

Big Data

Dr. Wenceslao Palma
wenceslao.palma@pucv.cl



www.theclinic.cl/2017/01/19/ma

MENÚ THE CLINIC

Martin Hilbert, experto en redes digitales: “Obama y Trump usaron el Big Data para lavar cerebros”

77 COMENTARIOS

Daniel Hopenhaym | 19 Enero, 2017 |
Tags: big data, Estados Unidos, Martin Hilbert, obama, redes digitales, Trump

El artículo discute el uso de Big Data en campañas políticas, específicamente mencionando a Obama y Trump, y cómo esto puede ser utilizado para influir en la opinión pública.

“Hay gente que cree que no se ha encontrado el remedio para el cáncer porque perjudicaría a las farmacéuticas”

Este artículo trata sobre el uso de Big Data en el ámbito de la salud y la industria farmacéutica, mencionando que algunas personas creen que no se ha encontrado un tratamiento para el cáncer debido a intereses económicos.



ciperchile.cl/2017/02/13/estudio

CIPER Chile

13 Febrero, 2017 • 13 comentarios

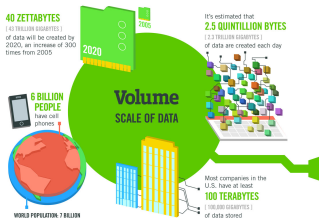
Estudio prevé que el 50% de los trabajadores chilenos será reemplazado por máquinas



Compartir Publicar en 3 + 1 El Correo

Este artículo de CIPER Chile presenta un estudio que predice que el 50% de los trabajadores chilenos serán reemplazados por máquinas. La imagen muestra varios robots humanoides en un entorno de museo o exposición.

Big Data: the four V's



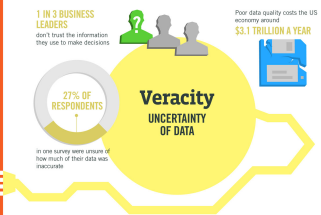
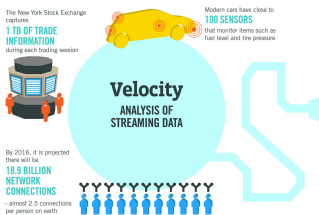
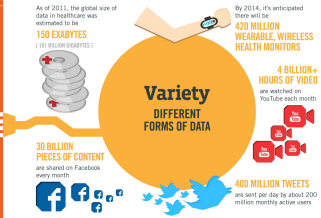
The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

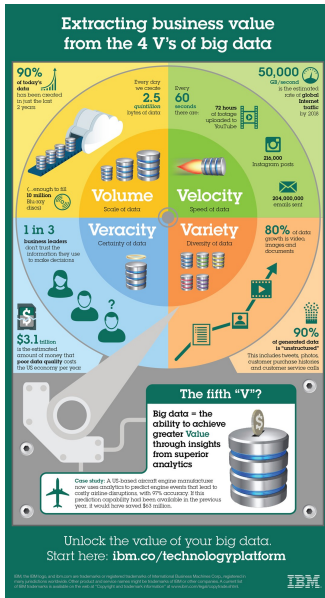
By 2015, **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.0 million in the United States



Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEFPEC, SAS



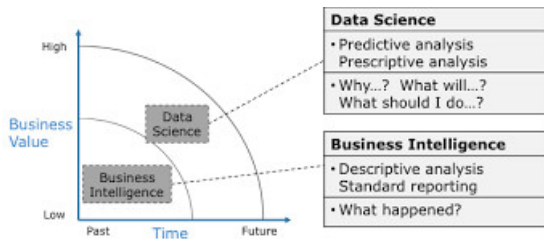
Big Data: the five V's



Big Data as stated by Cesar Hidalgo

- A lot of people actually are confused between Big Data and a lots of data.
- Big Data has to be big in three different ways: size, resolution and scope.

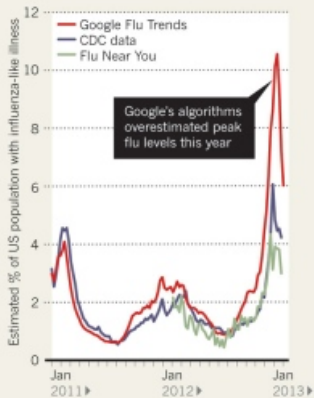
Big Data Mining: data science - business intelligence



Big Data Mining: Google Flu Trends

FEVER PEAKS

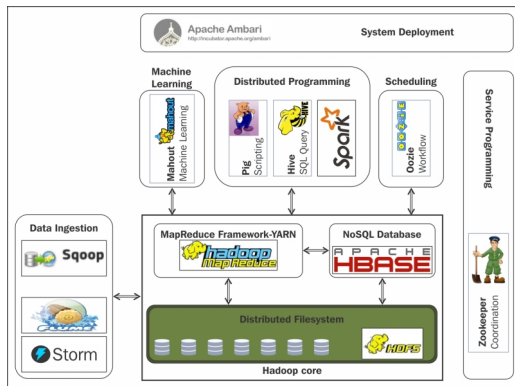
A comparison of three different methods of measuring the proportion of the US population with an influenza-like illness.



Big Data: how to tackle it?

- The only feasible approach to tackling large-data problems today is to divide and conquer.
- The general principles behind divide-and-conquer algorithms are broadly applicable to a wide range of problems in many different application domains.
- There are many issues that need to be addressed:
 - How to organize the data store?
 - How to break up a large problem into smaller tasks?
 - How to assign tasks across a potentially large number of computer nodes.?
 - How to coordinate synchronization among the different nodes?
 - How to share partial results from one node that is needed by another?
 - How do we accomplish all of the above in the face of software errors and hardware faults?

Big Data Mining: the hadoop ecosystem



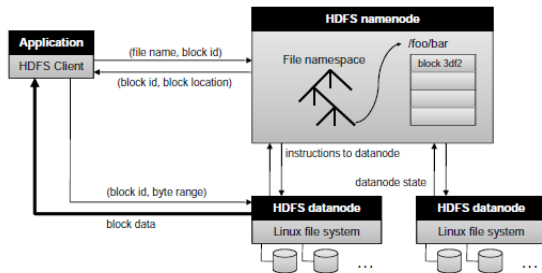
- Chemistry before the tubes.

Big Data: Why worry about foundations?

- Chemistry before the tubes.
- Computer science before big data mining tools.

HDFS

Hadoop Distributed File System (HDFS) is the primary storage system used by Hadoop applications. HDFS creates multiple replicas of data blocks and distributes them on compute nodes throughout a cluster to enable reliable, extremely rapid computations.



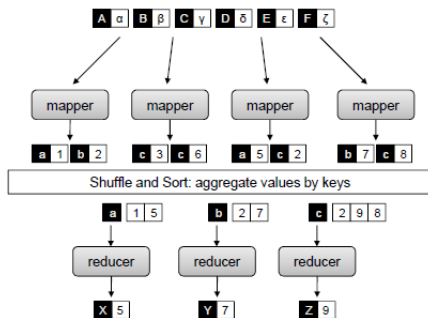
- Data replication makes the system more fault tolerant.
- Data replication provides scalability (w.r.t. data access).
- Data replication and partitioning provide high concurrency.

- Data replication makes the system more fault tolerant.
- Data replication provides scalability (w.r.t. data access).
- Data replication and partitioning provide high concurrency.

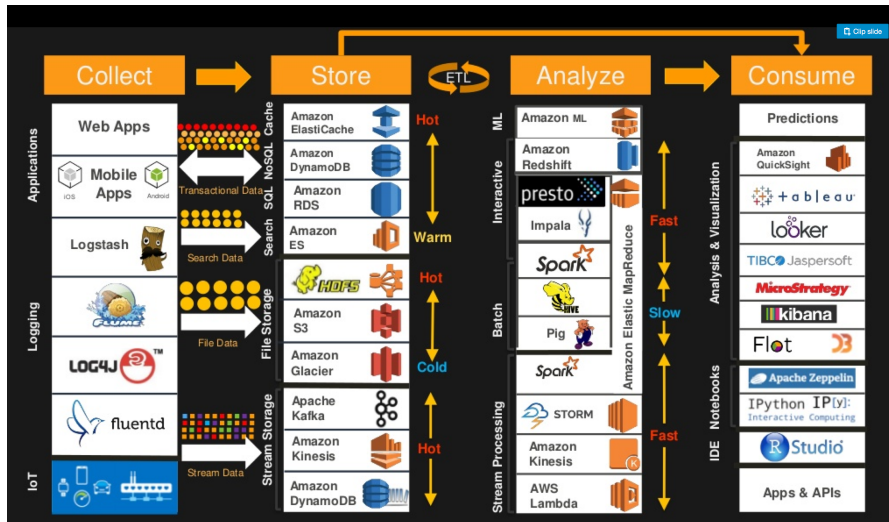
The problem with replication is that is hard to maintain data consistency over the time but in big data systems, the data is written once and the updates are stored as additional data sets over the time.

Big Data: MapReduce

MapReduce is a programming model for data processing introduced by Google (2004) to support parallel and fault-tolerant computations over large data sets on clusters of computers. It provides an abstraction that hides many system-level details from the programmer.



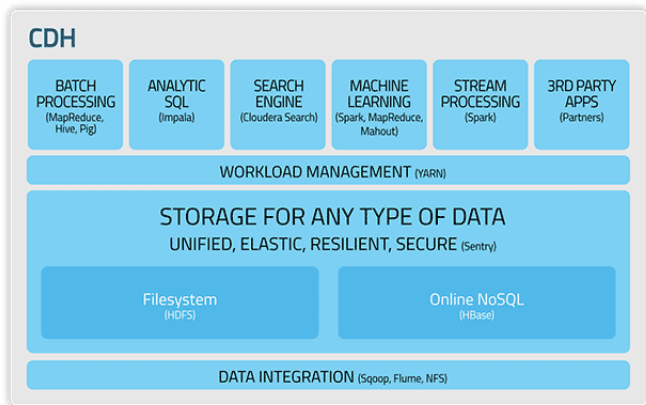
Big Data: Cloud Services



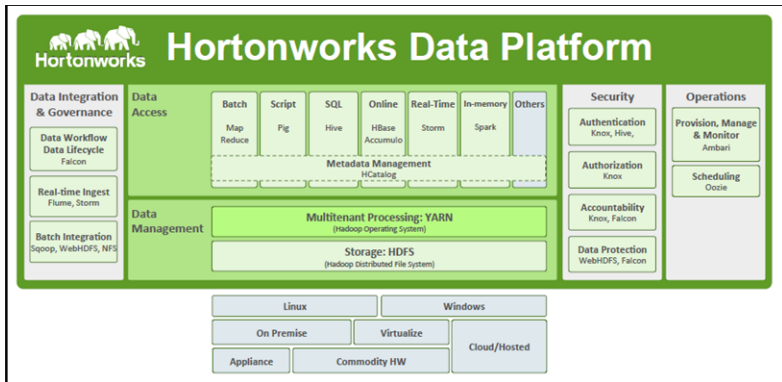


Built for developers. Deploy an SSD cloud server in 55 seconds!

Big Data: Hadoop ecosystem providers



Big Data: Hadoop ecosystem providers



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative



PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any visualization tools e.g. Flare, D3.js, Tableau