

# Spark by {Examples} (https://sparkbyexamples.com/)

## Spark Tutorial

[Spark – Installation on Windows \(https://sparkbyexamples.com/spark/apache-spark-installation-on-windows/\)](https://sparkbyexamples.com/spark/apache-spark-installation-on-windows/)

[Spark – Installation on Linux | Ubuntu \(https://sparkbyexamples.com/spark/spark-installation-on-linux-ubuntu/\)](https://sparkbyexamples.com/spark/spark-installation-on-linux-ubuntu/)

[Spark – Cluster Setup with Hadoop Yarn \(https://sparkbyexamples.com/spark/spark-setup-on-hadoop-yarn/\)](https://sparkbyexamples.com/spark/spark-setup-on-hadoop-yarn/)

[Spark – Web/Application UI \(https://sparkbyexamples.com/spark/spark-web-ui-understanding/\)](https://sparkbyexamples.com/spark/spark-web-ui-understanding/)

[Spark – Setup with Scala and IntelliJ \(https://sparkbyexamples.com/spark/spark-setup-run-with-scala-intellij/\)](https://sparkbyexamples.com/spark/spark-setup-run-with-scala-intellij/)

[Spark – How to Run Examples From this Site on IntelliJ IDEA \(https://sparkbyexamples.com/spark/how-to-run-spark-examples-from-intellij/\)](https://sparkbyexamples.com/spark/how-to-run-spark-examples-from-intellij/)

[Spark – SparkSession \(https://sparkbyexamples.com/spark/sparksession-explained-with-examples/\)](https://sparkbyexamples.com/spark/sparksession-explained-with-examples/)

[Spark – SparkContext \(https://sparkbyexamples.com/spark/spark-sparkcontext/\)](https://sparkbyexamples.com/spark/spark-sparkcontext/)

## Spark RDD Tutorial

[Spark RDD – Parallelize \(https://sparkbyexamples.com/apache-spark-rdd/how-to-create-an-rdd-using-parallelize/\)](https://sparkbyexamples.com/apache-spark-rdd/how-to-create-an-rdd-using-parallelize/)

[Spark RDD – Read text file \(https://sparkbyexamples.com/apache-spark-rdd/spark-read-multiple-text-files-into-a-single-rdd/\)](https://sparkbyexamples.com/apache-spark-rdd/spark-read-multiple-text-files-into-a-single-rdd/)

[Spark RDD – Read CSV \(https://sparkbyexamples.com/apache-spark-rdd/spark-load-csv-file-into-rdd/\)](https://sparkbyexamples.com/apache-spark-rdd/spark-load-csv-file-into-rdd/)

[Spark RDD – Create RDD \(https://sparkbyexamples.com/apache-spark-rdd/different-ways-to-create-spark-rdd/\)](https://sparkbyexamples.com/apache-spark-rdd/different-ways-to-create-spark-rdd/)

[PySpark \(https://sparkbyexamples.com/pyspark/\)](https://sparkbyexamples.com/pyspark/)

[Hive \(https://sparkbyexamples.com/apache-hive/\)](https://sparkbyexamples.com/apache-hive/)

[HBase \(https://sparkbyexamples.com/apache-hbase/\)](https://sparkbyexamples.com/apache-hbase/)

[Kafka \(https://sparkbyexamples.com/apache-kafka/\)](https://sparkbyexamples.com/apache-kafka/) [Sriram](https://sparkbyexamples.com/author/srirammimalapudi@gmail-com/)

[tutorials with examples \(https://sparkbyexamples.com/category/spark/\)](https://sparkbyexamples.com/category/spark/) [PySpark](https://sparkbyexamples.com/category/pyspark/) [FAQ's \(https://sparkbyexamples.com/category/sparksql/\)](https://sparkbyexamples.com/category/sparksql/)

[More \(https://sparkbyexamples.com/\)](https://sparkbyexamples.com/)

Apache Spark provides a suite of Web UI/User Interfaces ([Jobs](#), [Stages](#), [Tasks](#), [Storage](#), [Environment](#), [Executors](#), and [SQL](#)) to monitor the status of your Spark/PySpark application, resource consumption of Spark cluster, and Spark configurations.

To better understand how Spark executes the Spark/PySpark Jobs, these set of user interfaces comes in handy. In this article, I will run a small application and explain how Spark executes this by using different sections in Spark Web UI.

Before going into Spark UI first, learn about these two concepts.

- [Transformations \(https://sparkbyexamples.com/apache-spark-rdd/spark-rdd-transformations/\)](https://sparkbyexamples.com/apache-spark-rdd/spark-rdd-transformations/)
- [Action \(https://sparkbyexamples.com/apache-spark-rdd/spark-rdd-actions/\)](https://sparkbyexamples.com/apache-spark-rdd/spark-rdd-actions/)

Let me give a small brief on those two, Your application code is the set of instructions that instructs the driver to do a Spark Job

## Personalized Tissue Box

Custom Tissue Holder  
With Name & Charm in  
Wine, Tan, Black &  
Many More Variants.  
Shop Now!

The Messy Corner

## Spark Web UI – Understanding Spark Execution

[Spark RDD – Create Empty RDD \(https://sparkbyexamples.com/apache-spark-rdd/spark-how-to-create-an-empty-rdd/\)](https://sparkbyexamples.com/apache-spark-rdd/spark-how-to-create-an-empty-rdd/)

[Spark RDD – Transformations \(https://sparkbyexamples.com/apache-spark-rdd/spark-rdd-transformations/\)](https://sparkbyexamples.com/apache-spark-rdd/spark-rdd-transformations/)

[Spark RDD – Actions \(https://sparkbyexamples.com/apache-spark-rdd/spark-rdd-actions/\)](https://sparkbyexamples.com/apache-spark-rdd/spark-rdd-actions/)

[Spark RDD – Pair Functions \(https://sparkbyexamples.com/apache-spark-rdd/spark-pair-rdd-functions/\)](https://sparkbyexamples.com/apache-spark-rdd/spark-pair-rdd-functions/)

[Spark RDD – Repartition and Coalesce \(https://sparkbyexamples.com/spark/spark-repartition-vs-coalesce/\)](https://sparkbyexamples.com/spark/spark-repartition-vs-coalesce/)

[Spark RDD – Shuffle Partitions \(https://sparkbyexamples.com/spark/spark-shuffle-partitions/\)](https://sparkbyexamples.com/spark/spark-shuffle-partitions/)

[Spark RDD – Cache vs Persist \(https://sparkbyexamples.com/spark/spark-difference-between-cache-and-persist/\)](https://sparkbyexamples.com/spark/spark-difference-between-cache-and-persist/)

[Spark RDD – Persistence Storage Levels \(https://sparkbyexamples.com/spark/spark-persistence-storage-levels/\)](https://sparkbyexamples.com/spark/spark-persistence-storage-levels/)

[Spark RDD – Broadcast Variables \(https://sparkbyexamples.com/spark/spark-broadcast-variables/\)](https://sparkbyexamples.com/spark/spark-broadcast-variables/)

[Spark RDD – Accumulator Variables \(https://sparkbyexamples.com/spark/spark-accumulators/\)](https://sparkbyexamples.com/spark/spark-accumulators/)

[Spark RDD – Convert RDD to DataFrame \(https://sparkbyexamples.com/apache-spark-rdd/convert-spark-rdd-to-dataframe-dataset/\)](https://sparkbyexamples.com/apache-spark-rdd/convert-spark-rdd-to-dataframe-dataset/)

## Spark SQL Tutorial

[Spark SQL – Create DataFrame \(https://sparkbyexamples.com/spark/different-ways-to-create-a-spark-dataframe/\)](https://sparkbyexamples.com/spark/different-ways-to-create-a-spark-dataframe/)

[Spark SQL – Select Columns \(https://sparkbyexamples.com/spark/spark-select-columns-from-dataframe/\)](https://sparkbyexamples.com/spark/spark-select-columns-from-dataframe/)

[Spark SQL – Add and Update Column \(withColumn\)](#)

and let the driver decide how to achieve it with the help of executors.

Instructions to the driver are called Transformations and action will trigger the execution.

I had written a small application which does transformation and action.

```
//Transformation
val rawDF = spark.read //job(0) : Read
    .option("inferSchema", "true") //job(1) : InferSchema
    .option("header", "true")
    .csv( path = "data/survey.csv")

//Action
rawDF.count()// job(2): Get Count
```

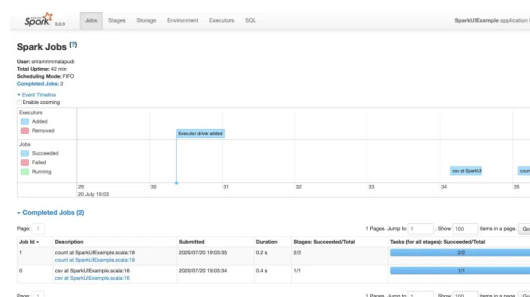
Application Code

Here we are creating a [DataFrame \(https://sparkbyexamples.com/spark/different-ways-to-create-a-spark-dataframe/\)](https://sparkbyexamples.com/spark/different-ways-to-create-a-spark-dataframe/) by reading a .csv file (<https://sparkbyexamples.com/apache-spark-rdd/spark-load-csv-file-into-rdd/>) and checking the count of the DataFrame. Let's understand how an application gets projected in Spark UI

Spark UI is separated into below tabs.

1. [Spark Jobs](#)
2. [Stages](#)
3. [Tasks](#)
4. [Storage](#)
5. [Environment](#)
6. [Executors](#)
7. [SQL](#)

If you are running the Spark application locally, Spark UI can be accessed using the <http://localhost:4040/> (<http://localhost:4040/>) . Spark UI by default runs on port 4040 and below are some of the additional UI's that would be helpful to track Spark application.



Spark Web UI

- Spark Application UI: <http://localhost:4040/> (<http://localhost:4040/>)

## Advanced Angular UI Components

Data Grid, Chart, Scheduler

No Dependencies, Easy to use, Responsive, Mobile ready

[htmlelements.com](http://htmlelements.com)

OPEN

## Advanced Angular UI Components

Data Grid, Chart, Scheduler

No Dependencies, Easy to use, Responsive, Mobile ready

[htmlelements.com](http://htmlelements.com)

OPEN

[\(https://sparkbyexamples.com/spark/spark-dataframe-withcolumn/\)](https://sparkbyexamples.com/spark/spark-dataframe-withcolumn/)

[Spark SQL – Rename Nested Column](#)

[\(https://sparkbyexamples.com/spark/rename-a-column-on-spark-dataframes/\)](https://sparkbyexamples.com/spark/rename-a-column-on-spark-dataframes/)

[Spark SQL – Drop column](#)

[\(https://sparkbyexamples.com/spark/spark-drop-column-from-dataframe-dataset/\)](https://sparkbyexamples.com/spark/spark-drop-column-from-dataframe-dataset/)

[Spark SQL – Where | Filter](#)

[\(https://sparkbyexamples.com/spark/spark-dataframe-where-filter/\)](https://sparkbyexamples.com/spark/spark-dataframe-where-filter/)

[Spark SQL – When Otherwise](#)

[\(https://sparkbyexamples.com/spark/spark-case-when-otherwise-example/\)](https://sparkbyexamples.com/spark/spark-case-when-otherwise-example/)

[Spark SQL – Collect data to Driver](#)

[\(https://sparkbyexamples.com/spark/spark-dataframe-collect/\)](https://sparkbyexamples.com/spark/spark-dataframe-collect/)

[Spark SQL – Distinct](#)

[\(https://sparkbyexamples.com/spark/spark-remove-duplicate-rows/\)](https://sparkbyexamples.com/spark/spark-remove-duplicate-rows/)

[Spark SQL- Pivot Table DataFrame](#)

[\(https://sparkbyexamples.com/spark/how-to-pivot-table-and-unpivot-a-spark-dataframe/\)](https://sparkbyexamples.com/spark/how-to-pivot-table-and-unpivot-a-spark-dataframe/)

[Spark SQL – Data Types](#)

[\(https://sparkbyexamples.com/spark/spark-sql-dataframe-data-types/\)](https://sparkbyexamples.com/spark/spark-sql-dataframe-data-types/)

[Spark SQL – StructType | StructField](#)

[\(https://sparkbyexamples.com/spark/spark-sql-structtype-on-dataframe/\)](https://sparkbyexamples.com/spark/spark-sql-structtype-on-dataframe/)

[Spark SQL – Schema](#)

[\(https://sparkbyexamples.com/spark/spark-schema-explained-with-examples/\)](https://sparkbyexamples.com/spark/spark-schema-explained-with-examples/)

[Spark SQL – Groupby](#)

[\(https://sparkbyexamples.com/spark/using-groupby-on-dataframe/\)](https://sparkbyexamples.com/spark/using-groupby-on-dataframe/)

[Spark SQL – Sort DataFrame](#)

[\(https://sparkbyexamples.com/spark/spark-how-to-sort-dataframe-column-explained/\)](https://sparkbyexamples.com/spark/spark-how-to-sort-dataframe-column-explained/)

[Spark SQL – Join Types](#)

[\(https://sparkbyexamples.com/spark/spark-sql-dataframe-join/\)](https://sparkbyexamples.com/spark/spark-sql-dataframe-join/)

[Spark SQL – Union and UnionAll](#)

[\(https://sparkbyexamples.com/spark/spark-dataframe-union-and-union-all/\)](https://sparkbyexamples.com/spark/spark-dataframe-union-and-union-all/)

- Resource Manager: <http://localhost:9870>  
(<http://localhost:9870/>)
- Spark JobTracker: <http://localhost:8088/>  
(<http://localhost:8088/>)
- Node Specific Info: <http://localhost:8042/>  
(<http://localhost:8042/>)

**Note:** To access these URLs, Spark application should in running state. If you wanted to access this URL regardless of your Spark application status and wanted to access Spark UI all the time, you would need to start [Spark History server](#) (<https://sparkbyexamples.com/hadoop/spark-setup-on-yarn/#spark-history-server>).

## 1. Spark Jobs Tab

### Spark Jobs (?)

**User:** sriramrimalapudi

**Total Uptime:** 2.5 h

**Scheduling Mode:** FIFO

**Completed Jobs:** 3

Jobs tab

The details that I want you to be aware of under the jobs section are [Scheduling mode](#), the [number of Spark Jobs](#), the [number of stages](#) it has, and [Description](#) in your spark job.

### 1.1 Scheduling Mode

We have three Scheduling modes.

1. **Standalone** mode
2. **YARN** mode
3. **Mesos**

Completed Jobs (3)					
Job ID	Description	Submitted	Duration	Progress: Successful/Total	Tasks (for all stages) Successful/Total
1	count of SparkDataFrame scala 20	2020/07/02 21:41:26	0.2 s	1/1	100/100
2	count of SparkDataFrame scala 22	2020/07/02 21:41:26	0.3 s	1/1	100/100
3	count of SparkDataFrame scala 18	2020/07/02 21:41:26	0.3 s	1/1	100/100

Spark Scheduling tab

As I was running in a local machine, I tried using Standalone mode

### 1.2 Number of Spark Jobs:

Always keep in mind, the number of Spark jobs is equal to the number of actions in the application and each Spark job should have

[Spark SQL – map\(\) vs mapPartitions\(\)](#)  
(<https://sparkbyexamples.com/spark/spark-map-vs-mappartitions-transformation/>).

[Spark SQL – foreach\(\) vs foreachPartition\(\)](#)  
(<https://sparkbyexamples.com/spark/spark-foreachpartition-vs-foreach-explained/>).

[Spark SQL – map\(\) vs flatMap\(\)](#)  
(<https://sparkbyexamples.com/spark/spark-map-vs-flatmap-with-examples/>).

[Spark SQL – Persist and Cache](#)  
(<https://sparkbyexamples.com/spark/spark-dataframe-cache-and-persist-explained/>).

[Spark SQL – UDF \(User Defined Functions\)](#)  
(<https://sparkbyexamples.com/spark/spark-sql-udf/>).

[Spark SQL – Array \(ArrayType\) Column](#)  
(<https://sparkbyexamples.com/spark/spark-array-arraytype-dataframe-column/>).

[Spark SQL – Map \(MapType\) column](#)  
(<https://sparkbyexamples.com/spark/spark-dataframe-map-maptype-column/>).

[Spark SQL – Flatten Nested Struct Column](#)  
(<https://sparkbyexamples.com/spark/spark-flatten-nested-struct-column/>).

[Spark SQL – Flatten Nested Array Column](#)  
(<https://sparkbyexamples.com/spark/spark-flatten-nested-array-column-to-single-column/>).

[Spark SQL – Explode Array & Map Columns](#)  
(<https://sparkbyexamples.com/spark/explode-spark-array-and-map-dataframe-column/>).

[Spark SQL – Sampling](#)  
(<https://sparkbyexamples.com/spark/spark-sampling-with-examples/>).

[Spark SQL – Partitioning](#)  
(<https://sparkbyexamples.com/spark/spark-partitioning-understanding/>).

## Spark SQL Functions

[Spark SQL String Functions](#)  
(<https://sparkbyexamples.com/spark/spark-sql-string-functions/>).

at least one Stage.

In our above application, we have performed 3 Spark jobs (0,1,2)

- Job 0. *read the CSV file.*
- Job 1. *Inferschema from the file.*
- Job 2. *Count Check*

So if we look at the fig it clearly shows 3 Spark jobs result of 3 actions.

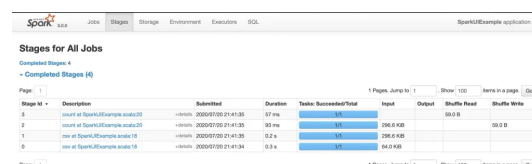
## 1.3 Number of Stages

Each [Wide Transformation](#) (<https://sparkbyexamples.com/apache-spark-rdd/spark-rdd-transformations/#wider-transformation>) results in a separate Number of Stages. In our case, Spark job0 and Spark job1 have individual single stages but when it comes to Spark job 3 we can see two stages that are because of the partition of data. Data is partitioned into two files by default.

## 1.4 Description

Description links the complete details of the associated SparkJob like Spark Job Status, DAG Visualization, Completed Stages I had explained the description part in the coming part.

## 2. Stages Tab



Stage ID	Description	Submitted	Duration	Tasks: Successful/Total	Input	Output	Shuffle Read	Shuffle Write
2	count of SparkRDDExample.scala:25	<= 2020-07-02 21:41:35	57 ms	1/1	0 B	0 B	0 B	0 B
2	count of SparkRDDExample.scala:25	<= 2020-07-02 21:41:35	50 ms	1/1	256.0 K B	0 B	0 B	0 B
1	case of SparkRDDExample.scala:13	<= 2020-07-02 21:41:35	0.2 s	1/1	0 B	256.0 K B	0 B	0 B
0	case of SparkRDDExample.scala:13	<= 2020-07-02 21:41:34	0.3 s	1/1	0 B	64.0 K B	0 B	0 B

Spark Stage Tab

We can navigate into Stage Tab in two ways.

1. Select the Description of the respective Spark job (Shows stages only for the Spark job opted)
2. On the top of Spark Job tab select Stages option (Shows all stages in Application)

In our application, we have a total of **4 Stages**.

The Stage tab displays a summary page that shows the current state of all stages of all Spark jobs in the spark application

[ark/usage-of-spark-sql-string-functions/](#))

[Spark SQL Date and Timestamp Functions](#)  
(<https://sparkbyexamples.com/spark/spark-sql-date-and-time-functions/>)

[Spark SQL Array Functions](#)  
(<https://sparkbyexamples.com/spark/spark-sql-array-functions/>)

[Spark SQL Map Functions](#)  
(<https://sparkbyexamples.com/spark/spark-sql-map-functions/>)

[Spark SQL Sort Functions](#)  
(<https://sparkbyexamples.com/spark/spark-sql-sort-functions/>)

[Spark SQL Aggregate Functions](#)  
(<https://sparkbyexamples.com/spark/spark-sql-aggregate-functions/>)

[Spark SQL Window Functions](#)  
(<https://sparkbyexamples.com/spark/spark-sql-window-functions/>)

[Spark SQL JSON Functions](#)  
(<https://sparkbyexamples.com/spark/spark-most-used-json-functions-with-examples/>)

## Spark Data Source API

[Spark – Read & Write CSV file](#)  
(<https://sparkbyexamples.com/spark/spark-read-csv-file-into-dataframe/>)

[Spark – Read and Write JSON file](#)  
(<https://sparkbyexamples.com/spark/spark-read-and-write-json-file/>)

[Spark – Read & Write Parquet file](#)  
(<https://sparkbyexamples.com/spark/spark-read-write-dataframe-parquet-example/>)

[Spark – Read & Write XML file](#)  
(<https://sparkbyexamples.com/spark/spark-read-write-xml/>)

[Spark – Read & Write Avro files](#)  
(<https://sparkbyexamples.com/spark/read-write-avro-file-spark-dataframe/>)

[Spark – Read & Write Avro files](#)  
([Spark version 2.3.x or earlier](#))  
(<https://sparkbyexamples.com/spark/using-avro-data-files-from-spark-sql-2-3-x/>)

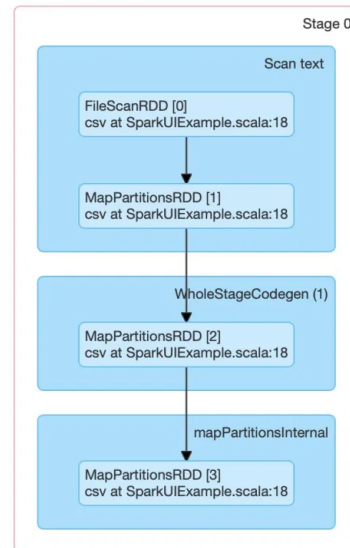
[Spark – Read & Write HBase using “hbase-spark” Connector](#)  
(<https://sparkbyexamples.com/spark/spark-sql-hbase/>)

The number of tasks you could see in each stage is the number of partitions that spark is going to work on and each task inside a stage is the same work that will be done by spark but on a different partition of data.

## Details for Stage 0 (Attempt 0)

**Total Time Across All Tasks:** 94 ms  
**Locality Level Summary:** Process local: 1  
**Input Size / Records:** 64.0 KiB / 1  
**Associated Job Ids:** 0

▼ DAG Visualization



Stage 0

## Stage detail

Details of stage showcase Directed Acyclic Graph (DAG) of this stage, where vertices represent the RDDs or DataFrame and edges represent an operation to be applied.

let us analyze operations in Stages  
Operations in Stage0 are

- 1.FileScanRDD
- 2.MapPartitionsRDD

## FileScanRDD

FileScan represents reading the data from a file.

It is given FilePartitions that are custom RDD partitions with PartitionedFiles (file blocks)

In our scenario, the *CSV file is read*

## MapPartitionsRDD

MapPartitionsRDD will be created when you use map Partition transformation

[ark/spark-read-write-using-hbase-spark-connector/](#)

[Spark – Read & Write from HBase using Hortonworks \(https://sparkbyexamples.com/spark/create-spark-dataframe-from-hbase-using-hortonworks/\)](#)

[Spark – Read & Write ORC file \(https://sparkbyexamples.com/spark/spark-read-orc-file-into-dataframe/\)](#)

[Spark – Read Binary File \(https://sparkbyexamples.com/spark/spark-read-binary-file-into-dataframe/\)](#)

## Spark Streaming & Kafka

[Spark Streaming – OutputModes \(https://sparkbyexamples.com/spark/spark-streaming-outputmode/\)](#)

[Spark Streaming – Reading Files From Directory \(https://sparkbyexamples.com/spark/spark-streaming-read-json-files-from-directory/\)](#)

[Spark Streaming – Reading Data From TCP Socket \(https://sparkbyexamples.com/spark/spark-streaming-from-tcp-socket/\)](#)

[Spark Streaming – Processing Kafka Messages in JSON Format \(https://sparkbyexamples.com/spark/spark-streaming-with-kafka/\)](#)

[Spark Streaming – Processing Kafka messages in AVRO Format \(https://sparkbyexamples.com/spark/spark-streaming-consume-and-produce-kafka-messages-in-avro-format/\)](#)

[Spark SQL Batch – Consume & Produce Kafka Message \(https://sparkbyexamples.com/spark/spark-batch-processing-produce-consume-kafka-topic/\)](#)

## PySpark Tutorial

[PySpark Tutorial For Beginners \(https://sparkbyexamples.com/py-spark-tutorial/\)](#)

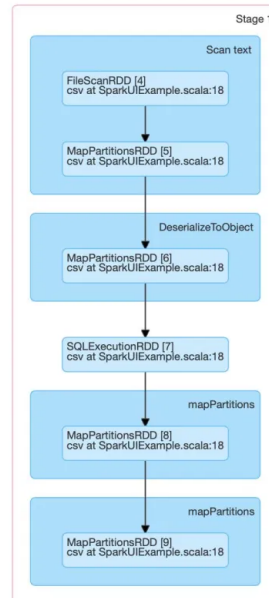
[PySpark – Features \(https://sparkbyexamples.com/py-spark-tutorial/#features\)](#)

[PySpark – Advantages \(https://sparkbyexamples.com/py-spark-tutorial/#advantages\)](#)

### Details for Stage 1 (Attempt 0)

Total Time Across All Tasks: 0.1 s  
Locality Level Summary: Process local: 1  
Input Size / Records: 296.6 KiB / 1260  
Associated Job Ids: 1

↪ DAG Visualization



Stage1

Operation in Stage(1) are

- 1.FileScanRDD
- 2.MapPartitionsRDD
- 3.SQLExecutionRDD

As File Scan and MapPartitionsRDD is already explained, let us look at SQLExecutionRDD

## SQLExecutionRDD

SQLExecutionRDD is Spark property that is used to track multiple Spark jobs that should all together constitute a single structured query execution.



[PySpark – Modules & Packages](https://sparkbyexamples.com/py-spark-tutorial/#modules-packages)  
(<https://sparkbyexamples.com/py-spark-tutorial/#modules-packages>)

[PySpark – Cluster Managers](https://sparkbyexamples.com/py-spark-tutorial/#cluster-manager)  
(<https://sparkbyexamples.com/py-spark-tutorial/#cluster-manager>)

[PySpark – Install on Windows](https://sparkbyexamples.com/py-spark-tutorial/#pyspark-installation)  
(<https://sparkbyexamples.com/py-spark-tutorial/#pyspark-installation>)

[PySpark – Web/Application UI](https://sparkbyexamples.com/spark/spark-web-ui-understanding/)  
(<https://sparkbyexamples.com/spark/spark-web-ui-understanding/>)

[PySpark – SparkSession](https://sparkbyexamples.com/py-spark/pyspark-what-is-sparksession/)  
(<https://sparkbyexamples.com/py-spark/pyspark-what-is-sparksession/>)

[PySpark – RDD](https://sparkbyexamples.com/py-spark-rdd)  
(<https://sparkbyexamples.com/py-spark-rdd>)

[PySpark – Parallelize](https://sparkbyexamples.com/py-spark/pyspark-parallelize-create-rdd/)  
(<https://sparkbyexamples.com/py-spark/pyspark-parallelize-create-rdd/>)

[PySpark – repartition\(\) vs coalesce\(\)](https://sparkbyexamples.com/py-spark/pyspark-repartition-vs-coalesce/)  
(<https://sparkbyexamples.com/py-spark/pyspark-repartition-vs-coalesce/>)

[PySpark – Broadcast Variables](https://sparkbyexamples.com/py-spark/pyspark-broadcast-variables/)  
(<https://sparkbyexamples.com/py-spark/pyspark-broadcast-variables/>)

[PySpark – Accumulator](https://sparkbyexamples.com/py-spark/pyspark-accumulator-with-example/)  
(<https://sparkbyexamples.com/py-spark/pyspark-accumulator-with-example/>)

## PySpark DataFrame

[PySpark – Create a DataFrame](https://sparkbyexamples.com/py-spark/different-ways-to-create-dataframe-in-pyspark/)  
(<https://sparkbyexamples.com/py-spark/different-ways-to-create-dataframe-in-pyspark/>)

[PySpark – Create an empty DataFrame](https://sparkbyexamples.com/py-spark/pyspark-create-an-empty-dataframe/)  
(<https://sparkbyexamples.com/py-spark/pyspark-create-an-empty-dataframe/>)

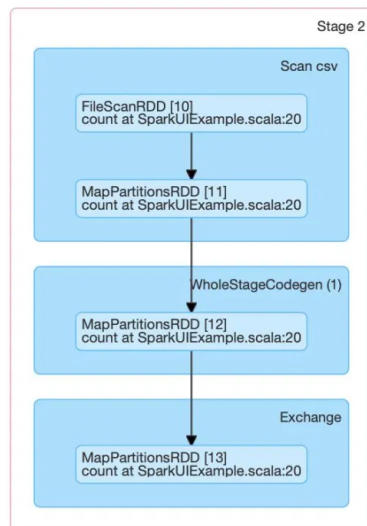
[PySpark – Convert RDD to DataFrame](https://sparkbyexamples.com/py-spark/convert-pyspark-rdd-to-dataframe/)  
(<https://sparkbyexamples.com/py-spark/convert-pyspark-rdd-to-dataframe/>)

[PySpark – Convert DataFrame to Pandas](https://sparkbyexamples.com/py-spark/convert-dataframe-to-pandas)  
(<https://sparkbyexamples.com/py-spark/convert-dataframe-to-pandas>)

## Details for Stage 2 (Attempt 0)

Total Time Across All Tasks: 75 ms  
Locality Level Summary: Process local: 1  
Input Size / Records: 296.6 KiB / 1259  
Shuffle Write Size / Records: 59.0 B / 1  
Associated Job Ids: 2

▼ DAG Visualization



Stage 2

Operation in Stage(2) and Stage(3) are

- 1.FileScanRDD
- 2.MapPartitionsRDD
- 3.WholeStageCodegen
- 4.Exchange

## Wholestagecodegen

A physical query optimizer in Spark SQL that fuses multiple physical operators

## Exchange

Exchange is performed because of the COUNT method.

*As data is divided into partitions and shared among executors, to get count there should be adding of the count of from individual partition.*

Represents the shuffle i.e data movement across the cluster(Executors).

It is the most expensive operation and if number of partitions is more exchange of data between executors will also be more.

## 3. Tasks

Tasks (1)										
Index	Task ID	Attempt	Status	Locality level	Executor ID	Host	Logs	Launched Time	Duration	GC Time
0	0	0	SUCCESS	NODE_LOCAL	driver	192.168.0.101		2025-07-03 15:11:08	463.0ms	0.0s / 1

Spark Tasks Tab

[spark/convert-pyspark-dataframe-to-pandas/](#)

[PySpark – show\(\)](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-show-display-dataframe-contents-in-table/>)

[PySpark – StructType & StructField](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-structtype-and-structfield/>)

[PySpark – Row Class](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-row-using-rdd-dataframe/>)

[PySpark – Column Class](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-column-functions/>)

[PySpark – select\(\)](#)  
(<https://sparkbyexamples.com/pyspark/select-columns-from-pyspark-dataframe/>)

[PySpark – collect\(\)](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-collect/>)

[PySpark – withColumn\(\)](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-withcolumn/>)

[PySpark – withColumnRenamed\(\)](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-rename-dataframe-column/>)

[PySpark – where\(\) & filter\(\)](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-where-filter/>)

[PySpark – drop\(\) & dropDuplicates\(\)](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-distinct-to-drop-duplicates/>)

[PySpark – orderBy\(\) and sort\(\)](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-orderby-and-sort-explained/>)

[PySpark – groupBy\(\)](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-groupby-explained-with-example/>)

[PySpark – join\(\)](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-join-explained-with-examples/>)

[PySpark – union\(\) & unionAll\(\)](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-union-and-unionall/>)

[PySpark – unionByName\(\)](#)  
(<https://sparkbyexamples.com/spark/union-by-name/>)

Tasks are located at the bottom space in the respective stage.

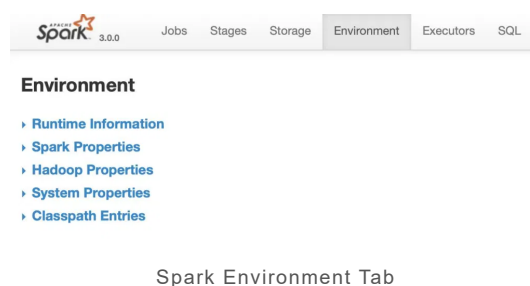
Key things to look task page are:

1. Input Size – Input for the Stage
2. Shuffle Write-Output is the stage written.

## 4. Storage

The Storage tab displays the persisted RDDs and DataFrames, if any, in the application. The summary page shows the storage levels, sizes and partitions of all RDDs, and the details page shows the sizes and using executors for all partitions in an RDD or DataFrame.

## 5. Environment Tab



This environment page has five parts. It is a useful place to check whether your properties have been set correctly.

1. **Runtime Information:** simply contains the runtime properties like versions of Java and Scala.
2. **Spark Properties:** lists the application properties like 'spark.app.name' and 'spark.driver.memory'.
3. **Hadoop Properties:** displays properties relative to Hadoop and YARN. **Note:** Properties like '<https://spark.apache.org/docs/3.0.0-preview/configuration.html#execution-behavior>' spark.hadoop' are shown not in this part but in 'Spark Properties'.
4. **System Properties:** shows more details about the JVM.
5. **Classpath Entries:** lists the classes loaded from different sources, which is very useful to resolve class conflicts.





[ark/pyspark-sql-date-and-timestamp-functions/](#)

[PySpark – JSON Functions](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/>)

## PySpark Datasources

[PySpark – Read & Write CSV File](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-read-csv-file-into-dataframe/>)

[PySpark – Read & Write Parquet File](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-read-and-write-parquet-file/>)


[PySpark – Read & Write JSON file](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-read-json-file-into-dataframe/>)


In our application, we performed read and count operation on files and DataFrame. So both read and count are listed SQL Tab

Some of the resources are gathered from <https://spark.apache.org/> (<https://spark.apache.org/>) thanks for the information.

“.....Keep learning and keep growing.....”

Share this:

 (<https://sparkbyexamples.com/spark/spark-web-ui-understanding/?share=facebook&nb=1>)


 (<https://sparkbyexamples.com/spark/spark-web-ui-understanding/?share=reddit&nb=1>)

 (<https://sparkbyexamples.com/spark/spark-web-ui-understanding/?share=pinterest&nb=1>)

 (<https://sparkbyexamples.com/spark/spark-web-ui-understanding/?share=tumblr&nb=1>)

 (<https://sparkbyexamples.com/spark/spark-web-ui-understanding/?share=pocket&nb=1>)

 (<https://sparkbyexamples.com/spark/spark-web-ui-understanding/?share=linkedin&nb=1>)

 (<https://sparkbyexamples.com/spark/spark-web-ui-understanding/?share=twitter&nb=1>)

**TAGS: PYSPARK DAG**  
([HTTPS://SPARKBYEXAMPLES.COM/TAG/PYSPARK-DAG/](https://sparkbyexamples.com/tag/pyspark-dag/)), **PYSPARK JOB**  
([HTTPS://SPARKBYEXAMPLES.COM/TAG/PYSPARK-JOB/](https://sparkbyexamples.com/tag/pyspark-job/)), **PYSPARK SCHEDULING**  
([HTTPS://SPARKBYEXAMPLES.COM/TAG/PYSPARK-SCHEDULING/](https://sparkbyexamples.com/tag/pyspark-scheduling/)), **PYSPARK TASK**  
([HTTPS://SPARKBYEXAMPLES.COM/TAG/PYSPARK-TASK/](https://sparkbyexamples.com/tag/pyspark-task/)), **PYSPARK UI**  
([HTTPS://SPARKBYEXAMPLES.COM/TAG/PYSPARK-UI/](https://sparkbyexamples.com/tag/pyspark-ui/)), **PYSPARK WEB UI**  
([HTTPS://SPARKBYEXAMPLES.COM/TAG/PYSPARK-WEB-UI/](https://sparkbyexamples.com/tag/pyspark-web-ui/)), **SPARK JOB**  
([HTTPS://SPARKBYEXAMPLES.COM/TAG/SPARK-JOB/](https://sparkbyexamples.com/tag/spark-job/)), **SPARK SCHEDULING**  
([HTTPS://SPARKBYEXAMPLES.COM/TAG/SPARK-SCHEDULING/](https://sparkbyexamples.com/tag/spark-scheduling/)), **SPARK SQL**  
([HTTPS://SPARKBYEXAMPLES.COM/TAG/SPARK-SQL/](https://sparkbyexamples.com/tag/spark-sql/)), **SPARK TASK**  
([HTTPS://SPARKBYEXAMPLES.COM/TAG/SPARK-TASK/](https://sparkbyexamples.com/tag/spark-task/)), **SPARK UI**  
([HTTPS://SPARKBYEXAMPLES.COM/TAG/SPARK-UI/](https://sparkbyexamples.com/tag/spark-ui/)), **SPARK WEB UI**  
([HTTPS://SPARKBYEXAMPLES.COM/TAG/SPARK-WEB-UI/](https://sparkbyexamples.com/tag/spark-web-ui/)).



**Sriram**

(<https://sparkbyexamples.com/author/sriramrim>)

(<https://Sparkbyexamples.Com/Author/Sriramrimmalapudi9Com/>).

Data Engineer. I write about BigData Architecture, tools and techniques are used to build Bigdata pipelines and other generic blogs.



malapudi  
9gmail-  
com/).

#### > THIS POST HAS 4 COMMENTS



**Chitra**

29 DEC 2020

[REPLY](#)

Thanks Sriram for this great job. It helped me a lot...



**buvana**

17 DEC 2020

[REPLY](#)

You just cleared all my greeks & Latin understanding about Spark UI. Thanks a lot for the very nice write!



**Anonymous**

7 NOV 2020

[REPLY](#)

Great job Sriram. This will be very helpful for lot of aspiring people who wants to learn Bigdata. Appreciate it.



**Shobhit Verma**

19 OCT 2020

[REPLY](#)

Appreciate your effort and deep information. Really helpful and thank you so much. Keep writing.

## Leave a Reply

### PySpark Built-In Functions

[PySpark – when\(\).](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-when-otherwise/>).

[PySpark – expr\(\).](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-sql-expr-expression-function/>).

[PySpark – lit\(\).](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-lit-add-literal-constant/>).

[PySpark – split\(\).](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-convert-string-to-array-column/>).

[PySpark – concat\\_ws\(\).](#)  
([https://sparkbyexamples.com/pyspark-convert-array-column-to-string-column/](https://sparkbyexamples.com/pyspark/pyspark-convert-array-column-to-string-column/)).

[PySpark – substring\(\).](#)  
([https://sparkbyexamples.com/pyspark-substring-from-a-column/](https://sparkbyexamples.com/pyspark/pyspark-substring-from-a-column/)).

[PySpark – translate\(\).](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-replace-column-values/#translate-replace-character-by-character>).

[PySpark – regexp\\_replace\(\).](#)  
([https://sparkbyexamples.com/pyspark-replace-column-](https://sparkbyexamples.com/pyspark/pyspark-replace-column-)

[values/#regexp\\_replace-replace-string-columns](#)).

[PySpark – overlay\(\)](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-replace-column-values/#overlay-function>).

[PySpark – to\\_timestamp\(\)](#)  
([https://sparkbyexamples.com/spark/pyspark-to\\_timestamp-convert-string-to-timestamp-type/](https://sparkbyexamples.com/spark/pyspark-to_timestamp-convert-string-to-timestamp-type/)).

[PySpark – to\\_date\(\)](#)  
([https://sparkbyexamples.com/pyspark-to\\_date-convert-timestamp-to-date/](https://sparkbyexamples.com/pyspark/pyspark-to_date-convert-timestamp-to-date/)).

[PySpark – date\\_format\(\)](#)  
([https://sparkbyexamples.com/pyspark-date\\_format-convert-date-to-string-format/](https://sparkbyexamples.com/pyspark/pyspark-date_format-convert-date-to-string-format/)).

[PySpark – datediff\(\)](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-difference-between-two-dates-days-months-years/#datediff>).

[PySpark – months\\_between\(\)](#)  
([https://sparkbyexamples.com/pyspark-pyspark-difference-between-two-dates-days-months-years/#months\\_between\(\)](https://sparkbyexamples.com/pyspark/pyspark-difference-between-two-dates-days-months-years/#months_between())).

[PySpark – explode\(\)](#)  
([https://sparkbyexamples.com/pyspark-pyspark-explode-nested-array-into-rows/](https://sparkbyexamples.com/pyspark/pyspark-explode-nested-array-into-rows/)).

[PySpark – array\\_contains\(\)](#)  
([https://sparkbyexamples.com/pyspark-pyspark-arraytype-column-with-examples/#array\\_contains](https://sparkbyexamples.com/pyspark/pyspark-arraytype-column-with-examples/#array_contains)).

[PySpark – array\(\)](#)  
([https://sparkbyexamples.com/pyspark-pyspark-arraytype-column-with-examples/#array](https://sparkbyexamples.com/pyspark/pyspark-arraytype-column-with-examples/#array)).

[PySpark – collect\\_list\(\)](#)  
([https://sparkbyexamples.com/pyspark-pyspark-aggregate-functions/#collect-list](https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/#collect-list)).

[PySpark – collect\\_set\(\)](#)  
([https://sparkbyexamples.com/pyspark-pyspark-aggregate-functions/#collect-set](https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/#collect-set)).

[PySpark – create\\_map\(\)](#)  
([https://sparkbyexamples.com/pyspark-pyspark-convert-dataframe-columns-to-maptype-dict/](https://sparkbyexamples.com/pyspark/pyspark-convert-dataframe-columns-to-maptype-dict/)).

[PySpark – map\\_keys\(\)](#)  
([https://sparkbyexamples.com/pyspark-pyspark-maptype-dict-examples/#map\\_keys](https://sparkbyexamples.com/pyspark/pyspark-maptype-dict-examples/#map_keys)).

[PySpark – map\\_values\(\)](#)  
(<https://sparkbyexamples.com/pyspark>

[ark/pyspark-maptypes-dict-examples/#map\\_values](#))

---

[PySpark – struct\(\)](#).  
(<https://sparkbyexamples.com/pyspark/pyspark-structtype-and-structfield/#update-struct-function>)

---

[PySpark – countDistinct\(\)](#).  
(<https://sparkbyexamples.com/pyspark/pyspark-count-distinct-from-dataframe/>)

---

[PySpark – sum\(\).avg\(\)](#).  
(<https://sparkbyexamples.com/pyspark/pyspark-dataframe-groupby-and-sort-by-descending-order/>)

---

[PySpark – row\\_number\(\)](#).  
([https://sparkbyexamples.com/pyspark/pyspark-window-functions/#row\\_number](https://sparkbyexamples.com/pyspark/pyspark-window-functions/#row_number))

---

[PySpark – rank\(\)](#).  
(<https://sparkbyexamples.com/pyspark/pyspark-window-functions/#rank>)

---

[PySpark – dense\\_rank\(\)](#).  
([https://sparkbyexamples.com/pyspark/pyspark-window-functions/#dense\\_rank](https://sparkbyexamples.com/pyspark/pyspark-window-functions/#dense_rank))

---

[PySpark – percent\\_rank\(\)](#).  
([https://sparkbyexamples.com/pyspark/pyspark-window-functions/#percent\\_rank](https://sparkbyexamples.com/pyspark/pyspark-window-functions/#percent_rank))

---

[PySpark – typedLit\(\)](#).  
(<https://sparkbyexamples.com/pyspark/pyspark-lit-add-literal-constant/#typedlit>)

---

[PySpark – from\\_json\(\)](#).  
([https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#from\\_json](https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#from_json))

---

[PySpark – to\\_json\(\)](#).  
([https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#to\\_json](https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#to_json))

---

[PySpark – json\\_tuple\(\)](#).  
([https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#json\\_tuple](https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#json_tuple))

---

[PySpark – get\\_json\\_object\(\)](#).  
([https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#get\\_json\\_object](https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#get_json_object))

---

[PySpark – schema\\_of\\_json\(\)](#).  
([https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#schema\\_of\\_json](https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#schema_of_json))

---

Apache Hadoop  
(<https://sparkbyexamples.com/category/hadoop/>)

Apache Spark  
(<https://sparkbyexamples.com/category/spark/>)

Apache Spark Streaming  
(<https://sparkbyexamples.com/category/spark/apache-spark-streaming/>)

Apache Kafka  
(<https://sparkbyexamples.com/category/kafka/>)

Apache HBase  
(<https://sparkbyexamples.com/category/hbase/>)

Apache Cassandra  
(<https://sparkbyexamples.com/category/cassandra/>)

Snowflake Database  
(<https://sparkbyexamples.com/category/snowflake/>)

H2O Sparkling Water  
(<https://sparkbyexamples.com/category/h2o-sparkling-water/>)

PySpark  
(<https://sparkbyexamples.com/category/pyspark/>)

Spark regexp\_replace() – Replace String Value  
([https://sparkbyexamples.com/spark/spark-regexp\\_replace-replace-string-value/](https://sparkbyexamples.com/spark/spark-regexp_replace-replace-string-value/))

How to Run a PySpark Script from Python?  
(<https://sparkbyexamples.com/pyspark/run-pyspark-script-from-python-subprocess/>)

Spark SQL like() Using Wildcard Example  
(<https://sparkbyexamples.com/spark/spark-sql-like-using-wildcard-example/>)

Spark isin() & IS NOT IN Operator Example  
(<https://sparkbyexamples.com/spark/spark-isin-is-not-in-operator-example/>)

Spark – Get Size/Length of Array & Map Column  
(<https://sparkbyexamples.com/spark/spark-get-size-length-of-array-map-column/>)

Spark Using Length/Size Of a DataFrame Column  
(<https://sparkbyexamples.com/spark/spark-using-length-size-of-a-dataframe-column/>)

Spark rlike() Working with Regex Matching Examples  
(<https://sparkbyexamples.com/spark/spark-rlike-regex-matching-examples/>)

Spark Check String Column Has Numeric Values  
(<https://sparkbyexamples.com/spark/spark-check-string-column-has-numeric-values/>)

Spark Check Column Data Type is Integer or String  
(<https://sparkbyexamples.com/spark/spark-check-column-data-type-is-integer-or-string/>)

SparkByExamples.com is a Big Data and Spark examples community page, all examples are simple and easy to understand, and well tested in our development environment Read more ..  
(<https://sparkbyexamples.com/about-sparkbyexamples/>)

Follow Us



[//www.](#) [//www.](#)



[//twitter](#) [ok.co](#) [n.com/](#) [//github](#)

[r.com/](#) [m/spar](#) [in/n-](#) [b.com/](#)

[sparkb](#) [kbyex](#) [nk-](#) [spark-](#)

[yexam](#) [ample](#) [b860a](#) [examp](#)

[ples\)](#) [s/\)](#) [8193/\)](#) [les/\)](#)

