Open Science Index, Computer and Information Engineering Vol:16, No:11, 2022 publications.waset.org/abstracts/146913.pdf

Machine Learning Data Architecture

Authors : Neerav Kumar, Naumaan Nayyar, Sharath Kashyap

Abstract : Most companies see an increase in the adoption of machine learning (ML) applications across internal and external-facing use cases. ML applications vend output either in batch or real-time patterns. A complete batch ML pipeline architecture comprises data sourcing, feature engineering, model training, model deployment, model output vending into a data store for downstream application. Due to unclear role expectations, we have observed that scientists specializing in building and optimizing models are investing significant efforts into building the other components of the architecture, which we do not believe is the best use of scientists' bandwidth. We propose a system architecture created using AWS services that bring industry best practices to managing the workflow and simplifies the process of model deployment and end-to-end data integration for an ML application. This narrows down the scope of scientists' work to model building and refinement while specialized data engineers take over the deployment, pipeline orchestration, data quality, data permission system, etc. The pipeline infrastructure is built and deployed as code (using terraform, cdk, cloudformation, etc.) which makes it easy to replicate and/or extend the architecture to other models that are used in an organization.

Keywords : data pipeline, machine learning, AWS, architecture, batch machine learning
Conference Title : ICBDET 2022 : International Conference on Big Data Engineering and Technology
Conference Location : San Francisco, United States
Conference Dates : November 03-04, 2022