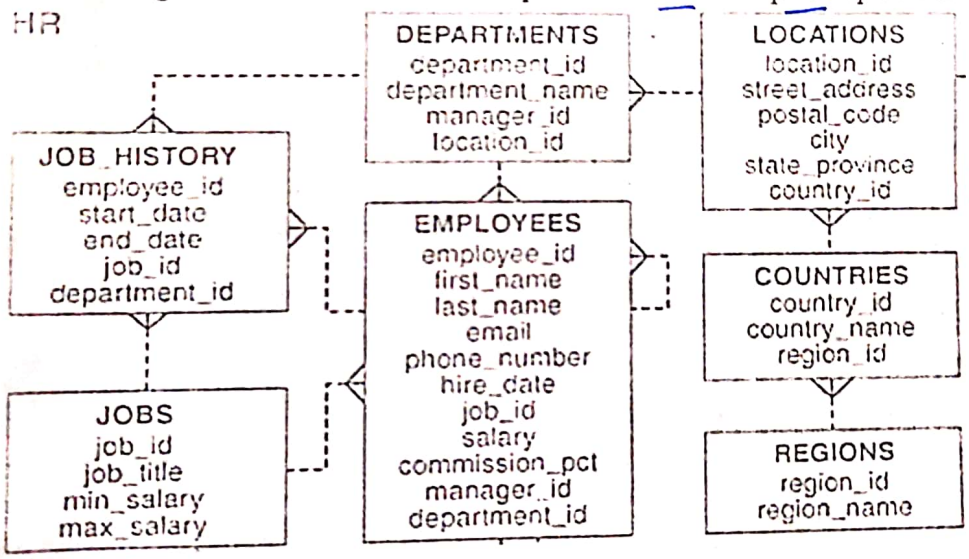
NED UNIVERSITY OF ENGINEERING & TECHNOLOGY

Midterm Examination (BSCS Final year) 2022

Data Warehouse and mining - (CT-463)

Instructions: Attempt all questions. Roll No 12

Q No. 1: Based on the following physical schema, design star schema which has minimum three dimension. using created star schema write queries for cube and pivot operation. (6)



Q No. 2: Why ERD is not suitable for data warehouse. Briefly define process of dimension model

Q No. 3: What is meant by metadata in the context of a Data warehouse? In what way ETL cycle can be used in typical data ware house. define with suitable example. (4)

Q No. 4: Write names of all De–normalization techniques. Based on given tables write queries using materialized view (any three techniques). (6)

Project

Project Code	Project Title	Project Manager	Project Budget
PC010	Pensions System	M Phillips	24500
PC045	Salaries System	H Martin	17400
PC064	HR System	K Lewis	12250

Project Team

Project Code	Employee No.	Hourly Rate
PC010	S10001	22.00
PC010	S10030	18.50
PC010	S21010	21.00
PC045	S10010	21.75
PC045	S10001	18.00
PC045	S31002	25.50
PC045	S13210	17.00
PC064	S31002	23.25
PC064	S21010	17.50
PC064	S10034	16.50

Employee

Employee No.	Employee Name	Department No. *
S10001	A Smith	L004
S10030	L Jones	L023
S21010	P Lewis	L004
S10010	B Jones	L004
S31002	T Gilbert	L023
S13210	W Richards	L008
S10034	B James	L0009

Department

Department No.	Department Name
L004	IT
L023	Pensions
L028	Database
L008	Salary
L009	HR

DWH Mid

CT-19012 Anabia Tahir

①

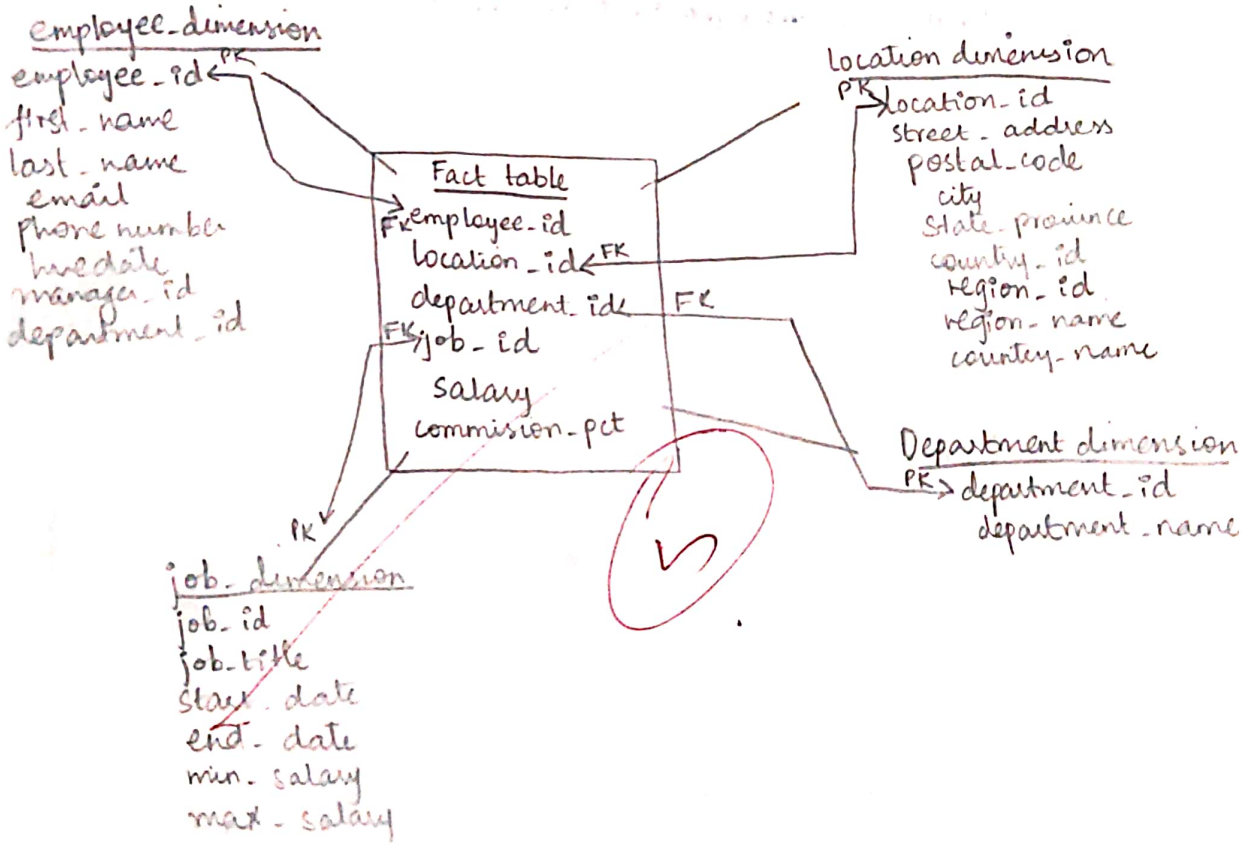
Star Schema:

① Identify business process:
The process is of employees, their job history and location and salaries management.

② identify facts:-
Salary, commision, and all dimension primary keys.

③ Identify ~~measures~~ dimensions:
job, employees, department, location.

④ Identify grains.
location_city, department_name, job_title



Cube operation:

```
select
  employee_id,
  department_id,
  location_id,
  job_id,
  sum (salary)
  sum (commision_pct)
  from fact table group by cube (employee_id, department_id,
                                 location_id), job_id)
```

Pivot operation :-

```
select * from (
 select
employee_id,
 location_id,
 job_id,
 department_id,
 ~~Sat~~ ~~Sun~~ salary, commission_pct
   from fact table) pivot
    sum(salary) for ~~employee~~ job_id ~~EIN~~ IN (-536, -533)
```

Entity relation diagram is not suitable for data warehouse ②
as number of tables and data is alot which makes the
ERD very complex and confusing. Furthermore ERD is used for
normalized data representation that is backed by the principle
of reducing storage space and using joins for ~~query~~ retrieval
of information. Whereas in data warehouse, analysis and retrieval of
data should take place within seconds. This is the reason why
redundant data needs to be added. However, redundant data isn't
supported by ERD modelling. ERD ~~also~~ shows relation b/w 2 or
more entities whereas in data warehousing we need multidimensio
-nal analysis for which dimension modelling suits best.

Process of Dimension Modelling:- (4)

① Identify business process.
In this step we need to understand that which ~~entity~~ attributes
are we measuring for e.g we need to do profitability analysis,
or fraud detection etc.

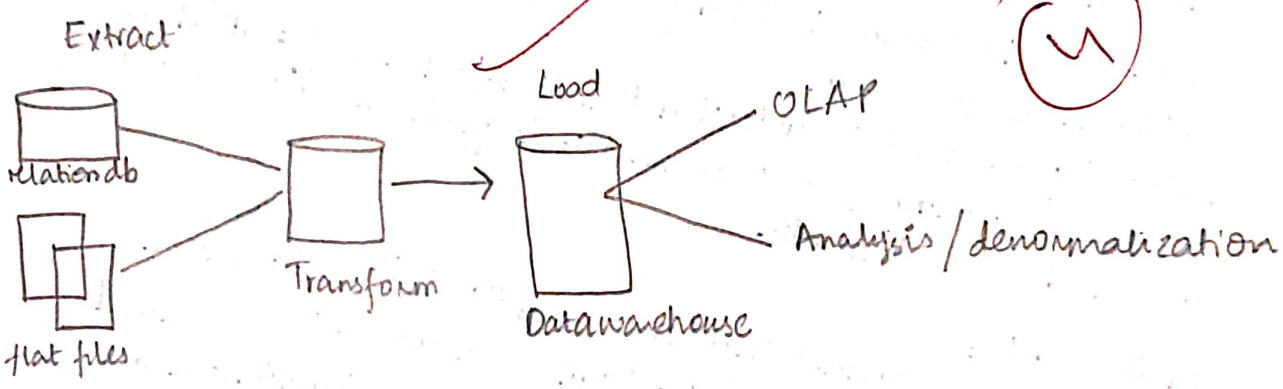
② Identify facts:
Facts are the measures that are quantitative, additive in
nature and are changing. Against these facts we check
our dimensions, to calculate the effect dimensions have
on facts.

③ Identify ~~measures~~ dimensions:-
~~Measures are the~~ Dimensions include location, time,
supplier, customer etc. Things that don't change or if so
they change, they do so gradually.

④ Identify grains:
Grain is the smallest unit of measurement in the
dimension modelling. E.g if we want to represent
location then the smallest grain we will use is street-
number. Higher granularity means more specifications.
and vice versa.

Qs.3. Metadata is the data about data. There are 3 types of metadeta. ① Structural ② Administrative ③ Descriptive.
In context of data warehouse it means that the data the warehouse ~~a~~ collects from various sources, the information about that is called metadata. Structural metadata is data stored in a structured form. Administrative metadata is used only by administrators. It may contain sensitive information. Descriptive metadata is data about data that is written in a comprehensive manner.

The ETL cycle stands for extract, transform and load.



The above figure is representative of how ETL cycle is used in a typical datawarehouse. For e.g an organization ~~collects~~ has multiple sources of information including its customer db, employee db and Sales db etc. Also all unstructured data from reviews to survey is extracted. Types of extraction include physical and logical. Then the company converts this data into a standard form and performs 5 steps of transformation including splitting / joining, summarization, collection, enrichment. This transforms the data into structured format ready to be used. The data is then loaded in the warehouse with any of the three techniques:
① full ② Incremental ③ Continuous / trickle. The history of data loaded onto the warehouse depends on the organization and its use case. If customer retention and product profitability is measured then the data is likely to be 5 years old or more.

CT-19012 ③

Denormalization is db optimization technique that is applied after normalization so that query retrieval is optimal while keeping storage as a tradeoff. There are quite a few denormalization techniques.

① Collapsing tables: It is applied on tables that either have one to one or many to many relation. And it is basically a join b/w the 2 tables.

② Splitting — Vertical: column wise splitting
— Horizontal: row wise splitting

6

③ Pre joining: Pre joining is again a joining, but b/w table that have a one to many relation. It is imperative that all columns are returned from both the tables.

④ Derived attributes: Are attributes that can be derived from one or more columns at run time but in denormalization they are calculated before hand.

) Redundant columns: Redundant columns which are mostly used for retrieval of information are used and ~~added~~ added in the table.

~~Splitting tables:~~
① Vertical split:
~~Collapsing tables:-~~

Queries:

Redundant column: (Adding hourly rate to project column)

```
create materialized view As(
Create table project_title_rate as (
Select p.*, project team hourly_rate from
project p, project team left join
p.project code = project team. project code))
```

Qs.3 Metadata ...

Splitting

horizontal:

```
create materialized view AS(
 Create table hourly_rate as(
 select project_code from Project team
  WHERE hourly_rate > 20.00))
```

vertical:-

```
create materialized view AS(
  Create table project_title AS(
  Select p.project code, p.Project title from project p))
```

Prejoining : (Joining project team and employee table)

```
create materialized view AS (
 create table employee_project as (
  select * from project team p left join employee.e
   on ~~employe~~ e.employee_number = p.employee_number))
```