



Introducción a Big Data y su aplicación en el entorno asegurador



Fernando Turrado García – Octubre 2016

Presentación

▶ Ponente

- ▶ Fernando Turrado García
 - ▶ Matemático, Máster en Investigación Informática (UCM)
 - ▶ Arquitecto de Información (OCASO S.A.)

▶ Objetivos

- ▶ Describir la problemática conocida como Big Data
- ▶ Introducción a la tecnología
- ▶ Discusión de cuatro casos de uso en el entorno asegurador

Índice

- ▶ **¿Qué es Big Data? (15')**
- ▶ **Tecnología (60')**
 - ▶ Motivación
 - ▶ Batch: Ecosistema Apache Hadoop
 - ▶ Tiempo real: Storm / Spark Streaming
 - ▶ Otras tecnologías: Pregel / Giraph / Machine Learning / ...
 - ▶ Líneas futuras
- ▶ **Casos de negocio (30')**
 - ▶ Social Networks: Sentiment Analysis
 - ▶ Social Networks: Networking & Fraud Detection
 - ▶ Internet Of Things: Black Box Insurance
 - ▶ Internet Of Things: Sensor networks & Healthcare
- ▶ **Comentarios (15')**

¿Qué es Big Data?



¿Qué es Big Data?

“Big Data is like teenage sex;
everybody talks about it,
nobody really knows how to do it,
everyone thinks everyone else is doing it,
so everyone claims they are doing it.”

Dan Ariely, Duke University

Enero 2013

¿Qué es Big Data?

- ▶ “**Big data**, **macrodatos** o **datos masivos** es un concepto que hace referencia al **almacenamiento de grandes cantidades de datos** y a los procedimientos usados para encontrar patrones repetitivos dentro de esos datos. El fenómeno del **big data** también se denomina a veces **datos a gran escala**.” (Wikipedia)
- ▶ Tipologías de datos:
 - ▶ Estructurados.
 - ▶ No estructurados.
 - ▶ Semiestructurados.
- ▶ Metadatos
- ▶ Datos vs Información

¿Qué es Big Data?

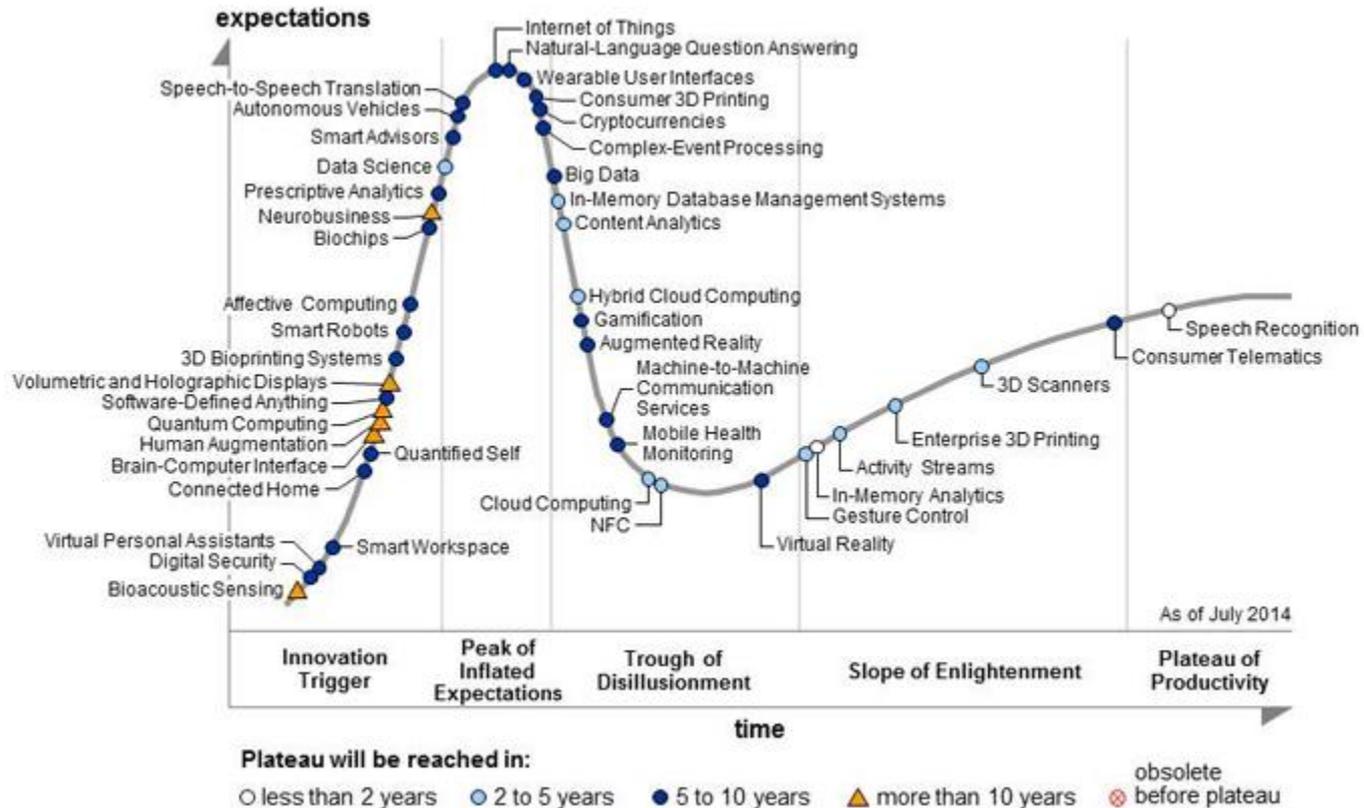
- ▶ Problemas calificados como Big Data:

- ▶ Análisis de redes sociales: marketing, elecciones, ...
- ▶ Medicina: Secuenciación de genes, tratamiento de imágenes ...
- ▶ Mantenimiento predictivo, redes de sensores, ...

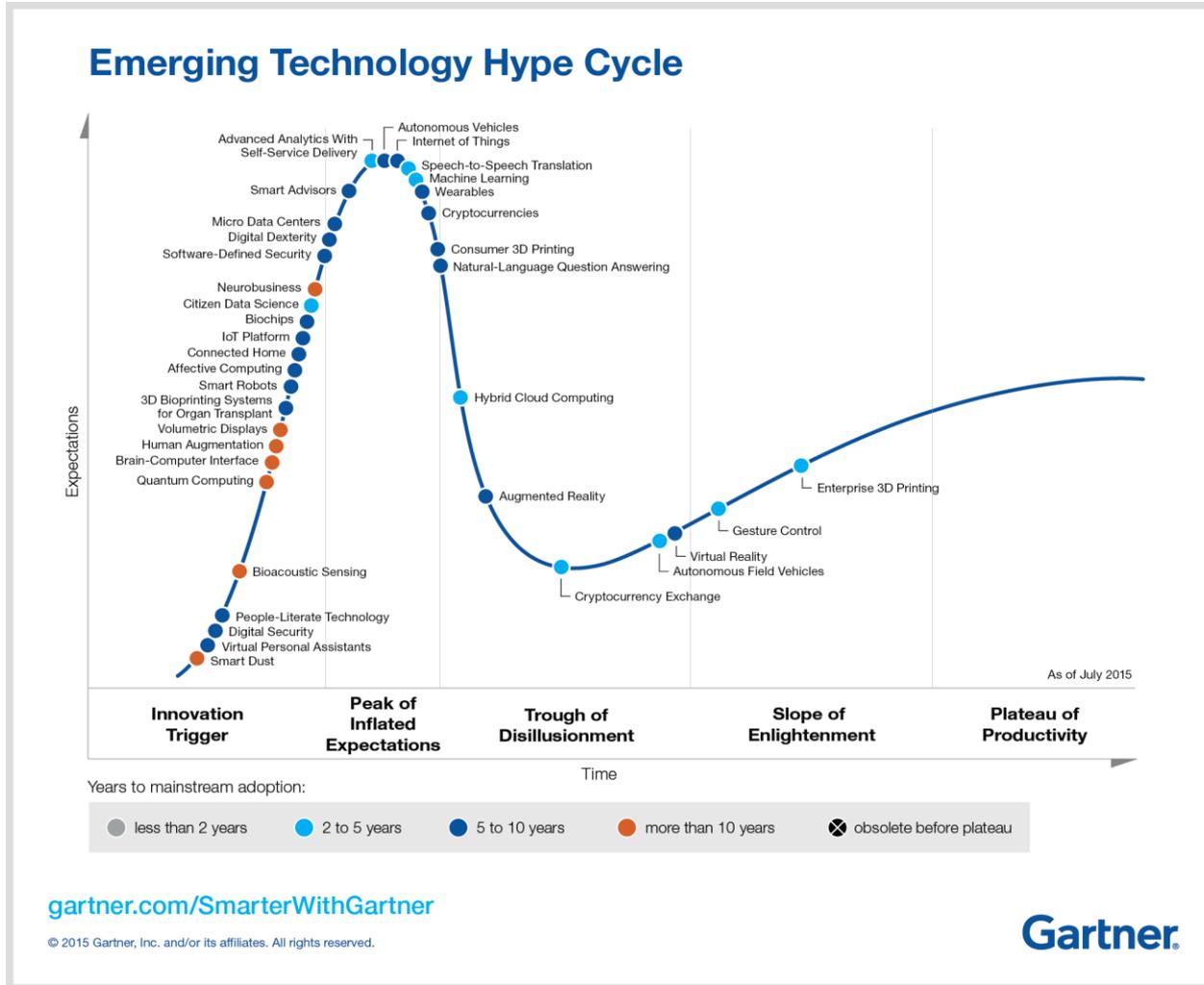
- ▶ En el año 2001, Gartner publicó en un informe denominado “3D Data Management: Controlling Data Volume, Velocity and Variety.” en el que se definieron las variables (hoy se conocen como las 3 V’s) que caracterizan los problemas de tipo BigData:

- ▶ Volume: Cantidad de datos a procesar.
- ▶ Velocity: Tiempo disponible para procesar los datos.
- ▶ Variety: Ausencia de homogeneidad en los datos a procesar.

¿Qué es Big Data?



¿Qué es Big Data?



Tecnología

Tecnología - Motivación

- ▶ Reto: ¿Cómo proporcionar soluciones tecnológicas a los problemas planteados?
- ▶ Requisitos:
 - ▶ Técnicos: Escalabilidad y tolerancia a fallos.
 - ▶ Económicos.
 - ▶ Legales: Protección de datos personales
 - ▶ ...
- ▶ Un sistema informático se considera escalable (o que posee la capacidad de escalar) si ante un aumento en volumen de los datos que maneja se le pueden añadir recursos (hardware) para mantener el rendimiento.
- ▶ Existen dos tipos de escalabilidad:
 - ▶ Vertical: Se obtiene ampliando los recursos individuales de cada uno de los servidores. Por ejemplo, RAM, número de procesadores, disco ...
 - ▶ Horizontal: Se obtiene ampliando el número de servidores presentes en el sistema.
- ▶ Los problemas “BigData”, por definición, necesitan sistemas que tengan escalabilidad horizontal.

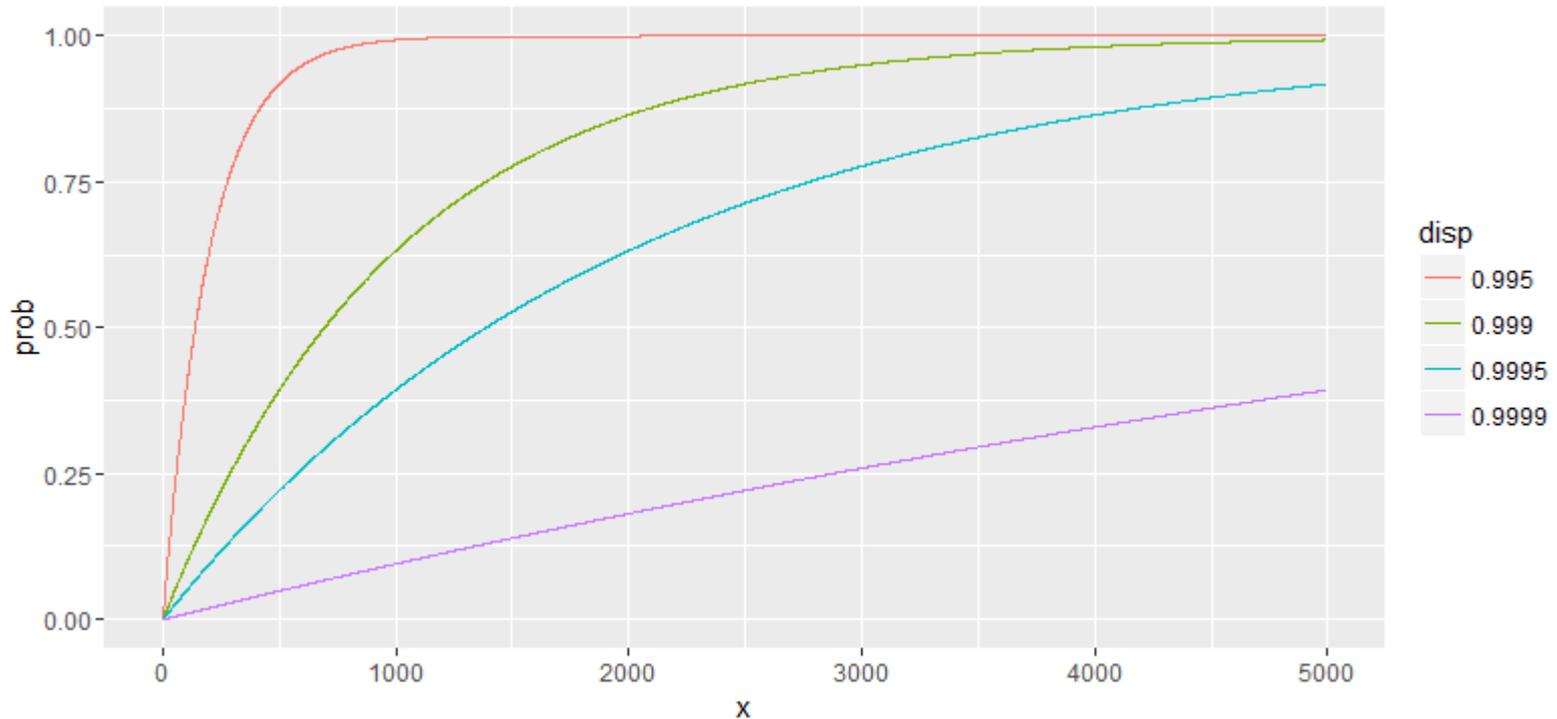
Tecnología - Motivación

- ▶ Según la ley de Moore, cada 18 meses se puede comprar el mismo hardware a la mitad de precio.
- ▶ Para minimizar los costes hardware de infraestructura, se definen los “commodity clusters” en oposición al modelo de supercomputadores.
- ▶ Los “commodity clusters” están compuestos por cientos o miles de servidores pequeños y es un modelo para las infraestructuras en la nube.
- ▶ Imaginemos que se está garantizando el 99,9% disponibilidad por cada servidor, menos de 9 horas de inactividad al año. => La probabilidad de que un servidor no esté disponible es de 0.1%.
- ▶ En un clúster de 1000 servidores, ¿cuál es la probabilidad de que al menos uno no esté disponible?.

Tecnología - Motivación

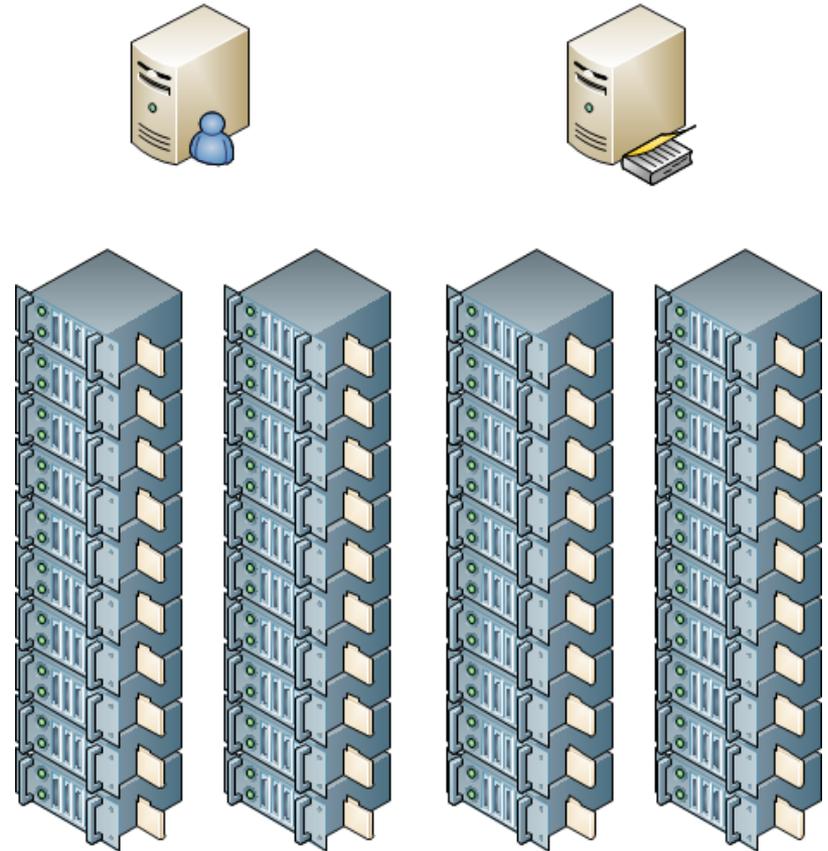
$$P(1 + \text{en error}) = 1 - P(0 \text{ en error}) = 1 - P(\text{no error})^N = 1 - 0.999^{1000} = 0.6323$$

63.23%



Tecnología – Sistemas distribuidos

- ▶ **DFS (Distributed File System):** Sistema de archivos distribuido que permite el almacenamiento de grandes volúmenes de datos.
- ▶ Cada fichero almacenado en un DFS es dividido en trozos, los cuales son repartidos (y replicados) entre los diferentes integrantes del clúster.
- ▶ **Servicio de directorio:** Almacena la relación de los ficheros creados en el DFS, las ubicaciones de los trozos y de las réplicas ...
- ▶ **Servicio de gestión de trabajos:** Se encarga de coordinar la ejecución distribuida de los trabajos.



Tecnología – Computación distribuida

- ▶ **Multihebra**

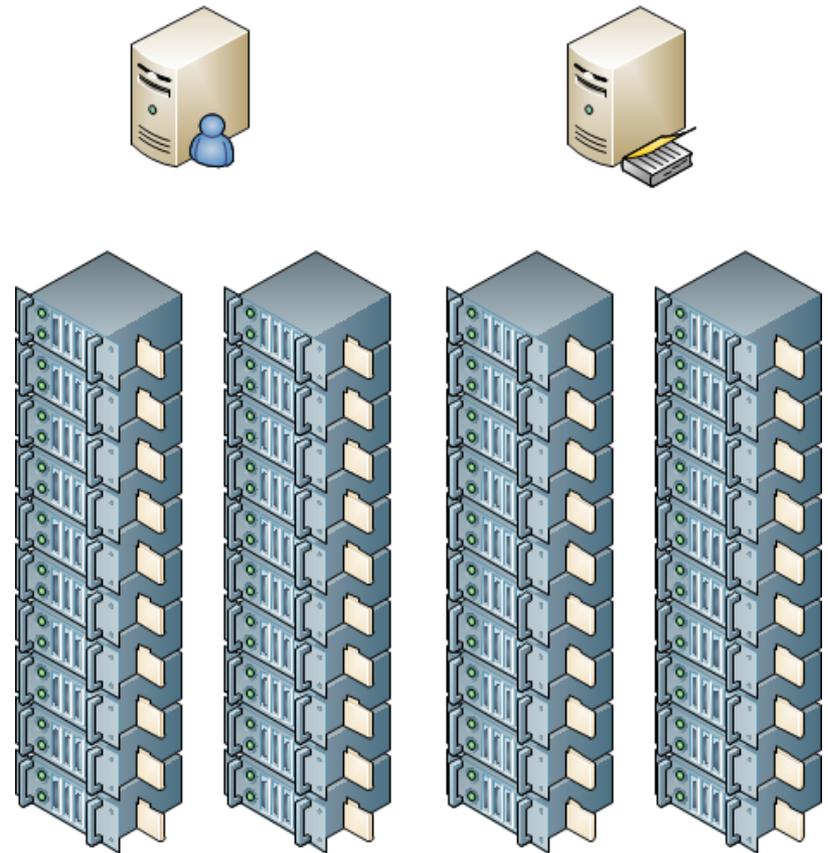
- ▶ **Multiproceso**

- ▶ **Gestión de lotes**

- ▶ Tolerancia a fallos
- ▶ Reintentos

- ▶ **Gestión de procesos**

- ▶ Tolerancia a fallos
- ▶ Reintentos



Tecnología – Hadoop

- ▶ En 2004, dos ingenieros de Google publican un artículo de investigación denominado “MapReduce: Simplified Data Processing on Large Clusters”.
- ▶ Este artículo describe un modelo funcional de computación distribuida pensado explícitamente para el tratamiento de grandes volúmenes de datos en clústeres compuestos por cientos o miles de servidores.
- ▶ Apache Hadoop es la implementación de referencia de este nuevo modelo de programación.
- ▶ Yahoo! ha sido uno de sus grandes impulsores y contribuidores.

Tecnología – Map Reduce

Apache Hadoop Wins Terabyte Sort Benchmark

By aanand – [Wed, Jul 2, 2008 11:42 AM EDT](#)

 Recommend

0

 Tweet

One of Yahoo's [Hadoop](#) clusters sorted 1 terabyte of data in **209 seconds**, which beat the previous record of 297 seconds in the annual general purpose (daytona) [terabyte sort benchmark](#). The sort benchmark, which was created in 1998 by Jim Gray, specifies the input data (10 billion 100 byte records), which must be completely sorted and written to disk. This is the first time that either a Java or an open source program has won. Yahoo is both the largest user of Hadoop with 13,000+ nodes running hundreds of thousands of jobs a month and the largest contributor, although non-Yahoo [usage](#) and [contributions](#) are increasing rapidly.

Tecnología – Map Reduce

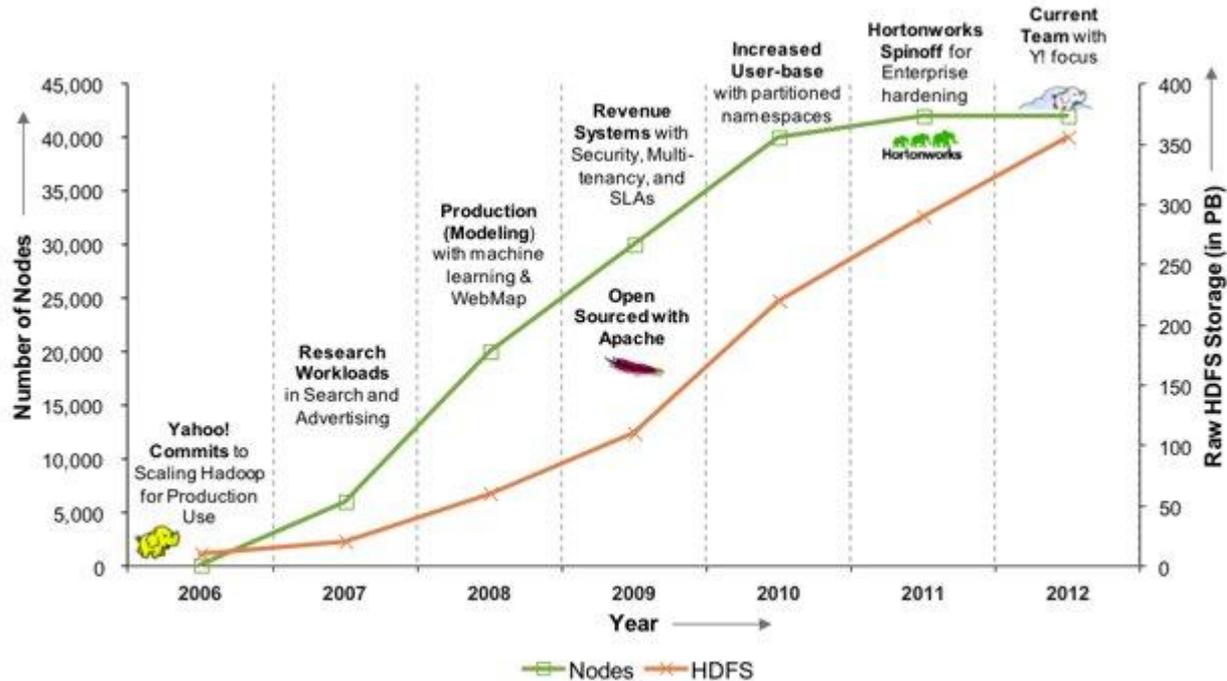
The cluster statistics were:

- 910 nodes
- 2 quad core Xeons @ 2.0ghz per a node
- 4 SATA disks per a node
- 8G RAM per a node
- 1 gigabit ethernet on each node
- 40 nodes per a rack
- 8 gigabit ethernet uplinks from each rack to the core
- Red Hat Enterprise Linux Server Release 5.1 (kernel 2.6.18)
- Sun Java JDK 1.6.0_05-b13

The benchmark was run with Hadoop trunk (pre-0.18) with a couple of optimization patches to remove intermediate writes to disk. The sort used 1800 maps and 1800 reduces and allocated enough memory to buffers to hold the intermediate data in memory. All of the code for the benchmark has been checked in as a Hadoop example.



Tecnología - Hadoop



Tecnología – Hadoop

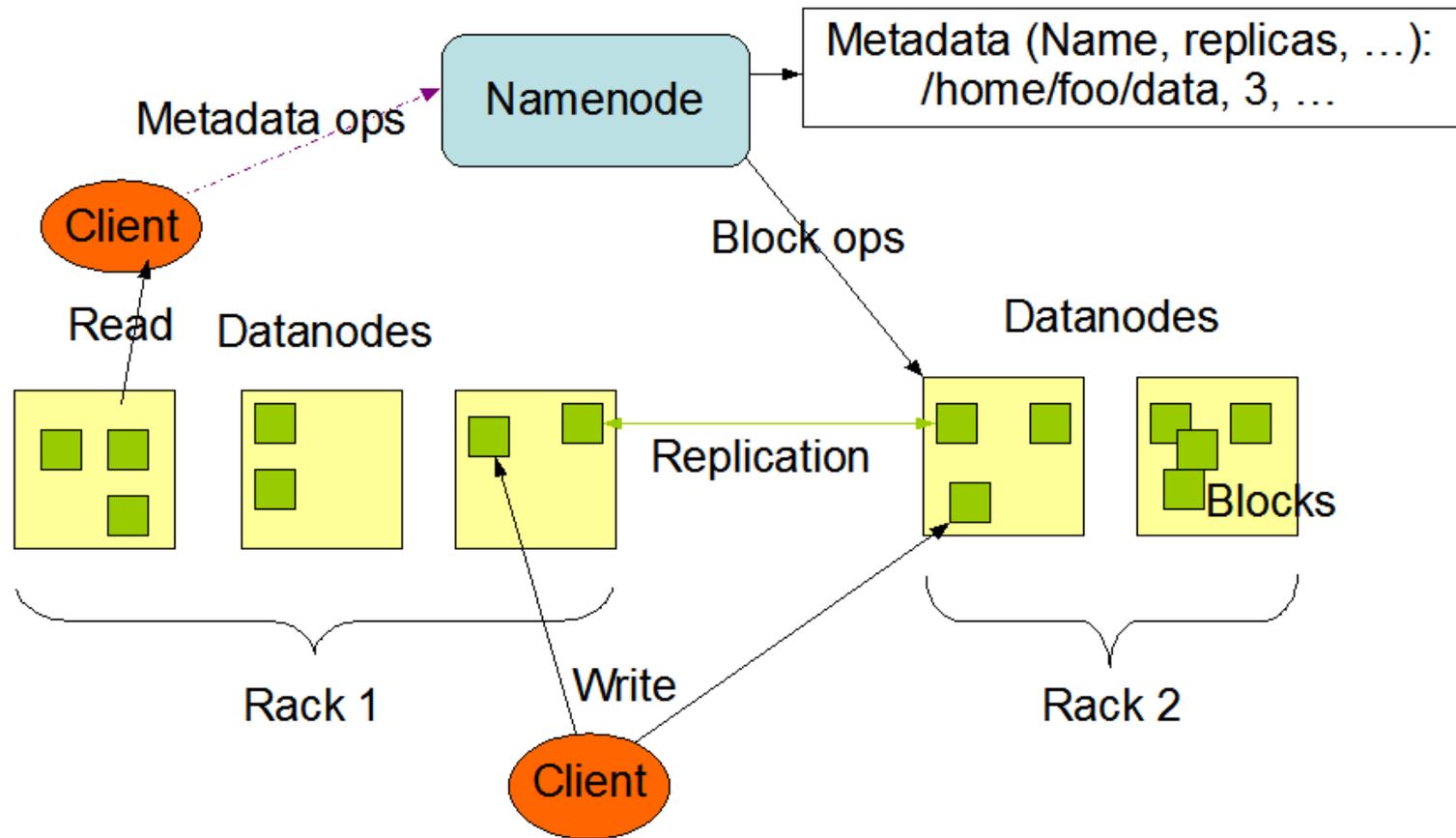
- ▶ Apache Hadoop está compuesto por varios módulos, siendo los principales:
 - ▶ HDFS: Sistema de almacenamiento distribuido
 - ▶ MapReduce: Framework de programación.
 - ▶ YARN: Gestor de trabajos y recursos
- ▶ En la versión 1 de Hadoop sólo se podían desarrollar programas siguiendo el patrón MapReduce.
- ▶ La versión 2 (actual), generaliza el concepto de programa permitiendo utilizar otros frameworks como Apache Spark ó Tez.

Tecnología – Hadoop

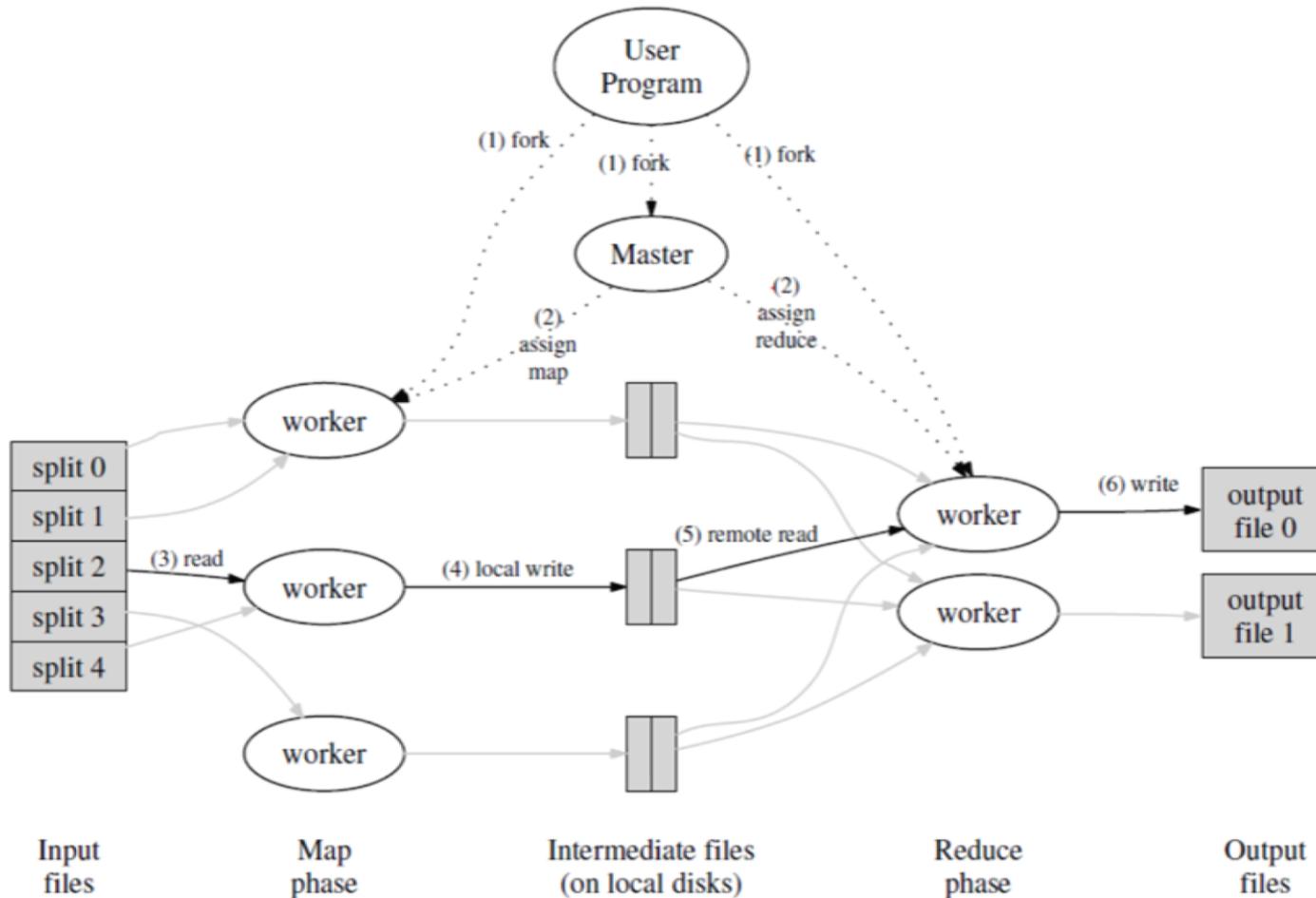
- ▶ **Framework MapReduce:** Framework encargado de ejecutar los programas en el clúster.
 - ▶ **JobTracker:** Servicio encargado de coordinar la ejecución de los trabajos MapReduce.
 - ▶ **TaskTracker:** Servicio encargado de la ejecución de las tareas MapReduce
- ▶ **DFS:** Hadoop DFS (HDFS). Los ficheros se dividen en bloques del mismo tamaño. Los bloques se distribuyen y replican en los diferentes servidores. Sólo funciona en modo “write once”. Está basado en dos servicios:
 - ▶ **NameNode:** Servidor encargado de gestionar el directorio de ficheros y el acceso a los mismos desde los clientes.
 - ▶ **DataNode:** Servidor encargado del almacenamiento y gestión de los bloques que componen los ficheros.

Tecnología – HDFS

HDFS Architecture



Tecnología – MapReduce



Tecnología – Map Reduce

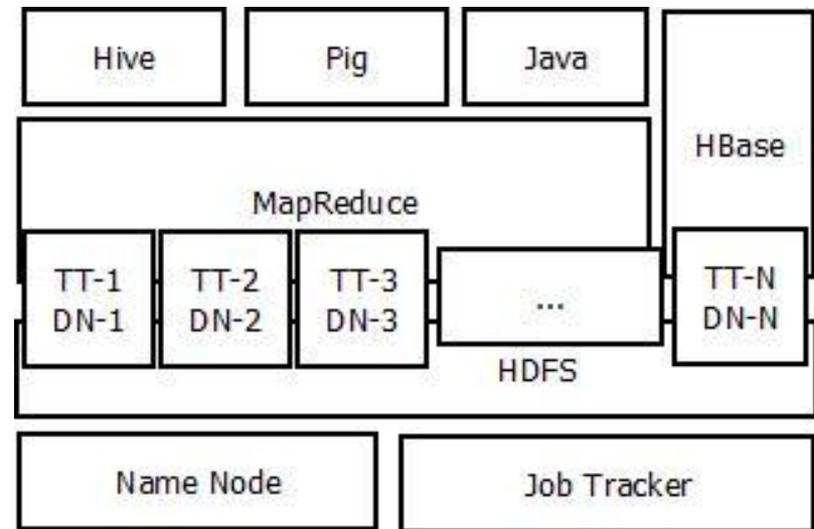
- ▶ Este patrón busca proporcionar un modelo de programación en el cual los desarrolladores no deban preocuparse de los aspectos relativos a la computación distribuida.
- ▶ En este modelo todo programa se debe escribir usando dos funciones:
 - ▶ **Map:** Recibe una entrada (un registro, en forma de tupla) y produce una serie de tuplas (clave, valor) relativas a esa entrada. Una vez acabada la función, el framework agrega las tuplas que comparten la misma clave.
 - ▶ **Reduce:** Recibe una lista con todos los valores generados para una clave dada y produce los resultados finales para esa clave.

Tecnología – Ejemplo

- ▶ **Crear un índice inverso:** Dada una colección de ficheros de texto, generar un índice que muestre cuántas veces aparece la misma palabra en él.
- ▶ El framework de ejecución se encarga de recolectar el contenido de los ficheros y ejecuta la función Map sobre cada línea de ellos.
- ▶ Pseudocódigo:
 - ▶ **Map(String nombreFichero, String contenido):**
para cada palabra (w) en documento (contenido):
ResultadoIntermedio(w, l)
 - ▶ **Reduce(String palabra, Iterator valores):**
int resultado=0;
para todo v en valores: resultado = resultado + v;
Resultado(palabra,resultado);

Tecnología – Ecosistema Hadoop

- ▶ **Apache Hive:** Interfaz SQL de acceso a los datos.
- ▶ **Apache Pig:** Lenguaje de scripting para la manipulación y transformación de datos.
- ▶ **Apache Hbase:** Gestor de base de datos (NoSql) de tipo columnar.
- ▶ **Otros: Sqoop, Flume, ...**



Tecnología – Tiempo real

- ▶ Tratamiento de grandes volúmenes de datos en tiempo real.
 - ▶ Redes de sensores (Internet Of Things)
 - ▶ Gestión de trazas (Clickstream)
- ▶ Dos frameworks de referencia:
 - ▶ Apache Storm
 - ▶ Apache Spark (Streaming)
- ▶ Otros: Apache Flink

Tecnología – Apache Spark



[Download](#)

[Libraries ▾](#)

[Documentation ▾](#)

[Examples](#)

[Community ▾](#)

[FAQ](#)

Spark wins Daytona Gray Sort 100TB Benchmark

We are proud to announce that Spark won the [2014 Gray Sort Benchmark](#) (Daytona 100TB category). A team from [Databricks](#) including Spark committers, Reynold Xin, Xiangrui Meng, and Matei Zaharia, [entered the benchmark using Spark](#). Spark won a tie with the Themis team from UCSD, and jointly set a new world record in sorting.

They used Spark and sorted 100TB of data using 206 EC2 i2.8xlarge machines in 23 minutes. The previous world record was 72 minutes, set by a Hadoop MapReduce cluster of 2100 nodes. This means that Spark sorted the same data 3X faster using 10X fewer machines. All the sorting took place on disk (HDFS), without using Spark's in-memory cache.

Outperforming large Hadoop MapReduce clusters on sorting not only validates the vision and work done by the Spark community, but also demonstrates that Spark is fulfilling its promise to serve as a faster and more scalable engine for data processing of all sizes.

For more information, see the [Databricks blog article](#) written by the Reynold Xin.



Tecnología – Apache Spark

	Hadoop World Record	Spark 100 TB *	Spark 1 PB
Data Size	102.5 TB	100 TB	1000 TB
Elapsed Time	72 mins	23 mins	234 mins
# Nodes	2100	206	190
# Cores	50400	6592	6080
# Reducers	10,000	29,000	250,000
Rate	1.42 TB/min	4.27 TB/min	4.27 TB/min
Rate/node	0.67 GB/min	20.7 GB/min	22.5 GB/min
Sort Benchmark Daytona Rules	Yes	Yes	No
Environment	dedicated data center	EC2 (i2.8xlarge)	EC2 (i2.8xlarge)

* not an official sort benchmark record



Tecnología – Tecnologías relacionadas

▶ Bases de datos NoSql:

- ▶ Apache Cassandra: Gestor de bases de datos NoSql de tipo columnar.
- ▶ Mongo DB: Gestor de bases de datos NoSql de tipo documental (json).

▶ Procesamiento de grafos:

- ▶ Google Pregel, Apache Giraph
- ▶ Neo4j: Gestor de base de datos NoSql

▶ Machine Learning:

- ▶ Apache Mahout
- ▶ SparkML

Tecnología – Distribuciones Comerciales

- ▶ **Hortonworks:**
 - ▶ Distribución más “fiel” a los componentes de Apache.
- ▶ **Cloudera:**
 - ▶ Impala como alternativa a Hive
- ▶ **IBM:**
 - ▶ BigR, Big SQL
 - ▶ DFS propietario alternativo
- ▶ **MapR:**
 - ▶ DFS propietario.
 - ▶ Gestor NoSql propietario.

Tecnología – Líneas futuras

- ▶ **Google Dremel: Google publica en 2010 un artículo describiendo un framework complementario a MapReduce para el tratamiento masivo de datos.**

Dremel: Interactive Analysis of Web-Scale Datasets

Venue

Proc. of the 36th Int'l Conf on Very Large Data Bases (2010), pp. 330-339

Publication Year

2010

Authors

Sergey Melnik, Andrey Gubarev, Jing Jing Long, Geoffrey Romer, Shiva Shivakumar, Matt Tolton, [Theo Vassilakis](#)

Abstract



Dremel is a scalable, interactive ad-hoc query system for analysis of read-only nested data. By combining multi-level execution trees and columnar data layout, it is capable of running aggregation queries over trillion-row tables in seconds. The system scales to thousands of CPUs and petabytes of data, and has thousands of users at Google. In this paper, we describe the architecture and implementation of Dremel, and explain how it complements MapReduce-based computing. We present a novel columnar storage representation for nested records and discuss experiments on few-thousand node instances of the system.

Tecnología – Google Big Query

Enterprise Cloud Data Warehouse

BigQuery is Google's **fully managed, petabyte scale, low cost enterprise data warehouse for analytics**. BigQuery is **serverless**. There is no infrastructure to manage and you don't need a database administrator, so you can focus on analyzing data to find meaningful insights using familiar SQL. BigQuery is a powerful Big Data analytics platform used by all types of organizations, from startups to Fortune 500 companies.



Casos de uso

Social Networks: Sentiment Analysis



Soc. Networks – Sentiment Analysis

- ▶ **Objetivo:** Medir la imagen de marca de la empresa.
- ▶ **Cómo:** Interpretación de los comentarios (ó textos) recogidos de las redes sociales sobre un tema dado.
- ▶ **Escala de valores de referencia:** positivo, neutro y negativo.
- ▶ **Necesario realizar un análisis sintáctico y semántico de los textos.**

Soc. Networks – Sentiment Analysis

► Ejemplos:

NO	ME	CAES	MAL	SENT.
-		-	-	+

ERES	UN	PERRO	PERO	TE	ADORO	SENT.
		-			+	+

Soc. Networks – Sentiment Analysis

- ▶ Twitter: > 500 millones de tuits al día
 - ▶ Textos cortos (140 caracteres).
 - ▶ Referencias explícitas (@ucm)
 - ▶ Temas destacados (#EconómicasBigData)

- ▶ ¿Están sesgadas las redes sociales?.

- ▶ ¿Impacto real en las cifras de la compañía?.

Soc. Networks – Sentiment Analysis

- ▶ Propuesta de tecnología
 - ▶ Librería para el procesamiento de lenguaje: Apache OpenLNP
 - ▶ Procesamiento en diferido: Apache Hadoop + MapReduce
 - ▶ Procesamiento en tiempo real: Apache Storm

Social Networks: Networking & Fraud Detection



Soc. Networks – Fraud Detection



El pueblo que fingía accidentes o cuando estafar a la aseguradora se profesionaliza

MARÍA ZUIL | 13

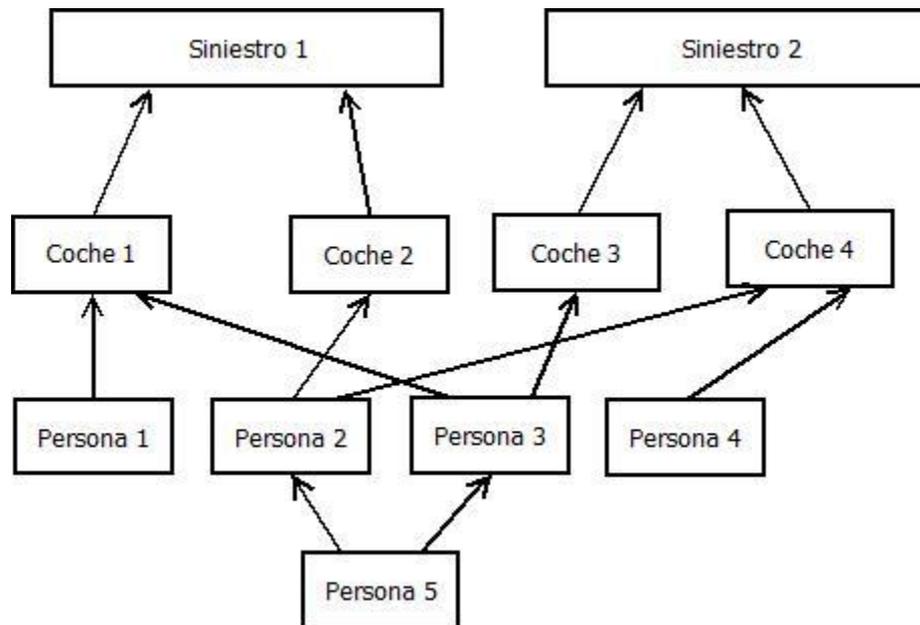
Soc. Networks – Fraud Detection

- ▶ **Desaparición de John Darwin.**
 - ▶ Seguro de vida, falsa desaparición
 - ▶ Mujer involucrada.
 - ▶ Hijos no involucrados.

- ▶ **500.000 libras defraudadas.**

- ▶ **Detectado el fraude por una fotografía tomada en una inmobiliaria.**

Soc. Networks – Fraud Detection



Soc. Networks – Fraud Detection

- ▶ Búsqueda de conjuntos de nodos que cumplan:
 - ▶ Densidad de aristas elevada.
 - ▶ Alto grado de interconexión
- ▶ Situación similar al problema de cliqué máximo
- ▶ Este último es NP-Completo

Soc. Networks – Fraud Detection

▶ Tecnología

▶ Análisis forense:

- ▶ Neo4j
- ▶ IBM Identity Insights
- ▶ SAS
- ▶ R

▶ Automatización:

- ▶ Apache Hadoop
- ▶ Apache Giraph

Internet Of Things: Black Box Insurance



IOT – Black Box

- ▶ ¿Cómo?
- ▶ Recopilación de telemetrías en tiempo real ó en diferido:
 - ▶ Posición
 - ▶ Fecha y hora
 - ▶ Velocidad
 - ▶ Aceleración (ó frenado)
 - ▶ Ángulo de giro
 - ▶ ...

IOT – Black Box

- ▶ Creación de un perfil de conductor (**modelo de scoring**) en base a ciertos indicadores:
 - ▶ % giros bruscos.
 - ▶ % de cambios de velocidad bruscos.
 - ▶ Horario de conducción
 - ▶ Localización (zonas geográficas)
 - ▶ ...
- ▶ Cálculo de la prima en base a los parámetros usuales junto con el perfil calculado anteriormente.

IOT – Black Box

▶ Tecnología:

- ▶ Apache Hadoop: Almacenamiento persistente.
- ▶ Apache Storm / Spark Streaming: Tratamiento en tiempo real.
- ▶ Apache Spark: Cálculo del perfil del conductor.

IOT – Black Box

- ▶ **Productos comercializados en España:**
 - ▶ MAPFRE YCar.
 - ▶ Verti Cuenta Kms.
 - ▶ Generali “Pago como conduzco”.
 - ▶ Zurich “Seguro inteligente”.

Internet Of Things: Healthcare



IOT - Healthcare



Comentarios

Enlaces – Formación gratuita

- ▶ <http://www.mmds.org/>
- ▶ MOOC's:
 - ▶ Coursera: <http://www.coursera.org>
 - ▶ Edx: <http://www.edx.org>
 - ▶ Big Data University (IBM): <http://www.bigdatauniversity.com>
- ▶ Documentación oficial de Apache.

Muchas gracias!