

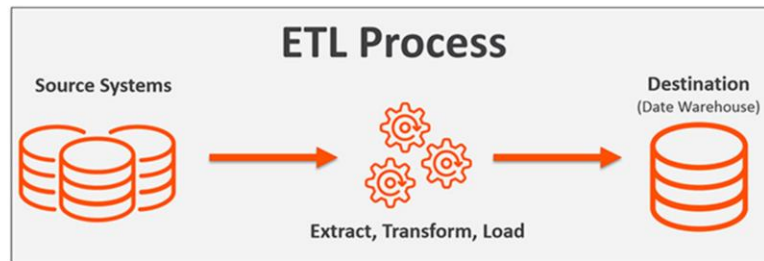
# Getting Data into the Data Warehouse

CS 537- Big Data Analytics

Dr. Faisal Kamiran

# ETL

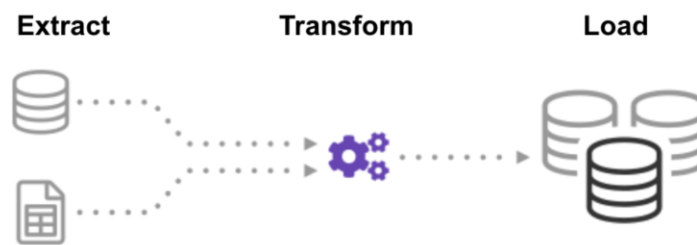
- Data transferred from source applications to the DWH or Data Mart
- Done with a process called ETL



Data sources include structured and unstructured data systems

# ETL

- Extract
- Transform
- Load



DR. FAISAL KAMIRAN

INFORMATION TECHNOLOGY UNIVERSITY

Extract: Collect data from data sources

Transform: Convert extracted data into a correct and common form

Load: Write data to the target Data Warehouse

## Extract

- Pull data from **multiple** source systems
- Traditionally done in “batches” (can be hourly, weekly etc.)
- Raw data is loaded including any existing errors
- Data transferred to a **staging area**



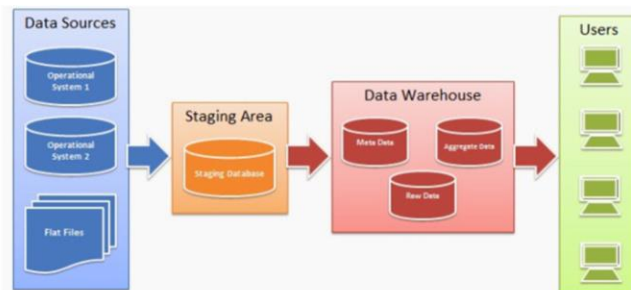
DR. FAISAL KAMIRAN

INFORMATION TECHNOLOGY UNIVERSITY

Batches: Extraction is not a continuous process. It is done at intervals  
Raw data is loaded including any errors (Error correction is done in the transform stage)

## Extract – Staging Area

- An intermediate storage area between the data sources and DWH
- The initial data is in different formats and may contain errors so it cannot be transferred directly to the DWH



DR. FAISAL KAMIRAN

INFORMATION TECHNOLOGY UNIVERSITY

The staging area acts as a buffer between the data warehouse and the source data.

Since data may be coming from multiple different sources, it's likely in various formats, and directly transferring the data to the warehouse may result in corrupted data. The staging area is used for transforming the data.

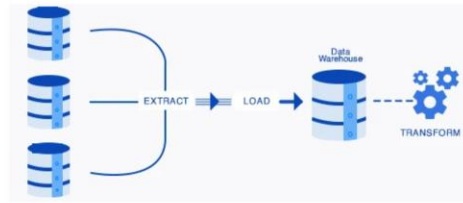
# Transform

- Convert data from multiple sources into a **uniform** format
- Performed on the extracted data in the staging area
- Transforms include
  - Cleaning
  - Filtering
  - Joining
  - Sorting
  - Splitting
  - Deduplication

# Load

- Final stage in the ETL process
- Involves transferring data into the DWH

# ELT



- ELT – Extract, Load, Transform
- Raw data stored in Hadoop HDFS, AWS S3 etc.
- No staging area
- Use big data environment computing power to **transform when needed**
- Used in cases where massive amounts of data need to be ingested quickly

With ELT, data is immediately available.



## ETL and ELT

- Data Warehouses work with relational SQL-like data structures
- Data must be transformed into a relational structure before it can be loaded into the Data Warehouse
- ETL used in Data Warehouses as **transformation must happen before loading**

[Online Analytical Processing \(OLAP\) data warehouses](#)—whether they are cloud-based or onsite—need to work with relational SQL-based data structures. Therefore, any data you load into your OLAP data warehouse must **transform** into a relational format *before* the data warehouse can ingest it. As a part of this data transformation process, data mapping may also be necessary to combine multiple data sources based on correlating information

## Variations of ETL

- Initial
- Incremental



## Initial Load ETL

- Done right before the Data Warehouse goes live
- Normally one time only
- Load all **relevant** data necessary for Analytics
- Redo if Data Warehouse corrupted

## Incremental ETL

- Incrementally “refreshes” the data warehouse
- New data: new employees, products, ...
- Modified data: employee promotions, product price change, ...
- Deleted data: employee resigns, customer unsubscribes
- Load only updated data instances

# Incremental ETL Patterns

## Append

- New data added at the end



# Incremental ETL Patterns

## In-place update

- Modify existing data (only some rows)



# Incremental ETL Patterns

## Complete replacement

- Overwrite existing data



Even if only a single row needs to be changed, entire data is modified

# Incremental ETL Patterns

## Rolling Append

- Maintain certain duration of history
- Wipe old data, when new data is appended



Like maintain only four weeks of data.  
The time window keeps rolling



## Incremental ETL Patterns

Modern data warehouses use

- ✓ Append
- ✓ In-place Update
- *Complete replacement*
- *Rolling Append*

Complete replacement and rolling append are not used in modern data warehouses. However, maybe found in very old DWHs.

# Data Transformation

## Goals

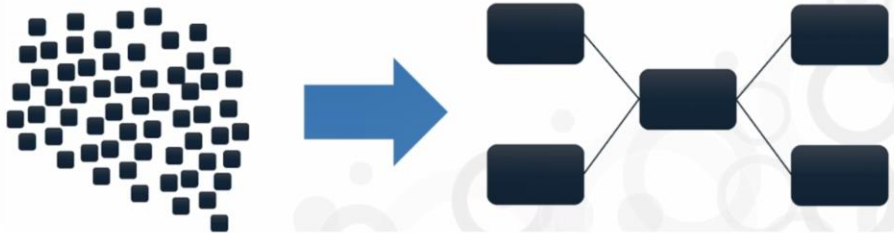
- Uniformity in data



# Data Transformation

## Goals

- Uniformity in data
- Restructuring



DR. FAISAL KAMIRAN

INFORMATION TECHNOLOGY UNIVERSITY

Restructure from raw form into a well-engineered data structure

## Data Transformation Models

- Data value unification
- Data type and size unification
- De-duplication
- Dropping columns (vertical slicing)
- Value-based row filtering (horizontal slicing)
- Correcting known errors

## Data value unification

- Merge data into a common format

Campus 1 New Faculty		
LastName	FirstName	Rank ...
Johnson	Susan	Professor
Wilson	Robert	Asst. Prof
Tolleson	Mary	Asst. Prof
Zimmerman	Todd	Professor
Marcus	Walter	Lecturer

Campus 2 New Faculty		
LastName	FirstName	Rank ...
Adleman	Robert	P
Bonvoy	Janice	AP
Clark	William	L
Douglas	Thomas	AP

Suppose that we have data from two different campuses, which use a different format for the Rank column

## Data value unification

- Choose one uniform format
- Transform other formats

LastName	FirstName	Rank	...
Johnson	Susan	P	
Wilson	Robert	AP	
Tolleson	Mary	AP	
Zimmerman	Todd	P	
Marcus	Walter	L	
Adleman	Robert	P	
Bonvoy	Janice	AP	
Clark	William	L	
Douglas	Thomas	AP	

The abbreviated format is our standard one, used in our dimension table in the DWH. So

## Data type and size unification

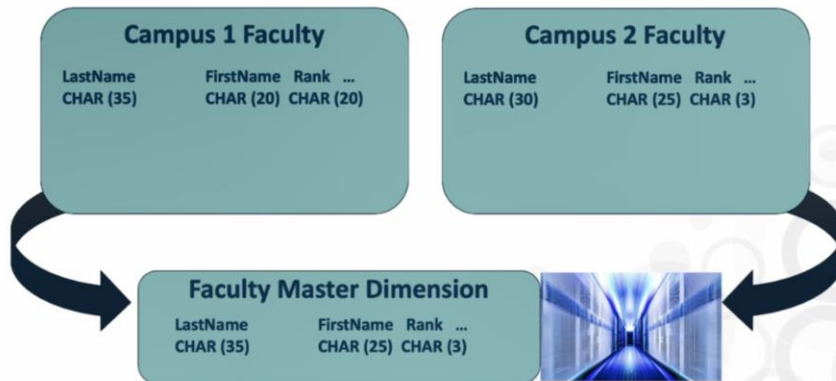
- Use one common set of data types and their sizes



Campus 1 and campus 2 used different data sizes for the columns in their source systems

## Data type and size unification

- Use one common set of data types and their sizes



DR. FAISAL KAMIRAN

INFORMATION TECHNOLOGY UNIVERSITY

Since previously, we chose to use Campus 2's abbreviated scheme, we will use their abbreviated data sizes in the dimension table



# De-duplication

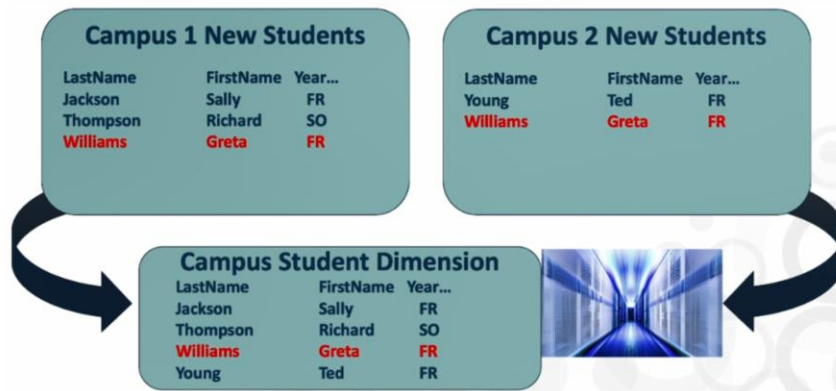
- Remove duplicate data

Greta Williams is  
taking classes on both  
campuses

Campus 1 New Students			Campus 2 New Students		
LastName	FirstName	Year...	LastName	FirstName	Year...
Jackson	Sally	FR	Young	Ted	FR
Thompson	Richard	SO	Williams	Greta	FR
Williams	Greta	FR			

Here, Greta Williams is taking classes on both campuses and needs to register for both campuses. She will have a record in source systems for both the campuses

# De-duplication




DR. FAISAL KAMIRAN

INFORMATION TECHNOLOGY UNIVERSITY

We need to detect the fact that Greta Williams appears in two different systems but is in fact a single student. This can be done through maybe the CNIC or any other natural key and then add de-duplicate it

## Dropping columns (Vertical slicing)

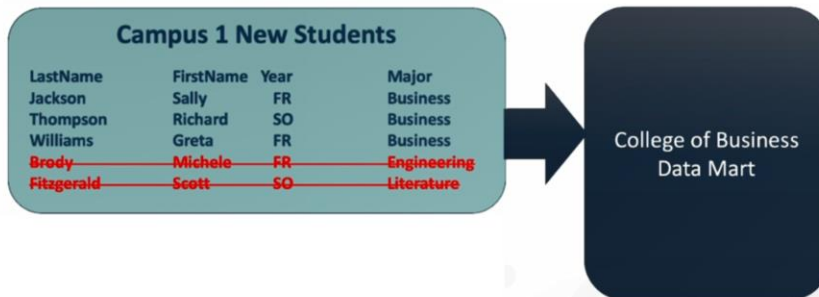


Campus 2 New Faculty				
LastName	FirstName	Rank	Column X	Column Y
Adleman	Robert	P	ABCDEF	XYZABC
Bonvoy	Janice	AP	RJTKWH	SLSHJS
Clark	William	L	QWERTY	ASDFGH
Douglas	Thomas	AP	ZXCVBN	CBNEUY

Data in Column X and Column Y  
not needed for analytic  
purposes

Slicing is done based on the columns

## Value-based row filtering (horizontal slicing)



Students with other majors will be filtered from the source data

Slicing is done based on certain columns in certain fields

Building a data mart containing information about business students. Only, students with business majors will be included. Other students will be filtered off

## Correcting known errors

- Fix errors in source data before loading



LastName	FirstName	Rank .....	Status
Adleman	Robert	P	F
Bonvoy	Janice	AP	X ← <i>Should be H</i>
Clark	William	L	A
Douglas	Thomas	AP	F

Status: Permissible Values  
F=Full-time | H=Half-time | A=Adjunct

## Correcting known Errors

Faculty Master Dimension			
LastName	FirstName	Rank .....	Status
Adleman	Robert	P	F
Bonvoy	Janice	AP	H
Clark	William	L	A
Douglas	Thomas	AP	F

DR. FAISAL KAMIRAN

INFORMATION TECHNOLOGY UNIVERSITY

Corrected data loaded in the DWH

## ETL best practices and guidelines

- Limit amount of incoming data to be processed



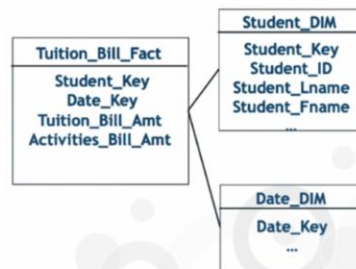
DR. FAISAL KAMIRAN

INFORMATION TECHNOLOGY UNIVERSITY

For incremental ETL, only ETL the data which is updated in the source systems

## ETL best practices and guidelines

- Limit amount of incoming data to be processed
- Process dimension tables before fact tables



If fact tables are processed first, then we might be trying to process a new student, whose entries are not present in the dimension table. Trying to do so will result in a foreign key error.



## ETL best practices and guidelines

- Limit amount of incoming data to be processed
- Process dimension tables before fact tables
- Opportunities for parallel processing



DR. FAISAL KAMIRAN

INFORMATION TECHNOLOGY UNIVERSITY

Preferable to incorporate parallel processing, like process dimension tables 1,2 and 3 in parallel then tables 4,5,6 and 7 and then the fact tables