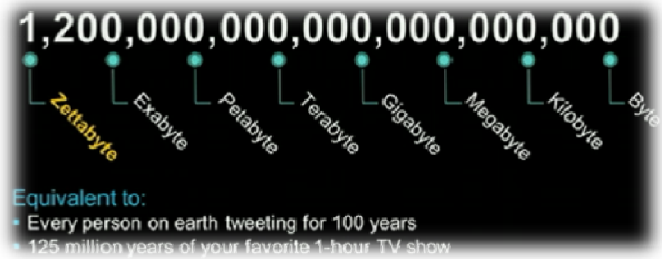
INVENTORS FOR THE DIGITAL WORLD

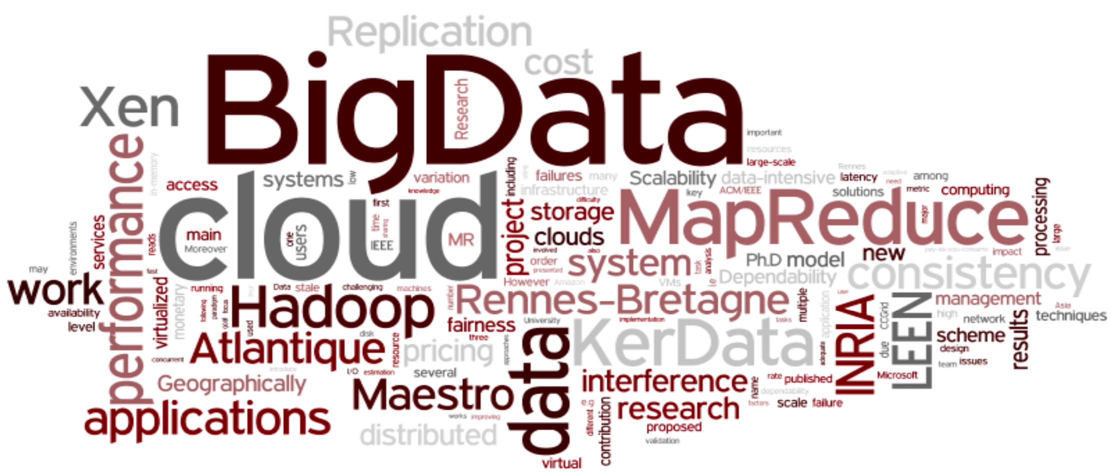
Big Data Processing in the Cloud: Hadoop and Beyond

SHADI IBRAHIM
Shadi.ibrahim@inria.fr

Lecture Outline



- Big Data overview
- Google MapReduce
- Hadoop
- Post-Hadoop



Big Data


POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL CENTERS
Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate
OFFICER ALGORITHM
Can a Crime Be Prevented Before It Begins?
NEW WAYS OF SEEING
A Gallery of Extraordinary Infographics
PLUS
Juan Enriquez Reprograms Life
James Gleick Unsplits the Bit
AND Lawrence Weschler Questions the Cloud
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE


01 Business
LE MAGAZINE QUI MET L'ENTREPRISE À L'HEURE DU NUMÉRIQUE
LES PROMESSES DU
BIG DATA
ACTUALITÉ
Le boom des Fab Labs donne un coup de fouet à notre R&D P.20


cnrs
international magazine
n° 28
Quarterly
January 2013
The
BIG DATA
REVOLUTION

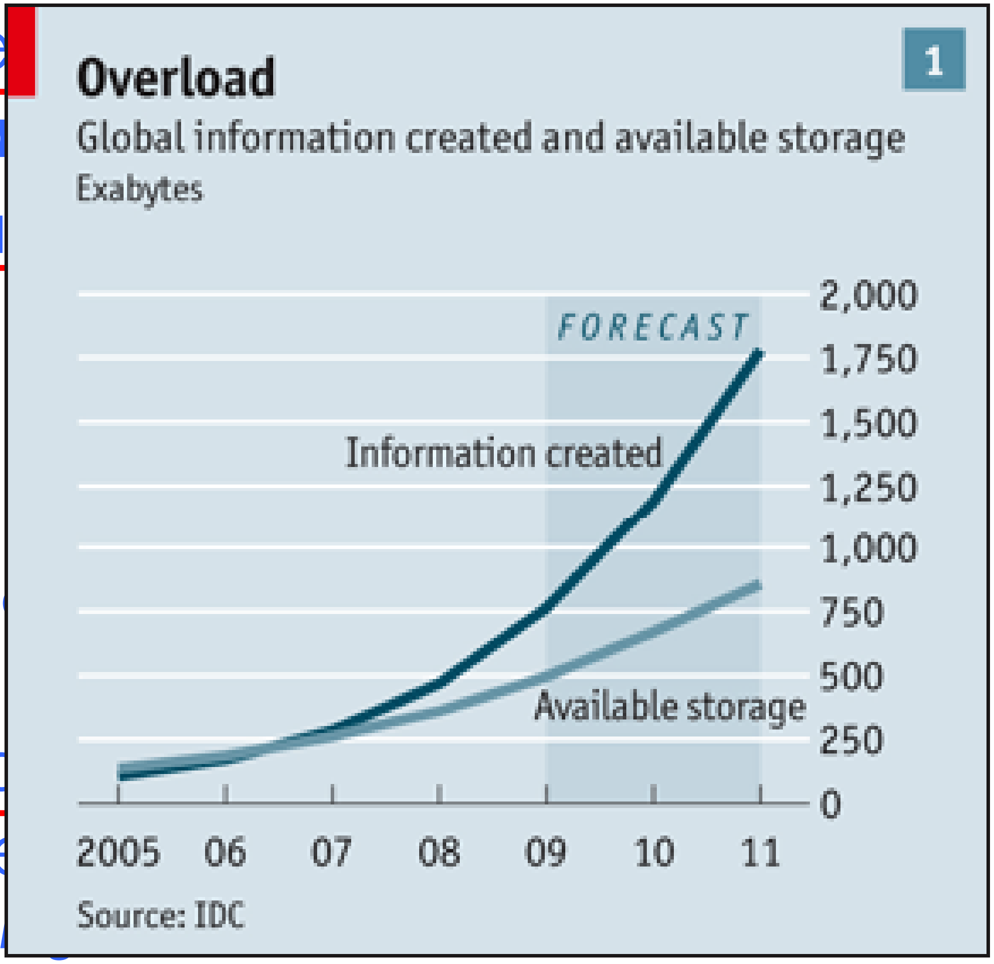

4 September 2008 | www.nature.com/nature | £10
nature
THE BITER BIT
Viral infections for viruses
TROPICAL CYCLONES
The strong get stronger
BLACK HOLE PHYSICS
A new window on the Galactic Centre
BIG DATA
NATUREJOBS
Minnesota musings
SCIENCE IN THE PETABYTE ERA


8-PAGE SPECIAL POSTGRADUATE SURVIVAL GUIDE
8TH BIRTHDAY ISSUE!
COSMOS
THE SCIENCE OF EVERYTHING
THE END OF VIOLENCE
Steven Pinker on the new peace
DEFEATING POLIO
Will politics jeopardise a cure?
FRAUDS AND FAKES
Science's biggest scams
GENIUS OF DOGS
Inside the canine brain
IS data THE NEW GOD?
How tracking your digital trail could predetermine your future – and why you'll benefit from today's data deluge. p46
GALAXIES AND NEBULAE


Pipeline | Volume 11 | Issue 3
Pipeline
The Unstructured Data Structure
GET SMARTER! NEXT-GEN ANALYSIS
BIG DATA TO THE RESCUE
THE BIG DATA HUBRIS
BIG DATA & ANALYTICS
EXPERT TIPS for Call Center CEM
Box Fresh: Real-time Contextual Awareness
Big Data Takes the C-suite
Big Data Infrastructure Demands
Machine Learning and Big Data

What is Big Data?

"Big data refers to data se
ability of typical databa
store, manage and ana
Institute, 2011



"Big data is the term for a
large and complex that
process using on-hand c
or traditional data proce
http://en.wikipedia.org/wiki/

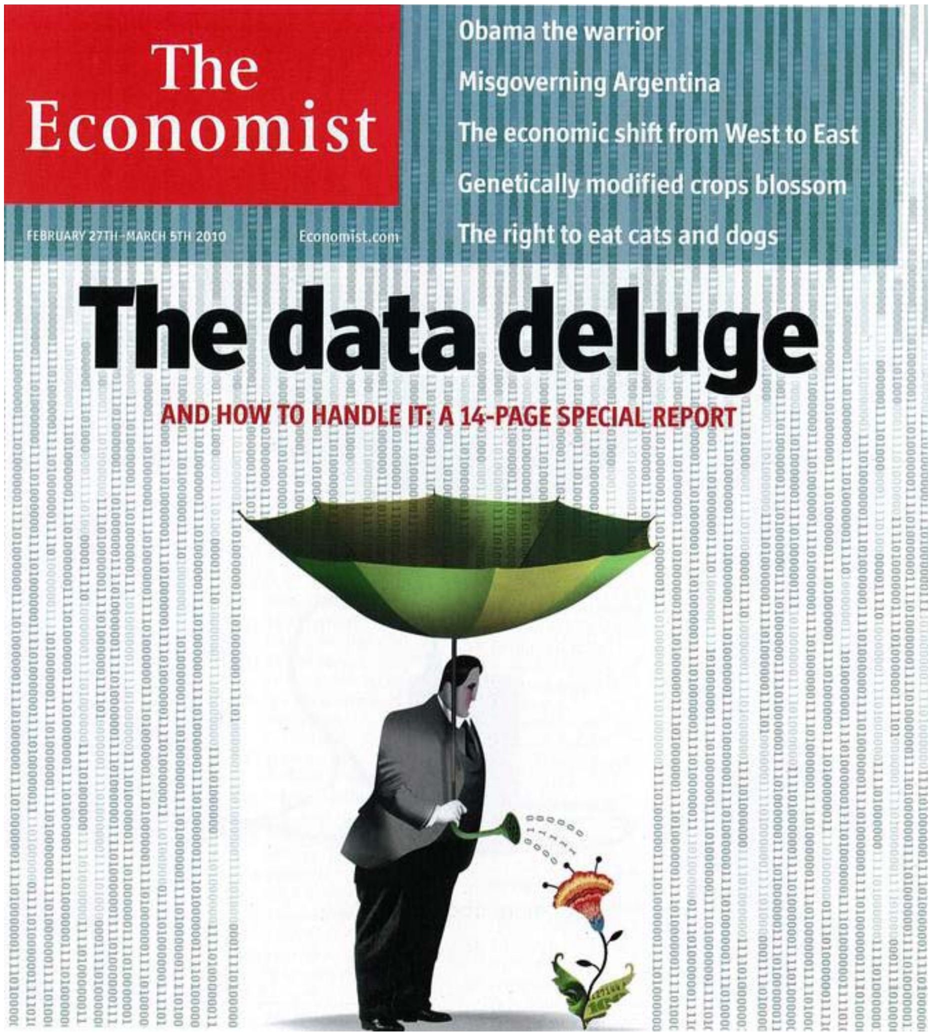
How *big* is Big Data ?

Earlier Berkeley studies estimated that by the end of 1999, the sum of human-produced information (including all audio, video recordings and text/books) was about 12 Exabytes of data.

Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up To 2003.

http://techcrunch.com/2010/08/04/schmidt-data/

Big picture of Big Data



Source: The economist Feb 25 2010

In 2010 The digital Universe was

1.2 ZettaBytes

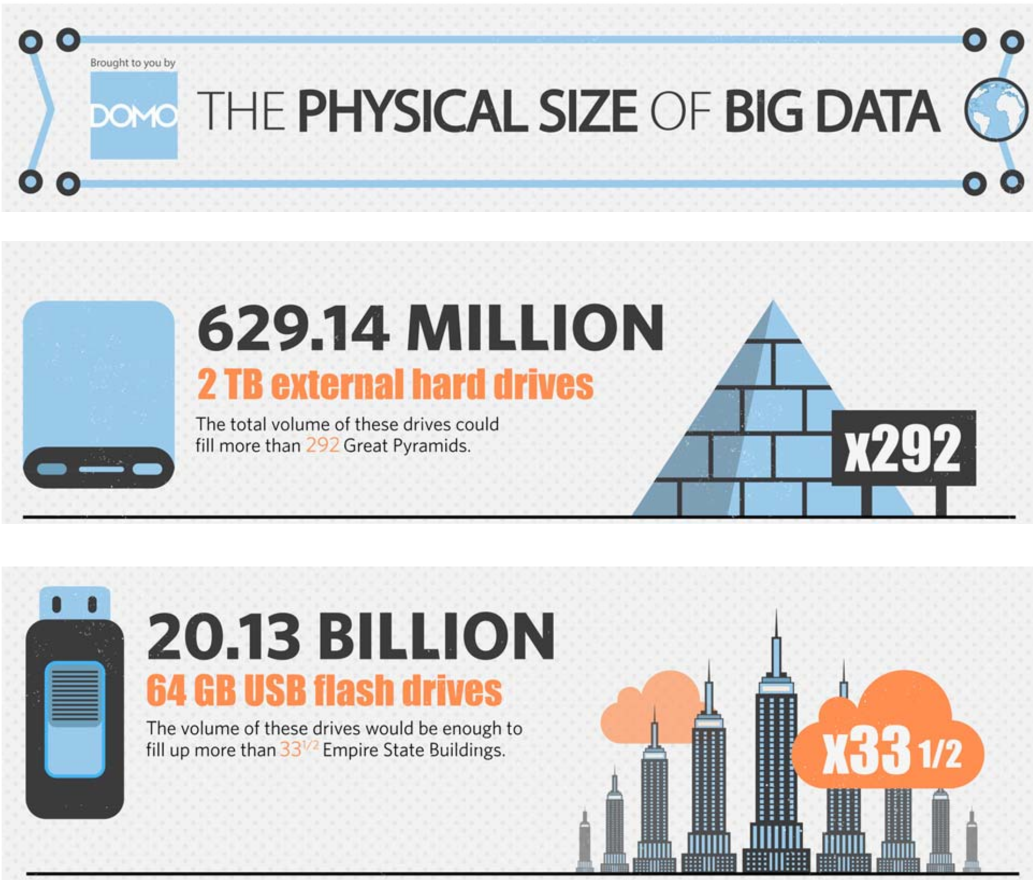
In 2015

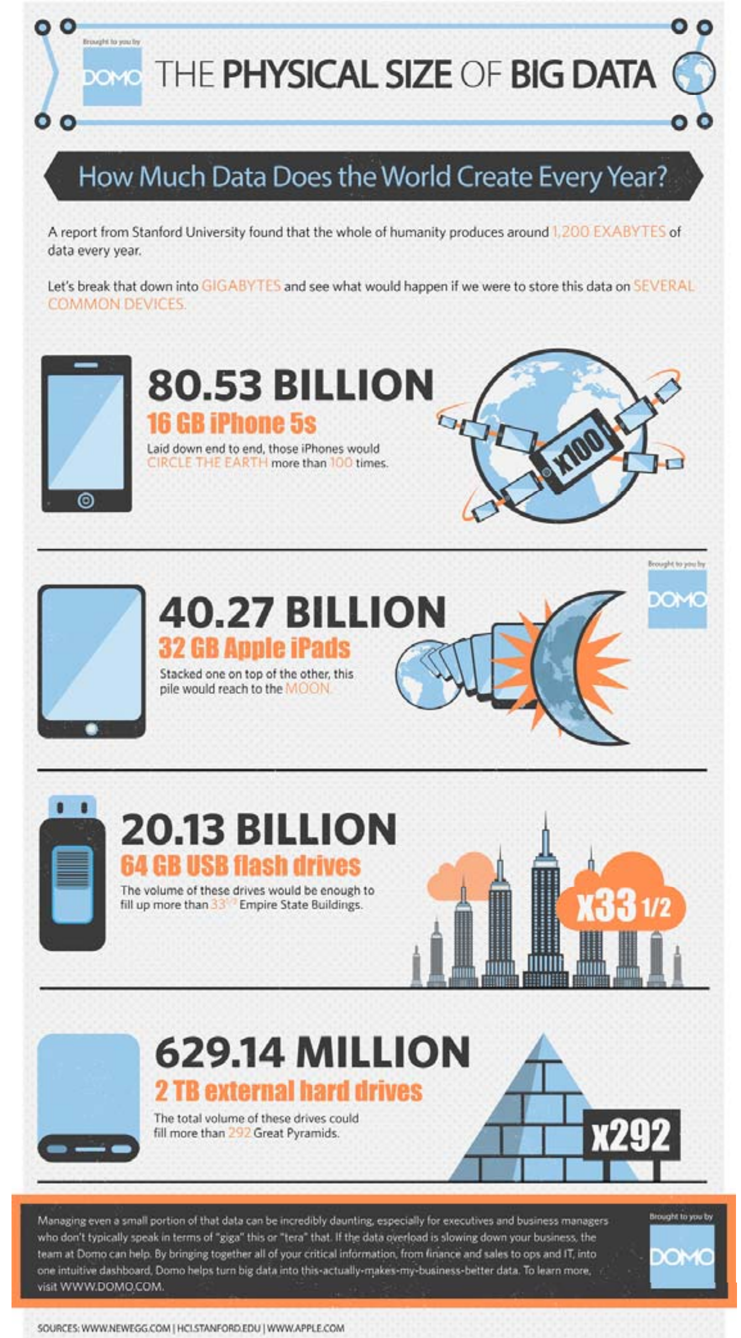
8 ZettaBytes

In a decade the Digital Universe will be

35 ZettaByte

1.2 Zettabyte?





Data inflation

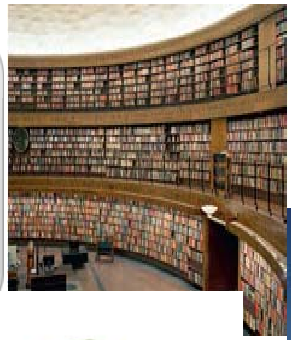
2

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Greek. All the catalogued books in America's Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America's postal service this year will amount to around 5PB. Google processes around 1PB every hour
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of *The Economist*
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information in existence this year is forecast to be around 1.2ZB
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine

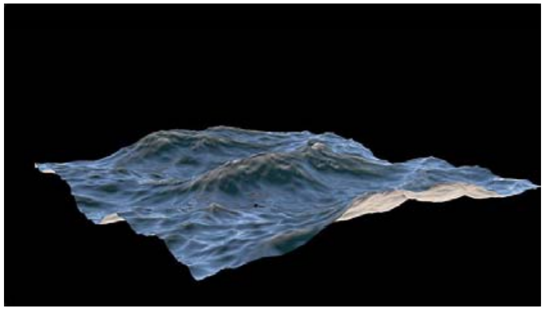
The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

Source: *The Economist*

What are the sources of Big Data?

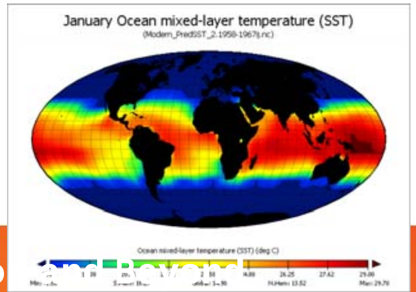













Our Data-Driven world: Science

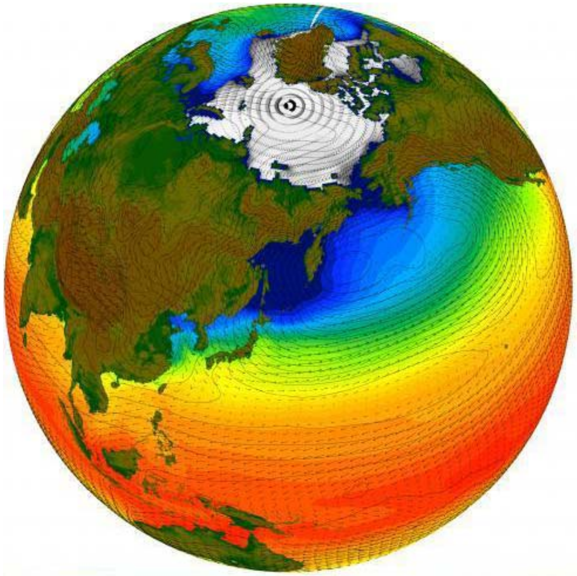
Particle accelerators in physics

*CERN's Large Hydron Collider (LHC) generates **15 PB** a year*

Astronomical instruments

*SQUARE KILOMETRE ARRAY (SKA) :The world's largest radio telescope) will collect **1 PB** a day ~ **400 PB** a year*

Climate Simulations



The NASA Center for Climate Simulation (NCCS) stores 32 petabytes of climate observations and simulations on the Discover supercomputing cluster.

Genome sequencers in biology

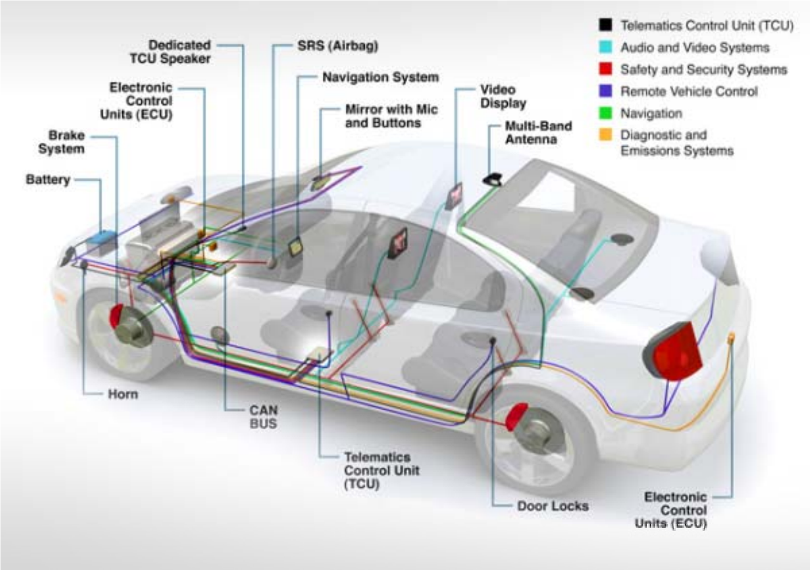
Web Data

- Google processes **20 PB** a day (2008)
- eBay has **6.5 PB** of user data + 50 TB/day (5/2009)

Social Networks

- Facebook has 2.5 PB of user data + 15 TB/day (4/2009)
- Twitter Generate approximately 12 TB of data per day

Our Data-Driven world Industry



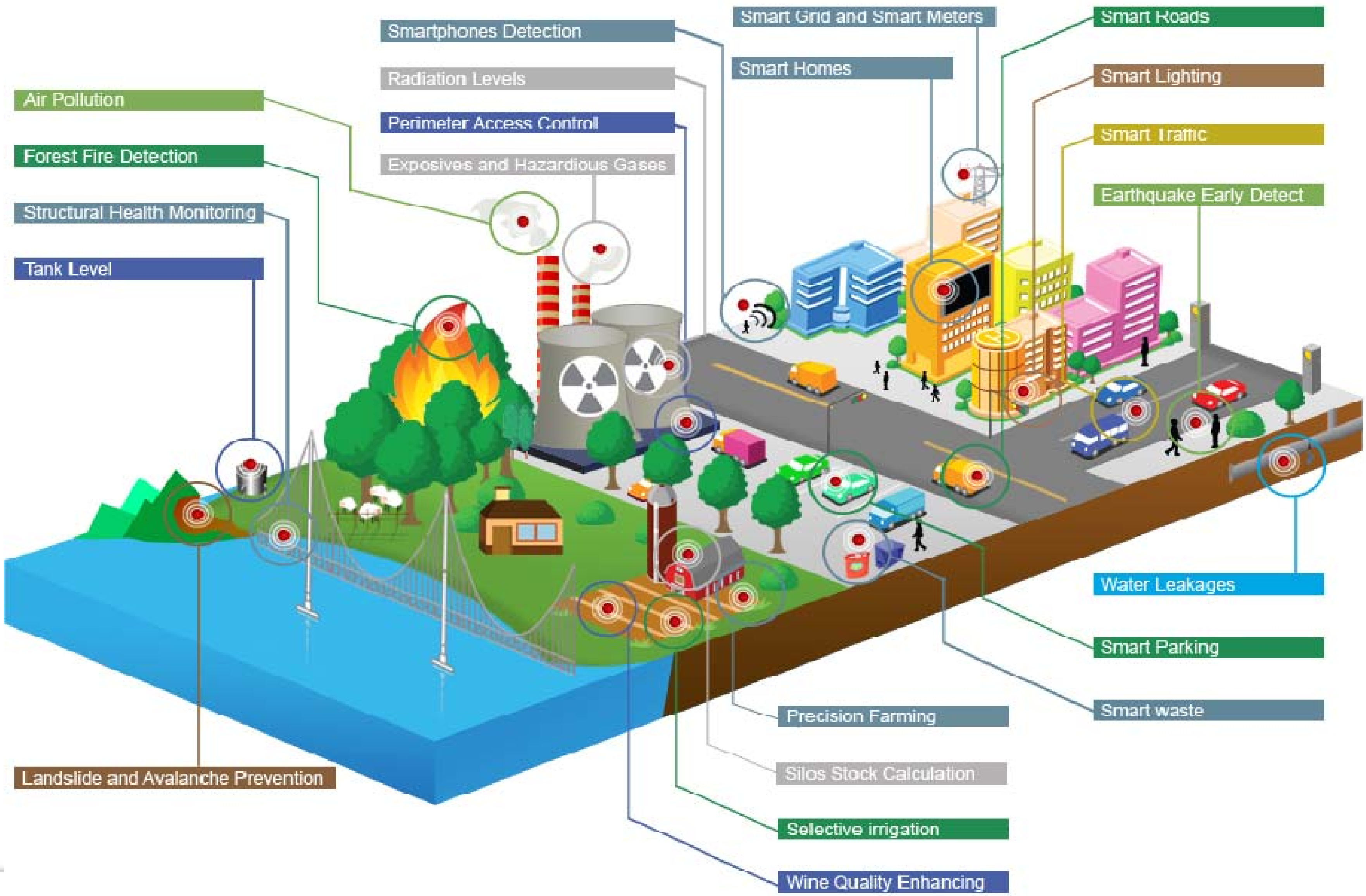
A single airplane engine generates more than 10 TB of data every 30 minutes.

Our Data-Driven world

Business & Commerce

- New York Stock Exchange **1TB** of data everyday
- Walmart’s customer transactions generate 2.5 petabytes of data every hour.

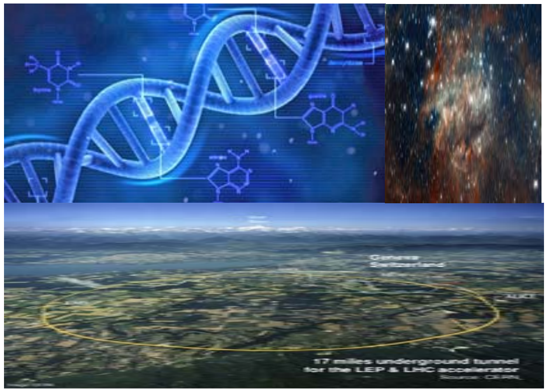
Internet of Things



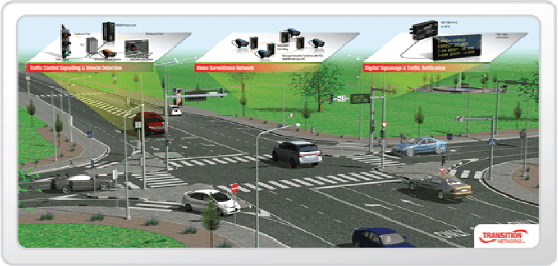
What is Big Data used for?

Big Data has the potential to *revolutionize* our lives in many ways

Scientific discoveries



Intelligent transportation



tripadvisor®
get the truth. then go.®

National and computer security

Early warning of natural disasters

Decisions

facebook

Google

AIRBUS

Diagnosis

Prevent failures

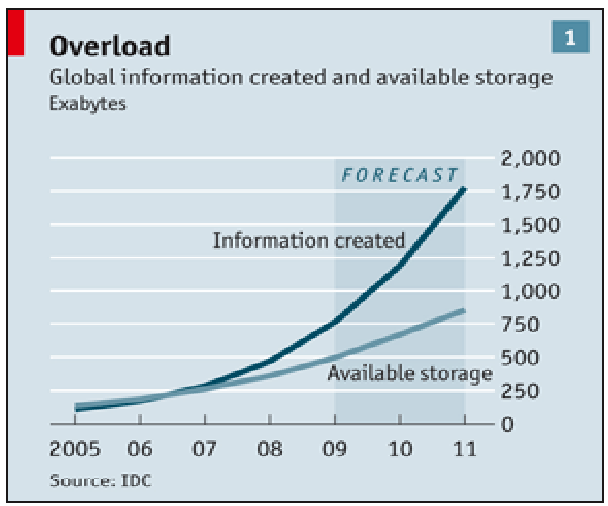
Why is the system slow?

Big Data Challenges

Big Data has the potential to *revolutionize* our lives in many ways

Volume	Variety	Velocity
8 ZettaBytes in 2015	**Many Forms**	In **60** seconds:
Big Data sizes will continue to grow at annual rate of 26.24%[1]	Structured, unstructured, Text , multimedia	694,445 search queries (G) 168 Million Emails are sent

- Storing
- Sharing
- Processing



1 IDC Study on Worldwide Big Data Technology and Services: 2014-2018 Forecast

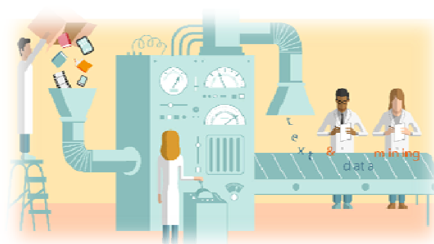
Big Data Challenges

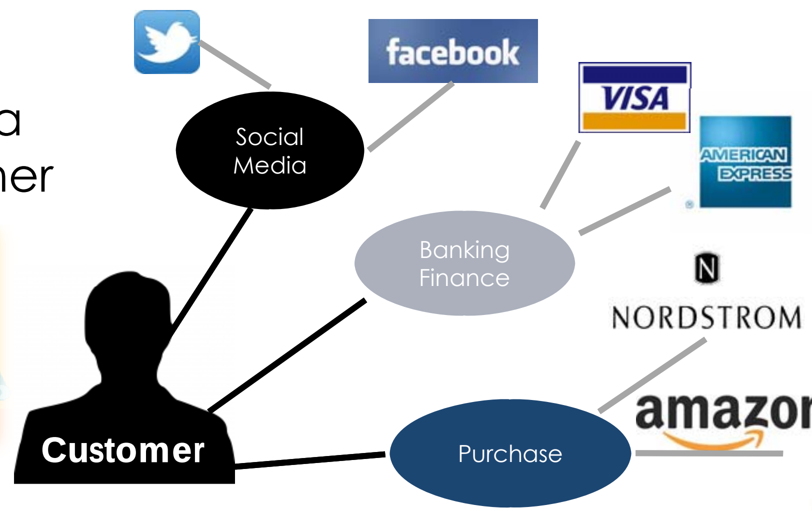
Big Data has the potential to *revolutionize* our lives in many ways

Volume	Variety	Velocity
8 ZettaBytes in 2015	**Many Forms**	In **60** seconds:
Big Data sizes will continue to grow at annual rate of 26.24%[1]	Structured, unstructured, Text , multimedia	694,445 search queries (G) 168 Million Emails are sent

- Storing
- Sharing
- Processing

- All these types of data need to linked together





1 IDC Study on Worldwide Big Data Technology and Services: 2014-2018 Forecast

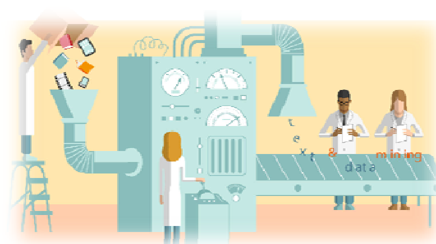
Big Data Challenges

Big Data has the potential to *revolutionize* our lives in many ways

Volume	Variety	Velocity
8 ZettaBytes in 2015	**Many Forms**	In **60** seconds:
Big Data sizes will continue to grow at annual rate of 26.24%[1]	Structured, unstructured, Text , multimedia	694,445 search queries (G) 168 Million Emails are sent

- Storing
- Sharing
- Processing

- All these types of data need to linked together



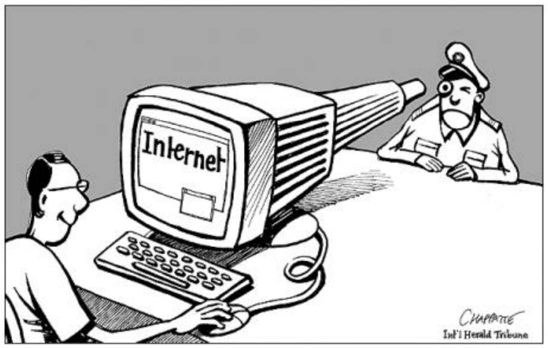
- Needs to be processed *fast*
- Product recommendations that are relevant & compelling
- Air traffic control
- Self driving cars

Late decisions ➔ missing opportunities

1 IDC Study on Worldwide Big Data Technology and Services: 2014-2018 Forecast

… and Privacy

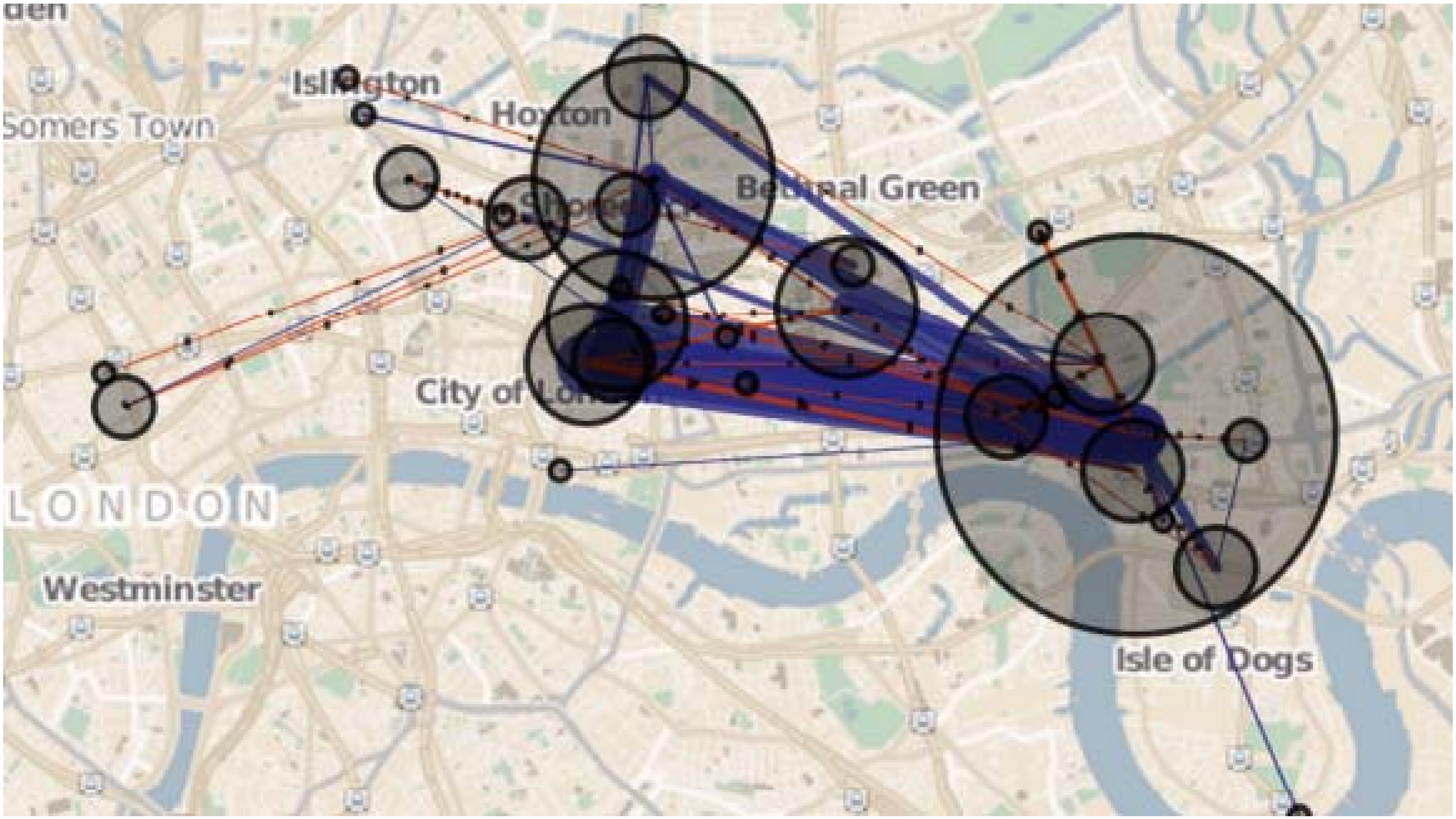

PRIVACY
PLEASE


Internet


DATA COLLECTION LISTS
CHEAP!
SOC. SEC. NUMBERS
DRIVER'S LICENSE NUMBERS
HOME OWNERSHIP
INCOME
MARITAL STATUS
INTERNET HABITS
AUTO
FAMILY STATUS
MEDICAL HISTORY
BANK RECORDS
INSURANCE RECORDS
FENCING STOLEN GOODS
KIRK


DOGBERT CONSULTS
CUSTOMER DATA IS AN ASSET THAT YOU CAN SELL.
Dilbert.com DilbertCartoonist@gmail.com
IT'S TOTALLY ETHICAL BECAUSE OUR CUSTOMERS WOULD DO THE SAME THING TO US IF THEY COULD.
10-13-10 © 2010 Scott Adams, Inc./Dist. by UFS, Inc.
SOUNDS FAIR.
IN PHASE ONE, WE'LL DEHUMANIZE THE ENEMY BY CALLING THEM "DATA."

Goodbye Anonymity



Turn Big Data into Big Value

Big Data opportunities?

- Programming model
  - *MapReduce: Simple yet scalable model*

Google hadoop Dryad

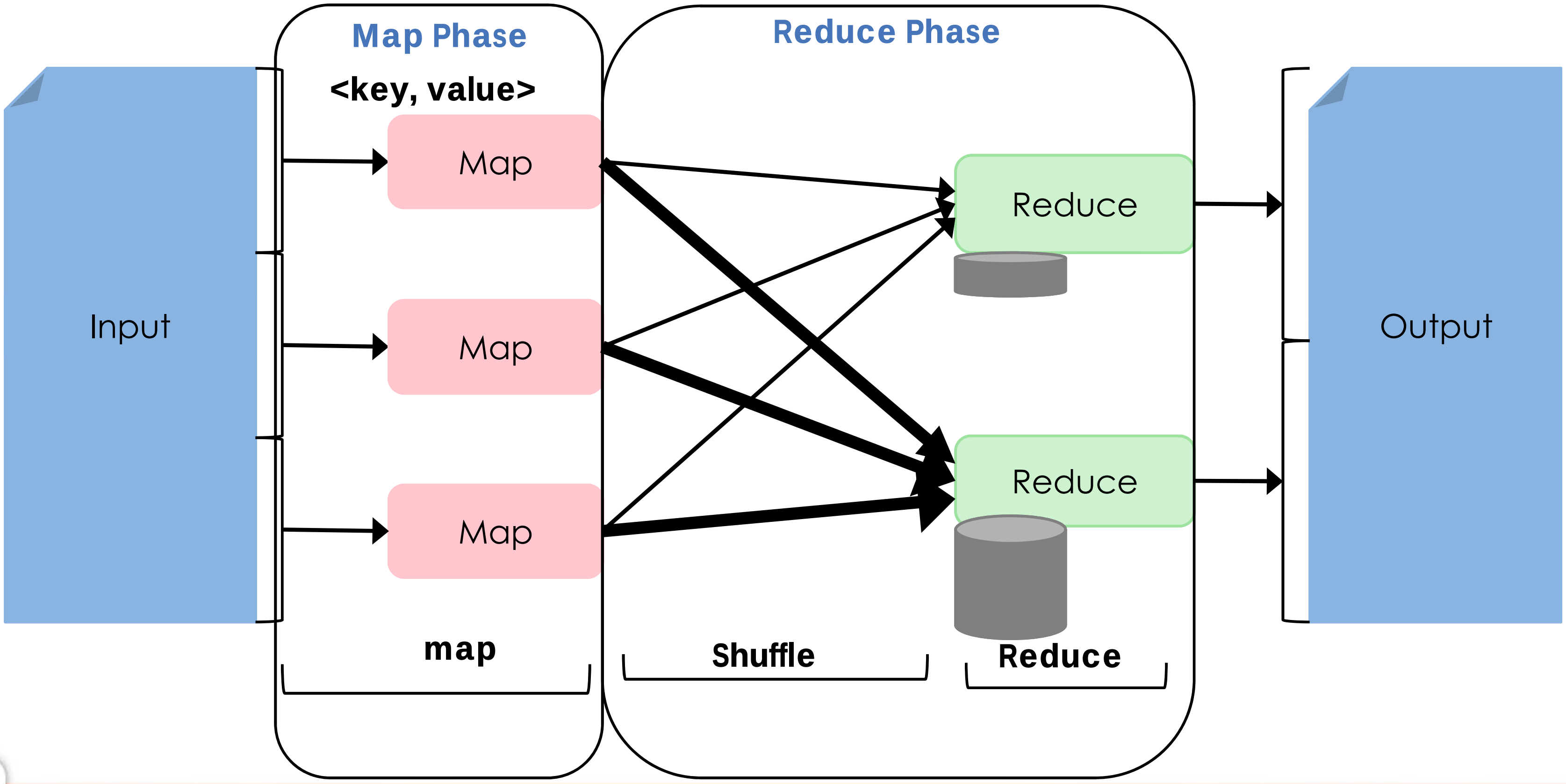
- Available computation/storage power
  - *Cloud computing: Allows users to lease computing and storage resources in a Pay-As-You-Go manner*

Windows Azure amazon webservices™

Data is ONLY as useful as the decisions it enables

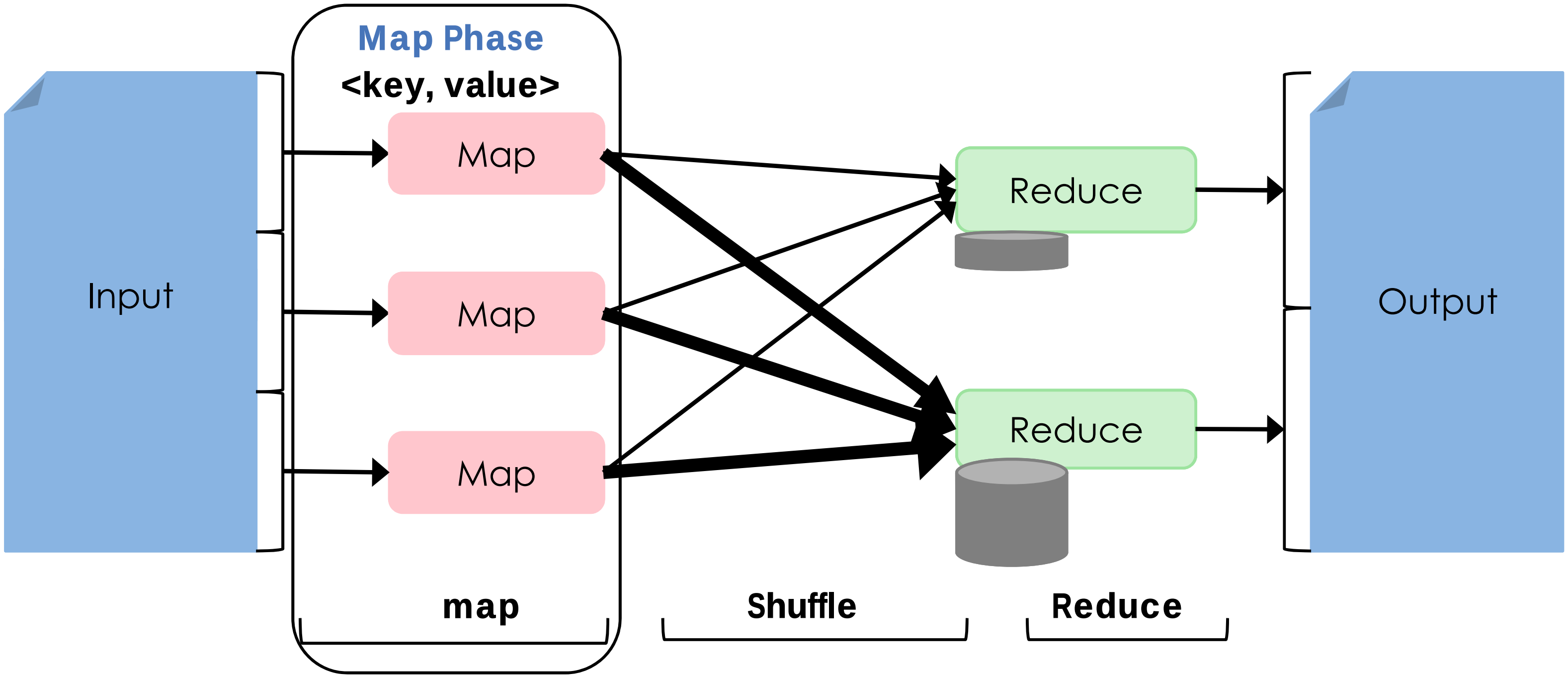
MapReduce System

MapReduce is a programming model and an associated implementation for processing and generating large data sets. *--- OSDI 2004 --*



MapReduce Model

MapReduce is a programming model and an associated implementation for processing and generating large data sets. *--- OSDI 2004 --*

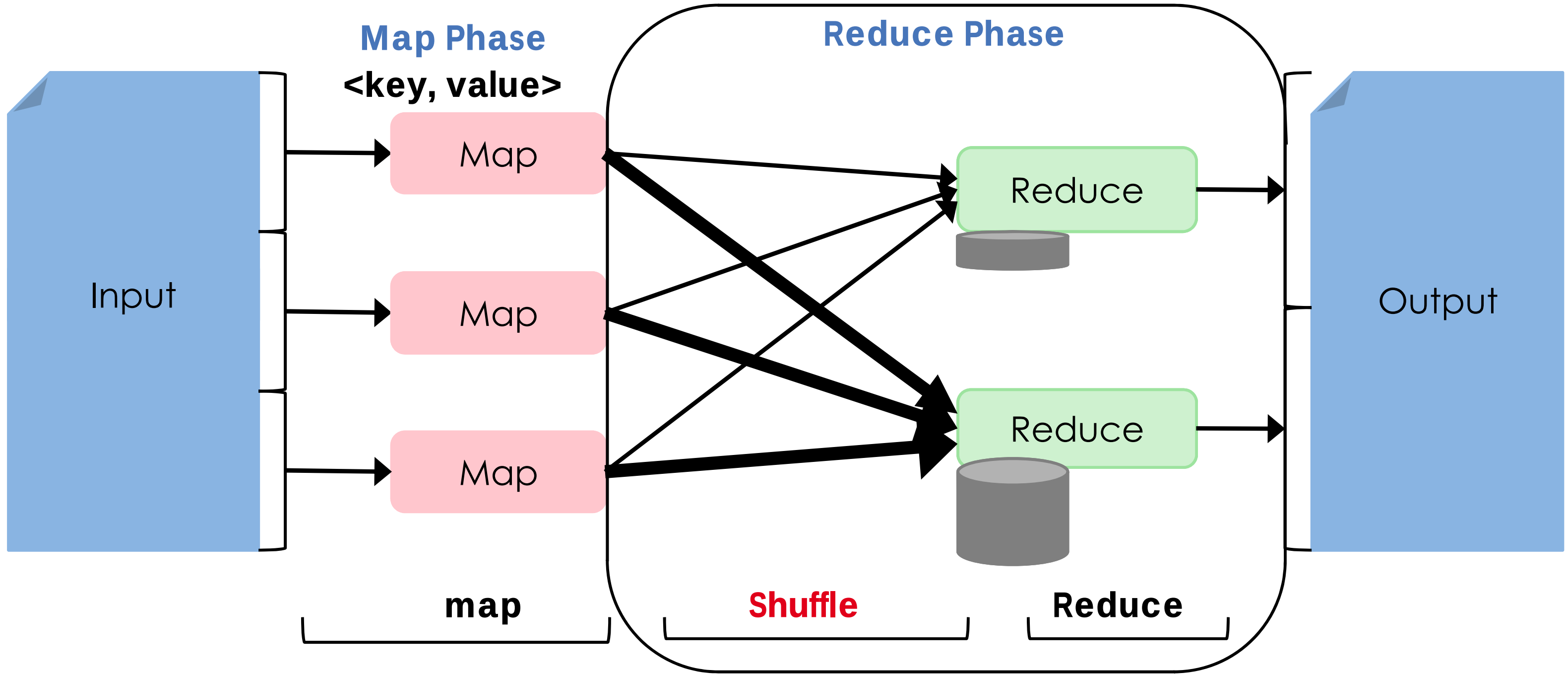


map(k, v) → <k', v'>*

Records from the data source (lines out of files, rows of a database, etc) are fed into the map function as key*value pairs: e.g., (filename, line)

MapReduce Model

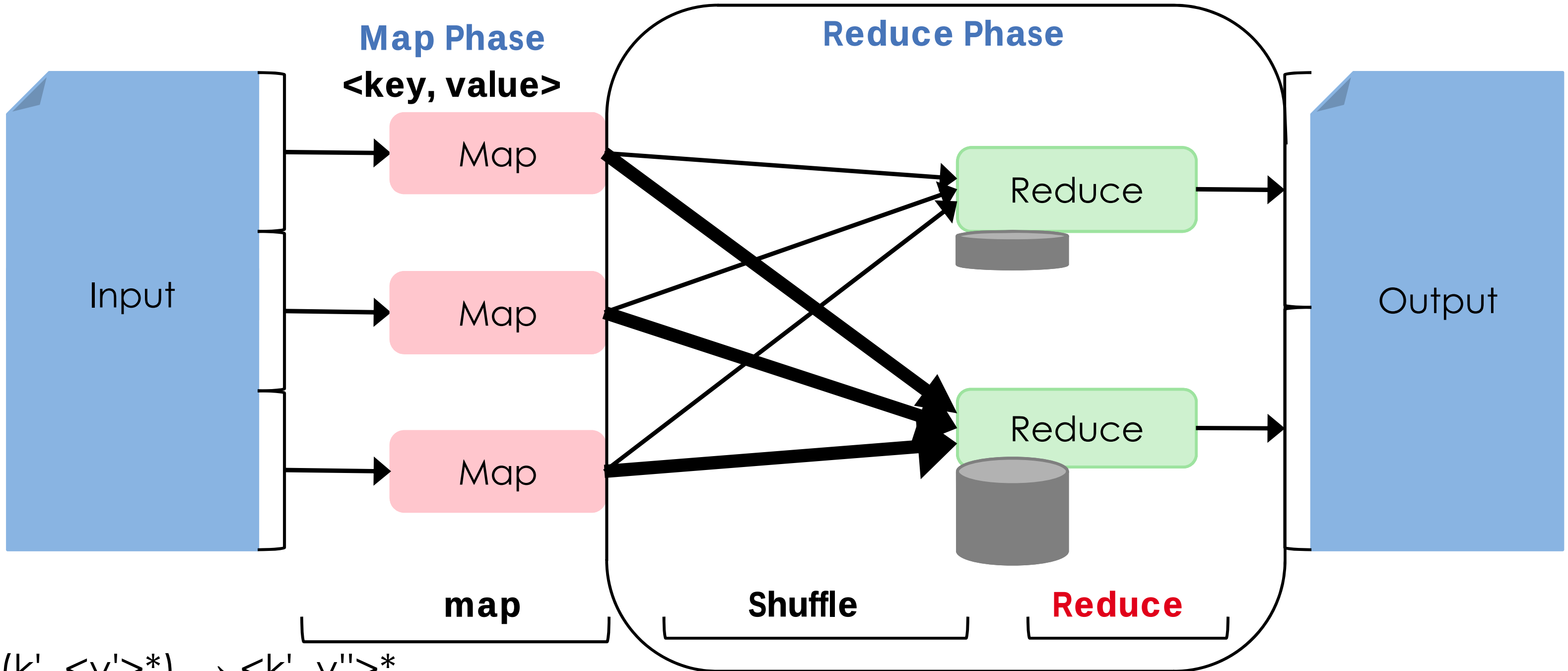
MapReduce is a programming model and an associated implementation for processing and generating large data sets. *--- OSDI 2004 --*



After the map phase is over, all the intermediate values for a given output key are combined together into a list

MapReduce Model

MapReduce is a programming model and an associated implementation for processing and generating large data sets. *--- OSDI 2004 --*



reduce(k', <v'>*) → <k', v''>*

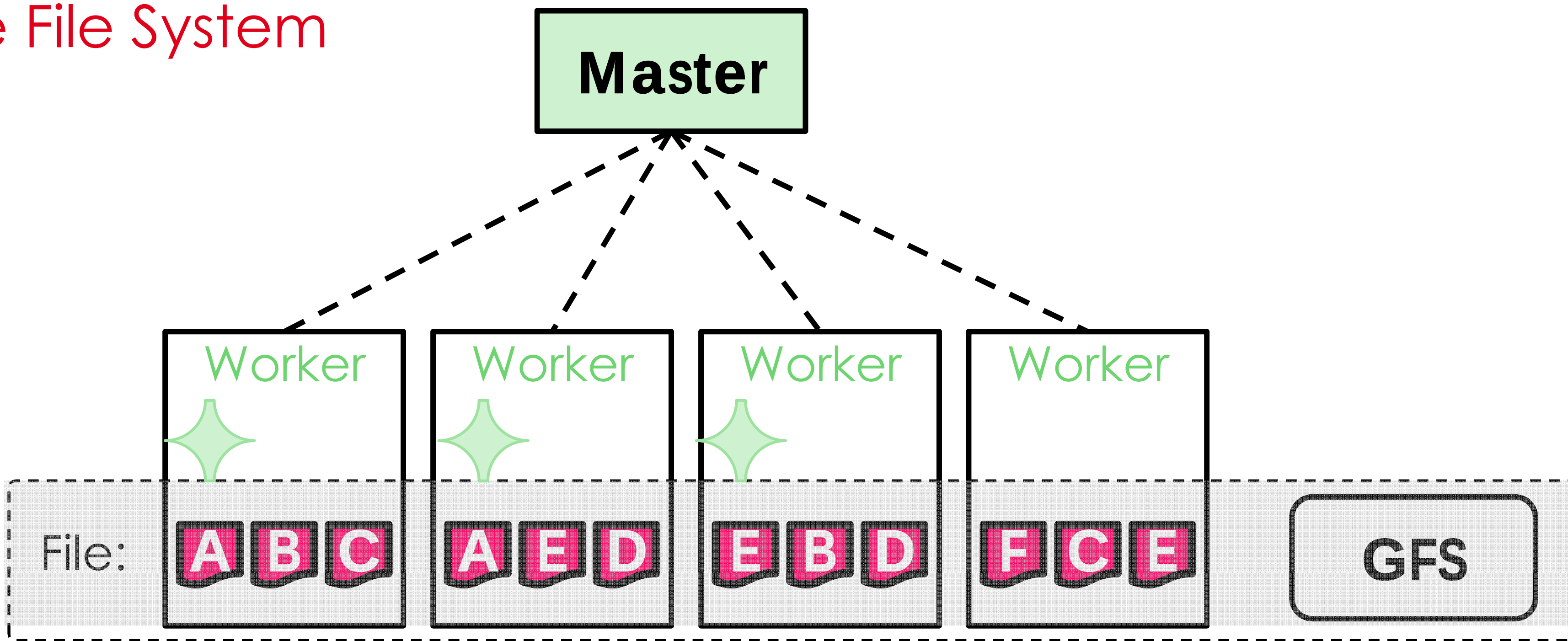
reduce() combines those intermediate values into one or more final values per key (usually only one)

MapReduce System

MapReduce is a programming model and an associated implementation for processing and generating large data sets. *--- OSDI 2004 --*

MapReduce: Runtime Environment

Google File System



MapReduce System

MapReduce is a programming model and an associated implementation for processing and generating large data sets. *--- OSDI 2004 --*

MapReduce: Runtime Environment

MapReduce Framework runs on the top of Google distributed File system (GFS)

- Files are divided into blocks (64MB)
- Blocks are replicated and distributed on many machines
  - Optimize read and write operations
  - Fault-tolerance

MR: Main features

MapReduce is a programming model and an associated implementation for processing and generating large data sets. *--- OSDI 2004 --*

MapReduce: Runtime Environment

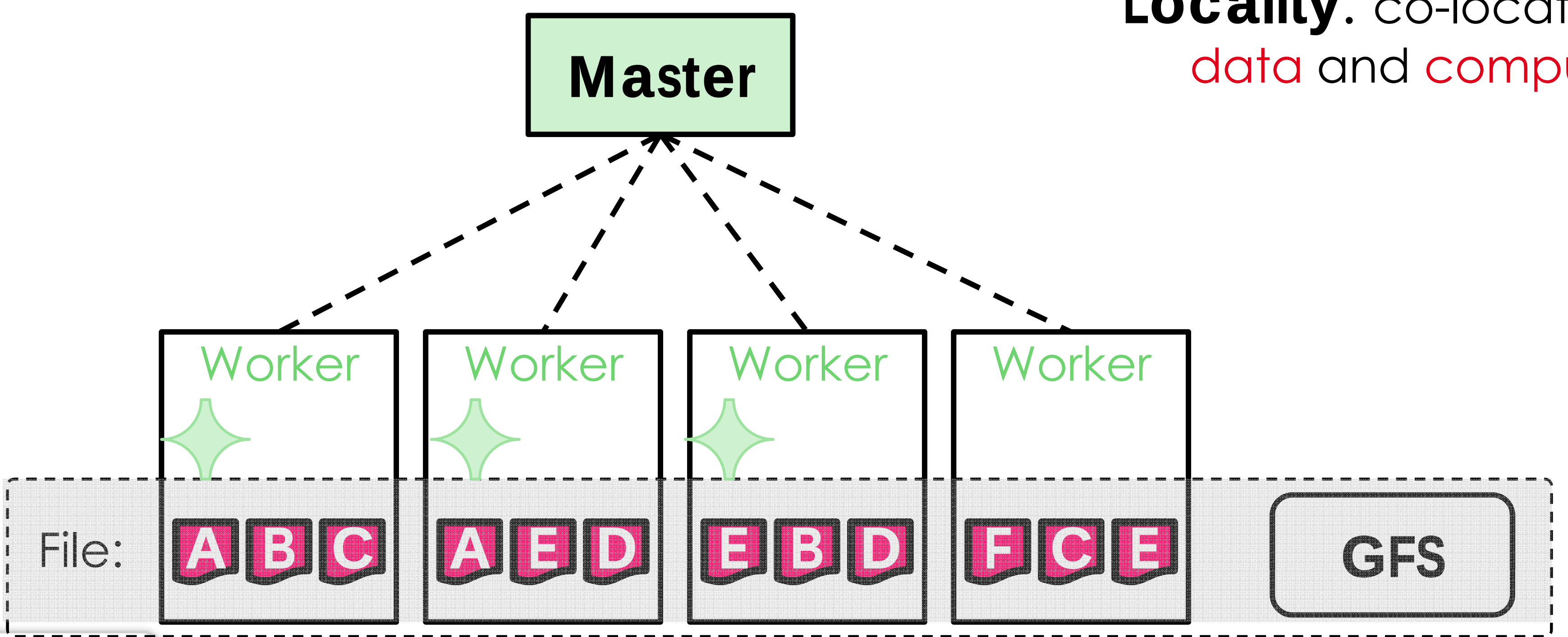
- Data locality
  - Co-locating data and computation
- Fault-tolerance
  - Re-execute tasks belonging to Failed node on another node
- Exploit Parallelism
  - Running independent Map (blocks) and Reduce (keys)

MR: Main features

MapReduce is a programming model and an associated implementation for processing and generating large data sets. *--- OSDI 2004 --*

MapReduce: Runtime Environment

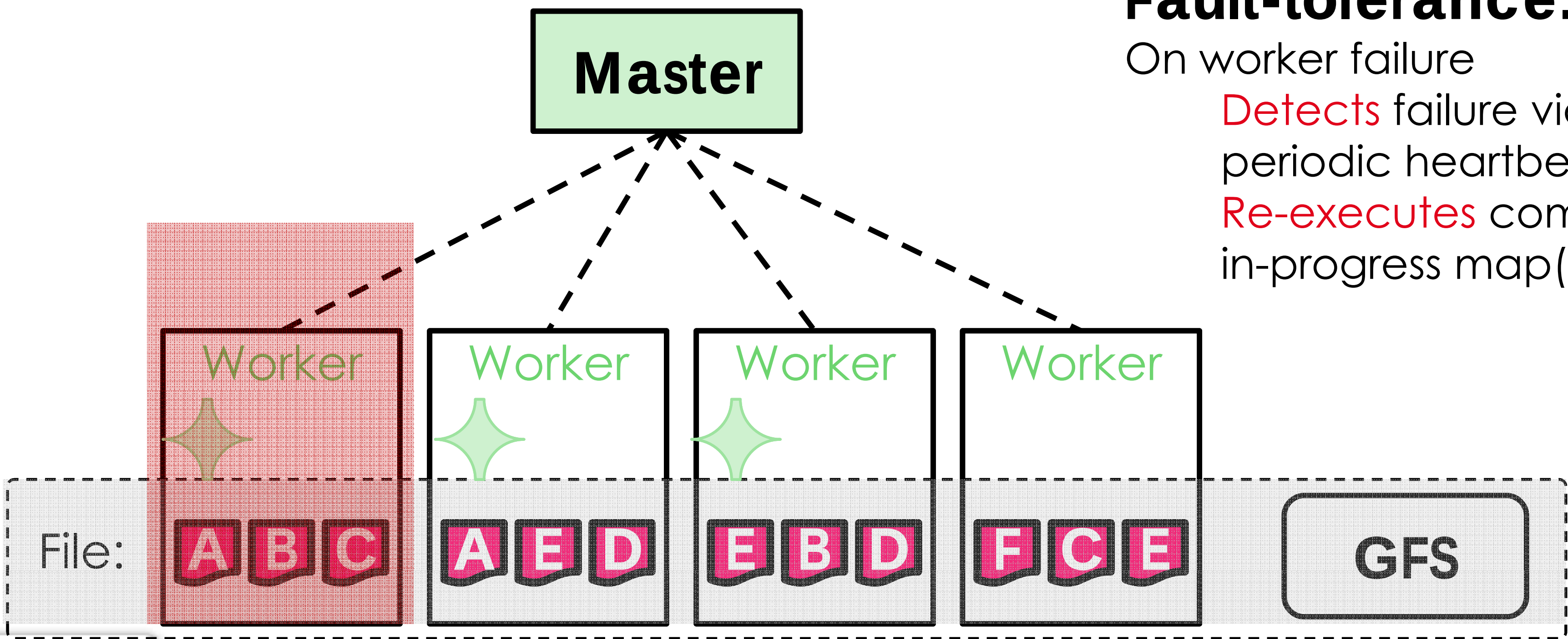
Locality: co-locating data and computation



MR: Main features

MapReduce is a programming model and an associated implementation for processing and generating large data sets. *--- OSDI 2004 --*

MapReduce: Runtime Environment



Fault-tolerance:

On worker failure

Detects failure via periodic heartbeats

Re-executes completed & in-progress map() tasks

MR: Main features

MapReduce is a programming model and an associated implementation for processing and generating large data sets. *--- OSDI 2004 --*

MapReduce: Runtime Environment

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

Actual Google MapReduce

Example is written in pseudo-code

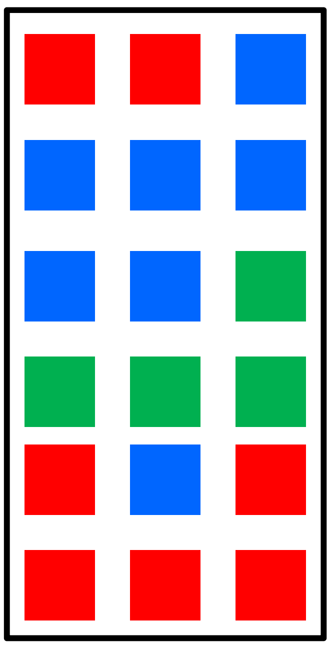
Actual implementation is in C++, using a MapReduce library

Bindings for Python and Java exist via interfaces

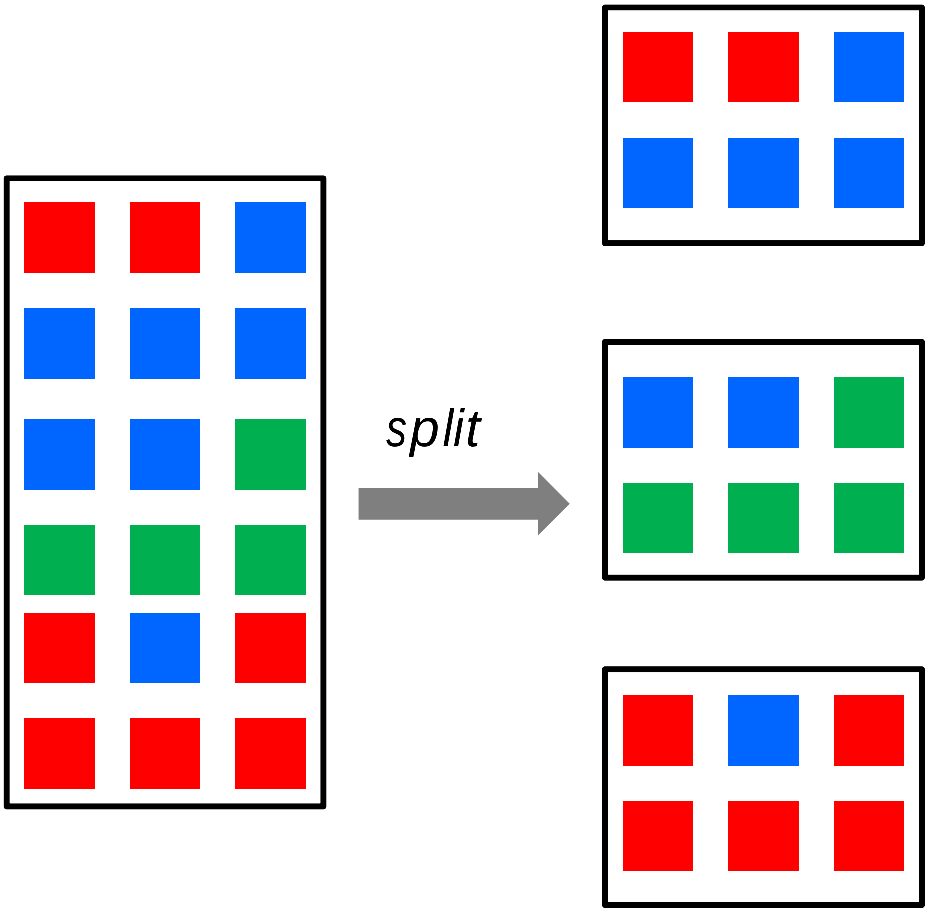
True code is somewhat more involved (defines how the input key/values are divided up and accessed, etc.)

"Google Infrastructure for Massive Parallel Processing", Walfredo Cirne, Presentation in the industrial track in CCGrid'2007.

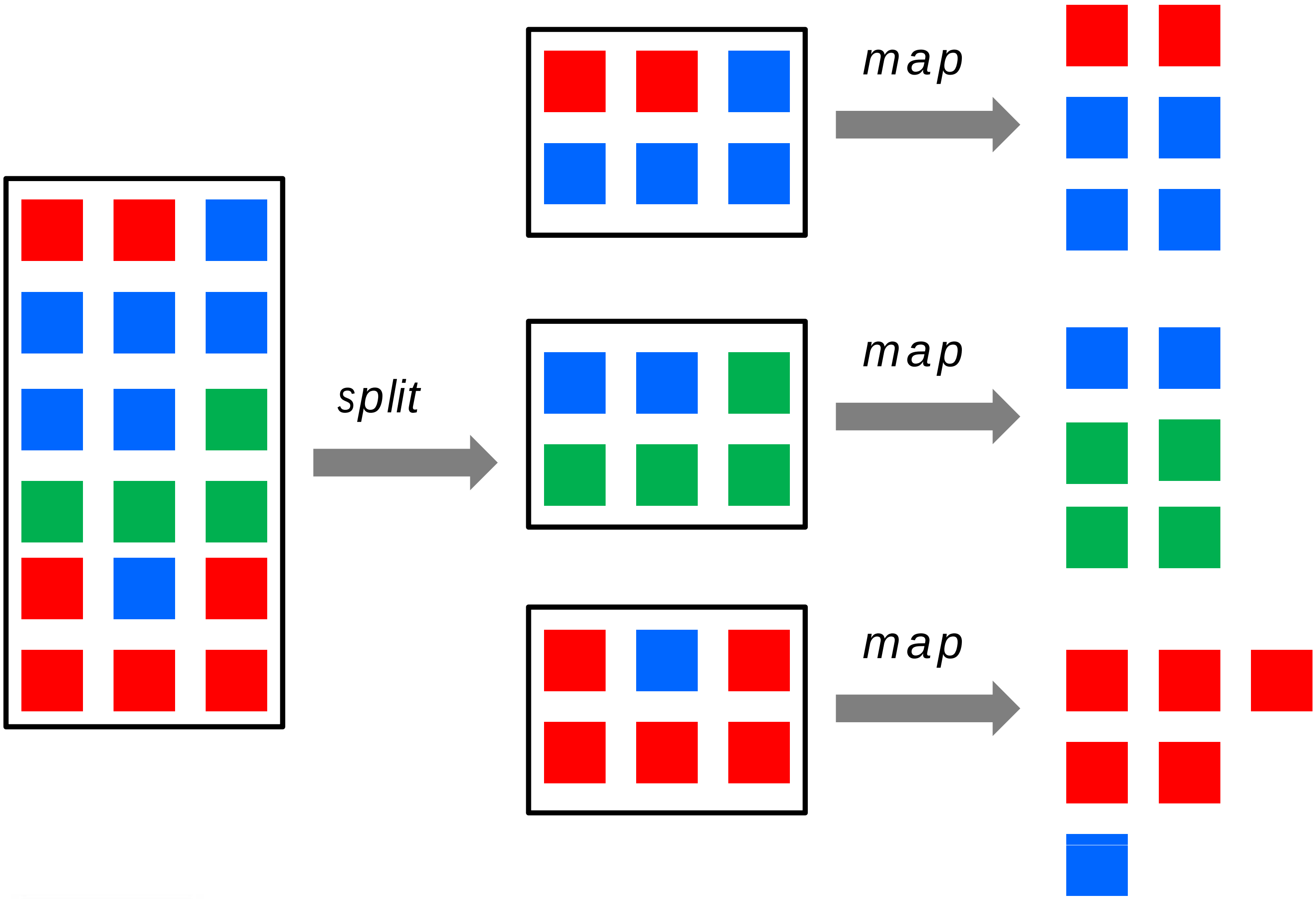
The Colored Squares Counter



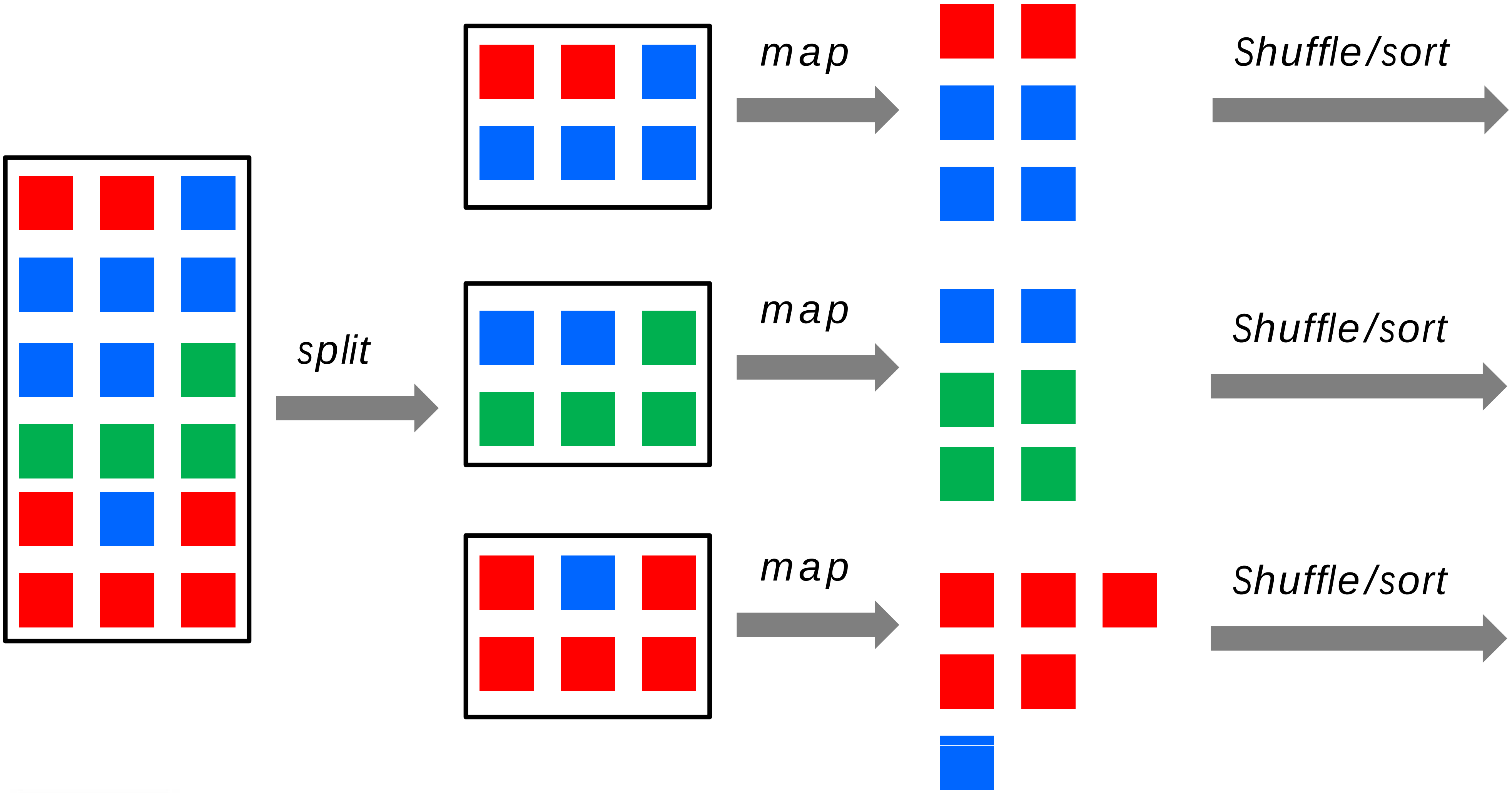
The Colored Squares Counter



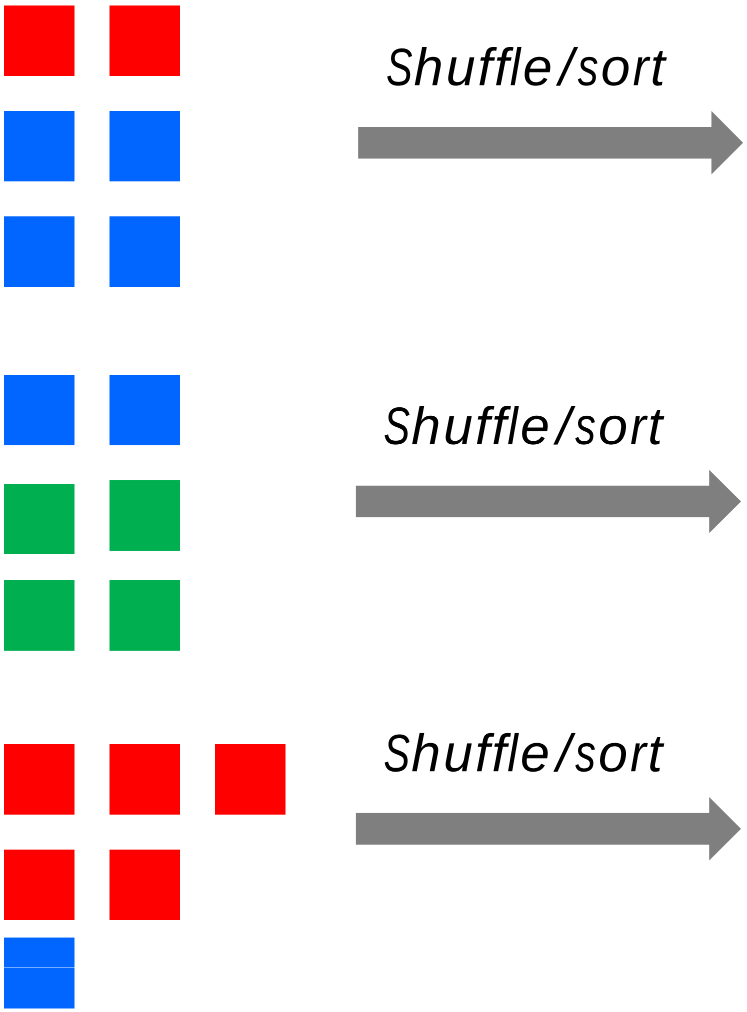
The Colored Squares Counter



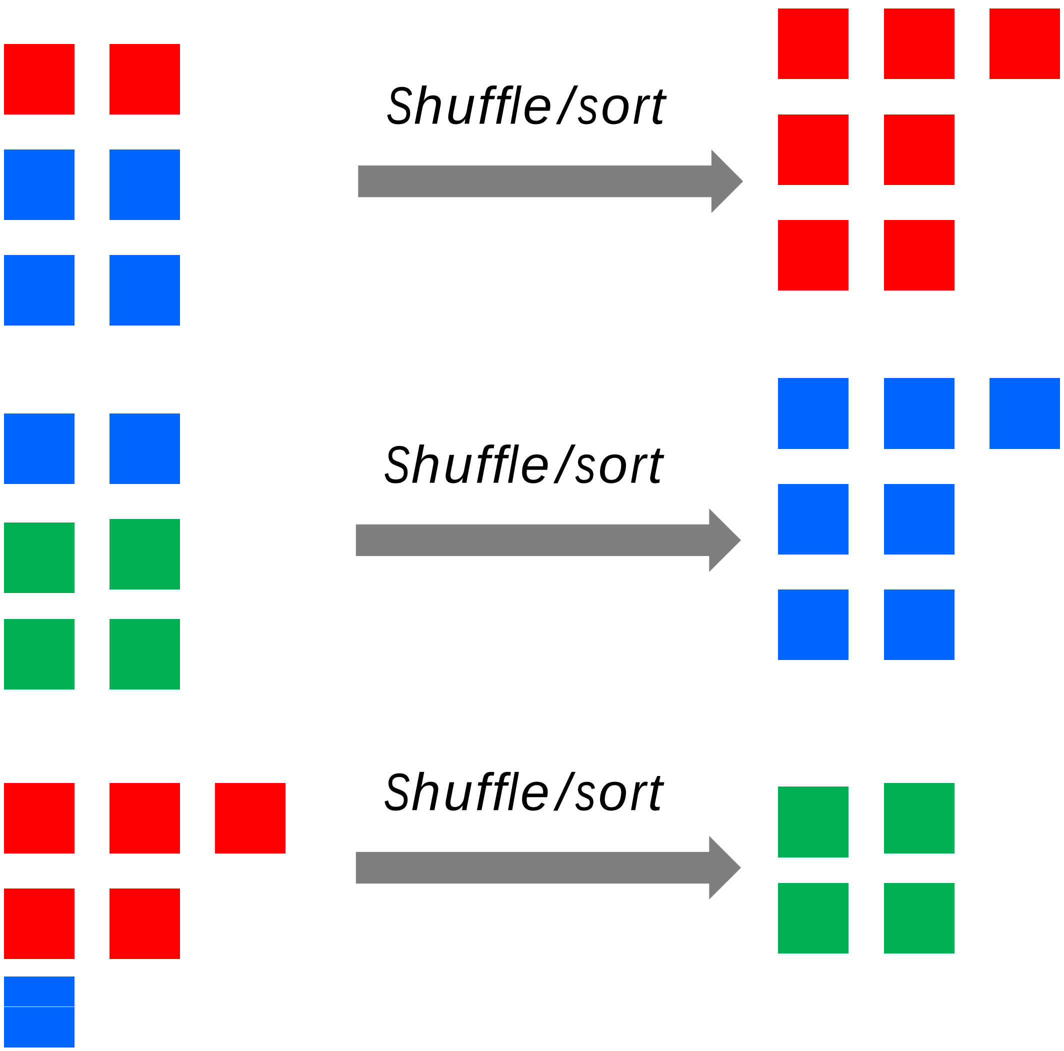
The Colored Squares Counter



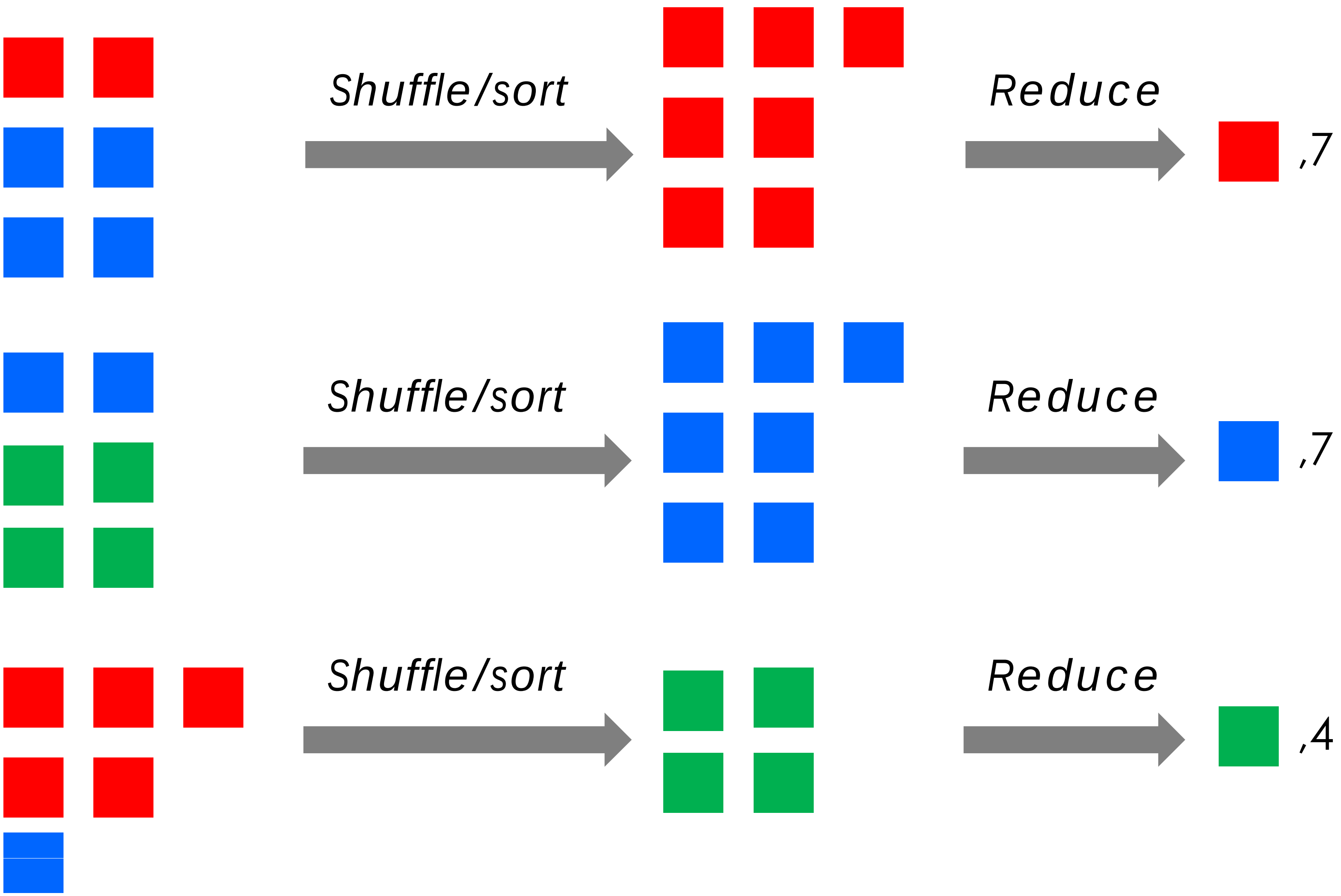
The Colored Squares Counter



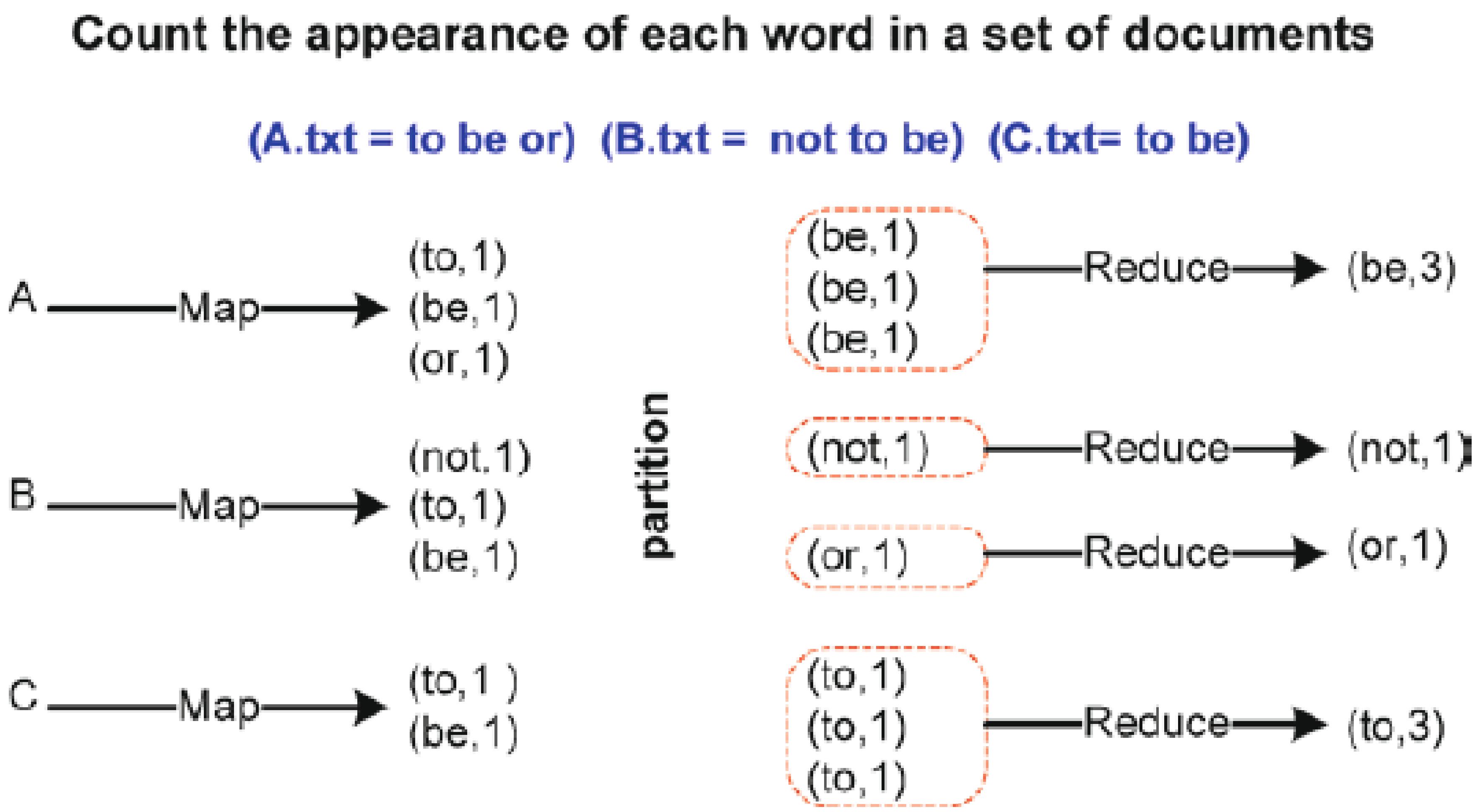
The Colored Squares Counter

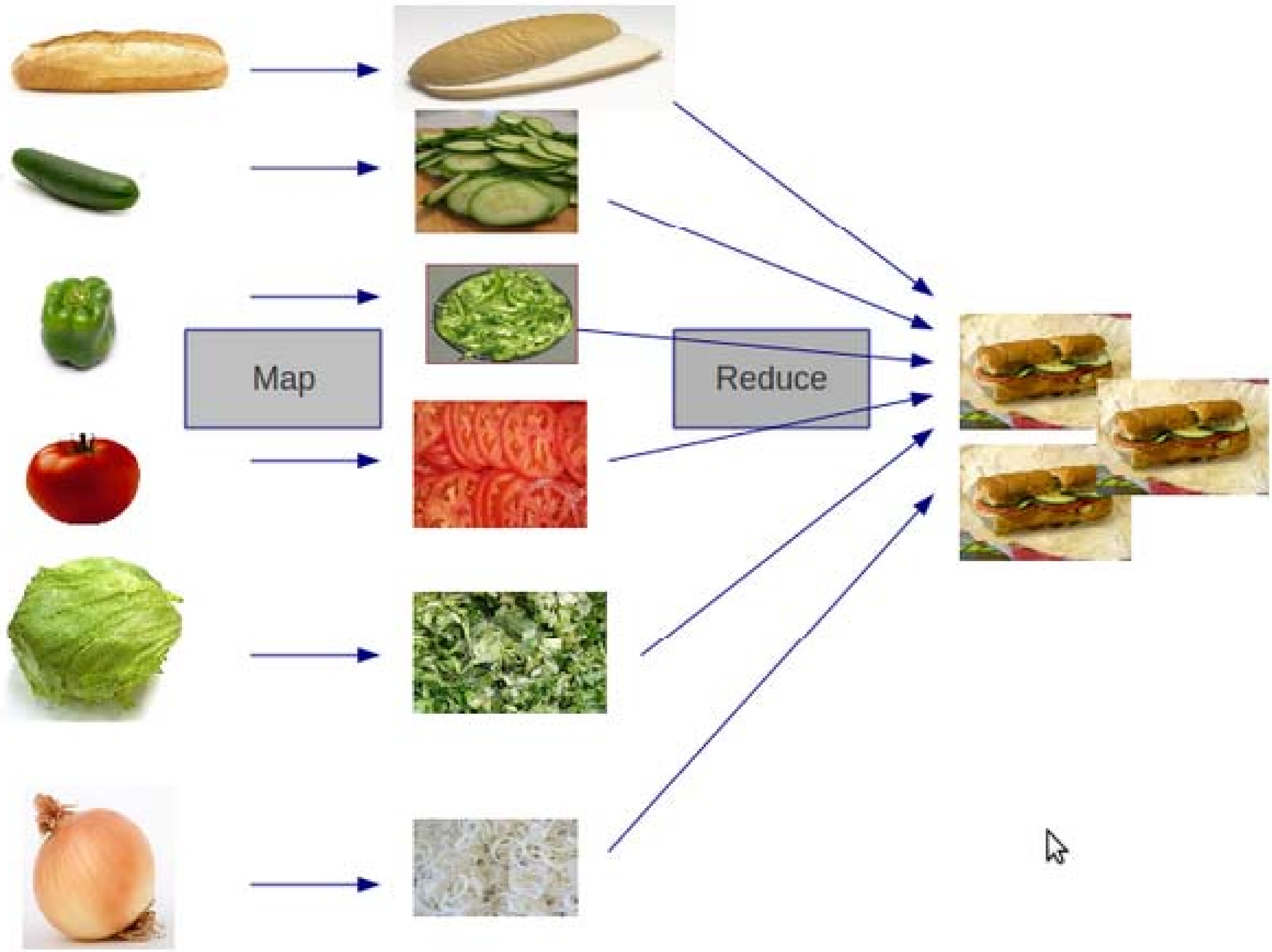


The Colored Squares Counter

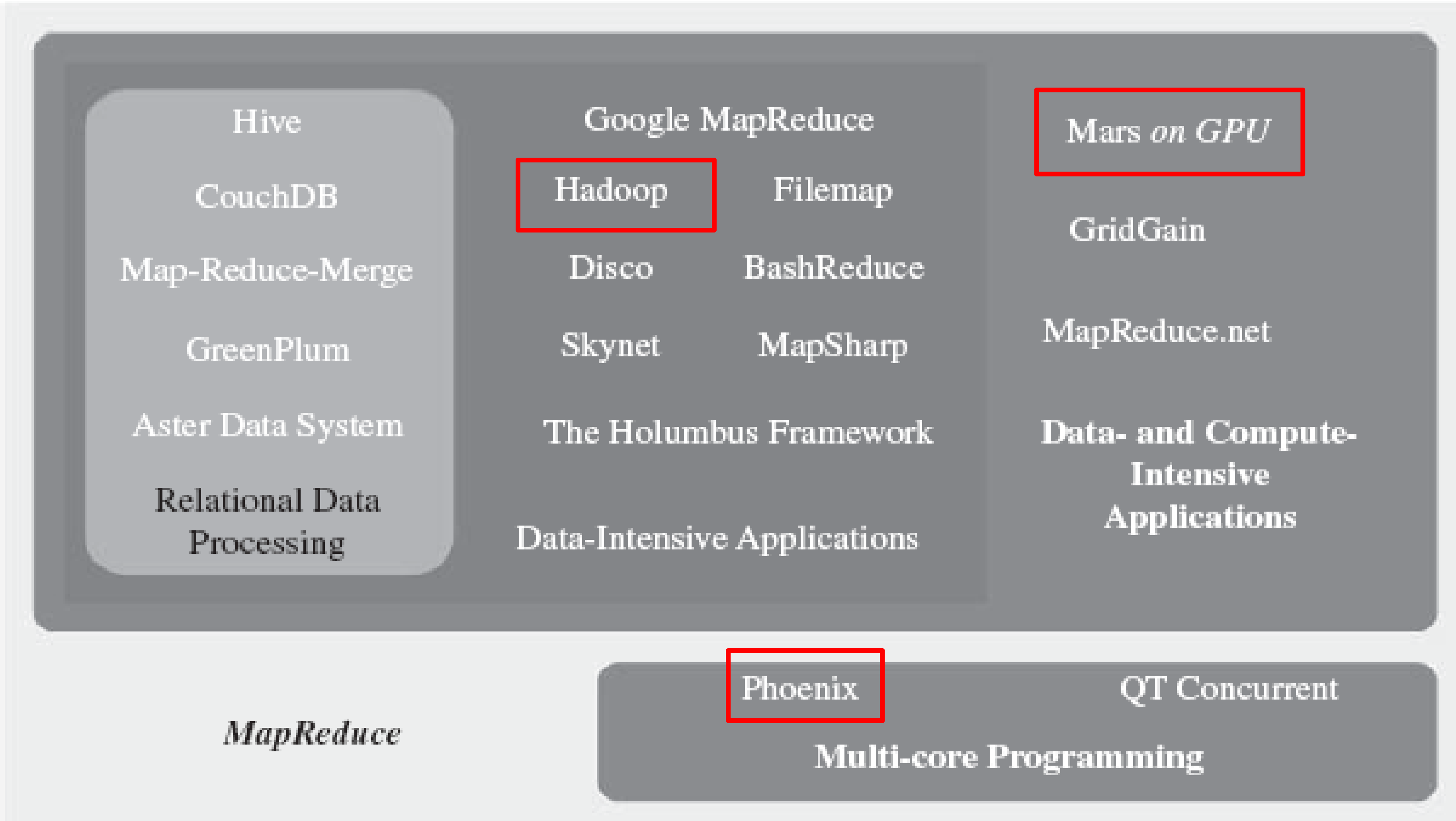


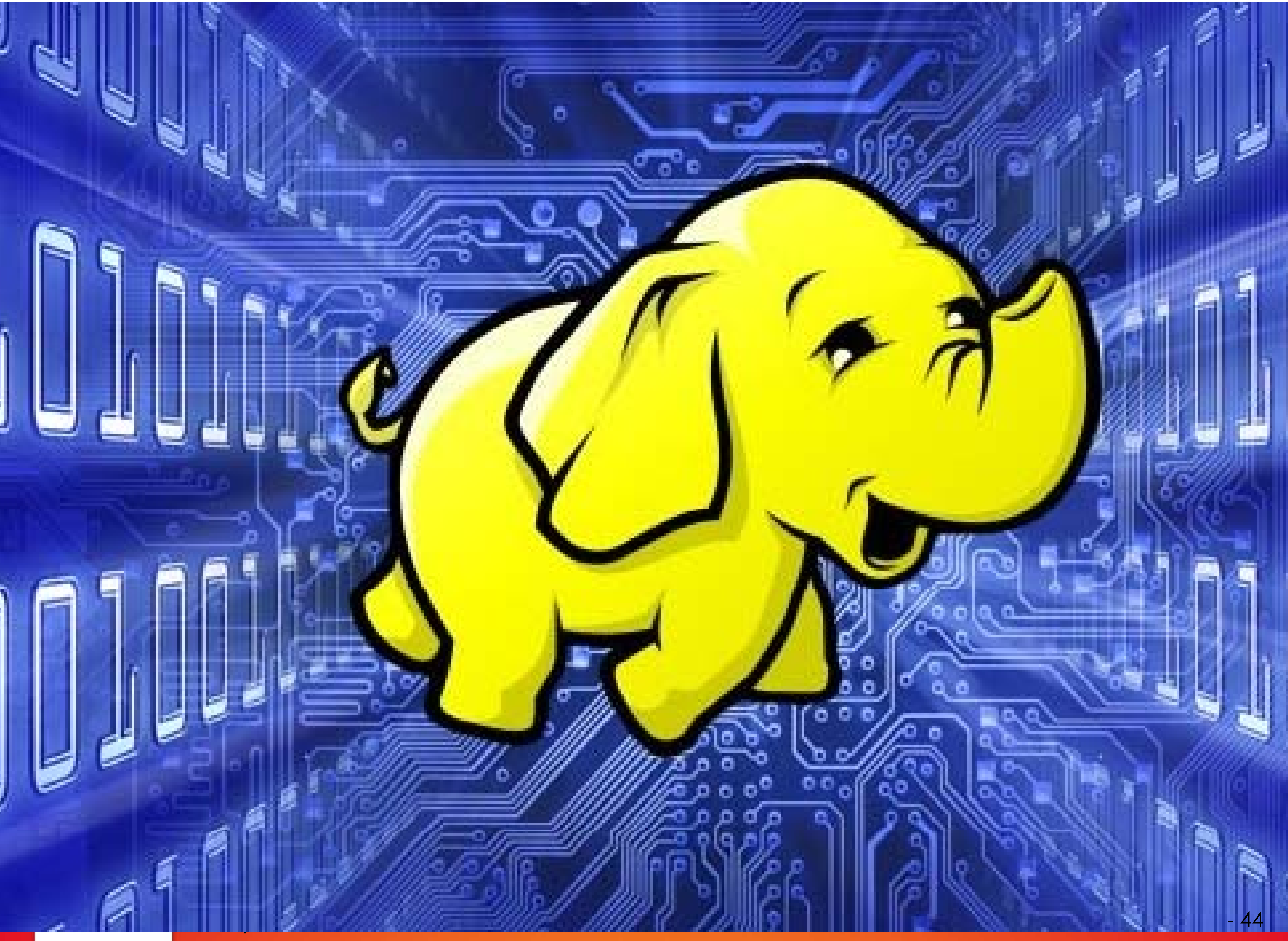
Word Count Example




Map
Reduce

MapReduce: Implementations





TODAY: the reign of Hadoop!

Hadoop is a top-level Apache project

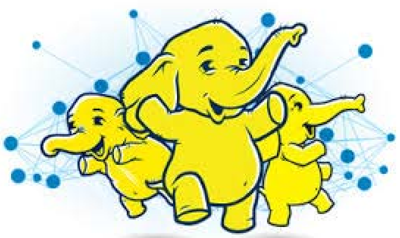
Hadoop is advocated by industry's premier Web players -- Google, Yahoo!, Microsoft, and Facebook-- *as the engine to power the cloud.*

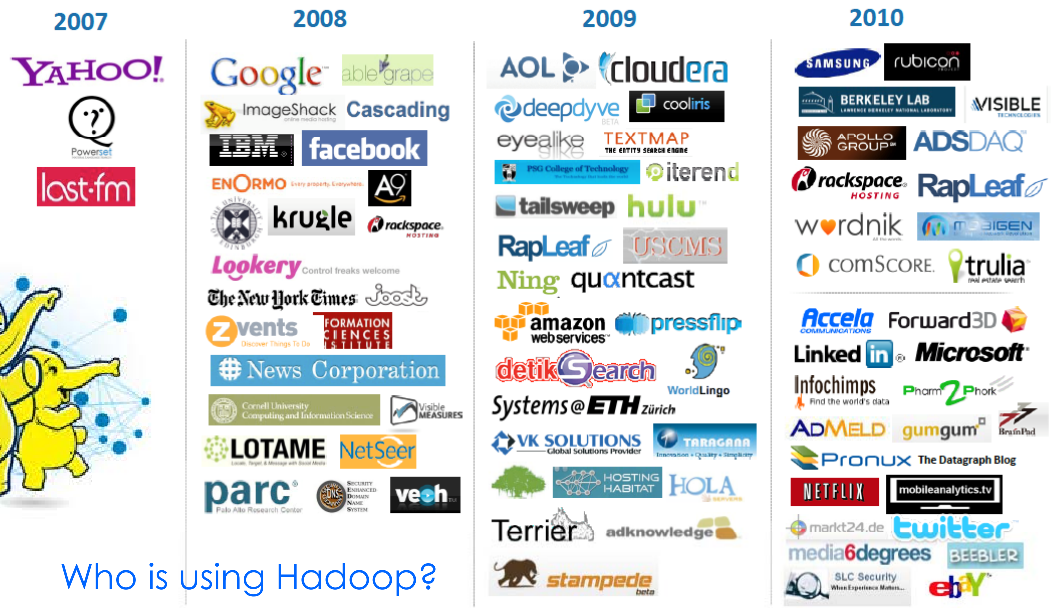
Web Indexing

Spam Detection

Smart Advertising

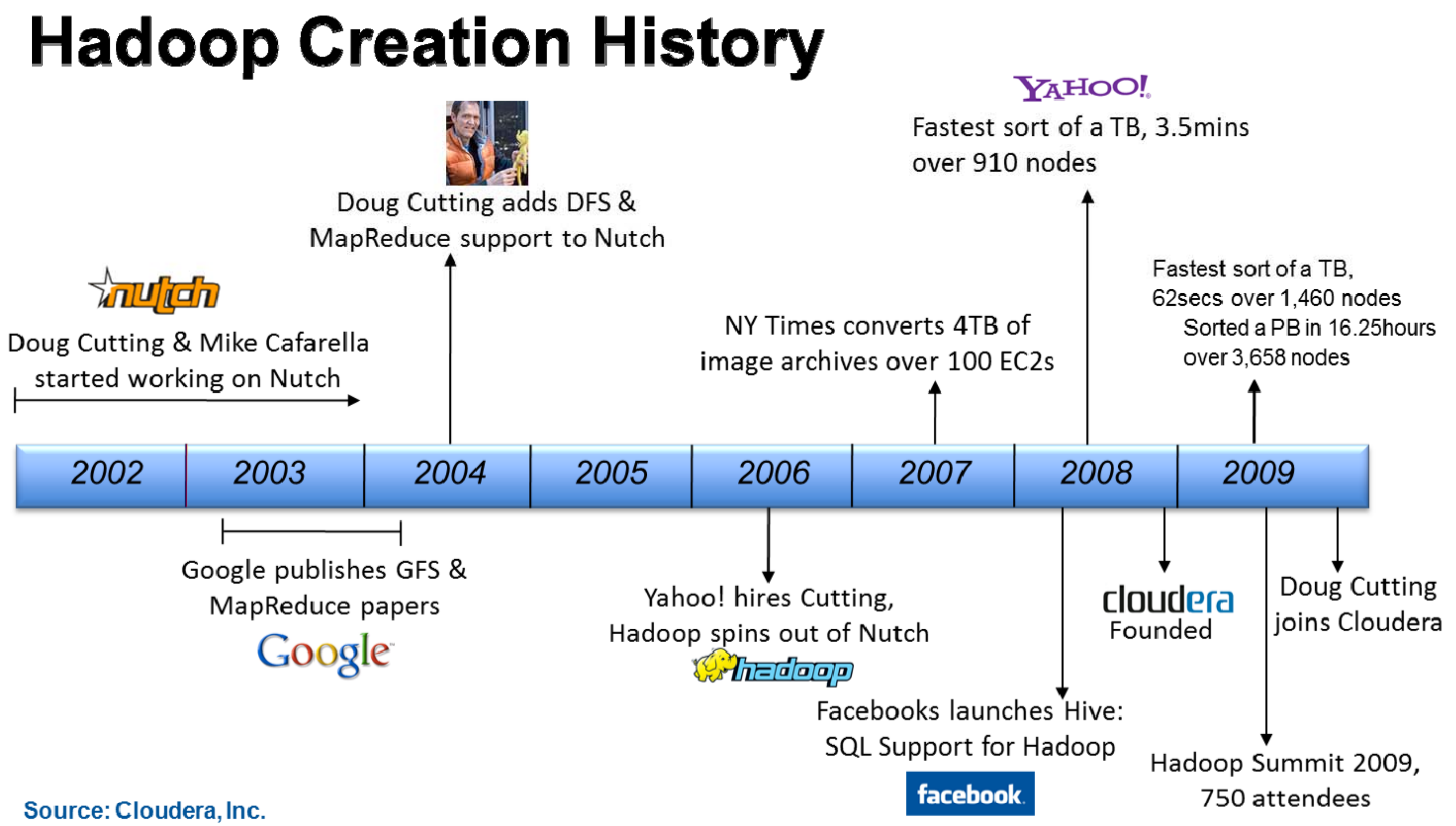
Market Analysis





Who is using Hadoop?

Hadoop History



NYTimes articles from 1851 -1922 available to the public online. There was a total of 11 million articles. They had to take sometimes several TIFF images and scale and glue them together to create one PDF version of the article. They used 100 EC2 instances to complete the job in just under 24 hours. They started with 4TB of data that was uploaded into S3 and through the conversion process created another 1.5TB.

Who is not Using Hadoop??

hadoop

2007

- Yahoo!
- Powerset
- last.fm

2008

- Google
- ablegrape
- ImageShack
- Cascading
- IBM
- facebook
- ENORMO
- A9
- University of Edinburgh
- krugle
- rackspace hosting
- Lookery
- The New York Times
- Joost
- Zvents
- Information Sciences Institute
- News Corporation
- Cornell University Computing and Information Science
- Visible Measures
- LOTAME
- NetSeer
- parc Palo Alto Research Center
- DNS Security Enhanced Domain Name System
- veoh

2009

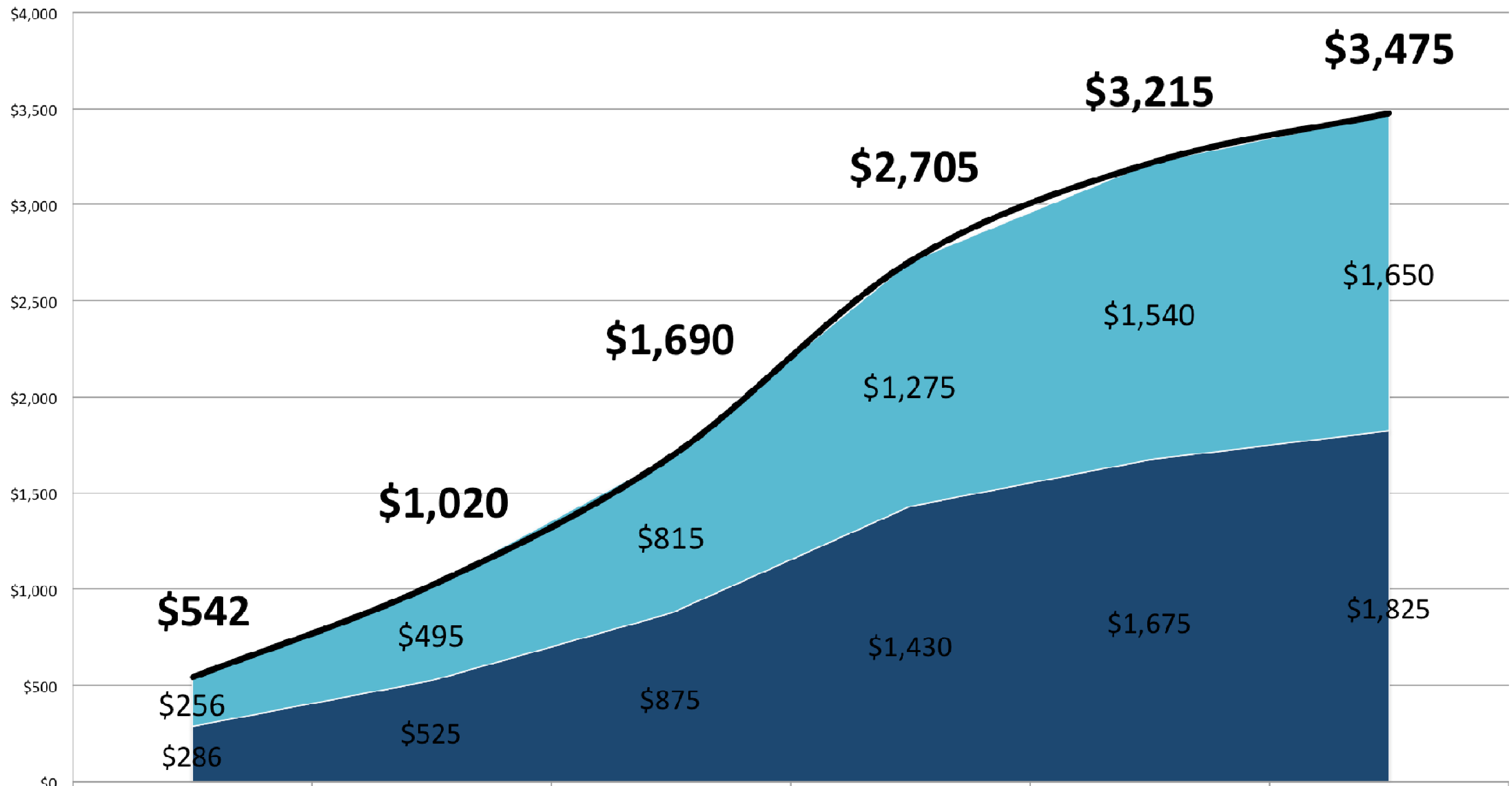
- AOL
- cloudera
- deepdyve
- cooliris
- eyealike
- TEXTMAP The entity search engine
- PSG College of Technology
- iterend
- tailsweep
- hulu
- RapLeaf
- USCMS
- Ning
- quantcast
- amazon webservices
- pressflip
- detikSearch
- WorldLingo
- Systems@ETH Zürich
- VK SOLUTIONS Global Solutions Provider
- TARAGANA
- HOSTING HABITAT
- HOLA servers
- Terrier
- adknowledge
- stampede beta

2010

- SAMSUNG
- rubicon project
- BERKELEY LAB
- VISIBLE TECHNOLOGIES
- APOLLO GROUP
- ADSDAQ
- rackspace hosting
- RapLeaf
- wordnik
- MOBIGEN
- comScore
- trulia real estate search
- Accela Communications
- Forward3D
- LinkedIn
- Microsoft
- Infochimps Find the world's data
- Pharm2Phork
- ADMELD
- gumgum
- BrainPad
- Pronux
- The Datagraph Blog
- NETFLIX
- mobileanalytics.tv
- markt24.de
- twitter
- media6degrees
- BEEBLER
- SLC Security
- ebay

Source: Hadoop Summit Presentations

Hadoop & NoSQL Software/Services Revenue Projection, 2012-2017 ($M)



	2012	2013	2014	2015	2016	2017
Hadoop	$256	$495	$815	$1,275	$1,540	$1,650
NoSQL	$286	$525	$875	$1,430	$1,675	$1,825
Total	$542	$1,020	$1,690	$2,705	$3,215	$3,475

Source: Wikibon 2013

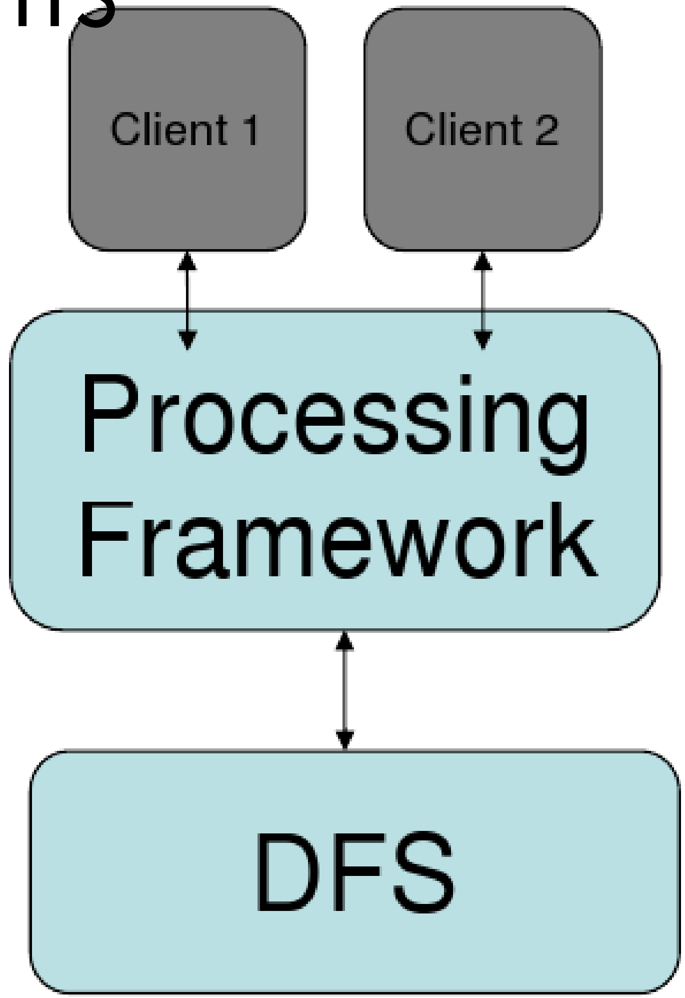
Hadoop Components

Distributed File System

- Name: HDFS
- Inspired by GFS

Distributed Processing Framework

- Modelled on the MapReduce abstraction



http://wiki.apache.org/hadoop/HadoopPresentations

HDFS Architecture: NameNode

Master-worker Architecture
HDFS Master “NameNode”

- Manages all file system metadata in memory
  - List of files
  - For each file name, a set of blocks
  - For each block, a set of DataNodes
  - File attributes (creation time, replication factor)
- Controls read/write access to files
- Manages block replication
- Transaction log: register file creation, deletion, etc.

http://wiki.apache.org/hadoop/HadoopPresentations

HDFS Architecture: DataNodes

HDFS servers "DataNodes"

A DataNode is a block server

- Stores data in the local file system (e.g. ext3)
- Stores meta-data of a block (e.g. CRC)
- Serves data and meta-data to Clients

Block Report

- Periodically sends a report of all existing blocks to the NameNode

Pipelining of Data

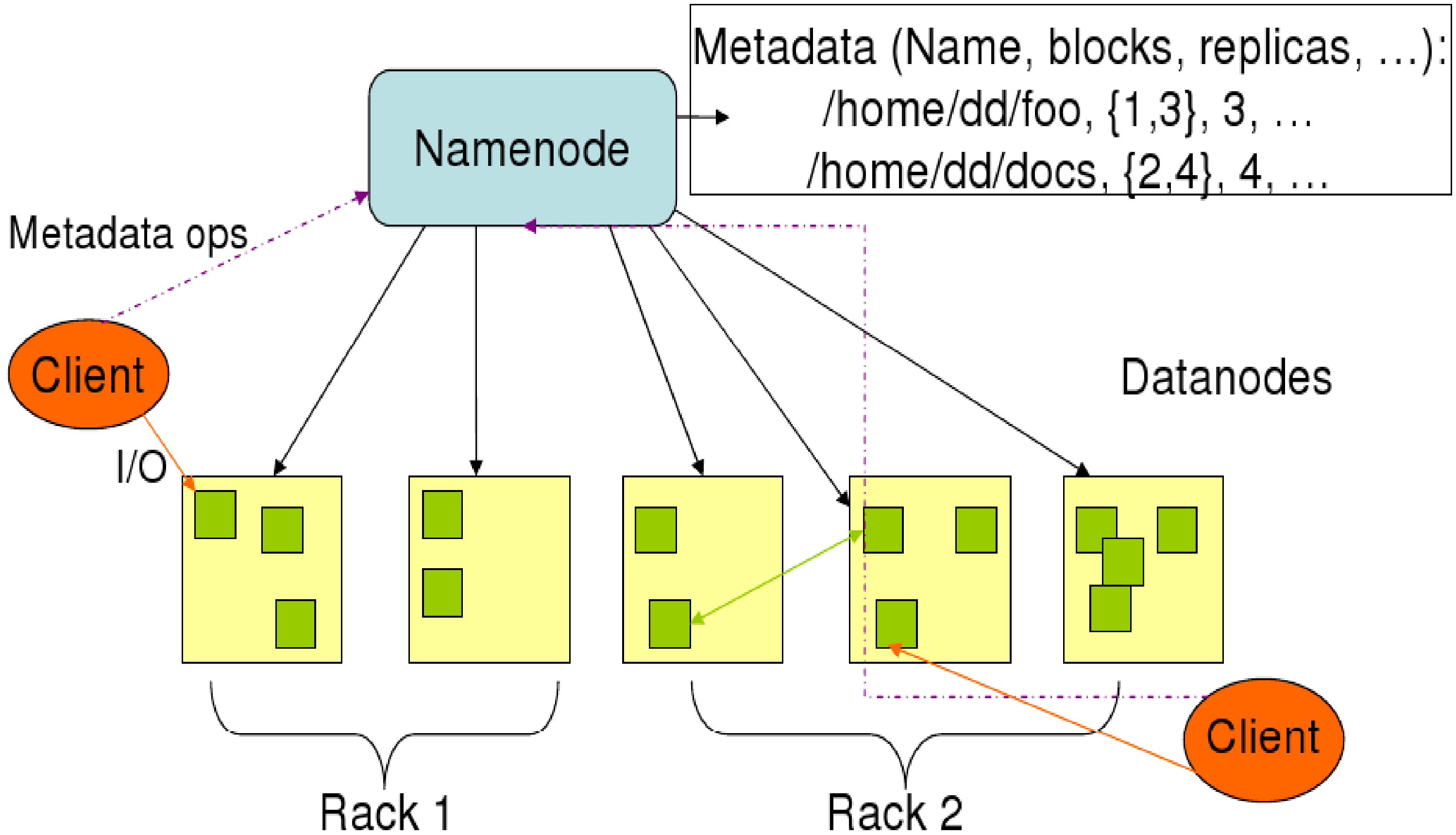
- Forwards data to other specified DataNodes

Perform replication tasks upon instruction by NameNode

Rack-aware

http://wiki.apache.org/hadoop/HadoopPresentations

HDFS Architecture



http://wiki.apache.org/hadoop/HadoopPresentations

Fault Tolerance in HDFS

DataNodes send heartbeats to the NameNode
- Once every 3 seconds

NameNode uses heartbeats to detect DataNode failures
- Chooses new DataNodes for new replicas
- Balances disk usage
- Balances communication traffic to DataNodes

http://wiki.apache.org/hadoop/HadoopPresentations

Data Pipelining

- Client retrieves a list of DataNodes on which to place replicas of a block
  - Client writes block to the first DataNode
  - The first DataNode forwards the data to the next
  - The second DataNode forwards the data to the next
- DataNode in the Pipeline
  - When all replicas are written, the client moves on to write the next block in file

http://wiki.apache.org/hadoop/HadoopPresentations

Hadoop MapReduce

Master-worker architecture

- Map-Reduce Master "JobTracker"
  - Accepts MR jobs submitted by users
  - Assigns Map and Reduce tasks to TaskTrackers
  - Monitors task and TaskTracker status, re-executes tasks upon failure

http://wiki.apache.org/hadoop/HadoopPresentations

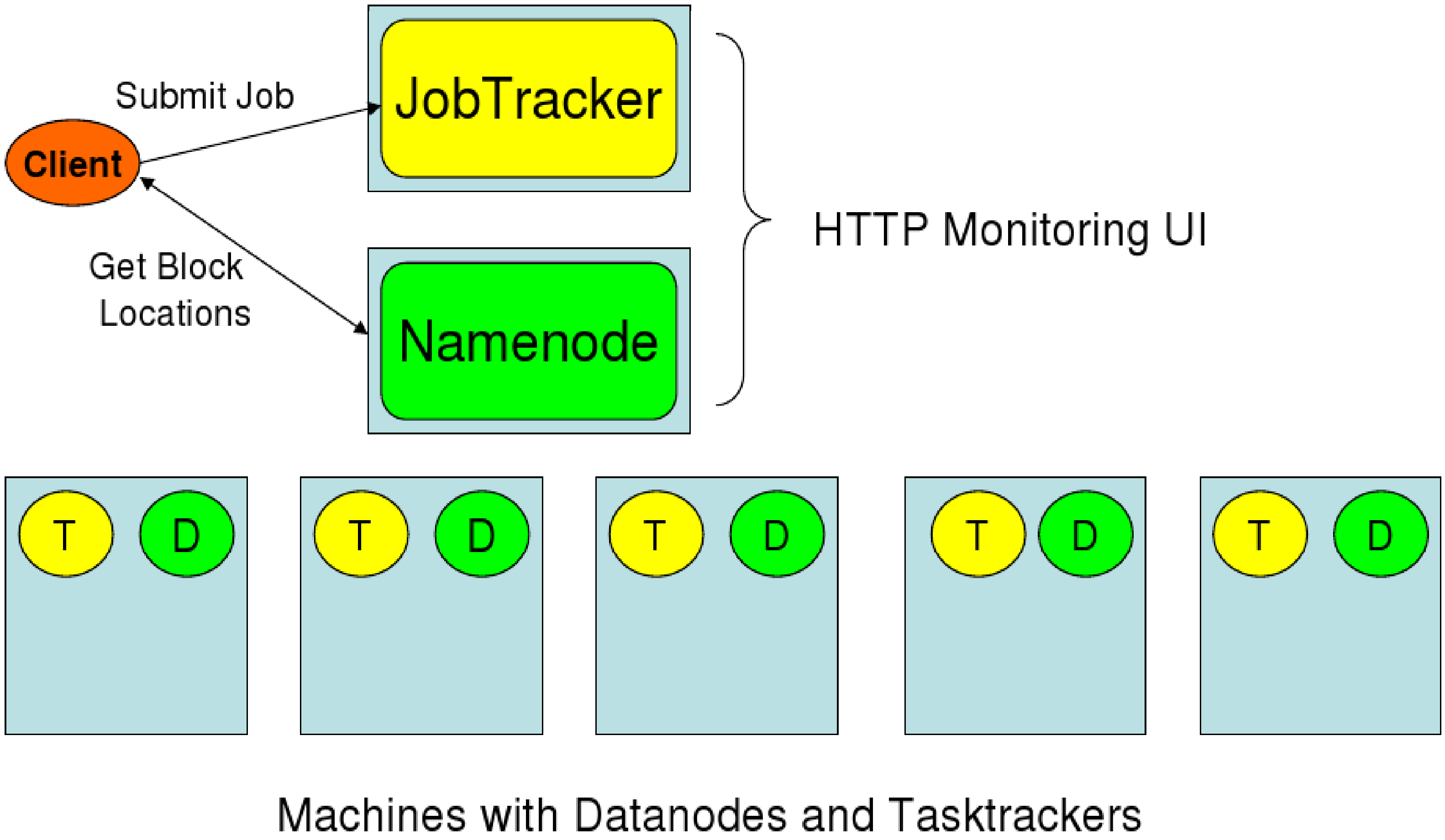
Hadoop MapReduce

Master-worker architecture

- Map-Reduce worker “TaskTrackers”
  - Run Map and Reduce tasks upon instruction from the JobTracker
- Manage storage and transmission of intermediate output

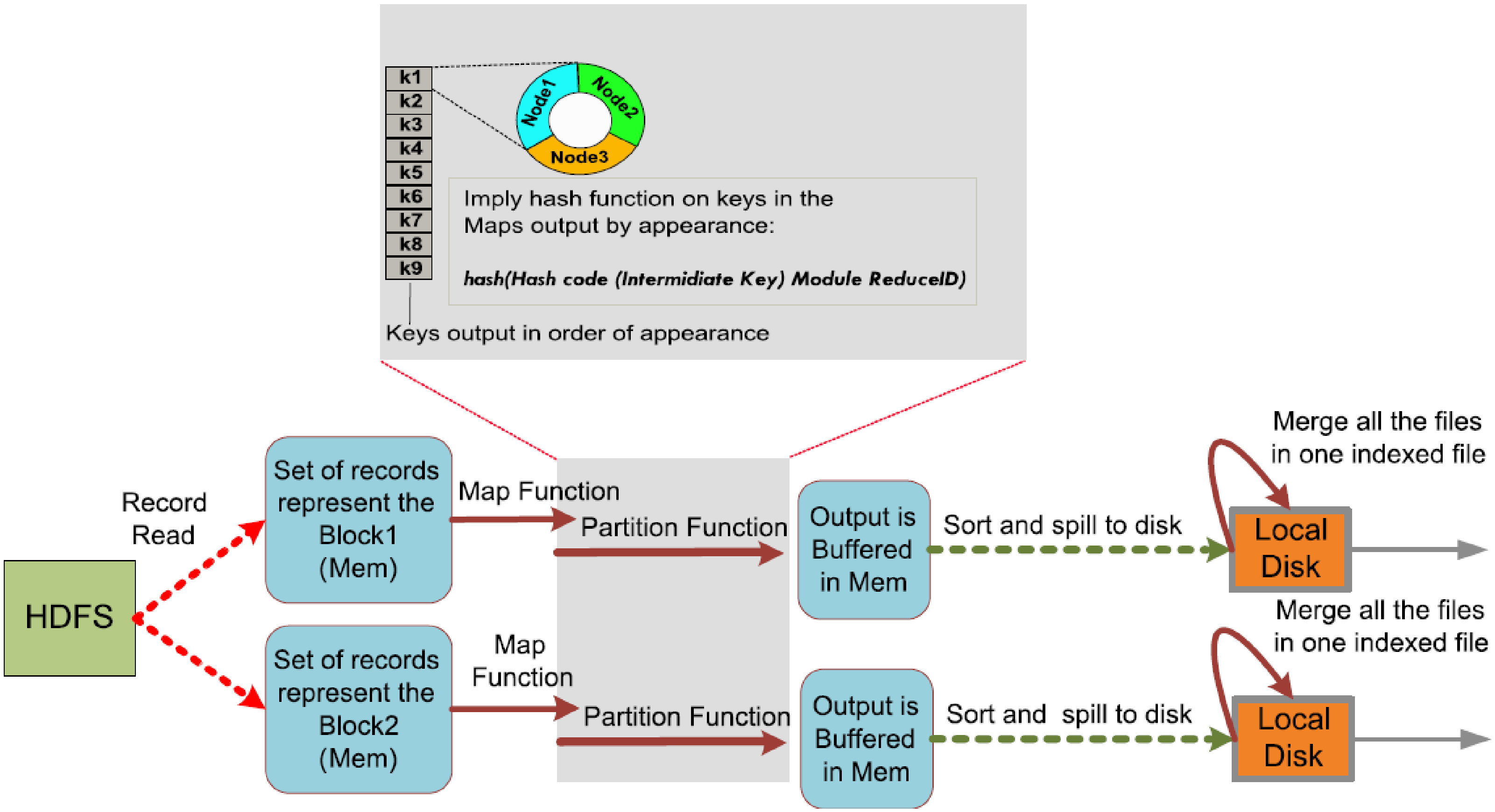
http://wiki.apache.org/hadoop/HadoopPresentations

Deployment: HDFS + MR



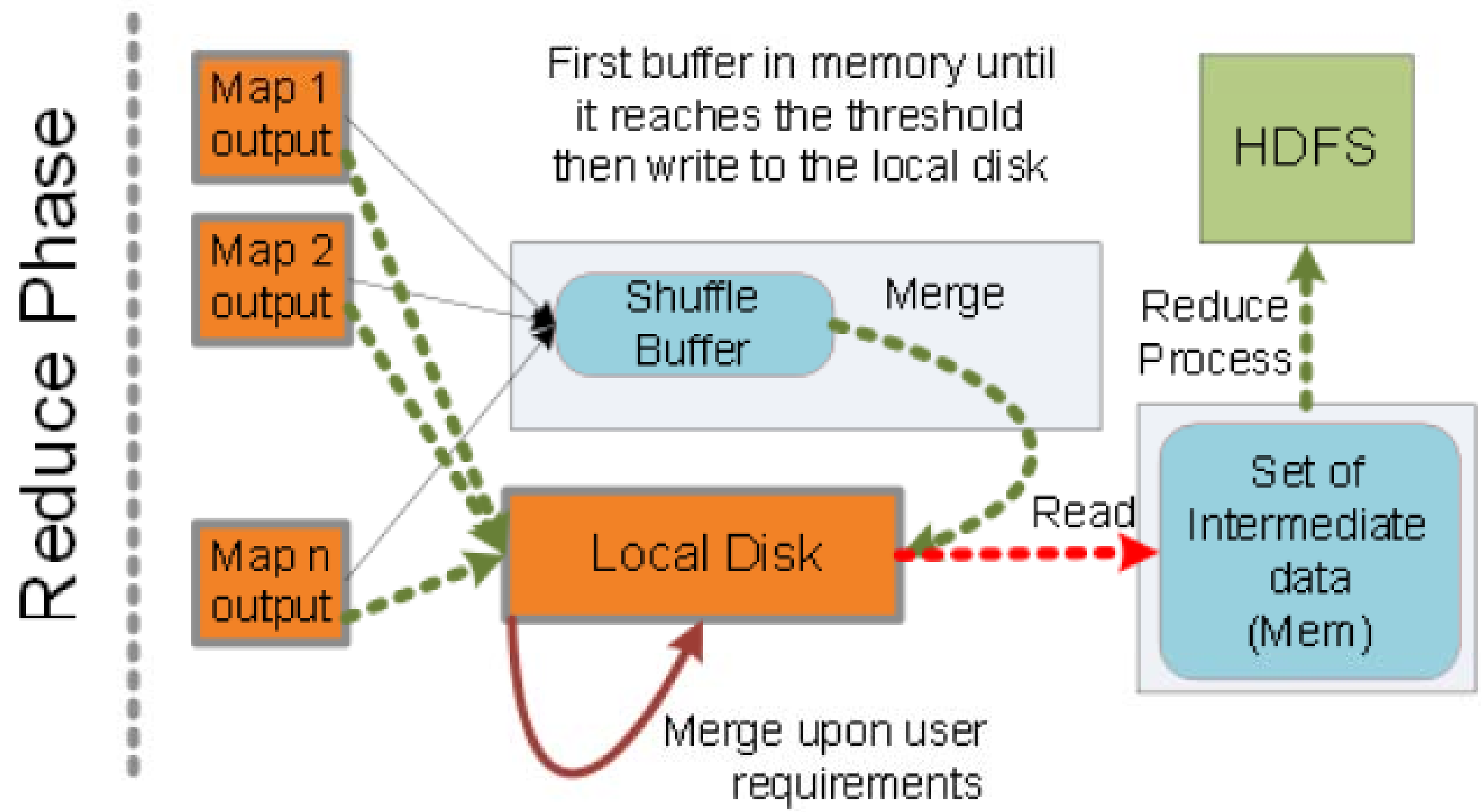
http://wiki.apache.org/hadoop/HadoopPresentations

Zoom on Map Phase



"Handling partitioning skew in MapReduce using LEEN" S Ibrahim, H Jin, L Lu, B He, G Antoniu, S Wu - Peer-to-Peer Networking and Applications, 2013

Zoom on Reduce Phase



Data Locality

Data Locality is exposed in the Map Task scheduling

Data are Replicated:

- Fault tolerance
- Performance : divide the work among nodes

Job Tracker schedules map tasks considering:

- Node-aware
- Rack-aware
- non-local map Tasks

Fault-tolerance

TaskTrackers send heartbeats to the Job Tracker

» Once every 3 seconds

TaskTracker uses heartbeats to detect

» Node is labled as failed If no heartbeat is recieved for a defined expiry time (Defualt : 10 Minutes)

Re-execute all the ongoing and complete tasks

Speculation in Hadoop

- Nodes slow (stragglers) → run backup tasks
  - Other jobs consuming resources on machine
  - Bad disks with soft errors transfer data very slowly
  - Weird things: processor caches disabled (!!)

Adopted from a presentation by Matei Zaharia "Improving MapReduce Performance in Heterogeneous Environments", OSDI 2008, San Diego, CA, December 2008.

Speculation in Hadoop

- Nodes slow (stragglers) → run backup tasks

  Other jobs consuming resources on machine

  Bad disks with soft errors transfer data very slowly

  Weird things: processor caches disabled (!!)

Node 1

Node 2

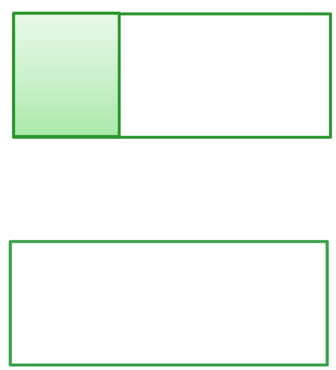
Adopted from a presentation by Matei Zaharia "Improving MapReduce Performance in Heterogeneous Environments", OSDI 2008, San Diego, CA, December 2008.

Speculation in Hadoop

- Nodes slow (stragglers) → run backup tasks
  - Other jobs consuming resources on machine
  - Bad disks with soft errors transfer data very slowly
  - Weird things: processor caches disabled (!!)

Node 1

Node 2



Adopted from a presentation by Matei Zaharia "Improving MapReduce Performance in Heterogeneous Environments", OSDI 2008, San Diego, CA, December 2008.

Speculation in Hadoop

- Nodes slow (stragglers) → run backup tasks

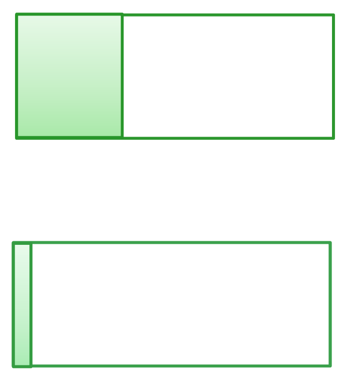
  Other jobs consuming resources on machine

  Bad disks with soft errors transfer data very slowly

  Weird things: processor caches disabled (!!)

Node 1

Node 2



Adopted from a presentation by Matei Zaharia "Improving MapReduce Performance in Heterogeneous Environments", OSDI 2008, San Diego, CA, December 2008.

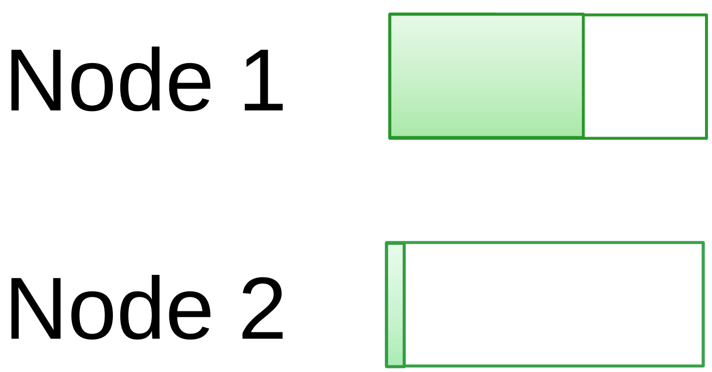
Speculation in Hadoop

- Nodes slow (stragglers) → run backup tasks

Other jobs consuming resources on machine

Bad disks with soft errors transfer data very slowly

Weird things: processor caches disabled (!!)



Adopted from a presentation by Matei Zaharia "Improving MapReduce Performance in Heterogeneous Environments", OSDI 2008, San Diego, CA, December 2008.

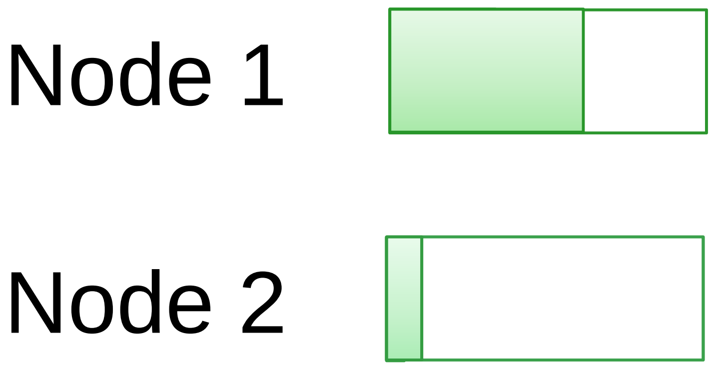
Speculation in Hadoop

- Nodes slow (stragglers) → run backup tasks

Other jobs consuming resources on machine

Bad disks with soft errors transfer data very slowly

Weird things: processor caches disabled (!!)



Adopted from a presentation by Matei Zaharia "Improving MapReduce Performance in Heterogeneous Environments", OSDI 2008, San Diego, CA, December 2008.

Speculation in Hadoop

- Nodes slow (stragglers) → run backup tasks

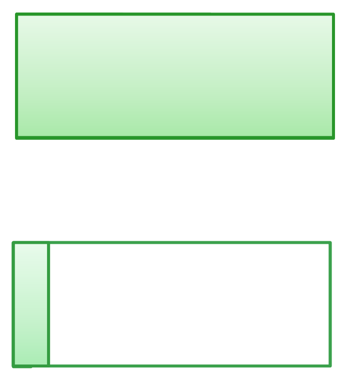
  Other jobs consuming resources on machine

  Bad disks with soft errors transfer data very slowly

  Weird things: processor caches disabled (!!)

Node 1

Node 2



Adopted from a presentation by Matei Zaharia "Improving MapReduce Performance in Heterogeneous Environments", OSDI 2008, San Diego, CA, December 2008.

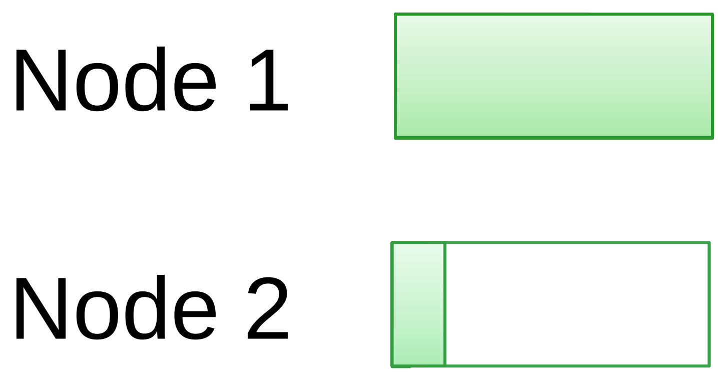
Speculation in Hadoop

- Nodes slow (stragglers) → run backup tasks

Other jobs consuming resources on machine

Bad disks with soft errors transfer data very slowly

Weird things: processor caches disabled (!!)



Adopted from a presentation by Matei Zaharia "Improving MapReduce Performance in Heterogeneous Environments", OSDI 2008, San Diego, CA, December 2008.

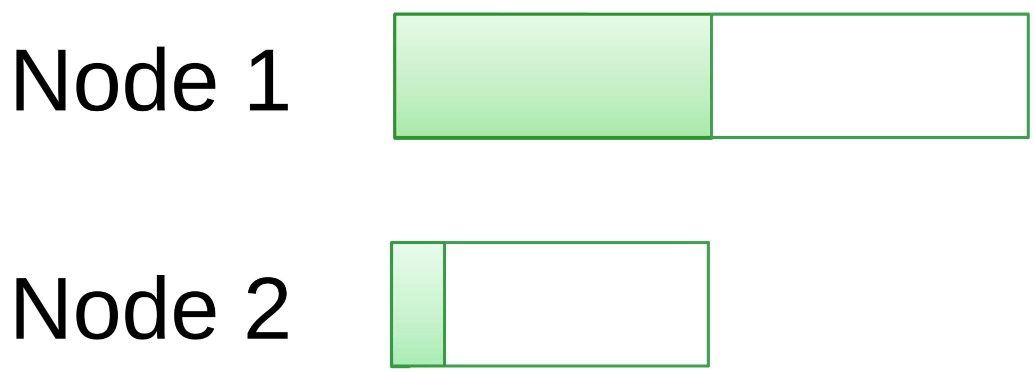
Speculation in Hadoop

- Nodes slow (stragglers) → run backup tasks

Other jobs consuming resources on machine

Bad disks with soft errors transfer data very slowly

Weird things: processor caches disabled (!!)



Adopted from a presentation by Matei Zaharia "Improving MapReduce Performance in Heterogeneous Environments", OSDI 2008, San Diego, CA, December 2008.

Speculation in Hadoop

- Nodes slow (stragglers) → run backup tasks

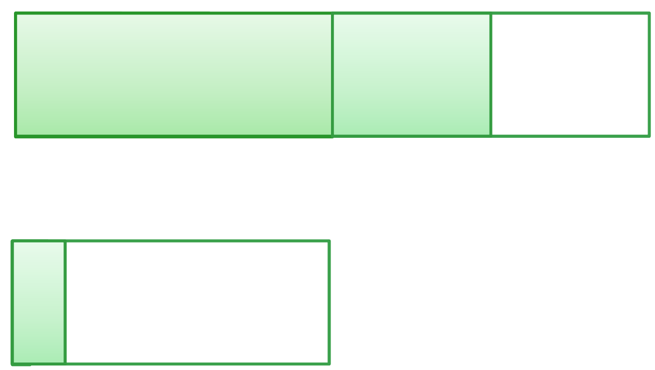
  Other jobs consuming resources on machine

  Bad disks with soft errors transfer data very slowly

  Weird things: processor caches disabled (!!)

Node 1

Node 2



Adopted from a presentation by Matei Zaharia "Improving MapReduce Performance in Heterogeneous Environments", OSDI 2008, San Diego, CA, December 2008.

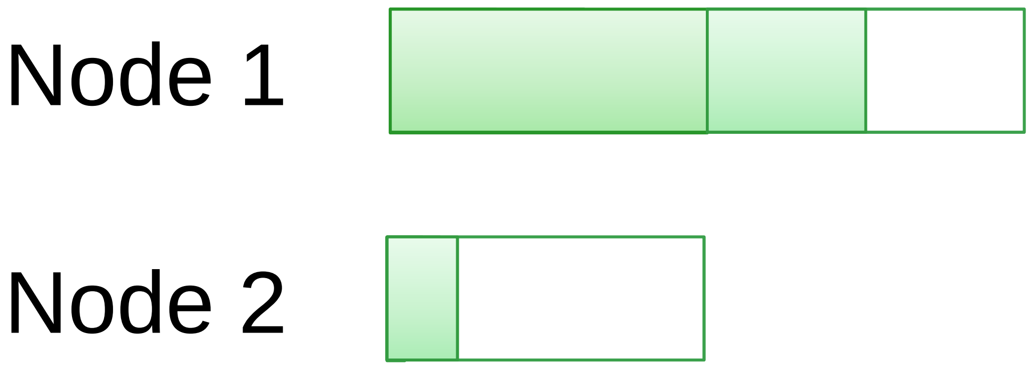
Speculation in Hadoop

- Nodes slow (stragglers) → run backup tasks

  Other jobs consuming resources on machine

  Bad disks with soft errors transfer data very slowly

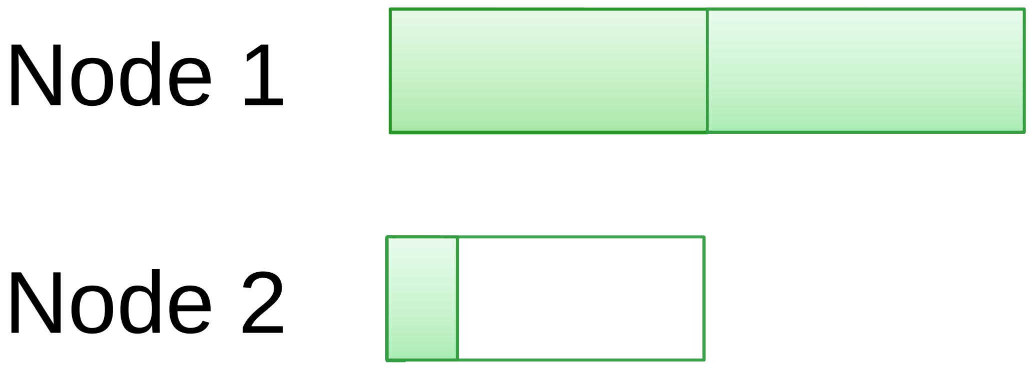
  Weird things: processor caches disabled (!!)



Adopted from a presentation by Matei Zaharia "Improving MapReduce Performance in Heterogeneous Environments", OSDI 2008, San Diego, CA, December 2008.

Speculation in Hadoop

- Nodes slow (stragglers) → run backup tasks
  - Other jobs consuming resources on machine
  - Bad disks with soft errors transfer data very slowly
  - Weird things: processor caches disabled (!!)



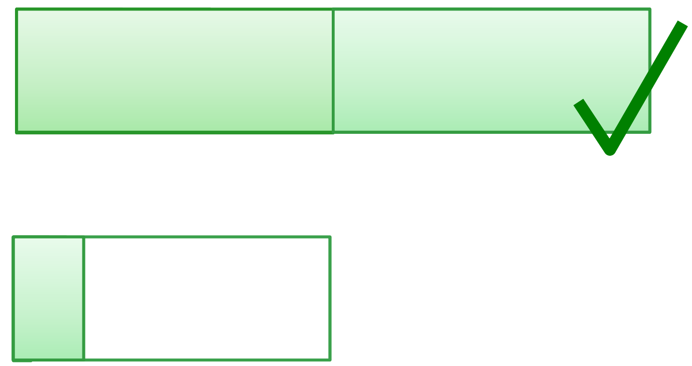
Adopted from a presentation by Matei Zaharia "Improving MapReduce Performance in Heterogeneous Environments", OSDI 2008, San Diego, CA, December 2008.

Speculation in Hadoop

- Nodes slow (stragglers) → run backup tasks
  - Other jobs consuming resources on machine
  - Bad disks with soft errors transfer data very slowly
  - Weird things: processor caches disabled (!!)

Node 1

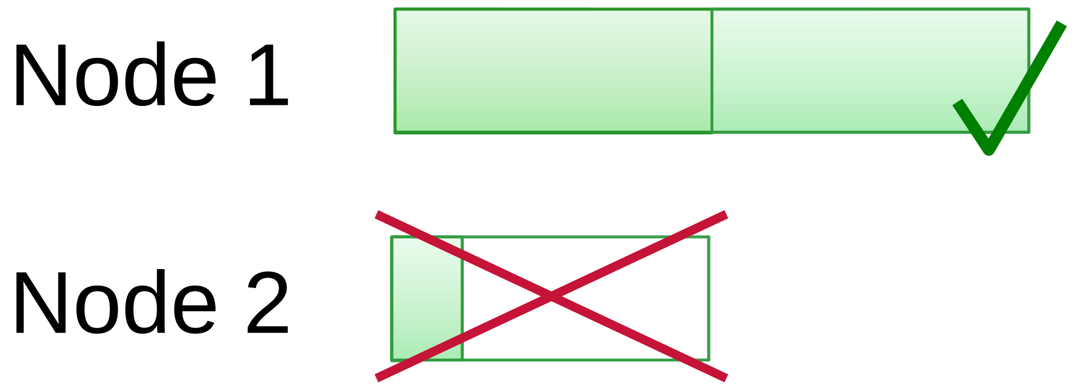
Node 2



Adopted from a presentation by Matei Zaharia "Improving MapReduce Performance in Heterogeneous Environments", OSDI 2008, San Diego, CA, December 2008.

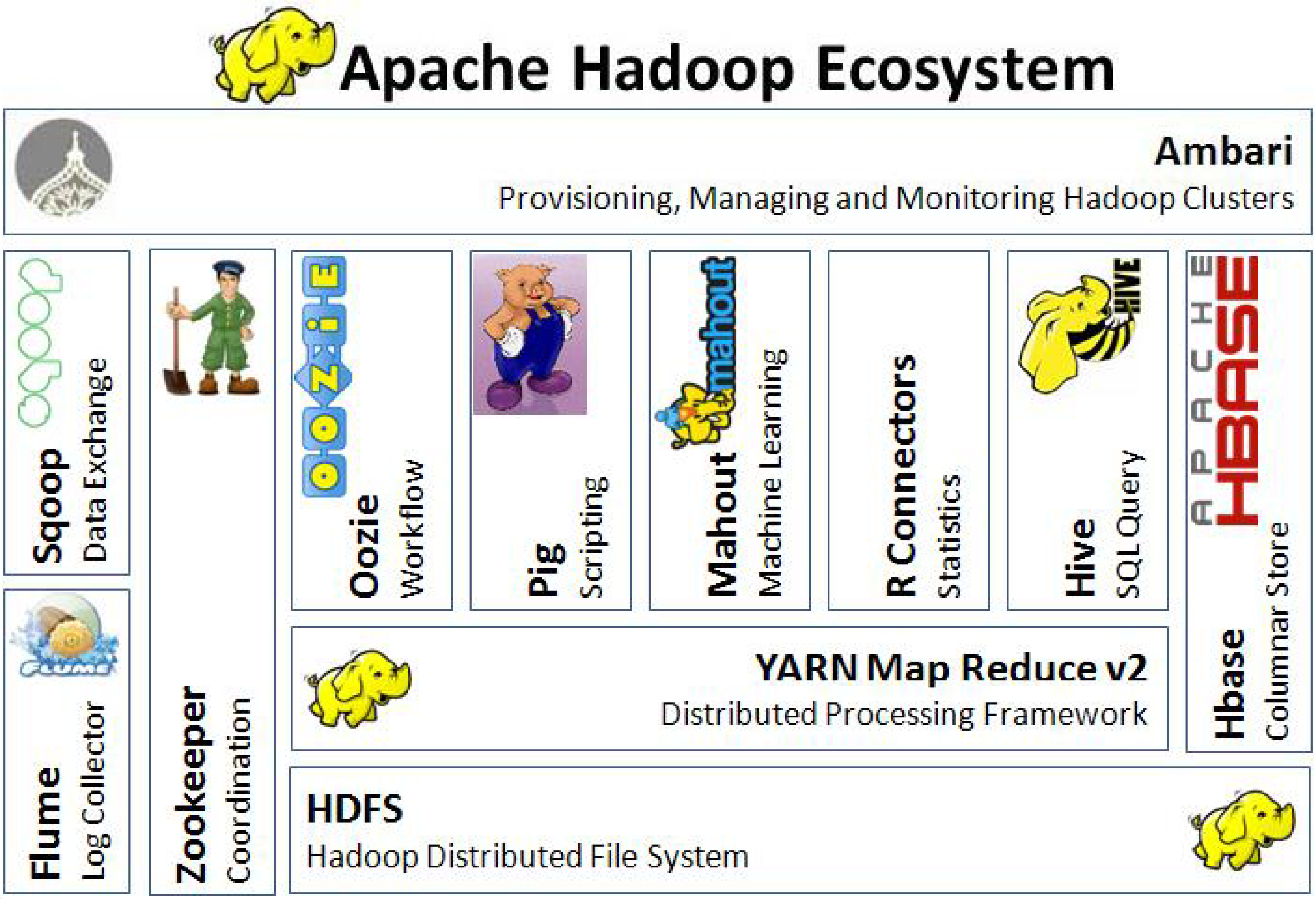
Speculation in Hadoop

- Nodes slow (stragglers) → run backup tasks



Adopted from a presentation by Matei Zaharia "Improving MapReduce Performance in Heterogeneous Environments", OSDI 2008, San Diego, CA, December 2008.

Hadoop ecosystem



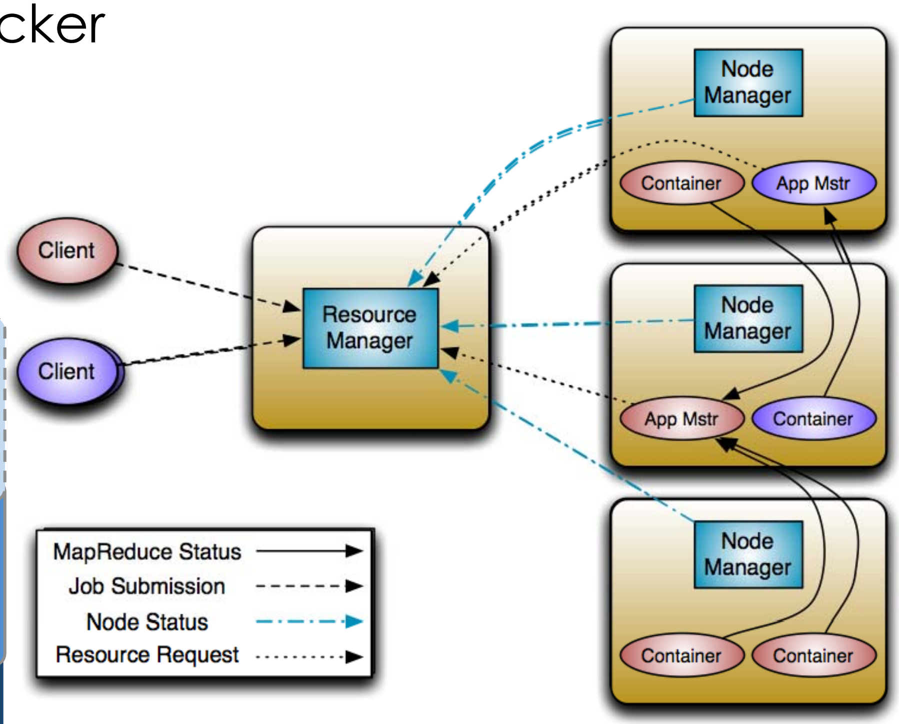
Hadoop ecosystem

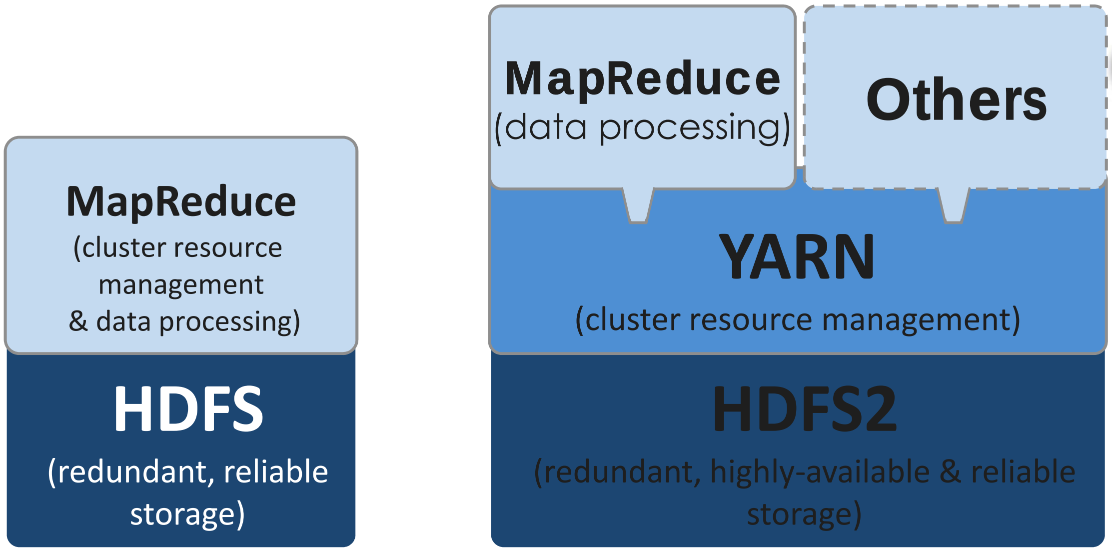
- Core: Hadoop MapReduce and HDF
- Data Access: Hbase, Pig, Hive
- Algorithms: Mahout
- Data Import: Flume, Sqoop and Nu

Hadoop 1.0 vs Hadoop 2.0

Split up the two major functions of JobTracker
- Cluster resource management
- Application life-cycle management

MapReduce becomes **user-land** library





Scalability-Fault-tolerance- Support for programming paradigms – Better utilization

Hadoop ecosystem

- Database primitives
- Hbase
  - » Wide column data structure built on HDFS

- Processing
- Pig
  - » A high level data-flow language and execution framework for parallel computatuon

Hadoop ecosystem

- Algorithm
- Mahout
  - » Scalable machine learning data mining
  - » Runs on the tip of Hadoop

- Processing
- Hive
  - » Data Warehouse Infrastructure
  - » Support for custom mappers and reducers for more sophisticated analysis

Hadoop ecosystem

- Data Import
  - Sqoop
    - Structured Data
    - Import and Export from RDBMS to HDFS

- Data Import
  - Flume
    - Import streams
    - Text files and systems logs

Hadoop ecosystem

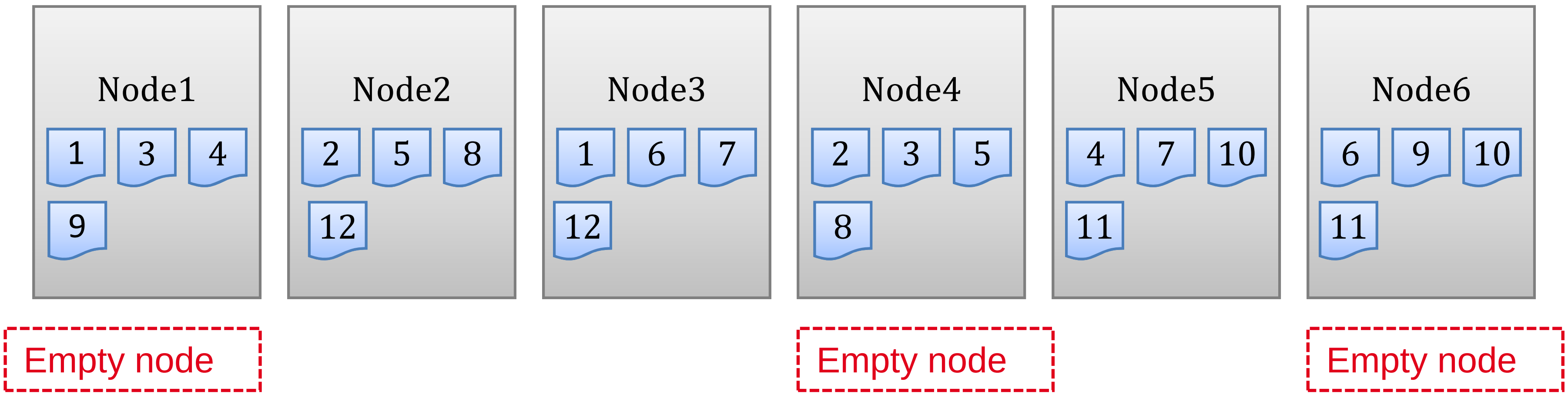
- Coordination
  - ZooKeeper
    - A high-performance coordination service for distributed applications

- Scheduling
  - Oozie
    - A workflow scheduler system to manage Apache Hadoop jobs

Research challenges

Data locality

Data locality in the Cloud



The simplicity of Map tasks Scheduling leads to

Non-local maps execution (25%)

Data locality: Side impacts

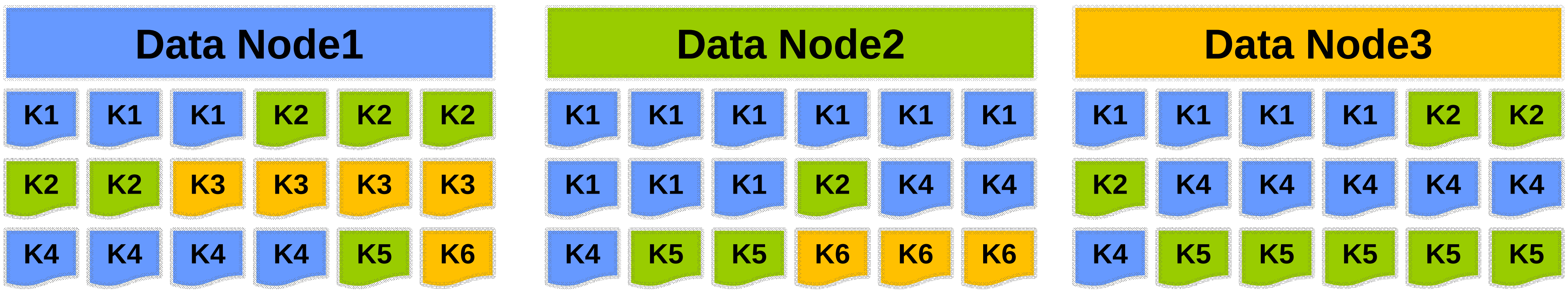
- Increase the execution time
- Increase the number of useless speculation
- Slot occupying
  - Imbalance in the Map execution among nodes

Data Skew

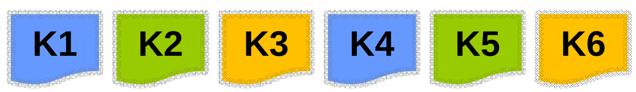
Data Skew

- The current Hadoop's hash partitioning works well when the keys are equally appeared and uniformly stored in the data nodes
- In the presence of Partitioning Skew:
  - Variation in Intermediate Keys' frequencies
  - Variation in Intermediate Key's distribution amongst different data node
- Native blindly hash-partitioning is to be inadequate and will lead to:
  - Network congestion
  - Unfairness in reducers' inputs ? Reduce computation Skew
  - Performance degradation

Data Skew



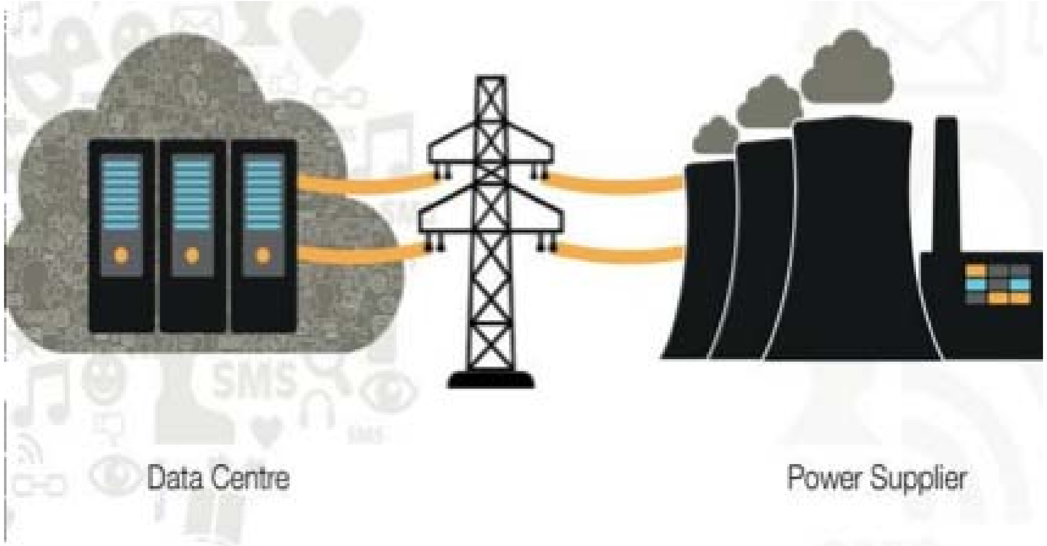
hash (Hash code (Intermediate-key) Modulo ReduceID)



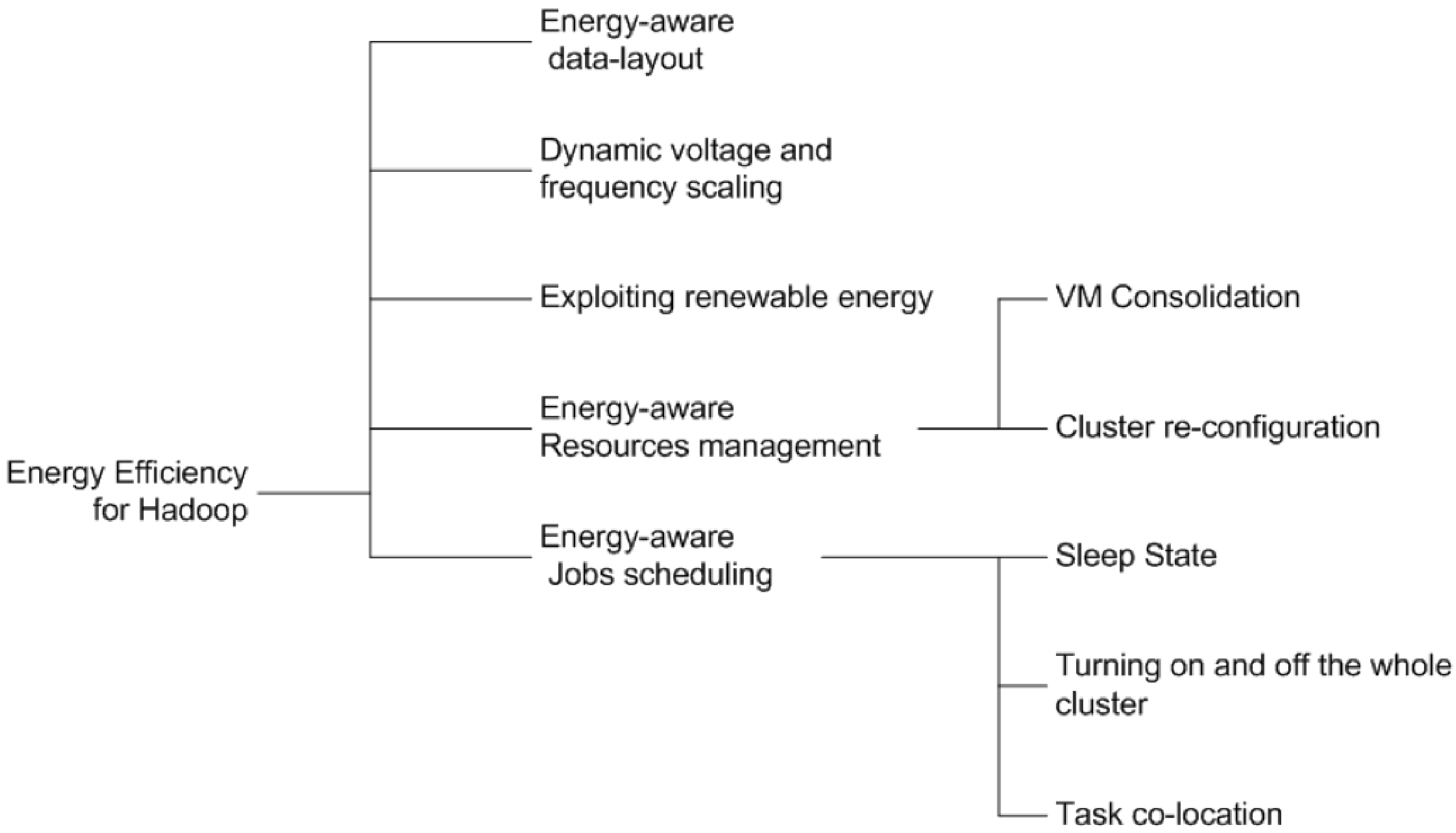
	Data Node1	Data Node2	Data Node3	
Total Data Transfer	**11**	**15**	**18**	Total 44/54
Reduce Input	**29**	**17**	**8**	cv 58%

Energy Efficency

- Data-centers consume large amount of energy
  - High monetary
    - Power bills become a substantial part of the total cost of ownership (TCO) of data-center (40% of the total cost)
  - High carbon emission

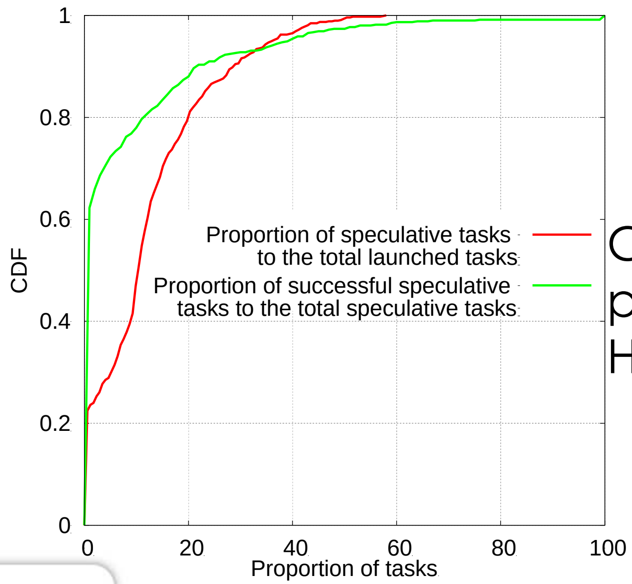


Energy-efficiency in Hadoop



Effective Stragglers Mitigation

Launching copies Resource consumption



CMU trace of production Hadoop cluster

Improve the detection and handling mechanisms used in Hadoop

Job Scheduling

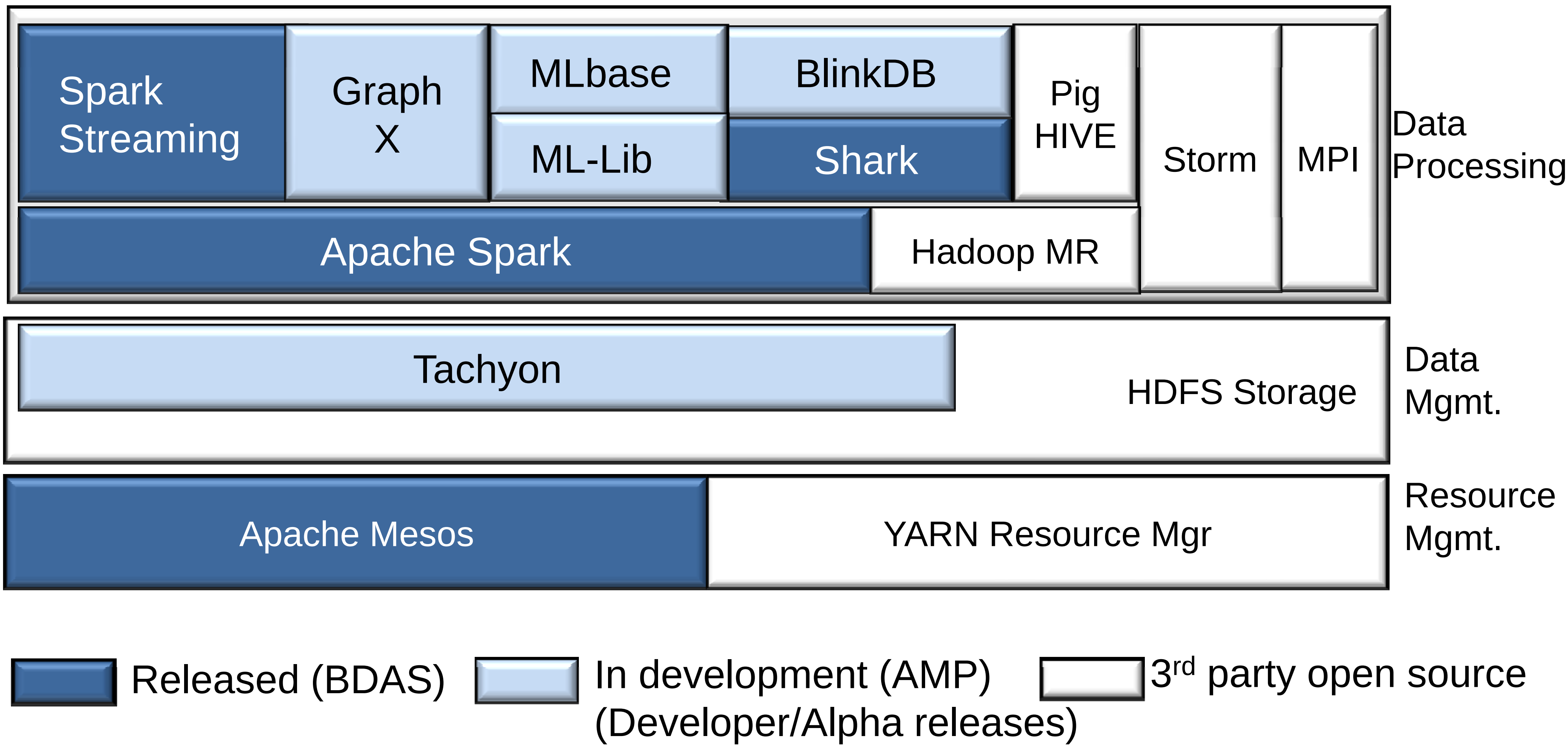
- MapReduce / Hadoop originally designed for high throughput batch processing
- Today's workload is far more diverse:
  - *Many users* want to share a cluster
    - Engineering, marketing, business intelligence, etc
  - Vast majority of jobs are *short*
    - Ad-hoc queries, sampling, periodic reports
  - *Response time* is critical
    - Interactive queries, deadline-driven reports

➢ How can we efficiently share MapReduce clusters between users?

- POST-HADOOP APPROACHES:

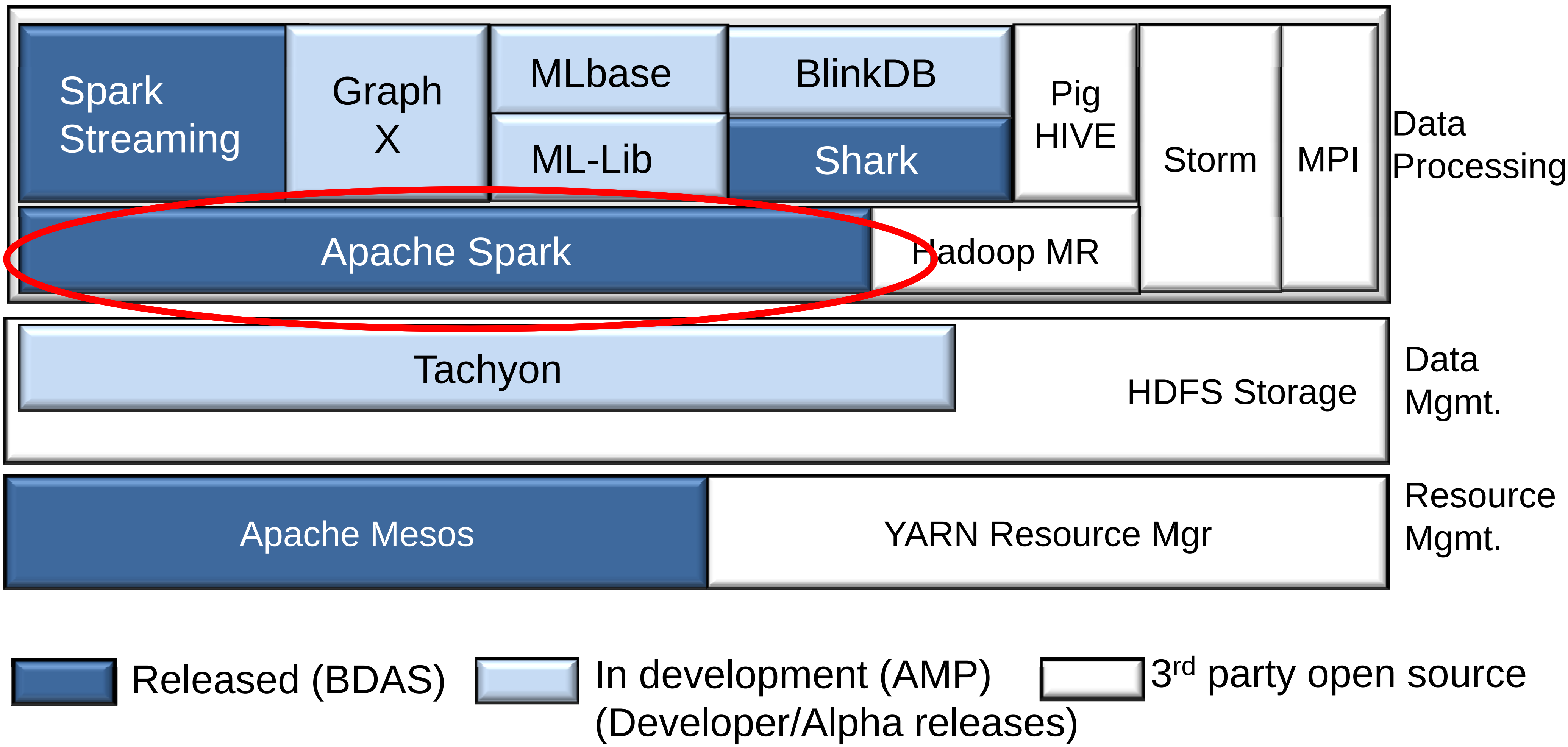
The Berkeley Data Analytics Stack

Berkeley Data Analytics Stack



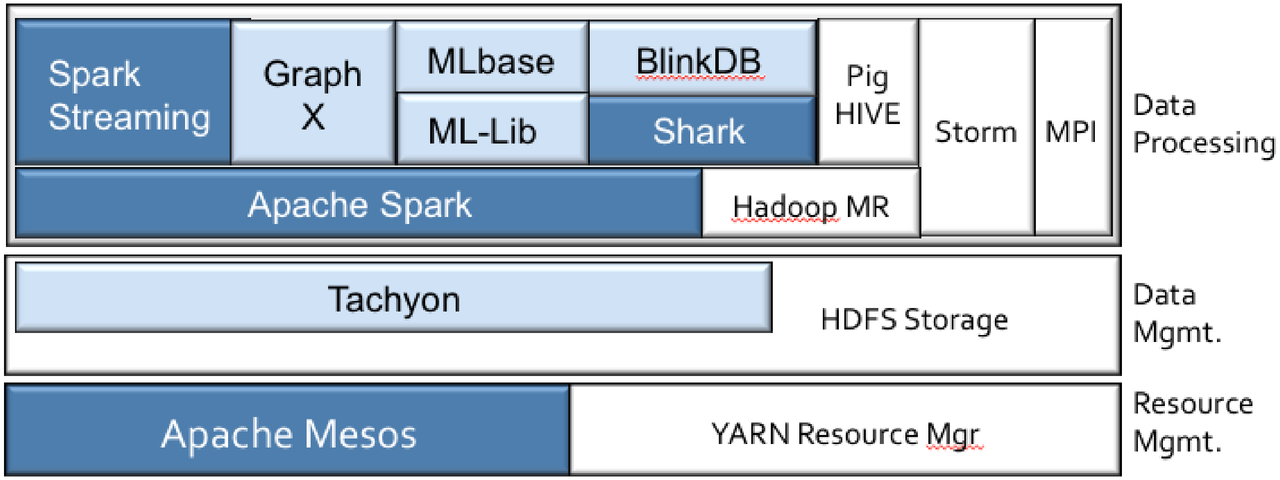
AMP BDAS Components being released under BSD or Apache Open Source License

Berkeley Data Analytics Stack



AMP BDAS Components being released under BSD or Apache Open Source License

BDAS Present - Summary



- **Spark Streaming** – Real-time Processing
- **Tachyon** – Memory-based layer on top of HDFS
- **BlinkDB** – Approximate Query Processing
- **MLlib** – Scalable Distributed Machine Learning Library in Spark
- **GraphX** – Graph processing integrated with Spark and Shark

See http://amplab.cs.berkeley.edu/publications for more information

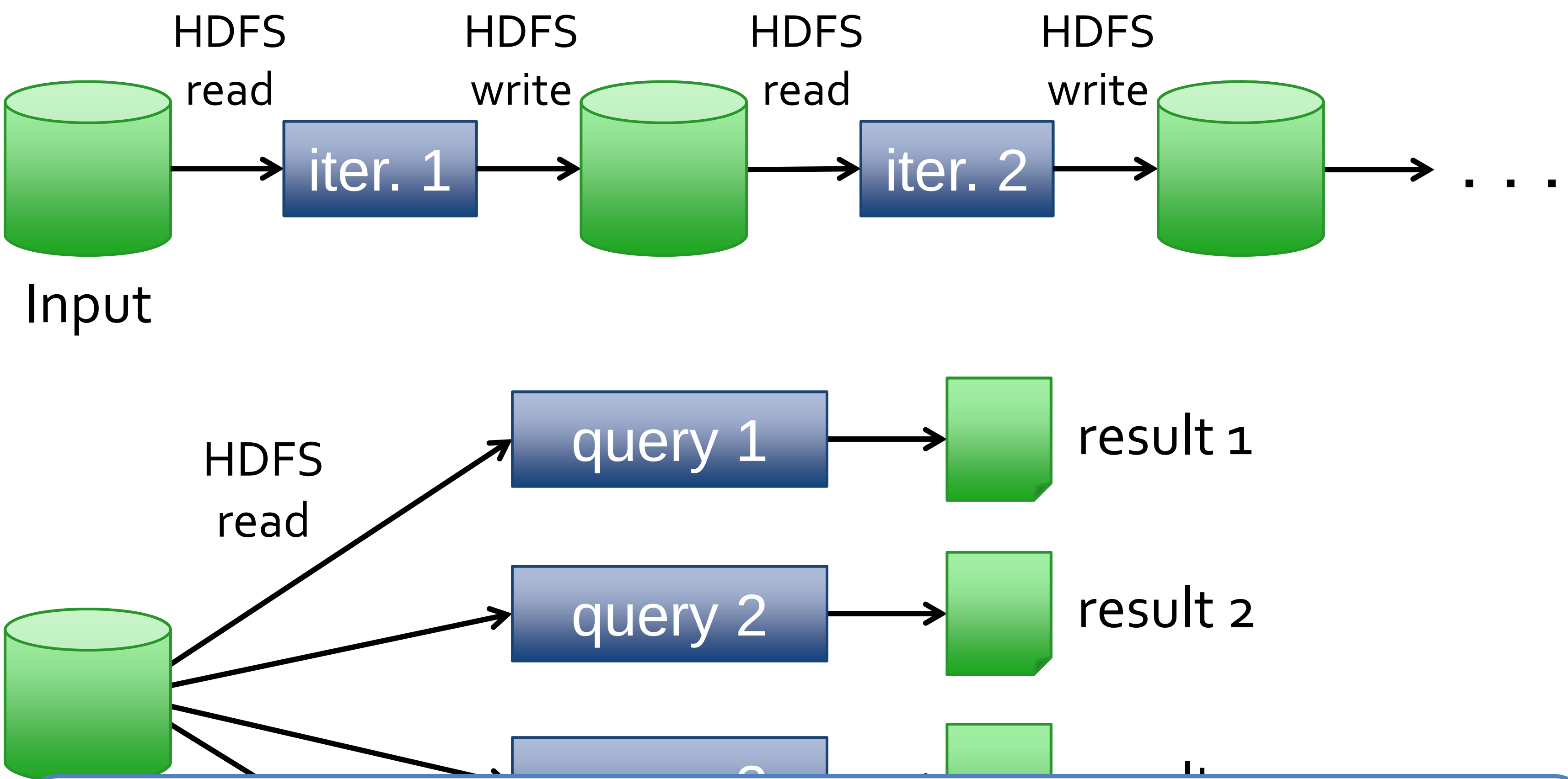
Beyond Hadoop

- Complex apps and interactive queries both need one thing that MapReduce lacks:
    - Efficient primitives for **data sharing**

In MapReduce, the only way to share data across jobs is stable storage ➔ slow!

M. Zaharia, et al, Resilient Distributed Datasets: A fault-tolerant abstraction for in-memory cluster computing, NSDI 2012.

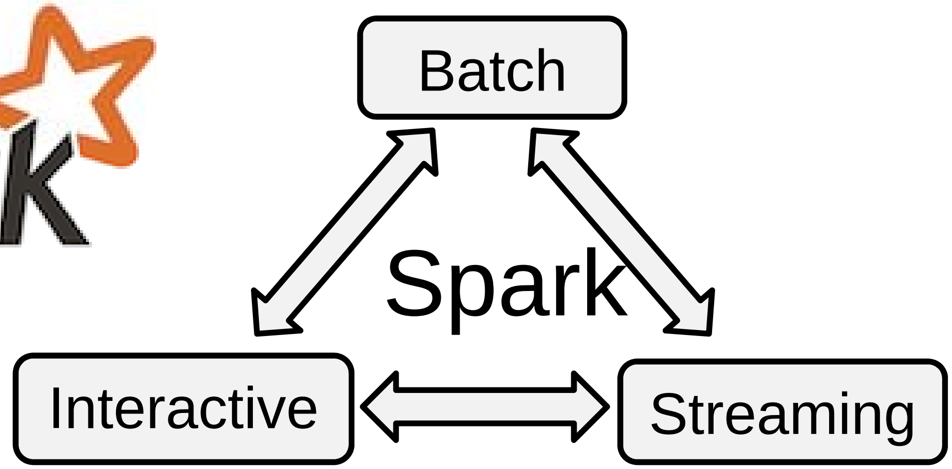
Examples



Slow due to replication and disk I/O, but necessary for fault tolerance

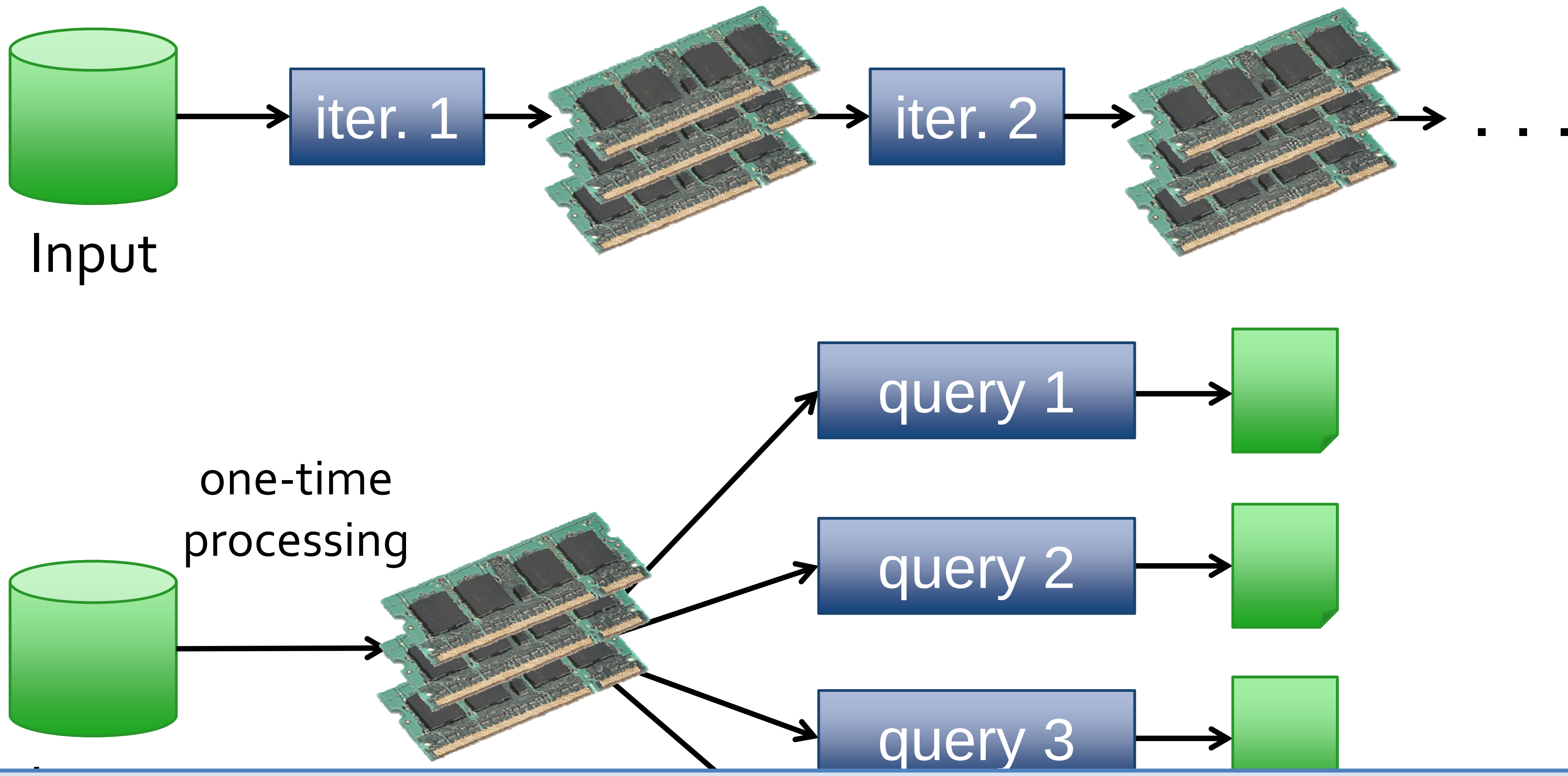
memory cluster computing, NSDI 2012.

Introducin Spark



- Fast, MapReduce-like engine
  - In-memory storage abstraction for iterative/interactive queries
  - General execution graphs
  - Up to 100x faster than Hadoop MR (2-10x even for on-disk)
- Compatible with Hadoop's storage APIs
  - Can access HDFS, HBase, S3, SequenceFiles, etc
- Great example of ML/Systems (and eventually DB) collaboration

In-Memory Data Sharing



10-100× faster than network/disk

M. Zaharia, et al, Resilient Distributed Datasets: A fault-tolerant abstraction for in-memory cluster computing, NSDI 2012.

Spark Streaming

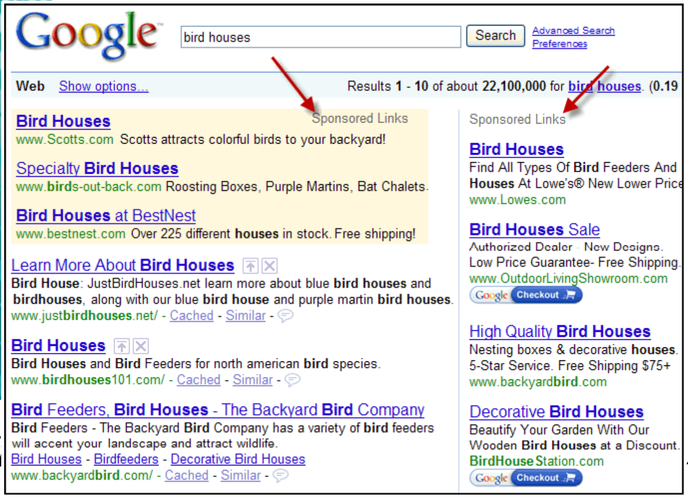
Why Spark Streaming?

Many big-data applications need to process large data streams in realtime

Website monitoring

Fraud detection

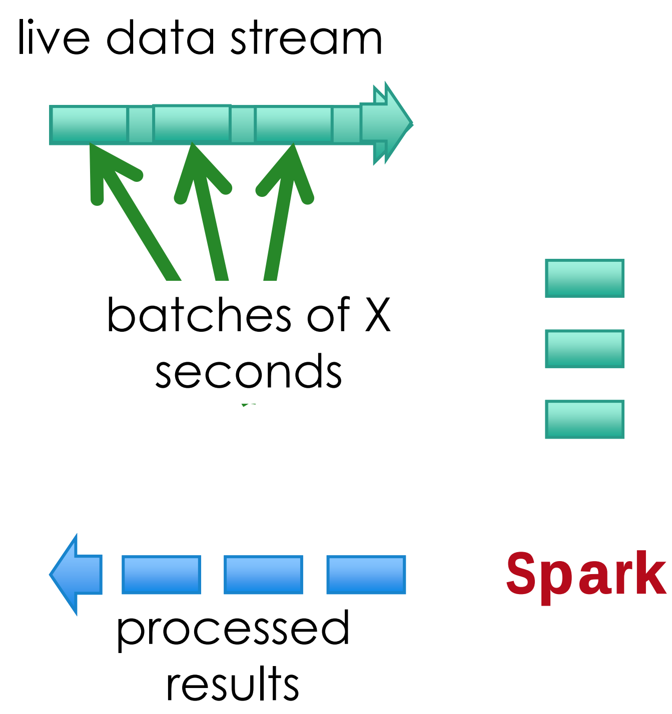
Ad monetization



M. Zaharia, et al, Resilient Distributed Datasets: A faul *uster computing, NSDI 2012.*

Discretized Stream Processing

Run a streaming computation as a **series of very small, deterministic batch jobs**



- Chop up the live stream into batches of X seconds
- Batch sizes as low as ½ second, latency ~ 1 second

M. Zaharia, et al, Resilient Distributed Datasets: A fault-tolerant abstraction for in-memory cluster computing, NSDI 2012.

INVENTORS FOR THE DIGITAL WORLD

Questions

Thank You!