THiNK BIG

ANALYTICS

Douglas Moore
Principal Consultant, Architect

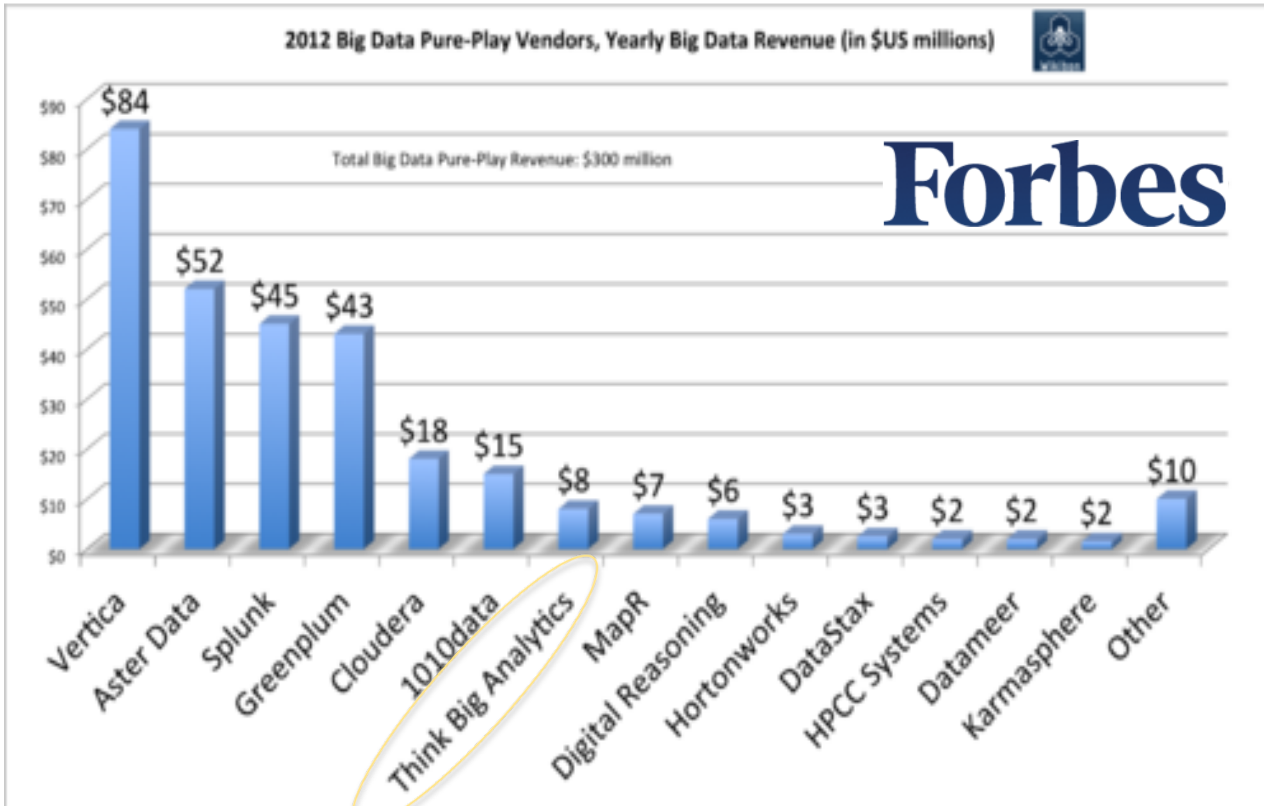
Big Data / Hadoop / NoSQL Agenda

- Who, Why?
- Data Processing Models
- Integration
- Common Uses
- Futures
- Summary

THiNK BIG ANALYTICS

Industry-Leading Big Data Solution Integrators

Big Data: $50 Billion Market by 2017



Source: Forbes February 2012

Think Big Recognized as a Top Pure-Play Big Data Vendor

100% Focus on Big Data consulting & Data Science solution services

Management Background:

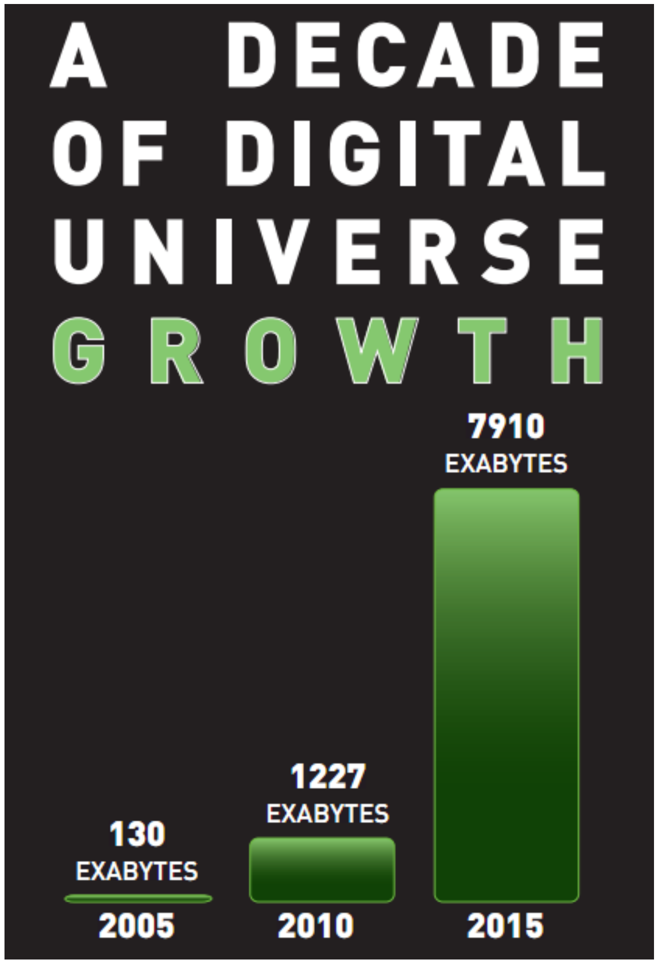
✓Cambridge Technology, C-bridge, Oracle, Sun Microsystems, Quantcast, Accenture

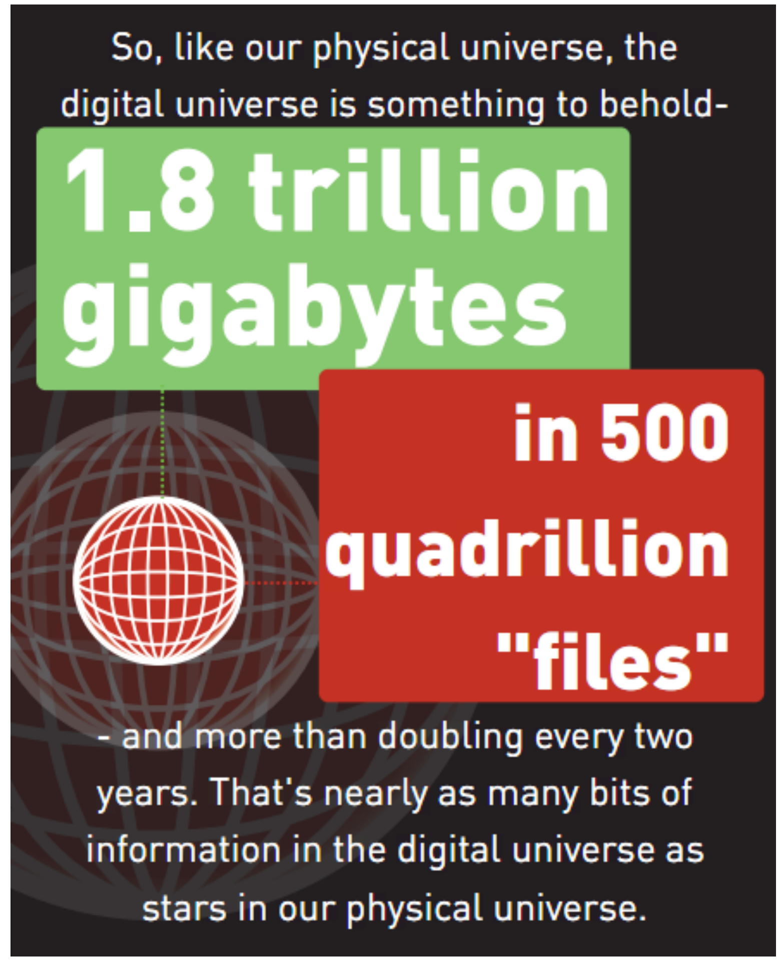
✓C-bridge Internet Solutions (CBIS) founder 1996 & executives, IPO 1999

THiNK BIG ANALYTICS

Industry-Leading Big Data Solution Integrators

THE 2011 IDC DIGITAL UNIVERSE STUDY sponsored by EMC





Source IDC

How Did Hadoop & MapReduce evolve?

Why Hadoop?	Online Industry: 2005-2008	Today: All Industries

Unstructured Data Analytics, Search & Recommendation for:

- Click Stream
- Log files
- Text
- Voice
- Pictures
- Video
- Docs
- Sensor Logs

Google
quantcast
YAHOO!
amazon.com
facebook
ebay

New Data Sources, Innovative Use Cases, Data Science & Predictive Analytics

+

Compute Processing $ & Time

ex. 26 Days→ 2 min

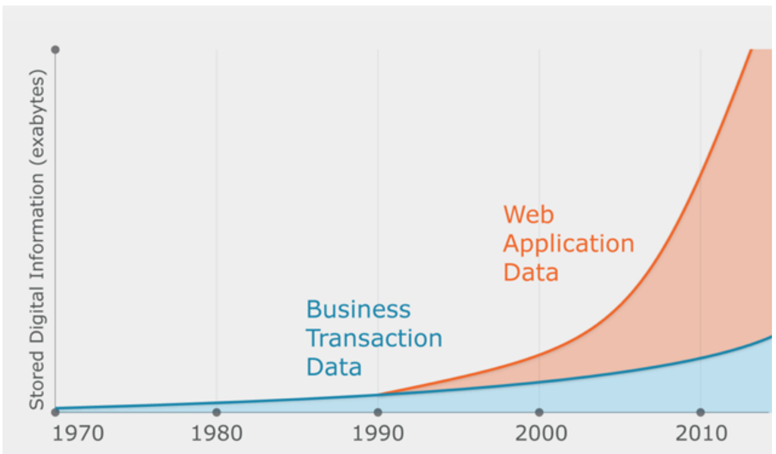
ex. 42 Hours → 40 min

ex. 18 Hours → 16 min

=

Business Innovation Velocity

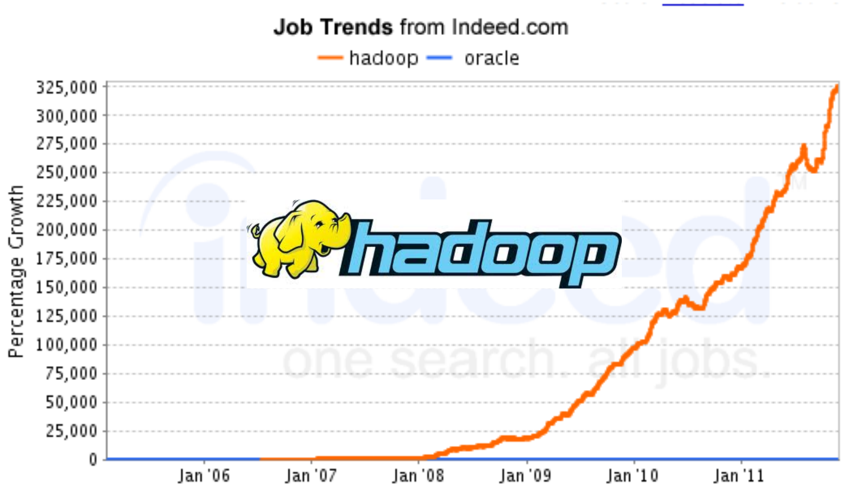
Big Data Growth Indicators



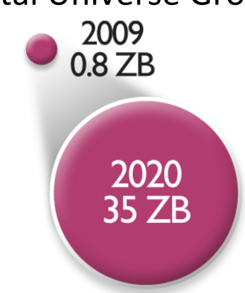
Scale: Absolute - Relative

Indeed.com searches millions of jobs from thousands of job sites.
This job trends graph shows relative growth for jobs we find matching your search terms.

Find SAS jobs, Mahout jobs



Digital Universe Growth



Growth Projections
Population
Mobile Phone
Machine Data

Source: IDC Digital Universe Study, sponsored by EMC, May 2010

THINK BIG ANALYTICS

Industry-Leading Big Data Solution Integrators

Real Companies:

- Enterprises 2008-2012 embracing Big Data
- Risk, Fraud, Acquisition, Products, Revenue Up-lift
- Creating Hadoop Big Data capabilities

Beyond unstructured

- Massive structured data
- Very large compute resources
- Unstructured data

Rapid Innovation:

- Data-driven: massive data not sample data
- Data-driven portfolio and product

VISA		scoring of individuals NOT groups.
JPMorganChase	2009	• 360° View of Customer • Analytics as a service to LOBs • Digital User Profiling and Segmentation • Digital User Lifecycle Management • Back End processing and analysis
Bank of America	2009	• Cross-Enterprise Risk for all LOBs • Massive Data analysis and processing
PayPal	2010	• Acquires Where for Mobile wallet and mobile advertising • GPS Location Data Services • Hyper-Local Analytics for Consumers and
ZIONS BANK	2011	
	2011	

Financial Services Big Dat	
Big Data Warehouse for B	
Ingestion and Streaming /	
IT Security Forensics	
Insight Lab, Data Science,	
Search - Structure, Semi S	
Right Time Ad Targeting +	
Geo Location + Hyper Loc	
Social Sentiment Analysis	
Advanced Segmentation	
360 Customer View with	
Loyalty Analysis - Brand I	
Customer Service Insights	
Recommendation Engine	
Risk Modeling - Bust Out	
Customer Churn Analysis	
Fraud Monitoring	◕
ML / Predictive Analytics w/ Actions	●

Big Data Use Case	Digital Adv
Big Data Warehouse for Batch Analysis & Compute	●
Simple Segmentation	◑
PII Structured Data Integration	◑
Offline Data Integration	◕
Scoring	◕
Campaign Perfomance	◑
Online Data / GEO & Social	●
Optimization	◕
Recommendation Engine	◕
Adv. Segmentation & Clone-a-like	●
X Channel Attribution	◕
Unstructured Data at Scale	●
Dynamic Content	◕
Dynamic Pricing	◕
Customer Lifetime Modeling	◕
Streaming Ingestion Analytics	●
ML / Predictive Analytics w/ Actions	●

Venture capital sees big returns in big data

Recommend 50 people recommend this. Be the first of your friends.

By Sarah McBride

SAN FRANCISCO | Fri Feb 17, 2012 3:03pm EST

Tweet 297

GIGAOM

HOME APPLE BROADBAND CLEANTECH CLOUD COLLABORATION

STRUCTURE

Obama's big data plans: Lots of cash and lots of open data

WSJ BLOGS

Venture Capital Dispatch

An inside look from VentureWire at high-tech start-ups and their investors.

November 8, 2011, 8:00 AM

Accel Makes Big Commitment To Big Data With $100M Fund

Pure Play Vendors

cloudera, MAPR Technologies, DataStax, 1010data, Datameer, 10gen, Digital Reasoning, splunk>, hortonworks, Karmasphere, tresata, wibidata

Large-Scale Vendors

Cisco, hp, Informatica, Microsoft, Oracle, Dell, Teradata, EMC², IBM, amazon webservices, intel, NetApp

THINK BIG ANALYTICS

Industry-Leading Big Data Solution Integrators

Why Hadoop + Big Data is changing the game?

- Previously impossible to do this analysis
- Analysis conducted at fraction of the cost
- Analysis conducted in less time
- Greater flexibility for future unknowns

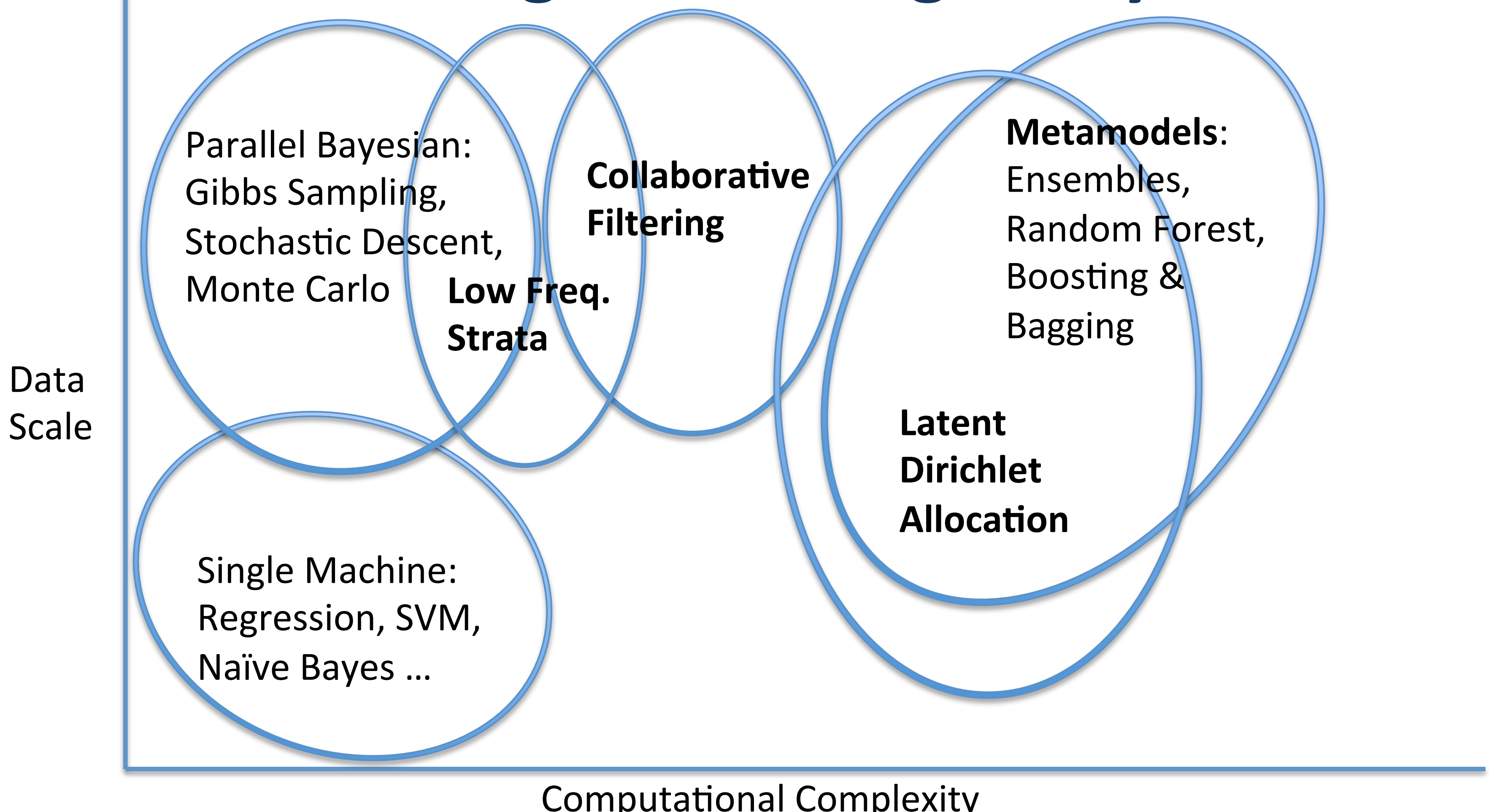
Big Analytics: Starting Simple

- Data-intensive cloud computing lets you work with massive data sets that vary
- Typical Process
  - Exploratory modeling: find patterns in **PBs** of data
  - Baseline modeling: simple models (e.g., Bayesian) iterated quickly
  - Live testing: frequent scoring with Champion/Challenger
  - Refinement: more sophisticated Machine Learning, features, etc.

Simple algorithms and lots of data trump complex models.

Halevy, Norvig, and Pereira (Google), *IEEE Intelligent Systems*

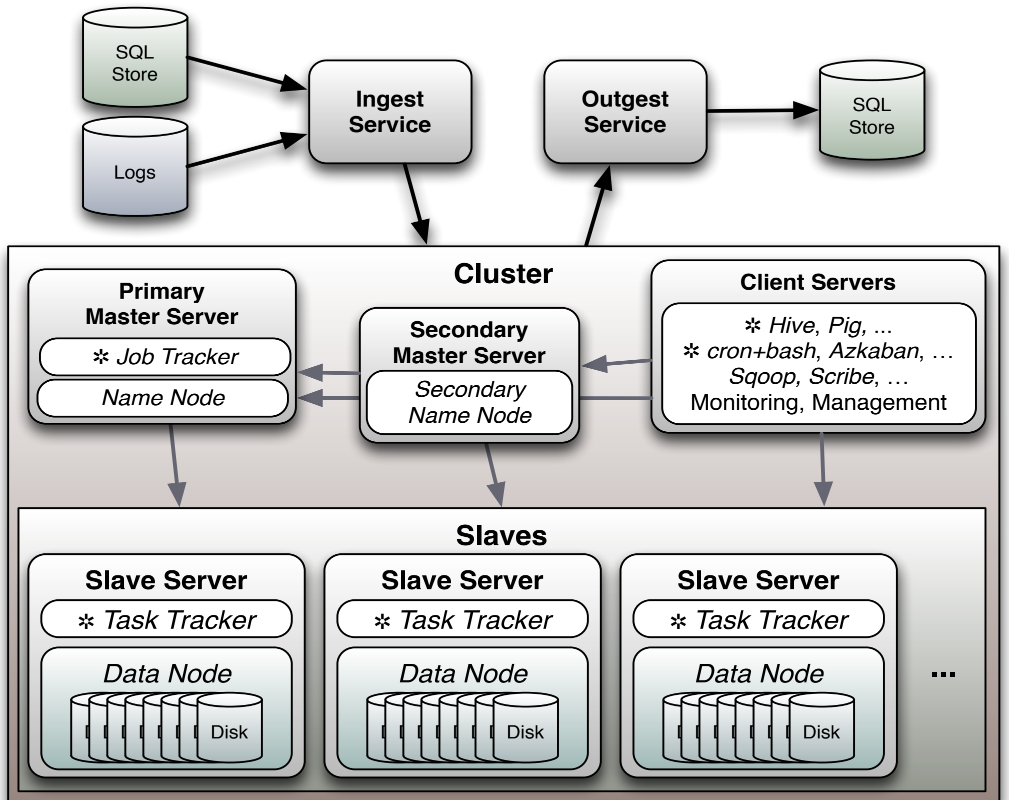
From Big Data to Big Analytics



Hadoop Origins

- Open Source Distributed Cluster Software
  - Distributed file system
  - Java-based MapReduce
  - Resource manager
- Started in Nutch project (open source crawler)
- Inspired by Google MapReduce and GFS

Hadoop Components



Key

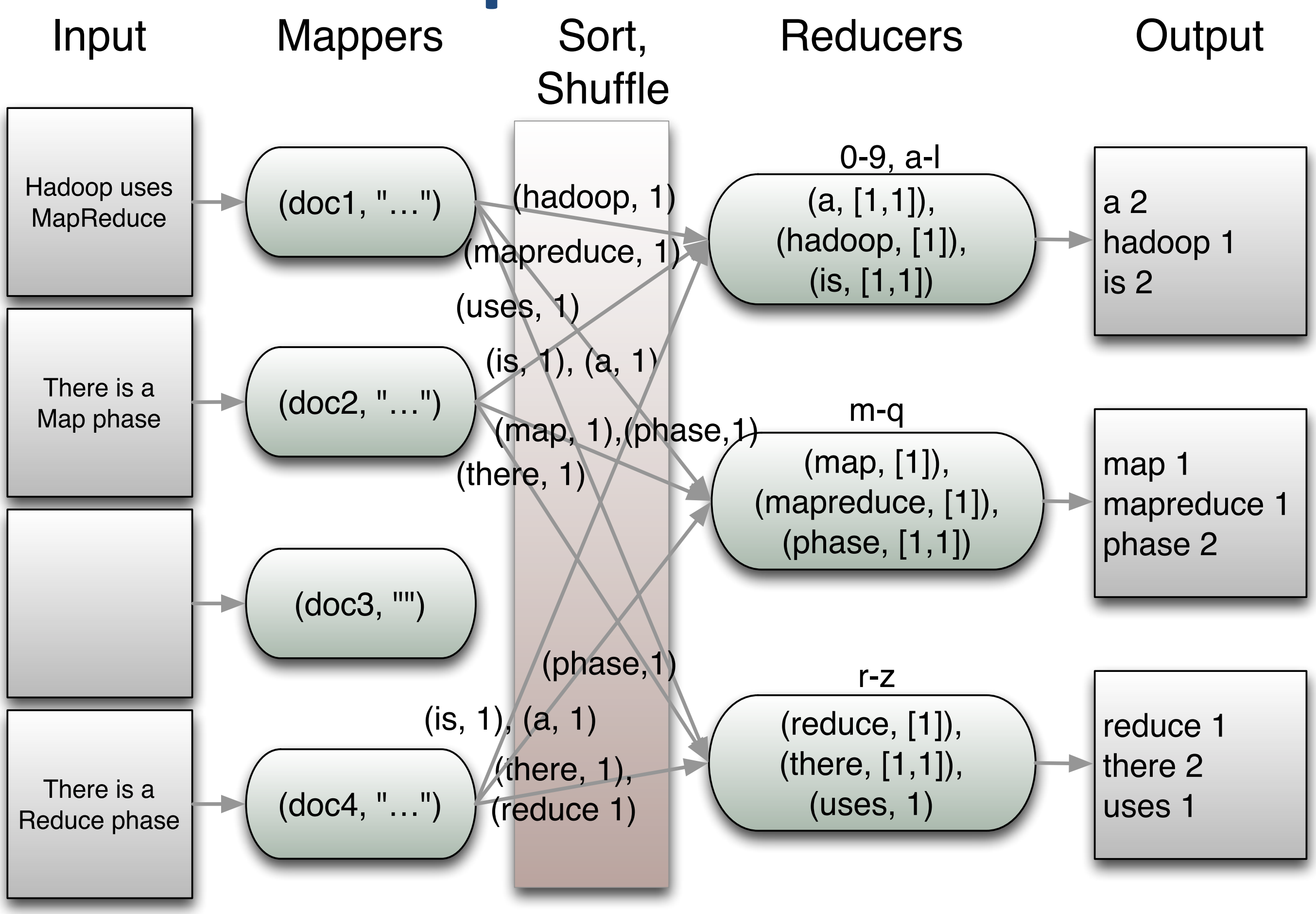
- *italics:* process
- ✲ : MR jobs

Data Processing Models

MapReduce 100

- Functional programming
  - Filter function $\underline{X}' = \underline{X}$ (if $X_i > 0$)
  - Map function: $\underline{Y} = \sin(\underline{X}')$
  - Reduce function $z = \mathrm{sum}(\underline{Y})$
- Hadoop
  - Spread the data out
  - Send the code to the data
  - Embarrassingly parallel problems work really well
- Many problems can be cast as a Map - Reduce

MapReduce 101



Word Count: Mapper

```
public static class TokenizerMapper
     extends Mapper<Object, Text, Text, IntWritable> {

  private final static IntWritable one = new IntWritable(1);
  private Text word = new Text();

  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);
    }
  }
}
```

Word Count: Reducer

```
public static class IntSumReducer
      extends Reducer<Text,IntWritable,Text,IntWritable> {
   private IntWritable result = new IntWritable();

   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);
   }
 }
```

MapReduce Wiring

```
public static void main(String[] args) throws Exception {
  Configuration conf = new Configuration();
  String[] otherArgs = new GenericOptionsParser(conf, args).
    getRemainingArgs();
  if (otherArgs.length != 2) {
    System.err.println("Usage: wordcount <in> <out>");
    System.exit(2);
  }
  Job job = new Job(conf, "word count");
  job.setJarByClass(WordCount.class);
  job.setMapperClass(TokenizerMapper.class);
  job.setCombinerClass(IntSumReducer.class);
  job.setReducerClass(IntSumReducer.class);
  job.setOutputKeyClass(Text.class);
  job.setOutputValueClass(IntWritable.class);
  FileInputFormat.addInputPath(job, new Path(otherArgs[0]));
  FileOutputFormat.setOutputPath(job, new Path(otherArgs[1]));
  System.exit(job.waitForCompletion(true) ? 0 : 1);
}
```

Hive Overview

- A *SQL*-based tool for *data warehousing* using Hadoop clusters.
- Lowers the *barrier* for Hadoop *adoption* for existing SQL apps and users..
  - Translates SQL to MapReduce
  - Provides an *optimizer*
- Extensible data types & UDFs
- The first popular *meta-data service* for Hadoop

Word Count in Hive

```
CREATE TABLE docs (line STRING);

LOAD DATA INPATH 'docs' OVERWRITE INTO
TABLE docs;

CREATE TABLE word_counts AS
SELECT word, count(1) as count from
  (SELECT explode(split(line, '\\s'))
   AS word FROM docs) w
GROUP BY word
ORDER BY count DESC, word;
```

Pig Overview

- *Pig Latin* is a higher-level map/reduce language
- A simple data flow language designed for productivity … not Turing complete (yet!)
- Built-in support for joins, filters, etc.
- Provides an *optimizer*
  - translates into Hadoop map reduce job steps
- Allows user-defined functions
- With *HCatalog* will share metadata with Hive

Sample Pig Script

```
lines = LOAD 'docs/*' USING TextLoader();

words = FOREACH lines GENERATE FLATTEN(TOKENIZE($0));

groups = GROUP words BY $0;

counts = FOREACH groups GENERATE $0, COUNT($1);

sorted = ORDER counts BY $1 desc, $0;

STORE sorted INTO 'output/wc' USING PigStorage('\t');
```

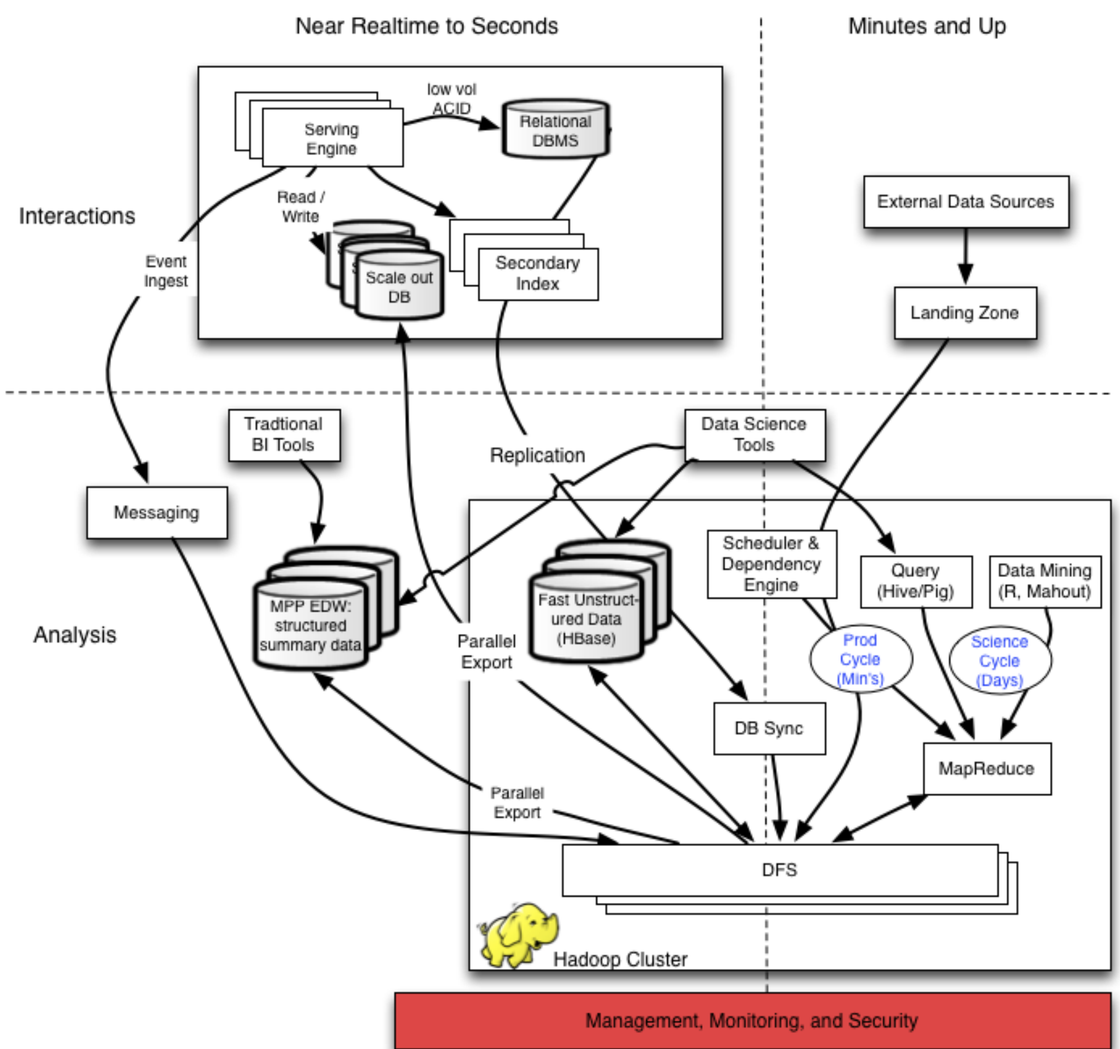
THiNK BIG ANALYTICS

Industry-Leading Big Data Solution Integrators

MapReduce Frameworks

- Cascading
  - Java-based optimizer & relational operators
- Crunch
  - Abstract collections and optimizer
- Streaming, Pipes
  - Non-Java integration (Perl, Python, Ruby, C/C++,…)
- Tap
  - Simplify time series processing, use of diverse tools and data formats

Integration



Source: Think Big Analytics

Reference Architecture

- Additive data processing power for flexibility.
- Big Data Strategy is integrated with HBase, relational, existing BI and data warehouse technology.
- Provides capability to create data science discipline using full set data.
- Analysis capability on "all" internal data with capability to add external data at will.

Unstructured Data Ingestion

Batch log shipping

- No distributed management and monitoring

Syslog forwarding

- No distributed management and monitoring

Apache Kafka

- Distributed message routing
- Distributed monitoring and management (agents)
- Written in Java

Apache Flume

- Pluggable sources, adapters, sinks
- Distributed monitoring and management (agents)
- Written in Java

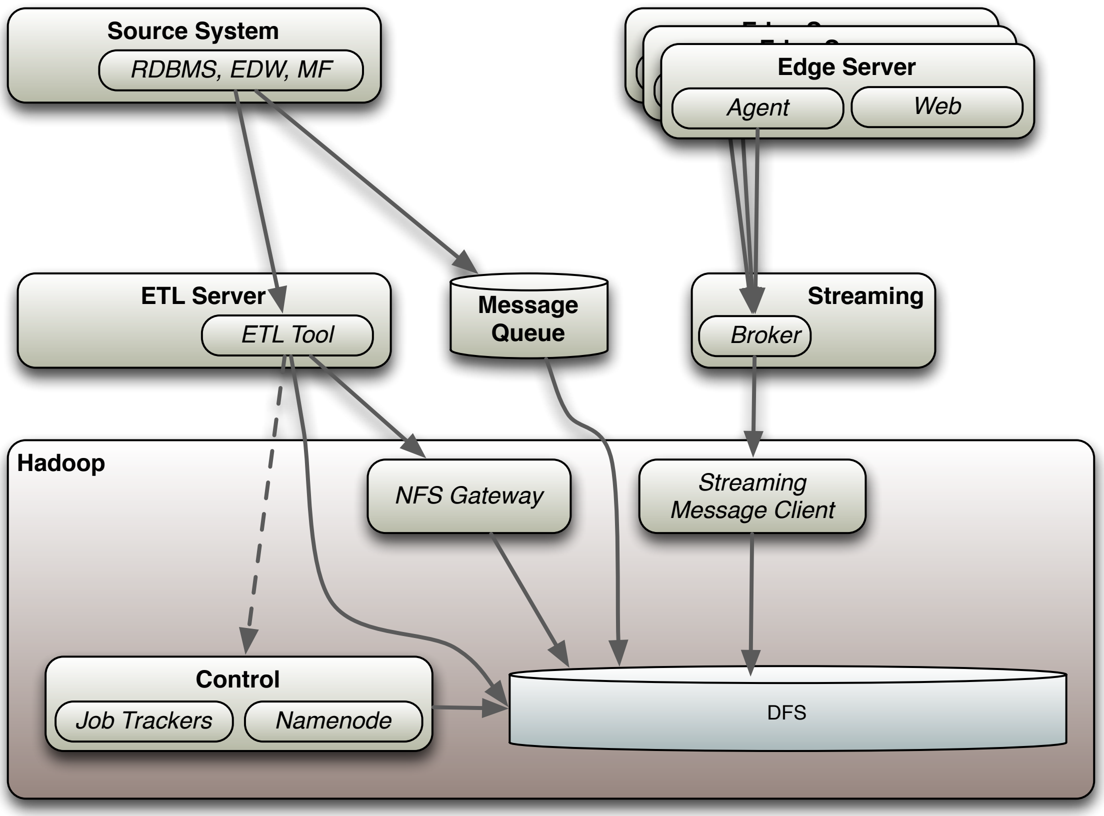
Other streaming frameworks

- Scribe, Chukwa, Honu

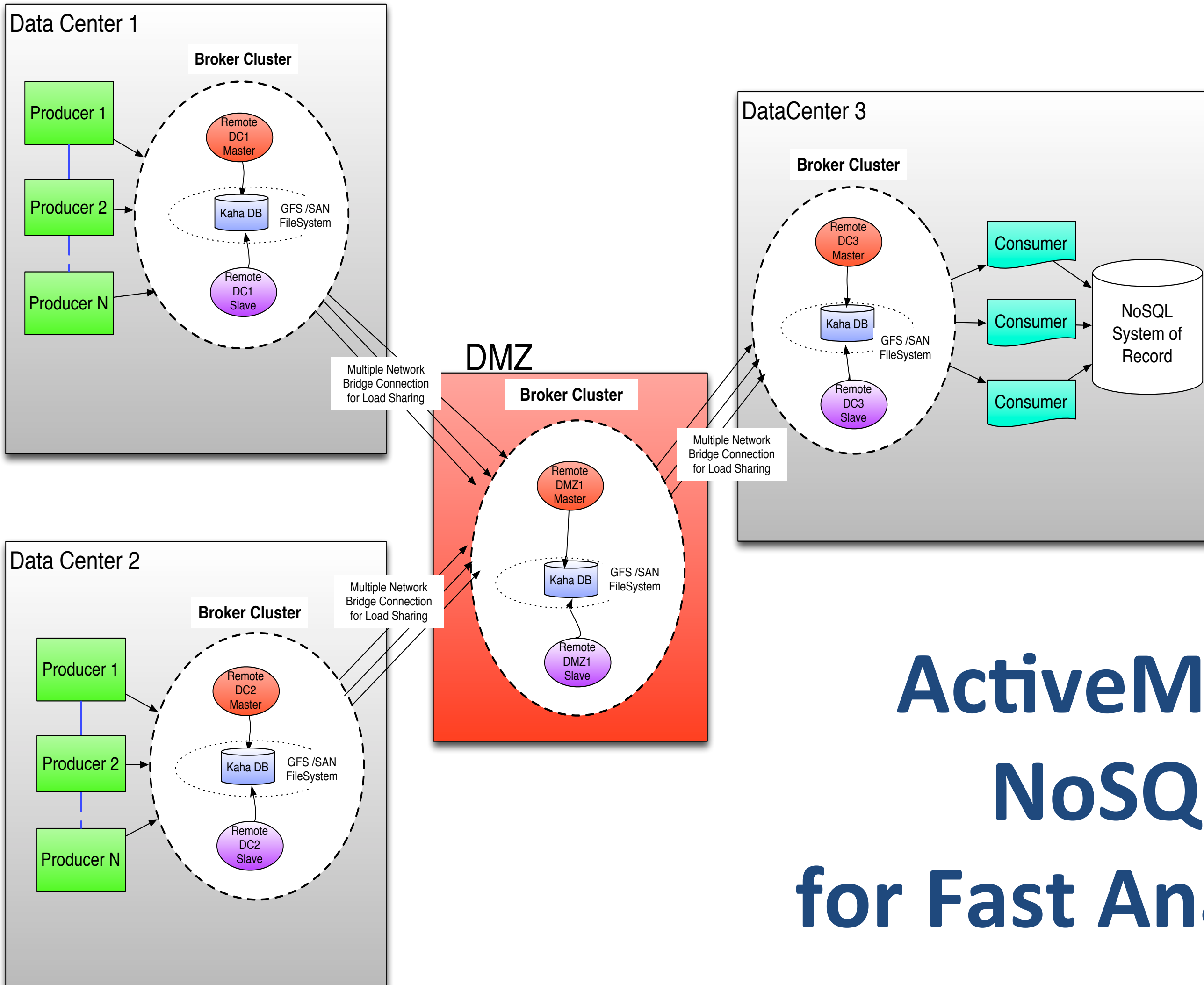
Message Queues

- ActiveMQ, ZeroMQ

Ingestion Architecture Alternatives



THiNK BIG ANALYTICS Industry-Leading Big Data Solution Integrators



ActiveMQ + NoSQL for Fast Analytics

THiNK BIG ANALYTICS Industry-Leading Big Data Solution Integrators

HBase

- Tables for Hadoop…
  inspired by Google's Big Table
- Supports both batch and random access
  - Ad hoc lookup
  - Website serving queries…
- High Consistency
- Maturing rapidly (e.g., reducing latency variance)
- Still a performance tax vs. DFS

Streaming Big Data

- Responding to incoming events at scale
- SQL-style
  - SQLStream, InfoSphere Streams
- MapReduce-style emerging
  - Kafka, S4, Storm, FlumeBase …

Common Uses

Common Workloads

- Batch processing
  - ETL
  - Model training
  - Model scoring
- Fast analytics
  - Search
  - Lookup

THINK BIG ANALYTICS

Industry-Leading Big Data Solution Integrators

IT Log & Security Forensics & Analytics

Find New Signal	100% Capture
Predict Events	Data Governance
React in real time	Shared Services

Automated Device Data Analytics

Failure Analysis	Cross Sell/Upsell
Proactive Fixes	Customer Analytics
Product Planning	Monetize Data

Advertising Analytics

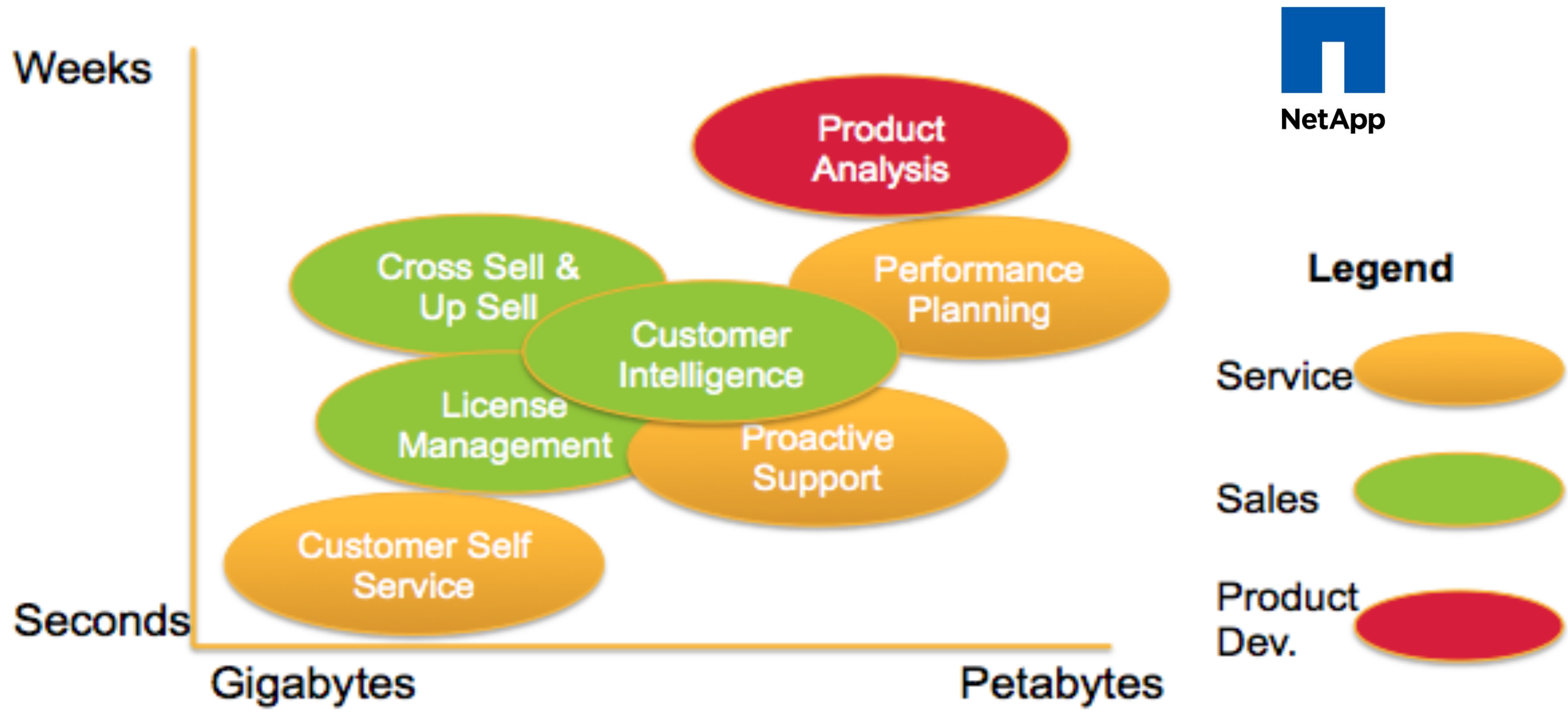
Attribution	Insights
Customer Value	Optimization
Segmentation	Social Media

Big Data Warehouse Analytics

Hadoop + MPP + EDW

Cost Reduction	Ad Hoc Insight
Flexibility	Predictive Analytics

Automated Device Support Case Study



Why Big Data Warehouse?

Challenges

- Cost to store unstructured data
- Poor response time to changing BI needs
- Data Warehouse access for departments

Goals

- Integrate unstructured data with data warehouse
- Predictive analytics based on data science
- Comprehensive access to cluster for all users

Hadoop's Role

- Support semi-structured and unstructured data
- Large scale storage
  - Transaction-level detail (e.g., clickstreams)
  - Archival
  - Integrated data: multiple warehouses, new data sources, …
- Powerful processing capacity
  - Perform large scale analyses/studies
  - Drill to detail in large fact tables
  - Query without structure: agility to analyze data without preprocessing
  - Transformation to build dimensional models, aggregates, and summaries
- Build predictive models

Data Agility

Classic Warehouse

- ETL
- Pre-parse all data
- Normalize up front
- Feed data marts
- New ideas need IT projects

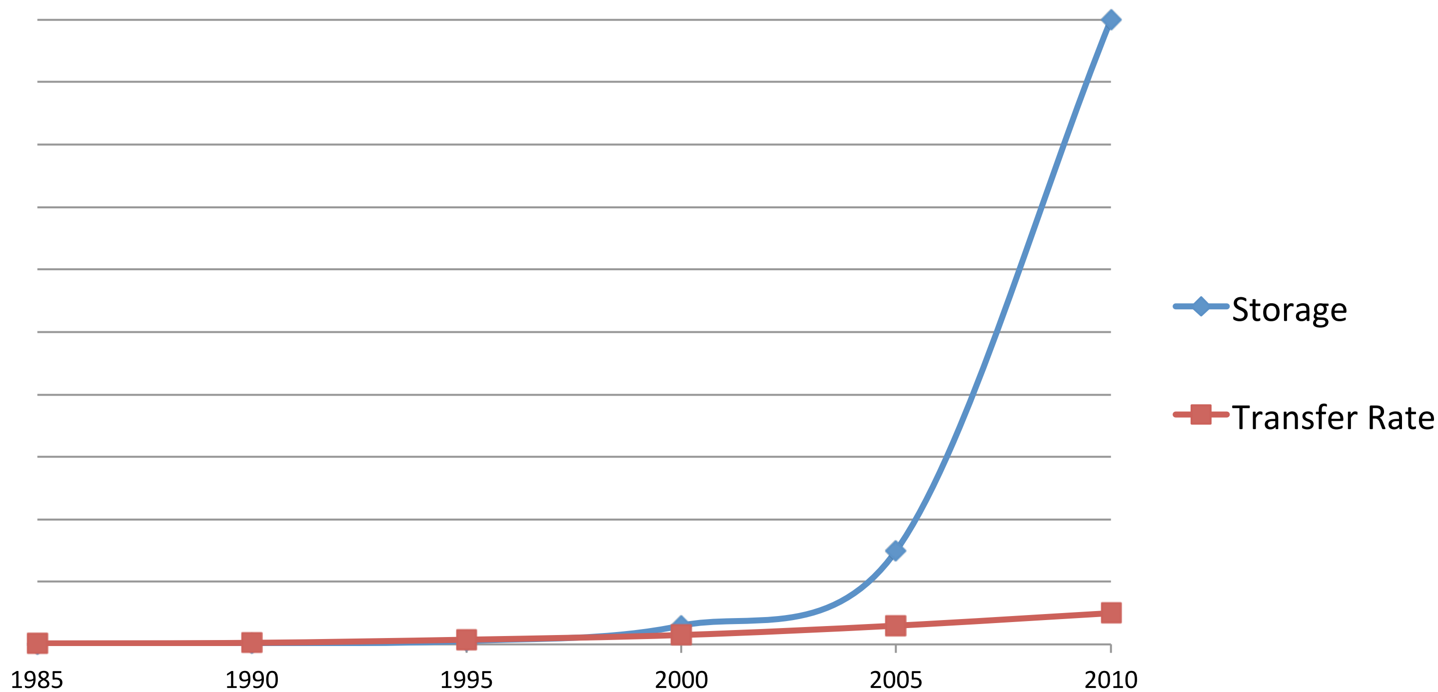
Big Data Warehouse

- Store raw data
- Parse only when proven
- Approximate parse on demand
- Capacity for analysis on demand
- Prove ideas *before* projects to optimize

Futures

Computing Trends

- The growth of storage density has well outpaced the growth of data transfer rates



Computing Trends, cont'd.

- In 1990, you could read all the data from a typical drive in about 5 minutes
- Today, it would take over 2 hours
- And, seek times have improved even more slowly than data transfer rates (SSDs improve this)
- Network speeds in the data center have improved at a comparable speed (60%/yr.)
- So clusters of commodity servers allow throughput
- Clusters of servers allow RAM density

Trends in Big Data for 2012

- Hadoop 0.23 (2.0?)
  - Explosion in New Application Models (e.g. MPI)
- HBase Prominence
- Data Science
  - Practices, Tools,
  - Technologies
- Integration
  - External Tools

Commodity Hardware in 2016?

- 512 GB of RAM
- 64 cores
- 15 TB spinning disks
- 1 TB SSDs for caching
- 100 Gigabit (InfiniBand?)

THiNK BIG ANALYTICS

Industry-Leading Big Data Solution Integrators

Summary

- Massive data volumes
  - Processing, Computation
- Ingestion is critical
  - High Volume
  - Reliable, Durable, HA, DR
  - Variety of sources
- All critical to delivering analytics
  - Low latency

THiNK BIG

ANALYTICS

douglas.moore@thinkbiganalytics..com

We're hiring…