

Project 2

Foundations of Data Warehousing (CS 378)
First Edition

September 13, 2024

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.

What is the output of 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 (BQ, GCS, Gemini, Composer).

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

