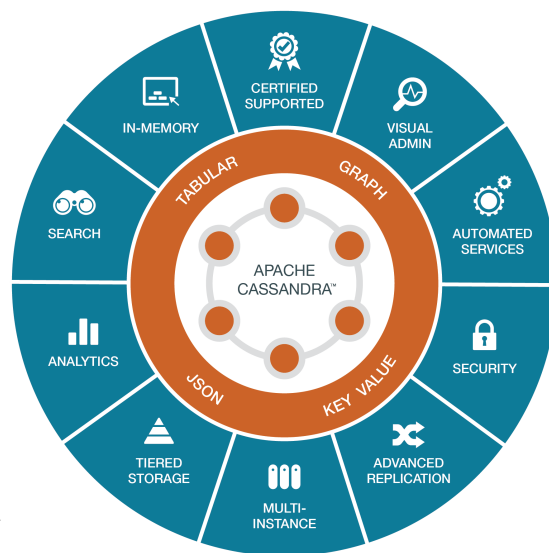# DS320: DataStax Enterprise Analytics with Apache Spark™

Instructor-Led Training

Overview

In this course, you will learn how to effectively and efficiently solve analytical problems with DataStax Enterprise Analytics. You will learn about Apache Cassandra™ and Apache Spark, including Spark API, Spark-Cassandra Connector, Spark SQL, Spark Streaming, and crucial performance optimization techniques.

Duration	3 days
Audience	Software professionals proficient with Cassandra core concepts and looking to learn DataStax Enterprise Analytics
Prerequisites	Completion of the DataStax Enterprise Foundations of Apache Cassandra™ online course. Linux command line and programming experience.

Course details

- Introduction
    - o Data Analytics with Cassandra and Spark
    - o Spark Architecture
    - o Spark Shell
    - o Spark Web UI
- Spark Essentials
    - o A Spark "Hello, World!"
    - o Resilient Distributed Dataset
    - o Creating an RDD
    - o RDD Transformations
    - o RDD Actions
- Connecting Spark to Cassandra

- Tuning and Partitioning
    - o Understanding Partitioning in Spark
    - o Partitioning Rules
    - o Controlling Partitioning
    - o Data Shuffling
- Optimizing with the Spark/Cassandra Connector
    - o Counting Rows in a Cassandra Table
    - o Grouping Rows Using Cassandra
    - o Joining Cassandra Tables
    - o Cassandra-Aware Partitioning
- Spark Streaming
    - o Discretized Stream

- o   Retrieving Data from Cassandra
- o   Working with Cassandra Rows
- o   Converting Cassandra Rows to Scala Tuples and Objects
- o   Saving Data to Cassandra
- Optimization is not Always Premature
  - o   Broadcast Variables
  - o   Accumulator Variables
  - o   RDD Persistence
- The Utility of Key/Value Pairs
  - o   Key-Value Pair RDD
  - o   Key-Value Pair RDDs: Aggregation
  - o   Key-Value Pair RDDs: Grouping and Sorting
  - o   Key-Value Pair RDDs: Joins
  - o   Key-Value Pair RDDs: Set Operations
- o   Spark Streaming Architecture
- o   Your First Spark Streaming Application
- o   DStream Stateless Transformations
- o   DStream Stateful Transformations
- o   DStream Window Transformations
- o   DStream Output Operations
- o   Checkpointing and Recovery
- o   DStream Persistence
- o   Controlling Parallelism
- Spark SQL
  - o   Spark SQL Basics
  - o   Creating DataFrames
  - o   Accessing DataFrame Schema and Rows
  - o   RDD Operations on DataFrames
  - o   Language-Integrated Queries
  - o   Saving DataFrames to Cassandra
  - o   Querying Cassandra with SQL
  - o   Writing Efficient SQL Queries

Benefits

Classroom training is delivered by DataStax experts, who will share first-hand experiences and practical use cases. A high level of interactivity with the trainer coupled with hands-on labs deliver an active learning experience.

A private, on-site training environment provides the option to tailor the delivery based on your needs and preferences. We will work with your team to determine areas requiring emphasis during delivery. Additionally, on-site training fosters teamwork and team learning that continue through the project.

For more information, please contact your account executive or send an email to training@datastax.com.

Terms

Please refer to the DataStax Services Terms for specific details regarding terms of Training delivery.

- This is a pre-paid Service offering.

- The Service expires 365 days after receipt of a valid customer order.

- The maximum class size is 20 attendees.

- This training is delivered onsite.

- A virtual classroom option is available, but we do not recommend mixing in-person and remote attendees. The remote attendees have limited learning experience in this environment.

DATASTAX