Big Data

Dr. Wenceslao Palma
wenceslao.palma@pucv.cl

Big Data


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


El uso de aplicaciones para contener el avance del Coronavirus en China, Corea y Singapur
El desarrollo de big data y geolocalización para optimizar la vigilancia sanitaria en tres países del Asia ha tenido consecuencias positivas en el control de la crisis epidemiológica. Sin embargo, el uso de datos personales no ha estado exento de críticas y desconfianzas.
POLITICAS PUBLICAS ASIA PACIFICO | 09 Abril 2020
Twittear
Recomendar 4
COVID-19


El Big Data hace irrelevante el pensamiento porque si todo es numerable, todo es igual...
Byung-Chul Han

Big Data: the seven V's

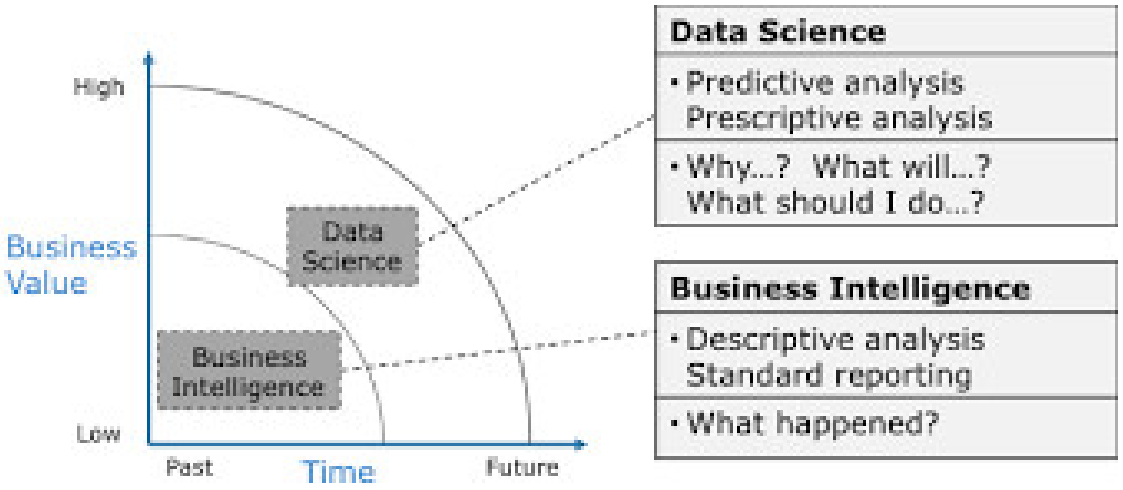
Volume Velocity Variety Variability Veracity Visualization Value

Big Data: How does it mean?

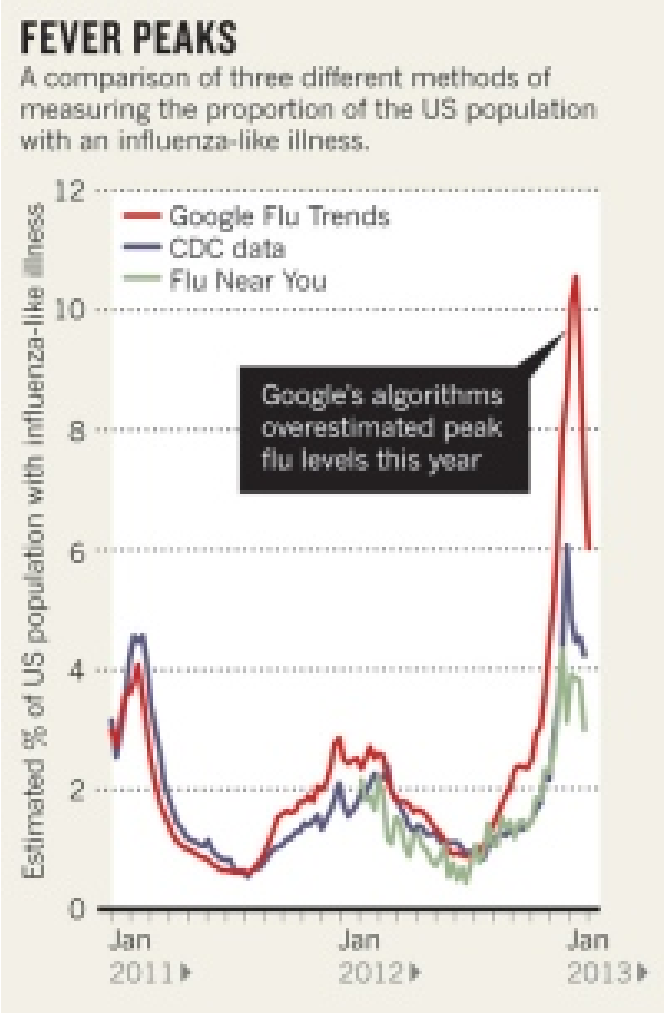
Big Data as stated by Cesar Hidalgo (MIT)

- 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: data science - business intelligence



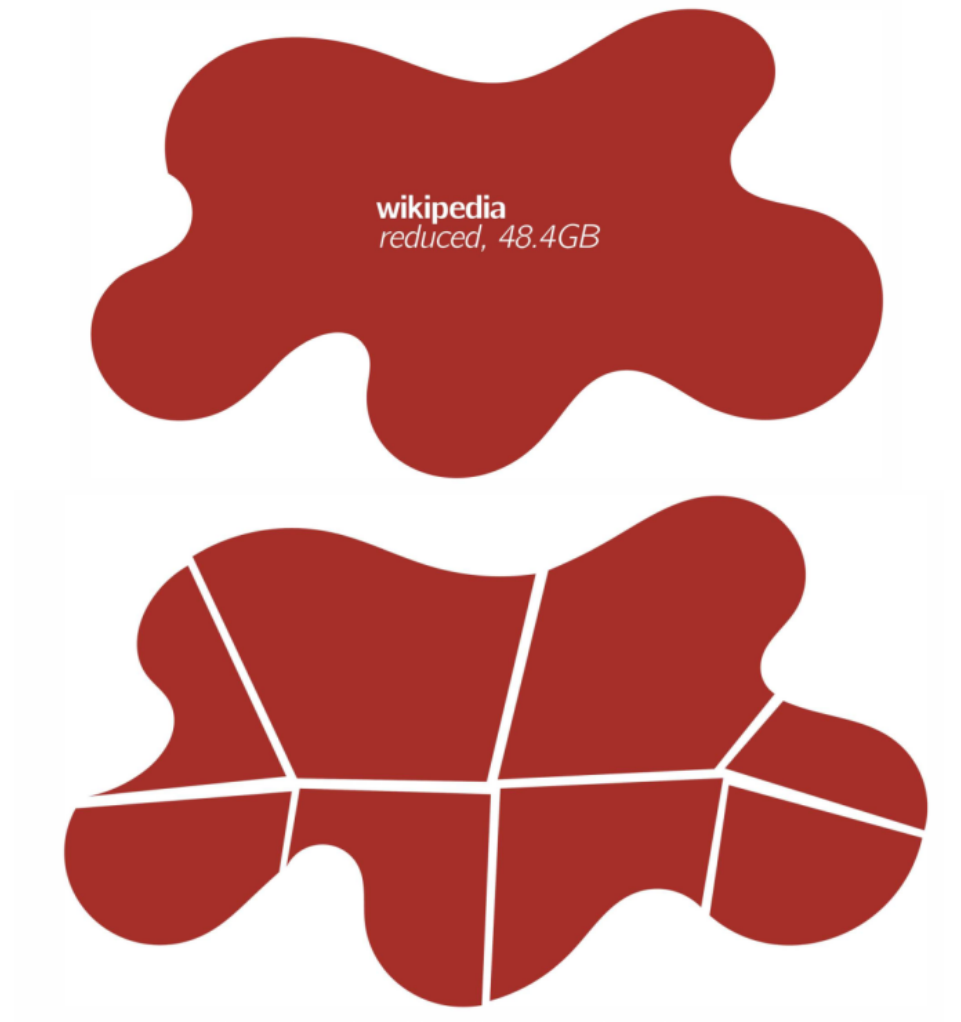
Big Data: Google Flu Trends



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.

Big Data: how to tackle it?

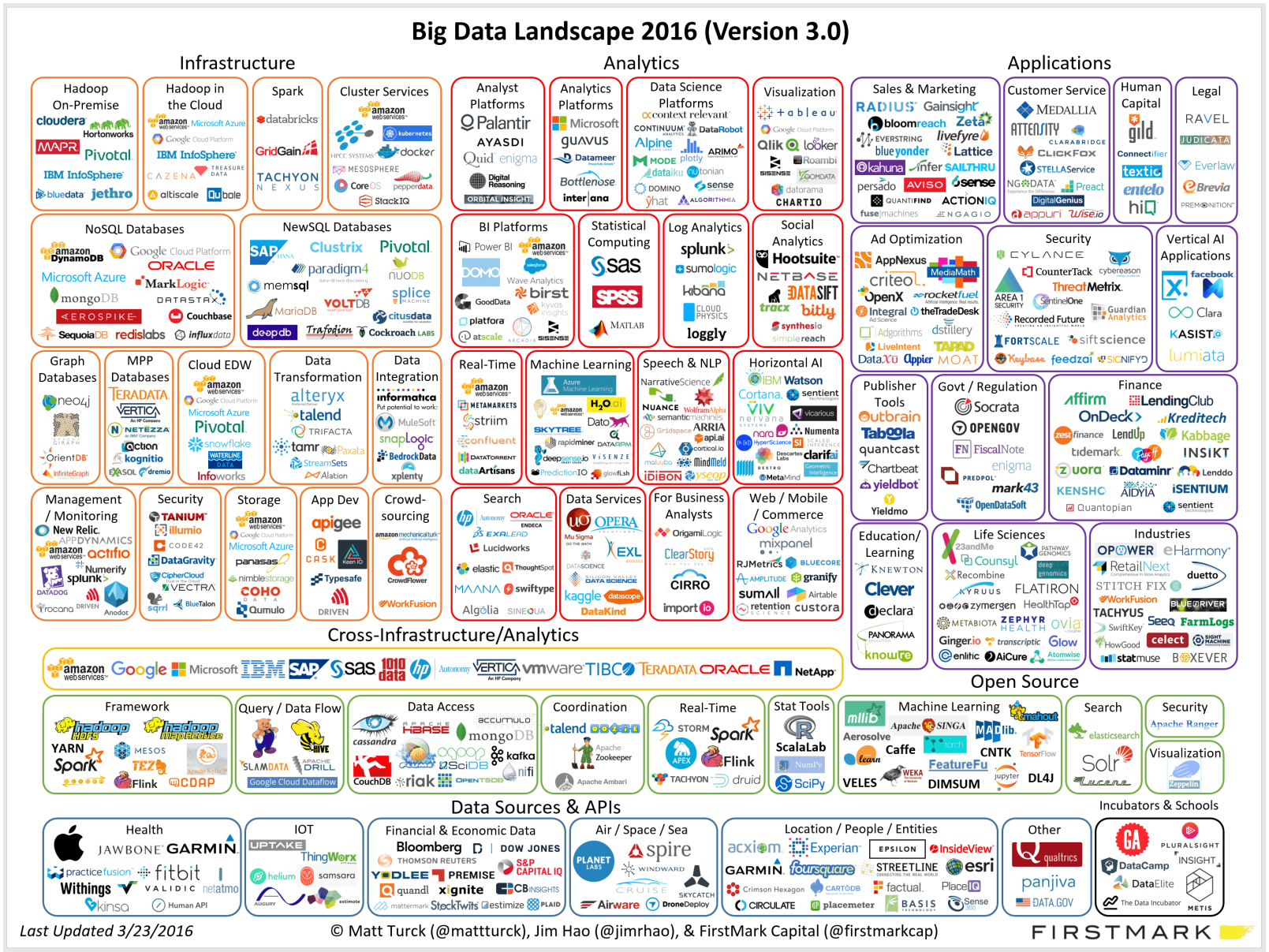


Big Data: how to tackle it?

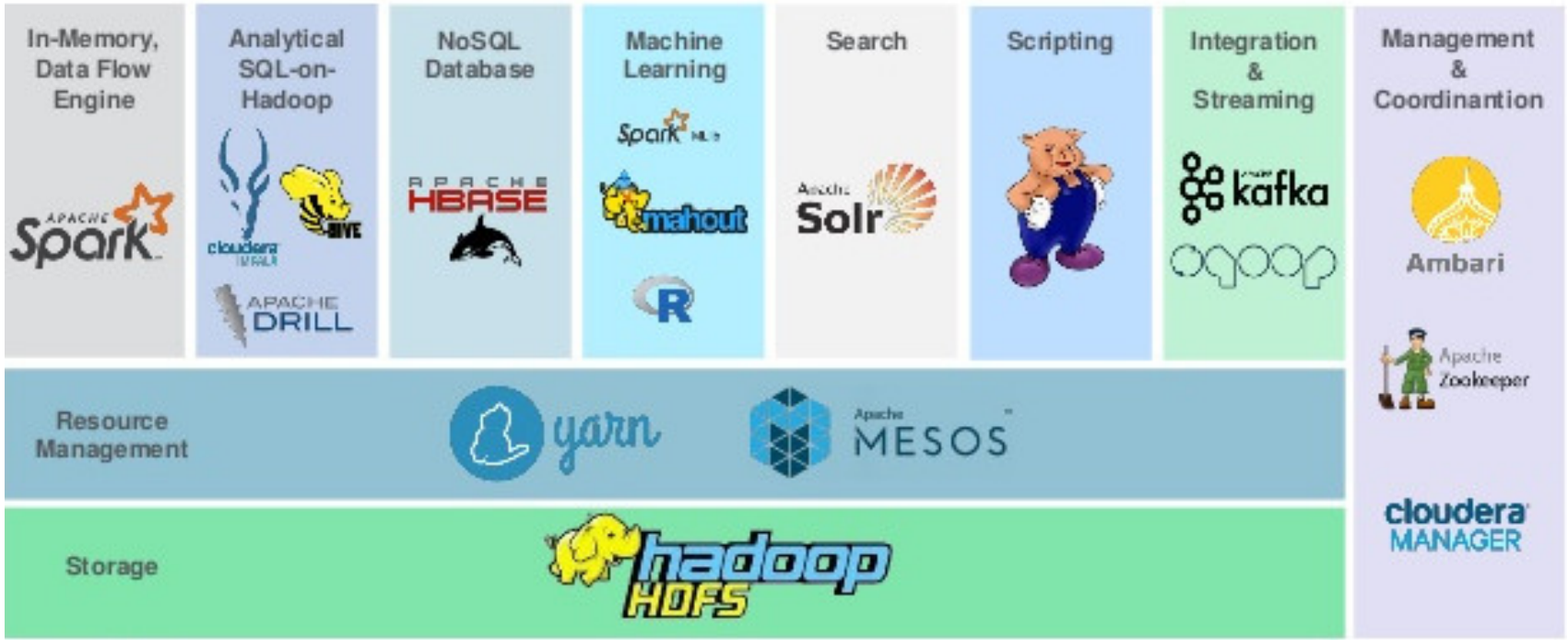
There are many issues that need to be addressed:

- 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: The Landscape



Big Data: the hadoop ecosystem



Big Data: Why worry about foundations?

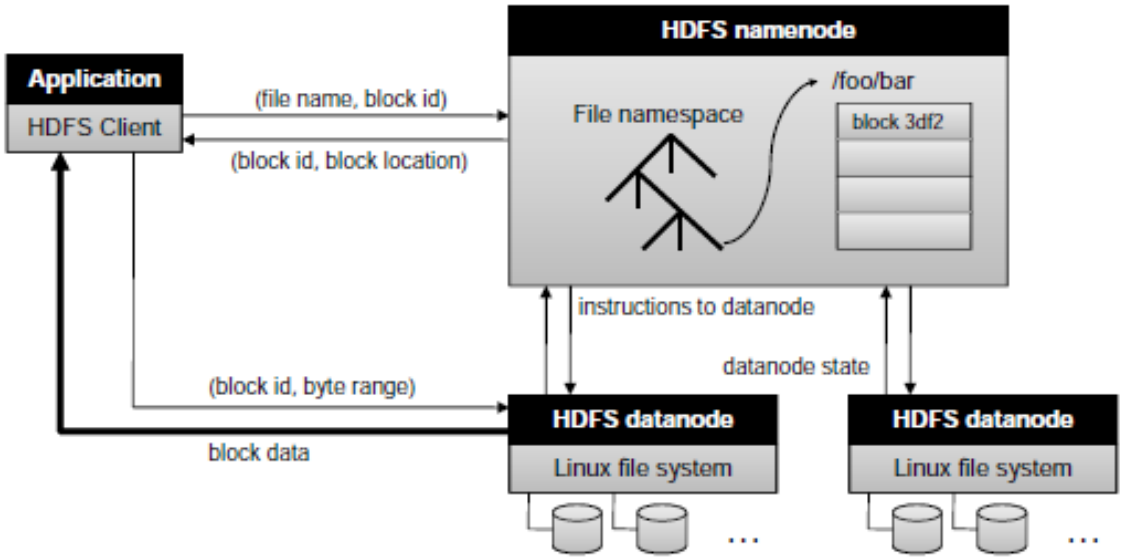
- Chemistry before the tubes.

Big Data: Why worry about foundations?

- Chemistry before the tubes.
- Computer science before big data/data science 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.



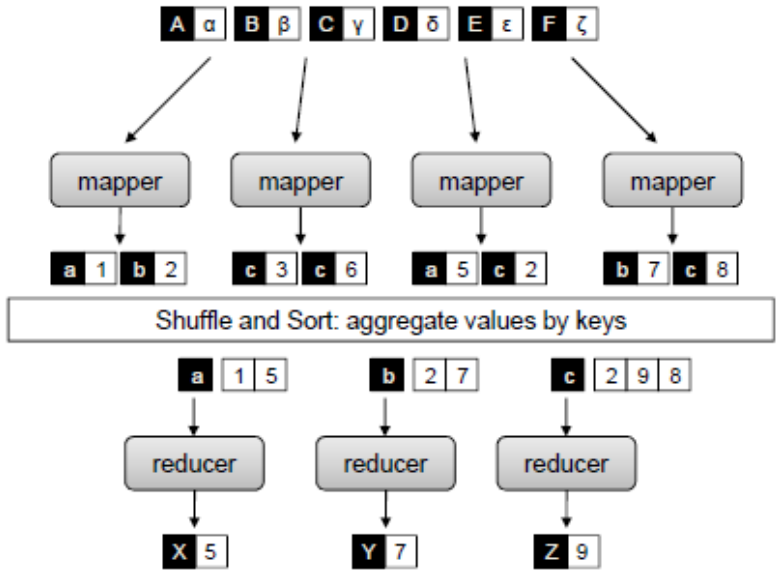
- Data replication makes the system more fault tolerant.
- Data replication provides scalability (w.r.t. data access).
- Data replication and partitioning provide high concurrency.

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

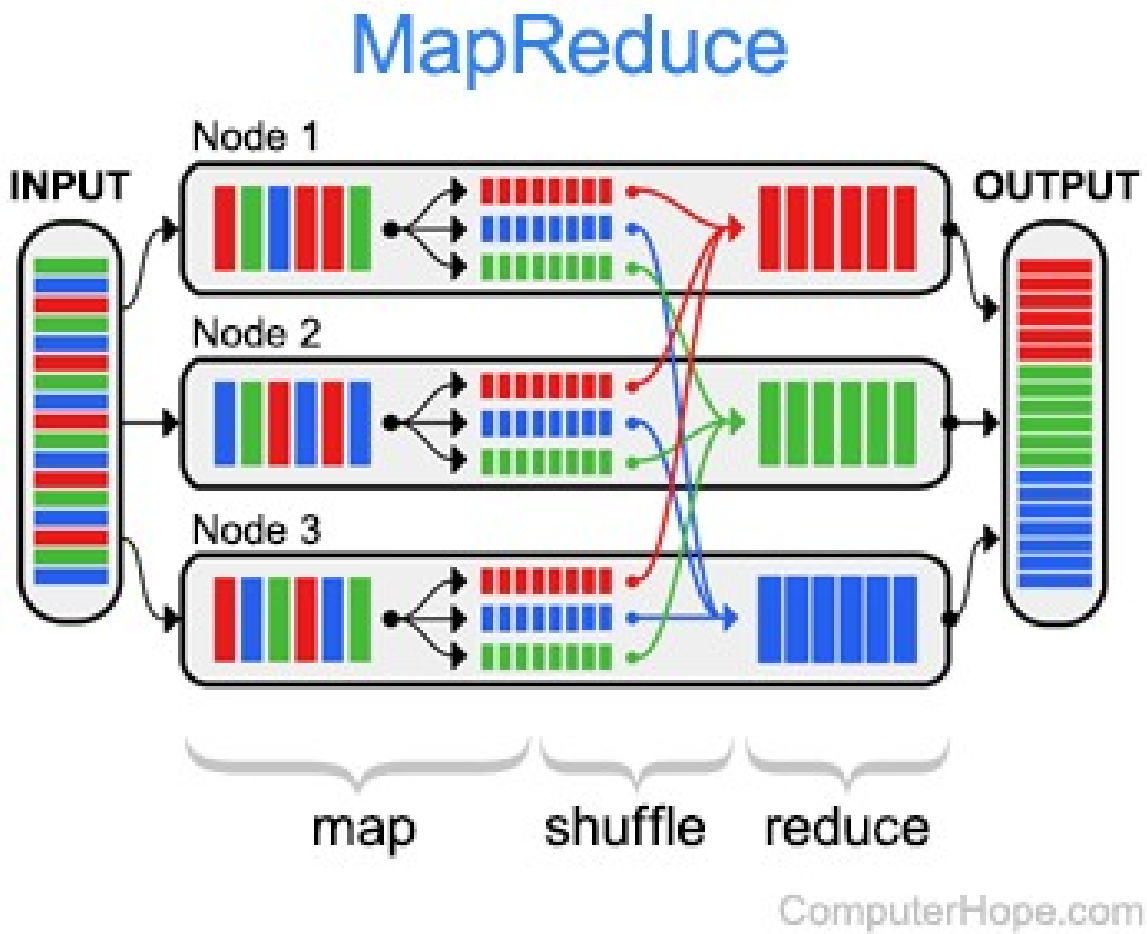
The problem with replication is that is hard to maintain data consistency over the time but in big data systems, the data is written once and the updates are stored as additional data sets over the time.

Big Data: Programming Model

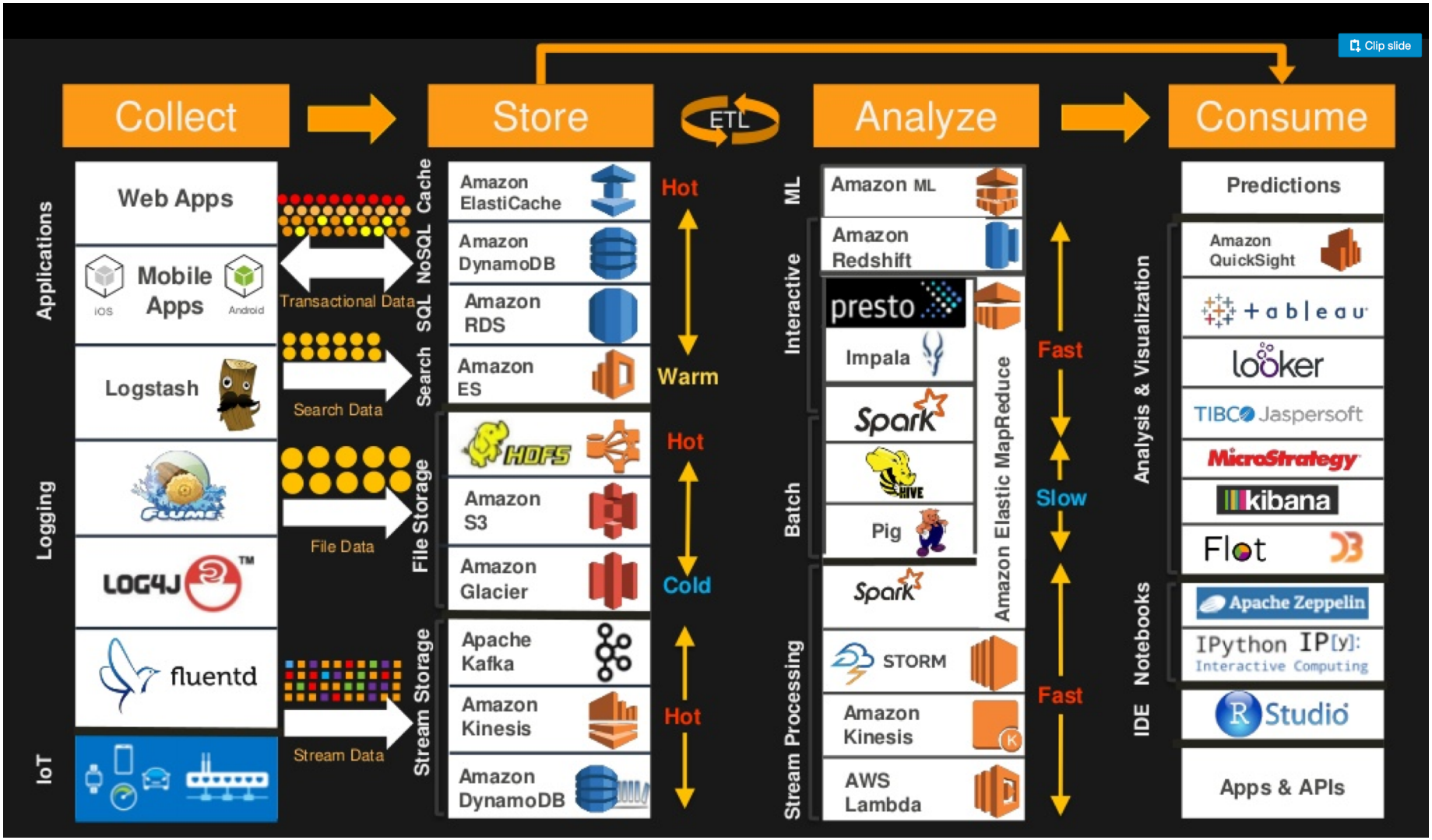
MapReduce is a programming model for data processing introduced by Google (2004) to support parallel and fault-tolerant computations over large data sets on clusters of computers. It provides an abstraction that hides many system-level details from the programmer.




```
map     (k1, v1)          ---> list(k2,v2)
reduce  (k2, list(v2))    ---> list(v2)
```

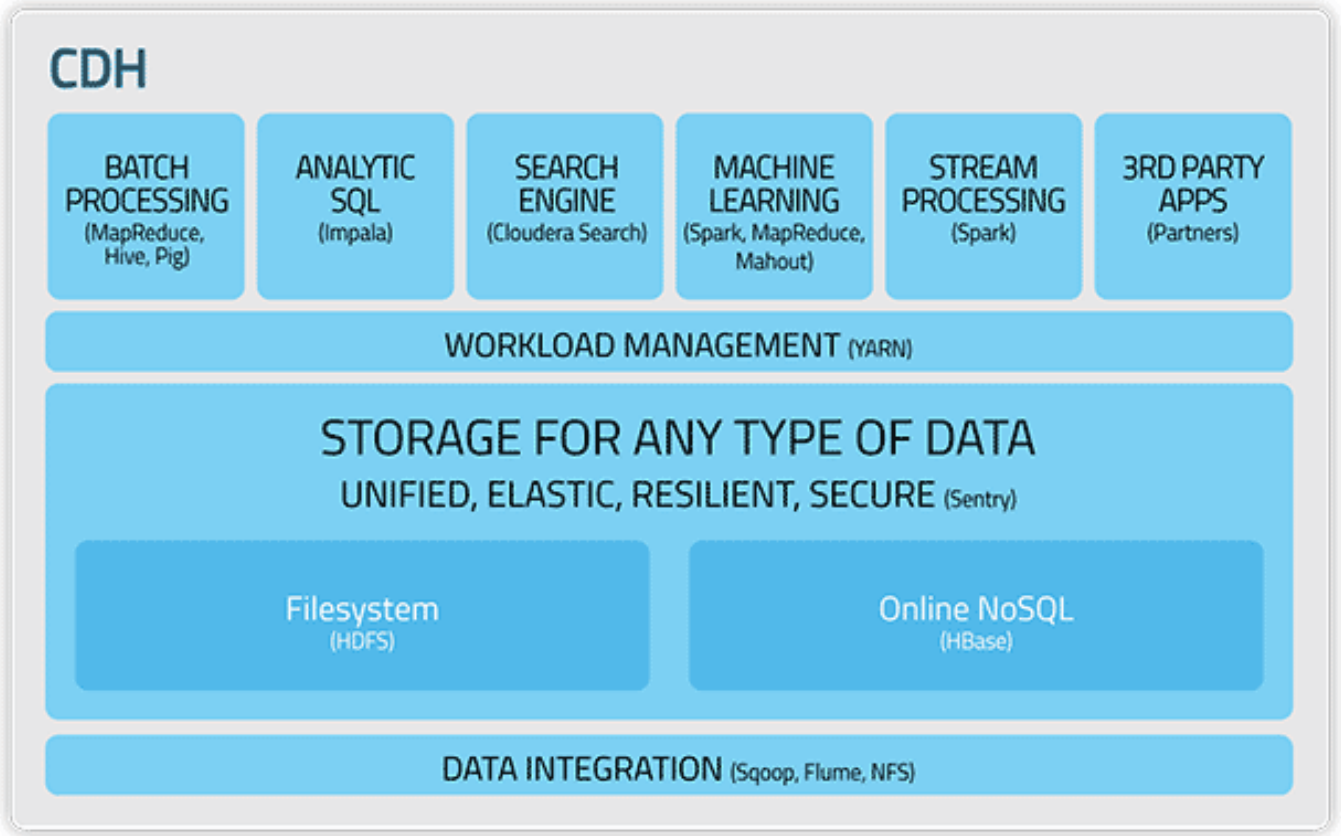

MapReduce
INPUT
Node 1
Node 2
Node 3
OUTPUT
map
shuffle
reduce
ComputerHope.com

Big Data: Cloud Services

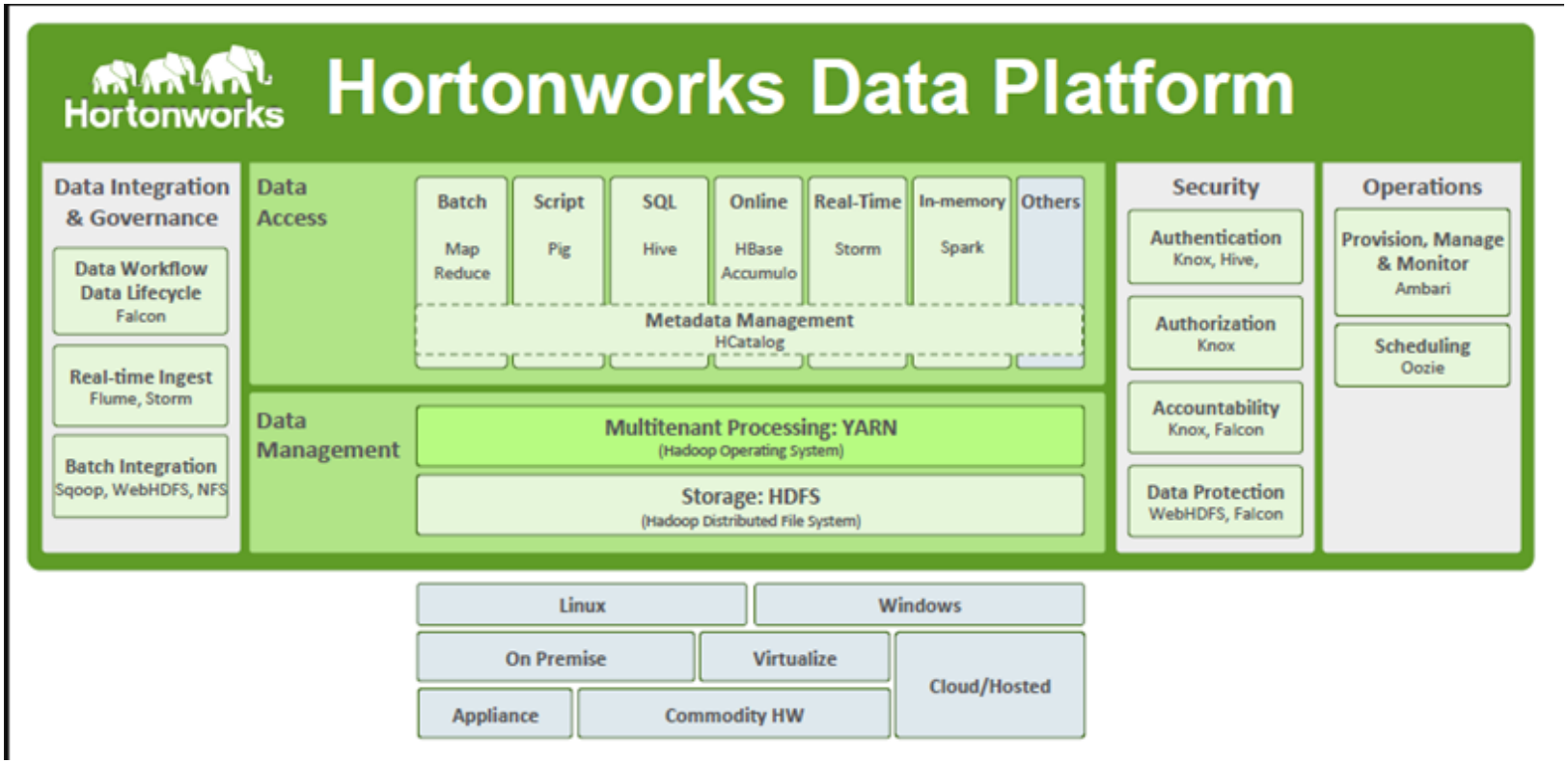


Big Data: Cloud Hosting

Big Data: Hadoop ecosystem providers



Big Data: Hadoop ecosystem providers



Big Data: the unicorn