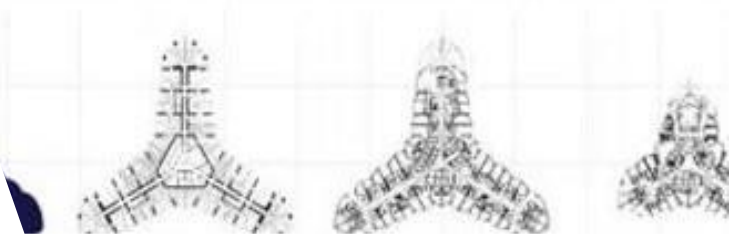
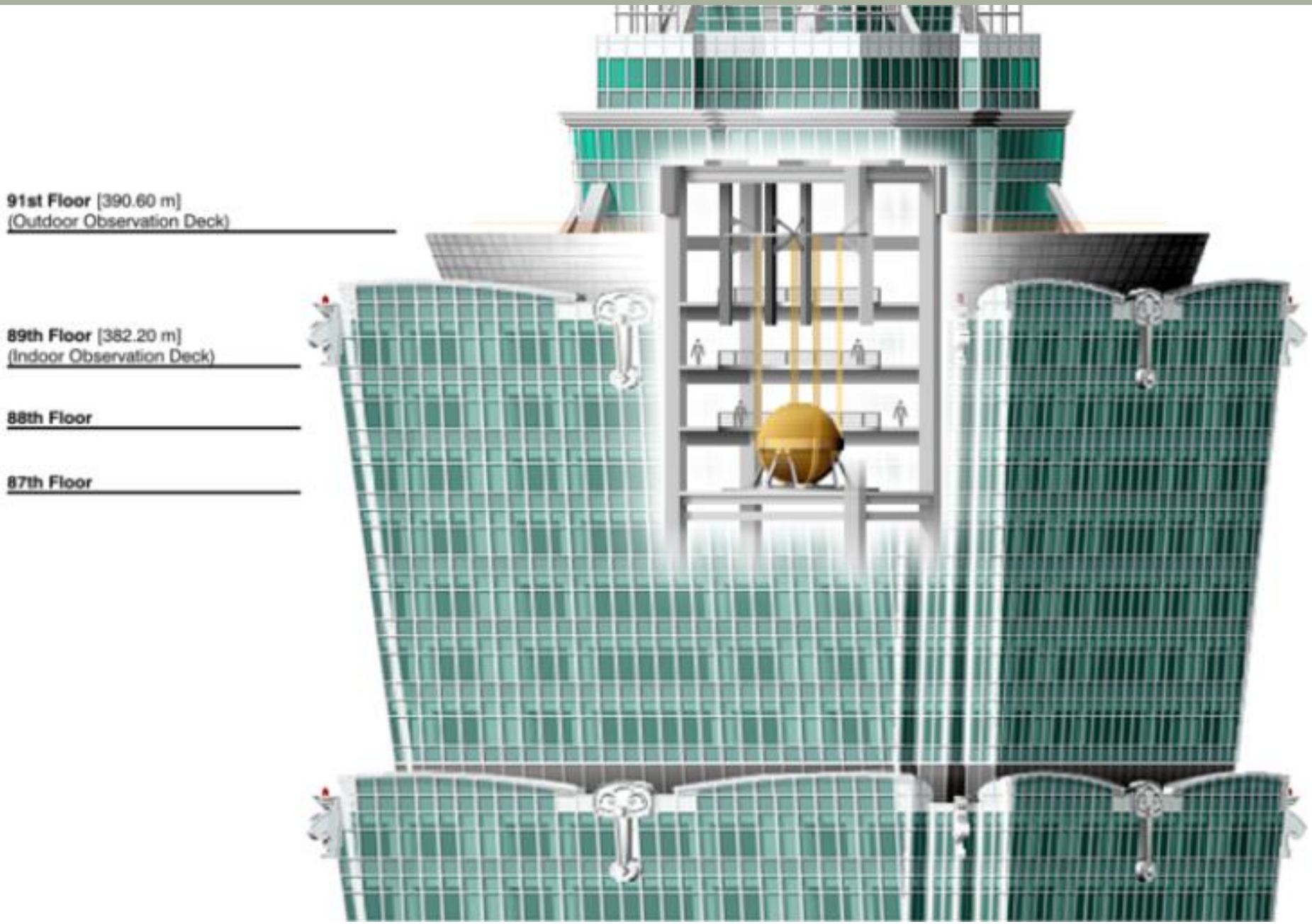
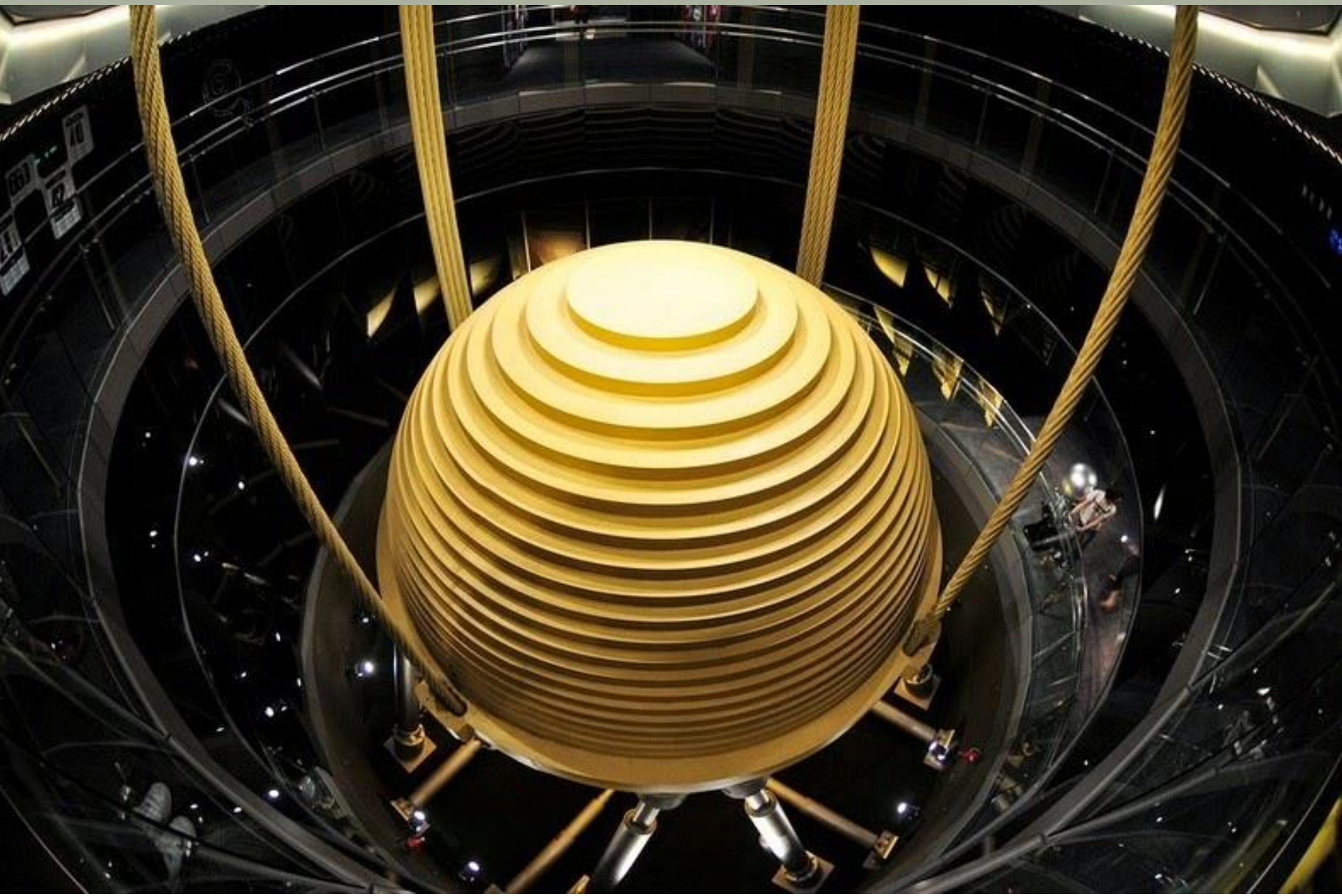
MIS781 Business Intelligence and Database

Module 7: Data Warehouse Architecture & Development

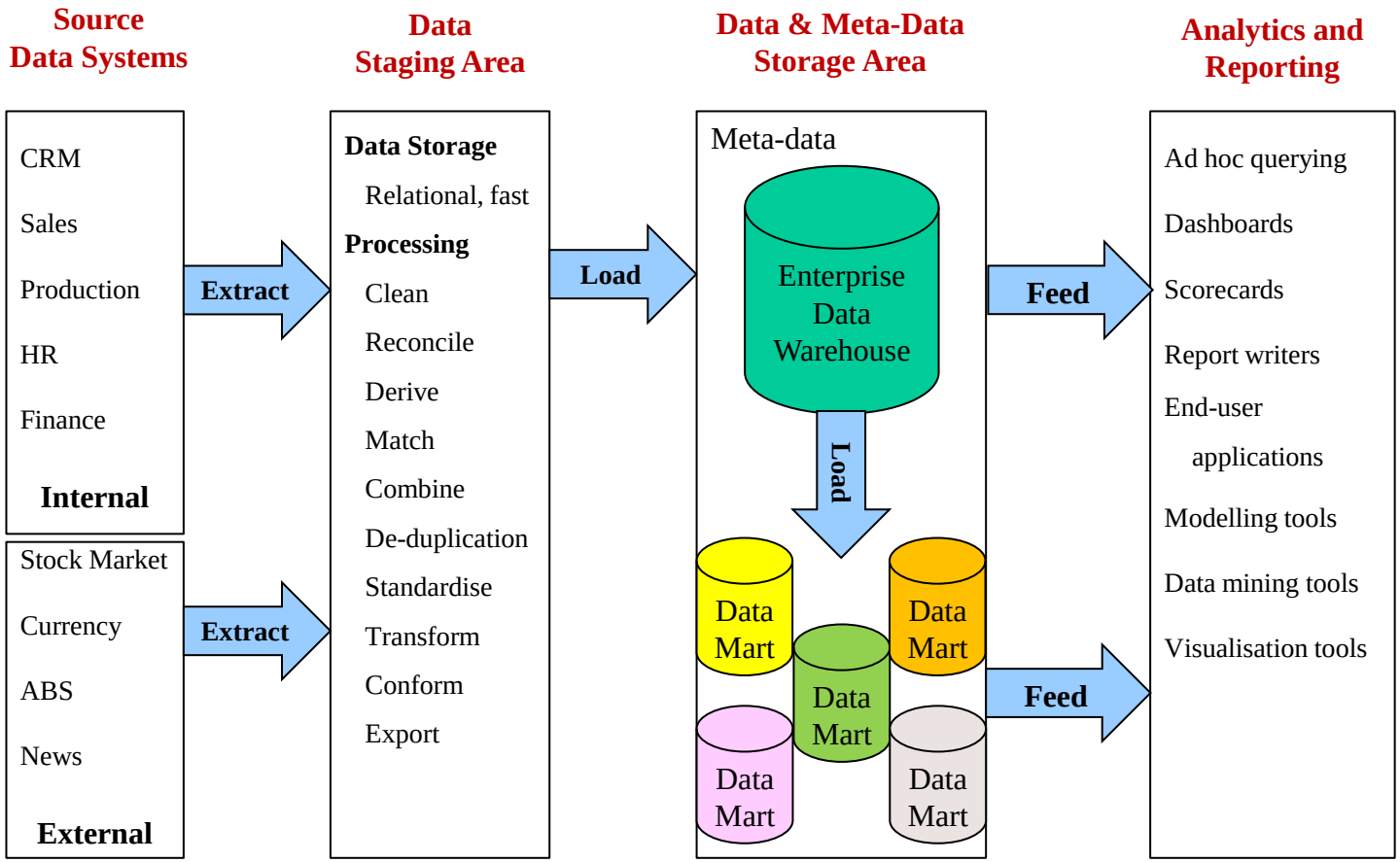



Asia
Japan
North America
Ring of Fire
Equator
South America
Australia
Pacific Ocean
Antartica


91st Floor [390.60 m]
(Outdoor Observation Deck)
89th Floor [382.20 m]
(Indoor Observation Deck)
88th Floor
87th Floor

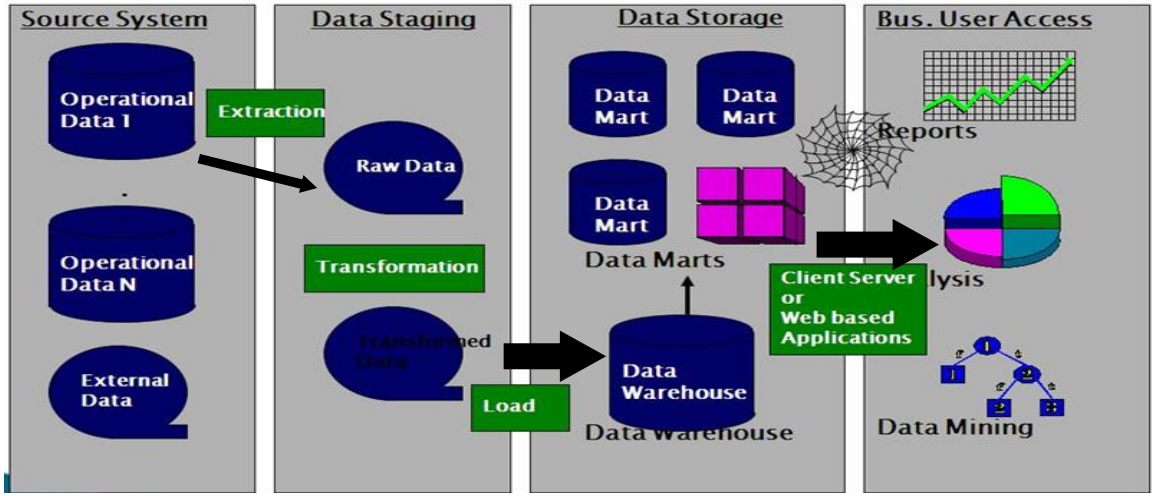


Data Warehouse Architecture



Data Warehouse Architecture

- Three parts of the data warehouse
  1. The **data warehouse** that contains the data and associated software
  2. Data acquisition (**back-end**) software that extracts data from legacy systems and external sources, consolidates and summarises them, and loads them into the data warehouse
  3. Client (**front-end**) software that allows users to access and analyse data from the warehouse



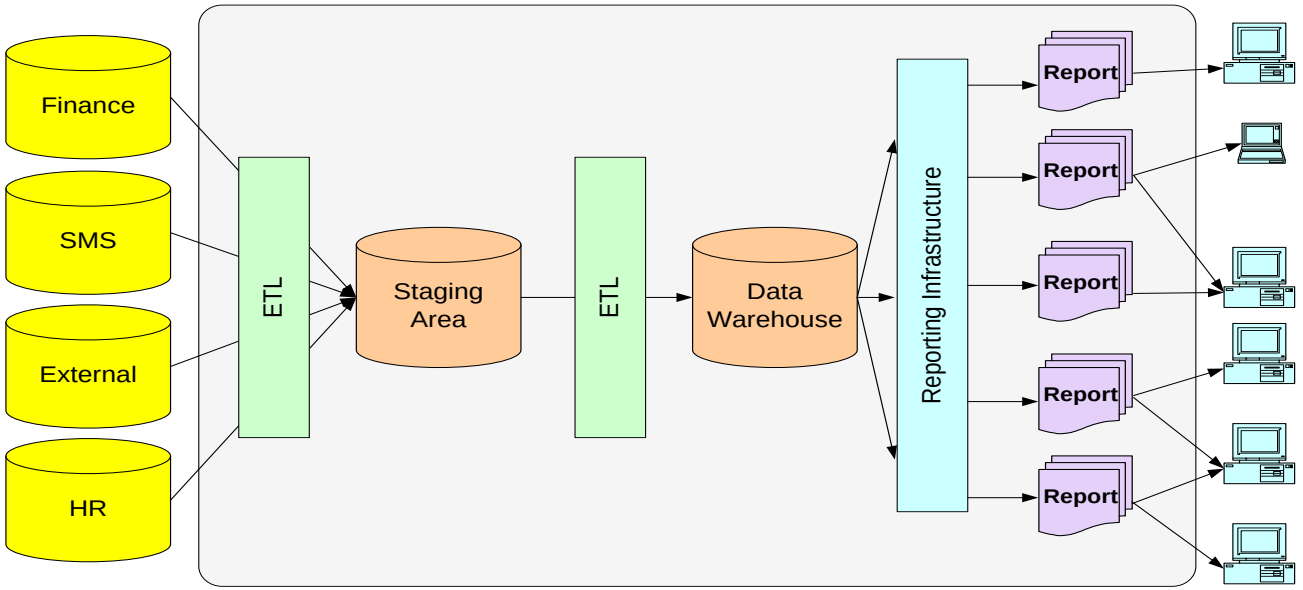
Benefits of a Sound DW Architecture

1. Provides an organising framework – shows where the components are and how they fit
2. Improved flexibility and maintenance – allows plug and play, permits quick addition of new data sources
3. Faster development and reuse – developers are better able to understand the DW process and content
4. Management and communications tool – sets our expectations and defines our responsibilities
5. Coordinate parallel efforts – multiple independent efforts have a better chance to converge

What's The Best DW Architecture?

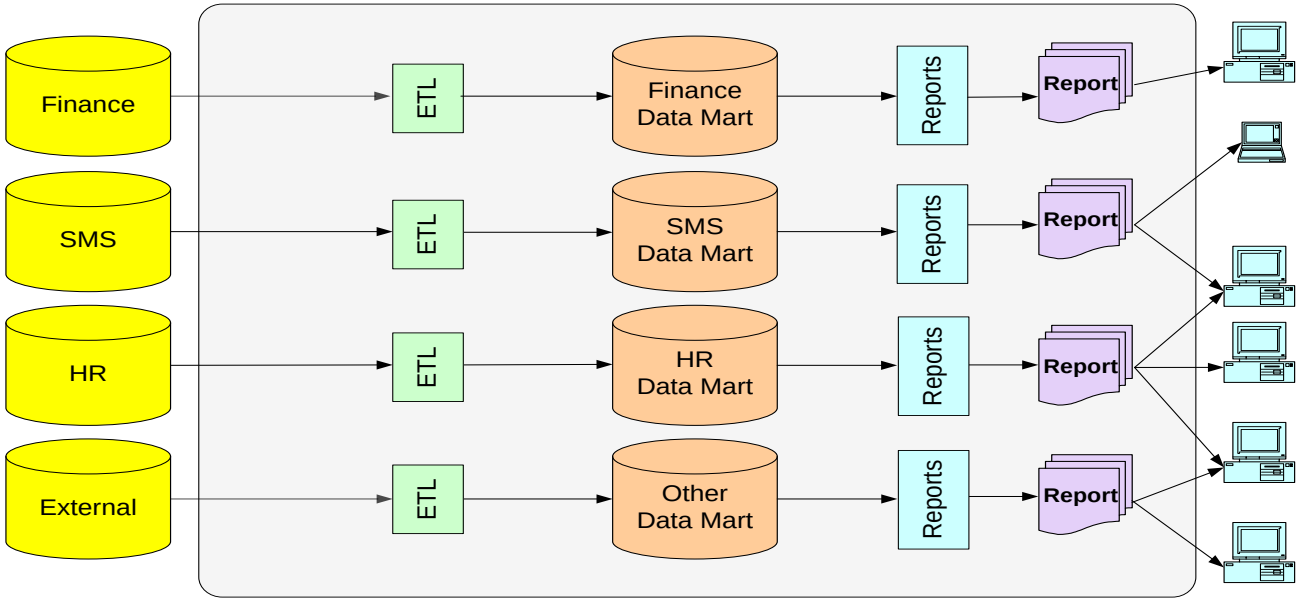
- There are different types of DW architecture…
- The basic types are
  - Independent Data Marts
  - Dependent Data Marts / Hub and Spoke
  - Bus Architecture
  - Federated
  - And more…

➢1. Central Data Warehouse



Highly variable with many individual approaches

➢2. Independent Data Marts

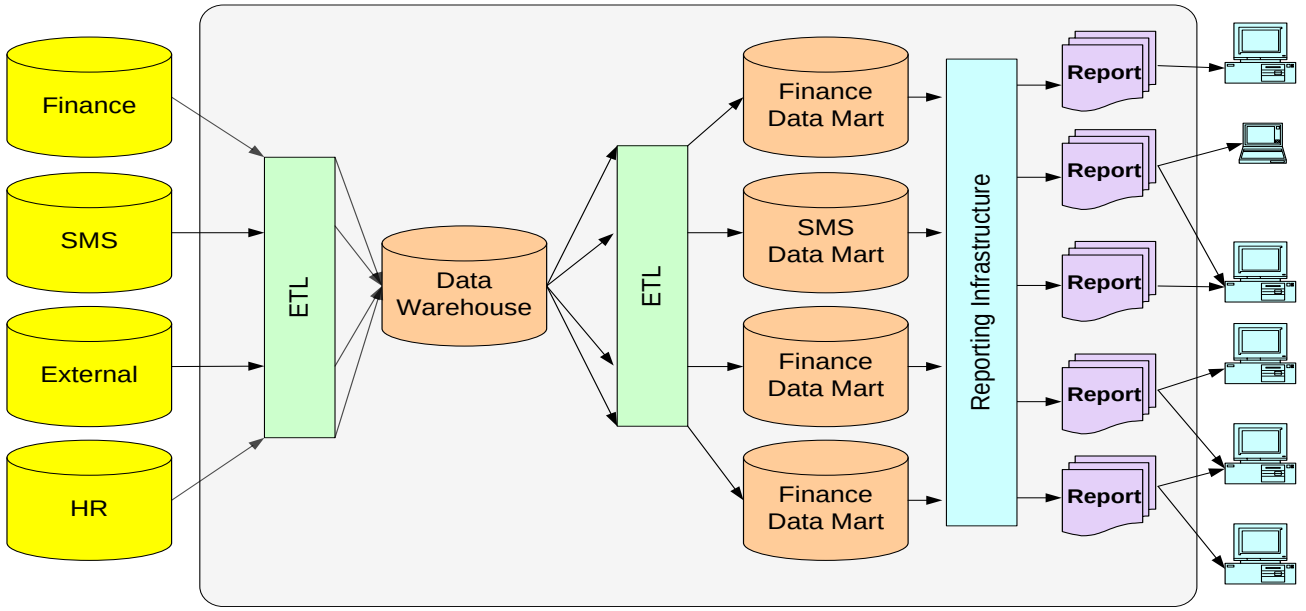


How Data Warehousing was often performed in the early days

- Individual projects developing solutions into functional silos
- No program / enterprise perspective
- No conformed dimensions

3. Dependent Data Marts / Hub & Spoke

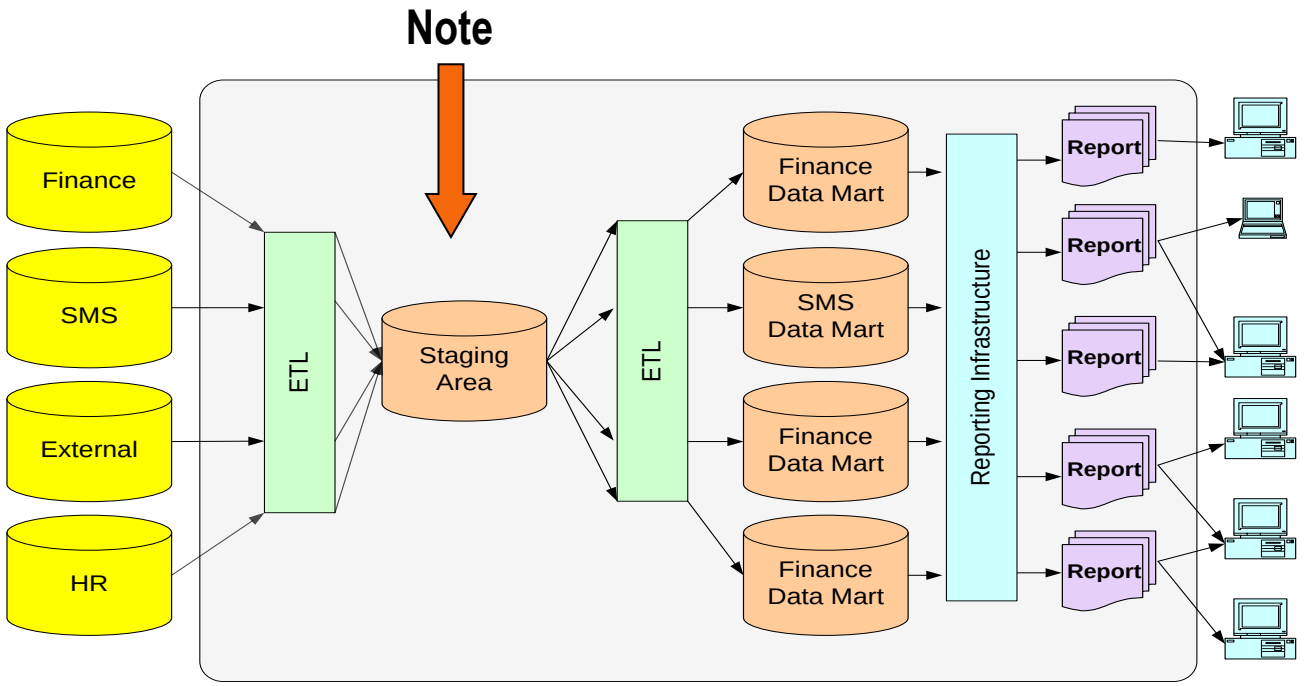
Usually employing a *Top-Down* approach



An approach also used in the early days, but refined over time

➢4. Data Mart Bus

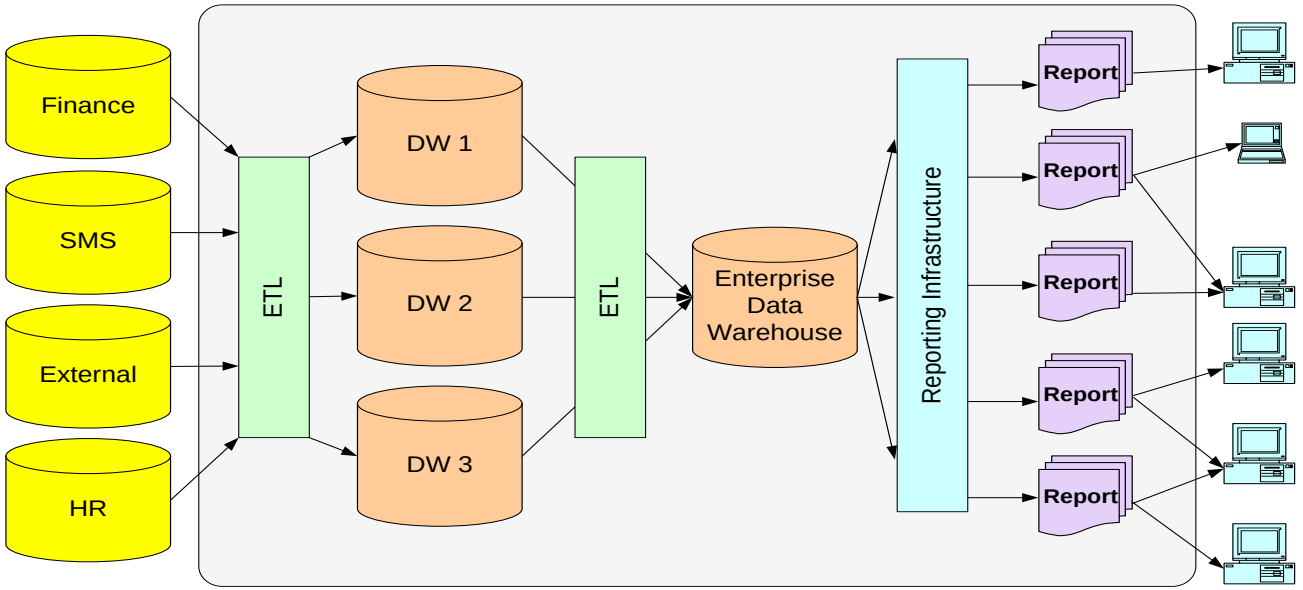
Usually employing a *Bottom-Up* approach



An approach also used in the early days, but refined over time

- Originally suggested building silos
- Now recommends enterprise perspective

➢5. Federated Data Warehouse



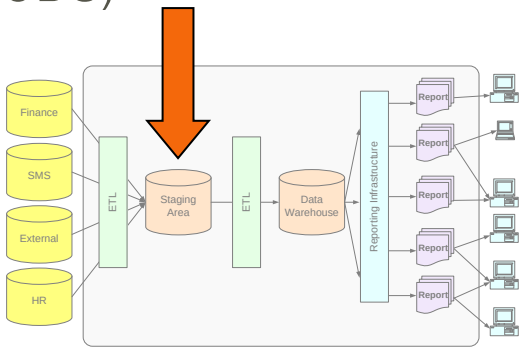
An attempt to consolidate legacy Data Marts

The Staging Area

- What is a staging area?
  - A **copy** of operational data from the transaction system

- Why would you build a staging area?
  - Less impact of loads on transaction system
  - 'Snapshot' allows repeatable DW ETL processes
  - Area to store 3rd party data
  - Employ data manipulation not appropriate for the DW
  - Foundation of a future Operational Data Store (ODS)

- Why not?
  - Data redundancy
  - Additional development time



Factors affecting DW architectures

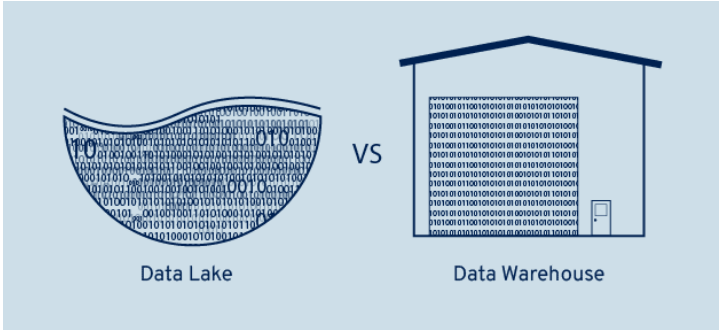
Ten factors that potentially affect the architecture selection decision:

1. Information interdependence between organisational units
2. Upper management's information needs
3. Urgency of need for a data warehouse
4. Nature of end-user tasks

5. Constraints on resources
6. Strategic view of the data warehouse prior to implementation
7. Compatibility with existing systems
8. Perceived ability of the in-house IT staff
9. Technical issues (e.g. DW vs Data Lake)
10. Social/political factors

DW vs Data Lake

Comparison	Data Warehouse	Data Lake
Data	Structured, processed data	Structured/semi-structured, unstructured data, raw data, unprocessed data
Processing	Schema-on-write	Schema-on-read
Storage	Expensive, reliable	Low cost storage
Agility	Less agile, fixed configuration	High agility, flexible configuration
Security	Matured	Maturing
Users	Business professional	Data Scientists (especially those familiar with domain)



DW vs Data Lake: https://www.youtube.com/watch?v=AwbKwcw7bgg

DB vs DW vs Data Lake: https://www.youtube.com/watch?v=-bSkREem8dM

Quiz

- Great Country consists of 7 states. The education department of each state operates under its respective state government and maintains its own education database. Now the federal education ministry of the country decides to have a national education data warehouse for its national-level BI system that is accessible by the officers of the federal and state education departments.
- Moreover, the BI system is also accessible by the federal ministry of finance. The finance minister of the country has allocated a budget of $300 million dollars and a project timeframe of 5 years for the BI endeavour.

- Propose the most appropriate DW architecture.

Slowly Changing Dimension (SCD) concept

- "Slowly Changing Dimension" is a common issue in data warehousing, because attribute for a record varies over time

E.g.:
Christina is a customer with XYZ Inc. She first lived in Chicago, Illinois. So, the original entry in the customer lookup table has the following record:

Customer Key	Name	State
1001	Christina	Illinois

At a later date, she moved to Los Angeles, California on 1 January, 2016. How should XYZ Inc now modify its customer table to reflect this change? This is the SCD problem.

Source: http://www.1keydata.com/datawarehousing/slowly-changing-dimensions.html

Slowly Changing Dimension (SCD) Types

- Data Warehouse designers have sorted out **three major approaches to SCDs**. These are called TYPE 1, TYPE 2 and TYPE 3.

1. A **Type 1** SCD is an **overwrite** of a dimensional attribute. The new record replaces the original record. No trace of the old record exists.

2. A **Type 2** SCD **creates a new record** for each change. A new record is added into the customer dimension table. Therefore, the customer is treated essentially as two people.

3. A **Type 3** SCD **adds a new field** in the dimension record but does not create a new record. The original record is modified to reflect the change

Customer Key	Name	State
1001	Christina	Illinois

Read: http://www.1keydata.com/datawarehousing/slowly-changing-dimensions.html

SCD Example

Type 1 SCD

Customer Key	Name	State
1001	Christina	California

Type 2 SCD

Customer Key	Name	State
1001	Christina	Illinois
1005	Christina	California

Type 3 SCD

Customer Key	Name	Original State	Current State	Effective Date
1001	Christina	Illinois	California	1 January, 2016

Slowly changing dimension (SCD) – Example 2

E.g. 2:

- Universities (and other organisations) are not static. Faculties are created/disbanded, schools are opened/closed. Courses are modified, Campuses open/close.
- People want to see their data with the relationships which existed at the time it was current – eg the school which is now closed.
- People also want to see their data with the relationships which exist today – eg units history – regardless of the fact it was in a different school.
- People want to see data in ways they haven't thought of yet!

Example of a real-world problem

- We have a school – Health and Behaviour sciences (HBS).
- The school has two units – HL84 (Health), and BS92 (Behaviour Science).
- HL84 started with 50 EFTSL in 2000 and BS92 started with 60 EFTSL. Each increased by 5 EFTSL a year.

Simple Load Report by School

	2000	2001	2002	2003	2004
HBS	110	120	130	140	150

Example of a real-world problem

- The School is now split into Health (HHL) and Behaviour Sciences (BSS) from 2005. The Two units are allocated accordingly to the new schools.

The Simple Load Report by School now looks like

	2000	2001	2002	2003	2004	2005
HBS	110	120	130	140	150	
HHL						75
BSS						85

Example of a real-world problem

- But of course, while the users/clients agree that the report is accurate.. What they really want is..

A "amended history" Simple Load Report by School

	2000	2001	2002	2003	2004	2005
HHL	50	55	60	65	70	75
BSS	60	65	70	75	80	85

- Oh.. But don't change the data or anything.. We might need to report it the other way as well!

Type 1,2,3 Problems

- These three types of slowly changing dimensions handle most of the situations faced by the DW Designer.

Like this….

	2000	2001	2002	2003	2004	2005
HBS	110	120	130	140	150	
HHL						75
BSS						85

Or like this….

	2000	2001	2002	2003	2004	2005
HHL	50	55	60	65	70	75
BSS	60	65	70	75	80	85

Using Standard Type 2 to represent the data

- If we use a standard type 2 approach to represent this data we would have the following.

Dimension Table

Key	Code	Description
1	HBS	Health and Behaviour sciences
2	HHL	Health
3	BSS	Behaviour Sciences

Fact Table

Key	Year	Course	EFTSL
1	2000	HL84	50
1	2000	BS92	60
1	2001	HL84	55
1	2001	BS92	65
1	2002	HL84	60
1	2002	BS92	70
1	2003	HL84	65
1	2003	BS92	75
1	2004	HL84	70
1	2004	BS92	80
2	2005	HL84	75
3	2005	BS92	85

Using Standard Type 3 to represent the data

- We could use a type 3 dimension and store the Original value of the records.. Eg..

Key	Code	Desc	Old Code	Old Desc
1	HBS	Health and Behaviour sciences	HBS	Health and Behaviour sciences
2	HHL	Health	HBS	Health and Behaviour sciences
3	BSS	Behaviour Sciences	HBS	Health and Behaviour sciences

The Data Dictionary

- collection of names, definitions, and attributes about data elements that are being used in a database

Data Dictionary

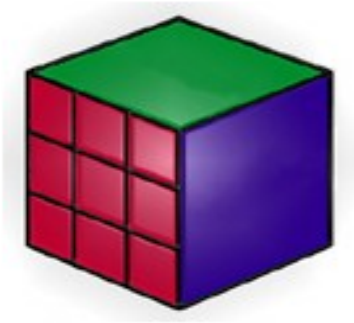
Data Dictionary outlining a Database on Driver Details in NSW

Field Name	Data Type	Data Format	Field Size	Description	Example
License ID	Integer	NNNNNN	6	Unique number ID for all drivers	12345
Surname	Text		20	Surname for Driver	Jones
First Name	Text		20	First Name for Driver	Arnold
Address	Text		50	First Name for Driver	11 Rocky st Como 2233
Phone No.	Text		10	License holders contact number	0400111222
D.O.B	Date / Time	DD/MM/YYYY	10	Drivers Date of Birth	08/05/1956

Example of EDW Glossary

Glossary

The following is a list of terms that are used to qualify the presentation of information found in cubes and reports. The **primary system** indicates where the term originated from and the **type** indicates how the term is used.



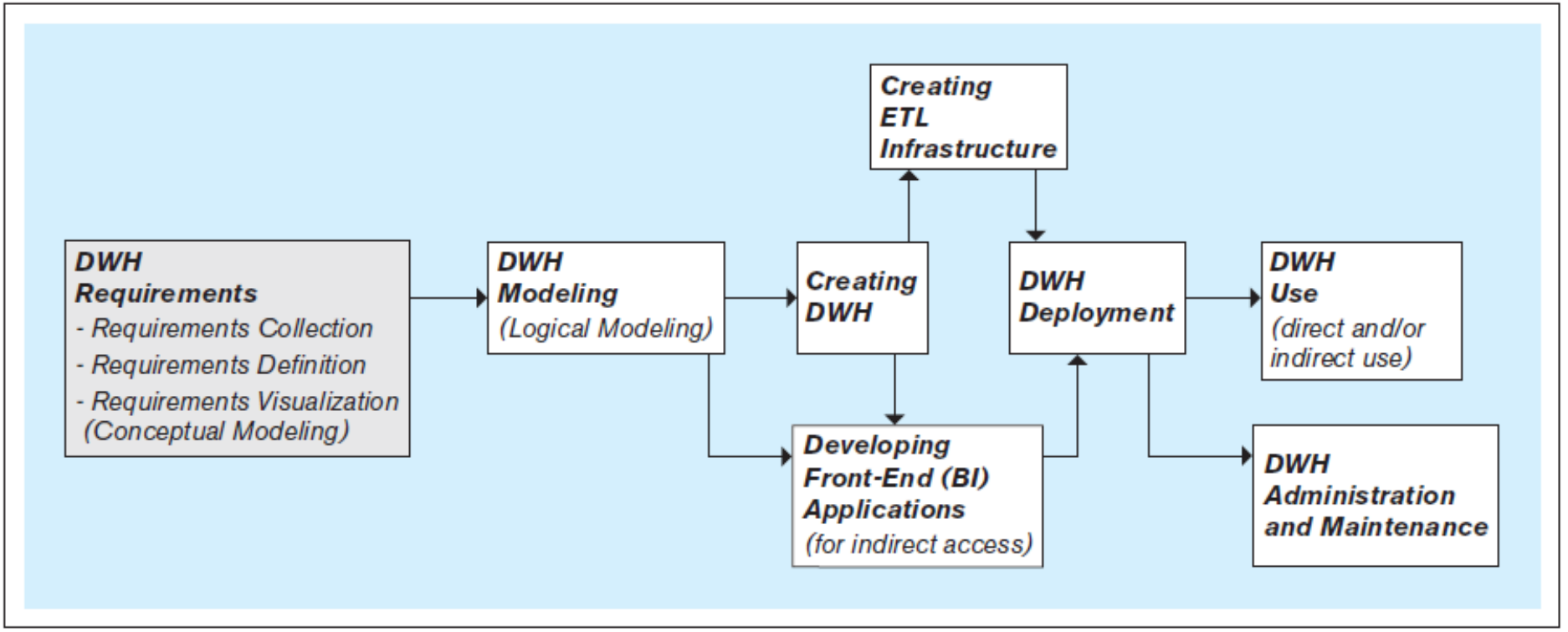
Logon here

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | XYZ

top^ **A**

Name	Primary system/s	Comments	Type
Acceptances	International Office	International onshore students who have accepted a University place.	Measure
Account	Human Resources		Dimension
Activity	Finance	Code to breakdown expenditure against certain types of activities.	Dimension
Activities (GDS)	PAS	In the GDS graduates are asked if they are working or have gone on to further study. 'Activities' provides a breakdown of the graduates' employment status.	Dimension

Steps in the Development of DW



e.g. Deakin SIPU ; https://planning.curtin.edu.au/

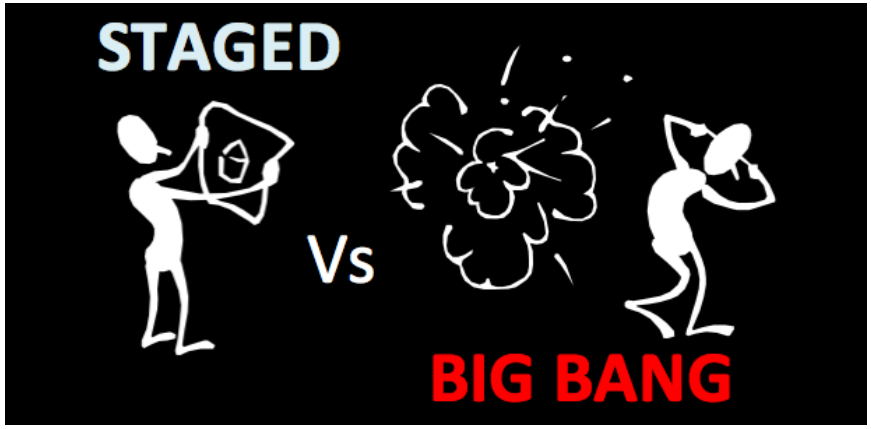
https://i.unisa.edu.au/staff/business-intelligence-and-planning/

Developing a DW Technical Architecture

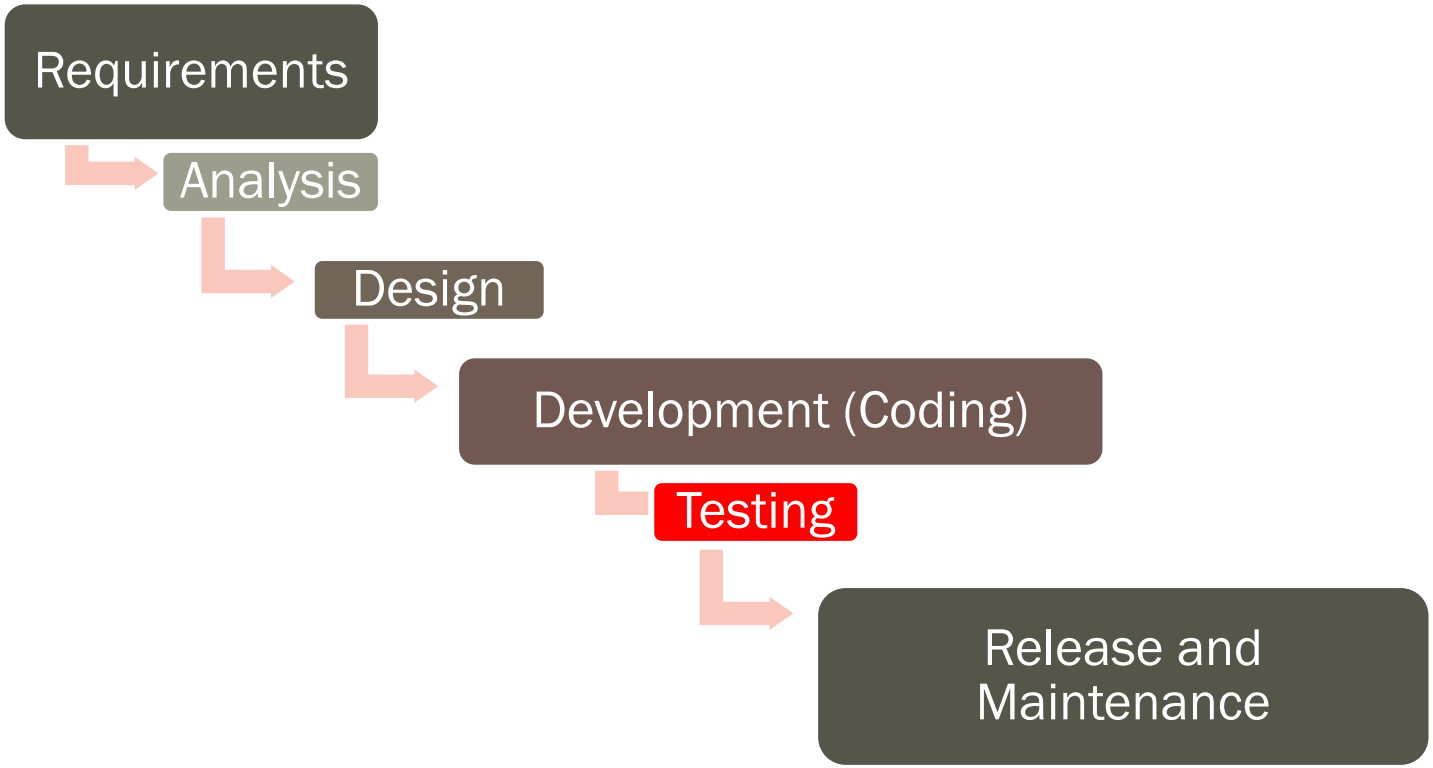
Developing a DW architecture is a difficult task and needs to be faced with the proper approach.

"Big Bang" Development Approach (Inmon)

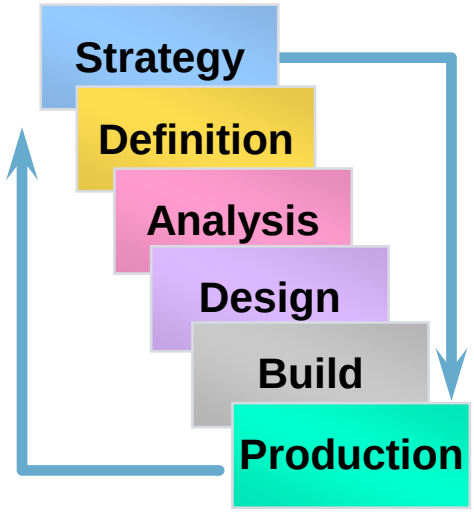
- Advantages:
  - warehouse built as part of major project (eg: BPR)
  - Having a "big picture" of the data warehouse before starting the data warehousing project
- Disadvantages:
  - Involves a high risk, takes a longer time
  - Runs the risk of needing to change requirements
  - Costly and harder to get support for from users



“Big Bang” Waterfall



Incremental Approach to Warehouse Development (Kimball)



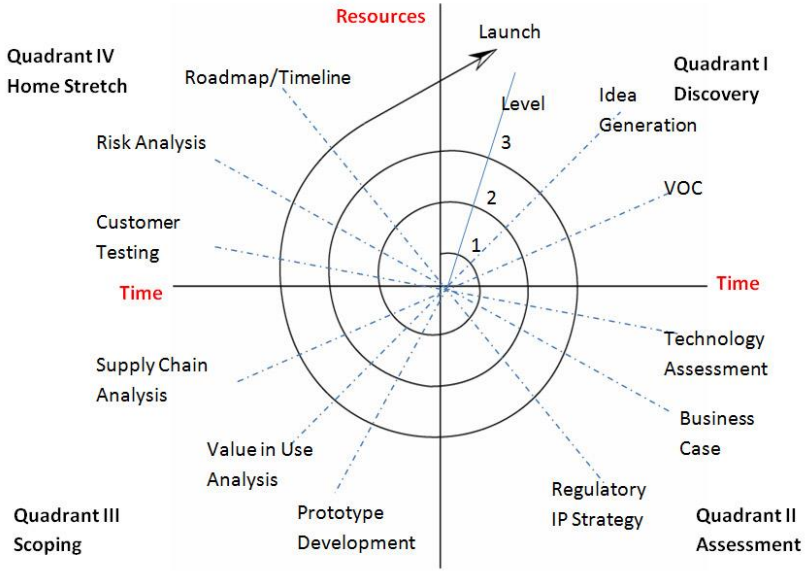
- Multiple iterations
- Shorter implementations
- Validation of each phase

- The incremental approach manages the growth of the data warehouse by developing incremental solutions that comply with the full-scale data warehouse architecture.

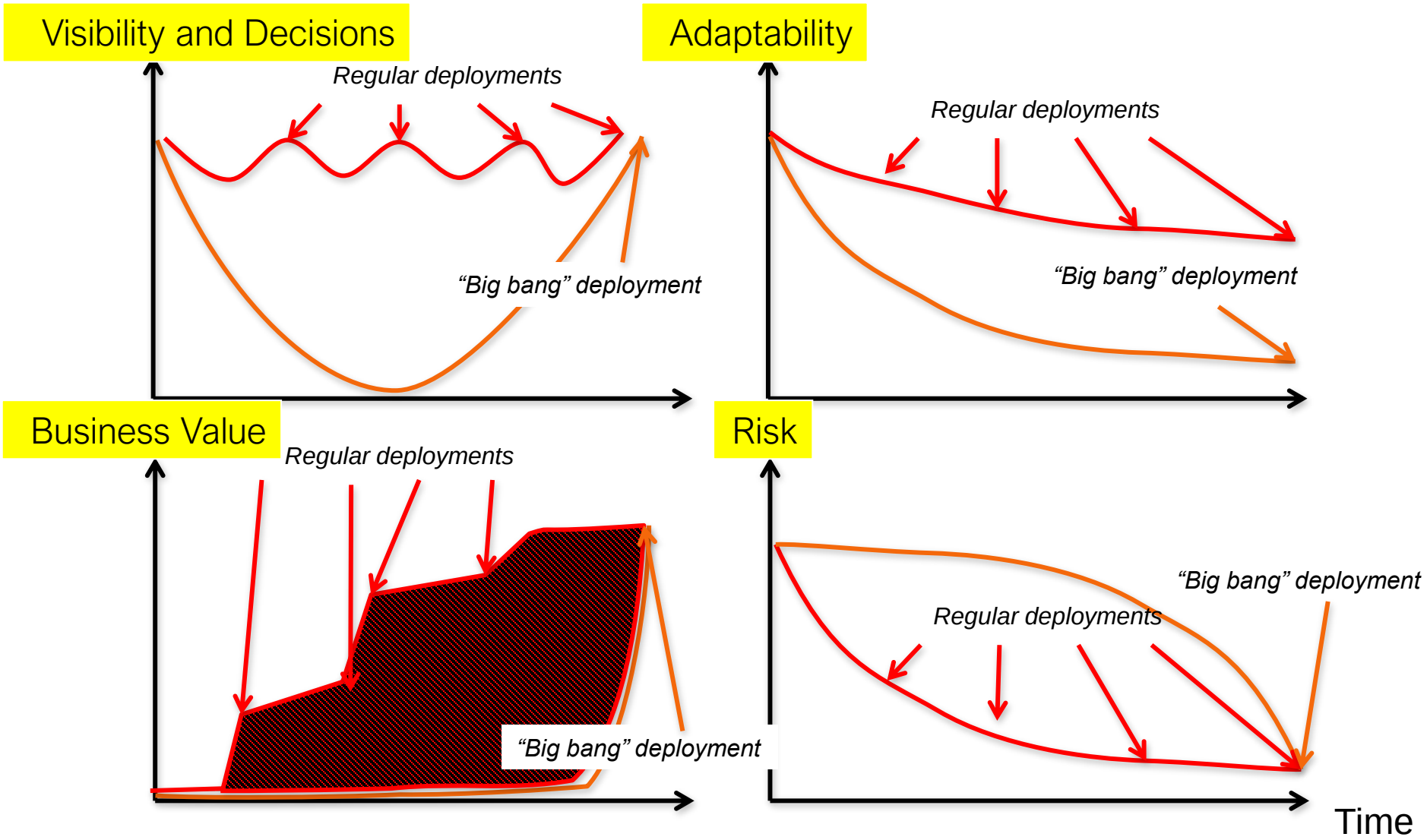
- **Think big and start small.** In other words, your strategy identifies the enterprise-wide warehouse which is delivered by small increments, in short timeframes.

Benefits of an Incremental Approach

- Delivers a strategic data warehouse solution through incremental development efforts
- Provides extensible, scalable architecture
- Quickly provides business benefits and ensures a much earlier return of investment
- Allows a data warehouse to be built based on a subject or application area at a time



Why Agile/Incremental Approach?



Big Bang (Inmon) vs Incremental (Kimball)

	Inmon	Kimball
Building Data warehouse	Time Consuming	Takes lesser time
Maintenance	Easy	Difficult, often redundant and subject to revisions
Cost	High initial cost; Subsequent project development costs will be much lower	Low initial cost; Each subsequent phase will cost almost the same
Time	Longer start-up time	Shorter time for initial set-up
Skill Requirement	Specialist team	Generalist team
Data Integration requirements	Enterprise-wide	Individual business areas

- No one-size-fits-all strategy to data warehousing!

Best practices for implementing DW

- Some best practices for implementing a data warehouse:

1. Project must fit with corporate strategy and business objectives
2. There must be complete buy-in to the project by executives, managers, and users
3. It is important to manage user expectations about the completed project
4. The data warehouse must be built incrementally
5. Build in adaptability
6. The project must be managed by both IT and business professionals
7. Develop a business/supplier relationship
8. Only load data that have been cleansed and are of a quality understood by the organisation
9. **User participation** in the development of data and access modeling is a critical success factor in data warehouse development
10. Be politically aware

Real DW https://www.youtube.com/watch?v=y5-3Pjbk8Zk

Three Bases of DW project justification (3F)

Facts

- "The data warehouse project will have a net present value of $753,000."
- "The data warehouse will yield a minimum reduction in operating cost of $193,000 annually."
- "The estimated increase in market share is 14.7 percent within the first 24 months of operation."

Faith

- "A data warehouse is part of the infrastructure, we can't cost justify it like a new fleet of trucks."
- "It seems reasonable to assume that this data warehouse will reduce our costs of servicing this market segment."
- "Trust me. This is why you hired me as BI director."

Fear

- "Our competitors are doing this even as we speak."
- "Our shareholders will view us as technologically behind if we don't do this now."
- "We have a small window of opportunity here and we are wasting precious time trying to decide."

Economic Feasibility Analysis
- Tangible and Intangible Benefits

- This activity is a type of cost-benefit analysis, complicated by the fact that many benefits are intangible and hard to measure.
- Tangible benefits are either cost reductions or revenue increases.
- Intangible benefits include improved morale, increased product quality, decrease in time to market, reduction in turnover, increased competitive advantage and more timely information.

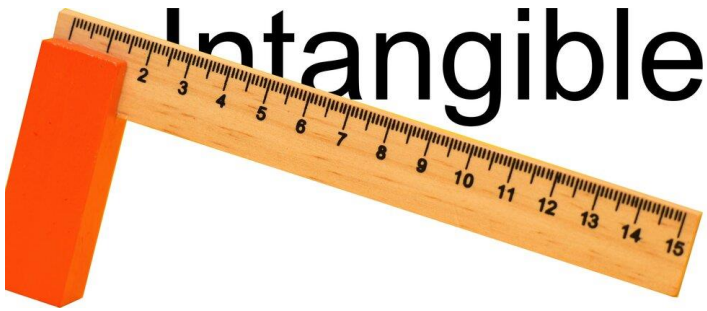
Tangible and Intangible Costs

- Tangible development costs include salaries, consultant fees, hardware and software purchases and data conversion.
- Tangible operating costs include annual licenses, upgrades, repairs, user training and depreciation.
- Intangible costs include disruption to environment, loss of goodwill, reduction in morale etc.

An Example of Measuring Intangibles

Suppose we believe that there is about a 60 percent chance that one out of every 10 workers will be distressed during the initial implementation. Our normal profits are $2,000,000, and we believe they would be zero if the entire workforce was distressed.

So, "distress" cost is: ??



An Example of Measuring Intangibles

Suppose we believe that there is about a 60 percent chance that one out of every 10 workers will be distressed during the initial implementation. Our normal profits are $2,000,000, and we believe they would be zero if the entire workforce was distressed.

Profit reduction is: (0.1)(2000000) = $200,000 and it has a 60% chance of occurring, so we estimate the "distress" cost at (0.6)(200000) = $120,000

Isn't it better to have an estimate based on reasonable assumption than making no estimate at all?

Spreadsheet Model of NPV Analysis

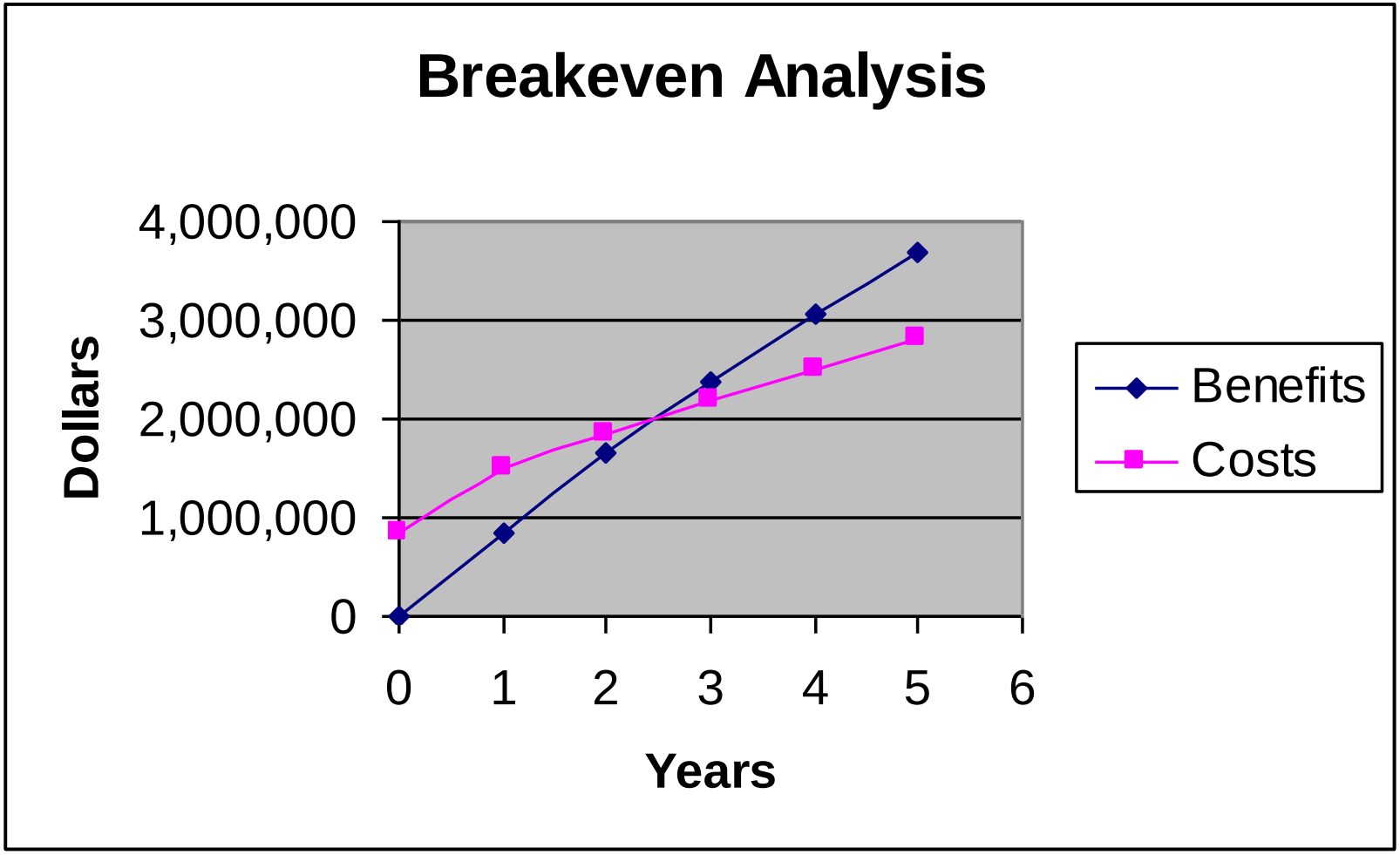
	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5	Totals
Net economic benefit		$920,168.30	$926,977.32	$934,286.92	$941,703.65	$949,228.60	
Discount rate (8.25%)	1.0000	0.9238	0.8534	0.7883	0.7283	0.6728	
PV of all benefits		$850,040.00	$791,067.06	$736,540.37	$685,808.14	$638,603.49	
NPV of all BENEFITS		$850,040.00	$1,641,107.06	$2,377,647.42	$3,063,455.56	$3,702,059.05	$3,702,059.05
One-Time COSTS	(831,579.65)	(322,659.82)					
PV of Equip. Deprec.		(56,086.46)	(51,811.97)	(47,863.25)	(44,215.47)	(40,845.70)	
Recurring Costs		(325,576.80)	(341,855.64)	(358,948.42)	(376,895.84)	(395,740.64)	
Discount rate (8.25%)	1.0000	0.9238	0.8534	0.7883	0.7283	0.6728	
PV of recurring costs		(300,763.79)	(291,733.93)	(282,975.17)	(274,479.38)	(266,238.66)	
NPV of all COSTS	(831,579.65)	(1,511,089.72)	(1,854,635.62)	(2,185,474.04)	(2,504,168.89)	(2,811,253.25)	(2,811,253.25)

Overall NPV	$890,805.80
Overall ROI	32%
IRR	38%

Breakeven Analysis						
Yearly NPV cash flow	(831,579.65)	549,276.21	499,333.13	453,565.20	411,328.76	372,364.83
Overall NPV cash flow	(831,579.65)	-661,049.72	-213,528.56	192,173.39	559,286.67	890,805.80

Project break-even occurs at 2.576 years

Example of Graphical Breakeven Analysis



Economic Feasibility Measures

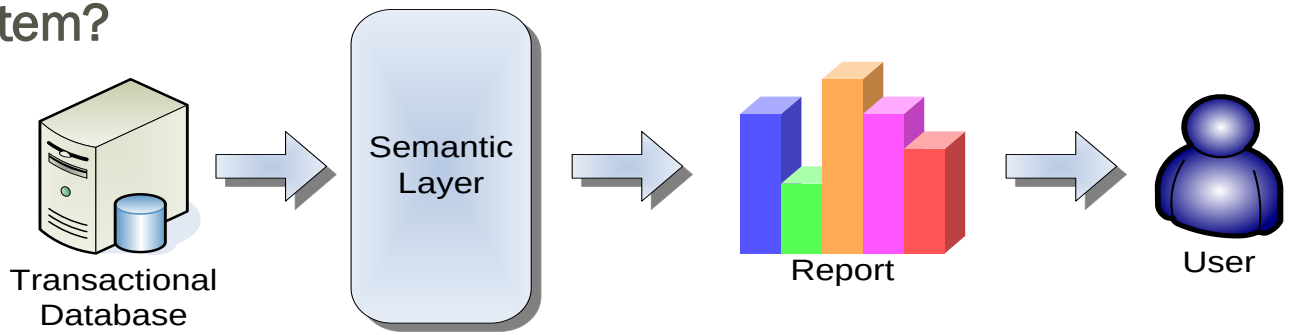
- Assessment of feasibility has to recognise the time value of money since many costs occur immediately while benefits accrue over several years.
- We show a project that returns a net present value of $890,806 over five years. The internal rate of return is 38% and breakeven is estimated at 2 years and 7 months.

Issues to consider

- Issues to consider to build a successful data warehouse:
  - Starting with the wrong sponsorship chain
  - Setting expectations that you cannot meet and frustrating executives at the moment of truth
  - Engaging in politically naive behavior
  - Loading the warehouse with information just because it is available
  - Believing that data warehousing database design is the same as transactional database design
  - Choosing a data warehouse manager who is technology oriented rather than user oriented
  - Focusing on traditional internal record-oriented data and ignoring the value of external data and of text, images, and, perhaps, sound and video

Summary: Benefits of Building a DW

- Why don't you just build reporting on top of the Transaction system?



- Has anyone ever tried this before?
  - Complex and time consuming to build
  - Slow to operate (poor performance)
  - Cumbersome to use (Ad hoc reporting is out of reach of the average user)
  - A nightmare to maintain
- It can be an almost impossible task, and doesn't always work!

Summary: Benefits of Building a DW

- A Data Warehouse lets you do stuff you wouldn't normally be able to do…

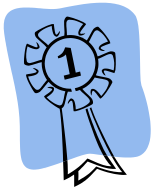
1. Speed

2. Ease of use

3. Ability to value add

4. Consolidate data

5. Consistency / single version of the truth



6. Manage data quality

Benefits of Building a DW

1. Speed

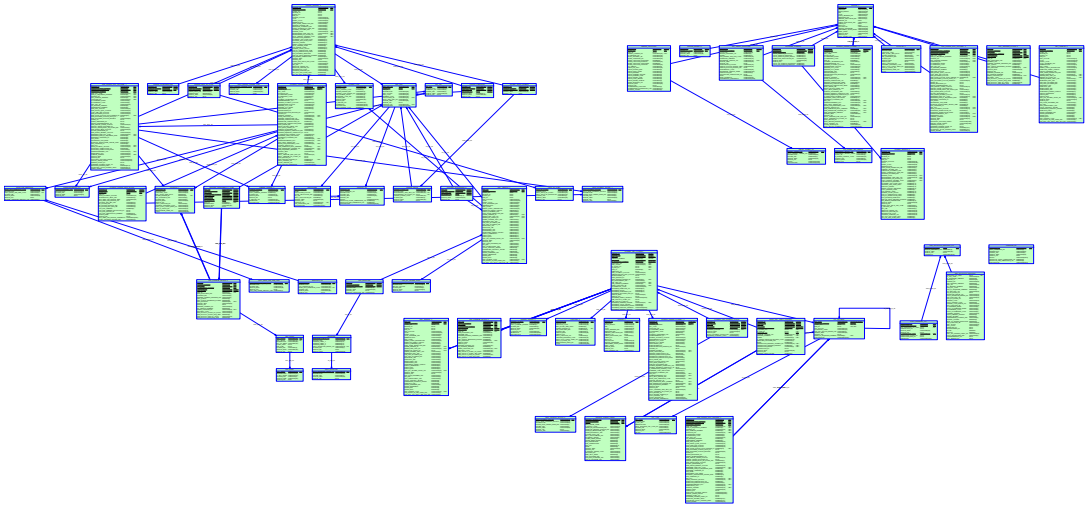
- A Data Warehouse is specifically designed and optimised for reporting
  - not to support the activities of a transaction system
  - OLTP systems are optimised for individual create, update or deletes

A Data Warehouse delivers many times greater performance than a normalised transactional DB

Benefits of Building a DW

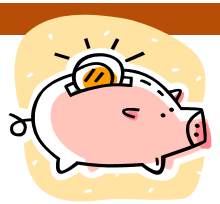
2. Ease of use

- DBs supporting OLTP systems are not meant for cross-organisational queries



A well architected Data Warehouse is much, much simpler to navigate

Benefits of Building a DW



3. Value Add

- A Data Warehouse can provide more data/information than is available from the OLTP system by incorporating additional business rules into the ETL processes or Semantic Layer.

- For example, arbitrary classification of students performed differently to how they are represented in the source.

A well designed DW can provide much more information to users than can be found in the source system

Benefits of Building a DW

- 4. Consolidation

- A Data Warehouse can provide data sourced from many disparate sources;

  Student management
  Human resources
  Finance
  Space management
  Marketing

- Example: Student to staff ratios

A DW can provide a unique perspective on many aspects of the organisation

Benefits of Building a DW

5. Consistency / single version of the truth

- A Data Warehouse can provide a commonly defined view of data and should become the **authoritative source**. This helps to minimise the distribution of conflicting data, and aids in common understanding.

A DW can deliver well understood and consistent data!

e.g. 'above track vs below track Corridor' in rail company

Benefits of Building a DW

6. Data quality

- Data Quality components can be built into a Data Warehouse to ensure **integrity is higher than the source**. For example, data validations.
- General principle should be to identify data issues via the DW, but to correct in the source.

A well designed DW can help ensure data quality meets the organisations needs

Failure factors

- Failure factors in data warehouse projects:

1. Cultural issues being ignored
2. **Inappropriate architecture**
3. Unclear business objectives
4. Missing information
5. Unrealistic expectations
6. Low levels of data summarisation
7. **Low data quality**
8. Missing ERP linkages

Practical assignment: Q & A

Gartner

Search

Advanced Search

My Tracks

My Library

2020 Magic Quadrant

Market Overview

From a financial perspective, the market for modern, self-service ABI platforms continues to grow at speed, but slower than before. According to Gartner's market share analysis, the market's revenue grew by 22.3% in 2018, compared with 35.0% in 2017. Pricing pressure and strong competition were broadly responsible for this deceleration. See "Market Share Analysis: Analytics and BI Software, Worldwide, 2018." But although spending is growing more slowly than before, the number of people using ABI platforms is accelerating massively. Microsoft alone now has millions of users around the world using its Power BI cloud service, which was launched just five years ago. The huge increase in user numbers is because the price per user is a fraction of what it was a decade ago.

2019 was a year of transition toward cloud ecosystem dominance. The rapid growth of the Microsoft Azure-based Power BI cloud service, along with Salesforce's acquisition of Tableau and Google's purchase of Looker, signaled a change whereby cloud stacks are now expected to come with a competitively priced ABI platform (see "Recent Acquisitions Signal Big Changes to the Analytics and Business Intelligence Platform Market"). Of course, along with this transition comes a natural concern about lock-in. The balancing factors here are vendors' attitudes toward, and implementation of, openness in their stacks and the growing importance of "multicloud" approaches, whereby customers can choose to run an application in, and spanning, multiple cloud IaaS offerings. The move to cloud platform as a service (PaaS)-aligned ABI as a norm is impacting how nonaligned vendors are positioning their offerings and competing. Two main strategies are emerging.


See you next
week!