Dimensional Data Modeling

Day 2

EcZachly Inc

What we’ll cover today

EcZachly Inc

- Idempotent pipelines
- Slowly-changing dimensions

Idempotent pipelines are CRITICAL

EcZachly Inc

Your pipeline produces the same results regardless of when it's ran!!!!

What does idempotent mean?

- **denoting an element of a set which is unchanged in value when multiplied or otherwise operated on by itself**

Pipelines should produce the same results

EcZachly Inc

- Regardless of the day you run it
- Regardless of how many times you run it
- Regardless of the hour that you run it

Why is troubleshooting non-idempotent pipelines hard?

EcZachly Inc

- Silent failure!
- You only see it when you get data inconsistencies and a data analyst yells at you

What can make a pipeline not idempotent

- INSERT INTO without TRUNCATE
  - Use MERGE or INSERT OVERWRITE every time please
- Using Start_date > without a corresponding end_date <
- Not using a full set of partition sensors
  - (pipeline might run when there is no/partial data)
- Not using **depends_on_past** for cumulative pipelines

What can make a pipeline not idempotent

- Relying on the “latest” partition of a not properly modeled SCD table
  - So much pain at Facebook, DAILY DIMENSIONS AND “latest” partition is a very bad idea
  - Cumulative table design AMPLIFIES this bug
- Relying on the “latest” partition of anything else

The pains of not having idempotent pipelines

EcZachly Inc

- Backfilling causes inconsistencies between the old and restated data
- Very hard to troubleshoot bugs
- Unit testing cannot replicate the production behavior
- Silent failures

Should you model as Slowly Changing Dimensions?

- Max, the creator of Airflow HATES SCD data modeling
  - Link to Max's article about why SCD's SUCK
- What are the options here?
  - Latest snapshot
  - Daily/Monthly/Yearly snapshot
  - SCD
- How slowly changing are the dimensions you're modeling?

Why do dimensions change?

EcZachly Inc

- Someone decides they hate iPhone and want Android now
- Someone migrates from team dog to team cat
- Someone migrates from USA to another country
- ETC ETC ETC

How can you model dimensions that change?

EcZachly Inc

- Singular snapshots
  - BE CAREFUL SINCE THESE ARE NOT IDEMPOTENT
- Daily partitioned snapshots
- SCD Types 1,2,3
-

The types of Slowly Changing Dimensions

- Type 0
    - Aren’t actually slowly changing (e.g. birth date)

The types of Slowly Changing Dimensions

EcZachly Inc

- Type 1
    - You only care about the latest value
    - **NEVER USE THIS TYPE BECAUSE IT MAKES YOUR PIPELINES NOT IDEMPOTENT ANYMORE**

The types of Slowly Changing Dimensions

EcZachly Inc

- Type 2
  - You care about what the value was from “start_date” to “end_date”
  - Current values usually have either an end_date that is:
    - NULL
    - Far into the future like 9999-12-31
  - Hard to use:
    - Since there’s more than 1 row per dimension, you need to be careful about filtering on time
  - **MY FAVORITE TYPE OF SCD**
    - The only type of SCD that is purely IDEMPOTENT

The types of Slowly Changing Dimensions

- Type 3
  - You only care about “original” and “current”
  - Benefits
    - You only have 1 row per dimension
  - Drawbacks
    - You lose the history in between original and current
  - Is this idempotent?
    - Partially, which means it’s not

Which types are idempotent?

- Type 0 and Type 2 are idempotent
  - Type 0 is because the values are unchanging
  - Type 2 is but you need to be careful with how you use the **start_date** and **end_date** syntax!
- Type 1 isn't idempotent
  - If you backfill with this dataset, you'll get the dimension as it is now, not as it was then!
- Type 3 isn't idempotent
  - If you backfill with this dataset, it's impossible to know when to pick "original" vs "current" and you'll either

SCD2 Loading

EcZachly Inc

- Load the entire history in one query
  - Inefficient but nimble
  - 1 query and you're done
- Incrementally load the data after the previous SCD is generated
  - Has the same "**depends_on_past**" constraint
  - Efficient but cumbersome

Let’s start the lab

EcZachly Inc