Blacklisted Users

Loading data from s3

```
var msg_gear = spark.read.parquet("s3://ha-prod-omnidata-us-east-1/marketing/email/ocelot/temp/training/input/message_gear")
```

```
msg_gear: org.apache.spark.sql.DataFrame = [email_hash: string, activity_type: string ... 2 more fields]
```

Viewing the loaded file

```
msg_gear.show(60,false)
```

```
+------------------------------------------------------------------+-------------+-----------------+-------------+
|email_hash                                                        |activity_type|activity_timestamp|category_code|
+------------------------------------------------------------------+-------------+-----------------+-------------+
|029357de2dd9c1a0db3562857fc48bdc94ab621ab9e2eef2c1c55e21937eb3f3|click        |null             |0            |
|0c00c5cd05f00698e276290d55e2dd32e42d26c9eb6438cde8a1deb4f8e03fe1|click        |null             |0            |
|1a096afddc5a4843073ce74ae91ee5eeddf7846a340b551faf8d16b0b19eff6d|click        |null             |0            |
|3548733db8061003bfd860f9694458d6c35a3638722809ea1b1d4a0db07afc77|click        |null             |0            |
|64e55417b38baf52f098f37300d59401a482bd7fad374fb865d29d7a972c3b3f|click        |null             |0            |
|798a53a38c1843b7a9716a2a4c39c2d7b5d447d28bf0eb8ae9668eccf3f82ad8|click        |null             |0            |
|7f5cf444f6b3925c30d3835d1452768088bae773b6c39752db3dc3ffca469468|click        |null             |0            |
|83352af54b70f29734720cd7bb9f4c1385704a495a59c3c3d5be3fdcf93ec95e|click        |null             |0            |
|87d0e3f5c6d8d3aa06a72dabf00da5156454a35eddefa70c7c44874b66763e23|click        |null             |0            |
|8f712990d65ade1424168d174fd11cc0d6efbe32f6539965db57c254873cd301|click        |null             |0            |
|9b57ddfd19c30fe2776dcf7c465cd4ee16755127d8db99d18d4c14ce58b71763|click        |null             |0            |
|a4483508b7237a1af67073e518a86deaed36e449c141aaa650ffa43a607d82e3|click        |null             |0            |
|beaee2edcc578a3ffe31dc9d27821ac8a7d6e1b6a14499a73f3bbd891212d1ed|click        |null             |0            |
|e75232a89c1ff6e80fdce3979be3031402cce89717e2ad1e69c5e6fb16ce486a|click        |null             |0            |
|ed94100228889360da97d8e78de0bdb3c263329f9eb22821f2fc4f48d151f5ab|click        |null             |0            |
```

Checking the schema and count

```
msg_gear.schema
msg_gear.count
```

```
res331: org.apache.spark.sql.types.StructType = StructType(StructField(email_hash,StringType,true), StructField(activity_type,StringType,true), StructField(activity_timestamp,TimestampType,true), StructField(category_code,IntegerType,true))
res332: Long = 1000000
```

Loading Data from s3

```
var sys_email_act = spark.read.parquet("s3://ha-prod-omnidata-us-east-1/marketing/email/ocelot/temp/training/input/system_email")
```

```
sys_email_act: org.apache.spark.sql.DataFrame = [email_hash: string, activity_type: string ... 2 more fields]
```

Viewing the loaded file

```
sys_email_act.show()
```

```
+--------------------+-------------+------------------+-------------+
|          email_hash|activity_type| activity_timestamp|category_code|
+--------------------+-------------+------------------+-------------+
|0007b3d7c2004b452...|      bounced|2021-04-04 18:23:47|           10|
|002f49351c5a11f51...|      bounced|2020-09-11 04:13:41|           25|
|00513323fb43ab877...|      bounced|2022-04-24 14:09:49|           21|
|006a201e1e1e9fcf6...|      bounced|2020-09-11 04:13:08|           25|
|007e51bad88701a2e...|      bounced|2020-09-09 17:37:37|           10|
|00835ba0f2df434bc...|      bounced|2022-04-03 14:06:48|           10|
|008ad1cf7f42f9d35...|      bounced|2020-09-11 04:12:40|           25|
|009ce6cef2e33b52b...|      bounced|2021-01-09 14:56:52|           10|
|00a3c8c3746fc008f...|      bounced|2021-01-28 17:12:06|           10|
|00d83ca1f4b4a1d2a...|      bounced|2021-12-14 17:00:26|           10|
|00ded3bfca8b8a7e3...|      bounced|2021-08-11 14:08:35|           10|
|00e0f2b6040e2b4dd...|      bounced|2020-08-05 15:18:36|           10|
|00f4a4f304774273...|      bounced|2020-12-13 17:30:05|           10|
|00f8e761b00f2a399...|      bounced|2020-11-24 14:56:30|           10|
|010c2476a0d2a850e...|      bounced|2022-03-20 13:50:08|           10|
```

Checking the schema and count

```
sys_email_act.schema
sys_email_act.count
```

```
res337: org.apache.spark.sql.types.StructType = StructType(StructField(email_hash,StringType,true), StructField(activity_type,StringType,true), StructField(activity_timestamp,TimestampType,true), StructField(category_code,IntegerType,true))
res338: Long = 500000
```

Union of message gear and system email activity

```
var uni = msg_gear.unionAll(sys_email_act)
```

```
warning: there was one deprecation warning; re-run with -deprecation for details
uni: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [email_hash: string, activity_type: string ... 2 more fields]
```

Viewing the data after union

```
uni.show()
```

```
+--------------------+-------------+-----------------+-------------+
|          email_hash|activity_type|activity_timestamp|category_code|
+--------------------+-------------+-----------------+-------------+
|029357de2dd9c1a0d...|        click|             null|            0|
|0c00c5cd05f00698e...|        click|             null|            0|
|1a096afddc5a48430...|        click|             null|            0|
|3548733db8061003b...|        click|             null|            0|
|64e55417b38baf52f...|        click|             null|            0|
|798a53a38c1843b7a...|        click|             null|            0|
|7f5cf444f6b3925c3...|        click|             null|            0|
|83352af54b70f2973...|        click|             null|            0|
|87d0e3f5c6d8d3aa0...|        click|             null|            0|
|8f712990d65ade142...|        click|             null|            0|
|9b57ddfd19c30fe27...|        click|             null|            0|
|a4483508b7237a1af...|        click|             null|            0|
|beaee2edcc578a3ff...|        click|             null|            0|
|e75232a89c1ff6e80...|        click|             null|            0|
|ed94100228889360da...|        click|             null|            0|
```

validating the data after union

```
uni.schema
uni.count
uni.intersect(msg_gear).count
uni.intersect(sys_email_act).count
```

```
res343: org.apache.spark.sql.types.StructType = StructType(StructField(email_hash,StringType,true), StructField(activity_type,StringType,true), StructField(activity_timestamp,TimestampType,true), StructField(category_code,IntegerType,true))
res344: Long = 1500000
res345: Long = 1000000
res346: Long = 500000
```

Blacklisted Users

```
//import org.apache.spark.sql.functions._
//var latest = uni.groupBy("email_hash").agg(max("activity_timestamp")).show(false)
```

```
import org.apache.spark.sql.functions._
+----------------------------------------------------------------+----------------------+
|email_hash                                                      |max(activity_timestamp)|
+----------------------------------------------------------------+----------------------+
|0000423e0ca72eb4ade676c82cf7c78fa24ef0abd4b767c14e592b3f997f281e|2022-05-22 19:47:12    |
|0000c1b88abc435879e55304090966b6d625340c875ea0db7517a0804eb738fb|2022-01-05 20:13:19    |
|0001d47f537d205149e3bf4a1b2e819c194feec0f23b6062e34cdc95a8e538d8|2022-05-25 04:55:33    |
|00026b34df9c37b068be3fb1a579316178a067a4f203a571693b1331ba80358b|2022-05-15 11:39:19    |
|0003624e9335623882fc59c700d5ec7d922b85ca4909b3efef752d4ae1bae9dc|2021-07-11 18:09:22    |
|0003fc856b996cf493b5c5128a5f86e4513b66a565dcbaf92b44a4d89eacf914|2021-04-25 05:06:53    |
|000711a90092ea3efb49da86849b6eb2a14af8c1b2ec5924f396d0afc343c036|2022-01-16 19:57:36    |
|0007cbd6325edea024e2ded012a4ea2f02ee39eb68fa9249a48f25760b9e00c9|2021-07-04 13:47:43    |
|00088915a7dba05ee1856e30e25c58e5e54d27dc91aa15097299f3a1d4ea3c9d|2022-05-18 05:49:01    |
|0009958bbea8c260fed1c849bfa17f1f79ba9d293ad5ac0ccdf14f91508ee9d6|2021-12-30 04:57:53    |
|000ac9c72742a063af4ae687b1a97a5b833ce8563a38242623bf4aa17aeb5fda|2022-02-09 19:19:28    |
|000c76ebdbbc43f7d8d3f7f473a2671be316ef1859b3f9a894268069f2e10005|2022-05-18 20:42:47    |
|000d55d8310f3898480465e17af66b86f58fc972c93d82bd7d01b218390a8da9|2022-04-19 21:00:28    |
|001055f933b991e974bfa415e9b8620cd7370fa8e40a45293b9c4bd2f8a8e7eb|2021-07-15 12:26:53    |
```

To show the data with latest activity timestamp

```
import org.apache.spark.sql.functions._
import org.apache.spark.sql.expressions.Window

var lat=Window.partitionBy("email_hash").orderBy(col("activity_timestamp").desc)
var latest= uni.withColumn("row_number",row_number.over(lat)).where($"row_number"===1).drop("row")

latest.show()
```

```
import org.apache.spark.sql.functions._
import org.apache.spark.sql.expressions.Window
lat: org.apache.spark.sql.expressions.WindowSpec = org.apache.spark.sql.expressions.WindowSpec@583ba38
latest: org.apache.spark.sql.DataFrame = [email_hash: string, activity_type: string ... 3 more fields]
+--------------------+-------------+-------------------+-------------+----------+
|          email_hash|activity_type| activity_timestamp|category_code|row_number|
+--------------------+-------------+-------------------+-------------+----------+
|0000423e0ca72eb4a...|    delivered|2022-05-22 19:47:12|            0|         1|
|0000c1b88abc43587...|    delivered|2022-01-05 20:13:19|            0|         1|
|0001d47f537d20514...|         open|2022-05-25 04:55:33|            0|         1|
|00026b34df9c37b06...|         open|2022-05-15 11:39:19|            0|         1|
|0003624e933562388...|    delivered|2021-07-11 18:09:22|            0|         1|
|0003fc856b996cf49...|    delivered|2021-04-25 05:06:53|            0|         1|
|000711a90092ea3ef...|    delivered|2022-01-16 19:57:36|            0|         1|
|0007cbd6325edea02...|    delivered|2021-07-04 13:47:43|            0|         1|
|00088915a7dba05ee...|    delivered|2022-05-18 05:49:01|            0|         1|
|0009958bbea8c260f...|    delivered|2021-12-30 04:57:53|            0|         1|
|000ac9c72742a063a...|    delivered|2022-02-09 19:19:28|            0|         1|
```

Creating an UDF

```
val is_blacklisted = udf((ip:Int) => {
    ip == 10 ||ip == 21 || ip == 30 || ip == 110})
var final_data = latest.withColumn("is_blacklisted", is_blacklisted(col("category_code")))
```

```
is_blacklisted: org.apache.spark.sql.expressions.UserDefinedFunction = UserDefinedFunction(<function1>,BooleanType,Some(List(IntegerType)))
final_data: org.apache.spark.sql.DataFrame = [email_hash: string, activity_type: string ... 4 more fields]
```

```
//checking data and count
final_data.show()
final_data.count
```

```
+--------------------+-------------+-------------------+-------------+----------+--------------+
|          email_hash|activity_type| activity_timestamp|category_code|row_number|is_blacklisted|
+--------------------+-------------+-------------------+-------------+----------+--------------+
|0000423e0ca72eb4a...|    delivered|2022-05-22 19:47:12|            0|         1|         false|
|0000c1b88abc43587...|    delivered|2022-01-05 20:13:19|            0|         1|         false|
|0001d47f537d20514...|         open|2022-05-25 04:55:33|            0|         1|         false|
|00026b34df9c37b06...|         open|2022-05-15 11:39:19|            0|         1|         false|
|0003624e933562388...|    delivered|2021-07-11 18:09:22|            0|         1|         false|
|0003fc856b996cf49...|    delivered|2021-04-25 05:06:53|            0|         1|         false|
|000711a90092ea3ef...|    delivered|2022-01-16 19:57:36|            0|         1|         false|
|0007cbd6325edea02...|    delivered|2021-07-04 13:47:43|            0|         1|         false|
|00088915a7dba05ee...|    delivered|2022-05-18 05:49:01|            0|         1|         false|
|0009958bbea8c260f...|    delivered|2021-12-30 04:57:53|            0|         1|         false|
|000ac9c72742a063a...|    delivered|2022-02-09 19:19:28|            0|         1|         false|
|000c76ebdbbc43f7d...|    delivered|2022-05-18 20:42:47|            0|         1|         false|
|000d55d8310f38984...|    delivered|2022-04-19 21:00:28|            0|         1|         false|
|001055f933b991e97...|         open|2021-07-15 12:26:53|            0|         1|         false|
```

Writing data into blacklisted

```
final_data.write.parquet("s3://ha-prod-omnidata-us-east-1/marketing/email/ocelot/temp/training/output/premnivas/blacklisted2")
```

```
// reading the uploaded file
var final_data = spark.read.parquet("s3://ha-prod-omnidata-us-east-1/marketing/email/ocelot/temp/training/output/premnivas/blacklisted2")
```

```
final_data: org.apache.spark.sql.DataFrame = [email_hash: string, activity_type: string ... 4 more fields]
```

```
// checking the schema and count
final_data.printSchema
final_data.count
final_data.show()
```

```
root
 |-- email_hash: string (nullable = true)
 |-- activity_type: string (nullable = true)
 |-- activity_timestamp: timestamp (nullable = true)
 |-- category_code: integer (nullable = true)
 |-- row_number: integer (nullable = true)
 |-- is_blacklisted: boolean (nullable = true)
res374: Long = 1139724
+--------------------+-------------+-------------------+-------------+----------+--------------+
|          email_hash|activity_type| activity_timestamp|category_code|row_number|is_blacklisted|
+--------------------+-------------+-------------------+-------------+----------+--------------+
|00008c41d735a9649...|    delivered|2021-05-12 21:10:33|            0|         1|         false|
|0001455ed37382646...|    delivered|2021-03-29 04:19:34|            0|         1|         false|
|000315734291d5d3c...|    delivered|2022-02-14 18:04:18|            0|         1|         false|
|00039804ec7882c1b...|    delivered|2022-04-28 17:10:32|            0|         1|         false|
|0003abea1e1b54f0b...|    delivered|2021-03-18 18:56:19|            0|         1|         false|
|00054d425d2b529ef...|    delivered|2022-04-03 20:18:41|            0|         1|         false|
```

```
// counting true and false value in the table
final_data.groupBy("is_blacklisted").count().show()
```

Blacklisted Users

```
+--------------+------+
|is_blacklisted| count|
+--------------+------+
|          true|340289|
|         false|799435|
+--------------+------+
```