1. Record Nr. UNINA9910506385403321

Autore	Hien Luu
Titolo	Beginning Apache Spark 3 : with DataFrame, Spark SQL, structured streaming, and Spark machine learning library / / Hien Luu
Pubbl/distr/stampa	New York, New York : , : Apress, , [2021] ©2021
ISBN	1-4842-7383-4
Edizione	[Second edition.]
Descrizione fisica	1 online resource (445 pages)
Disciplina	005.7
Soggetti	Big data Distributed databases Open source software Machine learning
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Note generali	Includes index.
Nota di contenuto	Intro -- Table of Contents -- About the Author -- About the Technical Reviewers -- Acknowledgments -- Introduction -- Chapter 1: Introduction to Apache Spark -- Overview -- History -- Spark Core Concepts and Architecture -- Spark Cluster and Resource Management System -- Spark Applications -- Spark Drivers and Executors -- Spark Unified Stack -- Spark Core -- Spark SQL -- Spark Structured Streaming -- Spark MLlib -- Spark GraphX -- SparkR -- Apache Spark 3.0 -- Adaptive Query Execution Framework -- Dynamic Partition Pruning (DPP) -- Accelerator-aware Scheduler -- Apache Spark Applications -- Spark Example Applications -- Apache Spark Ecosystem -- Delta Lake -- Koalas -- MLflow -- Summary -- Chapter 2: Working with Apache Spark -- Downloading and Installation -- Downloading Spark -- Installing Spark -- Spark Scala Shell -- Spark Python Shell -- Having Fun with the Spark Scala Shell -- Useful Spark Scala Shell Command and Tips -- Basic Interactions with Scala and Spark -- Basic Interactions with Scala -- Spark UI and Basic Interactions with Spark -- Spark UI -- Basic Interactions with Spark -- Introduction to Collaborative Notebooks -- Create a Cluster -- Create a Folder -- Create a Notebook -- Setting up Spark Source Code --

Summary -- Chapter 3: Spark SQL: Foundation -- Understanding RDD -- Introduction to the DataFrame API -- Creating a DataFrame -- Creating a DataFrame from RDD -- Creating a DataFrame from a Range of Numbers -- Creating a DataFrame from Data Sources -- Creating a DataFrame by Reading Text Files -- Creating a DataFrame by Reading CSV Files -- Creating a DataFrame by Reading JSON Files -- Creating a DataFrame by Reading Parquet Files -- Creating a DataFrame by Reading ORC Files -- Creating a DataFrame from JDBC -- Working with Structured Operations -- Working with Columns -- Working with Structured Transformations.
select(columns) -- selectExpr(expressions) -- filler(condition), where (condition) -- distinct, dropDuplicates -- sort(columns), orderBy (columns) -- limit(n) -- union(otherDataFrame) -- withColumn (colName, column) -- withColumnRenamed(existingColName, newColName) -- drop(columnName1, columnName2) -- sample (fraction), sample(fraction, seed), sample(fraction, seed, withReplacement) -- randomSplit(weights) -- Working with Missing or Bad Data -- Working with Structured Actions -- describe (columnNames) -- Introduction to Datasets -- Creating Datasets -- Working with Datasets -- Using SQL in Spark SQL -- Running SQL in Spark -- Writing Data Out to Storage Systems -- The Trio: DataFrame, Dataset, and SQL -- DataFrame Persistence -- Summary -- Chapter 4: Spark SQL: Advanced -- Aggregations -- Aggregation Functions -- Common Aggregation Functions -- count(col) -- countDistinct(col) -- min(col), max(col) -- sum(col) -- sumDistinct(col) -- avg(col) -- skewness(col), kurtosis(col) -- variance(col), stddev(col) -- Aggregation with Grouping -- Multiple Aggregations per Group -- Collection Group Values -- Aggregation with Pivoting -- Joins -- Join Expression and Join Types -- Working with Joins -- Inner Joins -- Left Outer Joins -- Right Outer Joins -- Outer Joins (a.k.a. Full Outer Joins) -- Left Anti-Joins -- Left Semi-Joins -- Cross (a.k.a. Cartesian) -- Dealing with Duplicate Column Names -- Use Original DataFrame -- Renaming Column Before Joining -- Using Joined Column Name -- Overview of Join Implementation -- Shuffle Hash Join -- Broadcast Hash Join -- Functions -- Working with Built-in Functions -- Working with Date Time Functions -- Working with String Functions -- Working with Math Functions -- Working with Collection Functions -- Working with Miscellaneous Functions -- Working with User-Defined Functions (UDFs) -- Advanced Analytics Functions.
Aggregation with Rollups and Cubes -- Rollups -- Cubes -- Aggregation with Time Windows -- Window Functions -- Exploring Catalyst Optimizer -- Logical Plan -- Physical Plan -- Catalyst in Action -- Project Tungsten -- Summary -- Chapter 5: Optimizing Spark Applications -- Common Performance Issues -- Spark Configurations -- Different Ways of Setting Properties -- Different Kinds of Properties -- Viewing Spark Properties -- Spark Memory Management -- Spark Driver -- Spark Executor -- Leverage In-Memory Computation -- When to Persist and Cache Data -- Persistence and Caching APIs -- Persistence and Caching Example -- Understanding Spark Joins -- Broadcast Hash Join -- Shuffle Sort Merge Join -- Adaptive Query Execution -- Dynamically Coalescing Shuffle Partitions -- Dynamically Switching Join Strategies -- Dynamically Optimizing Skew Joins -- Summary -- Chapter 6: Spark Streaming -- Stream Processing -- Concepts -- Data Delivery Semantics -- Notion of Time -- Windowing -- Stream Processing Engine Landscape -- Spark Streaming Overview -- Spark DStream -- Spark Structured Streaming -- Overview -- Core Concepts -- Data Sources -- Output Modes -- Trigger Types -- Data Sinks -- Watermarking -- Structured Streaming Applications --

Streaming DataFrame Operations -- Selection, Project, Aggregation Operations -- Join Operations -- Working with Data Sources -- Working with a Socket Data Source -- Working with a Rate Data Source -- Working with a File Data Source -- Working with a Kafka Data Source -- Working with a Custom Data Source -- Working with Data Sinks -- Working with a File Data Sink -- Working with a Kafka Data Sink -- Working with a foreach Data Sink -- Working with a Console Data Sink -- Working with a Memory Data Sink -- Output Modes -- Triggers -- Summary -- Chapter 7: Advanced Spark Streaming -- Event Time. Fixed Window Aggregation over an Event Time -- Sliding Window Aggregation over Event Time -- Aggregation State -- Watermarking: Limit State and Handle Late Data -- Arbitrary Stateful Processing -- Arbitrary Stateful Processing with Structured Streaming -- Handling State Timeouts -- Arbitrary State Processing in Action -- Extracting Patterns with mapGroupsWithState -- User Sessionization with flatMapGroupsWithState -- Handling Duplicate Data -- Fault Tolerance -- Streaming Application Code Change -- Spark Runtime Change -- Streaming Query Metrics and Monitoring -- Streaming Query Metrics -- Monitoring Streaming Queries via Callback -- Monitoring Streaming Queries via Visualization UI -- Streaming Query Summary Information -- Streaming Query Detailed Statistics Information -- Troubleshooting Streaming Query -- Summary -- Chapter 8: Machine Learning with Spark -- Machine Learning Overview -- Machine Learning Terminologies -- Machine Learning Types -- Supervised Learning -- Unsupervised Learning -- Reinforcement Learning -- Machine Learning Development Process -- Spark Machine Learning Library -- Machine Learning Pipelines -- Transformers -- Estimators -- Pipeline -- Pipeline Persistence: Saving and Loading -- Model Tuning -- Speeding Up Model Tuning -- Model Evaluators -- Machine Learning Tasks in Action -- Classification -- Model Hyperparameters -- Example -- Regression -- Model Hyperparameters -- Example -- Recommendation -- Model Hyperparameters -- Example -- Deep Learning Pipeline -- Summary -- Chapter 9: Managing the Machine Learning Life Cycle -- The Rise of MLOps -- MLOps Overview -- MLflow Overview -- MLflow Components -- MLflow in Action -- MLflow Tracking -- MLflow Projects -- MLflow Models -- MLflow Model Registry -- Model Deployment and Prediction -- Summary -- Index.

Sommario/riassunto

Take a journey toward discovering, learning, and using Apache Spark 3.0. In this book, you will gain expertise on the powerful and efficient distributed data processing engine inside of Apache Spark; its user-friendly, comprehensive, and flexible programming model for processing data in batch and streaming; and the scalable machine learning algorithms and practical utilities to build machine learning applications. Beginning Apache Spark 3 begins by explaining different ways of interacting with Apache Spark, such as Spark Concepts and Architecture, and Spark Unified Stack. Next, it offers an overview of Spark SQL before moving on to its advanced features. It covers tips and techniques for dealing with performance issues, followed by an overview of the structured streaming processing engine. It concludes with a demonstration of how to develop machine learning applications using Spark MLlib and how to manage the machine learning development lifecycle. This book is packed with practical examples and code snippets to help you master concepts and features immediately after they are covered in each section. After reading this book, you will have the knowledge required to build your own big data pipelines, applications, and machine learning applications. What You Will Learn Master the Spark unified data analytics engine and its various

components Work in tandem to provide a scalable, fault tolerant and performant data processing engine Leverage the user-friendly and flexible programming model to perform simple to complex data analytics using dataframe and Spark SQL Develop machine learning applications using Spark MLlib Manage the machine learning development lifecycle using MLflow Who This Book Is For Data scientists, data engineers and software developers.