An Oracle White Paper
February 2013

Information Management and Big Data
A Reference Architecture

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.

Introduction

In the original Oracle white paper on Information Management Reference Architecture we described how “information” was at the heart of every successful, profitable and transparent business in the world – something that’s as true today as it was then. Information is the lifeblood of every organization and yet Information Management (IM) systems are too often viewed as a barrier to progress in the business rather than an enabler of it. At best IM is an unsung hero.

What has changed in the last few years is the emergence of “Big Data”, both as a means of managing the vast volumes of unstructured and semi-structured data stored but not exploited in many organizations, as well as the potential to tap into new sources of insight such as social-media web sites to gain a market edge.

It stands to reason that within the commercial sector Big Data has been adopted more rapidly in data driven industries, such as financial services and telecommunications. These organizations have experienced a more rapid growth in data volumes compared to other market sectors, in addition to tighter regulatory requirements and falling profitability.

Many organizations may have initially seen Big Data technologies as a means to ‘manage down’ the cost of large scale data management or reduce the costs of complying with new regulatory requirements. This has changed as more forward-looking companies have understand the value creation potential when combined with their broader Information Management architecture for decision making, and applications architecture for execution. There is a pressing need for organizations to align analytical and execution capabilities with ‘Big Data’ in order to fully benefit from the additional insight that can be gained.

Received wisdom suggests that more than 80% of current IT budgets is consumed just keeping the lights on rather than enabling business to innovate or differentiate themselves in the market. Economic realities are squeezing budgets still further, making IT’s ability to change this spending mix an even more difficult task. For organizations looking to add some element of Big Data to their IT portfolio, they will need to do so in a way that complements existing solutions and does not add to the cost burden in years to come. An architectural approach is clearly what is required.

In this white paper we explore Big Data within the context of Oracle’s Information Management Reference Architecture. We discuss some of the background behind Big Data and review how the Reference Architecture can help to integrate structured, semi-structured and unstructured information into a single logical information resource that can be exploited for commercial gain.

Background

In this section, we will review some Information Management background and look at the new demands that are increasingly being placed on Data Warehouse and Business Intelligence solutions by businesses across all industry sectors as they look to exploit new data sources (such as social media) for commercial advantage. We begin by looking through a Business Architecture lens to give some context to subsequent sections of this white paper.

Information Management Landscape

There are many definitions of Information Management. For the purposes of this white paper we will use a broad definition that highlights the full lifecycle of the data, has a focus on the creation of value from the data and somewhat inevitably includes aspects of people, process and technology within it.

While existing IM solutions have focused efforts on the data that is readily structured and thereby easily analysed using standard (commodity) tools, our definition is deliberately more inclusive. In the past the scope of data was typically mediated by technical and commercial limitations, as the cost and complexities of dealing with other forms of data often outweighed any benefit accrued. With the advent of new technologies such as Hadoop and NoSQL as well as advances in technologies such as Oracle Exadata, many of these limitations have been removed, or at the very least, the barriers have been expanded to include a wider range of data types and volumes.

What we mean by Information Management:

Information Management (IM) is the means by which an organisation seeks to maximise the efficiency with which it plans, collects, organises, uses, controls, stores, disseminates, and disposes of its Information, and through which it ensures that the value of that information is identified and exploited to the maximum extent possible.

As an example, one of our telecommunications customers has recently demonstrated how they can now load more than 65 billion call data records per day into an existing 300 billion row relational table using an Oracle database. While this test was focused very squarely at achieving maximum throughput, the key point is that dealing with millions or even billions of rows of data is now much more common place, and if organised into the appropriate framework, tangible business value can be delivered from previously unimaginable quantities of data. That is the *raison d'être* for Oracle's IM Reference Architecture.

Although newer hardware and software technologies are changing what is possible to deliver from an IM perspective, in our experience the overall architecture and organising principles are more critical. A failure to organise data effectively results in significantly higher overall costs and the growth of a 'shadow IT' function within the business i.e. something that fills the gap between IT delivery capabilities and business needs. In fact, as part of a current state analysis we often try to measure the size of the 'shadow IT' function in our customers as a way of quantifying IM issues. How many people and how much time is spent preparing data rather than analysing it? How has the 'Shadow-IT' function influenced tools choices and the way in which IM is delivered? 'Shadow IT' can impose a significant additional burden in costs, time and tools when developing a transitional roadmap.

In many instances, we find existing IM solutions have failed to keep pace with growing data volumes and new analysis requirements. From an IT perspective this results in significant cost and effort in tactical database tuning and data reorganization just to keep up with ever changing business processes. Increasing data volumes also put pressure on batch windows. This is often cited by IT teams as the most critical issue, leading to additional costly physical data structures being built such as Operational Data Stores and Data Caches so a more real-time view of data can be presented. These structures really just serve to add cost and complexity to IM delivery. The real way to tackle the batch load window is not to have one.

Data in an IM solution tends to have a natural flow rate determined either by some technological feature or by business cycles (e.g. Network mediation in a mobile network may generate a file every 10 minutes or 10,000 rows, whichever is sooner, where as a business may re-forecast sales every 3 months).

By trickle feeding data at this underlying flow rate into the staging data layer, batch issues can be eliminated and the IM estate rationalised. We argue that by adopting Oracle's IM Reference Architecture you will be able to support rapid collaborative development, incorporating new data and new analysis areas, and thus keep pace with business change while dramatically reducing the size of 'shadow-IT' in your organization.

Extending the Boundaries of Information Management

There is currently considerable hype in the press regarding Big Data. Articles often feature companies concerned directly with social media in some fashion, making it very difficult to generalise about how your organization may benefit from leveraging similar tools, technology or data. Many of these social media companies are also very new, so questions about how to align Big Data technologies to the accumulated complexity of an existing IM estate are rarely addressed.

Big Data is no different from any other aspect of Information Management when it comes to adding value to a business. There are two key aspects to consider:

- How the new data or analysis scope can enhance your existing set of capabilities?
- What additional opportunities for intervention or processes optimisation does it present?

Figure 1 shows a simplified functional model for the kind of 'analyse, test, learn and optimise' process that is so key to leveraging value from data. The steps show how data is first brought together before being analysed and new propositions of some sort are developed and tested in the data. These propositions are then delivered through the appropriate mechanism and the outcome measured to ensure the consequence is a positive one.

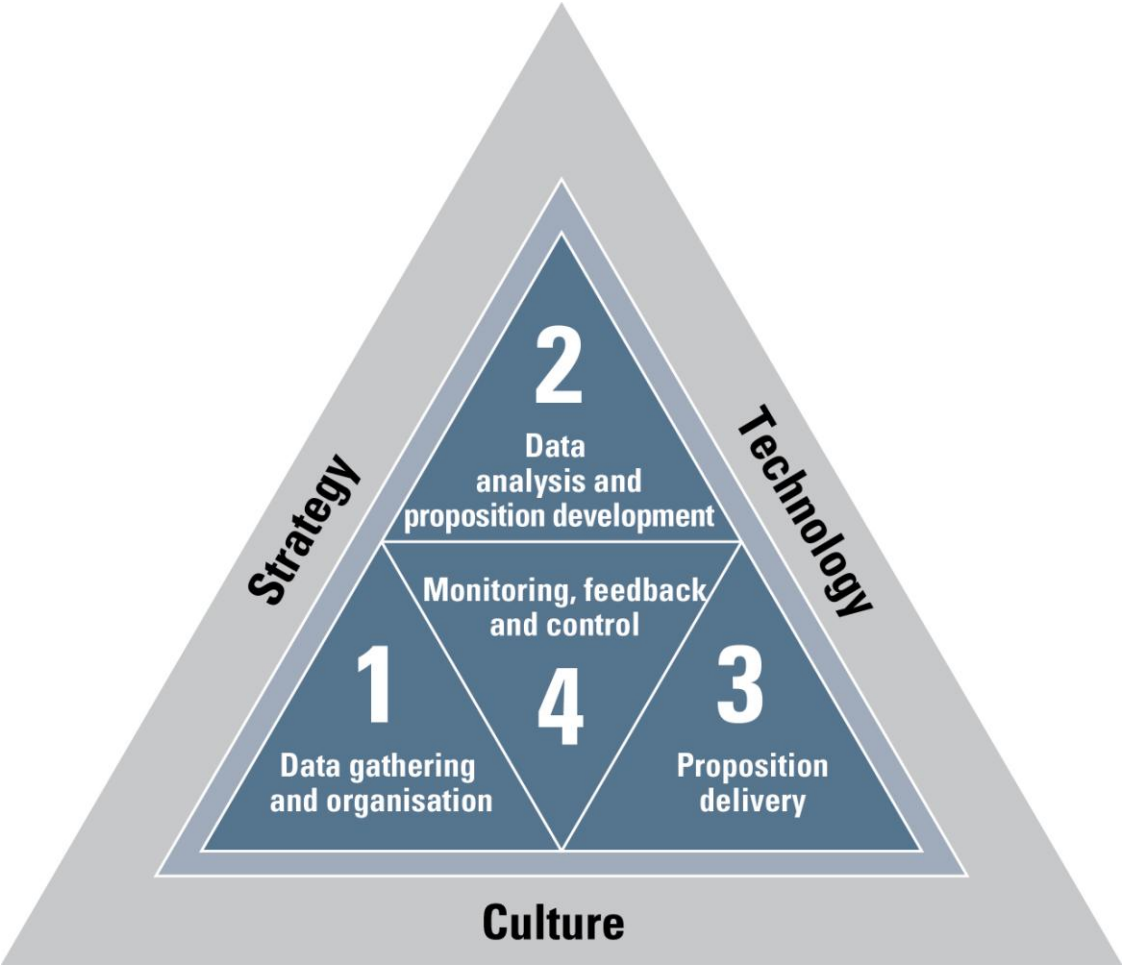
2
Data analysis and proposition development
Monitoring, feedback and control
4
1
Data gathering and organisation
3
Proposition delivery
Strategy
Technology
Culture

Figure 1. Simple functional model for data analysis

The model also shows how the operational scope is bounded by the three key dimensions of Strategy, Technology and Culture. To maximise potential, these three dimensions should be in balance. There is little point in defining a business strategy that cannot be supported by your organizations IT capacity or your employees ability to deliver IT.

Big Data Opportunity in Customer Experience Management

A common use-case for Big Data revolves around multi-channel Customer Experience Management (CX). By analysing the data flowing from social media sources we might understand customer sentiment and adapt service delivery across our channels accordingly to offer the best possible customer experience.

If we animate our simplified functional model (Figure 1) in a CX context we see the first task is to bring the data together from disparate sources in order to align it for analysis. We would normally do this in a Data Warehouse using the usual range of ETL tools. Next, we analyse the data to look for meaningful patterns that can be exploited through new customer propositions such as a promotion or special offer. Depending on the complexity of the data, this task may be performed by a Business Analyst using a BI toolset or a Data Scientist with a broader range of tools, perhaps both. Having defined a new proposition, appropriate customer interventions can be designed and then executed through channels (inbound/outbound) using OLTP applications. Finally, we monitor progress against targets with BI, using dashboards, reports and exception management tools.

Many modern 'next best offer' recommendations engines will automate each of the steps shown in our functional model and are integrated into OLTP applications that are responsible for the final offer delivery.

It's also interesting to note how the functional model shown in figure 1 maps against the different types of analysis and BI consumers shown in figure 2. In many organizations it falls to the Business Analyst to perform the required 'Data Analysis and Proposition Development' function using a standard BI toolset, rather than a Data Scientist using a more specialised suite of tools applied in a more agile fashion. It seems reasonable to suggest that the latter will be more successful in unlocking the full potential value of the data.

Another important point to make regarding this mapping is the need for the 'Monitoring, Feedback and Control' feedback loop which must link back at the Executive level through Enterprise Performance Management (EPM) to ensure that strategy is informed and adjusted based on operational realities.

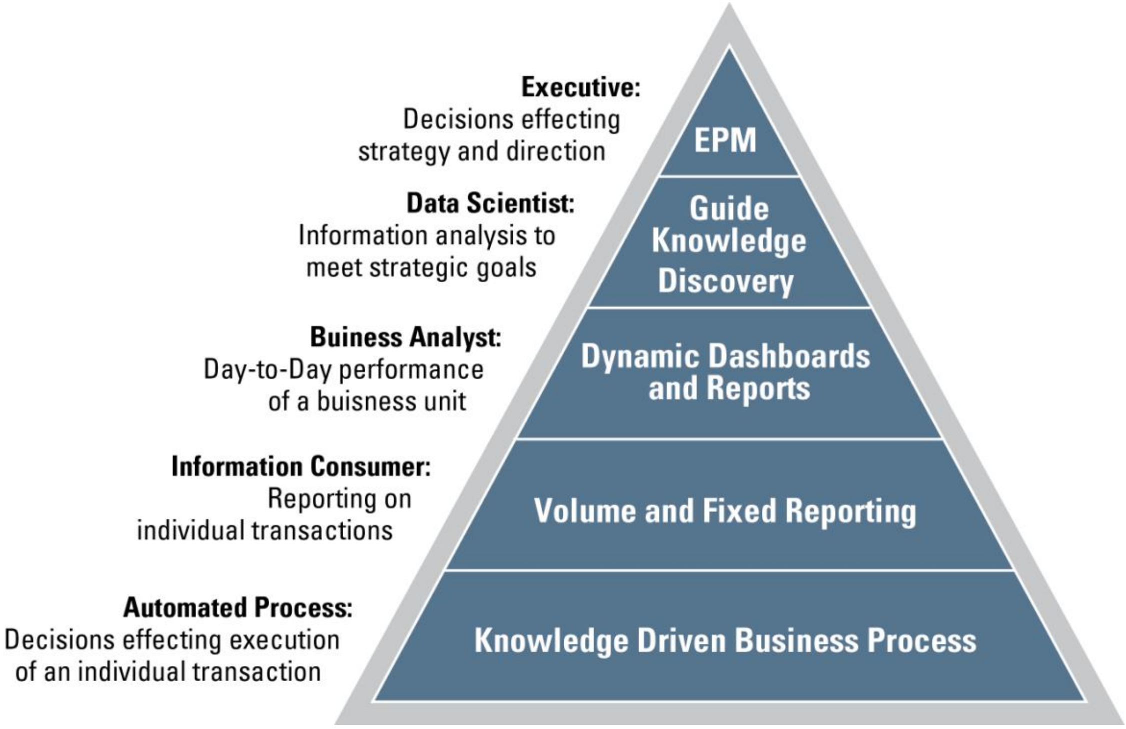
Executive:
Decisions effecting strategy and direction
EPM
Data Scientist:
Information analysis to meet strategic goals
Guide Knowledge Discovery
Buiness Analyst:
Day-to-Day performance of a buisness unit
Dynamic Dashboards and Reports
Information Consumer:
Reporting on individual transactions
Volume and Fixed Reporting
Automated Process:
Decisions effecting execution of an individual transaction
Knowledge Driven Business Process

Figure 2. Information consumers and types of analysis

To be successful in leveraging Big Data, organizations must do more than simply incorporate new sources of data if they are to capture its full potential. They must also look to extend the scope of their CRM Strategy and organizational culture as well as fit newer Big Data capabilities into their broader IM architecture. This point is shown conceptually in Figure 3 below. For example, telecoms companies who may have previously run a set number of fixed campaigns against defined target segments may now be able to interact with customers on a real-time basis using the customer's location as a trigger. But how should promotions be designed in order to be affordable and effective in this new world? How can we avoid fatiguing customers through the increased number of interventions?

What new reports must be created to track progress? How can these new opportunities for interaction and the data coming back from channels be used in other areas of customer management such as Brand Management, Price Management, Product and Offering Design, Acquisition and Retention Management, Complaint Management, Opportunity Management and Loyalty Management? These are all important questions that need to be answered, preferably before the customer has moved to a competitor or ticked the 'do not contact' box because they're fed up with being plagued by marketing offers.

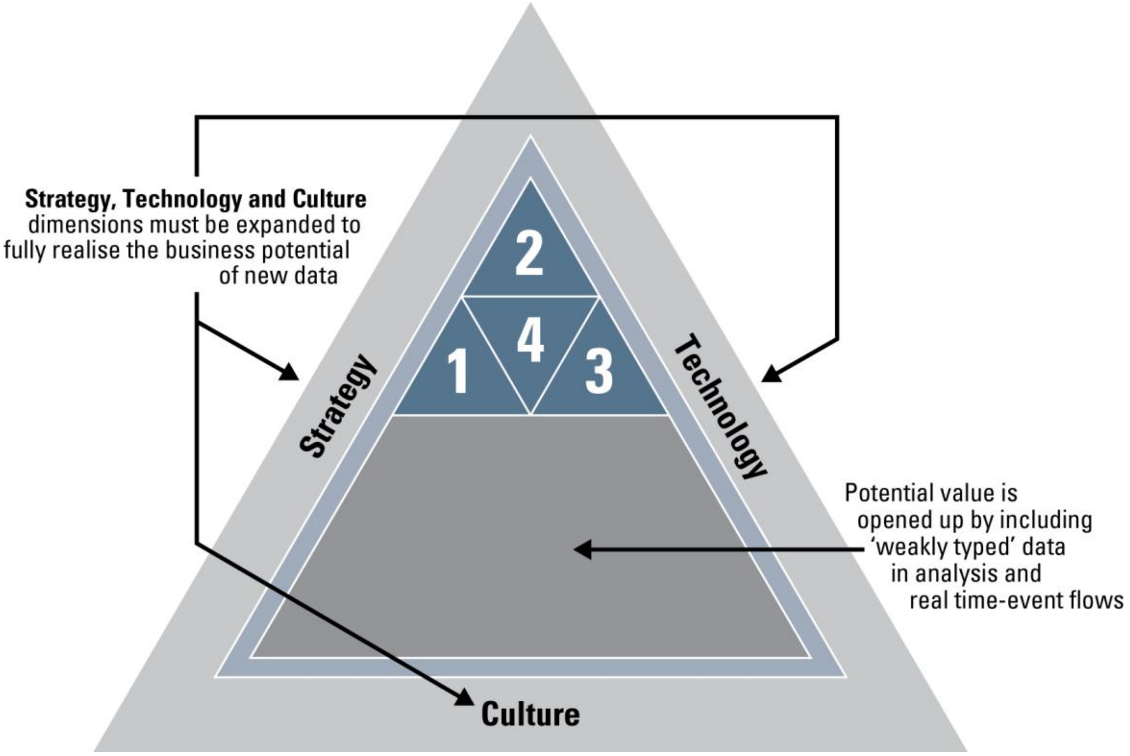
Strategy, Technology and Culture dimensions must be expanded to fully realise the business potential of new data
2
1
4
3
Strategy
Technology
Potential value is opened up by including 'weakly typed' data in analysis and real time-event flows
Culture

Figure 3. Conceptual expansion of functional model to include Big Data

It's also worth noting from Figure 1 that the data analysis and proposition development is separated from the proposition delivery (i.e. channel execution). While that seems self evident when represented in this fashion, we find that many people conflate the two functions when talking about technologies such as Data Mining. We will discuss this point again when looking at the role of Data Scientists, but we can see how the development of a Data Mining model for a problem such as target marketing is separate from the scoring of data to create a new list of prospects. These separate activities map well to the proposition analysis and proposition delivery tasks shown in our diagram Figure 1.

We would note that CX is just one example of a domain where Big Data and Information Management (more generally) can add value to an organization. You can see from our original definition that IM is all about data exploitation and applies equally to every other business domain.

Information Management Reference Architecture Basics

Oracle's Information Management Reference Architecture describes the organising principles that enable organizations to deliver an agile information platform that balances the demands of rigorous data management and information access. See the end of this white paper for references and further reading.

The main components of Oracle's IM Reference Architecture are shown in Figure 4 below.

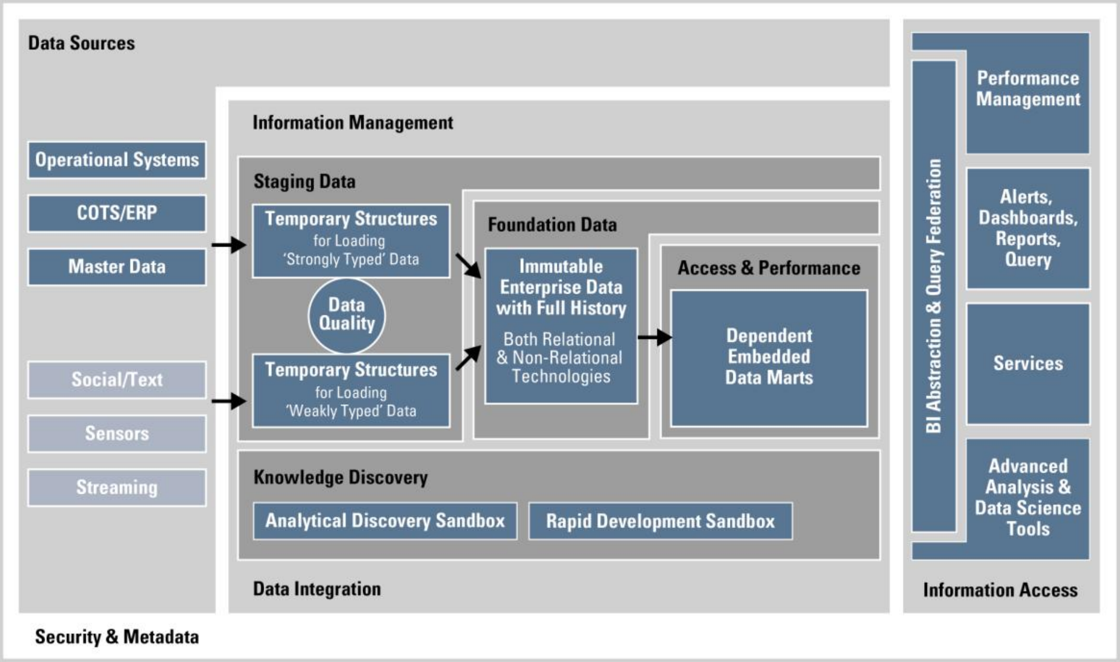
Data Sources
Information Management
Operational Systems
COTS/ERP
Master Data
Social/Text
Sensors
Streaming
Staging Data
Temporary Structures for Loading 'Strongly Typed' Data
Data Quality
Temporary Structures for Loading 'Weakly Typed' Data
Foundation Data
Immutable Enterprise Data with Full History
Both Relational & Non-Relational Technologies
Access & Performance
Dependent Embedded Data Marts
Knowledge Discovery
Analytical Discovery Sandbox
Rapid Development Sandbox
Data Integration
BI Abstraction & Query Federation
Performance Management
Alerts, Dashboards, Reports, Query
Services
Advanced Analysis & Data Science Tools
Information Access
Security & Metadata

Figure 4. Main components of the IM Reference Architecture

It's a classically abstracted architecture with the purpose of each layer clearly defined. In brief these are:

- Staging Data Layer. Abstracts the rate at which data is received onto the platform from the rate at which it is prepared and then made available to the general community. It facilitates a 'right-time' flow of information through the system.
- Foundation Data Layer. Abstracts the atomic data from the business process. For relational technologies the data is represented in close to third normal form and in a business process neutral fashion to make it resilient to change over time. For non-relational data this layer contains the original pool of invariant data.
- Access and Performance Layer. Facilitates access and navigation of the data, allowing for the current business view to be represented in the data. For relational technologies data may be logical or physically structured in simple relational, longitudinal, dimensional or OLAP forms. For non-relational data this layer contains one or more pools of data, optimised for a specific analytical task

or the output from an analytical process. e.g., In Hadoop it may contain the data resulting from a series of Map-Reduce jobs which will be consumed by a further analysis process.

- Knowledge Discovery Layer. Facilitates the addition of new reporting areas through agile development approaches and data exploration (strongly and weakly typed data) through advanced analysis and Data Science tools (e.g. Data Mining).
- BI Abstraction & Query Federation. Abstracts the logical business definition from the location of the data, presenting the logical view of the data to the consumers of BI. This abstraction facilitates Rapid Application Development (RAD), migration to the target architecture and the provision of a single reporting layer from multiple federated sources.

One of the key advantages often cited for the Big Data approach is the flexibility of the data model (or lack thereof) over and above a more traditional approach where the relational data model is seen to be brittle in the face of rapidly changing business requirements. By storing data in a business process neutral fashion and incorporating an Access and Performance Layer and Knowledge Discovery Layer into the design to quickly adapt to new requirements we avoid the issue. A well designed Data Warehouse should not require the data model to be changed to keep in step with the business and provides for rich, broad and deep analysis.

Over the years we have found that the role of sandboxes has taken on additional significance. In this (slight) revision of the model we have placed greater emphasis on sandboxes by placing into a specific Knowledge Discovery Layer where they have a role in iterative (BI related) development approaches, new knowledge discovery (e.g. Data Mining), and Big Data related discovery. These three areas are described in more detail in the following sections.

Knowledge Discovery Layer and the Data Scientist

What's the point in having useful data if you can't make it useful? The role of the Data Scientist is to do just that, using scientific methods to solve business problems using available data.

We begin this section by looking at Data Mining in particular. While Data Mining is only one approach a Data Scientist may use in order to solve data related issues, the standardised approach often applied is informative and, at a high level at least, can be generally applied to other forms of knowledge discovery.

Data Mining can be defined as the automatic or semiautomatic task of extracting previously unknown information from a large quantity of data. In some circles, especially more academic ones, it is still referred to as Knowledge Discovery in Large Datasets (KDD) which you might consider to be the forebear of Data Mining.

The Cross Industry Standard Process Model for Data Mining (CRISP-DM)© outlines one of the most common frameworks used for Data Mining projects in industries today. Figure 5 illustrates the main CRISP-DM phases. At a high level at least, CRISP-DM is an excellent framework for any knowledge discovery process and so applies equally well to a more general Big Data oriented problem or to a specific Data Mining one.

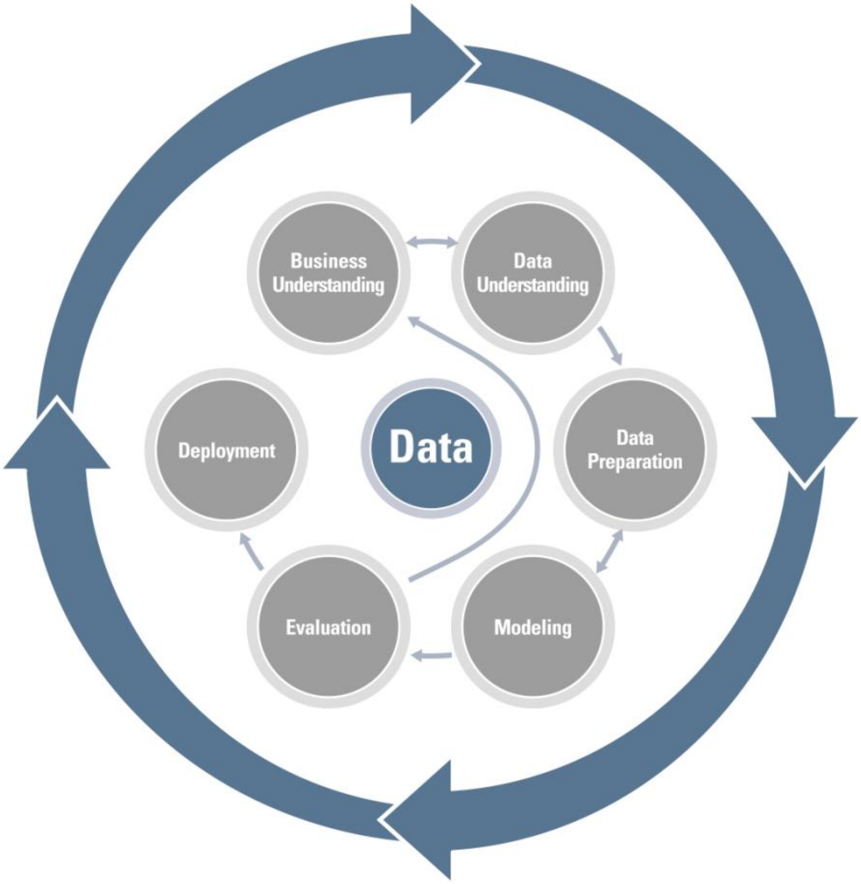

Figure 5. High level CRISP-DM process model

Figure 5 also shows how for any given task, an Analyst will first build both a business and data understanding in order to develop a testable hypothesis. In subsequent steps the data is then prepared, models built and then evaluated (both technically and commercially) before deploying either the results or the model in some fashion. Implicit to the process is the need for the Analyst to take factors such as the overall business context and that of the deployment into account when building and testing the model to ensure it is robust. For example, the Analyst must not use any variables in the model that will not be available at the point (channel and time) when the model will be used for scoring new data.

During the Data Preparation and Modelling steps it is typical for the Data Mining Analyst or Data Scientist to use a broad range of statistical or graphical representations in order to gain an understanding of the data in scope and may drive a series of additional transformations in order to emphasise some aspect of the data to improve the model's efficacy. Figure 5 shows the iterative nature of these steps and the tools required to do them well typically fall outside the normal range of BI tools organizations have standardised upon.

This highly iterative process can be supported within the Knowledge Discovery Layer using either a relational or Hadoop-based approach. For the sake of simplicity in this section, we will describe the relational approach. Please see the later sections on Big Data for a guide to a Hadoop-based approach.

When addressing a new business problem, an Analyst will be provisioned for a new project based sandbox. The Analyst will identify data of interest from the Access and Performance Layer or (less frequently) from the Foundation Data Layer as a starting point. The data may be a logical (view) rather than physical copy and may be sampled if the complete dataset is not required.

The Analyst may use any number of tools to present the data in meaningful ways to progress understanding. Typically, this might include Data Profiling and Data Quality tools for new or unexplored datasets, statistical and graphical tools for a more detailed assessment of attributes and newer contextual search applications that provide an agile mechanism to explore data without first having to solidify it into a model or define conformed dimensions.

A wide range of mining techniques may be applied to the data depending on the problem being tackled. Each of the steps in our process may create new data such as data selections, transformations, models or test results which are all managed within the Sandbox.

The actual form the knowledge takes will depend on the original business problem and the technique(s) adopted. For a target classification model it may be a simple list showing each customer's purchase propensity, whereas for a customer segmentation problem the result may be a cluster number used to identify customers with similar traits that can be leveraged for marketing purposes. In both cases the results of analysis may be written out as a list and consumed by MDM or operational systems (typically CRM in these cases) or the model itself deployed so results can be generated in real time by applications.

The output may not always be a physical list or a deployed model - It may be that the Analyst simply finds some interesting phenomena in the data, perhaps as a by-product of an analysis. In this case the only output may be an email or a phone call to share this new knowledge.

Knowledge Discovery Layer and Right to Left Development

Sandboxes also provide useful support for the rapid development of new reporting areas. Let's imagine that an important stakeholder has some new data available in a file that they want to combine with existing data and reports. This example is illustrated in Figure 6.

The goal is to deliver the new reports as quickly and simply as possible. However, it may be impossible to schedule the ETL team or get the work through formalised production control in the time available. We have also found that the majority of users understand and respond better to physical prototypes than they do relational data modelling and formalised report specifications.

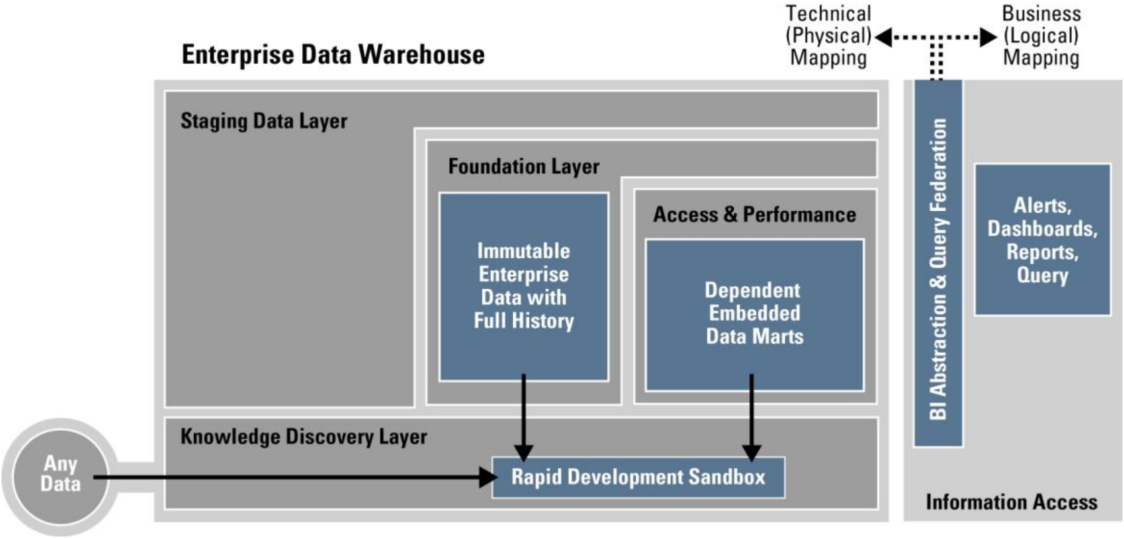
Technical
(Physical)
Mapping
Business
(Logical)
Mapping
Enterprise Data Warehouse
Staging Data Layer
Foundation Layer
Immutable Enterprise Data with Full History
Access & Performance
Dependent Embedded Data Marts
BI Abstraction & Query Federation
Alerts, Dashboards, Reports, Query
Any Data
Knowledge Discovery Layer
Rapid Development Sandbox
Information Access

Figure 6. Support for right-to-left Rapid Application Development

The starting point for a Rapid Application Development (RAD) approach is the provisioning of a new sandbox for the project. New data can then be identified and replicated into the sandbox and combined with data that already exists in any of the other layers of our model (not typically from the Staging Layer but this is also possible). The Business Analyst can then quickly make the new data available for reporting by mapping it (logically and physically) in the BI Abstraction Layer and then rapidly prototype the look and feel of the report until the stakeholder is satisfied with the results. Once functional prototyping is completed, the work to complete non-functional components and professionally manage the data through the formal layers of our architecture and put the data into a production setting must be completed. During this time the user can (optionally) continue to use the report in the sandbox until the work to professionally manage the data within the main framework is completed – switchover is simply a case of changing the physical mapping from the sandbox to the new location of the data in the Access and Performance Layer.

What is Big Data?

So what do we mean by Big Data and how can it create added insight or revenue generating opportunities?

Big Data is a term applied to data sets whose size is beyond the capability of commonly used software tools to capture, manage, and process. The sheer size of the data, combined with complexity of analysis and commercial imperative to create value from it, has led to a new class of technologies and tools to tackle it. The term Big Data tends to be used in multiple ways, often referring to both the type of data being managed as well as the technology used to store and process it. In the most part these technologies originated from companies such as Google, Amazon, Facebook and Linked-In, where they were developed for each company's own use in order to analyse the massive amounts of social media data they were dealing with. Due to the nature of these companies, the emphasis was on low cost scale-out commodity hardware and open source software.

The world of Big Data is increasingly being defined by the 4 Vs. i.e. these 'Vs' become a reasonable test as to whether a Big Data approach is the right one to adopt for a new area of analysis. The Vs are:

- Volume. The size of the data. With technology it's often very limiting to talk about data volume in any absolute sense. As technology marches forward, numbers get quickly outdated so it's better to think about volume in a relative sense instead. If the volume of data you're looking at is an order of magnitude or larger than anything previously encountered in your industry, then you're probably dealing with Big Data. For some companies this might be 10's of terabytes, for others it may be 10's of petabytes.
- Velocity. The rate at which data is being received and has to be acted upon is becoming much more real-time. While it is unlikely that any real analysis will need to be completed in the same time period, delays in execution will inevitably limit the effectiveness of campaigns, limit interventions or lead to sub-optimal processes. For example, some kind of discount offer to a customer based on their location is less likely to be successful if they have already walked some distance past the store.
- Variety. There are two aspects of variety to consider: syntax and semantics. In the past these have determined the extent to which data could be reliably structured into a relational database and content exposed for analysis. While modern ETL tools are very capable of dealing with data arriving in virtually any syntax they are less able to deal with semantically rich data such as free text. Because of this most organizations have restricted the data coverage of IM systems to a narrow range of data. It follows then that by being more inclusive, additional value may be created by an organization and this is perhaps one of the major appeals of the Big Data approach.

- Value. We need to consider what commercial value any new sources and forms of data can add to the business. Or, perhaps more appropriately, to what extent can the commercial value of the data be predicted ahead of time so that ROI can be calculated and project budget acquired. 'Value' offers a particular challenge to IT in the current harsh economic climate. It is difficult to attract funds without certainty of the ROI and payback period. The tractability of the problem is closely related to this issue as problems that are inherently more difficult to solve will carry greater risk, making project funding more uncertain.

  Later in this white paper, we will outline two distinctly different approaches to Big Data projects – to a large extent the approach you decide to adopt may be contingent on the level of certainty to which the value of the data can be predicted ahead of time and agreed upon by key stakeholders in your business.

To make data understandable it must be placed into a schema of some sort prior to analysis. One aspect that most clearly distinguishes Big Data from the relational approach is the point at which data is organized into a schema. In the relational approach we place data into a schema when it is initially written to the database, where as in a Big Data approach data is only organized into a schema immediately prior to analysis as it is read from disk. Thus Big Data can be said to be *'schema on read'* where as relational technology is *'schema on write'*.

Big Data is often seen as being more agile approach because in *'schema on read'* data is only structured immediately prior to analysis, but the approach also has hidden costs and risks which must be managed. For example, with *'schema on read'*, data quality is very dependent on the developer responsible for de-serialising / tokenizing the data from disk and this cost is potentially repeated for each program. It may also be difficult to find developers who are sufficiently knowledgeable about data streams written many years ago.

One area where there is almost no difference between *'schema on read'* and *'schema on write'* approaches is in the analysis technologies applied. Data Scientists will typically use a broad range of technologies such as Data Mining, Statistical and graphical analysis depending on the problem being tackled. It seems inevitable that in the future analysis tools will most likely work seamlessly across technologies, obfuscating the underlying storage technology.

In order to avoid any potential confusion between the physical representation of the data at rest or in-flight (syntax) and its inherent meaning (semantics), for the remainder of this white paper we will use the terms 'strongly typed' and 'weakly typed' data rather than overloaded terms such as structured, unstructured or semi-structured. By strongly typed, we mean data that is tabular in nature and can be readily placed into a set of relational tables, and weakly typed data refers to data where this is not the case. This will help us to distinguish between the way data is encoded (e.g. XML) from the contents of the data (e.g. free text versus well structured address data).

Big Data Technologies

There are endless articles, books and periodicals that describe Big Data from a technology perspective so we will instead focus our efforts here on setting out some basic principles and the minimum technology foundation to help relate Big Data to the broader IM domain.

The dominant Big Data technologies in use today commercially are Apache's Hadoop and No-SQL databases.

No-SQL databases are typically part of the real-time event detection process deployed to inbound channels (discussed in more detail in the section: "Big Data needs Big-Execution and Agile IM") but can also be seen as an enabling technology behind analytical capabilities such as contextual search applications. These are only made possible because of the flexible nature of the No-SQL model where the dimensionality of a query is emergent from the data in scope, rather than being fixed by the developer in advance. For the Data Scientist and Business Analysts in particular, this more agile approach can often lead to earlier insights into the data that may otherwise be obscured or constrained by a more formal development process.

Hadoop is a software framework for data intensive distributed applications and was developed from a number of academic papers published by Google who were researching in the area of parallel processing.

Hadoop has two major components:

- The Hadoop File System (HDFS). A highly scalable and portable file system for storing the data
- Map-Reduce. A programming model for processing the data in parallel.

The Map-Reduce framework allows analysis to be brought to the data in a distributed and highly scalable fashion, and the Hadoop ecosystem includes a wide range of tools to simplify analysis or manage data more generally. These tools create Map-Reduce programs which are then executed on top of HDFS. Analysis tools of note include:

- Apache Hive which provides a simple SQL-like interface
- Apache Pig which is a high level scripting language
- Apache Mahout for Data Mining.

Hadoop is designed for large volumes of data and is batch oriented in nature – even a simple query may take minutes to come back. In a typical Big Data oriented analysis scenario a Data Scientist may start by selecting a much smaller set of data and transforming it in some fashion and then combining this with the relational data from the Data Warehouse for analysis in a range of tools. Big Data analysis is also typically very explorative and iterative in nature so significantly more freedom is required than may traditionally be the case in Information Management. This is discussed in more detail in subsequent sections in this white paper.

While Hadoop offers native capabilities in the form of the Map-Reduce framework to analyse data as well as a wide range of analysis tools, Hadoop is more typically used as a preparatory step within an analysis process. Hadoop's low cost data storage model lends itself to providing a broad pool of data, each item of which may be of limited value to the organization, but for any given business problem may complete a missing link. Data may be selected, transformed and enhanced before it is moved to a relational setting and combined with additional corporate data where a more interactive analysis can be performed.

Given Hadoop is (currently at least) batch oriented other technologies are required in order to support real-time interactions. The most common technologies currently in use within this area are Complex Event Processing (CEP), In-Memory Distributed Data Grids, In-Memory Databases and traditional relational databases. These may be supported by other related technologies such as No-SQL databases, either sitting on top of a Hadoop cluster or using a specific data storage layer.

We will be discussing many of the topics outlined in this section such as inbound execution and knowledge discovery and the role of Big Data in more detail later on in this paper.

Big Data and the IM Reference Architecture

From the Big Data projects we have been involved in so far with some of our largest customers, two distinct patterns have emerged in terms of the approaches adopted. We have termed these 'Knowledge Stripping' and 'Knowledge Pooling' and to a large extent they describe the level of confidence and commitment surrounding the project. They differ in the extent to which organizations are willing to invest to pre-build both capability and capacity in order to more quickly discover and exploit information by making it available for analysis.

We describe Knowledge Stripping as a conservative approach to development where the minimum is undertaken in order to prove the underlying commercial value of the data. Knowledge Pooling is the antithesis of this approach, adopting a 'build it and they will come' approach.

Both Knowledge Stripping and Knowledge Pooling approaches have their merits and will appeal differently based on organizational context and maturity. They are described in more detail below.

Knowledge Stripping – Find the ROI Approach

In the 'Knowledge Stripping' approach companies are really looking to address the business problem as quickly and simply as possible. Having determined there is some commercial advantage they would move into a second phase of deploying the results of analysis within the confines of their existing IM framework.

It would be fair to say that companies adopting this approach want to see the ROI as a first step because they remain unconvinced by the arguments put forward regarding the possible value that can be derived from their data and are concerned about the hidden costs of Big Data technologies.

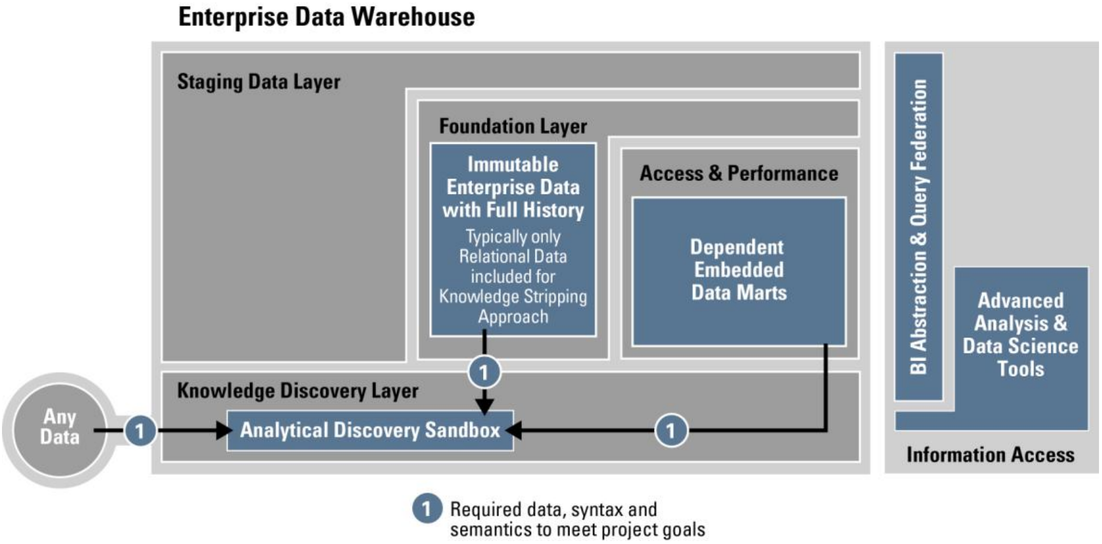
Enterprise Data Warehouse
Staging Data Layer
Foundation Layer
Immutable Enterprise Data with Full History
Typically only Relational Data included for Knowledge Stripping Approach
Access & Performance
Dependent Embedded Data Marts
BI Abstraction & Query Federation
Advanced Analysis & Data Science Tools
Information Access
Knowledge Discovery Layer
Analytical Discovery Sandbox
Any Data
1
1
1
1 Required data, syntax and semantics to meet project goals

Figure 7. Knowledge discovery for 'Knowledge Stripping' approach

Following the initial provisioning of a sandbox, data is identified from source systems and ingested into an HDFS cluster in our Sandbox. Additional data from the Staging, Foundation and Access and Performance Layers may also be included to add context in support of the business problem being tackled. This data is analysed using any and all tools at the disposal of the Data Scientist and additional data manipulations, transformations, aggregations and selections performed using Map-Reduce. See Figure 7.

Having completed some initial exploration and data preparation in Hadoop it is quite typical for the intermediate results produced to be ingested into a Discovery Sandbox in a relational database for further manipulation and analysis using relational tools. This is shown in Figure 8.

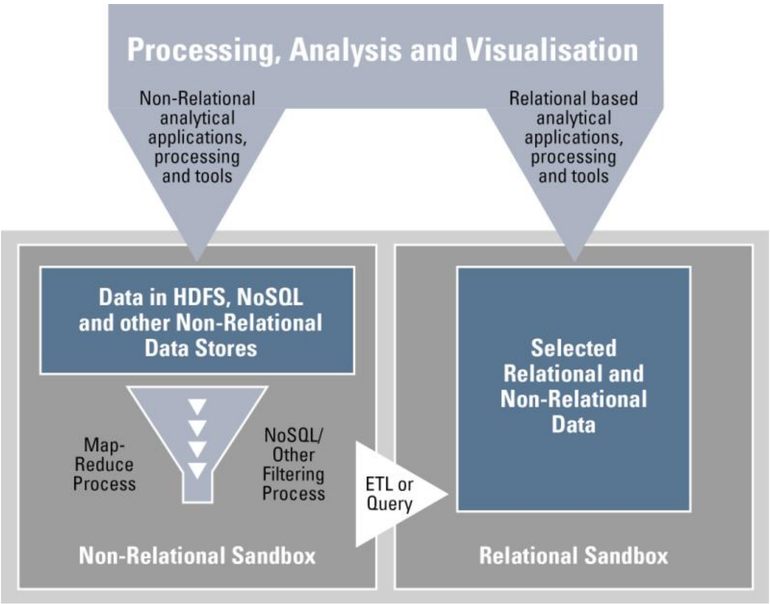
Processing, Analysis and Visualisation
Non-Relational analytical applications, processing and tools
Relational based analytical applications, processing and tools
Data in HDFS, NoSQL and other Non-Relational Data Stores
Map-Reduce Process
NoSQL/ Other Filtering Process
ETL or Query
Selected Relational and Non-Relational Data
Non-Relational Sandbox
Relational Sandbox

Figure 8. Combining relational and non-relational sandboxes for complete knowledge discovery

The choice of where to focus the main knowledge discovery task will be determined by the organizational context, the technical skills of the Analyst and the specific problem being tackled. That is to say it may only require analysis in the Hadoop cluster or on the relational infrastructure or both. It is also reasonable to suppose this choice may change over time as visualisation and analytical tools are further developed for Hadoop.

Once a result has been determined from the analysis it must be deployed. See Figure 9. Deployment can take several forms, and of these the most typical is for the knowledge to be deployed into inbound channels in order to change the flow or behaviour of a process in some fashion. This is shown in Figure 9 as the 'Real-Time Event Detection Process' and discussed in more detail in the section titled 'Big Data needs Big-Execution and agile IM'. This would be the case, for example, if the company was looking to build a system that could allow them to send an upgrade offer by SMS to customers that were within a short walk of a store if they were at risk of churning to a competitor. The offer itself may then also be based on the factors driving the churn risk as well as the value of the individual (lifetime value, Klout score etc.).

It is likely that the knowledge (e.g. a data mining model) will need to be re-coded in some fashion (i.e. it can't be used in the way it was originally generated) and will need to be wrapped in additional business logic to prevent conflicts with other deployments and comply with organizational governance. Multiple models may have overlapping rules, and other important factors must be considered such as offer reciprocity, control group optimisation, managing customer fatigue due to over marketing and respecting 'do not market' requests.

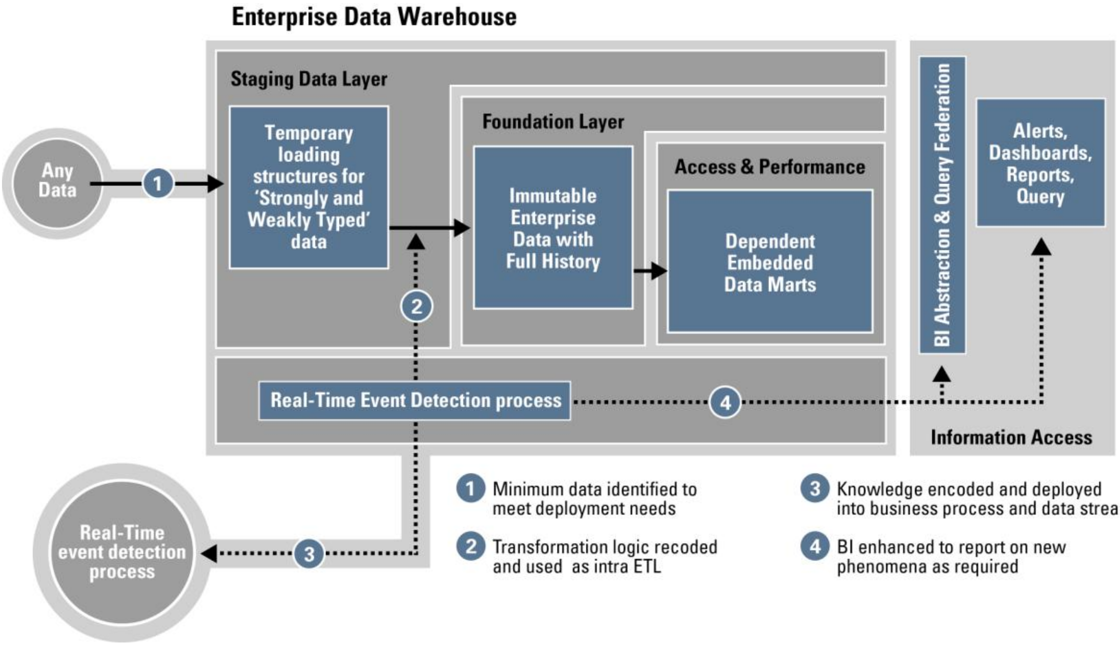
Enterprise Data Warehouse
Staging Data Layer
Any Data
1
Temporary loading structures for 'Strongly and Weakly Typed' data
Foundation Layer
Immutable Enterprise Data with Full History
Access & Performance
Dependent Embedded Data Marts
2
BI Abstraction & Query Federation
Alerts, Dashboards, Reports, Query
Real-Time Event Detection process
4
Information Access
Real-Time event detection process
3
1 Minimum data identified to meet deployment needs
2 Transformation logic recoded and used as intra ETL
3 Knowledge encoded and deployed into business process and data streams
4 BI enhanced to report on new phenomena as required

Figure 9. Final deployment stage for 'Knowledge Stripping' approach

Sometimes the new knowledge is not something that is directly relevant to inbound channels but is still of value to the business in later decision making, so needs to be captured within our 'corporate brain', the Data Warehouse. In these circumstances, we need to logically encode the data path in order to represent the result within the Data Warehouse. In other words, we need to somehow either find the value somewhere else or optimise the Map-Reduce steps to implement it as part of the ETL, to load data into the Foundation Data Layer.

Having identified new knowledge and changed the business process in some fashion it's also important that we can monitor the results. This is likely to require new reports to be added or existing reports enhanced in some fashion.

Knowledge Pooling – Assume the ROI Approach

Customers with a more fundamental belief in the value that can be derived from the additional weakly typed data typically opt for a Knowledge Pooling approach. The most obvious example of customers adopting this 'build it and they will come' approach is from the Intelligence agencies, but commercial organizations have also adopted this approach for specific use cases such as for pooling web logs.

In this approach, the primary task is to build a Hadoop cluster and fill it with the available data as a pool that can be dipped into to find whatever is required. Often this data is combined with strongly typed data coming from any of the Data Warehouse layers but most typically the Foundation or Access and Performance Layers.

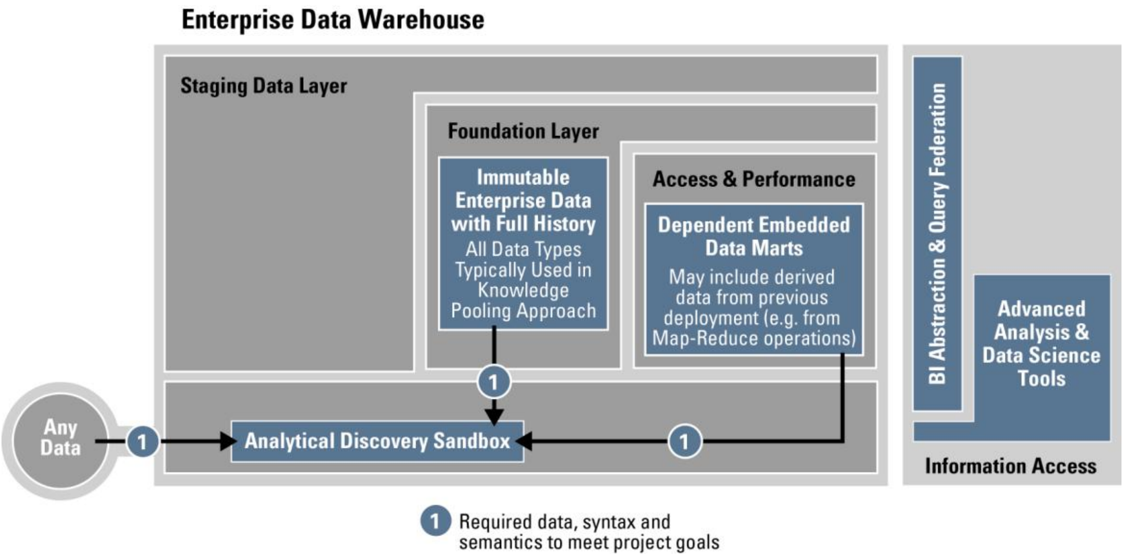
Enterprise Data Warehouse
Staging Data Layer
Foundation Layer
Immutable Enterprise Data with Full History
All Data Types Typically Used in Knowledge Pooling Approach
Access & Performance
Dependent Embedded Data Marts
May include derived data from previous deployment (e.g. from Map-Reduce operations)
BI Abstraction & Query Federation
Advanced Analysis & Data Science Tools
Information Access
Any Data
1
Analytical Discovery Sandbox
1
1
1
Required data, syntax and semantics to meet project goals

Figure 10. Knowledge discovery for 'Knowledge Pooling' approach

In many cases, the data required to address any particular business problem will already be present in the data pool. If not, the data pool can be augmented with this new data which may come from any source and will be stored in our cluster. The remaining tasks of analysing the data, building a model of some type and then deploying the knowledge to inbound channels as appropriate are very much the same as before, but there are some differences in subsequent deployment steps. See Figure 10.

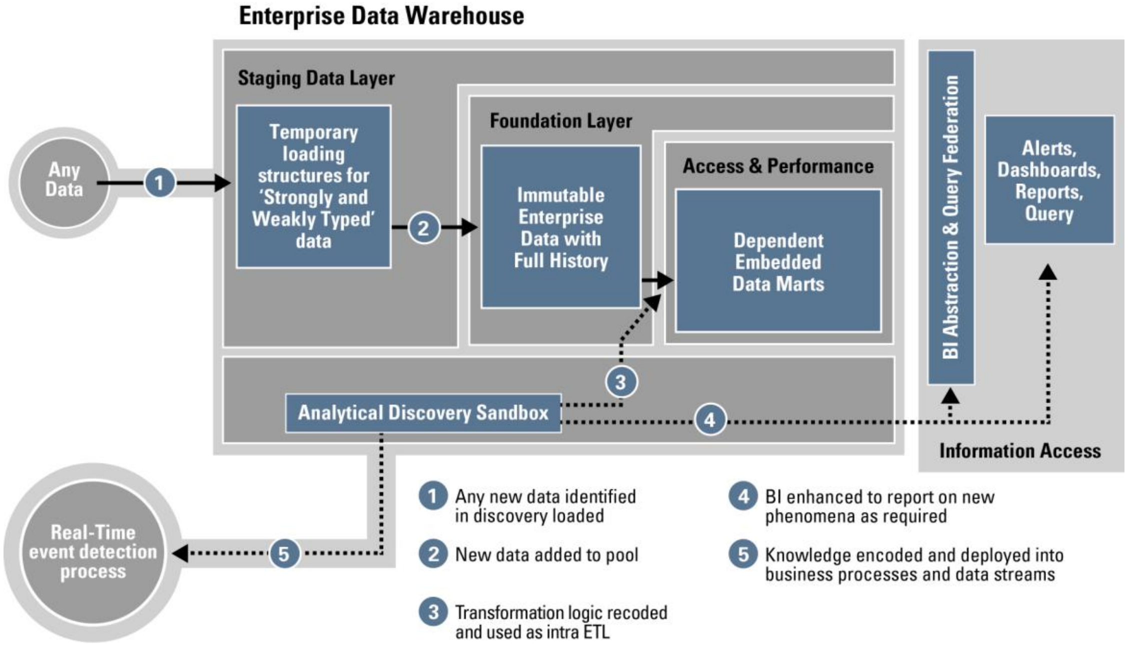
Enterprise Data Warehouse
Staging Data Layer
Any Data
1
Temporary loading structures for 'Strongly and Weakly Typed' data
2
Foundation Layer
Immutable Enterprise Data with Full History
Access & Performance
Dependent Embedded Data Marts
3
Analytical Discovery Sandbox
4
BI Abstraction & Query Federation
Alerts, Dashboards, Reports, Query
Information Access
Real-Time event detection process
5
1 Any new data identified in discovery loaded
2 New data added to pool
3 Transformation logic recoded and used as intra ETL
4 BI enhanced to report on new phenomena as required
5 Knowledge encoded and deployed into business processes and data streams

Figure 11. Final deployment stage for 'Knowledge Pooling' approach

We can consider our underlying pool of data to be part of the Foundation Layer of our Data Warehouse. While it will be physically deployed on a different set of technologies, logically it complements our strongly typed data with weakly typed data. The data is our immutable source of truth in just the same way. Our task then is to add any new data that has been used in the analysis to this pool of data; either to the relational store if strongly typed or the Hadoop store otherwise. Any subsequent transformation steps previously encoded in Map-Reduce jobs will need to be optimised and made suitable for a production setting and then included as part of the ETL feed of our Warehouse. This downstream data then logically becomes part of our Access and Performance Layer as it represents an interpretation of data and is not fact.

Just as before, we would also need to adjust the relevant reports or add new ones to track changes resulting from our deployment. The deployment of knowledge from the Knowledge Pooling approach is shown in Figure 11.

Choosing the Right Approach

It’s also interesting to look at other differences implied by the two approaches. Two dimensions of interest are the scope of the data represented in relational vs non-relational forms (i.e. the proportion of the model, not the volume of data) and the location of the analysis (i.e. to what extent will the analysis take place on a relational vs the non-relational technology platform.

Figure 12 shows these dimensions in graphical form. While the scale shown is somewhat arbitrary we suggest that companies adopting a Knowledge Pooling approach will have a broader range of data stored non-relationally and will prefer to perform a larger proportion of the actual analysis on that platform, perhaps before transferring the results into the relational world for combining with existing data.

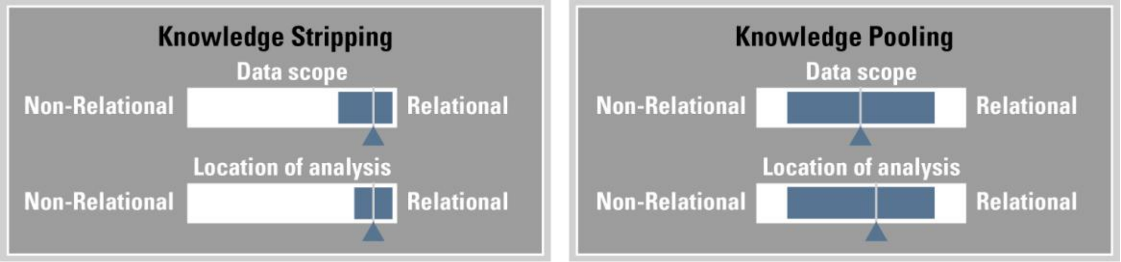

Figure 12. Comparison of Data Scope and Analysis Location between approaches

The opposite will be true for organizations more attracted to the Knowledge Stripping approach. As the approach is geared towards addressing a very specific business problem (i.e. tackling problems serially rather than building a more general capability), and given the analysis skills will already exist in the relational world, it is likely that non-relational technology will be used as a simple ‘bit-locker’ with result-sets filtered in place and pulled through to the relational platform for analysis. The ability to execute this kind of workflow from the relational domain with little to no understanding of how the data is physically stored or manipulated on the non-relational platform at enterprise scale should clearly inform your technology adoption choices.

Regardless of the approach you adopt initially it is very likely that the needle on both graphs will move to the left over time as vendors such as Oracle invest heavily in building new tools and capabilities that span both relational and non-relational domains.

Your choice of approach may also inform the way in which you deploy you Knowledge Discovery Layer. For companies adopting a Knowledge Stripping approach it is often simpler for them to isolate the Knowledge Discovery Layer onto a completely separate environment than on an operational (DW) platform. We have seen this to be the case in larger organizations with mature operational procedures as these can create a barrier to providing a Data Scientist the freedom they need in order to perform their job effectively. If it’s more practical for you to give your Data Scientists the flexibility they need using an isolated (relational) environment, then it makes sense to do so.

Big Data needs Big Execution and Agile IM

In order to make every decision as *'good'* as we can we need to bring the results of knowledge discovery to the business process and at the same time track any impact in the various dashboards, reports and exception analysis being monitored. New knowledge discovered through analysis may also have a bearing on business strategy, CRM strategy and financial strategy going forward. You can see these flows in figure 13 which shows a classification of various analysis types together with their scope and impact on the business.

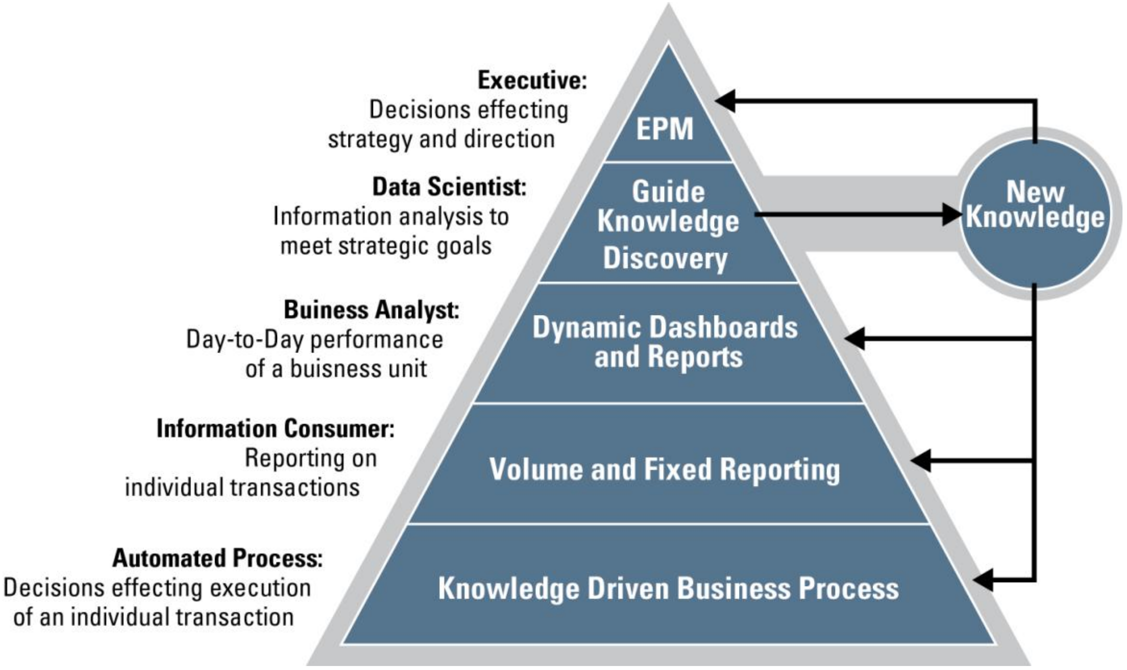
Executive:
Decisions effecting strategy and direction
Data Scientist:
Information analysis to meet strategic goals
Buiness Analyst:
Day-to-Day performance of a buisness unit
Information Consumer:
Reporting on individual transactions
Automated Process:
Decisions effecting execution of an individual transaction
EPM
Guide Knowledge Discovery
Dynamic Dashboards and Reports
Volume and Fixed Reporting
Knowledge Driven Business Process
New Knowledge

Figure 13. Scope of knowledge deployment to information consumers and types of analysis

Knowledge discovered through analysis may be deployed into inbound message flows to automate message routing (this was shown in previous figures as the 'real-time event detection process').

While all the linkages shown in Figure 13 are important, it is the link to business process that will have the most financial leverage and is also the link most often found missing! The challenge is in designing business processes to leverage the knowledge and creating the IT framework to orchestrate it.

In most instances knowledge will need to be re-coded in some fashion and blended with the knowledge gained from previous analysis. An additional governance wrapper must also be applied in order to protect the business from unintended consequences. For instance, in a CRM context we would want to protect the customer from being *'over-fatigued'* by marketing offers. Figure 13 shows a conceptual view of this inbound data flow and routing.

Figure 14 also shows how contextual data may be required to enhance the data within the inbound data stream in order to maximise the scope of the knowledge deployed. This data is typically either master data such as "customer" or contextual information such as current location.

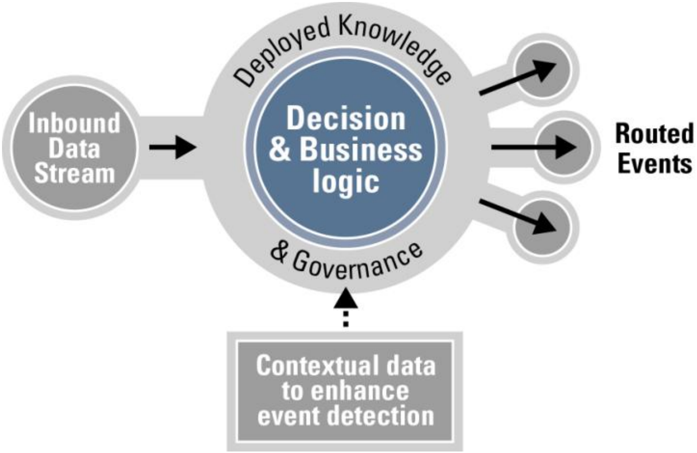
Inbound Data Stream
Deployed Knowledge
Decision & Business logic
& Governance
Routed Events
Contextual data to enhance event detection

Figure 14. Knowledge deployed into inbound data stream for Real-Time event detection

The performance of the solution to effectively route the inbound data stream is critical as in many instances the business benefit may be directly tied to it (e.g. the customer accepting the offer presented is tied to making the right offer and making it quickly). It follows that any related systems, such as the ones providing the contextual data, must have appropriate performance and availability characteristics.

As depicted in the simplified functional model for analysis shown in Figure 1, in order to track changes, either positive or negative, flowing from the additional knowledge we gained from our new data, it is imperative that the appropriate dashboards and reports are adjusted to perform the analysis feedback and control functions.

Just as your company builds out the capacity and knowledge to interact with customers more frequently and with more relevance, the climate for CRM among your customer base may also change. i.e., customers in general may be more accepting of the principals and may tolerate a broader range of offers. It follows then that the feedback to strategy and financial planning for your organization are also critical components.

Cautious First Steps

Big Data presents a significant opportunity to create new value from masses of data, much of which you may have access to already in your organization. Whether you choose to adopt a Knowledge Stripping or Knowledge Pooling approach from the start, it's clearly important that you quickly determine appropriate governance procedures in order to manage development and implementations over the life of the technology and data. Failure to consider the longer term implications of development will lead to productivity issues and cost escalations.

On the face of it, the cost of physically storing large quantities of data is dramatically reduced by the simplicity by which data can be loaded into a Big Data cluster because we no longer require a complex ETL layer seen in any more traditional Data Warehouse solution. The cluster itself is also typically built using low cost commodity hardware and analysts are free to write code in almost any contemporary language through the streaming API available in Hadoop. However true in part, the actual total cost of ownership may be far from "low cost" when considered over the full lifecycle of the solution:

- The business logic used within an ETL flow to tokenise a stream of data and apply data quality standards to it must be encoded (typically using Java) within each Map-Reduce program that processes the data and any changes in source syntax or semantics must be managed. The performance of the cluster and accuracy of the result will depend on the skills of your programmers in writing scalable bug free code and in their understanding of each data flow, even those created many years before.
- Although the storage nodes in a Hadoop cluster may be built using low cost commodity x86 servers, the master nodes (Name Node, Secondary Name Node and Job Tracker) requiring higher resilience levels to be built into the servers if disaster is to be avoided. Map-Reduce operations also generate a lot of network chatter so a fast private network is recommended. These requirements combine to add significant cost to a production cluster used in a commercial setting.
- Compression capabilities in Hadoop are limited because of the HDFS block structure and require an understanding of the data and compression technology to implement adding to implementation complexity with limited impact on storage volumes. This is not the case with commercial database offerings such as Oracle's Exadata platform which is able to achieve very high compression rates with low impact on CPU overhead.
- The story is similar for security which is also not built into Hadoop and so must be coded into the data loading and Map-Reduce programs. This is a significant issue in a commercial setting due to the potential value of the data and impact on the organization if data is accessed by unauthorised individuals or stolen.
- Other aspects to consider include the true cost of ownership of pre-production and production clusters such as the design build and maintenance of the clusters themselves, the transition to production of Map-Reduce code to the production cluster in accordance with standard operational procedures and the development of these procedures.

Whatever the true cost of Big Data compared to a relational data storage approach, it's clearly important that you develop your Big Data strategy with your eyes open, understanding the true nature of the costs and complexity of the infrastructure, practice and procedures you are putting in place. Failure to do so will inevitably lead to disappointment.

Conclusions

For many years now data driven companies have been struggling with just what to do with the data they collect but does not fit readily into relational databases such as text and web logs. Big Data offers the promise of unlocking the potential of this data and opens up new avenues for value creation such as through the correlation of social network and sentiment data with purchase behaviour to form a more complete picture of every customer.

In this white paper we have set out Oracle's Reference Architecture for Information Management, showing how weakly typed data can be incorporated into the value creation flow and highlighted the two dominant approaches being adopted by the data driven companies we are working with.

Although the tools used to manage and analyse Big Data can be downloaded for free and a cluster built using low cost commodity hardware, this does not necessarily equate to a dramatic reduction in the cost to own per TB of data. If your commercial reality requires you to also factor in non-functional requirements such business continuity, backup and recovery, security, 24 x 7 Enterprise Support infrastructure, patching support etc. which are all things you would normally consider as an important part of your IM infrastructure, then the real costs can grow considerably. The same is true for the real costs of building your own HDFS cluster and private network, which is a non-trivial exercise for a production cluster. As power and cooling costs have become significant you also need to factor these in when considering the efficiency of the platform.

There can be considerable overlap in the applicability of a traditional relational or HDFS approach to managing any particular set of data – so it is likely that you will always have a choice to make at the start of any new project. Factors such as the rate, mechanism and encoding type of the data being received onto the platform, how it needs to be analysed, non-functional considerations and others will all need to be considered. You will also need to look past other obvious temptations such as the fact that Hadoop is new and your staff are eager to experiment and build their skills.

We have long advocated the need for Information Management to be a first class system in IT – the way information is managed and exploited is becoming of increasing importance in today's ultra competitive world, so it has become critical to think through the IM implications when dealing with any OLTP system. In addition, as applications themselves become more operationally linked to IM systems, the characteristics of those systems must now also match.

It also follows that if IM system is to be considered as a first class system IT must do a better job of keeping up with the rate of change of the business, enabling it, not constraining it. Oracle's Reference Architecture does this through design. As well as the well defined abstraction layers that limit the impact of change we call out in particular the role of the BI Abstraction layer in managing roadmap and the Knowledge Discovery Layer to support a Right to Left RAD based development approach. By supporting the alignment between business need and IT delivery capability the need for 'shadow IT' is eliminated and costs reduced.

To profit fully from the potential additional insight offered by Big Data organizations must pay attention to matching the Big Data with the appropriate levels of investment in 'Big Insight' and 'Big Execution'. That is, just collecting large amounts of loosely typed data without also paying attention to the way you will analyse it to create insight and then execute against that insight to generate value is unlikely to be successful.

Big Data offers tremendous potential to add value to business. It also offers IT an equal opportunity to re-invent many of the significant problems in Information Management such as Data Quality, business alignment and Single Version of the Truth that we have overcome through frameworks such as Oracle's Reference Architecture for IM and technology platforms such as Exadata. Companies who fail to integrate Big Data into their overall Information Management strategy are likely to reduce their overall capability in this area over the longer terms once they have built a new generation of legacy.

Finding out more about Oracle's IM Reference Architecture

We recommend as a starting point you read the original white paper describing the Information Management Reference Architecture called 'Enabling Pervasive BI through a Practical Data Warehouse Reference Architecture'. Click here or you can find it on www.oracle.com either by entering the above title into the search facility or the document number which is '058925'.

There are also a number of short vignettes describing various aspects of the IM Reference Architecture that are available from YouTube. Click here or simply search YouTube for 'Oracle Information Management Masterclass' and select the videos by 'OracleCore.' You should see they are all numbered to help you progress through the material.

Oracle also runs a series of regular 'Master Classes' describing the IM Reference Architecture. These master classes are run as interactive whiteboard sessions hosted by Oracle Architects for Architects. The sessions are discussion-based and entirely PowerPoint free. To find out when the next Master Class is scheduled in your local country, please contact your local sales representative.

You can also find a further set of architectural materials to help you better plan, execute and manage your enterprise architecture and IT initiatives here. As well as the Information Management and Business Intelligence topics, these materials also cover other important areas such as SOA, Event-Driven Architecture, BPM and Cloud Computing.

ORACLE®

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

Information Management and Big Data,
A Reference Architecture
February 2013
Author: Doug Cackett
Contributing Authors:
Andrew Bond and John Gouk

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

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200

oracle.com

Hardware and Software, Engineered to Work Together