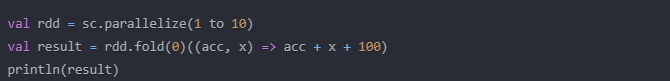
Analyze the application of fold() and aggregate() functions in Spark by considering a scenario where all the items in a collection are updated by a count of 100. Evaluate the efficiency, performance, and suitability of both.

A black background with white text

Description automatically generated

Consider a text file text.txt. Develop Spark code to read the file and count the number of occurrences of each word using Spark RDD. Store the result in a file. Display the words which appear more than 4 times.

A screenshot of a computer program

Description automatically generated

Consider the content of text file text.txt. Perform the counting of occurrences of each word using pair RDD.

A screenshot of a computer program

Description automatically generated

Write the Spark Code to print the top 10 tweeters. Tweet Mining: A dataset with the 8198 reduced tweets, reduced-tweets.json will be provided. The data contains reduced tweets as in the sample below: {"id":"572692378957430785", "user":"Srkian\_nishu smile", "text":"@always\_nidhi @YouTube no idnt understand bti loved of this mve is rocking", "place":"Orissa", "country":"India"} A function to parse the tweets into an RDD will be provided.



Simulate the following scenario using Spark streaming. There will be a process which will be streaming lines of text to a unix port using socket communication. The process we can use for this purpose is netcat. It will stream lines typed on the console to a unix socket. The spark application needs to read the lines from the specified port, and it needs to produce the word counts on the console. A batch interval of 5 second can be used.

A screen shot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

Develop the spark code to find the average of marks using the combineByKey() operation. Sample Input format: Array( ("Joe", "Maths", 83), ("Joe", "Physics", 74), ("Joe", "Chemistry", 91), ("Joe", "Biology", 82), ("Nik", "Maths", 69), ("Nik ", "Physics", 62), ("Nik ", "Chemistry", 97), ("Nik ", "Biology", 80))

A computer screen shot of a program code

Description automatically generated

Consider the Employee table with the fields as (EmpID, Dept, EmpDesg). Design the Spark program to partition the table using Dept and construct the hashed partition of 4.

A screenshot of a computer program

Description automatically generated

Consider a collection of 100 items of type integer given in the csv file. Write the Spark code to find the average of these 100 items.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

Consider a collection with items as (11,34,45,67,3,4,90). i. Illustrate how spark context will construct the RDD from the collection, assuming number of partitions to be made is 3. ii. Using mapPartitionsWithIndex return content of each partition along with partition index and apply a function, that increments the value of each element by 1, and returns an array.

A screenshot of a computer program

Description automatically generatedA screenshot of a computer program

Description automatically generated

**S**

Consider the Item object. Item = Map{(“Ball”:10), (“Ribbon”:50), (“Box”:20), (“Pen”:5), (“Book”:8), (“Dairy”:4),(“Pin”:20)} Design the spark program to perform the following i. Find how many partitions are created for the collection Item? ii. Display the content of the RDD Display the content of each partition separately

A screenshot of a computer program

Description automatically generated

The table 1b provides the distributed data of the scala object Item = Map{(“Ball”:10), (“Ribbon”:50), (“Box”:20), (“Pen”:5), (“Book”:8), (“Dairy”:4),(“Pin”:20)}

Partition1 {(“Ball”:10) (“Ribbon”:50) (“Box”:20) Partition2 (“Pen”:5) (“Book”:8) Partition3 (“Dairy”:4) (“Pin”:20)

Table 1b: Data distribution Design the spark program to perform the following i. Find how many partitions are created for the collection Item? ii. Display the content of the RDD iii. Display the content of each partition separately.

A screen shot of a computer program

Description automatically generated

Consider the text file words.txt as shown in the figure 1a. Write the spark code to perform the following. i) Count the number of occurrences of each word. ii) Arrange the word count in ascending order based on Key. iii) Display the words that begin with ‘s’. She sells sea shells by the seashore And the shells she sells by the seashore Are sea shells for sure Figure 1. A

A screenshot of a computer program

Description automatically generated

**S**

Illustrate the application of combineByKey to combine all the values of the same key in the following collection. ((“coffee”,2),(“cappuccino”,5),(“tea”,3),(“coffee”,10),(“ cappuccino”,15))

A screenshot of a computer program

Description automatically generated