Spork SOL
Scala + object oriented Spork 802 Supports older function Programming semi-structured or day/Date amount / M2/streaming/graphs. feautures.
APACHE SPARK (fast & general engine tox large-scare
programming -> RDO. date)
· Written in Scala
· Developed by AMPhab UC Berkely, now by Databricks.co
Spork Cove API
R Python Scala See Jeva.
· Spark Arthitecture
- Master-Slever, arthitecture > master > Driver' Sleves > "Norkers"
"- Transformation and actions are executed on
Worker node.
context (main entry point to speck A)
· Spark shell privides preconfigured spark context
00.10 00
Processed across Cluster
· ROD (Resilient Distributed Detaset)
data in memors
10st, it can be
reciated date
ROD fundamental units of data in spars (containe,
- In - memory computation - Computed results
· - Lazy Evaluation - not performed immediately
(RAm). Very fast. - Lazy Evaluation - action than the redictery - Fault tolerance + of feilure occurs in any partition of Roo, partition can be ve-computed
Roo, partition can be ve-computed
from axignal feult to letaner input data to create it.

of Roll	by accessor and Our con					
Day / Date	Cayying prisit by cache does not 11.03.4					
	Position > collection of verious date (ROD)					
	· Can not fit into a single mode					
1	o- Persistance > · Seve result of Rpb evaluation					
	Close by accepting underlying RDD. Gate Calling Persist or cache does not trisser preclution for computation. - Immutability > Nature can not be Offinged - Partitioning > Collection of veriety data (RDD) - Partitioning > Collection of veriety data (RDD) - Partitioning > Collection of RDD evaluation - Consultance > Store result of RDD evaluation - Collection new RDD Yetwan Value. + to beind on Criticat one Rod Yetwan Value. + to beind on Criticat one Rod texter the one of text ordered one - May or collection network of texter the one of text ordered the result of text ordered the result of text ordered the result of					
Preclation Computation. - Immutability + yelve can not be Changed - Partitioning - collection of veries data (ROD) - Can med fit into estingle needs - Persistance - sever result of RDD evaluation - Persistance - sever result of RDD evaluation - Colfine new RDD - Vetorn velve. + O - Descal on Civient - One of						
	input item	flatmap () = Yeturns contect () Flatmap () = Yeturns content () Flatmap () = Ye				
to 0 0 1	Calling prisit or choke does not trisser Preclution prisit or choke does not trisser Preclution computation. - Immutability + Velue can not be affirmed. - Partitioning + collection of verious data (ROD) - Partitioning + collection of verious data (ROD) - Partitioning + collection of verious data (ROD) - Partitioning + collection of PDD evaluation - Presistance + sever result of PDD evaluation - Connect Sever versult. - Transformation - Collection nesses tester to take product. - Private on Civilat delivery master node. - One confection nesses tester to take productly - Private that ness tester to take productly - Private versus fraction result. - Private versus fraction - Private versus fraction - Presistance from the perform sequences of the second of the lineage of the product of the performance of the lineage of t					
(resuns ser	Several Sequence (ste ()					
then	· RDD organized as pirected theyelic graph (DAG)					
does by accessing underlying RDD. By 19th Calling Persist or cache does not trisser Precution (competition.) - Immutability > Nelve can not be Charged - Partitioning > Collection of Verious casts (RDD) - Partitioning > Collection of Verious casts (RDD) - Partitioning > Collection of Verious casts (RDD) - Partitioning > Collection of RDD evaluation - Partitioning > Save result of RDD evaluation - Colfine RDD Verion of Verion value. To drivery master rules. Transforgation Action - Colfine RDD overation result of RDD evaluation - Count () text rules - Count () Count by Key () - Count by May () - Count by Key () - Count by Your So no chief () - Count by Key () - Count by Ke						
	-nodes are Roos					
1/2.						
	· Pipelining (Spark will perform sequences of					
	is just by you so no date					
	Calling persent by couche does not trisser precision from persent by couche does not trisser precision from persent by couche does not trisser precision from persent by value can not be alternated as a collection of versus and (ROD) - Partitioning - collection of versus and (ROD) - Partitioning - collection of person personal personal collection of the precision of the precis					
•	with best					
	Calling prisit or choke does not trisser Preclution prisit or choke does not trisser Preclution computation. - Immutability + Velue can not be affirmed. - Partitioning + collection of verious data (ROD) - Partitioning + collection of verious data (ROD) - Partitioning + collection of verious data (ROD) - Partitioning + collection of PDD evaluation - Presistance + sever result of PDD evaluation - Connect Sever versult. - Transformation - Collection nesses tester to take product. - Private on Civilat delivery master node. - One confection nesses tester to take productly - Private that ness tester to take productly - Private versus fraction result. - Private versus fraction - Private versus fraction - Presistance from the perform sequences of the second of the lineage of the product of the performance of the lineage of t					
	by accepting underlying RDD. Calling Prisit or cache does not trisser Precution (computation. - Immutability + yelve can not be changed - Partitioning + contection of veriors cathe (RDD) - Partitioning + contection of RDD evaluation - Partitioning + contection Action - Contection Action - Contection Partition - Count () + textumple() - Partition return to texture () + textumple() - Partition of transformation on thick () + textumple() - Partition of transformation on the texture of the definition of transformation on the performance () - Partition of texture () - Partition of texture () - Partition of textur					
	Contract of the contract of th					
	resource usige. To optimize, see that action is Almino					
	- 11VV/11/1					

	JVM > Java vertue & machine.
Day / Date	
	Roo. eache()
'	Cache Roo into memory
	By calling cache method, first action say Roo Keep
	values it has calculated in memory, ROD than
	uses cached values for calculating 2nd action.
	Unpersisting
	. As more and more Ropis are cached, memory elected. Spark starts expelling partition from cache. Zaida JVM garbadge collection time of Unavoidable.
\bigcap	·- Call un-persist method on Roo when caching is
	no longer needed,
1	
/w/	Why not to use Caching (Recomposation is faster as increase memory of more money.
	V ·
	verd dateset How many times data is accessed
11 - 1	then no point Amount Jof work involved.
	PySpark (interface for spark in python) pyspark shell.
	PySpark (interface for spark in python)
	Pyspark shell.
	Spark Data frame Pandas Dattaframe.
	· supports parallelization no pavallelization
	· Multiple Modes single mode
	· Lazy execution tager execution
	· Immutable Motable
Ł	Spark determe is for Parge deter
	Spark determe is for large dets.
	faster for large amount of date.
C-1 1 1	Maxim

۰	Simple Algo	for Reques	nt elements i	n stream Er begs
Day / Date •	Spark SQ1	is not a r	eplacement for	a databar
	· Natron	Transfer metion	Wide Tya	ns formation
	each input partition	output Partion	each ->	many outputs
	·- Faster		Slower	
	Not requir	re any data	might tr	equiver data
Day/Date o Spark SQL is not a replacement for Distributed to other application. Narpor Transformation Wide Tye or each one each of inpot perntion partition - Faster Slower - Not require any class might in Shuffling over cluster shuffling network. - Map, flatten map, Mep partition, River, can cample, Union returns direct of (K,V) returns grade Combined Seme to Combined in map reduce(). - Comparable to Rops in Core - Comparable to Rops in Core named columns. - are Crested of existing structor existing structor - are Crested of existing structor schema.	7			
	network		•	-11.4
	Map. flatte	en map	2	Lection Toin.
A company of the comp	Narpon Transformation Wide Tyensformation Narpon Transformation Wide Tyensformation - each One each - many inpot output inpot outputs pention Partition - Foster Slower - Foster Slower - Not require any data might trequisionally shoffling over cluster shuffling over cluster network. - May flatten may network. - May flatten may network. - May flatten may network. - Anterection, Join, Map partition, River, sample, Union - retorna desert of (kg) way by hely for each key your partition - are safet side combine. Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Combine Comb	esian, repartition		
	sample, ur	lion	- Vedu	ce by lease
		detalet o	fle Va) of i	of Rey Evetons
		Dais, m		pairs.
	Kv.	941 035A	eseted.	odo not map
	() Trail	Comb	ine	W .
				effect to
		map real	See ().	
	· Coout	0+0.	nain abstractio	n (m
	beark	Latatrames.	Spark SOL)	
-	1	0		0
***************************************	Named	columns.	xisting structure	al dete source
	· - are Cy	reuted > .	existing ROD	deficing 9
		. 0	schema.	Oction 5
	Example	tyem all coasts	(00)	entry point
	# to	ti Dize coastros	The day	[SAT COLICE!
	spark =	Spork Session.	builder 1	
		V V		Marin
	or e	aulc= (3) Conti	ext (sc)	

Day/Date deall with deta frome meta data	
· Dataframe Basic Operations	
- Schema - Schema object describing data.	
- print Schema - diplays schema as visual tree	
o- Cache/persist - dtupes-ever of	Col typ
o- Cache/persist o- Calumns (names of o- explain - print) of information	lebug
DataFrame Queries. Peturns new determine determine. - distinct - return new detaframe with different es	tion
- join - 1 dataframe join with another	ementy.
o- limit - o- select -	
0- f; ltey -	
DF. Select ("age")	
DF. where (" ase >21")	
, - Some queries take one or more call	
ageDF = DF. Select (DF. Ege)	
DF. select (Df. name , Df. age +1=)	
OF. Suit (DF. age . dese())	
	-
o- frequent Pattern mining => clf. frequitems.	
- Data can be stored to date source	
- Data can be Btored to clate source. • Boilt in support for JOBC & Parquet file	
· Crexte JOBCTable	•
o Instytinto a saveas Parquet File	
o Save As table	
• Base ROD contains von object.	
oute vad to get underlying reld.	
Row Roos have all standard actions a	nd
	xim