

```
In [1]: import findspark
findspark.init()
```

```
In [2]: data = [('James', '', 'Smith', '1994-04-01', 'M', 3000),
                ('Michael', 'Rose', '', '2000-05-19', 'M', 4000),
                ('Robert', '', 'Williams', '1978-09-05', 'F', 4000),
                ('Maria', 'Anne', 'Jones', '1967-12-01', 'F', 4000),
                ('Jen', 'Mary', 'Brown', '1980-02-17', 'F', -1)
                ]

columns = ["firstname", "middlename", "lastname", "dob", "gender", "salary"]

from pyspark.sql import SparkSession
spark = SparkSession.builder.appName('SparkByExamples.com').getOrCreate()
df = spark.createDataFrame(data = data, schema = columns)
```

1. Change DataType using PySpark withColumn()

```
In [3]: from pyspark.sql.functions import col
from pyspark.sql.types import IntegerType
df.withColumn("salary", col("salary").cast("Integer")).show()
```

| firstname | middlename | lastname | dob | gender | salary |
|-----------|------------|----------|------------|--------|--------|
| James | | Smith | 1994-04-01 | M | 3000 |
| Michael | Rose | | 2000-05-19 | M | 4000 |
| Robert | | Williams | 1978-09-05 | F | 4000 |
| Maria | Anne | Jones | 1967-12-01 | F | 4000 |
| Jen | Mary | Brown | 1980-02-17 | F | -1 |

2. Update The Value of an Existing Column

```
In [4]: df.withColumn("salary", col("salary")*100).show()
```

| firstname | middlename | lastname | dob | gender | salary |
|-----------|------------|----------|------------|--------|--------|
| James | | Smith | 1994-04-01 | M | 300000 |
| Michael | Rose | | 2000-05-19 | M | 400000 |
| Robert | | Williams | 1978-09-05 | F | 400000 |
| Maria | Anne | Jones | 1967-12-01 | F | 400000 |
| Jen | Mary | Brown | 1980-02-17 | F | -100 |

3. Create a Column from an Existing

```
In [5]: df.withColumn("CopiedColumn", col("salary")* -1).show()
```

| firstname | middlename | lastname | dob | gender | salary | CopiedColumn |
|-----------|------------|----------|------------|--------|--------|--------------|
| James | | Smith | 1994-04-01 | M | 3000 | -3000 |
| Michael | Rose | | 2000-05-19 | M | 4000 | -4000 |
| Robert | | Williams | 1978-09-05 | F | 4000 | -4000 |
| Maria | Anne | Jones | 1967-12-01 | F | 4000 | -4000 |
| Jen | Mary | Brown | 1980-02-17 | F | -1 | 1 |

4. Add a New Column using withColumn()

```
In [6]: from pyspark.sql.functions import lit

df.withColumn("Country", lit("USA")).show()
df.withColumn("Country", lit("USA")) \
  .withColumn("anotherColumn", lit("anotherValue")) \
  .show()
```

| firstname | lastname | dob | gender | salary | Country |
|-----------|----------|------------|--------|--------|---------|
| James | Smith | 1994-04-01 | M | 3000 | USA |
| Michael | Rose | 2000-05-19 | M | 4000 | USA |
| Robert | Williams | 1978-09-05 | F | 4000 | USA |
| Maria | Jones | 1967-12-01 | F | 4000 | USA |
| Jen | Brown | 1980-02-17 | F | -1 | USA |

| firstname | lastname | dob | gender | salary | Country | anotherColumn |
|-----------|----------|------------|--------|--------|---------|---------------|
| James | Smith | 1994-04-01 | M | 3000 | USA | anotherValue |
| Michael | Rose | 2000-05-19 | M | 4000 | USA | anotherValue |
| Robert | Williams | 1978-09-05 | F | 4000 | USA | anotherValue |
| Maria | Jones | 1967-12-01 | F | 4000 | USA | anotherValue |
| Jen | Brown | 1980-02-17 | F | -1 | USA | anotherValue |

5. Rename Column Name

```
In [7]: df.withColumnRenamed("gender", "sex") \
        .show(truncate = False)
```

| firstname | lastname | dob | sex | salary |
|-----------|----------|------------|-----|--------|
| James | Smith | 1994-04-01 | M | 3000 |
| Michael | Rose | 2000-05-19 | M | 4000 |
| Robert | Williams | 1978-09-05 | F | 4000 |
| Maria | Jones | 1967-12-01 | F | 4000 |
| Jen | Brown | 1980-02-17 | F | -1 |

6. Drop Column From PySpark DataFrame

```
In [8]: df.drop("salary") \
        .show()
```

| firstname | lastname | dob | gender |
|-----------|----------|------------|--------|
| James | Smith | 1994-04-01 | M |
| Michael | Rose | 2000-05-19 | M |
| Robert | Williams | 1978-09-05 | F |
| Maria | Jones | 1967-12-01 | F |
| Jen | Brown | 1980-02-17 | F |

7. PySpark withColumn() Complete Example

```
In [9]: import pyspark
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, lit
from pyspark.sql.types import StructType, StructField, StringType, IntegerType

spark = SparkSession.builder.appName('SparkByExamples.com').getOrCreate()

data = [('James', '', 'Smith', '1991-04-01', 'M', 3000),
        ('Michael', 'Rose', '', '2000-05-19', 'M', 4000),
        ('Robert', '', 'Williams', '1978-09-05', 'M', 4000),
        ('Maria', 'Anne', 'Jones', '1967-12-01', 'F', 4000),
        ('Jen', 'Mary', 'Brown', '1980-02-17', 'F', -1)]

columns = ["firstname", "lastname", "dob", "gender", "salary"]

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

df.printSchema()
df.show(truncate = False)

df2 = df.withColumn("salary", col("salary").cast("Integer"))
df2.printSchema()
```

```

df2.show(truncate = False)

df3 = df.withColumn("salary", col("salary") * 100)
df3.printSchema()
df3.show(truncate = False)

df4 = df.withColumn("CopiedColumn", col("salary") * -1)
df4.printSchema()

df5 = df.withColumn("Country", lit("USA"))
df5.printSchema()

df6 = df.withColumn("Country", lit("USA")) \
    .withColumn("anotherColumn", lit("anotherValue"))
df6.printSchema()

df.withColumnRenamed("gender", "sex") \
    .show(truncate = False)

df4.drop("CopiedColumn") \
    .show(truncate = False)

```

```

root
|-- firstname: string (nullable = true)
|-- middlename: string (nullable = true)
|-- lastname: string (nullable = true)
|-- dob: string (nullable = true)
|-- gender: string (nullable = true)
|-- salary: long (nullable = true)

```

```

+-----+-----+-----+-----+-----+-----+
|firstname|middlename|lastname|dob      |gender|salary|
+-----+-----+-----+-----+-----+-----+
|James    |          |Smith   |1991-04-01|M      |3000   |
|Michael  |Rose     |        |2000-05-19|M      |4000   |
|Robert   |         |Williams|1978-09-05|M      |4000   |
|Maria    |Anne     |Jones   |1967-12-01|F      |4000   |
|Jen      |Mary     |Brown   |1980-02-17|F      |-1     |
+-----+-----+-----+-----+-----+-----+

```

```

root
|-- firstname: string (nullable = true)
|-- middlename: string (nullable = true)
|-- lastname: string (nullable = true)
|-- dob: string (nullable = true)
|-- gender: string (nullable = true)
|-- salary: integer (nullable = true)

```

```

+-----+-----+-----+-----+-----+-----+
|firstname|middlename|lastname|dob      |gender|salary|
+-----+-----+-----+-----+-----+-----+
|James    |          |Smith   |1991-04-01|M      |3000   |
|Michael  |Rose     |        |2000-05-19|M      |4000   |
|Robert   |         |Williams|1978-09-05|M      |4000   |
|Maria    |Anne     |Jones   |1967-12-01|F      |4000   |
|Jen      |Mary     |Brown   |1980-02-17|F      |-1     |
+-----+-----+-----+-----+-----+-----+

```

```

root
|-- firstname: string (nullable = true)
|-- middlename: string (nullable = true)
|-- lastname: string (nullable = true)
|-- dob: string (nullable = true)
|-- gender: string (nullable = true)
|-- salary: long (nullable = true)

```

```

+-----+-----+-----+-----+-----+-----+
|firstname|middlename|lastname|dob      |gender|salary|
+-----+-----+-----+-----+-----+-----+
|James    |          |Smith   |1991-04-01|M      |300000|
|Michael  |Rose     |        |2000-05-19|M      |400000|
|Robert   |         |Williams|1978-09-05|M      |400000|
|Maria    |Anne     |Jones   |1967-12-01|F      |400000|
|Jen      |Mary     |Brown   |1980-02-17|F      |-100   |
+-----+-----+-----+-----+-----+-----+

```

```

root
|-- firstname: string (nullable = true)
|-- middlename: string (nullable = true)
|-- lastname: string (nullable = true)
|-- dob: string (nullable = true)
|-- gender: string (nullable = true)
|-- salary: long (nullable = true)
|-- CopiedColumn: long (nullable = true)

```

```

root
|-- firstname: string (nullable = true)
|-- middlename: string (nullable = true)
|-- lastname: string (nullable = true)
|-- dob: string (nullable = true)
|-- gender: string (nullable = true)
|-- salary: long (nullable = true)
|-- Country: string (nullable = false)

root
|-- firstname: string (nullable = true)
|-- middlename: string (nullable = true)
|-- lastname: string (nullable = true)
|-- dob: string (nullable = true)
|-- gender: string (nullable = true)
|-- salary: long (nullable = true)
|-- Country: string (nullable = false)
|-- anotherColumn: string (nullable = false)

```

```

+-----+-----+-----+-----+-----+
|firstname|middlename|lastname|dob      |sex|salary|
+-----+-----+-----+-----+-----+
|James    |          |Smith   |1991-04-01|M  |3000  |
|Michael  |Rose      |        |2000-05-19|M  |4000  |
|Robert   |          |Williams|1978-09-05|M  |4000  |
|Maria    |Anne      |Jones   |1967-12-01|F  |4000  |
|Jen      |Mary      |Brown   |1980-02-17|F  |-1    |
+-----+-----+-----+-----+-----+

```

```

+-----+-----+-----+-----+-----+
|firstname|middlename|lastname|dob      |gender|salary|
+-----+-----+-----+-----+-----+
|James    |          |Smith   |1991-04-01|M      |3000  |
|Michael  |Rose      |        |2000-05-19|M      |4000  |
|Robert   |          |Williams|1978-09-05|M      |4000  |
|Maria    |Anne      |Jones   |1967-12-01|F      |4000  |
|Jen      |Mary      |Brown   |1980-02-17|F      |-1    |
+-----+-----+-----+-----+-----+

```