

University of Essex

Department of Computing

Deciphering Big Data March

Individual Project - Executive Summary

Radha Murugan

Table of Contents

Overview	3
Problem Statement	3
Highlights of the Solution	3
Key Features	4
Dimensions	4
Data Types	5
Data Management Pipeline	6
GDPR and Compliance	8
Recommendation	9
Why Oracle?	9
Data Structure	9
Future considerations	10
Risks and Limitations	10
Conclusion	11
References	11

Overview

The Executive Summary provides details about the Human Resources (HR) Hiring process, including their findings, choice of the data model, strengths and limitations of the solution, risk mitigations for GDPR compliance, and input on the University's previous report.

A summary of the approach that aims to improve the efficiency of data management, streamline the hiring process, and enhance decision-making capabilities. Based on the business requirements, this executive summary generates the data model, data normalization, data transformation, suggested database builds, and logical designs, and evaluates the data management pipeline process.

Problem Statement

Some of the problems with the current approach have been identified based on needs collected and analysis performed.

- Data integrity issues due to unnormalized structure and data redundancy
- Access rights
- Storage
- Performance
- Integration with extendable tools
- Data masking (for use in third-party tool testing)
- Inappropriate Management Information (MI) reports

The above issues impact business in multiple areas, some are emphasized.

- Data repetition in multiple places from various data sources causes data integrity
- Having inappropriate access to data causes data protection issues
- Not having appropriate design and scalable tools impacting storage and performance
- Extending the job functions to work with external sites like LinkedIn, Totaljobs requires a versatile and maintainable DB
- For testing in non-production environments, data masking is not available for testing not adhere to data protection policies

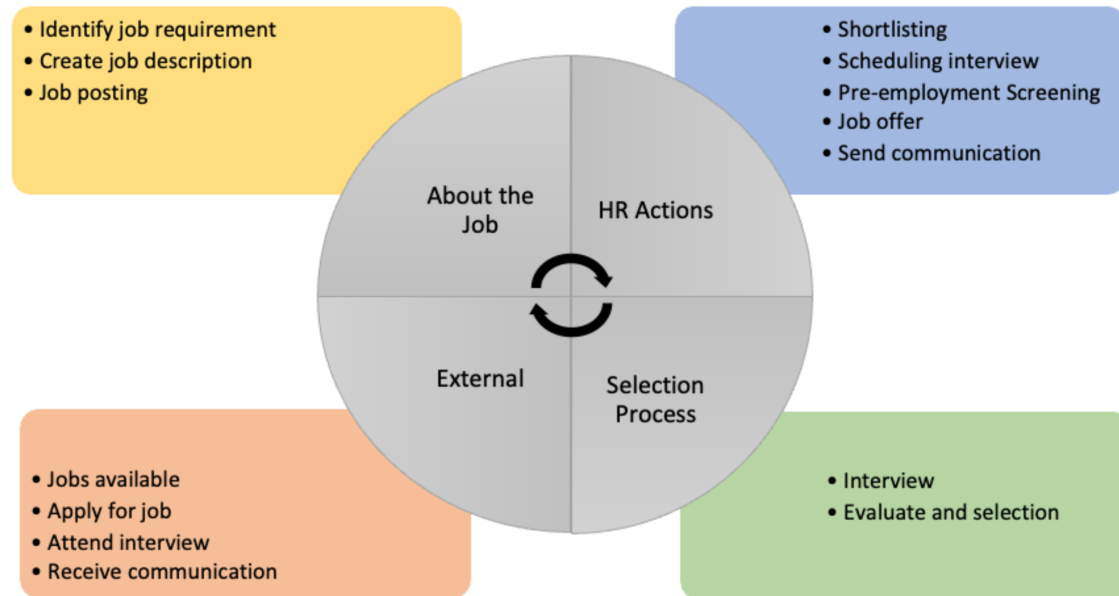
Highlights of the Solution

Oracle has been selected as the database solution after a detailed analysis of all the problems the organization has been having with the HR Hiring process (Nguyen, 2023) because Oracle eliminates information silos and makes data access and extraction simpler.

High availability is often achieved with Oracle Sharding/RAC, Oracle Multitenant/Pluggable databases (PDBs), and Oracle Active Data Guard. The Application Continuity tool is capable of handling both scheduled and unforeseen disruptions (Database Trends and Applications, 2021).

Key Features

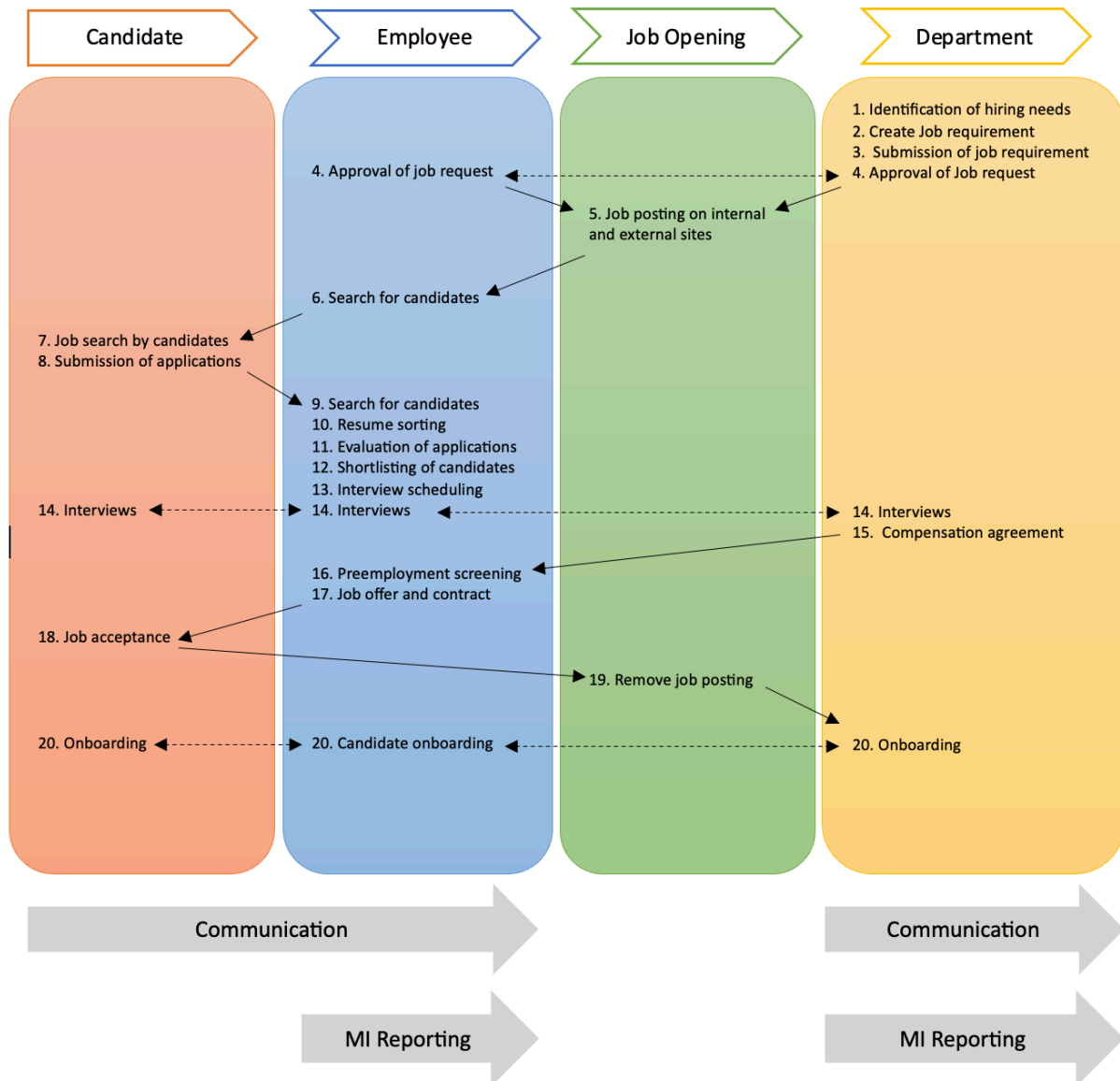
The solution implementation identified the key features.



Dimensions

Following a thorough study, it was determined that the hiring data needed to be rationally examined from a variety of angles.

- Candidate Dimension: Provides the data and status of the candidate
- Employee Dimension: Provides the information about the employee who is working with candidates and provides updates on the hiring process
- Department Dimension: Provide details of the department that requested talent acquisition from the job openings
- Job openings Dimension: Provide the details of the job created by the organization
- MI Reporting and Communication are horizontal processes across the various stages of recruitment activities



Data Types

As part of the indexing strategy, it is important to use the right data key values to ensure consistency across the different dimensions, hence understanding the data type is a key activity.

Following analysis, the various data kinds are divided into different groups.

- **Key:** The primary key for each dimension to provide uniqueness to the records
- **Key:** Foreign key to ensure consistency across the dimensions, so that identifying the connected records

- Dimension Contents: Multiple data types based on the details of the data such as varchar, numeric, and date. This is required to represent the data in the correct format for data processing and to produce accurate MI reports
- Record Timestamp: Date type (this is an important uniquely identify the records using type 2 Type 2 Slow Changing Dimension method (Oracle, ND). This is particularly useful for tracking changes over time, such as when a candidate is hired as an employee

Ensure each field is uniquely named and that all entries in the data cells are single-valued (Romansanta, 2020).

The dimensions, associated data structure with defined keys, contents of the dimensions, and audit-related information such as record timestamps are shown in the table below.

Dimensions											
CANDIDATE		EMPLOYEE		JOB_OPENING		DEPARTMENT		EE_LEVEL		EE_TYPE	
Attribute	Data Type	Attribute	Data Type	Attribute	Data Type	Attribute	Data Type	Attribute	Data Type	Attribute	Data Type
CAN_ID	NUMERIC	EE_ID	NUMERIC	JOB_ID	NUMERIC	DEPART_ID	NUMERIC	EL_ID	NUMERIC	EE_TYPE_ID	NUMERIC
JOB_ID	NUMERIC	DEPART_ID	NUMERIC	EE_ID	NUMERIC						
		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
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				
END_DATE	DATE	END_DATE	DATE	END_DATE	DATE	END_DATE	DATE				
LAST_UPDATE	DATE	LAST_UPDATE	DATE	LAST_UPDATE	DATE	LAST_UPDATE	DATE				

Data Management Pipeline

- The pipeline describes the process of how the HR Personnel (Employee) can input the data.
- Employee addition can be automated by providing an external interface to the employee DB of the organization, although in the current model, it can be a manual entry.

- Candidate additional can also be automated by providing an external interface to social sites like LinkedIn, although in the current model, it can be a manual entry.

Data Normalization

As part of the normalization strategy, apart from the above dimensions, it is important to ensure that the employee-level details and employee-type details are kept standard, consistent and need not be repeated in multiple tables.

- Applying normalization strategy, implementation of employee level and employee type additional dimensions designed to improve the access control and enhance the richness of data.
- The design focuses on the candidate's journey throughout the hiring process along with the initiation of the job request from the department and the employee working with the candidate and their data.
- The design ensures that every single attribute is unique
- The design makes sure that the key information is unique
- The foreign key implementation ensures that each row and its dependencies are addressed

Without these implementations in place, the data processing can lead to data integrity and the information sent across through MI reports can be incorrect and will require a manual activity in data correction.

Transformation

The data collected from external sites about the candidates may expected to have about 30% of operational expenses to correct the data (Kim, 2020).

The data errors can be avoided and made consistent when the data transformation happens. Some of the critical issues:

- Data missing during data entry and the storage can be modified to logically process it correctly. Null values can be incorporated when empty.
- Similar data requested from two different journeys can come back with two different data values, which can be transformed into one before it reaches the data dimensions.
- Data formatting is the most critical one, which can be addressed by this logical design.

Without these transformation processes, making sure that the data is cleaned is not possible. This ensures the database operates efficiently and that the data is in the correct format with the intended meaning (Antkowiak & Nowaczyk, 2021).

MI Reporting

The reporting tool was chosen as part of Oracle Reporting and Analytics. It also provides self-service exports to provide reports to different parts of the organization. It also provides a scheduled mailing facility that can be configured with required permissions.

Multiple sample reports are available as part of the Oracle HRMS module, which can be used to build MI reports on Jobs and positions (Oracle, 2015).

Here are some of the reports to be created and planned as part of this solution

1. Approved positions and their status
2. Position and the selected candidates within a department
3. Pre-hire report

Communication

Email communication to the candidates and the internal departments on the progress can be achieved through Oracle's tool Email Marketing (Oracle, 2021).

This tool allows us to provide the standard templates that can be designed to send to the candidates for interviews and other communications and providing the email-targeted audience and scheduling will help in the overall hiring process. The rejection emails will follow a standard template to ensure transparency with the candidates about the removal of their data collected as per GDPR compliance.

Data Masking

The Oracle data masking removes sensitive data and preserves data characteristics to support carrying out all of the hiring process after data masking. Pre-built masking templates can be built to reduce effort for deployment and enhance testing in the non-prod environments and mask or encrypt the PI details in the log information.

GDPR and Compliance

- Candidate's data provided to the organization. Candidates provide consent to process personal sensitive data like gender, culture, disability info, etc (Bika, 2018).
- Data controllers are put in place to ensure that the collection of data is restricted only related to the hiring process.
- Candidates having the right to access their data and allowed to rectify as necessary is made easy as the technical solution of the Candidate dimension will help to restrict any other data available for candidates
- Data processing and data distribution have been controlled by the 6 dimensions and the MI reporting is based on the data analysis report, not including any PI information.
- As part of GDPR compliance, it is required to have rejection emails follow a standard template implemented as part of the communication solution.
- The PI data removal planned as part of this solution as per GDPR guidelines

Recommendation

After a thorough analysis of the above problem statements, the solution recommended the usage of Oracle DB and its product features and tools which provides performance optimization, scalability, and security aspects of the database system.

Why Oracle?

MySQL has been analyzed carefully and identified that only a few of the problem statements are addressed but not all.

- This could be an ideal solution when dealing with a small amount of data
- MySQL is an open system, lot of code to be developed and managed to ensure it satisfies the data protection and GDPR requirements (Erik, 2022).

The following characteristics have been provided by Oracle are essential to the solution.

Data Structure

- Oracle allows multiple different ways to index the data, which means better ways to retrieve the data
- Oracle provides high scalability and huge amounts of data processing
- Oracle provides many concurrent transactions and provides large connection pool
- Oracle provides distributed databases, which provide access to different files from different networks
- The Database schema reflects a normalized structure that minimizes redundancy and ensures data integrity
- Implementation of referential integrity constraints to enforce data consistency

Data Transformation (ETL)

- Oracle Data Integrator (ODI) provides unique techniques for transporting data between Oracle DB for loading and transforming large volumes of data (Kush, Hermann Baer, and Potineni, 2003).

Security

- Oracle releases security updates and patches to protect against emerging threats with minimal or no service impact (Arsalan, 2023).
- Implementation of robust security measures including encryption and protocols to safeguard candidate sensitive data
- Oracle has more access security and it uses a combination of username, password, and valid profile. Protect data from unauthorized access through role-based control

Monitoring

- Easy for DBAs to manage and monitor their database, Oracle's autonomous database eliminates time-consuming administrative tasks and human error, which will free up the administrator's time.
- High availability and disaster recovery solutions are provided by Oracle to ensure maximum availability and minimize downtime (Database Trends and Applications, 2021).
- Continuous monitoring and evaluation of the database system to identify and address any emerging issues or bottlenecks

Performance

- Adoption of indexing strategies to improve query performance
- Regular database maintenance procedures to ensure optimal performance and reliability.
- Robust performance in addition to automatic performance management, compression, and encryption features to optimize data processing (Arsalan, 2023).

Reports

- Oracle Reporting and Analytics which has improved in user experience, new dashboards, and visualizations. Report-builder makes the creation of customized reports easy (Oracle MICROS Reporting and Analytics, 2023).

Cost

- The usage of MySQL is cost-effective compared to Oracle, due to Oracle's complex structure, and licensing cost.
- After careful consideration, it is recommended to use Oracle's DB as an enterprise-level Oracle Contract exists with the other parts of the organization in different departments and hence cost implementation would be minimal.

Future considerations

- Oracle provides flexible deployment options that are scalable when the organization would like to move into Cloud-based solutions.
- As the organization goes larger, talent acquisition needs to be very efficient, and relevant IT systems and tools needed to be maintained to the latest trend may become difficult Recruitment process outsourcing could be a cost-saving option to bring in qualified candidates, which will allow the HR team to focus on internal activities (Nick, 2023).

Risks and Limitations

Risks

Some of the critical risks identified in the solution can be mitigated.

Successful Implementation of the solution is very critical to solve the data protection issues. There is a high risk of data being accessible without specific role-based rights (Noss, 2023). As the candidate's personal information is very sensitive data, without the above solution in place, it can result in regulatory penalties.

Limitations

There are a few limitations to this solution, availability of common lookup tables.

As part of this solution managing the common Employee level and Employee type information across the organization, may vary specifically within different departments due to the varied functions carried out by the departments (Sen, 2009). Since the logical model is platform agnostic, performance optimizations such as query tuning, data storage, and retrieval will need to be taken care of during the conversion to the physical model.

Conclusion

In conclusion, the logical database solution of the HR hiring system is streamlined and satisfies the requirements of the organization by following best practices and incorporating the recommendations. The solution recommends how the security issues are being addressed effectively by conducting a critical evaluation of the overall solution. However, a detailed view of relationships and business rules is addressed. However, ongoing maintenance and monitoring are key to ensuring the integrity, performance, and security of the database.

References

Database Trends and Applications. (2021). Top 5 Reasons to Use an Oracle Database. [online] Available at: <https://www.dbta.com/Editorial/News-Flashes/Top-5-Reasons-to-Use-an-Oracle-Database-144191.aspx>.

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.>

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

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

Antkowiak M. & Nowaczyk M. (2021) Available from: <https://medium.com/transparent-data-eng/data-cleansing-examples-24581c3d14f1>

Kim, F. (2020). What is Data Accuracy, Why it Matters, and How Companies Can Ensure They Have Accurate Data? [online] Data Ladder. Available at: <https://dataladder.com/what-is-data-accuracy/>.

Oracle. (2015). Oracle Human Resources Management Systems Window Navigation and Reports Guide. [online] Available at: https://docs.oracle.com/cd/E18727_01/doc.121/e13517/T5677T5680.htm

Oracle. (2021). What is Email Marketing | Oracle? [online] Available at: <https://www.oracle.com/cx/marketing/email-marketing/>.

Bika, N. (2018). GDPR compliance guide for recruitment | Workable. [online] Recruiting Resources: How to Recruit and Hire Better. Available at: <https://resources.workable.com/tutorial/gdpr-compliance-guide-recruiting>.

Erik M. (2022) When Should You Choose Oracle vs SQL? [online] Available at: <https://www.cbtnuggets.com/blog/technology/system-admin/when-should-you-choose-oracle-vs-sql>.

Kush, F., Hermann Baer, M.B. and Potineni, P. (2003). Data Movement/ETL Overview. [online] Oracle Help Center. Available at: <https://docs.oracle.com/en/database/oracle/oracle-database/19/dwhsg/etl-overview.html#GUID-BE3AB415-4D45-45B2-843B-0FCC0567E991> [Accessed 27 May 2024].

Arsalan D. (2023). Why Oracle Database is Still a Top Choice for Enterprises. [online] Available at: <https://www.linkedin.com/pulse/why-oracle-database-still-top-choice-enterprises-dehghanisariyarghan/>.

Oracle MICROS Reporting and Analytics. (2023). Available at: https://docs.oracle.com/cd/E65545_01/reportinganalytics/20.1/CS.pdf

Nick B. (2023). What Is Recruitment Process Outsourcing (RPO)? | Definition from TechTarget. [online] Available at: [https://www.techtarget.com/searchhrsoftware/definition/recruitment-process-outsourcing-RPO#:~:text=Recruitment%20process%20outsourcing%20\(RPO\)%20is](https://www.techtarget.com/searchhrsoftware/definition/recruitment-process-outsourcing-RPO#:~:text=Recruitment%20process%20outsourcing%20(RPO)%20is)

Noss, S. (2023). Data Risk Mitigation: How To Keep Your Organization's Data Safe. [online] DataGrail. Available at: <https://www.datagrail.io/blog/data-privacy/data-risk-mitigation/>.

Sen, A. (2009). Five Simple Database Design Errors You Should Avoid. [online] Simple Talk. Available at: <https://www.red-gate.com/simple-talk/databases/sql-server/database-administration-sql-server/five-simple-database-design-errors-you-should-avoid/>.