# AWS technical stack for ELT and reporting

Amazon Web Services (AWS) offers a comprehensive suite of services that can be combined to build robust and scalable ELT (Extract, Load, Transform) and reporting pipelines. Here's a breakdown of the key components:

**For ELT (Extract, Load, Transform):**

1. **Data Ingestion & Extraction:**
	* **AWS Glue:** A fully managed, serverless data integration service that makes it easy to discover, prepare, move, and integrate data from various sources. It includes features for data cataloging, ETL job creation (visual or code-based in Python/Scala), and scheduling.
	* **AWS Data Pipeline:** A data workflow orchestration service for automating data movement and transformation across different AWS services and on-premises systems. While still available for existing customers, AWS recommends using AWS Glue for new ELT/ETL workloads.
	* **Amazon Kinesis Data Streams:** For real-time ingestion of streaming data.
	* **Amazon Kinesis Data Firehose:** For capturing, transforming, and loading streaming data into data lakes, data warehouses, and analytics services.
	* **AWS Snow Family (Snowball, Snowcone, Snowmedica):** For physical data transfer of large datasets into AWS.
	* **AWS Database Migration Service (DMS):** For migrating databases to AWS, which can be part of an initial data loading process.
	* **Amazon S3 (Simple Storage Service):** Often used as a staging area for raw data before transformation and loading into the data warehouse or data lake.
2. **Data Storage & Lakehouse:**
	* **Amazon S3:** A highly scalable and durable object storage service that forms the foundation for many data lakes on AWS.
	* **AWS Lake Formation:** A service that makes it easier to set up, secure, and manage data lakes in Amazon S3. It provides a central catalog, data governance, and simplifies data access control.
	* **Amazon Redshift:** A fully managed, petabyte-scale data warehousing service optimized for analytical workloads and complex SQL queries. It's often the target for transformed and structured data.
	* **Amazon RDS (Relational Database Service):** While primarily for operational databases, RDS can serve as a source for ELT processes and, in some cases, a target for specific reporting needs.
	* **Amazon DynamoDB:** A NoSQL key-value and document database that can be a source for ELT, especially for high-velocity data.
3. **Data Transformation:**
	* **AWS Glue:** As mentioned earlier, Glue provides powerful transformation capabilities through its visual interface (Glue Studio), Spark engine, and Python/Scala scripting.
	* **AWS Lambda:** A serverless compute service that can be used to run custom transformation logic in response to events or as part of a data pipeline.
	* **Amazon EMR (Elastic MapReduce):** A big data platform that allows you to run open-source frameworks like Apache Spark and Hadoop for large-scale data processing and complex transformations.
	* **Amazon Athena:** A serverless interactive query service that allows you to analyze data directly in Amazon S3 using standard SQL. While primarily a reporting tool, it can also be used for ELT-style transformations on data in the data lake.
	* **Redshift SQL:** For transformations that can be efficiently performed within the data warehouse using SQL. Redshift's MPP (Massively Parallel Processing) architecture makes it suitable for transforming large volumes of data.
4. **Orchestration & Workflow Management:**
	* **AWS Step Functions:** A serverless function orchestrator that makes it easy to sequence Lambda functions and other AWS services into flexible workflows. It's ideal for building resilient and auditable ELT pipelines.
	* **AWS Glue Workflows:** Provides a visual way to orchestrate Glue crawlers, ETL jobs, and triggers.
	* **Amazon EventBridge:** A serverless event bus that enables you to build event-driven data pipelines, triggering ELT processes based on data arrival or other events.

**For Reporting:**

1. **Business Intelligence & Visualization:**
	* **Amazon QuickSight:** A fast, cloud-powered BI service that makes it easy to create and share interactive dashboards and visualizations. It can connect to various data sources across AWS, including Redshift, S3 (via Athena or Direct Query), RDS, and more. QuickSight offers features like embedded analytics, natural language querying (QuickSight Q), and ML-powered insights.
2. **Query & Analysis:**
	* **Amazon Athena:** As mentioned earlier, Athena allows querying data directly in S3 using SQL, making it a powerful tool for ad-hoc analysis and building reports on data lake data.
	* **Amazon Redshift:** Optimized for high-performance SQL querying on structured data, serving as a backend for many reporting and BI tools.
	* **Amazon SageMaker:** While primarily an ML service, SageMaker Studio provides notebooks that data scientists and analysts can use for advanced data exploration and analysis that might feed into reporting.

**Key Characteristics of the AWS ELT and Reporting Stack:**

* **Scalability:** AWS services are designed to scale to handle massive volumes of data and users.
* **Serverless Options:** Many components like Glue, Lambda, Step Functions, Athena, and QuickSight offer serverless architectures, reducing the operational burden of managing infrastructure.
* **Integration:** AWS services are tightly integrated, allowing for seamless data flow and workflow automation.
* **Cost-Effectiveness:** Pay-as-you-go pricing for most services helps optimize costs based on actual usage.
* **Flexibility:** A wide range of services allows you to choose the best tools for your specific ELT and reporting needs.
* **Security:** AWS provides robust security features and compliance certifications to protect your data.

By strategically combining these AWS services, organizations can build highly efficient, scalable, and cost-effective solutions for their ELT and reporting requirements. The choice of specific services will depend on factors like data volume, velocity, complexity of transformations, reporting needs, and team expertise.