ARQUITECTURA DE SISTEMES BIG DATA PEL RECORD LINKAGE DE XARXES SOCIALS

TREBALL FINAL DE GRAU – Informe inicial

2018-2019

ALEJANDRO GARCIA CARBALLO
1423957

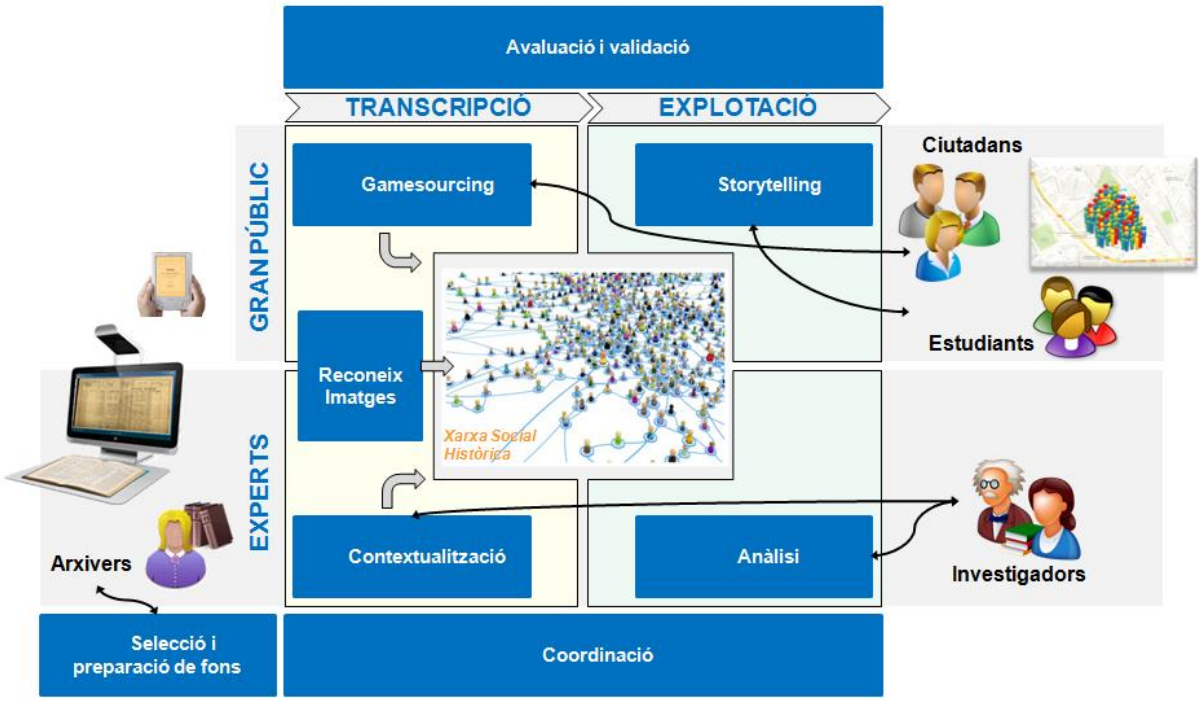
Índex:

1. Context del TFG:

Aquest treball ha sigut proposat per l'Oriol Ramos i està emmarcat dins del projecte de recerca XARXES, un projecte que s'està duent a terme en col·laboració amb el Centre de Visió per Computació (CVC) i el Centre d'Estudis Demogràfics (CED) de la UAB. Aquest projecte està finançat per Recercaixa de l'obra social La Caixa amb la col·laboració de l'ACUP (Associació Catalana d'Universitats Públiques).

Els documents demogràfics permeten estudiar un gran conjunt de comportaments en diferents àmbits com ara l'evolució social i econòmica del passat. Gràcies a les tècniques informàtiques de visió per computador es pretenen construir xarxes socials històriques aprofitant les dades demogràfiques de les que es disposen. Actualment es tenen dades de les poblacions catalanes de Sant Feliu de Llobregat, Begues, Castellví de Rosanes, Collbató, Corbera de Llobregat, El Papiol, Molins de Rei, Sant Vicenç de Castellet, Santa Coloma de Cervelló i Torrelles de Llobregat, però en un futur el nombre de poblacions pot augmentar i el sistema ha de suportar el possible creixement de dades. Les dades de les que es disposen actualment són alguns fitxers excels amb unes 25.000 entrades aproximadament i una part de la informació en MySQL amb dades dels habitants registrats en els padrons i en d'altres fonts de població de localitats veïnes.

Aquest projecte està pensat per a que sigui útil per a diferents usuaris, tant per a aquells usuaris que es dediquen a l'estudi de dades demogràfiques, històriques o qualsevol àmbit que pugui ser d'interès, com per a usuaris no experts, als quals se'ls pot oferir un servei didàctic més atractiu per tal d'apropar el patrimoni històric demogràfic de manera digital i facilitar també així el consum d'aquest coneixement.



Il·lustració 1 - Esquema general del funcionament global del projecte extret de la pàgina informativa del projecte XARXES [1]

En la il·lustració 1 podem veure un esquema general del funcionament global del projecte extret de la pàgina d'informació del projecte XARXES [1]. Podem observar que en aquest esquema es veuen els diferents tipus d'usuari que poden arribar a utilitzar aquesta aplicació i que és el que pot ser de més interès a cadascun d'aquests usuaris.

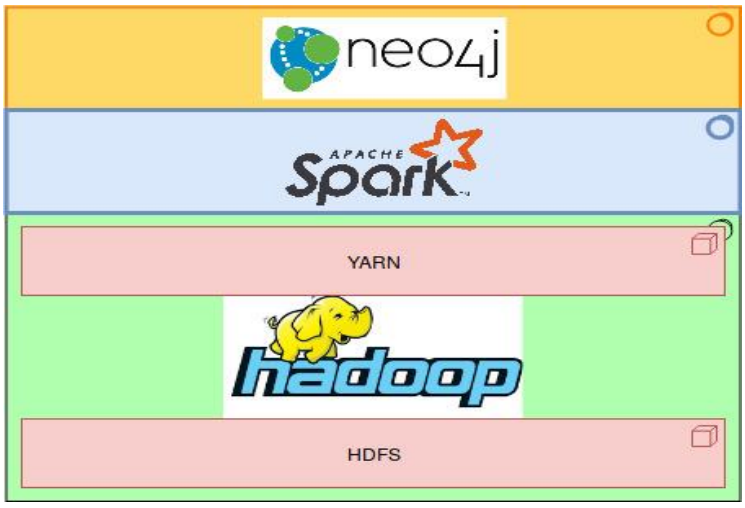
La feina en la que hem centraré dins d'aquest gran projecte és en la creació d'una arquitectura Big Data a través d'una arquitectura distribuïda. Col·laboraré amb Joana Maria Pujadas (del Centre d'Estudis Demogràfics) i amb un estudiant de màster que s'encarrega del processament de les dades el qual utilitzarà el nostre sistema Big Data.

2. En què consisteix aquest TFG?

Donat el context anterior, la meva feina està centrada en aconseguir una arquitectura que permeti una escalabilitat horitzontal per tal de poder anar ampliant amb tants nodes com sigui necessari tant l'espai com la potència de còmput. A diferència d'una escalabilitat vertical on el creixement consisteix en la millora del hardware i per tant, pot arribar al seu límit amb facilitat, amb un sistema amb escalabilitat horitzontal resolem aquest problema. Gràcies a aquest tipus d'arquitectura es podran anar afegint les dades de més poblacions sense haver-nos de preocupar per arribar a un màxim de capacitat.

Per dur a terme aquesta arquitectura utilitzaré els clústers proporcionats pel Centre de Visió per Computació els quals tenen instal·lats un sistema Linux amb la distribució CentOS als quals puc accedir a través de SSH. Treballaré en un node amb docker instal·lat, per tant, el projecte estarà basat en contenidors de docker i utilitzaré dockerfiles per a crear l'entorn de treball.

Faré ús d'algunes de les tecnologies més utilitzades en aquest àmbit. En la fase inicial del projecte he dedicat una part del temps a la investigació i a aprendre sobre aquest tipus de tecnologies i comparar diferents opcions possibles.



Il·lustració 2 - Tecnologies que utilitzaré durant el projecte

El diagrama que mostro a la il·lustració 2 representa una visió general de quines són les tecnologies que participaran en la realització de l'arquitectura. Com a base tenim un sistema Hadoop, com a capa de programació distribuïda utilitzaré Spark i finalment com a sistema de base de dades estructurada en graf utilitzaré Neo4j.

Hadoop és un framework d'Apache que permet el processament distribuït de grans conjunts de dades a través de clústers. Està dissenyat per a poder ser escalat des de servidors individuals fins a milers de màquines, cadascuna de les quals ofereix computació i almacenatge locals. La llibreria està dissenyada per detectar i manejar fallades a la capa d'aplicació, de manera que es proporciona un servei amb gran disponibilitat en els clústers, cadascun dels quals pot ser propens a fallades. No entraré en detall de l'arquitectura de Hadoop en aquest primer informe però explicaré breument els components pels quals està format:

-HDFS (Hadoop Distributed File System): Emmagatzema grans quantitats de dades a través de múltiples màquines, replicant les dades a través de diferents hosts per aconseguir tolerància a falles. Les dades es divideixen en blocs de mida fixa que acostumen a ser de 64 MiB, però el tamany d'aquests blocs es pot configurar. Com el mateix nom indica, es tracta d'un sistema de fitxers al igual que el sistema de fitxers d'un sistema operatiu, però a diferència d'aquests, aquest sistema de fitxers està pensat per a poder ser distribuït.

-YARN (Yet another Resource Negotiator): És una tecnologia de gestió de clústers que va sorgir en la segona versió de Hadoop. Amb YARN s'aconsegueix desacoblar la gestió de recursos als clústers del sistema de processament de dades com pot ser l'original de Hadoop MapReduce.

-MapReduce: Model de programació per processar grans datasets en paral·lel i de manera distribuïda. Està basat en dos passos principals que són Map i Reduce.

Gràcies al desacoblament que s'aconsegueix amb YARN, utilitzaré un altre sistema de processament de dades diferent de MapReduce anomenat **Spark**, ja que pot aconseguir executar un programa 100 vegades més ràpid en memòria o 10 vegades més ràpid en disc que utilitzant MapReduce. Algunes de les diferències principals entre aquestes dos tecnologies són:

Hadoop Mapreduce	Apache Spark
Ràpid	100x més ràpid que MapReduce
Processament en Batch	Processament en temps real
Emmagatzema les dades en disc	Emmagatzema les dades en memòria
Implementat en Java	Implementat en Scala

Com a base de dades estructurada en grafs utilitzaré **Neo4j**, ja que permet una integració simple entre la base de dades i Spark.

3. Planificació

En aquest apartat explicaré tant els objectius del projecte com les tasques que duré a terme al llarg del treball.

3.1. Objectius

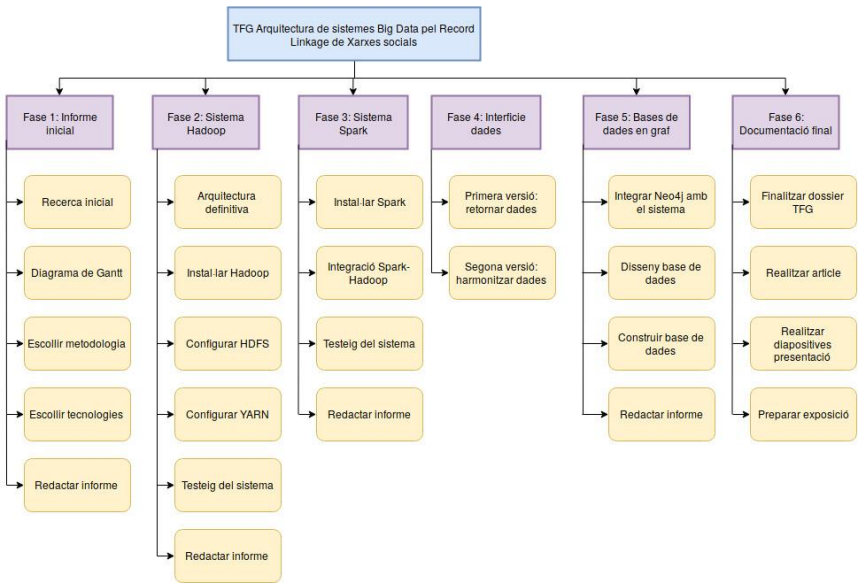
Donada l'explicació anterior on aclareixo en què consisteix el projecte i tenint un context per entendre on entra aquest treball dins del projecte global de recerca XARXES, podem establir que els objectius d'aquest TFG son els següents:

- Millorar la qualitat del projecte global de recerca XARXES a través d'un sistema amb arquitectura Big Data.
- Augmentar l'eficiència del projecte XARXES.
- Facilitar l'escalabilitat de les dades.

Per tant, el projecte quedarà finalitzat quan quedi implementada l'arquitectura Big Data amb cadascun dels components connectats, funcionant entre ells i garantint una qualitat en el sistema final.

3.2. Tasques

Per aconseguir dur a terme els objectius i finalitzar el projecte correctament hauré de realitzar un conjunt de tasques. Cada tasca pertany a una fase del projecte, el qual he dividit en 5 fases generals. Aquestes fases les visualitzaré a través d'un diagrama WBS (Work Breakdown Structure):



Il·lustració 3 - WBS per mostrar les fases del projecte

Fase 1: En aquesta fase tractaré de fer una planificació general del projecte, fer una recerca inicial per informar-me de tot el necessari per a poder fer tant la planificació com poder escollir correctament les tecnologies que utilitzaré, i escollir quina serà la metodologia que utilitzaré durant el projecte per dur-lo a terme. Finalment, com en totes les fases, documentaré tota la feina i tractaré de plasmar-la en la redacció de l'informe.

Fase 2: El sistema Hadoop serà la base que utilitzaran la resta de components, per tant, aquest serà el primer component a implementar. Hauré de fer una instal·lació en les màquines del CVC, configurar el HDFS i el YARN i testejar que la implantació s'ha fet correctament amb diferents tipus de tests. Tot el sistema que implementaré estarà en contenidors de docker.

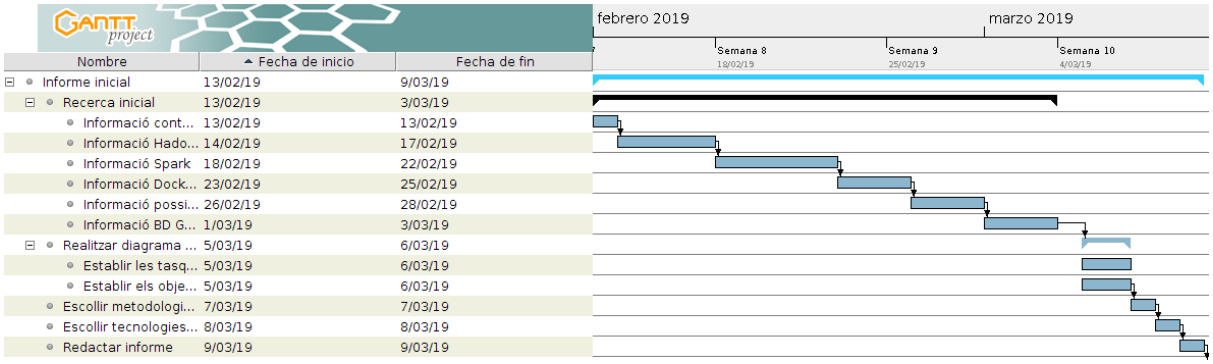
Fase 3: Una vegada tingui implementat el sistema Hadoop passaré a fer la implementació del sistema Spark, hauré de instal·lar-ho i integrar-ho amb el Hadoop, testejant que la integració és correcte.

Fase 4: En aquesta fase tractaré de fer un mòdul (de moment tinc pensat fer-ho en Python), per retornar les dades que estan distribuïdes per el HDFS. La primera versió d'aquest mòdul es centrarà només en retornar les dades que demana l'usuari, però si hi ha temps suficient, tractaré de fer una harmonització de les dades. Harmonitzar les dades consisteix en donar una versió actual i sense faltes ortogràfiques de les dades. Per exemple, alguns noms al llarg del temps poden evolucionar i canviar la manera en que s'escriuen o bé, han sigut anotats en les dades amb faltes ortogràfiques. A través d'un procés d'harmonització, es pot aconseguir el nom correcte. S'han de reservar els dos noms, tant l'harmonitzat com el que no ho està, ja que el sistema de visió per computador ha de detectar el que està escrit sense discriminar si es correcte ortogràficament.

Fase 5: Aquesta fase tracta tot el relacionat amb les bases de dades, on hauré d'integrar el sistema de bases de dades Neo4j, un sistema basat en graf, amb el sistema que vagi construint. Hauré de dissenyar la base de dades i construir-la, tenint en compte que les dades estan en formats diversos com pot ser fitxers excel o bases de dades MySQL.

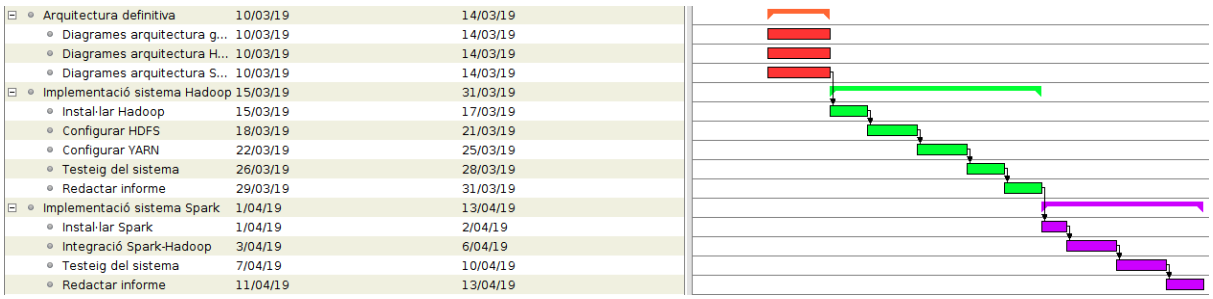
Fase 6: Finalment, hauré d'acabar el dossier del TFG entregable, fer l'article i les diapositives de la presentació i preparar-me tot el necessari per a fer l'exposició.

Aquestes tasques les he representat en el temps a través del diagrama de Gantt següent:



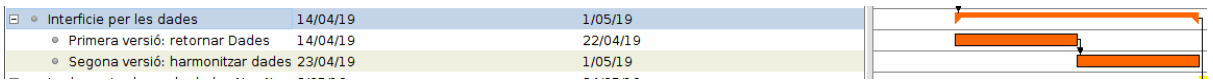
Il·lustració 4 - Diagrama de Gantt que inclou tasques de l'informe inicial

Aquesta part conté tota la feina duta a terme per a l'informe inicial, l'entrega que es fa el 10 de març. Conté la fase 1 del projecte, que està conformada per les tasques d'investigació i informació de tot el relacionat amb el TFG, fer una llista de tasques i objectius, fer el diagrama de Gantt per planificar en el temps les tasques, escollir la metodologia que utilitzaré al llarg del treball, les tecnologies que utilitzaré per començar el desenvolupament del treball i finalment la redacció de l'informe.



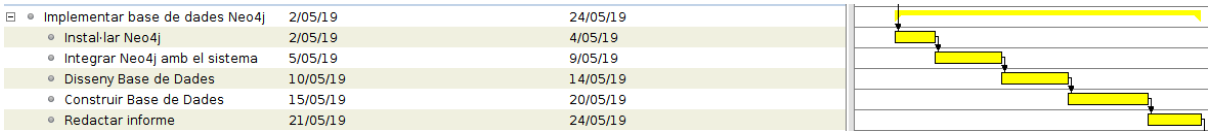
Il·lustració 5 - Diagrama de Gantt amb les tasques a realitzar per a la primera entrega de l'informe de progrés

La il·lustració 6 conté la feina que hauria d'estar finalitzada per a la segona entrega del TFG, amb el lliurament del primer informe de progrés. De manera resumida, en aquesta part del projecte hauria d'estar concretada l'arquitectura que es vol dur a terme i s'hauria de tenir instal·lat i configurat Hadoop i Spark per a poder realitzar la feina de base de dades en la següent entrega. Aquesta feina hauria d'estar finalitzada per al 14 d'abril, amb l'entrega del primer informe de progrés.



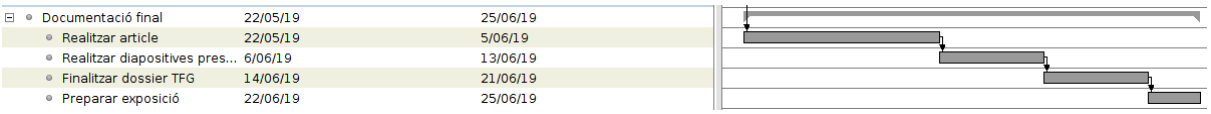
Il·lustració 6 - Diagrama de Gantt amb les tasques a realitzar per a la fase de interfície de dades

En la il·lustració 7 podem veure les dos versions que tractaré de fer de la interfície de dades. Poden haver-hi variacions en aquesta part segons com avanci el treball, però la versió inicial te contemplada poder realitzar aquestes dues versions.



Il·lustració 7 - Diagrama de Gantt de les tasques de bases de dades estructurada en graf

En aquesta part del projecte s'ha de finalitzar tota la instal·lació, el disseny i la construcció de la base de dades estructurada en graf. També s'hauria de realitzar una interfície per a les dades, que podrà ser aprofitada per l'alumne de master encarregat del processament de les dades.



Il·lustració 8 - Diagrama de Gantt amb les tasques per finalitzar el projecte

Aquesta fase final conté tot el relacionat amb la finalització de la documentació (tant l'article com el dossier final) i la realització de la presentació final del TFG.

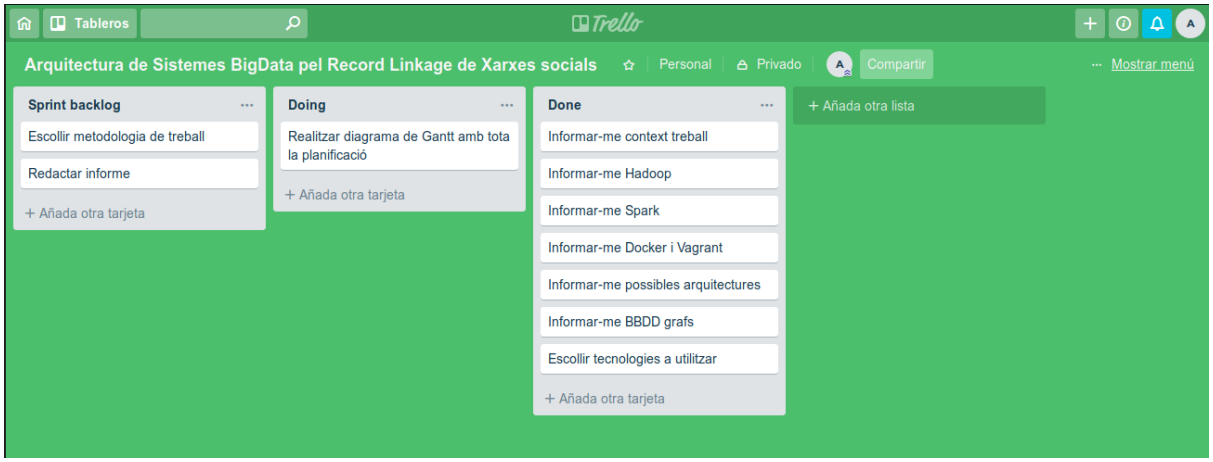
El temps dedicat a les tasques pot variar conforme vagi avançant en el TFG degut a que noti falta de temps o que alguna de les parts es compliqui més del que tenia pensat, per tant, si faig algun canvi, tot estarà reflectit en els futurs informes per poder fer un seguiment de quins canvis he hagut de fer conforme la planificació inicial.

El diagrama de Gantt s'ha realitzat amb una eina open source anomenada GanttProject [13].

4. Metodologia

Com a metodologia de treball faré ús d'una metodologia àgil per tal de poder adaptar qualsevol canvi necessari de manera ràpida i fer entregues cada cert temps per poder anar corregint i millorant el sistema.

Al ser una única persona, gran part de les metodologies conegudes a utilitzar haurien de ser adaptades per a poder fer-ne ús. En aquest treball utilitzaré una **metodologia Scrum adaptada** per tal de poder fraccionar el producte en entregues parcials. Això és necessari no només per la pròpia metodologia, sinó també perquè hem de fer entregues parcials en les diferents fites, i d'aquesta manera podré gestionar que quan faci una entrega parcial, aquesta contingui uns mòduls complets i entregables i que no quedi la feina a mitges.



Il·lustració 9 - Tauler amb les columnes que utilitzaré durant el projecte per a realitzar els sprints

Dividiré la feina en sprints de dues setmanes i crearé una llista de tasques en un tauler de l'eina Trello[14], que com es pot observar en la il·lustració tinc tres columnes: un sprint backlog, una columna de feina que està en procés i una altra per a la feina finalitzada. En les diferents fites adjuntaré informació de com ha anat el sprint, si no he arribat a temps per fer tota la feina i qualsevol modificació que sigui necessària.

Al fer les entregues hi hauran reunions amb el tutor per rebre feedback de com està anant la feina i de si és necessari fer canvis. També hi hauran algunes reunions amb un estudiant de master col·laborador que s'encarrega de la part de processament de les dades, amb el qual tindrem que concretar una part de la feina.

5. Legalitat

Les dades que s'utilitzaran en aquest projecte són dades de ciutadans reals, per tant, per a la realització d'aquest TFG hauré de signar un acord de confidencialitat. S'entén per informació confidencial qualsevol tipus de dades transmeses per part del CVC i que utilitzaré per a la realització del treball.

Les obligacions que he de seguir indicades en aquest acord impliquen que s'han d'utilitzar les dades de manera reservada, no puc divulgar ni comunicar la informació i també he d'evitar la copia o revelació d'aquesta informació a tercers.

Degut als grans danys i perjudicis que pot ocasionar la divulgació o ús no autoritzat d'aquestes dades, la part emissora (CVC) tindrà dret a reclamar davant els tribunals competents i a obtenir una indemnització pels danys generats.

6. Bibliografia

[1] Centre de Visió per Computació, “XARXES: Tecnologia i innovació ciutadana en la construcció de xarxes socials històriques per a la comprensió del llegat demogràfic”. [En línia]. Disponible a: http://dag.cvc.uab.es/xarxes/info Consultat: 13 de febrer de 2019

[2] B.Bahari, “Hadoop Fundamental”, Febrer 6, 2017. [En línia]. Disponible a: https://medium.com/sarccom/hadoop-fundamental-5179099f5b21 Consultat: 13 de febrer de 2019

[3] M.Rouse, “Database replication”, Juny, 2018. [En línia]. Disponible a: https://searchdatamanagement.techtarget.com/definition/database-replication Consultat: 14 de febrer de 2019

[4] B.Proffit, “Hadoop: What it is and how It works”, Maig 23, 2013. [En línia]. Disponible a: https://readwrite.com/2013/05/23/hadoop-what-it-is-and-how-it-works/ Consultat: 14 de febrer de 2019

[5] K.Shvachko, H.Kuang, S.Radia i R.Chansler, “The Hadoop Distributed File System”, [En línia]. Disponible a: http://storageconference.us/2010/Papers/MSST/Shvachko.pdf Consultat: 14 de febrer de 2019

[6] Apache, “HDFS Architecture”, 13 de novembre, 2018. [En línia]. Disponible a: https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html Consultat: 15 de febrer de 2019

[7] Diwakar, “Instaling Hadoop on Ubuntu”, 28 de juny, 2018. [En línia]. Disponible a: https://medium.com/devilsadvocatediwakar/installing-hadoop-on-ubuntu-9493af1af12f Consultat: 15 de febrer de 2019

[8] Apache, “Hadoop: Setting up a Single Node Cluster”, 13 de novembre, 2018. [En línia]. Disponible a: https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html Consultat: 15 de febrero de 2019

[9] N.Mahesh, “Setting up Hadoop – Mapreduce, HDFS, and YARN. Standalone and pseudo-distributed mode”, 7 de juliol, 2017. [En línia]. Disponible a: https://medium.com/@nidhinmahesh/getting-started-hadoop-mapreduce-hdfs-and-yarn-configuration-and-sample-program-febb1415f945 Consultat: 15 de febrer de 2019

[10] A.Bekker, “Spark vs Hadoop Mapreduce: Which big data framework to choose”, 14 de setembre, 2017. [En línia]. Disponible a: https://www.scnsoft.com/blog/spark-vs-hadoop-mapreduce Consultat: 18 de febrer de 2019

[11] Linode, “Install, configure, and Run Spark on Top of a Hadoop YARN Cluster”, 1 de juny, 2018. [En línia]. Disponible a: https://www.linode.com/docs/databases/hadoop/install-configure-run-spark-on-top-of-hadoop-yarn-cluster/ Consultat: 19 de febrer de 2019

[12] “Vagrant, Getting Started”, San Francisco, California. [En línia]. Disponible a: https://www.vagrantup.com/intro/getting-started/boxes.html [Consultat: 21 Feb. 2019]

[13] “GanttProject - Free project scheduling and management app for Windows, OSX and Linux”. (2019). *GanttProject*. [En línia] Disponible a: https://www.ganttproject.biz/download [Consultat 21 Feb. 2019]

[14] “Trello, la manera gratuita, flexible y visual de organizarlo todo con cualquiera.” (2019). Trello. [En línia] Disponible a: https://trello.com [Consultat 22 Feb. 2019]

[15] Business School OBS, “¿Qué es SCRUM? Características y ventajas”. [En línia]. Disponible a: https://www.obs-edu.com/es/blog-project-management/metodologia-agile/que-es-un-scrum-caracteristicas-y-ventajas [Consultat 20 Feb. 2019]

[16] “The #1 container industry conference for enterprise IT professionals, developers, architects and business leaders.” (2019). Docker. [En línia] Disponible a: https://www.docker.com/ [Consultat 1 Mar. 2019]