



WinterCorp

www.wintercorp.com

SPONSORED
RESEARCH
PROGRAM

BIG DATA: BUSINESS OPPORTUNITIES, REQUIREMENTS AND ORACLE'S APPROACH

RICHARD WINTER

December 2011

SUMMARY

NEW SOURCES OF DATA and distinctive types of data analysis are enabling a new set of business opportunities. Dubbed "Big Data," this new arena is characterized primarily by large and rapidly growing data volumes; varied data structures; and, new or newly intensified data analysis requirements.

As enterprises harness big data, they are discovering opportunities to better understand and predict the interests and behavior of their customers, especially in connection with e-commerce and social networking. In engineering and manufacturing, companies are finding new opportunities to predict maintenance problems, enhance manufacturing quality and manage costs via big data. And in healthcare, there are new opportunities to predict and/or more rapidly react to critical clinical events, yielding better health outcomes and more effective cost management.

This WinterCorp Executive Report provides an introduction to big data. The Report describes what big data really is and why big data is enabling new business opportunities. The Report then reviews requirements and describes the approach adopted by Oracle Corporation to provide its customers with enterprise class products for big data.

The development of this report by the independent data management expert WinterCorp has been sponsored by Oracle. A sidebar on page 3 describes the research methodology used.

WHAT IS "BIG DATA"?

The term "Big Data" first surfaced about five years ago in connection with the data management and analytical needs of extremely large internet businesses such as Google and Yahoo. These companies needed an economical and highly scalable architecture for the analysis of certain files, such as their logs of web activity. Their data and analytical requirements did not fit closely with the relational database products used in most commercial systems. And, because their data volumes were so extremely large and fast growing – between ten and one hundred times the size of the largest relational databases — they decided to create their own solutions.

As these techniques spread, principally via the use of open source licenses, it turned out that similar requirements existed in many enterprises.

A WINTER CORPORATION EXECUTIVE REPORT

About WinterCorp



WinterCorp is an independent consulting firm that specializes in the scalability of terabyte- and petabyte-scale data management systems and solutions, providing services to assist at every stage of the lifecycle.

Since our founding in 1992, we have architected solutions to some of the largest scale and most demanding requirements, world wide.

Our consulting services help enterprises define their requirements; architect their solutions; select their platforms; engineer their implementations; and, manage their growth to optimize business value.

Our seminars and structured workshops help client teams establish a shared foundation of knowledge and move forward to meet their challenges in database scalability, performance and availability.

Our expertise encompasses key products in commercial data warehousing as well as open source platforms such as Hadoop and other emerging technologies.

With decades of experience in large scale data management implementations, and in-depth knowledge of database products and technology, we deliver unmatched insight into the issues that impede performance and the technologies and solutions that enable business success.



WinterCorp

245 FIRST STREET,
SUITE 1800
CAMBRIDGE MA 02145
617-695-1800



The usual big data characteristics are:

- 1. Volume:** there is a lot of data to be analyzed and/or the analysis is extremely intense. Either way, a lot of hardware is needed;
- 2. Variety:** the data is not organized into simple, regular patterns as in a table; rather text, images and highly varied structures—or structures unknown in advance—are typical;
- 3. Velocity:** the data comes into the data management system rapidly and often requires quick analysis or decision making.

In addition, Oracle offers the idea that big data often has **low value density**. That is, most of the data in its originally received form may be of low value. Analytical processing may be required in order to transform the data into usable form or derive the usable portion.

For example, it may be impossible to derive much business value from the logs of a website prior to “sessionization.” That is, the logs must first be organized into segments, each of which describes the actions of one user making one website visit. Only after sessionization can the user behavior be analyzed for patterns meaningful for many types of business decision making.

Note that “low value density” does *not* mean “low value.” The opposite is true: you may have to analyze a lot of data to find what you want, but you do it because there is **very high value** to be found.

Taken together, these four characteristics—volume, variety, velocity and low value density—are viewed by Oracle as the defining characteristics of big data.

BIG DATA EXAMPLES

E-Commerce and Consumer Marketing. An example of a big data problem is storing, managing and acting on the sentiment expressed by consumers concerning a brand such as Toyota automobiles or children’s Tylenol. In the course of a day or a week, consumers can have millions of electronic interactions concerning a brand. While some of this is private, much of the interaction may be directly with the company via its websites, call centers, and stores; via email to the company; or, via open social media, such as Twitter or Facebook.

When something important is happening – for example, consumers reacting to an economic, safety or environmental incident—or to changes in price, product, supply or competition—the company needs to quickly understand how best to react to the situation. A slow reaction can greatly increase the cost or difficulty of solving a problem—or cause an important opportunity to be missed.

This has long been a challenge, but because of the volume, variety and speed of electronic communication today, the challenge is now often much greater and can require even faster and more decisive action. But, how to be decisive in the face of a tidal wave of unstructured consumer communications? How to understand what people are really saying, what they are reacting to, what the problem really is, what the solutions are?

The expectedly large and negative consumer reaction to new fees announced by large US banks in late 2011 is a case in point. Hundreds of thousands of accounts were lost in a few days—the equivalent of a year of losses under typical market conditions—by the time consumer sentiment was fully understood and corrective action could be taken.

A WINTER CORPORATION EXECUTIVE REPORT

Engineering and Manufacturing. Electronic communication by consumers is only one of many types of data which are either new entirely or are now present in much larger volumes than just a few years ago.

For example, the many sensors and data capture devices in consumer and industrial products—in engineering and testing processes—and, in manufacturing processes—now collectively generate truly enormous volumes of data. As a result of the declining cost of automatically collecting and delivering data, the physical world—and the daily activities of virtually every enterprise and its customers—are being measured in ways we could scarcely have imagined a few years ago.

There are hundreds of sensors in every car manufactured today, each generating data many times each second about some component or system of the vehicle when it is in use (e.g., brakes, accelerator, steering, transmission, lubrication, cooling, engine). This data is stored in the car; captured when the car is in for maintenance; and, then analyzed to diagnose immediate problems, but also for insight into engineering, manufacturing and maintenance issues.

Many commercial fleets feature devices in every vehicle to record every significant action regarding safety, efficiency and other factors—often generating many records per second per vehicle—for wireless transmission and immediate collection and analysis. Recall that there are commercial fleets of all kinds: cars, trucks, helicopters, construction equipment, road equipment and a vast array of specialty vehicles. The more costly or critical the vehicle, the more extensive the use of sensors and electronic devices to capture and emit data for safety, operator training, fleet maintenance and management.

Whatever the collection device—whether in manufacturing processes, engineering processes or end consumer or industrial products—there is only one certainty: the number of such devices will only increase every year and the amount of data each can generate will do likewise. The result is a certain hyper-exponential increase in the volume of data to be analyzed in the optimization of most of the large scale manufacturing, product design, engineering and maintenance processes in existence.

Healthcare. Perhaps the ultimate application for intelligent sensors—and a source for both a great volume of data and great value—is our own bodies and the medical devices that can increase and sustain our health, safety, and mobility.

Intelligent electronic devices—some used by people at home and some that travel with them as they go about their day—now capture and transmit data for analysis in managing chronic diseases and conditions; for dealing with sleep disorders; for monitoring exercise; and, for a rapidly growing array of health and medical uses.

In general, more frequent data about what is actually happening with the heart, the breathing process, the blood sugar or the blood pressure—as the patient goes about daily life—greatly increases the ability to make good clinical decisions. This is a large improvement over depending on only information that the doctor can gather in a quarterly examination and interview of the patient. Thus, a cardiologist may get data every day about every ambulatory patient—roughly one hundred times as often as he or she could get the information via quarterly office visits—thereby sometimes getting an early warning of a problem and preventing a heart attack.

Purpose and Methodology for this Report



This WinterCorp Executive Report describes the phenomenon known as “Big Data” and the approach to it recently introduced by Oracle Corporation.

In developing this report, WinterCorp drew on its own independent research and experience; interviewed Oracle employees; and, reviewed Oracle product materials.

Oracle was provided an opportunity to comment on the paper with respect to facts, in its capacity as the sponsor of this research.

WinterCorp has final editorial control over the content of this publication and is solely responsible for any opinions expressed.



A WINTER CORPORATION EXECUTIVE REPORT

The data gathered by such medical devices is voluminous and growing rapidly; it calls for intensive and complex analysis both to enhance clinical decisions and to guide research on better practices. While a relatively new area for “big data,” healthcare promises to yield large scale and valuable opportunities in the years to come.

Common Characteristics. These three big data examples have some things in common. In each case, there is a high volume to be analyzed. In all three cases, the data arrives for analysis at high velocity. In all three cases, there is a lot of data variety. The consumer sentiment comes in mostly unstructured messages and text. In the other cases, there are multiple sources—some structured, some not. And, in all three cases, there is low value density

When the analysis shows that a thousand customers have voiced the same sentiment; that a thousand patients have exhibited a similar pattern prior to heart attack; that a consistent pattern of engineering measurements correlate with a structural failure...then there can be value that has not been found elsewhere.

To some extent, all three of these examples can be attacked by storing and analyzing data in a relational data warehouse. But, a different type of data analysis environment—one designed for “big data”—one that features lower cost and higher flexibility—can provide a better fit for the purpose.

NEW OPPORTUNITIES

In addition to similar technical requirements, these three examples have something else in common: an extraordinarily high return on investment—based on a solution that was previously infeasible. In the past, it would have either cost too much or taken too long to get the analytical result, particularly in a situation where the outcome was uncertain and many different analytical techniques might have to be tried before the value is found.

In addition, note that these are three examples of a broad and far reaching set of opportunities. So, sentiment analysis applies equally to consumer and industrial products. The sensors described in the second example are just one of many types of electronic devices in pervasive use to automatically and remotely capture data for analysis in engineering, manufacturing and maintenance. Also in widespread use are audio and optical devices; location sensing devices; and, many other types of measurement devices.

Big data is thus linked with an enormous and immensely valuable set of new business, scientific and social opportunities in the years ahead.

NEW TECHNOLOGIES FOR BIG DATA

The new opportunities in front of us promise tremendous advances in commerce, engineering, manufacturing and healthcare—to name just a few examples.

But they require the solution to a tough problem: how do you handle these immense volumes of data? How do you deal with data that doesn't fit well structurally in a relational data warehouse? How can you afford to process and store data that arrives constantly 10-to-100 times faster than your transactional data? How can you accomplish the needed intensive, frequent and sometimes immediate analysis—analysis that is difficult or impossible to express in SQL? And, how do you deal with the economics of such large volumes of data, particularly when you may not know in advance whether it contains information of value?

The engineers at Google, Yahoo and other companies working with them—confronting such a problem—decided a new approach was needed. Their approach has principles in common with the highly parallel relational data warehouse—and some principles that are different.

First, as in the relational data warehouse, they envisioned large clusters of servers sharing the work to be done. And, they pictured distributing the data to be analyzed over the servers, so that each server could work on its own piece of data at the same time, shortening the time to completion. This principle, also, is in common with the relational database.

Second, from the start, they knew they wanted a solution that would work for very large clusters of servers—thousands, tens of thousands—even hundreds of thousands of servers. These components would have to be very compact and inexpensive. Otherwise, the solution would become unaffordable as it grew.

Third, they focused on how to simplify the writing of application programs that would operate on data distributed across the servers of a large cluster. Thus, they relieved the application programmer of concern with the distributed architecture of the cluster, but chose not to provide certain services we take for granted in a relational database, such as system management and

A WINTER CORPORATION EXECUTIVE REPORT

enforcement of data definitions or system optimization of complex non-procedural queries.

From these developmental efforts, two core database capabilities were created: a system called Cassandra for so-called “NoSQL” operational data management; and, a system called Hadoop for highly parallel intensive data analytics. These systems are now available as open source products from the Apache Foundation.

The following brief descriptions of Hadoop and Cassandra are from the Apache Foundation website:

Hadoop. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using a simple programming model. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

Hadoop supports MapReduce, a software framework for distributed processing of large data sets on compute clusters.

Cassandra. The Apache Cassandra Project develops a highly scalable second-generation distributed database, bringing together Dynamo's fully distributed design and Bigtable's Column Family-based data model.

Cassandra was open sourced by Facebook in 2008, and is now developed by Apache committers and contributors from many companies.

Some hundreds of companies have initiated programs to evaluate, experiment with and, in some cases, adopt these open source products for big data. But, they face new challenges. For example, using a relatively new open source product to manage and analyze large volumes of data—as part of a high value business process—is a difficult challenge for most companies.

ENTERPRISE REQUIREMENTS FOR BIG DATA

Now, what if you want these big data capabilities in your company, which is in retailing, financial services, manufacturing or some other industry? You are going to

install a new, large scale infrastructure. You will use it to manage important data to support a business purpose. You will create and maintain analytic software – Java programs employing MapReduce, for example—to use with the data. And, you will want to have a sustainable, adaptable resource that integrates into your existing IT environment and produces actionable, replicable results.

You need a way to create, use, manage adapt and sustain the system over a period of time, as individual employees come and go. You also need to be able to operate such a system reliably; avoid the loss of critical data; readily change the system as new technology and new requirements develop; and, apply the system to new problems as they arise. And, of course, the big data solution must be easy to deploy and integrate into their existing IT environment, and also be manageable, supportable and cost effective.

There are other enterprise requirements. Consider a project intended to increase customer retention in a credit card business. The project begins with sentiment analysis to understand what customers like and don't like about the current card, services, pricing, etc. This is performed in the big data environment in Hadoop and identifies customers who are expressing some concern. Information on those customers and their issues is then transferred to the data warehouse environment, where it can be combined with other customer data, transaction data and credit scores. From this combined analysis, incentives offers can be developed to retain the most valuable and credit worthy customers from the group that is unhappy but still enrolled. A campaign can then be conducted, measuring the impact of these actions. The campaign can be assessed partly by the customers' spending behavior and partly by the sentiment they express in email, website visits, etc.

Thus, the complete customer retention project involves a dozen or more steps in which data is used in—and moved between—three different environments: the big data environment, where there are facilities for capturing and analyzing customer sentiment; the data warehouse, where there is structured data on customers, transactions and credit score; and, the business intelligence environment, where there are tools for query, reporting and structured analysis.

So, an important part of the enterprise requirement is the need to carry out complete projects that deliver business

A WINTER CORPORATION EXECUTIVE REPORT

value and that apply the data and resources spanning big data, data warehousing and business intelligence.

ORACLE'S ENGINEERED SYSTEMS FOR DATA MANAGEMENT

Oracle addresses the enterprise requirements for big data via its overall approach to data management, including its Engineered Systems, as depicted in *Figure 1*.

Oracle's Engineered Systems for data management include Oracle Exadata Database Machine, Oracle Exalytics and Oracle Big Data Appliance. Each of these consists of integrated hardware and software, pre-configured, pre-tested and engineered for a certain class of data management requirements.

Exadata Database Machine is Oracle's solution to typical relational database requirements—both transaction processing and data warehousing. Exalytics In-Memory Machine is aimed at high performance business intelligence requirements. And, Oracle Big Data Appliance is Oracle's answer to the big data requirements that are the focus of this report.

Note that Oracle's data management software remains available as separate products. This is true for all the software components in the new Oracle Big Data Appliance, as well as for the software components of the other Oracle Engineered Systems for data management.

ORACLE'S BIG DATA APPLIANCE

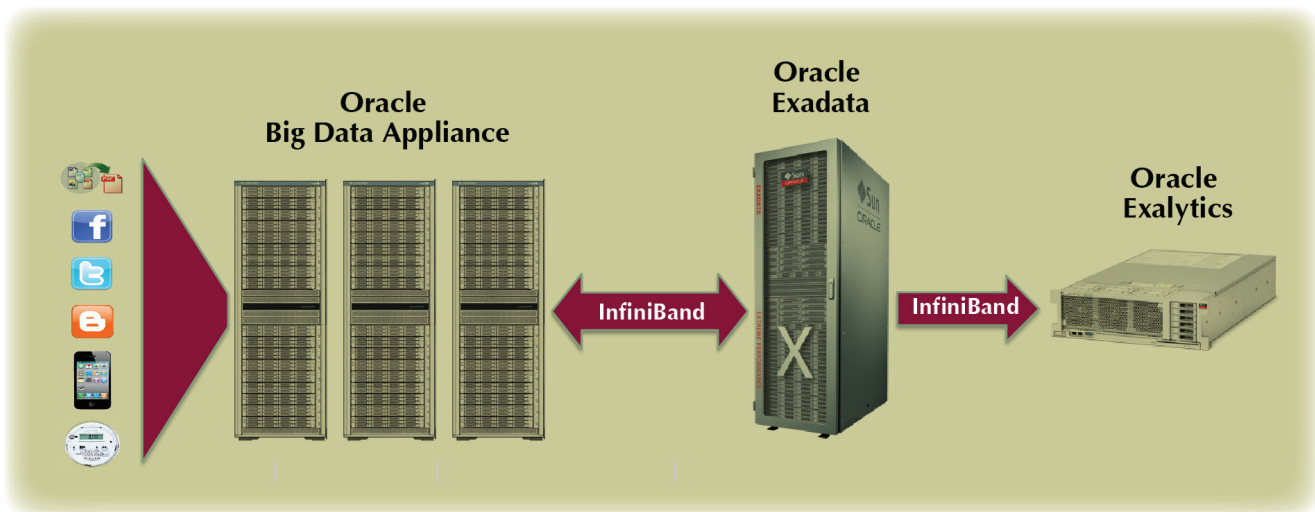
Oracle approaches big data as it does other data management requirements, believing that many customers reject the burden of integrating hundreds or thousands of software and hardware components to create a complete, scalable data management solution. By relieving the customer of the system integration, system maintenance and system support costs—and by adopting an enterprise approach to customer requirements—Oracle delivers significant value with this approach.

Oracle's key goals for Big Data Appliance are to enable its customers to jump start their big data projects by delivering a comprehensive and pre-integrated solution that is well integrated with Exadata and that the customer can readily deploy and support.

Oracle Big Data Appliance comes in a full rack configuration with 864GB of main memory and 432 TB of storage. The major hardware components are:

- 18 Nodes, each a Sun server consisting of
 - 2 CPUs (6-core Intel processors)
 - 48 GB memory per node (upgradable to 96 GB or 144 GB)
 - 12 x 2TB disks per node
- InfiniBand Networking
- 10 GbE connectivity

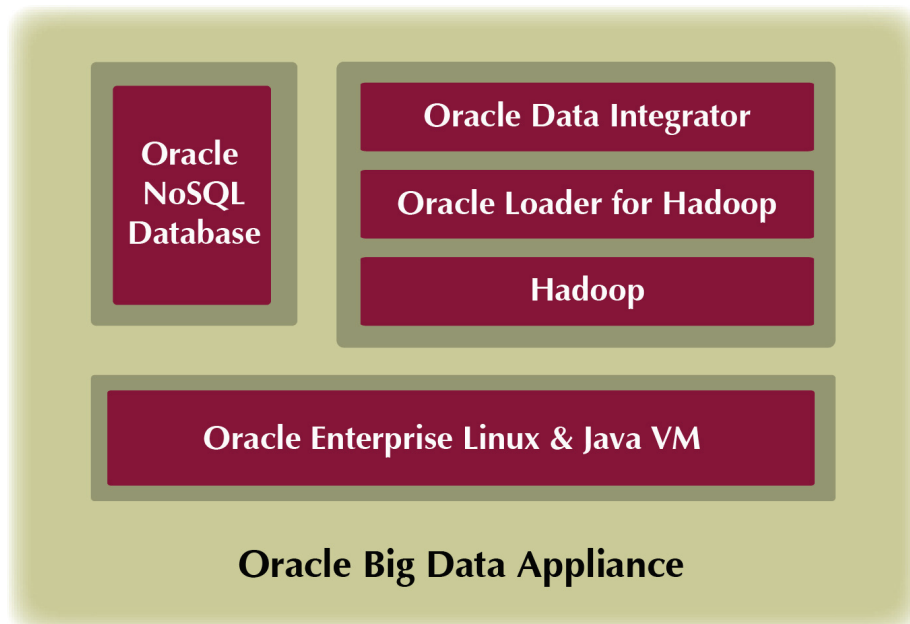
Figure 1: Oracle's Engineered Systems for Data Management



Source: Oracle Corporation

A WINTER CORPORATION EXECUTIVE REPORT

Figure 2: Oracle Big Data Appliance



Source: Oracle Corporation

Oracle Big Data Appliance includes a combination of open source components, packaged as system software with the appliance, and software developed by Oracle, packaged as Big Data Connectors. Key components include:

- **Open source distribution of Apache Hadoop**, the framework for distributed data analysis with programs written in Java or other procedural languages and interfacing to the data in Hadoop via MapReduce;
- **Open source distribution of R**, a programming language and software environment for statistical computing and graphics; R is part of the GNU project;
- **Oracle NoSQL Database** is a distributed, scalable, key-value database with a simple programming model; designed for ready deployment and management in an enterprise environment; the principal role for Oracle NoSQL Database here is low latency data capture/update and fast processing of simple operational queries;
- **Oracle Data Integrator Application Adaptor for Hadoop** simplifies the integration of data from Hadoop and data from Oracle databases; once the adaptor has been used to make Hadoop data accessible in the database, end users can access it with SQL and with business intelligence tools;
- **Oracle Loader for Hadoop** supports the use of MapReduce to load data from Hadoop into Oracle Database 11g. Unlike other Hadoop loaders, it generates Oracle internal formats to load data faster and use fewer database system resources;
- **Java HotSpot Virtual Machine**, a version of the Java virtual machine featuring just in time compilation and adaptive optimization, based in part on the automatic identification of hot spots in the executing Java program;
- **Oracle Enterprise Linux**, the version of Red Hat Linux enhanced by Oracle for enterprise use; certified to work with Oracle middleware; and,

employed as the operating system on Oracle's Engineered Systems.

To the extent that Oracle can deliver on its aims, Oracle Big Data Appliance will represent a sharp contrast to the customer experience of self-integrating, operating and managing a similar capability out of open source components, and independently sourced commercial hardware and software.

FINDINGS AND RECOMMENDATIONS

The much discussed phenomenon of "Big Data", while over promoted in some quarters, is about a very real set business opportunities with outstanding upside potential. Big data implementations are already a significant business force in hundreds of enterprises. They are likely to have a much greater impact in the coming years.

Oracle has outlined an approach to big data—focused on its Engineered Systems—aimed at making big data initiatives practical for the enterprise. The Oracle approach has three key strengths.

First, the customer is relieved of the integration involved in assembling a suitable set of hardware and software components to create a big data architecture. When Oracle Big Data Appliance is delivered and connected

A WINTER CORPORATION EXECUTIVE REPORT

appropriately to the customer's existing infrastructure, the customer receives a working hardware and software infrastructure for big data—pre configured, pre-tested—and ready to use.

Second, commercial quality support is available—and the entire system is supported by a single vendor. This is vastly different from the situation in which the customer must isolate the software and hardware components involved in any given problem and interact with multiple vendors—or perhaps with an open source support community—in order to diagnose and correct the problem.

Third, and more significantly for most companies already using Oracle Database: the new big data environment connects to the existing Oracle database environment at the data management software level. So, Oracle Big Data appliance can be used for staging and ETL processes that need to occur upstream of the data warehouse. The included Oracle Loader for Hadoop can then be used to transfer the data in parallel into Oracle Exadata. Oracle Exadata can readily use data that comes from an Oracle Big Data Appliance in conjunction with data already in the data warehouse.

For Oracle customers with an investment in Oracle software and hardware, it is a major benefit to be able to initiate work in big data, knowing that the data warehouse environment will be able to exchange and integrate data with the big data environment.

Note that all the software present in Oracle Big Data Appliance is also available separately for customers who prefer to custom configure their hardware and software systems.

SUMMING UP

Oracle Big Data Appliance—in connection with Oracle Exadata Database Machine and Oracle Exalytics In-Memory Machine—offers a new route for Oracle customers to leverage the business opportunities associated with “Big Data.” Big Data Appliance is aimed at addressing the technical issues of data variety, data velocity and data volume. At the same time, the Oracle approach to Big Data is focused on addressing the key enterprise requirements of integration, manageability and supportability.

As with any new product, customers must evaluate with care—and test rigorously—to ensure that their requirements for performance, scale, reliability and manageability are in fact satisfied by the products under consideration.

But, bearing the product maturity in mind, WinterCorp recommends that Oracle customers with an interest in big data take a look at Oracle Big Data Appliance.