

# Assignments

1) Perform inner Join operation on two dataframes fetched from CSV file

```
+-----+-----+-----+-----+-----+
|Band_Id|Guitarist_Id|      name|  hometown|      name|year|
+-----+-----+-----+-----+-----+
|      1|          1|  Angus Young|    Sydney|    AC/DC|1973|
|      2|          2|  Eric Clapton|    London|  Led Zeppelin|1968|
|      3|         39|  Aayush Tanuja|Los Angeles|  Metallica|1981|
|      3|          3|  Kirk Hammett|Los Angeles|  Metallica|1981|
|      5|         16|    Abhi Yadav|   Chennai|  The Metals|1964|
|      6|         20|  Kalpana Rathod|   Mumbai|The Rock band|1980|
|      8|         18|    Akash Gupta|    Delhi|  Magicians|1954|
|     12|        34|Neelam Choudhary|Chandigarh|  Party Hunks|1963|
|     13|        25|    Raju Irani|  Mangalore|  Superstars|1968|
|     13|        11|  Antony DSouza|  Mangalore|  Superstars|1968|
|     15|        14|    Karan Sharma|    Surat|    Ginny|1973|
|     20|          6|  Bucky Roberts|    Berlin|Melody Kings|1983|
|     23|        31|  Jimmy DCosta|  Bangalore|  Killer Rock|1980|
|     23|          9|    Rob Bucky|  Bangalore|  Killer Rock|1980|
|     24|        13|    Rahul Verma|Hyderabad|  Great Rock|1977|
|     25|        21|    Sheha Patil|   Chennai|   Bandits|1964|
|     26|        37|  Adithya Singh|   Kochin|Music Rulers|1962|
|     28|        27|    Rakesh Joshi|    Delhi|Gang Killers|1973|
|     32|          4|Stuarts Little|Hyderabad|  Band Name|1968|
|     35|          7|    Tim Rock|  Mangalore|  Classy Magic|1972|
+-----+-----+-----+-----+-----+
only showing top 20 rows
```

2) Read JSON file and parse the dataframe to display the data is expected format

```
20/09/07 12:26:29 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
20/09/07 12:26:29 INFO DAGScheduler: ResultStage 1 (show at ParseCascadedJson.scala:32) finished in 0.876 s
20/09/07 12:26:29 INFO DAGScheduler: Job 1 is finished. Cancelling potential speculative or zombie tasks for this job
20/09/07 12:26:29 INFO TaskSchedulerImpl: Killing all running tasks in stage 1: Stage finished
20/09/07 12:26:29 INFO DAGScheduler: Job 1 finished: show at ParseCascadedJson.scala:32, took 0.897078 s
20/09/07 12:26:29 INFO CodeGenerator: Code generated in 29.983001 ms
+-----+-----+-----+-----+-----+
|eventName|processName|hostname|filePath|dataFileName|
+-----+-----+-----+-----+-----+
|S2F_BWKL_RCR_CURRENTACTION|S2F_BWKL_RCR_CURRENTACTION|RN-EDWETLT-LAPP140|NFSRENO/STORM/CSX_CASE/06|bwkl_it_rcr_current_action_909051530.out.gz|
|S2F_BWKL_RCR_CURRENTACTION|S2F_BWKL_RCR_CURRENTACTION|RN-EDWETLT-LAPP140|NFSRENO/STORM/CSX_CASE/06|bwkl2_it_rcr_current_action_909051530.out.gz|
+-----+-----+-----+-----+-----+
20/09/07 12:26:29 INFO SparkContext: Invoking stop() from shutdown hook
```

### 3) Fetch Data from CSV file and dump it to a MySQL database

```
MariaDB [(none)]> use spark_scala
Database changed
MariaDB [spark_scala]> show tables;
+-----+
| Tables_in_spark_scala |
+-----+
| insurance              |
+-----+
1 row in set (0.00 sec)

MariaDB [spark_scala]> select * from insurance;
```

policyID	statecode	county	eq_site_limit	hu_site_limit	fl_site_limit	fr_site_limit	tiv_2011	tiv_2012	eq_site_deductible	hu_site_de
ductible	fl_site_deductible	fr_site_deductible	point_latitude	point_longitude	line	construction	point_granularity			
119736	FL	CLAY COUNTY	498960	498960	498960	498960	498960	792148.9	0	
9979.2		0	0	30.102261	-81.711777	Residential	Masonry		1	
448094	FL	CLAY COUNTY	1322376.3	1322376.3	1322376.3	1322376.3	1322376.3	1438163.57	0	
0		0	0	30.063936	-81.707664	Residential	Masonry	3		
206893	FL	CLAY COUNTY	190724.4	190724.4	190724.4	190724.4	190724.4	192476.78	0	
0		0	0	30.089579	-81.700455	Residential	Wood	1		
333743	FL	CLAY COUNTY	0	0	79520.76	0	79520.76	86854.48	0	
0		0	0	30.063236	-81.707703	Residential	Wood	3		
172534	FL	CLAY COUNTY	0	0	254281.5	0	254281.5	246144.49	0	
0		0	0	30.060614	-81.702675	Residential	Wood	1		
785275	FL	CLAY COUNTY	0	0	515035.62	0	515035.62	884419.17	0	
0		0	0	30.063236	-81.707703	Residential	Masonry	3		
995932	FL	CLAY COUNTY	0	0	19260000	0	19260000	20610000	0	
0		0	0	30.102226	-81.713882	Commercial	Reinforced Concrete	1		
223488	FL	CLAY COUNTY	328500	328500	328500	328500	328500	348374.25	0	

```
root
|-- policyID: integer (nullable = true)
|-- statecode: string (nullable = true)
|-- county: string (nullable = true)
|-- eq_site_limit: double (nullable = true)
|-- hu_site_limit: double (nullable = true)
|-- fl_site_limit: double (nullable = true)
|-- fr_site_limit: double (nullable = true)
|-- tiv_2011: double (nullable = true)
|-- tiv_2012: double (nullable = true)
|-- eq_site_deductible: double (nullable = true)
|-- hu_site_deductible: double (nullable = true)
|-- fl_site_deductible: double (nullable = true)
|-- fr_site_deductible: integer (nullable = true)
|-- point_latitude: double (nullable = true)
|-- point_longitude: double (nullable = true)
|-- line: string (nullable = true)
|-- construction: string (nullable = true)
|-- point_granularity: integer (nullable = true)

20/09/07 12:38:59 INFO FileSourceStrategy: Pruning directories with:
20/09/07 12:38:59 INFO FileSourceStrategy: Pushed Filters:
20/09/07 12:38:59 INFO FileSourceStrategy: Post-Scan Filters:
20/09/07 12:38:59 INFO FileSourceStrategy: Output Data Schema: struct<policyID: int, statecode: string, county: string, eq_site_limit: double,
hu_site_limit: double ... 16 more fields>
20/09/07 12:38:59 INFO MemoryStore: Block broadcast_6 stored as values in memory (estimated size 170.6 KiB, free 1036.2 MiB)
```

#### 4) Read file from HDFS and write the data to a file in Local file system

```
C:\WINDOWS\system32>cd ../../

C:\>hadoop fs -put files.csv /
2020-09-07 13:00:17,004 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false

C:\>hadoop fs -cat /files.csv
2020-09-07 13:00:43,399 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
band,id,name
1,1,Angus Young
2,2,Eric Clapton
3,3,Kirk Hammett
32,4,Stuarts Little
56,5,Benny Jobs
20,6,Bucky Roberts
35,7,Tim Rock
63,8,Dwey John
23,9,Rob Bucky
45,10,Jimmy John
13,11,Antony DSouza
C:\>
```






#### 5) Scala Code to perform GPG Encryption

```
File 'src/main/resources\data\gpgtest.txt' Being Encrypted
20/09/11 12:01:01 WARN ProcfsMetricsGetter: Exception when trying to compute pagesize, as a result reporting of ProcessTree metrics is stopped
Encryption successful !
Output File : src/main/resources\data\gpgtest.txt.gpg
Original file deleted !
File src/main/resources\data\gpgtest.txt.gpg Being Decrypted
gpg: WARNING: no command supplied. Trying to guess what you mean ...
gpg: AES.CFB encrypted data
gpg: encrypted with 1 passphrase
Decrypted Successfully
Decryption is completed ! Output File : src/main/resources\data\gpgtest.txt
```

#### 6) Find .txt files older than 7 days

```
C:\Java\jdk-11.0.8\bin\java.exe ...
File answers.txt is older than 7 Days (44 days)
File name.txt is older than 7 Days (44 days)
File newfile.txt is New (0 days)

Process finished with exit code 0
```

 answers.txt	7/24/2020 1:44 PM	Text Document	35 KB
 bands.csv	9/4/2020 7:35 PM	Microsoft Excel C...	2 KB
 name.txt	7/24/2020 1:44 PM	Text Document	24 KB
 newcsv.csv	9/7/2020 9:14 AM	Microsoft Excel C...	0 KB
 newfile.txt	9/7/2020 9:01 AM	Text Document	0 KB

7) Get a Hashmap of last modified time for each file

```
C:\Java\jdk-11.0.8\bin\java.exe ...
Map(name.txt -> Fri Jul 24 13:44:13 IST 2020, bands.csv -> Fri Sep 04 19:35:13 IST 2020,
  newcsv.csv -> Mon Sep 07 09:14:12 IST 2020, answers.txt -> Fri Jul 24 13:44:13 IST 2020,
  newfile.txt -> Mon Sep 07 09:01:14 IST 2020)

Process finished with exit code 0
```