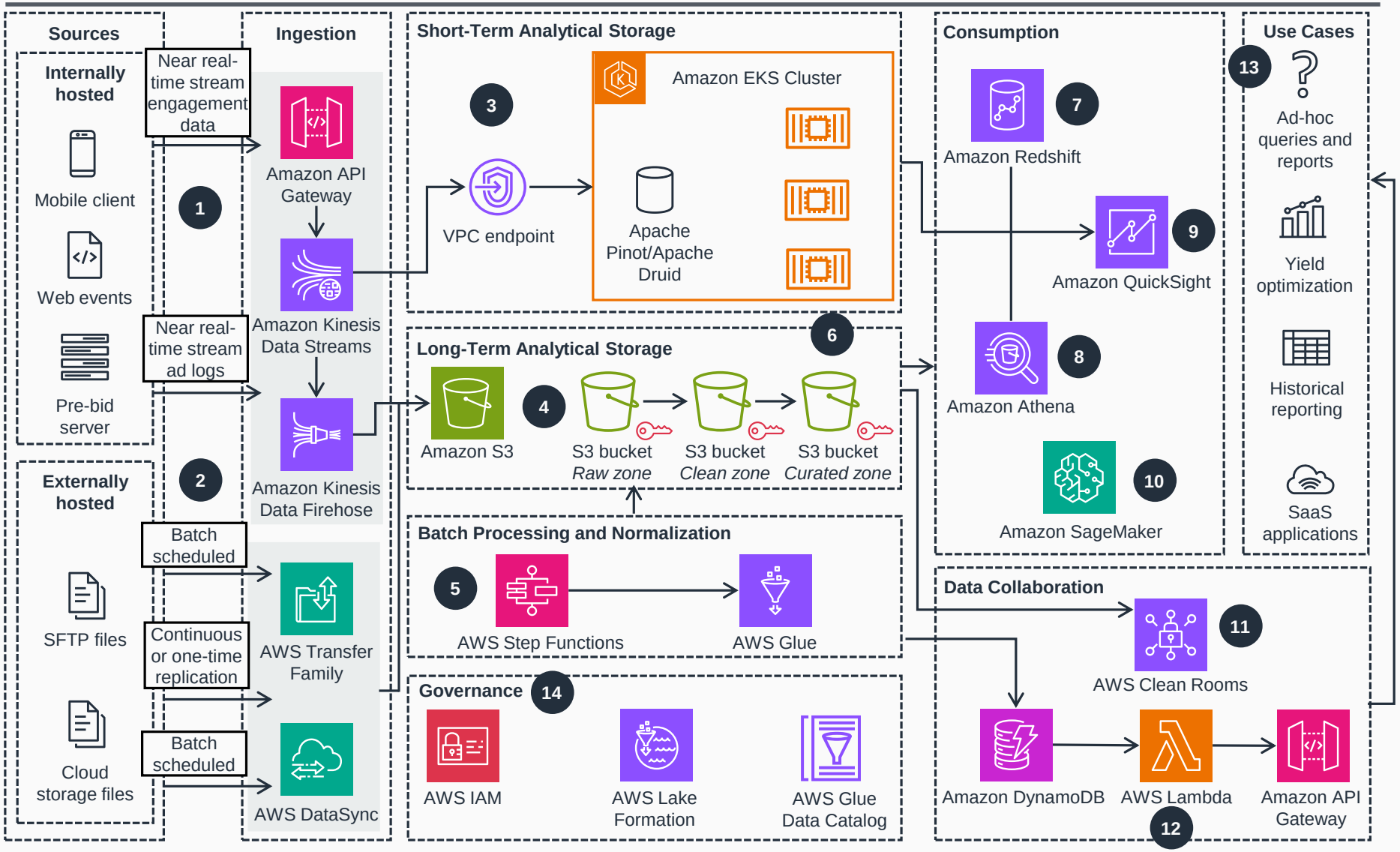
Guidance for Publisher Advertising Revenue Data Product on AWS

This architecture diagram creates an advertising revenue source of truth that can be viewed as a data product and feed into other publisher data products. This slide details steps 1-6 of the architecture diagram.



1. Stream the clickstream web and mobile engagement data and internally hosted pre-bid server data into AWS analytical infrastructure through **Amazon Kinesis** and **Amazon API Gateway**.

2. Collect the ad log data generated by third-party applications through the batch ingestion process. Use **AWS Transfer Family** for transferring data from files stored in SFTP servers. Use **AWS Data Sync** to copy files from external cloud data stores.

3. Store near real-time streaming data in a hot analytical storage layer that supports low latency ingestion and concurrent querying. Host analytical databases, such as Apache Pinot or Apache Druid Host, in **Amazon Elastic Kubernetes Service (Amazon EKS)** clusters. Time To Live (TTL) for events range from 7-30 days. The virtual private cloud (VPC) endpoint securely transfers data from AWS-managed services to VPC-hosted services.

4. Stream and batch load both external and internal data sets into the **Amazon Simple Storage Service (Amazon S3)** raw zone. Use VPC endpoints to securely transfer data.

5. **AWS Glue Apache Spark** jobs process large volumes of data using distributed computing. **AWS Step Functions** orchestrates data processing workflows that include **AWS Glue** and other services.

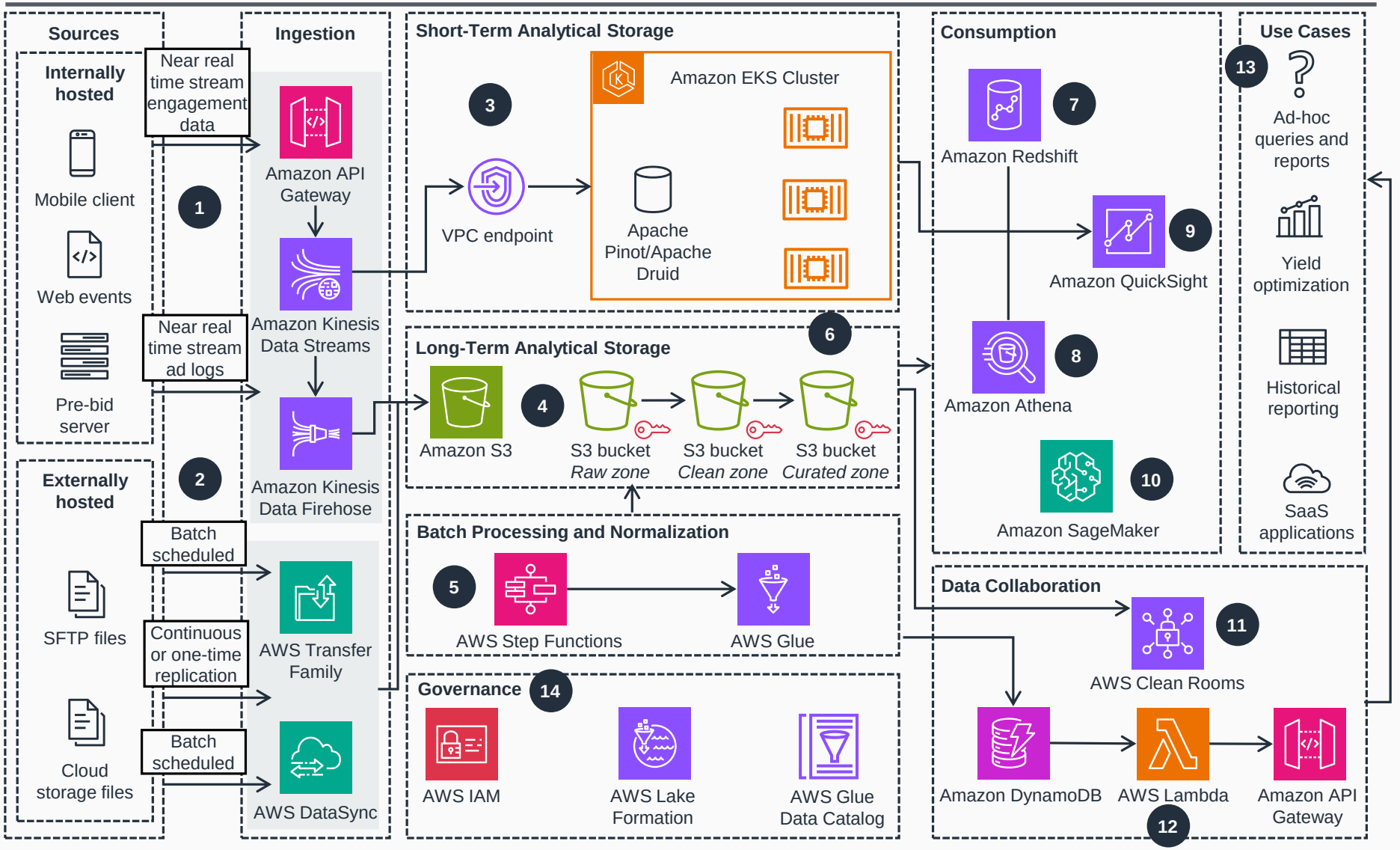
6. The logical clean zone in **Amazon S3** stores the source, such as data, in a read-optimized format. Use parquet format and apply proper partition and compression techniques. Use the logical curated zone to integrate data from various sources and store them in a normalized schema.

Reviewed for technical accuracy September 15, 2023

AWS Reference Architecture

Guidance for Publisher Advertising Revenue Data Product on AWS

This architecture diagram creates an advertising revenue source of truth that can be viewed as a data product and feed into other publisher data products. This slide details steps 7-14 of the architecture diagram.



7 Optionally, use **Amazon Redshift** to build a data warehouse that hosts curated or modeled data. This warehouse will handle repeated analytical queries and dashboards that can benefit from massively parallel processing (MPP) architecture and indexes.

8 **Amazon Athena** performs ad-hoc data discovery and analysis queries.

9 Connect the short-term analytical storage to **Amazon QuickSight** to build operational reporting dashboards. These dashboards are embedded in web applications that the operations team use. Historical data sets in **Amazon S3** and **Amazon Redshift** are accessed from **QuickSight** to build business intelligence (BI) dashboards.

10 Access the data lake using **Amazon SageMaker** to train, test, and deploy machine learning (ML) models. These ML models are deployed for use cases such as optimizing yield, making supply forecast, and shaping traffic.

11 Use **AWS Clean Rooms** to collaborate with advertisers and measurement providers. This enhances privacy and allows combined analysis of the curated datasets without exposing raw data.

12 Selectively load curated datasets into **Amazon DynamoDB** using **AWS Glue**. Build APIs using **AWS Lambda** and **API Gateway** to share data with stakeholders for use cases such as monetization.

13 Data in the Publisher Advertising data lake supports use cases such as yield optimization and acts as a source of truth for advertising revenue data in enterprise reporting.

14 Use **AWS Lake Formation** to define fine-grained access controls on **AWS Glue Data Catalog** tables, columns, and rows in the data lake. **AWS Identity and Access Management (IAM)** securely manages identities and access to AWS services and resources.

Reviewed for technical accuracy September 15, 2023

AWS Reference Architecture