Apache Storm para BigData

Jaime Esaú Zuleta Calzada
Estructura de Datos
Ingeniería en ciencias de la computación
Universidad Francisco Gavidia

Abstract

Debido a que las tecnologías de información constantemente están cambiando y creciendo da lugar a nuevos desafíos y oportunidades para procesar la información con diferentes técnicas o software. BigData es una de las tecnologías más competitiva en los negocios trayendo consigo grandes beneficios que la mayoría desconoce, al tratarse de BIgData estamos hablando que se tratan de bases de datos donde se almacenan millones de bytes en un solo nodo, al intentar procesar toda esta información sin las herramienta adecuadas podríamos tardas años queriendo procesar, organizar y presentar esta información, es allí donde surge la necesidad de ApacheStorm para el procesado de un gran volúmen de datos y en tiempo real.

Introducción

BigData destaca por ser una de las tecnologías más importantes en la forma en que las empresas utilizan analizan y procesan la información para beneficio de sus modelos de negocios, ya que combina la gestión de tecnologías de datos. BigData permite empresa para utilizar, almacenar y gestionar una gran cantidad de datos con la precisión tiempo para lograr la eficiencia una de las características más importante es que los datos recopilados deben administrarse de una manera que se mantengan los requisitos comerciales a pesar de trabajar con millones de datos por segundo.

Big data es un término moderno que cambió la forma de analizar todos los datos y sus métodos, la recopiclacion y análisis de datos va más allá d las técnicas tradicionales de recopilación de datos. ApacheStorm realiza las operaciones de procesamiento de datos de una manera más fácil y confiable,

no importando que tan grande sea el flujo de datos que se desea procesar, en comparación a Hadoop este realiza el procesamiento por lotes mientras que apache ApacheStorm lo realiza en tiempo real.

1. BigData

Big Data es una base de datos que aloja información, sin embargo, que posee característica como diferentes como volumen, variedad y la velocidad se combinan y enlazan para establecer grandes datos. Dicha información puede ser consultada en tiempo real la cual puede provenir diferentes fuentes o de diferentes tipos, por ejemplo; podría ser contenido de blog, imágenes, geolocalización, registros, etc. Comúnmente las bases de suelen ser de estructuras homogéneas y no pueden procesan datos heterogéneos, por lo tanto, esto dificulta la posibilidad de trabajar con ellas y produce incompatibilidad. Dichas información de la base de datos se puede utilizar en una base de datos relacionada.

2. Herramientas disponibles

Algunas de las herramientas de código abierto que se utilizan para analizar la BigData:

- **Apache HBase**

  Es un software basado en lenguaje Java que permite almacenar Big Data. Es no relacional y proporciona la funcionalidad similar a Bigtable de Google para almacenar datos.

- **Hadoop**

  Una de las principales características de este software son: fiabilidad, escalabilidad y su modelo de procesamiento, que permite procesar conjuntos de datos en grupos de máquinas utilizando el paradigma de programación distribuida.

- **Apache Spark**

  Este software permite procesar datos de forma más rápida y a gran escala, que se basan en la composición de clústeres, otra de las características es que puede operar en lotes, teniendo como

una desventaja que su tamaño de procesamiento por lotes es pequeño. Este software permite programar en tres diferentes lenguajes Java, Scala y Python.

3. Apache Storm

También es conocido como Hadoop, los desarrolladores crear sistemas de procesamiento distribuidos en tiempo real, con la capacidad de poder procesar flujos de datos grandes e ilimitados en tiempo real. Apache Storm es escalable, fácil de usar y ofrece baja latencia con datos garantizados procesados. Proporciona una arquitectura muy simple para la construcción de aplicaciones que se denomina *Topologías.* Una de las ventajas que ofrece ApacheStorm es que los desarrolladores pueden crear dichas aplicaciones en cualquier lenguaje que sea permita la comunicación con el formato de intercambio de datos JSON

Un clúster de Storm tiene tres conjuntos de nodos:

Nodo Nimbus: es el servidor principal donde el código de usuario debe estar cargado, distribuye el código en el clúster, lanza trabajadores en todo el clúster y monitorea el cálculo y reasigna trabajadores según sea necesario.

Nodos ZooKeeper: coordina el clúster Storm

Nodos de supervisor: se comunica con Nimbus a través de Zookeeper, inicia y detiene a los trabajadores según las señales de Nimbus.

4. Ventajas de ApacheStorm

- Facilidad para recuperar los datos en tiempo real no importando el tamaño de la información.
- La rapidez con la que pude procesar millones de bytes por segundo por cada nodo.
- Tolerante a fallas ya que realiza un seguimiento de todos los nodos trabajadores, cada vez que un nodo falla, el proceso es reiniciado en otro nodo.
- Fiabilidad de procesamiento de datos garantizado.
- Escalabilidad procese los datos en paralelo en un grupo de ordenadores.
- Modelo de procesamiento de flujo en tiempo real.
- Confiable cada una de las tuplas de datos debe ser procesada por lo menos una vez.

5. Casos de uso de ApacheStorm

- Servicios financieros

  Ayuda a la prevención de fraude, riesgos operacionales y violación de cumplimiento.
- Telecomunicaciones

  Prevención de brechas de seguridad y cortes de red.
- MYPE

  Prevención de contracción y desabastecimientos
- Fabricación

  Ayuda a la facilidad de mantenimiento preventivo y asegura la calidad.
- Transporte

  Facilita la supervisión de controladores y mantenimiento predictivo
- Web
- Previene fallas en la aplicación y problemas operacionales

6. Requisistos mínimos del hardware

2.4 GHz Intel Core i5, 8GB DDR3 SDRAM.

Oracle Java SDK 1.7.0_60

Apache Storm 0.9.2-incubating

Twitter4j 4.0.1

7. Ejemplo de procesamiento en tiempo real

Podemos crear una topología que cuente el número de palabras que nosotros elijamos, en este caso podríamos utilizar el Api de Streaming de Twitter. Esta topología al conectarse a twitter ira recolectando cada tweet que se vaya publicando y se eran agrupando por la palabra de manera que se puedan distribuir de una mejor manera y contabilizar el número de apariciones

Conclusión

Apache Storm en su última generación reconoce patrones particulares de una gran cantidad de datos en un periodo de tiempo particular, permite a los desarrolladores poder acoplarse fácilmente ya que es compatible con diferentes lenguajes de programación y no importando cual lenguaje se escoja apacheStorm no pierde ninguna de sus características para procesar los datos. La configuración de apachStorm es una tarea difícil, se debe realizar manualmente ya que no cuenta con una guía directa para la configuración, los desarrolladores esperan a futuro una versión más amigable y fácil de utilizar ya que es una de las herramientas mejor evaluada para el procesamiento de BigData en tiempo real.

Referencias

Iqbal, M. H., & Soomro, T. R. (2015). Big data analysis: Apache storm perspective. International journal of computer trends and technology, 19(1), 9-14.

Batyuk, A., & Voityshyn, V. (2016, August). Apache storm based on topology for real-time processing of streaming data from social networks. In 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP) (pp. 345-349). IEEE.