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.

Data warehouses contain structured data, typically for business intelligence. They require more work upfront but are easy to integrate with once operational. A good example is having a dashboard for APM tools or to monitor the profitability of a product line.

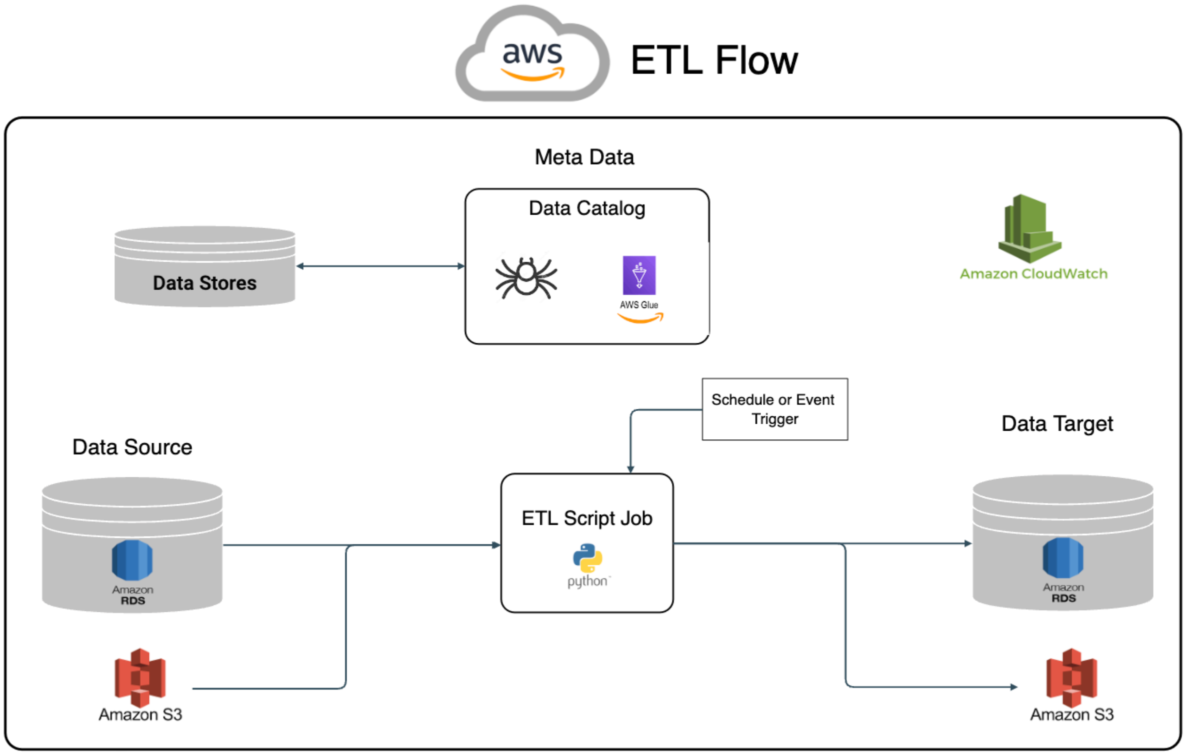
Explain serverless architecture. What are its pros and cons?

Serverless architecture allows engineering teams to launch applications and jobs without having to think about provisioning servers (hardware or virtual machines). This can save teams deployment time, maintenance, and usage cost in most cases (you are only charged for the time you run a function). You effectively eliminate the wasted resource 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. Also, because serverless infrastructure is multitenant, it may raise security concerns for your organization. Some applications 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. Another risk 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.

This example of a serverless flow incorporates AWS Glue for managing the ETL job. The Data Catalog contains meta data for all data sources and targets within the system. The data is obtained from provisioned crawlers within AWS Glue. When setting up a ETL job, users the select source and target tables from the Data Catalog. The sources and targets can be databases (SQL, and NoSQL) or S3. Amazon Cloudwatch is used to monitor the operation of the system.



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

Modern MLOps is the machine learning equivalent to DevOps. Organizations with growing machine learning requirements need to leverage concepts such as continuous integration (CI), continuous deployment (CD), microservices, monitoring, and technical project management into their ML workflows. Data engineers should be tightly coupled with data Scientists and DevOps professionals to ensure best practices and streamlined development.