# The structure spectrum

| Structured | Semi-structured | Unstructured |
| --- | --- | --- |
| Relational Databases | HTML | Plain text |
| Parquet | XML | Generic media |
| Formatted Messages | JSON | |

# Tabular data

- Simple data format
- Common variants:
  - Comma Separated Values (CSV),
  - Tab Separated Values (TSV)

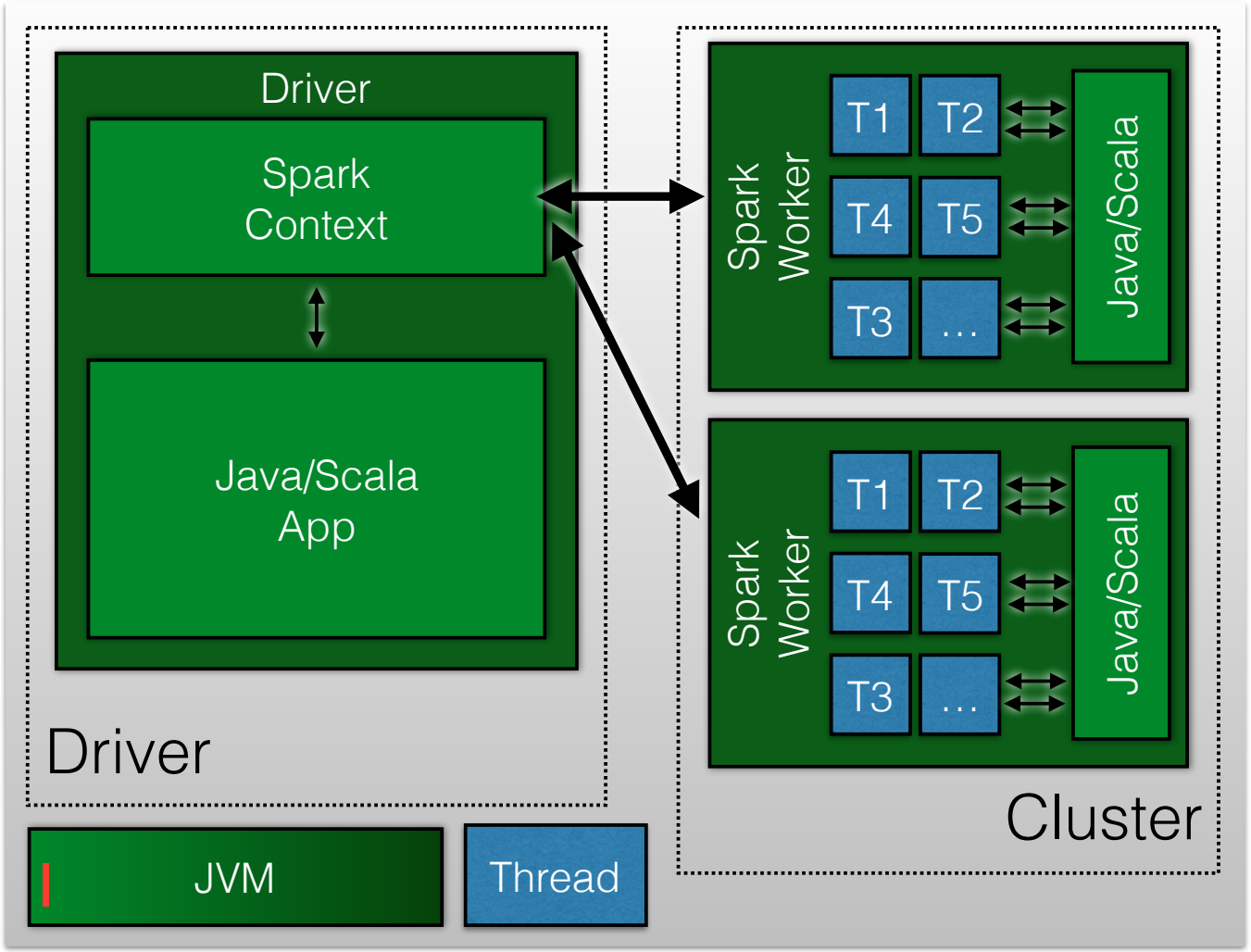
# Tabular data

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Genom | _ | PR | PR | _ | 3 | AA | _ | _ |
| 2 | skattereformen | _ | NN | NN | _ | 1 | PA | _ | _ |
| 3 | införs | _ | VV | VV | _ | 0 | ROOT | _ | _ |
| 4 | individuell | _ | AJ | AJ | _ | 5 | AT | _ | _ |
| 5 | beskattning | _ | VN | VN | _ | 3 | SS | _ | _ |
| 6 | ( | _ | IR | IR | _ | 5 | IR | _ | _ |
| 7 | särbeskattning | _ | VN | VN | _ | 5 | AN | _ | _ |
| 8 | ) | _ | IR | IR | _ | 5 | JR | _ | _ |
| 9 | av | _ | PR | PR | _ | 5 | ET | _ | _ |
| 10 | arbetsinkomster | _ | NN | NN | _ | 9 | PA | _ | _ |
| 11 | . | _ | IP | IP | _ | 3 | IP | _ | _ |

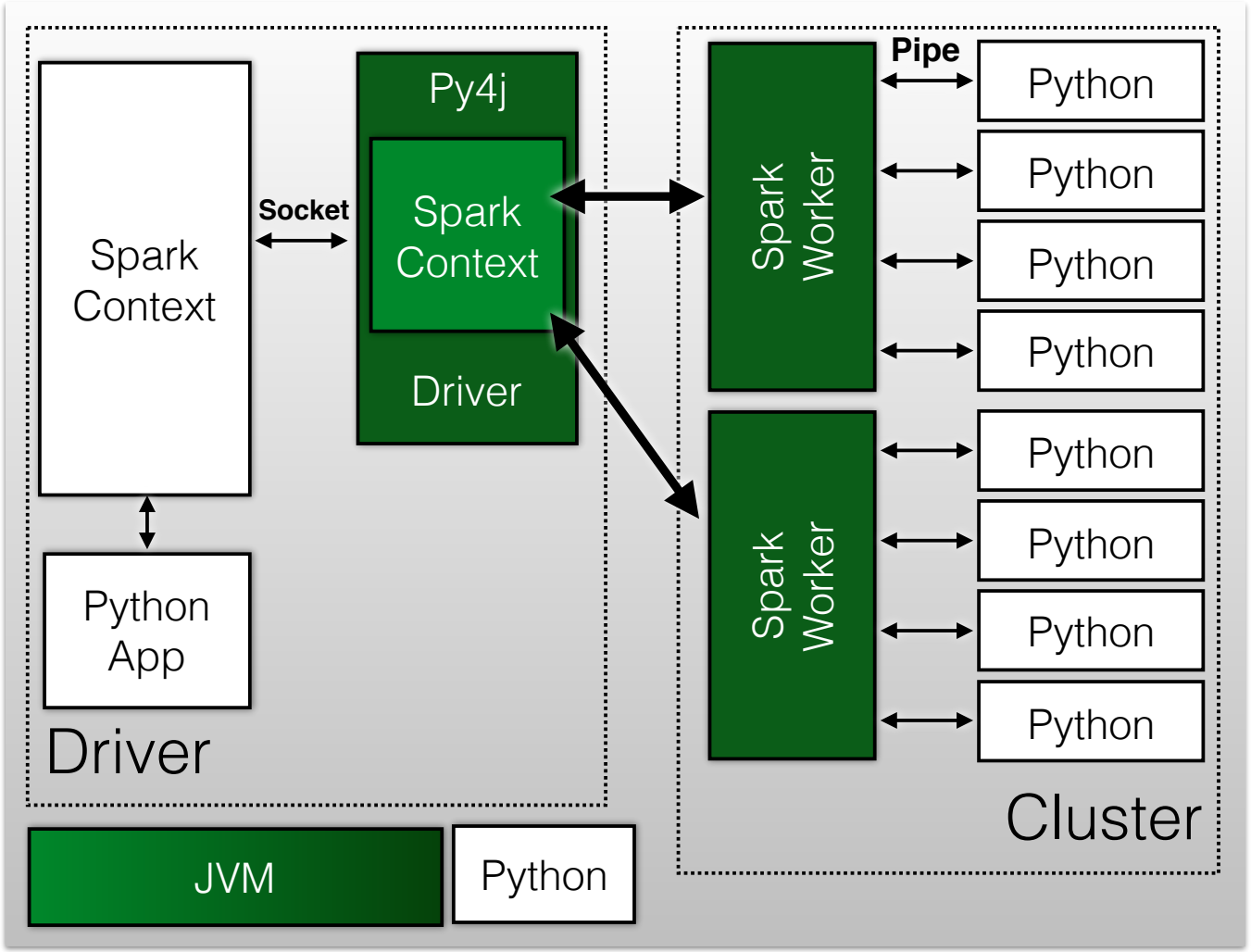
# Spark SQL

- **Why another API?**
  - Performance

# Spark: Java/Scala



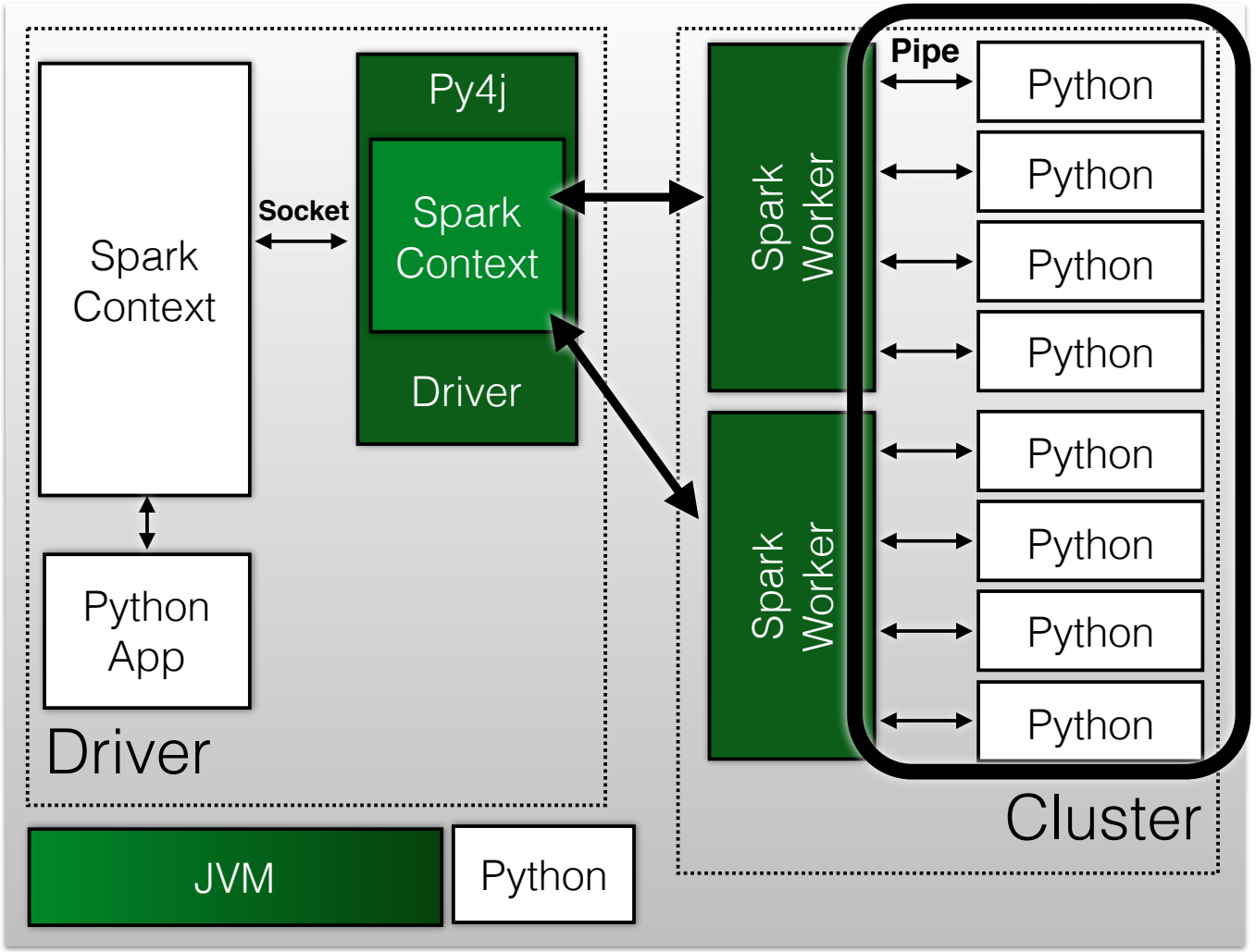
# Pyspark


Spark Context
Socket
Py4j
Spark Context
Driver
Python App
Driver
Pipe
Spark Worker
Spark Worker
Python
Python
Python
Python
Python
Python
Python
Python
Cluster
JVM
Python



# Pyspark

**Python and pipe Overhead**

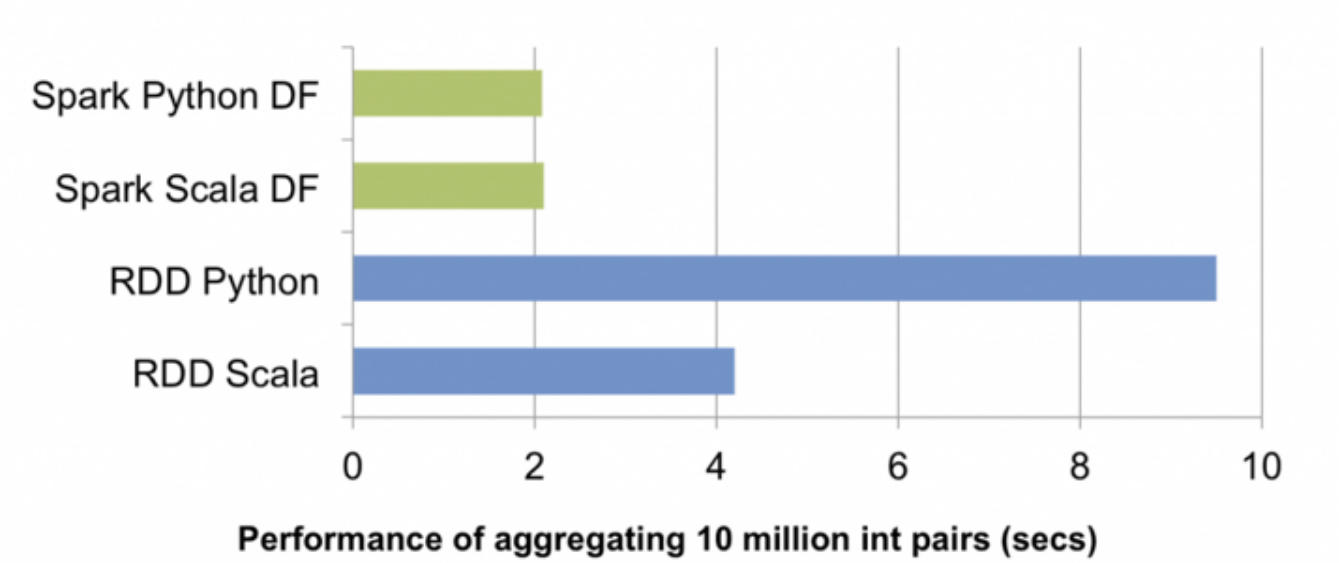


# Pyspark: Performance

- Serialization/Deserialization overhead
- CPython / PyPy is officially supported.
  CPython still has most support since numpy is not yet supported in PyPy
- Python code performance issues if most time spent computing is not within native libraries
- **How can most of the headache of overhead be reduced?**

# Pyspark SQL

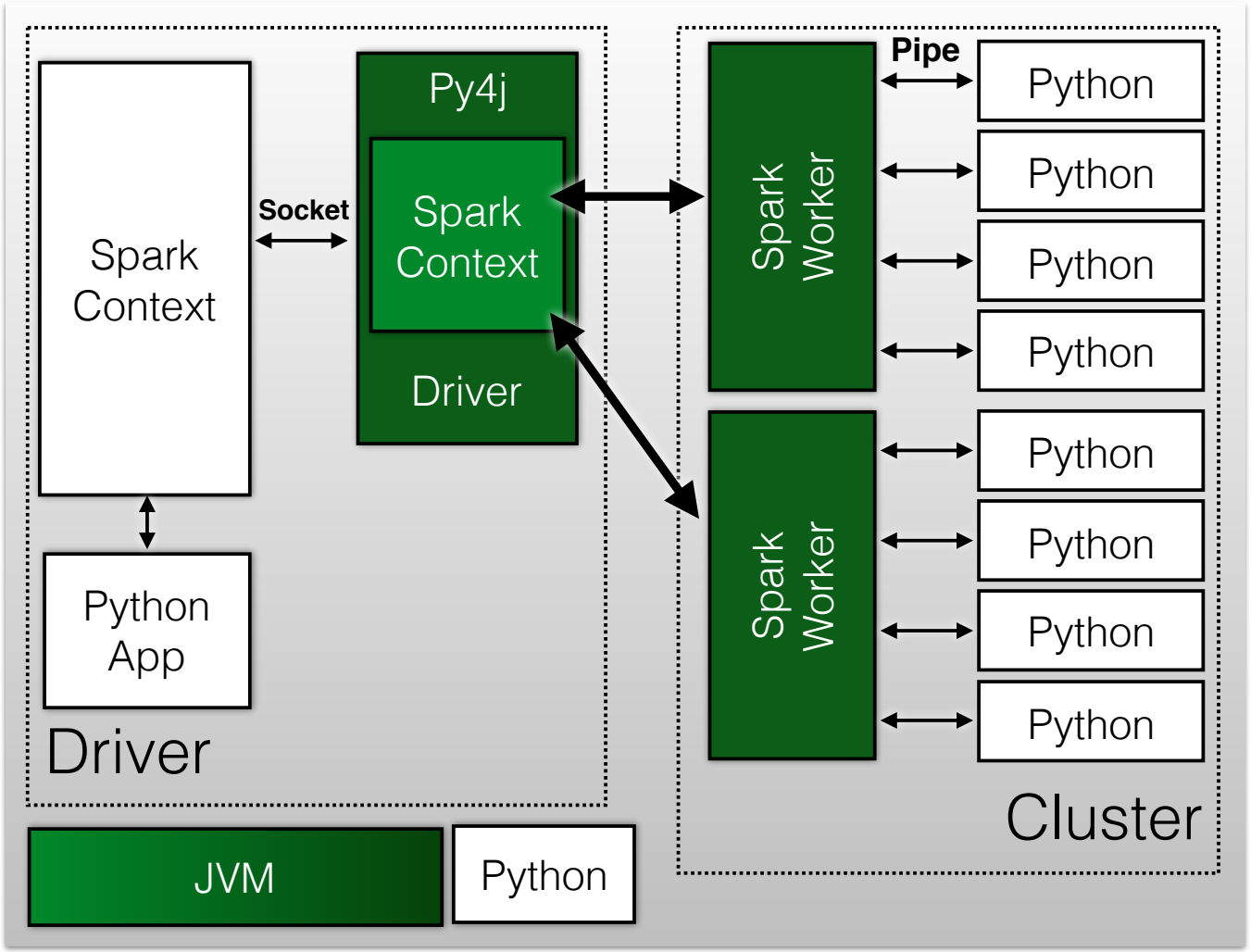
- Optimized data processing using optimized data structures and spark code, no extra overhead.
- New in Spark 1.3.0: Dataframes (SQL)




Source: https://databricks.com/blog/2015/02/17/introducing-dataframes-in-spark-for-large-scale-data-science.html

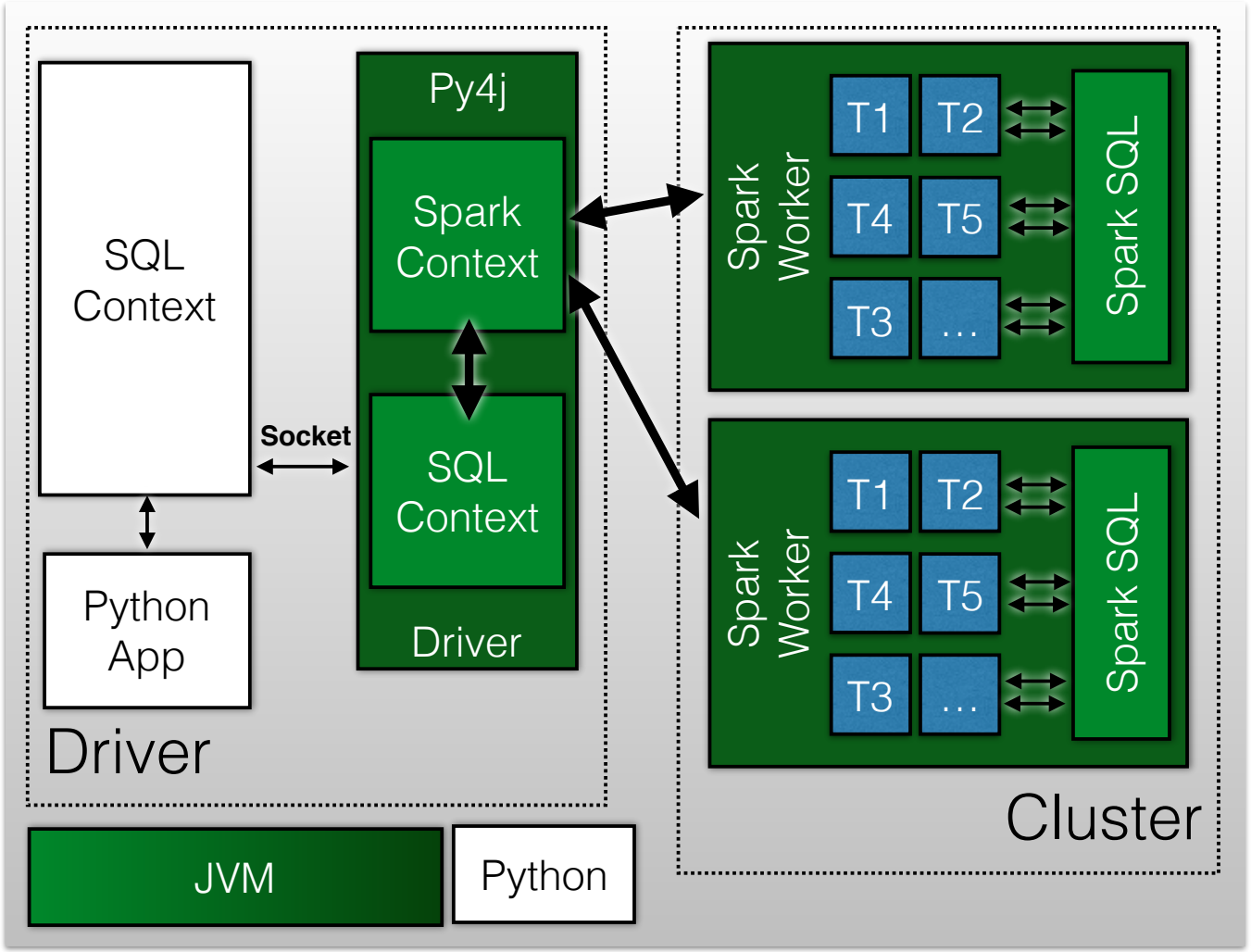


# Pyspark


Spark Context
Socket
Py4j
Spark Context
Driver
Python App
Driver
Pipe
Spark Worker
Spark Worker
Python
Python
Python
Python
Python
Python
Python
Python
Cluster
JVM
Python



# Pyspark SQL


SQL Context
Python App
Driver
Socket
Py4j
Spark Context
SQL Context
Driver
JVM
Python
Spark Worker
T1
T2
T4
T5
T3
…
Spark SQL
Spark Worker
T1
T2
T4
T5
T3
…
Spark SQL
Cluster



# Pyspark

**No overhead**
**+ Native Spark SQL**
**Code only.**