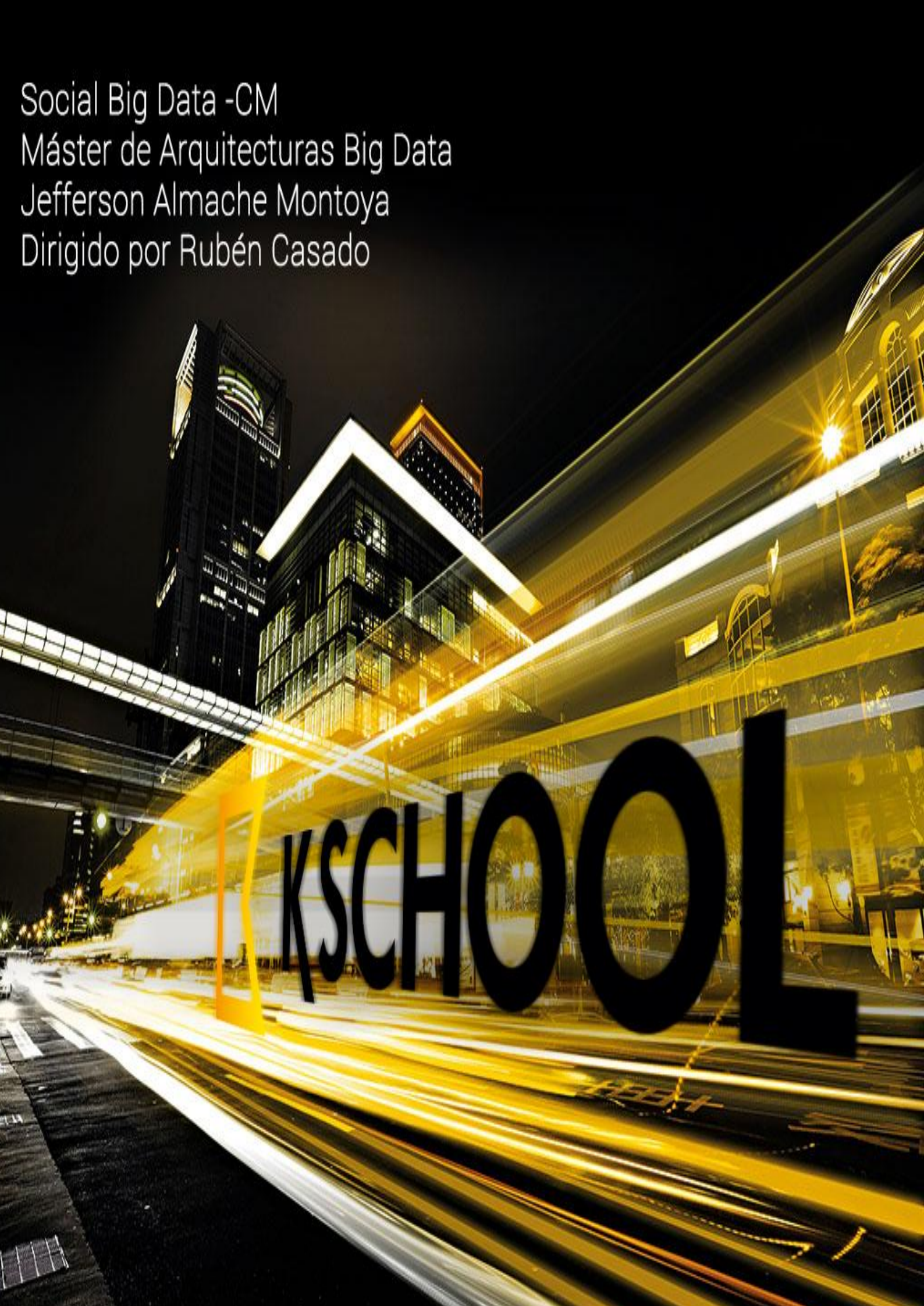
Social Big Data -CM
Máster de Arquitecturas Big Data
Jefferson Almache Montoya
Dirigido por Rubén Casado



Índice

1. INTRODUCCIÓN

Hoy en día, la palabra *Big Data* está de moda y es una de las principales tendencias tanto en el entorno empresarial como en los ámbitos de investigación y educación. Cada segundo, minuto, hora, día, etc.. se está generando un sinfín de información y se prevé que esto siga creciendo año tras año, debido a la digitalización del mundo y la llegada de nuevos dispositivos que van a originar una gran cantidad de información en *streaming*. Con el continuo avance de las distintas tecnologías enfocadas en manipular grandes volúmenes de información, tanto en *batch* como en *streaming,* esta nueva tendencia nos ayuda sacarle valor a los datos, nos permite ser más productivos e ingeniosos y a tomar decisiones que marquen la diferencia.

Este proyecto presenta un piloto que tiene como objetivo servir de base para ayudar a los distintos investigadores de la red social Social Big Data analizar el cambio social a través del Big Data. En este proyecto, se han llevado a cabo la fase de ingesta, procesamiento de datos en streaming, enriquecimiento de los datos y almacenamiento de los mismos. En el proyecto están implicados los grupos de investigación de prestigiosas universidades de Madrid como son la Universidad Complutense, la Universidad Politécnica y la Universidad Nacional de Educación a Distancia. Permiten cubrir las necesidades de este proyecto con la colaboración de expertos punteros en los campos de la Sociología, Geografía, Administración y Dirección de Empresas, (Psicología e Ingeniería) y empresas colaboradoras como Telefónica, ESRI, Indizen Kineo Monility Analytics y Nommon Solutions, además del Consorcio Regional de Transportes que coopera con el análisis de los datos masivos de la tarjeta inteligente del Consorcio, pero con vistas a la investigación y los trámites de la movilidad.

2. CONTEXTO

Los investigadores han creado una red, llamada Social Bigdata-CM que busca analizar los cambios generados por las nuevas tecnologías, así como las desigualdades de la sociedad que se construye en cimientos digitales, los movimientos sociales y la economía tanto colaborativa como social.

La transformación tecnológica se ve reflejada en la sociedad pues evolucionan juntas. La huella que deja en este caso es el BigData. Los *software* tradicionales no pueden ahora mismo manejar la gran cantidad de datos por cuestión de volumen (la capacidad de memoria va más allá de lo hasta ahora establecido), velocidad (pues se requieren resultados a tiempo real) y la variedad.[1] A partir del análisis de esta tecnología, se pueden determinar los patrones sociales de interacción entre individuos y entre estos y las instituciones y empresas. Estos campos afectan a tres pilares de la vida humana: el social, el político y el económico.

En términos sociales: la brecha social se pone de manifiesto con la tecnología, pues las prácticas comunicativas e interactivas evolucionan según los lugares. La interacción digital no puede ser analizada sin tener en cuenta la movilidad física, ya que nuestra forma de comunicación es digital (RRSS, llamadas telefónicas, etc.) que provienen de diversos lugares en los que desarrollan sus actividades dejando de esta forma un rastro digital. En 2020, según un estudio publicado por IDC Digital, cada persona habrá dejado 5247 GB de datos de huellas digitales.[2] La política también ha experimentado un cambio, una gran revolución provocada por las demandas de los ciudadanos y su participación activa en el sistema democrático. Gracias a las redes y la web, los ciudadanos pueden movilizarse y convocar la acción colectiva como movimientos sociales (como sería Black Lives Matter que lucha contra el racismo en Estados Unidos[3]), sindicatos, etc. Se abren nuevas vías para la participación ciudadana que provoca una mejora de los sistemas políticos. En 2012 tras las elecciones norteamericanas fue cuando el término de "Big Data político" cobró fuerza, pues se descubrió que un grupo de informáticos, sociólogos, politólogos y matemáticos (llamados La Cueva) habían estado acumulando datos de los equipos utilizados para la campaña demócrata de 2008 y sumaron una nueva capa de datos. Incluyeron gracias al Big Data Social, datos recogidos de las ciencias sociales. Consiguieron una gran base de datos que incluía los donantes, los datos recogidos por los encuestadores y voluntarios, así como la información y el análisis semántico de las redes sociales y demás credenciales. Los datos se almacenaron en un mismo

[1] Consultado en https://www.researchgate.net/profile/Ruben_Casado/publication/266373455_Emerging_trends_and_technologies_in_big_data_processing/links/55aca13208ae481aa7ff6524.pdf, disponible el 04/06/2017

[2] Consultado en http://www.juancmejia.com/marketing-en-redes-sociales/social-big-data-definicion-e-importancia-de-big-data-en-redes-sociales/, disponible el 04/06/17

[3] Consultado en http://blacklivesmatter.com/about/, disponible el 04/06/17

lugar para analizar la conducta de los votantes y analizar cómo se podía llegar a ellos de forma más directa, con la redacción de un solo mensaje. El objetivo del Big Data en la política consiste en saber qué sienten, por qué, con qué y cómo viven las identidades múltiples en una sociedad en continua conexión. Ocurrió lo mismo en 2015, cuando se aprobó la ley del matrimonio igualitario en EEUU y Facebook desarrolló una app que permitía colocar la bandera multicolor en la foto de perfil. The Atlantic sospechó que lo desarrolló para conocer el respaldo ciudadano de la ley, lo que la empresa negó rotundamente. Por tanto, el Big Data sirve para conocer y pronosticar patrones de conductas. De esta forma, Antoni Gutiérrez-Rubí, insta a evolucionar del Big Data al *Data Thinking* que otorga una oportunidad a la política de mayor sentido. [4]

Por último, cabe destacar el sector económico, pues se están dando una serie de cambios que han transformado la economía, que ha tomado los derroteros hacia la economía colaborativa o de pares. Se refiere a la capacidad de los ciudadanos, que gracias a las plataformas web, organizan servicios cuya base reside en la cooperación. De esta forma también ha cambiado la forma en la que el ciudadano se relaciona con las empresas. El ciudadano ya no es exclusivamente receptor, sino que, a través de las RRSS, impacta en la imagen de las empresas con las que se relaciona. Una de las principales empresas de comercio electrónico a nivel global, Alibaba, anuncia que el Big Data acabará con la economía tal y como la conocemos. Jack Ma (su presidente) anuncia que: "la economía de mercado y planificación quedarán atrás, ya que el Big Data llevará a que el mercado sea más inteligente, y la anticipación sea la norma de las empresas". Establece la metáfora: "un pescador no sale a pescar si las predicciones meteorológicas anuncian tormenta. Los mercados realizarán sus propias predicciones según los datos que recolectan y decidirán qué hacer y qué no". [5]

José Luis Zimmermann, director de la Asociación Española de Economía Digital afirma que: "vivimos en un mar de datos". Francisco Román (presidente de Vodafone España) señala: "al día de hoy, todo lo que viene acompañado de Big Data aporta grandes beneficios". [6]

En España, la gestión inteligente se ha establecido para mantenerse en el auge de las empresas pioneras. El sector ha crecido un 30% en este último año, según la consultora Synergic Partners y seis de cada diez empresas pretenden aplicar nuevas técnicas. China lleva la delantera con un 53% de empresas que utilizan esta tecnología, en un campo que mueve más de 116.000 millones de euros.

[4] Consultado en https://compolitica.com/wp-content/uploads/ACOPPapersN%C2%BA2.pdf, disponible el 04/06/17
[5] Consultado en http://www.economiadigital.es/tecnologia-y-tendencias/alibaba-big-data-economia-actual_408258_102.html, disponible el 04/06/2017
[6] Consultado en http://economia.elpais.com/economia/2016/06/03/actualidad/1464954943_672966.html, disponible el 04/06/2017

3. OBJETIVO

El objetivo principal del proyecto es el de analizar el cambio social a partir del Big Data, para ello se ha montado una arquitectura en *streaming* encargada de extraer, transformar, procesar, enriquecer y almacenar la información de *Twitter,* EMT Bus, EMT BiciMad, Tráfico y la calidad del aire en la ciudad de Madrid.

La arquitectura se pondrá a disposición de los distintos investigadores para analizar, explorar y entender datos de las distintas fuentes de información y sacar la utilidad de los mismos, así como incorporar nuevas ideas y futuras ampliaciones de la arquitectura base.

4. TECNOLOGÍAS UTILIZADAS

El desarrollo de la arquitectura se ha llevado a cabo en el lenguaje de programación *jvm: Scala(2.11)*, que nos aporta la combinación de la programación orientada a objetos y la programación funcional.

Se han utilizado las siguientes tecnologías Big Data:

- **Kafka (0.10.1):** es una cola de mensajes distribuida con el modelo publicación-suscripción que nos ofrece un alto rendimiento y baja latencia. Se ha utilizado Kafka para la parte de ingesta de los datos se ha utilizado kafka, de manera que, se hacen peticiones a web services y estos nos devuelven resultados en ficheros que son transformados y llevados a kafka, normalmente esto no debería ser así sino que se debe obtener la información directamente desde los sensores de las fuentes de información, por motivos de tiempo se ha utilizado la alternativa de hacer las peticiones get a los webservices.

- **Flink(1.3.0):** es una plataforma de streaming distribuido que nos aporta un api sencilla tanto para batch como para streaming. Se ha empleado Flink como motor de procesamiento del proyecto, éste se encarga de consultar las distintas fuentes de información en streaming desde Kafka para transformar y enriquecer los datos para posteriormente almacenarlos en Kafka dependiendo de las ventanas que se hayan definido, es conveniente aclarar que Flink tiene distintos conectores para guardar en un sistema de ficheros como hdfs, amazon s3 y gs.

- **Elasticsearch(5.2):** es una base de datos distribuida que nos permite indexar un gran volumen de datos, además podemos sacar ventaja(entre otras) de sus búsquedas de texto completo que se obtienen de una manera muy rápida gracias a la indexación de la información. Se ha aprovechado elasticsearch principalmente para guardar la información de los data stream de Flink, que tienen elementos convertidos a json, que es la unidad con la que trabaja elasticsearch. En la API nativa de Flink tenemos un conector directo a elasticsearch, proporcionandonos una comodidad a la hora de comunicar ambas tecnologías

- **Kibana(5.2):** es un dashboard que nos permite explotar los datos que se encuentran indexados en Elasticsearch, es importante mencionar que estos datos pueden venir en streaming y Kibana detecta los cambios automáticamente. Además posee una serie de gráficos muy útiles a la hora de visualizar la información.

- **Kafka connect ó secor**: esta herramienta nos permite llevar información desde Kafka hasta un sistema de ficheros como hdfs, s3, gs. Actualmente esta parte de la arquitectura está en modo de desarrollo.

- **Zeppelin(0.7):** es un notebook que nos permite explotar los datos y aplicar análitica más completa con tecnologias como R, Python, Spark, Flink, Elastic, shell entre otros. Se ha empleado esta herramienta principalmente para explorar los datos de Elasticsearch para posteriormente hacer analitica con Python numpy+pandas. Para la posterioridad se usará

Zeppelin para usar la librería de SparkML o FlinkML dependiendo del problema.

Tecnologías adicionales:

Git: Control de versiones; mantenimiento de versiones, eficiencia, centralización del código fuente. Se ha utilizado Git junto con BitBucket para almacenar los dos proyectos mencionados más adelante.

Maven: Para la construcción y gestión de los proyectos. Maven se ha aplicado para construir las dependencias necesarias de los proyectos que están involucrados en la arquitectura.

5. ARQUITECTURA

El esquema general de la arquitectura en streaming involucrando las tecnologías mencionadas en el capítulo anterior sería el siguiente:

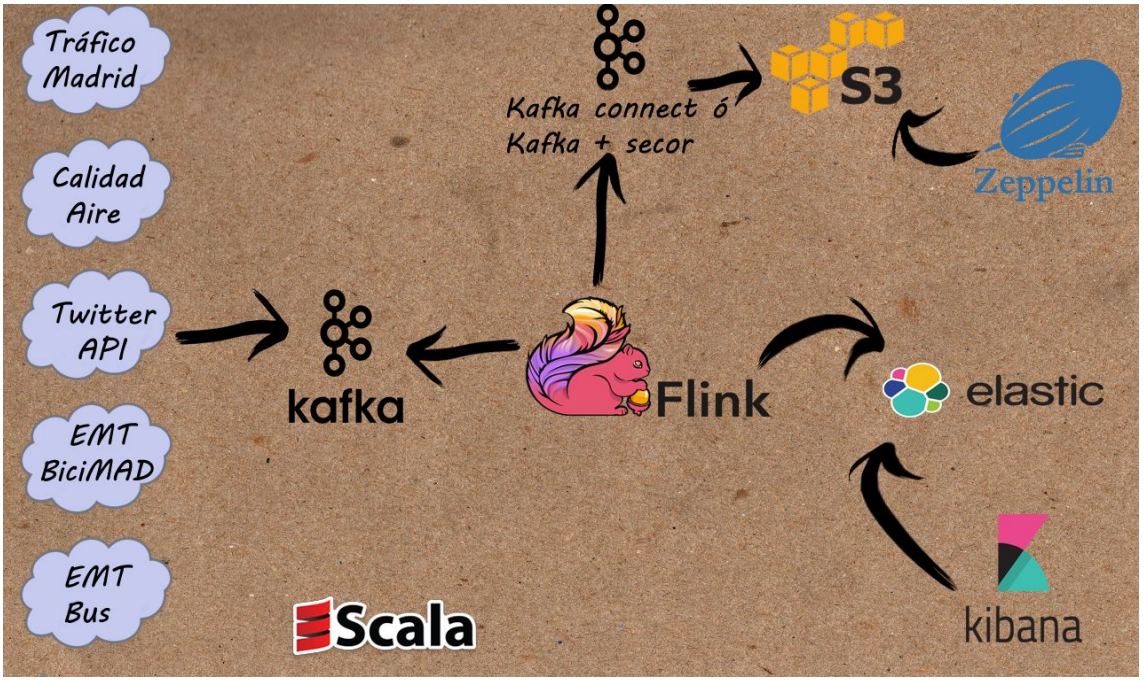


Figura 1. Esquema general arquitectura SocialBigData-CM

La arquitectura se divide en dos proyecto, uno para la parte de ingesta y otro para la parte de procesamiento, enriquecimiento y almacenamiento de los datos. Ambos proyectos se encuentran alojados en BitBucket.

- Ingest project: ssste proyecto se encarga de hacer peticiones a los distintas fuentes de información que devuelven su determinada respuesta, en el caso de que devuelvan ficheros xml, se encarga de transformarlos a formato json para almacenarlos en Kafka(evento a evento), para ello hace uso de la librería liftweb que nos proporciona bastante funcionalidad y que será clave en la arquitectura.

  Se ha decidido por parte de los investigadores que json era un formato flexible, así que se usará este formato como base. Además el proyecto está parametrizado por un archivo.conf que se encuentra en el proyecto, donde aparecen las distintas fuentes de información e información adicional como url, apikeys, topics, broker,etc.

- Engine project: este proyecto es el encargado de recoger la información en raw de las distintas fuentes de información para enriquecerlas utilizando Flink, para esto usa ficheros json adicionales con información acerca de nombres, tipos, coordenadas dependiendo de la fuente de información, una vez que los datos están enriquecidos se procede hacer el cálculo específico(si procede) de un conjunto de datos, así como su almacenamiento.

  Al igual en el proyecto de ingesta, el engine project también tiene su archivo.conf parametrizable donde recoge información como los topics de entrada, los topics de salida y la información acerca de elasticsearch y Kafka, además de los directorios donde tiene que guardar la información en formato json.

5.1 Conjunto de datos

A continuación se explica el conjunto de datos utilizados y un ejemplo de la información en crudo.

Aire:

Campos	Ejemplo en raw
ESTACIÓN MAGNITUD TÉCNICA HORARIO AÑO MES DIA HORA 1 VALIDO HORA 2 VALIDO HORA 3 VALIDO HORA 4 VÁLIDO ……etc	28,079,057,12,08,02,2017,05,21,00077,V,00025,V,00016,V,00020,V,00010,V,00009,V,00009,V,00013,V,00012,V,00014,V,00016,V,00010,V,00016,V,00016,V,00025,V,00022,V,00000,N,00000,N,00000,N,00000,N,00000,N,00000,N,00000,N,00000,N

La información del aire se actualiza cada hora y cada fila contiene los 24 valores horarios de un día, 30 ó 31 filas contiguas corresponde a los valores de los días del mes, repitiéndose con cada magnitud, técnica de todas las estaciones que lo miden.

Tráfico:

Campos	Ejemplo en raw
codigo descripcion accesoAsociado intensidad ocupación carga nivelServicio intensidadSat error subarea	<pm> <codigo>58001</codigo> <descripcion>AV.DE ABRANTES S-N ENTRE VIA LUSITANA Y BESOLLA</descripcion> <accesoAsociado>581803</accesoAsociado> <intensidad>0</intensidad> <ocupacion>71</ocupacion> <carga>71</carga> <nivelServicio>2</nivelServicio> <intensidadSat>1800</intensidadSat> <error>N</error> <subarea>1741</subarea> </pm>
codigo intensidad	<pm> <codigo>PM41261</codigo>

ocupación carga nivelServicio velocidad error	<intensidad>1260</intensidad> <ocupacion>4</ocupacion> <carga>28</carga> <nivelServicio>0</nivelServicio> <velocidad>84</velocidad> <error>N</error> </pm>

Para el tráfico es importante aclarar que nos viene información tanto de tráfico urbano como interurbano y no contienen los mismos campos.

EMTBus:

Campos	Ejemplo en raw
Line SecDetail OrderDetail Node Distance DistStopPrev Name PosxNode PosyNode	<REG> <Line>717</Line> <SecDetail>20</SecDetail> <OrderDetail>1</OrderDetail> <Node>5038</Node> <Distance>2517</Distance> <DistStopPrev>441</DistStopPrev> <Name>Cº POZO TIO RAIMUNDO-Cº HORMIGUERAS</Name> <PosxNode>445012</PosxNode> <PosyNode>4469519</PosyNode> </REG>

Se obtiene el itinerario de una línea, con los vértices para construir las rectas del recorrido y las coordenadas UTM de los ejes viales y los códigos de parada . Es importante aclarar que se van haciendo peticiones a las distintas líneas dependiendo de un archivo que se encuentra en el proyecto de ingesta donde se listan las estaciones, deben hacerse las peticiones con un cierto retraso para no provocar un ataque de denegación de servicio. Esta limitación se comentará en el apartado 6. Conclusiones y consideraciones.

BiciMad EMT:

Campos	Ejemplo en raw
id latitude longitude name	{ "id":84,"latitude":"40.4075685", "longitude":"-3.6902255","name":"AtochaA","light":1,"n

number address activate total_bases dock_bikes free_bases reservations_count	umber":"80a", "address":"Paseo de la Infanta Isabel nº 3", "activate":1,"no_available":0,"total_bases":24, "dock_bikes":16,"free_bases":6, "reservations_count":0 }

Obtiene la relación de todas las estaciones de Bicimad y su estado operacional, la información se va actualizando en tiempo real.

Twitter:

La información acerca de la información que contiene un tweet está disponible en este link donde explica cada uno de sus campos, funcionalidad, etc. En este proyecto sólo nos vamos a centrar en los básicos como el texto, usuario, retweets

5.2 Enriquecimiento

Para explicar los diferentes enriquecimientos se van describir con ejemplos y case classes que nos ayudan a manipular las distintas fuentes de información. Para cada fuente de información se han llevado a cabo los siguientes enriquecimientos:

Aire:

Raw	28,079,057,12,08,02,2017,05,21,00077,V,00025,V,00016,V,00020,V,00010,V,00009,V,00009,V,00013,V,00012,V,00014,V,00016,V,00010,V,00016,V,00016,V,00025,V,00022,V,00000,N,00000,N,00000,N,00000,N,00000,N,00000,N,00000,N,00000,N
Kafka-json	{"estacion" : "28079057","magnitud" : "12","tecnica": "08","horario": "02","fecha": "2017-05-21","listaHoras": [{"hora": "00:00", "valor": "00077", "isValid": "V"},{"hora": "01:00", "valor": "00025", "isValid": "V"},{"hora": "02:00", "valor": "00016", "isValid": "V"},{"hora": "03:00", "valor": "00020", "isValid": "V"},{"hora": "04:00", "valor": "00010", "isValid": "V"},{"hora": "05:00", "valor": "00009", "isValid": "V"},{"hora": "06:00", "valor": "00009", "isValid": "V"},{"hora": "07:00", "valor": "00013", "isValid": "V"},{"hora": "08:00", "valor": "00012", "isValid": "V"},{"hora": "09:00", "valor": "00014", "isValid": "V"},{"hora": "10:00", "valor": "00016", "isValid": "V"},{"hora": "11:00", "valor": "00010", "isValid": "V"},{"hora": "12:00", "valor": "00016", "isValid": "V"},{"hora": "13:00", "valor": "00016", "isValid": "V"},{"hora": "14:00", "valor": "00025", "isValid": "V"},{"hora": "15:00", "valor": "00022", "isValid": "V"},{"hora": "16:00", "valor": "00000", "isValid": "N"},{"hora": "17:00", "valor": "00000", "isValid": "N"},{"hora": "18:00", "valor": "00000", "isValid": "N"},{"hora": "19:00", "valor": "00000", "isValid": "N"},{"hora":

	"20:00", "valor": "00000", "isValid": "N"},{"hora": "21:00", "valor": "00000", "isValid": "N"},{"hora": "22:00", "valor": "00000", "isValid": "N"},{"hora": "23:00", "valor": "00000", "isValid": "N"}] }
Raw case class	Air(28079057,01,38,02,2017-05-21,List(GroupHour(00:00,00005,V), GroupHour(01:00,00004,V), GroupHour(02:00,00004,V), GroupHour(03:00,00005,V), GroupHour(04:00,00005,V), GroupHour(05:00,00005,V), GroupHour(06:00,00005,V), GroupHour(07:00,00005,V), GroupHour(08:00,00005,V), GroupHour(09:00,00005,V), GroupHour(10:00,00005,V), GroupHour(11:00,00004,V), GroupHour(12:00,00004,V), GroupHour(13:00,00004,V), GroupHour(14:00,00004,V), GroupHour(15:00,00004,V), GroupHour(16:00,00000,N), GroupHour(17:00,00000,N), GroupHour(18:00,00000,N), GroupHour(19:00,00000,N), GroupHour(20:00,00000,N), GroupHour(21:00,00000,N), GroupHour(22:00,00000,N), GroupHour(23:00,00000,N)))
Enrichment case class	EAir(Sanchinarro,28079057,-3,660502778,40,49420833,2017-05-21 15:00:00,Dioxido de Azufre(SO2),01,Fluorescencia ultravioleta,28079057,00004,PTE)

Tráfico: Es importante aclarar que se divide en tráfico urbano e interurbano, sin embargo el enriquecimiento que se aplica es el mismo, esto provoca que se lleve tanto a Kafka, Elasticsearch y disco de manera unívoca

- **Urbano:**

Raw	<pm> <codigo>58001</codigo> <descripcion>AV.DE ABRANTES S-N ENTRE VIA LUSITANA Y BESOLLA</descripcion> <accesoAsociado>581803</accesoAsociado> <intensidad>0</intensidad> <ocupacion>71</ocupacion> <carga>71</carga> <nivelServicio>2</nivelServicio> <intensidadSat>1800</intensidadSat> <error>N</error> <subarea>1741</subarea> </pm>
Kafka-json	{ "descripcion":"AV.DE ABRANTES S-N ENTRE VIA LUSITANA Y BESOLLA", "codigo":58001, "nivelServicio":2, "intensidad":0, "intensidadSat":1800, "accesoAsociado":581803, "ocupacion":71, "subarea":1741,

	"error":"N", "carga":71, "timestamp":"2017-05-21T17:05:43.000Z" }
Raw case class	UrbanTraffic(AV.DE ABRANTES S-N ENTRE VIA LUSITANA Y BESOLLA,58001,0,257,1800,581803,3,1741,N,17,2017-05-21T16:45:07.000Z)
Enrichment case class	ETraffic(1705,58001,AV.ABRANTES S-N(AV.POBLADOS-BESOLLA),2017-05-21T16:11:09.000Z,495,154,12,-1,N,-3,735009835,40,37529974)

- **Interurbano:**

Raw	<pm> <codigo>PM41261</codigo> <intensidad>1260</intensidad> <ocupacion>4</ocupacion> <carga>28</carga> <nivelServicio>0</nivelServicio> <velocidad>84</velocidad> <error>N</error> </pm>

Kafka-json	{
	"pm": {
	"codigo": "PM41261",
	"intensidad": "1260",
	"ocupacion": "4",
	"carga": "28",
	"nivelServicio": "0",
	"velocidad": "84",
	"error": "N"
	}
	}

Raw case class	InterUrbanTraffic(PM41261,0,540,81,2,N,12,2017-05-21T16:45:07.000Z)
Enrichment case class	ETraffic(354,PM41261,PM41261,2017-05-21T16:11:09.000Z,494,660,16,77,N,-3,690349781,40,38125022)

Twitter:

Raw

{
"createdAt": "May 21, 2017 5:21:19 PM",
"id": 866312960909737984,
"text": "https://t.co/63GBIkUWHj Para algunos medios informativos d extremaizda antidemocráticos q quieren una dictadura\"Madurista\",le jode la verdad",
"source": "\u003ca href\u003d\"http://twitter.com/download/android\" rel\u003d\"nofollow\"\u003eTwitter for Android\u003c/a\u003e",
"isTruncated": false,
"inReplyToStatusId": -1,
"inReplyToUserId": -1,
"isFavorited": false,
"isRetweeted": false,
"favoriteCount": 0,
"place": {
"name": "Córdoba",
"countryCode": "ES",
"id": "4ecb58704564d392",
"country": "España",
"placeType": “”
…...
……..
}

Kafka-json

{
"createdAt": "May 21, 2017 5:21:19 PM",
"id": 866312960909737984,
"text": "https://t.co/63GBIkUWHj Para algunos medios informativos d extremaizda antidemocráticos q quieren una dictadura\"Madurista\",le jode la verdad",
"source": "\u003ca href\u003d\"http://twitter.com/download/android\" rel\u003d\"nofollow\"\u003eTwitter for Android\u003c/a\u003e",
"isTruncated": false,
"inReplyToStatusId": -1,
"inReplyToUserId": -1,
"isFavorited": false,
"isRetweeted": false,
"favoriteCount": 0,
"place": {
"name": "Córdoba",
"countryCode": "ES",
"id": "4ecb58704564d392",
"country": "España",
"placeType": “”
…...
……..

	}
Raw StatusJSONImpl(TwitterFactory)	StatusJSONImpl{createdAt=Sun May 21 17:18:48 CEST 2017, id=866312325762076672, text='@Helancene au auuu', source='Twitter Web Client', isTruncated=false, inReplyToStatusId=866306883585540096, inReplyToUserId=199009096, isFavorited=false, isRetweeted=false, favoriteCount=0, inReplyToScreenName='Helancene', geoLocation=null, place=PlaceJSONImpl{name='Málaga', streetAddress='null', countryCode='ES', id='507a015cba6ef518', country='España', placeType='city', url='https://api.twitter.com/1.1/geo/id/507a015cba6ef518.json', fullName='Málaga, España' …….. }
Enrichment case class	ETweet(id_str:String, followers_count:String, friends_count:String, lang:String, location:String, screenName:String)

BiciMAD EMT:

Raw	{
	"address": "Paseo de la Infanta Isabel nº 3",
	"latitude": "40.4075685",
	"no_available": 0,
	"number": "80a",
	"light": 1,
	"name": "Atocha A",
	"activate": 1,
	"dock_bikes": 16,
	"free_bases": 6,
	"total_bases": 24,
	"id": 84,
	"reservations_count": 0,
	"longitude": "-3.6902255"
	}

Kafka-json	{ "address": "Paseo de la Infanta Isabel nº 3", "latitude": "40.4075685", "no_available": 0, "number": "80a", "light": 1, "name": "Atocha A", "activate": 1, "dock_bikes": 16, "free_bases": 6, "total_bases": 24, "id": 84, "reservations_count": 0,

	"longitude": "-3.6902255", "timestamp": "2017-05-21 17:01:22" }
Raw case class	BiciMAD(Paseo de la Infanta Isabel nº 3,40.4075685,0,80a,2.0,Atocha A,1,15,7,24,84,0,-3.6902255,2017-05-21 16:56:14)
Enrichment case class	No se ha definido un posible enriquecimiento, se transforma el json en raw

EMT Bus:

Raw	<REG> <Line>717</Line> <SecDetail>20</SecDetail> <OrderDetail>1</OrderDetail> <Node>5038</Node> <Distance>2517</Distance> <DistStopPrev>441</DistStopPrev> <Name>Cº POZO TIO RAIMUNDO-Cº HORMIGUERAS</Name> <PosxNode>445012</PosxNode> <PosyNode>4469519</PosyNode> </REG>

Kafka-json	{
	"Line": 717,
	"OrderDetail": 1,
	"Node": 5038,
	"DistStopPrev": 441,
	"PosxNode": 445012,
	"PosyNode": 4469519,
	"SecDetail": 20,
	"Distance": 2517,
	"Name": "Cº POZO TIO RAIMUNDO-Cº HORMIGUERAS",
	"timestamp": "2017-05-21 16:58:03"
	}

Raw case class	EMTBus(009,1,1846,179,445180,3,4480717,10,8492,MAR CASPIO-PZA.SANTOS HUMOSA,2017-05-21 16:58:03)
Enrichment case class	No se ha definido un posible enriquecimiento, se transforma el json en raw

5.3 Flujo de ejecución

El flujo de ejecución comienza con una consulta a un web service o a un api, dependiendo del caso, esta consulta nos devuelve un objeto que hay que modificarlo, transformarlo para luego ir almacenando evento a evento en Kafka(en su correspondiente topic, dependiendo del config), esta información está en formato json, luego Flink consulta esta información mediante un productor de Kafka, posteriormente Flink enriquece esta información y la guarda en Elasticsearch, Kafka o fichero dependiendo del archivo de configuración, que nos indica topic de salida, ruta del fichero y la información correspondiente a elasticsearch(index, type). La figura 2 muestra los topics que intervienen en el flujo ya descrito.

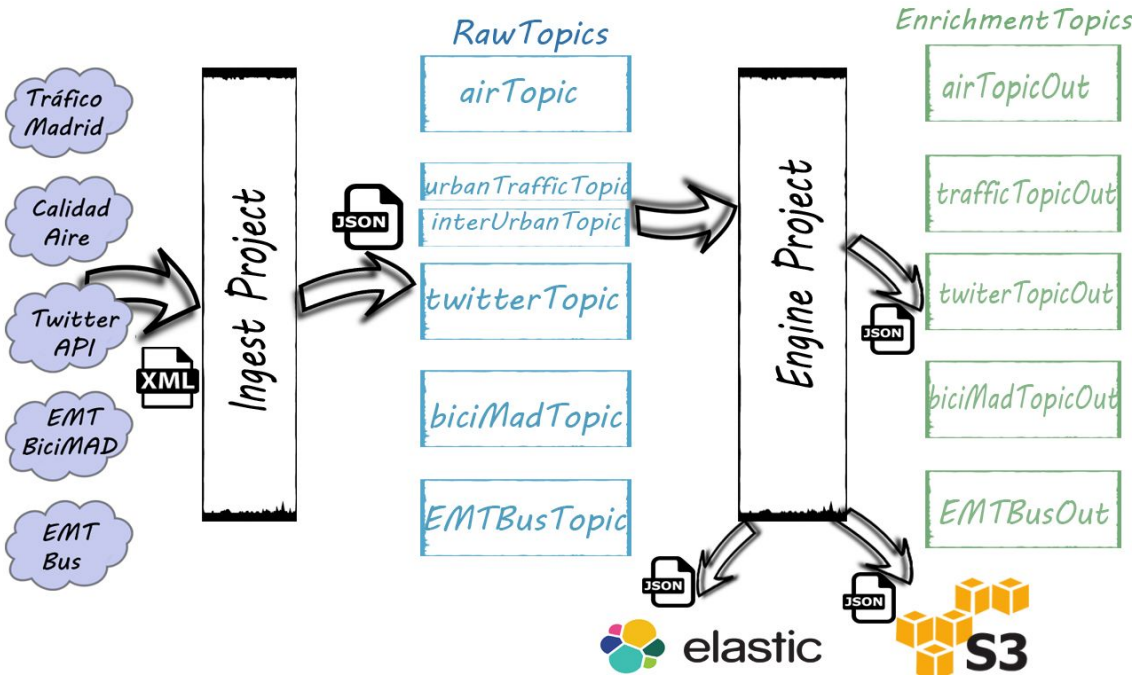


Figura 2. Diagrama arquitectura/topics ingest-engine

6. DIAGRAMAS

En la figura 3 se muestra el diagramas de estados entre los dos proyectos, solamente se ha realizado un diagrama general ya que la idea es la misma en todos los casos, este es un caso general pero se sigue este patrón excepto para *twitter* que realiza la petición a la API.

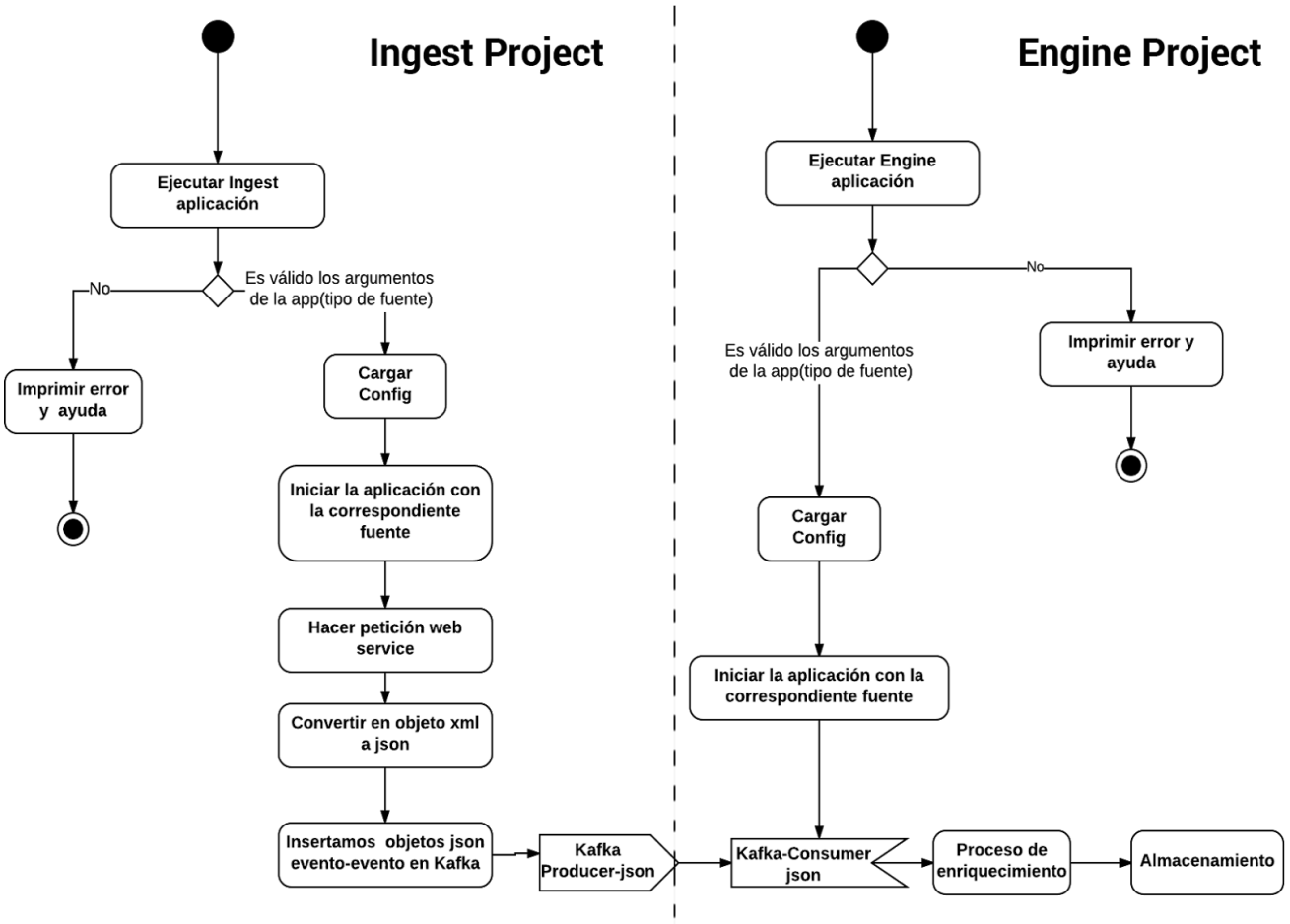


Figura 3 . Diagrama de estados del entre ingest-engine

Para los diagramas de clase se ha decidido hacer un diagrama general de los dos proyectos separados, por lo tanto se ha realizado un diagrama de clase para ingest project y otro para el engine project. Para no perder visibilidad en el diagrama se han puesto las siguientes imagenes: *classDiagram-Ingest.png* y *classDiagram-Ingest.png* dentro del mismo directorio de este documento.

7. CONCLUSIONES Y CONSIDERACIONES

Ha sido una experiencia enriquecedora formar parte de Social Big Data y poder realizar el trabajo de fin de máster desarrollando una arquitectura que sirva de base para los investigadores y sus futuras analíticas.

En cuanto a tecnología el desarrollo con Flink, ha sido bastante cómodo ya que su API es muy sencilla/amigable de usar. Asimismo es una ventaja la semejanza con la api de Spark que tiene mucho más uso en el mundo empresarial. Además, esta tecnología nos proporciona de manera simple los distintos conectores(*sinks/sources),* que tienen una madurez importante en el mundo de las tecnologías Big Data. Flink brinda la posibilidad, además, de implementar tu propio sink/source si hiciera falta.

En cuanto a la arquitectura ésta puede ser mejorada en muchas partes, ya sea añadiendo nuevas fuentes de información, nuevos enriquecimientos, parametrizando aún más los archivos de configuración, formatos de ficheros, etc, pero sin duda la principal mejora sería intentar ir a la fuente de información directamente y no haciendo una petición al web service como se muestra en la Figura 4, ya que esto dificulta que sea un streaming en condiciones.

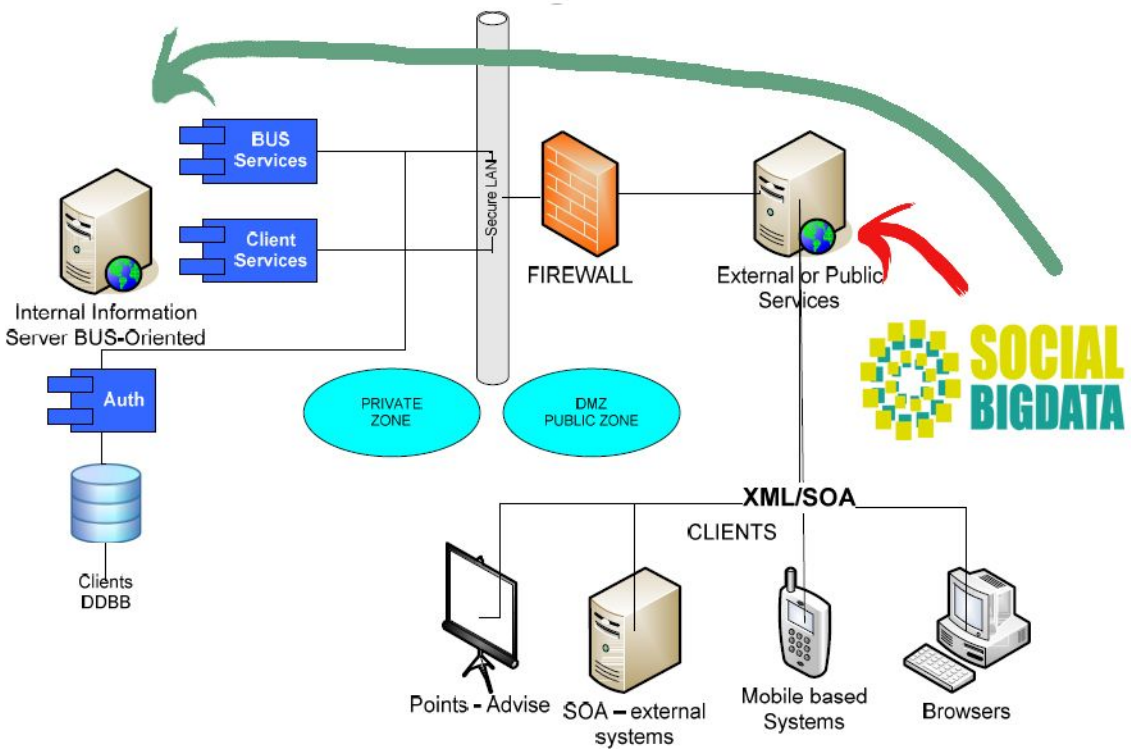


Figura 4. Peticiones web service

8.BIBLIOGRAFÍA Y WEBGRAFÍA

http://socialbigdata.transyt-projects.com/

http://datos.madrid.es/FWProjects/egob/contenidos/datasets/ficheros/Interprete_ficheros_%20calidad_%20del_%20aire_global.pdf

http://datos.madrid.es/FWProjects/egob/contenidos/datasets/ficheros/Transporte_trafico/PUNTOS%20MEDIDA%20TRAFICO_MADRID.pdf

https://dev.twitter.com/overview/api/tweets

http://opendata.emtmadrid.es/Servicios-web

http://www-05.ibm.com/services/es/gbs/consulting/pdf/El_uso_de_Big_Data_en_el_mundo_real.pdf