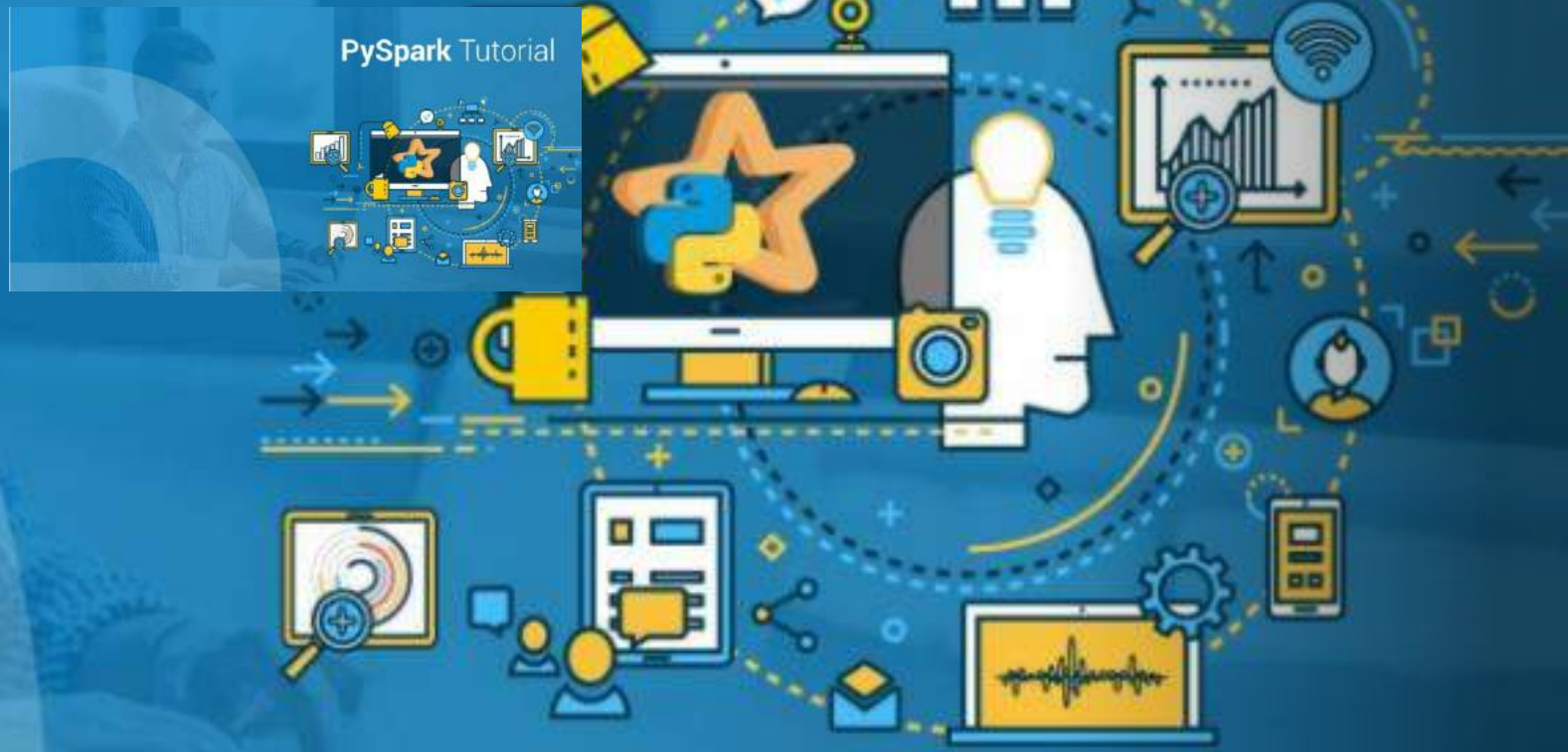
PySpark Tutorial

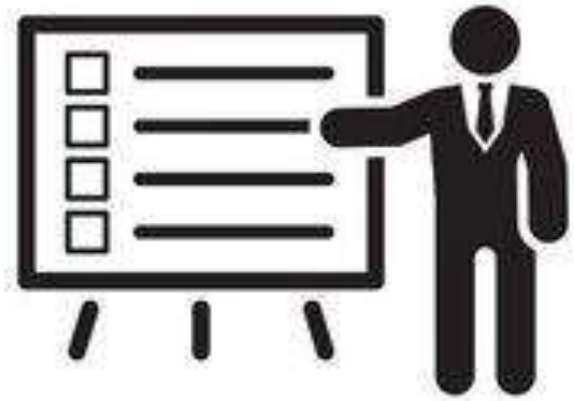


Presenters :

- **Subhan Khalid**
  - P200086
- **Usama Yazdani**
  - P200598

Objectives of Today's Training

- Problem statement
- Apache spark
- Define Pyspark
- Installation
- Key Components
- Architecture
- How it better than Hadoop
- Spark Session
- Demo



Problem Statement

- Efficiency Optimization for ML Model Training on Large Datasets
- SQL Query Execution Times
- Speed does matter in data processing

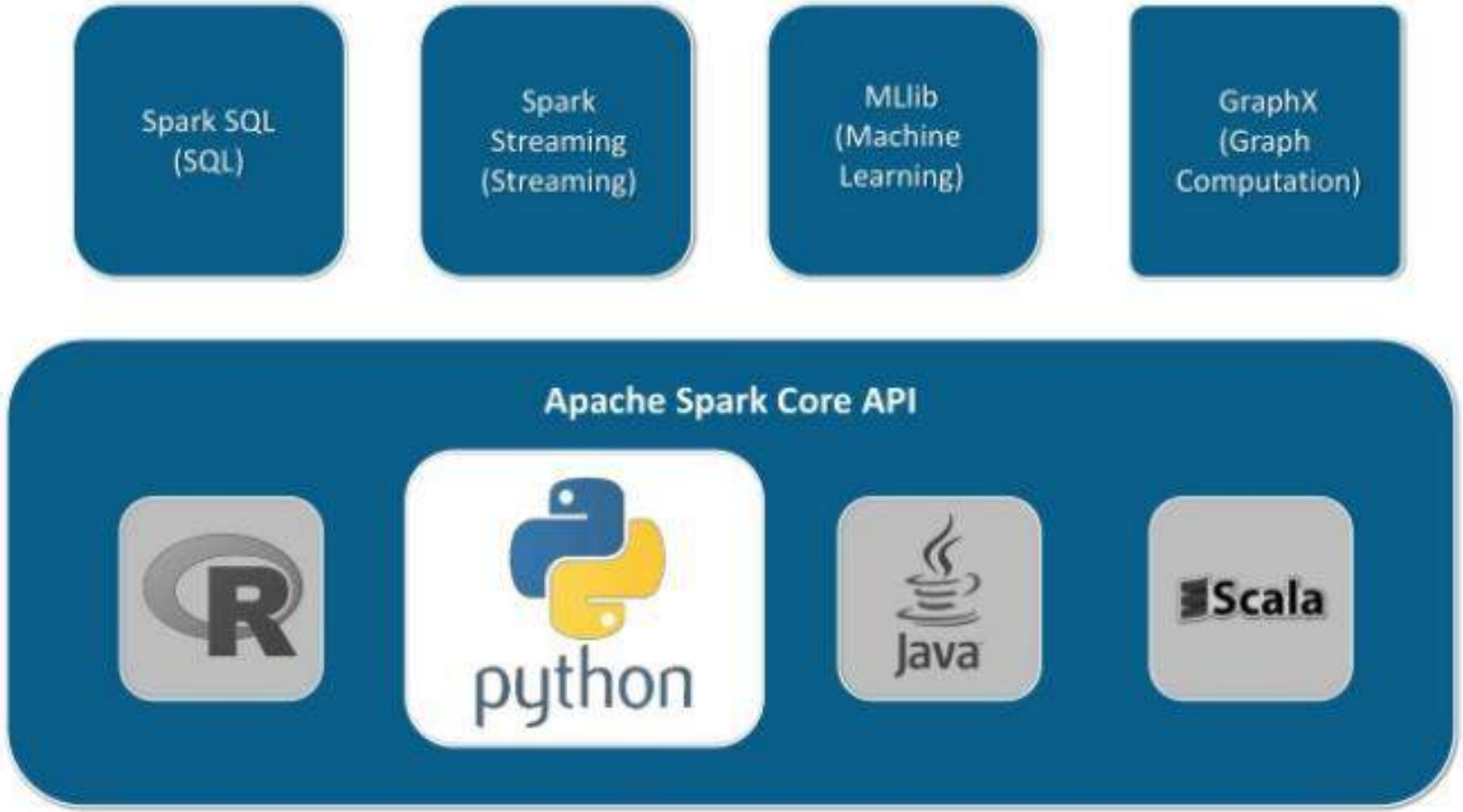
Apache Spark

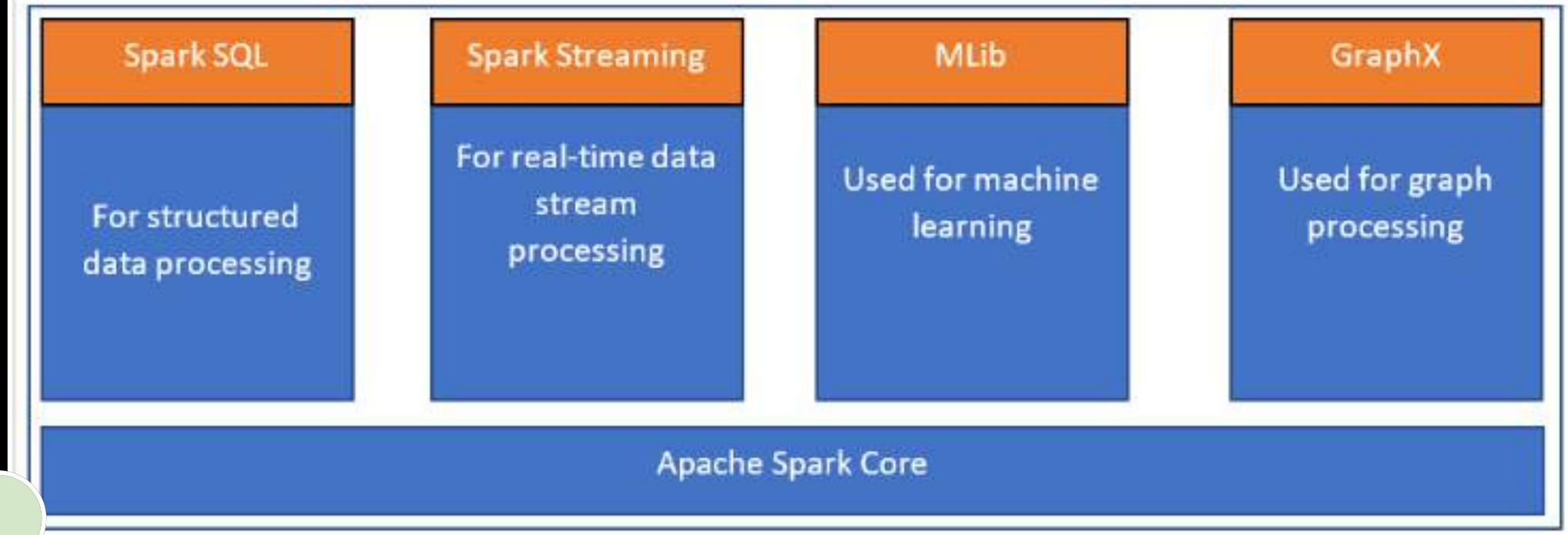
- Data processing Framework
- Designed for large scale data processing in distributed way

Pyspark def:

- Python API runs on Apache spark
- Deal with large datasets
- Used in ML like cleaning, data transformation, and data analysis
- Good choice for big data processing

Python in Spark Ecosystem

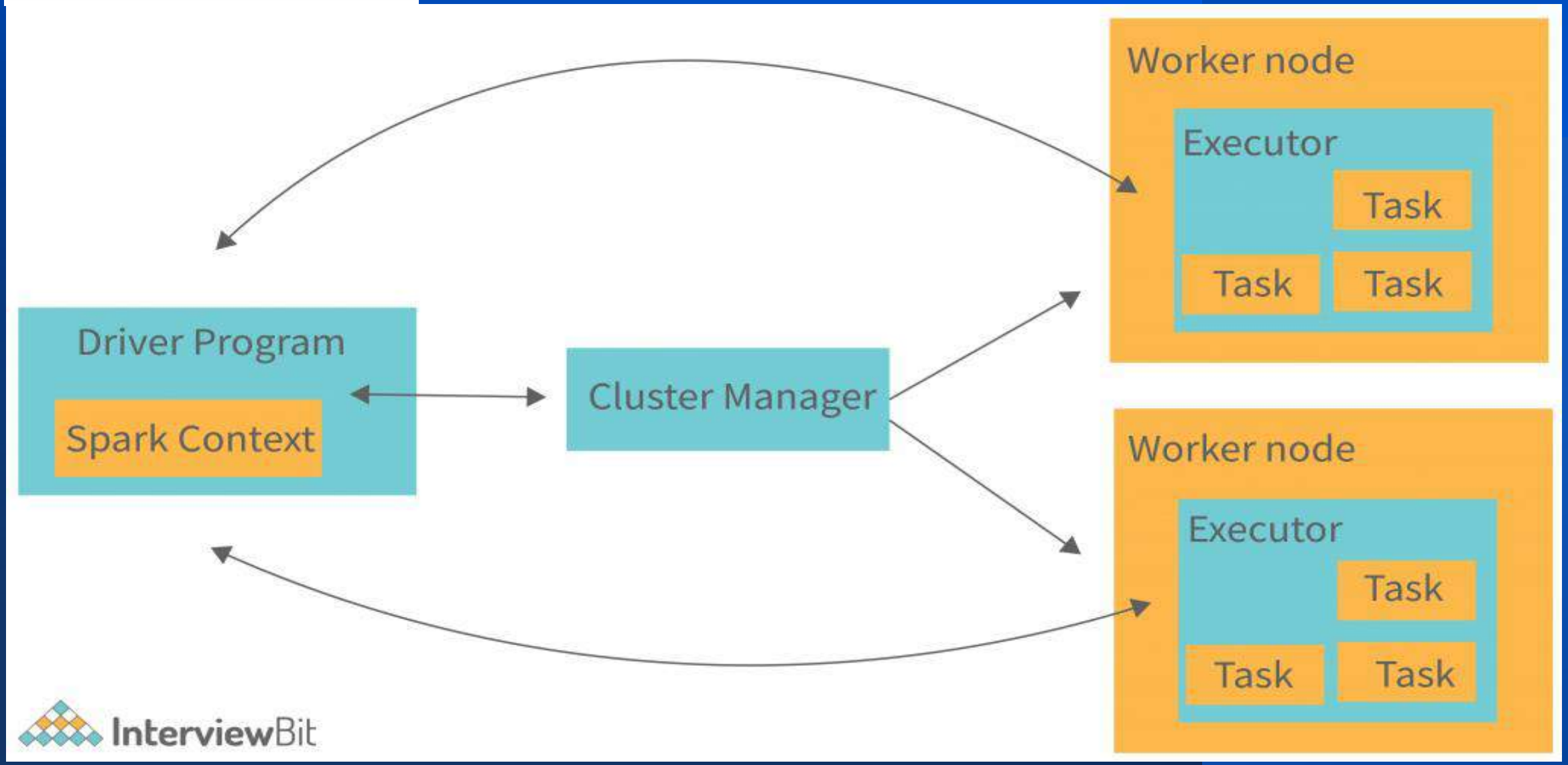



Spark SQL
For structured data processing
Spark Streaming
For real-time data stream processing
MLib
Used for machine learning
GraphX
Used for graph processing
Apache Spark Core

Pyspark

- **RDDs(**Resilient Distributed Dataset**):**
  - Data structure
  - Data portioned in such a way to ensure data processing in distributed way
  - Immutable and cannot modified
- **Data frame:** Similar to RDBMS well-structured and organized data

ARCHITECTURE



How it is better than Hadoop

- **Lazily evaluated :**
  - Doesn't Execute immediately
  - Start Execution with plan
  - Spark doesn't store all data in HDD use RAM and Hadoop store in HDD
  - Efficient due to less read write operation
  - Pre-decided and less I/O operation

Spark Session

- 1st step working with data
- Create data frame , RDD
- Apply complex operation

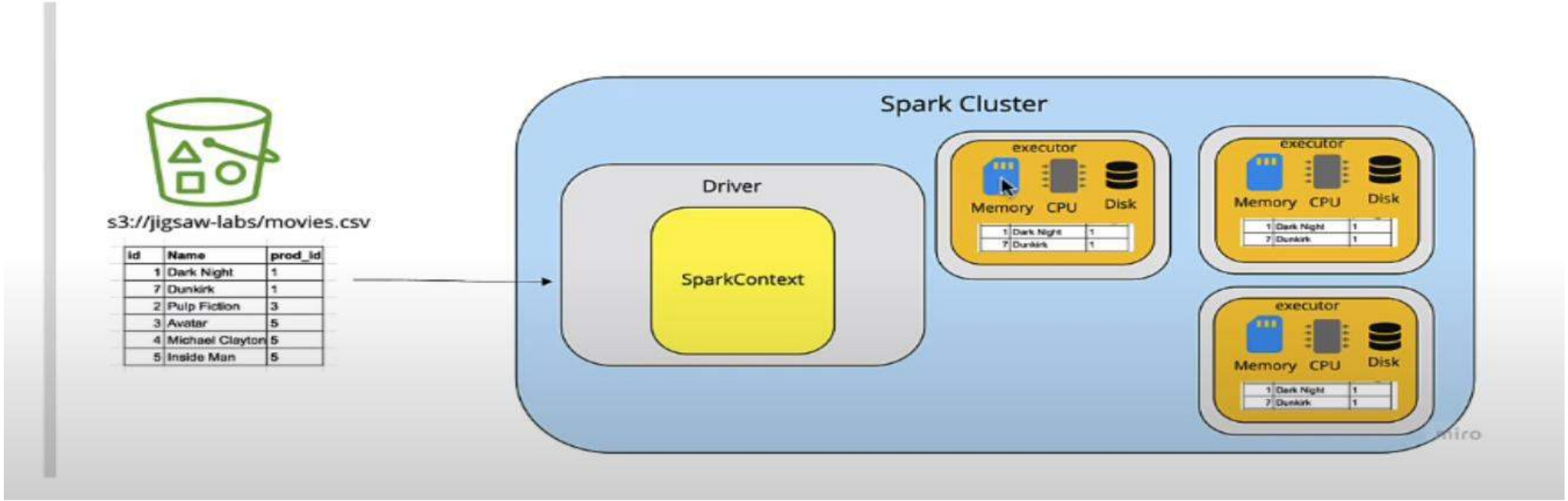
Pip install pyspark

```
(base) osama@osama-Vostro-14-3468:~$ pip install pyspark
Collecting pyspark
  Downloading pyspark-3.4.0.tar.gz (310.8 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 310.8/310.8 MB 1.4 MB/s eta 0:0
  Preparing metadata (setup.py) ... done
Collecting py4j==0.10.9.7
  Downloading py4j-0.10.9.7-py2.py3-none-any.whl (200 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 200.5/200.5 kB 2.9 MB/s eta 0:0
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.4.0-py2.py3-none-any.whl siz
1317129 sha256=7cbb8d10f03bd283c55cdf384a630007441765f03b1f1e0151e849b20929f
  Stored in directory: /home/osama/.cache/pip/wheels/db/1f/dc/108f626bfc37b2
2f400eb9cc6028ad9501cacad8dcce15
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9.7 pyspark-3.4.0
```

Word Count Ex:

```
[14]: in_text=sc.textFile('word_count.txt')

[15]: in_text.collect()

[15]: ['hi hello hi hello usama usama usama']

[16]: text_flat = in_text.flatMap(lambda x: x.split(' ')).map(lambda x: (x, 1)).reduceByKey(lambda x

[17]: text_flat.collect()

[17]: [('hi', 2), ('hello', 2), ('usama', 3)]

[ ]:
```

Pyspark parallel




THANK
YOU