An Oracle White Paper
January 2012

Oracle: Big Data for the Enterprise

ORACLE

Executive Summary

Today the term big data draws a lot of attention, but behind the hype there's a simple story. 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.

To derive real business value from big data, you need the right tools to capture and organize 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. Oracle offers the broadest and most integrated portfolio of products to help you acquire and organize these diverse data types and analyze them alongside your existing data to find new insights and capitalize on hidden relationships.

Introduction

With the recent introduction of Oracle Big Data Appliance and Oracle Big Data Connectors, Oracle is the first vendor to offer a complete and integrated solution to address the full spectrum of enterprise big data requirements. Oracle's big data strategy is centered on the idea that you can evolve your current enterprise data architecture to incorporate big data and deliver business value. By evolving your current enterprise architecture, you can leverage the proven reliability, flexibility and performance of your Oracle systems to address your big data requirements.

Defining Big Data

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.
- 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

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.

The Importance of 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.

For example, in the delivery of healthcare services, management of chronic or long-term conditions is expensive. Use of in-home monitoring devices to measure vital signs, and monitor progress is just one way that sensor data can be used to improve patient health and reduce both office visits and hospital admittance.

Manufacturing companies deploy sensors in their products to return a stream of telemetry. Sometimes this is used to deliver services like OnStar, that delivers communications, security and navigation services. Perhaps more importantly, this telemetry also reveals usage patterns, failure rates and other opportunities for product improvement that can reduce development and assembly costs.

The proliferation of smart phones and other GPS devices offers advertisers an opportunity to target consumers when they are in close proximity to a store, a coffee shop or a restaurant. This opens up new revenue for service providers and offers many businesses a chance to target new customers.

Retailers usually know who buys their products. Use of social media and web log files from their ecommerce sites can help them understand who didn't buy and why they chose not to, information not available to them today. This can enable much more effective micro customer segmentation and targeted marketing campaigns, as well as improve supply chain efficiencies.

Finally, social media sites like Facebook and LinkedIn simply wouldn't exist without big data. Their business model requires a personalized experience on the web, which can only be delivered by capturing and using all the available data about a user or member.

Building a Big Data Platform

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 to allow you to conduct deep analytics on the combined data set.

Infrastructure Requirements

The requirements in a big data infrastructure span data acquisition, data organization and data analysis.

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.

For example, NoSQL databases are often used to collect and store social media data. While customer facing applications frequently change, underlying storage structures are kept simple. Instead of designing a schema with relationships between entities, these simple structures often just contain a major key to identify the data point, and then a content container holding the relevant data. This simple and dynamic structure allows changes to take place without costly reorganizations at the storage layer.

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.

Analyze Big Data

Since data is not always moved during the organization phase, the analysis may also be done in a distributed environment, where some data will stay where it was originally stored and be transparently accessed from a data warehouse. 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.

For example, analyzing inventory data from a smart vending machine in combination with the events calendar for the venue in which the vending machine is located, will dictate the optimal product mix and replenishment schedule for the vending machine.

Solution Spectrum

Many new technologies have emerged to address the IT infrastructure requirements outlined above. At last count, there were over 120 open source key-value databases for acquiring and storing big data, with Hadoop emerging as the primary system for organizing big data and relational databases expanding their reach into less structured data sets to analyze big data. These new systems have created a divided solutions spectrum comprised of:

- Not Only SQL (NoSQL) solutions: developer-centric specialized systems
- SQL solutions: the world typically equated with the manageability, security and trusted nature of relational database management systems (RDBMS)

NoSQL systems are designed to capture all data without categorizing and parsing it upon entry into the system, and therefore the data is highly varied. SQL systems, on the other hand, typically place data in well-defined structures and impose metadata on the data captured to ensure consistency and validate data types.

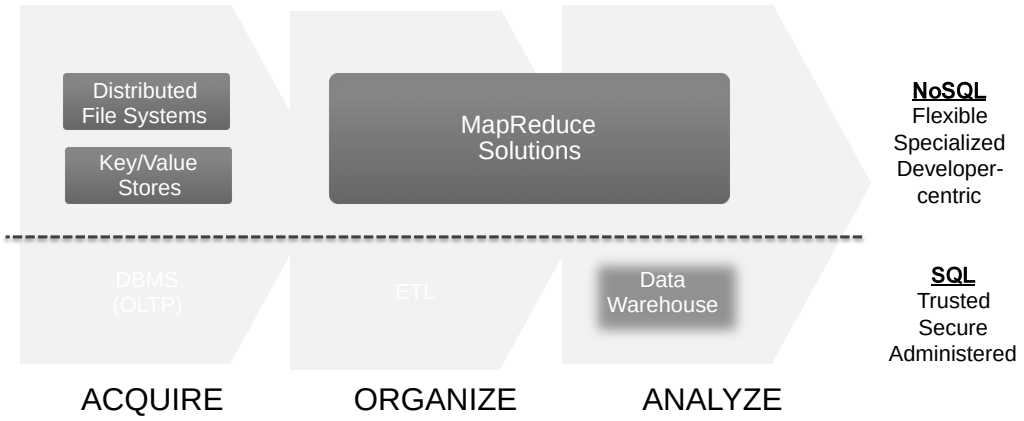

Figure 1 Divided solution spectrum

Distributed file systems and transaction (key-value) stores are primarily used to capture data and are generally in line with the requirements discussed earlier in this paper. To interpret and distill information from the data in these solutions, a programming paradigm called MapReduce is used. MapReduce programs are custom written programs that run in parallel on the distributed data nodes.

The key-value stores or NoSQL databases are the OLTP databases of the big data world; they are optimized for very fast data capture and simple query patterns. NoSQL databases are able to provide very fast performance because the data that is captured is quickly stored with a single indentifying key rather than being interpreted and cast into a schema. By doing so, NoSQL database can rapidly store large numbers of transactions.

However, due to the changing nature of the data in the NoSQL database, any data organization effort requires programming to interpret the storage logic used. This, combined with the lack of support for complex query patterns, makes it difficult for end users to distill value out of data in a NoSQL database.

To get the most from NoSQL solutions and turn them from specialized, developer-centric solutions into solutions for the enterprise, they must be combined with SQL solutions into a single proven infrastructure, that meets the manageability and security requirements of today's enterprises.

Oracle's Big Data Solution

Oracle is the first vendor to offer a complete and integrated solution to address the full spectrum of enterprise big data requirements. Oracle's big data strategy is centered on the idea that you can evolve your current enterprise data architecture to incorporate big data and deliver business value, leveraging the proven reliability, flexibility and performance of your Oracle systems to address your big data requirements.

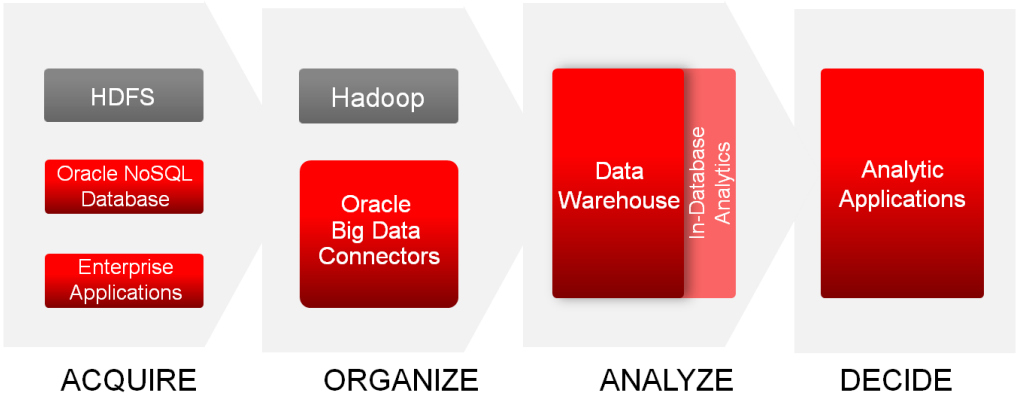

Figure 2 Oracle's Big Data Solutions

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 featuring specialized solutions developed by Oracle to deliver a complete, easy-to-deploy solution for acquiring, organizing and loading big data into Oracle Database 11g. It is designed to deliver extreme analytics on all data types, with enterprise-class performance, availability, supportability and security. With Big Data Connectors, the solution is tightly integrated with Oracle Exadata and Oracle Database, so you can analyze all your data together with extreme performance.

Oracle Big Data Appliance

Oracle Big Data Appliance comes in a full rack configuration with 18 Sun servers for a total storage capacity of 648TB. Every server in the rack has 2 CPUs, each with 6 cores for a total of 216 cores per full rack. Each server has 48GB[1] memory for a total of 864GB of memory per full rack.

[1] Upgradeable to 96GB or 144GB

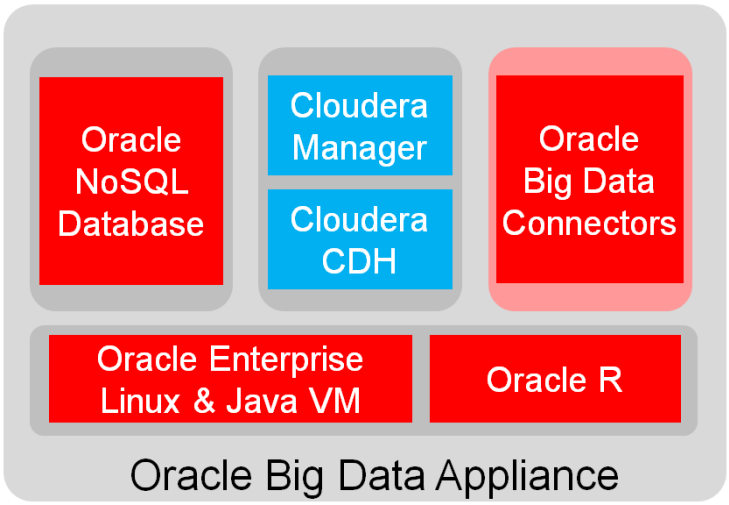

Figure 3 High-level overview of software on Big Data Appliance

Oracle Big Data Appliance includes a combination of open source software and specialized software developed by Oracle to address enterprise big data requirements.

The Oracle Big Data Appliance integrated software[2] includes:

- Full distribution of Cloudera's Distribution including Apache Hadoop (CDH)
- Cloudera Manager to administer all aspects of Cloudera CDH
- Open source distribution of the statistical package R for analysis of unfiltered data on Oracle Big Data Appliance
- Oracle NoSQL Database Community Edition[3]
- And Oracle Enterprise Linux operating system and Oracle Java VM

CDH and Cloudera Manager

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

[2] Oracle Big Data Connectors is a separately licensed product but Big Data Appliance can be pre-configured with Big Data Connectors

[3] Oracle NoSQL Database Enterprise Edition is available for Oracle Big Data Appliance as a separately licensed component

comprehensive set of open source software components needed to use Hadoop. Cloudera Manager is an end-to-end management application for CDH. Cloudera Manager gives a cluster-wide, real-time view of nodes and services running; provides a single, central place to enact configuration changes across the cluster; and incorporates a full range of reporting and diagnostic tools to help optimize cluster performance and utilization.

Oracle Big Data Connectors

Where Oracle Big Data Appliance makes it easy for organizations to acquire and organize new types of data, Oracle Big Data Connectors enables an integrated data set for analyzing all data. Oracle Big Data Connectors can be installed on Oracle Big Data Appliance or on a generic Hadoop cluster.

Oracle Loader for Hadoop

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. Unlike other Hadoop loaders, it generates Oracle internal formats to load data faster and use less database system resources. OLH is added as the last step in the MapReduce transformations as a separate map – partition – reduce step. This last step uses the CPUs in the Hadoop cluster to format the data into Oracle-understood formats, allowing for a lower CPU load on the Oracle cluster and higher data ingest rates because the data is already formatted for Oracle Database. Once loaded, the data is permanently available in the database providing very fast access to this data for general database users leveraging SQL or Business Intelligence tools.

Oracle Direct Connector for Hadoop Distributed File System

Oracle Direct Connector for Hadoop Distributed File System (HDFS) is a high speed connector for accessing data on HDFS directly from Oracle Database. Oracle Direct Connector for HDFS gives users the flexibility of querying data from HDFS at any time, as needed by their application.

It allows the creation of an external table in Oracle Database, enabling direct SQL access on data stored in HDFS. The data stored in HDFS can then be queried via SQL, joined with data stored in Oracle Database, or loaded into the Oracle Database. Access to the data on HDFS is optimized for fast data movement and parallelized, with automatic load balancing. Data on HDFS can be in delimited files or in Oracle data pump files created by Oracle Loader for Hadoop.

Oracle Data Integrator Application Adapter for Hadoop

Oracle Data Integrator Application Adapter for Hadoop simplifies data integration from Hadoop and an Oracle Database through Oracle Data Integrator's easy to use interface. Once the data is accessible in the database, end users can use SQL and Oracle BI Enterprise Edition to access data.

Enterprises that are already using a Hadoop solution, and don't need an integrated offering like Oracle Big Data Appliance, can integrate data from HDFS using Big Data Connectors as a stand-alone software solution.

Oracle R Connector for Hadoop

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 provides users of the open-source statistical environment R with the ability to analyze data stored in HDFS, and to run R models at scale against large volumes of data leveraging MapReduce processing – without requiring R users to learn yet another API or language. End users can leverage over 3500 open source R packages to analyze data stored in HDFS, while administrators do not need to learn R to schedule R MapReduce models in production environments.

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.

Oracle NoSQL Database

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. This intelligent driver keeps track of the underlying storage topology, shards the data and knows where data can be placed with the lowest latency. Unlike competitive solutions, Oracle NoSQL Database is easy to install, configure and manage, supports a broad set of workloads, and delivers enterprise-class reliability backed by enterprise-class Oracle support.

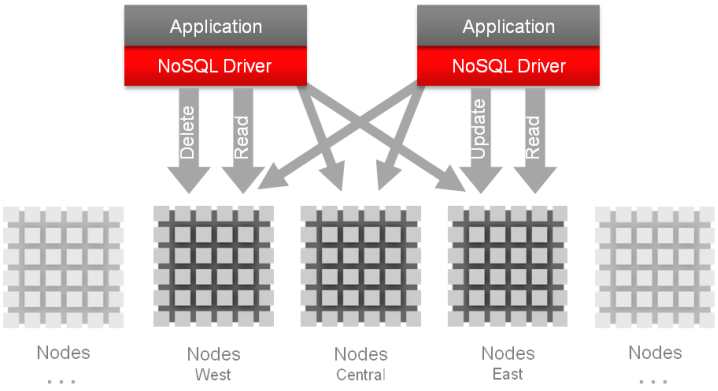

Figure 4 NoSQL Database Architecture

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. The product is available in both an open source community edition and in a priced enterprise edition for large distributed data centers. The former version is installed as part of the Big Data Appliance integrated software.

In-Database Analytics

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.

Connections between Oracle Big Data Appliance and Oracle Exadata are via InfiniBand, enabling high-speed data transfer for batch or query workloads. Oracle Exadata provides outstanding performance in hosting data warehouses and transaction processing databases.

Now that the data is in mass-consumption format, Oracle Exalytics can be used to deliver the wealth of information to the business analyst. Oracle Exalytics is an engineered system providing speed-of-thought data access for the business community. It is optimized to run Oracle Business Intelligence Enterprise Edition with in-memory aggregation capabilities built into the system.

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.

Figure 5 shows three Big Data Appliances streaming data – for example leveraging Apache Flume – from sensors and social media, acquiring this data, organizing it and leveraging Oracle Exadata for data analysis.

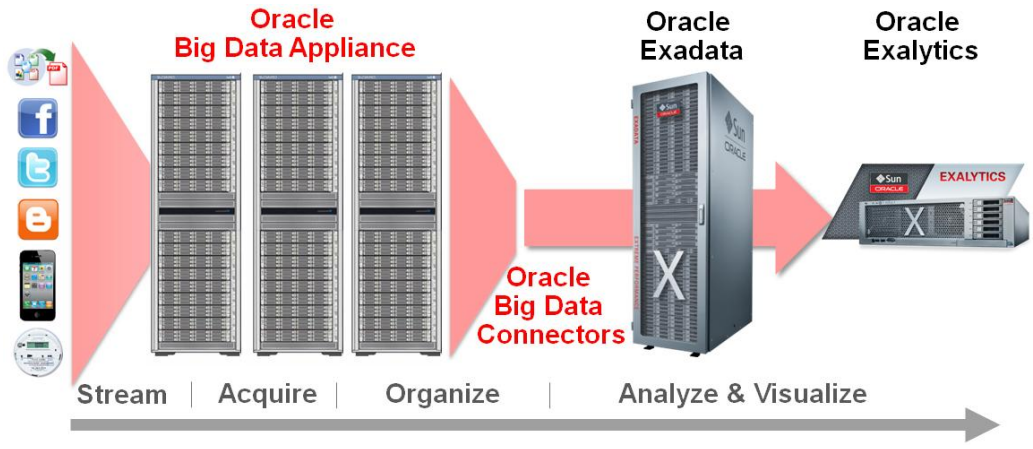

Figure 5 Usage model for Big Data Appliance and Exadata

Conclusion

Analyzing new and diverse digital data streams can reveal new sources of economic value, provide fresh insights into customer behavior and identify market trends early on. But this influx of new data creates challenges for IT departments. To derive real business value from big data, you need the right tools to capture and organize 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. 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.

Oracle: Big Data for the Enterprise
January 2012
Author: Jean-Pierre Dijcks

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200
oracle.com

Oracle is committed to developing practices and products that help protect the environment

0109