Scale for Big Schemas
For 10s-1000s of Tables, Columns, and Flows
Today’s Lecture

1. How to communicate DBs with
   ○ x-teams, customers, stakeholders?

2. What are good designs?
Database Design

• **Database design: Why?**
  • Agree on schema for use cases now (and later)

• **Consider issues such as:**
  • What entities to model
  • How entities are related
  • What constraints exist in the domain
  • How to achieve good designs

• **Several formalisms exist**
  • We discuss some flavors (ER diagrams, DAG diagrams)
Example 1: NCAA Basketball -- schema for 1 table in BigQuery

1. How did find relationships between columns?
2. How about 10x-100x tables? Columns?
Example 2: Shopify's simplified ERD (Entity Relation Diagram)
Example 3: Complex data flows

DBT’s DAG for Data Flow

Spark’s DAG for Data Flow

Note: 115 Stages in pipeline !!!
Problems

1. How to connect schemas across
   a. 10s-1000s of Tables, Columns, Relationships, and Flows?

2. How teams collaborate on Big Schemas?
   a. Different subsets useful to App team, Data Analysts, Data Eng...

3. When schemas change?
Intuition: Cooking Prep

Source (raw)
- lettuce
- onions
- cucumber
- tomatoes
- water
- vinegar
- Italian spices

Stage 1
- Chopped lettuce
- Sliced onions
- Sliced cucumber
- wedge-cut tomatoes
- Boiled water
- Italian dressing
- vinegar

Stage 2
- Basic salad
- Soup broth
- . . .
- Tortilla

Stage 3
- Caesar’s salad
- Tostada
- . . .
- Enchilada

Waiter -> User
- Combo 1
- Soup and Salad
- . . .
- Burrito Special

Idea
1. Organize prep stages for reuse
2. Show relations between food

Cooking credit: Christine at DBT
Problems

Personas

Schema Graph
(Guha’s talk on datacommons.org)

Big Schemas
Big Schema

1. Example1: Amazon Product orders

2. Example2: NCAA Basketball schema
   i. “Cooked” version
For Project 3

i. Use tool to convey below. Or something equivalent in text/figures.

ii. Important: **Start from** [https://bigschema.io](https://bigschema.io) to create new diagram. Don’t change the example links.

- **Analysis of your dataset (10%)**
  - Students show that they are using a meaningful dataset in terms of size and complexity. The overall dataset should be at least 250 MB.
  - Students clearly describe the information captured in the table.
  - It is clear that students understand the structure of their datasets such as data sizes and high-level relationships between tables.
  - Students list the keys and foreign keys between tables that will be used for exploration or describe connections between tables in some other way.
  - Students use a very simple or very small dataset (e.g. only one table with few columns or a dataset with very few tuples overall).
  - Little to no effort in explaining the dataset. It is not clear to the grader that the student(s) behind the project understands the structure of the data they are working with.