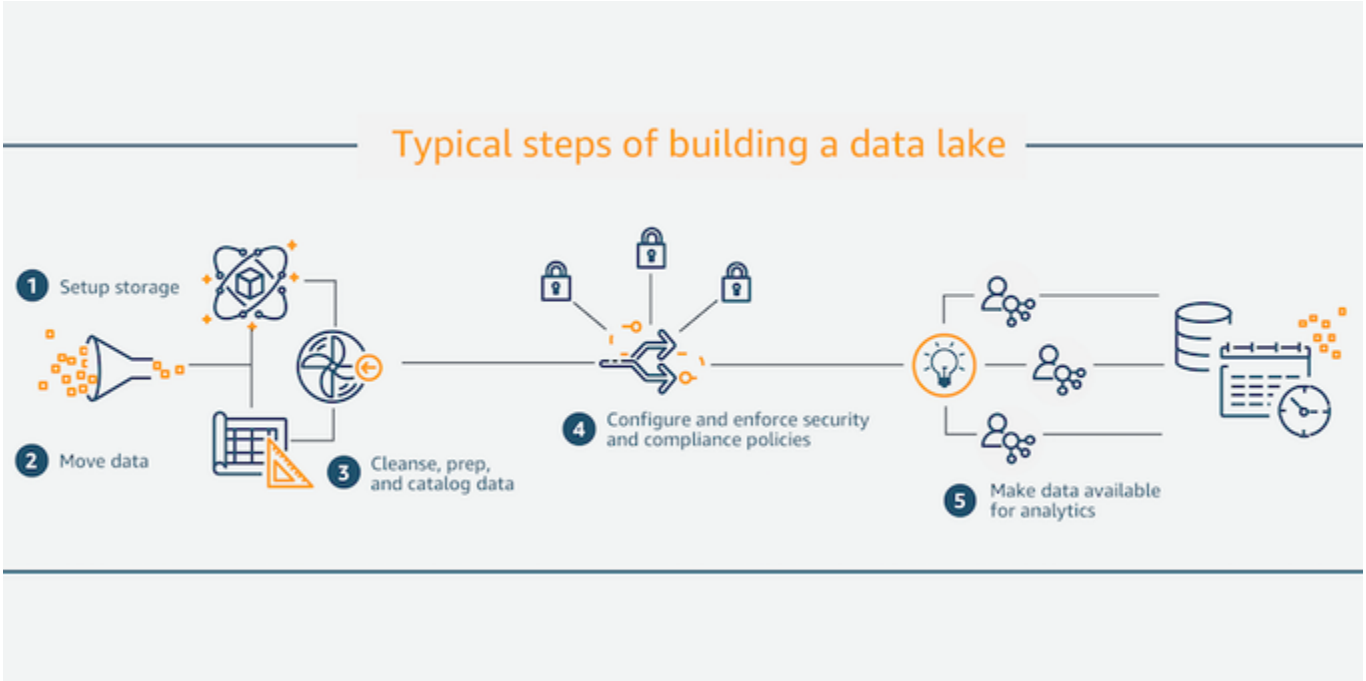
Building a Data Lake with AWS



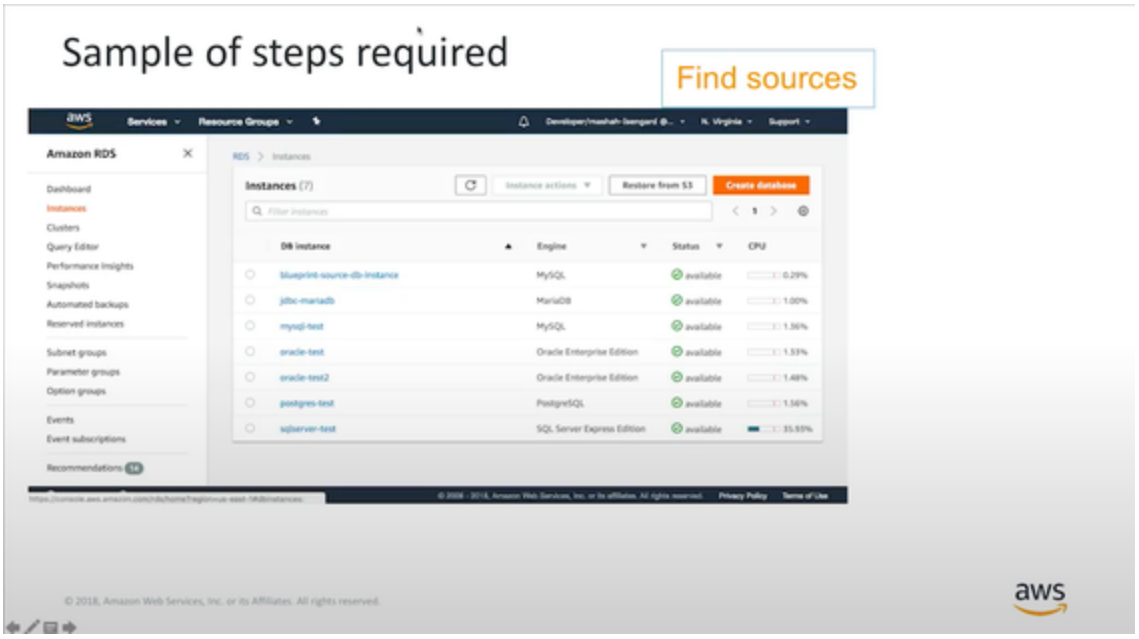
Building a data lake is not always easy and can sometimes take a lot of time before it can become automated. The following are steps that need to occur in order to build a functioning data lake:

1. Setup Storage
    - Data lakes hold an enormous amount of data. Because of this, we need to configure our s3 buckets and partitioners with in the cloud. For on-premise storage, we need to have hardware that can manage this task as well. .
2. Move Data
    - There is data that comes from multiple sources including on premise, in the cloud and IoT devices. In order to organize this data, we must use various ingestion methods, then enable crawlers to extract schemas and add metadata tags to a catalog. This is done with file transfers and ETL. .
3. Cleanse, Prep and Catalog Data
    - Once data is collected and organized, it must be partitioned, indexed and transformed in order to optimize performance and cost. It also must be deduplicated and matched with related records and re-cataloged to maintain cleanliness. .
4. Configure and Enforce Security and Compliance Policies
    - AWS allows us to enforce security policies around its tools so that certain people can access and transform certain data. We also must maintain compliance and protection policies in order to protect sensitive or identifiable information, encrypt data and keep audit logs of who is accessing what and when. .
5. Make Data Available for Analytics
    - We need to make the data easily accessible for analytics. Different people with in the organization need to be able to trust the data and easily find it with out putting in IT requests. This is done through proper curation, cataloging and easy to follow data lineage.

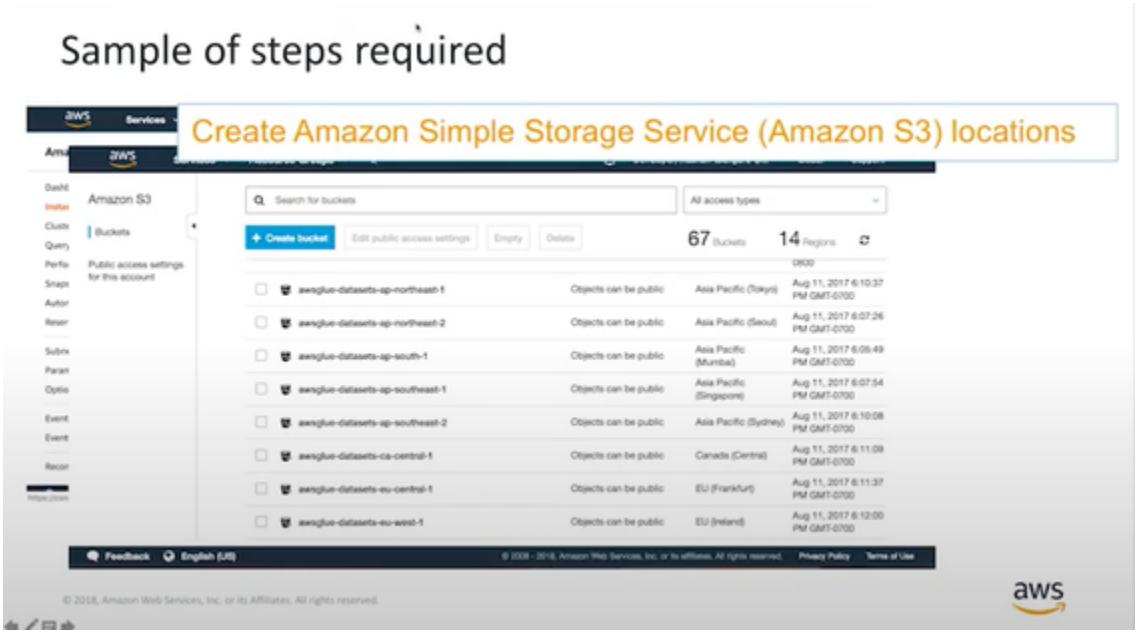
The reason why it can take a lot of time to build a data lake is because these steps are time consuming and usually done manually. It can take months to build workflows, pipelines, map security and create policy settings for an entire company. There's also a lot of time spent configuring tools and services for data movement, storage, cataloging, security, analysis and machine learning.

Although, when dealing with a small lake or beginning a business or new revenue stream, this process is fairly quick.

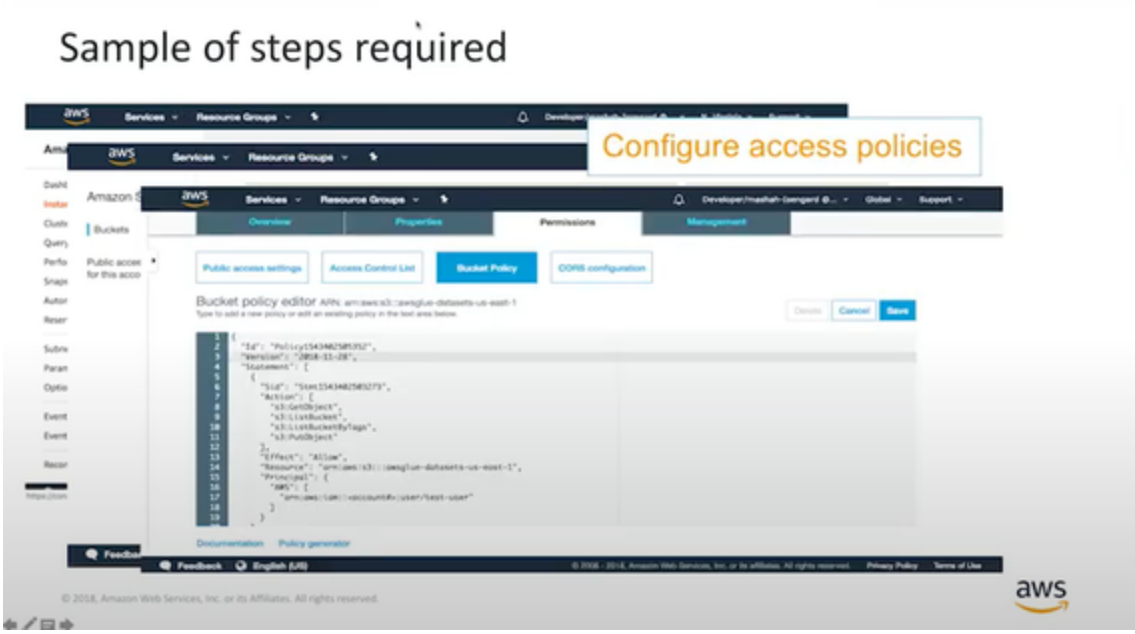
Let's take a look at the usual steps that are taken to set up and secure a basic data lake:



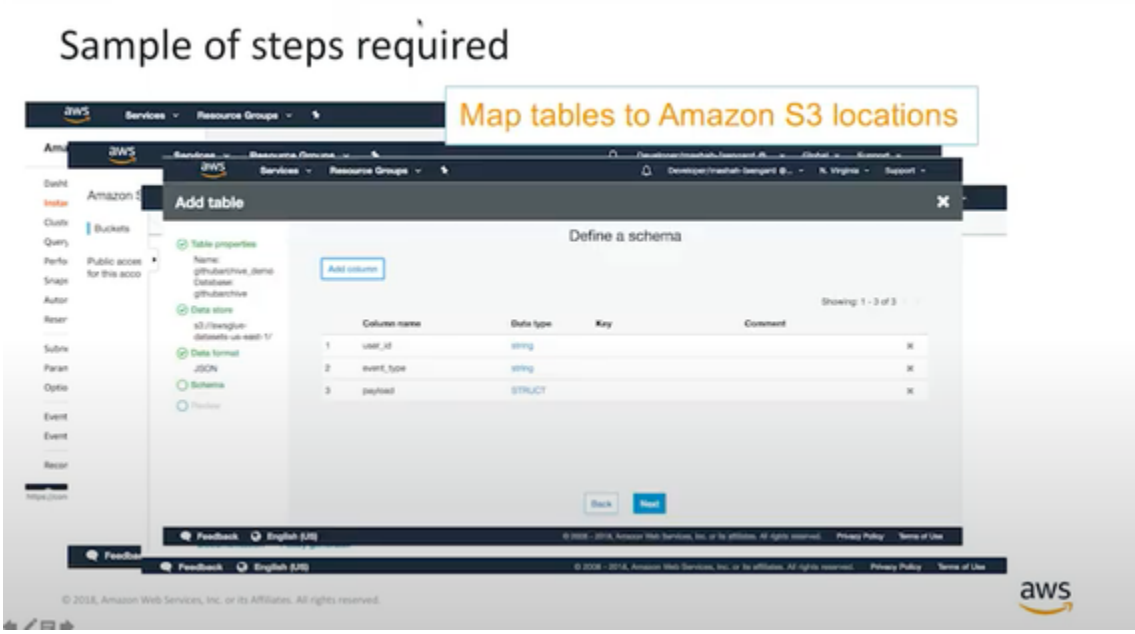
We first start with identifying our data sources from where we want to ingest data from, then select methods for ingesting the data.



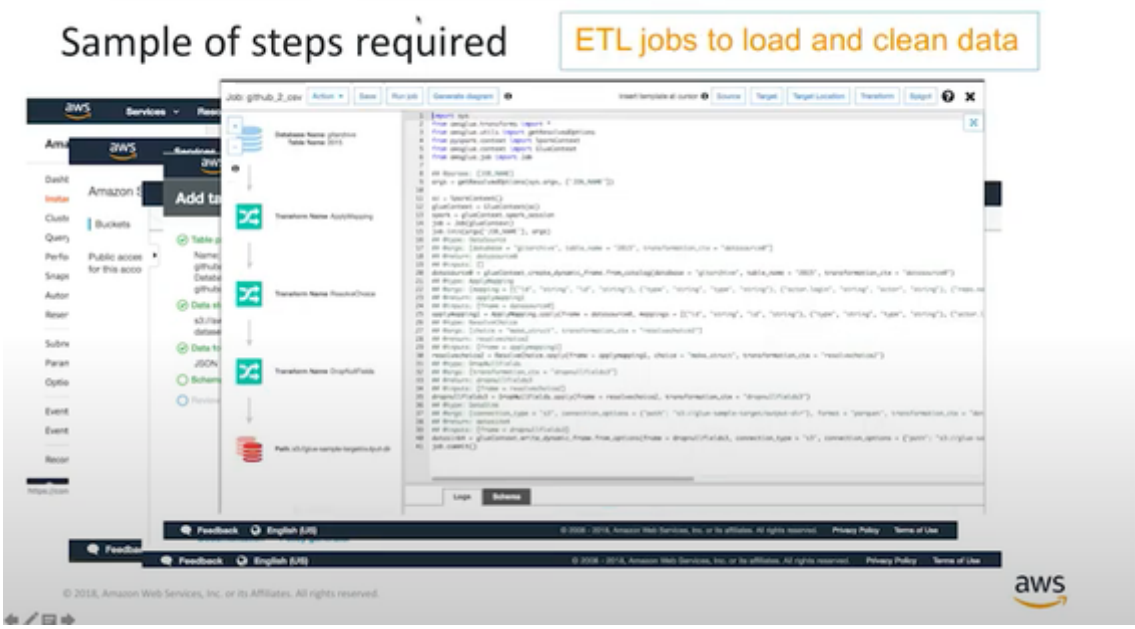
We then create the storage buckets in s3 to store this data.



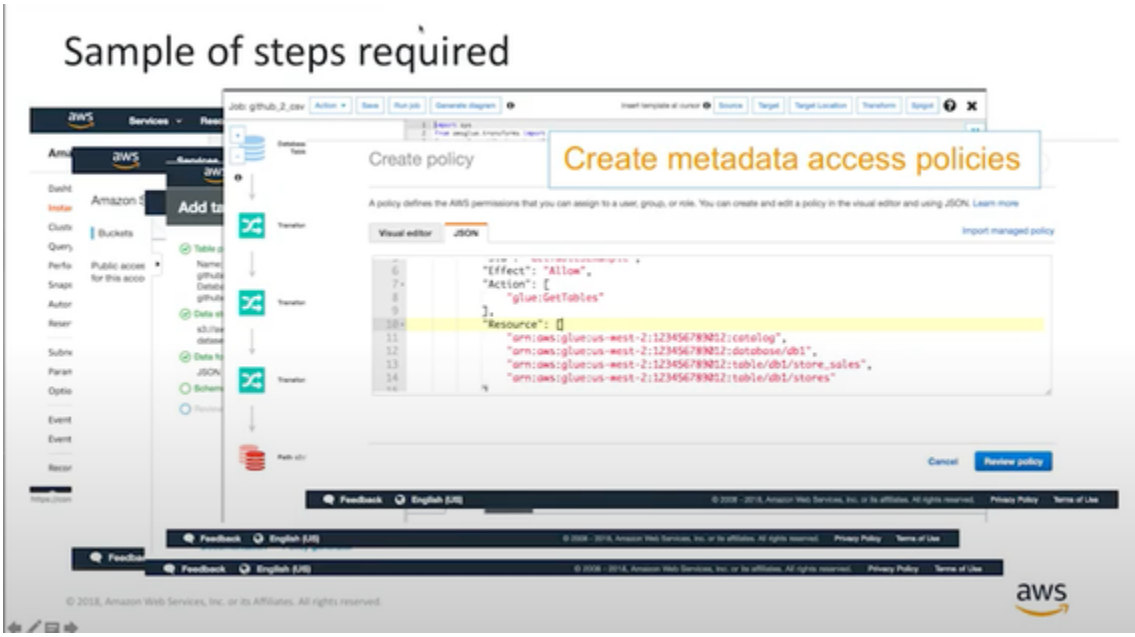
Next, we configure bucket policies to secure the buckets on the storage layer.



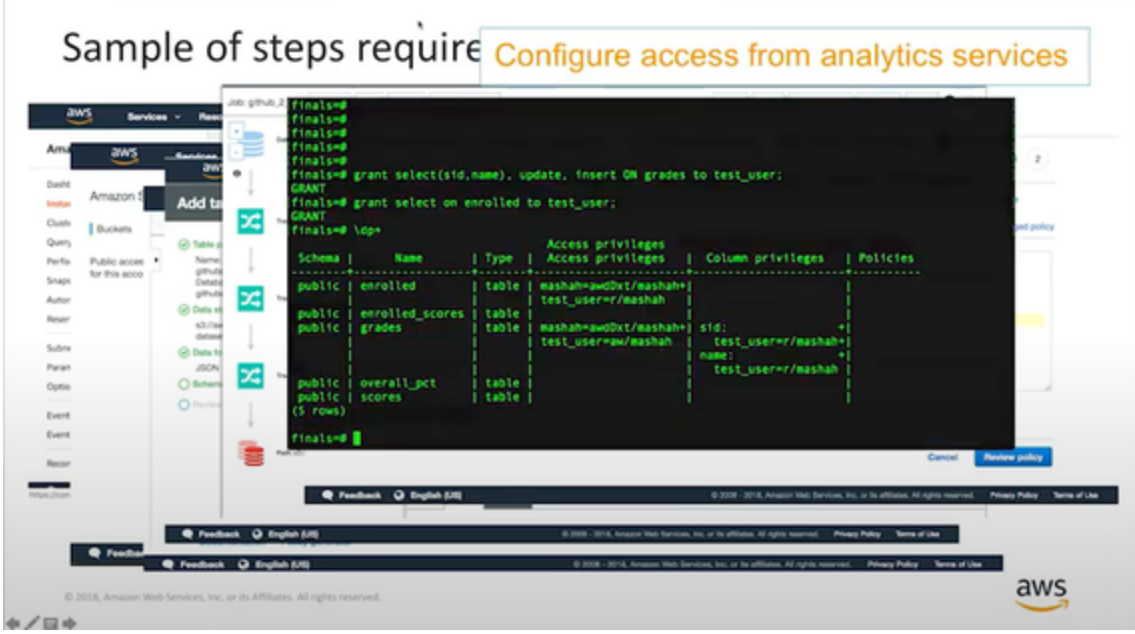
We then use glue to create tables and column mapping in order to use our data sets.



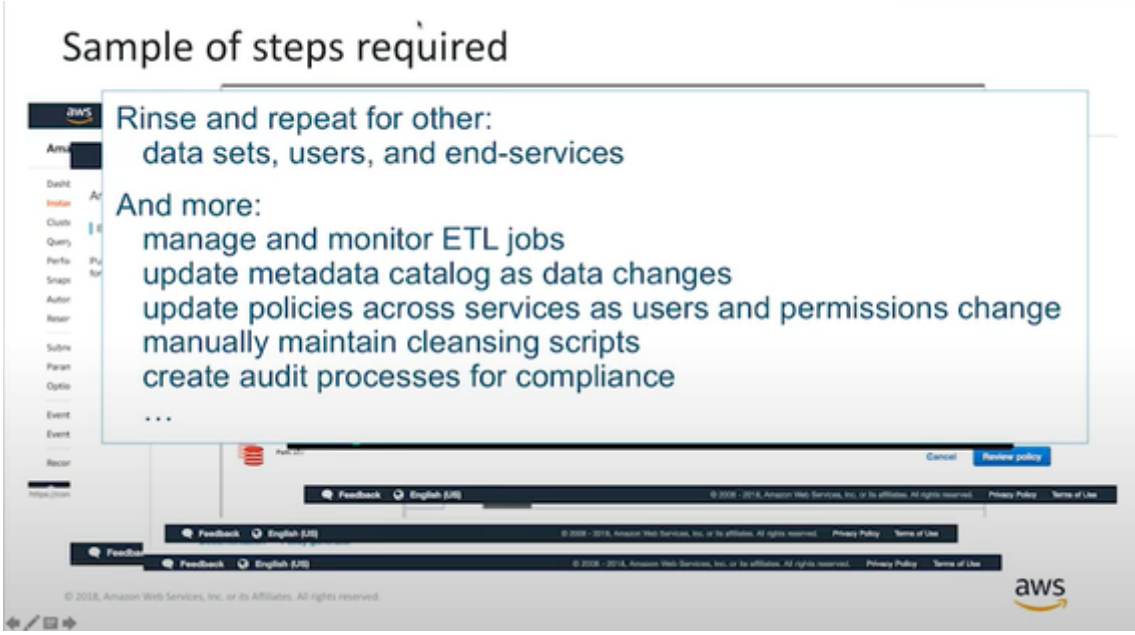
Then we create ETL jobs to load and clean the data.



Next are the IAM policies that need to be created to provide access to underlying AWS resources.



Finally, we configure the analytics services to grant specific user permissions to tables and columns.



All of these steps are just for one data set and we have to repeat everything over again according to the type of data, its ingestion method and the area it'll sit in with in the data lake.

AWS has created a solution for this which is a Data Lake Formation. It's a tool that helps us go through these processes easily. For a quick presentation on the Data Lake Formation please click here.