

Database Answers



Ruxley Towers, UK

DBA Manual of Good Practice

Barry Williams
barryw@databaseanswers.org

1.Management Summary.....	2
2.Analytics Services	5
3.Cloud Services.....	14
4.Data Services.....	17
5. Information Catalogue.....	24
6. Canonical Data Model.....	25
7.Data Sources.....	29
8. Master Data Management	30
9. Big Data (Lake).....	31
10. Data Mining.....	38
11. Data Modelling Theory.....	39
12. Data Vault	40
13. Data Integration Products.....	40
14. Microsoft Links.....	46
15.Proof-of-Concept – Airport Management.....	50
16.Conclusion	50

1.Management Summary

In this document we present an overview of the Services that we offer – Analytics, Cloud, Data Services and Data Feeds and how they work together to offer an A-to-Z Best Practice Approach to Enterprise Data Management.

This document is full of diagrams and can be read in less than 10 minutes.

1.1 Overview

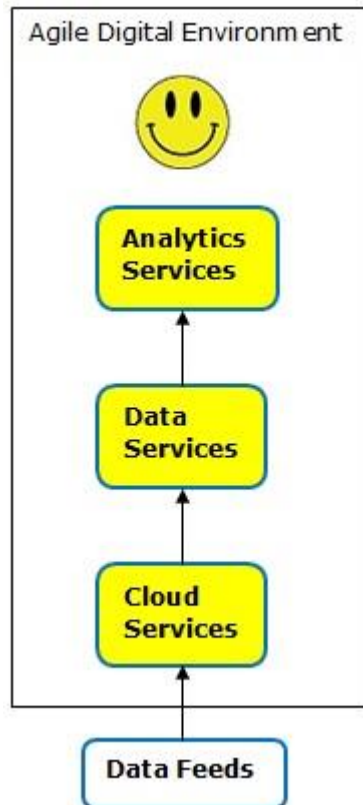
Our DBA Services provide a bridge to Digital Transformation and can integrate operational data such as a Landing Data Feed at an Airport.

This diagram shows how our Services are related – Cloud Services generates Data that is loaded into a Data Warehouse and Data Marts from where it is processed by Analytics to produce Reports, KPIs and so on.

Data Feeds load data from external sources, such as Airport Flight Arrivals.

This diagram is shown on our Database Answers Home Page :-

- <http://www.databaseanswers.org/index.htm>



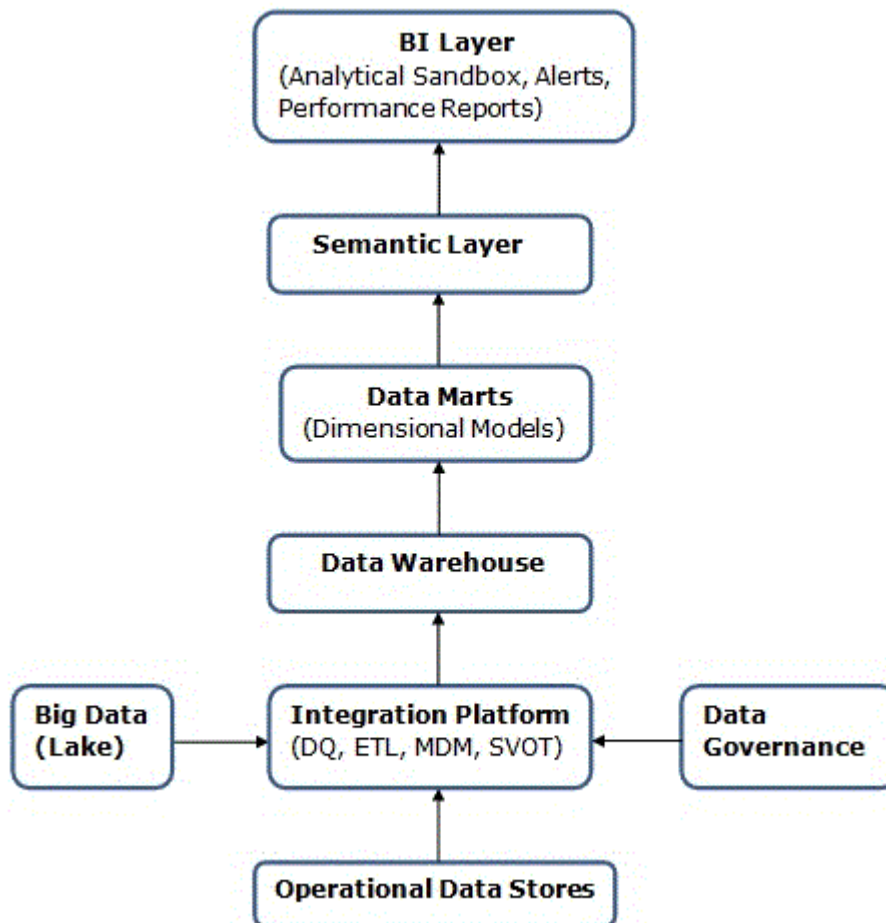
1.2 Benefits of this Approach

The benefits are that a common design Approach is used to support a wide variety of Industry specific Azure Services application.

This will appeal to Microsoft and Microsoft Partners because time spent becoming familiar with the Approach will support marketing to a wide range of end-users.

1.3 Reference Data Architecture

We use the Architecture to guide our thinking our thinking and help us with common points of reference.



1.4 Proof-of-Concept Projects ('POC')

I have produced Specifications for 21 of my 1,500+ Data Models that I have chosen to provide a Proof-of-Concept for our Canonical Data Model and Industry-Platforms.

Our target completion date is October 6th. (Barry's birthday !!!) and we have shown the first 3 priorities.

These links are clickable if you want to check the details of any particular Data Model.

POC Services Platforms 2020

1. [Air Transport](#)
2. [Assets](#)
3. [Banking - Investment](#)
4. [Banking - Self-Service \(Retail\)](#)
5. [Doctors and Patients](#)
6. [Dog Whisperer TV Show](#)
7. [Gym Training Diary](#)
8. [HR Self-Service](#)
9. [Insurance Self-Service](#)
10. [Law Enforcement](#)
11. [Local Government](#)
12. [Logistics](#)
13. [Olympic Games](#)
14. [Pharmaceutical Companies](#)
15. [Phone Bills](#)
16. [Restaurant Guides](#)
17. [Retail Sales](#)
18. [School Management Systems](#)
19. [Student Self-Service Registration](#)
20. [Telecomms](#)
21. [UN Global Compact Platform](#)

2. Analytics Services

This Section include a BI Layer, Key Performance Indicators (KPIs) , Performance Reports and Analytics.

2.1 BI Layer

2.1.1 What is it ?

The BI Layer sits between the Data Marts or Data Warehouse and the Presentation Layer.

It has become more important with the growth of big data and now incorporates aspects of compliance with data governance.

2.1.2 Why is it important ?

It is important because the need for more functionality has increased with Big Data analytics.

2.1.3 What will I learn ?

You will learn how to identify the requirements related to the user interface and analytical functionality.

2.1.4 Best Practice

Best Practice involves articulating user requirements and interactions in terms of user data structures.

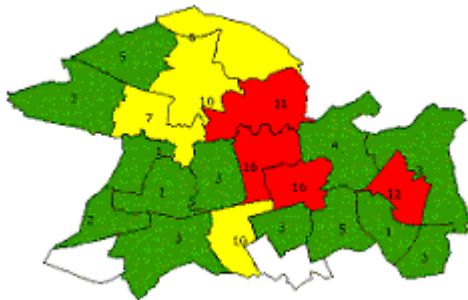
2.1.5 Templates

2.1.5.1 Map showing KPIs

This Map shows Key Performance Indicators (KPIs) for the Wards in a Local Authority

Each Ward is displayed in either Red, Amber or Green, depending in whether the KPIs Threshold values are reached or exceeded.

Red indicates a situation that requires urgent management attention, amber is a warning and green is within acceptable limits.



2.1.5.2 Reports at the Regional Level

This Report shows the total count of Customers gained and lost in an imaginary South-East Region

Rpt.1 Total Customers Gained and Lost by Week				
Date selected: End of May, 2012				
Week Ending	Location		Total Gained	Total Lost
March 6 th	SE Region		10	10
March 13 th	SE Region		20	20
March 20 th .	SE Region		30	30
March 27 th .	SE Region		40	40
April 3 rd /	SE Region		50	50
April 10 th .	SE Region		30	30
April 17 th .	SE Region		20	20
April 24 th .	SE Region		10	10

2.1.5.3 Reports at the City Level

This Report shows the total count of Customers gained and lost for London in the South-East Region.

RPT.1 Total Customers Gained and Lost by Week				
Date selected: End May, 2012				
Week Ending	Location		Total Gained	Total Lost
March 6 th 09	London		1	1
March 13 th 09	London		2	2
March 20 th . 09	London		3	3
March 27 th . 09	London		4	4
April 3 rd / 09	London		5	5
April 10 th . 09	London		3	3

April 17 th . 09		London		2		2
April 24 th . 09		London		1		1

2.1.5.4 Reports for Parking Tickets

This table shows a sample Template of unrealistic data for Parking Ticket Reports.

The Template is available on this page of the Database Answers Web Site :-

- http://www.databaseanswers.org/Parking_Rpts/PK06_TotalPaidPCNs_withPaymentMethod_demo_rpt.xls

PK.6 - Report on Total PCNs Paid with Payment Methods

Date selected: Month of January, 2011

PCN Type	Source	Payment Method	PCNs Paid	Amount Paid
PCN - BLE	H	Credit Card	5	£300.00
PCN - BLE	O	Cheque	186	£11,160.00
PCN - BLE	O	Credit Card	1	£60.00
PCN - BLE	O	Postal Order	4	£240.00
PCN - BLE	U	Auto Phone Payment	594	£35,700.00
PCN - CCTV	H	Credit Card	3	£150.00
PCN - CCTV	H	Debit Card	5	£250.00
PCN - CCTV	O	Cheque	171	£8,700.00
PCN - CCTV	O	Postal Order	2	£100.00
PCN - CCTV	U	Cash	50	£2,500.00
PCN - CCTV	U	Cheque	5	£250.00
PCN - DTE	H	Credit Card	28	£1,680.00
TOTAL			10,000	£500,000

2.1.6 FAQs

FAQ.1 Does your Chief Exec have Report requirements that you cannot meet ?

In order to respond to this situation appropriately, it is necessary to have an Information Catalogue, a Data Architecture and Data Lineage.

The solution then involves the following Steps :-

Step 1) Produce a draft Report for the Chief Execs approval

Step 2) Trace the lineage and perform a 'gap analysis' for all new data items.

Step 3) Talk to the Data Owners and establish when and how the data can be made available.

Step 4) Produce a Plan and timescale

Step 5) Review your Plan with the Chief Exec and obtain this agreement and formal sign-off.

Step 6) Deliver !!!

Performance Reports take data from Data Marts and many of the same considerations apply when it comes to determining **Best Practice**.

One difference is that is necessary to have a clearer understanding of the business operations and how the right kind of Performance Reports can provide insight to the business users.

This leads to the need for a management education process to be in place so that the evolution of Performance Reports can be planned in a logical manner, from basic summaries, to KPIs, Dashboards and so on.

FAQ.2 How do I produce Integrated Performance Reports for senior management ?

The key action here is to establish a unified Reporting Data Platform.

This will involve aspects previously discussed, including MDM, CMI and will certainly involve Data Lineage.

Senior Management will want to take a view of the integrated data and not focus on details of derivation.

Therefore, we have to follow the MDM approach with Data Lineage for each item in the Integrated Performance Reports.

FAQ.3 What are Key Performance Indicators ('KPIs')

Key Performance Indicators ('KPIs') are in common use and represent one aspect of Best Practice.

A variation of this approach are Key Quality Indicators,('KQIs') which are used to monitor and manage Data Quality.

Dashboards and Scorecards are often used in association with KPIs.

FAQ.4 Where can I find a Tutorial on Reporting ?

Here's a Tutorial from Database Answers on Integrated Performance Reporting –

-

http://www.databaseanswers.org/tutorial4_integrated_performance_reporting/index.htm

In broad terms, there are three areas involved :-

- i) Determine the Data Sources from the Data Marts
- ii) Choose the commercial Report-Writer
- iii) Create Data Validation and Transformation procedures

FAQ.5 How do I get certified as a Microsoft BI Specialist ?

Certification can be described as 'Necessary but not sufficient'. In other words, some employers consider it as evidence that you have the necessary technical knowledge and skills to be a Database Administrator, but without any experience, it will not guarantee you a job.

If you take your profession seriously and are committed to self-improvement, then you should certainly consider getting certified in the DBMS of your choice.

Here is a Web Link discussing the role of Microsoft Certified Technology Specialist in SQL Server Business Intelligence :-

<http://www.microsoft.com/learning/mcp/mcts/bi/default.aspx>

FAQ.6 How do I manage request for changes to Reports ?

When you are planning to produce Reports, it is vital to plan for changes to avoid disappointment.

The most common response when Users get their much-anticipated Reports for the first time, is for them to say – “Oh dear, that isn’t really what I wanted’.

Even when the Reports meet their Requirements, which will have been well-documented, and probably signed-off by the Users, they still want changes made.

There are some technical things you can do, including specifications for Report Templates which capture the features in families of similar Reports.

From a procedural point of view, you can discuss with the Users, how they see the patterns of future changes, and try to understand the operational environment. This will help you see how the Reports fit into their management style and

You can identify a progression from KPIs (Key Performance Indicators), Traffic Light Reports (using Red, Amber and Green to indicate the seriousness of situations being reported on), Dashboards, Scorecards

This will help you to arrange for the appropriate management education so that you and your Users are always in step, with your planning for what is just around the corner.

FAQ.7 What are the Qualities for Success in Performance Reporting

To be successful in this area of Performance Reporting it is useful to be able to see things from the User’s perspective and formulate the layout and content of the Reports accordingly

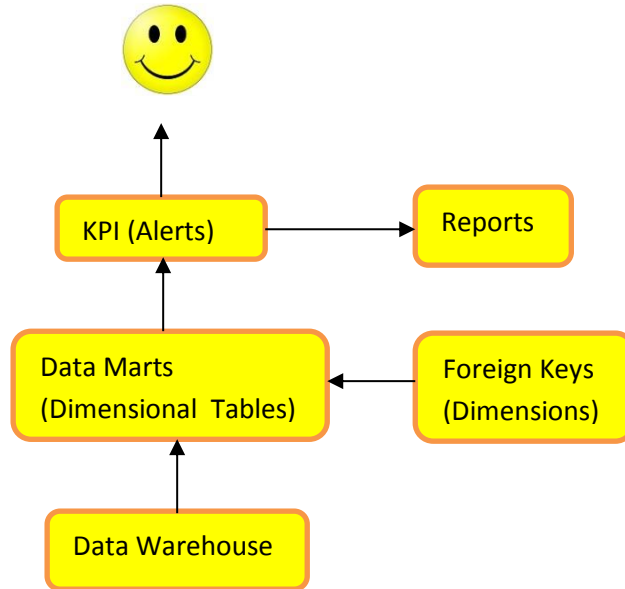
People who are successful working in this area are happy to work with End-Users and formulate Report requirements in a style that can be easily understood and implemented by the developers who might be the Report specialist.

They are subsequently able to implement the inevitable changes requests by the End-user and manage the expectations of the End-user and developers.

2.2 KPIs, Dashboards and Data Marts

We have defined a general approach to Key Performance Indicators (KPIs) and Data Marts.

It shows that our End-User receives an Alert when a KPI is triggered and can look at a Report to see the details.



We will generally consider using Dashboards to monitor KPIs, such as Alerts and Operational data.

2.3 Examples of KPIs

We have defined a general approach to Key Performance Indicators (KPIs) and Data Marts and here we show some examples.

2.3.1 Airport Management

Our KPI here is characterised by a time interval between Landings.

A cause for concern can be if the interval is greater than 1 minute and the CEO needs to be alerted immediately.

On this page, we show a Data Mart Design with KPIs :-

- [http://www.databaseanswers.org/bmews/bmews bi on the beach for airport in a box kpi data.htm](http://www.databaseanswers.org/bmews/bmews%20bi%20on%20the%20beach%20for%20airport%20in%20a%20box%20kpi%20data.htm)

2.3.2 Customers and Purchases

The beauty of this Data Mart is that applies to many kinds of Customers and Purchases.

It also shows very clearly a number of Dimensions all of which could be used as a KPI.

- [http://www.databaseanswers.org/data_models/customers and purchases data warehouse/index.htm](http://www.databaseanswers.org/data_models/customers_and_purchases_data_warehouse/index.htm)

2.3.3 Insurance

This Insurance example shows Dimensions of Customers, Dates, Locations and Policy Types :-

- [http://www.databaseanswers.org/data_models/POC insurance platform 2020/index.htm](http://www.databaseanswers.org/data_models/POC_insurance_platform_2020/index.htm)

2.3.4 Police

This simple example of a Star Schema for Police Information Reports analyses Criminal Activity by Addresses (or Locations), Crime Categories , People and Vehicles :-

- http://www.databaseanswers.org/data_models/police_information_reports/star_schema.htm

2.3.5 Restaurants

This shows a Data Mart for Restaurant Guide has Dimensions of Addresses, a Calendar for Dates, Types of Food, Star Gradings, and Types of Food :-

- http://www.databaseanswers.org/data_models/restaurant_guide/data_mart.htm

2.3.6 Retail

Retail is characterised by Customer Accounts and a need to report on total amounts and volumes by Account Types.

Here we show a Retail Data Mart that has Dimensions of

- http://www.databaseanswers.org/data_models/retail_customers/retail_customers_data_mart.htm

It has Dimensions of Calendar, Invoice, Mailshot Campaigns, Payment Methods, Products, Promotions, Staff and Stores.

This means that any of these Dimensions can be used to produce the appropriate KPIs.

2.3.7 United Nations

KPIs for the UN can reflect the Dimensions in the range of Reports that are available.

On this page, we show a Dimensional Model for FX (Foreign Exchange) Deals with three Dimensions of Currency, Deal and Date :-

- http://www.databaseanswers.org/data_models/investment_banking/financial_instruments_model.htm

This will become very useful when we get started with the Financial Planning Platform we are developing for the UN which is described on this page :-

- http://www.databaseanswers.org/data_models/un_global_compact_platforms_for_2017/index.htm

2.3.8 KPIs on the Golf Course

This shows an Alert for a major Airport where a KPI Monitoring System alerts the User on his Smartphone when there is an emergency at the Airport .

This picture shows the application of a Key Performance to send an Alert to the CEO of an Airport if an emergency occurs that requires his immediate attention.

This picture is on this page :-

- http://www.databaseanswers.org/bmews/bmews_bi_on_the_beach_for_airport_in_a_box_kpi_data.htm



3.Cloud Services

3.1 What is it ?

Cloud Services can be defined as a conceptual IT Delivery system with no concern of hardware, software or operating system.

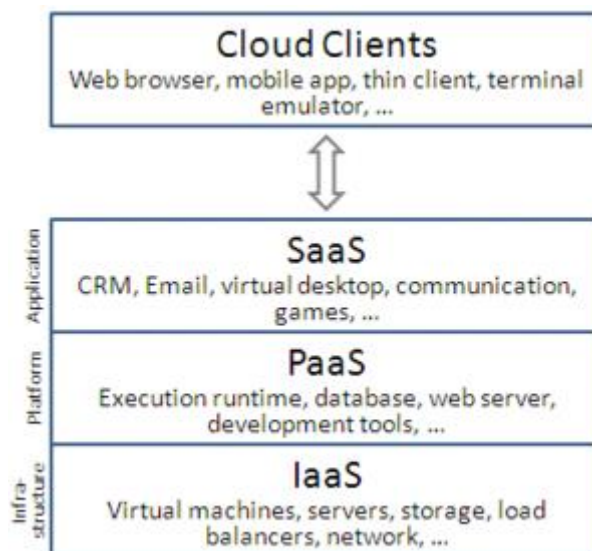
Wikipedia does not offer an entry for Cloud Services.

The nearest it suggests is Cloud Computing Service Models :-

- https://en.wikipedia.org/wiki/Cloud_computing#Service_models

It includes this diagram, which is describes as :-

“Cloud computing service models arranged as layers in a stack” :-



We define Cloud Services as a “Conceptual IT Delivery system with no concern of hardware, software or operating system”.

In other words, a User-oriented Business View such as Banking, Insurance, Retail and Travel.

- Event-Driven Platform
http://www.databaseanswers.org/data_models/event_driven_platform/index.htm

3.2 Why is it important ?

It is important because it represents a convenient way to think about the user and their interaction with IT Services.

In addition, we can combine this with a Model-View-Controller which is a well-established Application Architecture.

This provides us with a very powerful approach to discussing Cloud Services.

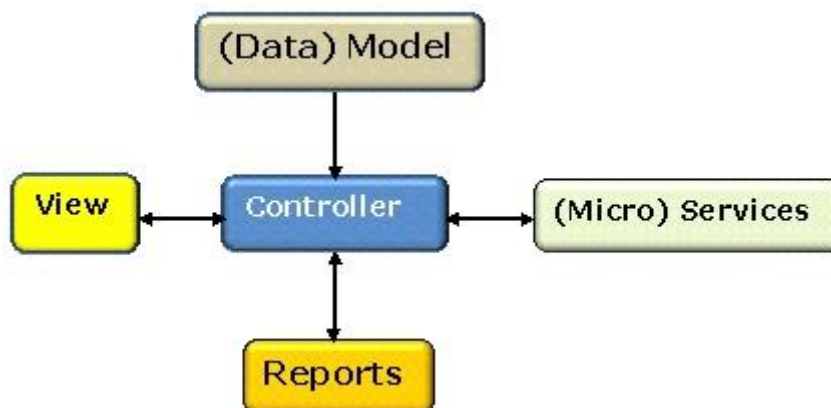
3.3 What will I learn ?

You will learn about a Model-View-Controller and how it can be the foundation for a Cloud Services Architecture.

3.4 Best Practice

This provides us with a very powerful that we show on our Database Answers Web site :-

- http://www.databaseanswers.org/data_models/mvc_model_view_controller/index.htm
which looks like this :-



This is where we run our operational Services based on our generic design based on our Canonical Data Model and Industry-specific Platforms.

Here we show three for illustration.

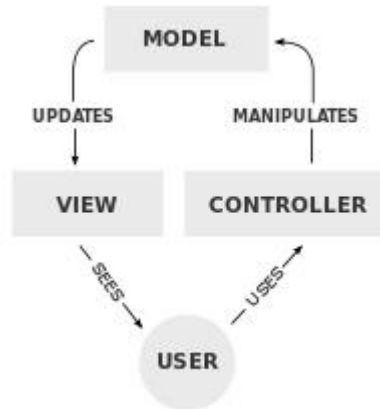
3.5 Model-View-Controller

We like the M-V-C as the Framework for our Solutions Architecture.

The Wikipedia version is on this page :-

- <https://en.wikipedia.org/wiki/Model%E2%80%93view%E2%80%93controller>

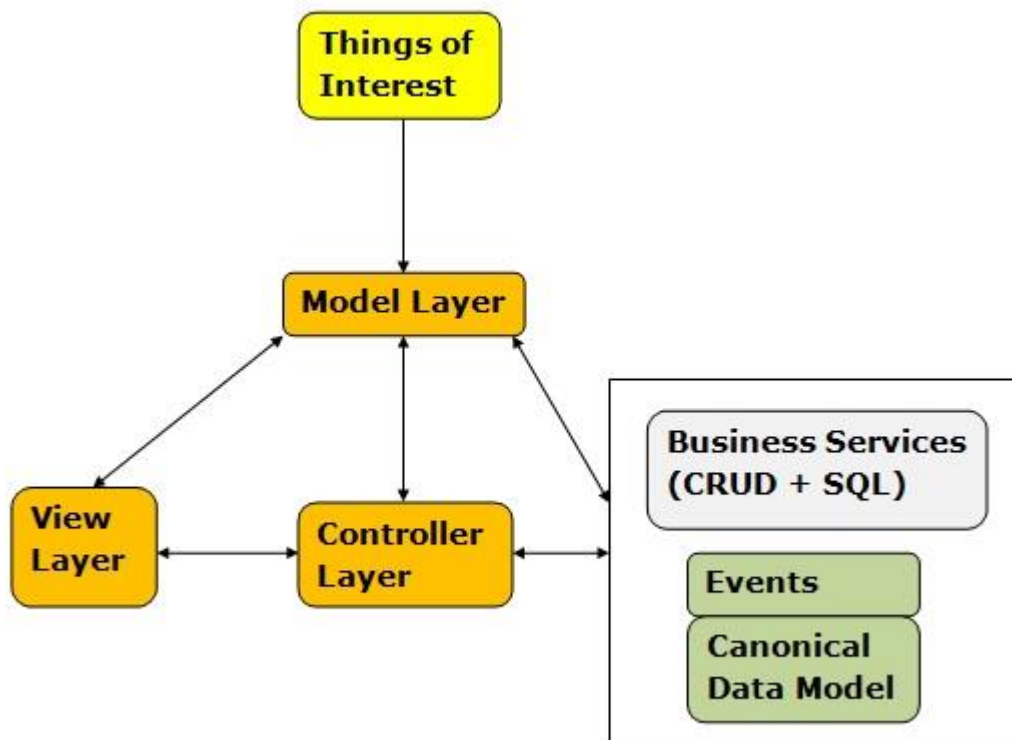
and looks like this :-



Our version is on the page :-

- http://www.databaseanswers.org/data_models/mvc_model_view_controller/index.htm

and looks like this :-



Wikipedia describes the three Components (MVC) in these terms :-

1. The *model* is the central component of the pattern. It expresses the application's behavior in terms of the problem, independent of the user interface. It directly manages the data, logic and rules of the application.
2. A *view* can be any output representation of information, such as a chart or a diagram. Multiple views of the same information are possible, such as a bar chart for management and a tabular view for accountants.
3. The *controller*, accepts input and converts it to commands for the model or view

4.Data Services

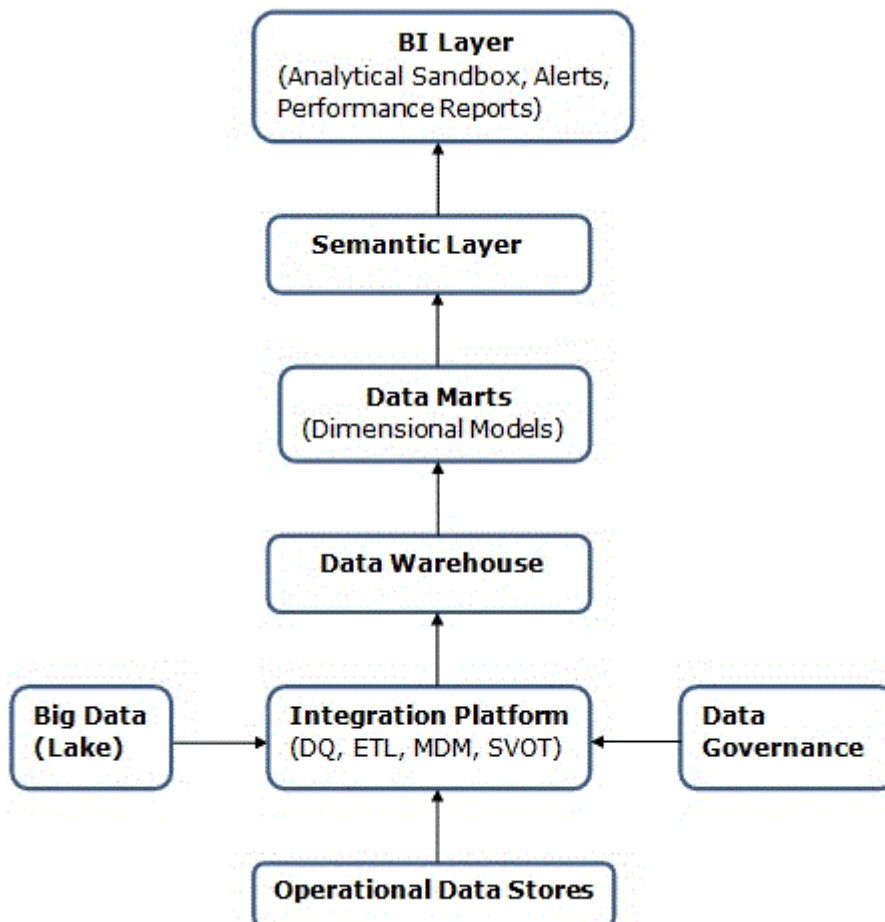
4.1 Reference Data Architecture

This is our Reference Data Architecture that we use to plan all our Data-related activities.

It is on our page :-

- http://www.databaseanswers.org/reference_data_architecture.htm

and looks like this :-



4.1.1 Data Marts

4.1.1.1 What is it ?

Wikipedia provides a good reference on Data Marts -

http://en.wikipedia.org/wiki/Data_mart

Data Marts are always built around Dimensions, such as Dates, Regions and Customers.

The other kinds of data are derived figures, such as totals.

4.1.1.2 Why is it important ?

Data Marts are important because they make it very easy to assemble data for Reports.

4.1.1.3 What will I learn ?

You will learn what a typical Data Mart looks like, how to design one and other useful facts.

4.1.1.4 Best Practice

Your Dimensions will always be Foreign Keys to Tables in your Data Warehouse.

Best Practice suggests that you position the Dimensions at the top of the Data Mart, and listed in alphabetical order.

4.1.1.5 Templates

4.1.1.6 Data Mart for Parking Tickets

This diagram shows a Data Model for a Data Mart to hold data about Parking Tickets issued by a Local Authority in the UK.

It was produced in a Word document from early discussion with the End-User and was very helpful in establishing communication and a collaborative method of working.

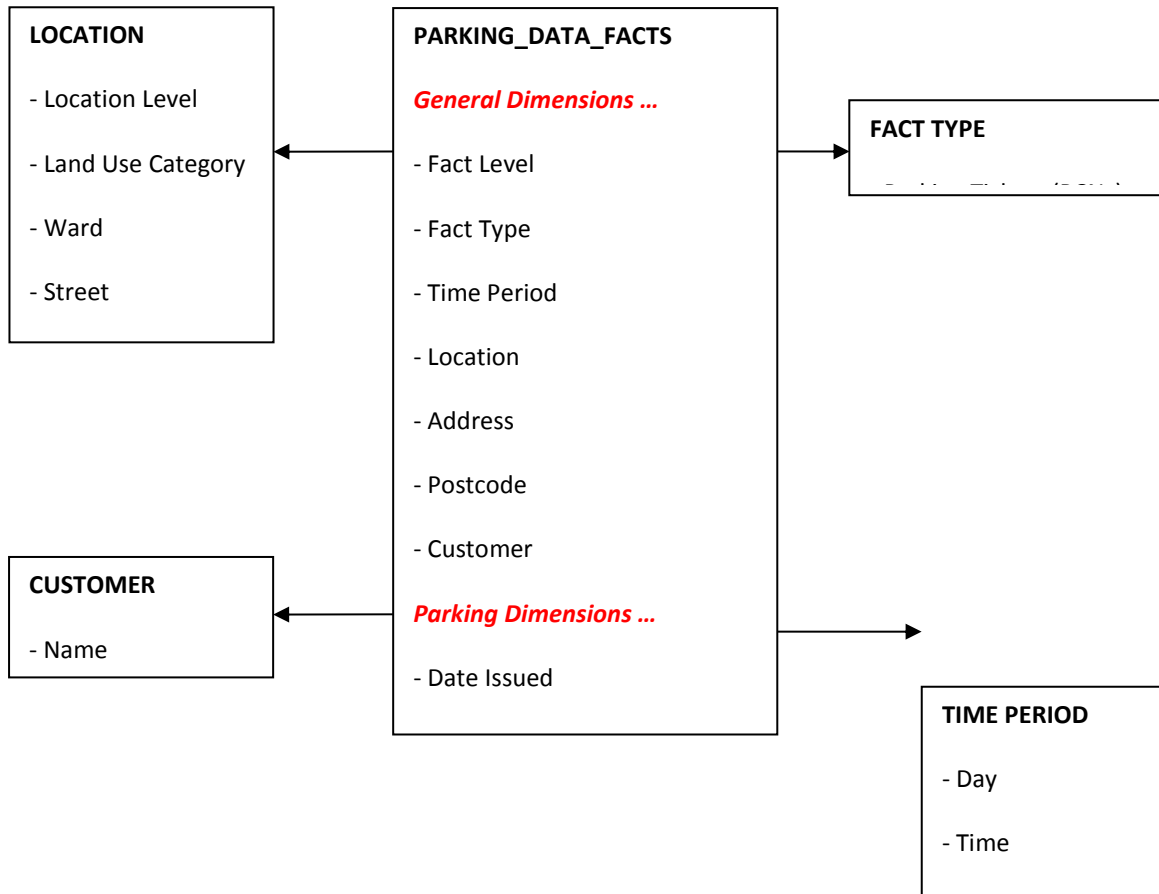
End-users find it easier to understand and agree to this kind of Data Model than a formal ERD.

This approach is therefore recommended.

Each Fact is associated with a number of Dimensions.

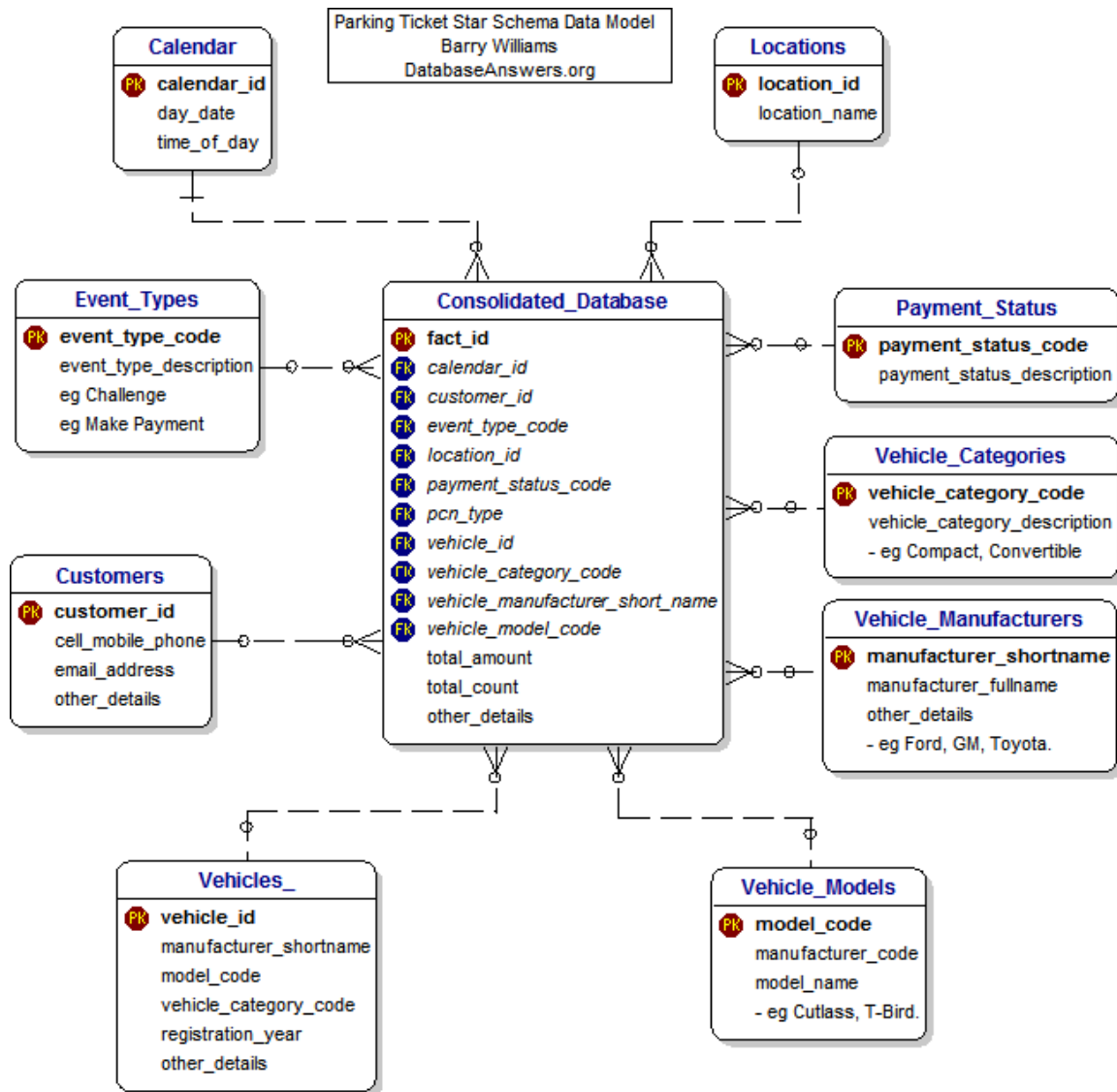
The 'FACTS' Table contains the list of data items which is available.

The other Tables are called 'Dimensions' and define how the Facts can be analysed.



4.1.2.5 Data Mart as a Data Model

This diagram was produced by a Data Modelling Tool and is the complete analysis of all the data required.



4.1.1.6 FAQs

FAQ.1 How do I design a Data Mart ?

The first step is to think about the Data Mart as a place where you simply throw all available data and provide 'hooks' so that any combination of data can easily be retrieved.

Briefly, the Key fields in Tables involved become Dimensions in a Data Mart.

Facts include all the basic data plus any derived data, typically averages, percentages and totals under various headings.

FAQ.2 What are the Qualities for Success in designing Data Mart ?

To be successful in designing Data Marts it is important to have a talent for visualizing the User's Requirements and for translating this to a formal design of Dimensions and Facts, together with the most important aspect, which is the derivation of the data required from the underlying basic data.

FAQ.3 How do I improve the performance of my Data Mart ?

Every DBMS produces what is called an Execution Plan for every SELECT statement.

The steps to improving the performance involve checking this Execution Plan against the Indexes that exist, and making sure that the Query Optimizer has used the appropriate Indexes to obtain the best performance.

This is a specialized area where DBA's spend a lot of their time when they are looking after production databases where speed is a mission-critical factor.

Data Marts are always created to support Business Intelligence, which includes Performance Reports, Balanced Scorecards, Dashboards, Key Performance Indicators and so on.

Best practice always requires user involvement and a generic design to support a flexible approach to meeting changing requirements.

Users will always want changes to their first specifications of their requirements.

The insight that they obtain from the first Reports helps them identify more precisely what their long-term requirements will be.

Therefore flexibility is important.

A well-designed Data Mart will anticipate the areas where flexibility is required.

The design process should always follow two steps :-

- Production of generic design for the Data Mart
- Implementation of the design with a specific Data Mart software product.

4.2 Data Warehouse

4.2.1 What is it ?

A Data warehouse is a repository of corporate data.

Wikipedia provides a good reference -

- https://en.wikipedia.org/wiki/Data_warehouse

I think of a Data Warehouse as a repository of centralised data from multiple source which has been transformed to be consistent with an Enterprise Data Model.

It is commonly used to provide a 'Single View of Corporate Data'

Wikipedia has a useful entry for a Data Warehouse at this page :-

- https://en.wikipedia.org/wiki/Data_Warehouse

It contains the following description :-

- A **data warehouse** is a used for [reporting](#) and [data analysis](#), and is considered a core component of [business intelligence](#) which stores integrated data from one or more disparate sources. They store current and historical data and are used for creating analytical reports for knowledge workers throughout the enterprise. Examples of reports could range from annual and quarterly comparisons and trends to detailed daily sales analysis.
- The data stored in the warehouse is [uploaded](#) from the [operational systems](#) (such as marketing or sales). The data may pass through an [operational data store](#) for additional operations before it is used in the DW for reporting.

4.3 Semantic Layer

4.3.1 What is it ?

Wikipedia provides a good reference on Semantic Layers -

- https://en.wikipedia.org/wiki/Semantic_layer

We have an entry for Semantic Layers on this page on Airport Operations :-

- http://www.databaseanswers.org/data_models/airport_operations2/index.htm

The Semantic Layer provides a 'User-Friendly' interface to a Database by translating Database terms into business terms.

4.3.2 Why is it important ?

A Semantic Layer is important because it make it very easy for business users to refer to data using terms that they are comfortable and familiar with.

4.3.3 What will I learn ?

You will learn what a typical Semantic Layer looks like, how to design one and other useful facts.

4.3.4 Best Practice

Best Practice suggests that you introduce the role of a Data Steward to be responsible for the contents of the Semantic Layer.

Here is a simple example, where we can see that the term 'Party' is translated into 'Passenger' in an Airport environment.

Approved by (Data Steward)	Approved Date	Database	Business
Joe Bloggs	April 1st. 2017	Party	Passenger

4.4 Agile Data Modelling

4.4.1 What is it ?

Wikipedia provides a good reference on Agile Modelling :-

- https://en.wikipedia.org/wiki/Agile_modeling

It is more flexible than traditional modeling methods.

4.4.2 Why is it important ?

Its increased flexibility makes it more appropriate to development projects.

4.4.3 What will I learn ?

It is part of a disciplined agile delivery :-

- https://en.wikipedia.org/wiki/Disciplined_agile_delivery

4.4.4 Best Practice

Best Practice dictates that agile delivery is part of a development-oriented discipline that focusses on deliverables.

An extension of data modelling patterns is the adaptive data model (ADM), a generalized data model used in data warehouse design and discussed in the Cutter report "The Message Driven Warehouse" :-

- <https://www.cutter.com/article/message-driven-warehouse-new-architectural-model-bi-systems-400661>

5. Information Catalogue

5.1 What is it ?

An Information Catalogue is a Repository of Information related to Information systems.

5.2 Why is it important ?

It is important because it provides a single point of reference and consistent definitions for data items, such as 'What is a Customer' – is it somebody who has actually purchased or simply made an enquiry ?

5.3 What will I learn ?

You will learn how to design and build an Information Catalogue.

5.4 Best Practice

You can get started on this page of our Database Answers Web Site that list some commercially available Data Dictionaries :-

- http://www.databaseanswers.org/data_dictionaries.htm

Here is the page listing our Data Models for Data Dictionaries :-

- http://www.databaseanswers.org/data_models/data_dictionary/index.htm

This page lists some very interesting Questions and Answers about Data Dictionary :-

- http://www.databaseanswers.org/data_models/data_dictionary/facts.htm

5.5 Templates

A suitable Template for getting started is a simple table.

An example for a Template for Entries in a Glossary is show on this page :-

- http://www.databaseanswers.org/template_for_new_style_pages.htm

5.6 FAQs

FAQ.1 How do I publish an Information Catalogue ?

It is good to start with a Spreadsheet and then move to a stand-alone Access Database before finally migrating to an Internet-based Database which is published on a Web Site.

FAQ.1 How can I be sure of Success with an Information Catalogue ?

To be successful in maintaining and publishing an Information Catalogue it is beneficial to enjoy detail and to have an interest in ensuring that all interested parties are on the same Page.

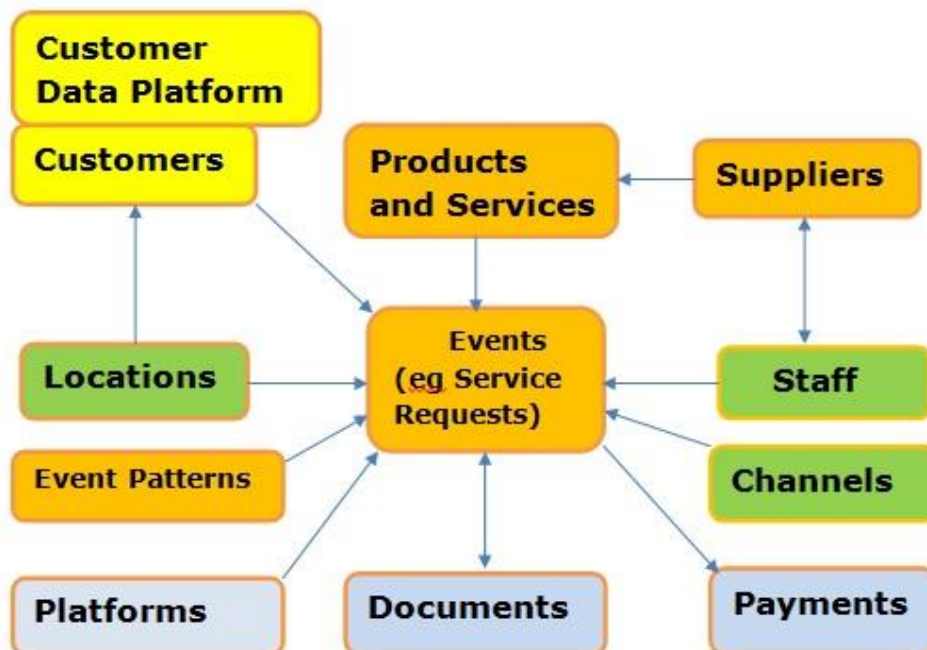
It is also useful to have an eye for detail and to have an appreciation for the way in which the separate Components within the Information Catalogue are interrelated.

6. Canonical Data Model

This is discussed in this page of my Database Answers Web Site :-

- http://www.databaseanswers.org/data_models/canonical_data_models/index.htm

And here is what it looks like :-

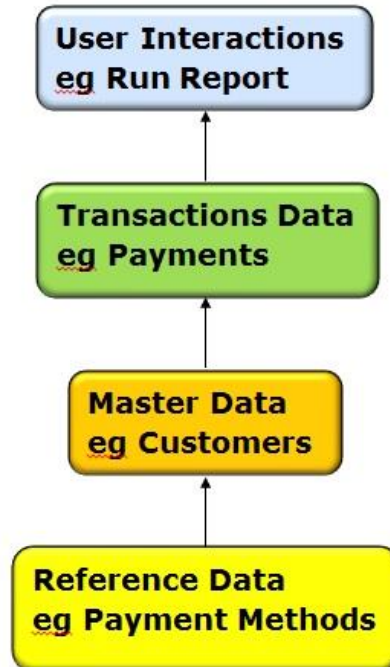


6.1 Generic Data Platform

Here is our Model :-

- http://www.databaseanswers.org/data_models/canonical_data_models/index.htm

Here is what it looks like that demonstrates the fundamental design of Layers :-

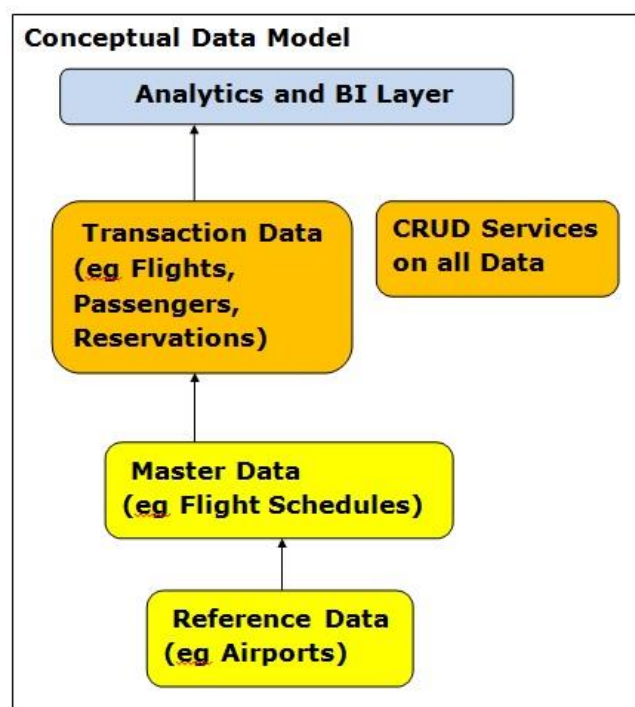


6.2 Industry Platforms

The term 'Platform' is commonly used but without an exact definition of its meaning, and our Data Platforms are defined as simply having many Layers of different kinds of Data.

6.2.1 Airport Management

- http://www.databaseanswers.org/data_models/airport_mgt_platform/index.htm



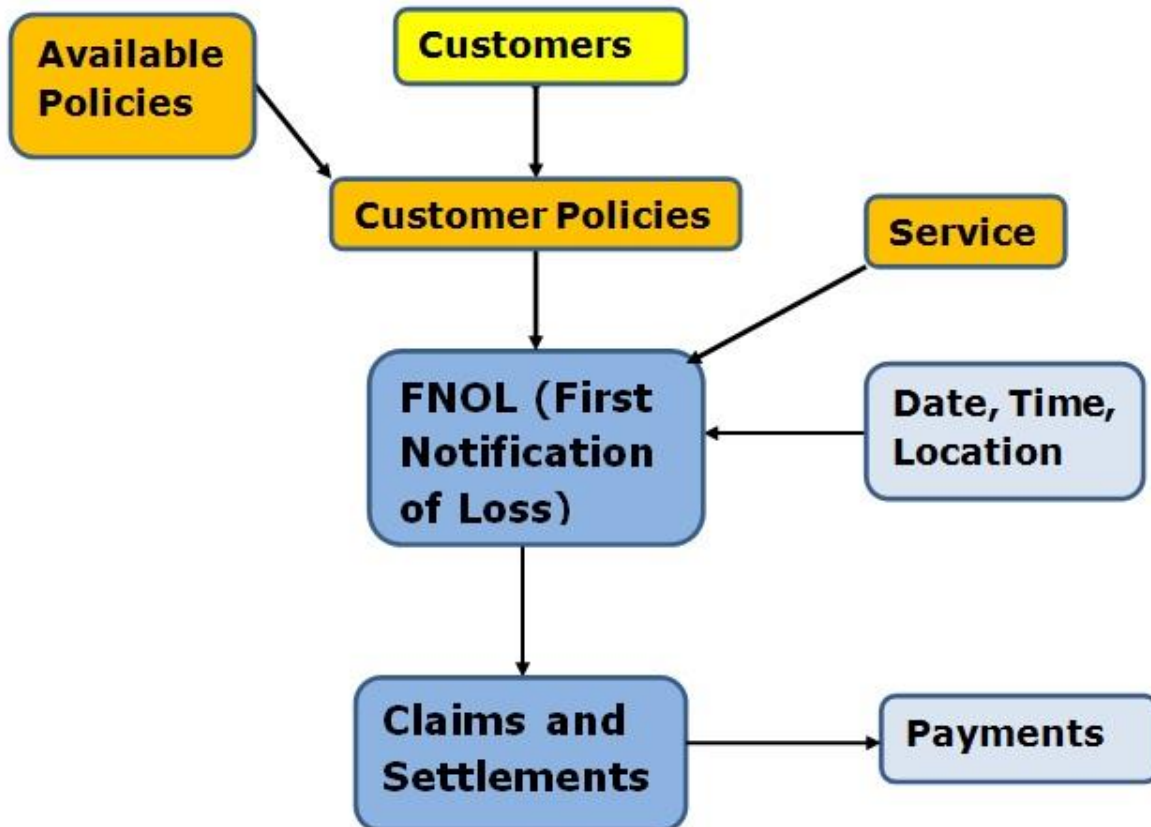
6.2.2 Insurance (FNOL)

FNOL is standard Insurance industry terminology for 'First Notification of Loss' and it means the data that a specific Customer has to provide as part of the first time a claim is made.

The Data Model is on this page :-

- http://www.databaseanswers.org/data_models/insurance_fnol/index.htm

And looks like this :-



6.2.3 United Nations

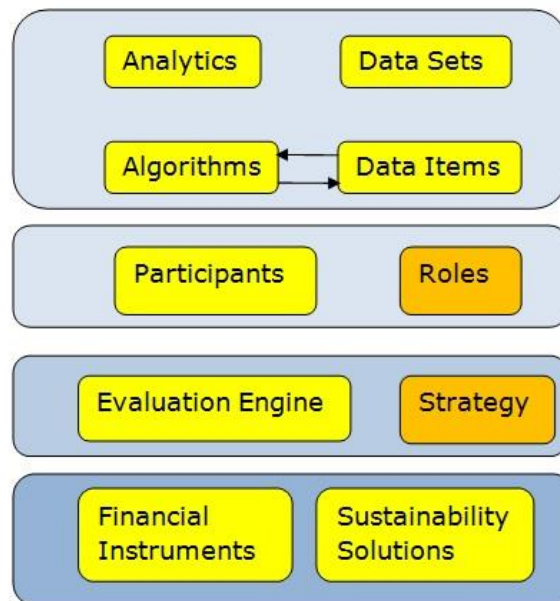
The UN Global Compact is a UN initiative to encourage business support of the UN activities and objectives.

We have been members since 2002 and here is a link to our Letter of Appreciation from the UN :-

- http://www.databaseanswers.org/letters/un_letter.htm

Here is the Web Link to our UN Platform :-

- http://www.databaseanswers.org/data_models/un_global_compact_platforms_for_2017/index.htm



7.Data Sources

7.1 What is it ?

The process of organising and managing data from different sources.

It typically involves cleaning-up and transforming data to a standard format for subsequent processing.

This can typically be part of a Data Integration activity.

7.2 Why is it important ?

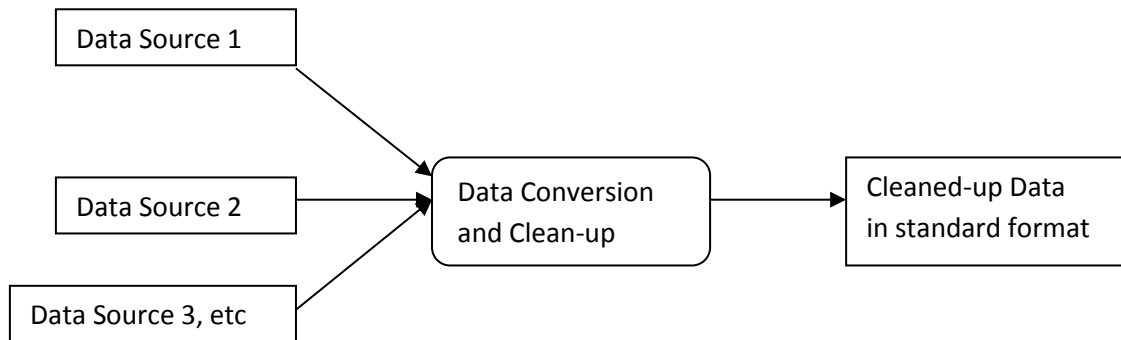
It is important because a common requirement is to identify multiple data sources.

7.3 What will I learn ?

You will learn how to identify multiples sources and formats in order to identify how to convert them to a common format for subsequent processing.

7.4 Best Practice

Best Practice looks like this :-



7.5 Templates

The Templates make it possible to record the details of all the data formats, sources and data stewards.

7.6 FAQs

This Wikipedia entry is a useful introduction :-

- https://en.wikipedia.org/wiki/Operational_data_store

8. Master Data Management

Wikipedia has a useful entry for Master Data Management (MDM) at this page :-

- https://en.wikipedia.org/wiki/Master_data_management

which says :-

- “The data that is mastered may include:
[reference data](#) – the business objects for transactions, and the dimensions for reports and analysis

MDM is a part of the ‘Single View of the Truth’.

It provides a solution to known problems and many major players, such as Informatica, offer commercially available solutions.

In a Blog by Erik Haahr posted on March 9th. 2017, he stated that the Gartner Group suggests four different styles of MDM Hub implementations :-

- (Gartner <http://www.gartner.com/technology/home.jsp>)

Here are the four Styles :-

- 1) Registry Style
The MDM Hub is the centre of Reference but stores only an Index used to retrieve master data from relevant backend systems
- 2) Centralised Style
The MDM Hub updates necessary in all backend system and is both a system of reference and a system of entry.
- 3) Coexistence Style
The MDM Hub stores all Master Data. The data entry procedure updates the Master data which then updates all the backend systems.
- 4) Consolidation Style
The MDM Hub is the system of reference for reporting purposes and stores all master data so there is no need to get data from backend systems.

My personal preference is the Registry Style because it is simple, neat and elegant and I have used it with great success.

Here is a link to my Data Model for a Customer Master Registry :-

- http://www.databaseanswers.org/data_models/customer_master_index/index.htm

8.1 MDM in Practice

This text is also on my Web Site at this page on my Tutorial on Cloud Services Architecture :-

- http://www.databaseanswers.org/data_models/tutorial_cloud_svcs_architecture/index.htm

In my Reference Data Architecture, I have identified three levels of Data :-

1. Transaction
2. Master
3. Reference

Of these, both Master and Reference are candidates for MDM and here we show some examples.

We might need to link to established sources of Master and Reference Data.

Master Data	Reference Data	Data Stewards	Date Changed	Nature of Change
Timetables		Joe Bloggs		
	Airlines	John Smith		
	Airports	John Smith		

9. Big Data (Lake)

9.1 What is it ?

Wikipedia has a useful entry for Data Lakes at this page :-

- https://en.wikipedia.org/wiki/Data_lakes

where it states :-

“A **data lake** is a method of storing data within a system or repository, in its natural format,^[u] that facilitates the collocation of data in various schemata and structural forms, usually object blobs or files.”

Two important features are

1. data stored ‘in its natural form’ – in other words, no data transformation
2. “collocation of data in various schemata” – in other words, data is simply stored in the format it arrives in.

I like the phrase ‘Data Lake’ because it is a simple and convenient way to describe something that would otherwise be difficult and complex to describe

Here are some guidelines for Best Practice -

1. Try to blend Traditional Enterprise data with Big Data in a Data LAKE
2. Add Self-Service functions
3. Ensure compliance with existing Data Governance to avoid losing the benefits of established procedures.
4. Choose modern data integration tools that will smooth the path to adopting , building on existing expertise.
5. Integrate your Data Lake with your existing Enterprise Data Architecture to co-exist with Data Warehouses and Marketing Sub-systems.

Using a commercial product such as Microsoft's Azure Data Lake, which we discuss below, will help you accelerate your learning curve.

9.2 Azure Data Lake

Microsoft has introduced Data Lakes for Azure and here is a good starting point :-

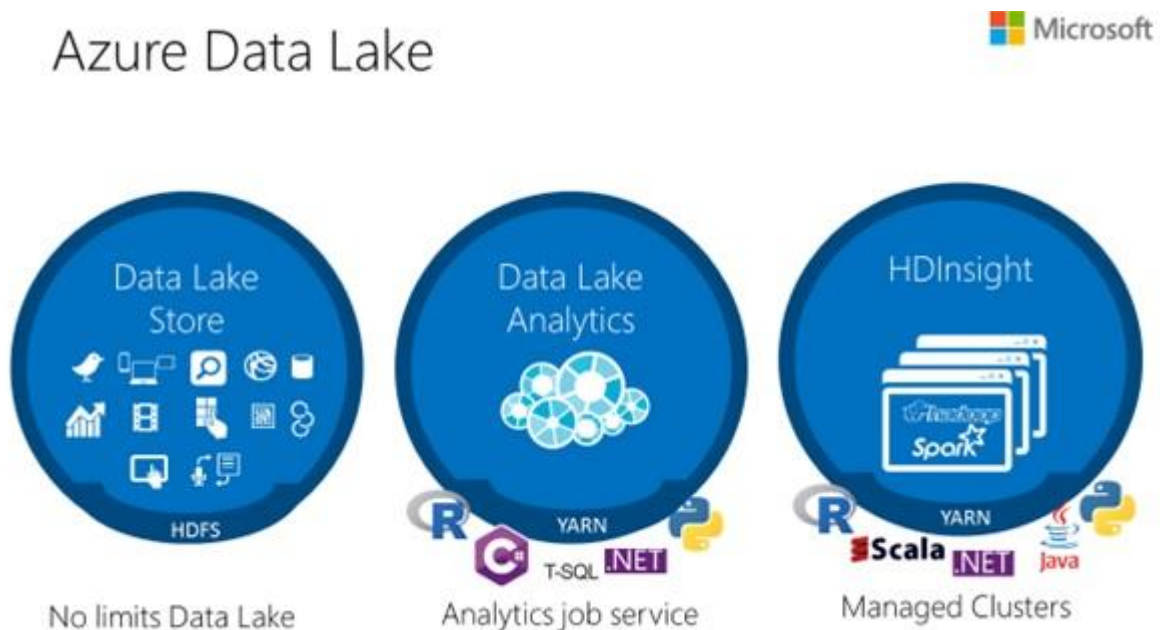
- <https://azure.microsoft.com/en-us/blog/the-intelligent-data-lake/>

Azure Data Lake Analytics is described in the following terms :-

“This is the first cloud analytics job service where you can easily develop and run massively parallel data transformation and processing programs in U-SQL, R, Python and .NET over petabytes of data.

It has rich built-in cognitive capabilities such as image tagging, emotion detection, face detection, deriving meaning from text, and sentiment analysis with the ability to extend to any type of analytics.”

This Microsoft diagram shows the three components of the Azure Data Lake :-



9.3 Data Integration Platform

9.3.1 What is it ?

Data Integration is the process of producing one stream of related data from a number of streams of data from different sources.

For example, Customer data can be obtained from retail purchases and telephone bills.

The key activity involves matching data for the same Customer from different streams.

In the case of Customers, this can be achieved by matching facts about a Customer such as names, addresses, gender and data of birth.

9.3.2 Why is it important ?

It is important because it provides a 'Single View of the Truth'

9.3.3 What will I learn ?

You will learn how to match data using appropriate external standards.

For external, in the UK, the Government maintains a so-called 'Post Office Address Format' or PAF-File and commercial software is available to simplify the matching process.

9.3.4 Best Practice

This page on our Database Answers Web Site is a good starting point :-

- http://www.databaseanswers.org/enterprise_data_integration.htm

9.3.5 Templates

This page on the IBM Web Site is a useful introduction to the more general topic of Master Data Management :-

- <http://www.ibm.com/analytics/us/en/technology/master-data-management/>

9.3.6 FAQs

This page has some very useful links to help you get started on Customer Data Integration:-

- http://www.databaseanswers.org/customer_data_integration.htm

Here is a useful set of links for commercial De-Duping software :-

- <http://www.databaseanswers.org/deduping.htm>

9.3.7 ETL – Extract, Transform and Load

ETL is a very important component of the Integration Platform.

This Wikipedia entry is a very useful introduction:-

- https://en.wikipedia.org/wiki/Extract,_transform,_load

The function of ETL is to take data from multiple sources, clean it up and transform it so that it can be loaded into a Data Warehouse.

It might include providing a 'Single View of the Truth', so that, for example, a Customer called John could be recognised as Johnny, Jon or Jonno.

It is common to find Libraries of Transformation Utilities being used that reflect corporate standards, such as closing Dates for Sales Orders.

9.4 Data Governance

9.4.1 What is it ?

Data Governance is the control of change within an IT and Data environment.

This requires some activities that are general, such as clearly-defined Roles and Responsibilities, and some that are specific to Data Management, such as Data Lineage.

The Sarbanes-Oxley Act came into law in the USA in 2002 and introduced much higher standards of financial compliance for any publicly-quoted company.

This means that Chief Executives have to state that the figures in their Annual Reports are 'the truth, the whole truth and nothing but the truth'.

This in turn has focussed a great deal of attention on Data Governance.

It means, for example, that Data Lineage is now 'front and centre' because it must be possible to trace every item of data in the Annual Reports down to the source of the data in an Operational System within the organisation.

9.4.2 Why is it important ?

It is important because there is an increasing need for accountability, for a 'Single View of the Truth' and for an understanding of how data is controlled within complex organisations.

All of these factors require a better awareness and better control of changes to the way is sourced, retrieved and processed between Data Sources (qv) and Performance Reports (qv).

9.4.3 What will I learn ?

You will learn how Data Governance can be implemented and how to determine if an organisation has good Governance in place.

You will also learn how to prepare for Certification and where to go for more detailed information.

9.4.4 Best Practice

Many people talk about Data Governance and many organisations talk about adopting it.

The reality is that progress is rather slow in implementing Governance.

This is primarily because it requires a commitment from the top to the bottom of management.

9.5 Useful Links

The Data Governance Institute –

- <http://www.datagovernance.com/>

Excellent review of Sarbanes-Oxley in Wikipedia –

- http://en.wikipedia.org/wiki/Sarbanes-Oxley_Act

LinkedIn has a number of relevant Groups, and here are just two of them :-

- Data Stewardship & Governance
- The Data Quality Association

9.6 Templates

This Section shows two examples of Templates.

9.6.1 Data Lineage

This table shows the example of Profitability.

Data Item	Source	Description	Lineage
Profitability	Annual Report	The difference between Revenue and Costs for a specified time-period.	Revenue is obtained from the Billing System. Costs are derived from the Purchasing System.

9.6.2 Roles and Responsibilities

This table shows the example of Profitability.

Role name	Incumbent	Responsibilities
Data Governance Mgr	Jane Doe	Responsible for changes to Roles and Responsibilities
Data Steward	John Doe	Responsible for changes to content of the Information Catalogue

9.7 FAQs

FAQ.1 What is Data Governance ?

Data Governance can be defined simply as 'Doing things right' in Enterprise Data Management by complying with the appropriate rules, policies and procedures.

These will all be designed to make sure that data used throughout the Enterprise is good-quality data, certainly when it appears in Performance reports.

FAQ.2 Why should my organisation have a Data Governance function ?

The existence of a **Data Governance function is a measure of the maturity of Data Management within an organization**

The first steps should be to establish a thin slice of Data Governance from top to bottom

- Wikipedia on Data Governance - http://en.wikipedia.org/wiki/Data_governance
- Alignment of Enterprise Architecture with Business Goals –
- http://www.information-management.com/infodirect/2009_115/enterprise_architecture_togaf-10015189-1.html?ET=informationmgmt:e886:2099687a:&st=email

If you are active in this area, you should consider joining a professional organizational.

This helps you to network with your peer group and will encourage you to keep up-to-date in knowledge and professional practice.

Here are two organisations that are planning active roles in Data Governance :-

i) The Data Governance Institute (Membership starts at \$150 for individuals) :-

<http://www.datagovernance.com/>

ii) The Data Governance and Stewardship Community of Practice (\$150/year) :-

- <http://www.datastewardship.com/>

It includes coverage of some very useful Case Studies :-

http://www.datastewardship.com/content.aspx?page_id=22&club_id=885168&module_id=37956

It also maintains a Data Governance Software Web Site :-

<http://www.datagovernancesoftware.com/>

and Sarbanes-Oxley Web Site - <http://www.sox-online.com/>

FAQ.3 How do I get a top-down view of Data Management in my organisation ?

Answers to this question are at different levels :-

- Data Governance at the top-level
- Master Data Management at the mid-level
- Data Integration at the mid-level
- Data Owners and Sources at the lowest level
- Information Catalogue mandated as the central repository of all this information
- Appropriate procedures in place to control all of these factors.

10. Data Mining

10.1 What is it ?

Wikipedia has a useful entry for Data Mining at this page :-

- https://simple.wikipedia.org/wiki/Data_mining

It states (in summary) :-

Data Mining is about finding new [information](#) in a lot of [data](#) with the aim of finding data that is both new and useful.

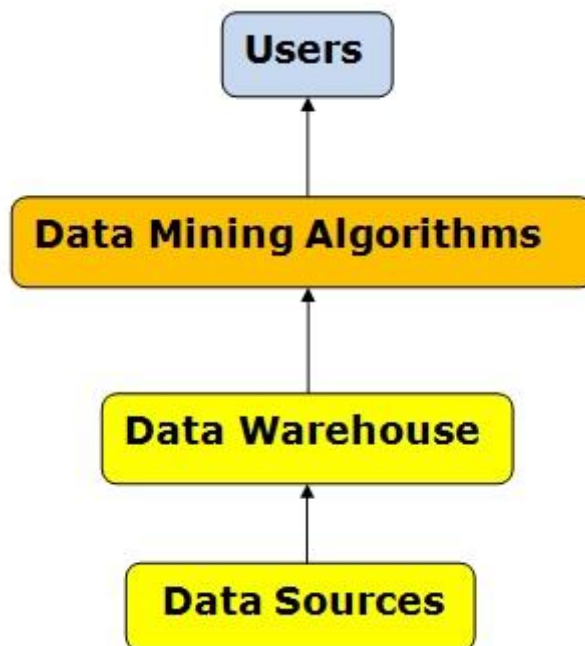
In many cases, data is stored so it can be used later. The data is saved with a goal. For example, a store wants to save what has been bought. They want to do this to know how much they should buy themselves, to have enough to sell later.

Saving this information, makes a lot of data.

The data is usually saved in a [database](#). The reason why data is saved is called the first use.

We have Data Models on this page :-

- http://www.databaseanswers.org/data_models/data_mining/index.htm and the Conceptual Model looks like this :-



11. Data Modelling Theory

11.1 Inheritance

Wikipedia has a useful entry for Inheritance at this page :-

- [https://en.wikipedia.org/wiki/Inheritance_\(object-oriented_programming\)](https://en.wikipedia.org/wiki/Inheritance_(object-oriented_programming))

In our Database Answers Web Site we have several Inheritance-related Data Models which provide valuable insight into the theory and practice of Inheritance, including these :-

- Aircraft –
 - http://www.databaseanswers.org/data_models/aircraft_and_inheritance/index.htm
- City Tourist Guide -
 - http://www.databaseanswers.org/data_models/city_tourist_guide/index.htm
- Insurance and eClaims –
 - http://www.databaseanswers.org/data_models/insurance_and_eclaims/index.htm
- Retail Customers –
 - http://www.databaseanswers.org/data_models/retail_customers/retail_customers_and_inheritance.htm
- Roles, Inheritance and Sub-types :-
 - http://www.databaseanswers.org/data_models/roles_inheritance_and_subtypes/index.htm
- Union Grievances –
 - http://www.databaseanswers.org/data_models/union_grievances/union_grievances_inheritance_model.htm
- Vehicle Maintenance -
 - http://www.databaseanswers.org/data_models/vehicle_maintenance_with_inheritance/index.htm

On this page, we have a detailed discussion :-

- Design Notes –
 - http://www.databaseanswers.org/inheritance_design_notes.htm

11.2 Semantic Models

Wikipedia has a descriptive entry for Semantic Models at this page :-

- https://en.wikipedia.org/wiki/Semantic_data_model

We also have this reference for Semantic Models on our Database Answers Web site :-

- http://www.databaseanswers.org/data_models/semantic_models/index.htm

Wikipedia states :-

Semantic models store information about Relations in the form of triples like this Object-RelationType-Object.

For example: the Buckingham Palace <is located in> London.

12. Data Vault

12.1 What is it ?

Wikipedia has a useful entry for Data Vault Modelling at this page :-

- https://simple.wikipedia.org/wiki/Data_vault_modeling

It states (in summary) :-

“Data vault modeling is a [database modeling](#) method to preserve different [sets](#) of historical data from different sources. It is also a method of looking at historical data that deals with issues such as auditing

Data Vault Modeling focuses on several things. First, it emphasizes the need to trace where all the data in the database came from. Each row has extra attributes that describe where the data came from, and at what time it was loaded. This feature lets auditors find the source of the values.

It is an approach developed by [Dan Linstedt](#).

13. Data Integration Products

13.1 What is it ?

Data Integration is the process of organising and managing data from different sources.

It typically involves cleaning-up and transforming data to a standard format for subsequent processing.

This can typically be part of a Data Integration activity.

This is an introduction to various Commercial Platforms that provide some enterprise Data Integration facilities.

13.2 Why is it important ?

It is important because a common requirement is to identify multiple data sources.

13.3 What will I learn ?

You will learn how to identify multiples sources and formats in order to identify how to convert

13.4 Some Commercial Products

13.4.1 Liaison Alloy Platform

Here are some quotes from the Liaison Web Site :-

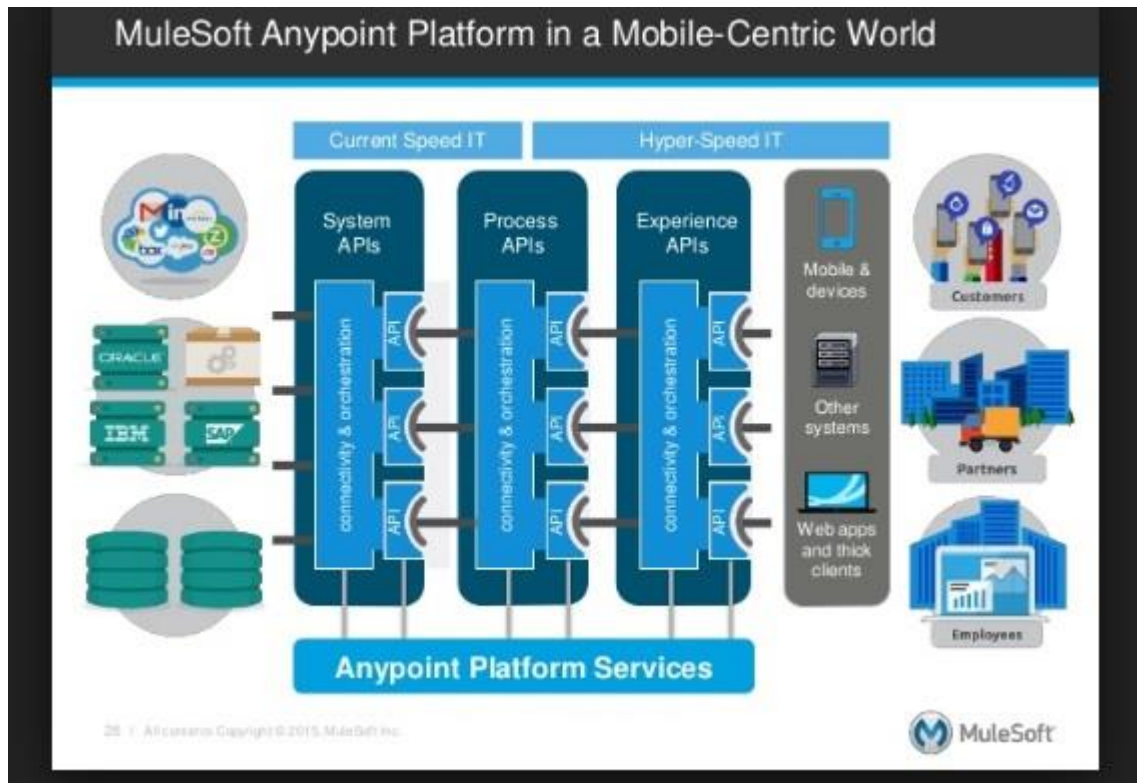
“Conceived from the ground up to address today’s technology disruptors, ALLOY is a next generation cloud platform for solving today’s integration and data management challenges.

- **ALLOY provides** unified integration and data management capabilities as managed services, buffering the complexities of increasing data volume and variety
- **ALLOY connects** any two application end points: cloud, mobile, device, on-premises, etc.
- **ALLOY persists** data in a big data repository, providing on-demand, self-service access to clean, quality data
- **ALLOY provides** built-in security and compliance
- **ALLOY is an efficient alternative** to DIY integration models such as ESB or iPaaS at a time when connections are growing exponentially
- <https://www.liaison.com/liaison-alloy-platform>
- <http://www.idevnews.com/stories/6515/Liaison-Alloy-Platform-Redefines-Integration-and-Data-Management>



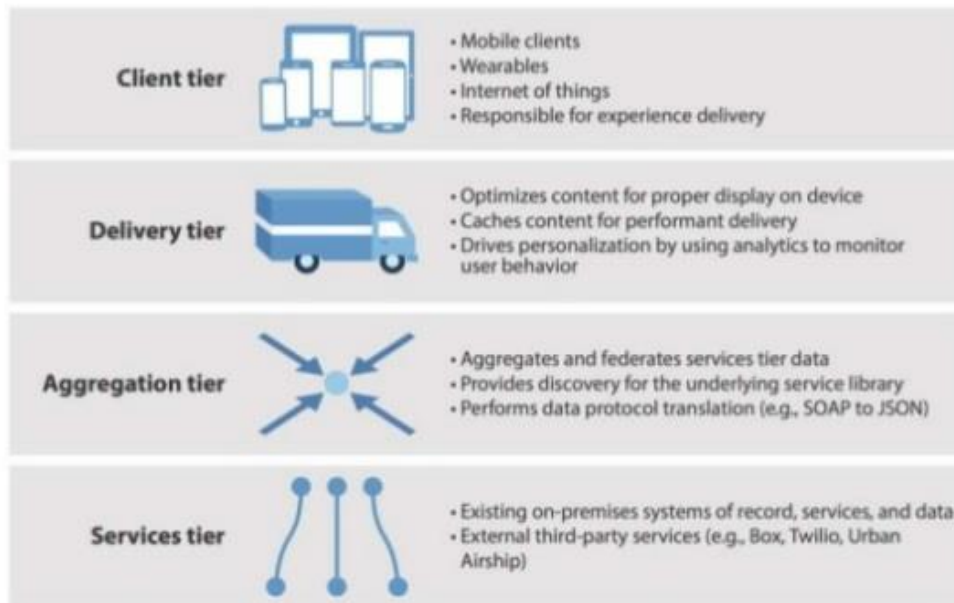
13.4.2 Mulesoft

13.4.2.1 Mulesoft’s Anypoint Platform



13.4.2.2 Mulesoft and Forrester

Forrester Research: The Four-Tier Engagement Platform



Source: Forrester Research, Inc.

12 | All contents Copyright © 2015, MuleSoft Inc.



13.4.3 Salesforce

13.4.3.1 Salesforce Cloud Platform

- <http://focusonforce.com/platform/salesforce-platform-overview/>



13.4.3.2 Salesforce and Events

On this page :-

- <https://www.slideshare.net/salesforcefoundation/georgetown-university-and-st-norbert-college-improving-recruiting-efficiency-webinar>

We like this slide because it combines the words Platform and event.



13.4.4 SAP and Google (Kronva)

They have a Netweaver Platform on this page :-

- <http://www.kronva.com/>

13.4.5 Software AG

Digital Business Platform for SAP :-

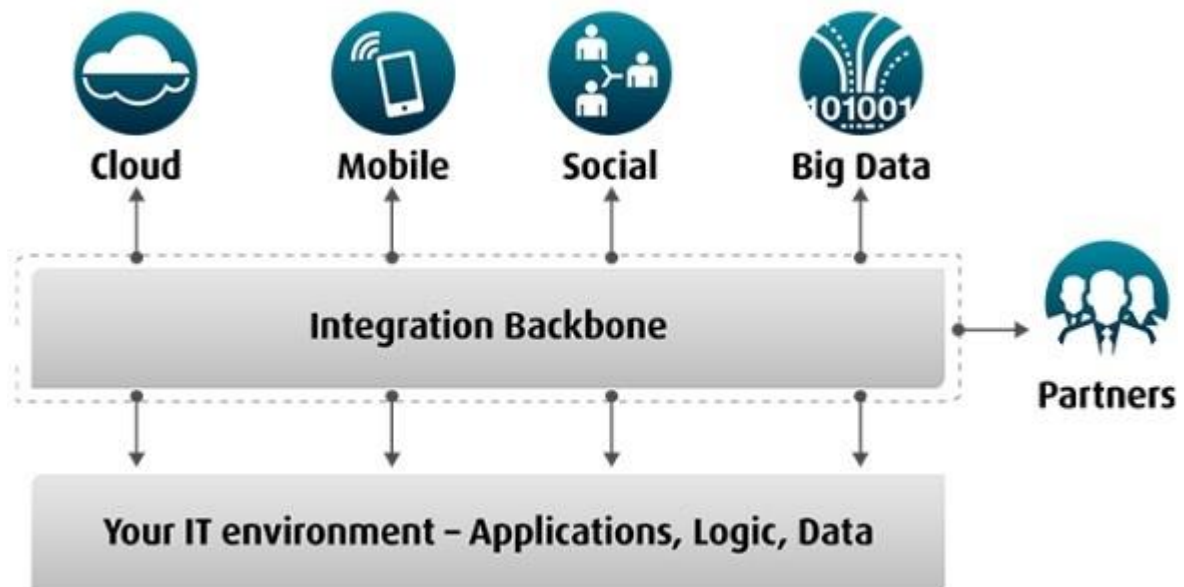
- https://marketplace.softwareag.com/apps/48105#!features/SAP_process_design

Claims and Policy Management for Insurance :-

- <https://marketplace.softwareag.com/apps/37895#!overview>

Here is their DBP Integration Platform or Webmethods Integration Platform on this page :-

- http://www2.softwareag.com/corporate/products/webmethods_integration/integration/default.aspx



14. Microsoft Links

Our initial analysis shows an overlap between Power BI and the Azure Data Factory.

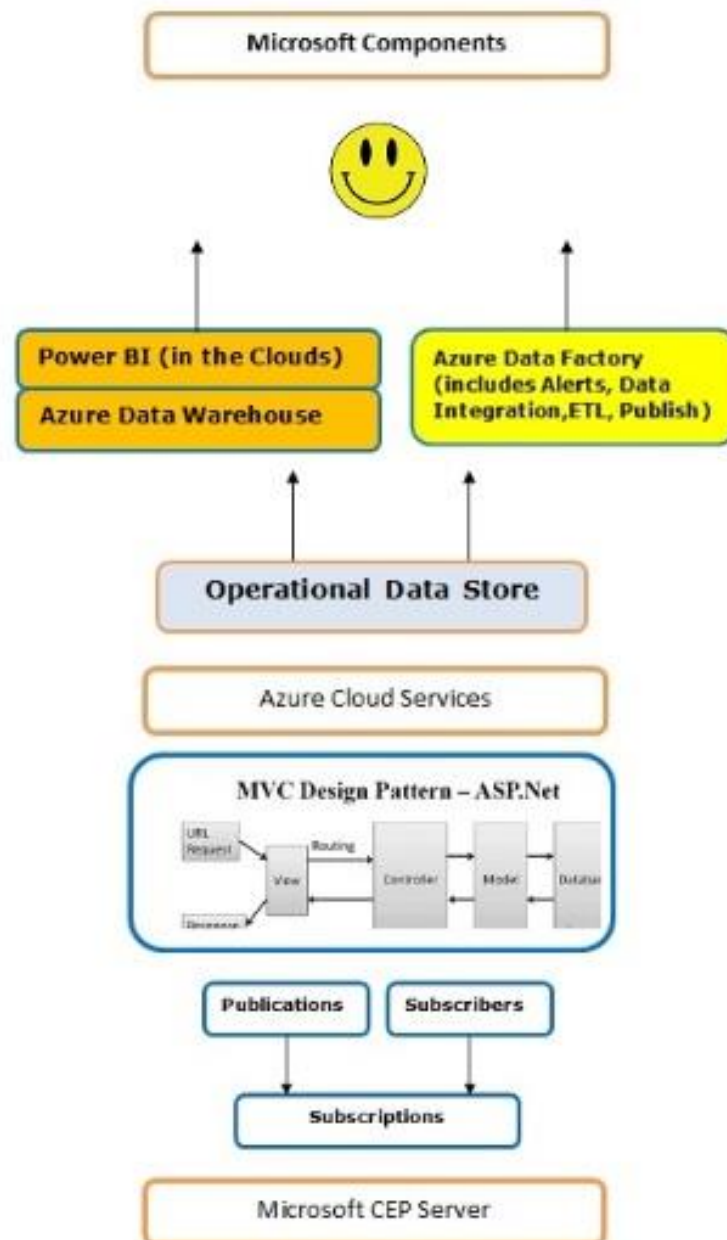
14.1 Microsoft Component Architecture

Here we show the Microsoft Components that we plan to use to implement the Reference Data Architecture.

The Microsoft documentation states that both the Azure Data Factory and Power BI are alternative ways of achieving the same objective of publishing data from Data Sources to the end-user. We will resolve this issue within a week.

We have discussed earlier the apparent overlap of the functionality offered by the Azure Data Factory and Power BI.

We look forward to resolving this overlap within a week.



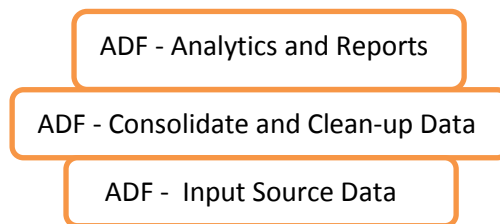
14.2 Generic Platform

There are three Components that apply to any Data Management activity :-



14.3 Azure Data Factory (ADF)

ADF can be used to implement all three Components :-



In this document entitled “What is Azure Data Factory”

- <https://docs.microsoft.com/en-us/azure/data-factory/data-factory-introduction>

Microsoft states :-

“Data Factory orchestrates and automates the **movement** and **transformation** of data” and “you can create data integration solutions to publish result data”.

A good introduction :-

- <https://azure.microsoft.com/en-gb/services/data-factory/>

and here's the architecture :-



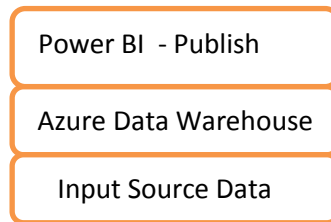
This is a summary of what Microsoft Web Site says :-

“Data Factory is a cloud-based data integration service for the **movement** and **transformation** of data. It can integrate data from various data stores, transform the data, and publish result data to the data stores.

It also provides rich visualizations to display the lineage between your data pipelines, and monitor the pipelines from a single unified view to easily pinpoint issues and setup monitoring **alerts**.”

14.4 Power BI

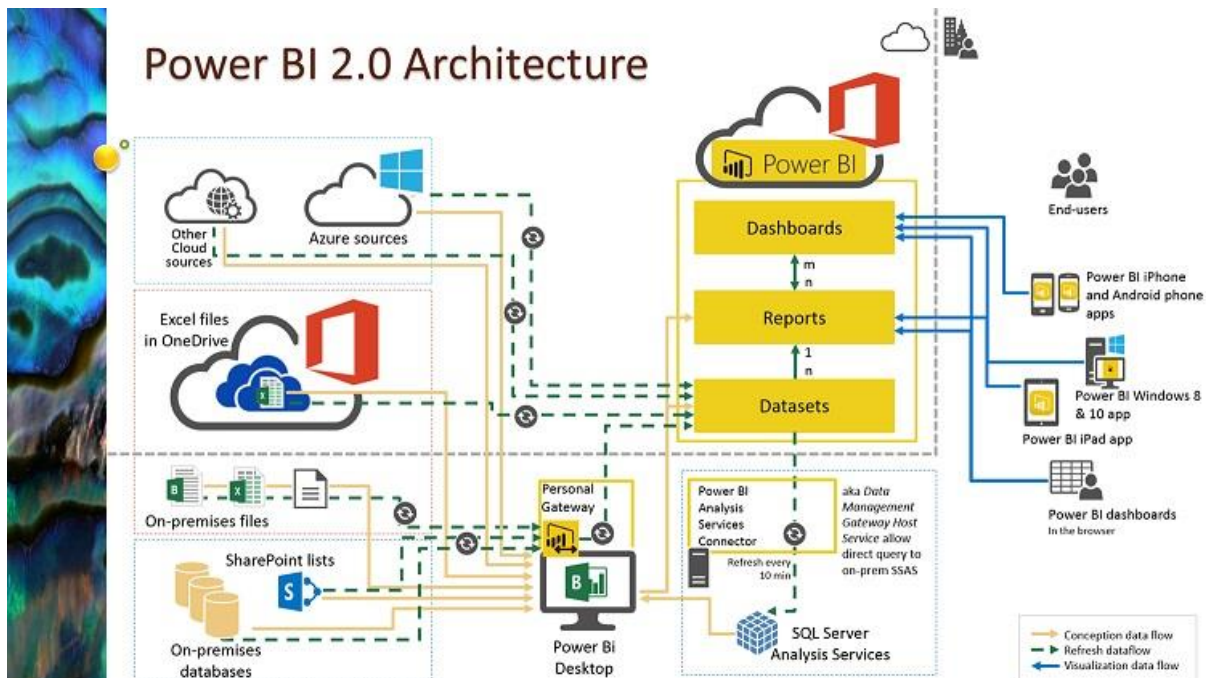
Power BI can be used to implement Analytics and Reports Component with Azure Data Warehouse Providing the Consolidate and Data Clean-up Data :-



Here's a good Architecture diagram :

- <https://www.itunity.com/article/power-bi-2-part-2-architectural-aspects-2339>

and looks like this :-



The text says :-

“In this picture, we can see from left to right, the sources, service and destinations:

- **The sources:** there are three types of sources in Power BI: files, databases and services
- **The service:** you can define three types of objects: datasets, reports and dashboards
- **The destinations:** dashboards and reports can be consumed in three ways

Data Sources are separated in three categories:

1. Files (which formats are Excel and Power BI)
2. Services, which are prepackaged models of data for popular services (like Google Analytics or GitHub)
3. Big Data and More (for Azure DBs and SQL Server Analysis on-premises)

Power BI 2.0 supports a new direct connectivity to Apache Spark (especially suited for Big Data scenarios) . Query performance over a Hadoop dataset can be 100 times faster with Spark.”

On this page :-

- <https://powerbi.microsoft.com/en-us/>

Microsoft states :-

“Power BI is a suite of **business analytics** tools to analyze data and share insights. Monitor your business and get answers quickly with rich dashboards available on every device”

And on this page :-

- <https://powerbi.microsoft.com/en-us/what-is-power-bi/>

It says (among other things) :-

“**Power BI Desktop** is a feature-rich data mashup and report authoring tool [which] combines data from disparate databases, files, and web services with visual tools that help you understand and fix data quality and formatting issues automatically.”

14.5 SQL Server

Connecting to SQL Server :-

- [https://technet.microsoft.com/en-us/library/ms190944\(v=sql.105\).aspx](https://technet.microsoft.com/en-us/library/ms190944(v=sql.105).aspx)

14.6 Microsoft Azure

A SQL Server edition in the Clouds is available called Azure SQL Data Warehouse.

It is architected for analytical workloads and can interface with unstructured data in an Azure Data Lake.

Here is a Microsoft reference to the Azure SQL Data Warehouse :-

- <https://azure.microsoft.com/en-us/services/sql-data-warehouse/>

and to the Data Lake :-

- <https://azure.microsoft.com/en-us/solutions/data-lake/>

There is also an Azure SQL Database which offers transactional consistency and high concurrency :-

- <https://azure.microsoft.com/en-us/services/sql-database/>

14.7 Jen Stirrup

Jen Stirrup is a Microsoft Guru worth following -

- <https://jenstirrup.com/>

Jen Stirrup on Power BI :-

- <https://jenstirrup.com/category/powerbi/>

15. Proof-of-Concept – Airport Management

Here we list some pages that give an insight into the POC.

A top-level Conceptual Model :-

- http://www.databaseanswers.org/data_models/airport_management/index.htm

And a Flights Dimensional UML Class Diagram :-

- http://www.databaseanswers.org/data_models/airport_management/flights_dimensional_model.htm

This link a Slide-Show showing each Step in the POC :-

- http://www.databaseanswers.org/slide_shows/Run_KPI_Trigger_for_Heathrow/slideshow_index.htm

On this Page, we show Data Models and test data in stages in the POC :-

- http://www.databaseanswers.org/data_models/POC_air_transport_platform_2020/print_version.htm

16. Conclusion

Our intention with this document is to publish Best Practice in Enterprise Data Management combined with Templates that demonstrate how Best Practice is applied in practical.