Inferred and explicit schemas

In [26]:

```
from pyspark.sql import SparkSession

spark = SparkSession \
    .builder \
    .appName("Inferred and explicit schemas") \
    .getOrCreate()
```

In [46]:

```
from pyspark.sql.types import Row
```

Inferring schema

In [33]:

```
lines = sc.textFile("../datasets/students.txt")
```

In [34]:

```
lines.collect()
```

Out[34]:

```
['Emily,44,55,78', 'Andy,47,34,89', 'Rick,55,78,55', 'Aaron,66,34,98']
```

In [35]:

```
parts = lines.map(lambda l: l.split(","))

parts.collect()
```

Out[35]:

```
[['Emily', '44', '55', '78'],
 ['Andy', '47', '34', '89'],
 ['Rick', '55', '78', '55'],
 ['Aaron', '66', '34', '98']]
```

In [36]:

```
students = parts.map(lambda p: Row(name=p[0], math=int(p[1]), english=int(p[2]), science
=int(p[3])))
```

In [37]:

```
students.collect()
```

Out[37]:

```
[Row(english=55, math=44, name='Emily', science=78),
 Row(english=34, math=47, name='Andy', science=89),
 Row(english=78, math=55, name='Rick', science=55),
 Row(english=34, math=66, name='Aaron', science=98)]
```

In [38]:

```
schemaStudents = spark.createDataFrame(students)

schemaStudents.createOrReplaceTempView("students")
```

In [39]:

```
schemaStudents.columns
```

Out[39]:

```
['english', 'math', 'name', 'science']
```

In [40]:

```
schemaStudents.schema
```

Out[40]:

```
StructType(List(StructField(english,LongType,true),StructField(math,LongType,true),Struct
Field(name,StringType,true),StructField(science,LongType,true)))
```

In [43]:

```
spark.sql("SELECT * FROM students").show()
```

```
+-------+----+-----+-------+
|english|math| name|science|
+-------+----+-----+-------+
|     55|  44|Emily|     78|
|     34|  47| Andy|     89|
|     78|  55| Rick|     55|
|     34|  66|Aaron|     98|
+-------+----+-----+-------+
```

Explicit schema

In [44]:

```
parts.collect()
```

Out[44]:

```
[['Emily', '44', '55', '78'],
 ['Andy', '47', '34', '89'],
 ['Rick', '55', '78', '55'],
 ['Aaron', '66', '34', '98']]
```

In [45]:

```
schemaString = "name math english science"
```

In [56]:

```
from pyspark.sql.types import StructType, StructField, StringType, LongType

fields = [StructField('name', StringType(), True),
          StructField('math', LongType(), True),
          StructField('english', LongType(), True),
          StructField('science', LongType(), True),
]
```

In [57]:

```
schema = StructType(fields)
```

In [58]:

```
schemaStudents = spark.createDataFrame(parts, schema)
```

In [61]:

```
schemaStudents.columns
```

Out[61]:

```
['name', 'math', 'english', 'science']
```

In [60]:

```
schemaStudents.schema
```

Out[60]:

```
StructType(List(StructField(name,StringType,true),StructField(math,LongType,true),StructF
ield(english,LongType,true),StructField(science,LongType,true)))
```

In [62]:

```
spark.sql("SELECT * FROM students").show()
```

```
+-------+----+-----+-------+
|english|math| name|science|
+-------+----+-----+-------+
|     55|  44|Emily|     78|
|     34|  47| Andy|     89|
|     78|  55| Rick|     55|
|     34|  66|Aaron|     98|
+-------+----+-----+-------+
```

In [ ]:

```

```