

Anna Plust, 36147

Duże zbiory danych

Lab 1, Sprawozdanie

1. Wykonać program dla platformy Spark w trybie interaktywnym.

Wykonanie programu lokalnie w trybie interaktywnym:

```
scala> val text = sc.textFile("hdfs://hadoop1:9000/students/st36148/plik.txt");
text: org.apache.spark.rdd.RDD[String] = hdfs://hadoop1:9000/students/st36148/plik.txt MapPartitionsRDD[9] at textFile at <console>:23

scala> val words = text.flatMap(line => line.split(" ", "\\."));
words: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[10] at flatMap at <console>:23

scala> val pairs = words.map(w => (w, 1));
pairs: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[11] at map at <console>:23

scala> val counts = pairs.reduceByKey((a, b) => a + b);
counts: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[12] at reduceByKey at <console>:23
```

```
scala> counts.foreach(c => println(c));
(1859,1)
(ono,1)
(szwedzka,1)
((przypisy,2)
(pryncypał,1)
(multiplicati,1)
(Gdy,1)
(motywy),1)
(stronie,2)
(cudzego,1)
(marnie,1)
(Transylwańczyk,1)
(Sama,1)
(merytorycznym,1)
(dalej,2)
(ci,1)
(poszła,1)
(pańs,2)
(książki,1)
(Kotwica,2)
(zawsze,1)
(wyd,1)
(Opracowanie,1)
(ucieszył,1)
(chociaż,1)
(którzy,1)
(techniczną,1)
(«Nie,1)
(wojny,2)
(pustoszył,1)
(wojnę,1)
(drzewo,1)
(sądzę,1)
(te,1)
```

Dane wynikowe można wydrukować i zobaczyć w konsoli, ponieważ wszystkie komendy są wykonywane lokalnie.

Wykonanie programu w trybie interaktywnym uruchomionego w trybie klastra: spark-shell --master spark://hadoop1:7077

```
scala> val text = sc.textFile("hdfs://hadoop1:9000/students/st36148/plik.txt");
text: org.apache.spark.rdd.RDD[String] = hdfs://hadoop1:9000/students/st36148/plik.txt MapPartitionsRDD[1] at textFile at <console>:23

scala> val words = text.flatMap(line => line.split(" ", "\\."));
words: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[2] at flatMap at <console>:23

scala> val pairs = words.map(w => (w, 1));
pairs: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[3] at map at <console>:23

scala> val counts = pairs.reduceByKey((a, b) => a + b);
counts: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[4] at reduceByKey at <console>:23

scala> counts.foreach(c => println(c));

scala>
```

W trybie klastra wynik nie jest widoczny w konsoli, ponieważ program (w tym metoda println) jest wykonywany na serwerze. Wynik można zobaczyć za standardowym wyjściu workerów, które zajmowały się wykonaniem programu.



stdout log page for app-20221215172603-0108/3

[Back to Master](#)

Showing 2529 Bytes: 0 - 2529 of 2529

Load More

```
(ono,1)
((przypisy,2)
(pryncypa1,1)
(Gdy,1)
(motywy),1)
(stronie,2)
(marnie,1)
(Transylwańczyk,1)
(Sama,1)
(merytorycznym,1)
(dalej,2)
(pańs,2)
(-i 1)
```



stdout log page for app-20221215172603-0108/0

[Back to Master](#)

Showing 2925 Bytes: 0 - 2925 of 2925

Load More

```
(1859,1)
(szwedzka,1)
(multiplicati,1)
(cudzego,1)
(poszła,1)
(chocia1,1)
(którzy,1)
(techniczną,1)
(«Nie,1)
(wojny,2)
(pustoszy1,1)
(wojnę,1)
(drzewo,1)
(sądzę,1)
(tę,1)
```

2) Stworzyć i uruchomić w języku Java aplikację dla platformy Spark

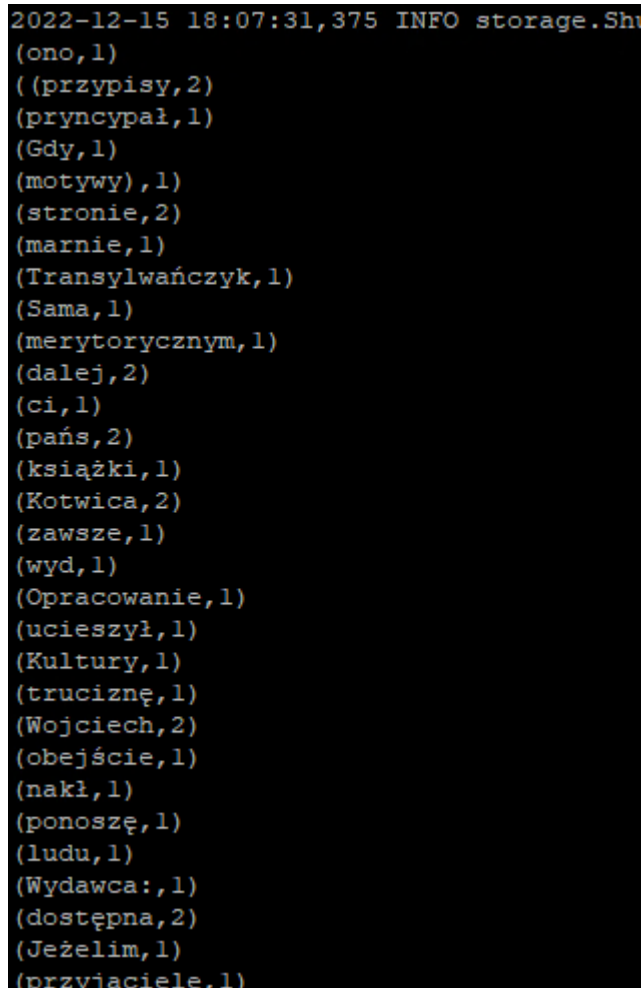
```
public class Main {  
    public static void main(String[] args) {  
  
        SparkConf conf = new SparkConf();  
        JavaSparkContext sc = new JavaSparkContext(conf);  
  
        JavaRDD<String> text = sc.textFile("hdfs://hadoop1:9000/students/st36148/plik.txt");  
        JavaRDD<String> words = text.flatMap(line -> Arrays.asList(line.split("[ ,;\\.]")).iterator());  
        JavaPairRDD<String, Integer> pairs = words.mapToPair(word -> new Tuple2<>(word, 1));  
        JavaPairRDD<String, Integer> counts = pairs.reduceByKey((a, b) -> a + b);  
        counts.foreach(c -> System.out.println(c));  
    }  
}
```

Tryb lokalny

Program uruchomiony komendą:

```
spark-submit --class Main lab1-1.0-SNAPSHOT.jar
```

Wynik jest widoczny w konsoli:



```
2022-12-15 18:07:31,375 INFO storage.Shu  
(ono,1)  
( (przypisy,2)  
(pryncypał,1)  
(Gdy,1)  
(motywy),1)  
(stronie,2)  
(marnie,1)  
(Transylwańczyk,1)  
(Sama,1)  
(merytorycznym,1)  
(dalej,2)  
(ci,1)  
(pańs,2)  
(książki,1)  
(Kotwica,2)  
(zawsze,1)  
(wyd,1)  
(Opracowanie,1)  
(ucieszył,1)  
(Kultury,1)  
(truciznę,1)  
(Wojciech,2)  
(obejście,1)  
(nakł,1)  
(ponoszę,1)  
(ludu,1)  
(Wydawca:,1)  
(dostępna,2)  
(Jeżelim,1)  
(przviaciele,1)
```

Tryb klastra, manager spark

spark-submit --class Main --master spark://hadoop1:7077 lab1-1.0-SNAPSHOT.jar

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20221215181736-0111	Main	16	1024.0 MiB		2022/12/15 18:17:36	st36148	FINISHED	5 s

Wynik nie jest widoczny w konsoli, ale na standardowym wyjściu workera:



stdout log page for app-20221215181736-0111/3

[Back to Master](#)

Showing 5454 Bytes: 0 - 5454 of 5454

Load More
(1859,1) (ono,1) (szwedzka,1) ((przypisy,2) (multiplicati,1) (pryncypał,1) (cudzego,1) (Gdy,1) (poszła,1) (motywy),1) (chociaż,1) (stronie,2) (którzy,1) (marnie,1) (techniczną,1) (Transylwańczk,1)

Tryb klastra, manager yarn, client

spark-submit --class Main --master yarn --deploy-mode client lab1-1.0-SNAPSHOT.jar

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus
application_1669516454756_0064	st36148	Main	SPARK		default	0	Thu Dec 15 19:25:47 +0100 2022	Thu Dec 15 19:25:47 +0100 2022	Thu Dec 15 19:26:01 +0100 2022	FINISHED	SUCCEEDED

Application Attempt appattempt_1669516454756_0064_000001

Application Attempt State:	FINISHED
Started:	Cz gru 15 18:25:47 +0000 2022
Elapsed:	14sec
AM Container:	container_1669516454756_0064_01_000001
Node:	N/A
Tracking URL:	History
Diagnostics Info:	
Nodes blacklisted by the application:	-
Nodes blacklisted by the system:	-

Driver program jest uruchomiony po stronie klienta, dlatego węzeł nie został przydzielony.

Wynik nie jest widoczny w konsoli.

Tryb klastra, manager yarn, cluster

spark-submit --class Main --master yarn --deploy-mode cluster lab1-1.0-SNAPSHOT.jar

Show 20 entries											
ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus
application_1669516454756_0065	st36148	Main	SPARK		default	0	Thu Dec 15 19:28:22 +0100 2022	Thu Dec 15 19:28:22 +0100 2022	Thu Dec 15 19:28:40 +0100 2022	FINISHED	SUCCEEDED

Application Attempt appattempt_1669516454756_0065_000001

Application Attempt State:	FINISHED
Started:	Cz gru 15 18:28:22 +0000 2022
Elapsed:	17sec
AM Container:	container_1669516454756_0065_01_000001
Node:	hadoop3:37373
Tracking URL:	History
Diagnostics Info:	
Nodes blacklisted by the application:	-
Nodes blacklisted by the system:	-

Driver program jest uruchomiony na nodzie klastra.

Wynik nie jest widoczny w konsoli.