Google Cloud

Readings: Architecting Production ML Systems

In this module, you further explore production ML systems and as well as training and model serving decisions you make to get the right performance profile. For more information, see the following readings and videos.

- Architecture of a real-world Machine Learning system
- ML Reference Architecture
- Machine Learning Pipeline: Architecture of ML Platform in Production
- Design Decisions for Architecting Production Machine Learning Systems
- Production ML Systems
- 3 Building Blocks of Machine Learning you Should Know as a Data Scientist
- MLOps: Continuous delivery and automation pipelines in machine learning
- Smart Decisions Game: Machine Learning for Architects
- Training and Serving CARET models using AI Platform Custom Containers and Cloud Run
- Using TensorFlow to predict product weight and dimensions
- Getting batch predictions
- How to extend a canned TensorFlow Estimator
- Introduction to loading data
- Google Cloud Vertex AI
- AI Simplified
- Cloud Blog AI & Machine Learning
- GitHub - Google Cloud Pipeline Components
- Vertex AI: Building a fraud detection model with AutoML