28:15

Data Engineering Assessment Questions - External

Multiple Choice

2. Which statement below accurately describes a star schema?

- ○ The (*) result of joining all fields selected from related tables
- ○ An appropriately indexed relational database
- ◉ A database schema centered around a core fact table and surrounding dimensional lookup tables
- ○ Any schema design in which all tables are related with at least one other table

3. Your friend asks you to define the natural primary key of your table. What is she referring to?

- ○ The system-generated key that is used to index the data in a table
- ◉ The field or fields that can uniquely identify rows in that table
- ○ The ID field to which all other dimension tables join
- ○ The constraint that ensures referential integrity in your schema design

4. Which is a characteristic of non-relational data?

- ○ Forced schema on data structures
- ○ Each row has the exact same columns
- ◉ Flexible storage of ingested data

- ◯ All of the above

5. Your VP keeps saying "orchestration layer." What is she talking about?

- ◯ A layer of code that facilitates open source data processing
- ◉ The automated configuration and management of systems and applications
- ◯ A systematic approach to deal with duplicates and data redundancy
- ◯ None of these

6. You're working in a git repo with two branches: main and develop. You're currently in the main branch. You want to (1) create a new feature branch off develop and (2) begin working from it. Which of the following commands would you use?

- ◉ git checkout develop; git checkout -b feature/my-feature
- ◯ git switch develop; git branch feature/my-feature
- ◯ git branch feature/my-feature develop; git start feature/my-feature
- ◯ All of the above will work

7. Your python code

```
items = {"Coffee": 2.2, "Tea": 1.5, "Chocolate": 2.5}
for item in items.keys()
  income = 0
  qty = input(f"How many {item}s have you sold?")
  income = income + qty * items[item]
  print(f"\nThe income today was {income:0.2f}")
```

Produces the following error

```
File "<path>/<filename>.py", line 3
for item in items.keys()
```

```
                ^
SyntaxError: invalid syntax
```

What's most likely the reason for the error?

- ○ You forgot a semicolon
- ◉ You forgot a colon
- ○ You forgot a comma
- ○ None of these

8. Which of the following is NOT an advantage of a publish/subscribe data model?

- ○ Loose coupling between software components
- ○ Near-real-time data transfer
- ○ Avoids polling
- ◉ Avoids concurrency problems

9. Which of the following is valid JSON?

- ○ {"benefits": ["plan1": {"ben_pkg_cd": "A666"}, "plan2": {"ben_pkg_cd": "C222"}}
- ◉ {"benefits": {"plan1": {"ben_pkg_cd": "A666"}, "plan2": {"ben_pkg_cd": "C222"}}}
- ○ {benefits: {plan1: {ben_pkg_cd: "A666"}, plan2: {ben_pkg_cd: "C222"}}}
- ○ {'benefits': [{'ben_pkg_cd': 'A666'}, {'ben_pkg_cd': 'C222'}]}

10. What is printed by the following python code snippet?

```
class MyClass(str):
```

```
    def join(self, x):
        return '+'.join([self + i for i in x])

    def output(self):
        print(self.join(['b','c']) + super().join(['b','c']))

t = MyClass('a')
t.output()
```

- ( ) ab+acbac
- (•) a+b+c+abc
- (•) a+b+c+abc
- ( ) bca+bac

11. Objects in which things about data should be captured and stored are called: _____.

- ( ) Tables
- (•) Entities
- ( ) Rows
- ( ) Columns

24:35

Data Engineering Assessment Questions - External

Suggest / write a question

12. Please enter one new question- and answer- that our leadership should consider adding to our Data Engineering assessment below. Include a brief comment describing why you feel this question would be helpful / useful.

I was tasked to designn a data pipeline to handle high-volume, real-time data ingestion from multiple sources, including IoT devices and web logs. The data was supposed to be processed, enriched, and stored in a data warehouse for further analysis. The pipeline was must ensure low-latency processing, handle schema evolution gracefully, and provide fault tolerance.

To design this data pipeline on Azure with Databricks I took a cloud-native, scalable, and resilient approach. for data ingestion, I used Azure Event Hubs nad Azure IoT Hub to ingest data streams from multiple sources, including IoT devices and web logs. Azure Event Hubs provides a high-throughput, low-latency event ingestion platform by which I can easily handle real-time data from diverse sources and scale as needed.
For real-time stream processing and enrichment, I used Azure Databricks with Apache Spark Structured Streaming. Databricks provided me a unified analytics

01:59

Data Engineering Assessment Questions - External

Written Responses

Please select **ONE** of the following questions to complete. Email your responses as attachments to your recruiter or interviewer.

13. Given the CMS provider data metastore, write a script that downloads all data sets related to the theme "Hospitals".

    The column names in the csv headers are currently in mixed case with spaces and special characters. Convert all column names to snake_case (Example: "Patients' rating of the facility linear mean score" becomes "patients_rating_of_the_facility_linear_mean_score").

    The csv files should be downloaded and processed in parallel, and the job should be designed to run every day, but only download files that have been modified since the previous run (need to track runs/metadata).

    Please email your code and a sample of your output to your recruiter or interviewer. Add any additional comments or description below.

    https://data.cms.gov/provider-data/api/1/metastore/schemas/dataset/items

```
import os
import re
import requests
import pandas as pd
from datetime import datetime
from concurrent.futures import ThreadPoolExecutor
import smtplib
from email.mime.text import MIMEText
from email.mime.multipart import MIMEMultipart

def convert_To_Snakecases(column_name):
    column_name = re.sub(r'[^a-zA-Z0-9\s]', '', column_name)  # Remove
special characters
    column_name = column_name.strip()  # Remove any leading/trailing spaces
    column_name = re.sub(r'\s+', '_', column_name)  # Replace spaces with
underscores
    return column_name.lower()

def downLoad_PRocess_CSV(dataset_url, output_dir, last_run_time):
    response = requests.head(dataset_url)
    file_modified_time = response.headers.get('Last-Modified')

    if file_modified_time:
        file_modified_time = datetime.strptime(file_modified_time, '%a, %d %b
%Y %H:%M:%S GMT')
        if file_modified_time > last_run_time:
            response = requests.get(dataset_url)
            if response.status_code == 200:

                df = pd.read_csv(pd.compat.StringIO(response.text))
                df.columns = [convert_To_Snakecases(col) for col in df.columns]

                filename = os.path.join(output_dir, dataset_url.split('/')[-1])
                df.to_csv(filename, index=False)
                print(f"Downloaded and processed {dataset_url}")
                return filename
            else:
                print(f"Failed to download {dataset_url}")
        else:
            print(f"No new data for {dataset_url}")
    return None

def send_Email(subject, body, to_email):
    from_email = 'your-email@example.com'
    password = 'your-email-password'

    msg = MIMEMultipart()
    msg['From'] = from_email
    msg['To'] = to_email
    msg['Subject'] = subject
    msg.attach(MIMEText(body, 'plain'))
```

```
    server = smtplib.SMTP('smtp.example.com', 587)
    server.starttls()
    server.login(from_email, password)
    text = msg.as_string()
    server.sendmail(from_email, to_email, text)
    server.quit()

def read_last_runTime(metadata_file):
    if os.path.exists(metadata_file):
        with open(metadata_file, 'r') as f:
            last_run_time = datetime.strptime(f.read().strip(), '%Y-%m-%d
%H:%M:%S')
    else:
        last_run_time = datetime(1970, 1, 1)  # If no metadata, assume the first
run
    return last_run_time


def update_last_runTime(metadata_file):
    with open(metadata_file, 'w') as f:
        f.write(datetime.now().strftime('%Y-%m-%d %H:%M:%S'))


def main():
    output_dir = './hospital_data'
    metadata_file = './last_run_time.txt'

    os.makedirs(output_dir, exist_ok=True)


    last_run_time = read_last_runTime(metadata_file)

    datasetPath = [ #////SOME PATH GOES HERE
    ]

    with ThreadPoolExecutor(max_workers=5) as executor:
        futures = [executor.submit(downLoad_PRocess_CSV, url, output_dir,
last_run_time) for url in datasetPath]
        results = [future.result() for future in futures]

    processed_files = [result for result in results if result]
    if processed_files:
        body = f"Downloaded and processed the following files:\n\n" +
"\n".join(processed_files)
        send_Email('Hospital Data Download Complete', body,
```

14. Your product owner has prioritized an epic "SMART on FHIR server integration for our Azure instance." The below context was provided :

Substitutable Medical Applications and Reusable Technologies (SMART on FHIR) is a healthcare standard through which applications can access clinical information through a data store. It adds a security layer based on open standards including OAuth2 and OpenID Connect, to FHIR interfaces to enable integration with EHR systems. Using SMART on FHIR provides at least three important benefits: Applications have a known method for obtaining authentication/authorization to a FHIR repository; Users accessing a FHIR repository with SMART on FHIR are restricted to resources associated with the user, rather than having access to all data in the repository; Users have the ability to grant applications access to a limited set of their data by using SMART clinical scopes. We need to create an integration with our Azure instance to realize these benefits.

As tech lead, please decompose this epic into development stories or tasks for sprint planning (you can do this in excel, send a picture of post-it notes, or send visuals from Jira/ any other tool you prefer).

Please email your response, and enter any comments below.

Story 1: Conduct research on SMART on FHIR and its components.Story 2: Define

15. Your friend is starting a restaurant business. She has ambitions to add franchise locations, and food trucks. She is "data-hungry" and wants to use data to support her operations / inform her strategy. Create a data model for her company (you can make up the details). At minimum, please include a logical and a physical model (you get to make up the tables, relationships, keys, etc).

    NOTE: if you use visio, please convert to a PDF.

    Please email your response, and enter any comments below.

    Tables Columns Restaurant restaurant_id (Primary Key) name

You can print a copy of your answer after you submit

Microsoft 365