Part 1. Infrastructure and Administration

Outline the structure of a data pipeline and the basic tools in each of the steps in the flow of information for cloud services (AWS).

○ ***Extract platform***

○ ***Staging***

○ ***Data Lake***

Basic structure of AWS Data Pipeline includes:

- Pipeline definition, it specifies the logic of data flows and treatments.
- EC2 instances are created to perform the defined work activities, once they have been scheduled within the pipeline.
- Task Runner polls for tasks and performs those tasks.

Although every case is different, a simple ETL flow will includes at least:

Step	Tools	Comments
Extract	AWS Glue	In fact, Glue achieves Extraction, Transformation and Load steps. It can be executed over AWS Lambda.
Staging	S3 DynamoDB RDS HBase	It depends on the data structure from the source, it can be used only S3, or Dynamo | HBase instead.
Data Lake	RedShift	Data ready for analysis, aggregation or any actions determined by user requirements.

Based on my experience, data extracted usually relies on RDBMS systems. It makes sense to load that information in Aurora, RDS, HBase or DynamoDB databases. Once the information is in there, it can be moved to RedShift for analysis purposes with QuickSight and other BI Tools. Nevertheless, users requirements will always be the guide for architecture design.

Explain the use case for Lambda Functions, API Gateway, and Step Functions in an ETL implementation.

Lamda Functions are serverless pieces of code. Not within a tool, but anonymous ones. Those kind of functions are used as a part of an orchestrated group of actions reacting to a specific event. In AWS, AWS Lambda is the tool that allows that response. In the case of API Gateway, is the AWS solution for creating, publishing, maintaining and monitoring APIs. Step Functions is the orchestration solution that makes it easy to sequence AWS Lambda functions and multiple AWS services into business-critical applications.

So, in an ETL context, Lambda Functions, API Gateway and Step Functions will react to any event, such as changes on a table, an API execution or a deleted file to trigger an ETL process or update an status in stage area or re-calculate an aggregation. Again, anything can be achieved, but always relying on user or enterprise requirements.