Name: Parth Nandedkar

Date: 10 Feb 2024 Topics: PySpark

Batch : Data Engineering Batch-1

Handwritten Notes:

12 55 mil 16	als) 1/1/14	10	0/02/ 2024
1 see 10 1344 1.	4./		
- Pyspark - Setting up Azure		-	- Soler o M
- Setting up Azure	lab.		
• . 9 . 1		William I.	- Die Bure
Cont. COL		The state of the s	
Sport SQL	4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	122 100	Feature of
- This a	module for sturctured data	processing	
- Sport SQ1	Us Tombonat Spork Con the	at introduces	a new doda
the second second section of the	- 11 01 010		
01.1		- 2 Mirchin	3) His long
Chellenges no n	in partial solution		
A COLUMN TO A COLU	1 20 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1		1 10 1 10 10 10
AFRICA DAMA 311	DINE 341 1 AC MOGG.		Annua (19
Perform ETL to and from	A Datafroine	APT that can po	thim relations
Perform & TL to and from	A Datafrone	APT that can pe	down relation
Perform RTL to and hom Vacalog (Semi) supportunition	A DataFrance A postions on huit-in RD	APT that can perboth esterned dake	How relations
Perform ETL to and from Vacabley (Semi) unformation Source	Date Fin RD	APT that can pe both externel dake	point and Sports
Perform RTL to and from Variog (Semi/Juptahrutry Source) Spark SQL Archi		APT that can pe both esterned dake D.	point and sports
	tecture (mothers)	tought well	UDF (
Spark SQL Archi		tought well	
	Lyton [Scale] Jana Hive	tought well) 7(1U - [ableo =
Spark SQL Archi	Lython Scalar Januar Hive &	tought well	UDF (
Spark SQL Archi	Python Scalar Tournal Hive & Spank SQL Schema R.DD	Use defined	VIDE (
Spark SQL Archi	Python Scalar Tournal Hive & Spank SQL Schema R.DD	tought well) 7(1U - [ableo =
Spark SQL Archi Larguege ARI	Spanle SQL Schema R.DD Data	frame 2 1018	VIDE (
Spark SQL Archi Language ARI	Schema RDD Data	france 2 (108)	VIDE (
Spark SQL Archi Larguse ARI	Spanle SQL Schema R.DD Data	france 2 (108)	VIDE (
Spark SQL Archi Largueze ART	Parquet JSON/HIVE C	france 2 (108)	VIDE (

	HIVE (groy Due 592).
19[02] 2020	(Hive tables are there)
-loomes AT -	Amount -
- Language API Schenza RDO -	- Settler up Aspelat.
- Data Sources -	1 ()
	120
	The mag
Features of Sparle SQL	1 1. dahar a 1 11 - 1
I Tokgrated mil with SDL & ray Alls	1. Hun Scalar and Jaca Con
2) Unified Dato Access (Look and qua	and data from vigorious of sources)
2) Him (and should be	y said sie
- 3) Hive Compatibility 3 To Run to compatibility	ed tive a word on existing when howe
5] S m ballity: a - Spark SQL	the reuser the title front-ind
Use to some croppe for and Meta Sto	or, giving you full competition
Use to same eighe for in and Heta's to both interactive be long to quoies to and the existing to a continuous from	of Mive Hate, quertes & VDFs
- Col > Simply	f Glendy Ma
1000 (10 10 10 1	1.1.1 0 000
- UDF (User defined functions)	Spale SOU Partitations
Tableau - Glidenii Innalii Isl	. S[_8,3]
, Sent of 1180017 11800	Liver Hill
Featuros	The part
- I MA sarks	20. 100
Spark RDD:	8 1 3
- It is foundamental data shucker	e of Spork
- It is immuniable distributed:	collection of objects that can't
- be stored in memory or disk - Each dolaset in RPP is disk	across cluster
Which may be computed an	all node of a land
- Parallel functional transformation	n 1009 G Chiller

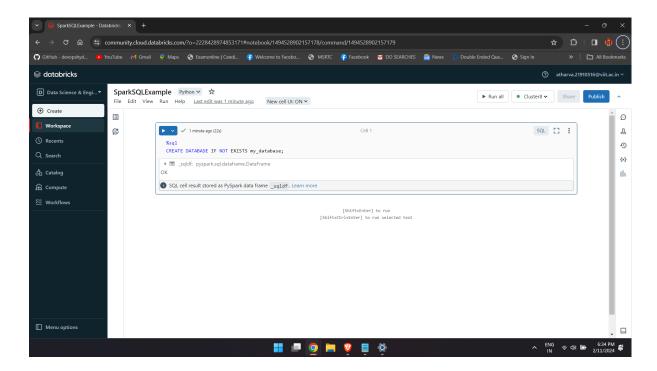
•	- Automatically ochuild Autience
	RDD contain any type Python, Java, Scala objects including-
-	RDD contains any type Python, Javas Scala objects including- wer defined clauses
101	
1	formally, on RDD is read-only, partitioned collection of records.
	amodelate to return
	- RDDs can be created through deterministic operations on either data on
1/2	Stable Blonge for where KOD i also all range of a public -
	On a single rode 7 always to long chiral
1.	- RDD is full -tolerant collection of elements that an be operated as i
1: 14	Storage States (HINS HIVE , POLY STORE)
	Samuel and Charles Court Action
,	Data set & a Mato frame 2 - 2 clos a grander to to date -2
AA T	Pataset & Data frame : solar of reducing to to to date -?
13-	A do distributed collection of Jata, which organized into named column Conceptually, it is required to scientism stables with good optimization technical frame can be constructed from an array of different sources such as five tables, Structured data file, external databases.
9 _ (i)	Conceptually, it is equivalent to relational tables with good applimization technical
E	A Data Frame can be constructed from an array of different sou
	Such as five fables, Structured data filely external databases.
	This API was designed for morden Big date . 10 . 1
750	
7.4	Data from ->
	Data frame -> Data is organized into roused columns, like table in a relational das database.
	relational end database.
_	

1 (8)

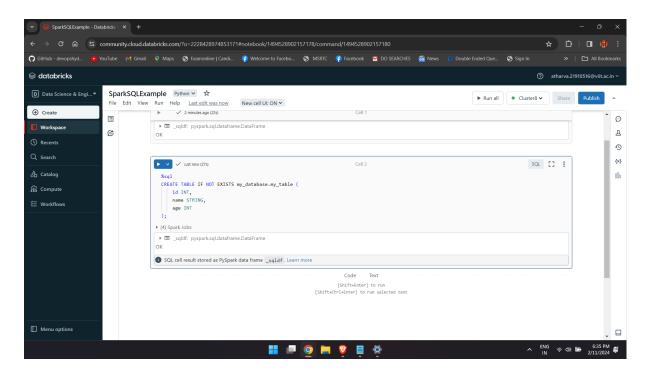
muld blide who	
The orbital Artison	
the orbital Astron	
month bling out	
	- Interestin
Television about Data France at a will some you was and	22 141×1
Whise douch	Will of
	11 2
i an EMD is a cod may suit board collection it counter	Moura! -
heature of Delaborare:	
con in create things defendate countries on sites due co	1(1(7)
- Ability o to spaces the data in the size of kilobytes on a single node s cluster to large cluster: - Supports diff data formates [Auro, csv, classic searchs] storage Systems [HDFS, HIVE, mysql etc)	to Petebytes
on a single nodes cluster to large cluster.	
I will apprece of a coff themselve to whole the tell of	101/148
- Supports diff data formates (Fluxo, csv, elaptic searchs	(Cassenda) and
storage Systems [HDFS, HIVE, Mysql etc)	
	W. H. 124
2 State of art optimization & code so generation through Cotalyte optimizer to be transformation frame work)	the Stark SQL
(Catalyst oftimizer l tree fransformation frame work)	
diplet relieve of total while regard in a named or lander believe	1
Cain her easily allograted with all 1859 date tools.	and framework
- Land to home of the before a day of the with all of Big date books. - Land to the Carry here easily whitegrated with all of Big date books. - Land to the black of beginning and was some a support to the support t	in A
as the table stution of dota file, evenual stabales.	rlau2
- Provides solAPI for Pythoni, Jan, Scala & R proj	granning.
	1 /
a mil iii	plott
The is criminal into meaned columns, like lable in a	1 1 1 1 1 1
- galatet tes	langilotis
††† 	
Till	

· ·	
Spark SQL	
1 337 1	
1. While less code	
2. Read les data	
5.	
Plan Optimization & Execution ?	
Analysis Logical Opinionis President Maning	Code
59 LAST Uprodued logical Ophinized Physical Copical Plan	Selected 185
Plan I I'm	Plan This
L DATA FORMET	
[atalog.]	
	*
	K

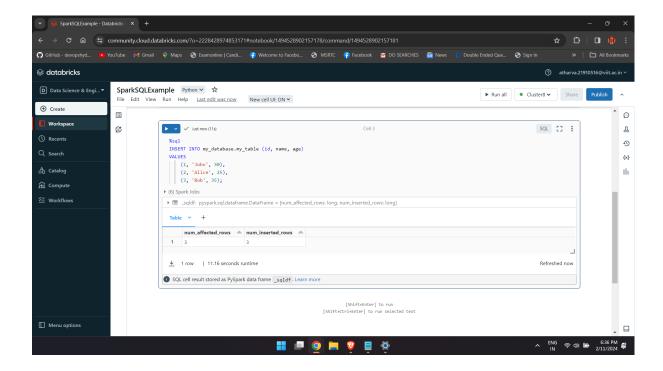
Creating Database using Spark SQL:



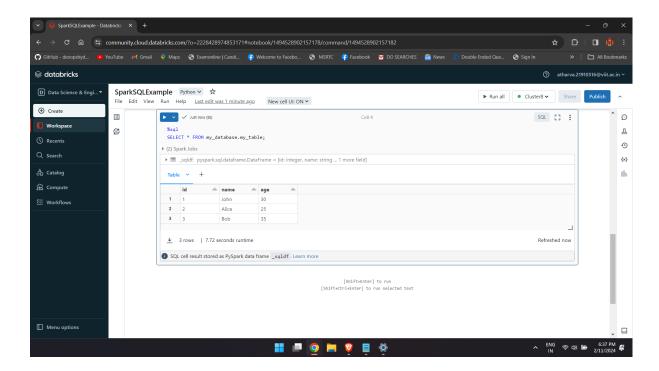
Creating Table using Spark SQL:



Adding data to table:



Fetching Data from table :



Filtering Data from table:

