

# CS 327E Class 9

April 13, 2020

# Announcements

- Exam Experience and Grading
  - Challenging exam + issues with Canvas
  - Dropped lowest score of the 3 parts (TF, MC, Coding)
  - Offering extra credit worth 10% of final grade
- Extra Credit Project:
  - Analyze two or more [COVID-related datasets](#) in BQ
  - Visualize results in Data Studio or [BQ Geo Viz](#)
  - Write [Medium article](#) with your findings (include code snippets + visualizations)
  - Individual assignment, request private repo from instructors
  - Due May 10th through Canvas

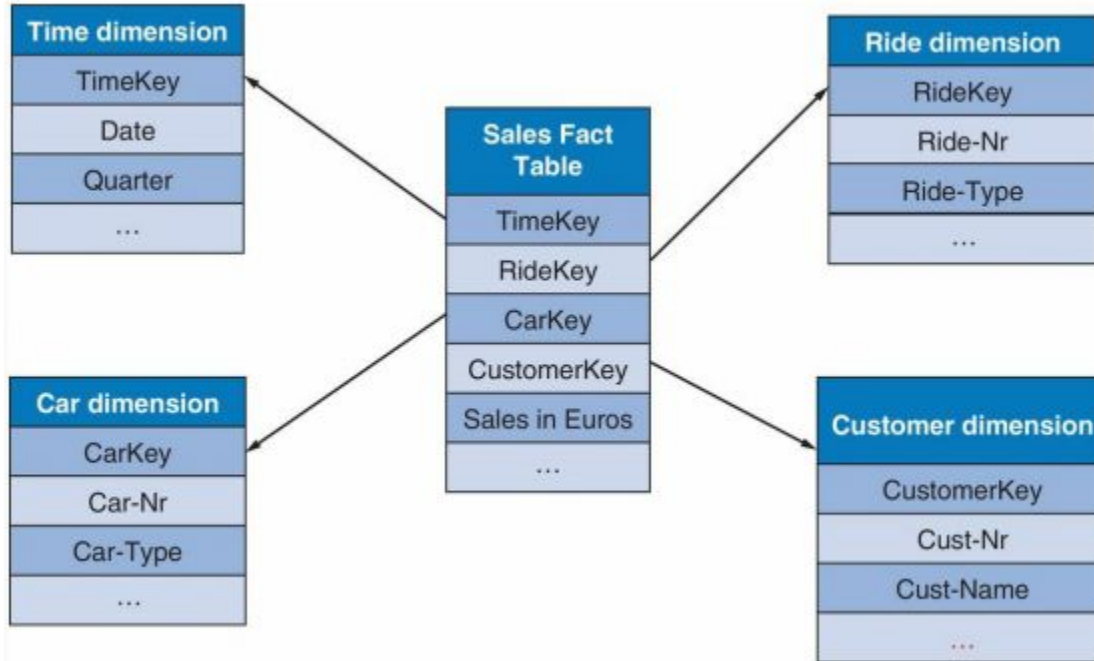
# Announcements

- Remaining Project Milestones:
  - **Milestone 9:** Secondary dataset ingestion and modeling pipeline
  - **Milestone 10:** Beam/SQL transforms + cross-dataset queries
  - **Milestone 11:** Workflow automation
  - **Milestone 12:** Demos and Presentations

1) A data warehouse is a specialized database which \_\_\_\_\_

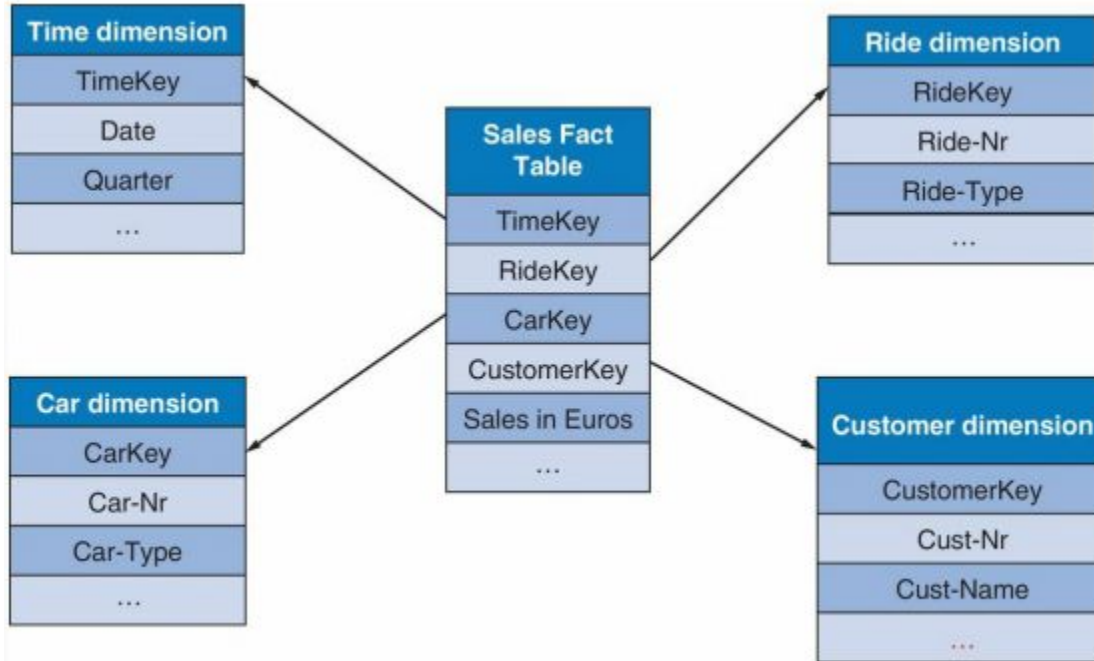
- A. integrates data from multiple different sources.
- B. processes a high volume of transactions per second.
- C. uses a 3NF schema.

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



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

3) In this Saber data warehouse schema, which column(s) form(s) the PK of the Sales Fact table?



- A. TimeKey
- B. RideKey
- C. CarKey
- D. CustomerKey
- E. All of the above

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

- A. Grain of the Fact table(s)
- B. Identifying the Dimension tables
- C. Slowly changing dimensions
- D. All of the above

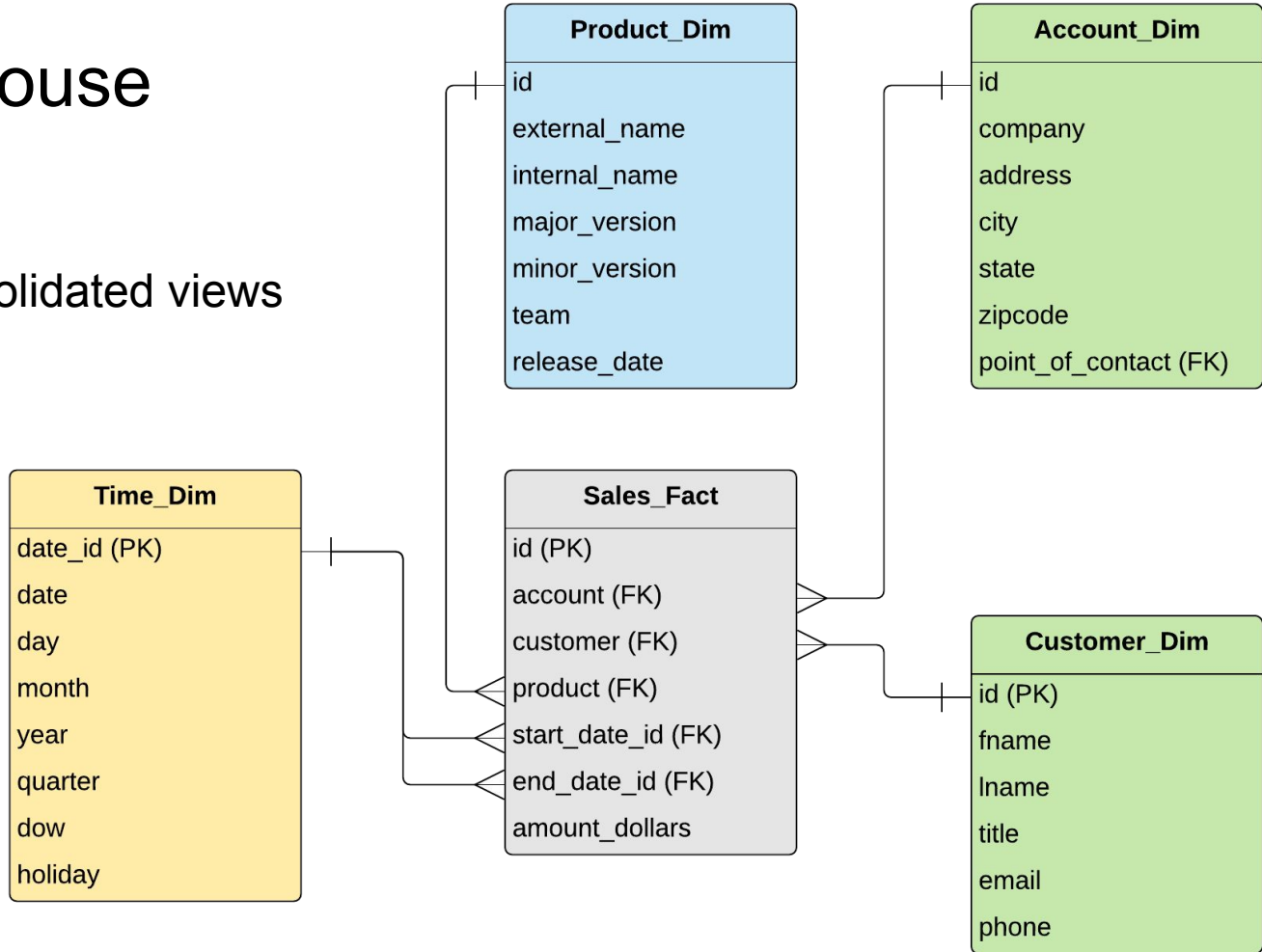
5) What activity can consume 80% of the time when building a data warehouse?

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



# Data Warehouse Challenges

- Creating consolidated views
- ETL pipelines

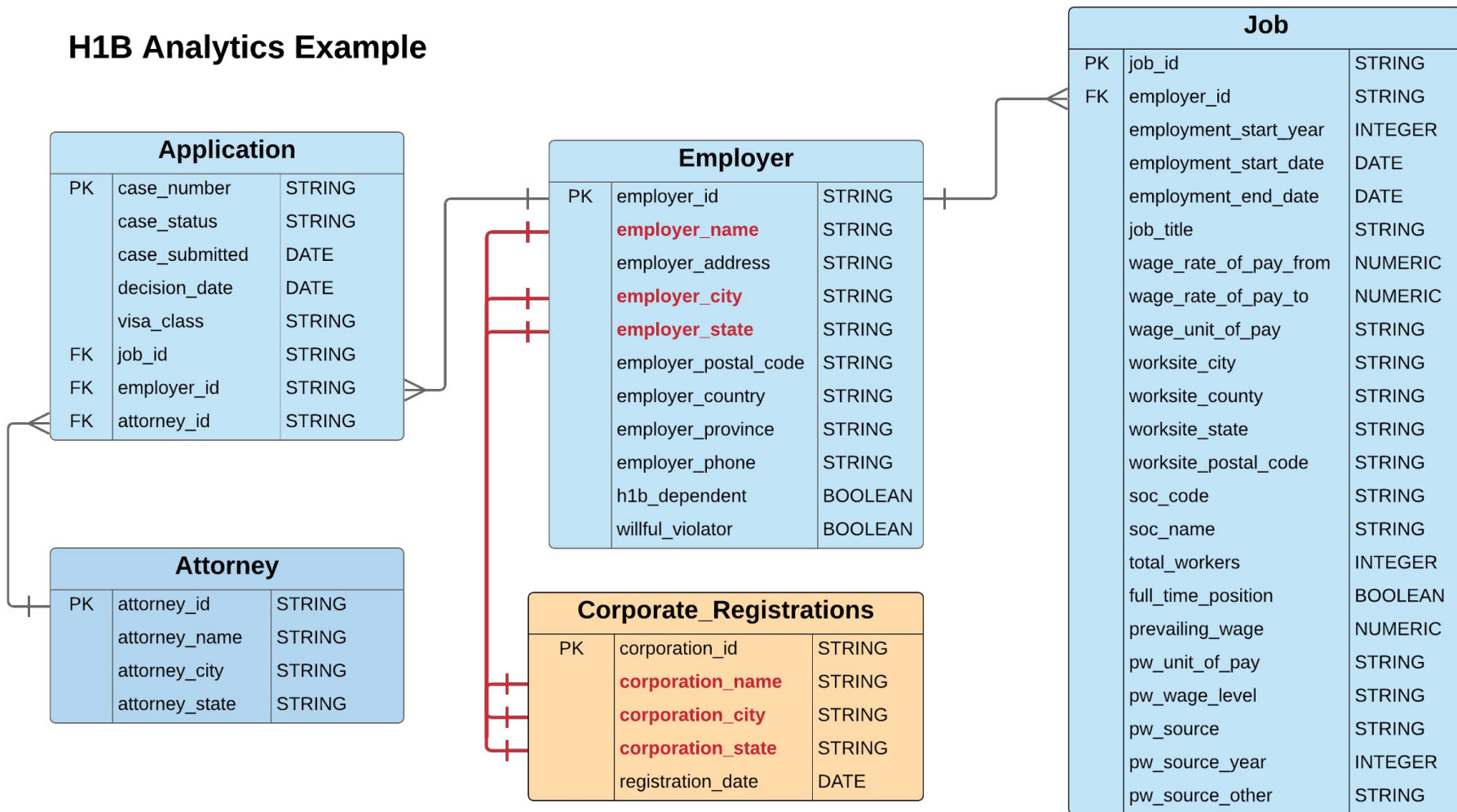


# Data Integration Patterns

## 1. Joining Independent Datasets:

```
SELECT a.foo, b.bar  
  
FROM Dataset_A.Table1 a  
  
JOIN Dataset_B.Table2 b  
  
ON a.foo = b.bar  
  
[WHERE ...]
```

# H1B Analytics Example



Application		
PK	case_number	STRING
	case_status	STRING
	case_submitted	DATE
	decision_date	DATE
	visa_class	STRING
FK	job_id	STRING
FK	employer_id	STRING
FK	attorney_id	STRING

Attorney		
PK	attorney_id	STRING
	attorney_name	STRING
	attorney_city	STRING
	attorney_state	STRING

Employer		
PK	employer_id	STRING
	employer_name	STRING
	employer_address	STRING
	employer_city	STRING
	employer_state	STRING
	employer_postal_code	STRING
	employer_country	STRING
	employer_province	STRING
	employer_phone	STRING
	h1b_dependent	BOOLEAN
	willful_violator	BOOLEAN

Corporate Registrations		
PK	corporation_id	STRING
	corporation_name	STRING
	corporation_city	STRING
	corporation_state	STRING
	registration_date	DATE

Job		
PK	job_id	STRING
FK	employer_id	STRING
	employment_start_year	INTEGER
	employment_start_date	DATE
	employment_end_date	DATE
	job_title	STRING
	wage_rate_of_pay_from	NUMERIC
	wage_rate_of_pay_to	NUMERIC
	wage_unit_of_pay	STRING
	worksite_city	STRING
	worksite_county	STRING
	worksite_state	STRING
	worksite_postal_code	STRING
	soc_code	STRING
	soc_name	STRING
	total_workers	INTEGER
	full_time_position	BOOLEAN
	prevailing_wage	NUMERIC
	pw_unit_of_pay	STRING
	pw_wage_level	STRING
	pw_source	STRING
	pw_source_year	INTEGER
	pw_source_other	STRING

Employer		
PK	employer_id	STRING
	<b>employer_name</b>	STRING
	employer_address	STRING
	<b>employer_city</b>	STRING
	<b>employer_state</b>	STRING
	employer_postal_code	STRING
	employer_country	STRING
	employer_province	STRING
	employer_phone	STRING
	h1b_dependent	BOOLEAN
	willful_violator	BOOLEAN

Corporate_Registrations		
PK	corporation_id	STRING
	<b>corporation_name</b>	STRING
	<b>corporation_city</b>	STRING
	<b>corporation_state</b>	STRING
	registration_date	DATE

```
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
```

### Engineering Tasks:

- Punctuation characters in join fields (e.g. corporation\_name, corporation\_city)
- Suffixes in corporation\_name (e.g. LLC, INC)
- Standardize join fields to improve matching accuracy

# Data Integration Patterns

## 2. Unioning Independent Datasets:

```
SELECT a, b, c  
  
FROM Dataset_A.Table1  
  
UNION DISTINCT  
  
SELECT x, y, z  
  
FROM Dataset_B.Table2
```

# Data Integration Patterns

## 2. Unioning Independent Datasets:

```
SELECT a, b, c  
  
FROM Dataset_A.Table1  
  
UNION ALL  
  
SELECT x, y, z  
  
FROM Dataset_B.Table2
```

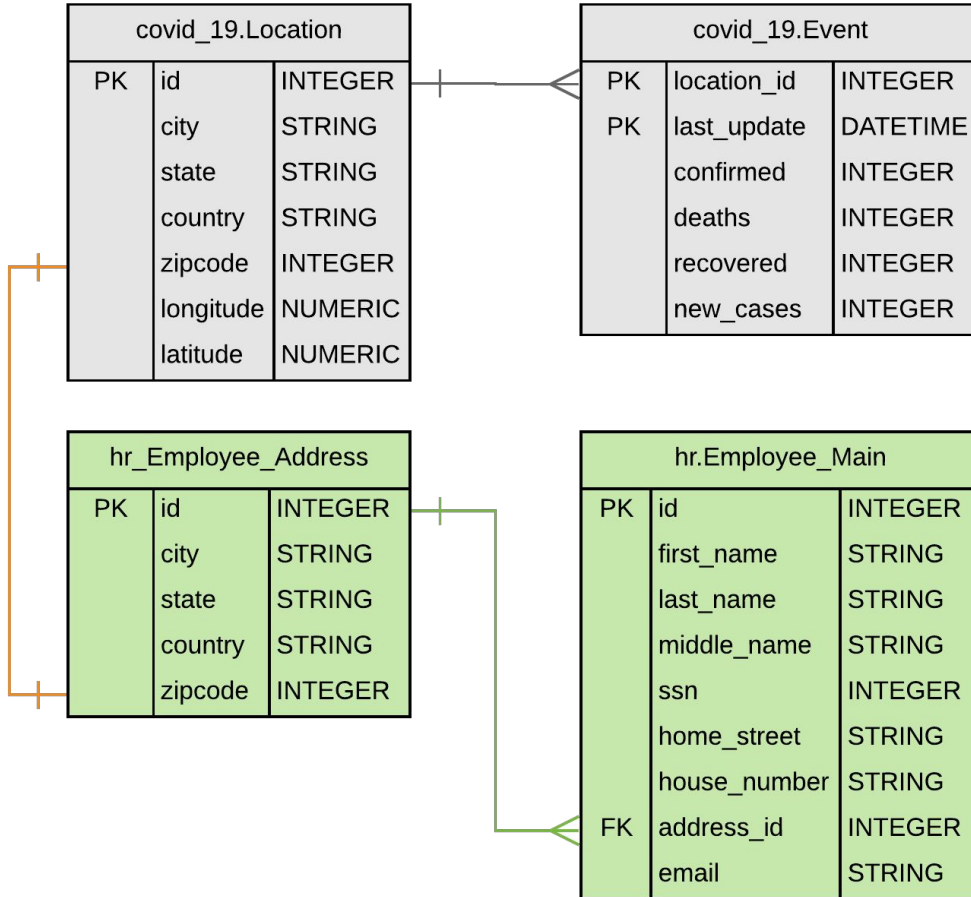
```
SELECT cand_name, office,  
        pty_affiliation,  
        SUM(cmte_amount) +  
        SUM(ind_amount) as amount  
FROM fec.Candidate ca  
JOIN fec.Contributions co  
ON ca.cand_id = co.cand_id  
GROUP BY cand_name, pty_affiliation  
UNION ALL  
SELECT cand_name_short,  
        seek_office, party,  
        SUM(contribution_amount) as  
        amount  
FROM tec.Cand ca  
JOIN tec.Contribs co  
ON ca.cand_id = co.cand_id  
GROUP BY cand_name, party
```

# Data Integration Patterns

## 3. Filtering on Independent Datasets:

```
SELECT a, b, c
FROM Dataset_A.Table1
[JOIN Dataset_A.Table2 ...]
WHERE d IN (SELECT x
            FROM Dataset_B.Table2 ...)
```

## COVID Employment Example



```
SELECT id, first_name, email
FROM hr.Employee_Main e
JOIN hr.Employee_Address a
ON e.address_id = a.id
WHERE zipcode IN
  (SELECT zipcode
   FROM covid_19.Location l
   JOIN covid_19.Event e
   ON l.id = e.location_id
   WHERE new_cases = 0)
ORDER by id
```

### Engineering Tasks:

- Obtain COVID data by zipcode
- Calculate new COVID cases



# Instructions for Partner Exercise

1. Go to sheet: <https://tinyurl.com/wldp9vr>
2. Search for **your group** in Column A
3. Start a Zoom meeting for your group
4. Add your Zoom meeting link to Column B next to your group
5. Go to your group's Zoom meeting

# Instructions for Partner Exercise

1. With your partner, **agree** on a secondary dataset
2. Describe your secondary dataset in `DATASETS.txt`
3. Decide how to **combine** your primary + secondary datasets
4. Go back to the Sheet and update Column C as **Done**
5. Wait for an instructor to join **your Zoom**
6. While you wait, read and discuss Milestone 9 with your partner
7. Review your plan with the instructor when they join your Zoom

# Milestone 9

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