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?

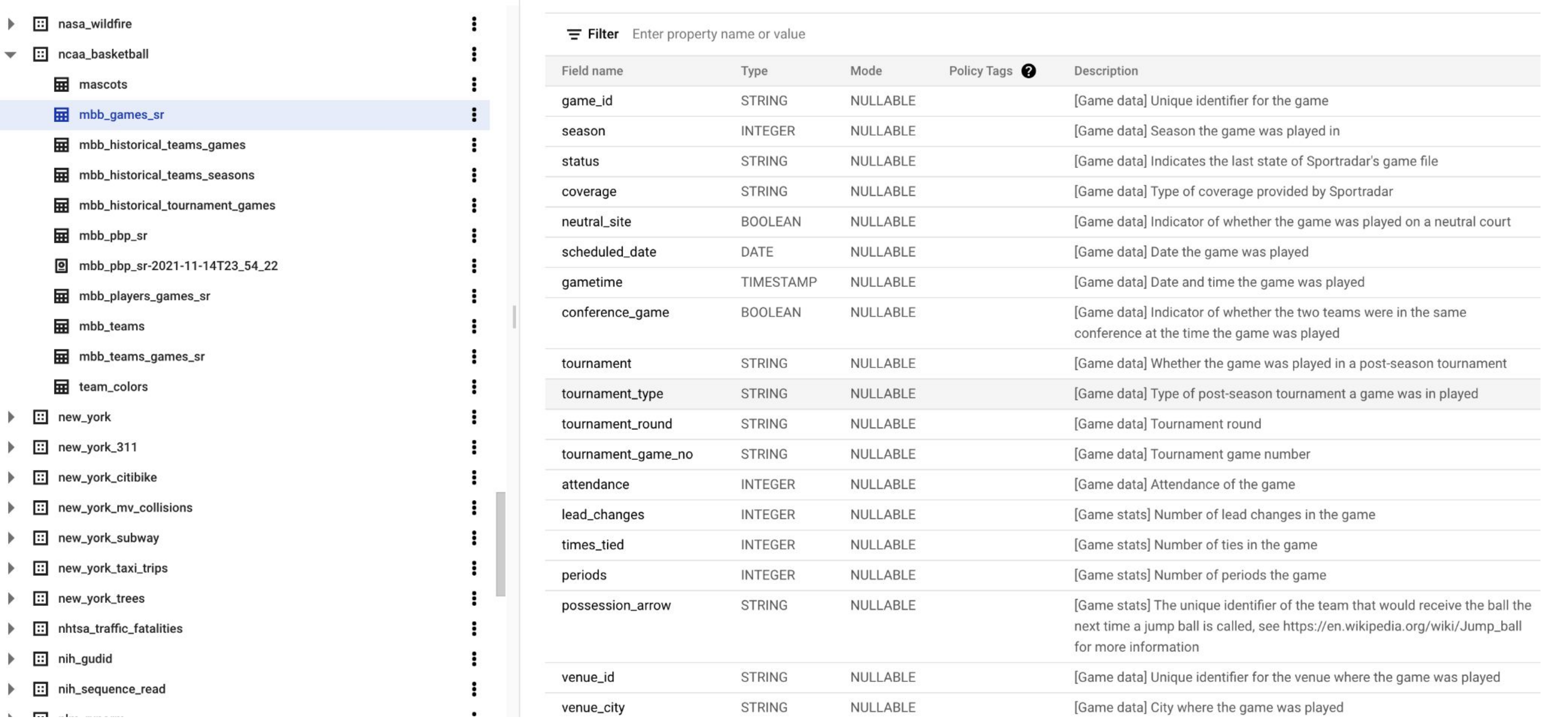
Database Design

- **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 <u>good</u> designs

- **Several formalisms exist**
  - We discuss some flavors (ER diagrams, DAG diagrams)

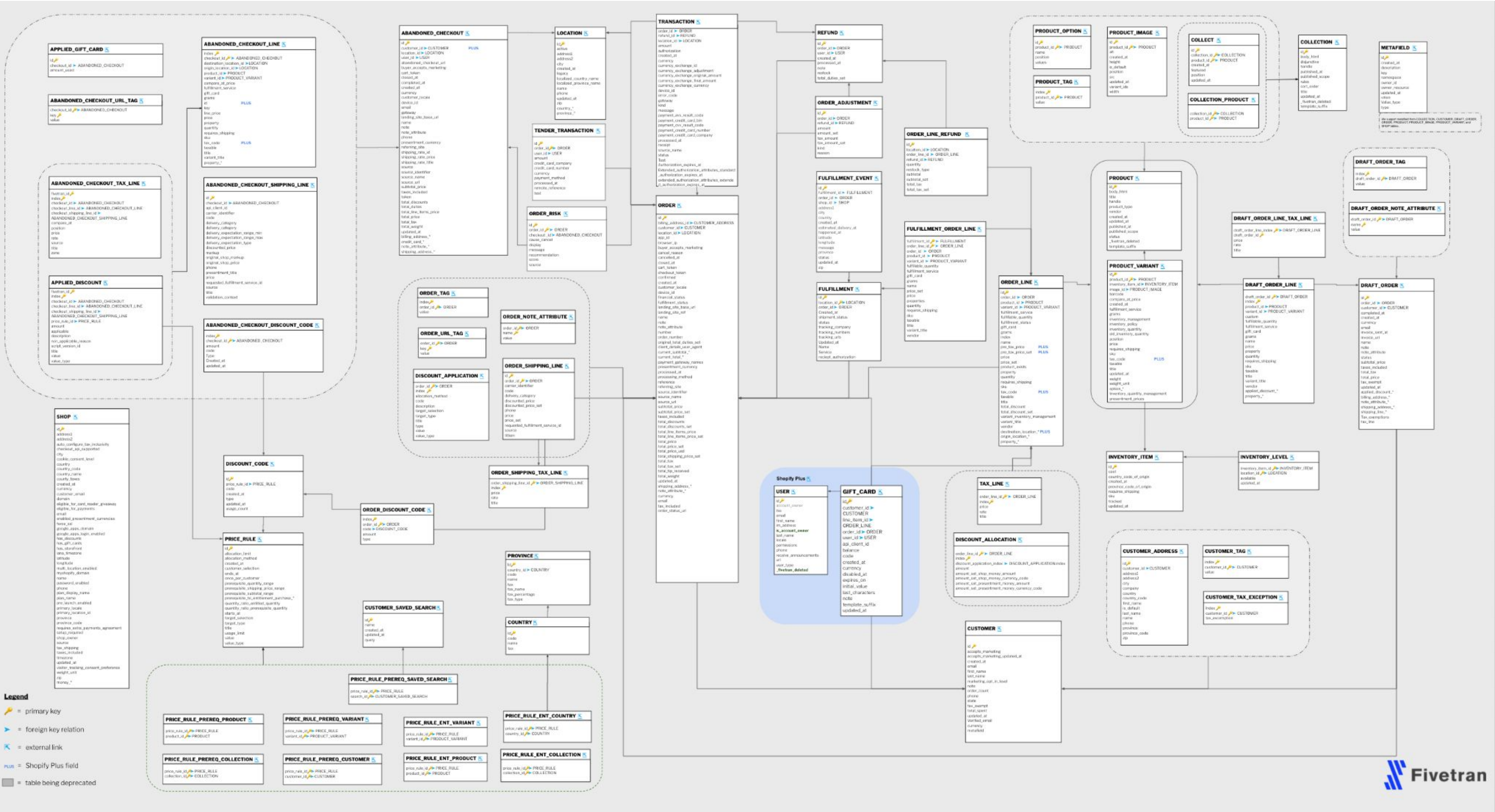
Example 1: NCAA Basketball -- schema for 1 table in BigQuery



Field name	Type	Mode	Policy Tags	Description
game_id	STRING	NULLABLE		[Game data] Unique identifier for the game
season	INTEGER	NULLABLE		[Game data] Season the game was played in
status	STRING	NULLABLE		[Game data] Indicates the last state of Sportradar's game file
coverage	STRING	NULLABLE		[Game data] Type of coverage provided by Sportradar
neutral_site	BOOLEAN	NULLABLE		[Game data] Indicator of whether the game was played on a neutral court
scheduled_date	DATE	NULLABLE		[Game data] Date the game was played
gametime	TIMESTAMP	NULLABLE		[Game data] Date and time the game was played
conference_game	BOOLEAN	NULLABLE		[Game data] Indicator of whether the two teams were in the same conference at the time the game was played
tournament	STRING	NULLABLE		[Game data] Whether the game was played in a post-season tournament
tournament_type	STRING	NULLABLE		[Game data] Type of post-season tournament a game was in played
tournament_round	STRING	NULLABLE		[Game data] Tournament round
tournament_game_no	STRING	NULLABLE		[Game data] Tournament game number
attendance	INTEGER	NULLABLE		[Game data] Attendance of the game
lead_changes	INTEGER	NULLABLE		[Game stats] Number of lead changes in the game
times_tied	INTEGER	NULLABLE		[Game stats] Number of ties in the game
periods	INTEGER	NULLABLE		[Game stats] Number of periods the game
possession_arrow	STRING	NULLABLE		[Game stats] The unique identifier of the team that would receive the ball the next time a jump ball is called, see https://en.wikipedia.org/wiki/Jump_ball for more information
venue_id	STRING	NULLABLE		[Game data] Unique identifier for the venue where the game was played
venue_city	STRING	NULLABLE		[Game data] City where the game was played

1. How did find relationships between columns?
2. How about 10x-100x tables? Columns?

Example2: Shopify's *simplified* ERD (Entity Relation Diagram)



APPLIED_GIFT_CARD
id 🔑
checkout_id ➤ ABANDONED_CHECKOUT
amount_used

ABANDONED_CHECKOUT_URL_TAG
checkout_id 🔑➤ ABANDONED_CHECKOUT
key 🔑
value

ABANDONED_CHECKOUT_LINE
index 🔑
checkout_id 🔑➤ ABANDONED_CHECKOUT
destination_location_id ➤ LOCATION
origin_location_id ➤ LOCATION
product_id ➤ PRODUCT
variant_id ➤ PRODUCT_VARIANT
compare_at_price
fulfillment_service
gift_card
grams
id PLUS
key
line_price
price
property
quantity
requires_shipping
sku
tax_code PLUS
taxable
title
variant_title
property_*

ABANDONED_CHECKOUT_TAX_LINE
fivetran_id 🔑
index 🔑
checkout_id ➤ ABANDONED_CHECKOUT
checkout_line_id ➤ ABANDONED_CHECKOUT_LINE
checkout_shipping_line_id ➤ ABANDONED_CHECKOUT_SHIPPING_LINE
compare_at
position
price
rate
source
title
zone

ABANDONED_CHECKOUT_SHIPPING_LINE
id 🔑
checkout_id ➤ ABANDONED_CHECKOUT
api_client_id
carrier_identifier
code
delivery_category
delivery_category
delivery_expectation_range_min
delivery_expectation_range_max
delivery_expectation_type
discounted_price
markup
original_shop_markup
original_shop_price
phone
presentment_title
price
requested_fulfillment_service_id
source
title
validation_context

APPLIED_DISCOUNT
fivetran_id 🔑
index 🔑
checkout_id ➤ ABANDONED_CHECKOUT
checkout_line_id ➤ ABANDONED_CHECKOUT_LINE
checkout_shipping_line_id ➤ ABANDONED_CHECKOUT_SHIPPING_LINE
price_rule_id ➤ PRICE_RULE
amount
applicable
description
non_applicable_reason
script_version_id
title
value
value_type

ABANDONED_CHECKOUT_DISCOUNT_CODE
index 🔑
checkout_id 🔑➤ ABANDONED_CHECKOUT
amount
code
Type
Created_at
updated_at

ABANDONED_CHECKOUT
id 🔑
customer_id ➤ CUSTOMER PLUS
location_id ➤ LOCATION
user_id ➤ USER
abandoned_checkout_url
buyer_accepts_marketing
cart_token
closed_at
completed_at
created_at
currency
customer_locale
device_id
email
gateway
landing_site_base_url
name
note
note_attribute
phone
presentment_currency
referring_site
shipping_rate_id
shipping_rate_price
shipping_rate_title
source
source_identifier
source_name
source_url
subtotal_price
taxes_included
token
total_discounts
total_duties
total_line_items_price
total_price
total_tax
total_weight
updated_at
billing_address_*
credit_card_*
note_attribute_*
shipping_address_*

LOCATION
id 🔑
active
address1
address2
city
created_at
legacy
localized_country_name
localized_province_name
name
phone
updated_at
zip
country_*
province_*

TENDER_TRANSACTION
id 🔑
order_id 🔑➤ ORDER
user_id ➤ USER
amount
credit_card_company
credit_card_number
currency
payment_method
processed_at
remote_reference
test

ORDER_RISK
id 🔑
order_id 🔑➤ ORDER
checkout_id ➤ ABANDONED_CHECKOUT
cause_cancel
display
message
recommendation
score
source

TRANSACTION
order_id ➤ ORDER
refund_id ➤ REFUND
location_id ➤ LOCATION
amount
authorization
created_at
currency
currency_exchange_id
currency_exchange_adjustment
currency_exchange_original_amount
currency_exchange_final_amount
currency_exchange_currency
device_id
error_code
gateway
kind
message
payment_avs_result_code
payment_credit_card_bin
payment_cvv_result_code
payment_credit_card_number
payment_credit_card_company
processed_at
receipt
source_name
status
Test
Authorization_expires_at
Extended_authorization_attributes_standard_authorization_expires_at
extended_authorization_attributes_extended_authorization_expires_at

ORDER
id 🔑
billing_address_id ➤ CUSTOMER_ADDRESS
customer_id ➤ CUSTOMER
location_id ➤ LOCATION
app_id
browser_ip
buyer_accepts_marketing
cancel_reason
cancelled_at
closed_at
cart_token
checkout_token
confirmed
created_at
customer_locale
device_id
financial_status
fulfillment_status
landing_site_base_url
landing_site_ref
name
note
note_attribute
number
order_number
original_total_duties_set
client_details_user_agent
current_subtotal_*
current_total_*
payment_gateway_names
presentment_currency
processed_at
processing_method
reference
referring_site
source_identifier
source_name
source_url
subtotal_price
subtotal_price_set
taxes_included
total_discounts
total_discounts_set
total_line_items_price
total_line_items_price_set
total_price
total_price_set
total_price_usd
total_shipping_price_set
total_tax
total_tax_set
total_tip_received
total_weight
updated_at
shipping_address_*
note_attribute_*
currency
email
tax_included
order_status_url

ORDER_TAG
index 🔑
order_id 🔑➤ ORDER
value

ORDER_NOTE_ATTRIBUTE
order_id 🔑➤ ORDER
name 🔑
value

ORDER_URL_TAG
order_id 🔑➤ ORDER
key 🔑
value

DISCOUNT_APPLICATION
order_id 🔑➤ ORDER
index 🔑
allocation_method
code
description
target_selection
target_type
title
type
value
value_type

ORDER_SHIPPING_LINE
id 🔑
order_id 🔑➤ ORDER
carrier_identifier
code
delivery_category
discounted_price
discounted_price_set
phone
price
price_set
requested_fulfillment_service_id
source
title

ORDER_SHIPPING_TAX_LINE
order_shipping_line_id 🔑➤ ORDER_SHIPPING_LINE
index 🔑
price
rate
title

ORDER_DISCOUNT_CODE
index 🔑
order_id 🔑➤ ORDER
code ➤ DISCOUNT_CODE
amount
type

PROVINCE
id 🔑
country_id ➤ COUNTRY
code
name
tax
tax_name
tax_percentage
tax_type

COUNTRY
id 🔑
code
name
tax

CUSTOMER_SAVED_SEARCH
id 🔑
name
created_at
updated_at
query

SHOP
id 🔑
address1
address2
auto_configure_tax_inclusivity
checkout_api_supported
city
cookie_consent_level
country
country_code
country_name
county_taxes
created_at
currency
customer_email
domain
eligible_for_card_reader_giveaway
eligible_for_payments
email
enabled_presentment_currencies
force_ssl
google_apps_domain
google_apps_login_enabled
has_discounts
has_gift_cards
has_storefront
iana_timezone
latitude
longitude
multi_location_enabled
myshopify_domain
name
password_enabled
phone
plan_display_name
plan_name
pre_launch_enabled
primary_locale
primary_location_id
province
province_code
requires_extra_payments_agreement
setup_required
shop_owner
source
tax_shipping
taxes_included
timezone
updated_at
visitor_tracking_consent_preference
weight_unit
zip
money_*

DISCOUNT_CODE
id 🔑
price_rule_id ➤ PRICE_RULE
code
created_at
type
updated_at
usage_count

PRICE_RULE
id 🔑
allocation_limit
allocation_method
created_at
customer_selection
ends_at
once_per_customer
prerequisite_quantity_range
prerequisite_shipping_price_range
prerequisite_subtotal_range
prerequisite_to_entitlement_purchase_*
quantity_ratio_entitled_quantity
quantity_ratio_prerequisite_quantity
starts_at
target_selection
target_type
title
usage_limit
value
value_type

PRICE_RULE_PREREQ_SAVED_SEARCH
price_rule_id 🔑➤ PRICE_RULE
search_id 🔑➤ CUSTOMER_SAVED_SEARCH

PRICE_RULE_PREREQ_PRODUCT
price_rule_id 🔑➤ PRICE_RULE
product_id 🔑➤ PRODUCT

PRICE_RULE_PREREQ_VARIANT
price_rule_id 🔑➤ PRICE_RULE
variant_id 🔑➤ PRODUCT_VARIANT

PRICE_RULE_ENT_VARIANT
price_rule_id 🔑➤ PRICE_RULE
variant_id 🔑➤ PRODUCT_VARIANT

PRICE_RULE_ENT_COUNTRY
price_rule_id 🔑➤ PRICE_RULE
country_id 🔑➤ COUNTRY

PRICE_RULE_PREREQ_COLLECTION
price_rule_id 🔑➤ PRICE_RULE
collection_id 🔑➤ COLLECTION

PRICE_RULE_PREREQ_CUSTOMER
price_rule_id 🔑➤ PRICE_RULE
customer_id 🔑➤ CUSTOMER

PRICE_RULE_ENT_PRODUCT
price_rule_id 🔑➤ PRICE_RULE
product_id 🔑➤ PRODUCT

PRICE_RULE_ENT_COLLECTION
price_rule_id 🔑➤ PRICE_RULE
collection_id 🔑➤ COLLECTION

REFUND
id 🔑
order_id ➤ ORDER
user_id ➤ USER
created_at
processed_at
note
restock
total_duties_set

ORDER_ADJUSTMENT
id 🔑
order_id ➤ ORDER
refund_id ➤ REFUND
amount
amount_set
tax_amount
tax_amount_set
kind
reason

ORDER_LINE_REFUND
id 🔑
location_id ➤ LOCATION
order_line_id ➤ ORDER_LINE
refund_id ➤ REFUND
quantity
restock_type
subtotal
subtotal_set
total_tax
total_tax_set

FULFILLMENT_EVENT
id 🔑
fulfillment_id ➤ FULFILLMENT
order_id ➤ ORDER
shop_id ➤ SHOP
address1
city
country
created_at
estimated_delivery_at
happened_at
latitude
longitude
message
province
status
updated_at
zip

FULFILLMENT
id 🔑
location_id 🔑➤ LOCATION
order_id ➤ ORDER
Created_at
shipment_status
status
tracking_company
tracking_numbers
tracking_urls
Updated_at
Name
Service
reciept_authorization

FULFILLMENT_ORDER_LINE
fulfillment_id 🔑➤ FULFILLMENT
order_line_id 🔑➤ ORDER_LINE
order_id ➤ ORDER
product_id ➤ PRODUCT
variant_id ➤ PRODUCT_VARIANT
fulfillable_quantity
fulfillment_service
gift_card
grams
name
price_set
price
properties
quantity
requires_shipping
sku
taxable
title
variant_title
vendor

ORDER_LINE
id 🔑
order_id ➤ ORDER
product_id ➤ PRODUCT
variant_id ➤ PRODUCT_VARIANT
fulfillment_service
fulfillable_quantity
fulfillment_status
gift_card
grams
index
name
pre_tax_price PLUS
pre_tax_price_set PLUS
price
price_set
product_exists
property
quantity
requires_shipping
sku
tax_code PLUS
taxable
title
total_discount
total_discount_set
variant_inventory_management
variant_title
vendor
destination_location_* PLUS
origin_location_*
property_*

Shopify Plus

USER
id 🔑
account_owner
bio
email
first_name
im_address
is_account_owner
last_name
locale
permissions
phone
receive_announcements
url
user_type
_fivetran_deleted

GIFT_CARD
id 🔑
customer_id ➤ CUSTOMER
line_item_id ➤ ORDER_LINE
order_id ➤ ORDER
user_id ➤ USER
api_client_id
balance
code
created_at
currency
disabled_at
expires_on
initial_value
last_characters
note
template_suffix
updated_at

TAX_LINE
order_line_id 🔑➤ ORDER_LINE
index 🔑
price
rate
title

DISCOUNT_ALLOCATION
order_line_id 🔑➤ ORDER_LINE
index 🔑
discount_application_index ➤ DISCOUNT_APPLICATION.index
amount
amount_set_shop_money_amount
amount_set_shop_money_currency_code
amount_set_presentment_money_amount
amount_set_presentment_money_currency_code

CUSTOMER
id 🔑
accepts_marketing
accepts_marketing_updated_at
created_at
email
first_name
last_name
marketing_opt_in_level
note
order_count
phone
state
tax_exempt
total_spent
updated_at
Verified_email
currency
metafield

PRODUCT_OPTION
id 🔑
product_id 🔑➤ PRODUCT
name
position
values

PRODUCT_TAG
index 🔑
product_id 🔑➤ PRODUCT
value

PRODUCT_IMAGE
id 🔑
product_id 🔑➤ PRODUCT
alt
created_at
height
is_default
position
src
updated_at
variant_ids
width

COLLECT
id 🔑
collection_id 🔑➤ COLLECTION
product_id 🔑➤ PRODUCT
created_at
featured
position
updated_at

COLLECTION_PRODUCT
collection_id 🔑➤ COLLECTION
product_id 🔑➤ PRODUCT

COLLECTION
id 🔑
body_html
disjunctive
handle
published_at
published_scope
rules
sort_order
title
updated_at
_fivetran_deleted
template_suffix

METAFIELD
id 🔑
created_at
description
key
namespace
owner_id
owner_resource
updated_at
value
Value_type
type

We support metafield from COLLECTION, CUSTOMER, DRAFT_ORDER ORDER, PRODUCT, PRODUCT_IMAGE, PRODUCT_VARIANT, and SHOP tables.

PRODUCT
id 🔑
body_html
title
handle
product_type
vendor
created_at
updated_at
published_at
published_scope
status
_fivetran_deleted
template_suffix

PRODUCT_VARIANT
id 🔑
product_id 🔑➤ PRODUCT
inventory_item_id ➤ INVENTORY_ITEM
image_id ➤ PRODUCT_IMAGE
barcode
compare_at_price
created_at
fulfillment_service
grams
inventory_management
inventory_policy
inventory_quantity
old_inventory_quantity
position
price
requires_shipping
sku
tax_code PLUS
taxable
title
updated_at
weight
weight_unit
option_*
inventory_quantity_management
presentment_prices

DRAFT_ORDER_TAG
index 🔑
draft_order_id 🔑➤ DRAFT_ORDER
value

DRAFT_ORDER_NOTE_ATTRIBUTE
draft_order_id 🔑➤ DRAFT_ORDER
name 🔑
value

DRAFT_ORDER_LINE_TAX_LINE
draft_order_line_index 🔑➤ DRAFT_ORDER_LINE
draft_order_id 🔑
price
rate
title

DRAFT_ORDER_LINE
draft_order_id 🔑➤ DRAFT_ORDER
index 🔑
product_id ➤ PRODUCT
variant_id ➤ PRODUCT_VARIANT
custom
fulfillable_quantity
fulfillment_service
gift_card
grams
name
price
property
quantity
requires_shipping
sku
taxable
title
variant_title
vendor
applied_discount_*
property_*

DRAFT_ORDER
id 🔑
order_id ➤ ORDER
customer_id ➤ CUSTOMER
completed_at
created_at
currency
email
invoice_sent_at
invoice_url
name
note
note_attribute
status
subtotal_price
taxes_included
total_tax
total_price
tax_exempt
updated_at
applied_discount_*
billing_address_*
note_attribute_*
shipping_address_*
shipping_line_*
Tax_exemptions
tax_line

INVENTORY_ITEM
id 🔑
cost
country_code_of_origin
created_at
province_code_of_origin
requires_shipping
sku
tracked
updated_at

INVENTORY_LEVEL
inventory_item_id 🔑➤ INVENTORY_ITEM
location_id 🔑➤ LOCATION
available
updated_at

CUSTOMER_ADDRESS
id 🔑
customer_id ➤ CUSTOMER
address1
address2
city
company
country
country_code
first_name
is_default
last_name
name
phone
province
province_code
zip

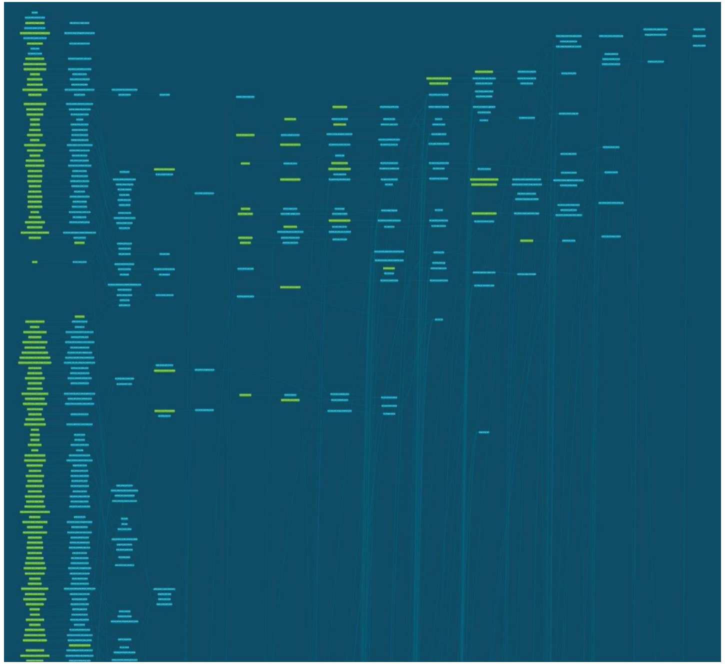
CUSTOMER_TAG
index 🔑
customer_id 🔑➤ CUSTOMER
value

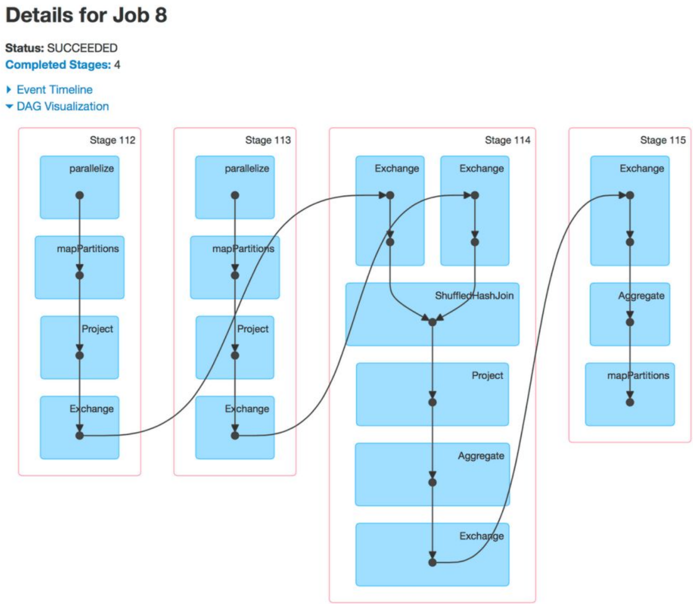
CUSTOMER_TAX_EXCEPTION
index 🔑
customer_id 🔑➤ CUSTOMER
tax_exemption

Legend
🔑 = primary key
➤ = foreign key relation
↖ = external link
PLUS = Shopify Plus field
▨ = table being deprecated

Fivetran

Example3: Complex data flows





Details for Job 8

Status: SUCCEEDED

Completed Stages: 4

▸ Event Timeline

▾ DAG Visualization

Note: 115 Stages in pipeline !!!

DBT’s DAG for Data Flow

Spark’s DAG for Data Flow

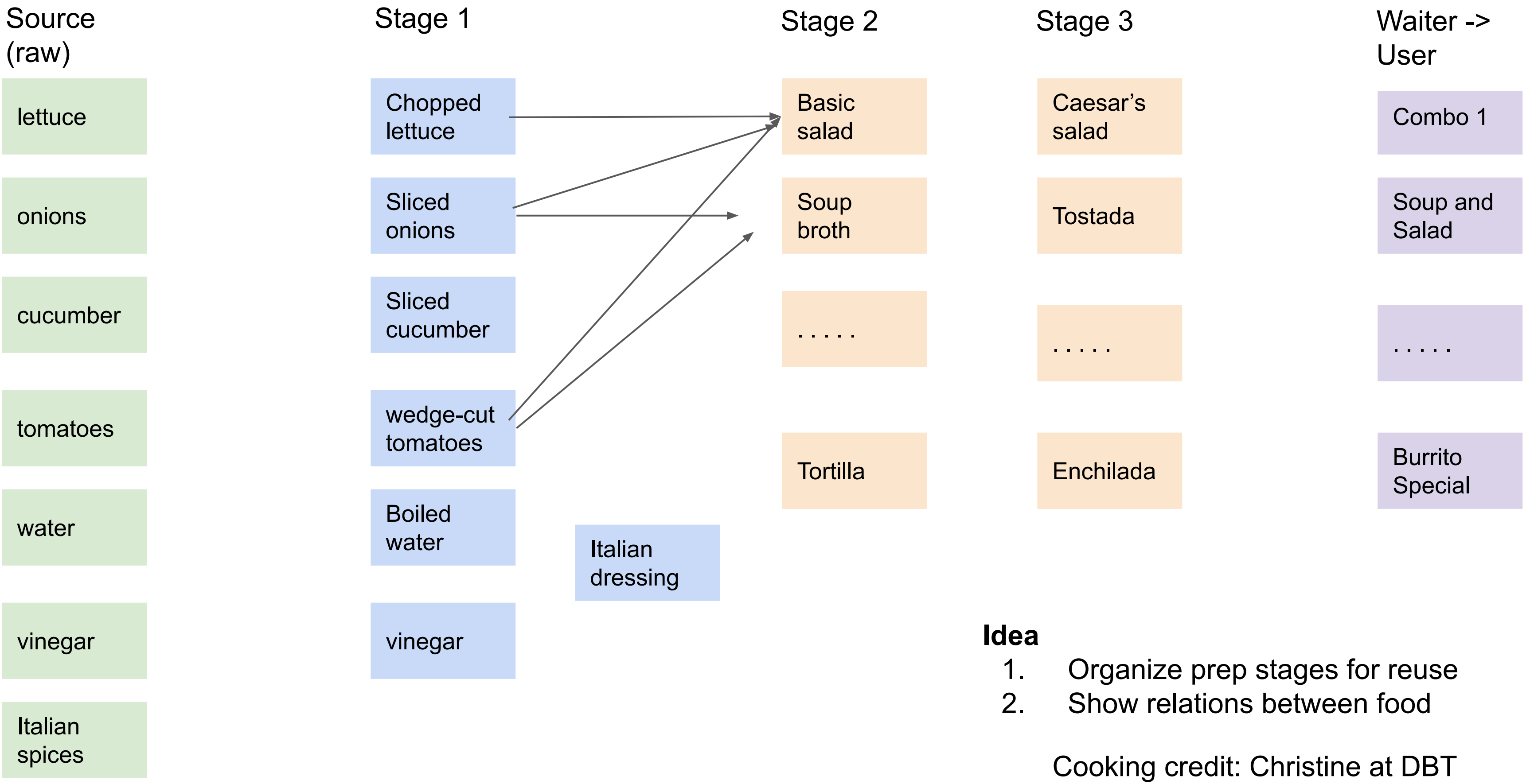
Problems

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

vinegar

Italian dressing

Stage 2

Basic salad

Soup broth

. . . . .

Tortilla

Stage 3

Caesar’s salad

Tostada

. . . . .

Enchilada

Waiter -> User

Combo 1

Soup and Salad

. . . . .

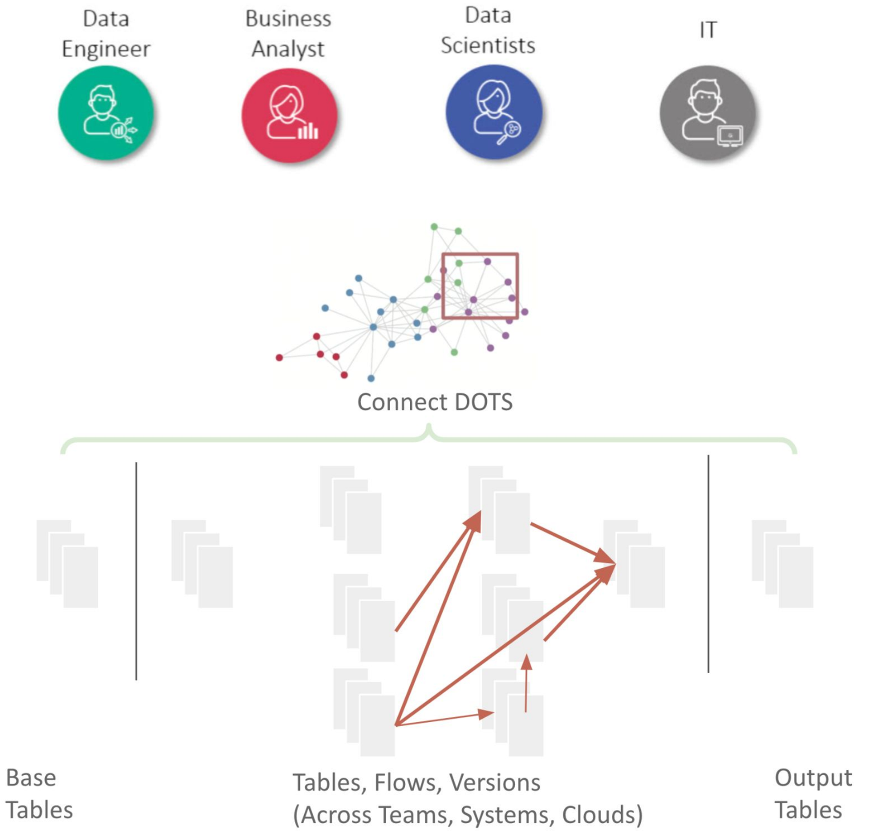
Burrito Special

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

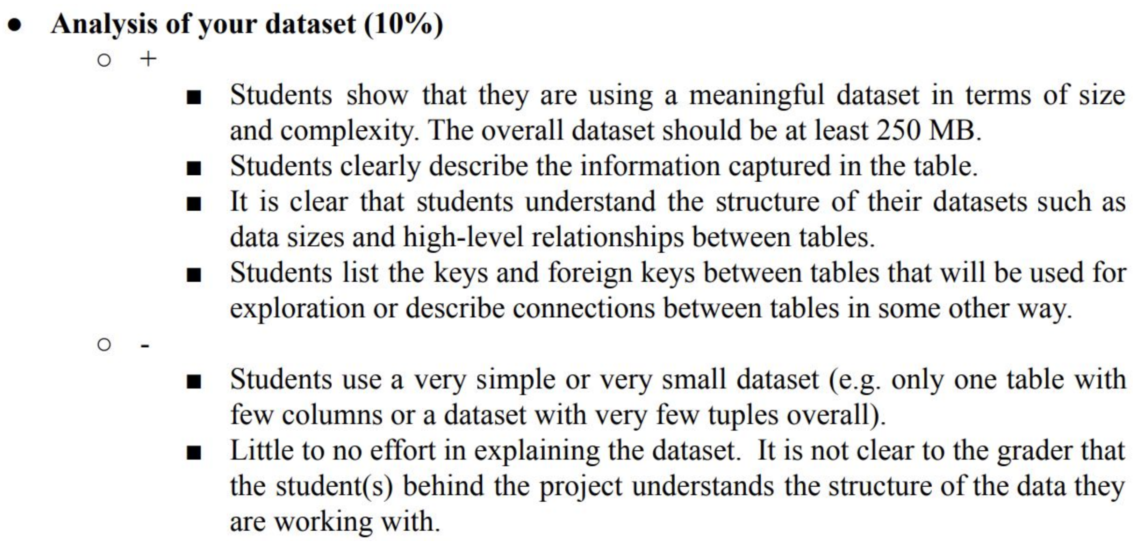
1. Example1: Amazon Product orders

2. Example2: NCAA Basketball schema
    i. "Cooked" version

For Project3

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

ii. Important: **Start from** 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.