

**MANUAL PRACTICO PARA EL APRENDIZAJE DEL BIG DATA**

DIEGO FERNANDO SANCHEZ RIOS

UNIVERSIDAD TECNOLOGICA DE PEREIRA  
INGENIERIA DE SISTEMAS Y COMPUTACION

PEREIRA

2018

**MANUAL PRACTICO PARA EL APRENDIZAJE DE BIG DATA**

DIEGO FERNANDO SANCHEZ RIOS

Monografía para optar al título de Ingeniero de Sistemas y Computación

Asesor

Roberto Guillermo Solarte

**UNIVERSIDAD TECNOLOGICA DE PEREIRA  
INGENIERIA DE SISTEMAS Y COMPUTACION**

**PEREIRA**

**2018**

## TABLA DE CONTENIDO

Pág.

### Tabla de contenido

INTRODUCCION .....	4
1. GENERALIDADES .....	7
1.2 PLANTEAMIENTO DEL PROBLEMA .....	7
1.2 OBEJTIVOS .....	7
1.2.1 OBJETIVO GENERAL .....	7
1.2.2 OBJETIVOS ESPECIFICOS .....	7
1.3 JUSTIFICACIÓN .....	8
2. MARCO TEORICO .....	9
2.1 ¿QUE ES BIG DATA? .....	9
2.2 4V's DE BIG DATA .....	10
2.3 HISTORIA Y EVOLUCION DE BIG DATA .....	11
2.4 DATOS EN BIG DATA .....	12
3. AREAS DE BIG DATA .....	13
3.1 RECOLECCIÓN .....	13
3.2 ALMACENAMIENTO .....	13
3.3 ANALISIS .....	15
3.4 VISUALIZACION .....	15
4. PARADIGMAS DE BIG DATA .....	16
4.1 MAPREDUCE .....	16
4.2 PROCESAMIENTO MASIVO EN PARALELO (MPP) .....	18
5. PLATAFORMAS DE BIG DATA .....	19
6. Instalacion de un ambiente de Big Data y desarrollo de un caso practico .....	20
6.1 Instalacion de un ambiente de Big Data .....	20
6.2 Caso practico .....	37
7. METODOLOGIA PARA EL APRENDIZAJE .....	53
8. CONCLUSIONES .....	55
9. BIBLIOGRAFIA .....	56

## LISTA DE FIGURAS

	Pag.
Ilustración 1 install jdk .....	21
Ilustración 2 version java .....	21
Ilustración 3 crear grupo .....	22
Ilustración 4 verificar grupo .....	22
Ilustración 5 install ssh .....	22
Ilustración 6 configure ssh .....	23
Ilustración 7 acceso ssh .....	24
Ilustración 8 ipv6 .....	25
Ilustración 9 ipv6 confi .....	25
Ilustración 10 enlace simbolico .....	26
Ilustración 11 whereis .....	27
Ilustración 12 java Configuración de hadoop .....	27
Ilustración 13 hadoop-env.....	27
Ilustración 14 hadoop-env2.....	28
Ilustración 15 core-site .....	29
Ilustración 16 coresite2 .....	30
Ilustración 17 hdfs .....	32
Ilustración 18 mapresite .....	33
Ilustración 19bashrc .....	34
Ilustración 20 bashrc2 .....	34
Ilustración 21 format namenode .....	35
Ilustración 22 localhost 50070.....	36
Ilustración 23 localhost 50090.....	37
Ilustración 24 ppa .....	39
Ilustración 25 actualizar .....	39
Ilustración 26 eliminar binarios.....	40
Ilustración 27 couchbd .....	40
Ilustración 28 detener couchbd .....	40
Ilustración 29 encender couchbd .....	41
Ilustración 30 estado couchbd .....	41
Ilustración 31 pip.....	43
Ilustración 32 tweepy .....	43
Ilustración 33 descarga de couchbd-09 .....	44
Ilustración 34 descomprimir couchbd.....	44
Ilustración 35 instalar paquetes de python.....	44
Ilustración 36 import couchbd .....	45
Ilustración 37 import tweepy .....	45
Ilustración 38 plataforma twitter .....	46
Ilustración 39 crear nueva app.....	46
Ilustración 40 ejemplo tweets.....	47
Ilustración 41 claves tweet.....	47

Ilustración 42 claves de acceso .....	48
Ilustración 43 crear database.....	48
Ilustración 44 codigo python .....	49
Ilustración 45 Cordenadas colombia.....	50
Ilustración 46 Recoleccion de tweets.....	51
Ilustración 47 datos tweet .....	51
Ilustración 48 datos de tweet .....	52

## INTRODUCCION

Big data hace referencia a las combinaciones de datos cuyo tamaño, complejidad y velocidad dificultan la captura, gestión, procesamiento y análisis por parte de las herramientas de software tradicionales. El estudio de Big data se hace complejo debido principalmente a que la mayoría de los datos generados por las tecnologías modernas, como los web logs, la maquinaria, los vehículos, las búsquedas en Internet, las redes sociales como Facebook, computadoras portátiles, teléfonos inteligentes entre muchas otras, son datos no estructurados.

El objetivo de Big data es convertir el dato en información, para facilitar su análisis y así aprovechar su contenido para ayudar en la toma de decisiones, mejorar el marketing de las empresas y otras ventajas que se obtiene cuando se procesan los datos de la manera correcta; Actualmente el análisis de los datos se ha convertido en una ayuda para muchas empresas en la toma de decisiones. El big data es una herramienta moderna que requiere mucha atención, es por esto que se hace necesario elaborar un manual práctico para su aprendizaje y que este brinde un apoyo fundamental en el estudio de lo que es en realidad y de lo que se puede hacer con el Big data.

## **1. GENERALIDADES**

### **1.2 PLANTEAMIENTO DEL PROBLEMA**

Big data ha surgido como un área de estudio importante tanto para los profesionales como para los investigadores y es gracias a la tecnología moderna que su desarrollo e investigación se hace más importante y es de vital importancia que los ingenieros de sistemas o estudiantes interesados tengan conocimientos sobre él, ya que cada vez más las investigaciones tecnológicas combinan Big data y analistas con la nube, internet de las cosas, redes sociales, movilidad, etc.... Y es gracias a esto que muchas empresas han visto una oportunidad de negocio en el análisis de datos y han decidido implementarlo en sus operaciones de comercio; los documentos informativos sobre este tema son muy variados, confusos y la gran mayoría de artículos e investigaciones se encuentran en otros idiomas, las implementaciones que se han hecho son aisladas y no hay un documento claro que brinde la información de cómo se debe integrar el Big data utilizando las herramientas open source.

### **1.2 OBEJTIVOS**

#### **1.2.1 OBJETIVO GENERAL**

Realizar un manual práctico que facilite el proceso de enseñanza-aprendizaje de lo que es Big data, de cómo debe ser implementado y de la importancia que tiene en la actualidad.

#### **1.2.2 OBJETIVOS ESPECIFICOS**

- Identificar y analizar cuáles son las tecnologías requeridas, herramientas de software y el hardware necesarias para la correcta implementación de Big data.
- Configurar e integrar las herramientas para la elaboración del ambiente de Big data
- Elaborar el caso de estudio el cual se va a utilizar para realizar el testeado en el ambiente configurado

### **1.3 JUSTIFICACIÓN**

En la actualidad, con el auge de redes sociales, el internet de la cosas. El crecimiento de la población, entre otros muchos factores, han provocado gran crecimiento de los datos, por lo cual, las empresas y organizaciones han tenido que enfrentarse a nuevos retos que les permitan descubrir, analizar y sobretodo, entender el funcionamiento de herramientas no tan tradicionales. Es por esto, que resulta adecuado e interesante profundizar en este tema y proponer un manual práctico que permita entender lo que es Big Data, lo que significa y su importancia actualmente.

la información que se genera cada segundo y que circula en las plataformas web puede ser vista por muchas empresas como una oportunidad de negocio, si esta información se traduce en cifras quien las utilice podrá detectar tendencias de mercadeo, orientar acciones que se van a llevar a cabo mediante la toma de decisiones entre muchas otras oportunidades que pueden ser provechosas si se aplica Big Data.

La elaboración de este manual sobre Big Data brinda una ventaja en el desarrollo profesional a todas las personas que están involucradas en tecnologías de información, Ingenieros de Sistemas, científicos de datos, analistas, directores de TI que vean en Big Data un elemento competitivo y que deseen tener nuevas estrategias en sus negocios, ya que se podrá conocer más detalladamente de que se trata este tema y de cómo puede ser implementado



## 2. MARCO TEORICO

### 2.1 ¿QUE ES BIG DATA?

Desde su origen han existido diversas definiciones y explicaciones de lo que se significa Big Data; IBM que es una de las empresas más importantes a nivel mundial sobre tecnología define a Big Data como:

“la tendencia en el avance de la tecnología que ha abierto las puertas hacia un nuevo enfoque de entendimiento y toma de decisiones, la cual es utilizada para describir enormes cantidades de datos (estructurados, no estructurados y semi estructurados) que tomaría demasiado tiempo y sería muy costoso cargarlos a un base de datos relacional para su análisis. De tal manera que, el concepto de Big Data aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales”. [1]

Oracle compañía de software especializada en el desarrollo de aplicaciones locales o en la nube que aporta soluciones a muchas empresas da la siguiente definición:

“Big Data describe una estrategia holística de gestión de la información que incluye e integra muchos nuevos tipos de datos y de gestión de datos junto con datos tradicionales”. [2]

Entonces, podemos denominar Big Data como el análisis y gestión de grandes volúmenes de datos los cuales no pueden ser tratados de la manera convencional, y los cuales deben cumplir con la ley de las 4V's del Big Data.

## 2.2 4V's DE BIG DATA

**Volumen:** hace referencia a la cantidad de los datos que se generan por segundo, los datos son generados automáticamente por máquinas, redes e interacciones personales en sistemas como redes sociales; datos de valores desconocidos, como mensajes de Twitter, flujos de clics en páginas web, aplicaciones móviles, tráfico de red, equipos con sensores que capturan datos a la velocidad de la luz, etc. [3]

El volumen delimita el concepto de datos masivos ya que no se pueden almacenar en ordenadores simples por el contrario requieren de una tecnología específica para el almacenamiento.

**Velocidad:** se refiere al ritmo con el que los datos fluyen de fuentes como procesos de negocios, máquinas, redes e interacción humana con cosas como redes sociales, dispositivos móviles, etc. El flujo de datos es masivo y continuo. Estos datos en tiempo real pueden ayudar a los investigadores y las empresas a tomar decisiones valiosas que brindan ventajas competitivas estratégicas y retorno de la inversión si puede manejar la velocidad. [4]

Para aprovechar al máximo el significado de los datos su procesamiento debe realizarse en tiempo real o en el menor tiempo posible. Para mejorar el análisis y la extracción de conclusiones se requiere una velocidad para acceder o visualizar los datos.

**Variedad:** se refiere a las diferentes formas, fuentes y tipos que tienen los datos tanto estructurados como no estructurados o semiestructurados entre los que se incluye documentos de texto, audios, videos, emails, fotos, videos, sistemas de monitorización, PDFs, ficheros de sonido etc.[3]

**Veracidad:** La veracidad es cuán exacto o verdadero puede ser un conjunto de datos. Pero en el contexto de Big Data, la definición de veracidad, adquiere un poco más de significado. Cuando se trata de la precisión del Big Data, no solo se trata de la calidad de los datos en sí, sino también de la confiabilidad de la fuente de datos, el tipo y el procesamiento. Eliminar aspectos como los prejuicios, las

anomalías o incoherencias, la duplicación y la volatilidad son solo algunos de los aspectos que contribuyen a mejorar la precisión de los grandes datos.

La veracidad de los datos implica asegurar que el método de procesamiento de los datos reales tenga sentido en función de las necesidades del negocio y que el resultado sea pertinente a los objetivos. La interpretación de grandes datos de la manera correcta garantiza que los resultados sean relevantes y procesables. [5]

## **2.3 HISTORIA Y EVOLUCION DE BIG DATA**

El nombre de Big Data es un nombre novedoso y el cual ha tenido un auge muy importante en esta era de la tecnología, pero su concepto ha sido implementado muchos años atrás.

En 1880 se realiza un censo en los Estados Unidos de América, censo que tardó 8 años en tabularse, está sobre carga de información como fue denominada, fue fundamental para que se enfocaran en la importancia que tiene el tratamiento de la información y de la necesidad de desarrollar avances en la metodología para el tratamiento de los datos.

Herman Hollerith desarrolló una máquina capaz de tomar la información depositada en tarjetas perforadas y analizarlos; la máquina de Hollerith como fue nombrada implementó un sistema que revolucionó el valor de los datos y disminuyó el tiempo de análisis de estos. [6]

La primera máquina de procesamiento de datos apareció en 1943 y fue desarrollada por los británicos para descifrar los códigos nazis durante la Segunda Guerra Mundial. Este dispositivo, llamado Colossus buscaba patrones en mensajes interceptados a una velocidad de 5.000 caracteres por segundo. De ese modo, se reduce la tarea de semanas enteras a solo unas pocas horas. [7]

En 1965 El gobierno de los Estados Unidos planea el primer centro de datos del mundo para almacenar 742 millones de declaraciones de impuestos y 175 millones de juegos de huellas dactilares en cinta magnética [8]

En la década de los 70 el análisis de los datos empieza hacer prioridad para las predicciones y la toma de decisiones, el modelo Black-Sholes que se crea en 1973 y su propósito era poder predecir el precio óptimo de las acciones en el futuro.[6]

En el año 1999, aparece el término Big Data en Visually Exploring Gigabyte Datasets in Real Time, publicado por la Association for Computing Machinery. En esta publicación se describe el problema que se genera al almacenar grandes cantidades de datos sin una forma adecuada de analizarla. [8]

En el año 2005 la web generada por los usuarios empieza a implementarse con mayor rapidez, la web 2.0 como fue denominada se logra implementando páginas web de estilo HTML con bases de datos basadas en SQL.

En este año también es creada una herramienta de código abierto hadoop cuyo objetivo principal es el almacenamiento y el análisis de grandes datos.

## **2.4 DATOS EN BIG DATA**

La categorización de los datos es importante para analizarlos de la manera adecuada, en Big Data se definen tres tipos, datos estructurados, datos no estructurados y datos semi-estructurados los cuales son definidos teniendo en cuenta su precedencia y forma. A continuación se definirán los tipos de datos:

- Datos Estructurados: se define datos estructurados a los datos que tienen definido su longitud, formato, números, fechas y que tienen esta información almacenada en bases de datos relacionales como SQL. Los podríamos ver como si fuese un archivador perfectamente organizado donde todo está identificado, etiquetado y es de fácil acceso [9]

Así que los datos estructurados son cualquier tipo de dato que se encuentre en un campo fijo dentro archivo o registro.

- Datos no Estructurados: Es aquella información que no está almacenada en tablas de bases de datos y no tiene definida una estructura interna. Estos datos son generados en su mayoría por los usuarios e incluyen

mensajes de correo electrónico mensajes de redes sociales, mensajes instantáneos y otras comunicaciones en tiempo real como documentos, imágenes, audio y vídeo. [10]

- Datos Semi-Estructurados: Son aquellos datos que no residen de bases de datos relacionales, pero que presentan una organización interna que facilita su tratamiento, tales como documentos XML, CSV y datos almacenados en bases de datos NoSQL. [11]

### **3. AREAS DE BIG DATA**

#### **3.1 RECOLECCIÓN**

La recolección de datos hace referencia a una de las disciplinas de big data la cual en muy poco tiempo ha variado con mayor rapidez, esto es debido a que los datos son generados en grandes volúmenes, y son provenientes de muchas fuentes y de diversos dispositivos distribuidos por todo el mundo que transmiten, procesan y recolectan los datos que son generados por las diversas actividades como la información generada por las redes sociales, plataformas digitales, datos de geolocalización, entre muchos otros.

#### **3.2 ALMACENAMIENTO:**

La información se ha convertido en una materia prima de gran valor. El almacenamiento masivo de datos y las nuevas fuentes de obtención de los mismos como las redes sociales, plataformas digitales, buscadores en internet entre otros, no sólo afectan al mundo de los negocios, sino también al ámbito académico y a las Administraciones públicas. Es por esto que se debe llevar a cabo un almacenamiento escalable, es decir se debe implementar un sistema

de almacenamiento que pueda variar su tamaño sin afectar el rendimiento general del sistema.

Debido a esta necesidad han aparecido diferentes soluciones para tratar el almacenamiento masivo de datos, estas soluciones permiten establecer perfiles y hacer clasificaciones de datos, algunas de estas herramientas son:

**Data Warehousing:** Un data warehouse es un repositorio unificado para todos los datos que recogen los diversos sistemas de una empresa. El repositorio puede ser físico o lógico y hace hincapié en la captura de datos de diversas fuentes sobre todo para fines analíticos y de acceso. [12]

En otras palabras un data Warehouse es una arquitectura de almacenamiento de datos que le brinda a las empresas la capacidad de comprender y utilizar sus datos para tomar decisiones estratégicas.

**Business Intelligence:** hace referencia al uso de estrategias y herramientas que sirven para transformar información en conocimiento, con el objetivo de mejorar el proceso de toma de decisiones en una empresa. [13]

Las herramientas de Business Intelligence se han convertido en una potente herramienta, capaz de analizar y procesar infinidad de datos, de infinidad de fuentes y así ayudar a las empresas a extraer conclusiones para mejorar sus cifras de negocio.

**Cloud computing:** ofrece servicios a través de la conectividad y gran escala de Internet. La computación en la nube democratiza el acceso a recursos de software de nivel internacional, pues es una aplicación de software que atiende a diversos clientes. La multilocación es lo que diferencia la computación en la nube de la simple tercerización y de modelos de proveedores de servicios de aplicaciones más antiguos. [14]

La computación en la nube ofrece a quienes adquieren su servicio la capacidad de un pool de recursos de computación, mantenimiento, seguridad de los datos, y fácil acceso a la información.

### 3.3 ANALISIS

El análisis de datos es el proceso de examinar grandes cantidades de datos y así extraer información para descubrir patrones ocultos, correlaciones desconocidas y otra información útil, Lo más importante del análisis de datos es procesar la información de manera eficaz y en un tiempo razonable, de tal manera, que se puedan obtener resultados óptimos, Tal información puede proporcionar ventajas competitivas a través de organizaciones rivales y resultar en beneficios para el negocio. [15]

El análisis de Big data puede hacerse con herramientas de software de uso común en el marco de disciplinas analíticas avanzadas, como el análisis predictivo y la minería de datos.

Algunas de esas herramientas son, Analytics esta herramienta permite coleccionar datos desde un dominio web hasta poder visualizarlos en formas de reporte, Otra herramienta muy utilizada es R, “un lenguaje de programación que facilita tanto el análisis de datos como el desarrollo de nuevo software de estadística. Esta área de Big data se encarga de extraer información relevante hacia el usuario

### 3.4 VISUALIZACION

La visualización de datos permite representar cualquier tipo de información de una forma visual y sencilla, permite difundir el análisis previo de manera precisa y consistente, para ser visualizada al igual que proporciona la opción de comunicar el significado de los datos de una manera más entendible. Algunas herramientas que nos permiten la visualización de los datos son:

**Datawrapper:** Permite visualizaciones interactivas y responsivas. Ofrece desde los clásicos gráficos de barra, tablas y mapas hasta visualizaciones más complejas. También permite personalizar los colores, las fuentes y otros elementos gráficos para adaptar la apariencia del gráfico a la web donde se va a insertar. [16]

**Jupyter:** Es un proyecto de código abierto que permite el análisis de big data, la visualización y la colaboración en tiempo real en el desarrollo de software en

más de una docena de lenguajes de programación. la interfaz contiene el campo para la entrada de código, y la herramienta ejecuta el código para entregar la imagen visualmente legible en función de la técnica de visualización elegida.[17]

**Tableau:** Es un software con una simplicidad de uso y capacidad para producir visualizaciones interactivas más allá de las proporcionadas por soluciones generales. es especialmente adecuado para manejar los grandes y cambiantes conjuntos de datos que se utilizan en las operaciones de Big Data, incluidas las aplicaciones de inteligencia artificial y aprendizaje automático, gracias a la integración con una gran cantidad de soluciones de bases de datos avanzadas, incluidas hadoop, amazon aws, my sql , sap y teradata. [18]

#### **4. PARADIGMAS DE BIG DATA**

El análisis de big data es diferente del análisis tradicional debido al gran aumento en el volumen de datos, debido a esto se hace imposible de manejar los datos usando los sistemas tradicionales de administración de bases de datos relacionales, para esto se necesitan paradigmas de programación que brindan una ayuda en el proceso y manejo de los datos de big data; en esta sesión se estudiarán los paradigmas que se centran en el desarrollo de aplicaciones y la gestión de grandes datos los cuales son MapReduce y MMP (Procesamiento Masivo en Paralelo)

##### **4.1 MAPREDUCE**

MapReduce es un framework que proporciona un sistema de procesamiento de datos paralelo y distribuido. Este paradigma se basa en enviar el proceso computacional al sitio donde residen los datos que se van a tratar, los cuales se coleccionan en un clúster Hadoop. MapReduce posee una arquitectura maestro / esclavo, la cual cuenta con un servidor maestro (JobTracker) y varios servidores esclavos (TaskTrackers), uno por cada nodo del clúster. Cuando se lanza un proceso de MapReduce se distribuyen las tareas entre los diferentes servidores



del cluster y, es el propio framework Hadoop quien gestiona el envío y recepción de datos entre nodos. Una vez se han procesado todos los datos, el usuario recibe el resultado del clúster. [19]

## Algoritmo

El programa MapReduce se ejecuta en tres fases, fase de mapa, fase de mezcla y fase de reducción .

**Fase map o mapeo:** en esta fase se ejecuta la función map, la cual recibe como parámetros una dupla (clave, valor) y esta retorna una lista de pares. El trabajo de la función map es procesar los datos de entrada. Esta función se encarga del mapeo y es aplicada a cada elemento de la entrada de datos, debido a esto se obtendrá una lista de pares por cada llamada a la función map. . En general, los datos de entrada están en forma de archivo o directorio y se almacenan en el sistema de archivos de Hadoop (HDFS) Después se agrupan todos los pares con la misma clave de todas las listas, creando un grupo por cada una de las diferentes claves generadas. La función map procesa los datos y crea varios fragmentos pequeños de datos. [12]

**Fase reduce:** esta etapa es la combinación de la etapa mezcla y la etapa Reduce. La función Reduce se aplica en paralelo para cada grupo de datos que son retornados por la función map. Esta función es llama una vez para cada clave única de la salida de la función map. Con esta clave, se pasa una lista de todos los valores asociados con la clave para que pueda realizar alguna fusión para producir un conjunto más pequeño de los valores. Al iniciar 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 key-value sean contiguos, gracias a esto se obtiene que la operación Reduce se simplifique ya que el archivo se lee secuencialmente. 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 Después del procesamiento, produce un nuevo conjunto de resultados, que se almacenarán en el HDFS. [18]

- Durante un trabajo de MapReduce, Hadoop envía las tareas de Asignar y Reducir a los servidores apropiados del clúster.
- El marco gestiona todos los detalles del paso de datos, como la emisión de tareas, la verificación de la finalización de tareas y la copia de datos en el clúster entre los nodos.
- Después de completar las tareas dadas, el clúster recopila y reduce los datos para formar un resultado apropiado y lo envía de vuelta al servidor de Hadoop.

#### **4.2 PROCESAMIENTO MASIVO EN PARALELO (MPP)**

MPP es un paradigma que permite hacer cálculos para el procesamiento de consultas distribuidas, En MPP el procesamiento de datos se distribuye a través de un banco de nodos de cálculo, estos nodos están separados y procesan los datos en paralelo, los conjuntos de salida a nivel de nodo se ensamblan entre sí para producir un conjunto de resultados final. [20]

Mpp (procesamiento paralelo masivo) es el procesamiento coordinado de un programa por múltiples procesadores que trabajan en diferentes partes del programa, con cada procesador usando su propio sistema operativo y memoria. Típicamente, los procesadores mpp se comunican usando alguna interfaz de mensajería. En algunas implementaciones, hasta 200 o más procesadores pueden trabajar en la misma aplicación. Una disposición de "interconexión" de rutas de datos permite enviar mensajes entre procesadores. [21]

## 5. PLATAFORMAS DE BIG DATA

La plataforma de Big Data es un tipo de solución de tecnologías de información que combina las funciones y capacidades de varias aplicaciones y utilidades de Big Data en una única solución.

La plataforma de Big Data generalmente consiste en almacenamiento de Big Data, servidores, bases de datos, administración de Big Data, inteligencia comercial y otras utilidades de administración de Big Data.

Una plataforma de análisis de datos grandes ayuda a extraer el valor de los datos. Los datos solo son útiles cuando se pueden derivar resultados comerciales beneficiosos, y para extraer los objetos valiosos de los datos, se deben adoptar las medidas adecuadas. [22]

Algunas herramientas que nos permiten una buena práctica de Big Data son:

- **HADOOP:** La biblioteca Hadoop ofrece un framework que utiliza modelos de programación simples para el procesamiento distribuido de un gran conjunto de datos a través de varias máquinas conectadas. Esta herramienta permite el procesamiento de grandes volúmenes de datos a través de clusters, usando un modelo simple de programación. Además su diseño permite pasar de pocos nodos a miles de nodos de forma ágil. Fue diseñado para superar fallos y errores en la capa de aplicaciones, proporcionando de este modo una alta precisión. Hadoop es reconocida por su enorme capacidad para analizar y gestionar datos de dimensiones inconmensurables, brindando información final de gran utilidad y en perfecta organización, además es la herramienta utilizada por gigantes de Internet como Yahoo! y Facebook. [23]
- **SPARK:** Apache Spark es un motor de código abierto desarrollado específicamente para el procesamiento y análisis de datos a gran escala. Spark ofrece la capacidad de acceder a los datos en una variedad de fuentes, incluido Hadoop Distributed File System (HDFS), OpenStack Swift, Cassandra entre otros. Apache Spark está diseñado para acelerar los análisis en Hadoop y ofrece un conjunto completo de herramientas complementarias que incluyen una biblioteca de aprendizaje automático

con todas las funciones (MLlib), un motor de procesamiento de gráficos (GraphX) y procesamiento de flujo.[25]

## **6. Instalación de un ambiente de Big Data y desarrollo de un caso practico**

El objetivo de este capítulo es elaborar una serie de procedimientos paso a paso que describa el proceso de instalación de Hadoop, además de explicar el funcionamiento de cada uno de los servicios con los que cuenta esta herramienta para proporcionar un mejor entendimiento y un fácil aprendizaje de lo que es y lo que hace Hadoop.

### **6.1 Instalación de un ambiente de Big Data**

Para la creación del ambiente de Big Data se utilizara la herramienta hadoop la cual permite el procesamiento distribuido de grandes volúmenes de datos mediante un cluster. Esta herramienta permite correr todas las aplicaciones dentro del mismo cluster, por lo tanto toda la información quedara alojada allí y podrá ser utilizada sin ningún inconveniente.

Hadoop cuenta con tres modelos de arquitectura de cluster:

**Modo No Distribuido:**El modo no distribuido también es conocido como modo de un solo nodo (single node), el cual se ejecuta como un solo proceso de JAVA y es más utilizado para depuración; single nodo será el mono utilizado para el ejemplo de esta guía.

**Modo Pseudo-distribuido:** El modo pseudo-distribuido es aquel en el cual un único nodo es configurado para trabajar como una simulación de una arquitectura distribuida, es ideal para desarrollo y probar aplicaciones.

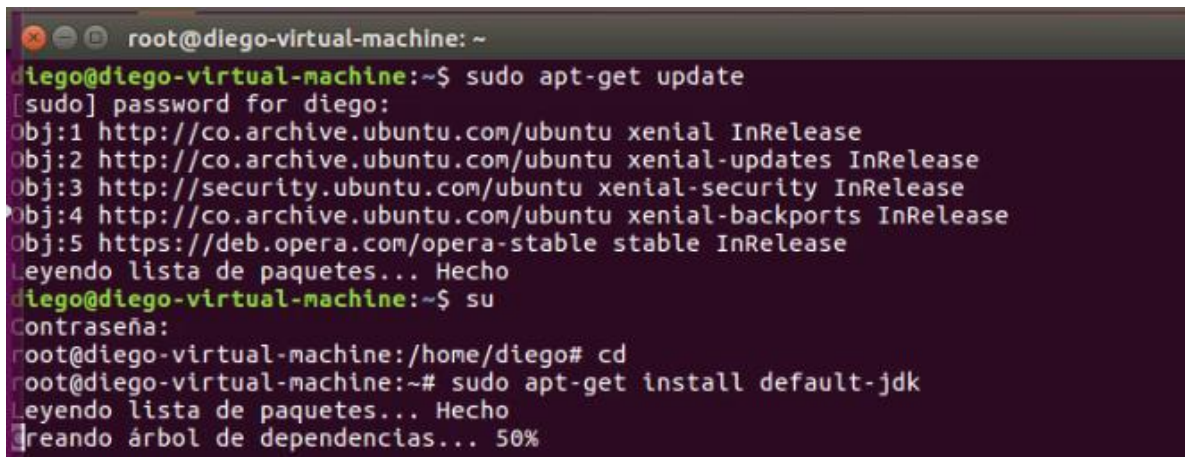
**Modo Completamente Distribuido:** El modo completamente distribuido es aquel en el cual un clúster se configura como una arquitectura distribuida con todos los servicios maestro-esclavos funcionando y es apropiado para un entorno de producción.

## Pasos para instalar hadoop

Antes de empezar con la instalación de Hadoop se debe tener instalado el JDK de Java el cual brinda una serie de herramientas de Desarrollo para la creación de programas en Java. En caso de no tenerlo instalado en el equipo, se puede instalar ejecutando los siguientes comandos:

```
sudo apt-get update
```

```
sudo apt-get install default-jdk
```

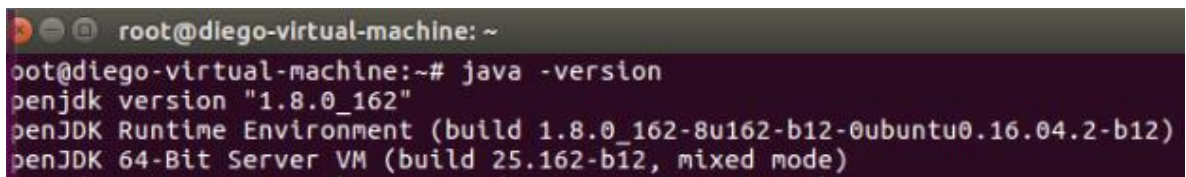


```
root@diego-virtual-machine: ~
diego@diego-virtual-machine:~$ sudo apt-get update
[sudo] password for diego:
Obj:1 http://co.archive.ubuntu.com/ubuntu xenial InRelease
Obj:2 http://co.archive.ubuntu.com/ubuntu xenial-updates InRelease
Obj:3 http://security.ubuntu.com/ubuntu xenial-security InRelease
Obj:4 http://co.archive.ubuntu.com/ubuntu xenial-backports InRelease
Obj:5 https://deb.opera.com/opera-stable stable InRelease
Leyendo lista de paquetes... Hecho
diego@diego-virtual-machine:~$ su
Contraseña:
root@diego-virtual-machine:/home/diego# cd
root@diego-virtual-machine:~# sudo apt-get install default-jdk
Leyendo lista de paquetes... Hecho
Creando árbol de dependencias... 50%
```

Ilustración 1 install jdk Fuente: autor

Comprobamos que se halla instalado correctamente o si ya está instalado verificar que versión corre en el equipo, ejecutando el siguiente comando:

```
java -versión
```

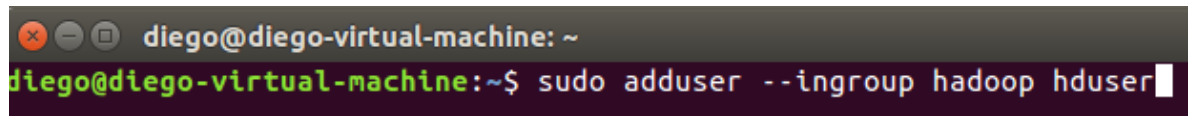


```
root@diego-virtual-machine: ~
root@diego-virtual-machine:~# java -version
openjdk version "1.8.0_162"
openJDK Runtime Environment (build 1.8.0_162-8u162-b12-0ubuntu0.16.04.2-b12)
openJDK 64-Bit Server VM (build 25.162-b12, mixed mode)
```

Ilustración 2 version java Fuente: autor

Para administrar los permisos de usuarios de una forma más flexible se debe agregar un usuario dedicado a hadoop con el siguiente comando:

```
sudo adduser --ingroup hadoop hduser
```

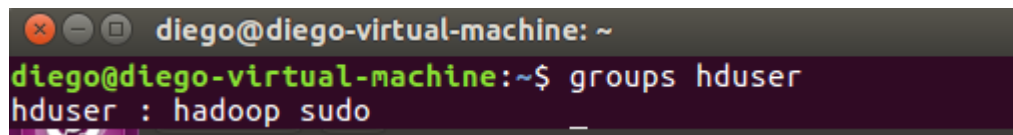


```
diego@diego-virtual-machine: ~  
diego@diego-virtual-machine:~$ sudo adduser --ingroup hadoop hduser
```

Ilustración 3 crear grupo fuente: autor

Verificar que el grupo hadoop y el usuario hduser se hayan creado correctamente

```
groups hduser
```



```
diego@diego-virtual-machine: ~  
diego@diego-virtual-machine:~$ groups hduser  
hduser : hadoop sudo
```

Ilustración 4 verificar grupo fuente: autor

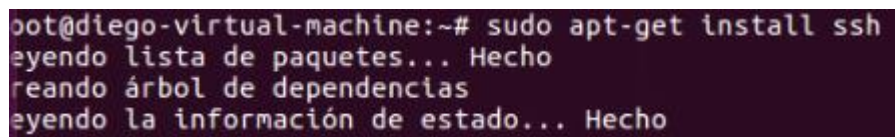
### Configurar el acceso SSH

Para realizar la gestión de los nodos, necesitamos acceso ssh. Para ello, hay que hacer la configuración del acceso ssh antes de instalar Apache Hadoop.

En caso de no tener instalado ssh, bastará con ejecutar los siguientes comandos.

```
sudo apt-get update
```

```
sudo apt-get install ssh
```



```
bot@diego-virtual-machine:~# sudo apt-get install ssh  
leyendo lista de paquetes... Hecho  
reando árbol de dependencias  
leyendo la información de estado... Hecho
```

Ilustración 5 install ssh fuente:autor

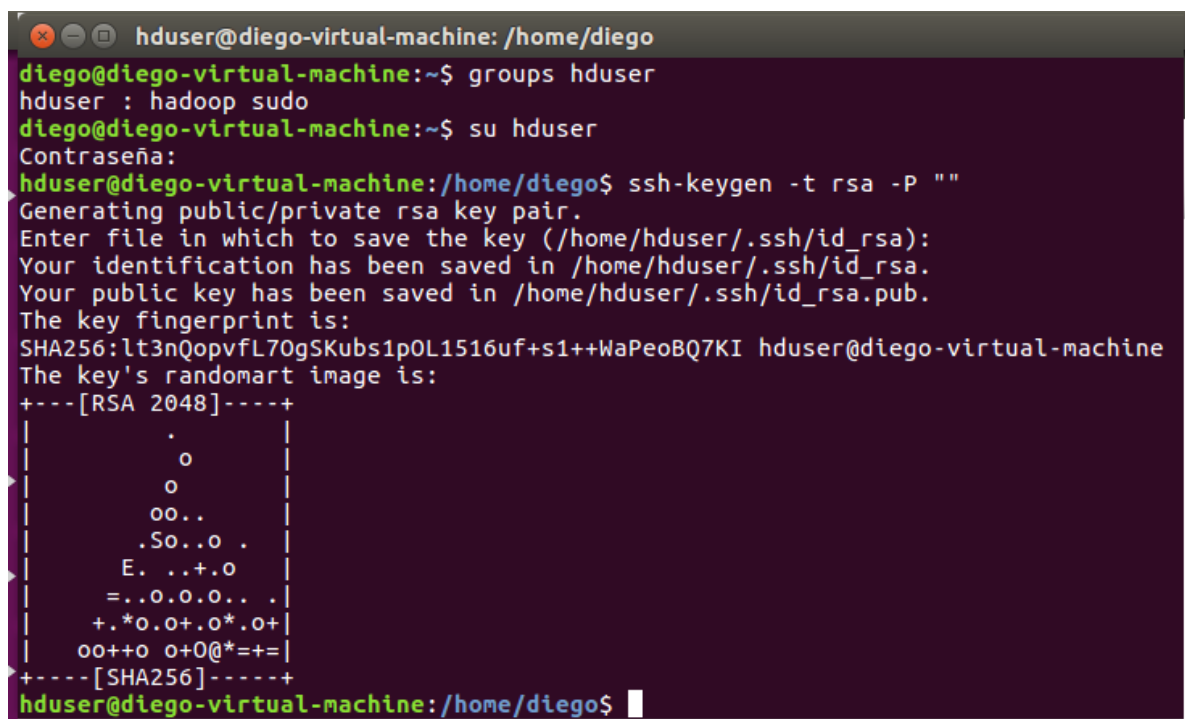
Para configurar el acceso ssh del usuario hduser a localhost, lo primero que hay que hacer es conectarse al sistema como el usuario hduser.

```
su - hduser
```

A continuación, hay que generar una clave SSH con contraseña vacía para el usuario hduser.

```
ssh-keygen -t rsa -P ""
```

El sistema preguntará por la ruta y el nombre del fichero que contendrá la clave ssh, a lo que se le puede indicar: /home/hduser/.ssh/id\_rsa



```
hduser@diego-virtual-machine: /home/diego
diego@diego-virtual-machine:~$ groups hduser
hduser : hadoop sudo
diego@diego-virtual-machine:~$ su hduser
Contraseña:
hduser@diego-virtual-machine:~/home/diego$ ssh-keygen -t rsa -P ""
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hduser/.ssh/id_rsa):
Your identification has been saved in /home/hduser/.ssh/id_rsa.
Your public key has been saved in /home/hduser/.ssh/id_rsa.pub.
The key fingerprint is:
SHA256:lt3nQopvfl70gSKubs1p0L1516uf+s1++WaPeoBQ7KI hduser@diego-virtual-machine
The key's randomart image is:
+---[RSA 2048]-----+
|
|   o
|  o
| oo..
| .So..o .
|  E. .+.o
| =..o.o.o.. .
| +.*o.o+.o*.o+
| oo++o o+o@*+=+
+-----[SHA256]-----+
hduser@diego-virtual-machine:~/home/diego$
```

Ilustración 6 configure ssh fuente: autor

El hecho de crear la clave ssh sin contraseña es necesario para poder desbloquear la clave sin necesidad de interacción por parte de un usuario cada vez que se conecta a un nodo.

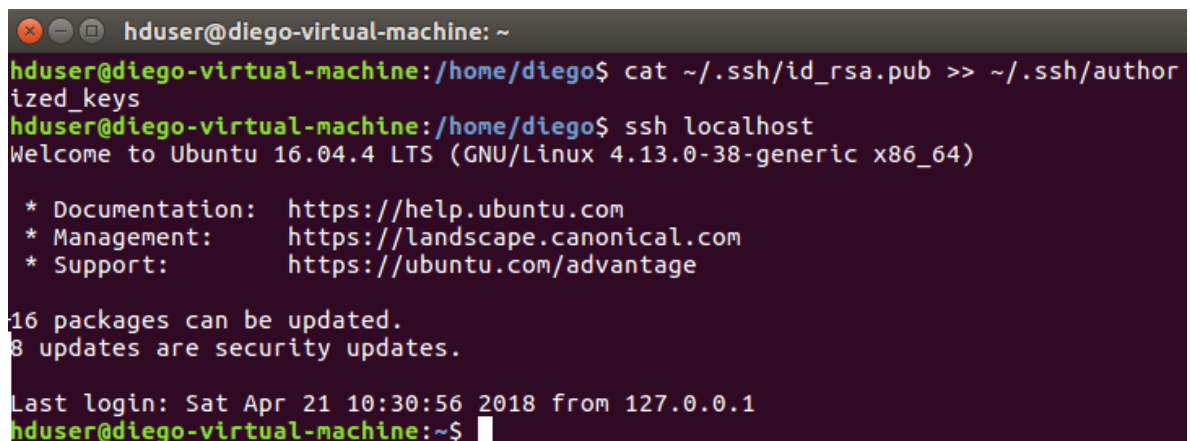
Una vez creada la clave ssh, se debe habilitar el acceso SSH a la máquina local utilizando dicha clave.

```
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
```

Una vez realizado los pasos anteriores se comprueba que se ha configurado correctamente el acceso SSH, y para ello se debe conectar a la máquina local con el comando

```
ssh localhost
```

y debe de mostrar una respuesta como lo vemos en la siguiente imagen

A terminal window titled 'hduser@diego-virtual-machine: ~' showing the execution of 'cat ~/.ssh/id\_rsa.pub >> ~/.ssh/authorized\_keys' and 'ssh localhost'. The output of the ssh command shows the Ubuntu 16.04.4 LTS login screen, including system information, update notifications, and the last login time.

```
hduser@diego-virtual-machine:~/home/diego$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
hduser@diego-virtual-machine:~/home/diego$ ssh localhost
Welcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.13.0-38-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:    https://landscape.canonical.com
 * Support:       https://ubuntu.com/advantage

16 packages can be updated.
8 updates are security updates.

Last login: Sat Apr 21 10:30:56 2018 from 127.0.0.1
hduser@diego-virtual-machine:~$
```

Ilustración 7 acceso ssh fuente: autor

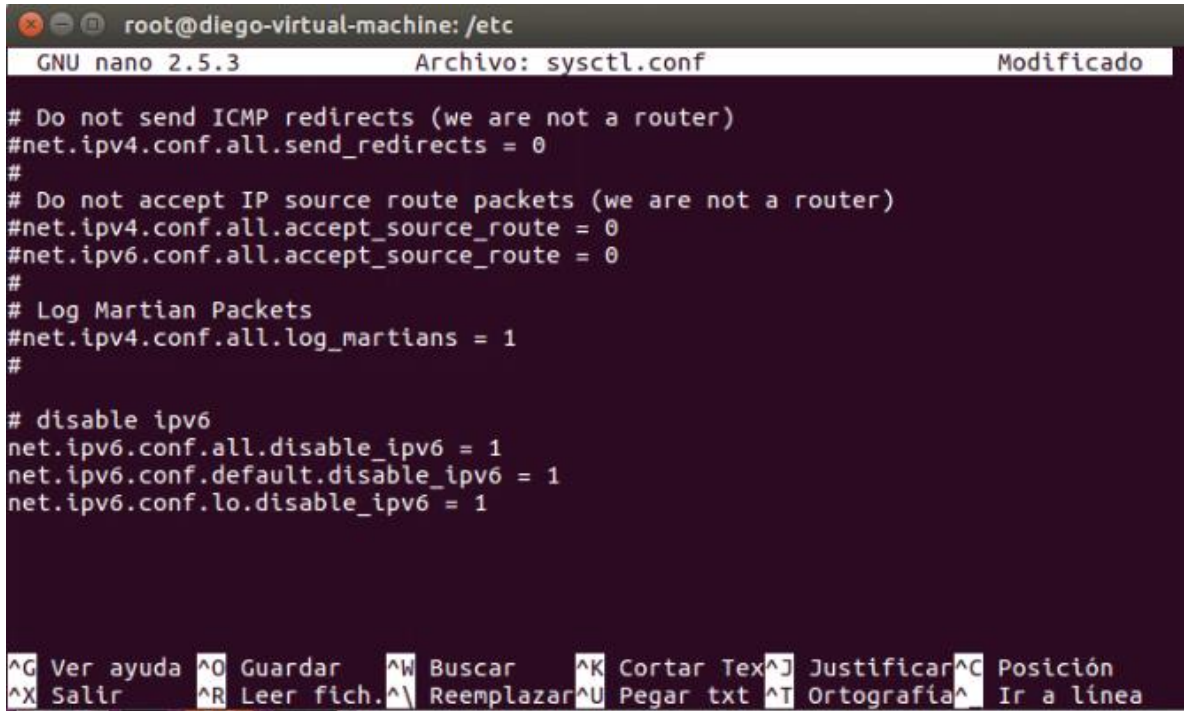
## Desactivar IPv6

Uno de los problemas de utilizar IPv6 en Ubuntu es que utilizando la dirección IP 0.0.0.0 en la configuración de varios Hadoop relacionados provoca que Hadoop mapee las direcciones IPv6. Como no es necesario tener esto activo a menos que esté conectado a una red IPv6, conviene desactivarlos antes de instalar Apache Hadoop.

Para desactivar IPv6 en Ubuntu Linux 14.04 LTS basta con abrir el fichero /etc/sysctl.conf y añadir las siguientes líneas al final del fichero.



```
# disable ipv6
net.ipv6.conf.all.disable_ipv6 = 1
net.ipv6.conf.default.disable_ipv6 = 1
net.ipv6.conf.lo.disable_ipv6 = 1
```



```
root@diego-virtual-machine: /etc
GNU nano 2.5.3          Archivo: sysctl.conf          Modificado

# Do not send ICMP redirects (we are not a router)
#net.ipv4.conf.all.send_redirects = 0
#
# Do not accept IP source route packets (we are not a router)
#net.ipv4.conf.all.accept_source_route = 0
#net.ipv6.conf.all.accept_source_route = 0
#
# Log Martian Packets
#net.ipv4.conf.all.log_martians = 1
#

# disable ipv6
net.ipv6.conf.all.disable_ipv6 = 1
net.ipv6.conf.default.disable_ipv6 = 1
net.ipv6.conf.lo.disable_ipv6 = 1

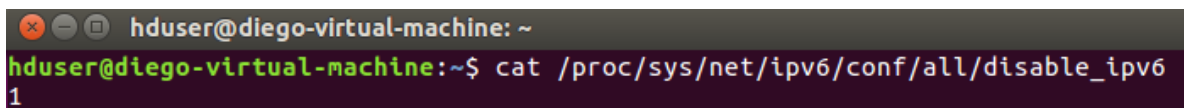
^G Ver ayuda  ^O Guardar   ^W Buscar   ^K Cortar Tex^J Justificar^C Posición
^X Salir      ^R Leer fich.^_ Reemplazar^U Pegar txt  ^T Ortografía^_ Ir a línea
```

Ilustración 8 ipv6 fuente: autor

Para que los cambios se apliquen, se debe reiniciar el sistema. Una vez reiniciado, se puede comprobar si se ha desactivado correctamente IPv6 ejecutando el siguiente comando.

```
cat /proc/sys/net/ipv6/conf/all/disable_ipv6
```

que retorna los valores: 0 si IPv6 está activo, o 1 si está desactivado



```
hduser@diego-virtual-machine: ~
hduser@diego-virtual-machine:~$ cat /proc/sys/net/ipv6/conf/all/disable_ipv6
1
```

Ilustración 9 ipv6 confi fuente: autor

## Instalación y configuración de hadoop

Descargamos la versión deseada de hadoop desde la página de apache hadoop, para este caso utilizaremos la versión 2.6.0 y la descargamos del siguiente enlace

<https://archive.apache.org/dist/hadoop/core/hadoop-2.6.0/> 2014-11-30  
[26]

Descomprimir el fichero tar que se descarga del link anterior. Para ello, ejecutar el siguiente comando desde un terminal:

```
tar -xzf hadoop-2.6.0.tar.gz
```

Para realizar el proceso de una manera más cómoda movemos la instalación de Hadoop al directorio /usr/local/hadoop. Entonces, debemos primero crear el directorio y lo hacemos con el comando:

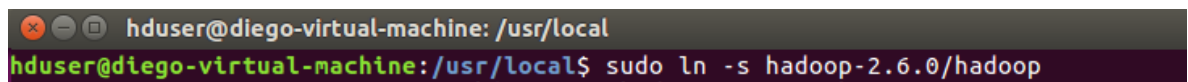
```
sudo mkdir -p /usr/local/hadoop
```

Y movemos el archivo a esta nueva carpeta

```
sudo mv hadoop-2.6.0/ /usr/local/hadoop
```

Ahora creamos un enlace simbólico apuntando a la carpeta que contiene los ficheros de hadoop:

```
sudo ln -s hadoop-2.6.0/ hadoop
```

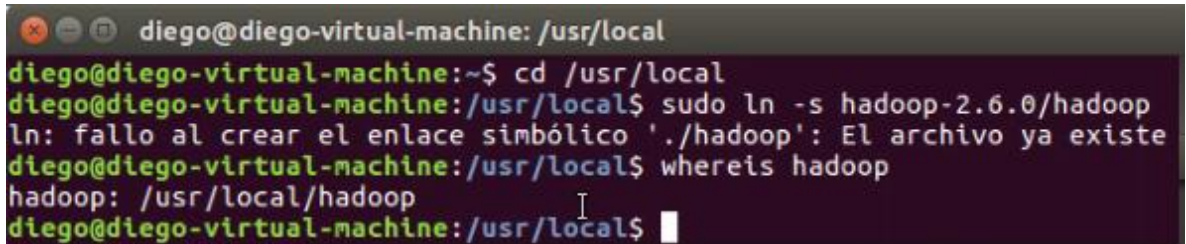
A screenshot of a terminal window with a dark background. The prompt is 'hduser@diego-virtual-machine: /usr/local'. The command being entered is 'sudo ln -s hadoop-2.6.0/hadoop'. The output of the command is not visible.

```
hduser@diego-virtual-machine: /usr/local
hduser@diego-virtual-machine: /usr/local$ sudo ln -s hadoop-2.6.0/hadoop
```

Ilustración 10 enlace simbólico fuente: autor

De esta forma se puede hacer referencia directamente a hadoop sin necesidad de tener que hacer referencia a la versión que está utilizando.

Comprobar donde está apuntando el enlace simbólico con el comando  
whereis hadoop

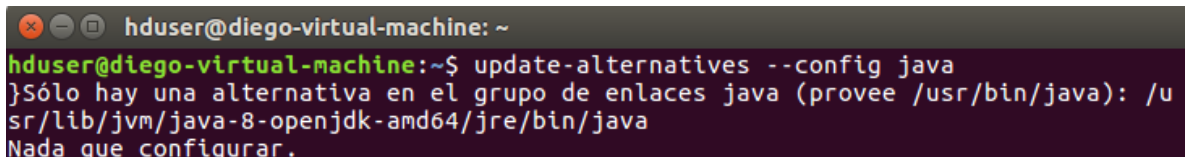


```
diego@diego-virtual-machine: /usr/local
diego@diego-virtual-machine:~$ cd /usr/local
diego@diego-virtual-machine:/usr/local$ sudo ln -s hadoop-2.6.0/hadoop
ln: fallo al crear el enlace simbólico './hadoop': El archivo ya existe
diego@diego-virtual-machine:/usr/local$ whereis hadoop
hadoop: /usr/local/hadoop
diego@diego-virtual-machine:/usr/local$
```

Ilustración 11 whereis fuente: autor

A continuación configurar algunos archivos de hadoop para que funcione correctamente para esto debemos saber la ubicación donde se encuentra instalado java JDK y así poder configurar la variable de entorno JAVA\_HOME, para esto se utiliza el siguiente comando:

update-alternatives --config java



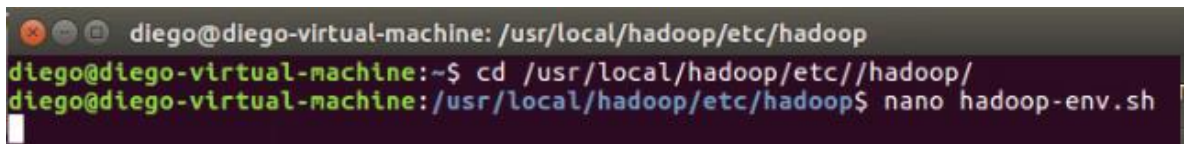
```
hduser@diego-virtual-machine: ~
hduser@diego-virtual-machine:~$ update-alternatives --config java
}Sólo hay una alternativa en el grupo de enlaces java (provee /usr/bin/java): /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/java
Nada que configurar.
```

Ilustración 12 java fuente: autor

Configuración de hadoop

hadoop-env.sh

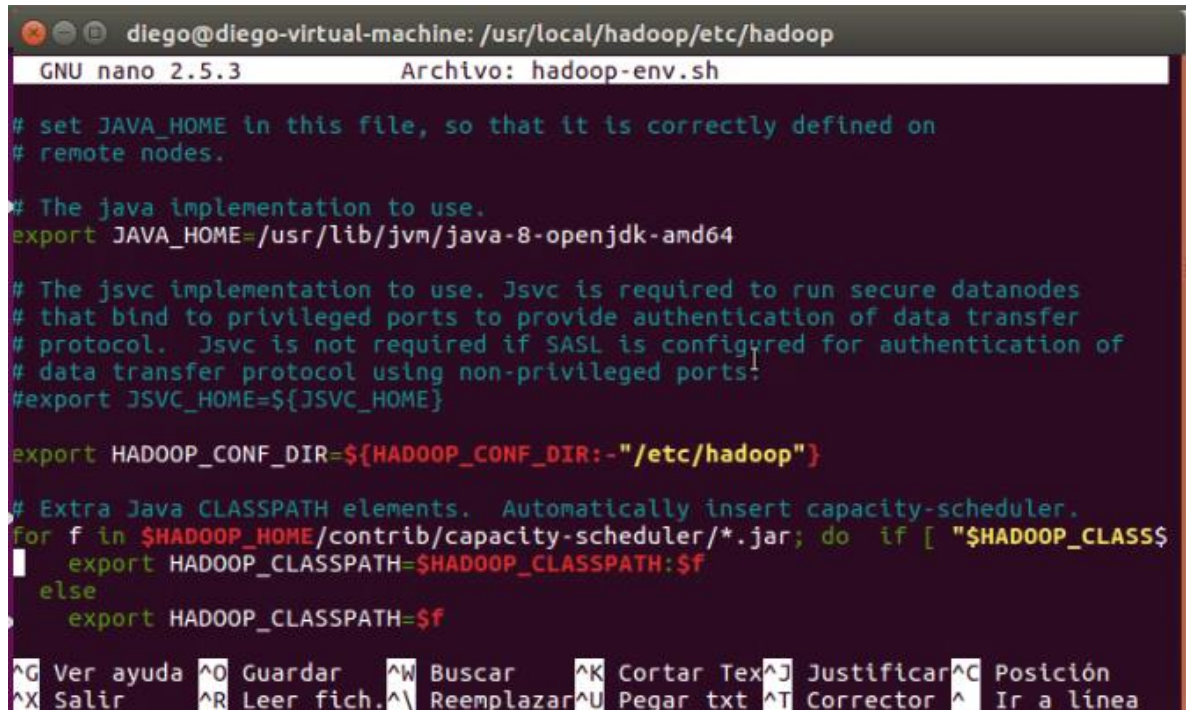
Con el editor de texto abrir el archivo hadoop-env.sh El cual establece las variables de entorno específicas de Hadoop que se encuentra en la dirección hadoop/etc/hadoop



```
diego@diego-virtual-machine: /usr/local/hadoop/etc/hadoop
diego@diego-virtual-machine:~$ cd /usr/local/hadoop/etc/hadoop/
diego@diego-virtual-machine:/usr/local/hadoop/etc/hadoop$ nano hadoop-env.sh
```

Ilustración 13 hadoop-env fuente: autor

Para configurar la variable de entorno de hadoop se quita el comentario en la dirección del JAVA HOME, y se cambia la dirección que se encuentra allí por la dirección donde se encuentra instalado el java JDK.



```
diego@diego-virtual-machine: /usr/local/hadoop/etc/hadoop
GNU nano 2.5.3 Archivo: hadoop-env.sh

# set JAVA_HOME in this file, so that it is correctly defined on
# remote nodes.

# The java implementation to use.
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

# The jsvc implementation to use. Jsvc is required to run secure datanodes
# that bind to privileged ports to provide authentication of data transfer
# protocol. Jsvc is not required if SASL is configured for authentication of
# data transfer protocol using non-privileged ports!
#export JSVC_HOME=${JSVC_HOME}

export HADOOP_CONF_DIR=${HADOOP_CONF_DIR:-"/etc/hadoop"}

# Extra Java CLASSPATH elements. Automatically insert capacity-scheduler.
for f in $HADOOP_HOME/contrib/capacity-scheduler/*.jar; do if [ "$HADOOP_CLASS$
| export HADOOP_CLASSPATH=$HADOOP_CLASSPATH:$f
| else
| export HADOOP_CLASSPATH=$f
|

^G Ver ayuda ^O Guardar ^W Buscar ^K Cortar Text ^J Justificar ^C Posición
^X Salir ^R Leer fich. ^\ Reemplazar ^U Pegar txt ^T Corrector ^_ Ir a línea
```

Ilustración 14 hadoop-env2 fuente: autor

Guardar y salir.

Crear las variables HADOOP\_HOME en el path para que el sistema operativo pueda encontrar lo ejecutables necesario, para crear esta variable se utiliza los siguientes comandos:

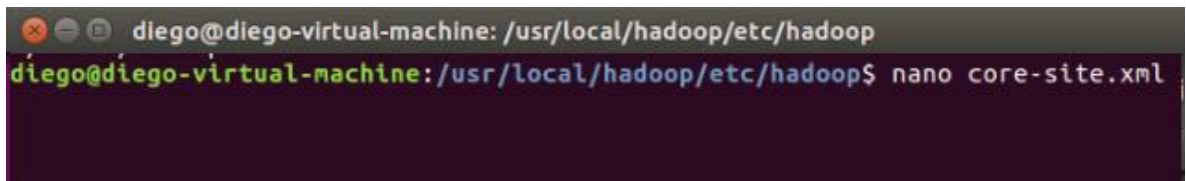
```
$ export HADOOP_HOME=/usr/local/hadoop/bin
```

```
$ export HADOOP_IN_PATH=/usr/local/hadoop/bin
```

core-site.xml

En este archivo de hadoop se pueden indicar numerosas opciones de configuración, para este ejemplo se configura el directorio HDFS por defecto en el localhost.

Abrir el archivo core-site.xml que se encuentra en la dirección /usr/local/hadoop/etc/hadoop/



```
diego@diego-virtual-machine: /usr/local/hadoop/etc/hadoop
diego@diego-virtual-machine: /usr/local/hadoop/etc/hadoop$ nano core-site.xml
```

Ilustración 15 core-site fuente: autor

Insertar las siguientes líneas dentro de las etiquetas

<configuration> </configuration>.

<property>

<name>hadoop.tmp.dir</name>

<value>/app/hadoop/tmp</value>

</property>

“base para otros directorios temporales”

<property>

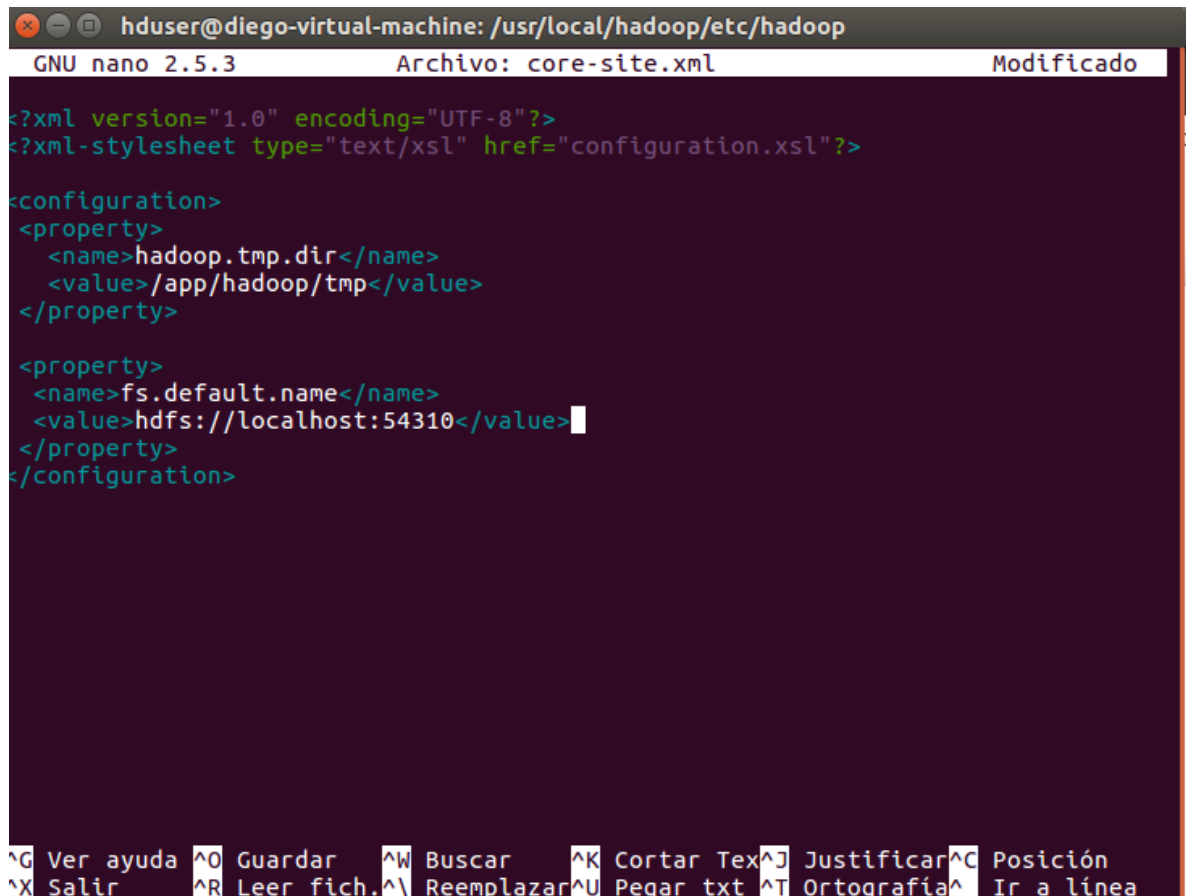
<name>fs.default.name</name>

<value>hdfs://localhost:54310</value>

</property>

“Nombre del sistema de archivos predeterminado. Una URL cuyo esquema y autoridad determinan la implementación del sistema de archivos. El esquema de la URL determina la propiedad config (fs.scheme.impl) nombrando la clase de

implementación del sistema de archivos. la autoridad de URL se usa para determinar el host, el puerto, etc. para un sistema de archivos”



```
hduser@diego-virtual-machine: /usr/local/hadoop/etc/hadoop
GNU nano 2.5.3 Archivo: core-site.xml Modificado

<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>
  <property>
    <name>hadoop.tmp.dir</name>
    <value>/app/hadoop/tmp</value>
  </property>

  <property>
    <name>fs.default.name</name>
    <value>hdfs://localhost:54310</value>
  </property>
</configuration>
```

Ilustración 16 coresite2 fuente: autor

Guardar y salir del archivo

hdfs-site.xml

Este archivo contiene información sobre como Hadoop almacenará la información en el clúster

Antes de editar este archivo, se debe crear dos directorios que contendrán el namenode y el nodo de datos para esta instalación de Hadoop.

Este archivo debe configurarse para cada host en el clúster que se está utilizando.

Especificar los directorios que se usarán como el namenode y el nodo de datos en ese host.

Esto se puede hacer usando los siguientes comandos:

```
hduser@laptop:~$ sudo mkdir -p /usr/local/hadoop_store/hdfs/namenode
hduser@laptop:~$ sudo mkdir -p /usr/local/hadoop_store/hdfs/datanode
hduser@laptop:~$ sudo chown -R hduser:hadoop /usr/local/hadoop_store
```

Abrir el archivo hdfs-site.xml que se encuentra en la dirección /hadoop/etc/hadoop y agregamos la siguiente configuración:

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
```

“Factor de replicación. Como se tiene una única computadora en el clúster lo ponemos a 1”

```
<property>
  <name>dfs.namenode.name.dir</name>
  <value>file:/home/hadoop/workspace/dfs/name</value>
</property>
```

“Ruta del sistema de archivos donde el NameNode almacenará los metadatos”

```
<property>
  <name>dfs.datanode.data.dir</name>
  <value>file:/home/hadoop/workspace/dfs/data</value>
</property>
```

```
</configuration>
```

Ruta del sistema de archivos donde el DataNode almacenara los bloques

```
hduser@diego-virtual-machine: /usr/local/hadoop/etc/hadoop
GNU nano 2.5.3 Archivo: hdfs-site.xml

-->
<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
    <description>Default block replication.
    The actual number of replications can be specified when the file is created.
    The default is used if replication is not specified in create time.
    </description>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>file:/usr/local/hadoop_store/hdfs/namenode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>file:/usr/local/hadoop_store/hdfs/datanode</value>
  </property>
</configuration>

^G Ver ayuda ^O Guardar ^W Buscar ^K Cortar Tex ^J Justificar ^C Posición
^X Salir ^R Leer fich. ^\ Reemplazar ^U Pegar txt ^T Ortografía ^_ Ir a línea
```

Ilustración 17 hdfs fuente: autor

Guardar y salir del archivo

Mapred-site.xml

Se utiliza para especificar quien realiza el MapReduce y el lugar donde se lleva a cabo. Como solo se tiene un único nodo en nuestro clúster, solo habrá una Job Map y otro Reduce.

Antes de editar el archivo, se debe renombrarlo. Dicho archivo se encuentra en la dirección /hadoop/etc/hadoop El nombre por defecto es mapred-site.xml.template así que es necesario crear un nuevo archivo con el nombre mapred-site.xml.

Insertar las siguientes líneas entre las etiquetas <configuration> </configuration>:



```

<property>
  <name>mapreduce.framework.name</name>
  <value>yarn</value>
</property>
<property>
  <name>mapred.job.tracker</name>
  <value>localhost:54311</value>
</property>

```

Guardar como un nuevo archivo y lo nombrarlo mapred-site.xml

```

diego@diego-virtual-machine: /usr/local/hadoop/etc/hadoop
GNU nano 2.5.3 Archivo: mapred-site.xml Modificado
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->
<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>mapred.job.tracker</name>
    <value>localhost:54311</value>
  </property>
</configuration>

```

Ilustración 18 mapresite fuente: autor

Archivo .bashrc

Abrir el archivo ./bashrc con el editor y agregar las siguientes líneas al final

```

#HADOOP VARIABLES START
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export HADOOP_INSTALL=/usr/local/hadoop
export PATH=$PATH:$HADOOP_INSTALL/bin

```

```

export PATH=$PATH:$HADOOP_INSTALL/sbin
export HADOOP_MAPRED_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_HOME=$HADOOP_INSTALL
export HADOOP_HDFS_HOME=$HADOOP_INSTALL
export YARN_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_INSTALL/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_INSTALL/lib"
#HADOOP VARIABLES END

```

```

hduser@diego-virtual-machine: ~
hduser@diego-virtual-machine:~/home$ cd
hduser@diego-virtual-machine:~$ nano .bashrc

```

Ilustración 19 bashrc fuente: autor

```

GNU nano 2.5.3 Archivo: .bashrc

if ! shopt -oq posix; then
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi

#HADOOP VARIABLES START
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export HADOOP_INSTALL=/usr/local/hadoop
export PATH=$PATH:$HADOOP_INSTALL/bin
export PATH=$PATH:$HADOOP_INSTALL/sbin
export HADOOP_MAPRED_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_HOME=$HADOOP_INSTALL
export HADOOP_HDFS_HOME=$HADOOP_INSTALL
export YARN_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_INSTALL/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_INSTALL/lib"
#HADOOP VARIABLES END

^G Ver ayuda ^O Guardar ^W Buscar ^K Cortar Tex ^J Justificar ^C Posición
^X Salir ^R Leer fich. ^\ Reemplazar ^U Pegar txt ^T Ortografía ^_ Ir a línea

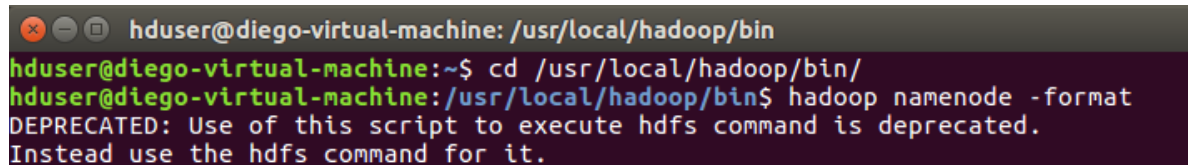
```

Ilustración 20 bashrc2 fuente: autor

Terminada la configuración del archivo .bashrc se debe formatear el namenodo que utiliza hadoop para su aplicaion y esta ejecución se realiza en el nuevo sistema de archivos hadoop.

Antes de ejecutar el comando de formateo se debe ingresar al sistema operativo como usuario root, y se ejecuta el siguiente comando en la carpeta bin de hadoop para formatear el nodo

```
hadoop namenode -format
```



```
hduser@diego-virtual-machine: /usr/local/hadoop/bin
hduser@diego-virtual-machine:~$ cd /usr/local/hadoop/bin/
hduser@diego-virtual-machine:/usr/local/hadoop/bin$ hadoop namenode -format
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
```

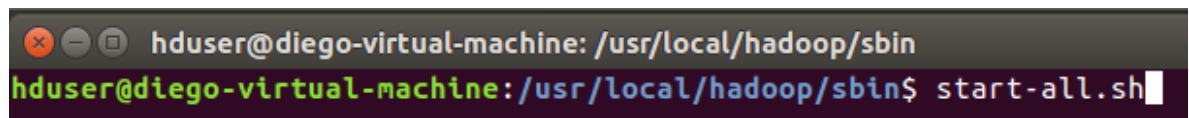
Ilustración 21 format namenode fuente: autor

El comando `hadoop namenode -format` solo debe ejecutarse cuando se inicia un nuevo proyecto de hadoop, ya que este comando destruye todos los datos en el sistema de archivos de hadoop.

### Inicializando Hadoop

Ahora es el momento de iniciar el clúster de nodo único recién instalado.

Para inicialización de hadoop se debe acceder desde el terminal a la carpeta `sbin` que se encuentra en la dirección `/usr/local/hadoop/sbin` y se ejecuta el comando `start-all.sh` para inicializar todos los nodos



```
hduser@diego-virtual-machine: /usr/local/hadoop/sbin
hduser@diego-virtual-machine:/usr/local/hadoop/sbin$ start-all.sh
```

Verificamos que si está funcionando de la manera correcta, con el comando `jps` podemos obtener el PID del proceso y la clase en ejecución

```
hduser@diego-virtual-machine: /usr/local/hadoop/sbin
hduser@diego-virtual-machine: /usr/local/hadoop/sbin$ jps
11137 ResourceManager
11249 NodeManager
10677 NameNode
10981 SecondaryNameNode
12159 Jps
hduser@diego-virtual-machine: /usr/local/hadoop/sbin$
```

La salida significa que ahora existe una instancia funcional de Hadoop ejecutándose en el servidor privado virtual.

Ahora se verifica la conexión accediendo de hadoop a la dirección

<http://localhost:50070> [27]

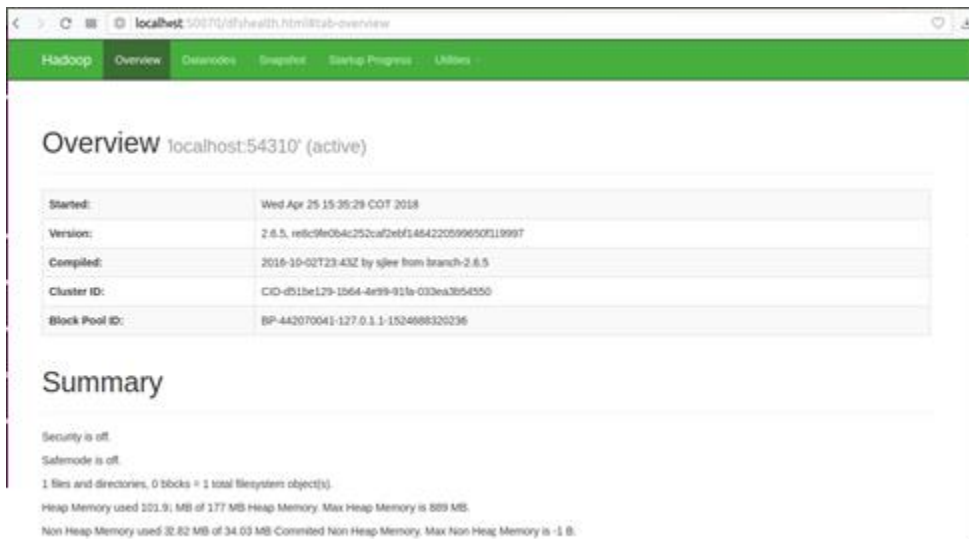


Ilustración 22 localhost 50070 fuente: autor

Acceder a <http://localhost:50090/status.jsp> [28] para verificar que el nodo secundario este activo.



Ilustración 23 localhost 50090 fuente: autor

## 6.2 Caso practico

El termino de Big data engloba una gran cantidad de tecnologías técnicas, y de algoritmos; pero también hacen parte de él componentes visuales que nos permiten plasmar grandes cantidades de datos, o indicadores elaborados y servicios de almacenamiento avanzados que tienen un formato distinto a las relaciones tabulares de las bases de datos SQL. Las bases de datos NoSQL son estructuras que permiten almacenar información en aquellas situaciones en las que las bases de datos relacionales generan ciertos problemas debido principalmente a problemas de escalabilidad y rendimiento.

En este caso en particular se realizará la práctica de recolección de Tweets como una demostración de lo que es un ambiente de Big Data, es decir, la recolección de grandes volúmenes de datos, y el almacenamiento en una base de datos NOSQL. Más adelante se explica paso por paso como es el proceso de instalación y configuración de las herramientas utilizadas para este ejemplo.

## Twitter

Twitter es una de las redes sociales más importantes en la actualidad, en el 2017 twitter registro que por cada día se generaban 500 millones de tuits, es decir 6000 tuist por segundo, para este caso práctico se utilizara esta red social recolectando tweets de una determinada zona geográfica y almacenándolos en una base de datos Nosql.[34]

Twitter es una red de información conformada por 280 caracteres llamados tweets, los tweets están conformados por texto, hashtags, @”nombres de usuarios” o urls.

Esta red social ofrece APIs, que permiten a los desarrolladores adaptarse a diferentes necesidades. Por ejemplo, existe el Streaming API que permite el acceso en tiempo real a los Tweets que han sido publicados; el Rest API permite acceder al núcleo central donde se encuentran los datos de Twitter y, el Search API el cual ofrece una información del acceso a los datos del autor como el id, el nombre del usuario con el que aparece en Twitter.

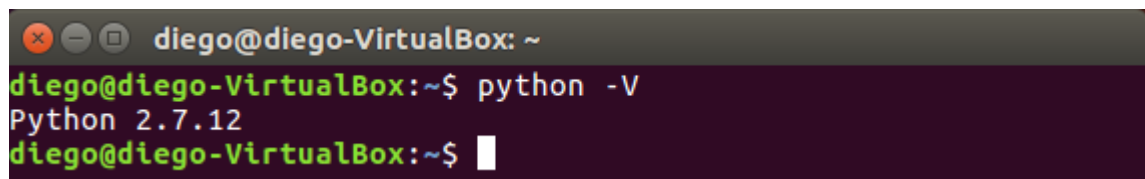
A continuación se muestran los pasos a seguir para realizar esta práctica:

- **Python**

Para realizar este ejercicio se necesita tener instalado Python en la máquina, la cadena de distribución GNU/LINUX tiene instalado por defecto esta herramienta.

Verificamos la versión de Python con el comando

```
python -V
```



```
diego@diego-VirtualBox: ~  
diego@diego-VirtualBox:~$ python -V  
Python 2.7.12  
diego@diego-VirtualBox:~$
```

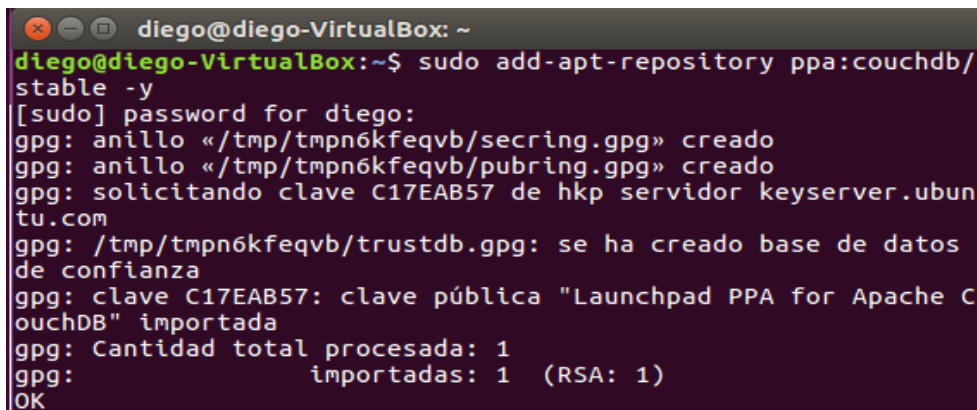
- **CouchBD**

Apache CouchDB es una base de datos NoSQL. Utiliza JSON para almacenar datos, JavaScript como lenguaje de consulta y MapReduce y HTTP como API. Permite la creación de vistas, que son el mecanismo que permite la combinación

de documentos para retornar valores de varios documentos, es decir, CouchDB permite la realización de las operaciones JOIN típicas de SQL. [29]

Se requiere tener instalado CouchDB, y para ello se utiliza los repositorios ppa que permiten adquirir la versión actualizada de cualquier programa:

```
$sudo add-apt-repository ppa:couchdb/stable -y
```

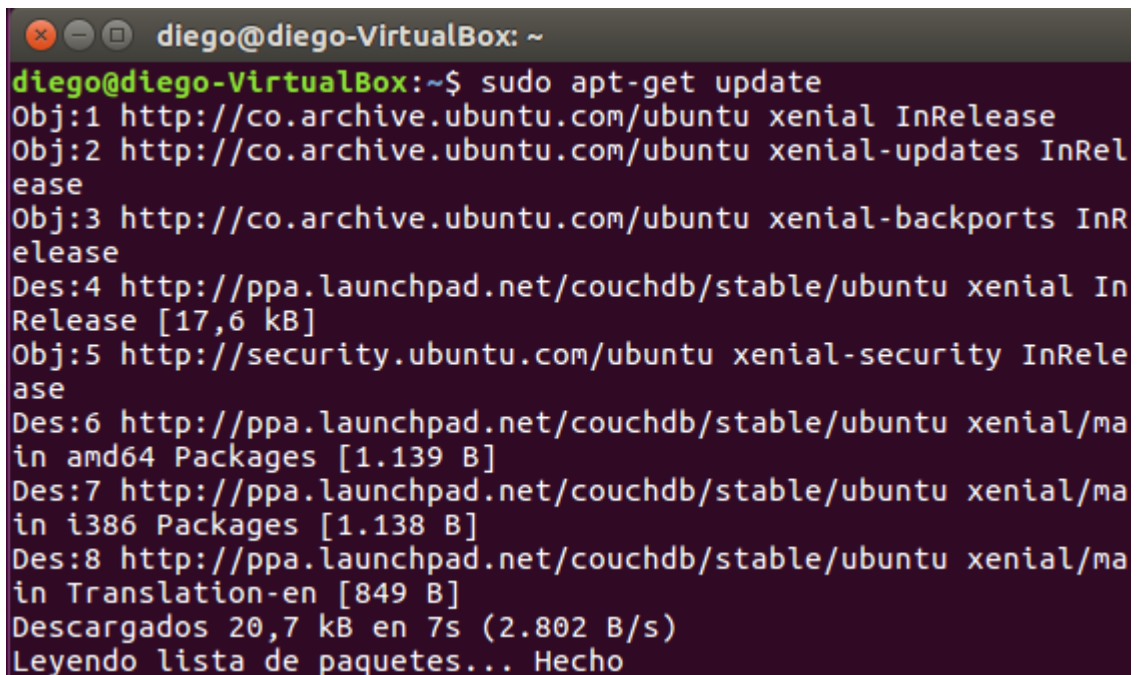


```
diego@diego-VirtualBox: ~
diego@diego-VirtualBox:~$ sudo add-apt-repository ppa:couchdb/
stable -y
[sudo] password for diego:
gpg: anillo «/tmp/tmpn6kfeqvb/secring.gpg» creado
gpg: anillo «/tmp/tmpn6kfeqvb/pubring.gpg» creado
gpg: solicitando clave C17EAB57 de hkp servidor keyserver.ubun
tu.com
gpg: /tmp/tmpn6kfeqvb/trustdb.gpg: se ha creado base de datos
de confianza
gpg: clave C17EAB57: clave pública "Launchpad PPA for Apache C
ouchDB" importada
gpg: Cantidad total procesada: 1
gpg:          importadas: 1 (RSA: 1)
OK
```

Ilustración 24 ppa fuente: autor

Actualizar la lista de paquetes del sistema operativo para cargar actualizaciones instaladas:

```
Sudo apt-get update
```



```
diego@diego-VirtualBox: ~
diego@diego-VirtualBox:~$ sudo apt-get update
Obj:1 http://co.archive.ubuntu.com/ubuntu xenial InRelease
Obj:2 http://co.archive.ubuntu.com/ubuntu xenial-updates InRel
ease
Obj:3 http://co.archive.ubuntu.com/ubuntu xenial-backports InR
elease
Des:4 http://ppa.launchpad.net/couchdb/stable/ubuntu xenial In
Release [17,6 kB]
Obj:5 http://security.ubuntu.com/ubuntu xenial-security InRel
ease
Des:6 http://ppa.launchpad.net/couchdb/stable/ubuntu xenial/ma
in amd64 Packages [1.139 B]
Des:7 http://ppa.launchpad.net/couchdb/stable/ubuntu xenial/ma
in i386 Packages [1.138 B]
Des:8 http://ppa.launchpad.net/couchdb/stable/ubuntu xenial/ma
in Translation-en [849 B]
Descargados 20,7 kB en 7s (2.802 B/s)
Leyendo lista de paquetes... Hecho
```

Ilustración 25 actualizar fuente: autor

Para mejorar el rendimiento de couchBD se debe Eliminar cualquier posible existencia de los binarios de esta base de datos.

```
sudo apt-get remove couchdb couchdb-bin couchdb-common -yf
```

```
diego@diego-VirtualBox:~$ sudo apt-get remove couchdb couchdb-bin couchdb-common -yf
Leyendo lista de paquetes... Hecho
Creando árbol de dependencias
Leyendo la información de estado... Hecho
El paquete «couchdb» no está instalado, no se eliminará
El paquete «couchdb-bin» no está instalado, no se eliminará
El paquete «couchdb-common» no está instalado, no se eliminará
0 actualizados, 0 nuevos se instalarán, 0 para eliminar y 260 no actualizados.
```

#### Ilustración 26 eliminar binarios

Instalar CouchBD

Para instalar couchdb se utiliza el comando

```
$sudo apt-get install -V couchdb
```

```
root@diego-VirtualBox:~# sudo apt-get install -V couchdb
Leyendo lista de paquetes... Hecho
Creando árbol de dependencias
Leyendo la información de estado... Hecho
Se instalarán los siguientes paquetes adicionales:
 couchdb-bin (1.6.1-0ubuntu6ppa2~xenial1)
 couchdb-common (1.6.1-0ubuntu6ppa2~xenial1)
```

#### Ilustración 27 couchbd

Para que la base de datos de CouchBD funcione de la manera correcta es necesario detener el servicio de couchdb y encenderlo de nuevo.

Utilizamos los siguientes comandos:

Detener el servicio de CouchBD:

```
$sudo systemctl stop couchdb
```

```
root@diego-VirtualBox:~# sudo systemctl stop couchdb
```

#### Ilustración 28 detener couchbd



Encender la base de datos de couchbd:

```
$sudo systemctl start couchdb
```

```
root@diego-VirtualBox:~# sudo systemctl start couchdb
```

Ilustración 29 encender couchbd

Cuando se realice el paso anterior se debe comprobar el estado de CouchBD y verificar que se instaló de la manera correcta

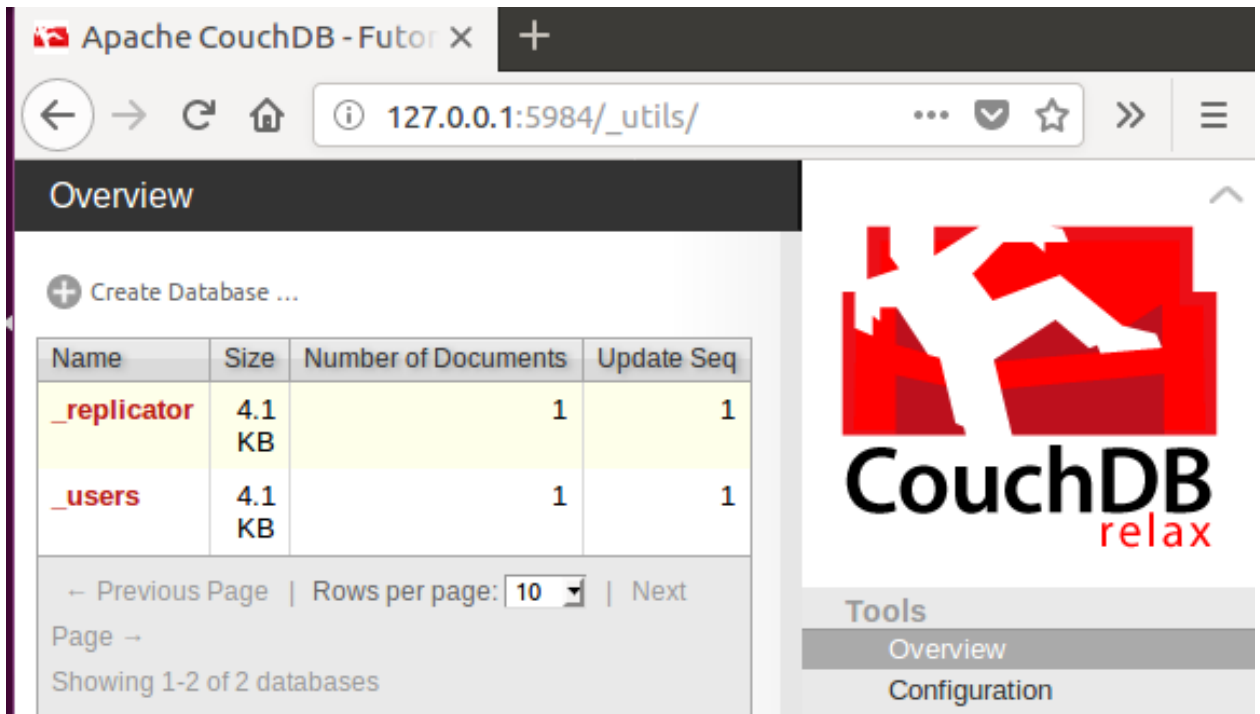
```
$sudo systemctl status couchdb
```

```
root@diego-VirtualBox:~# sudo systemctl status couchdb
● couchdb.service - Apache CouchDB
   Loaded: loaded (/lib/systemd/system/couchdb.service; enable
   Active: active (running) since lun 2018-04-30 20:02:01 -05;
   Main PID: 4446 (beam)
   CGroup: /system.slice/couchdb.service
           └─4446 /usr/lib/erlang/erts-7.3/bin/beam -Bd -K tru
             └─4461 sh -s disksup

abr 30 20:02:01 diego-VirtualBox systemd[1]: Started Apache Co
abr 30 20:02:03 diego-VirtualBox couchdb[4446]: Apache CouchDB
abr 30 20:02:04 diego-VirtualBox couchdb[4446]: Apache CouchDB
abr 30 20:02:04 diego-VirtualBox couchdb[4446]: [info] [<0.33.
...skipping...
● couchdb.service - Apache CouchDB
   Loaded: loaded (/lib/systemd/system/couchdb.service; enable
   Active: active (running) since lun 2018-04-30 20:02:01 -05;
   Main PID: 4446 (beam)
   CGroup: /system.slice/couchdb.service
           └─4446 /usr/lib/erlang/erts-7.3/bin/beam -Bd -K tru
             └─4461 sh -s disksup
```

Ilustración 30 estado couchbd

Acceder a la dirección [http://127.0.0.1:5984/\\_utils/](http://127.0.0.1:5984/_utils/) [30] que es la dirección por defecto de couchdb para ver las configuraciones de este motor de bases de datos.



The screenshot shows the Apache CouchDB web interface. The browser tab is titled "Apache CouchDB - Futor". The address bar shows the URL "127.0.0.1:5984/\_utils/". The page title is "Overview". There is a "Create Database ..." button. A table lists two databases:

Name	Size	Number of Documents	Update Seq
<b>_replicator</b>	4.1 KB	1	1
<b>_users</b>	4.1 KB	1	1

Below the table, there are navigation links: "← Previous Page", "Rows per page: 10", and "Next Page →". At the bottom, it says "Showing 1-2 of 2 databases". On the right side, there is a "CouchDB relax" logo and a "Tools" menu with "Overview" and "Configuration" options.

## Configurar Python

Python es un lenguaje de programación interpretado, orientado a objetos y de alto nivel con semántica dinámica. Su estructura de datos integrados de alto nivel, combinados con el tipado dinámico y el enlace dinámico, lo hacen atractivo para el rápido desarrollo de aplicaciones, así como para su uso como scripting o lenguaje de pegado para conectar componentes existentes.

Python admite módulos y paquetes, lo que permite la modularidad del programa y la reutilización del código. Para este caso de estudio se utilizara una biblioteca de Python para couchdb; CouchDB-0.9 es un paquete que proporciona una interfaz de alto nivel conveniente para el servidor de couchdb. También se utilizara la librería tweepy, este paquete de software proporciona una comunicación con la plataforma de la red social twitter.

## Instalar tweepy

Esta herramienta nos permite acceder a las llaves generadas por Twitter y así se podrán recolectar los Tweets. Para instalar esta librería debemos instalar el paquete pip que nos permite instalar y administrar paquetes de software hechos en Python.

Para la instalación de pip se digita el siguiente comando:

```
apt install python-pip
```

```
root@diego-VirtualBox:~# apt install python-pip
Leyendo lista de paquetes... Hecho
Creando árbol de dependencias
Leyendo la información de estado... Hecho
Se instalarán los siguientes paquetes adicionales:
```

Ilustración 31 pip

Para la instalación de la librería tweepy ejecutamos el comando:

```
pip install tweepy
```

```
root@diego-VirtualBox:~# pip install tweepy
Collecting tweepy
  Downloading https://files.pythonhosted.org/packages/05/f1/2e8c7b202dd04117a378ac0c55cc7dafa80280ebd7f692f1fa8f27fd6288/tweepy-3.6.0-py2.py3-none-any.whl
Collecting requests>=2.11.1 (from tweepy)
```

Ilustración 32 tweepy

## CouchDB-0.9

Descargar y descomprimir esta librería para acceder a CouchDB desde python, la descargamos con el siguiente comando:

```
$wget https://pypi.python.org/packages/source/C/CouchDB/CouchDB-0.9.tar.gz  
[31]
```

```
root@diego-VirtualBox:~# wget https://pypi.python.org/packages  
/source/C/CouchDB/CouchDB-0.9.tar.gz  
--2018-04-30 20:13:08-- https://pypi.python.org/packages/sour  
ce/C/CouchDB/CouchDB-0.9.tar.gz  
Resolviendo pypi.python.org (pypi.python.org)... 151.101.0.223  
, 151.101.64.223, 151.101.128.223, ...
```

Ilustración 33 descarga de couchbd-09

Descomprimir los archivos descargados con el comando:

```
$tar zxvf CouchDB-0.9.tar.gz
```

```
root@diego-VirtualBox:~# tar zxvf CouchDB-0.9.tar.gz  
CouchDB-0.9/  
CouchDB-0.9/doc/  
CouchDB-0.9/doc/conf.py
```

Ilustración 34 descomprimir couchbd

Es necesario instalar archivos que correspondan a la librería couchdb, pero deben ser ejecutados con Python para que puedan ser reconocidos por este lenguaje de programación. Entonces se digita el comando:

```
python setup.py install
```

```
root@diego-VirtualBox:~/CouchDB-0.9# python setup.py install  
running install  
running bdist_egg  
running egg_info  
writing CouchDB.egg-info/PKG-INFO
```

Ilustración 35 instalar paquetes de python

A continuación se accede a Python desde la terminal y se importa la librería CouchDB a python.

```
python
```

```
>>> import couchdb  
>>>exit()
```

```
root@diego-VirtualBox:~/CouchDB-0.9# python  
Python 2.7.12 (default, Dec 4 2017, 14:50:18)  
[GCC 5.4.0 20160609] on linux2  
Type "help", "copyright", "credits" or "license" for more info  
rmation.  
>>> import couchdb  
>>> exit ()
```

Ilustración 36 import couchbd

Importamos tweepy, de esta manera el sistema operativo cuenta con todos los requisitos para poder recolectar los Tweets

```
import tweepy
```

```
root@diego-VirtualBox:~/CouchDB-0.9# import tweepy
```

Ilustración 37 import tweepy

## Ingreso API de Twitter

Acceder a Twitter como desarrollador para poder tener acceso a las claves generadas por la red social y así poder desarrollar la recolección de tweets

Ingresar a la aplicación web de desarrollador de Twitter desde el siguiente enlace:

<https://apps.twitter.com/> [32]

Ingresar a la plataforma con una cuenta de Twitter ya creada:

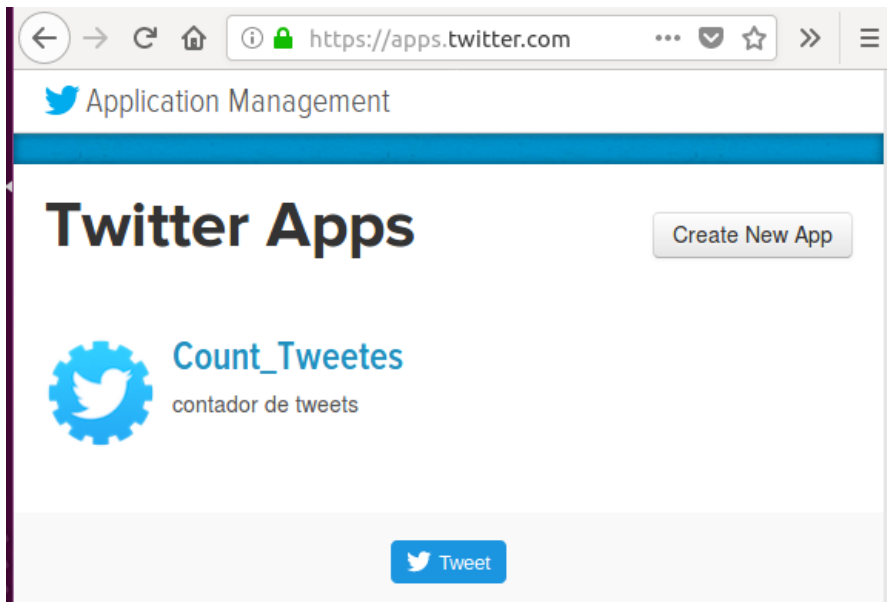


Ilustración 38 plataforma twitter

Luego de acceder a la plataforma se debe dar click en el botón de “Create New App” para crear una nueva aplicación, en la cual se despliegan ciertos campos que son obligatorios llenarlos.

---

## Create an application

Application Details

**Name \***  
  
Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.

**Description \***  
  
Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

**Website \***  
  
Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens.  
(If you don't have a URL yet, just put a placeholder here but remember to change it later.)

**Callback URL**  
  
Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth\_callback URL on the request token step, regardless of the value given here. To restrict your application from using callbacks, leave this field blank.

Ilustración 39 crear nueva app

Después de completar los datos requeridos, se deben aceptar los términos y condiciones y dar click en la opción Create your Twitter Application.

Your application has been created. Please take a moment to review and adjust your application's settings.

## ejemplo\_tweets

Test OAuth

Details Settings Keys and Access Tokens Permissions



contar tweets

https://www.utp.edu.co

### Organization

Information about the organization or company associated with your application. This information is optional.

Organization None

Organization website None

### Ilustración 40 ejemplo tweets

En esta parte se muestran los datos de creación de la aplicación así como el acceso a las claves que proporciona twitter para la creación de las nuevas aplicaciones, para acceder a ellas es necesario dar click a la opción “Keys and Access Tokens”

## ejemplo\_tweets

Test OAuth

Details Settings Keys and Access Tokens Permissions

### Application Settings

Keep the “Consumer Secret” a secret. This key should never be human-readable in your application.

Consumer Key (API Key) wXIB5jP2B8tiRLTiL2ffjkV25

Consumer Secret (API Secret) 1pzuBZBYGsLk647IT8r5joDS3vjLmCpNmnpCg1973X26120CKa

Access Level Read and write (modify app permissions)

Owner Koko1384

Owner ID 302379045

### Ilustración 41 claves tweet

Esta esta parte se debe dar click en el botón “Your Access Token” para generar las claves de acceso:

### Your Access Token

*This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.*

Access Token	002676645 aQ16n06LH764mK0M070jyA1n0BD70mZKRL000b
Access Token Secret	DjntCEbrDgizLIDomixars00bsmIn045ZQ05AKn0W0S2z4p0
Access Level	Read and write
Owner	Koko1384
Owner ID	302379045

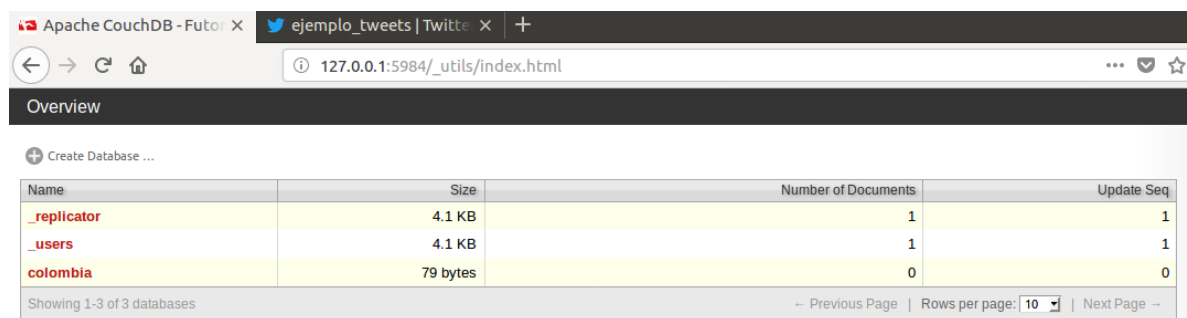
#### Ilustración 42 claves de acceso

Estas claves son necesarias al momento de escribir el código que permitirá la recolección de tweets

Crear la base de datos

Se debe crear una base de datos en el entorno de Couchdb, se debe acceder al navegador y digitar la siguiente dirección.

[http://127.0.0.1:5984/\\_utils/](http://127.0.0.1:5984/_utils/) [30]



The screenshot shows a web browser window with the URL `127.0.0.1:5984/_utils/index.html`. The page title is "Overview" and it displays a table of databases. The table has columns for Name, Size, Number of Documents, and Update Seq. The databases listed are `_replicator` (4.1 KB, 1 document, update seq 1), `_users` (4.1 KB, 1 document, update seq 1), and `colombia` (79 bytes, 0 documents, update seq 0). The page also shows a "Create Database ..." button and navigation controls.

Name	Size	Number of Documents	Update Seq
<code>_replicator</code>	4.1 KB	1	1
<code>_users</code>	4.1 KB	1	1
<code>colombia</code>	79 bytes	0	0

#### Ilustración 43 crear database

Para este caso creamos una base de datos llamada “Colombia”



Creación del Código en Python.

El desarrollo de esta aplicación se puede realizar en cualquier IDE que soporte Python.

Las claves requeridas para este paso se encuentran en la api de twitter como Consumer Key API, Consumer Secret API, Access Token, Access Token Secret.

```
GNU nano 2.5.3 Archivo: tweepys.py
import couchdb
from tweepy import Stream
from tweepy import OAuthHandler
from tweepy.streaming import StreamListener
import json

### API###
ckey = "wXlB5jP2B8tiRLtL2ffjkV25"
csecret = "1pzuBZBYGsLk647lT8r5j0D53vjLmCpNmnPcg1973X26120CKa"
atoken = "302379045-iwQ17c8RZ5bdnyxMyVuASmQkbiH4mDlMHXTeKZK"
asecret = "VMbW9S1V9CrOmzaFaRwKXuRbMcr5FaQ29h7Dd3qWrgC6I"

## Recoleccion de los Tweets y almacenaminto##

class listener(StreamListener):

    def on_data(self, data):
        dictTweet = json.loads(data)
        try:
            dictTweet["_id"] = str(dictTweet['id'])
            doc = db.save(dictTweet)
            print "duardo" + str(doc) + "=>" + str(data)
        except :
            print "ya existe"
            pass
        return True
    def on_error(self, status):
        print status

class listener(StreamListener):

    def on_data(self, data):
        dictTweet = json.loads(data)
        try:
            dictTweet["_id"] = str(dictTweet['id'])
            doc = db.save(dictTweet)
            print "duardo" + str(doc) + "=>" + str(data)
        except :
            print "ya existe"
            pass
        return True
    def on_error(self, status):
        print status

auth = OAuthHandler(ckey, csecret)
auth.set_access_token(atoken, asecret)
twitterStream = Stream(auth, listener())

server = couchdb.Server('http://localhost:5984/')
try:
    db = server.create('colombia')
except:
    db = server['colombia']

###locacion###

twitterStream.filter(locations=[-78.4,0.6,-69.1,8.8])
```

Ilustración 44 codigo python

Para la identificación de las coordenadas se utilizó la página web <https://boundingbox.klokantech.com> [33] la cual nos permite extraer las coordenadas de una delimitada zona geográfica.

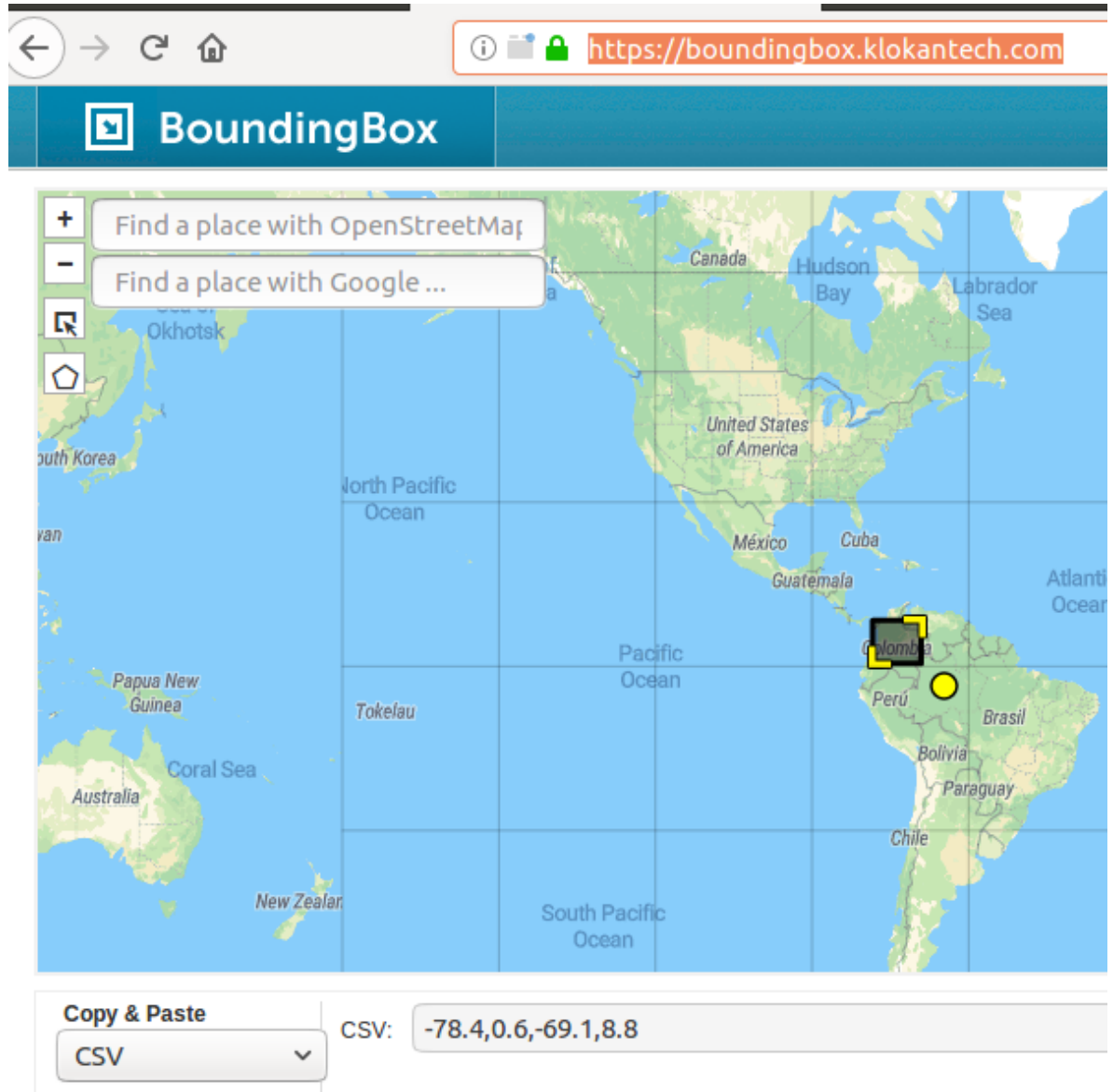


Ilustración 45 Cordenadas colombia

Se recomienda guardar este código de python en la carpeta de CouchDB-09 con el nombre deseado.

Ejecución del programa.

A continuación se debe ejecutar el programa desarrollado y verificar como se almacenan los datos.

Para ejecutar el programa se debe estar en el directorio donde se guardó el código realizado y ejecutar el comando:

python tweepys.py

```
root@dlego-VirtualBox:~/CouchDB-0.9# python tweepys.py
duardo(u'991140517290364928', u'1-2c7204b728ccd8d9db3353c6437cc82c')=>{"created_at": "Tue May 01 02:21:28 +0000 2018", "id": "991140517290364928", "id_str": "991140517290364928", "text": "@Hora20 PREOCUPADA POR TRATAR DE DEJAR EN CLARO, SEG\u00daN ELLA, EL PELIGRO QUE @petrogustavo representa para los ricos\u2026 https://t.co/TKiVw60ts", "source": "\u003ca href='\"http://twitter.com\"' rel='\"nofollow\"'\u003eTwitter Web Client\u003c/a\u003e", "truncated": true, "in_reply_to_status_id": null, "in_reply_to_status_id_str": null, "in_reply_to_user_id": 143285665, "in_reply_to_user_id_str": "143285665", "in_reply_to_screen_name": "Hora20", "user": {"id": "3406232428", "id_str": "3406232428", "name": "DESPE\u00da LOSADERO", "screen_name": "DPlosadero", "location": null, "url": null, "description": "En Colombia el periodismo est\u00e1 plagado d desinformantes sin principios ni car\u00e1cter traficantes d mentiras mercenarios d venganzas esbirros dl poder y del dinero", "translator_type": "none", "protected": false, "verified": false}}
```

#### Ilustración 46 Recoleccion de tweets

En la terminal aparecerán varias estructuras que contienen la información de los Tweets recolectados. De esta manera se van almacenando los datos en la base de datos NoSQL, en formato JSON.

Para verificar el almacenamiento se debe dirigir al explorador y acceder a la base de datos creada anteriormente.

The screenshot shows a web browser window with the address bar displaying '127.0.0.1:5984/\_utils/database.html?colombia'. The browser tabs include 'ejemplo\_tweets | Twitte' and 'Bounding Box Tool: Me'. The main content area shows a CouchDB interface for a database named 'colombia'. It features a table with two columns: 'Key' and 'Value'. The 'Key' column contains tweet IDs, and the 'Value' column contains JSON objects representing tweet data. The table shows 10 rows of data, with a 'Showing 1-10 of 33 rows' indicator at the bottom. The JSON objects in the 'Value' column include fields like 'created\_at', 'id', 'id\_str', 'text', 'source', 'truncated', 'in\_reply\_to\_status\_id', 'in\_reply\_to\_status\_id\_str', 'in\_reply\_to\_user\_id', 'in\_reply\_to\_user\_id\_str', 'in\_reply\_to\_screen\_name', 'user', 'location', 'url', 'description', 'translator\_type', and 'protected'.

Key	Value
"991140517290364928" ID: 991140517290364928	{rev: "1-2c7204b728ccd8d9db3353c6437cc82c"}
"991140526236798977" ID: 991140526236798977	{rev: "1-459187a3e0b311351078d6405c6f740e"}
"991140529135071232" ID: 991140529135071232	{rev: "1-82940562fdd29561888f75545e5699af"}
"991140534755430400" ID: 991140534755430400	{rev: "1-ed48f732d9b2dee2dbf4a8a0639295af"}
"991140538530353152" ID: 991140538530353152	{rev: "1-a56ffddb2ec09d7f1c9d4a88ddeb966b"}
"991140542271680512" ID: 991140542271680512	{rev: "1-312a4858ebc24e892e69a5d67fc992ab"}
"991140545991991296" ID: 991140545991991296	{rev: "1-d31f37c9db79c2df4640e1ae32eb3fe7"}
"991140552094769152" ID: 991140552094769152	{rev: "1-f76271c52bb8985699be4147214d5af"}
"991140562794315776" ID: 991140562794315776	{rev: "1-a0b091a0eada791eb19f85e3360f9258"}
"991140563033493504" ID: 991140563033493504	{rev: "1-ad801adf73854d15f15c471e20d4c464"}

#### Ilustración 47 datos tweet

Seleccionamos uno de los datos recolectados y al abrirlo se puede ver atributos como el nombre de usuario, la fecha en la que realizó el Tweet y, una gran variedad de contenido como el texto, Hashtag, latitud y longitud de donde se realizó esta transacción.

Field	Value
_id	"991140517290364928"
_rev	"1-2c7204b728ccd8d9db3353c6437cc82c"
contributors	null
coordinates	null
created_at	"Tue May 01 02:21:28 +0000 2018"
entities	<ul style="list-style-type: none"> <li>user_mentions</li> <li>symbols [ ]</li> <li>hashtags [ ]</li> <li>urls</li> </ul>
extended_tweet	<ul style="list-style-type: none"> <li>display_text_range</li> <li>entities</li> <li>full_text "@Hora20 PREOCUPADA POR TRATAR DE DEJAR EN CLARO, SEGÚN ELLA, EL PELIGRO QUE @petrogustavo representa para los ricos de este país, DIANA CALD..."</li> </ul>
favorite_count	0
favorited	false
filter_level	"low"
geo	null
id	991140517290364900
id_str	"991140517290364928"
in_reply_to_screen_name	"Hora20"

Ilustración 48 datos de tweet

## 7. METODOLOGIA PARA EL APRENDIZAJE

Se realizaran unas matrices de planificación sobre debe ser el proceso de enseñanza de Big Data, las cuales sirvan como un manual de apoyo para las personas que deseen aprender sobre cómo utilizar esta herramienta.

**Descripción:** La gran variedad y la extrovertida cantidad de datos junto con el análisis de estos, son la base fundamental sobre la cual se desarrollan estrategias de extracción de la información, este documento pretender brindar una ayuda a quien quiera aprender o enseñar los conceptos basados en el procesamiento de grandes volúmenes de datos los cuales proceden de muchas áreas y esto hace que su formato y velocidad sean distintos, este proceso se llevara a cabo haciendo uso de herramientas y tecnologías que facilitan el manejo de los datos y nos permiten realizar el análisis y la extracción de la información que es en pocas palabras lo que hace Big Data.

De esta manera, se pretende da a conocer a los interesados las posibles herramientas que pueden ser utilizadas al momento de desarrollar casos prácticos de Big Data, con lo cual se busca que quien ponga en práctica este manual sean capaces de usar este conocimiento para el desarrollo de nuevos casos, problemas, investigaciones, etc., que sean útiles en su formación personal y profesional.

Objetivo General del manual: Explicar a los interesados los conocimientos necesarios sobre Big Data partiendo de conceptos relevantes, herramientas y tecnologías que expongan su utilidad.

## MATRIZ DE PLANIFICACION

1. TIEMPO ESTIMADO : 1 SEMANA
2. OBEJTIVO: Brindar a los interesados los conceptos básicos acerca de Big Data, para que se familiaricen con esta herramienta.
3. TITULO DEL CAPITULO: Marco teorico

CONTENIDO DEL CAPITULO	METODOLOGIAS	RECURSOS	ESTANDARES DE EVALUACION
<p><b>Conceptos</b></p> <ul style="list-style-type: none"> <li>• Definicion de Big Data</li> <li>• 4V's de Big Data</li> <li>• Historia y evolución de Big Data</li> <li>• Tipos de datos</li> </ul> <p><b>Procedimiento</b></p> <ul style="list-style-type: none"> <li>• Leer el capítulo 2 de este manual practico</li> <li>• Practicar la lectura en libros de big data para aumentar los conocimientos, algunos de estos son:</li> <li>• Big data : la revolución de los datos masivos (Kenneth Cukier y Viktor Mayer-Schönberger)</li> <li>• Analytics: el uso de big data en el mundo real</li> </ul>	<p>ver conferencias acerca de Big Data y de cuál es su historia</p> <p><b>Conceptos:</b> Realizar un resumen con la información más relevante del capítulo 2</p>	<ul style="list-style-type: none"> <li>• Capítulo 2 del presente manual practico</li> <li>• Conferencias sobre Big Data</li> <li>• Textos de apoyo</li> <li>• Reforzar los conceptos de Big Data con Exposiciones de los temas expuestos en el capítulo 2</li> </ul>	<p>Al finalizar el capítulo 2 se deberán tener claro los siguiente:</p> <ul style="list-style-type: none"> <li>• Definir en sus propias palabras que es Big Data</li> <li>• Saber que hace Big Data</li> <li>• Inicios y Evolución de Big Data</li> </ul>

## MATRIZ DE PLANIFICACION

1. TIEMPO ESTIMADO : 1 SEMANA
2. OBEJTIVO: Brindar a los interesados conceptos objetivos de elementos de Big Data
3. TITULO DEL CAPITULO: Areas de Big Data Y Paradigmas del Big Data

CONTENIDO DEL CAPITULO	METODOLOGIAS	RECURSOS	ESTANDARES DE EVALUACION
<p><b>Conceptos</b></p> <ul style="list-style-type: none"> <li>• Areas de Big Data</li> <li>• Paradigmas de Big Data</li> </ul> <p><b>Procedimiento</b></p> <ul style="list-style-type: none"> <li>• Leer el capítulo 3 y 4 de este manual practico</li> <li>• Practicar la lectura en libros sobre Big Data para aumentar los conocimientos, algunos de estos son:</li> <li>• hadoop soluciones big data</li> </ul>	<p>Ver videos sobre cómo funciona Big Data, como está estructurado y los paradigmas expuestos en este documento.</p> <p><b>Conceptos:</b> Realizar un resumen sobre sobre las áreas de Big Data y dar una breve explicación de los paradigmas mencionados</p>	<ul style="list-style-type: none"> <li>• Capítulo 3 y 4 del presente manual practico</li> <li>• Conferencias sobre Big Data</li> <li>• Aprendizaje autónomo acerca de las herramientas y tecnologías de big data</li> <li>• Realizar ejercicios donde se puedan aplicar y resolver utilizando MapReduce</li> </ul>	<p>Al finalizar el capítulo 3 y 4 se deberán tener claro lo siguiente:</p> <ul style="list-style-type: none"> <li>• Definir en sus propias palabras cuales son las áreas de Big Data</li> <li>• conocer como funciona Mapreduce</li> <li>• conocer algunas plataformas en la que se puede implementar Big Data</li> </ul>

## MATRIZ DE PLANIFICACION

4. TIEMPO ESTIMADO : 2 SEMANAS
5. OBEJTIVO: lograr que los interesados puedan instalar un ambiente de Big Data y que puedan ejecutar algunos ejemplos y así puedan conocer como es el funcionamiento de Big Data
6. TITULO DEL CAPITULO: Instalación de un ambiente Big Data

CONTENIDO DEL CAPITULO	METODOLOGIAS	RECURSOS	ESTANDARES DE EVALUACION
<p><b>Conceptos</b></p> <ul style="list-style-type: none"> <li>• Instalacion de un ambiente Big Data</li> </ul> <p><b>Procedimiento</b></p> <ul style="list-style-type: none"> <li>• Seguir los pasos expuestos en el capítulo 7 del manual sobre Big Data</li> <li>•</li> </ul>	<p>Ver tutoriales sobre cómo se debe instalar Hadoop</p> <p>Leer documentos sobre Hadoop, para facilitar el proceso de instalación</p> <p><b>Conceptos:</b> Realizar un laboratorio siguiendo los pasos expuestos en el manual,</p>	<ul style="list-style-type: none"> <li>• Capítulo 7 del presente manual practico</li> <li>• Tutoriales sobre la instalación y ejecución de Hadoop</li> <li>• Aprendizaje autónomo acerca de las herramientas y tecnologías de big data</li> <li>• Realizar proyectos en los cuales se puedan implementar casos propios de Big Data</li> </ul>	<p>Al finalizar el capítulo 7 se deberán tener claro lo siguiente:</p> <ul style="list-style-type: none"> <li>• Como instalar un ambiente de Big Data</li> <li>• Ejecutar casos prácticos para conocer el funcionamiento de Big Data</li> </ul>



## **8. CONCLUSIONES**

Big data es una tendencia que brinda una ayuda para el manejo de grandes volúmenes de información, principalmente es utilizado por grandes empresas, pero gracias a las nuevas tecnologías y su fácil acceso es posible ser utilizado por cualquier empresa o institución que desee vincularse con esta herramienta

Las plataformas de Big Data al permitir el manejo de datos estructurados y no estructurados, presentan un gran beneficio para la toma de decisiones gracias a la facilidad de manejar todos esos tipos de datos, lo cual proporciona ventajas tanto para la vida profesional como para los diferentes campos de la ciencia

La estructura de un ambiente Big Data ayuda a mejorar la manipulación de los datos, optimizando la gestión de la información respecto a tiempo y costo, logrando obtener mejores resultados en las estadísticas para una buena toma de decisiones.

## 9. BIBLIOGRAFIA

[1] IBM. que es el Big Data. Disponible en:

<https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/>

[2] Oracle. Plataforma Oracle Big Data

<https://www.oracle.com/es/big-data/index.html>

[3] Baos Analytics Everywhere. Las 4V's del Big Data

<https://www.baoss.es/las-4-vs-del-big-data/>

[4] INSIDE BIGDATA

<https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/>

[5] Veracity: the most important "V" of Big Data

<https://www.gutcheckit.com/blog/veracity-big-data-v/>

[6] WINSHUTTLE. BIG DATA Y HISTORIA DEL ALMACENAMIENTO DE LA INFORMACION.

<https://www.winshuttle.es/big-data-historia-cronologica/>

[7] A Short History of Big Data

<https://dataflog.com/read/big-data-history/239>

[8] A brief history of big data everyone should read

<https://www.weforum.org/agenda/2015/02/a-brief-history-of-big-data-everyone-should-read/>

[9] KYOCERA Document Solutions. Diferencia entre datos estructurados y no estructurados

<https://smarterworkspaces.kyocera.es/blog/diferencia-datos-estructurados-no-estructurados/>

[10] Datos no estructurados 101

<https://www.ondata.com/nuix/blog/2015/06/25/datos-no-estructurados-101/>

[11] BIGDATAHOY

<https://bigdatahoy.wordpress.com/2015/05/04/a-que-nos-referimos-con-informacion-no-estructurada/>

[12] Power Data Especialistas en gestión de datos. Data Warehouse: todo lo que necesitas saber sobre almacenamiento de datos

<https://www.powerdata.es/data-warehouse>

[13] Signaturit sing anywhere, anytime. Qué es Business Intelligence (BI) y qué herramientas existen

<https://blog.signaturit.com/es/que-es-business-intelligence-bi-y-que-herramientas-existen>

[14] salesforce. Cloud Computing- Aplicaciones en un solo tacto

<https://www.salesforce.com/mx/cloud-computing/>

[15] ¿ Qué es el Big Data?

<http://www.fundacionctic.org/sat/articulo-que-es-el-big-data>

[16] BBVA. 10 herramientas para la visualización de datos

<https://www.bbva.com/es/10-herramientas-visualizacion-datos/>

[17] Towards Data Science

<https://towardsdatascience.com/top-4-popular-big-data-visualization-tools-4ee945fe207d>

[18] The 7 Best Data Visualization Tools In 2017 “artículos”

<https://www.forbes.com/sites/bernardmarr/2017/07/20/the-7-best-data-visualization-tools-in-2017/#17a461756c30>

[19] SolidQ. Que es MapReduce?

<http://blogs.solidq.com/es/big-data/que-es-mapreduce/>

[20] ZDNet. MapReduce and MPP: Two sides of the Big Data coin?

<https://www.zdnet.com/article/mapreduce-and-mpp-two-sides-of-the-big-data-coin/>

[21] MPP (massively parallel processing)

<https://whatis.techtarget.com/definition/MPP-massively-parallel-processing>

[22] techopedia. Big Data Analytics Platform

<https://www.techopedia.com/definition/31750/big-data-analytics-platform>

[23] Baos Analytics Everywhere. 10 Herramientas para manejar Big Data Analytics

<https://www.baoss.es/10-herramientas-para-manejar-big-data-analytics/>

[25] webopedia

<https://www.webopedia.com/TERM/A/apache-spark.html>

[26]

<https://archive.apache.org/dist/hadoop/core/hadoop-2.6.0/>

[27]

<http://localhost:50070>

[28]

<http://localhost:50090/status.jsp>

[29] acens. Bases de datos NoSQL

<https://www.acens.com/wp-content/images/2014/02/bbdd-nosql-wp-acens.pdf>

[30]

[http://127.0.0.1:5984/\\_utils/](http://127.0.0.1:5984/_utils/)

[31]

<https://pypi.python.org/packages/source/C/CouchDB/CouchDB-0.9.tar.gz>

[32]

<https://apps.twitter.com/>

[33]

<https://boundingbox.klokantech.com>

[34] NoSmoke. NoSQL

<http://nosmoke.cycle-it.com/2014/03/31/nosql/>

[35] brandwatch. 98 estadísticas de las redes sociales para 2017

<https://www.brandwatch.com/es/blog/98-estadisticas-de-las-redes-sociales-para-2017/>