Lernmodul zur Bearbeitung großer Datenmengen mit Apache Spark

SparkSession

Starten einer SparkSession namens <name> in einem Cluster mit <clustertyp>	spark = SparkSession.builder.master(clustertype) .appName(name).getOrCreate()
Stoppen einer SparkSession	spark.stop()

RDD Basics

RDD aus Collection erstellen (ParallelizeCollectionRDD)	spark.sparkContext.parallelize(collection)
Anzahl Partitionen eines RDDs ausgeben	rdd.getNumPartitions()
Gesamtes RDD anzeigen	rdd.collect()
Die ersten n Zeilen eines RDDs anzeigen	rdd.take(n)
Sortieren eines RDDs	rdd.sortByKey()
Gruppieren eines RDDs	rdd.groupByKey()

DataFrame Basics

DataFrame aus Collection erstellen	spark.createDataFrame(data=collection, schema=schema)
DataFrame aus CSV-Datei erstellen (mit Schema und Spaltennamen)	spark.read.csv(filepath, *inferSchema=True*, *header=True*)
DataFrame aus Datenbanktabelle	spark.read.jdbc(url, tablename)
Schema erzeugen	StructType([StructField(columnname, type, nullable), ...])
DataFrame anzeigen	df.show()
Die ersten n Zeilen eines DataFrames anzeigen	df.show(n)
Spalte(n) eines DataFrames auswählen	df.select(columnname(s)) /df.select(df.columnname, ..)
Schema eines DataFrames anzeigen	df.printSchema()

Statistische Informationen zu einer Spalte/ mehreren Spalten anzeigen	df.describe(columnname(s))
Anzahl der Einträge eines DataFrames ausgeben	df.count()
Sortieren eines DataFrames	df.sort(columnname) / df.orderBy(columnname)
Gruppieren eines DataFrames	df.groupBy(columnname)
SQL-Abfrage auf temporäre View eines DataFrames	df.createOrReplaceTempView(viewname) spark.sql(query)
Zusammenfügen von DataFrames über einer Spalte	df1.join(df2, df1.columnname == df2.columnname, *how*)
Zusammenfügen von DataFrames mit gleichem Schema	df1.union(df2)

Analyse, Filterung und Bereinigung von DataFrames

Filtern von DataFrames	df.filter(function) / df.where(function)
Duplikate entfernen	df.dropDuplicates()
Nullwerte ersetzen	df.fillna(replacement) / df.na.fill(replacement)
Nullwerte entfernen	df.dropna() / df.na.drop()
Einzigartige Werte einer Spalte ausgeben	df.select(columnname).distinct()

Transformation von DataFrames

Spalten eines DataFrames verändern	df.withColumn(columnname, function)
Datentyp einer Spalte ändern	df.columnname.cast(type)
Stringwerte ersetzen	df.withColumn(columname, regexp_replace(columnname, oldvalue, newvalue)
Numerische Werte ersetzen	df.withColumn(columnname, replace(columnname, oldvalue, newvalue)
Spalten eines DataFrames umbenennen	df.withColumnRenamed(oldname, newname)
Spalte eines DataFrames entfernen	df.drop(columnname)
Einem DataFrame eine Spalte hinzufügen	df.select(columnname(s), function.alias(columnname))
User Defined Function erstellen	udf(lambda function)