Course Intro

Elements of Data Integration (CS 329E)

Second Edition

Jan 17, 2025

What is an entity?

In a DW, every table is called an entity.

Most entities represent real-world things. Examples:

Places

Movies

Products

They represent distinct, identifiable concepts in the world.

Each entity is represented with a schema.

Each entity has a unique identifier called a Primary Key (abbreviated to PK).

Entities do not contain other entities.

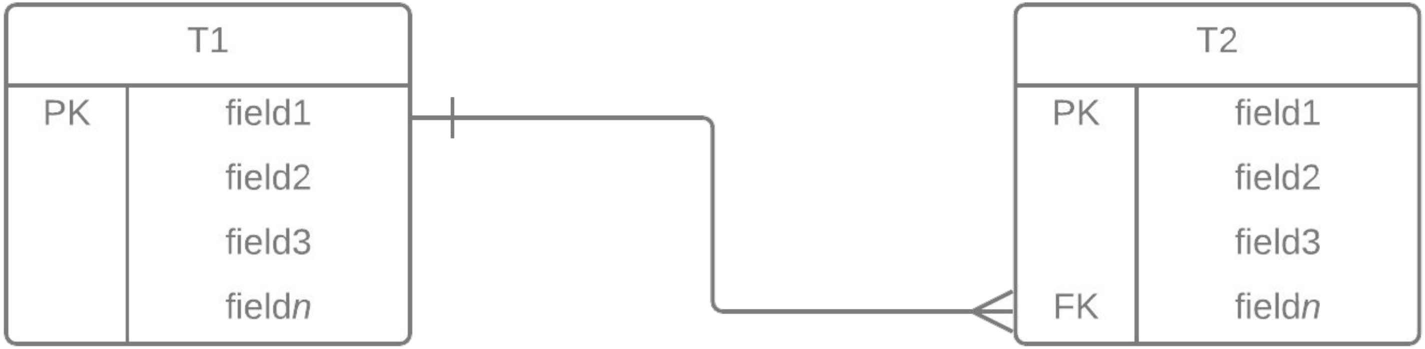
Entities link to other entities. These links are called relationships.

There are three relationship types: one-to-one, one-many, and many-to-many.

One-to-one and one-to-many relationships are represented with Foreign Keys (abbreviated to FK).

Many-to-many relationships are represented as their own tables called Junction Tables.

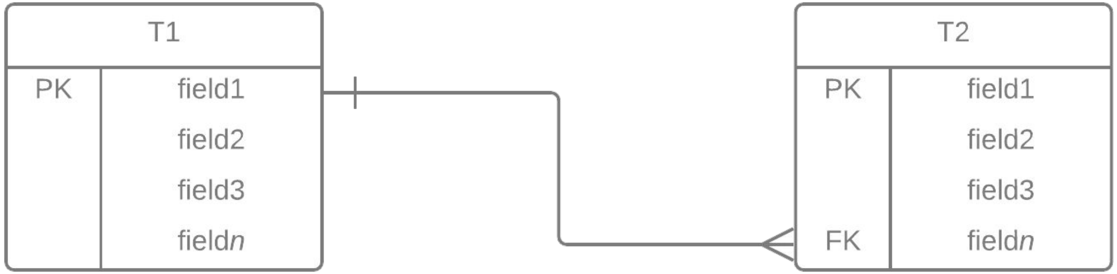
Table Relationship: One-to-Many (1:*m*)



T1	
PK	field1
	field2
	field3
	field*n*

T2	
PK	field1
	field2
	field3
FK	field*n*

Table Relationship: One-to-Many (1:*m*)



T1	
PK	field1
	field2
	field3
	field*n*

T2	
PK	field1
	field2
	field3
FK	field*n*

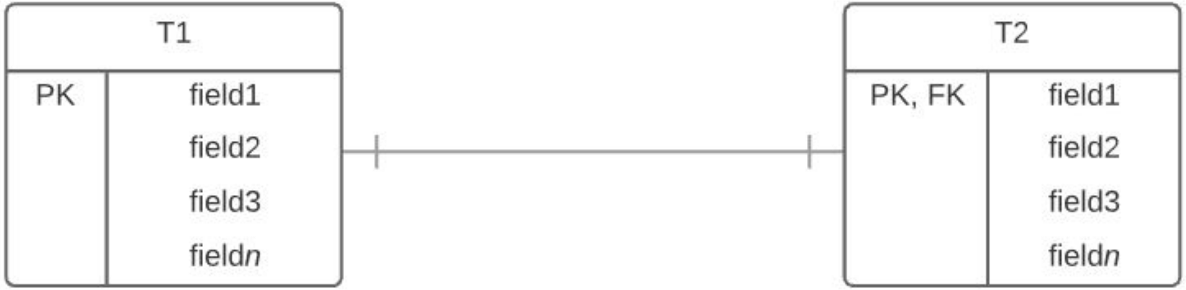
Author

id	name	section
1	Mary Tuma	news
2	Michael King	arts
3	Nina Hernandez	news
4	Sunil Kumar	music

Article

id	title	date	*authid*
1	Turmoil at the Zoo	2019-01-26	1
2	CodeNEXT's New Friend	2019-01-27	1
3	Quote of the Week	2019-01-27	3
4	SXSW News	2019-01-28	2
5	More from Steve Adler	2019-01-28	1

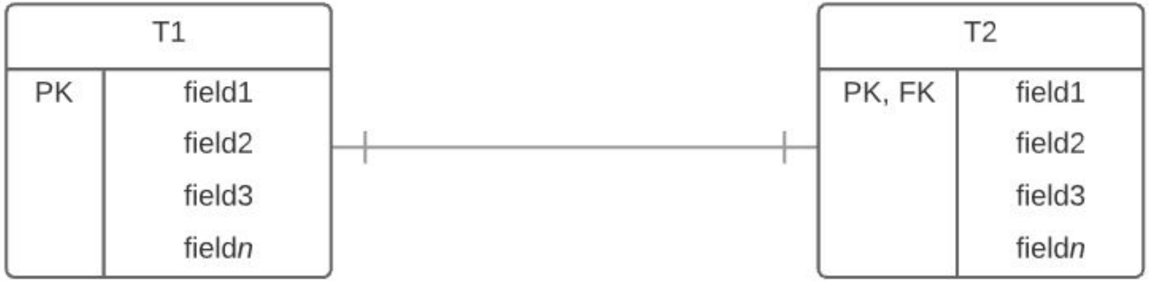
Table Relationship: One-to-One (1:1)



T1	
PK	field1
	field2
	field3
	field*n*

T2	
PK, FK	field1
	field2
	field3
	field*n*

Table Relationship: One-to-One (1:1)



T1

PK

field1

field2

field3

field*n*

T2

PK, FK

field1

field2

field3

field*n*

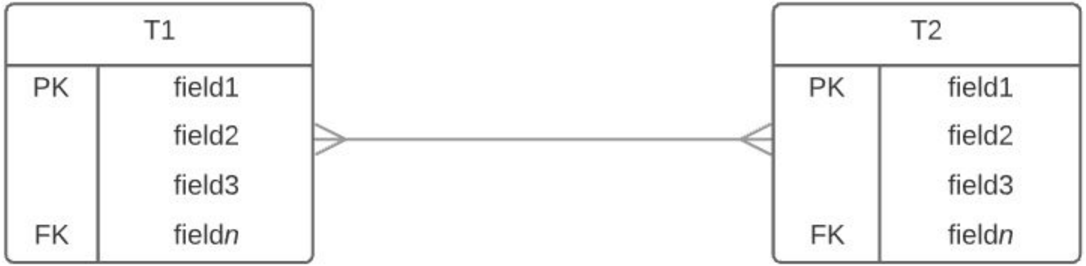
Article

id	title	date	*authid*
1	Turmoil at the Zoo	2019-01-26	1
2	CodeNEXT's New Friend	2019-01-27	1
3	Quote of the Week	2019-01-27	3
4	SXSW News	2019-01-28	2
5	More from Steve Adler	2019-01-28	1

Article_Stats

id	clicks	likes	dislikes	comments
1	120	45	9	13
2	0	0	0	0
3	8	0	0	2
4	30	4	0	1
5	9	1	3	3

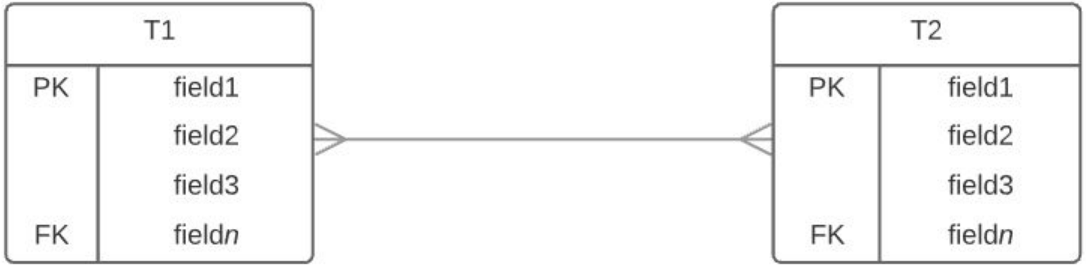
Table Relationship: Many-to-Many (*m*:*n*)



T1	
PK	field1
	field2
	field3
FK	field*n*

T2	
PK	field1
	field2
	field3
FK	field*n*

Table Relationship: Many-to-Many (m:n)



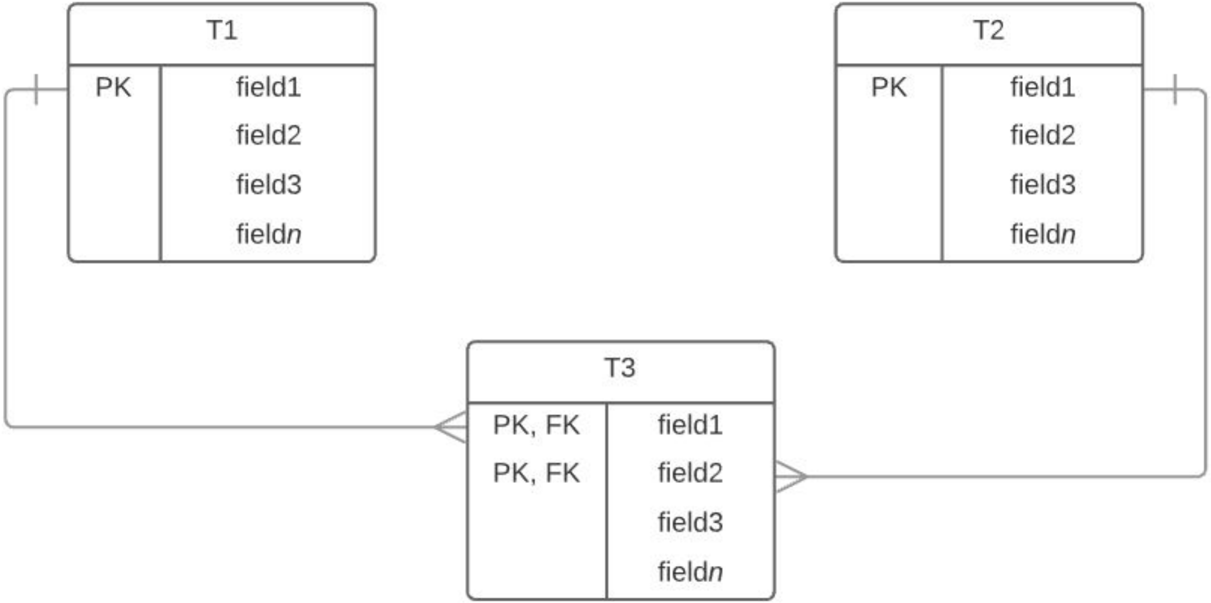
Tag

id	tag	*aids*
1	Politics	1, 2, 5
2	Austin	1, 2, 3, 4, 5
3	Mayor	3, 5
4	Business	1, 2, 5
6	Land Development	2
37	Animals	1

Article

id	title	date	*authid*	*tids*
1	Turmoil at the Zoo	2019-01-26	1	4, 37, 2
2	CodeNEXT's New Friend	2019-01-27	1	2, 6
3	Quote of the Week	2019-01-27	3	2, 3
4	SXSW News	2019-01-28	2	2, 40, 7
5	More from Steve Adler	2019-01-28	1	2, 3, 4

Representation of Many-to-Many (*m*:*n*)



T1	
PK	field1
	field2
	field3
	field*n*

T2	
PK	field1
	field2
	field3
	field*n*

T3	
PK, FK	field1
PK, FK	field2
	field3
	field*n*

Table Relationship: Many-to-Many (*m*:*n*)

Tag

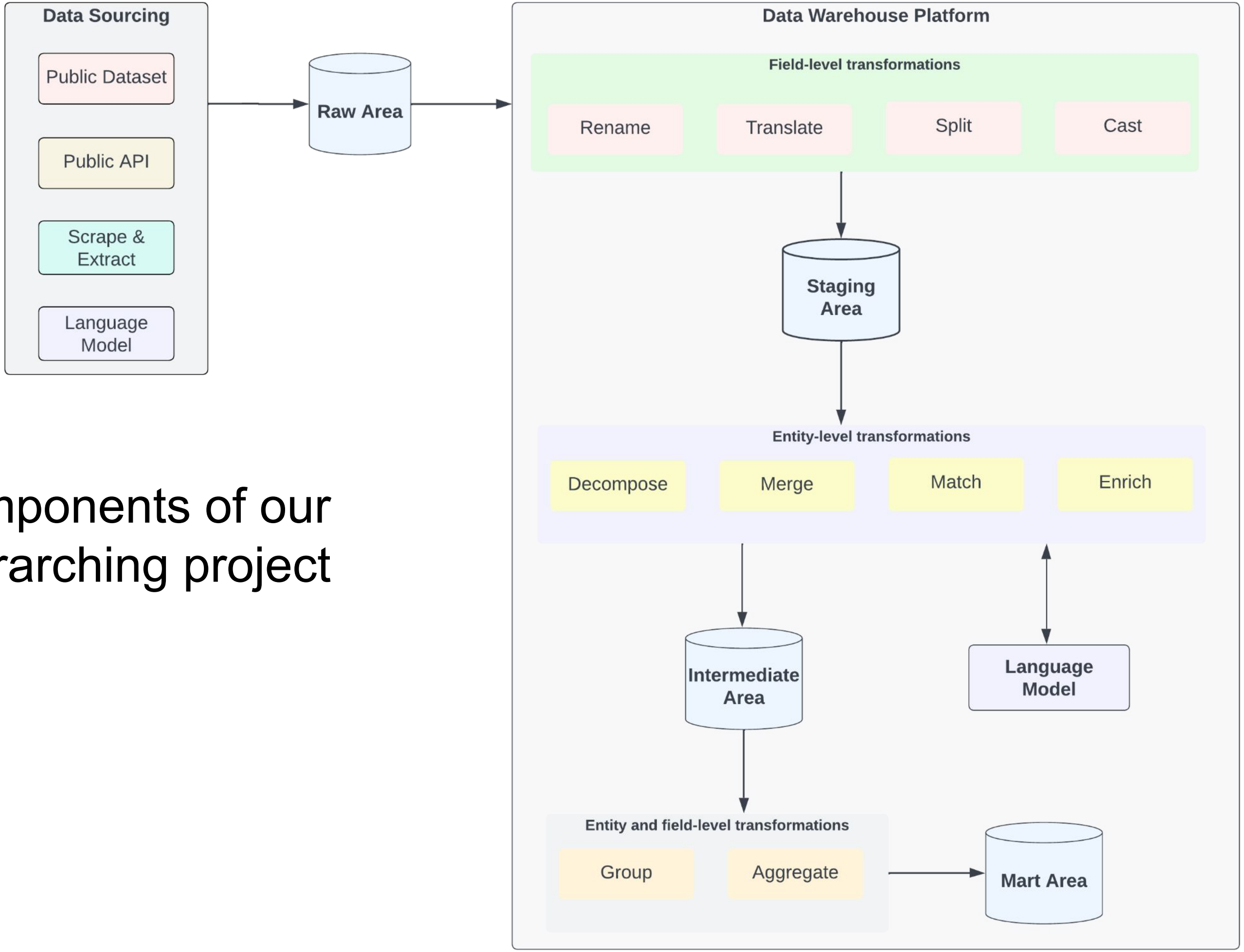
id	tag
1	Politics
2	Austin
3	Mayor
4	Business
6	Land Development
37	Animals

Article

id	title	date	*authid*
1	Turmoil at the Zoo	2019-01-26	1
2	CodeNEXT's New Friend	2019-01-27	1
3	Quote of the Week	2019-01-27	3
4	SXSW News	2019-01-28	2
5	More from Steve Adler	2019-01-28	1

Tagged_Article

tid	*aid*
4	1
37	1
2	1
2	2
6	2



Components of our overarching project

Scoping our overarching project

You will end up with a logical data model that unifies independent sources of data for a particular subject area.

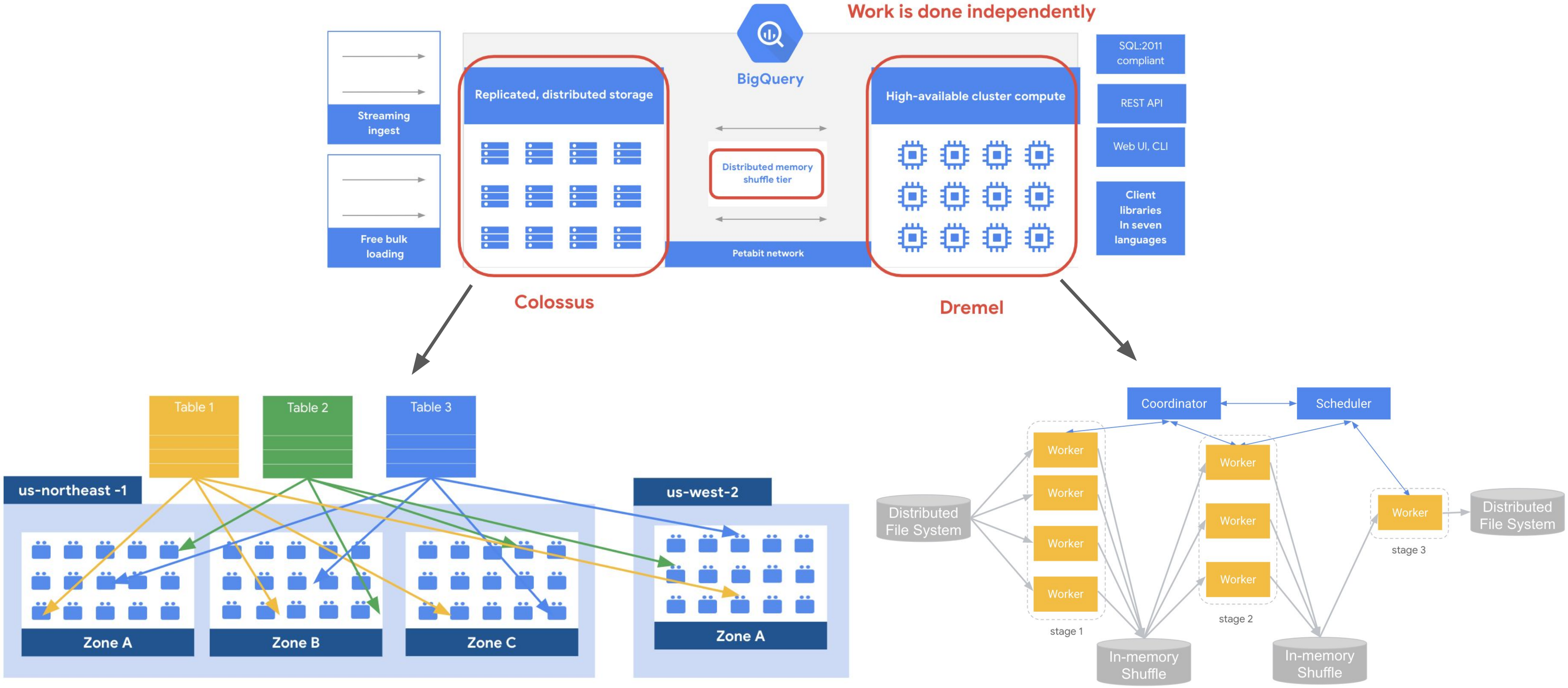
You will end up with a data pipeline that populates your data model.

Our implementation will use Google Cloud products (BigQuery, Google Cloud Storage, Gemini, Vertex AI).

Although we will be using the language model for extracting structured data and for enriching data, we probably won't have enough time to explore different grounding techniques.

We also won't have the time to build data visualizations, dashboards or reports on top of our warehouse.

BigQuery under the hood



How this class will work

- A major portion of our class periods will be spent on project work.
- You will work collaborative with your partner on each project.
- You and your partner should block off time outside of class to finish off any remaining aspects of the project.
- If you have to miss class, you should email me, the TA, and your partner in advance.
- If you are having some collaboration challenges with your partner, please email me and/or come to office hours.
- Assignments will come with code samples to help you get started with your project.
- Assignments won’t come with a complete playbook, you are expected to figure some things out on your own.
- Once you get your grade back, you will have the option to resubmit your project if you lost points as long as have submitted your work on time and completed the requirements, by the assignment deadline.
- You will have 7 days to resubmit from the time you get your grade back.
- The instructors will be doing regular reviews of your in-progress work during class to reduce the possibility of rework.
- Our week-by-week schedule is tentative. Timelines may be adjusted based on the needs of the class.
- Please use Ed for most outside class communication. However, be careful not to overshare your work.
- As this is a newish course, please share your feedback with me early and often via email and/or office hours.