“Housekeeping”

- Welcome to today’s ACM Webinar. The presentation starts at the top of the hour.
- If you are experiencing any problems/issues, refresh your console by pressing the **F5** key on your keyboard in **Windows**, **Command + R** if on a **Mac**, or refresh your browser if you’re on a mobile device; or close and re-launch the presentation. You can also view the Webcast Help Guide, by clicking on the “Help” widget in the bottom dock.
- To control volume, adjust the master volume on your computer.
- If you think of a question during the presentation, please type it into the **Q&A** box and click on the submit button. You do not need to wait until the end of the presentation to begin submitting questions.
- At the end of the presentation, you’ll see a **survey** open in your browser. Please take a minute to fill it out to help us improve your next webinar experience.
- You can download a copy of these slides by clicking on the **Resources** widget in the bottom dock.
- This presentation is being recorded and will be available for on-demand viewing in the next 1-2 days. You will receive an **automatic e-mail notification** when the recording is ready.

Simplifying Big Data ...with Hadoop

Denny Lee, Senior Director, Data Platform
Concur Technologies

Association for Computing Machinery

Advancing Computing as a Science & Professior

ACM Learning Center

http://learning.acm.org

- 1,400+ trusted technical books and videos by leading publishers including O'Reilly, Morgan Kaufmann, others
- Online courses with assessments and certification-track mentoring, member discounts on tuition at partner institutions
- Learning Webinars on big topics (Cloud/Mobile Development, Cybersecurity, Big Data, Recommender Systems, SaaS, Agile, Machine Learning, Natural Language Processing, Parallel Programming, etc.)
- ACM Tech Packs on top current computing topics: Annotated Bibliographies compiled by subject experts
- Popular video tutorials/keynotes from ACM Digital Library, A.M. Turing Centenary talks/panels
- Podcasts with industry leaders/award winners

“Housekeeping”

- Welcome to today’s ACM Webinar. The presentation starts at the top of the hour.
- If you are experiencing any problems/issues, refresh your console by pressing the **F5** key on your keyboard in **Windows**, **Command + R** if on a **Mac**, or refresh your browser if you’re on a mobile device; or close and re-launch the presentation. You can also view the Webcast Help Guide, by clicking on the “Help” widget in the bottom dock.
- To control volume, adjust the master volume on your computer.
- If you think of a question during the presentation, please type it into the **Q&A** box and click on the submit button. You do not need to wait until the end of the presentation to begin submitting questions.
- At the end of the presentation, you’ll see a **survey** open in your browser. Please take a minute to fill it out to help us improve your next webinar experience.
- You can download a copy of these slides by clicking on the **Resources** widget in the bottom dock.
- This presentation is being recorded and will be available for on-demand viewing in the next 1-2 days. You will receive an **automatic e-mail notification** when the recording is ready.

Talk Back

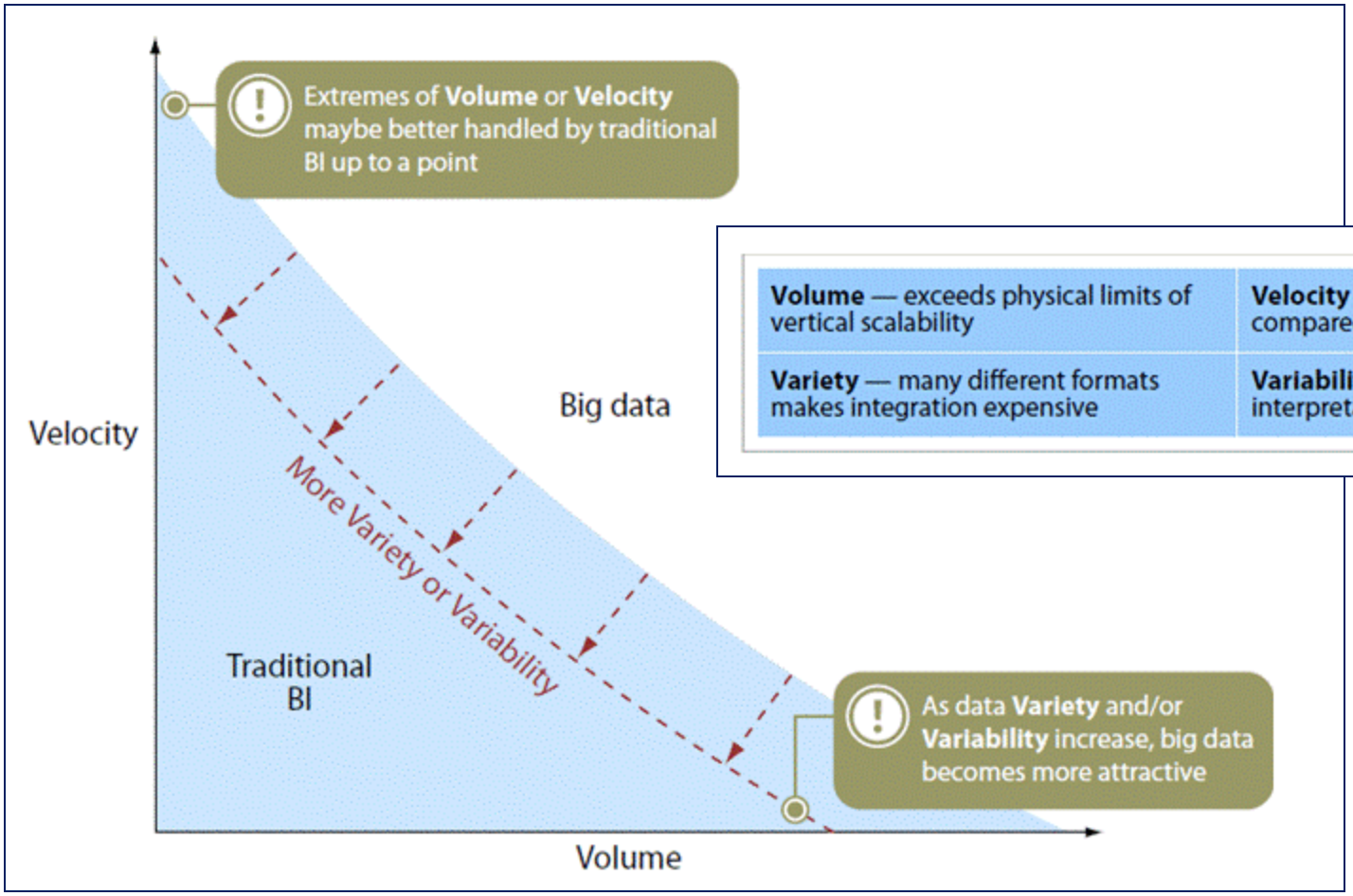
- Use the Facebook widget in the bottom panel to share this presentation with friends and colleagues
- Use Twitter widget to Tweet your favorite quotes from today's presentation with hashtag #ACMWebinarHadoop
- Submit questions and comments via Twitter to @acmeducation – we're reading them!

Large

Complex

Unstructured

What is Big Data?



Volume — exceeds physical limits of vertical scalability	**Velocity** — decision window small compared to data change rate
Variety — many different formats makes integration expensive	**Variability** — many options or variable interpretations confound analysis

A Definition of Big Data - 4Vs: volume, velocity, variability, variety

Big data: techniques and technologies that make handling data at extreme scale economical.

Scale Up!

With the power of the Hubble telescope, we can take amazing pictures 45M light years away

Amazing image of the Antennae Galaxies (NGC 4038-4039)

Analogous with scale up:

- non-commodity
- specialized equipment
- single point of failure*

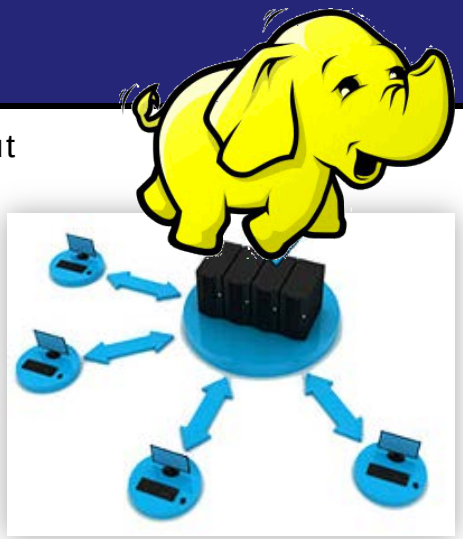
Scale Out | Commoditized Distribution

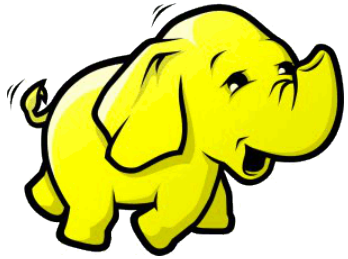
Hubble can provide an amazing view Giant Galactic Nebula (NGC 3503) but how about radio waves?

- Not just from one area but from *all areas viewed* by observatories
- SETI @ Home: 5.2M participants, 10^{21} floating point operations[1], 769 teraFLOPS[2]

Analogous with commoditized distributed computing

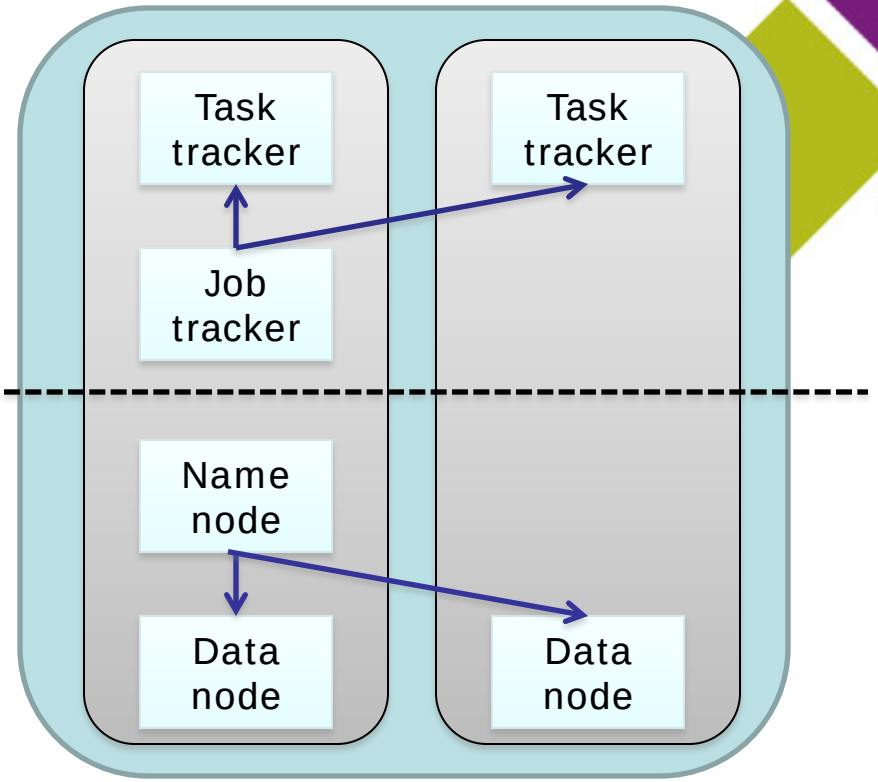
- Distributed and calculated locally
- Engage with hundreds, thousands, + machines
- Many points of failure, auto-replication prevents this from being a problem





Map Reduce Layer

HDFS Layer



Reference:
http://en.wikipedia.org/wiki/File:Hadoop_1.png

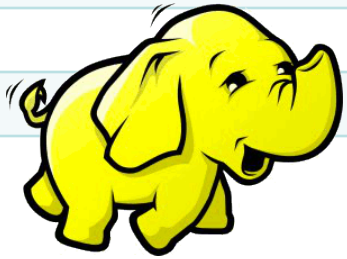
What is Hadoop?

- Synonymous with the Big Data movement
- Infrastructure to automatically distribute and replicate data across multiple nodes and execute and track map reduce jobs across all of those nodes
- Inspired by Google's Map Reduce and GFS papers
- Components are: Hadoop Distributed File System (HDFS), Map Reduce, Job Tracker, and Task Tracker
- *Based on the Yahoo! "Nutch" project in 2003, became Hadoop in 2005 named after Doug Cutting's son's toy elephant*

Comparing RDBMS and MapReduce

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

Reference: Tom White's
Hadoop: The Definitive Guide



example

- WordCount demo
- Map/Reduce
- HiveQL on a weblog sample

Sample Java MapReduce WordCount Function

```
// Map Reduce is broken out into a Map function and reduce function
// ----------------------------------------------------------------

// Sample Map function: tokenizes string and establishes the tokens
// e.g. “a b\tc\nd” is now an key value pairs representing [a, b, c, d]

public void map(Object key, Text value, Context context
                                ) throws IOException, InterruptedException {
        StringTokenizer itr = new StringTokenizer(value.toString());
        while (itr.hasMoreTokens()) {
                word.set(itr.nextToken());
                context.write(word, one);
        }
}

// Sample Reduce function: does the count of these key value pairs

public void reduce(Text key, Iterable<IntWritable> values,
                                        Context context
                                       ) throws IOException,
                                      InterruptedException {
        int sum = 0;
        for (IntWritable val : values) {
                sum += val.get();
        }
        result.set(sum);
        context.write(key, result);
}
```

Executing WordCount against sample file

The **Project** Gutenberg eBook of The Notebooks of Leonardo Da Vinci, Complete
by Leonardo Da Vinci
(#3 in our series by Leonardo Da Vinci)

Copyright **laws** are changing all over the world. Be sure to check the
copyright **laws** for your country before downloading or redistributing
this or any other **Project** Gutenberg eBook.

This header should be the first thing seen when viewing this **Project**
Gutenberg file. Please do not remove it. Do not change or edit the
header without written permission.

Please read the "legal small print," and other information about the
eBook and **Project** Gutenberg at the bottom of this file. Included is
important information about your specific rights and restrictions in
how the file may be used. You can also find out about how to make a
donation to **Project** Gutenberg, and how to get involved.

**Welcome To The World of Free Plain

**eBooks Readable By Both Humans and

*****These eBooks Were Prepared By T

Title: The Notebooks of Leonardo Da

Author: Leonardo Da Vinci

...

Code to execute:

```
hadoop jar AcmeWordCount.jar
AcmeWordCount /test/davinci.txt
/test/davinci_wordcount
```

Purpose: To perform count of number of words within the said davinci.txt

```
laws      2
Project   5
```

…

Query a Sample WebLog using HiveQL

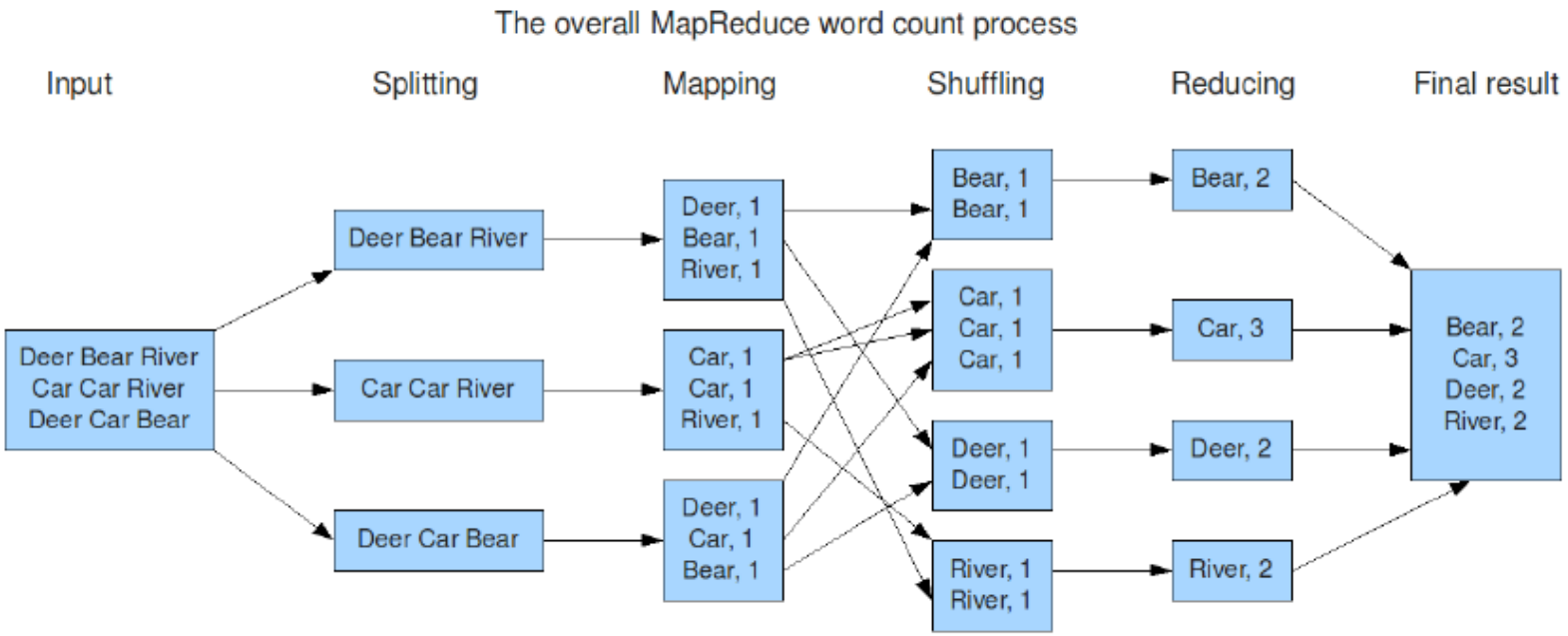
```
// Sample Generated Log
588.891.552.388,-,08/05/2011,11:00:02,W3SVC1,CTSSVR14,-,-,0,-
,200,-,GET,/c.gif,Mozilla/5.0 (Windows NT 6.1; rv:5.0)
Gecko/20100101 Firefox/5.0,http://foo.live.com/cid-
4985109174710/blah?fdkjafdf,[GUID],-,MSFT,
&PageID=1234&Region=89191&IsoCy=BR&Lang=1046&Referrer=hotmail.c
om&ag=2385105&Campaign=&Event=12034
```

GUID	PUID	Parameters
[GUID]	[PUID]	&PageID=1234&Region=89191&IsoCy=BR&Lang=1046&Referrer=hotmail.com&ag=2385105&Campaign=&Event=12034

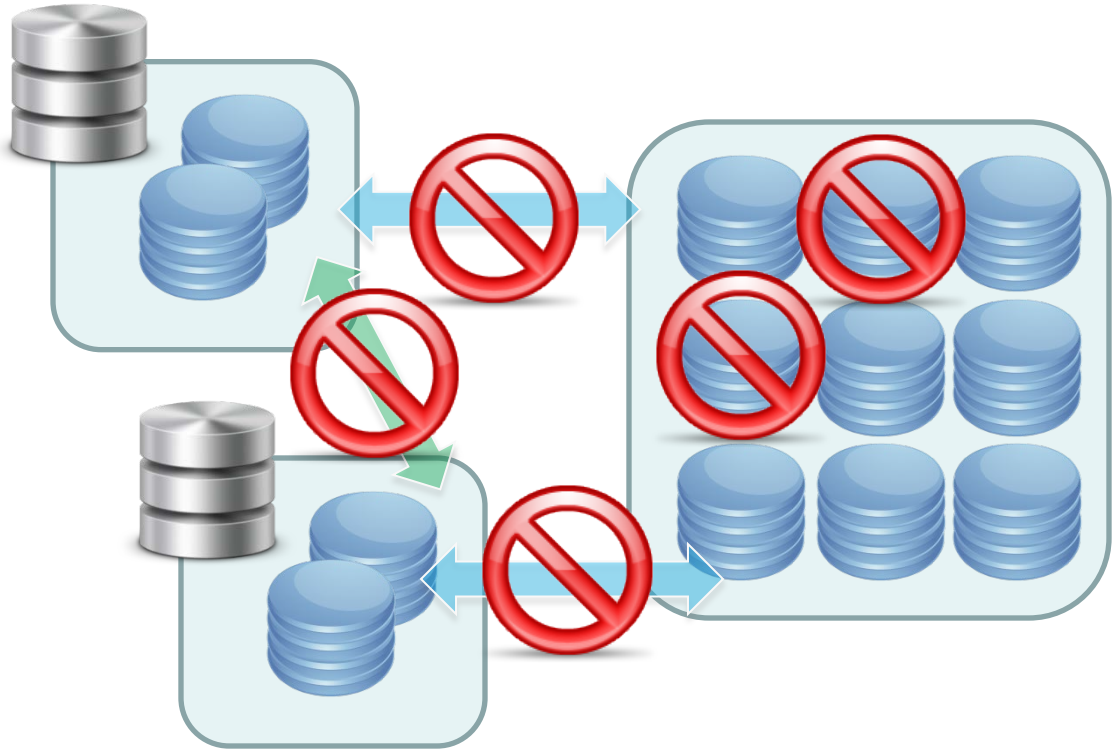
```
select
   GUID,
   str_to_map("param", "&", "=")['IsoCy'],
   str_to_map("param", "&", "=")['Lang']
from
weblog;
```

HiveQL: SQL-*like* language

- Write SQL-like query which becomes MapReduce functions
- Includes functions like **str_to_map** so one can perform parsing functions in HiveQL



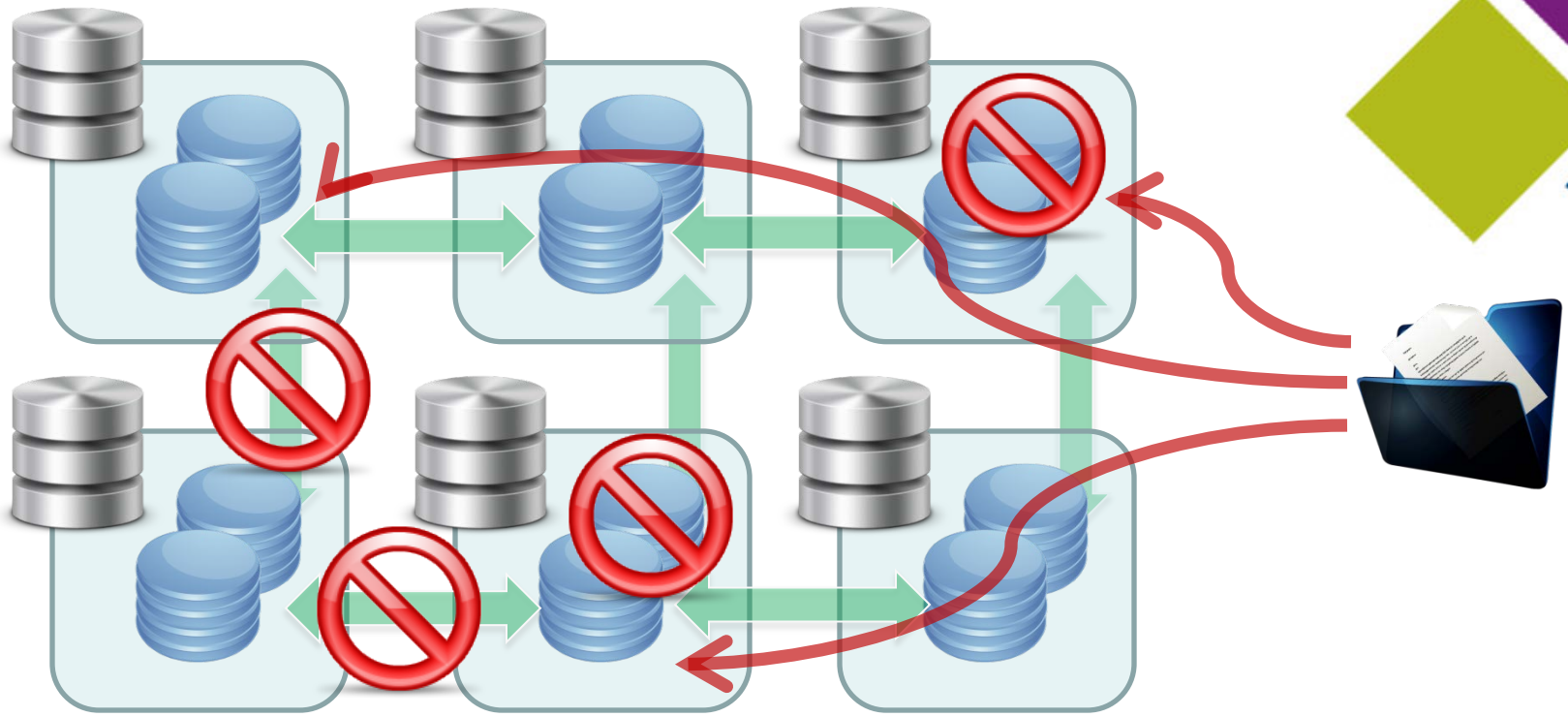
A great Map Reduce Tutorial can be found at: http://hci.stanford.edu/courses/cs448g/a2/files/map_reduce_tutorial.pdf



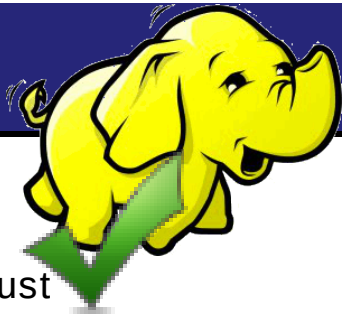
Traditional RDBMS: Move Data to Compute

As you process more and more data, and you want interactive response

- Typically need more expensive hardware
- Failures at the points of disk and network can be quite problematic
- It's all about ACID: atomicity, consistency, isolation, durability
- Can work around this problem with more expensive HW and systems
- Though distribution problem becomes harder to do



Hadoop / NoSQL: Move Compute to the Data

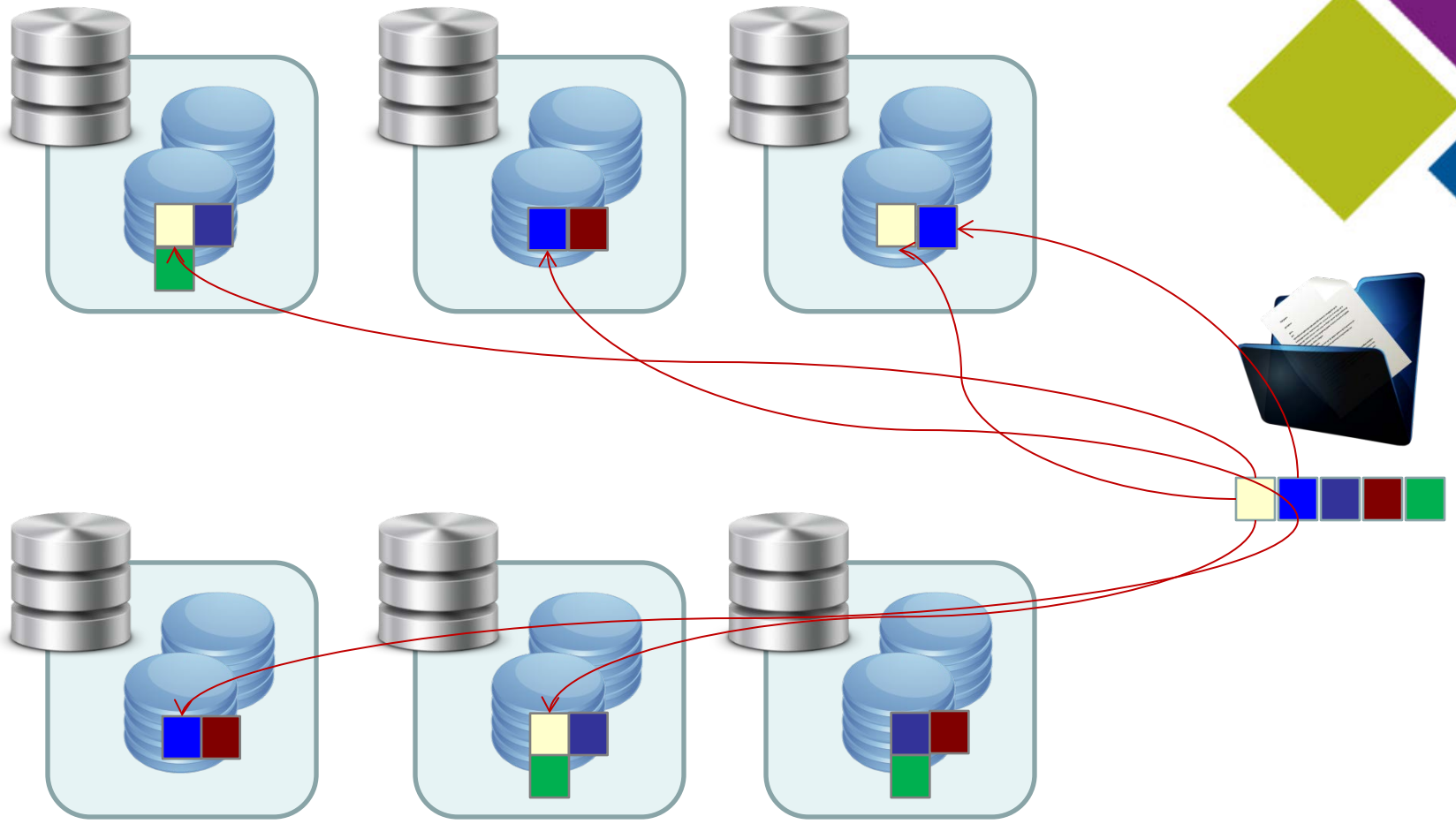


Hadoop (and NoSQL in general) follows the Map Reduce framework

- Developed initially by Google -> Map Reduce and Google File system
- Embraced by community to develop MapReduce algorithms that are very robust
- Built Hadoop Distributed File System (HDFS) to auto-replicate data to multiple nodes
- And execute a single MR task on all/many nodes available on HDFS

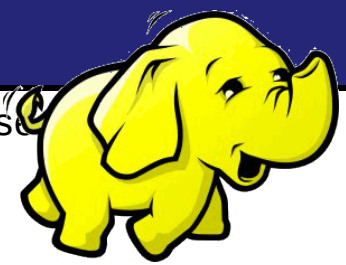
Use commodity HW: no need for specialized and expensive network and disk

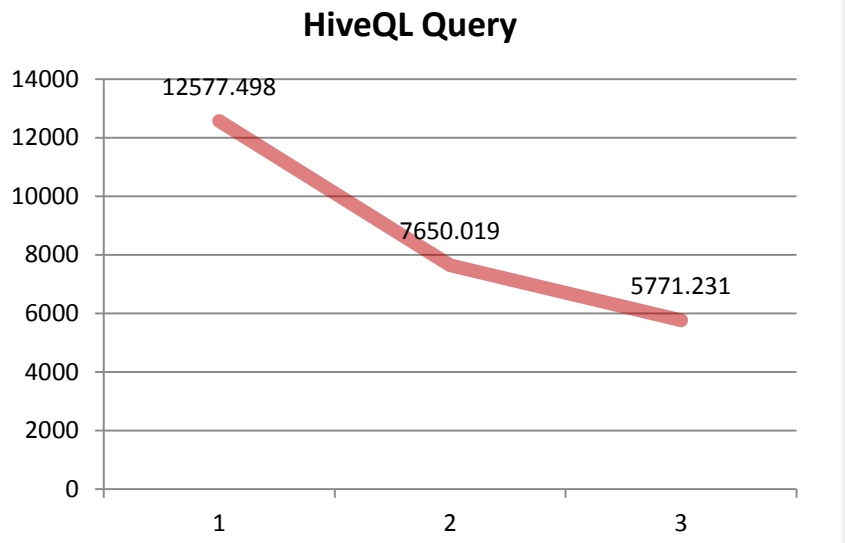
Not so much ACID, but BASE (basically available, soft state, eventually consistent)

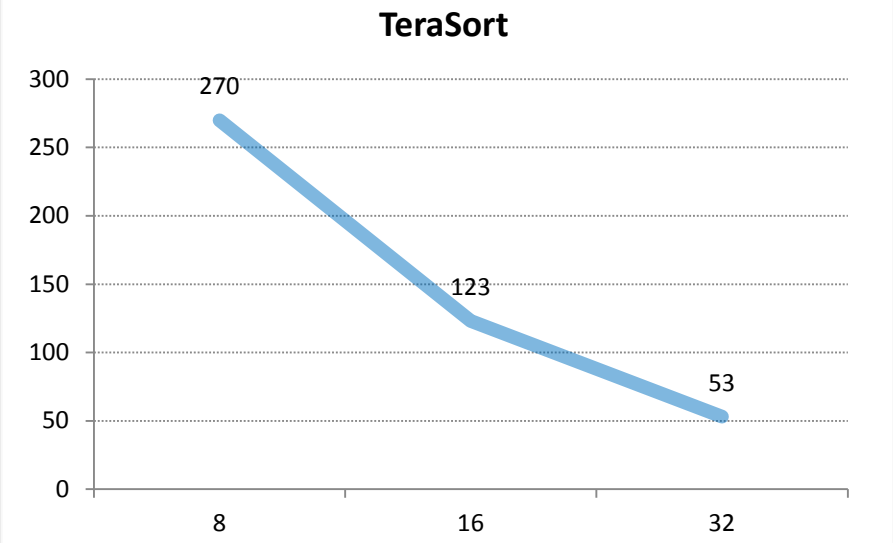


Hadoop: Auto-replication

- Hadoop processes data in 64MB chunks and then replicates to different se
- Replication value set at the hdfs-site.xml, dfs.replication node







Distribution, Linear Scalability

- Key facet is that as more nodes are added, the more can be calculated, faster (*minus overhead*)
- Designed for Batch systems (though this will change in the near future)
- Roughly linear scalability for HiveQL query as well as TeraSort

Cassandra
Hive
Scribe
Hadoop

Hadoop
Oozie
Pig (-latin)

BackType
Hadoop
Pig / Hbase
Cassandra

MR/GFS
Bigtable
Dremel
...

SimpleDB
Dynamo
EC2 / S3
...

NoSQL ecosystem | open source, commodity

Mahout	|	Scalable machine learning and data mining
MongoDB	|	Document-oriented database (C++)
Couchbase	|	CouchDB (doc dB) + Membase (memcache protocol)
Hbase	|	Hadoop column-store database
R	|	Statistical computing and graphics
Pegasus	|	Peta-scale graph mining system
Lucene	|	full-featured text search engine library

15 out of 17
sectors in the US have more data stored per company than the US Library of Congress

140,000-190,000
more deep analytical talent positions

1.5 million
more data savvy managers in the US alone

50-60%
increase in the number of Hadoop developers within organizations already using Hadoop within a year

€250 billion
Potential annual value to Europe's public sector

$300 billion
Potential annual value to US healthcare

NoSQL ecosystem | business value

How Facebook moved 30 petabytes of Hadoop data

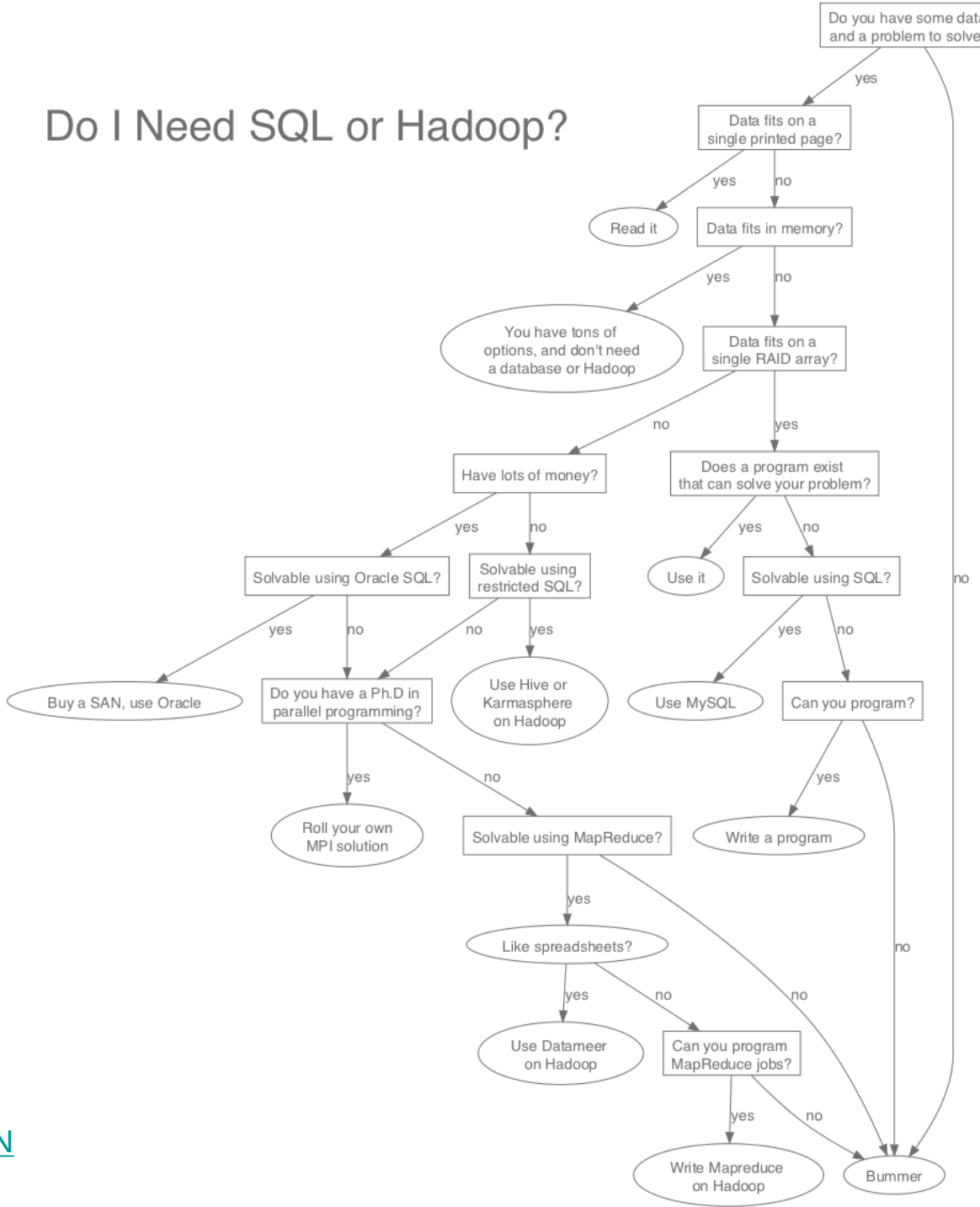
… the migration proved that disaster recovery is possible with Hadoop clusters. This could be an important capability for organizations considering relying on Hadoop (by running Hive atop the Hadoop Distributed File System) as a data warehouse, like Facebook does. As Yang notes, "Unlike a traditional warehouse using SAN/NAS storage, HDFS-based warehouses lack built-in data-recovery functionality. We showed that it was possible to efficiently keep an active multi-petabyte cluster properly replicated, with only a small amount of lag."

This illustrates a new thesis or collective wisdom emerging from the Valley: If a technology is not your core value-add, it should be open-sourced because then others can improve it, and potential future employees can learn it. This rising tide has lifted all boats, and is just getting started.

Kovas Boguta: Hadoop & Startups: Where Open Source Meets Business Data

APPENDIX

Do I Need SQL or Hadoop?



http://bit.ly/1jOfq7N

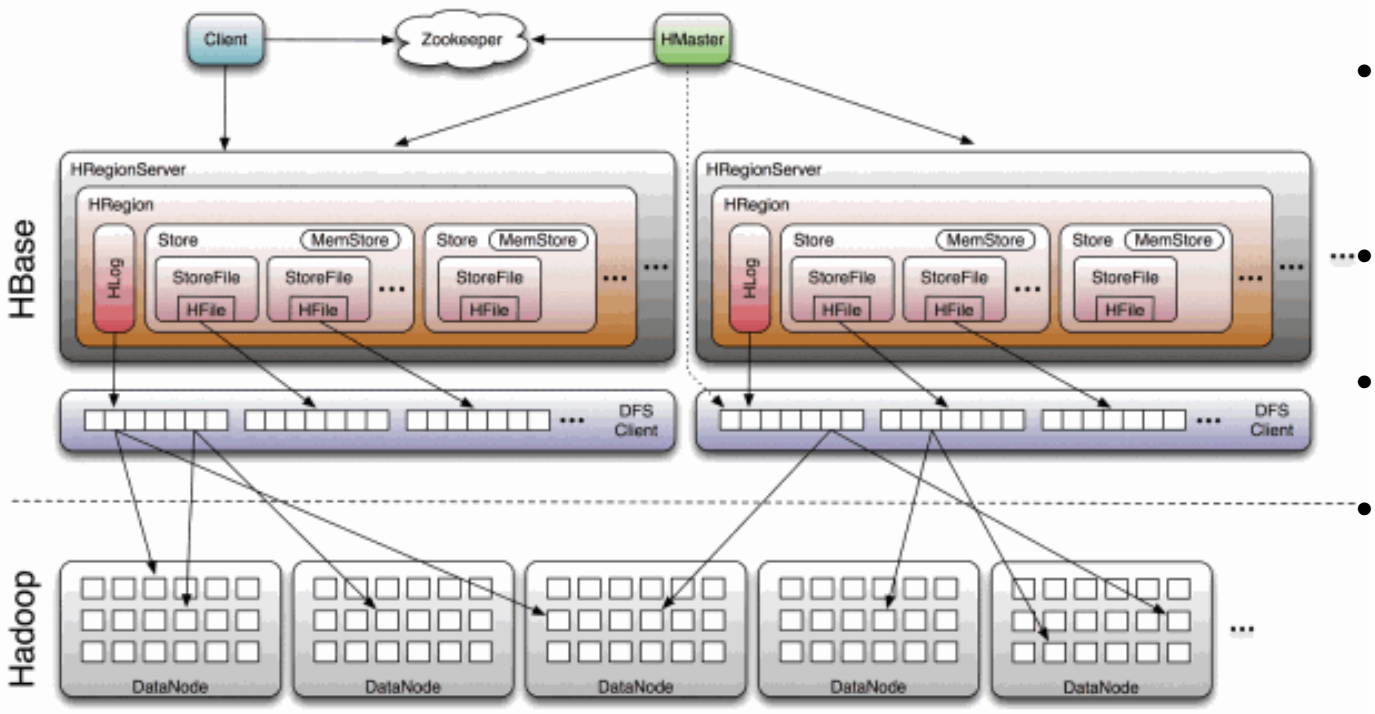
To Hadoop...and Beyond!

- Hadoop: The Definitive Guide
- Apache Hadoop YARN: Moving beyond MapReduce and Batch Processing with Apache Hadoop 2
- Impala: http://impala.io
- Spark: http://spark.apache.org
- Kafka: https://kafka.apache.org
- Mesos: http://mesos.apache.org

NoSQL

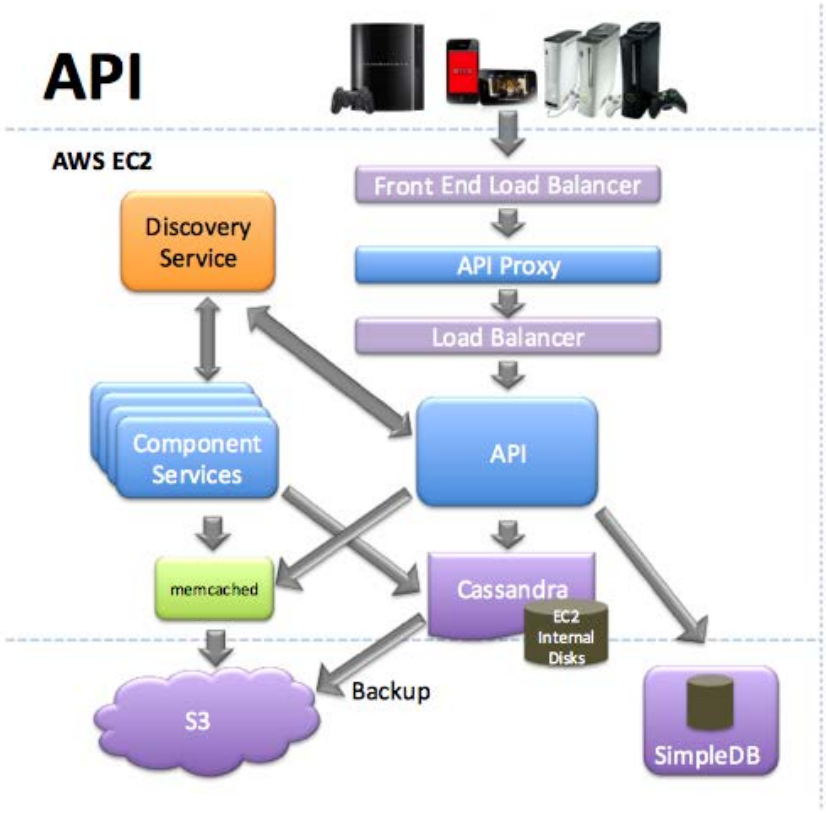
- As David DeWitt described in the PASS 2011 Day 3 Keynote, the NoSQL is analogous to OLTP (as Hadoop is to BI)
- Comprised of many components: HBase, Cassandra, MongoDB, CouchBase, MemcacheD, etc.
- Implementations of Google's BigTable – distributed storage system for managing structured data at very large sizes

HBase



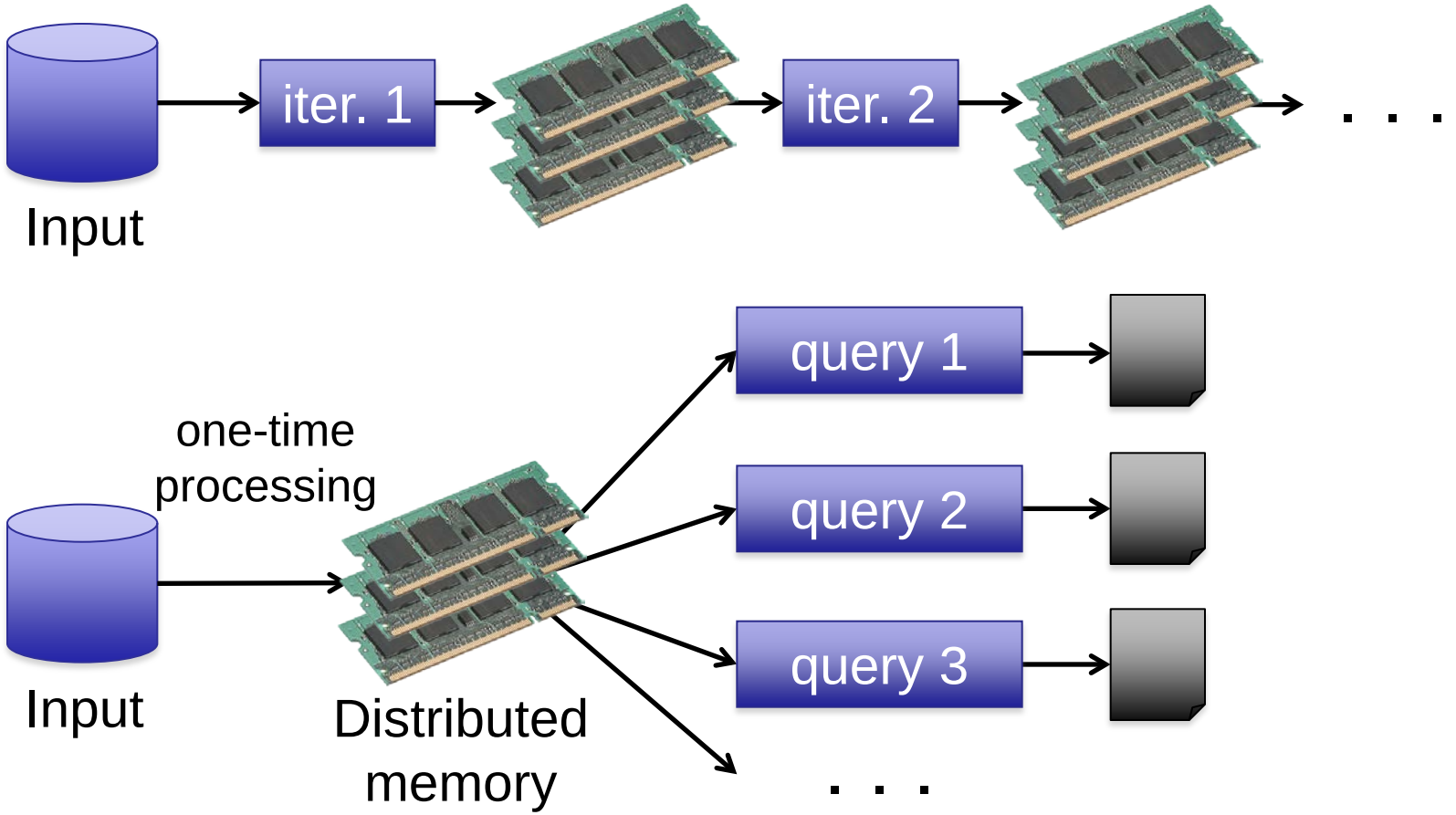
- Efficient at Random Reads/Writes
- Distributed, large scale data store
- Utilizes Hadoop for persistence
- Both HBase and Hadoop are distributed

Cassandra (Netflix)



- Reference: Migrating Netflix from Oracle to Global Cassandra (http://www.slideshare.net/adrianco/migrating-netflix-from-oracle-to-global-cassandra)
- Migrate from Oracle to Cassandra
- Move to the Cloud (AWS in this case)
- Showcase on how Big Data and Cloud work hand-in-hand
- Cassandra is based on Google's GFS and Amazon's Dynamo papers
- Does NOT require Hadoop

Spark



10-100× faster than network and disk


Q&A

ACM: The Learning Continues…

- Questions about this webcast? learning@acm.org
- ACM Learning Webinars (on-demand archive): http://learning.acm.org/webinar
- ACM Learning Center: http://learning.acm.org
- ACM Queue: http://queue.acm.org/
- ACM SIGKDD: http://www.kdd.org/
- ACM Tech Pack on Business Intelligence and Data Management: http://techpack.acm.org/bi/