Big Data and Advanced Analytics Technologies and Use Cases

Colin White
President, BI Research
DAMA Portland
February 2013

Agenda

There is considerable interest at present on the topic of big data. Much of the discussion about this topic, however, is focused on the technology supporting big data, rather than on how analytics generated from big data can be leveraged for business benefit. One of the most exciting aspects of big data technology is that it allows organizations to support advanced analytic workloads and applications that were not previously possible for cost or performance reasons. New and evolving big data solutions provide significant business benefits because they help remove these cost and performance barriers. The objectives of this presentation are to discuss the benefits of big data and to present use cases and case studies that demonstrate the value of advanced analytics. It also explains how the existing data warehousing environment can be extended to support big data solutions. Topics that will be covered include:

- Review the history and evolution of big data and advanced analytics
- Explain the role of the data scientist in developing advanced analytics
- Look at the technologies that support big data
- Explain how the existing data warehousing environment can be extended to support big data and advanced analytics
- Discuss big data use cases and the benefits they bring to the business

The Evolution of Digital Data

First OLTP systems

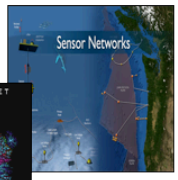
Early decision support products

First commercial RDBMSs

Early data warehousing

Big data & advanced analytics

1960 — 1970 — 1980 — 1990 — 2012

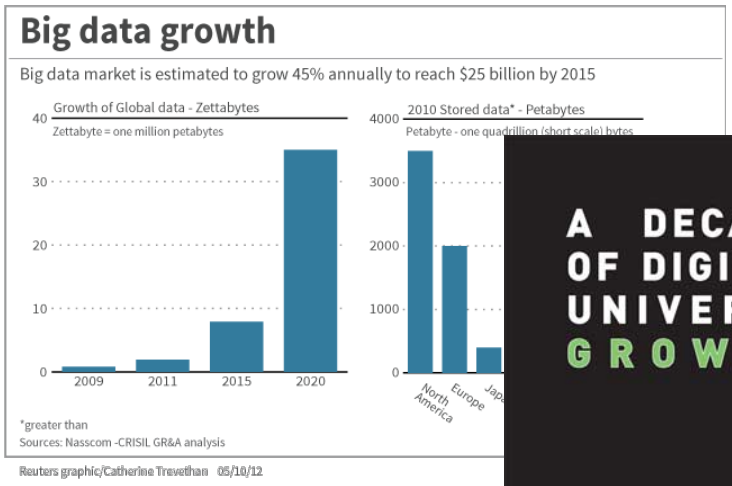


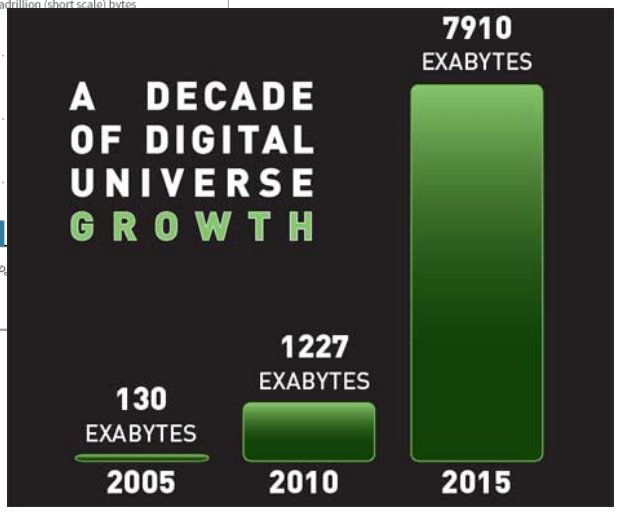
Sabre
84,000 txs/day

Increasing Data Volumes

Sabre
60,000 txs/sec
25 TB EDW

Data Growth: Choose an Analyst!





Data Growth: Multi-Structured Data

Definition: data that has unknown, ill-defined or overlapping schemas

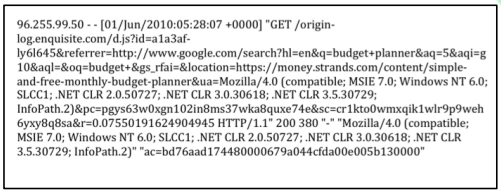
- Machine generated data, e.g., sensor data, system logs
- Internal/external web content including social computing data
- Text, document and XML data
- Graph, map and multi-media data

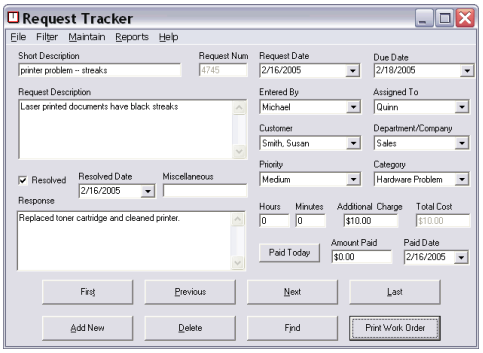
Volume increasing faster than structured data

Usually not integrated into a data warehouse

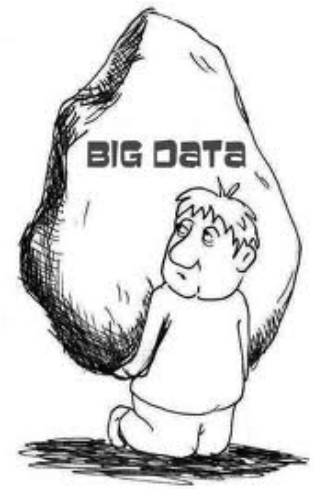
Increasing number of analytical techniques to extract useful information from this data

This information can be used to extend traditional predictive models and analytics





Data Growth: Big Data



Big data technologies apply to all types of digital data not just multi-structured data

"Big" is a relative term and is different for each organization and application

What you do with big data and how you use it for business benefit should be the main consideration – analytics play a key role here

The Value of Data: IBM 2012 Study

IBM Center for Applied Insights

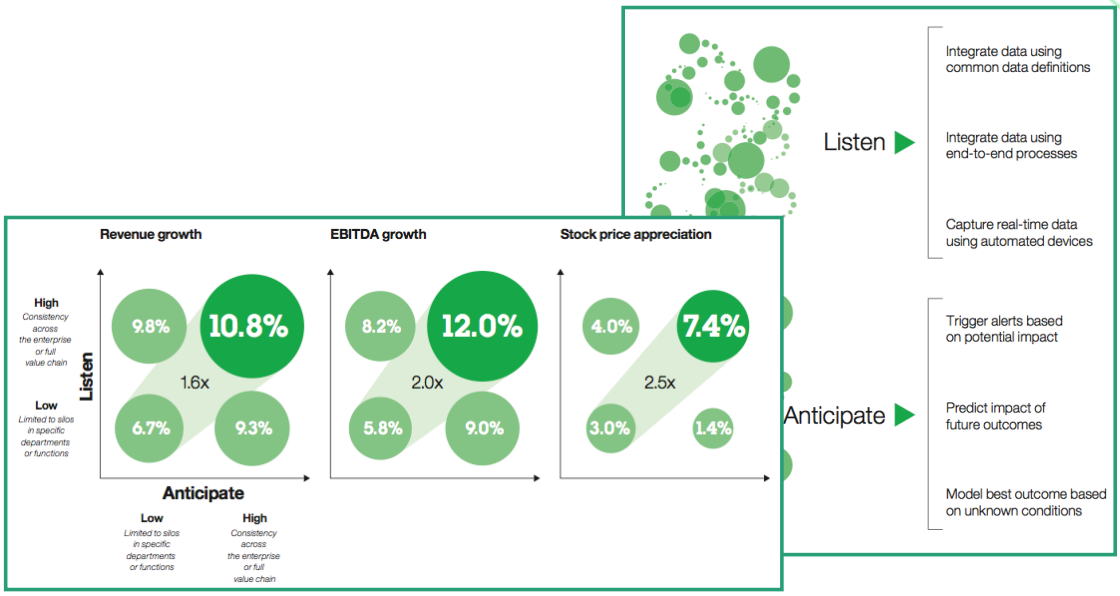
Outperforming in a data-rich, hyper-connected world

We live in the era of "big data" and connectivity. Collectively, our planet generates **fifteen petabytes** of new information every day[1] – about eight times the information housed in all the academic libraries in the United States.[2] People are connecting and communicating more than at any time in human history. In 2011 alone, an estimated **seven trillion** text messages were sent.[3] There are almost **six billion** mobile phone subscriptions.[4] Today, **two billion** people worldwide are plugged into the Internet,[5] interacting and sharing information on sites like Facebook and Twitter.

About the Study

The IBM Center for Applied Insights, in cooperation with the Economist Intelligence Unit and the IBM Institute of Business Value, recently surveyed 1,168 executives (two thirds of which were CxOs) across nine industries in 64 countries. We investigated their ability to capture data, share insights and take actions based on what they learn, and used a binary logistic regression analysis to understand the statistically significant correlations with financial performance.

The Value of Data: IBM 2012 Study



The Changing World of BI Analytics

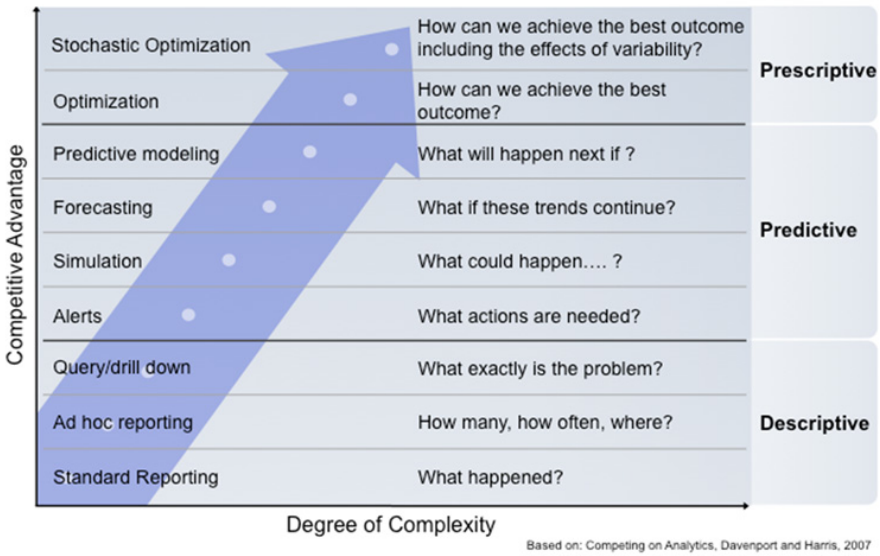
Advanced Analytics

- Improved analytic tools and techniques for statistical and predictive analytics
- New tools for exploring and visualizing new varieties of data
- Operational intelligence with embedded BI services and BI automation

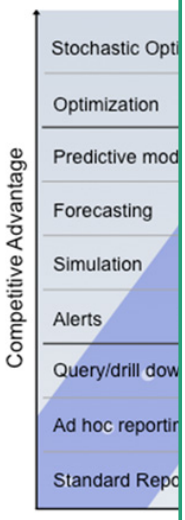
Data Management

- Analytic relational database systems that offer improved price/performance and libraries of analytic functions
- In-memory computing for high performance
- Non-relational systems such as Hadoop for handling new types of data
- Stream processing/CEP systems for analyzing in-motion data

Advanced Analytics Example: SC Digest 2012



Advanced Analytics Example: SC Digest 2012



1. **Descriptive analytics** – using historical data to describe the business. This is usually associated with Business Intelligence (BI) or visibility systems. In supply chain, you use descriptive analytics to better understand your historical demand patterns, to understand how product flows through your supply chain, and to understand when a shipment might be late.

2. **Predictive analytics** – using data to predict trends and patterns. This is commonly associated with statistics. In the supply chain, you use predictive analytics to forecast future demand or to forecast the price of fuel.

3. **Prescriptive analytics** – using data to suggest the optimal solution. This is commonly associated with optimization. In the supply chain, you use prescriptive analytics to set your inventory levels, schedule your plants, or route your trucks.

Advanced Analytics in Supply Chain, Dr. Michael Watson, Supply Chain Digest, November 2012

The Role of the Data Scientist: CITO Interviews

"Data scientists turn big data into big value, delivering products that delight users, and insight that informs business decisions.

Strong analytical skills are a given: above all, a data scientist needs to be able to derive robust conclusions from data. But a data scientist also needs to possess creativity and strong communication skills."

Daniel Tunkelang, Principal Data Scientist, LinkedIn

"A data scientist is someone who can obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning. Data scientists not only are adept at working with data, but appreciate data itself as a first-class product."

Hilary Mason, Chief Scientist at bitly

"... someone who has the both the engineering skills to acquire and manage large data sets, and also has the statistician's skills to extract value from the large data sets and present that data to a large audience."

John Rauser, Principal Engineer, Amazon.com

Source: citoresearch.com/content/growing-your-own-data-scientists

Data Science Skills Requirements

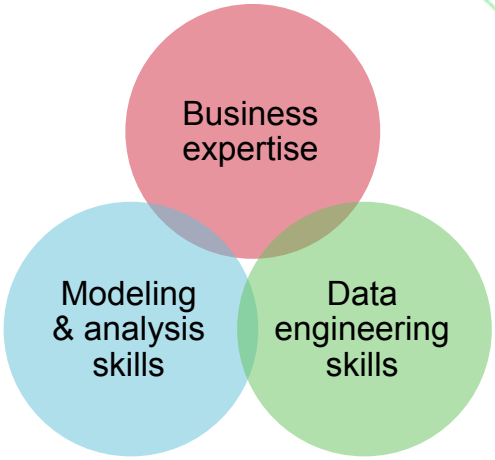
Business domain subject matter expert with strong analytical skills

Creativity and a good communications

Knowledgeable in statistics, machine learning and data visualization

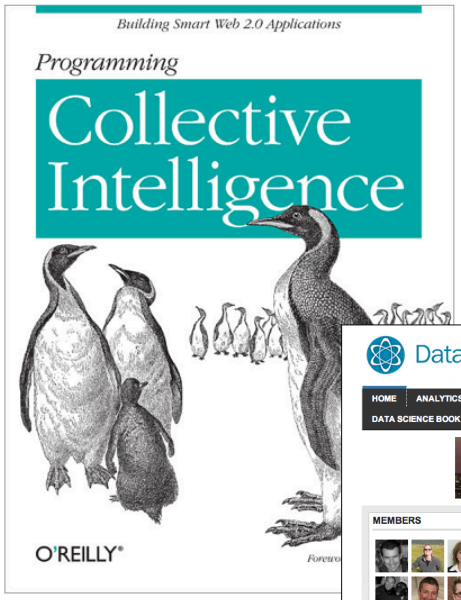
Able to develop data analysis solutions using modeling/analysis methods and languages, such as MapReduce, R, SAS, etc.

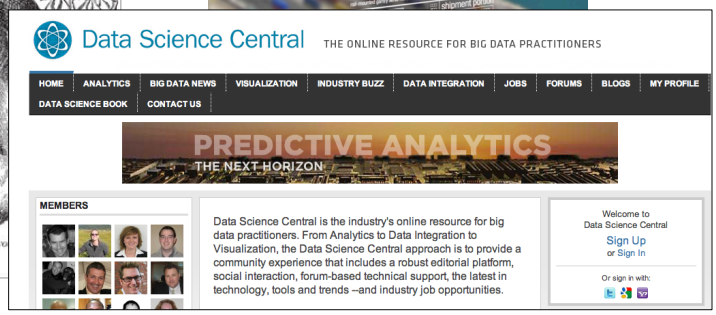
Adept at data engineering, including discovering and mashing/blending large amounts of data



Is this one person or a team of specialists?

Data Science: Further Reading





What Then is Big Data?

Represents ***analytic and data management solutions*** that could not previously be supported because of:

- Technology limitations – poor performance, inadequate analytic capabilities, etc.
- High hardware and software costs
- Incomplete or limited data for generating the required solutions

Set of overlapping technologies that enable customers to deploy analytic systems *optimized* to suite specific business needs and workloads

Optimization may involve improving performance, reducing costs, enabling new types of data to be analyzed, etc.

Big Data Application Examples

Source: Microsoft

Big Data and Data Life Cycle Management

Data Management and Analytic Performance

- Capacity planning
- Managing data warehouse growth
- Analytic performance management and optimization
- Service level agreements

Data Governance

- Security: user access, encryption, masking, etc.
- Quality: governed/ungoverned data
- Backup and recovery
- Archiving and retention: historical analysis, compliance

The Impact of Big Data on the Data Life Cycle

Need fast time to value to quickly gain business benefits from big data

- Impractical to use traditional EDW approach for all analytic solutions
- Extend existing data warehousing environment to support big data and accommodate data growth

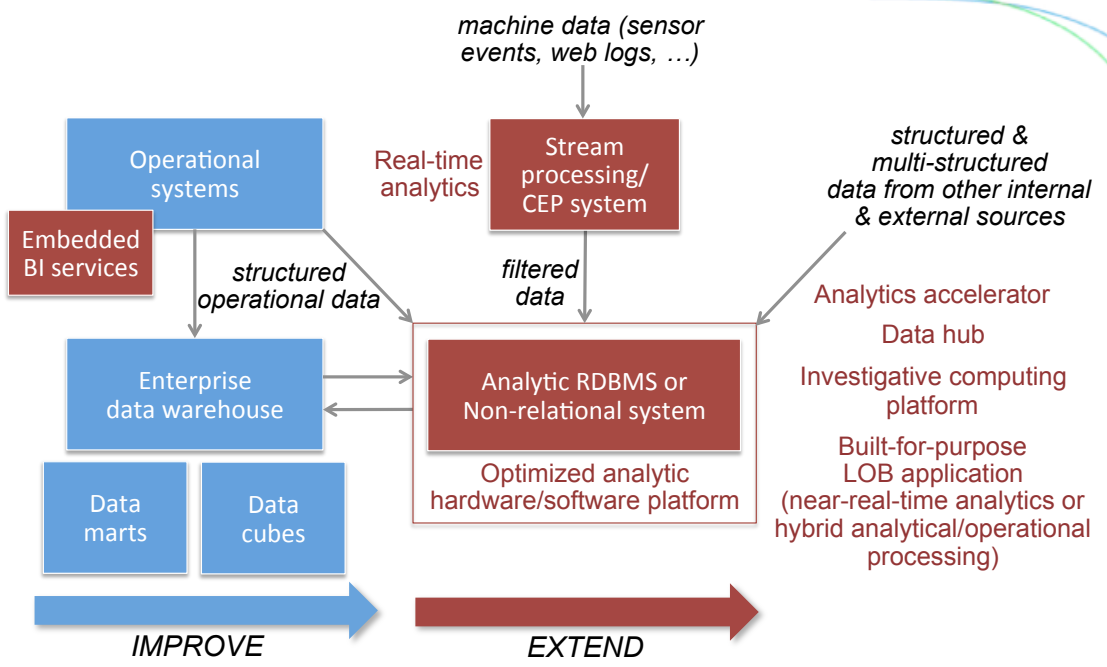
Need high performance solutions for supporting big data analytic workloads

- One-size fits all data management is no longer viable
- Match technologies and costs to business needs and analytic workloads

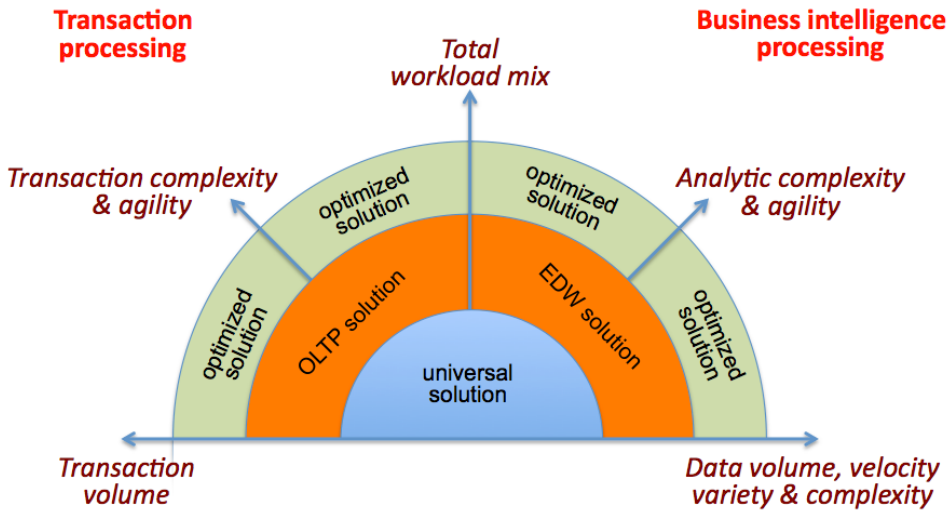
Need improved data governance to handle big data

- No longer practical to rigidly control and govern all forms of data – implement different levels of governance based on security, compliance and quality needs
- Determine data archiving and policies based on the possible future need to analyze historical data and data compliance requirements

The Extended (or Logical) Data Warehouse



Beyond the EDW: Optimized Platforms



Optimized Analytic Platforms: Variables

Analytics Required to Meet Business Needs

- Complexity – reporting, OLAP or advanced analytics
- Agility – latency of data, analytics, decisions, recommendations and actions
- Workload mix – complexity of overall analytic workload; concurrent data modification

Data Required to Meet Business Needs

- Volume – amount of data to be managed
- Velocity – rate of data generation or change
- Variety – types of data to be managed
- Complexity – number of data sources and relationships; quality and structure of data

Optimized Platforms: Analytic RDBMSs

Extend traditional RDBMSs with features designed specifically for analytic processing and new analytic techniques

Hardware exploitation
Parallel computing
New data types
New storage structures
Data compression
Support for hybrid storage
Intelligent workload management
In-memory data
In-memory analytics
In-database aggregation & analytics

Analytic RDBMSs: Hardware Exploitation

Faster processors

Multi-core processors

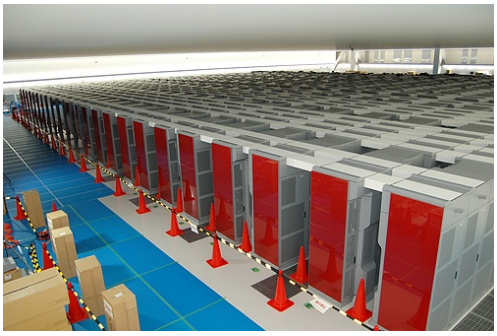
Intelligent hardware

64-bit memory spaces

Large-capacity disk drives

Fast hard-disk and solid-state drives

Hybrid storage configurations



Scale-up/out parallel processing configurations

Lower-cost hardware (blades, clusters)

Reduced power and cooling requirements

Packaged hardware/software appliances

Analytic RDBMSs: New Storage Structures

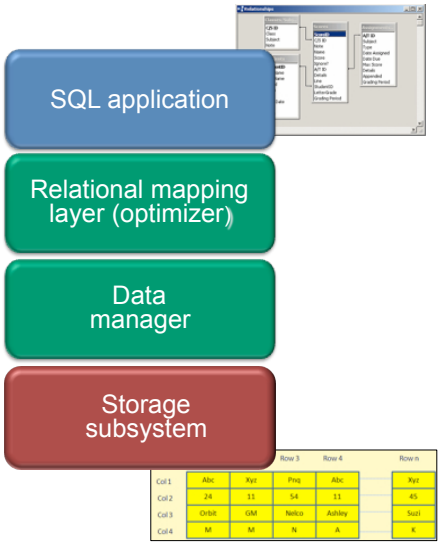
DBMS vendors are enabling new *physical storage structures* to improve performance, reduce storage requirements and support new types of analyses

Examples: compressed columnar, XML, time-series, multi-media

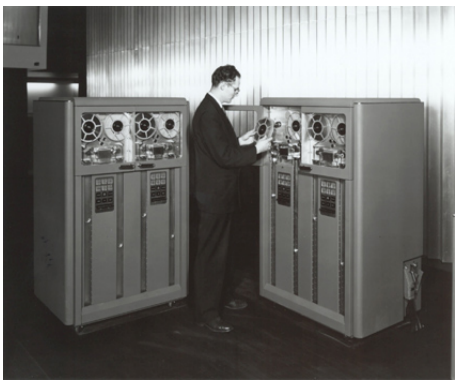
These enhancements and their implementation vary by vendor

It is important to recognize that *physical* storage structures should be independent of the *logical* data model and the data manipulation language (DML)

- True for the relational model and SQL
- Often not true for non-relational systems



Analytic RDBMSs: Data Storage Options



Large-capacity hard-disk drives (HDD)

- More economical, less reliable, slower performance, e.g., SATA drives in white-box H/W
- Often “short-stroked” to improve performance

High-performance hard-disk drives (HDD)

- More expensive, more reliable, better performance, e.g., enterprise SAS drives

Solid-state drives (SSDs)

- High and consistent performance
- Better reliability and more energy efficient
- Distinguish between commodity SSDs and enterprise SSDs

Dynamic RAM (DRAM)

- Best performance - eliminates I/O overheads
- Use by in-memory computing systems

Memory versus Storage

Memory

- Data is directly addressable by a CPU via a memory bus
- Eliminates I/O overhead and provides fast access to data
- Types of solid-state memory:
  - Processor cache(s) – very fast, volatile data
  - Dynamic RAM – fast (nanoseconds), volatile data

Storage

- Data is addressable via a device interconnect or network protocol
- Several types of storage for persisting data:
  - Commodity HDD: high capacity, less reliable, low cost (e.g., SATA)
  - Enterprise HDD: more reliable, higher cost (e.g., SAS)
  - NAND flash memory devices: fast, very reliable, high cost (e.g., PC SSD, enterprise SSD, flash storage array, hybrid SSD/HDD)

What is In-Memory Computing?

A workload where all the data being processed is stored in a computer memory that is directly addressable via the CPU's memory bus

Provides high-speed performance for OLTP and BI workloads by eliminating I/O to storage devices

Especially beneficial for *interactive* and *iterative* BI analytic workloads

Several types of BI-related in-memory computing

The Changing World of BI Analytics

Advanced Analytics ← In-memory analytics

- Improved analytic tools and techniques for statistical and predictive analytics
- New tools for exploring and visualizing new varieties of data
- Operational intelligence with embedded BI and BI automation

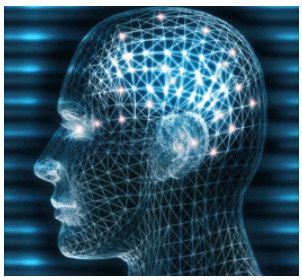
Big Data Management ← In-memory data

- Analytic relational database systems that offer improved price/performance and libraries of analytic functions
- Non-relational systems such as Hadoop for handling new types of data
- Stream processing/CEP systems for analyzing in-motion data

Why In-Memory Computing for BI Analytics?

Benefits

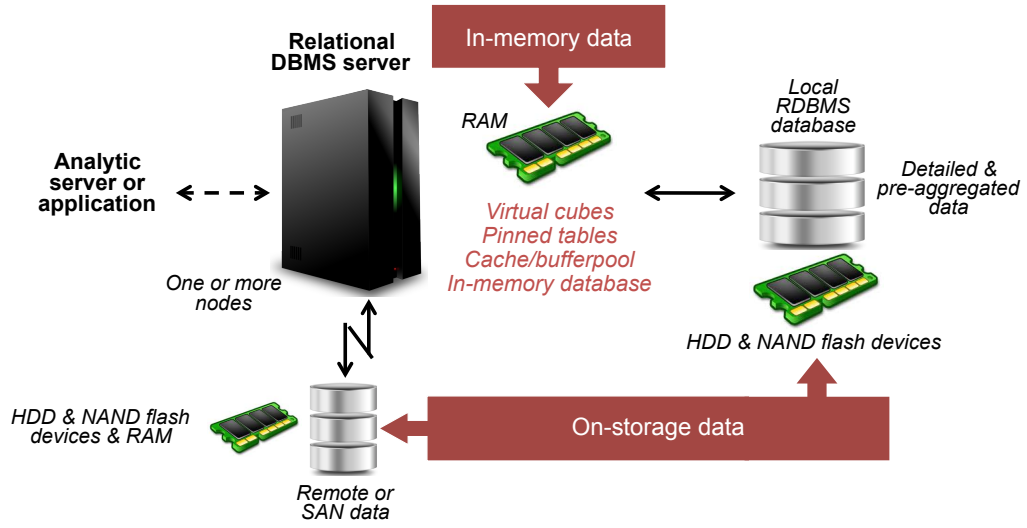
- Technology answer: Improved speed and performance, e.g., quickly run complex analyses on the fly
- Business answer: What if you could do ...? e.g., real-time fraud detection



Considerations

- Types of *in-memory data* and *in-memory analytics* and their benefits
- Relationship to in-database processing, e.g., in-database aggregation and in-database analytics

In-Memory Data



In-Memory Database Systems: Vendor Examples

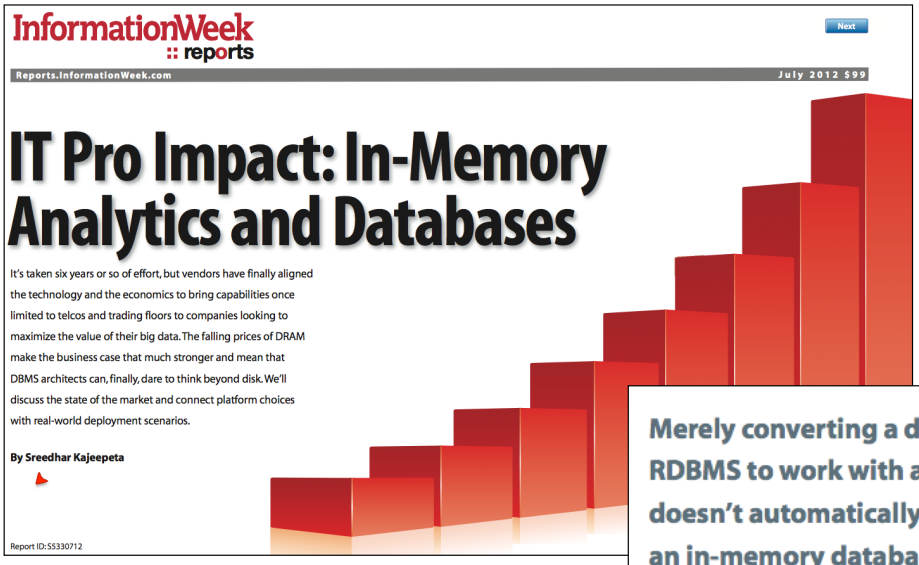
Relational DBMSs
- EXASOL EXASolution
- IBM solidDB and Informix Warehouse Accelerator
- Kognitio Analytical Platform
- Microsoft xVelocity and “Hekaton”
- Oracle TimesTen
- SAP HANA and Sybase ASE
- VoltDB
- Many vendors also support “pinned” tables

Non-Relational DBMSs
- Memcached

Multi-Dimensional DBMSs
- IBM Cognos TM1

In-Memory Data Grids
- Cloud platforms from Amazon, Google, IBM, Microsoft, VMware, etc.

In-Memory Data: Important to Note



Merely converting a disk-based RDBMS to work with a RAM disk doesn't automatically make it an in-memory database.

In-Database Technologies

In-Database Aggregation

- Some RDBMSs support pre-aggregation to enhance BI performance
- Should be transparent to user – optimizer decides when to use aggregate
- Various names – materialized views, materialized query tables, etc.

In-Database Analytics

- Brings the processing to the data rather than the data to the processing
- Consists primarily of predefined analytic functions – created by RDBMS vendor, third-party vendor, open source community, user developed

In-Database Analytic Functions

Analytical functions stored in an RDBMS offer several benefits

- Users (e.g., data scientists) only need to understand what a function does and how to use it - they do not need to know how to develop the functions
- Functions can exploit the parallel processing capabilities of an RDBMS – moves the processing to the data rather than the data to the processing
- Important to understand the level of parallel processing and how a function is run, e.g., external to the RDBMS, in RDBMS protected memory, etc.

Several approaches to using in-database functions

- RDBMS built-in functions (arithmetic, string, date, statistical functions)
- Functions provided by a 3rd-party vendor, e.g., FuzzyLogix
- Open source functions, e.g., Apache Mahout, R
- Development options for creating user-defined functions (scripting language, Java, C++, SQL MapReduce, etc.)

Analytic RDBMSs: Vendor Examples

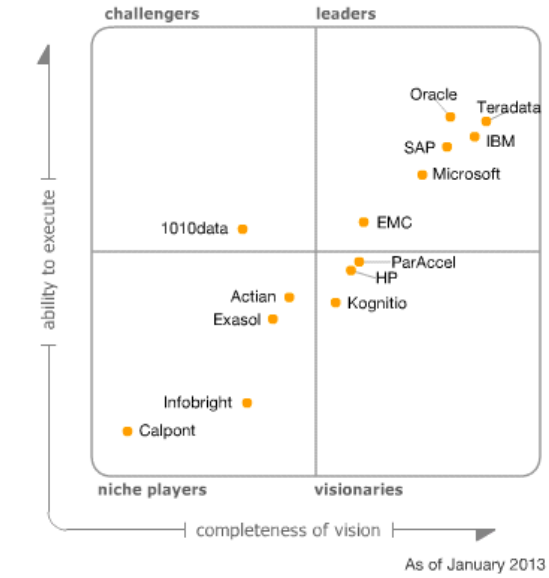
Traditional RDBMS Products

- IBM DB2: PureData for Operational Analytics
- IBM Informix: Ultimate Warehouse Edition (with Warehouse Accelerator)
- Microsoft SQL Server: Parallel Data Warehouse
- Oracle Database: Exadata
- SAP Sybase ASE and IQ
- Teradata Database: Active EDW, DW Appliance, Extreme Data Appliance, etc.

Other Solutions

- EMC Greenplum Database and Distributed Computing Appliance
- HP Vertica Analytics Platform
- IBM Netezza: PureData for Analytics, DB2 Analytics Accelerator
- Kognitio Analytical Platform
- Oracle Exalytics (Oracle TimesTen & Oracle Essbase)
- ParAccel Analytic Platform
- SAP HANA
- Teradata Aster Database: MapReduce Platform, Big Data Analytics Appliance
- InfoBright, MySQL, PostgreSQL, etc.

Data Warehouse DBMSs: Gartner 2013 MQ



Source: Gartner (January 2013)

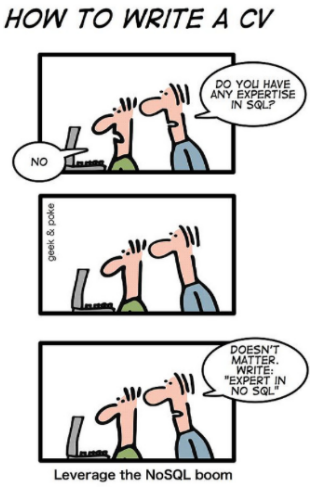
Optimized Platforms: Non-Relational Systems - 1

Several Internet companies developed their own non-relational (NoSQL or NewSQL) systems to support extreme data volumes

- Google example: Google file system, MapReduce, BigTable, BigQuery
- Main goal was the processing of large volumes of multi-structured data
- Several of these developments found their way into the open source community

Non-relational systems are not new, but modern versions are often open source

- Deployed on low-cost white-box hardware in a large-scale distributed computing environment
- Several types of systems & data stores
- Key industry focus area is Hadoop

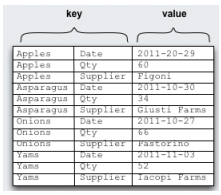


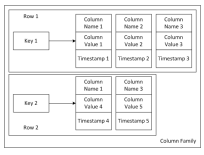
Optimized Platforms: Non-Relational Systems - 2

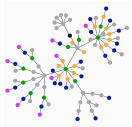
Many types of products, APIs and languages

Volume ← → *Complexity*

Key/Value Pair	Column Family	Document	Graph





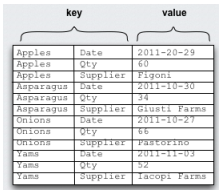
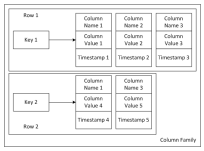
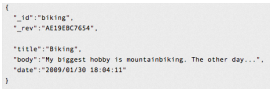
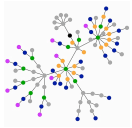


Can handle varieties of data and processing that are difficult to support using a traditional RDBMS

Optimized Platforms: Non-Relational Systems - 2

Many types of products, APIs and languages

Volume ← → *Complexity*

Key/Value Pair	Column Family	Document	Graph
			
Hadoop HDFS Amazon Dynamo Redis Riak Oracle NoSQL DB	Google BigTable HBase Cassandra MS Azure Tables	CouchDB MongoDB	Neo4J Google Freebase

"A framework for running applications on a large hardware cluster built of commodity hardware." *wiki.apache.org/hadoop/*

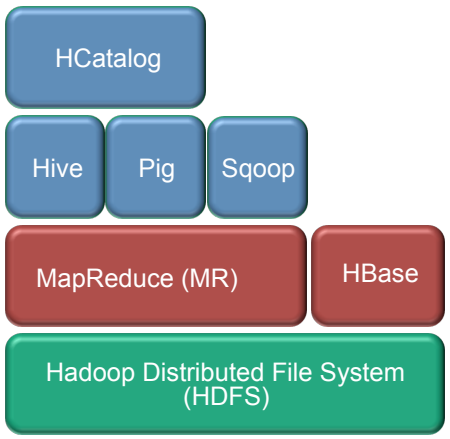
Provides a distributed file system (**HDFS**) that stores data across the nodes of the cluster to provide high performance

Includes a programming model called **MapReduce (MR)** where the processing is divided into small fragments of work that can be executed on any node in the cluster

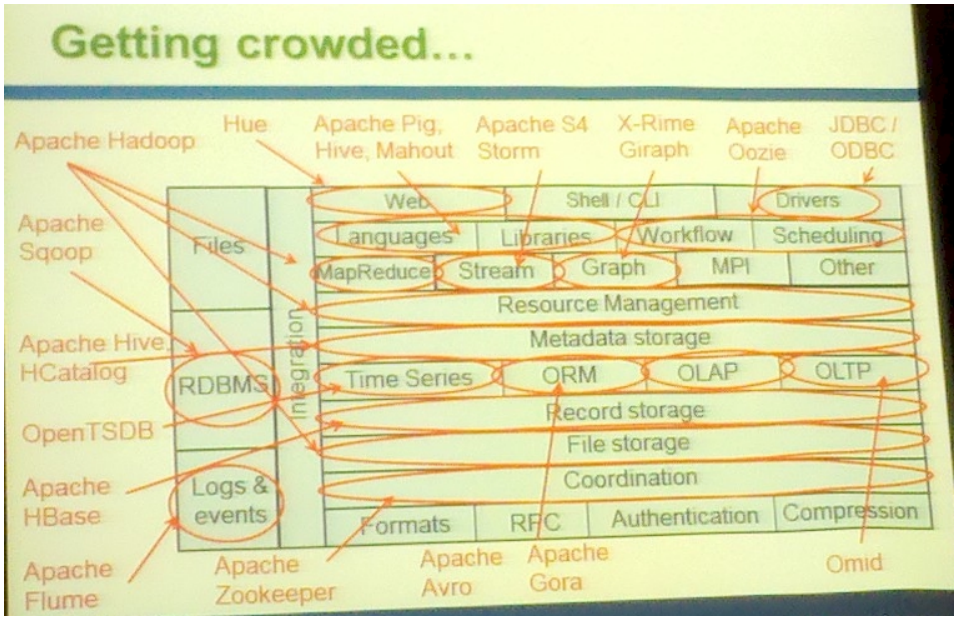
Hive and **Pig** are high-level languages for MR development

Related components include HBase, Sqoop, HCatalog, Flume, Storm, Mahout, Impala, etc.

Major distributions come from Apache, Cloudera, Hortonworks and MapR



Components



The Hadoop Ecosystem

Source: Hortonworks

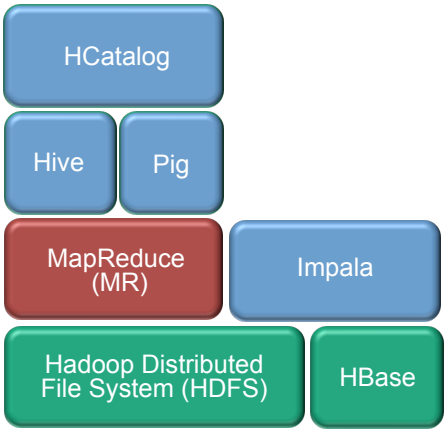
Data Management: Hadoop Option

Distributions

- Apache: Hadoop, HBase, Hive, Pig, Sqoop, Cassandra, Mahout
- Cloudera: Enterprise Free (CDH), Enterprise Core, RTD for HBase, RTQ for Impala
- HortonWorks Data Platform (includes Talend)
- MapR: M3 (free), M5, M7

Other solutions

- EMC Greenplum: HD, Isilon NAS for HD, HD Distributed Computing Appliance (DCA)*
- Hadapt Adaptive Analytical Platform*
- HP AppSystem for Apache Hadoop
- IBM InfoSphere BigInsights Basic and Enterprise Editions
- Microsoft HDInsight Server & Service
- Oracle Big Data Appliance
- SAS High-Performance Analytics Server
- Teradata Aster Big Data Analytics Appliance*



* *Hybrid Hadoop/RDBMS system*

Data Management: Relational vs Non-Relational

Given the number of options and a fast changing marketplace comparisons are difficult

Focus is on analytic RDBMSs versus Hadoop HDFS – DBMS versus file system, which is an “apples to oranges” comparison

At a high-level, an analytic DBMS is suited to complex interactive workloads and Hadoop HDFS for batch processing of multi-structured

From an Hadoop perspective, HBase is becoming more important, but lack of SQL support is an inhibitor

- Hive support for HBase is in development but it still uses batch Map/ Reduce (MR)
- Cloudera is developing Impala which supports Hive SQL syntax but eliminates MR – supports both HDFS and HBase

Workload suitability and performance are important, but development and administration effort, and tools support are also key considerations

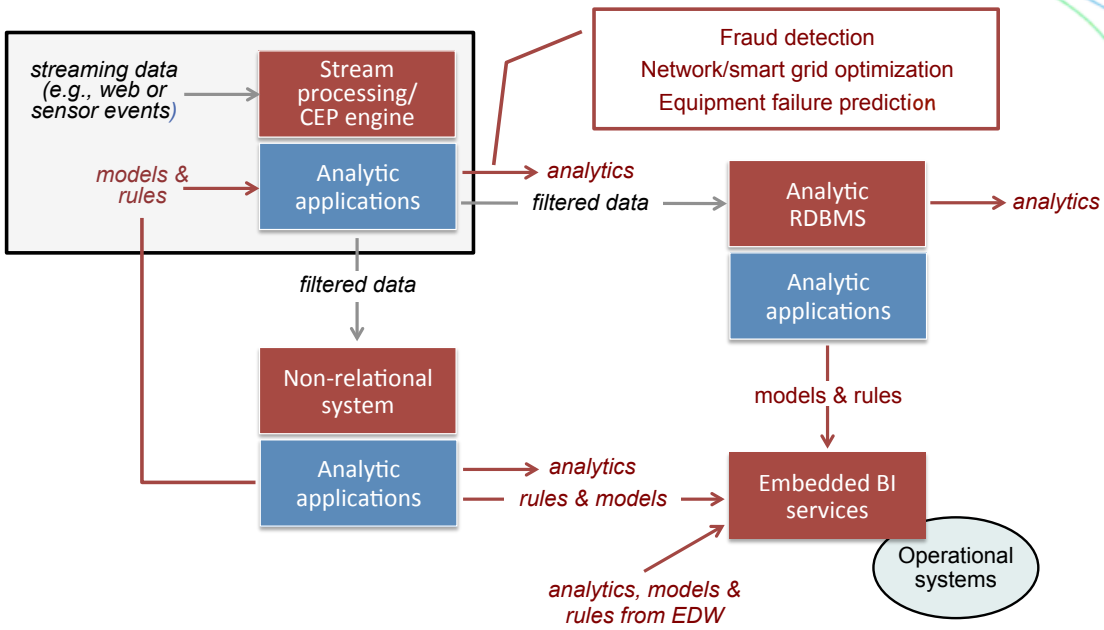
Data Management: Language Considerations

```
package org.myorg;

import java.io.IOException;
import java.util.*;

import org.apache.hadoop.fs.Path;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapred.*;
import org.apache.hadoop.util.*;

public class WordCount {

  public static class Map extends MapReduceBase implements Mapper<LongWritable, Text, Text, IntW
    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(LongWritable key, Text value, OutputCollector<Text, IntWritable> output, Rep
      String line = value.toString();
      StringTokenizer tokenizer = new StringTokenizer(line);
      while (tokenizer.hasMoreTokens()) {
        word.set(tokenizer.nextToken());
        output.collect(word, one);
      }
    }
  }

  public static class Reduce extends MapReduceBase implements Reducer<Text, IntWritable, Text, I
    public void reduce(Text key, Iterator<IntWritable> values, OutputCollector<Text, IntWritable>
      int sum = 0;
      while (values.hasNext()) {
        sum += values.next().get();
      }
      output.collect(key, new IntWritable(sum));
    }
  }

  public static void main(String[] args) throws Exception {
    JobConf conf = new JobConf(WordCount.class);
    conf.setJobName("wordcount");

    conf.setOutputKeyClass(Text.class);
    conf.setOutputValueClass(IntWritable.class);

    conf.setMapperClass(Map.class);
    conf.setCombinerClass(Reduce.class);
    conf.setReducerClass(Reduce.class);

    conf.setInputFormat(TextInputFormat.class);
    conf.setOutputFormat(TextOutputFormat.class);

    FileInputFormat.setInputPaths(conf, new Path(args[0]));
    FileOutputFormat.setOutputPath(conf, new Path(args[1]));

    JobClient.runJob(conf);
  }
}
```

```
#### define the explanatory variable with two levels:
#### 1=one or more parents smoke, 0=no parents smoke

parentsmoke=as.factor(c(1,0))

#### NOTE: if we do parentsmoke=c(1,0) R will treat this as
#### a numeric and not categorical variable

#### need to create a response vector so that it has counts for both "success" and "failure"

response<-cbind(yes=c(816,188),no=c(3203,1168))
response

#### fit the logistic regression model

smoke.logistic<-glm(response~parentsmoke, family=binomial(link=logit))

#### OUTPUT

smoke.logistic
summary(smoke.logistic)
anova(smoke.logistic)
```

```
--selfjoin.pig
-- For each stock, find all dividends that increased between two dates
divs1     = load 'NYSE_dividends' as (exchange:chararray, symbol:chararray,
                 date:chararray, dividends);
divs2     = load 'NYSE_dividends' as (exchange:chararray, symbol:chararray,
                 date:chararray, dividends);
jnd       = join divs1 by symbol, divs2 by symbol;
increased = filter jnd by divs1::date < divs2::date and
                 divs1::dividends < divs2::dividends;
```


Optimized Platforms: Stream Processing/CEP



Optimized Platforms: Cloud Option

Two approaches:

- Virtual machine image – may be provided by the user, a DBMS vendor, or a cloud vendor (e.g., Amazon)
- Database as a Service (DBaaS) offered by a public cloud vendor

Relational DBMS DBaaS

- Amazon Redshift (ParAccel) and Relational Database Service (MySQL, Oracle, MS SQL Server)
- Google Cloud SQL (MySQL) and BigQuery Service
- HP Cloud Relational Database for MySQL
- Kognitio Cloud
- Microsoft Azure SQL Database
- Oracle Cloud

Non-Relational DBMS DBaaS

- Amazon DynamoDB, SimpleDB, ElastiCache
- Microsoft Azure Tables and Blob storage
- SalesForce database.com

Data Management in the Cloud Example

NETFLIX

Netflix's on-premises IT infrastructure was too fragile and the traditional operations model didn't respond fast enough to business needs

- Rapidly growing and highly variable data-center requirements
- Inability to automate data-center operations

As a result the company migrated from an on-premises environment to an Amazon Web Services infrastructure

It evaluated how the new environment would affect the IT infrastructure and redesigned applications as appropriate

Spreads its processing across many different Amazon data centers and regions to enhance reliability and availability

Different service environments are randomly taken offline to confirm that the environment can continue operating in the face of a resource failure

Conclusion: Netflix changed its approach because it recognized that the future of its business required a different way of doing things

Source: Netflix (slideshare.net/adrianco/netflix-in-the-cloud-at-sv-forum) and CIO Magazine

Summary: Big Data Benefits

Traditional Decision-Making Environment *(determine and analyze current business situation)*	**Big Data Extensions** *(provide more complete answers, predict future business situations, investigate new business opportunities)*
Integrated data sources	Virtualized and blended data sources
Structured data	Multi-structured data
Aggregated and detailed data (with limits)	Large volumes of detailed data (no limits)
Relational EDW with at rest data Dimensional cubes/marts with at rest data	Non-relational stores with at rest data Streaming/CEP systems with in motion data
One-size fits all data management	Flexible & optimized data management
Reporting and OLAP	Advanced analytic functions & predictive models
Dashboards and scorecards	Sophisticated visualization of large result sets
Structured navigation (drill, slice/dice)	Flexible exploration of large result sets
Humans interpret results, patterns and trends	Sophisticated trend and pattern analysis
Manual analyses, decisions and actions	Analytics & model-driven recommendations & actions

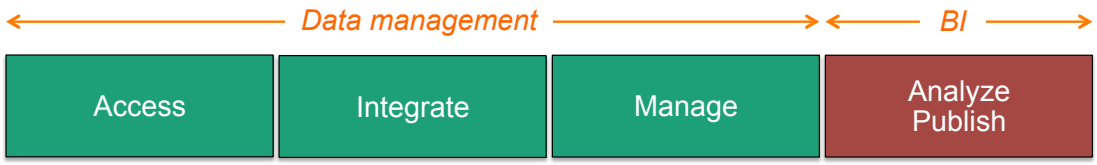
Choosing the Right Solution

Organizations will likely use multiple analytic solutions and data management systems – the challenge is deciding which to use when and how to interconnect the systems

Use Cases and Application Examples

Use Case	Application Example
Real-Time Monitoring & Analytics	In-line fraud detection to reduce financial losses caused by stolen credit cards
Near-Real-Time Analytics	Next best customer offer to the channel to increase customer satisfaction & reduce churn
Data Integration Hub	Collect and manage all sales-related detailed data (POS, web, supply chain) for down stream analysis
Analytics Accelerator	Offload & boost the performance of selected financial analyses to increase satisfaction/retention of key clients
New LOB Analytic Application	Manage & monitor spot buying on web advertising exchanges
Investigative Computing Platform	Evaluate the effectiveness of different social computing channels

Software Selection: Some Key Options



Data access
- languages & APIs
- connectors
- data virtualization
- search

Data integration
- modeling
- profiling
- cleansing
- transforming
- loading

Data manager
- relational
- non-relational
- stream processing/ CEP
- in-memory
- cloud-based

Reporting
- batch
- interactive

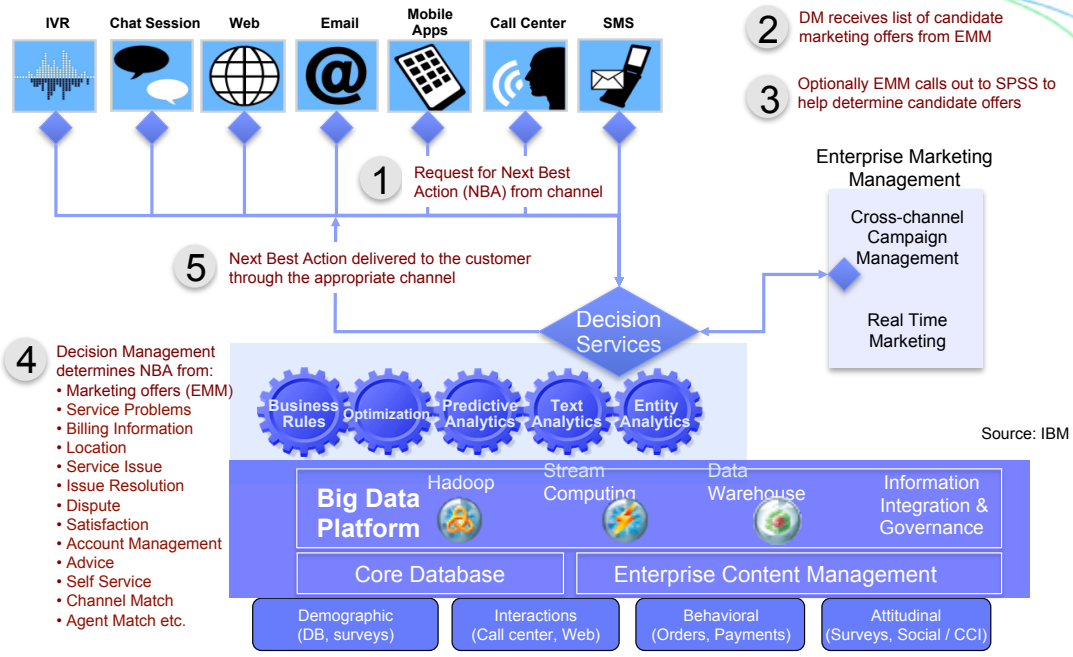
Multidimensional analysis
- ROLAP
- MOLAP

Advanced analytics
- data mining
- machine learning
- advanced functions
- data exploration
- data visualization
- operational intelligence

Use Cases and Technologies

Use Case	Stream Processing/ CEP system	Embedded BI Services	Enterprise Data Warehouse	Analytic Relational DBMS	Hadoop System
Real-Time Monitoring & Analytics	✔	✔			
Near-Real-Time Analytics			✔	✔	✔
Data Integration Hub					✔
Analytics Accelerator				✔	✔
New LOB Analytic Application				✔	✔
Investigative Computing Platform				✔	✔

Example Telco Provider: Real-Time Embedded BI



Source: IBM

Example SEARS HOLDINGS MetaScale

$43 billion retail organization with over 4,000 stores (Sears and Kmart)

Numerous legacy systems with applications written in COBOL and Assembler (over 100 million lines)

Running out of capacity but at the current cost of $3K-$7K per MIP per year another solution had to be found

Requirements:

- Cost effectively manage increasing data volume
- Reduce the number of data warehouses and ETL jobs
- Reduce analytical processing times and provide intra-day analytics
- Capture and store all detailed transaction data (POS data, web clicks, supply chain events, etc.) for analysis
- Limit changes to existing user interfaces

Primary source: Presentation by Dr. Phillip Shelley (CTO Sears Holdings and CEO MetaScale) at the Hadoop Summit, June 2012

Example SEARS HOLDINGS MetaScale

Solution: Hadoop Data Hub and Analytics Accelerator

Enhanced pricing application

- Issue: only about 10% of the sales data is in the EDW; pricing models were taking taking 8 weeks to setup and run
- Hadoop MapReduce solution analyzes price elasticity based on 100% of the sales data
- Pricing models can now be run weekly (or daily if required)

Improved customized offers to loyal customers

- Issue: existing system was not scalable; only a small subset of the data could be analyzed
- Replaced 6,000 COBOL application with 400 lines of Pig and Java UDFs; implemented in 6 weeks
- Application can now be run multiple times per day per store per line item per customer – reduces impact of competitors such as Amazon

Example SEARS HOLDINGS MetaScale

Solution: Hadoop Data Hub and Analytics Accelerator /cont.

Reduce time to run batch BI applications

- Existing mainframe window of 3.5 hours was becoming insufficient to run 64 batch pricing jobs against 500 million rows of data
- Batch jobs were rewritten in Pig and run on data FTP'd to Hadoop from the mainframe and results FTP'd back to the mainframe
- Jobs run 100% faster (run in 8 minutes; FTP is the main overhead)

Reduce time to run batch and interactive BI applications

- Existing batch and interactive BI applications were taking too long to run and could handle only a subset of the data
- Data from over 50 sources now stored and analyzed on Hadoop
- Datameer is used for analytics
- Pig is used for ETL and for creating output for Excel

Example SEARS HOLDINGS MetaScale

Conclusions

Pleased with Hadoop's ability to run enterprise workloads – enables detailed data to be stored and analytics to be run that were not previously possible

Hadoop is only one component of the BI/DW ecosystem and strategy

Hadoop requires significant education and implementation effort and is lacking tools for enterprise integration

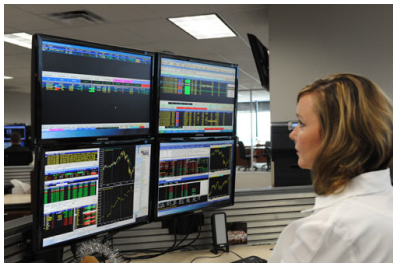
New Sears subsidiary (MetaScale) formed to help enterprises integrate existing systems with Hadoop because "75% of CEOs and CIOs don't even know what Hadoop is"

Example: International Bank Trading Desk

Solution: Analytic RDBMS as an Analytics Accelerator

This international bank offers a wide range of services to its over 40 million customers

One of the bank's trading desks uses the appliance for an analytic solution that handles the ad hoc analysis of billions of rows of detailed loan/bond data



Month-end loading was reduced from days to 2 hours

Key customer queries were reduced from 3-4 days to about 7 minutes

Appliance was treated as a black box by the IT group for compliance reasons

Example

MediaMath

Solution: Analytic RDBMS for New LOB Application

MediaMath is a leader in the billion dollar display advertising business

Provides a platform called TerminalOne that enables ad agencies and large-scale advertisers to identify, bid on, buy, and optimize ad impressions

Automatically matches each impression in real time with ads that are meaningful and relevant to users

Analyzes upwards of 15 billion ad impressions a day and calculates the fair value of more than 50,000 ads/sec

Source: LUMA Partners

Example @WalmartLabs

Solution: Hadoop System for Investigative Computing

Phase 1 of the project was to use Hadoop and MapReduce to consolidate web site data (from 10 sites) for e-commerce analysis

Phase 2 involved providing users with Hive access to the web (and social) data for investigative purposes

- Demand for access and use of the system grew dramatically
- People forgot this was an experimental system!
- Requirements grew: larger cluster, resource management, SLAs, real-time data, metadata catalog
- Extended Hadoop to support a data hub containing 10 years of detailed data and reduce data stored in existing DW systems

Primary source: Presentation by Stephen O'Sullivan and Jeremy King at the Hadoop Summit, June 2012

Barriers to Success

Educating IT and the business about the use cases and business benefits of big data

Lack of skills for enabling data science and investigative computing projects

Understanding and selecting the components that are required to build and support a big data analytics ecosystem

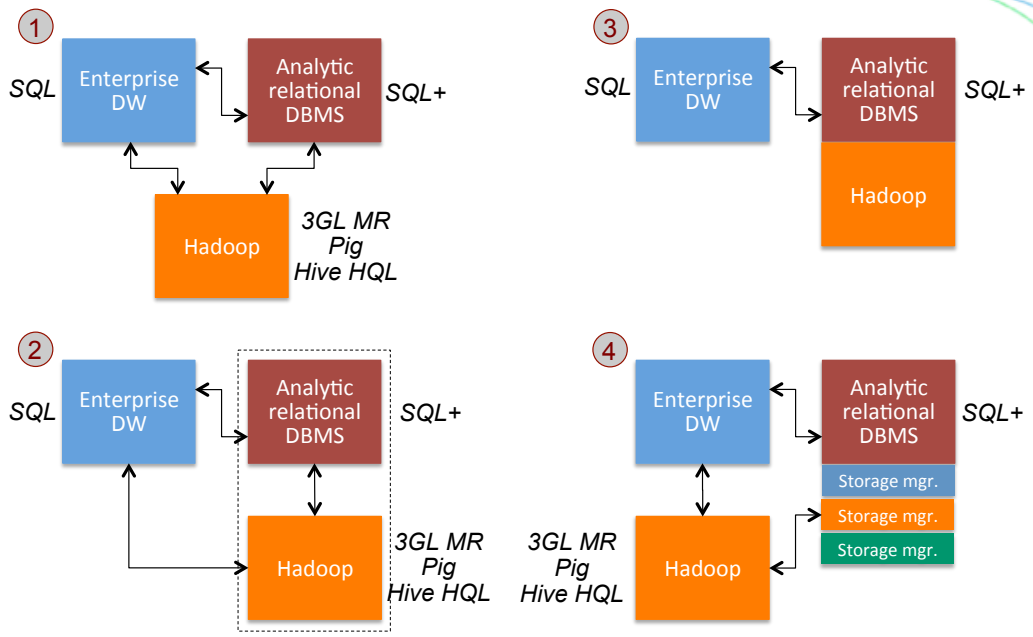
The immaturity of new non-relational systems and the level of IT development and administration resources and skills required for supporting them

The amount of data integration and level of data movement required in a big data environment

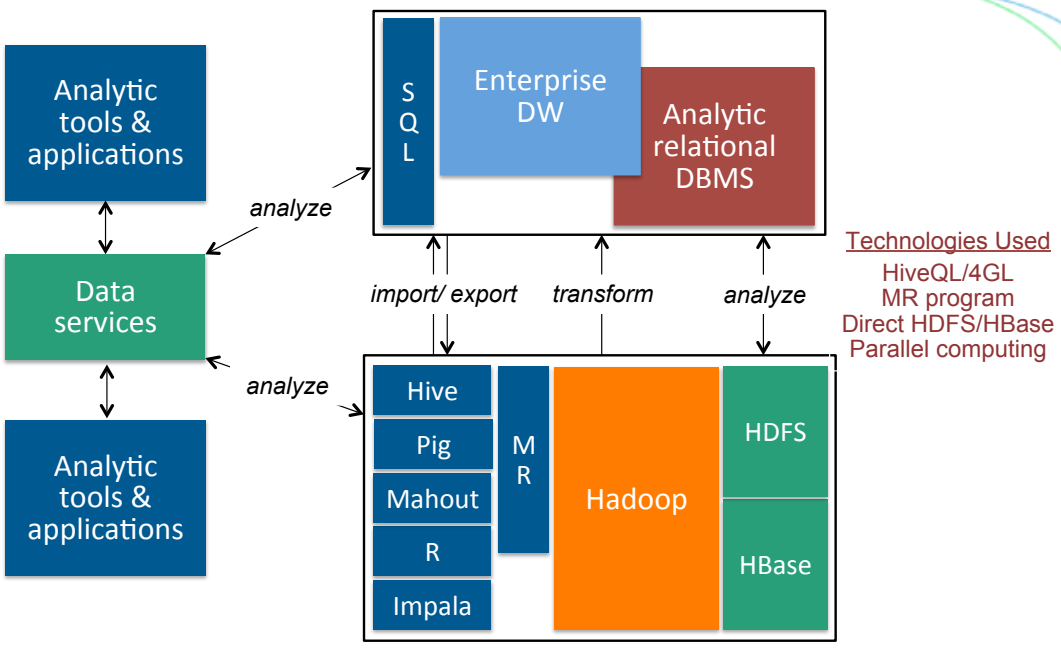
Developing data governance and data retention approaches to support the big data environment

Providing business users with a single and seamless user interface

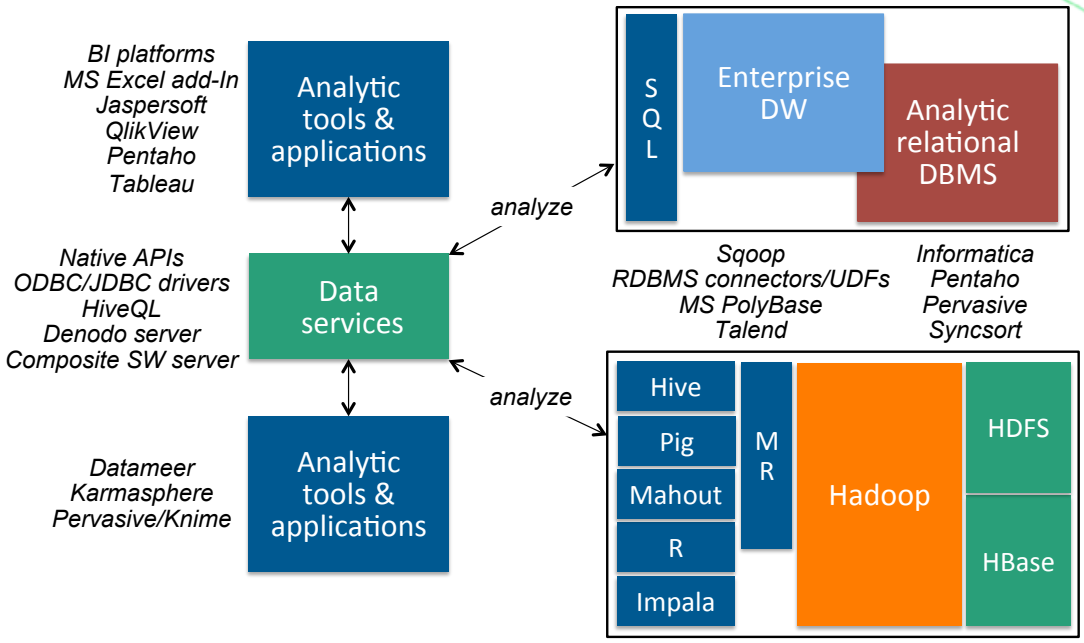
Integration Example: RDBMS + Hadoop



Hybrid Option 1: Techniques



Hybrid Option 1: Product Examples



Conclusions

Big data represents a new and evolving analytics ecosystem – it is not one technology, but a set of overlapping technologies

Big data represents a new wave of analytic value for organizations

Big data initiatives at present are primarily LOB use-case driven

True value is gained from a hybrid of existing and new data systems

One size does not fit all and it is becoming impossible for any given vendor to satisfy all requirements

Integration with existing enterprise systems will be a key vendor differentiator

The vendors that win will be those that can best tackle cost, product and data integration, and/or advanced analytics needs

It is important for organizations to design an extended (or logical) data warehouse and advanced analytics ecosystem in order to be able to evolve and grow the BI environment