Babbel Datalake Architectural Digram

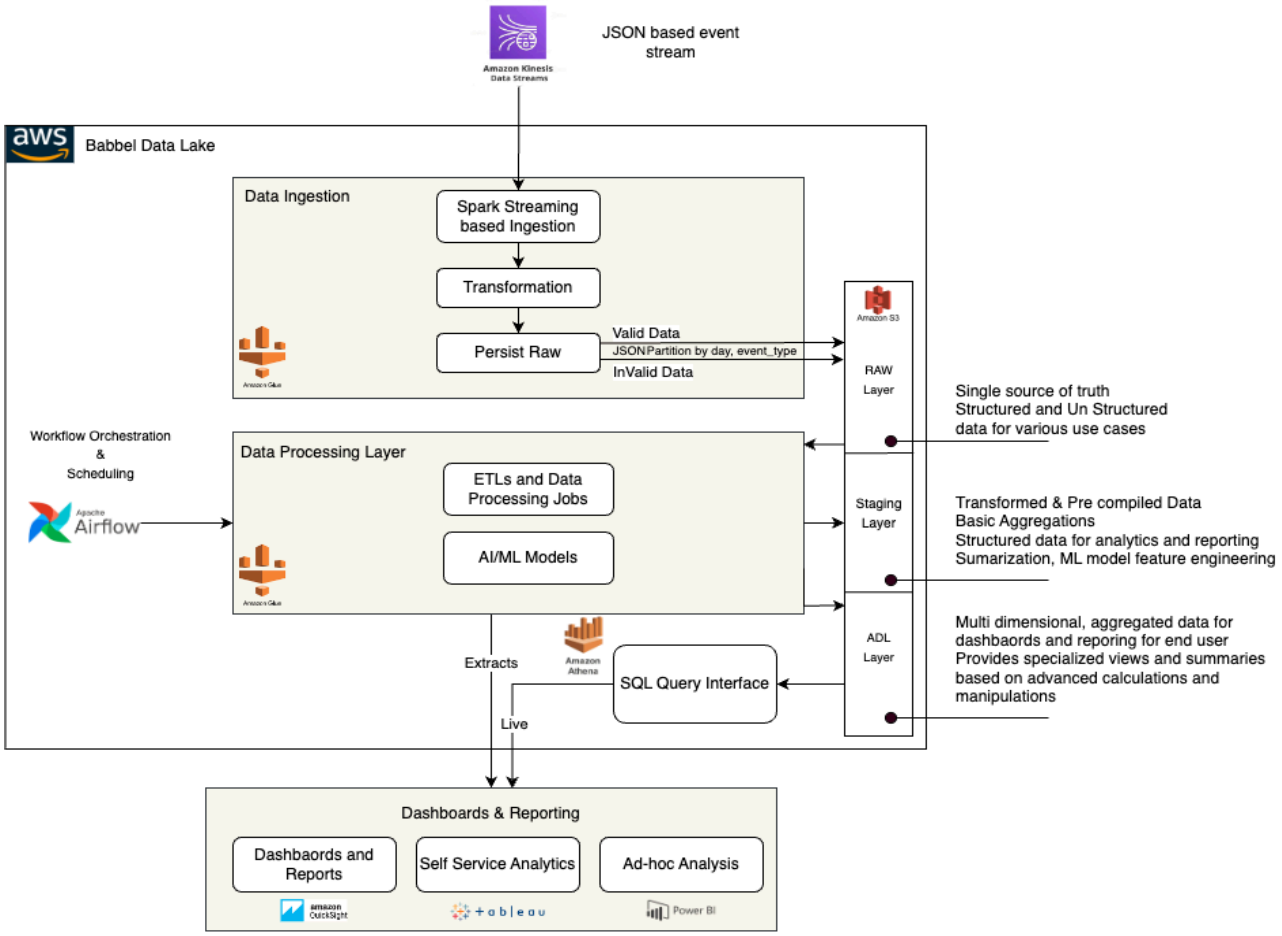
Architectural Design:

This document provides a proposed Data Lake architecture for different analytical use-cases. It can process streaming data from kinesis stream, transform and persist it. It is designed to be resilient, scalable, performant and cost effective to support range of analytical use-cases for stakeholders including Specialized Dashboards and Reports, Self Service Analytics, Ad-hoc Analysis and AI/ML model development.

Proposed Design recommends range of managed services and technologies in AWS cloud. This will help develop and deliver this system faster for stakeholders as well as get cost benefits of pay as you go model and reduced maintenance of those services in future.

Assumptions:

- Payload of events is not known and Schema registry is needed to support schema evolution.
- Monitoring, Governance and Security Systems are in place separately (not part of this document) and can be integrated as needed.



Data Storage:

Data is stored Amazon S3 which provides cost effective, large scale, distributed storage layer to Data lake. Data is stored in purposed built layered approach to serve different purposes at different stages. These layers are Raw and Staging and ADL (Aggregated Data Layer). Data is moved from Kinesis Stream to different layers in the storage layer as it is cleaned, transformed and aggregated.

Raw events are stored in JSON formatted and partitioned by date (day), event_type to support better data retrieval and support incremental processing later on. Data is stored into valid and invalid paths after performing initial validation checks.

Staging and Aggregated data layers store processed data in Parquet format. These layers provide query capabilities by Athena and benefits from column format and compression of parquet.

Data Ingestion and Processing:

This architecture is based on Apache Spark engine for distributed, large scale fault tolerant batch and streaming data processing. Historical data will be pulled in Glue batch jobs and Glue Streaming Job will be used to ingest realtime data from Event Log File during operational hours. Data will be validated, cleaned and persisted in Staging layer in S3. Glue based ETLs and Pipelines will transform data as per business needs and create specialized and aggregated datasets in ODL and ADL layers. All workflows including ETLs and Pipelines will be orchestrated and managed by Airflow (MWAA). Realtime Monitoring platform will be integrated to monitor all workflows and they will log and publish metrics as well as will be designed to be resilient to failures with auto retries and self startups.

Dashboards and Reporting:

Different Analytical systems will provide access to data and visualization for historical analysis. Critical dashboards will be created in Quicksight with historical trends and predictions for specific audience. Stakeholders will have access to critical ADL and ODL datasets for Self Service analtics and Ad-hoc historical analysis using Tableau / Power BI. Data will be pushed/pulled into these dashboards on defined schedules and SLAs, as well as Live query access will be provided using Athena Engine.