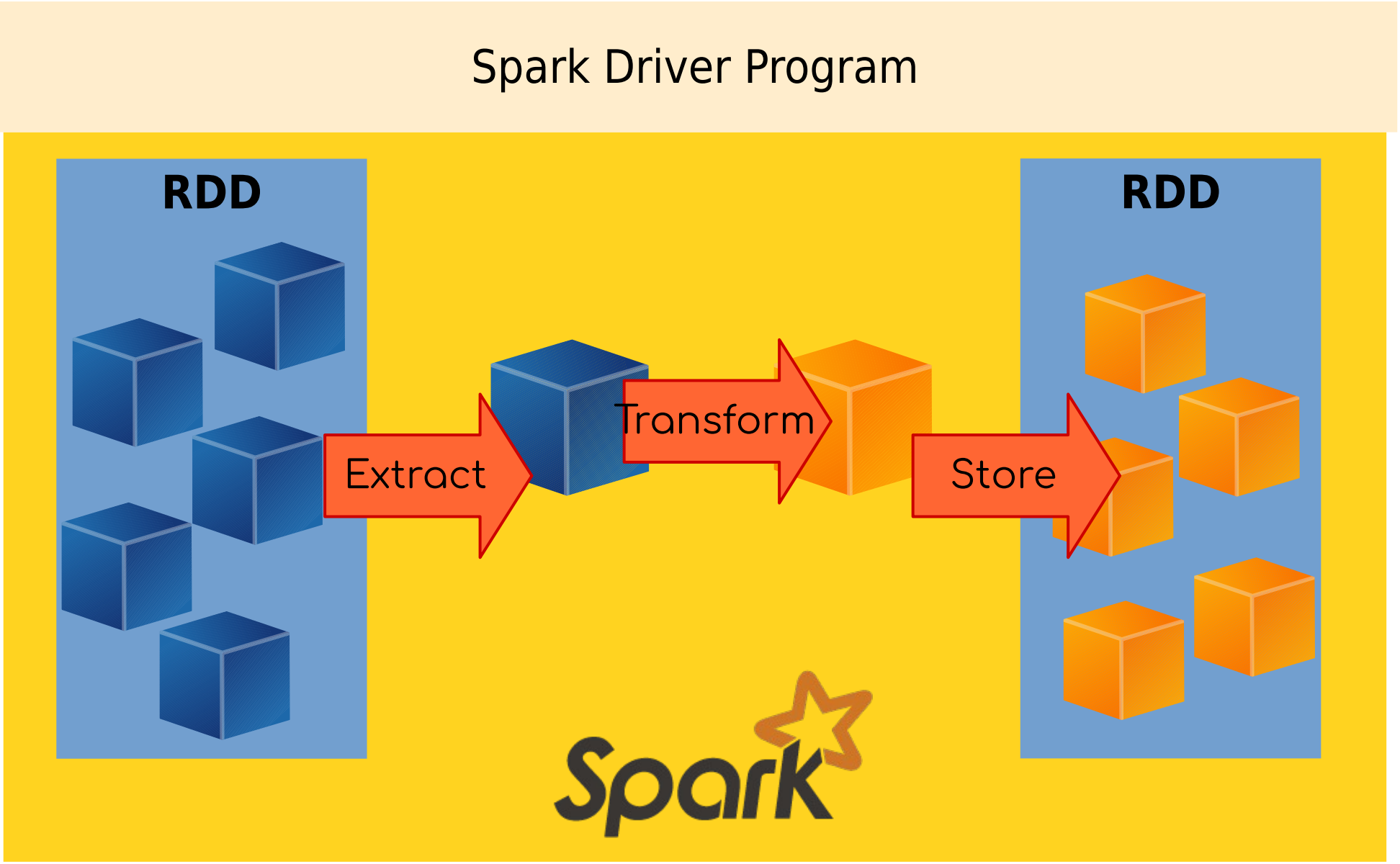
dimajix

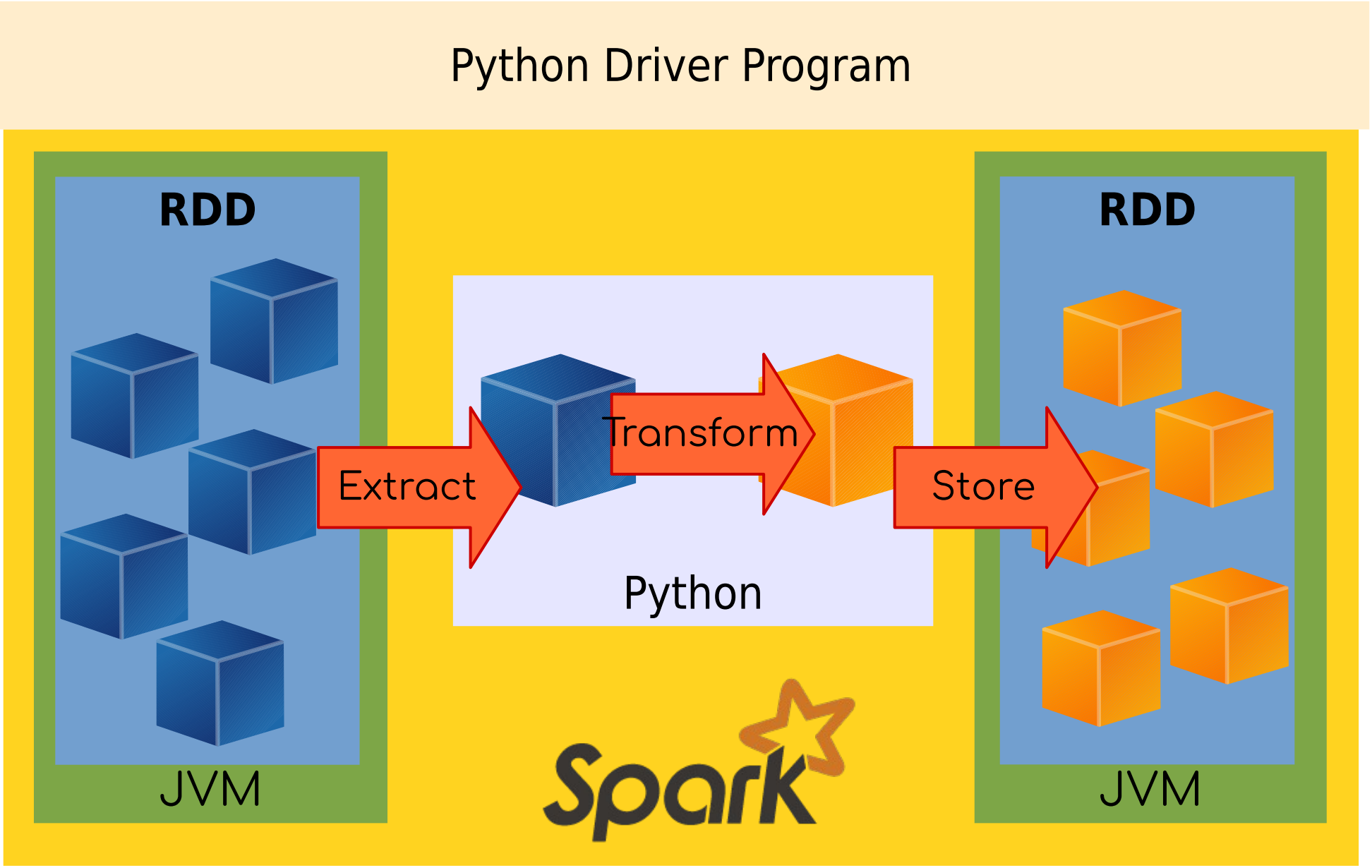
Big Data | Analytics | Algorithms

Spark Runtime Architecture

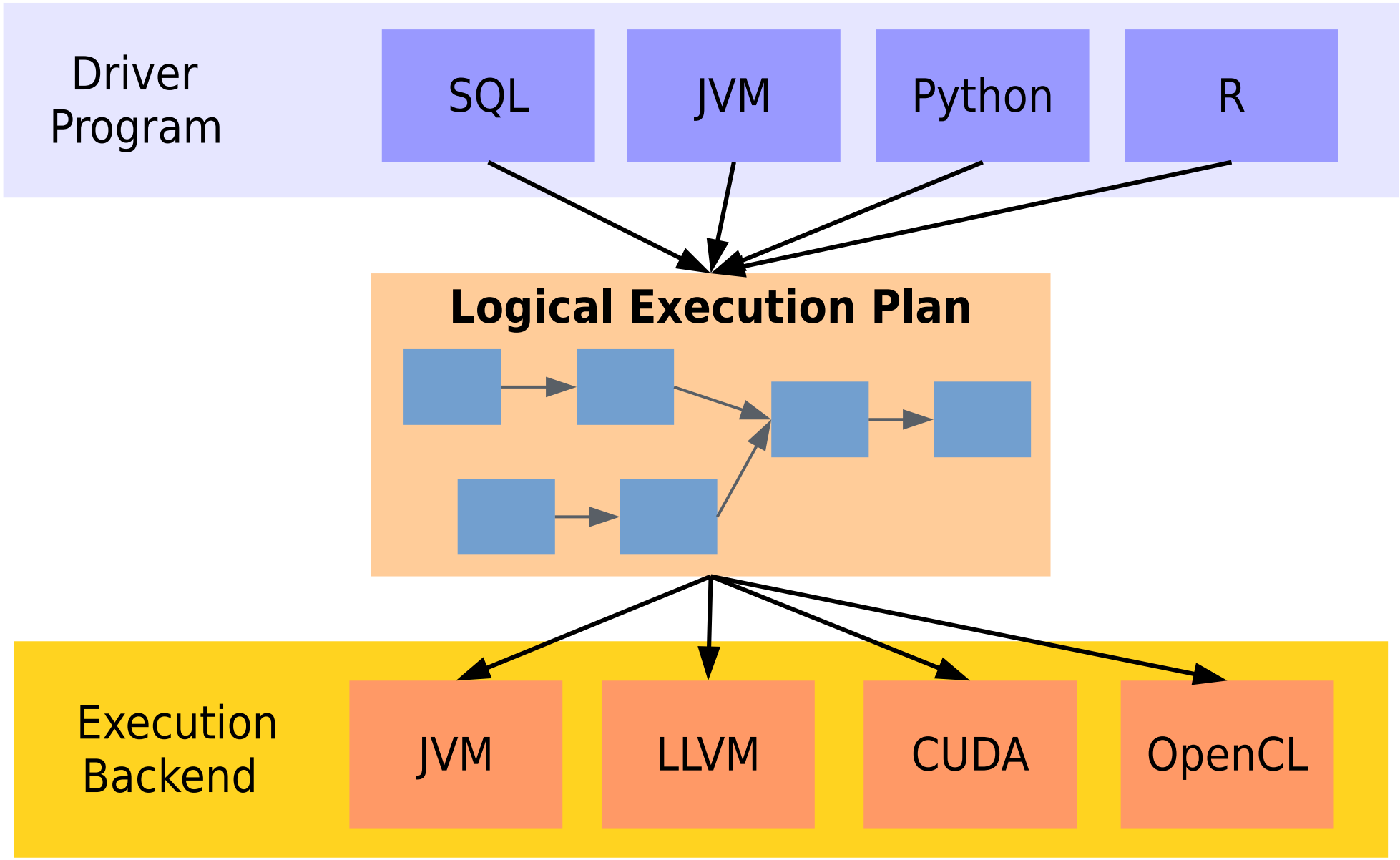
dimajix | Dr. Kaya Kupferschmidt | k.kupferschmidt@dimajix.de


Spark Driver Program
RDD
Extract
Transform
Store
RDD
Spark

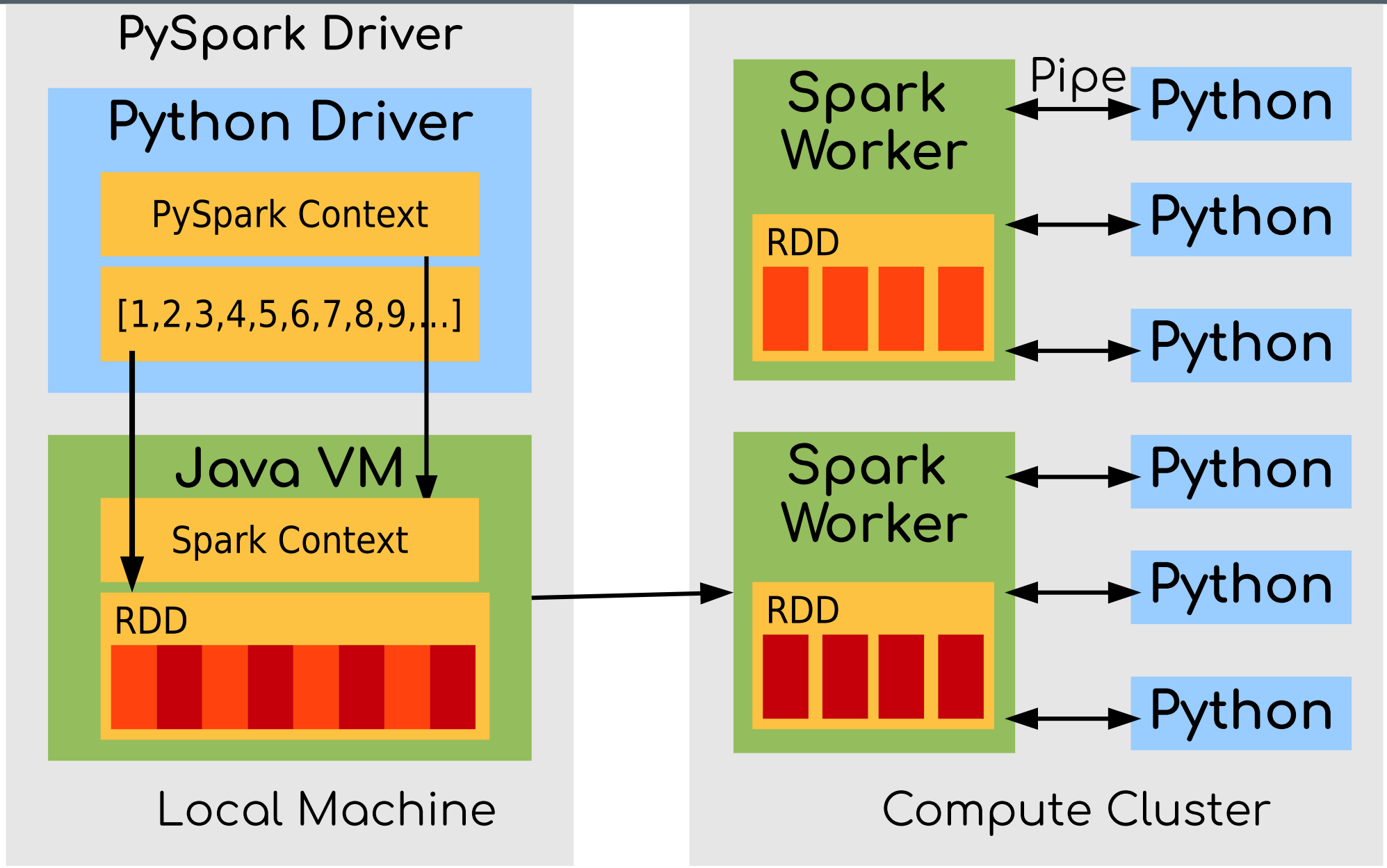
RDD Execution (PySpark)



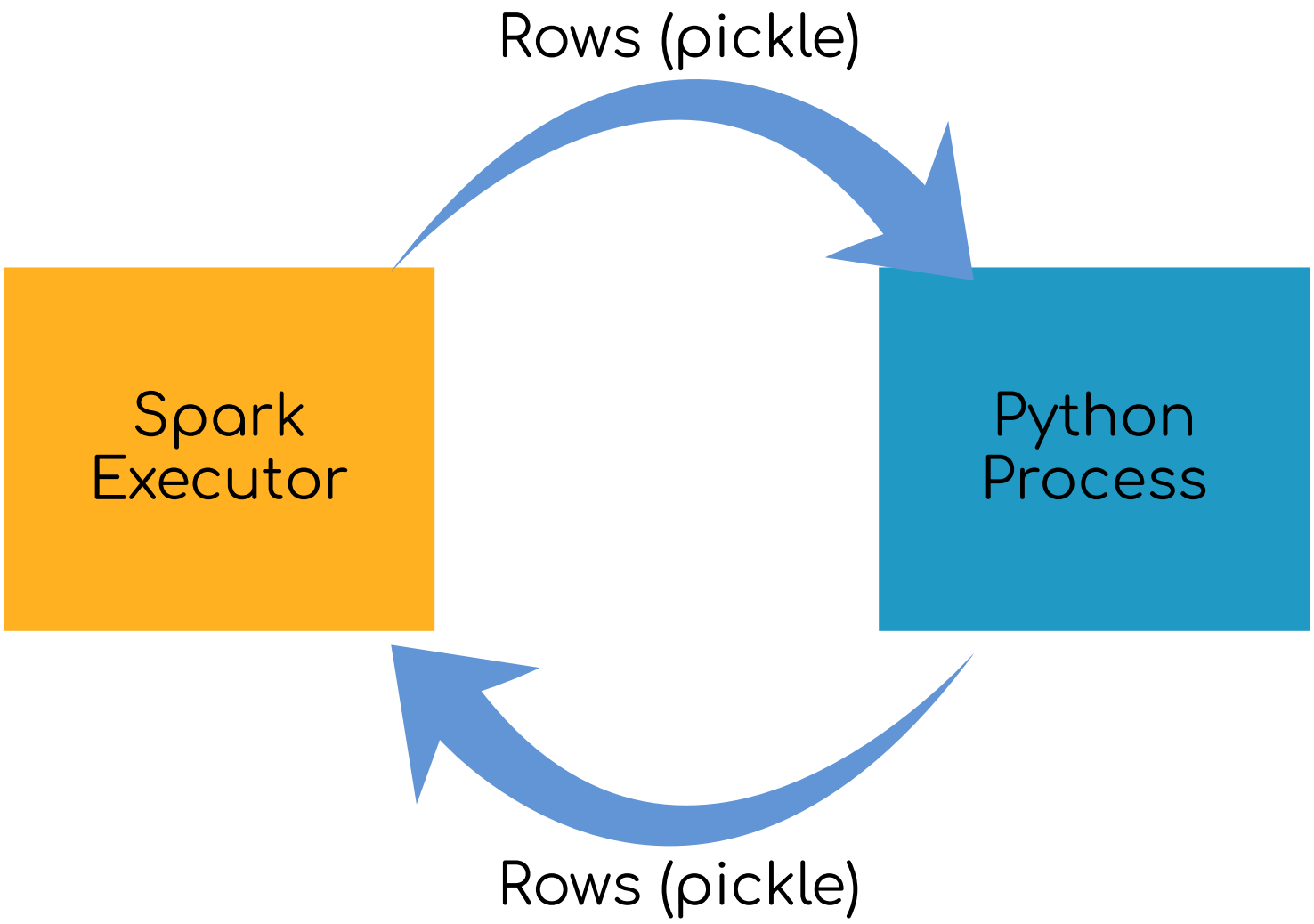
DataFrame Execution



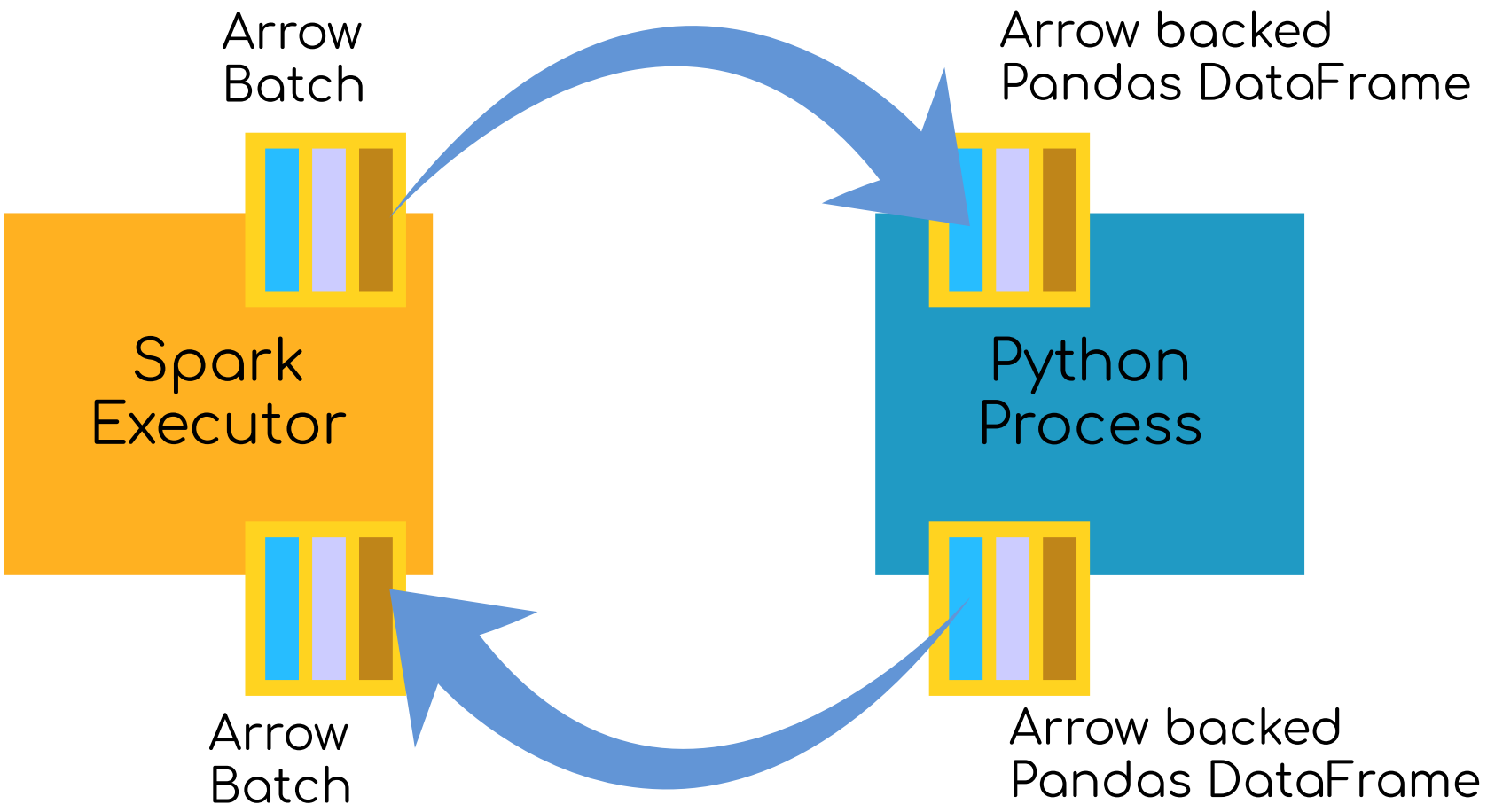
PySpark Architecture



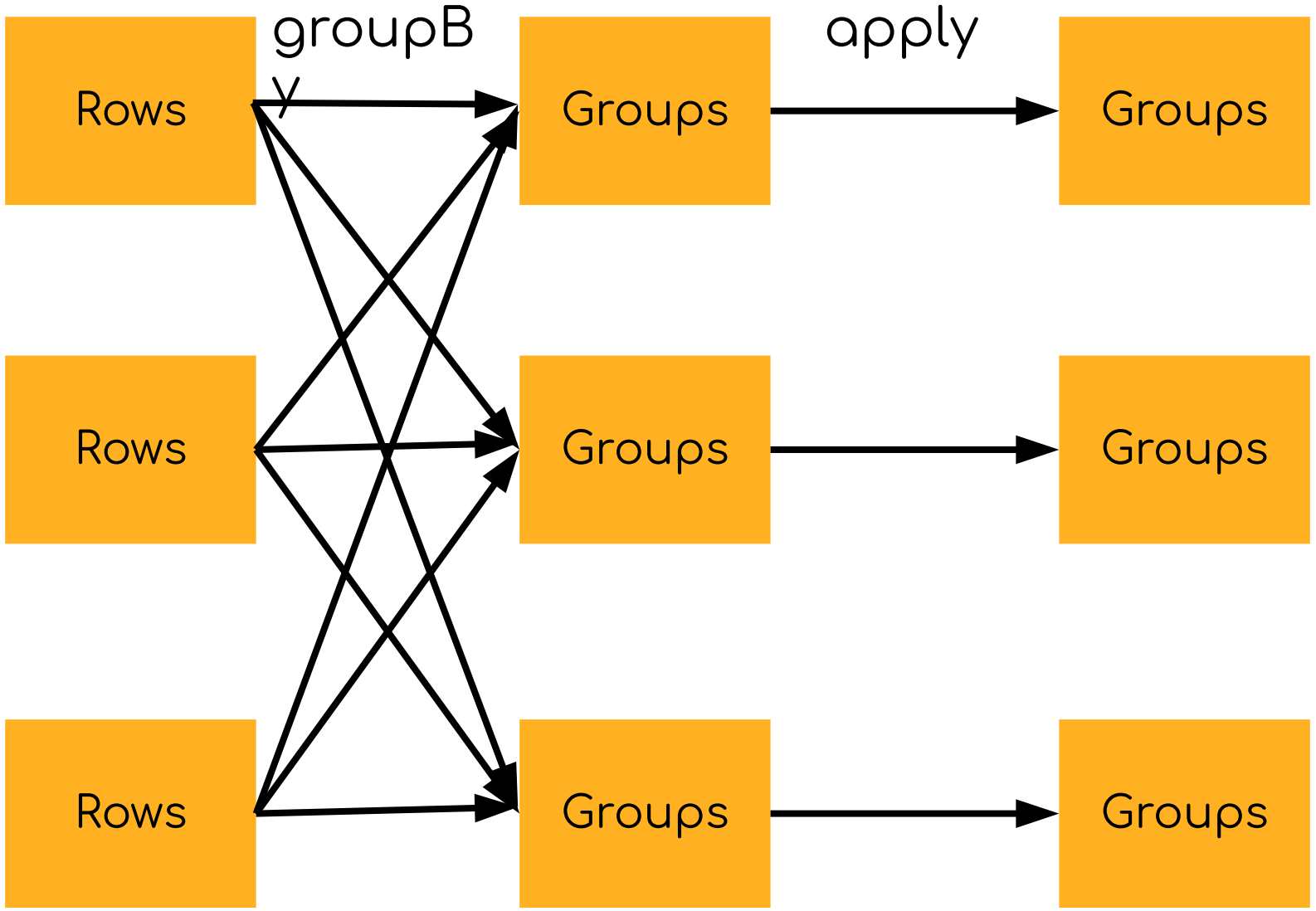
User Defined Functions


Rows (pickle)
Spark Executor
Python Process
Rows (pickle)

PySpark Vectorized UDF



PySpark Grouped UDF



Columnar Memory Model

	session_id	timestamp	source_ip
Row 1	1331246660	3/8/2012 2:44PM	99.155.155.225
Row 2	1331246351	3/8/2012 2:38PM	65.87.165.114
Row 3	1331244570	3/8/2012 2:09PM	71.10.106.181
Row 4	1331261196	3/8/2012 6:46PM	76.102.156.138

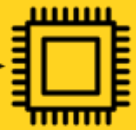
Traditional Memory Buffer

Row 1	1331246660
	3/8/2012 2:44PM
	99.155.155.225
Row 2	1331246351
	3/8/2012 2:38PM
	65.87.165.114
Row 3	1331244570
	3/8/2012 2:09PM
	71.10.106.181
Row 4	1331261196
	3/8/2012 6:46PM
	76.102.156.138

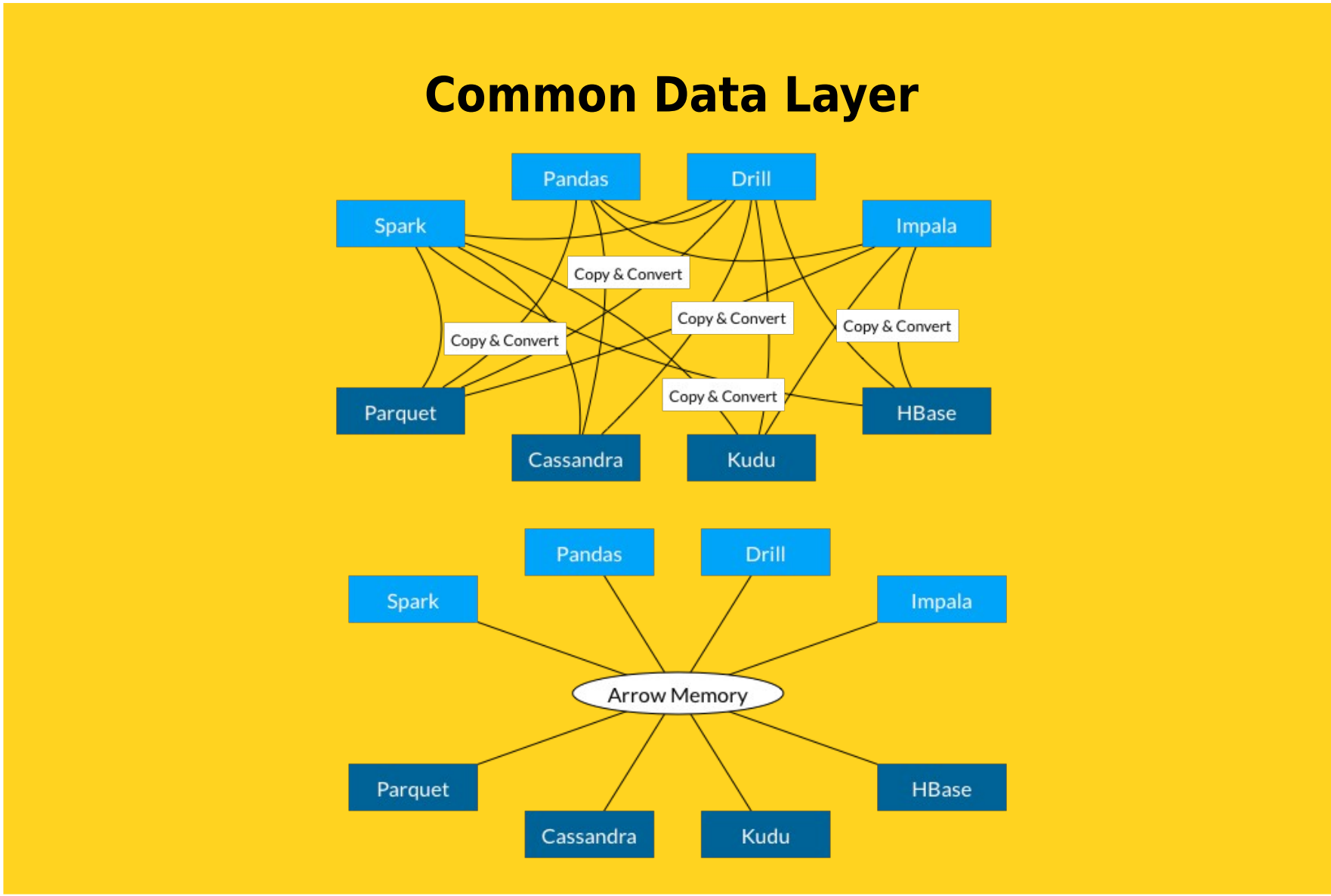
Arrow Memory Buffer

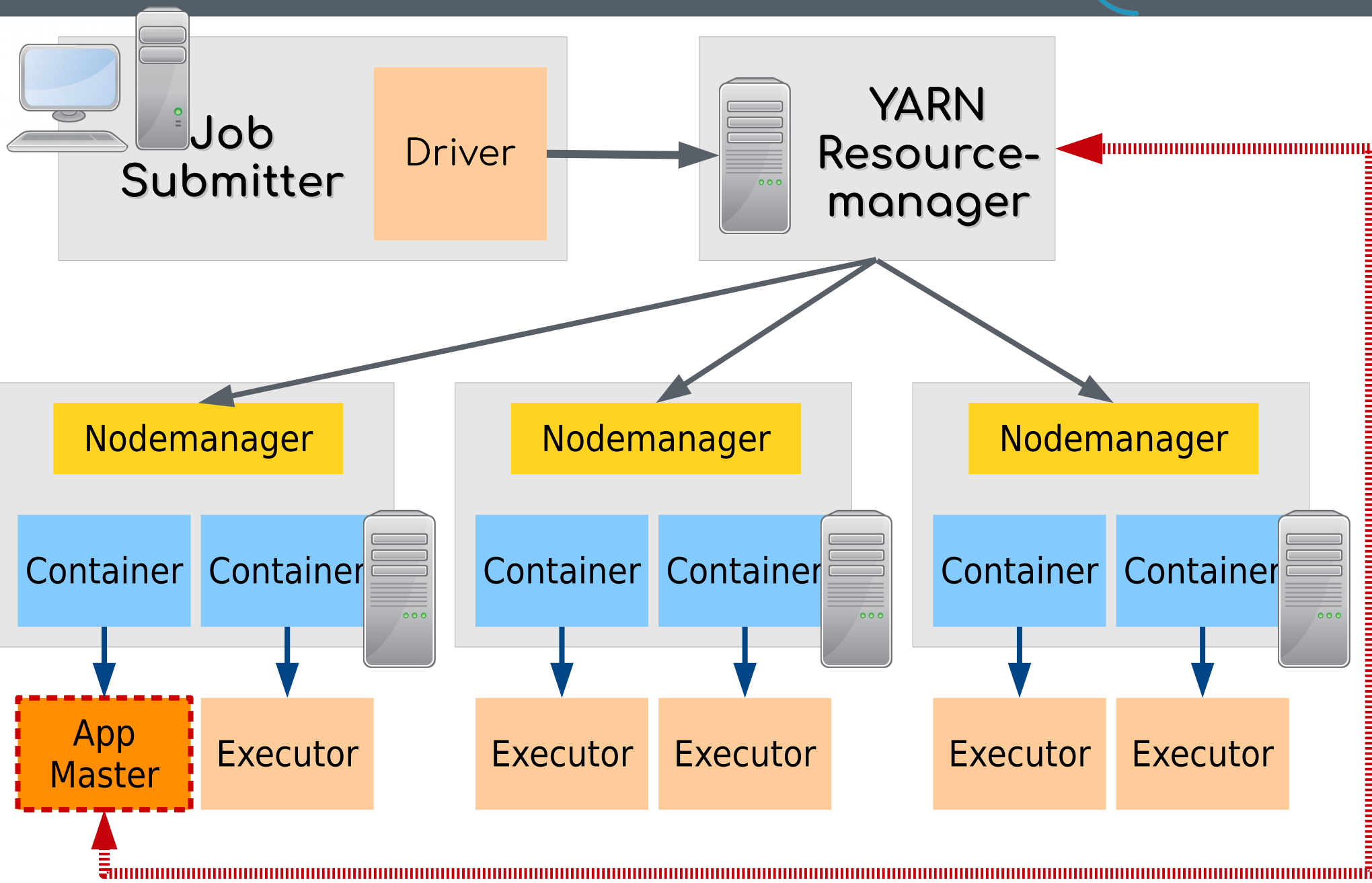
session_id	1331246660
	1331246351
	1331244570
	1331261196
timestamp	3/8/2012 2:44PM
	3/8/2012 2:38PM
	3/8/2012 2:09PM
	3/8/2012 6:46PM
source_ip	99.155.155.225
	65.87.165.114
	71.10.106.181
	76.102.156.138

```
SELECT * FROM clickstream
WHERE session_id = 1331246351
```



Intel CPU


Common Data Layer
Pandas
Drill
Spark
Impala
Copy & Convert
Copy & Convert
Copy & Convert
Copy & Convert
Parquet
HBase
Cassandra
Kudu
Pandas
Drill
Spark
Impala
Arrow Memory
Parquet
HBase
Cassandra
Kudu

Spark on YARN

Apache Spark – Runtime Architecture