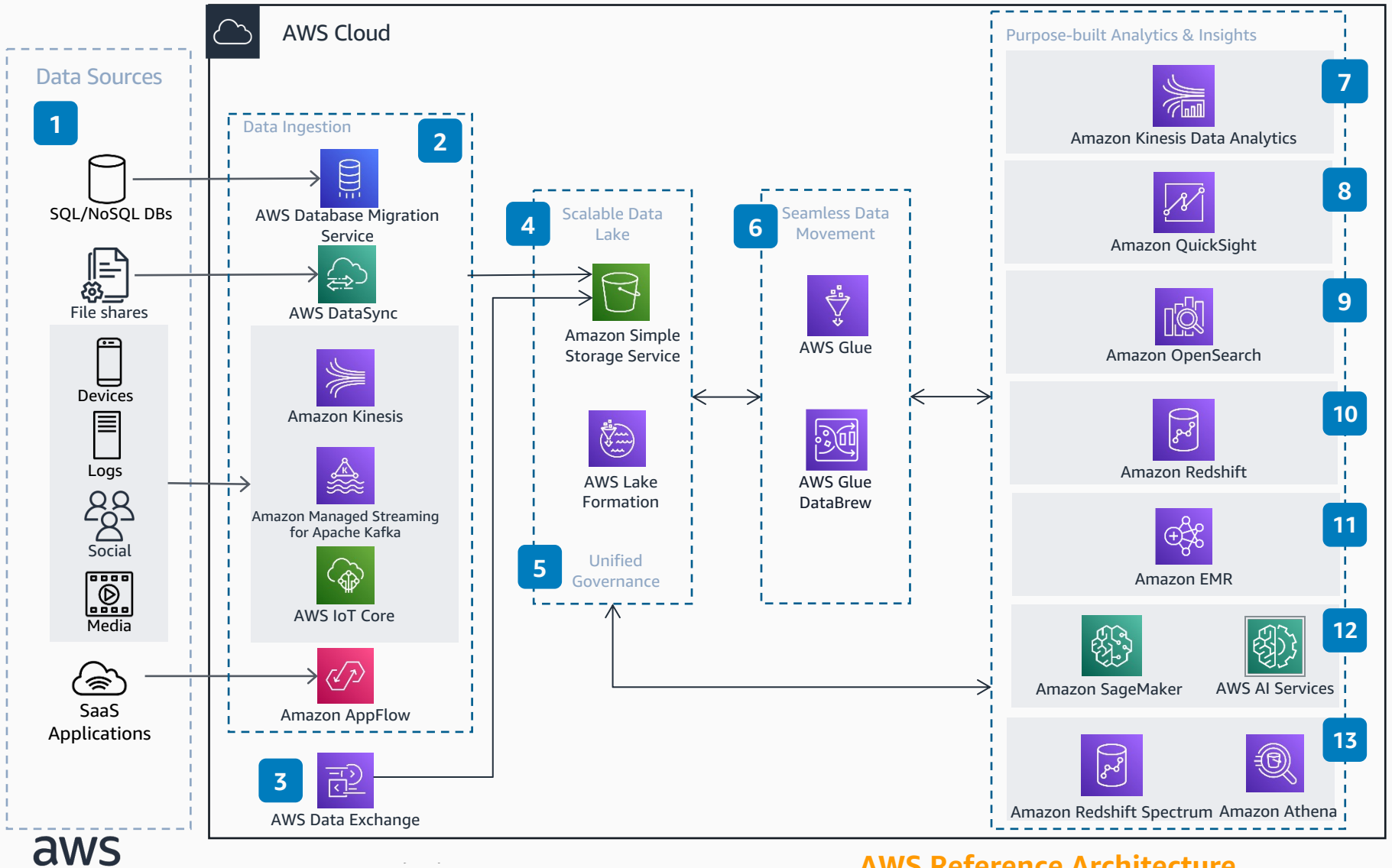
Modern Data Analytics Reference Architecture on AWS

This architecture enables customers to build modern data analytics pipeline using the Lake House approach to derive insights from the data.



1. Data is collected from multiple data sources across the enterprise, software as a service (SaaS) applications, edge devices, logs, streaming media, and social networks.
2. Based on the type of data source, **AWS Database Migration Service**, **AWS DataSync**, **Amazon Kinesis**, **Amazon Managed Streaming for Apache Kafka**, **AWS IoT Core,** and **Amazon AppFlow** are used to ingest the data into a data lake in AWS.
3. **AWS Data Exchange** is used for integrating third-party data into the data lake.
4. **AWS Lake Formation** is used to build the scalable data lake, and **Amazon Simple Storage Service (Amazon S3)** is used for data lake storage.
5. **AWS Lake Formation** is also used to enable unified governance to centrally manage security, access control, and audit trails.
6. **AWS Glue** is used to extract, transform, catalog, and ingest data across multiple data stores. **AWS Glue DataBrew** could be used for visual data preparation.
7. **Amazon Kinesis Data Analytics** is used to transform and analyze streaming data in real time.
8. **Amazon QuickSight** provides machine learning (ML) powered business intelligence.
9. **Amazon OpenSearch** can be used for operational analytics.
10. **Amazon Redshift** is used as a cloud data warehouse.
11. **Amazon EMR** provides the cloud big data platform for processing vast amounts of data using open source tools.
12. **Amazon SageMaker** and **AWS AI services** can be used to build, train, and deploy ML models, and add intelligence to your applications.
13. **Amazon Redshift Spectrum** and **Amazon Athena** enable interactive querying, analyzing, and processing capabilities.

AWS Reference Architecture