CS 327E Class 9

November 11, 2019

Announcements

- Grading update

- What to expect from remaining Milestones:
  - Milestone **9**: Find `dataset2` + ingest into BQ + model the data
  - Milestone **10**: Create Beam pipelines + cross-dataset queries
  - Milestone **11**: Orchestrate workflow
  - Milestone **12**: Present your project

- Review your `dataset2` selection: sign-up sheet

1) A data warehouse provides _____

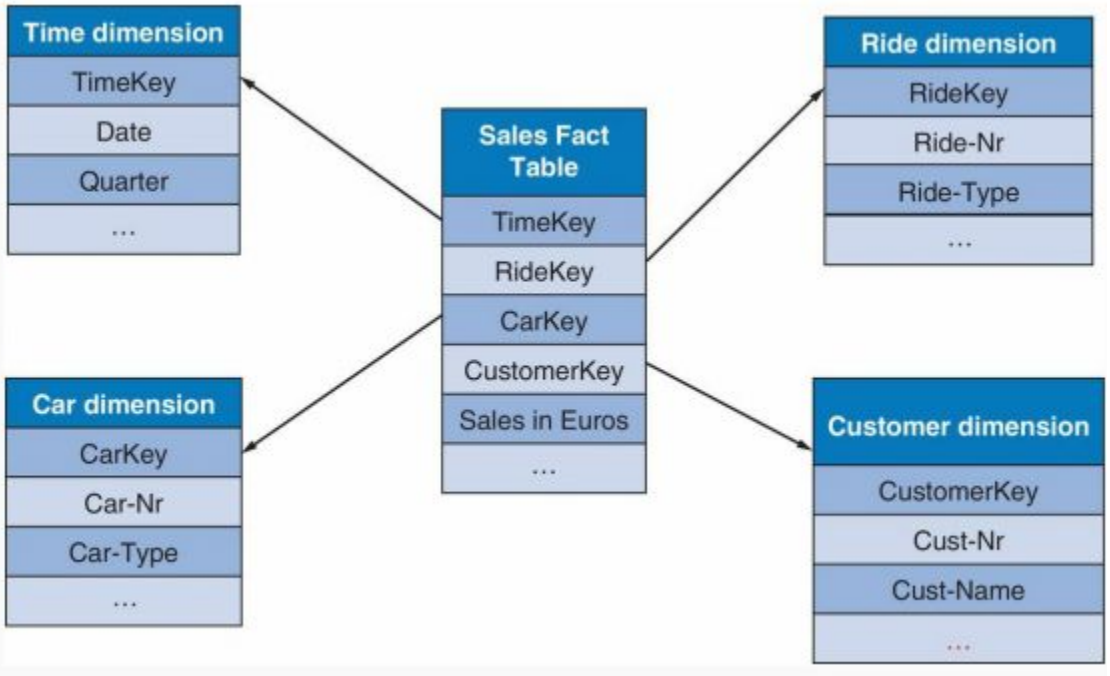
A. a centralized and consolidated data platform by integrating data from different sources and in different formats.

B. an operational data platform with guaranteed consistency during transaction processing.

2) What are the most common schemas of a data warehouse?

A. Star and Snowflake schemas
B. Fact and Dimension schemas
C. Normalized and Denormalized schemas

3) In this Saber data warehouse schema, which column stores a fact/measure?



A. Car-Nr
B. Cust-Nr
C. Sales in Euros
D. None of the above

4) What are some important considerations when designing a data warehouse schema?

A. The grain of the Fact table(s)
B. Identifying the Dimension tables
C. Handling slowly changing dimensions
D. All of the above

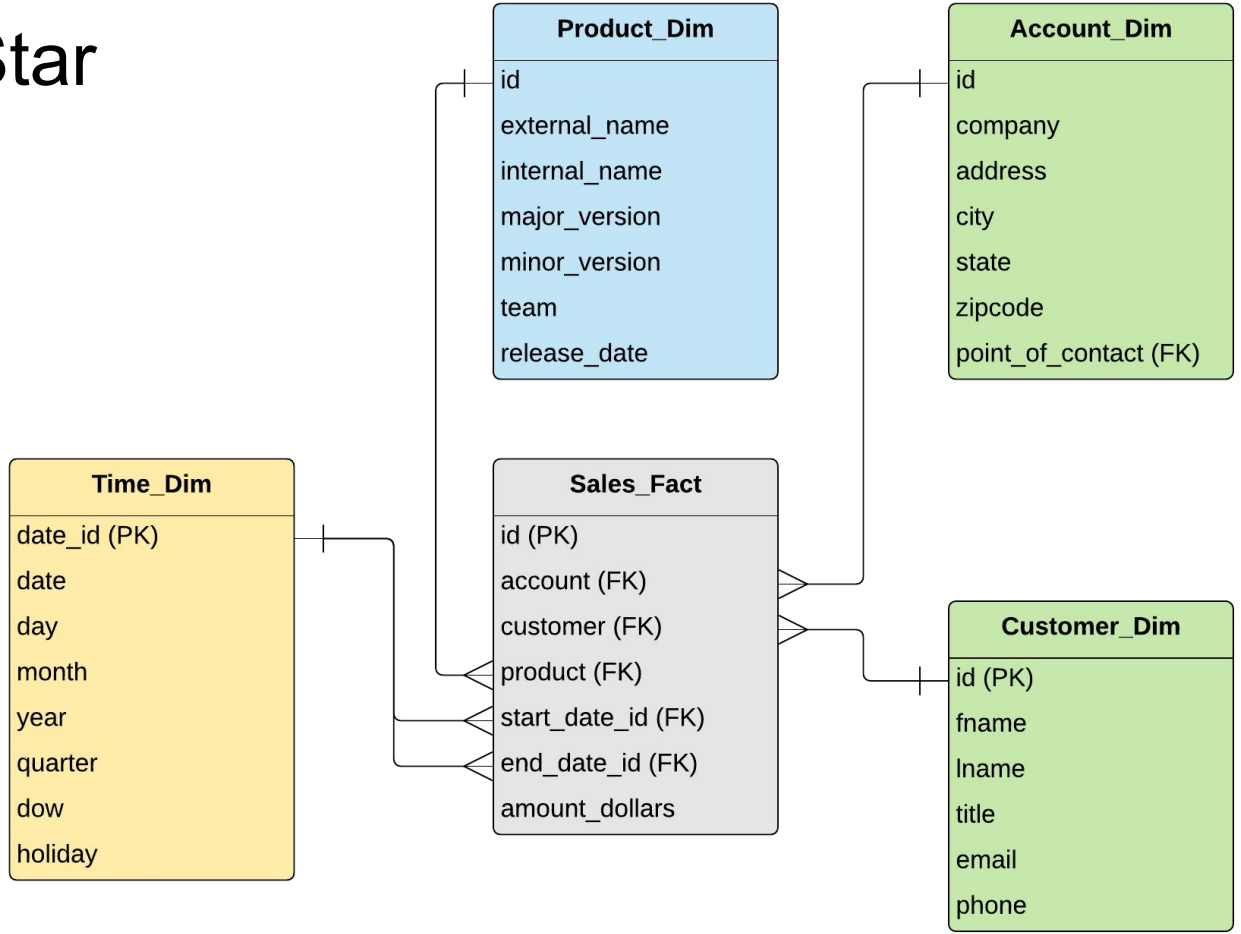
5) What activity can consume 80% of the time when building a data warehouse?

A) Designing the data warehouse schema
B) Building the ETL process
C) Creating the BI reports

6) Just like a data warehouse, a data lake is a central repository of data. Unlike a data warehouse, a data lake stores data in its raw form and its primary users are data scientists.

A) True
B) False

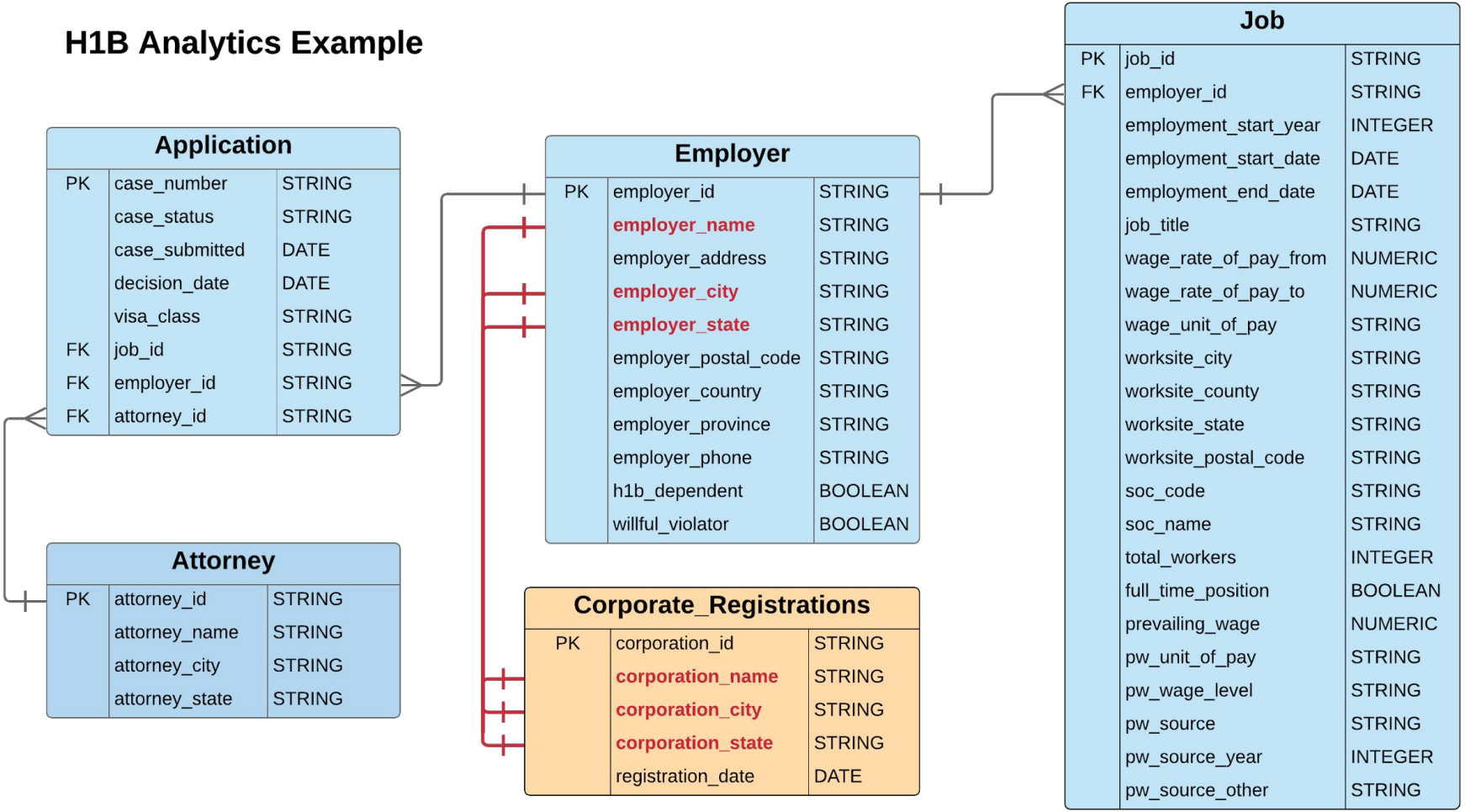
Classic Star Schema

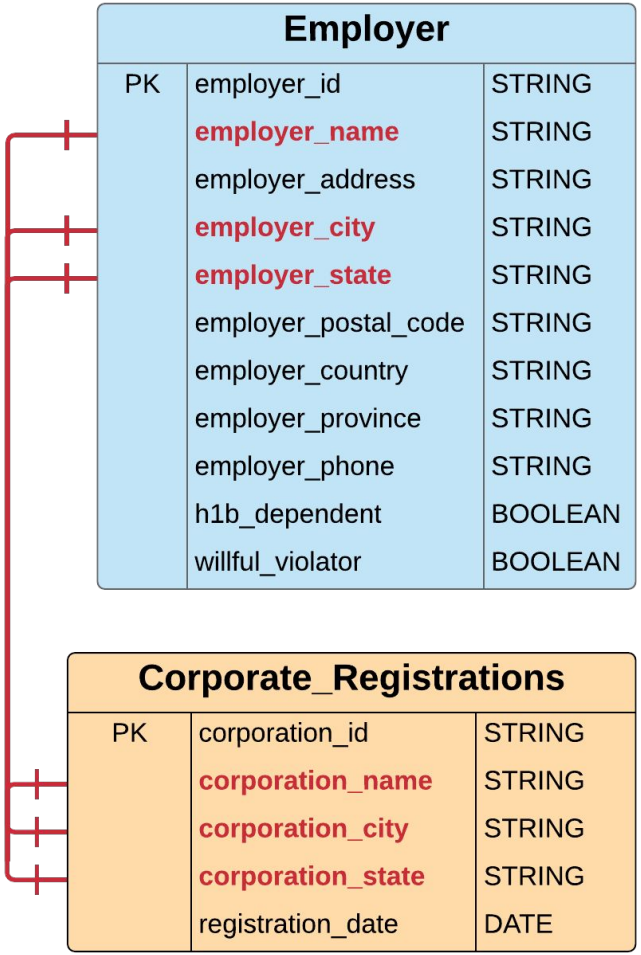


Data Integration Challenge

```
SELECT ...

FROM Source1.Account as A1 JOIN Source2.Account as A2

ON A1.c1 = A2.c1 AND A1.c2 = A2.c2

...
```

H1B Analytics Example






```
SELECT employer_name, registration_date
FROM Employer
JOIN Corporate_Registrations
on employer_name = corporation_name
and employer_city = corporation_city
and employer_state = corporation_state
```

Results:

- 2% matches between Employer and Corporate_Registrations
- Punctuation characters in corporation_name and corporation_city
- Suffixes in corporation_name (e.g. LLC, INC)

Creating the data pipeline for `dataset2`

1. Upload `dataset2` files to Cloud Storage bucket
2. Create staging area in BigQuery
3. Load data files into BigQuery as staging tables
4. Create modeled area in BigQuery
5. Identify Entity Types and create modeled tables
6. Identify relationships between tables
7. Identify Primary and Foreign Keys

Same steps as `dataset1`, except using a Jupyter Notebook.

Jupyter Notebooks

- Project Jupyter is open-source software
- Widely used for developing data science projects
- A web-based environment for creating notebooks
- Integrates code and its output into a single document, saved in `.ipynb` file
- Notebook is made up of cells
- Cell: block of code to be executed or container for text to be displayed
- Two types of cells: Code and Markdown
- Kernel: computation engine that executes the code in a notebook

Jupyter Notebook Demo

Milestone 9

http://www.cs.utexas.edu/~scohen/milestones/Milestone9.pdf