ARCH-492

Architecting Cloudera Edge to AI

Course Overview

Course Type
Instructor-Led Training

Level
Intermediate to Advanced

Duration
4 days

Platform
Cloudera

Topics Covered
- Cloudera Platform
- Big Data Architecture
- Building Scalable applications
- Building Fault Tolerant Solutions
- Security and Privacy
- Deployment on Public, Private, and Hybrid Cloud
- Architecting AI
- AI Studios

About This Training

Architecting Cloudera from Edge to AI is a 4–day learning event that addresses advanced big data architecture topics for building edge to AI applications to cover streaming, operational data processing, analytics, and machine learning. The workshop brings together technical contributors into a group setting to design and architect solutions to a challenging business problem. The workshop addresses big data architecture problems in general, and then applies them to the design of a challenging system.

Throughout the highly interactive workshop, participants apply concepts to real-world examples resulting in detailed synergistic discussions. The workshop is conducive for participants to learn techniques for architecting big data systems, not only from Cloudera's experience but also from the experiences of fellow participants.

More specifically, this workshop addresses advanced big data architecture topics, including, data formats, transformation, transactions, real-time, batch and machine learning processing, scalability, fault tolerance, security, and privacy, minimizing the risk of an unsound architecture and technology selection.

Who Should Take This Course?

Participants should mainly be architects, developer team leads, big data developers, data engineers, senior analysts, dev ops admins and machine learning developers who are working on big data or streaming applications and have an interest in how to design and develop such applications on Cloudera. To gain the most from the workshop, participants should have working knowledge of popular Big Data and streaming technologies such as HDFS, Spark, Kafka, Hive/Impala, Data Formats, and relational database management systems. Detailed API level knowledge is not needed, as there will not be any programming activities and instead the focus will be on architecture design.

The workshop will be divided into small groups to discuss the problems, develop solutions, and present their solutions.

ARCH-492

Architecting Cloudera Edge to AI

Introduction

- Team activity: Team Introductions

Technology Review

- Data Hubs
  - Important Architecture trends
  - Data Lifecycle
- Data Flow & Streaming
  - Spark Streaming, Flink, Kafka Streams/Connect
  - Comparing Streaming solutions
- Data Engineering
  - Spark
  - HDFS, Ozone
  - YARN, Kubernetes, Yunikorn
- Data Warehouse
  - Hive, Impala, DataViz
  - Real Time Data warehouse architectures
  - Comparing Databases and storage engines
- Operational Database
  - Hbase, Phoenix, Kudu, Solr
- Cloudera AI
  - Machine Learning
- Cloudera Observability
- Replication Manager

Workshop Application Use Cases

- Oz Metropolitan
- Architectural questions
- Team activity: Review Metroz Use Cases and Logical Architecture

Application Vertical Slice

- Definition
- Minimizing risk of an unsound architecture
- Selecting a vertical slice
- Team activity: Metroz Vertical Slice

Application Processing

- Real time, near real time processing
- Batch processing
- Data access patterns
- Delivery and processing guarantees
- Data consistency and ACID transactions
- Stream processing guarantees
- Machine Learning pipelines
- Team activity: Metroz Processing

Application Data

- Three V's of Big Data
- Data Lifecycle
- Data Formats
- Transforming Data
- Team activity: Metroz Data Requirements

Scalable Applications

- Scale up, scale out, scale to X
- Determining if an application will scale
- Poll: scalable airport terminal designs
- Spark scalability and parallel processing
- Scalable storage engines: HDFS, Ozone, Kafka, and Kudu
- Team activity: Scaling Metroz

CLOUDERA

ARCH-492

Architecting Cloudera Edge to AI

Fault-Tolerant Distributed Systems

- Principles
- Transparency
- Hardware vs. Software redundancy
- Tolerating disasters
- Stateless functional fault tolerance
- Stateful fault tolerance
- Replication and group consistency
- Application tolerance for failures
- Team activity: Failures in Metroz

Security and Privacy

- Principles
- Security Architecture
- Knox Security Architecture
- Ranger Security Architecture
- Setting security policies with Ranger
- Threat Analysis
- Team activity: Securing Metroz

Deployment

- Cluster sizing and evolution
- On-premise vs. Cloud
- Cost comparison
- Edge computing
- Cloudera on Cloud Architecture
- Introduction to containers and kubernetes
- Team activity: Deploying Metroz

Software Architecture

- Architecture artifacts
- Team activity: Metroz Physical Architecture

Machine Learning and AI

- Introduction to ML and AI in Big Data Applications
- Architect's Role in ML and AI-Driven Projects
- High-Level View of Machine Learning (ML) and Artificial Intelligence (AI)
- Big Data and ML/AI in Public Cloud vs. Private Cloud
- Common Challenges in ML/AI Architectures
- Best Practices for Architecting ML and AI in Big Data
- Emerging Trends

AI Studios

- Learn about AI Studios
- Explain core features of RAG Studio
- Explain core features of Agent Studio
- Build and Deploy Context-Aware Chatbots
- AI Agent tools

Potential Cloudera Solutions

- Review of Uber and Lyft Big data platforms
- Review of Metroz Cloudera solution architectures

Wrap Up