In [1]:

```
import pandas as pd
from pyspark.sql import SparkSession
import findspark
from pyspark.sql import Row
```

In [2]:

```
findspark.init()
findspark.find()
```

Out[2]:

```
'/Users/youngjinseo/anaconda3/lib/python3.10/site-packages/pyspark'
```

In [3]:

```
spark = SparkSession.builder.appName('practice1').config("spark.driver.bindAddress", "127.0.0.1").getOrCreate()
spark
```

```
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/09/25 21:53:22 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

Out[3]: **SparkSession - in-memory**

SparkContext

Spark UI

Version	`v3.5.2`
Master	`local[*]`
AppName	`practice1`

Df1, Df2 데이터프레임 만들기

In [4]:

```
df1 = [
    Row(id=1, name="John", age=28, city="New York", salary=5000),
    Row(id=2, name="Jane", age=35, city="Los Angeles", salary=6000),
    Row(id=3, name="Michael", age=42, city="Chicago", salary=7000),
    Row(id=4, name="Chris", age=31, city="New York", salary=4000)
]

df2 = [
    Row(id=3, name="Michael", age=42, city="Chicago", salary=7000),
    Row(id=4, name="Chris", age=31, city="Boston", salary=4500),
    Row(id=5, name="Jessica", age=25, city="Los Angeles", salary=3500),
    Row(id=6, name="David", age=29, city="Miami", salary=5500)
]
```

In [5]:

```
# 두 개의 DataFrame 생성
df1 = spark.createDataFrame(df1)
df2 = spark.createDataFrame(df2)
```

In [6]:

```
df1.show()
```

```
+---+-------+---+-----------+------+
| id|   name|age|       city|salary|
+---+-------+---+-----------+------+
|  1|   John| 28|   New York|  5000|
|  2|   Jane| 35|Los Angeles|  6000|
|  3|Michael| 42|    Chicago|  7000|
|  4|  Chris| 31|   New York|  4000|
+---+-------+---+-----------+------+
```

In [7]:

```
df2.show()
```

```
+---+-------+---+-----------+------+
| id|   name|age|       city|salary|
+---+-------+---+-----------+------+
|  3|Michael| 42|    Chicago|  7000|
|  4|  Chris| 31|     Boston|  4500|
|  5|Jessica| 25|Los Angeles|  3500|
|  6|  David| 29|      Miami|  5500|
+---+-------+---+-----------+------+
```

groupby (Max, Min, Count, Sum, Avg)

In [10]:

```
df1.groupBy("city").max("salary").show()
df1.groupBy("city").min("salary").show()
df1.groupBy("city").sum("salary").show()
df1.groupBy("city").avg("salary").show()
```

```
+-----------+-----------+
|       city|max(salary)|
+-----------+-----------+
|   New York|       5000|
|Los Angeles|       6000|
|    Chicago|       7000|
+-----------+-----------+

+-----------+-----------+
|       city|min(salary)|
+-----------+-----------+
|   New York|       4000|
|Los Angeles|       6000|
|    Chicago|       7000|
+-----------+-----------+

+-----------+-----------+
|       city|sum(salary)|
+-----------+-----------+
|   New York|       9000|
|Los Angeles|       6000|
|    Chicago|       7000|
+-----------+-----------+

+-----------+-----------+
|       city|avg(salary)|
+-----------+-----------+
|   New York|     4500.0|
|Los Angeles|     6000.0|
|    Chicago|     7000.0|
+-----------+-----------+
```

join

In [12]:

```
join_df = df1.join(df2,df1.id == df2.id,'outer')
join_df.show()
```

```
+----+-------+----+-----------+------+----+-------+----+-----------+------+
|  id|   name| age|       city|salary|  id|   name| age|       city|salary|
+----+-------+----+-----------+------+----+-------+----+-----------+------+
|   1|   John|  28|   New York|  5000|NULL|   NULL|NULL|       NULL|  NULL|
|   2|   Jane|  35|Los Angeles|  6000|NULL|   NULL|NULL|       NULL|  NULL|
|   3|Michael|  42|    Chicago|  7000|   3|Michael|  42|    Chicago|  7000|
|   4|  Chris|  31|   New York|  4000|   4|  Chris|  31|     Boston|  4500|
|NULL|   NULL|NULL|       NULL|  NULL|   5|Jessica|  25|Los Angeles|  3500|
|NULL|   NULL|NULL|       NULL|  NULL|   6|  David|  29|      Miami|  5500|
+----+-------+----+-----------+------+----+-------+----+-----------+------+
```

In [13]:

```
join_df = df1.join(df2,df1.id == df2.id,'inner')
join_df.show()
```

```
+---+-------+---+--------+------+---+-------+---+-------+------+
| id|   name|age|    city|salary| id|   name|age|   city|salary|
+---+-------+---+--------+------+---+-------+---+-------+------+
|  3|Michael| 42| Chicago|  7000|  3|Michael| 42|Chicago|  7000|
|  4|  Chris| 31|New York|  4000|  4|  Chris| 31| Boston|  4500|
+---+-------+---+--------+------+---+-------+---+-------+------+
```

union

In [14]:

```
union_df = df1.union(df2)
union_df.show()
```

```
+---+-------+---+-----------+------+
| id|   name|age|       city|salary|
+---+-------+---+-----------+------+
|  1|   John| 28|   New York|  5000|
|  2|   Jane| 35|Los Angeles|  6000|
|  3|Michael| 42|    Chicago|  7000|
|  4|  Chris| 31|   New York|  4000|
|  3|Michael| 42|    Chicago|  7000|
|  4|  Chris| 31|     Boston|  4500|
|  5|Jessica| 25|Los Angeles|  3500|
|  6|  David| 29|      Miami|  5500|
+---+-------+---+-----------+------+
```

intersect

In [15]:

```
intersect_df = df1.intersect(df2)
intersect_df.show()
```

```
+---+-------+---+-------+------+
| id|   name|age|   city|salary|
+---+-------+---+-------+------+
|  3|Michael| 42|Chicago|  7000|
+---+-------+---+-------+------+
```

In [16]:

```
df1.show()
```

```
+---+-------+---+-----------+------+
| id|   name|age|       city|salary|
+---+-------+---+-----------+------+
|  1|   John| 28|   New York|  5000|
|  2|   Jane| 35|Los Angeles|  6000|
|  3|Michael| 42|    Chicago|  7000|
|  4|  Chris| 31|   New York|  4000|
+---+-------+---+-----------+------+
```

In [17]:

```
df2.show()
```

```
+---+-------+---+-----------+------+
| id|   name|age|       city|salary|
+---+-------+---+-----------+------+
|  3|Michael| 42|    Chicago|  7000|
|  4|  Chris| 31|     Boston|  4500|
|  5|Jessica| 25|Los Angeles|  3500|
|  6|  David| 29|      Miami|  5500|
+---+-------+---+-----------+------+
```

crosstab

In [18]:

```
crosstab_df = df1.crosstab("city","age")
crosstab_df.show()
```

```
+-----------+---+---+---+---+
|   city_age| 28| 31| 35| 42|
+-----------+---+---+---+---+
|Los Angeles|  0|  0|  1|  0|
|    Chicago|  0|  0|  0|  1|
|   New York|  1|  1|  0|  0|
+-----------+---+---+---+---+
```

dtypes & columns

In [20]:

```
dtypes_df = df1.dtypes
print(dtypes_df)
```

```
[('id', 'bigint'), ('name', 'string'), ('age', 'bigint'), ('city', 'string'), ('salary', 'bigint')]
```

In [22]:

```
columns_df = df1.columns
print(columns_df)
```

```
['id', 'name', 'age', 'city', 'salary']
```

Spark 세션 종료

In [23]:

```
spark.stop()
```