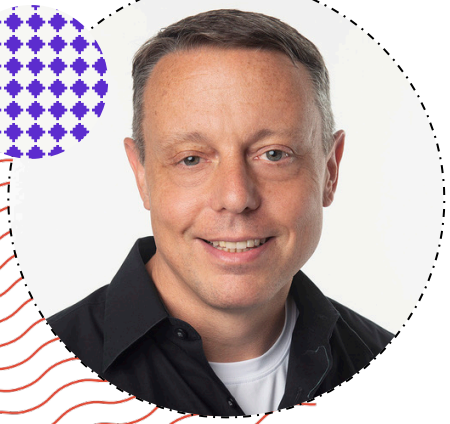
HubSpot MDM Contact Matching & Merging

by Jeff Huth



Am I already in your database?

Hi. Here is my Contact record.

Name: Jeff Huth, Phone Number: 800-867-5309x042

Title: Principal Analytics Engineer

Company Name: Wind River Systems

Company Industry: Tech, Employees: 1,500

Email: jeff.huth@gmail.com, Twitter: @drinkdata

LinkedIn: https://www.linkedin.com/in/huthjeff/

Agenda

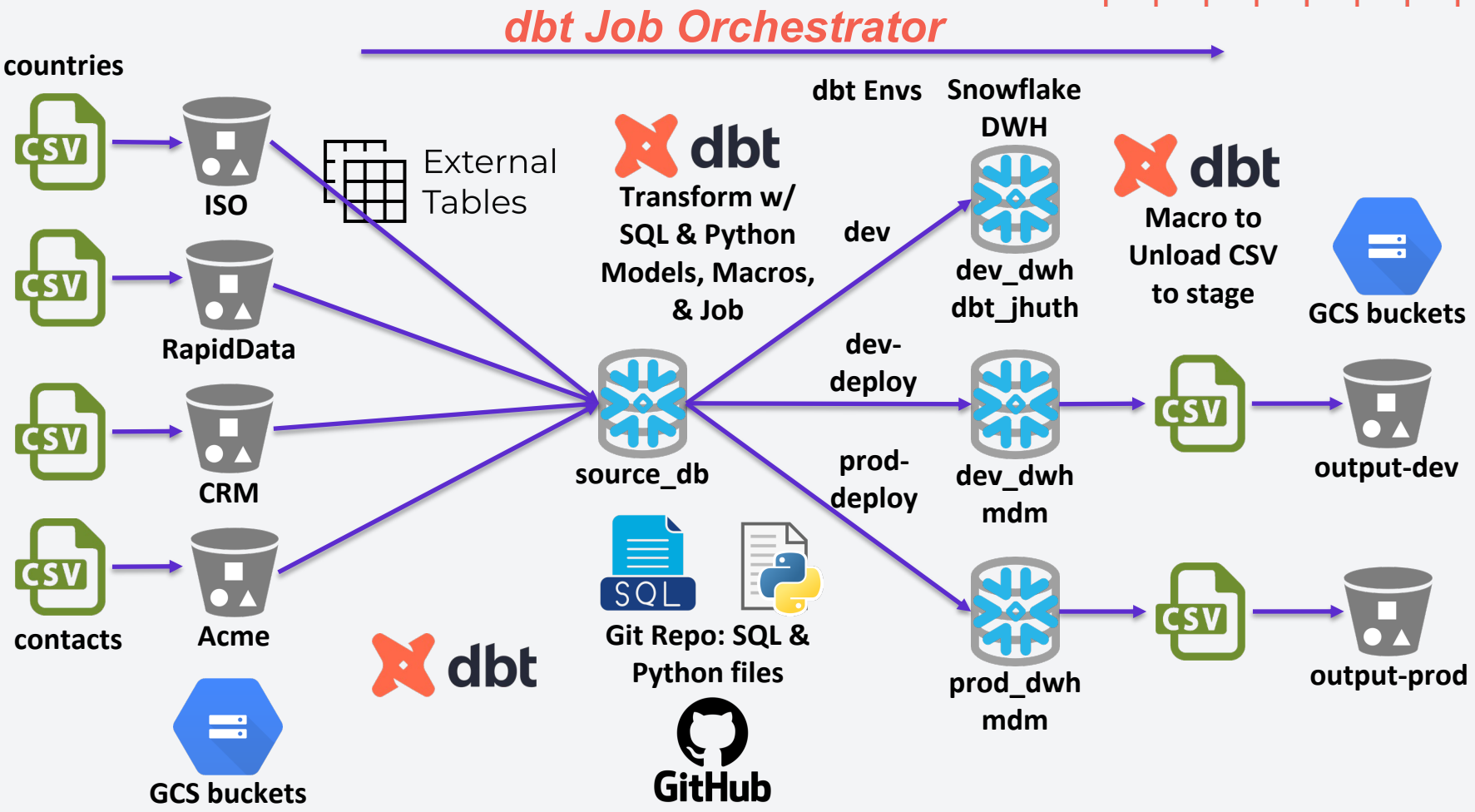
Introductions

Candidate-led code review

Interviewer-led feedback session

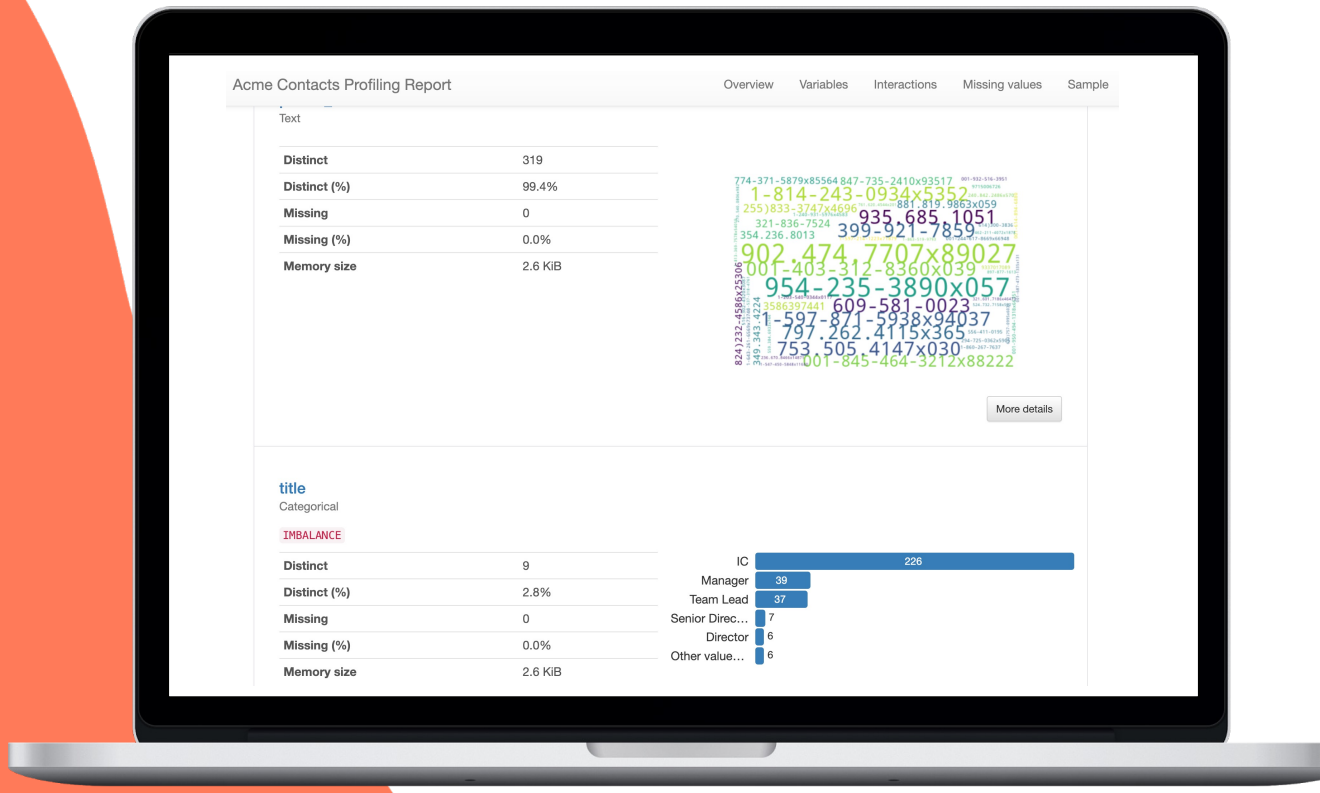
Candidate questions & wrap-up

Data Architecture Diagram



Data Profiling with Python + ydata-profiling

Using a Jupyter Notebook



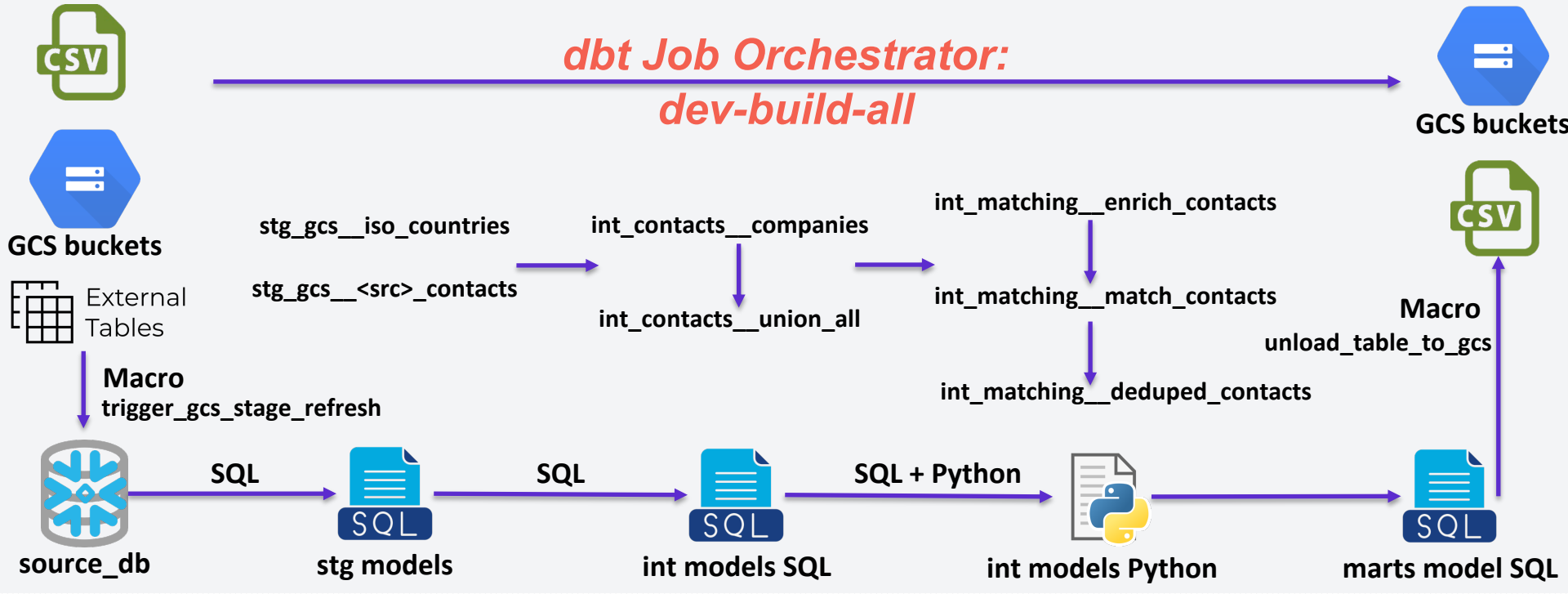
Common fields:
Name, Title, Email, Phone, Company Name + Domain

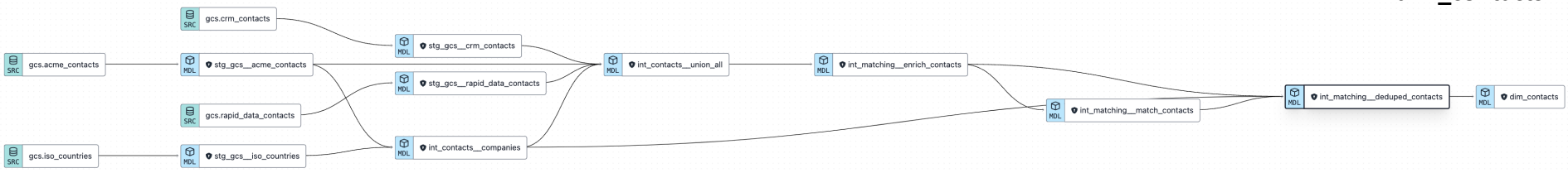
Unique field: Email?

Discover common fields, unique fields, attribute values, NULLs

dbt Model Transformations

dbt





Python Matching & Merging

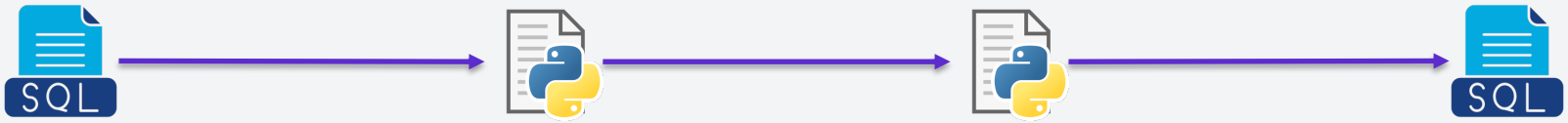
dbt +

int_contacts__union_all
- Union data
- Lookup company/country
- Basic pre-processing

int_matching__enrich_contacts
- Python DataFrames + libraries
- Validate: email, phone, domain, IP address
- Parse first/last name + first initial
- Clean + phonetics (metaphones)

int_matching__match_contacts
- Python DataFrames
- recordlinkage library:
  - Index
  - Compare (rules)
  - Features
  - Matches
- De-duplicate match pairs → array

int_matching__deduped_contacts
- Join enriched + merge keys
- De-duplicate Contacts
- Entity resolution
  - Contact ID (first record)
  - Latest record: name, etc.
  - Latest valid email & phone
  - Latest src specific fields
  - Company info lookup



Normalize & pre-process w/ common fields

Clean, Validate, & Parse

Match & Merge

Assumptions

- Files delivered to GCS from sources
- Data is USA, Canadian, European
  - UTF-8, Latin alphabet
  - Name/phone/email formats
  - Dates in UTC, Currency in $$
- Different people based on differences in: name, email, phone, company, title
- Most recent record is the best & most up-to-date
- Matching rules: 4 (or 5) of 7 compared attributes

Challenges

- BigQuery vs. Snowflake? Python
- Snowflake Snowpark Conda libraries available
- dbt Python models on Snowflake
- recordlinkage (matching) library – getting distinct & unique match key sets
- Airbyte community connectors (GCS)
- GCS to Snowflake: stages, external tables

Next Steps

- Improve matching rules (indexes, rule types)
- Add scoring/weighting rules
- AI/ML: Investigate other algorithms
- Reporting: Visualize and alert
- Orchestration w/ Dagster + Airbyte + Snowflake + dbt
- Data security: Roles, Grants, PII, encryption
- CI/CD Pipelines: dbt, Dagster, GitHub workflows

Thank You!