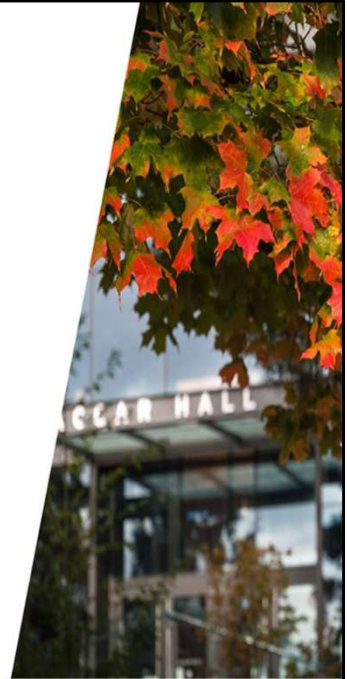


ETL AND TEXTUAL ANALYTICS

Data Analytics for Professional Accountants
(ACCTG 522)

Class 3 | MPAcc class of 2026



Welcome!

Class Agenda

- Review and ETL Overview
- ETL Cases in the FSB Remote Labs
- Conclusion and look ahead

Review: ETL Overview

Results - Input Data (1) - Output

3 of 3 Fields | Cell Viewer | 597 records displayed

Record	EmployeeCode	FirstName	LastName
14	Canada342-9W	Sean	Morgan
15	US476-10W	Julie	Johnson
16	MEX93-5M	Samuel	Perry
17	ENG442-9Mo	Jarrold	Beckett
18	Canada17-7W	Lillian	Nelson
19	ARG559-10W	Delfina	Brandon
20	ENG389-10W	Shaun	Jacobs
21	MEX121-11W	Theresa	Barnes
22	ENG489-12W	Stephen	Powell
23	AUS88-10W	Queenie	Thibodeau
24	AUS455-3W	Juan	King
25	AUS491-13M	Laureen	Varney
26	ENG193-10Mo	Anthony	Griffin
27	ARG508-14W	Olga	Rowe

5 new output fields again

Results - Output Data (4) - Input

8 of 8 Fields | Cell Viewer | 597 records displayed

Record	EmployeeCode	FirstName	LastName	Location	EmplID	PlantID	PayPeriod	LocationPlantID
14	Canada342-9W	Sean	Morgan	CAN	342	9	W	CAN-9
15	US476-10W	Julie	Johnson	USA	476	10	W	USA-10
16	MEX93-5M	Samuel	Perry	MEX	93	5	M	MEX-5
17	ENG442-9Mo	Jarrold	Beckett	ENG	442	9	M	ENG-9
18	Canada17-7W	Lillian	Nelson	CAN	17	7	W	CAN-7
19	ARG559-10W	Delfina	Brandon	ARG	559	10	W	ARG-10
20	ENG389-10W	Shaun	Jacobs	ENG	389	10	W	ENG-10
21	MEX121-11W	Theresa	Barnes	MEX	121	11	W	MEX-11
22	ENG489-12W	Stephen	Powell	ENG	489	12	W	ENG-12
23	AUS88-10W	Queenie	Thibodeau	AUS	88	10	W	AUS-10
24	AUS455-3W	Juan	King	AUS	455	3	W	AUS-3
25	AUS491-13M	Laureen	Varney	AUS	491	13	M	AUS-13
26	ENG193-10Mo	Anthony	Griffin	ENG	193	10	M	ENG-10
27	ARG508-14W	Olga	Rowe	ARG	508	14	W	ARG-14

Extension

Using RegEx101 tool attempt to Transform these similar EmployeeCodes:

Canada1234-1Mo
Mexico0001234-121Monthly
Australia45789-13Annual
AUS123-13M
FRA334455-145weekly
GRE67 1a

Take aways

Regular expressions:

[A-Za-z] lists of acceptable terms (alphabetic)

[A-Za-z0-9] as above, alphanumeric

*,+ zero, one or more

^,\$ beginning, end of the line

| or

\d any digit, \d{N} a specific "N" number of digits

. Any character, \. A full stop (escaping)

Extra RegEx examples

Dates:

- Invoice Date: 03/15/2025
- Payment Due: 04/14/2025

Invoices:

- INV-100234
- INV-567890
- Reference: INV-456

Extra RegEx examples

Firm Identifiers:

- CIK: 0000320193
- Ticker: TSLA
- Ticker: AAPL

Matching dollar amounts and accounts:

- Total: \$12,450.00
- Late Fee: \$25
- Refund: (\$1,200)

Extra RegEx examples

Phone numbers, email, credit card numbers:

- 206-123-4567
- real_person@example.com
- Potential fraud detected: 4532-9876-1234-5678

ETL Cases in the FSB

Remote Labs

ETL Cases, Continued

Today we will examine “Joining Data”:

- Joining data is another **fundamental step** in the ETL process, as in most cases, developing insight will require data from different databases.
- Joining data can sometimes be referred to as merging data, blending data, or combining data.

Why is joining data important?

Joining data is part of the transforming step in ETL.

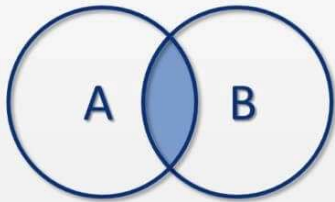
- Incorrect joins are a **data quality issue**, incorrect joins can:
 - **crash** (or really slow down) our programs!
 - have us **lose** important data.
 - Have us generate **incorrect conclusions/insight** from our analysis.

Joining using unique keys

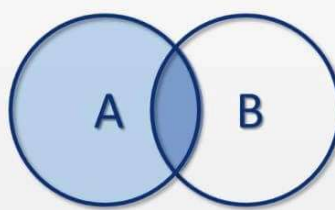
In all cases we will use unique keys to join the data.

- All the unique keys are described in the case
- One of them was tricky, can you remember which one?

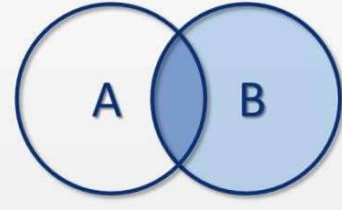
INNER JOIN



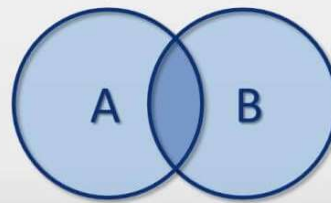
OUTER JOIN



LEFT

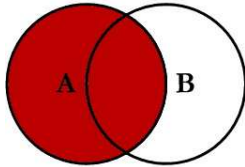


RIGHT

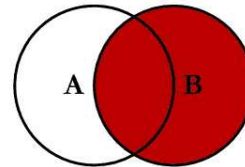


FULL

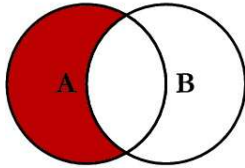
SQL JOINS



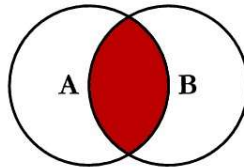
```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
```



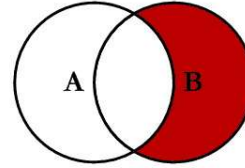
```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
```



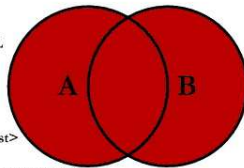
```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
WHERE B.Key IS NULL
```



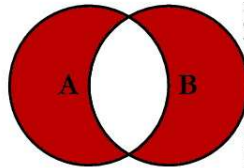
```
SELECT <select_list>
FROM TableA A
INNER JOIN TableB B
ON A.Key = B.Key
```



```
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.J. Morritt, 2008

Joining using unique keys

In all joins we will use “one to many” matches.

- Joining the large information source (“JELinItems”) to the smaller set of information in the other data sheets.

Why join across multiple software?

Because joining data is so fundamental, we want to work towards becoming software agnostic.

Working with cutting-edge software allows you to:

- Become adaptable and resilient to **new software**.
- Impress recruiters with practical discussions of cutting-edge software.

Joining across different tools:

Some terminology:

- Alteryx: drag and drop “Join” and “Union” tools
- PowerBI: data relationships tab
- Python: code-level with package “pandas”
- SQL: base code-level
- Tableau: Both relationships and joins.

Alteryx Joins:

Join (1) - Configuration

☐ Join by Record Position

☒ Join by Specific Fields

Left Right

Options TIP: To reorder multiple rows: select, right-click and drag.

	Input	Field	Type	Size	Rename	Description
☒		*Unknown	Unknown	0		Dynamic or Unknown Fields



SQL/Python:

```
SELECT
  JEL.*,
  BU.*
FROM JELineitems AS JEL
LEFT JOIN BusinessUnits AS BU
  ON JEL.BusinessUnitID = BU.BusinessUnitID;
```

```
SELECT
  JEL.*,
  BU.*
FROM JELineitems AS JEL
RIGHT JOIN BusinessUnits AS BU
  ON JEL.BusinessUnitID = BU.BusinessUnitID;
```

```
import pandas as pd
```

```
# Example: read from Excel file (each sheet as
DataFrame)
file_path = "your_file.xlsx"
```

```
# Load the two sheets
JELineitems = pd.read_excel(file_path,
sheet_name="JELineitems")
BusinessUnits = pd.read_excel(file_path,
sheet_name="BusinessUnits")
```

```
# Perform SQL-style join on BusinessUnitID
# Equivalent to: SELECT * FROM JELineitems
#             INNER JOIN BusinessUnits
#             ON JELineitems.BusinessUnitID =
BusinessUnits.BusinessUnitID;
merged_df = JELineitems.merge(
  BusinessUnits,
  on="BusinessUnitID",
  how="inner" # can also use 'left', 'right', or 'outer'
)
```

Lab Agenda

1. Open files in excel / open case and identify the keys
2. Compare Tableau and Power BI

Tableau Relationships

⊞ JELinItems+ (Analytics_mindset_case_studies_ET...

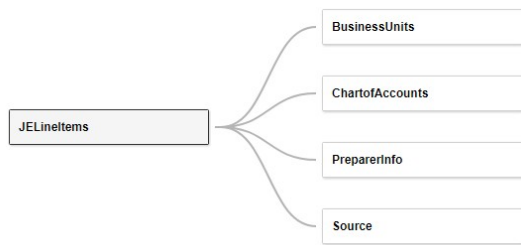
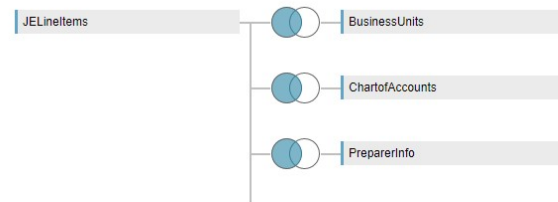
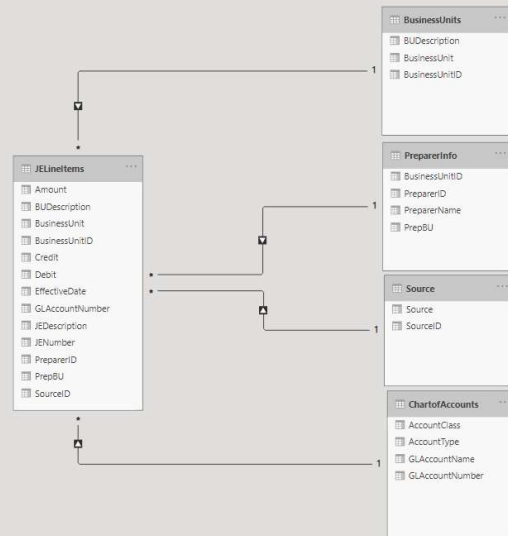


Tableau Joins

JELineItems is made of 5 tables. ⓘ

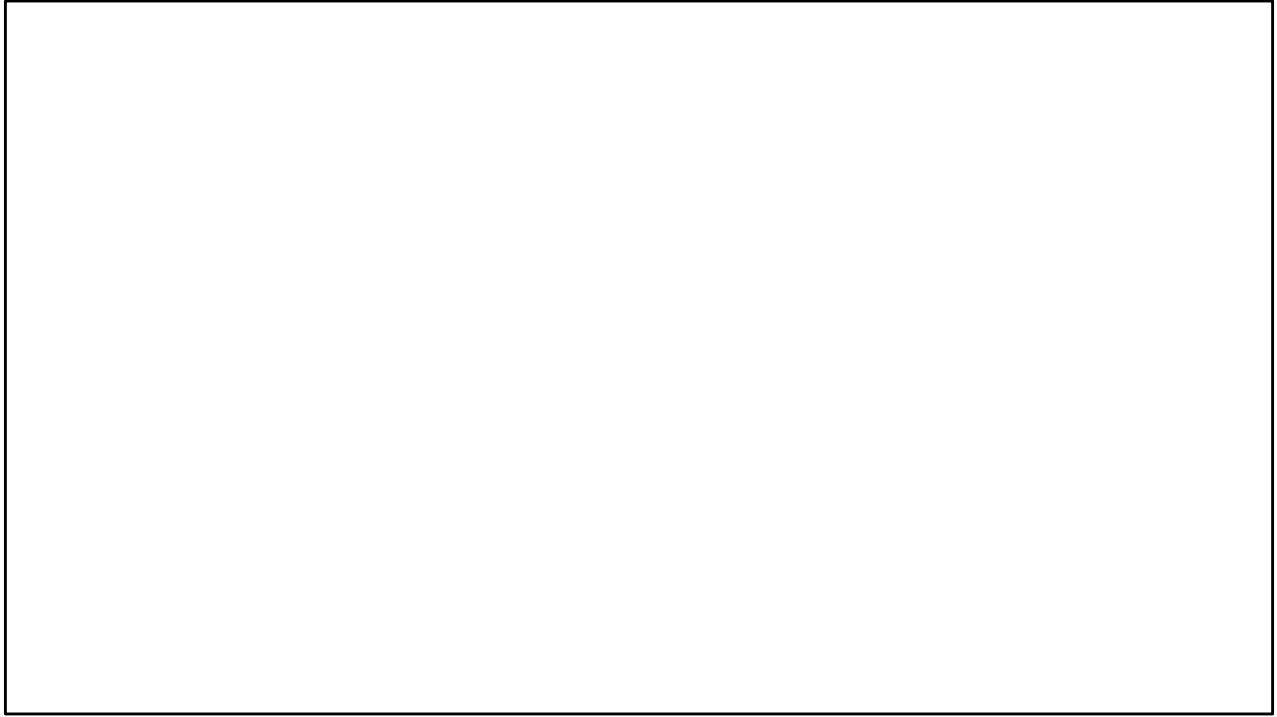


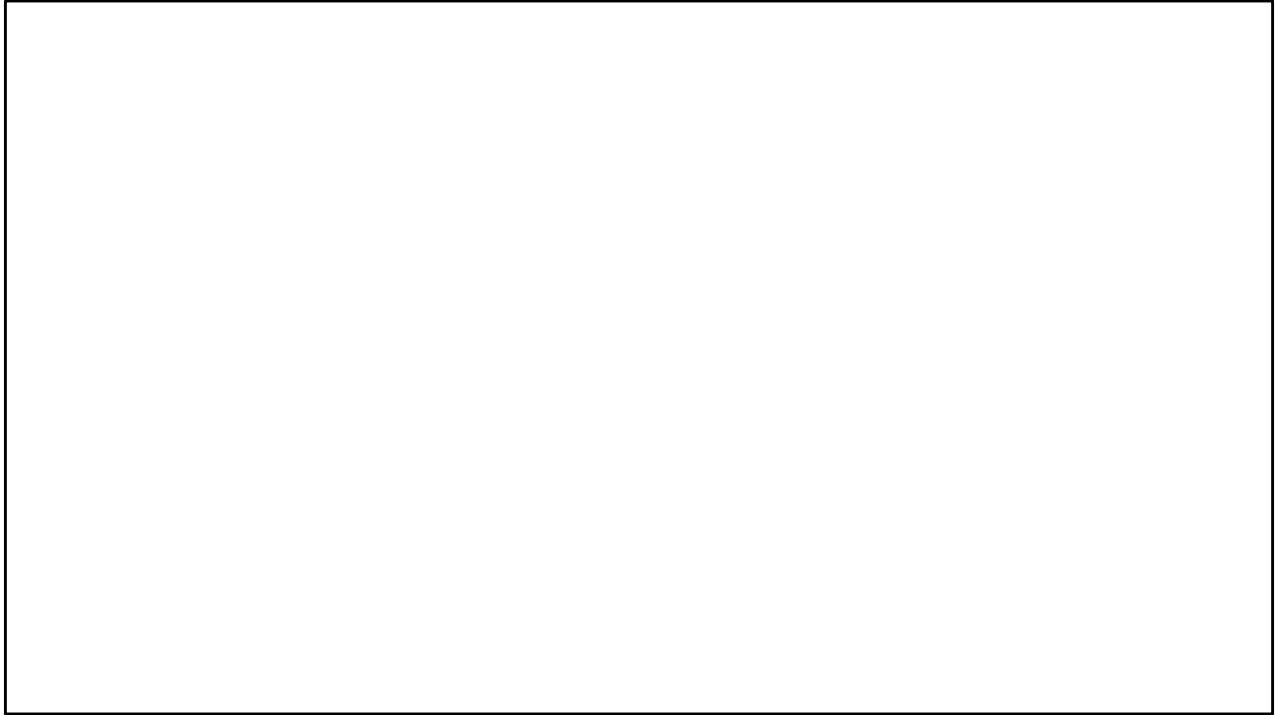
Power BI Data Relationships



To FSB remote labs

Again I recommend **saving screenshots** of your work in Alteryx (I've added blank slides if you want to save them in this ppt document, ctrl+m adds more slides)





Conclusion and look-ahead

ETL is the first step in the analytical process – we cannot undertake data analytics without data!

ETL is especially important when the data needs to be cleaned or transformed for use in analytics.

Conclusion and look-ahead

Wednesday:

- Cases: Enron Emails Case – RegEx Heavy

Thank you and Congratulations!

