Pyspark Problems based on realtime SQL Questions

SWIPE

Dhanushkumar Palani

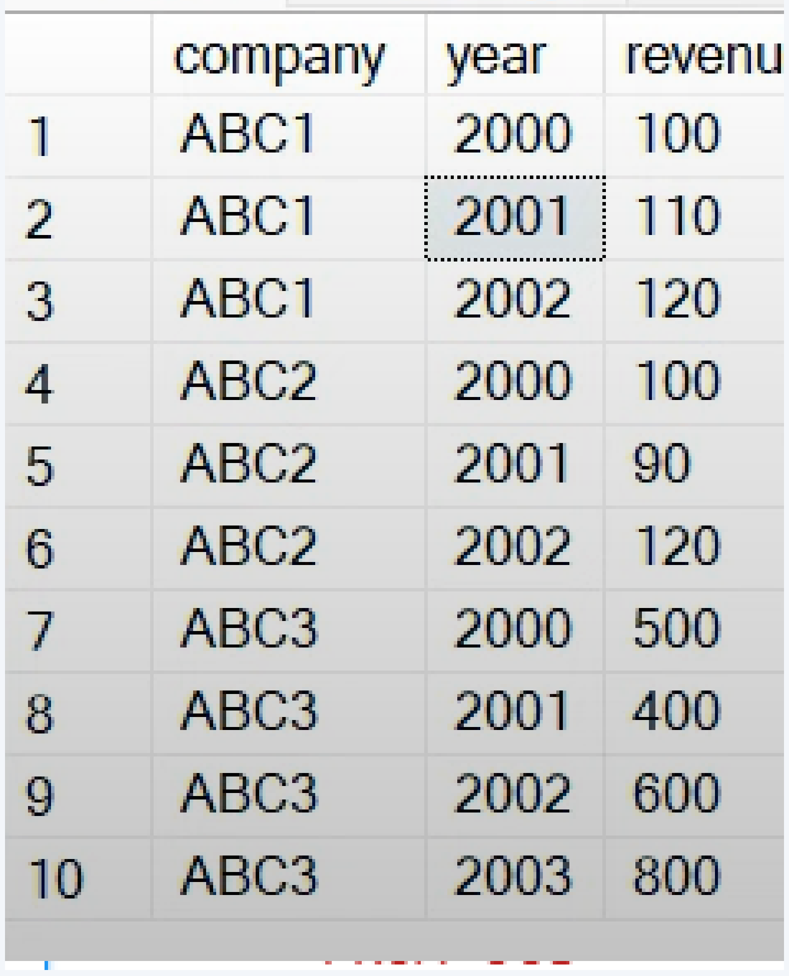
@dhanushpalani1120@gmail.com

PROBLEM STATEMENT

FIND THE ONLY COMPANY WHOSE REVENUE IS INCREASING EVERY YEAR.

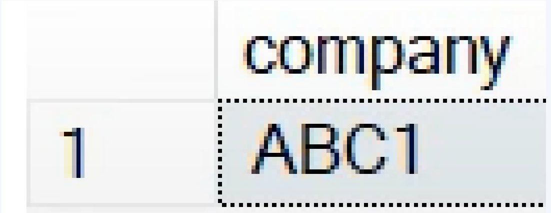
suppose a company's revenue is increasing fro 3 years and
a very next year revenue is dipped,
in that case it should not come in output

Sample Input



	company	year	revenu
1	ABC1	2000	100
2	ABC1	2001	110
3	ABC1	2002	120
4	ABC2	2000	100
5	ABC2	2001	90
6	ABC2	2002	120
7	ABC3	2000	500
8	ABC3	2001	400
9	ABC3	2002	600
10	ABC3	2003	800

Sample ouput



	company
1	ABC1

KEEP SWIPING

DATAFRAME CREATION

```
from pyspark.sql.types import StructType, StructField, StringType, IntegerType
from pyspark.sql import SparkSession
import getpass
username=getpass.getuser()
spark=SparkSession.\
builder.\
config('spark.ui.port','0').\
config("spark.sql.warehouse.dir",f"/user/itv008777/warehouse").\
enableHiveSupport().\
master('yarn').\
getOrCreate()
```

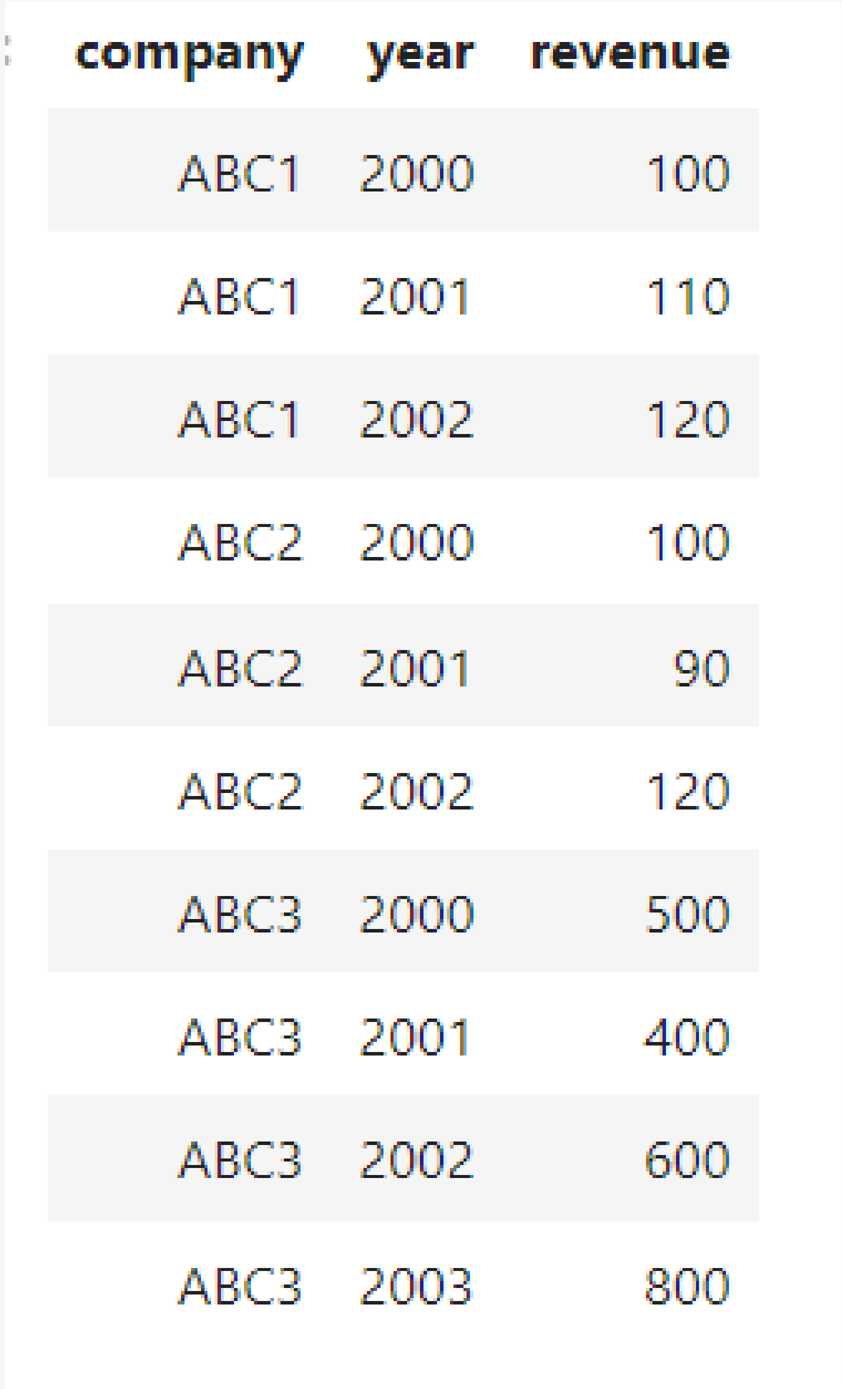
SCHEMA & DATA DEFINITION:

```
schema = StructType([
            StructField("company", StringType(), True),
            StructField("year", IntegerType(), True),
            StructField("revenue", IntegerType(), True)
         ])
```

KEEP SWIPING →

```
data = [("ABC1", 2000, 100),
            ("ABC1", 2001, 110),
            ("ABC1", 2002, 120),
            ("ABC2", 2000, 100),
            ("ABC2", 2001, 90),
             ("ABC2", 2002, 120),
             ("ABC3", 2000, 500),
             ("ABC3", 2001, 400),
             ("ABC3", 2002, 600),
             ("ABC3", 2003, 800)]

df = spark.createDataFrame(data,schema)
```



company	year	revenue
ABC1	2000	100
ABC1	2001	110
ABC1	2002	120
ABC2	2000	100
ABC2	2001	90
ABC2	2002	120
ABC3	2000	500
ABC3	2001	400
ABC3	2002	600
ABC3	2003	800

KEEP SWIPING

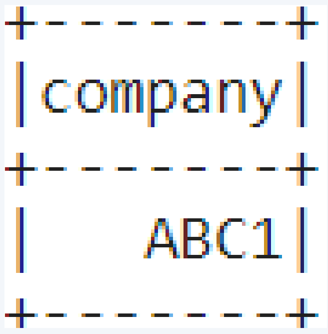
PYSPARK CODE

IN DATAFRAME

```
 window= Window()\
    .partitionBy("company")\.
     .orderBy("year")

result_df = df.withColumn("flag",
             F.col("revenue") - F.lag("revenue", 1, 0).
                  over(window)) \
         .groupBy("company") \
         .agg(F.min("flag").alias("min_flag")) \
          .filter("min_flag >= 0")
            .select("company") \
                  .distinct()


result_df.show()
```



```
+--------+
| company|
+--------+
|    ABC1|
+--------+
```

"MORE TO COME! SHARE YOUR INSIGHTS IN THE COMMENTS."

THANK YOU !