Spark and Machine Learning at Scale

Course Number: WA3290

Duration: 4 days

Course Description

This Spark and Machine Learning training teaches participants how to build, deploy, and maintain powerful data-driven solutions using Spark and its associated technologies. The course begins with an introduction to Spark, its architecture, and how it fits into the Hadoop and Cloud-based ecosystems. Participants will learn to set up Spark environments using DataBricks Cloud, AWS EMR clusters, and SageMaker Studio. In addition, students will learn about Spark's core functionalities, including RDDs, DataFrames, transformations, and actions.

Skills Gained

- Work with Spark's machine learning (ML) libraries, focusing on data preprocessing, feature engineering, model training, and evaluation.
- Perform stream processing and graph analysis with GraphX and Graphframes
- Deploy Spark ML artifacts
- Understand machine learning at scale
- Implement distributed training, hyperparameter tuning, model selection, and performance optimization for machine learning pipelines

Who Can Benefit

This course targets data scientists, machine learning engineers, big data engineers, and other professionals with experience in data analysis who wish to leverage Spark for scalable machine learning solutions. It is also suitable for those who want to enhance their large-scale data processing and machine learning knowledge.

Prerequisites

- Basic understanding of Python programming
- Familiarity with data processing and analysis concepts
- Familiarity with Python Pandas
- Familiarity with basic machine learning concepts and algorithms is recommended

Course Details

Outline

Chapter 1 - Introduction to Spark. Overview of Spark and its Architecture

- Big Data and the Analytics Process
- What is Big Data?
- Volume
- Velocity
- Variety
- Veracity
- Too large to fit into memory
- Big data and analytic process
- Scaling and Distributed Computing
- How to Actually Scale?
- Bring the Data to the Compute
- Bring the Compute to the Data
- Introduction to the Spark Platform
- History of Spark and Hadoop
- Spark vs. Hadoop MapReduce
- Supported Languages
- Pandas API on Spark
- Spark Architecture: Cluster Manager
- Standalone cluster manager
- Apache Hadoop YARN
- Apache Mesos
- Spark Architecture: Driver Process
- Spark Architecture: Executor Process and Workers
- Spark Building Blocks

- Spark SQL and the Catalyst

Chapter 2 - Introduction to Spark - Setting up a Spark Environment

- Set Up On-Premise Spark Environment (Ubuntu 20.04, Docker)
- Set Up DataBricks Community Cloud and Compute Cluster
- Set Up EMR Cluster and Attach Notebook

Chapter 3 - Basic Spark Operations and Transformations

- Spark Session and Context
- Loading Data
- Actions and Transformations
- More on Actions in Spark
- More on Transformations in Spark
- Persistence and Caching

Chapter 4 - Introduction to Spark SQL

- What is Spark SQL?
- Uniform Data Access with Spark SQL
- Integration with cloud storage
- Using JDBC Sources
- Hive Integration
- What is a DataFrame?
- Creating a DataFrame in PySpark
- Commonly Used DataFrame Methods and Properties in PySpark
- Grouping and Aggregation in PySpark
- The "DataFrame to RDD" Bridge in PySpark
- The SQLContext Object
- Examples of Spark SQL / DataFrame (PySpark Example)
- Converting an RDD to a DataFrame Example
- Example of Reading / Writing a JSON File
- Performance, Scalability, and Fault-tolerance of Spark SQL

Chapter 5 - Spark's ML libraries - Lecture: Introduction to Spark's ML libraries

- Spark MLlib
- Algorithms

- Classification
- Binary Classification
- Multi-Class Classification
- Multi-Label Classification
- Imbalanced Classification
- Regression
- Linear Regression
- Simple Linear Regression
- Multiple Linear Regression
- Polynomial Regression
- Support Vector Regression
- Decision Tree Regression
- Random Forest Regression
- Feature Engineering
- TF-IDF - PySpark example
- Word2Vec - PySpark example
- Count Vectorizer - PySpark example
- Feature Transformers of Spark MLlib
- Tokenizer - PySpark example
- Stopwords Remover
- Stopwords Remover - PySpark example
- N-gram - PySpark example
- Binarizer - PySpark example
- Principal Component Analysis
- What is PCA used for?
- Advantages and disdvantagesof PCA
- PCA - PySpark example
- String Indexing - PySpark example
- Why One-Hot Encoding is used for nominal data?
- One-Hot Encoding - PySpark Example
- Bucketizer - PySpark example
- Standardization and Normalization
- Difference between Standardization and Normalization

- Standard Scaler
- Robust Scaler
- Min Max Scaler
- Max Abs Scaler
- Imputer
- Feature Selectors in Spark MLlib
- Vector Slicer - PySpark example
- Chi-Squared selection - PySpark example
- Univariate Feature Selector
- Variance Threshold Selector
- Locality Sensitive Hashing
- Locality Sensitive Hashing in Spark MLlib
- LSH Operations
- Locality Sensitive Hashing in Spark MLlib
- Bucketed Random Projection for Euclidean Distance
- MinHash for Jaccard Distance
- Pipeline
- Transformer
- Estimator
- Persistence
- Introduction to Hyperparameter Tuning
- Hyperparameter tuning methods
- Random Search
- Grid Search
- Bayesian Optimisation
- Hyperparameter Tuning with Spark

Chapter 6 - Streaming and Graphs

- Stream Analytics
- Tools for Stream Analytics: Kafka, Storm, Flink, Spark
- Timestamps in stream analytics
- Windowing Operations

Chapter 7 - Deploying Spark ML Artifacts - Introduction to deploying Spark ML Artifacts

- How the Spark system works
- What is Deployment?
- Spark Deployment Artifacts
- Packaging Spark (ML) for Production
- Deploy Spark ML to EMR
- Deploy Spark (ML) with Sagamaker
- Serving and Updating Spark ML Models
- Model Versioning with AWS Model Registry

Chapter 8 - Machine learning at Scale - Introduction to Machine Learning at Scale

- Introduction to Scalability
- Common Reasons for Scaling Up ML Systems
- How to Avoid Scaling Infrastructure?
- Benefits of ML at Scale
- Challenges in ML Scalability
- Data Complexities - Challenges
- ML System Engineering - Challenges
- Integration Risks - Challenges
- Collaboration Issues - Challenges

Chapter 9 - Machine learning at Scale - Distributed Training of Machine Learning models

- Introduction to Distributed Training
- Data Parallelism
- Steps of Data Parallelism
- Data Parallelism vs. Random Forest
- Model Parallelism
- Frameworks for Implementing Distributed ML
- Introduction to Distributed Training vs. Distributed Inference
- Introduction to Training
- Introduction to Inference
- Key components of Inference
- Inference Challenges
- Training vs. Inference
- Introduction to GPUs

- Inference - Hardware
- AWS Inferentia Chip vs GPU

Chapter 10 - Machine learning at Scale - Hyperparameter tuning and model selection at scale

- Hyperparameter Tuning at Scale
- Hyperparameter Tuning Challenges
- Distributed Hyperparameter Tuning
- Bayesian Optimization
- Distributed Hyperparameter Tuning
- Spark Based Tools
- TensorFlowOnSpark
- Advantages of TensorFlowOnSpark
- BigDL
- Advantages of BigDL
- Horovod
- Advantages of Horovod
- H2O Sparkling Water
- Advantages of Sparkling Water over H2O

Lab Exercises

- Lab 1. Spark Introduction Lab
- Lab 2. Spark Setup Lab
- Lab 3. Installing graphframes in DCC