Ch01

Fundamentals of Data Engineering

CHAPTER ASSIGNMENT

Name _____ Class _____ No. _____ Date _____

> Answer every question in the space given. Show your working. You may use the chapter page to help you.

Questions from *Data Engineering with Google Cloud Platform*, 1st edition, Adi Wijaya (Packt, 2022), and from the Guru Jazzy chapter page, for classroom use.

Exercise from the book

Ex You are a data engineer at a book publishing company and your product manager has asked you to build a dashboard to show the total revenue and customer satisfaction index in a single dashboard.

Your company doesn't have any data infrastructure yet, but you know that your company has these three applications that contain TBs of data:

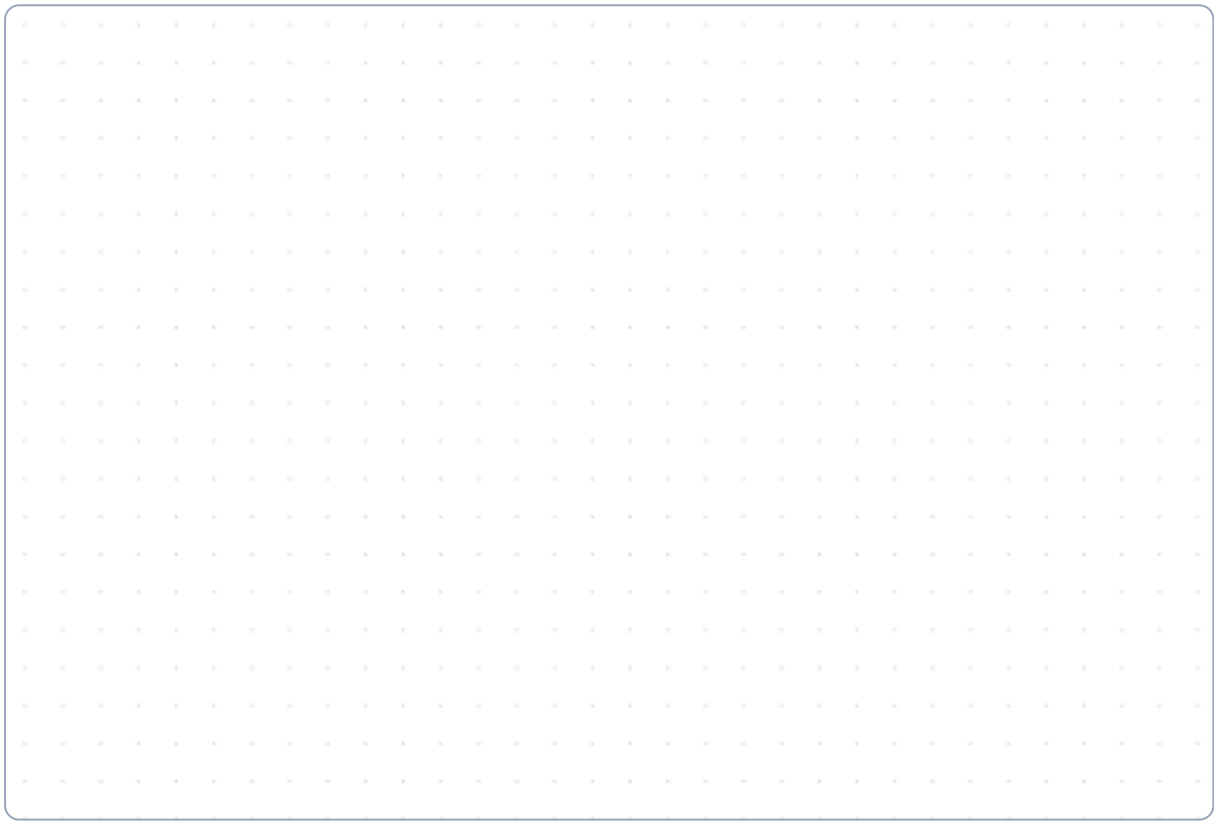
- The company website
- A book sales application using MongoDB to store sales transactions, including transactions, book ID, and author ID
- An author portal application using MySQL Database to store authors' personal information, including age

Do the following:

(1) List important follow-up questions for your manager.

(2) List your technical thinking process of how to do it at a high level.

(3) Draw a data pipeline architecture.



There is no right or wrong answer to this practice. The important thing is that you can imagine how the data flows from upstream to downstream, how it should be processed at each step, and finally, how you want to serve the information to end users.

Sorting activities from the chapter page

S1 **ETL or ELT?**

Options: **ETL** · **ELT**

No.	Scenario	Answer
1	A 200 GB MySQL export needs to reach BigQuery, and analysts will shape it with SQL afterwards.	
2	The target is a small reporting database that struggles with heavy joins and aggregations.	
3	Customer names and phone numbers must be masked before the data leaves the source network.	
4	IoT devices emit JSON constantly. You want to keep everything now and decide on a schema later.	

Quiz from the chapter page: Check yourself

1 A bank cannot tell how many credit-card customers also hold a mortgage. What is the root cause?

A. Its SQL is too slow

B. The data lives in separate systems (silos), so it must first be brought together

C. Credit-card data is unstructured

D. It has too few dashboards

---

2 Which statement best captures the key technical difference between a lake and a warehouse?

A. Lakes hold unstructured data and warehouses never can

B. Warehouses are always cheaper

C. A lake separates storage, compute and schema; a warehouse bundles them as one package

D. Lakes cannot run SQL

---

3 Your target is BigQuery and your source is a 200 GB MySQL dump. Which pattern is the sensible default?

A. ETL, transforming on a small VM

B. ELT, loading raw and transforming inside BigQuery

C. Neither, because BigQuery cannot transform data

D. ETL, because ELT is only for lakes

---

4. Is 50 GB of tweets "big data"?

   A. Yes, because it is social media

   B. Yes, because it is unstructured

   C. Only if it satisfies all five Vs

   D. Probably not. Big is relative to your system, and 50 GB fits comfortably on one machine

---

5. A 1.15 GB file is stored in 128 MB chunks. About how many chunks, and what records where they live?

   A. 9 chunks, tracked in metadata

   B. 10 chunks, tracked in metadata

   C. 10 chunks, tracked by the compute engine

   D. 128 chunks, tracked in metadata

End of assignment · Check your answers against the chapter at gurujazzy.com/subjects/gcp-data-engineering/chapter-01-fundamentals-of-data-engineering/