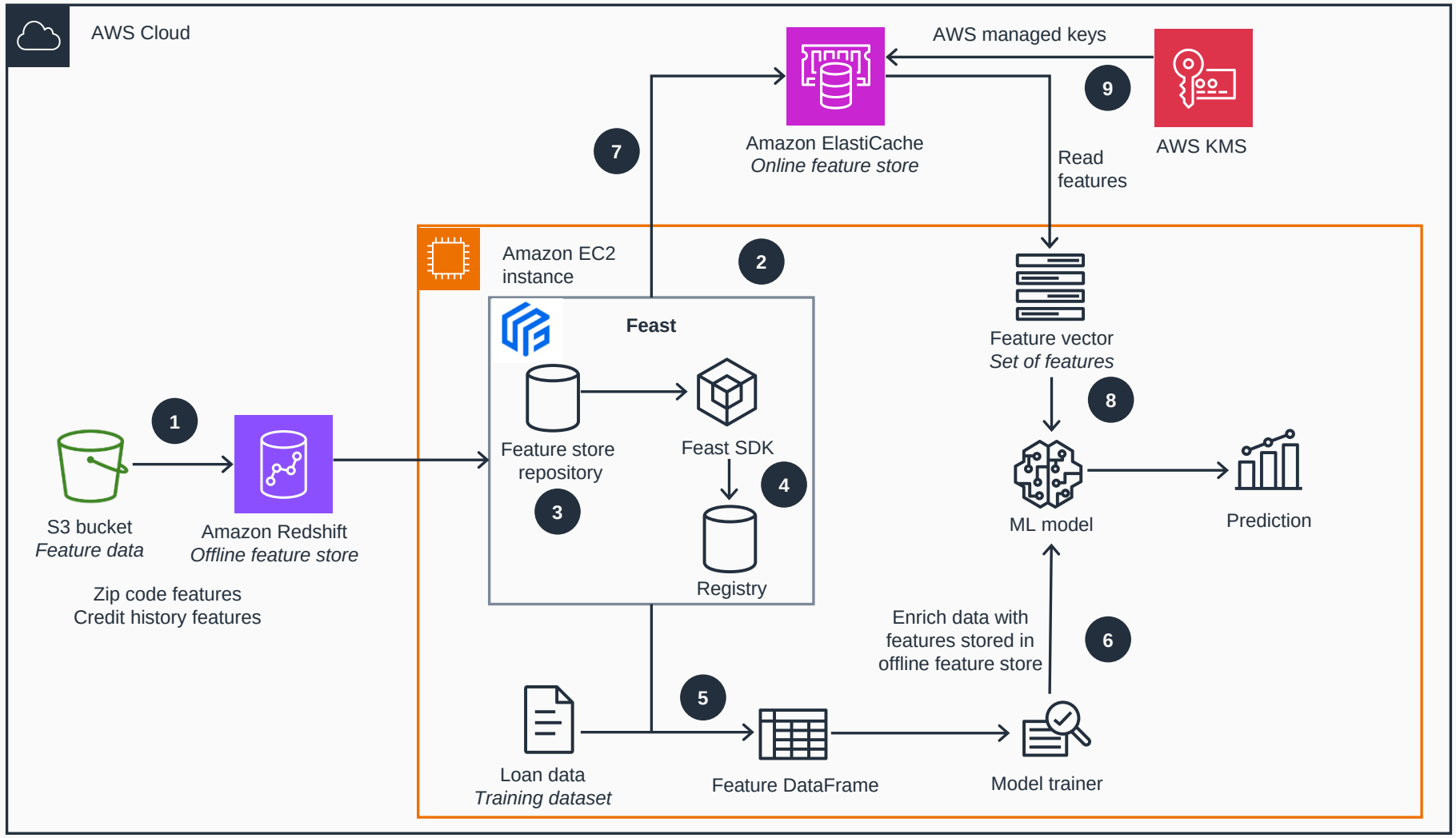
Guidance for Ultra-Low Latency, Machine Learning Feature Stores on AWS

This architecture diagram shows how to make online predictions with a real-time credit scoring system.



1. Set up data infrastructure to deploy **Amazon Redshift**, an **Amazon Simple Storage Service (Amazon S3)** bucket containing zip code and credit history parquet files, and **AWS Identity and Access Management (IAM)** roles. Additionally, set up policies for **Amazon Redshift** to access **Amazon S3**, and create an **Amazon Redshift** table that can query the parquet files.
2. Deploy Feast infrastructure.
3. Create a feature store repository, and configure **Amazon ElastiCache** as the online feature store and **Amazon Redshift** as the offline feature store. Create feature definitions.
4. Register the feature definitions and the underlying infrastructure into a Feast registry using the Feast SDK.
5. Generate training data using features and labels from the data and features from Feast. The features from Feast enrich the historical data and create a Feature DataFrame.
6. Train the ML model using the training dataset and a model trainer.
7. Ingest batch features into the **ElastiCache** online feature store. These online features are used to make online predictions with our trained model.
8. Read feature vector from **ElastiCache** for making predictions.
9. Use **AWS Key Management Service (AWS KMS)** to encrypt **ElastiCache** data at rest.

Reviewed for technical accuracy May 15, 2024

AWS Reference Architecture