CIS 9440 - Data Warehousing and Analytics

Class #3

What’s on Blackboard?

- HW #1, due before class on Wednesday, 9/28
  - Please submit on Blackboard using *submission template*
- Final Project groups assigned
- Next week - class <u>asynchronous</u>
- Install python with Anaconda
- If you do not already have a Text Editor, consider downloading one:
  - Notepad ++, Atom, Sublime, etc.
    - This way we can more easily share SQL, python, etc.
- Office Hours - Poll

Data Warehousing in the News

“Turnkey” Data Warehouse Solution

Week 3 Class Overview:

1. Getting Started with NYC Open Data
1. Analytical SQL (continued)
2. OLTP vs OLAP
3. The 5 C’s of Data Warehousing
4. The Kimball Lifecycle
5. Data Warehouse Project Planning
6. CIS 9440 Final Project Milestones

Week 3 Class Overview:

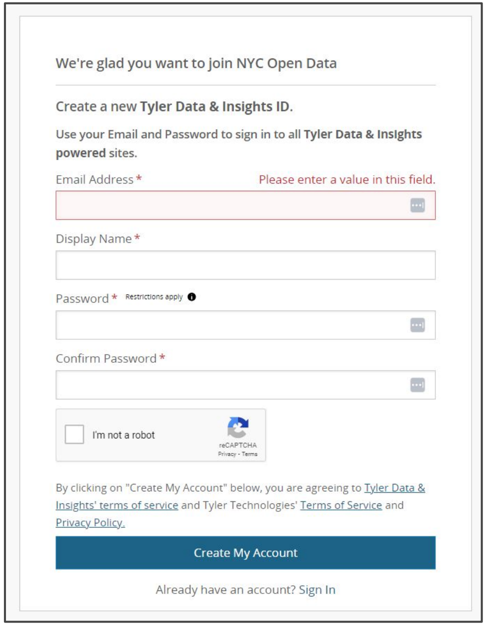
1. **Getting Started with NYC Open Data**
1. Analytical SQL (continued)
2. OLTP vs OLAP
3. The 5 C’s of Data Warehousing
4. The Kimball Lifecycle
5. Data Warehouse Project Planning
6. CIS 9440 Final Project Milestones

NYC Open Data

- NYC Open Data is a government website that provides public, updated New York City datasets
- To use NYC Open Data for your final project, you will need to leverage the websites API
- Let’s setup your API now, go to https://data.cityofnewyork.us
  - After creating a Developer Account, go to https://data.cityofnewyork.us/profile/edit/developer_settings to create an API Token
  - Then, we can use python to pull data with sodapy https://pypi.org/project/sodapy/

NYC Open Data

- https://data.cityofnewyork.us/signup



We're glad you want to join NYC Open Data

Create a new **Tyler Data & Insights ID.**

Use your Email and Password to sign in to all **Tyler Data & Insights powered** sites.

Email Address * Please enter a value in this field.

Display Name *

Password * Restrictions apply

Confirm Password *

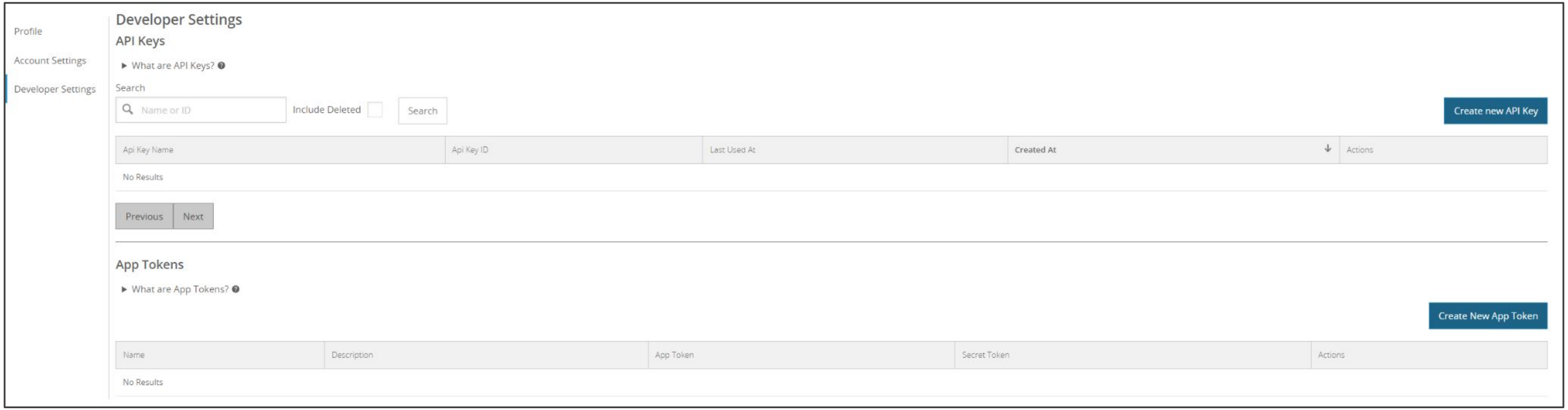
☐ I'm not a robot

By clicking on "Create My Account" below, you are agreeing to Tyler Data & Insights' terms of service and Tyler Technologies' Terms of Service and Privacy Policy.

Already have an account? Sign In

NYC Open Data

- https://data.cityofnewyork.us/profile/edit/developer_settings
- Click “Create new API Key”



Developer Settings

API Keys

▸ What are API Keys?

Search

Name or ID

Include Deleted □

Api Key Name	Api Key ID	Last Used At	Created At ↓	Actions
No Results				

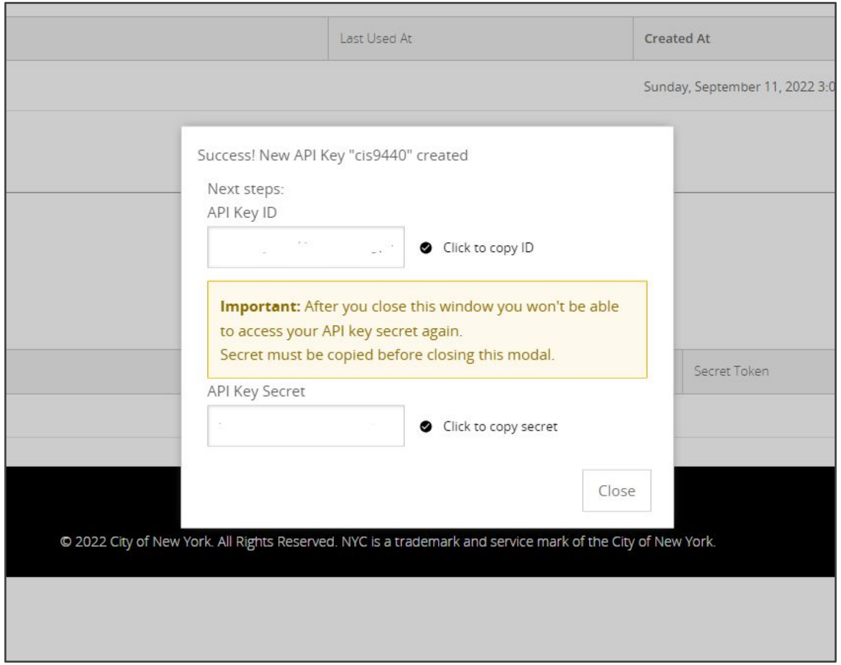
App Tokens

▸ What are App Tokens?

Name	Description	App Token	Secret Token	Actions
No Results				

NYC Open Data

- Copy your API Key ID and API Key Secret to a file on your computer



Sunday, September 11, 2022 3:0

Success! New API Key "cis9440" created

Next steps:

API Key ID

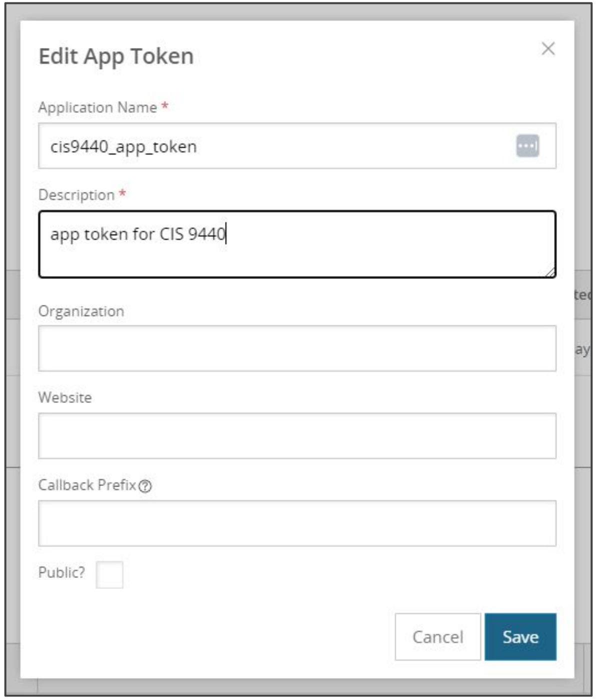
Important: After you close this window you won't be able to access your API key secret again.
Secret must be copied before closing this modal.

API Key Secret

© 2022 City of New York. All Rights Reserved. NYC is a trademark and service mark of the City of New York.

NYC Open Data

- Click “Create new App Token”



Edit App Token

Application Name *

cis9440_app_token

Description *

app token for CIS 9440

Organization

Website

Callback Prefix

Public? □

NYC Open Data

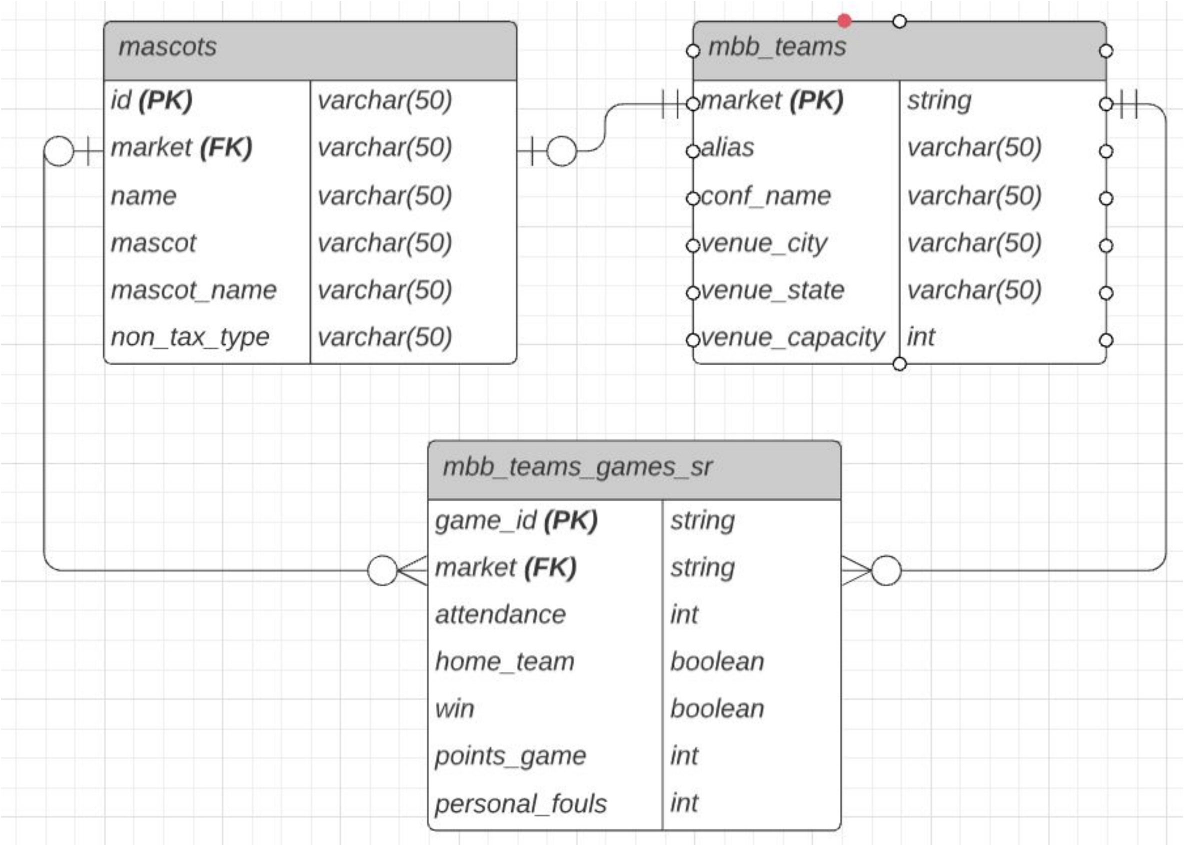
- You’re all set with the API for now!
- Remember your login info, and your API Key ID and Secret for future classes
- To access your API Key Token you can always go back to https://data.cityofnewyork.us/profile/edit/developer_settings

Now, let’s explore how to find NYC Public Datasets: https://opendata.cityofnewyork.us/data/

Week 3 Class Overview:

1. Getting Started with NYC Open Data
1. **Analytical SQL (continued)**
2. OLTP vs OLAP
3. The 5 C’s of Data Warehousing
4. The Kimball Lifecycle
5. Data Warehouse Project Planning
6. CIS 9440 Final Project Milestones

Data Model for today’s class



mascots	
id **(PK)**	varchar(50)
market **(FK)**	varchar(50)
name	varchar(50)
mascot	varchar(50)
mascot_name	varchar(50)
non_tax_type	varchar(50)

mbb_teams	
market **(PK)**	string
alias	varchar(50)
conf_name	varchar(50)
venue_city	varchar(50)
venue_state	varchar(50)
venue_capacity	int

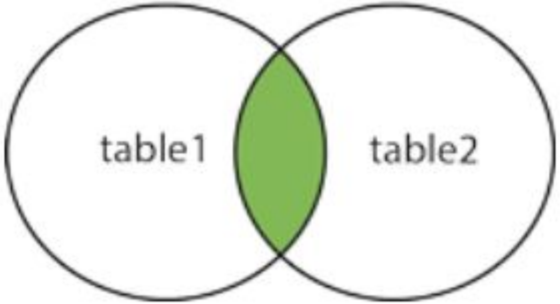
mbb_teams_games_sr	
game_id **(PK)**	string
market **(FK)**	string
attendance	int
home_team	boolean
win	boolean
points_game	int
personal_fouls	int

*Made with Lucidchart

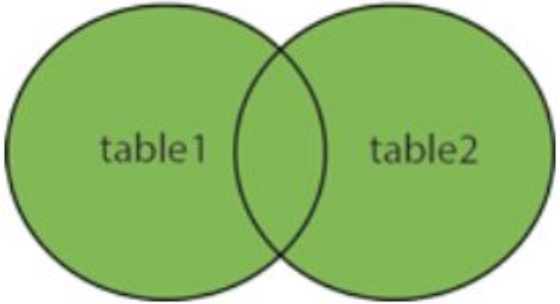
What is a JOIN in SQL?

We use a JOIN to query columns from multiple tables in SQL. We specify how we link a JOIN. For example, if we wanted information about a college from both the mascots and mbb_teams table, we could JOIN these tables on the college name.

INNER JOIN

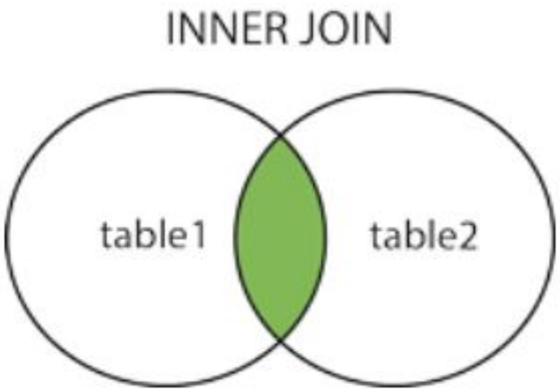


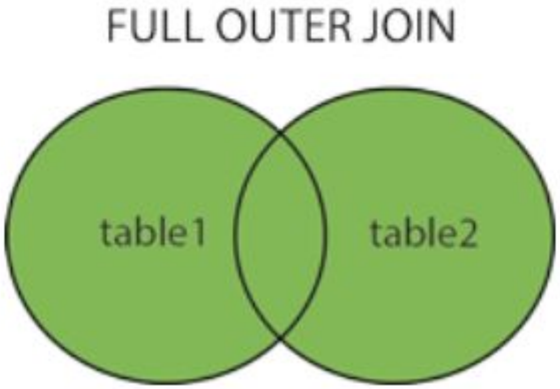
FULL OUTER JOIN



Types of Joins

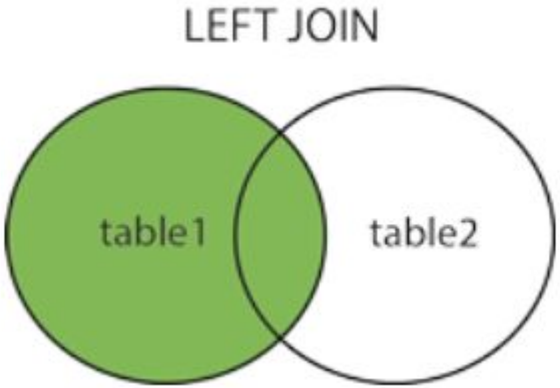
- Inner Joins return records that have matching values in both tables.
- Outer Joins return all records for both left and right tables.

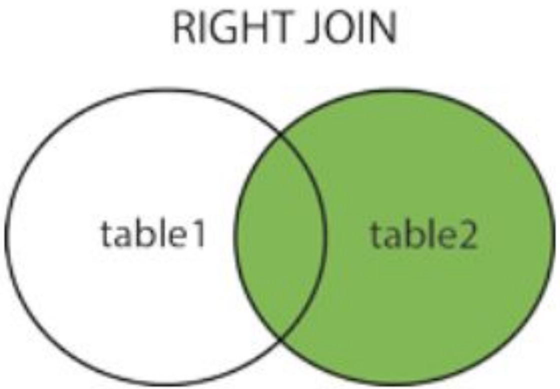




Types of Joins (continued)

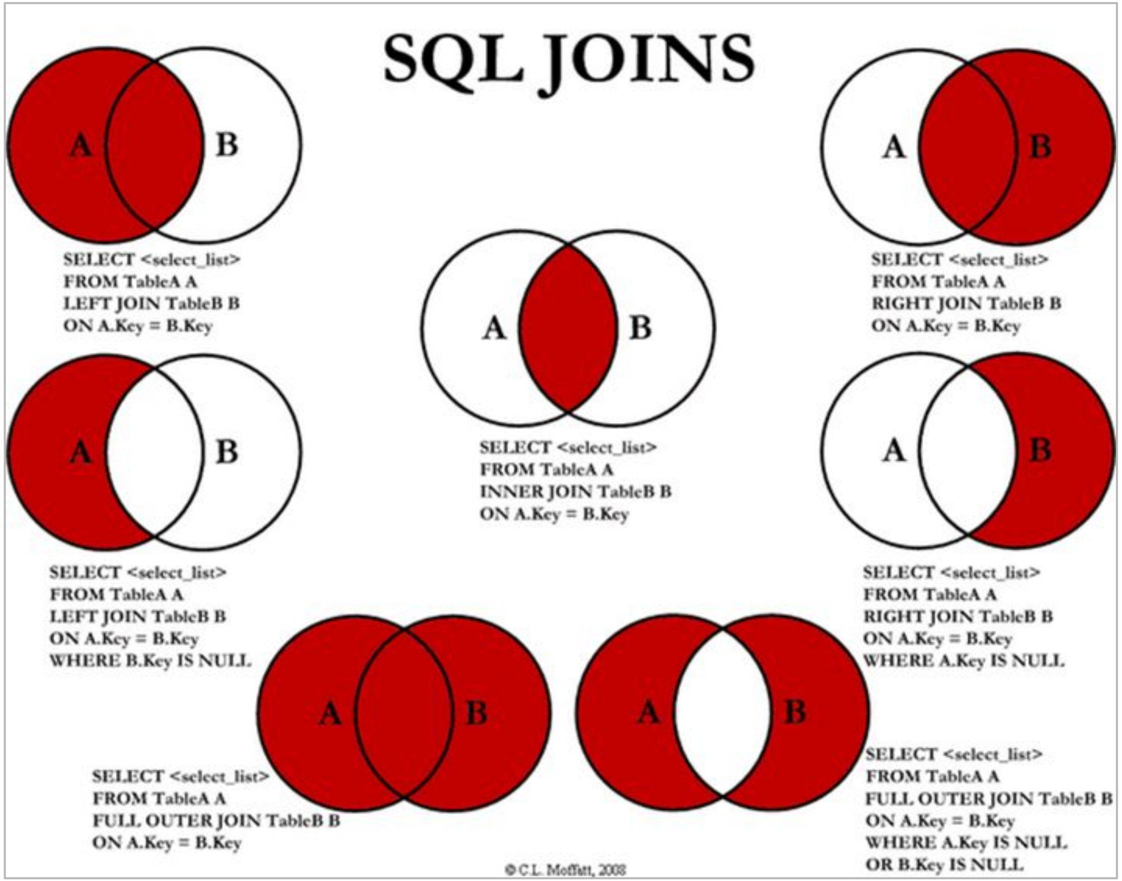
- Left Joins return all records from left table, and matching records from right table.
- Right Joins return all records from right table, and matching records from left table.





Joins Syntax (good to print)

Notice the “ON” statement in each JOIN.



SQL JOINS

```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
```

```
SELECT <select_list>
FROM TableA A
INNER JOIN TableB B
ON A.Key = B.Key
```

```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
```

```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
WHERE B.Key IS NULL
```

```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
```

```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
```

```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
OR B.Key IS NULL
```

© C.L. Moffatt, 2008

JOIN

When you have multiple tables that have fields in common and you want to see attributes from both tables.

```
SELECT a.market,

a.mascot_name,

b.conf_name

FROM `bigquery-public-data.ncaa_basketball.mascots` a

JOIN `bigquery-public-data.ncaa_basketball.mbb_teams` b

ON a.market = b.market;
```

JOIN (continued)

When you have multiple tables that have fields in common and you want to see attributes from both tables.

```
SELECT a.market,

a.mascot_name,

b.conf_name

FROM `bigquery-public-data.ncaa_basketball.mascots` a

JOIN `bigquery-public-data.ncaa_basketball.mbb_teams` b

ON a.market = b.market

WHERE b.conf_name IN ("Ivy", "Big Sky")

ORDER BY b.venue_capacity DESC;
```

JOIN (continued)

Top 5 teams and mascots with most total points scored

```
SELECT a.market,

a.mascot,

sum(b.points_game) total_points

FROM `bigquery-public-data.ncaa_basketball.mascots` a

JOIN `bigquery-public-data.ncaa_basketball.mbb_teams_games_sr` b

ON a.market = b.market

GROUP BY a.market, a.mascot

ORDER BY 3 DESC

LIMIT 5;
```

JOIN (continued 2)

Most Common Type of Mascot in North Carolina

```sql
SELECT a.mascot,
COUNT(a.market) AS number_of_mascots
FROM `bigquery-public-data.ncaa_basketball.mascots` a
JOIN `bigquery-public-data.ncaa_basketball.mbb_teams` b
ON a.market = b.market
WHERE b.venue_state = "NC"
GROUP BY a.mascot
ORDER BY number_of_mascots DESC;
```

CASE statement

Create an IF statement in SQL

```
SELECT market, venue_name,

CASE WHEN venue_capacity > 15000 THEN 'large arena'

WHEN venue_capacity > 7500 THEN 'medium arena'

ELSE 'small arena'

END AS arena_size

FROM `bigquery-public-data.ncaa_basketball.mbb_teams`
```

Practice Question 10:

(5 minutes)

Write a query that returns the `market`, `mascot`, and `mascot_name` from all colleges with `conf_name` equal to “Big 12”.

Practice Question Set #3

Use keywords such as **COUNT, DISTINCT, LIMIT, ORDER BY**, **WHERE**, **IN**, **NOT IN**, **AND**, **OR**, **AVG**, **SUM**, **MAX**, **GROUP BY**, **JOIN** to answer the following:

(*Refer to the ER Diagram)

1. What is the market, mascot, and venue_state of the 5 teams with the largest venue_capacity?
2. Which conf_name had the most total wins?

Week 3 Class Overview:

1. Getting Started with NYC Open Data

1. Analytical SQL (continued)

2. OLTP vs OLAP

3. The 5 C’s of Data Warehousing

4. The Kimball Lifecycle

5. Data Warehouse Project Planning

6. CIS 9440 Final Project Milestones

The Goal of the Data Warehouse Project: Technical Success

- Technical Success
  - Go from **OLTP** system to **OLAP** system
    - System of Record to a System of History
    - This requires Data Warehousing processes from source systems to Data Warehouse

Will you hear OLAP or OLTP in the workplace?

OLTP to OLAP

Online Transaction Processing (OLTP)

1. Transaction Processing Systems
2. Operational Systems
3. System of Record (SoR)
4. “Source”
   a. HR
   b. Purchasing
   c. Manufacturing
   d. Inventory
   e. Sales
   f. Customer Support

Online Analytical Processing (OLAP)

1. Analytical Systems
2. Business Intelligence Systems
3. System of History (SoH)
4. “Target”
   a. Data Lake
   b. Data Mart
   c. Data Warehouse (EDW)

OLTP to OLAP (continued)

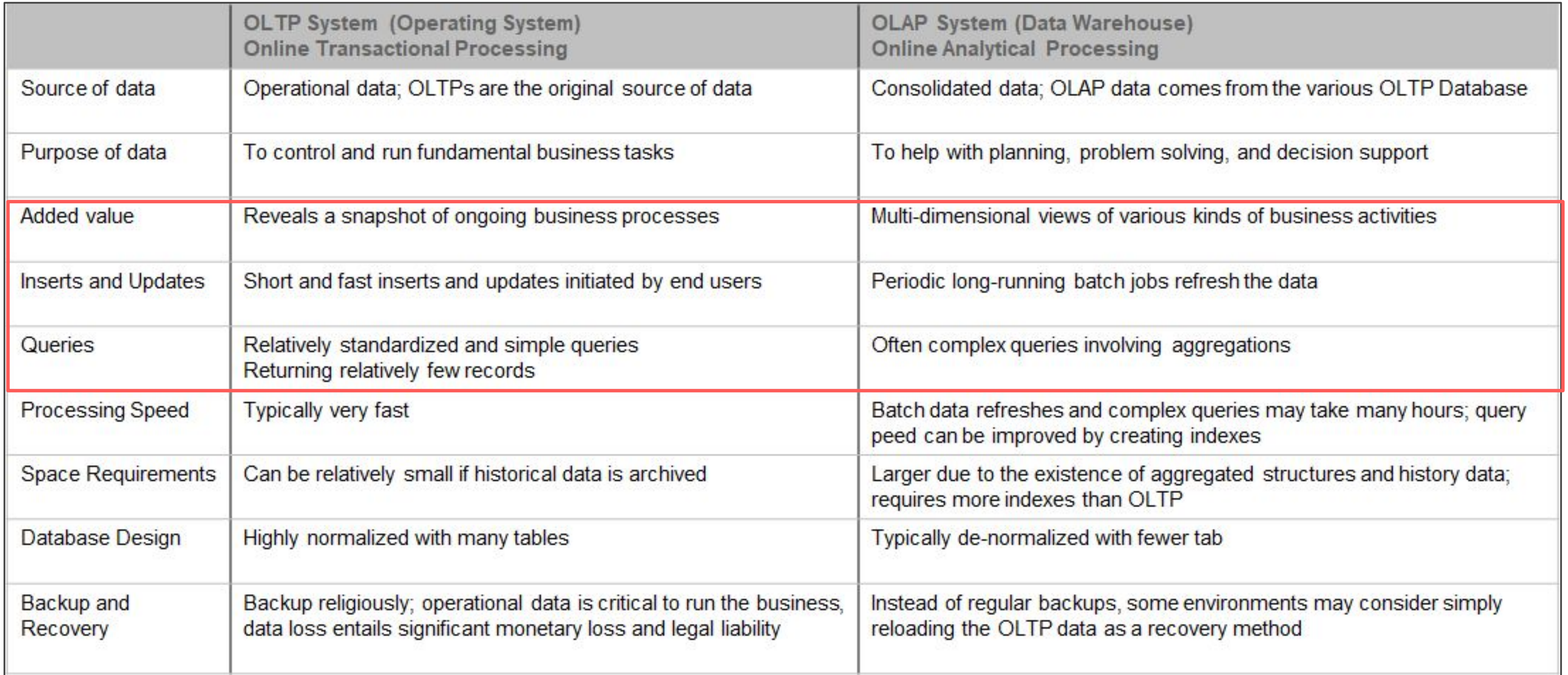
Online Transaction Processing (OLTP)

1. Support a single business functional area
2. Normalized Schema
3. Optimized for small, frequent writing
4. Maintains only recent data

Online Analytical Processing (OLAP)

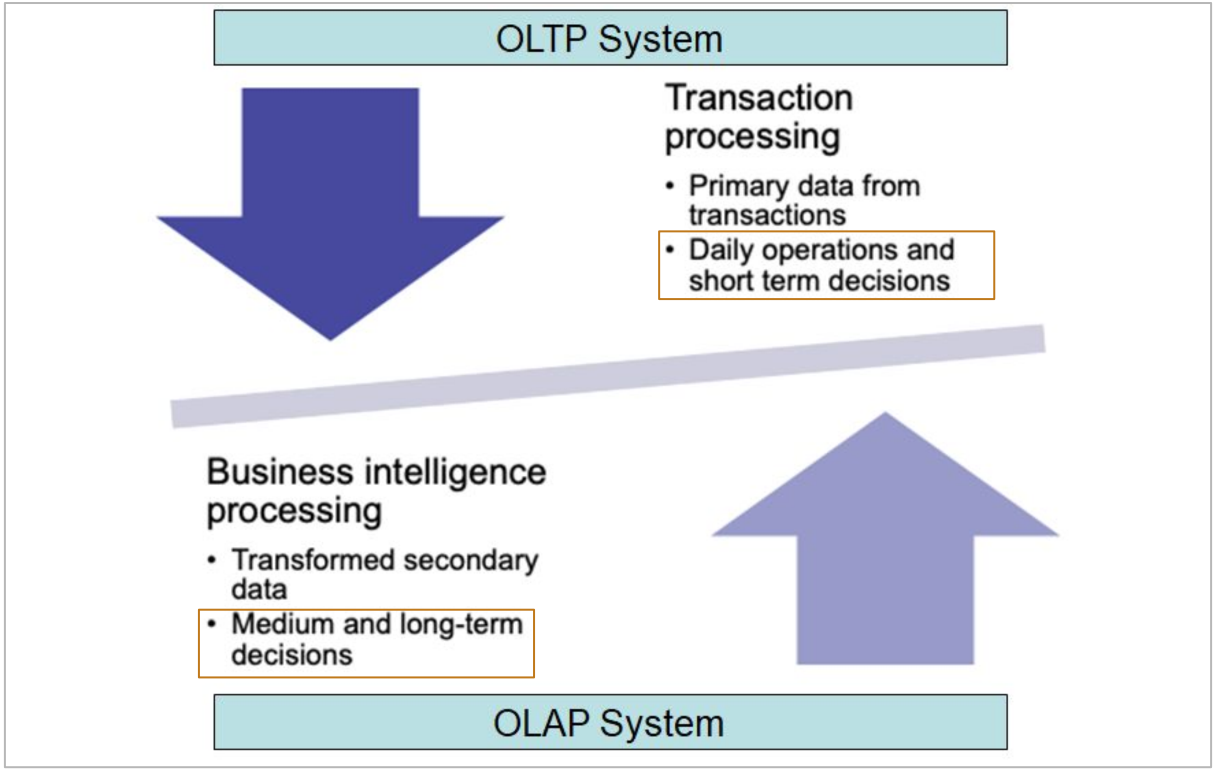
1. Integrate data from multiple business functional areas
2. Logically centralized repository for decision making (de-normlized)
3. Optimized for reading, longer, ad-hoc queries
4. Maintains deep historical data
   - Good for *ML/AI*

OLTP vs OLAP systems:



	OLTP System (Operating System) Online Transactional Processing	OLAP System (Data Warehouse) Online Analytical Processing
Source of data	Operational data; OLTPs are the original source of data	Consolidated data; OLAP data comes from the various OLTP Database
Purpose of data	To control and run fundamental business tasks	To help with planning, problem solving, and decision support
Added value	Reveals a snapshot of ongoing business processes	Multi-dimensional views of various kinds of business activities
Inserts and Updates	Short and fast inserts and updates initiated by end users	Periodic long-running batch jobs refresh the data
Queries	Relatively standardized and simple queries Returning relatively few records	Often complex queries involving aggregations
Processing Speed	Typically very fast	Batch data refreshes and complex queries may take many hours; query peed can be improved by creating indexes
Space Requirements	Can be relatively small if historical data is archived	Larger due to the existence of aggregated structures and history data; requires more indexes than OLTP
Database Design	Highly normalized with many tables	Typically de-normalized with fewer tab
Backup and Recovery	Backup religiously; operational data is critical to run the business, data loss entails significant monetary loss and legal liability	Instead of regular backups, some environments may consider simply reloading the OLTP data as a recovery method

OLTP vs OLAP systems (continued)



OLTP System

Transaction processing

- Primary data from transactions
- Daily operations and short term decisions

Business intelligence processing

- Transformed secondary data
- Medium and long-term decisions

OLAP System

Activity (3 minutes):

Think about a company/organization that must currently go from a OLTP to an OLAP system with their data.

Why do they have to go from OLTP to OLAP?

Week 3 Class Overview:

1. Getting Started with NYC Open Data
1. Analytical SQL (continued)
2. OLTP vs OLAP
3. **The 5 C’s of Data Warehousing**
4. The Kimball Lifecycle
5. Data Warehouse Project Planning
6. CIS 9440 Final Project Milestones

Data in your OLAP system:
The Five C's of Data Warehousing Data

- Clean
- Consistent
- Conformed
- Current
- Comprehensive

- Why are the 5 C's most important to an OLAP system? (Data Warehouse)

Five C’s of Data Warehousing Data Errors

Examples of data that is not Clean:

- **Invalid Entry**: In an HR system, a person’s first record is a Termination rather than a Hire (e.g., conversions)
- **Missing Item**: In an active students directory, a student is listed with a minor but no major
- **Duplicate Rows**: In a table of upcoming events, one event is listed twice. Each record has a different number of listed attendees.

(Think about data cleansing and constraints during Data Integration)

```
SELECT spc_common,
count(*) total_trees
FROM `bigquery-public-data.new_york_trees.tree_census_2015`
GROUP BY spc_common;
```

Five C’s of Data Warehousing Data Errors

Examples of data that is not Consistent:

- **Conflicting results**: An organization produces multiple sales reports - by category, by month, and by vendor. The total sales of the reports are not equal.

How does this happen?

- Data is stored in multiple places, it should be stored in one conformed dimension
- Data is aggregated at different grains in different tables
  - There should only be one source of truth! At the (atomic) lowest grain.

```
SELECT count(distinct repo_name) FROM `bigquery-public-data.github_repos.licenses`;

SELECT count(distinct repo_name) FROM `bigquery-public-data.github_repos.files`;
```

Five C’s of Data Warehousing Data Errors

Examples of data that is not Conformed:

- **Different units**: Some educational reports show students tenure by semesters, others show by total credits completed
- **Units vs Dollars**: Half the sales reports are printed in Sales Units, while the other half shows Sales Dollars
- **Multiple Hierarchies**: HR uses the reporting hierarchy but Sales uses a geographical hierarchy

(this is alleviated by Dimensional Modeling)

Five C’s of Data Warehousing Data Errors

Examples of data that is not <u>Comprehensive or Current:</u>

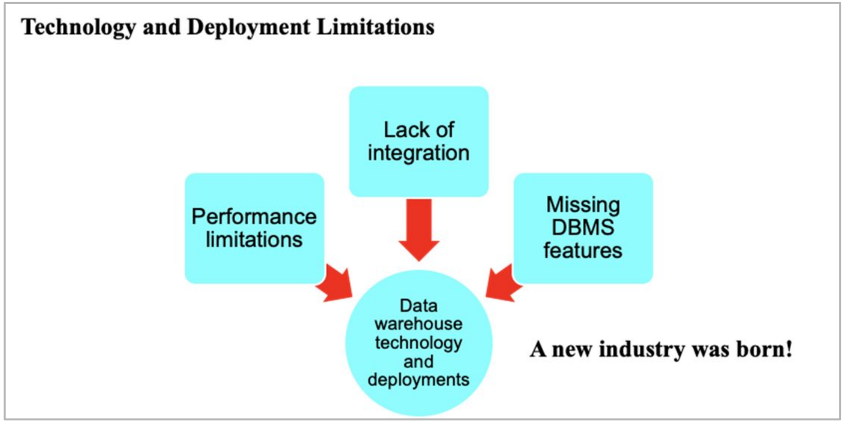
- **Out of date**: The data is not updated on a frequent enough basis, so the data is not useful
- **Not in an available system**: The data for returns does not live in the sales database, and needs a special request to attain

```sql
SELECT year, count(distinct gameId)
FROM `bigquery-public-data.baseball.games_wide`
group by year
order by year desc;
```

```sql
SELECT * FROM `nba-project-301800.nba_data.nba_finals_teams`
ORDER BY Year DESC;
```

How Data Warehousing ensures the Five C's?

Technology and Deployment Limitations



A new industry was born!

- Characteristics:
  - Shares dimensions
  - Table constraints
  - Data Stewards
  - Logically centralized repository for decision making
  - Integrated and Transformed data optimized for reporting

Activity (3 minutes):

With the company you thought of that goes from OLTP to OLAP, what are 3 ways they ensure the 5 C’s of Data Warehousing?

Week 3 Class Overview:

1. Getting Started with NYC Open Data

1. Analytical SQL (continued)

2. OLTP vs OLAP

3. The 5 C’s of Data Warehousing

4. **The Kimball Lifecycle**

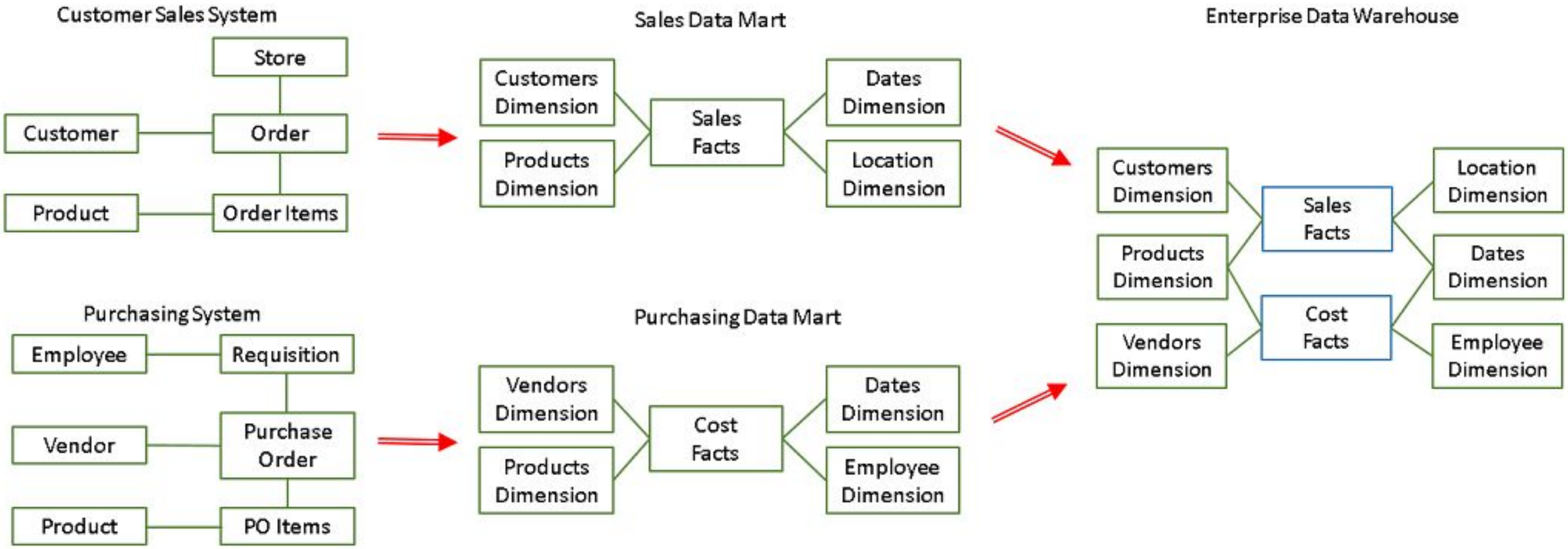
5. Data Warehouse Project Planning

6. CIS 9440 Final Project Milestones

The Kimball Lifecycle

- The Kimball Lifecycle is a roadmap of high-level tasks that drives a successful DW/BI Project
- The Kimball Lifecycle developed in the **1980’s** and has been utilized by thousands of data warehouse and business intelligence (DW/BI) project teams across virtually every industry, application area, business function, and technical platform (DWTK).
- The core focus of the Kimball Lifecycle / “Business Dimensional Lifecycle approach”:
    - Focus on adding business value across the enterprise
    - Dimensionally structure the data that’s delivered to the business
    - **Iteratively develop** the DW/BI environment in manageable lifecycle increments rather than attempting a galactic Big Bang approach

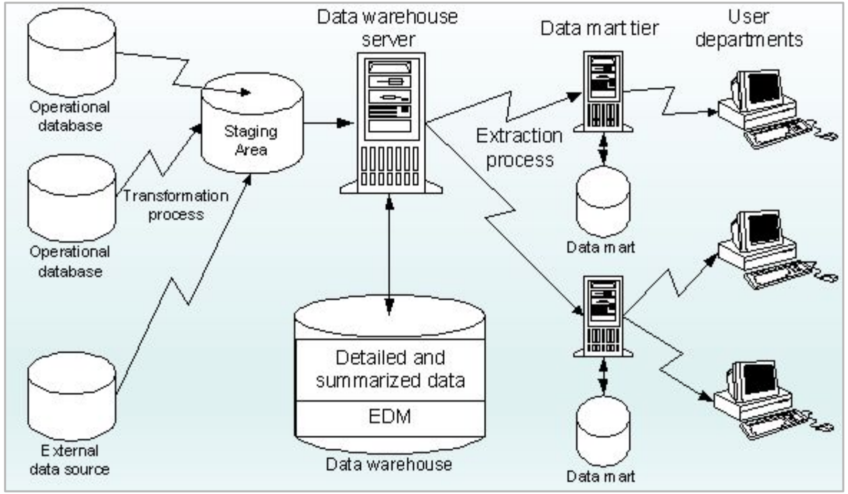
Kimball Architecture (more details)



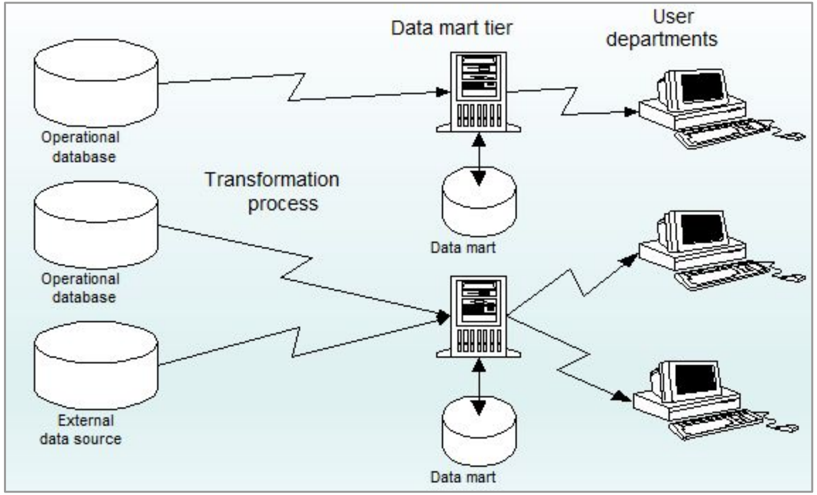
How does this compare to Inmon?

Are there other approaches?
Kimball vs Inmon Architecture

Inmon: Top-Down



Kimball: Bottom-Up



Difference between Kimball Lifecycle and Systems Development Lifecycle (SDLC)

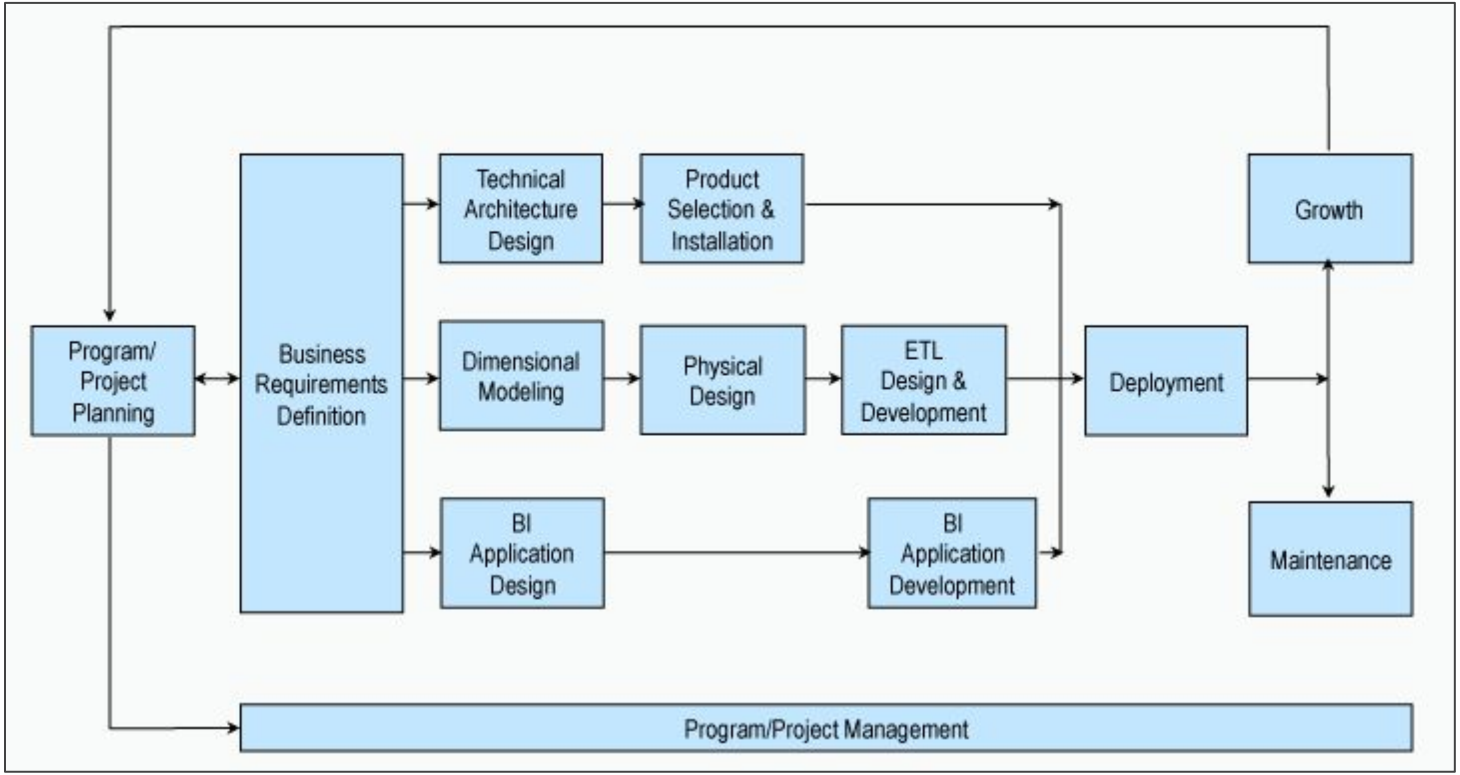
Typical SDLC

1. Identifying business opportunity or problem
2. Perform feasibility study
3. Gather user requirements
4. Develop data and application models
5. Select deployment hardware and software
6. Code data models and applications
7. Write documentation
8. Deploy testing environment
9. Deploy production system
10. Maintain production system

Kimball Lifecycle differences

1. More flexible front-end applications
2. Driven by business users
3. Data profiling and integration are much more critical
   a. Cannot afford messy data
4. Have more complicated analysis/KPI’s

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

The Kimball Lifecycle is a roadmap of high-level tasks required for successful DW/BI Projects

Refer to this roadmap for your Final Project!

DWTK, Chapter 17

Week 3 Class Overview:

1. Getting Started with NYC Open Data
1. Analytical SQL (continued)
2. OLTP vs OLAP
3. The 5 C’s of Data Warehousing
4. The Kimball Lifecycle
5. **Data Warehouse Project Planning**
6. CIS 9440 Final Project Milestones

Kimball Lifecycle: Project Planning

Business and Technical Justification: In this phase, the project's sponsors detail the business justification, opportunities and benefits as well as the technical justifications for the DW project. Staffing and other necessary resources are identified.

- Business Justification: What is the business benefit?
- Technical Justification: Is this technically feasible?

DWTK, page 406

Business Intelligence Guidebook, list on page 46

Kimball Lifecycle: Project Planning

Business Justification:

- Review business initiatives and processes
- Enlist BI sponsors and stakeholders (e.g., potential BI users)
- Document benefits and outcomes in terms of adding business value
- Project scope and Budgeting

Technical Justification:

- Assessing necessary technical skills/expertise
- Assessing data current quality
- Product evaluations for proof-of-concept and technology roadmaps

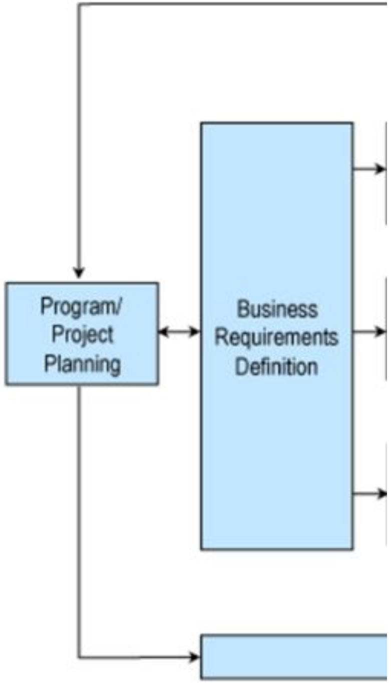
DWTK, page 406

Business Intelligence Guidebook, list on page 46

Project Planning Example 1 - Amazon

Project Planning Example 2 - NYS

Kimball Lifecycle: Requirements Gathering

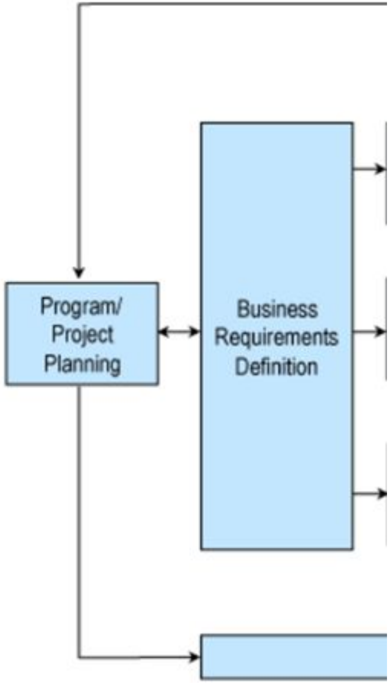


Business Requirements Definition (KPI’s): In this phase business users are interviewed to determine what measurements / metrics they require. Essentially, you’re drilling down a layer deeper than the previous phase.

These **Key Performance Indicators** (KPI’s) are often calculated by summing, transforming, and combining OLTP transactions data. This stage requires deep involvement of the business managers and decision makers.

DWTK, page 410

Kimball Lifecycle: Requirements Gathering



KPI Examples: Suppose you're an Outdoors Retail Company, embarking on DW/BI project to optimize inventory and drive sales. Some KPI's you may create:

- Sales per square foot
- Conversion Rate
- Gross and Net profits
- Average basket value
- Year over year growth
- Stock to sales (inventory turnover)
- Gross Margin ROI

DWTK, page 410

KPI examples

More KPI’s:

Spotify Podcast Data Warehouse

1. Podcasts/Week
2. Average duration/podcast
3. Audience Growth WoW
4. Podcast Rating Monthly Percentage Change

Triathlon Club Data Warehouse

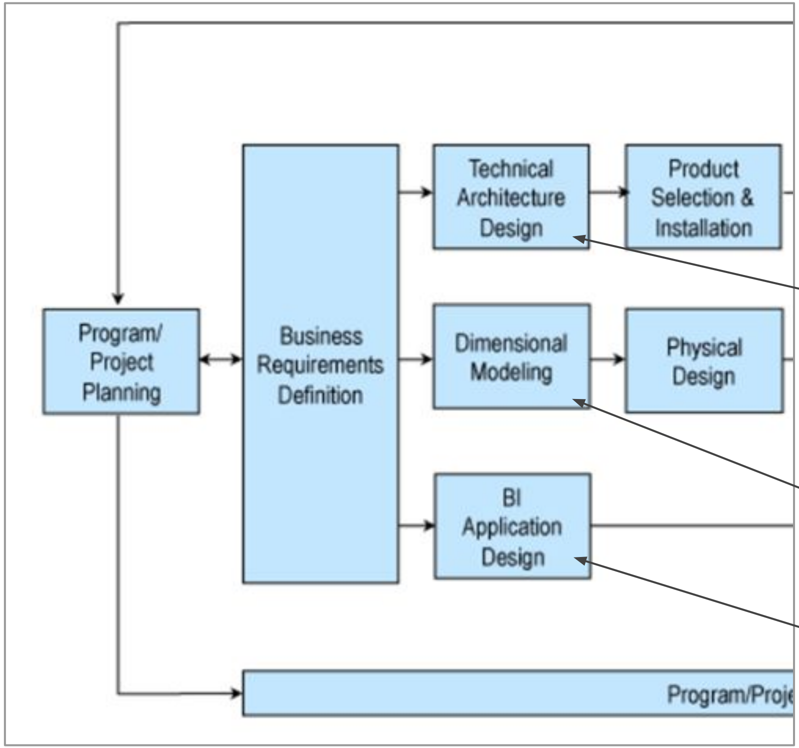
1. Runs/Week
2. Distance/Run
3. Time distribution of events
4. Average Heart Rate per training session

Requirements Gathering Example

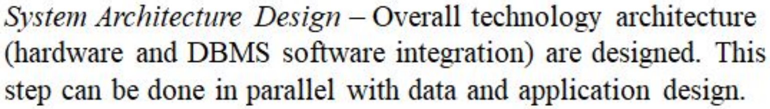
Activity:

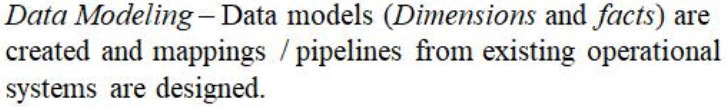
1. Keeping in mind the company you came up with, open the following template: https://docs.google.com/document/d/16Am4jMG09442pV0-wB6xOdt9ewW3YJNY5G9AXVcnfL4/edit?usp=sharing
2. Create a copy in Google Docs
3. On your own, spend 15 minutes planning a Data Warehouse for this company

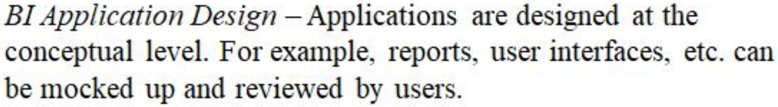
Kimball Lifecycle: System Design


Technical Architecture Design
Product Selection & Installation
Program/ Project Planning
Business Requirements Definition
Dimensional Modeling
Physical Design
BI Application Design
Program/Proje

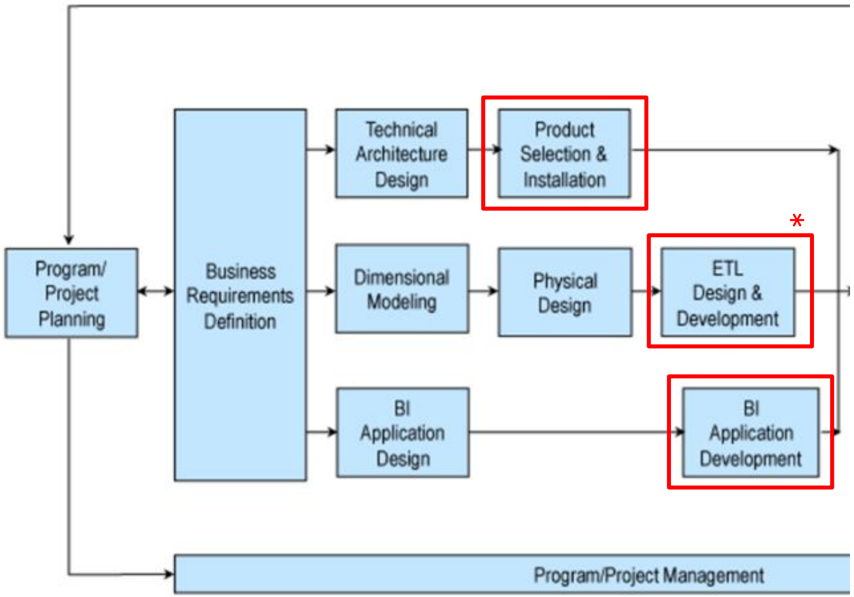
System Design: This phase, the overall system is designed using conceptual models at three levels:


System Architecture Design – Overall technology architecture (hardware and DBMS software integration) are designed. This step can be done in parallel with data and application design.


Data Modeling – Data models (*Dimensions* and *facts*) are created and mappings / pipelines from existing operational systems are designed.


BI Application Design – Applications are designed at the conceptual level. For example, reports, user interfaces, etc. can be mocked up and reviewed by users.

Kimball Lifecycle: System Development

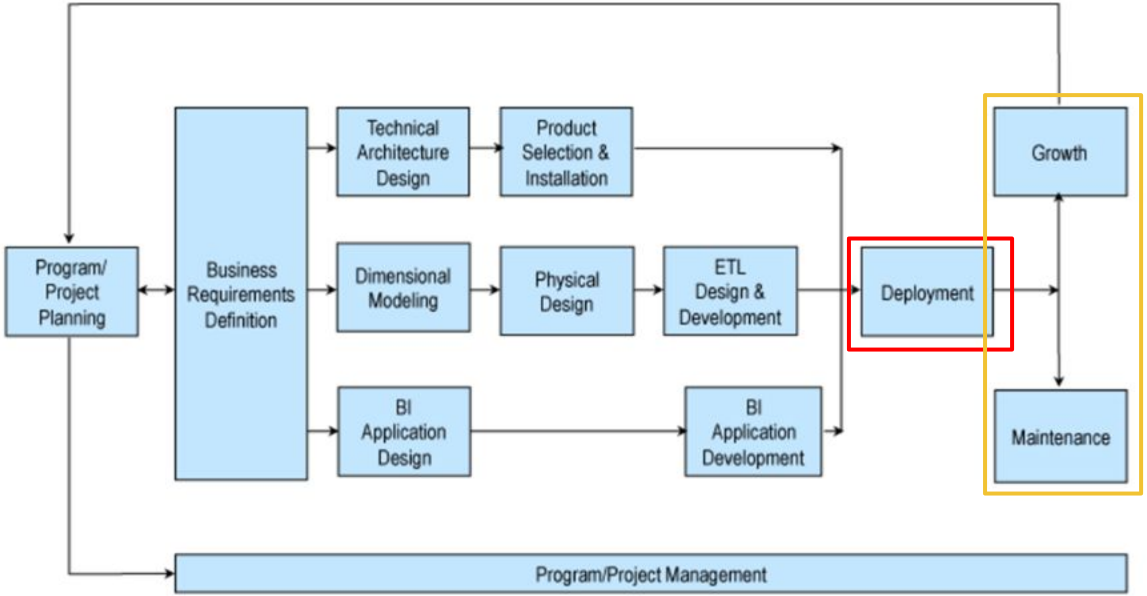

Program/ Project Planning
Business Requirements Definition
Technical Architecture Design
Product Selection & Installation
Dimensional Modeling
Physical Design
*
ETL Design & Development
BI Application Design
BI Application Development
Program/Project Management

System Development: In this phase, the designs are implemented in hardware and software. DBMS vendors are selected, data warehouse schemas are created, ETL code is written/configured, and BI applications are coded.

This stage is carried out almost exclusively by technologists (programmers, DBAs, etc.) although business users may be called upon for testing.

*Note: ETL Design & Development encompasses 34 subsystems (link)

Kimball Lifecycle: Deployment and Growth

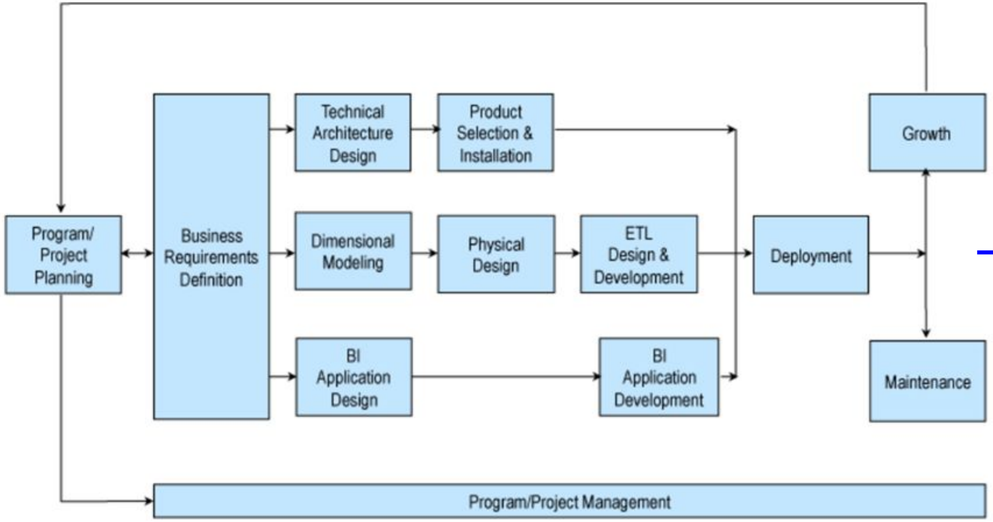

Program/ Project Planning
Business Requirements Definition
Technical Architecture Design
Product Selection & Installation
Dimensional Modeling
Physical Design
ETL Design & Development
BI Application Design
BI Application Development
Deployment
Growth
Maintenance
Program/Project Management

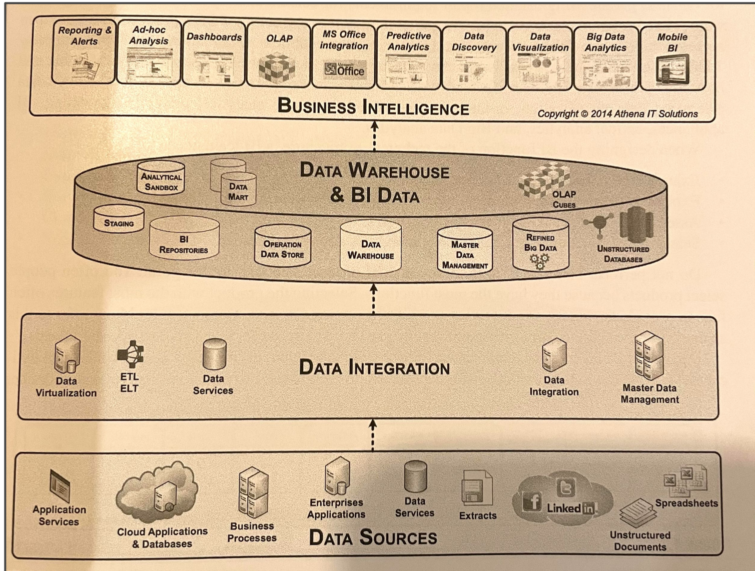
Phased Deployment: In this phase, users are brought on-line (*on boarded*) to the data warehouse.

Maintenance and Growth: Data warehouses undergo continuous evolution as new KPI’s and data sources are defined and integrated.

Kimball Lifecycle

Following the Kimball roadmap will get you to your desired BI destination.

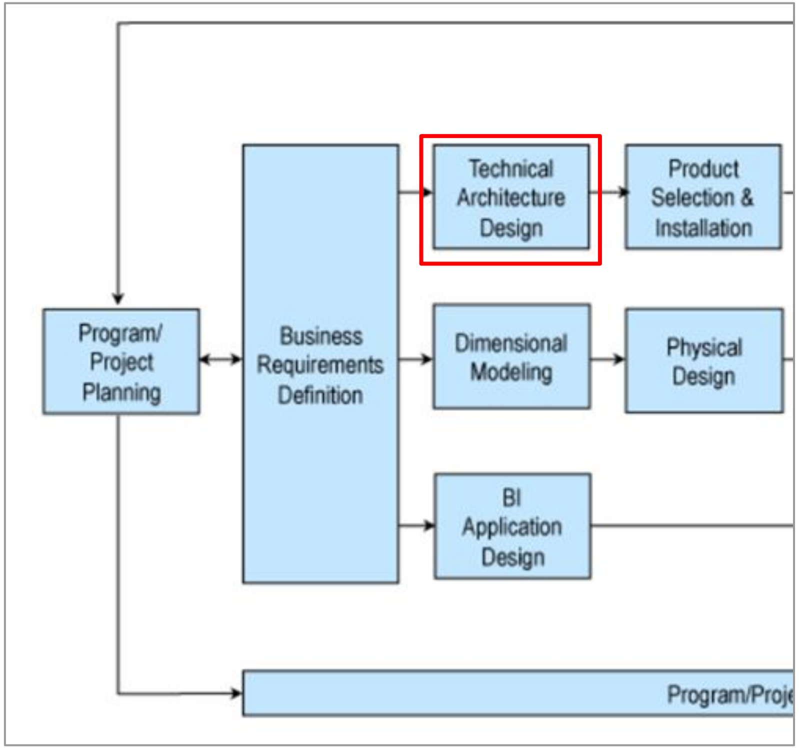

Program/ Project Planning
Business Requirements Definition
Technical Architecture Design
Product Selection & Installation
Dimensional Modeling
Physical Design
ETL Design & Development
BI Application Design
BI Application Development
Deployment
Growth
Maintenance
Program/Project Management



IGURE 4.6

I technical architecture categories.

Kimball Lifecycle: Technical Architecture Design


Technical Architecture Design
Product Selection & Installation
Program/ Project Planning
Business Requirements Definition
Dimensional Modeling
Physical Design
BI Application Design
Program/Proje

The **Technical Architecture Design** is the blueprint for the DW/BI environment’s technical services and infrastructure. This will serve as the framework to support integration for technologies and applications.

In your Business Intelligence Guidebook, the Technical Architecture Design is broken into 4 pieces:

- Information Architecture
- Data Architecture
- Technical Architecture
- Product Architecture

DWTK, page 416

Business Intelligence Guidebook, Page 66

Technical Architecture Design End Result

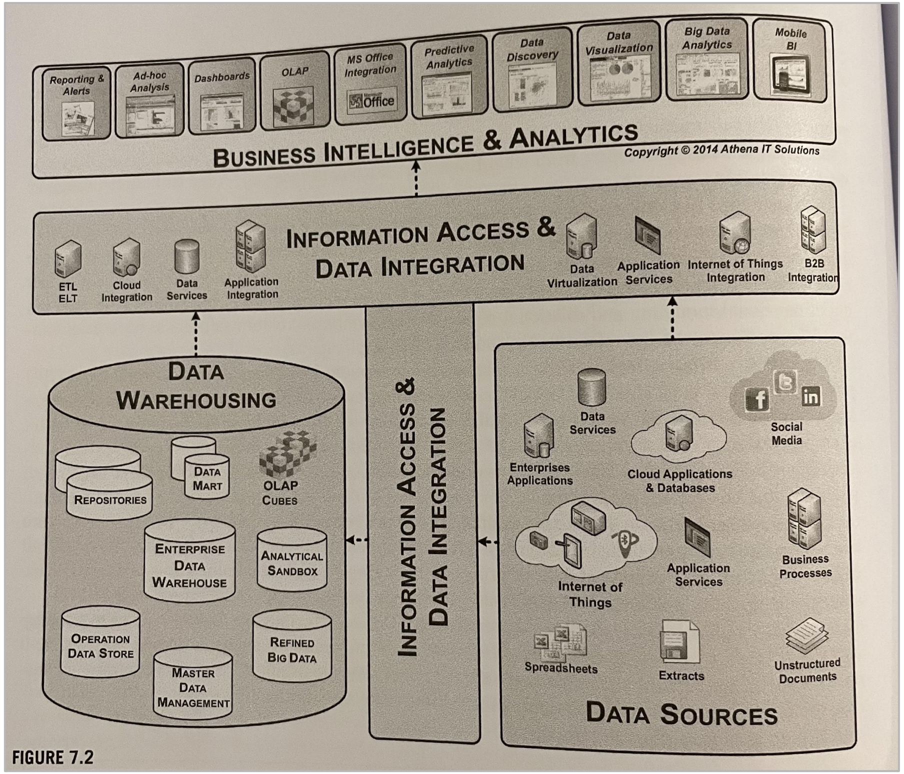
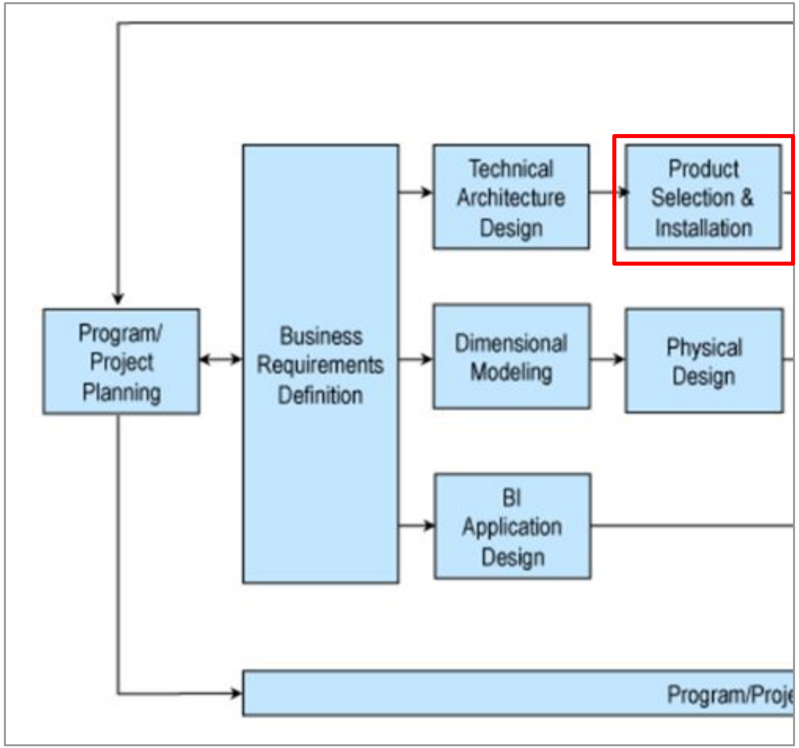


FIGURE 7.2

The product of **Technical Architecture Design** is a blueprint of the pieces you will use:

- How many Data Marts? What types of Reporting? How many data sources? Types of Data Integration tools?
- No Vendor specifics yet. Example: Enterprise push-button reporting, not MicroStrategy

Kimball Lifecycle: Product Selection


Technical Architecture Design
Product Selection & Installation
Program/ Project Planning
Business Requirements Definition
Dimensional Modeling
Physical Design
BI Application Design
Program/Proj

The **Product Selection** is a byproduct of the Technical Architecture Design. The design gave you the shopping list, the product selection must contain vendors that fit the shopping list.

Week 3 Class Overview:

1. Getting Started with NYC Open Data
1. Analytical SQL (continued)
2. OLTP vs OLAP
3. The 5 C’s of Data Warehousing
4. The Kimball Lifecycle
5. Data Warehouse Project Planning
6. **CIS 9440 Final Project Milestones**

Final Project Milestones

1. Project Proposal - *due soon, check your project groups and Bb*
2. Final Dimensional Model
3. Completed ETL Pipeline
4. BI Dashboard Foundation
5. Final Project and Report

Milestone 1 - Project Proposal

Groups submit project ideas for approval

a. Title of your project (include full name and e-mail address of all members), the course number and course section.

b. A narrative description of the business the DW/BI project is intended for. This should include a description of the problem or opportunity being addressed.

c. Business Justification (like done in class)

d. Technical Justification (like done in class)

e. Identification of the information needs – what data sources and information would help solve the problem or allow one to take advantage of the opportunity. Include an initial list of KPI's your group will use to address the issue or opportunity.

f. Informal discussions with the professor can help to refine the project and proposal.

Milestone 2 - Final Dimensional Model

1. Final list of KPI’s
2. All data sources identified
3. Dimensional Model started. Initial list of dimensions and facts that have been identified. This should come naturally from the above discussions. For a group project, I would expect a project to contain at least 5 dimensions and 2 fact tables. (these are your datasets)
    a. This may be done with Lucidchart.com if easiest.

Milestone 3 - Completed ETL Pipeline

1. Source code OR screenshots of application used during ETL
2. Location and size of your datasets
3. Explanation of each step of ETL process, and why you chose to do it that way (prepare to defend your method, think about ways it could break)
4. Overview/Picture of your Data Warehousing Project
    a. With the ETL complete, what does it look like? Data Sources to ETL to where?

Milestone 4 - BI Dashboard Foundation

1. Goal: Create a dashboard application that displays at least 3 different data representations (example: a table, heatmap, and chart). This may be implemented in a BI tool like Tableau.
2. At this Milestone, you must connect your Data Warehouse to a BI tool and show at least the foundation of one of the data representations described above

Milestone 5 - Final Project and Report

1. A separate cover page listing the course, group member's names and project title.
2. An introduction section similar to the proposal section including the narrative description of the business or organization used for the data warehouse being created and description of the source data. Be sure to include appropriate citations of each data source (web site, curator, date accessed, etc.)
3. The dimensional model diagram.
4. A description and screen pictures of each of the ETL processes.
5. Screen shots and brief descriptions of the final schema that the business analytics tools are working with.
6. Screen shots and descriptions of the analytics (at least 3) on the dashboard application developed based on the data warehouse data.
7. Descriptions of the tools (databases, analytics, ETL, programming languages, etc.) used to complete the project.
8. A narrative conclusion section that describes:
9. a) the software and database tools the group used to coordinate and manage the project as well as carry out the programming tasks (list of bullet points with software or service and one sentence of what it was used for)
10. b) the group's experience with the project (which steps were the most difficult? Which were the easiest? what did you learn that you did not imagine you would have? if you had to do it all over again, what would you have done differently?)
11. c) if the proposed benefits can be realized by the new system.
12. d) any final comments and conclusions
13. A References list that provides the web sites and other sources for data, techniques, methods, software, etc. used to complete the project.

Homework:

1. Final Project Milestone 1
2. Reading, BI Guidebook Ch 4, 5, 6, 7
3. Continue on HW #1