



Education

# Introduction to Analytics and Big Data - Hadoop

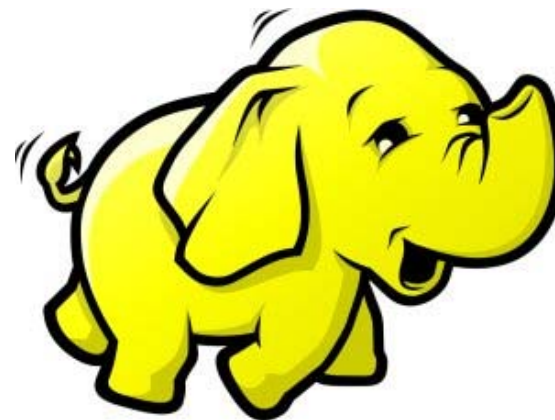
Rob Peglar  
EMC Isilon

- The material contained in this tutorial is copyrighted by the SNIA.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - ◆ Any slide or slides used must be reproduced in their entirety without modification
  - ◆ The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

**NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.**

# BIG DATA AND HADOOP

Data Challenges  
Why Hadoop





The Economist, Feb 25, 2010

IN 2010 THE DIGITAL UNIVERSE WAS  
**1.2 ZETTABYTES**

IN A DECADE THE DIGITAL UNIVERSE WILL BE  
**35 ZETTABYTES**

**90%** OF THE DIGITAL UNIVERSE IS  
**UNSTRUCTURED**

IN 2011 THE DIGITAL UNIVERSE IS  
**300 QUADRILLION** FILES

**WIRED**

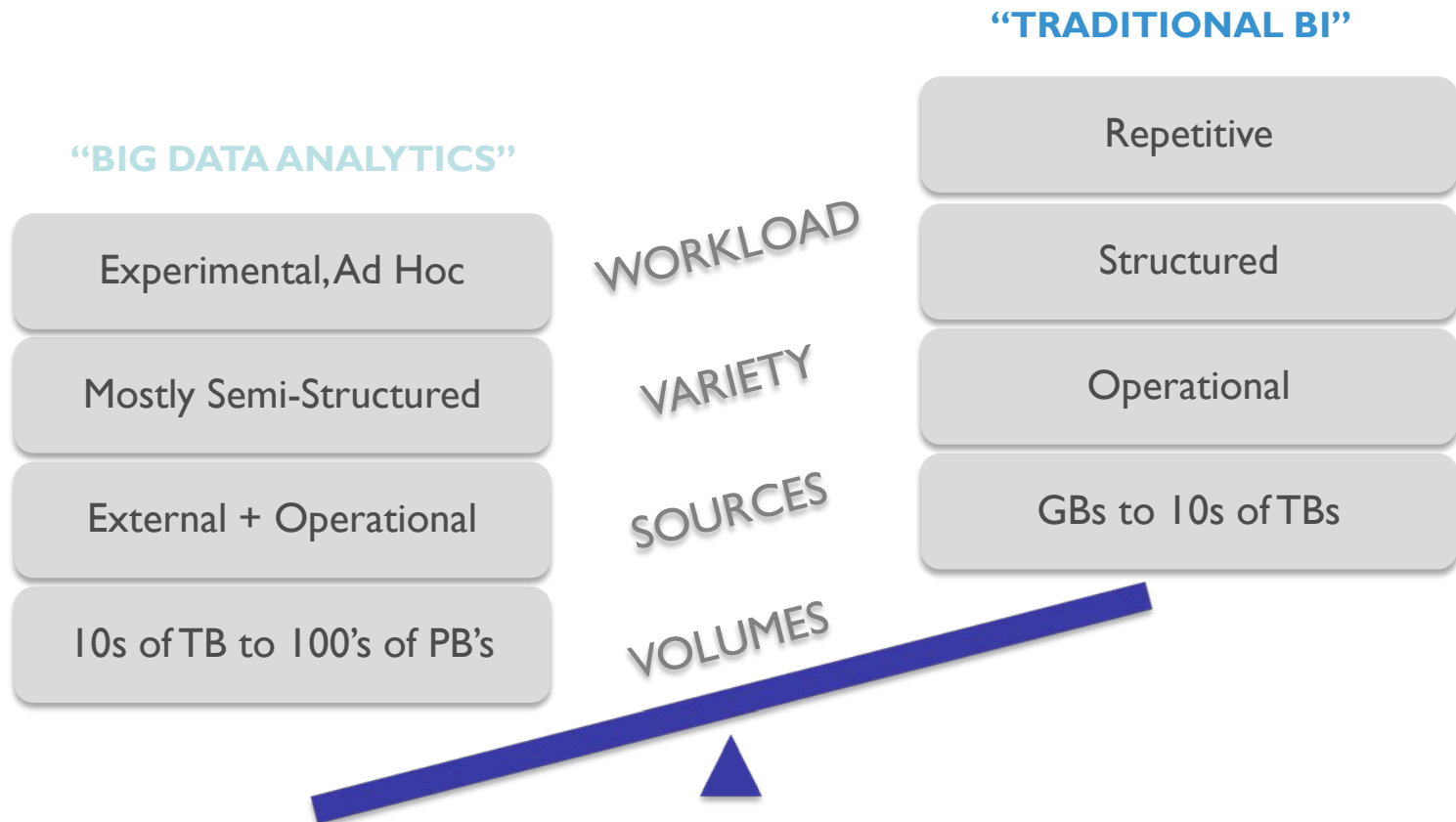
The New York Times

Bloomberg  
Businessweek

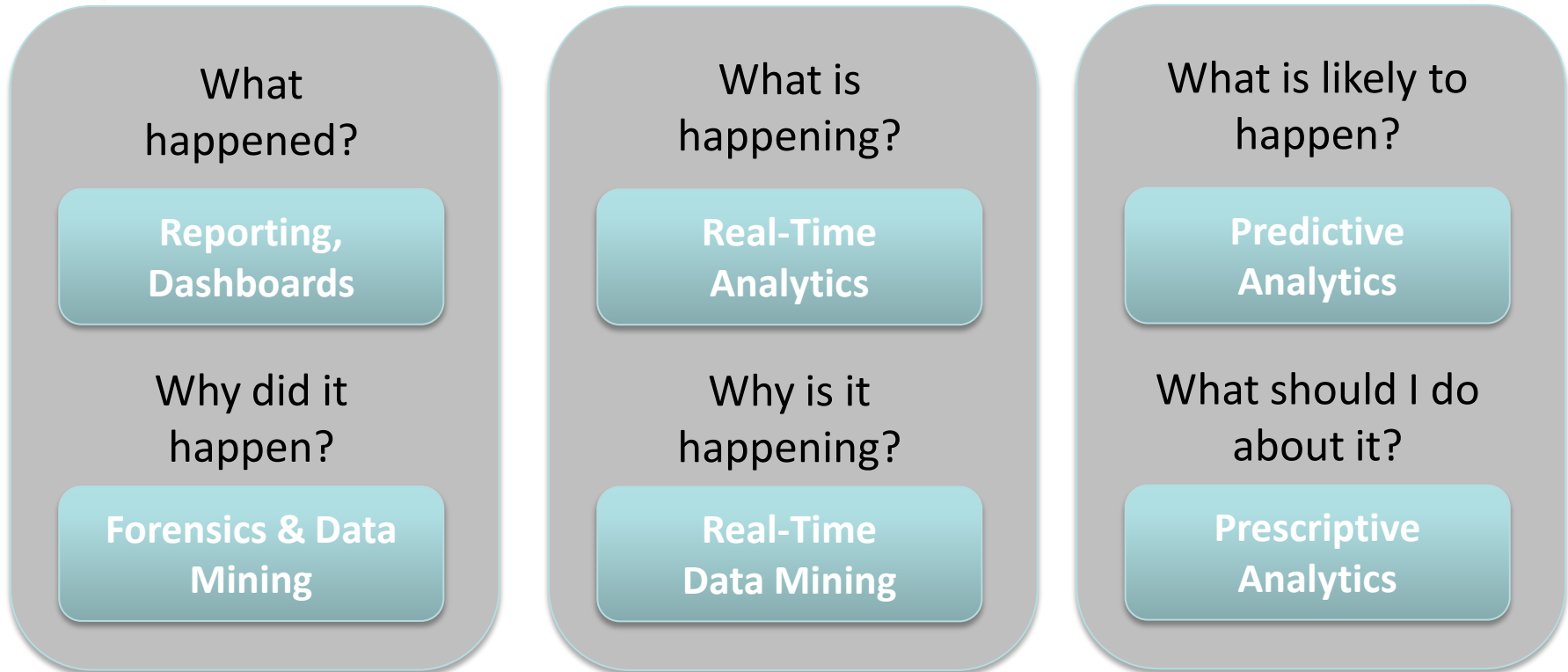
Forbes

WALL STREET JOURNAL

# Big Data Is Different than Business Intelligence



# Questions from Businesses will Vary



# Web 2.0 is “Data-Driven”



“The future is here, it’s just not evenly distributed yet.”  
William Gibson

# The world of Data-Driven Applications

google.org Flu Trends

[Google.org home](#)

Flu Trends

[Home](#)

United States

National

[Download data](#)

[How does this work?](#)

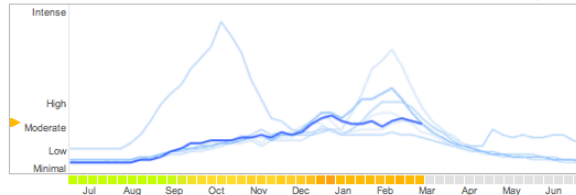
[FAQ](#)

## Explore flu trends - United States

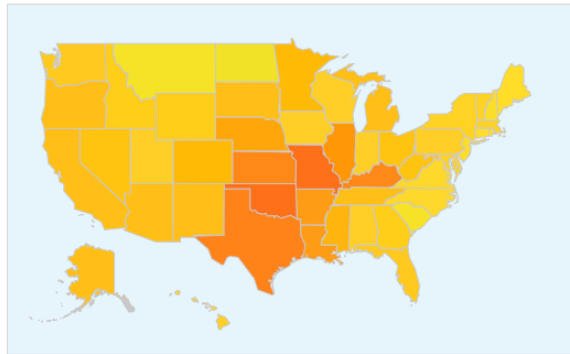
We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. [Learn more >](#)

National

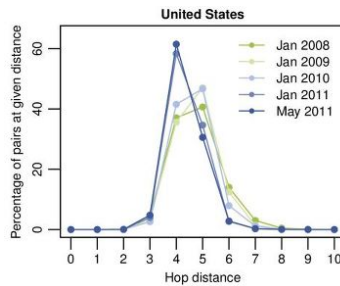
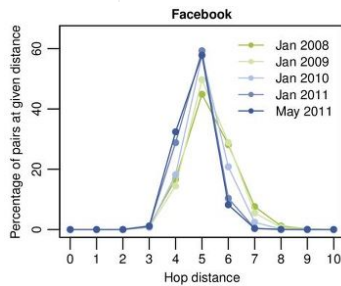
2011-2012 Past years



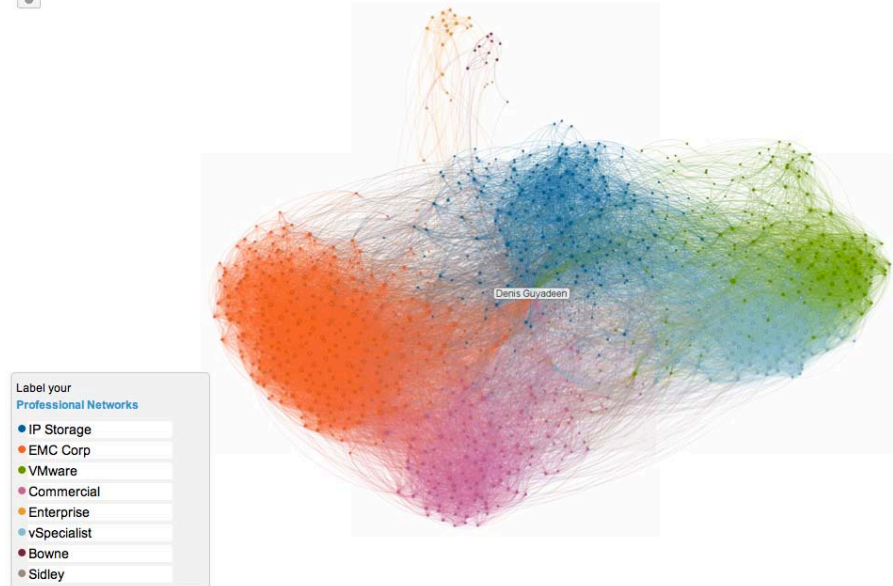
States [Cities](#) (Experimental)



Estimates were made using a model that proved accurate when compared to historic official flu activity data. Data current through March 5, 2012.



LinkedIn Maps [Share](#)



### More Top Picks for You

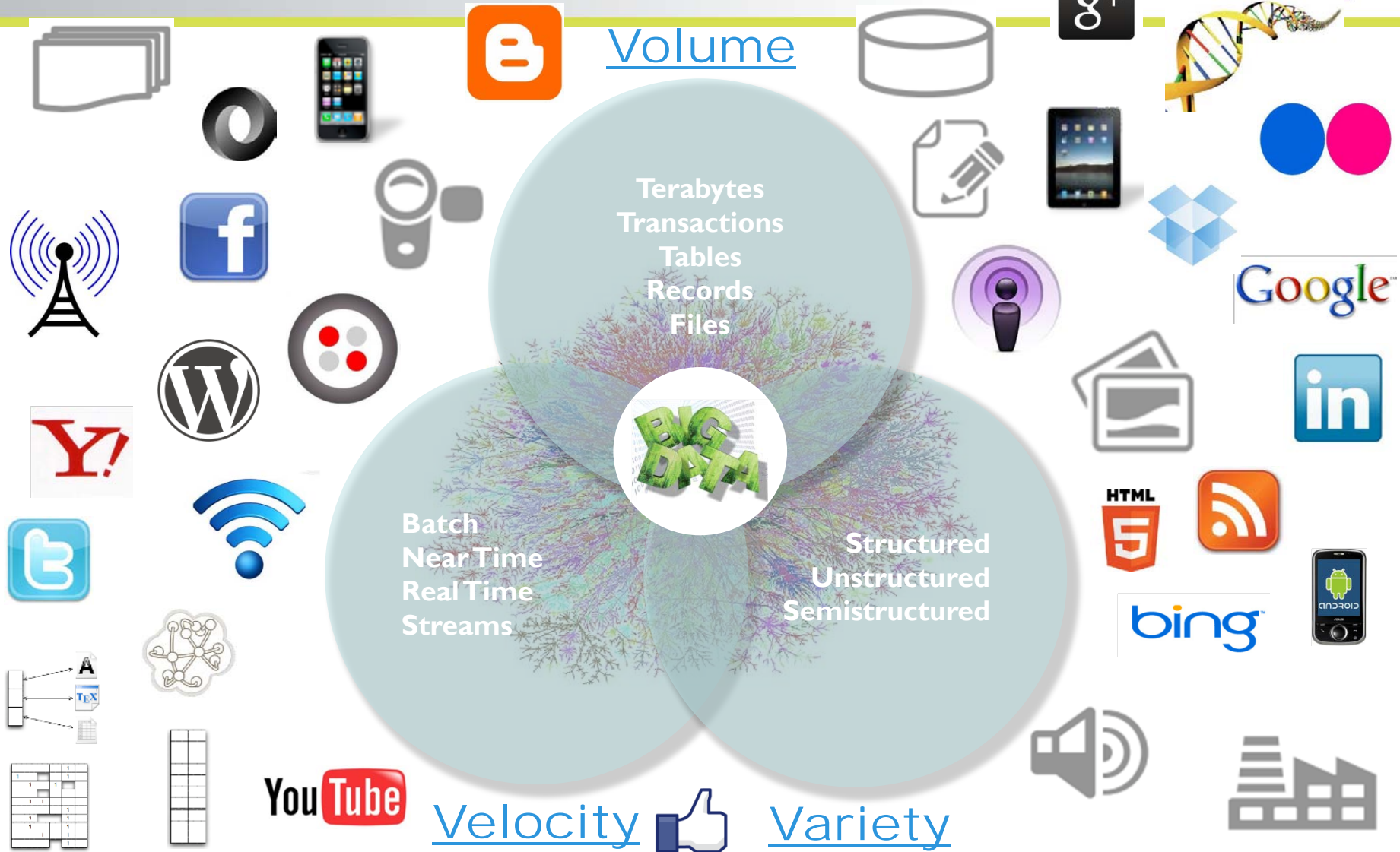
\$2.95	\$24.17	\$20.99	\$27.99	\$39.17	\$19.95

### More Items to Consider

\$19.37	\$15.18	\$14.88	\$13.50	\$23.10	\$16.13









# Attributes of Big Data



# Ten Common Big Data Problems

1. Modeling true risk
2. Customer churn analysis
3. Recommendation engine
4. Ad targeting
5. PoS transaction analysis
6. Analyzing network data to predict failure
7. Threat analysis
8. Trade surveillance
9. Search quality
10. Data "sandbox"

# The Big Data Opportunity

<b>Financial Services</b> 	<b>Healthcare</b> 
<b>Retail</b> 	<b>Web/Social/Mobile</b> 
<b>Manufacturing</b> 	<b>Government</b> 



## Retail

- CRM – Customer Scoring
- Store Siting and Layout
- Fraud Detection / Prevention
- Supply Chain Optimization



## Advertising & Public Relations

- Demand Signaling
- Ad Targeting
- Sentiment Analysis
- Customer Acquisition



## Financial Services

- Algorithmic Trading
- Risk Analysis
- Fraud Detection
- Portfolio Analysis



## Media & Telecommunications

- Network Optimization
- Customer Scoring
- Churn Prevention
- Fraud Prevention



## Manufacturing

- Product Research
- Engineering Analytics
- Process & Quality Analysis
- Distribution Optimization



## Energy

- Smart Grid
- Exploration



## Government

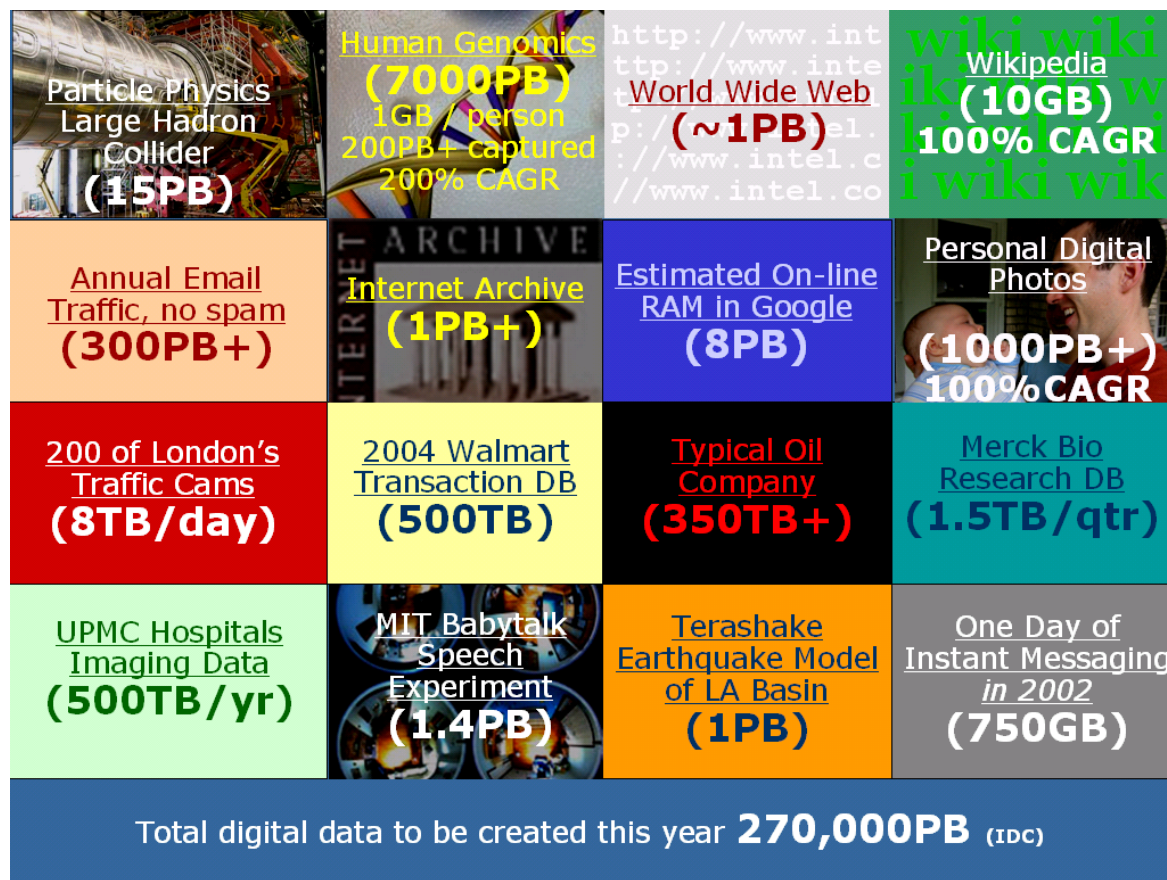
- Market Governance
- Counter-Terrorism
- Econometrics
- Health Informatics



## Healthcare & Life Sciences

- Pharmaco-Genomics
- Bio-Informatics
- Pharmaceutical Research
- Clinical Outcomes Research

# Why Hadoop?



Answer: Big Datasets!

# Why Hadoop?

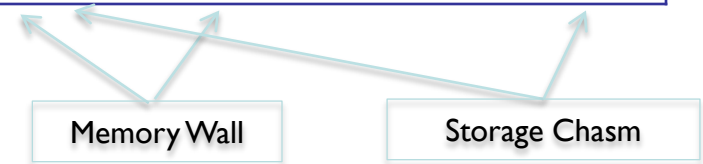
Big Data analytics and the Apache Hadoop open source project are rapidly emerging as the preferred solution to address business and technology trends that are disrupting traditional data management and processing.

**Enterprises can gain a competitive advantage by being early adopters of big data analytics.**

**Gartner**

# Storage & Memory B/W lagging CPU

	CPU	DRAM	LAN	Disk
Annual bandwidth improvement (all milestones)	1.5	1.27	1.39	1.28
Annual latency improvement (all milestones)	1.17	1.07	1.12	1.11



- CPU B/W requirements out-pacing memory and storage
- Disk & memory getting “further” away from CPU
- Large sequential transfers better for both memory & disk

# Commodity Hardware Economics

For **\$1000**  
One computer can

**Process**  
~32GB

**Store**  
~15TB

**99.9%**  
Of data is **Underutilized**

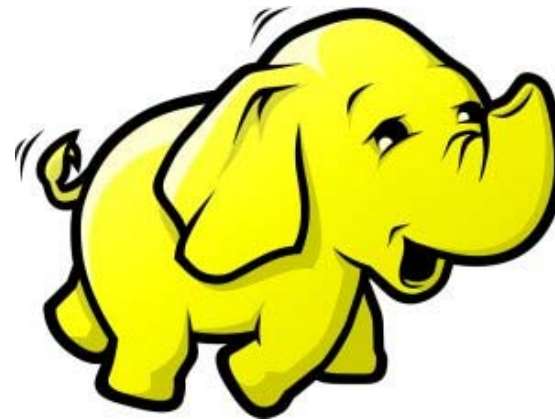


# Enterprise + Big Data = Big Opportunity



# WHAT IS HADOOP

Hadoop Adoption  
HDFS  
MapReduce  
Ecosystem Projects



# Hadoop Adoption in the Industry

2007

**YAHOO!**



Powerset  
NATURAL LANGUAGE SEARCH

**last.fm**

2008

**Google** **able grape**

**ImageShack** **Cascading**

**IBM** **facebook**

**ENORMO** Every property. Everywhere. **A9**

**THE UNIVERSITY OF EDINBURGH** **krugle** **rackspace HOSTING**

**Lookery** Control freaks welcome

**The New York Times** **Joost**

**Zvents** Discover Things To Do **FORMATION SCIENCES INSTITUTE**

**News Corporation**

**Cornell University** Computing and Information Science **Visible MEASURES**

**LOTAME** Locate, Target, & Message with Social Media **NetSeer**

**parc** Palo Alto Research Center **SECURITY ENHANCED DOMAIN NAME SYSTEM** **vech**

2009

**AOL** **cloudera**

**deepdyve** **cooliris**

**eyealike** **TEXTMAP** THE ENTITY SEARCH ENGINE

**PSG College of Technology** **iterend**

**tailsweep** **hulu**

**RapLeaf** **USCMS**

**Ning** **quxntcast**

**amazon web services** **pressflip**

**detikSearch** **WorldLingo**

**Systems@ETH** Zürich

**VK SOLUTIONS** Global Solutions Provider **TARAGANA** Innovation + Quality + Simplicity

**HOSTING HABITAT** **HOLA** SERVERS

**Terrier** **adknowledge**

**stampede** beta

2010

**SAMSUNG** **rubicon PROJECT**

**BERKELEY LAB** LAWRENCE BERKELEY NATIONAL LABORATORY **VISIBLE TECHNOLOGIES**

**APOLLO GROUP** **ADSDAQ**

**rackspace HOSTING** **RapLeaf**

**wordnik** All the words. **MOBILIGN** Mobile Network Population

**COMSCORE** **trulia** real estate search

**Accela COMMUNICATIONS** **Forward3D**

**LinkedIn** **Microsoft**

**Infochimps** Find the world's data **Pharm 2Phork**

**ADMELD** **gumgum** **BrafnPad**

**Pronux** The Datagraph Blog

**NETFLIX** **mobileanalytics.tv**

**markt24.de** **twitter**

**media6degrees** **BEEBLER**

**SLC Security** When Experience Matters... **ebay**

# What is Hadoop?



- A scalable fault-tolerant distributed system for data storage and processing
- Core Hadoop has two main components
  - ◆ Hadoop Distributed File System (HDFS): self-healing, high-bandwidth clustered storage
    - Reliable, redundant, distributed file system optimized for large files
  - ◆ MapReduce: fault-tolerant distributed processing
    - Programming model for processing sets of data
    - Mapping inputs to outputs and reducing the output of multiple Mappers to one (or a few) answer(s)
- Operates on unstructured and structured data
- A large and active ecosystem
- Open source under the friendly Apache License
  - ◆ <http://wiki.apache.org/hadoop/>



# HDFS 101

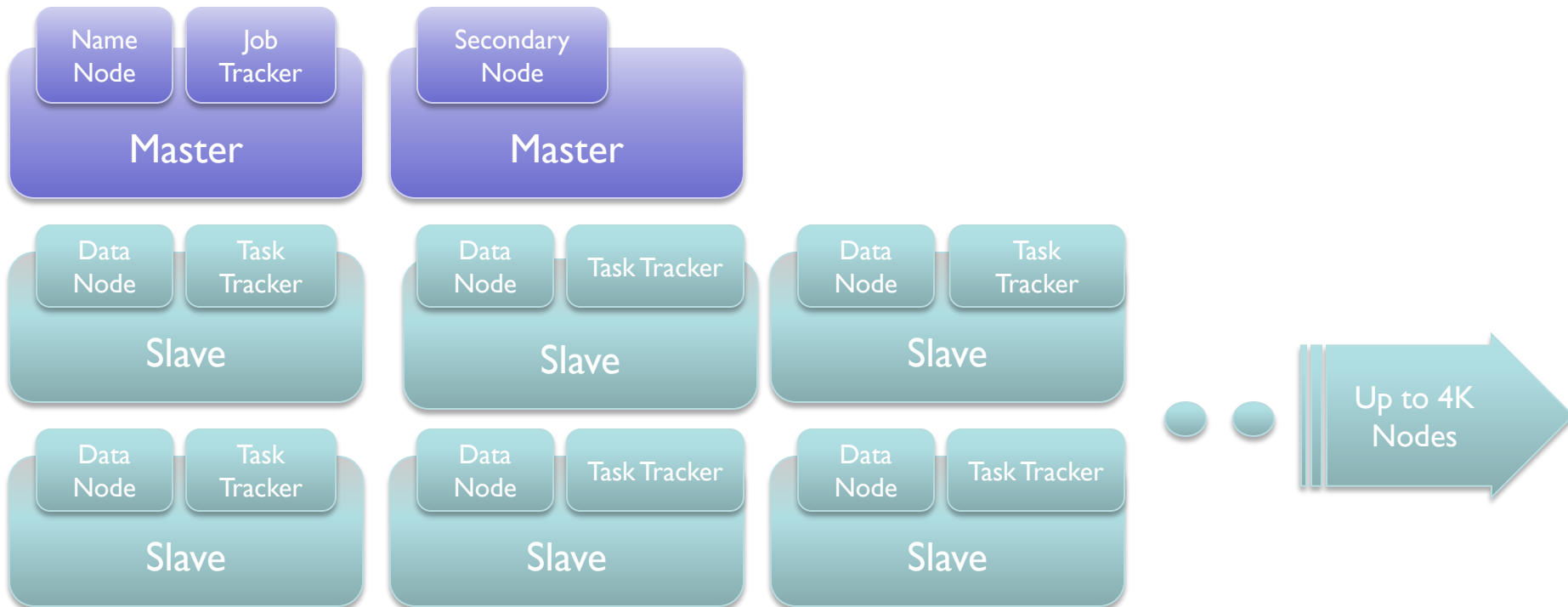
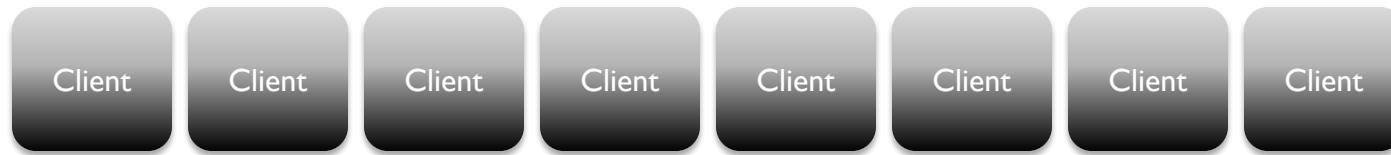
## The Data Set System



- Sits on top of a native (ext3, xfs, etc..) file system
- Performs best with a 'modest' number of large files
- Files in HDFS are 'write once'
- HDFS is optimized for large, streaming reads of files

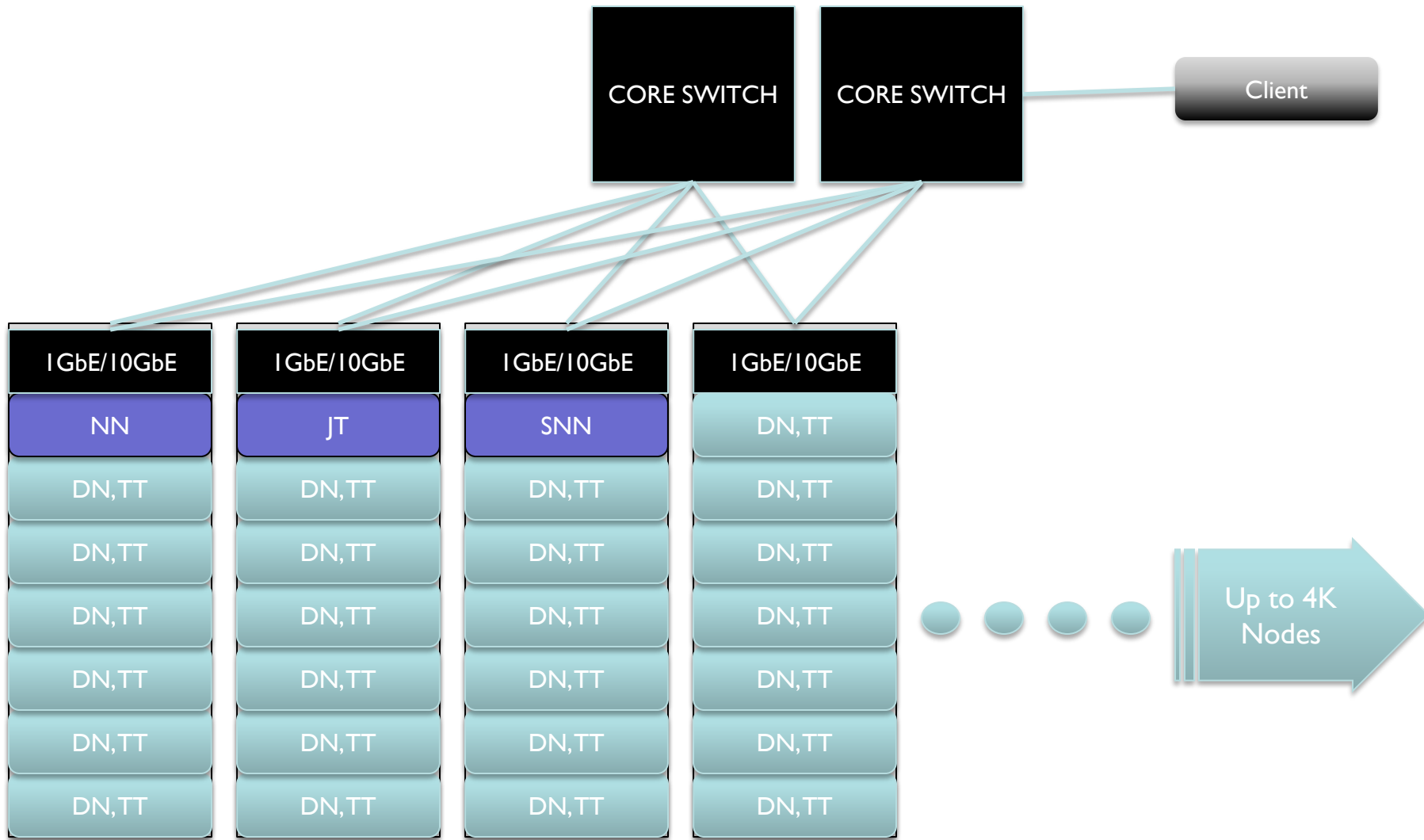
- Hadoop Distributed File System
  - Data is organized into files & directories
  - Files are divided into blocks, distributed across cluster nodes
  - Block placement known at runtime by map-reduce = computation co-located with data
  - Blocks replicated to handle failure
  - Checksums used to ensure data integrity
- Replication: one and only strategy for error handling, recovery and fault tolerance
  - Self Healing
  - Make multiple copies

# Hadoop Server Roles

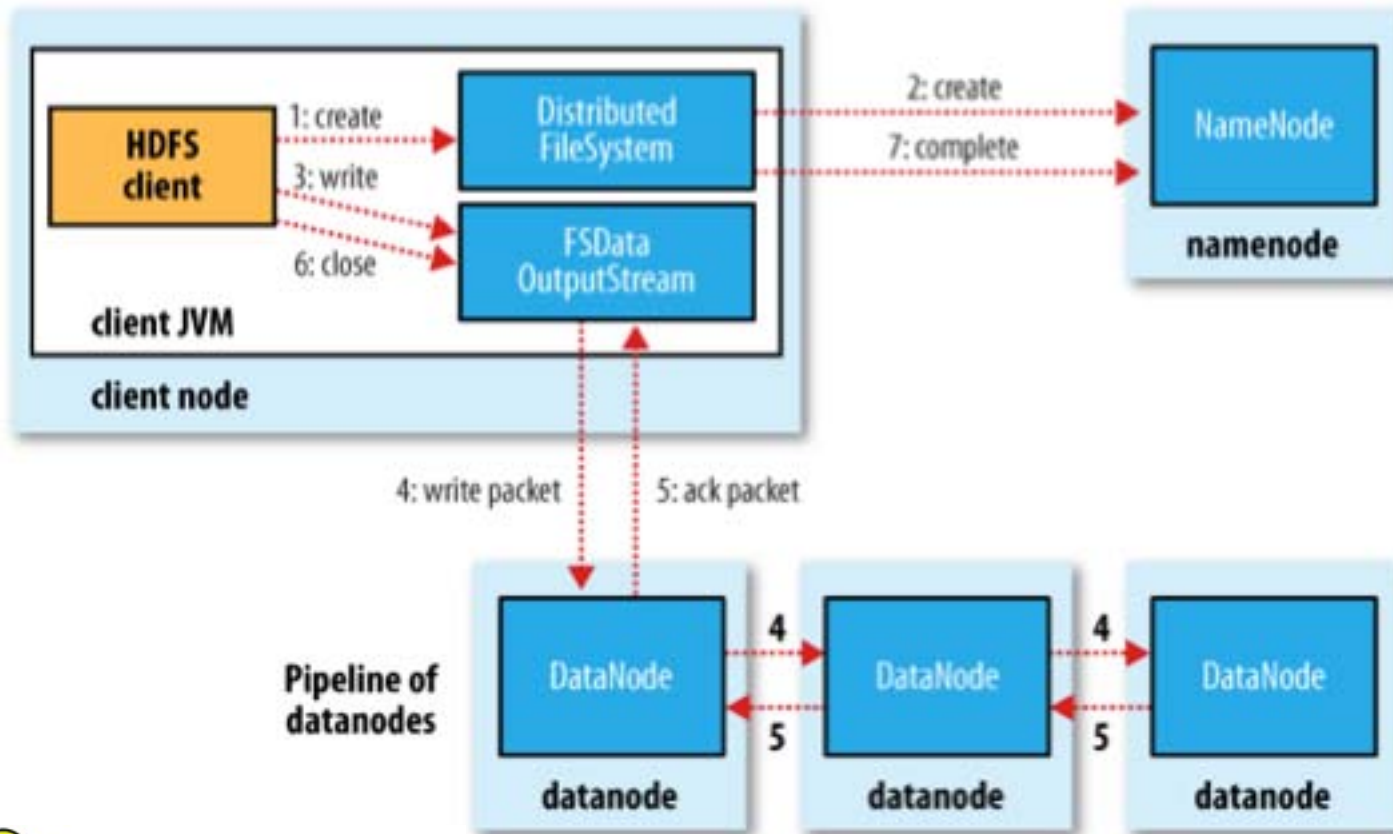




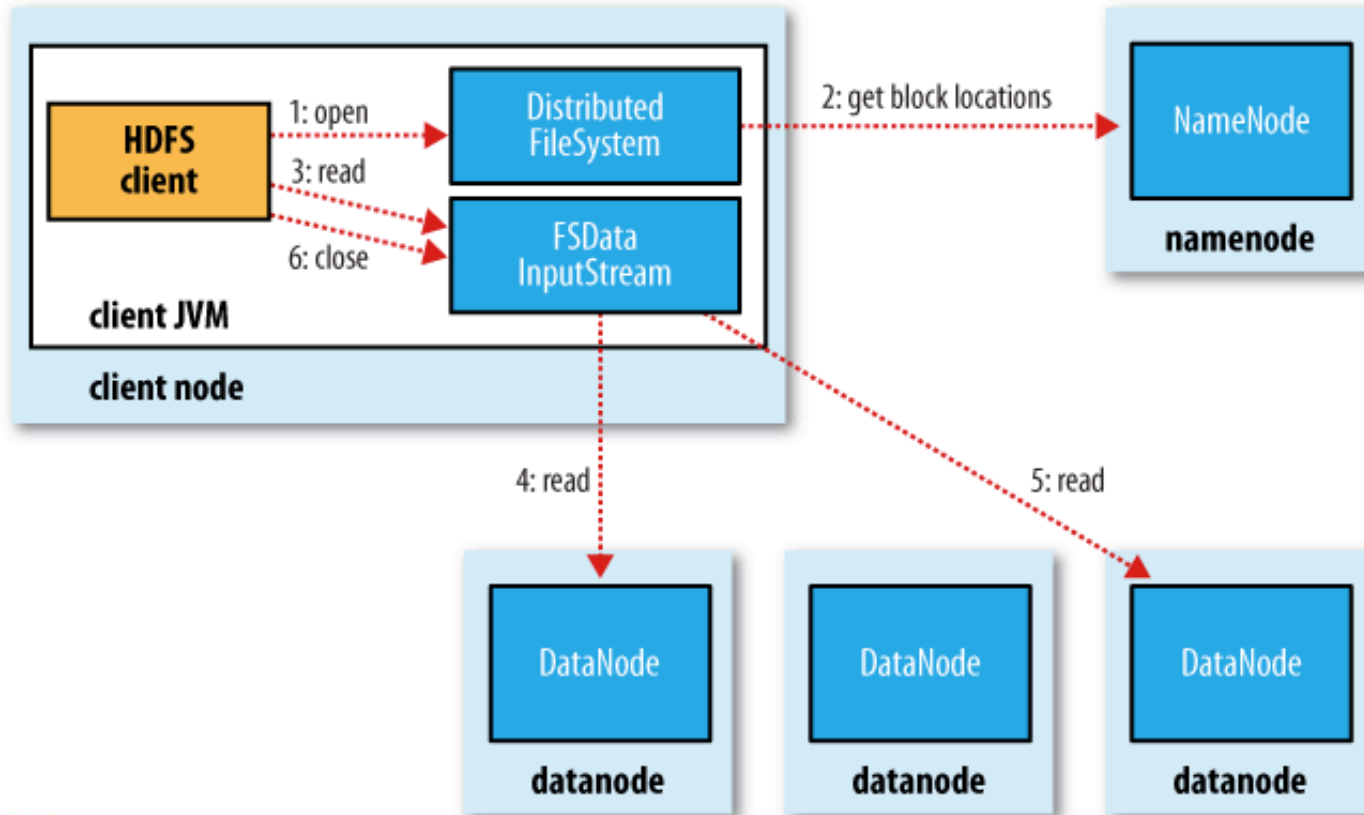
# Hadoop Cluster



# HDFS File Write Operation



# HDFS File Read Operation





# MapReduce 101

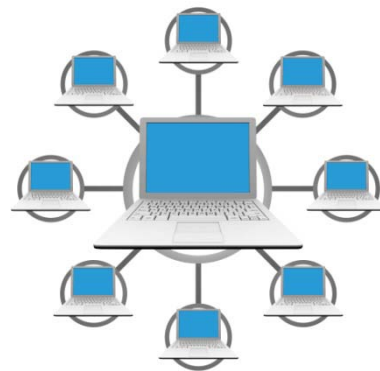
Functional Programming meets  
Distributed Processing

# What is MapReduce?

- A method for distributing a task across multiple nodes
- Each node processes data stored on that node
- Consists of two developer-created phases
  1. Map
  2. Reduce
- In between Map and Reduce is the Shuffle and Sort

# MapReduce Provides:

- Automatic parallelization and distribution
- Fault Tolerance
- Status and Monitoring Tools
- A clean abstraction for programmers
- Google Technology RoundTable: MapReduce

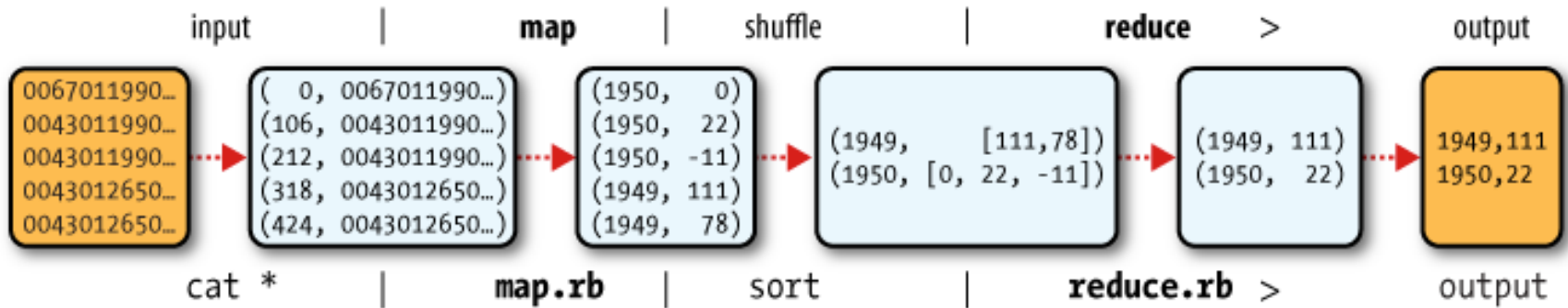


- A user runs a client program on a client computer
- The client program submits a job to Hadoop
- The job is sent to the JobTracker process on the Master Node
- Each Slave Node runs a process called the TaskTracker
- The JobTracker instructs TaskTrackers to run and monitor tasks
- A task attempt is an instance of a task running on a slave node
- There will be at least as many task attempts as there are tasks which need to be performed

- Each Mapper processes single input split from HDFS
- Hadoop passes developer's Map code one record at a time
- Each record has a key and a value
- Intermediate data written by the Mapper to local disk
- During shuffle and sort phase, all values associated with same intermediate key are transferred to same Reducer
- Reducer is passed each key and a list of all its values
- Output from Reducers is written to HDFS



# MapReduce Operation



What was the max/min temperature for the last century?

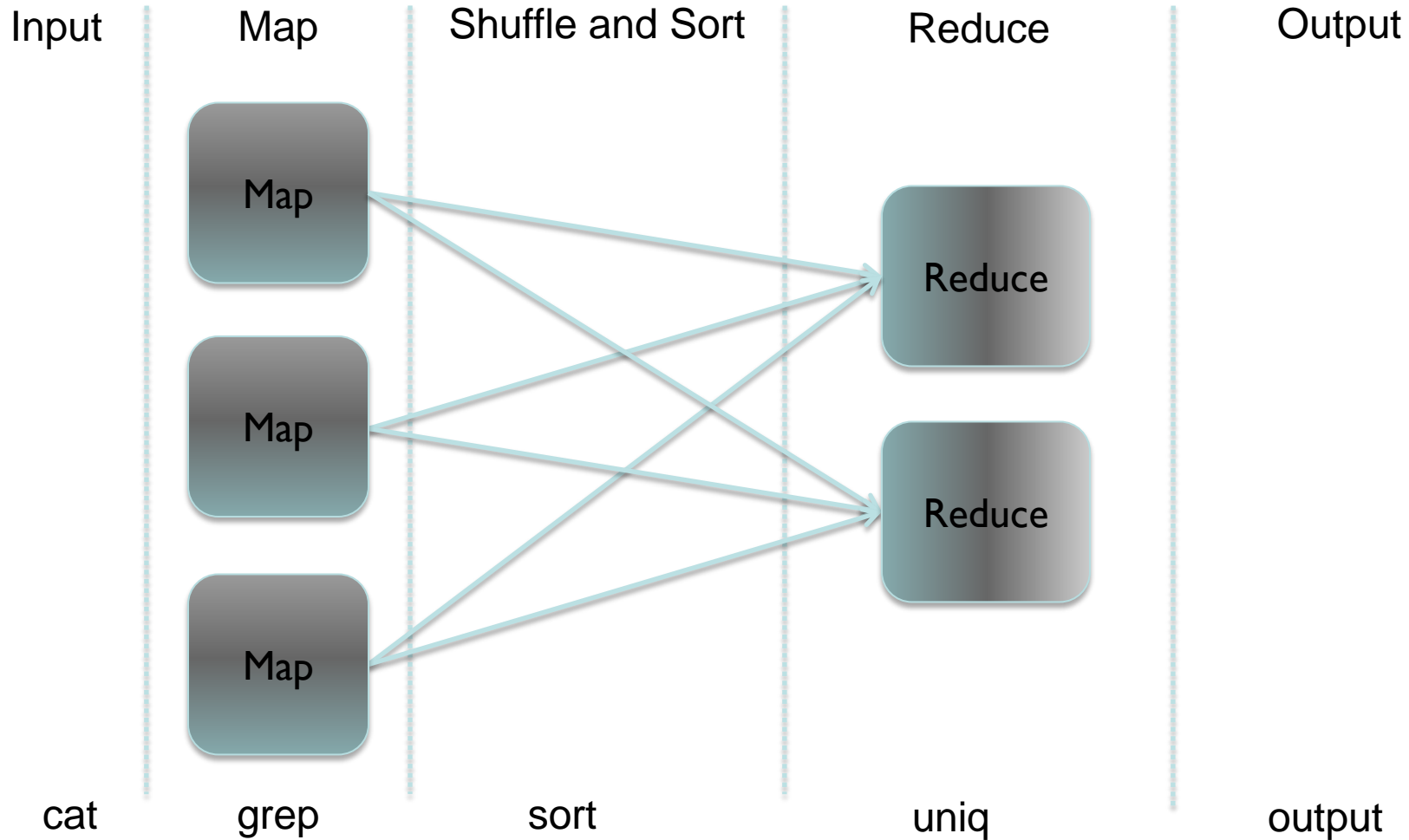
## ➤ The requirement:

- ◆ you need to find out grouped by type of customer how many of each type are in each country with the name of the country listed in the `countries.dat` in the final result (and not the 2 digit country name). Each record has a key and a value

## ➤ To do this you need to:

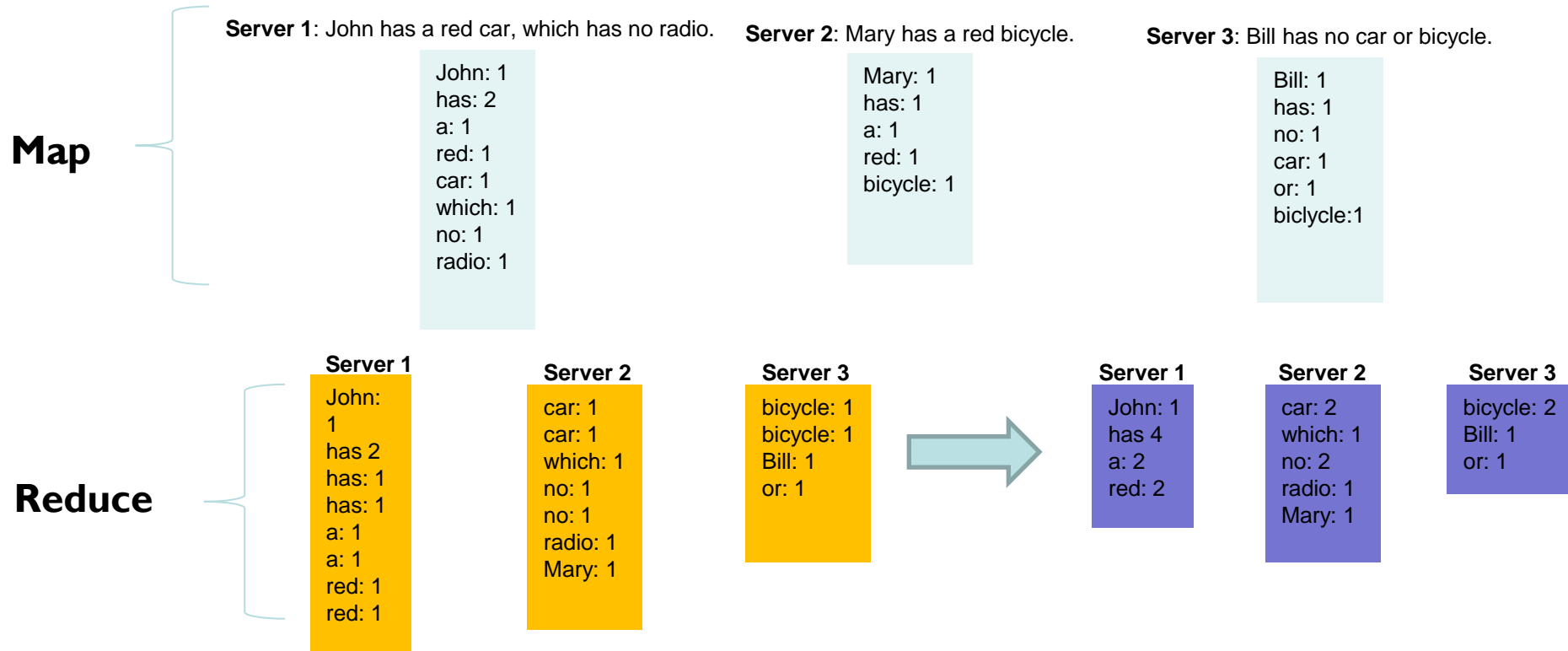
- ◆ Join the data sets
- ◆ Key on country
- ◆ Count type of customer per country
- ◆ Output the results

# MapReduce Paradigm

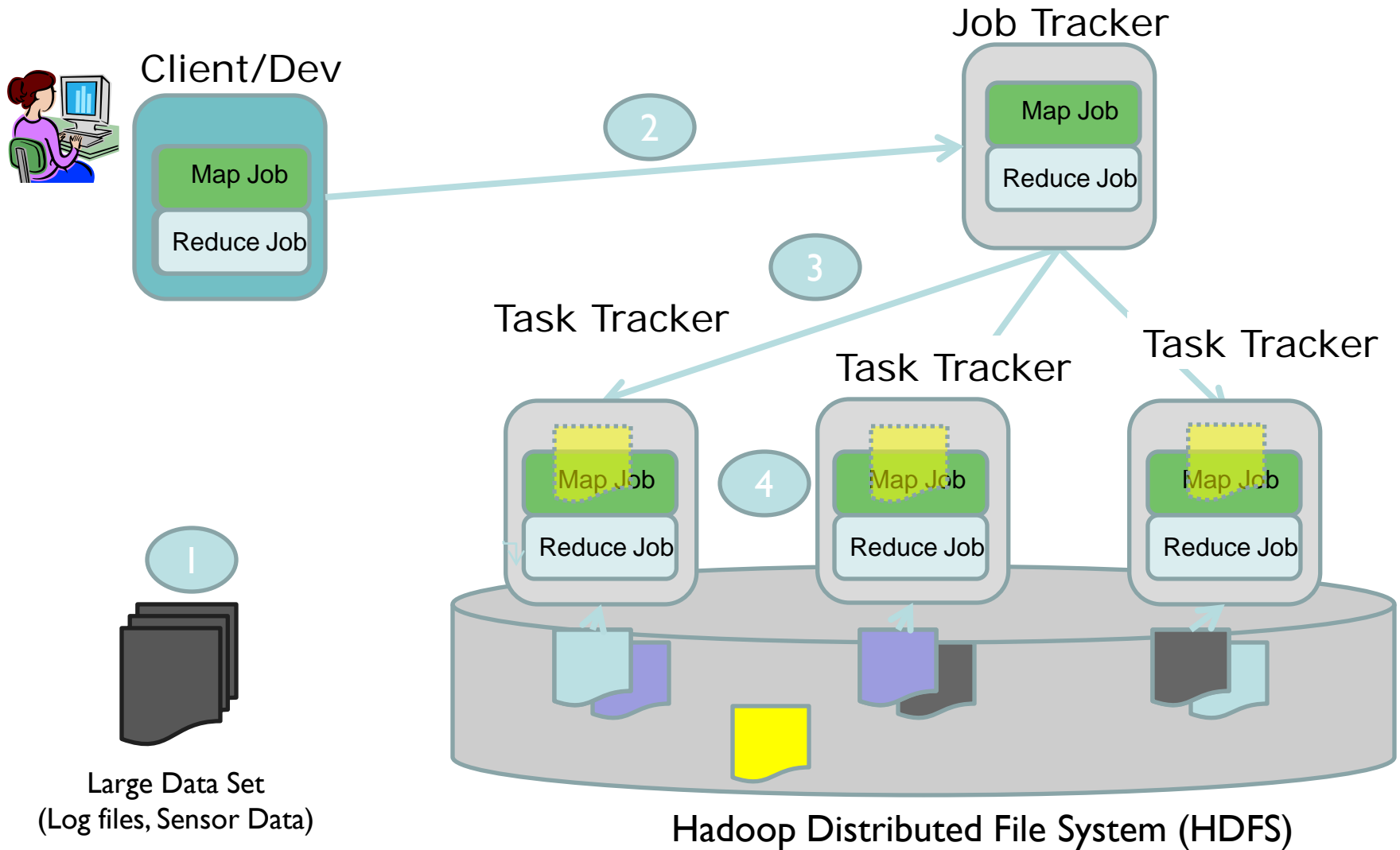


# MapReduce Example

**Problem:** Count the number of times that each word appears in the following paragraph:  
John has a red car, which has no radio. Mary has a red bicycle. Bill has no car or bicycle.



# Putting it all Together: MapReduce and HDFS



# Hadoop Ecosystem Projects

- Hadoop is a ‘top-level’ Apache project
  - Created and managed under the auspices of the Apache Software Foundation
- Several other projects exist that rely on some or all of Hadoop
  - Typically either both HDFS and MapReduce, or just HDFS
- Ecosystem Projects Include
  - Hive
  - Pig
  - HBase
  - Many more.....



# Hadoop, SQL & MPP Systems

<b>Hadoop</b>	<b>Traditional SQL Systems</b>	<b>MPP Systems</b>
Scale-Out	Scale-Up	Scale-Out
Key/Value Pairs	Relational Tables	Relational Tables
Functional Programming	Declarative Queries	Declarative Queries
Offline Batch Processing	Online Transactions	Online Transactions

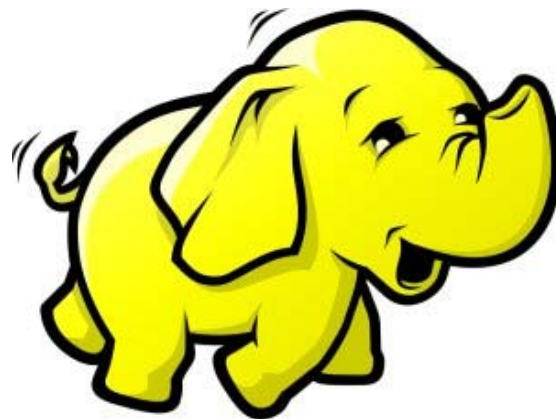
# Comparing RDBMS and MapReduce

	Traditional RDBMS	MapReduce
Data Size	Gigabytes ( <i>Terabytes</i> )	Petabytes ( <i>Exabytes</i> )
Access	Interactive and Batch	Batch
Updates	Read / Write many times	Write once, Read many times
Structure	Static Schema	Dynamic Schema
Integrity	High (ACID)	Low
Scaling	Nonlinear	Linear
DBA Ratio	1:40	1:3000

*Reference: Tom White's Hadoop: The Definitive Guide*



# Hadoop Use Cases



## ➤ Issues

- ◆ What make and model systems are deployed?
- ◆ Are certain set top boxes in need of replacement based on system diagnostic data?
- ◆ Is there a correlation between make, model or vintage of set top box and customer churn?
- ◆ What are the most expensive boxes to maintain?
- ◆ Which systems should we pro-actively replace to keep customers happy?

## ➤ Big Data Solution

- ◆ Collect unstructured data from set top boxes—multiple terabytes
- ◆ Analyze system data in Hadoop in near real time
- ◆ Pull data in to Hive for interactive query and modeling
- ◆ Analytics with Hadoop increases customer satisfaction

## ➤ Issues

- ◆ Fixed inventory of ad space is provided by national content providers. For example, 100 ads offered to provider for 1 month of programming
  - ◆ Provider can use this space to advertise its products and services, such as pay per view
  - ◆ Do we advertise “The Longest Yard” in the middle of a football game or in the middle of a romantic comedy?
  - ◆ 10% increase in pay per view movie rentals = \$10M in incremental revenue
- Big Data Solution
    - ◆ Collect programming data and viewer rental data in a large data repository
    - ◆ Develop models to correlate proclivity to rent to programming format
    - ◆ Find the most productive time slots and programs to advertise pay per view inventory
    - ◆ Improve ad placement and pay-per-view conversion with Hadoop

- Risk Modeling
  - Bank had customer data across multiple lines of business and needed to develop a better risk picture of its customers. i.e, if direct deposits stop coming into checking acct, it's likely that customer lost his/her job, which impacts creditworthiness for other products (CC, mortgage, etc.)
  - Data existing in silos across multiple LOB's and acquired bank systems
  - Data size approached 1 petabyte
- Why do this in Hadoop?
  - Ability to cost-effectively integrate + 1 PB of data from multiple data sources: data warehouse, call center, chat and email
  - Platform for more analysis with poly-structured data sources; i.e., combining bank data with credit bureau data; Twitter, etc.
  - Offload intensive computation from DW

- Sentiment Analysis
  - Hadoop used frequently to monitor what customers think of company's products or services
  - Data loaded from social media sources (Twitter, blogs, Facebook, emails, chats, etc.) into Hadoop cluster
  - Map/Reduce jobs run continuously to identify sentiment (i.e., Acme Company's rates are "**outrageous**" or "**rip off**")
  - Negative/positive comments can be acted upon (special offer, coupon, etc.)
- Why Hadoop
  - Social media/web data is unstructured
  - Amount of data is immense
  - New data sources arise weekly

- ◆ World Economic Forum: “Personal Data: The Emergence of a New Asset Class” 2011
- ◆ McKinsey Global Institute: Big Data: The next frontier for innovation, competition, and productivity
- ◆ Big Data: Harnessing a game-changing asset
- ◆ IDC: 2011 Digital Universe Study: Extracting Value from Chaos
- ◆ The Economist: Data, Data Everywhere
- ◆ Data Science Revealed: A Data-Driven Glimpse into the Burgeoning New Field
- ◆ O’Reilly – What is Data Science?
- ◆ O’Reilly – Building Data Science Teams?
- ◆ O’Reilly – Data for the public good
- ◆ Obama Administration “Big Data Research and Development Initiative.”

- ▶ Please send any questions or comments on this presentation to the SNIA at this address:  
[tracktutorials@snia.org](mailto:tracktutorials@snia.org)

**Many thanks to the following individuals  
for their contributions to this tutorial.**

*SNIA Education Committee*

**Denis Guyadeen  
Rob Peglar**