Big Data and the Analytic Race

What is Big Data?

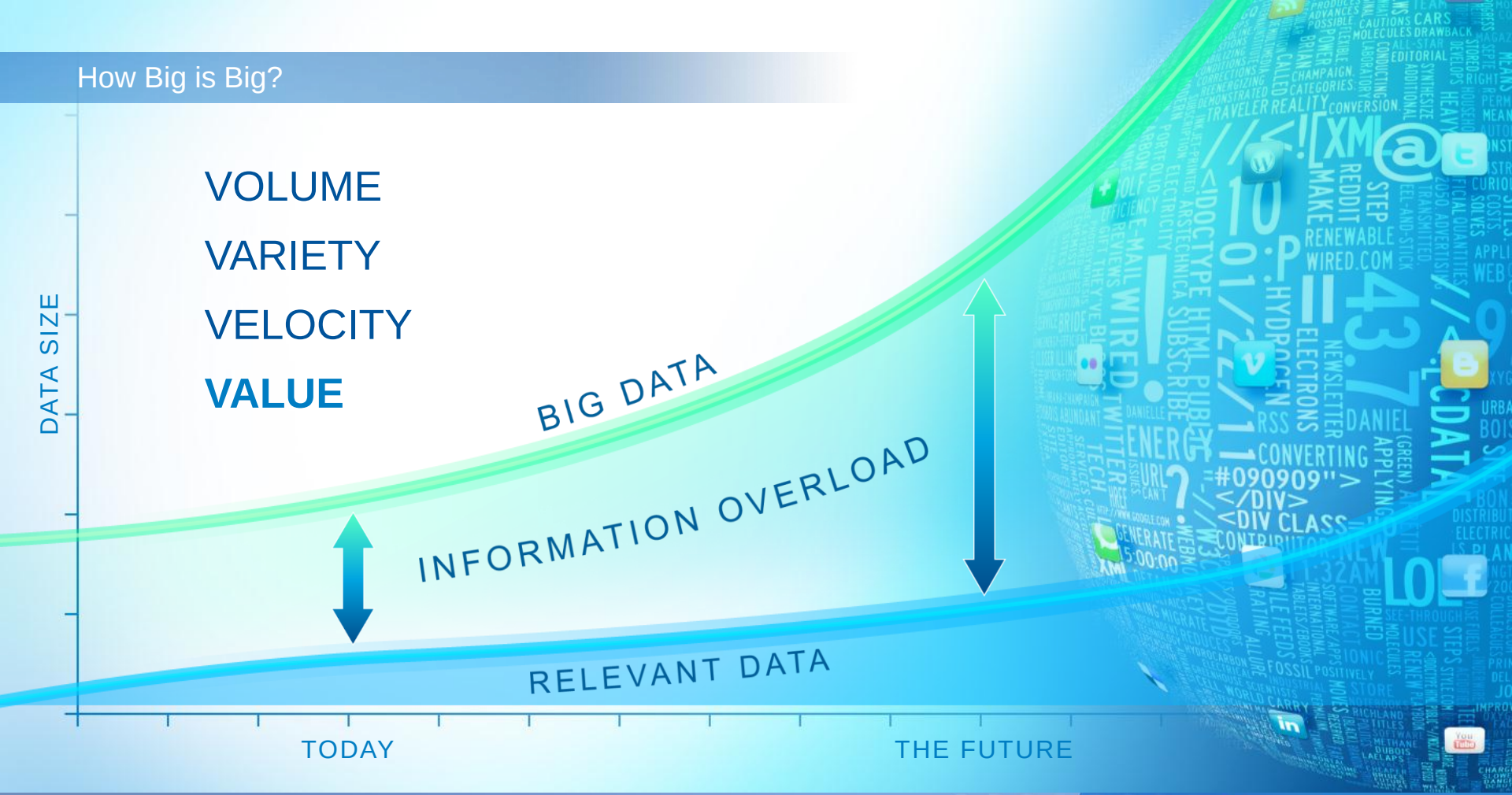
Big Data

When volume, velocity and variety of data exceeds an organization's storage or compute capacity for accurate and timely decision-making

The Data Explosion

Vast Quantities of Data Are Being Generated Like Never Before

How Big is Big?



Big Data Examples

- Each Swipe of a Credit Card
- Each Banking Transaction
- Insurance Claims / Telematics data / Voice Transcription
- Mobile devices, signal and location data
- Each Cell phone ping, call, text, each RFID tag
- Each Web Site, Each Blog
  - Yahoo alone has 200 petabytes of data!

They Are All Producing More and More Data.

Smart Meters

Just One Small Example: Smart Meters

- Before Meters Read Once a Month
- Now Every 15 Minutes is Typical
- 3,000 Times as Much Data.
- For a Utility With One Million Customers it's Like an IT Shop Tracking CPU for One Million Servers.

Smart GRID

- But Wait! There's More!!
- New Smart Grid Initiative
  - Control Down to the Appliance Level.
  - So More Like Tracking 10 Million Servers.

So What?

- So There Is Lot's of Data Out There
- Why Do We Care?
- What Is The Big Deal About Big Data?

An Analytical Gold Mine

Tens If Not Hundreds of Millions of Dollars!

- Actionable Information is Buried Deep in the Mountains of Data
- Information About Customers Can Help Retain Them and Help Them Spend More
- Then There is FRAUD…

Fraud

Annual Losses From Credit Fraud Estimated

At $48 Billion in 2008!

- Reducing This Fraud by Just 1% Saves $480 Million / Year!
- One Large Bank Estimated Their Losses at $6.4 Billion in One Year

Government Fraud

Annual Medicare Fraud Estimated Between $60 Billion and $90 Billion per Year

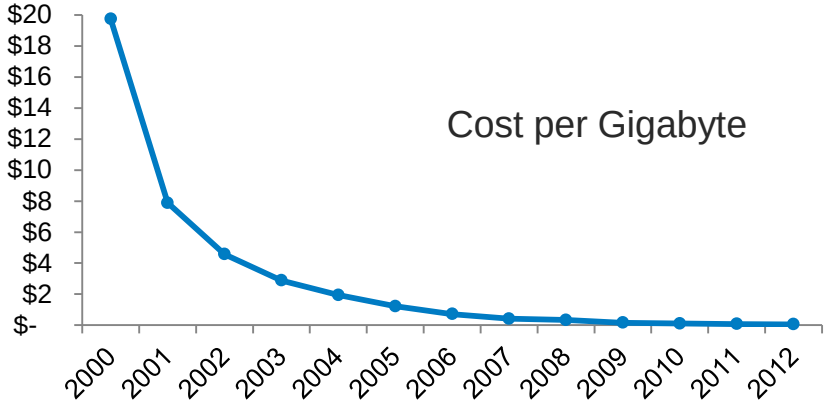
- Reducing Medicare Fraud by Just 1% Saves $600-$900 Million / Year!

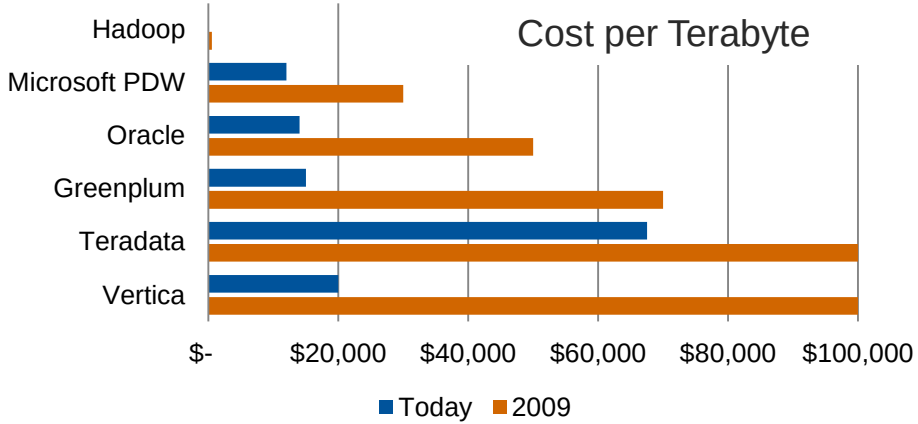
The Problem

The Problem Is: How Do You Mine All This DATA?

- Too Big To Store and Use
- Too Slow To Analyze

…Or At Least That Is How It Used To Be…



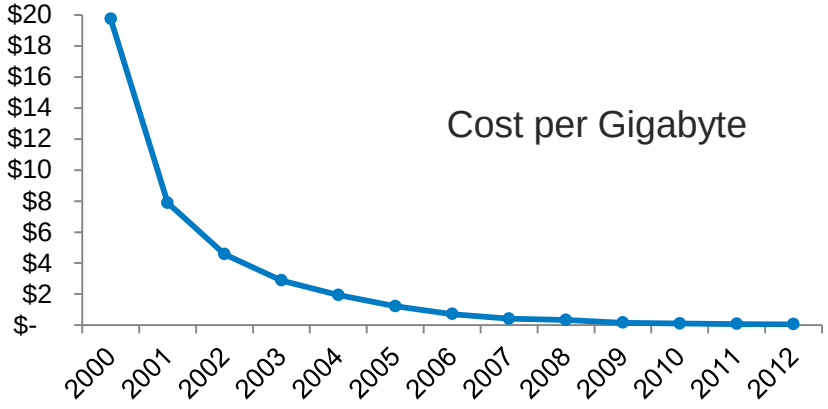


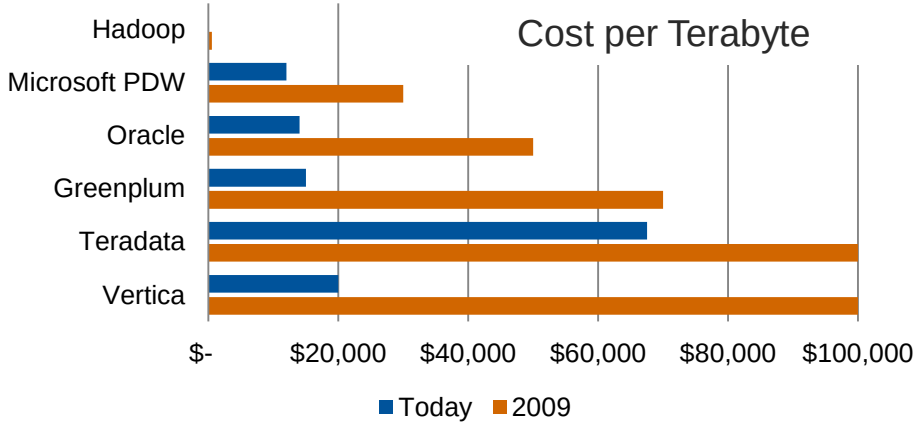
Cost of Storage, Memory, Computing

- In 2000 a GB of **Disk** $17 today < $0.07
- In 2000 a GB of **Ram** $1800 today < $1
- In 2009 a TB of **RDBMS** was $70K today < $ 20K

10 Terabytes files now reasonable to move In-memory

TRENDS IN BIG DATA, STORAGE, HADOOP & IN-MEMORY TECHNOLOGY





2011 – 2012 Major Players & Hadoop

- Greenplum MapR (May ‘11)
- IBM Big Insights (May ‘11)
- Microsoft and Hadoop (Oct ‘11)
- SAP Sybase IQ & Hadoop (November ‘11)
- Oracle & Cloudera Appliance (Jan ‘12)
- Teradata Partners w. Hortonworks (Feb ‘12)
- SAS LASR Server on Hadoop (Mar‘12)

In-Memory Technology

- SAS HP Solutions
- SAS LASR In-memory Server
- SAP HANA
- Oracle Exalytics

IT needs to be involved in setting up these MPP environments

Types of Big Data Processing

- Grid
- Clustered Databases
- In-Database Analytics
- In-Memory Analytics

Grid

- Grid
  - Spreading one or more workloads over multiple, possibly heterogeneous servers
  - Scheduling, prioritization, redundancy, flexibility
  - Shared access to data
  - Some large jobs run in parallel on multiple servers
  - Mainframe Sysplex

Clustered Database

- Clustered Database
  - Spread large datasets across many servers
  - Spreads database processing across many servers
  - Analytical Database, Not Transactional
    - Read/Write massive amounts of data at once, not individual records.
  - Examples:
    - Teradata
    - EMC Greenplum
    - Hadoop

In-Database

- In-Database
  - Regular SQL DBs like Oracle and DB2
  - Moving computing into the database instead of pulling the data out
  - Useful for scoring massive amounts of data for models

In-Memory

- In-Memory
  - Now take that distributed data with distributed processing and “Snap” it into memory
  - Processing was fast before … but now is crazy fast!
  - Tens of terabytes can be handled in memory!
    - 100 servers with 96 GB each is nearly 10 TB
    - Commodity hardware is not nearly as expensive as before

In-Memory Results

Business Problem	Data Size and Analysis	Before	In-Memory
Probability of Loan Default	• 1 billion rows of data • Regression analysis	11 to 20 hours depending on hardware configuration	Less than 54 seconds
Optimize Response to Marketing Campaign across multiple channels	• 100 million rows of historical contact information • 15 million customers • 900 offers • 20 offers per customer • Many business rules	2.5 to 5 hours	Less than 90 seconds
Calculate Credit Risk Exposure across entire bank	• 10s of Millions of rows of customer data • Regression analysis	167 hours (a week)	84 seconds

In-Memory Difference In Analytics

Banks's Current Process

- 5 hours
- Model lift of 1.6%
- 1 model per day per modeler
- One algorithm (NN)
- 7 iterations of NN training

In-Memory Data Mining

- 3 minutes
- Model lift of 2.5%
- 1 model per 30 minutes conservatively
- Random Forest, SVM, Logistic and other challenger methods
- More complex network 5000 iterations in 70 minutes

Apache Hadoop

WHAT IS THE SCOOP ON HADOOP?

What exactly is Hadoop?

Think of it as an **infinitely expandable filing cabinet** that has the ability to help you summarize what is stored in it

- Can store any kind of data in it
- When it gets filled up, just buy more drawers…
- It has built in some nifty “space organizers”!
- Hadoop is partitioned into compartments (called ‘clusters’) that can be used for analysis
- It has its own set of languages
- Basic versions are “free” to download and use
- Different vendors offer their own custom versions

WHAT IS HADOOP?

“Open Source Software that allows for the distributed processing of large data sets across clusters of commodity computers”

It isn’t a database, it is a file system with a parallel programing model.

WHAT IS THE SCOOP ON HADOOP?

- Big Internet Companies Are Using Hadoop
  - Origins in Google's MapReduce and Google File System (GFS)
  - Yahoo Is Largest Contributor
  - Amazon Has Its Version (Elastic MapReduce)
- Also Used by eBay, IBM, SAP, Twitter, Netflix, LinkedIn, Apple, AOL, HP, Intuit, Microsoft and SAS, As Well As Many Others.
- In 2011 Facebook Had a 30 Petabyte Hadoop Cluster / Yahoo over 200 PB

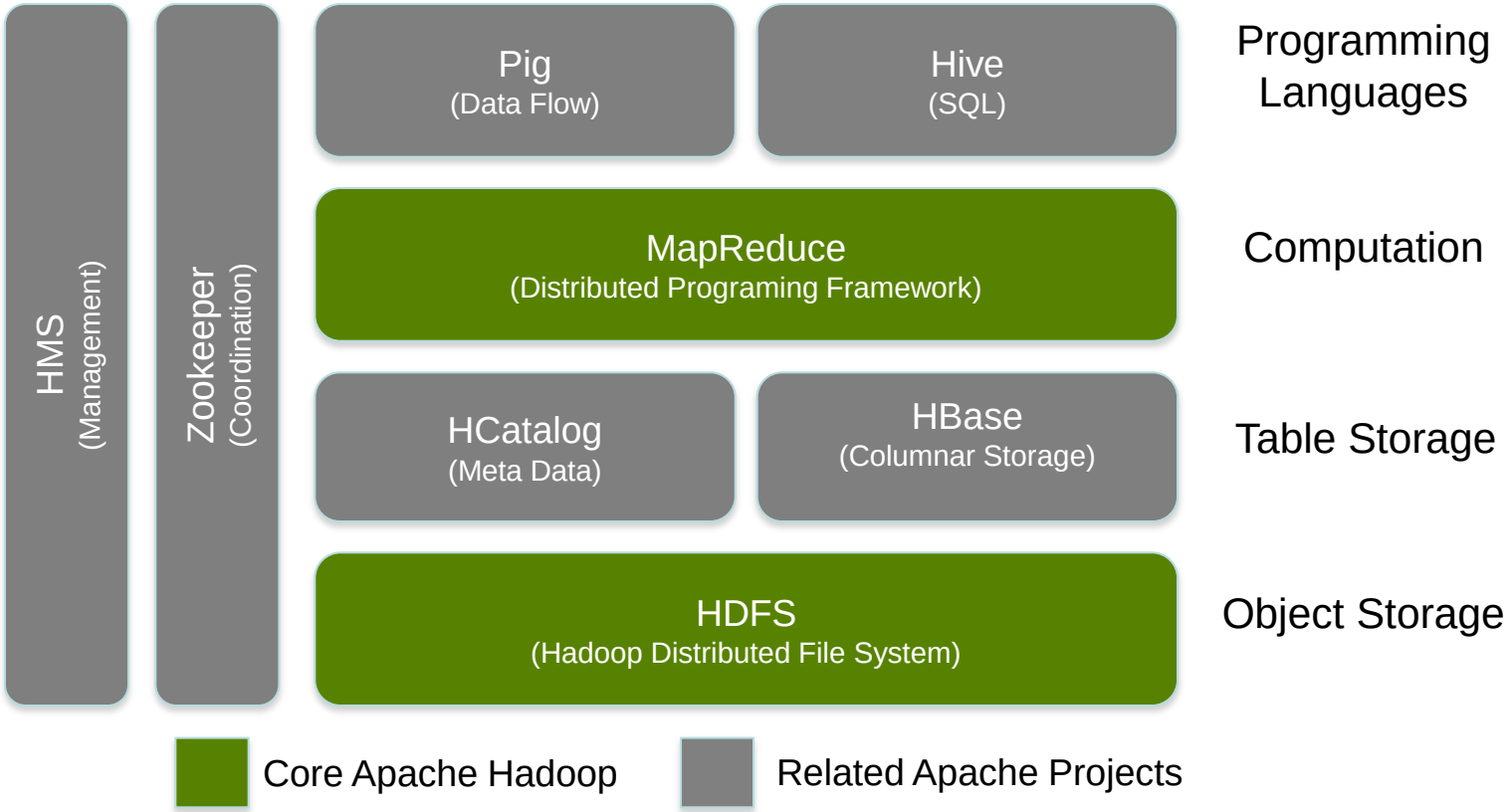
KEY COMPONENTS OF HADOOP

- HDFS – Stores petabytes of data reliably
  - Simple – Just a bunch of disks ~ no RAID
  - Reliable and Redundant ~ expect server failure
    - » Doesn't slow down or lose data even as hardware fails
  - Open Source So Other File Systems Can Be Used

- MapReduce – Allows huge distributed computations
  - Batch processing centric
    - » Hence its great simplicity and scalability, not a fit for all use cases

- Hadoop-Related Components
  - Pig – Programing Language To Simplify Creation of MapReduce Programs
    - » Grunt – interactive shell
  - Hive – SQL-Like Front-End to MapReduce
    - » Data still stored as sequential files, not database
  - Hbase – Database Built on HDFS
    - » Real-time random read/write
    - » Linearly scalable
    - » Ironically not SQL
  - ZooKeeper – Centralized Service for Distributed Applications

HADOOP ECOSYSTEM & LINGO



WHAT YOU NEED TO KNOW ABOUT HADOOP

- RDBMS Databases = Connectors & adaptors to Hadoop (ie. Oracle,SAP)
- IBM = Big Insights / Big Sheets
- SAS/Access to Hadoop
- SAS/EDI Server 4.4 provides Hadoop Transformations for Data Integration
- SAS/Metadata & Lineage provide governance of Hadoop data
- SAS Proc Hadoop enables users to intermix MR & Pig in-line with SAS code
- R & Revolution = Bunch of MR Packages for Hadoop
  - RHIPE - interface between R and Hadoop
  - RHIVE – connect R to HIVE (similar to SAS/Access)
  - Rhadoop - is a collection of three R packages:
- Mahout = Data Mining for Hadoop
  - Main use today ~ Recommender engines (e.g. Amazon)

Managing Hadoop

Where Does IT Fit In?

- Business Needs IT To Manage Big Data Systems
- Possibly Hundreds or Even Thousands of Nodes In One Big Data Cluster
  - Yahoo has 42,000 Hadoop Nodes…
  - Spread Over 20 Hadoop Clusters…
  - Holding 200 Petabytes of Data

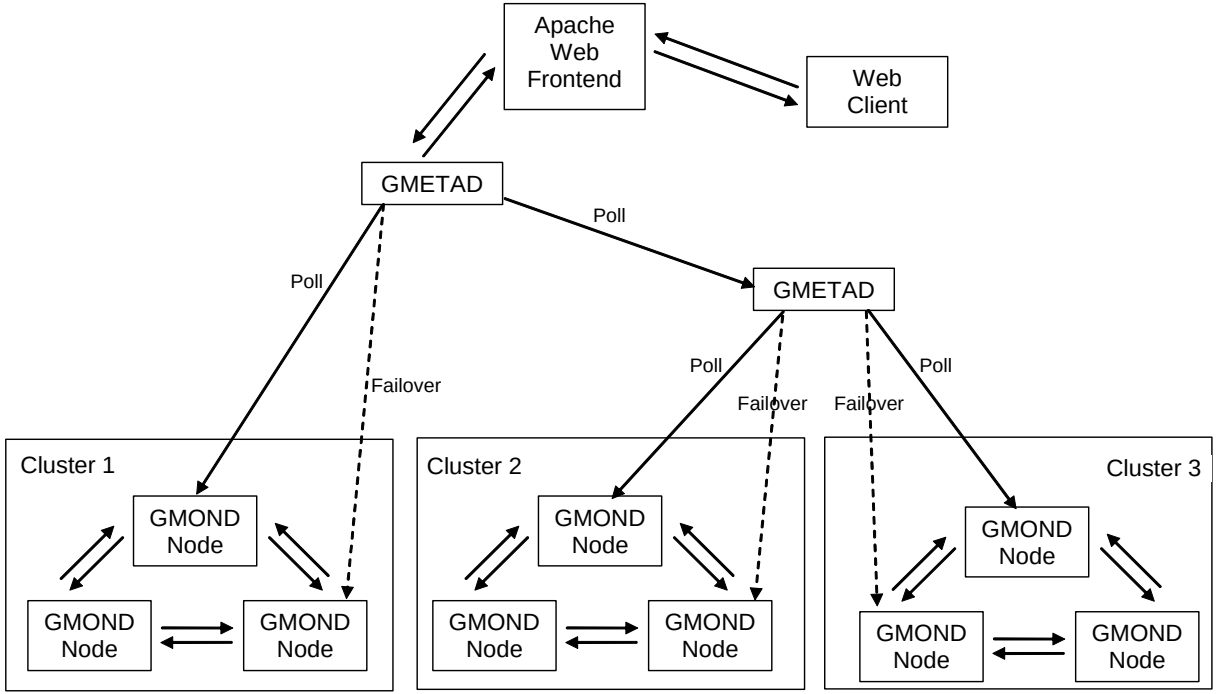
Management of the Hadoop Hive/Cluster

- How is IT going to manage the workload being dispersed within the Hive
- Which Hive/cluster components are limiting capacity/performance
- Where are the bottlenecks or room for increased utilization
- Any Nodes/Slaves not used at all or hardly engaged
- Resource statistics and metrics
- Where is it all going to come from and can we get a 3D view

Operations Monitoring and Reporting

- Introducing Ganglia
- Scalable Distributed Monitoring System
- Targeted at monitoring clusters and grids
- View Live or Historic Statistics
- Multicast-based Listen/Announce protocol
- Leverages widely used technologies such as XML for data representation, XDR for compact portable data transport, RRDtool for data storage and visualization
- http://ganglia.sourceforge.net or http://www.ganglia.info

Ganglia Architecture



Ganglia Web Frontend

Ganglia Gmond – Metric Gathering Agent

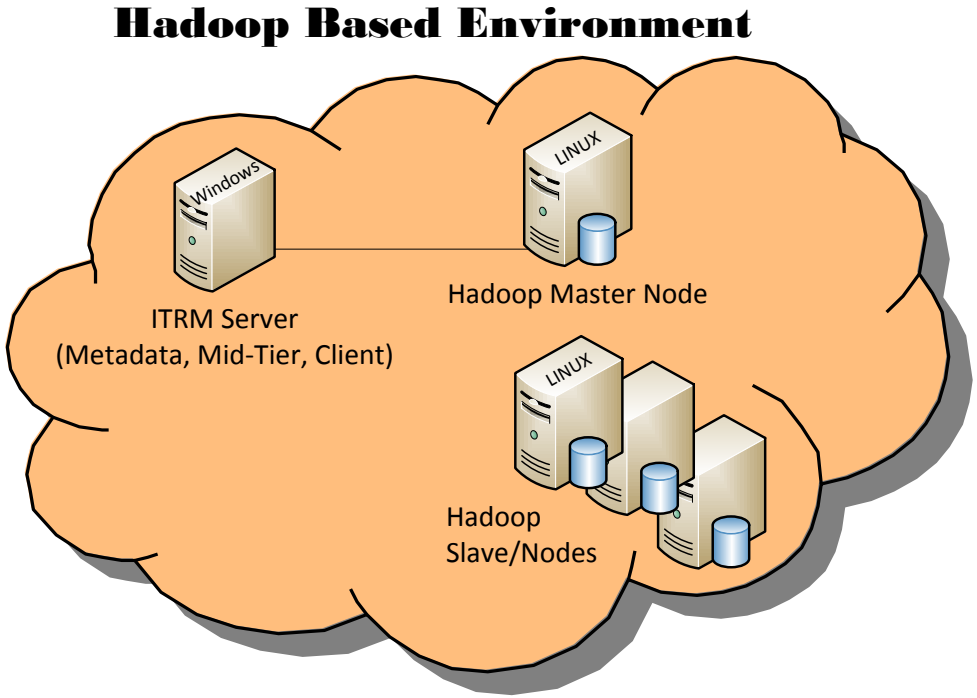
- Built-in metrics
  - Various CPU, Network I/O, Disk I/O and Memory
- Extensible
  - Gmetric – Out-of-process utility capable of invoking command line based metric gathering scripts
  - Loadable modules capable of gathering multiple metrics or using advanced metric gathering APIs
- Built on the Apache Portable Runtime
  - Supports Linux, FreeBSD, Solaris and more…

Gmond – Metric Gathering Agent (continued)

- Automatic discovery of nodes
  - Adding a node does not require configuration file changes
  - Each node is configured independently
  - Each node has the ability to listen to and/or talk on the multicast channel
  - Can be configured for unicast connections if needed
  - Heartbeat metric determines the up/down status
- Thread pools
  - Multicast listeners – Listen for metric data from other nodes in the same cluster
  - Data export listeners – Listen for client requests for cluster metric data

The Hadoop Datamart

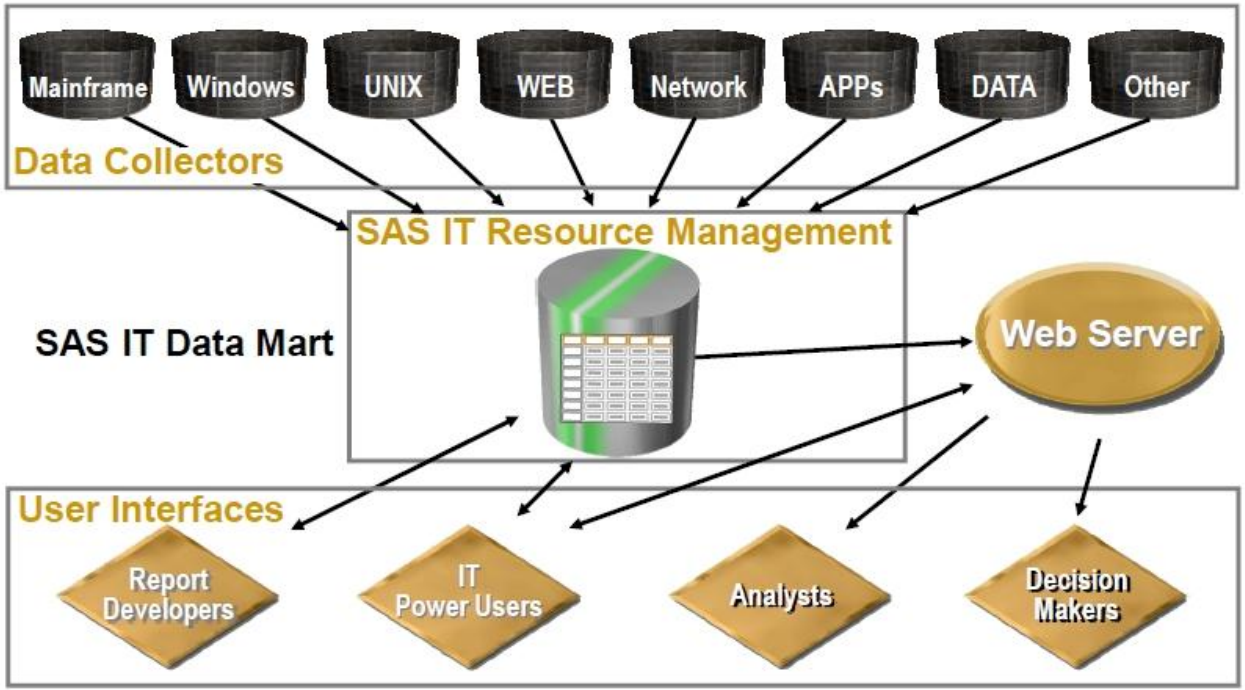
Case Study – SAS ITRM / Hadoop



SAS/ITRM Custom Adapter to Integrate Hardware Metrics
with Hadoop Performance Information

SAS IT Resource Management

Logical Architecture

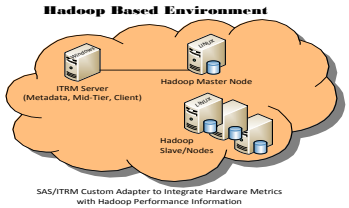


Case Study – Hadoop / Ganglia Adapter for SAS/IT Resource Management

Comprehensive Management of the Hadoop Environment with a Field ITRM Adapter

Providing:

- Analysis
- Reporting
- Metrics reporting from Ganglia

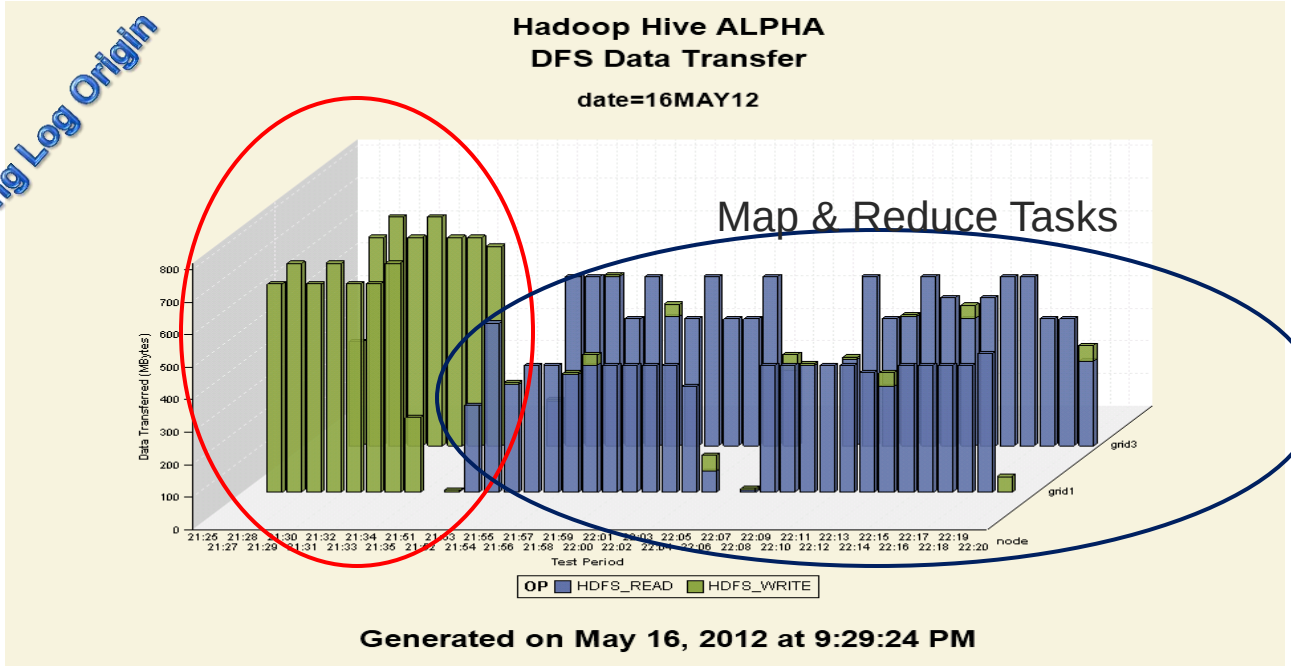


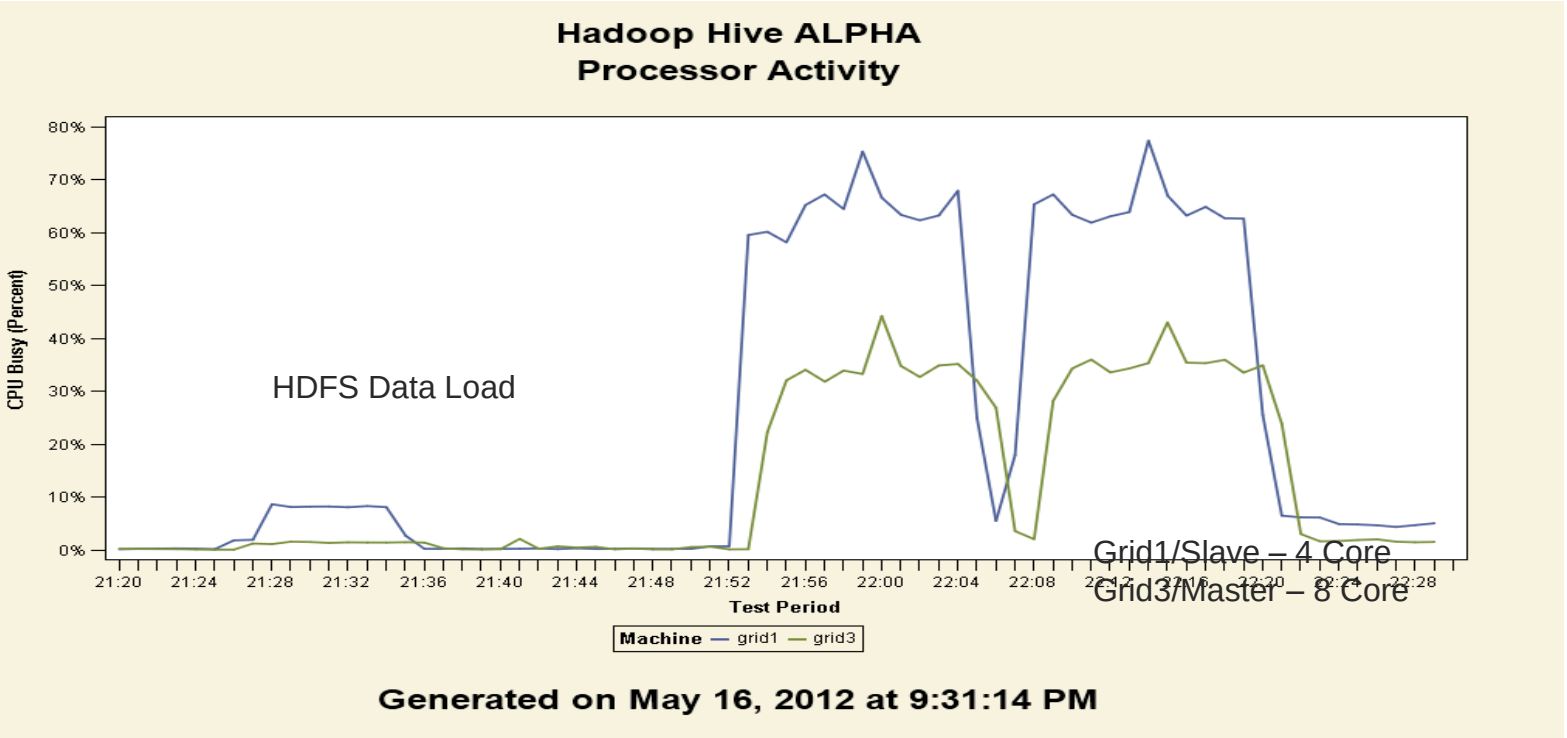
- ITRM adapter for Hadoop logs
- ITRM Datamart based on Hadoop data
- Integrated LINUX/UNIX OS performance metrics
  - SAR or Ganglia
- Analysis routines for both memory and storage based Hadoop environments
- Reports for both engineer and senior management

Accessing RRD Files

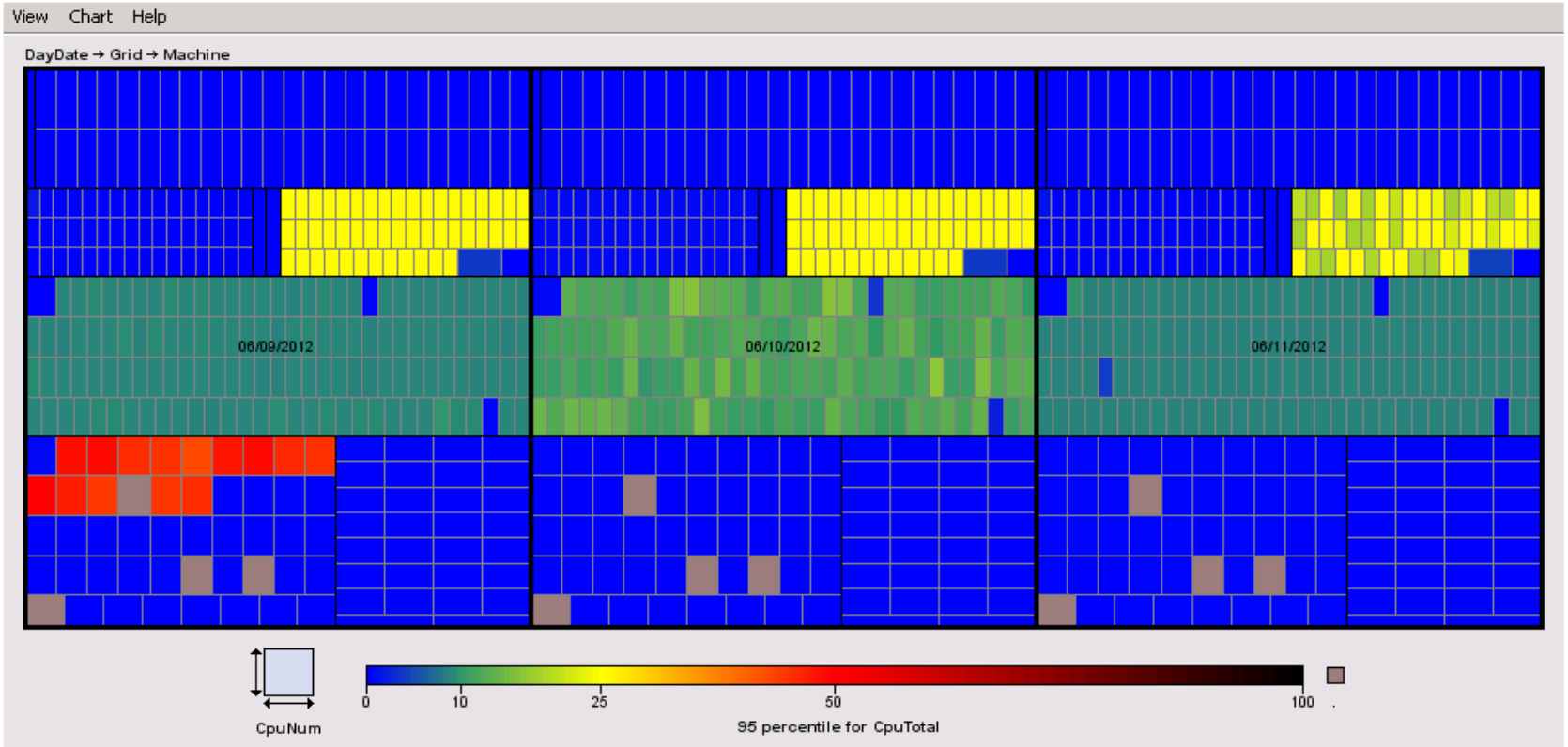
- RRDtool Adapter
  - The RRDtool adapter is a new adapter for ITRM 3.3
  - Reads any data from a RRD that has been created using the RRDtool software
  - Creates a Stage Table based on the contents of the user's RRDs
  - Reads the data even if it has been consolidated.
  - Will read a single round-robin database, or will read all round-robin databases in directory.
  - If multiple round-robin databases are read, the data will be combined into a single staging table.

Data-Node Tracking Log Origin



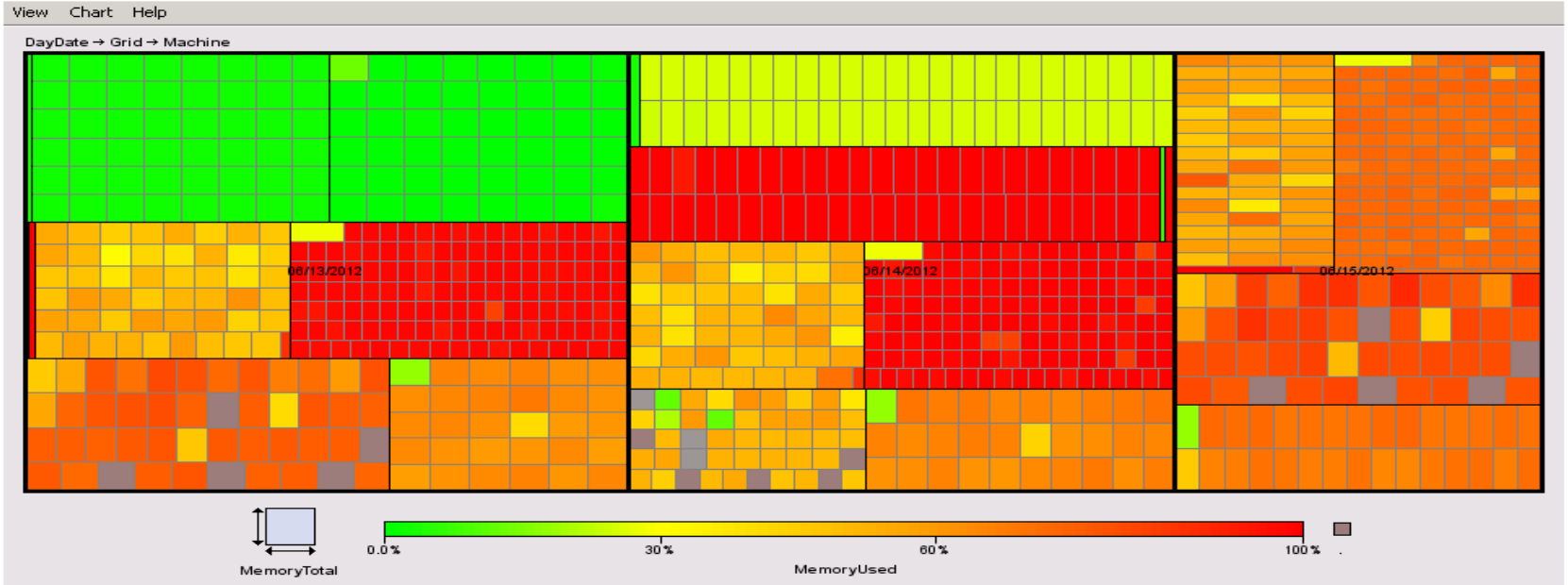


EEC Grid Performance - Last 3 days
Processor Busy - 95 Percentile



Generated on June 14, 2012 at 7:48:20 PM

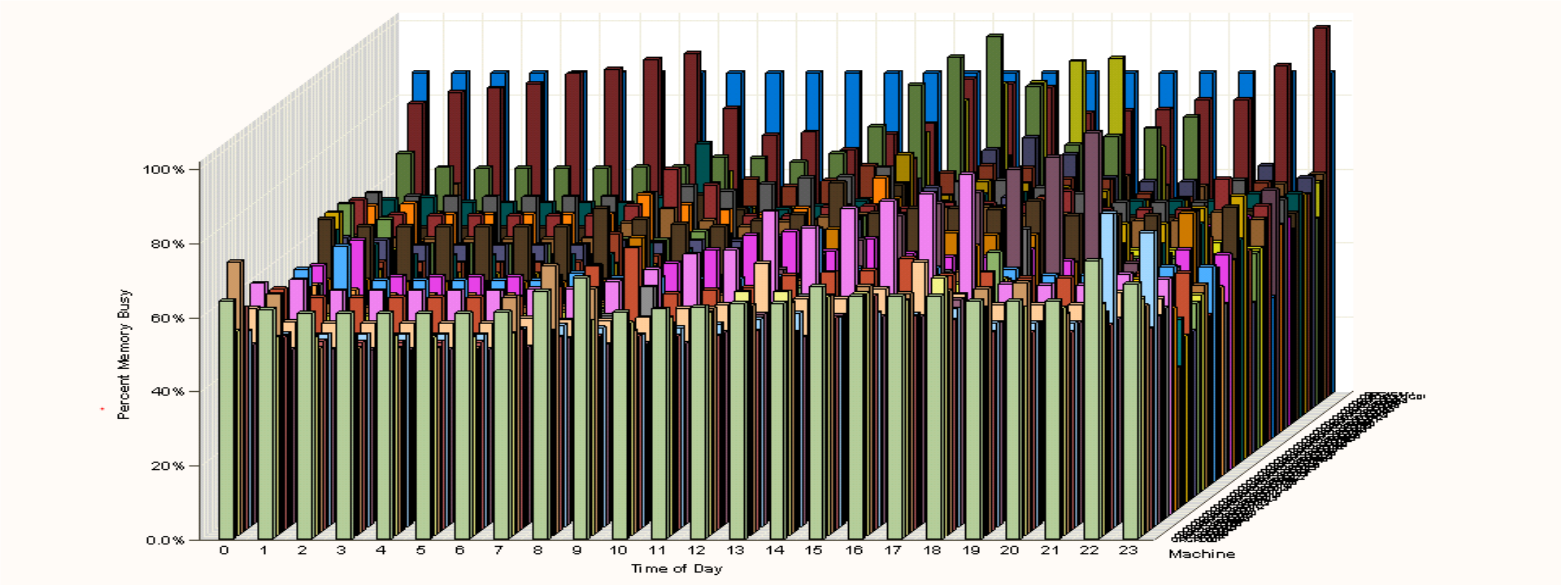
EEC Grid Performance - Last 3 days
Memory Percent Used - 95th Percentile

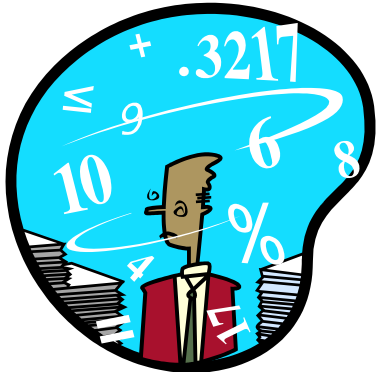


Generated on June 15, 2012 at 6:20:59 PM

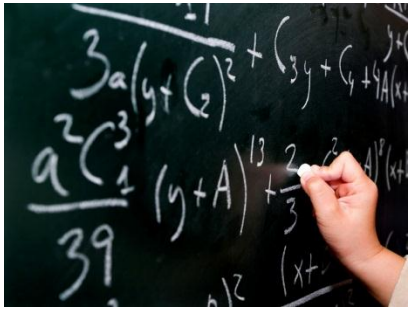
EEC Grid Performance Utilization
Memory Analysis

DayDate=06/14/2012 Grid=ORGRID





Analytics – From Zero to Insight



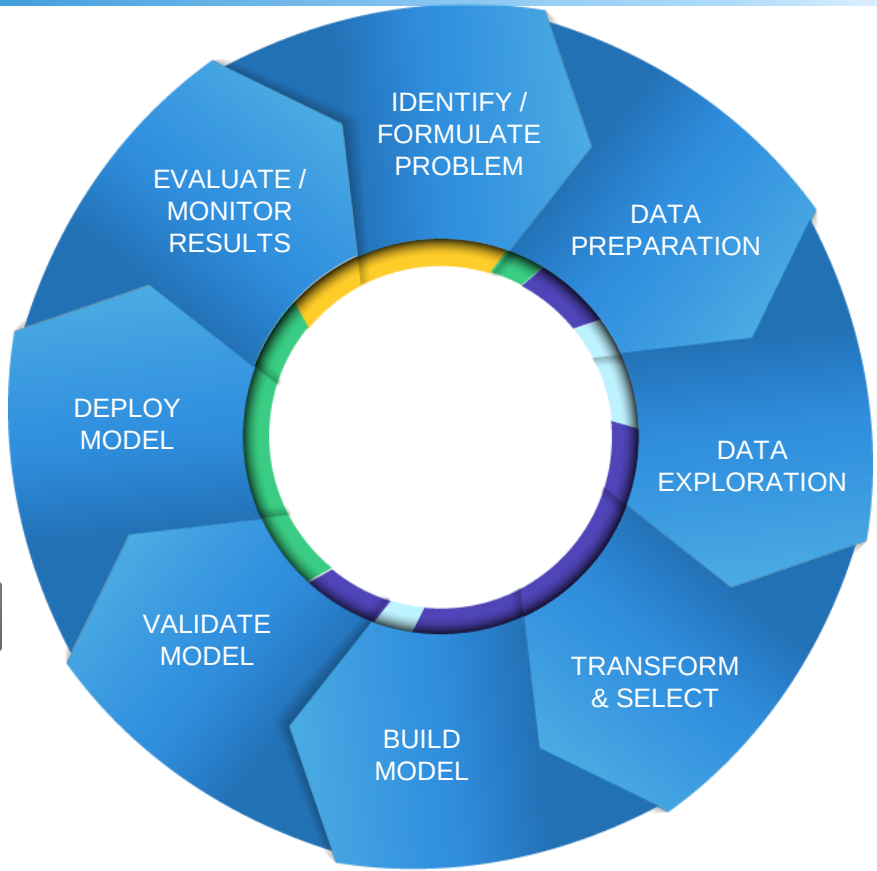
TRADITIONAL ANALYTICS LIFECYCLE

BUSINESS MANAGER

Domain Expert
Makes Decisions
Evaluates Processes and ROI

IT SYSTEMS / MANAGEMENT

Model Validation
Model Deployment
Model Monitoring
Data Preparation



BUSINESS ANALYST

Data Exploration
Data Visualization
Report Creation

DATA MINER / STATISTICIAN

Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

PREDICTIVE ANALYTICS TECHNIQUES - EXAMPLES

- **Statistics**:

  Lot of math in the tools available for analytics, methods applied to business problems.

- **Data Mining Models**
  - **Which products are customers likely to buy?**
  - **Which workers are likely to quit/resign/be fired?**
- **Text Models**
  - **What are people saying about my products and services? Can I detect emerging issues from customer feedback or service claims?**
- **Forecasting Models**
  - **How many products will be sold this year, next year?**
  - **How does this break down into each product over the next 3 months, 6 months?**
- **Operations Research**
  - **What is the least cost route for transporting goods from warehouses to final destinations? (PRESCRIPTIVE)**

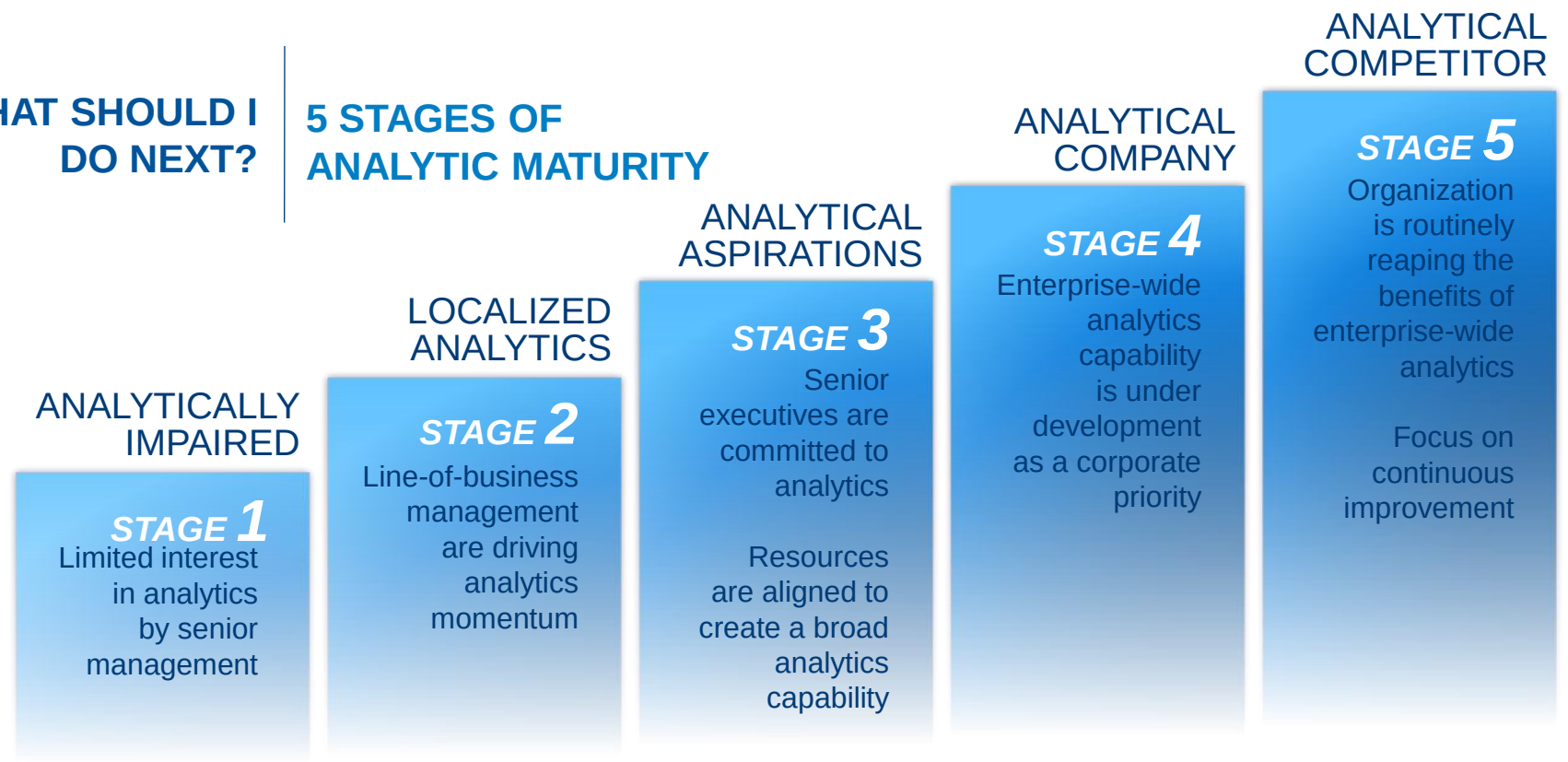
EXTERNAL VIEWPOINT

CHALLENGES IN ANALYTICS ADOPTION

Source: The Current State of Business Analytics: Where Do We Go From Here?
Prepared by Bloomberg Businessweek Research Services, 2011

WHAT SHOULD I DO NEXT?

5 STAGES OF ANALYTIC MATURITY



Source: Davenport, Thomas and Harris, Jeanne. "Competing on Analytics: The New Science of Winning"

High Performance Analytic Trends

- Commercial Support and Acceptance for R and Open Source software
- HPA offerings (Oracle, SAS , IBM, Revolution Analytics, SAP (HANA), RapidMiner, Apache Mahout)
- Partnerships (ex: Tableau and Cloudera, Revolution Analytics and Cloudera or Netezza, SAS & Teradata, SAS & EMC Greenplum, Teradata and Alpine Miner, Microsoft and Hortonworks)
- Everything to Everyone
  - Data Visualization vendors become Big Data vendors
  - Hardware and appliance makers become Analytics experts
- Product stacks become blurry
  - BI vs. Data Viz vs. Big Data vs. Analytics
- FREE Software (R, Hadoop)

High Performance Analytics Vendors

- IBM (InfoSphere Big Insights, InfoSphere Streams, SPSS, Cognos, Netezza)
- SAS (High Performance Analytics / In-Memory Visual Analytics / Access to Hadoop)
- Oracle (Exalytics, Advanced Analytics, OBIEE, Business Applications)
- SAP (HANA, Business Objects, Business Applications)
- Microsoft (Microsoft SQL Server 2012, PowerPivot)
- MicroStrategy (MicroStrategy 9.x)
- QlikTech (QlikView10)
- Tableau (Server and Desktop Edition)
- Tibco (Spotfire Professional Edition and Server)
- Revolution Analytics (RevoScale R, Netezza, Cloudera)

References

- APR (http://apr.apache.org/)
- libConfuse (http://www.nongnu.org/confuse/)
- expat (http://expat.sourceforge.net/)
- pkg-config (http://www.freedesktop.org/wiki/Software/pkg-config)
- python (http://www.python.org/)
- PCRE (http://www.pcre.org/)
- RRDtool (http://oss.oetiker.ch/rrdtool/)
- Hadoop - The Definitive Guide – 3rd Edition
- SAS High Performance Analytics and Visual Analytics - http://www.sas.com/reg/gen/corp/1909596?gclid=COWH-8D8srECFUXc4Aod9wMAXw

Questions