

An Oracle White Paper in Enterprise Architecture
March 2012

Oracle Information Architecture: An Architect's Guide to Big Data

Disclaimer

The following is intended to outline our general product direction. It is intended for informational purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

What Makes Big Data Different?.....	4
Comparing Information Architecture Operational Paradigms	6
Big Data Architecture Capabilities	7
Storage and Management Capability	7
Database Capability	8
Processing Capability	9
Data Integration Capability	9
Statistical Analysis Capability	9
Big Data Architecture.....	10
Traditional Information Architecture Capabilities	10
Adding Big Data Capabilities	10
An Integrated Information Architecture	11
Making Big Data Architecture Decisions	13
Key Drivers to Consider.....	13
Architecture Patterns in Three Use Cases.....	14
Big Data Best Practices	20
#1: Align Big Data with Specific Business Goals.....	20
#2: Ease Skills Shortage with Standards and Governance	21
#3: Optimize Knowledge Transfer with a Center of Excellence	21
#4: Top Payoff is Aligning Unstructured with Structured Data	21
#5: Plan Your Sandbox For Performance	22
#6: Align with the Cloud Operating Model.....	22
Summary.....	22
Enterprise Architecture and Oracle.....	23

Introduction

If your organization is like many, you're capturing and sharing more data from more sources than ever before. As a result, you're facing the challenge of managing high-volume and high-velocity data streams quickly and analytically.

Big Data is all about finding a needle of value in a haystack of unstructured information. Companies are now investing in solutions that interpret consumer behavior, detect fraud, and even predict the future! McKinsey released a report in May 2011 stating that leading companies are using big data analytics to gain competitive advantage. They predict a 60% margin increase for retail companies who are able to harvest the power of big data.

To support these new analytics, IT strategies are mushrooming, the newest techniques include brute force assaults on massive information sources, and filtering data through specialized parallel processing and indexing mechanisms. The results are correlated across time and meaning, and often merged with traditional corporate data sources. New data discovery techniques include spectacular visualization tools and interactive semantic query experiences. Knowledge workers and data scientists sift through filtered data asking one unrelated explorative question after another. As these supporting technologies emerge from graduate research programs into the world of corporate IT, IT strategists, planners, and architects need to both understand them and ensure that they are enterprise grade.

Planning a Big Data architecture is not about understanding just what is different. It's also about how to integrate what's new to what you already have – from database-and-BI infrastructure to IT tools, and end user applications. Oracle's own product announcements in hardware, software, and new partnerships have been designed to change the economics around Big Data investments and the accessibility of solutions. The real industry challenge is not to think of Big Data as a specialized science project, but rather integrate it into mainstream IT.

In this paper, we will discuss adding Big Data capabilities to your overall information architecture, planning for adoption using an enterprise architecture perspective, and describe some key use cases.

What Makes Big Data Different?

Big data refers to large datasets that are challenging to store, search, share, visualize, and analyze. At first glance, the orders of magnitude outstrip conventional data processing and the largest of data warehouses. For example, an airline jet collects 10 terabytes of sensor data for every 30 minutes of flying time. Compare that with conventional high performance computing where New York Stock Exchange collects 1 terabyte of structured trading data per day. Compare again to a conventional structured corporate data warehouse that is sized in terabytes and petabytes. Big Data is sized in peta-, exa-, and soon perhaps, zetta-bytes! And, it's not just about volume, the approach to analysis contends with data content and structure that cannot be anticipated or predicted. These analytics and the science behind them filter low value or low-density data to reveal high value or high-density data. As a result, new and often proprietary analytical techniques are required. Big Data has a broad array of interesting architecture challenges.

It is often said that data volume, velocity, and variety define Big Data, but the unique characteristic of Big Data is the manner in which the value is discovered. Big Data is unlike conventional business intelligence, where the simple summing of a known value reveals a result, such as order sales becoming year-to-date sales. With Big Data, the value is discovered through a refining modeling process: make a hypothesis, create statistical, visual, or semantic models, validate, then make a new hypothesis. It either takes a person interpreting visualizations or making interactive knowledge-based queries, or by developing 'machine learning' adaptive algorithms that can discover meaning. And in the end, the algorithm may be short-lived.

The growth of big data is a result of the increasing channels and variety of data in today's world. Some of the new data sources are user-generated content through social media, web and software logs, cameras, information-sensing mobile devices, aerial sensory technologies, genomics, and medical records.

Companies have realized that there is competitive advantage in this information and that now is the time to put this data to work.

A Big Data Use Case: Personalized Insurance Premiums

To begin, let's consider the business opportunity that Big Data brings to a conventional business process in the insurance industry – calculating a competitive and profitable insurance premium.

In an effort to be more competitive, an insurance company wants to offer their customers the lowest possible premium, but only to those who are unlikely to make a claim, thereby optimizing their profits. One way to approach this problem is to collect more detailed data about an individual's driving habits and then assess their risk.

In fact, insurance companies are now starting to collect data on driving habits utilizing sensors in their customers' cars. These sensors capture driving data, such as routes driven, miles driven, time of day, and braking abruptness. This data is used to assess driver risk; they compare individual driving patterns with other statistical information, such as average miles driven in your state, and peak hours of drivers on the road. Driver risk plus actuarial information is then correlated with policy and profile information to offer a competitive and more profitable rate for the company. The result? A personalized insurance plan. These unique capabilities, delivered from big data analytics, are revolutionizing the insurance industry.

To accomplish this task, a great amount of continuous data must be collected, stored, and correlated. Hadoop is an excellent choice for acquisition and reduction of the automobile sensor data. Master data and certain reference data including customer profile information are likely to be stored in the existing DBMS systems, and a NoSQL database can be used to capture and store reference data that are more dynamic, diverse in formats, and change frequently. Project R and Oracle R Enterprise are appropriate choice for analyzing both private insurance company data as well as data captured from public sources. And finally, loading the MapReduce results into an existing BI environment allows for further data exploration and data correlation. With these new tools, the company is now able to address the storage, retrieval, modeling, and processing side of the requirements.

In this case, the traditional business process and supporting data (master, transaction, analytic data) are able to add statistically relevant information to their profit model and deliver an industry innovative result.

Every industry has pertinent examples. Big data sources take a variety of forms: logs, sensors, text, spatial, and more. As IT strategists, planners, and architects, we know that our businesses have been trying to find the relevant information in all this unstructured data for years. But the economic choices around getting to these requirements have been a barrier until recently.

So, what is your approach to offer your line of business a set of refined data services that can help them not just find data, but improve and innovate their core business processes? What's the impact to your information architecture? Your organization? Your analytical skills?

Comparing Information Architecture Operational Paradigms

Big data differs from other data realms in many dimensions. In the following table you can compare and contrast the characteristics of big data alongside the other data realms described in [Oracle’s Information Architecture Framework \(OIAF\)](#).

Data Realm	Structure	Volume	Description	Examples
Master Data	Structured	Low	Enterprise-level data entities that are of strategic value to an organization. Typically non-volatile and non-transactional in nature.	Customer, product, supplier, and location/site
Transaction Data	Structured & semi-structured	Medium – high	Business transactions that are captured during business operations and processes	Purchase records, inquiries, and payments
Reference Data	Structured & semi-structured	Low – Medium	Internally managed or externally sourced facts to support an organization’s ability to effectively process transactions, manage master data, and provide decision support capabilities.	Geo data and market data
Metadata	Structured	Low	Defined as “data about the data.” Used as an abstraction layer for standardized descriptions and operations. E.g. integration, intelligence, services.	Data name, data dimensions or units, definition of a data entity, or a calculation formula of metrics.
Analytical Data	Structured	Medium-High	Derivations of the business operation and transaction data used to satisfy reporting and analytical needs.	Data that reside in data warehouses, data marts, and other decision support applications.
Documents and Content	Unstructured	Medium – High	Documents, digital images, geo-spatial data, and multi-media files.	Claim forms, medical images, maps, video files.
Big Data	Structured, semi-structured, & unstructured	High	Large datasets that are challenging to store, search, share, visualize, and analyze.	User and machine-generated content through social media, web and software logs, cameras, information-sensing mobile devices, aerial sensory technologies, and genomics.

Table 1: Data Realms Definitions (Oracle Information Architecture Framework)

These different characteristics have influenced how we capture, store, process, retrieve, and secure our information architectures. As we evolve into Big Data, you can minimize your architecture risk by finding synergies across your investments allows you to leverage your specialized organizations and their skills, equipment, standards, and governance processes.

Here are some capabilities that you can leverage:

Data Realms	Security	Storage & Retrieval	Modeling	Processing & Integration	Consumption
Master data Transactions Analytical data Metadata	Database, app, & user access	RDBMS / SQL	Pre-defined relational or dimensional modeling	ETL/ELT, CDC, Replication, Message	BI & Statistical Tools, Operational Applications
Reference data	Platform security	XML / xQuery	Flexible & Extensible	ETL/ELT, Message	System-based data consumption
Documents and Content	File system based	File System / Search	Free Form	OS-level file movement	Content Mgmt
Big Data - Weblogs - Sensors - Social Media	File system & database	Distributed FS / noSQL	Flexible (Key Value)	Hadoop, MapReduce, ETL/ELT, Message	BI & Statistical Tools

Table 2: Data Realm Characteristics (Oracle Information Architecture Framework)

Big Data Architecture Capabilities

Here is a brief outline of Big Data capabilities and their primary technologies:

Storage and Management Capability

Hadoop Distributed File System (HDFS):

- An Apache open source distributed file system, <http://hadoop.apache.org>
- Expected to run on high-performance commodity hardware
- Known for highly scalable storage and automatic data replication across three nodes for fault tolerance
- Automatic data replication across three nodes eliminates need for backup
- Write once, read many times

Cloudera Manager:

- Cloudera Manager is an end-to-end management application for Cloudera's Distribution of Apache Hadoop, <http://www.cloudera.com>
- 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.

Database Capability

Oracle NoSQL: [\(Click for more information\)](#)

- Dynamic and flexible schema design. High performance key value pair database. Key value pair is an alternative to a pre-defined schema. Used for non-predictive and dynamic data.
- Able to efficiently process data without a row and column structure. Major + Minor key paradigm allows multiple record reads in a single API call
- Highly scalable multi-node, multiple data center, fault tolerant, ACID operations
- Simple programming model, random index reads and writes
- Not Only SQL. Simple pattern queries and custom-developed solutions to access data such as Java APIs.

Apache HBase: [\(Click for more information\)](#)

- Allows random, real time read/write access
- Strictly consistent reads and writes
- Automatic and configurable sharding of tables
- Automatic failover support between Region Servers

Apache Cassandra: [\(Click for more information\)](#)

- Data model offers column indexes with the performance of log-structured updates, materialized views, and built-in caching
- Fault tolerance capability is designed for every node, replicating across multiple datacenters
- Can choose between synchronous or asynchronous replication for each update

Apache Hive: [\(Click for more information\)](#)

- Tools to enable easy data extract/transform/load (ETL) from files stored either directly in Apache HDFS or in other data storage systems such as Apache HBase
- Uses a simple SQL-like query language called HiveQL
- Query execution via MapReduce

Processing Capability

MapReduce:

- Defined by Google in 2004. ([Click here for original paper](#))
- Break problem up into smaller sub-problems
- Able to distribute data workloads across thousands of nodes
- Can be exposed via SQL and in SQL-based BI tools

Apache Hadoop:

- Leading MapReduce implementation
- Highly scalable parallel batch processing
- Highly customizable infrastructure
- Writes multiple copies across cluster for fault tolerance

Data Integration Capability

Oracle Big Data Connectors, Oracle Loader for Hadoop, Oracle Data Integrator:

[\(Click here for Oracle Data Integration and Big Data\)](#)

- Exports MapReduce results to RDBMS, Hadoop, and other targets
- Connects Hadoop to relational databases for SQL processing
- Includes a graphical user interface integration designer that generates Hive scripts to move and transform MapReduce results
- Optimized processing with parallel data import/export
- Can be installed on Oracle Big Data Appliance or on a generic Hadoop cluster

Statistical Analysis Capability

Open Source Project R and Oracle R Enterprise:

- Programming language for statistical analysis ([Click here for Project R](#))
- Introduced into Oracle Database as a SQL extension to perform high performance in-database statistical analysis ([Click here for Oracle R Enterprise](#))
- Oracle R Enterprise allows reuse of pre-existing R scripts with no modification

Big Data Architecture

In this section, we will take a closer look at the overall architecture for big data.

Traditional Information Architecture Capabilities

To understand the high-level architecture aspects of Big Data, let’s first review a well formed logical information architecture for structured data. In the illustration, you see two data sources that use integration (ELT/ETL/Change Data Capture) techniques to transfer data into a DBMS data warehouse or operational data store, and then offer a wide variety of analytical capabilities to reveal the data. Some of these analytic capabilities include: dashboards, reporting, EPM/BI applications, summary and statistical query, semantic interpretations for textual data, and visualization tools for high-density data. In addition, some organizations have applied oversight and standardization across projects, and perhaps have matured the information architecture capability through managing it at the enterprise level.



Figure 1: Traditional Information Architecture Capabilities

The key information architecture principles include treating data as an asset through a value, cost, and risk lens, and ensuring timeliness, quality, and accuracy of data. And, the EA oversight responsibility is to establish and maintain a balanced governance approach including using center of excellence for standards management and training.

Adding Big Data Capabilities

The defining processing capabilities for big data architecture are to meet the volume, velocity, variety, and value requirements. Unique distributed (multi-node) parallel processing architectures have been created to parse these large data sets. There are differing technology strategies for real-time and batch processing requirements. For real-time, key-value data stores, such as NoSQL, allow for high performance, index-based retrieval. For batch processing, a technique known as “Map Reduce,” filters data according to a specific data discovery strategy. After the filtered data is discovered, it can be analyzed directly, loaded into other unstructured databases, sent to mobile devices, or merged into traditional data warehousing environment and correlated to structured data.



Figure 2: Big Data Information Architecture Capabilities

In addition to new unstructured data realms, there are two key differences for big data. First, due to the size of the data sets, we don't move the raw data directly to a data warehouse. However, after MapReduce processing we may integrate the "reduction result" into the data warehouse environment so that we can leverage conventional BI reporting, statistical, semantic, and correlation capabilities. It is ideal to have analytic capabilities that combine a conventional BI platform along with big data visualization and query capabilities. And second, to facilitate analysis in the Hadoop environment, sandbox environments can be created.

For many use cases, such as real time foot traffic management and in-store marketing, the key distinction for big data is that is continuously changing and unpredictable. To track the effectiveness of floor displays and promotions, customer movement and behavior must be interactively explored with visualization or query tools.

In other use cases, the analysis cannot be complete until you correlate it with other enterprise data - structured data. In the example of consumer sentiment analysis, capturing a positive or negative social media comment has some value, but associating it with your most or least profitable customer makes it far more valuable. So, the needed capability with Big Data BI is context and understanding. Using powerful statistical and semantic tools allow you to find the needle in the haystack, and will help you predict the future.

In summary, the Big Data architecture challenge is to meet the rapid use and rapid data interpretation requirements while at the same time correlating it with other data.

What's important is that the key information architecture principles are the same, but the tactics of applying these principles differ. For example, how do we look at big data as an asset. We all agree there's value hiding within the massive high-density data set. But how do we evaluate one set of big data against the other? How do we prioritize? The key is to think in terms of the end goal. Focus on the business values and understand how critical they are in support of the business decisions, as well as the potential risks of not knowing the hidden patterns.

Another example of applying architecture principles differently is data governance. The quality and accuracy requirements of big data can vary tremendously. Using strict data precision rules on user sentiment data might filter out too much useful information, whereas data standards and common definitions are still going to be critical for fraud detections scenarios.

To reiterate, it is important to leverage your core information architecture principles and practices, but apply them in a way that's relevant to big data. In addition, the EA responsibility remains the same for big data. It is to optimize success, centralize training, and establish standards.

An Integrated Information Architecture

One of the obstacles observed in Hadoop adoption in enterprise is the lack of integration with existing BI eco-system. At present, the traditional BI and big data ecosystems are separate causing integrated data analysis headaches. As a result, they are not ready for use by the typical business user or executive.

Earlier adopters of big data have often times written custom code to move the processed results of big data back into database or developed custom solutions to report and analyze on them.

These options might not be feasible and economical for enterprise IT. First of all, it creates proliferations of one-off code and different standards. Architecture impacts IT economics. Big Data done independently runs the risk of redundant investments. In addition, most businesses simply do not have the staff and skill level for such custom development work.

A better option is to incorporate the Big Data results into the existing Data Warehousing platform. The power of information lies in our ability to make associations and correlation. What we need is the ability to bring different data sources, processing requirements together for timely and valuable analysis.

Here is Oracle's holistic capability map that bridges traditional information architecture and big data architecture:

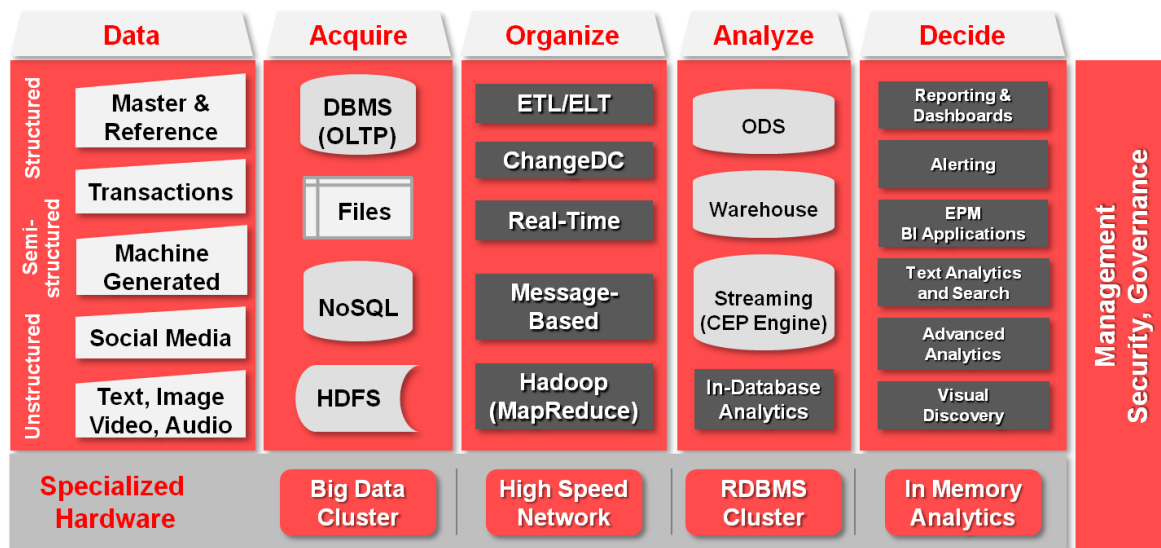


Figure 3: Oracle Integrated Information Architecture Capabilities

As various data are captured, they can be stored and processed in traditional DBMS, files, in the Hadoop Clustered File System, and NoSQL.

Architecturally, the critical component that breaks the divide is the integration layer in the middle. This integration layer needs to extend across all of the data types and domains, and bridge the gap between the traditional and new data acquisition and processing framework. The data integration capability needs to cover the entire spectrum of velocity and frequency. It needs to handle extreme and ever-growing volume requirements. And it needs to bridge the variety of data structures. You need to look for technologies that allow you to integrate Hadoop / MapReduce with your data warehouse and transactional data stores in a bi-directional manner.

The next layer is where you load the “reduction-results” from Big Data processing output into your data warehouse for further analysis. You also need the ability to access your structured data such as customer profile information while you process through your big data to look for patterns such as detecting fraudulent activities.

The Big Data processing output will be loaded into the traditional ODS, data warehouse, and data marts, for further analysis, just as the transaction data. The additional component in this layer is the Complex Event Processing engine to analyze stream data in real time.

The Business Intelligence layer will be equipped with advanced analytics, in-database statistical analysis, and advanced visualization, on top of the traditional components such as reports, dashboards, and queries.

Governance, security, and operational management also cover the entire spectrum of data and information landscape at the enterprise level.

With this architecture, the business users do not see a divide. They don't even need to be made aware that there is a difference between traditional transaction data and Big Data. The data and analysis flow would be seamless as they navigate through various data and information sets, test hypothesis, analyze patterns, and make informed decisions.

Making Big Data Architecture Decisions

Information Architecture is perhaps the most complex area of IT. It is the ultimate investment payoff. Today's economic environment demands that business be driven by useful, accurate, and timely information. And, the world of Big Data adds another dimension to the problem. However, there are always business and IT tradeoffs to get to data and information in a most cost-effective way.

Key Drivers to Consider

Here is a summary of various business and IT drivers you need to consider when making these architecture choices.

BUSINESS DRIVERS	IT DRIVERS
<ul style="list-style-type: none"> • Better insight • Faster turn-around • Accuracy and timeliness 	<ul style="list-style-type: none"> • Reduce storage cost • Reduce data movement • Faster time-to-market • Standardized toolset • Ease of management and operation • Security and governance

Architecture Patterns in Three Use Cases

The business drivers for Big Data processing and analytics are present in every industry and will soon be essential in the delivery of new services and in the analysis of a process. In some cases, they enable new opportunities with new data sources that we have become familiar, such as tapping into the apps on our mobile devices, location services, social networks, and internet commerce. Industry nuances on device generated data can enable remote patient monitoring, personal fitness activity, driving behavior, location-based stored movement, and predicting consumer behavior. But, Big Data also offers opportunities to refine conventional enterprise business processes, such as text and sentiment-based customer interaction through sales and service websites and call center functions, human resource resume analysis, engineering change management from defect through enhancement in product lifecycle management, factory automation and quality management in manufacturing execution systems, and many more.

In this section, we will explore three use cases and walk through the architecture decisions and technology components. Case 1: Retail-weblog analysis. Case 2: Financial Services-real-time transaction detection. Case 3: Insurance-unstructured and structured data correlation.

Use Case #1: Initial Data Exploration

The first example we are going to look at is from the retail sector. One of nation's leading retailers had disappointing results from its web channels during the Christmas season and is looking to improve customers' experience with their online shopping site. Some of the potential areas to investigate include the web logs and product/site reviews for shoppers. It would be beneficial to understand the navigation pattern, especially related to abandoned shopping carts. The business needs to determine the value of these data versus the cost before making a major investment.

This retailer's IT department faces challenges in a number of areas: skill set (or the lack thereof) for these new sets of data and processing power requirements to process such large volume. Traditional SQL tools are preferred choice for business and IT, however, it is not economically feasible to load all the data into a relational database management platform.

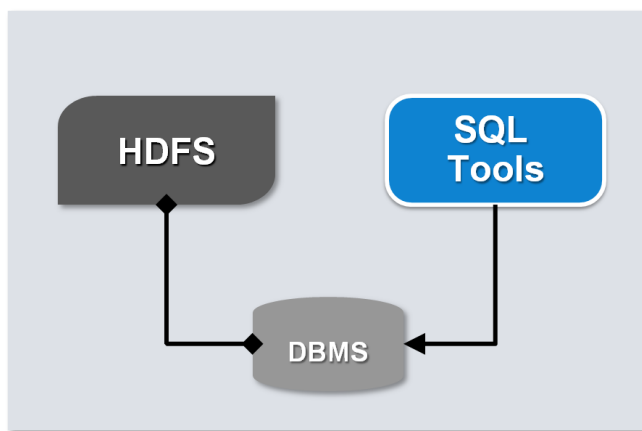


Figure 4: Use Case #1: Initial Data Exploration

As the diagram above shows, this design pattern calls for mounting the Hadoop Distributed File System through a DBMS system so that traditional SQL tools can be used to explore the dataset. The key benefits include:

- No data movement
- Leverage existing investments and skills in database and SQL or BI tools

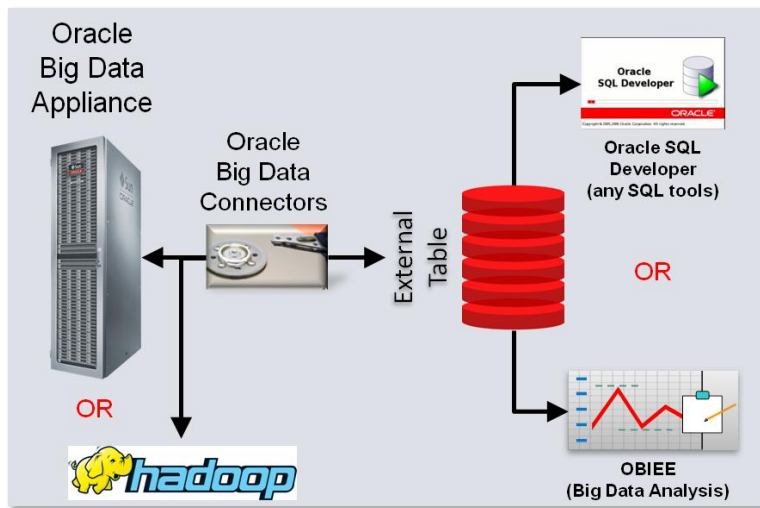


Figure 5: Use Case #1: Architecture Decisions

The diagram above shows the logical architecture using Oracle products to meet these criteria.

The key components in this architecture:

- Oracle Big Data Appliance (or other Hadoop Solutions):
 - Powered by the full distribution of Cloudera's Distribution including Apache Hadoop (CDH) to store logs, reviews, and other related big data
- Oracle Big Data Connectors:
 - Create optimized data sets for efficient loading and analysis in Oracle Database 11g
- Oracle Database 11g:
 - External Table: A feature in Oracle database to present data stored in a file system in a table format and can be used in SQL queries transparently.
- Traditional SQL Tools:
 - Oracle SQL Developer: Development tool with graphic user-interface that allows users to access data stored in a relational database using SQL.

- Business Intelligence tools such as OBIEE can be also used to access data through Oracle Database

In summary, the key architecture choice in this scenario is to avoid data movement, minimize processing requirement and investment, until after the initial investigation. Through this architecture, the retailer mentioned above is able to access Hadoop data directly through database and SQL interface for the initial exploration.

Use Case #2: Big Data for Complex Event Processing

The second use case is relevant to financial services sector. Large financial institutions play a critical role in detecting financial criminal and terrorist activity. However, their ability to effectively meet these requirements are affected by their IT departments' ability to meet the following challenges:

- The expansion of anti-money laundering laws to include a growing number of activities, such as gaming, organized crime, drug trafficking, and the financing of terrorism
- The ever growing volume of information that must be captured, stored, and assessed
- The challenge of correlating data in disparate formats from an multitude of sources

Their IT systems need to provide abilities to automatically collect and process large volumes of data from an array of sources including Currency Transaction Reports (CTRs), Suspicious Activity Reports (SARs), Negotiable Instrument Logs (NILs), Internet-based activity and transactions, and much more.

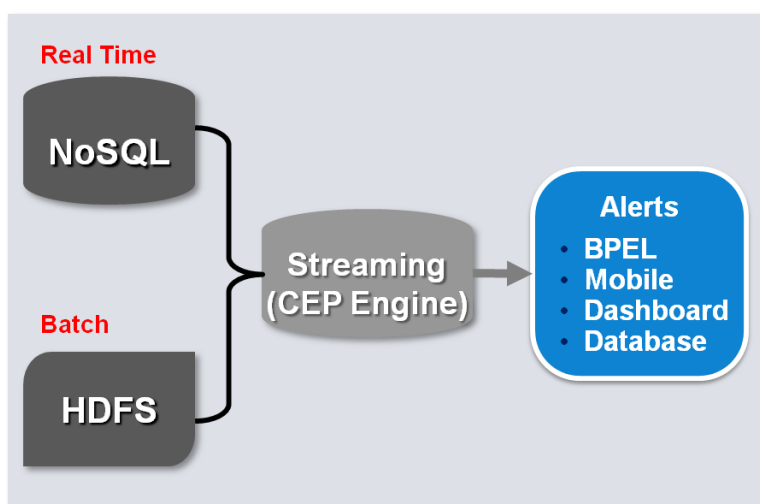


Figure 6: Use Case #2: Big Data for Complex Event Processing

The ideal scenario would have been to include all historic profile changes and transaction records to best determine the rate of risk for each of the accounts, customers, counterparties, and legal entities, at various levels of aggregation and hierarchy. However, it was not traditionally possible due to constraints in processing power and cost of storage. With HDFS, it is now possible to incorporate all the detailed data points to calculate such risk profiles and send to the CEP engine to establish the basis for the risk model.

NoSQL database in this scenario will capture and store low latency and large volume of data from various sources in a flexible data structure, as well as real-time data integration with Complex Event Process engine to enable automatic alerts, dashboard, and trigger business process to take appropriate actions.

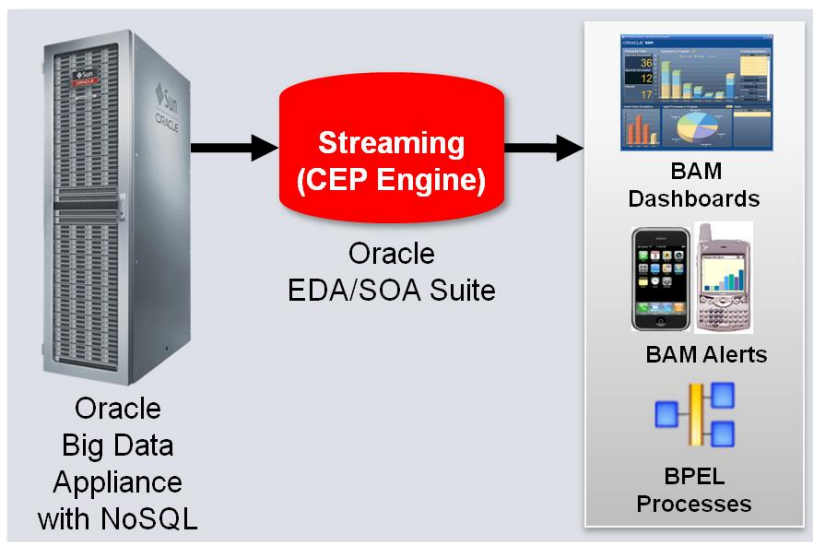


Figure 7: Use Case #2: Architecture Decisions

The logical diagram above highlights the following main components of this architecture:

- Oracle Big Data Appliance (or other Hadoop Solutions):
 - Powered by the full distribution of Cloudera's Distribution including Apache Hadoop (CDH) to store logs, reviews, and other related big data
 - NoSQL to capture low latency data with flexible data structure and fast querying
 - MapReduce to process large amount of data for reduced and optimized dataset to be loaded into database management system
- Oracle EDA:
 - Oracle CEP: Streaming complex event engine to continuously process incoming data, analyze and evolve patterns, and raise events if suspicious activities are detected

- Oracle BPEL: Business Process Execution Language engine to define processes and appropriate actions based on the event raised
- Oracle BAM: Real-time business activity monitoring dashboards to provide immediate insight and generate actions

In summary, the key principle of this architecture is to integrate big data with event driven architecture to meet complex regulatory requirements. Although database management systems are not included in this architecture depiction, it is expected that raised events and further processing transactions and records will be stored in the database either as transactions or for future analytical requirements.

Use Case #3: Big Data for Combined Analytics

The third use case is to continue our discussion of the insurance company mentioned in the earlier section of this paper. In a nutshell, the insurance giant has a need to capture the large amount of sensor data that track their customers' driving habits, store them in a cost effective manner, process this data to determine trends and identify patterns, and to integrate end results with existing transactional, master, and reference data they are already capturing.

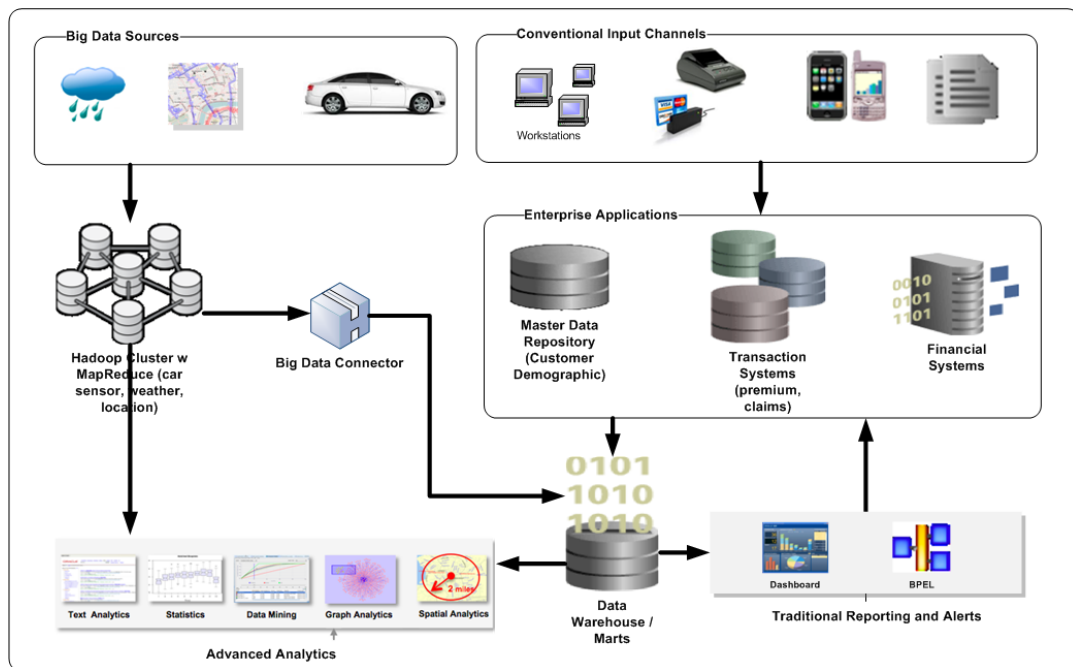


Figure 8: Use Case #3: Data Flow Architecture Diagram

The key architecture challenge of this architecture is to integrate Big Data with structured data.

The diagram below is a high-level conceptual view that reflects these requirements.

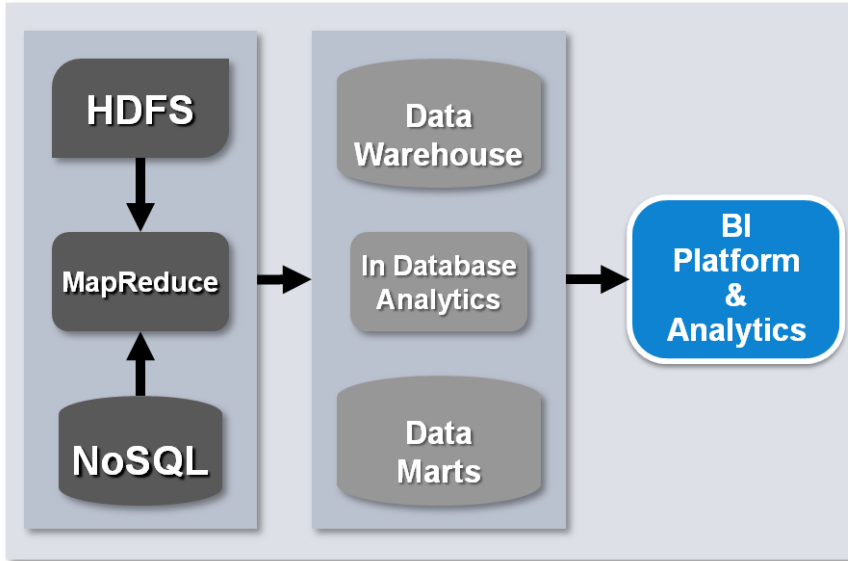


Figure 9: Use Case #3: Conceptual Architecture for Combined Analytics

The large amount of sensor data needs to be transferred to and stored at the centralized environment that provides flexible data structure, fast processing, as well as scalability and parallelism. MapReduce functions are needed to process the low-density data to identify patterns and trending insights. The end results need to be integrated into the database management system with structured data.

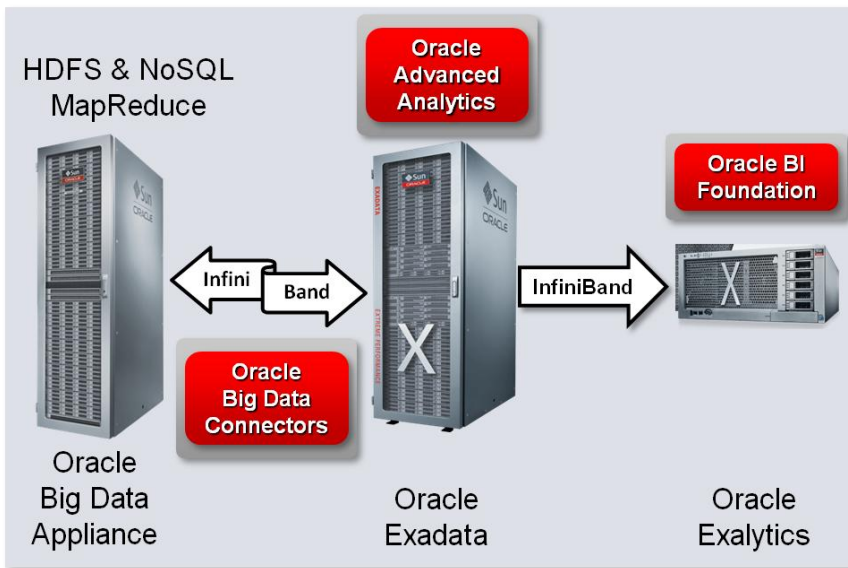


Figure 10: Use Case #3: Physical Architecture for Combined Analytics

Leveraging Oracle engineered systems including Oracle Big Data Appliance, Oracle Exadata, and Oracle Exalytics reduces implementation risks, provides fast time to value and extreme performance and scalability to meet a complex business and IT challenge.

The key components of this architecture include:

- Oracle Big Data Appliance (or other Hadoop Solutions):
 - Powered by the full distribution of Cloudera's Distribution including Apache Hadoop (CDH) to store logs, reviews, and other related big data
 - NoSQL to capture low latency data with flexible data structure and fast querying
 - MapReduce to process large amount of data for reduced and optimized dataset to be loaded into database management system
- Oracle Big Data Connectors: Provides an adapter for Hadoop that integrates Hadoop and Oracle Database through easy to use graphical user interface.
- Oracle Exadata: Engineered database system that supports mixed workloads for outstanding performance of the transactional and/or data warehousing environment for further combined analytics.
- Oracle Exalytics: Engineered BI system that provides speed-of-thought analytical capabilities to end users.
- Infiniband: Connections between Oracle Big Data Appliance, Oracle Exadata, and Oracle Exalytics are via InfiniBand, enabling high-speed data transfer for batch or query workloads.

Big Data Best Practices

Here are a few general guidelines to build a successful big data architecture foundation:

#1: Align Big Data with Specific Business Goals

The key intent of Big Data is to find hidden value - value through intelligent filtering of low-density and high volumes of data. As an architect, be prepared to advise your business on how to apply big data techniques to accomplish their goals. For example, understand how to filter weblogs to understand eCommerce behavior, derive sentiment from social media and customer support interactions, understand statistical correlation methods and their relevance for customer, product, manufacturing, or engineering data. Even though Big Data is a newer IT frontier and there is an obvious excitement to master something new, it is important to base new investments in skills, organization, or infrastructure with a strong business-driven context to guarantee ongoing project investments and funding. To know if you are on the right track, ask yourself, how does it support and enable your business architecture and top IT priorities?

#2: Ease Skills Shortage with Standards and Governance

McKinsey Global Institute¹ wrote that one of the biggest obstacles for big data is a skills shortage. With the accelerated adoption of deep analytical techniques, a 60% shortfall is predicted by 2018. You can mitigate this risk by ensuring that Big Data technologies, considerations, and decisions are added to your IT governance program. Standardizing your approach will allow you to manage your costs and best leverage your resources. Another strategy to consider is to implement appliances that would provide you with a jumpstart and quicker time to value as you grow your in-house expertise.

#3: Optimize Knowledge Transfer with a Center of Excellence

Use a center of excellence (CoE) to share solution knowledge, planning artifacts, oversight, and management communications for projects. Whether big data is a new or expanding investment, the soft and hard costs can be an investment shared across the enterprise. Another benefit from the CoE approach is that it will continue to drive the big data and overall information architecture maturity in a more structured and systematic way.

#4: Top Payoff is Aligning Unstructured with Structured Data

It is certainly valuable to analyze Big Data on its own. However, by connecting high density Big Data to the structured data you are already collecting can bring even greater clarity. For example, there is a difference in distinguishing *all* sentiment from that of only your *best* customers. Whether you are capturing customer, product, equipment, or environmental Big Data, an appropriate goal is to add more relevant data points to your core master and analytical summaries and lead yourself to better conclusions. For these reasons, many see Big Data as an integral extension of your existing business intelligence and data warehousing platform.

Keep in mind that the Big Data analytical processes and models can be human and machine based. The Big Data analytical capabilities include statistics, spatial, semantics, interactive discovery, and visualization. They enable your knowledge workers and new analytical models to correlate different types and sources of data, to make associations, and to make meaningful discoveries. But all in all, consider Big Data both a pre-processor and post-processor of related transactional data, and leverage your prior investments in infrastructure, platform, BI and DW.

¹ McKinsey Global Institute, May 2011, The challenge—and opportunity—of ‘big data’, https://www.mckinseyquarterly.com/The_challenge_and_opportunity_of_big_data_2806

#5: Plan Your Sandbox For Performance

Discovering meaning in your data is not always straightforward. Sometimes, we don't even know what we are looking for initially. That's completely expected. Management and IT needs to support this "lack of direction" or "lack of clear requirement." So, to accommodate the interactive exploration of data and the experimentation of statistical algorithms we need high performance work areas. Be sure that 'sandbox' environments have the power they need and are properly governed.

#6: Align with the Cloud Operating Model

Big Data processes and users require access to broad array of resources for both iterative experimentation and running production jobs. Data across the data realms (transactions, master data, reference, summarized) is part of a Big Data solution. Analytical sandboxes should be created on-demand and resource management needs to have a control of the entire data flow, from pre-processing, integration, in-database summarization, post-processing, and analytical modeling. A well planned private and public cloud provisioning and security strategy plays an integral role in supporting these changing requirements.

Summary

Big Data is here. Analysts and research organizations have made it clear that mining machine generated data is essential to future success. Embracing new technologies and techniques are always challenging, but as architects, you are expected to provide a fast, reliable path to business adoption.

As you explore the 'what's new' across the spectrum of Big Data capabilities, we suggest that you think about their integration into your existing infrastructure and BI investments. As examples, align new operational and management capabilities with standard IT, build for enterprise scale and resilience, unify your database and development paradigms as you embrace Open Source, and share metadata wherever possible for both integration and analytics.

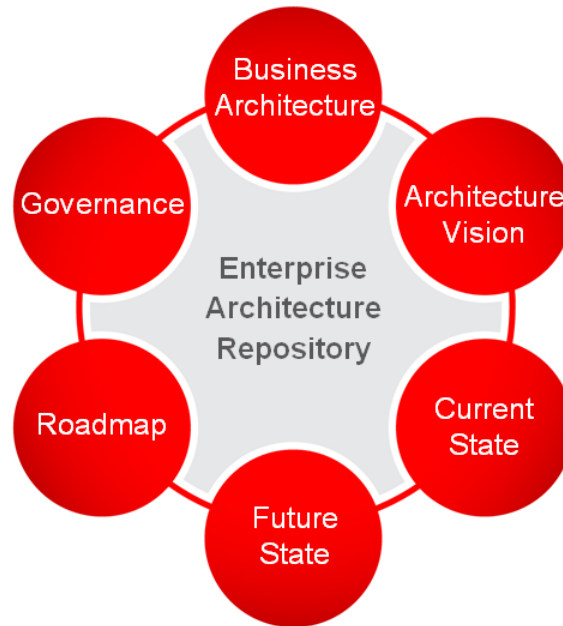
Last but not least, expand your IT governance to include a Big Data center of excellence to ensure business alignment, grow your skills, manage Open Source tools and technologies, share knowledge, establish standards, and to manage best practices.

For more information about Oracle and Big Data, visit www.oracle.com/bigdata.

You can listen to Helen Sun discuss the topics in this white paper at this webcast. Select session 6, Conquering Big Data. [Click here](#).

Enterprise Architecture and Oracle

Oracle has created a streamlined and repeatable process to facilitate the development of your big data architecture vision.



The Oracle Architecture Development Process divides the development of architecture into the phases listed above. Oracle Enterprise Architects and Information Architects use this methodology to propose solutions and to implement solutions. This process leverages many planning assets and reference architectures to ensure every implementation follows Oracle's best experiences and practices.

For additional white papers on the [Oracle Architecture Development Process \(OADP\)](#), the associated [Oracle Enterprise Architecture Framework \(OEAF\)](#), read about Oracle's experiences in enterprise architecture projects, and to participate in a community of enterprise architects, visit the www.oracle.com/goto/EA

To understand more about Oracle's enterprise architecture and information architecture consulting services, please visit, www.oracle.com/goto/EA-Services.

Watch our webcast on Big Data, by clicking [here](#).



An Oracle White Paper
in Enterprise Architecture:

Enterprise Information Management:
Oracle Information Architecture:
An Architect's Guide to Big Data

March 2012

Authors: Helen Sun, Peter Heller

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

www.oracle.com



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

Copyright © 2012, Oracle and/or its affiliates. All rights reserved. This document is provided for information purposes only and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.

Oracle and Java are registered trademarks of Oracle and/or its affiliates. Cloudera, Cloudera CDH, and Cloudera Manager are registered and unregistered trademarks of Cloudera Inc. Other names may be trademarks of their respective owners.

AMD, Opteron, the AMD logo, and the AMD Opteron logo are trademarks or registered trademarks of Advanced Micro Devices. Intel and Intel Xeon are trademarks or registered trademarks of Intel Corporation.

All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. UNIX is a registered trademark licensed through X/Open Company, Ltd. 1010

Hardware and Software, Engineered to Work Together