Big Data – The end of Data Warehousing?

Hermann Bär
Oracle USA
Redwood Shores, CA

Schlüsselworte
Big data, data warehousing, advanced analytics, Hadoop, unstructured data

Introduction
If there was an "Unwort" for the IT industry, the winner of last year would be undoubtedly 'Big Data". Who has not heard about this new term, seen any ad, or had any vendor talking about it. Even Wikipedia has its own page already, and searching for this terms shows about 40 times more hits than searching for "data warehousing".
But what does it really mean? Is it hype, or is there a real story behind it? What will happen to data warehousing, an area we all worked on for the last decades? This article will shed some light on the simple yet powerful story behind this new term and what it means for data warehousing.

What is Big Data?
Big Data is actually very simple. For decades, companies have been making business decisions based on transactional data stored in relational databases. Beyond that critical data, however, is a potential treasure trove of non-traditional, less structured data: weblogs, social media, email, sensors, and photographs that can be mined for useful information. Decreases in the cost of both storage and compute power have made it feasible to collect this data - which would have been thrown away only a few years ago. As a result, more and more companies are looking to include non-traditional yet potentially very valuable data with their traditional enterprise data in their business intelligence analysis. Big Data is the term that is used to describe the platform and concept to derive business value from these non-traditional data sets in a much a deeper and broader sense than we ever thought about in the classical (structured) relational data warehousing world.

Big data is about capturing and organizing a wide variety of data types from different sources, and to be able to easily analyze it within the context of all your enterprise data to find new insights and capitalize on hidden relationships.

So what data types are we talking about? Big data typically refers to the following types of data:

- Traditional enterprise data – includes customer information from CRM systems, transactional ERP data, web store transactions, general ledger data. This is the data we mainly focus on with today's data warehousing environments

- Machine-generated /sensor data – includes Call Detail Records ("CDR"), weblogs, smart meters, manufacturing sensors, equipment logs (often referred to as digital exhaust), trading systems data.

- Social data – includes customer feedback streams, micro-blogging sites like Twitter, social media platforms like Facebook

As you can see right away, big data is not a replacement for data warehousing, but goes just way beyond the data sets we are used to consider, which leads to a potential data explosion. The McKinsey Global Institute estimates that data volume is growing 40% per year, and will grow 44x between 2009 and 2020. But while it's often the most visible parameter, volume of data is not the only characteristic that matters. In fact, there are four key characteristics that define big data:

- **Volume.** Machine-generated data is produced in much larger quantities than non-traditional data. For instance, a single jet engine can generate 10TB of data in 30 minutes. With more than 25,000 airline flights per day, the daily volume of just this single data source runs into the Petabytes. Smart meters and heavy industrial equipment like oil refineries and drilling rigs generate similar data volumes, compounding the problem.
- **Velocity.** Social media data streams – while not as massive as machine-generated data – produce a large influx of opinions and relationships valuable to customer relationship management. Even at 140 characters per tweet, the high velocity (or frequency) of Twitter data ensures large volumes (over 8 TB per day).
- **Variety.** Traditional data formats tend to be relatively well described and change slowly. In contrast, non-traditional data formats exhibit a dizzying rate of change. As new services are added, new sensors deployed, or new marketing campaigns executed, new data types are needed to capture the resultant information.
- **Value.** The economic value of different data varies significantly. Typically there is good information hidden amongst a larger body of non-traditional data; the challenge is identifying what is valuable and then transforming and extracting that data for analysis.

To make the most of big data, enterprises must evolve their IT infrastructures to handle the rapid rate of delivery of extreme volumes of data, with varying data types, which can then be integrated with an organization's other enterprise data to be analyzed

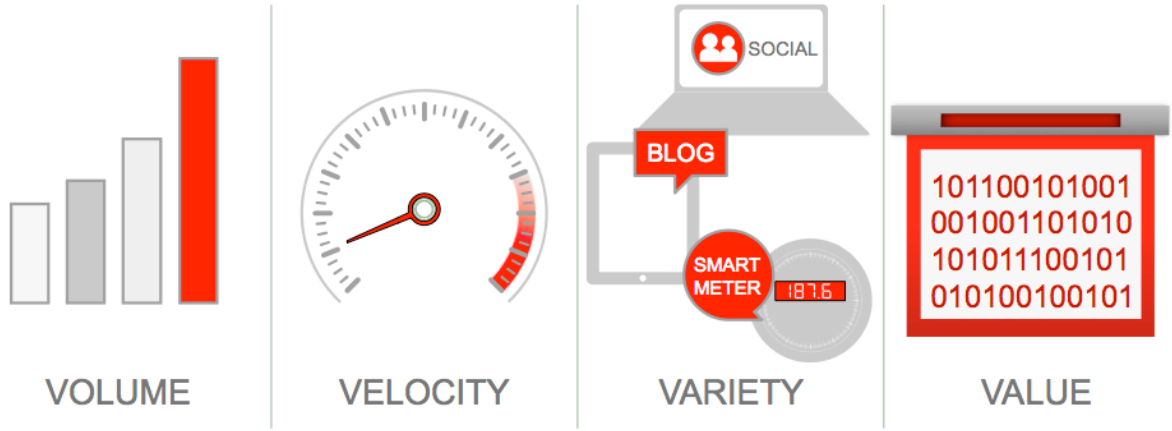

Abb. 1: What defines Big Data?

When big data is distilled and analyzed in combination with traditional enterprise data, enterprises can develop a more thorough and insightful understanding of their business, which can lead to enhanced productivity, a stronger competitive position and greater innovation – all of which can have a significant impact on the bottom line.

Oracle's Big Data Platform – the end-to-end solution

As with data warehousing, web stores or any IT platform, an infrastructure for big data has unique requirements. In considering all the components of a big data platform, it is important to remember that the end goal is to easily integrate your big data with your enterprise data - your classical data warehousing target - to allow you to conduct deep analytics on the combined data set.

Acquire Big Data

The acquisition phase is one of the major changes in infrastructure from the days before big data. Because big data refers to data streams of higher velocity and higher variety, the infrastructure required to support the acquisition of big data must deliver low, predictable latency in both capturing data and in executing short, simple queries; be able to handle very high transaction volumes, often in a distributed environment; and support flexible, dynamic data structures.

NoSQL databases are frequently used to acquire and store big data. They are well suited for dynamic data structures and are highly scalable. The data stored in a NoSQL database is typically of a high variety because the systems are intended to simply capture all data without categorizing and parsing the data. Unlike in a classical data warehousing system where data underlies more or less strict and well-defined schema, data collected through NoSQL has no schema defined at this stage. This is the data that was often ignored and not considered for analysis. Collecting this additional data and making it available is a complementary (evolutionary) addition to your data warehouse.

Oracle NoSQL Database is a distributed, highly scalable, key-value database based on Oracle Berkeley DB. It delivers a general purpose, enterprise class key value store adding an intelligent driver on top of distributed Berkeley DB. The primary use cases for Oracle NoSQL Database are low latency data capture and fast querying of that data, typically by key lookup. Oracle NoSQL Database comes with an easy to use Java API and a management framework.

Organize Big Data

In classical data warehousing terms, organizing data is called data integration. Because there is such a high volume of big data, there is a tendency to organize data at its original storage location, thus saving both time and money by not moving around large volumes of data. The infrastructure required for organizing big data must be able to process and manipulate data in the original storage location; support very high throughput (often in batch) to deal with large data processing steps; and handle a large variety of data formats, from unstructured to structured.

Apache Hadoop is a new technology that allows large data volumes to be organized and processed while keeping the data on the original data storage cluster. Hadoop Distributed File System (HDFS) is the long-term storage system for web logs for example. These web logs are turned into browsing behavior (sessions) by running MapReduce programs on the cluster and generating aggregated results on the same cluster. These aggregated results are then loaded into a Relational DBMS system. Unlike today's data warehousing environments, this is an additional enriching step in the Big Data processing.

Oracle is uniquely qualified to combine everything needed to meet the big data challenge – including software and hardware – into one engineered system. The **Oracle Big Data Appliance** is an

engineered system that combines optimized hardware with the most comprehensive software stack. Oracle Big Data Appliance contains Cloudera's Distribution including Apache Hadoop (CDH) and Cloudera Manager. CDH is the #1 Apache Hadoop-based distribution in commercial and non-commercial environments. CDH consists of 100% open source Apache Hadoop plus the comprehensive set of open source software components needed to use Hadoop.

In addition, Oracle Big Data Appliance features specialized solutions developed by Oracle to deliver a complete, easy-to-deploy solution for acquiring, organizing and loading big data into Oracle Database 11g, incl.

- **Oracle Loader for Hadoop (OLH)** enables users to use Hadoop MapReduce processing to create optimized data sets for efficient loading and analysis in Oracle Database 11g.
- **Oracle Direct Connector for Hadoop Distributed File System** (HDFS) is a high speed connector for accessing data on HDFS directly from Oracle Database.
- **Oracle Data Integrator Application Adapter for Hadoop** to simplify data integration from Hadoop and an Oracle Database through Oracle Data integrator's easy to use interface.

Note that all of the above-mentioned technology products are in addition to the classical, well-known tools and technologies of today's data warehouses. Those work in conjunction with your data warehousing environment, ensuring an end-to-end solution to easily integrate your big data with your enterprise data to allow you to conduct deep analytics on the combined data set.

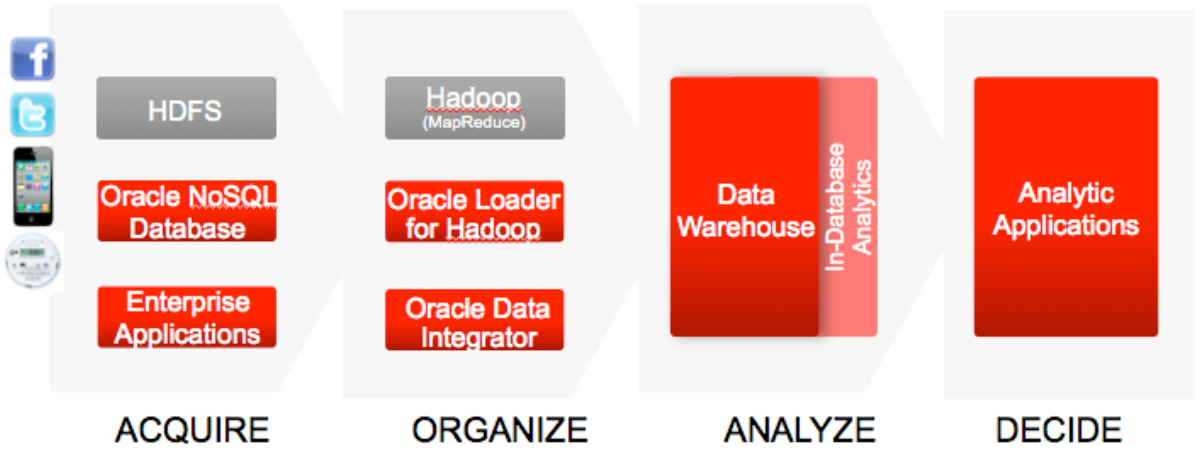

Abb. 2: Oracle's end-to-end Big Data and Data Warehousing platform

Analyze Big Data

Analyzing data to gain business insight is not a new concept; that's ultimately what data warehousing is about. However, the infrastructure required for analyzing big data must be able to support deeper analytics such as statistical analysis and data mining, on a wider variety of data types stored in diverse systems; scale to extreme data volumes; deliver faster response times driven by changes in behavior; and automate decisions based on analytical models. Most importantly, the infrastructure must be able to integrate analysis on the combination of big data and traditional enterprise data. New insight comes not just from analyzing new data, but from analyzing it within the context of the old to provide new perspectives on old problems. With Oracle Big Data Appliance and the storage of data in non-traditional form – such as in HDFS – provides yet-another data set to gain insight from. In addition to

the before-mentioned software solutions provided on the Big Data Appliance, one additional software component is part of the solution that enables data insight right on the Big Data Appliance: **Oracle R Connector for Hadoop** is an R package that provides transparent access to Hadoop and to data stored in HDFS. R Connector for Hadoop can optionally be used together with the Oracle Advanced Analytics Option for Oracle Database. The Oracle Advanced Analytics Option enables R users to transparently work with database resident data without having to learn SQL or database concepts but with R computations executing directly in-database.

While a lot of analysis and data processing will take place on the Oracle Big Data Appliance, most if not all of the adhoc-analysis and data exploration will take place after the data is well-defined and in a structured from, in a relational system. Once data has been loaded from Oracle Big Data Appliance into Oracle Database or Oracle Exadata, end users can use one of the following easy-to-use tools for in-database, advanced analytics:

- **Oracle R Enterprise** – Oracle's version of the widely used Project R statistical environment enables statisticians to use R on very large data sets without any modifications to the end user experience. Examples of R usage include predicting airline delays at a particular airports and the submission of clinical trial analysis and results.

- **In-Database Data Mining** – the ability to create complex models and deploy these on very large data volumes to drive predictive analytics. End-users can leverage the results of these predictive models in their BI tools without the need to know how to build the models. For example, regression models can be used to predict customer age based on purchasing behavior and demographic data.

- **In-Database Text Mining** – the ability to mine text from micro blogs, CRM system comment fields and review sites combining Oracle Text and Oracle Data Mining. An example of text mining is sentiment analysis based on comments. Sentiment analysis tries to show how customers feel about certain companies, products or activities.

- **In-Database Semantic Analysis** – the ability to create graphs and connections between various data points and data sets. Semantic analysis creates, for example, networks of relationships determining the value of a customer's circle of friends. When looking at customer churn customer value is based on the value of his network, rather than on just the value of the customer.

- **In-Database Spatial** – the ability to add a spatial dimension to data and show data plotted on a map. This ability enables end users to understand geospatial relationships and trends much more efficiently. For example, spatial data can visualize a network of people and their geographical proximity. Customers who are in close proximity can readily influence each other's purchasing behavior, an opportunity which can be easily missed if spatial visualization is left out.

- **In-Database MapReduce** – the ability to write procedural logic and seamlessly leverage Oracle Database parallel execution. In-database MapReduce allows data scientists to create high-performance routines with complex logic. In-database MapReduce can be exposed via SQL. Examples of leveraging in-database MapReduce are sessionization of weblogs or organization of Call Details Records (CDRs).

Every one of the analytical components in Oracle Database is valuable. Combining these components creates even more value to the business. Leveraging SQL or a BI Tool to expose the results of these

analytics to end users gives an organization an edge over others who do not leverage the full potential of analytics in Oracle Database. The somewhat funny – or ironic part for that matter – is that every single of the above-mentioned analytical capabilities (with the exclusion of Oracle R Enterprise) were available in the Oracle Database for many years, but it needed 'Big Data' to make people aware of it.

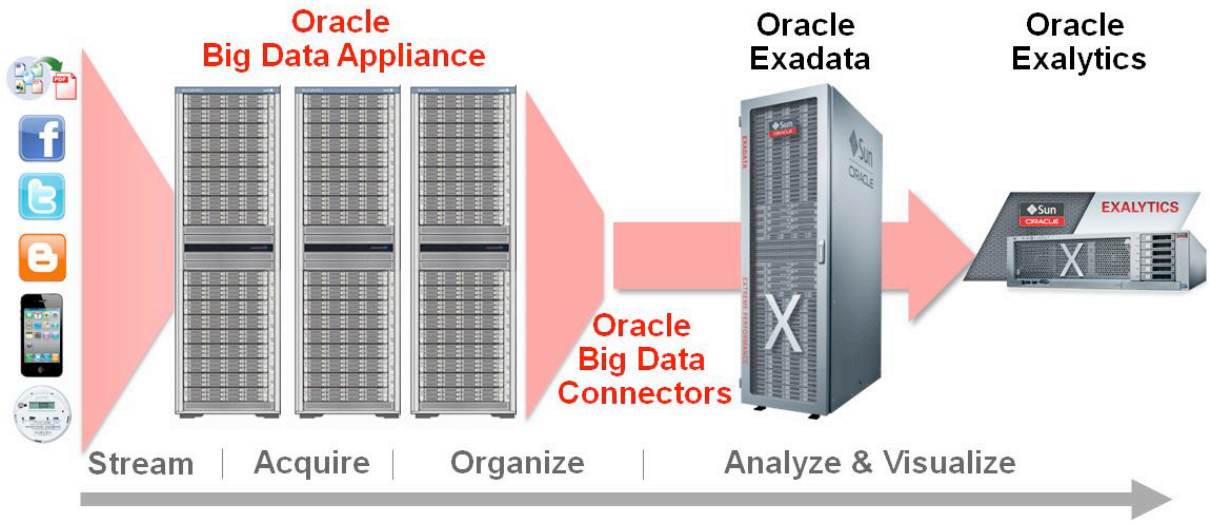

Abb. 3: Oracle's Engineered Systems for Big Data and Data Warehousing platform

Oracle Big Data Appliance – the newest member in Oracle's family of Engineered Systems – represents the 'missing piece' in providing and end-to-end solution – hardware and software – for Big Data and Data Warehousing. Oracle Big Data Appliance, in conjunction with Oracle Exadata Database Machine and the new Oracle Exalytics Business Intelligence Machine, delivers everything customers need to acquire, organize, analyze and maximize the value of Big Data within their enterprise.

Big Data – more than a platform and technology

Data warehouses built using relational technology like Oracle Database and Oracle Exadata provide insight into how the business is running and how to improve it. Big data in some ways is a natural evolution of that. What's different in big data is the new data sources; new types of data not previously captured, stored, or analyzed. Very large volumes of data are acquired very rapidly, and may not be neatly structured, which can make storing and analyzing it a challenge.

What is also different is the mindset of what to do with the data and how to react on the insight you are gaining: with the sheer volumes of data being generated you have to conquer the (somewhat revolutionary) task of implementing – and trusting – automated decisions. Also, all the new technologies make it more cost effective and easier to store any kind of data, whether its values is known from the get-go or not. A new breed of business analysts- so-called data scientists – will add curiosity about the data and its hidden gems to the mix. The not-so-well-defined first step of data exploration and analysis will change the way of how businesses deal with all their data to really improve the bottom line and to make an impact on overall profitability.

To derive real business value from big data, you not only need the right tools to capture and organize a wide variety of data types from different sources; you not only must be able to easily analyze it within the context of all your enterprise data. But it's still needed: by using the Oracle Big Data Appliance and Oracle Big Data Connectors in conjunction with Oracle Exadata, enterprises can acquire, organize and analyze all their enterprise data – including structured and unstructured – to make the most informed decisions.

Kontaktadresse:
Hermann Bär
Oracle USA
500 Oracle Parkway, MS 4op742
Redwood Shores, CA 94065
USA

Telefon: +1 650-506-6833
Fax: +1 650-506-6833
E-Mail hermann.baer@oracle.com
Internet: www.oracle.com