An Enterprise Architect’s Guide to Big Data

Reference Architecture Overview

ORACLE ENTERPRISE ARCHITECTURE WHITE PAPER | MAY 2015

Disclaimer

The following is intended to outline our general product direction. It is intended for information 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.

Table of Contents

Executive Summary

According to the Forum for Innovation, 90% of the world's data was created in the last 2 years. To find the value in all this data, businesses and IT have been eagerly experimenting with a host of new analytical techniques in addition to storage, processing, and integration technologies. Everyone is keenly aware of Big Data as it is at the heart of nearly every digital transformation. For example, new customer experiences are entirely powered by smart devices and a cultural obsession to respond in the moment. Smart products are not only providing conventional diagnostic codes in real time, but able to capture their entire environmental context. The result will be real-time enterprises that are more competitive by increasing brand relevancy and resiliency. Consequently, enterprise analytical and business processes are being reimagined, and longer, faster, and personalized value chains are emerging. Already, Big Data is a commercial success; in thousands of cases, it has increased brand loyalty, uncovered truths, predicted the future, revealed product reliability, and discovered real accountability.

Big Data presents IT with more than a few challenges. But, there are more similarities than differences. For one, IT is learning that as proofs of concept become operational, these new capabilities must align with the SLAs of the related operational and decision support systems and processes that they are supporting. Big Data operations must perform at scale, be resilient, secure, and governable by standardizing architecture decisions, integration and development. To meet the operate-faster mandate, the analytics need to avoid excessive ETL/transformations and redundant data stores – all of which slow overall throughput as well as are costly to operate and maintain. To deliver what the business needs, architects needs to think clearly about how to efficiently manage the volume, variety, velocity of this new data across the entire enterprise information architecture. Big Data goals are not any different than the rest of your information management goals – it's just that now, the economics to process this data are doable.

There is a good reason that you should look to Oracle as the foundation for your Big Data capabilities. From its inception, 35 years ago, Oracle has invested deeply across nearly every element of information management – from software to hardware and to the innovative integration of both. Not only has Oracle's family of databases continue to solve the toughest technological problems delivering the fastest, highest performance, and most reliable, available, scaleable data platform, it has migrated data through generations of technology change. Furthermore, delivering an information platform requires more than the database itself. It requires all the ancillary capabilities, such as data capture, transformation, movement, quality, security, and management while providing robust data access, visualization, and analytics. Oracle's unique value has been its long history of engineering the broadest stack of enterprise-class information technology to work together—to simplify complex IT environments, reduce TCO, and to minimize the risk when new areas emerge – like Big Data.

Oracle's perspective is that Big Data is not an island. It is merely the latest aspect of an integrated enterprise-class information management capability. Without care, Big Data can easily add to the complexity of a corporate IT environment as it continues to evolve through frequent open source contributions, expanding cloud services, and true innovation in analytic strategies. In addition to relying on a strong company, like Oracle, for best-of-breed products, support, and services, it is even more important to adopt an enterprise architecture approach to navigate your way to the safest and most successful future state. By taking an enterprise architecture approach, decisions can be made ensuring business alignment, a value centric roadmap, and ongoing governance.

This paper is an introduction to the Big Data ecosystem and the architecture choices that an enterprise architect will likely face. This paper defines key terms, capabilities, a reference architecture, describes key Oracle and Open Source products, provides perspectives and principles and ties it together with real-world use cases. The approach and guidance offered is the byproduct of hundreds of customer projects and highlights the decisions that customers faced in the course of their architecture planning and implementations. Oracle's architects work across many industries and government agencies and have developed a standardized methodology based on enterprise architecture best practices. Oracle's enterprise architecture approach and framework are articulated in the Oracle Architecture Development Process (OADP) and the Oracle Enterprise Architecture Framework (OEAF).

A Pointer to Additional Architecture Materials

Oracle offers additional documents that are complementary to this white paper. A few of these are described below:

IT Strategies from Oracle (ITSO) is a series of practitioner guides and reference architectures designed to enable organizations to develop an architecture-centric approach to enterprise-class IT initiatives. ITSO presents successful technology strategies and solution designs by defining universally adopted architecture concepts, principles, guidelines, standards, and patterns.

The Big Data and Analytics Reference Architecture (39 pages) offers a logical architecture and Oracle product mapping. The Information Management Reference Architecture (200 pages) covers the information management aspects of the Oracle Reference Architecture and describes important concepts, capabilities, principles, technologies, and several architecture views including conceptual, logical, product mapping, and deployment views that help frame the reference architecture. The security and management aspects of information management are covered by the *ORA Security* (140 pages) and *ORA Management and Monitoring* (72 pages) Other related documents in this ITSO library include cloud computing, business analytics, business process management, or service-oriented architecture.

The Information Management and Big Data Reference Architecture (30 pages) white paper offers a thorough overview for a vendor-neutral conceptual and logical architecture for Big Data. This paper will help you understand many of the planning issues that arise when architecting a Big Data capability.

A series of short industry specific overviews highlighting specific big data analytical and transaction opportunities for Communications Service Providers, Education, Financial Services, Healthcare Payers and Providers, Logistics, Manufacturing, Media and Entertainment. Oil & Gas, Pharmaceuticals and Life Sciences, Retail, and Utilities.

Lastly, numerous Big Data materials can be found on Oracle Technology Network (OTN) and Oracle.com/BigData.

Fundamental Concepts

What is Big Data?

Historically, a number of the large-scale Internet search, advertising, and social networking companies pioneered Big Data hardware and software innovations. For example, Google analyzes the clicks, links, and content on 1.5 trillion page views per day (www.alexa.com) – and delivers search results plus personalized advertising in milliseconds! A remarkable feat of computer science engineering.

As Google, Yahoo, Oracle, and others have contributed their technology to the open source community, broader commercial and public sector interest took up the challenge of making Big Data work for them. Unlike the pioneers, the broader market sees big data slightly differently. Rather than the data interpreted independently, they see the value realized by adding the new data to their existing operational or analytical systems.

So, Big Data describes a holistic information management strategy that includes and integrates many new types of data and data management alongside traditional data. While many of the techniques to process and analyze these data types have existed for some time, it has been the massive proliferation of data and the lower cost computing models that have encouraged broader adoption. In addition, Big Data has popularized two foundational storage and processing technologies: Apache Hadoop and the NoSQL database.

Big Data has also been defined by the four "V"s: Volume, Velocity, Variety, and Value. These become a reasonable test to determine whether you should add Big Data to your information architecture.

- **Volume.** The amount of data. While volume indicates *more* data, it is the granular nature of the data that is unique. Big Data requires processing high volumes of low-density data, that is, data of unknown value, such as twitter data feeds, clicks on a web page, network traffic, sensor-enabled equipment capturing data at the speed of light, and many more. It is the task of Big Data to convert low-density data into high-density data, that is, data that has value. For some companies, this might be tens of terabytes, for others it may be hundreds of petabytes.
- **Velocity.** A fast rate that data is received and perhaps acted upon. The highest velocity data normally streams directly into memory versus being written to disk. Some Internet of Things (IoT) applications have health and safety ramifications that require real-time evaluation and action. Other internet-enabled smart products operate in real-time or near real-time. As an example, consumer eCommerce applications seek to combine mobile device location and personal preferences to make time sensitive offers. Operationally, mobile application experiences have large user populations, increased network traffic, and the expectation for immediate response.
- **Variety.** New unstructured data types. Unstructured and semi-structured data types, such as text, audio, and video require additional processing to both derive meaning and the supporting metadata. Once understood, unstructured data has many of the same requirements as structured data, such as summarization, lineage, auditability, and privacy. Further complexity arises when data from a known source changes without notice. Frequent or real-time schema changes are an enormous burden for both transaction and analytical environments.
- **Value.** Data has intrinsic value—but it must be discovered. There are a range of quantitative and investigative techniques to derive value from data – from discovering a consumer preference or sentiment, to making a relevant offer by location, or for identifying a piece of equipment that is about to fail. The technological breakthrough is that the cost of data storage and compute has exponentially decreased, thus providing an abundance of data from which statistical sampling and other techniques become relevant, and meaning can be derived. However, finding value also requires new discovery processes involving clever and insightful analysts, business users, and executives. The real Big Data challenge is a human one, which is learning to ask the right questions, recognizing patterns, making informed assumptions, and predicting behavior.

The Big Questions about Big Data

The good news is that everyone has questions about Big Data! Both business and IT are taking risks and experimenting, and there is a healthy bias by all to learn. Oracle's recommendation is that as you take this journey, you should take an enterprise architecture approach to information management; that big data is an enterprise asset and needs to be managed from business alignment to governance as an integrated element of your current information management architecture. This is a practical approach since we know that as you transform from a proof of concept to run at scale, you will run into the same issues as other information management challenges, namely, skill set requirements, governance, performance, scalability, management, integration, security, and access. The lesson to learn is that you will go further faster if you leverage prior investments and training.

Here are some of the common questions that enterprise architects face:

THE BIG DATA QUESTIONS

Areas	Questions	Possible Answers
Business Context		
Business Intent	How will we make use of the data?	» *Sell new products and services* » *Personalize customer experiences* » *Sense product maintenance needs* » *Predict risk, operational results* » *Sell value-added data*
Business Usage	Which business processes can benefit?	» *Operational ERP/CRM systems* » *BI and Reporting systems* » *Predictive analytics, modeling, data mining*
Data Ownership	Do we need to own (and archive) the data?	» *Proprietary* » *Require historical data* » *Ensure lineage* » *Governance*
Architecture Vision		
Ingestion	What are the sense *and* respond characteristics?	» *Sensor-based real-time events* » *Near real-time transaction events* » *Real-time analytics* » *Near real time analytics* » *No immediate analytics*
Data Storage	What storage technologies are best for our data reservoir?	» *HDFS (Hadoop plus others)* » *File system* » *Data Warehouse* » *RDBMS* » *NoSQL database*
Data Processing	What strategy is practical for my application?	» *Leave it at the point of capture* » *Add minor transformations* » *ETL data to analytical platform* » *Export data to desktops*
Performance	How to maximize speed of ad hoc query, data transformations, and analytical modeling?	» *Analyze and transform data in real-time* » *Optimize data structures for intended use* » *Use parallel processing* » *Increase hardware and memory* » *Database configuration and operations* » *Dedicate hardware sandboxes* » *Analyze data at rest, in-place*

Areas	Questions	Possible Answers
Latency	How to minimize latency between key operational components? (ingest, reservoir, data warehouse, reporting, sandboxes)	» *Share storage* » *High speed interconnect* » *Shared private network* » *VPN - across public networks*
Analysis & Discovery	Where do we need to do analysis?	» *At ingest – real time evaluation* » *In a raw data reservoir* » *In a discovery lab* » *In a data warehouse/mart* » *In BI reporting tools* » *In the public cloud* » *On premises*
Security	Where do we need to secure the data?	» *In memory* » *Networks* » *Data Reservoir* » *Data Warehouse* » *Access through tools and discovery lab*
Current State		
Unstructured Data Experience	Is unstructured or sensor data being processed in some way today? (e.g. text, spatial, audio, video)	» *Departmental projects* » *Mobile devices* » *Machine diagnostics* » *Public cloud data capture* » *Various systems log files*
Consistency	How standardized are data quality and governance practices?	» *Comprehensive* » *Limited*
Open Source Experience	What experience do we have in Open Source Apache projects? (Hadoop, NoSQL, etc)	» *Scattered experiments* » *Proof of concepts* » *Production experience* » *Contributor*
Analytics Skills	To what extent do we employ Data Scientists and Analysts familiar with advanced and predictive analytics tools and techniques?	» *Yes* » *No*
Future State		
Best Practices	What are the best resources to guide decisions to build my future state?	» *Reference architecture* » *Development patterns* » *Operational processes* » *Governance structures and polices* » *Conferences and communities of interest* » *Vendor best practices*
Data Types	How much transformation is required for raw unstructured data in the data reservoir?	» *None* » *Derive a fundamental understanding with schema or key-value pairs* » *Enrich data*
Data Sources	How frequently do sources or content structure change?	» *Frequently* » *Unpredictable* » *Never*
Data Quality	When to apply transformations?	» *In the network* » *In the reservoir* » *In the data warehouse* » *By the user at point of use* » *At run time*

Areas	Questions	Possible Answers
Discovery Provisioning	How frequently to provision discovery lab sandboxes?	» *Seldom* » *Frequently*
Roadmap		
Proof of Concept	What should the POC validate before we move forward?	» Business use case » New technology understanding » Enterprise integration » Operational implications
Open Source Skills	How to acquire Open Source skills?	» Cross-train employees » Hire expertise » Use experienced vendors/partners
Analytics Skills	How to acquire analytical skills?	» Cross-train employees » Hire expertise » Use experienced vendors/partners
Governance		
Cloud Data Sources	How to guarantee trust from cloud data sources?	» Manage directly » Audit » Assume
Data Quality	How to clean, enrich, dedup unstructured data?	» *Use statistical sampling* » *Normal techniques*
Data Quality	How frequently do we need to re-validate content structure?	» *Upon every receipt* » *Periodically* » *Manually or automatically*
Security Policies	How to extend enterprise data security policies?	» *Inherit enterprise policies* » *Copy enterprise policies* » *Only authorize specific tools/access points* » *Limited to monitoring security logs*

What's Different about Big Data?

Big Data introduces new technology, processes, and skills to your information architecture and the people that design, operate, and use them. With new technology, there is a tendency to separate the new from the old, but we strongly urge you to resist this strategy. While there are exceptions, the fundamental expectation is that finding patterns in this new data enhances your ability to understand your existing data. Big Data is not a silo, nor should these new capabilities be architected in isolation.

At first glance, the four "V"s define attributes of Big Data, but there are additional best-practices from enterprise-class information management strategies that will ensure Big Data success. Below are some important realizations about Big Data:

Information Architecture Paradigm Shift

Big data approaches data structure and analytics differently than traditional information architectures. A traditional data warehouse approach expects the data to undergo standardized ETL processes and eventually map into pre-defined schemas, also known as "schema on write". A criticism of the traditional approach is the lengthy process to make changes to the pre-defined schema. One aspect of the appeal of Big Data, is that the data can be captured without requiring a 'defined' data structure. Rather, the structure will be derived either from the data itself or through other algorithmic process, also known as "schema on read." This approach is supported by new low-cost, in-memory parallel processing hardware/software architectures, such as HDFS/Hadoop and Spark.

In addition, due to the large data volumes, Big Data also employs the tenet of "bringing the analytical capabilities to the data" versus the traditional processes of "bringing the data to the analytical capabilities through staging, extracting, transforming and loading," thus eliminating the high cost of moving data.

Unifying Information Requires Governance

Combining Big Data with traditional data adds additional context and provides the opportunity to deliver even greater insights. This is especially true in use cases where with key data entities, such as customers and products. 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.

Hence, organizations have the governance responsibility to align disparate data types and certify data quality. Decision makers need to have confidence in the derivation of data regardless of its source, also known as data lineage. To design in data quality you need to define common definitions and transformation rules by source and maintain through an active metadata store. The powerful statistical and semantic tools can enable you to find the proverbial needle in the haystack, and can help you predict future events with relevant degrees of accuracy, but only if the data is believable.

Big Data Volume Keeps Growing

Once committed to Big Data, it is a fact that the data volume will keep growing – maybe even exponentially. In your throughput planning, beyond estimating the basics, such as storage for staging, data movement, transformations, and analytics processing, think about whether the new technologies can reduce latencies, such as parallel processing, machine learning, memory processing, columnar indexing, and specialized algorithms. In addition, it is also useful to distinguish which data could be captured and analyzed in a cloud service versus on-premises.

Big Data Requires Tier 1 Production Guarantees

One of the enabling conditions for big data has been low cost hardware, processing, and storage. However, high volumes of low cost data on low cost hardware should not be misinterpreted as a signal for reduced service level agreement (SLA) expectations. Once mature, production and analytic uses of Big Data carry the same SLA guarantees as other Tier 1 operational systems. In traditional analytical environments users report that, if their business analytics solution were out of service for up to one hour, it would have a material negative impact on business operations. In transaction environments, the availability and resiliency commitment are essential for reliability. As the new Big Data components (data sources, repositories, processing, integrations, network usage, and access) become integrated into both standalone and combined analytical and operational processes, enterprise-class architecture planning is critical for success.

While it is reasonable to experiment with new technologies and determine the fit of Big Data techniques, you will soon realize that running Big Data at scale requires the same SLA commitment, security policies, and governance as your other information systems.

Big Data Resiliency Metrics

Operational SLAs typically include two key related IT management metrics: Recovery Point Objective (RPO) and Recovery Time Objective (RTO). RPO is the agreement for acceptable data loss. RTO is the targeted recovery time for a disrupted business process. In a failure operations scenario, hardware and software must be recoverable to a point in time. While Hadoop and NoSQL include notable high availability capabilities with multi-site failover and recovery and data redundancy, the ease of recovery was never a key design goal. Your enterprise design goal should be to *provide for resiliency across the platform.*

Big Data Security

Big Data requires the same security principles and practices as the rest of your information architecture. Enterprise security management seeks to centralize access, authorize resources, and govern through comprehensive audit practices. Adding a diversity of Big Data technologies, data sources, and uses adds requirements to these practices. A starting point for a Big Data security strategy should be to align with the enterprise practices and policies already established, avoid duplicate implementations, and manage centrally across the environments.

Oracle has taken an integrated approach across a few of these areas. From a governance standpoint, Oracle Audit Vault monitors Oracle and non-Oracle (HDFS, Hadoop, MapReduce, Oozie, Hive) database traffic to detect and block threats, as well as improve compliance reporting by consolidating audit data from databases, operating systems, directories, files systems, and other sources into a secure centralized repository. From data access standpoint, Big Data SQL enables standard SQL access to Hadoop, Hive, and NoSQL with the associated SQL and RBAC security capabilities: querying encrypted data and rules enforced redaction using the virtual private database features. Your enterprise design goal should be to *secure all your data and be able to prove it.*

Big Data and Cloud Computing

In today's complex environments, data comes from everywhere. Inside the company, you have known structured analytical and operational sources in addition to sources that you may have never thought to use before, such as log files from across the technology stack. Outside the company, you own data across your enterprise SaaS and PaaS applications. In addition, you are acquiring and licensing data from both free and subscription public sources – all of which vary in structure, quality and volume. Without a doubt, cloud computing will play an essential role for many use cases: as a data source, providing real-time streams, analytical services, and as a device transaction hub. Logically, the best strategy is move the analytics to the data, but in the end there are decisions to make. The physical separation of data centers, distinct security policies, ownership of data, and data quality processes, in addition to the impact of each of the four Vs requires architecture decisions.   So, this begs an important distributed processing architecture. Assuming multiple physical locations of large quantities of data, what is the design pattern for a secure, low-latency, possibly real-time, operational and analytic solution?

Big Data Discovery Process

We stated earlier that data *volume*, *velocity*, *variety* and *value* define Big Data, but the unique characteristic of Big Data is the *process* in which value is discovered. Big Data is unlike conventional business intelligence, where the simple reporting of a known value reveals a fact, such as summing daily sales into year-to-date sales. With Big Data, the goal is to be clever enough to discover patterns, model hypothesis, and test your predictions. For example, value is discovered through an investigative, iterative querying and/or modeling process, such as asking a question, make a hypothesis, choose data sources, create statistical, visual, or semantic models, evaluate findings, ask more questions, make a new hypothesis – and then start the process again. Subject matter experts interpreting visualizations or making interactive knowledge-based queries, can be aided by developing 'machine learning' adaptive algorithms that can further discover meaning. If your goal is to stay current with the pulse of the data that surrounds you, you will find that Big Data investigations are continuous. And your discoveries may result in one-off decisions or may become the new best practice and incorporated into operational business processes.

The architectural point is that the discovery and modeling processes must be fast and encourage iterative, orthogonal thinking. Many recent technology innovations enable these capabilities and should be considered, such as memory-rich servers for caches and processing, fast networks, optimized storage, columnar indexing, visualizations, machine learning, and semantic analysis to name a few. Your enterprise design goal should be to *discover and predict fast.*

Unstructured Data and Data Quality

Embracing data *variety*, that is, a variable schema in a variety of file formats requires continuous diligence. While variety offers flexibility, it also requires additional attention to understand the data, possibly clean and transform the data, provide lineage, and over time ensure that the data continues to mean what you expect it to mean. There are both manual and automated techniques to maintain your unstructured data quality. Examples of unstructured files: an XML file with an accompanying text-based schema declarations, text-based log files, standalone text, audio/video files, and key-value pairs – a two column table without predefined semantics.

For use cases with an abundance of public data sources, whether structured, semi-structured, or unstructured, you must expect that the content and structure of data to be out of your control. Data quality processes need to be automated. In the consumer products industry, as an example, social media comments not only come from predictable sources like your website and Facebook, but also the next trendy smartphone which may appear without any notice. In some of these cases, machine learning can help keep schemas current.

Mobility and Bring Your Own Device (BYOD)

Users expect to be able to access their information anywhere and anytime. To the extent that visualizations, analytics, or operationalized big data/analytics are part of the mobile experience, then these real-time and near real-time requirements become important architectural requirements.

Talent and Organization

A major challenge facing organizations is how to acquire a variety of the new Big Data skills. Apart from vendors and service partners augmenting staff, the most sought-after role is the data scientist — a role that combines domain skills in computer science, mathematics, statistics, and predictive modeling. By 2015, Gartner predicts that 4.4 million jobs will be created around big data. At a minimum, it is time to start cross-training your employees and soon - recruiting analytic talent. And lastly, organizations must consider how they will organize the big data function—as departmental resources or centralized in a center of excellence.

It is important to recognize that the world of analytics has its own academic and professional language. Due to this specialization, it is important to have individuals that can easily communicate among the analytics, business management and technical professionals. Business analysts will need to become more analytical as their jobs evolve to work closely with data scientists.

Organizational and Technical Resource Resistance to Change

Organizations implementing new Big Data initiatives need to be sensitive to the potential emotional and psychological impact to technical resources when deploying these new technologies. The implication of deploying new Big Data technologies and solutions can be intimidating to existing technical resources and fear of change, lack of understanding, or fear for job security could result in resistance to change, which could derail Big Data initiatives. Care should be taken to educate technical resources with traditional relational data skill sets on the benefits of Big Data solutions and technologies. Differences in architectural approaches, data loading and ETL processes, data management, and data analysis, etc. should be clearly explained to existing technical resources to help them understand how new Big Data solutions fit into the overall information architecture.

An Architecture Framework for Big Data

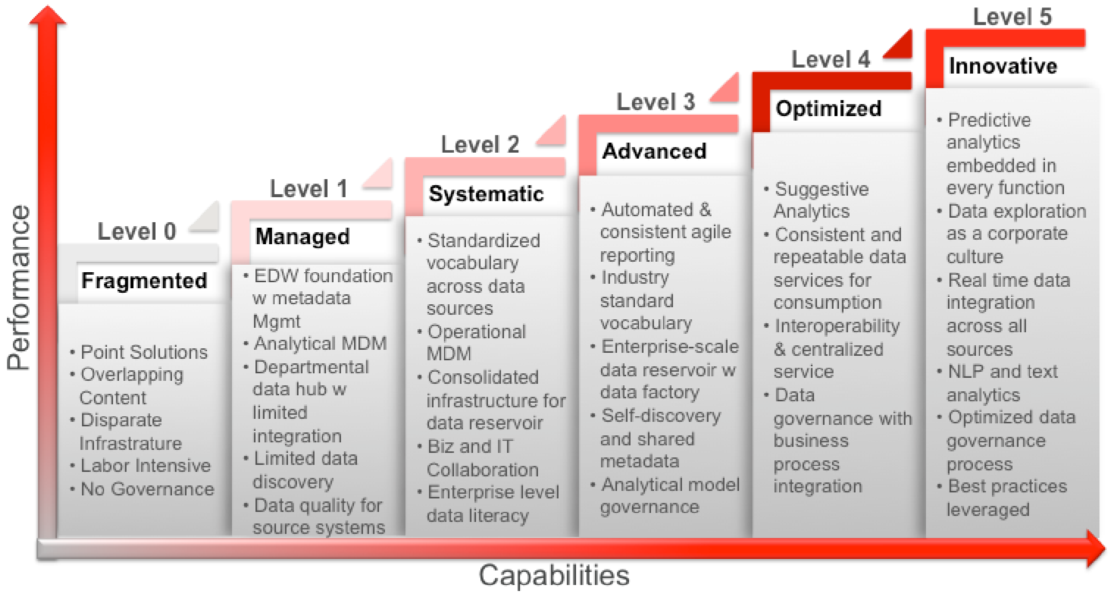
In this section, we will set up some perspective from making inroads into a Big Data-inclusive analytics culture. We will start by discussing an Oracle originated maturity model that can help you recognize where you are and what you need to do in order to reach your goals. And, then we will discuss Oracle's enterprise architecture approach for information architecture development.

Big Data and Analytics Maturity Model

Thomas H. Davenport was perhaps the first to observe in his Harvard Business Review article published in January 2006 (“Competing on Analytics”) how companies who orientated themselves around fact based management approach and compete on their analytical abilities considerably out-performed their peers in the marketplace.

The reality is that it takes continuous improvement to become an analytics-driven organization. To help customers build their roadmaps, Oracle developed the Big Data and Analytics Maturity Model. The Model helps customers determine a desired capability state, and can identify specific capabilities that are lacking or lagging and are therefore inhibiting Big Data and Analytics initiatives. A remediation approach for each of the identified inhibitors can then be determined from industry best practices and prior experiences. These remedies can be prioritized and used to create a plan, called the Big Data and Analytics Roadmap, to put Big Data and Analytics initiatives back on track.

Let’s take a detailed look at each stage of the Big Data and Analytics Maturity Model:



Performance
Level 0
Fragmented
• Point Solutions
• Overlapping Content
• Disparate Infrastrature
• Labor Intensive
• No Governance
Level 1
Managed
• EDW foundation w metadata Mgmt
• Analytical MDM
• Departmental data hub w limited integration
• Limited data discovery
• Data quality for source systems
Level 2
Systematic
• Standardized vocabulary across data sources
• Operational MDM
• Consolidated infrastructure for data reservoir
• Biz and IT Collaboration
• Enterprise level data literacy
Level 3
Advanced
• Automated & consistent agile reporting
• Industry standard vocabulary
• Enterprise-scale data reservoir w data factory
• Self-discovery and shared metadata
• Analytical model governance
Level 4
Optimized
• Suggestive Analytics
• Consistent and repeatable data services for consumption
• Interoperability & centralized service
• Data governance with business process integration
Level 5
Innovative
• Predictive analytics embedded in every function
• Data exploration as a corporate culture
• Real time data integration across all sources
• NLP and text analytics
• Optimized data governance process
• Best practices leveraged
Capabilities

Figure 1: Big Data and Analytics Capability Maturity Model

In the following six tables, we will describe each of these stages through the following perspectives: definition, characteristics, incorporating of data sources, architecture, infrastructure, organization and people, and state of governance.

BIG DATA AND ANALYTICS CAPABILITY MATURITY MODEL

Level 0 – Fragmented	
Definition	Level 0 is defined by fragmented point solutions with inefficient and inconsistent versions of the truth.
Characteristics	The main characteristic of this stage is the multitude of vendor-based and internally developed applications that are used to address specific analytic needs as they arise.
Data Sources	Overlapping data content leads to multiple versions of analytic truth
Architecture	The fragmented point solutions are neither co-located in a data warehouse nor architecturally integrated with one another.
Infrastructure	Technologies and tools are informal and not standardized. Decisions are heavily influenced by project team and/or technology-savvy individuals, instead of based on a set of commonly agreed upon guiding principles.
Organization and People	» Reports are labor intensive and inconsistent. » Analysts spend majority of their time assembling reports and crosscheck numbers to avoid mistakes and correct data quality issues manually. » Analytical skills are spread throughout different parts of the organization. Skill levels are inconsistent. Sharing of expertise and best practices are rare and ad hoc.
Governance	» Data governance is limited or non-existent. » Analytical and data standards do not exist.

Level 1 – Managed	
Definition	Level 1 is identified by the existence of enterprise data warehouse as the foundation for data analysis
Characteristics	» Searchable metadata repository is available across the enterprise. » Data warehouse is updated within one month of source system changes.
Data Sources	» Business, revenue, and financial data are co-located in a single data warehouse. » Unstructured data is not being leveraged, or is minimally leveraged for decision-making
Architecture	» Analytical master data capabilities are established for customers, products, and locations if applicable. » Integration capability is primarily ETL tool-based for enterprise data warehouse data loading needs. » Metadata is managed through the combination of data dictionary of the data warehouse, ETL tool repository, as well as the BI metadata repository. » Initial exploration of data hub and Hadoop at departmental level with limited integration capabilities » Few to none data exploration and discovery capabilities
Infrastructure	» The organization has started to set enterprise-wide standards for information management technologies. » IT organization has started to look into the consolidation and optimization of technology portfolios around business intelligence, data warehouse, and integration tools.
Organization and People	» The enterprise data warehouse team reports organizationally to the global CIO instead of regional or lines of business IT. » BI Center of Excellence starts to form in some organizations cross lines of businesses to share best practices and establish standards.
Governance	» Data governance is forming around the data quality of source systems. » Functional data stewardship is taking shape.

Level 2 – Systematic	
Definition	Level 2 is defined with standardized vocabulary across the enterprise and customer master data move from analytical only to focus on the operational level.
Characteristics	» Master data and reference data identified and standardized across disparate source system content in the data warehouse. » Naming, definition, and data types are consistent with local standards.
Data Sources	» Customer master data is defined and integrated at the operational level. » Additional Data Sources are incorporated into the enterprise data warehouse including costing model, supply chain information (if applicable), and customer experience, including some unstructured content.
Architecture	» Analytic objective is focused on consistent and efficient production of reports supporting basic management and operations of the organization. » Key performance indicators (KPI) are easily accessible from the executive level to the front-line managers and knowledge workers. » Standards on the data reservoir around Hadoop distribution and NoSQL are forming.
Infrastructure	» Information management technology services are optimized and consolidated with other infrastructure and application technology services into enterprise-wide, holistic, and integrated enabling services.
Organization and People	Corporate and business unit data analysts meet regularly to collaborate and prioritize new features for the enterprise data reservoir.
Governance	» Data governance forms around the definition and evolution of customer and other master data management including products, location/facilities (if applicable), and employees. » Data governance expands to raise the data literacy of the organization and develop a data acquisition strategy for additional Data Sources.

Level 3 – Advanced	
Definition	Level 3 is defined by automated reporting with efficient, consistent, and agile production.
Characteristics	» The main characteristic of this level is manifested in reduction of waste and customer variability. » There is more focus toward consistent and efficient production of reports required for regulatory requirements.
Data Sources	» Text data content is available for simple key word searches. » Precision of customer master is improved by including data from multiple touch points including mobile and social.
Architecture	» Enterprise data warehouse content is organized into standardized data marts. » Enterprise data reservoir capabilities are maturing with more data subject areas, data types, and integration capabilities at both ingestion and consumption levels. » Enterprise information discovery unifies the warehouse and reservoir
Infrastructure	Integration with existing systems across LOBs, virtualization, and user-centric tools.
Organization and People	» Permanent multi-disciplinary teams are in-place that continuously monitor opportunities to improve quality as well as reduce risk and cost across different business processes.
Governance	» Centralized data governance exists for review and approval of externally released data. » Metadata management includes self-discovery, user-friendly search, and self-publishing capabilities. » Data Governance functions expand to manage analytical model accuracy.

Level 4 – Optimized	
Definition	The key marker for Level 4 of analytical maturity is actionable analytics to improve operational and risk intervention
Characteristics	» The “accountable organization” shares in the financial risk and reward that is tied to business outcomes. » Analytics are available at the point of customer touch points and operational level to support the objectives of maximizing the quality of products, services, and customer care.
Data Sources	Data content expands to include telemetry and streaming Data Sources from internal and external content providers as applicable.
Architecture	» On average, the enterprise data warehouse is updated multiple times per day as source system data changes. » Data services are defined, built, managed, and published to provide consistent service-based data management capabilities.
Infrastructure	» Infrastructure capabilities focus on interoperability and service-oriented architecture. » Centralized services verses local autonomy is also balanced to achieve economy of scale and agility.
Organization and People	The enterprise data reservoir and analytical teams report organizationally to a C-level executive who is accountable for balancing cost and quality of products and services.
Governance	Data governance plays a major role in the accuracy of metrics supporting quality-based compensation plans for executives.

Level 5 – Innovative	
Definition	Level 5 of the analytical maturity model is defined in the terms of industry advances in product and service capabilities with mature and governed prescriptive analytics capabilities.
Characteristics	Knowledge-sharing and data exploration are core of the corporate culture.
Data Sources	Analytics expands to include natural language processing (NLP) of text, prescriptive analytics, and interventional decision support.
Architecture	» Prescriptive analytics are seamlessly embedded in every level of strategic and operational business decisions to maximize outcome. » The enterprise data warehouse is updated real time to reflect changes in the source systems.
Infrastructure	Information Architecture group actively review, evaluate technologies trends and jointly work with business to adopt new technologies and tools in evaluating effectiveness of initiatives such as creating new business models and driving new business capabilities.
Organization and People	Employees from various LOBs collaborate to share risk and reward.
Governance	» Governance councils, information owners, and data stewards have optimized processes in resolving problems pertaining to cross-functional information management issues. » Best practices are identified, documented, and communicated. » Information architecture team ensures that these best practices are extended across the enterprise through reference architecture development and sharing, as well as architecture review activities.

Taking an Enterprise Architecture Approach

Taking an enterprise architecture (EA) approach ensures that all technology decisions are driven from an agreed upon business context. This strategy may sound obvious, but often, IT departments make "the right" choices only to learn that the business has shifted its priorities. An EA approach maintains business alignment as you develop your architecture vision and build the corresponding investment roadmap. We recommend that this is the perspective you should take as you experiment and evolve your Big Data initiatives.

The Oracle Architecture Development Process (OADP) is the standardized methodology that ensures business alignment. This process is supported by planning artifacts, reference architectures, and tools managed from the Oracle Enterprise Architecture Framework (OEAF).

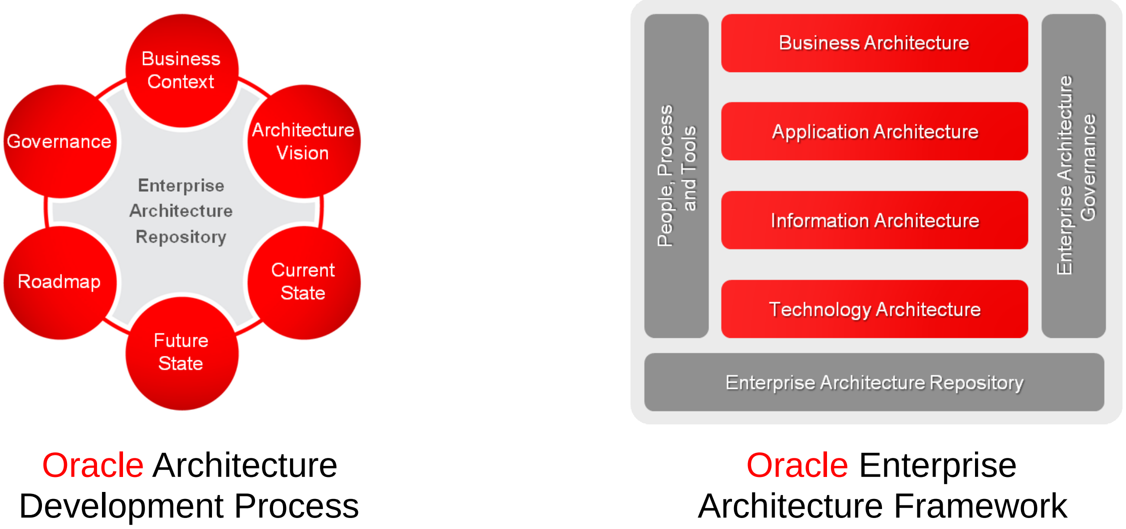

Business Context
Architecture Vision
Current State
Future State
Roadmap
Governance
Enterprise Architecture Repository
Oracle Architecture Development Process
People, Process and Tools
Business Architecture
Application Architecture
Information Architecture
Technology Architecture
Enterprise Architecture Governance
Enterprise Architecture Repository
Oracle Enterprise Architecture Framework

Figure 2:   Oracle Architecture Development Process (OADP) and Oracle Enterprise Architecture Framework (OEAF)

Oracle Enterprise Architecture Framework

The Oracle Enterprise Architecture Framework provides a streamlined, technology agnostic framework that helps Oracle collaboratively work with customers to develop strategic roadmaps and architecture solutions. Oracle emphasizes a “just enough” and “just in time” practical approach to Enterprise Architecture, which may be used standalone or as to complement a customer’s EA methodology. By focusing on business results and leveraging Oracle’s unique EA assets and reference architectures, the Oracle Enterprise Architecture Framework can be employed to efficiently create a strategic roadmap with sound technology solutions backed by a business case for business and IT alignment.

Big Data is supported in this framework with practitioner guides and reference architectures for Information Architecture and Business Analytics. You can download these white papers here: www.oracle.com/goto/ITStrategies

For more information about Oracle’s Information Architecture approach, please refer to Oracle Enterprise Architecture Framework: Information Architecture Domain white paper at the “Information Architecture and Governance” section at http://www.oracle.com/goto/ea.

Architecture Development for Big Data and Analytics

Oracle Architecture Development Process

The Oracle Architecture Development Process (OADP) is a practical methodology developed to ensure business results and not just articulate technology design. OADP's core operating principle is to achieve clarity and alignment between business objectives and IT commitments. And from there, all planning, roadmaps, and technology decisions are constrained by an agreed upon business strategy, accountability, and governance. This disciplined approach can be applied to individual segments in the business, or be used to align enterprise initiatives.

For larger transformation initiatives, OADP aids enterprise architects and planners to anticipate complexity, disruption, and risk. Using OADP, Oracle's Enterprise Architects can rely on proven planning and technology design patterns that accelerate projects. Oracle also use the OADP as the foundation for an internal education and certification program that teaches Oracle and industry best practices.

In the following section, we will drill into the elements of OADP as applied to Big Data.

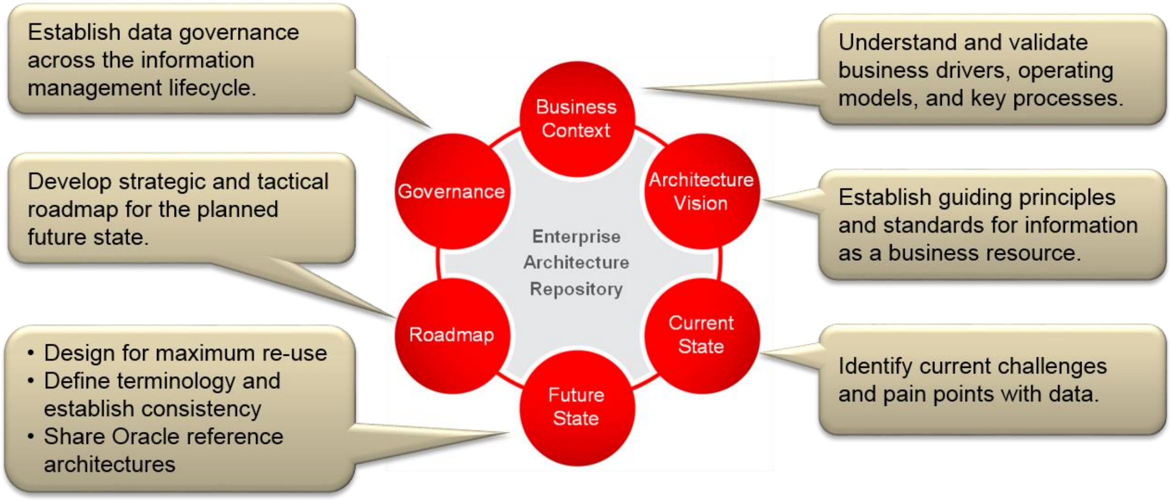

Establish data governance across the information management lifecycle.
Develop strategic and tactical roadmap for the planned future state.
• Design for maximum re-use
• Define terminology and establish consistency
• Share Oracle reference architectures
Business Context
Governance
Architecture Vision
Enterprise Architecture Repository
Roadmap
Current State
Future State
Understand and validate business drivers, operating models, and key processes.
Establish guiding principles and standards for information as a business resource.
Identify current challenges and pain points with data.

Figure 3: The overall process to develop a big data and analytical architecture, applying OADP.

The 6 main steps of the architecture development approach include:

- Align with business context
- Define architecture vision
- Assess current state architecture
- Establish future state architecture
- Determine strategic roadmap
- Implement governance

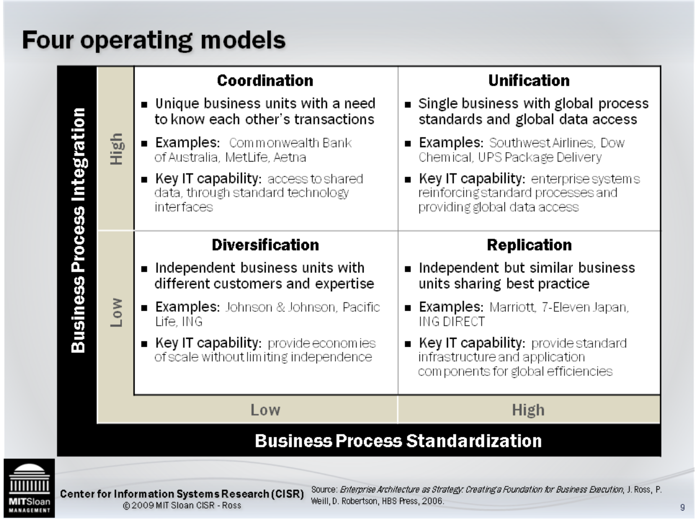
Align with Business Context

The Business Context phase is the first step in architecture development process. It is the most important phase in the process since it establishes the basis for the overall architecture discussion. It is critical to understand and align with the main business priorities and objectives that are driving your organization.

In the case of Big Data, every industry is stepping up to innovating with the new data coming from people and things enabled by *smart* consumer devices and *smart* products. Consequently, enterprise business processes are being reimagined and longer and faster value chains are emerging.

The initial business questions that set business context describe the business operating model. It is important to evaluate an organization's current and anticipated future state operating model to ensure that it will support the big data architecture recommendations. For example, a coordinated operating model might imply higher value in an investment in analytics as a service than investment in consolidated storage hardware platforms. Similarly, a diversified operating model has low level of standardization and data integration. It enables maximum agility in the business. Companies moving from a replicated operating model to a unified model will require significant change in breaking down the data-sharing barrier, which drives the need for enhanced data integration capabilities. It is important to include business model operational realities and contingencies in your architecture planning.

The business operating model is a key element of business architecture. It is used to define the degree of business process standardization versus the degree of business process and data integration. Jeanne Ross in her book "Enterprise Architecture as Strategy: Creating a Foundation for Business Execution" discusses the four types of operating models: Diversification, Replication, Coordination, and Unification.



Four operating models

Business Process Integration	Low (Business Process Standardization)	High (Business Process Standardization)
High	**Coordination** ■ Unique business units with a need to know each other's transactions ■ **Examples:** Commonwealth Bank of Australia, MetLife, Aetna ■ **Key IT capability:** access to shared data, through standard technology interfaces	**Unification** ■ Single business with global process standards and global data access ■ **Examples:** Southwest Airlines, Dow Chemical, UPS Package Delivery ■ **Key IT capability:** enterprise systems reinforcing standard processes and providing global data access
Low	**Diversification** ■ Independent business units with different customers and expertise ■ **Examples:** Johnson & Johnson, Pacific Life, ING ■ **Key IT capability:** provide economies of scale without limiting independence	**Replication** ■ Independent but similar business units sharing best practice ■ **Examples:** Marriott, 7-Eleven Japan, ING DIRECT ■ **Key IT capability:** provide standard infrastructure and application components for global efficiencies

Center for Information Systems Research (CISR)
© 2009 MIT Sloan CISR - Ross

Source: *Enterprise Architecture as Strategy: Creating a Foundation for Business Execution*, J. Ross, P. Weill, D. Robertson, HBS Press, 2006.

Figure 4: Business Operating Models described by Jeanne Ross, MIT Sloan

Examples of business contexts are included in these big data industry overviews: Communications Service Providers, Education, Financial Services, Healthcare Payers and Providers, Logistics, Manufacturing, Media and Entertainment. Oil & Gas, Pharmaceuticals and Life Sciences, Retail, and Utilities.

Oracle can help you articulate your business context by developing the following assets:

- *Business Architecture Models*: These models describe the business goals and objectives, business functions, and the organizations involved.

- *Business Architecture Principles:* The Architecture Principles Map is a collection of principles organized by the domains of an EA framework (Business, Application, Information, Technology, and Governance). The principles and their relationships serves as a foundation for ongoing decision-making.
- *Strategy Map and Balanced Score Card:* The Strategy Map and Balanced Score Card allow for the identification of key knowledge perspectives that contribute to the organizations business goals and objectives. This provides the ability to define knowledge and performance metrics to capture and evaluate the contribution from the addition of Big Data capabilities to the organizations IT portfolio.
- *Information Asset Inventory:* The Information Asset Inventory provides an organizational level perspective of the core information assets that drives decision making and supports core business processes. Understanding how information flows through the organization and who shares the information is an essential element in understanding how changes and improvements to the information architecture will support the operations of the business.
- *Big Data Value Hypothesis:* Development of a formal value hypothesis provides the enterprise with an early view into the benefits to be achieved through the adoption of Big Data and ancillary changes to the information architecture in the context of the organization strategy and objectives. Provides the enterprise with gross analysis of projected value for the investment in these new capabilities.

Define Architecture Vision

The Architecture Vision phase is the second step in the architecture development process. It relies on the business context previously defined and maps it to a supporting IT architecture. Big Data and Information Management cover a wide range of topics and interests. In this phase, decisions about the scope of the following items should be articulated:

ARCHITECTURE VISION

Vision Content	Architecture Principles
Alignment with Enterprise Information Architecture	Big Data analytics is maximized when it can be integrated and aligned with existing enterprise information stores and business and analytic processes.
Actionable Insights	Analysis is most effective when appropriate action (human or automated) can be taken at the time of event, insight, or discovery.
Business Process Integration	Analytics architecture (reservoir, warehouse, access) should be designed into critical business process flows for operational, knowledge, and analytics users.
Business Continuity	Analytics processes and applications availability must be aligned with business processes
Widespread Access	Information and analysis should be available to all roles, processes, and applications. The architecture must enable end users who are not familiar with data structures and analytical tools to search and view information pertinent to their needs and authorization.
Data Quality and Metadata	Standardizing and sharing metadata at the enterprise level improves the accuracy of analysis, reduced costs of reconciliation, and improves communication.
Data Service Provisioning	Analytics data and systems should be available as a shared service in the IT catalog
Security	Stratify data assets based on security requirements and sensitivity; apply tight controls through defense-in-depth. Provide forensic analysis capabilities to detect, report, and prevent uncharacteristic usage and access
Infrastructure Availability	Analytic processes and systems are Tier 1 and require an appropriate SLA
Context Based Governance	Establish data governance practices for all data-including unstructured.

Oracle can help you articulate your architecture vision by developing the following assets:

- *Architecture Vision:* A view of IT architecture which supports the business architecture created in the previous phase.
- *Architecture Maturity Assessment:* An assessment of their current capabilities and ability to achieve their business objectives evaluating current IT capabilities, IT management structures, funding models, and governance. Make a recommendation and a hypothesis for where the enterprise needs to be, in terms of maturity level, to be able to meet their business objectives.
- *Architecture Principles:* Translate business architecture principles into the application, information, and technology architecture domains. The alignment of business and IT principles helps guide future decisions.
- *Architecture Scope*: Define the concrete scope of the architecture engagement.
- *Information Architecture Domain Model:* Capture the key capabilities and assets within the Information Architecture that will be added or changed and how they contribute and support the business strategy and objectives.
- *Conceptual Architecture*: One or more conceptual architecture models may be defined and presented to help communicate the high level relationships between the proposed capabilities to be implemented in support of the architecture vision. A conceptual model provides a basis for a systemic implementation of capabilities without specific details on how the technology will be implemented.

Assess Current State Architecture

The Current State Architecture phase assesses the key analytical and reporting processes, existing data and infrastructure investments, information flows, resource skills, and governance practices in the context of the architecture vision. This is particularly important since the Big Data components, such as the Data Reservoir and Discovery Labs will need to integrate with existing data stores, other BI/Analytics investments, and other infrastructure.

Oracle's approach for current state architecture focuses the efforts on high value targets and iteratively drills down into details as needed. This practical approach avoids collecting unnecessary details of the current state and focuses on the minimum required to accurately develop an architecture plan and build a business case for change in future phases.

Oracle can help you assess your current state by developing the following assets:

- *Current Business Architecture:* Inventories and relationships around the business functions, processes, and organizations around the problem or initiative we are addressing
- *Current Application Architecture*: The set of applications and processes supporting a particular architecture scope
- *Current Information Architecture:* An information asset inventory, a high-level enterprise data model and the data flow patterns around a particular architecture scope.
- *Information Flow Diagrams:* Diagrammatic depiction and the major flows of an information asset during its lifecycle. Shows data sources, data integration, mapping/transformation and consumption. Can also be extended to show organizational ownership and governance.
- *Information Asset Inventory:* Identify and capture key information assets and sources in order to define a baseline architecture view. Provides the ability to define asset and capability relationships to support the integration of Big Data services within the current architecture.
- *Current Technology Architecture:* An inventory and physical architectures of the technical services and capabilities supporting the information and application architectures

Establish Future State Architecture

The Future State Architecture phase describes a specific set of technical recommendations or blueprint. This phase relies upon reference architectures and design patterns. Oracle's long history in information management has allowed its architects to inventory best practices across of spectrum of use cases and with consideration of specific Oracle product capabilities. In the *Oracle Big Data Platform* section of this paper is an overview of the primary product components. And, the following two documents describe Oracle's reference architecture in detail: Information Management and Big Data Reference Architecture (30 pages) and Information Management Reference Architecture (200 pages). In this section, we will just highlight the need to establish a blueprint to unify your information landscape.

Learning from the Past

Most organizations have data located across a large number of heterogeneous data sources. Analysts spend more time finding, gathering and processing data than analyzing it. Data quality, access, and security are very key issues for most organizations. The most common challenge analysts' face is data collection across multiple systems, and the tasks of cleaning, harmonizing, and integrating it. It is estimated that 80% of the time analysts spend is around data preparation. Just imagine if they spent that time finding new insights.

According to The Data Warehouse Institute (TDWI), the trend in data warehouse architecture is the movement toward more diversified data platforms and related tools within the physical layer of the extended data warehouse environment. While there is a clear benefit to a centralized physical data warehousing strategy, more and more organizations are looking to establish a logical data warehouse, which is composed of a data reservoir that combines different databases, relational, NoSQL, and HDFS.

The logical architecture is supported by automated data federation and data virtualization with tooling to ensure schema mapping, conflict resolution, and autonomy policy enforcement. Success comes from a solid logical design based mostly on business structures. In our experience, the best success occurs when the logical design is based primarily on the business— how it is organized, its key business processes, and how the business defines prominent entities like customers, products, and financials.

The key is to architect a multi-platform data environment without being overwhelmed by its complexity, which a good architectural design can avoid. Hence, for many user organizations, a multi-platform physical plan coupled with a cross-platform logical design is a new data platform and architecture that's conducive to the new age of big data and analytical solutions.

Here are some guidelines on building this hybrid platform.

- First, define logical layering, and organize and model your data asset according to purpose, business domain, and existing implementation.
- Integrate between technologies and layers so both IT-based and user-based efforts are supported.
- Consolidate where it makes sense and virtualize access as best as possible, so that you can reduce the number of physical deployments and operational complexity.

The logical architecture below represents the desired blueprint for Big Data featuring the data reservoir, discovery labs, and a progression of data staging layers that protect raw data and enable data consumers to access the data they need.

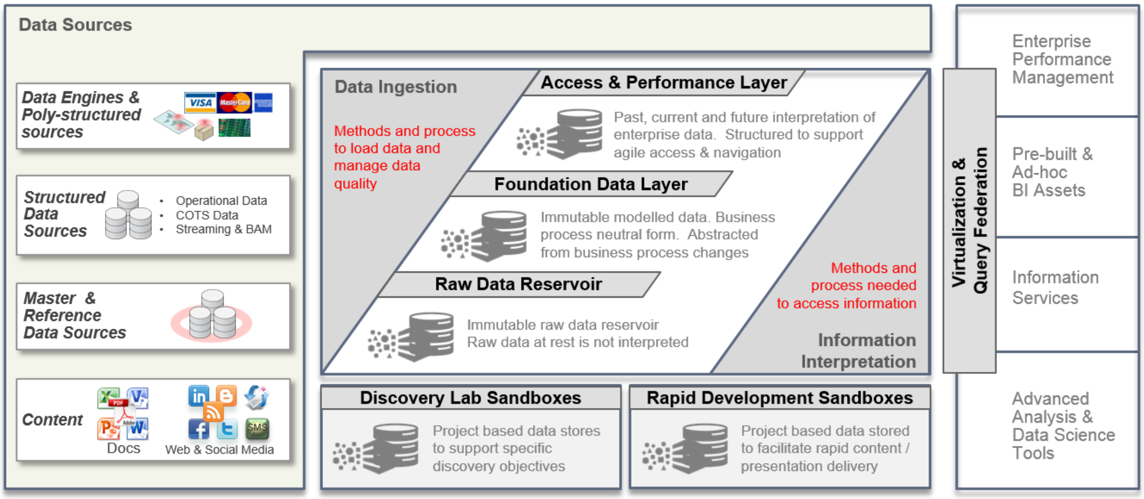

Data Sources
Data Engines & Poly-structured sources
Structured Data Sources
Operational Data
COTS Data
Streaming & BAM
Master & Reference Data Sources
Content
Docs
Web & Social Media
Data Ingestion
Methods and process to load data and manage data quality
Access & Performance Layer
Past, current and future interpretation of enterprise data. Structured to support agile access & navigation
Foundation Data Layer
Immutable modelled data. Business process neutral form. Abstracted from business process changes
Raw Data Reservoir
Immutable raw data reservoir
Raw data at rest is not interpreted
Methods and process needed to access information
Information Interpretation
Discovery Lab Sandboxes
Project based data stores to support specific discovery objectives
Rapid Development Sandboxes
Project based data stored to facilitate rapid content / presentation delivery
Virtualization & Query Federation
Enterprise Performance Management
Pre-built & Ad-hoc BI Assets
Information Services
Advanced Analysis & Data Science Tools

Figure 5: Oracle's Information Management Reference Architecture: Information Transformation and Staging Layers

Here are the primary data staging layers of an information management platform:

- **Raw Data Reservoir layer**: a landing pad for raw data at rest.
- **Foundation layer:** focuses on historic data management at the atomic level. Process-neutral, it represents an enterprise data model with a holistic vision.
- **Access and Performance layer:** also known as analytical layer, with business/functional-specific models, snapshots, aggregations, and summaries.
- **Discovery Lab Sandboxes:** where data exploration occurs with a significant amount of user autonomy.

It takes careful consideration to decide on the placement of data assets into the each of the layers. Factors to consider include data subject area, level of data detail and granularity, and level of time relevance (whether or not it is operational or historic). The focus of consideration needs to be on data and information consumption. These layers will correspond to different types of data storage and retrieval mechanism with most effective integration capabilities, performance characteristics, and security measures.

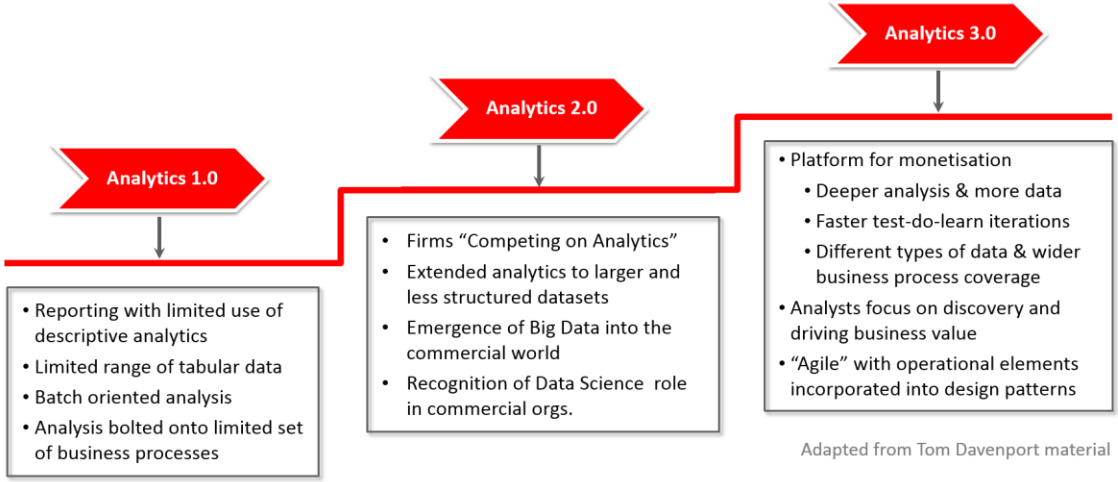
Oracle can help you design and evaluate your planned future state by developing the following assets:

- The future state architecture artifacts
- A set of architectural gaps that exist between the current and future states
- A set of architectural recommendations and a cost-benefit analysis for those recommendations

Determine Strategic Roadmap

The Roadmap phase creates a progressive plan to evolve toward the future state architecture. Key principles of the roadmap include technical and non-technical milestones designed to deliver business value and ultimately meet the original business expectations.

Companies vary in their level of big data and analytical maturity, but some say most are eager for the monetization of this new data. Examples include: increasing revenue by improving customer intimacy through finer tuned marketing, interpreting social sentiment, text, geo-fencing with mobile devices, and having products automatically request service calls. However, until your use cases come knocking on your door, the trick to respond quickly is to understand your current information architecture and develop the skills you need in the new technologies. The potential benefits are compelling and companies are motivated to improve their analytic strength. The illustration below contains an overview of an incremental approach to improving business analytical capabilities overtime.


Analytics 1.0
• Reporting with limited use of descriptive analytics
• Limited range of tabular data
• Batch oriented analysis
• Analysis bolted onto limited set of business processes
Analytics 2.0
• Firms "Competing on Analytics"
• Extended analytics to larger and less structured datasets
• Emergence of Big Data into the commercial world
• Recognition of Data Science role in commercial orgs.
Analytics 3.0
• Platform for monetisation
• Deeper analysis & more data
• Faster test-do-learn iterations
• Different types of data & wider business process coverage
• Analysts focus on discovery and driving business value
• "Agile" with operational elements incorporated into design patterns

Adapted from Tom Davenport material

Figure 6: An Incremental Approach to Analytical Maturity

If you are new to Big Data, the initial decisions are which business area to invest in first and how to build up these capabilities. The rule of thumb is to choose a business area that is mature for success based on business priority, perceived value and return of investment, availability of required historic data sources to produce meaningful insights, and adequate skills level of analysts. For the architect, the most common activities include use case development and prioritization, proof of concept with a few use cases, building of a small Hadoop cluster or NoSQL database to supplement the existing data warehouse, and initiation of data discovery capabilities to show early business benefit. The key is to take small steps, and be aware of the overall future state architecture, focus on quick-time-to market, and allow for continue adoption of open source architecture and capabilities.

As the business develops confidence in new processes and technologies, then you can expect to increase operational scale. More data sources are introduced that require expansion of the overall platform. Integration amongst these components (between RDBMS and HDFS, for example) is a critical capability to ensure success. Data discovery becomes enterprise-scale and mainstream. This is the phase when data and analytics as a service will mature.

Eventually, once the mechanics of data acquisition and quality are mastered, and the data has been integrated into business processes, dashboards, and discovery labs, you have reshaped your business into an analytics driven culture. With this solid foundation, the enterprise is fully prepared to adopt cutting-edge solutions and embrace new analytical trends.

Oracle can help you build a roadmap that includes:

- A set of architectural gaps that exist between the current and future states
- A cost-benefit analysis to close the gaps
- Maximizes the value from each phase of the roadmap
- Minimizes the risk and cost
- Considers technology dependencies across phases
- Provides the flexibility to adapt to new business priorities and to changing technology.

Implement Governance

The Governance phase is designed to ensure a successful transition to the desired Future State Architecture. The secret to managing governance is to ensure alignment across the overall business objectives, architecture principles, and the roadmap. OADP has been designed with this in mind. However, with Big Data Analytics, there is an additional cautionary note, around data governance.

As Daniel J. Boorstin pointed out, "The greatest obstacle to discovery is not ignorance - it is the illusion of knowledge." With the rise of machine learning, data mining, and other advanced analytics, IDC predicts that "decision and automation solutions, utilizing a mix of cognitive computing, rules management, analytics, biometrics, rich media recognition software and commercialized high-performance computing infrastructure will *proliferate*."

So data governance, in the context of big data, will need to focus on determining when automated decisions are appropriate and when human intervention and interpretation are required. Using strict data precision rules on user sentiment data might filter out too much useful information, whereas data standards and common definitions are still critical for fraud detections scenarios. Quality standards need to be based on the nature of consumption. In summary, the focus and approach for data governance need to be relevant and adaptive to the data types in question and the nature of information consumption.

People and Skills

Success with Big data and analytics requires more than just mastering technology. In order to put the data to good use, organizations must increase their ability to communicate analytically, that is, articulate business problems in the context of data, then apply statistical modeling, understand results, and make refinements.

Success depends on a diverse and well-orchestrated team of people. However, a "skills gap" is one of the largest barrier to success with new data and analytical platforms. Various sources including US Bureau of Labor Statistics, US Census, Dun & Bradstreet, and McKinsey Global Institute analysis, indicated that by 2018, the demand for deep analytical talent in the US could be 50 to 60 percent greater than the supply. According to TDWI, majority of the organizations investing in big data and analytics intend to upgrade their analytical strengths by improving the skills of existing staff and hiring new analytical talent.

Emerging roles and responsibilities typically required to support Big Data solutions:

BIG DATA ROLES AND RESPONSIBILITIES

Roles	Responsibilities
Big Data Champion	Big data and analytics are powerful business resources and need a leader to champion and oversee their usage. Early on, champions typically came from the IT organization. However, the shift from viewing big data as a technology to a business enabler for new market strategies and business models has created a shift in ownership of big data initiatives to marketing and product development executives.
Information Architects and Enterprise Architects	Data warehouses and the need for data quality, integration, and metadata are not going away. Rather, big data and analytics are becoming extensions of the existing information architecture. Information architects will be challenged to design and support an ecosystem now consisting of diversified information assets and related tools. Enterprise architects must collaborate with Information Architects to develop a cross-platform design that implements the expanded information architecture capabilities and does not create a big data silo for the organization.
Data Scientists	Data Scientist uses scientific methods to solve business problems using available data. Data Scientists will mine data, apply statistical modeling and analysis, interpret the results, and drive the implication of data results to application and to prediction.

Roles	Responsibilities
Business Analysts	Business analysts use their business knowledge to help the data scientist understand the business context and the business understand the analytic results. As data and models become more sophisticated, the need to synthesize the two becomes critical
Hadoop Administrators	Administrators responsible for managing Hadoop distributions and platforms
Java / Hadoop Programmers	Java / Hadoop programmers experienced in writing MapReduce code, PIG, and other languages required to support Big Data platforms

In order to be most effective, a big data team must be connected by a common mission. An emerging organizational best practice is the creation of a Big Data and Analytics Center of Excellence (COE) designed to facilitate communication, collaboration, and significantly increase productivity on behalf of the organization. The COE is a cross-functional team of analytic and domain specialists (i.e. the roles described above) who plan and prioritize analytics initiatives, manage and support those initiatives, and promote broader use of information and analytics best practices throughout the enterprise. The COE typically does not have direct ownership of all aspects of the big data and analytics framework; instead, it provides coordination and execution for strategic, enterprise, and line of business initiatives – both from an infrastructure and support perspective as well as governance perspective. The key benefits experienced by organizations that implement a COE are:

- Cross-training of professionals from diverse data science disciplines
- Data Scientist/Business Analyst alignment - provides advanced analytics professionals with a forum where they can partner with business-savvy subject matter experts to promote faster iterative processes
- Mobilizes resources for the good of the organization and has the flexibility and scalability to serve enterprise-centric initiatives as well as specific business units
- Ultimately changes the culture of the organization to appreciate the value of analytics-driven decisions and continuous learning.

Big Data Reference Architecture Overview

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.

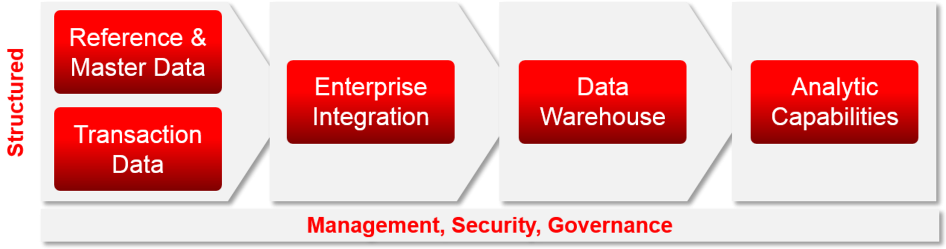


Figure 7: Traditional Information Architecture Components

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 enterprise architecture oversight responsibility is to establish and maintain a balanced governance approach including using a 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 storage 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 or semi-structured databases, sent to mobile devices, or merged into traditional data warehousing environment and correlated to structured data.

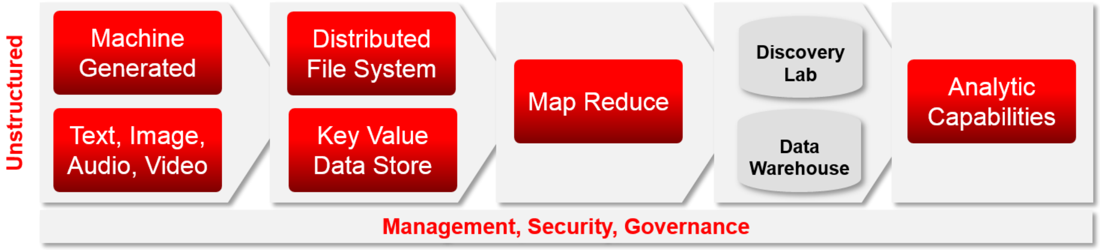


Figure 8: Big Data Information Architecture Components

Many new analytic capabilities are available that derive meaning from new, unique data types as well as finding straightforward statistical relevance across large distributions. Analytical throughput also impacts the transformation, integration, and storage architectures, such as real-time and near-real time events, ad hoc visual exploration, and multi-stage statistical models. Nevertheless, it is common after Map Reduce processing to move the "reduction result" into the data warehouse and/or dedicated analytical environment in order to leverage existing investments and skills in business intelligence reporting, statistical, semantic, and correlation capabilities. Dedicated analytical environments, also known as Discovery Labs or sandboxes, are architected to be rapidly provisioned and de-provisioned as needs dictate.

One of the obstacles observed in enterprise Hadoop adoption is the lack of integration with the existing BI eco-system. As a result, the analysis is not available to the typical business user or executive. When traditional BI and big data ecosystems are separate they fail to deliver the value added analysis that is expected. Independent Big Data projects also runs the risk of redundant investments which is especially problematic if there is a shortage of knowledgeable staff.

A Unified Reference Architecture

To draw out these ideas a bit further, let's consider an integrated information architecture.

Oracle's Information Management Architecture is shown below with key components and flows. One highlight is the separation and integration of the Discovery Lab alongside various forms of new and traditional data collection. See the reference architecture white paper for a full discussion. Click here. Click here for an Oracle product map.

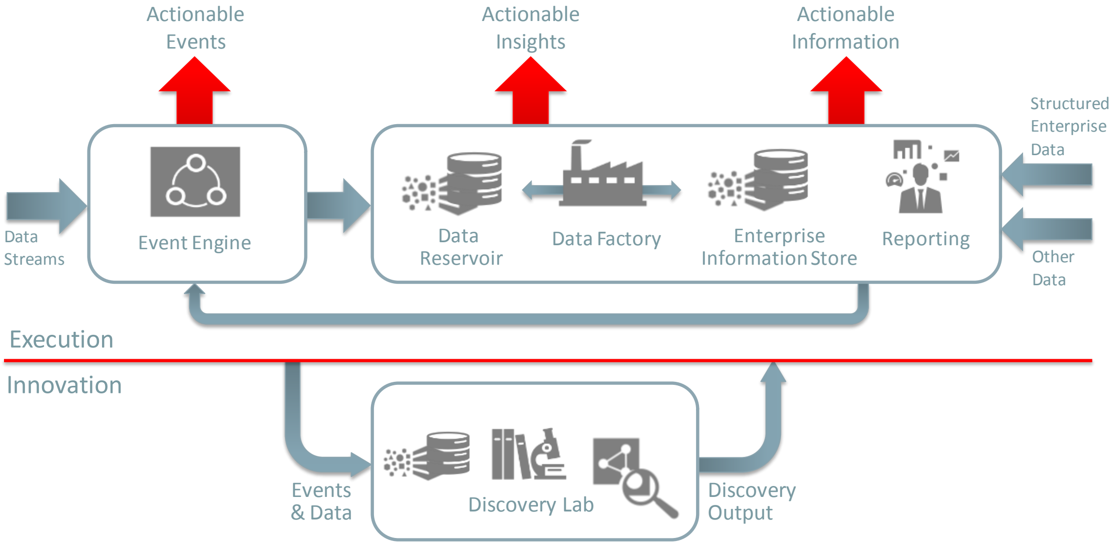


Figure 9: Conceptual Model for *The Oracle Big Data Platform* for unified information management and Big Data

A description of these primary components follows:

» **Event Engine:** Components which process data in-flight to identify actionable events and then determine *next-best-action* based on decision context and event profile data and persist in a durable storage system. The decision context relies on data in the data reservoir or other enterprise information stores.

- **Data Reservoir:** Economical, scale-out storage and parallel processing for data which does not have stringent requirements for data formalization or modelling. Typically manifested as a Hadoop cluster or staging area in a relational database.
- **Data Factory:** Management and orchestration of data into and between the Data Reservoir and Enterprise Information Store as well as the rapid provisioning of data into the Discovery Lab for agile discovery.
- **Enterprise Information Store**: Large scale formalized and modelled business critical data store, typically manifested by a Data Warehouse or Data Marts.
- **Reporting:** Business Intelligence, faceted navigation, and data mining analytic tools including dashboards, reports, and mobile access for timely and accurate reporting.
- **Discovery Lab:** A set of data stores, processing engines, and analysis tools separate from the everyday processing of data to facilitate the discovery of new knowledge. This includes the ability to provision new data into the Discovery Lab from any source.

The interplay of these components and their assembly into solutions can be further simplified by dividing the flow of data into execution -- tasks which support and inform daily operations -- and innovation – tasks which drive new insights back to the business. Arranging solutions on either side of this division (as shown by the horizontal line) helps inform system requirements for security, governance, and timeliness.

An additional comment about data flow:

> *"The reality is often more complex than these well-defined situations, and it's not uncommon for the data flow between Hadoop and the relational database to be described by circles and arcs instead of a single straight line. The Hadoop cluster may, for example, do the preprocessing step on data that is then ingested into the RDBMS and then receive frequent transaction dumps that are used to build aggregates, which are sent back to the database. Then, once the data gets older than a certain threshold, it is deleted from the database but kept in Hadoop for archival purposes. Regardless of the situation, the ability to get data from Hadoop to a relational database and back again is a critical aspect of integrating Hadoop into your IT infrastructure."*

Source: Hadoop Beginner's Guide, Chapter 9: Working with Relational Databases, Packt Publishing, 2013, Web ISBN-13: 978-1-84951-731-7

Enterprise Information Management Capabilities

Drilling a little deeper into the unified information management platform, here is Oracle's holistic capability map:

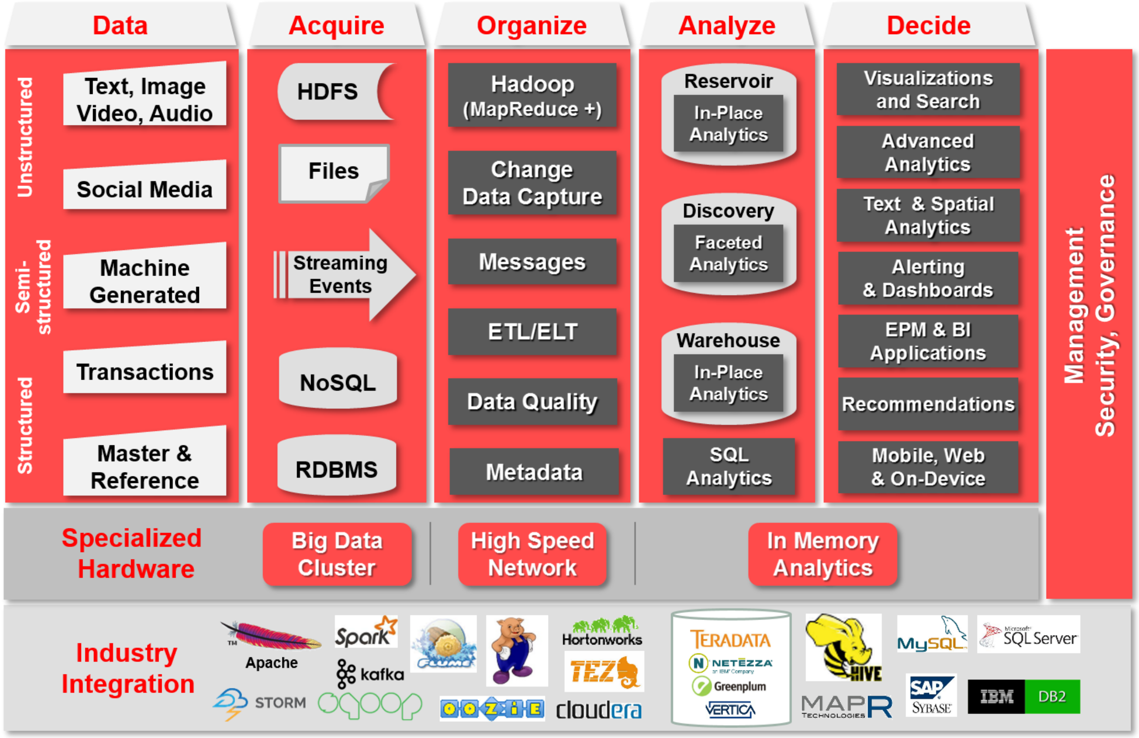


Figure 10: Oracle's Unified Information Management Capabilities

A brief overview of these capabilities beginning on the left hand side of the diagram.

As various data types are ingested (under Acquire), they can either be written directly (real-time) into memory processes or can be written to disk as messages, files, or database transactions. Once received, there are multiple options on where to persist the data. It can be written to the file system, a traditional RDBMS, or distributed-clustered systems such as NoSQL and Hadoop Distributed File System (HDFS). The primary techniques for rapid evaluation of unstructured data is by running map-reduce (Hadoop) in batch or map-reduce (Spark) in-memory. Additional evaluation options are available for real-time streaming data.

The integration layer in the middle (under Organize) is extensive and enables an open ingest, data reservoir, data warehouse, and analytic architecture. It extends across all of the data types and domains, and manages the bi-directional gap between the traditional and new data acquisition and processing environments. Most importantly, it meets the requirements of the four Vs: extreme volume and velocity, variety of data types, and finding value where ever your analytics operate. In addition, it provides data quality services, maintains metadata, and tracks transformation lineage.

The Big Data processing output, having converted it from low density to high density data, will be loaded into the a foundation data layer, data warehouse, data marts, data discovery labs or back into the reservoir. Of note, the discovery lab requires fast connections to the data reservoir, event processing, and the data warehouse. For all of these reasons a high speed network, such as Infiniband provides data transport.

The next layer (under Analyze) is where the “reduction-results” are loaded from Big Data processing output into your data warehouse for further analysis. You will notice that the reservoir and the data warehouse both offer ‘in-place’ analytics which means that analytical processing can occur on the source system without an extra step to move the data to another analytical environment. The SQL analytics capability allows simple and complex analytical queries optimally at each data store independently and on separate systems as well as combining results in a single query.

There are many performance options at this layer which can improve performance by many orders of magnitude. By leveraging Oracle Exadata for your data warehouse, processing can be enhanced with flash memory, columnar databases, in-memory databases, and more. Also, a critical capability for the discovery lab is a fast, high powered search, known as faceted navigation, to support a responsive investigative environment.

The Business Intelligence layer (under Decide) is equipped with interactive, real-time, and data modeling tools. These tools are able to query, report and model data while leaving the large volumes of data in place. These tools include advanced analytics, in-database and in-reservoir statistical analysis, and advanced visualization, in addition to the traditional components such as reports, dashboards, alerts and queries.

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

With a unified architecture, the business and analytical users can rely on richer, high quality data. Once ready for consumption, 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.

Big Data Architecture Capabilities

Here is a brief outline of a few Apache projects (www.apache.org) and Oracle Big Data products. A complete product listing is included in *The Oracle Big Data Platform* product table. Click here.

Ingest Capability

There are a number of methods to introduce data into a Big Data platform.

Apache Flume (Click for more information)

- A distributed, reliable, and available service for efficiently moving large amounts of log data and other data.
- Captures and processes data asynchronously. A data event captures data in a queue (channel) and then a consumer dequeues the event (sink) on demand. Once consumed, the data in the original queue is removed which forces writing data to another log or HDFS for archival purposes.
- Data can be reliably advanced through multiple states by linking queues (sinks to channels) with 100% recoverability
- Data can be processed in the file system or in-memory. However, in memory processing is not recoverable.

Apache Storm (Click for more information)

- A distributed real-time, parallelized computation system that runs across a cluster of machines.
- Topology consumes streams of data and processes those streams in arbitrarily complex ways, repartitioning the streams between each stage of the computation
- Use cases: real-time analytics, online machine learning, continuous computation, distributed RPC, ETL, and more.

Apache Kafka (Click for more information)

- A publish-subscribe messaging system.
- Messages are immediately written to file system and replicated within the cluster to prevent data loss.
- Messages are not deleted when they are read but retained with a configurable SLA
- A single cluster serves as the central data backbone that can be elastically expanded without downtime.

Apache Spark Streaming: (Click for more information)

- *Spark Streaming.* Spark Streaming is an extension of Spark. It extends Spark for doing large scale stream processing, and is capable of scaling to 100's of nodes and achieves second scale latencies. Spark Streaming supports both Java and Scala, which makes it easy for users to map, filter, join, and reduce streams (among other operations) using functions in the Scala/Java programming language. It integrates with Spark's batch and interactive processing while maintaining fault tolerance similar to batch systems that can recover from both outright failures and stragglers. In addition, Spark streaming provides support for applications with requirements for combining data streams with historical data computed through batch jobs or ad-hoc queries, providing a powerful real-time analytics environment.

Oracle Event Processing (Click for more information)

- Processes multiple event streams to detect patterns and trends in real time
- Offers flexible deployment options: standalone, integrated in the SOA stack, lightweight on embedded Java
- Provides visualization into emerging opportunities or risk mitigation through seamless integration with Oracle Business Activity Monitoring
- Ensures downstream applications and service-oriented and event-driven architectures are driven by true, real-time intelligence

Oracle Golden Gate (Click for more information)

- Log-based change data capture, distribution, transformation, and delivery
- Support for heterogeneous databases and operating systems
- Bidirectional replication without distance limitation
- Transactional integrity
- Reliable data delivery and fast recovery after interruptions

Storage and Management Capability

Hadoop Distributed File System (HDFS): (Click for more information)

- An Apache open source distributed file system
- 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: (Click for more information)

- Cloudera Manager is an end-to-end management application for Cloudera's Distribution of Apache Hadoop
- 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.

Data Management Capability

Oracle NoSQL Database: (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.
- Only contains metadata that describes data access in Apache HDFS and Apace HBase, not the data itself.
- Uses a simple SQL-like query language called HiveQL
- Query execution via MapReduce

Processing Capability

MapReduce/MapReduce 2:

- 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/MapReduce 2 implementation
- Highly scalable parallel batch processing
- Highly customizable infrastructure
- Writes multiple copies across cluster for fault tolerance

Apache Spark:

- Memory based MapReduce. Operates on HDFS directly or through Hadoop

- Faster and more expressive than MapReduce
- Enables real-time streaming workloads

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

SQL Data Access

Oracle Big Data SQL (Click for more information)

- Provides single, optimized SQL query for distributed data
- Supports multiple data sources, including Hadoop, NoSQL and Oracle Database
- Includes automatic, extensible Oracle Database external table generation
- Provides Smart Scan on Hadoop to minimize data movement and maximize performance
- Ensures advanced security for Hadoop and NoSQL data (redaction, virtual private database)

Data Discovery

Oracle Big Data Discovery (Click for more information)

- Interactive discovery capabilities reading Hadoop, Hive, and NoSQL data
- A single, easy to use product, built natively on Hadoop to transform raw data into business insight in minutes, without the need to learn complex products or rely only on highly skilled resources.

Statistical Analysis Capability

Open Source Project R and Oracle R Enterprise (part of Oracle Advanced Analytics:

- 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

Highlights of Oracle’s Big Data Architecture

There are a few areas to explore further in Oracle’s product capabilities:

Big Data SQL

Oracle Big Data SQL provides architects a great deal of flexibility when making decisions about data access, data movement, data transformation, and even data analytics. Rather than having to master the unique native data access methods for each data platform, Big Data SQL standardizes data access with Oracle’s industry standard SQL. It also inherits many advanced SQL analytic features, execution optimization, and security capabilities. Big Data SQL honors a key principle of Big Data – bring the analytics to the data. By reducing data movement, you will obtain analytic results faster.

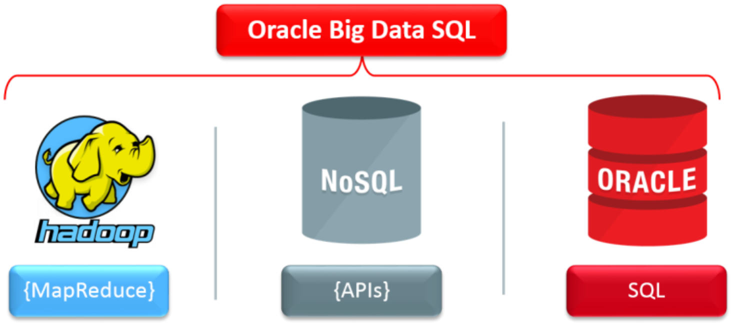
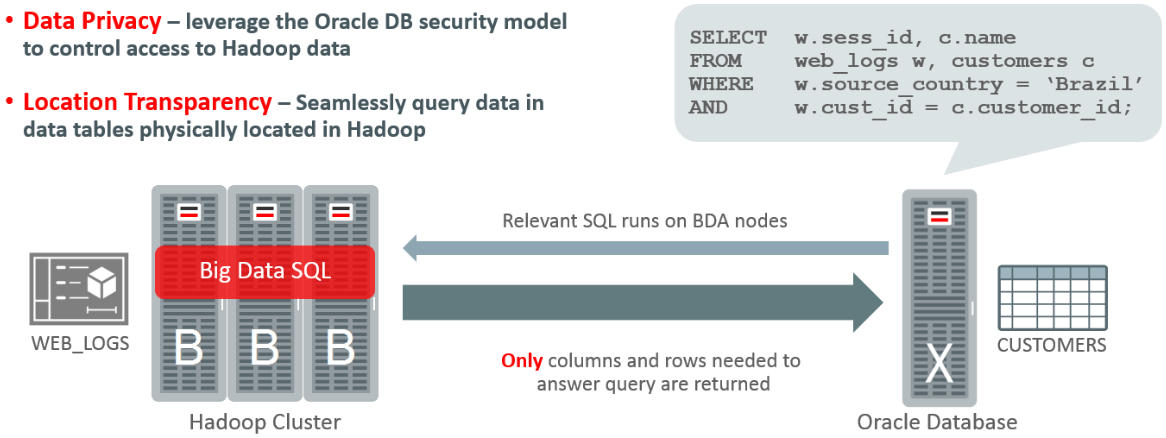


Figure 11: Illustration of Big Data SQL bypassing native access methods

Oracle Big Data SQL runs on the Oracle Big Data Appliance and Oracle Exadata. Big Data SQL enables one SQL query to join across Hadoop, NoSQL, and Oracle RDBMS.

How it works: Big Data SQL references the HDFS metadata catalog (hcatalog) to discover physical data locations and data parsing characteristics. It then automatically creates external database tables and provides the correct linkage during SQL execution. This enables standard SQL queries to access the data in Hadoop, Hive, and NoSQL as if it were native in the Oracle Database.



- **Data Privacy** – leverage the Oracle DB security model to control access to Hadoop data
- **Location Transparency** – Seamlessly query data in data tables physically located in Hadoop

```
SELECT  w.sess_id, c.name
FROM    web_logs w, customers c
WHERE   w.source_country = ‘Brazil’
AND     w.cust_id = c.customer_id;
```

Relevant SQL runs on BDA nodes
Big Data SQL
B
B
B
WEB_LOGS
Hadoop Cluster
Only columns and rows needed to answer query are returned
X
CUSTOMERS
Oracle Database

Figure 12: Illustration of Big Data SQL execution with SQL join showing data source transparency

Key benefits of Big Data SQL are:

Leverage Existing SQL Skills

- Users and developers are able to access Big Data data and applications without learning new SQL skills.

Rich SQL Language

- Big Data SQL is the same multi-purpose query language for analytics, integration, and transformation as the Oracle SQL language that manages the Oracle Database. Big Data SQL is not a subset of Oracle's SQL capabilities, rather it is an extension of Oracle's core SQL engine that operates natively on the Oracle Big Data Appliance.

Engineered for Performance using Oracle Smart Scan

- When used with the Oracle Big Data Appliance or Oracle Exadata, it has been engineered to run at optimum performance using Oracle Smart Scan. During SQL execution, Oracle Smart Scan is able to filter desired data at the storage layer, thus minimizing the data transfer through the backplane or network interconnect to the compute layer.

Speed of Discovery

- Organizations no longer have to copy and move data between platforms, construct separate queries for each platform and then figure out how to connect the results.
- Oracle Smart Scan technology implemented on the Oracle Big Data Appliance increases query performance by minimizing data transfer from the local storage system into the database cache. Oracle Smart Scan is also used on Oracle Exadata.
- SQL-enabled business intelligence tools and applications can access Hadoop and NoSQL data sources.

Governance and Security

- Big Data SQL extends the advanced security capabilities of Oracle Database such as redaction, privilege controls, and virtual private database to limit privileged user access to Hadoop and NoSQL data.

Data Integration

With the surging volume of data being sourced from an ever growing variety of data sources and application, many streaming with great velocity, organizations are unable to use traditional data integration mechanisms such as ETL (extraction, transformation, and load). Big data requires new strategies and technologies designed to analyze big data sets at terabyte or even petabyte scale. As mentioned earlier in this paper, big data has the same requirements for quality, governance, and confidence as conventional data.

Oracle's family of Integration Products supports nearly all of the Apache Big Data technologies as well as many competitive products to Oracle. Our core integration capabilities support an entire infrastructure around data movement and transformation that include integration orchestration, data quality, data lineage, and data governance.

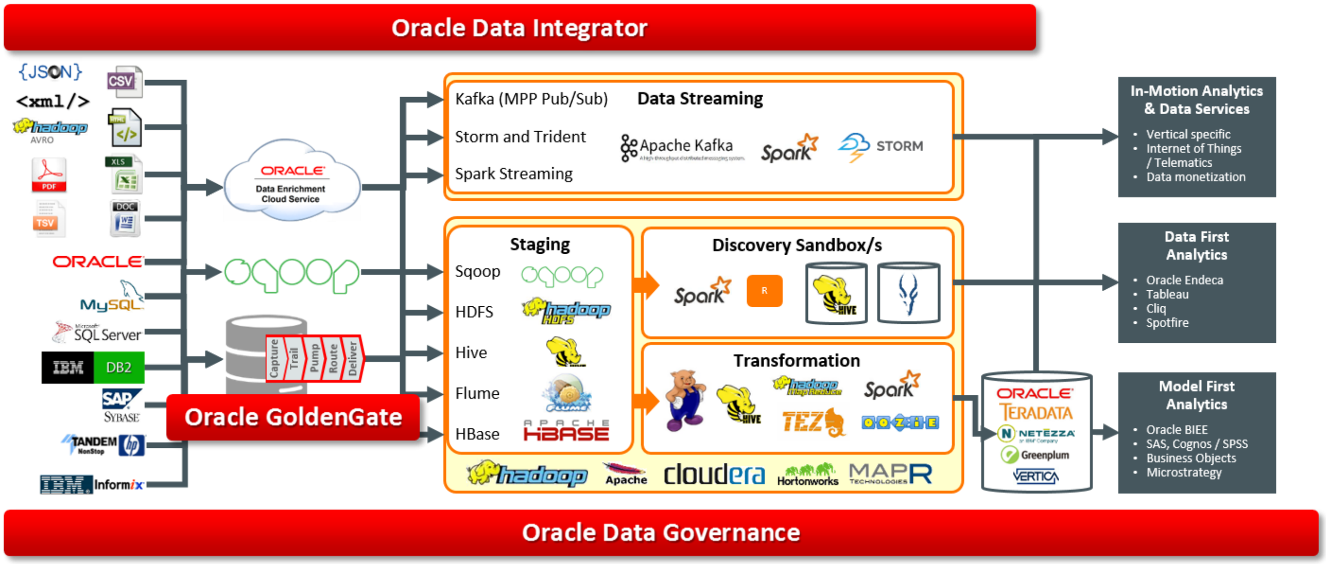


Figure 13: Oracle Open Integration Architecture

Oracle Big Data Connectors

Oracle Big Data Connectors enable the integration of data stored in Big Data platforms, including HDFS and NoSQL databases, with the Oracle RDBMS, facilitating data access to quickly load, extract, transform, and process large and diverse data sets. The Big Data Connectors provide easy-to-use graphical environments that can map sources and targets without writing complicated code, supporting various integration needs in real-time and batch.

Oracle's Big Data Connector offerings include:

- Oracle SQL Connector for HDFS: Enables Oracle Database to access data stored in Hadoop Distributed File System (HDFS). The data can remain in HDFS, or it can be loaded into an Oracle database.
- Oracle Loader for Hadoop: A MapReduce application, which can be invoked as a command-line utility, provides fast movement of data from a Hadoop cluster into a table in an Oracle database.
- Oracle Data Integrator Application Adapter for Hadoop: Extracts, transforms, and loads data from a Hadoop cluster into tables in an Oracle database, as defined using a graphical user interface.
- Oracle R Connector for Hadoop: Provides an interface between a local R environment, Oracle Database, and Hadoop, allowing speed-of-thought, interactive analysis on all three platforms.
- Oracle XQuery for Hadoop: Provides native XQuery access to HDFS and the Hadoop parallel framework.

Apache Flume

Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming event data from streaming systems and applications to Hadoop. Examples of streaming applications include application logs, GPS tracking, social media updates, and digital sensors. It has a simple and flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms. It uses a simple extensible data model that allows for online analytic application.

Multiple technologies exist that can be used to process and store fast-moving streams like application logs, machines sensors, and social media updates in the Hadoop Distributed File System (HDFS) . Flume deploys as one or more *agents*, each contained within its own instance of the Java Virtual Machine (JVM). The Flume agent, via its three pluggable components: Sources, Sinks, and Channels, can be configured to listen (Sources) for any events ranging from strings in stdout to HTTP Posts and RPC calls, transfer (Channel) events from their sources and write (Sink) outputs to storage.

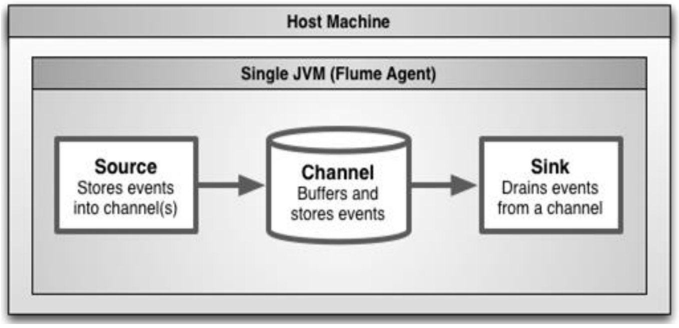

Host Machine
Single JVM (Flume Agent)
Source
Stores events into channel(s)
Channel
Buffers and stores events
Sink
Drains events from a channel

Figure 14: Flume Architecture

Some typical use cases include:

- Sentiment Analysis and Brand Reputation -- Financial service firms deploy Flume agents in various applications such as Call Centers, CRM, Branch operations, Web applications, and Mobile apps to pull data from all channels into HDFS, then analyzing these data streams to help deliver insights.
- Quality Control and Production Improvement – In factories, many machines in production produce log files. Manufacturers deploy Flume agents to collect these log files from production machines and store them in HDFS. The analysis of these large volumes of log files can provide significant info to pin-point potential problems and improve quality control.

Apache Sqoop

Sqoop is a big data integration tool designed for efficiently transferring bulk data between Hadoop ecosystems (HDFS, NoSQL, etc) and structured data stores such as relational databases (Oracle, SQL Server, Teradata, MySQL, etc). Organizations deploy Sqoop to load data from a production database into a HDFS and/or HBase for analysis. In addition, Sqoop export jobs are used to export data sets from HDFS, load into production database, and combine with production transactions for business insights analysis.

Sqoop uses MapReduce to import and export the data, which provides parallel operation as well as fault tolerance, Sqoop can import data to HDFS from RDBMS and also export the data from Hadoop to any RDBMS in form of CSV file or direct export to databases. The input to the import process is a database table. Sqoop will read the table row-by-row into HDFS. The output of this import process is a set of files containing a copy of the imported table.

Event Processing

Oracle Event Processing is a complete, open architecture that enables the sourcing, processing, and publishing of complex events. This allows filtering, correlation, and processing of events in real-time so that downstream applications are driven by true, real-time intelligence. Oracle Event Processing provides the ability to join incoming streaming events with persisted data, thereby delivering contextually aware filtering, correlation, aggregation, and pattern matching. Oracle Event Processing can support very low latency and high throughput in an application context that assumes a large amount of data and a high requirement for immediate response.

Oracle Event Processing is based on an open architecture with support for industry-standards including ANSI SQL, Java, Spring DM and OSGi. Oracle Event Processing makes it easy for organizations to deal with event processing needs, with a visual development environment as well as standard Java-based tooling.

Oracle Event Processing key features include:

- Deployable stand-alone, integrated in the SOA stack, or on lightweight Embedded Java
- Comprehensive event processing query language supports both in-memory and persistent query execution based on standard SQL syntax
- Runtime environment includes a lightweight, Java-based container that scales to high-end event processing use cases with optimized application thread and memory management
- Enterprise class High Availability, Scalability, Performance and Reliability with an integrated in-memory grid and connectivity with Big Data tools
- Advanced Web 2.0 management and performance monitoring console

Oracle Event Processing targets a wealth of industries and functional areas. The following are some use cases:

- Telecommunications: Ability to perform real-time call detail record monitoring and distributed denial of service attack detection.
- Financial Services: Ability to capitalize on arbitrage opportunities that exist in millisecond or microsecond windows. Ability to perform real-time risk analysis, monitoring and reporting of financial securities trading and calculate foreign exchange prices.
- Transportation: Ability to create passenger alerts and detect baggage location in case of flight discrepancies due to local or destination-city weather, ground crew operations, airport security, etc.
- Public Sector/Military: Ability to detect dispersed geographical enemy information, abstract it, and decipher high probability of enemy attack. Ability to alert the most appropriate resources to respond to an emergency.
- Insurance: In conjunction with Oracle Real Time Decisions, ability to learn to detect potentially fraudulent claims.
- IT Systems: Ability to detect failed applications or servers in real-time and trigger corrective measures.
- Supply Chain and Logistics: Ability to track shipments in real-time and detect and report on potential delays in arrival.

Security Architecture

Without question, the Big Data ecosystem must be secure. Oracle's comprehensive data security approach ensures the right people, internal or external, get access to the appropriate data and information at right time and place, within the right channel. Defense-in-depth security prevents and safeguards against malicious attacks and protects organizational information assets by securing and encrypting data while it is in-motion or at-rest. It also enables organizations to separate roles and responsibilities and protect sensitive data without compromising privileged user access, such as DBAs administration. Furthermore, it extends monitoring, auditing and compliance reporting across traditional data management to big data systems.

Oracle Big Data Appliance includes data at rest and network encryption capabilities. The Big Data Appliance also includes enterprise-grade authentication (Kerberos), authorization (LDAP and Apache Sentry project), and auditing (Oracle Audit Vault and Database Firewall) that can be automatically set up on installation, greatly simplifying the process of hardening Hadoop.

Below is the logical architecture for the big data security approach:

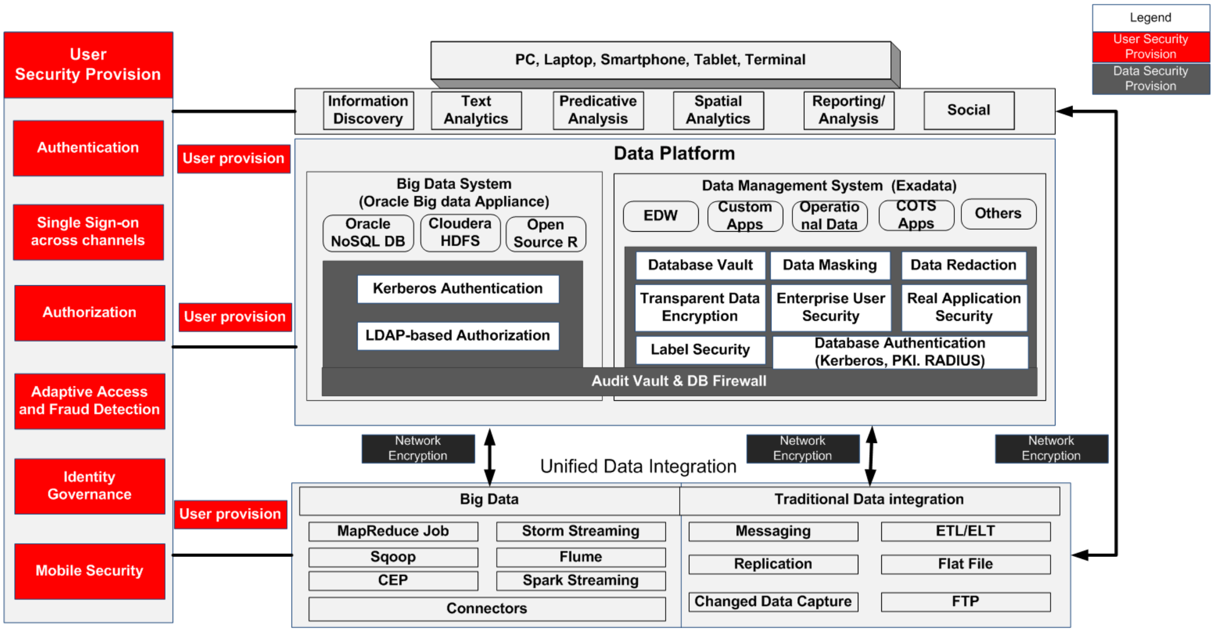


Figure 15: Oracle Security Architecture for the Oracle Big Data Platform

The spectrum of data security capabilities are:

- Authentication and authorization of users, applications and databases
- Privileged user access and administration
- Data encryption (Cloudera Navigator Encrypt) and redaction
- Data masking and subsetting
- Separation of roles and responsibilities
- Transport security
- API security (Database Firewall)
- Database activity monitoring, alerting, blocking, auditing and compliance reporting

Comparing Business Intelligence, Information Discovery, and Analytics

Many analytical and visualization tools have surfaced in recent years. It is important to understand the similarities and differences of the key capabilities to help select the right tool for the right analysis and user base.

Business Intelligence and Information Discovery

An important concept to remember is that Business Intelligence and Information Discovery are peers. They solve different problems and create different types of value. Business Intelligence provides proven answers to known questions. Information Discovery provides fast answers to new questions. The key performance indicators (KPIs), reports, and dashboards produced by the Business Intelligence tools drive the need for exploration and discovery using Information Discovery Platform.

For example, when a report says that warranty claims on the top-selling product grew by 15 percent last month, the new questions that surface could include “What has changed?”, “What’s the root cause of this change?”, and “What are customers saying about this product recently?” In this example, Oracle Endeca or Oracle Big Data Discovery can be leveraged as a discovery application for data exploration.

The relationship can go the other direction as well. Information Discovery creates new KPIs for the Business Intelligence tools to deliver. For example, a consumer packaged goods (CPG) company learned through Information Discovery that the preference for seemingly unrelated brands was highly correlated with certain customer segments. This came from a social media discovery application and established new KPIs that the company should track on an ongoing basis. These new KPIs will then be incorporated into standard reports and dashboards within the Business Intelligence applications.

The development processes of Information Discovery versus Business Intelligence applications could vary as well. With traditional BI, the users know the answer they are looking for. They define requirements and build to the objective. It starts with requirement documentation, data model definition, and then implementation and development of a data model, metadata definition, data integration, and finally, reports and dashboards development. When new business requirements surface, a new iteration of this process is triggered.

With an Information Discovery application, the users may have an idea or interest, but they don’t necessarily know what to anticipate. The answer for the initial question will lead to the next set of questions. As a result, the development process is more fluid. It requires that the data analysts explore the data as they develop and refine their hypothesis.

An example:

For data discovery use cases, analysts explore data seeking to discover relationships and correlations. In a discovery lab, many structured and unstructured datasets, which typically are not integrated, are combined without traditional data integration to find a solution. In the example of preventative maintenance, disparate data sources provide structured part specifications, performance history, service history, and repair processes. These traditional structured data sources are then combined with unstructured text sourced from comment fields from social applications or customer relationship management systems to anticipate maintenance requirements and to better understand the root cause of specific maintenance issues. Discovery processes like this work best when the domain expert can continuously pivot the data in real-time. In these cases, highly specialized indexes across optimized data structures, semantic processing, large memory caches, and fast networks are required for instantaneous data searching and model building.

However, it should be noted that traditional Business Intelligence and Information Discovery are not mutually exclusive. Instead, they are complementary and when combined they provide the user community with a more powerful decision support foundation.

Advanced Analytics

Advanced Analytics is similar to information discovery in that it is based on a more iterative approach compared to traditional BI.

Users typically start with a hypothesis or a set of questions. They define and collect data sources including structured data from their existing data warehouse and unstructured data from weblogs, clickstreams, social media channels, and internal text logs and documents. They use tools to explore results and reduce ambiguity. Information Discovery tools are a great fit for this stage.

Once the datasets are refined, analysts can apply analytical models using analytic tools like Oracle Advanced Analytics, SAS, and R to eliminate outliers, find concentrations, and make correlations. They then interpret the outcome and continuously refine their queries and models and establish an improved hypothesis.

In the end, this analysis might lead to a creation of new theories and predictions based on the data. Again, the process is very fluid in nature and is different from traditional Software Development Life Cycle (SDLC) and BI development.

Here's a comparison chart that clarifies the three areas of analytics and reporting capabilities:

COMPARISON OF ORACLE ANALYTICS AND REPORTING CAPABILITIES

	Oracle **BI Foundation Suite**	**Oracle** **Endeca Information Discovery or Oracle Big Data Discovery**	**Oracle** **Advanced Analytics**
Key Concept	**Proven answers to known questions**	**Fast answers to new questions**	**Uncover trends based on hypothesis**
Approach	Semantic model integrates data sources and provides strong governance, confidence, and reuse	Ingest sources as needed for discovery. Model derived from incoming data	Various statistical and machine learning algorithms for identifying hidden correlations
Data Sources	Data warehouse plus federated sources, mostly structured, with the ability to model the relationships	Multiple sources that may be difficult to relate and may change over time including structured, semi-structured, and unstructured data sources	Structured data sources leveraging Oracle Data Mining (a component of Oracle Advanced Analytics) and structured and unstructured data leveraging the Oracle R Distribution data in RDBMS databases and Hadoop
Users	Broad array of enterprise consumers via reports, dashboards, mobile, embedded in business processes, …	Technical users with an understanding of the business and business requirements	Data scientists and technical users who understand statistical modeling, text mining and analytics, predictive modeling, etc.
Timing	Company has months to complete	Company has weeks to complete	Weeks to months to analyze and fit the model

In short, understanding the differences and the complementary nature of Business Intelligence, Information Discovery, and Advanced Analytics will help you choose the right tool for the right requirement based on the business context. To learn more, click on the links in the table heading.

Big Data Architecture Patterns in Three Use Cases

In this section, we will explore the following 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: Data Exploration

The first example 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. One area to investigate is the website navigation pattern, especially related to abandoned shopping carts.

The architecture challenge in this use case was to quickly implement a solution with as much existing tools, skills, and infrastructure as possible in order to minimize cost and to quickly deliver a solution to the business. The number of skilled Hadoop programmers on staff was very few but they did have SQL expertise. One option was to load all the data into a relational database management platform so that the SQL programmers could access the data, but the data movement was extensive and required so much processing power and storage that it did not make economic sense. The 2nd option was to load the data into Hadoop and directly access the data in HDFS using SQL.

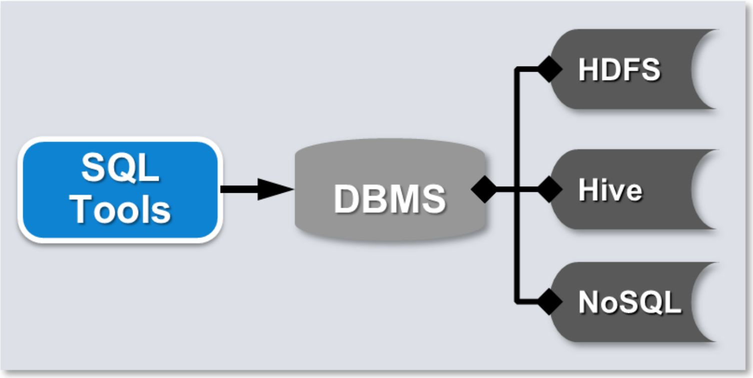


Figure 16: Use Case #1: Data Exploration

The conceptual architecture decision/approach as shown above was to provide direct access to the Hadoop Distributed File System by simply associating it with an Oracle Database External Table. Once connected, Oracle Big Data SQL or the Oracle SQL Connector for HDFS, one of the Oracle Big Data Connectors enabled traditional SQL tools to explore the dataset.

The key benefits include:

- Low cost Hadoop storage
- Ability to leverage existing investments and skills in SQL and BI tools
- No client side software installation
- Leverage Oracle's robust SQL language
- No data movement into the relational database
- Fast ingestion and integration of structured and unstructured datasets

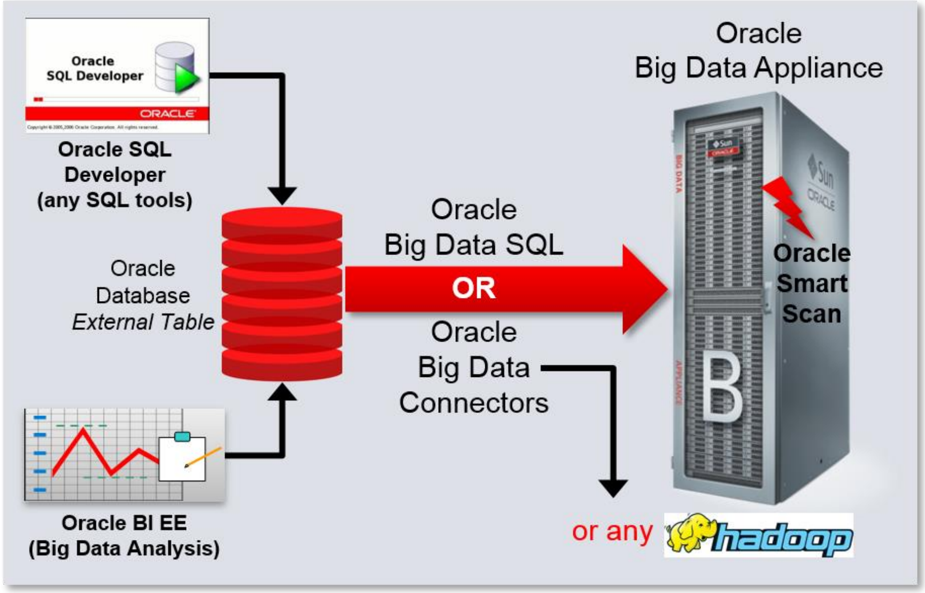

Oracle
SQL Developer
ORACLE
Oracle SQL Developer (any SQL tools)
Oracle Database External Table
Oracle BI EE (Big Data Analysis)
Oracle Big Data SQL
OR
Oracle Big Data Connectors
Oracle Big Data Appliance
Oracle Smart Scan
or any hadoop

Figure 17: Use Case #1: Data Exploration and the Architecture Decisions

The diagram above shows the architectural components used to meet this challenge.

The key Oracle components in this architecture:

- 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 Oracle BI Enterprise Edition can be used to access data through the Oracle Database
- Oracle Database External Table:
  - An Oracle database feature that presents data stored in a file system in a row and column table format. Then, data is accessible using the SQL query language.
- Oracle Big Data SQL:
  - The preferred SQL access method that provides advanced connectivity between the Oracle Big Data Appliance (data reservoir) and Oracle Exadata (data warehouse).
  - Makes use of Oracle’s Smart Scan feature that intelligently selects data from the storage system directly rather than moving the data into main memory and then evaluating it.
  - Uses the ‘hcatalog’ metadata store, to automatically create database external tables for optimal operations. Big Data SQL can connect to multiple data sources through this catalog.
- Oracle Big Data Connectors:
  - Connectors that access HDFS data through the Oracle external table capability.
  - Also creates optimized data sets for efficient loading and analysis in Oracle Database 12c and Oracle Enterprise R. These connectors can connect to any Hadoop solution.
- Oracle Big Data Appliance:
  - An Oracle Engineered System. Powered by the full distribution of Cloudera’s Distribution including Apache Hadoop (CDH) to store logs, reviews, and other related big data

In summary, the key architecture choice in this scenario is to avoid data movement and duplication, minimize storage and processing requirements and costs, and leverage existing SQL tools and skill sets.

Use Case #2: Real Time Alerts

A large financial institution had regulatory obligations to detect potential financial crimes and terrorist activity. However, there were challenges:

- Correlating data in disparate formats from an multitude of sources – this requirement arose from 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
- Capturing, storing, and accessing the ever growing volume of information that was constantly streaming in to the institution. Their IT systems needed 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. Some of these sources were real time, some were in batch.

The architecture challenge in this use case was to use a streaming processing engine to evaluate a variety of data sources simultaneously. And to leverage their existing BI platform for the associated regulatory reporting requirements. The required architecture capabilities were multi-channel processing (both real time and batch) and the right tools for each channel. For example, some data went through the real time channel. However, there was also a certain set of data that needed to go through both the real time channel as well as the batch channel to be processed and analyzed through different angles.

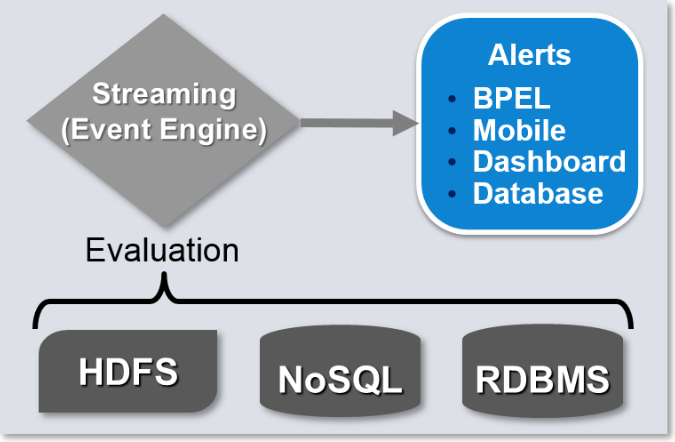


Figure 18: Use Case #2: Big Data for Real Time Alerts

The ideal scenario is 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. Historically, the volume and variety had never been used to its fullest extent due to constraints in processing power and the cost of storage. With Hadoop, Spark, and/or Storm processing, it is now possible to incorporate all the detailed data points to calculate continuous risk profiles. Profile access and last transactions can be cached in a NoSQL database and then be accessible to real-time event processing engine on-demand to evaluate the risk. After the risk is evaluated, transaction actions and exceptions update the NoSQL cached risk profiles in addition to publishing event messages. Message subscribers would include various operational and analytical systems for appropriate reporting, analysis and action.

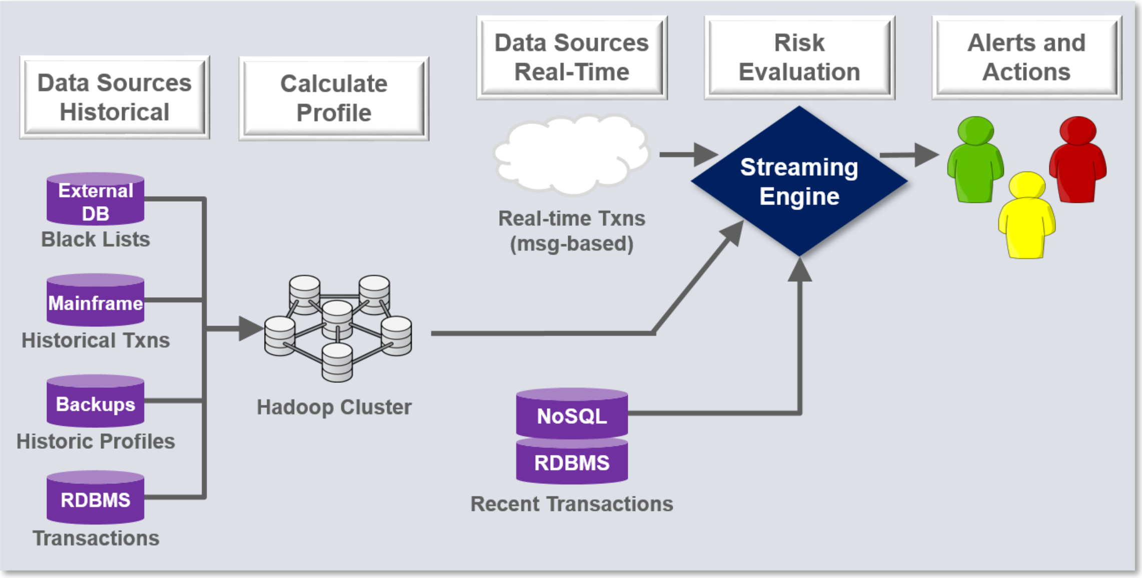


Figure 19: Use Case #2: Real Time Alerts - Data Flow

In this scenario, the Hadoop Cluster consolidates data from real-time, operational, and data warehouse sources in flexible data structures. A periodic batch-based risk assessment process, operating on top of Hadoop, calculates risk, identifies trends, and updates an individual customer risk profile cached in a NoSQL database. As real-time events stream from the network, the event engine evaluates risk by testing the transaction event versus the cached profile, then triggers appropriate actions, and logs the evaluation.

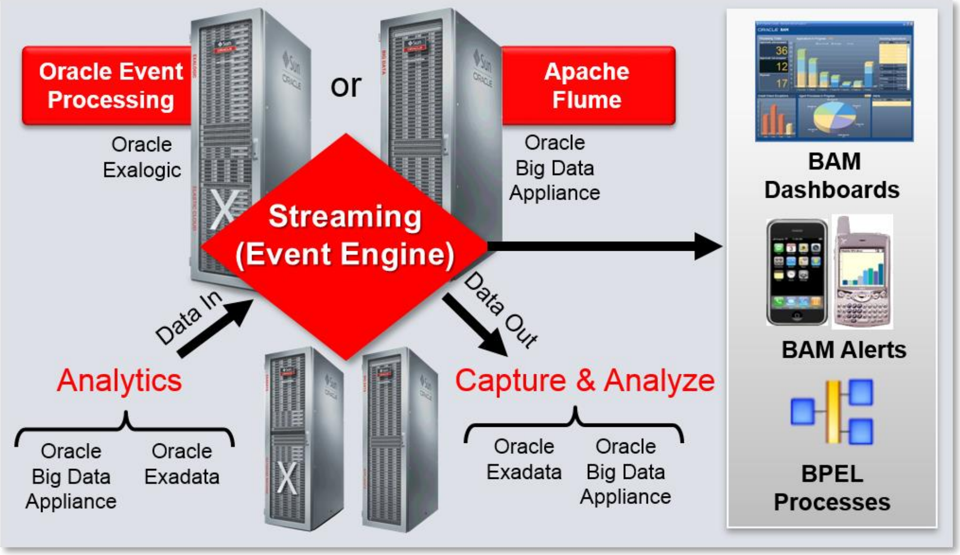


Figure 20: Use Case #2: Real-Time Alerts - Architecture Decisions

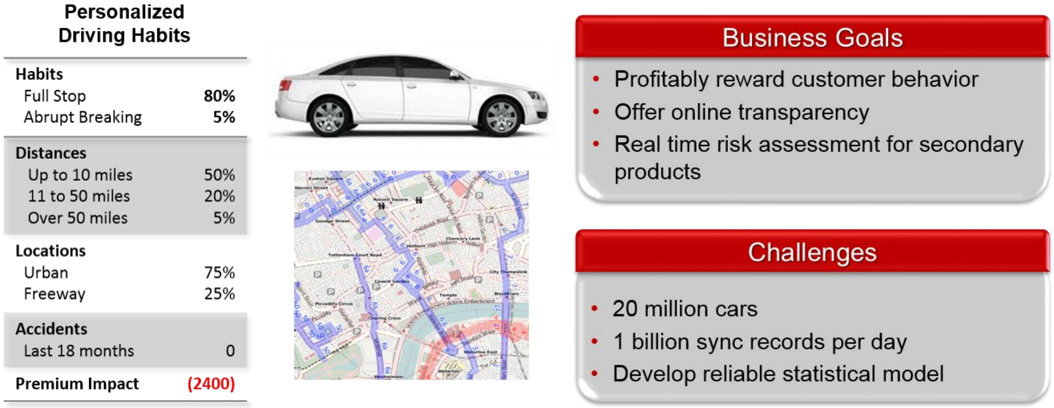
The diagram above illustrates the following components of this architecture:

- Event Processing
  - Oracle Event Processing. Oracle Event Processing continuously processes incoming data, analyzes and evolves patterns, and raises events if conditions are detected. Oracle EP runs in an Open Service Gateway (OSGi) container and can operate on any Java Runtime Environment. It is part of the Oracle SOA Suite and can operate across Oracle's Engineered System including Oracle Exalogic, Sparc T5, and more.
  - Oracle Stream Explorer. A business level user interface allowing interpreting data streams without requiring knowledge of underlying event technology characteristics.
  - Apache streaming options. Powered by the Cloudera Distribution on the Oracle Big Data Appliance and Apache, multiple streaming options are available including Spark Streaming, Flume, and Storm.
- Oracle Big Data Appliance (or other Hadoop Solutions):
  - Capture events (various options, such as Flume, Spark Streaming)
  - Powered by the full distribution of Cloudera's Distribution including Apache Hadoop (CDH) to store logs, reviews, and other related big data
  - NoSQL database to capture low latency data with flexible data structure and fast querying
  - MapReduce to process high volume, high variety data from multiple data sources and then reduce and optimize dataset to calculate risk profiles. Profile data can be evaluated by an event engine and the transaction actions and exceptions can be stored in Hadoop.
- Oracle Exalogic:
  - Oracle Event Processing: Streaming event engine to continuously process incoming data, analyze and evolve patterns, and raise events if suspicious activities are detected. Event activity can be captured for analysis.
  - Oracle Business Process Execution Language (BPEL): The BPEL engine defines processes and appropriate actions when events occur.
  - Oracle Business Activity Monitoring (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 disparate data with an 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 an insurance company seeking to personalize insurance coverage and premiums based on individual driving habits. The insurance company needs to capture a large amount of vehicle-created sensor data reflecting their customers' driving habits, store it 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.



Personalized Driving Habits

Habits	
Full Stop	**80%**
Abrupt Breaking	**5%**
Distances	
Up to 10 miles	50%
11 to 50 miles	20%
Over 50 miles	5%
Locations	
Urban	75%
Freeway	25%
Accidents	
Last 18 months	0
Premium Impact	**(2400)**

Business Goals

- Profitably reward customer behavior
- Offer online transparency
- Real time risk assessment for secondary products

Challenges

- 20 million cars
- 1 billion sync records per day
- Develop reliable statistical model

Figure 21: Use Case #3: Business Objectives

The architecture challenge in this use case was to bridge the gap between a Big Data architecture and their existing information architecture and existing investments. First, unstructured driving data needed to be matched up and correlated to the structured insured data (demographics, in-force policies, claims history, payment history, etc.). The users had to be able to consume the results using the existing BI eco-system. And lastly, data security had to be in place to meet regulatory and compliance requirements.

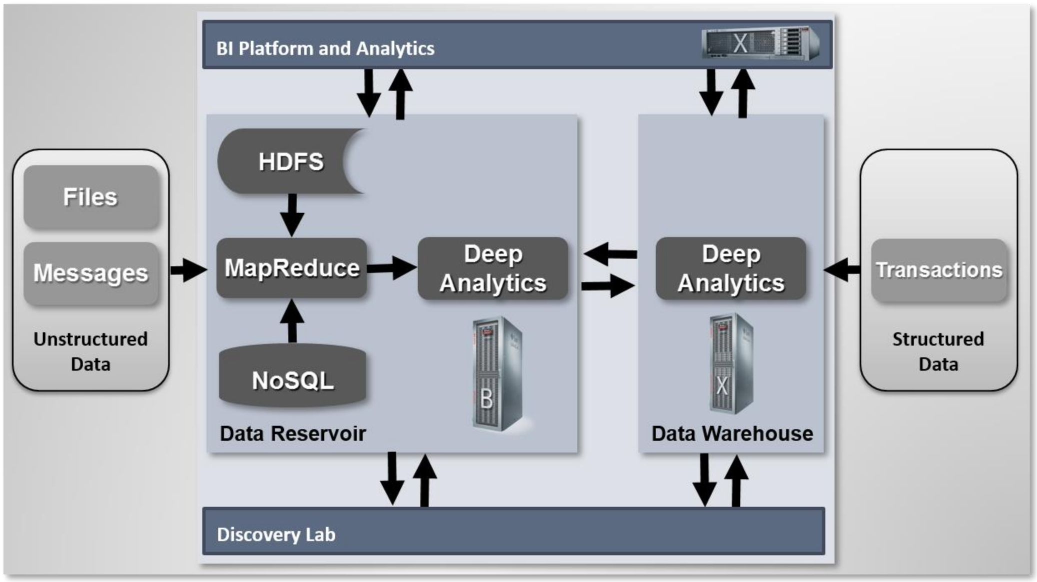


Figure 22: Use Case #3: Oracle Big Data Platform for Combined Analytics

In this scenario, the solution must accomplish multiple goals: update the customer profile, calculate new premiums, update the data warehouse, and contribute data to a discovery lab where profitability and competiveness can be analyzed. The architecture decisions were designed to minimize data movement across platforms, integrate BI and analytic processes, and enable deep analysis, and ensure access/identity management and data security capabilities would be applied consistently.

Due to the volume and variety of sensor data, HDFS was chosen to store the raw data. MapReduce processing filtered the low-density data into meaningful summaries. In-reservoir SQL and "R" analytics calculated initial premium scoring with data "in-place." Customer profiles were updated in the NoSQL database, and exported to the operational and data warehouse systems. The driving behavior data, derived profiles, and other premium factors, were loaded into the discovery lab for additional research. Using conventional BI and information discovery tools, some enabled by Big Data SQL, data was accessible across all these environments.

As a result of this architecture approach, the business users did not experience a "Big Data" divide. That is, they did not even need to know there was a difference between traditional transaction data and big data. Everything was seamless as they navigated through the data, tested hypotheses, analyzed patterns, and made informed decisions.

In summary, the key architecture choice in this use case was the integration of unstructured Big Data with structured RDBMS and BI data. Leveraging Oracle engineered systems including Oracle Big Data Appliance, Oracle Exadata, and Oracle Exalytics reduced implementation risks, provided fast time to value, and extreme performance and scalability to meet a very complex business and IT challenge.

The Oracle Big Data Platform

The overall capabilities in the prior use case also summarize the operational components of *The Oracle Big Data Platform* (see Figure 9) which are engineered to work together and deliver enterprise grade capabilities for performance, scalability, availability, and management. The key principles of the platform are:

- **Discover and Predict Fast** – Make daily discovery and data science easier and faster. Unique hardware and software engineering enables huge volumes of data to be processed at extreme speeds. This is a critical requirement for ad hoc data exploration and data intensive statistical modeling. These techniques allow for exploring big data in-place whether in the Data Reservoir or the Data Warehouse further reducing time to benefit.
- **Simplify Access to All Data** – Access data across Hadoop, NoSQL, and relational databases as if in a single system. Oracle has simplified access by enabling SQL to read data in additional databases. This capability extends to tools that use SQL to read data thus preserving existing business intelligence, analytics, and reporting investments.
- **Govern and Secure All Data** – Integrate data from inside and outside the firewall without sacrificing data quality standards. Oracle provides additional features beyond Apache Sentry, MIT Kerberos, and native encryption capabilities. Oracle provides transparency and lifecycle management for data transformation orchestration, data quality, and data linage. Secure access to HDFS data is extended from your existing Oracle investments in Oracle Identity Management and Oracle Database. Oracle also provides a single, secure audit repository that monitors access to open source in addition to the Oracle software on the Big Data Appliance.

In the table below are the key hardware and software components of the Oracle Big Data Platform:

ORACLE BIG DATA PLATFORM

Oracle Engineered System	Oracle Product	Description
Event Engine		
Oracle Exalogic Elastic Cloud	Website and product data sheet	Oracle Exalogic is hardware and software engineered together to provide extreme performance, reliability, and scalability for middleware services. It works with Oracle, Java, and other applications.
Oracle SuperCluster	Website and product data sheet	Oracle SuperCluster is Oracle's most powerful database machine. It includes an integrated server, storage, networking and software system for end-to-end database and application operations.
	Oracle Event Processing	An event processor to capture data streams. Can be installed on Oracle Exalogic, Oracle SuperCluster, and other Java Runtime Environments.
Data Reservoir		
Oracle Big Data Appliance	Website and product data sheet	Optimized hardware combined with the most comprehensive Big Data software stack, delivering complete, easy-to-deploy solutions for acquiring, organizing, and loading big data into Oracle Database.
	Cloudera Enterprise	Leading Open Source distribution supported by Oracle. Cloudera includes "Map Reduce" parallel processing capability and supports various Apache Projects, such as Impala, Spark, and more.
	Oracle NoSQL Database	Database with flexible data structure and fast performance.
	Oracle R Distribution	An open source analytic library supported by Oracle
	Oracle Big Data Discovery	A visual, intuitive product that runs natively on Hadoop and enables business insight in minutes without data specialists.
Data Factory		
Multiple	Oracle Big Data Connectors	Provides adapters that enable Hadoop to integrate with an Oracle Database across Big Data and Exadata engineered systems. Managed through an easy to use graphical user interface.
Multiple	Oracle Data Integrator	A suite of products enabling bulk data integration and transformation with next-generation Extract Load and Transform (ELT) technology with declarative design for seamless batch and real-time integration
Multiple	Oracle Data Integrator Big Data Option	A flexible and heterogeneous data integration platform for big data processing that enables you to transform, enrich, and govern raw data
Multiple	Oracle Golden Gate	Low-impact, real-time change data capture, distribution, and delivery for transactional data across heterogeneous systems for continuous availability, zero downtime migration, and business intelligence.
Multiple	Oracle Golden Gate for Big Data	Streams transactional data into big data systems in real time
Multiple	Oracle Data Quality	Enables organizations to govern, measure, improve, and manage the quality of data. Data quality includes profiling, cleansing, matching, monitoring capabilities.
Multiple	Oracle Enterprise Metadata Mgmt	Captures metadata from Oracle and third-party solutions enabling exploration, reporting, analysis, and governing of metadata.
Data Warehouse		
Oracle Exadata	Website and product data sheet	Database machine designed to provide extreme performance for both data warehousing and OLTP (online transaction processing)

Oracle Engineered System	Oracle Product	Description
		applications, and an ideal platform for consolidating database workloads in private clouds as well as in the data center.
Oracle Flash Storage	Website and product data sheet	SAN storage, architected to exploit the power of flash for greater performance, and better economics through automated tiering of flash and disk.
	Oracle Database Enterprise Edition 12c	Oracle's latest version of its relational database that includes in-memory data processing capabilities delivering breakthrough analytical performance.
	A total of 16 database options: (Database In-Memory, Multitenant, Spatial and Graph, text, and more)	Additional database capabilities in the areas of performance and scalability, high availability, security and compliance, data warehousing and big data, and manageability.
	Oracle Advanced Analytics	A family of analytical products that perform advanced data mining functions containing: Oracle Data Miner and Oracle Enterprise R
Business Analytics		
Oracle Exalytics In-Memory Machine	Website and product data sheet	An in-memory hardware and software platform that provides the fastest business intelligence (BI) solution for modeling, forecasting, and planning applications, along with optimized, advanced data visualization for actionable insight into large data sets
	Oracle BI Foundation Suite	An enterprise metadata store that underlies a family of business intelligence reporting and development tools.
	Oracle Endeca Information Discovery	Oracle Endeca Information Discovery (EID) is a complete enterprise data discovery platform that combines information of any type, from any source. EID uses faceted navigation to allow business users to explore all forms of data through an intuitive search engine.
	Oracle Essbase	The industry-leading multi-dimensional online analytical processing (OLAP) server, designed to help business users forecast business performance.
Discovery Lab		
Oracle Big Data Appliance Oracle Exadata Oracle Exalytics In-Memory Machine Oracle Exalogic Elastic Cloud		Discovery Labs can run on dedicated or shared infrastructure.
	Oracle Products useful in a Lab:	» *Analytical SQL functions: Spatial, RDF/SPARQL, SQL Pattern Matching* » *Oracle Advance Analytics: Oracle R + Oracle Data Mining* » *Oracle Spatial and Graph* » *Oracle Big Data Discovery* » *Oracle Endeca* » *Oracle Data Quality* » *Oracle Metadata Manager*
Cross System Components		
	Infiniband Networking	Connections between Oracle Big Data Appliance, Oracle Exadata, and Oracle Exalytics are via Infiniband (40GB per second), enabling high-speed data transfer for batch or query workloads.

Oracle Engineered System	Oracle Product	Description
	Oracle Big Data SQL	Enables the Oracle SQL language to access data from data stores on the Oracle Big Data Appliance including Hadoop, Hive, and Oracle NoSQL.
	Oracle Enterprise Manager	Oracle Enterprise Manager is Oracle’s integrated enterprise IT management product line, which provides complete, integrated and business-driven enterprise management.
	Oracle Audit Vault and Database Firewall	Oracle Audit Vault and Database Firewall monitors Oracle and non-Oracle database traffic to detect and block threats, as well as improves compliance reporting by consolidating audit data from databases, operating systems, directories, and other sources

Big Data Best Practices

Guidelines for building 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 determine 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 Institute[1] 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. Organizations implementing Big Data solutions and strategies should assess skills requirement early and often and should proactively identify any potential skills gaps. Skills gaps can be addressed by training / cross-training existing resources, hiring new resources, or leveraging consulting firms. Implementing Oracle Engineered systems can also jumpstart Big Data implementations and can provide quicker time to value as you grow your in-house expertise. In addition, leveraging Oracle Big Data solutions will allow you leverage existing SQL tools and expertise with your Big Data implementation, saving time, money, while allowing you to use existing skill sets.

#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. Leveraging a CoE approach can help 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 and integrating low density Big Data with the structured data you are already using today, you can bring even greater business 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, which can lead to better conclusions. For these reasons, many see Big Data as an integral extension of your existing business intelligence and data warehousing platform and information architecture.

[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

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, coupled with 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.

#5: Plan Your Discovery Lab 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." That being said, it's important for Analysts and Data Scientists doing the discovery and exploration of the data to work closely with the business to understand key business knowledge gaps and requirement. 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, and summarized) is part of a Big Data solution. Analytical sandboxes should be created on-demand and resource management is critical to ensure control of the entire data flow, including 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.

Final Thoughts

It's not a leap of faith that we live in a world of continuously increasing data, nor will we as data consumers ever expect less. The effective use of big data, with the rise of social media, mobile devices, and sensors, etc. is now recognized as a key differentiator for companies to gain competitive advantage and to outperform their peers. Tom Peters, bestselling author on business management, once said, "Organizations that do not understand the overwhelming importance of managing data and information as tangible assets in the new economy, will not survive."

The Big Data promise has motivated the business to invest. The information architect is on the front lines as researcher, designer, and advisor. 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 'what's new' across the spectrum of Big Data capabilities, we suggest that you think about a platform but deliver a project. It's important to align new operational and management capabilities with standard IT processes and capabilities, leverage prior investments, 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 leverage best practices where ever possible.

Oracle has leveraged its 30 year leadership in information management and significant investments in research and development to bring the latest innovations and capabilities into enterprise-class Big Data products and solutions. You will find that Oracle's Big Data platform is unique – it is engineered to work together, from the data reservoir to the discovery lab to the data warehouse to business intelligence, delivering the insights that your business needs. Now is the time to work with Oracle to build a Big Data foundation for your company and your career. These new elements are quickly becoming a core requirement for planning your next generation information architecture.

This white paper introduced you to Oracle Big Data products, architecture, and the nature of Oracle's one-on-one architecture guidance services. To understand more about Oracle's enterprise architecture and information architecture consulting services, please visit, www.oracle.com/goto/EA-Services and the specific information architecture service here.

For additional white papers on the Oracle Architecture Development Process (OADP), the associated Oracle Enterprise Architecture Framework (OEAF), or 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 delve deeper into the Oracle Big Data reference architecture consisting of the artifacts, tools and samples, contact your local Oracle sales representative and ask to speak to Oracle's Enterprise Architects.

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

Oracle Corporation, World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065, USA

Worldwide Inquiries
Phone: +1.650.506.7000
Fax: +1.650.506.7200

CONNECT WITH US

blogs.oracle/enterprisearchitecture

facebook.com/OracleEA

twitter.com/oracleEAs

oracle.com/goto/EA

May 2015
An Enterprise Architecture White Paper – An Enterprise Architect's Guide to Big Data — Reference Architecture Overview
Author: Peter Heller, Dee Piziak
Contributing Author: Jeff Knudsen