Week 2 Quiz

1. Which of the following statements is true about the transformation stage of the data engineering lifecycle?

   - The transformation stage of the data engineering lifecycle is made up of two parts: queries and transformation.

   - <mark>Data transformation is the "turn data into something useful" stage of the data engineering lifecycle.</mark>

   - Data transformation only occurs after data has been ingested and before it is stored.

   Correct: That's right. The transformation phase of the data engineering lifecycle is where you as a data engineer really start to add value to data by transforming it into a useful form that serves downstream use cases.

2. Which of the following statements are true about data generation and source systems? Select all that apply.

   - <mark>As a data engineer, you will typically not create or maintain the source systems from which you ingest data.</mark>

     Correct: In general, source systems are created and maintained by upstream stakeholders so they're typically out of your control.

   - <mark>Databases are the most common source systems you'll interact with.</mark>

     Correct: The most common source systems are databases, which could be a relational database or other type of NoSQL systems.

   - <mark>As a data engineer, you'll be most successful if you work directly with source system owners to understand how those systems work.</mark>

     Correct: That's right. You'll be most successful if you work with the source system owner to understand how those systems generate data, how the data might change over time, and how those changes will impact the downstream systems you build.

   - As a data engineer, you are typically responsible for setting up a source system and generating data befo re the ingestion stage.

3. Which of the following are examples of storage abstractions?
   - <mark>Data Lake</mark>
   - Relational Database
   - <mark>Data Warehouse</mark>
   - Object Storage

4. Imagine you are a data engineer at a company that provides online courses through a mobile app. Which of the following examples represents an embedded analytics use case?

- A real-time dashboard that tracks critical performance metrics in the mobile app - like the load time for course content pages and videos - and sends alerts to engineers when errors occur or app features are not performing as expected.

- An internal dashboard showing course enrollments and ratings.

- A recommendation engine that provides learners with suggestions on the next course they should take.

- ==A user-facing dashboard that shows learners how many courses they have completed and the total time they have spent learning each week.==

5. When it comes to the security of your data systems, what does the principle of least privilege imply?

- ==You give users or applications access to only the essential data and resources to perform their job or int ended function, and only for the duration that is required.==

- Only the most senior members of the team should be given admin access to systems, while more junior members should be given restricted access.

- You should always ingest sensitive data, but only allow access to sensitive data to those individuals or applications that require it.

- Only the data engineers should operate as "admin", "superuser" or from the root shell.

6. Which of the following statements are true about magnetic disk drives, solid-state drives, and Random Access Memory (RAM), which are some of the raw ingredients of storage systems? Select all that apply.

- Solid-state drives typically have faster read and write speeds than RAM.

- Magnetic disk drives are not used anymore in modern storage systems.

- ==Magnetic disk drives are significantly cheaper than solid-state drives.==
- ==RAM is volatile, meaning that if your system loses power, data stored in RAM is typically lost very quickly.==

7. True or False: data engineers need to know how to code.

- ==True==

- False

8. For which of the following use cases would you consider streaming ingestion over batch ingestion? Select all that apply.

- To serve data to a data analyst who wants to create dashboards showing online user activity on the company's website within a second or less after the activity data has been recorded.

  Correct: Since the data has to be shown on the dashboard shortly after it is produced, this is a good use case for streaming ingestion.

- In a pipeline to serve data to data analysts who are interested in analyzing the company's weekly sales. Each Monday, the data analysts would like to look at the sales from the previous week.

- To serve data for an operational analytics use case, where instantaneous alerts must be issued if a website goes down or a live product feature stops working.

  Correct: Operational analytics is typically about monitoring real-time data for immediate action, so this is a good use case for streaming ingestion.

- To serve data for a fraud detection system in bank transactions, where fraudulent transactions or attempted transactions must be detected and mitigated immediately while they are underway.

  Correct: Since the transaction fraud has to be detected and dealt with as soon as possible after it occurs, this is a good use case for streaming ingestion.

- To serve data to a data scientist who is interested in training a machine learning model. The data scientist would like to use historical data to train the model.

9. What are the three pillars of DataOps?

- Orchestration, Version Control, and Incident Response

- Automation, Observability, and Monitoring

- Automation, Observability & Monitoring, and Incident Response

- Automation, Version Control, and Data Quality

10. In the lab this week, you used AWS Glue, which is an ETL service that helps you prepare and integrate data from multiple sources. Which stage(s) of the data engineering lifecycle did you implement using AWS Glue? Select all that apply.

- Ingestion

- Data generation in source systems

- Transformation

- Storage

11. In the lab, one of the AWS services you used was Amazon Athena. Which of the following statements best describes this service?
    - Amazon Athena is a relational database service for normalized data.
    - Amazon Athena is a query service that allows you to directly query data from Amazon S3 using SQL queries.
    - Amazon Athena is a visualization tool that allows you to create Business Intelligence dashboards.
    - Amazon Athena is a type of object storage system.