

# An Enterprise Architect's Guide to Big Data

## Reference Architecture Overview

ORACLE ENTERPRISE ARCHITECTURE WHITE PAPER | MARCH 2016





## 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

<b>Executive Summary</b>	<b>1</b>
A Pointer to Additional Architecture Materials	3
<b>Fundamental Concepts</b>	<b>4</b>
What is Big Data?	4
The Big Questions about Big Data	5
What's Different about Big Data?	7
<b>Taking an Enterprise Architecture Approach</b>	<b>11</b>
<b>Big Data Reference Architecture Overview</b>	<b>14</b>
Traditional Information Architecture Capabilities	14
Adding Big Data Capabilities	14
A Unified Reference Architecture	16
Enterprise Information Management Capabilities	17
Big Data Architecture Capabilities	18
<b>Oracle Big Data Cloud Services</b>	<b>23</b>
<b>Highlights of Oracle's Big Data Architecture</b>	<b>24</b>
Big Data SQL	24
Data Integration	26
Oracle Big Data Connectors	27
Oracle Big Data Preparation	28
Oracle Stream Explorer	29
Security Architecture	30
Comparing Business Intelligence, Information Discovery, and Analytics	31
Data Visualization	33
Spatial and Graph Analysis	35
<b>Extending the Architecture to the Internet of Things</b>	<b>36</b>
<b>Big Data Architecture Patterns in Three Use Cases</b>	<b>38</b>
Use Case #1: Retail Web Log Analysis	38
Use Case #2: Financial Services Real-time Risk Detection	39
Use Case #3: Driver Insurability using Telematics	41
<b>Big Data Best Practices</b>	<b>43</b>
<b>Final Thoughts</b>	<b>45</b>



## Executive Summary

Today, Big Data is commonly defined as data that contains greater variety arriving in increasing volumes and with ever higher velocity. Data growth, speed and complexity are being driven by deployment of billions of intelligent sensors and devices that are transmitting data (popularly called the Internet of Things) and by other sources of semi-structured and structured data. The data must be gathered on an ongoing basis, analyzed, and then provide direction to the business regarding appropriate actions to take, thus providing value.

Most are keenly aware that Big Data is at the heart of nearly every digital transformation taking place today. For example, applications enabling better customer experiences are often powered by smart devices and enable the ability to respond in the moment to customer actions. Smart products being sold can capture an entire environmental context. Business analysts and data scientists are developing a host of new analytical techniques and models to uncover the value provided by this data. Big Data solutions are helping to increase brand loyalty, manage personalized value chains, uncover truths, predict product and consumer trends, reveal product reliability, and discover real accountability.

IT organizations are eagerly deploying Big Data processing, storage and integration technologies in on premises and Public Cloud-based solutions. Cloud-based Big Data solutions are hosted on Infrastructure as a Service (IaaS), delivered as Platform as a Service (PaaS), or as Big Data applications (and data services) via Software as a Service (SaaS) manifestations. Each must meet critical Service Level Agreements (SLAs) for the business intelligence, analytical and operational systems and processes that they are enabling. They must perform at scale, be resilient, secure and governable. They must also be cost effective, minimizing duplication and transfer of data where possible. Today's architecture footprints can now be delivered consistently to these standards. Oracle has created reference architectures for all of these deployment models.

There is good reason for you to look to Oracle as the foundation for your Big Data capabilities. Since 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 on premises and Cloud-based solutions. Oracle's family of data management solutions continue to solve the toughest technological and business problems delivering the highest performance on the most reliable, available and scalable data platforms. Oracle continues to deliver ancillary data management capabilities including data capture, transformation, movement, quality, security, and management while providing robust data discovery, access, analytics and visualization software. Oracle's unique value is 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 – such as Big Data.

Oracle thinks that Big Data is not an island. It is merely the latest aspect of an integrated enterprise-class information management capability. Looked at on its own, Big Data can easily add to the complexity of a corporate IT environment as it evolves through frequent open source contributions, expanding Cloud-based offerings, and emerging analytic strategies. Oracle's best-of-breed products, support, and services can provide the solid foundation for your enterprise architecture as you navigate your way to a safe and successful future state.

To deliver to business requirements and provide value, architects must evaluate 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 and technology are mature enough to process and analyze this data.

This paper is an introduction to the Big Data ecosystem and the architecture choices that an enterprise architect will likely face. We define key terms and capabilities, present reference architectures, and describe key Oracle products and open source solutions. We also provide some perspectives and principles and apply these in 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. These should look familiar to architects familiar with TOGAF and other best architecture 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](#) paper (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](#) paper (140 pages) and [ORA Management and Monitoring](#) paper (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.

Examples of the business context for Big Data implementations for many companies and organizations appears in the industry whitepapers posted on the [Oracle Enterprise Architecture web site](#). Industries covered include agribusiness, communications service providers, education, financial services, healthcare payers, healthcare providers, insurance, logistics and transportation, manufacturing, media and entertainment, 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](http://www.alexa.com)) – and delivers search results plus personalized advertising in milliseconds! This is 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
<b>Business Context</b>		
Business Intent	How will we make use of the data?	<ul style="list-style-type: none"> <li>» <i>Sell new products and services</i></li> <li>» <i>Personalize customer experiences</i></li> <li>» <i>Sense product maintenance needs</i></li> <li>» <i>Predict risk, operational results</i></li> <li>» <i>Sell value-added data</i></li> </ul>
Business Usage	Which business processes can benefit?	<ul style="list-style-type: none"> <li>» <i>Operational ERP/CRM systems</i></li> <li>» <i>BI and Reporting systems</i></li> <li>» <i>Predictive analytics, modeling, data mining</i></li> </ul>
Data Ownership	Do we need to own (and archive) the data?	<ul style="list-style-type: none"> <li>» <i>Proprietary</i></li> <li>» <i>Require historical data</i></li> <li>» <i>Ensure lineage</i></li> <li>» <i>Governance</i></li> </ul>
<b>Architecture Vision</b>		
Ingestion	What are the sense <i>and</i> respond characteristics?	<ul style="list-style-type: none"> <li>» <i>Sensor-based real-time events</i></li> <li>» <i>Near real-time transaction events</i></li> <li>» <i>Real-time analytics</i></li> <li>» <i>Near real time analytics</i></li> <li>» <i>No immediate analytics</i></li> </ul>
Data Storage	What storage technologies are best for our data reservoir?	<ul style="list-style-type: none"> <li>» <i>HDFS (Hadoop plus others)</i></li> <li>» <i>File system</i></li> <li>» <i>Data Warehouse</i></li> <li>» <i>RDBMS</i></li> <li>» <i>NoSQL database</i></li> </ul>
Data Processing	What strategy is practical for my application?	<ul style="list-style-type: none"> <li>» <i>Leave it at the point of capture</i></li> <li>» <i>Add minor transformations</i></li> <li>» <i>ETL data to analytical platform</i></li> <li>» <i>Export data to desktops</i></li> </ul>
Performance	How to maximize speed of ad hoc query, data transformations, and analytical modeling?	<ul style="list-style-type: none"> <li>» <i>Analyze and transform data in real-time</i></li> <li>» <i>Optimize data structures for intended use</i></li> <li>» <i>Use parallel processing</i></li> <li>» <i>Increase hardware and memory</i></li> <li>» <i>Database configuration and operations</i></li> <li>» <i>Dedicate hardware sandboxes</i></li> <li>» <i>Analyze data at rest, in-place</i></li> </ul>
Latency	How to minimize latency between key operational components? (ingest, reservoir, data warehouse,	<ul style="list-style-type: none"> <li>» <i>Share storage</i></li> <li>» <i>High speed interconnect</i></li> </ul>

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

Areas	Questions	Possible Answers
<b>Roadmap</b>		
Proof of Concept	What should the POC validate before we move forward?	<ul style="list-style-type: none"> <li>» Business use case</li> <li>» New technology understanding</li> <li>» Enterprise integration</li> <li>» Operational implications</li> </ul>
Open Source Skills	How to acquire open source skills?	<ul style="list-style-type: none"> <li>» Cross-train employees</li> <li>» Hire expertise</li> <li>» Use experienced vendors/partners</li> </ul>
Analytics Skills	How to acquire analytical skills?	<ul style="list-style-type: none"> <li>» Cross-train employees</li> <li>» Hire expertise</li> <li>» Use experienced vendors/partners</li> </ul>
<b>Governance</b>		
Cloud Data Sources	How to guarantee trust from cloud data sources?	<ul style="list-style-type: none"> <li>» Manage directly</li> <li>» Audit</li> <li>» Assume</li> </ul>
Data Quality	How to clean, enrich, dedup unstructured data?	<ul style="list-style-type: none"> <li>» <i>Use statistical sampling</i></li> <li>» <i>Normal techniques</i></li> </ul>
Data Quality	How frequently do we need to re-validate content structure?	<ul style="list-style-type: none"> <li>» <i>Upon every receipt</i></li> <li>» <i>Periodically</i></li> <li>» <i>Manually or automatically</i></li> </ul>
Security Policies	How to extend enterprise data security policies?	<ul style="list-style-type: none"> <li>» <i>Inherit enterprise policies</i></li> <li>» <i>Copy enterprise policies</i></li> <li>» <i>Only authorize specific tools/access points</i></li> <li>» <i>Limited to monitoring security logs</i></li> </ul>

## 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.

## Taking an Enterprise Architecture Approach

A best practice is to take an enterprise architecture (EA) approach to transformational initiatives in order to maintain business alignment and maximize return on investment. Big Data is a transformation initiative. According to McKinsey, “The payoff from joining the Big-Data revolution is no longer in doubt. The broader research suggests that when companies inject data and analytics deep into their operations, they can deliver productivity and profit gains that are higher than those of the competition.”

Typically, organizations know the set of capabilities they wish to deliver and they can articulate an end-to-end roadmap. They can identify the platforms and resources needed to accomplish the objectives. They’ve got a firm grasp on the required People, Process, and Technology. Big Data disrupts the traditional architecture paradigm. With Big Data, organizations may have an idea or interest, but they don’t necessarily know what will come out of it. The answer or outcome for an initial question will trigger the next set of questions. It requires a unique combination of skill sets, the likes of which are new and not in abundance. The architecture development process needs to be more fluid and very different from SDLC-like architecture process so many organizations employ today. It must allow organizations to continuously assess progress, correct course where needed, balance cost, and gain acceptance.

The Oracle Enterprise Architecture Development Process (OADP) was designed to be a flexible and a “just-in-time” architecture development approach. It also addresses the People, Process, and Technology aspects of architecture; hence, it is well-suited to building out a holistic Big Data Architecture incrementally and iteratively. The Technology footprint should be familiar to followers of TOGAF, incorporating business architecture, application architecture, information architecture, and technology architecture. Oracle Enterprise Architects contribute their industry experience across nearly every technology stack in addition to their expertise in the Oracle portfolio.

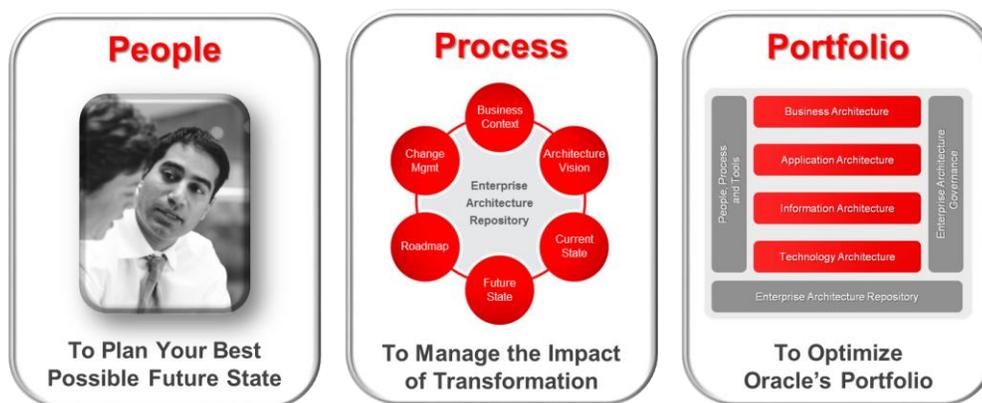


Figure 1: People, Process, and Portfolio Aspects of Oracle's Enterprise Architecture Program

Key people in the process include business project sponsors and potential users (including data scientists), enterprise architects, and Big Data engineers. Data scientists mine data, apply statistical modeling and analysis, interpret the results, and drive the implication of data results to application and to prediction. Big Data administrators and engineers manage and monitor the infrastructure for security, performance, data growth, availability, and scalability.

The six key steps in the process outlined here are to establish business context and scope, establish architecture vision, assess the current state, assess the future state and economic model, define a strategic roadmap, and establish governance over the architecture. This tends to be a closed loop process as illustrated since successful deployment leads to new ideas for solving business needs. We'll next briefly walk through these steps.

## Step 1 – Establish Business Context and Scope

In this step, we incubate ideas and uses cases that would deliver value in the desired timeframe. This is typically the most difficult step for organizations as they frequently experience the “we don’t know what we don’t know” syndrome. It is also challenging to put boundaries around scope and time so as to avoid “boiling the ocean” or scope creep.

Oracle Big Data practitioners and Business Architects are a valuable resource during this step, helping to uncover potential business value and return on investment that a project might generate.

## Step 2 – Establish an Architecture Vision

We illustrate the steps in establishing an architecture vision in Figure 2.

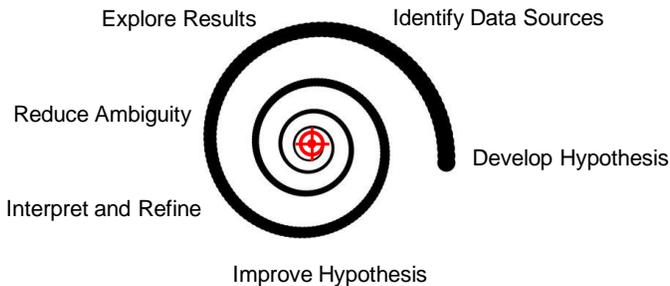


Figure 2: Steps in Establishing an Architecture Vision

We begin our architecture vision by developing the hypothesis or the “Big Idea” we created in the previous step. Based on the problem we are solving, we can now identify the data sources, including how we will acquire, access, and capture the data. We next outline how we’ll explore the data producing results including how we’ll reduce the data and use information discovery, interactive query, analytics, and visualization tools. We apply these to reduce ambiguity, for example by applying statistical models to eliminate outliers, find concentrations, and make correlations. We next define how and who will interpret and refine results and establish an improved hypothesis.

## Step 3 – Assess the Current State

As we assess our current state, we return to the technology illustration in Figure 1 as a guide. We evaluate our current business architecture including processes, skill sets, and organizations already in place. We review our application architecture including application processes. When evaluating the information architecture, we review current assets, our data models, and data flow patterns. Of course, we also evaluate the technology architecture including platforms and infrastructure that might include traditional data warehouses and Big Data technologies already deployed. We also look at other aspects in the current footprint such as platform standards, system availability and disaster recovery requirements, and industry regulations for data security that must be adhered to.



## Step 4 – Establish Future State and Economic Model

In our future state planning, we evaluate how our business architecture, application architecture, information architecture, and technology architecture will need to change consistent with our architecture vision. We begin to determine how we might deliver business value early and often to assure project success and evaluate the technology changes and skills that will be needed at various steps along the way. At this point, we likely evaluate whether Cloud-based solutions might provide a viable alternative, especially where time to market is critical. As part of that evaluation, we take another look at where critical data sources reside today and are likely to reside in the future. And we will evaluate the impact any current platform standards already in place, system availability and disaster recovery mandates, and industry regulations for data security that must be adhered to in our future state.

## Step 5 – Develop a 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.

The roadmap should contain:

- » A set of architectural gaps that exist between the current and future state
- » A cost-benefit analysis to close the gaps
- » The value from each phase of the roadmap and suggestions on how to maximize value while minimizing risk and cost
- » Consideration of technology dependencies across phases
- » Flexibility to adapt to new business priorities and to changing technology
- » A plan to eliminate any skills gaps that might exist when moving to the future state (e.g. training, hiring, etc.)

## Step 6 – Establish Governance over the Architecture

Governance, in the context of Big Data, focuses on who has access to the data and data quality but also on whether data quality measures are desirable before analysis takes place. For example, 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.

Focus might also be applied to determining when automated decisions are appropriate and when human intervention and interpretation are required. 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. Thus, in most deployment examples today, there is a hybrid strategy leveraging Big Data solutions for exploration of all data (regardless of quality) among a small group of trusted data scientists and traditional data warehouses as the repository of truth and cleansed data for ad-hoc queries and reporting to the masses.

## Big Data Reference Architecture Overview

### Traditional Information Architecture Capabilities

To understand the high-level architecture aspects of Big Data, let's first review 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.

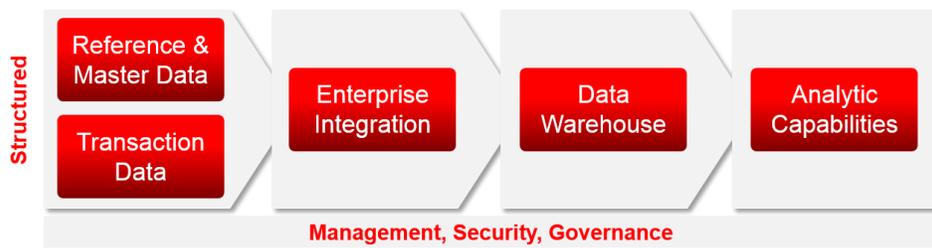


Figure 3: Traditional Information Architecture Components

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.

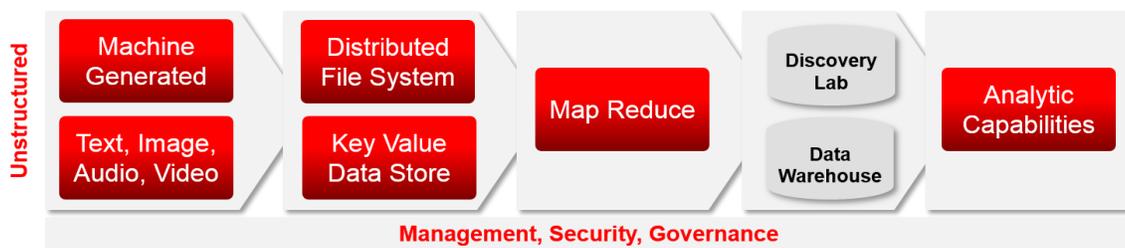


Figure 4: Big Data Information Architecture Components

In addition to the new components, new architectures are emerging to efficiently accommodate new storage, access, processing, and analytical requirements. First, is the idea that specialized data stores, fit for purpose, are able to store and optimize processing for the new types of data. A Polyglot strategy suggests that big data oriented architectures will deploy multiple types of data stores. Keep in mind that a polyglot strategy does add some complexity in management, governance, security, and skills.

Second, we can parallelize our MPP data foundation for both speed and size, this is crucial for next-generation data services and analytics that can scale to any latency and size requirements. With this Lambda based architecture we're now able to address fast data that might be needed in an Internet of Things architecture.

Third, MPP data pipelines that allow us to treat data events in a moving time windows at variable latencies; in the long run this will change how we do ETL for most use cases.

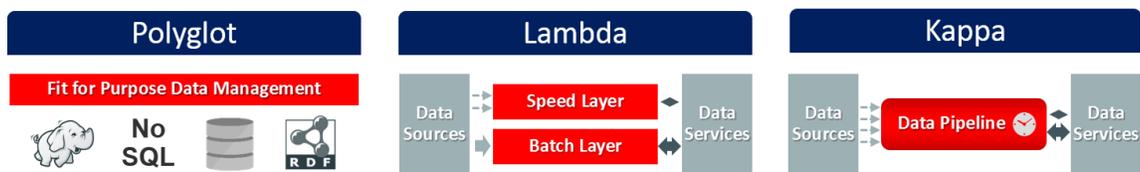


Figure 5: Big Data Architecture Patterns

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.

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 ecosystem. 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

Oracle has a defined view of a unified reference architecture based on successful deployment patterns that have emerged. Oracle's Information Management Architecture, shown in Figure 6, illustrates key components and flows and highlights the emergence of the Data Lab and 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.

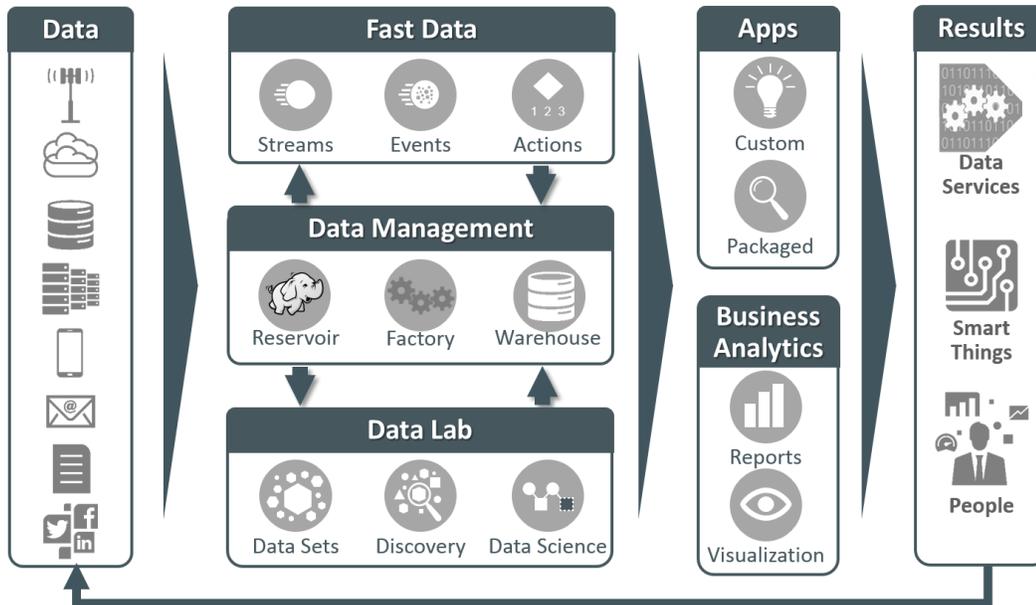


Figure 6: Conceptual model for *The Oracle Big Data Platform* for unified information management

A description of these primary components:

- » **Fast Data:** Components which process data in-flight (streams) 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.
- » **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.
- » **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.
- » **Warehouse:** Large scale formalized and modelled business critical data store, typically manifested by a Data Warehouse or Data Marts.
- » **Data Lab:** A set of data stores, processing engines, and analysis tools separate from the data management activities to facilitate the discovery of new knowledge. Key requirements include rapid data provisioning and subsetting, data security/governance, and rapid statistical processing for large data sets.
- » **Business Analytics:** A range of end user and analytic tools for business Intelligence, faceted navigation, and data mining analytic tools including dashboards, reports, and mobile access for timely and accurate reporting.
- » **Apps:** A collection of prebuilt adapters and application programming interfaces that enable all data sources and processing to be directly integrated into custom or packaged business applications.

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.

## Enterprise Information Management Capabilities

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

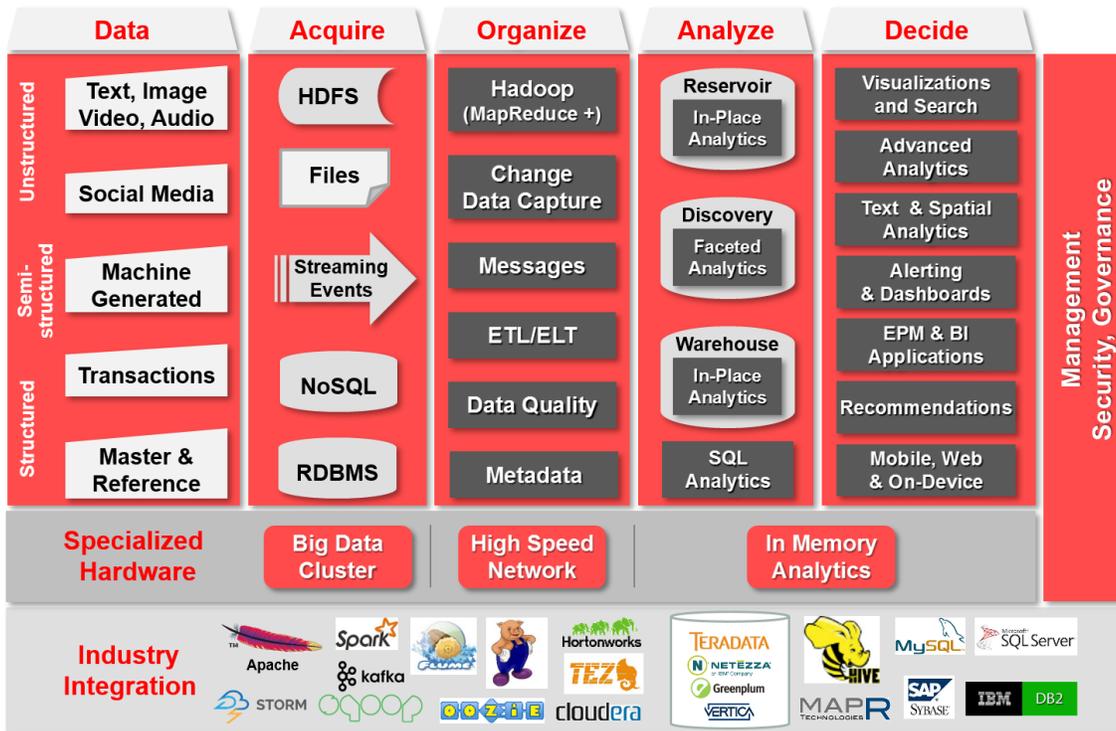


Figure 7: Oracle's Unified Information Management Capabilities

A brief overview of these capabilities appears 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

Required Big Data architecture capabilities can be delivered by a combination of solutions delivered by Apache projects ([www.apache.org](http://www.apache.org)) and Oracle Big Data products. Here we will take a look at some of the key projects and 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](#))

- » Flume provides a distributed, reliable, and available service for efficiently moving large amounts of log data and other data. It captures and processes data asynchronously. A data event will capture data in a queue (channel) and then a consumer will dequeue the event (sink) on demand. Once consumed, the data in the original queue is removed which forces writing of 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](#))

- » Storm provides a distributed real-time, parallelized computation system that runs across a cluster of nodes. The topology is designed to consume streams of data and process those streams in arbitrarily complex



ways, repartitioning the streams between each stage of the computation. Use cases can include real-time analytics, on-line machine learning, continuous computation, distributed RPC, ETL, and more.

Apache Kafka ([Click for more information](#))

- » Kafka is an Apache publish-subscribe messaging system where 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 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 Stream Explorer ([Click for more information](#))

- » Stream Explorer can process multiple event streams, detecting patterns and trends in real time, and then initiating an action. It can be deployed as standalone, integrated in the SOA stack, or in a lightweight fashion on embedded Java. Stream Explorer can ensure downstream applications and service-oriented and event-driven architectures are driven by true, real-time intelligence.

Oracle GoldenGate ([Click for more information](#))

- » Golden Gate enables log-based change data capture, distribution, transformation, and delivery. It has support for heterogeneous data management systems and operating systems and provides bidirectional replication without distance limitation. GoldenGate can ensure transactional integrity and reliable data delivery and fast recovery after interruptions.

### Distributed File System Capability

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

- » HDFS is an Apache open source distributed file system that runs on high-performance commodity hardware and appliances built with such (e.g. Oracle Big Data Appliance). It is designed to be deployed highly scalable nodes and associated storage. HDFS provides automatic data replication (usually deployed as triple replication) for fault tolerance. Most organizations deploy data and manipulate it directly in HDFS for write once, read many applications such as those common in analytics.

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

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

- » Apache HBase is designed to provide random read/write access to very large non-relational tables deployed in Hadoop. Among the features are linear and modular scalability, strictly consistent reads and writes, automatic and configurable sharding, automatic failover between Region Servers, base classes for Hadoop MapReduce jobs and Apache HBase tables, Java API client access, and a REST-ful web service.

Apache Kudu: ([Click for more information](#))

- » Kudu provides a combination of fast inserts/updates and efficient columnar scans to enable multiple real-time analytic workloads across a single storage layer. As a more recent complement to HDFS and Apache HBase, Kudu gives architects the flexibility to address a wider variety of use cases without exotic workarounds. For example, Kudu can be used in situations that require fast analytics on fast (rapidly changing) data. Kudu promises to lower query latency significantly for Apache Impala and Apache Spark initially, with other execution engines to come.

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

- » For high transaction environments (not just append), where data models call for table based key-value pairs, and consistency is defined by policies needing superior availability in NoSQL environments, Oracle's NoSQL DB excels in web scale-out and click-stream type low latency environments.
- » Oracle NoSQL Database is designed as a highly scalable, distributed database based on Oracle Berkeley DB. Sleepycat Software. Oracle NoSQL Database is a general purpose, enterprise class key value store that adds an intelligent driver on top of an enhanced distributed Berkeley database. This intelligent driver keeps track of the underlying storage topology, understands and uses data shards where necessary, and knows where data can be placed in a clustered environment for the lowest possible latency. Unlike competitive solutions, Oracle NoSQL Database is easy to install, configure and manage. It supports a broad set of workloads and delivers enterprise-class reliability backed by enterprise-class Oracle support.
- » Using Oracle NoSQL Database allows data to be more efficiently acquired, organized and analyzed. Primary use cases include low latency capture plus fast querying of the same data as it is being ingested, most typically by key-value lookup. Examples of such use cases include Credit Card transaction environments, high velocity, low latency embedded device data capture, and high volume stock market trading applications. Oracle NoSQL Database can also provide a near consistent Oracle Database table copy of key-value pairs where a high rate of updates is required. It can serve as a target for Oracle GoldenGate change data capture and used in conjunction with event processing using Oracle Stream Explorer and Oracle Real Time Decisions. The product is available in both an open source community edition and an enterprise edition for large distributed data centers. The latter version is part of the Big Data Appliance.
- » Oracle NoSQL DB Enterprise Edition distinguishes itself from the NoSQL Community Edition version with Oracle stack integration. Specifically the EE edition includes, or is required with the following:
  - » Oracle Database External Table integration
  - » Oracle Big Data SQL integration
  - » Oracle Coherence integration
  - » Oracle Stream Explorer (Event Processing) integration
  - » Oracle Enterprise Manager integration
  - » Oracle Semantic Graph integration
  - » Oracle Wallet integration
  - » SNMP administrative interface

## Processing Capability

Apache Hadoop: ([Click for more information](#))

- » The *Apache Hadoop* software library is a framework that allows for the distributed processing of large data sets across clusters of nodes using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Processing capabilities for query and analysis are primarily delivered using programs and utilities that leverage MapReduce and Spark. Other key technologies include the Hadoop Distributed File System (HDFS) and YARN (a framework for job scheduling and cluster resource management).

MapReduce: ([Click for more information](#))

- » MapReduce relies on a linear dataflow structure evenly allocated as highly distributed programs. Apache provides MapReduce in Hadoop as a programming model and an implementation designed to process large data sets in parallel where the data resides on disks in a cluster.

Apache Spark: ([Click for more information](#))

- » Spark provides programmers with an application programming interface centered on a data structure called the resilient distributed dataset (RDD). Spark's RDDs function as a fast working set or cache for distributed programs that in essence provide a form of distributed shared memory. The speed of access and availability of this working dataset facilitates high performance implementations of two common algorithm paradigms: (1) iterative algorithms, that must reuse and access data multiple times, as well as (2) newer analytics and exploratory type of processing that are akin to processing and query models found with traditional databases. Leading the class in the category of iterative algorithms is machine learning.

## Data Integration Capability

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

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

- » The Oracle Loader for Hadoop enables parallel high-speed loading of data from Hadoop into an Oracle Database. Oracle Data Integrator Enterprise Edition in combination with the Big Data Connectors enables high performance data movement and deployment of data transformations in Hadoop. Other features in the Big Data Connectors include an Oracle SQL Connector for HDFS, Oracle R Advanced Analytics for Hadoop, and Oracle XQuery for Hadoop.

## SQL Data Access

Oracle Big Data SQL ([Click for more information](#))

- » Big Data SQL enables Oracle SQL queries to be initiated against data also residing in Apache Hadoop clusters and NoSQL databases. Oracle Database 12c provides the means to query this data using external tables. Smart Scan capabilities deployed on the other data sources minimizes data movement and maximizes performance. Because queries are initiated through the Oracle Database, advanced security, data redaction, and virtual private database capabilities are extended to Hadoop and NoSQL databases.

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

- » Hive provides a mechanism to project structure onto Hadoop data sets and query the data using a SQL-like language called HiveQL. The language also enables traditional MapReduce programmers to plug in their custom mappers and reducers when it is inconvenient or inefficient to express this logic in HiveQL. It only contains metadata that describes data access in Apache HDFS and Apache HBase, not the data itself. More recently, HiveQL query execution is commonly used with Spark for better performance.



Apache Impala - Cloudera ([Click for more information](#))

- » With Impala, you can query data, whether stored in HDFS or Apache HBase, using typical SQL query functions of select, join, and various aggregate functions an order of magnitude faster than using Hive with MapReduce. To avoid latency and improve application speed, Impala circumvents MapReduce to directly access the data through a specialized distributed query engine that is very similar to those found in commercial parallel RDBMSs.

### Statistical Analysis Capability

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

- » R is a programming language for statistical analysis ([Click here for Project R](#)). Oracle first enabled running R algorithms in parallel without the need to move data out of the data store in the Oracle Database ([Click here for Oracle R Enterprise](#)). Oracle R Advanced Analytics for Hadoop (ORAAH) is bundled into Oracle's Big Data Connectors and provides high performance in-Hadoop statistical analysis capabilities leveraging Spark and MapReduce.

### Spatial & Graph Capability

Oracle Big Data Spatial and Graph: ([Click for more information](#))

- » Oracle Big Data Spatial and Graph provides analytic services and data models supporting Big Data workloads on Apache Hadoop and NoSQL database technologies. Oracle Big Data Spatial and Graph includes two main components: A property graph database and 35 built-in graph analytics that discover relationships, recommendations and other graph patterns in big data and a wide range of spatial analysis functions and services to evaluate data based on how near or far something is to one another, whether something falls within a boundary or region, or to process and visualize geospatial map data and imagery.

### Information Discovery

Oracle Big Data Discovery ([Click for more information](#))

- » Oracle Big Data Discovery provides an interactive interface into Apache Hadoop (e.g. Cloudera, Hortonworks) to easily find and explore data, quickly transform and enrich data, intuitively discover new insights by combining data sets, and share the results through highly visual interfaces.

### Business Intelligence

Oracle Business Intelligence Suite ([Click for more information](#))

- » Oracle Business Intelligence Suite provides a single platform for ad-hoc query and analysis, report publishing, and data visualization from you workstation or mobile device. Direct access to Hadoop is supported via Hive and Impala.

### Real-Time Recommendation Engine

Oracle Real-Time Decisions ([Click for more information](#))

- » Using Oracle RTD, business logic can be expressed in the form of rules and self-learning predictive models that can recommend optimal courses of action with very low latency using specific performance objectives. This is accomplished by using data from different “channels”. Whether it's on the web in the form of click-stream data, in a call center where computer telephony integration provides valuable insight to call center agents, or at the point-of-sale, Oracle RTD can be combined with the complex event processing of Oracle Streams Explorer to create a complete event based Decision Management System.

## Oracle Big Data Cloud Services

As noted previously in this paper, organizations are eagerly deploying Big Data processing, storage and integration technologies in on premises and Public Cloud-based solutions. These solutions are often seen as providing faster time to market, more flexibility in deployment, and as cost effective alternatives to further in-house investments in undifferentiated skills and infrastructure. Cloud-based Big Data solutions are hosted on Infrastructure as a Service (IaaS), delivered as Platform as a Service (PaaS), or as Big Data applications (and data services) via Software as a Service (SaaS) manifestations.

Oracle IaaS deployment might include Oracle and non-Oracle software components and be deployed on premises or in the Oracle Public Cloud. Installation and management of “platform” software is typically your responsibility in IaaS deployment models.

Key Oracle PaaS Cloud Services that might become part of your Big Data deployment strategy can include:

- » Big Data Cloud Service: Hadoop and Spark are delivered as an automated Cloud Service. This offering includes the Cloudera Data Hub Edition, Oracle Big Data Connectors, Oracle Spatial and Graph, Oracle Data Integrator with Advanced Big Data Option and Database Cloud Service (via Connectors).
- » Big Data Discovery Cloud Service: Hosted in the Big Data Cloud Service, Big Data Discovery is used in the exploration and transformation of data residing in Hadoop and can help uncover new business insight that the data can provide through its discovery capabilities.
- » Business Intelligence Cloud Service: Provides ad-hoc query and analysis dashboards and data visualization to data residing in Schema as a Service, Database as a Service, through REST APIs, and from various other sources.
- » Big Data SQL Cloud Service: An optimal solution for using Oracle Database SQL to query data residing in the Big Data Cloud Service linked through a data warehouse residing in the Exadata Cloud Service.
- » Exadata Cloud Service: Streamlines implementation and management of Oracle relational databases while greatly improving query performance by using additional optimization provided by Exadata Storage Server Software.
- » Big Data Preparation Cloud Service: Combines machine learning with a Natural Language Processing engine to ingest, enrich, publish, govern and monitor data. During the ingest and import process, schema and duplicate data can be detected, data can be cleansed normalized, and sensitive data can be detected and masked. The enrich process includes profiling, annotation, data classification, semantic enrichment, and missing data interpolation. Data can be published on demand, scheduled, or event driven. Govern and monitoring capabilities can include automated alerts, system controls, and reusable user policies.
- » Internet of Things Cloud Service: Provides device virtualization, endpoint management and an event store, supports high speed messaging and stream processing, and provides enterprise connectivity including support of REST APIs.

Oracle SaaS offerings in the Oracle Public Cloud are typically not easily identified as running on Hadoop since they are provided as applications. For example, Oracle’s Marketing Cloud includes BlueKai technology that is deployed on a Hadoop cluster.

Discussions about on premises and Public Cloud deployment scenarios often revolve around infrastructure considerations, security, and networking required between sites. Of course, leveraging Oracle’s Public Cloud puts infrastructure considerations (floor space, equipment, power consumption, and environmental control) in Oracle’s hands and can meet the primary goals of many organizations’ Cloud redeployment strategies. Networking security (including firewalls, encryption, etc.) and management of the various software layers and data (e.g. who has access) is a consideration regardless of location. If data is being moved between sites, such as from on premises to a Public Cloud or vice versa, then data volumes and needed data transfer rates must be considered in the architecture of the envisioned solution. Oracle’s architects can provide guidance in all of these areas.

## Highlights of Oracle's Big Data Architecture

In this section, we will further explore some of Oracle's product capabilities previously introduced.

### Big Data SQL

Oracle Big Data SQL provides 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 (represented in Figure 8), 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.

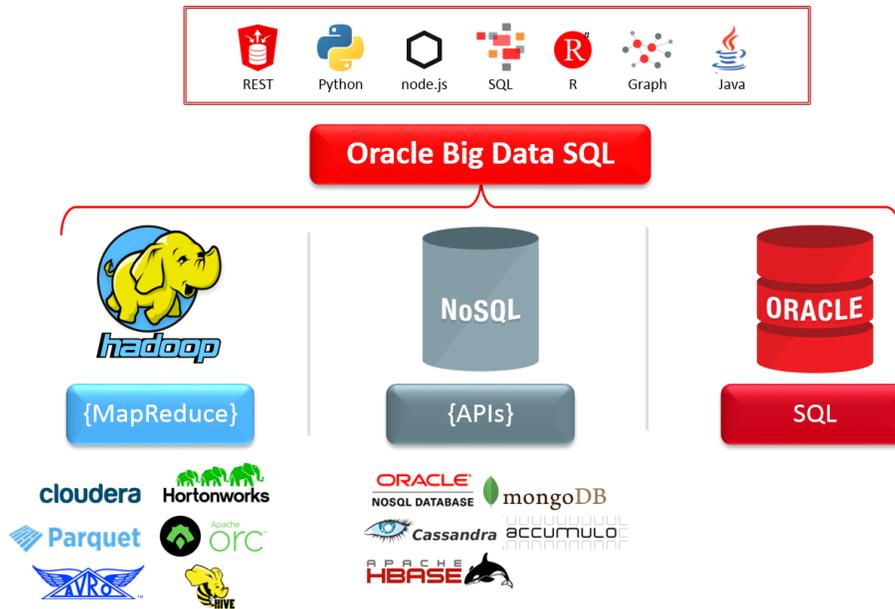


Figure 8: Big Data SQL offers a standards-based SQL language interface accessible from many programming languages

Oracle Big Data SQL is a software product that has a component that runs inside the Hadoop cluster and a component that runs inside the Oracle database. Big Data SQL enables one SQL query to join data residing in Hadoop (Cloudera, Hortonworks), NoSQL, and Oracle databases simultaneously. Big Data SQL provides a familiar processing interface into Hadoop for the vast array of SQL programmers and SQL tools in use today.

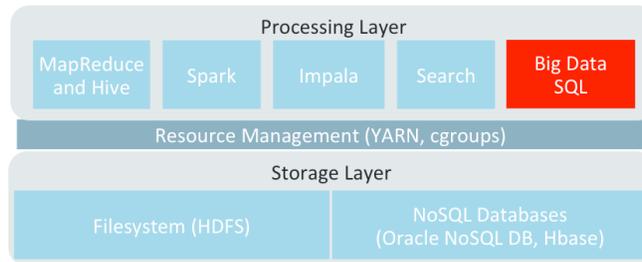


Figure 9: Big Data SQL operates efficiently and natively alongside other Hadoop services

*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. Additional capabilities include: Automatic discovery of Hive table metadata, automatic translation from Hadoop types, automatic conversion from any input format, and fan-out parallelism across the cluster

- **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

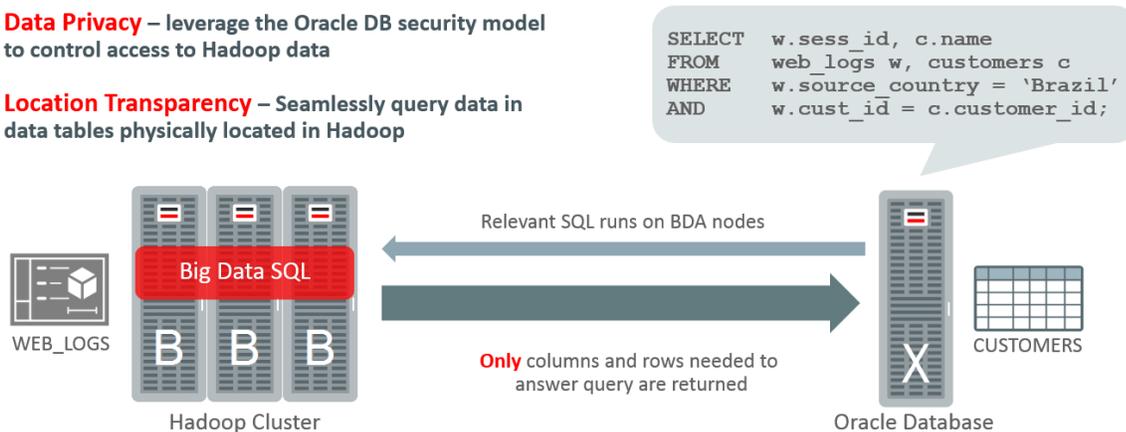


Figure 10: One high performance deployment option: Oracle Big Data Appliance connected to Oracle Exadata over Infiniband. Shows Big Data SQL execution with SQL join showing data source transparency

Key benefits of Big Data SQL are:

- » Leverage Existing SQL Skills - Users and developers are able to access data in Hadoop and NoSQL database 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 accesses 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 then operate in Hadoop and NoSQL databases.
- » Performance Optimization - 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 interconnection to the compute layer. For example, storage indexes provide query speed-up through transparent I/O elimination of HDFS Blocks.
- » When deployed on the Oracle Big Data Appliance and Oracle Exadata, InfiniBand’s high bandwidth enables queries to return results to the Oracle with optimum performance.
- » Faster 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. Familiar 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.

Big Data SQL also includes two useful utilities. *Copy2BDA* enables you to rapidly copy tables from an Oracle database into Hadoop. *Oracle Table Access for Hadoop and Spark (OTA4H)* is an Oracle Big Data Appliance feature that converts Oracle tables to Hadoop and Spark . OTA4H allows direct, fast, parallel, secure and consistent access to master data in Oracle database using Hive SQL and Spark SQL. There are a set of APIs that support SerDes, HCatalog, InputFormat, and StorageHandler.

## Data Integration

With the surging volume of data being sourced from an ever growing variety of data sources and applications, 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, in order for big data to deliver value, it has the same requirements for quality, governance, and confidence as traditional data sources.

The growing data volume from structured and semi-structured data sources is leading many to explore a Big Data solution as an augmentation to an existing ETL environment. Many enterprise data warehouses consume over half of their processing cycles performing batch ETL. Real-time or near real-time feeds further increase the processing requirements, leading many to settle for a traditional nightly batch load. Enterprise Data Warehouse processing cycles are better spent delivering value with actual analytics, instead of transformation. Big Data solutions represent an economic way to off-load many of these processing intensive jobs, freeing resources on the EDW for analytics.

Oracle's family of Integration Products supports nearly all of the Apache Big Data technologies as well as many non-Oracle products. Core integration capabilities support an entire infrastructure around data movement and transformation that include integration orchestration, data quality, data lineage, and data governance. The modern data warehouse is no longer confined to a single physical solution, so complementing technologies that enable this new logical data warehouse are more important than ever.

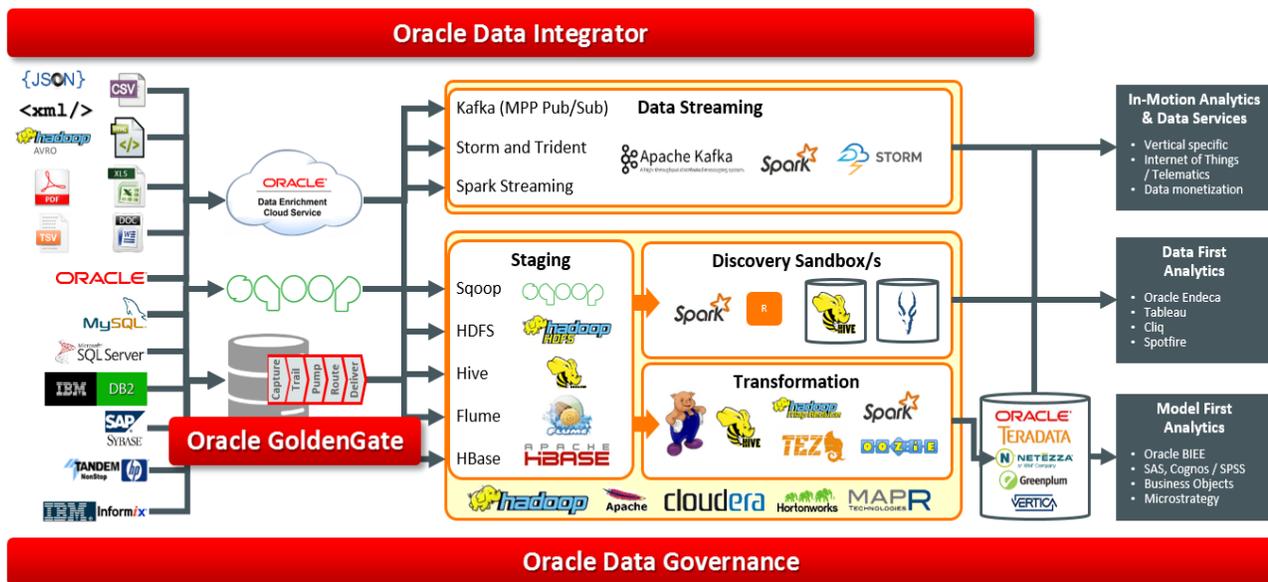


Figure 11: Oracle Open Integration Architecture

As Big Data solutions continue to mature, so do tools supporting its integration with other enterprise platforms. ETL tools such as Oracle Data Integrator are continually evolving to support Big Data as both a destination for data and an intermediary transformation powerhouse. SQL-like and SQL-on-Hadoop technologies such as Spark SQL and Hive allow SQL transformations to be more easily pushed to an Apache Hadoop platform (e.g. Cloudera and Hortonworks), and powerful, flexible technologies like Spark, Pig, and MapReduce can enable complex transformations. For example, Oracle Data Integrator transformations can be deployed into Hadoop and the cluster can be used as a high-speed transformation engine and significant ETL processing workload can be removed from the data warehouse.



High speed ingestion tools such as Oracle's GoldenGate, Sqoop, and Flume can deliver data to Hadoop and make it an efficient landing zone for data sources. These tools can help enable real-time/near real-time load and online archiving. The continuous collection of data from source systems and technologies such as Kafka, Storm, and Spark Streaming allow action real-time processing of data. This data can be transformed and streamed into an RDBMS, or event processing can take action on the data.

As data volumes grow and Big Data storage costs decrease, using Hadoop clusters as an Enterprise Data Warehouse online deep archive is becoming a popular use case. Retiring records to Hadoop or retaining them after initial ETL are two ways to accomplish this. Utilizing query tools such as Oracle's Big Data SQL can enable analysts to reach into the online archive to access data they would previously have to request be restored from tape archives.

## 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 support Apache Hadoop (e.g. Cloudera and Hortonworks) and 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 into an Oracle Database, as defined using a graphical user interface.
- » Oracle R Advanced Analytics for Hadoop: Provides the ability to run R scripts in Hadoop directly against data stored there leveraging Spark and MapReduce for parallelism.
- » Oracle XQuery for Hadoop: Provides native XQuery access to HDFS and the Hadoop parallel framework.



## Oracle Big Data Preparation

Oracle Big Data Preparation (BDP) Cloud Service gives you an easy-to-use way to work with your data. With its coordinated features, you can automate, streamline, and guide the error-prone process of data ingestion, preparation, repair, enrichment, and governance without costly manual intervention. This service is all about presenting an easy-to-use way to interact and work with data. To make sense of data, you define a structure and correlate the disparate data sets. This important step involves both understanding and standardizing your data. Big Data Preparation facilitates the development lifecycle of data. It provides the following capabilities:

- **Ingest:** Automatically ingest structured, semi-structured and unstructured data from multiple sources in a variety of formats. Within the ingestion step one can create standard statistical analysis of numerical data and frequency and term analysis of text data. Data can be cleaned, duplicates identified and data can be repaired to remove inconsistencies. At ingestion BDP can detect and identify schema and metadata this is explicitly defined in headers, fields or tags.
- **Enrich:** Create statistical profiles of your data, identify attribute and property schemata and automatically enrich data with a reference knowledge base. BDP's machine learning system working with reference data sets will make a recommendation on how best to enrich and correlate the data.
- **Govern:** Interactive dashboard all the creation of user policies and system controls, adjust to automated alerts and viewing of job details
- **Publish:** Define sources and targets, schedule events and decide which formats you want to use to export your data.

Finally, BDP can automate this process on a daily, weekly or monthly basis against a predetermined data source. RESTful APIs help automate the entire data preparation process, from file movement to preparation to publishing.

## Oracle Stream Explorer

To address the Fast Data requirements of the Oracle Big Data reference architecture, Oracle includes an integrated, complex event processing solution that can source, process, and publish events. Oracle Stream Explorer provides the ability to join incoming streamed events with persisted data, thereby delivering contextually aware filtering, correlation, aggregation, and pattern matching. Oracle Stream Explorer can support very low latency and high data volume environments in an application context.

Oracle Stream Explorer is based on an open architecture that supports industry-standards including ANSI SQL, Java, Spring DM and OSGi. It includes a real-time visual development environment to facilitate developing effective continuous SQL. As a platform, Stream Explorer ensures that your IT team can develop event-driven applications without the hurdle of specialized training or unique skill-set investment.

Oracle Stream Explorer's 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
- » Language constructs for Fast Data integration with Hadoop and Oracle NoSQL
- » 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 for Java Embedded provides a uniquely small disk and memory footprint enabling distributed intelligence within Internet-of-Things infrastructures

Oracle Stream Explorer also targets a wealth of industries and functional including the following 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, aids in a fraud detection architecture, monitoring and reporting of financial securities trading and calculation of 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.
- » *Supply Chain and Logistics*: Ability to track shipments in real-time and detect and report on potential delays in arrival.
- » *IT Systems*: Ability to detect failed applications or servers in real-time and trigger corrective measures.

## 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.

Apache Hadoop projects enable data at rest and network encryption capabilities. For example, the Cloudera Distribution of Hadoop includes enterprise-grade authentication (Kerberos), authorization (LDAP and Apache Sentry project), and auditing 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:

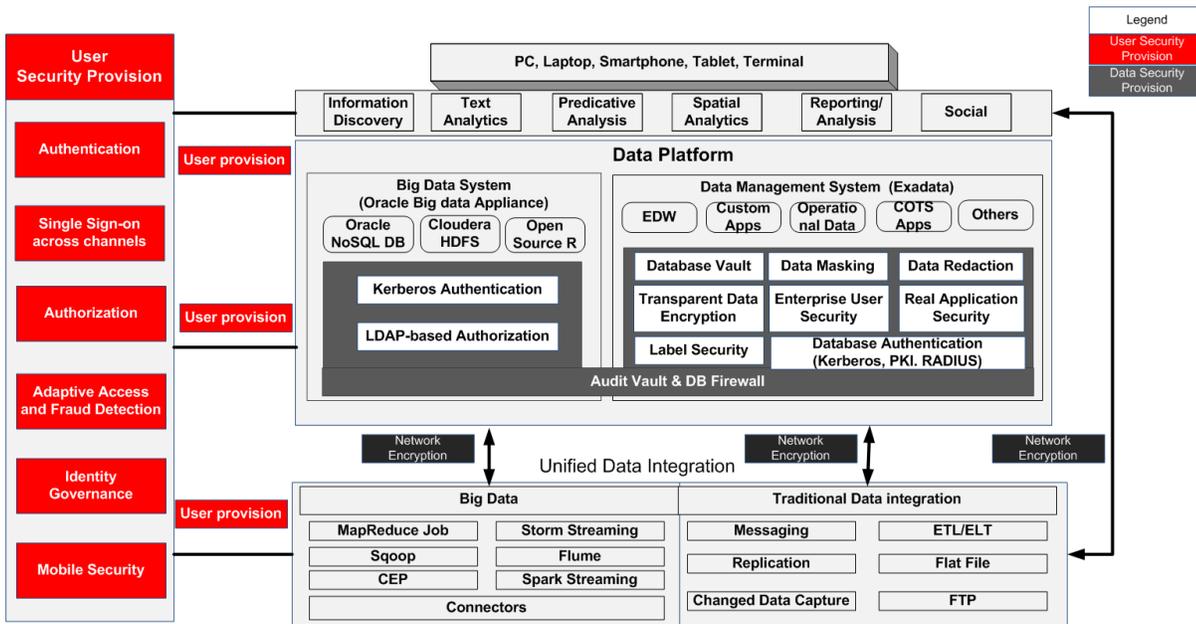


Figure 12: Oracle Security Architecture for the Oracle Big Data Platform

The spectrum of data security capabilities are:

- » Authentication and authorization of users, applications and databases (typically using Kerberos)
- » 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

Analyzing the data to reveal insights that help organizations meet their business objectives is critical for success in an increasingly data-driven economy. The type of analytics carried out may span a spectrum of data science, from traditional business intelligence, to information discovery or data mining, and culminating in machine learning and advanced analytics. An organization, mature in its analytics capabilities, will employ all three forms of analytics since they complement one another. The common thread is always how the analytics helps the different lines of business meet their business objectives quickly and easily.

**Business Intelligence (BI)** provides proven answers to known questions - the key performance indicators (KPIs), reports, and dashboards – providing a view into the health of business operations. BI users know the answer they are looking for, and use tools such as Oracle Business Intelligence to quickly identify structured datasets and combine them to generate reports. They setup dashboards providing decision makers with situational awareness about their company's operations for monitoring general trends and spotting unexpected changes.

**Information Discovery**, also referred to as *Data Mining*, focuses on explaining the root causes of what is observed by the business. Often this involves discovering previously unknown relationships amongst the various business indicators. Discovery also expands beyond the traditionally structured datasets and into semi-structured (e.g. application logs) or unstructured (e.g. customer reviews) data. For example, shifts in sentiment around a brand on social media (i.e. Twitter, Facebook, etc.) may have a strong correlation with the sales for that brand. Traditionally sentiment analysis lay in the realms of advanced analytics due to the sophisticated nature of Natural Language Processing (NLP). Oracle Social Relationship Management and sentiment analysis algorithms within Oracle Big Data Discovery (BDD), simplify the visualization of social sentiment's correlation with known business metrics. BDD makes data science of discovery agnostic to the volume of data, providing a natural intuitive interface for visual exploration of data backed by the power of Hadoop.

**Advanced Analytics** or *machine learning* algorithms analyze the data to build mathematical models that describes the patterns or relationships within the data. Once learnt, the mathematical models can then be used to explain the relationships or make predictions about the future. For example, a machine learnt model that analyzes the real-time sensor data stream to predict the likelihood of failure can provide sufficient warning so that preventative measures can be taken to avoid costly production downtime. Oracle's philosophy has always been to enable analytics and actions where data resides – whether it is in the database or Hadoop data lakes. Data Scientists use R, a popular statistical modeling environment, for machine learning modeling against data in database via Oracle R Enterprise or against the data on the Hadoop cluster via the Oracle R Advanced Analytics for Hadoop (ORAAH). ORAAH using Spark can provide 100 to 200 times speed up for training generalized linear regression and neural network models over pure MapReduce implementations. This allows data scientists to build models on large volumes of data even faster. Once a model has been trained using R, it can be deployed into production via the Database, Hadoop or Oracle Stream Explorer to make predictions on real-time event streams.

A summary comparison appears in the following chart:

## COMPARISON OF ORACLE BUSINESS INTELLIGENCE, INFORMATION DISCOVERY AND ADVANCED ANALYTICS

	Oracle <a href="#">Business Intelligence Suite</a>	Oracle <a href="#">Oracle Big Data Discovery</a>	Oracle <a href="#">Advanced Analytics</a>
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; direct access to Hadoop via Hive and Impala	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

Analytical insights are sometimes limited by the nature of data, and at some point organizations must augment their proprietary data with external datasets to develop even richer insights. For example, a retail brand looking to build deeper insights about its customers may be limited to only the interactions the customer has with their website or purchases made at their bricks & mortar stores. Oracle Marketing Cloud (OMC) is an example of a **Data as a Service** offering and allows retailers to understand their customer's behaviors and interests beyond what the retailer can observe, enabling better personalization for the online, offline and mobile marketing campaigns. OMC is one of the largest 3<sup>rd</sup> party marketplace for data, capturing online behavior for 700 million profiles and offline behaviors for 110 million households in the US. Whether the analytics is done in house or in the cloud, using the right kind of data leads to even richer actionable insights.

## Data Visualization

A picture is worth a thousand words (or a billion rows of data) and so data visualization is not new. It has been used for thousands of years as a way for humans to tell stories. Today, visualization is useful in understanding the massive varied data sets that reside in Hadoop clusters. Tools such as Oracle's Big Data Discovery enable visualization of data stored in Hadoop enabling exploration of the data, making new discoveries, and sharing these findings with others. But this is only the beginning of the exploration process.

Data Visualization must be provided across the entire data analysis environment including Hadoop clusters and traditional data stores. The Oracle Business Intelligence Suite provides such visualization. Figure 13 illustrates some of the traditional ways of representing data through data visualization.



Figure 13: Traditional Data Visualization delivered in Oracle Business Intelligence Suite

Today, new visualization methods have been developed to explain big data volume and variety. Figure 14 illustrates data volumes by type over time.

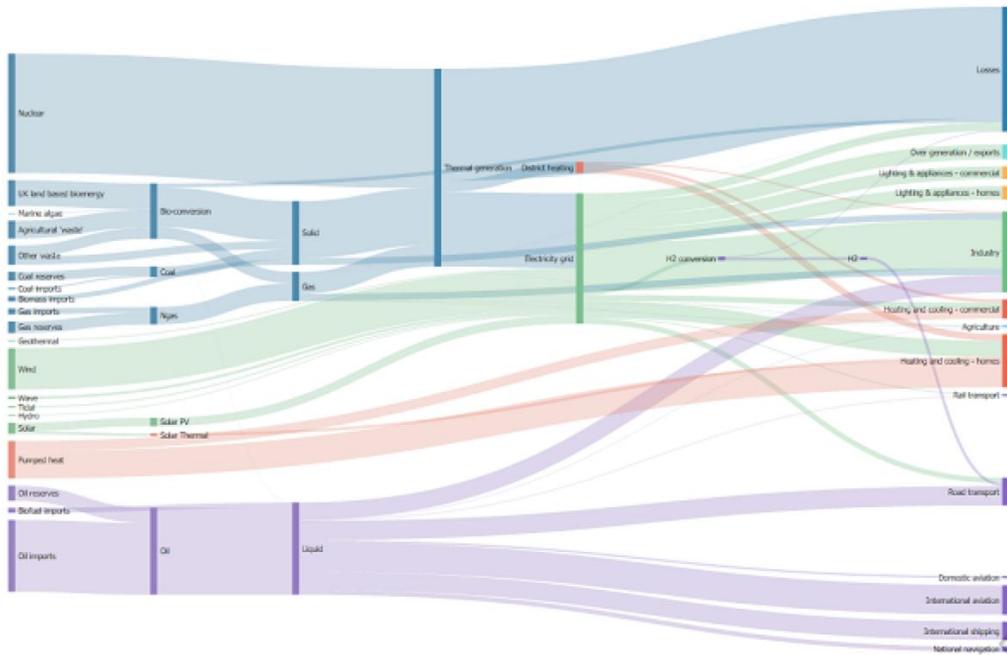


Figure 14: Data Visualization delivered in Oracle Business Intelligence Suite

Of course, geo-spatial data might be displayed as a map, date data might be displayed as a timeline, and other data sets could be displayed in a different visual that the visualization tool might recommend. The illustration below shows geo-spatial data linked to sales data to graphically show sales by region.

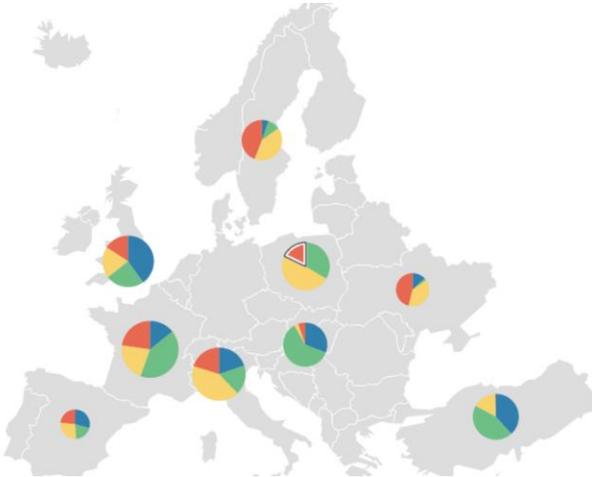


Figure 15: Data Visualization of sales data that includes spatial information

Extensive data visualizations are available in Oracle's Business Intelligent Suite. Data visualization is also included in Oracle Business Intelligence Cloud Service.

## Spatial and Graph Analysis

Advanced Graph analytics opens up a new set of possibilities for understanding relationships that go beyond traditional relational data. Previously analytics were restricted to simple one-to-one, one-to-many, or many-to-many relationships. Graph analytics allow us to analyze many-to-many-to-many and represents networks, such as the simple social network below.

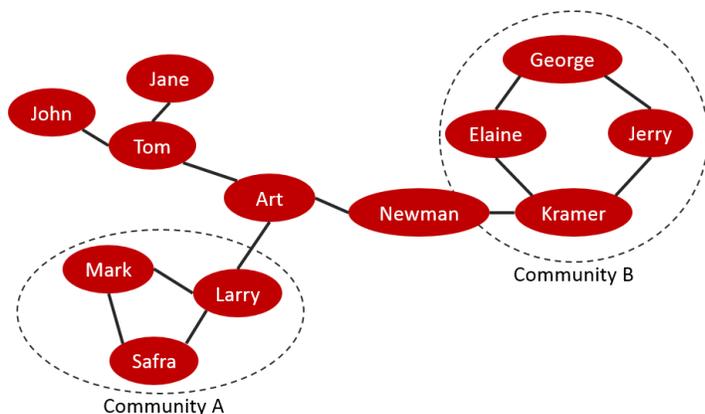


Figure 16: A Simple Social Network

In traditional analytics, representing these relationships is simple: John is friends with Tom, Tom is friends with Art, and so forth. Analyzing and finding insight into more complex relationships becomes a challenge. Oracle's Big Data Spatial and Graph capabilities feature built-in analytics to allow us to easily persist these entities and relationships in either Oracle NoSQL or Apache HBase. Graph algorithms can quickly identify that Mark, Larry, and Safra form a strong relationship or that Mark is connected to Newman through Larry and Art. While this example may seem simple, real world relationships can be dauntingly complex.

Typical use cases for Graph databases include:

- » Identify key influences, bridge entities, and clusters in social network relationships
- » Intelligently identify item affinity to enhance a customer's experience and make smarter and simpler recommendations
- » Identify patterns and connections that indicate fraudulent activity.

The spatial analytics in Oracle's Big Data Spatial and Graph enable analysis based on locations and processing of image data. Linking disparate sets of location data such as GPS coordinates, descriptive location ("near Big Ben"), addresses, and geographical names can provide deeper insights into understanding data sets containing rich location data. Image processing for raster or vector graphics allow us to efficiently analyze digital maps and photographs in a massively parallel Hadoop environment.

Typical use cases include:

- » Identifying when customers enter a certain area for location-based advertising.
- » Identify droughts, rainfall, and other changes in satellite images

## Extending the Architecture to the Internet of Things

Deployment of intelligent sensors and devices transmitting data and the intelligent capture and analysis of that data is now often referenced as the Internet of Things (IoT). Industry analysts point to tens of billions of such devices currently deployed rapidly growing into the hundreds of billions over the next few years. These devices are producing Zetabytes of data every month. The transmissions typically consist of high velocity semi-structured data streams that must land in highly scalable data management systems. Hadoop provides the ideal platform for analyzing such data.

A typical IoT capability map is shown below. Sensors and other data transmission sources are pictured in the Device Domain. Data typically flows to and through a Communications Gateway as pictured. Intelligent devices (including device status and software updates) are handled in the Device Management layer. Data is sometimes routed into NoSQL databases that front-end Hadoop clusters, or directly into the Hadoop clusters in the Enterprise Domain. The Enterprise Domain is where data discovery, predictive analytics, and basic query and reporting needs are met.

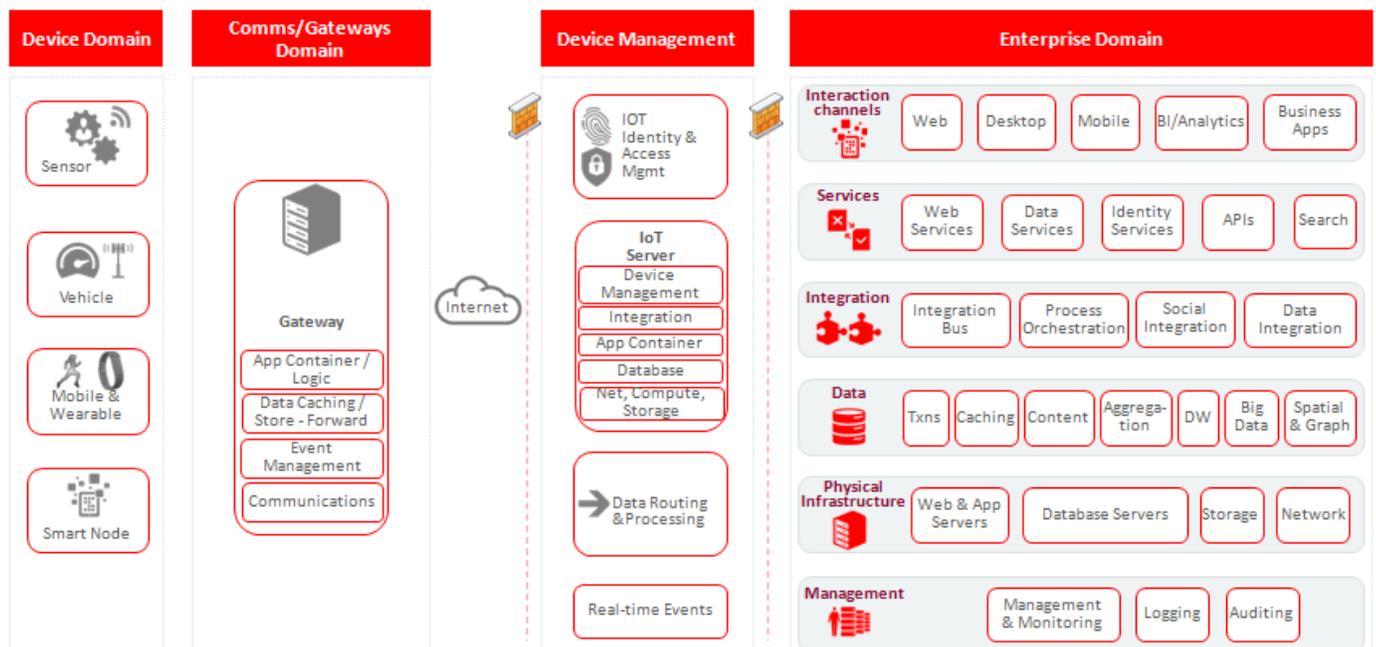


Figure 17: Connected Devices Capability Map (Internet of Things)

In some scenarios, immediate action must be taken when data is first transmitted (as when a sensor reports a critical problem that could damage equipment or cause injury) or where it would be possible alleviate some other preventable situation (such as relieving a highway traffic jam). Event processing engines are designed to take certain pre-programmed actions quickly by analyzing the data streams while data is still in motion or when data has landed in NoSQL database front-ends or Hadoop. The rules applied are usually based on analysis of previous similar data streams and known outcomes.

Some of the Oracle products that map to the Capability Map appear in the next figure. Many of the Big Data products from Oracle are described elsewhere in this paper, including both on premises and Cloud-based solutions.

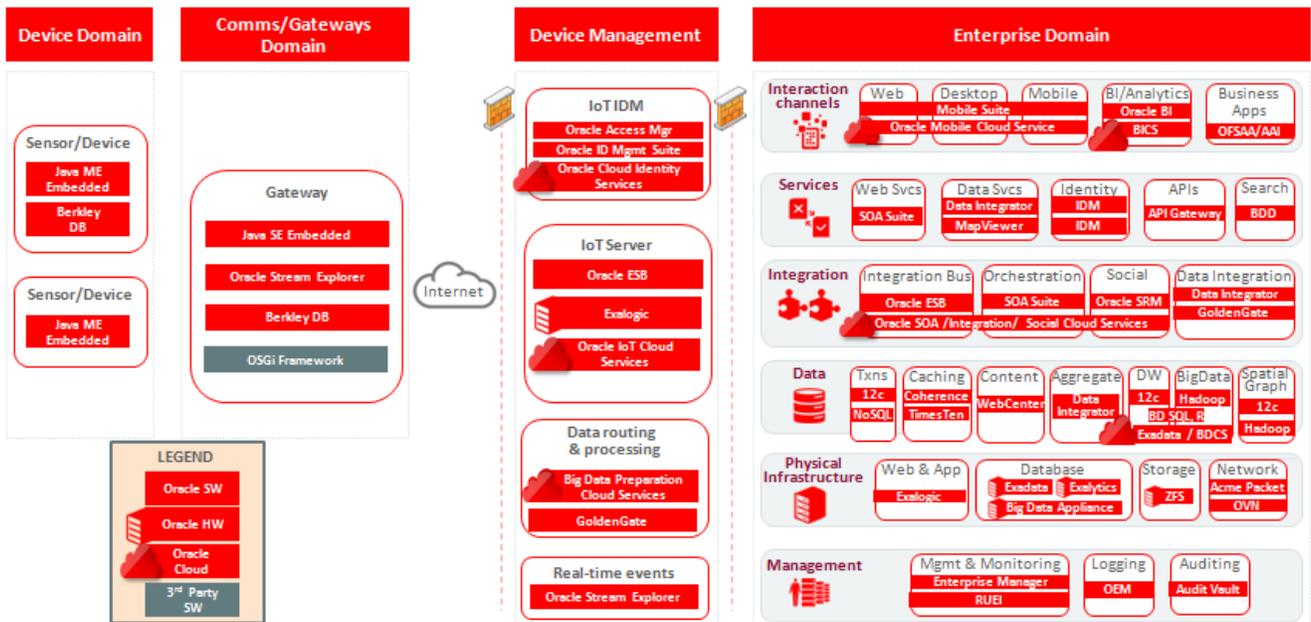


Figure 18: Oracle Products and the Capability Map (Internet of Things)

## Big Data Architecture Patterns in Three Use Cases

In this section, we will explore three use cases and walk through the architecture decisions and technology components:

- » Case 1: Retail web log analysis
- » Case 2: Financial Services real-time risk detection
- » Case 3: Driver insurability using telematics

### Use Case #1: Retail Web Log Analysis

In our first example, a leading retailer reported disappointing results from its web channels during the Christmas season. It is looking to improve customers' experience at the online shopping site. Analysts at the retailer will investigate the website navigation pattern, especially abandoned shopping carts.

The architecture challenge is to quickly implement a solution using mostly existing tools, skills, and infrastructure in order to minimize cost and to quickly deliver a solution to the business. The number of skilled Hadoop programmers on staff is very few but they do have SQL expertise. Loading all of the data into the existing Oracle data warehouse enabling the SQL programmers to access the data there is rejected because the data movement would be extensive and the processing power and storage required would not make economic sense. The 2nd option was to load the data into Hadoop and directly access the data in HDFS using SQL.

The conceptual architecture as shown in Figure 19 provides direct access to the Hadoop Distributed File System by simply associating it with an Oracle Database external table. Once connected, Oracle Big Data SQL enables traditional SQL tools to explore the data set.

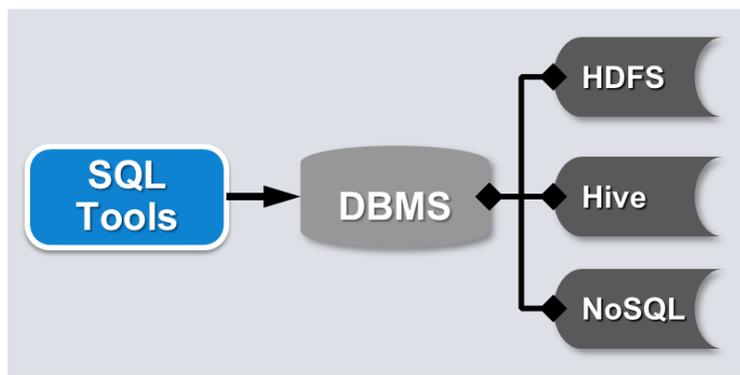


Figure 19: Use Case #1: Retail Web Log Analysis

The key benefits of this architecture include:

- » Low cost Hadoop storage
- » Ability to leverage existing investments and skills in Oracle SQL and BI tools
- » No client side software installation
- » Leverage Oracle data warehouse security
- » No data movement into the relational database
- » Fast ingestion and integration of structured and unstructured data sets



Key Oracle architectural components used to meet this challenge include:

- » 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 Business Intelligence Enterprise Suite 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.
- » Hadoop:
  - » Cloudera Hadoop Distribution deployed on Oracle's Big Data Appliance or in Oracle's Public Cloud as the Big Data Cloud Service or Apache Hadoop distribution (for example, on IaaS).
- » Oracle Big Data SQL:
  - » A SQL access method that provides advanced connectivity between the Oracle Big Data Appliance (data reservoir) and Oracle Exadata (data warehouse) or deployed in the Oracle Public Cloud as the Big Data SQL Cloud Service with the Big Data Cloud Service and Exadata Cloud Service.
  - » 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.

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: Financial Services Real-time Risk Detection

A large financial institution has regulatory obligations to detect potential financial crimes and terrorist activity. However, there are 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 data that is constantly streaming in to the institution. IT systems must 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 provide data in real-time, some provide data in batch mode.

The institution wants to use their existing business intelligence tools to meet regulatory reporting requirements. Because of a mix of real-time and batch data feeds, a streaming event processing engine must be part of the solution to evaluate the variety of data sources.

Figure 20 illustrates the proposed solution. It will enable analysis of 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 hierarchies. Previously, the volume and variety of data meant that it could not be used to its fullest extent due to constraints in processing power and the cost of storage required. With Hadoop, Spark, and/or Storm processing, we will 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 include various operational and analytical systems for appropriate reporting, analysis and action.

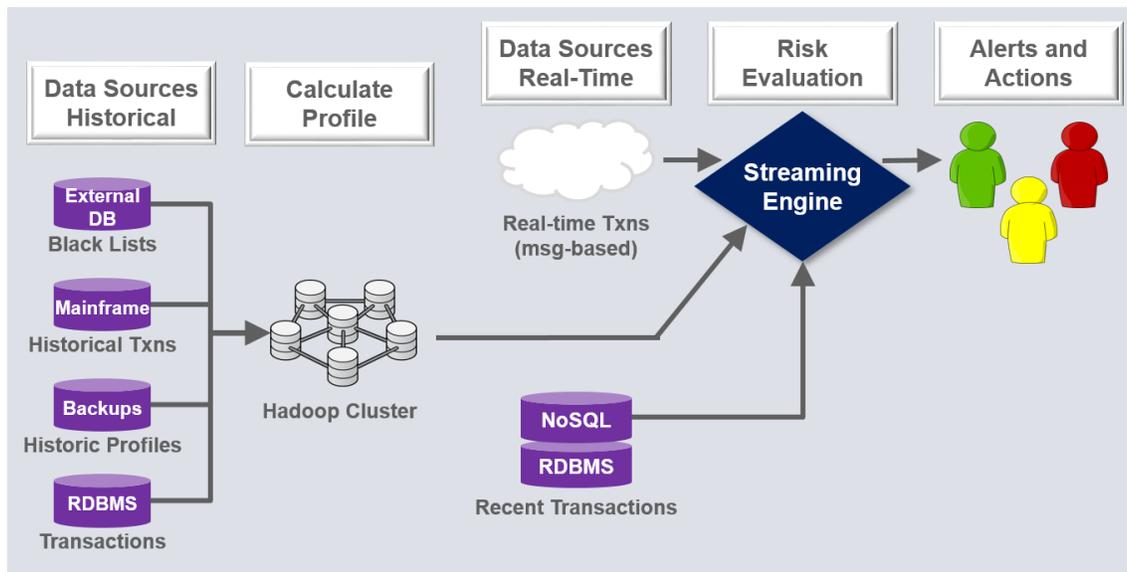


Figure 20: Use Case #2: Financial Services Real-time Risk Detection

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.

The following components are included in the architecture:

- » Stream / Event Processing
  - » Oracle Stream Explorer continuously processes incoming data, analyzes and evolves patterns, and raises events if conditions are detected. Stream Explorer runs in an Open Service Gateway (OSGi) container and can operate on any Java Runtime Environment. It provides a business level user interface allowing interpreting data streams without requiring knowledge of underlying event technology characteristics. It can be deployed on premises or in the Oracle Public Cloud (Internet of Things Cloud Service)
  - » Apache streaming options could also be deployed including [Spark Streaming](#), [Flume](#), and [Storm](#).
- » Hadoop:
  - » Cloudera Hadoop Distribution deployed on Oracle's Big Data Appliance or in Oracle's Public Cloud as the Big Data Cloud Service or Apache Hadoop distribution (for example, on Oracle IaaS).
  - » Spark or MapReduce processing of 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 R Advanced Analytics for Hadoop for data mining / statistical detection of fraud.
- » Oracle Big Data Appliance (or other Hadoop Solutions):
  - » Capture events (various options, such as Flume, Spark Streaming)

- » Oracle NoSQL Database to capture low latency data with flexible data structure and fast querying (deployed on Oracle's Big Data Appliance or in Oracle's Public Cloud as NoSQL Database as a Service or Apache or other NoSQL distribution (for example, on Oracle IaaS).

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: Driver Insurability using Telematics

The third use case is an insurance company seeking to personalize insurance coverage and premiums based on individual driving habits. The insurance company will capture a large amount of vehicle-created sensor data (e.g. telematics / Internet of Things) reflecting their customers' driving habits. They must 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.



Figure 21: Use Case #3: Auto Insurance Company Business Objectives

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

Figure 22 illustrates the new architecture. Internet of Things architectures rely on middleware components to gather data from sensors, manage the devices, and analyze streaming data. As in our previous example, streaming data might make its way first into NoSQL databases or directly into Hadoop. Analyzed data eventually makes its way into the pre-existing data warehouse.

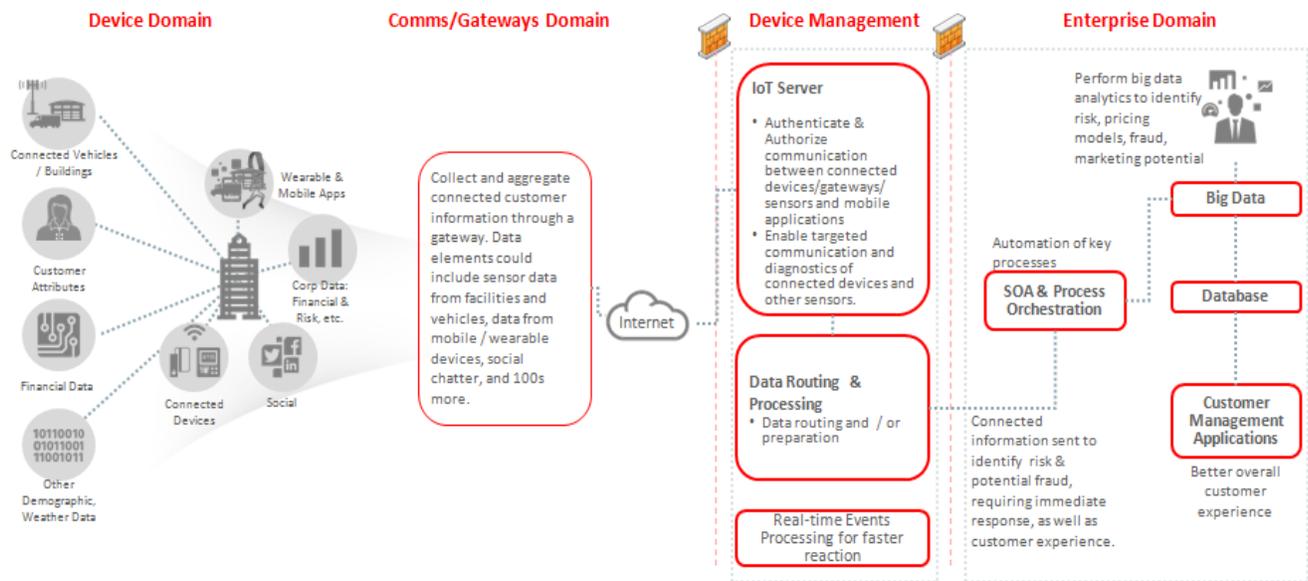


Figure 22: Use Case #3: Driver Insurability using Telematics (Internet of Things) Sensor Data

The solution can accomplish multiple goals. It can be used to update the customer profile, calculate new premiums, update the data warehouse, and contribute data to a discovery lab where profitability and competitiveness can be analyzed. The architecture is designed to minimize data movement across platforms, integrate business intelligence and analytic processes, enable deep analysis, and ensure access / identity management and data security capabilities are applied consistently.

Due to the volume and variety of sensor data, HDFS is chosen to store the raw data. Spark and 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 business intelligence and information discovery tools, some enabled by Big Data SQL, data is 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 data and our data warehouse. The solution pictured can be deployed on premises, on IaaS platforms, or in Oracle’s Public Cloud on PaaS platforms.



## Big Data Best Practices

Guidelines for building a successful big data architecture foundation:

### #1: Align Big Data with Specific Business Goals

A key intent of Big Data is to find new value from more extensive data sets - 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. Examples include understanding how to filter web logs to understand eCommerce behavior, deriving sentiment from social media and customer support interactions, and understanding 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 how Big Data supports and enables your top business and IT priorities.

### #2: Ease Skills Shortage with Standards and Governance

McKinsey Global Institute<sup>1</sup> 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's Big Data related Cloud Services 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,

---

<sup>1</sup> McKinsey Global Institute, May 2011, 'The challenge—and opportunity—of 'big data'', [https://www.mckinseyquarterly.com/The\\_challenge\\_and\\_opportunity\\_of\\_big\\_data\\_2806](https://www.mckinseyquarterly.com/The_challenge_and_opportunity_of_big_data_2806)



many see Big Data as an integral extension of your existing business intelligence and data warehousing platform and information architecture.

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 intelligence that can be gained from social media, sensors and other mobile devices that form the Internet of Things, is recognized by many organizations as key to gaining a competitive advantage and outperforming 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 businesses 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 will provide a fast, reliable path to business adoption.

As you explore the spectrum of Big Data capabilities, we suggest that you think about a platform but deliver projects impactful to the business. 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. As you do this, you'll be expected to align new operational and management capabilities with standard IT processes and capabilities, leverage prior investments, and build for enterprise scale and resilience.

Oracle has over 30 years of leadership in information management and continues to make 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. Oracle solutions can be delivered in on premises deployment models or in Oracle's Public Cloud.

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](http://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](http://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](http://www.oracle.com/bigdata).



CONNECT WITH US

 [blogs.oracle.com/enterprisearchitecture](https://blogs.oracle.com/enterprisearchitecture)

 [facebook.com/OracleEA](https://facebook.com/OracleEA)

 [twitter.com/oracleEAs](https://twitter.com/oracleEAs)

 [oracle.com/goto/EA](https://oracle.com/goto/EA)

**Oracle Corporation, World Headquarters**

500 Oracle Parkway  
Redwood Shores, CA 94065, USA

**Worldwide Inquiries**

Phone: +1.650.506.7000  
Fax: +1.650.506.7200

**Hardware and Software, Engineered to Work Together**

Copyright © 2016, Oracle and/or its affiliates. All rights reserved. This document is provided for information purposes only, and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document, and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.

Oracle and Java are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.

Intel and Intel Xeon are trademarks or registered trademarks of Intel Corporation. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. AMD, Opteron, the AMD logo, and the AMD Opteron logo are trademarks or registered trademarks of Advanced Micro Devices. UNIX is a registered trademark of The Open Group. 0316

March 2016

An Enterprise Architecture White Paper – An Enterprise Architect’s Guide to Big Data — Reference Architecture Overview

Author: Peter Heller, Dee Piziak, Robert Stackowiak, Art Licht, Tom Luckenbach, Bob Cauthen, Avishkar Misra, John Wyant, Jeff Knudsen



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