Data Journey

1. Machine learning pipelines for production have become prominent in several industries. They introduce complexity to the ML lifecycle due to the large amount of data, tools, and workflows involved. If data and models are not tracked properly during the life cycle, it becomes infeasible to recreate an ML model from scratch or to explain to stakeholders how it was created. What can be done to prevent these shortcomings?

   - Debug the model so it becomes reliable.
   - Use a relational database.
   - Establish data and model provenance tracking mechanisms.
   - Use a TFX publisher

   Provenance will track the chain of artifacts and transformations at play in a given pipeline.

2. ML Metadata (MLMD) is a library for recording and retrieving metadata associated with ML production pipelines among other applications. What is the definition of attribution in this library?

   - Is a record of the relationship between artifacts and executions.
   - Is a record of the relationship between artifacts and contexts.
   - Is a record of a component run or a step in an ML workflow and the runtime parameters
   - Is a record of the relationship between executions and contexts.

3. Every run of a production ML pipeline generates metadata about its components, their executions (e.g. training runs), and the resulting artifacts (e.g. trained models). **ML metadata (MLMD)** registers this information in a database called the **Metadata Store**. The *MetaDataStore* object receives a connection configuration that corresponds to the storage backend used. Which of the following configurations will you use for fast experimentation and local runs??

   - Fake Database
   - MongoDB
   - SQL
   - MySQL

   This provides a fast in-memory DB that can be easily destroyed after experimentation.