HPC and Large Scale Data Analytics

SOS17 Conference
Jekyll Island, Georgia

Bill Blake
CTO
Cray Inc

March 26, 2013

HPC and Large-scale Data Analytics

Divergence or Convergence?

Supercomputing

- Highest performance parallel computing
- Highly integrated processor-memory-interconnect & network storage
- “Basketball court sized”

And/Or

Large-scale Data Analytics aka Cloud Computing

- Highest performance distributed computing at largest scale
- Lowest cost processor-memory-interconnect & local storage
- “Warehouse sized”

OSTP Press Release on Big Data
(Text excerpted from Full Release issued on March 29, 2012)

OBAMA ADMINISTRATION UNVEILS "BIG DATA" INITIATIVE: ANNOUNCES $200 MILLION IN NEW R&D INVESTMENTS

Aiming to make the most of the fast-growing volume of digital data, the Obama Administration today announced a "Big Data Research and Development Initiative." By improving our ability to extract knowledge and insights from large and complex collections of digital data, the initiative promises to help solve some the Nation's most pressing challenges.

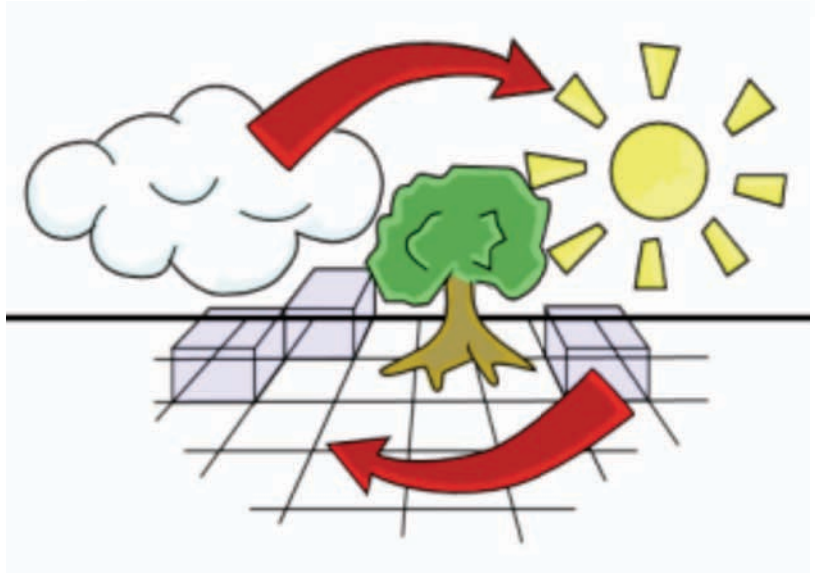
To launch the initiative, six Federal departments and agencies today announced more than $200 million in new commitments that, together, promise to greatly improve the tools and techniques needed to access, organize, and glean discoveries from huge volumes of digital data.

"In the same way that past Federal investments in information-technology R&D led to dramatic advances in supercomputing and the creation of the Internet, the initiative we are launching today promises to transform our ability to use Big Data for scientific discovery, environmental and biomedical research, education, and national security," said Dr. John P. Holdren, Assistant to the President and Director of the White House Office of Science and Technology Policy.

Bold emphasis added by presenter

Need to Create a “Virtuous Cycle”

Cloud provides new distributed programming models that utilize “divide and conquer” approaches with massive scale-out Service Oriented Architectures using local storage and low cost hardware, and new data analytics algorithms where data scientists claim “the larger the data the simpler the algorithm”



HPC provides new parallel programming models, highly scalable Global Memory Architectures supported by highest BW, lowest latency interconnects, new algorithms for high fidelity modeling and simulation that assimilate (sensor) data and highly iterative processing of both capability and capacity workloads that additionally support data mining and knowledge discovery

Image courtesy of University of Michigan –Atmospherics Dynamics Modeling Group

Exploring the Big Data Space

Big Data

Big Data and Why We Should Care

Big Data refers to data that is not easily captured, managed and analyzed by traditional tools due to:

- *Volume* (growing > 60%/yr),
- *Velocity* (often real time streaming),
- *Value* (data sizes beginning at > 100TB)
- *Variety* (all forms of unstructured data: logs, docs, images)

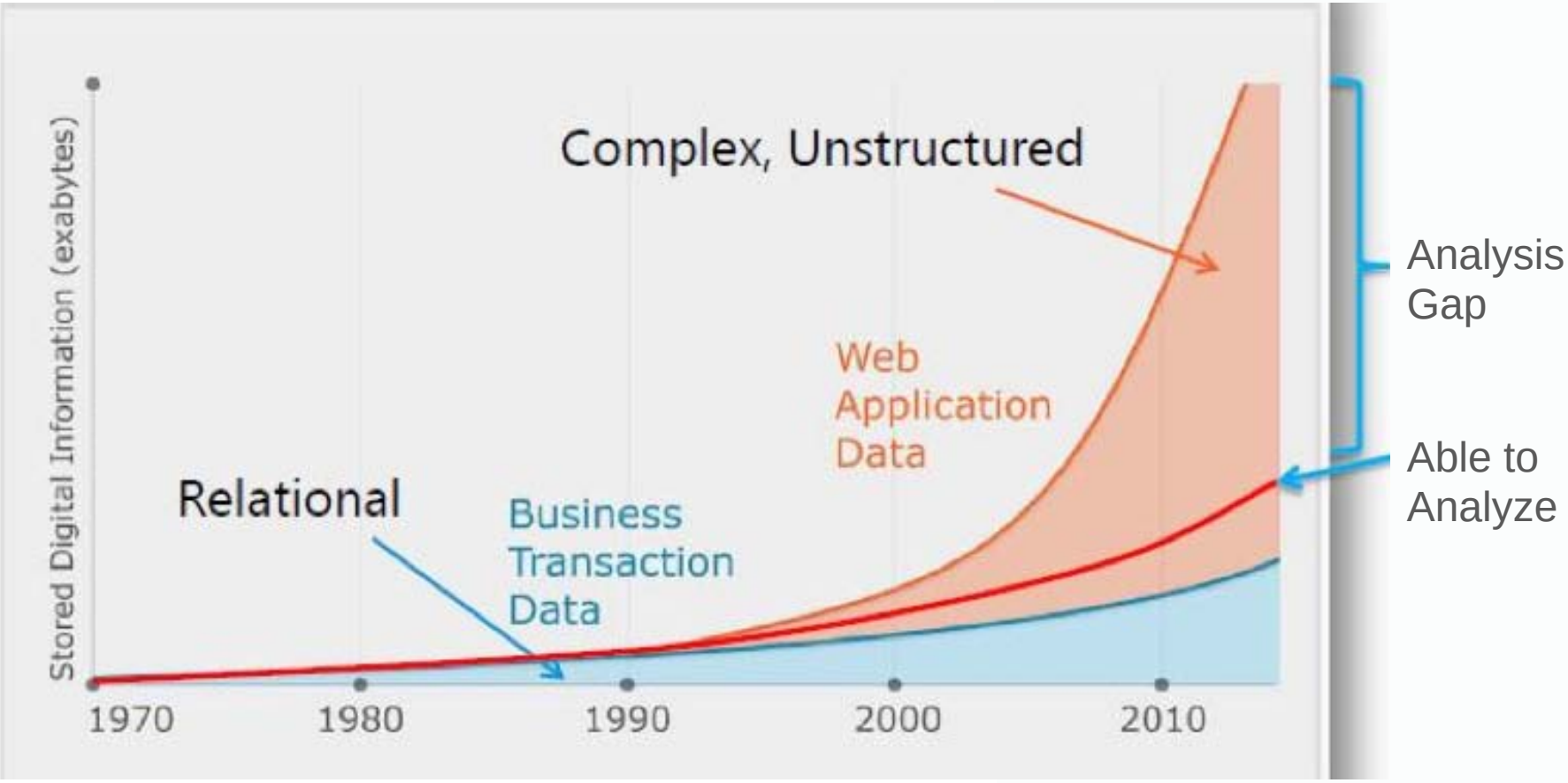
IDC expects Hadoop, an enabling technology, to run on over 50% of Big Data Projects over time representing a $8.5B market by 2015

Science will increasingly be (sensor) data-driven to understand the world

Business will increasingly be data-driven to understand customers

Big Data Means New Kinds of Data

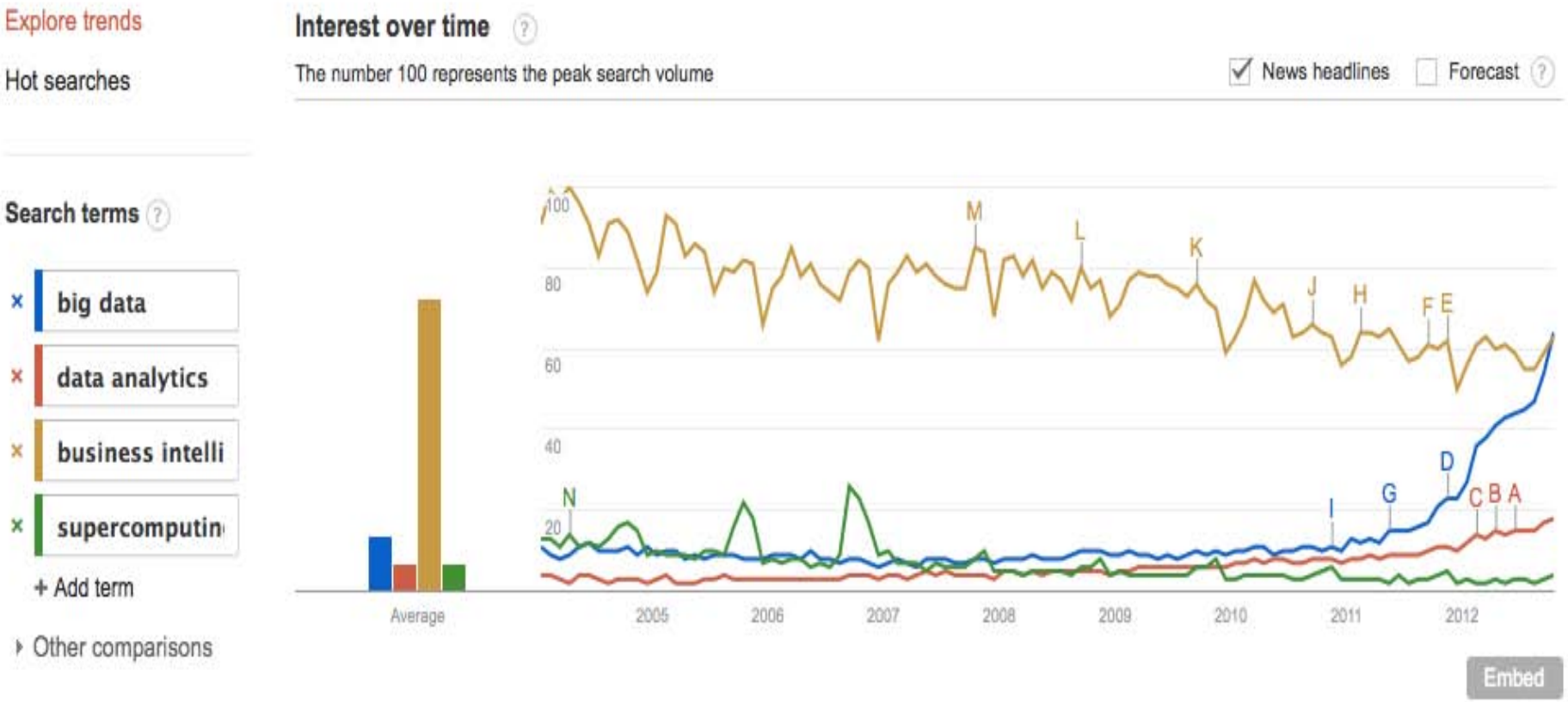
EMC estimates that by 2020 there will be 40,000 Exabytes of data created, although the majority of that data will not be created by humans but sensors



Source: IDC White Paper sponsored by EMC May 2009

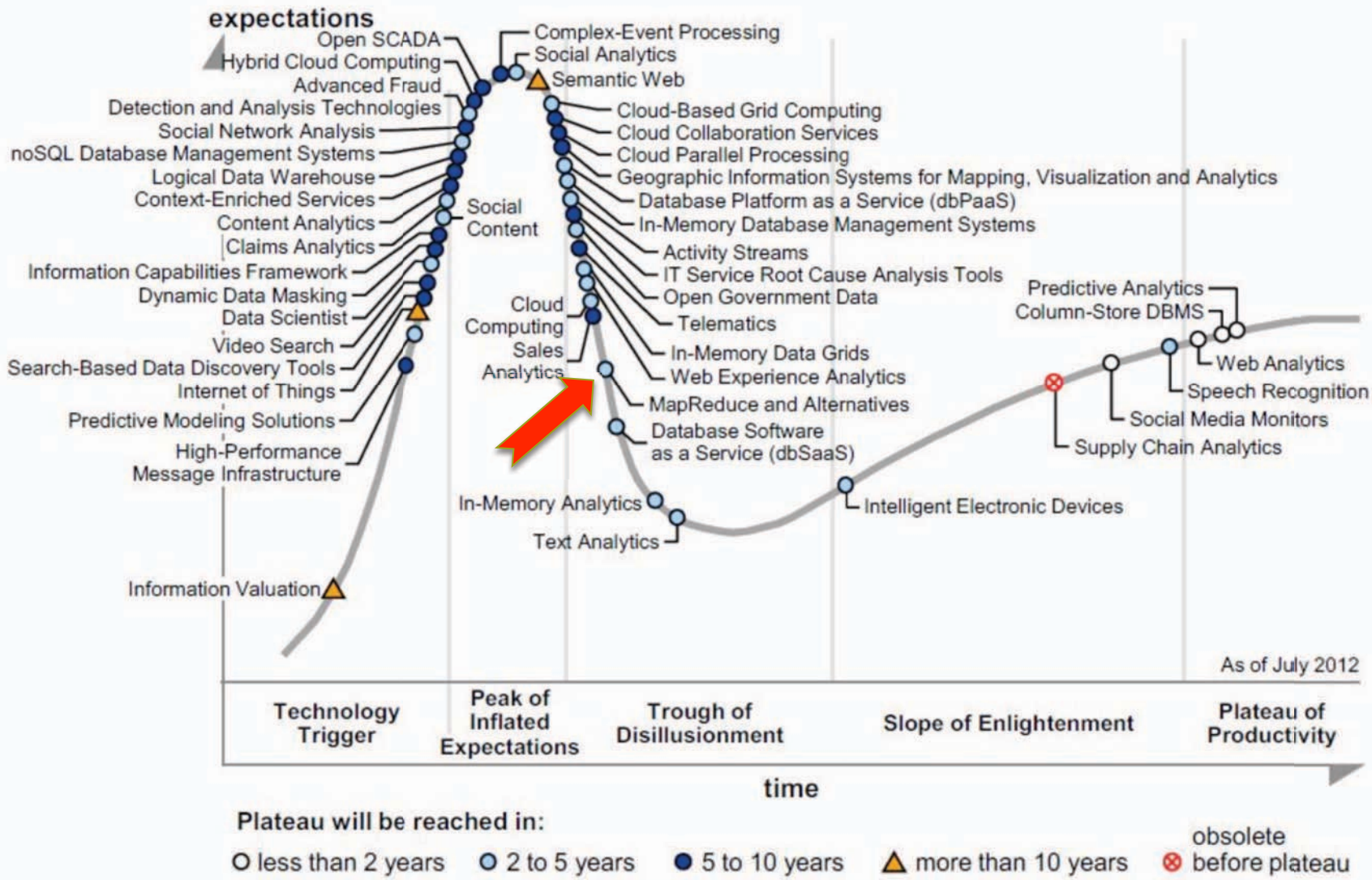
Interesting Trends

Google Trends shows "Big Data" search volume recently exceeded "Business Intelligence" search volume while "Supercomputing" search volume drifts lower



Descending Into Gartner's "Trough"

Figure 1. Hype Cycle for Big Data, 2012



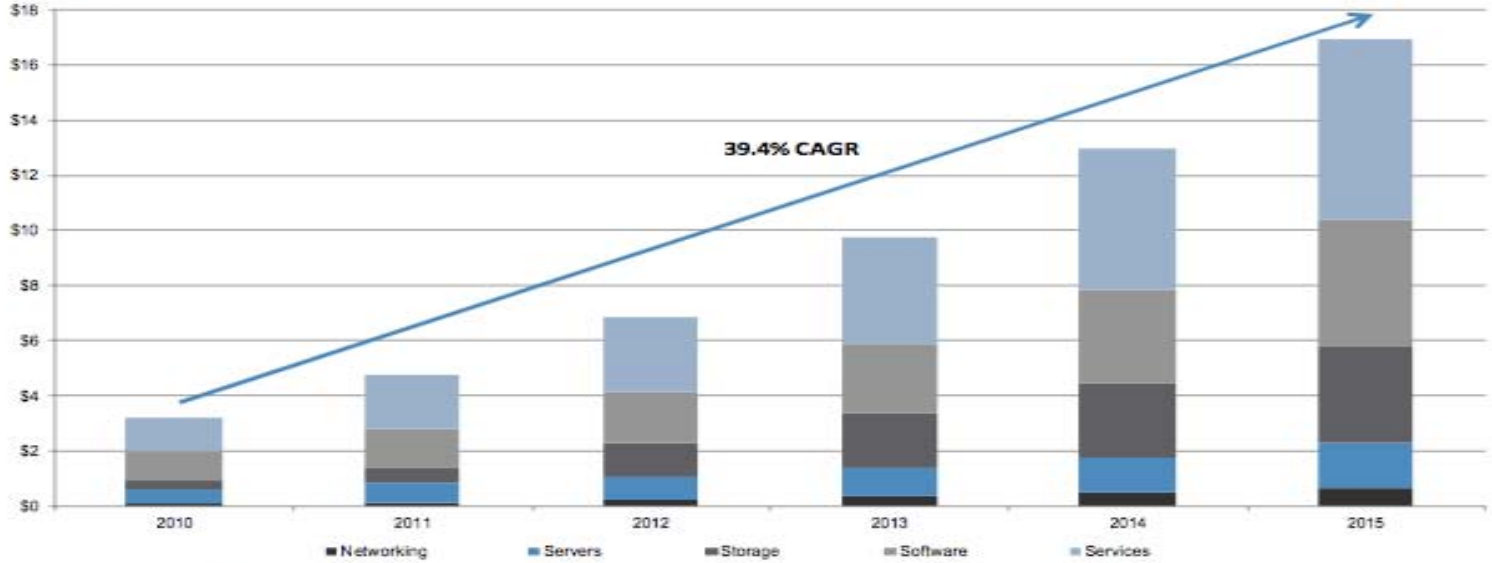
Source: Gartner (July 2012)

Long Term View of Possible Revenue/Spending

(Note: WW Big Data growth of 40% CAGR is forecast to continue to 2018 and reach $48B)

Figure 5: Worldwide Big Data Technology and Services Revenue by Segment (2010-2015)

$ billions



Source: IDC and J.P. Morgan.

	2013	2014	2015
Server BD TAM	$4.0B	$5.0B	$7.0B
Storage BD TAM	$3.0B	$3.5B	$4.0B
Hadoop 50% of BD TAM	$3.5B	$4.25B	$5.5B

Gartner Group that enterprises that invest and use Big Data will be 20% more effective in all financial metrics by 2015

Big Data Landscape

Vertical Apps

PREDICTIVE POLICING
bloomreach
MYRRIX

Log Data Apps

splunk> loggly sumologic

Ad/Media Apps

rocketfuel
collective[i]
bluefin
Recorded Future
LuckySort
Media Science
TURN
DataXu

Business Intelligence

ORACLE | Hyperion
SAP Business Objects
RJMetrics
Microsoft | Business Intelligence
IBM COGNOS
birst
MicroStrategy
Autonomy
QlikView
bime
Chart.io
GoodData

Analytics and Visualization

tableau
Palantir
OPERA metaLayer
METAMARKETS
dataspora
centrifuge
TERADATA ASTER
sas
TIBCO
KARMASPHERE
panopticon Real-Time Visual Data Analysis
pentaho
Datameer
platfora
ClearStory
CIRRO
alteryx
visual.ly
AYATA

Data As A Service

factual
kaggle
knoema beta
GNIP
DATASIFT
Windows Azure Marketplace
INRIX
LexisNexis
LOQATE

Analytics Infrastructure

Hortonworks
VERTICA
MAPR
cloudera
INFOBRIGHT
PARACCEL
EMC²
GREENPLUM
NETEZZA
kognitio
DATASTAX
EXASOL
calpont

Operational Infrastructure

COUCHBASE
10gen
TERADATA
HADAPT
TERRACOTTA
VoltDB
MarkLogic
INFORMATICA

Infrastructure As A Service

amazon web services
Windows Azure
infochimps
Google BigQuery

Structured Databases

ORACLE
MySQL
Microsoft SQL Server
PostgreSQL
IBM DB2
SYBASE
memsql

Technologies

hadoop
hadoop MapReduce
mahout
APACHE HBASE
Cassandra

dave@vcdave.com
blogs.forbes.com/davefeinleib

Big Data Changes Over Time

Initially, Transaction Analytics (OLAP) with ad hoc SQL queries on structured data in relational databases by *analysts* producing BI Reports

Looking at all the data, O(100) TB, all the time

Recently, Textual Analytics (MapReduce) providing API for analysis of unstructured data in log files by programmers seeking “long tail” results/insights

Looking at all the data, O(1000) TB, once at a time

And emerging, Graph Analytics (RDF, OWL) with ad hoc SPARQL queries on linked data by analysts seeking discovery via hypothesis

Looking at all the data, O(100) TB, and relationships

It is Really About Decision Making through Fact Finding and Equation Solving

Key Function	Language	Data Approach	"Airline" Example
OLTP	Declarative (SQL)	Structured (relational)	ATM transactions Buying a seat on an airplane
OLAP Ad Hoc	Declarative (SQL+UDF) or NoSQL	Structured (relational)	Business Intelligence analysis of bookings for new ad placements or discounting policy
Semantic Ad hoc	Declarative (SPARQL)	Linked, Open (graph-based)	Analyze social graphs and infer who might travel where
API for analysis	Procedural (MapReduce)	Unstructured (Hadoop files)	Application Framework for large scale weblog analysis
Data Assimilation	Procedural (C++, Fortran)	Data merged With simulations	Sensor data incorporated into the computer simulation
Optimize Models	Procedural (Solver Libs)	Optimization <-> Simulation	Complex Scheduling Estimating empty seats
Simulate Models	Procedural (Fortran, C++)	Matrix Math (Systems of Eq's)	Mathematical Modeling and simulation (design airplane)

Analyst Query

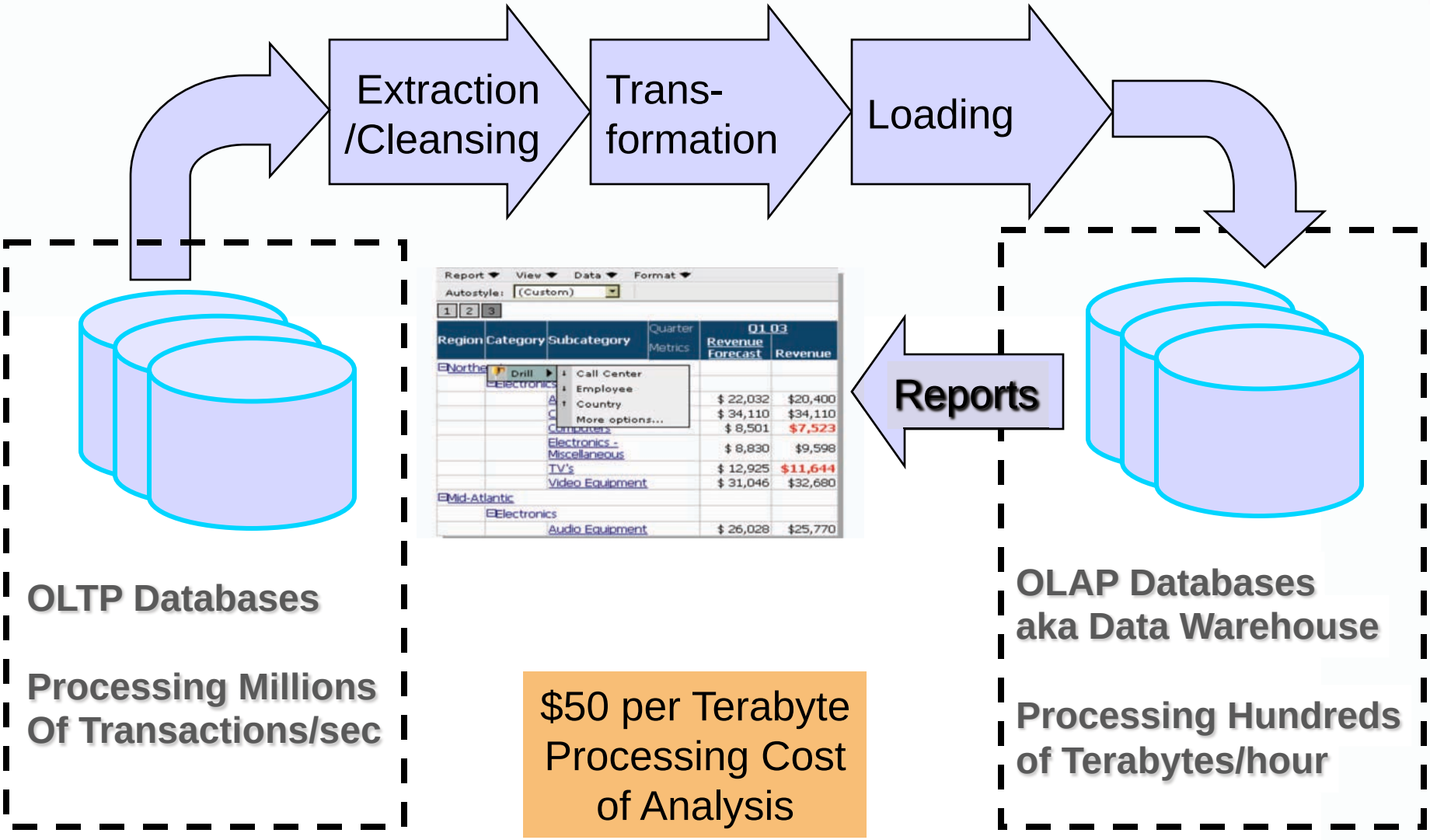
Languages & Tools for Programmers

The Online Analytical Processing Problem

Key Function	Language	Data Approach	SMP Server	Cluster And MPP	Cloud And Grid	Web Scale
OLAP Ad Hoc	Declarative (SQL+UDF)	Structured (relational)				

- Business Intelligence Applications generate reports of aggregations
  - Need to **read** at all the data all the time (telecom, retail, finance, advertising, etc)
- BI Analytics require **ad hoc queries** since you don't know the next question to ask until you understand the answer to the last question
- Standard SQL is limited by the Algebraic Set Theory basis of RDBMS, if you need Calculus then insert User Defined Functions into the SQL
- Programming models in conflict as Modeling and Simulation combine with Business Intelligence in Predictive Analytics

Operational vs. Analytical RDBMS



Shifting from Analyst to Programmer

Key Function	Language	Data Approach	SMP Server	Cluster And MPP	Cloud And Grid	Web Scale
OLAP Ad Hoc	Procedural (MapReduce)	Unstructured (Hadoop Files)				

Google, Yahoo and their friends are using a data-intensive application frameworks to analyze very large datasets (e.g.,weblogs) and not transactions or structured data. What will this look like in 10 years?

- MapReduce/Hadoop: an programming model/application framework performing "group-by (map), sort and aggregation (reduce)"
- Involves not queries, but programs willing to forgo the need for transactional integrity or the performance of using structured data (**4X-5X disadvantage on equal hardware**, but with excellent scaling on cheap hardware)
- An increasingly popular approach with organizations that have the programming talent to use it, especially research organizations

What's so important about MapReduce and Hadoop?

Map and Reduce are artifacts of functional programming from the days of the Lisp Language

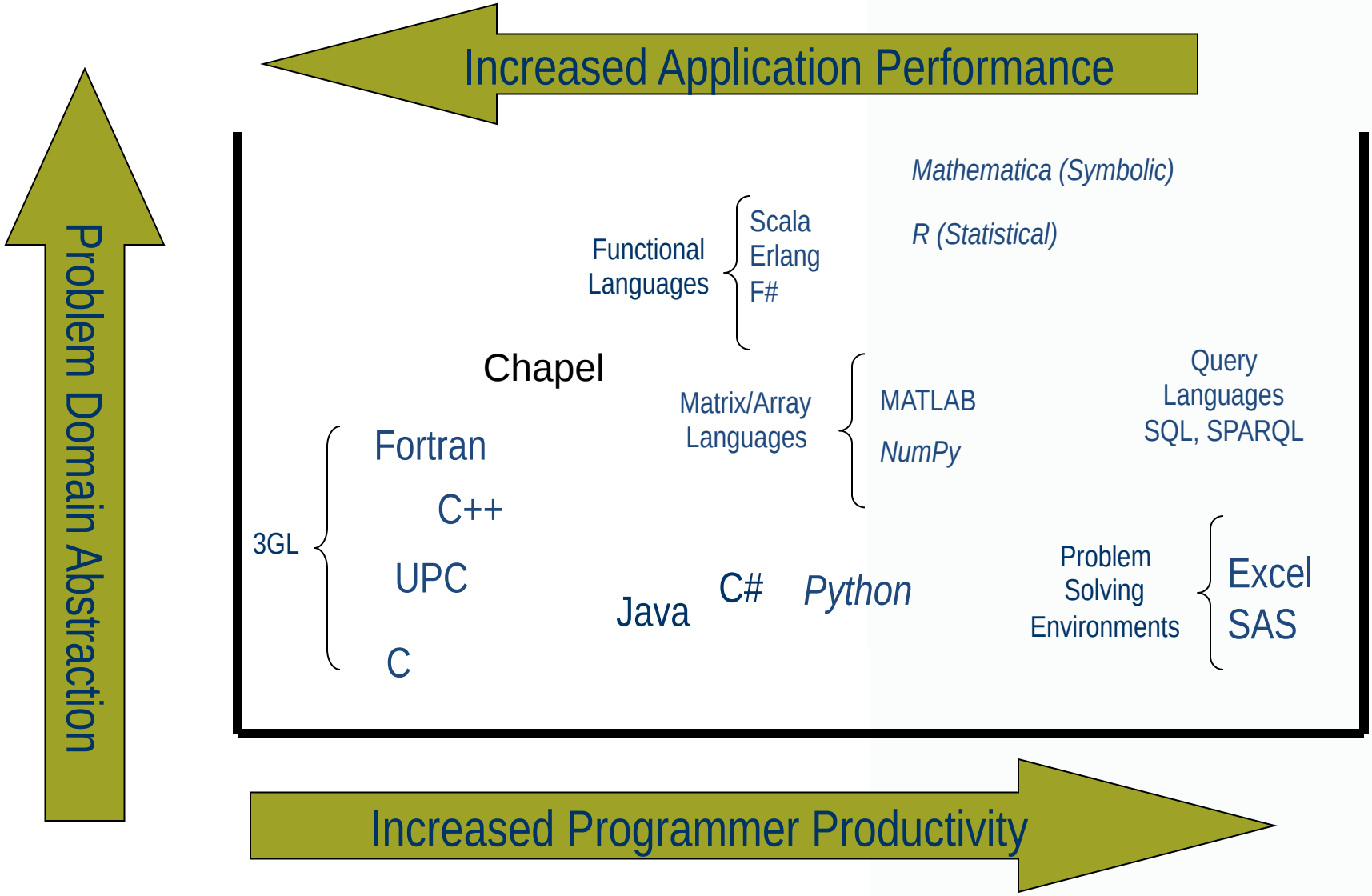
- Cleverly adapted to data-intensive distributed computing by Google and Yahoo
- **Programming model** separates the "what" is being computed from the "how" it is being computed
- "Our goal was to achieve highly scalable, high performance analysis ***without*** supercomputers" Raymie Stata, ex-CTO Yahoo

Ideally suited for the analysis of textual data without the requirement to fit the structured data of the transactional DB world

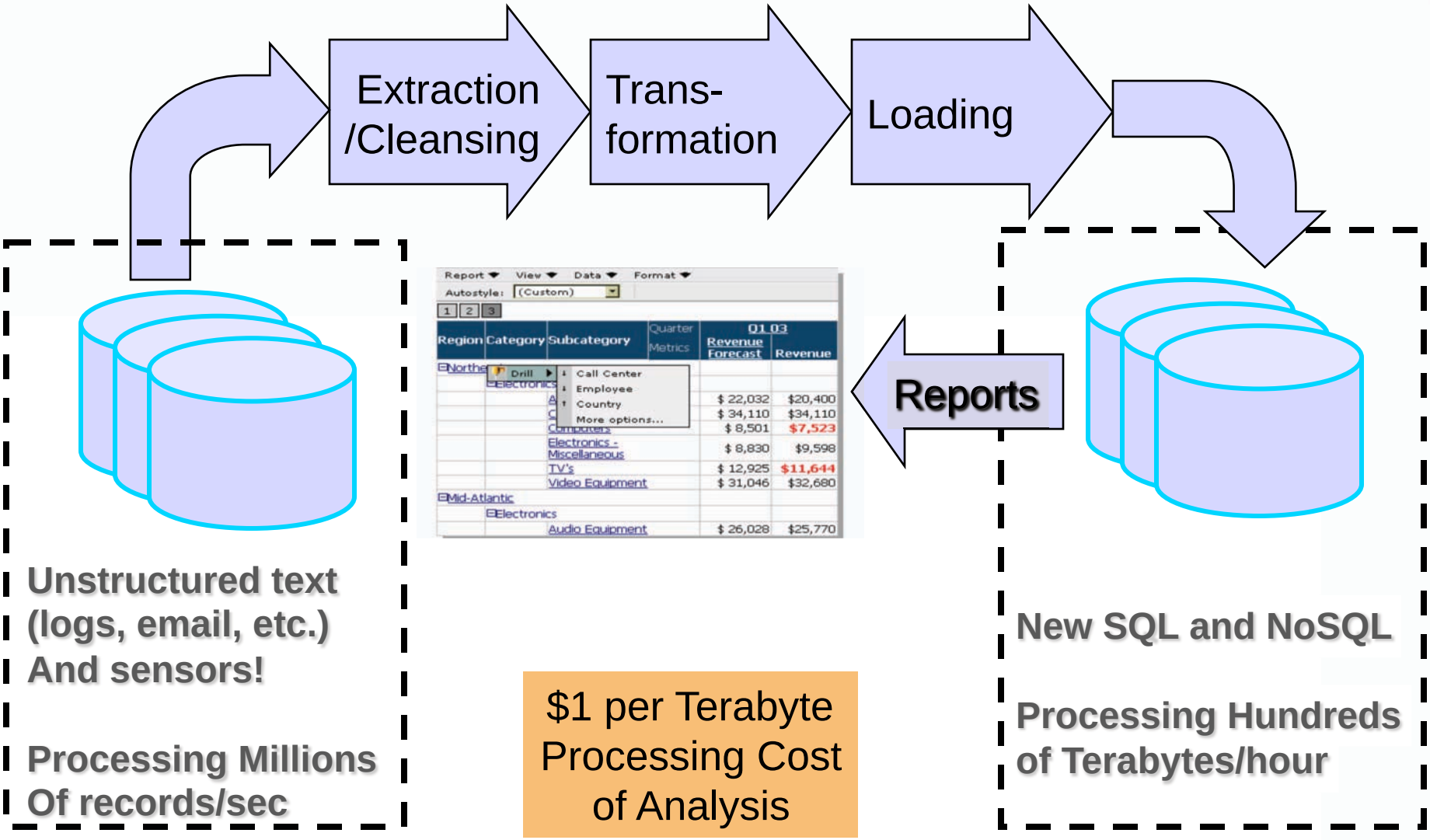
- Lots of data redundancy (3X in HDFS) helps with the frequent failures of cheap hardware at very high node counts

MR-Hadoop is a data-intensive application development framework for C++ programmers not a tool for BI analysts

A Better Way to Express Analytics?
Queries and Program Code Do Not Mix Well



Hadoop's Role Will Focus on Data Ingest



MapReduce on Cray Systems

Excellent progress at NERSC providing MapReduce capability on their XE6 systems

- New capabilities for job policies and the run-time environment to support very large numbers of Joint Genome Institute jobs.
- Utilize the Cray Cluster Compatibility Mode (CCM) to support tools like Java and support a throughput oriented scheduling environment
- An excellent real-world example of meeting the needs of the data-intensive community in the world of traditional simulation and modeling

Sandia's Development of MapReduce in MPI (MR-MPI)

- MapReduce functionality implemented in MPI context (no Java)
- MR-MPI library performs data movement between processors and supports and requires local disks for "out of core" large data sets

Active Collaborations underway with DOE labs and NSF sites

Comparison of MapReduce Runtimes

ref: Steve Plimpton and Karen Devine at Sandia National Labs (MR-MPI)

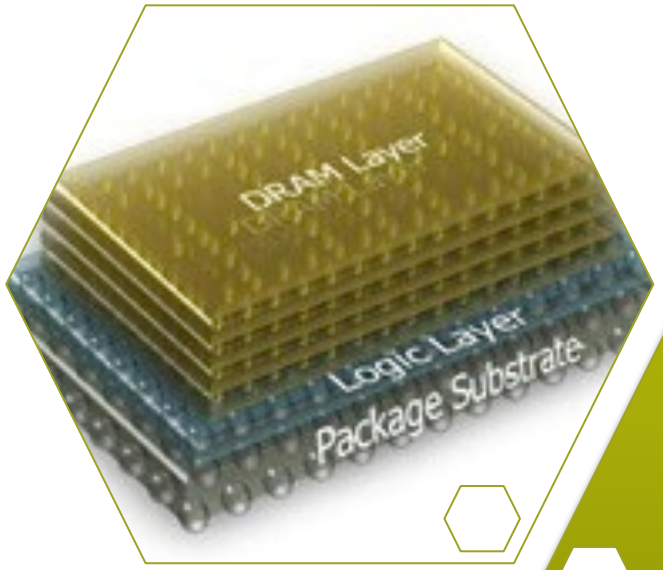
	MapReduce	MapReduce	Custom Code
Runtime	HADOOP	MR-MPI	XMT-1/C
SW Stack	Java+Linux	XT CLE/MPI	XMT C+MTGL
Storage	Local (HDFS)	Network-based	In core only
Resilience	3X redundant	Limited by MPI	No failover
Coding Effort	Low (Java)	Medium (C++)	High
Processors	48	48	32
Sandia result: Single-source shortest path algorithm using WebGraphB 187M vertices 532M edges	38,925 sec Run on x86 low cost cluster	13,505 sec 8,031 sec with enhanced algorithm on XC-30	37 sec < 10 sec results expected on uRiKA

Advancing from Generating Reports to Inference and Discovery

Key Function	Language	Data Approach	SMP Server	Cluster And MPP	Cloud And Grid	Web Scale
Semantic Ad hoc	Declarative (SPARQL)	Linked, Open (graph-based)				

- The International W3C standards body has approved the key standards, called RDF and OWL to support the Semantic Web aka Web 3.0 with "machine readable" open linked data
- Future databases will use triples (subject-predicate-object) vs tables and with RDF/OWL federate heterogeneous data
- Future databases will support **reasoning** not just **reporting**
- This work started as a combined European Defense and DARPA effort
- **Major RDBMS vendors are admitting Relational and XML are ill-suited to the needs of the semantic web of the future**

Shifting Focus to HPC



HPC

The “Big Data” Challenge

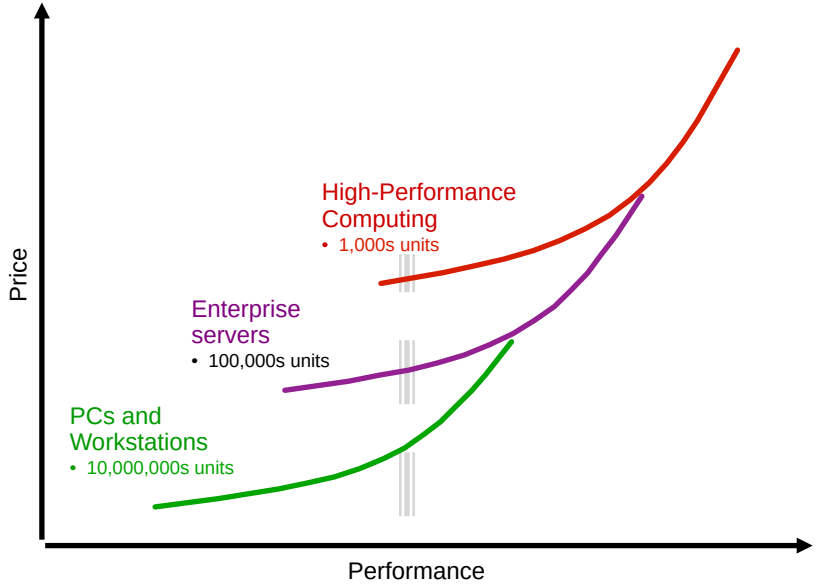
Supercomputing minimizes data movement
the focus is loading the “mesh” in distributed memory, computing the answer with fast message passing, and visualizing the answer in memory – the high performance “data movement” is for loading, check pointing or archiving.

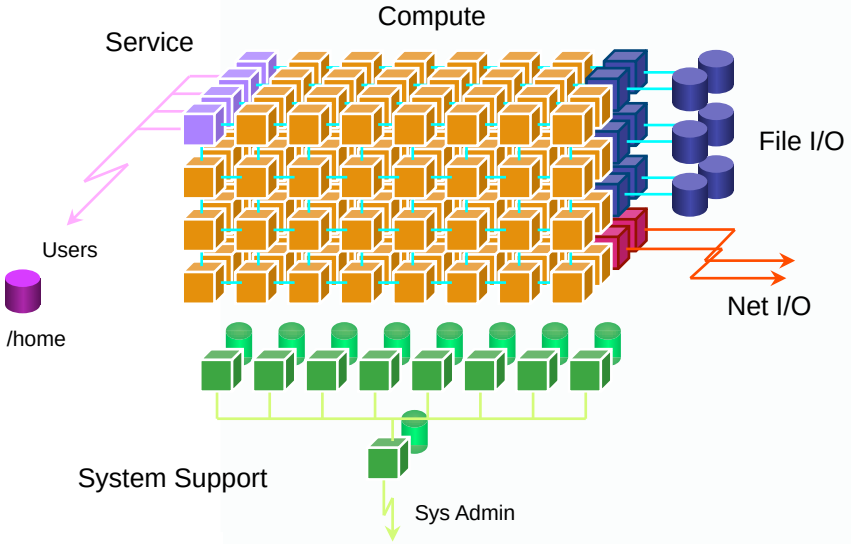
Data-intensive computing is all about data movement
scanning, sorting, streaming and aggregating *all the data all the time* to get the answer or discover new knowledge from unstructured or structured data sources.

Maximum Scalability: System Node Specialization

Key to Cray's MPP scalability is system node o/s specialization combined with very high bandwidth, low latency interconnects

8





Courtesy of Dr. Bill Camp, Sandia National Laboratories circa 2000

Big Data is “Data in Motion”

Set the stage for the fusion of computationally intensive and data intensive computing in future Cray systems

Build on Cray's success of **delivering the most scalable systems through heterogeneous and specialized nodes**

- Nodes not only optimized for compute, but also storage and network I/O, all connected with the highest level of interconnect performance. Add system capability to the edge of the fabric

There is an opportunity to **increasing key analytic application performance with an "appliance style" approach** using Cray's primary supercomputing products with extensions configured as optimized HW/SW stacks – adding value around the edge of the high performance system network

The Netezza Approach to Appliances: Moving Processing to the Data

Active Disk architectures

- Integrated processing power and memory into disk units
- Scaled processing power as the dataset grew

Decision support algorithms offloaded to Active Disks to support key decision support tasks

- Active Disk architectures use stream-based model ideal for the software architecture of relational databases

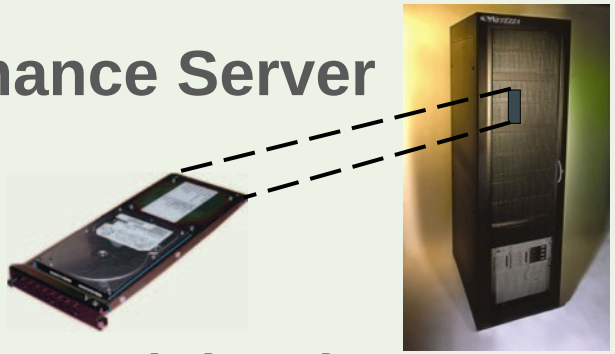
Influenced by the success of Cisco and NetApp appliances, the approach combined software, processing, networking and storage leading to the first database warehouse appliance!

Netezza is an IBM Company

E. Riedel, “Active disks: remote execution for network-attached storage,” Ph.D. dissertation, Pittsburgh, PA, 1999.

Active Disks as Intelligent Storage Nodes

Netezza Performance Server



Netezza added:

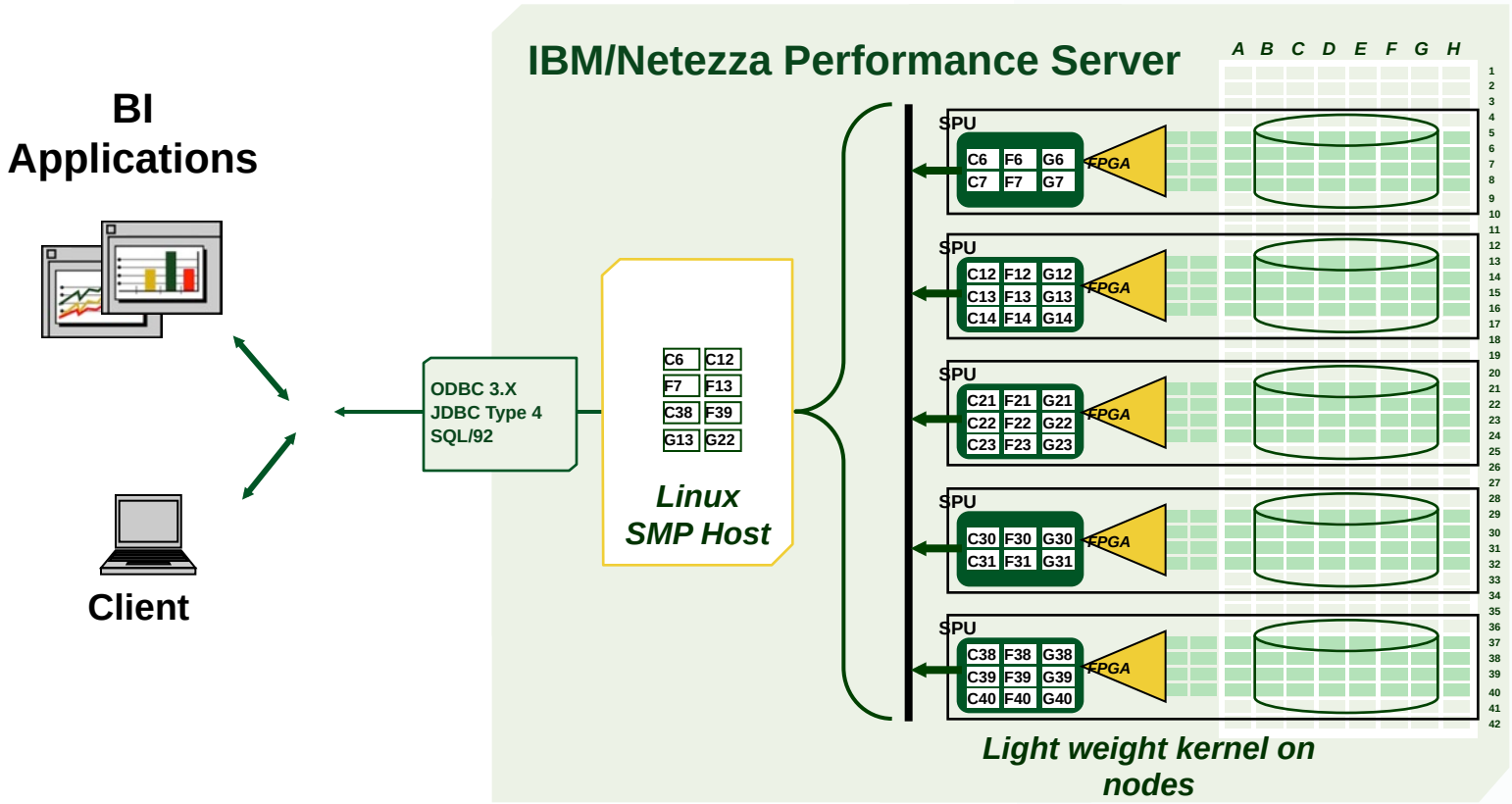
- Highly optimized query planning
- Code generation
- Stream processing

Result: 10X to 100X performance speedup over existing systems

Snippet Processing Unit (SPU)

A compute node for directly processing SQL queries on tables

Streaming Data Flow Unclogs the Bottleneck



Postgresql RDBMS is sole application running on the server with complete control over active disk storage nodes → DW appliance

Key to performance of ad hoc queries on the VLDB was delivering the highest bandwidth access to tables on disk

Another Active Storage Idea

Very Interesting research has been published from HP Labs, ORNL, U Michigan and Georgia Institute of Technology on future Data-Centric system architectures:

- Utilization of “active” NVRAM in storage hierarchy
- On-line processing and “data staging” where I/O and data movement actions are enhanced with computation to better filter, reduce, sort, compress data.
- Applications running with this “offload” have shown 3X to 4X speedups
- The next thing after “active” Disks?

S. Kannan, A. Gavrilovska, et al. “Using Active NVRAM for IO Staging” Proceeding PDAC '11
Proceedings of the 2nd international workshop on Petascale data analytics: challenges and opportunities

Future File Systems and Storage

Cray is engaging the research and vendor community to address major Exascale and Big Data file system issues

- How do we address the bandwidth requirements at scale?
- POSIX metadata requirements and alternatives

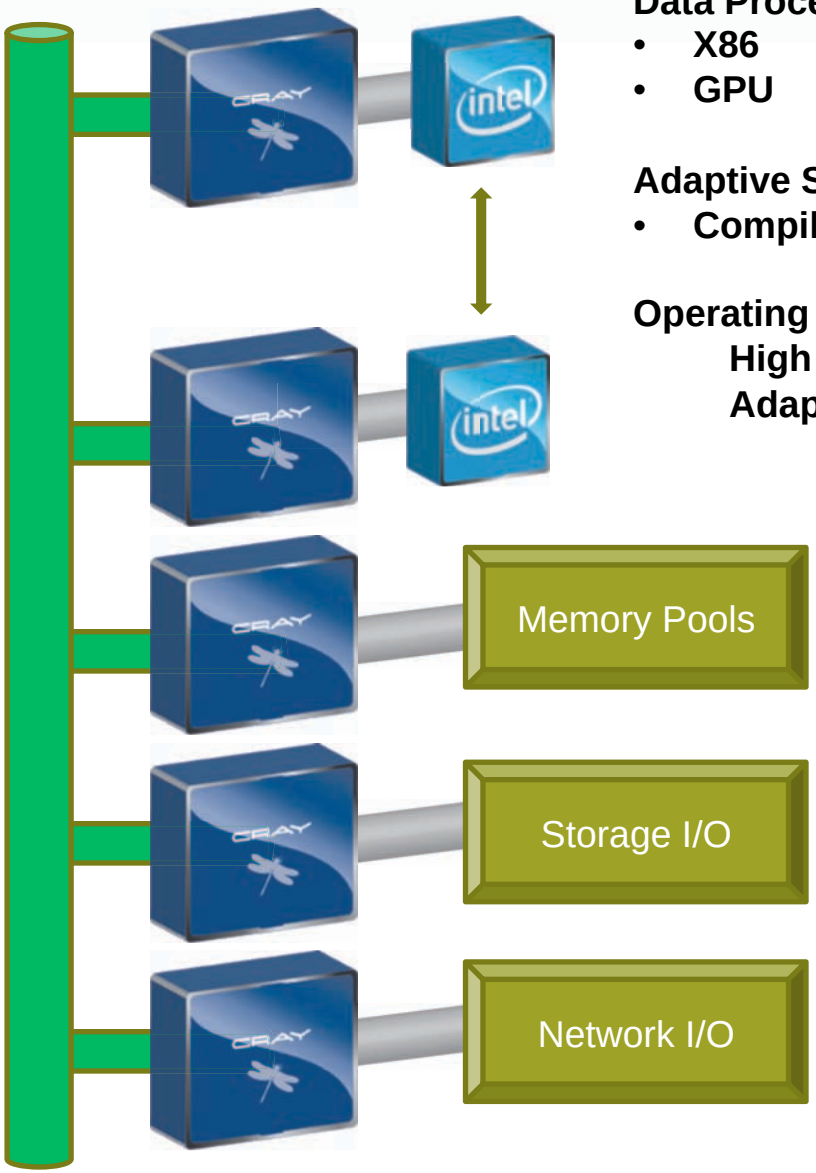
Hierarchical Local Storage Strategies

- Burst Buffers optimize compute node access to reduce required external I/O bandwidth – checkpoint to buffer as archive process
- Checkpoint to archive for saving global memory state

Dramatic Changes in Memory/Storage Hierarchy over next 5 years

- Possibilities include:
  - DRAM becomes just another level of cache?
  - 1+ TB NVRAM per socket becomes “working memory”?
  - Possibility of “active NVRAM” with addition of local processing
  - Multi Terabyte SSD
  - 50+ Terabyte per spindle HDD (HAMR)

Adapting to Data-Intensive Computing: Adding Value at the Edge of the Network



Data Processors
- **X86**
- **GPU**

Adaptive Supercomputing
- **Compilers, Auto-tuning,**

Operating System
High throughput Scheduling
Adaptive Runtime, Network

High Performance Memory
- **DRAM**
- **Active NVRAM?**

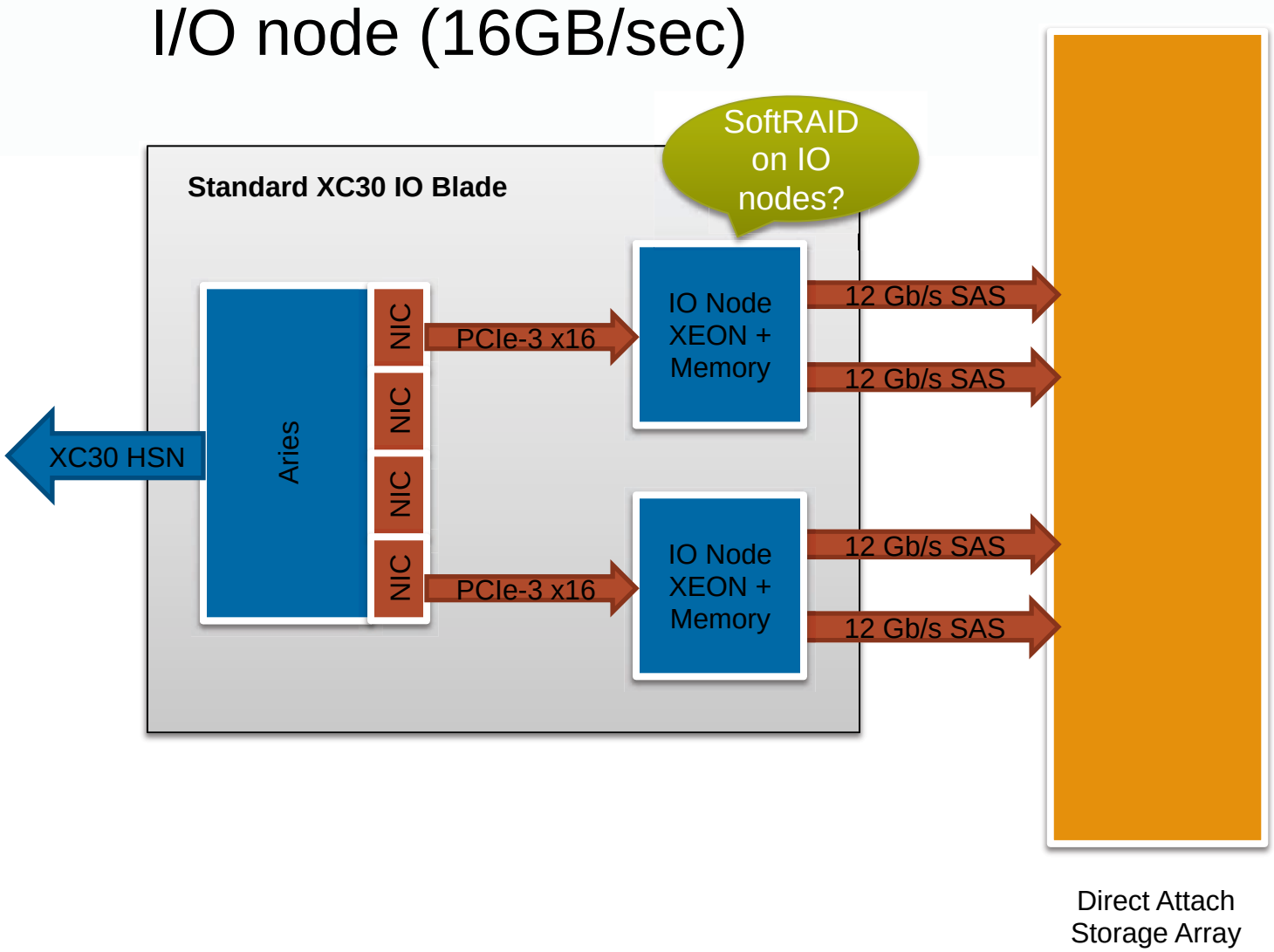
Storage
- **SAS to PCIe**
- **Tightly coupled Software Stack for RAID**
- **Very Low Latency**

Intelligent I/O
- **Protocol Load Balancing**
- **Specialized Data Ingest**
- **Policy Based Data Movement**

Software Defined Network

The fastest Interconnects will be key to analytic workloads

Direct Attached Disks and Software Controlled RAID



Concluding Comments

- Warehouse scale distributed computing, aka Cloud, provides an excellent multi-tenancy resource for high throughput capacity computing
- Cray expects capacity workloads, that run fine on up to 500 to 1000 node Ethernet connected clusters, will increasingly migrate from cluster to cloud
- But highly parallel **analytic** workloads, especially those that require low latency messaging and/or global memory operations that benefit greatly from the high performance interconnects and tight integration of MPP machines, will not migrate from MPP to Cloud
- We do expect many Cloud developments to “condense” into future MPP systems, including programming models, software defined networks, and hypervisors that combined with the high performance message passing and global atomic memory support in networks such as the Cray Aries network will best support **the fusion of HPC and large-scale analytics**

Thank You!

bill.blake@cray.com

Declarative Versus Imperative (Language)

SQL

SELECT *list* FROM *table*
WHERE *condition*
GROUP BY category

MapReduce

Map() function selects data
<There are no tables>
Reduce() handles grouping