



Introduction to Cask Data Application Platform

The **Introduction to Cask Data Application Platform** (CDAP) course delivers the key concepts and expertise needed to build real-time and batch application on CDAP. This course is comprised of a combination of lectures and labs to reinforce core concepts. At the end of the course participants will be able to write CDAP applications, integrate applications in their CI environments and possess the skills to test and debug applications.

Audience & Prerequisites

This course is designed for developers and architects who are looking to use CDAP. Participants must be familiar with the basics of Java programming, debugging skills and should have basic knowledge about HDFS and MapReduce.

Materials Required

- Laptop with Mac, Windows or Linux on 64-bit OS.
- Virtual Box installed (<https://www.virtualbox.org/wiki/Downloads>)
- CDAP Standalone VM installed (Instructions provided separately)

Course Overview

- CDAP Concepts & Capabilities
- Data Ingestion
- Data Exploration
- Understanding Datasets
- Data Serving
- Batch Processing
- Scheduling & Sequencing Jobs using Workflow
- Spark Integration with CDAP
- Real-time processing



Course Outline

<p>1. The Motivation for CDAP</p> <ul style="list-style-type: none">• Context• The Hadoop Ecosystem Today• Challenges with Hadoop• The need for CDAP• CDAP benefits for developers <p>2. CDAP Overview</p> <ul style="list-style-type: none">• Functional view of CDAP• CDAP architecture• Building Blocks <p>3. Data Ingestion</p> <ul style="list-style-type: none">• Ingesting data in real-time• Ingesting data in batch• Managing the data lifecycle <p>4. Data Exploration</p> <ul style="list-style-type: none">• Exploring ingested data• Attaching schemas <p>5. Introduction to Datasets</p> <ul style="list-style-type: none">• The need for Datasets• Data abstraction using Datasets• Using Datasets in applications <p>6. Datasets on HBase</p> <ul style="list-style-type: none">• Introduction to TTable• Using Table Datasets• Comparison to relational databases	<p>7. Datasets on HDFS</p> <ul style="list-style-type: none">• Creating and using file based Datasets• Partitioned file Datasets <p>8. Data Serving</p> <ul style="list-style-type: none">• Using services in applications• Writing handlers to serve data <p>9. Batch Processing</p> <ul style="list-style-type: none">• Brief intro to MapReduce• Writing batch jobs in CDAP• Reading from Streams & Datasets• Writing data to Datasets <p>10. Scheduling using Workflow</p> <ul style="list-style-type: none">• Scheduling jobs using workflows• Creating time-based and size-based triggers <p>11. Spark Integration with CDAP</p> <ul style="list-style-type: none">• Writing Spark Programs in CDAP• Spark programming model• Managing Spark programs <p>12. Real-time Processing</p> <ul style="list-style-type: none">• Introduction to Tigon• Reading from Streams & Datasets• Writing to Datasets from Tigon Flows
---	--