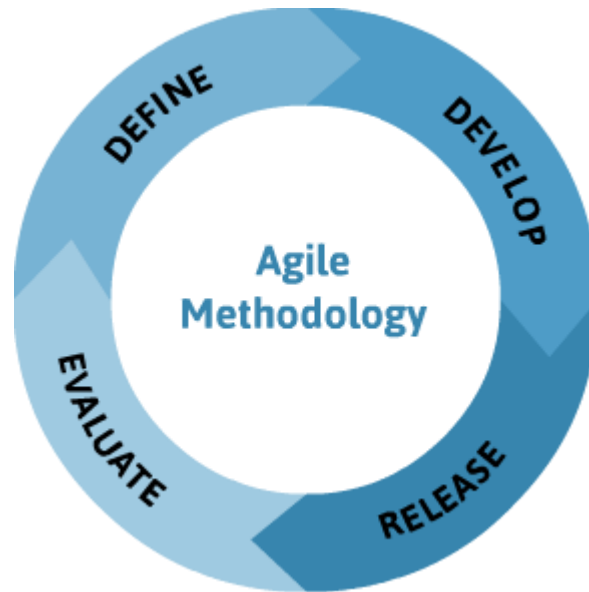


Figure 6: The Agile cycle

The Agile methodology is very common in the Software Project Management as it enables the team to provide value as soon as possible. Instead of committing to yearlong projects, the agile methodology suggests to work in “Sprints”. Each Sprint having a duration about three weeks and no more than four, has unchangeable objectives defined at the beginning of the Sprint and the development has to commit to them, nothing more and nothing less. Once the Sprint is finished, the “product” is released for the customer to see the evolution and the Sprint is evaluated in order to think about possible improvements to add in the next Sprint.

In this project the first sprints are organized by cities and analytics layers, the first one dedicated to the city of Santander creating the base for the description layer, being followed by Murcia, Barcelona and Madrid and getting also the predictive and prescriptive layers in following sprints.

As seen in Figure 7, iteration1 tasks correspond to the first layer of analytics: Retrieve the data for the corresponding city, cleaning it and inserting it into the database with the correct formatting. As a result the platform will be gathering real



time and historic data for the specific city to which the sprint is dedicated to where is possible to start the discussion about first view trends and relationships between sensor information and service operation.

In the second iteration for each city, the data visualization results from the first iteration will be improved upon the conclusions obtained from the previous iteration. Deeper analytics will also be performed in order obtain more insight about the city's behavior and make data driven decisions about the services offered in the city affected.

RESOURCES

The main resources used to carry out this project are:

- Agile methodology.
- Cloud: AWS is the chosen cloud platform to use. Amazon became the first company to really invest in the cloud platform market and is nowadays the most affordable, reliable, supported and complete platform. It keeps adding new services to their outstanding platform such as the Machine Learning Service, or the Internet of Things platform.
- Processing / Big Data:
 - Amazon RDS instance running a PostGreSQL database to keep the data.
 - Elastic Mapreduce: Cluster of machines that perform parallel processing of the raw data in order to clean and transform the data to fit the data model in the database.
 - Amazon EC2 instance: A machine will act as the server hosting the visualization tool, reachable from any location. Another machine running Python in order to perform the data acquisition and parsing will also be used.
 - SQL and Python programming languages to design ETL
- Predictive analytics – still in definition



- FIWARE APIs: In order to retrieve and collect data about some cities, for example Santander, we will make use of FIWARE's REST API to connect with the smart city's controller and retrieve the last data from all the sensors.
- User View: Qlik Sense Enterprise: In order to present the results and analytics to the user, the visualization tool chosen is Qlik Sense in the server, or Enterprise, version. The tool enables the developers and analysts to transform data into a responsive dashboard containing charts, maps or KPIs indicators very easily.
- Documentation and SW version control: internal SVN server enabling the whole team to share documents and retrieve previous versions if needed. Wiki is also use to exchange opinions and information regarding the different topics.



CHAPTER 2 SMART CITIES PLATFORM

INTRODUCTION

One of Ferrovial's main objectives when launching this project is to have a common platform to make all smart cities' data available to the different areas whom might be interested inside the company.

Ferrovial manages city services in different cities across the whole Iberian Peninsula (e.g. Madrid, Barcelona, Santander, Murcia...) but there is not current connection between the teams responsible for each the cities.

Centralizing the data related to city services in one unique platform might bring new insights to the company as the ability to compare cities and therefore, understand why some decisions taken in a particular city might affect another or simply help understand all the cities as a whole.

PLATFORM DESIGN

INFRASTRUCTURE

The whole platform is to be cloud-based, using Amazon Web Services as the cloud services provider. The two main parts of the platform are, as for any web based application or platform, front-end and back-end.

Back-end

The Back-end of the platform consists of a PostgreSQL database engine running on an RDS (Relational Database Service) instance inside Ferrovial's Virtual Private



Cloud. The relational database storing all the data for the platform should be accessible by three types of machines:

- The machines collecting the data for the platform: Machines launching processes to collect data from APIs or drop areas inside the VPC.
- The machines hosting the front-end of the platform. The visualization layer of the platform will read the data from the database in order to create graphs and charts to show to the user.
- The machines responsible for data analysis and modelling. The instances performing the predictive or exploratory analytics need access to the data.

The main goal when designing the database schema resides in the universality of the latter, having to host data from a variety of sources, formats and relations.

Front-end

The front-end of the platform represents the visualization layer presented to the users of the platform, in this case, the analysts responsible for the cities' are inside Ferrovial. The platform uses Qlik Sense, a visualization tool enabling very easy development and usage of interactive and connected sheets of visualization.

The tool implements a credentials system in order to give different levels of access to different users, for example, the responsible for the whole are should have full access but a junior analyst staffed in the Santander project should only have access to the data related to his project. All these security measures are available through the administrator console and are changed very intuitively.

Analysis

The different analytics performed on the smart cities database are performed in Amazon EC2 instances inside Ferrovial's VPC. The different machines used to perform exploratory analysis, database validation and predictive modelling are hosted in the cloud and accessed through ssh from the local computers in the office in Madrid. The different analysis performed in the project are built with python (see Chapter 4), R (see Chapter 3) and SQL.



ONLINE ANALYTICS PLATFORM

The first module of the platform consists of an “Online analytics module”. This module contains the latest measurements for the city chosen. In the case of the city of Santander it can be almost considered as a real time platform as the data is refreshed every ten minutes at most. The module always has available for the user the latest information regarding the cities he’s interested in.

The main role of this module is to enhance considerably the reactive behavior the company can have towards changes in the city’s behavior or anomalies in the system.

An important topic concerning the city of Madrid for example, this year the city it was forbidden for cars to park inside the city due to the high levels of pollution in the air. The use of this platform could improve considerably the ability of the company to react to meteorological changes of this kind and thus, lowering the effect of the measures taken as the problem would have been detected before.

Another example of an interesting use of the platform is the detection of anomalies in the sensor network. For example, in the city of Barcelona where Ferrovial is responsible for waste collection, a sensor could register a very unusual high temperature inside a container, possibly meaning the existence of fire inside the container. Instead of discovering it the next day when the garbage truck passes to pick the waste up, the platform could raise an alarm and dispatch a team in order to save the container from burning and save not only the container but the sensor inside it.

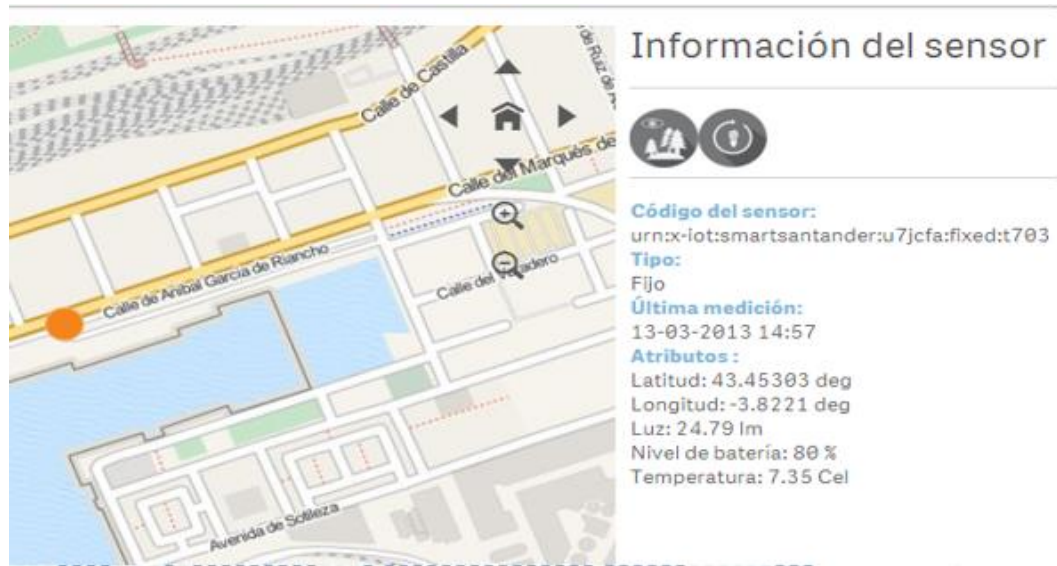


Figure 7: Illustration of online platform for the city of Santander

DEEP ANALYTICS PLATFORM

The main module in the created platform is the deep analytics module, hosting historical data for all of the cities inside Ferrovial's network. The module shows the user descriptive analytics concerning the historical data of the measurements regarding the selected city. The descriptive analytics presented to the user consist of summary statistics and visualizations bringing new insights to Ferrovial's analysts as the amount of data inside the platform amounts to several million of measurements and was never analyzed together due to computation limitations existing previously inside the firm.

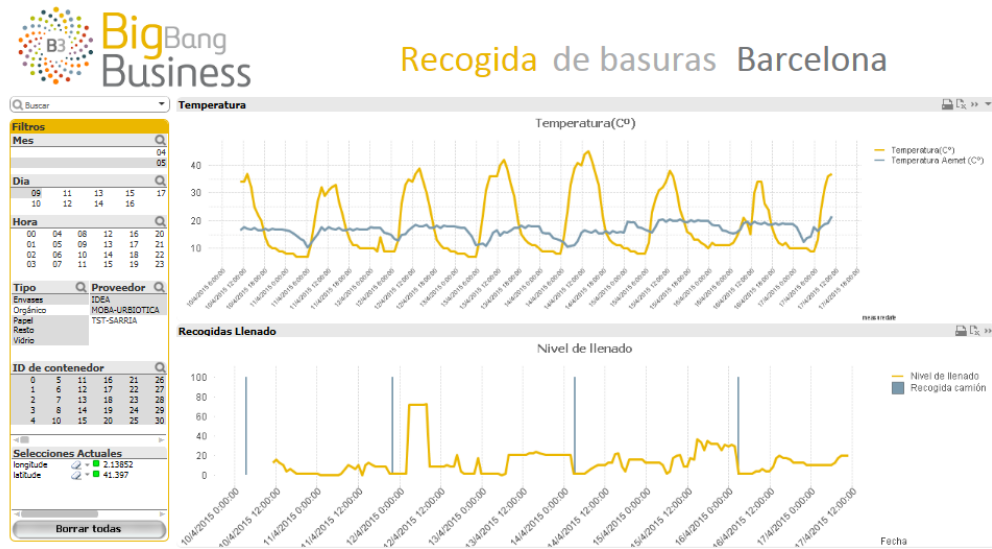


Figure 8: Illustration of the deep analytics platform

The users are able, not only to visualize and detect patterns in data they could never analyze together, but also to perform aggregations and statistics on historical data never available to them before.

The real innovation and distinctive value is brought to the platform by the ability to perform machine learning or advanced analytics on the historical data of the cities. This approach enables the analysts to link to areas that were very separated until very recently. Using big data technologies, the historical data is not only stored, but read from the machines performing the analysis at a very high throughput (up to 10 Gb/s for Amazon Web Services) and analyzed using clusters of computers in order to parallelize the operations to perform on the dataset.

RESULTS

Using the Online analytics platform, the analysts working in the different cities where Ferrovial provides services, can detect and react very rapidly to any anomalies which may occur.

Moreover, using the Deep analytics module, the data regarding the city of Santander was collected from August 2015 to February 2015. 1200 sensors measuring up to



54 different variables in the city measured more than 150 million observations. The data was analyzed using the platform and qualified as unready to be used for any kind of predictive analytics due to various reasons regarding the maintenance of the API serving the data:

1. Lack of documentation regarding how to access the data and subscribe to real time events.
2. Almost 20% of sensors not measuring since 2014 or 2013.
3. Measures concerning the status of parking spots always measuring True (100% of the times).
4. Sensors present long times of inactivity not measuring anything and starting to measure again for another two or three months (Figure 9).

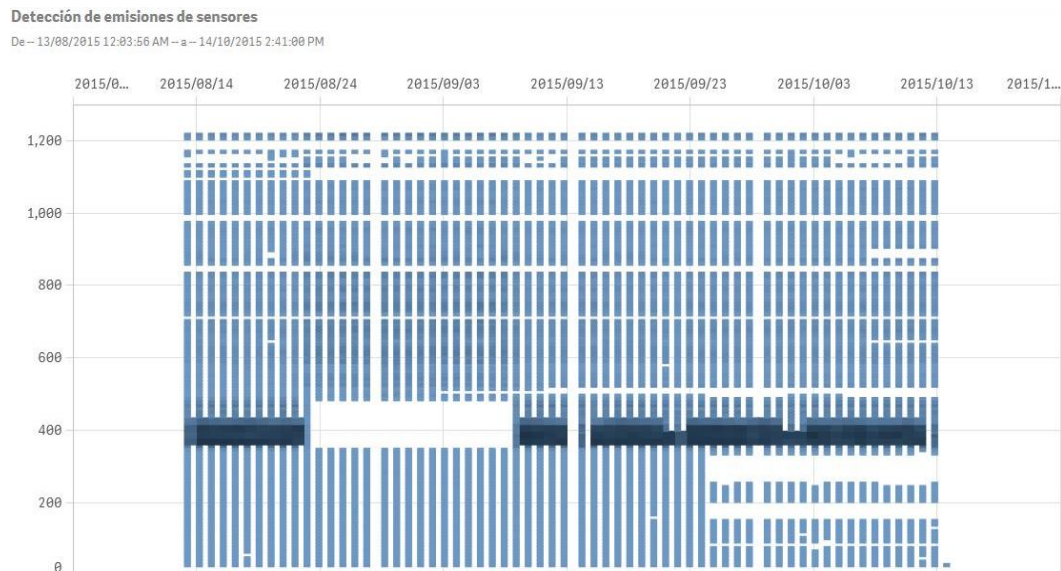


Figure 9: Sensors showing inactivity

The next city in the project roadmap was the city of Barcelona where Ferrovial, as mentioned previously is responsible for the waste collection.

The information generators in this case is a network of 303 sensitized containers in the neighborhood of Sarriá measuring every hour the temperature inside the container and the fill level of the latter.

In this case, the measurements' quality were far better than the previous case and were deemed fit to a predictive analysis.



The rest of the project focuses on the predictive analyses carried out: the main goal is to optimize how Ferrovial handles the waste collection routes. The optimization of the routes is divided into two main categories: first, predicting which container will be full the next day and then optimizing the route followed by the truck in order to only pick up the containers flagged as full following the shortest available.

The prediction phase was first regarded as a classification algorithm labelling whether a container will be full or not (0 or 1, binary classification) and in a second approach, a regression algorithm was implemented in order to estimate the fill level of every container in the network.



UNIVERSIDAD PONTIFICIA COMILLAS
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)
INGENIERO INDUSTRIAL



CHAPTER 3 CLASSIFICATION

INTRODUCTION

As mentioned in **Error! Reference source not found.**, the first approach is a classification algorithm. The level of the containers is threshold in order to obtain a binary target variable representing whether the container is considered full or not. The binary target variable is the one the classification algorithm will predict in order to determine each day, if the container will be full (target variable is 1) or not (target variable is 0).

The classification problem is coded in the R language, using the RStudio IDE and the MySQL database instance described in **Error! Reference source not found.**

DATA PREPARATION / WRANGLING

The data concerning the level of the containers resides in a MySQL instance hosted in the Amazon Relational Database System service inside Ferrovial's Virtual Private Cloud.

After downloading the whole dataset, using a SQL connector for R, some operations need to be performed in order to obtain a tidy dataset (Hadley Wickham, Tidy data):

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

In order to achieve such a dataset, in the scope of this problem, the main transformations to carry out are:



Calculate laborality from measure date instead of having the full date as an input for the algorithm. The date laborality is obtained from the weekday, grouping into three factors:

1. Working days: Monday – Friday
2. Saturday
3. Sunday

The behavior of the citizens is supposed to be alike for the first five days of the week when most of the people has to go to work and perhaps take out the trash at night or early in the morning whereas the Saturdays and Sundays behave differently than this previous days.

A new binary variable indicating whether the day of the measurements is a national/regional festivity because people is supposed to behave differently regarding the trash in this type of days too.

The containers having less than twenty measurements per day (maximum is twenty four as they measure once every hour) are dropped from the model as they lack important information that other containers have.

An important synthetic variable for this problem is the average filling level for each container up to the previous weekday (e.g. if the measurements are on a Monday, the average filling level up to the previous Monday).

An important aspect to take into account in the scope of this project is the fact that the measurement of the container levels are being distorted by the garbage trucks that pick up the garbage each and every day.

In order to mitigate this effect, the variable taken into account is the difference between consecutive measurements instead of the actual value measured. When a strong negative increment is detected (the garbage was picked up), the incremental value is calculated as the average of the previous and the next positive incremental values.



By doing this, the uncertainty concerning the pick-up trucks is removed from the model and is not taken into account in the analysis as it would only add noise and distortion.

Once all the previous operations are performed, the input variables for the classification model are described in table.

Feature	Description
Month	Month of the measurement (1-12)
Day	Day of measurement (1-31)
Laborality	Laborality (W, Sat, Sun)
Hour	Hour of the measurement (0-24)
Container type	Type of container (R, V, P, O)
Inc_prev	Average of previous increments

Table 1: Classification features

MODELING

The algorithm chosen to try to solve the Classification problem is a Support Vector Machine as it is a very versatile algorithm and the implementation in R (under the library e1071) is very fast to train and make predictions.

SUPPORT VECTOR MACHINE

The support vector machine maps the original space in which the observations are located to a much higher-dimensional space where, presumably, the separation between the two classes becomes easier to perform using a linear classification boundary.

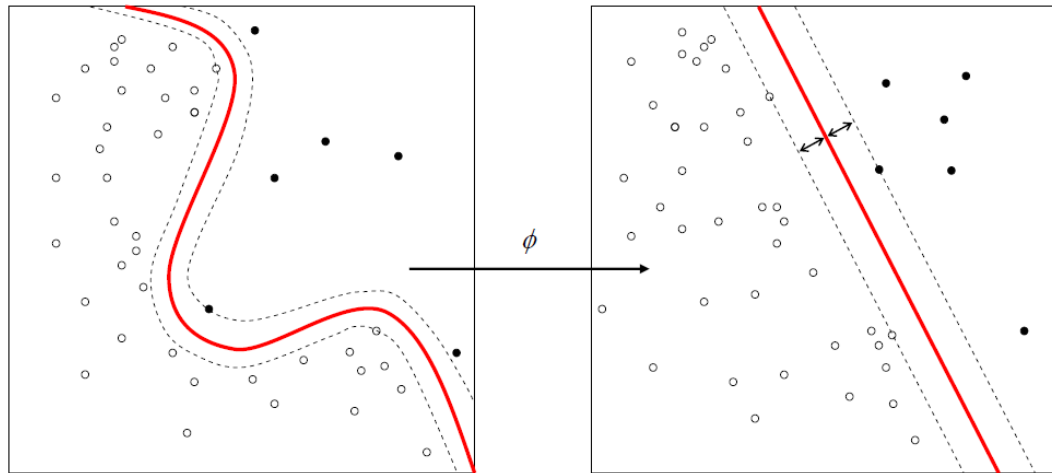


Figure 10: Kernel mapping

The performance metric to be used in the scope of this project is the Area Under the Curve (AUC) of the Receiver Operation Characteristic (ROC).

When a binary classifier predicts, four scenarios are possible:

1. The algorithm predicts 1 and the true label is 1: This case is called a true positive as the algorithm predicted a correct outcome of 1.
2. The algorithm predicts 0 and the true label is 0: This case is called a true negative as the algorithm predicted a correct outcome of 0.
3. The algorithm predicts 1 and the true label is 0: This case is called false positive as the algorithm predicted a positive outcome but the true label is not.
4. The algorithm predicts 0 and the true label is 1: This case is called false negative as the algorithm predicted a negative outcome but the true label is positive.

The main goal is to minimize both types of prediction error but depending on the use case one may be “less damaging” than the other. In the case of this project where the support vector machine predicts whether the container will be full or not, the false negative case is way more damaging for Ferrovial as the algorithm predicts an empty container when it is really full and the waste may start to pile up outside the container.



As the support vector machine really predicts the probability of an observation to have the label 1, varying the threshold to apply to the predictions we can directly affect the error rates.

The Receiver Operation Characteristic curve represents the true positive rate against the false positive rate at various threshold settings.

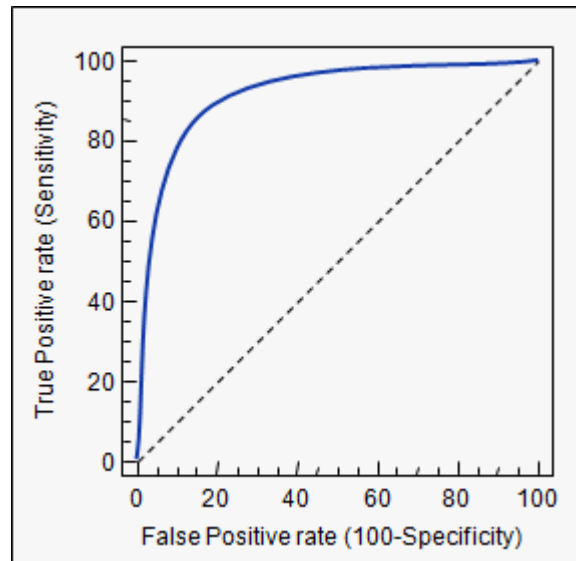


Figure 11: ROC curve example

Observing Figure 11, it seems logical that the best scenario is the top left corner where the true positive rate is 1 (all the labeled true cases are equal to all the real true cases) and the false positive rate is 0 (no positive case is predicted wrong).

The random case would have 50% accuracy and thus, the true positive rate would be equal to the false positive rate no matter the threshold. The random case is represented by the diagonal line in Figure 11 and the algorithm should aim to, at least, perform better than random guessing.

Once the curve is plotted, a measure of the performance is the area under the ROC curve as an area of 1 can only be achieved in the best case scenario described above and 50% of the area is achieved by random guessing.

The support vector from the R library e1071 (used in this project) machine needs to tune the following hyper parameters:



- Kernel: the type of function to map the actual observations to a higher dimensional space in order to make the data linearly separable. Choices are: linear, polynomial, radial or sigmoid.
- Gamma: Can be seen as the inverse of the radius of influence of samples selected by the model as support vectors.
- C: C for cost, represents the regularization term in order to force the model to be simpler, avoiding over fitting.

RESULTS

The Area Under the Curve of the support vector machine classifying whether a container will be full or not for each hour of the day is 75%.

The AUC performance measure can also be re-scaled in order to make the random case equivalent to 0 instead of 50%, making it more intuitive. This re-scaling corresponds to the Gini index and is calculated by performing:

$$Gini = 2 * AUC - 1$$

The Gini index for the support vector machine in this project is 50%.

The classification model is scalable as we have only one model for the 303 containers but may be also losing some details as it has to deal with a lot of variance from all the different types of containers.



CHAPTER 4 REGRESSION

INTRODUCTION

The second approach to the garbage level prediction is a regression approach. Instead of classifying if the container is full or not, the intention in the new approach is to estimate the average increment for a specific container for a specific day and a moment of that day.

The problem radically pivots in this new approach from a binary classification to an estimation of a continuous variable.

DATA PREPARATION / WRANGLING

The dataset needs to be cleaned and prepared to obtain a tidy and prepared dataset for the regression problem.

The first change implement is to round the value of the measurements in order to decrease variability without losing too much information. The measurements are rounded to a resolution of 5%.

Negative increments are assigned with a zero value as a pick up is supposed or error in the measurement.

Date variables are created in order to avoid the inclusion of the date variable in the model. A binary variable describing whether the date corresponds to a festive day or to a weekday (0: Monday, 6: Sunday).

Again, in order to decrease variability in the data, measurements are grouped into eight different moments of the day. The data shifted from an hourly format to an aggregated state. The loss of information is not critical and the loss in variability is considerable.



The last variable to be created is the average of three previous increments for a specific container for the same day of the week (e.g. if the measurement is a Monday in the period x of the day, the average of the increments of the three previous Mondays on the same period is computed).

After cleaning and preparing the dataset, the input of the model consists of six different variables:

Feature	Description
Month	Month of the measurement (1-12)
Day	Day of measurement (1-31)
Day_period	Moment of the day (1-8)
Weekend	Binary indicator (0-1)
Weekday	Day of the week (0-6)
Inc_prev	Average of previous increments

Table 2: Regression features

The final dataset is finally divided into two different sets: training and test set. The training set corresponds to the 67% of the tidy dataset and the split is done randomly in order to assure robustness.

The training set will be used to train the models using 10-fold cross-validation for hyper-parameter tuning and the test set will be used to evaluate model performance as it represents observations the model has never seen.

MODELING



The regression task can be done using an infinity of different models and combination of parameters, in the scope of this project, three models were tried in order to choose, using cross-validation, the best model and set of hyper parameters.

KNN REGRESSOR

The KNN Regressor is an approach based on the fact that very similar points will behave similarly with regards to the target variable.

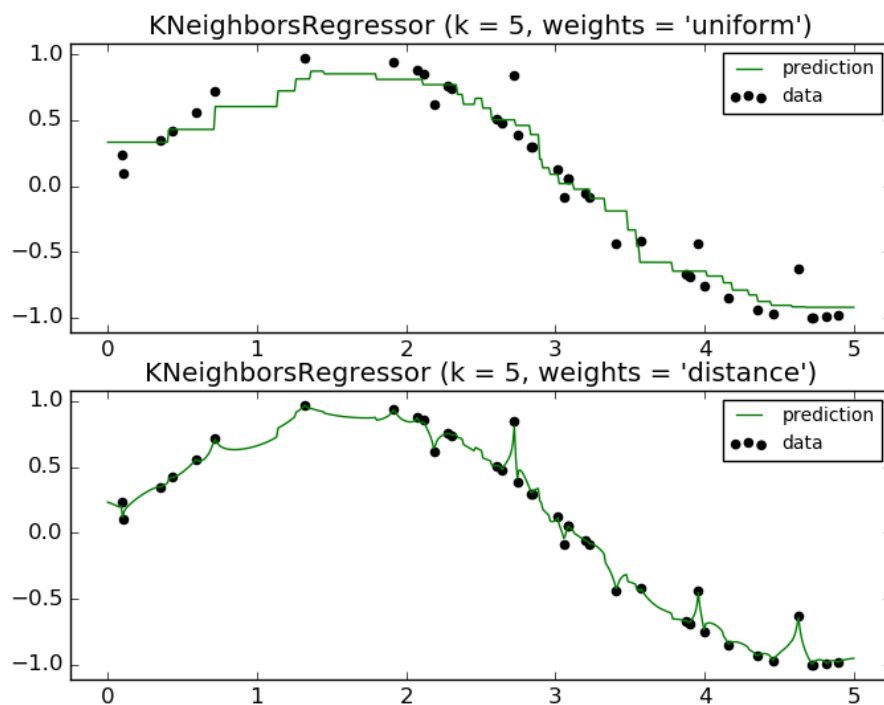


Figure 12: KNN Regression example

The performance metric used to estimate whether a model is better or if a set of parameters is more appropriate is the RMSE (Root Mean Squared Error).

$$RMSE = \sqrt{\frac{\sum_{t=1}^n (\hat{y}_t - y_t)^2}{n}}$$

In the equation, \hat{y}_t represents the prediction while y_t represents the real value of the target variable in the specific observation, t. The measure represents is in the



same units as the target variable, how much does the model deviates, on average, for every observation.

The hyper parameter tuning of the model is performed using cross-validation on the training set. The parameters to be tuned for this model are:

1. `n_neighbors`: the number of neighbors to use for regression.
2. `Weights`: weight function used in prediction. Possible choices are:
 - Uniform: all points are weighted equally
 - Distance: The inverse of the distance is used to weight the observation.
 - User defined function.

The hyper parameter tuning is performed using a grid search across the two possible dimensions and cross-validation on the training set in order to get an unbiased performance metric.

The values for the `n_neighbors` hyper parameter are from 2 to 5. The estimated value is obtained by calculating the mean of the `n_neighbors` weighted target variable's value.

The results obtained with the KNN Regressor are presented in table.

Average training set RMSE **1.35**

Average test set RMSE	2.00
------------------------------	-------------

Table 3: KNN Regression results

DECISION TREES

The decision tree algorithm represents the second approach taken towards the regression problem. It learns simple decision rules on the dataset in order to infer the value of the continuous target variable.

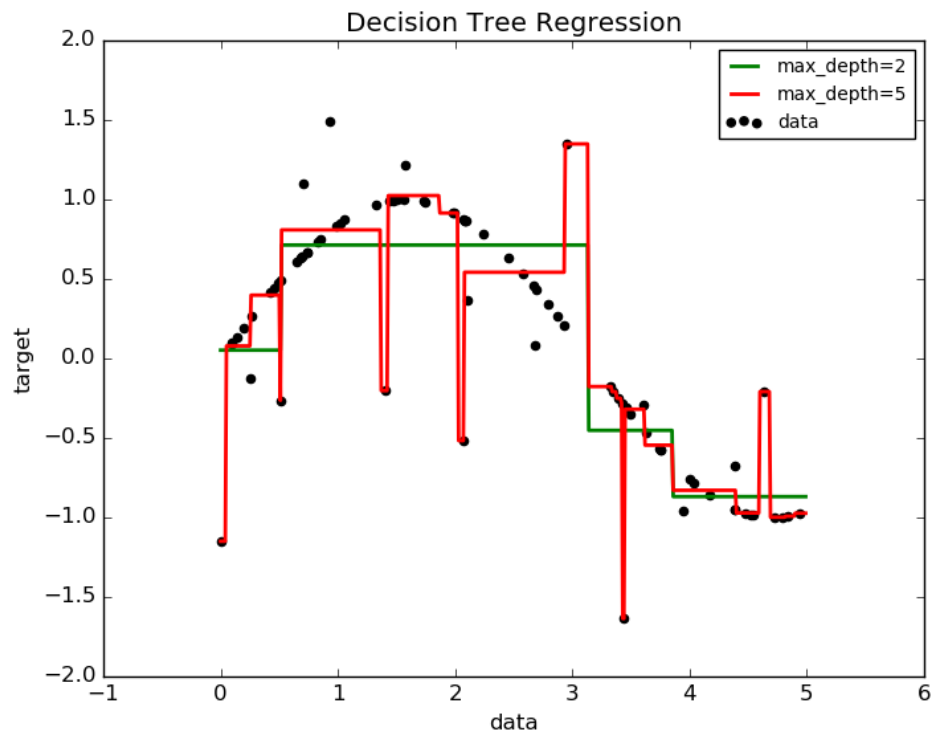


Figure 13: Decision tree regression example

The performance metric is again the RMSE, in order to be capable to compare the performance of both algorithms to decide which one to implement.

Again, the hyper parameter tuning of the model is performed by using a grid search and cross-validation in the training set.

The parameters to be tuned for this model are:

1. Max_depth: The depth of the tree, more depth will be able to catch more details but may tend to overfit the training set.
2. Min_samples_split: The minimum number of samples required to split an internal node.

The results obtained with the Decision Tree Regressor are presented in table.

Average training set RMSE **1.47**

Average test set RMSE	2.3
------------------------------	------------

Table 4: Decision tree regression results



GRADIENT BOOSTING REGRESSOR

The Gradient Boosting Machine (gbm) fits an ensemble model to the training set in order to estimate the incremental target variable.

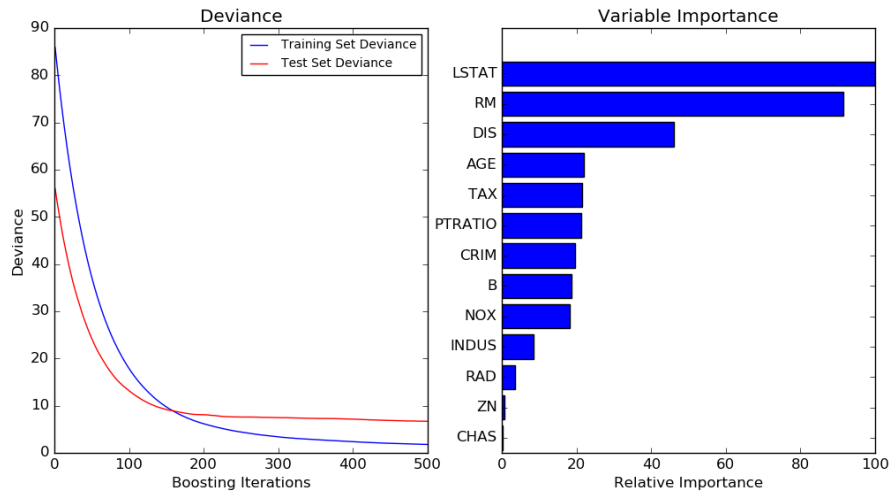


Figure 14: GBM Regression example

Again, the performance metric is the RMSE and the hyper parameter tuning is performed using cross-validation in the training set and grid search.

The parameters to be tuned for the model are:

1. `N_estimators`: The number of boosting stages to perform.
2. `Learning_rate`: The shrinkage of the contribution of each tree.
3. `Loss`: loss function to be optimized.

The results of the Gradient Boosting estimator are shown in table.

Average training set RMSE **1.2**

Average test set RMSE **1.9**

Table 5: GBM Regression results



RESULTS

The results for the regression approach seem better than the ones for the classification approach as not only the models are better on average, but the granularity of the information is bigger, the result of the prediction is the actual level of waste of the container and not whether the container will be full or not. The regression approach is one step more rich in information gained but has the drawback of not being scalable at all. Each container has its own model trained exclusively and as the project expands to the whole city and the number of unseen containers grow, the implemented models will be unable to make any predictions and new model would have to be trained, pushing the possible insights deriving from the analysis back in time.



UNIVERSIDAD PONTIFICIA COMILLAS
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)
INGENIERO INDUSTRIAL

CHAPTER 5 ROUTE OPTIMIZATION

INTRODUCTION

Once the predictions are calculated (using the classification or regression approaches), they are stored in the platform and presented to the user. The user sees a comparison between the route followed by the truck when picking up all the containers and the optimized route when picking up the containers predicted to be full by the algorithm,

ROUTE OPTIMIZATION

The route optimization is an implementation of the travelling salesman problem. An optimization problem consisting of travelling a graph from start to bottom, visiting all the nodes minimizing the total distance travelled.

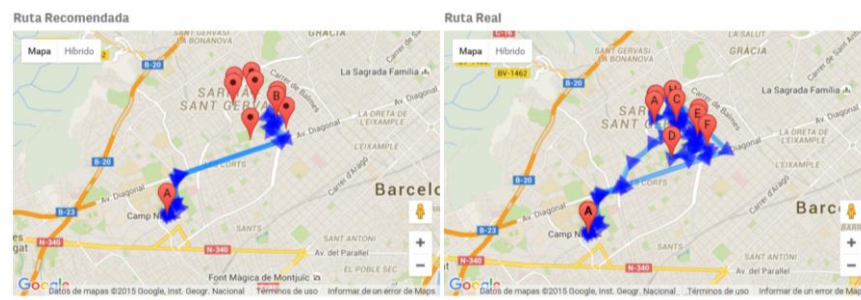


Figure 15: Comparison of optimized vs. normal route

The implementation of the algorithm as well as the representation of the optimum route is provided by Google Maps through an Application Programming Interface (API). The platform ingests the predicted values for the containers, thresholds



which ones among them will be labelled as full for the next shift and then sends a requests to Google Maps servers. The information returned by the API is a map containing Google Maps features (zoom, movement...) and the calculated optimum route printed inside.

RESULTS

The optimum route algorithm was not implemented in the scope of the project as it was not one of the objectives and the problem, being an NP-hard problem would require a timeframe not available in the project.

The API provided by Google solves two issues at the same time, the route calculation and representation but presents one major drawback as it has a limitations in term of number of points to calculate the optimum route (50 is the maximum).



CHAPTER 6 CONCLUSIONS

Throughout the project, three main deliverables were created: a full analytics platform enabling Ferrovial to gather, store and analyze in the same tool all its data concerning Smart cities, a classification algorithm capable of predicting for every container whether it will be full or not for every hour of every day and 303 regression algorithms capable of predicting, for a given container the rounded incremental of waste for every moment of the day. The results of the created algorithms are presented to the user using the Google Maps API through the deep analytics platform.

As the current technological revolution takes place, the companies nowadays have to seek for innovative and unseen ways of creating or maintaining their competitive advantages. One way to do so is to use the power of the new computational technologies to deploy advanced analytics inside the company. This project represents a first step inside Ferrovial's strategy to do so and a proof that innovation in the strategic decisions of this era is crucial to maintain advantage.



UNIVERSIDAD PONTIFICIA COMILLAS
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)
INGENIERO INDUSTRIAL



CHAPTER 7 FUTURE DEVELOPMENTS

The project has fulfilled the objectives and all the different connecting parts inside it are stable and robust but can be widely developed.

The future developments of the project are to be brought mainly in two areas:

- Algorithms
- Platform

ALGORITHMS

The accuracy obtained by both classification and regression algorithms is acceptable, even outstanding in some cases but an optimizer could be used to find the best set of hyper parameters as a grid search covering all the possibilities is computationally impossible (genetic algorithms, Bayesian optimizer...).

In one hand the classification algorithms presents a problem of being too general and not being able to capture all the details for all the container as there is only one algorithm trying to capture all the variance brought by the 303 containers. On the other hand, in the regression step, the approach is unscalable as each container has its own trained model and adding new containers as the pilot project expands to the whole city of Barcelona the amount of training and predictions to handle will become unmanageable.

One interesting future development is to find an equilibrium in the tradeoff created by the two different approaches, clusterize the containers in order to create groups of containers which behave very similarly and train a model for each of the clusters. This will not only reduce considerably the number of models to train and maintain but also resolve the problem of the new containers arriving as very quickly a cluster prediction could assign them to a specific cluster and since that moment, predictions for that container would be available.



One last improvement to be brought to the system is bagging. Instead of having just one model predicting the binary label for classification or the incremental waste level for regression, have different 10 models predicting and then averaging or taking the mode in order to get a single value. The bagging systems provides robustness to the models and very often and increase of model performance.

PLATFORM

The main development to be brought to the platform is regarding the optimum route calculation. The Google Maps API shouldn't be a solution for the future as the number of sensors grow, the API won't be able to solve the problem as it has a limited number of points. One possible solution would be to use the R library TSP which is developed exclusively to solve this precise problem.

The calculation would have to be performed on a powerful machine but this is not a problem as the platform is cloud based and can have any kind of machine up and running in less than a minute. The proposed solution would give the machine the prediction data and coordinates of the containers, perform the calculations to achieve an optimum route and pass the coordinates of the route to the deep analytics platform to present to the user. The data exchanges wouldn't represent a problem as the throughput inside Amazon Web Services is very high.



PART II ANNEXES



UNIVERSIDAD PONTIFICIA COMILLAS
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)
INGENIERO INDUSTRIAL

ANNEX 1 KNN ALGORITHM

The k-Nearest Neighbors (kNN) algorithm follows the maxim “things that are alike are likely to have properties that are alike”. This approach is used in machine learning for both classification and regression. The main advantages of the kNN algorithm are its simplicity and effectiveness (very fast training phase) but is computationally expensive (requires a large amount of memory).

The algorithm will be explained for classifications tasks as it is more intuitive and then the regression approach will be presented.

The kNN algorithm begins with a training dataset made up of examples that are classified into several categories, as labeled by a nominal variable. Assume that we have a test dataset containing unlabeled examples that otherwise have the same features as the training data. For each record in the test dataset, kNN identifies k records in the training data that are the "nearest" in similarity, where k is an integer specified in advance. The unlabeled test instance is assigned the class of the majority of the k nearest neighbors. The kNN algorithm treats the features as coordinates in a multidimensional feature space.

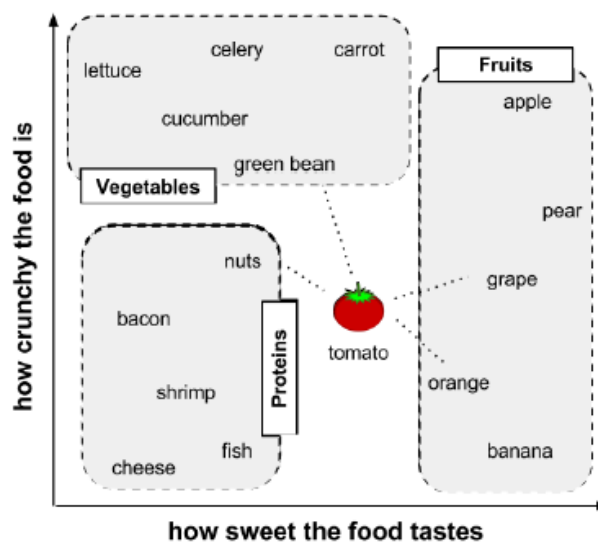


Figure 16: Illustration of kNN prediction [7]



When a new unlabeled observation is available, the algorithm computes the similarity between the new unlabeled observation and the observations in the training set. To measure similarity, a distance metric in the space defined by the features is used. Traditionally, kNN uses Euclidean distance but almost every implementation of the algorithm allows for distance selection as another metric may be suitable depending on the use case.

$$\text{dist}(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2}$$

Equation 1: Euclidean distance

After the distance calculation, the algorithm classifies the new unlabeled observation as the mode between its k nearest observations in the training set.

Deciding how many neighbors to use for kNN determines how well the model will generalize to future data. The balance between overfitting and underfitting the training data is a problem known as the bias-variance tradeoff. Choosing a large k reduces the impact or variance caused by noisy data, but can bias the learner such that it runs the risk of ignoring small, but important patterns.

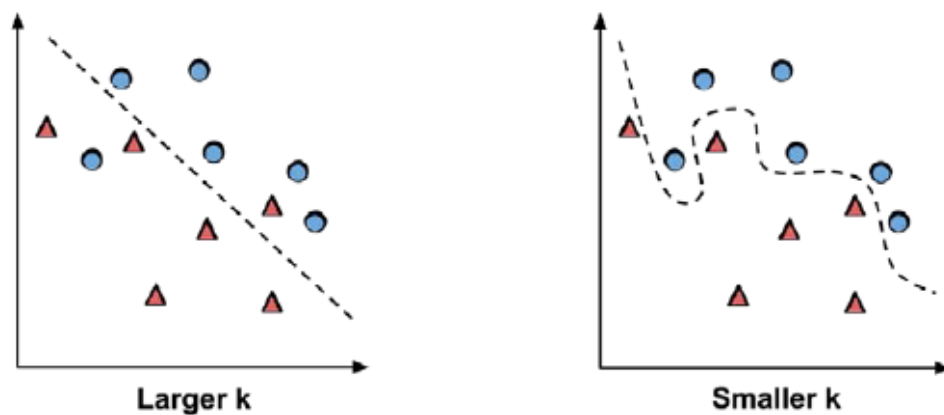


Figure 17: kNN bias-variance tradeoff [7]

Suppose we took the extreme stance of setting a very large k, equal to the total number of observations in the training data. As every training instance is represented in the final vote, the most common training class always has a majority



of the voters. The model would, thus, always predict the majority class, regardless of which neighbors are nearest.

On the opposite extreme, using a single nearest neighbor allows noisy data or outliers, to unduly influence the classification of examples. For example, suppose that some of the training examples were accidentally mislabeled. Any unlabeled example that happens to be nearest to the incorrectly labeled neighbor will be predicted to have the incorrect class, even if the other nine nearest neighbors would have voted differently.

Obviously, the best k value is somewhere between these two extremes. Typically the k parameter is set between 3 and 10 or equal to the square root of the number of observations in the training set. However, the best approach to tune the hyper parameter k is to use cross validation on the training set and choose the number of neighbors that gives the best prediction accuracy.

The k NN algorithm can also be used for regression, the only difference in the regression approach is the operation performed to predict the target continuous variable. In a classification task, the majority class among the neighbors is returned whereas for regression, the target variable is predicted by local interpolation of the targets associated of the nearest neighbors in the training set.



UNIVERSIDAD PONTIFICIA COMILLAS
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)
INGENIERO INDUSTRIAL



ANNEX 2 DECISION TREES

Decision trees are perhaps the single most widely used machine learning technique, and can be applied for modeling almost any type of data—often with unparalleled performance.

As for the kNN algorithm, the decision tree model will be explained for classification as it is more intuitive and then present the minor changes made to make it available as a regression algorithm.

Decision tree learners build a model in the form of a tree structure. The model itself comprises a series of logical decisions, similar to a flowchart, with decision nodes that indicate a decision to be made on an attribute.

These split into branches that indicate the decision's choices. The tree is terminated by leaf nodes (also known as terminal nodes) that denote the result of following a combination of decisions.

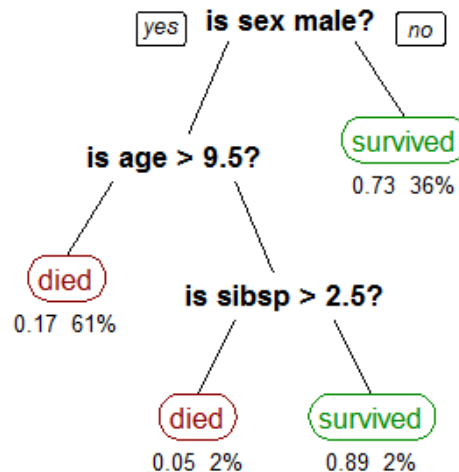


Figure 18: Decision tree example

Data that is to be classified begin at the root node where it is passed through the various decisions in the tree according to the values of its features. The path that



the data takes funnels each record into a leaf node, which assigns it a predicted class.

Decision trees are built using a heuristic called recursive partitioning. This approach is generally known as divide and conquer because it uses the feature values to split the data into smaller and smaller subsets of similar classes.

Beginning at the root node, which represents the entire dataset, the algorithm chooses a feature that is the most predictive of the target class. The examples are then partitioned into groups of distinct values of this feature; this decision forms the first set of tree branches. The algorithm continues to divide-and-conquer the nodes, choosing the best candidate feature each time until a stopping criterion is reached.

This might occur at a node if:

- All (or nearly all) of the examples at the node have the same class
- There are no remaining features to distinguish among examples
- The tree has grown to a predefined size limit

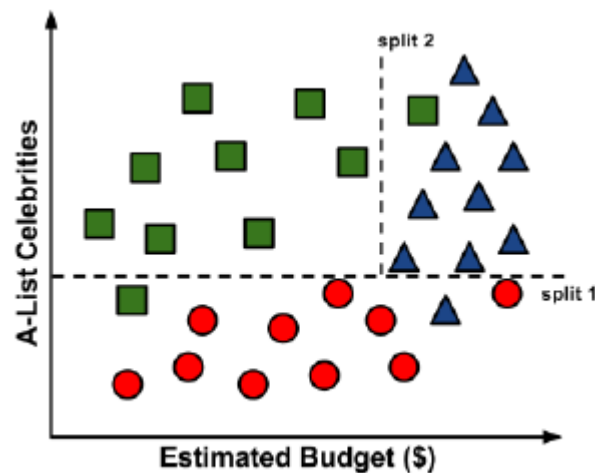


Figure 19: Classification based on decision trees [7]

The regression approach to decision tree learning varies from classification as always, having a continuous target variable instead of a labeled one. In this case, at each node, a multiple linear regression model is built from the examples reaching



that node. The split is then made considering the node impurity being the sum of squared deviations about the mean and the node predicting the sample mean of the continuous target variable.

Another approach to regression trees is to simply predict the average of the observations falling into the node but, even it is a much more common practice, the results are relatively worse compared to the multiple linear regression approach. The upside of the second approach being the decrease of training times as the multiple linear regression approach has to train a model for every node split which may be very time and memory consuming for large trees.



UNIVERSIDAD PONTIFICIA COMILLAS
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)
INGENIERO INDUSTRIAL

ANNEX 3 GRADIENT BOOSTING

REGRESSION

In order to improve model performance, a widespread approach in machine learning is to implement ensemble methods on the algorithms already tried to solve the problem at hand.

In the scope of this project, an ensemble method called boosting was applied to the decision tree regression algorithm already implemented.

All ensemble methods are based on the idea that by combining multiple weaker learners, a stronger learner is created. Using this simple principle, a large variety of algorithms has been developed distinguished largely by two questions:

- How are the weak learning models chosen and/or constructed?
- How are the weak learners' predictions combined to make a single final prediction?

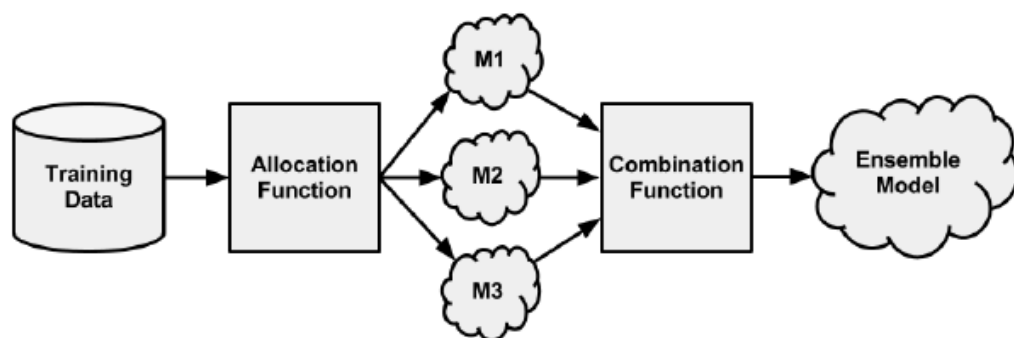


Figure 20: Ensemble methods flowchart [7]

The allocation function in Figure 20 dictates whether each model receives the full training dataset or merely a sample.

Since the ideal ensemble includes a diverse set of models, the allocation function could increase diversity by artificially varying the input data to train a variety of learners.



After the models are constructed, they can be used to generate a set of predictions, which must be managed in some way. The combination function governs how disagreements among the predictions are reconciled. For example, the ensemble might use a majority vote to determine the final prediction, or it could use a more complex strategy such as weighting each model's votes based on its prior performance.

A very widespread practice in ensemble learning is called boosting, because it boosts the performance of weak learners to attain the performance of stronger learners. Boosting is shown to perform quite better and certainly no worse than the best of the weaker models. Given the obvious utility of this finding, boosting is thought to be one of the most significant discoveries in machine learning.

Boosting uses ensembles of models trained on resampled data and a vote to determine the final prediction. The key difference is that the resampled datasets in boosting are constructed specifically to generate complementary learners, and the vote is weighted based on each model's performance rather than giving each an equal vote.

Beginning from an unweighted dataset, the first classifier attempts to model the outcome. Examples that the classifier predicted correctly will be less likely to appear in the training dataset for the following classifier, and conversely, the difficult-to-classify examples will appear more frequently. As additional rounds of weak learners are added, they are trained on data with successively more difficult examples. The process continues until the desired overall error rate is reached or performance no longer improves. At that point, each classifier's vote is weighted according to its accuracy on the training data on which it was built.

Though boosting principles can be applied to nearly any type of model, the principles are most commonly used with decision trees, as in the scope of this project.

ANNEX 4 SUPPORT VECTOR MACHINES

A Support Vector Machine (SVM) can be imagined as a surface that defines a boundary between various points of data which represent examples plotted in multidimensional space according to their feature values. The goal of an SVM is to create a flat boundary, called a hyperplane, which leads to fairly homogeneous partitions of data on either side. This linear boundary (hyperplane) is used to separate the linearly separable data into homogeneous groups (Figure 21).

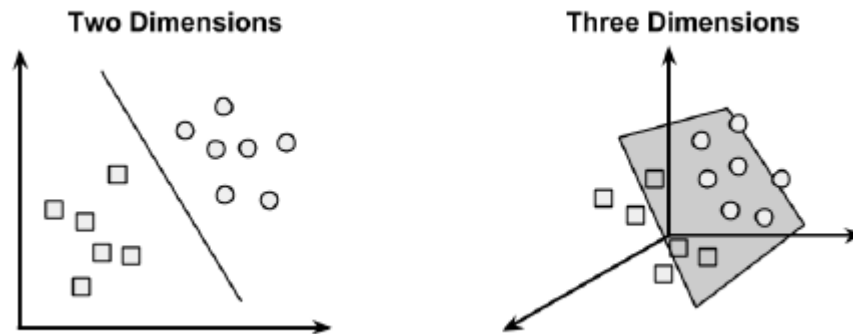


Figure 21: Linear separation in 2D and 3D [7]

There is an infinity of possibilities of lines separating the two classes, the SVM algorithm calculates the Maximum Margin Hyperplane (MMH) that creates the greatest separation between the two classes as it is likely that the line that leads to the greatest separation will generalize the best to future data. This is because slight variations in the positions of the points near the boundary might cause one of them to fall over the line by chance.

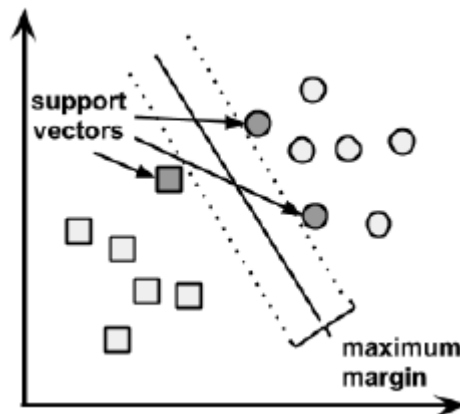


Figure 22: Support vectors [7]

The support vectors are the closest points to the MMH, each class must have at least one but can have more. One key feature of SVMs is that the boundary is calculated from the support vectors and then, it represents a very compact way to store the models even if the amount of training data is very large.

When the data is linearly separable, the SVM algorithm is very simple, calculates the MMH, which is a mathematical operation available since a very long time ago, and stores the model in a compact way.

One of the key features of Support Vector Machines is not only its ability to find the MMH but to be applicable to non-linearly separable data.

In many real-world applications, the relationships between variables are non-linear. A key feature of SVMs is their ability to map the problem into a higher dimension space using a process known as the kernel trick. In doing so, a non-linear relationship may suddenly appear to be linear.

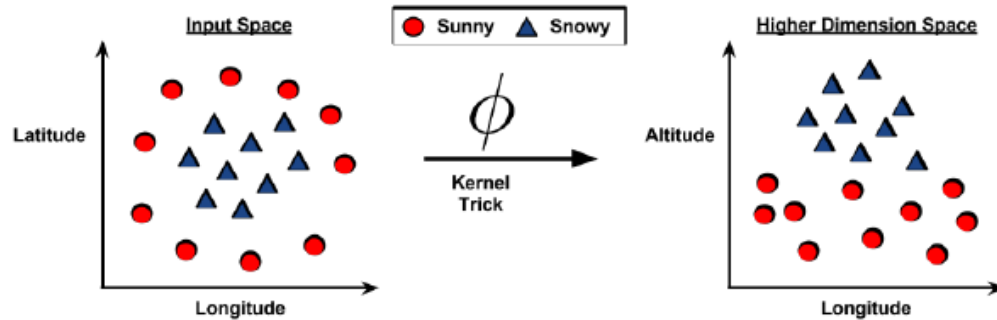


Figure 23: Kernel trick [7]

SVMs with non-linear kernels add additional dimensions to the data in order to create separation in this way. Essentially, the kernel trick involves a process of adding new features that express mathematical relationships between measured characteristics.

There is no reliable rule for matching a kernel to a particular learning task. The fit depends heavily on the concept to be learned as well as the amount of training data and the relationships among the features. Often, a bit of trial and error is required by training and evaluating several SVMs on a validation dataset. That said, in many cases, the choice of kernel is arbitrary, as the performance may vary only slightly.



UNIVERSIDAD PONTIFICIA COMILLAS
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)
INGENIERO INDUSTRIAL



ANNEX 5 AMAZON WEB SERVICES

The world is currently experimenting an undeniable technological revolution, affecting all technological markets one of the most important new tools brought to the public is undoubtedly **the cloud**. From file storage services (e.g. Dropbox, S3...) to PaaS and SaaS platform services, the cloud offers nowadays infinite computation power and storage. One of the strongest players in the cloud computing services market is Amazon Web Services, subsidiary of Amazon.com.

Since 2006, AWS offers an ever evolving and state of the art cloud computing platform to its clients. One of the key benefits of cloud computing is the opportunity to replace up front capital infrastructure expenses with low variable costs that scale with your business. With the cloud, businesses no longer need to plan for and procure servers and other IT infrastructure weeks or months in advance. Instead, they can instantly spin up hundreds or thousands of servers in minutes and deliver results faster.

Today, Amazon Web Services provides a highly reliable, scalable, low cost infrastructure platform in the cloud that powers hundreds of thousands of businesses in 190 countries around the world.

AWS provides an abstraction layer from the upfront investments in hardware and platform maintenance, it enables you to provision exactly the right type and size of computing resources you need to power your project.

Cloud computing provides a simple way to access servers, storage, databases, and a broad set of application services over the Internet. Cloud computing providers such as AWS own and maintain the network connected hardware required for these application services, while you provision and use what you need using a web application.

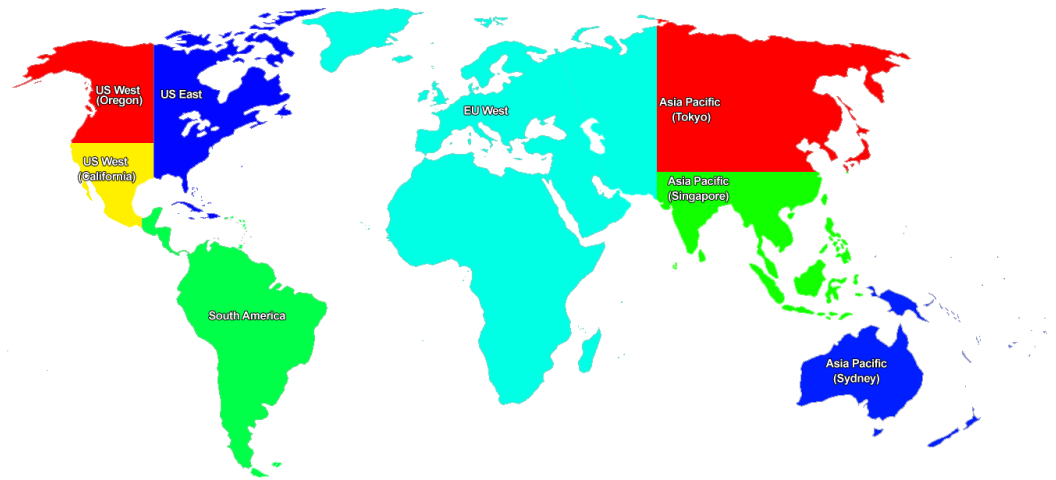


Figure 24: AWS availability zones

AWS is available in multiple locations worldwide. These locations are composed of regions and Availability Zones. A region is a named set of AWS resources in the same separate geographic area. Each region has multiple, isolated locations known as Availability Zones. AWS enables the placement of resources, such as instances, and data in multiple locations. Resources aren't replicated across regions unless you chose to do so.

Each region is completely independent and is designed to be completely isolated from the other regions. This achieves the greatest possible fault tolerance and stability. Each Availability Zone is isolated, but the Availability Zones in a region are connected through low-latency links. Availability Zones are physically separated within a typical metropolitan region and are located in lower risk flood plains (specific flood zone categorization varies by region). In addition to utilizing discrete uninterruptable power supply (UPS) and onsite backup generators, they are each fed via different grids from independent utilities to further reduce single points of failure. Availability Zones are all redundantly connected to multiple tier-1 transit providers.

One of the main concerns with the cloud is security. ¿How secure is the platform? ¿Is it secure enough for me to put my information inside? Are example of questions asked when a cloud-based platform is discussed. AWS takes care of security as all AWS customers benefit from data center architecture and network architecture built



to satisfy the requirements of the most security-sensitive organizations. AWS and its partners offer hundreds of tools and features to help meet security objectives for visibility, auditability, controllability, and agility. This means that you can have the security you need, but without the capital outlay, and with much lower operational overhead than in an on-premises environment.

AWS provides you with guidance and expertise through online resources, personnel, and partners.

To access the services available, Amazon Web Services offers an internet portal, the AWS management portal or the Amazon Command Line Interface.



Figure 25: Amazon Web Services

Amazon offers a wide set of services, this annex will only cover the ones used in the scope of the project.



AMAZON EC₂

Amazon Elastic Compute Cloud (Amazon EC₂) is a web service that provides resizable compute capacity in the cloud. It is designed to make web-scale computing easier for developers.



Amazon EC2

Figure 26: Amazon EC2

Amazon EC₂ enables the user to obtain and configure virtual servers in the cloud in less than 10 seconds. It offers complete control over the computing resources of the instances and edit them at any time in minutes. Amazon EC2 changes the economics of computing by allowing you to pay only for capacity that you actually use. Amazon EC2 provides developers and system administrators the tools to build failure resilient applications and isolate themselves from common failure scenarios. In the scope of the project, the machines used both to gather data from APIs (Santander), run the machine learning algorithms and store the visualization software are EC2 instances residing inside Ferrovial's own Virtual Private Cloud.

AMAZON S3

Amazon Simple Storage Service (Amazon S3) provides developers and IT teams with safe, secure, and highly-scalable object storage. Amazon S3 is easy to use.



It has a simple web service interface for storage and retrieval of any amount of data from anywhere on the web.



Figure 27: Amazon S3

Amazon S3 offers a directory system just as Dropbox does, with high replication and unlimited capacity and low latency data transfer between services.

In the scope of the project all the data used resides in S3 as a backup and staging area. When the data is analyzed in batch, and doesn't have to be stored in the database, the files containing it are stored in the S3 file system while waiting to be processed.

AMAZON EMR

Amazon Elastic MapReduce (Amazon EMR) is a web service that makes it easy to quickly and cost-effectively process vast amounts of data.



Figure 28: Amazon EMR



Amazon EMR simplifies big data processing, providing a managed Apache Hadoop framework that makes it easy, fast, and cost-effective for you to distribute and process vast amounts of your data across dynamically scalable Amazon EC2 instances. You can also run other popular distributed frameworks such as Apache Spark and Presto in Amazon EMR, and interact with data in other AWS data stores such as Amazon S3 and Amazon DynamoDB.

Amazon EMR provides a similar interface as EC2 but instead of booting a single computing instance, you boot an entire cluster of machines. The configuration is, once again, entirely tunable in terms of computation power of each instance, number of instances... It also enables the user to choose a cluster manager and install some of the most popular frameworks while booting (Apache Hive, Apache Pig...).

In the scope of the project, the ETL for Santander and Barcelona's measurements is performed using Apache Hive on top of a Hadoop cluster running in an EMR cluster instance.



ANNEX 6 HADOOP AND MAPREDUCE

As stated in Annex 5, one of the most important aspects of the technological revolution the world is experiencing is the increase of available computing power. One of the uses of the cloud is to perform analyses which were impossible to do before by analyzing datasets impossible to analyze before. Hadoop is one of the many frameworks available nowadays helping the companies to analyze enormous datasets in the cloud.

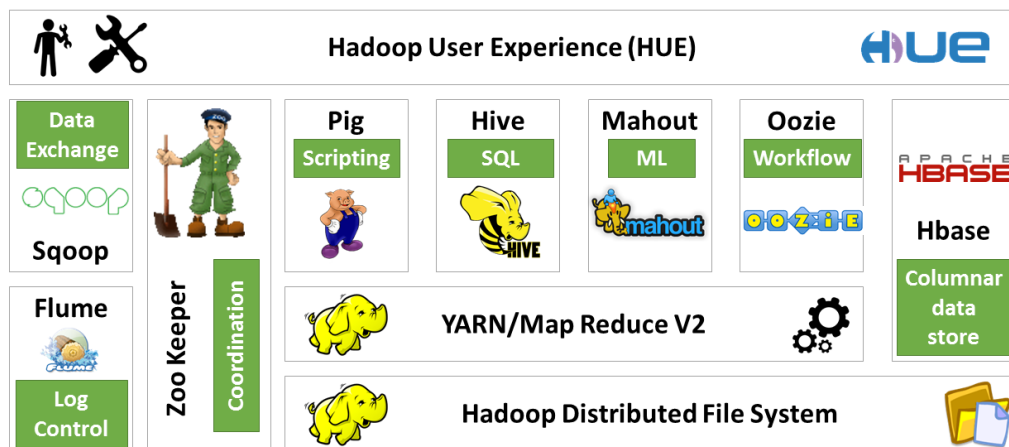


Figure 29: Hadoop ecosystem [6]

Hadoop, can't and should not be seen as a single tool, it is defined officially as an open-source software framework for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware.

The last part of the definition is key, Hadoop is designed to run on cheap hardware and thus, should be able to detect and overcome node failures in every component.

The annex will present the three most important modules inside the Apache Hadoop framework and comment briefly some of the others, used in the current project.



HADOOP DISTRIBUTED FILE SYSTEM

The core module in the framework, allows the storage of enormous files, abstracting the user from the partitioning and distribution.

It's a scalable file system that distributes and stores data across all machines in a Hadoop cluster (a group of servers). Each HDFS cluster contains the following:

- NameNode: Runs on a “master node” that tracks and directs the storage of the cluster.
- DataNode: Runs on “slave nodes,” which make up the majority of the machines within a cluster. The NameNode instructs data files to be split into blocks, each of which are replicated three times and stored on machines across the cluster. These replicas ensure the entire system won't go down if one server fails or is taken offline—known as “fault tolerance.”
- Client machine: neither a NameNode nor a DataNode, Client machines have Hadoop installed on them. They're responsible for loading data into the cluster, submitting MapReduce jobs and viewing the results of the job once complete.

As every file is replicated 3 times in the system, the breakdown of one of the servers is easily overcome as the NameNode knows exactly where the lost information is replicated.

MAPREDUCE

MapReduce is the programming paradigm used to treat the vast amounts of data stored in HDFS. A MapReduce program is composed of a Map procedure that performs filtering and sorting (such as sorting students by first name into queues, one queue for each name) and a Reduce method that performs a summary operation (such as counting the number of students in each queue, yielding name frequencies). The "MapReduce System" (also called "infrastructure" or "framework") orchestrates the processing by marshalling the distributed servers, running the

various tasks in parallel, managing all communications and data transfers between the various parts of the system, and providing for redundancy and fault tolerance.

The main advantage is that the map step is run in parallel, across all machines in the cluster on the data available locally. This is one of the key aspects of MapReduce, the computation is run locally, and there is no data transfer through the network which would be time consuming as the amount of data is huge. The reduce operation takes the partial results of the different mapping stages and sums up the operations.

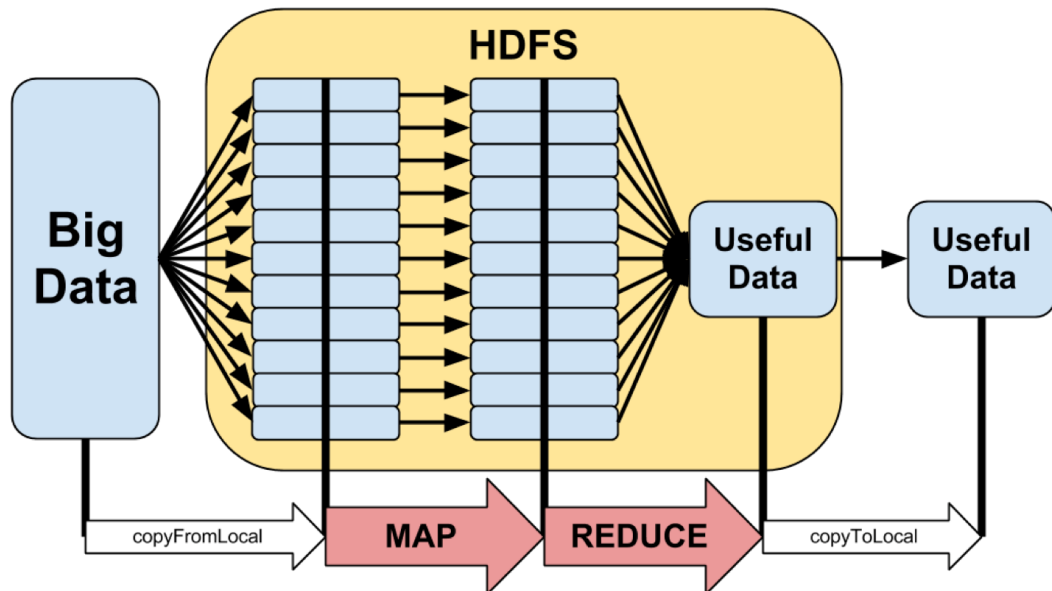


Figure 30: MapReduce paradigm [6]

The delegation of tasks across the cluster is performed by “daemons,” the JobTracker and TaskTracker:

- JobTracker: The JobTracker oversees how MapReduce jobs are split up into tasks and divided among nodes within the cluster.
- TaskTracker: The TaskTracker accepts tasks from the JobTracker, performs the work and alerts the JobTracker once it’s done. TaskTrackers and DataNodes are located on the same nodes to improve performance.

In Hadoop, MapReduce jobs are developed in Java but also can be called via the Hadoop API.



YET ANOTHER RESOURCE NEGOTIATOR

Apache Hadoop YARN (Yet Another Resource Negotiator) is a cluster management technology sitting on top of HDFS and manages the delegation of resources for MapReduce jobs.

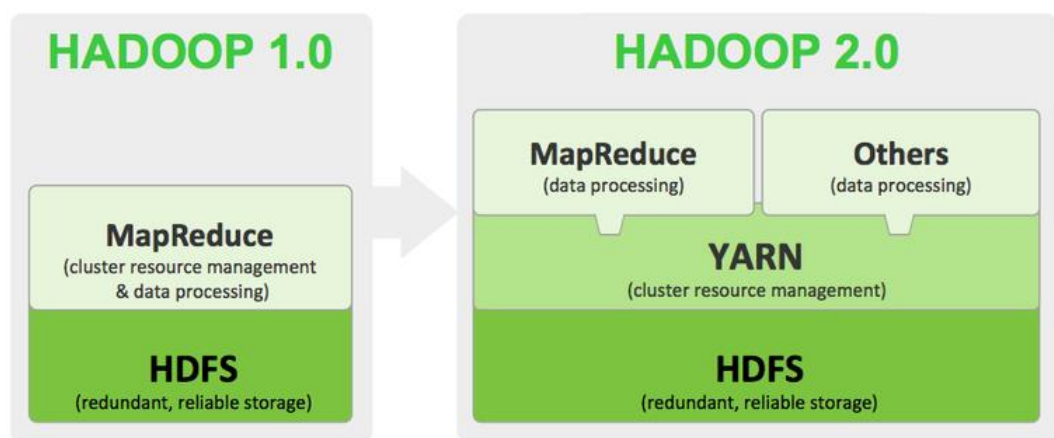


Figure 31: Hadoop 2.0 architecture [6]

YARN is often called the operating system of Hadoop because it is responsible for managing and monitoring workloads, maintaining a multi-tenant environment, implementing security controls, and managing high availability features of Hadoop. Like an operating system on a server, YARN is designed to allow multiple, diverse user applications to run on a multi-tenant platform.

The applications available to sit above YARN are applications which form an abstraction layer for the user. As writing MapReduce jobs in java is a very tedious task, YARN translates the programs the programs running in the applications above it to MapReduce jobs to be run inside the cluster.



Figure 32: Apache Hive [6]

One example of such application is Hive, the most popular SQL-in-Hadoop option and the Hadoop community has invested heavily in making Hive faster, more scalable, and supportive of more SQL operations. Hive was used in the scope of this project when designing and developing the ETL process for the cities of Santander and Barcelona.



UNIVERSIDAD PONTIFICIA COMILLAS
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)
INGENIERO INDUSTRIAL



BIBLIOGRAPHY

- [1] https://web-assets.domo.com/blog/wp-content/uploads/2014/04/DataNeverSleeps_2.0_v2.jpg
- [2] <http://www.slideshare.net/mjft01/big-data-landscape-version-20?related=1>
- [3] <http://digitaltechdiary.com/gartners-2015-hype-cycle-for-digital-marketing/2241/>
- [4] http://www.slideshare.net/IIG_HES/mining-in-the-middle-of-the-city-the-needs-of-big-data-for-smart-cities
- [5] <http://futurecity.glasgow.gov.uk/dashboards/>
- [6] <http://www.tomsitpro.com/articles/hadoop-2-vs-1,2-718.html>
- [7] <https://aws.amazon.com/es/whitepapers/overview-of-amazon-web-services/>
- [8] www.chinastor.org/upload/2014-05/14050416114822.pdf