



extended Datavault Framework

Briefing Paper

Version 1.1

©Business Thinking Limited 2019

Contents

[What is the problem?](#)

[Why modernise your data analytics service?](#)

[The extended Datavault Framework](#)

[The result](#)

[The Datavault offering](#)

[Contact us](#)



**“It is not the beauty of a building
you should look at; its the
construction of the foundation that
will stand the test of time.”**

David Allan Coe

What is the problem?

Analytics today is all about speed of change and increased scope of application. Analytics has grown to become critical to your business operations. Your business needs to keep up with the competition and your business managers have been asking for new features for some time.

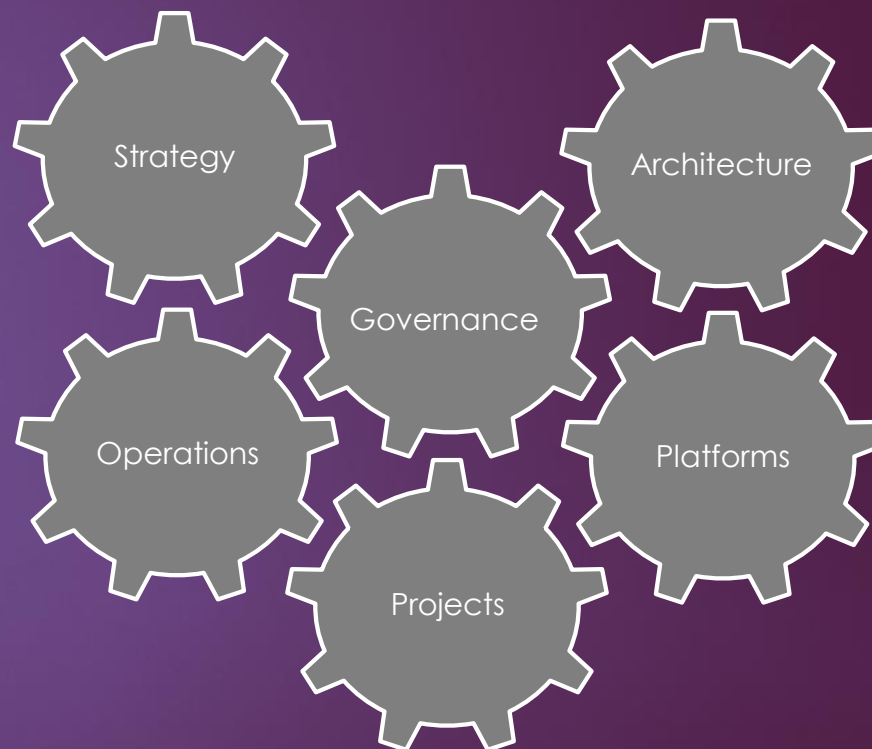
You may have concluded that you need to modernise your data service, it is no longer sufficient to keep patching legacy systems.

You need to get started on change as soon as possible but there are a number of barriers holding you back:

- ▶ An inflexible legacy analytics architecture
- ▶ Lack of capability to use analytics
- ▶ Poor quality data

There is a lot to do and it is difficult to prepare and construct a comprehensive plan so that work can get started as soon as possible.

Make your analytics work better



Why Modernise your Data Analytics Service?

You have a traditional Data Warehouse(s) that is failing to keep up with changing business requirements. We have identified five areas where you may need to target improvements:

1. Responsiveness to Business Needs

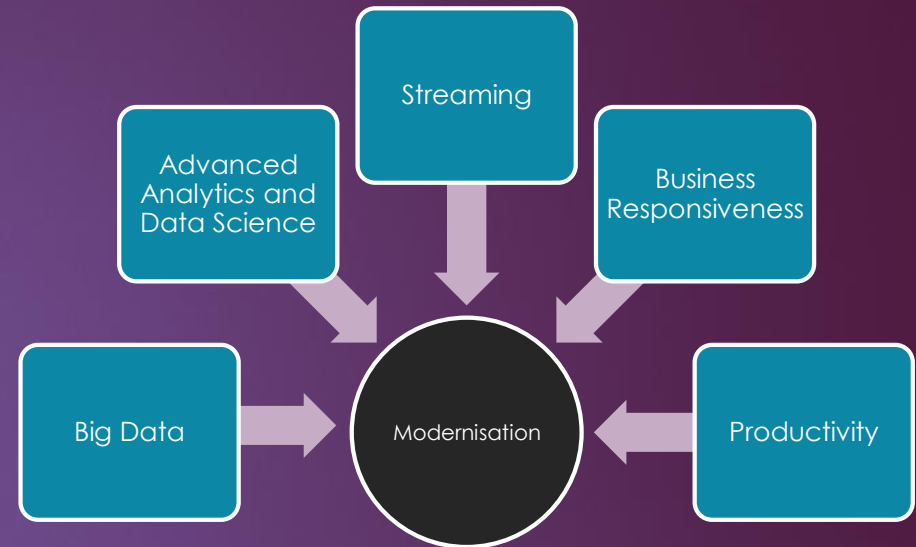
The development process can take weeks or months which makes the service unresponsive to user needs. This is made worse because the business itself is under ever-increasing pressure to move faster.

2. Advanced Analytics and Data Science

Your business wants to use the latest techniques for analysing data such as predictive modelling, machine learning and artificial intelligence. Unfortunately your existing Data Warehouse isn't engineered to give your data scientists what they need.

3. Productivity

Your existing Data Warehouse uses older technology that requires a large team to run the service. There is no easy way to improve the productivity of the team with the existing technology stack.



4. Big Data

The scale required by massive new data sets from web data to IoT applications, drives different infrastructure requirements and techniques especially if the data is unstructured.

5. Streaming and Real Time Engines

Increasingly, data-driven organisations need real time access to their data to enable features such as real-time transactional responses in an operational environment.

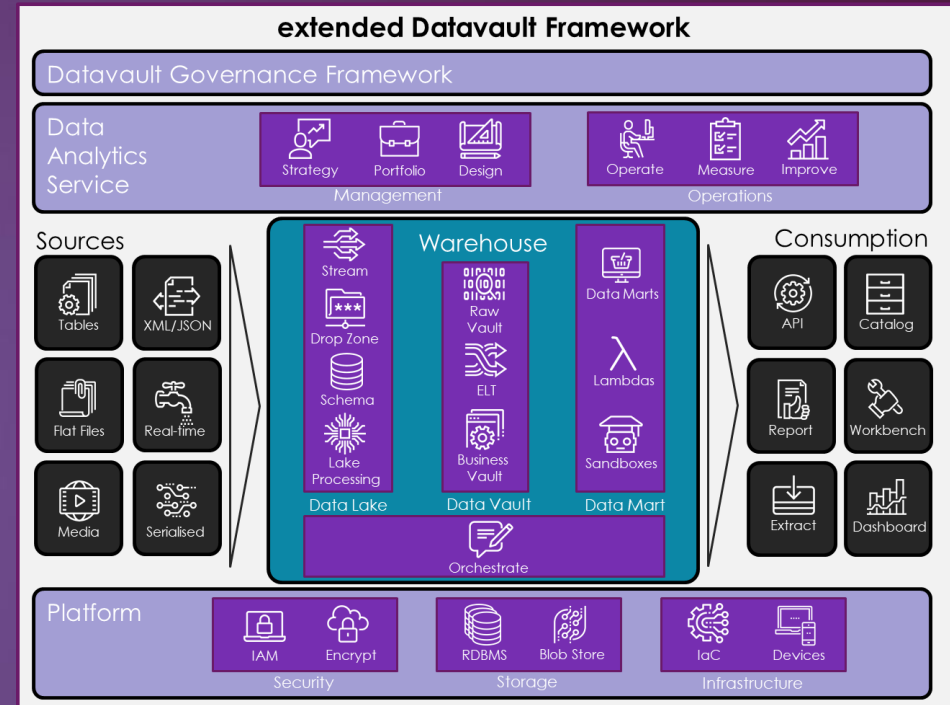
extended Datavault Framework

Our extended Datavault Framework provides a reference model against which you can work out the changes needed to deliver a modernised data analytics service.

The framework uses the Data Vault 2.0 method for agile Data Warehousing, and extends it to include management, dev-ops and information governance.

The framework:

- ▶ Helps you plan your modernisation programme
- ▶ Expresses, in a clear way what, at a lower level of detail could be a difficult and diverse change initiative
- ▶ Covers all the components needed to deliver the change including new and emerging use cases.

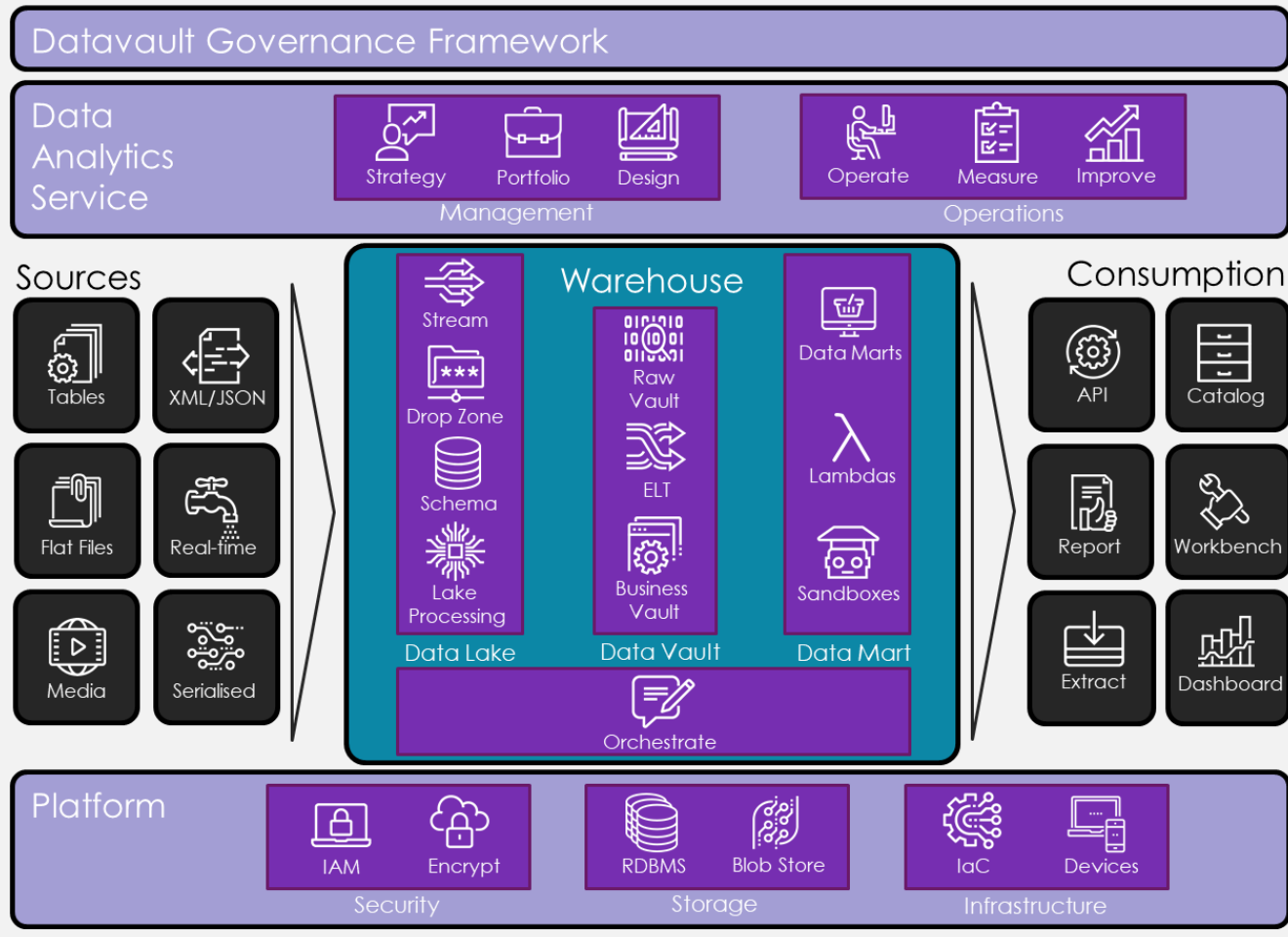




extended Datavault Framework

Our extended Datavault Framework

extended Datavault Framework



The extended Datavault Framework was devised to help organisations to deliver a modern, agile, data analytics service.

The Framework combines real-world client experiences to deliver:

- Strategic alignment and a business case for Business Intelligence and Analytics
- A Data Vault 2.0 agile data warehouse with a data lake
- Information Governance to address underlying data management issues
- The latest tools and trends in cloud-based Data Warehousing, Big Data and analytics

Data Analytics Service



“Knowledge has become the key economic resource and the dominant, if not the only, source of competitive advantage.”

Peter F. Drucker
Management Guru

Data Analytics Service

Data Analytics Service



Strategy



Portfolio



Design

Management



Operate



Measure



Improve

Operations

A modern Business Intelligence team offers a Data Analytics Service, itself made up of a set of lower-level services. The Data Analytics Service should deliver operational excellence through applying good practice and will participate in the business change needed to modernise itself. Management and operations consists of six activities.



Manage user expectations by defining a Data Service Catalogue and delivering change through incremental improvements to service levels rather than by communicating the technology changes that are happening underneath due to modernisation.

Service Management

Data Analytics Service



Strategy



Portfolio



Design

Management



Operate



Measure



Improve

Operations



Strategy

Strategy means having an agreed vision, implementation plan, and blueprint looking out 3-5 years into the future. It is important to align the vision with business value.



Portfolio

A Service Portfolio is a list of data services offered to the business. Managing the Portfolio involves adding new services, improving others and withdrawing underused services based on statistics about usage, quality, cost, etc.



Service Design

Design how services operate so they deliver reliably within constraints (target cost, SLAs). Automate design documentation to keep it current to meet regulatory needs.

Service Operations

Data Analytics Service



Strategy



Portfolio



Design

Management



Operate



Measure



Improve

Operations



Operate

The Data Analytics Service must fit into the business management cycle: planning and securing a budget; managing the team; managing operational delivery within SLAs; reporting and reviewing performance; risk management; service continuity planning; issue tracking and resolution.



Measure

There needs to be a mechanism to measure the Data Analytics Service performance. Assemble measures in the data warehouse and use dashboards to visualise performance.



Improve

Business as usual includes continuous improvement. The team needs project management capability to deliver change.

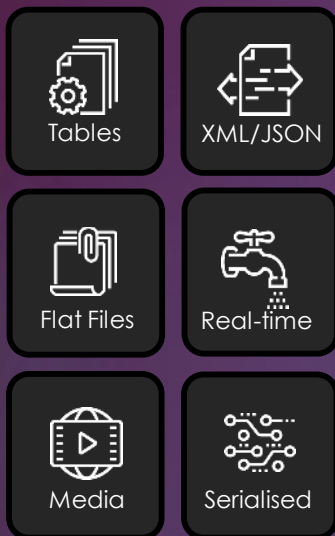
Data Sources



“The amount of data we produce every day is truly mind-boggling. There are 2.5 quintillion bytes of data created each day at our current pace, but that pace is only accelerating with the growth of the Internet of Things (IoT). Over the last two years alone 90 percent of the data in the world was generated.”

Forbes, May 2018

Data Sources




A modernised Data Service needs to ingest and process increasing volumes and variety of data. Depending on circumstances, much of that data may need to be processed in real-time. It also needs to serve up data in a variety of ways to meet an increasing number of use cases. The only way to achieve this is by loading raw data and deferring the application of business rules to the point where data is extracted.



The design aim is to load raw data into the Data Lake and Vault with minimal processing. The Data Vault 2.0 system will standardise input data, such as flattening XML structures or deriving deltas to facilitate automation of ETL.

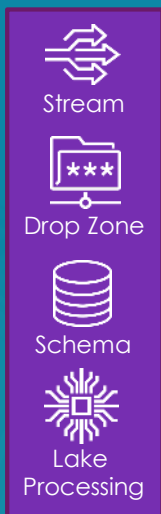
Data Lake



**“The price of light is
less than the cost of
darkness.”**

Arthur C. Nielson

Data Lake



A data lake in essence contains data organised in a file system and performs the role of a persistent data store. Some users may find value from accessing individual data files for simple operational reporting and analytics. It is also a source for the downstream Data Warehouse where more sophisticated, value adding processing is applied.



In the end a Data Lake is a file based system and needs to be organised into directories and sub-directories that can hold data from potentially hundreds of systems and store millions of files such that an individual file can be found easily. There are common strategies for achieving this but ensure that this aspect of the design is addressed up-front because changing it later can be a pain.

Data Lake



Streaming

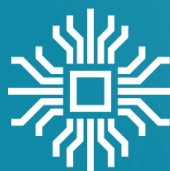
The Lake needs to accept and store individual records, or sets of records, sent as a stream of events with data attached. It also needs to be capable of accepting streamed media, such as audio or video as a continuous message, accumulated and chunked into files by the streaming service.



Drop Zone

Even with streaming of data the need to process batch files does not go away. A drop zone provides a secure location for sources to deliver files to the data warehouse. Files delivered here are detected and moved as quickly as possible into an internal, secure area.

Data Lake



Schema

Data should be stored with its metadata such as details of schemas, search tags, and classifications. (Avro is a good format to hold metadata for some data sets).

Data Lakes have several uses as a persistent data store, as a system of record, as a feed to the data vault, as a raw data set for access by data scientists, and as an operational data store.

Lake Processing

We need to get data into the data lake, and we need to perform housekeeping on the lake data sets. Lake processing validates source data, calculates diffs, formats it, adds tags and saves it into the lake. It may also concatenate small files (to make processing efficient) and apply data management rules such as retention periods, or mask sensitive data.

Data Vault



**“Information is the
oil of the 21st
century, and
analytics is the
combustion
engine.”**

Peter Sondergaard
Gartner

Data Vault



A data vault adds value to the data held in the data lake – by constructing a point in time, integrated view of data assembled from different tables of data taken from different systems. It should only process a subset of data from the data lake where the vault can add value through integration and deriving business rules.



Analytics tends to yield the greatest benefit delivering valuable and surprising insights when combining data from multiple sources. Data Vault assembles an integrated data set while retaining the raw nature of the data, performing the hard work needed to prepare to data for downstream processing. This avoids each consumer having to work out the integration themselves which can result in data sets disagreeing with each other.

Data Vault



Raw Vault

The Raw Data Vault is a normalised, point in time, logical view of the business. It is made up of standard building blocks (hubs, links, satellites, etc.) and it integrates raw data around entities of interest to the business, assembling data from different systems into a common view.

ELT Automation

The Data Vault is built from a small number of patterns. Each pattern requires similar code to perform its data load process. This means we only need to code each pattern once. By substituting metadata we can generate all the SQL statements we need to perform the load. Data Vault projects become more about data modelling and metadata management and less about coding ELT.



Business Vault

Business rules derive new data from data held in the raw data vault. The business rule engine should be thought of as another data warehouse source system that happens to draw its data from the data warehouse. Each rule should be coded separately, and resultant data fed to the front end of the data warehouse for storage in the data lake.



Data Marts

A close-up photograph of a hand painting a traditional Indian deity, likely Lord Venkateswara, on a ceramic pot. The deity is depicted in a standing pose, adorned with elaborate jewelry and a crown, holding a mace. The background of the painting is a dark, circular frame. The pot itself is covered in intricate, light-colored floral patterns. The hand is using a fine brush to add details to the deity's attire.

**“Without big data
analytics,
companies are
blind and deaf”**

Geoffrey Moore
Author

Data Marts



Data Marts



Lambdas



Sandboxes



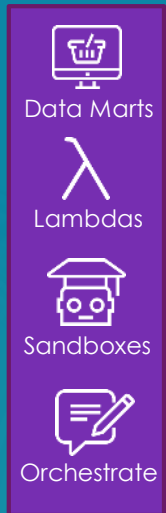
Orchestrate

A Data Mart contains a subset of the Data Vault data dedicated for a particular end-user use case. Marts are pre-formatted and pre-calculated for ease of use and ensure users can only access data that they are allowed to access. In a Data Vault solution there may be dozens of Marts, and they are increasingly virtualised as views on the Vault making them extremely agile.



Mart virtualisation makes it possible to develop many small dedicated Marts which gives you fine control to meet user needs. However it is important to plan these in detail to ensure that derived values are properly versioned to ensure synchronisation between Marts.

Data Marts



Data Mart

The Data Mart is a reporting data set, taken as a subset of data from the data vault and used to support a defined set of users with specific reporting and analysis needs. This can be structured as a star schema , cube, or other reporting-friendly format.



Lambda

A lambda is a cloud function that runs without the need to set up or specify a server platform. All details are managed by the cloud provider – all we need to is define the function and how it is triggered. For example, a lambda could be used to send data to a location on demand, to load a spreadsheet to the data lake when a file is detected in the drop zone, or to perform real-time calculations on a data stream to detect if the business needs to respond to a detected business event.

Data Marts



Sandbox

A sandbox is a separated, prepared data set generally used by data scientists for their experimental work. For example, a researcher may want to analyse customer buying behaviour and want to work with a stratified sample of masked customer and purchase data. This data is extracted from the data lake or data vault and placed in a private store for the exclusive use of the data scientist. If such a data set is useful and requested many times then it could become a routine service offered through a data mart.



Orchestration

The Data Vault architecture requires multiple data pipelines to work as a coordinated set. Some are triggered by an external event others are scheduled. The orchestration component ensures that this co-ordination happens as well as recovering from errors when they occur.


Orchestration



Orchestration

The Data Vault architecture requires multiple data pipelines to work as a coordinated set. Some are triggered by an external event others are scheduled. The orchestration component ensures that this co-ordination happens as well as recovering from errors when they occur.

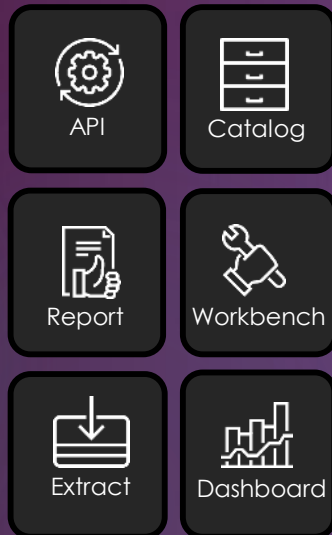
Consumption



**“You can have data
without information,
but you cannot
have information
without data.”**

Daniel Keys Moran
Author

Consumption



Data curated and presented in Marts can be consumed using a variety of mechanisms. These include the use of analytics dashboards, data feeds, reports and programmatic interfaces.



Legacy systems focus heavily on push-based enterprise reporting. With the move towards self-service BI, users will come to get data when they need it which has an impact on the pattern of use spreading the demand across the day perhaps with a peak in the first hour of business as managers download their datasets.

Consumption



API

Application Programmable Interfaces serve data to other applications. It may execute a lambda, or it may reply with one or more data records. Cloud providers offer excellent mechanisms to implement an API service layer.



Report

A report is primarily a static set of data, formatted for consumption by business management or for issue to an external party. Reports may be printed, emailed, converted to pdf, or downloaded into Excel.



Extract

An extract is a simple data set taken from the lake or mart formatted as a csv file or xml/json file. Extracts may be stratified samples, archives, or feeds to other data processing steps.

Consumption



Catalog

A Catalog contains a glossary of business terms and business rules (based on the governance glossary), a list of data sets available to the business, a link between terms and their related data sets and a search function.



Workbench

A workbench is a data manipulation and analysis tool used to work with a data set extracted from the Lake or Mart.



Dashboard

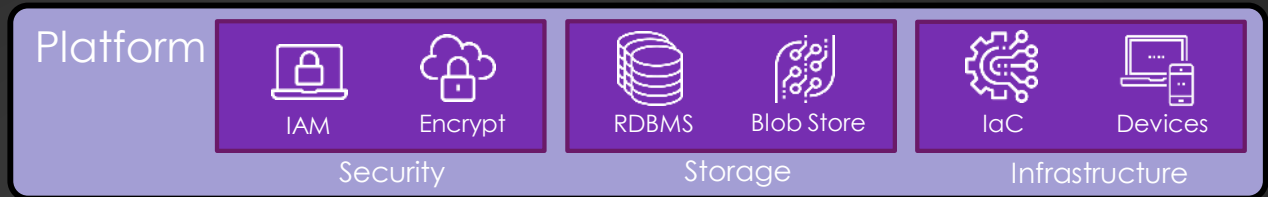
An interactive visualisation of data, usually from the mart, but also from the lake. This may draw graphs, trend lines, bar charts, or other forms of data visualisation to help an end user analyse and digest a date set.

Platform

**“There’s no way that
company exists in a
year.”**

Tom Siebel, Founder of Siebel
speaking about Salesforce.com

Platform



Cloud providers enable virtual data centres to be generated on-demand. With infrastructure as code languages such as Terraform it is possible to specify the data centre we want, run the code and have a working data centre built in minutes. Data centres can be created, tested, destroyed and rebuilt repeatedly and no longer are a bottleneck in warehouse development.



Hashicorp offer several tools to allow support infrastructure as code including Packer and Terraform. There is even a Golang test suite you can use to test your Terraform code and prove that your data centre is working correctly.

Platform Security

Platform



IAM



Encrypt



RDBMS



Blob Store



IaC



Devices

Security

Storage

Infrastructure



Identity and Access Management

Identity and access management (IAM) is a key aspect of data security. The data centre should be build on a 'trust no one' approach. All components should be given roles with policies applied to allow them to access just the applications and services necessary. The use of IaC (infrastructure as code) will ensure it is effectively managed and maintained.



Encryption

Encryption is used to ensure that data is kept in an encrypted form while still being accessible. Cloud services facilitate this by configuration settings that are generally invisible to developers. These powerful features offer protection to data at rest and access with little or no impact on development.

Platform Storage

Platform



IAM



Encrypt



RDBMS



Blob Store



IaC



Devices

Security

Storage

Infrastructure



RDBMS

Cloud services offer administration free versions of common databases. This means that the cloud provider manages the database administration, including security patches, optimising storage, automatic back-up and recovery with data spread across multiple availability zones. Many options provide on-demand scaling and the ability to pause the database when not in use.



Blob Store

A data lake needs to store large numbers of files and a blob store is the cheapest way to hold them. Formats are available that support Map Reduce and Apache Hive, for example, can treat file systems as a relational database allowing SQL queries to be run.

Platform Infrastructure

Platform



IAM



Encrypt



RDBMS



Blob Store



IaC



Devices

Security

Storage

Infrastructure



Infrastructure as Code

Infrastructure as code (IaC) allows the entire data centre to be constructed automatically from code (we use Terraform and Packer) and ensure the code is thoroughly tested with the code used to build it. In this way an entire data centre can be staged from the start of the project in a matter of minutes and the data centre can be parameterised to grow as needed.



Devices

Provision needs to be made for an IoT type architecture where removable devices feed in data in real time or small batches. Real time requires the ability to stream from remote devices – cloud providers provide excellent support for data flows where devices are able to support the relevant code.

Governance



“Quality data does NOT guarantee quality information, but quality information is impossible without quality data”

Peter Benson

Project Leader of ISO 8000

Datavault Governance Framework

Datavault Governance Framework

Management



Value



Policy



Structure



Projects

Delivery



Inventory



Lifecycle



Quality



Security

Disciplines



Audit



Metadata

Effective Data Warehouses require Data Governance to:

- ▶ Improve the quality of data pipelines
- ▶ Provide clear definitions of business terms
- ▶ Provide a framework to organise and categorise data at scale.

In turn, Data Governance benefits from having a Data Warehouse that can:

- ▶ Act as a long-term, invariant system of record for point in time data independent of ever-changing source systems
- ▶ Provide metadata for dashboarding data quality and lifecycle measures
- ▶ Provide evidence of Data Governance benefits.

We have published a separate Briefing Paper that sets out our Datavault Governance Framework. This complements and dovetails into the extended Datavault Framework.

More details are available at:

www.data-vault.com/stop-wasting-data

THE RESULT:

The extended Datavault Framework enables business problems to be solved better, faster and cheaper than traditional techniques.

Datavault Offering

Datavault is a specialist consultancy dedicated to helping our customers unlock the value in their data.

We can help you and support your Data Vault 2.0 journey every step of the way.

Unlike the big consultancies, we supply a small team, use automation and our agile Data Analytics Service method to deliver a new data service without unnecessary drama, cost or risk.

We work alongside, coach and develop your team so they can deliver the new Data Analytics Service and we are around to help you through the process for as long as you need us.

We are also passionate about information governance and will help ensure that there is a sound governance wrapper around your data.

Analytics for decision making

Sensible **Business Cases**

Integration for end to end solutions

Efficiency from **Automation**

Cloud for cost effectiveness

Expert legacy **Migration**

Machine Learning for new insight

Fast deliverables with **Agile**

Governance for control of your data



Talk to us about implementing Data Vault 2.0

www.data-vault.co.uk

 enquiries@Data-vault.co.uk

 +44 (0)23 9263 7171

