EBook Gratis

APRENDIZAJE bigdata

Free unaffiliated eBook created from
Stack Overflow contributors.

#bigdata

Tabla de contenido

Acerca de 1

Capítulo 1: Empezando con bigdata 2

Observaciones 2

Un desarrollador podría estar interesado en las capacidades de procesamiento de big data p 3

Examples 3

¿Qué es Big Data? 3

Ejemplo de Big Data 4

¿Qué viene bajo Big Data? 5

Capítulo 2: Comenzando con Big Data / Hadoop Security 7

Introducción 7

Observaciones 7

Examples 8

Ejemplo para ACL 8

Creditos 9

Acerca de

You can share this PDF with anyone you feel could benefit from it, downloaded the latest version from: bigdata

It is an unofficial and free bigdata ebook created for educational purposes. All the content is extracted from Stack Overflow Documentation, which is written by many hardworking individuals at Stack Overflow. It is neither affiliated with Stack Overflow nor official bigdata.

Use the content presented in this book at your own risk; it is not guaranteed to be correct nor accurate, please send your feedback and corrections to info@zzzprojects.com

Capítulo 1: Empezando con bigdata

Observaciones

Esta sección proporciona una descripción general de qué es bigdata y por qué un desarrollador puede querer usarlo.

Big data son los datos caracterizados por las 4 V's. Estos son Volumen, Velocidad, Variedad y Veracidad.

1. Volumen: cuando la cantidad de datos está en un volumen enorme como Terabytes o Petabytes. Como dice un informe, hemos generado datos del 90% del mundo en los últimos 2 o 3 años.
2. Velocidad: la velocidad a la que fluyen los datos en el sistema. Por ejemplo, millones de usuarios que cargan su contenido en sitios de redes sociales al mismo tiempo generan datos tan altos como en el rango de Terabytes / seg.
3. Variedad - Diferentes tipos de datos en función de su naturaleza. Puede estar estructurado (como la mayoría de los RDBMS anteriores), Semiestructurado (correo electrónico, XML, etc.) y No estructurado (Videos, Audios, Datos de sensores, etc.).
4. Veracidad: es el medio con el que obtenemos una visión significativa de nuestros datos disponibles. Esto puede considerarse como el aspecto más importante de los datos, ya que la mayoría de las decisiones de negocios dependen de la utilidad de los datos.

La plataforma más general utilizada para almacenar y procesar big data es el **Marco de Hadoop.** Se compone de 2 cosas:

1. Sistema de archivos distribuidos de Hadoop (HDFS): los datos se almacenan en el Sistema de archivos distribuidos de Hadoop (HDFS), que en realidad es un conjunto de hardware básico, a diferencia de la forma primitiva de almacenamiento en servidores. Los datos residen en HDFS y pueden procesarse para obtener información utilizando diversas herramientas. y marcos.
2. MapReduce (MR): este es el marco de procesamiento predeterminado para Hadoop. MapReduce (es parte de Apache Hadoop)

Con un avance en Hadoop, comenzaron a surgir nuevas herramientas de procesamiento en la comunidad de Hadoop. Algunas de las herramientas / marcos más populares:

1. Chispa de apache
2. Tormenta apache
3. Apache Flink

   Y muchos más..

Algunos de los mecanismos de almacenamiento distintos de HDFS simple:

1. Colmena
2. HBase
3. Casandra

Y muchos más..

Un desarrollador podría estar interesado en las capacidades de procesamiento de big data para que pueda ser una gran diferencia en la forma en que vemos nuestros datos. En un universo paralelo, también podemos llamar big data como Rich-Untamed-Data. Tenemos que domesticar esta gran cantidad de datos. Con big data podríamos ser capaces de procesar el potencial oculto de los datos ya existentes.

Se puede citar un mejor ejemplo en el comportamiento de clics de los clientes en los sitios web de compras en los que sus visitas, clics y la cantidad de tiempo que pasan en ese sitio web, le dice al minorista en línea que adquiera el producto y envíe recomendaciones basadas en el comportamiento del usuario.

Examples

¿Qué es Big Data?

Big Data, en su forma más básica, se puede describir como el término general métrico por diferentes aspectos de los datos. Estos diferentes aspectos son

> Volumen (Gran cantidad de Datos), Velocidad (Mayores velocidades de flujo de datos), Variedad (Datos estructurados, no estructurados y semiestructurados) y Veracidad (Tomar decisiones correctas basadas en datos).

Estas métricas eran difíciles de cuidar por las bases de datos relacionales de la vejez. Surgió la necesidad de un nuevo sistema y el procesamiento de Big Data llegó al rescate. Si bien muchas personas tienen una comprensión diferente sobre qué es Big Data, aquí hay algunas de las definiciones de Big Data dadas por los líderes de la industria en el sector de Datos:

Definiciones:

- "Big Data supera el alcance de los entornos de hardware y herramientas de software más utilizados para capturarlo, administrarlo y procesarlo en un tiempo transcurrido tolerable para su población de usuarios" (Artículo de la revista Teradata, 2011)
- "Big data se refiere a conjuntos de datos cuyo tamaño está más allá de la capacidad de las herramientas típicas de software de base de datos para capturar, almacenar, administrar y analizar". (The McKinsey Global Institute, 2012)
- "Big Data es una colección de conjuntos de datos tan grandes y complejos que se vuelve

difícil de procesar con herramientas de administración de bases de datos disponibles". (Wikipedia, 2014)
- "Los Big Data son activos de información de alto volumen, alta velocidad y / o gran variedad que requieren nuevas formas de procesamiento para permitir una mejor toma de decisiones, recuperación de información y optimización de procesos" (Gartner, 2012)

Cuando los datos se vuelven "grandes"?


```
IOPS:Input/Output Operations Per Second
```

Ejemplo de Big Data

Big data es un término para conjuntos de datos que son tan grandes o complejos que las aplicaciones tradicionales de procesamiento de datos son inadecuadas para tratarlos. Los desafíos incluyen análisis, captura, curación de datos, búsqueda, intercambio, almacenamiento, transferencia, visualización, consulta, actualización y privacidad de la información.

Un ejemplo general de big data:

Datos recogidos por la red social facebook. Facebook recopila cientos de terabytes (TB) de datos

todos los días. Los datos recopilados pueden ser imágenes, videos, publicaciones, actualizaciones, etc. Los datos varían de estructurados a no estructurados. A me gusta, compartir o reaccionar puede ser datos estructurados, ya que sabemos claramente la estructura de los mismos. Mientras que las actualizaciones o publicaciones son datos no estructurados que no siguen exactamente una estructura. ¡Todos estos datos juntos forman BigData!

¿Qué viene bajo Big Data?

Big data involucra los datos producidos por diferentes dispositivos y aplicaciones. A continuación se presentan algunos de los campos que están bajo el paraguas de Big Data.

- Datos de la caja negra: es un componente del helicóptero, aviones y aviones, etc. Captura las voces de la tripulación de vuelo, las grabaciones de micrófonos y auriculares, y la información de rendimiento de la aeronave.
- Datos de las redes sociales: las redes sociales como Facebook y Twitter contienen información y las opiniones publicadas por millones de personas en todo el mundo.
- Datos de la bolsa de valores: los datos de la bolsa de valores contienen información sobre las decisiones de "compra" y "venta" tomadas sobre una parte de las diferentes compañías que los clientes tomaron.
- Datos de la red eléctrica: Los datos de la red eléctrica contienen información consumida por un nodo en particular con respecto a una estación base.
- Datos de transporte: los datos de transporte incluyen el modelo, la capacidad, la distancia y la disponibilidad de un vehículo.
- Datos del motor de búsqueda: los motores de búsqueda recuperan gran cantidad de datos de diferentes bases de datos.
- Datos de sensores: datos de diferentes dispositivos que trabajan con sensores, por ejemplo: datos meteorológicos (meteorológicos y climáticos), datos sísmicos (terremotos), datos oceánicos (mareas, tsunamis, etc.).

Por lo tanto, Big Data incluye gran volumen, alta velocidad y amplia variedad de datos. Los datos en él serán de tres tipos.

```
1. Structured data : Mostly data from Relational Databases.

2. Semi Structured data : XML data, email data.

3. Unstructured data : Word, PDF, Text, Media Logs.
```

Lea Empezando con bigdata en línea: https://riptutorial.com/es/bigdata/topic/6890/empezando-con-bigdata

Capítulo 2: Comenzando con Big Data / Hadoop Security

Introducción

Podemos asegurar los datos en Hadoop usando diferentes métodos. Cada método tiene sus propias ventajas. También podemos combinar más de un método para obtener mejores resultados. Este tema cubre las ventajas y limitaciones de cada método.

Observaciones

1. Kerberos es un protocolo de autenticación de red:

a. Ventaja: Autenticar usuarios en el nivel de entrada.

segundo. Limitación: Kerberos impide el acceso de usuarios no autorizados al entorno. Pero después de iniciar sesión, no proporcionará autenticaciones de nivel detallado como tabla, columna, carpeta, nivel de archivo, etc.

2. Apache Sentry es un sistema para imponer bases de funciones de grano fino

a. Ventaja: Autentificaciones de nivel de aplicación como Hive, Impala, Solr, etc. Puede controlar el acceso en la base de datos, tabla, nivel de columna para un usuario / grupo en particular.

segundo. Limitación: no puede controlar las carpetas HDFS que están subrayadas detrás de aplicaciones como Hive, Impala, etc. Ej .: Hive table prod.table1 almacenada en /user/hive/warehouse/prod.db/table1. La configuración de la función de centinela en Hue puede controlar solo el acceso a la tabla / columna en Hue, pero es posible que el usuario pueda administrar para acceder a las carpetas directamente en HDFS

do. Limitación: las carpetas HDFS que no estén relacionadas con Hive, Impala, etc. no serán controladas

3. Una lista de control de acceso (ACL) es una lista de entradas de control de acceso (ACE). Cada ACE en una ACL identifica a un administrador y especifica los derechos de acceso permitidos, denegados o auditados para ese administrador.

a. Ventaja: el acceso a nivel de carpeta es posible para los usuarios que usan

4. El cifrado HDFS implementa un cifrado transparente de extremo a extremo de los datos leídos y escritos en HDFS.

a. Ventaja: Cifrar los datos proporcionará seguridad de nivel adicional. En General, el cifrado de datos es requerido por un número de diferentes entidades gubernamentales, financieras y reguladoras

Examples

Ejemplo para ACL

```
hadoop fs -setfacl
```

Lea Comenzando con Big Data / Hadoop Security en línea:
https://riptutorial.com/es/bigdata/topic/9869/comenzando-con-big-data---hadoop-security

Creditos

S. No	Capítulos	Contributors
1	Empezando con bigdata	Ani Menon, Community, Mr. P, NeoWelkin, Sayali Sonawane
2	Comenzando con Big Data / Hadoop Security	saranvisa