VoltDB the open source database you'll never outgrow

VoltDB Streamlining Hadoop for Enterprise Adoption

August 2012

Mark Hydar

Market Technology and Strategy

Agenda

- “Big Data” and the Data Landscape
- Our Thoughts on Data Pipelines
- VoltDB Streaming Overview
- Addressing the Topics
  - Hadoop is too complex and expensive for mainstream enterprises.
  - It’s taking too long to find useful insights amid an ocean of low quality, disconnected data.
  - How can my organization reduce costs and mitigate data risks?
  - How can I gain quicker access to operational insights?
  - What can I do to improve data quality and reduce total pipeline processing times?
- Q&A

What is “Big Data”?

Velocity = VoltDB
Big Data = VoltDB + Volume + Variety

- The old equation of Big Data

  big data = volume = warehouse (OLAP)

- This has changed

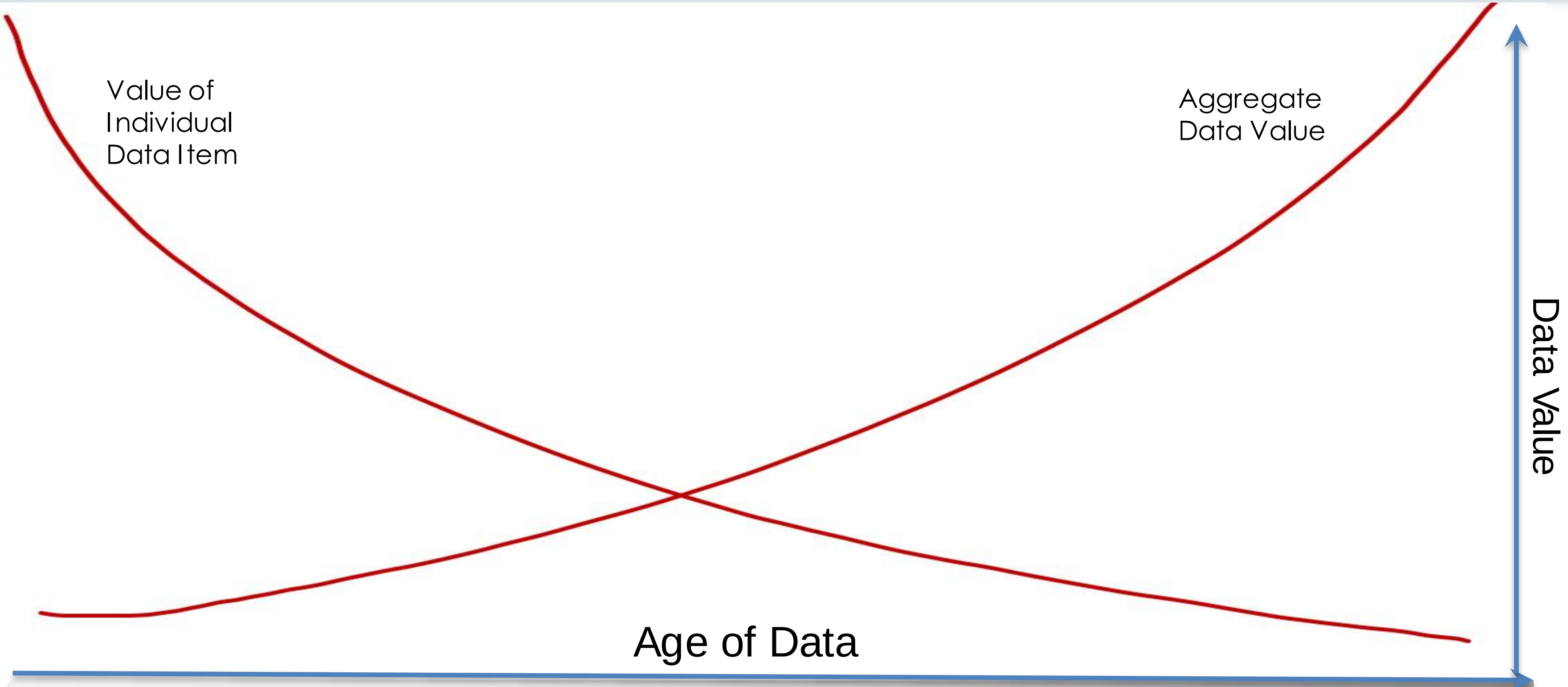
  big data = velocity + volume = transactions (OLTP) + warehouse (OLAP)

- Big Data is fast and deep

- As it arrives, you probably do something with it (or wish you could)

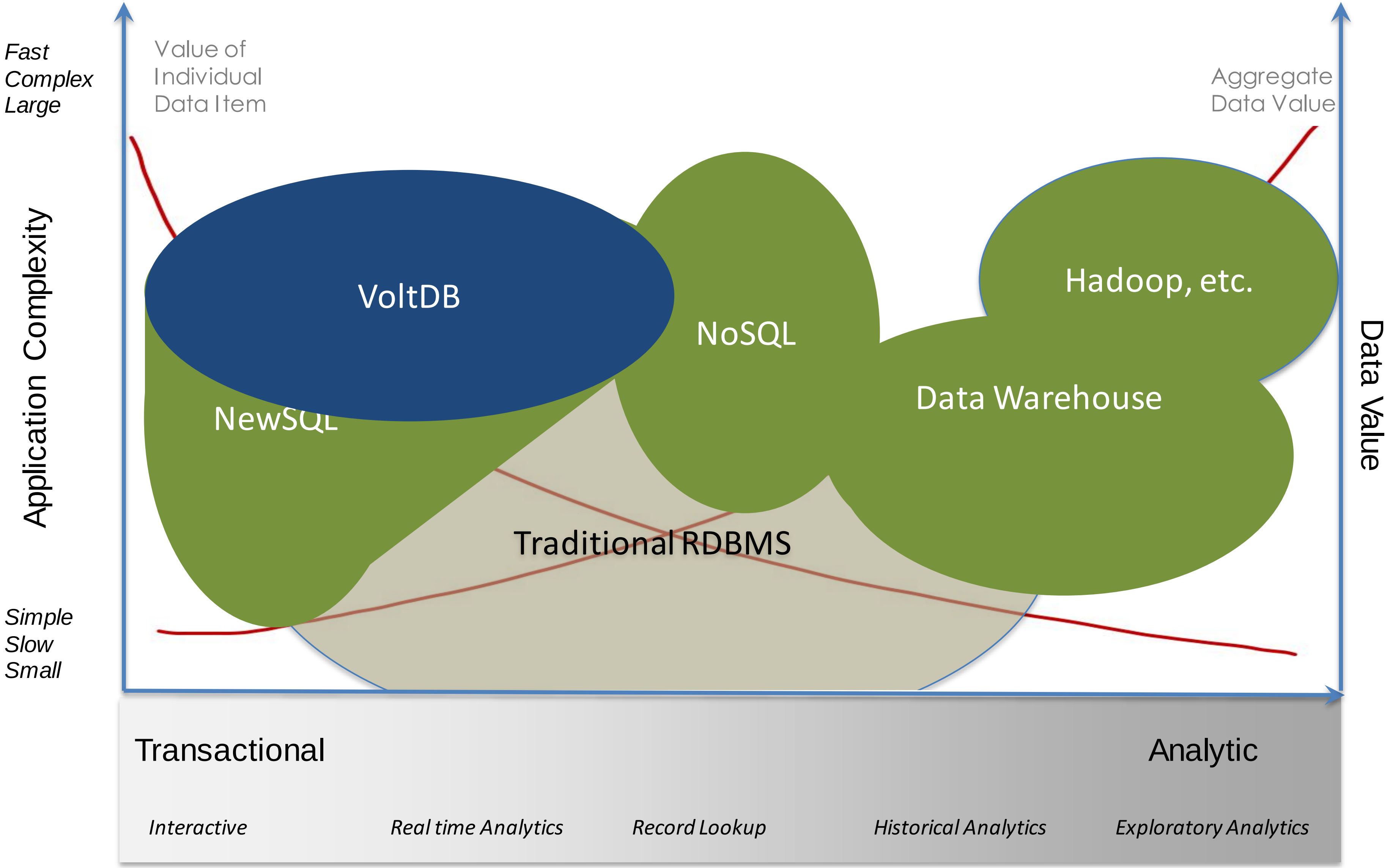
  you may just want to slow it down!

Data Value Chain

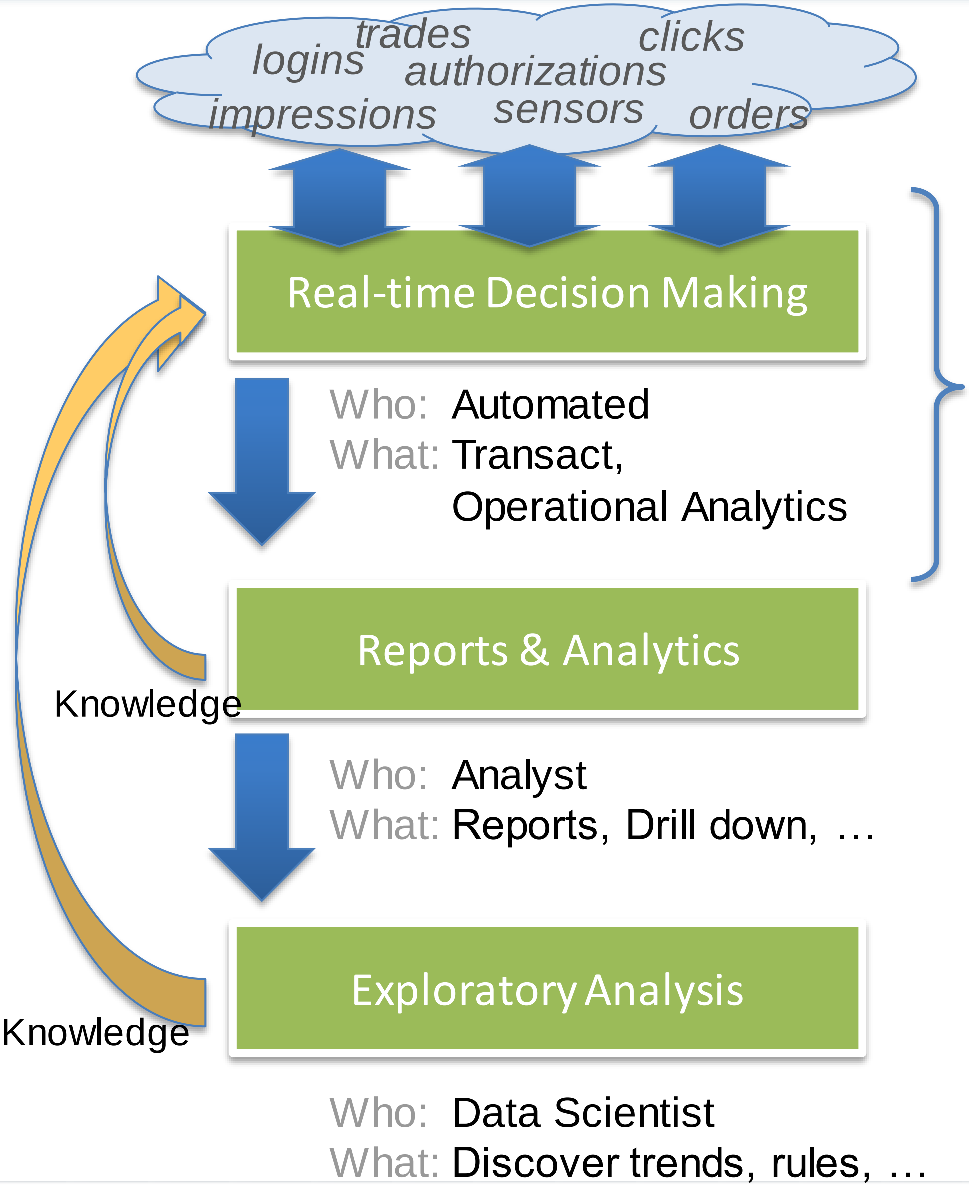


Interactive	Real time Analytics	Record Lookup	Historical Analytics	Exploratory Analytics
Milliseconds	*Hundredths of seconds*	*Second(s)*	*Minutes*	*Hours*
. Place trade . Serve ad . Enrich stream . Examine packet . Approve trans.	. Calculate risk . Leaderboard . Aggregate . Count	. Retrieve click stream . Show orders	. Backtest algo . Business Intelligence . Daily reports	. Algo discovery . Log analysis . Fraud pattern match

Database Landscape



Lifecycle for Big Data



trades
logins
clicks
authorizations
impressions
sensors
orders

Real-time Decision Making

Who: Automated
What: Transact,
Operational Analytics

Reports & Analytics

Knowledge

Who: Analyst
What: Reports, Drill down, …

Exploratory Analysis

Knowledge

Who: Data Scientist
What: Discover trends, rules, …

- Make the most informed decision every time there is an interaction
- Real-time decisions are informed by operational analytics & past knowledge
- Sometimes called OLTP

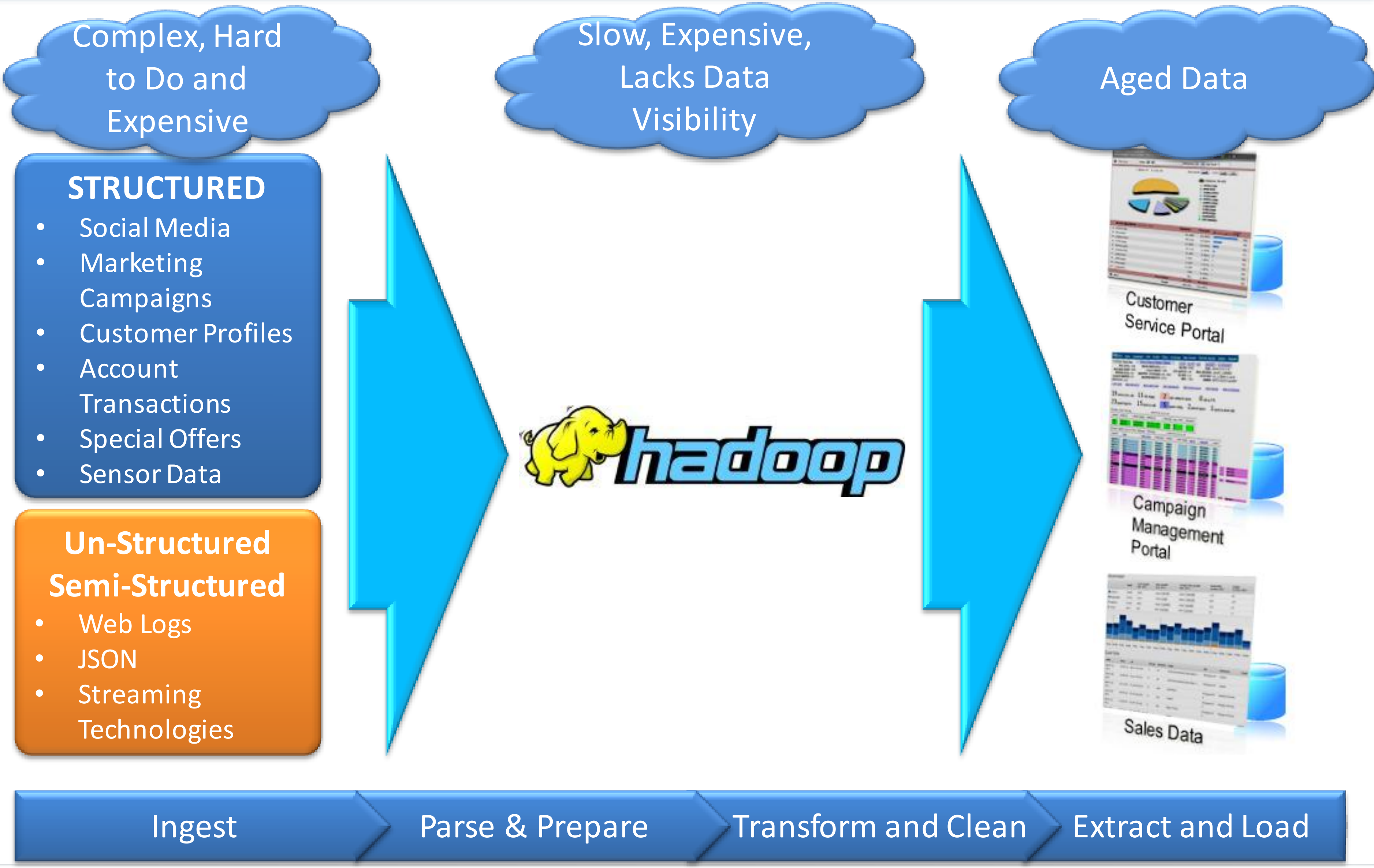
"It is not enough to capture massive amounts of data; organizations must also sift through the data, extract information and transform it into ***actionable knowledge****."*

The Value Available in Fast Data

- Data-driven decisions in real-time
- Better decisions by using more information sources
- Faster decisions
- Insights into real-time operational analytics

If I could "**Access**" to my data sooner, I would be **more Insightful** than the next guy. *-What your competition is thinking right now*

The Typical Hadoop Data Pipeline



Complex, Hard to Do and Expensive

Slow, Expensive, Lacks Data Visibility

Aged Data

STRUCTURED

- Social Media
- Marketing Campaigns
- Customer Profiles
- Account Transactions
- Special Offers
- Sensor Data

Un-Structured Semi-Structured

- Web Logs
- JSON
- Streaming Technologies

Ingest → Parse & Prepare → Transform and Clean → Extract and Load

The VoltDB Database

- High-performance RDBMS
- In-memory database
- Automatic scale-out on commodity servers
- Built-in high availability
- Relational structures, ACID and SQL

VoltDB Performance Advantage	
TPC-C single_node (Oracle)	**45 X**
TPC-C single node (MySQL)	**100 X**

Cost Disruption		
	Ex. Traditional RDBMS	VoltDB
System	SPARC SuperCluster/Oracle 11g	18 , 8-core Intel servers
Price/tpmC	$ 1.01	**$0.012**

VoltDB *is Faster, Better, Cheaper than the competition*

How VoltDB is Used

- High throughput, relentless data feeds
- Fast operations on high value data
- Real-time analytics present immediate visibility

	Data Feed	Real-time	Real-time
Network Traffic Monitoring	Network packets	Examine packet by source / destination	Identify bandwidth outliers
Financial Trade Support	Market orders	Ingest trade data	Recall post trade order groupings
Sensor tracking & analytics	Sensor position feed	Identification and cleansing of tag info	Notification and groupings
Mobile Gaming	Online game	Game state updates and usage patterns	Leaderboard lookups
Digital Ad Tech	Ad bid / click stream	Bid, optimize content	Report ad performance

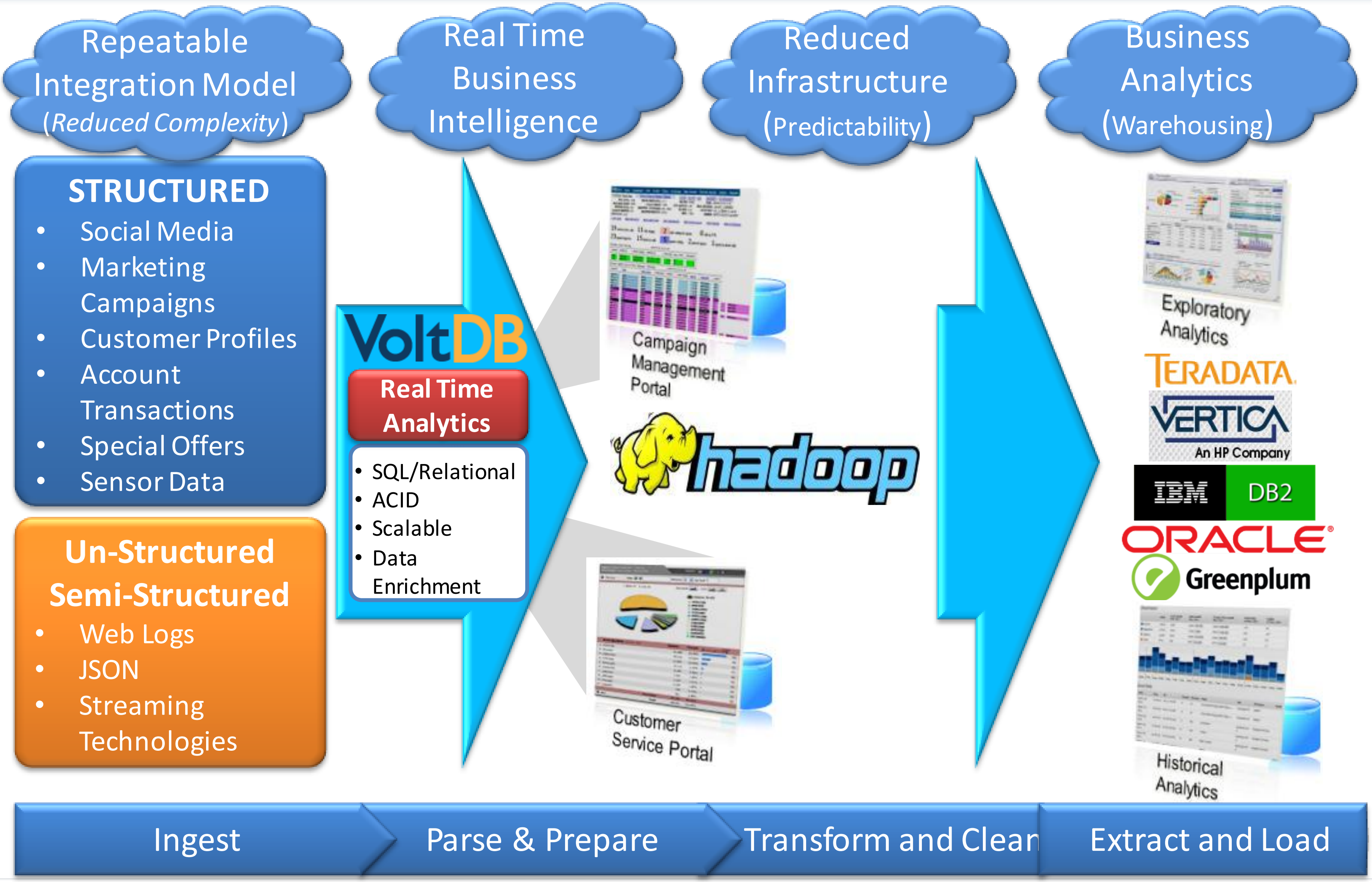
Why Address the Streaming Gap

- VoltDB Hadoop Data Streaming
  - + Real Time Business Decisioning
  - + Data Quality and Enrichment
  - + Simplifies Data Integration
  - + Shortens Time to Market
- Increasing Productivity
  - + Common Development and Data Environment
    - — Provides Reusability (data flows and computations)
    - — Provides Universal Access to Real Time Data
    - — Uses well known data access utilities

Hadoop Integration

- Motivation
  - + Big Data = high velocity (VoltDB) + high volume (Hadoop)
    - — VoltDB ingests fire hose, manages state, supports real-time analytics, spools to Hadoop
    - — Hadoop imports from VoltDB (via Sqoop)
- Technologies
  - + VoltDB Export
    - — Real-time streaming export
    - — Data consolidation, aggregation, enrichment
    - — Buffering and overflow to eliminate “impedance mismatches”
    - — Bi-directional durability
  - + Sqoop
    - — Cloudera-authored DBMS=>HDFS importer
    - — Pull-based technology

The Optimized Data Pipeline



Repeatable Integration Model (*Reduced Complexity*)

Real Time Business Intelligence

Reduced Infrastructure (Predictability)

Business Analytics (Warehousing)

STRUCTURED

- Social Media
- Marketing Campaigns
- Customer Profiles
- Account Transactions
- Special Offers
- Sensor Data

Un-Structured Semi-Structured

- Web Logs
- JSON
- Streaming Technologies

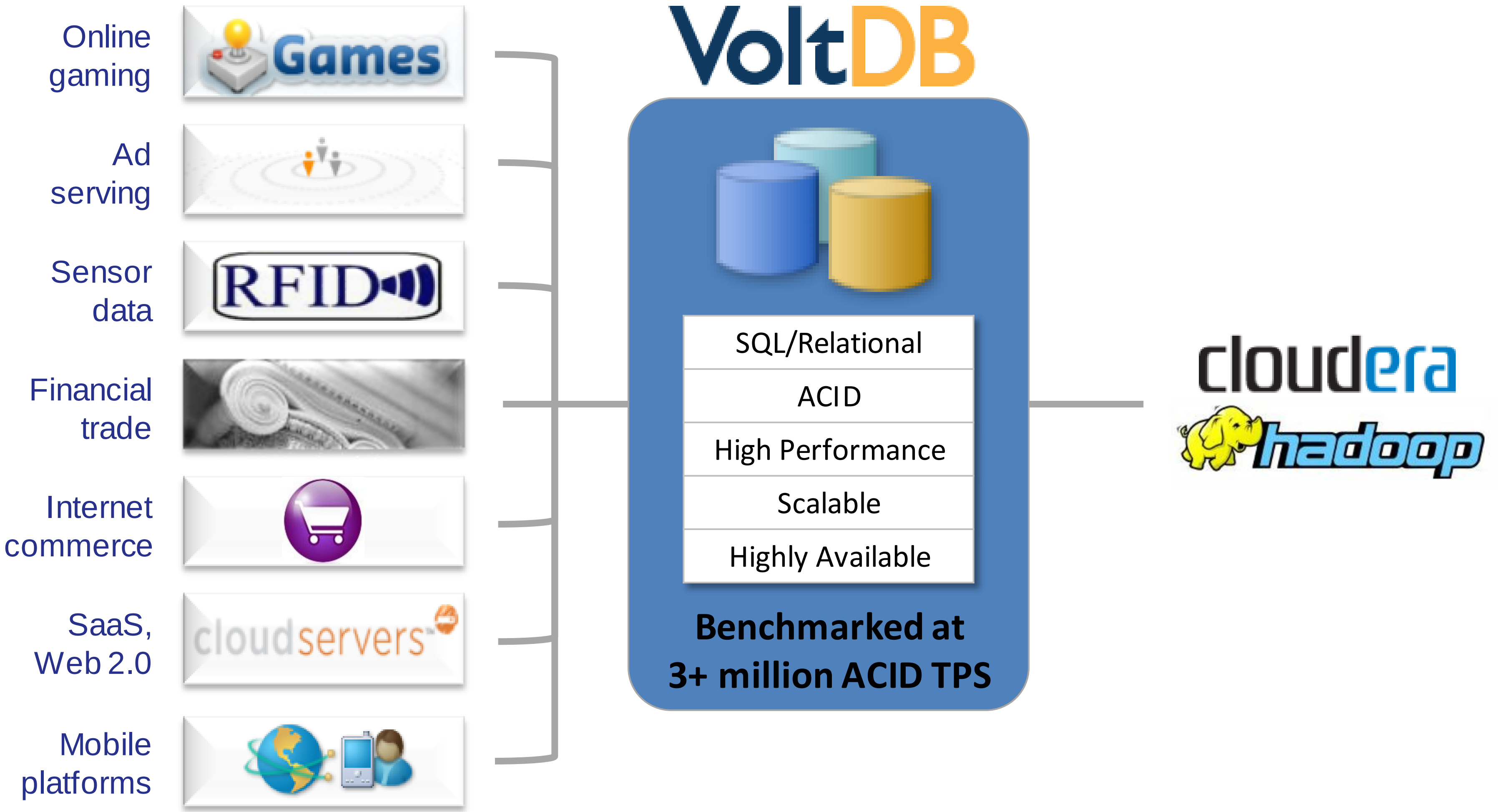
VoltDB

Real Time Analytics

- SQL/Relational
- ACID
- Scalable
- Data Enrichment

Ingest → Parse & Prepare → Transform and Clean → Extract and Load

VoltDB in the Big Data Landscape (Today)



Online gaming

Ad serving

Sensor data

Financial trade

Internet commerce

SaaS, Web 2.0

Mobile platforms

SQL/Relational
ACID
High Performance
Scalable
Highly Available

Benchmarked at 3+ million ACID TPS

The Close

- As Hadoop adoption increases, it has become evident that programming Hadoop is too complex and expensive for mainstream enterprises.
- It’s taking too long to find useful insights amid an ocean of low quality, disconnected data. How do I address the key barriers to Hadoop adoption?
- How can organizations reduce costs and mitigate data risks?
- How can they gain quicker access to operational insights?
- What can I do to improve data quality and reduce total pipeline processing times?
- Did we explore strategies that leading organizations are using to streamline Hadoop processing and eliminate adoption challenges?

Questions?

email mhydar@voltdb.com
twitter @mhydar

Download the VoltDB Enterprise Edition Trial
http://voltdb.com/products-services/downloads

Join the VoltDB Community
http://community.voltdb.com

More information on VoltDB Blog
http://voltdb.com/company/blog

Follow @VoltDB on twitter