DWMA Assignment 1

1. Select one of the KPI (Key Performance Indicator) and decide who is the 'stakeholder' for the project you intend to represent. Propose a centralised database solution, architecture and methodology for the organisation.

Introduction and Current Problems

The North and West Yorkshire Clinical Commissioning Group (CCG) aims to create a health and social care system that incorporates data from social care providers and care homes. The Leeds City Council (LCC) runs six care facilities that assist senior citizens while they heal, three in North Yorkshire (NYR) and three in West Yorkshire (WYR). The services provided by hospitals are extended by these care facilities.

The primary function of the system is the evaluation of care home effectiveness through the tracking and analysis of data associated to the length of recovery and bed occupancy in the care homes, as well as the integration of social care and health data from care homes to enable healthcare professionals, social workers, and service providers to access patient records more easily.

The problems from the current system can be mentioned as this, the social care and healthcare systems that are now in place are not well linked. Because of this, it is difficult for medical professionals and social workers to easily access one other's data, which makes it difficult to offer senior patients comprehensive care. The other issue that we can find is all older people are regarded as care home patients, however not all of them might be on the social care database. This may result in informational gaps. Moreover, according to the case study, social workers are more valuable in having access to patient data than medical professionals like doctors and nurses are to the data of the social care system. This indicates that to improve the standard of care and support services, social care providers should give priority to gaining access to patient data.

KPI and Stakeholders

The System allows us to work with a large number of stakeholders and KPIs, but the contents of the report will be covered below.

KPI: Improve the effectiveness of the bed occupancy of wards and each care home.
Stakeholder: North and West Yorkshire CCG (Clinical Commission Group)

The reports listed below can give the CCG information regarding the efficiency of ward bed occupancy for the chosen KPI.

Reports:

- Total no. of beds occupied last month between each care home.
- Number of beds that were unoccupied last year for each care home.
- Bed type most requested in each ward for last month for each care home.
- Wards that beds are fully occupied to the max capacity last 4 months.
- Wards that least number of beds got occupied in last month.

Centralised database solution, Architecture and Methodology

Given that the North and West Yorkshire Clinical Commission Groups are using two distinct databases, as stated in the case study, in order to satisfy the above-mentioned KPI, This organization would benefit from a data warehouse since it offers a platform for analytics and reporting and enables the integration of data from multiple sources, such as social and health care systems. All relevant information will be housed centrally in the data warehouse.

To address the above KPI the only relevant data tables need to be included. The data tables are Ward, Bed, Care Center. The proposed solution can be mentioned as below.

- Ward, Bed, Care Center Tables will be dimension tables with only relevant data that need to address the KPI and the Fact table will be designed from the relevant data from the dimension tables.

To design a Datawarehouse we can use star schema or Snowflex schema in the context of the North and West Yorkshire Clinical Commissioning Groups (CCG), the decision that what design we go with for the data warehouse in terms of a star or snowflake schema is based on the demands and requirements of the company because Each schema offers benefits and compromises.

Data is arranged in a star schema around a main fact table that is linked to dimension tables. When it comes to advantages of Star schemas, they are easier to build, use, and comprehend implicitly. When working with intricate data related to social care and healthcare, this can be quite helpful. Moreover, User-Friendly Because of its straightforward structure (Kimball et al., 2011). When we talk about show flex schema to lessen data redundancy, the dimension tables in a snowflake schema are normalized, or divided into several related tables. Snowflake schemas eliminate redundancy and ensure consistency in data, which can be beneficial in the healthcare industry where data accuracy is critical. On the other hand, can be more storage efficient (Inmon et al., 1996).

According to the requirements it seems like the data accuracy and storage efficiency is not a big deal, but query speed and user-friendliness are important considerations there for the solution we consider designing the data warehouse is star schema.

Methodology

There are two management approaches that we must consider while creating the data warehouse. Details about the approaches and its technique are discussed below.

The top-down method: When using a top-down method, need to build the data warehouse as a whole first, encompassing all its components, before putting any particular ones into practice. It requires creating a comprehensive data architecture, data modelling, and ETL procedures that go from the highest level to the lowest degree of detail. When we weigh the benefits of this, we can see that it provides a strategic and comprehensive picture of data needs, guarantees alignment with organizational goals and objectives, and is generally easier to implement in terms of consistency and integration (Inmon et al., 1996). This technique works effectively for organizations who wish to guarantee consistency and alignment with their strategic goals and have a well-defined enterprise-wide data strategy.

On the other hand, In a **bottom-up method**, the initial stage is to build individual data marts or smaller data warehouses to meet specific departmental or functional needs. Over time, the integration of these smaller units into a bigger data warehouse is driven by the demands of the firm.

This methodology's advantages include a faster time-to-value because of the quick deployment of individual components, simplicity in establishing the project's success early on, and adaptability to changing organizational requirements (Kimball et al., 2011). This strategy is suitable in situations when departmental or functional data requirements are particular and urgent resolutions are needed. It permits adaptability and gradual advancement. Long-term, however, integrating data marts into a unified data warehouse can demand more work.

Taking into account the point, the **Kimball approach bottom-up approach** will be appropriate based on these requirements due to these advantages. When CCG departments or care facilities have critical data-related needs or difficulties, a bottom-up strategy can be used to swiftly resolve these problems. It makes it possible to quickly implement data solutions that are customized to the needs of social workers and other healthcare professionals, who are more inclined to accept solutions that they helped build and increase the likelihood of successful adoption.

Architecture

The primary options available to us when selecting a data warehouse architecture are Different strategies for arranging the parts and tiers of a data warehouse system are represented by **two-tier and three-tier data warehouse architectures**.

In a **two-tier data warehouse architecture**, client applications such as reporting, analytics, and business intelligence tools can directly access data that is kept in a centralized repository called the data warehouse. This architecture is straightforward and reasonably simple. and the benefits of this include direct access to data for reporting and analytics apps and a simplified architecture with fewer layers.
An extra layer known as the application layer, or middle tier, is introduced in **three-tier data warehouse architecture** between the client applications and the data warehouse. Request management, business logic, and data processing are handled by this intermediate layer (Kimball et al., 2011). The benefits of this are Scalability and maintainability are enhanced by the architecture's separation of the user interface, application logic, and data storage layers.

There are various architectural strategies for structuring and arranging data marts and data integration in the context of data warehousing. Each strategy has benefits and is appropriate for various use situations. Let us take a quick look at a few popular data mart and data integration architectures. A stand-alone data warehouse created especially for a particular division or business unit is called an independent data mart. Multiple separate data marts are formed in a data mart bus architecture, each supporting a distinct department or business sector. The primary hub (data warehouse) in a hub-and-spoke architecture acts as the primary repository for integrated data. A single, comprehensive data warehouse that compiles information from multiple sources throughout the company is the focal point of a centralised architecture. Data is still stored in the original source systems in a federated architecture, and real-time query distribution occurs to these systems. When we consider each approach the Hub and spoke method is most appropriate with the designing the data warehouse because to facilitate data integration and cross-functional analysis among the centres, a hub and spoke architecture could prove advantageous. It offers some liberty with a certain degree of data integration.

2. Produce a star schema (SS) design for a decision support system to support your KPI and the reports (2-4 as a guide). Document by producing a SS data dictionary.

a. Dimension Table

Table Name	**Attribute Name**	**Data Type**
Time	*Time_Id*	Int (PK)
	Year	Int
	Month	Int
	Date	Int
Care_Centre	*Care_Id*	Int (PK)
	Name	Varchar (20)
Ward	*Ward_Id*	Int (PK)
	Name	Varchar (20)
	Capacity	Int
Bed	Bed_No	Int (PK)
	Status	Varchar (50)
	Type	Varchar (50)

b. Fact Table

Table Name	Attribute Name	Data Type	Keys	Measure
BedOccupency_FACT_Table	Number	Int	Primary Key	Unique serial number
	Time_Id	Int	Foreign Key	Refers to Time table.
	Care_Id	Int	Foreign Key	Refers to Care_centre table
	Ward_Id	Int	Foreign Key	Refers to Ward table
	TotalNo_occupied_beds	Int	Non key	

c. Data Sample for Dimension Tables

Ward

Ward_Id	Ward_Name
001	GENERAL care
002	GENERAL care
003	GENERAL care
004	ICU

Time

Time_id	**Year**	**Month**	**Day**
1	2023	1	12
2	2023	2	15
3	2023	3	23
4	2022	5	10
5	2022	12	4

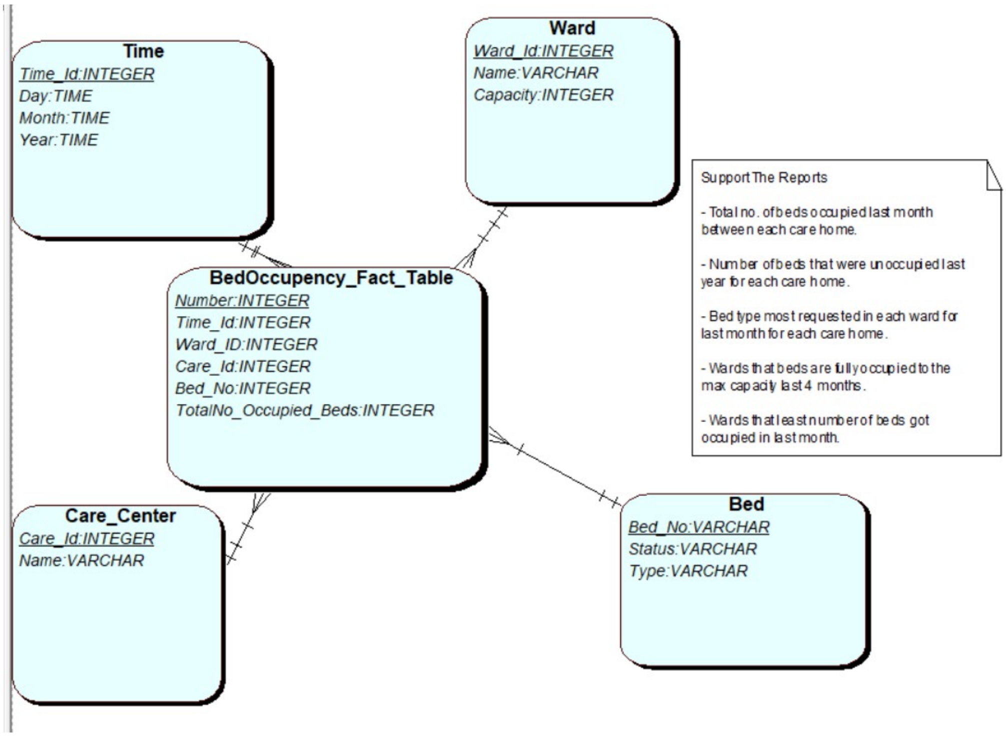
Care_Centre

Care_Id	**Name**
1	LBU CareHome
2	OSCAR CareHome
3	JUNO CareHome
4	BEWAN CareHome
5	ALTORN CareHome
6	WELLBEING CareHome

Bed

Bed_No	**Bed_Type**	**STATUS**
1102	Manual	Available
1024	Single	Occupied
1502	Semi-electric	Occupied
1401	Fully Electric	Occupied

d. QSEE Star Schema



Support The Reports

- Total no. of beds occupied last month between each care home.

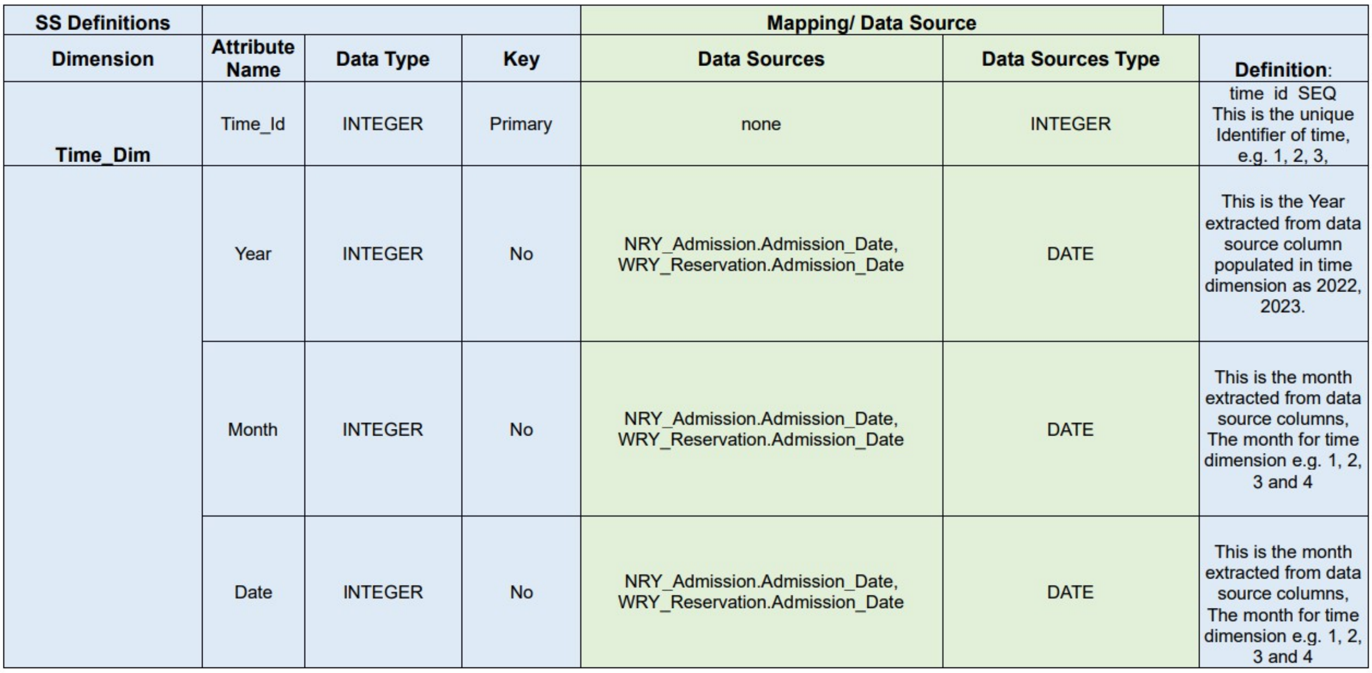
- Number of beds that were unoccupied last year for each care home.

- Bed type most requested in each ward for last month for each care home.

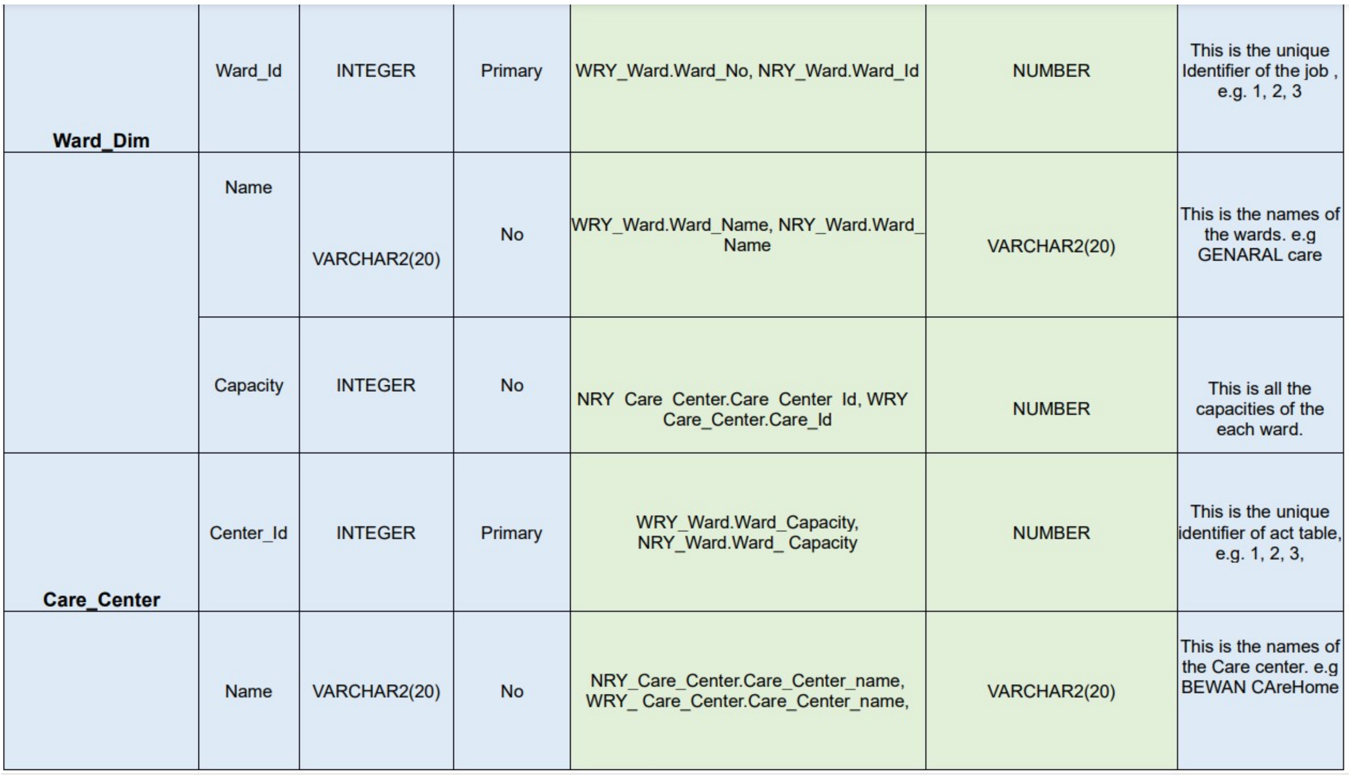
- Wards that beds are fully occupied to the max capacity last 4 months.

- Wards that least number of beds got occupied in last month.

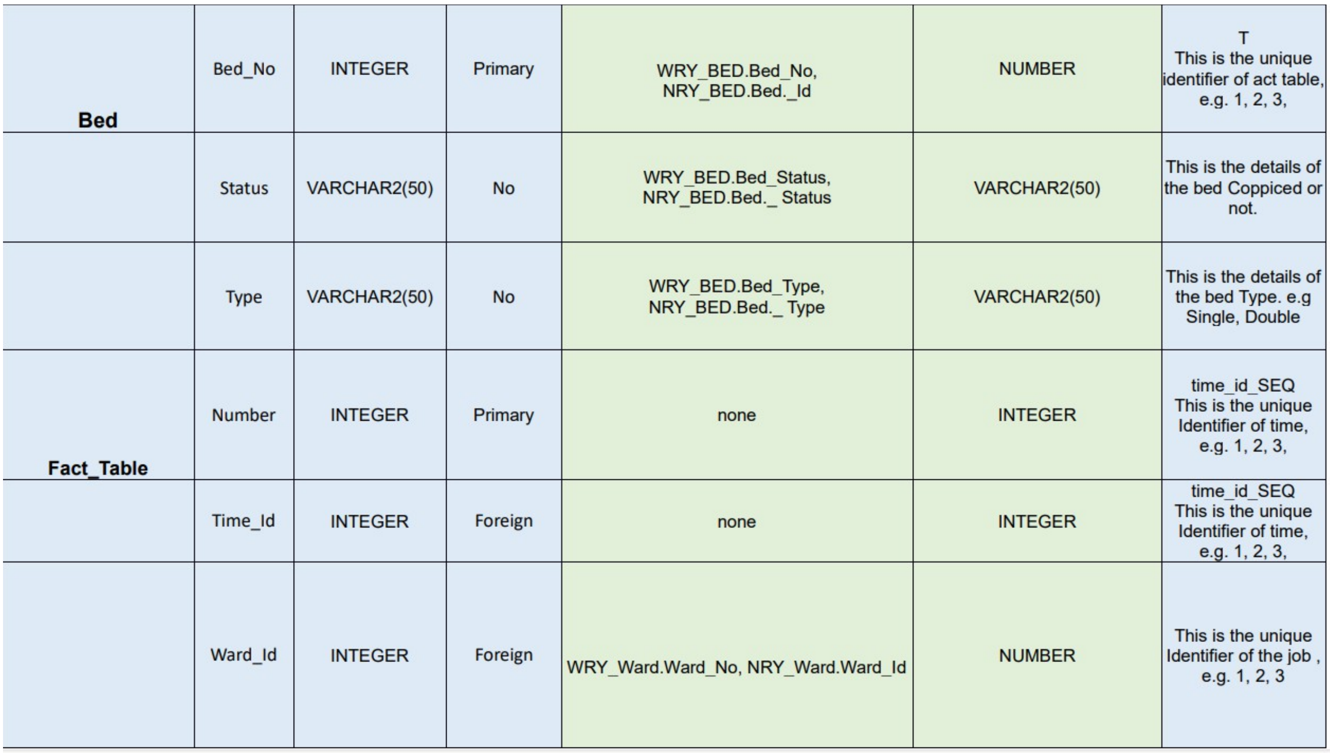
e. Data Dictionary



SS Definitions				Mapping/ Data Source		
Dimension	**Attribute Name**	**Data Type**	**Key**	**Data Sources**	**Data Sources Type**	**Definition**:
Time_Dim	Time_Id	INTEGER	Primary	none	INTEGER	time id SEQ This is the unique Identifier of time, e.g. 1, 2, 3,
	Year	INTEGER	No	NRY_Admission.Admission_Date, WRY_Reservation.Admission_Date	DATE	This is the Year extracted from data source column populated in time dimension as 2022, 2023.
	Month	INTEGER	No	NRY_Admission.Admission_Date, WRY_Reservation.Admission_Date	DATE	This is the month extracted from data source columns, The month for time dimension e.g. 1, 2, 3 and 4
	Date	INTEGER	No	NRY_Admission.Admission_Date, WRY_Reservation.Admission_Date	DATE	This is the month extracted from data source columns, The month for time dimension e.g. 1, 2, 3 and 4



Ward_Dim	Ward_Id	INTEGER	Primary	WRY_Ward.Ward_No, NRY_Ward.Ward_Id	NUMBER	This is the unique Identifier of the job , e.g. 1, 2, 3
	Name	VARCHAR2(20)	No	WRY_Ward.Ward_Name, NRY_Ward.Ward_ Name	VARCHAR2(20)	This is the names of the wards. e.g GENARAL care
	Capacity	INTEGER	No	NRY Care Center.Care Center Id, WRY Care_Center.Care_Id	NUMBER	This is all the capacities of the each ward.
Care_Center	Center_Id	INTEGER	Primary	WRY_Ward.Ward_Capacity, NRY_Ward.Ward_ Capacity	NUMBER	This is the unique identifier of act table, e.g. 1, 2, 3,
	Name	VARCHAR2(20)	No	NRY_Care_Center.Care_Center_name, WRY_ Care_Center.Care_Center_name,	VARCHAR2(20)	This is the names of the Care center. e.g BEWAN CAreHome



Bed	Bed_No	INTEGER	Primary	WRY_BED.Bed_No, NRY_BED.Bed._Id	NUMBER	T This is the unique identifier of act table, e.g. 1, 2, 3,
	Status	VARCHAR2(50)	No	WRY_BED.Bed_Status, NRY_BED.Bed._ Status	VARCHAR2(50)	This is the details of the bed Coppiced or not.
	Type	VARCHAR2(50)	No	WRY_BED.Bed_Type, NRY_BED.Bed._ Type	VARCHAR2(50)	This is the details of the bed Type. e.g Single, Double
Fact_Table	Number	INTEGER	Primary	none	INTEGER	time_id_SEQ This is the unique Identifier of time, e.g. 1, 2, 3,
	Time_Id	INTEGER	Foreign	none	INTEGER	time_id_SEQ This is the unique Identifier of time, e.g. 1, 2, 3,
	Ward_Id	INTEGER	Foreign	WRY_Ward.Ward_No, NRY_Ward.Ward_Id	NUMBER	This is the unique Identifier of the job , e.g. 1, 2, 3

3. Produce an ETL Design for your SS, include an ETL data dictionary and a diagram (using excel) to show sample data for each of the SS tables.

Data Problems and Solutions

The common data quality issues listed below can have a big influence on how reliable and usable the information is in a data warehouse. It is imperative to tackle these concerns in order to guarantee the accuracy, significance, and dependability of healthcare and social care data.

Inaccurate Data: Data that is erroneous, incomplete, or misstated.
Solution: To find and fix errors in the data, need to apply data cleansing procedures, data profiling, and data validation guidelines (Redman et al., 1997).

Irrelevant data: Data that is not relevant or helpful to the objectives of the company.
Solution: Review and update data frequently to get rid of information that is out of date or unnecessary (Wang et al., 1996).

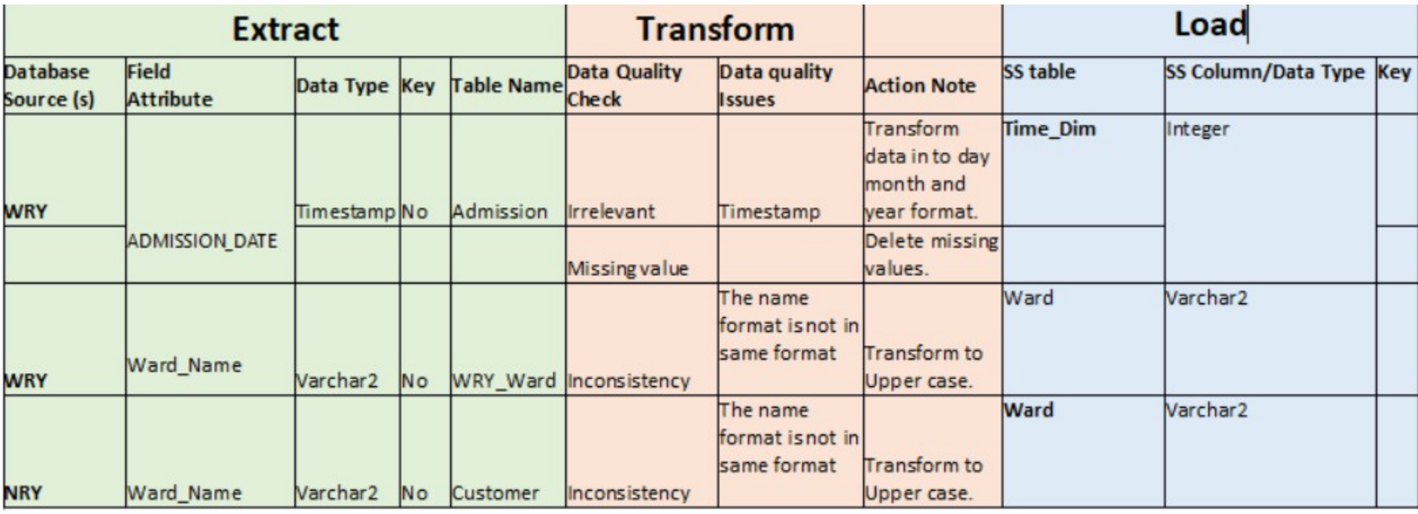
Out Date Data: Information that ages over time.
Solution: To maintain the data up to date, establish maintenance and update methods. Use data quality monitoring to find and fix out-of-date information (Redman et al., 1997).

Inconsistent Data: Variations and conflicts arise from data that is not consistent or standardized.
Solution: To guarantee consistency, define and enforce data standards, put data governance procedures into place, and apply data integration and transformation procedures (Kimball et al., 2011).

Invalid Data: Information that doesn't follow established guidelines for data integrity and quality.
Solution: To discover and fix erroneous data, apply quality checks and data profiling, enforce data limitations, and implement data validation guidelines (Wang et al., 1998).

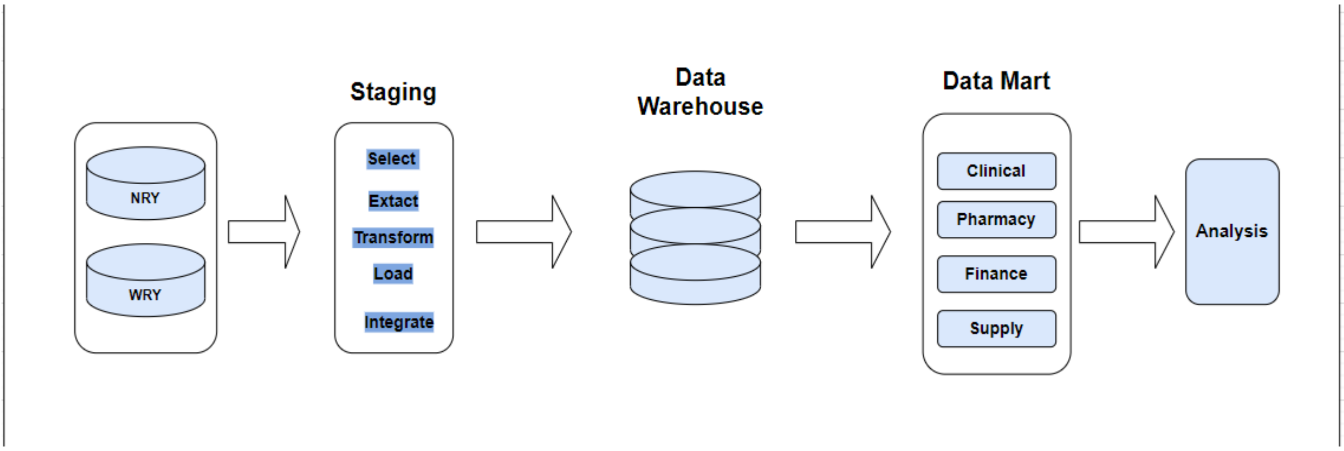
Unclear Data: Information with inadequate documentation or unclear interpretation.
Solution: Implementing metadata management, standards for data documentation, and precise data definitions (Redman et al., 1997).

ELT Data Dictionary



Extract					Transform			Load		
Database Source (s)	Field Attribute	Data Type	Key	Table Name	Data Quality Check	Data quality Issues	Action Note	SS table	SS Column/Data Type	Key
WRY	ADMISSION_DATE	Timestamp	No	Admission	Irrelevant	Timestamp	Transform data in to day month and year format.	Time_Dim	Integer	
					Missing value		Delete missing values.			
WRY	Ward_Name	Varchar2	No	WRY_Ward	Inconsistency	The name format is not in same format	Transform to Upper case.	Ward	Varchar2	
NRY	Ward_Name	Varchar2	No	Customer	Inconsistency	The name format is not in same format	Transform to Upper case.	Ward	Varchar2	

ELT Flow



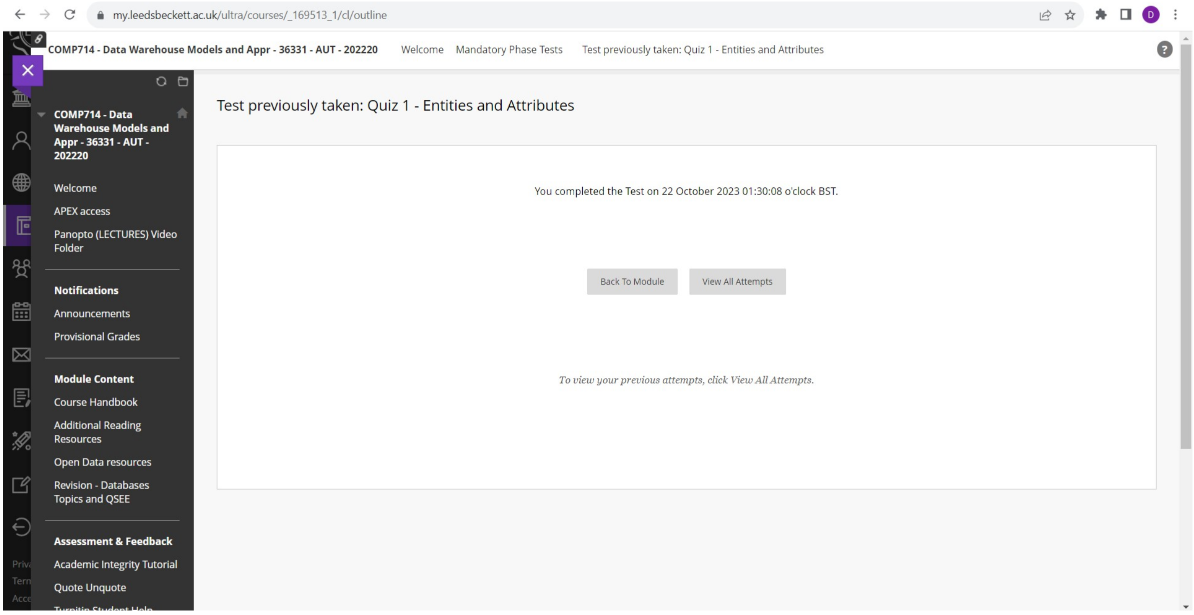
Creating the Extract, Transform, and Load (ETL) process for your healthcare and social care system is one of the most important steps in the data warehousing process. An overview of the ETL design factors for such a system is provided below:

Determine the data's source as a first step by locating its databases, files, and other systems. The NRY and WRY databases are what the healthcare system uses. In order to retain current data, recurrent extraction plans must be established. Custom scripts or ETL solutions will be used to extract data from source systems in the staging area. Moreover, data cleaning, data integration, Transformation and loading will be conduct in the staging area. Integrate data from various forms and sources into a single, cohesive structure through data integration. Data cleansing is the process of finding and fixing flaws, inconsistencies, and inaccuracies in data this includes improve data by including pertinent features or missing information. Data integrity and quality are ensured by applying validation criteria (Kimball et al., 2011).

.

4. Four Tutorials work upload.

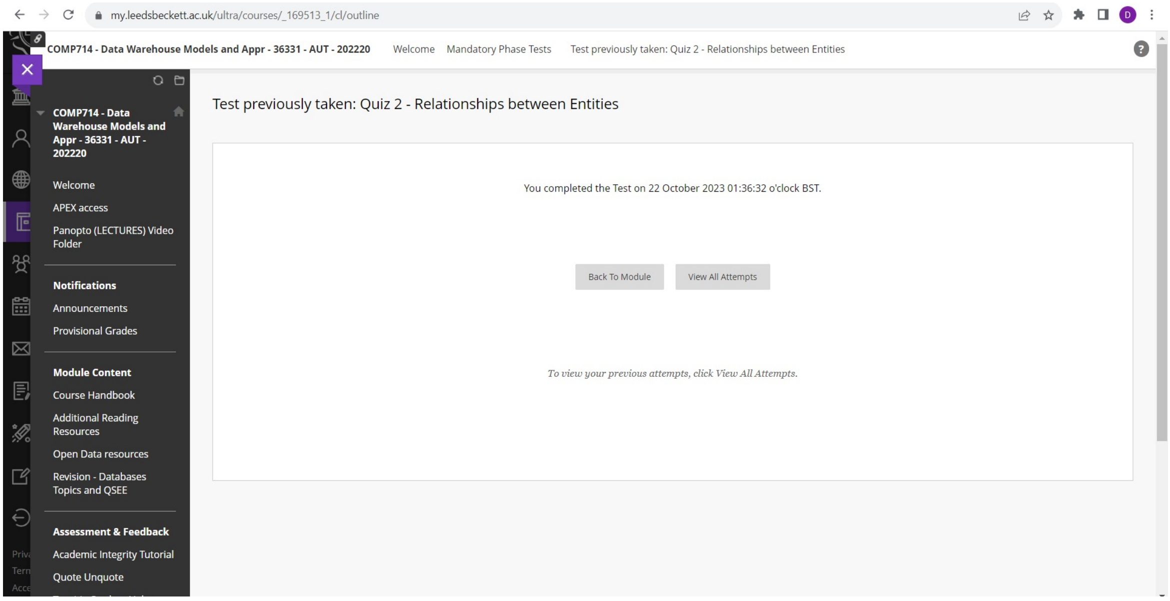
Tutorial Week 1


my.leedsbeckett.ac.uk/ultra/courses/_169513_1/cl/outline
COMP714 - Data Warehouse Models and Appr - 36331 - AUT - 202220
Welcome
Mandatory Phase Tests
Test previously taken: Quiz 1 - Entities and Attributes
COMP714 - Data Warehouse Models and Appr - 36331 - AUT - 202220
Welcome
APEX access
Panopto (LECTURES) Video Folder
Notifications
Announcements
Provisional Grades
Module Content
Course Handbook
Additional Reading Resources
Open Data resources
Revision - Databases Topics and QSEE
Assessment & Feedback
Academic Integrity Tutorial
Quote Unquote
Back To Module
View All Attempts

Test previously taken: Quiz 1 - Entities and Attributes

You completed the Test on 22 October 2023 01:30:08 o'clock BST.

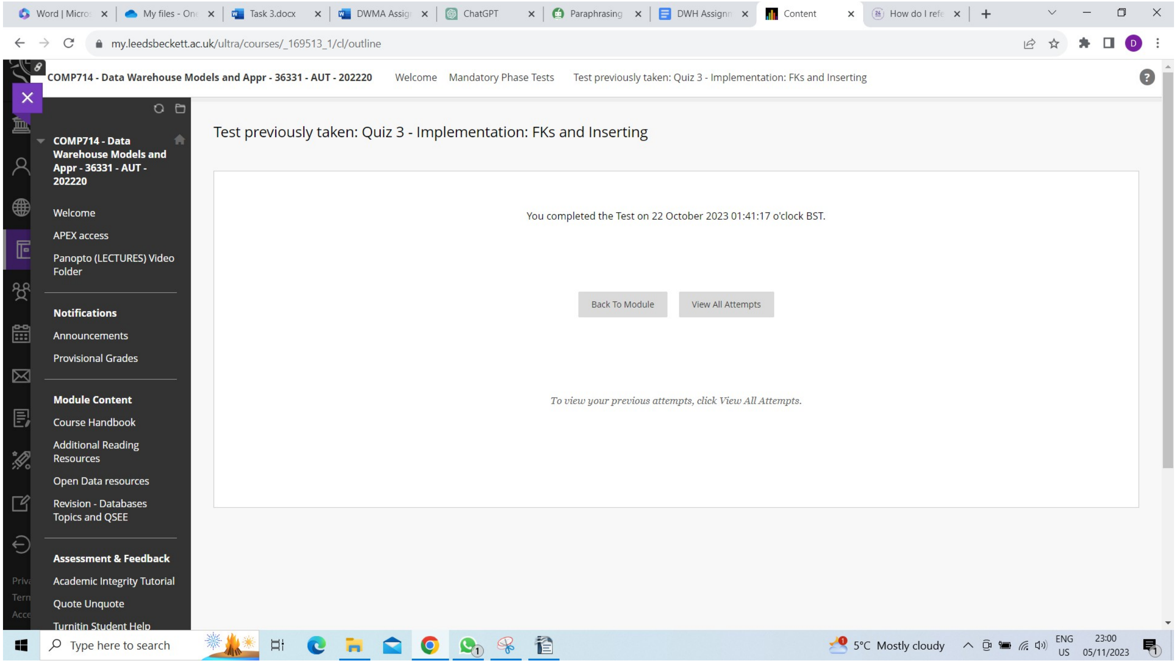
To view your previous attempts, click View All Attempts.


my.leedsbeckett.ac.uk/ultra/courses/_169513_1/cl/outline
COMP714 - Data Warehouse Models and Appr - 36331 - AUT - 202220
Welcome
Mandatory Phase Tests
Test previously taken: Quiz 2 - Relationships between Entities
COMP714 - Data Warehouse Models and Appr - 36331 - AUT - 202220
Welcome
APEX access
Panopto (LECTURES) Video Folder
Notifications
Announcements
Provisional Grades
Module Content
Course Handbook
Additional Reading Resources
Open Data resources
Revision - Databases Topics and QSEE
Assessment & Feedback
Academic Integrity Tutorial
Quote Unquote
Back To Module
View All Attempts

Test previously taken: Quiz 2 - Relationships between Entities

You completed the Test on 22 October 2023 01:36:32 o'clock BST.

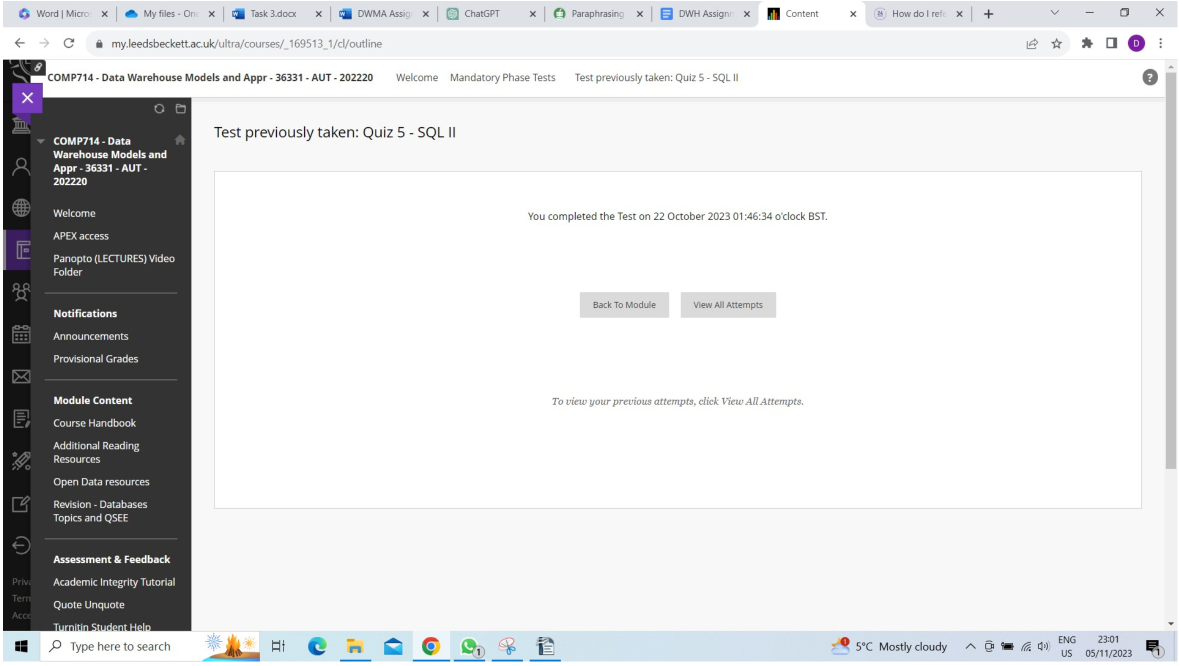
To view your previous attempts, click View All Attempts.



Test previously taken: Quiz 3 - Implementation: FKs and Inserting

You completed the Test on 22 October 2023 01:41:17 o'clock BST.

To view your previous attempts, click View All Attempts.



Test previously taken: Quiz 5 - SQL II

You completed the Test on 22 October 2023 01:46:34 o'clock BST.

To view your previous attempts, click View All Attempts.

Tutorial Week 2 (LA Shop)

1. Who are the stakeholders of LA Shop?
    - CEO, Borad Directors.
2. What would be some potential aims and objectives for LA Shop?
    - Increase revenue and sales
    - Improve customer satisfaction
    - Enhance employee performance
3. Choose one of these aims/objectives further consider.

a. What could they investigate to understand better the present situation?

- To Increase revenue and sales: The database data of the customer and the booking has been collected over the years from the business.

b. What reports would help this this?

- Which shop got the most bookings this month according to last month.
- Compcare the number of the bookings in last year of each location.
- Most popular items in the bookings for January month in each location.

c. What data is required for the report?

- Booking (Id, Product_Id, Time)
- Country, Region, shop

d. Draw out an example illustration of this report (maybe use excel).

- Which shop got the most bookings this month according to last Three months.

shop	Country	Region	Bookings this month	Bookings this last month
A	UK	Europe	1150	2075
B	France	Europe	2563	2889
C	USA	North America	3421	3120
D	Canada	North America	785	1340
E	Australasia	Australasia	897	721
F	Australasia	Australasia	1820	1560

4. How would LA shop know if they had been effective in reaching their objective?

a. Can the above report(s) include year on year trends? yes

b. Is there other information that would be useful? yes

5.Consider specifically for the reports the dimensions, i.e. variables to consider against each other (time … cost … product … location …) and the facts (no_of_shops, no_creche_bookings …)

Dimensions

- Time
- Booking
- Country
- Region

Facts

- Booking_id, Time_id, Contry_name, Region_id

Measures:

- Number of bookings of each shop

Tutorial Week 3

Design - Data warehouse requirements, OLTP/OLAP, the star schema, Data Warehouse approaches, methodology

The presentation covered essential aspects of data warehousing, including data warehouse requirements, OLTP (Online Transaction Processing), OLAP (Online Analytical Processing), the star schema, data warehouse approaches, and methodologies. Here's a concise summary of the key points:

Data Warehousing: A data warehouse is a subject-oriented, integrated, time-variant, and non volatile collection of data that supports management decision-making. The star schema, a data model used in data warehousing, consists of Fact Tables (quantitative data from events) and Dimension Tables (descriptive data) structured in a simple and accessible manner.

Advantages and Challenges of Dimensional Modeling: Dimensional modeling offers advantages such as a straightforward architecture, efficient data extraction, and the ability to drill down into data. Challenges include slower drill-up/down, complex drill-across and through operations, potential scaling issues, and the need for specialized tools.

OLTP vs. OLAP: OLTP systems are customer-focused and handle daily operational tasks like purchases, payroll, and banking. OLAP systems are market-focused and excel in data analysis and decision-making. They process large queries and are essential for historical data analysis and learning from experience.

Data Warehouse Approaches: Various methodologies are available for data warehousing, including Inmon's Corporate Information Factory, Kimball's Dimensional Data Warehousing, Bus Architecture, and Stand-Alone Data Marts. The choice of approach is influenced by factors such as business type, departmental requirements, data strategy, and project timelines.

In summary, data warehousing plays a vital role in enabling effective decision-making by providing well-structured data. Understanding the distinctions between OLTP and OLAP, the star schema, and various data warehouse approaches is crucial for designing and implementing successful data warehousing solutions.

virtual data warehousing

The idea of virtual data warehousing (VDW) is examined and contrasted with conventional data warehousing (DW) in the study by Khurram Shahzad and Ghulam Mustafa. Below is a summary of the main ideas covered in the paper:

The necessity of Decision Support Systems (DSS) for business performance analysis and decision support is emphasised at the outset of the study. A crucial part of DSS is introduced: data warehousing (DW), which is the process of transforming and storing operational data for decision support. perceptions.

Problem Statement: Difficulties with managing content and schema updates, as well as restricted access to the most recent data, are highlighted as issues with traditional DW. Information display and retrieval system improvements may be required as a result of structural changes in DW.

Further Requirements for DW: The document emphasise the necessity for DW to support schema changes without causing data loss, adjust to content changes without changing the schema, give users access to the most recent operational data, and allow for conventional SQL data retrieval.

Evaluation of Current Solutions: In order to meet DW issues brought on by modifications, the study assesses current solutions, such as schema evolution and schema versioning techniques. The efficiency of these techniques in handling changes to the schema and content is evaluated.

Results and Comparison: It is determined that the methodologies of schema versioning and schema evolution both meet the requirements for handling changes in content and schema. However, the schema evolution strategy is more adept at adjusting to changes, whereas the schema versioning approach could be resource-intensive and confusing for users.

The suggested remedy is virtual data warehousing, or VDW. In order to address the goals listed, the paper presents VDW as a virtualization technique. This strategy allows for quick data retrieval, on-the-fly data availability, schema adjustments, and content changes. A virtual, instance-less, subject-oriented, time-variant warehouse (VDW) is characterised as one that facilitates decision-making.

VDW-Tool, or Virtual Data Warehouse: The operational data sources (ODS), virtual schema, metadata repository, and mapping rules that comprise a VDW-Tool's architecture are covered. The conceptual framework is represented by the virtual schema, while operational data is via ODS. Standard SQL queries can get data by bridging the gap between virtual schema and operational sources through the use of metadata and mapping rules.

The 4-DEF Data Warehouse Evaluation Framework Four dimensions—content flexibility, schema flexibility, reaction speed, and versioning space—are used to create a framework for assessing and contrasting VDW with conventional DW. To carry out the investigation, parameters and hypotheses are developed.

Experimentation: With VDW and CDW, four case studies are utilised to test queries, changes to content and schema, and record count computations. Based on the developed hypotheses, the outcomes are examined.

The impact of data warehouses on decision-making and the significance of trustworthy and consistent information are highlighted in the paper's conclusion. It illustrates how VDW successfully addresses issues and improves flexibility in handling schema and content revisions, and it presents the 4-DEF framework for analysing VDW as well as the VDW-Tool for implementation.

Managing Data Lakes in Big Data Era

The paper explores the idea of data lakes and how useful they are for solving large data problems. Here is a succinct summary of the study's key concepts:

The purpose of this paper is to present the idea of data lakes and talk about real-world experiences with their application in big businesses. The growing volume and velocity of data has made data lakes popular as a way to manage and use large and varied amounts of data. The idea behind a data lake is to gather, process, store, and analyse unstructured and multi-structured data—which frequently has unrealized value for businesses—using reasonably priced technologies. It offers an approachable and adaptable platform for data analysis and is closely related to Apache Hadoop.

The capabilities of a data lake are essential for the low-cost archiving and storage of large volumes of raw data. They make data modelling and integration easier by supporting data transformations, handling different data types in the same repository, and using a "schema on read" methodology.

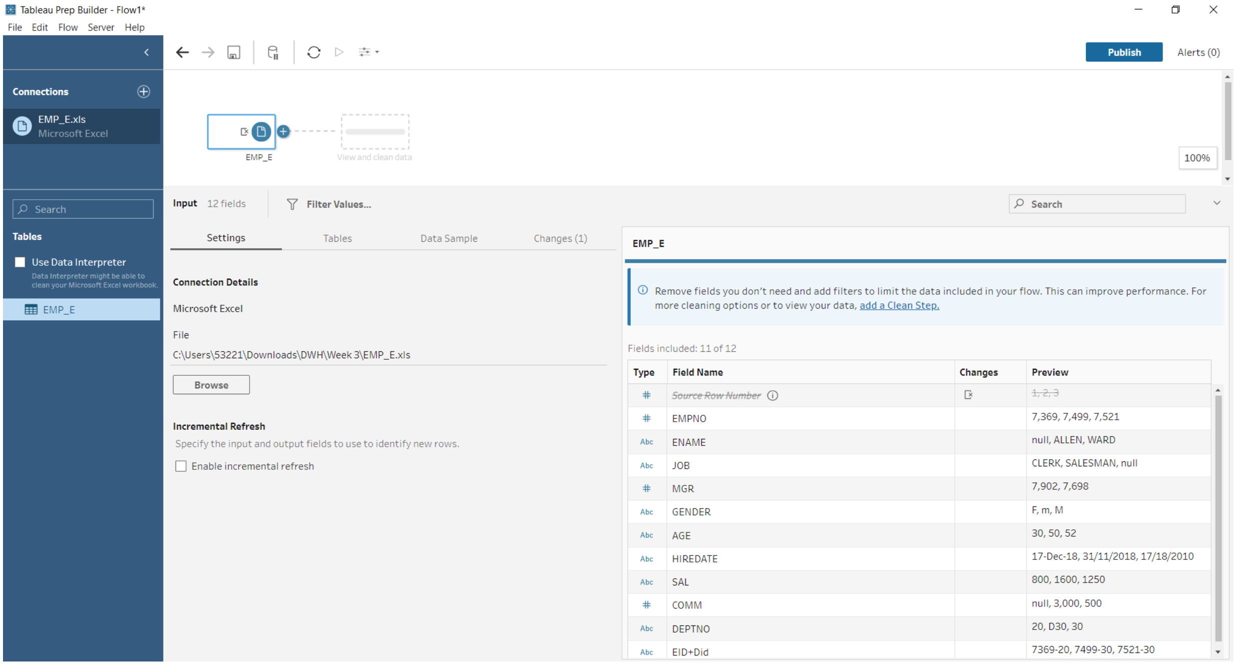
Advantages of Having No Schema: When handling raw, unstructured data, data lakes offer more flexibility by loading data without first defining its structure. There is support for non-traditional data types and speedier data loading. As needed, a schema might be created; this is known as "late binding."

The Comparison of Data Lakes and Data Warehouses: On the basis of workload, schema, scale, access, benefits, querying, data type, cost, and complexity, data lakes and enterprise data warehouses (EDWs) are contrasted. They have unique qualities and fulfil various functions. How Data Lakes Affect the Data Management Ecosystem: Data lakes provide low-level code languages and make use of programming frameworks like MapReduce, which have an impact on data management, especially in the extract-transform-load (ETL) process. Integration between data lakes and EDWs is necessary because processed data from the data lake is frequently transferred to a data warehouse for additional analysis.

Establishing a Data Lake: Information silos are addressed, and big data projects are supported by the implementation of data lakes. To reduce risks associated with data quality, lineage, security, access control, and performance, information governance is crucial. Organisational data strategy, data-driven culture, and IT infrastructure all have an impact on how mature a data lake becomes over time.

Advice for IT Executives: In the field of information management, data lakes are becoming more and more popular, especially in larger businesses. They have a lot of promise, but in order to properly extract value, they should be implemented with additional services and disciplines. IT executives should be aware of the competencies needed to manage data lakes and make sure their organisation has them.

Tutorial Week 4



Connection Details

Microsoft Excel

File

C:\Users\53221\Downloads\DWH\Week 3\EMP_E.xls

Incremental Refresh

Specify the input and output fields to use to identify new rows.

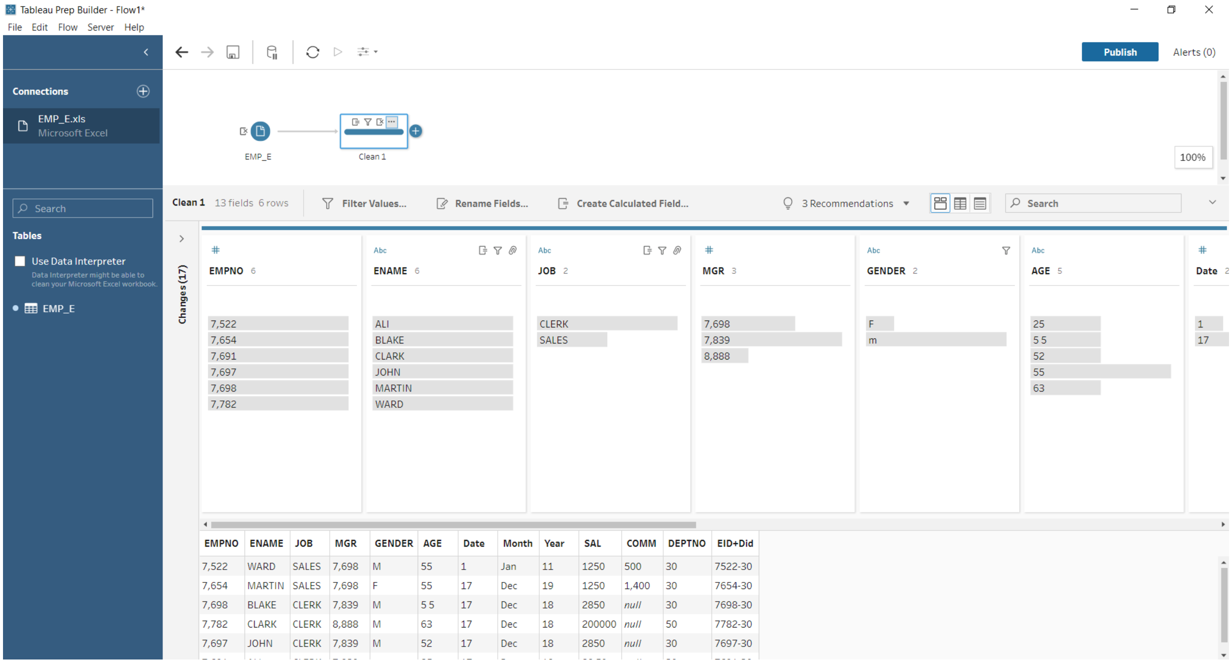
□ Enable incremental refresh

EMP_E

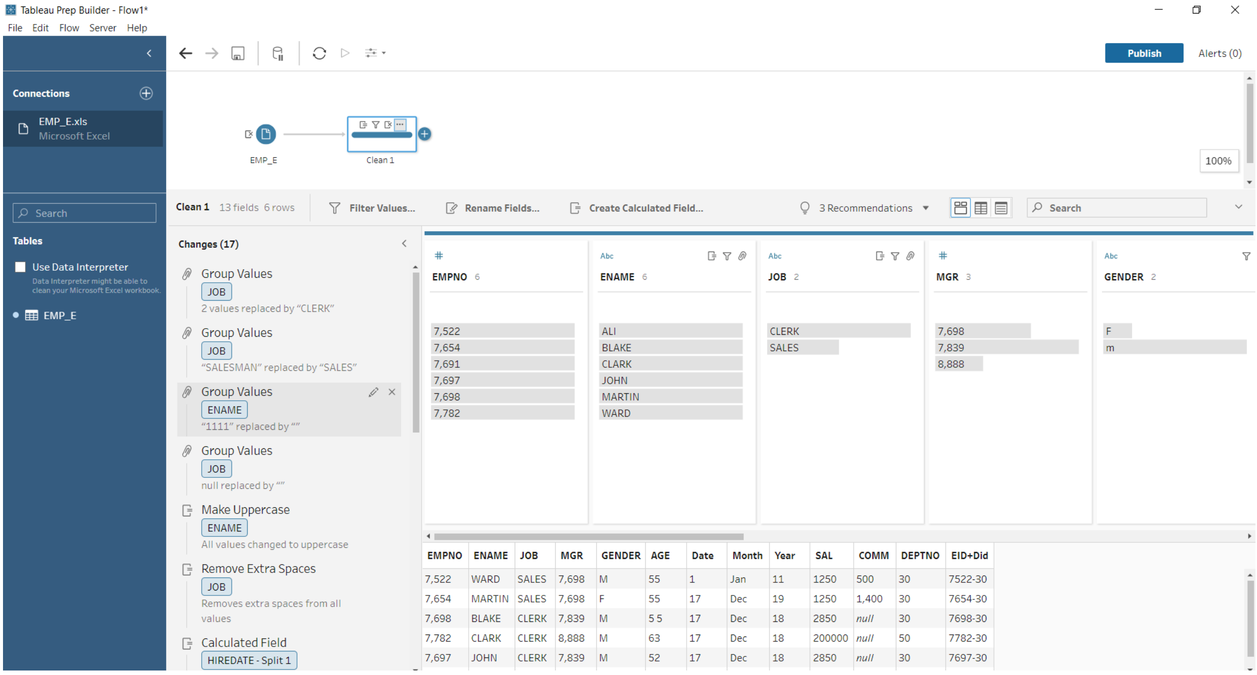
ⓘ Remove fields you don't need and add filters to limit the data included in your flow. This can improve performance. For more cleaning options or to view your data, add a Clean Step.

Fields included: 11 of 12

Type	Field Name	Changes	Preview
#	~~Source Row Number~~ ⓘ		~~1, 2, 3~~
#	EMPNO		7,369, 7,499, 7,521
Abc	ENAME		null, ALLEN, WARD
Abc	JOB		CLERK, SALESMAN, null
#	MGR		7,902, 7,698
Abc	GENDER		F, m, M
Abc	AGE		30, 50, 52
Abc	HIREDATE		17-Dec-18, 31/11/2018, 17/18/2010
Abc	SAL		800, 1600, 1250
#	COMM		null, 3,000, 500
Abc	DEPTNO		20, D30, 30
Abc	EID+Did		7369-20, 7499-30, 7521-30



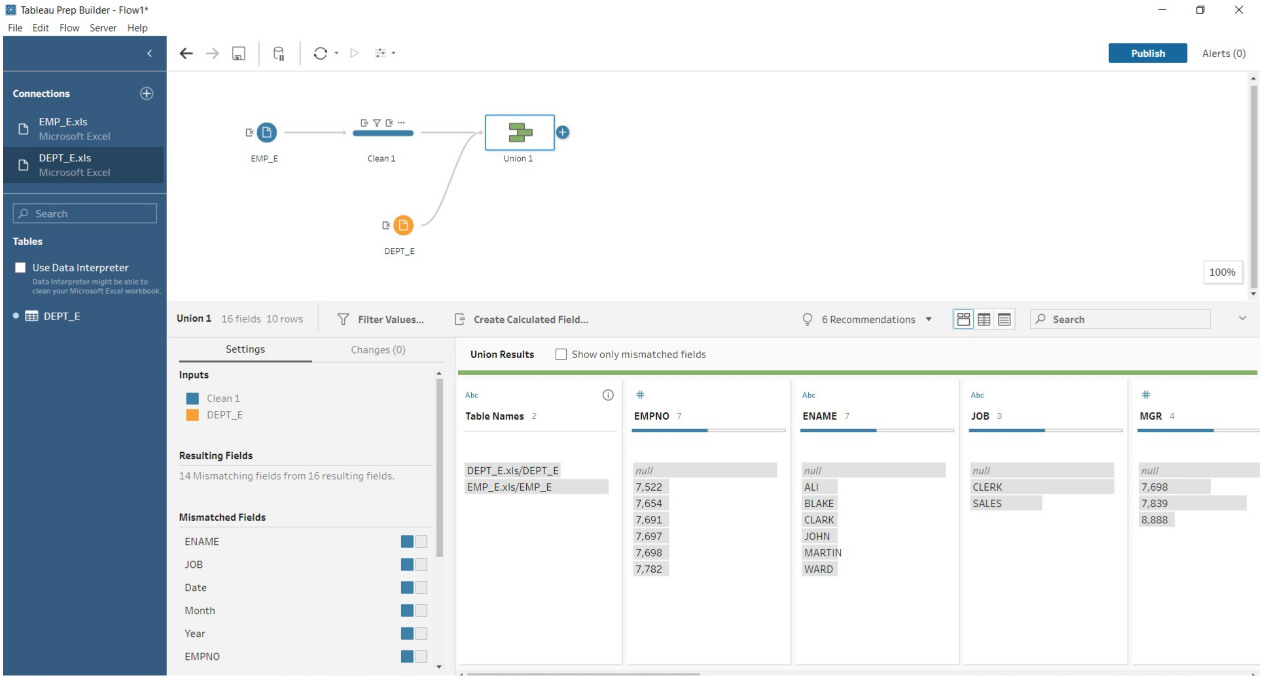
EMPNO	ENAME	JOB	MGR	GENDER	AGE	Date	Month	Year	SAL	COMM	DEPTNO	EID+Did
7,522	WARD	SALES	7,698	M	55	1	Jan	11	1250	500	30	7522-30
7,654	MARTIN	SALES	7,698	F	55	17	Dec	19	1250	1,400	30	7654-30
7,698	BLAKE	CLERK	7,839	M	5 5	17	Dec	18	2850	*null*	30	7698-30
7,782	CLARK	CLERK	8,888	M	63	17	Dec	18	200000	*null*	50	7782-30
7,697	JOHN	CLERK	7,839	M	52	17	Dec	18	2850	*null*	30	7697-30



Changes (17)

- Group Values
  JOB
  2 values replaced by "CLERK"
- Group Values
  JOB
  "SALESMAN" replaced by "SALES"
- Group Values
  ENAME
  "1111" replaced by ""
- Group Values
  JOB
  null replaced by ""
- Make Uppercase
  ENAME
  All values changed to uppercase
- Remove Extra Spaces
  JOB
  Removes extra spaces from all values
- Calculated Field
  HIREDATE - Split 1

EMPNO	ENAME	JOB	MGR	GENDER	AGE	Date	Month	Year	SAL	COMM	DEPTNO	EID+Did
7,522	WARD	SALES	7,698	M	55	1	Jan	11	1250	500	30	7522-30
7,654	MARTIN	SALES	7,698	F	55	17	Dec	19	1250	1,400	30	7654-30
7,698	BLAKE	CLERK	7,839	M	5 5	17	Dec	18	2850	*null*	30	7698-30
7,782	CLARK	CLERK	8,888	M	63	17	Dec	18	200000	*null*	50	7782-30
7,697	JOHN	CLERK	7,839	M	52	17	Dec	18	2850	*null*	30	7697-30



Inputs

- Clean 1
- DEPT_E

Resulting Fields

14 Mismatching fields from 16 resulting fields.

Mismatched Fields

- ENAME
- JOB
- Date
- Month
- Year
- EMPNO

References/bibliography

Kimball, R., Ross, M., Mundy, J., & Thornthwaite, W. (2011). "The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling." Wiley.

Inmon, W. H. (1996). "Building the Data Warehouse." Wiley.

Redman, T. C. (1997). "Data Quality: The Field Guide." Digital Press.

Wang, R. Y., & Strong, D. M. (1996). "Beyond Accuracy: What Data Quality Means to Data Consumers." Journal of Management Information Systems, 12(4), 5-33.

Wang, R. Y. (1998). "A Product Perspective on Total Data Quality Management." Communications of the ACM, 41(2), 58-65.