End-to-End Data Management in Support of an ML RecSys

(Keynote)

David Cohen,
Senior Principal Engineer, Intel
david.e.cohen@intel.com

Abstract

Since the inception of the World Wide Web, companies have been mining web-server logs to analyze user interactions. The rise of Smart-Phones, their video and photo capabilities, and Smart-Phone-resident applications continues to drive year-on-year data growth at phenomenal rates. By the mid 2010s, Google observed that data generated annually by their Youtube service was growing faster than annual improvements in CPU performance. By 2016 the industry began to employ offload accelerators (e.g. GPU, TPU, etc) to augment CPU processing capabilities in order to keep up with the deluge of incoming data. It has been noted that today data growth continues to outstrip performance improvements in the computational plant, even after incorporating offload accelerators.

Over the same period, companies such as Alibaba, Amazon, Google, Meta, Microsoft and others have aggressively adopted, and innovated with machine learning. These companies have seen an explosion in the dataset sizes used for training these models as well as the frequency of decisions the models produce in the serving environment. The number of parameters in a single model has grown to be in the billions to trillion range. Models of this size consume significant amounts of infrastructure resources at training and serving time. Training a model of this size can take weeks to complete.

These data growth and machine learning trends are reshaping the data management discipline within companies that operate at this scale. For example, queries originating from machine learning activities have become a driving force in the utilization of data warehouse resources. In this talk we'll discuss how data management tools are used throughout the data processing cycle: from log capture, data ingestion into the data warehouse, preprocessing, and feature engineering in support of model training and serving. The focus will be on positioning each of these steps within a broader Recommendation System scenario. We'll conclude with a discussion on the changing landscape of the infrastructure that hosts machine learning and data management workloads.

Biography

Dave is a Senior Principal Engineer in Intel's Data Center and Artificial Intelligence (DCAI) business unit. He focuses on large scale data management challenges being faced by Intel's Loud customers. Prior to Intel, David was a Director in the Office of the CTO at EMC where he lead efforts related to integrating storage systems with network virtualization. David also has a long-history of working on building distributed systems in industry, most recently working for the investment banks: Goldman Sachs and Merrill Lynch. An experienced practitioner, Dave's active connections to commercial and academic research and development labs insure Intel’s Data Management Solutions are both well-grounded and cutting-edge. An acknowledged industry expert in system architecture and development, Dave is a sought after speaker and published author.