Alfredo Moreno Muñoz[1] y
Juan Alfonso Lara Torralbo[2]

Análisis de actividad de un servicio de teleasistencia social mediante Big Data y Data Mining

Extracto:

La utilización de Big Data en el momento tecnológico en el que nos encontramos está adquiriendo una gran fuerza e importancia. En las grandes empresas existentes en los principales sectores sociales y de servicios ya han sido implantados sistemas de Big Data que permiten almacenar y tratar toda la información que poseen e incorporarla al día a día de sus clientes y mercados para mejorar los servicios ofrecidos y dar un paso más allá en la relación cliente/empresa.

En teleasistencia, con la llegada de tecnologías IP a los terminales domiciliarios, la comunicación que realizan con la central tendrá lugar a través de internet en lugar de línea de teléfono. Esto permitirá que se empiecen a utilizar sistemas de Big Data debido al incremento de información que se envía desde el terminal al centro de atención. Con el volumen de información que se espera recibir, se podrán descubrir patrones de comportamiento de los usuarios, detectar enfermedades, como, por ejemplo, el alzhéimer, pero, sobre todo, se podrá recibir una información más detallada del estado de todos los dispositivos y sensores instalados en la vivienda del usuario en el momento en el que se produce una llamada de emergencia.

Palabras clave: Big Data, Hadoop, MapReduce, Mahout, minería de datos (Data Mining), teleasistencia, usuarios, llamadas.

Sumario

Fecha de entrada: 21-09-2016
Fecha de aceptación: 25-11-2016

[1] A. Moreno Muñoz, arquitecto de *software* en Tunstall Ibérica.
[2] J. Alfonso Lara Torralbo, profesor de la Universidad a Distancia de Madrid (UDIMA).

Telecare service activity analysis using Big Data and Data Mining

Abstract:

In the current moment that we are living now, the use of Big Data is taken a strength and a very important relevance. The biggest companies of social sector and service sector are using Big Data technologies that allow to store and treat all the information that they have of users and, in a second way, the incorporation of the knowledge of the treatment of this information in the life of the users, in the way of improve the services offered and go to the next step in the relationship of customer/company.

In telecare, with the IP technology in Telecare Unit, the communication between the unit and control centre will be done using internet instead of telephony cable. The companies will start to use these technologies to store all the information that the unit will send to the control center. With all this information, the companies will be able to discover patterns of user's behavior, detect some illnesses like, for example, alzheimer. The most important action that the companies will be able to have is to have more information related to the situation of all devices and sensors installed in user's home when the emergency alarm is raised.

Keywords: Big Data, Hadoop, MapReduce, Mahout, Data Mining, telecare, users, calls.

1. INTRODUCCIÓN

1.1. ¿Qué es la teleasistencia?

La teleasistencia domiciliaria es un sistema tecnológico de atención en la casa de la persona usuaria que puede ser utilizado en situaciones de urgencias. Mediante la línea de teléfono de la vivienda de la persona usuaria se envían alarmas de emergencia a las centrales de teleasistencia. Además de la alarma de emergencia, suelen existir diferentes periféricos asociados al servicio, como pueden ser detectores de humo, gas, agua, etc.

Básicamente, podríamos definir la teleasistencia como un servicio que, a través de la línea telefónica, y con un equipamiento de comunicaciones e informático ubicado en el centro de atención y en el domicilio de la persona usuaria, permite que se establezca una comunicación mediante manos libres con solo presionar el botón del dispositivo. Activo durante las 24 horas del día y los 365 días del año, los usuarios que requieran teleasistencia serán atendidos por personal cualificado que dará la respuesta adecuada en el menor tiempo posible.

Las personas usuarias del servicio de teleasistencia deben poder comunicarse directamente con el centro de atención tantas veces como quieran y a la hora que deseen; por ello, el centro de atención debe permanecer operativo las 24 horas del día, todos los días del año.

Las comunicaciones que pueden establecerse entre el centro de atención y la persona usuaria/terminal de teleasistencia son:

- Comunicaciones producidas por la activación del sistema ante una emergencia.
- Comunicaciones de atención y comunicación interpersonal.
- Comunicaciones de control técnico del sistema.

Los servicios que se prestan pueden ser de dos tipos diferentes:

- **Reactivos.** La persona usuaria envía una alarma desde su domicilio.
- **Proactivo.** El centro de atención se pone en contacto con la persona usuaria.

Dentro del catálogo de servicios que una central de teleasistencia puede ofrecer, destacaremos cuatro:

- Alarma.
- Seguimiento.
- Recordatorio.
- Videoconferencia.

Los servicios de alarma y seguimiento se prestan en situaciones críticas que requieran atención inmediata, como pueden ser caídas, desorientaciones, etc. Los servicios de recordatorio y videoconferencia están enfocados a evitar situaciones de soledad, abandono y aislamiento de la persona usuaria.

Teniendo en cuenta los diferentes servicios que pueden prestarse desde una central de atención y los tipos de servicios, las alarmas serían un tipo de servicio reactivo, mientras que el seguimiento, el recordatorio y la videoconferencia serían servicios proactivos.

La teleasistencia domiciliaria es un sistema tecnológico de atención en la casa de la persona usuaria que puede ser utilizado en situaciones de urgencias

En la prestación del servicio de teleasistencia pueden diferenciarse dos modelos diferentes de la misma:

- Utilizando unidad móvil.
- Sin utilizar unidad móvil.

En la teleasistencia sin unidad móvil la persona usuaria recibe apoyo a distancia desde el centro de atención; mientras que en la teleasistencia con unidad móvil los servicios prestados desde el centro de atención son complementados con la intervención a domicilio en ciertos casos. Lo usual a la hora de prestar el servicio es iniciar la asistencia en remoto desde el centro de atención y movilizar, si fuera necesario, una unidad móvil que se desplace hasta el domicilio de la persona usuaria.

El objetivo principal de la propuesta es el análisis de los datos asociados a las diferentes llamadas, tanto entrantes como salientes, entre el centro de atención y el usuario de teleasistencia.

1.2. Términos relacionados con la teleasistencia

A continuación exponemos una relación de los términos más habituales en el ámbito de la teleasistencia:

- **Usuario de teleasistencia.** Persona que dispone en su casa de un terminal de teleasistencia y es beneficiaria de la prestación del servicio.
- **Terminal de teleasistencia.** Dispositivo mediante el cual el usuario se pone en contacto con el centro de atención.

- **Centro de atención.** *Call centr*e donde se reciben y emiten llamadas de teleasistencia.
- **Llamadas de teleasistencia.** Llamadas entre el centro de atención y los usuarios relacionadas con la prestación del servicio de teleasistencia.
- **Tipo de llamada.** Clasificación de los diferentes tipos de llamadas de teleasistencia que pueden ser recibidas y emitidas entre el centro de atención y los usuarios.
- **Acciones de llamadas.** Acciones derivadas de la llamada entre el usuario de teleasistencia y el centro de atención.
- **Tipo de acciones.** Clasificación de las diferentes acciones que pueden derivarse de una llamada de teleasistencia.
- **Motivo de cierre de una llamada.** Razón de la finalización de la llamada entre el usuario de teleasistencia y el centro de atención.
- **Protocolo.** Lenguaje de comunicación utilizado por el terminal de teleasistencia para ponerse en contacto con la central de telefonía del centro de atención.
- **Línea.** Línea telefónica de la central de telefonía por la que entra o por la que son emitidas las llamadas entre el usuario de teleasistencia y el centro de atención.
- **Unidad móvil.** Empleados cuyo trabajo es ir a atender a los usuarios de teleasistencia a sus domicilios.
- **Operador de teleasistencia.** Empleados cuyo trabajo es recibir las alarmas de los usuarios de teleasistencia y emitir llamadas a los mismos.
- **Sala.** Ubicación donde se encuentran los operadores de teleasistencia.
- **Intervención domiciliaria.** Desplazamiento de un recurso a la vivienda del usuario de teleasistencia.

Activo durante las 24 horas del día y los 365 días del año, los usuarios que requieran teleasistencia serán atendidos por personal cualificado que dará la respuesta adecuada en el menor tiempo posible

2. ESTADO DE LA CUESTIÓN

2.1. Teleasistencia

2.1.1. Descripción del servicio

El nacimiento de la teleasistencia está directamente ligado al inicio de las telecomunicaciones en la medida en que han apoyado la prestación de ayuda a distancia en momentos de urgencia. La teleasistencia es un servicio de atención a distancia, basado en tecnologías de la información y la comunicación (TIC), que es fiable, estable y está siempre disponible desde dondequiera que sea para cualquier persona que necesite recibir apoyo social, sanitario o de otra índole.

La teleasistencia es un servicio que está orientado a personas que se encuentran en situación de dependencia, agrupadas en función de la necesidad que presenten:

- Viven solas o pasan gran parte del día sin compañía.
- Tienen un aislamiento geográfico o desarraigo social.
- Sufren los riesgos causados por la avanzada edad.
- Personas con discapacidad.
- Personas con enfermedades graves o parcialmente dependientes.
- Familiares y/o cuidadores informales.

Desde la perspectiva de la calidad de vida que perciben los usuarios de este servicio, está constatado que la teleasistencia incrementa el bienestar de dichas personas en lo referente a:

- Reducción de sensación de aislamiento.
- Aumento de la seguridad.
- Mejora del acceso a los cuidados sociales o sanitarios.
- Promueve una atención sociosanitaria más continuada.

2.1.2. Terminal de teleasistencia domiciliaria

Los terminales de teleasistencia son dispositivos que se encuentran en los domicilios de las personas usuarias y que deben permitir la comunicación telefónica de voz con el centro de atención, tanto en modo manos libres como en modo normal.

La comunicación entre el centro de atención y el domicilio debe poder realizarse de forma bidireccional y ha de poder activarse tanto por la persona usuaria como por el centro de atención.

Además de la comunicación de voz, los terminales deben permitir:

- Enviar alarmas técnicas a los centros de atención, como, por ejemplo, baterías bajas, autochequeo, etc.
- Configuración remota desde el centro de atención.

Estos terminales deben cumplir los estándares de telefonía implantados en cada país para las llamadas de voz y para las llamadas de datos.

Figura 1. **Tunstall «Lifeline Vi»**

Fuente: http://www.tunstall.com.

Figura 2. **Tunstall «Amie»**

Fuente: http://www.tunstall.com.

La teleasistencia es un servicio de atención a distancia, basado en tecnologías de la información y la comunicación (TIC), que es fiable, estable y está siempre disponible desde dondequiera que sea para cualquier persona que necesite recibir apoyo social, sanitario o de otra índole

En la figura 1 se puede ver un ejemplo de terminal de teleasistencia. En este caso se trata del terminal «Lifeline Vi», de Tunstall Healthcare.

2.1.3. Unidad de control remoto

El terminal de teleasistencia domiciliaria suele ir acompañado por una unidad de control remoto (UCR).

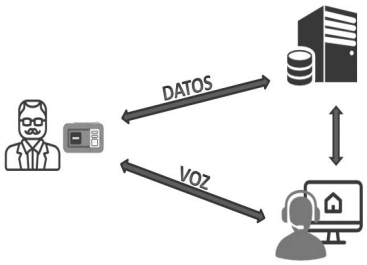
La UCR es un dispositivo en forma de colgante o pulsera que el usuario deberá llevar constantemente puesto. En la figura 2 se puede ver una imagen del pulsador «Amie», de Tunstall Healthcare.

La comunicación entre la UCR y el terminal de teleasistencia se realiza vía radio. Una vez que se pulsa el botón del colgante, este transmite la señal al terminal automáticamente y realiza una llamada de emergencia al centro de atención.

La UCR deberá disponer, al menos, de los siguientes botones de llamada:

- Botón de emergencia para contactar con el centro de atención.
- Botón de llamada configurable para llamar a un número predefinido.

Figura 3. **Comunicaciones teleasistencia**



Fuente: autoría propia.

2.1.4. Central de comunicaciones y atención

Es un centro provisto de tecnología suficiente y capacidad de respuesta para dar cobertura total al servicio de teleasistencia. Es el responsable de la recepción de alarmas y de la emisión de avisos.

El centro de atención y comunicaciones tiene que estar preparado para recibir y enviar las comunicaciones de voz y de datos de forma bidireccional con los domicilios de las personas usuarias del servicio de teleasistencia. En la figura 3 se muestra cómo se bifurcan las comunicaciones desde los terminales domiciliarios de teleasistencia hasta el centro de atención.

Las funciones de un centro de atención son las siguientes:

- Recepción de alarmas.
- Emisión de llamadas.
- Enlazar las comunicaciones entrantes con los datos de los usuarios que las provocan.
- Comunicación con los servicios de emergencias y recursos externos.
- Transferencia de información/llamadas en caso de ser necesario.

2.2. Big Data

En el mundo actual de la informática se tiene la creencia de que todo lo que deseamos hacer con bases de datos podemos llevarlo a cabo con el modelo relacional, pero existen una serie de problemáticas relacionadas con estas bases de datos, por lo que se han creado una serie de herramientas y sistemas que aportar una forma alternativa de atacar problemas y/o mejorar nuestros sistemas.

Esta serie de problemáticas son:

- **Variedad de tipos de datos.** Han aparecido nuevos tipos de datos, como, por ejemplo, los datos no estructurados, que las bases de datos relacionales no pueden almacenar.
- **Escalabilidad.** En la actualidad, las bases de datos relacionales no pueden estar distribuidas en nodos diferentes de forma sencilla y transparente para el usuario, por lo que se debe aplicar una escalabilidad vertical añadiendo CPU y memoria. Pero, lo que se busca en escalabilidad es la escalabilidad horizontal, para poder tener todos los servidores que queramos trabajando en paralelo sin límite alguno.
- **Modelo relacional.** Con este modelo no es posible optimizar al 100 % los sistemas, ya que, por ejemplo, podríamos necesitar tener herencia de objetos, columnas variables según las filas, etc.
- **Velocidad.** En la actualidad, se generan datos de forma muy elevada, por ello, es necesario tener una velocidad de procesado que sea capaz de escalar de forma horizontal, para poder trabajar en paralelo y ahorrar tiempo, siguiendo la técnica del divide y vencerás.

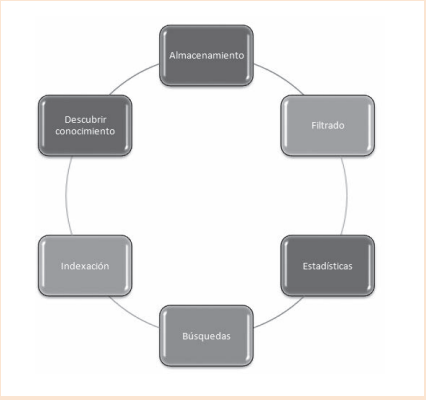
Por todas estas problemáticas surge Big Data, para poder resolver con mayor eficacia estos problemas.

Big Data se caracteriza por las «3 V», que no es otra cosa que «velocidad» a la hora de procesar los datos, gran «volumen» de datos y «variedad» de datos. La figura 4 representa una imagen descriptiva de las «3 V».

Sobre este tipo de datos suelen realizarse diferentes tareas, entre las que destacan las mostradas en la figura 5.

Figura 4. **Propiedades Big Data**

Figura 5. **Tareas Big Data**

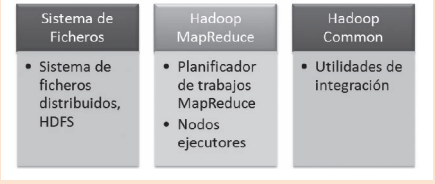


Fuente: autoría propia.

Figura 6. **Logotipo de Hadoop**

Fuente: http://hadoop.apache.org.

Figura 7. **Arquitectura Hadoop**



Fuente: autoría propia.

2.3. MapReduce

Es un proceso de extracción de valores de un gran número de orígenes de datos distintos que está compuesto por:

- **Map.** Extraer y asignar valores a determinadas claves para un único documento.
- **Reduce.** Acumulación y combinación de claves de múltiples documentos para crear un valor reducido único para cada clave a partir de los múltiples valores generados.

2.4. Apache Hadoop

Es una implementación MapReduce, de código abierto e inspirada en los documentos de Google sobre MapReduce y Google File System. El proyecto Hadoop es administrado por Apache Software Foundation y permite el desarrollo de aplicaciones de procesamiento paralelo, permitiendo trabajar con miles de nodos y *petabytes* de datos.

La plataforma Hadoop fue creada por Doug Cutting. El nombre y el logotipo de Hadoop surge por el nombre del elefante del hijo de Doug (véase figura 6).

2.4.1. Arquitectura

La arquitectura de Hadoop está basada en tres pilares fundamentales, que son los que se describen en la figura 7.

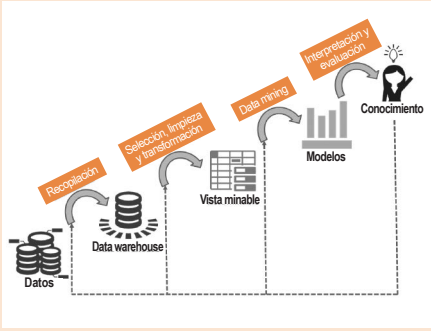
Sobre HDFS (Hadoop Distributed File System) podemos localizar el motor de MapReduce, compuesto por un planificador de trabajos, JobTracker, mediante el cual las aplicaciones envían los trabajos MapReduce. El planificador envía las tareas a los diferentes nodos TaskTracker, que están disponibles en el *cluster*, donde se ejecutarán las operaciones Map y Reduce correspondientes.

2.5. KDD y Data Mining

El proceso de descubrimiento de conocimiento en bases de datos [*knowledge discovery in databases* (KDD)] puede definirse como el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos.

Figura 8. **Propiedades de los datos extraídos**

Figura 9. **Proceso KDD**



Fuente: autoría propia.

Las propiedades de los datos obtenidos a partir del proceso de descubrimiento de conocimiento de una base datos se muestran en la figura 8.

Con las propiedades descritas en la figura anterior, se puede decir que el conocimiento extraído debe ser conocimiento relevante que está oculto en la base de datos y que reporta beneficios su extracción, y, obviamente, se trata de un conocimiento previamente desconocido.

El proceso KDD está compuesto por diferentes etapas que se realizan de forma secuencial; aunque, por la naturaleza del proceso, tiene carácter iterativo, ya que es posible tener que aplicar varias veces el proceso KDD hasta obtener el conocimiento que estamos buscando. Las fases del proceso KDD se observan en la figura 9.

Las fases del proceso KDD se definen de la siguiente forma:

- **Recopilación.** Consiste en la integración de diferentes fuentes de datos en un mismo almacén de datos (*data warehouse*).
- **Selección, limpieza y transformación de datos.** Los datos integrados deben ser tratados antes de realizar el proceso de Data Mining. Hay que realizar una selección de aquellos datos que van a utilizarse y, sobre ese subconjunto de datos, se lleva a cabo un proceso de limpieza y transformado para dejarlos en condiciones de ser tratados en fases posteriores. El objetivo de esta fase es obtener una vista minable para la fase siguiente.
- **Data Mining.** Es considerada la fase más importante del proceso de KDD. Se define como el proceso de exploración y de análisis, por medios automáticos o semiautomáticos, de los datos existentes en la vista minable obtenida en la fase anterior con el fin de descubrir patrones/modelos significativos y reglas. El resultado de la fase son los patrones/modelos de esa minería.
- **Interpretación y evaluación de modelos.** El primer paso de esta fase es la evaluación de los patrones y de los modelos obtenidos, ya que, antes de ser interpretados para la obtención de conocimiento, debe comprobarse que tienen la calidad suficiente para poder realizar la interpretación.

2.6. Apache Mahout

2.6.1. Introducción

Apache Mahout arrancó en 2008 como un subproyecto de Apache Lucene. Se trata de un *framework* de desarrollo para realizar *machine learning* y Data Minig. Apache Mahout incluye algoritmos de clasificación, *clustering* y asociación.

La característica más importante de Mahout es que está compuesto por un conjunto de librerías de *software* libre que permiten llevar a cabo análisis de grandes cantidades de datos en entornos distribuidos. Es frecuente emplear conjuntamente Mahout y Hadoop en proyectos de Big Data.

Figura 10. **Logotipo de Mahout**

Fuente: https://mahout.apache.org.

El nombre de Mahout proviene del significado que tiene, que no es otro que «persona que maneja y conduce elefantes», y, a su vez, el logotipo es una persona encima de un elefante, lo que indica un claro guiño al uso conjunto de Hadoop y Mahout en los proyectos. Podemos ver su logotipo en la figura 10.

Mahout es un proyecto orientado a la investigación. Posee una amplia comunidad por detrás, documentación y ejemplos. Además, se encuentra bajo Apache License. Todos estos puntos convierten a Mahout en uno de los *frameworks* de *machine learning* más potentes y más respaldados de todos los que nos podamos encontrar en el abanico de posibilidades de elección que se tiene en la actualidad.

3. DESARROLLO

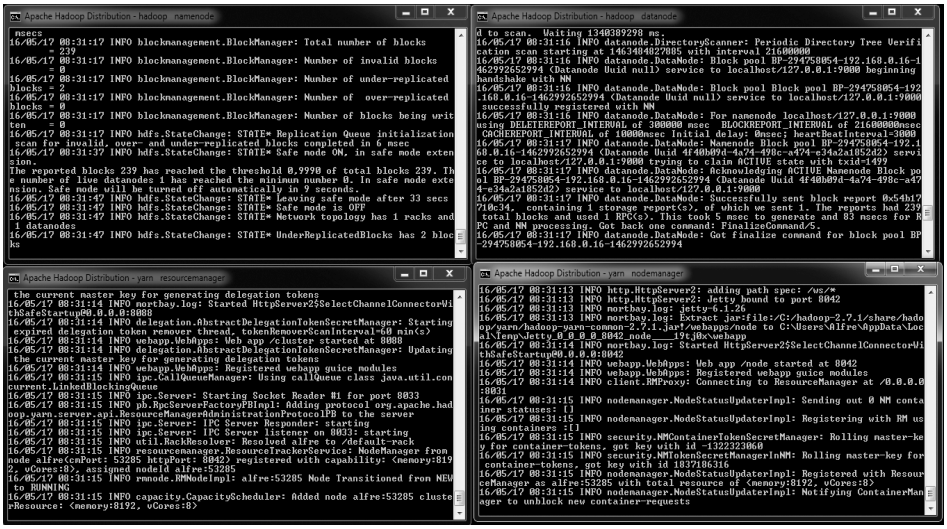
En los siguientes apartados se explicarán todos los puntos referentes al desarrollo del proyecto.

3.1. Big Data

En el desarrollo se ha utilizado Hadoop 2.7.1, instalado sobre Microsoft Windows 7. Una vez se haya instalado y configurado Hadoop, el siguiente paso es arrancarlo. En la figura 11 se muestra la instancia de Hadoop arrancada.

Para el desarrollo se han utilizado ficheros CSV (*comma-separated values*) que contienen información referente a llamadas de teleasistencia. Las acciones que se han realizado sobre esas llamadas y datos demográficos no son de carácter personal.

Figura 11. **Hadoop en funcionamiento**



Fuente: autoría propia.

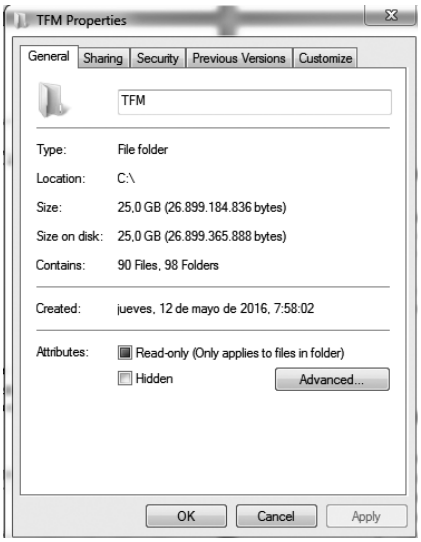
El tamaño de la información con la que se trabajará contiene llamadas hasta mayo de 2016. Contiene un total de 25 GB de información, tal y como puede verse en la figura 12.

La estructura de los ficheros es la siguiente:

- Fecha y hora de la llamada.
- Identificador de expediente de teleasistencia.
- Protocolo de comunicaciones.
- Código de llamada.
- Motivo de cierre de la llamada.
- Texto descriptivo de la llamada.
- Línea de entrada de la llamada.
- Acción realizada sobre la llamada.
- Modelo del terminal de teleasistencia que realiza/recibe la llamada.
- Ciudad del usuario de la llamada.
- Provincia del usuario de la llamada.
- Fecha de nacimiento del usuario de la llamada.

En la figura 13 podemos observar la carga de ficheros en Hadoop.

Figura 12. **Tamaño de datos**



Fuente: autoría propia.

Figura 13. **Datos cargados en Hadoop**

```
C:\hadoop-2.7.1\bin>hadoop fs -put C:/TFM/ /TFM

C:\hadoop-2.7.1\bin>hadoop fs -ls /TFM
Found 9 items
drwxr-xr-x   - Alfre supergroup          0 2016-05-18 19:11 /TFM/2008
drwxr-xr-x   - Alfre supergroup          0 2016-05-18 19:11 /TFM/2009
drwxr-xr-x   - Alfre supergroup          0 2016-05-18 19:12 /TFM/2010
drwxr-xr-x   - Alfre supergroup          0 2016-05-18 19:14 /TFM/2011
drwxr-xr-x   - Alfre supergroup          0 2016-05-18 19:16 /TFM/2012
drwxr-xr-x   - Alfre supergroup          0 2016-05-18 19:18 /TFM/2013
drwxr-xr-x   - Alfre supergroup          0 2016-05-18 19:21 /TFM/2014
drwxr-xr-x   - Alfre supergroup          0 2016-05-18 19:25 /TFM/2015
drwxr-xr-x   - Alfre supergroup          0 2016-05-18 19:26 /TFM/2016

C:\hadoop-2.7.1\bin>hadoop fs -ls /TFM/2015
Found 12 items
-rw-r--r--   1 Alfre supergroup  493149613 2016-05-18 19:21 /TFM/2015/01.csv
-rw-r--r--   1 Alfre supergroup  440568506 2016-05-18 19:21 /TFM/2015/02.csv
-rw-r--r--   1 Alfre supergroup  487152548 2016-05-18 19:22 /TFM/2015/03.csv
-rw-r--r--   1 Alfre supergroup  448892071 2016-05-18 19:22 /TFM/2015/04.csv
-rw-r--r--   1 Alfre supergroup  451084730 2016-05-18 19:22 /TFM/2015/05.csv
-rw-r--r--   1 Alfre supergroup  447800374 2016-05-18 19:23 /TFM/2015/06.csv
-rw-r--r--   1 Alfre supergroup  465973484 2016-05-18 19:23 /TFM/2015/07.csv
-rw-r--r--   1 Alfre supergroup  445964027 2016-05-18 19:24 /TFM/2015/08.csv
-rw-r--r--   1 Alfre supergroup  448610057 2016-05-18 19:24 /TFM/2015/09.csv
-rw-r--r--   1 Alfre supergroup  483419491 2016-05-18 19:24 /TFM/2015/10.csv
-rw-r--r--   1 Alfre supergroup  466891878 2016-05-18 19:25 /TFM/2015/11.csv
-rw-r--r--   1 Alfre supergroup  482925491 2016-05-18 19:25 /TFM/2015/12.csv

C:\hadoop-2.7.1\bin>
```

Fuente: autoría propia.

3.2. Data Mining con Mahout

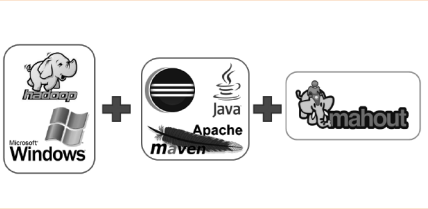
3.2.1. Decisiones de implementación

Para desarrollar la parte de Data Mining se han tomado las siguientes decisiones tecnológicas:

- La versión de Apache Mahout que se va a utilizar será la 0.12.1, liberada el 19 de mayo de 2016.
- El lenguaje de programación para realizar la Data Mining será Java.
- El IDE (*integrated development environment*) para desarrollar será Eclipse Mars.2.
- Las referencias de Mahout serán cargadas utilizando Apache Maven desde Eclipse.

En un primer momento se iba a utilizar la línea de comandos para la ejecución de Mahout, pero ha sido deshabilitada a partir de la versión 0.10.0.

Figura 14. **Diagrama tecnológico**



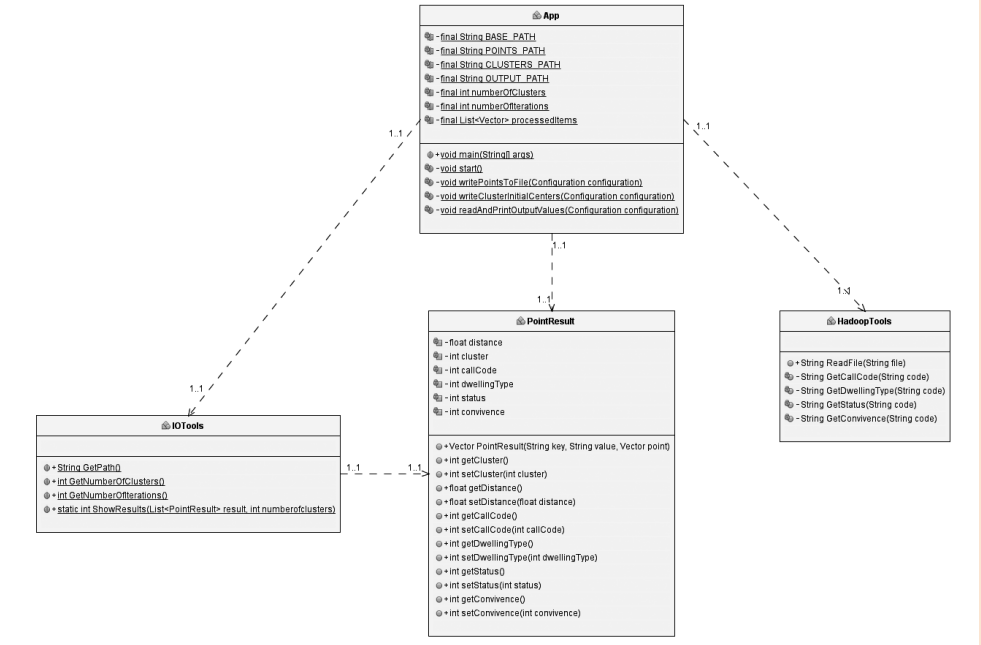
Fuente: autoría propia.

La figura 14 muestra el diagrama tecnológico utilizado.

En la figura 15 se puede ver el diagrama de clases del proyecto.

El algoritmo seleccionado para realizar el proceso de Data Mining es el algoritmo de clusterización KMeans.

Figura 15. **Diagrama de clases**



Fuente: autoría propia.

3.2.2. *Arquitectura de la solución*

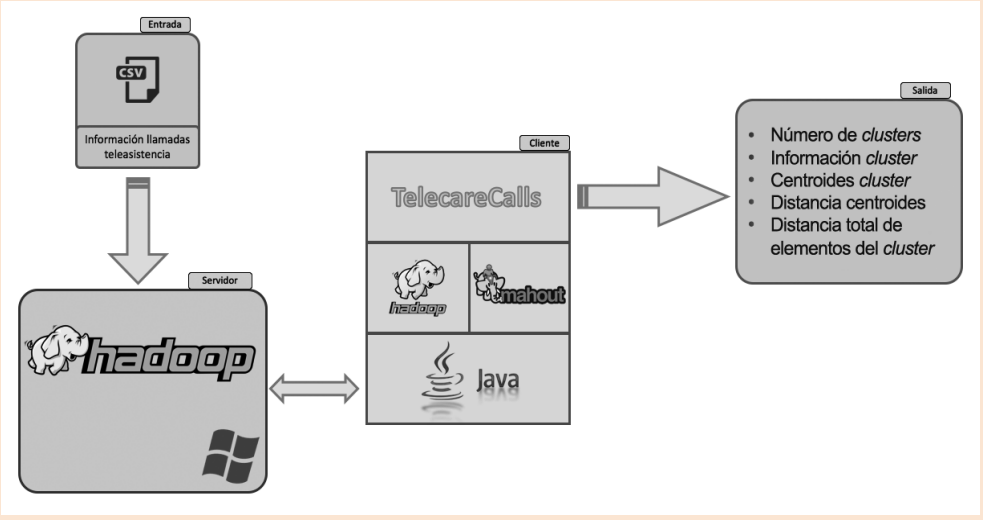
La figura 16 representa la arquitectura modular y funcional de la solución desarrollada.

Las entradas al sistema están compuestas por el conjunto de ficheros CSV que tienen la información relativa a las llamadas. Estos ficheros son almacenados en el servidor con plataforma Microsoft Windows y utilizando Apache Hadoop para dicho almacenamiento.

En la parte cliente se tiene un programa desarrollado en Java que utiliza los módulos correspondientes de Apache Hadoop para la interacción con el sistema de ficheros almacenado en el servidor. También es usado para la lectura de la información y para el procesamiento de dicha información utilizando el algoritmo KMeans proporcionado por las librerías de Apache Mahout.

Por último, el aplicativo desarrollado en Java realiza un análisis de los datos obtenidos en el procesamiento de KMeans y prepara la información de salida por pantalla.

Figura 16. **Arquitectura**



Fuente: autoría propia.

4. RESULTADO DEL ANÁLISIS

A continuación se va a realizar una ejecución con los siguientes parámetros:

- 6 *clusters*.
- 10 iteraciones.
- Distancia euclídea.

El *clustering* ha agrupado los elementos leídos desde el fichero indicado en 6 *clusters*. Cada *cluster* tiene las siguientes características:

- ***Cluster* 0.** Un 12 % del total son alarmas de llamadas técnicas, de usuarios con un tipo de vivienda desconocida, un estado civil desconocido y que viven con una persona.

- ***Cluster* 1.** Un 11 % del total son llamadas de gestión del servicio de usuarios viudos que viven solos en una vivienda unifamiliar urbana.
- ***Cluster* 2.** Un 10 % del total son alarmas de gestión del servicio de usuarios casados que viven con una persona en un edificio de vecinos.
- ***Cluster* 3.** Un 39 % del total son llamadas en segundo plano de usuarios en situación desconocida que viven con una persona en un edificio de vecinos.
- ***Cluster* 4.** Un 19 % del total son llamadas de gestión del servicio de usuarios viudos que viven solos en un edificio de vecinos.
- ***Cluster* 5.** Un 6 % del total son alarmas de emergencias social o sanitaria, en situación y compañía desconocida y en un tipo de vivienda también desconocido.

Tal y como puede observarse en los indicadores de bondad de las particiones, la distancia entre los diferentes centroides de los distintos *clusters* es parecida. Eso indica que la distancia entre ellos es equitativa respecto a todos los centroides de *cluster*.

Figura 17. **Resultados de la ejecución**

```
################## RESULTADOS ##################

--- PARTICIONES ---

Número de clusteres: 6

El clúster número 0 está formado por 99138 elementos. 12% del total.
El clúster número 1 está formado por 93696 elementos. 11% del total.
El clúster número 2 está formado por 80015 elementos. 10% del total.
El clúster número 3 está formado por 315866 elementos. 39% del total.
El clúster número 4 está formado por 154482 elementos. 19% del total.
El clúster número 5 está formado por 49824 elementos. 6% del total.

------------------

--- CENTROIDES ---

Descripción de valores:
        1.- Indica el código de llamada
        2.- Indica el tipo de vivienda
        3.- Indica el estado civil del usuario de la llamada
        4.- Indica el número de convivientes

El elemento central del clúster 0 es: Llamada técnica,Otra/Desconocida,Otro,Vive con 1 persona
El elemento central del clúster 1 es: Gestión del servicio,Unifamiliar urbana,Viudo/a,Vive solo
El elemento central del clúster 2 es: Gestión del servicio,Edificio de vecinos,Casado/a,Vive con 1 persona
El elemento central del clúster 3 es: Llamada segundo plano,Edificio de vecinos,Otro,Vive con 1 persona
El elemento central del clúster 4 es: Gestión del servicio,Edificio de vecinos,Viudo/a,Vive solo
El elemento central del clúster 5 es: Emergencia social/sanitaria,Otra/Desconocida,Otro,Sin datos

------------------

--- INDICADORES DE BONDAD DE LAS PARTICIONES ---

La distancia total entre todos los puntos del cluster 0 es 242555.0491475221
La distancia total entre todos los puntos del cluster 1 es 135482.55791842422
La distancia total entre todos los puntos del cluster 2 es 130452.9215241147
La distancia total entre todos los puntos del cluster 3 es 981484.4583260308
La distancia total entre todos los puntos del cluster 4 es 127840.29056102465
La distancia total entre todos los puntos del cluster 5 es 284151.0810550536

La distancia total entre el centro del cluster 0 y el resto de centros es 37.77012621789048
La distancia total entre el centro del cluster 1 y el resto de centros es 38.539457622078714
La distancia total entre el centro del cluster 2 y el resto de centros es 37.33380552670093
La distancia total entre el centro del cluster 3 y el resto de centros es 44.66537066882352
La distancia total entre el centro del cluster 4 y el resto de centros es 36.69236625064385
La distancia total entre el centro del cluster 5 y el resto de centros es 49.705638079973546

------------------------------------------------

################################################
```

Fuente: autoría propia.

5. CONCLUSIONES

Big Data y Machine Learning son un valor muy al alza en empresas tecnológicas que quieren tener un aspecto diferenciador respecto al resto del sector.

En teleasistencia, la llegada de terminales IP producirá un incremento muy considerable de la información que ahora mismo se envía a través de líneas de teléfono mediante la negociación de protocolo entre la central de telefonía y el terminal de teleasistencia. A través de IP, dicha información será enviada en forma de paquete sin coste de llamada, por lo que se tenderá a disponer de grandes volúmenes de información de cada persona usuaria de teleasistencia.

La nueva información que está por llegar al mundo de la teleasistencia producirá que las grandes compañías necesiten de sistemas de Big Data para su almacenamiento y procesado optimizado. Al disponer de tantos datos, las empresas podrán empezar a utilizar Machine Learning para predecir comportamientos de usuarios, detectar situaciones análogas de usuarios, etc.

Podemos afirmar que Big Data y Machine Learning están a punto de llegar a nuestras vidas en el entorno sociosanitario y que van a llegar para quedarse, ya que mejorarán la calidad de vida de las personas usuarias del servicio.

Tecnológicamente hablando, y centrados en las dos plataformas tecnológicas que se han utilizado en este trabajo, Hadoop y Mahout, es posible decir que Hadoop es un producto maduro que puede ser utilizado en entornos de producción y que es empleado por empresas punteras del sector tecnológico; mientras tanto, Mahout es un producto estrella en sectores de investigación, que es usado también por empresas punteras para realizar investigaciones. Pero en entornos muy potentes y grandes, las empresas están utilizando otras tecnologías que incluyen Machine Learning y que pueden ser utilizadas con objetivos diferentes, no únicamente para aprendizaje.

6. LÍNEAS FUTURAS

El estudio desarrollado en este artículo es un trabajo totalmente funcional que permite la clusterización de llamadas para obtener información sobre los diferentes grupos de llamadas principales que tiene el servicio de teleasistencia en el periodo de tiempo que contenga el fichero CSV que se le pasa por parámetro.

Además de ser totalmente funcional, el trabajo es totalmente ampliable, pudiendo seguir la línea de trabajo desde diferentes posibilidades, entre las que destacan:

- Utilización de otros algoritmos incluidos en Mahout.
- Mostrar los resultados gráficamente.
- Almacenar los resultados en Hadoop.
- Presentar los resultados en fichero pdf.
- Presentar los resultados en fichero de hoja de cálculo.
- Enviar resultados por correo electrónico.
- Permitir navegar en la estructura de ficheros de Hadoop para indicar el fichero a procesar.
- Integración con terceros que permita recibir ficheros CSV mediante un servicio web y devolver el procesado del mismo.
- Ampliar el sistema a telemedicina.
- Integrar Hadoop y Mahout en la plataforma Microsoft .NET.

7. BIBLIOGRAFÍA

Alex Ott´s Blog: *Getting started with examples from «Mahout in Action»*. Disponible en: http://alexott.blogspot.com.es/2012/07/getting-started-with-examples-from.html [Consultado: abril de 2016].

Apache Hadoop: http://hadoop.apache.org/ [Consultado: abril de 2016].

Chimpler Blog: *Using the Mahout Naive Bayes classifier to automatically classify Twitter messages*. Disponible en: https://chimpler.wordpress.com/2013/03/13/using-the-mahout-naive-bayes-classifier-to-automatically-classify-twitter-messages/ [Consultado: abril de 2016].

Cook, S. [2013]: «Mahout item recommender tutorial using Java and Eclipse», *YouTube*. Disponible en: https://www.youtube.com/watch?v=yD40rVKUwPI [Consultado: abril de 2016].

Data Mining: *Tareas del Data Mining*. Disponible en: http://datamining-ucm.github.io/docs/tareas.html [Consultado: mayo de 2016].

Fayyad, U.; Piatetsky-Shapiro, G. y Smyth, P. [1996]: *Knowledge discovery and data mining: towards a unifying framework*. Disponible en: https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf [Consultado: abril de 2016].

[1997]: *From data mining to knowledge discovery in databases*. Disponible en: http://www.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf [Consultado: mayo de 2016].

Ficheros de Windows necesarios para ejecutar Hadoop: https://github.com/sardetushar/hadooponwindows/archive/ master.zip [Consultado: abril de 2016].

GitHub: *CSVToMahout.java*. Disponible en: https://github.com/josephmisiti/hadoop-examples/blob/master/mahout/clustering/CSVToMahout.java [Consultado: abril de 2016].

Source code for «Mahout in Action» book. Disponible en: https://github.com/tdunning/MiA [Consultado: mayo de 2016].

Grupo Fivasa: *Tareas en Data Mining*. Disponible en: http://grupofivasa.blogspot.com.es/2009/09/tareas-en-data-mining.html [Consultado: mayo de 2016].

Hadoop on the Road: *Interfaz Java (FileSystem)*. Disponible en: http://hadoopontheroad.blogspot.com.es/2013/02/hdfs-interfaz-java.html [Consultado: abril de 2016].

IBM Developer Works: *Introducing Apache Mahout*. Disponible en: http://www.ibm.com/developerworks/java/library/j-mahout/ [Consultado: mayo de 2016].

Imserso [2005]: *Guía de teleasistencia domiciliaria*, España: Ministerio de Sanidad, Servicios Sociales e Igualdad.

Java: http://www.oracle.com/technetwork/java/index.html [Consultado: abril de 2016].

Lara, J. A. [2014]: *Integración de bases de datos*, Madrid: Centro de Estudios Financieros.

Marín, J. M.: *Introducción a Data Mining*. Disponible en: http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/introduccion-DM.pdf [Consultado: abril de 2016].

Oliver, A. C. [2014]: «Enjoy machine learning with Mahout on Hadoop», *JavaWorld*. Disponible en: http://www.javaworld.com/article/2241046/big-data/enjoy-machine-learning-with-mahout-on-hadoop.html [Consultado: mayo de 2016].

Redko, A. [2012]: «Apache Mahout getting started», *Java Code Geeks*. Disponible en: https://www.javacodegeeks.com/2012/02/apache-mahout-getting-started.html [Consultado: mayo de 2016].

Safe Living (blog): https://safeliving.wordpress.com/ [Consultado: abril de 2016].

Soft Computing and Intelligent Information System: *Big Data: algorithms for data preprocessing, computational intelligence, and imbalanced classes*. Disponible en: http://sci2s.ugr.es/BigData [Consultado: abril de 2016].

SolidQ: *Big Data mining with Mahout*. Disponible en: http://summit.solidq.com/big-data-mining-mahout/ [Consultado: mayo de 2016].

The Big Data BlOG: http://thebigdatablog.weebly.com/ [Consultado: abril de 2016].

TooDey: http://toodey.com y http://toodey.com/2015/08/10/hadoop-installation-on-windows-without-cygwin-in-10-mints/ [Consultado: abril de 2016].

Tunstall: http://www.tunstall.com/ [Consultado: abril de 2016].

Universidad Carlos III de Madrid [octubre 2015]: *Fundamentals of Big Data Software and Hardware Technologies*.

Wikipedia: *Bid Data*. Disponible en: https://es.wikipedia.org/wiki/Big_data [Consultado: abril de 2016].

K-means. Disponible en: https://es.wikipedia.org/wiki/K-means [Consultado: mayo de 2016].

White, T. [2015]: *Hadoop the definitive guide*, EE. UU.: O'Really.

YouTube: *Apache Mahout tutorial for beginners*. Disponible en: https://www.youtube.com/watch?v=zvfKH9Yb0s0 [Consultado: mayo de 2016].