UNIPI Tech Seminars: Big Data Analytics

PhD Candidate Nikos Koutroumanis

What is Apache Spark?

- Apache Spark is an open-source **parallel processing framework**

- Works with any Hadoop-supported storage system (HDFS, S3, Avro, …)

- Expressive development APIs (in multiple languages) that allows for batch processing, real-time streaming and machine learning on **Big** datasets

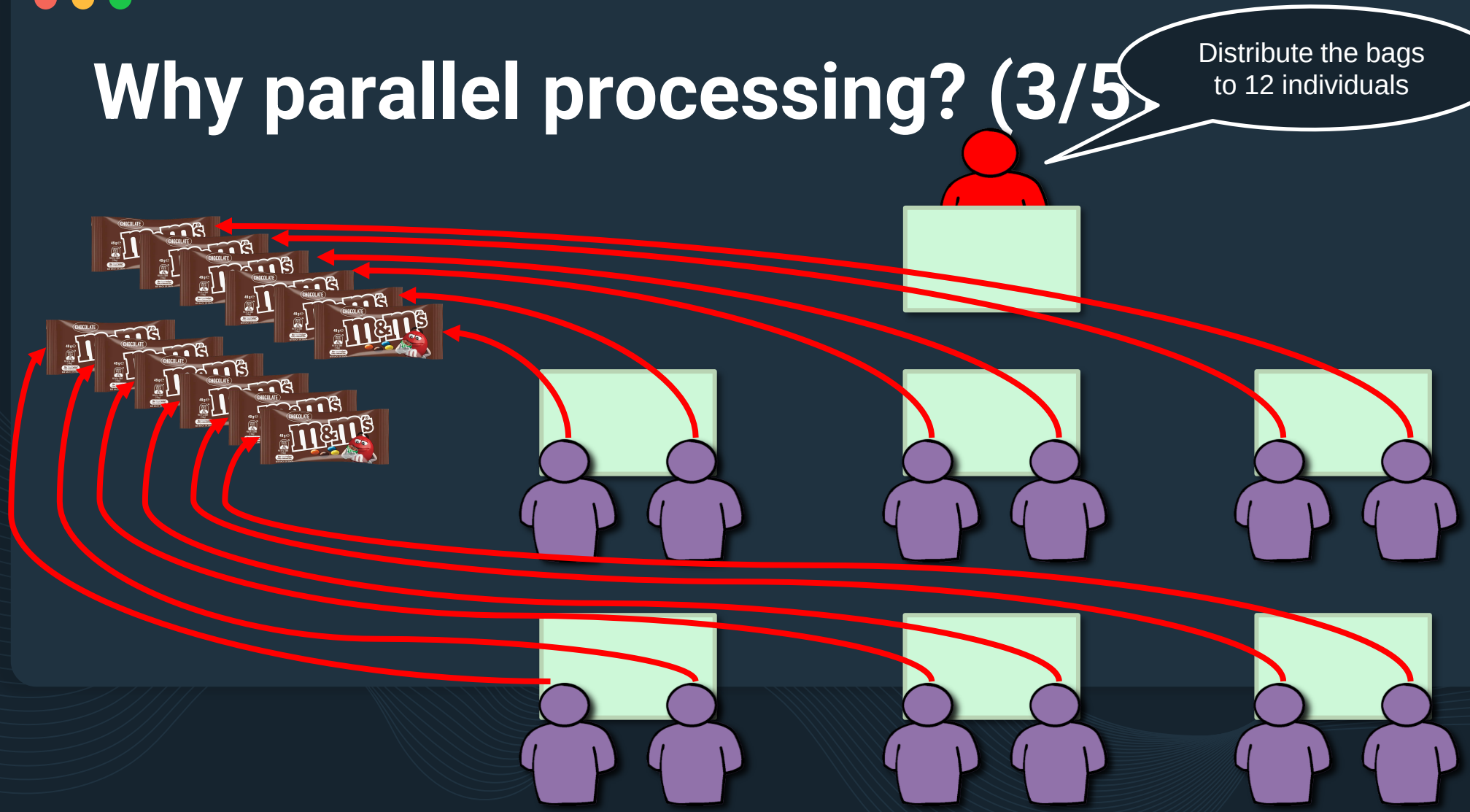
Why parallel processing? (1/5)

- Assume we have 24 packets of candies
- We want to know how many red candies exist

Why parallel processing? (2/5)

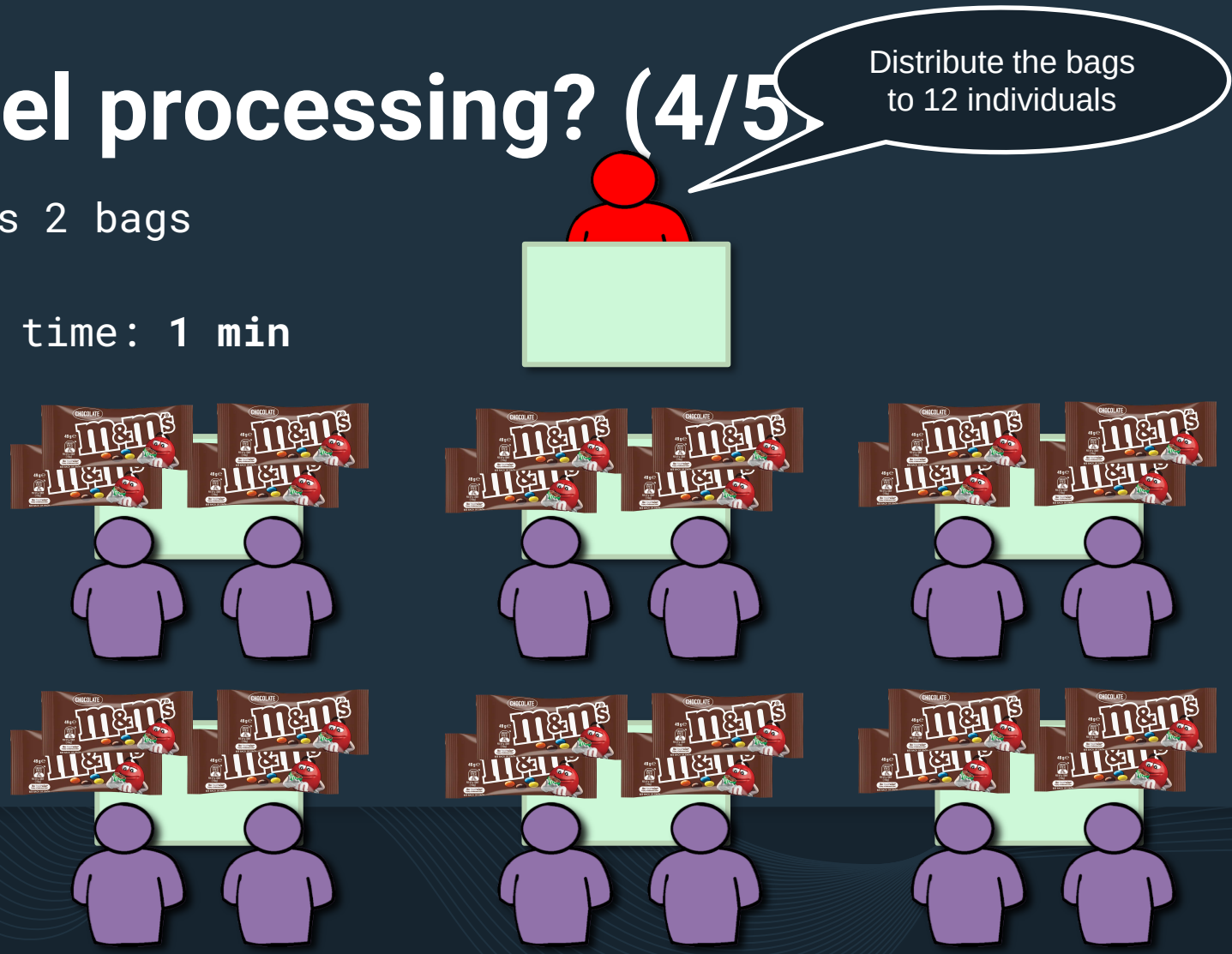
- 30 sec every bag
- Counting total time
  - 30 * 24 = 720 sec = 12 min

Why parallel processing? (3/5



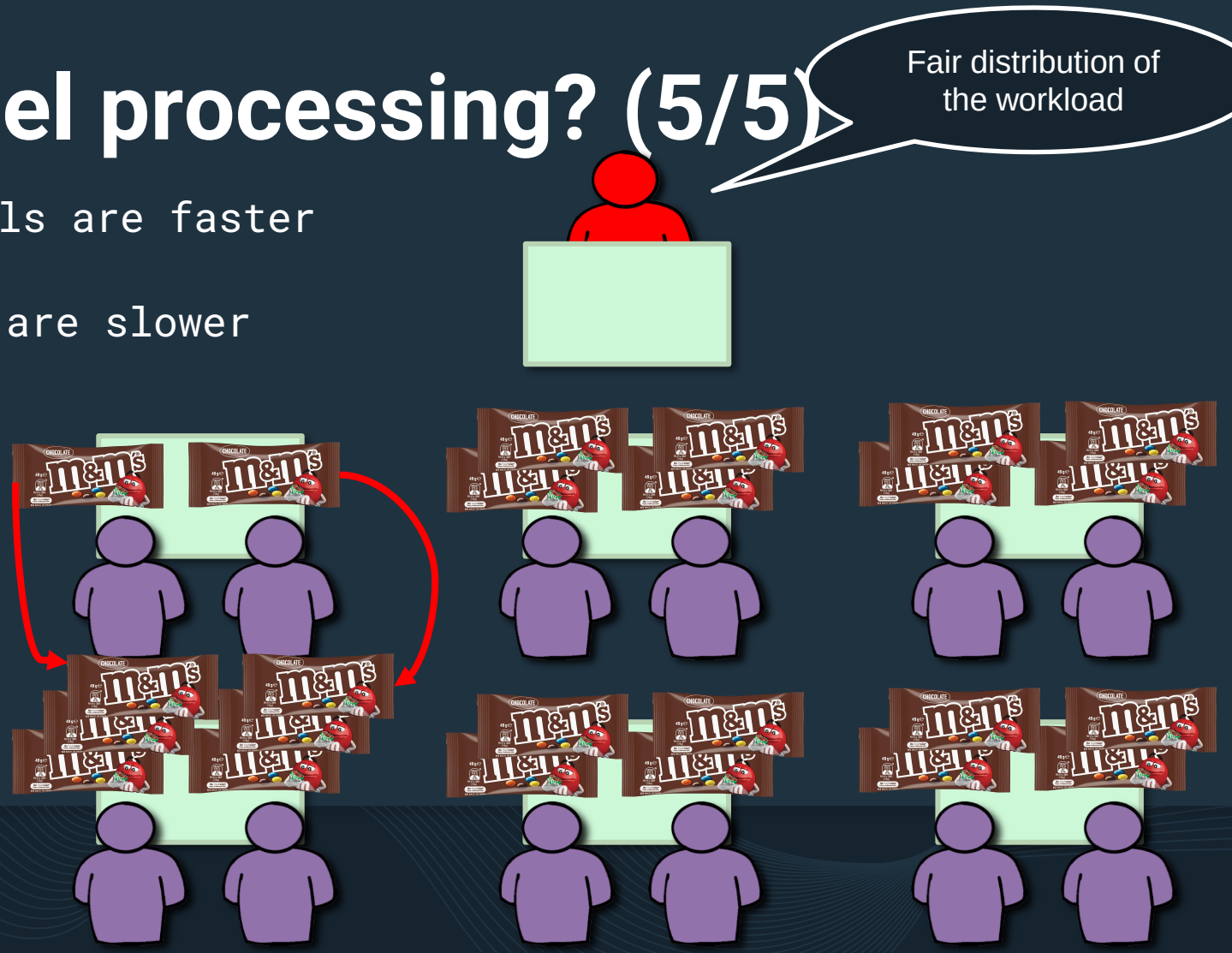
Why parallel processing? (4/5)

- Everybody takes 2 bags
- Total counting time: **1 min**

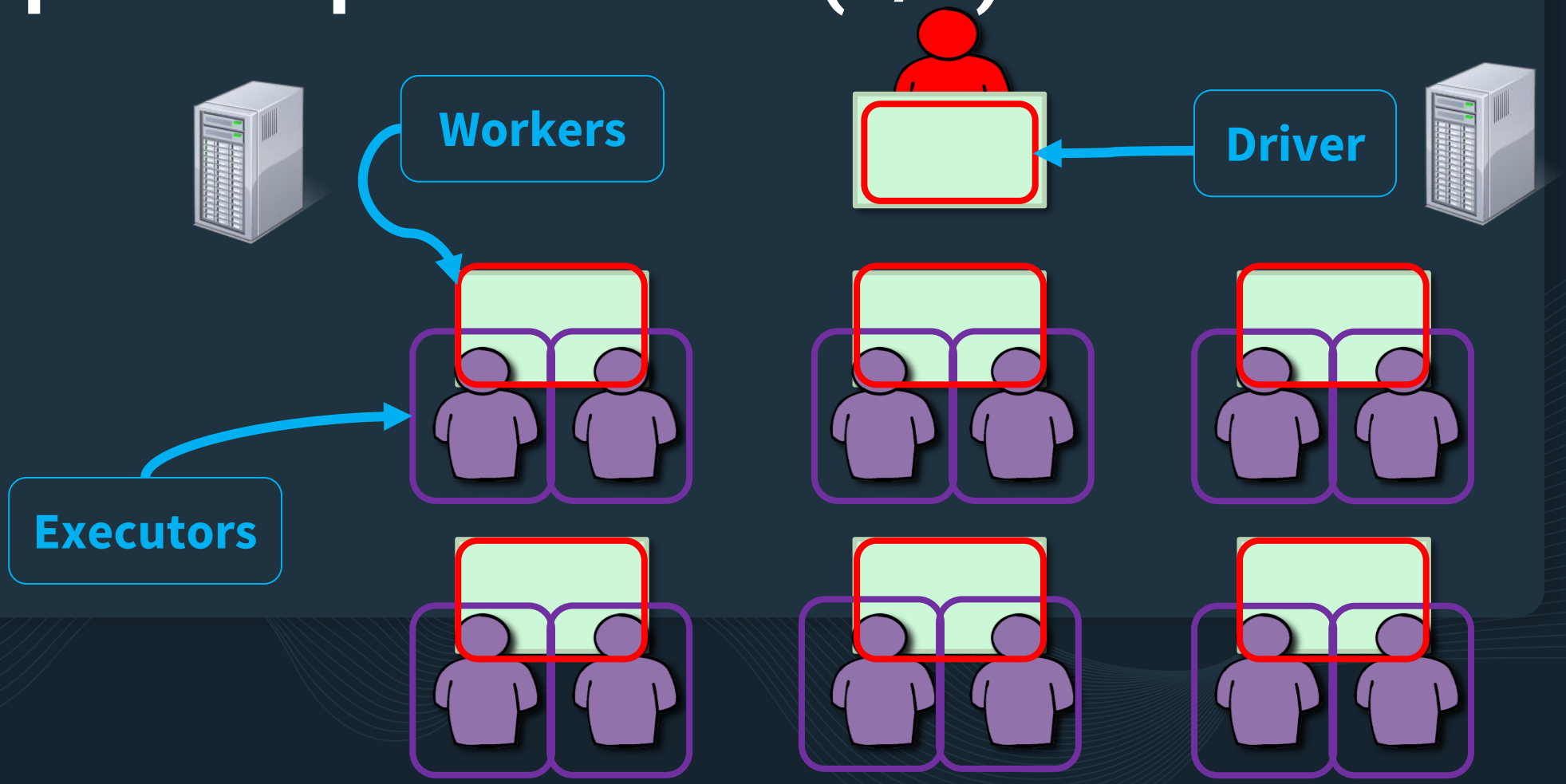


Why parallel processing? (5/5)

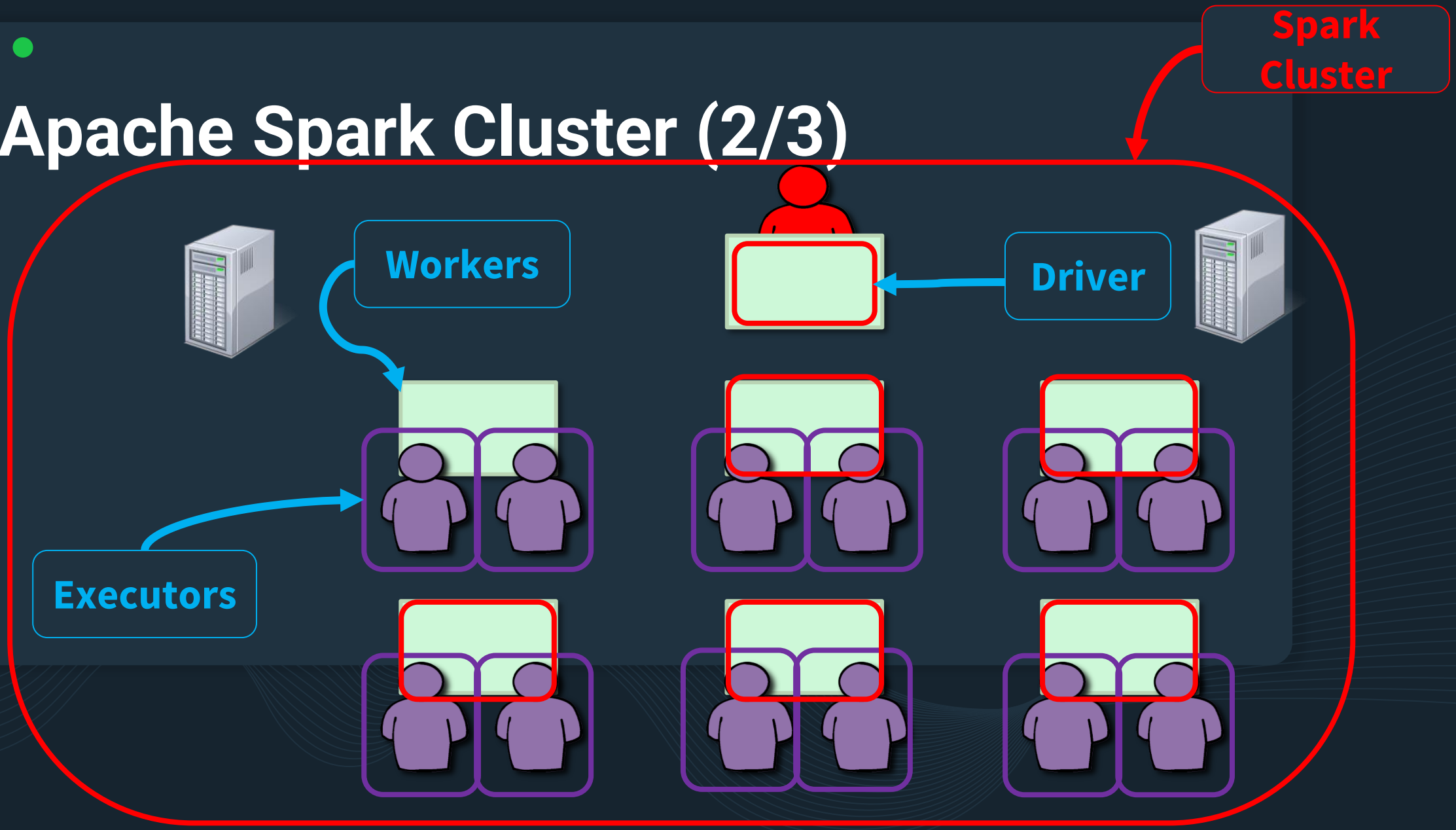
- Some individuals are faster
  - While some are slower



Apache Spark Cluster (1/3)



Apache Spark Cluster (2/3)

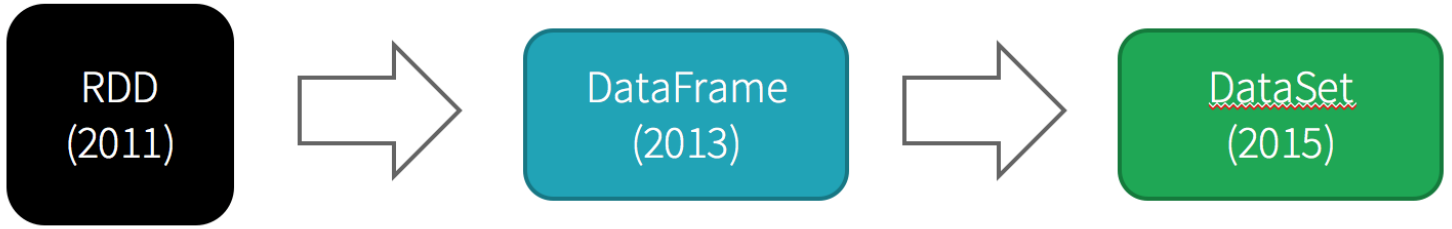


Apache Spark Cluster (3/3)

- Used with **H**adoop **D**istributed **F**ile **S**ystem – HDFS
- Hadoop is ideal for large-scale batch processing

Apache Spark APIs

History of Spark APIs



Distribute collection of JVM objects

Functional Operators (map, filter, etc.)

Distribute collection of Row objects

Expression-based operations and UDFs

Logical plans and optimizer

Fast/efficient internal representations

Internally rows, externally JVM objects

Almost the "Best of both worlds": **type safe + fast**

But slower than DF
Not as good for interactive analysis, especially Python

databricks

Dataframe

What is a Dataframe?

A distributed collection of data organized into named columns, similar to a table in a relational database.

Use Cases: Data manipulation and aggregation, ETL operations, SQL queries.

Features:

- Optimized execution through Catalyst optimizer.
- Support various data sources (JSON, Parquet, JDBC, etc.).
- High-level abstraction.

Dataframe Operations

Transformations

- **Select**
- **Filter**
- **GroupBy**
- **Aggregate**
- **Join**
- **WithColumn**
- **Drop**

Create a new DataFrame from an existing one

Actions

- **Show**
- **Collect**
- **Count**
- **Take**
- **Write**

Trigger the execution of transformations and return results to the driver program or write data to external storage

Visit the link

https://gist.github.com/nkoutroumanis

Copy and paste the three displayed links

https://pithos.okeanos.grnet.gr/public/Kj8RBwX20MzC3SsKV6wOl2

https://pithos.okeanos.grnet.gr/public/ow3M82qHWgwGkbsX8KDAn6

https://pithos.okeanos.grnet.gr/public/Kx8dj2N00JtmJWAUSOTSO3

Visit Databricks Platform

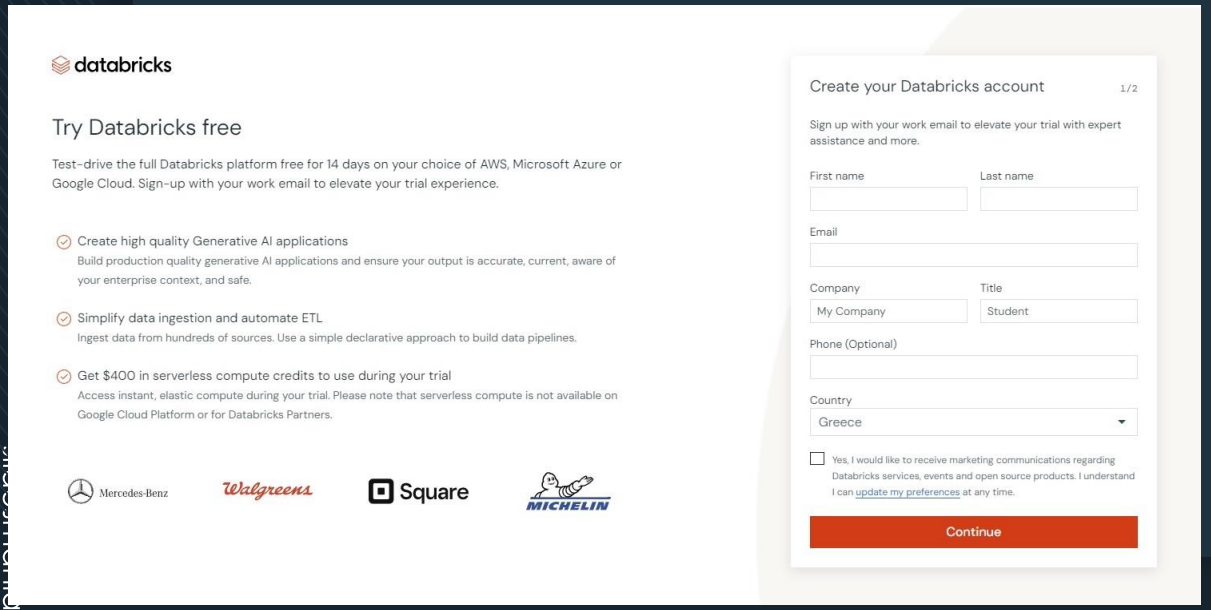
https://community.cloud.databricks.com

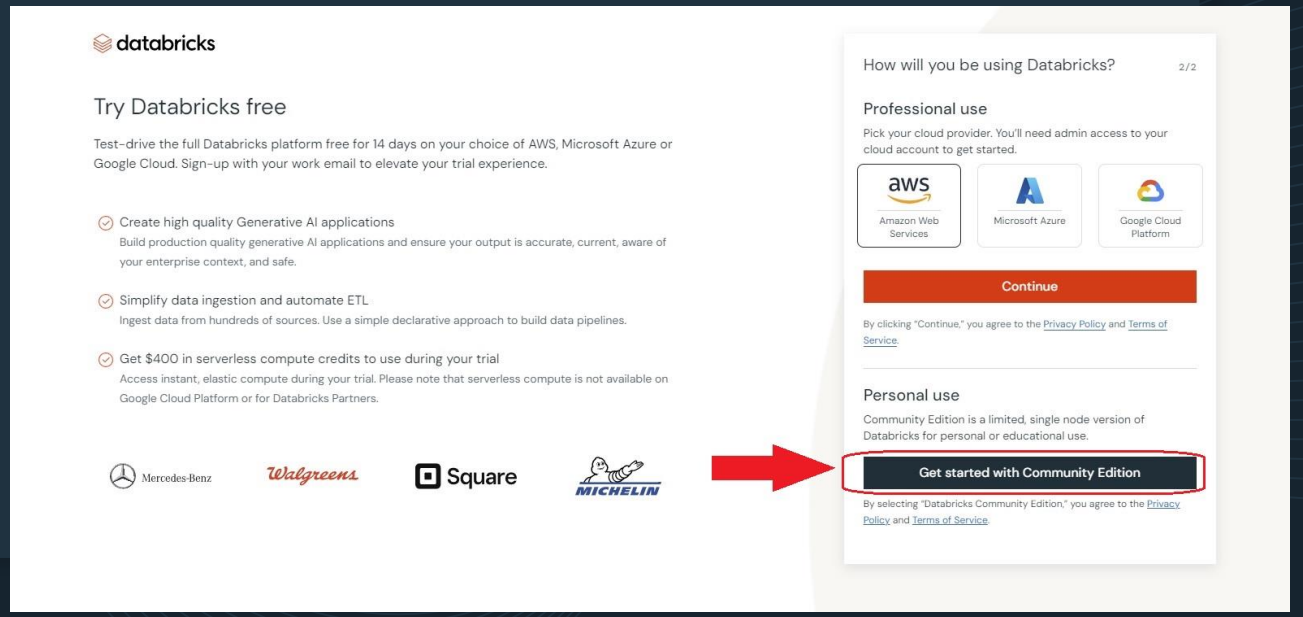
databricks

Register to Databricks

To register:

https://www.databricks.com/try-databricks#account





THANK YOU!

Do you have any questions?

koutroumanis@unipi.gr