© 2011 by The 451 Group. All rights reserved.

NoSQL, NewSQL and Beyond

Open Source Driving Innovation in Distributed Data Management

Matthew Aslett, The 451 Group
matthew.aslett@the451group.com

© 2011 by The 451 Group. All rights reserved

Overview

- Open source impact in the database market
- NoSQL and NewSQL databases
  - Adoption/development drivers
  - Use cases
- And beyond
  - Data grid/cache
  - Cloud databases
  - Total data

© 2011 by The 451 Group. All rights reserved

Building Your Big Data Future with Open Source

451 Research is focused on the business of enterprise IT innovation. The company's analysts provide critical and timely insight into the competitive dynamics of innovation in emerging technology segments.

Tier1 Research is a single-source research and advisory firm covering the multi-tenant datacenter, hosting, IT and cloud-computing sectors, blending the best of industry and financial research.

The Uptime Institute is 'The Global Data Center Authority' and a pioneer in the creation and facilitation of end-user knowledge communities to improve reliability and uninterruptible availability in datacenter facilities.

TheInfoPro is a leading IT advisory and research firm that provides real-world perspectives on the customer and market dynamics of the enterprise information technology landscape, harnessing the collective knowledge and insight of leading IT organizations worldwide.

ChangeWave Research is a research firm that identifies and quantifies 'change' in consumer spending behavior, corporate purchasing, and industry, company and technology trends.

© 2011 by The 451 Group. All rights reserved

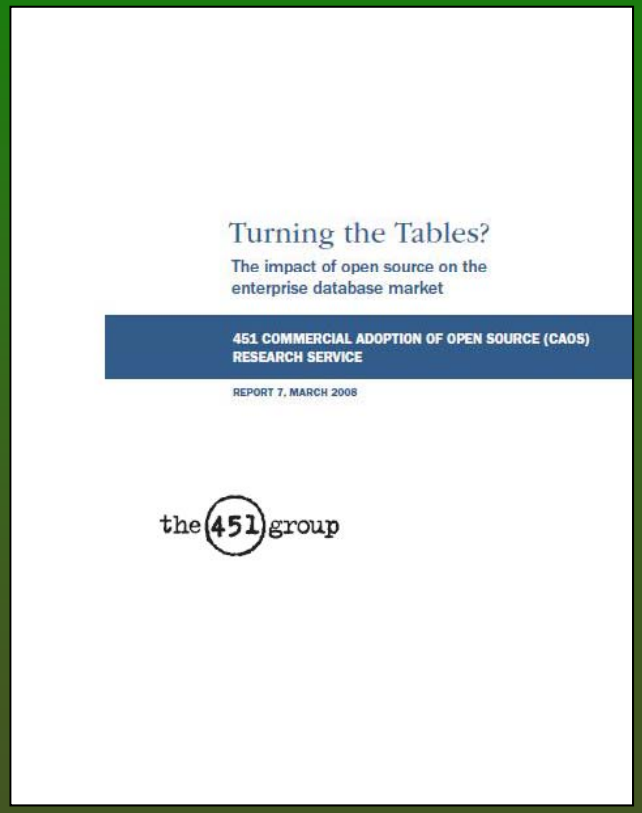
Coverage areas

- Senior analyst, enterprise software
- Commercial Adoption of Open Source (CAOS)
  - Adoption by enterprises
  - Adoption by vendors
- Information Management
  - Databases
  - Data warehousing
  - Data caching

© 2011 by The 451 Group. All rights reserved

Relevant reports

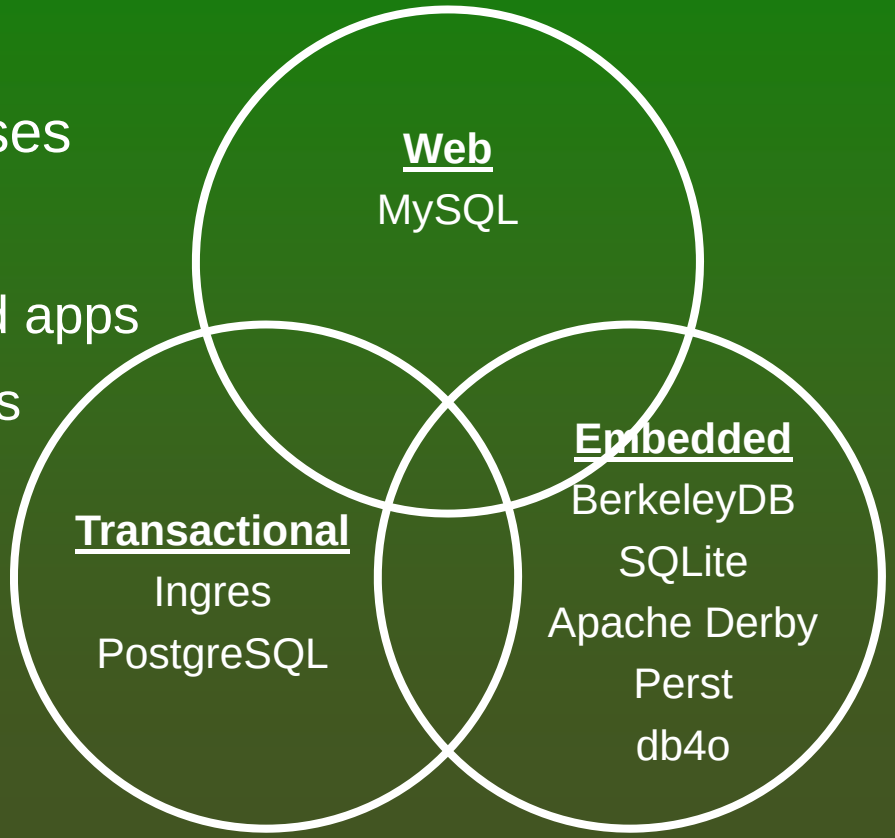
- Turning the Tables
  - The impact of open source on the enterprise database market
  - Published March 2008
  - Survey of 368 database purchasers
  - sales@the451group.com



© 2011 by The 451 Group. All rights reserved

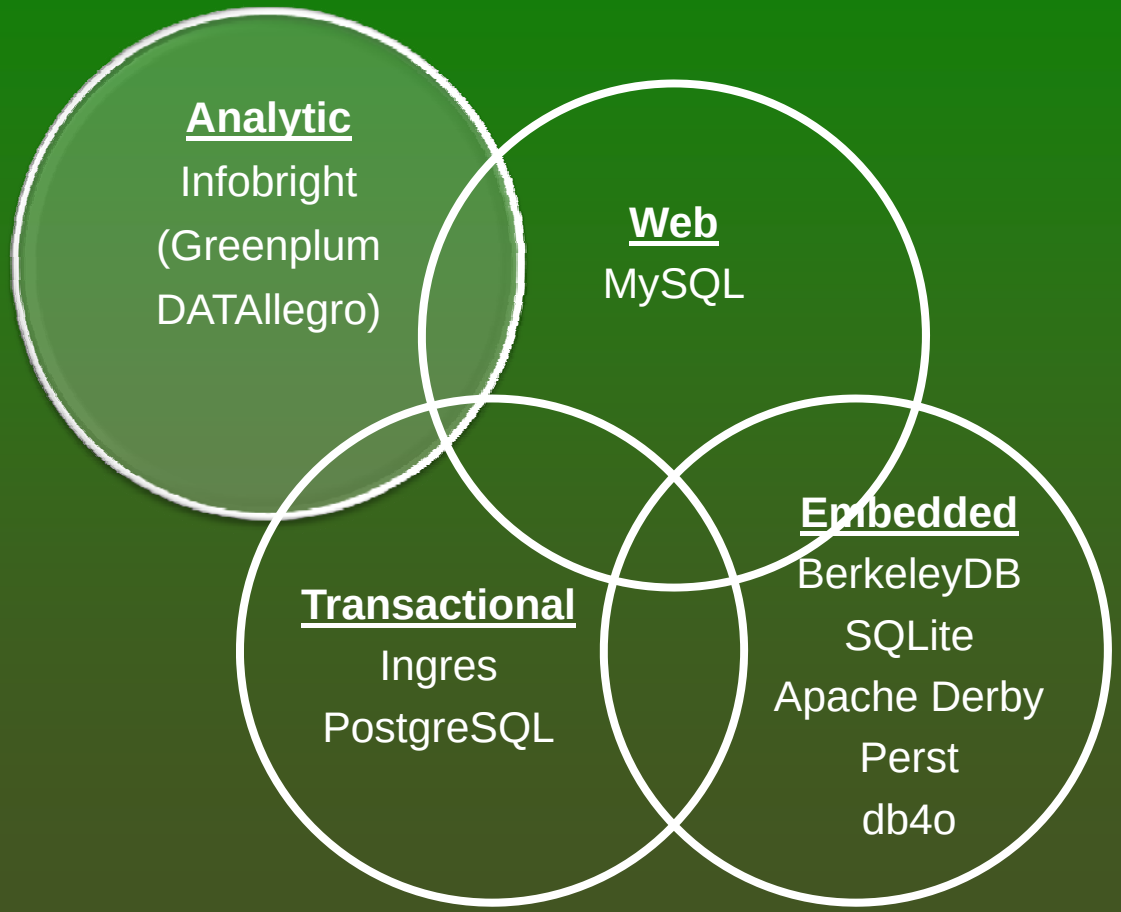
Open source database landscape 2008

- Open source adoption widespread but shallow
- 30.4% using OSS databases
- Main usage areas
  - 45% In-house-developed apps
  - 41% Single-function apps
  - 38% Development
  - 36% Web apps



© 2011 by The 451 Group. All rights reserved

Open source database landscape 2008



© 2011 by The 451 Group. All rights reserved

Relevant reports

- Warehouse Optimization
  - Ten considerations for choosing/building a data warehouse
  - Published September 2009
  - The role of open source and the emergence of Hadoop
  - sales@the451group.com


the 451 group

WAREHOUSE OPTIMIZATION

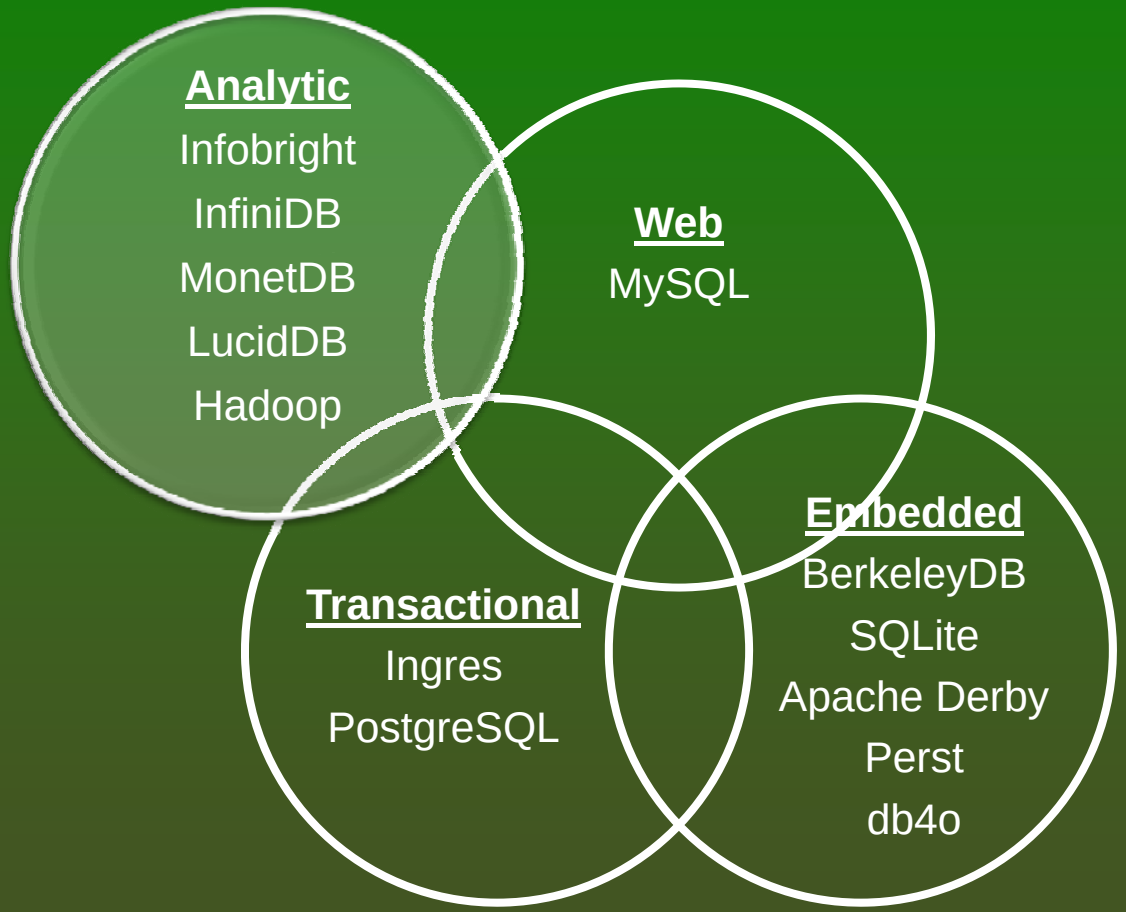
Ten considerations for choosing/building a data warehouse

IM INFORMATION MANAGEMENT

4 FINDINGS | 5 IMPLICATIONS | 1 BOTTOM LINE

© 2011 by The 451 Group. All rights reserved

Open source database landscape 2009



© 2011 by The 451 Group. All rights reserved

Relevant reports

- Data Warehousing 2009-2013
  - Market Sizing, Landscape and Future
  - Published August 2010
  - The potential impact of Hadoop
  - sales@the451group.com


the 451 group
Analyzing the Business of Enterprise IT Innovation

DATA WAREHOUSING: 2009-2013

Market Sizing, Landscape and Future

This report provides market-sizing estimates for the data-warehouse sector from 2009 to 2013. It includes revenue estimates and growth projections, and examines the business and technology trends driving this market.

IM INFORMATION MANAGEMENT

4 FINDINGS

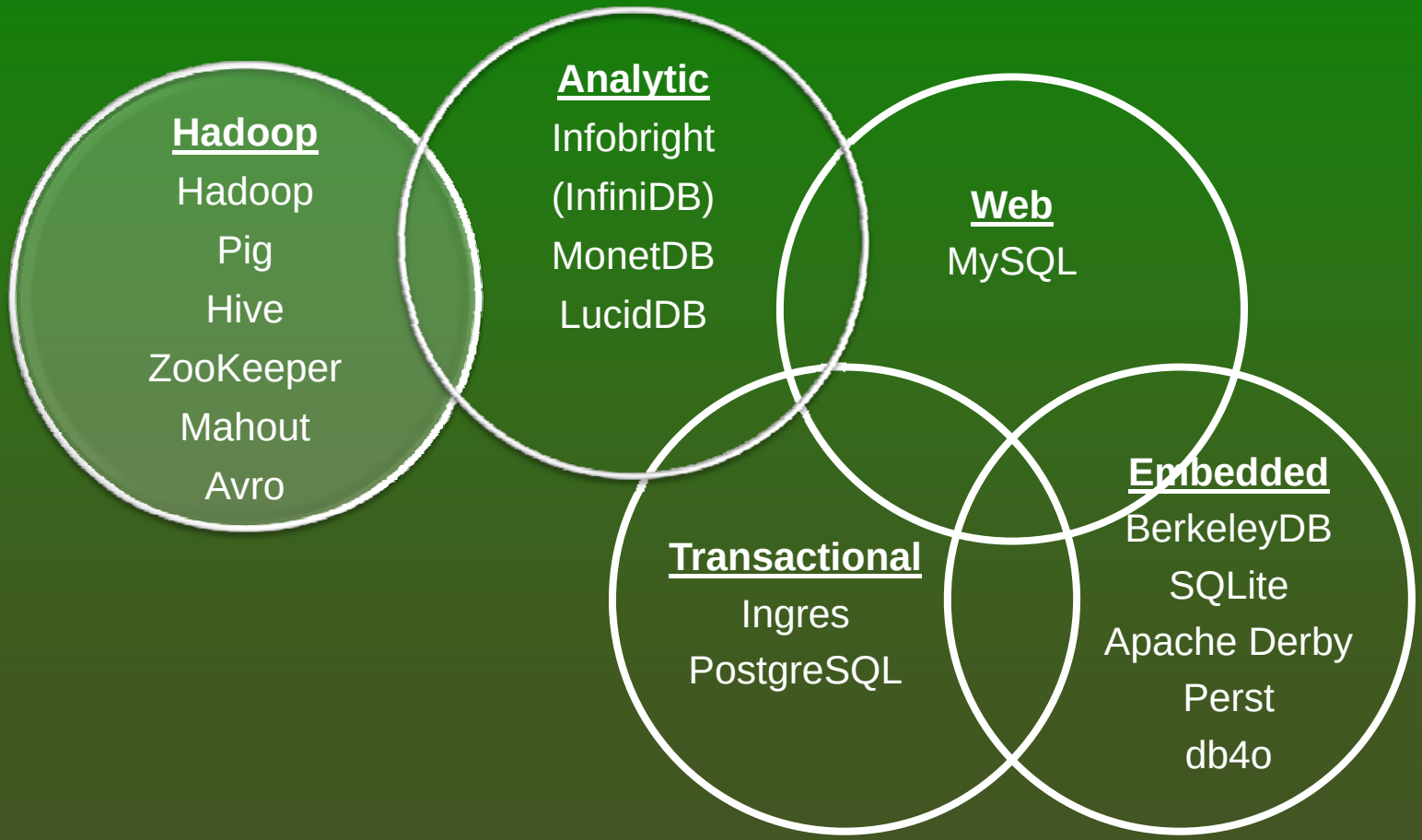
5 IMPLICATIONS

1 BOTTOM LINE

AUGUST 2010

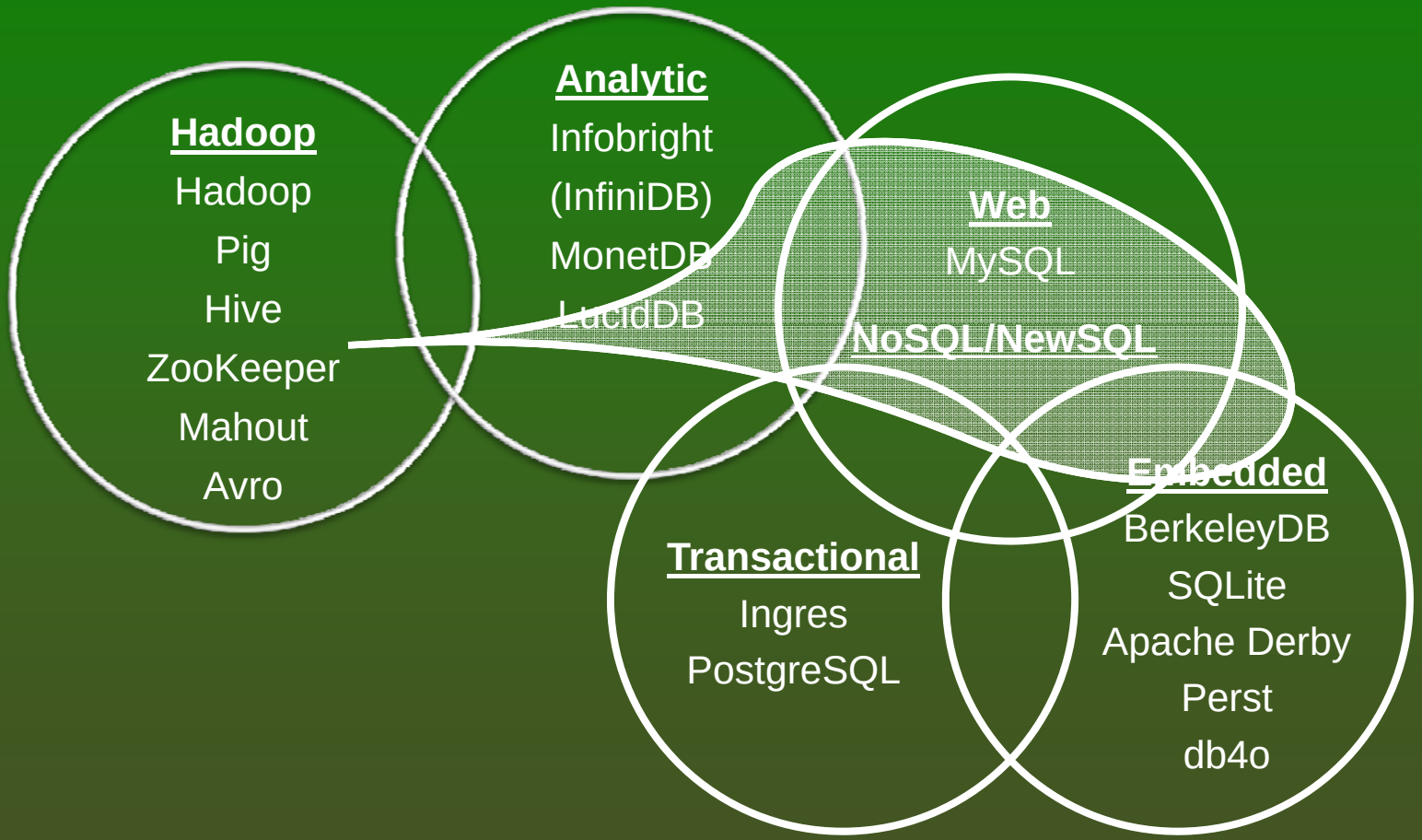
© 2011 by The 451 Group. All rights reserved

Open source database landscape 2010



© 2011 by The 451 Group. All rights reserved

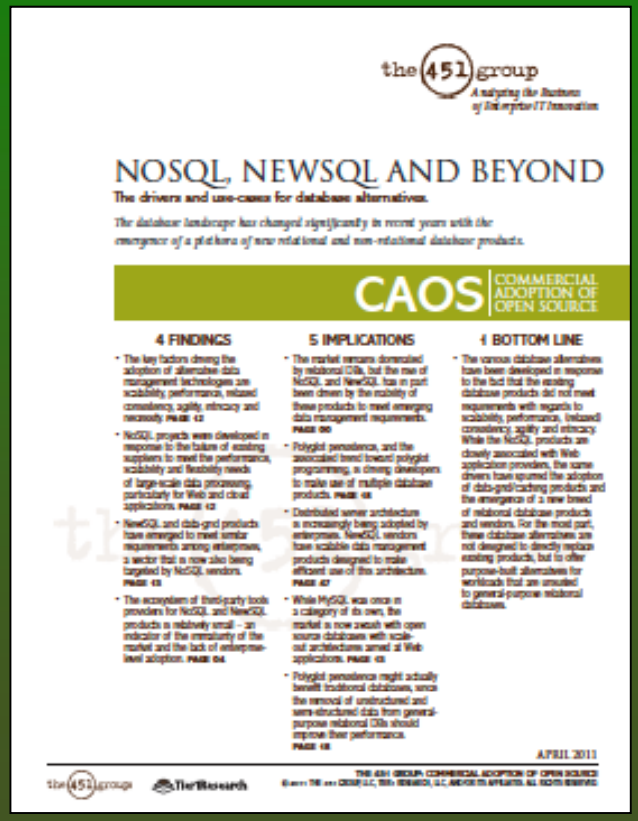
Open source database landscape 2011



© 2011 by The 451 Group. All rights reserved

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
  - 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

- The key factors driving the adoption of alternative data management technologies are scalability, performance, relaxed consistency, agility, efficacy and necessity.
- NoSQL projects were developed in response to the failure of existing suppliers to meet the performance, scalability and flexibility needs of large-scale data processing, particularly for Web and cloud applications.
- NewSQL and data-grid products have emerged to meet similar requirements among enterprises, a sector that is now also being targeted by NoSQL vendors.
- The ecosystem of third-party tools providers for NoSQL and NewSQL products is relatively small – an indication of the immaturity of the market and the lack of enterprise-level adoption.

5 IMPLICATIONS

- The market remains dominated by relational DBs, but the rise of NoSQL and NewSQL has in part been driven by the inability of these products to meet emerging data management requirements.
- Polyglot persistence, and the associated trend toward polyglot programming, is driving developers to make use of multiple database products.
- Distributed server architecture is increasingly being adopted by enterprises. NewSQL vendors have scalable data management products designed to make efficient use of this architecture.
- While MySQL was once in a category of its own, the market is now awash with open source databases with scale-out architectures served at Web applications.
- Polyglot persistence might actually benefit traditional databases, since the removal of unstructured and semi-structured data from general-purpose relational DBs should improve their performance.

1 BOTTOM LINE

- The various database alternatives have been developed in response to the fact that the existing database products did not meet requirements with regards to scalability, performance, (relaxed) consistency, agility and efficacy. While the NoSQL products are closely associated with Web application providers, the same drivers have spurred the adoption of data-grid/caching products and the emergence of a new breed of relational database products and vendors. For the most part, these database alternatives are not designed to directly replace existing products, but to offer purpose-built alternatives for workloads that are unsuited to general-purpose relational databases.

APRIL 2011

the 451 group TheResearch

© 2011 by The 451 Group. All rights reserved

NoSQL, NewSQL and Beyond

- NoSQL
  - New non-relational databases
  - Rejection of fixed table schema and join operations
  - Meet scalability requirements of distributed architectures
  - And/or schema-less data management requirements
- NewSQL
  - New relational databases
  - Retain SQL and ACID compliance
  - Meet scalability requirements of distributed architectures
  - Or to improve performance such that horizontal scalability is no longer a necessity
- Data grid/cache
  - Store data in memory to increase app and database performance
  - Potential alternative to relational databases as the primary platform for distributed data management

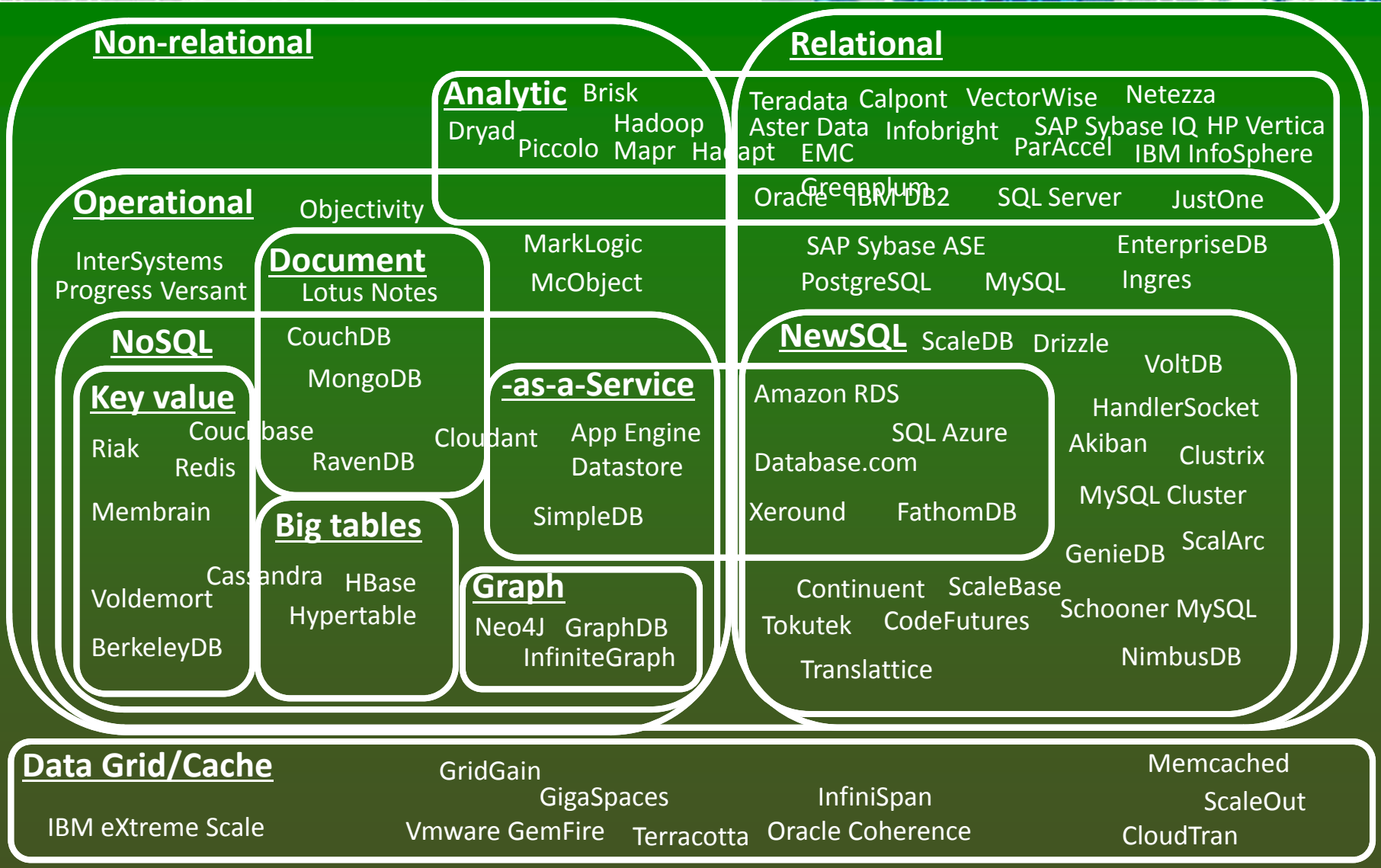
© 2011 by The 451 Group. All rights reserved

NoSQL, NewSQL and Beyond

- NoSQL
  - Big tables
  - Key value stores
  - Document stores
  - Graph databases
- NewSQL
  - MySQL storage engines
  - Transparent sharding
  - Appliances: software and hardware
  - New databases
- Data grid/cache
  - Spectrum of data management capabilities
  - From non-persistent data caching to persistent caching, replication, and distributed data and compute grid functionality

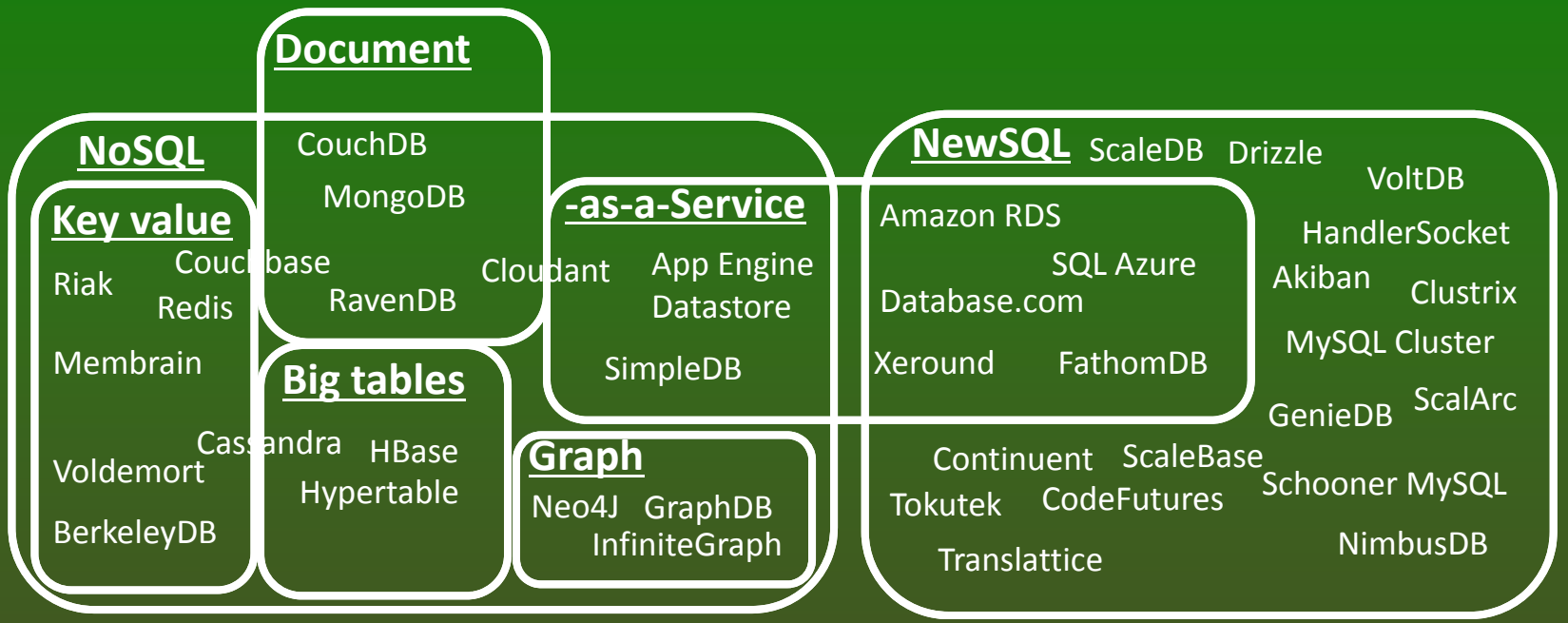
© 2011 by The 451 Group. All rights reserved

Building Your Big Data Future with Open Source



© 2011 by The 451 Group. All rights reserved

NoSQL, NewSQL and Beyond



Data Grid/Cache

GridGain
GigaSpaces
InfiniSpan
Memcached
ScaleOut
IBM eXtreme Scale
Vmware GemFire
Terracotta
Oracle Coherence
CloudTran

© 2011 by The 451 Group. All rights reserved

Database SPRAIN

- "An injury to ligaments… caused by being stretched beyond normal capacity"

  Wikipedia

- Six key drivers for NoSQL/NewSQL/DDG adoption
  - **S**calability
  - **P**erformance
  - **R**elaxed consistency
  - **A**gility
  - **I**ntricacy
  - **N**ecessity

© 2011 by The 451 Group. All rights reserved

Scalability

- Associated sub-driver: Hardware economics
  - Scale-out across clusters of commodity servers
- Example project/service/vendor
  - BigTable, HBase, Riak, MongoDB, Couchbase, Hadoop
  - Amazon RDS, Xeround, SQL Azure, NimbusDB
  - Data grid/cache
- Associated use case:
  - Large-scale distributed data storage
  - Analysis of continuously updated data
  - Multi-tenant PaaS data layer

© 2011 by The 451 Group. All rights reserved

Scalability

- User: StumbleUpon
- Problem:
  - Scaling problems with recommendation engine on MySQL
- Solution: HBase
  - Started using Apache HBase to provide real-time analytics on Su.pr
  - MySQL lacked the performance headroom and scale
  - Multiple benefits including avoiding declaring schema
  - Enables the data to be used for multiple applications and use cases

© 2011 by The 451 Group. All rights reserved

Performance

- Associated sub-driver: MySQL limitations
  - Inability to perform consistently at scale
- Example project/service/vendor
  - Hypertable, Couchbase, 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

© 2011 by The 451 Group. All rights reserved

Performance

- User: AOL Advertising
- Problem:
  - Real-time data processing to support targeted advertising
- Solution: Membase Server
  - Segmentation analysis runs in CDH, results passed into Membase
  - Make use of its sub-millisecond data delivery
  - More time for analysis as part of a 40ms targeted ad response time
  - Also real time log and event management

© 2011 by The 451 Group. All rights reserved

Relaxed consistency

- Associated sub-driver: CAP theorem
  - The need to relax consistency in order to maintain availability
- Example project/service/vendor
  - Dynamo, Voldemort, Cassandra
  - Amazon SimpleDB
- Associated use case:
  - Multi-data center replication
  - Service availability
  - Non-transactional data off-load

© 2011 by The 451 Group. All rights reserved

Relaxed consistency

- User: Wordnik
- Problem:
  - MySQL too consistent –blocked access to data during inserts and created numerous temp files to stay consistent.
- Solution: MongoDB
  - Single word definition contains multiple data items from various sources
  - MongoDB stores data as a complete document
  - Reduced the complexity of data storage

© 2011 by The 451 Group. All rights reserved

Agility

- Associated sub-driver: Polyglot persistence
  - Choose most appropriate storage technology for app in development
- Example project/service/vendor
  - MongoDB, CouchDB, Cassandra
  - Google App Engine, SimpleDB, SQL Azure
- Associated use case:
  - Mobile/remote device synchronization
  - Agile development
  - Data caching

© 2011 by The 451 Group. All rights reserved

Agility

- User: Dimagi BHOMA (Better Health Outcomes through Mentoring and Assessments) project
- Problem:
  - Deliver patient information to clinics despite a lack of reliable Internet connections
- Solution: Apache CouchDB
  - Replicates data from regional to national database
  - When Internet connection, and power, is available
  - Upload patient data from cell phones to local clinic

© 2011 by The 451 Group. All rights reserved

Intricacy

- Associated sub-driver: Big data, total data
  - Rising data volume, variety and velocity
- Example project/service/vendor
  - Neo4j, GraphDB, InfiniteGraph
  - Apache Cassandra, Hadoop,
  - VoltDB, Clustrix
- Associated use case:
  - Social networking applications
  - Geo-locational applications
  - Configuration management database

© 2011 by The 451 Group. All rights reserved

Intricacy

- User: Evident Software
- Problem:
  - Mapping infrastructure dependencies for application performance management
- Solution: Neo4j
  - Apache Cassandra stores performance data
  - Neo4j used to map the correlations between different elements
  - Enables users to follow relationships between resources while investigating issues

© 2011 by The 451 Group. All rights reserved

Necessity

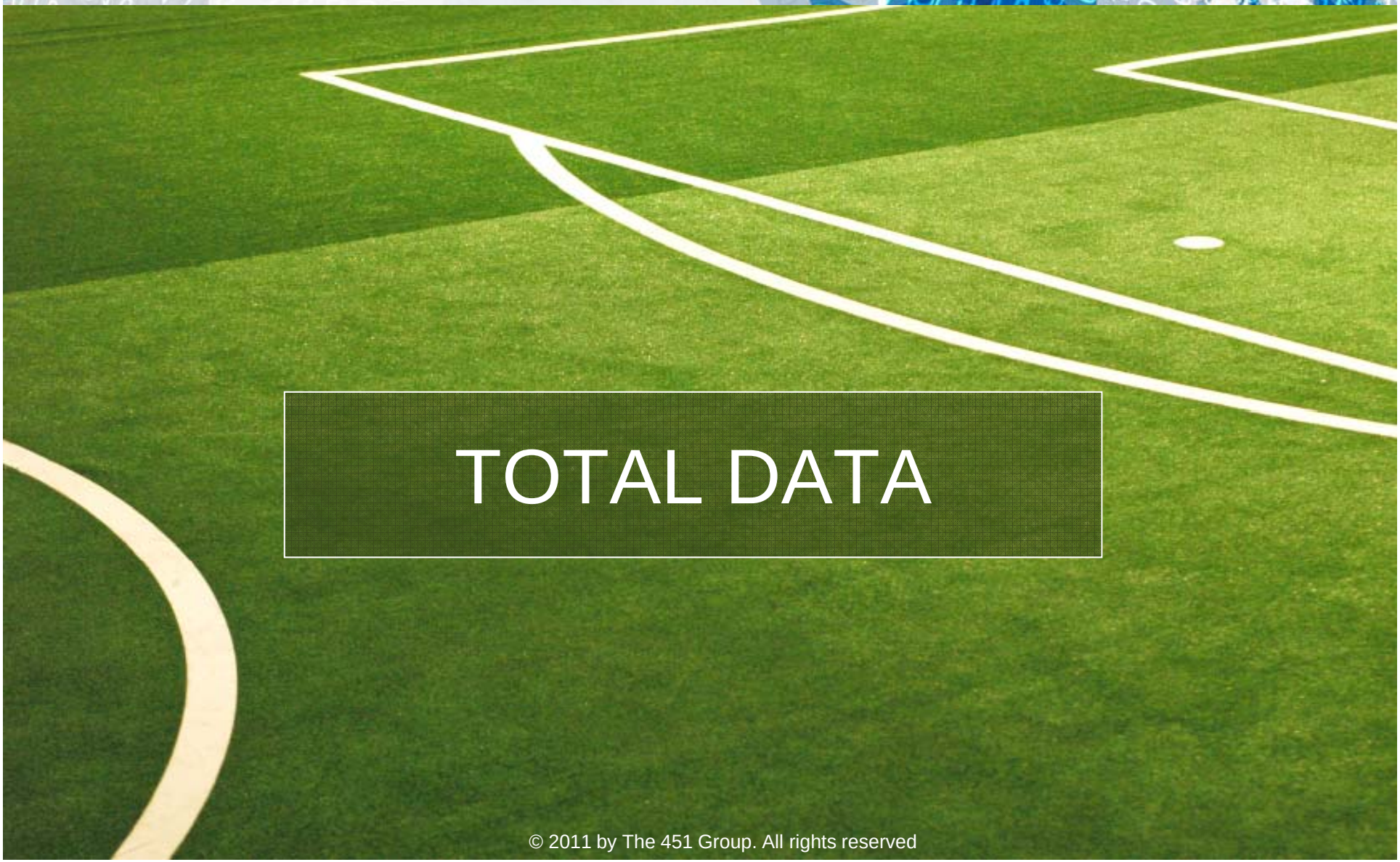
- Associated sub-driver: Open source
  - The failure of existing suppliers to address the performance, scalability and flexibility requirements of large-scale data processing
- Example project/service/vendor
  - BigTable, Dynamo, MapReduce, Memcached
  - Hadoop, HBase, Hypertable, Cassandra, Membase
  - Voldemort, Riak, BigCouch
  - MongoDB, Redis, CouchDB, Neo4J
- Associated use case:
  - All of the above

© 2011 by The 451 Group. All rights reserved

Necessity

- BigTable: Google
- Dynamo: Amazon
- Cassandra: Facebook
- HBase: Powerset
- Voldemort: LinkedIn
- Hypertable: Zvents
- Neo4j: Windh Technologies
  - Yahoo: Apache Hadoop and Apache HBase
  - Digg: Apache Cassandra
  - Twitter: Apache Cassandra, Apache Hadoop and FlockDB

© 2011 by The 451 Group. All rights reserved



© 2011 by The 451 Group. All rights reserved

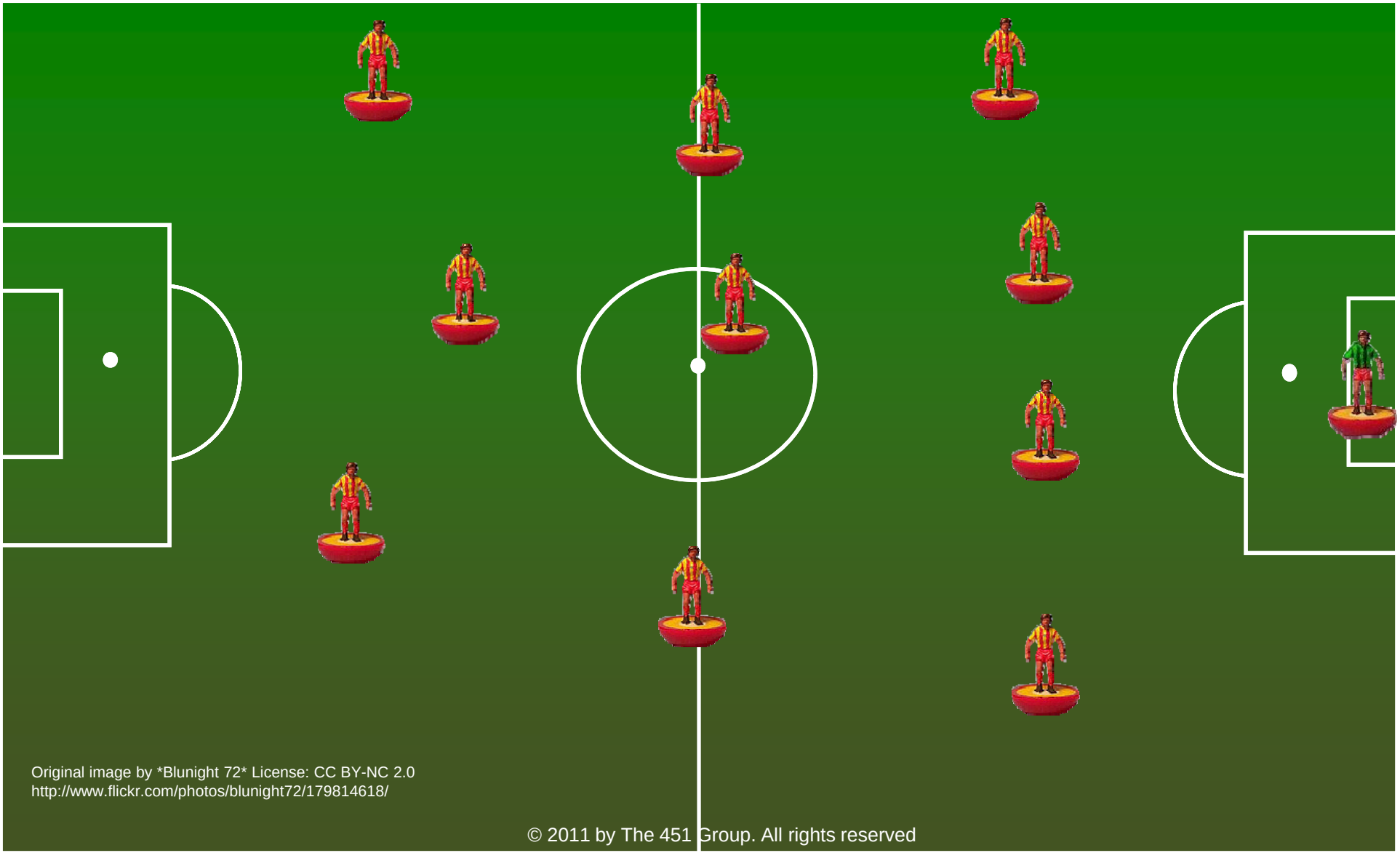
Introducing Total Data

- Defined by The 451 Group to describe new approaches to data management
- Reflects the changing data management landscape as pragmatic choices are made about data storage and analysis techniques
- Processing any data applicable to analytic query
  - in database, data warehouse, or Hadoop, or archive
  - structured or unstructured, relational or non-relational
  - on-premise or in the cloud
- Inspired by ‘Total Football’

© 2011 by The 451 Group. All rights reserved

COMPUTERWORLD OSBC SAN FRANCISCO

Building Your Big Data Future with Open Source



Original image by *Blunight 72* License: CC BY-NC 2.0
http://www.flickr.com/photos/blunight72/179814618/

© 2011 by The 451 Group. All rights reserved

Total Football

– "You make space, you come into space. And if the ball doesn't come, you leave this space and another player will come into it."

Bernadus Hulshoff, Ajax 1966-77

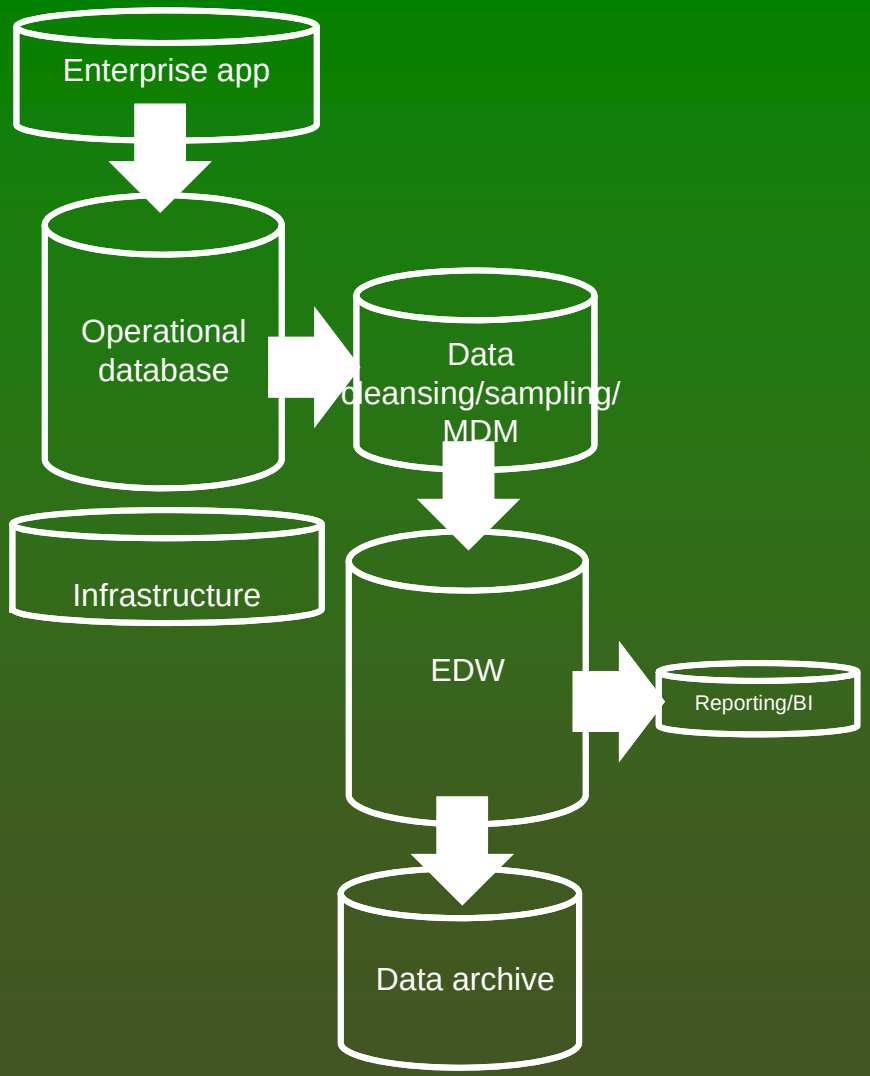
© 2011 by The 451 Group. All rights reserved

Total Football meets Total Data

- Abandonment of restrictive (self-imposed) rules about individual roles and responsibility
- Enabled and relied on fluidity and flexibility to respond to changing requirements
- Reliant on, and exploited, improved performance levels

© 2011 by The 451 Group. All rights reserved

Building Your Big Data Future with Open Source

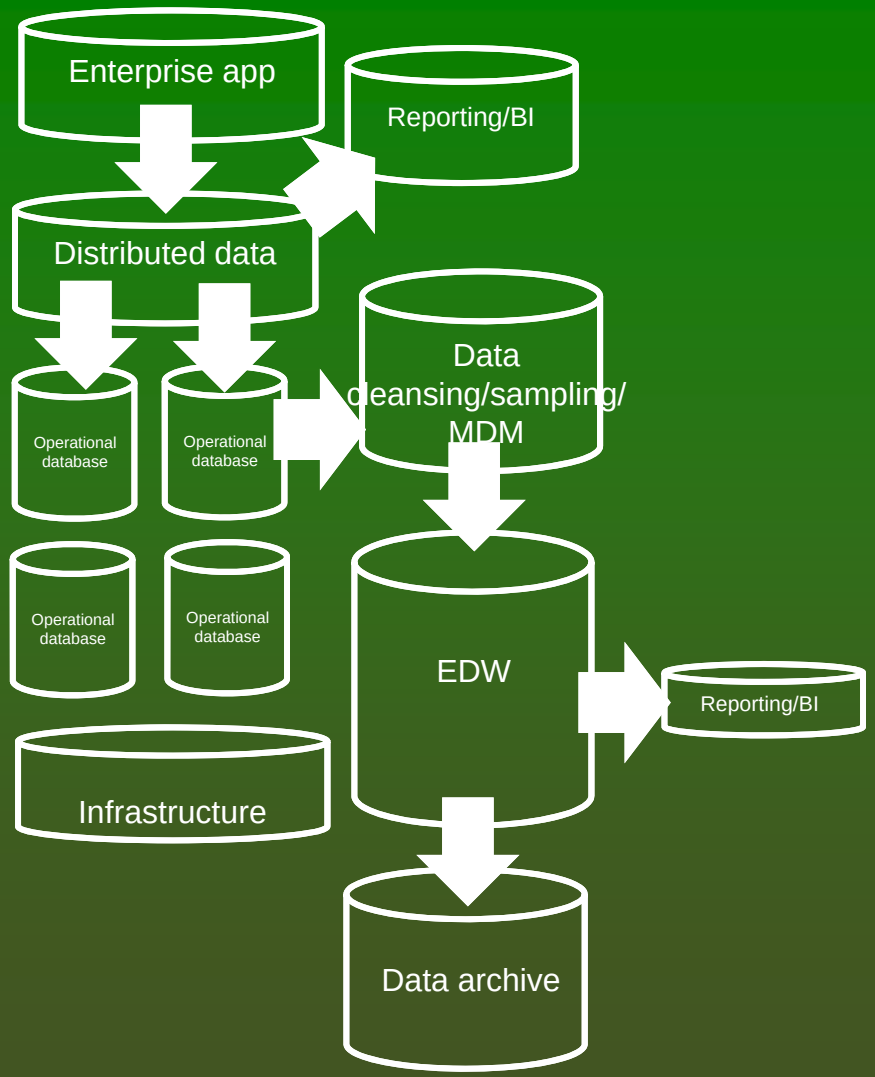


Data management – in theory

- The application is the primary source of data
- The relational database is sacrosanct
- The enterprise data warehouse is the single source of the truth (or is supposed to be)
- Data archived to tape
- Infrastructure primarily exists to support the data/application layer

© 2011 by The 451 Group. All rights reserved

Building Your Big Data Future with Open Source

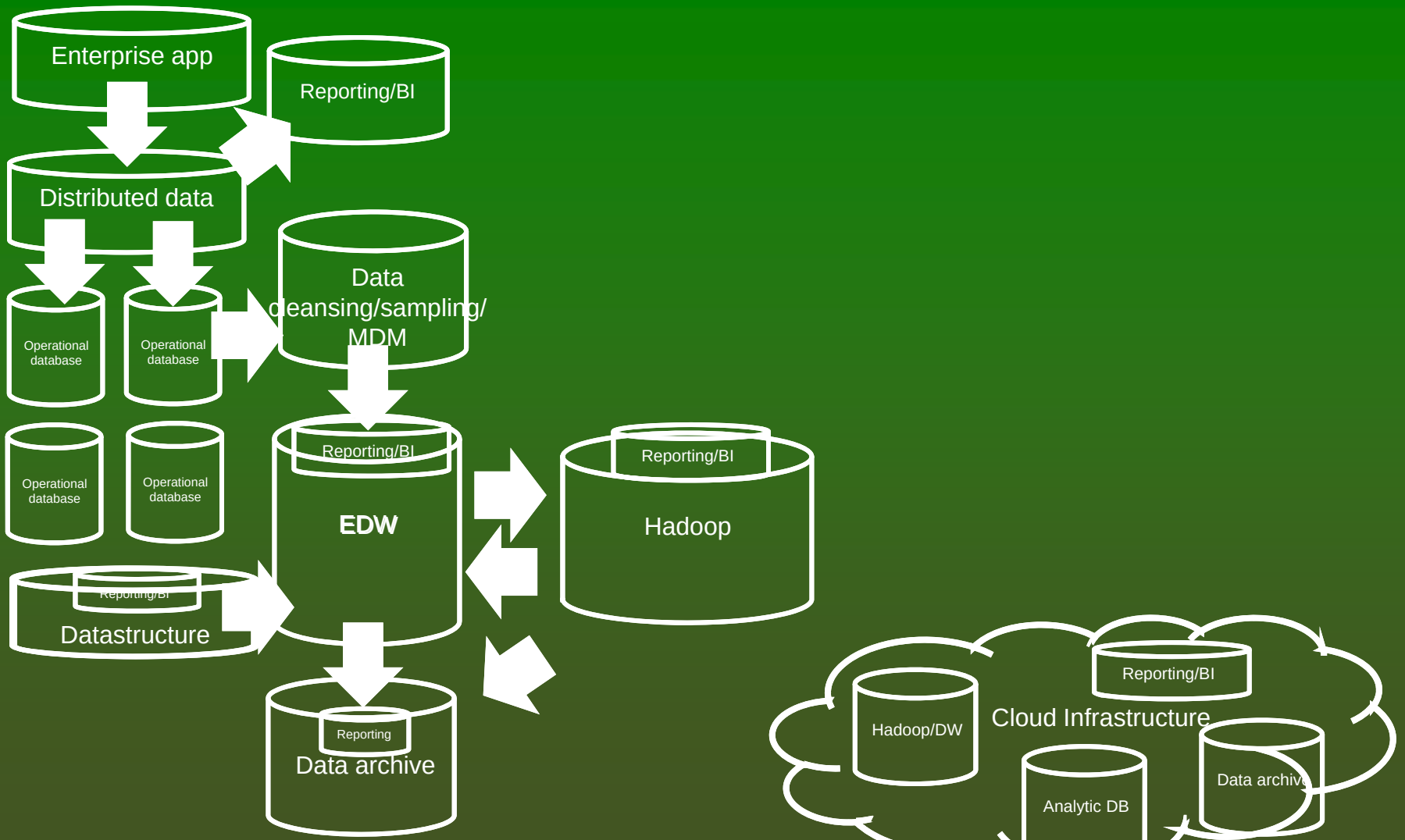


Data management – in practice

- ~~The relational database is sacrosanct~~
- Distributed data layer to meet the scalability and performance demands
- New opportunities for real-time BI
- Polyglot persistence – use the most appropriate data storage for the application

© 2011 by The 451 Group. All rights reserved

Building Your Big Data Future with Open Source



© 2011 by The 451 Group. All rights reserved

Databases in the cloud

- 2008: Operational database moved to the cloud: EnterpriseDB, MySQL, Oracle, SQL Server, DB2 (09)
- 2009: Analytic databases moved to the cloud: Vertica (08), Aster Data, Netezza, RainStor, Greenplum, Teradata
- Some dev and test, some early adopters, but limited take-up
- The software, and licensing, is not designed to be elastic
- Data security, management concerns
- A true “cloud database” must be designed to take advantage of elastic scaling and multi-tenancy

© 2011 by The 451 Group. All rights reserved

Cloud databases

- 2008: Introduction of SimpleDB, Microsoft SDS, Google App Engine
- 2009: Amazon RDS, Microsoft SQL Azure, Rackspace FathomDB
- 2010: NoSQL database vendors: 10gen, DataStax, Hypertable, Basho, Couchbase, Neo Technology,
- 2010: Google High Replication Datastore, NewSQL projects.
- 2011: More from Amazon, Google, Microsoft, Rackspace, NewSQL vendors, distributed data grid providers – GigaSpaces CEAP
- 2012: True cloud databases emerge. PaaS emerges as a potential battleground for NoSQL, NewSQL and DDG providers

© 2011 by The 451 Group. All rights reserved

Data location, data location, data location

- Not the end of the EDW, but EDW is one of many sources of BI, rather than the only source of BI
- Choose the right storage technology: software and hardware
  - EDW, Hadoop or archive
  - On-premise or on the cloud
  - Memory, disk or SSD

© 2011 by The 451 Group. All rights reserved

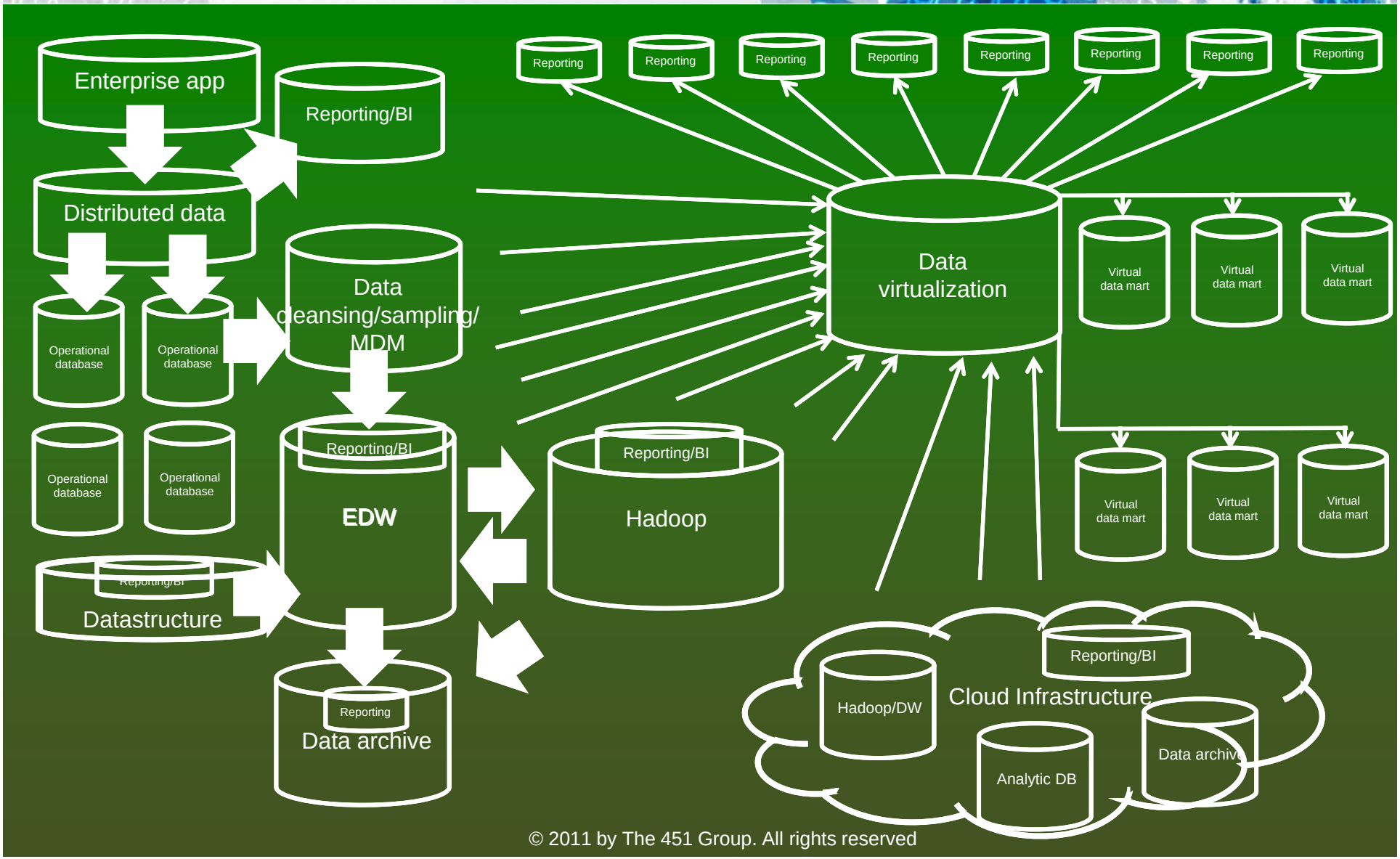
Data location, data location, data location

- Understand the requirements:
  - Value and temperature of the data
  - Ensure data can be queried using existing tools/skills
  - Cost
- The issue of data location becomes paramount
  - Avoid data movement and duplication – retain governance
  - Virtual data marts and data cloud
  - Data virtualization to provide access to multiple data sources

© 2011 by The 451 Group. All rights reserved

COMPUTERWORLD OSBC SAN FRANCISCO

Building Your Big Data Future with Open Source



© 2011 by The 451 Group. All rights reserved

Relevant reports

- Total Data
  - Explaining the total data management approach to dealing with the impact of big data on the data management landscape
  - Coming late 2011
  - Including the growing Hadoop ecosystem and real-time
  - sales@the451group.com

COMING LATE 2011

© 2011 by The 451 Group. All rights reserved

Conclusion

- The key factors driving the adoption of alternative data management technologies are scalability, performance, relaxed consistency, agility, intricacy and necessity (SPRAIN).
- NoSQL projects were developed in response to the failure of existing suppliers to meet the performance, scalability and flexibility needs of large-scale data processing, particularly for Web and cloud computing applications.
- NewSQL and data-grid products have emerged to fulfill a similar requirement among enterprises, a sector that is also now being targeted by NoSQL vendors.

© 2011 by The 451 Group. All rights reserved

Conclusion

- The database market remains dominated by relational DBs and incumbent industry giants, but the rise of NoSQL and NewSQL has in part been driven by the failure of these products to address emerging data management requirements.

- For the most part these database alternatives are not designed to directly replace existing database products, but to offer purpose-built alternatives for workloads that are unsuited to general-purpose relational databases.

- However, PaaS – and the distributed data layer that will power future PaaS – is emerging as a potential battleground for NoSQL, NewSQL and DDG providers.

© 2011 by The 451 Group. All rights reserved

© 2011 by The 451 Group. All rights reserved