

# Get Started with PySpark Programming



# Module Agenda

## Get Started with PySpark Programming

### Spark SQL Overview

DE 0.1 – Spark SQL

DE 0.2L – Spark SQL Lab

DE 0.3 – DataFrame & Column

DE 0.4L – Purchase Revenues Lab

DE 0.5 – Aggregation

DE 0.6L – Revenue by Traffic Lab

# Spark SQL Overview

# Spark SQL is a module for structured data processing with multiple interfaces

**SQL**

**DataFrame API**

Python, Scala, Java, R



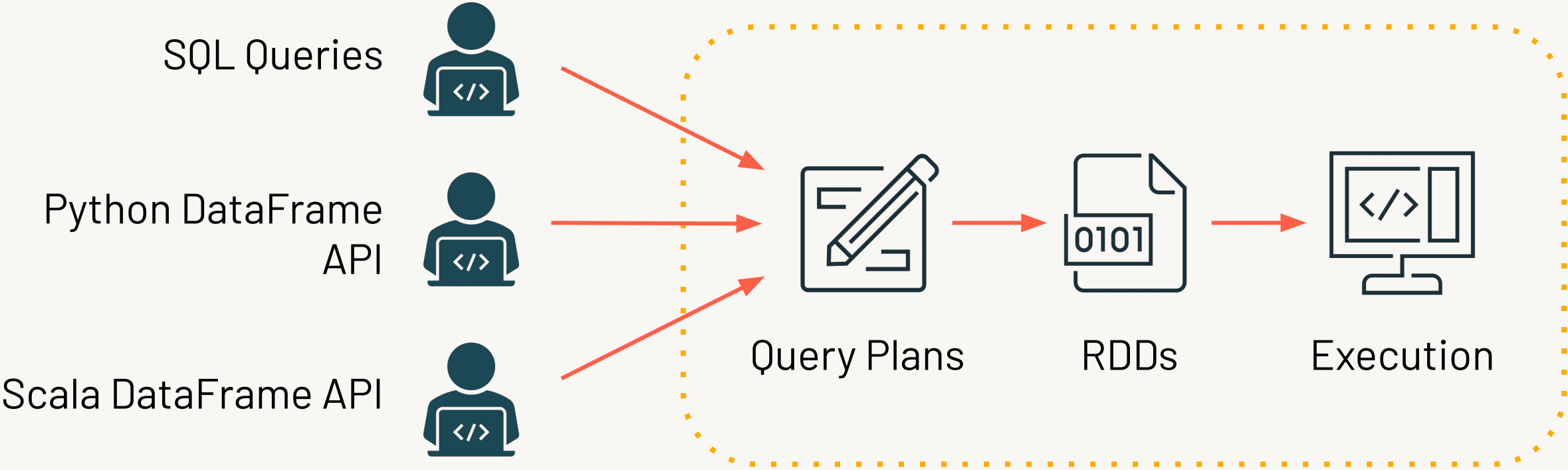
# The same Spark SQL query can be expressed with SQL and the DataFrame API

```
SELECT id, result
FROM exams
WHERE result > 70
ORDER BY result
```

```
spark.table("exams")
  .select("id", "result")
  .where("result > 70")
  .orderBy("result")
```



# Spark SQL executes all queries on the same engine



# Spark SQL optimizes queries before execution

