

Big Data: A través de una implementación

Diego Krauthamer

Universidad Abierta Interamericana
Av. San Juan 951 Buenos Aires, Argentina
diego.krauthamer@uai.edu.ar

Abstract. El presente tutorial se centrará en una puesta al día de las tecnologías presentes en el mercado relacionadas con Big Data; es decir que el análisis de datos en la actualidad requiere de herramientas especializadas basadas en dichas tecnologías. Diariamente se producen 2 millones de búsquedas en google y se genera un volumen de datos que no puede almacenarse, ni procesarse con tecnologías de bases de datos tradicionales. En los últimos años han tenido gran auge los sistemas Big Data, impulsados por el crecimiento exponencial de datos, y por la política de datos abiertos u Open Data que distintas organizaciones han adoptado.

Keywords: Big Data, Hadoop, Análisis de Datos.

1 Introducción

El tutorial tiene como principal objetivo introducir al asistente en una de las últimas tendencias tecnológicas como lo es Big Data.

Diariamente se generan en internet millones de datos no estructurados; esto representa un gran desafío tanto a nivel de almacenamiento, como de procesamiento. En virtud de ello se requieren tecnologías que posibiliten cumplir con dichos propósitos.

El tutorial estará dividido en dos partes: la primera de ellas será una introducción a Big Data y las tecnologías asociadas necesarias para la creación, almacenamiento y explotación de datos; la segunda consistirá en la demostración de una implementación concreta de Big Data.

En la primera parte se abordará Big Data desde lo teórico realizando un breve síntesis de la evolución de las tecnologías de bases de datos, para luego ahondar en el surgimiento y criterios que llevaron al concepto de Big Data, como así también de las distintas tecnologías que permiten el almacenamiento, procesamiento y explotación de datos (Hadoop, Hive, Pig, etc.).

En la segunda parte se mostrará en vivo a los asistentes HDInsight de Microsoft, que es una de las implementaciones de Big Data que existen en el mercado. Se creará un clúster Hadoop, luego se ejecutarán scripts de Powershell, Hive y Sqoop, y se finalizará con un ejemplo de análisis de datos mediante una herramienta de usuario final como es Microsoft Excel. De esta manera se brindará a los asistentes la posibilidad de conocer el funcionamiento de la tecnología Big Data a través de una implementación en la nube.

2 Resumen de Contenidos

2.1 Primera parte - Introducción

- 2.1.1 ¿Qué es Big Data?
- 2.1.2 Hadoop
 - 2.1.1 ¿Qué es Hadoop?
 - 2.1.2 HDFS
 - 2.1.3 MapReduce
- 2.1.3 Tecnologías asociadas.
 - 2.1.3.1 Hive
 - 2.1.3.2 Sqoop
 - 2.1.3.3 Pig
 - 2.1.3.4 Big SQL

2.2 Segunda parte - Demostración

- 2.2 Microsoft HDInsight
 - 2.2.1 Introducción
 - 2.2.2 Programación en HDInsight
 - 2.2.2.1 Mapreduce Jobs y Windows Powershell
 - 2.2.3 Análisis de Datos
 - 2.2.3.1 Usando Apache Hive
 - 2.2.3.2 Usando Apache Pig
 - 2.2.3.3 Usando Microsoft Excel y Power Pivot.

3 Resultados del aprendizaje

Como resultado el asistente podrá introducirse en el mundo de Big Data, como así también de sus tecnologías asociadas Hadoop, Hive, Pig, etc. Asimismo; podrá conocer en vivo una de tantas implementaciones de Big Data en la existentes en el mercado, como lo es HDInsight de Microsoft.

Referencias

1. Chauhan, A., Fontana, V., Hart M, M.S.: Introducing Microsoft Azure HDInsight, pp. 16 Microsoft Press (2014)
2. Schoenborn, B.: Big Data Analytics Infrastructure for Dummies, pp. 6--7, Wiley (2014)
3. PowerData: Del Bit a Big Data, <http://landings.powerdata.es/del-bit-a-big-data-guia-gratuita>.
4. White, T.: Hadoop the Definitive Guide. Second Edition, pp. 13--45, O' Reilly /Yahoo Press (2011).
5. HortonWorks : Introducing Apache Hadoop for Developers, <http://hortonworks.com/hadoop-tutorial/introducing-apache-hadoop-developers/>