Febrero 2014

atsistemas
Consulting, IT Services & Software Development

Big Data con nombres propios

Al hablar de tecnología Big Data se está obligado, sin duda alguna, a hablar de programación paralela y procesamiento distribuido, ya que éstas serán las características que permitirán ofrecer escalabilidad, alta disponibilidad y capacidad de replicación de estos sistemas.

Una tarea básica de una tecnología Big Data es el procesamiento de las fuentes de datos, donde se distinguiría entre el procesamiento Batch de las fuentes de datos; y, por otro lado, el procesamiento en tiempo real, o pseudo tiempo real, de las fuentes de datos.

Para hacerlo de una forma más didáctica, se hará un símil respecto a un mar de datos, donde bucear en el océano buscando el tesoro sería el procesamiento Batch típico del Big Data; y lanzar unas redes de arrastre esperando atrapar algo valioso sería el procesamiento en tiempo real o pseudo tiempo real.

Procesamiento Batch

Para este procesamiento Batch, la solución tecnológica de referencia es, sin duda alguna, las que están basadas en Apache Hadoop, un framework de procesamiento distribuido que utiliza el paradigma map/reduce, donde los datos son mapeados, agrupados y reducidos como forma de procesamiento; y que, además, utiliza un sistema de ficheros distribuido llamado HDFS, consiguiendo un sistema escalable, de alta disponibilidad y con replicación de datos, y todo ello sobre hardware barato. El proyecto Hadoop está compuesto, a su vez, por un buen número de subproyectos como Sqoop, Pig, Hive, HBase, Flume, Zookeeper y otros más que le dotan de una gran funcionalidad. A esto se le denomina ecosistema Hadoop.

Tomando como base Hadoop, varios fabricantes han decidido realizar distribuciones más o menos mejoradas y optimizadas, en una estrategia similar a las distintas distribuciones de Linux. De todas ellas, destacan:

cloudera
Ask Bigger Questions

- Instalación más sencilla
- Interfaz web para su gestión
- Incluye Impala, para consultas en tiempo real sobre Hadoop

Hortonworks

- Instalación sencilla
- Soporta Hadoop 2.0 con YARN, para la gestión de jobs y tasks
- Incluye Storm

Apache Stark, un proyecto con su origen en la Universidad de Berkeley, está ganando bastante terreno, y no sin razón.

Sin embargo, aunque se podría calificar a Hadoop como la tecnología más madura sobre las que se asientan la mayoría de las soluciones Big Data para procesamiento Batch; Apache Stark, un proyecto con su origen en la Universidad de Berkeley, está ganando bastante terreno, y no sin razón. Ahora acogido por Apache, se basa en procesamiento distribuido map/reduce en memoria, permitiendo que se mejoren entre 10 y 100 veces los resultados realizados sobre Hadoop. Se integra fácilmente con HDFS y con bases de datos NoSQL; y dispone de frameworks en Java, Scala y Python. Su único delito es que aún es demasiado joven.

Tiempo real o pseudo real

Para poder realizar el procesamiento en tiempo real mencionado anteriormente, tendríamos que recurrir a soluciones que sean capaces de trabajar con flujos de datos. Existen algunas tecnologías con concepciones diferentes, como pueden ser S4, Akka, Storm o Trident.

Aquí, sin duda alguna, la apuesta es por Storm, quien se basa en la definición de los productores de datos, llamados Spouts; los procesadores de los datos, que llama Bolts y la definición de una topología en la que se definen cómo se relacionan los productores y los procesadores. A todo eso lo llama Topology. De esta forma tan sencilla, y al estar construido sobre Zookeeper, consigue una velocidad de procesamiento espectacular, se programa en Java de una manera muy simple, y se puede conectar a sistemas de colas, bases de datos relacionales o NoSQL, HDFS, etc.

Mundo NoSQL

Forma parte de cualquiera de las dos formas de procesar los datos, pues son necesarias sí o sí. Pero la verdadera cuestión es cuál de las dos utilizar, ya que existen multitud de opciones, todas ellas muy interesantes.

Se podrían clasificar de acuerdo a la forma en que organizan los datos: columnares, clave-valor, documental, etc., Sin embargo, ahora mismo ya existen tres de referencia que van un paso por delante, no en términos tecnológicos, sino más por el respaldo del mercado y las buenas estrategias de sus fabricantes y comunidades. Éstas son:

Pero la verdadera cuestión es cuál de las dos utilizar, ya que existen multitud de opciones, todas ellas muy interesantes.

atsistemas
Consulting, IT Services & Software Development

Big Data con nombres propios

Para todos estos componentes ya existen múltiples opciones, y a diario están apareciendo nuevos frameworks y herramientas.

APACHE HBASE

- Base de datos basada en Google BigTable
- Funciona muy bien con Hadoop, pues forma parte de su ecosistema
- Fácil de usar desde otros componentes Hadoop como Hive o Pig
- Está bien documentada y es bastante madura

Cassandra

- Base de datos columnar que ofrece varios APIs en distintos lenguajes
- Ha desarrollado su propio lenguaje, CSQL, que permite consultas tipo SQL
- Eficiente en el tiempo de respuesta
- Fácilmente integrable con Hadoop

mongoDB

- Utiliza documentos BSON, de tipo JSON, aumentando su flexibilidad con respecto a los esquemas de datos; y reduciendo el esfuerzo de los desarrolladores
- Es distribuida y dispone de recopilación de datos
- Puede realizar sharding de los documentos, tareas map/reduce; y dispone funciones específicas para sistemas de geolocalización
- Fácilmente integrable con Hadoop

Juventud, divino pecado.

Finalmente, una arquitectura Big Data debería contar de ciertos componentes de manera estándar, como son:

- El aprovisionamiento de datos
- Capacidad de procesamiento distribuido
- Herramientas para consultas en tiempo real
- Sistemas para el almacenamiento de datos y mecanismos de publicación

Para todos estos componentes ya existen múltiples opciones, y a diario están apareciendo nuevos frameworks y herramientas. Sin embargo, todos pecan de una gran juventud a excepción de Hadoop, y esto significa que nos encontramos en un momento en el que no hay claramente tecnologías de referencia. Esto llegará con el tiempo y con el respaldo que les ofrezcan las comunidades de desarrolladores y el mundo empresarial.

Big Data con nombres propios

atsistemas
Consulting, IT Services & Software Development

Referencias de Tecnologías

Hadoop	Framework OpenSource desarrollado para el procesamiento distribuido implementando el paradigma map/reduce
HDFS	Hadoop Distributed File System, sistema de ficheros distribuidos utilizado por el framework Hadoop, permite escalabilidad y replicacion de los ficheros
Pig	Subproyecto Hadoop, es un lenguaje de programación de mas alto nivel que permite generar procesos map/reduce
Hive	Subproyecto Hadoop, es un lenguaje de programación similar a SQL que permite generar procesos map/reduce
HBase	Subproyecto Hadoop, es una base de datos NoSQL basada en BigTable de Google
Flume	Subproyecto Hadoop, es una herramienta para poder manejar flujos de datos en sistemas de ficheros HDFS
Zookeeper	Subproyecto Hadoop, es un framework que ofrece utilidades y primitivas para el desarrollo de servicios de coordinación en arquitecturas de procesamiento distribuido
Impala	Componente propietario de las últimas distribuciones Cloudera Hadoop, que permite realizar consultas en tiempo real
YARN	Yet Another Resource Negotiator, forma parte de la version 2.0 de Hadoop, y es una tecnología para una mejor y más potente gestión de los recursos y procesos de un cluster Hadoop
S4	Simple Scalable Streaming System, S4, framework de desarrollo para procesamiento continuo de streams que utiliza el paradigma de programación basado en actores
Akka	Framework para programación concurrente basado el paradigma de programación basado en actores
Storm	Framework de desarrollo Java para procesamiento de streams con base en Zookeeper
Trident	Frameworks de programación construido sobre Storm que permite agrupaciones, joins, agregaciones, funciones y filtros sobre streams
Cassandra	Base de datos NoSQL de tipo columnar, soporta fuerte desnormalización, y posee vistas. Dispone de un lenguaje propio CSQL, que permite realizar Querys al estilo SQL
MongoDB	Base de datos NoSQL, de tipo documental, utiliza objetos BSON(Binary JSON), que ofrece flexibilidad total de esquemas de datos

atsistemas
Consulting, IT Services & Software Development

En atSistemas somos más de 500 profesionales dedicados desde 1994 a la consultoría, servicios de IT y desarrollo de software. Nuestros servicios se caracterizan por la flexibilidad y la agilidad, lo que nos permite ayudar a grandes empresas de todos los sectores, aportando conocimiento y experiencia sobre el más amplio abanico de tecnologías.

Nuestra cartera de clientes incluye más de 200 de las principales empresas del país, con representación de todos los sectores de actividad, a los que prestamos servicio desde nuestras oficinas de Madrid, Barcelona, Cádiz, Zaragoza y A Coruña.

Nuestro portfolio de servicios abarca desde el desarrollo de software a medida hasta la integración de grandes soluciones de software empresarial, en áreas que van desde la más compleja arquitectura de sistemas hasta las soluciones más novedosas de comercio electrónico o aplicaciones móviles.

Valle de Alcudia, 3
28230 Las Rozas, Madrid

Passeig de Gràcia 55, 8º - 4ª
08007 Barcelona

902 888 902

atsistemas.com

info@atsistemas.com