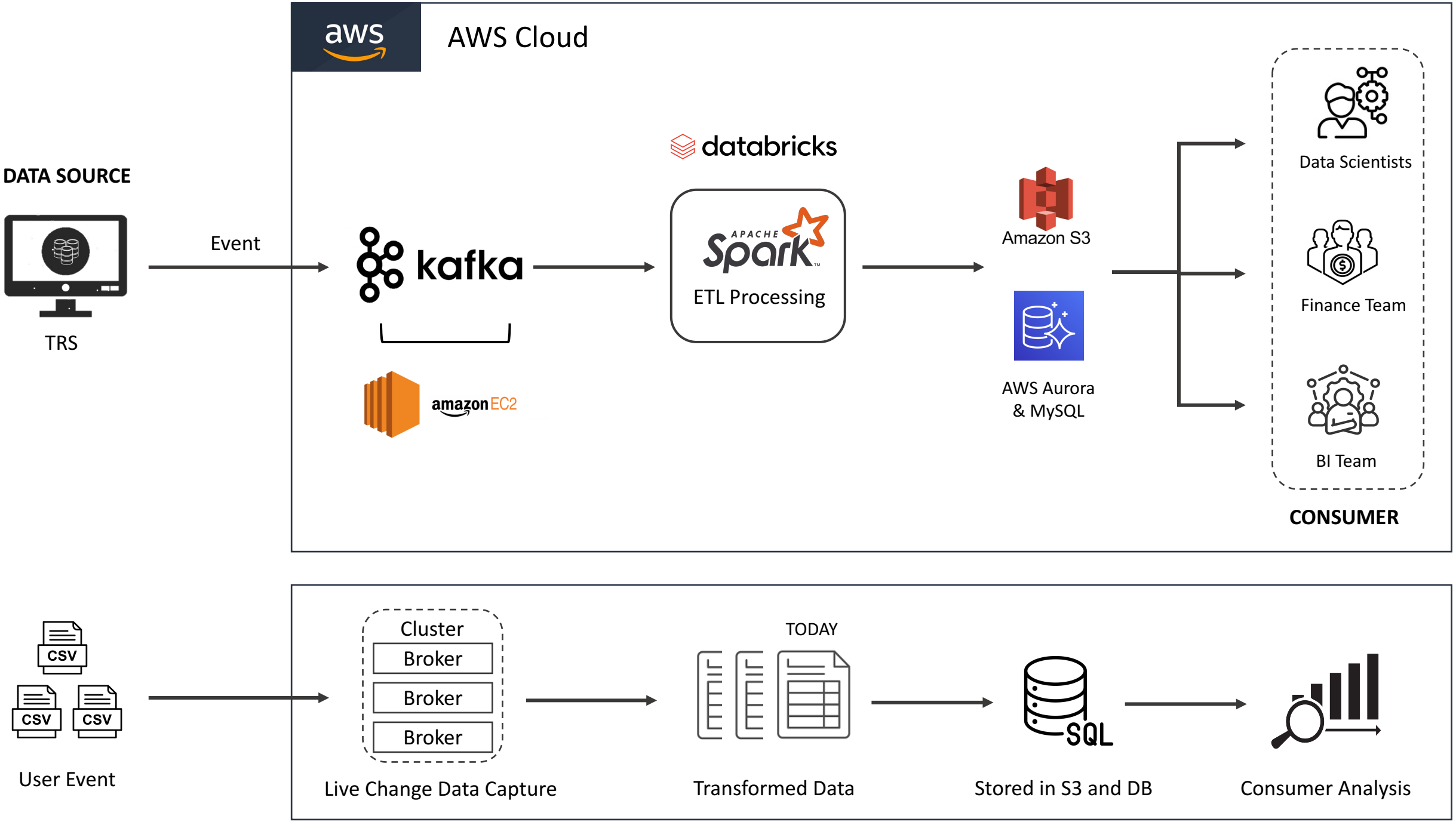
AWS Cloud
DATA SOURCE
TRS
Event
kafka
amazon EC2
databricks
APACHE Spark
ETL Processing
Amazon S3
AWS Aurora & MySQL
Data Scientists
Finance Team
BI Team
CONSUMER
CSV
CSV
CSV
User Event
Cluster
Broker
Broker
Broker
Live Change Data Capture
TODAY
Transformed Data
SQL
Stored in S3 and DB
Consumer Analysis

Production System Architecture

Kafka: Implementing CDC with Kafka allows for seamless capture and processing of real-time changes from TRS. Kafka acts as a reliable message broker that can handle large volumes of data.

Spark: Using Spark for the ETL pipeline allows us to process large volumes of data efficiently and tailor them to suit specific teams.

AWS S3 and MySQL: Kept up to date with near real time data and changes made in TRS allowing Data Scientists, Finance, and BI teams to access live data for their respective use cases.

Scalability: The production pipeline uses Elastic Computing (EC2) instances in AWS that can be scaled up to handle larger and more complex datasets. Storing the datasets in a data lake (AWS S3) also provides a scalable and cost-effective solution.