Aplicaciones de la Big Data a la Computación Urbana.

Robert Sneyder García Moreno.

Universidad Tecnológica de Pereira
Facultad de Ingenierías
Programa de Ingeniería de Sistemas y Computación
Pereira, Risaralda
2017

Aplicaciones de la Big Data a la Computación Urbana.

Robert Sneyder García Moreno.

Director,

Ingeniero Omar Iván Trejos Buriticá

Universidad Tecnológica de Pereira
Facultad de Ingenierías
Programa de Ingeniería de Sistemas y Computación
Pereira, Risaralda
2017

1.Generalidades

1.1 Titulo

Aplicaciones de la Big data a la computación urbana.

1.2. Introducción

La acelerada evolución tecnológica que se presenta actualmente en las sociedades es el resultado del surgimiento de nuevas necesidades que parecen, que en la anterioridad no existían, es por eso que se puede observar que antes de la aparición de la telefonía móvil las personas simplemente se limitaban a comunicarse por teléfonos fijos ,y aunque ahora esto podría parecer algo inútil pues por medio de un smartphone se puede contactar a cualquier persona en el momento que se desee y sin importar la distancia a la que se encuentre, antes este tipo de comunicación era prácticamente el único medio al que se acudía para realizar dicha acción. A medida que incrementa la cantidad de dispositivos y las necesidades que estos van generando con su aparición ,también aumenta el volumen de la información generada ya que estos están en constante proceso de recepción y envío de datos. En los últimos años han aparecido muchas personas y organizaciones (publicas y privadas) que han empezado a sacar provecho de esta información creando estrategias para solucionar problemas que se presentan en la sociedad.

Debido a la gran cantidad de datos generados en la actualidad, se hace necesario buscar soluciones óptimas para procesar dicha información de la mejor manera posible, es por esto que hace relativamente poco tiempo aparece lo que ahora se conoce con el nombre de BIG DATA , que básicamente consiste en un conjunto de herramientas informáticas destinadas a la

manipulación, gestión y análisis de grandes volúmenes de datos de todo tipo, los cuales no pueden ser gestionados por las herramientas informáticas tradicionales [1]. Haciendo uso de estas tecnologías (y de otras que están estrechamente relacionadas) y enfocados a los entornos urbanos nace el concepto de computación urbana que aprovecha toda la información arrojada por una inmensa cantidad de dispositivos para crear soluciones a los diferentes problemas que se pueden presentar en la urbe , soluciones que pueden ser aplicaciones o servicios que permitan a la gente desplazarse, coordinarse, interrelacionarse socialmente en las ciudades, y que deberían incrementar la calidad de vida urbana[2].

La siguiente monografía está diseñada de forma práctica y sencilla para empezar a conocer el reciente mundo de la computación urbana, recorriendo los conceptos y características que hacen parte de esta herramienta y mostrando los beneficios de la incursión en esta área de la informática, también se pretende mostrar las aplicaciones de esta y la utilización de tecnologías de Big Data para la creación de soluciones en entornos urbanos. Del mismo modo se intenta proporcionar una serie de conclusiones que sirvan como base a quien le pueda interesar para posteriores estudios en este tema.

1.3 Planteamiento Del Problema

En la actualidad, la cantidad de información que se genera por segundo es impresionante, pues se puede ver que la cantidad de dispositivos aumentan cada vez más, entonces nace la necesidad de hacer de estos datos un “diamante en bruto” que al ser aprovechados se pueda convertir en un modelo de negocio muy rentable y que ayuden a proporcionar soluciones especificas dentro de una sociedad. Anteriormente la heterogeneidad y el gran volumen en la información suponía un gran problema que costaba mucho solucionar, pero gracias a las últimas herramientas tecnológicas

que han emergido, esta situación ha dado un beneficioso giro , pues la aplicación de tecnologías como lo es big data y sus distintas plataformas han simplificado en gran medida tal hecho, brindando la oportunidad de resolver de manera optima a asuntos muy específicos .La creciente ola de tecnologías emergentes hace que todos los días se deban actualizar los conocimientos que ya se han adquirido; una de las recientes herramientas surgidas es la computación urbana, que se apoya de otras plataformas tecnológicas para crear soluciones en entornos urbanos.

El campo de la ingeniería de sistemas y computación, y en general todas las áreas afines a la informática supone una gran variación en sus contenidos en lapsos de tiempos relativamente cortos, es por esto que surge la necesidad de estar en constante actualización en los temas que afectan directa e indirectamente a estas áreas del conocimiento. Parte de este proceso de actualización es el estudio e investigación de nuevas tecnologías, por esto en esta monografía se pretende establecer una base bien fundamentada apoyada por fuentes confiables para quien o quienes se encuentre interesados en el tema de computación urbana, dando claras definiciones a los conceptos relacionados al mismo, pasando también por las características de todo este entorno y las aplicaciones de este dentro del ecosistema urbano

1.4. Objetivo General

Realizar un estudio elaborado acerca de la computación urbana y las aplicaciones de los entornos big data en esta.

1.5. Objetivos Específicos

- Realizar una investigación sobre todos los conceptos relevantes relacionados con la computación urbana.

- Conocer las aplicaciones de los entornos big data en la computación urbana.
- Identificar las necesidades que la computación urbana intenta resolver.
- Establecer una serie de conclusiones que apoyen procesos de investigaciones futuras.

1.6 Metodología

Para lograr los objetivos específicos propuestos a continuación, se especifican las actividades propuestas para cada uno de ellos.

Objetivo 1: Realizar una investigación sobre todos los conceptos relevantes relacionados con la computación urbana:

- Realizar la lectura y análisis de los artículos referentes a computación urbana encontrados en bases de datos suscritas a la universidad como lo son la IEEE y ELSEVIER, y otros recursos donde haya información pertinente al tema.
- Identificar los conceptos básicos y relevantes que abarca la computación urbana.

Objetivo 2: Conocer las aplicaciones de los entornos big data en la computación urbana:

- Identificar las plataformas y herramientas de big data existentes utilizadas en la computación urbana.
- Comparar las diferentes tecnologías y técnicas de big data utilizadas en la computación urbana

Objetivo 3: Identificar las necesidades que la computación urbana intenta resolver:

- Conocer los objetivos del desarrollo de la computación urbana.
- Averiguar algunos proyectos a nivel mundial que utilizan o pretenden usar la computación urbana como medio de desarrollo del mismo.

Objetivo 4: Establecer una serie de conclusiones que apoyen procesos de investigaciones futuras:

- Elaborar un cuadro comparativo entre las diferentes herramientas utilizadas en la computación urbana.
- Analizar diversas investigaciones,publicaciones y articulos relacionados con la computación urbana.

2.Estado Del Arte.

2.1 Concepto de Big Data.

En los últimos años, los datos se han convertido en uno de los bienes más preciados para las compañías en casi todos los campos. No solo es importante para las empresas relacionadas con la industria de la informática, sino también para otros tipos de organizaciones, tales como las entidades gubernamentales, el sector salud, educación, el sector de las diferentes ingenierías, entre muchas otras más. Los datos generados de diversas fuentes son esenciales para la ejecución de las actividades diarias que se llevan a cabo dentro de la organización, ayudando así a la alta gerencia lograr la consecución de sus objetivos y tomando las mejores decisiones basadas en la información extraída de dichos datos [1]. Se estima que de todo los datos generados en la historia de la humanidad el 90 por ciento sido creado sólo en los últimos años. En el año 2003, se estimó que los datos creados por la humanidad era cerca de cinco exabytes, y esta cantidad en la actualidad se puede generar en tan solo dos días.[2].

Esta tendencia hacia el aumento del volumen y el detalle de los datos que son colectados por las compañías no cambiará en un futuro cercano, el masivo uso de las redes sociales, multimedia y la aparición del internet de las cosas está produciendo una abrumador flujo de datos nunca antes visto. La cantidad de datos que se genera actualmente en su mayoría no posee estructura alguna y es heterogénea , lo que hace más difícil el proceso de análisis de los mismos.

Es por esto que nace el término Big Data el cual hace referencia a un conjunto de datos de tal tamaño y estructura que excede las capacidades de las herramientas tradicionales de programación para la recolección, almacenamiento y procesamiento en un tiempo razonable y con mayor razón también sobrepasa la capacidad de percepción humana. Una de las razones por

las que administrar efectivamente tal volumen de datos es difícil es debido a que los datos generados pueden ser estructurados, semi-estructurados y no estructurados, los cuales se explicaran a continuación [3].

2.2 Tipos de Datos.

Debido a la falta de homogeneidad en los datos generados por las diferentes fuentes de información y ya que, por ejemplo, a la aparición de bases de datos no relacionales, los datos se pueden ser clasificados de acuerdo a su estructura:

2.2.1 Datos estructurados.

En gran medida, los datos generados por las tradicionales fuentes de información son estructurados, pues poseen ciertos formatos y esquemas determinados con campos fijos y rigurosamente establecidos. Este tipo de datos son los que conforman las bases de datos relacionales, hojas de cálculo y diferentes clases de archivos, y se componen de piezas de información que se conoce con anterioridad, además de que se producen en un especifico orden. La ventaja de este tipo de datos es que facilitan en gran manera su manipulación y procesamiento. Ejemplos de datos estructurados son: Fecha de nacimiento (DD/MM/AAA), Documento de identidad, número telefónico, entre otros**.**

2.2.2 Datos semiestructurados.

Este tipo de datos poseen un flujo lógico y un formato en gran medida definido, pero que no es fácil para la comprensión del usuario. Son datos que poseen etiquetas y/o marcadores que permiten realizar una separación sobre los elementos de datos, se pude decir que no tienes formatos fijos. La lectura de datos semiestructurados requiere el uso de reglas complejas que

determinan cómo proceder después de la lectura de cada pieza información. Un ejemplo típico de datos semiestructurados son los registros Web logs de las conexiones a internet; un Web log se compone de diferentes piezas de información, cada una de las cuales sirve para un propósito específico. Ejemplos típicos son el texto de etiquetas de lenguajes XML y HTML.

2.2.3 Datos No Estructurados

Los datos no estructurados son aquellos que no tienen un tipo definido. Normalmente se almacenan en "documentos" u "objetos" sin estructura uniforma, y se tiene poco o ningún control sobre ellos. Datos de texto, video, audio, fotografía son datos no estructurados. Por ejemplo, las imágenes se clasifican por su resolución en píxeles. Algunos ejemplos de datos que no poseen campos fijos son: audio, video, fotografía, documentos impresos, formularios especiales, mensajes de correo electrónico y de texto, artículos, mensajería instantánea, entre muchos otros. Se puede afirmar que los datos más difíciles de analizar son los datos no estructurados, y es gracias a su continuo crecimiento que se han desarrollado herramientas para su manipulación como son el caso de MapReduce, Hadoop o bases de datos NoSQL, de las cuales se hablarán más adelante.

2.3 Características de Big data

Según lo planteado por IBM[4] cuando se habla de Big Data se puede identificar tres características determinantes que aclara el panorama acerca de si el problema que se aborda se trata o no de Big Data, estas características se encuentran inmersas en lo que también es conocido como el "modelo de las tres V" (3 V) y hace referencia a Volumen, Velocidad y Variedad, aunque IBM mismo ha propuesto una cuarta característica (lo que implica un "modelo

de cuarto V") y se refiere a la Veracidad de los datos. Otras fuentes notables añaden una quinta característica y es el Valor, lo que supondría un "modelo de cinco V".

2.3.1 Volumen. El tamaño de los datos disponibles están creciendo a un ritmo cada vez mayor, y esto afecta tanto a las empresas como a las personas particulares. En la época actual existen más fuentes que generan datos de forma continua. Anteriormente, los datos en las empresas eran generados solo por los empleados internos, pero en la actualidad también se crea datos a partir de los socios, clientes y todo aquel que esté relacionado directa o indirectamente con la organización, además de las máquinas y dispositivos involucrados en la infraestructura empresarial [5]. Si se echa un vistazo hacia hace un poco más de una década, los almacenes de datos (data warehouse) de las grandes empresas que tenían de uno a diez terabytes se consideraban enormes; hoy en día cualquier persona puede adquirir unidades de disco de uno a cinco terabytes por precios relativamente bajos.

Según las previsiones de la empresa consultora y de investigación de las tecnologías de la información Gartner Inc., en 2020 más de 25 mil millones de dispositivos estarán conectados a internet, acrecentando un volumen de datos que a finales de 2013 ya se estimaba en 4,4 billones de GB y que llegará, según los pronósticos, a multiplicarse por 10 en tan solo seis años [6].

2.3.2 Velocidad. Para muchas aplicaciones, la velocidad de generación de datos es incluso más importante que el volumen de los mismos. La creación en tiempo real y cercano a esto hace posible que una empresa sea mucho más ágil que sus competidores, lo que le da una fuerte posición en el mercado en el cual se desempeña [7].Inicialmente, las empresas analizaban los datos en procesamiento por lotes(el cual se explicará detalladamente más adelante) , el cual

consiste en tomar pedazos, secciones o como el término lo expones lotes de datos que ya han sido previamente almacenados y se envían a un servidor el cual se encarga de procesarlos y posteriormente enviar el resultado. Este tipo de procesamiento sirve siempre y cuando la transmisión de datos sea más lenta que la velocidad de procesamiento y cuando el resultado es útil a pesar del retardo. Para las nuevas fuentes de datos, como las diferentes aplicaciones móviles y redes sociales, el procesamiento por lotes es de poca utilidad [5] debido a la velocidad de creación, procesamiento y acceso de datos que requieren.

Lo anterior supone que las empresas deben contar en su arquitectura TI con herramientas que le permitan manipular grandes volúmenes de datos con la más óptima velocidad, lo que quiere decir que con los sistemas tradicionales se hace una tarea prácticamente imposible.

2.3.3 Variedad. Desde tablas de Excel, pasando por bases de datos relacionales hasta archivos multimedia, la estructura de los datos ha cambiado para dar paso a cientos de formatos que han aparecido durante los últimos tiempos. Cuando se trata de big data en la mayoría de ocasiones no se tiene control sobre los formatos de datos de entrada, es por eso que se puede considerar datos puros de texto, fotos, audio, video, datos de GPS, SMS, pdf, flash, y muchos otros cientos de formatos existentes en la actualidad. La variedad de los datos entonces está dada por la estructura de los mismos, y considerando que la fuente de estos puede ser de cualquier tipo; así tenemos que los datos como se dijo anteriormente en el apartado 2.2 donde se habla acerca de los tipos, pueden mencionarse la existencia de datos estructurados, semiestructurados y no estructurados.

El volumen de los datos asociado a Big Data implica nuevos retos para los centros de datos que intentan tratar con la variedad de los mismos. Con la explosión de sensores y dispositivos

inteligentes (los cuales están estrechamente relacionados con el internet de las cosas-IoT) , así como las tecnologías de colaboración social, los datos en las empresas se han convertido en un asunto muy complejo, ya que no solo existen los datos relacionales tradicionales, sino que también priman en bruto, datos semiestructurados y no estructurados procedentes de páginas Web, archivos de registro Web (Web log), incluyendo datos de flujos de clic, foros de medios sociales, correo electrónico, documentos, datos de sensores de sistemas activos y pasivos, entre otros.

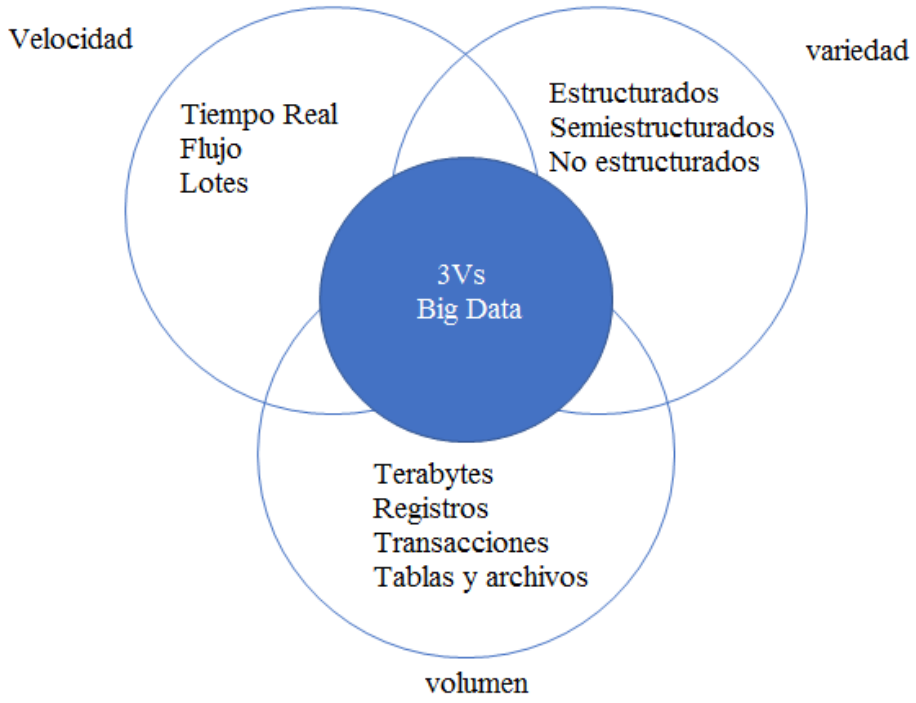

Figura 1. Modelo de 3Vs de Big Data. Moreno, J.,Serrano, M. A.,Fernández E.(2016)., Main Issues in Big Data Security, Future Internet, 8,44.

2.3.4 Veracidad. A medida que el volumen y la variedad de los datos aumentan , una característica como la veracidad toma gran relevancia, pues es la variable menos uniforme a lo largo de los distintos tipos de datos, ya que lleva implícito el sesgo, ruido y la alteración de los mismos. La veracidad hace referencia al nivel de fiabilidad asociado a ciertos tipos de datos. Esforzarse por conseguir unos datos de alta calidad es un requisito importante y un reto fundamental de big data, pero incluso los mejores métodos de limpieza no pueden eliminar la imprevisibilidad inherente de algunos datos, como el tiempo, la economía o las futuras decisiones de compra de un cliente; a pesar de la incertidumbre , los datos siguen conteniendo información valiosa [8].Realizar una óptima gestión de los datos con respecto a la veracidad es imprescindible a la hora de realizar el proceso de toma de decisiones , pues administrar esta variable de forma correcta supone gran beneficio para las empresas ya que eleva la calidad de la información obtenida.

2.3.5 Valor. El éxito de una empresa está en la toma de las mejores decisiones que puedan llevar a la organización a estar fuertemente posicionados en el mercado y entorno en el cual se encuentran. El dato no es valor, pues esta característica no está presente por el solo hecho de recolectar inmensas cantidades de datos. El verdadero valor de los datos reside la transformación de estos en acción o decisión mediante el proceso de conversión de la información arrojada por los datos en conocimiento; o sea, el valor de los datos está en cuán accionables son los mismos para poder tomar las mejores decisiones en base a estos.

Es conveniente entonces saber que no todos los datos de los cuales partimos se tienen que convertir en acción o decisión, y es por eso que existen tecnologías aplicadas que facilitan el proceso de extracción de valor de los datos[9].

2.4 Fuentes de grandes volúmenes de Datos

En la actualidad los datos proceden de innumerables fuentes provenientes de todos los lugares posibles, ya sea de un sensor ubicado en una remota granja o de un celular moderno. Es por esto que la cantidad de datos va creciendo de manera exponencial, lo que ha hecho que en las empresas y organizaciones pasen de terabytes a peta bytes en un tiempo considerablemente corto.

De todas la fuentes, los datos procedentes de la Web son, sin duda alguna, los que ocupan la mayoría a nivel mundial, pues es la fuente más utilizada y reconocida en la actualidad, y probablemente así sea durante muchos años más. Pero, existen muchas otras fuentes que adhieren y aumentan el volumen de datos, algunos de los orígenes más comunes son:

- Datos de la Web,
- Datos de los medios sociales (redes sociales, blogs, wikis, entre otras),
- Datos de Internet de las Cosas (IoT),
- Datos de interconexión entre máquinas, M2M (IoT),
- Datos industriales de organizaciones y empresas,
- Datos de redes y telecomunicaciones,
- Datos de geolocalización,
- Datos personales,
- Datos de texto,
- Otros

Una tendencia clara que se observa a diario es que las tecnologías fundamentales, que contienen y transportan datos, conducen a múltiples fuentes de grandes datos en las industrias más diferentes. A la inversa, diferentes industrias pueden aprovecharse de numerosas fuentes de datos.

2.4.1 Tipos de fuentes de Big Data. Tratar de clasificar todos los tipos de fuentes generadoras de Big Data no es una tarea muy difícil, pero a continuación se trata de dar una clasificación que trate de abarcar los distintos tipos de fuentes:

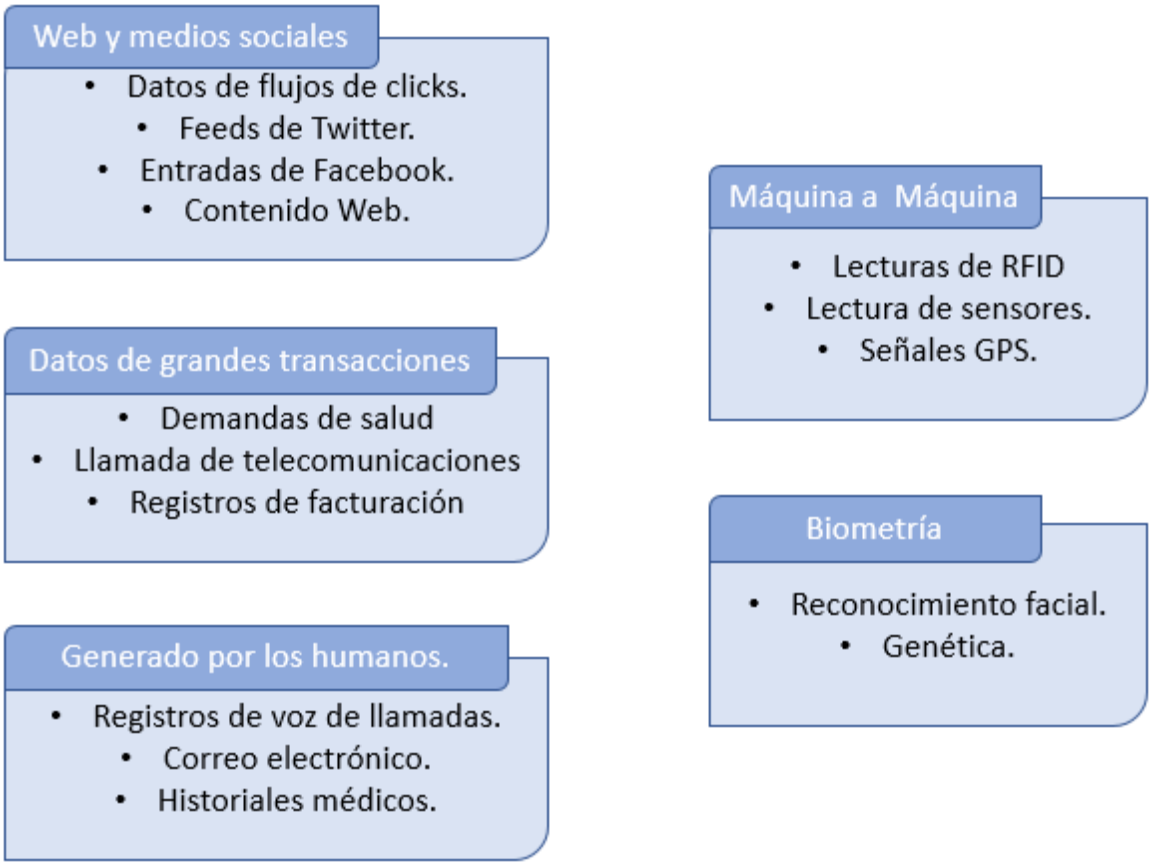

Figura 2.Tipos de fuentes Big Data

2.4.1.1 Web y medios sociales. Incluye el contenido web e información generada en las redes sociales como Facebook, Twitter, Instagram, LinkedIn, Foursquare; blogs, periódicos , televisión; wikis como MediaWiki, Wikipedia; marcadores sociales como Del.icio.us. Stumbleupon, entre muchos otros más. En está categoría, los datos se capturan, almacenan o distribuyen teniendo en cuenta las siguientes características: La información incluye datos procedentes de los flujos de clics, tuits, retuits o entradas en general (feeds) de Twitter , Tumblr, postings de Facebook y sistemas de gestión de contenidos Web diversos como Youtube, Flickr, Picasa; o sitios de almacenamiento de información como Dropbox, Skydrive o OneDrive.

Los datos de la Web y los medios sociales se analizan con herramientas de analítica Web y analítica social mediante el uso de métricas de indicadores KPI.

2.4.1.2 Máquina-a-Máquina (M2M). Se refiere a las tecnologías que permiten conectarse a otros diferentes dispositivos entre sí, M2M utiliza dispositivos como sensores o medidores que capturan algún evento en particular (humedad, velocidad, temperatura, presión, variables meteorológicas, variables químicas como la salinidad, entre otros), los cuales transmiten a través de redes cableadas, inalámbricas y móviles a otras aplicaciones, que traducen estos eventos en información significativa. Entre los dispositivos que se emplean para capturar datos de esta categoría podemos considerar chips o etiquetas RFID, chips NFC, medidores inteligentes (de temperaturas, de electricidad, presión), sensores, dispositivos GPS que ocasionan la generación de datos mediante la lectura de los medidores, lecturas de los chips RFID y NFC, lectura de los sensores, señales GPS, señales de GIS ,etcétera.

La comunicación M2M ha originado el conocido Internet de las cosas o de los objetos, que representa a los miles de millones de objetos que se comunican entre sí y que pueden acceder si es necesario a internet.

2.4.1.3 Transacciones de grandes datos. Son los grandes datos transaccionales procedentes de operaciones normales de transacciones de todo tipo. Incluye registros de facturación, en telecomunicaciones y registros detallados de las llamadas (CDR), entre otros. Estos datos transaccionales están disponibles en formatos tanto semiestructurados como no estructurados. Los datos generados procederán de registros de llamada de callcenters, departamentos de facturación, reclamaciones de las personas, presentación de documentos, etcétera.

2.4.1.4 Biometría. La biometría o reconocimiento biométrico se refiere a la identificación automática de una persona basada en sus características anatómicas o trazos personales. Los datos anatómicos se crean a partir del aspecto físico de una persona, incluyendo huellas digitales, iris, escaneo de la retina, reconocimiento facial, genética, ADN, reconocimiento de voz, incluso olor corporal. Los datos de comportamiento incluyen análisis de pulsaciones y escritura a mano. Los avances tecnológicos han incrementado considerablemente los datos biométricos disponibles. En el área de la seguridad e inteligencia, los datos biométricos han sido información importante para las agencias de investigación. En el área de negocios y de comercio elctronico, los datos biométricos se pueden combinar con datos procedentes de medios sociales lo que hace aumentar el volumen de datos contenidos en los datos biométricos. Los datos generados por la biometría se pueden agrupar en dos grandes categorías: Genética y reconocimiento facial.

2.4.1.5 Datos generados por la personas. Personas generan enormes y diversas cantidades de datos como la información que guarda un centro de llamadas telefónicas (call center) al establecer una llamada, notas de voz, correos electrónicos, documentos electrónicos, estudios y registros médicos electrónicos, recetas médicas, documentos papel, faxes, entre muchos otros. El problema que acompaña a los documentos generados por las personas es que pueden contener información sensible que necesita, normalmente, quedar oculta, enmascarada o cifrada de alguna forma para conservar la privacidad. Por eso, estos datos necesitan ser protegidos por las leyes nacionales o supranacionales (como es el caso de la Unión Europea o el Mercosur) relativas a protección de datos y privacidad.

Los datos contenidos en la clasificación anterior los explicaremos ahora con más detenimiento, pero agrupados ya en categorías más próximas a la actividad diaria de la web.

2.5 Integración de los Datos y oportunidades de negocio con Big Data.

A raíz de toda esta explosión alrededor de Big Data las empresas se empiezan a plantear cómo pueden tomar ventaja de las grandes oportunidades que esto trae, que pueden hacer con las tecnologías Big Data, cómo pueden evitar riesgos, entre muchos otros planteamientos. Un número creciente de organizaciones les hacen frente a dichas cuestiones desplegando herramientas especializadas como bases de datos de procesamiento masivamente paralelo (MPP, Massively Parallel Processing), sistemas de archivos distribuidos Hadoop, algoritmos MapReduce, computación en la nube, etcétera. La pieza clave es la integración de los datos; por esto es crucial para las organizaciones facilitar que los negocios accedan a todos los datos de modo que se puede aplicar sobre ellos infraestructura de Big Data.

La integración de datos facilita a las organizaciones combinar toda una plataforma Big Data con los datos transaccionales tradicionales para generar valor y conseguir la mayor eficacia posible. Por esta razón uno de los aspectos más interesantes no es tanto lo que harán ellos mismos por el negocio, sin i lo que se podrá conseguir para el negocio cuando se combinan con otros datos de la organización. Un buen ejemplo puede ser enriquecedor: utilizar las preferencias y rechazos de los perfiles de los clientes en los medios sociales con el objetivo de mejorar la comercialización de destino.

El mayor valor de Big Data puede producirse cuando se combina con datos corporativos; colocándolos en un contexto más grande se puede obtener que la calidad del conocimiento del negocio se incrementa exponencialmente, incluso la estrategia de Big Data dentro de una estrategia global de la compañía es mucho más rentable que tener una estrategia independiente.

Se considera que es muy importante que la organización no desarrolle una estrategia de Big Data distinta de su estrategia tradicional de datos, ya que en ese caso fallará toda la estrategia del negocio. Big Data y Datos tradicionales son ambas partes de la estrategia global. Para que las organizaciones tengan éxito se necesita desarrollar una estrategia cohesiva donde Big Data no sea un concepto distinto y autónomo.

Es importante insistir en la importancia para las organizaciones de desarrollar una estrategia de Big Data que no sea distinta de la estrategia de datos tradicionales y conseguir una idónea integración de datos. Esta circunstancia es vital ya que ambos forman parte de una estrategia global, aunque Big data irá creciendo de manera exponencial deberá coexistir de modo híbrido con los datos tradicionales durante muchos años. Dicen las grandes consultoras de datos que los Big Data deben ser otra faceta de una buena estrategia de datos de la empresa.

Son numerosos los ejemplos que se pueden dar sobre la integración de datos de todo tipo en estrategias corporativas.

En el caso de la industria eléctrica, los datos de las redes inteligentes (Smart grids) son una herramienta muy poderosa para las compañías de este sector, que conociendo los patrones históricos de facturación de los clientes, sus tipos de vivienda y otros indicadores, unidos con los datos proporcionados por los medidores inteligentes (Smart meterá) instalados en las viviendas pueden conseguir ahorros de coste considerables para la compañía proveedora del servicio eléctrico , y grandes reducciones del consumo eléctrico de los clientes.

Otro caso típico se da en el sector del comercio electrónico donde el análisis de los textos de los correos electrónicos, mensajes de texto SMS o de aplicaciones como WhatsApp , se integran junto con el conocimiento de las especificaciones detalladas del producto que se está examinando; los datos de ventas relativas a esos productos, y una información histórica del producto proporcionan un gran poder al contenido de los textos citados cuando se ponen en un contexto global.

Big Data como oportunidad de negocio en una organización se puede ver a sí mismo como un proceso organizacional dividido en dos etapas diferentes:

1. Etapa de Preparación:

- Fase Inicial: Es imprescindible valorar que la implementación de una solución de Big Data en la organización es una alternativa factible y realista. En caso de que así sea, tendremos que preparar todo lo necesario y conseguir las autorizaciones pertinentes para poder llevarlo a cabo.
    - Detectar necesidades: tienen que ver con el volumen de datos a almacenar, su variedad, velocidad de recogida, procesamiento y escalabilidad horizontal. En este

proceso también se revelan carencias cuando se confronta la nueva tecnología con la existente en la compañía.

  - Justificar la inversión: con el Big Data se pretenden mejorar las cuestiones técnicas, al igual que crear un entorno de alto rendimiento que posibilite el ahorro de costes.
  - Evaluar las limitaciones: se deberá tener en cuenta la infraestructura de la empresa, su madurez tecnológica, sus recursos, pero, sobre todo, los aspectos legales en relación a la privacidad de datos.

- Fase de planificación: en esta fase se determinará el presupuesto con el vamos a contar durante el proceso y los recursos que van a intervenir en el mismo, a saber:
  - Gestores: sponsors, directores de proyecto, coordinadores y gestores de calidad.
  - Diseñadores y arquitectos de datos: perfiles técnicos con los objetivos muy claros en cuanto a la implementación del proyecto.
  - Implementadores: personal cualificado, analistas y desarrolladores, con conocimientos del sector y de tecnología.
  - Operadores de datos: de entrada, intermedios y de resultado.
- Fase de diseño: se parte de un diseño acorde a las necesidades de la organización y se va optimizando teniendo en cuenta el coste, la escalabilidad y las distintas opciones del mercado. Consta de 2 etapas:
  - Infraestructura: son las redes, los equipos o los servidores, es decir, el soporte físico de la solución.
  - Arquitectura: es el apoyo lógico de la solución, formado por los protocolos, las comunicaciones o los procedimientos, entre otros.

- Fase de implementación: en este punto ya se han tenido que dejar cubiertos aspectos como el de la administración, el mantenimiento o la seguridad para poder poner en marcha la solución de Big Data. Los pasos, para ello, son los siguientes:
    - Instalación de servidores y componentes y puesta en marcha de la infraestructura.
    - Configuración de dicha infraestructura para su correcto funcionamiento.
    - Ingesta, transformación y explotación de datos.

2. Etapa de recopilación, análisis de datos y generación de valor:

- Fase de recopilación: los datos son los componentes básicos del Big Data. Durante el proceso se trasformarán en información y, con las técnicas adecuadas, aportarán conocimiento. Aquí ya podremos hablar de datos inteligentes. Dicho proceso comienza con estos pasos:
    - Evaluar los datos: se realiza un estudio de la utilidad de los datos, una evaluación de su volumen y frecuencia de explotación, y una definición de accesos y restricciones para proteger información confidencial.
    - Absorber los datos para su explotación: se preparan y estandarizan los datos.
    - Gestión de datos: en relación a su seguridad, visibilidad, mantenimiento y disponibilidad.
- Fase de análisis: es el núcleo de la solución de Big Data y se concreta en dos acciones:
    - Generar los cálculos y algoritmos que se precisen para implementar la solución.
    - Intervención de especialistas para detectar patrones, tendencias y oportunidades y/o amenazas.

- Fase de agregación de valor: la investigación de los datos por parte de los analistas permite la elaboración de conclusiones y la identificación de nuevas vías de desarrollo del negocio.

Este proceso debe estar integrado con los demás procesos organizacionales de manera sinérgica, pues cómo se dijo anteriormente, Big Data no es un concepto aislado.

2.6 Casos de éxito de Big Data.

El Big Data se ha convertido en una herramienta poderosa para aquellas compañías que desean estar fuertemente posicionadas y ser lideres en el mercado, es por esto que las empresas le han apostado a herramientas y tecnologías que los ayude a anticiparse a un mercado que cada día que pasa se vuelve más y más competitivo. A continuación se presentaran tres casos de éxito del Big Data en organizaciones mundialmente reconocidas.

2.6.1 Reducción de portabilidades de T-Mobile. La gran compañía de telecomunicación celular consiguió reducir a la mitad el numero de portabilidades (de 100.000 el primer semestre de 2011 a 50.000 en el segundo semestre) gracias a la aplicación de técnicas Big Data. Las operadoras de telefonía móvil e internet tienen un número impresionante de datos sobre sus clientes: La cantidad de llamadas que realizan, las horas en las que realizan estas, a quienes llaman con más frecuencia, números preferidos, tiempos que duran sin cobertura, entre muchos otros más. Con estos datos a su disposición y analizando las interacciones de sus clientes en medios sociales, en T-Mobile se propusieron rebajar sustancialmente el número de portabilidades

hacia otros competidores en Estados Unidos. Para ello la empresa utilizó tres herramientas básicas: Sus propios sistemas de cobro (Billing systems), herramientas de monitorización social, además de Spluk y Tableau Software para analizar la información y presentarla de una forma visual.

Combinando toda esta información en T-Mobile descubrieron que las expectativas de portabilidades pueden determinarse a través del análisis de tres factores:

- Facturas,
- Llamadas que se cortan debido a mala cobertura,
- Conversaciones de los clientes: positivas, negativas o neutrales.

Todos estos factores asociados a la influencia o reputación en medios sociales de cada uno de sus clientes, partiendo de la hipótesis de que clientes con un gran número de seguidores o influencia podrán tener un efecto positivo o negativo (según las circunstancias) en otros potenciales clientes de la marca.

La combinación de todos los aspectos mencionados anteriormente llevó a T-Mobile a calcular para cada cliente un "Customer Lifetime Value", un valor monetario individual según las expectativas de negocio y permanencia. Esta información era transmitida en tiempo real a cada agente de la compañía para presentar a los clientes ofertas personalizadas en función de su valor personal.

De esta forma la empresa pasó de casi 100.000 portabilidades en el primer trimestre de 2011 **pasaron a tan sólo 50.000** en el segundo trimestre, una reducción del 50% gracias a un buen aprovechamiento del Big Data y de todos los datos e información que la operadora tiene de sus clientes.

2.6.2 Unilever: de las conversaciones en redes sociales al comportamiento real del los consumidores. Unilever se planteó el reto de conocer si las conversaciones de los potenciales consumidores en medios sociales difieren mucho de su comportamiento real de compra. Para ello se aprovecharon de software proporcionado por Compete (para medir el ROI), Cymfony (*'social media listening'*) y de su propia herramienta de analítica, CybrTrack90, que realiza dos funciones principales: seguir las menciones de sus marcas (o a las categorías de productos en los casos en los que la marca no fuese mencionada expresamente) en medios sociales y también analizar el comportamiento de los usuarios en las búsquedas que éstos realizan en internet.

Lo que los directivos de marketing de Unilever descubrieron es que existen tres actividades diferentes en el proceso de compra: conversaciones, soluciones y compras. Las conversaciones sobre las diversas marcas de comida de Unilever tenían lugar fuera de los comercios y espacios de compra, por lo que la empresa debería encontrar una forma de acercarse a ellos en esos momentos.

Otra conclusión a la que llegaron en la compañía fue que las decisiones de compra de los consumidores están influenciadas por una serie de factores que van más allá del precio o la comunicación, como por ejemplo factores relacionados con la salud, el bienestar o la preparación de las comidas. Por ello desde la empresa se plantearon atacar también estos momentos, con el objetivo de facilitar el consumo a los potenciales compradores y que éstos tuviesen una experiencia más simple y satisfactoria con los alimentos de la marca.

En definitiva, con la combinación de tres herramientas como CybrTrack90, Compete y Cymfony, Unilever fue capaz de comprobar que lo que los consumidores dicen en entornos online difiere, en ocasiones, de la vida real. Y que acercarse a sus consumidores más fieles para

simplificar la compra y el consumo pueden ser claves en el futuro de la marca y en el bienestar de sus clientes.

Unilever combinó los resultados provenientes del sistema de ‘social media listening‘ y los datos de clicks en sus diferentes páginas web asociadas para entender el comportamiento de sus potenciales consumidores y adaptar su oferta y comunicación a dichos hábitos en tiempo real, definiendo estrategias diferentes en sus canales según el día y momento de la semana para que los consumidores encontrasen aquello que se ajusta a sus necesidades.

2.6.3Moneyball: el Big Data aplicado al baseball. Las decisiones en el mundo de los deportes siempre han estado basadas en dos tipos de factores: personales/subjetivos y monetarios. Sin embargo Billy Beane, general manager de los Oakland Athletics de la Major League Baseball estadounidense, decidió poner fin a estas limitaciones.

Billy Beane utilizó una serie de métodos estadísticos propios de los mercados financieros para determinar la valía de sus jugadores y de otros potenciales. El general manager llegó a la conclusión de que en el mundo del baseball se prestaba demasiada atención a una serie de estadísticas y se dejaba de lado otras muchas que tenían un gran valor intrínseco a la hora de seleccionar jugadores, como por ejemplo los porcentajes ‘on-base‘ o ‘slugging‘.

Esta aplicación poco convencional del Big Data en el mundo de los deportes llevó a Billy Beane a sentar cátedra entre sus compatriotas, a crear una nueva escuela de pensamiento en el mundo del deporte y convirtió a Moneyball en un éxito de masas con su adaptación en el cine. Pero los éxitos no se quedaron aquí, ya que el extrovertido directivo de los Oakland A’s llevó a su equipo a competir con otros como los New York Yankees con un presupuesto mucho más

ajustado: $45 millones de dólares frente a $125 millones. Una utilización efectiva del Big Data que muestra que con recursos limitados también pueden obtenerse grandes resultados.[11].

3. Big Data.

3.1 Elementos Primordiales de Big Data

Para llevar a cabo todo este proceso sin fisuras, es vital que una solución de Big Data cuente, al menos, con los siguientes componentes:

3.1.1 Fuentes. Las más habituales son los registros históricos de la compañía, los almacenes de datos, los dispositivos inteligentes, los sistemas de gestión de datos, Internet y el Internet de las Cosas. Para poder determinarlas es necesaria la puesta en común de los conocimientos técnicos, por parte de los desarrolladores, y la perspectiva del negocio, por parte de los analistas.

3.1.2 Capa de almacenamiento. Su función es la de recoger y transformar los datos sin perder de vista la normativa legal. Además, tiene que dar acceso a los datos independientemente de su formato, volumen, frecuencia u origen.

3.1.3 Capa de análisis. Se encarga de leer los datos almacenados. Mediante la utilización de los modelos, los algoritmos y las herramientas adecuadas, proporciona visibilidad sobre los datos para que puedan ser consultados en la capa de consumo.

3.1.3 Capa de consumo. Son muchos los proyectos y usuarios que se benefician del conocimiento extraído en todo este proceso. La forma de consumir los datos dependerá del destinatario, pero será habitual verlos en forma de *reporting* o visualización en tiempo real.

3.2 Arquitectura de referencia (AR) para los sistemas Big Data

A continuación, se muestra una arquitectura de referencia la cual pretende representar de alguna manera el tratamiento y transformación que sufre los datos en su cadena de valor a través de una serie de procesos que pueden ser visto como una descomposición de módulos que

integrados entre sí realizan una labor colaborativa y procedimental con el fin de ayudar en el proceso de toma de decisiones que se llevan en una organización cualquiera[14].

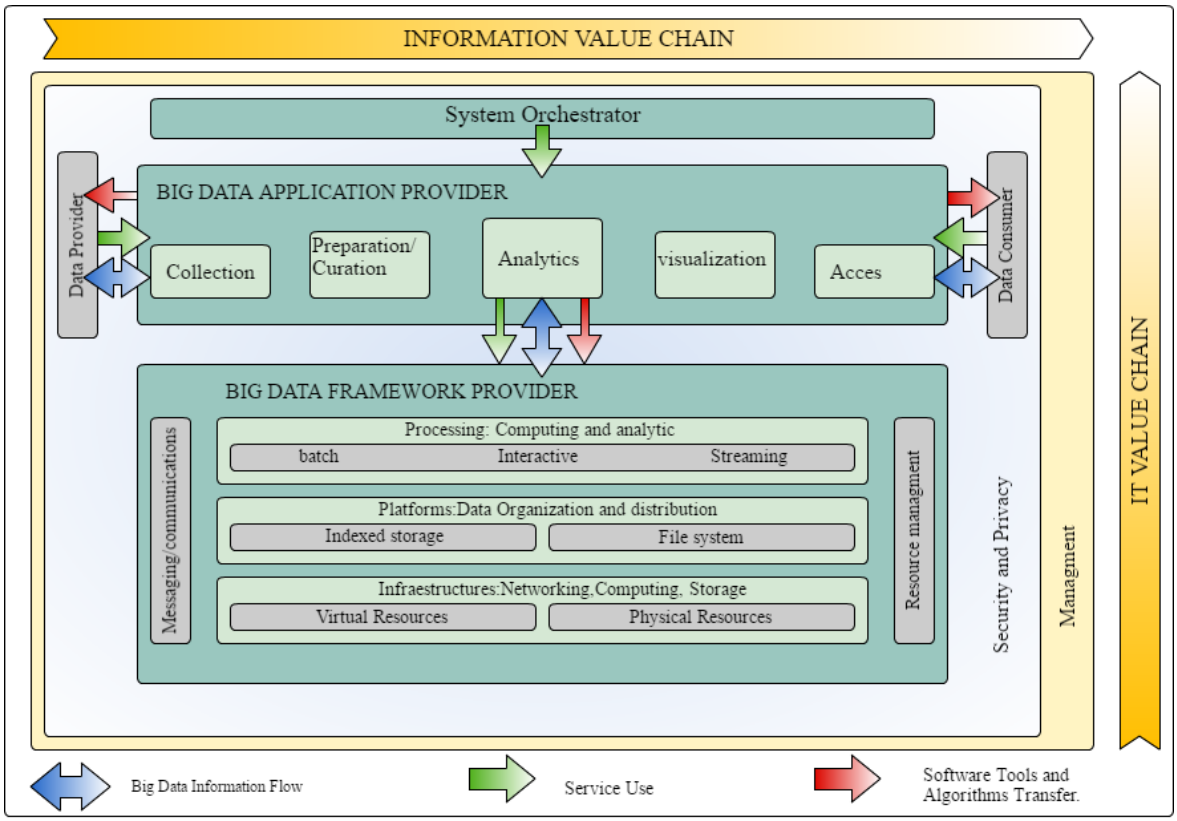

Figura 3. Modelo Arquitectura de Referencia. NIST Big Data Working Group (NBD-PWG) (2013), NIST Big Data Interoperability Framework: Volume 6, Reference Architecture. U.S, 100 Bureau Drive (Mail Stop 8900) Gaithersburg.

Esta arquitectura de referencia se puede organizar en dos ejes determinados por la cadena de valor de Big Data, en donde podemos encontrar en el eje horizontal la información y en el eje vertical las tecnologías involucradas en todo este proceso. Como se ha de notar existen tres componentes primordiales dentro de esta arquitectura: Data Provider (proveedor de datos), Big Data Application Provider (Proveedor de aplicaciones Big Data) y Data Consumer (Consumidor

de datos) , en donde el primero y el último nombrados en cuestón son instancias externas. Sumergiéndonos en lo que respecta a Big Data Application Provider qué es el componente arquitectural encargado de intermediar de forma transparente entre el Data Provider y Data Consumer, se puede observar que a través del eje de la información el valor de los datos que son proveidos en primera instancia son creados por la colección (Data Collection) , integración (Preparation/Curation) , análisis (Analytics) y posteriormente los resultados arrojados por este proceso son aplicados en la visualización (Visualization) y acceso (Acces) de la información para así ser utilizados en la última instancia.

En la figura 2 se puede evidenciar cinco componentes principales que representan diferentes roles técnicos con los que cuentan todos los sistemas Big Data . Estos componentes funcionales son:

3.2.1 System Orchetrator. Este componente se encarga de configurar los modulos que interactúan en el Big Data Application , integrando y definiendo las actividades dentro de una operación vertical del sistema , y esto lo logra con la inclusión de una colección de los roles específicos de los módulos del componente en el cual influye, operando sobre uno o más actores y en donde administra y orquesta (De ahí la definición del nombre del módulo) la operación del sistema Big Data. Estos actores pueden ser componentes humanos, componentes de software o una combinación de ambos.

La función del System Orchestrator es básicamente administrar y configurar los demás componentes de la arquitectura, implementando una o más cargas de trabajo que están diseñadas a ejecutar en la misma. La carga de trabajo administrada por el System Orchestrator puede ser asignando componentes de framework a nodos individuales físicos o virtuales, todo a bajo nivel,

o proveyendo una interfaz de usuario grafica que soporte las especificaciones de un flujo de trabajo enlazando múltiples aplicaciones y componentes a un alto nivel. El system Orchestrator puede también, a través de una estructura administrativa, monitorear la carga de trabajo y el sistema en general en donde se pueda confirmar que los requerimientos específicos de calidad de servicio son conocidos por cada módulo, además de que ocurra una actualización automática en donde se pueda asignar recursos físicos y virtuales adicionales conforme los requerimientos vayan sufriendo cambios.

3.2.2 Data Provider. El Data Provider o proveedor de datos es quien se encarga de introducir nuevos datos e información, alimentando el sistema Big Data. Se puede decir que es el componente que se encarga de suplir de materia prima al sistema, para posteriormente procesar y transformar esta en información valiosa para la organización. Cabe la pena resaltar que los nuevos datos ingresados son diferentes de los que ya están almacenados en repositorios y en uso por el sistema, pero ambos pueden ser accedidos por tecnologías similares. Los actores de involucrados dentro del Data Provider puede ser cualquiera, desde un sensor, hasta una persona que ingresa datos manualmente.

Como se vio anteriormente, una de las características principales de Big Data es la variedad. Los datos importados por el Data Provider pueden ser de innumerables fuentes, tales como imágenes, audios, videos , HTTP Cookies , datos de sensores y muchos otras más. Tanto el Data Provider como el Big Data Application Provider normalmente se asocia a diferentes organizaciones, a menos que la organización implemente su propia plataforma Big Data con sus propias y únicas fuentes de datos. Es por esto que se deben tener diferentes consideraciones de seguridad y privacidad en la arquitectura del sistema Big Data.

Dentro de las funciones que debe cumplir el Data Provider encontramos las siguientes:

- Colección de datos;
- Persistencia de datos;
- Provisión de funciones de transformación para depuración de datos de información sensitiva tales como información personal privada;
- Creación de metadatos describiendo las fuentes de datos, políticas de uso/ derechos de acceso y otros atributos relevantes;
- Verificación de cumplimiento de derechos de acceso sobre el acceso de datos;
- Establecimiento de contratos formales o informales para autorizaciones de acceso de datos;
- Provisión mecanismos de inyección y extracción de datos; y
- Publicación de disponibilidad de información y recursos de acceso.

3.2.3 Big Data Application Provider. El Big Data Application Provider reúne un conjunto especifico de operaciones a lo largo del ciclo de vida de los datos teniendo en cuenta los requerimientos establecidos por el System Orchestrator y también los requerimientos y políticas de seguridad y privacidad. Este componente funcional se encarga de encapsular toda la lógica de negocio y funcionalidad a ser ejecutadas por la arquitectura. Las funciones (separadas en subcomponentes funcionales como se muestra en la Figura 2.) que se asocian a este elemento arquitectural son los siguientes:

- Colección.
- Preparación.

- Análisis.
- Visualización.
- Acceso.

Los datos propagados a través de toda la arquitectura son transformados en diferentes maneras en orden de extraer valor de la información. Cada función del Biga Data Application Provider puede ser implementada por entidades independientes y presentado todo como un servicio autónomo.

3.2.3.1 Colección. El subcomponente de colección se encarga de la conexión e interacción de con el Data Provider. Esto se puede ver como un servicio general , tal como un servidor de archivos o un servidor web, configurado por el System Orchestrator para aceptar y tratar datos específicos, o puede ser un servicio de aplicación especifico diseñado para extraer o recibir datos del Data Provider. Desde que este módulo recibe los datos , estos deben ser almacenados en un buffer (establecido en el diseño de la arquitectura) hasta que los datos sean almacenados por el Big Data Framework Provider. En una etapa inicial de la colección, conjuntos de datos de estructura similar son recolectados , estableciendo una uniforme seguridad, políticas y otras consideraciones.

3.2.3.2 Preparación. Este modulo funcional es donde empieza la transformacón de los datos adquiridos, aunque se observará que en el análisis en donde se produce una transformación más avanzada. Aquí se produce los procesos de limpieza, validación, estandarización , formateo y/o encapsulamiento de datos. En la verificación se puede incluir un trabajo de optimización

mediante manipulaciones e indexación que ayude el proceso de análisis. Es importante anotar que aquí se puede agregar datos de diferentes Data Provider, acoplando mediante claves principales (es una opción) los diferentes metadatos para crear un expandido y mejorado conjunto de datos.

3.2.3.3 Análisis. El análisis involucrado en el Big Data Application Provider incluye la codificación de la lógica de negocia en un bajo nivel del sistema Big Data (El System Orchestrator se encarga del proceso a un alto nivel).Aquí se implementa las técnicas de extracción del valor de los datos basados en los requerimientos de la aplicación vertical donde se ve envuelto el Big Data Framework Provider. Los requerimientos especifican los algoritmos de procesamiento de datos para producir los resultados que se requieren. Tipicamente en este módulo del sistema se cuenta con software que implementa el análisis lógico como procesamiento batch(lotes) o streaming(flujo) que puede variar dependiendo de ciertos factores determinados, también se puede observar el componente messaging/communication el cual se encarga de realizar el intercambio de información a través de todo el framework de procesamiento mediante funciones de control o paso de datos.

3.2.3.4 Visualización. El módulo de visualización básicamente se encarga de presentar los datos procesados y los resultados del proceso de análisis a el Data Consumer en un formato que informe y comunique un significado y aprendizaje a partir de los resultados arrojados, es importante que se cuente con una interfaz que facilite la interpretación humana. Algunas técnicas de visualización pueden producir documentos estáticos, información en caché para un acceso posterior (ejemplos. Reportes o gráficos), sin embargo otra técnicas de a menudo incluyen una

generación de una interfaz interactiva que ayudan al proceso de visualización mediante filtros de búsqueda u otras herramientas útiles y de fácil interacción.

3.2.3.5 Acceso. El acceso dentro de el Big Data Provider es el módulo que permite la comunicación con el Big Data Consumer. Este subcomponente podría interactuar con los módulos de visualización y análisis permitiendo al consumidor obtener resultados de la información solicitada. Además, el módulo de acceso permite confirmar que los metadatos descritos y administrados, y los esquemas de los metadatos son capturados y mantenidos por el Data Consumer , así como lo s datos son transferidos al mismo.

3.2.4 Big Data Framework Provider. El Big Data Framework provider consiste básicamente en un o más instancias de componentes organizados jerárquicamente en el eje vertical de la arquitectura de referencia . Muchas de las implementaciones desarrolladas en este componente son el resultado de la combinación de múltiples tecnologías enfocadas en proporcionar flexibilidad o reunir un completo rango de requerimientos, los cuales son impulsados desde el Big Data Application Provider.

El Big Data Framework Provider contiene los siguientes tres subcomponentes:

- Frameworks de procesamiento.
- Frameworks de plataforma de datos.
- Frameworks de infraestructura.

3.2.4.1 Frameworks de procesamiento. Los frameworks de procesamiento proveen toda la infraestructura lógica y de software necesaria que soporta la implementación de aplicaciones que tratan con el volumen, la velocidad, la variedad y demás características de los datos en un contexto Big Data.

El framework de procesamiento generalmente se enfoca en la manipulación de los datos, donde se pueden clasificar entre un procesamiento batch (por lotes) y streamig(en tiempo real).Sin embargo, también existe procesamiento near real time (cercano a tiempo real).

Las líneas que dividen y establecen si un framework pertenece a un procesamiento por lotes o tiempo real no son muy sólidas o distinguidas. En general, muchas de las arquitecturas Big Data incluirán frameworks que soporten un amplio rango de requerimientos. A continuación de se describe la clasificación de los frameworks de procesamiento.

3.2.4.1.1 Batch Frameworks. En esta clasificación de frameworks de procesamiento van aquellas herramientas y tecnologías que soportan el procesamiento por lotes, el cual los datos ya almacenados son procesados (Sólo los datos que ya están almacenados). El procesamiento por lotes ha sido una herramienta muy utilizada desde que ha aparecido todo este estallido de big data gracias a su eficiencia, en donde su objetivo se basa en acumular enormes cantidades de datos, procesarlos y producir resultados por lotes. Este procesamiento obliga al uso de diferentes tecnologías para la gestión de entrada de los datos, su procesamiento y la gestión de las salidas y resultados generados. Existen herramientas que ayudan la realización de dicho proceso, una de estas es Hadoop que ha permitido almacenar grandes cantidades de datos y escalarlos horizontalmente mediante la adición de más nodos en la medida que así sea necesario; alrededor

de esta poderosa herramienta han surgido muchas otras como HBase, Hive, Pig, entre otras, que lo que hacen es ayudar en el proceso de análisis de los datos a estudiar.

A continuación, se presenta un gráfico en donde se muestra el comportamiento de una arquitectura de procesamiento por lotes:

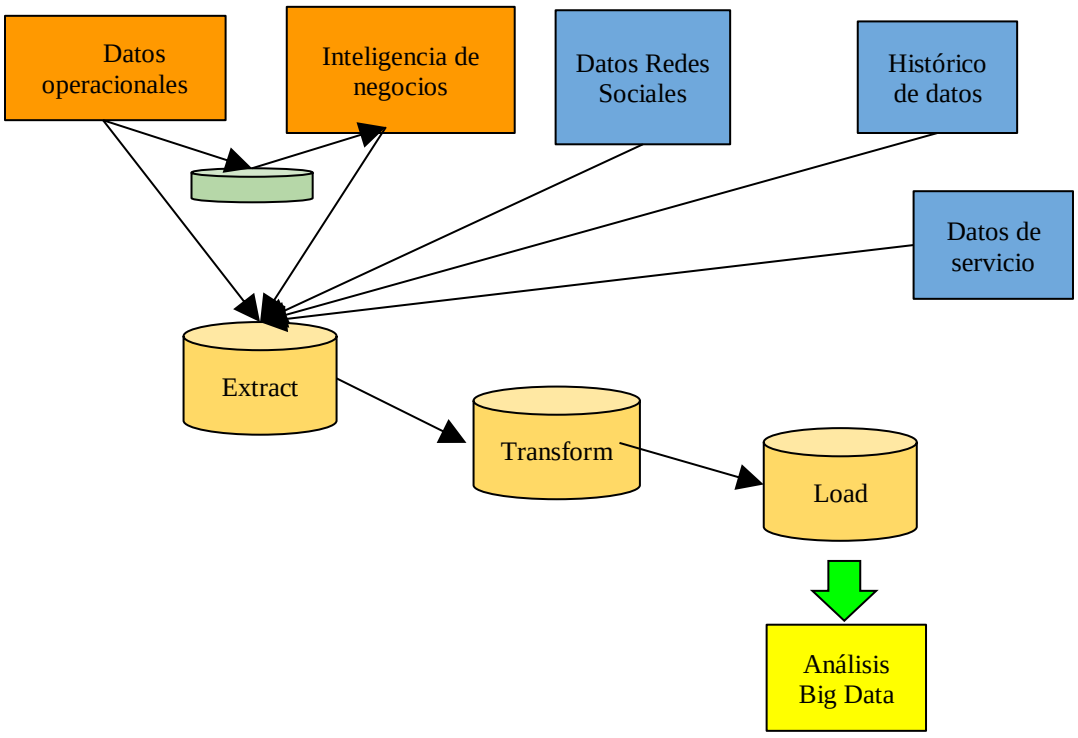

Figura 4. Arquitectura de procesamiento por lotes.

3.2.4.1.2 Streaming frameworks. Este tipo de técnicas de procesamiento y análisis de dato se basan en la implementación de un modelo de datos en el que los datos asociados a series de tiempos (hechos) fluyen continuamente a través de una red de entidades de transformación que componen el sistema.

No hay limitaciones de tiempo obligatorias en este tipo de procesamiento, al contrario, con las técnicas de procesamiento y análisis de datos en tiempo real. Por ejemplo. Un sistema que se ocupe del recuento de las palabras incluidas encada tweet para el 99.9% de los tweets procesados es un sistema de procesamiento en stream válido. Tampoco existe una obligación en cuanto al plazo de tiempo de generación del output recibido del sistema. Las únicas limitaciones son:

- Se debe disponer de suficiente memoria para almacenar entradas en cola.
- La tasa de productividad del sistema a largo plazo debería ser más rápida, o por lo menos igual, a la tasa de entrada de datos en ese mismo período. Si esto no fuese así, los requisitos de almacenamiento del sistema crecerían sin límite.

Este tipo de técnicas de procesamiento y análisis de dato s no está destinado a analizar un conjunto completo de grandes datos, por lo que generalmente no presentan esa capacidad, salvo excepciones.

3.2.4.2 Messaging/communications frameworks. Este tipo de frameworks tienen sus raíces en los ambientes de computación de alto rendimiento (High Performance Computing) en las comunidades científicas y de investigación. Los frameworks de comunicación y mensajería fueron desarrollados para proveer APIs para un seguro encolamiento, transmisión y recepción entre los nodos en un cluster escalado horizontalmente. Este módulo típicamente está implementado sobre modelos punto a punto donde los datos son transferidos desde un transmisor a un receptor, aunque también hay variaciones como son los modelos multicast (distribución

uno- muchos o muchos-muchos), que permiten a los transmisores enviar datos a todos los receptores de interés.

3.2.4.3 Resource management framework. Con el paso del tiempo los sistemas Big Data se han vuelto más complejos lo cual ha llevado a la utilización de cada vez muchos más recursos computacionales para cubrir todos los requerimientos demandados. Este módulo se encarga de gestionar y administrar todos los recursos y herramientas necesarias para apoyar todo el proceso de extracción, procesamiento y análisis con conlleva los ambientes Big Data.

3.2.5 Data consumer. Al igual que el Data Provider, el rol del Data consumer en esta arquitectura de referencia puede sugerir un usuario final u otro sistema. Las actividades asociadas a este módulo son las siguientes:

- Búsqueda y recuperación de información,
- Descarga,
- Analizar localmente,
- Reporte,
- Visualización.

Data Consumer usa las interfaces o servicios provistos por el Big Data Application Provider para obtener acceso a la información de interés. Estas interfaces pueden incluir reporte, recuperación y representación de datos.[10].

3.3 Técnicas de Análisis y Procesamiento Big Data

El análisis de la información como parte fundamental de una organización permite a la misma tomar decisiones críticas que le ayude a tener una fuerte posición en el mercado[18].

A continuación se presentan algunas técnicas de análisis utilizadas en el entorno Big Data:

3.3.1 Test A/B. Es una técnica en la que se compara un grupo de control con una variedad de grupos de test para determinar qué cambios o tratamientos producirán una mejora dada una variable objetiva (por ejemplo, un ratio de respuesta de una acción de marketing). Un ejemplo de este experimento de testing A/B (también llamado split testing o bucket testing), es determinar qué texto, maquetación, imágenes o colores producen una mejora en los ratios de conversión de una tienda online o una acción de marketing por email. Big data permite ejecutar y analizar una gran cantidad de pruebas, siempre asegurando que los grupos son de un tamaño suficiente para detectar diferencias estadísticamente significativas entre el grupo de control y los grupos de pruebas. Cuando se manipila más de una variable en el experimento simultáneamente, la generalización multivariante de esta técnica, que se aplica a modelos estadísticos, se le llama A/B/N testing [14].

3.3.2 Reglas de asociación. El objetivo de las reglas de asociación es encontrar asociaciones o correlaciones entre los elementos u objetos de bases de datos transaccionales, relacionales o datawarehouses. Las reglas de asociación tienen diversas aplicaciones como:

- Soporte para la toma de decisiones
- Diagnóstico y predicción de alarmas en telecomunicaciones

- Análisis de información de ventas
    - Diseño de catálogos
    - Distribución de mercancías en tiendas.
    - Segmentación de clientes en base a patrones de compra

Las reglas de asociación son parecidas a las reglas de clasificación. Se encuentran también usando un procedimiento de covering. Sin embargo, en el lado derecho de las reglas, puede aparecer cualquier par o pares atributo-valor.

Para encontrar ese tipo de reglas se debe de considerar cada posible combinación de pares atributo-valor del lado derecho.

Para posteriormente podarlas usando cobertura (número de instancias predichas correctamente) y precisión (proporción de número de instancias a las cuales aplica la regla).

Ejemplo: Encontrar las reglas de asociación X=>Z de la tabla 1 con la restricción de cumplir con un mínimo de cobertura y de precisión.

Las reglas con:

- Cobertura mínima de 50%
- Precisión mínima de 50%
    - A=>C (50%, 66.6%)
    - C=>A (50%, 100%).

Una regla de asociación es una expresión de la forma X=>Z donde X y Z son conjuntos de elementos.

El significado intuitivo:

Las transacciones de la base de datos que contienen X tienden a contener Z [15].

Transacción	Elementos Comprados.
1	A,B,C
2	A,C
3	A,D
4	B,E,F

Tabla 1. Reglas de asociación.

3.3.3 Fusión e integración de datos. Son una serie de técnicas que permiten integrar y analizar datos de múltiples fuentes con el objeto de realizar descubrimientos entre la información de manera más eficiente y potencialmente más precisa que si fueran analizados utilizando una sola fuente de datos. Un ejemplo práctico sería la aplicación combinada de diversos sensores de datos de dispositivos conectados en la llamada Internet de las cosas, integrado con el rendimiento de sistemas complejos distribuidos en una explotación petrolífera. Otro ejemplo sería el análisis vía procesamiento de lenguaje natural de datos de redes sociales combinados con datos de ventas en tiempo real, con el objetivo de determinar el efecto que está teniendo una campaña de marketing en el sentimiento de los clientes y su comportamiento reflejado en las decisiones de compra.

3.3.4 data Mining . Consiste en extraer patrones de grandes datasets mediante la combinación de métodos estadísticos y de aprendizaje automático con la gestión de las bases de datos. Entre las técnicas de datamining se incluyen técnicas de aprendizaje de reglas de asociación, análisis de agrupamiento, clasificación y regresión. Como ejemplos de aplicaciones

prácticas estarían la minería de datos de clientes para determinar qué segmentos son más proclives a responder a una oferta, minar datos de recursos humanos para identificar características de los empleados de más éxito, o el análisis de cestas de compras para modelar el comportamiento de compras de los clientes.

3.3.5Algoritmos Genéticos . Es una técnica utilizada para optimizar datos inspirada en el proceso de la evolución natural o supervivencia de los mejor adaptados. Con esta técnica las soluciones posibles son codificadas como si fueran cromosomas que pueden combinarse y mutar. Estos cromosomas son seleccionados y separados para sobrevivir dentro de un ecosistema modelado que determina la adaptabilidad o el rendimiento de cada uno dentro del conjunto. Estos algoritmos evolutivos funcionan bien para solucionar problemas no lineales, como por ejemplo, mejorar la planificación de tareas en la industria manufacturera, o la optimización del rendimiento de una cartera de inversión.

3.3.6 Aprendizaje automático. Una especialidad dentro de la ciencia computacional también conocida como inteligencia artificial, que se ocupa del diseño y desarrollo de algoritmos por los cuales se permite a los ordenadores pueden hacer evolucionar comportamientos basados en datos empíricos. Uno de los objetivos principales de esta técnica es aprender de forma autónoma a reconocer patrones complejos y tomar las decisiones basándose en los datos. Un ejemplo sería el procesamiento de lenguaje natural, como Siri o Google Now, que ya llevan los smartphones.

3.3.7 Redes Neuronales. Los modelos computacionales, inspirados por los trabajos de redes neuronales biológicas, como las conexiones de las células del cerebro, que buscan patrones entre

datos. Las redes neuronales son apropiadas para buscar patrones no lineales y optimización. Entre las aplicaciones prácticas de esta técnica, por ejemplo, la identificación de los clientes de alto valor que están en riesgo de cambiar de proveedor, o la identificación de partes de seguro fraudulentos.

3.3.8 Analisis de Redes . Son técnicas empleadas para caracterizar relaciones entre nodos separados en un gráfico o red. Al analizar las conexiones entre individuos de una comunidad en las redes sociales podemos extraer cómo fluye la información o quién ejerce la mayor influencia y sobre quiénes. Entre las aplicaciones prácticas están la identificación de los líderes de opinión para realizar una acción de marketing precisa, o identificar los cuellos de botella en los flujos de información de las compañías.

3.3.9 Análisis de sentimientos. Consiste en la aplicación de técnicas como la de procesamiento de lenguaje natural así como otras técnicas analíticas para identificar y extraer información subjetiva de las fuentes. Pueden identificar el sentimiento hacia una marca, producto o característica por su tipo, como la polaridad (pudiendo ser un sentimiento positivo, negativo o neutral), así como el grado y fuerza del sentimiento. En big data esta técnica se utiliza sobre todo en blogs, microblogs, y redes sociales para determinar cómo los segmentos y el mercado reaccionan ante acciones previstas e imprevistas.

3.3.10 Análisis espacial . Son una serie de técnicas, sobre todo estadísticas, que permiten analizar las propiedades topológicas, geométricas o geográficas codificadas dentro de un conjunto de datos. A menudo estos datos de ubicación son capturados gracias a un GIS (sistemas

de información geográfica) que registran, por ejemplo, coordenadas de longitud y latitud. Incorporando datos espaciales en regresiones espaciales podemos averiguar la correlación entre consumidores que desean adquirir un producto y su localización. También se emplean en simulaciones, por ejemplo, una empresa que desee expandirse puede averiguar cómo respondería la red de una cadena de suministro según donde estuviera ubicada.

3.3.11 Simulación. El modelado del comportamiento de sistemas complejos se utiliza para previsión, predicciones y planificación de escenarios futuros. El método Monte Carlo, por ejemplo, consiste en una serie de algoritmos basados en la repetición de sampleados aleatorios, permitiendo ejecutar miles de simulaciones, cada una con supuestos diferentes. Se obtiene así una muestra del histograma con la distribución probabilística de los resultados. Se aplica mucho en el sector financiero, por ejemplo, para realizar una evaluación de las opciones de llegar a objetivos de resultados dada la incertidumbre sobre el éxito de las iniciativas aprobadas.

3.4 Herramientas para Big Data

Analizar y procesar grandes cantidades de datos no es tarea fácil, por eso a medida que ha aparecido toda esta explosión de Big Data también se ha desarrollado múltiples herramientas que sin dudad alguna facilitan todo el proceso por el cual se ve sometido los datos. Debido a la variedad de problemas a los que una organización puede enfrentarse en sus ánimos por mejorar y competir en todo este ámbito Big Data para lograr sus objetivos, las herramientas existentes ofrecen soluciones y posibilidades que dependen de la problemática a la que se vea obligado tratar. La oferta de proveedores está respondiendo proporcionando arquitecturas altamente distribuidas y escalables, integrando también soluciones en cuanto a memoria y robustez de

procesamiento. Cabe la pena resaltar que también el modelo de código abierto es cada vez más aceptado por los profesionales en gestión de datos. A continuación se presentan unas de las herramientas más aceptadas en los entornos Big Data.

3.4.1 Hadoop. Es un proyecto de software abierto bajo licencia Apache,su creador es Doug Cutting el cual inició el desarrollo de esta infraestructura digital cuando trabajaba en Yahoo! Inspirándose en tecnologías liberadas por Google, concretamente MapReduce Y Google File System (GFS), con el fin de utilizarla como base para un motor de búsqueda distribuido [20].

Hadoop permite el procesamiento de grandes volúmenes de datos bajo clusters de servidores básicos. Está diseñado para convertir un sistema de servidor único de servidor a miles de máquinas, lo que lo hace tener un comportamiento altamente escalable y con un muy alto grado de tolerancia a fallas. En lugar de depender del hardware de alta gama, la fortaleza de estos clusters depende de la capacidad de detectar y manejar fallas a nivel de aplicación del software. Apache Hadoop es un sistema distribuido que utiliza una arquitectura Master-Slave, también hace uso de su propio sistema de almacenamiento de datos Hadoop Distributed File System (HDFS) y ejecuta algoritmos de MapReduce para realizar cálculos.

3.4.1.1 Arquitectura principal de Hadoop.

3.4.1.1.1 Hadoop Distributed File System. El sistema de archivos de Hadoop se ha desarrollado usando diseño de sistema de archivos distribuidos. Se ejecuta en hardware de

productos básicos. A diferencia de otros sistemas distribuidos, HDFS es muy tolerante y diseñado utilizando hardware de bajo coste[21].

HDFS proporciona un acceso fácil en medio de inmensas cantidades de datos. Para almacenar estos grandes volúmenes de información, este sistema de archivos los almacena en varias máquinas. Vale la pena resaltar que los datos se almacenan de manera redundante para evitar las pérdidas de datos por posibles fallos.

El sistema de archivos distribuidos de Hadoop sigue los lineamientos de la arquitectura maestro-esclavo, en donde presenta un NameNode que hace de Maestro y un conjunto de DataNode que hacen de esclavos.HDFS está diseñado para almacenar grandes cantidades de datos a través de un gran número de máquinas, implementando también un sistema de replicación de datos para hacer frente a un mal funcionamiento o daño de equipos del cluster [16].

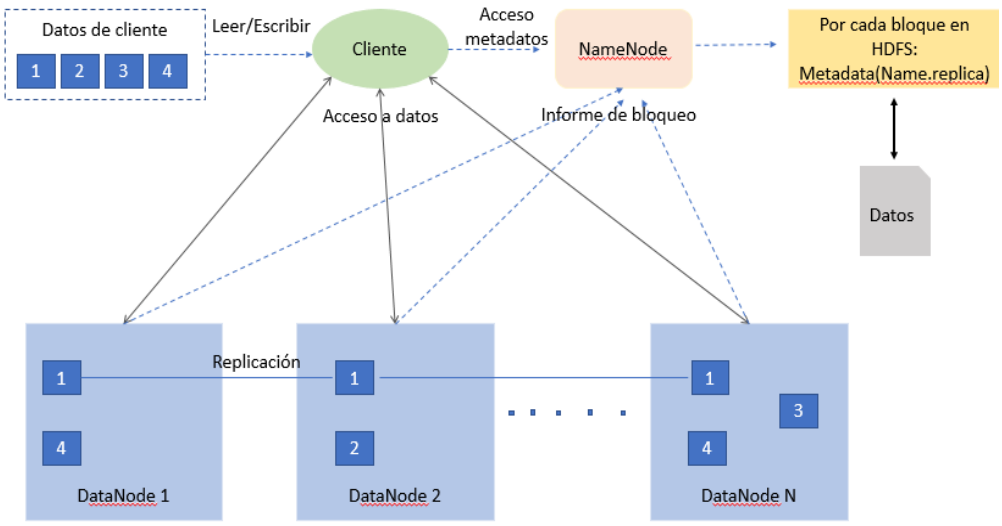

Figura 5. Representación HDFS. Hvivanir, (2014).Aquitectura HDFS. Recuperado desde: .https://hvivani.com.ar/2014/07/19/arquitectura-hdfs/

En primera instancia los archivos de entrada se dividen en bloques de tamaño fijo, estos son almacenados de manera distribuida en el cluster Hadoop. Un archivo puede estar dividido en varios bloques los cuales se almacenan en diferentes DataNode ó nodos del cluster, esto trae como resultado que cuando se quiera acceder a un archivo, se debe consultar a múltiples nodos, esto quiere decir que el HDFS soporta archivos de tamaños más grandes que la capacidad de disco de una sola máquina. El NameNode, es el encargado de almacenar todos los metadatos, gestionando el espacio de nombres del sistema de archivos. HDFS esta implementado bajo el paradigma arquitectural Maestro /Esclavo, estableciendo el NameNode como el maestro y los DataNodes como los esclavos.

Cabe la pena resaltar que la existencia de un solo NameNodo aunque puede simplificar la arquitectura del sistema, también puede ser un aspecto vulnerable, pues si este falla, todo el resto también lo hará, esto se conoce como punto único de falla (Single Point of Failure).

Debido a que el DataNode almacena relativamente baja cantidad de datos, pues solo maneja los metadatos (nombres de los archivos, permisos y ubicación) , los cuales son almacenados en la memoria principal, lo que permite el rápido acceso. HDFS maneja un proceso de replicación , esto quiere decir que los datos que un archivo está almacenado en varios DataNode en caso de que algún nodo del cluster presente fallas. Todos los DataNode del sistema envían periódicamente (normalmente cada 10 minutos) una señal al NameNode , cuando este no recibe ningún informe de bloqueo (como se conoce) el nodo es referenciado como muerto y se deja de enviar peticiones de entrada y salida al mismo.

3.4.1.1.2 MapReduce. Se puede referenciar como el corazón de Hadoop[17], es un paradigma de programación que proporciona un sistema de procesamiento paralelo y distribuido a través de cientos de servidores en un cluster Hadoop. MapReduce fue diseñado para resolver problemas que pueden ser paralelizados haciendo uso de dos funciones Map y Reduce (a lo que se debe el nombre). Es de resaltar que no todos los problemas pueden ser resueltos eficientemente mediante este paradigma, ya que MapReduce se pensó para trabajar sobre conjunto de datos de gran tamaño, por lo que hace uso del Hadoop Distributed File System[22].

MapReduce cuenta es está concebido bajo la arquitectura Maestro/esclavo; cuenta con un servidor maestro denominado jobTracker y varios servidores esclavos llamados taskTracker, uno por cada nodo del clúster. Básicamente el jobTracker es el encargado de gesionar la interacción del usuario con el framework. El usuario envía trabajos mapReduce al jobTracker y este se encarga de asignar las peticiones en orden de llegada a los esclavos, esto quiere decir que los taskTracker ejecutan tareas bajo la orden del nodo Maestro y manejan el movimiento de datos entre la fase Map y Reduce[23].

Las funciones map y reduce están definidas ambas con respecto a datos estructurados en tuplas del tipo (clave,valor).

3.4.1.1.2.1 Función Map. La función map toma una tupla del tipo (clave,valor) con un tipo en un dominio de datos y devuelve una lista de pares en un dominio diferente:

Map(k_1, V_1) → List (k_2,V_2)

Esta función se encarga del mapeo y es aplicada en paralelo para cada ítem en la entrada de datos, por lo que se obtendrá una lista de pares por cada llamada a la función Map, después de esto se agrupan todos los pares que posean las misma clave de todas las listas, creando un grupo de por cada una de las diferentes claves generadas. Desde un punto de vista de arquitectura el nodo maestro toma un entrada de datos, lo divide en varios segmentos o problemas menores y los distribuye sobre los nodos esclavos, los cuales también pueden subdividir un segmento si así se determina, cada esclavo se encarga de procesar el problema y transmitir la respuesta al nodo maestro.El conjunto de archivos de entrada se divide en varias tareas lladas fileSplit, el tamaño fijo de bloque por lo general es de 128MB.

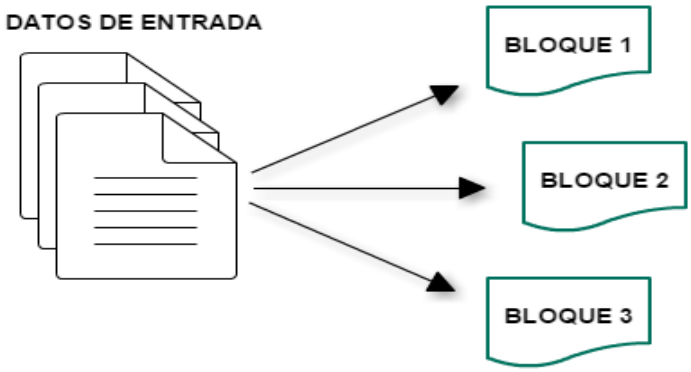

Figura 6. Operación Split.

3.4.1.1.2.2 Función Reduce. La función reduce se aplica en paralelo para cada grupo creado por la función map, produciendo una colección de valores para cada dominio:

$$\text{Reduce } (k_2, \text{list}(V_2)) \rightarrow \text{list}(V_3)$$

Cada llamada a Reduce produce un valor V_3 o una llamada vacía, aunque llamada puede retornar más de un valor. El retorno de todas esas llamadas se recoge como la lista de resultado deseado produciendo un conjunto más pequeño de los valores.

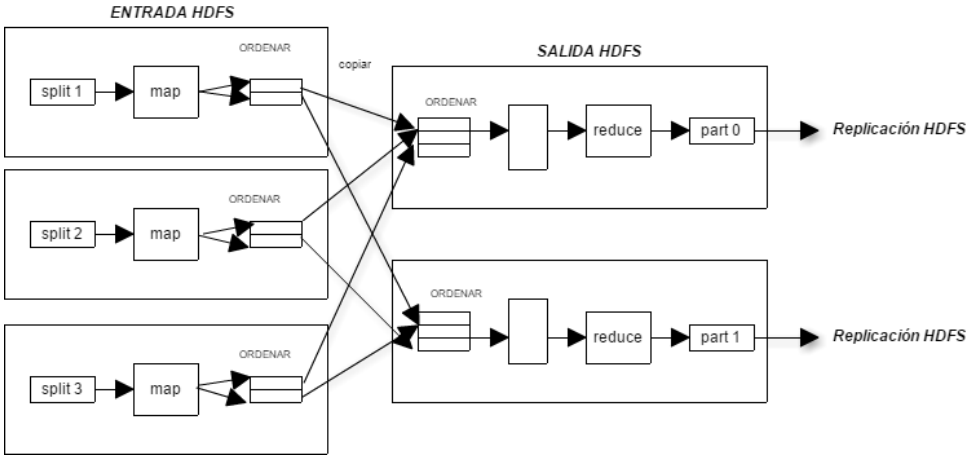

Figura 7. Esquema General MapReduce. What is MapReduce?. (s.f). Recuperado desde : https://www-01.ibm.com/software/data-infosphere/hadoop/mapreduce/

Cuando se inicia la tarea reduce, la entrada se encuentra dispersa en varios archivos a través de los nodos en las tareas de Map. Los datos obtenidos de la fase Map se ordenan para que los pares clave-valor sean contiguos (fase de ordenación,), esto hace que la operación Reduce se simplifique ya que el archivo se lee secuencialmente. Si se ejecuta el modo distribuido estos necesitan ser primero copiados al filesystem local en la fase de copia. Una vez que todos los datos están disponibles a nivel local se adjuntan a una fase de adición, el archivo se fusiona de forma ordenado. Al final, la salida consistirá en un archivo de salida por tarea reduce ejecutada.

Por lo tanto, N archivos de entrada generará M mapas de tareas para ser ejecutados y cada mapa de tareas generará tantos archivos de salida como tareas Reduce hayan configuradas en el sistema.

3.4.1.2 Ecosistema Hadoop. Hadoop cuenta con un ecosistema muy diverso, el cual crece cada día más, por lo que es complicado conocer todos los proyectos que hacen interacción don Hadoop. A continuación se muestra los más comunes [20]:

3.4.1.2.1 Chukwa. Es un sub-proyecto dedicado a la carga masiva de varios ficheros texto dentro de un clúster Hadoop (ETL). Chukwa se construye bajo el sistema de archivos distribuido (HDFS) y el marco MapReduce y hereda la escalabilidad y robustez de Hadoop. También incluye un conjunto de herramientas flexible y potente para la visualización y análisis de los resultados.

3.4.1.2.2 Apache Flume. Igualmente hace parte de Hadoop y surge para subir datos de aplicaciones al sistema de archivos de Hadoop (HDFS). Su Arquitectura se basa en flujos de streaming de datos, ofrece mecanismos para asegurar la entrega y mecanismos de recuperación.[24]

3.4.1.2.3 Hive. Apache Hive es una infraestructura de almacenamiento de datos construida sobre Hadoop para proporcionar agrupación, consulta, y análisis de datos.1 Inicialmente desarrollado por Facebook, Apache Hive es ahora utilizada y desarrollado por otras empresas como Netflix y la Financial Industry Regulatory Authority (FINRA). Amazon mantiene una derivación de software de Apache Hive incluida en Amazon Elastic MapReduce en sus servicios Amazon Web Services [3].

3.4.1.2.4 Apache HBase. Es un sistema de bases de datos orientado a columnas que se ejecuta en HDFS y a diferencia de los sistemas de bases de datos relacionales, HBase no soporta un lenguaje de consulta estructurado como SQL. Cada tabla contiene filas y columnas como una base de datos relacional y cada tabla tiene definida su clave principal de acceso. HBase permite que muchos atributos sean agrupados llamándolos familias de columnas, de tal manera que los elementos de una familia de columnas son almacenados en un solo conjunto [3]

3.4.1.2.5 Apache Avro. Es un proyecto de Apache que provee servicios de serialización. Cuando se guardan datos en un archivo, el esquema que define ese archivo es guardado dentro del mismo; de este modo es más sencillo para cualquier aplicación leerlo posteriormente puesto que el esquema está definido dentro del archivo [3].

4. Computación Urbana.

4.1 Definición de computación Urbana

La computación Urbana es un proceso de adquisición, integración y análisis de gran cantidad de datos heterogéneos generado por diversas fuentes en espacio urbanos como sensores, dispositivos, vehículos, edificios, y los mismos humanos, con el objetivo de abordar las problemáticas y situaciones que afronta una ciudad como lo puede ser la polución del air, congestión de tráfico vehicular o el incremento del consumo energético. La computación urbana conecta tecnologías de sensores ubicuos y discretos, administración avanzada de datos y modelos de análisis de los mismos, y métodos de visualización para crear soluciones que mejoren el medio ambiente, la calidad de vida humana y los sistemas de operaciones enla ciudades. La computación urbana también ayuda a entender el fenómeno urbano e incluso a predecir el futuro de las ciudades. La computación urbana es un campo interdisciplinario que fusiona la ciencia computacional con áreas tradicionales como el transporte, ingeniería civil, economía, ecología y sociología en el contexto de espacios urbanos [25].

4.2 Framework de la Computación Urbana.

4.2.1 Framework General. Un Framework general para la computación urbana es una arquitectura compuesta de los siguientes componentes:

- Detección de datos urbanos.
- Gestión de datos urbanos.

- Análisis de datos urbanos.
- Oferta de servicio.

A diferencia de los sistemas convencionales en el entorno web, que suelen estar basados en arquitecturas de datos de tipo único y unas pocas tareas, en la computación urbana se tiene múltiples datos y muchas tareas. Muchas veces es necesario hacer fusiones de datos diversos para llevar a cabo una sola tarea.

La Figura 8 representa un framework general para la computación urbana en donde se comprenden cuatro capas determinadas por los componentes anteriormente mencionados: Detección de datos urbanos, Gestión de datos urbanos, Análisis de datos urbanos y oferta de servicio.

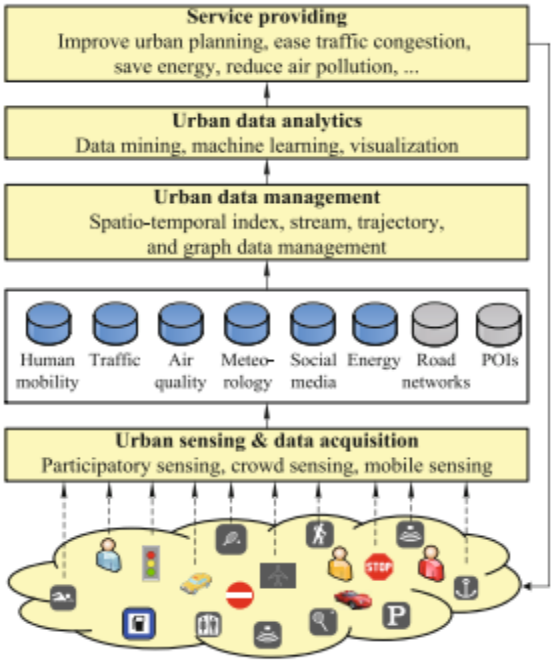

Figura 8. Framework general de la computación urbana. Zheng Y. (2017) , Urban computing: enabling urban intelligence with big data,Beijing,China, 1-3 , doi:10.1007/s11704-016-6907-2

A continuación se presenta y se intenta dar una breve reseña de cada una de las capas implicadas en el framework general propuesto[25].

4.2.1.1 Detección de datos urbanos. La detección de datos urbanos se encarga de recolectar los mismos de diversas fuentes a través de sensores o de los mismos humanos en una ciudad. Existen dos principales modos de detección de datos, detección mediante sensores y detección a través de los humanos actuando como sensores. La primera forma de detección hace referencia al despliegue de sensores en sitios fijos, esto con el fin de detectar el comportamiento de cierto fenómeno en un lugar específico; ejemplos de esto son los sensores ubicados en las estaciones meteorológicas o un sensor fijado en objetos en movimiento como un bus o un taxi. Estos sensores continuamente envían y reciben información desde y hacia un sistema Backend sin necesidad de involucrar a personan en el circuito. El segundo modo de detección es el establecimiento de los humanos como sensor, el cual saca provecho de la exploración de la dinámica urbana cuando los mismos se mueven e interactúan en los entornos urbanos. La información recolectada de las individualidades es luego usada para resolver problemas colectivos; ejemplos de esto son las personas comentando en redes sociales, manejando o caminando y enviando información posicional.

4.2.1.2 Gestión de datos urbanos. La capa de gestión de datos urbanos administra los datos urbanos a gran escala y de manera dinámica, donde usualmente asocia la coordinación espacial con una marca de tiempo en la cual fue generada la información (propiedades espacio-

temporales), usando normalmente plataformas de computación en la nube, indexación de estructuras, y algoritmos de recuperación. Actualmente las plataformas de computación en la nube no soportan muy bien los datos con propiedades espacio-temporal por tres principales razones: la primera es por la estructura de los datos espacio-temporales como por ejemplo los datos de trayectoria, en segundo lugar por la dificultades que se presentan para realizar consultas sobre los datos con propiedades espacio-temporales, y la tercer razón es por que en una aplicación de computación urbana se necesita consumir simultáneamente múltiples conjunto de datos de diferentes dominios.

La capa de gestión, primero cuenta con diversos mecanismos de almacenamiento en la nube para diferentes tipos de datos urbanos. Este componente arquitectural diseña estructuras de indexación únicas y algoritmos de recuperación para datos con propiedades espacio temporales como bien pueden ser indexación hibrida de estructuras para organizar los datos a través de diferentes dominios. Estas técnicas de indexación y algoritmos de recuperación tienen sus bases en algoritmos de machine learning y minería de datos. Esta capa también usa una serie de algoritmos avanzados para la administración de datos como lo son el map-matching , algoritmo de cobertura máxima y despacho dinámico, en donde se puede resolver problemas de computación urbana.

4.2.1.3 Análisis de datos urbano. En esta capa se aplica modelos de minería de datos y algoritmos de aprendizaje automatico para desbloquear el poder del conocimiento de los datos a través de diferentes dominios. Esta capa adapta los datos a modelos de minería de datos y machine learning, como procesamiento mediante clúster, clasificación y algoritmos de regresión, manipulando las

propiedades espacio-temporales únicas de los datos , la consistencia de la destancia espacial, jerarquía espacial, cercanía temporal, periodo y tendencia. Esta capa también fusiona el conocimiento de múltiples conjunto de datos basados en métodos de fusión en los dominios, tales como los basados en aprendizaje en profundidad ,multi vista, fusión de datos basados en dependencia probabilística . Como muchas aplicaciones en l acomputación urbana necesita servicios instantáneos, es también importante combinar técnicas de bases de datos con algoritmos de aprendizaje automático dentro de tareas de minería de datos. Una de las funciones importantes de esta capa es llenadas los valores perdidos en los datos espacio temporales, modelos predictivos e inferencia de causalidad

4.2.1.4 Oferta de servicio. Esta capa ofrece una interface que permite a los sistemas de dominios llamar el conocimiento adquirido sobre los datos desde una aplicación de computación urbana , a través de plataformas de computación en la nube. Como la computación urbana es un campo interdisciplinario, el conocimiento de los datos debe ser integrado en un sistema de dominio existente para informar su toma de decisiones. En adición a esto, es de vital importancia habilitar una interfaz que permita la visualización del análisis de los datos, donde se combine la sabiduría humana con la inteligencia de las máquinas. En términos de tiempo, los servicios basados en computación urbana ofrecen sus funciones para ayudar en tres situaciones : Para entender la situación actual, predecir el futuro, diagnosticar el pasado. Por ejemplo, inferir la calidad del aire en tiempo real, predecir cómo será en el futuro o diagnosticar la causas del estado de esta analizando su comportamiento a través de periodos de tiempos.Basados en los dominios los servicios son creados para ayudar el análisis en áreas como el transporte ciudadano,

la protección ambiental, planeación urbana, ahorro de energía, entretenimiento o seguridad pública.

4.2.2 Desafíos de la computación urbana. Los objetivos que busca cumplir la computación urbana mediante el establecimiento de un framework general resulta en afrontar tres desafíos a los cuales se ve enfrentado[25]:

4.2.2.1 Detección urbana y adquisición de datos. Lo primero es la adquisición de datos mediante diversas técnicas que puedan discretamente y continuamente recolectar datos a lo largo y ancho de la ciudad. Esto no es tarea fácil . Monitorear el flujo de tráfico en un segmento de la carretera es fácil, pero hacerlo de manera continua explorando el tráfico de toda la ciudad es realmente un reto que no se aborda completamente solo poniendo sensores en todas las calles. Construyendo una nueva infraestructura de detección mediante sensores podría ayudar a cumplir los objetivos que se plantea la computación urbana, pero a su vez agravaría la carga de información en la ciudad.Pretender a los humanos como un sensor es un nuevo concepto que tal vez ayude a encarar dicha problemática. Por ejemplo, cuando las personas publican en una red social , están ayudando a entender los eventos que están pasando alrededor de la misma. Cuando muchas personas sobre una avenida, sus rutas trazadas en sus GPS pueden reflejar los patrones y anomalías presentadas en dicha carretera. Sin embargo, como una moneda de dos caras, a pesar de la felxibilidad e inteligencia de los sensores humanos, estos a su vez también traen tres desafíos que se tienen que abordar:

4.2.2.1.1 Consumo de energía y privacidad. Esto no es un problema no trivial para aplicaciones de detección participativa, donde los usuarios proactivamente contribuyan sus datos (usualmente usando un celular inteligente) para ahorrar de un celular y proteger su privacidad durante dicho proceso. Existe un intercambio entre energía, privacidad, y la utilidad de compartir los datos.

4.2.2.1.1 pérdida de control y sensores distribuidos sin uniformidad . Se pueden poner sensores tradicionales en cualquier parte y configurarlos para que frecuentemente estén enviando información. Sin embargo, no se puede controlar a las personas , quienes pueden enviar información al tiempo que ellos quieran, o si lo prefieren no compartir información. En algunos lugares, puede que no haya personas en algunos momentos, y eso inevitablemente resulta en la perdida de datos y en un problema esparcido. También se puede presentar el caso en que una persona genere contenido desde algún lugar (con mucha gente) y puede presentarse información redundante, adicionando más carga, comunicación y almacenamiento innecesariamente.

4.2.2.1.2 Datos implícitos, sin estructura, y con ruido. Los datos generados tradicionalmente por el sistema de sensores son estructurados, explícitos y limpios, además de que son relativamente fáciles de entender. Sin embargo, los datos contribuidos por los usuarios son usualmente libres de formato, tales como textos e imágenes, o no son explícitos como lo son los sensores tradicionales. Algunas veces, la información de generada por los sensores humanos es también propensa al ruido.

4.2.2.2 Computación con datos heterogéneos. Resolver problemas que se presentan en un entorno urbano involucra un ancho rango de factores. Sin embargo, existen técnicas de aprendizaje automático y minería de datos que se encargan de un tipo de dato; por ejemplo, la computación de visualización trata con imágenes, y lenguaje natural basado en textos. Tratar de extraer solo características de diferentes fuentes de información no trae implícitamente un mejoramiento en el rendimiento de un modelo, pues esto aumenta el riesgo de agravar el problema de obtener datos esparcidos. Por eso es necesario tratar de manera adecuada la extracción de datos de múltiples fuentes , y esto se logra mediante el uso de modelos analíticos avanzados que refuerce el aprendizaje y conocimiento entre los datos heterogéneos que se extraen de estos.

Muchos escenarios presentes en la computación urbana requieren de respuestas instantáneas. Además de aumentar el número de maquinas dentro de un clúster, también se hace necesario la implementación de algoritmos de aprendizaje automático y de minería de datos que promuevan la eficiencia y eficacia de los modelos mediante el conocimiento entre datos heterogéneos.

4.2.2.3 Sistemas híbridos que combinan mundos físicos y virtuales. A diferencia de los motores de búsquedas o videos juegos donde los datos son generados en un mundo digital, la computación urbana integra datos generados en ambos mundos; como por ejemplo combina el tráfico con las redes sociales. Alternativamente los datos son generados en el mundo físico y enviados de alguna manera al mundo digital, como un sistema en la nube; después estos datos pueden ser procesados con otros datos en dicha nube, y el conocimiento obtenido de diferentes tipos de datos puede ser enviado a usuarios del mundo físico mediante programas clientes vía web o móvil. Esto es uno de los grandes retos de la computación urbana, pues es necesario

contar con un sistema que interactúe de manera fluida entre ambos mundos, que permita la comunicación de muchos dispositivos de manera concurrente y de forma simultánea, además que se pueda enviar y recibir datos en diferentes formato en diversas frecuencias.

4.3 Datos Urbanos

A continuación, se muestra las fuentes de datos frecuentemente presentes en la Computación Urbana[25].

4.3.1 Datos geográficos. Son datos generados dentro de un dominio espacial en donde se puede ver representar por ejemplo una red que muestre todas las avenidas principales de una ciudad mediante un grafo, en el cual las carreteras se identifiquen como las aristas y las intersecciones como los nodos. Estos datos son usados comúnmente en el monitoreo y predicción del trafico, planeación urbana, direccionamiento urbano, y análisis de consumo de energía. También se representa otras propiedades como la longitud, restricción de velocidad, tipo de carretera, numero de líneas, entre muchas otras. Todas estas propiedades arrojan un gran número de datos que tienen que ser de alguna manera almacenados, procesados, analizados y representados para que puedan tener un verdadero valor para tomar decisiones en base a los resultados arrojados.

4.3.2 Datos de tráfico. Existen muchas maneras de recolectar datos del tráfico de una ciudad, entre ellas se tiene los sensores, cámaras de vigilancia, monitoreo de vehículos mediante GPS. Los sensores se ubican generalmente en las avenidas principales , pues es muy costoso en términos de dinero tener sensores en cada calle de la ciudad y más si es una el área metropolitana

es de gran tamaño. Mediante la instalación de un par de sensores y conociendo la distancia entre estos, se puede conocer entonces la velocidad en la que viaja un auto, y si se cuenta el número de carros que pasa en un determinado intervalo de tiempo se puede establecer el volumen del tráfico que hay en una avenida. Con las cámaras de seguridad existe la posibilidad entonces de conocer como se viaja un vehículo (Esto es importante en términos de seguridad social), como por ejemplo las condiciones de un auto, cuantas personas viajan dentro de uno, que se transporta en ello, entre otros. Esto da como resultado un monitoreo de las condiciones del trafico dentro de una ciudad.

El monitoreo del tráfico usando el GPS de los autos genera constantemente datos que deben ser analizados en tiempo real en donde usualmente se envía a un sistema central que tiene la capacidad de recomendar trayectorias, establecer la ruta más corta o simplemente proporcionar la forma de llegar a un destino.

4.3.3 Señales de teléfonos móviles. Son datos generados por causa de la comunicación celular, en donde se puede evidenciar detalles de una llamada como de qué lugar se está realizando la llamada o de donde se recibe, el número telefónico de quienes establecen comunicación, duración de la llamada y la hora en la que comienza la misma.

4.3.4 Datos de conmutación. La movilidad de las personas genera diariamente inmensas cantidades de datos de conmutación (cambio), como el intercambio de información en las transacciones bancarias, o en la utilización de tarjetas en las estaciones de buses. En la actualidad casi toda esta información se genera a través de tarjetas que son utilizadas mediante dispositivos lectores, que se encargan de almacenar información y enviarla a centrales de información y así

establecer un mundo virtualizado y digital. Toda estos datos arroja información valiosa que al ser correctamente analizada ayuda en gran manera a tomar decisiones en aspectos fundamentales de una ciudad como el comportamiento financiero de una ciudad, la movilidad, la planeación urbana y ambiental y muchos otros más.

4.3.5 Datos de monitoreo ambiental. Se incluye datos generados por monitoreo meteorológico, donde se puede calcular y obtener valores que hacen referencia a la humedad relativa, presión, velocidad del viento, condiciones climáticas y los cuales pueden ser publicados en sitios y aplicaciones web. Se puede determinar entonces condiciones naturales como la calidad del aire mediante el calculo cuantitativo que permita conocer la concentración del $N0_2$ o de SO_2.

4.3.6 Datos de redes sociales. Los datos de redes sociales consiste básicamente en dos partes. La primera hace referencia a la estructura social, representado usualmente por un grafo, lo que denota la reación, interdependencia o interación entre usuarios. La segunda parte está determinada por el contenido generado por los usuarios de las redes sociales, tales como textos, videos ,imágenes y fotografías, en donde contiene abundante información acerca de un usuario pues mediante esta se puede establecer comportamientos e intereses de los mismos. Cuando se añade ubicación a las redes sociales, se puede modelar el movimiento de las personas en áreas urbanas, lo que nos ayuda a entender anomalías urbanas.

4.3.7 Datos de economía. Estos datos hacen referencia a los datos generados por movimientos de utilización de transferencias bancarias, pagos de facturas, compras de inmuebles

y compras de diferentes índoles en donde se representa la dinámica económica de una ciudad. Estos datos ayudan a entender el ritmo financiero que se lleva en el entorno urbano y también son un punto de partida para predecir el futuro económico de una sociedad.

4.3.8 Energía. Normalmente en este tipo de datos puede comunicar aspectos como el consumo de gas en los vehículos mediante la concurrencia en las estaciones de gas. Estos datos se pueden obtener mediante sensores o inferir de otras fuentes de datos implícitamente. Estos datos pueden ser usados en la evaluación de la infraestructura energética de una ciudad, o en el calculo de la emisión de los gases y polución generados por los vehículos. Además el consumo energético también cabe dentro de este tipo de datos, pues la información arrojada por el análisis de estos puede ayudar a optimizar el uso de energía residencial por ejemplo cambiando los picos de alta carga energética a periodos de baja demanda de consumo.

4.3.9 Datos del sector Salud. Uno de los usos más altruista de la computación urbana está en el sector de la salud. En la actualidad uno de los objetivos de los entes gubernamentales en temas de tecnología es ayudar a la investigación en la cura y tratamiento de enfermedades. Este tipo de datos puede ser utilizado para el análisis de enfermedades y ayudar a diagnosticar las mismas de la mejor manera, estableciendo tratamientos que mejoren la calidad de vida de las personas.

4.4 Aplicaciones de la computación urbana

En esta sección se muestra la aplicación de la Computación Urbana en donde se destacan seis categorías las cuales se nombran a continuación [25].

4.4.1 Planeación Urbana. Una efectiva planeación es de gran importancia para construir una ciudad inteligente. Formular la planeación urbano requiere de un ancho rengo de factores, tales como el flujo de tráfico, movilidad de la personas, e infraestructura de movilidad. Los métodos tradicionales de recolección de información como las encuestas han resultado en grandes esfuerzos que en la mayoría de ocasiones no arroja suficiente información acerca de muchos aspectos que se deben abordar en una buena planeación urbana. La computación urbana se enfoca en proveer oportunidades de mejorar la formulación de planeaciones futuras buscando siempre en arrojar información que facilite los procesos como el mejoramiento de los problemas fundamentales en las redes de transporte, descubrimiento de regiones funcionales y detección de fronteras de una ciudad.

4.4.2 Sistemas de transporte. Entre las muchas aplicaciones de la computación urbana, el monitoreo constante de los sistemas de transporte es de los más comunes entre la población de una ciudad, en los últimos años se ha observado surgir nuevas aplicaciones que permiten a las personas tener mejores experiencias en este aspecto. Es por esto que la Computación Urbana debe proveer mecanismos que faciliten la movilidad en la ciudad, apuntándole a mejorar cada día elementos como la experiencia de conducción, los servicios de taxis y transporte publico en general.

4.4.3 Medio Ambiente. En la actualidad el ambiente es uno de los temas de conversación en los diálogos gubernamentales que se desarrollan en el mundo, pues durante los últimos años la contaminación ha venido creciendo de manera exponencial. Es aquí donde la computación urbana cumple un papel que puede ser de gran relevancia para la ciudad, pues mediante un

monitoreo constante de las condiciones ambientales en el entorno urbano se pueden prevenir, atender y corregir muchas situaciones ambientales indeseadas. La Computación Urbana debe ayudar a mejorar aspectos como el monitoreo de la calidad del aire para toma de decisiones tempranas, contaminación auditiva y reducción de generación de basura.

4.4.4 Consumo de energía. El rápido progreso de las urbanizaciones ha traído consigo más consumo de energía. Debido a esto es necesario implementar tecnologías que ayuden a la disminución del consumo de energía. Se deben desarrollar entonces tecnologías que permitan mediante el uso de la computación Urbana controlar de alguna manera el consumo energético y de gas a través de la ciudad, apoyándose de herramientas como son los sensores en donde se envié información constante acerca de cuánto gas se abastece en una estación de gas diariamente o cuanto energía consume un sector residencial; esto con el fin de tomar medidas y mejores decisiones que beneficien a la colectividad urbana.

4.4.5 Redes Sociales. Las redes sociales en la actualidad hacen parte de la vida de las personas, y para algunas (por no decir que la mayoría) se han convertido en una necesidad. El contenido generado por las redes sociales como se dijo anteriormente se puede dividir en dos partes. La primera es la interacción entre usuarios y la segunda es lo que publican los mismos. El uso contante de las redes sociales por los habitantes puede traer una serie de ventajas si se sabe aprovechar esto.

Cuando las personas agregan a sus contenidos ubicación permiten entonces que mediante ciertas herramientas se detecten anomalías (en caso de que se presenten) y ayuden a tomar medidas al respecto. Además de esto , el análisis basado en datos arrojados por las redes sociales

ayuda a entender de mejor manera los diferentes comportamientos que se presenta dentro de una ciudad determinando así patrones que ayuden a entender ciertas situaciones presentadas en la urbanización.

4.4.6 Economía. La economía como aspecto importante dentro de una sociedad siempre ha sido constantemente monitoreada. Aún así existen temporadas en las que ciertas sociedades se han visto golpeadas por diferentes sucesos que han terminado en crisis económicas. Mediante la computación urbana se puede entender la dinámica financiera de una ciudad, hallando tendencias económicas que ayuden a predecir los comportamientos de futuros del aspecto en cuestión, esto con el ánimo de tomar las mejores medidas y decisiones, para evitar situaciones no deseadas o tomar una elección que mejore la calidad económica de la urbanización.

5.Urban Big Data: Aplicación Del Big Data A La Computación Urbana.

5.1 Definición y características de Urban Big Data

El término de Urban Big Data hace referencia a la cantidad masiva de datos estáticos y dinámicos generados de diferentes sujetos y objetos en contextos urbanos, los cuales son recolectados por los gobiernos de las ciudades, instituciones públicas, empresas , y personas utilizando nuevas tecnologías[27].

Grandes cantidades de datos pueden ser compartidos,integrados, procesados y analizados para proporcionar a las personas un conocimiento más profundo sobre el estado de las operaciones urbanas y ayudar a tomar mejores decisiones en la administración ciudadana con bases más estables y enfoques más científicos, y de este modo optimizar la utilización de los recursos urbanos, reducir los costos de operación en los sistemas urbanos, y promocionar un ambiente seguro,armonioso e inteligente.

Además de la características nombradas en el capítulo dos (velocidad,variedad,veracidad y valor), el Urban Big Data también posee unas características adicionales , las cuales se mencionan a continuación:

5.1.1 Jerarquía. La jerarquía presente en el Urban Big Data refleja profundamente la organización jerárquica en los sistemas físicos y sociales de una ciudad. Por ejemplo, los registros médicos electrónicos son categorizados por hospital o región, mientras que las imágenes médicas pueden ser categorizadas en términos de dispositivos médicos individuales o propios de los hospitales.

5.1.2 Integridad. Mantener la integridad de los datos es parte importante en el objetivo de lograr un mejoramiento en las condiciones urbanas de una ciudad. Es cierto que los sistemas urbanos evolucionan cada día de manera más rápida debido también a la aparición de nuevas tecnologías que facilitan este proceso, y es debido a la rápida mejora que se presenta en la integridad de los datos,que el Urban Big Data ha adquirido la capacidad de descubrir con mayor precisión la dinámica urbana global.

5.1.3 Correlación. Todos los tipos de datos urbanos están altamente relacionados con otros. Por ejemplo, la información acerca de la logística urbana puede estrechamente relacionar los datos de logísticas de diferentes empresas de manufactura, comercial, de transporte, e incluso industrias fnancieras. Dichas correlaciones pueden ser utilizadas para mutuas colaboraciones, también para realizar minería de datos de las operaciones urbanas.

Debido a estás características únicas y generales, Urban Big Data debe ser aplicado usando nuevas tecnologías de procesamiento de datos. Todo el proceso desde la adquisición de los datos hasta el modelamiento de los datos procesados , debe ser entonces automatizado. Este proceso ocurre de la siguiente manera:

El primer paso es la recolección y almacenamiento de los datos originales, donde se incluyen los patrones de extracción y filtrado de los datos obtenidos de las fuentes de datos deseados de acuerdo a los requerimientos , después de esto se debe realizar un proceso de limpieza y preprocesamiento de los datos adquiridos mediante la utilización de mecanismos como optimización, organización y normalización de los datos , así como llenar los campos de datos faltantes, los cuales permitan el establecimiento del data set o conjunto de datos a ser procesados.

El segundo paso es el procesamiento y análisis del dataset mediante técnicas como el análisis lineal, análisis no lineal , análisis secuencial, análisis de curva variable, entre muchas otras; y con esto proceder con una categorización de los datos y un análisis de las relaciones de los inter datos e inter categorías con apoyo de procedimientos como la regresión logística. El tercer paso es la identificación de las relaciones inherentes entre los datos categorizados y descubrir la existencia de patrones, reglas, y aprendizaje a través de algoritmos de redes neuronales o algoritmos genéticos. Finalmente las relaciones halladas entre las diferentes variables son explicadas en una interactiva y visual manera que exprese y profundice el entendimiento de los resultados.

5.2 Categorización de la Urban Big data

Urban Big Data describe el estado en tiempo real de varios elementos urbanos, incluyendo edificios, calles, alcantarillado, medio ambiente, empresas, economía, productos, tiendas, medicina, cultura, educación, trafico, orden público, entre otros [27].

Urban Big data puede ser categorizada en cinco tipos:

- Datos de sensores sobre infraestructura y objetos en movimiento,
- Datos de usuarios,
- Datos de administración gubernamental,
- Datos de registros de clientes y transacciones,
- Datos de artes y humanidades.

A continuación se muestra los cinco tipos de Urban Big Data con un respectivo ejemplo y un grupo de usuario.

Tipo	Ejemplo	Grupo de usuarios.
Datos de sensores sobre infraestructura y objetos móviles	Internet de las cosas; Sistemas de sensores para administración de tráfico, telefonía celular y edificios; cámaras de monitoreo	Operaciones urbanas públicas y privadas, investigadores científicos de ingeniería.
Datos de usuarios	Sistemas de detección participativa, redes sociales, GPS.	Empresas privadas, organizaciones publicadas enfocadas a clientes, desarrolladores independientes, investigadores en ciencias sociales y ciencia de datos.
Datos de administración gubernamental	Datos de administración pública sobre transacciones, impuestos, pagos; datos básicos públicos sobre la población, trafico, geografía.	Innovadores, hackers, instituciones de datos gubernamentales, científicos sociales.
Datos de clientes y registros de transacciones	Tarjetas de crédito, datos de clientes, datos de utilidades publicas e instituciones publicas	Organizaciones privadas, instituciones publicas, desarrolladores independientes.
Datos de artes y humanidades	Textos, imágenes, audios, videos, datos artísticos y material cultural, objetos digitales.	Diseñadores Urbanos, organizaciones de historia, arquitectura, arte y de humanidades

Tabla 2. Tipos de Urban Big data

5.3 Aplicaciones de Urban Big Data en el Desarrollo Urbano

La llegada de la Urban Big data no solo proporciona un nuevo enfoque hacia el estudio de las operaciones y desarrollo urbano, también ofrece una nueva oportunidad de renovar las ventajas competitivas de las ciudades. En el contexto de la rápida informatización y digitalización global, la Big Data se ha convertido en un recurso estratégico vital para las urbanizaciones.El fortalecimiento de las competencias urbanas requiere que cada ciudad haga uso adecuado de sus ventajas y libere el valor potencial de sus recursos, buscando la mejora de los beneficios socioeconómicos del Big Data. El Big Data Juega un importante papel en el desarrollo urbano por las siguientes razones[27]:

- Promoción de compartidos, intensivos, integrados y colaborativos factores de producción basados en la web,
- Facilitación de negocios de innovación, tecnología y recursos humanos ,
- Mejoramiento de los núcleos de negocio de las empresas.

Además de esto el uso constante del Big Data fortalece el surgimiento de nuevos patrones de negocio y nuevos puntos de crecimiento económico.Básicamente, el uso de Urban Big Data explora el contexto urbano y el proceso de urbanización es analizado, visualizado.

5.4 Urban Big Data y las Ciudades Inteligentes.

Originalmente propuesto por IBM en 2008, el concepto de “Smart City” o ciudad inteligente se enfoca en la medida, interconexión e inteligencia, buscando aplicar sistemas específicos de tecnologías de la información a las herramientas administración urbana. El desarrollo de lo

relacionado con las ciudades inteligentes comúnmente es instruido desde los campos de la administración gubernamental e inteligencia de servicios.

La promoción de una ciudad inteligente hace referencia al proceso del desarrollo de un espacio ternario, comprendido por instalaciones urbanas físicas, sociedad y datos urbanos, con un enfoque científico basado en la inteligecia consolidad de los ciudadanos, empresas, y gobiernos. El desarrollo de ciudades inteligentes es un proceso que va desde la descentralización hasta centralización, y desde lo superficial hasta lo más profundo del contexto. Un apropiado punto de partida para la construcción de una ciudad inteligentes es la implementación de un sistema inteligente basado en los datos urbanos ya existentes. La tarea subsecuente es mejorar la automatización de la cadena de valor de los datos en la infraestructura urbana física que de manera gradual integre y comparta datos de forma innovadora con aplicaciones de Urban Big data. El objetivo es la promoción del desarrollo de la toma de decisiones a nivel macro y micro servicios. Se puede entonces proponer un desarrollo de un modelo de una ciudad inteligente el cual consta de cinco partes:

1. Sistema de soporte infraestructural.
2. Sistema de aplicaciones.
3. Sistema industrial.
4. Sistema de indexación.
5. Sistema de aseguramiento operativo

El proceso de desarrollo ocurre de la siguiente manera:

1. Grandes volúmenes de datos estructurados y no estructurados son generados por diferentes fuentes urbanas, y consolidados dentro de una plataforma de datos urbanos unificada, generando así una base de datos de información urbana.
2. Por correlación, integración , limpieza, procesamiento, análisis, minería, y visualización, se obtiene información de valor de los datos que puedan reflejar de mejor manera el curso de los eventos en las ciudades en aras de satisfacer las necesidades de asuntos gubernamentales, comercio, y administración urbana, y que también ayude al mejoramiento de las capacidades de las tomas de decisiones.
3. La transformación y mejoramientos de otras industrias son promocionadas al lado de el desarrollo de la industria de Big Data, y el desarrollo de la informatización e inteligencia urbana es acelerada.
4. Una adecuada indexación para el desarrollo de mediciones del nivel y efecto de la Urban Big Data son creados, y un mecanismo de aseguramiento operativo de Big Data es desarrollado para mejorar la estabilidad y confiabilidad de la operación de la arquitectura de servicios Urban Big Data.

5.5 Puntos Claves del desarrollo Urban Big Data

La clave para el desarrollo de una ciudad inteligentes el mejoramiento a un alto nivel de el diseño de un modelo de Urban Big Data y la definición de los puntos clave para el desarrollo del mismo. Se propone entonces algunos puntos clave para el desarrollo de una ciudad inteligente basado en cuatro perspectivas: Soporte Infraestructural, gobernación urbana, servicios urbanos, y desarrollo económico[27].

5.5.1 Unificación de los sistemas de soporte infraestructural de Big Data. Una ciudad debe construir una plataforma de Internet de las cosas (IoT) para instalaciones públicas. La arquitectura tecnológica de esta plataforma está caracterizada por una plataforma de servicios integrados y una aplicación de módulos interoperables, mientras que el sistema de administración está caracterizado por operaciones profesionales y servicios abiertos. La plataforma debe estar destinada a detectar dinámicamente los cambios fluctuantes en las condiciones de infraestructura urbana, estar bien informada acerca del estado y las características de las operaciones urbanas, proveer diagnósticos precisos del desarrollo de tendencias de elementos urbanos y proporcionar un importante apoyo en el proceso de toma de decisiones para el desarrollo urbano. Para dicho fin, la ciudad debe inspeccionar y determinar status quo de los datos de los recursos gubernamentales, los campos de los datos que pueden ser compartidos o de libre acceso, y listar los datos que son confidenciales o que involucra un componente de seguridad pública y privacidad personal. La ciudad también debe construir una arquitectura de gobierno de datos urbanos, y establecer un mecanismo donde se puedan compartir e intercambiar servicios que permitan limpiar, procesar, integrar, aplicar minería de datos, y compartir los datos urbanos. Finalmente se debe construir una plataforma abierta unificada para el gobierno de datos y desarrollas un centro de servicios que permita al público acceder a los datos para ayudar en aspectos como el tráfico, medicina, seguridad social, geografía, cultura, educación, ciencia y tecnología, entre muchos otros más.

5.5.2 Unificación de los sistemas de soporte infraestructural de Big Data. La gobernación Urbana involucra una variedad de campos, incluyendo la administración de la seguridad pública, planeación y construcción urbana, supervisión de mercado, y protección ambiental. Por lo tanto, es importante aplicar Big Data en la gobernación urbana. La plataforma de aplicación de Big

Data para la gobernanza social implica la referencia de datos en diversos campos, incluyendo energía, electricidad, protección ambiental, administración ambiental, conservación del agua, salud, trafico, meteorología, entre otros. Los recursos de datos deben ser procesados, analizados y aplicados más profundamente para identificar problemas de administración social rápidamente, proveer un sistema de alertas tempranas, y resolver problemas satisfactoriamente, mejorando así la gobernación urbana de base. Para lograr un integración multidireccional, y una colaboración interdepartamental, una ciudad debe construir una plataforma de gestión para los datos e información de registro de bienes raíces, desarrollar un sistema unificado de servicios de Big Data y construir una gran plataforma de aplicaciones Big Data para la planificación Urbana en apoyo de la gestión interdepartamental para proporcionar un soporte a todo el proceso, desde la formulación y revisión, hasta la implementación de la planificación urbana.

6.CONCLUSIONES

- Big Data no hace referencia únicamente a los datos, sino que también comprende todo el espectro de técnicas, métodos, herramientas y tecnologías alternativas, que permiten resolver problemas que involucran cierta complejidad, de una forma más eficaz y eficiente que los métodos tradicionales. A pesar de que con Big Data se puede lograr estos beneficios (eficacia y eficiencia) en problemas complejos de análisis de datos, Las técnicas utilizadas en los entorno Big Data no está desplazando las técnicas tradicionales. En lugar de esto, ambos métodos (los tradicionales y los relacionados con Big Data) se están complementando en pro de implementar soluciones que involucren todos los escenarios posibles.
- El desarrollo de las ciudades inteligentes mediante la implementación de mecanismos que involucre Urban Big Data en la planificación urbana pueden tener un gran impacto sobre el desarrollo nacional en un país. Estos esfuerzos pueden aumentar el poder de decisión de la sociedad al permitirles tomar decisiones inteligentes y efectivas en tiempos oportunos.
- El Urban Big Data desempeña un papel fundamental en el desarrollo de la inteligencia urbana. A medida que la población y la producción económica se concentra en las ciudades, el desarrollo urbano con éxito indicaría que el cuerpo principal nacional está bien desarrollado.
- El uso de grandes volúmenes de datos urbanos administrados y abiertos con éxito promoverá el desarrollo de una industria de servicios basada en el conocimiento urbano,

creará nuevos mercados y oportunidades de negocios y promoverá el desarrollo de la inteligencia urbana.

7.BIBLIOGRAFÍA.

1. Big data - Explicación y definición de big data (s.f). Recuperado desde : http://www.quees.info/que-es-big-data.html.
2. Computación urbana: Aplicaciones tecnológicas para mejorar la vida en la ciudad (2008). Recuperado desde http://gestionpublicave.blogspot.com.co/2008/05/computacin-urbana-aplicaciones.html.
3. Barranco, R. (2012)., ¿Qué es Big Data?, IBM Software Group México, recuperados desde https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/index.html
4. Moreno, J.,Serrano, M. A.,Fernández E.(2016)., *Main Issues in Big Data Security,* Future Internet, 8,44, doi:10.3390/fi8030044.
5. Sagiroglu, S.; Sinanc, D. Big data: A review. In Proceedings of the 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA, USA, 20–24 May 2013; pp. 42–47.
6. Joyane Aguilar , L. (2013) , *Big Data: Analisis de grandes volumenes de datos en Organizaciones*, Mexico: AlfaOmega grupo editor.
7. Eaton, C. , Deroos, D., Deutsch, T., Lapis, G. (2012). , Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, Estados Unidos: The McGraw-Hill Companies.
8. Soubra, D. (2012,julio). The 3Vs that define Big Data. Data Science Central . Recuperado de http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data .

9. Velocidad, variedad y volumen, las tres magnitudes clave de Big data. (s.f). Recuperado de https://blog.es.logicalis.com/analytics/velocidad-variedad-y-volumen-las-3-magnitudes-clave-de-big-data .
10. McAfee, A., Brynjolfsson, E. (2012). Big Data: The Management Revolution. Harvard Busins Review. Recuperado el 31 enero de 2017, desde http://www.rosebt.com/uploads/8/1/8/1/8181762/big_data_the_management_revolution.pdf .
11. Schroeck, M.,Schroeck, R., Smart, J., Tufano,P.(2012). Analitycs: El uso de big data en el mundo real.
12. Las 7 V del Big data: Características más importantes. (junio 28, 2016). Recuperado el 07 de febrero de 2017 , desde http://www.iic.uam.es/innovacion/big-data-caracteristicas-mas-importantes-7-v/#valor-datos.
13. Mora L.(*2016*). Qué es Big Data: fases y elementos. Recuperado el 05 de mayo de 2016 desde https://www.veinteractive.com/es/blog/que-es-big-data-fases-elementos/
14. NIST Big Data Working Group (NBD-PWG) (2013), NIST Big Data Interoperability Framework: Volume 6, Reference Architecture. U.S, 100 Bureau Drive (Mail Stop 8900) Gaithersburg, MD 20899-8930.
15. Big Data, tres casos de éxito: T- Mobile, Unilever y MoneyBall ,(26 Marzo 2013), recuperado de http://blogginzenith.zenithmedia.es/big-data-tres-casos-de-exito-t-mobile-unilever-y-moneyball.
16. Rayo, A. (2016) Tipos de datos en Big Data: clasificación por categoría y por origen. Recuperado desde: http://www.bit.es/knowledge-center/tipos-de-datos-en-big-data/

17. Rayo, A. (2016). Análisis de Datos en Big Data: tipos y fases del análisis. Recuperado desde: http://www.bit.es/knowledge-center/analisis-de-datos-en-big-data/.

18. Archanco (2016). Las 13 mejores técnicas de análisis de datos que todo directivo debe conocer. Recuperado desde: http://papelesdeinteligencia.com/tecnicas-de-analisis-de-datos/

19. Reglas de Asociación. (2012). Recuperado desde: https://ccc.inaoep.mx/emorales/Cursos-NvoAprend/node18.html

20. Introducción a Hadoop y su ecosistema (s.f). Recuperado desde: http://www.ticout.com/log2013/04/02introduccion-a-hadoop-y-su-ecosistema/

21. Hvivanir, (2014).Aquitectura HDFS. Recuperado desde: .https://hvivani.com.ar/2014/07/19/arquitectura-hdfs/

22. What is MapReduce?. (s.f). Recuperado desde : https://www-01.ibm.com/software/data-infosphere/hadoop/mapreduce/

23. Olmedo, Y. (2012). ¿Qué es MapReduce?. Recuperado desde: http://blogs.solidq.com-es/big-data/que-es-mapreduce/ .

24. Gracia, L. M., (2012). ¿Qué es Apache Flume ?., Recuperado desde: https://unpocodejava.wordpress.com/2012/10/25/que-esapache-flume/ .

25. Zheng Y. (2017) , Urban computing: enabling urban intelligence with big data,Beijing,China, 1-3 , doi:10.1007/s11704-016-6907-2

26. Lam, Chuck (28 de julio de 2010). Hadoop in Action (1st edición). Manning Publications. p. 325. ISBN 1935182196.

27. Pan, Y.,Tian, Y,.Liun, X., Gu, D., Hua G. (2016). Urban Big Data and the Develooment of City Intelligence. China . Department of Automation, Tsinghua University, Beijing 100084, China.