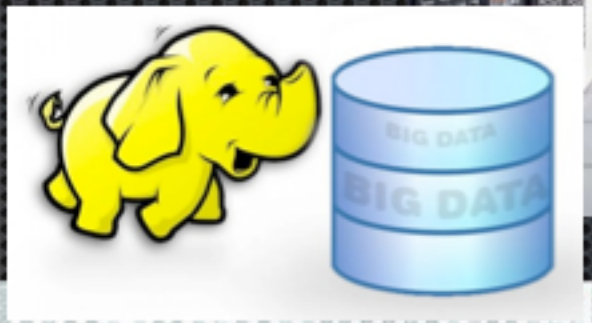
Big Data: Tremendous challenges, great solutions



Luc Bougé
ENS Rennes

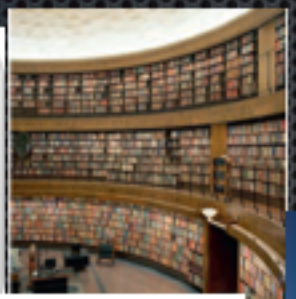
Alexandru Costan
INSA Rennes
Gabriel Antoniu
INRIA Rennes

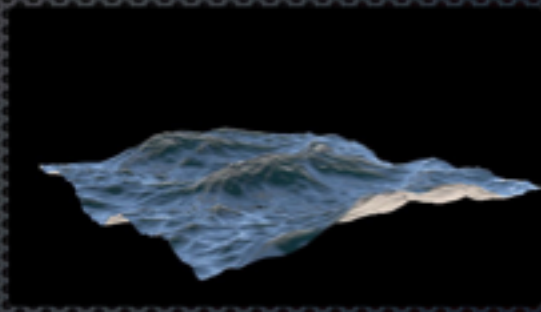
Équipe KerData

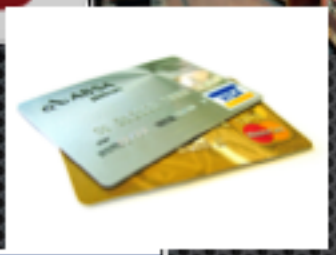
Survive the data deluge!

Big Data?

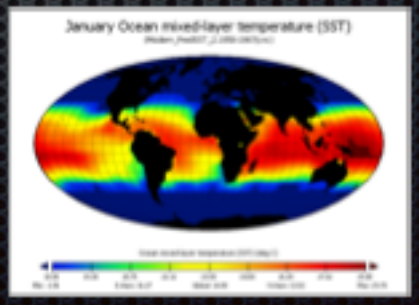




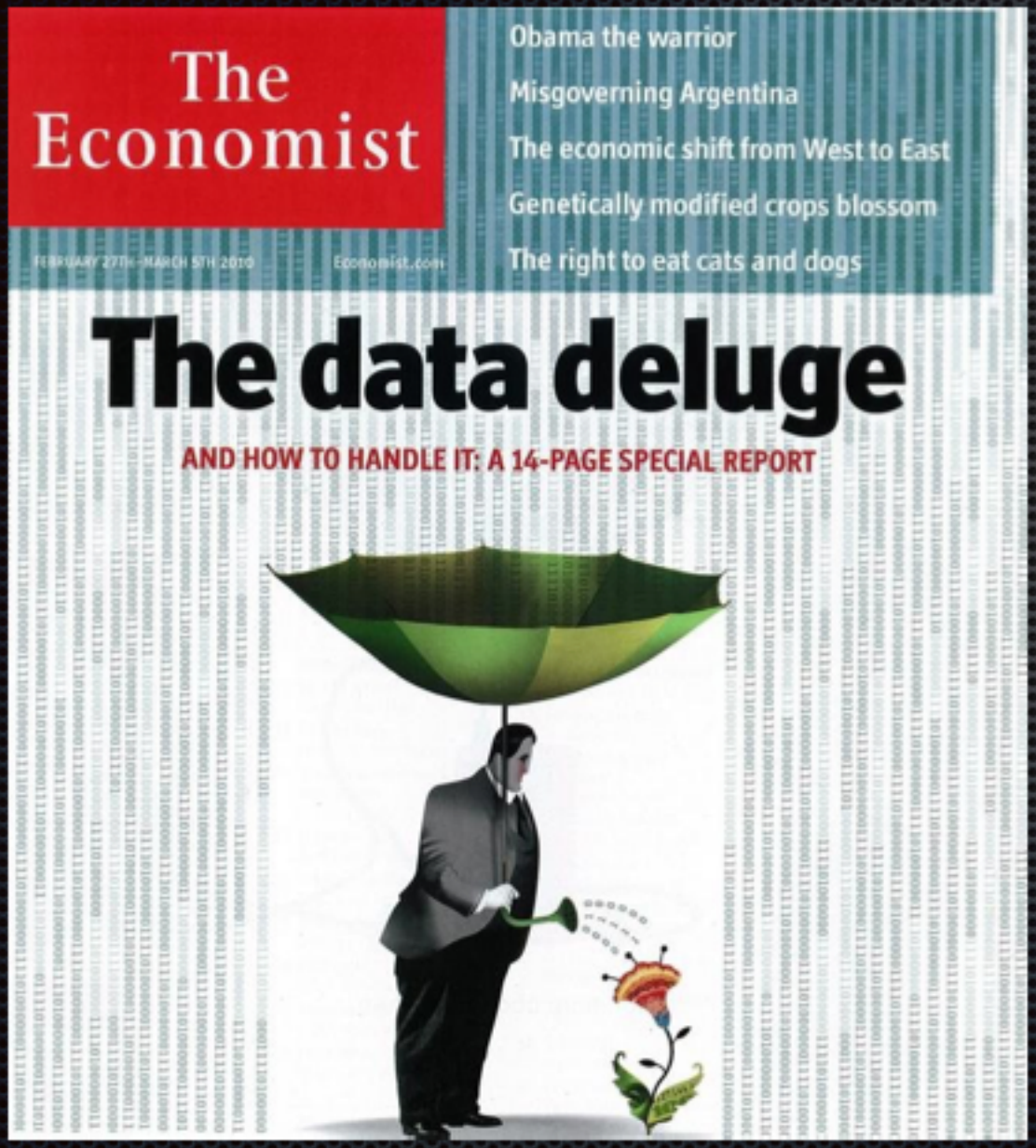






Big Picture



The digital universe:
1.2 ZettaBytes in 2010

35 ZettaBytes in 2020

1 Zetta = 10^{21}

Exponential growth

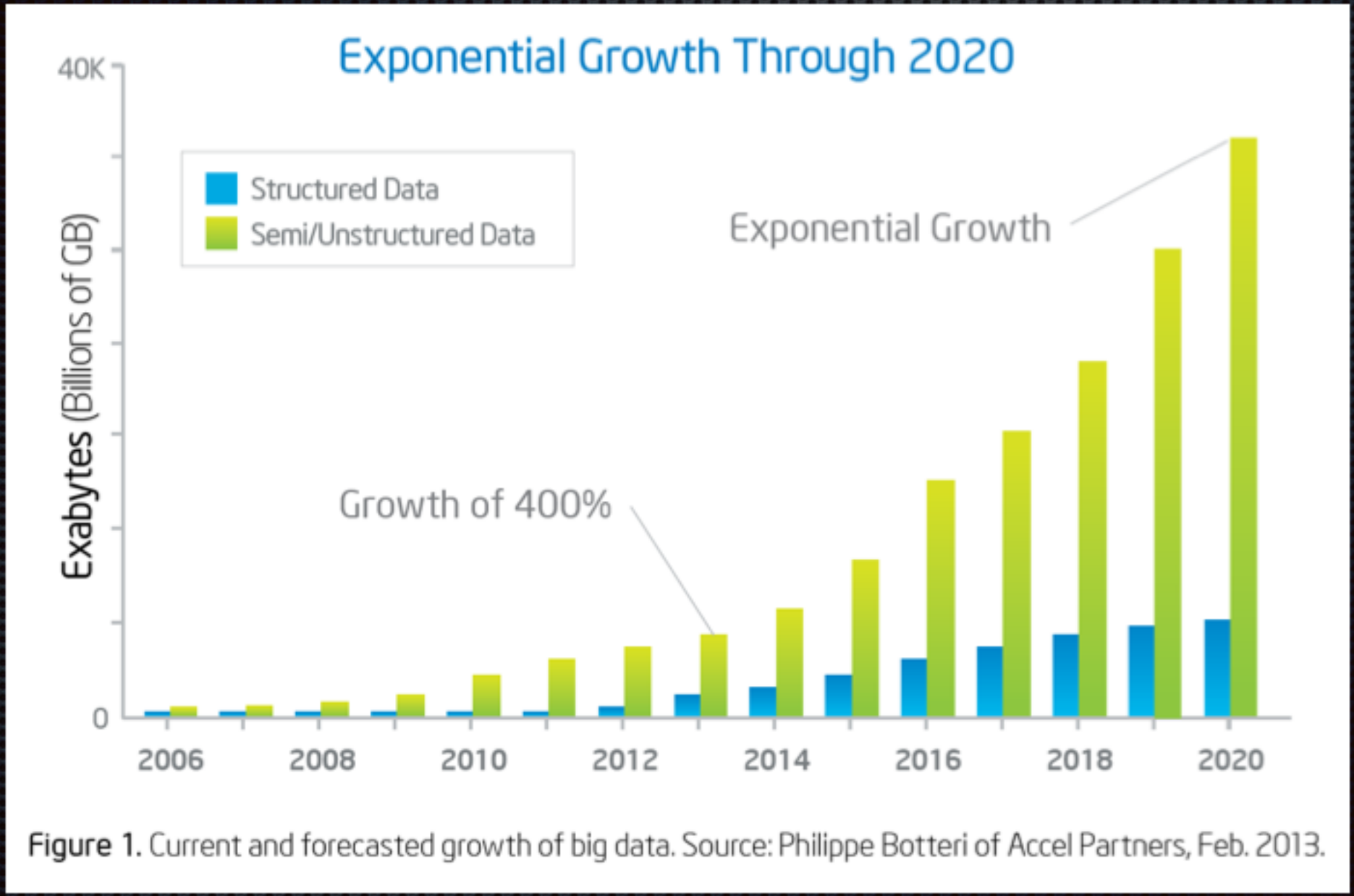
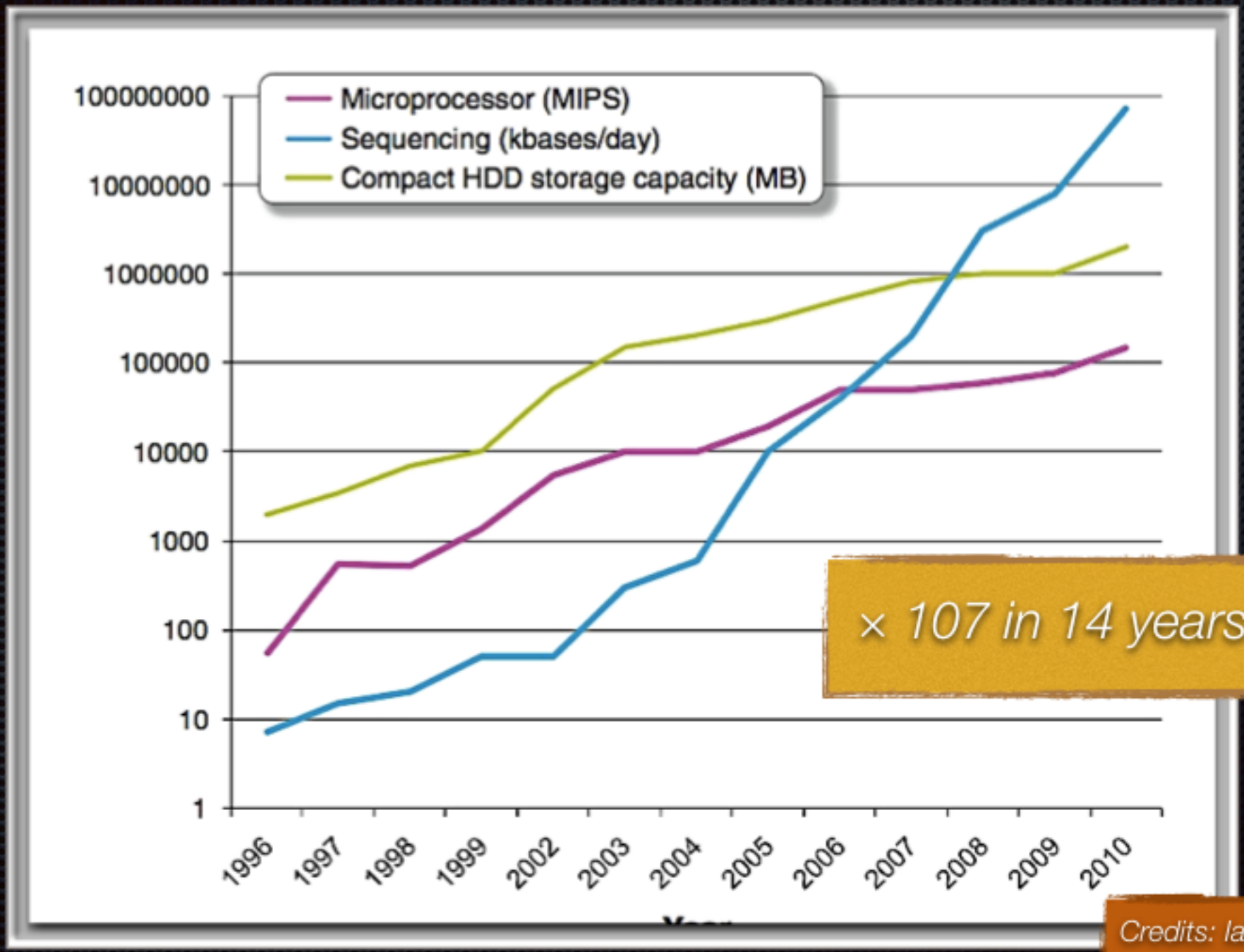


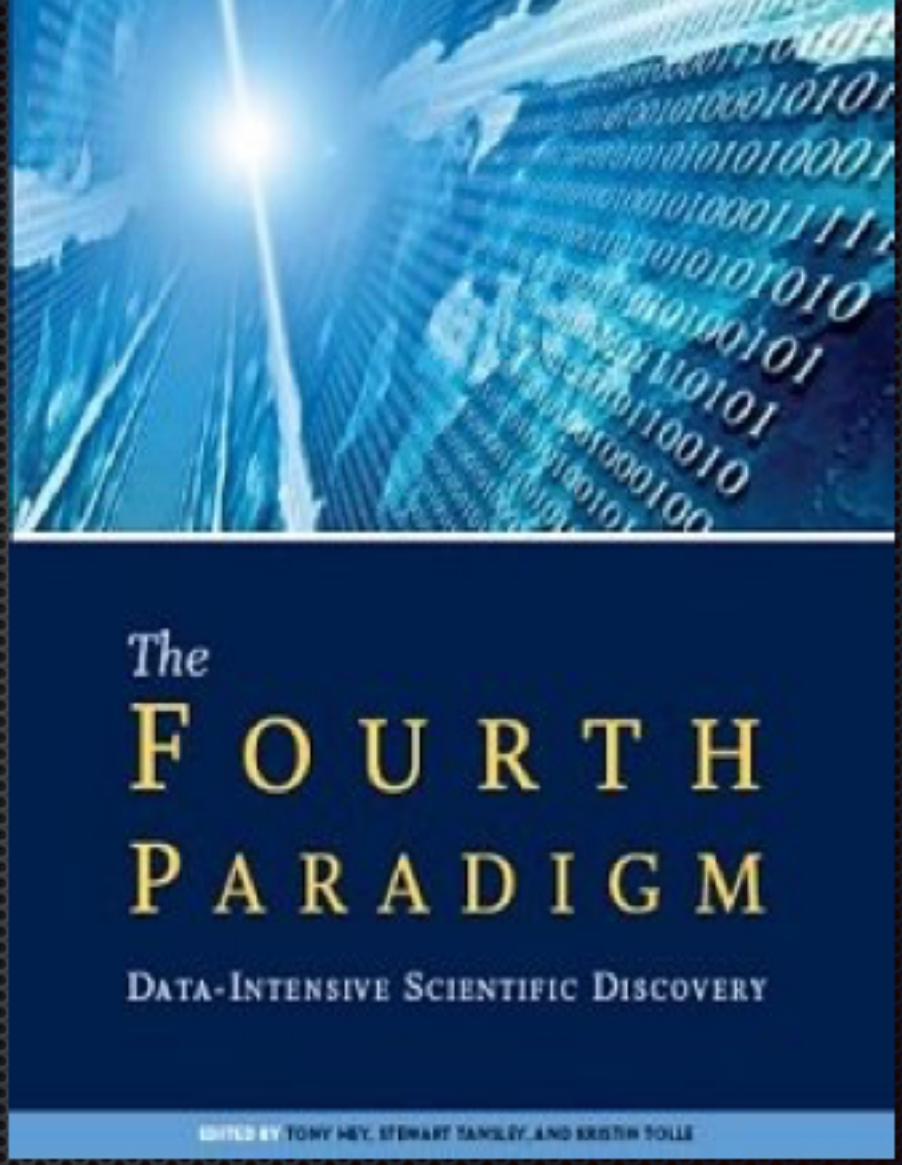
Figure 1. Current and forecasted growth of big data. Source: Philippe Botteri of Accel Partners, Feb. 2013.

Bioinformatics



Credits: Ian Foster

From Big Data to Data Science



Big Data = ?

Big Data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze.

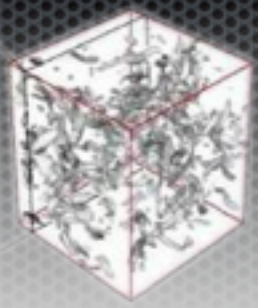
(McKinsey Global Institute)

Big Data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

Wikipedia

Paradigms for scientific discovery

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$



Experimental	Theoretical	Computational	The Fourth Paradigm
Description of natural phenomena	Newton's laws, Maxwell's equations…	Simulation of complex phenomena	?
Thousand years ago	Last few hundred years	Last few decades	Today and the Future

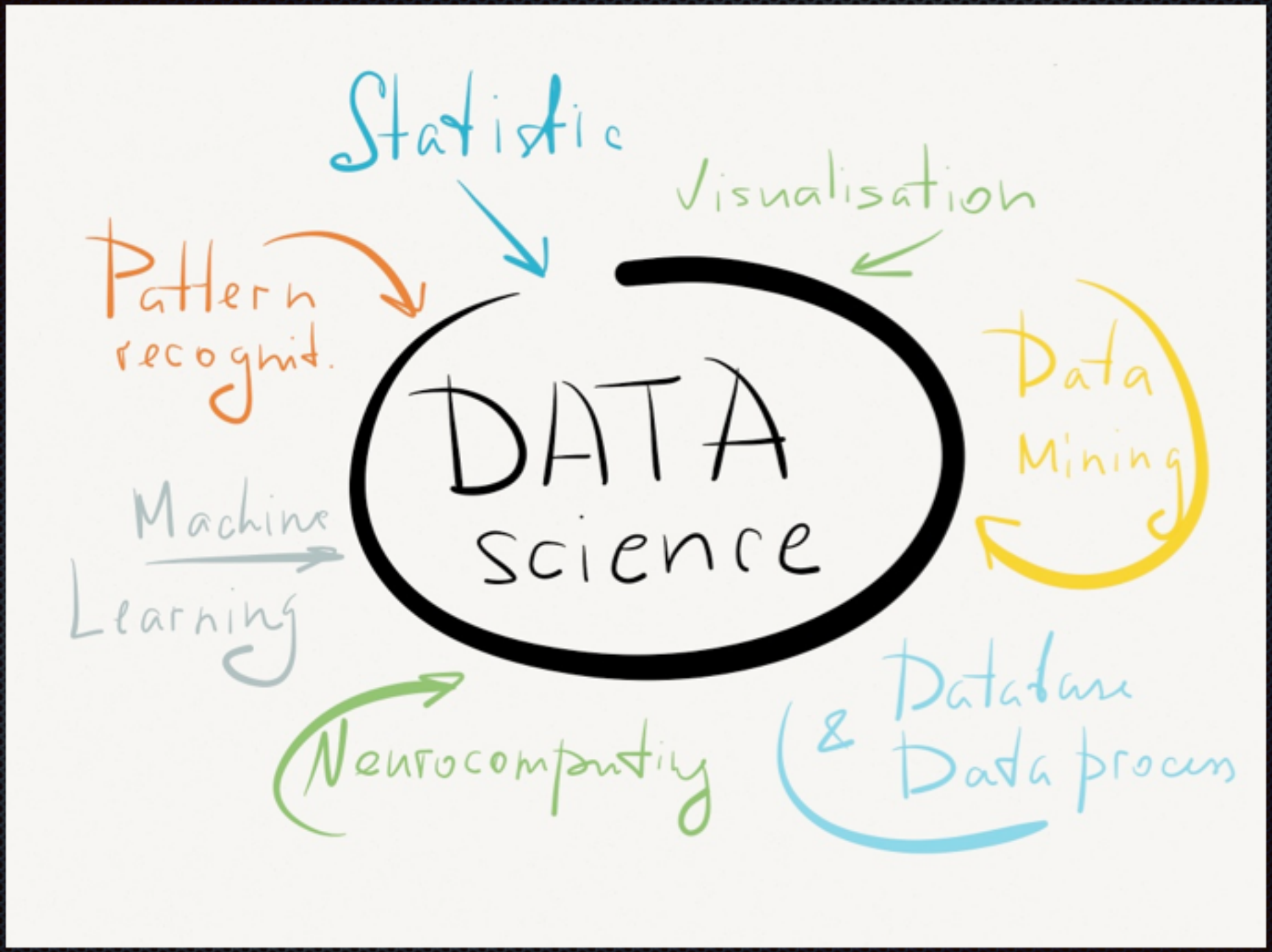
Data Science

Office of Science and Technology Policy
Executive Office of the President
New Executive Office Building
Washington, DC 20502

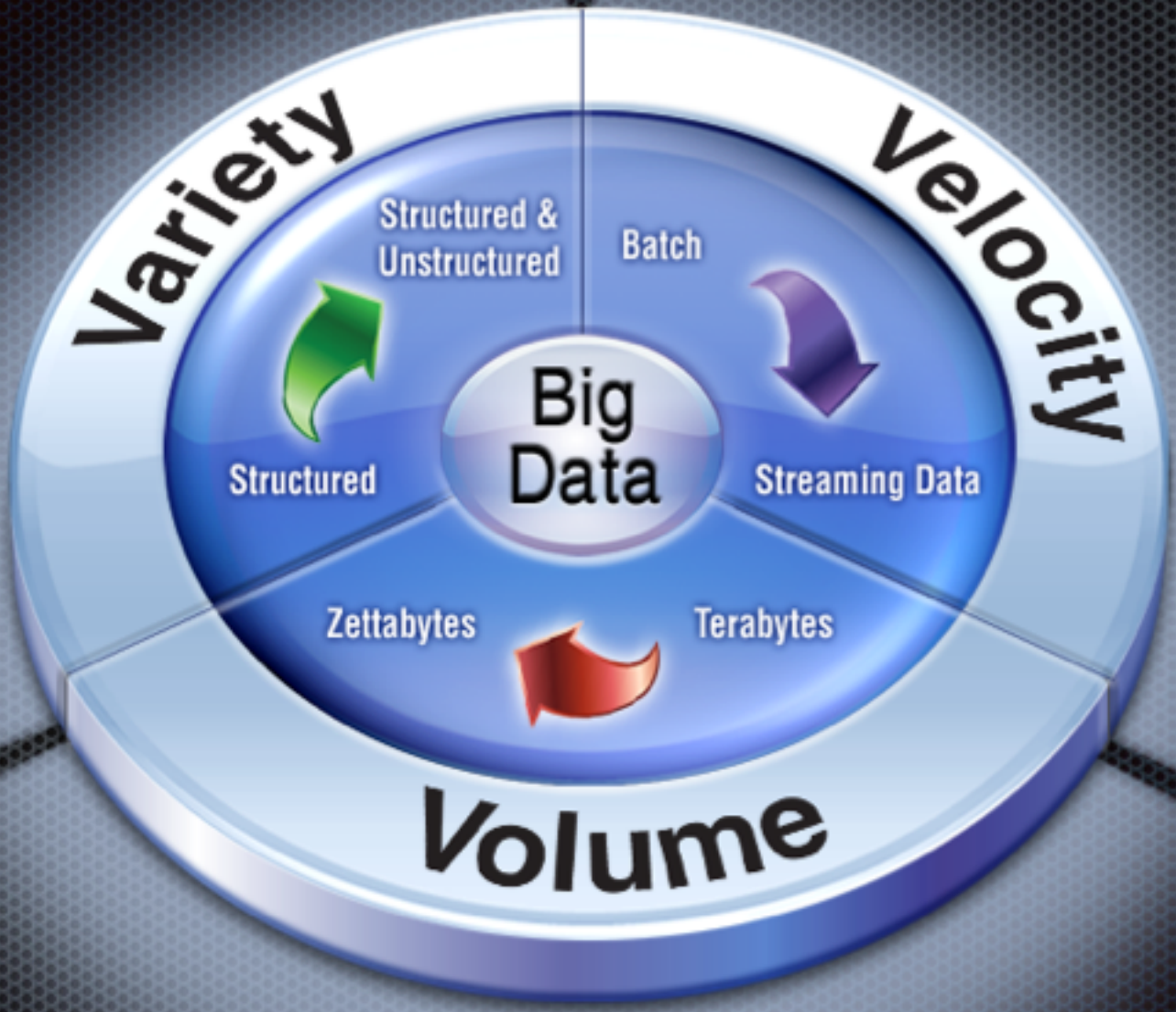
FOR IMMEDIATE RELEASE
March 29, 2012

Contact: Rick Weiss 202 456-6037 rweiss@ostp.eop.gov
Lisa-Joy Zgorski 703 292-8311 lisajoy@nsf.gov

OBAMA ADMINISTRATION UNVEILS "BIG DATA" INITIATIVE: ANNOUNCES $200 MILLION IN NEW R&D INVESTMENTS


Statistic
Visualisation
Pattern recognit.
DATA science
Data Mining
Machine Learning
Neurocomputing
Database & Data procss

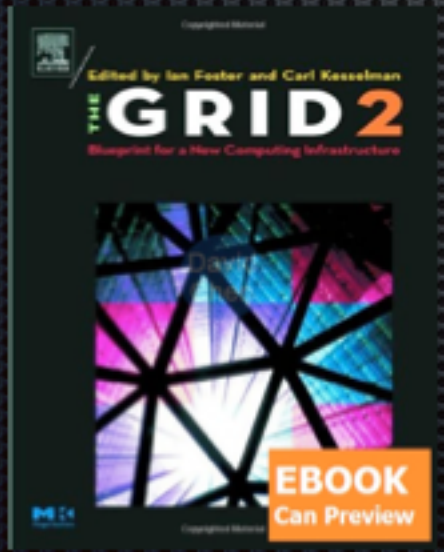
Big Data: The 3 Vs



Clouds for Big Data

The Power-Grid metaphor

The Power and Metaphor



Electricity

Elasticity

Clouds among us!

Clouds among us!

Gmail

Google Drive

Microsoft® Office 365

Key #1: Clusters

Key #1: Clusters

Thomas Sterling (1994)

IRISA (2001)

Key #2: Grids

Grid 5000

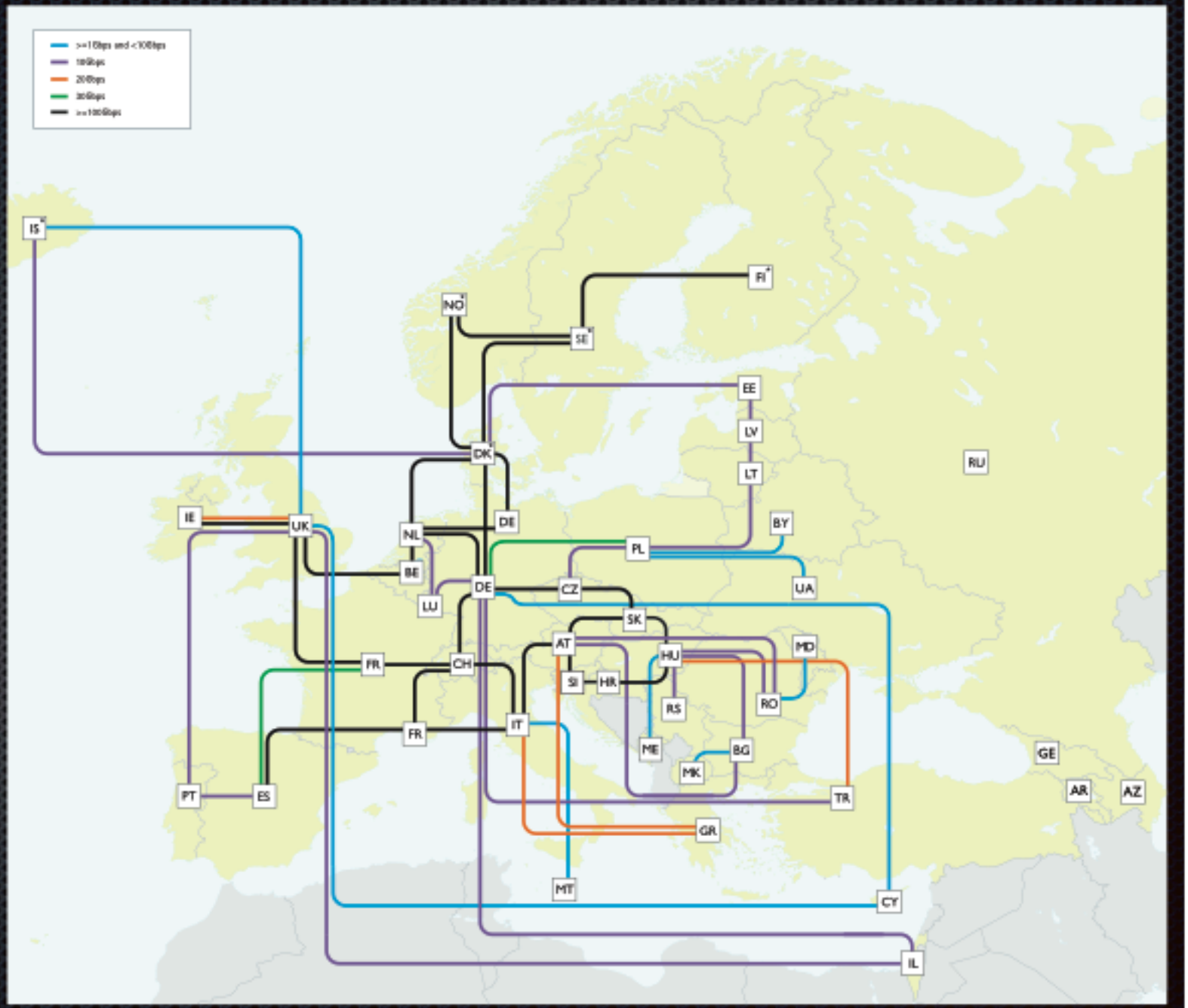
Key #2: Grids

Grid 5000

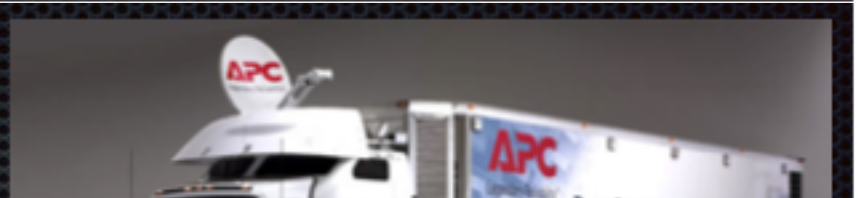
Key #3: Virtual machines

Key #4: High-speed networks

GÉANT



Key #5: Data centers



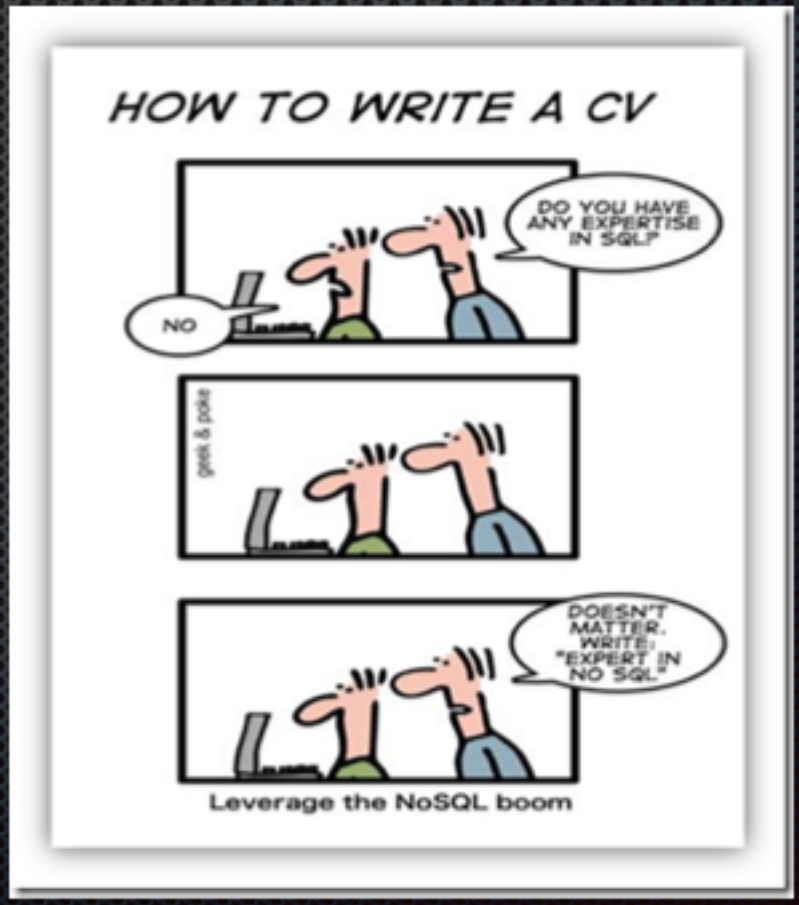
Google data centers today


Plan
Sat.
Rel.
Nunaat
(Greenland)
Island
(Iceland)
Suomi
(Finland)
Sverige
(Sweden)
Norge
(Norway)
Россия
(Russia)
Canada
Polska
(Poland)
Deutschland
(Germany)
Україна
(Ukraine)
Казахстан
(Kazakhstan)
Монгол Улс
(Mongolia)
France
España
(Spain)
Italia
(Italy)
Türkiye
(Turkey)
中国
(China)
대한민국
(South Korea)
日本
(Japan)
North Atlantic Ocean
North Pacific Ocean
العراق
(Iraq)
ایران
(Iran)
افغانستان
(Afghanistan)
پاکستان
(Pakistan)
الجزائر
(Algeria)
ليبيا
(Libya)
مصر
(Egypt)
السعودية
(Saudi Arabia)
India
ประเทศไทย
(Thailand)
México
(Mexico)
Mali
Niger
تشاد
Tchad
(Chad)
السودان
(Sudan)
Nigeria
ኢትዮጵያ
(Ethiopia)
Venezuela
Colombia
Kenya
DR Congo
Tanzania
Indonesia
Papua New Guinea
Brasil
(Brazil)
Perú
(Peru)
Bolivia
Angola
Namibia
Botswana
Madagasikara
(Madagascar)
Indian Ocean
Australia
South Atlantic Ocean
South Africa
Argentina
©2014 Google
Conditions d'utilisation
Google

Storing Data?

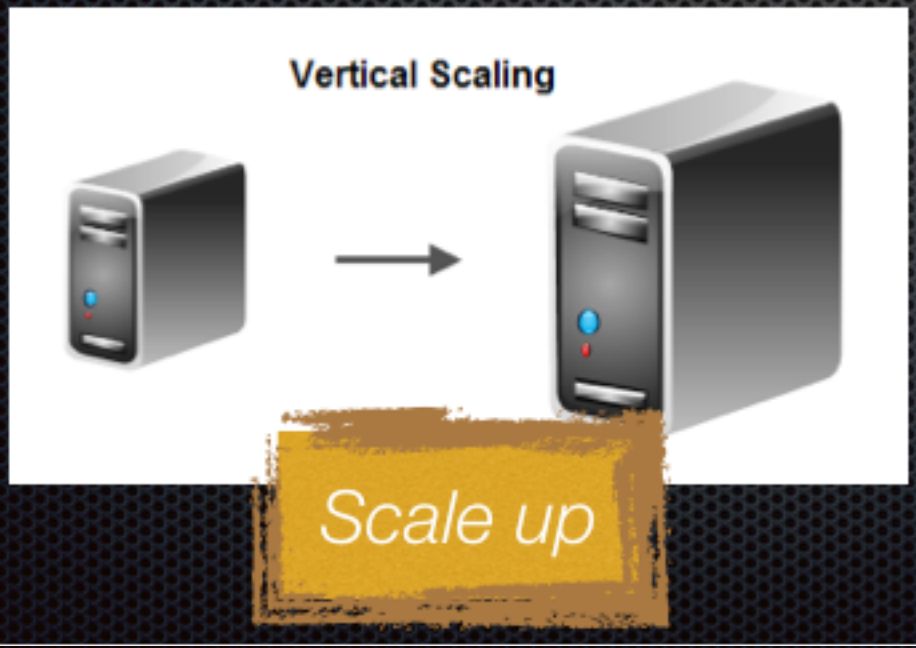
Not SQL or

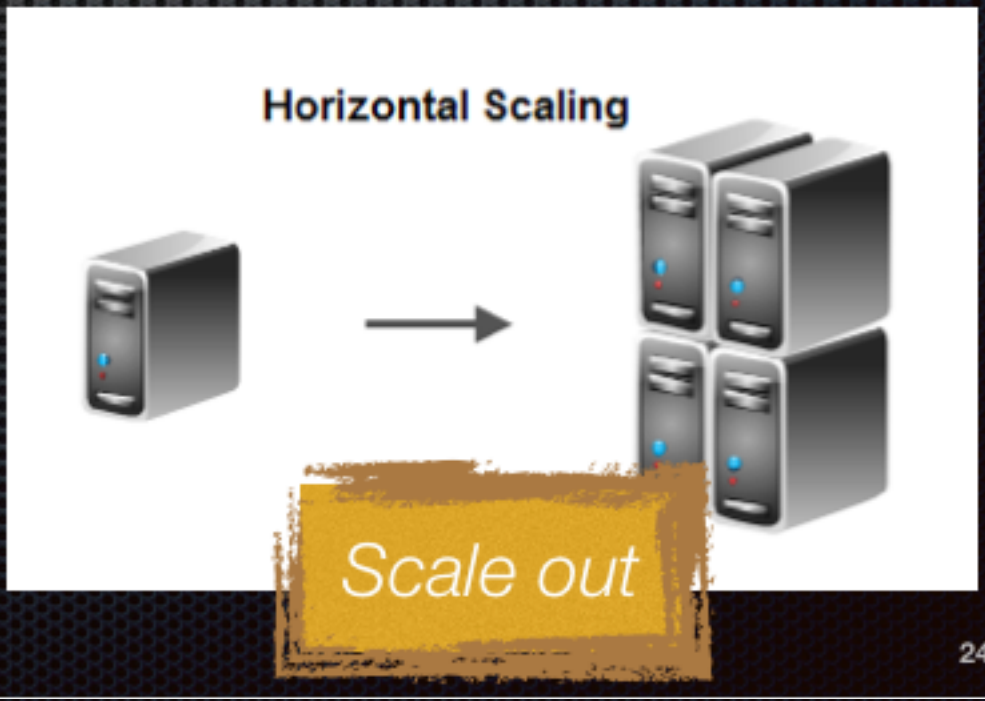
Storing Data: SQL vs. NoSQL



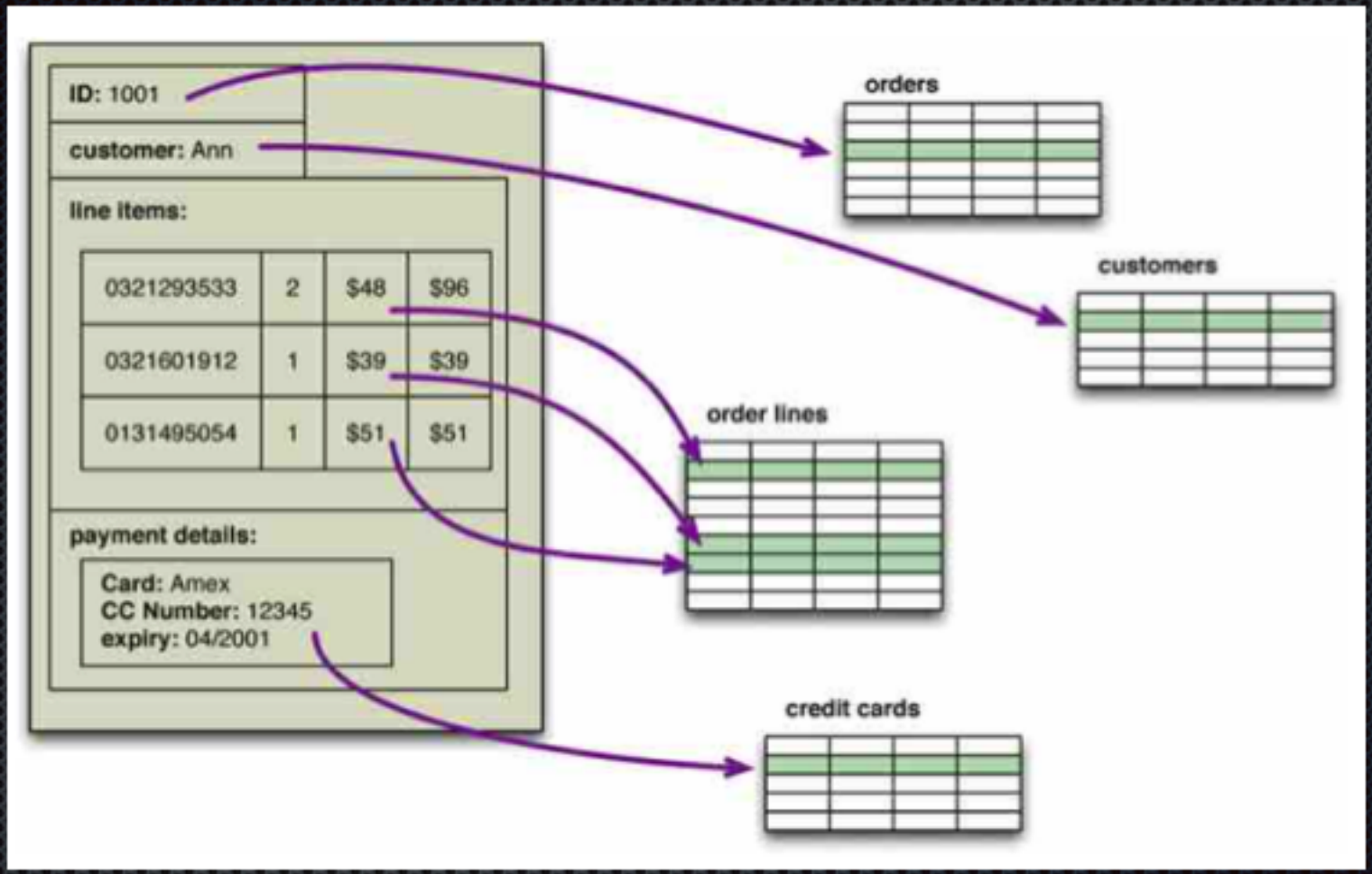
Relational databases

- Not designed to run on multiple nodes (clusters)
- Favor vertical scaling
- Cannot cope with large volumes of data and operations

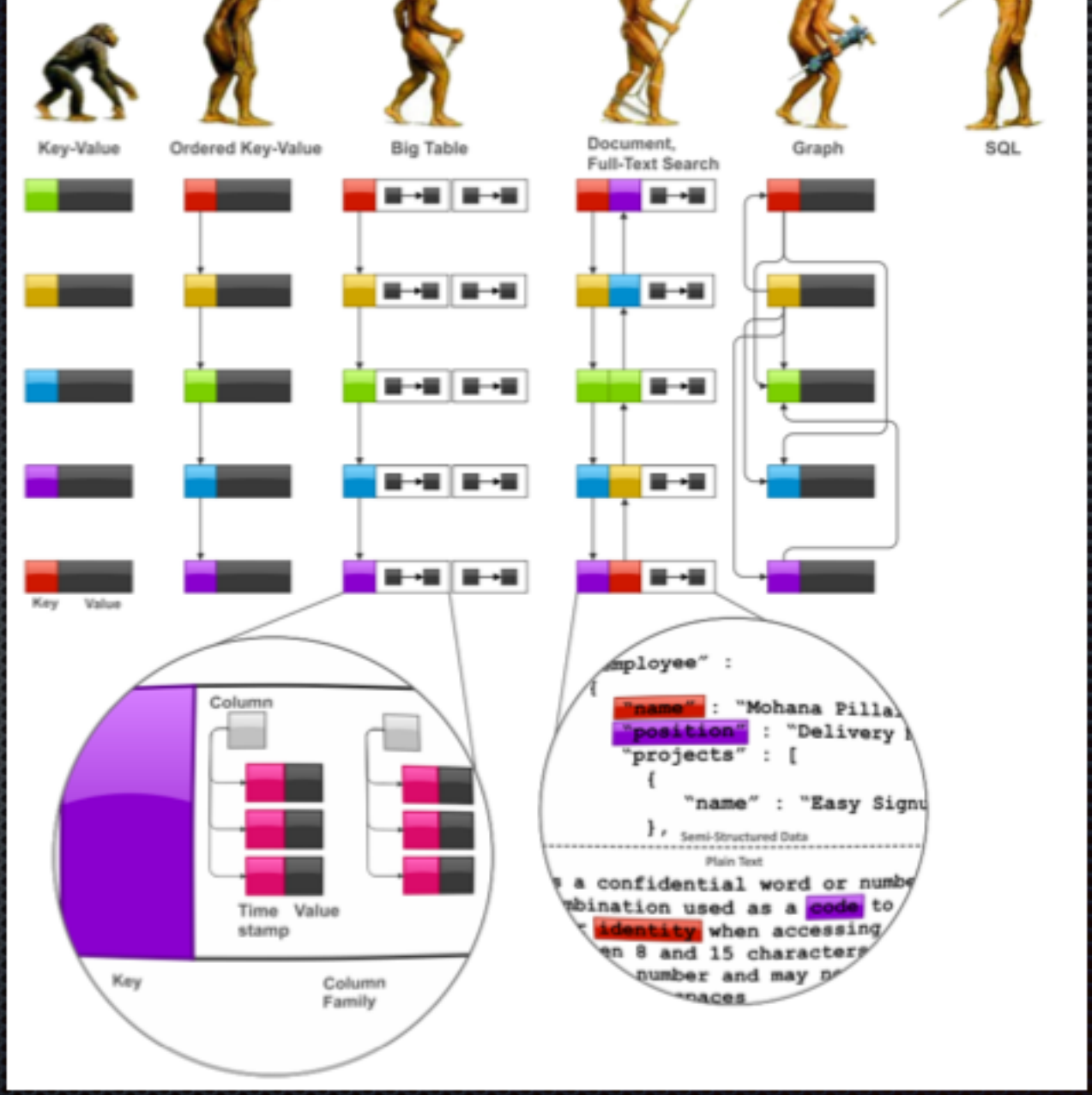


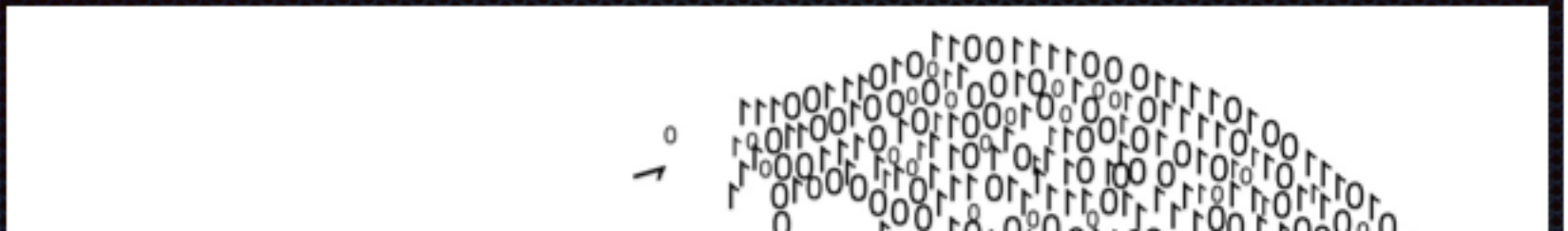


SQL


ID: 1001
customer: Ann
line items:
0321293533 2 $48 $96
0321601912 1 $39 $39
0131495054 1 $51 $51
payment details:
Card: Amex
CC Number: 12345
expiry: 04/2001
orders
customers
order lines
credit cards


STOP
FOLLOWING
ME


Key-Value
Ordered Key-Value
Big Table
Document,
Full-Text Search
Graph
SQL
Key
Value
Column
Time stamp
Value
Key
Column Family
"name" : "Mohana Pilla
"position" : "Delivery
"projects" : [
{
"name" : "Easy Signu
},
Semi-Structured Data
Plain Text
a confidential word or numbe
bination used as a code to
identity when accessing
en 8 and 15 characters
number and may
spaces


0

Map-Reduce for Big Data

Processing the Web

- Introduced by Google in 2004
- Big Data @ Google:
  - 20+ billion web pages x 20KB = 400+ TB
- One computer can read 30-35 MB/sec from disk
  - ~4 months to read the web
  - ~1,000 hard drives just to store the web
- Even more time per HDD, to do something with the data!

Solution: Use many machines!

- Good news: “easy” parallelization
  - Reading the web with 1000 machines ⇒ less than 3 hours
- Bad news: programming challenge
  - Communication and coordination
  - Debugging
  - Fault tolerance
  - Management and monitoring
  - Optimization
- Worse news: redo all the work for every problem you want to solve
- More users, happier users: more data
  - Bigger web, mailbox, blog, etc.
  - Find the right information, and find it faster!

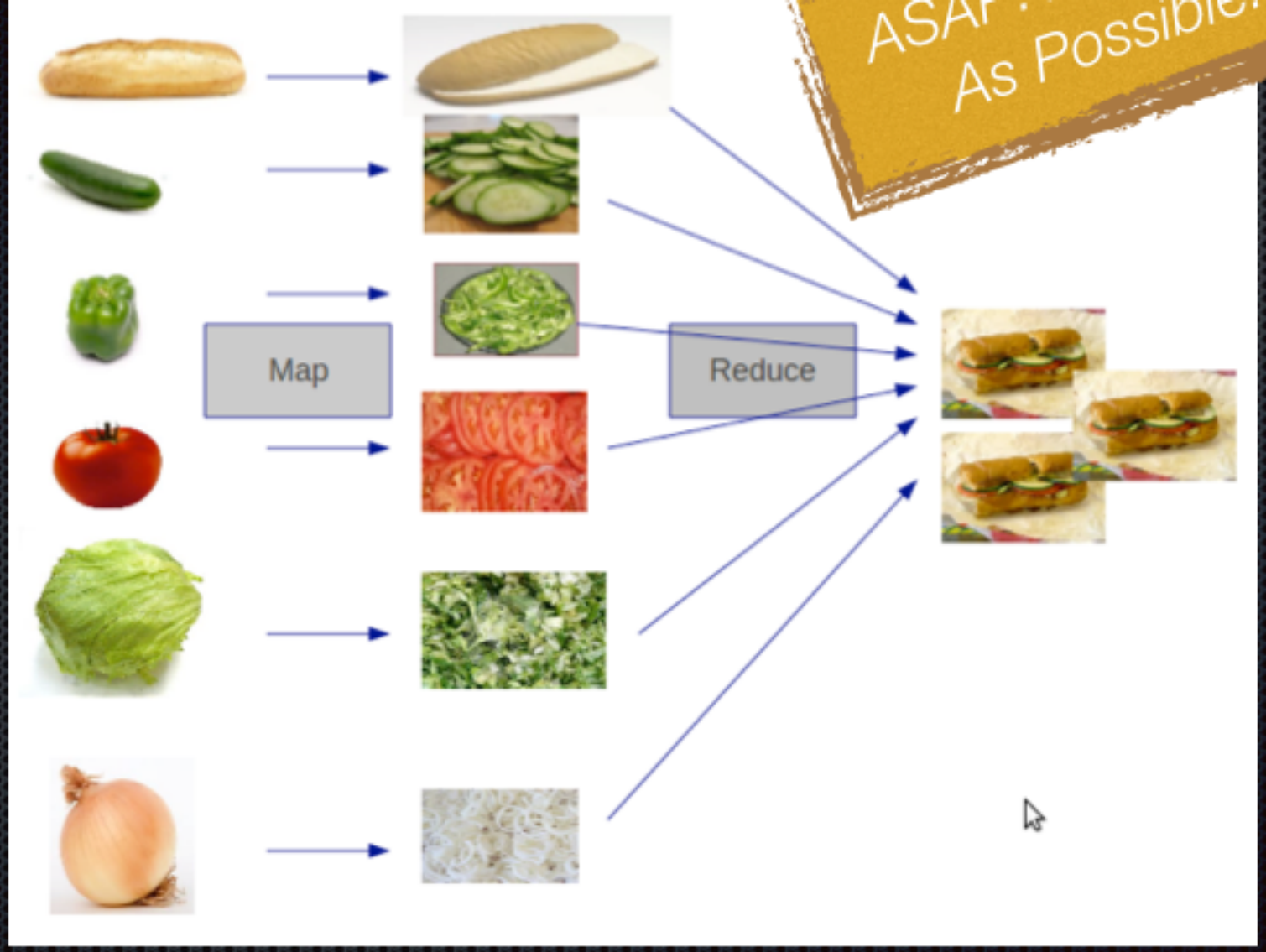
Programming Model

- Scalable
- Simple
- Resilient
- Monitorable
- Versatile
- Abstract
  - Automatic parallelization
  - Load balancing
  - Network and disk transfer optimization
  - Handling of machine failures
  - Robustness
  - Improvements to core library benefit all users of library!

ASAP: As Possible!

Map-Reduce

ASAP: As Simple



Friday, November 21, 2008

Google | Official Blog

Insights from Googlers into our products, technology, and the Google culture

Sorting 1PB with MapReduce

Posted: Friday, November 21, 2008

+1 59

At Google we are fanatical about organizing the world's information. As a result, we spend a lot of time finding better ways to sort information using MapReduce, a key component of our software infrastructure that allows us to run multiple processes simultaneously. MapReduce is a perfect solution for many of the computations we run daily, due in large part to its simplicity, applicability to a wide range of real-world computing tasks, and natural translation to highly scalable distributed implementations that harness the power of thousands of computers.

In our sorting experiments we have followed the rules of a standard terabyte (TB) sort benchmark. Standardized experiments help us understand and compare the benefits of various technologies and also add a competitive spirit. You can think of it as an Olympic event for computations. By pushing the boundaries of these types of programs, we learn about the limitations of current technologies as well as the lessons useful in designing next generation computing platforms. This, in turn, should help everyone have faster access to higher-quality information.

We are excited to announce we were able to sort 1TB (stored on the Google File System as 10 billion 100-byte records in uncompressed text files) on 1,000 computers in 68 seconds. By comparison, the previous 1TB sorting record is 209 seconds on 910 computers.

Sometimes you need to sort more than a terabyte, so we were curious to find out what happens when you sort more and gave one petabyte (PB) a try. One petabyte is a thousand terabytes, or, to put this amount in perspective, it is 12 times the amount of archived web data in the U.S. Library of Congress as of May 2008. In comparison, consider that the aggregate size of data processed by all instances of MapReduce at Google was on average 20PB per day in January 2008.

It took six hours and two minutes to sort 1PB (10 trillion 100-byte records) on 4,000 computers. We're not aware of

Google

google.com/+google

News and updates on Google's products, technology and more

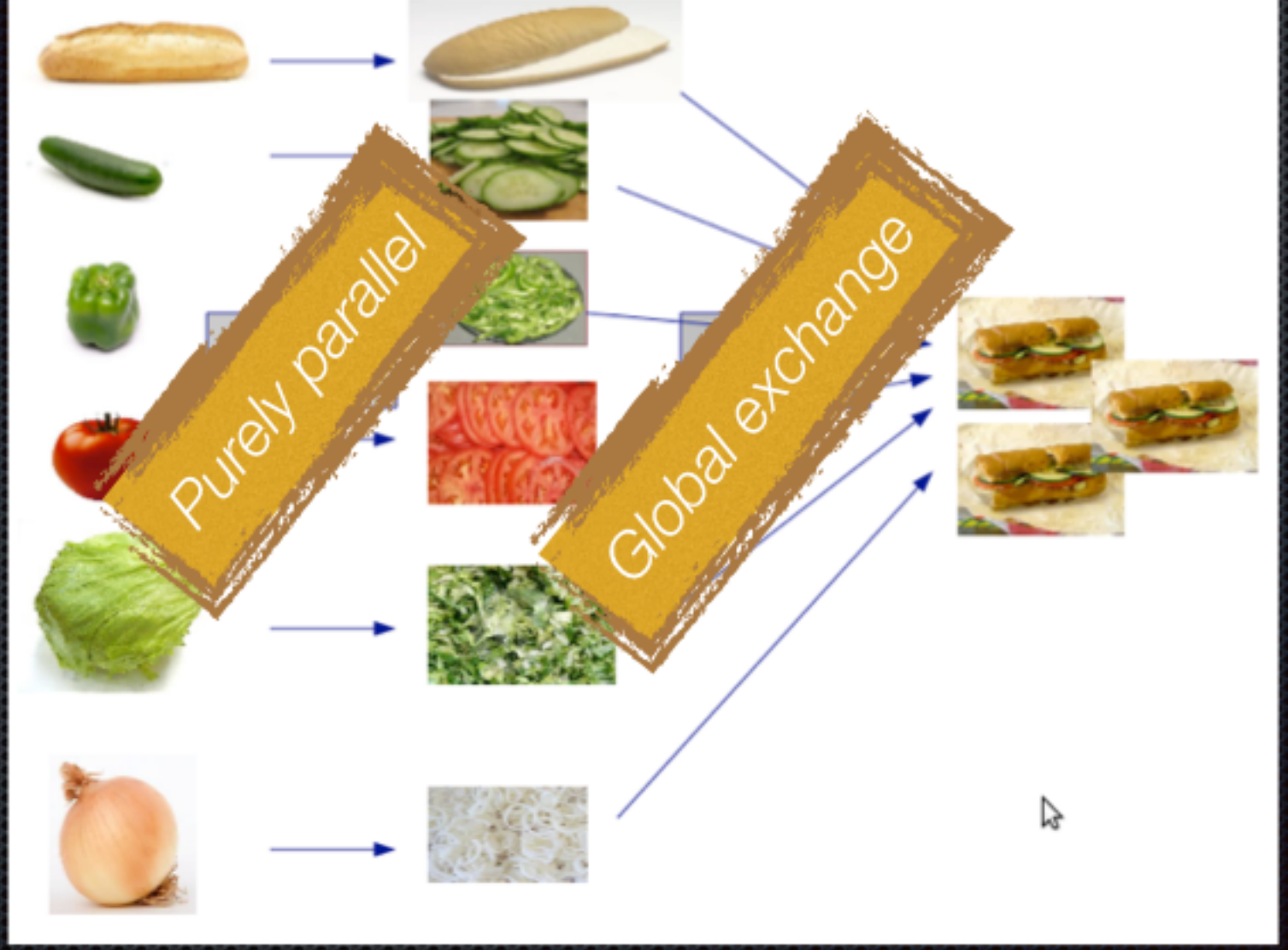
Suivre +1

+ 12 650 192

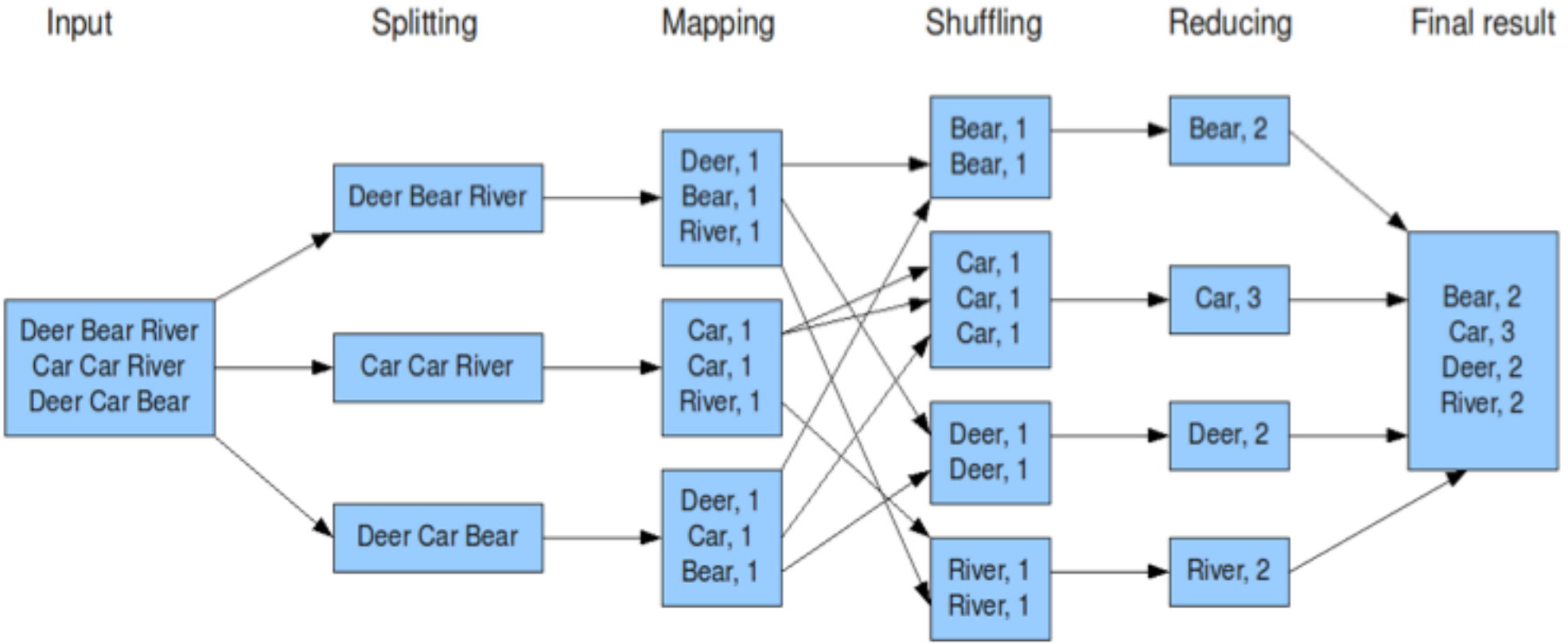
Labels

Archive

Feed



Word Count

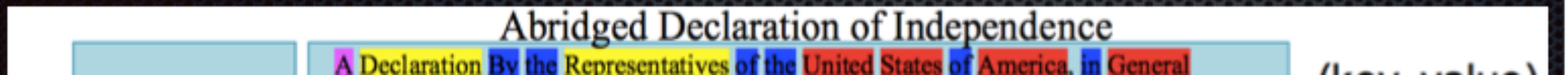


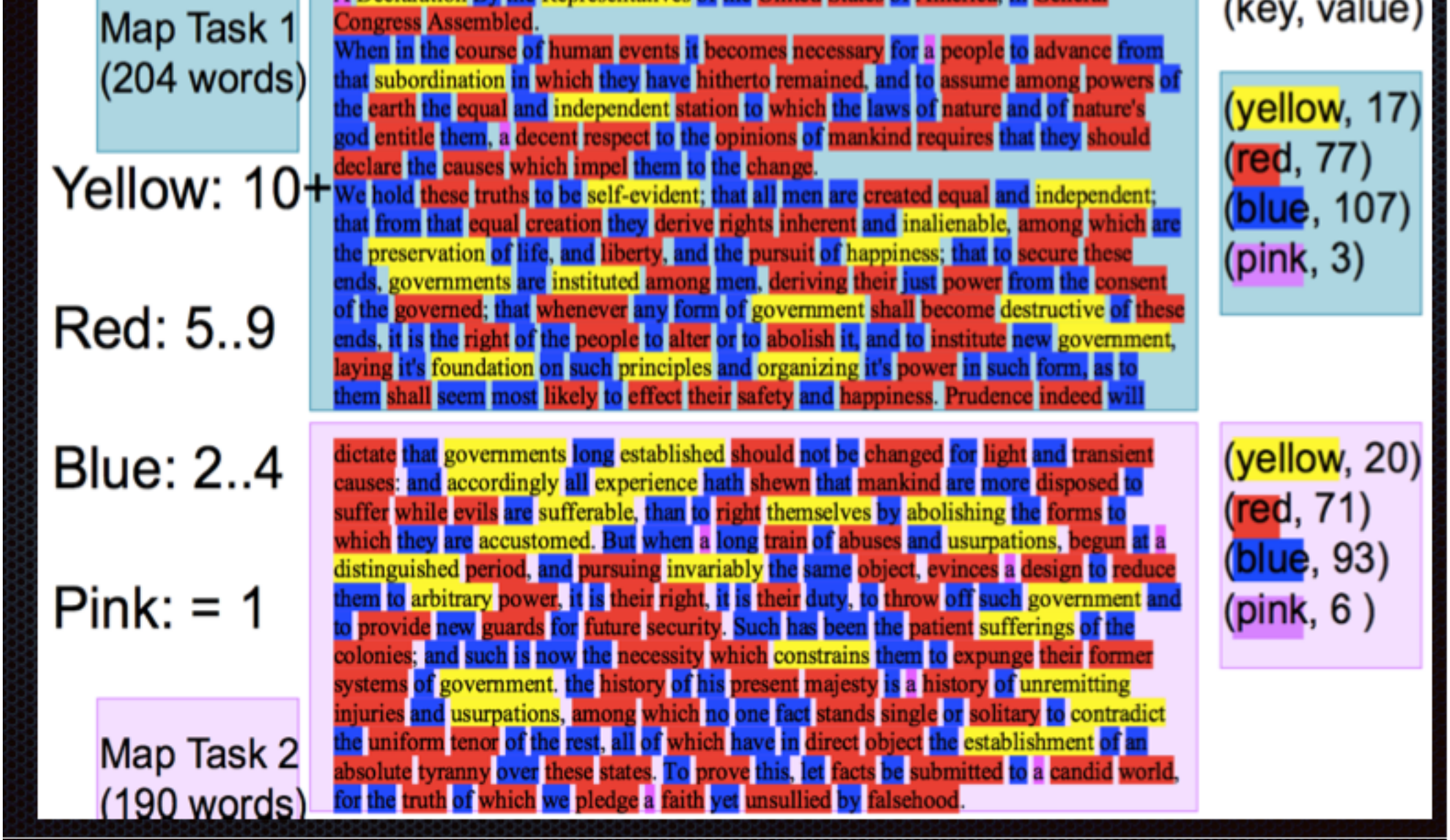
Word Length

Abridged Declaration of Independence

A Declaration By the Representatives of the United States of America, in General Congress Assembled.
When in the course of human events it becomes necessary for a people to advance from that subordination in which they have hitherto remained, and to assume among powers of the earth the equal and independent station to which the laws of nature and of nature's god entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the change.
We hold these truths to be self-evident; that all men are created equal and independent; that from that equal creation they derive rights inherent and inalienable, among which are the preservation of life, and liberty, and the pursuit of happiness; that to secure these ends, governments are instituted among men, deriving their just power from the consent of the governed; that whenever any form of government shall become destructive of these ends, it is the right of the people to alter or to abolish it, and to institute new government, laying it's foundation on such principles and organizing it's power in such form, as to them shall seem most likely to effect their safety and happiness. Prudence indeed will dictate that governments long established should not be changed for light and transient causes: and accordingly all experience hath shewn that mankind are more disposed to suffer while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, begun at a distinguished period, and pursuing invariably the same object, evinces a design to reduce them to arbitrary power, it is their right, it is their duty, to throw off such government and to provide new guards for future security. Such has been the patient sufferings of the colonies; and such is now the necessity which constrains them to expunge their former systems of government. the history of his present majesty is a history of unremitting injuries and usurpations, among which no one fact stands single or solitary to contradict the uniform tenor of the rest, all of which have in direct object the establishment of an absolute tyranny over these states. To prove this, let facts be submitted to a candid world, for the truth of which we pledge a faith yet unsullied by falsehood.

Word Length





Congress Assembled.
When in the course of human events it becomes necessary for a people to advance from that subordination in which they have hitherto remained, and to assume among powers of the earth the equal and independent station to which the laws of nature and of nature's god entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the change.
We hold these truths to be self-evident; that all men are created equal and independent; that from that equal creation they derive rights inherent and inalienable, among which are the preservation of life, and liberty, and the pursuit of happiness; that to secure these ends, governments are instituted among men, deriving their just power from the consent of the governed; that whenever any form of government shall become destructive of these ends, it is the right of the people to alter or to abolish it, and to institute new government, laying it's foundation on such principles and organizing it's power in such form, as to them shall seem most likely to effect their safety and happiness. Prudence indeed will

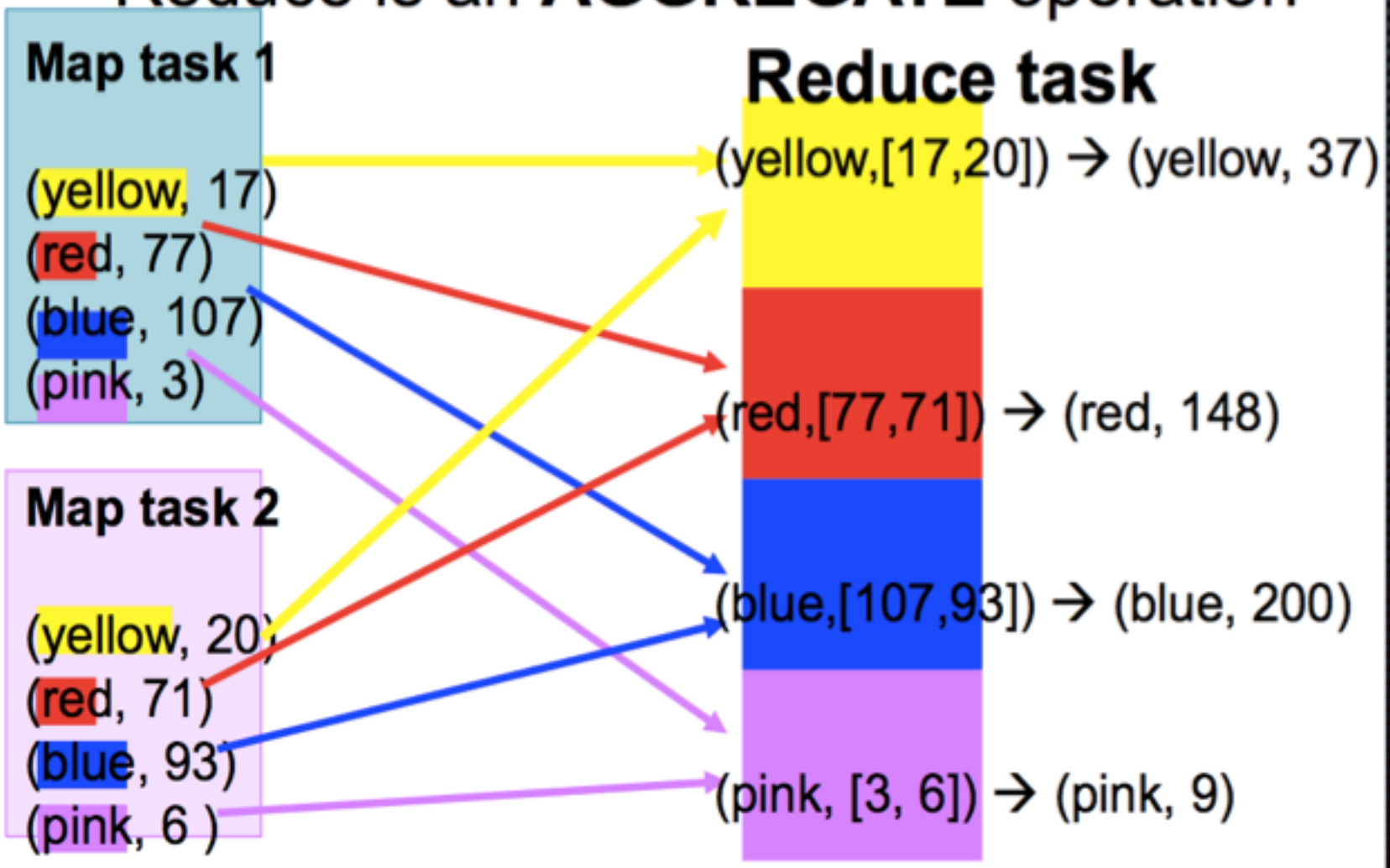
dictate that governments long established should not be changed for light and transient causes: and accordingly all experience hath shewn that mankind are more disposed to suffer while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, begun at a distinguished period, and pursuing invariably the same object, evinces a design to reduce them to arbitrary power, it is their right, it is their duty, to throw off such government and to provide new guards for future security. Such has been the patient sufferings of the colonies; and such is now the necessity which constrains them to expunge their former systems of government. the history of his present majesty is a history of unremitting injuries and usurpations, among which no one fact stands single or solitary to contradict the uniform tenor of the rest, all of which have in direct object the establishment of an absolute tyranny over these states. To prove this, let facts be submitted to a candid world, for the truth of which we pledge a faith yet unsullied by falsehood.

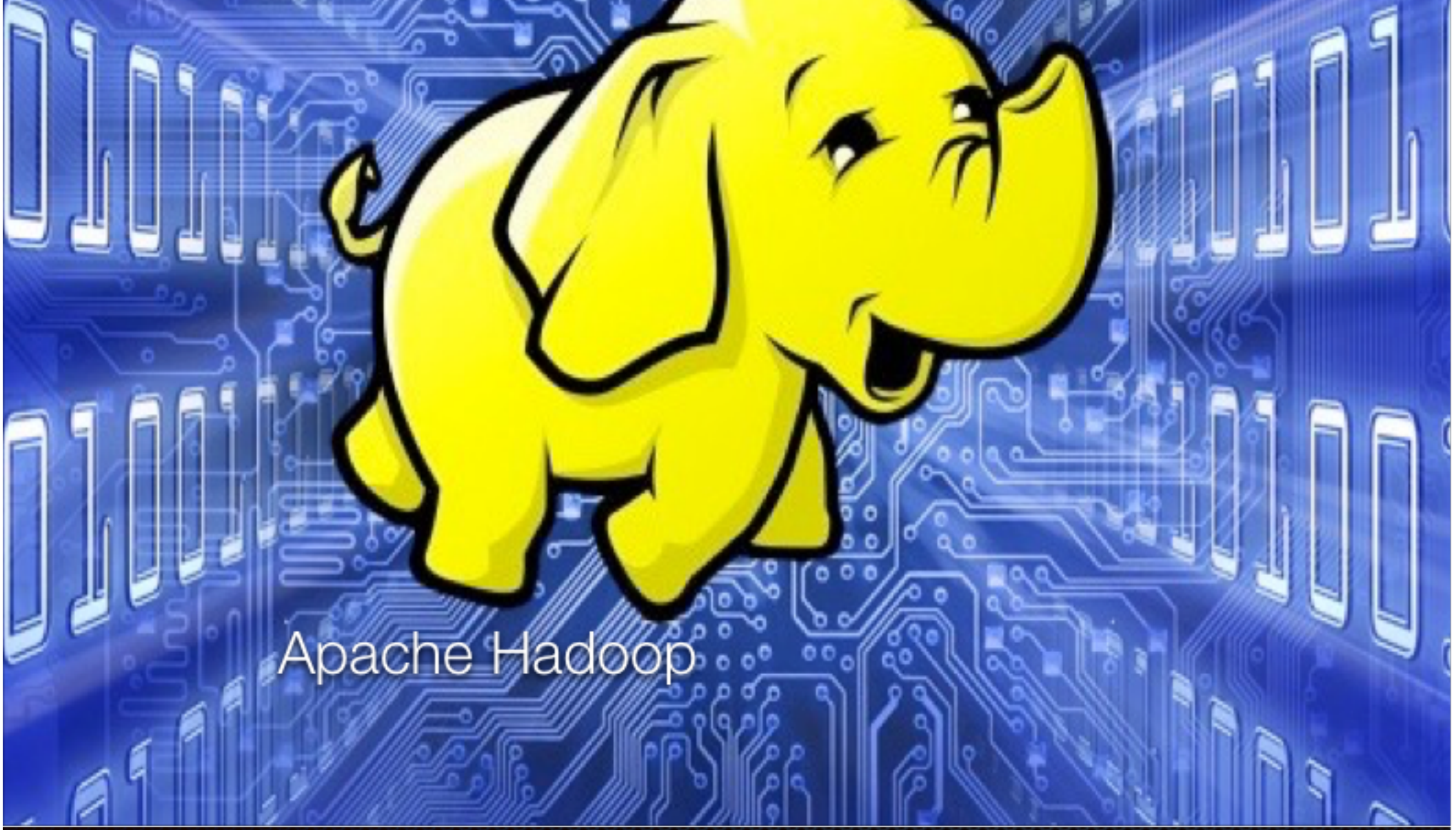
Word Length

Map is a **GROUP BY** operation

Map is a **GROUP BY** operation

Reduce is an **AGGREGATE** operation

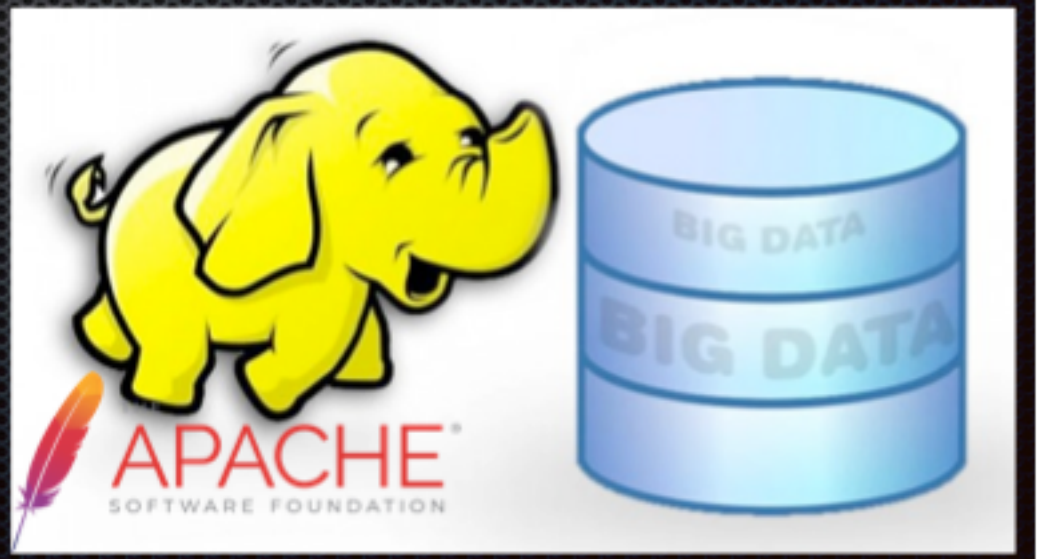




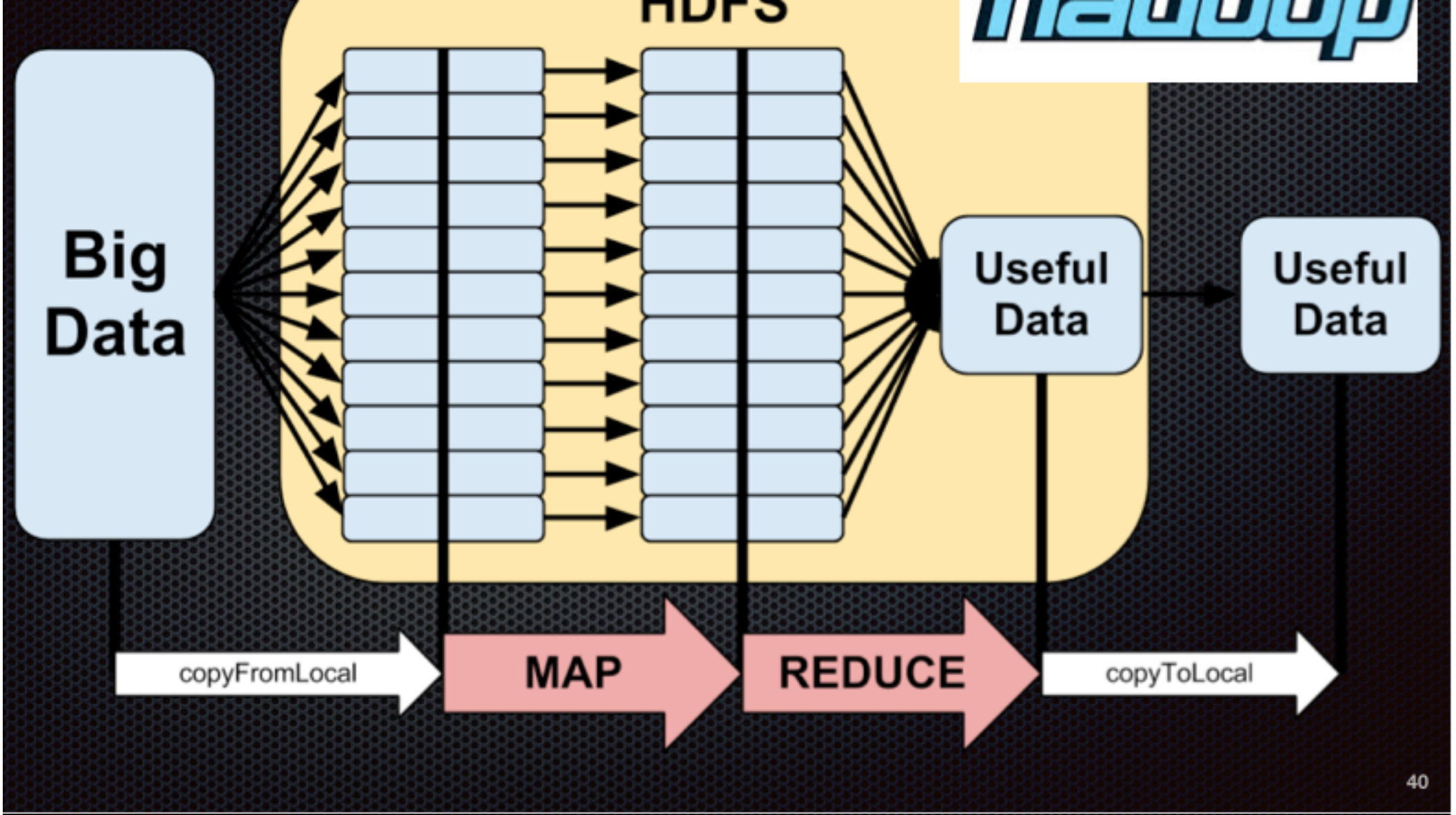
What is Hadoop?

- Hadoop is a top-level Apache project
  - Open source implementation of MapReduce
  - Developed in Java

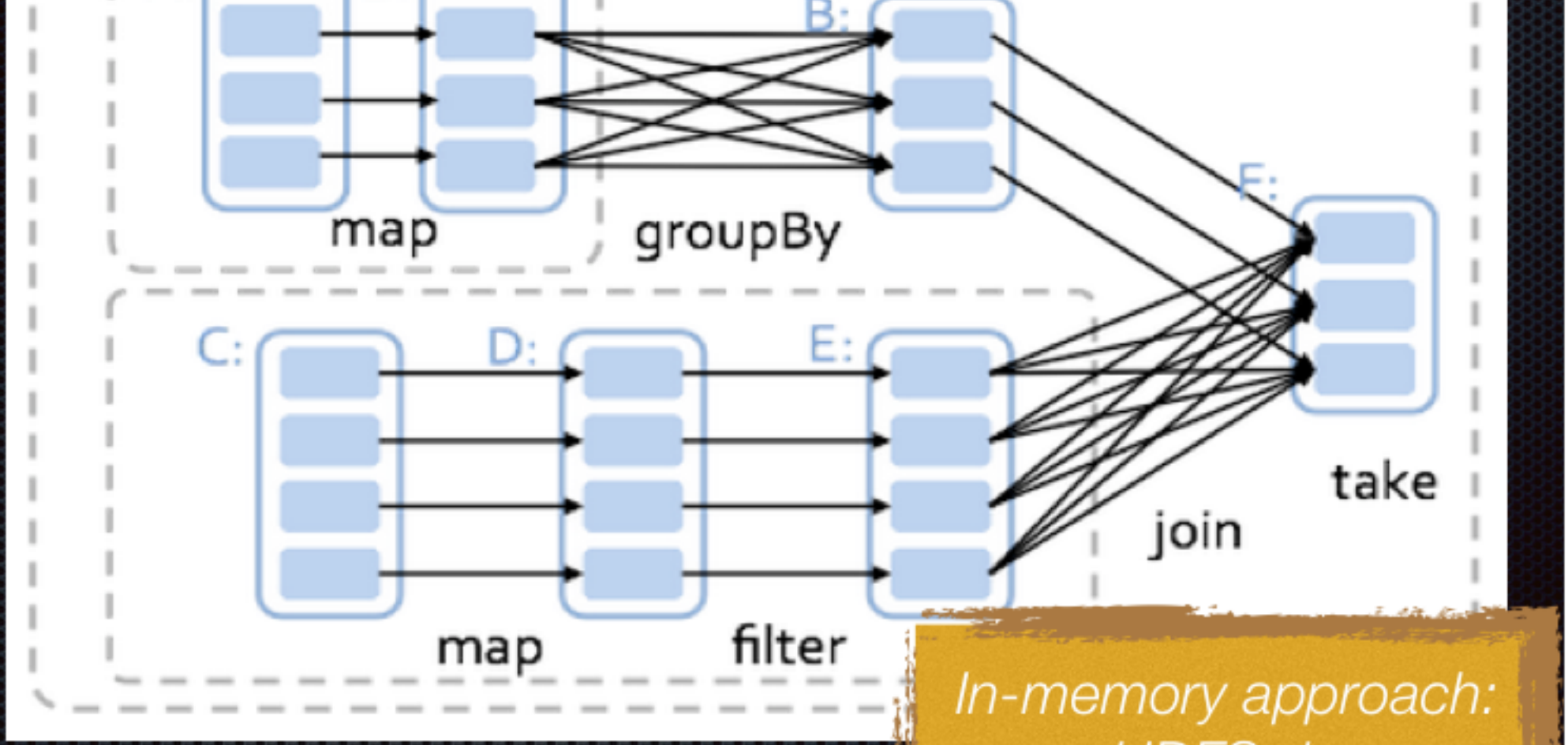
- Platform for data storage and processing
  - Scalable
  - Fault tolerant
  - Distributed
  - Any type of complex data



Hadoop: Single Map-Reduce



Spark: Complex flow of actions

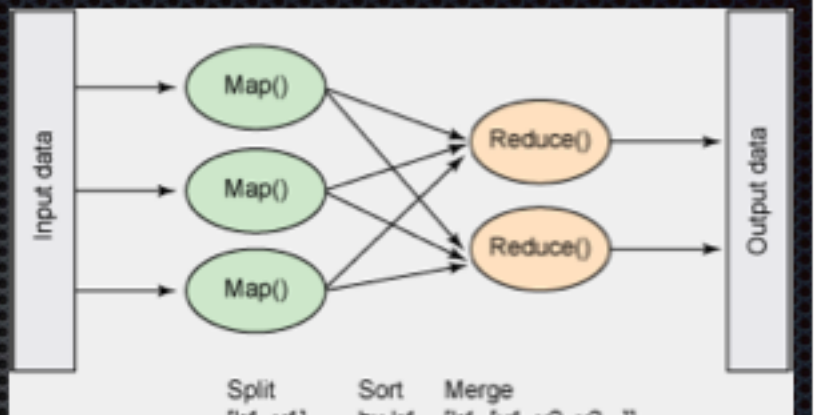


In-memory approach: no HDFS dump between actions

Map-Reduce: Architecture

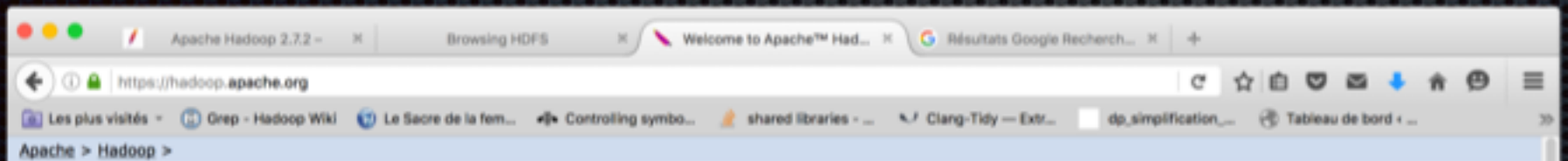
- One master, many workers
- Input data split into M map tasks (typically 64 MB in size)
  - M = 200,000
- Reduce phase partitioned into R reduce tasks
  - R = 4,000
- Tasks are assigned to workers dynamically
  - Workers = 2,000

Parallelism

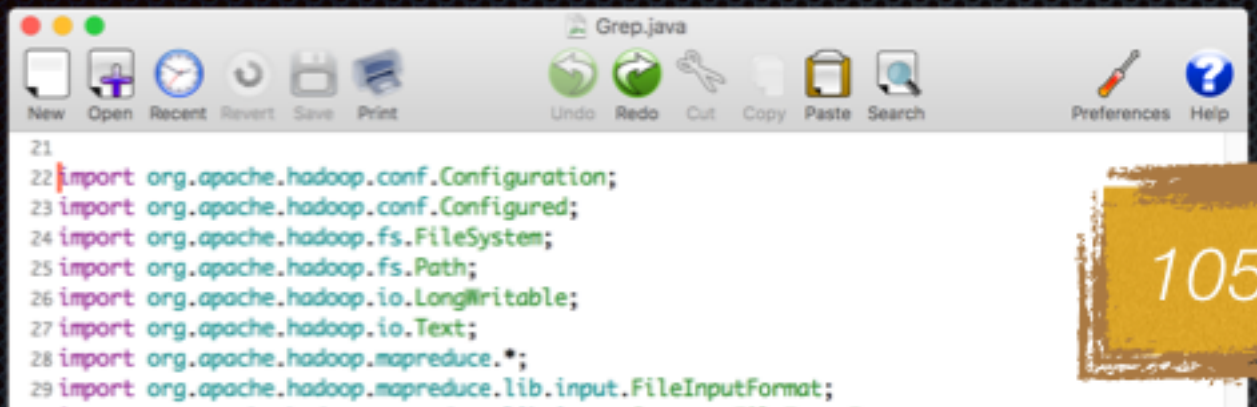


- map() functions run in parallel, creating different intermediate values from different input data sets
- reduce() functions also run in parallel, each working on a different output key
- All values are processed independently
- Bottleneck: reduce phase cannot start until map phase is completely finished*
  - For the original version of MapReduce

Hadoop Programming




hadoop
Top
Wiki
Search with Apache Solr
Search
Last Published: 02/13/2016 08:31:55
About
Welcome
What Is Apache Hado...
Getting Started ...
Download Hadoop
Who Uses Hadoop?...
News
Releases
Mailing Lists
Issue Tracking
Who We Are?
Who Uses Hadoop?
Buy Stuff
Sponsorship
Thanks
Privacy Policy
Bylaws
License
Documentation
Related Projects
Welcome to Apache™ Hadoop®!
What Is Apache Hadoop?
The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.
The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.
The project includes these modules:
Hadoop Common: The common utilities that support the other Hadoop modules.
Hadoop Distributed File System (HDFS™): A distributed file system that provides high-throughput access to application data.
Hadoop YARN: A framework for job scheduling and cluster resource management.
Hadoop MapReduce: A YARN-based system for parallel processing of large data sets.
Other Hadoop-related projects at Apache include:
Ambari™: A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig and Sqoop. Ambari also provides a dashboard for viewing cluster health such as heatmaps and ability to view MapReduce, Pig and Hive applications visually alongwith features to diagnose their performance characteristics in a user-friendly manner.
Avro™: A data serialization system.
Cassandra™: A scalable multi-master database with no single points of failure.
Chukwa™: A data collection system for managing large distributed systems.
HBase™: A scalable, distributed database that supports structured data storage for large tables.
Hive™: A data warehouse infrastructure that provides data summarization and ad hoc querying.
Mahout™: A Scalable machine learning and data mining library.
Pig™: A high-level data-flow language and execution framework for parallel computation.
Spark™: A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide


Grep.java
New
Open
Recent
Revert
Save
Print
Undo
Redo
Cut
Copy
Paste
Search
Preferences
Help
21
22 import org.apache.hadoop.conf.Configuration;
23 import org.apache.hadoop.conf.Configured;
24 import org.apache.hadoop.fs.FileSystem;
25 import org.apache.hadoop.fs.Path;
26 import org.apache.hadoop.io.LongWritable;
27 import org.apache.hadoop.io.Text;
28 import org.apache.hadoop.mapreduce.*;
29 import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
105 lines

```
30 import org.apache.hadoop.mapreduce.lib.input.SequenceFileInputFormat;
31 import org.apache.hadoop.mapreduce.lib.map.InverseMapper;
32 import org.apache.hadoop.mapreduce.lib.map.RegexMapper;
33 import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
34 import org.apache.hadoop.mapreduce.lib.output.SequenceFileOutputFormat;
35 import org.apache.hadoop.mapreduce.lib.reduce.LongSumReducer;
36 import org.apache.hadoop.util.Tool;
37 import org.apache.hadoop.util.ToolRunner;
38
39 /* Extracts matching regexs from input files and counts them. */
40 public class Grep extends Configured implements Tool {
41   private Grep() {}                                // singleton
42
43   public int run(String[] args) throws Exception {
44     if (args.length < 3) {
45       System.out.println("Grep <inDir> <outDir> <regex> [<group>]");
46       ToolRunner.printGenericCommandUsage(System.out);
47       return 2;
48     }
49
50     Path tempDir =
51       new Path("grep-temp-"+
52           Integer.toString(new Random().nextInt(Integer.MAX_VALUE)));
53
54     Configuration conf = getConf();
55     conf.set(RegexMapper.PATTERN, args[2]);
56     if (args.length == 4)
57       conf.set(RegexMapper.GROUP, args[3]);
58
59     Job grepJob = new Job(conf);
60
-:---  Grep.java    24% (22,0)    (Java/l Abbrev)
Undo!
```


grepJob.setJobName("grep-search");

FileInputFormat.setInputPaths(grepJob, args[0]);

*grepJob.setMapperClass(**RegexMapper**.class);*

*grepJob.setCombinerClass(**LongSumReducer**.class);*
*grepJob.setReducerClass(**LongSumReducer**.class);*

```
FileOutputFormat.setOutputPath(grepJob, tempDir);
grepJob.setOutputFormatClass(SequenceFileOutputFo
grepJob.setOutputKeyClass(Text.class);
grepJob.setOutputValueClass(LongWritable.class);

grepJob.waitForCompletion(true);

Job sortJob = new Job(conf);
sortJob.setJobName("grep-sort");

FileInputFormat.setInputPaths(sortJob, tempDir);
sortJob.setInputFormatClass(SequenceFileInputFormat.class);

sortJob.setMapperClass(InverseMapper.class);

sortJob.setNumReduceTasks(1);              // write a single file
FileOutputFormat.setOutputPath(sortJob, new Path(args[1]));
sortJob.setSortComparatorClass(        // sort by decreasing freq
  LongWritable.DecreasingComparator.class);

sortJob.waitForCompletion(true);
```

```
public void map(K key, Text value,
                Context context)
  throws IOException, InterruptedException {
  String text = value.toString();
  Matcher matcher = pattern.matcher(text);
  while (matcher.find()) {
    context.write(new Text(matcher.group(group)),
                  new LongWritable(1));
  }
```


Map-Reduce and Hadoop in a Nutshell

- Data-parallel programming model
  - Automatic division of job into tasks

- Automatic partition and distribution of data
- Automatic placement of computation near data
- Recovery from failures

- Hadoop
  - Open source implementation
  - Many useful subprojects
  - Very large user community

Let users focus on application, not on complexity of distributed computing

A-Brain: A Bioinformatics Application

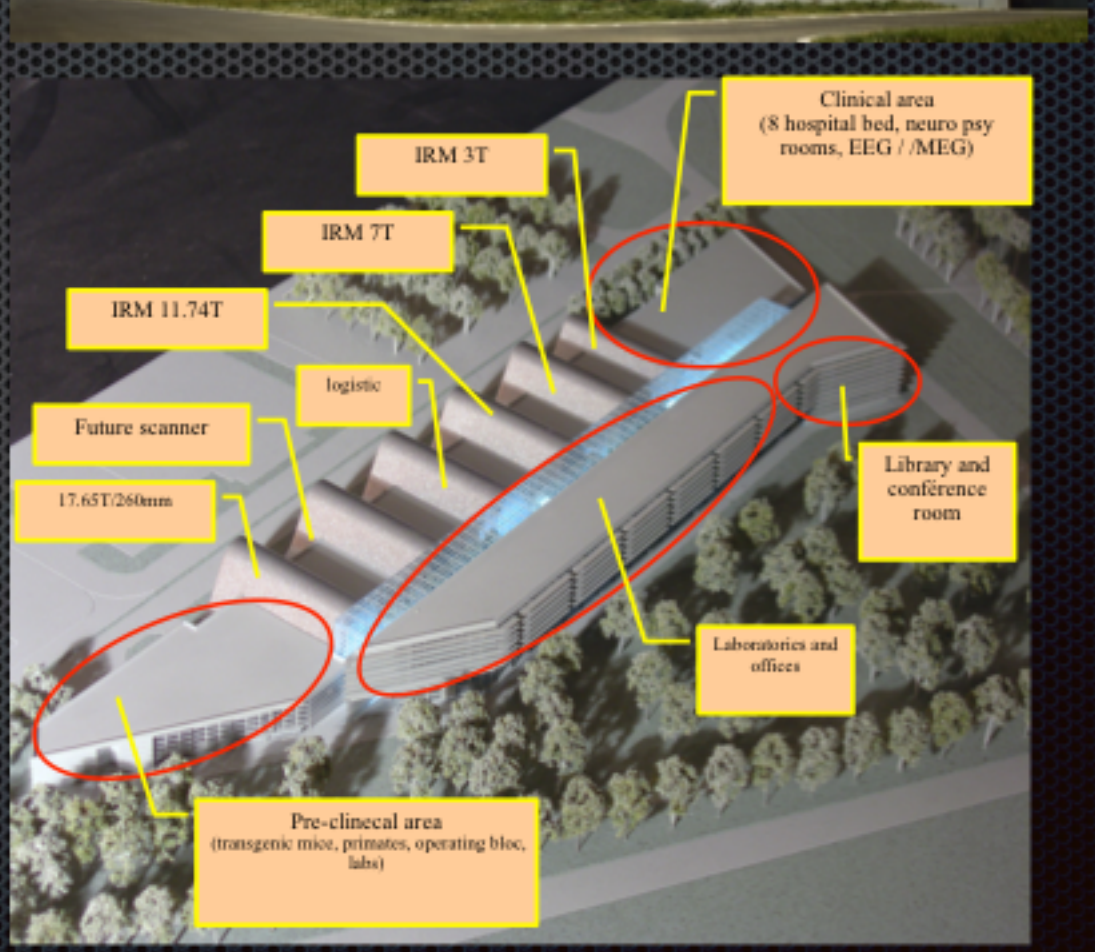
A Bioinformatics Application Benefiting from Map-Reduce on Clouds

The A-Brain Project
Data-Intensive Processing on Microsoft Azure Clouds

- Application
  - Large-scale joint genetic and neuroimaging data analysis

NeuroImaging data analysis

- Goals
  - Application: assess and understand the variability between individuals
  - Infrastructure: assess the potential benefits of Azure
- Approach
  - Optimized data processing on Microsoft's Azure clouds
- Framework
  - Joint MSR-Inria Research Center
  - KerData and Parietal Inria teams

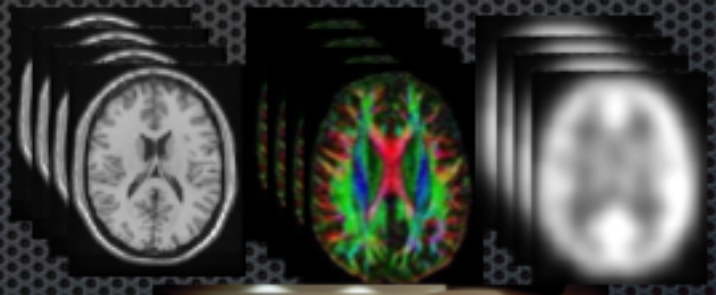


Goal: detect risk factors for psychiatric diseases/ behaviors

Clinical / Behavior

Genetic Data

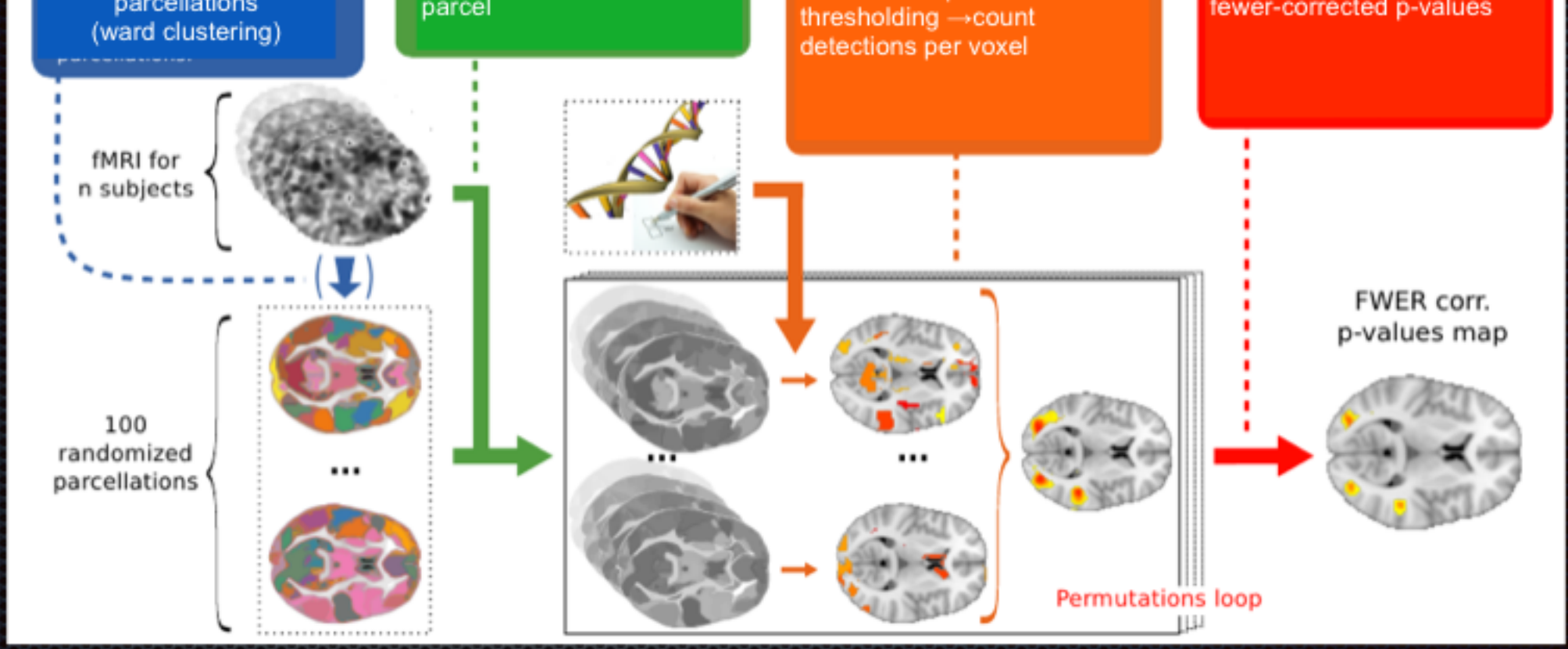
G G
T G
T T
T G
G G



Focus on this link

Brain Data

Randomized-parcellation based inference

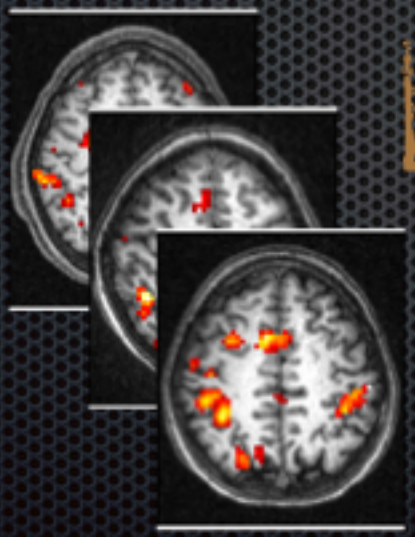


Statistical analysis for large-scale neuroimaging-genetics

- Image data → 4D to 2D, dimension nvoxels × nsubjects

- Genetic data → dimension nsnps × nsubjects
- Statistical confidence?

nvoxels= 10^5
nsnps= 10^6



Correlations

SNP data

A really Big Data challenge!

Data: $.8 \times 10^4 \times 5 \times 10^4 \times 5 \times 10^5$ =>

5%-10%

useful

double permutation voxels SNPs

Computation: 10^4 × 5×10^4 × 5×10^5 ➔ 2.5×10^{14} associations

The algorithm: 1.5×10^6 associations / second

Estimate timespan on single machine:

1.67×10^8 seconds ➔ 5.3 years

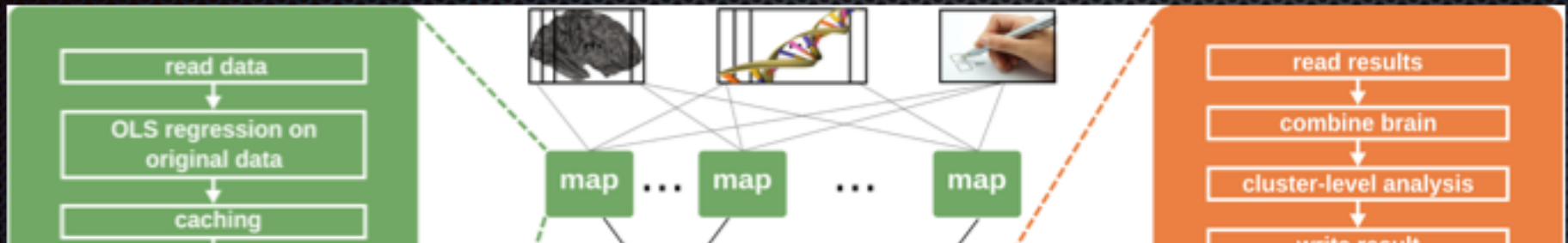
The cloud can help

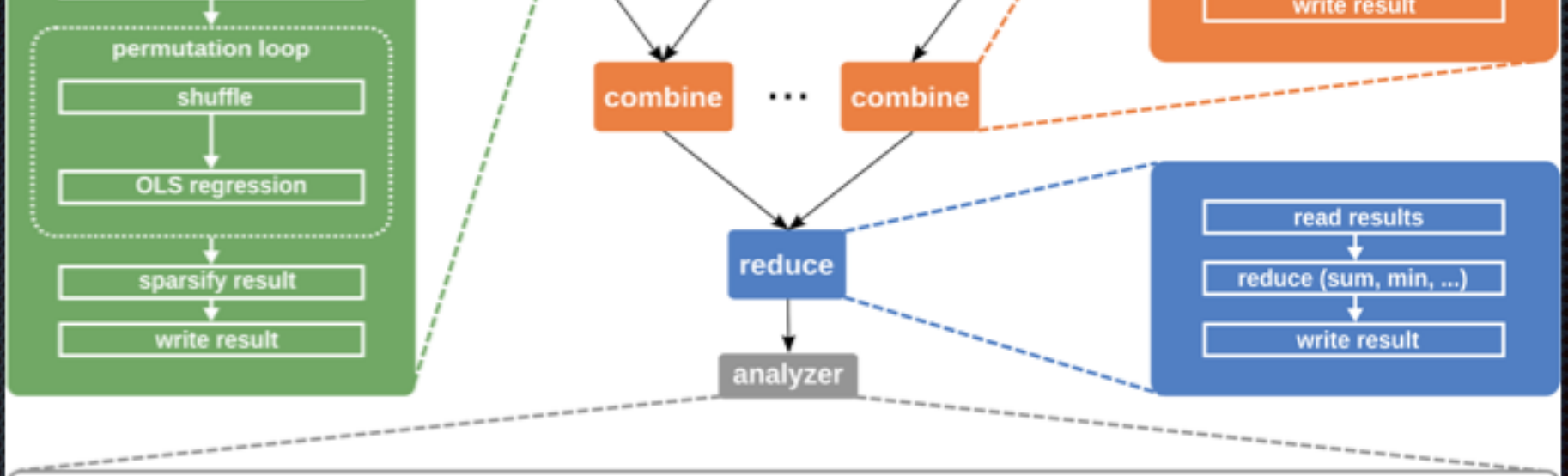
Estimation for A-Brain on Azure (350 cores)

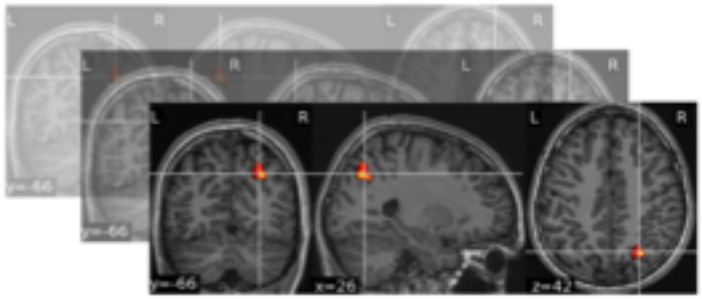
$2.5 \times 10^{14} / 350 \times 1.47 \times 10^{6} = 485 \times 10^{3}$ seconds

5.3 years ➔ ***5.6 days***

Approach: A-Brain as Map-Reduce Processing







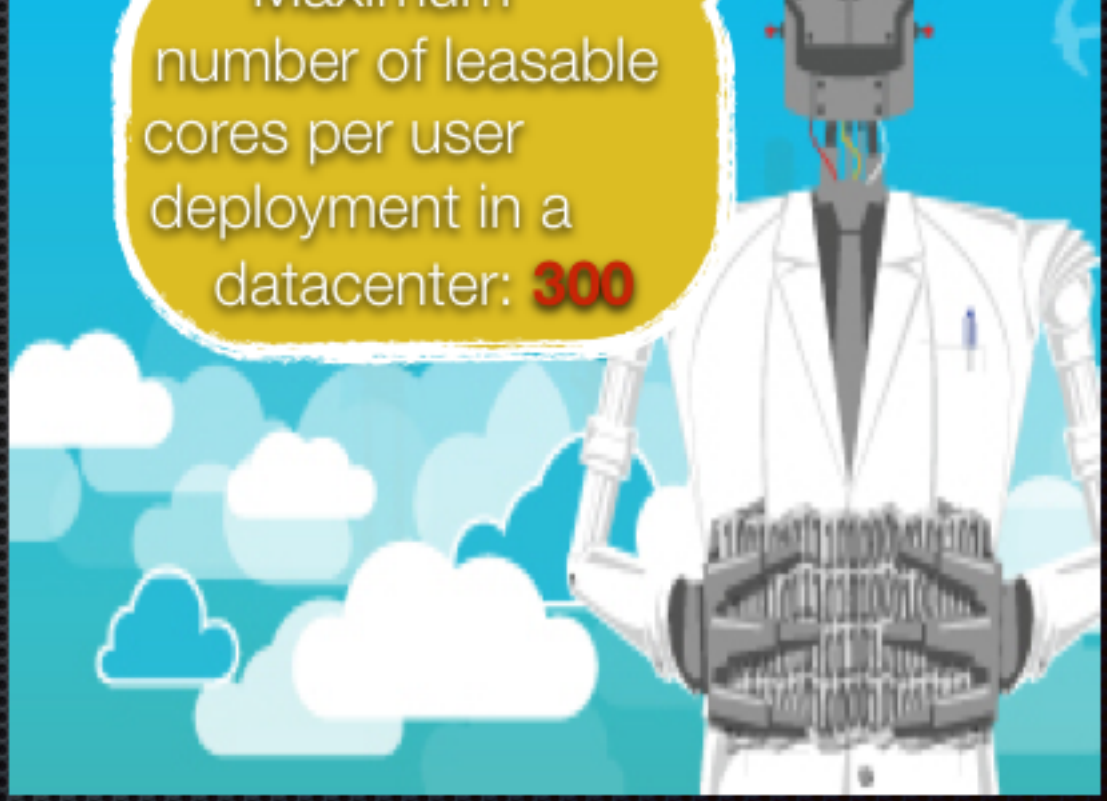
SNP rs	Chr	Volume (mm³)	Corr. P-value	Cluster Volume (mm³)	Corr. P-value	Minimum P-value	Corr. P-value
rs2788031	1	810	1	810	1	1,43E-006	1
rs2185077	1	648	1	540	1	3,18E-005	1
rs10888855	1	1053	1	1026	1	2,51E-005	1
rs10888857	1	2241	1	2241	0,9995	3,76E-007	1
rs3753009	1	2592	0,9888	2565	0,9651	1,90E-007	1
rs7551981	1	540	1	540	1	5,98E-005	1
rs4433442	1	783	1	783	1	5,69E-005	1

Deploying A-Brain on the cloud

- Data: 1.8 PB

- Data: 1.81 TB (5%-10% useful)
- Estimated makespan on 350 cores: 5 days

Microsoft Azure



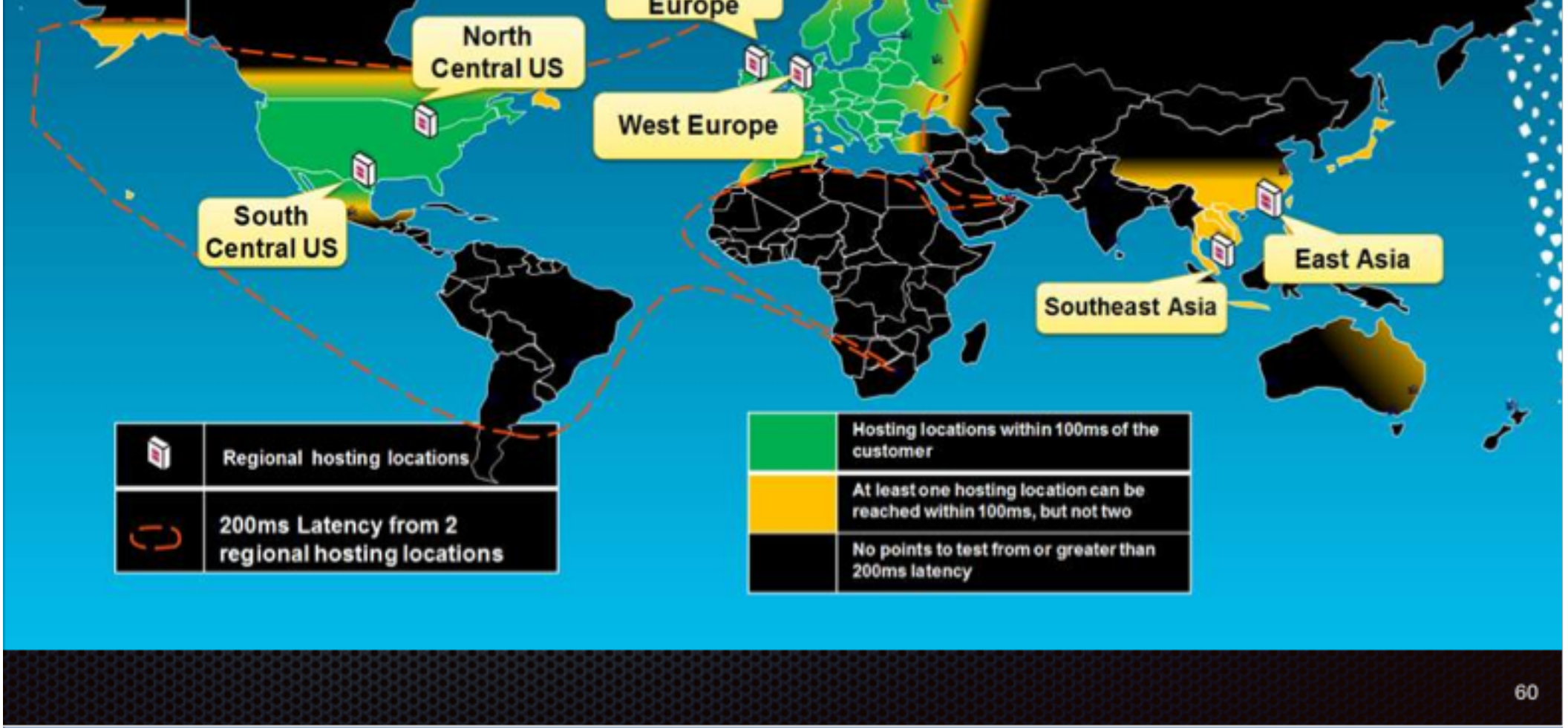
Enabling MapReduce Processing across Cloud Data centers

across Cloud Data-centers

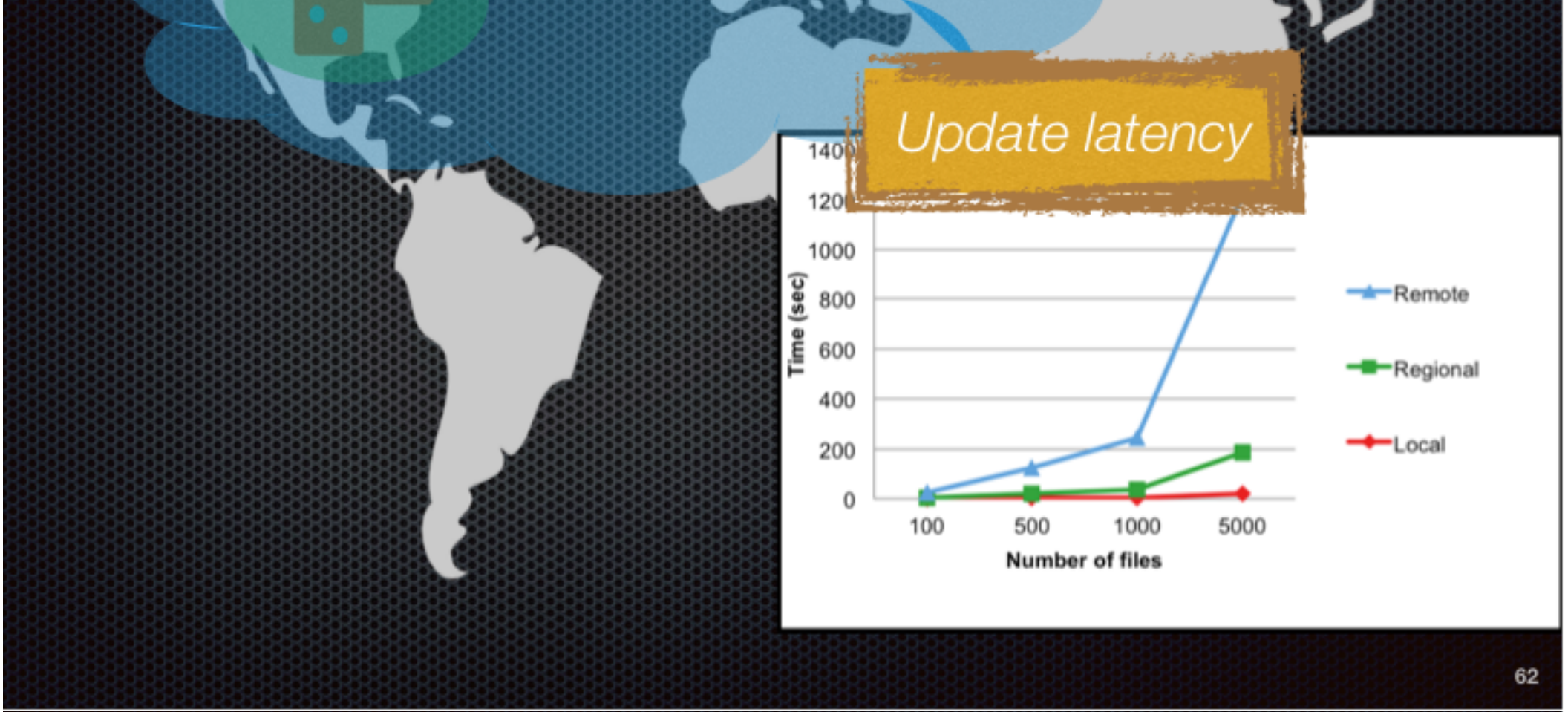
What is the cloud? One datancenter?

What is the cloud? One datacenter?

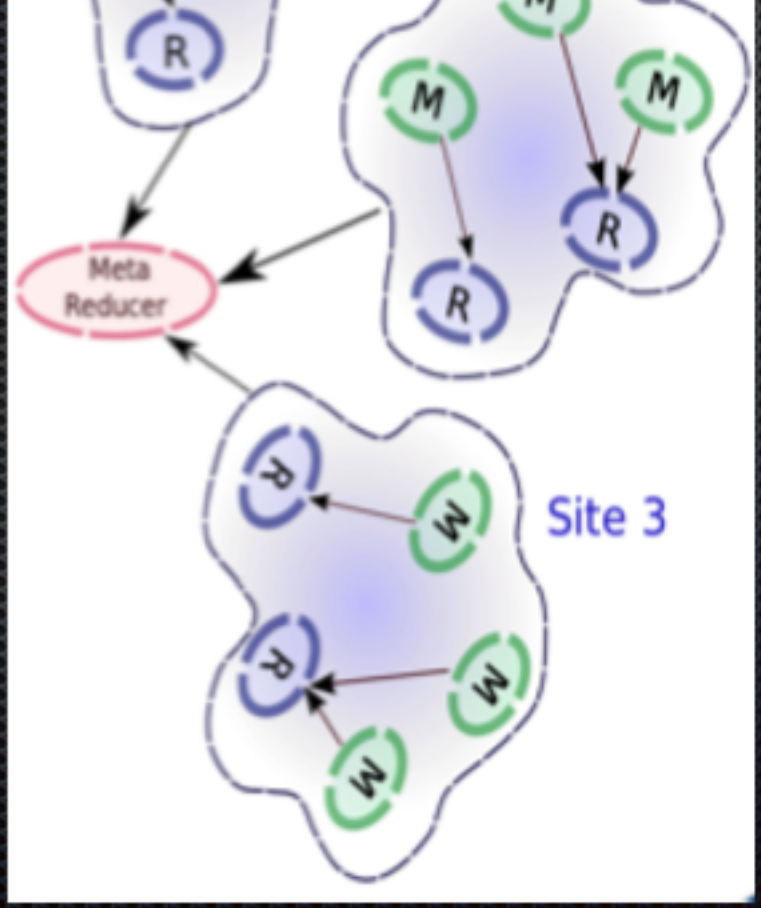
What is the cloud? Many geographically distributed datacenters!


North Central US
West Europe
South Central US
East Asia
Southeast Asia
Regional hosting locations
200ms Latency from 2 regional hosting locations
Hosting locations within 100ms of the customer
At least one hosting location can be reached within 100ms, but not two
No points to test from or greater than 200ms latency

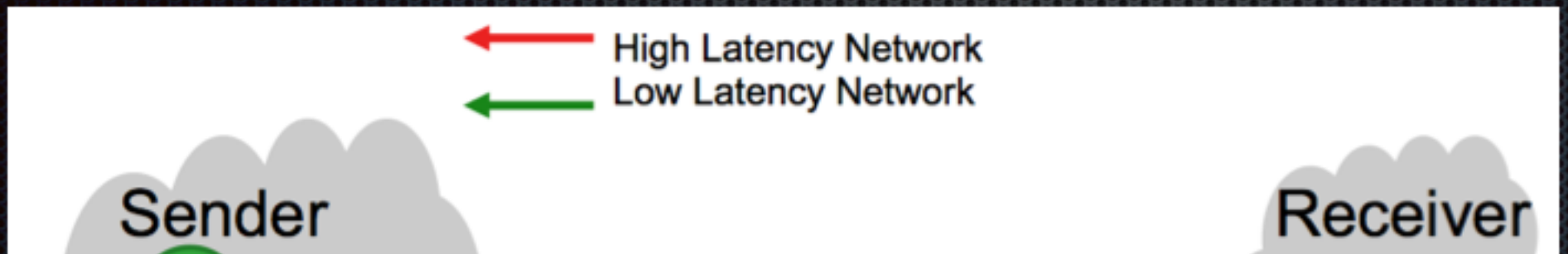
Main obstacle: the network!

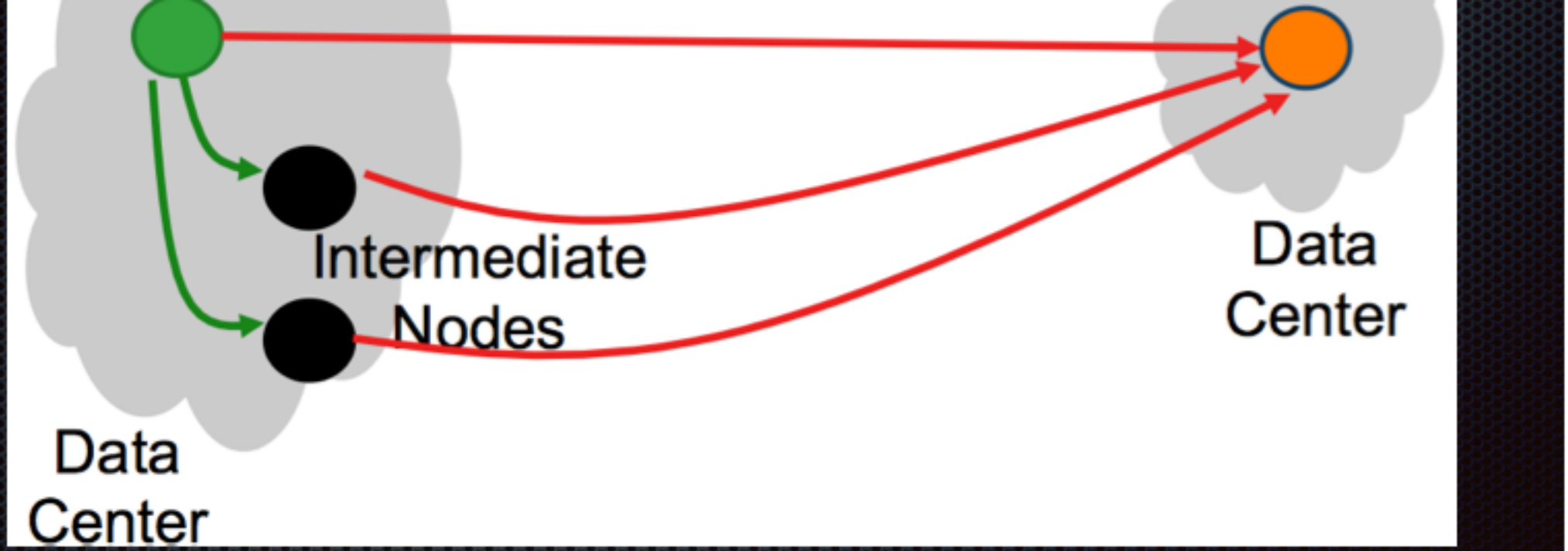


Beyond single site processing



Leverage network parallelism: multi-path transfers





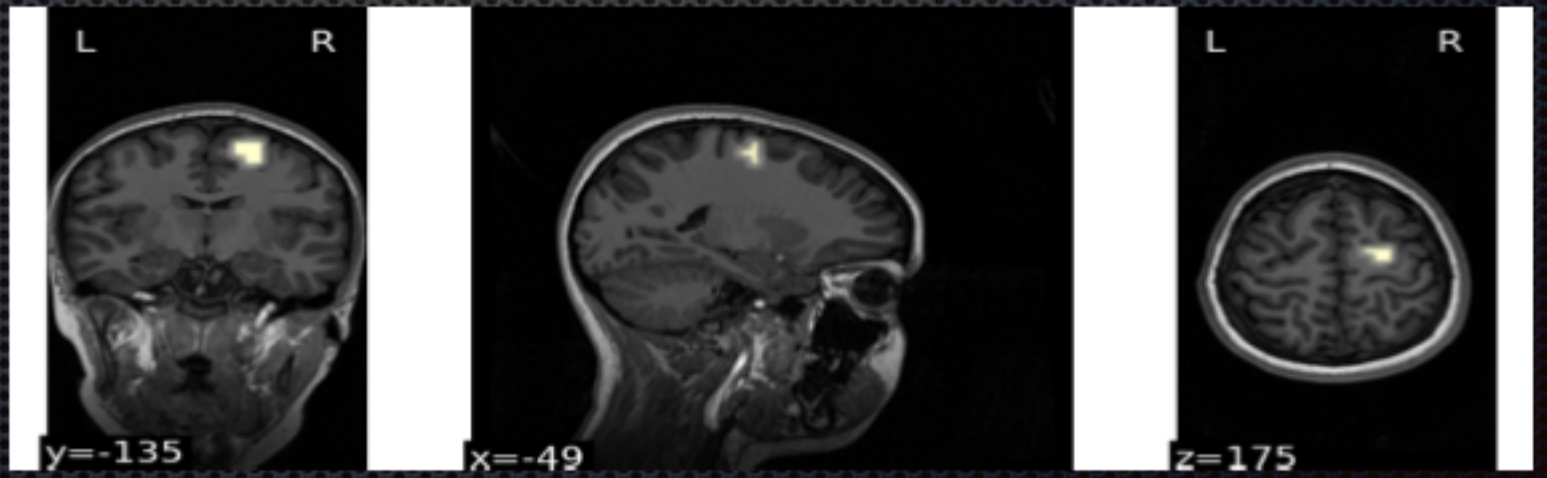
It worked!

finding associations:

p(Brain image , Genetic data)

associations.

>2000 subjects



To take away

- Map-Reduce is not THE solution
  - Efficient for some classes of applications
  - Inefficient for others (e.g. real-time processing)

- New platforms build on MapReduce to propose more efficient processing
  - Spark: in-memory processing
  - Support for streaming
- Multi-site data management remains an open-issue
  - Overlap data upload with data processing
  - Exploit network parallelism
  - Play with the parameters

Thank you!


DATA
www.nuxeo.com
Soulcié