

# Information Management and Big Data

## A Reference Architecture

ORACLE WHITE PAPER | SEPTEMBER 2014





## Table of Contents

Introduction	1
Background	2
Information Management Landscape	2
What is Big Data?	3
Extending the Boundaries of Information Management	5
Information Management Conceptual Architecture	8
Information Management Logical Architecture view	10
Data Ingestion Process	13
Understanding schema-on-read and schema-on-write	15
Information Provisioning Process	17
Understanding differences in query concurrency costs and information quality	18
Discovery Lab Sandboxes and the Data Scientist	19
Rapid Development Sandboxes and iterative development methods	22
Technology approach and the Big Data Management System	24
Big Data Adoption	26
Conclusions	29
Finding out more about Oracle's Information Management Reference Architecture	30



## Introduction

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

In the last few years we have seen a significant rise in the number of organisations adopting Big Data technologies as a means by which they can economically manage more of their data and analyse it to support more *fact based* decision making.

So given the strategic business imperative and increasing technology capability, it is important to understand how these technologies relate to the existing Information Management estate and how best to combine them (if at all) into a coherent platform that will not only support a management by fact approach, but also facilitate the discovery of new facts and ideas and set these into the business process.

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 these solutions by organisations across all industry sectors as they seek to exploit new categories and types of data 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 (IM). 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 Database Machine and in-memory technologies such as Oracle Exalytics and the In-Memory Database Option, 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, volumes and possibilities.

Modern hardware and software technologies are changing what it's possible to deliver from an Information Management perspective. In our experience the overall architecture and organising principles are more critical, as, a failure to organise data effectively will undoubtedly result in significantly higher costs and poor business alignment.

In the past couple of years we have seen a significant increase in the use of Big Data and analytical technologies as our customers bring more data under management (this is often the rather "messy" data that has hitherto been too big or complex to manage) and do more with that data through analytical discovery.

#### 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.

## What is Big Data?

This section is intended as a simple primer to Big Data for those who are not as confident in what is meant by the term and how the technologies it encompasses might be used to create additional insights and business value. If you are familiar with Big Data approach and technologies we suggest you skip this section.

Big Data is a term often applied by people to describe 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 100'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, in the past they were less able to deal with semantically rich data such as free text. As a result many organisations restricted the data coverage of IM systems to a narrow range of data. Deferring the point at which this kind of rich data, which is often not fully understood by the business, also has significant appeal and avoids costly and frustrating modelling mistakes. It follows then that by being more inclusive and allowing greater model flexibility additional value may be created - this is perhaps one of the major appeals of the Big Data approach
- » **Value.** The commercial value of any new data sources must also be considered. Or, perhaps more appropriately, we must consider the extent to which the commercial value of the data can 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. Big Data technologies can have a significant impact on the overall picture of "information ROI" as well as more specific project viability by minimising up-front investment prior to developing a more complete understanding of value through the discovery process. See *Extending the Boundaries of Information Management* for more details of this discovery process.

We discuss the topic of technology adoption in "*Big Data Adoption*" later in this white paper, but it is worth noting that the adoption pattern adopted is often contingent on the level of certainty to which the value of the data can be ascribed and agreed upon ahead of time in order to attract project funding from the business.



To make data understandable a schema must be applied to it 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. This topic is discussed in more detail in the section *Understanding schema-on-read and schema-on-write*.

One area where there is almost no difference between the approaches is in the analysis technologies applied. Data Scientists will typically use a broad range of technologies such as SQL, Data Mining, statistical and graphical analysis depending on the problem being tackled.

Another area of confusion is to do with the processing of unstructured and semi-structured data. This is mostly because the terms conflate two key differences of the physical representation of the data at rest or in-flight (syntax) and its inherent meaning (semantics). A file containing JSON or XML data is as easily processed by relational and Big Data technologies, but if the meaning of the data is not fully understood or could change over time, having some schema flexibility will be important.

Many of the points raised in this primer on Big Data are discussed in more detail on the sections *“Understanding schema-on-read and schema-on-write”* and *“Understanding differences in concurrency costs and quality”*.

Perhaps the most critical difference between relational and Big Data technologies is really more to do with a philosophical positioning than anything technical. The technologies come from different places and the people involved have a different world view. Because of the schema freedom offered by Big Data technologies we are more able to accept and process really messy data and evolve the schema as we go. This makes it ideal for data discovery and leads to a greater level of agility.

But agility comes at a cost! You can imagine what might happen in your business if every time you analysed total sales for the year you received a different result! It would not take long before trust in the data had broken down and the IM solution abandoned. We discuss Governance and differences between the approaches to schema in a number of sections later in this white paper.

## Extending the Boundaries of Information Management

Few organisations would suggest their Information Management systems are perfect or could not be improved. Many consist of multiple fragmented silos of information, each of which contains a narrow scope of data and serves a limited community of Business Analysts and users. As a result of the fragmentation and lack of architectural vision, making simple changes (provision new data, new reports, changing hierarchies etc) becomes difficult resulting in both business and IT frustrations.

So what should contemporary IM systems strive to achieve?

---

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

*Thomas H. Davenport and others have gone on to observe that organisations who are able to go one step further and operationalise decisions based on facts will be more profitable again. It stands to reason then that those organisations who are able to apply facts to more business processes and more steps in each process will have a more optimised and profitable outcome.*

---

Given the complicated IM estate that already exists, the key question you are probably asking yourself right now is “just how can Big Data technologies be applied to create additional business value or reduce the costs of delivering Information Management?” After all, simply adding Big Data technology to your existing estate will do nothing in of itself, other than add additional costs and complexity to an already costly and complex environment. It is only through decommissioning some of your existing estate or enabling additional insights that would hitherto not have been possible, will you actually add value to the business.

While there are undoubtedly some examples of how Big Data technologies can reduce costs, for many of our customers, Big Data has acted as a lever to focus management attention on the organisation’s analytical capabilities to first find, and then monetise the additional insights gained from new data or data at a lower grain. In this way, Big Data has focused management attention and accelerated the move towards a *fact based* management approach, often by offering additional opportunities for analysis (more data or more points within a business process) or more opportunities for applying analysis to operational decision making.

IM solutions can often become very fragile over time, leading to a significant burden from an operational costs and IT productivity standpoint. Implemented appropriately, Big Data can have a profound impact on these aspects due to the schema-on-read approach and a reduction in integration efforts. You will find more discussion on these topics in the sections on “*Understanding* schema-on-read and schema-on-write” and “Discovery Lab Sandboxes and the Data Scientist”.

What we mean by “analytics”:

The systematic computational analysis of data or statistics.

The role of the Data Scientist is also closely related to analytics. A Data Scientist is focused on extracting knowledge from data and using a scientific approach and is typically skilled in computer science, statistics and advanced analytical tools.

Others prefer the term Quant, Data Miner or Advanced Data Analyst when describing the role.

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 developed and tested on the data. These propositions are then delivered through the appropriate mechanism and the outcome measured to ensure the consequence is a positive one.

Note how the operational scope of our solution is bounded by the three key dimensions of Strategy, Technology and Culture. To maximise potential, these three dimensions need to be in balance. There is little point in defining a business strategy that cannot be supported by an organisation’s IT capacity or your employees ability to deliver it.

A lack of balance between these three dimensions may explain why analytics is rarely pervasive across an entire organisation –the use of analytics is mostly patchy, applied to a limited set of problems or business processes in a narrow range of departments.

In order for a company to continue to lead the market it must continually innovate to find new opportunities to add value to existing customers, further optimise existing business processes, cut costs, or find new markets. If data is to lie at the heart of your *fact based* management approach it follows that considerable focus should be placed on enabling new discoveries as well as the process of taking insights gained to operationalise them quickly and efficiently.

Through the use of Big Data technologies a company may be able to store data at a lower granularity, keep the data for longer or apply a range of analytical tools that would until now have been too expensive in more traditional relational technologies. There are reasonable grounds to say that more data will outperform a cleverer algorithm almost every time!

Another view of this functional model is shown in Figure 2. It shows how, when driven by a new business challenge or the availability of new data, a Data Scientist will investigate the data by bringing it together with other enterprise data using a variety of tools and techniques.

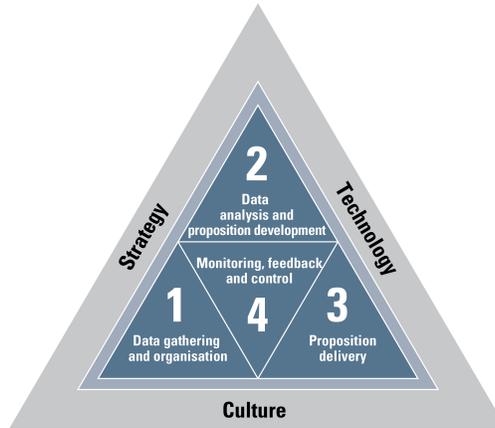


Figure 1. Simplified functional model for data analysis

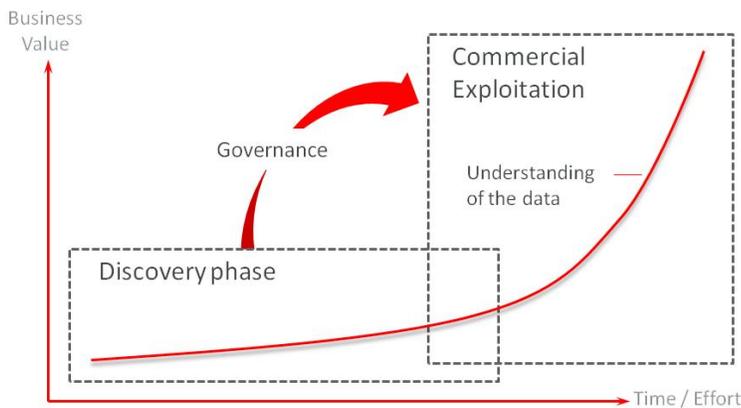


Figure 2. Discovery and commercial exploitation for new data.



Having discovered something of value in the data that can be exploited, the next challenge is to operationalise the insight in some fashion, often by including it within a business process, and then monitoring the effect on the business process to ensure the behaviour first observed on the workbench is exhibited in real life and the results are positive.

It is important to observe that as the people, process and tools used during the discovery phase will be different to those required for commercial exploitation there is a step required to go from one to the other. This is shown as a governance step in the diagram due to the important role it plays to the success and management of information overall. The goal in design is to minimise the time taken for the discovery part of the process and reduce the size of the step between discovery and commercial exploitation.

The most successful Data Science teams use a wide array of tools and techniques in discovery and program in a partisan manner. In contrast, most successful organisations have a very limited and rational set of tools for exploitation and enforce strict coding standards. The governance step is required to reproduce the insight using this rationalised production toolset and within coding standards and security constraints. As part of the *“route to production”* for every new insight, careful consideration needs to be applied to ensure the behaviour is reproducible in the broader commercial context, and how / when the model needs refreshing due to a decline in performance over time.

Business Intelligence Competency Centres (BICC) continue to play an important role in many organisations to drive adoption and increase in the value and effectiveness of BI tools. As we have previously discussed in this white paper, adoption of “analytics” is often patchy, limited to a small range of departments or business problems. If the value of Big Data is bound by an organisations ability to deliver new insights, it makes clear business sense to pay particular attention to broadening the adoption of analytics across the organisation. An Analytical Competency Centre (ACC) provides the organisational context (including senior executive patronage) and skilled resource pool required to achieve this.

## Information Management Conceptual Architecture

Oracle's Information Management Conceptual Architecture is shown in Figure 3 and shows key components and flows in a compact form. It highlights in particular how the innovation derived from analytical discovery is somewhat separated from the day to day execution, with the governance function previously discussed linking the two.

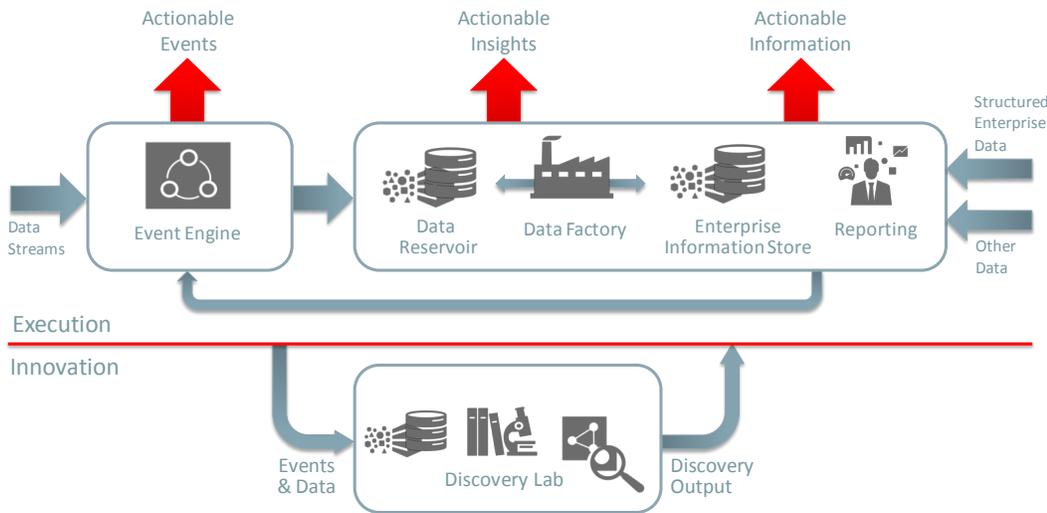


Figure 3. Information Management Conceptual Architecture view.

When defining the role of Big Data components within information architectures, it is helpful to align components and their purposes along the flow of data into the larger data environment. In this compact, flow-based conceptual model, we define a set of components, many of which may be present in your existing information architectures. These components are:

- » **Event Engine:** Components which process data in-flight to identify actionable events and then determine *next-best-action* based on decision context and event profile data and persist in a durable storage system.
- » **Data Reservoir:** Economical, scale-out storage and parallel processing for data which does not have stringent requirements for formalisation or modelling. Typically manifested as a Hadoop cluster or staging area in a relational database.
- » **Data Factory:** Management and orchestration of data into and between the Data Reservoir and Enterprise Information Store as well as the rapid provisioning of data into the Discovery Lab for agile discovery.
- » **Enterprise Information Store:** Large scale formalised and modelled business critical data store, typically manifested by an (Enterprise) Data Warehouse. When combined with a Data Reservoir, these form a Big Data Management System.
- » **Reporting:** BI tools and infrastructure components for timely and accurate reporting.
- » **Discovery Lab:** A set of data stores, processing engines, and analysis tools separate from the everyday processing of data to facilitate the discovery of new knowledge of value to the business as previously shown in Figure 2. This includes the ability to provision new data into the Discovery Lab from outside the architecture.



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 red line in Figure 3) helps inform system requirements for security, governance, and timeliness.

Through our recent work with a wide range of customers we have observed how an organisation's priorities will define the scope of the deployed solution, especially in the initial project phases. For many organisations the most important first step is often to deliver a Discovery Lab in order to prove the potential value of their data and analytics overall. For others where investment decisions and the project drive have come from senior business executives, a broader implementation scope involving more components has been adopted.

Our work in this area has led us to identify a number of distinct implementation scopes which focus in particular on different components of our Conceptual Architecture. These are discussed further in the section on "*Big Data Adoption*" at the end of this white paper.

## Information Management Logical Architecture view

Oracle's Information Management Reference Architecture describes the organising principles that enable organisations to deliver an agile information platform that balances the demands of rigorous data management and information access.

It's an abstracted architecture with the purpose of each layer clearly defined. The main components shown in Figure 4 include:

- » **Data Sources.** These represent all potential sources of raw data which are required by the business to address its information requirements. Sources include both internal and external system. Data from these systems will vary in structure and presentation method.
- » **Data Ingestion and Information Interpretation.** These are the methods and processes required for ingestion and interpretation of information to and from each of the data layers in our architecture. Importantly, their shape is intended to illustrate the differences in processing costs for storing and interpreting data at each level and for moving data between them.
- » **Raw Data Reservoir.** An immutable and un-modelled data store with data at the lowest level of granularity. The Raw Data Reservoir may comprise both relational and non-relational components with data typically stored in the same form as the underlying data source. In the case of relational data the reservoir acts as a transient area to assist in the data loading process into upper layers, or as a classical Operation Data Store of near real-time un-integrated data.
- » **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 that the business has decided to manage formally.
- » **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 to aggregate data which will be consumed by a further analysis process.
- » **Discovery Lab and Rapid Development Sandboxes.** These sandboxes facilitates the addition of new reporting areas through agile development approaches and the analytical discovery process to identify new knowledge that can be commercially exploited in some form. This is discussed in much more detail in the sections "*Discovery Lab Sandboxes and the Data Scientist*" and "*Rapid Development Sandboxes and iterative development methods*".
- » **Virtualisation & 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 agile approaches to development, migration to the target architecture and the provision of a single reporting layer from multiple federated sources such as is typically found in large multi-national organisations with multiple autonomous operational units and a governing parent entity.
- » **Enterprise Performance Management.** Includes tools such as Financial Performance Management, financial forecasting and Balanced Scorecard tools. These are distinct from other BI Assets as they typically include analytical modelling and may read/write to a data store in a managed fashion.
- » **Pre-Built and Ad-hoc BI Assets.** The standard set of BI tools, reports and dashboards across a range of access devices.
- » **Information Services.** These information services enable the seamless operationalisation of information within the organisation and to the wider trading community. For example, a new customer segmentation might be generated each month and the results written back to operational systems such as the event engine. Other links might include the updating of data in solutions such as Master Data Management or technologies such as BPEL.

- » **Advanced Analysis and Data Science Tools.** These tools are distinct from standard BI tools as they typically have a much more limited user community (Data Scientists, Statisticians and more advanced Business Analysts).

Master Data Management (MDM) solutions are considered to hold the 'master' for any given business entity and are one of the many sources for our Information Management platform. The role of this platform in respect of MDM is to preserve a record of changes across time, enable an analysis of these changes and provide the data as context to other analyses. Further data quality related checking and enrichment may also occur in the Information Management platform as a separate process (as part of Discovery rather than Data Ingestion) which may result in changes to master data. These changes may be pushed back to the MDM solution and the new update received back into the Information Management platform through the standard flow.

Unlike standard BI tooling, Advanced Analysis tools and applications such as forecasting and Data Mining may create new data as part of the analysis process. Under the control of the tool or application, data can be read and written to and from the Analysis Sandbox area of the Access and Performance Layer. Once the final result is obtained a further step in the process may take a result set (such as a customer list or new customer segmentation) and move it to the required operational systems (e.g. Master Data Management or an Event Engine).

The Foundation Data Layer and Access and Performance Layers offer two additional levels of abstraction that further reduce the impact of schema changes in the data platform while still presenting a single version of the truth to consumers.

As previously discussed, in order for a business to remain competitive it must continually innovate by finding new insights in the data that can be commercially exploited. This implies that business processes and the reporting requirements required to support them will be in a constant state of flux, evolving continuously. To meet these competing demands, we abstract data above our Raw Data Reservoir into two modelled layers:

- » **Foundation Data Layer:** abstracts the data away from the business process through the use of a business process neutral canonical data model. This gives the data longevity, so that changes in source systems or the interpretation placed on the data by current business processes does not necessitate model or data changes.
- » **Access and Performance Layer:** allows for multiple interpretations (e.g. past, present and future) over the same data as well as to simplify the navigation of the data for different user communities and tools. Objects in this layer can be rapidly added and changed as they are derived from the Foundation Data Layer.

It is worth noting that many commercial-off-the-shelf packages (COTS) such as HR, Finance, CRM, Sales and Service include a pre-packaged reporting solution, often based on a dedicated data mart approach. These packages typically omit the Foundation Data Layer model from the design, preferring to populate the data mart directly using a dedicated ETL process. These applications do not require the additional level of abstraction offered by the Foundation Data Layer as the developers are in control of the application source data, ETL, data mart design and migration process.

Although COTS solutions may include a full range of reports and analysis capabilities relating to a given application, some data may also be required to provide context to a broader range of analysis. For example, HR data (staff numbers and training per department) might prove an interesting addition to a sales analysis in a department store. In these circumstances, as well as the out of the box data feed to the data mart, an additional feed will also be required into the Foundation Data Layer for the data in question.

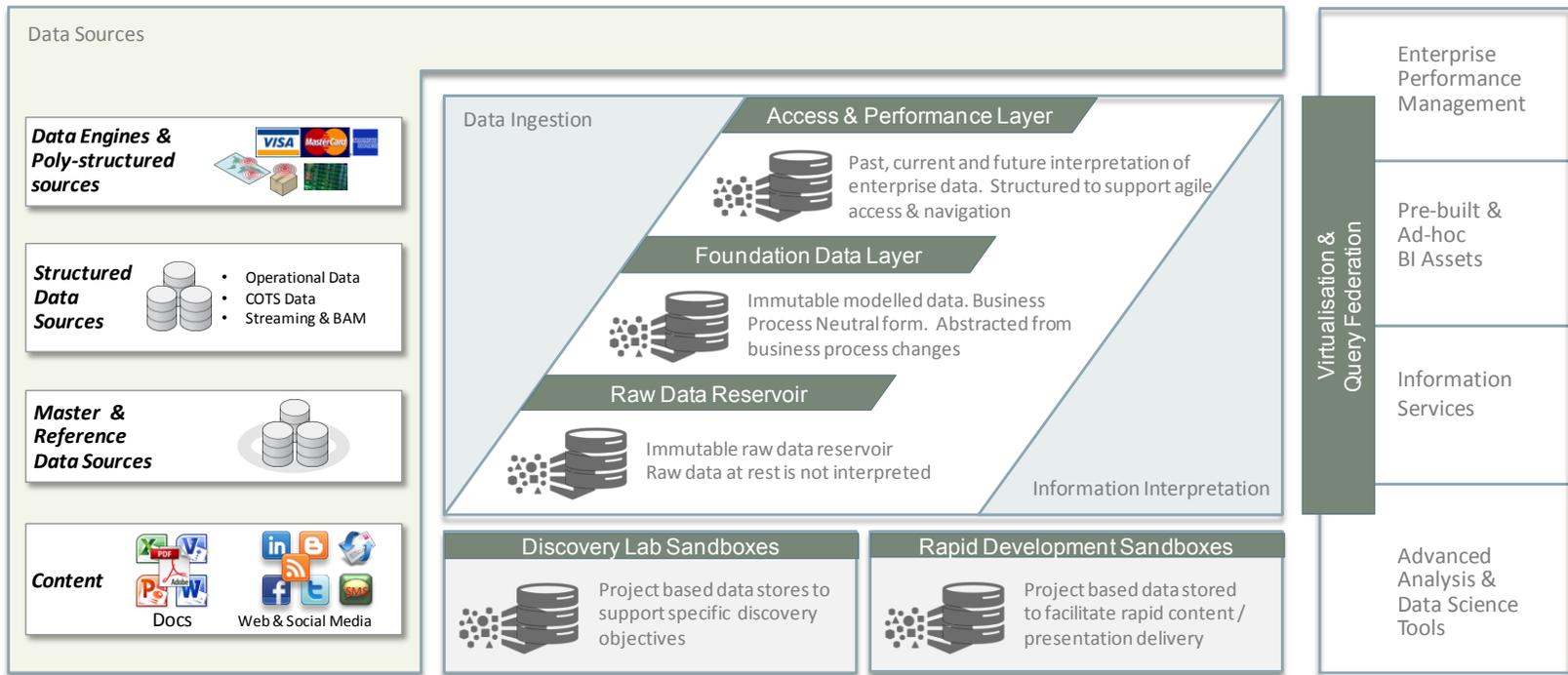


Figure 4. Main components of Oracle's Information Management Reference Architecture.

## Data Ingestion Process

Data is loaded and made available for querying through the Data Ingestion process (see Figure 5).

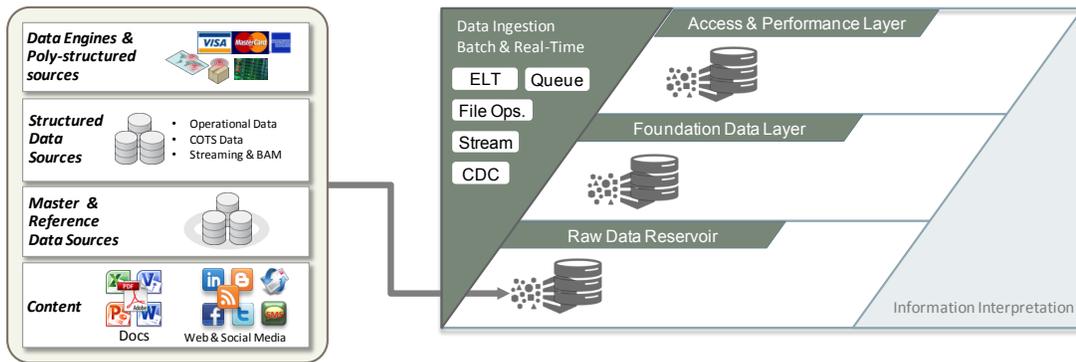


Figure 5. Data ingestion from sources.

Data will be received through a variety of mechanisms onto the data management platform (synchronously and asynchronously) and is processed by the appropriate mechanisms and methods of the data ingestion layer.

The Raw Data Reservoir provides an immutable pool of data directly reflecting the operational environment from which the data was sourced.

In the majority of cases data will be landed and then further processed into other modelled layers from the Raw Data Reservoir (remembering that the Raw Data Reservoir may comprise both relational and non-relational technologies). These flows are depicted in Figure 6. It is worth noting that the flows can be implemented either logically (views) or physically (intra-ETL) regardless of the underlying storage technology although performance and query concurrency costs must also be considered.

Different data will arrive at different rates and will be processed as appropriate. The rate at which data is received onto the IM platform and the frequency at which data is refreshed and made available for general query is driven by business needs.

In the most part data will have its own (natural) rate of flow which often dictates the approach adopted, but best practice advice is to:

- » Adopt the simplest approach possible for sourcing and transporting data. Simple is best.
- » Take all the data, not just a selection. You will only want more later. It is better to at least have the data flowing onto the IM platform even if you don't retain it.
- » Never do "big-batch" if you can "micro-batch". The micro-batch approach allows data to be dripped onto the IM platform and prepared for query access, eliminating most batch load window issues.

Due to cost and ETL batch loading pressures, many designers chose a restrictive approach to the data that was stored, retained the data for a minimum period of time or simplified the model so attributes that were subject to more frequent changes were not included. The result of these choices has typically been:

- » New reporting requirements trigger an IT development process rather than a data discovery process as additional data items need to be identified, sourced, source system changes made, ETL developed, data modelled and finally reported on.

» The possible value of any analysis is reduced either because of the paucity of data available for analysis or due to the extended timeline and costs involved. i.e. the business is discouraged from looking for opportunities to exploit data due to these factors.

Data is typically partitioned using a suitable attribute such as date range and region. This allows for a more granular management throughout the data lifecycle. For instance, indexes can be built on each partition rather than the complete dataset and partitions of data loaded and unloaded to the Foundation Data Layer from the Raw Data Reservoir and compressed differently on a per-partition basis to meet lifecycle requirements.

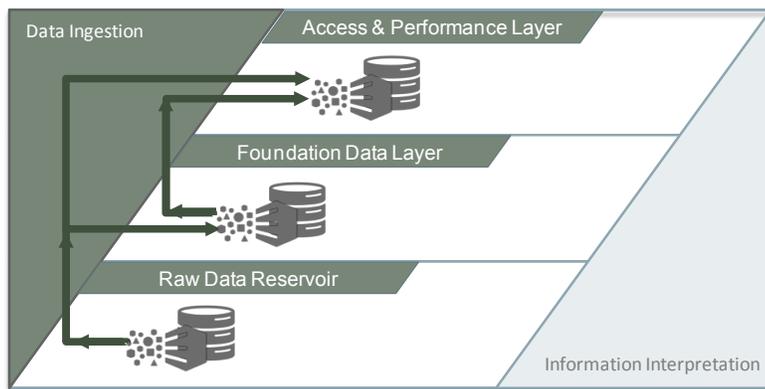


Figure 6. Logical or physical movement of data between layers.

The majority of the Access and Performance Layer will be made up of objects that will refresh automatically from data in the Foundation Data Layer. Depending on the underlying technology adopted, this could occur either when the data becomes stale or when a user queries the view, thus deferring the aggregation until sometime later. For objects requiring load scripts to run, an intra-ETL job will follow the initial ETL process to refresh Access and Performance Layer objects.

In order for business users to have confidence in the Data Warehouse and for it to serve as the basis for regulatory reporting, the quality of the data and accuracy of queries are paramount. Although Oracle guarantees it will not read or write dirty data through the multi-version read consistency mechanism for the database (which is unique in the industry), this would not be the case for non-relational source, so care must be taken to ensure any differences can be understood by users.

## Understanding schema-on-read and schema-on-write

One key advantage often cited for the Big Data approach is the flexibility afforded by the “schema-on-read” approach when compared to the more structured, formal and rigid approach imposed by a “schema-on-write” canonical data model typically employed in more traditional Data Warehouses.

It seems clear that “schema-on-read” and “schema-on-write” approaches both have a place in a contemporary Information Management platform. The differences between the approaches and the typical boundary of IT’s responsibility is illustrated in Figure 7 below.

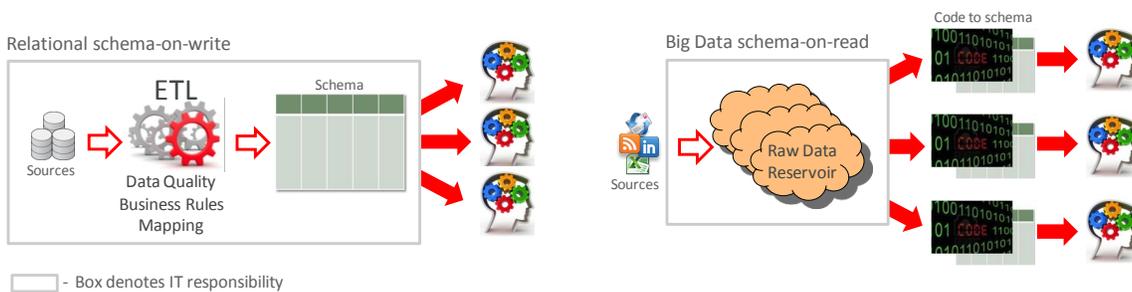


Figure 7. Differences between schema-on-write and schema-on-read approaches.

Schema-on-write describes the standard ETL approach for landing data into a relational schema. In this approach data is sourced and processed in a number of distinct steps. In most Data Warehouses data quality is validated as part of this process and data is integrated together before writing into the Foundation Data Layer schema. The ETL process used to achieve this would have gone through the usual Systems Development Lifecycle including a full suite of testing.

Hadoop and other Big Data technologies most typically use a schema-on-read approach where data is loaded onto the storage platform (typically) unmodified by the mechanism used to land the data. A number of mechanisms can then be used to interpret the data and present it in a tabular fashion. The best example of this is perhaps Map:Reduce, where the mapper will parse the data and make it available for further processing. In this model, the quality of the data output is clearly a function of the code written (or generated). Crucially, as the interpretation happens at the point in time the data is physically read, the reader must manage any data differences over time. e.g. where source data has changed over time, a small change to the Map:Reduce code could result in significant differences in results if left unchecked.

The two approaches to schema also have a bearing on processing costs and supportability. As we have already described, in schema-on-read, the data quality is dependent on the serialisation / de-serialisation code and the cost of code execution may well be repeated each time the data is accessed. Also, as the data being accessed may well be from many years ago, it may well be difficult to find people who still understand the data sufficiently to advise on any changes required to the accessing programs. See “*Understanding differences in concurrency costs and quality*” for a more detailed discussion on this topic.



It's worth noting some newer approaches such as the use of Apache Avro can abstract schema changes through versioning, but the general point is still true – there are important differences between the two different approaches which we must be aware of so we can mitigate for them.

In order for a business to have a shared understanding of the meaning of information, it follows that the measures and KPI's applied to it must be widely agreed by the business, as must the underlying qualities and characteristics of the data itself. The act of designing, agreeing and socialising the schema used to manage the data can often form part of the process of building this shared understanding. Some level of schema design is inevitable, even if new developments might only require logical rather than physical representation.

So what about the data that is not currently understood by the business? For new data that has not been fully explored or understood by the business it makes absolutely no sense to attempt to model and manage it in a formal schema before it can be used. Relational technology is extremely efficient in managing and manipulating data, but the costs of making schema changes are high and should be avoided.

Another important aspect of the schema-on-read vs schema-on-write debate relates to information quality and accountability from an Information Governance standpoint. Decisions based on poor or unreliable data will be poor decisions, so if our intention is to fully support any type of decision (operational, tactical and strategic) through our Information Management system we must pay attention to information quality.

We impose a minimum data quality standard on our data through the ETL process which transforms and loads data into our schema using a schema-on-write approach. Ideally we only load data into our Foundation Data Layer and beyond into the Access and Performance Layer that passes a minimum acceptable standard. We account for missing data, enrich and correct data where possible and constantly report on data quality and availability through a standard set of BI dashboards to improve quality over time. All this is assured through the formalised development and release procedures that are adhered to as part of the agreed development process.

For schema-on-read the picture is less clear. Here, information quality is a function of the code/tool used to access and manipulate the data. For example, if we were using Map:Reduce then it would be a function of the code implemented in the Map:Reduce job. While on the face of it this looks no different than the SQL code typically used to access schema-on-write data, in schema-on-read the code is run when the query is executed and not when the data is stored. In this way the code must continue to function although the data “signature” may in fact be changing over time. Thus information quality is really a function of how well the access code stands up over time and how coding standards are enforced.

## Information Provisioning Process

As business is continually evolving as far as its data needs are concerned, it follows that a realistic ambition is to be able to deliver the data in order to support any monitoring and decision making process, even if its not necessarily in an ideal form. The reference architecture supports this by enabling managed access to data from any data layer as illustrated in Figure 8 below.

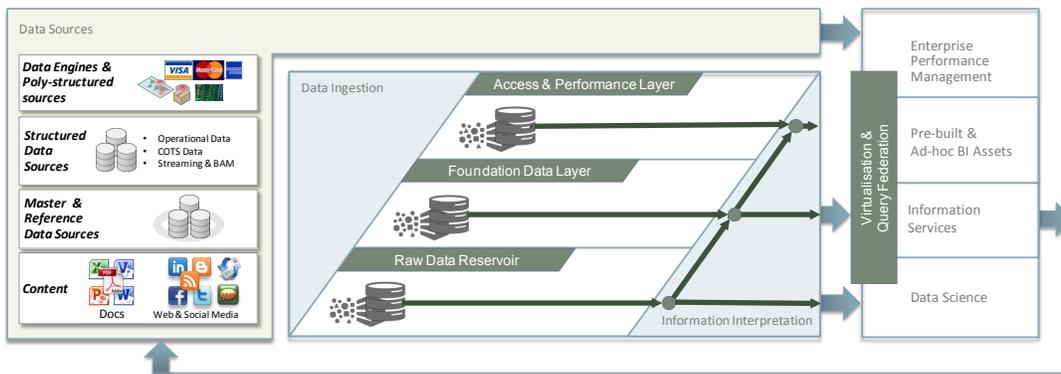


Figure 8. Outbound provisioning into information access tools and beyond.

Data can be referenced by any BI tool and includes any DW layer as well as ETL and Data Quality process metadata. This allows for a broader analytical capability to be offered, providing depth of analysis as well as width of business process coverage. That said, the majority of queries will of course be against the Access and Performance Layer, as its entire raison d'être is to simplify user access and navigation.

Enterprise Performance Management Applications are also able to query the underlying source systems directly.

The Virtualisation and Query Federation layer is able to dynamically map a single logical query to multiple sources based on metadata and make this available through protocols such as ODBC/JDBC to a variety of BI and other tools. For instance, a real-time picture of intra-day sales may be generated by joining the data in the Access and Performance Layer with data in the Raw Data Reservoir that is loaded but not yet made available in the Foundation Layer.

The additional services capability enables seamless operationalisation of the data within the organisation and to the wider trading community for solutions such as Master Data Management and technologies such as BPEL.

Advanced Analysis tools and applications such as forecasting and Data Mining may also access data directly or through the Virtualisation and Query Federation Layer. They are also able to create new data in a tightly controlled fashion through the discovery process – this is explained in more detail in Discovery Lab Sandboxes and the Data Scientist.

## Understanding differences in query concurrency costs and information quality

Data will typically be landed into our data management platform and stored in our Raw Data Reservoir at the lowest possible level of granularity and most typically in a form matching the data source. In the relational context, the Raw Data Reservoir can serve as both a data staging mechanism as well as an Operational Data Store. In the Big Data context the reservoir acts as the primary store of un-modelled data.

Data may also be directly loaded into the Foundation and Access and Performance Layers – this is often the case when an external ETL server is used or for COTS based Business Intelligence Applications.

Having been landed at a particular layer, data can be elevated to higher levels through intra-ETL processing. In doing so, although we are investing more in processing, we are at the same time increasing the quality and assurance we have in the data as the definition of the data is also formalised.

Through this process of formalisation and enrichment, typical query concurrency costs are also significantly reduced which is an important consideration for many organisations. In this way we can balance the cost to store data initially with the query concurrency costs. For example, although the costs to store a JSON file in Hadoop is extremely small you might only be able to support a small handful of concurrent queries on the data. On the other hand, we might use an intra-ETL process to model and aggregate the data instead, and in doing so you might add several orders of magnitude of query concurrency for the same relative cost.

These facets are illustrated in Figure 9 with the differences in the line length showing conceptually the relative difference between ingestion and interpretation costs. Note though that in the case of the intra-ETL, to elevate data from one layer to the next the cost is born just once and not for each query execution.

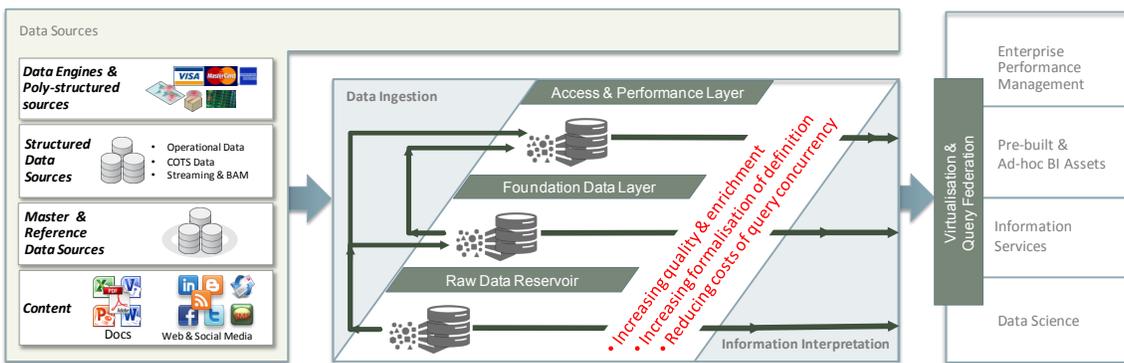


Figure 9. Illustrates the relative differences in query concurrency costs between data at different levels.

## Discovery Lab Sandboxes and the Data Scientist

What's the point in having lots of 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.

The Cross Industry Standard Process Model for Data Mining (CRISP-DM)<sup>©</sup> outlines one of the most common frameworks used for Data Mining projects in industries today. Figure 10 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 regardless of the tools applied.



Figure 10. High level CRISP-DM process model.

Figure 10 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 prepared, models built and then evaluated (both technically and commercially) before deploying 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 to gain an understanding of the data in scope. To answer the business problem further data may be required or multiple data re-coding and transformations to emphasise aspect of the data to improve model efficacy. Data discovery is almost always highly iterative.

It is often said of Data Mining that with more time you will get a better answer. Performance is an important consideration for the Discovery Lab, especially for more challenging problems and larger data sets. By iterating through the development cycle faster more model variants can be tried and a more optimal solution arrived at. Solving problems faster means your Data Science team can tackle more problems and drive further business value.

This highly iterative process is supported through the implementation of a Discovery Lab Sandbox as shown in figure 11. It shows how a mixture of relational and non-relational data from both the current data under management as well as from operational and external systems can be combined and quickly provisioned to address a new discovery opportunity. The data may be a logical view rather than a physical copy depending on implementation details, and for some problems it may be sufficient to use a sample rather than use the complete dataset in the first instance.

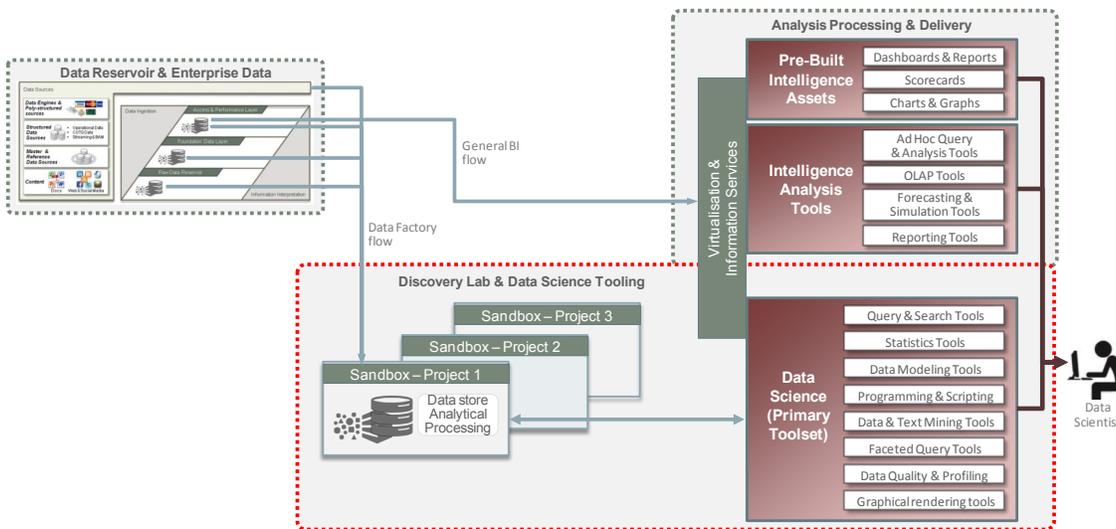


Figure 11. Logical view of Discovery Lab operation.

The Data Scientist 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, statistical or visual techniques may be applied to the data depending on the problem being addressed. 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. This write-back capability is perhaps what marks the difference between this kind of analytical discovery and more traditional BI and reporting which is a read-only function.



Information discovery is a data understanding problem and not an IT development problem, so it is important that we remove or minimise the involvement of IT from the steps in our CRISP-DM process between the definition of the business problem being addressed and the deployment to operational elements. Other important factors include:

- » Discovery Labs should function on real, ideally up to date, production data. That is, they are an offline production system and not a non-production environment. In this way we will avoid the pitfalls of redaction and masking which can inhibit value discovery.
- » The ability to rapidly provision and then de-provision at defined points in time are seen as critical to avoid the Discovery Lab environment being used to fill gaps that should ideally be addressed in the usual systems development process, or a single lab being used for everything.
- » Demand management and provisioning automation should ideally happen under the control of the business (or quasi business) and not IT. Companies who are able to focus on the right business areas, try new ideas and abandon them or quickly put them into a production context will out perform their slower or less agile peers.

The actual form the insight or 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 or expected value, 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 our Master Data Management system or operational system such as the event engine depicted in the Conceptual Architecture pictured in Figure 3. In some instances it is also possible for the model itself (rather than the data) to be deployed to these operational systems so results can be generated in real time by these applications.

Not every discovery exercise will result in directly deployable code or scores. On many occasions it may be that the Data Scientist 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.

The collection of tools and technologies used in the Discovery Lab will most likely be determined by the organisational 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.

As technologies and organisational capabilities are developing rapidly in the area of Analytics, it seems reasonable to suppose that the technology choices you make today will change over time. The recent trend for analytical tools to abstract away from the underlying storage technology of the data will also make the choice of implementation less critical.

If the future prosperity of a business is contingent on a constant stream of innovation then it follows that the organisation should focus attention on optimising the discovery process where possible. This includes the removal of mundane tasks from the Data Scientists by automating process steps such as the provisioning/de-provisioning of sandboxes and data for a new discovery challenge and the ongoing management of deployed models.

## Rapid Development Sandboxes and iterative development methods

Rapid development sandboxes provide critical support for iterative / agile development techniques to quickly develop new reporting areas or make adjustments to existing ones to continuously align to business needs.

To follow our previous example, let's imagine that our Data Scientist has progressed through a discovery exercise using a new feed of customer behaviour data derived from weblogs to build some additional customer segmentation models that have now been deployed to our operational systems and will be used in a cross/up-sell process. Our challenge is to ensure we adjust our reporting systems in time so that we are able to report on campaign effectiveness to ensure we are seeing the expected behaviour as well as trigger a refresh of the data mining models once we see its effectiveness declining.

As the requirement is for new data not currently modelled in our system, a traditional waterfall approach would require the close cooperation of a number of teams (including operations, Applications, ETL, Data Warehouse and BI) before the data can be presented to the business user. Moreover, in this kind of approach requirements are often documented in such a way as to make it difficult for users to really express their requirements effectively or check for completeness.

Even for simpler changes that do not require source systems and ETL changes, a waterfall development approach is rarely a good match to business user expectations. On the whole these users respond much more positively to physical prototypes than they do to relational data modelling and formalised report specifications.

The starting point for this kind of iterative development approach is the provisioning of a new sandbox for the development work – Figure 12 shows the general operation of a Development Sandbox.

Having provisioned the Sandbox, the next task is to identify any new data and make it available in the sandbox, either logically or physically replicated, so it can be combined with other data currently under management as required.

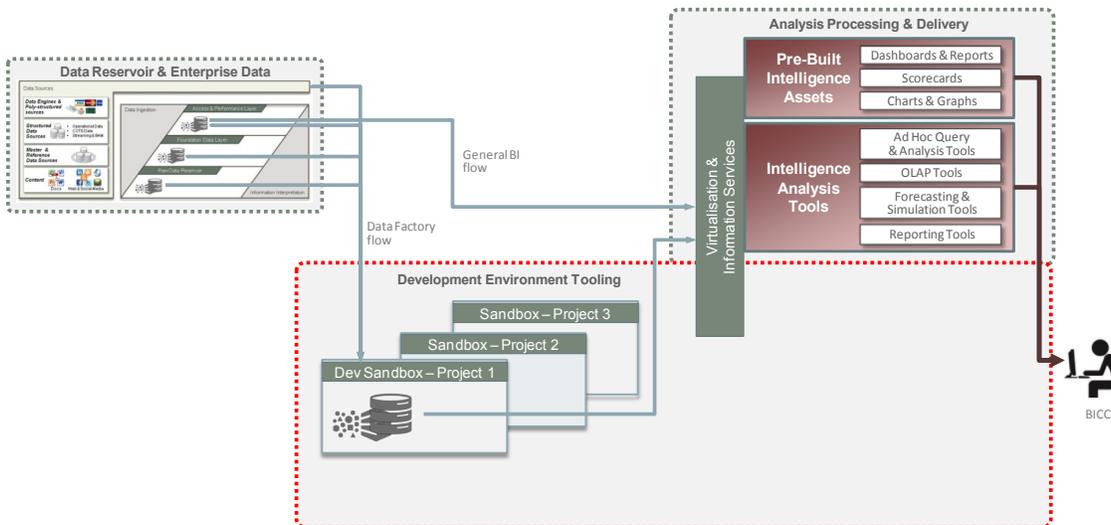


Figure 12. Support for agile development methods that require data changes.



Once the data is made available to the Business Analyst it can be mapped in the Virtualisation & Query Federation Layer so it can be made available to the user's choice of BI tools and the iterative development cycle can begin.

Through the combination of the Virtualisation & Query Federation Layer and the BI tooling the Business Analyst can confirm the look and feel of the reports, check for coverage and show the broader context to the user by combining it with existing reports.

Once functional prototyping is finished, the work to add non-functional components and professionally manage the data through the formal layers of the architecture and so put the data into a production setting must be completed. During this time, and providing there is sufficient business value, it may be possible for the user to continue to use the new reports that are based on the Sandbox data until the work to professionally manage the data has been 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.

As before with Discovery Sandboxes, it is vital that the additional activity to professionally manage the data is completed following the prototyping activity or the overall system will eventually be brought down under the weight of supporting these kind of developments. Therefore, governance of this activity is critical, as is the project interlocks between the Business Analyst and IT development community.

## Technology approach and the Big Data Management System

In the longer term it is likely that any contemporary Information Management System will include a blend of relational, Big Data and NoSQL technologies to manage the wide range of functional and non-functional requirements demanded. Experience also suggests that a minimum set of technologies and vendors is sensible if development and management teams are to manage costs and focus more on the solution than the technologies.

While the Information Management System overall may be made up of a blend of technologies, it is still important to be able to rationally select the optimal primary storage mechanism for any given project. This is especially the case for poly-structured data where a number of storage choices can be made. For examples, a file containing JSON data could be stored “as is” in the Raw Data Reservoir on Hadoop, each row could be stored in native JSON in the Oracle database or NoSQL, or the JSON data could be stripped and the data stored in a relational modelled set of tables.

It is important for the primary storage location for data is selected using a rational set of criteria. We have found the criteria shown in the spider diagram (Figure 13) to be a useful guide for any new project. For any given set of data the application criteria can be plotted against an agreed reference set for each technology (only relational and Hadoop are shown in the diagram but others such as NoSQL are easily added).

Once the primary storage mechanism is selected any identified weakness can be mitigated for through the application of specific approaches. For examples, assuming relational technology is selected for the “my Application” example shown in Figure 13, the design team must take care to ensure that the ingestion rate, ingestion simplicity and data scarcity criteria can be met appropriately

Table 1 gives a brief explanation of the criteria used in the spider diagram. These are not meant to be the last word in determining criteria, and you may have your own view of others that should be include to meet your specific requirements as well as the baseline values for each given your own implementation pattern and technology choices.

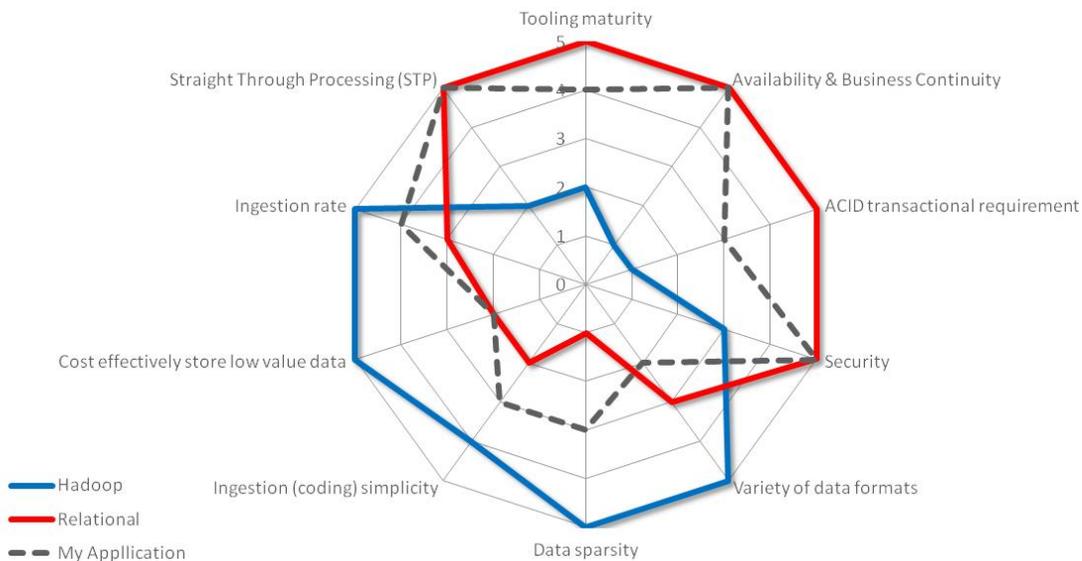


Figure 13. Primary storage location selection criteria.

**TABLE 1. DEFINITION AND CALIBRATION GUIDE FOR SELECTING THE PRIMARY STORAGE LOCATION FOR DATA**

Criteria	Description	Calibration Guide
Tooling maturity	Describes the level of maturity required to deliver the solution and maintain it over time. Low maturity will result in higher maintenance costs. Important for solutions with longer lifecycles or with broader use of additional utilities/tools.	0 = simple application requiring little use of utilities/tools or with short lifecycle. 5 = complex application that will be around of years.
Availability and Business Continuity	Describes the desired level of business continuity and availability , RPO/RTO, High Availability, etc.	0 = Batch oriented loads, no DR, no backup 5 = minutes RTO, 0 RPO, etc.
ACID transaction requirement	Does the solution require ACID compliance?	0 = Not required 5 = Full ACID compliance required (i.e. No relaxation of ACID)
Security	The security level required from a governance and risk perspective	0 = No real security restrictions 5 = Fully robust security including data at rest and separation of duties etc. Label level access.
Variety of data formats	Seeks to describe the level of variety on the data. Influences classical ETL costs.	0 = Discrete number of well understood types 5 = Highly variable range of data types and styles
Data sparsity	Describes how sparse the data might be for any given collection. In a classical relational table this describes how many columns might be completed in any given row. Indication of data model and analytical complexity.	0 = Data is very dense. 5 = infinitely large range of attributes and values within those attributes.
Ingestion (coding) simplicity	Level of complexity of ingestion process from a design point of view. Implies a high cost for ETL if tackled in a traditional relational fashion.	0 = very simple. No complexity in parsing or cleansing data. 5 = Very high level of complexity.
Cost effectively store low value data	Determines whether the emphasis of the solution is the reduction in the cost of storage or on some other aspect such as performance or analysis.	0 = No emphasis on costs 5 = it is all about costs
Ingestion rate	A measure of the loading bandwidth required.	0 = <i>Volume not considered to be challenging or relevant.</i> 5 = <i>Volume is the problem!</i>
Straight through processing (STP)	A measurement of the end-to-end latency from the data being available for loading to the time at which it must be made available to downstream systems in query.	0 = <i>Not a concern for the system.</i> 5 = <i>Real-Time STP</i>

## Big Data Adoption

Adoption in the context of Big Data can be a challenge – just how do you introduce disruptive technologies into a highly structured and regimented organisation without losing all the benefits you were seeking in the first place.

This white paper has outlined many of the core architectural principles and practices we that contribute to a successful Information Management platform that combines the technology benefits of Big Data and relational technologies, but other issues relating to technical competencies and current organisational standards are also important to address.

How do you select the right technology components and build a skilled team to develop, operate, govern and support Big Data technologies without going through painful mistakes? Big Data technologies are evolving rapidly so leveraging the right components is key to minimise the risk of investing in areas that subsequently fall out of favour. The same is true of how you decide to physically build the infrastructure and combine the Big Data technologies with your existing Information Management components – these are big decisions that will most likely have significant impact on future capabilities.

How can you push through a new approach when the organisation and standards developed to reduce spiralling IT costs makes adding new tools and vendors a significant challenge? The way you currently fund projects, impose development standards, procure hardware and outsource operations will all have an impact on adoption.

Evidence from our customers has shown two distinctly different adoption paths. Broadly speaking we would suggest that one is IT lead and one more business lead. As well as the difference in sponsorship, the major difference lies in the scope of the initial projects, with IT lead adoption choosing a much narrower scope as a way of learning about the technologies and forcing through required changes to the organisational processes previously highlighted. Business led adoption has been strategic in nature, more far reaching in scope and impact and driven by business strategy. These projects are better able to sweep aside organisational resistance because of executive sponsorship and don't have the same need to prove ROI at each phase – meeting the business need is sufficient evidence of value.

Somewhat related to the technology adoption approach and project sponsorship is the scope and phasing of the initial projects as the points of integration and additional capabilities will have an impact on other systems.

From our work we have identified a range of implementation patterns (see Figure 14), each of which has a different scope and focus. These include:

1. **Discovery Lab.** Focus on rolling out discovery capabilities only. No broader integration with operationalising insight.
2. **Re-conceptualise IM.** Broader implementation that adds Big Data capabilities, especially in enabling the Discovery Lab capability, and re-visits the overall design of the IM solution.
3. **Big Data Application.** Adoption of Big Data technologies to address a specific application. For example, a pharmaceuticals company who now store and process genome data using Hadoop. The data is only required for research so there is no broader integration with the rest of the company's enterprise data.
4. **Big Data Technology Pilot.** Typically focused on adding Big Data specific tooling to an existing IM estate. For example, adding "messy" data that was previously unmanaged to the existing enterprise data store.
5. **Operationalise Insights.** Here the focus is more on operationalising insights, often using NoSQL and next-best-action tooling. This kind of project often follows the successful roll-out of additional Discovery Lab capabilities.



Adoption in the wider context is also important to consider. We have long advocated the need for Information Management to be a first class system in IT – as we have discussed, 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 operational application. In addition, as applications themselves become more operationally linked to IM systems, the characteristics of those systems must now also be matched.

It follows that if an Information Management 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 clearly defined abstraction layers that limit the impact of change we call out in particular the role of the Virtualisation & Query Federation in addition to the Discovery Sandboxes to support iterative development and discovery.

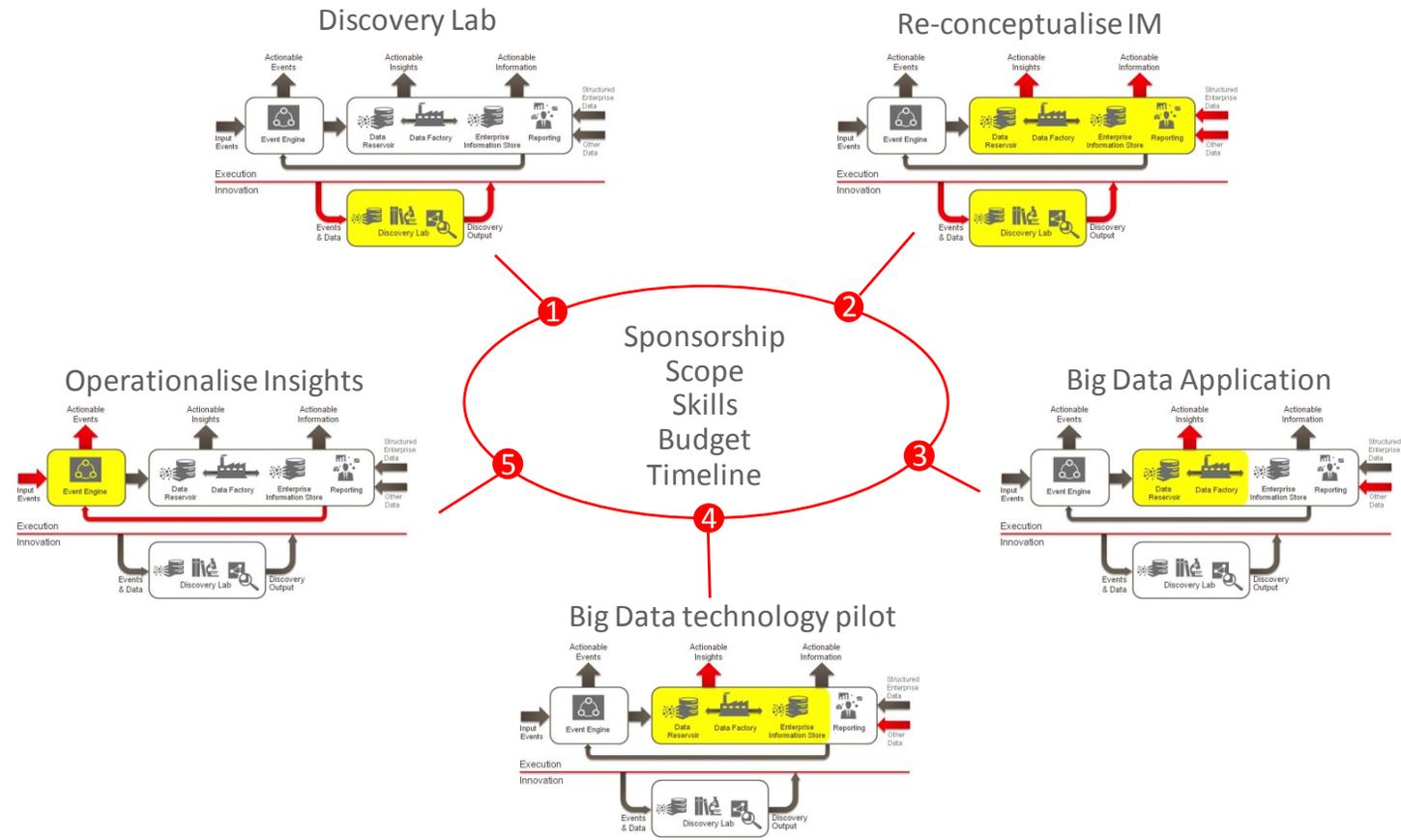


Figure 14. Big Data adoption design patterns.



## 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 technology is just that – a technology. And just like any technology it's not the technology itself that's as important as the solutions they can enable and business value that can be derived.

Big Data technologies, when combined with relational technologies and others such as spatial, BI, OLAP and many more, offer the promise of unlocking the untapped potential that exists within the broader set of data that has hitherto been too difficult or too expensive to manage.

Big Data also opens up a range of new design and implementation patterns that can make Information Management solutions less brittle, speed development, reduce costs and generally improve business delivery. When designed appropriately, they can combine these benefits without giving up the things the business has come to expect and value such as good governance, data quality and robustness.

Modern businesses are evolving and are constantly demanding more from their Information Management systems. No longer satisfied with standardised reporting by a limited set of users, modern businesses manage by fact, demanding faster and more pervasive access to information on which to base critical business decisions. This change to the volume, velocity and reach of the information is in turn forcing changes to the solution architecture and technology that underpins the solutions.

This white paper has sought to outline a practical reference architecture that can enable the delivery of Information Management pervasively in such a manner. The architecture strikes a balance between the data management and information access requirements of data in a single design concept, to ensure business value can be delivered in a sustainable fashion over time, without major re-engineering or loss of service while data is being re-engineered, regardless of the underlying storage model: Big Data or Relational.

The Reference Architecture is a useful device, which can be used as both a design template for new Information Management solution designs, as well as a 'measuring stick' from which you can assess an existing implementation and upon which roadmap options can also be developed.

The basic principles of the Reference Architecture are useful regardless of the precise technologies deployed to deliver it. However, Oracle Corporation is uniquely able to deliver integrated components from the disk to the dashboard of the reference architecture. Please contact your local Oracle account team for more information about how Oracle technology can be used to help you deliver your next generation of Information Management systems.



## Finding out more about Oracle's Information Management Reference Architecture

You may like to take a look at a technology briefing note on “Information Architecture for Big Data Management Systems” which reviews a number of aspects of the conceptual architecture and design patterns also outlined in this white paper. You can find it [here](#) or search for the title using your favourite search engine.

We have also recorded a number of short vignettes describing important 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.’ Some of the videos present a previous version of the Reference Architecture but the underlying principles remain unchanged.

We collaborated closely with Rittman Mead Consulting on the development of this white paper. You can find a number of useful blog posts describing specific aspects of Big Data technology implementations on their web site [here](#).

Oracle also runs a series of regular ‘Master Classes’ describing the IM Reference Architecture. These 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.

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

**Worldwide Inquiries**  
Phone: +1.650.506.7000  
Fax: +1.650.506.7200

CONNECT WITH US

-  [blogs.oracle.com/oracle](http://blogs.oracle.com/oracle)
-  [facebook.com/oracle](http://facebook.com/oracle)
-  [twitter.com/oracle](http://twitter.com/oracle)
-  [oracle.com](http://oracle.com)

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

Copyright © 2014, 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. 0914