A to Z

Curated by
Deepa Vasanthkumar

A

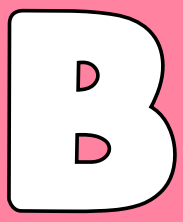
Action	Operations that trigger the execution of the RDD transformations and return a result to the driver program or write it to storage.
API	A set of functions and protocols for building software and applications, allowing interaction with Apache Spark.
Apache Spark	An open-source distributed computing system for big data processing and analytics.

DEEPA VASANTHKUMAR

A

Adaptive Query Execution (AQE)	A feature in Spark that dynamically adjusts query plans at runtime based on the actual data being processed, leading to optimized performance.
Aggregation	A process of combining multiple values into a single value. Spark provides various aggregation functions such as sum(), avg(), count(), and custom aggregations using aggregate().

DEEPA VASANTHKUMAR



Broadcast variables	Variables cached on each machine, instead of shipping a copy of it with tasks, to improve performance when tasks across stages need the same data.
Batch Processing	Processing of large blocks of data at once, as opposed to real-time or streaming data processing. Spark is often used for batch processing in ETL workflows.
Bucketing	A technique for partitioning data into a fixed number of buckets to optimize join operations and aggregation tasks.

C

Caching	The process of storing data in memory to speed up retrieval during processing.
Cluster Manager	Manages resources in a cluster, such as YARN, Mesos, or Spark's standalone cluster manager.
Core	The basic unit of parallelism in Spark. It is a thread running a task.

DEEPA

C

Catalyst Optimizer	The query optimization framework in Spark SQL that applies a series of optimization rules to improve the execution plan of queries.
Checkpointing	A process of saving the intermediate state of an RDD or DataFrame to reliable storage, which is useful for long-running computations and fault tolerance.
Columnar Storage	A data storage format that stores data in columns rather than rows, which is beneficial for analytical queries that typically access a subset of columns. Spark supports columnar formats like Parquet and ORC.

DEEPA

D

Dataframe	A distributed collection of data organized into named columns, similar to a table in a relational database.
Dataset	A strongly-typed, immutable collection of objects that can be manipulated using functional transformations (map, flatMap, filter, etc.).
Driver Program	The main program that creates the SparkContext, connects to the cluster, and coordinates the execution of the tasks.

DEEPA
VASANTHKUMAR

D

Delta Lake	An open-source storage layer that brings ACID transactions to Apache Spark and big data workloads, enabling reliable data lakes with data versioning and time travel.
DAG (Directed Acyclic Graph)	A representation of a sequence of computations to be performed on data. In Spark, a DAG is used to track the lineage of operations and optimize execution.

DEEPA
VASANTHKUMAR

E

Executor	A distributed agent responsible for executing a task on a worker node and returning the results to the driver program.
ETL (Extract, Transform, Load)	A process in data warehousing and data integration that involves extracting data from source systems, transforming it to fit business needs, and loading it into a target data store.
Event Time	The time at which events actually occurred, as opposed to processing time when events are processed by the system. Spark Structured Streaming supports event-time processing.

DEEPA VASANTHKUMAR

F

Fault Tolerance	The ability of Spark to recover from node failures and recompute lost data.
Functional Programming	A programming paradigm in Spark where functions are treated as first-class citizens and operations are performed using transformations.
FlatMap	A transformation that applies a function to each element of an RDD or DataFrame and returns a new RDD or DataFrame by flattening the results.

DEEPA VASANTHKUMAR

F

Framework	A collection of libraries and tools that provides a structured approach to building and managing applications. Spark is a big data processing framework.
Filter	A transformation that returns a new RDD or DataFrame containing only the elements that satisfy a given condition.

FlatMap

DEEPA VASANTHKUMAR

G

GraphX	A component of Spark for graph processing and analysis.
Gradient Descent	An optimization algorithm used in machine learning to minimize a function by iteratively moving towards the steepest descent.
GroupByKey	A transformation that groups the values of a key-value pair RDD by key. It returns a new RDD of (key, Iterable<values>) pairs.

DEEPA

SANTHKUMAR

H

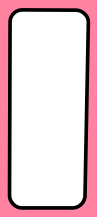
Hadoop	An open-source framework for distributed storage and processing of large datasets. Spark can run on Hadoop clusters and use Hadoop's HDFS.
Hive	A data warehousing solution built on top of Hadoop that allows querying and managing large datasets using SQL. Spark can use Hive for reading and writing data.

DEEPA VASANTHKUMAR

H

HDFS (Hadoop Distributed File System)	A distributed file system designed to store large datasets across multiple nodes. Spark can read from and write to HDFS.
HiveContext	A class in Spark that allows querying data using the HiveQL language. It is part of the Spark SQL module and provides compatibility with Hive.
Hive Metastore	A database that stores metadata about Hive tables, such as schema and location. Spark SQL can integrate with Hive Metastore to access this metadata.

DEEPA
VASANTHKUMAR



In-memory Computing	Storing data in memory to improve performance instead of reading from disk storage.
Iterator	An object in Spark that allows traversing through a collection of elements one at a time.
InputFormat	A class in Hadoop (and used by Spark) that defines how input data is split and read into the system. Examples include TextInputFormat and SequenceFileInputFormat.

DEEPA

J

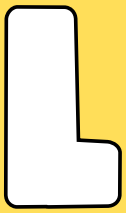
Job	A sequence of tasks submitted to the cluster for execution, generated by a Spark action.
Join	A transformation that combines two RDDs or DataFrames based on a common key. Types of joins include inner join, outer join, left join, and right join.
Job Scheduler	The component of Spark that handles the scheduling of jobs, dividing them into stages and tasks, and distributing them across the cluster for execution.

DEEPA VASANTHKUMAR

K

Kryo	A serialization library used by Spark for fast and efficient serialization of objects.
Kafka	A distributed streaming platform that Spark can integrate with for processing real-time data streams.
Kinesis	A real-time data streaming service provided by AWS. Spark can read data from and write data to Kinesis streams for real-time data processing.

DEEPA VASANTHKUMAR



Lazy Evaluation	A technique used in Spark where transformations on RDDs are not immediately executed but are recorded in a lineage graph for optimization.
Lineage	A record of the transformations applied to an RDD, used for fault tolerance and recomputation.
Logical Plan	An abstract, high-level representation of a query that describes what operations need to be performed. Spark's Catalyst Optimizer generates and optimizes logical plans before converting them into physical plans for execution.

DEEPA

M

MapReduce	A programming model for processing large datasets. Spark can execute MapReduce tasks much faster due to its in-memory computing capabilities.
MLlib	A machine learning library in Spark providing various algorithms and utilities for scalable machine learning.

DEEPA VASANTHKUMAR

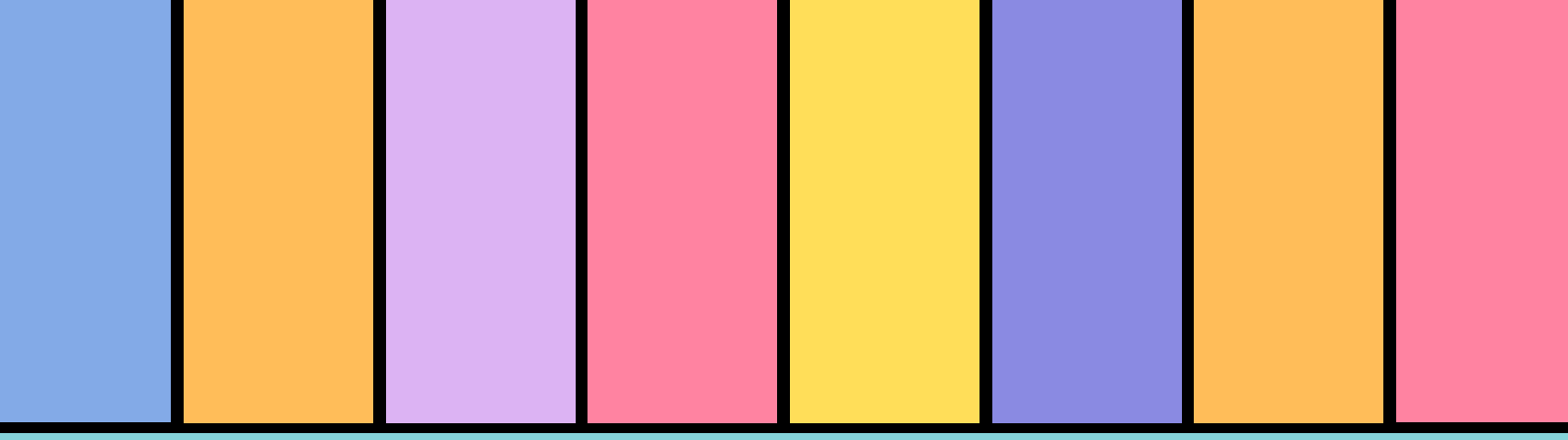
M

Map	A narrow transformation that applies a function to each element of an RDD or DataFrame, returning a new RDD or DataFrame with the results.
MapPartitions	A transformation that applies a function to each partition of an RDD or DataFrame, rather than to each element. This can be more efficient for certain operations.

DEEPA VASANTHKUMAR

N

Node	A single machine in a Spark cluster that can be a driver or a worker.
Notebook	An interactive environment, such as Databricks Notebooks or Jupyter Notebooks, where users can write and execute Spark code, visualize data, and document their analysis.
Normalization	The process of structuring data to reduce redundancy and improve data integrity. In Spark, this often involves transforming and cleaning data before analysis.



O

Operations	Functions in Spark that can be applied to RDDs, such as transformations and actions.
Optimization	Techniques used to improve the performance of Spark jobs, including query optimization and execution plan tuning.
OutputFormat	A class in Hadoop (and used by Spark) that defines how the output data is written. Examples include TextOutputFormat and SequenceFileOutputFormat.

DEEPA

VASANTHKUMAR

P

Partition	A division of data in an RDD or DataFrame that can be processed in parallel.
Persist	Storing an RDD in memory across operations for faster access.
PySpark	The Python API for Spark, allowing the use of Spark with Python.

DEEPA

P

Pipeline	A sequence of data processing steps, often used in machine learning workflows. Spark's MLlib provides APIs to create and manage pipelines.
Physical Plan	A detailed, low-level representation of how a query will be executed in Spark. It is generated by the Catalyst Optimizer from the logical plan.

DEEPA

Q

Query	A request for data retrieval and processing in Spark, often written in SQL or using DataFrame/Dataset APIs.
Query Execution Plan	The series of steps Spark takes to execute a query, which includes both the logical plan and the physical plan.

DEEPA
VASANTHKUMAR

R

RDD (Resilient Distributed Dataset)	The fundamental data structure in Spark, representing an immutable, distributed collection of objects.
Resource Manager	Manages the allocation of resources in a cluster for Spark jobs.
Range Partitioning	A partitioning strategy where data is divided into ranges based on a key. This can optimize performance for range queries and joins.

DEEPA

VASANTHKUMAR

R

ReduceByKey	A wide transformation that merges the values of each key using an associative reduce function. It is more efficient than groupByKey because it performs partial aggregation locally before shuffling the data.
Repartition	A transformation that reshuffles the data in an RDD or DataFrame into a specified number of partitions. This can be used to increase or decrease the level of parallelism.
ROW	A record in a DataFrame, representing a single entry with fields corresponding to the columns of the DataFrame.

DEEPA
VASANTHKUMAR

S

SparkContext	The main entry point for Spark functionality, responsible for connecting to the cluster and creating RDDs.
Spark SQL	A Spark module for working with structured data using SQL queries.
Streaming	Processing real-time data streams in Spark, using the Spark Streaming API.

DEEPA

VASANTHKUMAR

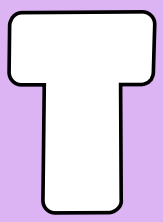
S

Schema	The structure that defines the organization of data in a DataFrame or Dataset, including column names and data types.
Shuffle	A process of redistributing data across partitions that involves moving data between executors. It typically occurs during wide transformations like groupByKey, reduceByKey, and join.
SparkSession	The entry point for programming Spark applications with the DataFrame and Dataset API. It replaces the older SQLContext and HiveContext.

S

Stage	A set of tasks that can be executed in parallel during the execution of a Spark job. A job is divided into stages based on shuffle boundaries.
Structured Streaming	An API in Spark for stream processing that allows you to process data in real-time using high-level declarative queries similar to batch processing.

DEEPA VASANTHKUMAR



Task	A unit of work that runs on a single executor and is a part of a job.
Transformation	Operations that create a new RDD from an existing one, such as map, filter, and reduceByKey.
Triggers	A mechanism in Structured Streaming that specifies when the system should process the next set of data. Examples include continuous processing and micro-batch intervals

DEEPA
VASANTHKUMAR

U

UDF (User-Defined Function)	Custom functions defined by users to extend the capabilities of Spark SQL.
Unpersist	Releasing the memory used by a cached RDD.
Unit Tests	Tests that validate the functionality of individual components of Spark applications. Libraries like spark-testing-base can help write unit tests for Spark applications.

V

View	A temporary table in Spark SQL created from a DataFrame.
Vectorization	A technique used in Spark MLlib to represent data in a format that can be efficiently processed by machine learning algorithms.

DEEPA VASANTHKUMAR

W

Worker Node	A node in a Spark cluster that executes tasks and returns results to the driver.
Wide Transformation	Transformations that require data shuffling across nodes, such as groupByKey and reduceByKey.
Window Function	A function that performs a calculation across a set of table rows related to the current row. Spark SQL supports window functions for operations like ranking, aggregations, and analytic functions.

DEEPA

X

XML	A markup language that Spark can read and write using libraries and custom parsers.

DEEPA VASANTHKUMAR

Y

YARN (Yet Another Resource Negotiator)	A cluster management technology used by Spark for resource allocation and job scheduling.

DEEPA VASANTHKUMAR

Z

Zookeeper	A centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services, often used in Spark streaming for managing offsets in Kafka.
Z-Order	A multidimensional clustering method used in Delta Lake to optimize data skipping and improve query performance by clustering data in a way that enhances locality for multiple columns.
Zeppelin	An open-source web-based notebook that enables interactive data analytics. Apache Zeppelin supports multiple languages and can be used with Spark for interactive data exploration and