



Shwetank Singh
GritSetGrow - GSGLearn.com



Data

Engineering 101

SPARK

ALL CONCEPTS TO GET STARTED

APACHE
spark™

Driver Program

The main program that runs the user's application and creates the `SparkContext`.

Example

Initializing `SparkContext`:
`from pyspark import
SparkContext; sc = SparkContext("local",
"App").`



Shwetank Singh
GritSetGrow - GSGLearn.com



Executors

Worker nodes that execute tasks and store data for the application.

Example

Executors process tasks and cache data in memory.



Shwetank Singh
GritSetGrow - GSGLearn.com



Cluster Manager

Manages resources and schedules jobs across the cluster.

Example

Using YARN, Mesos, or Standalone mode for resource management.



Shwetank Singh
GritSetGrow - GSGLearn.com



SparkContext

The entry point to Spark functionality.

Example

```
sc = SparkContext("local", "App")
```



Shwetank Singh
GritSetGrow - GSGLearn.com



SparkSession

The entry point for DataFrame and SQL functionality.

Example

```
from pyspark.sql import SparkSession;  
spark = SparkSession.builder.appName("App")  
.getOrCreate()
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Resilient Distributed Datasets (RDDs)

Immutable distributed collections of objects
that can be processed in parallel.

Example

```
rdd = sc.parallelize([1, 2, 3, 4, 5]).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Transformations

Operations that create a new RDD from an existing one.

Example

```
rdd2 = rdd.map(lambda x: x * 2).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Actions

Operations that trigger computation and return a result to the driver.

Example

`count = rdd.count()`.



Shwetank Singh
GritSetGrow - GSGLearn.com



Lazy Evaluation

Transformations are not executed until an action is called.

Example

Defining transformations but no computation until action:

```
rdd2 = rdd.map(x => x * 2);
```

```
rdd2.count().
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Directed Acyclic Graph

(DAG)

A sequence of computations represented as a graph.

Example

Viewing the DAG in Spark UI.



Shwetank Singh
GritSetGrow - GSGLearn.com



Spark SQL

A module for working with structured data using SQL.

Example

```
df = spark.sql("SELECT * FROM table").
```



Shwetank Singh
GritSetGrow - GSGLearn.com



DataFrame

A distributed collection of data organized into named columns.

Example

```
df = spark.read.json("file.json").
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Dataset

A strongly-typed collection of objects.

Example

```
ds = spark.createDataset([(1, "Alice"), (2, "Bob")]).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Catalyst Optimizer

Spark's query optimizer that generates efficient execution plans.

Example

Analyzing query plans with `df.explain()`.



Shwetank Singh
GritSetGrow - GSGLearn.com



PySpark

The Python API for Spark, allowing Python developers to use Spark's features.

Example

```
from pyspark.sql import SparkSession;  
spark =  
SparkSession.builder.appName("App").getOrCreate().
```



Shwetank Singh
GritSetGrow - GSGLearn.com



SQL

Allows execution of SQL queries in PySpark.

Example

```
spark.sql("SELECT COUNT(*) FROM table").
```



Shwetank Singh
GritSetGrow - GSGLearn.com



DataFrame

Similar to pandas DataFrame, but distributed across a cluster.

Example

```
df = spark.read.csv("file.csv").
```



Shwetank Singh
GritSetGrow - GSGLearn.com



RDD

Low-level API for distributed data manipulation.

Example

```
rdd = sc.textFile("file.txt").
```



Shwetank Singh
GritSetGrow - GSGLearn.com



UDF

User-defined functions for extending PySpark's built-in functions.

Example

```
spark.udf.register("add_one", lambda x: x + 1,  
IntegerType()).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Window Functions

Functions that allow calculations across a set of rows related to the current row.

Example

```
df.withColumn("rank", row_number()  
.over(Window.partitionBy("dept")  
.orderBy("salary"))).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Machine Learning

Using MLlib for scalable machine learning in PySpark.

Example

```
from pyspark.ml.classification import  
LogisticRegression; lr = LogisticRegression().
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Streaming

Processing real-time data streams with Spark Streaming.

Example

```
from pyspark.streaming import StreamingContext;  
ssc = StreamingContext(sc, 1).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



GraphFrames

A library for graph processing using PySpark.

Example

```
from graphframes import GraphFrame; g =  
GraphFrame(vertices, edges).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Accumulators

Variables that can only be "added" to, used for aggregating information across executors.

Example

```
accum = sc.accumulator(0).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Broadcast Variables

Variables that are cached on each machine to efficiently distribute large values.

Example

```
broadcastVar = sc.broadcast([1, 2, 3]).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Partitioning

Dividing data into partitions for parallel processing.

Example

`rdd = rdd.repartition(4).`



Shwetank Singh
GritSetGrow - GSGLearn.com



Shuffling

Redisistributing data across different nodes,
required by certain transformations like
`reduceByKey`.



Shwetank Singh
GritSetGrow - GSGLearn.com



Caching

Storing RDDs in memory for faster access during iterative algorithms.

Example

rdd.cache().



Shwetank Singh
GritSetGrow - GSGLearn.com



Persisting

Storing RDDs with different storage levels
(memory, disk, etc.).

Example

```
rdd.persist(StorageLevel.MEMORY_AND_DISK).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Checkpointing

Saving RDDs to reliable storage to truncate lineage graphs and provide fault tolerance.

Example
`rdd.checkpoint()`.



Shwetank Singh
GritSetGrow - GSGLearn.com



Fault Tolerance

Recovering lost data using lineage information.

Example

Handling node failures by recomputing lost partitions.



Shwetank Singh
GritSetGrow - GSGLearn.com



Lineage

Tracking the sequence of transformations that produced an RDD.

Example

Viewing lineage with `rdd.toDebugString`.



Shwetank Singh
GritSetGrow - GSGLearn.com



MapReduce

A programming model for processing large datasets with a distributed algorithm on a cluster.

Example

Implementing a simple MapReduce job:
`rdd.map(...).reduceByKey(...).`



Shwetank Singh
GritSetGrow - GSGLearn.com



HDFS

Integration with Hadoop Distributed File System for data storage.

Example

```
rdd = sc.textFile("hdfs://path/to/file").
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Data Sources

Reading and writing data from various sources like HDFS, S3, JDBC, and more.

Example

```
df = spark.read.jdbc(url, table, properties).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



SQL Functions

Built-in functions for data manipulation and analysis in Spark SQL.

Example

```
df.select(concat(col("name"), lit(" "),  
col("surname"))).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Column Pruning

An optimization technique that only reads necessary columns from the data source.

Example

Using column pruning with DataFrame API.



Shwetank Singh
GritSetGrow - GSGLearn.com



Predicate Pushdown

An optimization technique that pushes filter conditions to the data source.



Shwetank Singh
GritSetGrow - GSGLearn.com



Catalyst Optimization

Spark's internal optimization framework for generating efficient execution plans.

Example

Viewing optimized logical plan with
`df.queryExecution.optimizedPlan`.



Shwetank Singh
GritSetGrow - GSGLearn.com



Tungsten

An execution engine in Spark SQL for memory management and binary processing.

Example

Enabling Tungsten optimization:
`spark.sql.tungsten.enabled = true.`



Shwetank Singh
GritSetGrow - GSGLearn.com



Job

A complete computation expressed as a high-level action like count or saveAsTextFile.

Example

rdd.count().



Shwetank Singh
GritSetGrow - GSGLearn.com



Stage

A set of parallel tasks that will be executed as a unit.

Example

Viewing stages in Spark UI for a job.



Shwetank Singh
GritSetGrow - GSGLearn.com



Task

The smallest unit of work in Spark, executed by an executor.

Example

Monitoring task execution in Spark UI.



Shwetank Singh
GritSetGrow - GSGLearn.com



Cache

Storing RDDs in memory to speed up repeated access.

Example
`rdd.cache()`.



Shwetank Singh
GritSetGrow - GSGLearn.com



Persist

Storing RDDs with different storage levels
(memory, disk, etc.).

Example

```
rdd.persist(StorageLevel.MEMORY_AND_DISK).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Checkpoint

Saving an RDD to a reliable storage to truncate the lineage graph and provide fault tolerance.

Example
`rdd.checkpoint()`.



Shwetank Singh
GritSetGrow - GSGLearn.com



Fault Tolerance

Spark's ability to recompute lost data using lineage information.

Example

Handling node failures by recomputing lost partitions.



Shwetank Singh
GritSetGrow - GSGLearn.com



Lineage

The logical execution plan that Spark builds to keep track of the transformations applied to an RDD.

Example

Viewing lineage with `rdd.toDebugString`.



Shwetank Singh
GritSetGrow - GSGLearn.com



MapReduce

A programming model for processing large datasets with a parallel, distributed algorithm.

Example

Implementing a simple MapReduce job in Spark: `rdd.map(...).reduceByKey(...)`.



Shwetank Singh
GritSetGrow - GSGLearn.com



HDFS

Integration with Hadoop Distributed File System for data storage.

Example

Reading data from HDFS:

```
val rdd = sc.textFile("hdfs://path/to/file").
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Data Sources

Interfaces for reading and writing data from various sources like HDFS, S3, JDBC, and more.

Example

Reading data from a JDBC source:

```
val df = spark.read.jdbc(url, table, properties).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



SQL Functions

Built-in functions for data manipulation and analysis in Spark SQL.

Example

Using concat function:

```
df.select(concat(col("name"), lit(" "),  
col("surname")).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



User-Defined Functions (UDFs)

Custom functions defined by users to extend
Spark SQL capabilities.

Example

```
spark.udf.register("add_one", lambda x: x + 1,  
IntegerType()).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Window Functions

Functions that perform calculations across a set of rows related to the current row.

Example

```
df.withColumn("rank", row_number()  
.over(Window.partitionBy("dept")  
.orderBy("salary"))).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



GroupBy

Grouping data by one or more columns for aggregation.

Example

```
df.groupBy("department").agg(avg("salary")).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Join

Combining rows from two or more
DataFrames based on a related column.

Example

```
df1.join(df2, df1("id") === df2("id")).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Sorting

Ordering rows in a DataFrame based on column values.

Example

`df.sort("age", ascending=True).`



Shwetank Singh
GritSetGrow - GSGLearn.com



Aggregations

Computing summary statistics of data groups.

Example

```
df.groupBy("department").agg(sum("salary"))
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Pivot

Transforming unique values from one column
into multiple columns.

Example

```
df.groupBy("year").pivot("quarter").sum("revenue").
```



Shwetank Singh
GritSetGrow - GSGLearn.com



DataFrame Operations

Common operations on DataFrames, such as selecting, filtering, and grouping.

Example

```
df.select("name").filter(df("age") > 21).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Schema

Defining the structure of a DataFrame with column names and types.

Example

```
schema = StructType([StructField("name",  
StringType(), True), StructField("age",  
IntegerType(), True)]).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



DataFrame Creation

Creating DataFrames from various data sources like CSV, JSON, and RDDs.

Example

```
df = spark.read.csv("file.csv").
```



Shwetank Singh
GritSetGrow - GSGLearn.com



DataFrame Transformation

Modifying DataFrames using operations like
withColumn, filter, groupBy, and agg.

Example

```
df.withColumn("new_col", df("col1") + df("col2"))
```



Shwetank Singh
GritSetGrow - GSGLearn.com



DataFrame Actions

Triggering computations and returning results to the driver.

Example
`df.show()`.



Shwetank Singh
GritSetGrow - GSGLearn.com



SQL Queries

Executing SQL queries on DataFrames.

Example

```
df.createOrReplaceTempView("table");  
spark.sql("SELECT * FROM table").
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Joins

Combining DataFrames based on common columns.

Example

```
df1.join(df2, df1("id") === df2("id")).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Aggregations

Aggregating data in DataFrames using functions like sum, avg, and count.

Example

```
df.groupBy("department").agg(sum("salary"))
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Filtering

Filtering rows based on column values.

Example

`df.filter(df("age") > 21).`



Shwetank Singh
GritSetGrow - GSGLearn.com



Selection

Selecting specific columns from a DataFrame.

Example

```
df.select("name", "age").
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Renaming

Renaming columns in a DataFrame.

Example

```
df.withColumnRenamed("old_name",  
"new_name")
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Dropping Columns

Removing columns from a DataFrame.

Example

```
df.drop("column_name").
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Merging

Combining multiple DataFrames into one.

Example

`df1.union(df2).`



Shwetank Singh
GritSetGrow - GSGLearn.com



Pivoting

Transforming data to show unique values from one column as multiple columns.

Example

```
df.groupBy("year")  
.pivot("quarter")  
.sum("revenue").
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Unpivoting

Transforming data from columns into rows.

Example

```
melted_df = df.selectExpr("stack(3, 'Q1', Q1,  
'Q2', Q2, 'Q3', Q3) as (quarter, revenue)").
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Writing

Writing DataFrames to external storage systems like HDFS, S3, or databases.

Example

```
df.write.csv("output.csv").
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Reading

Reading data from various external storage systems into DataFrames.

Example

```
df = spark.read.json("input.json").
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Repartitioning

Changing the number of partitions of a DataFrame.

Example

`df.repartition(10).`



Shwetank Singh
GritSetGrow - GSGLearn.com



Coalescing

Reducing the number of partitions of a DataFrame.

Example

`df.coalesce(1).`



Shwetank Singh
GritSetGrow - GSGLearn.com



Serialization

Converting DataFrame rows into a format that can be stored and transmitted.

Example

Using PySpark's built-in serializers like pickle and Kryo.



Shwetank Singh
GritSetGrow - GSGLearn.com

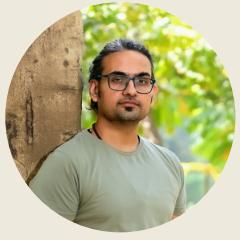


Deserialization

Converting serialized data back into
DataFrame rows.

Example

Reading serialized data and converting it back
to DataFrame.



Shwetank Singh
GritSetGrow - GSGLearn.com



Functions

Built-in functions for performing operations on DataFrames.

Example

from pyspark.sql.functions

```
import col, lit; df.select(col("name"), lit(1)).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Column Operations

Performing operations on DataFrame columns.

Example

```
df.withColumn("new_col", df("col1") + df("col2"))
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Exploding

Transforming a DataFrame with nested structures into a flat DataFrame.

Example

```
df.select("name", explode("hobbies"))
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Aggregation Functions

Built-in functions for aggregating data in
DataFrames.

Example

```
from pyspark.sql.functions import sum, avg;  
df.groupBy("dept")  
.agg(sum("salary"), avg("salary"))
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Statistics

Computing summary statistics of DataFrames.

Example

`df.describe().show()`.



Shwetank Singh
GritSetGrow - GSGLearn.com



Union

Combining rows from two DataFrames into one.

Example

`df1.union(df2).`



Shwetank Singh
GritSetGrow - GSGLearn.com



Intersect

Finding common rows between two
DataFrames.

Example

`df1.intersect(df2).`



Shwetank Singh
GritSetGrow - GSGLearn.com



Except

Finding rows in one DataFrame that are not in another.

Example

`df1.except(df2).`



Shwetank Singh
GritSetGrow - GSGLearn.com



Distinct

Removing duplicate rows from a DataFrame.

Example

`df.distinct()`.



Shwetank Singh
GritSetGrow - GSGLearn.com



Drop Duplicates

Removing duplicate rows based on specific columns.

Example

```
df.dropDuplicates(["col1", "col2"]).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Missing Data Handling

Handling missing values in DataFrames.

Example

`df.fillna(0).`



Shwetank Singh
GritSetGrow - GSGLearn.com



Date Functions

Working with date and time data in
DataFrames.

Example

```
df.withColumn("year", year("date_col"))
```



Shwetank Singh
GritSetGrow - GSGLearn.com



String Functions

Performing operations on string columns in
DataFrames.

Example

```
df.withColumn("uppercase_name",  
upper("name"))
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Conversion

Converting between different data types in
DataFrames.

Example

```
df.withColumn("int_col", col("string_col")  
.cast("int")).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Explode

Converting an array column into multiple rows, one for each element in the array.

Example

```
df.select("name", explode("hobbies")).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



JSON Functions

Working with JSON data in DataFrames.

Example

```
df.selectExpr("json_tuple(json_col, 'key1',  
'key2')").
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Advanced Analytics

Using advanced analytical functions available
in PySpark.

Example

```
df.stat.freqItems(["col1", "col2"]).
```



Shwetank Singh
GritSetGrow - GSGLearn.com



Vectorized UDFs

UDFs that operate on entire columns at once, improving performance.

Example

```
from pyspark.sql.functions import pandas_udf;  
@pandas_udf("int") def add_one(s: pd.Series) -> pd.Series: return s + 1.
```



Shwetank Singh
GritSetGrow - GSGLearn.com



THANK YOU
A P A C H E
S p a r k ™

