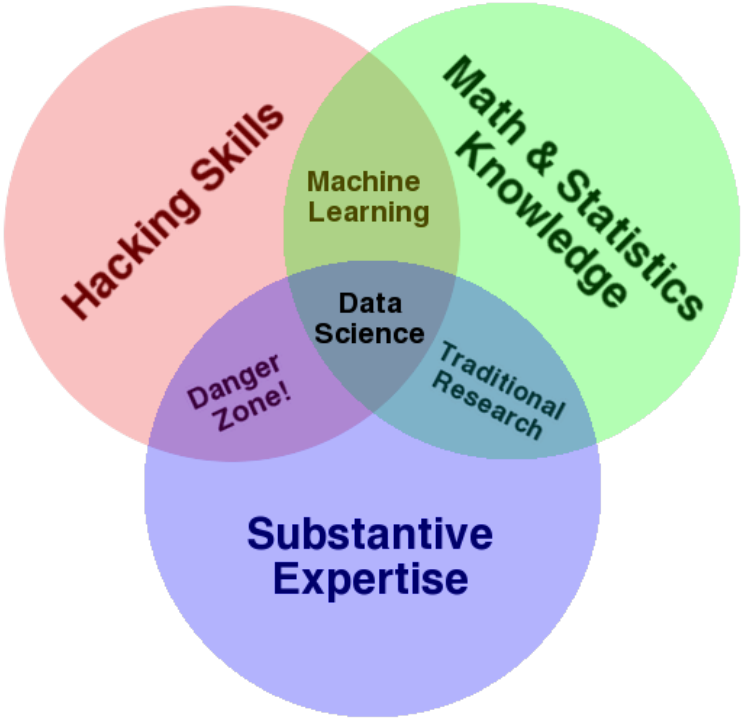
Big Data & Machine Learning

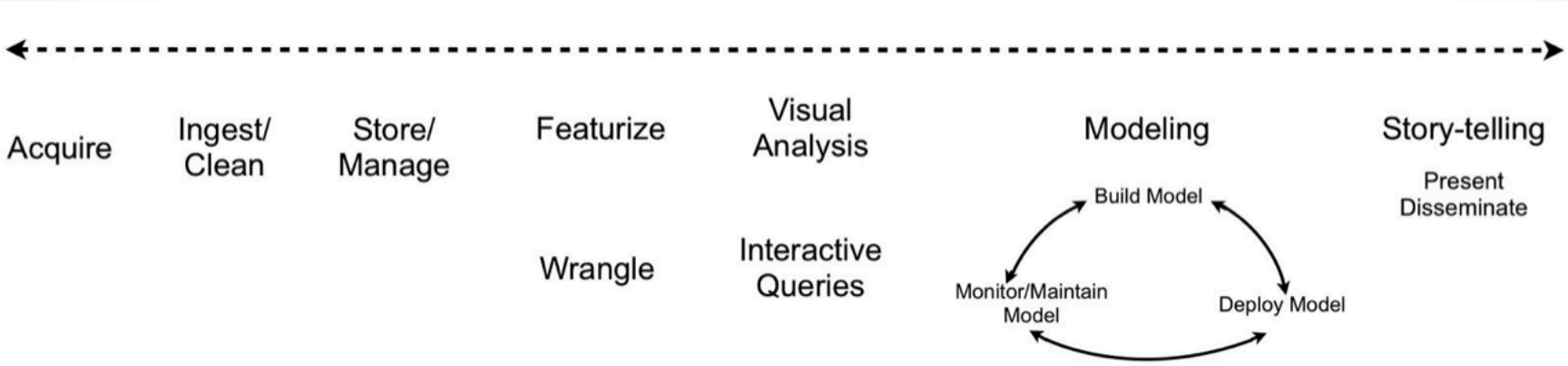
MSc. Ing. Máximo Gurméndez
Universidad de Montevideo

¿Qué es Big Data?

¿Qué es Machine Learning?

¿Qué es Data Science?





Ejemplo: Predecir origen de artículos

República.com.uy

EL PAIS.com.uy

¿QUÉ DIARIO LO ESCRIBIÓ ?

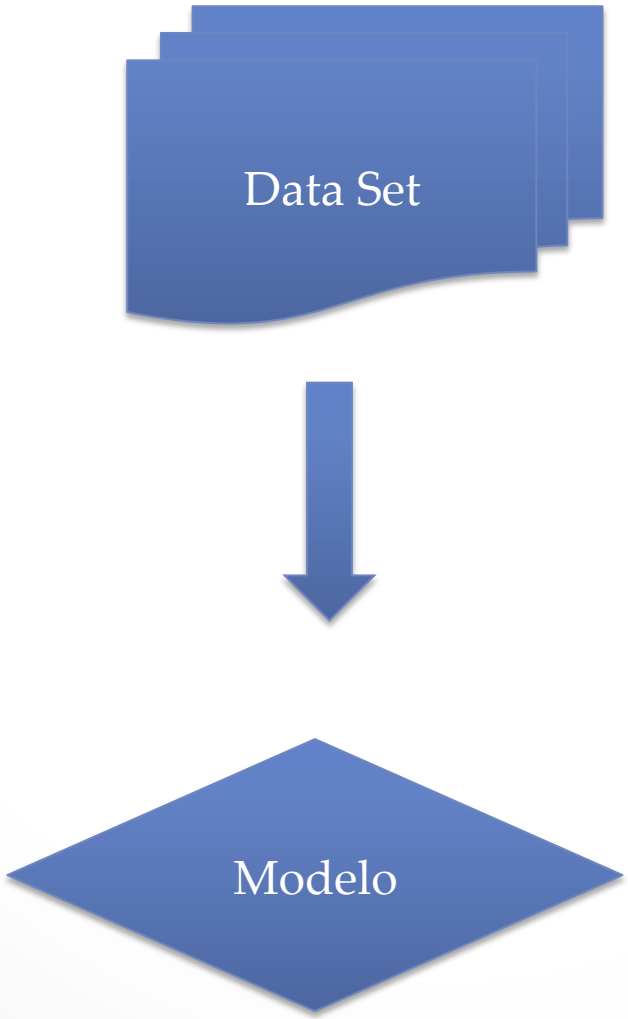
ARTÍCULO X

Armamos un Data Set

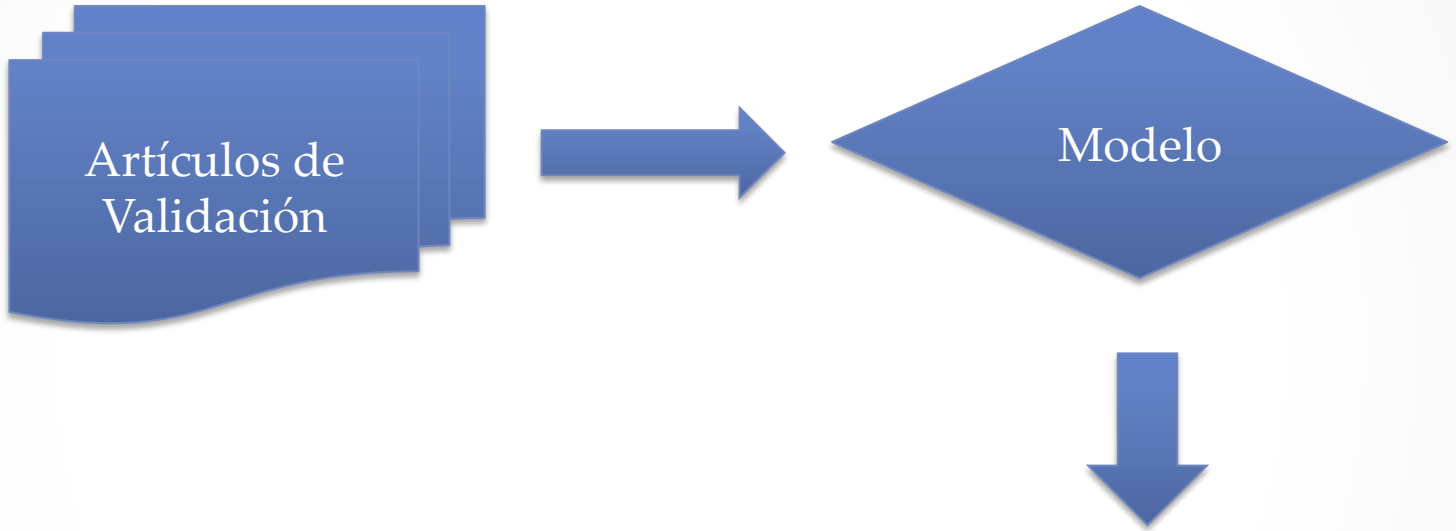
CRAWLER

Data Set

Generamos un Modelo



Validamos el Modelo

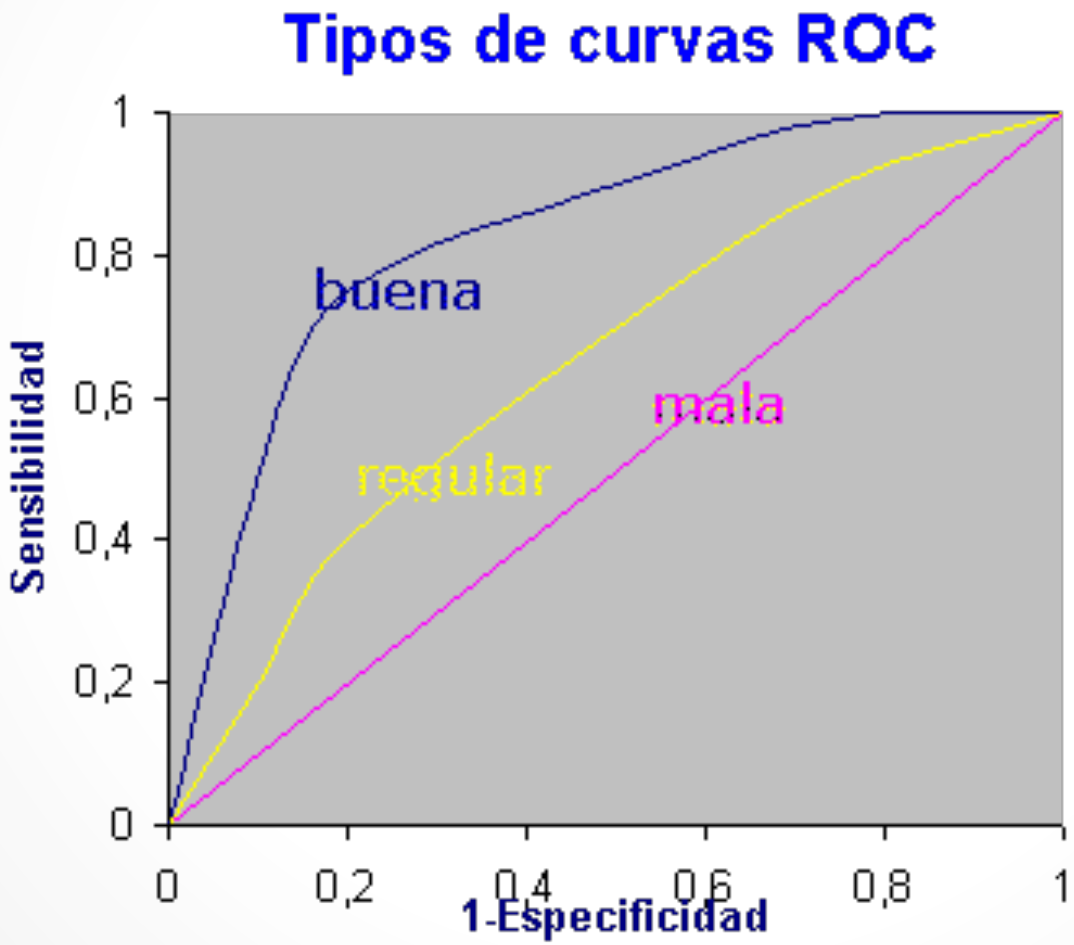


Origen de Artículo	LR	EP	LR	EP	LR	EP	LR
Predicción:	LR	EP	EP	EP	LR	LR	LR
Precisión 71%	✓	✓	✗	✓	✓	✗	✓

Matriz de Confusión

Diario	Aciertos	Errores	Precisión
El Pais	885	249	78%
La República	2521	201	92%
Precisión			88% (ROC: 0.85)

Curva ROC


Tipos de curvas ROC
Sensibilidad
1
0,8
0,6
0,4
0,2
0
buena
regular
mala
0
0,2
0,4
0,6
0,8
1
1-Especificidad

Modelo

- Algoritmo:
  - Support Vector Machines (SVM)
- Atributos:
  - n-gramas de las palabras más frecuentes
- Sentiment Analysis:
  - Identificar palabras positivas / negativas
  - Identificar palabras asociadas a los partidos políticos

Proceso

- Selección de los datos
- Pre-procesamiento
- Transformación
- Learning
- Interpretación / Evaluación

¿Cómo funciona?

- Ejemplo:

  Predecir el interés que esta charla pueda tener en la audiencia.

Tipo	Edad	Perfil	Interesa_charla
ESTUDIANTE	< 21	TÉCNICO	SI
TRABAJADOR	< 21	EMPRESARIAL	SI
ESTUDIANTE	< 21	EMPRESARIAL	NO
ESTUDIANTE	< 21	ACADÉMICO	SI
ESTUDIANTE	> 21 < 25	EMPRESARIAL	SI
TRABAJADOR	> 21 < 25	ACADÉMICO	NO
ESTUDIANTE	> 21 < 25	EMPRESARIAL	NO
TRABAJADOR	> 21 < 25	EMPRESARIAL	NO
ESTUDIANTE	> 21 < 25	ACADÉMICO	NO
ESTUDIANTE	> 25	TÉCNICO	SI
TRABAJADOR	> 25	TÉCNICO	NO
ESTUDIANTE	> 25	TÉCNICO	SI
ESTUDIANTE	> 25	TÉCNICO	SI
TRABAJADOR	> 25	EMPRESARIAL	NO

Modelo (Árbol de Decisión)

```
Edad = < 21
|   Perfil = TÉCNICO: SI
|   Perfil = EMPRESARIAL
|   |   Tipo = ESTUDIANTE: NO
|   |   Tipo = TRABAJADOR: SI
|   Perfil = ACADÉMICO: SI
Edad = > 21 < 25
|   Tipo = ESTUDIANTE
|   |   Perfil = EMPRESARIAL: SI
|   |   Perfil = ACADÉMICO: NO
|   Tipo = TRABAJADOR: NO
Edad = > 25
|   Tipo = ESTUDIANTE: SI
|   Tipo = TRABAJADOR: NO
Edad = > 25 : NO
```

Tipo	Edad	Perfil	Interesa_charla
ESTUDIANTE	< 21	TÉCNICO	SI
TRABAJADOR	< 21	EMPRESARIAL	SI
ESTUDIANTE	< 21	EMPRESARIAL	NO
ESTUDIANTE	< 21	ACADÉMICO	SI
ESTUDIANTE	> 21 < 25	EMPRESARIAL	SI
TRABAJADOR	> 21 < 25	ACADÉMICO	NO
ESTUDIANTE	> 21 < 25	EMPRESARIAL	NO
TRABAJADOR	> 21 < 25	EMPRESARIAL	NO
ESTUDIANTE	> 21 < 25	ACADÉMICO	NO
ESTUDIANTE	> 25	TÉCNICO	SI
TRABAJADOR	> 25	TÉCNICO	NO
ESTUDIANTE	> 25	TÉCNICO	SI
ESTUDIANTE	> 25	TÉCNICO	SI
TRABAJADOR	> 25	EMPRESARIAL	NO

Tipo	Edad	Perfil	Interesa_charla
ESTUDIANTE	< 21	TÉCNICO	SI
ESTUDIANTE	< 21	EMPRESARIAL	NO
ESTUDIANTE	< 21	ACADÉMICO	SI
ESTUDIANTE	> 21 < 25	EMPRESARIAL	SI
ESTUDIANTE	> 21 < 25	EMPRESARIAL	NO
ESTUDIANTE	> 21 < 25	ACADÉMICO	NO
ESTUDIANTE	> 25	TÉCNICO	SI
ESTUDIANTE	> 25	TÉCNICO	SI
ESTUDIANTE	> 25	TÉCNICO	SI
TRABAJADOR	< 21	EMPRESARIAL	SI
TRABAJADOR	> 21 < 25	ACADÉMICO	NO
TRABAJADOR	> 21 < 25	EMPRESARIAL	NO
TRABAJADOR	> 25	TÉCNICO	NO
TRABAJADOR	> 25	EMPRESARIAL	NO

Tipo	Edad	Perfil	Interesa_charla
ESTUDIANTE	< 21	ACADÉMICO	SI
ESTUDIANTE	> 21 < 25	ACADÉMICO	NO
TRABAJADOR	> 21 < 25	ACADÉMICO	NO
ESTUDIANTE	< 21	EMPRESARIAL	NO
ESTUDIANTE	> 21 < 25	EMPRESARIAL	SI
ESTUDIANTE	> 21 < 25	EMPRESARIAL	NO
TRABAJADOR	< 21	EMPRESARIAL	SI
TRABAJADOR	> 21 < 25	EMPRESARIAL	NO
TRABAJADOR	> 25	EMPRESARIAL	NO
ESTUDIANTE	< 21	TÉCNICO	SI
ESTUDIANTE	> 25	TÉCNICO	SI
ESTUDIANTE	> 25	TÉCNICO	SI
ESTUDIANTE	> 25	TÉCNICO	SI
TRABAJADOR	> 25	TÉCNICO	NO

Edad	Perfil	Tipo	Interesa_charla
< 21	ACADÉMICO	ESTUDIANTE	SI
< 21	EMPRESARIAL	ESTUDIANTE	NO
< 21	EMPRESARIAL	TRABAJADOR	SI
< 21	TÉCNICO	ESTUDIANTE	SI
> 21 < 25	ACADÉMICO	ESTUDIANTE	NO
> 21 < 25	ACADÉMICO	TRABAJADOR	NO
> 21 < 25	EMPRESARIAL	ESTUDIANTE	SI
> 21 < 25	EMPRESARIAL	ESTUDIANTE	NO
> 21 < 25	EMPRESARIAL	TRABAJADOR	NO
> 25	TÉCNICO	ESTUDIANTE	SI
> 25	TÉCNICO	ESTUDIANTE	SI
> 25	TÉCNICO	ESTUDIANTE	SI
> 25	TÉCNICO	TRABAJADOR	NO
> 25	EMPRESARIAL	TRABAJADOR	NO

Algoritmo (ID3)

```
Id3(Ejemplos, Atributo-objetivo, Atributos )
   Si todos los ejemplos son positivos devolver un nodo positivo
   Si todos los ejemplos son negativos devolver un nodo negativo
   Si Atributos está vacío devolver el voto mayoritario del valor del atributo objetivo en
                                                                                  Ejemplos
   En otro caso
        Sea A Atributo el MEJOR de atributos
        Para cada v valor del atributo hacer
              Sea Ejemplos(v) el subconjunto de ejemplos cuyo valor de atributo A es v
              Si Ejemplos(v) esta vacío devolver un nodo con el voto mayoritario del
                                                          Atributo objetivo de Ejemplos
              Sino Devolver Id3(Ejemplos(v), Atributo-objetivo, Atributos/{A})
```

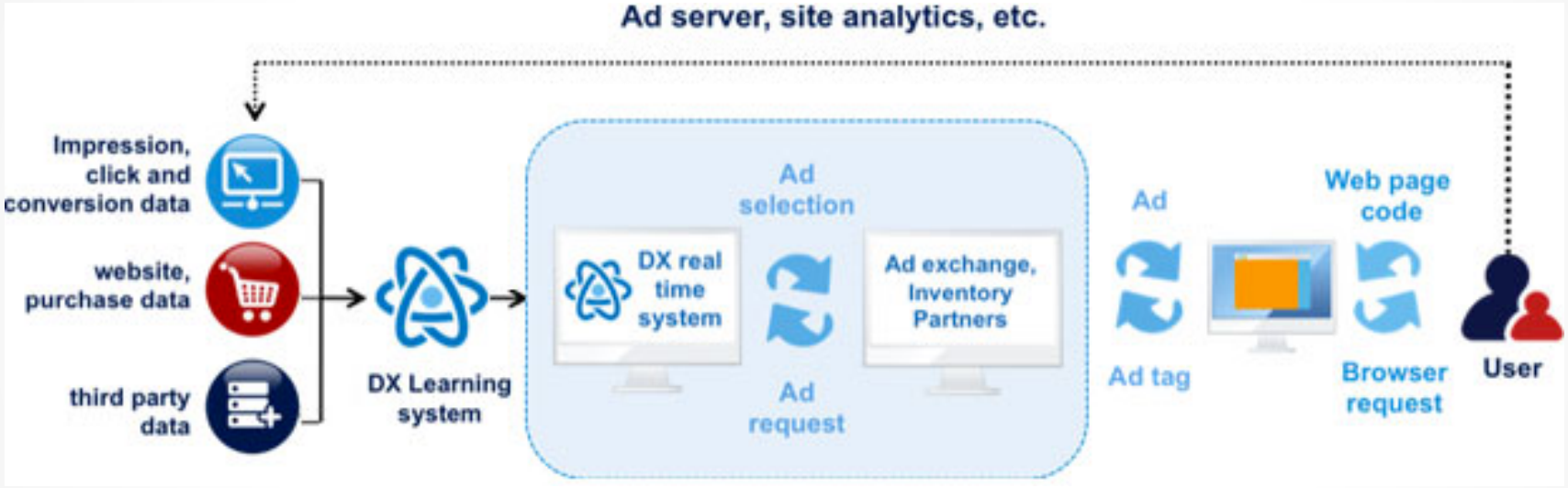
Machine Learning

- Aprendizaje Supervisado
- Aprendizaje No Supervisado
- Aprendizaje por Refuerzo

Big Data

- 4 billones de páginas web indexadas
- 100 horas de video subidas a YouTube por minuto
- Walmart maneja 1 millón de transacciones por hora
- 90% de los datos del mundo fueron creados en los últimos 2 años
- Facebook: 30 petabytes de información almacenada, analizada y accedida.
- Twitter: 230 millones de tweets por día
- 300 billones de emails enviados todos los días
- 1 billón de búsquedas por día en Google

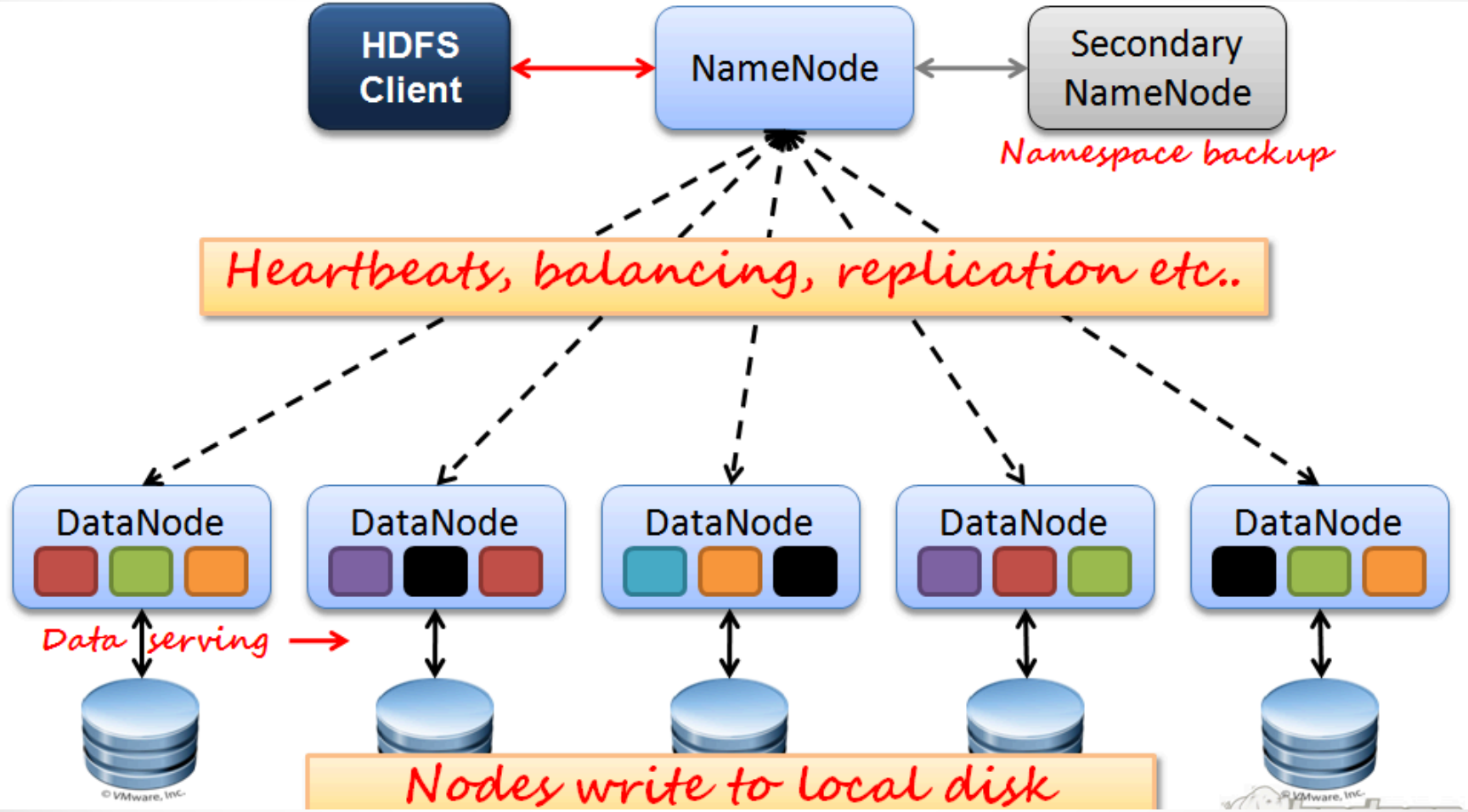
Caso: DataXu

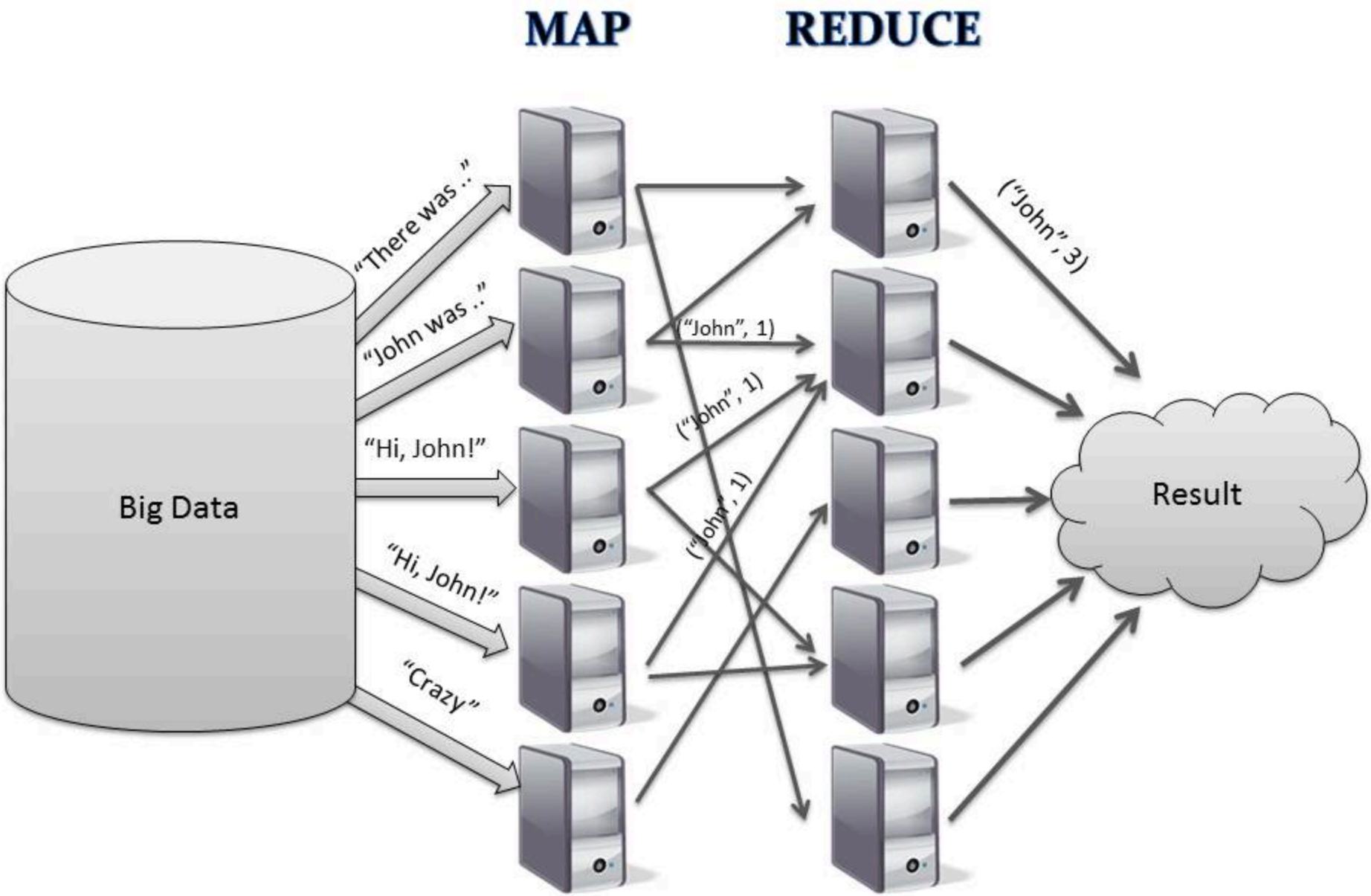

Ad server, site analytics, etc.
Impression, click and conversion data
website, purchase data
third party data
DX Learning system
Ad selection
DX real time system
Ad exchange, Inventory Partners
Ad request
Ad
Ad tag
Web page code
Browser request
User

Big Data

- Google
  - File System (GFS)
  - Map Reduce
- Hadoop
  - FileSystem (HDFS)
  - Map Reduce, YARN
- Amazon
  - S3
  - Elastic Map Reduce

HDFS

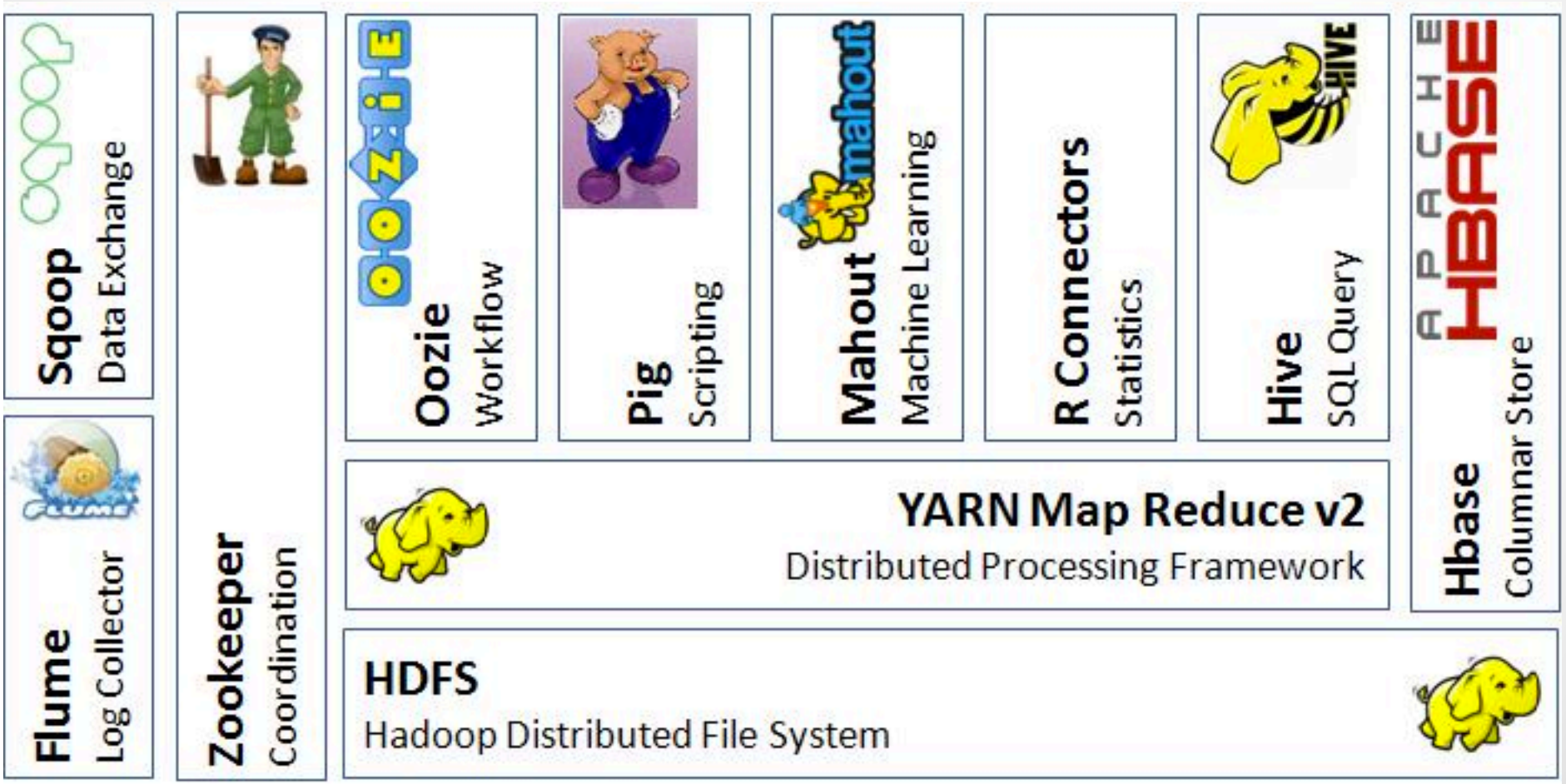

HDFS Client
NameNode
Secondary NameNode
Namespace backup
Heartbeats, balancing, replication etc..
DataNode
DataNode
DataNode
DataNode
DataNode
Data serving
Nodes write to local disk


MAP
REDUCE
Big Data
"There was .."
"John was .."
"Hi, John!"
"Hi, John!"
"Crazy"
("John", 1)
("John", 1)
("John", 1)
("John", 3)
Result

Hadoop Resuelve

- Almacenamiento Distribuido
- Particionar la Información
- Procesamiento distribuido
- Agrupación de datos
- Tolerante a fallas

Hadoop Eco System



Apache Mahout

- Collaborative Filtering
  - Item-Based Collaborative Filtering
  - Matrix Factorization with Alternating Least Squares
  - Matrix Factorization with Alternating Least Squares on Implicit Feedback
- Classification
  - Naive Bayes / Complementary Naive Bayes
  - Random Forest
  - Hidden Markov Models
  - Multilayer Perceptron
- Clustering
  - k-Means Clustering
  - Fuzzy k-Means
  - Streaming k-Means
  - Spectral Clustering
- Dimensionality Reduction
  - Lanczos Algorithm
  - Principal Component Analysis
- Miscellaneous
  - Frequent Pattern Mining - MapReduce
  - RowSimilarityJob
  - ConcatMatrices
  - Collocations

Deseo de los científicos e ingenieros

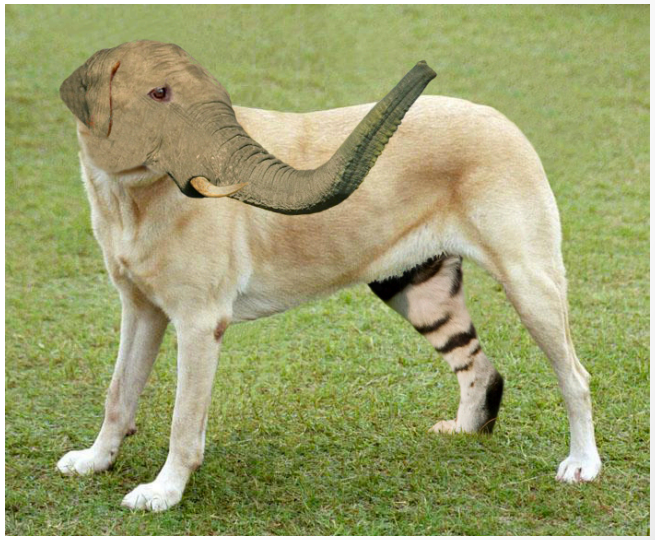
Escribir un algoritmo en un lenguaje de alto nivel y que la tecnología se encargue de la paralelización y optimización

Deseo de los ingenieros ML

- Escribir un algoritmo en un lenguaje de alto nivel y que la tecnología se encargue de la paralelización y optimización
- No existe todavía

Nuestro Algoritmo

Nuestro Software

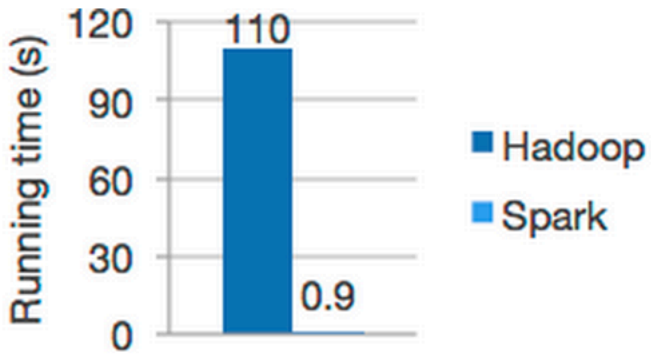


- Ley del instrumento: si lo único que tienes es un martillo, todo se verá como un clavo

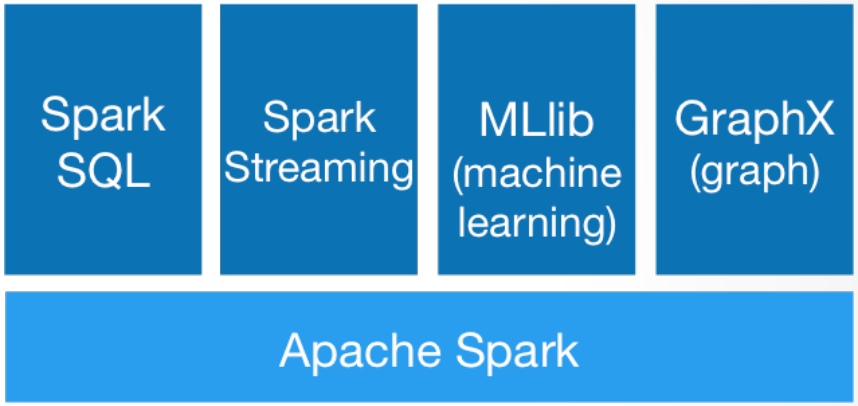
Apache Spark

- Generalización de MapReduce
- Optimizado para guardar los datos en memoria
- Compatible con Hadoop / YARN
- Algoritmos se describen como transformaciones funcionales
- Optimización transparente
- Transformaciones son declarativas
- Compatible con Python, Java y Scala.
- Mahout migrando a Spark

Apache Spark



Logistic regression in Hadoop and Spark




```
file = spark.textFile("hdfs://...")

file.flatMap(lambda line: line.split())
    .map(lambda word: (word, 1))
    .reduceByKey(lambda a, b: a+b)
```

Word count in Spark's Python API

Impacto Big Data & Machine Learning

- Análisis del comportamiento de los clientes (y acción!)
- Optimización de procesos de negocio
- Salud
- Deportes
- Ciencia e Investigación
- Optimización de dispositivos y máquinas
- Seguridad Cuidadana
- Bolsa

Muchas Gracias

¿Preguntas?

Links

Links

- http://mobile.blogs.wsj.com/cio/2014/07/10/germanys-12th-man-at-the-world-cup-big-data/?mg=blogs-wsj&utm_content=buffer378d4&utm_medium=social&utm_source=linkedin.com&utm_campaign=buffer
- http://en.wikipedia.org/wiki/Big_data
- https://www.linkedin.com/pulse/article/20131113065157-64875646-the-awesome-ways-big-data-is-used-today-to-change-our-world
- https://spark.apache.org/
- www.dataxu.com
- https://mahout.apache.org/users/basics/algorithms.html
- http://hortonworks.com/hadoop/yarn/
- http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html
- http://www.ibmbigdatahub.com/gallery/quick-facts-and-stats-big-data
- https://yoyoclouds.wordpress.com/tag/hdfs/
- http://www.baselinemag.com/analytics-big-data/slideshows/surprising-statistics-about-big-data.html
- http://www.forbes.com/sites/tedgreenwald/2012/03/29/big-data-small-decisions-rendered-very-quickly-under-the-hood-at-dataxu/
- http://www.worldwidewebsize.com/
- http://papers.nips.cc/paper/3150-map-reduce-for-machine-learning-on-multicore.pdf
- http://www.cs.ubc.ca/~murphyk/MLbook/
- http://en.wikipedia.org/wiki/Data_mining#Process

Imágenes

- http://www.pakwheels.com/forums/attachments/guess-humor-hobbies/545611d1210727280-how-many-people-can-u-fit-ur-car-volkswagen-beetle_1938_800x600_wallpaper_1f_bob_pakwheels-com-.jpg
- http://www.enterrasolutions.com/media/images/2013/08/6a00d8341c4ebd53ef0191047c2f5c970c-pi.png
- http://www.baselinemag.com/analytics-big-data/slideshows/surprising-statistics-about-big-data.html
- http://www.ibmbigdatahub.com/gallery/quick-facts-and-stats-big-data
- https://yoyoclouds.wordpress.com/tag/hdfs/
- http://www.hrc.es/bioest/roc_21.gif
- http://techblog.baghel.com/index.php?itemid=132