Herramientas para el Análisis de Datos

Herramientas para el Análisis de Datos

DOCENTE VIRTUAL

U de Cataluña

Diplomado en Big Data – Data Science

TABLA DE CONTENIDO

Herramientas para el Análisis de Datos

Uno de los objetivos del uso de las tecnologías Big Data es el de transformar los datos en conocimiento útil para la empresa, y para ello se necesitan herramientas Big Data que nos ayuden a analizar, procesar y almacenar todos los datos recogidos. Un gran número de entre las mejores herramientas usadas en Big Data son: open source, lo que da fe del éxito de este modelo de desarrollo, además de las alternativas de pago. A continuación, se enlista una selección de herramientas open source que ofrecen soluciones para la explotación de software de Big Data en todos sus procesos: almacenamiento, procesamiento y análisis:

1. Hadoop

Esta herramienta Big Data open source se considera el framework estándar para el almacenamiento de grandes volúmenes de datos; se usa también para analizar y procesar, y es utilizado por empresas como Facebook y Yahoo!. La biblioteca Hadoop utiliza modelos de programación simples para el almacenamiento y procesamiento distribuido de grandes conjuntos de datos en clústeres, dando redundancia para no perder nada y, al mismo tiempo, aprovechando muchos procesos a la vez. Dispone de un sistema de archivos distribuido en cada nodo del clúster: el HDFS (Hadoop Distributed File System), y se basa en el proceso de MapReduce de dos fases. Soporta diferentes sistemas operativos y también se usa frecuentemente sobre cualquiera de las principales plataformas en la nube, como Amazon EC2/S3 o Google Cloud.

2. MongoDB

Dentro de las bases de datos NoSQL, una de las más famosas es MongoDB. Con un concepto muy diferente al de las bases de datos relacionales, se está convirtiendo en una interesante alternativa para almacenar los datos de nuestras aplicaciones. MongoDB es una base de datos orientada a documentos (guarda los datos en documentos, no en registros). Estos

documentos son almacenados en BSON, que es una representación binaria de JSON. A pesar de que las bases de datos NoSQL no tienen una extensa variedad de uso, MongoDB tiene un ámbito de aplicación más amplio en diferentes tipos de proyectos: es especialmente útil en entornos que requieran escalabilidad. Con sus opciones de replicación y sharding, podemos conseguir un sistema que escale horizontalmente sin demasiados problemas.

3. ElasticSearch

Elasticsearch es una potente herramienta para la búsqueda entre grandes cantidades de datos, especialmente cuando los datos son de tipo complejo. Nos permite indexar y analizar en tiempo real un gran volumen de datos y hacer consultas sobre ellos. Un ejemplo de uso son las consultas de texto completo; al estar los datos indexados, los resultados se obtienen de forma muy rápida. En el IIC utilizamos esta herramienta para indexar datos dentro de nuestras soluciones de entorno digital. A diferencia de otros sistemas parecidos, no necesita declarar un esquema de la información que añadimos, no sabemos exactamente qué forma van a tener los datos. Con Elasticsearch podemos hacer búsquedas de texto complicadas, visualizar el estado de nuestros nodos y escalar sin demasiadas necesidades, si se diera el caso de que necesitáramos más potencia.

3. Apache Spark

Apache Spark es un motor de procesamiento de datos de código abierto realmente rápido. Creado por Matei Zaharia en la Universidad de Berkeley, se considera el primer software open source que hace la programación distribuida (muy en esencia, consiste en distribuir el trabajo entre un grupo de ordenadores, "clúster", que trabajan como uno realmente accesible a los científicos de datos. Se pueden programar aplicaciones usando diferentes lenguajes como Java,

Scala, Python o R. pudiendo ser, según el programa, hasta 100 veces más rápido en memoria o 10 veces más en disco que Hadoop MapReduce.

4. Apache Storm

Apache Storm es un sistema de computación distribuida en tiempo real orientado a procesar flujos constantes de datos, por ejemplo, datos de sensores que se emiten con una alta frecuencia o datos que provengan de las redes sociales, donde a veces es importante saber qué se está compartiendo en este momento. Aunque Hadoop sea un gran sistema para el procesado de un gran volumen de datos, no está pensado para hacerlo en tiempo real, ya que tiene una alta latencia. Apache Storm está siendo una revolución para procesar grandes cantidades de información en tiempo real, ya que es capaz de procesar millones de mensajes por segundo. En el IIC utilizamos Apache Storm para nuestra herramienta Lynguo, que requiere esta tecnología Big Data para procesar en tiempo real los comentarios de las redes sociales para su monitorización y análisis. Apache Storm puede ser utilizado para procesar los logs de nuestras aplicaciones para ver el uso que se hace de los distintos servicios y gestión de errores; para extraer información de redes sociales a través de sus APIs y analizar un fenómeno en tiempo real; recoger y procesar datos de sensores; buscadores verticales, web analytics, etc.

5. Lenguaje R

R es un lenguaje de programación y entorno de software para cálculo estadístico y gráficos. El lenguaje R es de los más usados por los estadistas y otros profesionales interesados en la minería de datos, la investigación bioinformática y las matemáticas financieras. R se parece más al lenguaje de las matemáticas que a otros lenguajes de programación, lo que puede ser un inconveniente para los programadores a la hora de elegir programar en R para temas de Big Data. Lo que está claro es que si eliges usar R podrás disponer de una gran cantidad de librerías

creadas por la comunidad de R y otras tantas herramientas de altísima calidad (por ejemplo, RStudio).

6. Python

Python es un lenguaje avanzado de programación con la ventaja de ser relativamente fácil de usar para usuarios que no estén familiarizados con la informática de manera profesional, pero que necesitan trabajar con análisis de datos (estadistas, biólogos, físicos, lingüistas…). Es una herramienta para Big Data muy eficiente, en parte debido a la gran comunidad existente, por lo que Python dispone de muchas librerías ya hechas por otros usuarios. Sin embargo, tiene en su contra que no es un lenguaje muy rápido en su ejecución, por lo que suele ser empleado para tareas de integración o tareas donde no haya cálculos pesados.

BIBLIOGRAFÍA

- http://www.iic.uam.es/innovacion/herramientas-big-data-para-empresa/

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

  https://www.universidadviu.es/las-herramientas-big-data-mas-conocidas/

© U de Cataluña, 2020

Todos los derechos reservados. Prohibida la reproducción total o parcial sin permiso o autorización de la Universidad, Bogotá - Colombia.