**Spark SQL: Programmatically specifying the   
schema**

**Limitations in Case classes:**

There are certain limitations of using case class for defining the schema:

* Case classes cannot take more than 22 fields
* You may not know the schema beforehand

**About this activity:**

* Data is loaded as an RDD of the Row objects
* Schema is created separately using the StructType and StructField objects, which represent a table and a field respectively.
* Schema is applied to the Row RDD to create a DataFrame

1. Create a new Scala notebook and import the below packages

Import for the implicit conversion

Shape

Description automatically generated with medium confidence

Import the spark SQL datatypes and row objects



A picture containing chart

Description automatically generated

1. Use the file which we already loaded into hive (activity – 1). Load the employee data into RDD



1. Split each line into array of string based on comma as a delimiter

A picture containing chart

Description automatically generated



1. Convert the RDD of array[string] to the RDD of the row objects:



1. Create schema using the struct type and struct field objects. The struct field object takes parameters in the form of param name, param type and nullability

Text, letter

Description automatically generated

1. Apply schema to create the employee data frame



1. Register the employeeDF as a view

A picture containing company name

Description automatically generated

1. Run a SQL query against it



1. Get the output values from emp.

A picture containing chart

Description automatically generated