Universidad Central de Venezuela
Facultad de Ciencias
Escuela de Computación

Tema 1: Introducción a la Ciencia de Datos

Prof. Jesús Lares

jesus.lares@ciens.ucv.ve

@cienciadedatos

+58 414 331.43.99

Agenda

- Big Data
- Científico de Datos
- Diagrama de Venn
- Arquitectura de Big Data
- Apache Hadoop
- Escenario: Fuentes de datos
- Casos de uso de Big Data
- Tendencias en Big Data

Agenda

- Big Data
- Científico de Datos
- Diagrama de Venn
- Arquitectura de Big Data
- Apache Hadoop
- Escenario: Fuentes de datos
- Casos de uso de Big Data
- Tendencias en Big Data

Big Data: Definición desde el punto de vista de capacidad de cómputo

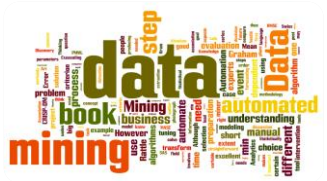
Big Data:

Es una plataforma tecnológica (hardware y software) que permite almacenar (de manera distribuida) y procesar (en forma paralela y distribuida) conjuntos de datos, que por su gran volumen, superan las capacidades de las plataformas TI tradicionales, ya sea porque tomaría demasiado tiempo procesar dichos datos o porque sería muy costoso implementar una arquitectura que soporte tal cantidad de datos.

Big Data: Definición desde el punto de vista científico (Ciencia de Datos)

Ciencia de Datos:

Es la generación de conocimiento a partir de grandes volúmenes de datos, aplicando técnicas de procesamiento paralelo y distribuido, para implementar algoritmos que permitan predecir o detectar patrones sobre los datos almacenados. A partir de los resultados obtenidos se podrán construir herramientas que permitan analizar los resultados y apoyar los procesos de toma de decisiones.



Big Data: Unidades de Medida

Unidades de Medidas

IEC prefixes and symbols for binary multiples				
Factor	Name	Symbol	Origin	SI Derivation
2^{10}	kibi	Ki	kilobinary $(2^{10})^1$	kilo: $(10^3)^1$
2^{20}	mebi	Mi	megabinary $(2^{10})^2$	mega: $(10^3)^2$
2^{30}	gibi	Gi	gigabinary $(2^{10})^3$	giga: $(10^3)^3$
2^{40}	tebi	Ti	terabinary $(2^{10})^4$	tera: $(10^3)^4$
2^{50}	pebi	Pi	petabinary $(2^{10})^5$	peta: $(10^3)^5$
2^{60}	exbi	Ei	exabinary $(2^{10})^6$	exa: $(10^3)^6$
2^{70}	zebi	Zi	zettabinary $(2^{10})^7$	zetta: $(10^3)^7$
2^{80}	yobi	Yi	yottabinary $(2^{10})^8$	yotta: $(10^3)^8$

International System of Units (SI)

Big Data: V's


5 V's Tradicionales
Otras V's
Valor
Utilidad
Impacto Social
Impacto Económico
Visualización
Despliegue
Cuadro de Mandos
Variabilidad
Generación
Cambios
Visión
Data-Driven
Cultura del Dato
Volumen
Grandes volúmenes
Velocidad
Generación de datos
Procesamiento
Variedad
Diversos Formatos
Data Stream
Veracidad
Calidad de datos
Precisión

Agenda

- Big Data
- Científico de Datos
- Diagrama de Venn
- Arquitectura de Big Data
- Apache Hadoop
- Escenario: Fuentes de datos
- Casos de uso de Big Data
- Tendencias en Big Data

Científico de Datos

Profesional que debe provenir y dominar las ciencias matemáticas y la estadística, además de tener avanzados conocimientos en programación (y sus múltiples lenguajes), ciencias de la computación y analítica predictiva.

Profesional dedicado a analizar e interpretar grandes almacenes o bases de datos

Científico de Datos

- Conocimientos en Matemáticas
- Conocimientos en Estadística
- Conocimientos en Ciencias de la Computación
- Conocimientos en Programación
- Conocimientos en Base de Datos
- Conocimientos en Analítica Predictiva
- Conocimientos en Big Data

Científico de Datos

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th 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

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

DOMAIN KNOWLEDGE & SOFT SKILLS

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

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 of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization: customer tracking and on-site analytics: predictive analytics and econometrics: data warehousing and big data systems: marketing channel insights in Paid Search, SEO, Social, CRM and brand.

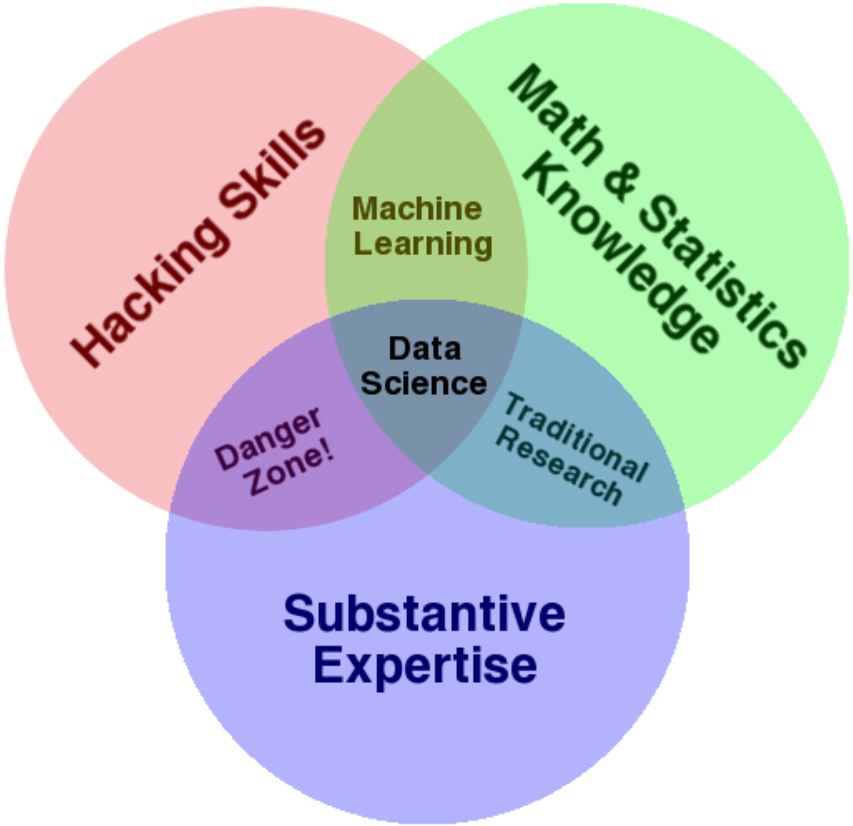
Marketing DISTILLERY

Científico de Datos: Trabajo en Equipo

Agenda

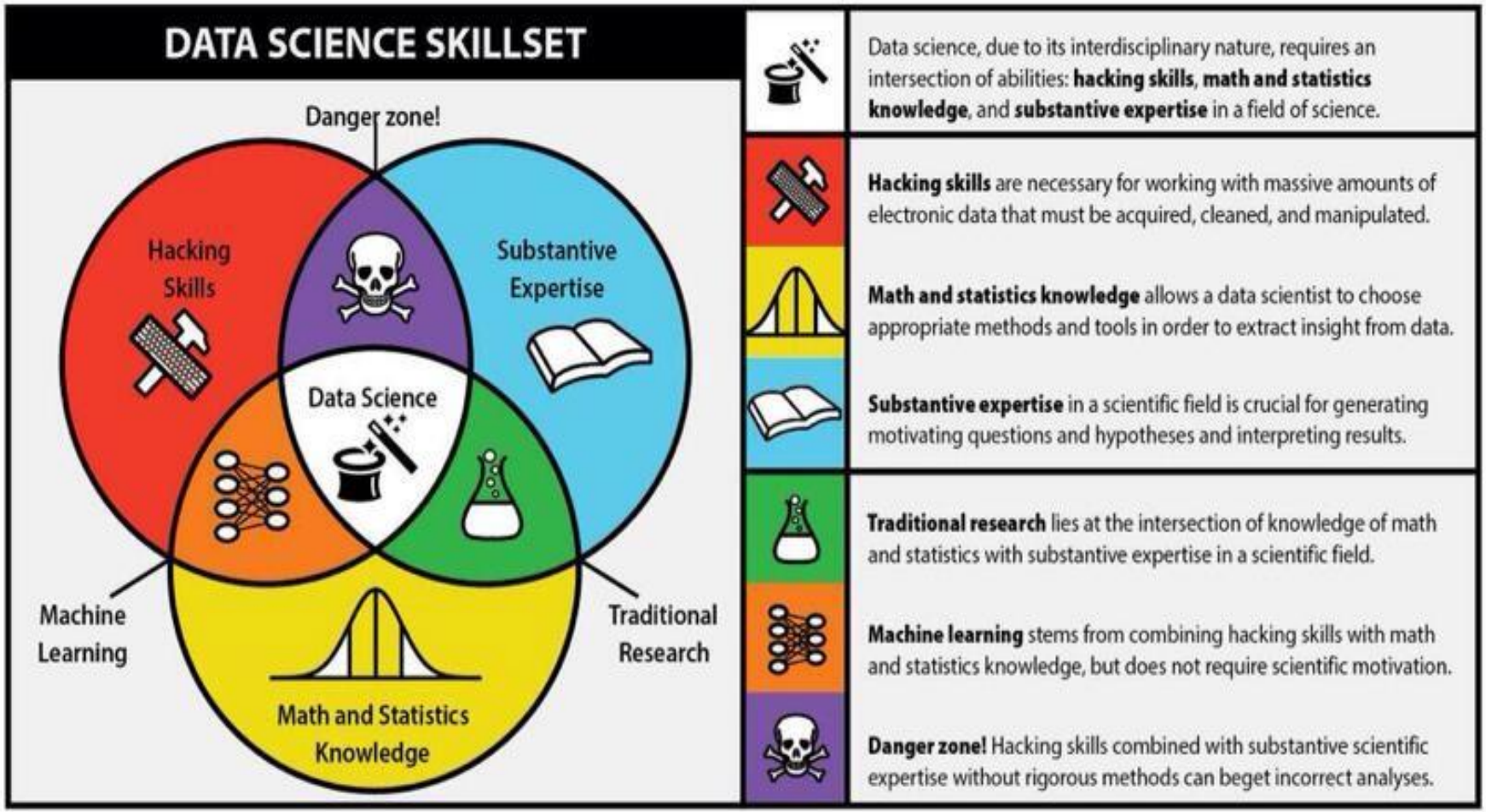
- Big Data
- Científico de Datos
- Diagrama de Venn
- Arquitectura de Big Data
- Apache Hadoop
- Escenario: Fuentes de datos
- Casos de uso de Big Data
- Tendencias en Big Data

Diagrama de Venn



Fuente: Drew Conway

Diagrama de Venn



Agenda

- Big Data
- Científico de Datos
- Diagrama de Venn
- **Arquitectura de Big Data**
- Apache Hadoop
- Escenario: Fuentes de datos
- Casos de uso de Big Data
- Tendencias en Big Data

Hardware: Cluster Maestro-Esclavo

Comienza simple

Escala a medida que lo necesites

Nodos

- RAM: >= 4 GB
- CPU: >= 2 Cores
- HDD: >= 512 GB

Escalabilidad Horizontal

- Desempeño proporcional al tamaño de la Plataforma
- Compatibilidad con la **Nube**

Software: Open Source

Agenda

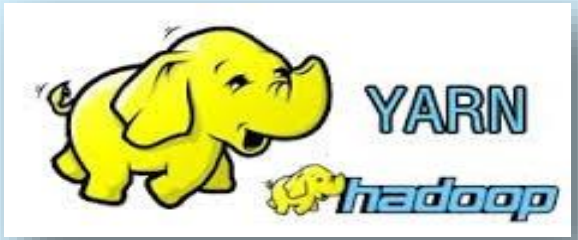
- Big Data
- Científico de Datos
- Diagrama de Venn
- Arquitectura de Big Data
- **Apache Hadoop**
- Escenario: Fuentes de datos
- Casos de uso de Big Data
- Tendencias en Big Data

Apache Hadoop: Antecedentes

Antecedentes Apache Hadoop

Publicaciones de Google (Papers):

- Google File System (2003)
- MapReduce (2004)
- Bigtable (2006)



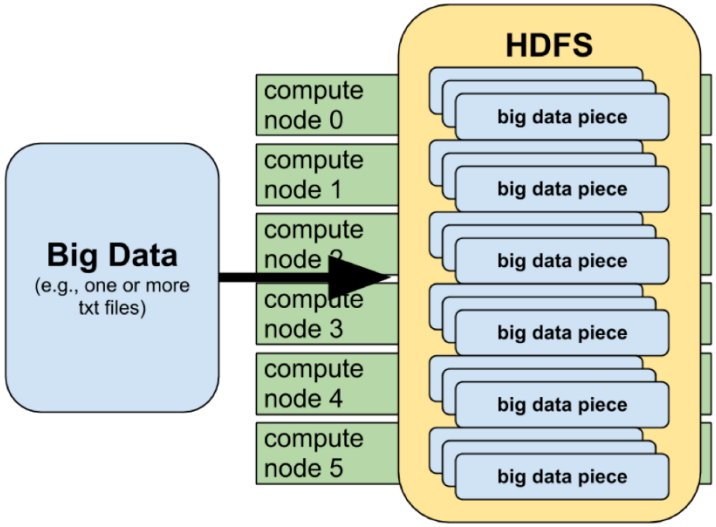
Doug Cutting (2006)

Apache Hadoop

Plataforma de software de código abierto para el almacenamiento distribuido y procesamiento distribuido de grandes volúmenes de datos (Big Data) en clusters de ordenadores construidos a partir de *"Commodity hardware"*.

Los servicios de Hadoop proporcionan almacenamiento de datos, procesamiento de datos, acceso a datos, la gestión de datos, seguridad y operaciones.

Apache Hadoop: HDFS

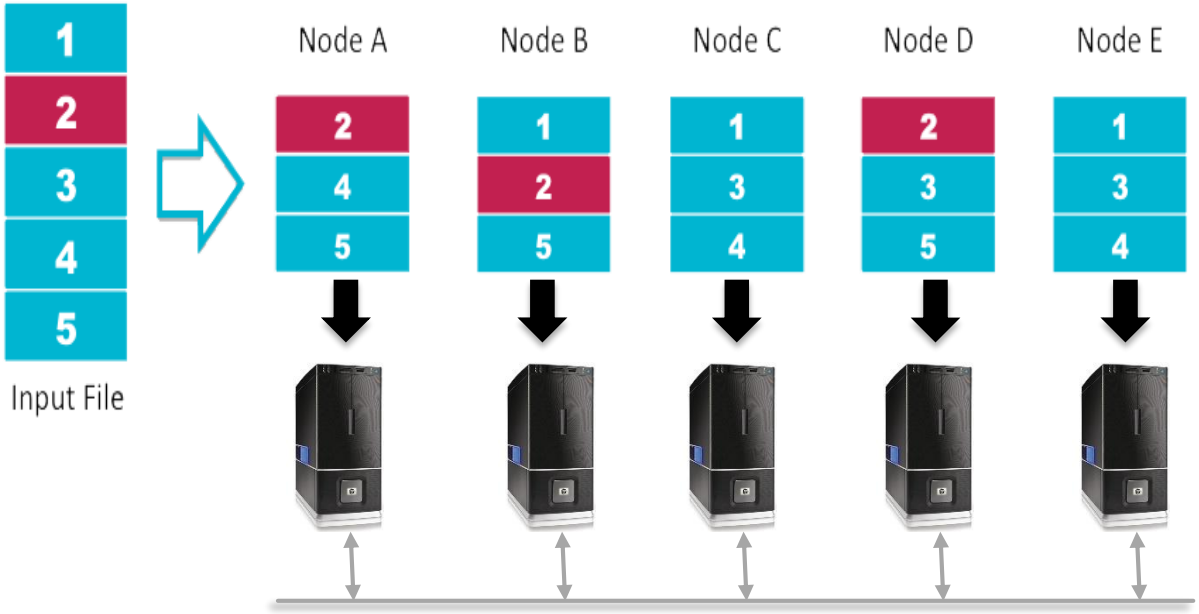


HDFS: Hadoop Distributed File System

- Almacenamiento Distribuido
- Soporta grandes volúmenes de datos
- Replicación de datos
- Soporta escalabilidad horizontal
- Integración con diversas plataformas
- Integración con almacenes de datos
- Código abierto (Open Source)

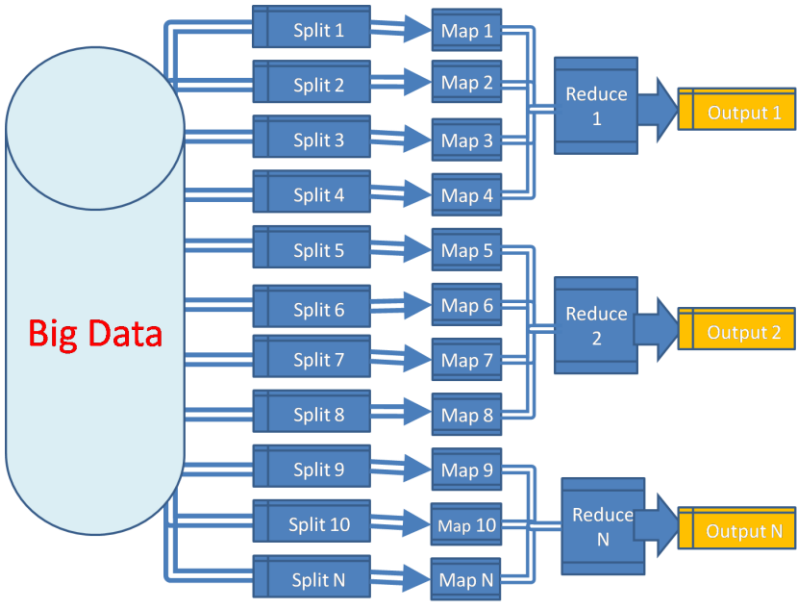
Apache Hadoop: HDFS

HDFS Data Distribution



- Bloques (64 MB o 128 MB)
- Replicación (3 por defecto)
- Tolerancia a fallos

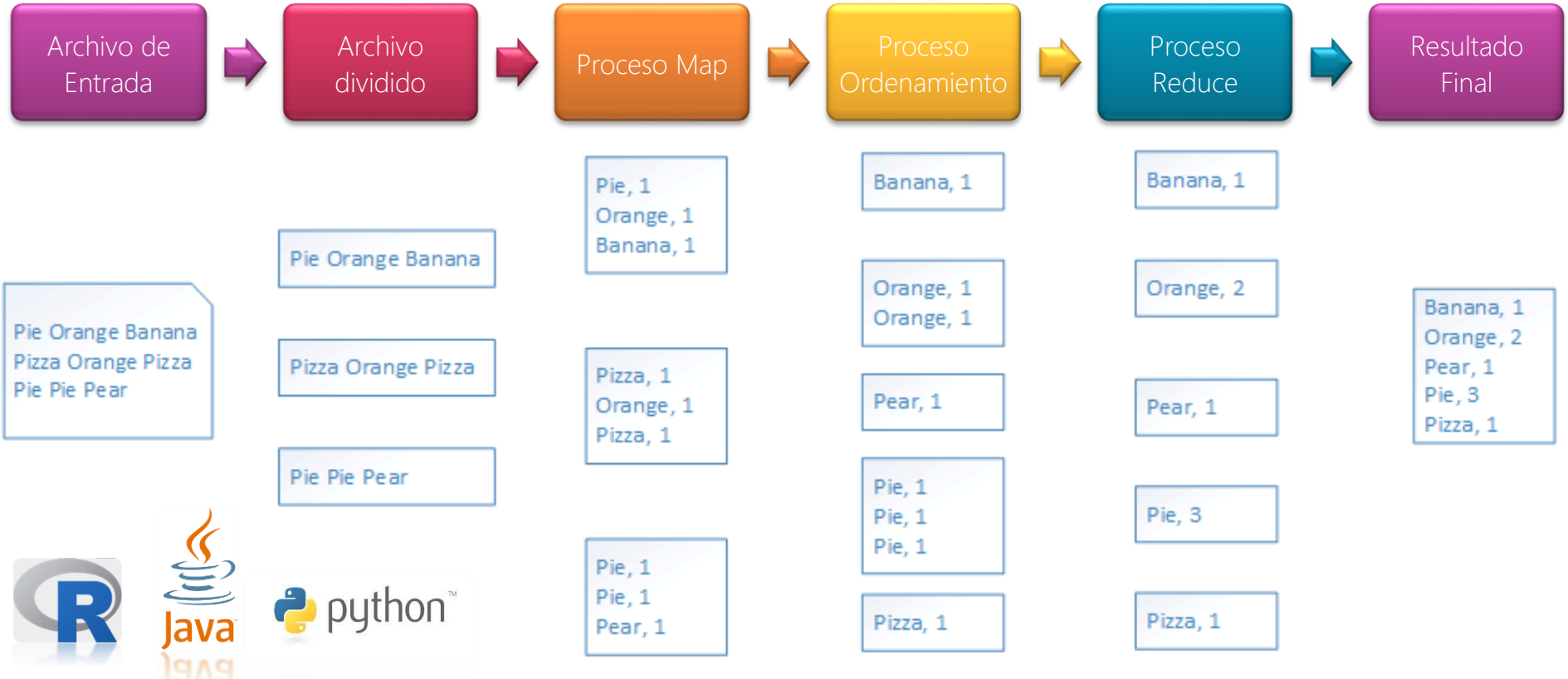
Apache Hadoop: MapReduce



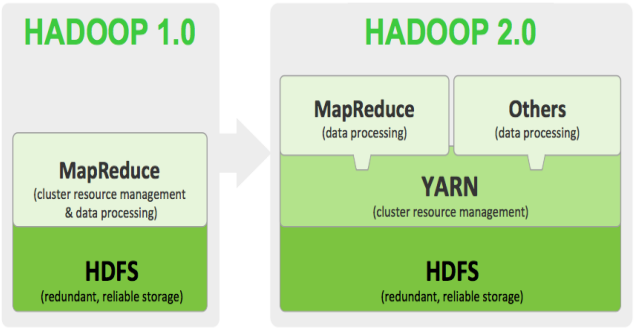
- Procesamiento paralelo y distribuido
- Eficiente sobre grandes volúmenes de datos
- Multilenguaje: Java, R, Python
- Esquema Divide y Vencerás
- Función Mapeo (Map)
- Función de Reducción (Reduce)
- Cambio de Paradigma de programación

Apache Hadoop: MapReduce

Ejemplo: Conteo de palabras



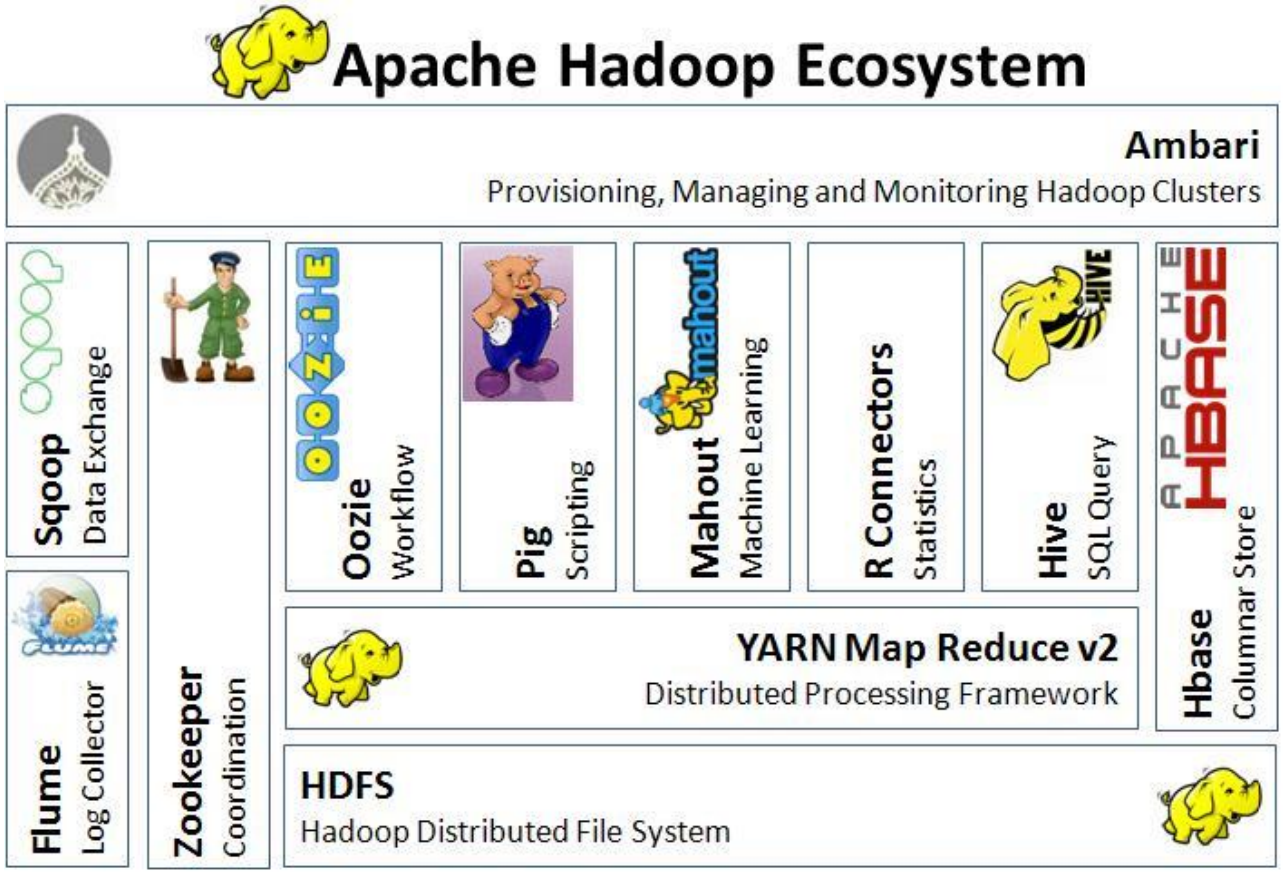
Apache Hadoop: YARN



Características	Descripción
Tenencia Múltiple (Multi-tenancy)	Permite que múltiples motores de acceso utilicen Hadoop como el estándar común para los procesamientos en lotes, interactivo y en tiempo real, permitiendo el acceso simultáneamente el mismo conjunto de datos.
Utilización de Cluster	Asigna dinámicamente los recursos del cluster
Escalabilidad	Soporta la escalabilidad o creciemiento del Cluster
Compatibilidad	Mantiene la compatibilidad hacia atras de los procedimientos MapReduce de las versiones anteriores de Hadoop

YARN: Yet Another Resource Negotiator

Apache Hadoop: Ecosistema

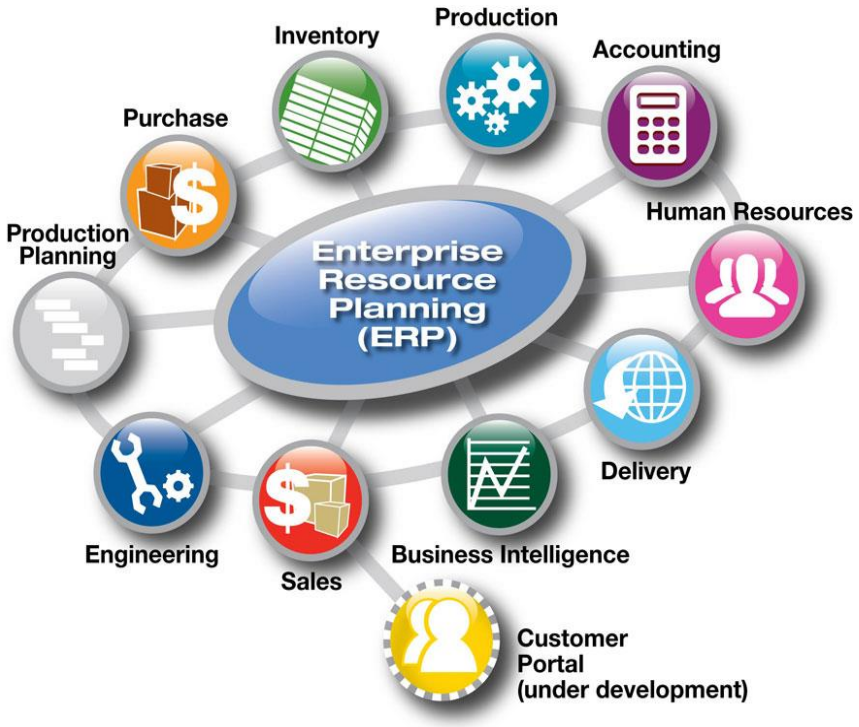


Agenda

- Big Data
- Científico de Datos
- Diagrama de Venn
- Arquitectura de Big Data
- Apache Hadoop
- **Escenario: Fuentes de datos**
- Casos de uso de Big Data
- Tendencias en Big Data

Escenario: Fuentes de datos

Integración de fuentes de datos empresariales



- Diversas tecnologías: Sistemas, BD
- Diversos formatos de datos
- Grandes volúmenes de datos
- Problemas de integración de datos
- Problemas de Calidad de Datos
- Complejidad en la generación de reportes
- Consultas ineficientes
- Esquemas de almacenamiento ineficientes

Escenario: Fuentes de datos

Grandes volúmenes de datos

Esquemas de almacenamiento ineficientes

Sistema de archivos distribuidos

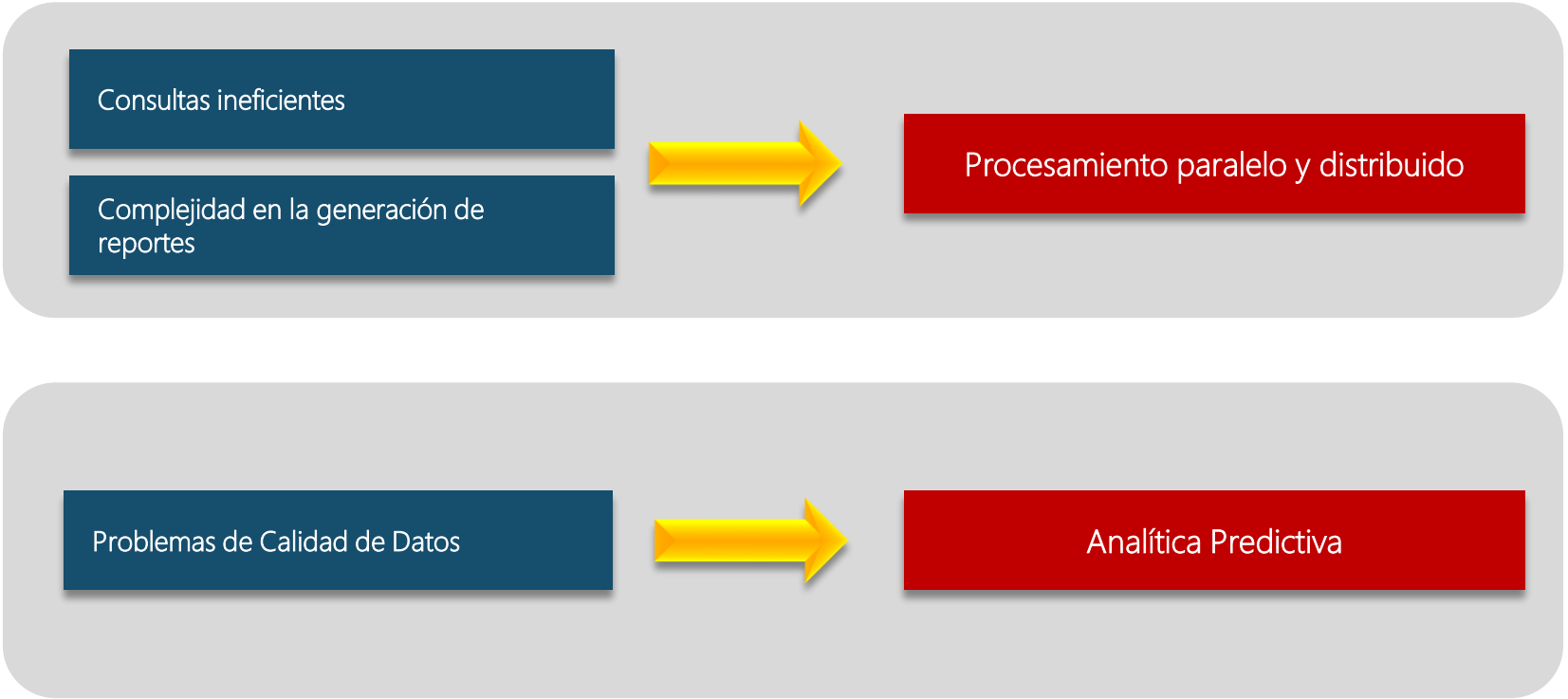
Diversas tecnologías: Sistemas, BD

Problemas de integración de datos

Diversos formatos de datos

Base de Datos NoSQL

Escenario: Fuentes de datos



Escenario: Fuentes de datos

Procesamiento paralelo y distribuido

Base de Datos NoSQL

Analítica Predictiva

Sistema de archivos distribuidos

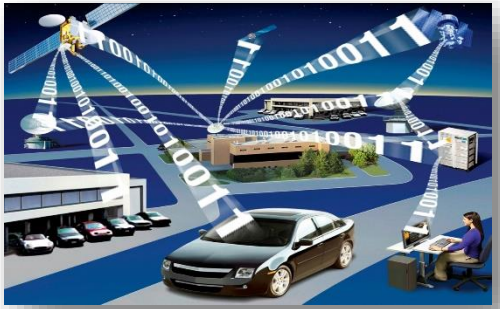
Agenda

- Big Data
- Científico de Datos
- Diagrama de Venn
- Arquitectura de Big Data
- Apache Hadoop
- Escenario: Fuentes de datos
- Casos de uso de Big Data
- Tendencias en Big Data

Casos de uso de Big Data

Redes Sociales

Internet de las cosas



Investigaciones científicas

Patrones de Compra

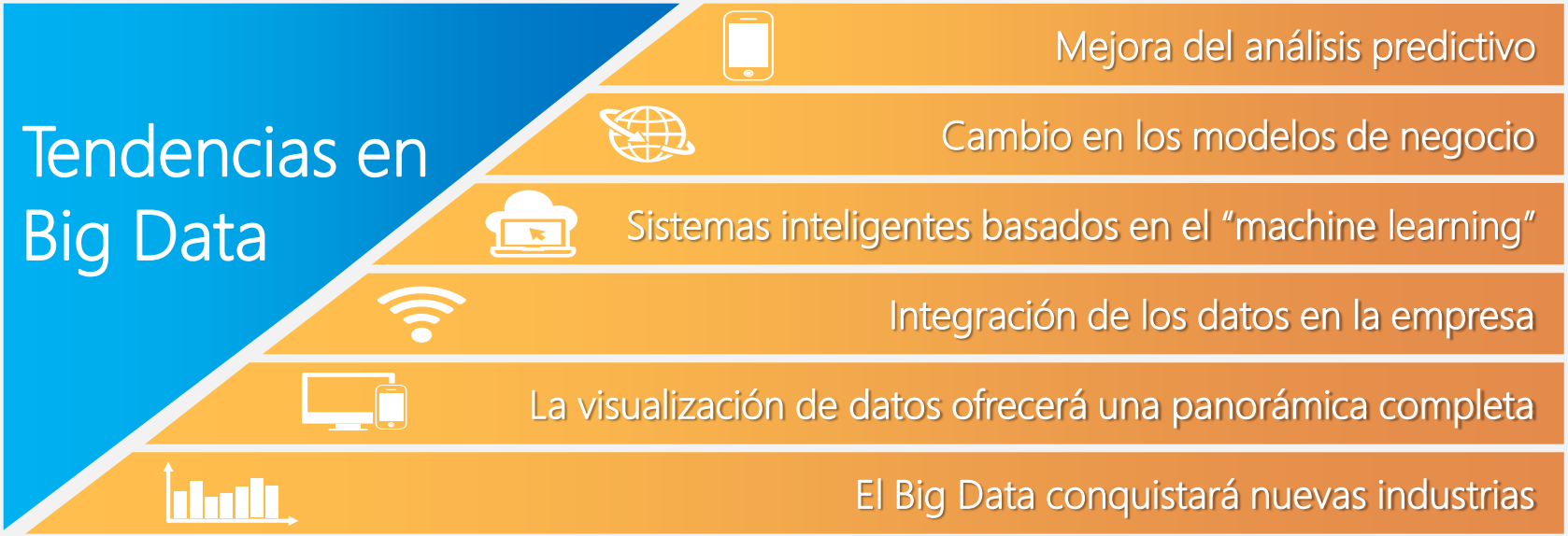
Detección de Fraudes

Ataques informáticos

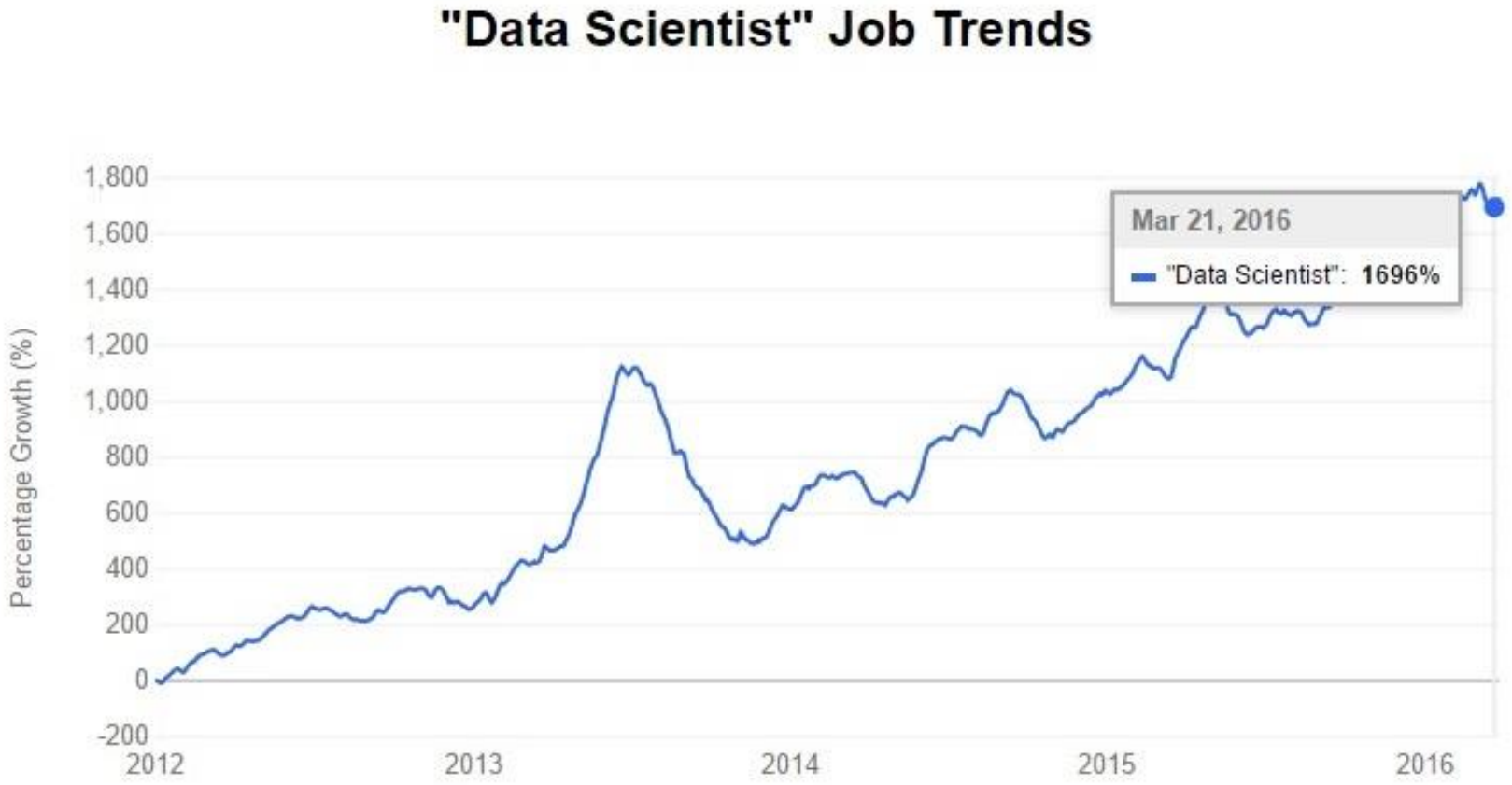
Agenda

- Big Data
- Científico de Datos
- Diagrama de Venn
- Arquitectura de Big Data
- Apache Hadoop
- Escenario: Fuentes de datos
- Casos de uso de Big Data
- Tendencias en Big Data

Tendencias en Big Data

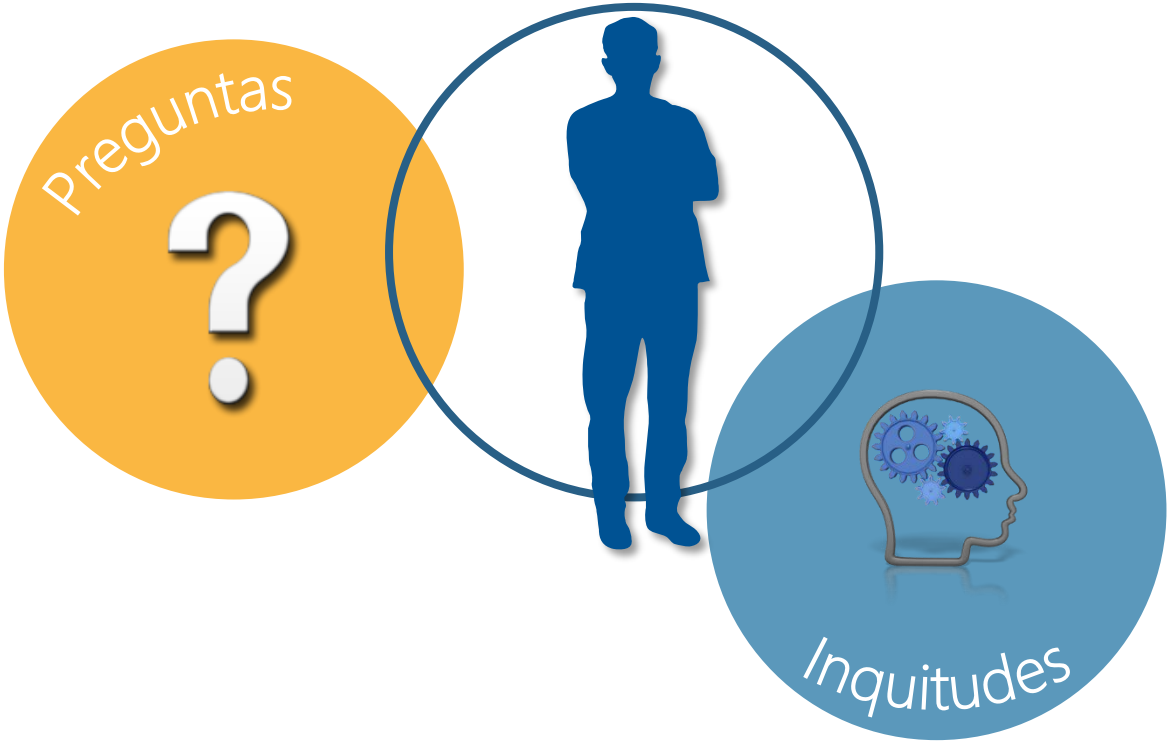


Tendencias en Big Data



Fuente: www.indeed.com

Preguntas/Inquietudes




شكراً جزيلاً
ngiyabonga
рахмат
danke
謝謝
teşekkür ederim
Баярлалаа
mersi
kia ora
barka
welalin
spas
tack
mahalo
tapadh leat
спасибо
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
thank you
gracias
kiitos
dankie
nanni
nandri
dhanyavad
хвала
asante
manana
bedankt
bayarlalaa
gràcie
hvala
mauruuru
köszönöm
akun dankon
aciū
obrigada
murakoze
tenki
enkosi
djiere dieuf
tau
mochchakkeram
дякую
mamnun
dziękuję
chnorakaloutioun
gratias ago
gràcies
sulpáy
taiku
go raibh maith agat
chokrane
sobodi
dėkuji
obrigado
sukriya
kop khun krap
sagolun
arigatô
takk
dakujem
trugarez
mèsi
didi madloba
ありがとう
tanemirt
grazie
najis tuke
terima kasih
rahmet
ευχαριστώ
diolch
dhanyavadagalu
shukriya
mercé
мерси
kam sah hamnida
rahmat
তোমাকে ধন্যবাদ
감사합니다
xièxie
merci