

Java-Success.com

Prepare to fast-track, choose & go places with 800+ Java & Big Data Q&As with lots of code & diagrams.

[Home](#) [Why? ▾](#) [300+ Java FAQs ▾](#) [300+ Big Data FAQs ▾](#) [Courses ▾](#)[👤 Membership ▾](#) [Your Career ▾](#)[Home](#) > [bigdata-success.com](#) > [Tutorials - Big Data](#) > [TUT - Spark Scala on Zeppelin](#) >

11: Spark on Zeppelin – Dataframe groupBy, collect_list, explode & window

11: Spark on Zeppelin – Dataframe groupBy, collect_list, explode & window

 Posted on [September 15, 2018](#)

Pre-requisite: Docker is installed on your machine for Mac OS X (E.g. \$ brew cask install docker) or Windows 10. [Docker interview Q&As](#). This extends [setting up Apache Zeppelin Notebook](#).

Step 1: Pull this from the docker hub, and build the image with the following command.

```
1 $ docker pull apache/zeppelin:0.7.3
2
```

300+ Java Interview FAQs

300+ Java FAQs



16+ Java Key Areas Q&As



150+ Java Architect FAQs



80+ Java Code Quality Q&As



150+ Java Coding Q&As



300+ Big Data Interview FAQs

300+ Big Data FAQs



Tutorials - Big Data

TUT -  Starting Big Data

TUT - Starting Spark & Scala

You can verify the image with the “docker images” command.

Step 2: Run the container with the above image.

```
1 $ docker run --rm -it -p 8080:8080 apache/zeppelin
2
```

Step 3: Open Zeppelin notebook via a web browser “http://localhost:8080”. Create a note book with “spark” as a default interpreter.

How to aggregate values into collection after groupBy?

collect_list(expr) – Collects and returns a list of non-unique elements.

collect_set(expr) – Collects and returns a set of unique elements.

concat(str1, str2, ..., strN) – Returns the concatenation of str1, str2, ..., strN.

concat_ws(sep, [str | array(str)]+) – Returns the concatenation of the strings separated by sep.

```
1 %spark
2
3
4 import org.apache.spark.sql.functions.collect_list
5
6 case class Employee (id: Integer, name: String, ex
7
8 val employees = Seq(
9     Employee(1, "John", "Java" ),
10    Employee(2, "Peter", "Scala"),
11    Employee(2, "Peter", "Python"),
12    Employee(2, "Peter", "Spark"),
```

TUT - Starting with Python

TUT - Kafka

TUT - Pig

TUT - Apache Storm

TUT - Spark Scala on Zeppelin

TUT - Cloudera

TUT - Cloudera on Docker

TUT - File Formats

TUT - Spark on Docker

TUT - Flume

TUT - Hadoop (HDFS)

TUT - HBase (NoSQL)

TUT - Hive (SQL)

TUT - Hadoop & Spark

TUT - MapReduce

TUT - Spark and Scala

TUT - Spark & Java

TUT - PySpark on Databricks

TUT - Zookeeper

800+ Java Interview Q&As

300+ Core Java Q&As



300+ Enterprise Java Q&As



150+ Java Frameworks Q&As



120+ Companion Tech Q&As



Tutorials - Enterprise Java



```
13     Employee(1, "John", "JEE"),
14     Employee(6, "Elliot", "Unix")
15 )
16
17 val employeeDF = spark.createDataFrame(
18     spark.sparkContext.parallelize(employees)
19 )
20
21 val expertise = collect_list($"expertise").alias("expertise")
22
23 val resultDF = employeeDF.groupBy($"id")
24     .agg(expertise)
25
26 resultDF.show(false)
27
```

Output:

```
1 import org.apache.spark.sql.functions.collect_list
2 defined class Employee
3 employees: Seq[Employee] = List(Employee(1,John,Java),Employee(2,Peter,Scala),Employee(2,Peter,Python),Employee(2,Peter,Spark),Employee(6,
4 employeeDF: org.apache.spark.sql.DataFrame = [id: int,expertise: string]
5 expertise: org.apache.spark.sql.Column = collect_list(expertise)
6 resultDF: org.apache.spark.sql.DataFrame = [id: int,expertise: string]
7 +---+-----+
8 |id|expertise|
9 +---+-----+
10 |1 |[Java, JEE]|
11 |6 |[Unix]|
12 |2 |[Scala, Python, Spark]|
13 +---+-----+
14
```

What if expertise is an array? explode first

```
1 %spark
2
3
4 import org.apache.spark.sql.functions.{collect_list, explode}
5
6 case class Employee (id: Integer, name: String, expertise: Array[String])
7
8 val employees = Seq(
9     Employee(1, "John", Array("Java", "Scala")),
10    Employee(2, "Peter", Array("Scala")),
11    Employee(2, "Peter", Array("Python", "Git")),
12    Employee(2, "Peter", Array("Spark")),

```

```

13     Employee(1, "John", Array("JEE", "Spring")),
14     Employee(6, "Elliot", Array("Unix", "DevOps"))
15 )
16
17 val employeeDF = spark.createDataFrame(
18     spark.sparkContext.parallelize(employees)
19 )
20
21 employeeDF.show()
22
23 val resultDF = employeeDF.select($"id", explode($"expertise"
24     .groupBy($"id")
25     .agg(collect_list($"expertise")))
26
27 resultDF.show(false)
28

```

Output:

```

1 import org.apache.spark.sql.functions.{collect_list, explode}
2 defined class Employee
3 employees: Seq[Employee] = List(Employee(1,John,[JEE, Spring]),
4 employeeDF: org.apache.spark.sql.DataFrame = [id: int, name: string, expertise: array]
5 +---+-----+-----+
6 | id | name | expertise |
7 +---+-----+-----+
8 |  1 | John | [Java, Scala] |
9 |  2 | Peter | [Scala] |
10 |  2 | Peter | [Python, Git] |
11 |  2 | Peter | [Spark] |
12 |  1 | John | [JEE, Spring] |
13 |  6 | Elliot | [Unix, DevOps] |
14 +---+-----+-----+
15 resultDF: org.apache.spark.sql.DataFrame = [id: int, collect_list(expertise): array]
16 +---+-----+-----+
17 | id | collect_list(expertise) |
18 +---+-----+-----+
19 |  1 | [Java, Scala, JEE, Spring] |
20 |  6 | [Unix, DevOps] |
21 |  2 | [Scala, Python, Git, Spark] |
22 +---+-----+-----+
23

```

Cumulative expertise the window function

```

1 %spark
2

```

```

3
4 import org.apache.spark.sql.functions.{collect_set, collect_list}
5 import org.apache.spark.sql.expressions.Window
6
7 case class Employee (id: Integer, name: String, expertise: List[String])
8
9 val employees = Seq(
10   Employee(1, "John", Array("Java", "Scala")),
11   Employee(2, "Peter", Array("Scala")),
12   Employee(2, "Peter", Array("Python", "Git")),
13   Employee(2, "Peter", Array("Spark")),
14   Employee(1, "John", Array("JEE", "Spring")),
15   Employee(6, "Elliot", Array("Unix", "DevOps"))
16 )
17
18 val employeeDF = spark.createDataFrame(
19   spark.sparkContext.parallelize(employees)
20 )
21
22 employeeDF.show()
23
24 val resultDF = employeeDF.select($"id", explode($"expertise" as expertise_array)
25                               .withColumn("expertise_count", count($"expertise_array")))
26
27 resultDF.show(false)
28

```

Output:

```

1 import org.apache.spark.sql.functions.{collect_set, collect_list}
2 import org.apache.spark.sql.expressions.Window
3 defined class Employee
4 employees: Seq[Employee] = List(Employee(1,John,[Java, Scala]), Employee(2,Peter,[Scala]), Employee(2,Peter,[Python, Git]), Employee(2,Peter,[Spark]), Employee(1,John,[JEE, Spring]), Employee(6,Elliot,[Unix, DevOps]))
5 employeeDF: org.apache.spark.sql.DataFrame = [id: integer, name: string, expertise: array<string>]
6 +----+-----+-----+
7 | id | name | expertise |
8 +----+-----+-----+
9 |  1 | John | [Java, Scala] |
10 |  2 | Peter | [Scala] |
11 |  2 | Peter | [Python, Git] |
12 |  2 | Peter | [Spark] |
13 |  1 | John | [JEE, Spring] |
14 |  6 | Elliot | [Unix, DevOps] |
15 +----+-----+-----+
16 resultDF: org.apache.spark.sql.DataFrame = [id: integer, expertise_count: integer, expertise_array: array<string>]
17 +----+-----+-----+
18 | id | expertise_count | expertise_array |
19 +----+-----+-----+
20 |  1 | 2 | [JEE] |
21 |  1 | 2 | [Java, JEE] |
22 |  1 | 2 | [Scala, Java, JEE] |
23 |  1 | 2 | [Scala, Java, Spring, JEE] |

```

```
24 | 6 | DevOps | [DevOps] |
25 | 6 | Unix | [Unix, DevOps] |
26 | 2 | Git | [Git] |
27 | 2 | Python | [Python, Git] |
28 | 2 | Scala | [Scala, Python, Git] |
29 | 2 | Spark | [Scala, Spark, Python, Git] |
30 | +---+-----+-----+
31 |
```

◀ 10: Spark on Zeppelin – union, udf and explode

12: Spark on Zeppelin – Dataframe pivot ▶

Disclaimer

The contents in this Java-Success are copyrighted and from EmpoweringTech pty Ltd. The EmpoweringTech pty Ltd has the right to correct or enhance the current content without any prior notice. These are general advice only, and one needs to take his/her own circumstances into consideration. The EmpoweringTech pty Ltd will not be held liable for any damages caused or alleged to be caused either directly or indirectly by these materials and resources. Any trademarked names or labels used in this blog remain the property of their respective trademark owners. Links to external sites do not imply endorsement of the linked-to sites. [Privacy Policy](#).