Introduction

School of Information Studies
Syracuse University

Agenda:

1. Define data warehousing.
2. Explain four characteristics of a data warehouse.
3. Discuss the relationship between data warehouse, business intelligence, and analytics.
4. Explain the five types of analytics.
5. Demonstrate how the process of data warehousing works.
6. Learn about the fathers of data warehousing.
7. Cover our case studies.

Connect Activity: Introduction to Data Warehouses

When you hear the word "data warehouse," describe in a few words what comes to mind.

The Data Warehouse Defined

School of Information Studies
Syracuse University

What is the most important asset of any organization?

Answer

DATA

Why?

Without Data:

- Do you know your customers?
- Understand their needs?
- Can you figure out what products to put on sale?
- Which ones to discontinue?
- Do you know your expenses?
- Your profitability?

NOPE

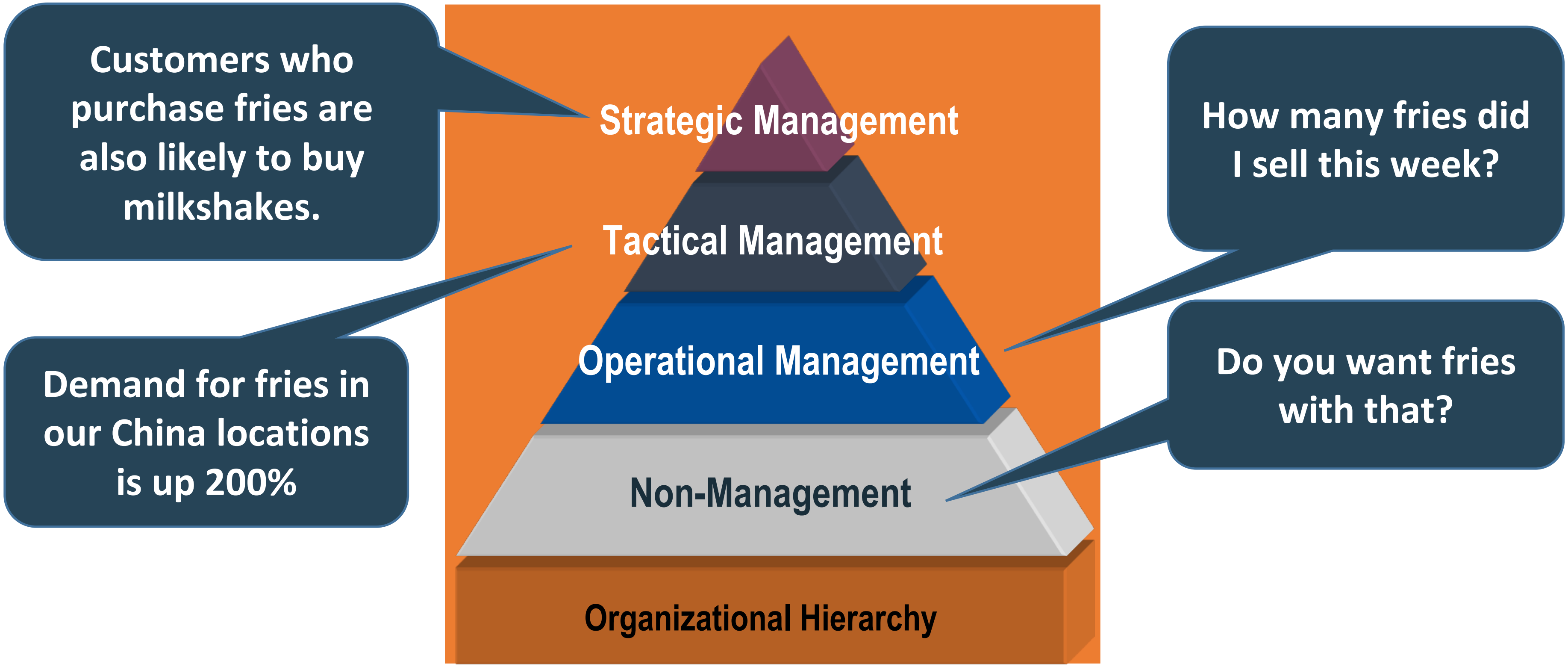
This reminds me of a story…

The information needs of an organization…

```
        MCDONALD'S RESTAURANT
   21 UPPER CENTENNIAL PARKWAY
          STONEY CREEK
          905-664-6826
       GST No. 112968409
  YOUR  ORDER  #20720
2008-08-02                1:33 PM
        29117 2 16 20720
1 ANGUS BURGER V             3.11
   No MAYO
1 4-NUGGET HM                2.54
1 MEDIUM FRY                 2.09
1 MED FRUITOPIA              1.79
   No Ice
1 CHILD ICE TEA              1.29
   No Ice
1 BOY TOY
1 HAPPY TOY                  0.96

GST                          0.59
PST                          0.94
```

The information needs of an organization…

Each level of an organization has different informational needs and requirements



Strategic Management

Tactical Management

Operational Management

Non-Management

Organizational Hierarchy

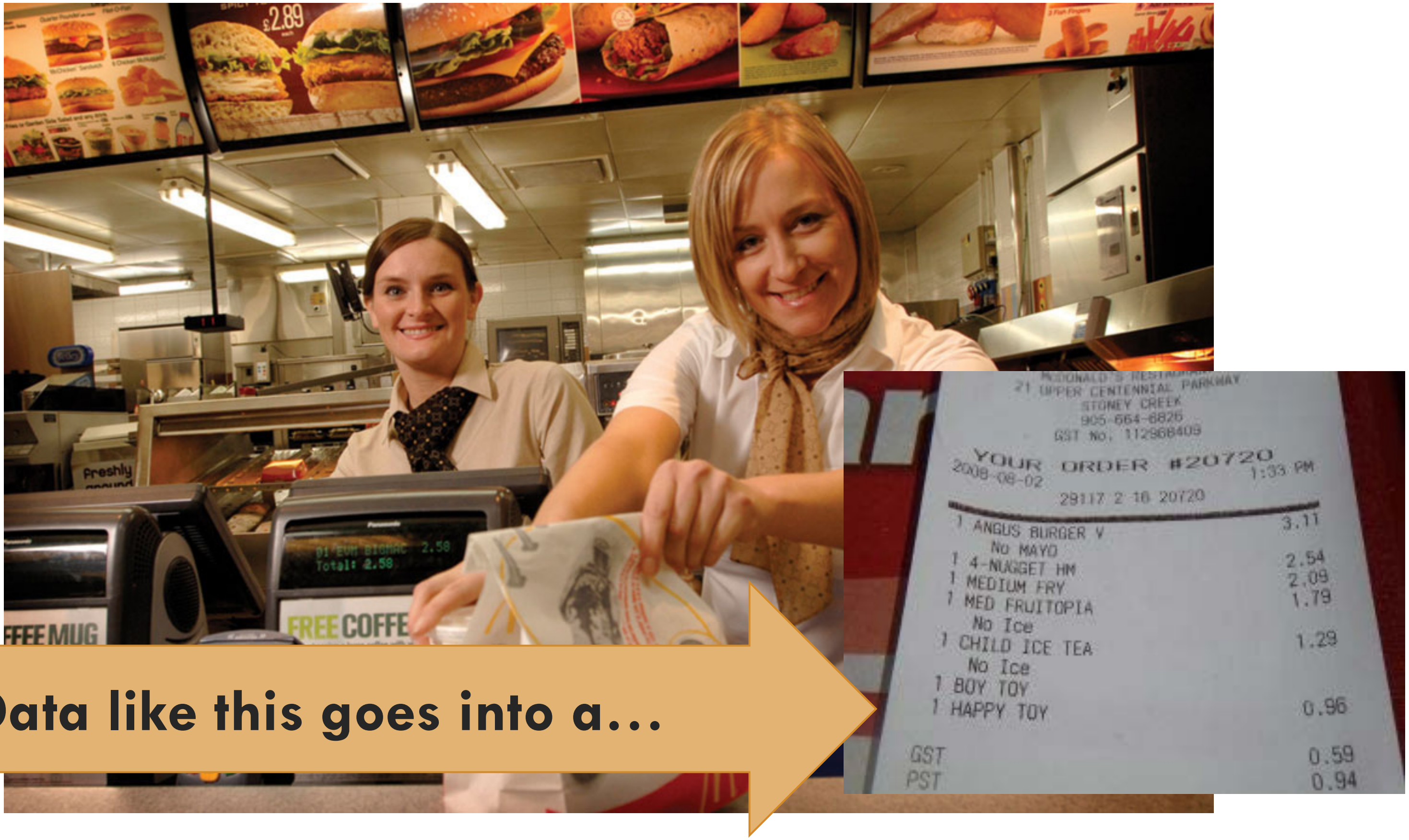
Customers who purchase fries are also likely to buy milkshakes.

Demand for fries in our China locations is up 200%

How many fries did I sell this week?

Do you want fries with that?

The information needs of an organization…



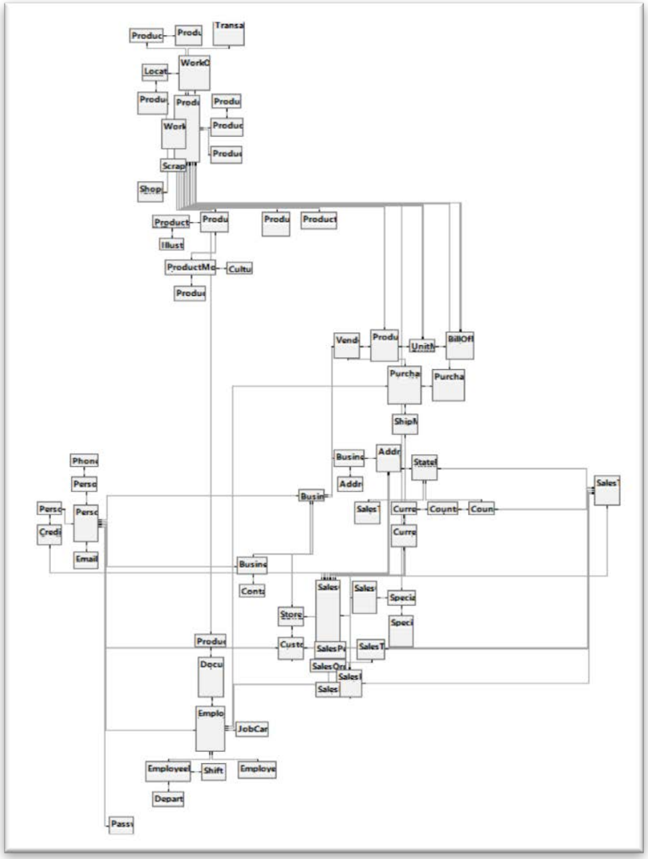
```
          MCDONALD'S RESTAURANT
      21 UPPER CENTENNIAL PARKWAY
             STONEY CREEK
             905-664-6826
          GST No. 112968409
   YOUR    ORDER    #20720
2008-08-02                   1:33 PM
          29117 2 16 20720
 1 ANGUS BURGER V                3.11
     No MAYO
 1 4-NUGGET HM                   2.54
 1 MEDIUM FRY                    2.09
 1 MED FRUITOPIA                 1.79
     No Ice
 1 CHILD ICE TEA                 1.29
     No Ice
 1 BOY TOY
 1 HAPPY TOY                     0.96

GST                              0.59
PST                              0.94
```

Data like this goes into a…

Starts With the OPERATIONAL Database (OLTP)

- Online Transaction Processing System
- Typically stored in a relational database or files
- Highly normalized (data stored as efficiently as possible, lots of tables)
- Optimized for processing speed and handling the “now”
- Designed for capturing data, not for reporting on it
- Designed to support the operational needs of the organization

Transactional Databases Are *Complex*

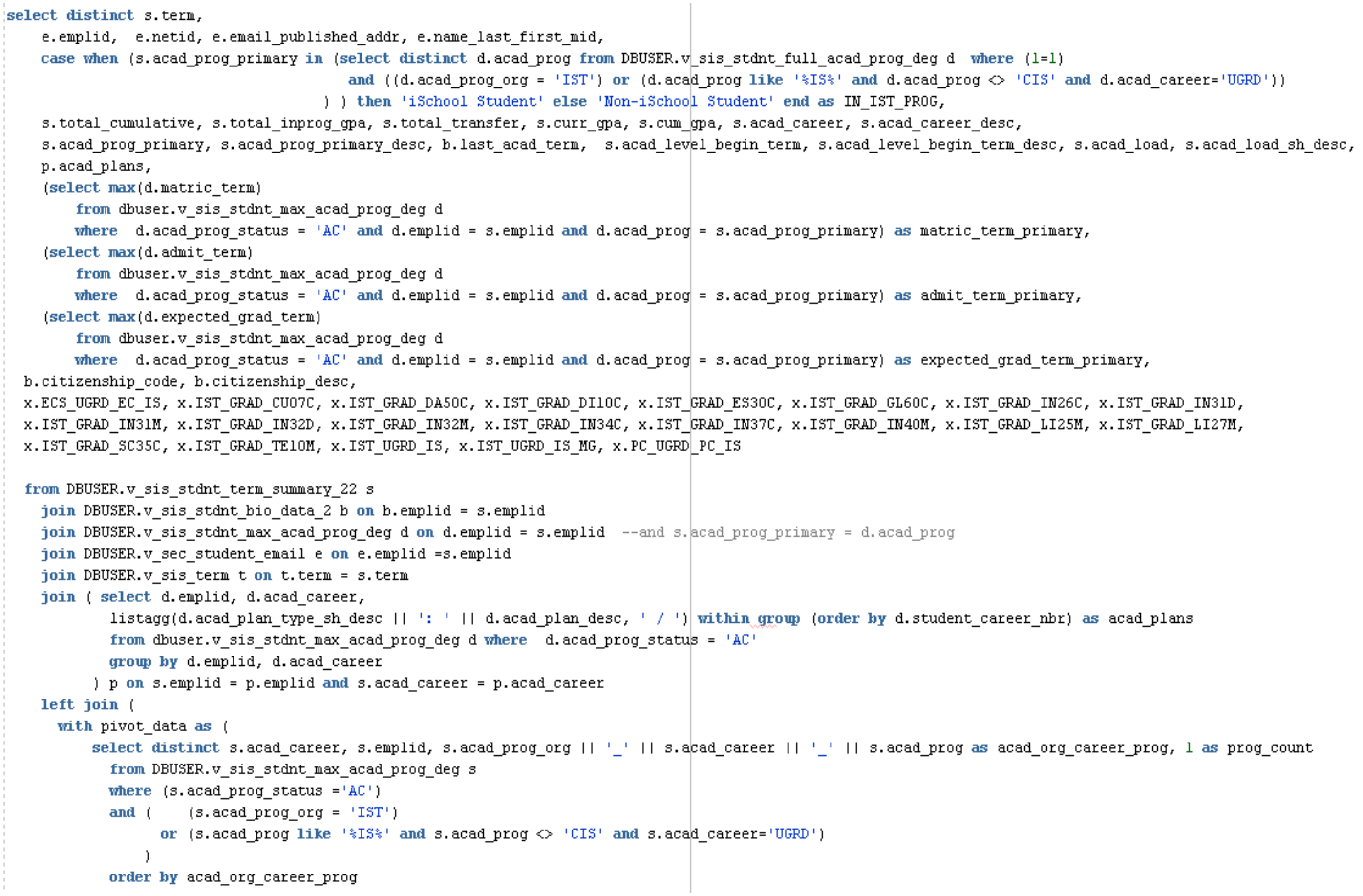


← Adventure Works **fictitious** bicycle manufacturer: **72 tables.**

Blackboard learning management system: **592 tables.**

SU’s Oracle PeopleSoft ERP implementation: **40,000+ tables.**

Example: a Query of “iSchool Students”



```sql
select distinct s.term,
    e.emplid,  e.netid, e.email_published_addr, e.name_last_first_mid,
    case when (s.acad_prog_primary in (select distinct d.acad_prog from DBUSER.v_sis_stdnt_full_acad_prog_deg d  where (1=1)
                                        and ((d.acad_prog_org = 'IST') or (d.acad_prog like '%IS%' and d.acad_prog <> 'CIS' and d.acad_career='UGRD'))
                                     ) ) then 'iSchool Student' else 'Non-iSchool Student' end as IN_IST_PROG,
    s.total_cumulative, s.total_inprog_gpa, s.total_transfer, s.curr_gpa, s.cum_gpa, s.acad_career, s.acad_career_desc,
    s.acad_prog_primary, s.acad_prog_primary_desc, b.last_acad_term,  s.acad_level_begin_term, s.acad_level_begin_term_desc, s.acad_load, s.acad_load_sh_desc,
    p.acad_plans,
    (select max(d.matric_term)
        from dbuser.v_sis_stdnt_max_acad_prog_deg d
        where  d.acad_prog_status = 'AC' and d.emplid = s.emplid and d.acad_prog = s.acad_prog_primary) as matric_term_primary,
    (select max(d.admit_term)
        from dbuser.v_sis_stdnt_max_acad_prog_deg d
        where  d.acad_prog_status = 'AC' and d.emplid = s.emplid and d.acad_prog = s.acad_prog_primary) as admit_term_primary,
    (select max(d.expected_grad_term)
        from dbuser.v_sis_stdnt_max_acad_prog_deg d
        where  d.acad_prog_status = 'AC' and d.emplid = s.emplid and d.acad_prog = s.acad_prog_primary) as expected_grad_term_primary,
  b.citizenship_code, b.citizenship_desc,
  x.ECS_UGRD_EC_IS, x.IST_GRAD_CU07C, x.IST_GRAD_DA50C, x.IST_GRAD_DI10C, x.IST_GRAD_ES30C, x.IST_GRAD_GL60C, x.IST_GRAD_IN26C, x.IST_GRAD_IN31D,
  x.IST_GRAD_IN31M, x.IST_GRAD_IN32D, x.IST_GRAD_IN32M, x.IST_GRAD_IN34C, x.IST_GRAD_IN37C, x.IST_GRAD_IN40M, x.IST_GRAD_LI25M, x.IST_GRAD_LI27M,
  x.IST_GRAD_SC35C, x.IST_GRAD_TE10M, x.IST_UGRD_IS, x.IST_UGRD_IS_MG, x.PC_UGRD_PC_IS

  from DBUSER.v_sis_stdnt_term_summary_22 s
    join DBUSER.v_sis_stdnt_bio_data_2 b on b.emplid = s.emplid
    join DBUSER.v_sis_stdnt_max_acad_prog_deg d on d.emplid = s.emplid  --and s.acad_prog_primary = d.acad_prog
    join DBUSER.v_sec_student_email e on e.emplid =s.emplid
    join DBUSER.v_sis_term t on t.term = s.term
    join ( select d.emplid, d.acad_career,
            listagg(d.acad_plan_type_sh_desc || ': ' || d.acad_plan_desc, ' / ') within group (order by d.student_career_nbr) as acad_plans
            from dbuser.v_sis_stdnt_max_acad_prog_deg d where  d.acad_prog_status = 'AC'
            group by d.emplid, d.acad_career
          ) p on s.emplid = p.emplid and s.acad_career = p.acad_career
    left join (
      with pivot_data as (
          select distinct s.acad_career, s.emplid, s.acad_prog_org || '_' || s.acad_career || '_' || s.acad_prog as acad_org_career_prog, 1 as prog_count
            from DBUSER.v_sis_stdnt_max_acad_prog_deg s
            where (s.acad_prog_status ='AC')
            and (    (s.acad_prog_org = 'IST')
                  or (s.acad_prog like '%IS%' and s.acad_prog <> 'CIS' and s.acad_career='UGRD')
                )
            order by acad_org_career_prog
```

Issues Reporting With Transactional Databases

Difficult, time-consuming, and error prone

- Many joins, subselects, due to vast number of tables.
- *How do you know your query is correct?*

Resource-intensive

- The database is not optimized for this purpose.
- *Multi-table joins are RAM and CPU hogs.*

Impossible

- Transactional systems are flushed or archived frequently to maintain performance.
- *You can't query data you no longer have.*

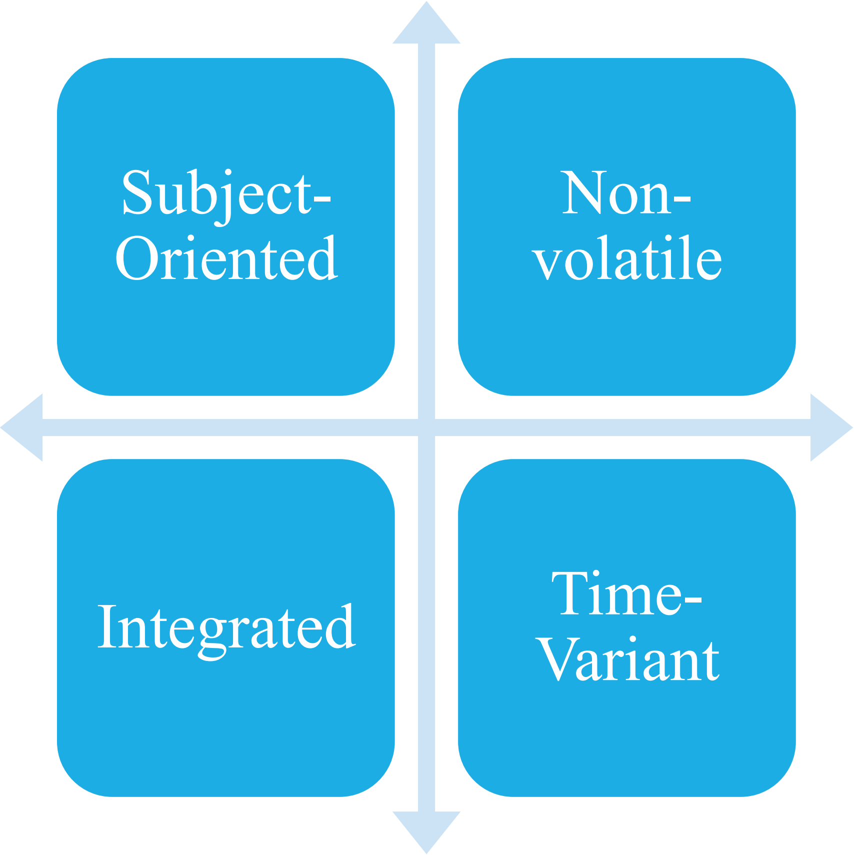
Solution? The Data Warehouse

- Designed to support an organization's informational needs.
- Data is restructured and conducive to reporting and analytic applications.
- OLTP databases are data sources for the data warehouse.
- Data grow over time; existing data in the warehouse never changes.

Characteristics of the Data Warehouse

School of Information Studies
Syracuse University

Four Characteristics of the Data Warehouse



Subject-Oriented

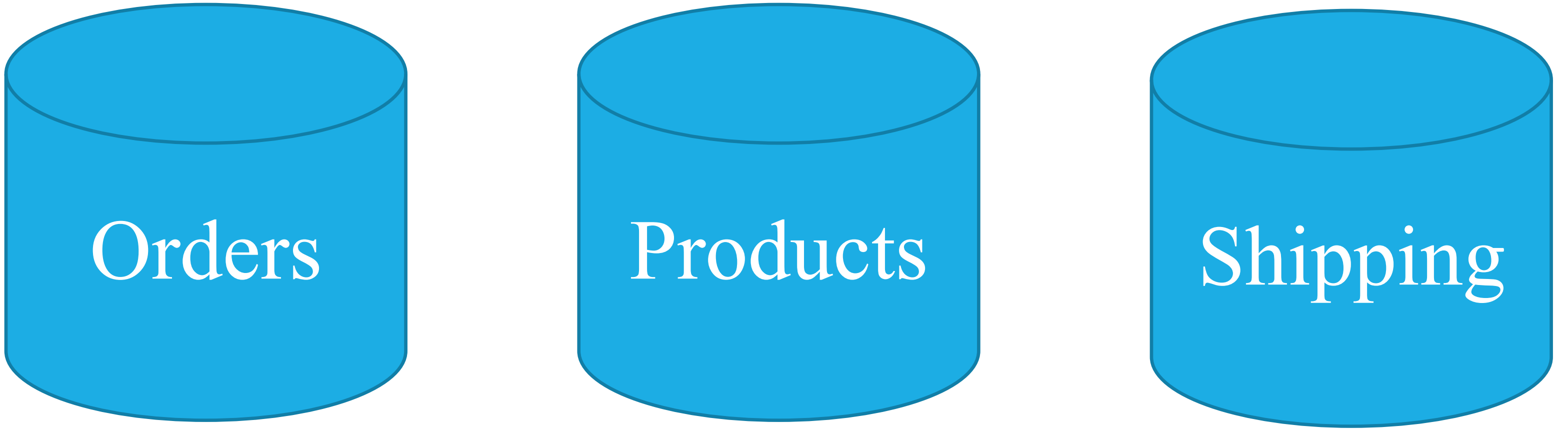
Non-volatile

Integrated

Time-Variant

Subject-Oriented

- Optimized to give answers to diverse questions
- Used by all functional areas
- Built around business entities and processes



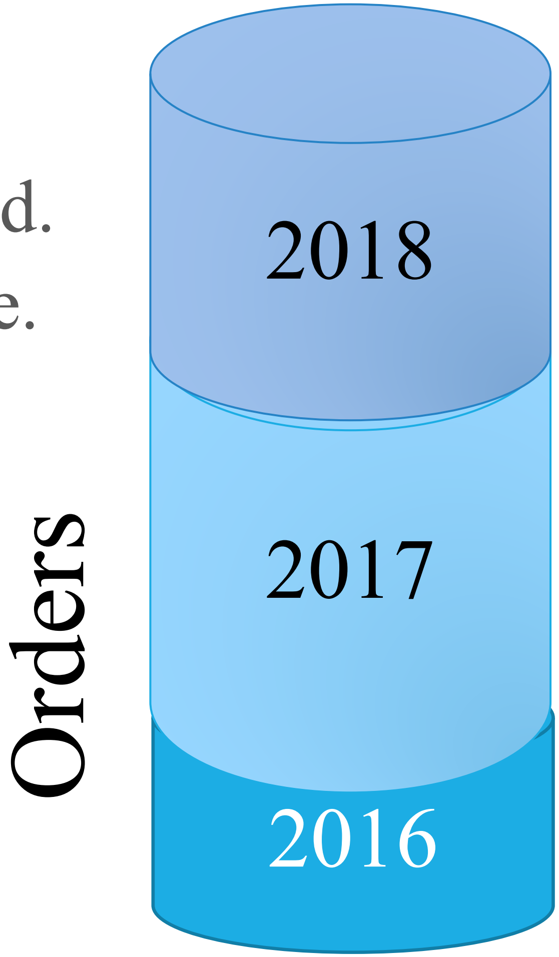
Orders

Products

Shipping

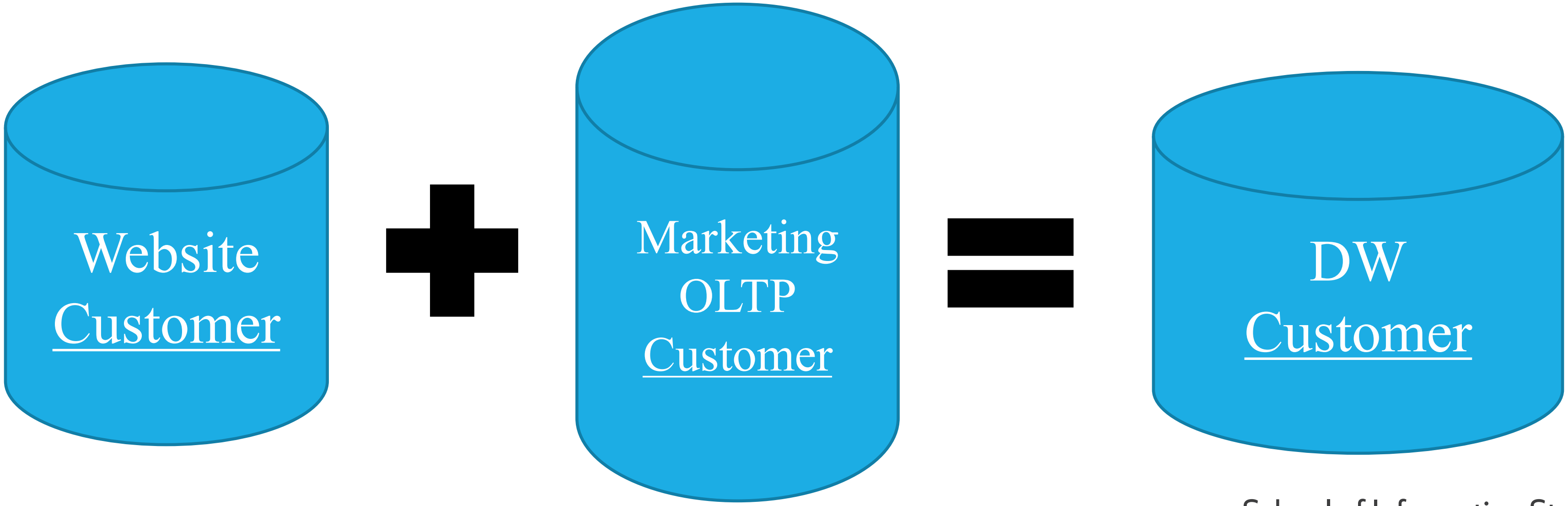
Nonvolatile

- Data are never removed or changed.
- Data are always growing over time.
- Historical data. It happened!
  We do not rewrite history!



Integrated

- Centralized in one place
- Holds data retrieved from entire organization
- "Single version of the truth"



Time-Variant

- Flow of data through time
- Data reflect as it was at that point in time

Invoice #: 12345
Amount: $55.90
Date: 4/1/2016
Customer: Michael Fudge
Address: 1313 Mockingbird

Invoice #: 52949
Amount: $95.50
Date: 11/4/2017
Customer: Michael Fudge
Address: 1600 Pennsylvania

What Is Business Intelligence?

School of Information Studies
Syracuse University

Business Intelligence

- Analytical and decision-support capabilities of the data warehouse
- Informed decision-making
- The presentation of actionable information
- The "glitz and glam" of data warehousing

Data Warehouse or Business Intelligence?

Is the **data warehouse** a component of **business intelligence**?

or

Is **business intelligence** a component of the **data warehouse**?

"No, *you* back off! I was here before you!"

DW Is the Foundation for BI

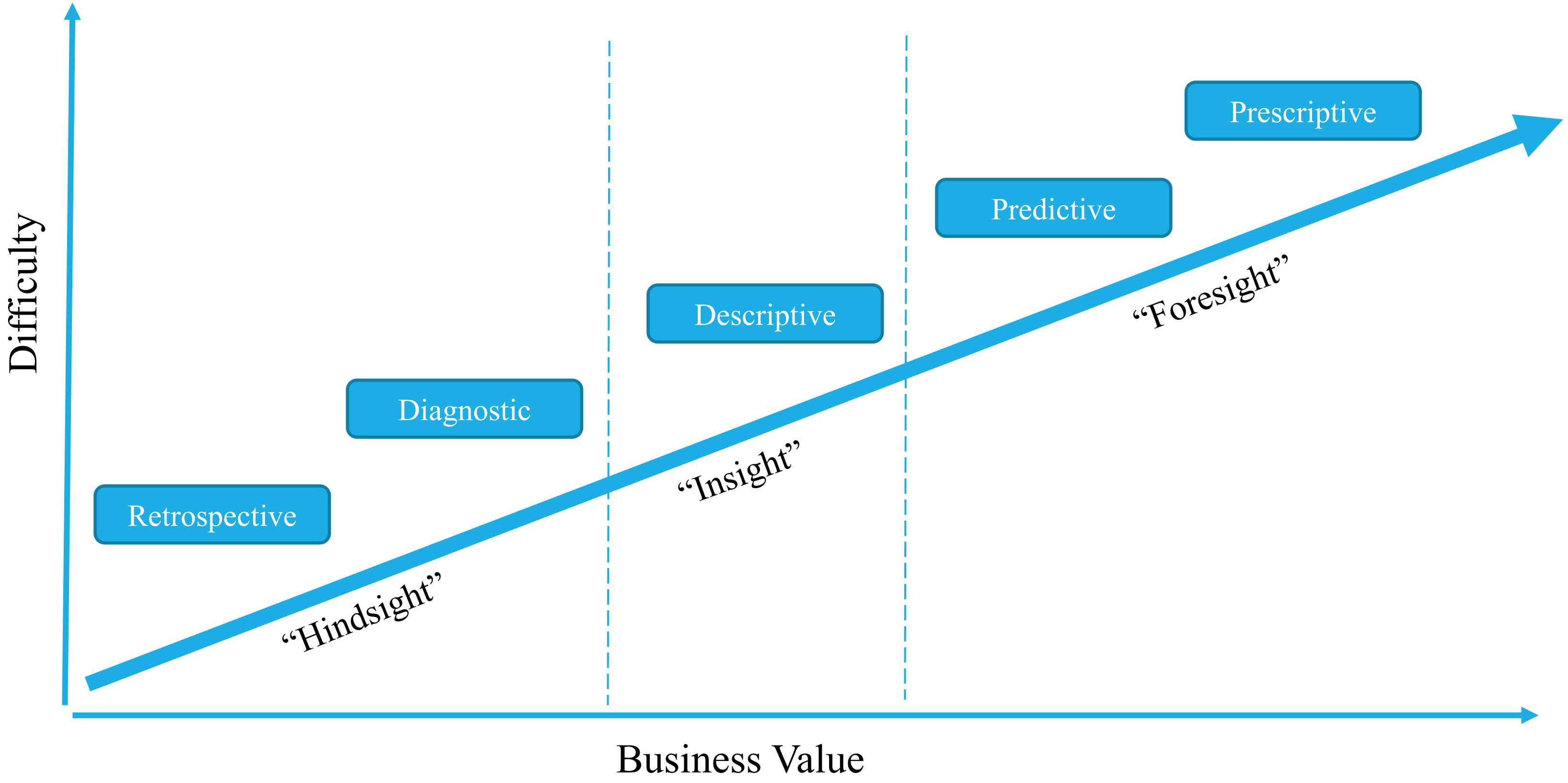
Five Types of Analytics

School of Information Studies
Syracuse University

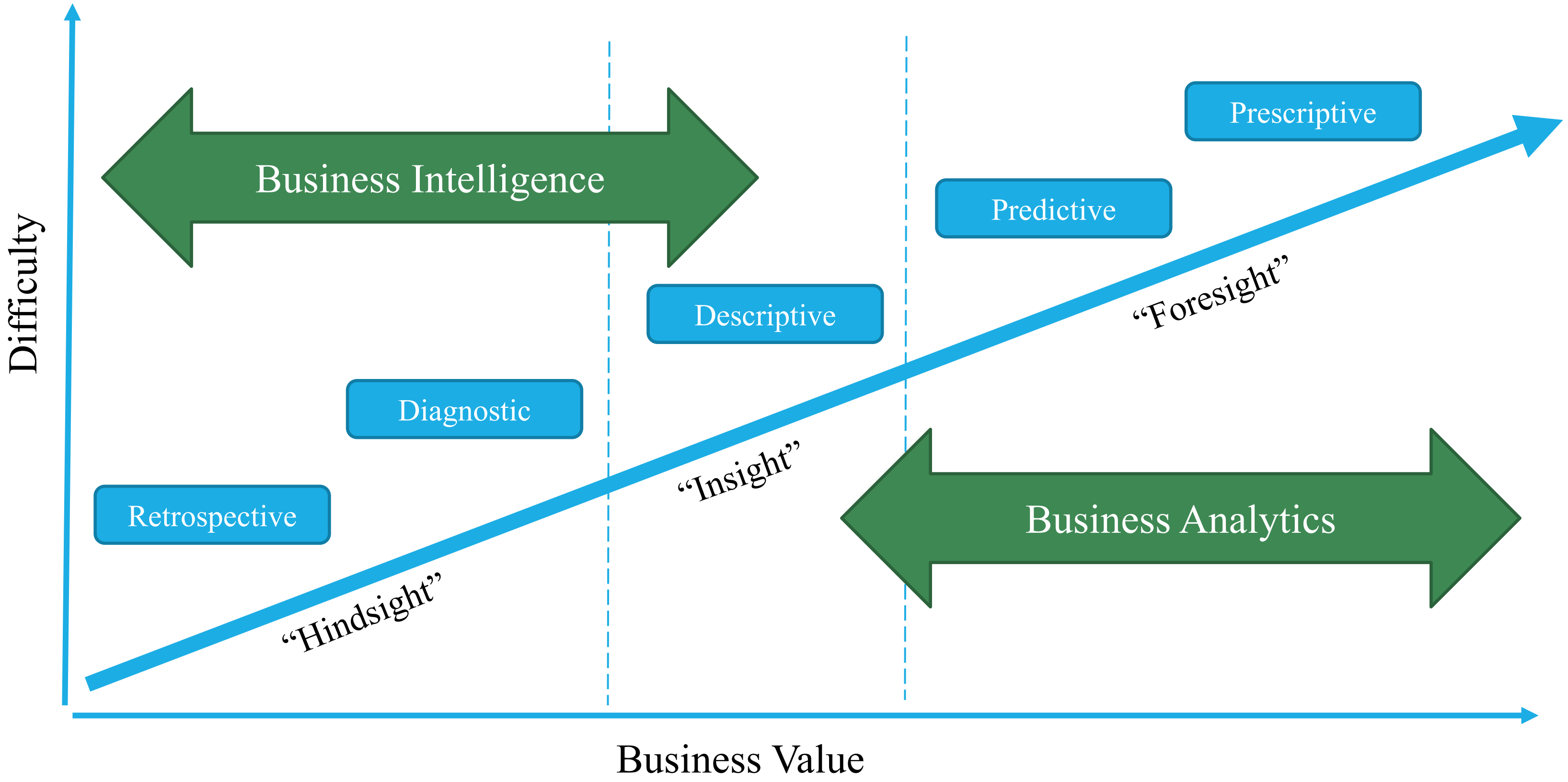
Analytics Is the Technology-Driven Analysis of Data

1. **Retrospective**: traditional business intelligence/reporting
   “What happened?”
2. **Diagnostic**: analytic dashboard/drill-down
   “Why did it happen?”
3. **Descriptive**: Real-time dashboard
   “What is happening now?”
4. **Predictive**: machine learning/forecasting
   “What is likely to happen?”
5. **Prescriptive analytics**: make a decision or take action
   “What should I do about it?”

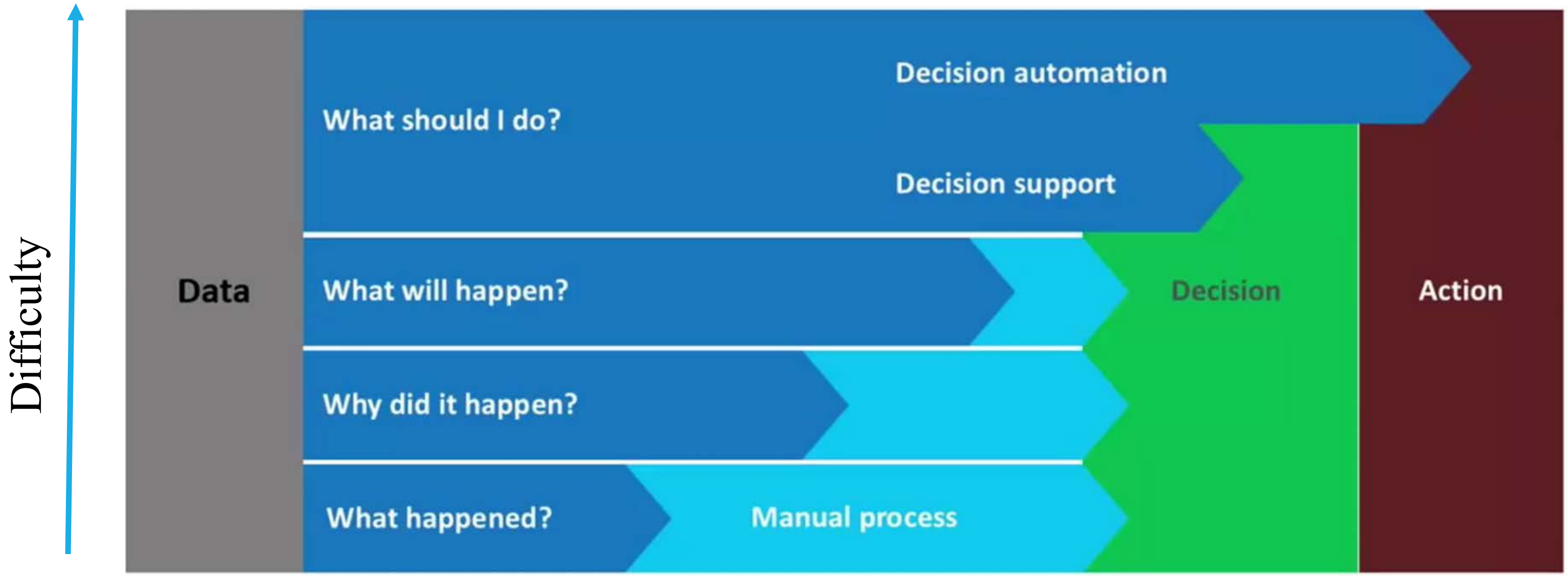
Comparison of Analytics



Comparison of Analytics



The Evolution of the Analytics Process



School of Information Studies
Syracuse University

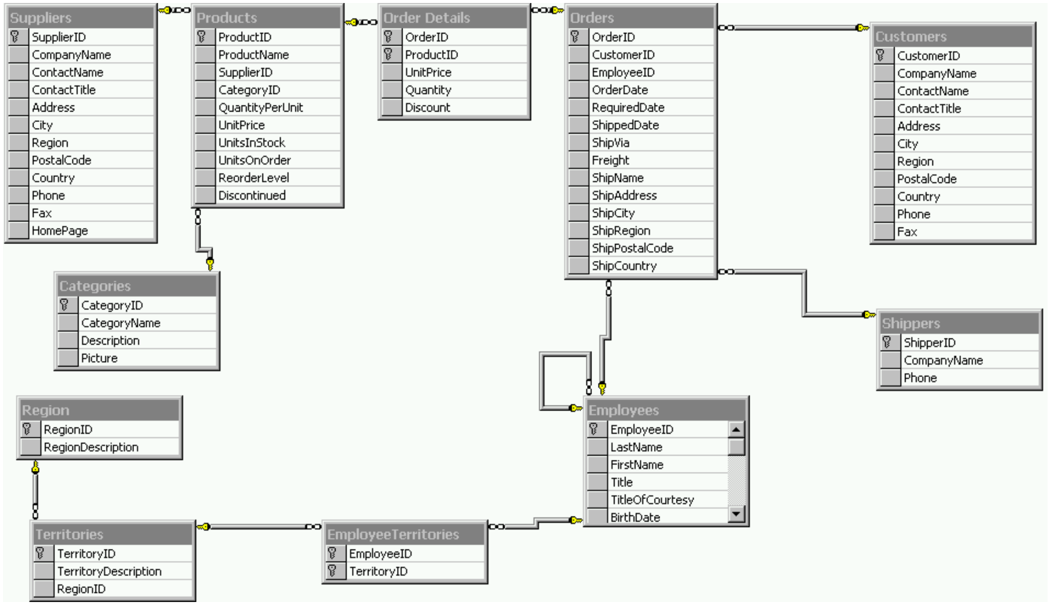
10,000-Foot View of the Process

School of Information Studies
Syracuse University

But How Does This Work?

Here’s a hyper abridged example…

1: We Have an OLTP Database



- Insufficient reporting capabilities.
- Can report only “in the now.”
- It takes complex queries to get questions answered.
- Database optimized for CRUD, not analytics.

2: Identify Business Process to Model

Business Process and Gain

- Orders: products sold to customers over time by sale
- One row per product order (product on the order)

Dimensions

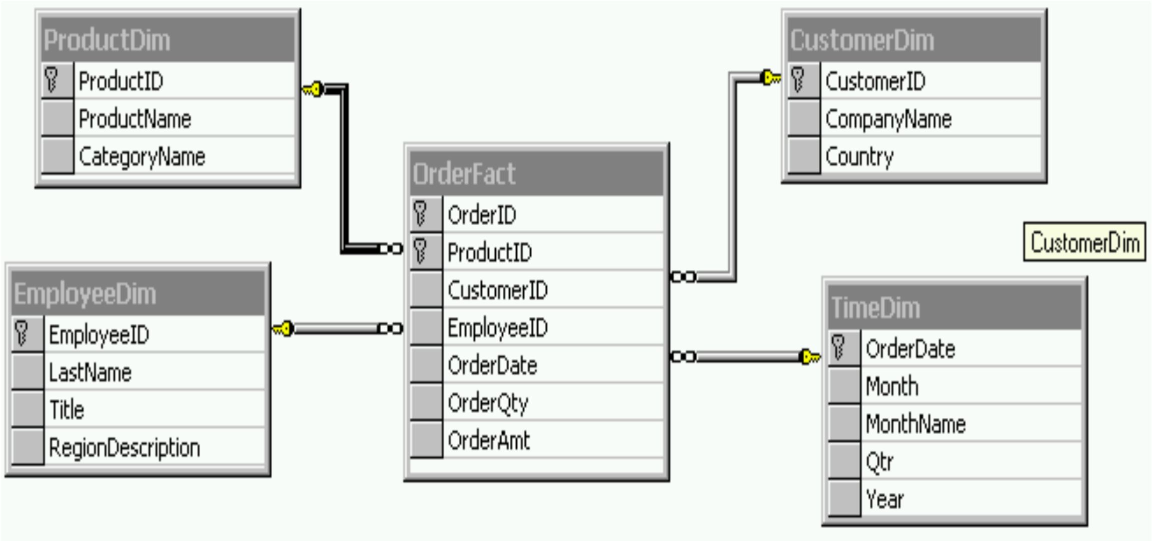
- Products, employees (sales), time (order date), customer
- Denormalized so they are easier for business users to understand

Facts

- Order quantity, order amount
- Things we can measure or quantify across dimensions

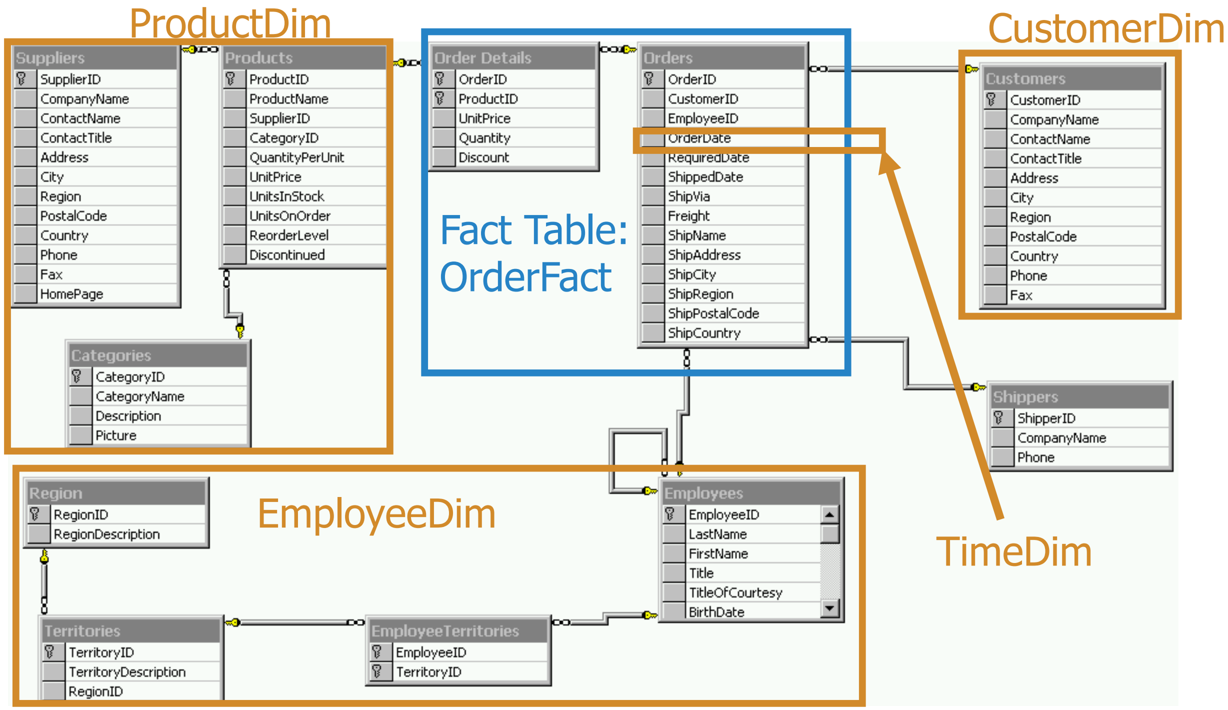
This represents our **data mart** in the data warehouse

3: Create Northwind Orders Star Schema



- **Data mart** is implemented as a **star schema** in a RDBMS.
- This is called ROLAP.
- Fact table + outer dimensions.
- No data (yet).
- Fields are based on what's available in the source data.

4: Create Northwind Source to Target Map

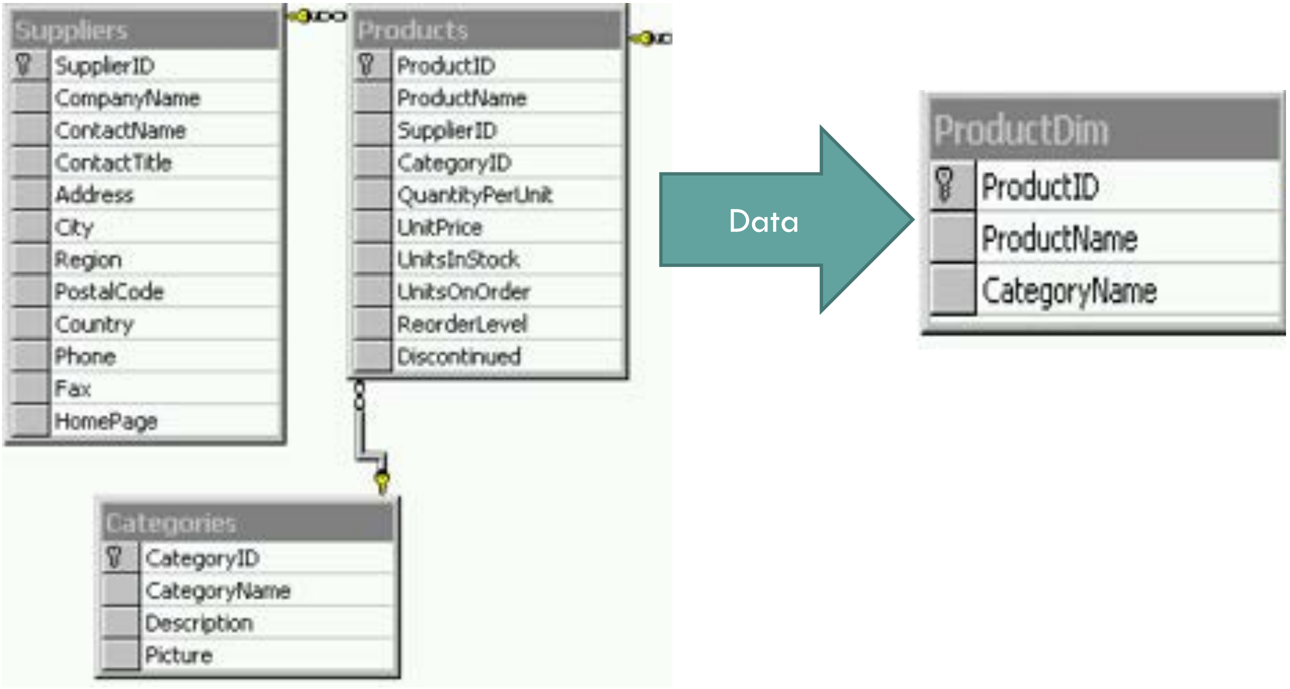


- How does the OLTP align with OLAP?
- Helps us define the ETL process

5: Populate Targets With ETL

Products Source

ProductsDim



Suppliers
- SupplierID
- CompanyName
- ContactName
- ContactTitle
- Address
- City
- Region
- PostalCode
- Country
- Phone
- Fax
- HomePage

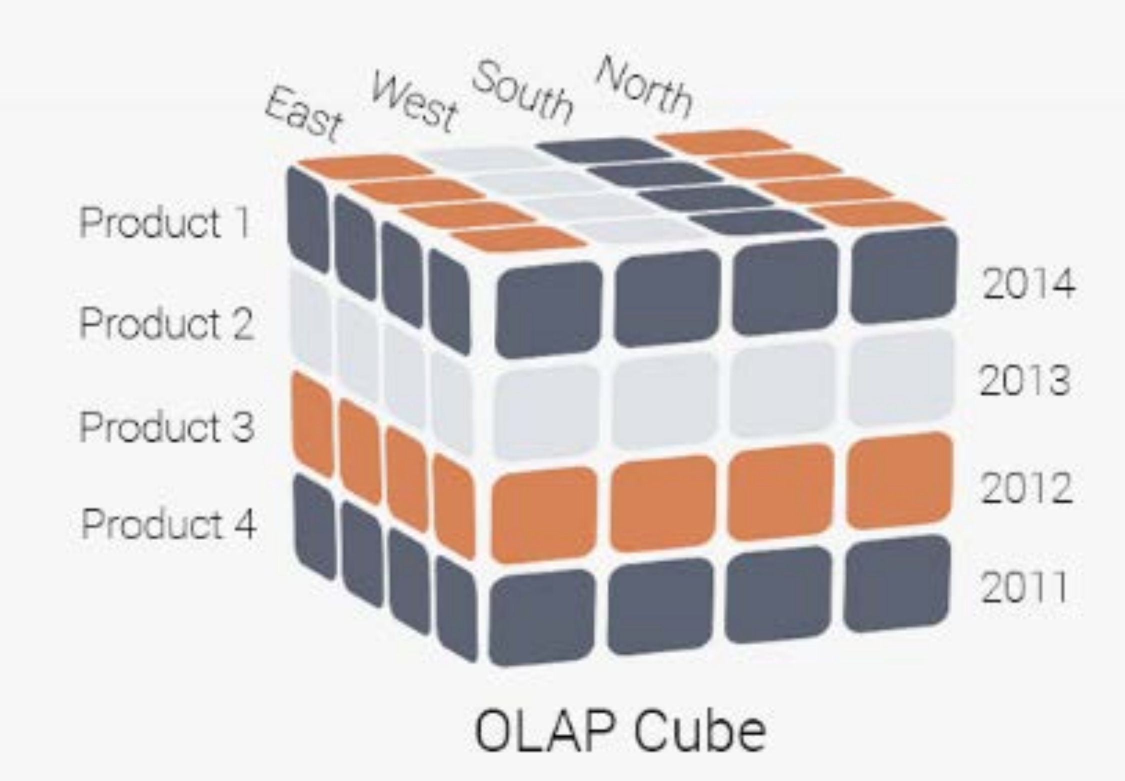
Products
- ProductID
- ProductName
- SupplierID
- CategoryID
- QuantityPerUnit
- UnitPrice
- UnitsInStock
- UnitsOnOrder
- ReorderLevel
- Discontinued

Categories
- CategoryID
- CategoryName
- Description
- Picture

ProductDim
- ProductID
- ProductName
- CategoryName

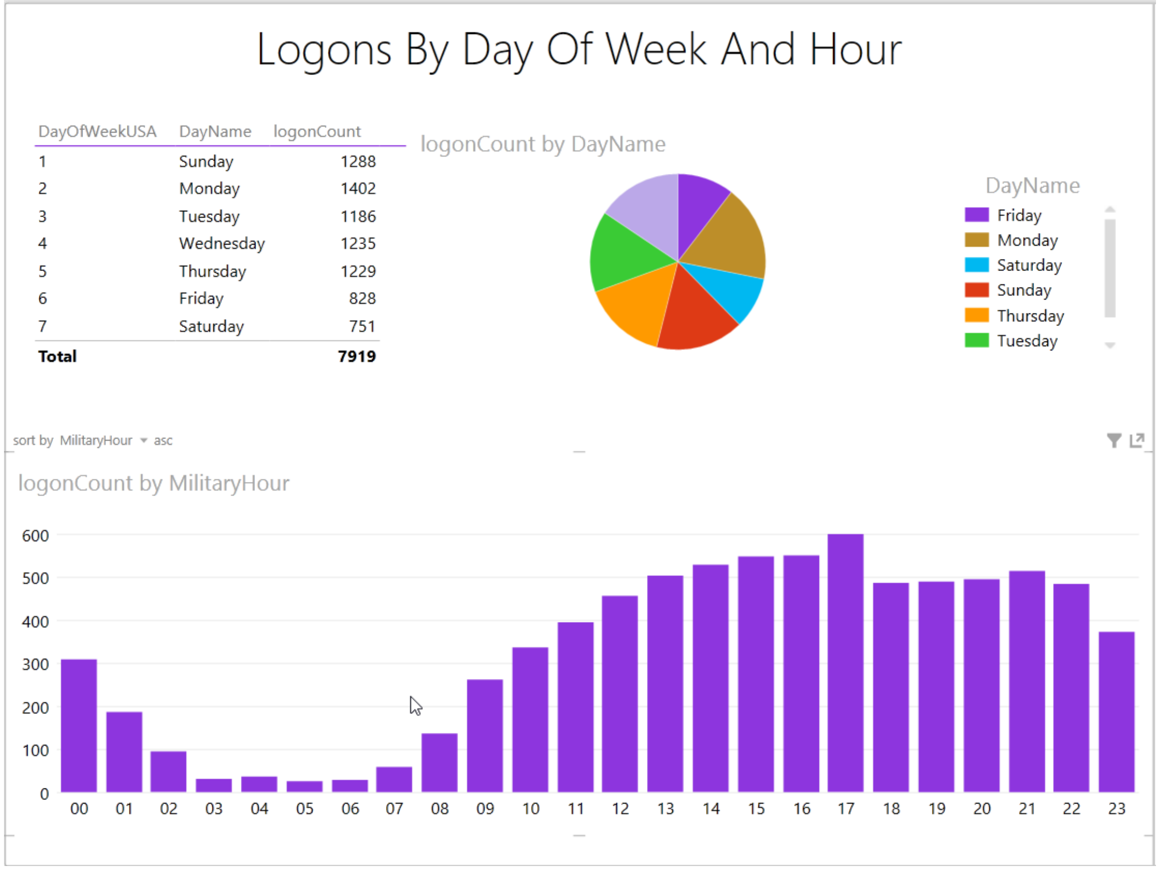
- ETL stands for extract-transform-load.
- Dimensions before facts.
- Need a strategy to handle changes to data.
- Tooling exists to assist with the process.

6: Build a Cube (MOLAP)



- Build aggregations of facts across dimensions.
- Static ad hoc reporting structure.
- Add a semantic model to address hierarchies and formatting.
- Uses a special database: MOLAP.

7: Visualize With a BI Tool



Logons By Day Of Week And Hour

DayOfWeekUSA	DayName	logonCount
1	Sunday	1288
2	Monday	1402
3	Tuesday	1186
4	Wednesday	1235
5	Thursday	1229
6	Friday	828
7	Saturday	751
Total		**7919**

You can easily query star schemas and cubes in a variety of BI tools like **Excel, PowerBI,** or **Tableau.**

Demo: The Data Warehouse in Action

School of Information Studies
Syracuse University

Demo: Visualizing Adventure Works Internet Orders With Excel

1. Explain "Adventure Works"
2. Explore OLTP
3. Explore the DW
4. Example of BI in action

As you watch the demo take notes:

- Write down any questions you have as they arise during the demo
- Did the demo help clarify doubts or misconceptions you may had had from earlier in the lesson? Which ones?

Kimball and Inmon

School of Information Studies
Syracuse University

The Fathers of Data Warehousing

	W. H. Inmon	Ralph Kimball
The "Father" of…	Data warehousing	Business intelligence
Invented:	Data warehousing	Dimensional models
Data warehouse is:	Normalized tables	Dimensional models
Purpose of a data warehouse:	Data integration	Query
Million-dollar idea:	"Corporate information factory"	"Kimball lifecycle"
Approach: How is the Data Warehouse built?	Data-first (iterative, bottom-up)	Process-first (waterfall, top-down)

Kimball vs. Inmon

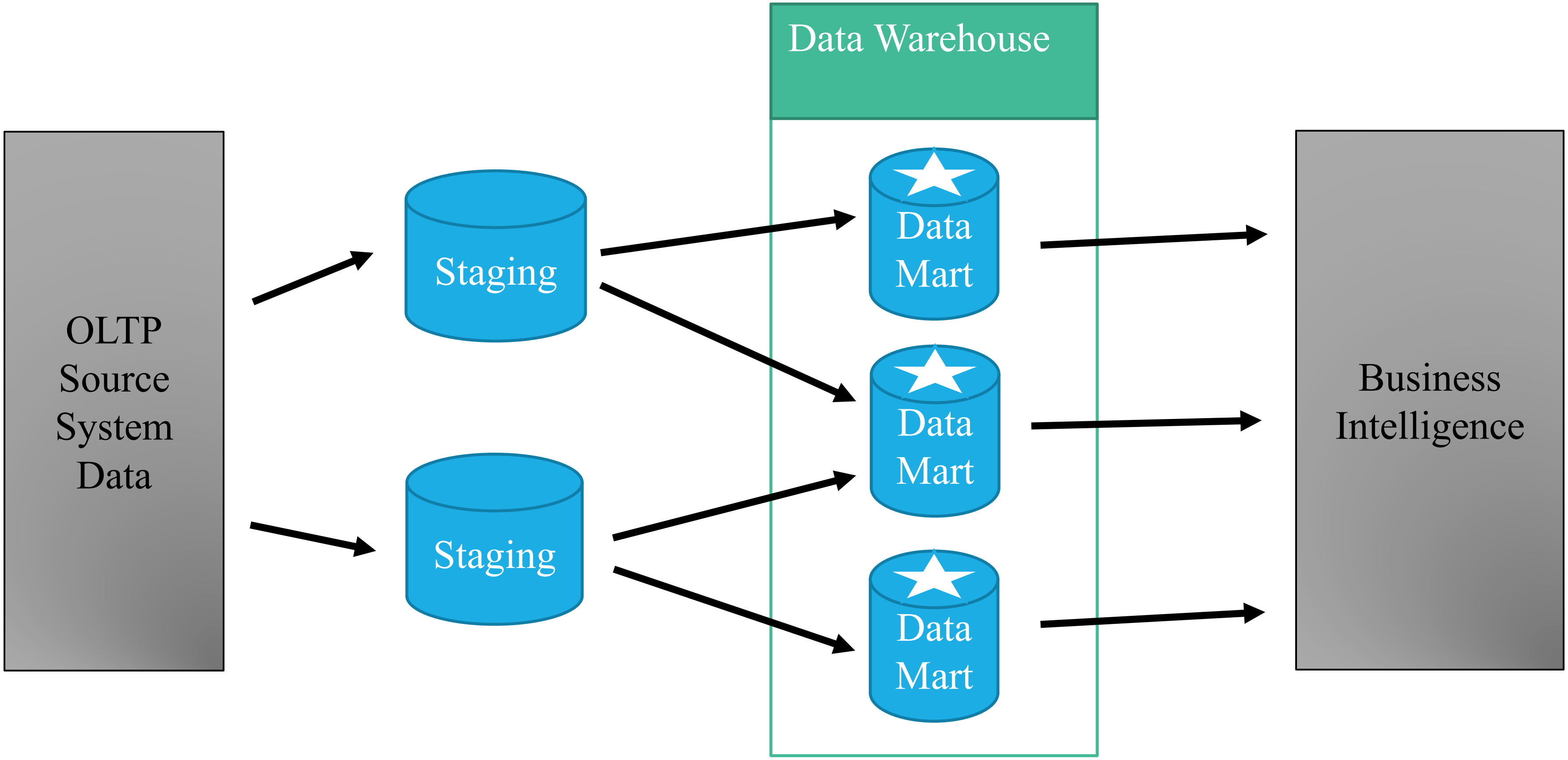
Inmon Data Warehouse

- **Relational modeling**
- Entity-relationship model
- Tables in third normal form
- Many tables using joins
- Built for data integration
- Indirect access of data by users

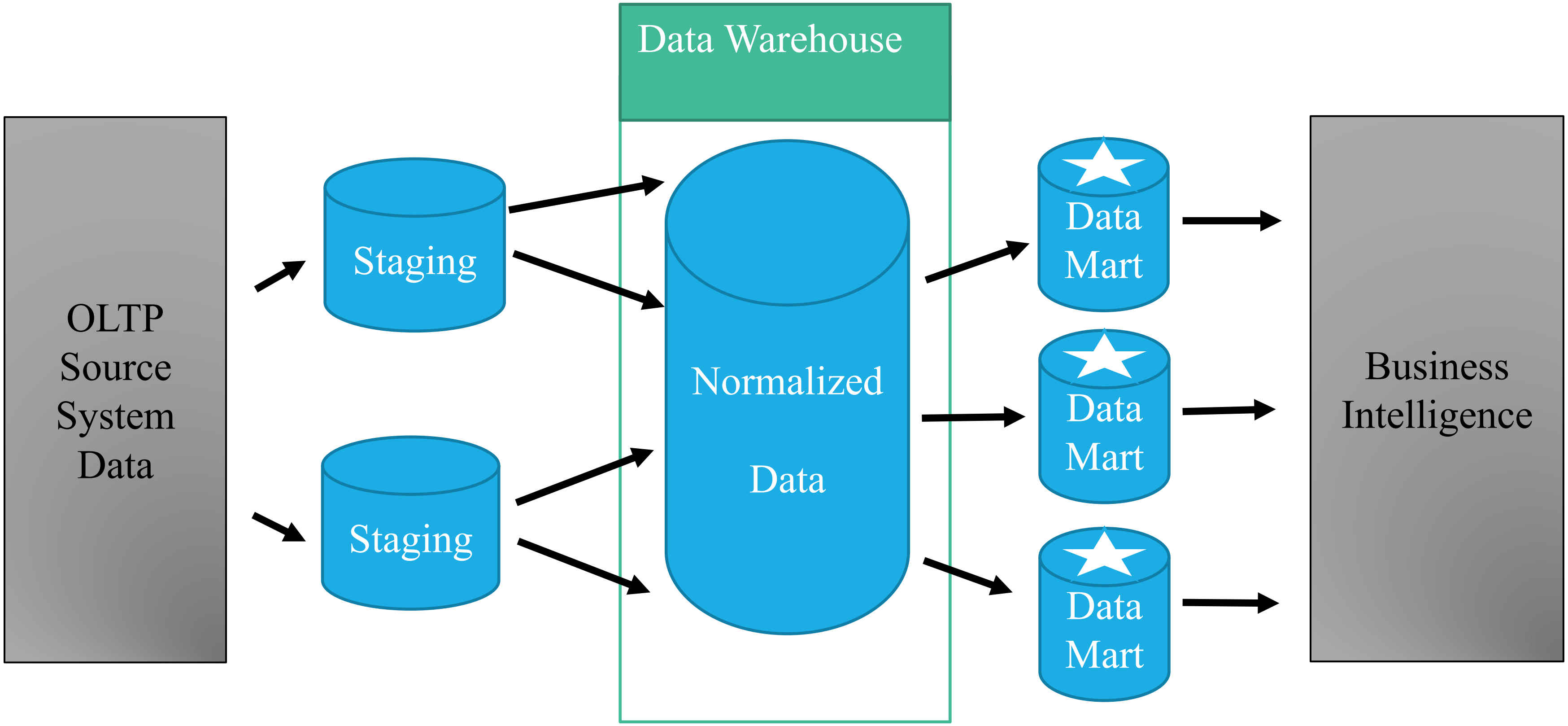
Kimball Data Warehouse

- **Dimensional modeling**
- Fact tables and dimensions/star schema
- Tables are denormalized
- Easier for end users to understand
- Built for ad hoc queries
- Direct access of data by users

Kimball Data Warehouse



Inmon Data Warehouse



School of Information Studies
Syracuse University

Why Inmon Data Warehouse?

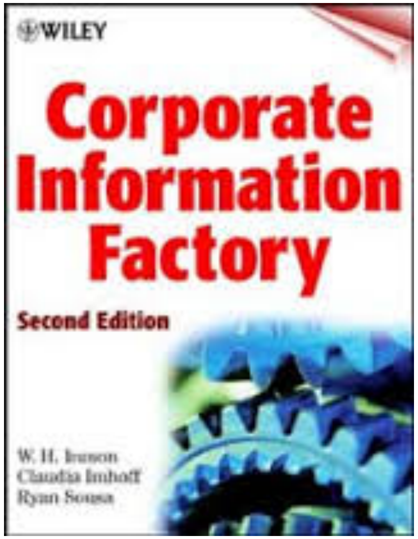
If the goal is data mart, why go through the added complexity?

- Building data marts from source data is more difficult than from data warehouse data.
- Normalized DW data can be used for a variety of purposes beyond analytical queries.
- Less stress on source systems.
- "Single version of the truth."

Your Textbooks

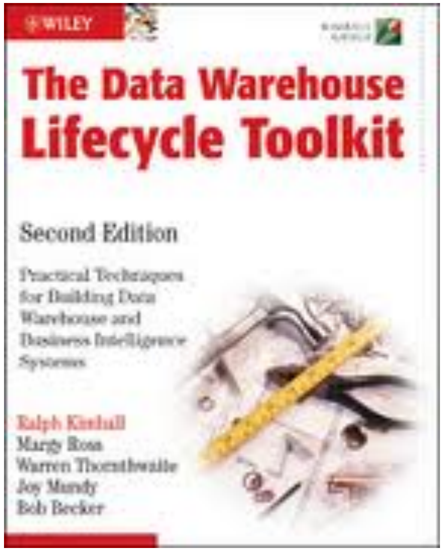
“What”

Inmon



“How To”

Kimball


WILEY
The Data Warehouse Lifecycle Toolkit
Second Edition
Practical Techniques for Building Data Warehouse and Business Intelligence Systems
Ralph Kimball
Margy Ross
Warren Thornthwaite
Joy Mundy
Bob Becker

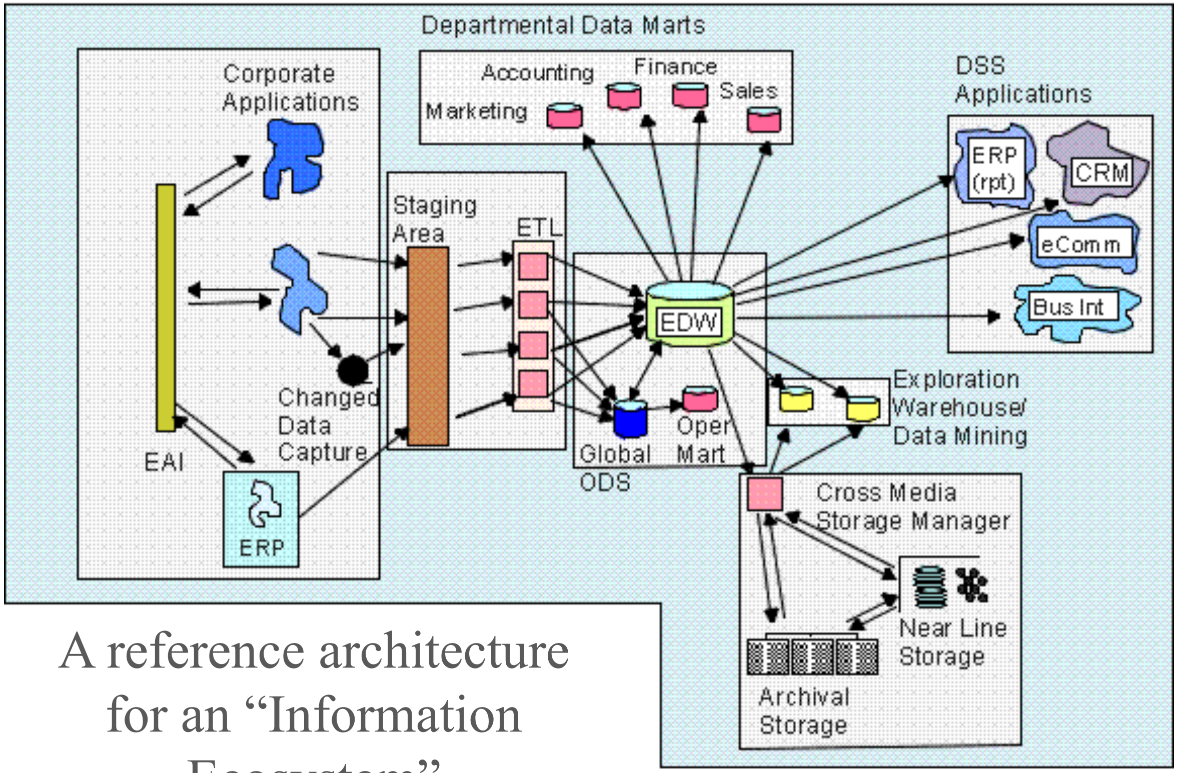
We’ll use the Inmon definitions and apply the Kimball approach.

The Corporate Information Factory

School of Information Studies
Syracuse University

Inmon’s Corporate Information Factory

Corporate Information Factory



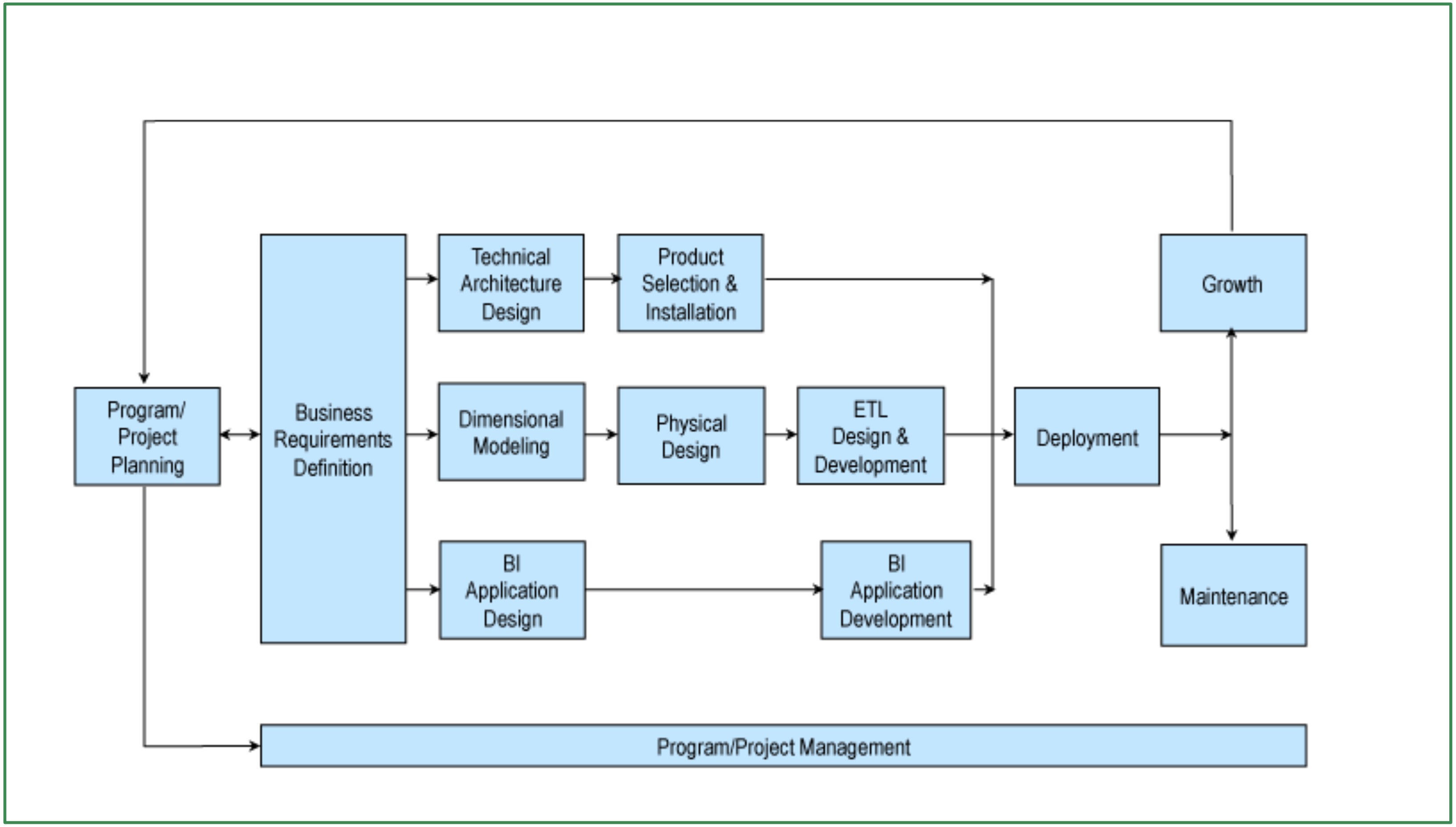
A reference architecture for an “Information Ecosystem”

by Bill Inmon and Claudia Imhoff

The Kimball Lifecycle

School of Information Studies
Syracuse University

The Kimball Lifecycle


Program/ Project Planning
Business Requirements Definition
Technical Architecture Design
Product Selection & Installation
Dimensional Modeling
Physical Design
ETL Design & Development
BI Application Design
BI Application Development
Deployment
Growth
Maintenance
Program/Project Management

Class Case Studies Overview

School of Information Studies
Syracuse University

Our Case Studies

- Sample OLTP systems
- Highly normalized
- Represent actual business and their processes
- Used in your homework, labs, for in-class demos, and for your group project

OLTP Databases Used in This Class

Northwind

- Fictitious company called Northwind Traders, which deals in the import/export of specialty foods
- Used in homework and labs

Fudgemart and FudgeFlix

- Fictitious conglomerate Fudgemart, Inc., with two subsidiaries: one in e-commerce and the other in the movie rental business
- Used for in-class demos and student projects

Data Profiling

- Examining your data so that you understand its characteristics:
  - Purpose of the data
  - What "one row" of the data means
  - How the tables connect to each other
  - Business keys
  - Assess the quality of the data.
- "Getting to know your data" because…
- "You cannot model that which you do not understand."

Walkthrough: OLTP Databases

Walk through the case study databases:

1. Northwind
2. Fudgemart
3. FudgeFlix
4. External sources