Part 1: Answer the questions

Engineering Basic:

OS

1. How many CPUs and RAM in VM
2. Find which directory is using most of disk storage
3. Change owner and permission of directory
4. Show process name “airflow”

GIT

1. Change branch into development
2. Exclude all files under /config and /vendors
3. Commit change with “test exclude”
4. Push the change to /features/de-test

Programming Language

1. Rank and give the rating of your top 3 favourite programming languages (5 highest - 1 lowest)
2. What is the difference between list and tuples in python ?
3. Write a program that prints the numbers from 1 to 100.
4. For multiples of three print “Fizz” instead of the number and for the multiples of five print “Buzz”. For numbers which are multiples of both three and five print “FizzBuzz

Code:

Data Engineering

1. What’s the difference between data lake and data warehouse ?
2. What tools do you use to build the whole pipeline ?
3. How do you monitor data pipeline jobs ?
4. How do you manage secrets (password/credential) being seen from people outside your team ?
5. How would you design and implement data pipeline if requirement is load data from gbq to redshift
6. What would you do if you get requirement to add encrypt citizen_id to the data warehouse

Part 2: DE Exercise:

Please code by do following below and sent code by GitHub

1. Load and transform data from CSV file to RDS
2. The target table (total_netsale) will have table structure following below

column_name	data type
customer_id	Integer
first_name	Varchar(300)
last_name	Varchar(300)
total_sale_thb	Float
shipping_thb	Float
tax_thb	Float
created_date	Datetime (UTC+7)
updated_date	Datetime (UTC+7)

3. Show the sum of total_sale (total field), shipping and tax in THB to total_netsale table
4. Add created_date and updated_date column into target table (total_netsale)
5. Load product.csv into RDS by removing "^un^" in name field

*** Please send the answer sheet in (pdf file) for Part1 and upload code to GitHub for Part2 ***