CIS 9440 - Data Warehousing and Analytics

Class #6

Refocus: Project Overview

Milestone #2: Dimensional Modeling

- Create the Dimensional Models for the data you found in Milestone #1
- You will use the BUS Matrix to map any Conformed Dimension(s)

Milestone #4: BI Application Foundation

- Map out the Reports/Dashboards/Scorecards you will create to deliver the KPI's you created in Milestone #1
- Properly connect your physical data from Milestone #3 to a BI Application (Tableau)

Milestone #1: Project Planning

- Generate justification for project
- Create >= 5 KPI's that guide data-driven decisions for your audience

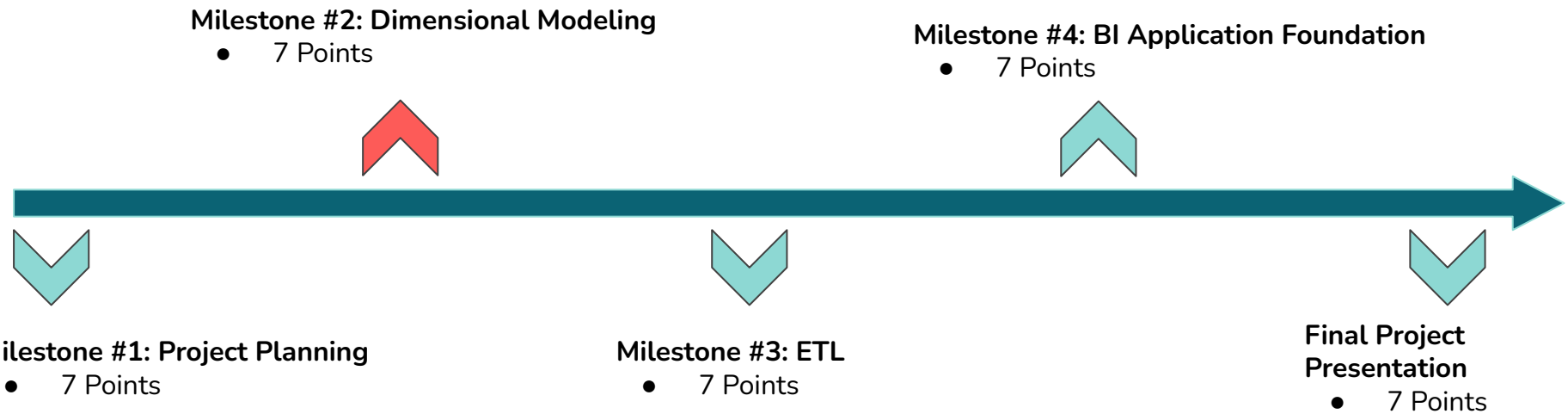
Milestone #3: ETL

- Either manually (with code) or with an application, physically get your data from Milestone #1 into the dimensional models you created in Milestone #2
- Document all steps along the way, and test that your Fact Tables and dimensions connect

Final Project Presentation

- 10 min max
- Use template to submit your Final Project

Project Grading, 35% of Final Grade



Week 6 Class Overview:

1. Technical Interview Practice #2
2. python API - Yelp
3. Introduction to Data Integration
4. ETL Workshop #1

Week 6 Class Overview:

1. **Technical Interview Practice #2**
2. python API - Yelp
3. Introduction to Data Integration
4. ETL Workshop #1

Today’s Technical Interview Practice

Question #1:

(2.5 minutes)

https://www.testdome.com/questions/sql/pets/68916

SQL Practice:

In the BigQuery public datasets there a table called `bigquery-public-data.fdic_banks.institutions` with the following fields:

- state_name (this is a string)
- institution_name (this is a string)
- active (this field is either true or false)
- address (this is a string)
- total_deposits (this is an integer)

Write a query to find the top 3 states (state_name) with the most active institutions.

SQL Practice:

In the BigQuery public datasets there a table called `bigquery-public-data.fdic_banks.institutions` with the following fields:

- state_name (this is a string)
- institution_name (this is a string)
- active (this field is either true or false)
- address (this is a string)
- total_deposits (this is an integer)

Write a query to find the number of states that have more than 1 Billion dollars in total_deposits.

Midterm Exam details

- Date: Wednesday, October 26th
- Time: 6:00pm ET
- Location: Online, on Blackboard
- Format: Short answer, multiple choice, matching
- Open book and notes: yes, you may use all class material
- Percent of Final Grade: 10%

Midterm Exam topics

- Definition and purpose of Data Warehouse
- Source Systems vs Target Systems
- Types of SQL Statements
  - Analyzing SQL statements
  - Will not be writing any SQL queries
- Kimball Lifecycle Project Planning and KPI’s
- Kimball Lifecycle BUS Matrix
- Dimensional Modeling
- Slowly Changing Dimensions
- ETL Definition

Week 6 Class Overview:

1. Technical Interview Practice #2
2. **python API - Yelp**
3. Introduction to Data Integration
4. ETL Workshop #1

Why look at the Yelp API?

- **API**: Application Programming Interface
- Data Acquisition from API's is very common in Data Warehousing
- Great to have experience with an API for interviews
- Many Project Teams in this course may leverage the Yelp (or other social APIs) for their Final Project

Where to start?

- To use an API you typically need authentication credentials
  - Oftentimes, a API Key is generated for each API each
  - Start here to make a Yelp API Key: https://www.yelp.com/developers/v3/manage_app
  - Then, checkout the documentation: https://www.yelp.com/developers/documentation/v3/business_search

Week 6 Class Overview:

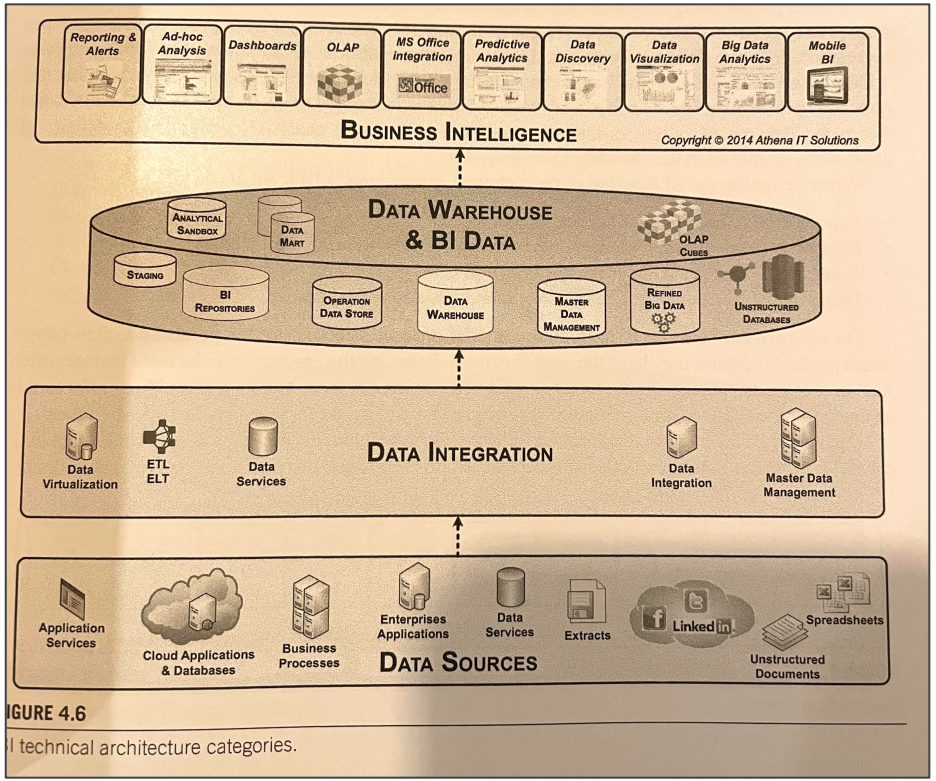
1. Technical Interview Practice #2
2. python API - Yelp
3. **Introduction to Data Integration**
4. ETL Workshop #1

There’s an entire field around Data Integration

https://dataanalytics.buffalostate.edu/job-titles-data-scientists

What is Data Integration (DI)

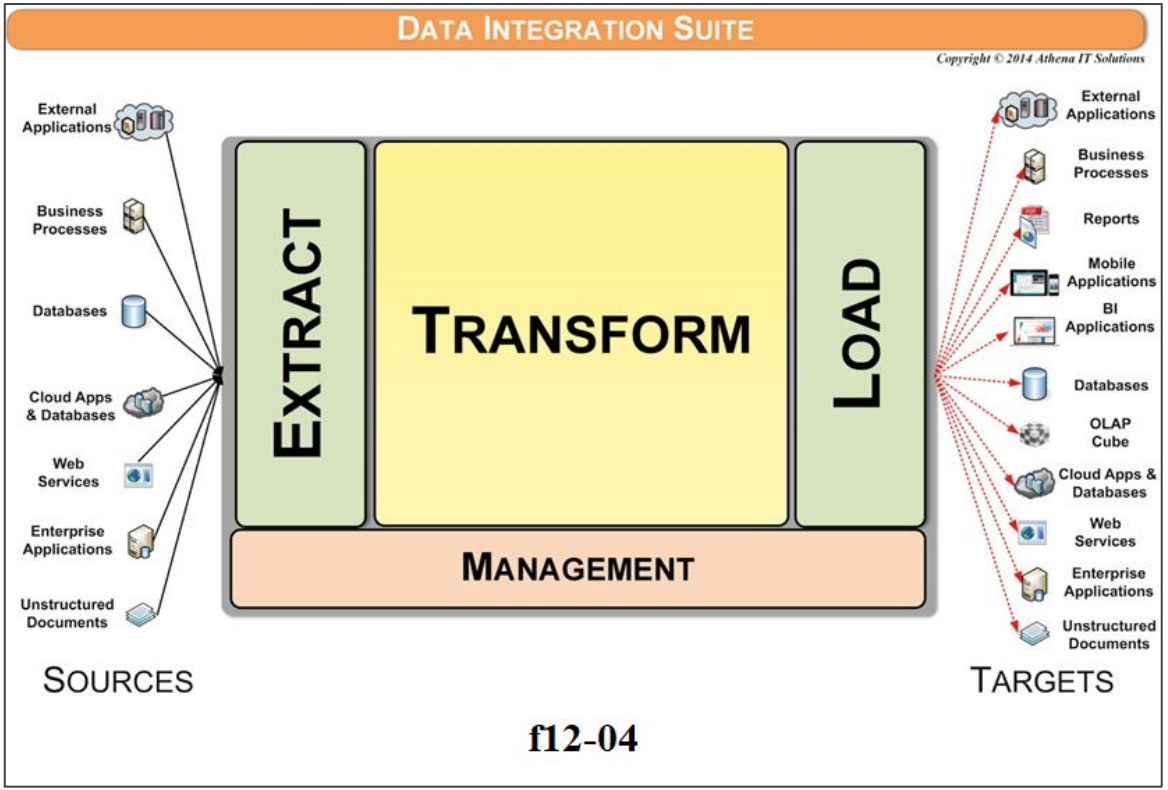
- **Data Integration:** combining data from different sources and brings it together to provide a unified view in Dimensional Models
  - This is the crux of Data Warehousing
  - DI is often described as the DW “iceberg”
  - Think about the word “integrate”



IGURE 4.6

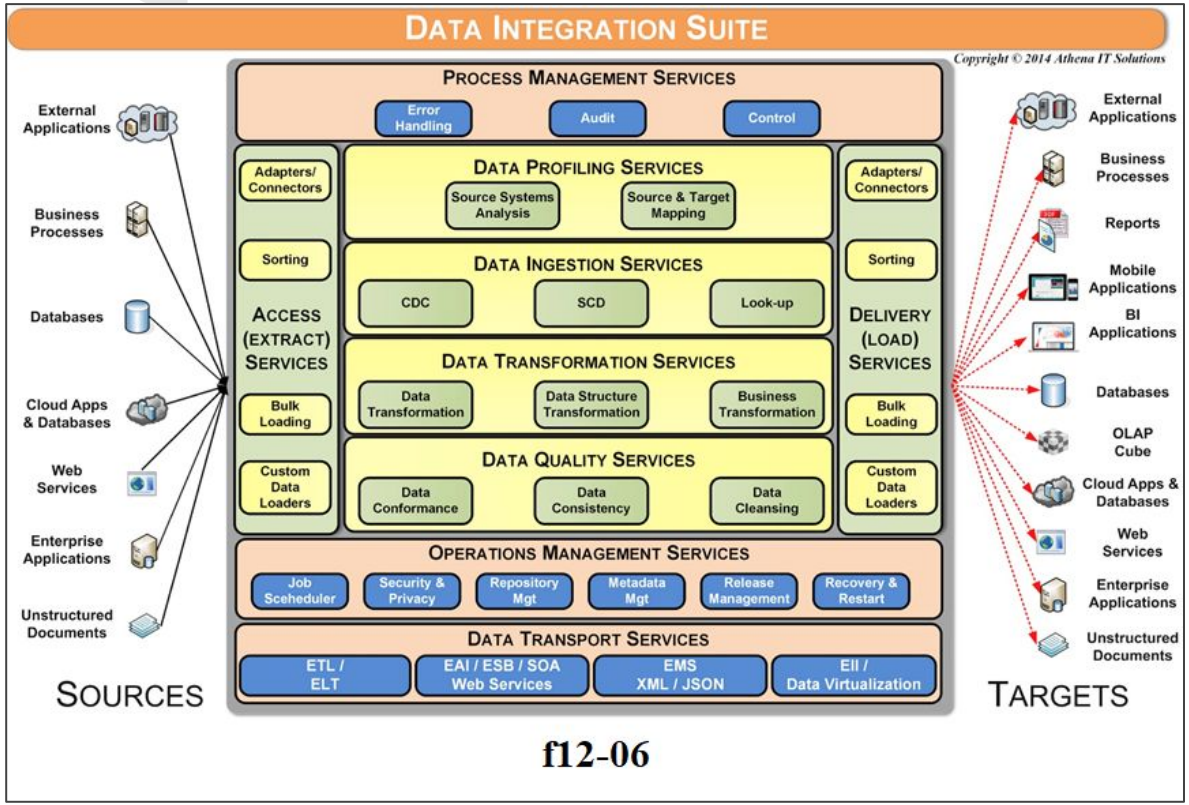
l technical architecture categories.

Overview of ETL



Goal: data will go from “source” systems to “target” systems

Details of ETL



f12-06

Today we will focus on:

- Data Profiling
- Data Transformation
- Data Quality

Data Integration Industry Tools

- Many **Vendors** in this space:
  - Oracle Warehouse Builder and Oracle Data Integrator (ELT)
  - Microsoft SQL Server Integration Services
  - Talend
  - Alteryx
  - IBM InfoSphere Information Server and DataStage
  - Informatica PowerCenter
  - Clover ETL
  - Pentaho (Enterprise and Community Editions)

- **Build your own**:
  - Many organizations elect instead to write their own ETL tools using scripting languages, custom software, cron jobs etc.

ETL Pipeline questions:

1. Should you build your own ETL Pipeline, or use a Vendor tool?
2. Is there a middle ground?
3. What do industry professionals use?

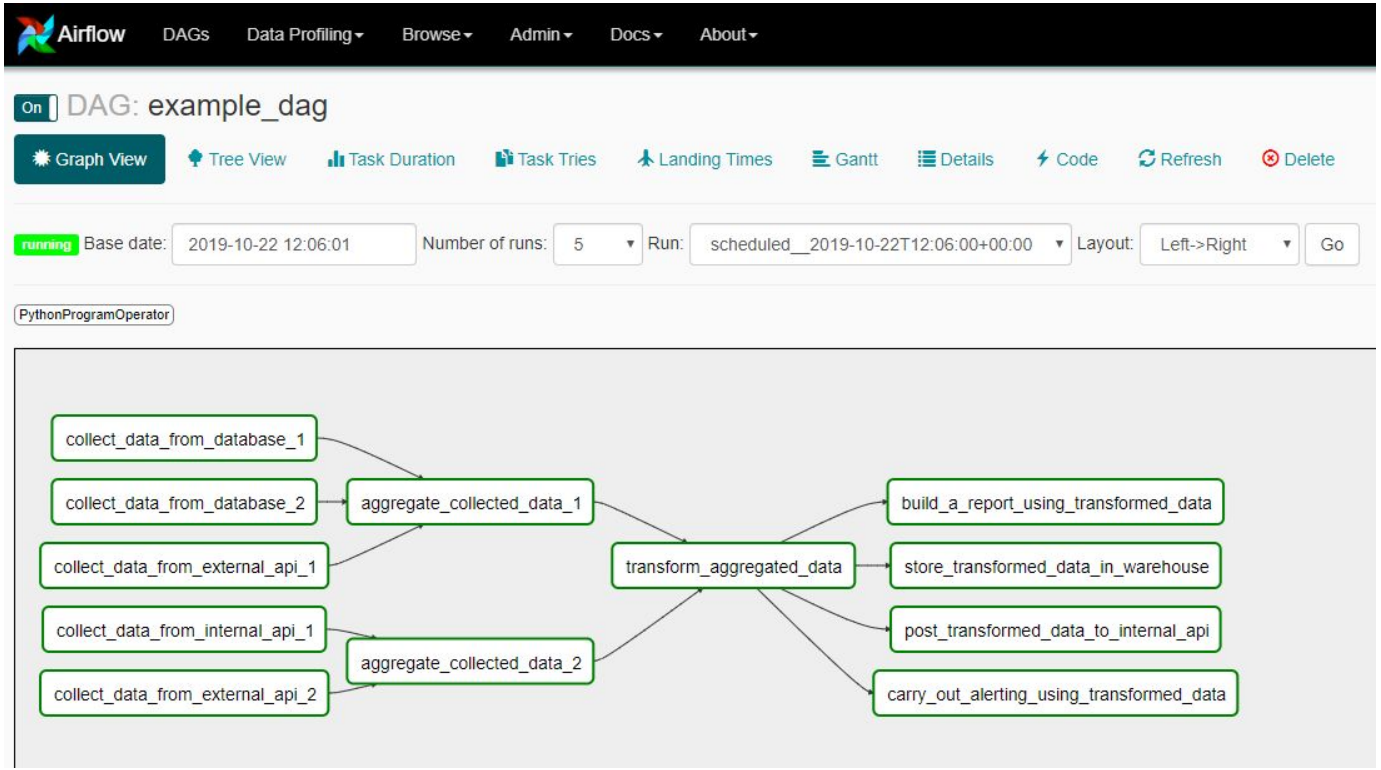
DI: Use a Vendor or Build Your Own?

- Advantages of using a Vendor tool:
  - Visual data flow and self-documentation
    - vs. a list of stored procedures, SQL, OS scripts [no flow diagrams!]
  - Structured design process
    - Tool is dedicated for the sole purpose of ETL and can guide you
  - Management features and resilience
    - Hand Coded systems may be fragile or may need to be extended when change happens
  - Data Lineage and Data Dependency functionality
  - High Performance (e.g., parallel processing)

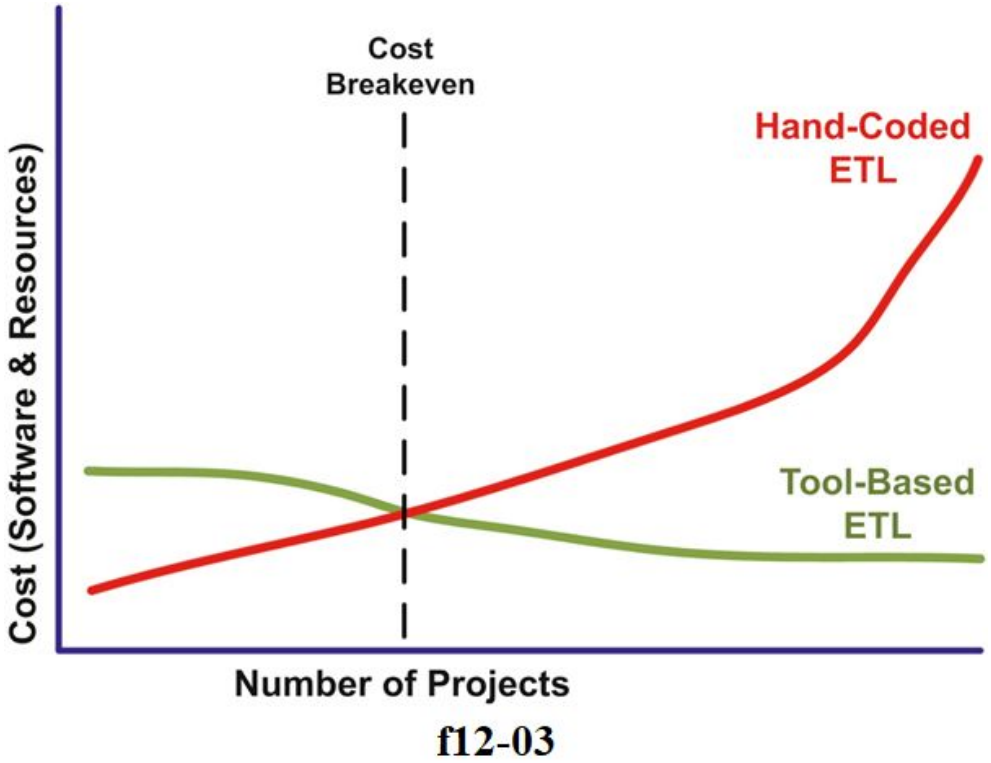
DI: Use a Tool or Build Your Own?

- Advantages of building your own ETL Pipeline:
  - Far cheaper, no licensing costs
  - Requires no Vendor tool training
    - Requires ETL Pipeline building skill sets (more common)
  - 100% customizable
    - Will support advanced transformations (no limitations)

DI: Middle Ground, example: Apache Airflow



ETL Cost Comparison



Should you Design an In-Memory or Disk-based Pipeline?

- Each stage of the ETL Pipeline has the potential to generate a new, intermediate set of data.
- A major consideration is whether or not to write these intermediate data sets to disk or simply pass them on (in-memory) to the next stage of the ETL pipeline.
- Advantages to in-memory ETL pipelines are primarily performance based. It simply takes less time to process if intermediate results are not written to disk.
- However a risk of in-memory pipelining is if and when failures occur. Then the ETL must be re-started at the most recent disk-based copy.

ETL vs ELT

ETL vs ELT, how to choose?

The method you use to load data depends on how much transformation is needed

Extract and Load

Extract, Load, Transform

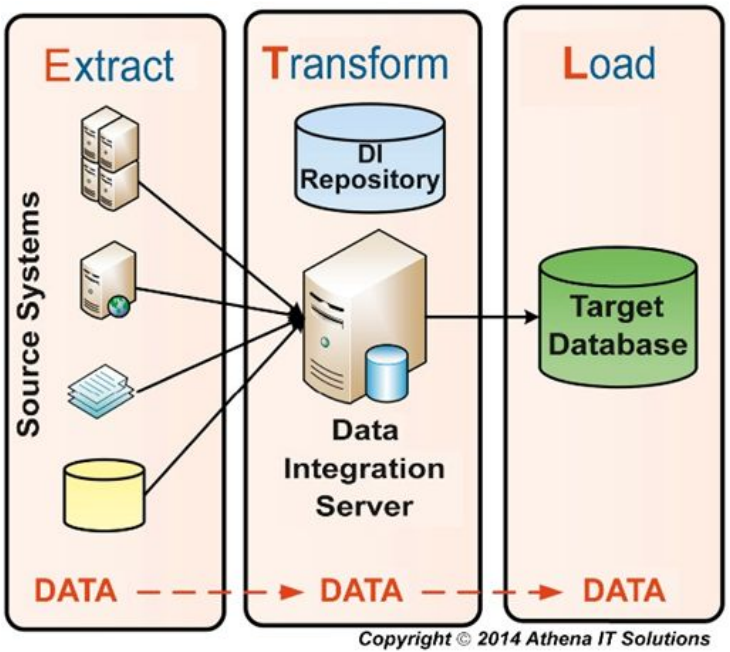
Extract, Transform, Load

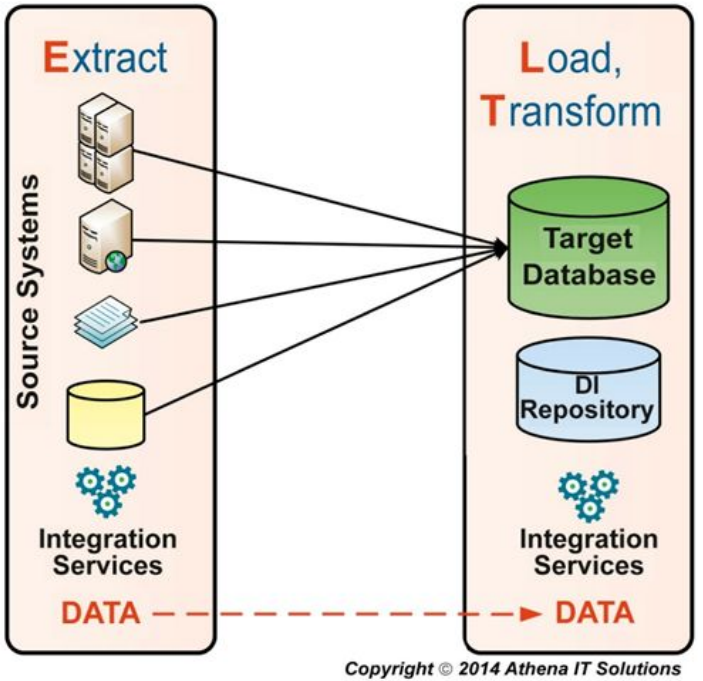
Google Cloud

ETL vs ELT, how to choose? (continued)

Type	When to use
EL	Batch load of historical data Scheduled periodic loads of log files (e.g. once a day) •But only if the data are already clean and correct!
ELT	Experimental datasets where you are not yet sure what kinds of transformations are needed to make the data useable. Any production dataset where the transformation can be expressed in SQL
ETL	Everything else, anything more complex

ETL vs ELT Architectures:

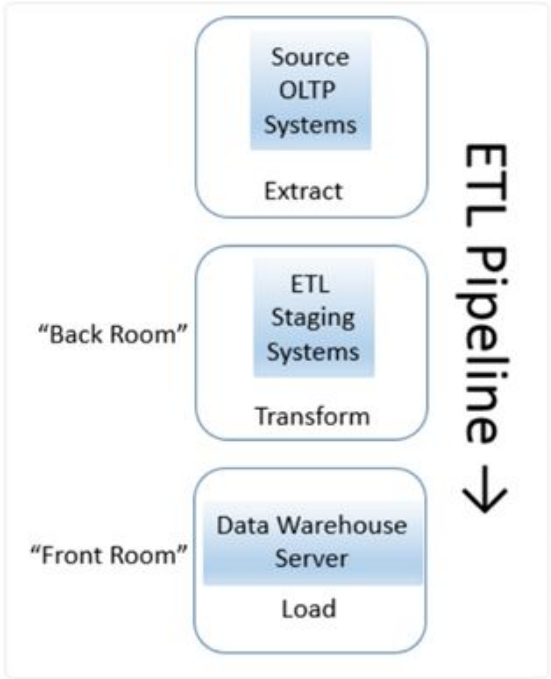


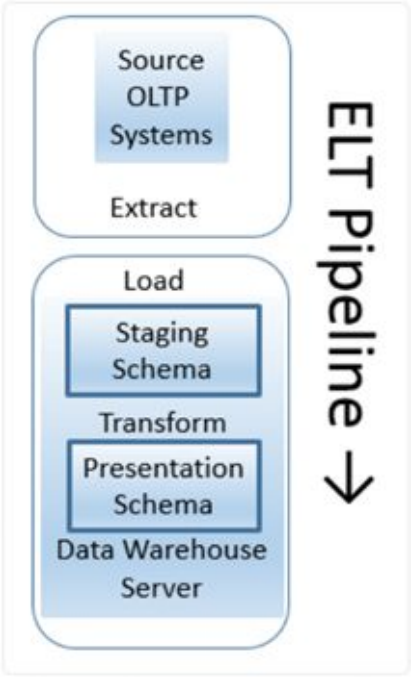


ETL vs ELT

- The traditional ETL process can be viewed as a rigid set of steps that may break when any change occurs in the source systems or target warehouse
- Advocates of ELT:
  - Load the data directly from the source systems into the data warehouse database, then:
  - operate on the loaded data to transform it into a data warehouse “presentation” schema.

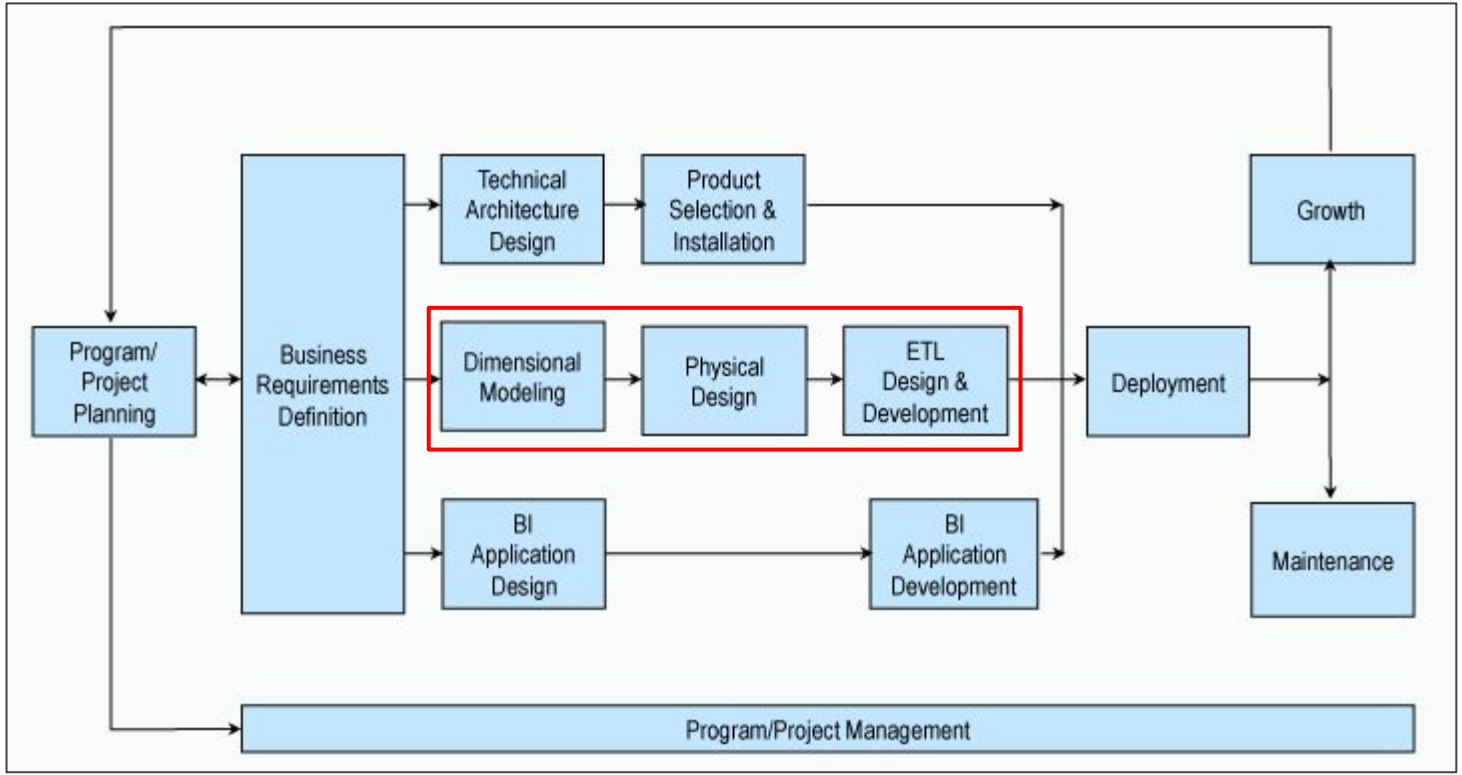
- Advocates of ETL:
  - Combining the Load and Transform step makes your master data vulnerable
  - You can do more validation with 3 steps! From Source to Back Room. Then from Back Room to Front Room.
  - Storage is cheap!





ETL in the Kimball Lifecycle

ETL in Kimball Lifecycle

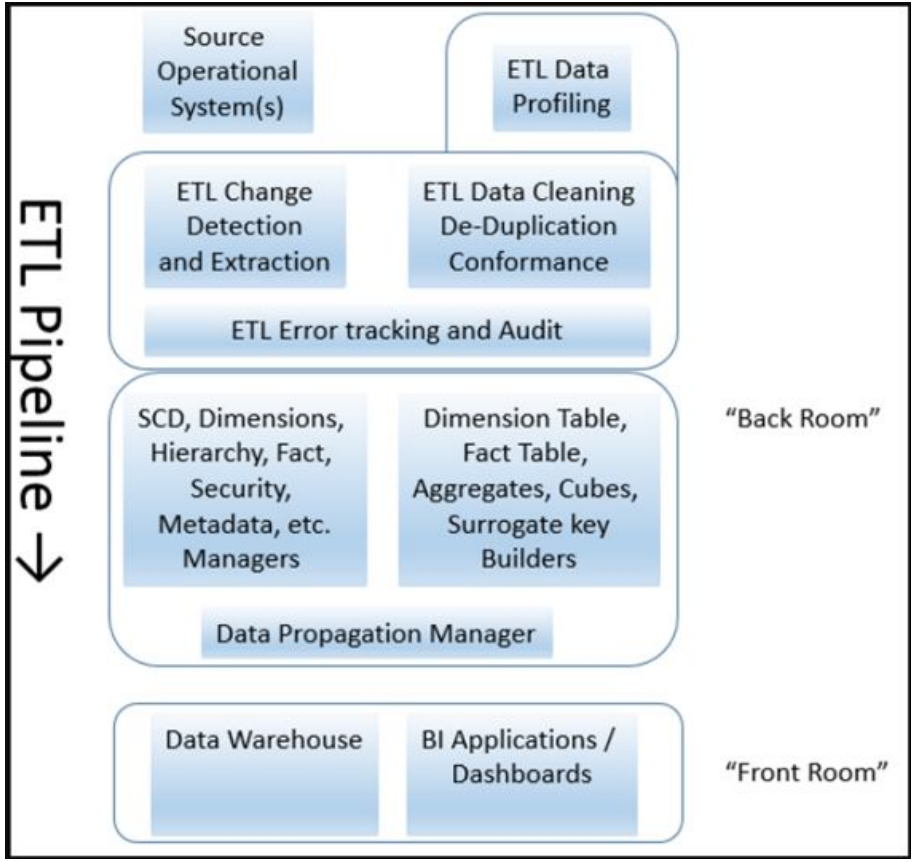


Dimensional Modeling is the first step in the “middle track” of the Kimball Lifecycle.

All tracks after the “Business Requirements Definition” can be done in parallel

DWTK, Chapter 17

More Detailed view of ETL



ETL will take the majority of time for a Data Warehousing project

Kimball ETL, 34 “Subsystems”

- Over many years and being exposed to thousands of systems, Kimball identified 34 ETL “Subsystems”
  - **1-3**: Extracting, getting data into the Data Warehouse
    - What is Data Profiling?
  - **4-8**: Cleaning and Conforming the data
    - Reusable cleaning tools
    - What if there’s an error?
  - **9-21**: Delivering the data into Dimensions and Facts
    - Surrogate Key Generators
    - Deliver Facts and Dimensions
  - **22-34**: Managing the ETL environment
    - What’s a cron job?

Week 6 Class Overview:

1. Technical Interview Practice #2
2. python API - Yelp
3. Introduction to Data Integration
4. **ETL Workshop #1**

ETL Workshop, before beginning

What is our ETL goal?

1. Extract data
2. Profile and Clean data
3. Turn data into Facts and Dimensions
4. Deliver data to our Data Warehouse (BigQuery)

Let’s look at an example of what our data will look like:

Spotify data example from last week:

https://drive.google.com/drive/folders/13Xs3tgT4gArntoYp7O5fq5K0kWf1iOpo?usp=sharing

ETL Workshop Details

Using a Jupyter Notebook, we will:

1. Connect to NYC Open Data with your API token
2. Pull the “311 Dataset” into python
3. Clean the Dataset
4. Data Profiling
5. Conform the Dataset (look at DWTK)
6. Deliver the Dataset into Dimensions
7. Create the Fact Table
8. Connect python to BigQuery
9. Put our entire script into a .py file
10. Run our ETL process and deliver the results to BigQuery

To practice ETL:

1. Open the 311 Dataset on NYC Open Data: link
2. Open up your NYC Open Data developer account: link
3. Open a new Jupyter Notebook

Homework:

1. Final Project Milestone #2 on Blackboard, due 10/15/22