Big Data:
Un punto de vista tecnológico[1]

MDW, 2018

Diego Sevilla Ruiz

Dpto. de Ingeniería y Tecnología de Computadores
Facultad de Informática
Universidad de Murcia

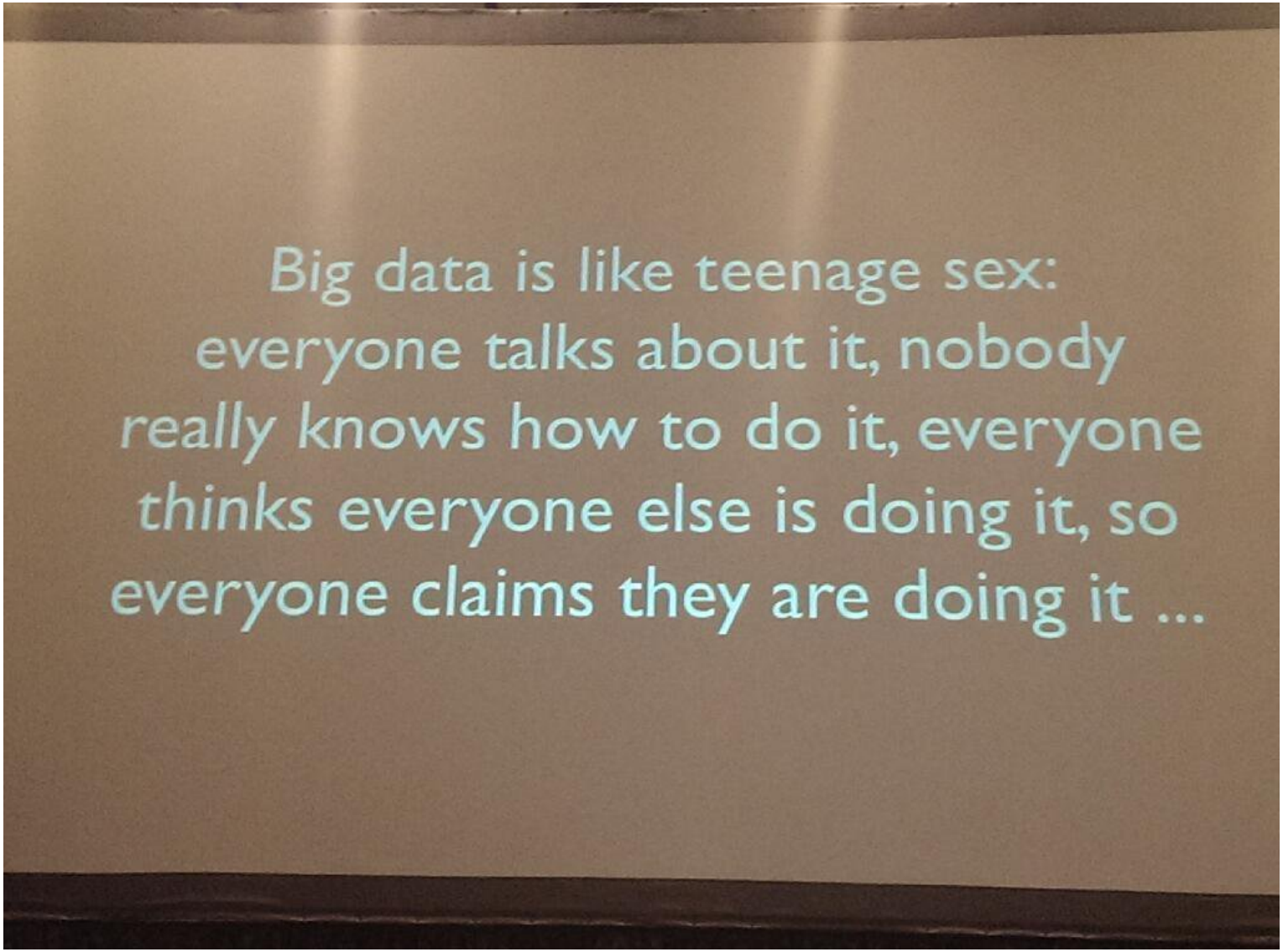
dsevilla@um.es

Junio de 2018

[1] https://github.com/dsevilla/murciadigitalweek18

Big Data

https://twitter.com/jmibl/status/390768769259163648.

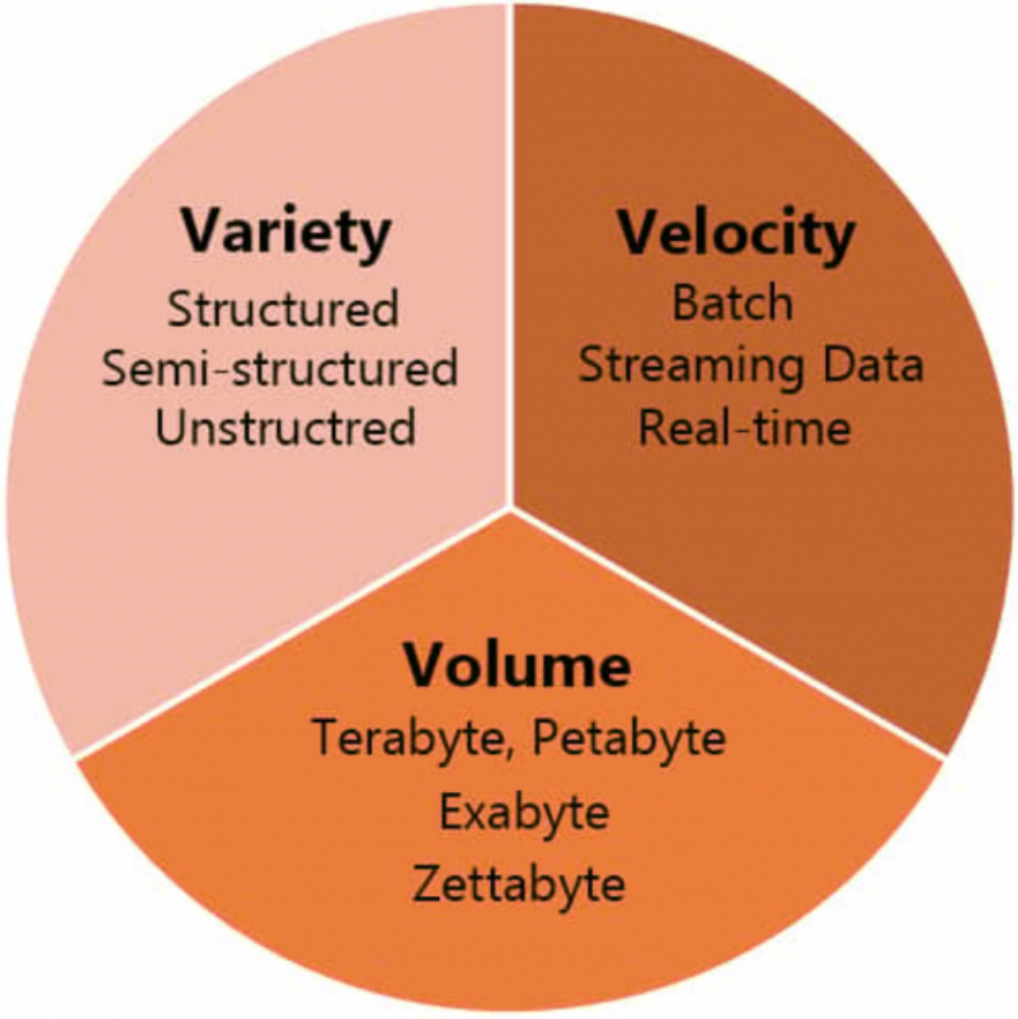


Big Data


BIG
DATA
INFORMATION
PETABYTES
STORAGE
LARGE
SETS
SEARCH
ANALYSIS
RESEARCH
MANAGEMENT
TECHNOLOGIES
VOLUME
SOFTWARE
SAN
NAS
GOVERNMENT
PROCESS
MILLION
SCIENCE
DEVELOPMENT
TECHNOLOGY
NEW
BILLION
FUTURE
PARALLEL
INTERNET
LEARNING
WORLD
BASED
EXABYTES
SIZE
DATABASES
PROCESSING
ANALYTICS
INITIATIVE
TERABYTES

Big Data – 3Vs

https://www.theviable.co/how-big-data-impact-to-corporate/3v-model-of-big-data/

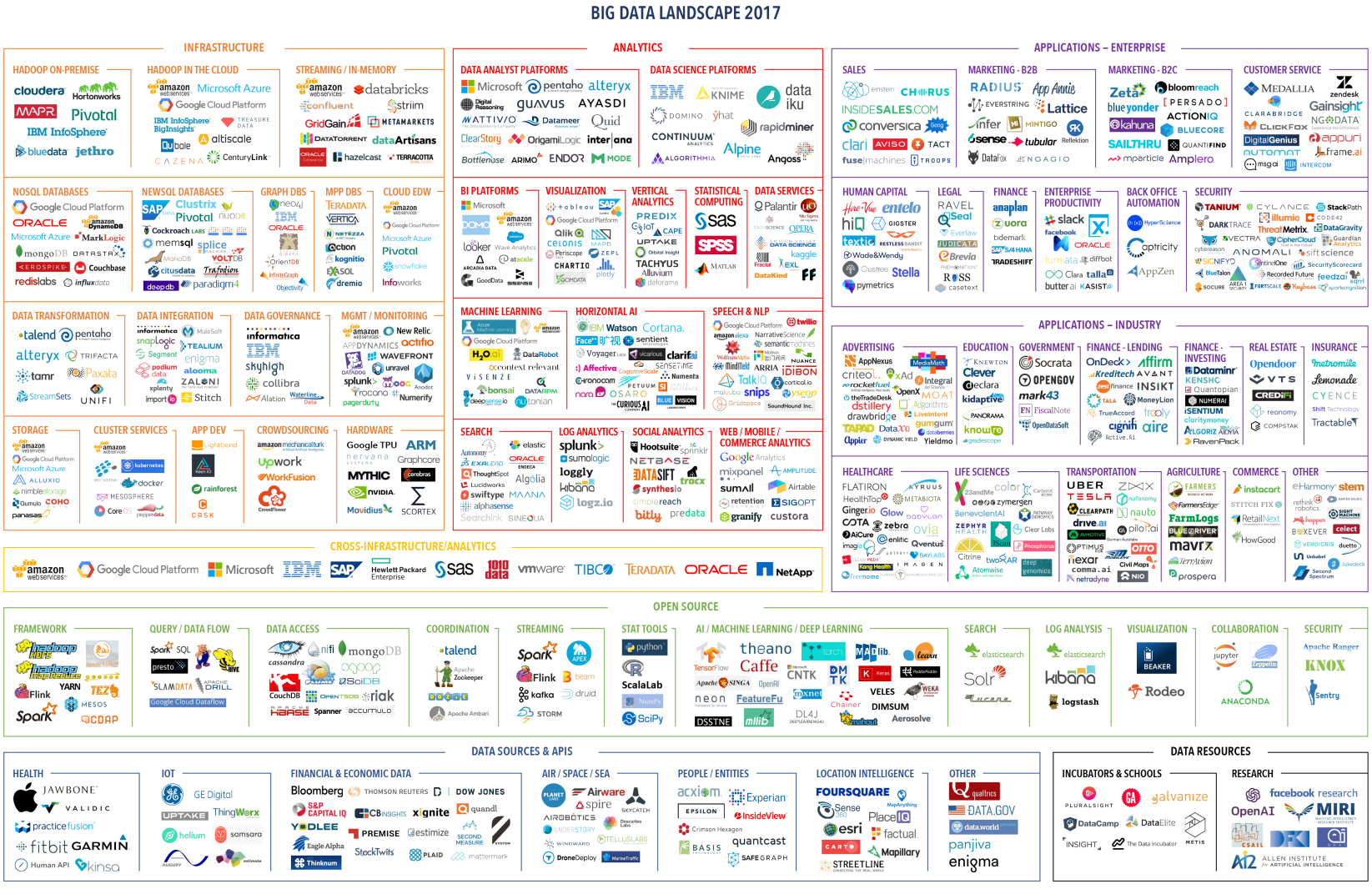


Big Data – ¡¡¡8Vs!!!


BIG DATA
with 8 V's
01
VOLUME
Can you find the information you are looking for?
02
VALUE
Can you find it when you most need it?
03
VERACITY
Are you dealing with information or disinformation?
04
VISUALISATION
Can you make sense at a glance? Does it trigger a decision?
05
VARIETY
Is a picture worth a thousand words in 70 languages? Is your information balanced?
06
VELOCITY
Information gains momentum and crises & opportunities evolve in real time. How is outlook for today?
07
VISCOSITY
Does it stick with you? Does it call for action?
08
VIRALITY
AHA to-go? Does it convey a message that can be pasted into a presentation or Instgrammed?

Big Data Landscape 2017

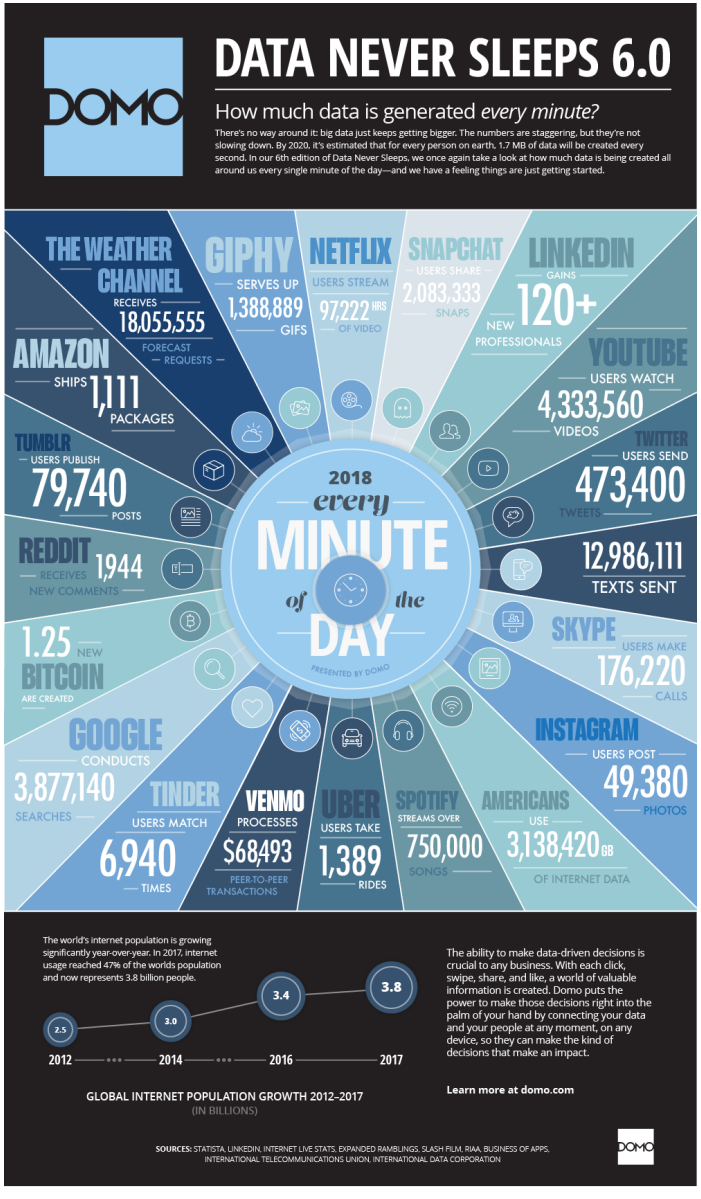
http://mattturck.com/bigdata2017/



FIRSTMARK
EARLY STAGE VENTURE CAPITAL

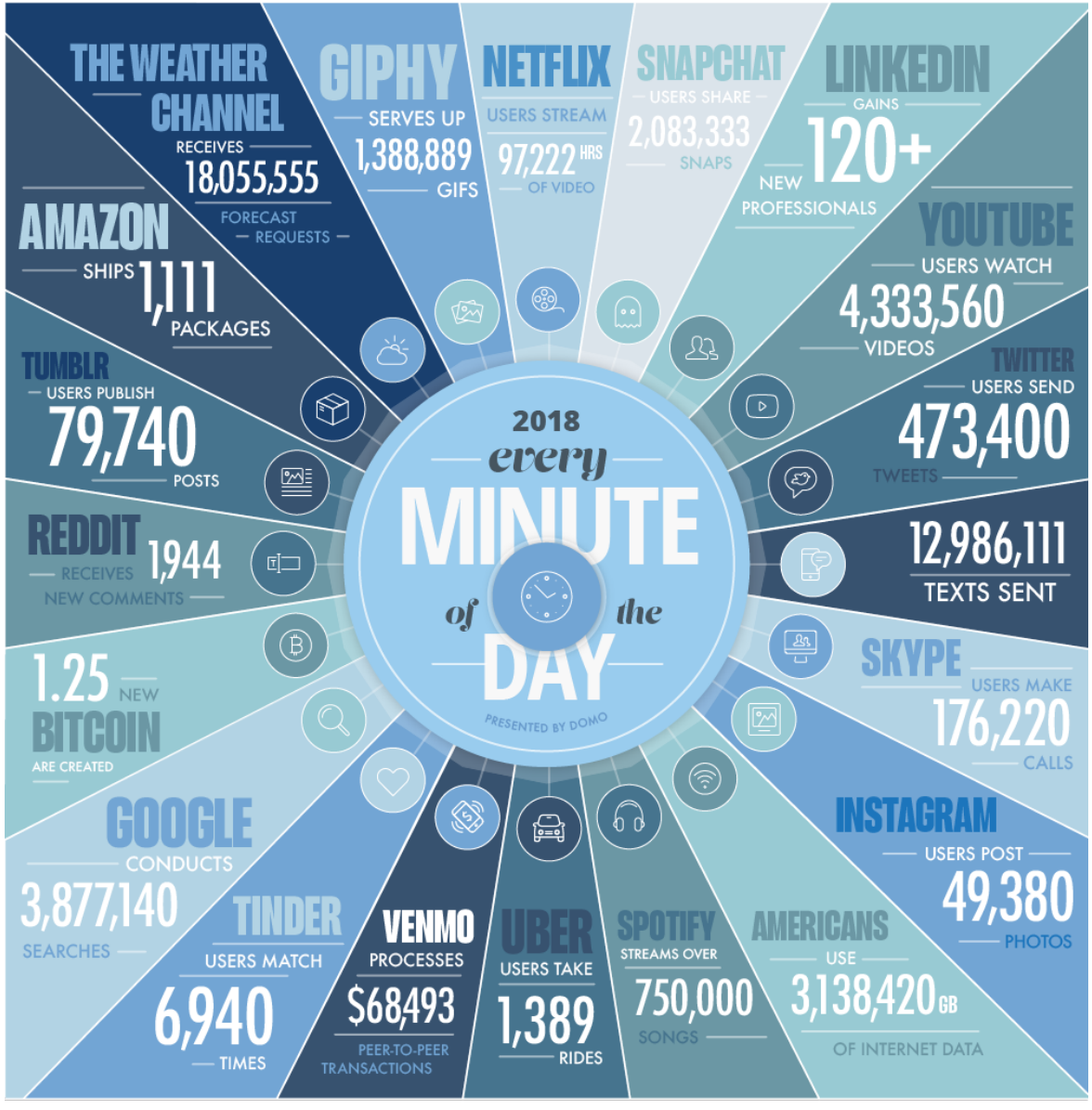
Big Data – Un minuto del día

https://www.domo.com/learn/data-never-sleeps-6

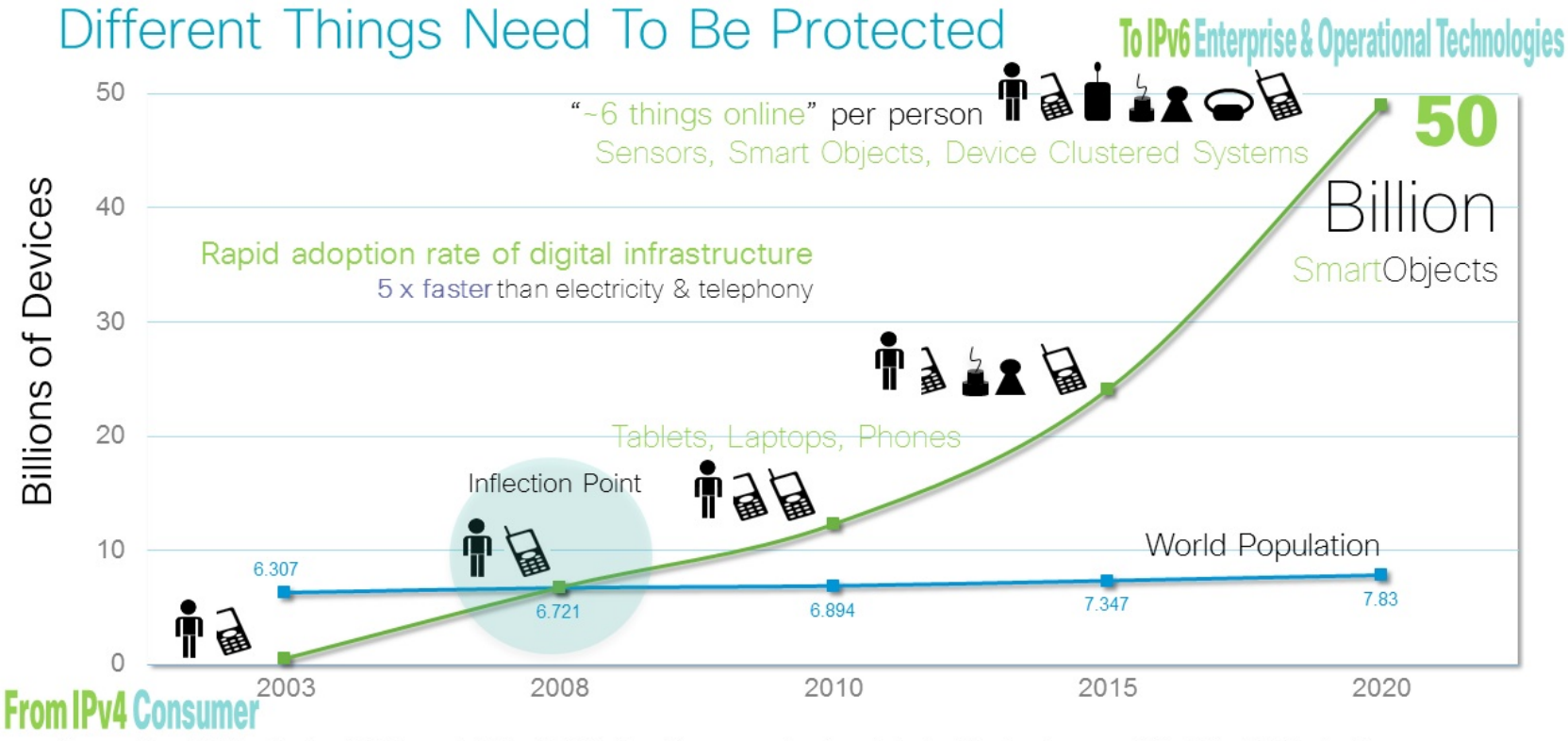


Big Data – Un minuto del día

https://www.domo.com/learn/data-never-sleeps-6



Big Data – ¡Y eso sin contar IoT!



Source: Cisco IBSG projections, UN Economic & Social Affairs http://www.un.org/esa/population/publications/longrange2/WorldPop2300final.pdf

Big Data – Airbus A350

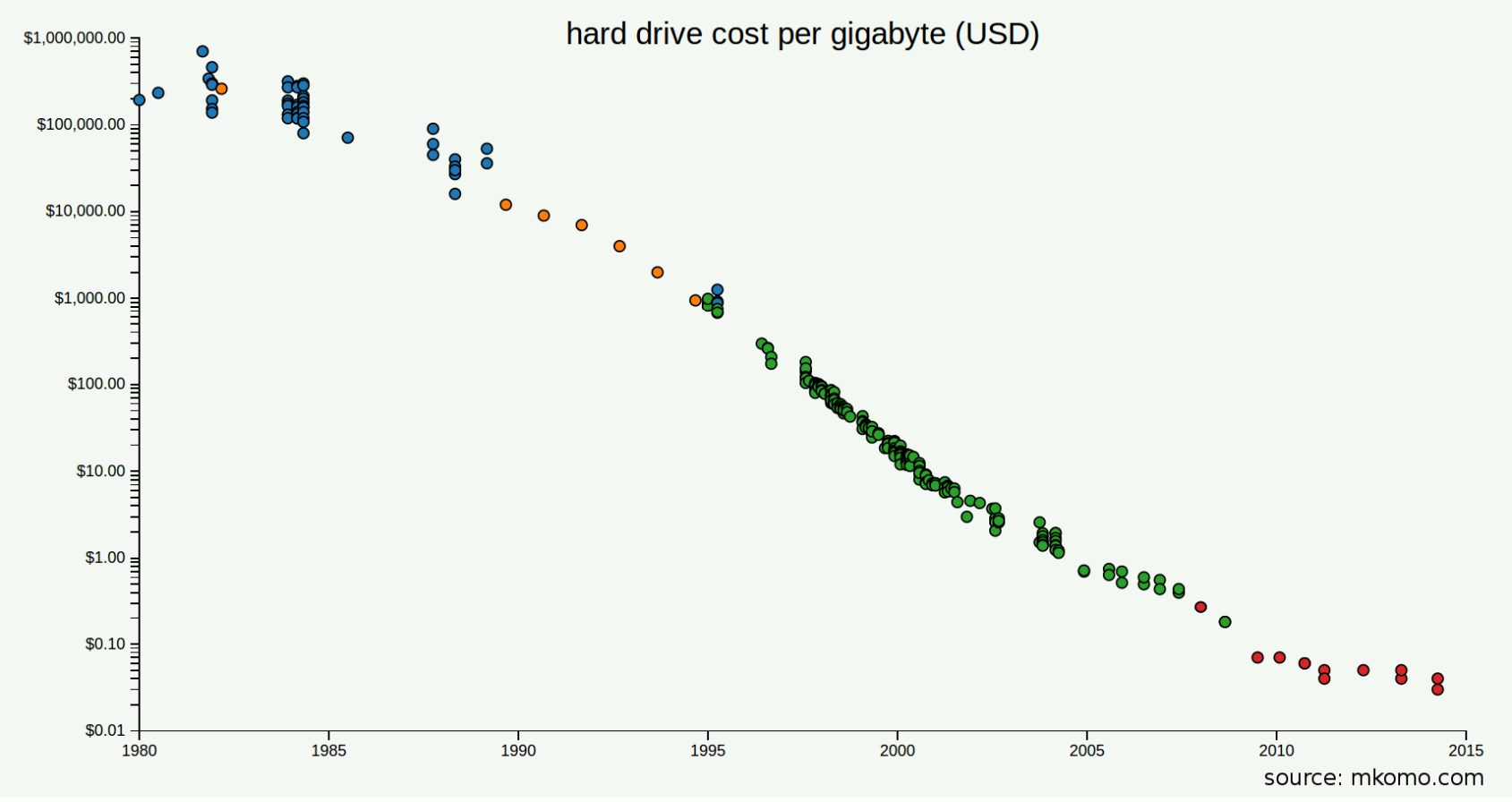
https://siliconsemiconductor.net/article/102842/Aviation_depends_on_sensors_and_big_data

While the need for ever-increasing sensor data is to be expected in such a time sensitive and safety conscious industry as air travel, it may be surprising to realize that each and every Airbus A350 generates 2.5 Tb of data each day it operates. When the new A380-1000 debuts, that data cache will more than triple. Aircraft operational data is also complex: imagine the data from 10,000 wing sensors entwined with data streams from literally thousands of other sensors.

Big Data, entonces...

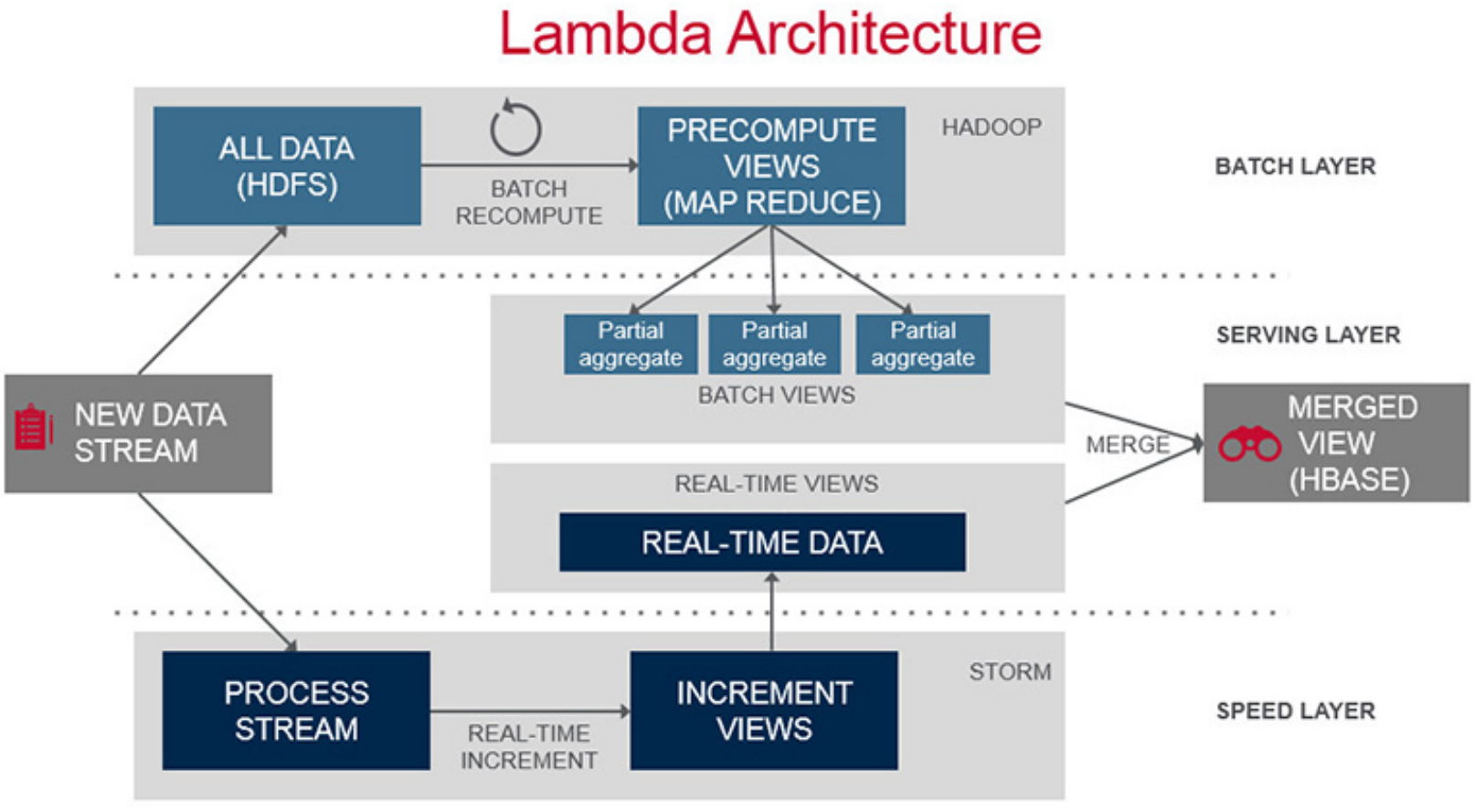
- Recopilar todos los datos posibles
  - Después serán útiles, se podrán analizar
  - El coste de almacenamiento cada vez es menor
  - Si no $\Rightarrow$ coste de oportunidad
- *Data Science*
  - El análisis ofrecerá ***conocimiento*** para mejorar (**valor**)
- Imposible procesar todo ***online***:
  - Separación entre capa *batch* y capa *online*
  - Lambda Architecture

Coste por GB


hard drive cost per gigabyte (USD)
$1,000,000.00
$100,000.00
$10,000.00
$1,000.00
$100.00
$10.00
$1.00
$0.10
$0.01
1980
1985
1990
1995
2000
2005
2010
2015
source: mkomo.com

Lambda Architecture

https://mapr.com/developercentral/lambda-architecture/



Data Science

Howe, 2013

"Next sexy job"

"The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill."

— Hal Varian, Google

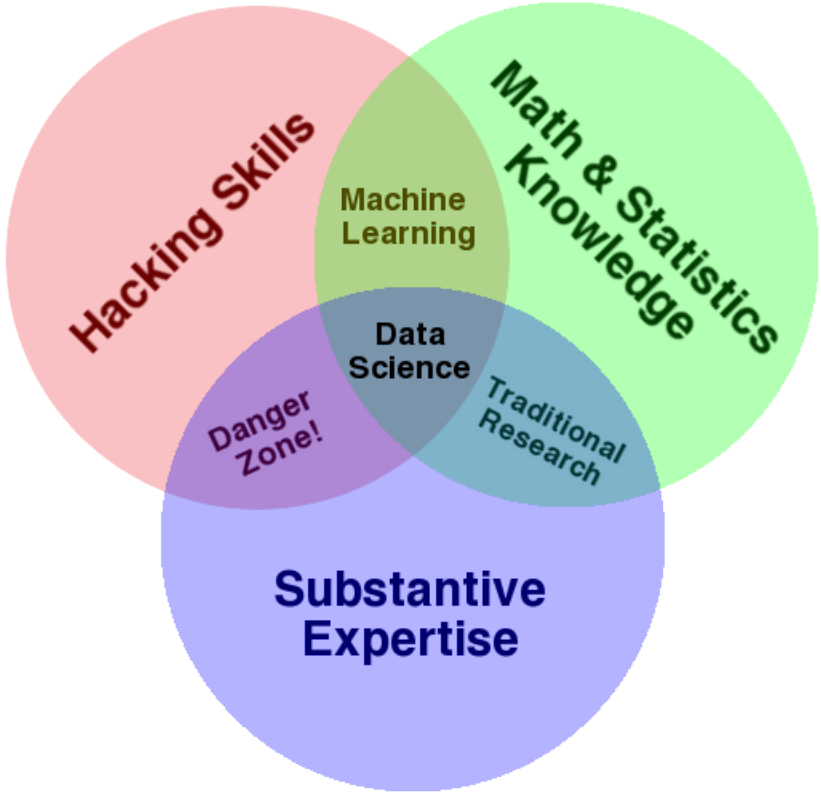
"Data science is the civil engineering of data. Its acolytes possess a practical knowledge of tools & materials, coupled with a theoretical understanding of what's possible."

— Mike Driscoll, Metamarkets

Data Science

http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Drew Conway Venn Diagram of Big Data:



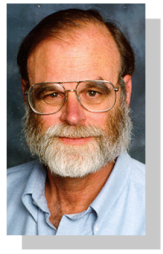
Data Science

Mike Driscoll's three sexy skills of data geeks

- Statistics
  - traditional analysis
- Data Munging
  - parsing, scraping, and formatting data
- Visualization
  - graphs, tools, etc.

(data wrangling, data jujitsu, data munging)

Data Access Hitting a Wall



Current practice based on data download (FTP/GREP)
Will not scale to the datasets of tomorrow

- You can GREP 1 MB in a second
- You can GREP 1 GB in a minute
- You can GREP 1 TB in 2 days
- You can GREP 1 PB in 3 years.

- Oh!, and 1PB ~5,000 disks

- At some point you need
  indices to limit search
  parallel data search and analysis
- This is where databases can help

- You can FTP 1 MB in 1 sec
- You can FTP 1 GB / min (~1$)
- … 2 days and 1K$
- … 3 years and 1M$

[slide src: Jim Gray]

eScience, Data Science, 4º paradigma

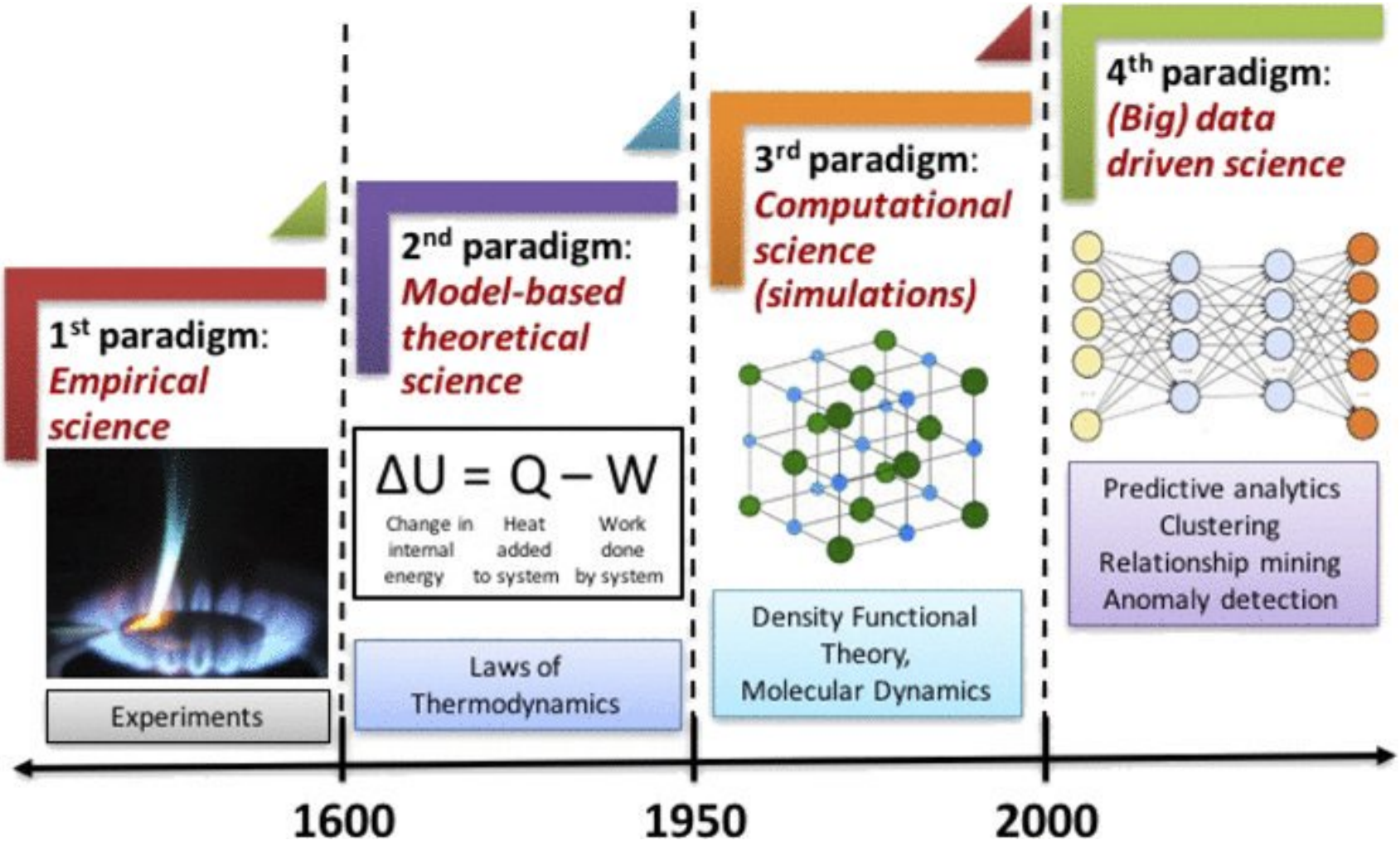
- Tradicionalmente, la ciencia se desarrollaba de forma **empírica**, por observación, o reproduciendo condiciones en el laboratorio
- Desde hace unos cientos de años, los **modelos teóricos** también se han aceptado como una forma de explicar sucesos, y sugerir nuevos experimentos
- En los últimos ~50 años, **la simulación** se ha usado para reproducir condiciones especiales o no reproducibles. Modelos teóricos demasiado complejos para resolverlos analíticamente, parte de un estado inicial y comprueba a dónde se llega

4º paradigma (ii)

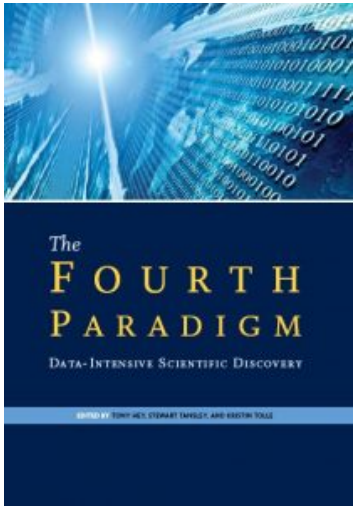
- Hoy: **exploración basada en datos**
  - Unifica teoría, experimentación y simulación
  - Los datos se capturan por instrumentos o bien se generan por simulaciones
  - Son procesados por software
  - La información (y el conocimiento extraído) se almacenan en un ordenador
  - Los científicos analizan los ficheros/bases de datos usando nuevas herramientas estadísticas y bases de datos capaces de gestionar cada vez más datos
  - En vez de "**preguntar al mundo**", se obtienen resultados de combinar conjuntos "**descargados**" de datos de formas no previstas anteriormente

The Fourth Paradigm

https://twitter.com/aip_publishing/status/856825353645559808



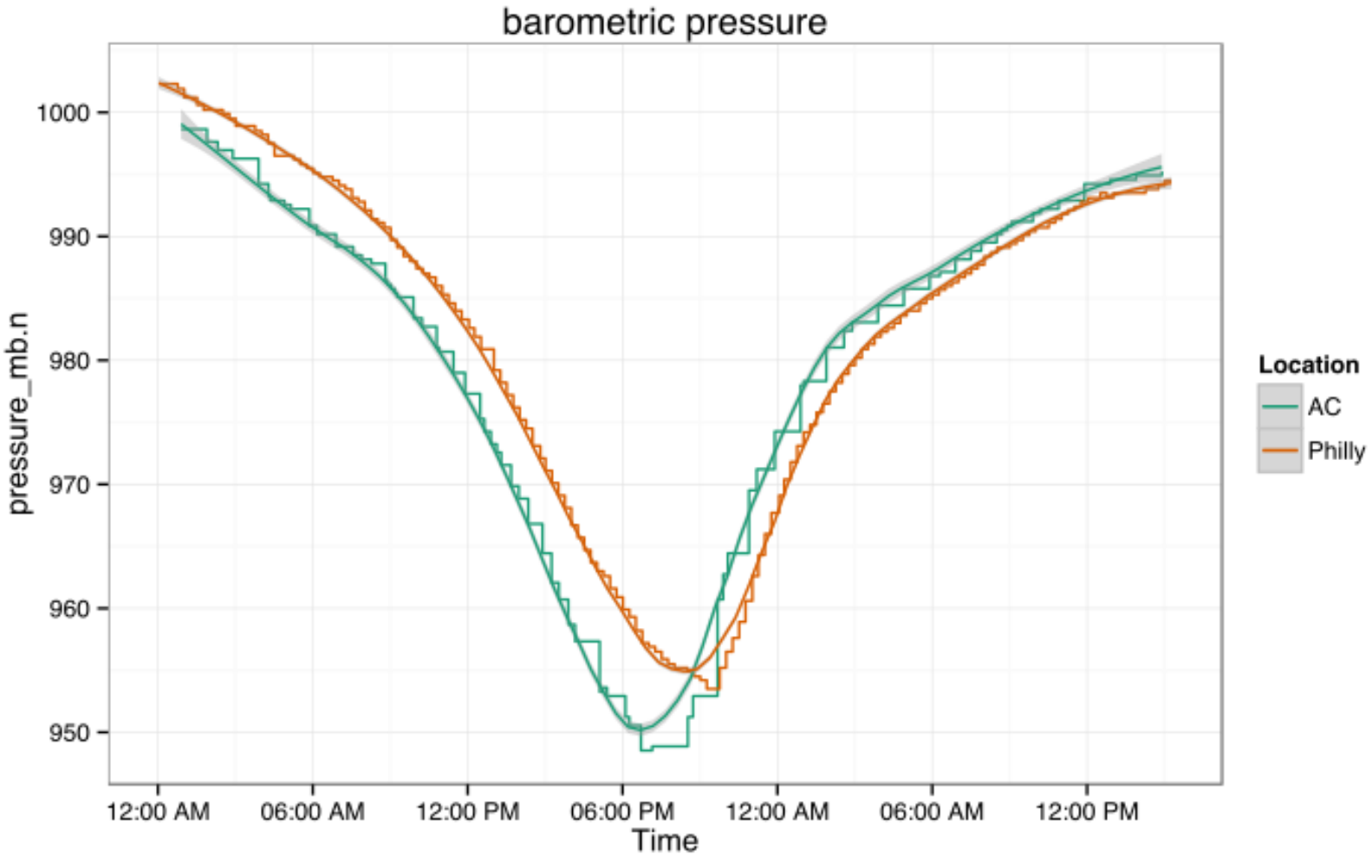
The Fourth Paradigm



Jim Gray. *The Fourth Paradigm*, Microsoft Research, 2009

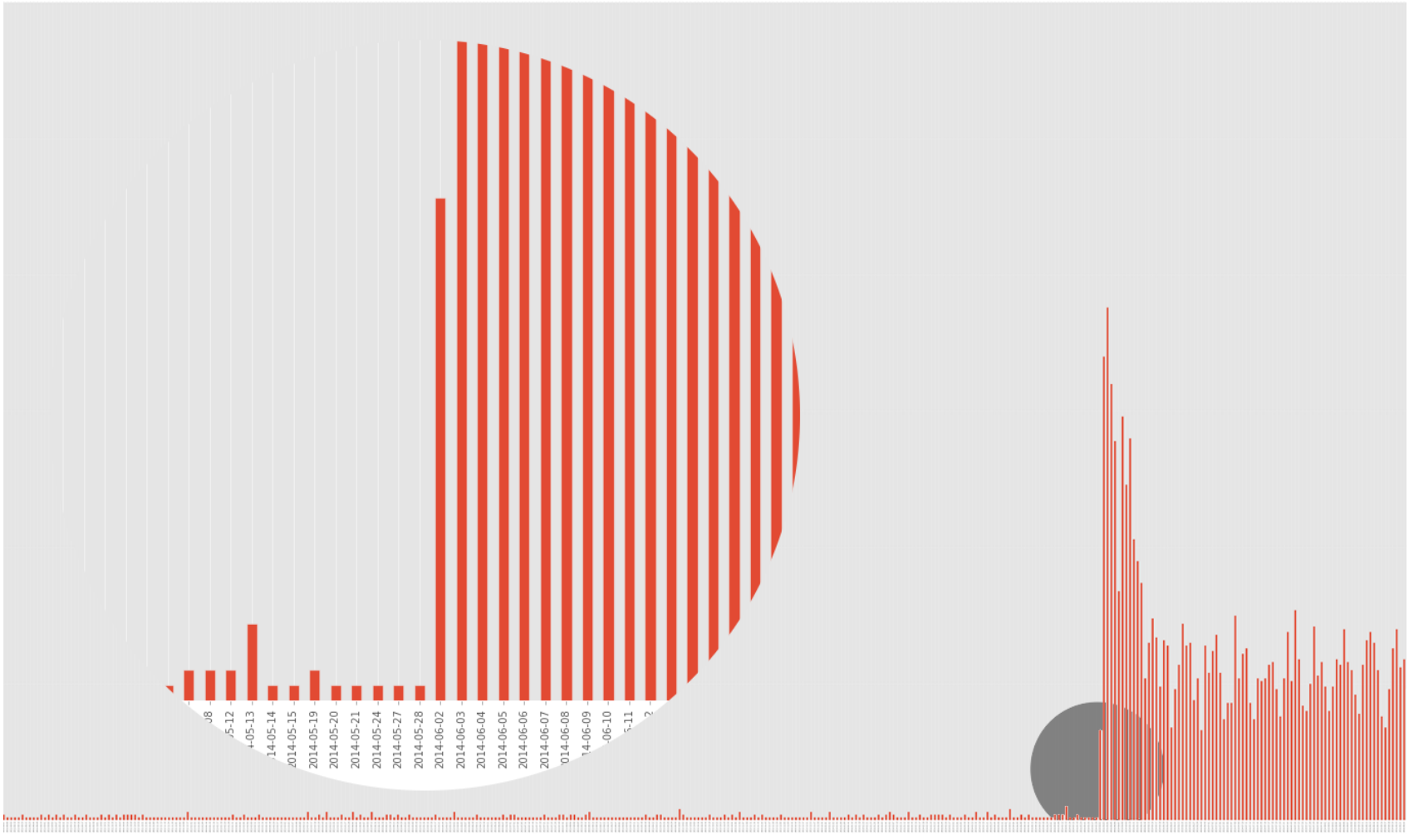
Huracán Sandy, 2012

`http://rpubs.com/JoFrhwld/sandy` (Howe, 2013)



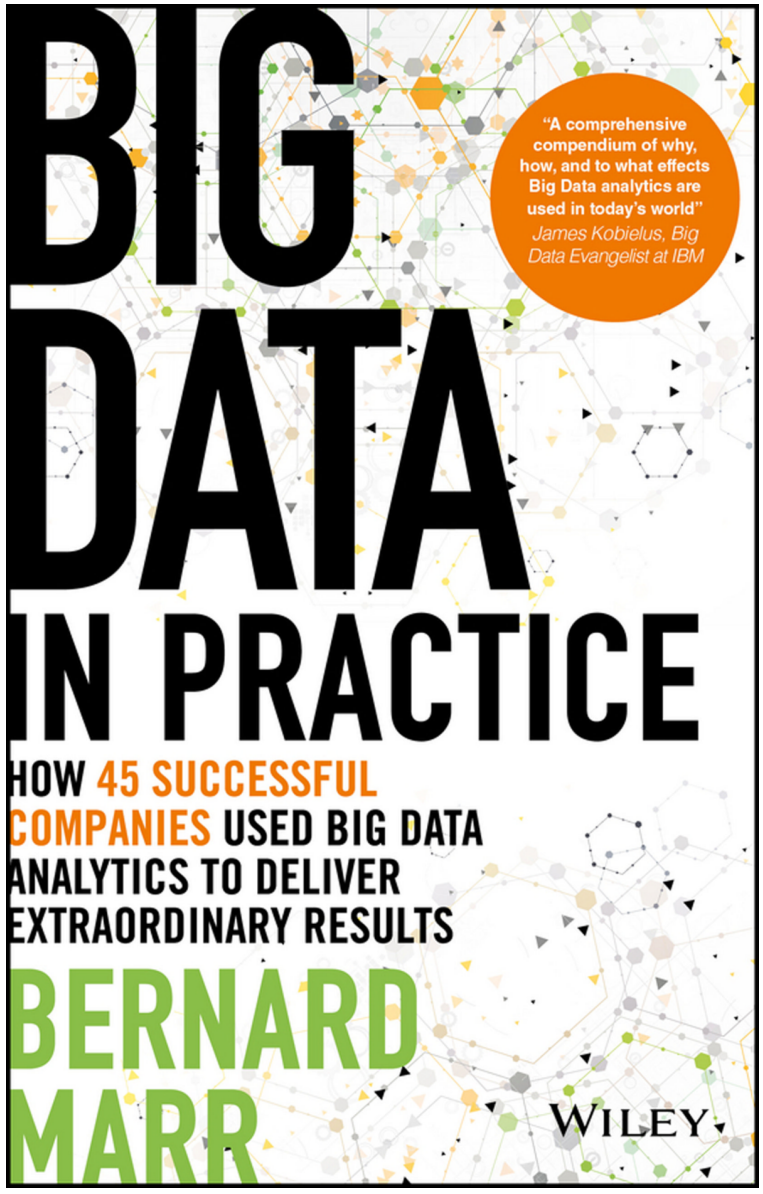
¿Cuándo introdujo Apple el “Swift”?

Una mirada al tag “swift” de todas las preguntas de Stackoverflow



Big Data – Casos de éxito

Bernard Marr – Big Data in Practice



Big Data – Caso de éxito: Walmart

- Walmart tiene 20.000 tiendas en 28 países
- En 2015 se propusieron procesar 2,5 Petabytes de información cada hora
- Formó lo que denominó “Data Café”, que procesaba la base de datos on-line de 40 Petabytes
- Redujeron tiempo de respuesta a problemas de semanas a **20 minutos**
- Por ejemplo, detectaron que no se vendía un producto que sí se vendía en otras sucursales $\Rightarrow$ Los empleados habían olvidado ponerlo en las estanterías

Big Data – Caso de éxito: CERN

- Sólo el *Large Hadron Collider* (LHC) genera 30 Petabytes al año
- Los sensores del LHC detectan cientos de millones de colisiones de partículas
- Hay que procesar esos datos para encontrar patrones de colisión para detectar partículas
- 300 GB de datos por segundo $\Rightarrow$ 300 MB/s filtrados

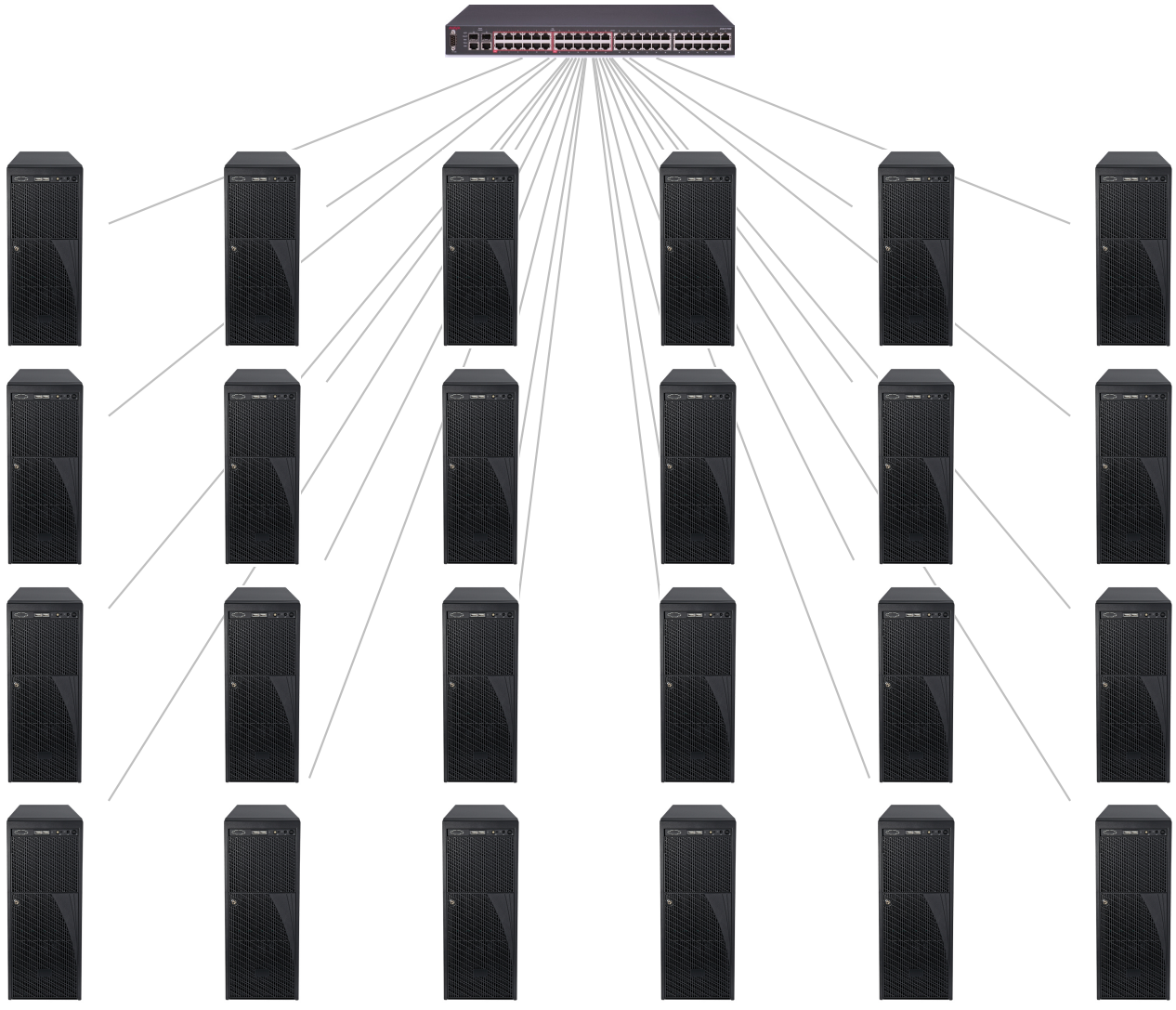
Big Data – Caso de éxito: Netflix

- Inicialmente sólo tenía datos que asociaban un cliente, una película alquilada, la calificación y la fecha
- Conforme se convirtió en *streaming*, pudo recopilar un conjunto mayor de datos
- Añadió *tags* a cada película y episodio, lo que permitió generar contenido que, previsiblemente, iba a ser bien recibido
- Oracle, después NoSQL y Cassandra
- También utiliza tecnologías de Big Data como Hadoop, Pig y Hive (más después)

Big Data – Casos de éxito

Y además: Rolls-Royce, Shell, Apixio, Lotus F1 Team, Pendleton & Son Butchers, US Olympic Women's Cycling Team, ZSL, Facebook, John Deere, Royal Bank of Scotland, LinkedIn, Microsoft, Acxiom, US Immigration And Customs, Nest, GE, Etsy, Narrative Science, BBC, Milton Keynes, Palantir, Airbnb, Sprint, Dickey's Barbecue Pit, Caesars, Fitbit, Ralph Lauren, Zynga, Autodesk, Walt Disney Parks and Resorts, Experian, Transport for London, The US Government, IBM Watson, Google, Terra Seismic, Apple, Twitter, Uber, Electronic Arts, Kaggle, Amazon, etc.

Cambio de perspectiva: Red



Cambio de perspectiva: Red

Procesamiento distribuido

- Necesidad de **paralelización máxima**
- **Escalabilidad**
- Explotación de la **localidad de los datos**:
  - Datos producidos en cada nodo se utilizan en siguientes iteraciones
  - Cada nodo hace de **servidor** para recibir datos

Cambio de perspectiva: Red

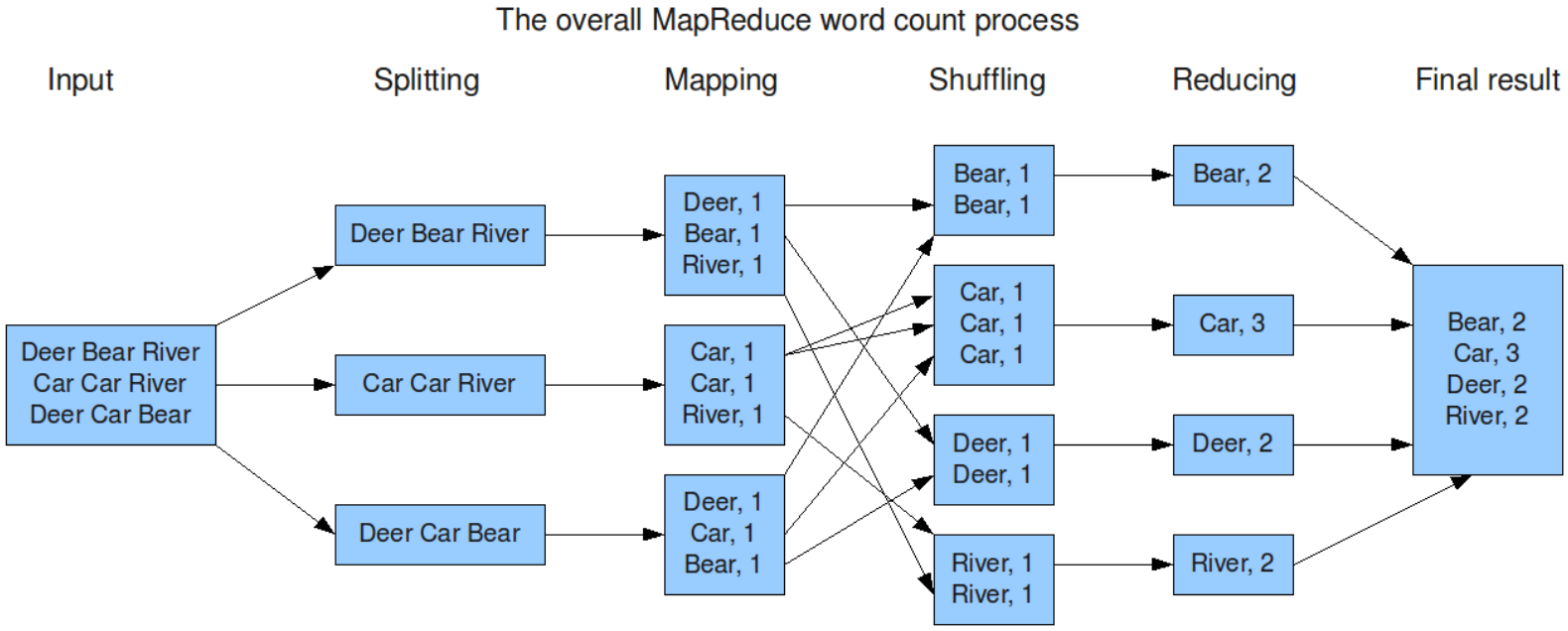
Procesamiento distribuido

- Vuelta al modelo funcional inherentemente paralelo: (e.g. **Map-Reduce**)
- Almacenamiento distribuido: (e.g. **HDFS**)
- Coordinación distribuida: (e.g. **Zookeeper**)

Map-Reduce

- Se aplica una misma operación `map()` a cada dato residente en un nodo se realiza de forma paralela en **todos** los nodos simultáneamente
- Con los resultados parciales de cada nodo, una función `reduce()` genera un resultado (o un conjunto de resultados) final
- Hay un proceso intermedio de *shuffle* para agrupar valores relacionados antes del `reduce()`
- Resultados parciales en el mismo nodo (localidad) $\Rightarrow$ procesamientos **en cadena**

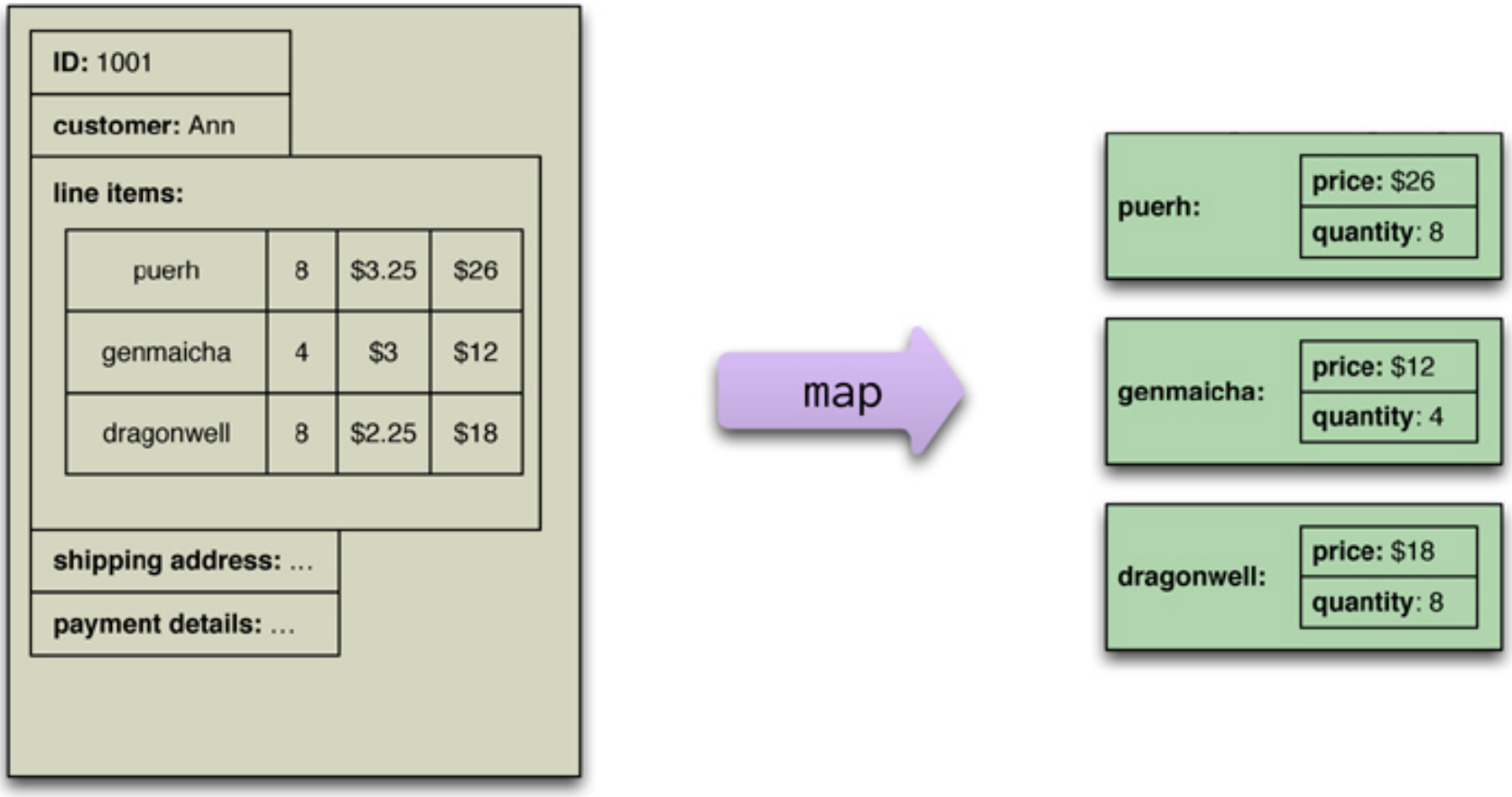
Map-Reduce (II)



(de `http://www.milanor.net/blog/an-example-of-mapreduce-with-rmr2/`)

Map-Reduce aplicado a la empresa

NoSQL Distilled. Sadalage, Fowler, Addison-Wesley, 2012

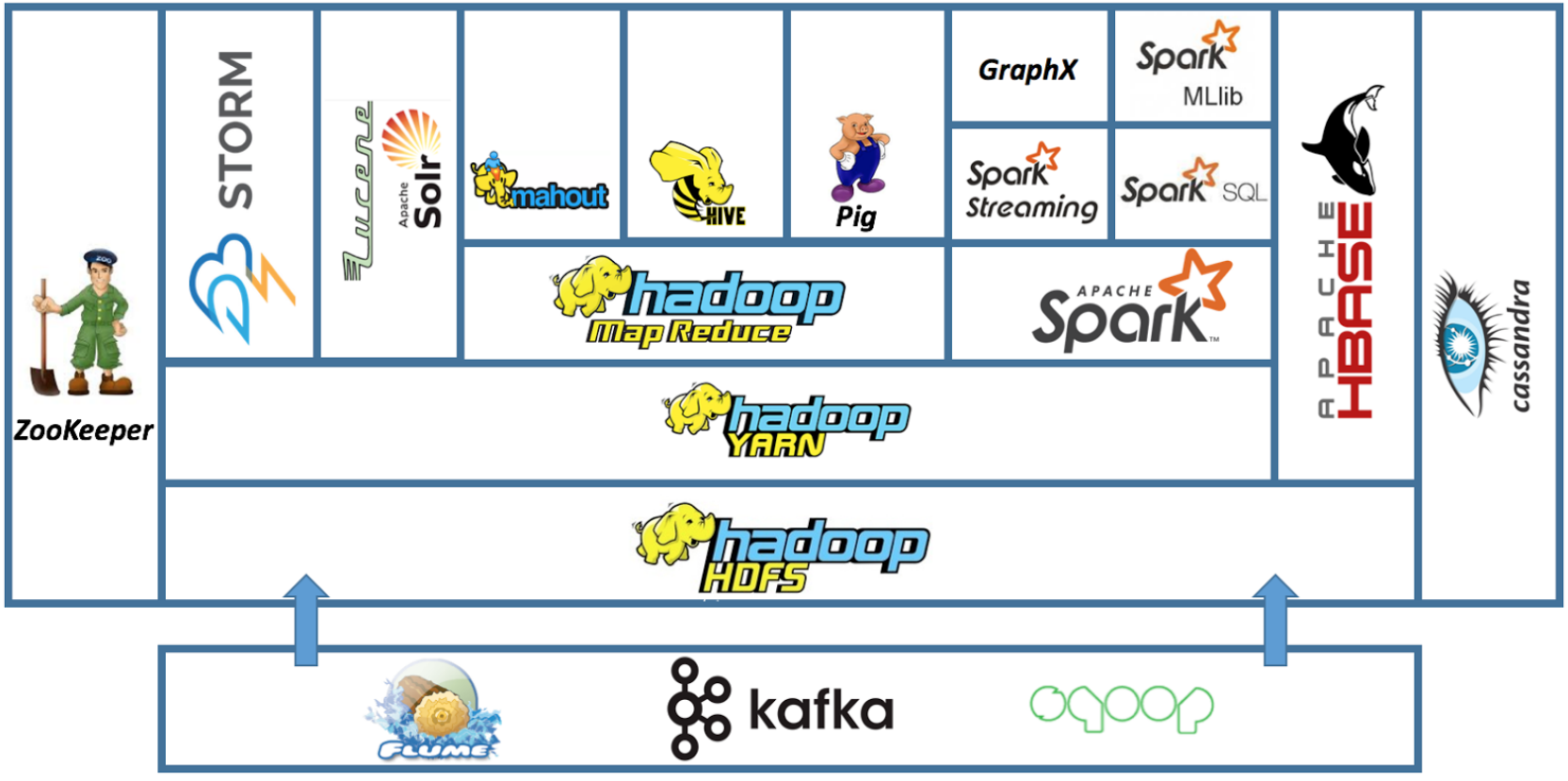


Map-Reduce aplicado a la empresa (II)

NoSQL Distilled. Sadalage, Fowler, Addison-Wesley, 2012

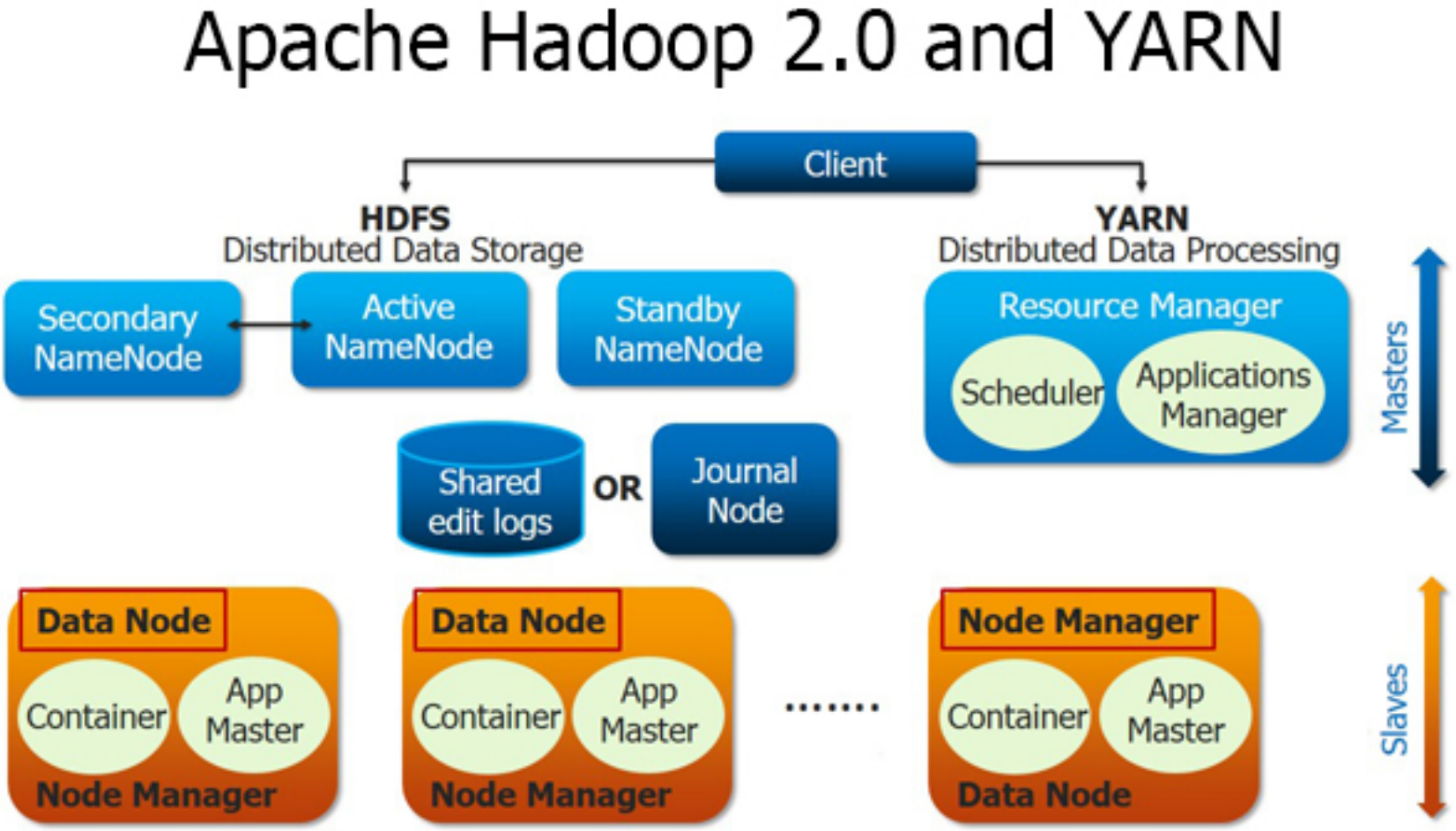
Hadoop

http://blog.newtechways.com/2017/10/apache-hadoop-ecosystem.html

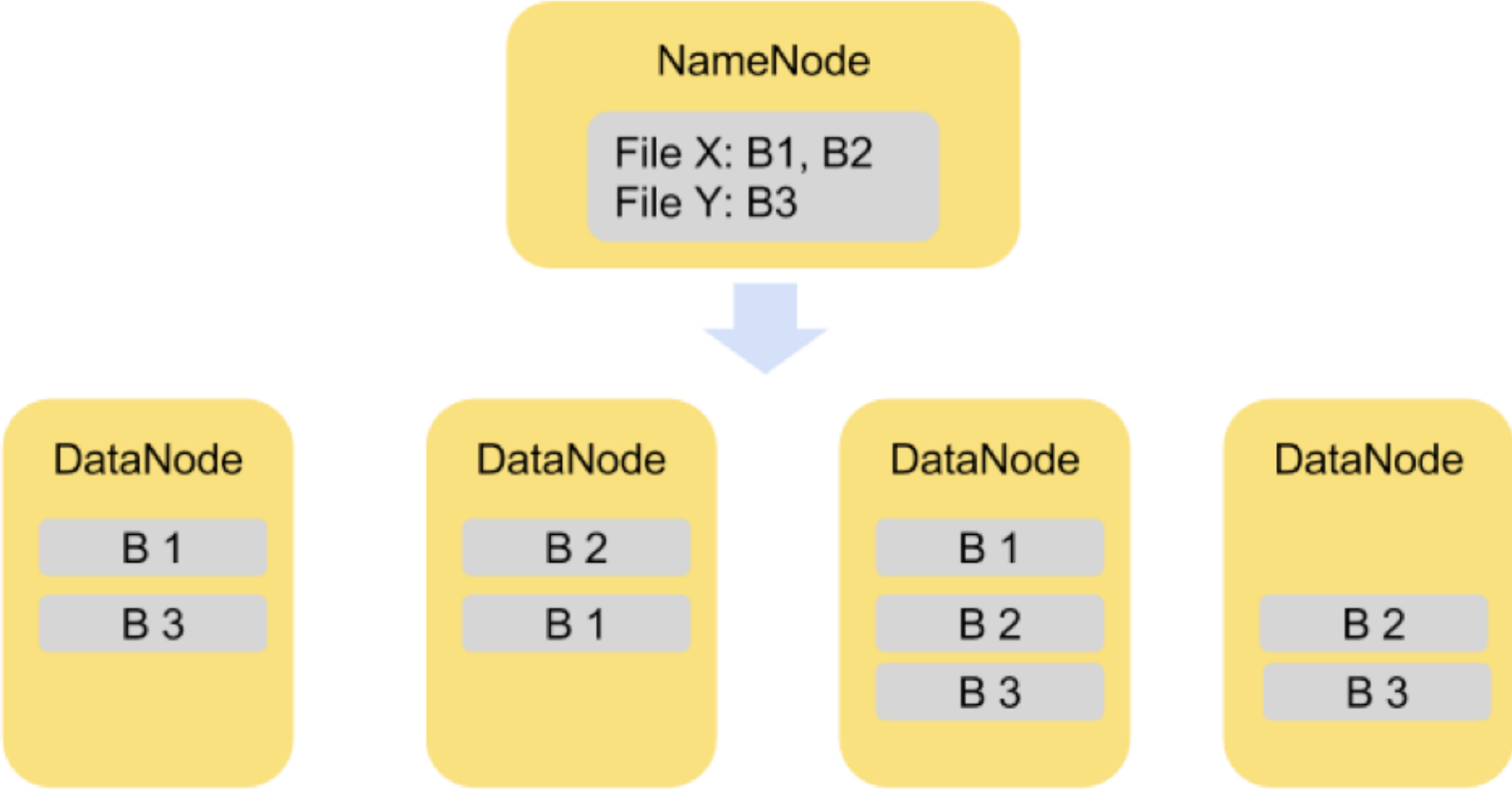


Hadoop – YARN

http://bigdataanalyticsnews.com/hadoop-yarn-adds-application-threads-big-data-users/

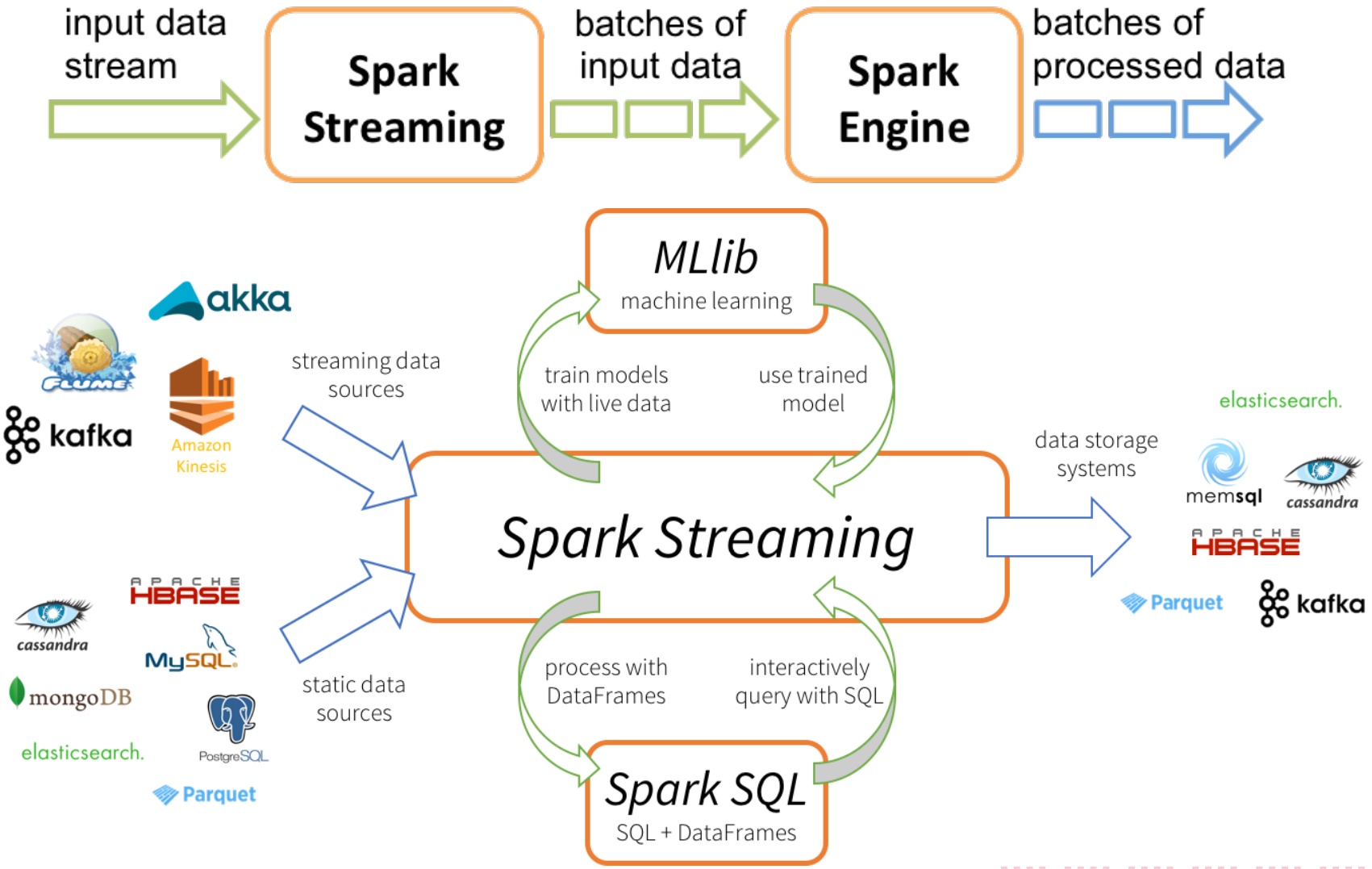


HDFS


NameNode
File X: B1, B2
File Y: B3
DataNode
B 1
B 3
DataNode
B 2
B 1
DataNode
B 1
B 2
B 3
DataNode
B 2
B 3

Spark Streaming

https://www.datanami.com/2015/11/30/spark-streaming-what-is-it-and-whos-using-it/



NoSQL


OrientDB
DynamoDB
CouchDB
Redis
Solr
MarkLogic
RavenDB
MongoDB
xDB
NoSQL
Cassandra
Voldemort
Riak
eXist-db
Elasticsearch
Berkeley
CouchBase
Hadoop
ObjectDB
Neo4J

NoSQL

- Bases de datos **orientadas a la escalabilidad**
- Normalmente **de código abierto**
- **No utilizan SQL** como lenguaje de consulta
- Utilizan alguna variación de **Map-Reduce**
- Proponen un modelo de datos más flexible que el relacional
- Se definen varios tipos:
  - Documentales
  - Columnares
  - Grafos

Representación relacional de un CV

Kleppmann, 2016. *Designing Data Intensive Applications*

http://www.linkedin.com/in/williamhgates

Bill Gates

Greater Seattle Area | Philanthropy

Summary

Co-chair of the Bill & Melinda Gates Foundation. Chairman, Microsoft Corporation. Voracious reader. Avid traveler. Active blogger.

Experience

Co-chair • Bill & Melinda Gates Foundation
2000 – Present

Co-founder, Chairman • Microsoft
1975 – Present

Education

Harvard University
1973 – 1975

Lakeside School, Seattle

Contact Info

Blog: thegatesnotes.com
Twitter: @BillGates

users table

user_id	first_name	last_name	summary
251	Bill	Gates	Co-chair of … blogger.

region_id	industry_id	photo_id
us:91	131	57817532

regions table

id	region_name
us:7	Greater Boston Area
us:91	Greater Seattle Area

industries table

id	industry_name
43	Financial Services
48	Construction
131	Philanthropy

positions table

id	user_id	job_title	organization
458	251	Co-chair	Bill & Melinda Gates F…
457	251	Co-founder, Chairman	Microsoft

education table

id	user_id	school_name	start	end
807	251	Harvard University	1973	1975
806	251	Lakeside School, Seattle	NULL	NULL

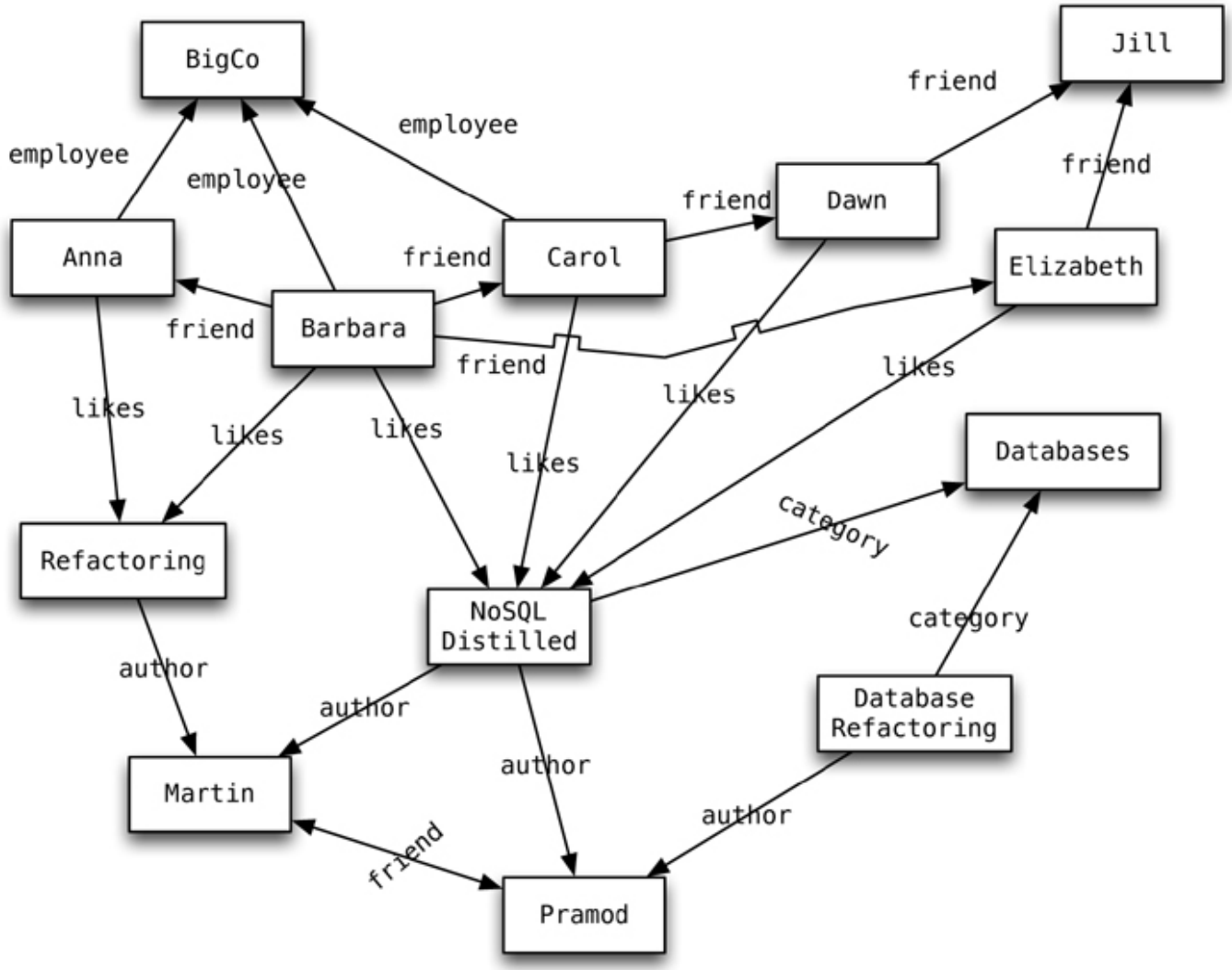
contact_info table

id	user_id	type	url
155	251	blog	http://thegatesnotes.com
156	251	twitter	http://twitter.com/BillGates

```
{
  ”user_id”: 251,
  ”first_name”: ”Bill”,
  ”last_name”: ”Gates”,
  ”summary”: ”Co-chair of the Bill & Melinda Gates... Active blogger.”,
  ”region_id”: ”us:91”,
  ”industry_id”: 131,
  ”photo_url”: ”/p/7/000/253/05b/308dd6e.jpg”,
  ”positions”: [
    {
      ”job_title”: ”Co-chair”,
      ”organization”: ”Bill & Melinda Gates Foundation”
    },
    {
      ”job_title”: ”Co-founder, Chairman”,
      ”organization”: ”Microsoft”
    }
  ],
  ”education”: [
    {
      ”school_name”: ”Harvard University”,
      ”start”: 1973,
      ”end”: 1975
    },
    {
      ”school_name”: ”Lakeside School, Seattle”,
      ”start”: null,
      ”end”: null
    }
  ],
  ”contact_info”: {
    ”blog”: ”http://thegatesnotes.com”,
    ”twitter”: ”http://twitter.com/BillGates”
  }
}
```

Grafos

NoSQL Distilled. Sadalage, Fowler, Addison-Wesley, 2012



Datos y consultas en Neo4J

Datos:

```
CREATE
  (NAmerica:Location {name:'North America', type:
     'continent'}),
  (USA:Location {name:'United States', type:'
     country'}),
  (Idaho:Location {name:'Idaho', type:'state'}),

  (Lucy:Person {name:'Lucy' }),

  (Idaho)-[:WITHIN]->(USA)-[:WITHIN]->(NAmerica),
  (Lucy) -[:BORN_IN]-> (Idaho)
```

Datos y consultas en Neo4J (II)

Consulta:

```
MATCH
(person)-[:BORN_IN]->()-[:WITHIN*0..]->(us:
    Location {name:'United States'}),
(person)-[:LIVES_IN]->()-[:WITHIN*0..]->(eu:
    Location {name:'Europe'})
RETURN person.name
```

(con esta consulta tan cercana al lenguaje natural, estamos buscando los emigrantes de EEUU en Europa)