

Big Data Summit

2010 Bay Area



Everything New Is Old Again

Merv Adrian, IT Market Strategy

February 19, 2010

aster data
big data. fast insights.™

Question Everything. It's All Changed. Or Has It?

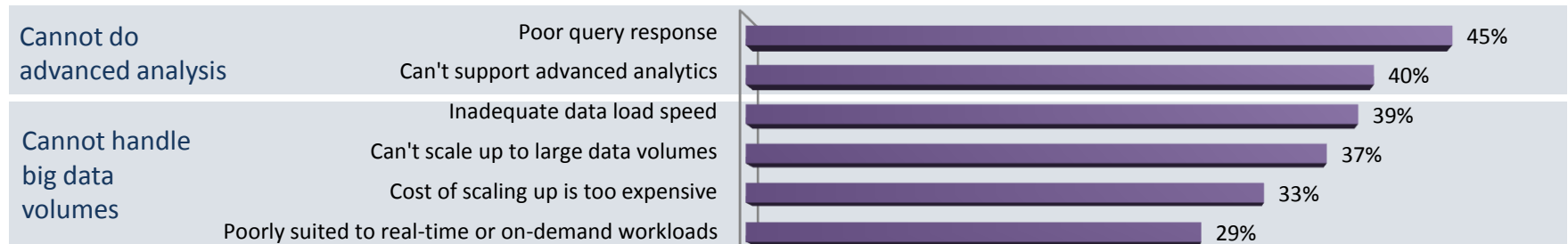
In computing's early days, we wrote programs and loaded them to run against files we loaded at the same time. Then we cleared it all out to start over again. Today, some people are doing it again. Where and why? For what they call "Big Data." Some questions:

- What Does "Big Data" mean to me?
- Do I want to keep it all? Why?
- Where should I physically store it?
- How should I "process" it? SQL or NoSQL?
- DB or not DB? That is the question.
- Where should analytics be done? A Wish List

What Does "Big Data" Mean To Me?

Data Volume, But Also Complexity of Analysis

What problems are driving you to look at new Data Warehouse/Data Management Solutions?



- Half of TDWI Survey respondents will replace their DW platform in next three years. Why?
- These are the responses above 25%

Source: TDWI 2009 Company Survey of 417 Companies

Do I Want To Keep It All? Why?

- Will it be repeatedly used? How often?
- Must I use all of it, or are samples enough?
- Are there regulatory retention considerations?
- Will it be updated?
- Do I need it in (or near) real time?

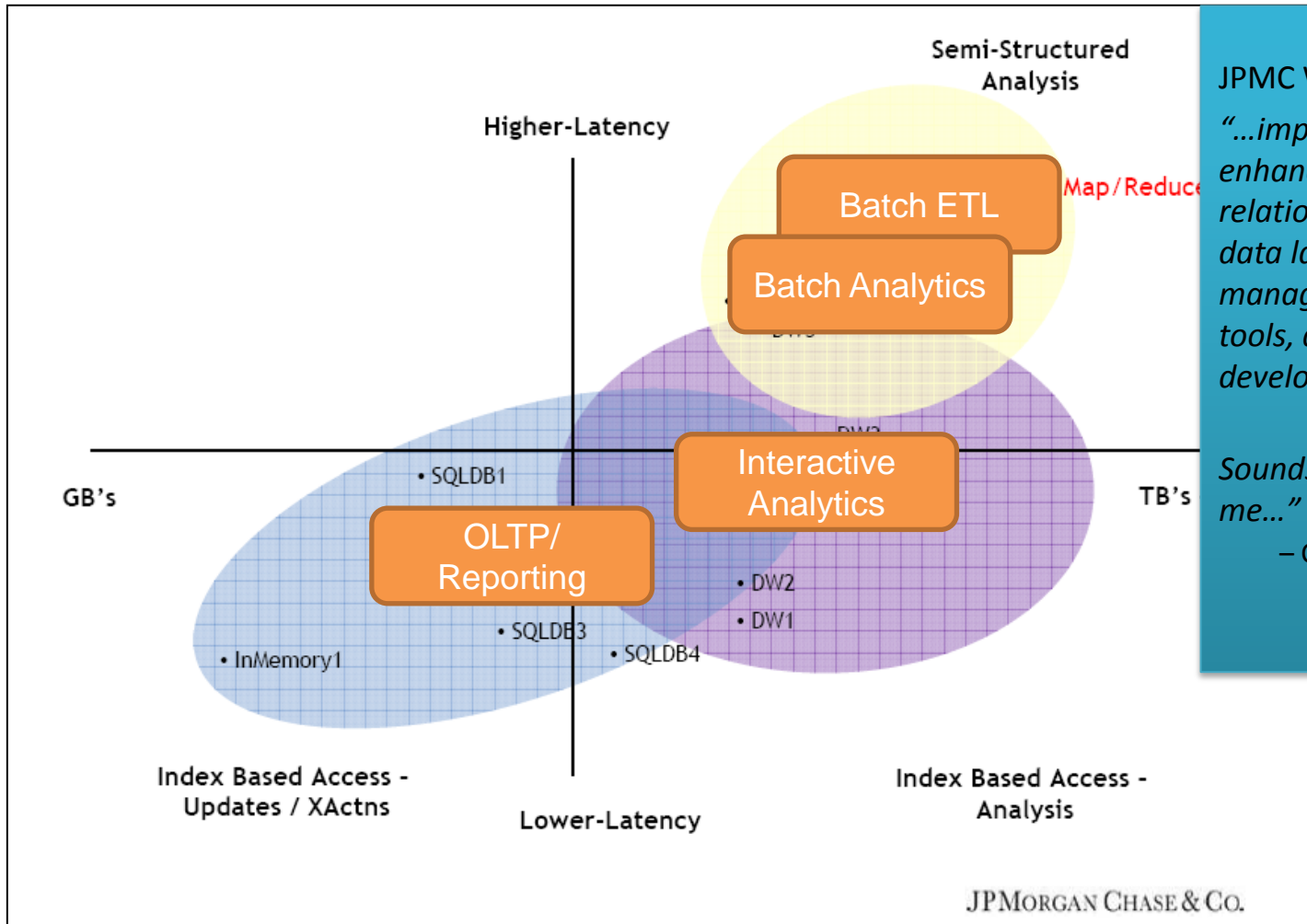
Where Should I Physically Store It?

- Considerations: storage cost curve keeps coming down. But it still costs a LOT of money
 - Acquisition costs
 - Costs of space, power
 - Skills, people, time to manage it all
- Onsite? Offsite? Primary and backup? How many copies?
- In a private cloud? A public one?
- Online? Near-line? Archive?

How Should I “Process” It? SQL or NoSQL?

- Much processing can be done outside DBMSs
 - Debate rages over whether SQL is always the “right” language
 - It’s almost moot if your team’s skills are in other languages
- Hence, NoSQL
 - a movement that promotes the use of MapReduce, often atop Hadoop (more on that shortly), for “big data” programs
- Note that DBMSs can increasingly support other kinds of code execution (more on that later)

Processing Scenarios: J.P. Morgan Chase Explores MapReduce



JPMC Wish List:

"...improved SQL interfaces, enhanced security, support for a relational container, reduced data latency, better management and monitoring tools, and an easier to use developer programming model."

"Sounds like a relational DBMS to me..."

– Colin White, Oct 2009

What is Hadoop?

- An alternative method for working on (usually Big) data outside of DBMSs
- A scalable fault-tolerant grid operating system with:
 - **MapReduce:** Fault-Tolerant Distributed Processing
 - **HDFS:** Self-Healing High-Bandwidth Clustered Storage
- A collection of programs and utilities
 - Available as open source (under Apache License)



DB or not DB? That Is The Question

- *Managing* data is not the same as *using* data
 - Transactional requirements impose needs
 - Analytical requirements often relax them, but add others
- Where needed, DBMSs afford substantial benefits:
 - Enforcing data quality, security, privacy
 - Encryption, compression
 - Backup, recovery
 - Programming – and parallelization – of complex analytics

Hadoop and RDBMS: A High-Level View

Hadoop	RDBMS
Designed for very large hardware clusters (leading to possibly higher power consumption)	Designed for multipurpose hardware environments
High data availability (thru S/W)	High data availability (thru H/W & S/W) High data integrity High data compression (in some cases)
Includes MapReduce	Some products support MapReduce and/or Hadoop system connectors
Focused towards in-house developed applications (Java, Python, Perl, etc.)	Provide many SQL-based vendor tools and packaged solutions
Programmatic data access to underlying file system	Sophisticated SQL query optimizer determines access to file system
Limited understanding of data and data relationships, e.g., dynamic data	Well understood data and data relationships
Suited to large batch jobs and data exploration of very large data files	Suited to batch and online analysis of complex data systems

Where Should Analytics be Done? A Wish List

- **Close to the data** – minimize movement latency, complexities
 - DBMS primitives, UDFs optimize processing
- In a **parallel processing** environment
 - Parallelization is challenging, although products exist
- **In memory** to minimize I/O
- Where tools exist
 - Ideally: familiar, manageable ones
 - Build. AND test. AND debug. AND deploy.
 - Manage resources

Summary

- It's not at all clear that "no database" makes sense except in specific cases
- The NoSQL movement is adding – surprise – DBMS-like features as fast as it can
- DBMSs continue to improve support for in-memory, (parallel), MPP, and in-database analytics
- What are the decision variables? As always:
 - Skills
 - Money
 - Time
 - Fit

Big Data Summit

2010 Bay Area



Thank You!

Merv Adrian, IT Market Strategy
www.itmarketstrategy.com

February 19, 2010

aster data
big data. fast insights.™