University of Essex

Department of Computing

Deciphering Big Data March 2024

Development Team Project: Project Report

Date: 21/Apr/2024

1. Introduction

The aim of this project is to design a single logical database to organize and optimize the hiring process for the HR department of an organization. Without a robust data system, managing the data of multiple applicants becomes arduous and costly. This report will produce a proposal for the database build, cover the logical design of the proposed database, and critically evaluate the data management pipeline process.

2. Proposal of the Database Build and Logical Flow

After careful consideration of factors such as access rights, storage, and integration with extendable tools, Oracle was chosen to be the database solution (Nguyen S., 2023). Oracle is renowned for its seamless integration with ETL tools like Oracle Data Integrator for data transformation and cleaning, as well as visualization tools like Oracle Business Intelligence. This choice enhances security and reduces complexity, supporting organizational expansion (Oracle, 2021).

Creating a database involves considering the path that data follows within the system. Figure 1 illustrates the anticipated logical data flow within the hiring system, across four dimensions, and outlines access rights as follows:

- HR staff will have access to the “Candidate Dimension”, facilitating efficient management of applicant data.
- Management-level staff in each department will be able to access the “Employee Dimension”, but only for data pertaining to their own department staff, ensuring row-level security (Huey P., 2017).
- All staff will have read access to the “Department Dimension”, while only HR staff will be authorized to have write access to edit it.

- Management-level staff will have the authority to create positions in the “Job Openings Dimension’, which can then be reviewed, commented on, and then potentially modified by HR personnel, fostering collaborative decision-making and efficient resource allocation.

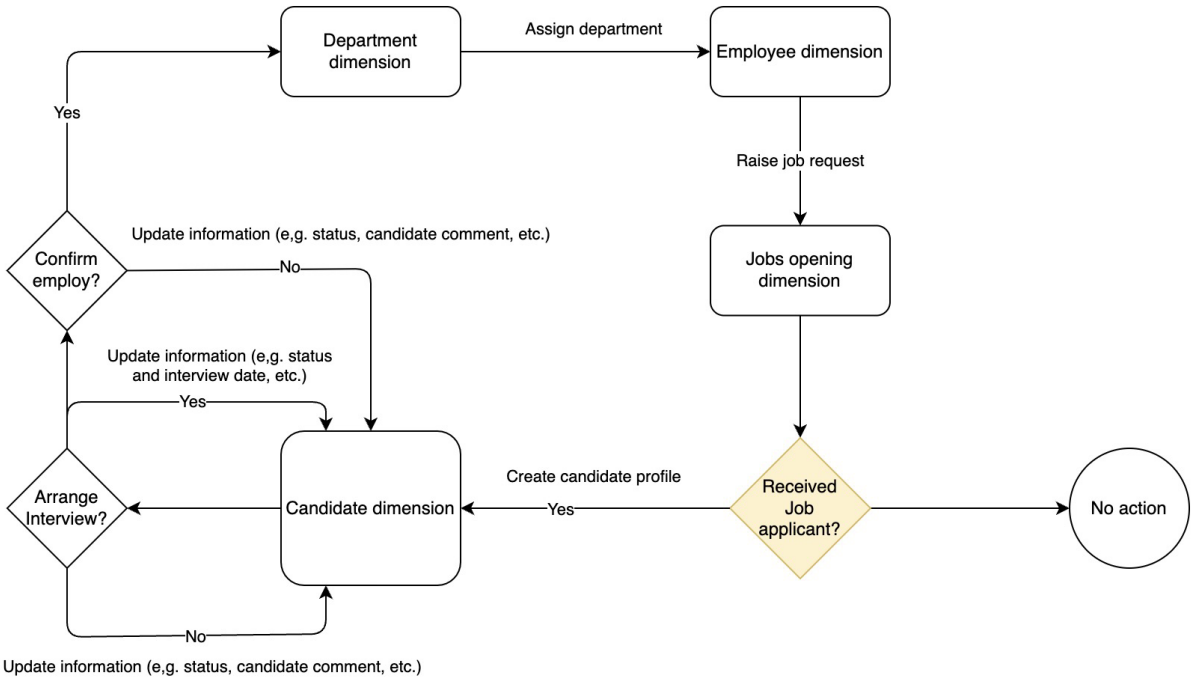

Figure 1. Logical Flow of the hiring system and its dimensions.

3. The Design of the DBMS and Data Management Process

This section delves into the design of the database, expanding its scope to accommodate essential dimensions and exploring its structure and implications for effective data management.

3.1 The Design of the Database

Rather than create a database purely for candidates, this design accommodates the tracking of candidates throughout the hiring process, along with existing employees and their data. Shown below is Figure 2, which illustrates the potential inputs for each dimension, along with their attributes, data types, and characteristics.

Dimensions												
CANDIDATE		EMPLOYEE		JOB_OPENING		DEPARTMENT		EE_LEVEL		EE_TYPE		
Attritube	Data Type	Attritube	Data Type	Attritube	Data Type	Attritube	Data Type	Attritube	Data Type	Attritube	Data Type	
CAN_ID	NUMERIC	EE_ID	NUMERIC	JOB_ID	NUMERIC	DEPART_ID	NUMERIC	EL_ID	NUMERIC	EE_TYPE_ID	NUMERIC	Primary Key
JOB_ID	NUMERIC	DEPART_ID	NUMERIC	EE_ID	NUMERIC							Foreign Key
		CAN_ID	NUMERIC	DEPART_ID	NUMERIC							
		EL_ID	NUMERIC	EE_TYPE_ID	NUMERIC							
		EE_TYPE_ID	NUMERIC	EL_ID	NUMERIC							
CAN_FIRST	VARCHAR	EE_FIRST	VARCHAR	POSITION	VARCHAR	DEPART_CODE	VARCHAR	EE_LEVEL	VARCHAR	EE_TYPE	VARCHAR	Dimension content
CAN_MID	VARCHAR	EE_MID	VARCHAR	DESC	VARCHAR	DEPARTMENT	VARCHAR					
CAN_LAST	VARCHAR	EE_LAST	VARCHAR	MAX_SAL	NUMERIC							
CAN_NICK	VARCHAR	NICK_NAME	VARCHAR									
CAN_TEL	NUMERIC	EE_IC	VARCHAR									
CAN_MAIL	VARCHAR	COM_TEL	NUMERIC									
DOB	DATE	COM_MAIL	VARCHAR									
AGE	INTEGER	POSITION	VARCHAR									
EXPECT_SAL	NUMERIC	SALARY	NUMERIC									
AVAILABILITY	VARCHAR											
FRESH_GRAD	VARCHAR											
REL_EXP_MON	NUMERIC											
WORK_EXP_MON	NUMERIC											
EDU_LEVEL	VARCHAR											
CAN_DESC	VARCHAR											
APP_DATE	DATE											
IN_DATE	DATE											
STATUS	VARCHAR											
START_DATE	DATE	START_DATE	DATE	START_DATE	DATE	START_DATE	DATE					Record timestamp
END_DATE	DATE	END_DATE	DATE	END_DATE	DATE	END_DATE	DATE					
LAST_UPDATE	DATE	LAST_UPDATE	DATE	LAST_UPDATE	DATE	LAST_UPDATE	DATE					

Figure 2. Dimension's structure, attributes and data types.

In section 2, the logical flow of the database was introduced, establishing four core dimensions: “CANDIDATE”, “EMPLOYEE”, “JOB_OPENING”, and “DEPARTMENT”. However, the design of the database has been expanded to incorporate six dimensions by including the "EE_LEVEL" and "EE_TYPE" dimensions. These enhancements are designed to improve access control and enhance the richness of the data. In addition, the table data types are categorized into four groups:

- The “Primary keys” and “Foreign Keys” groups are both numeric, ensuring their uniqueness and consistency.
- The “Dimension Content” group, which includes a mix of data types such as VARCHAR, NUMERIC, and DATE that represent row level data meaning.
- The “Record Timestamp” group, which helps uniquely identify records using the Type 2 Slow Changing Dimension method (Oracle, N.D.). This is particularly useful for tracking changes over time, such as when a candidate transitions to an employee in two distinct periods.

Dimension	Table	Usage
Candidates	CANDIDATE	Data and status of candidate
Employees	EMPLOYEE	Data and status of employee
Job Opening	JOB_OPENING	Available job opening of the organization
Department	DEPARTMENT	All departments of the organization
Employee level	EE_LEVEL	The level of employee, e.g. Entry, Senior, etc…
Employee Type	EE_TYPE	The type of employee, e.g.. Contractor, Perm, Part-Time, etc…

Table 1. Dimension’s implications.

3.2 The Data Management Pipeline

Figure 3 outlines the process by which employees will input data into the database. When new information about an employee or candidate needs to be added to the database, the relevant information will be manually input by the employees of the organization. This ensures that the database remains accurate and compliant with the organization's standards and protocols.

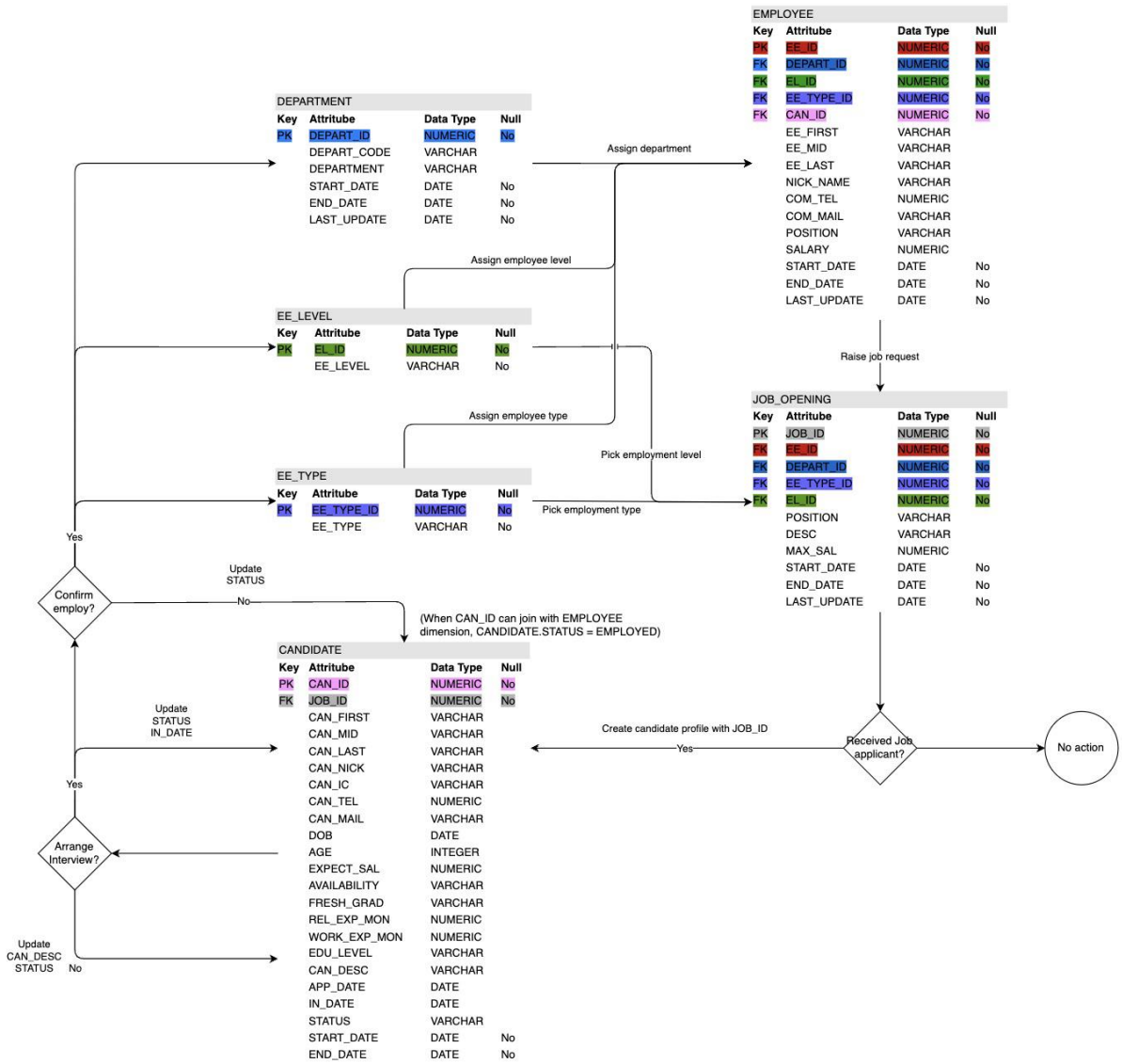

Figure 3. DBMS Design with attributes.

When processing and handling data for the database, it is essential to implement a data management pipeline. This pipeline consists of two crucial components: data normalization and data transformation.

3.2.1 Data normalization

When handling data with diverse sources and formats, like unstructured or semi-structured, it is imperative to implement normalization procedures to minimize data redundancy. These processes include:

- Reviewing metadata implications and ensuring data is atomic.
- Adding unique identifiers to prevent repetition of rows or columns.
- Ensure each field is uniquely named and that all entries in the data cells are single-valued (Romansanta J., 2020).
- Eliminating partial dependencies, ensuring all attributes in each row are dependent on their primary key.
- Eliminating transitive dependencies, which ensures that all attributes are in their most optimal arrangements in respect to the primary key.

Following these processes ensures that the data conforms to the first, second, and third normal forms.

3.2.2 Data transformation

When employees input new data into the database, the data collected may contain errors or inconsistencies. The team has considered some issues that may arise because of this manual input or human errors, such as leading or trailing spaces in strings, which are common in data collection. Data cleaning is a crucial step in the

data management pipeline when it comes to ensuring smooth database operation and accurate data representation. This involves:

- Data validation to ensure that nonsensical data - such as leading or trailing whitespace in VARCHAR or DATE fields - is correctly stored in the database.
- Handling missing data by converting whitespace-only values to NULL to avoid misleading data interpretation.
- Fixing structural errors by rectifying any structural inconsistencies in the data.
- Removing irrelevant and duplicate entries by eliminating data entries that are redundant or irrelevant to maintain database integrity.
- Data formatting, which is done by correcting unexpected symbols or formatting issues, such as in email addresses or phone numbers (Antkowiak M. & Nowaczyk M., 2021).

These data cleaning processes ensure that the database operates efficiently and that the data is in the correct format with the intended meaning (Antkowiak M. & Nowaczyk M., 2021).

4. Risk Assessment

Furthermore, effective risk management is pivotal for ensuring the success of the database implementation. Potential risks such as data loss, security breaches, and system outages have been identified (Noss, 2023). Implementing reliable backup procedures, strict security controls, and failover systems are some of the mitigating techniques (United IT Consultants, 2023). Through proactive mitigation of these risks, the objective is to guarantee a seamless deployment process, protect data integrity, and maintain business operations without interruptions.

5. Conclusion

This project's objective was to develop a logical database that streamlines and enhances the HR department's hiring process. This report covered the database build, the logical flow and design of the proposed database, and conducted a critical evaluation of the data management pipeline process. Implementing the proposed database could lead to significant time and resource savings for the organization while fostering organizational growth.

References:

Antkowiak M. & Nowaczyk M. (2021) Available from: https://medium.com/transparent-data-eng/data-cleansing-examples-24581c3d14f1 [Accessed 15 Apr 2024]

Huey P. (2017). Available from: https://docs.oracle.com/en/database/oracle/oracle-database/12.2/tdpsg/enforcing-row-level-security-with-oracle-label-security.html#GUID-14A8D6C6-629B-46C5-86A9-2AED3A46E64D [Accessed 12 Apr 2024]

Nguyen S. (2023). Available from: https://blog.dreamfactory.com/the-benefits-of-oracle-dbms-for-your-organization/#:~:text=This%20popular%20and%20powerful%20relational,a%20secure%20hybrid%20cloud%20environment. [Accessed 11 Apr 2024]

Noss S. (2023) Available from: Data Risk Mitigation: How To Keep Your Organization's Data Safe. Available from: https://www.datagrail.io/blog/data-privacy/data-risk-mitigation/ [Accessed 17 Apr 2024]

Oracle (2021). Available from: https://www.oracle.com/a/ocom/docs/10-benefits-of-oracle-data-management-platform.pdf

Oracle (N.D.) Available from: https://www.oracle.com/webfolder/technetwork/tutorials/obe/db/10g/r2/owb/owb10gr2_gs/owb/lesson3/slowlychangingdimensions.htm [Accessed 15 Apr 2024]

Romansanta J. (2020). Available from: https://medium.com/@jromasanta/a-step-by-step-database-normalization-for-dummies-or-not-1b32b725e1be

United IT Consultants (2023) The Importance of Data Backup: Mitigating the Risks of Losing Critical Information. Available from: https://www.linkedin.com/pulse/importance-data-backup-mitigating-risks-losing-critical [Accessed 17 Apr 2024]

Bibliography:

Kowieski J. (2022) Available from: https://www.thoughtspot.com/data-trends/data-science/what-is-data-cleaning-and-how-to-keep-your-data-clean-in-7-steps [Accessed 14 Apr 2024]

Tableau (N.D.) Available from: https://www.tableau.com/learn/articles/what-is-data-cleaning [Accessed 15 Apr 2024]

The Upwork Team (2022) Available from: https://www.upwork.com/resources/data-cleaning-basics [Accessed 14 Apr 2024]