Oracle Database 11g for Data Warehousing & Big Data: Strategy, Roadmap

Jean-Pierre Dijcks, Hermann Baer
Oracle
Redwood City, CA, USA

Keywords:

Big Data, Oracle Big Data Appliance, Hadoop, NoSQL, Oracle NoSQL DB, Analytics, Parallel Processing

Introduction

Many things have been said about the topic of Big Data. This paper will share a vision of how Oracle delivers big data for the enterprise and how an enterprise can leverage that solution to deliver business value and competitive advantage. That vision boils down to a very basic premise: Evolve the current enterprise data architecture to incorporate big data and deliver business value.

The three main data source categories for big data are:

- Traditional enterprise data sources – CRM systems, transactional or ERP data, web store transactions, customer information etc.
- Machine generated / sensor data – Call detail records, weblogs, smart meters, manufacturing systems, financial risk calculations etc.
- Social data – customer feedback streams, micro-blogging sites like Twitter, social media platforms like Facebook

Traditional enterprise data source have grown rapidly in size over the past 10 years. Data warehouses have grown from between 1 – 10 Terabyte size on average to the 50 – 100 TB size range.

Machine generated data is generated in even large quantities. Public sources quote a single jet engine to generate 10TB of data in 30 minutes. With more than 25,000 flights per day the volume of just this single data source runs into the petabytes. Smart meters, heavy industrial equipment like oil refineries and drilling rigs generate similar data volumes compounding the problem.

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. Twitter users produce 8TB of data every day.
To manage these data streams IT infrastructures must evolve to handle the data types, the speed of delivery and the capacity to store this data.
To derive business value from these data streams IT capabilities must evolve to allow for deep analytics on all of the organizations data.

Therefore, big data at its essence is the evolution of infrastructure and processes to deliver deep analytics on extremely large, complex data streams flowing into the enterprise from a very large number of data source.

Building a Big Data Platform

As with data warehousing, web stores or any IT platform, big data requires specialized components, technology and people. Current data management infrastructure must evolve into a big data platform.

To build an enterprise big data platform, a wide ranging set of requirements must be covered. New technologies as well as well-understood ones should be considered in satisfying the requirements for a big data platform.
Infrastructure Requirements

As with all large infrastructure endeavors, a big data platform has to deal with a set of requirements. These requirements are roughly divided into three categories as is shown in **Figure 1**.

Figure 1 Acquire, organize and analyze big data

Acquire Big Data
The acquisition phase is one of the major changes in infrastructure from the days before big data. Big data often deals with data streams of higher velocity and higher variety as a lot of the data streams are generated by machines or sensors. Examples of those data types are raw call detail records, web logs, smart meter readings and other data generated by sensors.

Acquiring that type of data requires a number of requirements to be satisfied, generally summarized by the following categories:

- Low, predictable latency in both capturing data and in executing short, simple queries
- The ability to handle very high transaction volumes, often in a distributed environment
- A reliance on flexible, sometimes even dynamic data structures

In social media, a key component to acquire and manage is a user profile. Facebook's 750 Million users each have a unique profile linked to the content generated by these users. Web stores capture profiles, shopping cards and wish lists. All of this profile information is combined with browsing behavior generating very large data stores.

As the customer facing applications change frequently, the underlying storage structures are kept dynamic or simple. This dynamic ability allows changes to take place without massive reorganizations of storage structures.

Smart meters continuously generate data points tied to a meter id or household allowing for up-to-the-minute billing capability like peak demand pricing. Industrial equipment data can be captured to determine when maintenance is required, or how to detect potential failures before they happen.

Organize Big Data
In classical data warehousing organizing data is called data integration. Because of the high volumes in big data, there is a tendency to organize data for consumers at its original storage location. By not moving around large volumes of data, both time and cost is saved.

Requirements for this infrastructure component are:

- Ability to process and manipulate data in the original storage location
- Very high throughput (often in batch) to deal with large data processing steps
- Ability to handle a large variety of data formats, from unstructured to structured

Hadoop is a new technology that allows large data volumes to the organized and processed while keeping the data on the original data storage cluster. Hadoop Distributed File System (HDFS) is the primary 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 and then typically loaded into a Relational DBMS system.

Analyze Big Data
Once the acquisition and organization components of data are in place the analysis of big data can start. As the organization phase did not move data, analysis will often be in a distributed environment as well. Some data stays where it was stored originally and is combined with data in a data warehouse.

The main requirements for analyzing big data are:

- Deeper analytics on a wider variety of data types stored in diverse systems
- Scale to previously impossible data volumes
- Faster response times driven by changes in behavior
- Automate decisions based on analytical models

Statistical analysis, data mining and the analysis of relationships between users are a few examples of deeper analytics. While a lot of these techniques have been available for years, they are now applied to a much wider and larger data set.

Social data, which often shows trends in real-time, implies that analytics are no longer a means to analyze historical patterns. Analytics are now real time decision engines alerting or changing behavior as social media patterns dictate. If an item in a particular region or city generates a sudden burst of activity on social platforms, direct action is required for either damage control or to capture as much of the buzz as is possible.

Smart vending machines are another example. If a specific vending machine is in a location that is about to host a sports tournament the vending machine should be filled with more

sports drinks or should be replenished sooner than scheduled. To achieve this active replenishment a mix of inventory data for the machine and the events calendar for the venue must automatically adjust the schedule to restock the machine.

Solution Spectrum

The requirements shown in the previous section lead to technology solutions focused on solving these specific requirements. Systems to solve data acquisition requirement sprout like mushrooms, shown by the over 120 open source key-value databases available at last count. Hadoop has emerged as a system to organize those data streams. Relational database are expanding their reach into less structured data sets.

These new systems created a divided solutions spectrum shown in **Figure 2**. These two schools of thought are often grouped into:

- Not Only SQL (NoSQL) solutions, typically developer centric environments creating highly specialized systems for an organization
- SQL - the world typically equated with the security and trusted nature of relational database management systems (RDBMS)

The data equated to the NoSQL systems is typically of high variety, meaning that these systems are designed as large buckets to simply capture any data without categorizing and parsing the data upon entry into the system.

SQL solutions typically place data in well-defined structures and impose metadata on the data captured. SQL solutions are carefully designed to ensure consistency and validate data types. NoSQL stores first and foremost store data and worry about what that data is at a later point, where a SQL store defines what data is expected and rejects data not in that format. This leads to higher variety of data types in NoSQL stores, and much higher information density in the SQL world. This is shown in the y-axis in **Figure 2**.

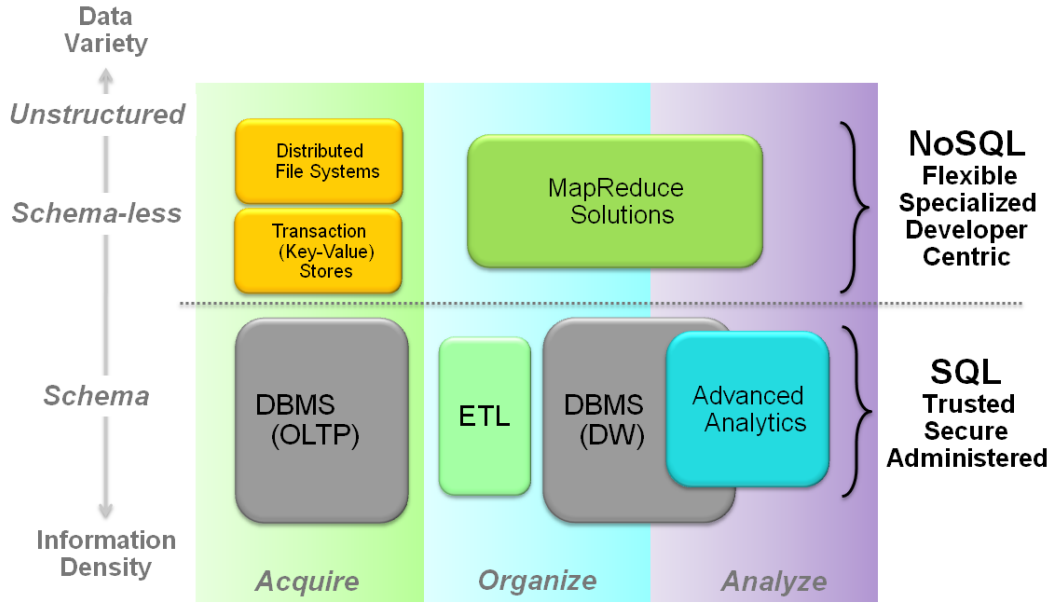

Figure 2 Divided solution spectrum

Distributed file systems, transaction (key-value) stores – as is shown in **Figure 2** – are mostly used to capture data and are generally in line with the requirements discussed earlier in this paper. These NoSQL solutions are the OLTP databases of the big data world optimized for very fast data capture and only concerned with simple query patterns. To achieve the performance required data is not interpreted and cast into a schema, but as quickly as possible stored with a single identifying key. Doing so allows the NoSQL solution to rapidly store large numbers of transactions.

However, due to the changing nature of the data in the NoSQL solutions, any data organization effort requires programming to interpret the storage logic used. That fact and the fact that no complex query patterns are supported make it harder for end users to distil value out of NoSQL data.

To turn big data from a highly flexible, specialized and developer centric system into an enterprise ready solution we must combine the two solutions into a single infrastructure and bring the managed, secure and trusted world of databases to the entire ecosystem in an evolutionary matter.

Oracle's Big Data Solution

Bridging the divide between NoSQL and SQL is much simpler than it seems if all the rhetoric is disregarded and the capabilities of technologies are matched to business requirements.

Bridging the Gap

One way of working with both NoSQL and SQL in a single infrastructure is by leveraging technologies to bridge the two worlds. To achieve this Oracle is introducing Oracle Loader for Hadoop, allowing fast and easy data loads from Hadoop into the Oracle RDBMS.
Oracle Loader for Hadoop – as shown in **Figure 3** – bridges the Hadoop world into Oracle Database by generating optimized data files for loading into Oracle Database.

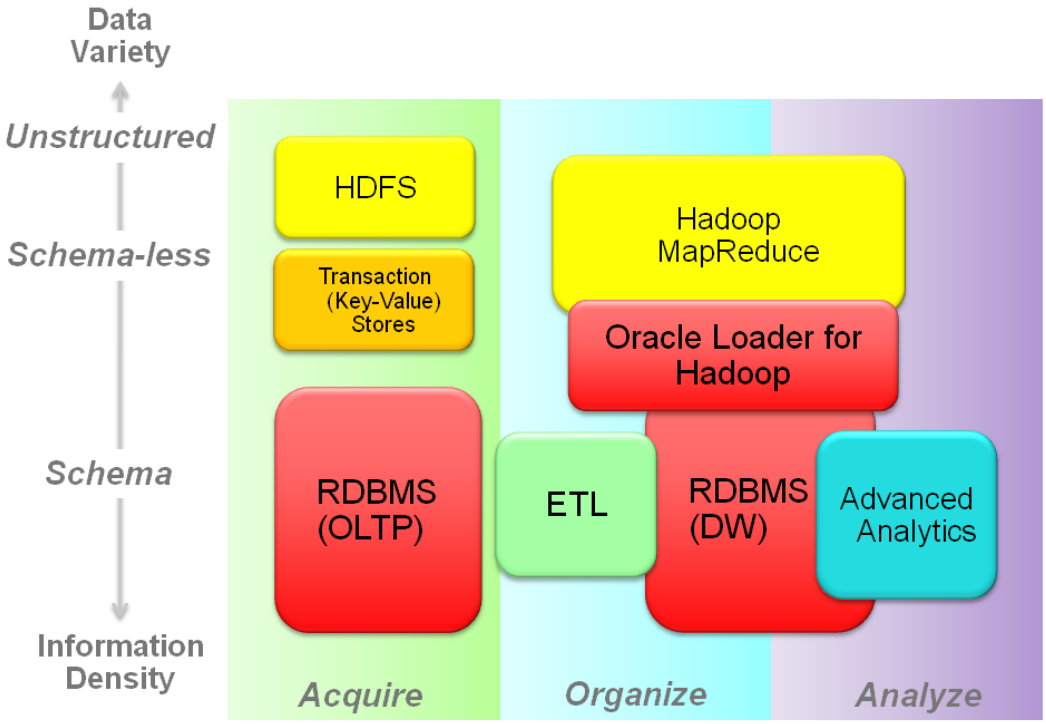

Figure 3 Oracle Loader for Hadoop

Oracle Loader for Hadoop (OLH) does load data into Oracle, but only does so when the data is reduced down from its initial size. 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.

To load data, OLH is added as the last step in the MapReduce transformations as a separate map – partition – reduce step. That last step uses the CPUs in the Hadoop cluster to do the formatting into Oracle understood formats. Doing this formatting on the Hadoop cluster allows for lower CPU load on the Oracle cluster and allows for higher data ingest rates because the data is already formatted for Oracle Database.

Another possible solution to integrate data from HDFS leverages Oracle external tables. To create an external table based solution, mount the HDFS file system via an open source solution called Filesystem in User SpacE (FUSE) and define an external table on a data set. Once HDFS is visible to the database via the external table, the data on HDFS can be accessed via SQL queries and can be joined with database resident data. There is no need to first load the data into the RDBMS. More details can be found in this article.

External tables and direct access are best used in scenarios where initial data exploration is the goal or where incidental access to large volumes of data on HDFS is required. Once the data analyst understands the data and has created solutions with that data, it can be loaded into the database using Oracle Loader for Hadoop. By loading it permanently into Oracle Database, higher performance data access and better availability of that data for large numbers of end users is easier to guarantee.

Oracle Engineered Systems

The unique capability of Oracle – combining software with hardware into engineered systems – allows for the creation of a much simpler to implement solution for big data. Engineered systems make deployment easier and speed to value quicker.

Oracle Big Data Appliance

To allow quicker deployment of a big data system in the enterprise, Oracle Big Data Appliance combines optimized hardware components with a set of new or enhanced big data software components.

Oracle Big Data Appliance delivers:

- Complete and Optimized solution for big data
- Single Vendor support for both hardware and software
- Easy to deploy solution
- Tight integration with Oracle Exadata and Oracle Database

Figure 4 shows how Oracle Big Data Appliance fits within the entire ecosystem of Oracle engineered systems for big data. Oracle Big Data Appliance intends to be the data capture and organization platform for data stored in NoSQL solutions.

The integration with Oracle Exadata is particularly important as the goal is to support access to all data. By integrating Oracle Big Data Appliance with Oracle Exadata a single unified data access pattern is available to all data analysts.

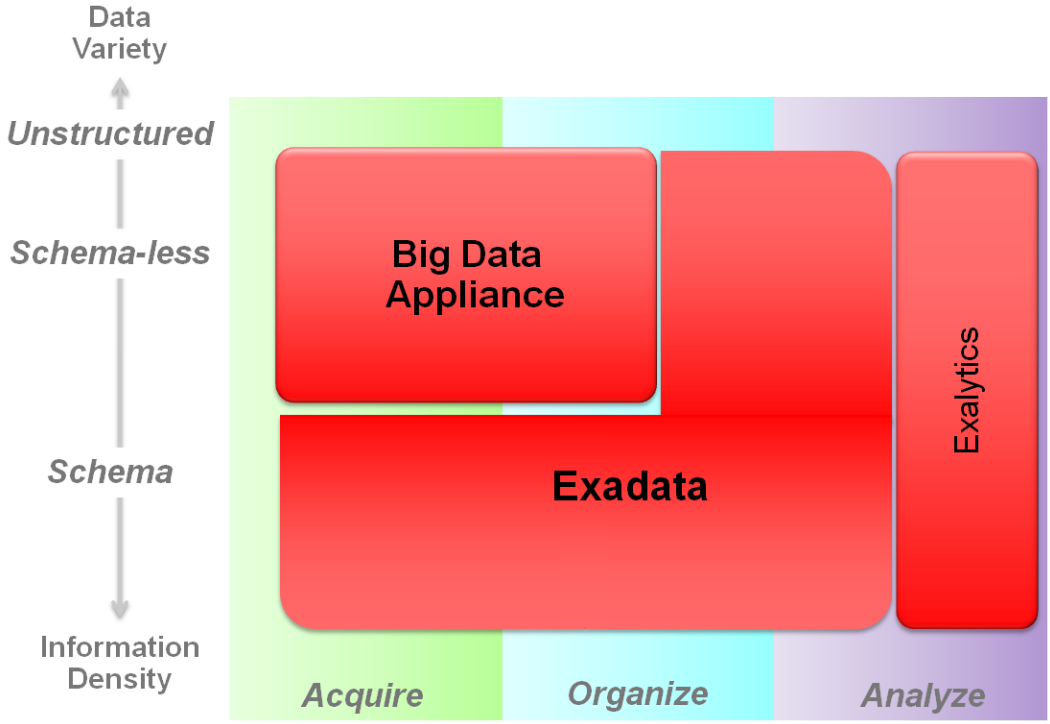

Figure 4 Engineered systems for big data

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 where Oracle Exalytics is an an engineered system providing speed of thought data access for the business community.

Oracle Big Data Appliance – Hardware Overview

To acquire and organize data Oracle Big Data Appliance comes in a full rack configuration with 18 Sun X4270MS servers for a total capacity of 648TB raw storage. Each server provides 2 CPUs, each with 6 cores and 48GB memory.

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

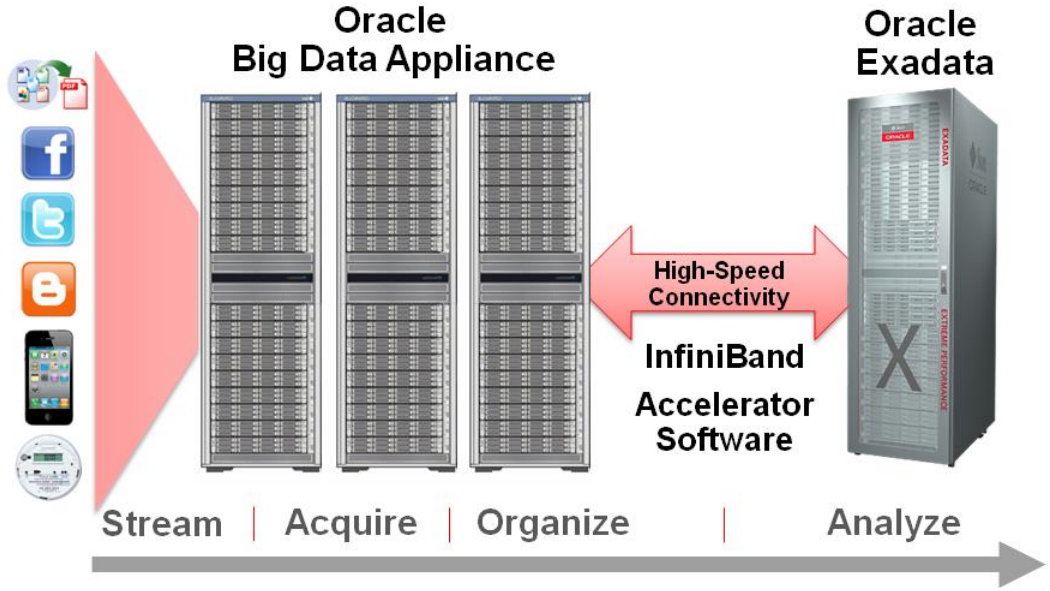

Figure 5 Usage model for Big Data Appliance and Exadata

Oracle Big Data Appliance – Software Overview

The list of software components on Oracle Big Data Appliance includes (see **Figure 6**):

- Oracle Enterprise Linux 5.6 serves as the operating system on Oracle Big Data Appliance
- Oracle JRockit serves as the Java virtual machine
- Distribution including Apache Hadoop, including HDFS and other components[1] fully supported by Oracle when leveraged on Big Data Appliance
- Oracle Loader for Hadoop. OLH is newly developed at Oracle to efficiently load data into Oracle Database. OLH leverages proprietary technology and is not a repackaged SQOOP component

[1] Oracle distribution which includes Hadoop will also include components such as Hive, HBase, Zookeeper and other components. This distribution will only be available on Oracle Big Data Appliance.

- Oracle NoSQL Database Enterprise Edition. A new product developed on top of Oracle Berkeley DB Java Edition. Capture high volume transactions in a key-value stroe and service up simple queries to applications.
- Oracle Data Integrator Application connectors for Hadoop to provide a graphical way of managing ETL loads leveraging Hadoop data
- Seamless integration with Oracle Exadata via newly developed Accelerator Software only available on Oracle Big Data Appliance

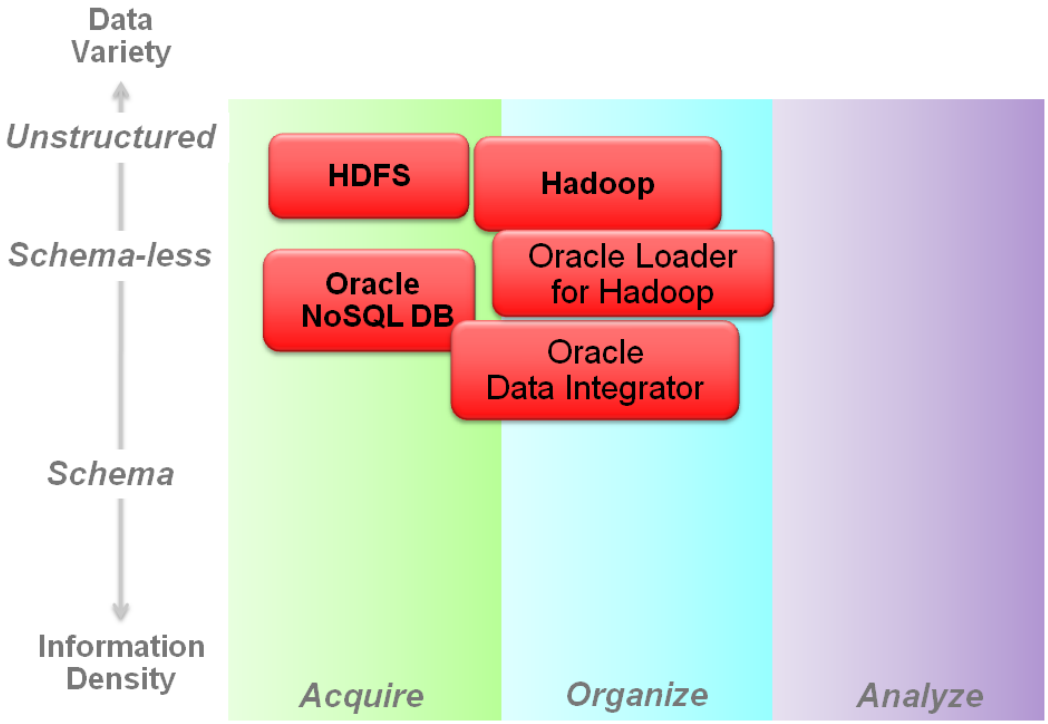

Figure 6 High-level overview of Big Data Appliance Software

Overall, Oracle engineered systems enable customers to build big data solution for the enterprise without the effort and risk of building custom systems.

Integrated Oracle Big Data Software

Figure 7 shows how Oracle delivers a complete and integrated software stack based on the previously discussed engineered systems.

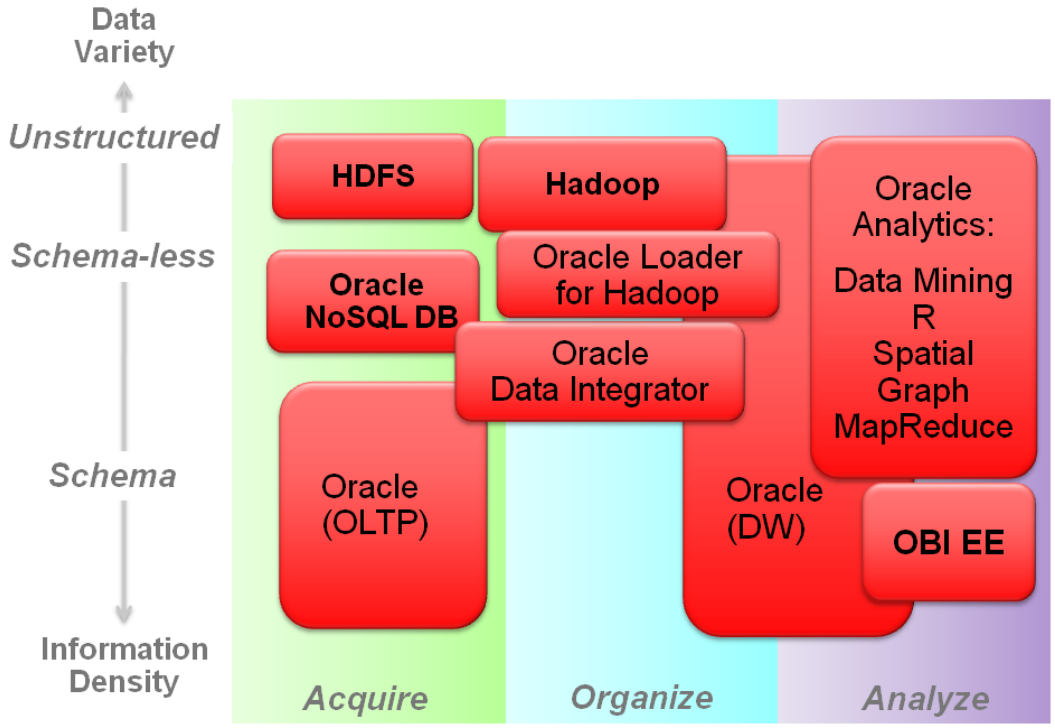

Figure 7 Integrated Oracle software stack for big data

In the acquire phase Oracle's newly introduced key-value solution, NoSQL Database (NoSQL DB) is shown as well as Oracle distribution for Hadoop. NoSQL DB is based on the proven storage technology of Berkeley DB, itself a storage engine of choice underneath open source NoSQL databases.

Oracle NoSQL DB 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.

The primary use cases for NoSQL DB are low latency data capture and fast querying of that data, typically by key lookup. NoSQL DB 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 the large distributed data center setup. The latter version is part of the Big Data Appliance.

As part of the Big Data Appliance, Oracle also distributes Apache Hadoop, including HDFS and other components. Coupled with the before mentioned Oracle Loader for Hadoop customers can create an integrated system managing data within Oracle solutions and allowing easy flow between the NoSQL and HDFS data stores and Oracle database.

Oracle Data Integrator (ODI) allows for an easy design of integration tasks in the Oracle stack. Once the data is accessible in the database, end users can use SQL and Oracle BI Enterprise Edition to access data.

In-Database Analytics

One of the abilities of Oracle Database, shown in **Figure 7**, is to deliver deep analytics to end users. Oracle offers easy to use and high performance tools for analytics, deeply embedded within the database. A small sampling of capabilities:

- 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. An example of leveraging in-database MapReduce is sessionization of weblogs or organization of Call Details Records (CDRs)
- 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 where based on comments. Sentiment analysis tries to show how customers feel about certain companies, products or activities
- 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 the prediction of airline delays at a particular airport
- In-Database Semantic Analysis – the ability to create graphs and connections between various data points and data sets. Semantic analysis creates for examples social networks 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 for example plotted on a map. Enables end users to understand geospatial relationships and trends much easier. For example, spatial data can visualize a network of people and their geographical proximity. Close proximity of a set of customers can lead to influencing of each other's purchasing behavior which can be easily missed if spatial visualization is left out

Figure 8 Combining Churn Models and Sentiment Analysis in OBI EE

Every one of the analytical components in Oracle Database is valuable. Combining a few of these 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.

Conclusion

Big data at its essence is the evolution of infrastructure and processes to deliver deep analytics on extremely large, complex data streams flowing into the enterprise from a very large number of data source.

By delivering an engineered system with new big data software components like Oracle NoSQL Database, Oracle Loader for Hadoop and a distribution including Apache Hadoop and accelerator software for integrating tightly with Oracle Database, Oracle brings big data to the enterprise.

Oracle big data solutions deliver:

- The most comprehensive software stack for big data
- Deep analytics within a single system
- Enterprise ready solutions
- Engineered to work together

On top of that, Oracle Big Data Appliance delivers:

- Complete and optimized solution for big data in a single rack
- Single Vendor support for both hardware and software
- Easy to deploy solution eliminating risk and reducing time to market
- Tight integration with Oracle Exadata and Oracle Database

Oracle – Big data for the enterprise.

Contact address:

Jean-Pierre Dijcks
500 Oracle Parkway, M/S 4op7
Redwood City, CA, 94065

Phone: +1 650 607 5394
Email jean-pierre.dijcks@oracle.com
Internet: blogs.oracle.com/datawarehousing