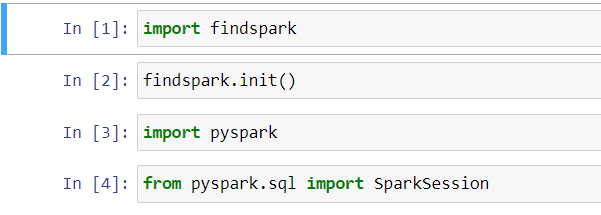
**PYSPARK SQL**

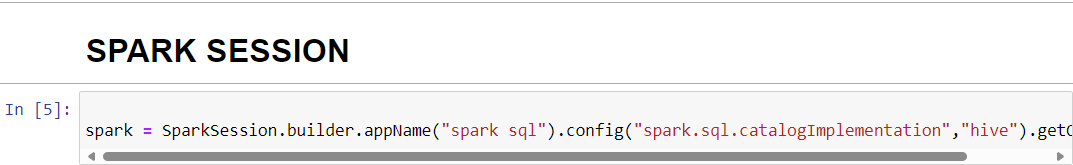
Spark SQL brings native support for SQL to Spark and streamlines the process of querying data stored both in RDDs (Spark’s distributed datasets) and in external sources. Spark SQL conveniently blurs the lines between RDDs and relational tables. Unifying these powerful abstractions makes it easy for developers to intermix SQL commands querying external data with complex analytics, all within in a single application. Concretely, Spark SQL will allow developers to:

* Import relational data from Parquet files and Hive tables
* Run SQL queries over imported data and existing RDDs
* Easily write RDDs out to Hive tables or Parquet files

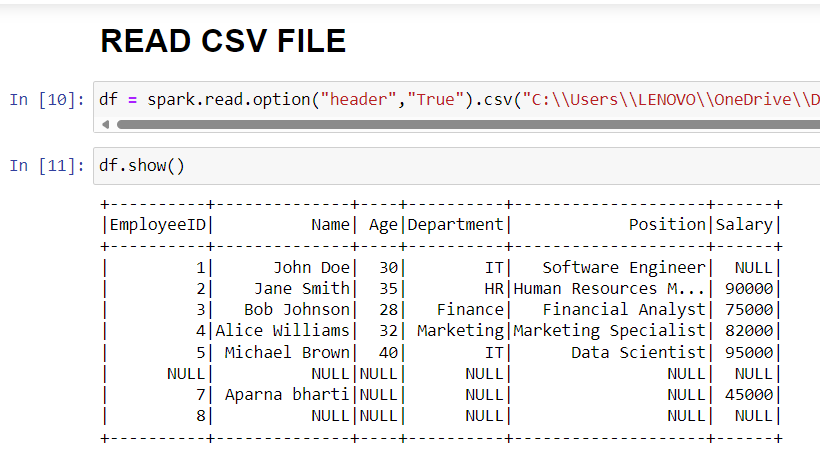
## **Architecture of Spark SQL**

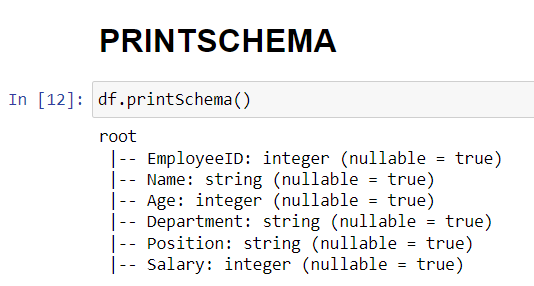
Spark SQL is a library on top of the Spark core execution engine, as shown in *Figure 4.2*. It exposes SQL interfaces using JDBC/ODBC for Data Warehousing applications or through a command-line console for interactively executing queries. So, any **Business Intelligence** (**BI**) tools can connect to Spark SQL to perform analytics at memory speeds. It also exposes a Dataset API and DataFrame API, which are supported in Java, Scala, Python, and R. Spark SQL users can use the Data Source API to read and write data from and to a variety of sources to create a DataFrame or a Dataset. *Figure 4.2* also indicates the traditional way of creating and operating on RDDs from programming languages to the Spark core engine.

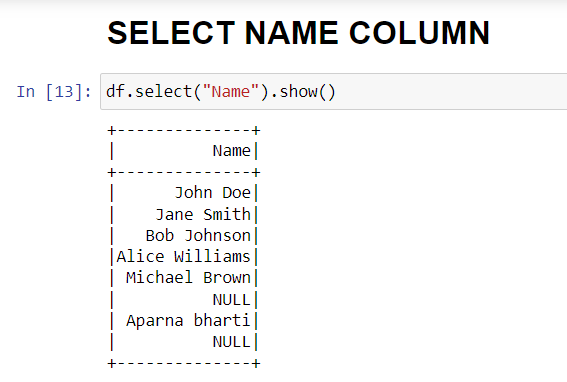
****

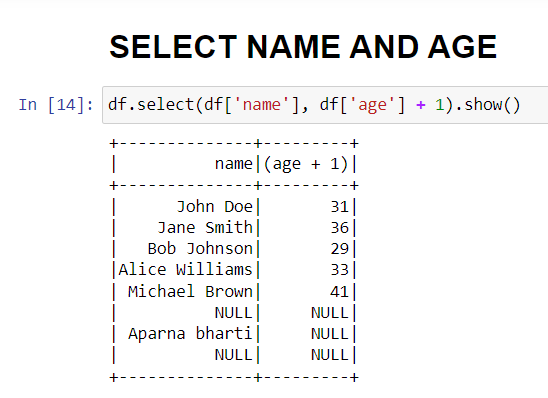
****

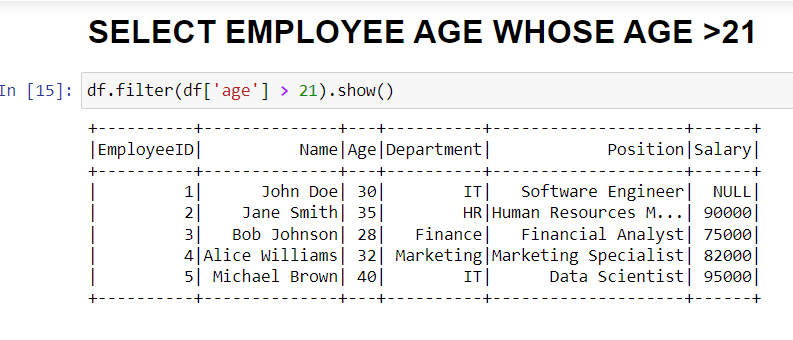
****

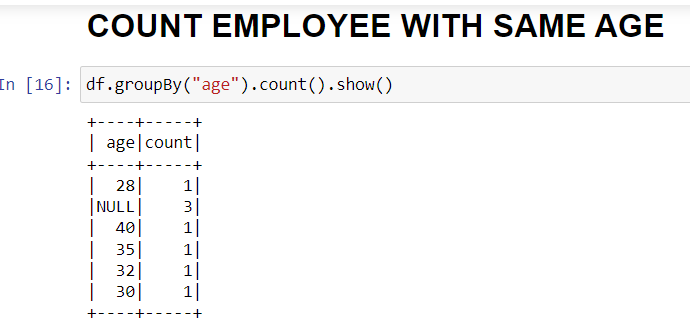
****

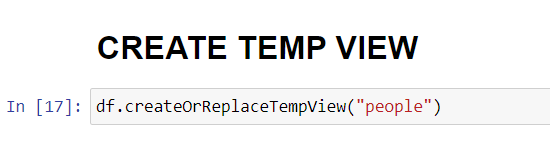
****

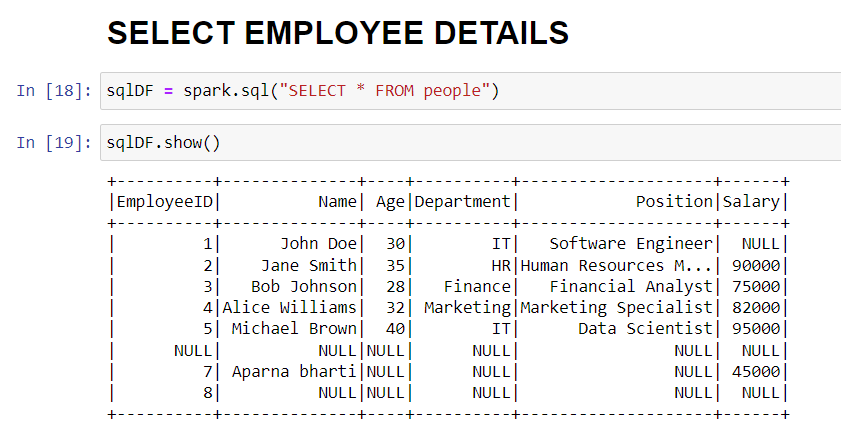
****

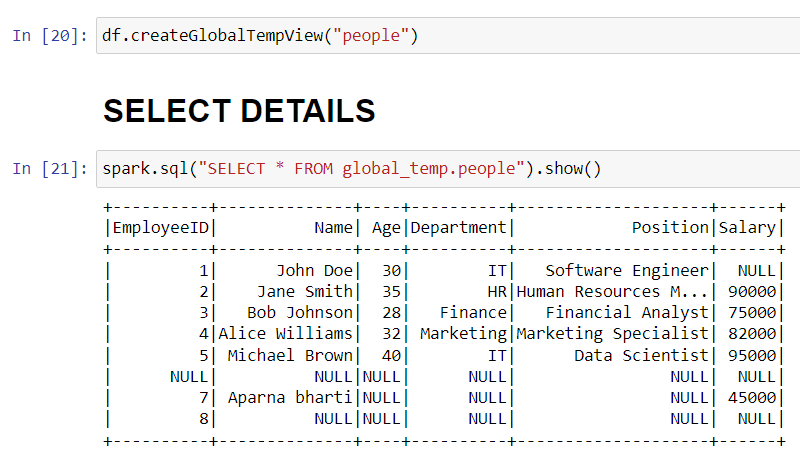
****

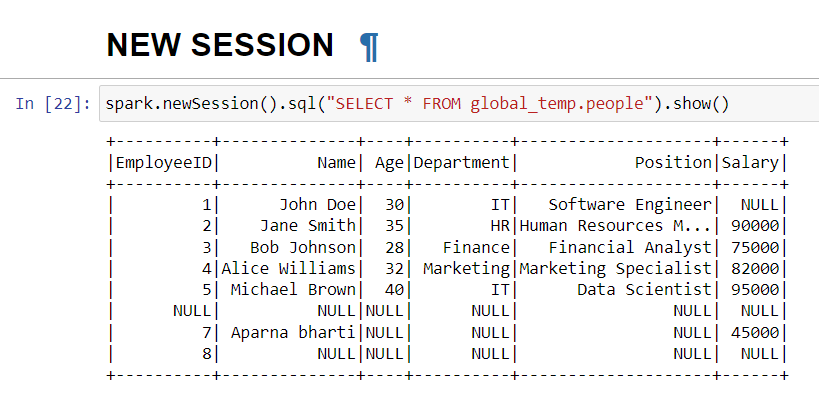
****

****

****

****

****

****