COMM1822: Introduction to Databases for Business Analytics

Take-Home Exam Answer Sheet

Term 3 2020

By submitting your Take-Home Exam Answer sheet for assessment, you agree that:

- This is a Take-Home Exam, your responses are your own original work.
- You have attempted this Take-Home Exam by yourself without any help from others. Thus, you have NOT worked, collaborated or colluded with any other persons in the formulation of your responses. The work that you are submitting for your Take-Home Exam is your **OWN** work.
- In accordance with the Declaration form, this Take-Home Exam paper cannot be copied, forwarded or shared.
- Students are reminded of UNSW's rules regarding Academic Integrity and Plagiarism. Plagiarism is a serious breach of ethics at UNSW and is not taken lightly. For details see Examples of plagiarism.
- This Take-Home Exam is an open book/open web, further information is available "Here".
    - You are permitted to refer to your course notes, any materials provided by the course convenor or lecturer, books, journal articles, or other material/s from the Internet.
    - It is sufficient to use in-text citations that include the following information: the name of the author or authors; the year of publication; the page number (where the information/idea can be located on a particular page when directly quoted), For example, (McConville, 2011, p.188).
    - You are required to cite your sources and attribute direct quotes appropriately when using external sources (other than your course materials).
    - When citing Internet sources, please use the following format: website/page title and date.
    - If you provide in-text citations, you must provide a Reference List. The Reference list will **NOT** be counted towards your word limit.
- The work that you are submitting for your Take-Home Exam has **NOT** been previously submitted for academic credit in this, or any other course, at UNSW or elsewhere.
- The examiner may, for examination and grading purposes, reproduce and provide a copy of your submission to another member of the University.
- ***Turnitin*** may retain a copy of this Take-Home Exam submission on its database for the purpose of future plagiarism checking.
- You have **NOT** used third-party essay/writing services. These are **NOT** permitted by UNSW and where detected, your submission **WILL** be treated as plagiarism in accordance with applicable policies and procedures.
- Plagiarism detected in submitted work will be treated as a failure to act responsibly, ethically and with academic integrity and in accordance with student responsibilities found in Student Code of Conduct. The University may take disciplinary action under the Student Misconduct Procedures.
- Reproducing, publishing, posting, distributing or translating this Take-Home Exam is an infringement of copyright and will be referred to the Student Conduct & Integrity Unit for action.

By signing below, you acknowledge and agree with each of the above statements.

Student ID: _______z5183946________

Signature (this can be either your digital signature or simply type your Full Name:

______Yiyan Yang_____________________

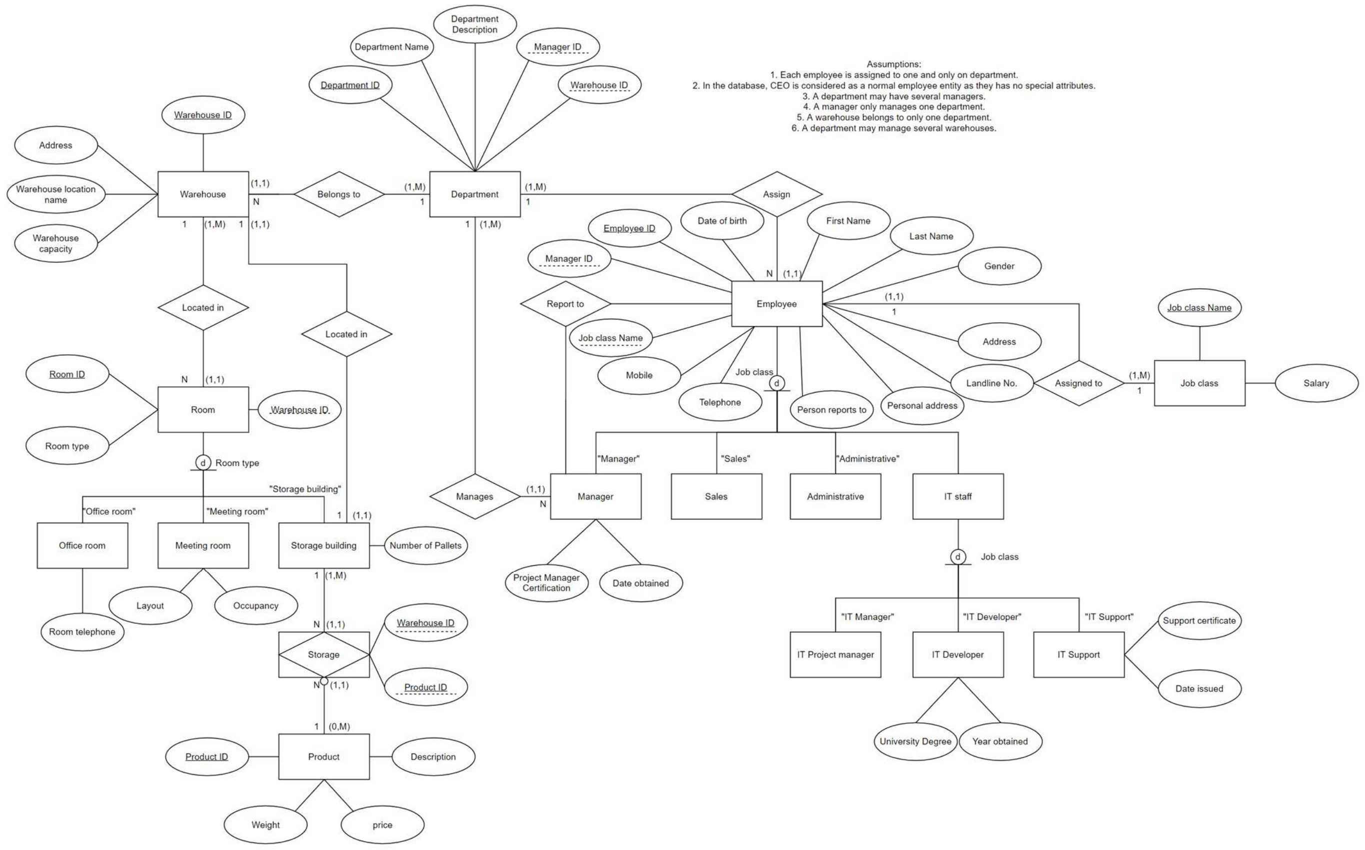
Question 1a (word count: 58)

Your answer:

Assumptions:

1. Each employee is assigned to one and only on department.
2. In the database, CEO is considered as a normal employee entity as they has no special attributes.
3. A department may have several managers.
4. A manager only manages one department.
5. A warehouse belongs to only one department.
6. A department may manage several warehouses.

Diagram:



Assumptions:
1. Each employee is assigned to one and only on department.
2. In the database, CEO is considered as a normal employee entity as they has no special attributes.
3. A department may have several managers.
4. A manager only manages one department.
5. A warehouse belongs to only one department.
6. A department may manage several warehouses.

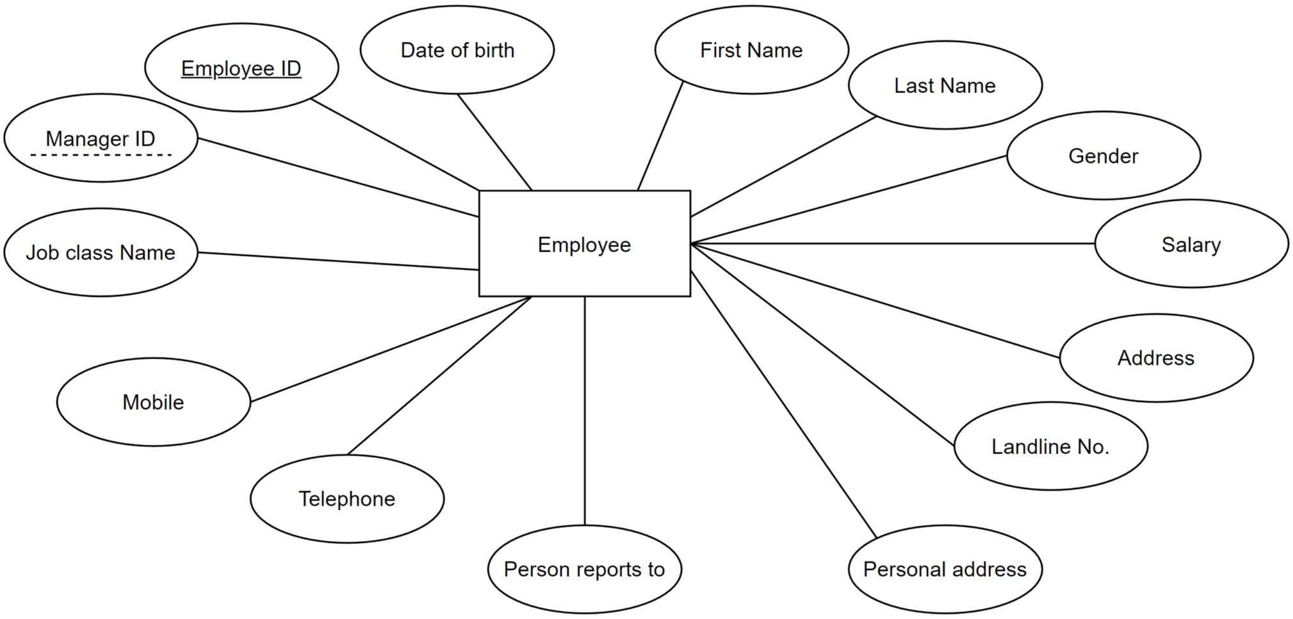
Question 1b (word count: 123)

Your answer:

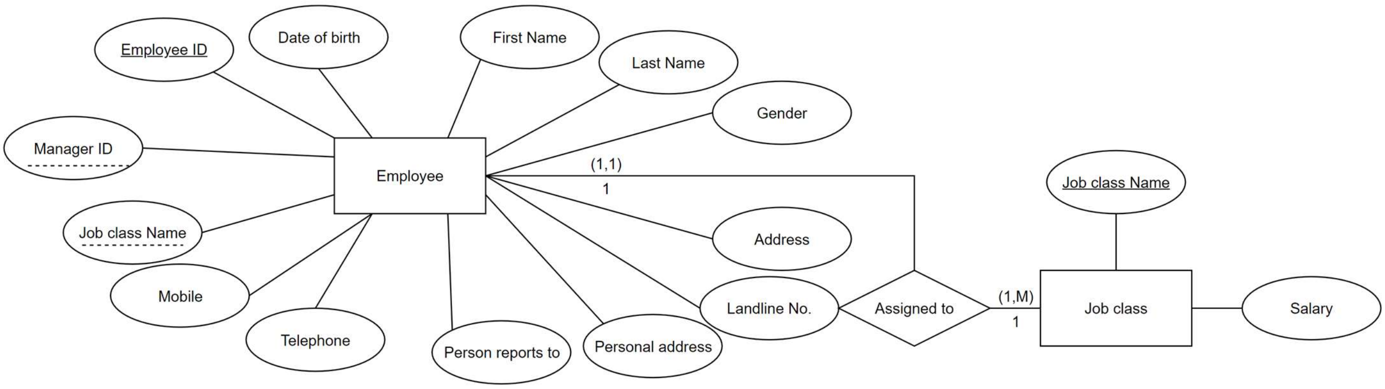
Example 1:

Employees that have the same job class may have different salaries, instead of paying equally to employees with same job class. Modification incurred is to make salary attribute belong to employees and delete job class entity as it doesn't hold any attributes, which means it's no need to keep an entity to record job class.

The diagram below is the resulting employee entity



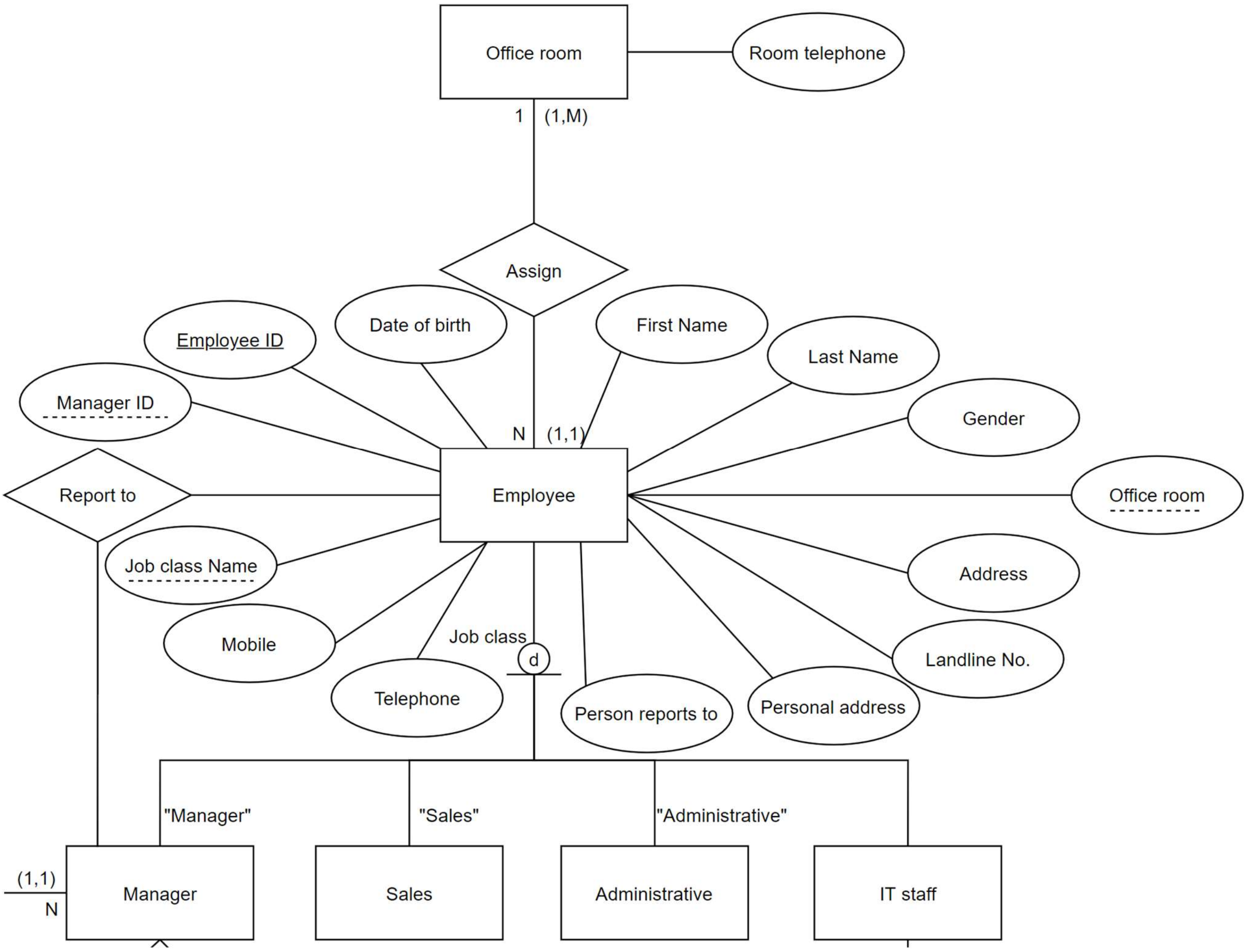
Which is different from the original ER diagram:



Example 2:

Some employees are assigned to office rooms, where each employee can only be assigned to up to one office room. Thus, an attribute must be created to record the assignment of office room and a relationship is needed between office room and employee.

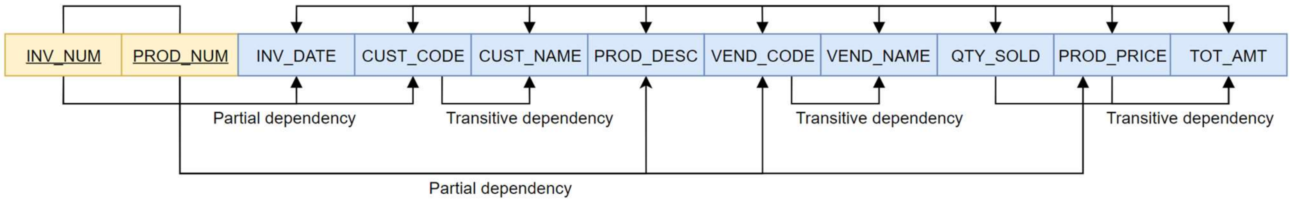
This is the resulting diagram:



Question 2a (word count: 51)

Your answer:

Primary key is labelled in yellow which are INV_NUM and PROD_NUM. Because union of the set of attributes determined by them is the whole set of the remaining keys while they cannot cover the whole set solely. Dependency graph is attached below with Partial dependencies and transitive dependencies labelled and listed.



1NF(INV_NUM, PROD_NUM, INV_DATE, CUST_CODE, CUST_NAME, PROD_DESC, VEND_OCDE, VEND_NAME, QTY_SOLD, PROD_PRICE, TOT_AMT)

Partial Dependencies:
(INV_NUM->INV_DATE, CUST_CODE)
(PROD_NUM->PROD_DESC, VEND_CODE, PROD_PRICE)

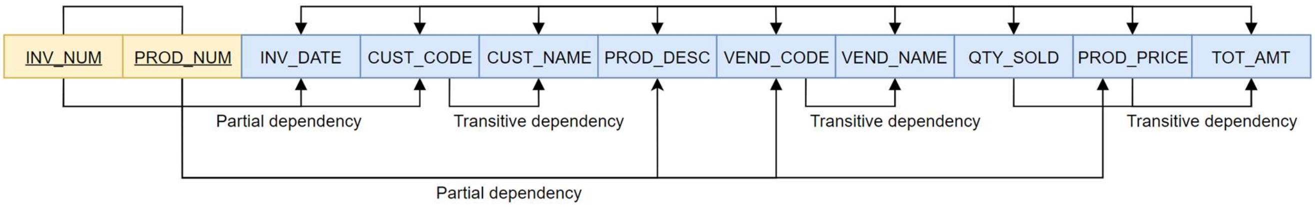
Transitive Dependency:
(CUST_CODE->CUST_NAME)
(VEND_CODE->VEND_NAME)
((QTY_SOLD, PROD_PRICE), TOT_AMT)

Question 2b (word count: 110)

Your answer:

1NF:

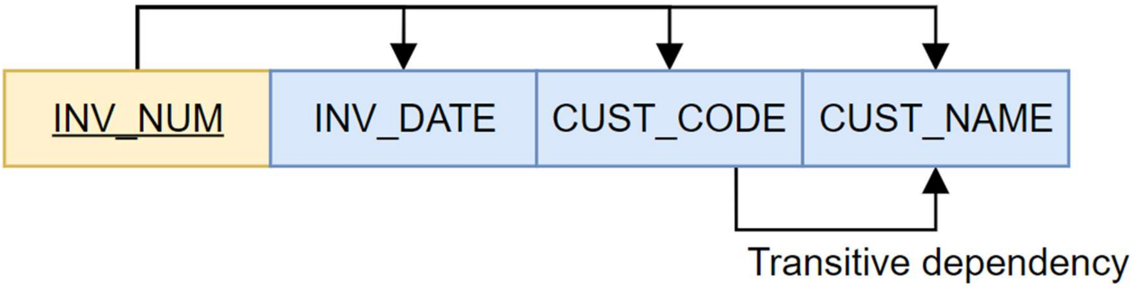
As shown is Q2a, the graph is:

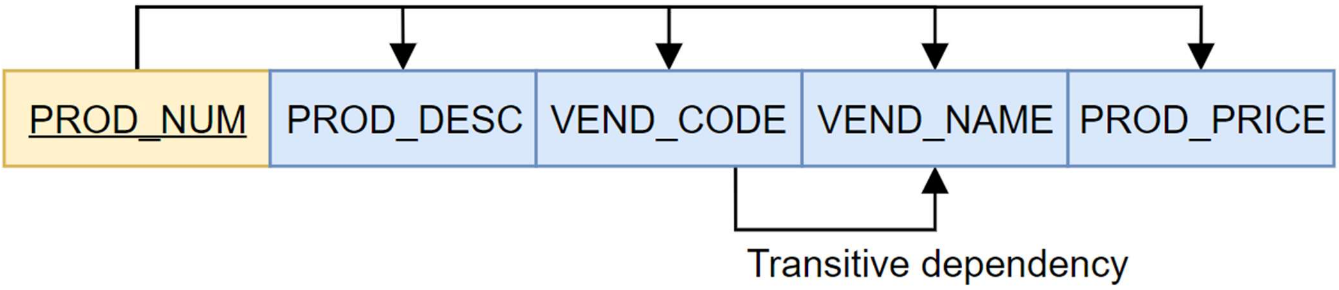


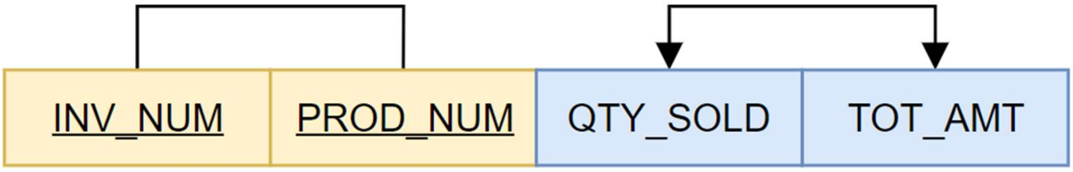
2NF:

1. Create a table for INV_NUM, INV_DATE, CUST_CODE and CUST_NAME with INV_NUM as PK.
2. Create a table for PROD_NUM, PROD_DESC, VEND_CODE, VEND_NAME, PROD_PRICE with PROD_NUM as PK.
3. Create a table for INV_NUM, PROD_NUM, QTY_SOLD and TOT_AMT with (INV_NUM and PROD_NUM) as a composite primary key.

The diagram is shown below:



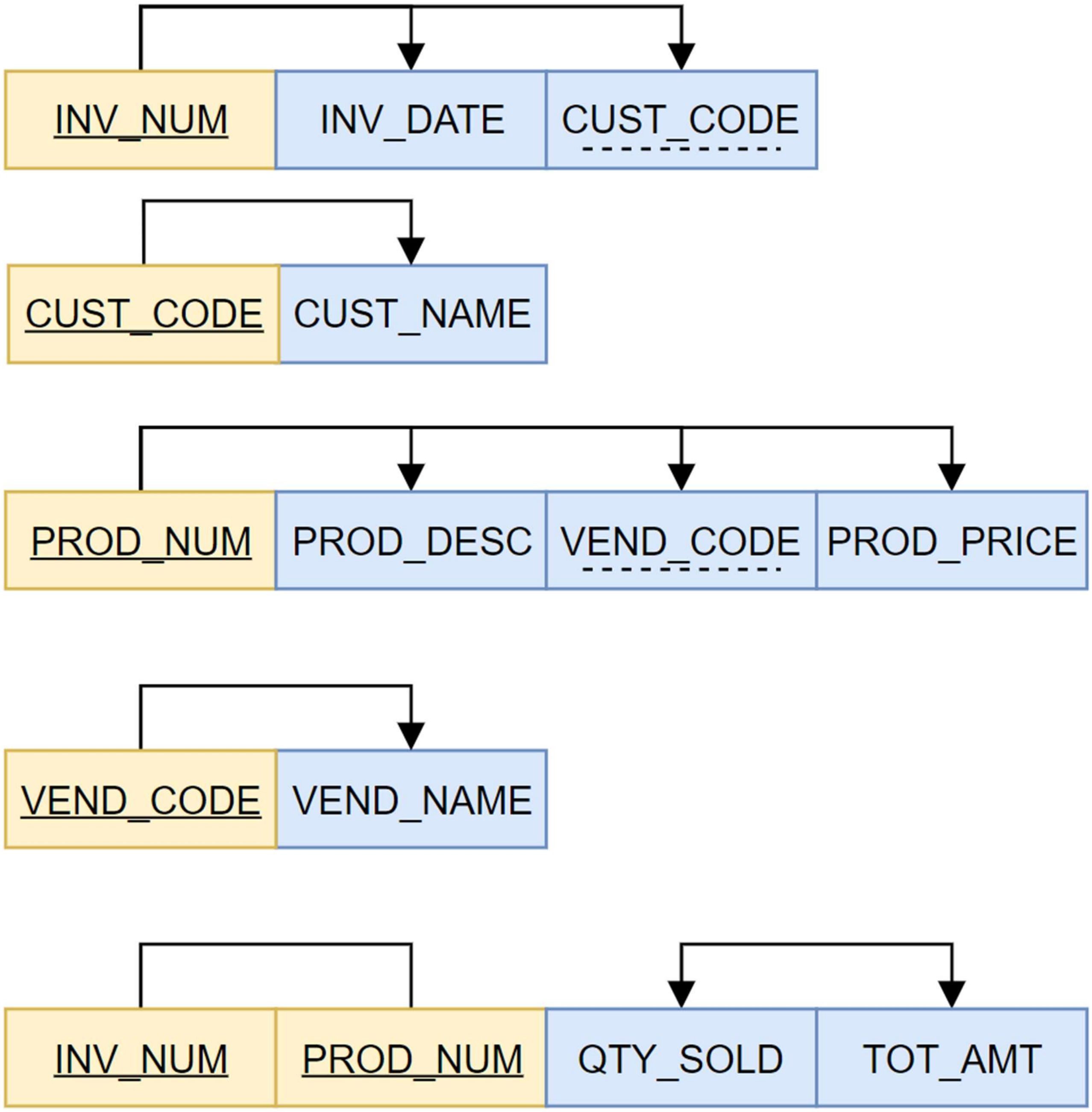




3NF:

1. Removed transitive dependency from Invoice table to form customer table containing CUST_CODE and CUST_NAME with CUST_CODE as PK.
2. Removed transitive dependency from production table to form vendor table containing VEND_CODE and VEND_NAME with VEND_CODE as PK.
3. Keep CUST_CODE as FK in Invoice table and VEND_CODE as FK in Product table.

The diagram is shown below:

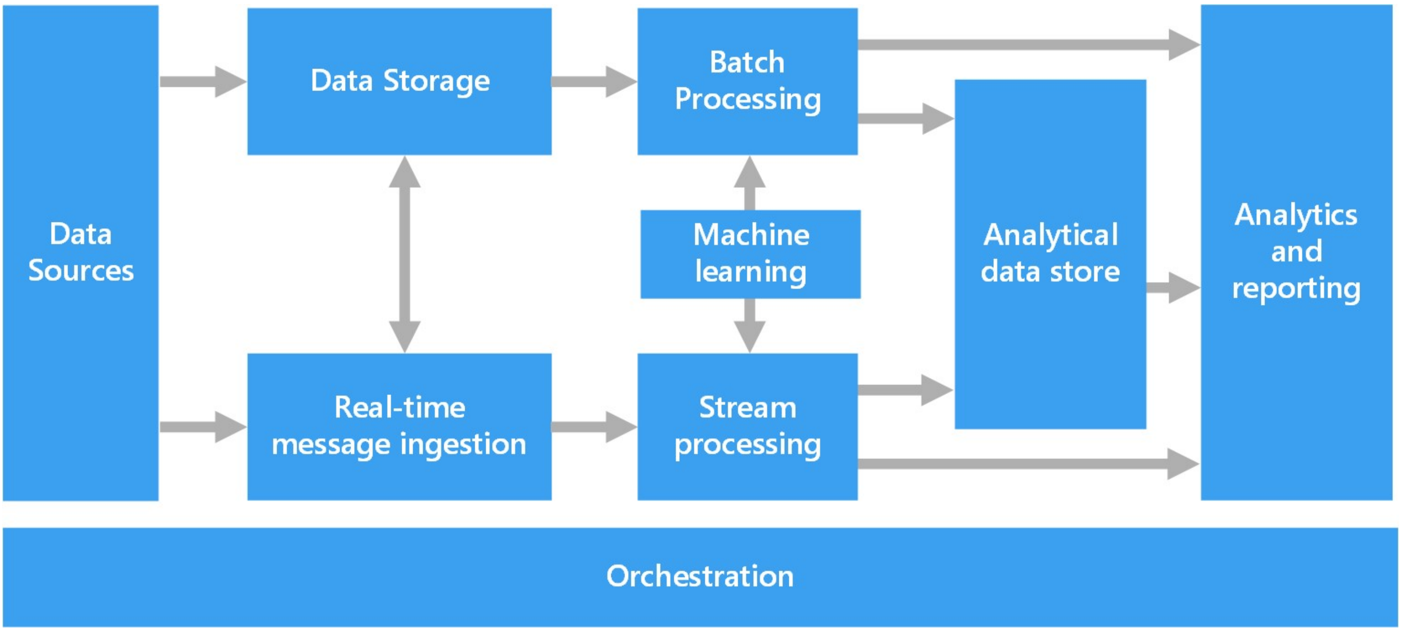


Question 3 (word count: 430)

Your answer:

As showed in one diagram in the big data architecture guide found in Microsoft Azure official document, there are five stages to handle the big data generated by users and other data sources such as web API and sensor:

1. Data collection
2. Data warehousing
3. Data processing
4. Analytical data organizing
5. Analytics and reporting



The first step, data collection is not complicated comparing to the following steps. The main purpose is to collect timely data from various sources that covers all the potential important information that are needed to perform analysis. This company may use APP to get essential user-specific information such as location, time and the riders' journey. And, API is a decent source of traffic and weather information while weather data can be obtained from sensors installed on the bikes as supplementary to APIs.

Data warehousing is the crucial step to build the information base that will be used for processing and analysing. Data lake is a distributed store which holds high volumes of large files in various formats for batch processing operations. Azure Data Lake Store is a practical and stable way to implement a data lake for start-ups. Because GPS and traffic information are real-time data that are generated frequently, Real-time message ingestion is also needed to be handled methodically.

There are two topics in data processing: batch processing and stream processing. As the scale of big data is too large for a single processing unit, long-running batches are needed to filter, aggregate and prepare data for analysis and Map/Reduce in Hadoop cluster is the most common way to implement batch processing. On the contrary, real-time streaming data is simpler so the computing capability is not the main issue, rather a message ingestion store is required to buffer the enormous amount of data generated, often IoT (Internet of Things) is applied to solve this task.

Usually, data for analysis are organised in a structured format. Relational data warehouse is the most traditional approach to provide the required organized data. Besides, NoSQL can guarantee low-latency access and interactive Hive database is able to organize a distributed data store.

Analysis is the ultimate purpose of Big Data and it is the final step of the flow of the data. Data modelling layer, visualization technologies and interactive data explorations are common ways to empower users to analyse the data where programming languages such as python and R, Excel and Spark are widely used.

Finally, orchestration is the process of transforming and passing the data between different steps. Automation is the main topic to elevate productivity and efficiency.

Diagram from:

https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/