Experiment-04

1) spak frame work is from apache rendol which is open source 2) spark is only meant for sig data processing (using in-numory cluster computing primitive) -- however it's not meant for storage 3) spark does not have any storage component --- and it go read-the input data from localtile system LLFS). HDFS, an FDBMS, NOSQL, cloud Vendors (Azure (Aws)

-> one of the processing is done from spark, it can write the olp data to any of the above storage systems.

Then we install spark we will see (spark core, sparks) spark streaming, sparkmillib, spark Graph-X)

3 modules

> spark is a united stack

-) spark processing can be achieved through

1. scala language

2. Python epysparle four compatable

3. P. Canguage

ce. Java language

Advantage!

The main advantage of spark is because of its In-memory duster computing a processing which is not present in our hadop hadoop is a desk based-processing system.

-> Hadoop follows top to bottom approach othere as spark tollows bottom to top approach!

- It we want to see the different methods in spark, we need to type sealar sc. (And press tab)

If well show the different mothods in spark, -) In spark we have a different RD operations like

1 Transformation DAction

1) Transformation is a method of an PDD exacts a new PDD by	
fring a computation on the source pop	
1) Map 5) Union	-
a) filter o) Intersection	-
3) Nat map 7) substract	-
1) Map 5) Union 2) Filter b) Intersection 3) Flat map 7) substract W Map partitions 8) sort by	+
plactions eve the me mode that return a value to a disvey.	+
O collect - method returns climents in source pop as an array.	1
a count - The method returns a count of eliments in source ROD	-
To sun spark processing in a standatione cluster manage no	KTP2
1) Using scala -> spark-shell 2) Using python -> pyspark	
2) Using python -> pyspark 3) Using R >> spark P	
3) Using F > spart	
I scalas har file = sc. text File ("file: 111 homel vbox enser [prac [spart]	
1001 I MO	- 10-1
scala > mapfile - file map (x =) (x, x. lingth)) scala > mapfile collect · for each (printly)	N
scales day mapa = to contact of the collect of the	
output: (Apache spark is an open source, ,32)	
The saide range data proposing enont of	
(It is wide range data processing engine, 38) (There is a common belief that Apache spart, 42) (15 an extension of Hadoop, 28)	
(15 cm extension of Hadoop, 28)	
It uses thatoop for storage parpose only, 41)	
Il scala z vox data = sc. paralleliz (List ("spark", "scala", "c",	
It uses Hadoop for Storage parpose only. All) Il scala > Var data = St. paralleliz (List ("spark", "scala", "c", "cobol", "ve+t", "Hadoop", 'fuby", 'Information", "block chain "generative AI", "chat Gpt", "java", "python", "ct+t")) Scala > var mapdata = data. map (x => (x,x. length)) Scala > randata. colut. breach (prontlo))))
" generative AI" "chot Got" "iava" "python", "ctt")	-
scala s var mardata = data. map (x => (x,x. length))	
scala > Mapdata. collect. breach (printle)	

```
autput: (spart, 5)
          (scala, 5)
          (C,1)
           (cobol,5)
          (uct+, u)
          (hadoop, b)
          (raby 14)
          (information, 11)
          (blockchain, to)
          (generativeAI,12)
          (chatGPT, 7)
         (jong,u)
          (python, 6)
-) Here collect gathers all -lu elements of RDD or dotatrane

-) for each is an action that implies a tunction to each
-) printle is simply printing out each element
   element of collected data.
 > It collideall elements from Mapdata & then prints each
  taple to the consol.
scala , var Mapdata = data. map (x=) (x, x. reverse, x. lingth)
scala = mapdata. collect. for cach (printle)
output: (spark, traps, 5)
         (scala, alacs, 5)
          (cobol, loboc, 5)
          (VC++,++cV., 4)
          (hadoop, poodah, 6)
          (Ruby, your, u)
          (information, noitam rothi
          (blockchain, niahclecolb, 10)
           (generative AI, IA exit areneg, 12)
           (chater, TPG take, 7)
           (java, avaj, 4)
           (python, nothlyp, 6
```

filler transformation! - filter is used like where clause type scalas var file = sc. textfile ("file; Illhome luboxuser (prac) scalas var filfile - tile. filter (x=)x.length >=10) Spart / Test. log ") scala > file. count OP: - 1887: hong = 19 scala > filthe count 888: hong = 17 scale > filtile. collect. foreach (printlp) output: spark sol is Apache spark's module for coorling with structured data. The soil syntax section describes the SQL syntax in detail along with usage examples when applicable. Spark SOL> To work with structured & semi. structured data by making use of tabular kind of view, we are going ahead with spark sql. => Whod is main entry point in spark core => sc =) What is main programming abstraction in spark corr is - 3 RDD (Resilient pistributed database) > What is main entry point in SparksQL -> spark =) What is main programming abstraction in spark SQL-> Octobrane (DF) > What is a Dataframe (DF) DataFrame is a now-column object wollich is a schema based pop schema + PDD => Its a kind a table in RDBMs => notaframe in spark. we create an ROD by making use of sc. tixtlife sc. parallelize

=> from the created poo how to convert poo to pala frame. convert rad to dataframe. syntax: - ndd - object . to of oclaselt method available a which is automatically Eg:-scala > var data = sc. parallelize (1 tour) scala >var mapdata = data. map (x=)(x,(x+x)) scala > var dfobj = mapdada. to DF scala > dtobj. columns output: resta: Array [org. apach. spark. sql. mw] = Array [11,13] [2,4], [3,9], [4,16], [5,25], [6,36], [7,49], [8,64], [9,81], [10,100] - - - - [ao, 1600]) scala > dtobj . show() octput! · Dhorthin dute when only showing top 20 mos scalas Hobj. show(40)

The course only shace it coil display only to records in given tall it or having more than all records, are should newton to	output:- +	
1		
1 5 126 120 12		3 9
1	1	4 16
10 81 100 101 100		
10 181 11 121 11 181 11		8
13 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1		? 81
13 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1		
18 196 18 250 19 520 19		
16 255 18 12 259 19 510 20 361 21 400 22 404 22 434 23 656 26 625 27 656 28 656 29 656 29 150 20 150 20	1	
19		16
196 100	1201	H 1 288 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
26 361 000 0001 0001 101 101 101 101 101 101	0	
22 USY 12 ST6 656 24 154 25 150 26 150 27 160 28 170 29 160 29 160 20 10	1	20 3611 00012 (000125-1101
23 1369 1 24 1369 1 36 1369 1 36 1369 1 36 1369 1 36 1369 1 37 1464 1 36 1369 1 37 1464 1 37 1464 1 38 1369 1 36 1444 1 37 1444 1 37 1444 1 37 1444 1 37 1444 1 37 1444 1 37 1444 1 37 1444 1 38 1 369 1 36 1444 1 37 1444 1 37 1444 1 37 1444 1 38 1 369	i	
1		22 484
29 1 + 64 29 1 + 64 29 1 + 64 29 1 + 64 21 1 900 21 1 900 22 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1	24 529 1
30 1 24 125 1069 106 1		
29 1 + 84 30 1 841 31 900 1 32 196 1 33 1024 1 36 1126 1 36 1126 1 36 1126 1 37 1266 1 38 1 369 1 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4	1111111	
30 1 841 31 900 132 196 133 1089 10 10 10 10 10 10 10 10 10 10 10 10 10		2.5
1 32 1 96° 1 33 1024 1089 10 1081 1081 1081 1081 1081 1081 1	3	
1 35 1156 36 1225 1 34 1246 1 36 1246 1	. (01)	31 900
1 35 1156 36 1225 1 34 1246 1 36 1246 1	To de San	(32 1 % 1 (0) (0) (0) (0) (0) (0) (0) (0)
36 1225 37 1296 38 1369 White Halling And Andrew Market	10	33 (024)
33 1296 38 1369 White Halling And Market Ma	'	TOUR TOUR DE LA COMPANIE DE LA COMPA
38 1369 Who Ho Who Ho Who Ho Who Se Who		36 1125 P. Malo robst by Mach Deck Very part 120000
38 1 369 Who had had the state of the state		34 1296
+	(29 1 36 9 10 11 11 11
+)	36 1444 - 101 - 10
The use only show it will display only to records in given tall, it we having more than do records, use should newtion to purply is I show I is an action.	(10012-11	03 40 100 100 100 100 100 100 100 100 100
The use only show it will display only to records in given table, if we having more than all records, use should newtion to purply is I show II is an action.	+	
table, if we having more other do records, use should newtion to	246	I at a st well disday and in records in given
table, if we having more than a records, we should me worth to	=) It we	use only show the wife of the state of the s
number to I should -) it is an action.	table if	we having more than all records, we should newton to
	nti mle	= I short I -> it is an action.

```
Eg: the number should be above to & the square square square square the square to show the stable scales speak square to show tables of stables of the square to stables of the square to stable square should be above to & the square
```

		(" select " AND ++	0.415	mn 11) e	hasa	
. 4,			-d<1	300 1.3	1000	AREA BREE
output	+	4+				
	1 _1	1-21				
	+	1				
	1 11	121				
	1 12	169				
		THE PERSON NAMED IN COLUMN TWO IS NOT THE OWNER.				
	1 14					
	16	256				
	+ (2	289			The state of the s	
	1 18					
	1 19	324				
	20	361				
	21	000				
	21 22 23 24 25	1 484 1				
	1 24	1 529 1				
	, 25	571				
	26	625				
	1 27	676				STARKE TO
	. 28	1 729 1				
	29	1 841				
	30	900				