Data Warehouse Design

“Employee Domain”

Star Schema

Hijir Della Wirasti

17 November 2024

https://github.com/hijirdella/Data-Warehouse-Design.git
https://www.linkedin.com/in/hijirdella/

Table of contents

01

Objectives of the project

Objectives

Objective:

1. Design a Data Warehouse (DWH) schema for the selected domain.
2. Create an Entity-Relationship Diagram (ERD).
3. Describe the schema as either a Star Schema.
4. Provide three sample queries for Data Mart tables.

The goal

1. ERD Diagram
2. Star Schema Description
3. Sample Queries (Data Mart Tables)

02

Dataset Selection

Dataset Selection

```
1  SELECT * FROM public.dim_employee
2  ORDER BY employee_id ASC
```

Data Output | Messages | Notifications

	employee_id [PK] integer	first_name character varying (50)	last_name character varying (50)	birt_date date	hire_date date	gender character varying (6)	department_id integer	education_id integer
1	1	Ignazio	Wharby	1994-11-25	2022-08-29	Female	1	4
2	2	Oralla	Minnette	1999-09-30	2024-02-05	Male	6	3
3	3	Uriah	Plowell	1994-02-24	2022-03-09	Male	7	1
4	4	Jecho	Laraway	1991-04-29	2022-07-14	Male	6	5
5	5	Nettle	Kindleysides	1992-10-12	2023-07-27	Male	8	1
6	6	Emmie	Simper	1995-10-10	2023-12-18	Male	4	5
7	7	Annamaria	Camden	1998-06-22	2023-02-03	Male	7	4
8	8	Hashim	Raper	1997-01-05	2022-03-20	Female	3	4
9	9	Ezequiel	Speedy	1996-11-27	2023-06-23	Male	10	1
10	10	Bernard	Whatsize	1994-04-25	2023-09-14	Male	6	5
11	11	Neils	Boland	1992-04-11	2022-06-08	Male	2	3
12	12	Rudiger	Warlow	1999-11-19	2022-09-17	Male	8	1
13	13	Allistir	Byron	1995-08-10	2023-08-14	Female	7	4
14	14	Paquito	Naisey	1998-08-17	2023-12-05	Male	3	3
15	15	Winne	Mellei	1999-05-16	2023-09-28	Male	6	4

Total rows: 1000 of 1000 | Query complete 00:00:00.459

Selected Dataset:

Employee SQL Dataset

Domain Overview:

Employee records for HR analytics

Dataset Selection

```
1  SELECT * FROM public.dim_department
2  ORDER BY department_id ASC
```

Data Output | Messages | Notifications

	department_id [PK] integer	department_name character varying (100)
1	1	HR
2	2	Finance
3	3	Marketing
4	4	IT
5	5	Operations
6	6	Sales
7	7	Legal
8	8	Research and Development
9	9	Customer Service
10	10	Administration

```
1  SELECT * FROM public.dim_education
2  ORDER BY education_id ASC
```

Data Output | Messages | Notifications

	education_id [PK] integer	education_level character varying (50)
1	1	High School
2	2	Associate Degree
3	3	Bachelor's Degree
4	4	Master's Degree
5	5	PhD

```
1  SELECT * FROM public.fact_employee_performance
2  ORDER BY performance_id ASC
```

Data Output | Messages | Notifications

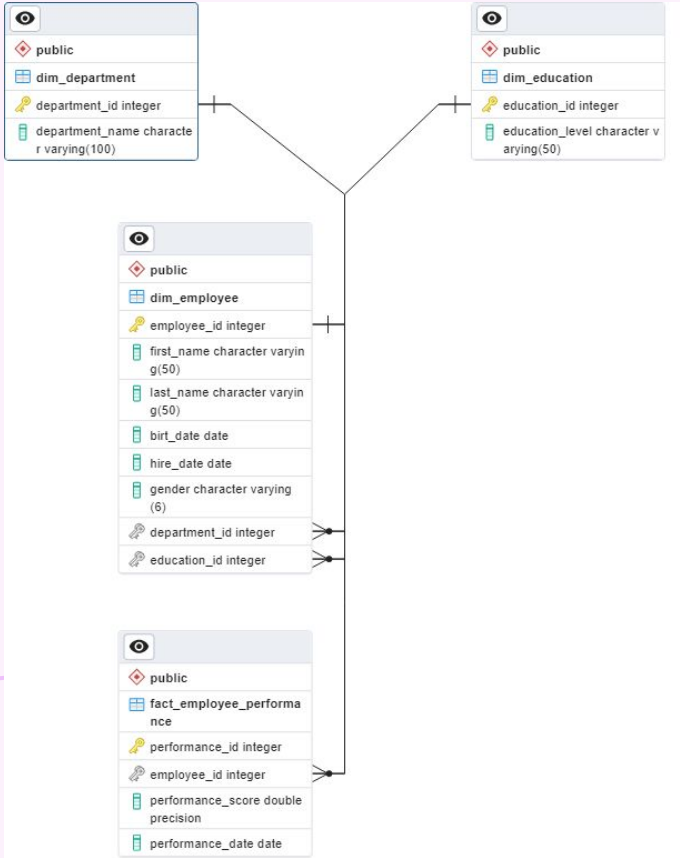
	performance_id [PK] integer	employee_id integer	performance_score double precision	performance_date date
1	1	1	36.4784694903924	2022-10-13
2	2	2	38.40222708316501	2020-07-31
3	3	3	61.0410367577725	2022-06-12
4	4	4	15.068316886181066	2020-02-22
5	5	5	54.08936348060487	2020-02-11
6	6	6	80.67333418394043	2022-03-24
7	7	7	45.78888591104575	2022-04-21
8	8	8	48.15436852138222	2022-08-28
9	9	9	94.05782915152795	2022-03-25
10	10	10	95.32636516472932	2021-07-26
11	11	11	83.25610647569377	2020-11-26
12	12	12	28.01159941513005	2021-06-15
13	13	13	45.386553647844565	2020-01-09
14	14	14	53.4976948478143	2020-12-27

Total rows: 1000 of 1000 | Query complete 00:00:00.317

03

ERD
Diagram

ERD Diagram (1)



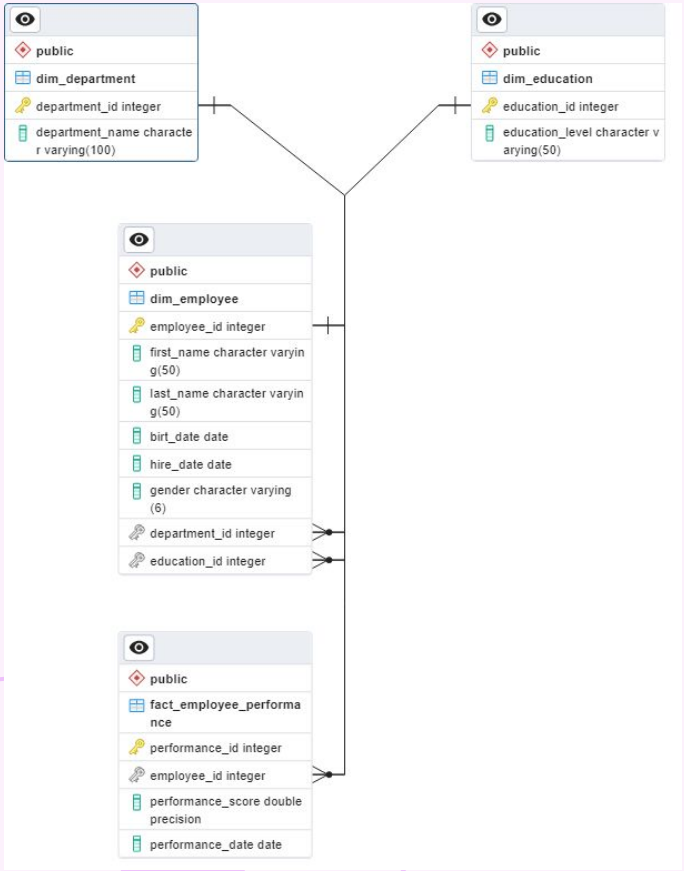
The ERD (Entity-Relationship Diagram) represents a **star schema** design for an employee performance data warehouse. It consists of one fact table and three dimension tables, structured as follows:

Fact Table

1. **fact_employee_performance**
   - **performance_id** (Primary Key): Unique identifier for each performance record.
   - **employee_id** (Foreign Key): Links to the dim_employee table, representing the employee being evaluated.
   - **performance_score**: Numerical score measuring the employee's performance.
   - **performance_date**: Date of the performance evaluation.

This table stores the core business metrics and links to the dimension tables for detailed context.

ERD Diagram (2)



Dimension Tables

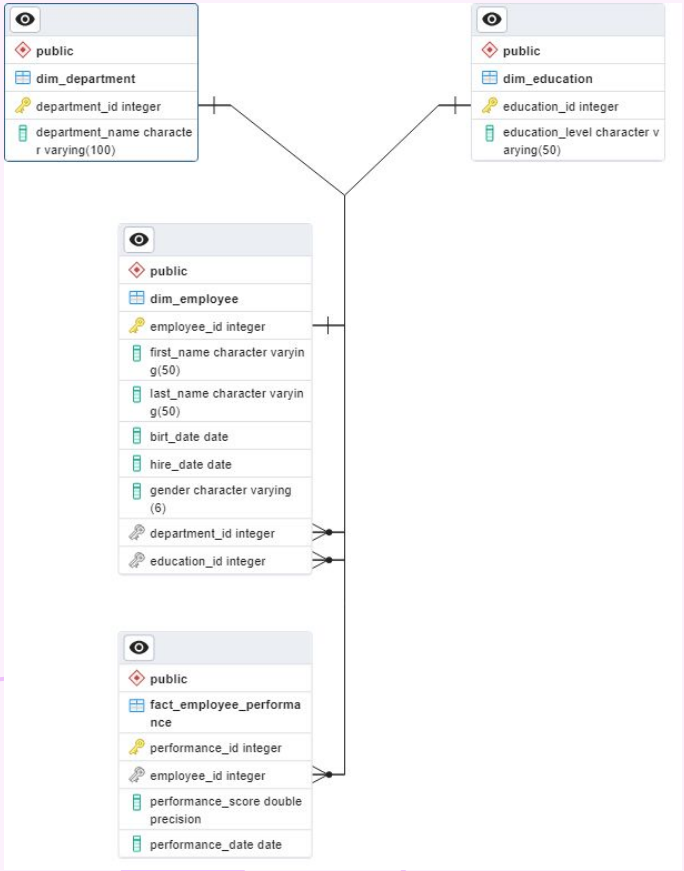
1. **dim_department**
   - **department_id** (Primary Key): Unique identifier for each department.
   - **department_name**: Name of the department (e.g., HR, IT, Sales).

   Provides descriptive information about departments employees are associated with.

2. **dim_education**
   - **education_id** (Primary Key): Unique identifier for each education level.
   - **education_level**: Describes the education qualification (e.g., Bachelor's Degree, Master's Degree).

   Contains data on employee education levels.

ERD Diagram (3)

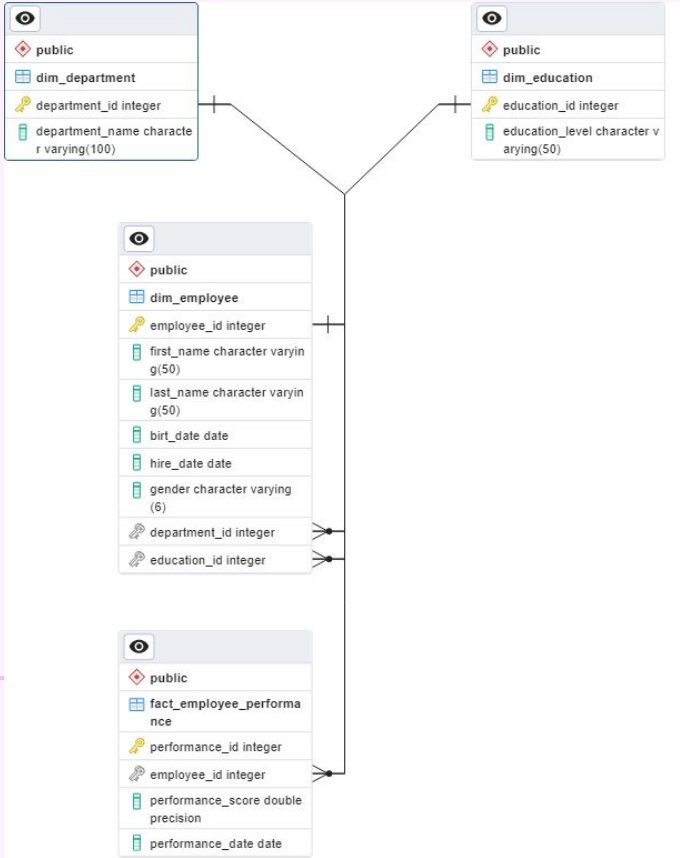


dim_employee

- **employee_id** (Primary Key): Unique identifier for each employee.
- **first_name** and **last_name**: Basic personal information.
- **birt_date**: Date of birth.
- **hire_date**: Date the employee was hired.
- **gender**: Gender of the employee.
- **department_id** (Foreign Key): Links to the dim_department table.
- **education_id** (Foreign Key): Links to the dim_education table.

Provides detailed information about employees, including their demographic details, department, and education.

ERD Diagram (4)



Relationships

- The **fact_employee_performance** table connects to the **dim_employee** table using employee_id, enabling analysis of employee performance scores.
- The **dim_employee** table connects to:
  - **dim_department** through department_id to relate employees to their departments.
  - **dim_education** through education_id to associate employees with their education qualifications.

This schema structure allows analytical queries to easily combine performance metrics with employee details, department context, and educational background, making it ideal for reporting and analysis.

04

Schema Description

Schema Description

Schema Description

The schema is designed as a star schema to facilitate efficient analytical queries and reporting for employee performance data.

Relationships

1. **fact_employee_performance → dim_employee**: Linked via employee_id, providing employee-specific context to performance metrics.
2. **dim_employee → dim_department:** Linked via department_id, enabling grouping or filtering performance data by department.
3. **dim_employee → dim_education:** Linked via education_id, allowing performance comparisons based on education levels.

Purpose of the Schema

This schema enables comprehensive analysis of employee performance by connecting it to demographic, departmental, and educational contexts. It is optimized for:

1. Aggregating performance metrics (e.g., average scores by department or education level).
2. Tracking trends over time (via performance_date).
3. Cross-dimensional reporting to identify patterns and insights for strategic decision-making.

This star schema structure ensures simplicity, scalability, and ease of use for reporting and analysis.

05

Data Mart

Data Mart
1. Performance by Department

```sql
SELECT *
FROM data_mart.dm_performance_by_department
ORDER BY avg_performance_score DESC;
```

Data Output | Messages | Notifications

	department_id integer	department_name character varying (100)	total_performance_records bigint	avg_performance_score double precision	max_performance_score double precision	min_performance_score double precision
1	10	Administration	93	53.89606946797804	98.53116992937917	1.1635648925892195
2	5	Operations	102	51.323257590568915	98.40633043792111	0.40534835236372135
3	2	Finance	95	50.706955256809906	99.99345392794964	0.15092794211510885
4	3	Marketing	87	50.18430453138673	99.14297853905703	1.7370520309907267
5	4	IT	73	50.10862996994088	97.95346518123212	4.220286700932396
6	6	Sales	117	49.45306361044393	99.128372523017	0.45168581411014763
7	9	Customer Service	111	49.034953497140314	99.95289276267422	0.039422602567884546
8	7	Legal	109	47.764074868224895	98.54524518444978	1.71110360014457
9	1	HR	104	47.00113417872205	99.47554551437605	0.5512051713364441
10	8	Research and Development	109	46.75672340014453	98.3939999617546	0.40690617173697596

Data Mart
1. Performance by Department

Performance Analysis Summary:

- **Top Performers**: *Administration* (53.89 avg, 98.53 max) and *Operations* (51.32 avg, 98.40 max) exhibit consistent high performance, setting benchmarks for other departments.
- **Mid Performers**: *Finance* (50.70), *Marketing* (50.18), and *IT* (50.10) maintain steady averages, reflecting reliable performance.
- **Low Performers**: *R&D* (46.75 avg), *Customer Service*, and *Legal* demonstrate lower averages, with wide gaps between min and max scores, signaling potential issues.
- **High Variance Across Departments**: Departments like *Legal* and *Customer Service* show significant differences between max and min scores, suggesting diverse performance levels. Conversely, *Administration* and *Operations* have more consistent scores.
- **Insights**: Improve underperforming departments through training or team restructuring, investigate causes of variance in Legal and Customer Service, and leverage top-performing departments as models for success.

Data Mart
2. Performance by Education

```sql
SELECT *
FROM data_mart.dm_performance_by_education
ORDER BY avg_performance_score DESC;
```

Data Output | Messages | Notifications

	education_id integer	education_level character varying (50)	total_performance_records bigint	avg_performance_score double precision	max_performance_score double precision	min_performance_score double precision
1	2	Associate Degree	214	51.57069875349984	99.48211848805857	0.5512051713364441
2	4	Master's Degree	193	49.86262604346513	99.99345392794964	0.039422602567884546
3	1	High School	194	49.68513783950901	99.47554551437605	0.15092794211510885
4	3	Bachelor's Degree	209	49.228151487257485	99.128372523017	0.40690617173697596
5	5	PhD	190	46.95257079630433	98.73801632977144	0.24081114864653586

Data Mart
2. Performance by Education

Performance Analysis Summary:

- **Top Performers**:
  a. **Associate Degree** has the highest average score (51.57), reflecting consistent performance.
  b. **Master's Degree** follows with an average of 49.86 and the highest max score (99.99), showing exceptional individual potential.
- **Mid-Level Performers**:
  a. **High School** graduates (49.68) outperform Bachelor's (49.22), showcasing strong capabilities despite lower education levels.
- **Low Performers**:
  a. **PhD** holders have the lowest average (46.95), potentially due to role mismatch or overqualification.
- **High Variance**:
  a. PhD and Bachelor's show wider score ranges, suggesting performance inconsistencies.

Recommendations:

- Leverage Associate and Master's Degree employees for consistent results.
- Investigate performance gaps for PhD holders and align roles better.
- Target training for Bachelor's and High School groups to reduce variability.

Data Mart
3. Performance by Gender (1)

```
1  SELECT * FROM data_mart.dm_performance_by_gender
```

Data Output | Messages | Notifications

	gender character varying (6)	total_performance_records bigint	avg_performance_score double precision	max_performance_score double precision	min_performance_score double precision
1	Female	271	49.33558038574339	99.47554551437605	0.45168581411014763
2	Male	729	49.57237713045363	99.99345392794964	0.039422602567884546

Insights from Gender-Based Performance Data:

1. **Performance Comparison**:
   - Males have a slightly higher average performance score (49.57) than females (49.33), indicating comparable performance across genders with minor differences.
2. **Max and Min Scores**:
   - Males achieved the highest maximum performance score (99.99), while females reached 99.47, showing high-performing individuals in both groups.
   - Females have a higher minimum score (0.45) compared to males (0.03), suggesting fewer outliers or underperformers in the female group.

Data Mart
3. Performance by Gender (2)

3. Participation:

a. Males contribute significantly more to total performance records (729 vs. 271), possibly reflecting a larger workforce or more evaluated roles.

Recommendations:

- Ensure equal growth opportunities and address participation gaps to enhance diversity.
- Investigate factors contributing to the narrower score range in females for potential best practices.

Data Mart
4. Performance Trends Over Time

```
127  --Performance Trends Over Time
128  SELECT *
129  FROM data_mart.dm_performance_trends
130  ORDER BY performance_month;
```

Data Output | Messages | Notifications

	performance_month timestamp with time zone	total_performance_records bigint	avg_performance_score double precision	max_performance_score double precision	min_performance_score double precision
1	2020-01-01 00:00:00+07	28	44.052530019505234	98.40633043792111	6.147366379212982
2	2020-02-01 00:00:00+07	24	46.01416770354516	97.60180611491944	0.15092794211510885
3	2020-03-01 00:00:00+07	30	51.41198085569371	94.7649855534654	0.5512051713364441
4	2020-04-01 00:00:00+07	25	47.458935590209734	94.38668820188005	4.700922114065942
5	2020-05-01 00:00:00+07	27	42.62559169162599	88.39794289375895	4.195714017058982
6	2020-06-01 00:00:00+07	31	54.02221764790315	96.32558183269526	2.556657798173201
7	2020-07-01 00:00:00+07	30	45.269302793283984	97.07536915010795	6.962532627002149
8	2020-08-01 00:00:00+07	40	52.554763701350886	98.3939999617546	0.40690617173697596
9	2020-09-01 00:00:00+07	43	54.86500964135037	96.0307629090104	2.909867588641335
10	2020-10-01 00:00:00+07	18	40.26845274843077	98.34831252250262	0.9858322954326759
11	2020-11-01 00:00:00+07	21	51.524758209309915	86.62816016205352	15.455561539278007
12	2020-12-01 00:00:00+07	31	57.61032734755107	99.128372523017	2.2629230016597335

Total rows: 36 of 36 | Query complete 00:00:00.175 | Ln

Data Mart
4. Performance Trends Over Time

Insights:

1. **Performance Fluctuations:** Average scores vary, peaking in October 2020 (54.86) and dipping in April 2020 (42.62), indicating seasonal or workload impacts.
2. **Consistent Top Scores:** Maximum scores stay near 99, showing strong top performers throughout.
3. **Low Scores:** Minimum scores highlight occasional underperformance, requiring targeted interventions.
4. **Growing Records:** An increase in performance records suggests expanding workforce or improved tracking.

Focus on low-performance periods for improvements and replicate strategies from high-performing months.

06

Conclusion

Conclusion on Schema and Data Mart Design

The schema follows a **star schema** structure, enabling efficient querying and data aggregation. The **dimensional tables** (e.g., dim_department, dim_education, dim_employee) provide descriptive attributes, while the **fact table** (fact_employee_performance) captures measurable performance metrics, establishing clear relationships.

Key Benefits of the Design:

1. **Optimized Query Performance**:
    - The star schema minimizes joins by centralizing performance metrics in the fact table, making aggregations straightforward.
2. **Scalability and Flexibility**:
    - Additional dimensions (e.g., `dim_project`, `dim_region`) can be added without disrupting the schema's core structure.
3. **Clear Analytical Objectives**:
    - The schema focuses on employee performance, allowing targeted analysis across departments, education levels, and gender.

Thanks!

Do you have any questions?

hijirdw@gmail.com
https://github.com/hijirdella
https://www.linkedin.com/in/hijirdella/