SFL Scientific Questions

# What is a Data Lake? Explain its benefits, how it differs from a data warehouse, and how it might benefit a client.

Data lakes are large repositories for unstructured or semi-structured data. They are great to store things like logs, raw IoT data, images, historical archived data, and any other information you may want to use in the future. They are low cost and easy to setup. The most popular example would be Amazon S3.

# Explain serverless architecture. What are its pros and cons?

Serverless architecture allows engineering teams to launch applications and processes without having to think of provisioning servers (hardware or virtual machines). This can save team deployment time, maintenance and DevOps cost, and usage cost in most cases (you are only charged for the time you run a function). You effectively eliminate the wasted space problem that traditional server architectures introduce.

There are a few downsides. First, you have less control of the underlying infrastructure. If there is a failure, debugging and traceability can be tough. You lose transparency into the system. Also, because serverless infrastructure is multitenant, it may raise security concerns for your company. Some application use-cases require data to be isolated from other tenants. Another problem, depending on the use case, is performance latency. Serverless functions aren’t constantly running. The extra “boot time” may cause a drop in user experience or delay pipeline processes. One last issue is vendor lock-in. Migrating from one CSP serverless option to another is tedious. Some of this risk is now mitigated with serverless functions now supporting container deployments.

# Please provide a diagram for an ETL pipeline (ex: Section 2) using serverless AWS services. Describe each component and its function within the pipeline.

# Describe modern MLOps and how organizations should be approaching management from a tool and system perspective.