Produced by:
Dr. Brenda Mullally bmullally@wit.ie
Ruth Barry rbarry@wit.ie

Department Computing Maths and Physics
Waterford Institute of Technology
www.wit.ie
moodle.wit.ie

MSc Enterprise Software Systems

Business Intelligence

Learning Outcomes

- Reasons to create an EDW
- Design
  - Entity relationship modelling
  - Dimensional modelling
- Data Understanding
- ETL
- Challenges
- Implementation
- Future of DW

Reasons for creating an EDW

- Operational queries
- Database structure
- Data Integration

Data Warehouse Design

Inmon:

EDW approach, top down, ER Modelling

Kimball:

Data Mart approach, bottom up, Dimensional Modelling

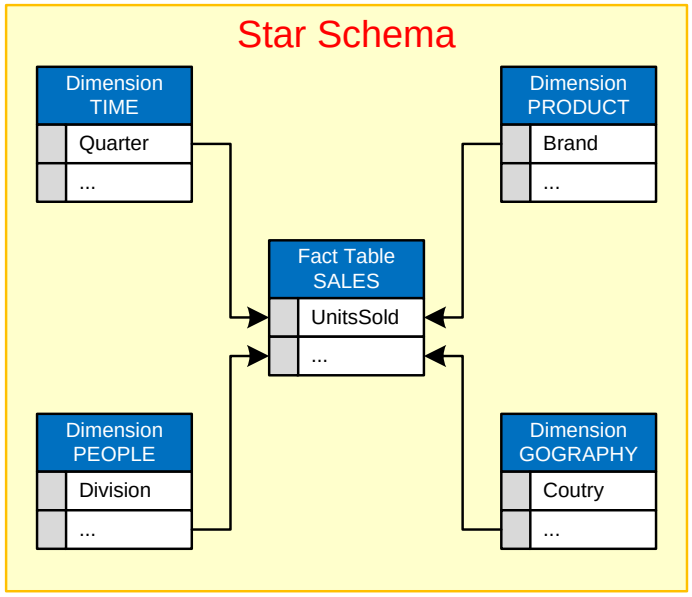
Packaged

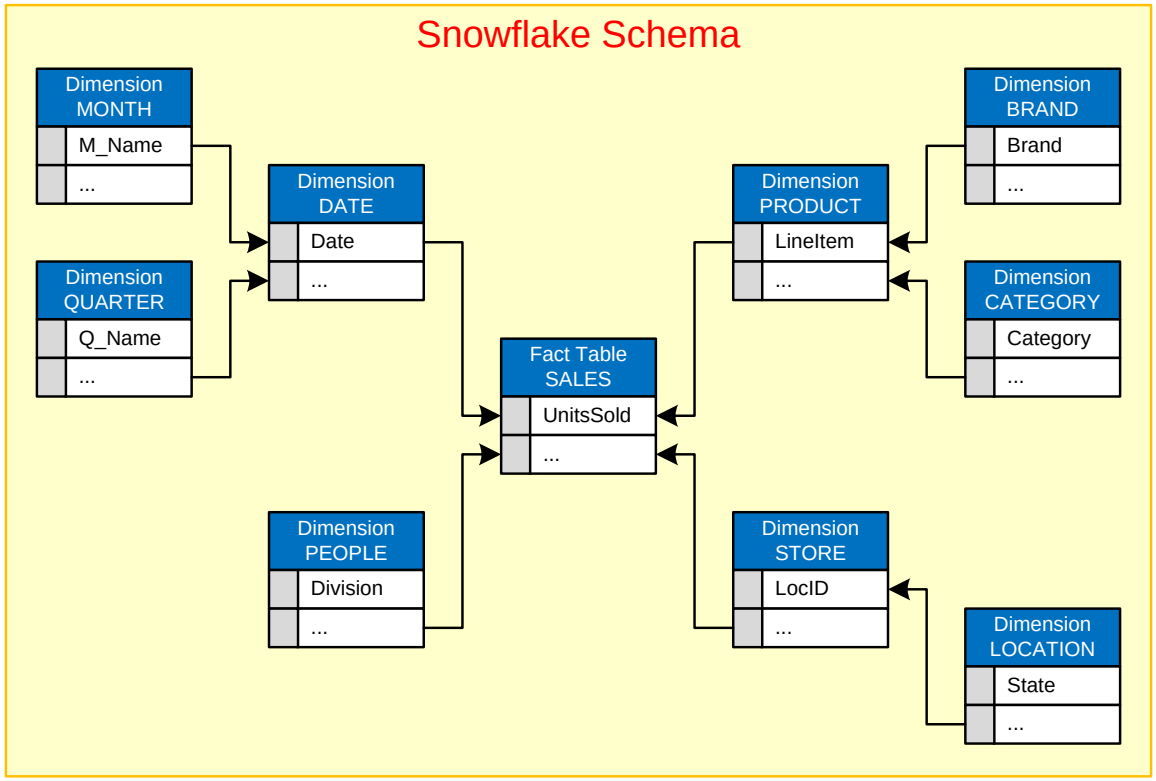
Which is best?

Representation of Data in DW

- Dimensional Modeling – a retrieval-based system that supports high-volume query access
- Star schema – the most commonly used and the simplest style of dimensional modeling
  - Contain a fact table surrounded by and connected to several dimension tables
  - Fact table contains the descriptive attributes (numerical values) needed to perform decision analysis and query reporting
  - Dimension tables contain classification and aggregation information about the values in the fact table
- Snowflakes schema – an extension of star schema where the diagram resembles a snowflake in shape

Star vs Snowflake Schema





Multidimensionality

- Multidimensionality

  The ability to organize, present, and analyze data by several dimensions, such as sales by region, by product, by salesperson, and by time (four dimensions)

- Multidimensional presentation
  - Dimensions: products, salespeople, market segments, business units, geographical locations, distribution channels, country, or industry, Time: daily, weekly, monthly, quarterly, or yearly
  - Facts: money, sales volume, head count, inventory profit, actual versus forecast

Data Understanding

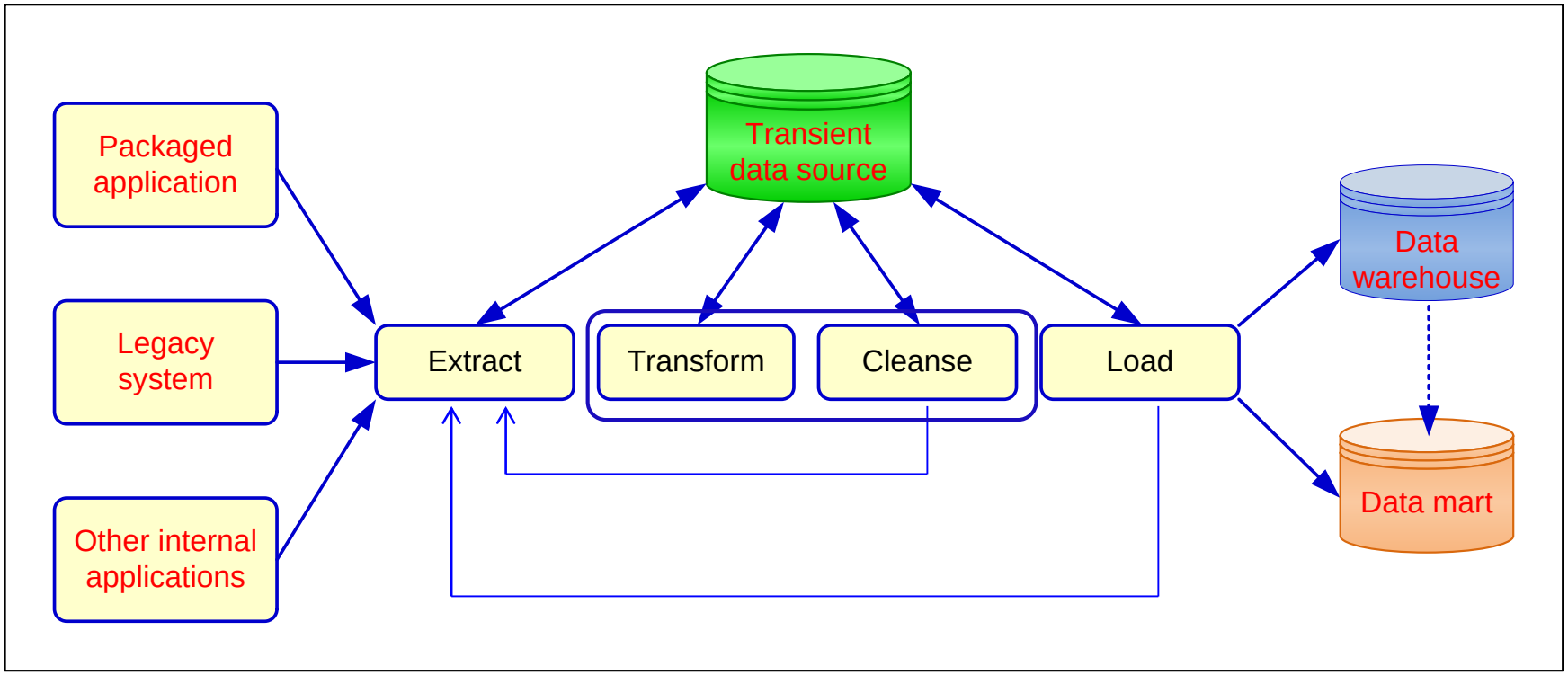
- Data understanding is crucial to the success of BI projects.

- Data collection
- Data description
- Data quality and verification

- Understanding and preparing the data consumes most of the time and resources in the implementation of a BI project. It may consume from 50-80% of the total resources.

Data Integration and the Extraction, Transformation, and Load (ETL) Process

Extraction, transformation, and load (ETL)



ETL

- Extraction
  - sources
- Transform & Clean
  - rules
- Load

- ETL tool – purchased or developed?

ETL

- Issues affecting the purchase of ETL tool
  - Data transformation tools are expensive
  - Data transformation tools may have a long learning curve
- Important criteria in selecting an ETL tool
  - Ability to read from and write to an unlimited number of data sources/architectures
  - Automatic capturing and delivery of metadata
  - A history of conforming to open standards
  - An easy-to-use interface for the developer and the functional user

ETL

- Careful of pitfalls:
  - Too many ETL processes
  - Redundant data
  - Poor design of ETL - costs

DW Challenges

- Technical Options
- Changes in technologies and vendors
- Integration requirements
- Knowledge transfer challenges
- Data Quality
- Time to market
- Buy versus build
- Aggregation

DW Implementation

- Establishment of service-level agreements and data-refresh requirements
- Identification of data sources and their governance policies
- Data quality planning
- Data model design
- ETL tool selection
- Relational database software and platform selection
- Data transport
- Data conversion
- Reconciliation process
- Purge and archive planning
- End-user support

DW Implementation

- Project must fit with corporate strategy & business objectives
- There must be complete buy-in to the project by executives, managers, and users
- It is important to manage user expectations about the completed project
- Build in adaptability, and scalability
- The project must be managed by both IT and business professionals
- Only load data that have been cleansed and are of a quality understood by the organization.
- Do not overlook training requirements
- Be politically aware

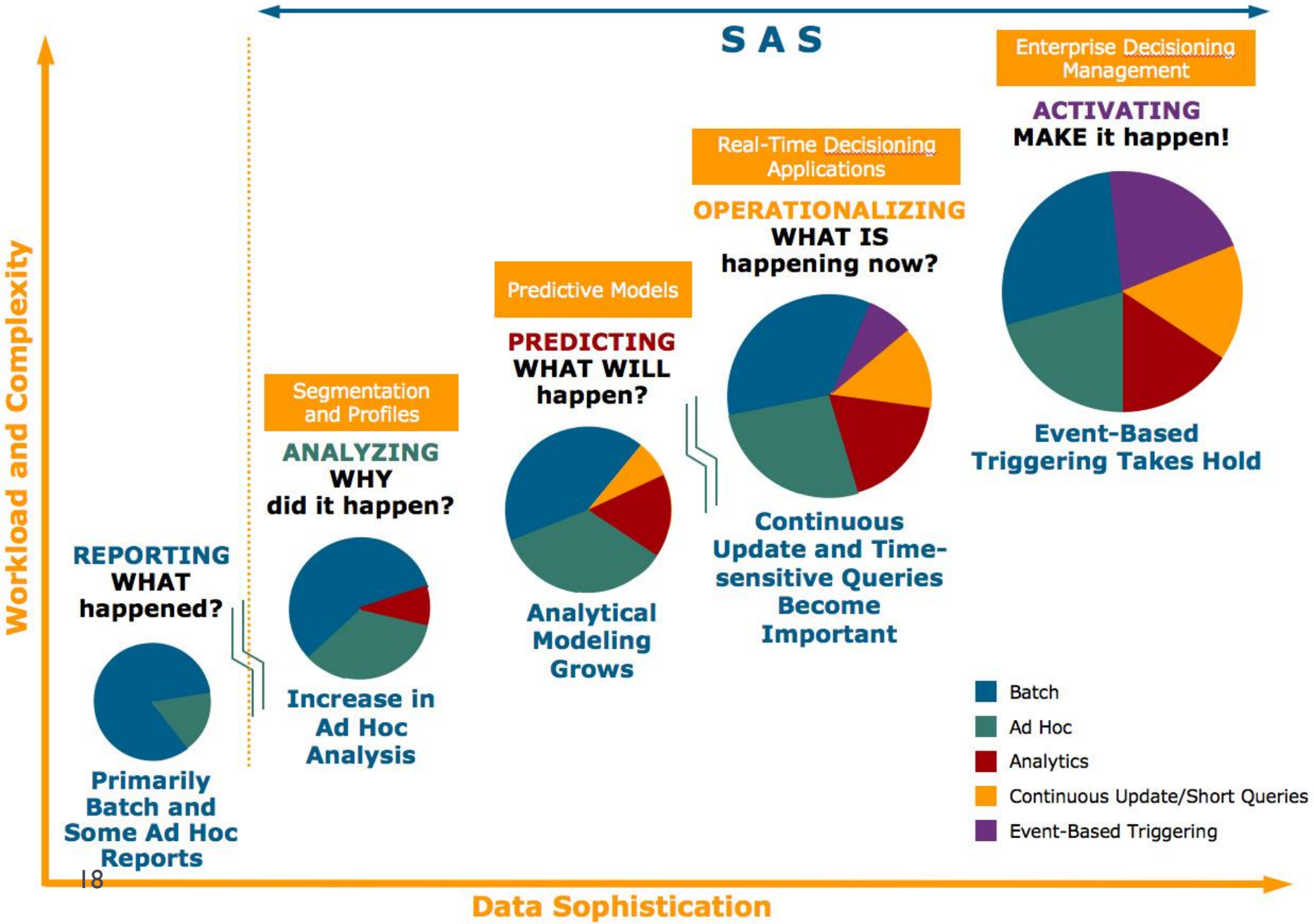
DW Implementation: Things to Avoid

- Starting with the wrong sponsorship chain
- Setting expectations that you cannot meet
- Engaging in politically naive behavior
- Loading the data 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

DW Implementation: Things to Avoid

- Focusing on traditional internal record-oriented data and ignoring the value of external data and of text, images, etc.
- Delivering data with confusing definitions
- Believing promises of performance, capacity, and scalability
- Believing that your problems are over when the data warehouse is up and running
- Focusing on ad hoc and periodic reporting instead of alerts

Enterprise Decision Evolution and DW



Massive DW and Scalability

- Scalability
  - The main issues pertaining to scalability:
    - The amount of data in the warehouse
    - How quickly the warehouse is expected to grow
    - The number of concurrent users
    - The complexity of user queries
  - Good scalability means that queries and other data-access functions will grow linearly with the size of the warehouse

Real-time/active/right-time DW/BI

- Enabling real-time data updates for real-time analysis and real-time decision making is growing rapidly
  - Push vs. Pull (of data)
- Concerns about real-time BI
  - Not all data should be updated continuously
  - Mismatch of reports generated minutes apart
  - May be cost prohibitive
  - May also be unachievable

Teradata active DW

Real time/active/right time

For many people, the "real-time" term is synonymous with "instantaneous."

This interpretation, however, is incorrect when applied to data warehousing. While some warehouse data may be captured and entered into the warehouse in seconds or minutes, much of it is not.

For example, some source systems, such as a legacy COBOL program that is updated once a month, can never be more real-time than when last updated. Some data may be prohibitively expensive or difficult to make real-time. Most importantly, there may not be a business need for real-time data. Data only needs to be as fresh as the business requirements. For these reasons, some people prefer the "right time" term.

DW Administration and Security

- Data warehouse administrator (DWA)
  - DWA should…
    - have the knowledge of high-performance software, hardware and networking technologies.
    - possess solid business knowledge and insight.
    - be familiar with the decision-making processes so as to suitably design/maintain the data warehouse structure.
    - possess excellent communications skills.
- Security and privacy is a pressing issue in DW & BI
  - Safeguarding the most valuable assets
  - Government regulations (Data protection, HIPPA, Sarbanes Oxley etc.)
  - Must be explicitly planned and executed

https://www.dataprotection.ie/documents/facebook%20report/final%20report/report.pdf

The Future of DW

- Sourcing…
  - Web, social media, big data
  - Open source software
  - SaaS (software as a service)
  - Cloud computing
  - DW appliances
- Infrastructure…
  - Real-time DW
  - Data management practices/technologies
  - In-memory storage
  - In-memory processing
  - Sandboxes
  - Advanced analytics