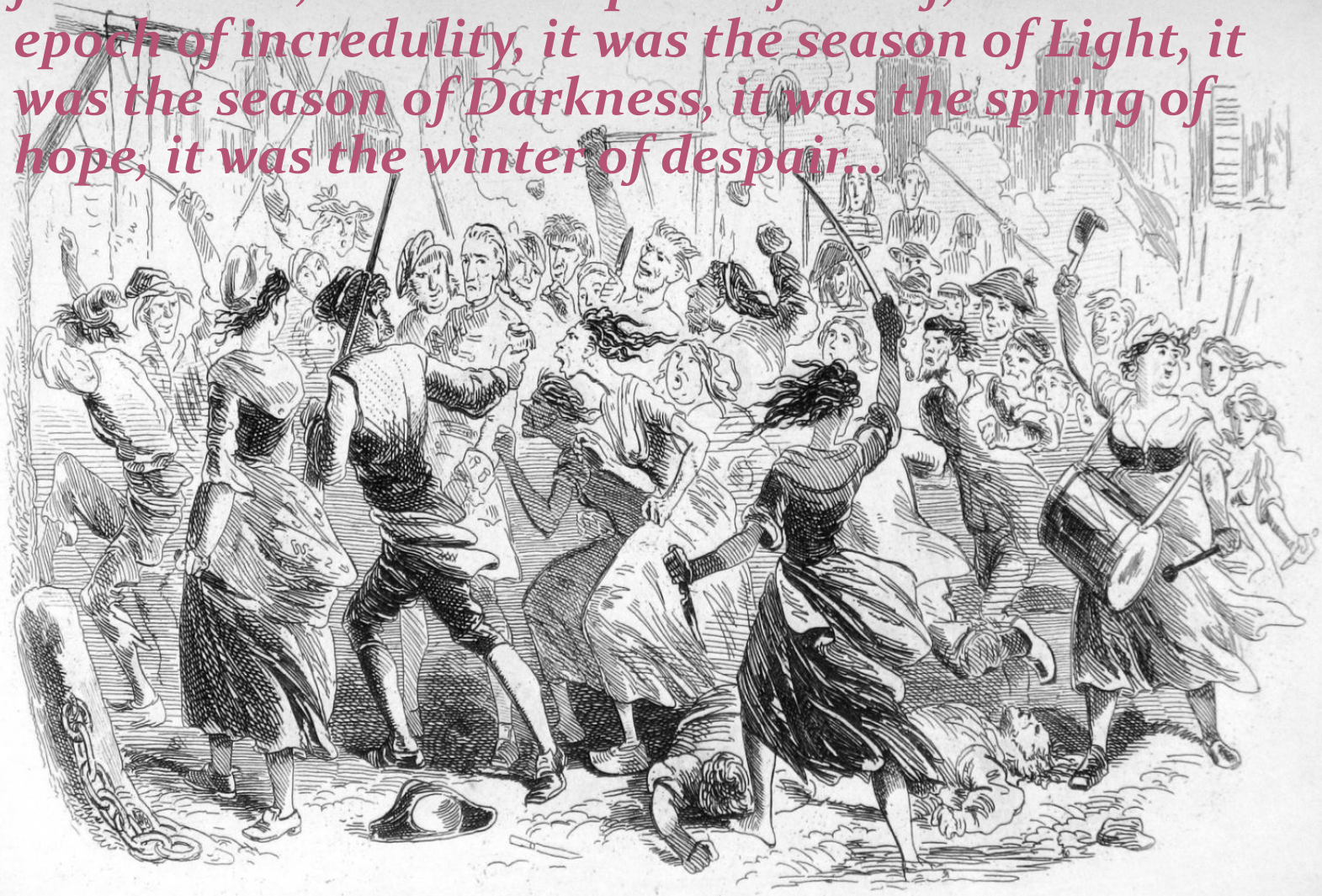
"Database"?? Research Directions The User's View

Todd Walter
Chief Technologist
Teradata Corporation

Disclaimer

- The following solely represents the opinions of Todd Walter not the opinions of Teradata Corporation
- Nothing in this document may be construed to be a promise of future functionality, capability or product from Teradata Corporation at any point in the future.

- *"It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair…*



- *--- Charles Dickens – "A Tale of Two Cities"*

Lot's to Celebrate – Real Implementations

- Temporal
- Integrated Spatial
- Hybrid Storage
- Columnar
- In Memory
- Optimization algorithms and techniques
- Very large relational databases implementations
- …

Lot's of Opportunity
Marvelous Time to be a Researcher

- Open Source
- Commodity platforms
- Explosion of fast Research in industry
- Explosion of desire to analyze and get value from data

But Only Available to the 0.1%

- Developers building tools for developers
  - Users quickly being left behind
- Articles about how data is being locked away behind a small number of data gurus
  - Data Scientist quickly becoming a bad word
  - An expensive person who hands out crumbs
- Computer scientist required to use the new tools
  - And bring all the parts together
- Remember the charts about how everyone in the world was going to have to become a Cobol programmer to catch up with the backlog of applications?
  - Not sustainable or scalable
  - What about the other 99.9% of organizations??

Organizations Face Several Obstacles in Achieving a Unified Data Environment

Difficulty deploying and integrating new systems

Difficulty managing multiple systems, new types of data

Difficulty providing accessibility to fast insights on big data

Hard to find right skills; Lack of supportability for new systems & "data scientists"

Source: Big Analytics 2012 Survey, Teradata

Data Integration is a Big Part of IT Being Turned Upside Down

- "CMO will spend more on IT than the CIO"
  - Means that the users are rebelling against the IT stranglehold
- Slow waterfall development model doesn't match speed of business
- Data Integration is one of the key impedances
  - 6 months to add a new attribute to the data warehouse!!!
  - Tools and processes not aligned to deliver at rate of business requirements
  - Big pushback against data modeling, data quality,… - perceived to be the cause of lack of responsiveness
- Much wider range of Data Quality expectations/tolerances than ever before
  - Sometime correctly, sometimes not
- Non-integrated data leads to analysts spending their time doing it on demand
  - Many times worse in the big data space – but people feel freedom/control from schema on read

Users are in Desperation Mode And Companies are Taking Big Risks

- "Good Enough" is the macro trend
- Companies that before were ultra conservative in adopting a new release let alone a new technology are betting the business on beta versions of stuff
- Huge backlash coming when the inevitable failures happen
- But people have jobs to get done
- 433 data management products in our inventory, changes daily!

Where is the New Frontier?

- Bring the embarrassing riches of the modern analytic environments to the masses
  - And that does not mean putting a weak veneer of SQL on top of Hadoop…
- Where is all the analysis really done today?
  - Excel, Access, R, Matlab
  - Where are the scalable but usable algorithms?
- Collaboration among analysts
  - Chorus, Sharepoint
  - Sharing at the level of the analytics being done
  - Weak but a start
- New tools
  - Visualization (eg Tableau, Spotfire,…)
  - GUI query/orchestration (eg Alteryx, BioFortis,…)
  - New interfaces to describe the problems and orchestrate the underlying technologies

Thoughts on Technologies

Platform Technologies Specialization vs Cost

- Macro trend is to minimize cost per server
  - Cloud, Hadoop,…
  - Organizations minimizing cost today at all costs
  - Huge clusters/clouds running <<20% of capacity!
  - Vastly over provisioning to cover for inefficiency
    - But the servers are cheap by all IT metrics
- When will the backlash occur? Will it?
- Until then, specialization is counter-indicated
  - GPUs, hybrid storage, SSDs,…
  - Space optimized server packaging

Big Discontinuity Coming

- Disk capacity growth slowing or stopping
  - And new tradeoffs for disk users
- But CPU power (cores) continue on Moore's law curve
  - But current clusters, clouds are already over provisioned on CPU
- Current standard architecture for cluster server will begin to break down
  - Where will it go?
  - SSDs,… too expensive to add IO per server
  - Small space efficient servers have less disk space
  - Network attached storage?
    - Breaks the fundamental model but might be necessary
    - Required network bandwidth/scalability is available but expensive

CPUs

- CPUs flat
  - Potentially a downshift on a per core basis
    - ARM, micro servers, et al
  - Puts a lot of pressure on massively scaling algorithms
  - On single threaded things like query optimization
  - And exacerbates skew
- CPU over provisioned in many cases
  - How do we change our view to utilize CPU to cover for weakness elsewhere?
  - To save IO
  - To save network
  - To protect data
- Many researchers still trying to optimize CPU – why bother?

GPUs and Co-Processors

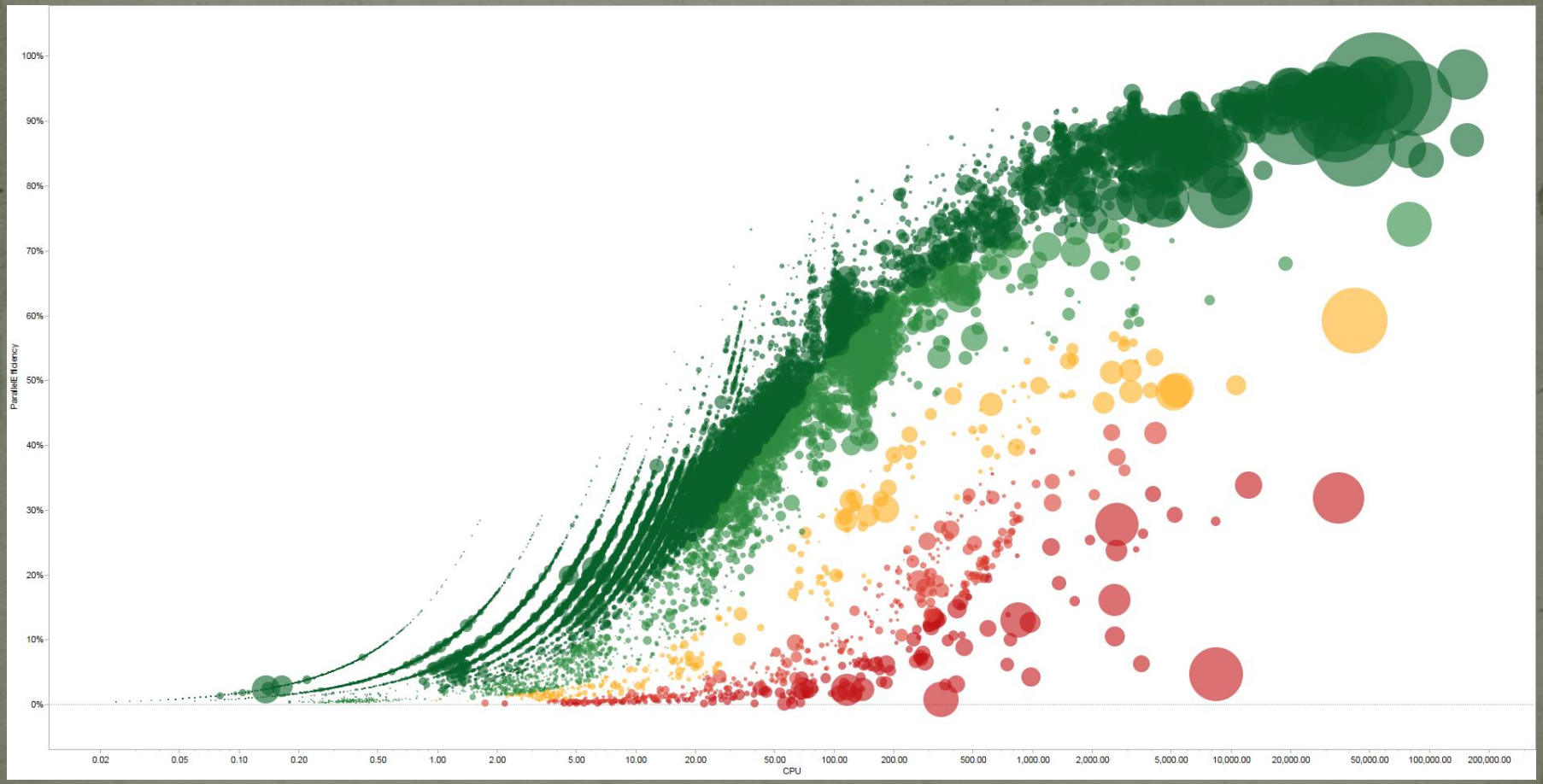
- Interesting opportunities but …
  - If already overprovisioned on CPU, who cares?
  - If Intel pulls the functionality into the next chip, obsolete in one or two generations
  - If no one will put them in because they are too expensive times 10,000 or 100,000…
  - Are we optimizing for the old days of CPU being the critical resource?
- With enough network bandwidth/scalability, are specialized co-processors a workable model?
  - How about optimization algorithms for when to ship and when to do at home?
  - But will anyone adopt co-processors in this commodity only world?

Query Optimization

- Stuck optimizing qualification, joins, counts and sums
- Where are the scalable algorithms for the questions people want to ask?
  - Look at the R, Matlab library!
  - Thousands of algorithms that need to operate on large data sets
  - None of them written to scale beyond a single CPU
  - How do we create scalable algorithms in a replicable, repeatable way?
  - And this is just the beginning
    - Text, sensor data, image, video,…
  - New kinds of problems – iterative, path, graph,… - that do not fit SQL well
- Skew
  - The giant (cluster) killer
  - Where is the work on skew?
  - Real world data

Skew – New Ways of Measuring and Thinking Required

0.1% of jobs are responsible for 53% of the PE overhead of the system



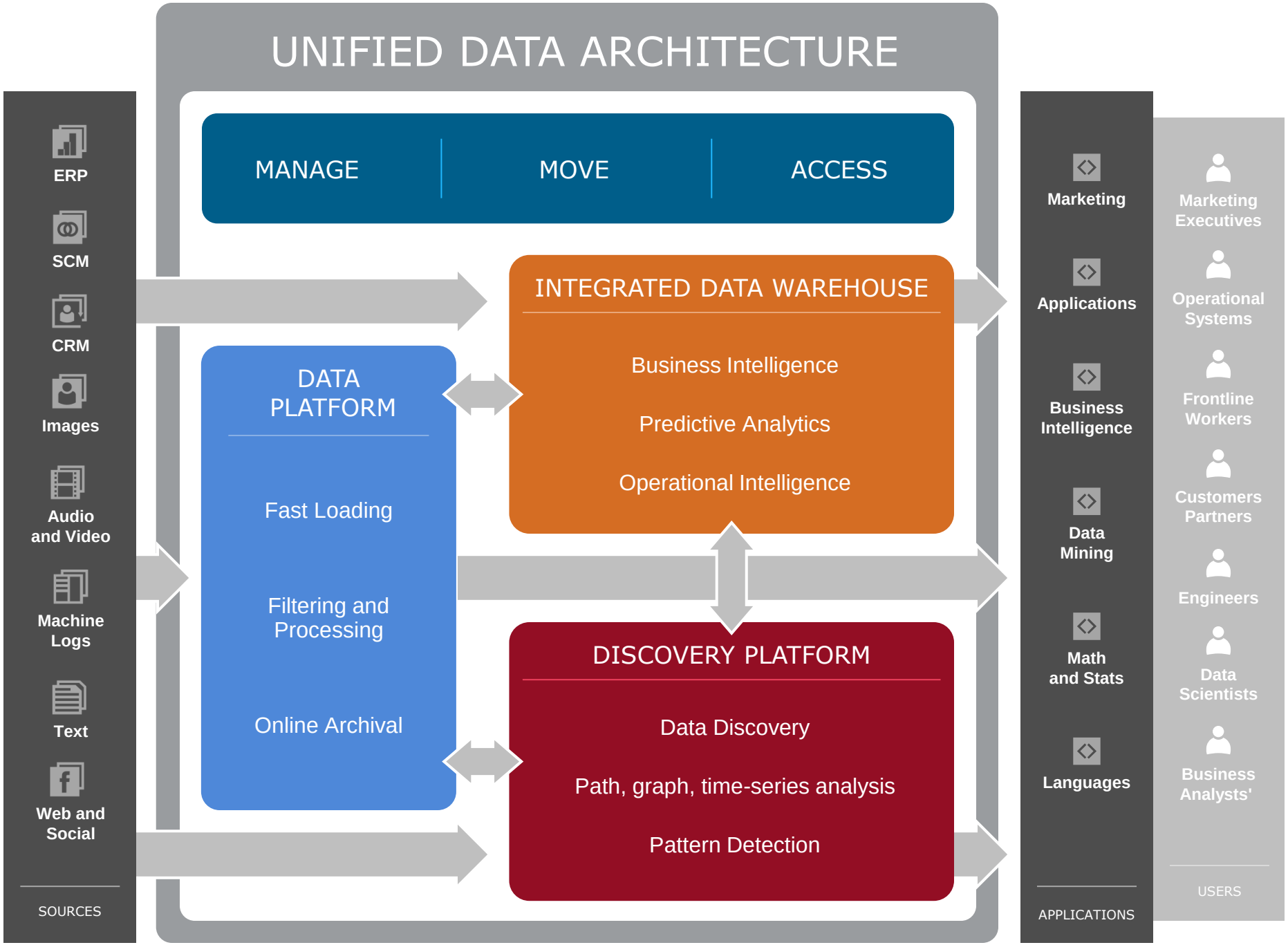
http://yottascale.com/entry/the-colorful-secrets-of-bigdata-platforms

In Memory, Cubes, …

- Cubes are walking dead
  - Already moved to hybrid a decade ago
  - In Memory finishes them off
- In Memory for analytics just scratching the surface
  - Not scaling yet
  - Will it? Network latency >>> memory access times for the forseeable future
    - Any interesting problem needs to move (a lot of) data
    - New algorithms and methods for locality of reference needed
- Replacing the cubes with front edge caching
  - More manageable, relieves cube build times
  - Still hitting scale walls and going hybrid
  - Need great synchronization, smart caching, workload pattern analysis and detection of real world workloads,…

Right Data Management Tool for the Right Job

- Currently in a place where no tool does everything
- Need to assemble many tools to get a job done
- Multiple or many data stores required even for analytic environment
- Big question
  - Will these converge?
  - If yes, what do we need to do to make that happen?
    - Combine SQL, MapReduce, algorithms, new platform capabilities,…
    - Handle radically mixed workloads
  - If no – what do we need to do to make it usable?
    - Virtualization
      - Heterogeneous processing models
      - Not just heterogeneous SQL that we still can't do
      - And where is the metadata management?


UNIFIED DATA ARCHITECTURE
MANAGE
MOVE
ACCESS
INTEGRATED DATA WAREHOUSE
Business Intelligence
Predictive Analytics
Operational Intelligence
DATA PLATFORM
Fast Loading
Filtering and Processing
Online Archival
DISCOVERY PLATFORM
Data Discovery
Path, graph, time-series analysis
Pattern Detection
ERP
SCM
CRM
Images
Audio and Video
Machine Logs
Text
Web and Social
SOURCES
Marketing
Applications
Business Intelligence
Data Mining
Math and Stats
Languages
APPLICATIONS
Marketing Executives
Operational Systems
Frontline Workers
Customers Partners
Engineers
Data Scientists
Business Analysts'
USERS

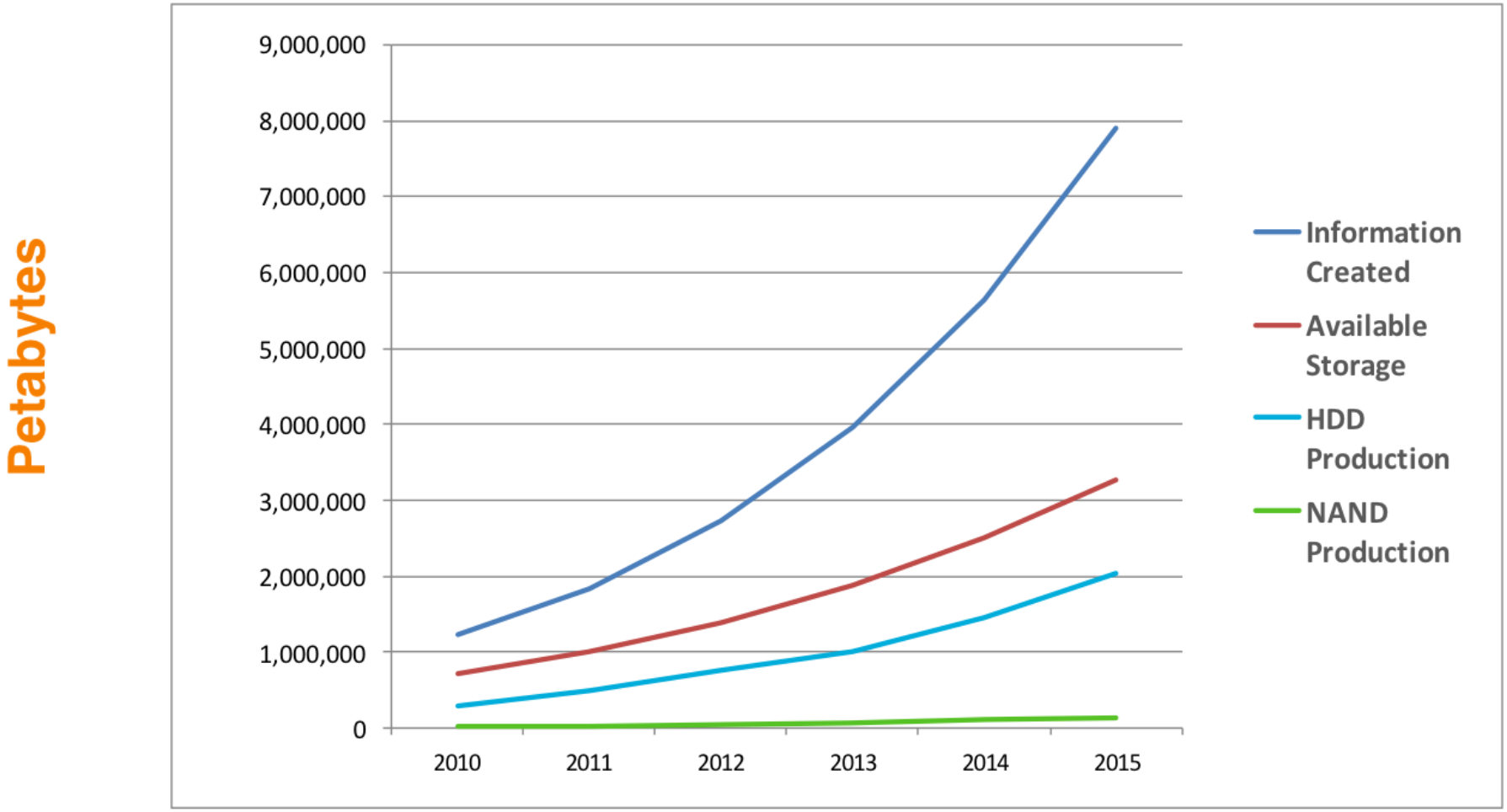
Converged Analytic Systems??

- As these environments evolve and converge, radically mixed workloads must be handled
  - Front edge sub-second lookups
  - Back edge resource hogs
  - Streamed filtering and analysis
  - On the fly analytics
  - Ad hoc analytics
- Will we build one ring to bring them all and bind them to one copy of the data?
  - Or continue on todays trajectory of many, many copies of data, many disparate systems,…??
- If convergence then:
  - Drastic improvements in workload management
  - And all the –ilities!!
- If ~convergence:
  - Then superhuman improvements in deployment, data movement, synchronization, …
  - And Orchestration, Virtualization
  - And all the -ilities

If We Had a Clean Slate?

- What would a cloud data store look like?
  - With all of today's requirements for data types, algorithm types
  - Assuming massive sharing
  - Assuming massive multi-tenancy
  - Assuming radically mixed workloads
  - And flexible enough to deal with a platform environment that is changing fundamental balances on an annual basis?
  - With all the -ilities

Data Generation Outstrips Storage Capacity



Source: IDC IVIEW, "Extracting Value from Chaos," June 2011, and
Gartner, Market Trends: "Enlarging the Library of Forms in Which Storage is a Necessary Element," June 30, 2011

And Extra Credit

- The Anti-Database
- Signal to Noise ratio really bad in much of the Big Data world
- How to throw away the RIGHT data?
  - In an intelligent way
  - Learning
  - Adjusting
  - Related, connected streams
  - Easy rules specification