Procesamiento de Datos Masivos
03MBID
Tema 2: Big Data – MapReduce

Yudith Cardinale, PhD

Profesor, Universidad Internacional de Valencia
Profesor Titular, Universidad Simón Bolívar, Departamento de Computación, Venezuela
Investigador Asociado, Universidad Católica San Pablo, Perú

Noviembre 2022

Outline

Agenda

The Big Data Boom

Motivación

- La producción de datos por parte de usuarios en la Web (blogs, redes sociales, etc.) y el compartimiento de información ubicua (sensores y dispositivos móviles, cámaras, micrófonos, fotografías, etc.), aumenta drásticamente la cantidad de datos que pueden ser procesados y las perspectivas de interpretación.
- El crecimiento de la cantidad de datos disponibles se dispara de una manera sin precedentes: En 2017 la IDC (International Data Corporation) predijo que para el 2025 se alcanzarían los 163 ZB (trillones de GB) de datos.
- Según IDC, sólo en el 2020 se creó 64.2 ZB de datos (producto de la situación COVID-19).
- "The amount of digital data created over the next five years will be greater than twice the amount of data created since the advent of digital storage. The question is: How much of it should be stored?"

The Big Data Boom

Motivación (cont.)

- Sólo el 2% de los datos creados en 2020, se mantuvo (se almacenó).
- En 2020, IDC anunció:
  - Un crecimiento de 23% anual para el período 2020-2025.
  - Que los datos de IoT (sin incluir cámaras de vídeo-vigilancia) son el segmento de datos que crece más rápido, seguido de las redes sociales.
  - Los datos creados en el *cloud* no crecen tan rápido como los datos almacenados en el *cloud*, pero sigue siendo uno de los segmentos que crece rápido.
  - La creación de datos en *the edge* crece casi tan rápido como en el *cloud*.
- Predicciones de IDC 2021: https://www.idc.com/research/viewtoc.jsp?containerId=US46920420

The Big Data Boom

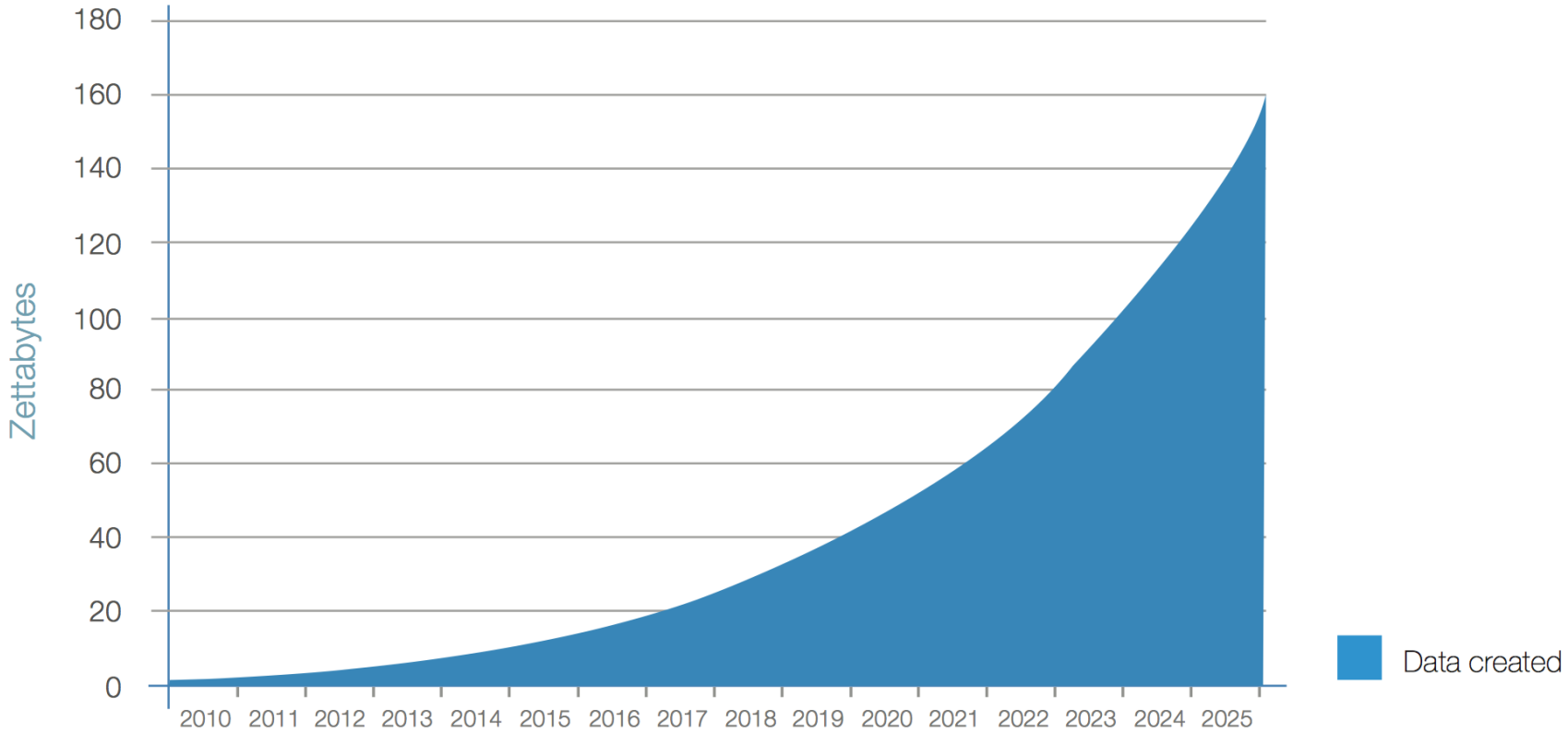
Motivación (cont.)

- Facebook: 31,25 millones de mensajes enviados y 2,77 millones de vídeos reproducidos cada segundo.
- Youtube: 300 horas de vídeo se cargan cada minuto.
- En 5 años habrá más de 50 billones de dispositivos inteligentes conectados.

The Big Data Boom

Predicción en 2017:

Figure 2. Annual Size of the Global Datasphere

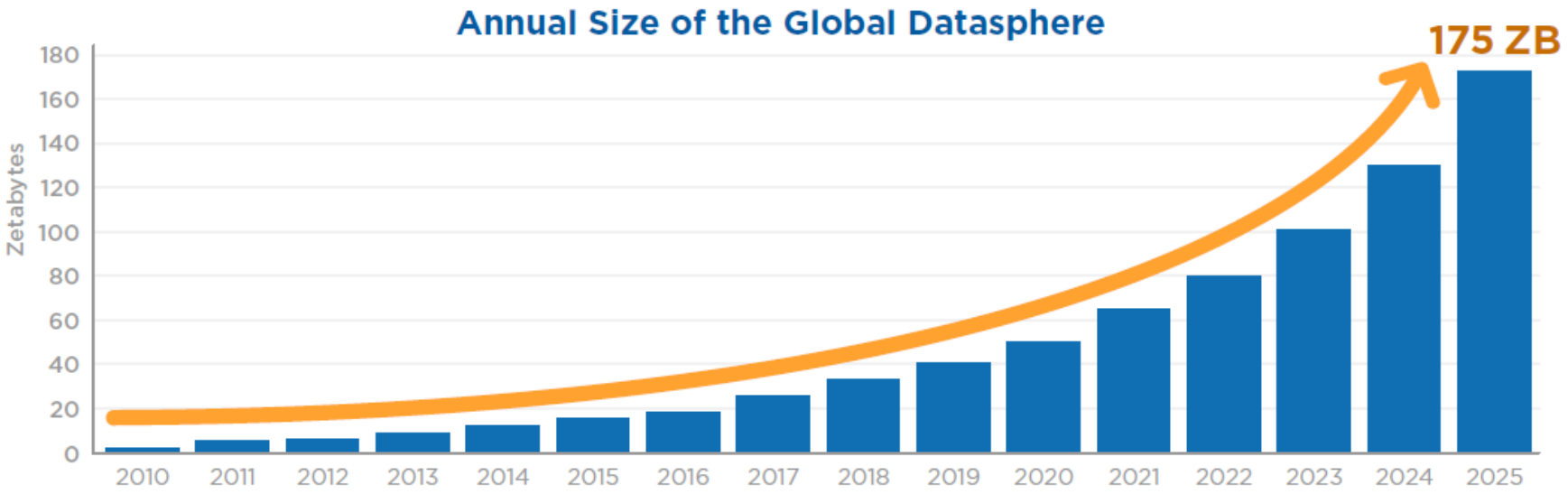


Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

The Big Data Boom

Predicción en 2018:

Figure 1 – Annual Size of the Global Datasphere



The Big Data Boom: Data Explosion

Data Explosion

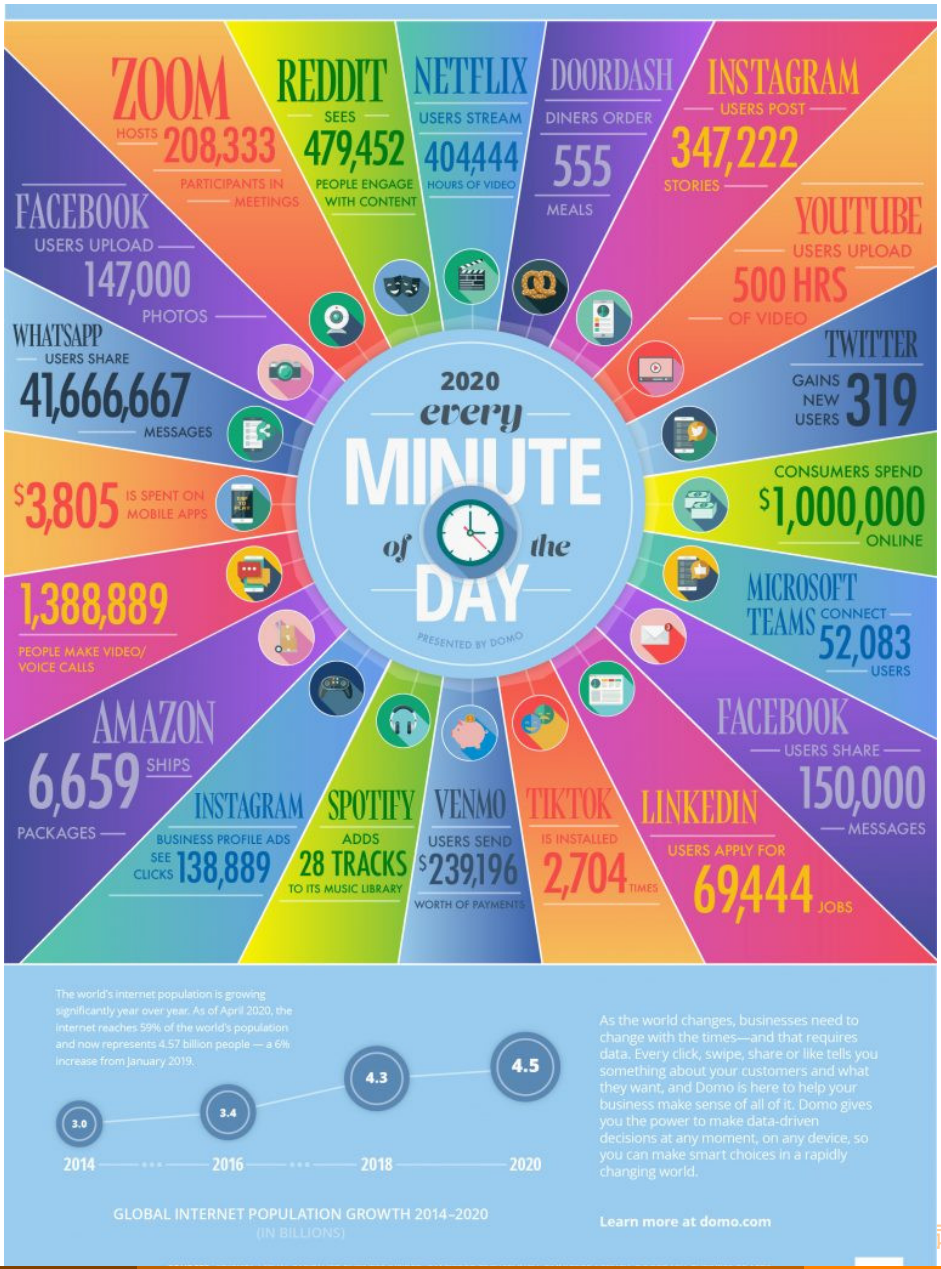
- ¿ Qué genera este crecimiento explosivo de datos? $\Rightarrow$ Innovación
  - Transformación del modelo de negocio
    - Globalización y conectividad
    - De *product-oriented* a *service-oriented*, personalización de servicios
    - Modelos B2B, B2B2C y B2C2C.
  - Nuevas fuentes de datos (social media, dispositivos móviles, redes de sensores, ...) $\Rightarrow$ Evolución social y cultural
    - Cada día creamos 2.5 quintillones de bytes de datos; más de 90% de los datos existentes hoy en el mundo, fueron creados solamente en los dos últimos años!
  - Tecnología avanzada en
    - Dispositivos móviles
    - Redes de procesamiento de datos a gran escala
    - La "commoditization" del hardware
    - Cloud Computing, IoT, ciencia de los datos
    - Seguridad, virtualización, *open-source software*, ...

The Big Data Boom: Data Explosion

https://www.domo.com/learn/data-never-sleeps-3-0

http://www.adweek.com/prnewser/how-many-times-do-the-worlds-social-media-users-click-every-minute/117427

The Big Data Boom: Data Explosion



The Big Data Boom: Definiciones

Definiciones

- "Big Data can be defined as volumes of data available in varying degrees of complexity, generated at different velocities and varying degrees of ambiguity, that cannot be processed using traditional technologies, processing methods, algorithms, or any commercial off-the-shelf solutions. Data defined as Big Data includes machine-generated data from sensor networks, nuclear plants, X-ray and scanning devices, and airplane engines, and consumer-driven data from social media. Big Data producers that exist within organizations include legal, sales, marketing, procurement, finance, and human resources departments"[1].
- "Big Data refers to datasets and flows large enough that has outpaced our capability to store, process, analyze, and understand"[2].

The Big Data Boom: Definiciones

Definiciones

- "Big Data are high–volume, high–velocity, and high–variety information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization" (Gartner 2012).
- "Building new analytic applications based on new types of data, in order to better serve your customers and drive a better competitive advantage" (David McJannet, Hortonworks).
- "Big Data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it..."

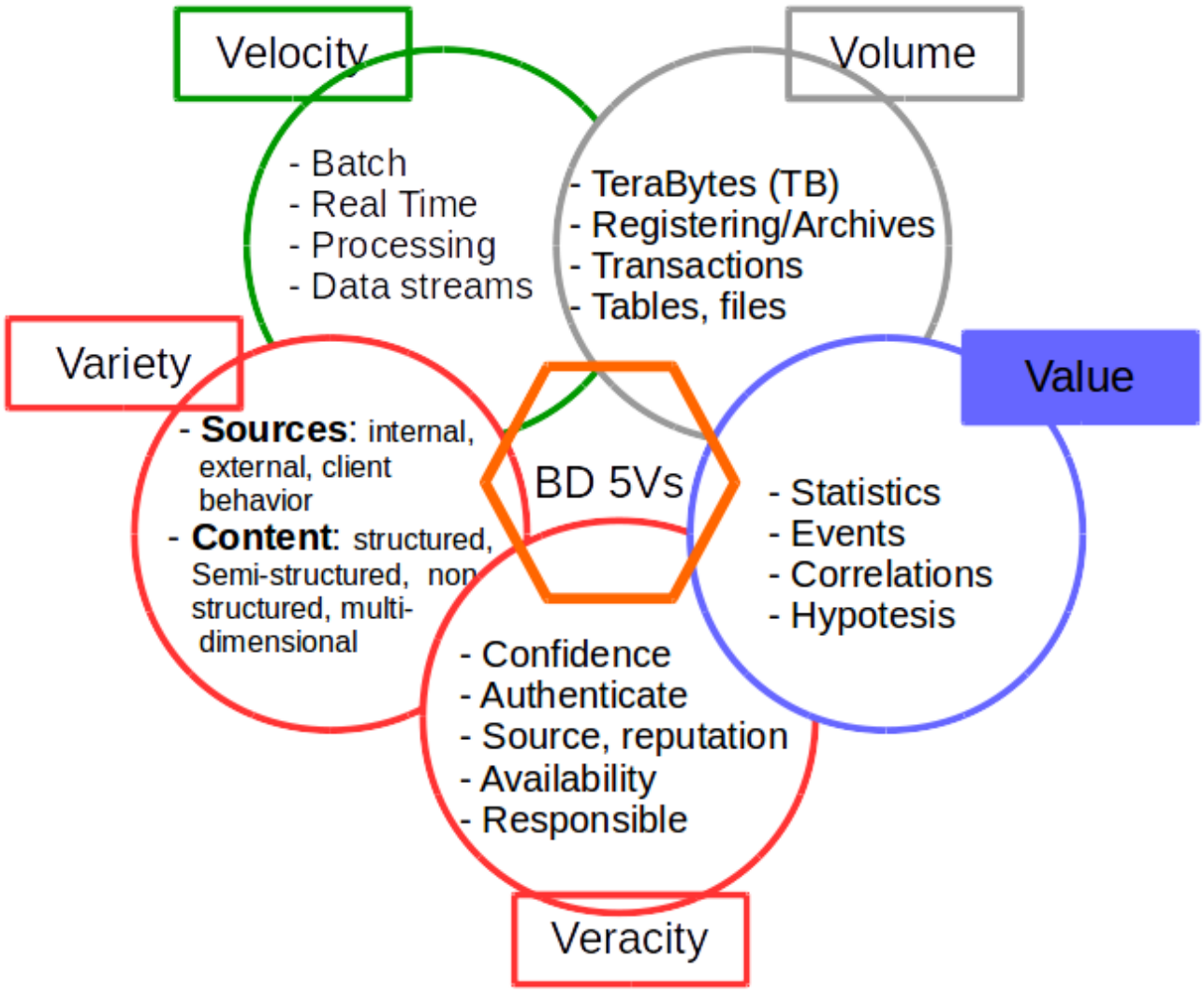
Dan Ariely
Professor of Psychologie
Duke University – NC - USA

The Big Data Boom: Las Vs

Big Data Dimensions

- En 2001, un analista de META Group (ahora Gartner) definió, en un reporte de investigación, los retos y oportunidades del crecimiento de los datos desde una perspectiva de tres dimensiones: Las "3Vs" (Volume, Velocity, and Variety);
- En 2012, IBM agregó una cuarta dimensión: Veracity;
- La dimensión más importante (para el negocio), fue identificada más recientemente: Value $\Rightarrow$ **Ciencia de los datos con "5Vs"**;
- Versión de "7Vs": "5Vs" + Variability, Visualization
- Versión de "10Vs": "5Vs" + Variability, Validity, Venue, Vocabulary, Vagueness
- ¿Se te ocurre otra V? Agrégala aquí .....

The Big Data Boom: 5 Vs



The Big Data Boom: 7 Vs

- Volume: enorme cantidad de datos generados
- Velocity: rapidez con la que se generan y mueven esos datos
- Variety: datos de múltiples tipos, estructurados y no-estructurados; texto, datos de sensores, audio, vídeo, click streams, ficheros de log, etc.
- Veracity, datos correctos o incorrectos.
- Value: capacidad de extraer valor de los datos.
- Variability: el significado de los datos puede variar con el tiempo.
- Visualization, los datos deben poder ser comprendidos.

The Big Data Boom: Más Vs

Figure: https://www.learnbigdatatools.com/learn-big- data-databases/

The Big Data Boom: Data Explosion

Retos

- Estos datos se presentan en formatos que difícilmente pueden ser tratados por DBSM tradicionales
  - no están organizados en formatos de tablas y la estructura puede variar (no-estructurados);
  - se generan en tiempo real en flujos contínuos;
  - provienen de distintas fuentes (dispositivos móviles, sensores, PCs, Laptops, objetos, ...) de forma desordenada y no predecible;
- La captura, almacenamiento, búsqueda, compartimiento, análisis y visualización de datos se debe redefinir.
  - Coleccionar grandes volúmenes de datos, variados, para encontrar nuevas ideas;
  - Capturar rápidamene los datos creados;
  - Almacenar todos esos datos;
  - Tratar, analizar y usar esos datos.

The Big Data Boom: Data Explosion

Structured data	Semi-structured data	Unstructured data
Databases	XML / JSON data Email Web pages	Audio Video Image data Natural language Documents

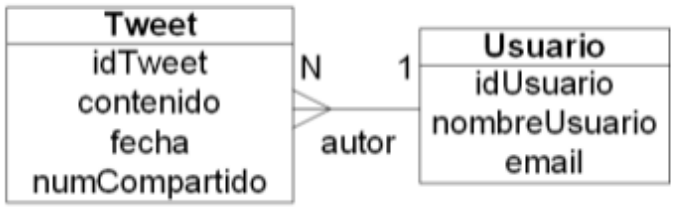
The Big Data Boom: Data Explosion

Retos

- Estos datos se presentan en formatos que difícilmente pueden ser tratados por DBSM tradicionales
  - no están organizados en formatos de tablas y la estructura puede variar (no-estructurados);
  - se generan en tiempo real en flujos contínuos;
  - provienen de distintas fuentes (dispositivos móviles, sensores, PCs, Laptops, objetos, ...) de forma desordenada y no predecible;
- La captura, almacenamiento, búsqueda, compartimiento, análisis y visualización de datos se debe redefinir:
  - Coleccionar grandes volúmenes de datos, variados, para encontrar nuevas ideas;
  - Capturar rápidamene los datos creados;
  - Almacenar todos esos datos;
  - Tratar, analizar y usar esos datos.

Modelos Big Data

- Modelos de datos
  - Relacional



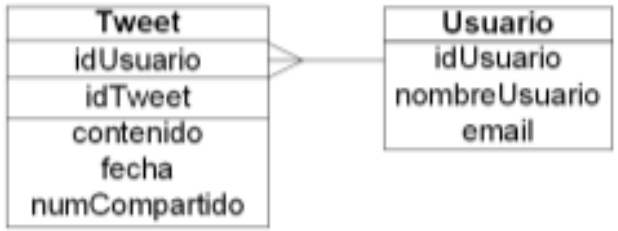
idTweet	contenido	fecha	numCompartidos	idAutor
1	et temp...	07/08/2018	20	1
2	Semp...	07/08/2018	12	2
3	est qua...	08/08/2018	178	1
4	Phare...	09/08/2018	6	1
5	magna...	10/08/2018	2	3

idUsuario	nombreUsuario	email
1	flarcher0yo	flarcher0@nytimes.com
2	cbatchel1972	cbatchelour1@exblog.jp
3	serPailTwo	spail2@csmonitor.com

Modelos de datos

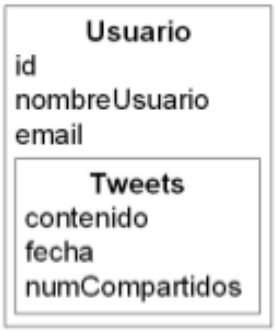
NoSQL: orientado a pares <clave, valor>

Clave de partición	Clave de ordenación	Atributos
idUsuario 1	idTweet 1	contenido et temp... fecha 07/08/2018 numCompartidos 20
idUsuario 2	idTweet 2	contenido Semp... fecha 07/08/2018 numCompartidos 12
idUsuario 1	idTweet 3	contenido est qua... fecha 08/08/2018 numCompartidos 178
idUsuario 1	idTweet 4	contenido Phare... fecha 09/08/2018 numCompartidos 6
idUsuario 3	idTweet 5	contenido magna... fecha 10/08/2018 numCompartidos 2



Clave de partición	Atributos
idUsuario 1	nombreUsuario flarcher0yo email flarcher0@nytimes.com
idUsuario 2	nombreUsuario cbatchel1972 email cbatchelour1@exblog.jp
idUsuario 3	nombreUsuario serPailTwo email spail2@csmonitor.com

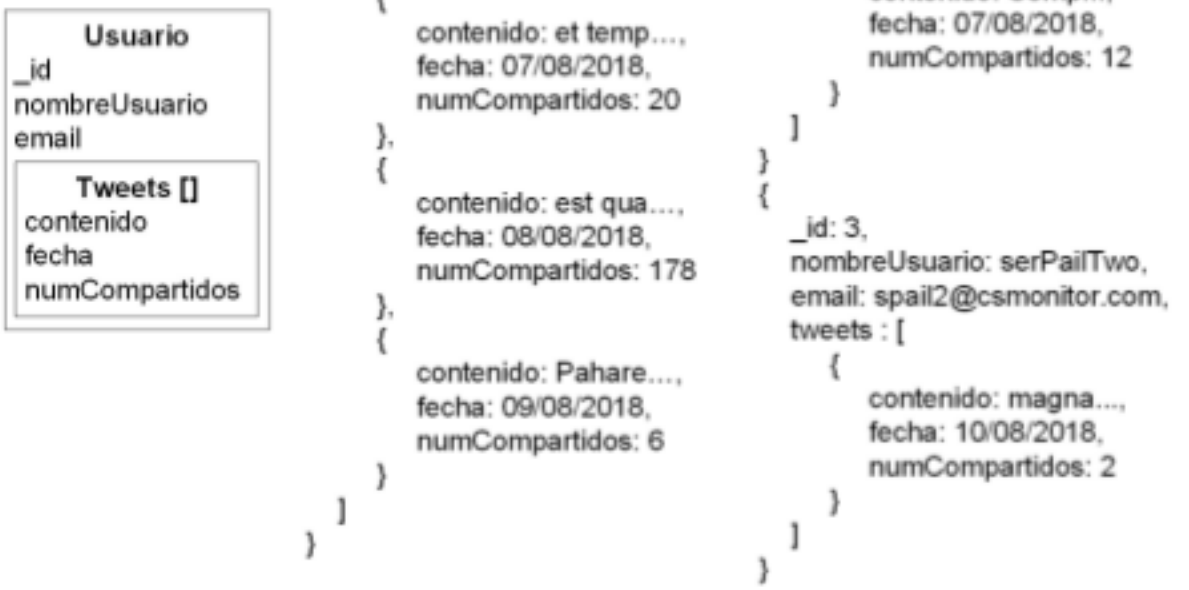
- Modelos de datos
  - NoSQL: orientado a columnas



Clave	Datos usuario		Datos Tweet					
idUsuario	nombreUsuario	email	contenido		fecha		numCompartidos	
1	flarcher0yo	flarcher0@nytimes.com	contenido1	et temp...	fecha1	07/08/2018	numCompartidos1	20
			contenido3	est qua...	fecha3	08/08/2018	numCompartidos3	178
			contenido4	Phare...	fecha4	09/08/2018	numCompartidos4	6
2	cbatchel1972	cbatchelour1@exblog.jp	contenido2	Semp...	fecha2	07/08/2018	numCompartidos2	12
3	serPailTwo	spail2@csmonitor.com	contenido5	magna...	fecha2	10/08/2018	numCompartidos2	2

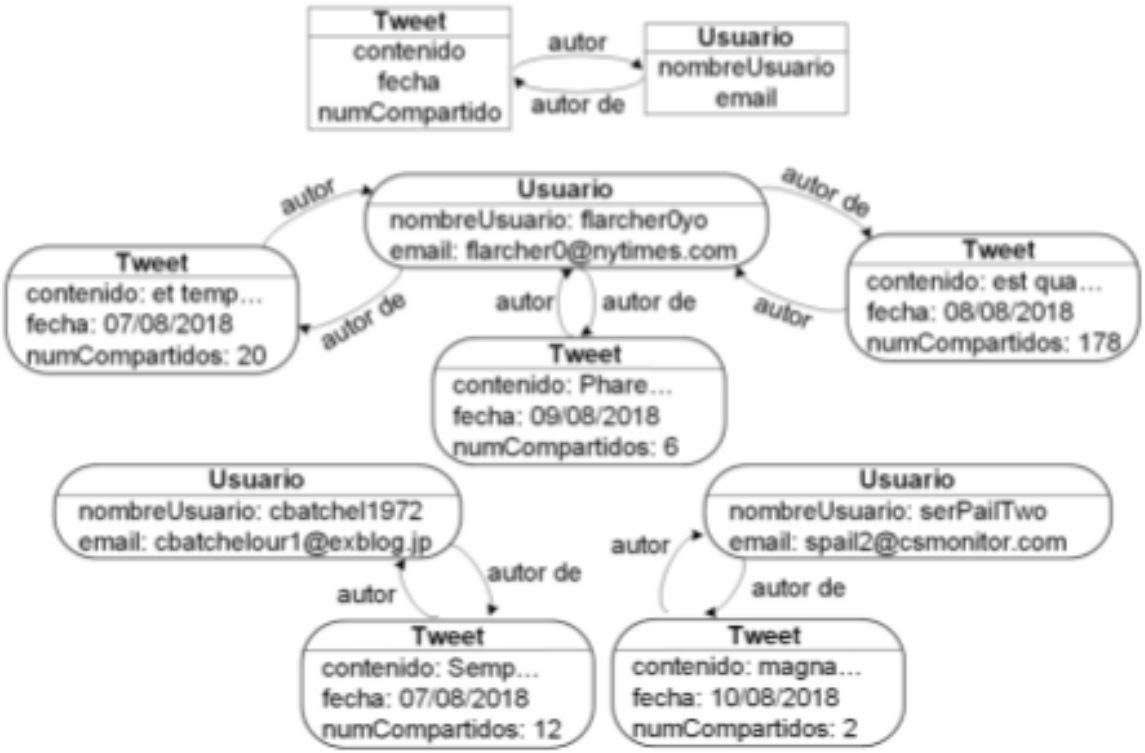
Modelos de datos

NoSQL: orientado a documentos



Modelos de datos

NoSQL: orientado a grafos



The Big Data Boom: Data Explosion

Modelos de Procesamiento

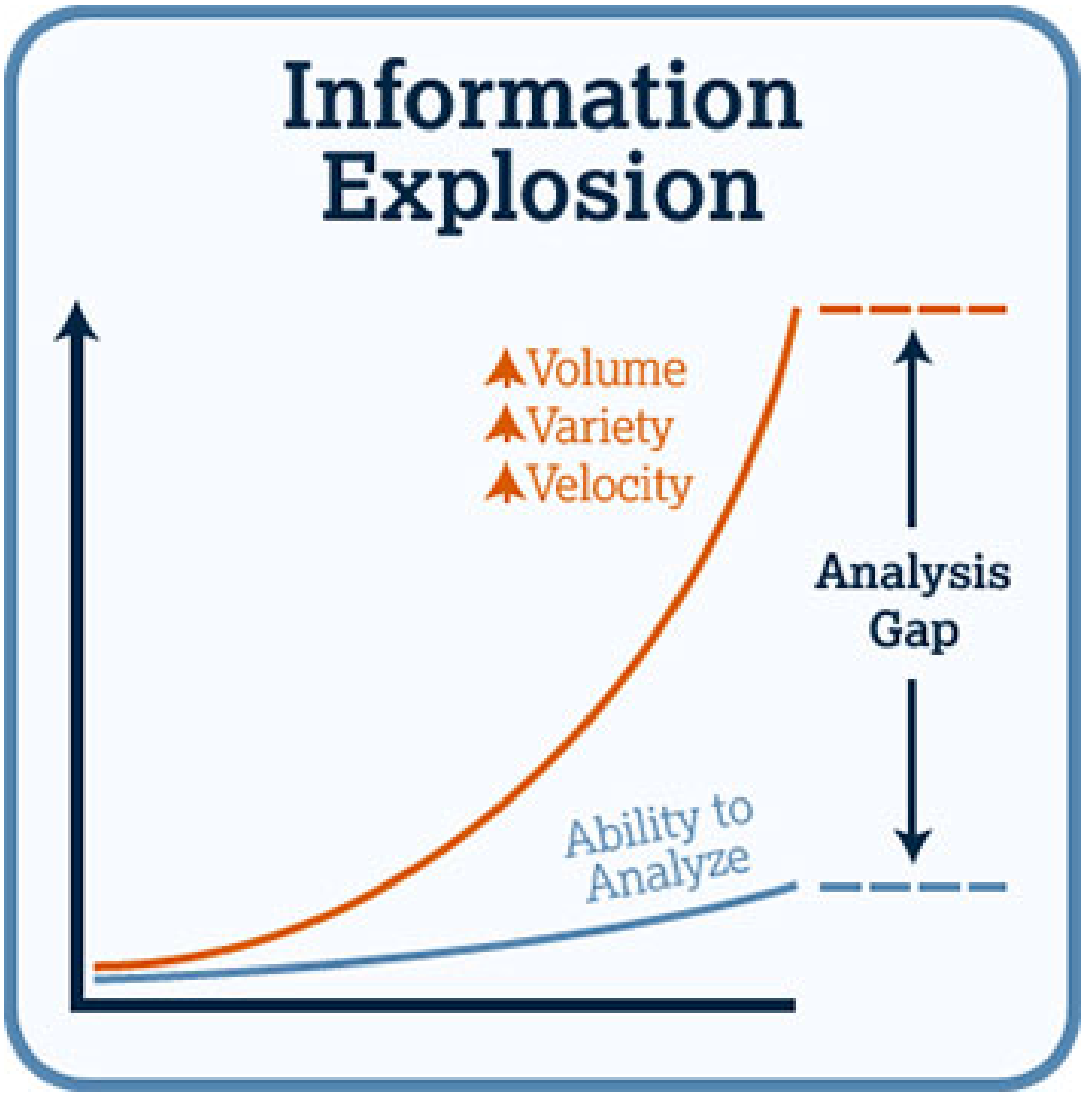
- Batch
- Streaming
- Transaccional
- Arquitectura Lambda: Streaming + Batch
  - Speed Layer: procesamiento en tiempo real
  - Batch Layer: procesamiento batch
  - Usuario: consulta las vistas de speed + batch
  - Desventaja: mantener dos "sistemas"
- Arquitectura Kappa: Streaming + Batch
  - Datos en tiempo real: streaming
  - Datos históricos: streaming grande
  - Usuario consulta las vistas del streaming histórico + tiempo real
  - Desventaja: Si hay fallo, se tiene que re-ejecutar el histórico – A veces se puede evitar

Agenda

Ciencia de los datos

Ciencia de los datos = Big Data Analytics

Ciencia de los datos: Big Data Analytics

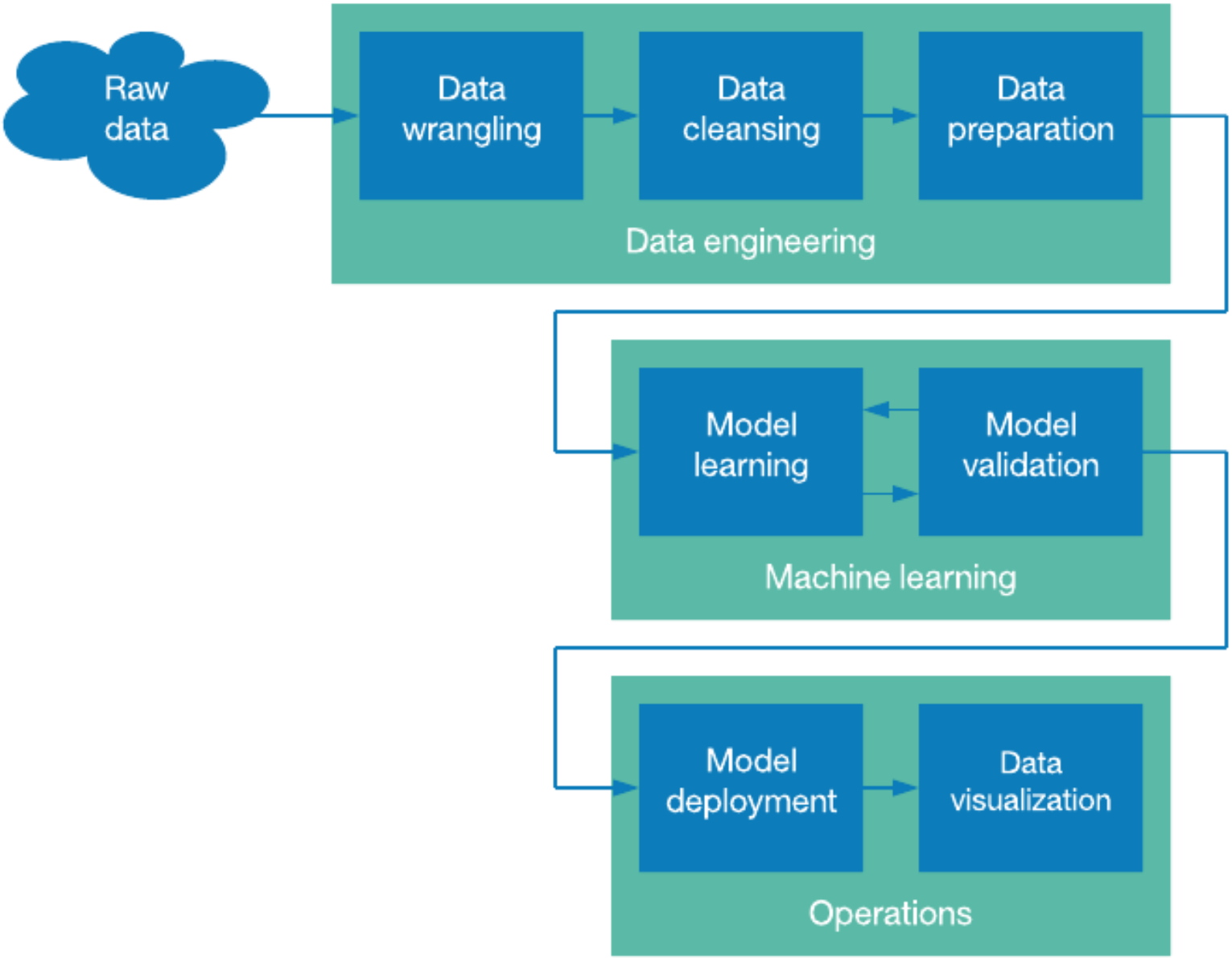


Ciencia de los datos: Big Data Analytics

Big Data Analytics

- Para obtener valor de los datos es necesario procesarlos, analizarlos;
- *Data Science* es un campo multidisciplinario cuyo objetivo es extraer valor de los datos;
- *Data Science* es un proceso que transforma datos crudos en significados, en información interpretada;
- El proceso es un *pipeline* que incluye ingeniería de los datos, *machine learning* y operaciones sobre los resultados.

Ciencia de los datos: Big Data Analytics



Ingeniería de Datos

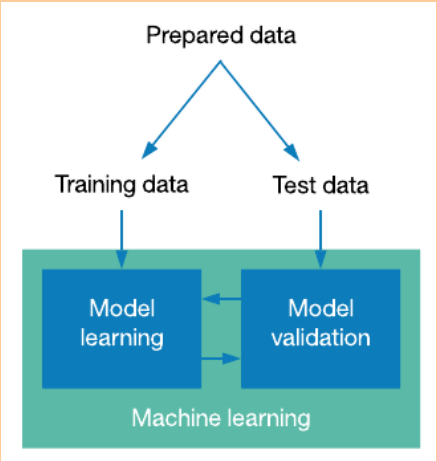
- Adquisición de datos:
  - Identificar y recolectar los datos crudos;
  - Integrar datos de diversas fuentes;
  - Representarlos en un formato común y consistente;
- Limpieza de los datos:
  - Identificar valores erróneos, inconsistencias o parámetros insuficientes;
  - La corrección de datos se puede hacer manual o automáticamene; si los datos no se pueden reparar, se eliminan;
  - Identificar valores atípcos (*outliers*) a través de análisis estadístico, por ejemplo.
- Pre-procesamiento o preparación de los datos:
  - Aún si los datos están "limpios", puede que requieran una peparación adicional antes de pasar a la fase de *machine learning;*
  - Normalización de los datos;
  - Convertir datos categóricos en valores numéricos.

Machine Learning

- Modelo de Aprendizaje:

Machine learning approaches		
Supervised learning	Unsupervised learning	Reinforcement learning
Backprop neural networks	K-means clustering	Q-learning
Decision tree learning	Principal component analysis	Temporal diff learning
Bayesian statistics	Generative adversarial network	SARSA
Support vector machines	Adaptive resonance theory	Monte Carlo methods
Random decision forests	Hierarchical clustering	Inverse reinforcement learning

- Validación del Modelo:

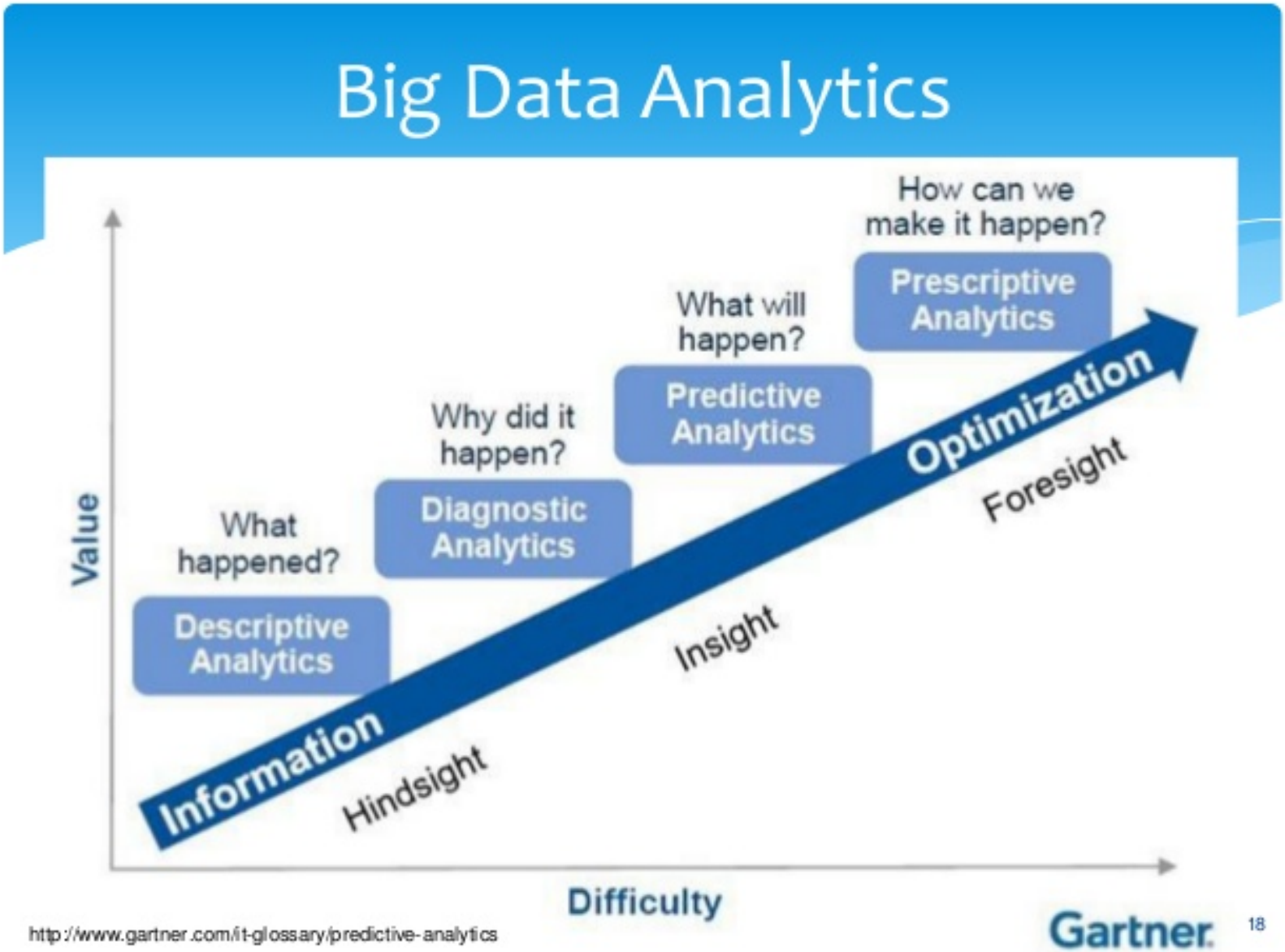


Ciencia de los datos: Big Data Analytics

Operaciones sobre los resultados

- Visualización:
  - Gráficos;
  - Reportes.
- Despliegue del modelo:
  - Predicciones;
  - Recomendaciones.

Ciencia de los datos: Big Data Analytics



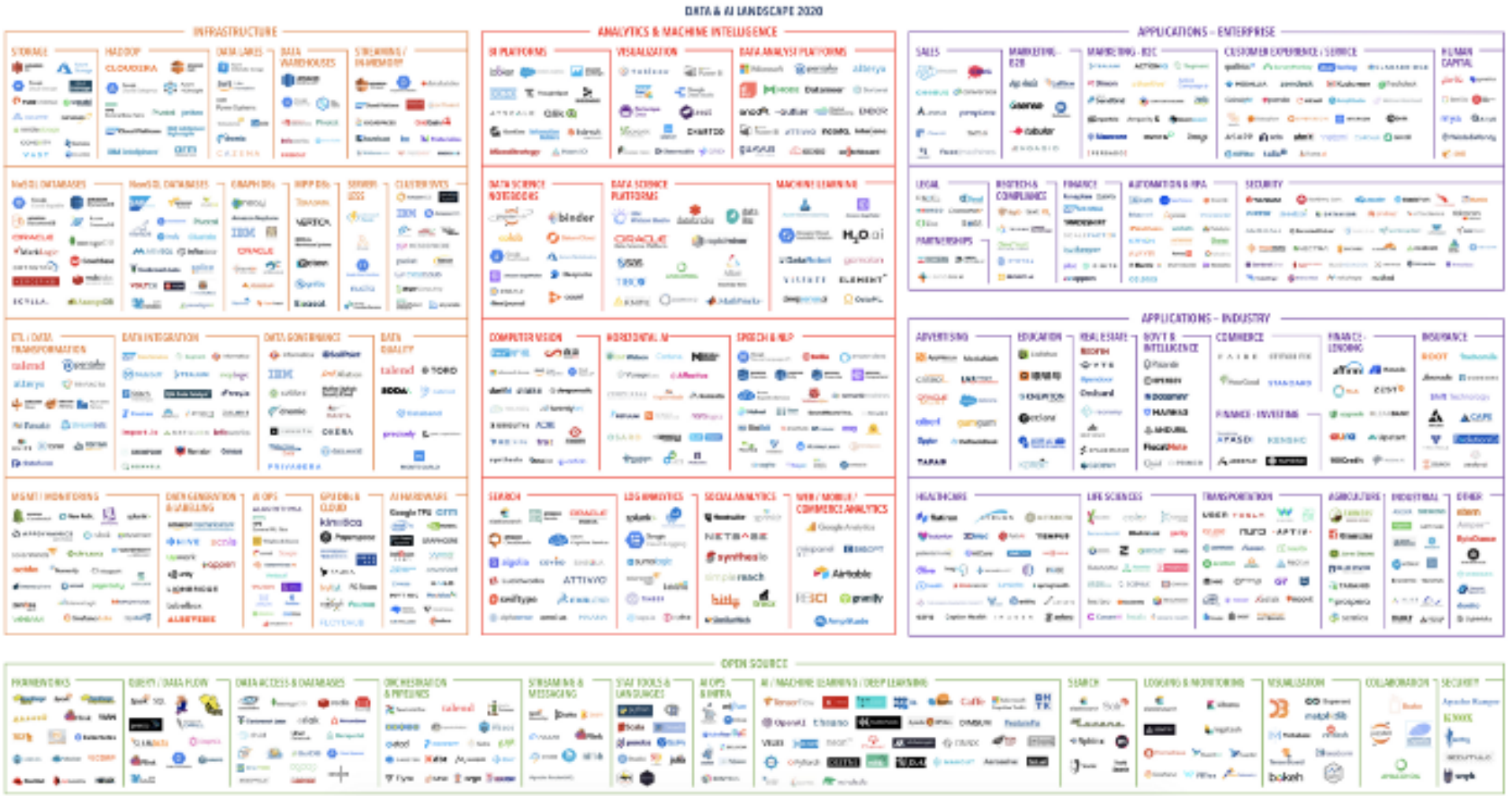
Retos Tecnológicos I

- Búsqueda y refinamiento (Lucene, Solr, Nutch, Elasticsearch, OpenRefine)
- Serialización (JSON, BSON, Apache Thrift, Apache Avro, Google Protocol Buffers)
- Sistemas de almacenamiento (HDFS, GFS, Lustre, Amazon S3)
- Servidores (Amazon EC2, Google Cloud Platform, Azure, OpenShift, Heroku, Tanzu, Ambari)
- Procesamiento (Hadoop, Hive, Pig, Spark, Flink, Beam, Apex, Disco, Dask, Tez, Cascading, Azkaban, Oozie, mrjob, Flume, Storm, Kinesis, Samza)

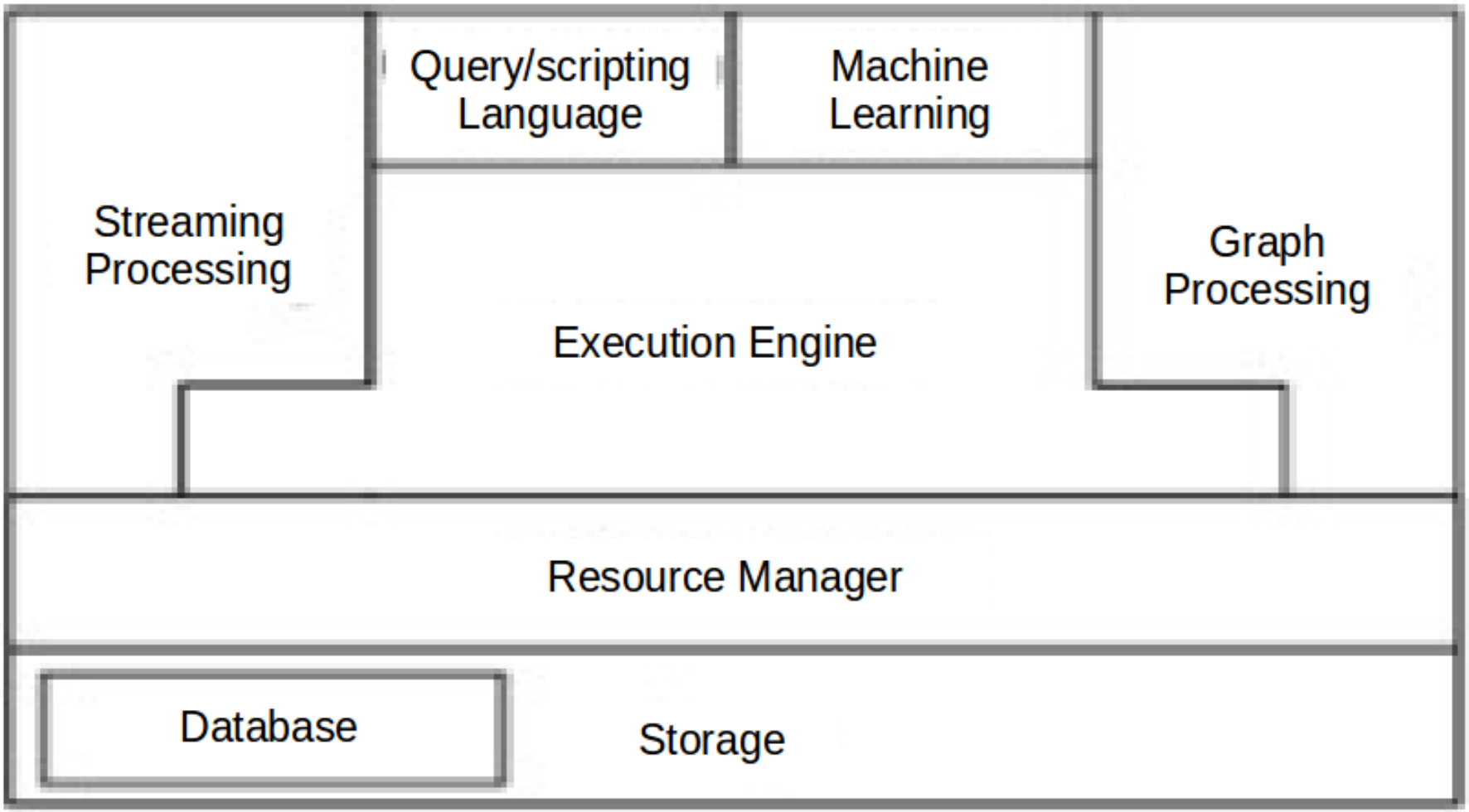
Retos Tecnológicos II

- Bases de datos (HBase, Cassandra, Accumulo, MongoDB, CouchDB, Riak, Amazon DynamoDB, Google BigTable, Drill, Kylin, Parquet, Sqoop)
- Análisis y BI (R, Greenplum, Splunk, Datameer, Kylin, Tableau, Jethro)
- Lenguaje natural (Natural Language Toolkit, Apache OpenNLP, Perldoop, Open Calais)
- Machine learning y deep learning (Weka, Apache Mahout, SciKit-learn, Pytorch, TensorFlow, caffe, Keras, Microsoft Cognitive Toolkit, BigDL, MXNet, Edward)
- Visualización (R, D3.js, Google Data Studio)

Ciencia de los datos: Big Data Analytics



Data Science: Big Data Analytics Stack



Data Science: Big Data Analytics Stack

Big Data Analytics – Storage (Filesystem)

- Los sistemas de ficheros tradicionales no están diseñados para funcionar con sistemas de procesamiento de datos de gran escala
- La eficiencia tiene mayor prioridad que otras características, e.g., servicio de directorios
- Cantidades masivas de datos normalmente se almacenan en múltiples máquinas de una manera distribuida
- HDFS, Amazon S3 (Amazon Simple Storage Service), Quantcast File Sytem, GFS, GlusterFS, ...

Data Science: Big Data Analytics Stack

Big Data Analytics – Databases

- Los sistemas de gestión de bases de datos relacionales (RDMS) no fueron diseãdos paras ser distribuidos.
- Las bases de datos NoSQL relajan una o más de las propiedades ACID, y proponen las propiedades BASE (Basically Available, Soft state, Eventual consistency)
- Se requieren diferentes modelos de datos: key/value, column-family, graph, document.
- Dynamo, Scalaris, BigTable, Hbase, Cassandra, MongoDB, Voldemort, Riak, Neo4J, ...

Big Data Analytics – Resource Management

- Distintos frameworks demandan diferentes recursos computacionales.
- Grandes organizaciones requieren la posibilidad de compartir datos y recursos entre múltiples frameworks.
- Los Resource Managements (gestores de recursos) permiten compartir los recursos de un cluster entre múltiples frameworks, asegurando isolation (aislamiento) de los recursos.
- Mesos, YARN, Quincy, ...

Data Science: Big Data Analytics Stack

Big Data Analytics – Execution Engine

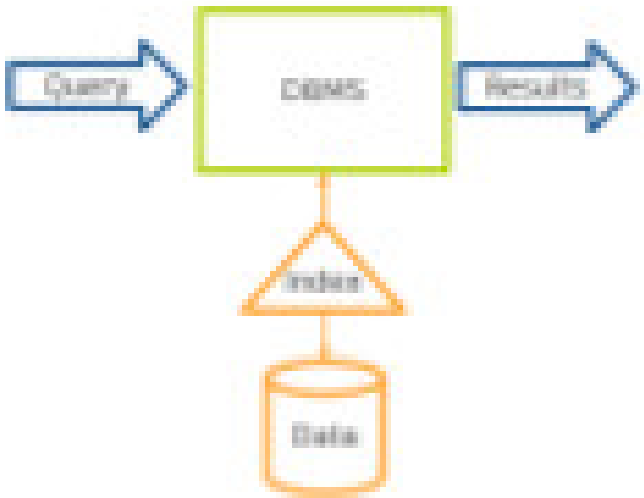
- Permiten procesamiento de datos escalable y tolerante a fallos en clusters de nodos susceptibles.
- Responden a un modelo de programación para clusters de nodos commodity.
- MapReduce, Spark, Stratosphere, Dryad, Hyracks, ...

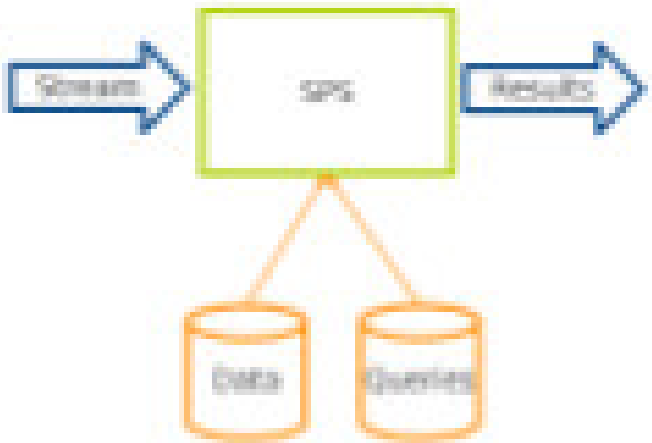
Big Data Analytics – Query/Scripting Languages

- La programación de bajo nivel para motores de ejecución, como MapReduce, no resulta fácil para usuarios finales.
- De ahí la necesidad de lenguajes de programación de alto nivel para mejorar las capacidades de queries de los motores de ejecución.
- Son capaces de traducir las funciones definidas por el usuario (user-defined functions) a la API de bajo nivel de los motores de ejecución.
- Pig, Hive, Shark, Meteor, DryadLINQ, SCOPE, ...

Big Data Analytics – Stream Processing

- Permite obtener resultados en tiempo real y con bajo latencia;
- Database Management Systems (DBMS) vs. Stream Processing Systems (SPS):





Data Science: Big Data Analytics Stack

Big Data Analytics – Graph Processing

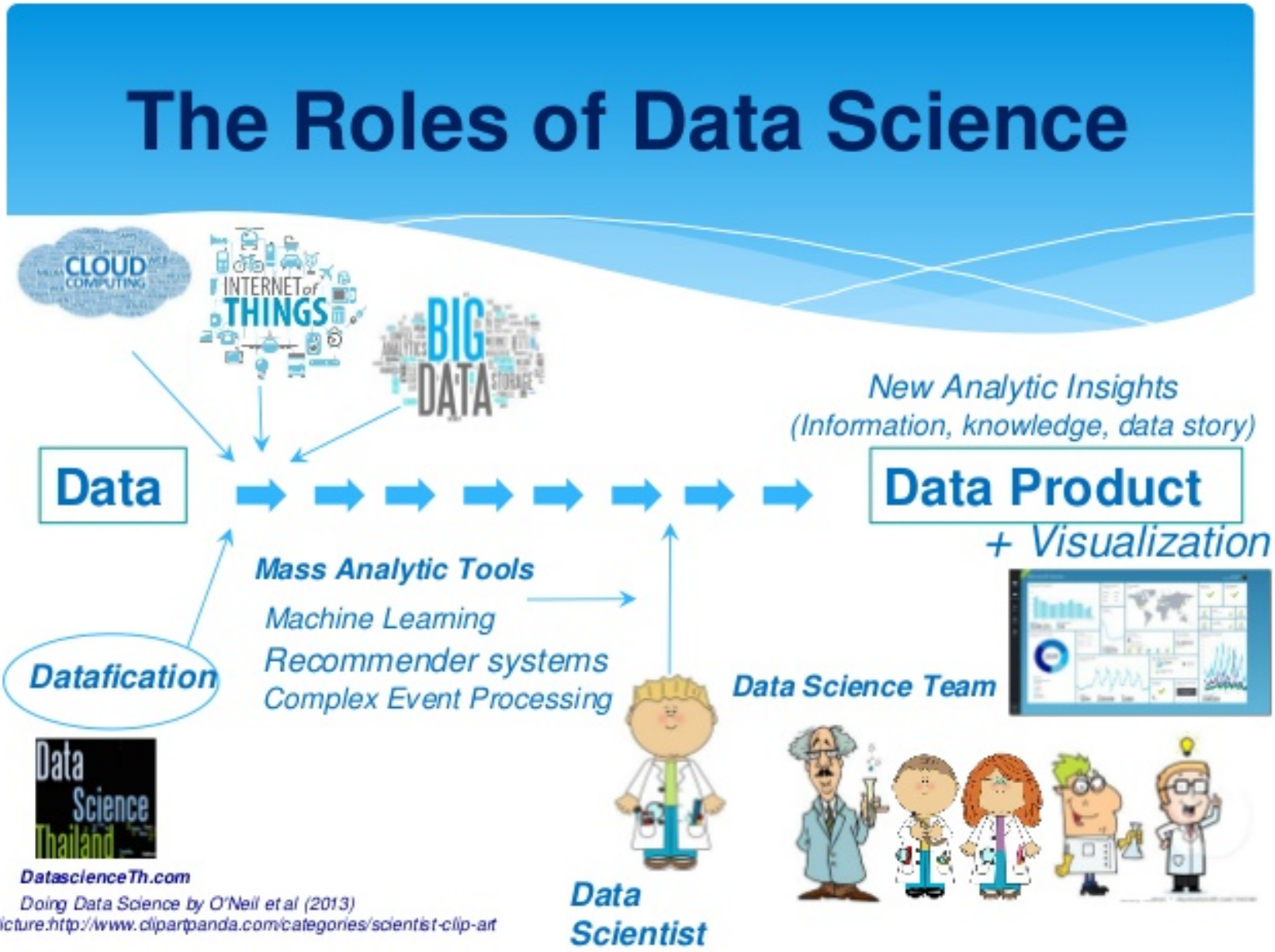
- Muchos problemas se pueden expresar usando grafos: implica dependencias computacionales dispersas y múltiples iteraciones para converger.
- Los frameworks de procesamiento paralelo de datos , como MapReduce, no son ideales para estos problemas: resulta muy lento.
- Los frameworks de procesamiento de grafos están optimizados para problemas basados en grafo.
- Pregel, Giraph, GraphX, GraphLab, PowerGraph, GraphChi, ...

Data Science: Big Data Analytics Stack

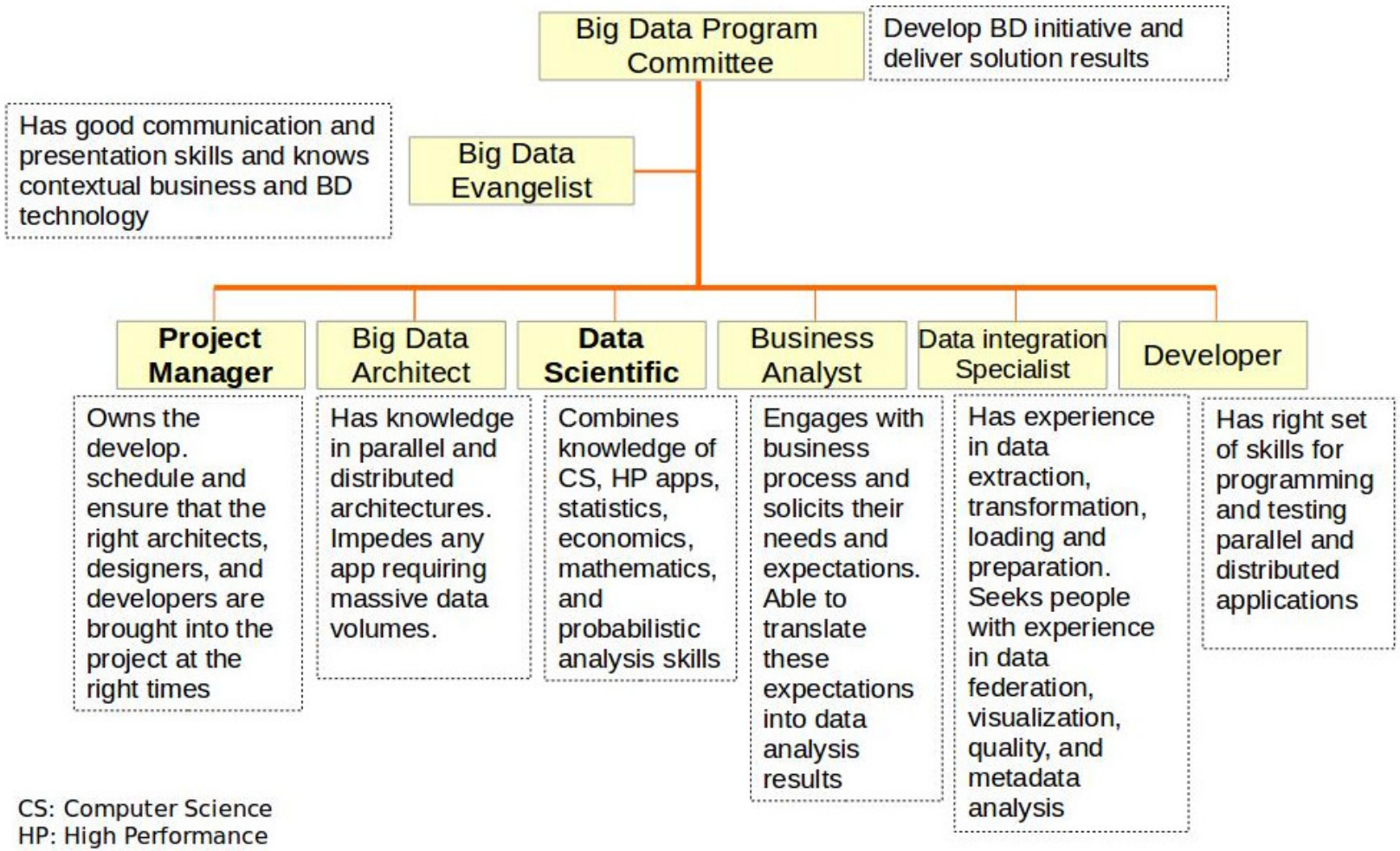
Big Data – Machine Learning

- Implementar y usar técnicas de machine learning a gran escala puede resultar en tareas complejas para desarrolladores y usuarios finales.
- Existen plataformas que facilitan tales tareas, ofreciendo librerías escalables de machine learning y minería de datos.
- Mahout, MLBase, SystemML, Ricardo, Presto, ...

Data Science: Big Data Analytics



Data Scientist



Veamos este video:

https://youtu.be/yR2wWQYiVKM

Fin

¡Muchas Gracias por su atención!

Yudith Cardinale
ycardinale@usb.ve