
451 Research
ANALYZING THE BUSINESS
OF ENTERPRISE IT INNOVATION
Photo credit: swisscan on Flickr http://bit.ly/HCZCRQ
NoSQL Road Show, Zurich
NoSQL, NewSQL, Big Data… Total Data
The Future of Enterprise Data Management
Matthew Aslett
Research Manager, Data Management and Analytics
WWW.451RESEARCH.COM
NEW YORK · BOSTON · WASHINGTON DC · SAN FRANCISCO · SEATTLE · DENVER · LONDON · SAO PAULO · DUBAI · SINGAPORE

Overview

NoSQL and NewSQL databases

- Adoption and development drivers

Big data and Total Data

- Definition and implications

The 451 Group

451 Research

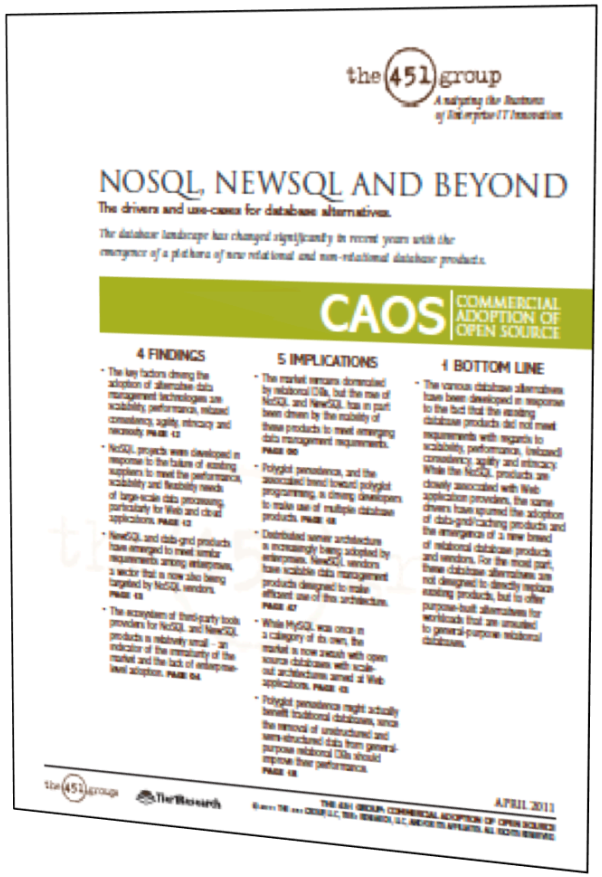
- Matthew Aslett
  - Research manager, data management and analytics
  - With The 451 Group since 2007
  - www.twitter.com/maslett

Information Management
- Operational databases
- Data warehousing
- Data caching
- Event processing

Commercial Adoption of Open Source (CAOS)
- Open source projects
- Adoption of open source software
- Vendor strategies

Relevant reports

- NoSQL, NewSQL and Beyond
  - Assessing the drivers behind the development and adoption of NoSQL and NewSQL databases, as well as data grid/ caching technologies
  - Released April 2011
  - Role of open source in driving innovation
  - sales@the451group.com


the 451 group
Analyzing the Business of Enterprise IT Innovation

NOSQL, NEWSQL AND BEYOND

The drivers and use-cases for database alternatives

The database landscape has changed significantly in recent years with the emergence of a plethora of new relational and non-relational database products.

CAOS COMMERCIAL ADOPTION OF OPEN SOURCE

4 FINDINGS

5 IMPLICATIONS

6 BOTTOM LINE

APRIL 2011

NoSQL, NewSQL and Beyond

NoSQL

- New breed of non-relational database products
- Rejection of fixed table schema and join operations
- Designed to meet scalability requirements of distributed architectures
- And/or schema-less data management requirements

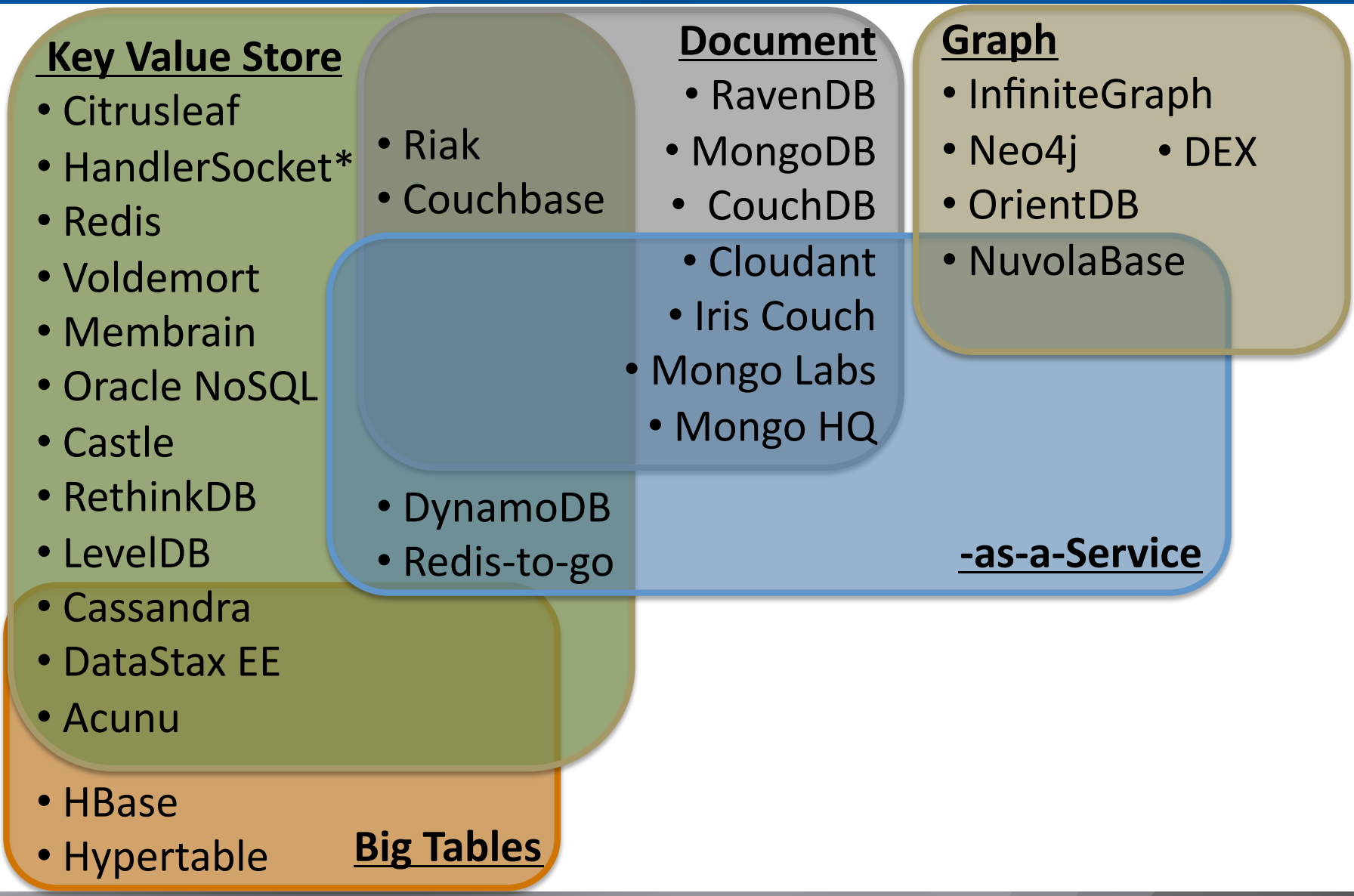
NewSQL

- New breed of relational database products
- Retain SQL and ACID
- Designed to meet scalability requirements of distributed architectures
- Or improve performance so horizontal scalability is no longer a necessity

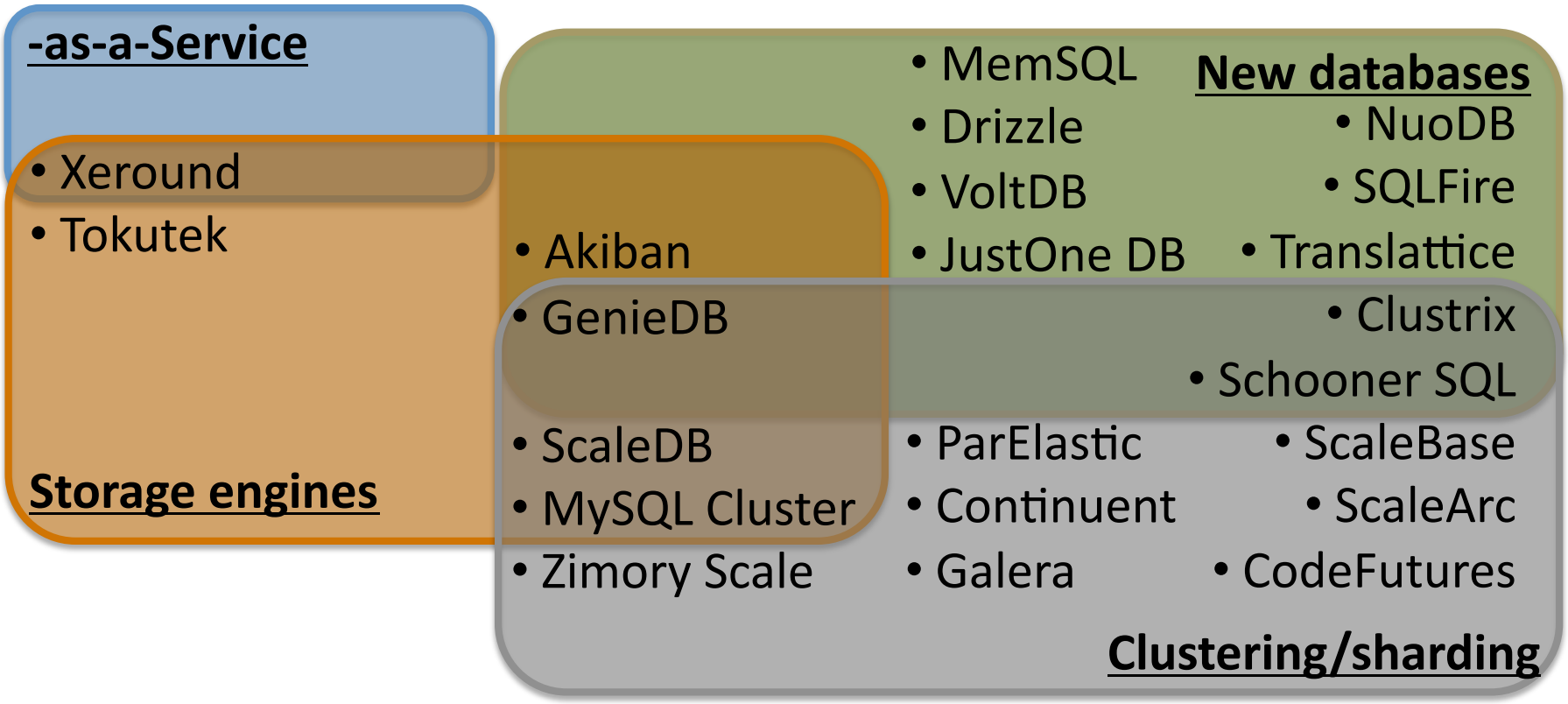
… and Beyond

- In-memory data grid/cache products
- Potential primary platform for distributed data management

The NoSQL landscape



The NewSQL ecosystem



SPRAINED RELATIONAL DATABASES

Photo credit: Foxtongue on Flickr
http://www.flickr.com/photos/foxtongue/4844016087/

SPRAIN

Necessity

- The failure of existing suppliers to address emerging requirements
- "We couldn't bet the company on other companies building the answer for us."

  Werner Vogels, Amazon CTO
- The motivation for creating Dynamo was enabling choice and not forcing the relational database to do something it was not designed to do.

Necessity - open source

- Example projects:
- BigTable: Google
- Dynamo: Amazon
- Cassandra: Facebook
- HBase: Powerset
- Voldemort: LinkedIn
- Hypertable: Zvents
- Neo4j: Windh Technologies

Scalability - Hardware economics

- Example project/service/vendor:
  - BigTable, HBase, Riak, MongoDB, Couchbase, Hadoop
  - Xeround, NuoDB
  - Data grid/cache

- Associated use case:
  - Large-scale distributed data storage
  - Analysis of continuously updated data
  - Multi-tenant PaaS data layer

Performance - MySQL limitations

- Example project/service/vendor:
- Hypertable, Couchbase, Riak, Membrain, MongoDB, Redis
- Data grid/cache
- VoltDB, Clustrix

- Associated use case:
- Real time data processing of mixed read/write workloads
- Data caching
- Large-scale data ingestion

SPRAIN

Relaxed consistency - CAP Theorem

- Example project/service/vendor:
- Dynamo, Voldemort, Cassandra, Riak
- Amazon DynamoDB

- Associated use case:
- Multi-data center replication
- Service availability
- Non-transactional data off-load

SPRAIN

Agility - polyglot persistence, schema-less

- Example project/service/vendor:
- MongoDB, CouchDB, Cassandra, Riak
- Google App Engine, SimpleDB,

- Associated use case:
- Mobile/remote device synchronization
- Agile development
- Data caching

SPRAIN

Intricacy - big data, total data

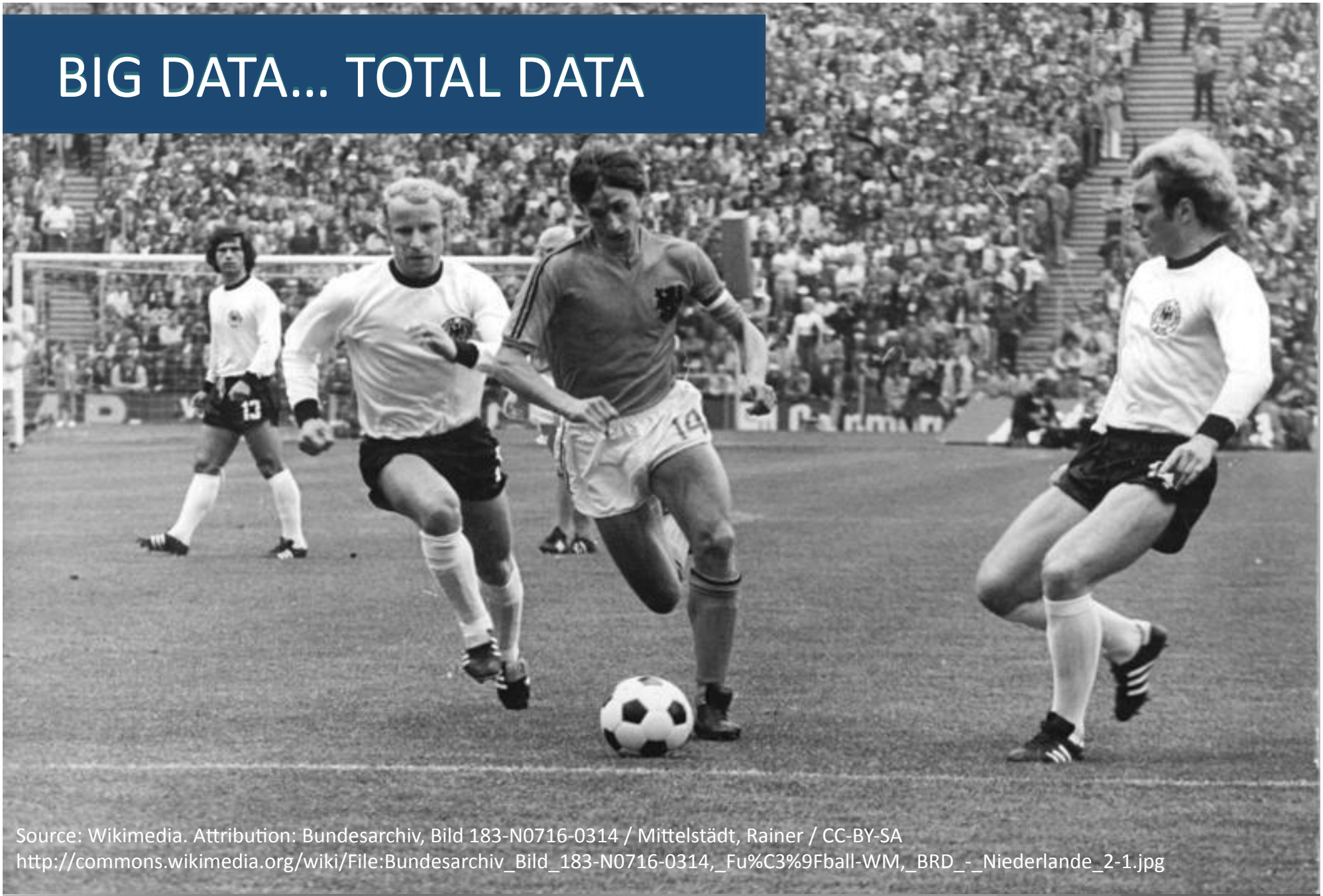
- Example project/service/vendor:
- Neo4j, GraphDB, InfiniteGraph
- Apache Cassandra, Hadoop, Riak
- VoltDB, Clustrix

- Associated use case:
- Social networking applications
- Geo-locational applications
- Configuration management database

Relevant reports

- MySQL, NoSQL and NewSQL
  - Assessing the competitive dynamic between the MySQL ecosystem, NoSQL and NewSQL technologies
  - Due May 2012
  - Including market sizing of the three database segments
  - Survey of 200+ database users
  - sales@the451group.com

BIG DATA… TOTAL DATA



Source: Wikimedia. Attribution: Bundesarchiv, Bild 183-N0716-0314 / Mittelstädt, Rainer / CC-BY-SA
http://commons.wikimedia.org/wiki/File:Bundesarchiv_Bild_183-N0716-0314,_Fu%C3%9Fball-WM,_BRD_-_Niederlande_2-1.jpg

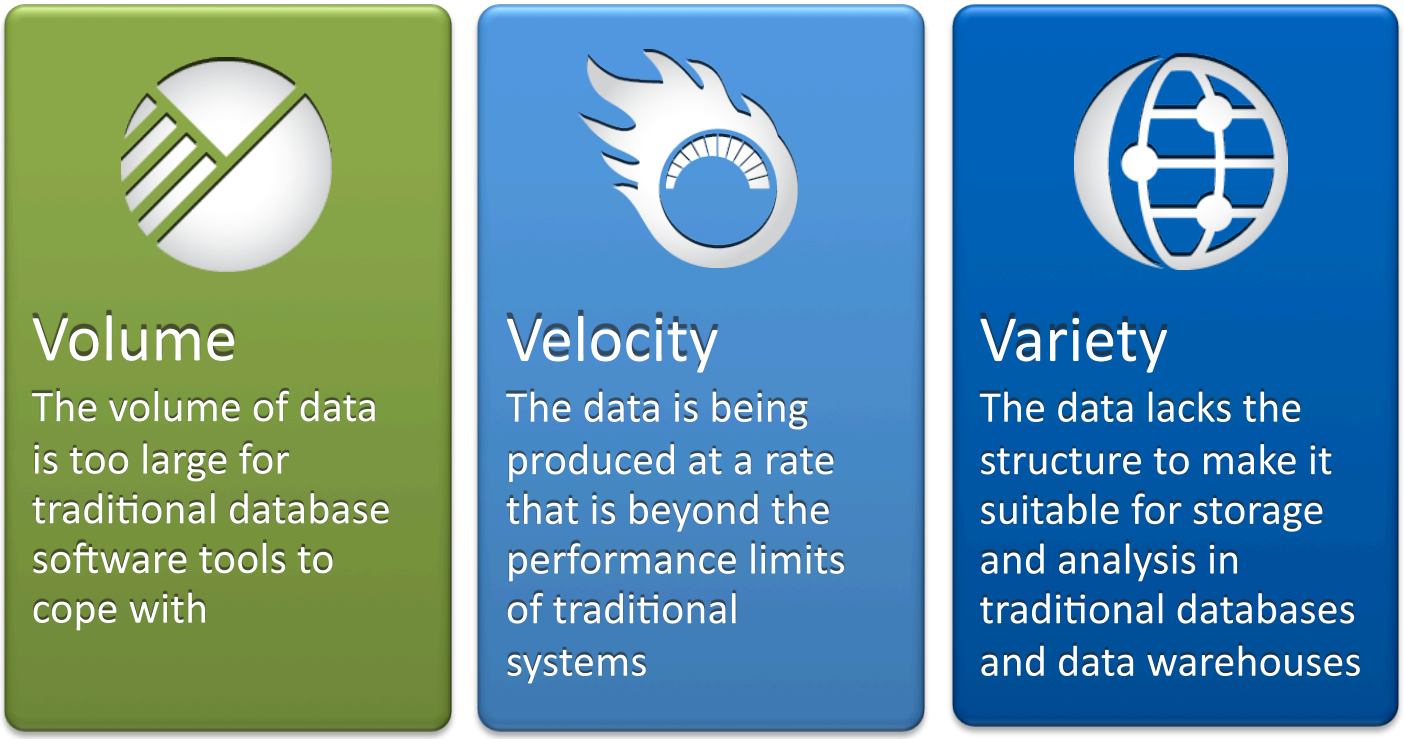
‘Big Data’

- The cost of storage, processing and bandwidth has dropped enormously, while network access has increased significantly.
- It is now more economically feasible to store and process many data sets that were previously ignored using clusters of commodity servers and advanced data processing software.

- The increased use of interactive applications and websites – as well as sensors, meters and other data-generating machines – has increased the amount and variety of data to store and process.
- “Big data” is a phrase that describes the realization of greater business intelligence by storing, processing and analyzing the increased volume, velocity and variety of data.

‘Big Data’

- The cost of storage, processing and bandwidth has dropped enormously, while network access has increased significantly.
- It is now more economically feasible to store and process many data sets that were previously ignored using clusters of commodity servers and advanced data processing software.



Beyond ‘big data’

- To realize value from data you need to look beyond the data itself:
- Generating value from data is about more than just the volume, variety, and velocity of data.

- “Total Data”

- Not just another name for Big Data
- Inspired by ‘Total Football’ – a new approach to soccer that emerged in the late 1960s
- If your data is big, the way you manage it should be total
- The adoption of non-traditional data processing technologies is driven not just by the nature of the data, but also by the user’s particular data processing requirements.

Beyond ‘big data’

- To realize value from data you need to look beyond the data itself:
- Generating value from data is about more than just the volume, variety, and velocity of data.

Totality

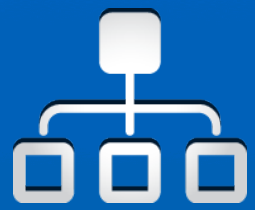
The desire to process and analyze data in its entirety, rather than analyzing a sample of data and extrapolating the results.

Exploration

The interest in exploratory analytic approaches, in which schema is defined in response to the nature of the query.

Frequency

The desire to increase the rate of analysis in order to generate more accurate and timely business intelligence.



Dependency

The reliance on existing technologies and skills, and the need to balance investment in those existing technologies and skills with the adoption of new techniques.

Totality

Totality

The desire to process and analyze data in its entirety, rather than analyzing a sample of data and extrapolating the results.

- 100 million users, more than double that amount of items for sale at any time
- Generates 50TB of new data a day
- Processes over 100PB of data a day
- The company collects everything, including historical data related to every item offered for sale and every purchase made.

Totality

Totality

The desire to process and analyze data in its entirety, rather than analyzing a sample of data and extrapolating the results.

ORBITZ

- Prior to adopting Hadoop, only had transactional and summarized non-transactional data stored in its EDW
- The vast majority of its log data was discarded as not valuable enough to be efficiently processed in an enterprise data warehouse
- Now using Hadoop to process hundreds of GBs of log data produced by the millions of searches and transactions performed on its site each day

Exploration

Exploration

The interest in exploratory analytic approaches, in which schema is defined in response to the nature of the query.

ORBITZ

- The company wanted to perform analysis on customer data in order to create geo-targeted advertising
- The required data was already present in its data warehouse but was modeled in a way that would not allow Orbitz to efficiently process the query
- Extracting the data into Hadoop enabled the company to query it in a way that the data warehouse was never designed for

Relevant reports

- Total Data – published December 2011
  - Examines the trends behind 'big data'
  - Explains the new and existing technologies used to store and process and deliver value from data
  - Outlines a Total Data management approach focused on selecting the most appropriate data storage and processing technology to deliver value from big data
  - sales@the451group.com

Thank you. Questions? Comments?

matt.aslett@451research.com
@maslett

451 Research

Photo credit: swisscan on Flickr http://bit.ly/HCZCRQ