



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)

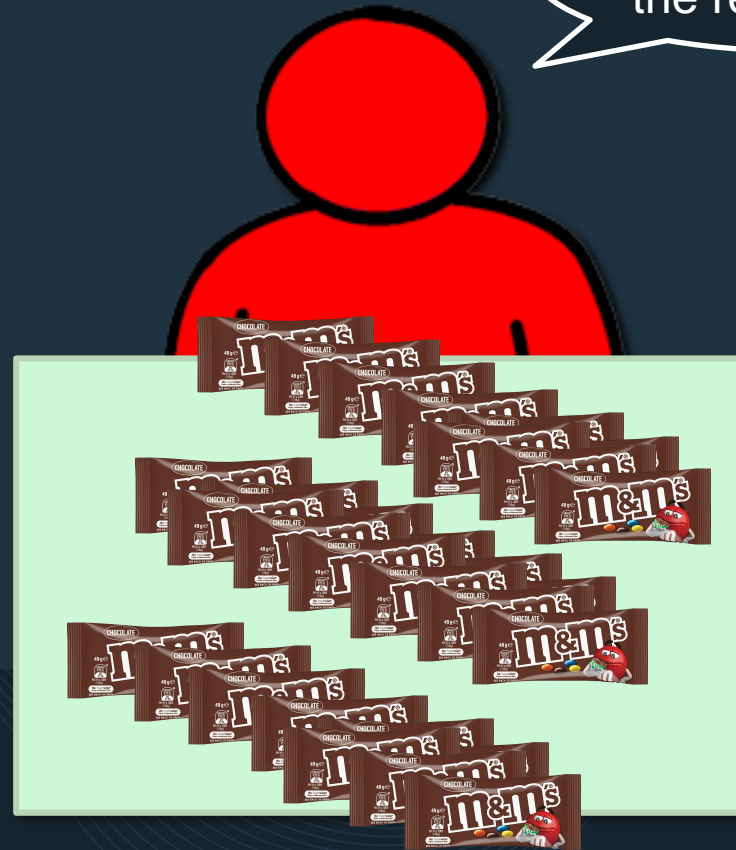
- Assume we have 24 packets of candies
- We want to know how many red candies exist



Why parallel processing? (2/5)

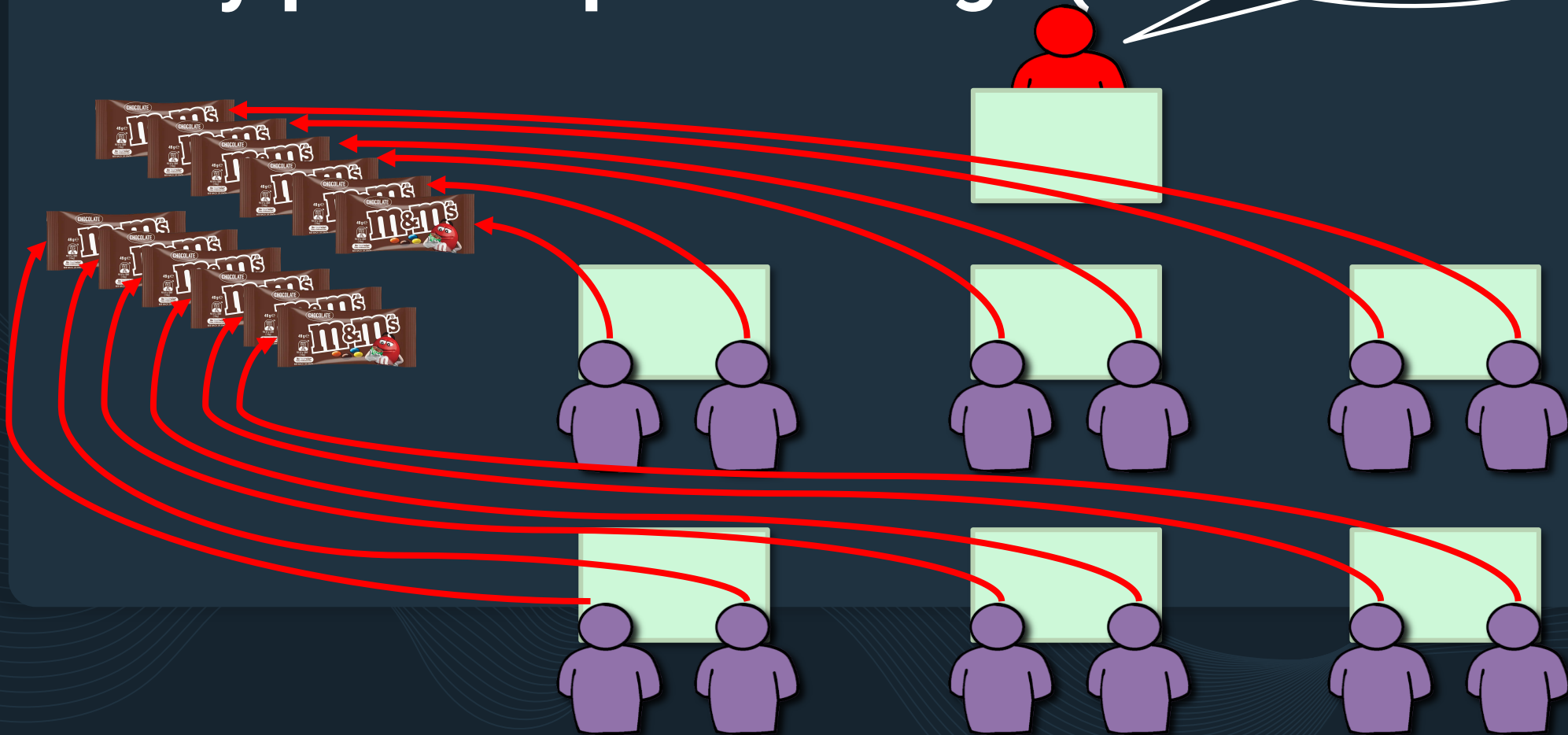
- 30 sec every bag
- Counting total time
 - $30 * 24 = 720 \text{ sec} = 12 \text{ min}$

Start counting
the red candies



Why parallel processing? (3/5)

Distribute the bags to 12 individuals



Why parallel processing? (4/5)

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

Distribute the bags to 12 individuals



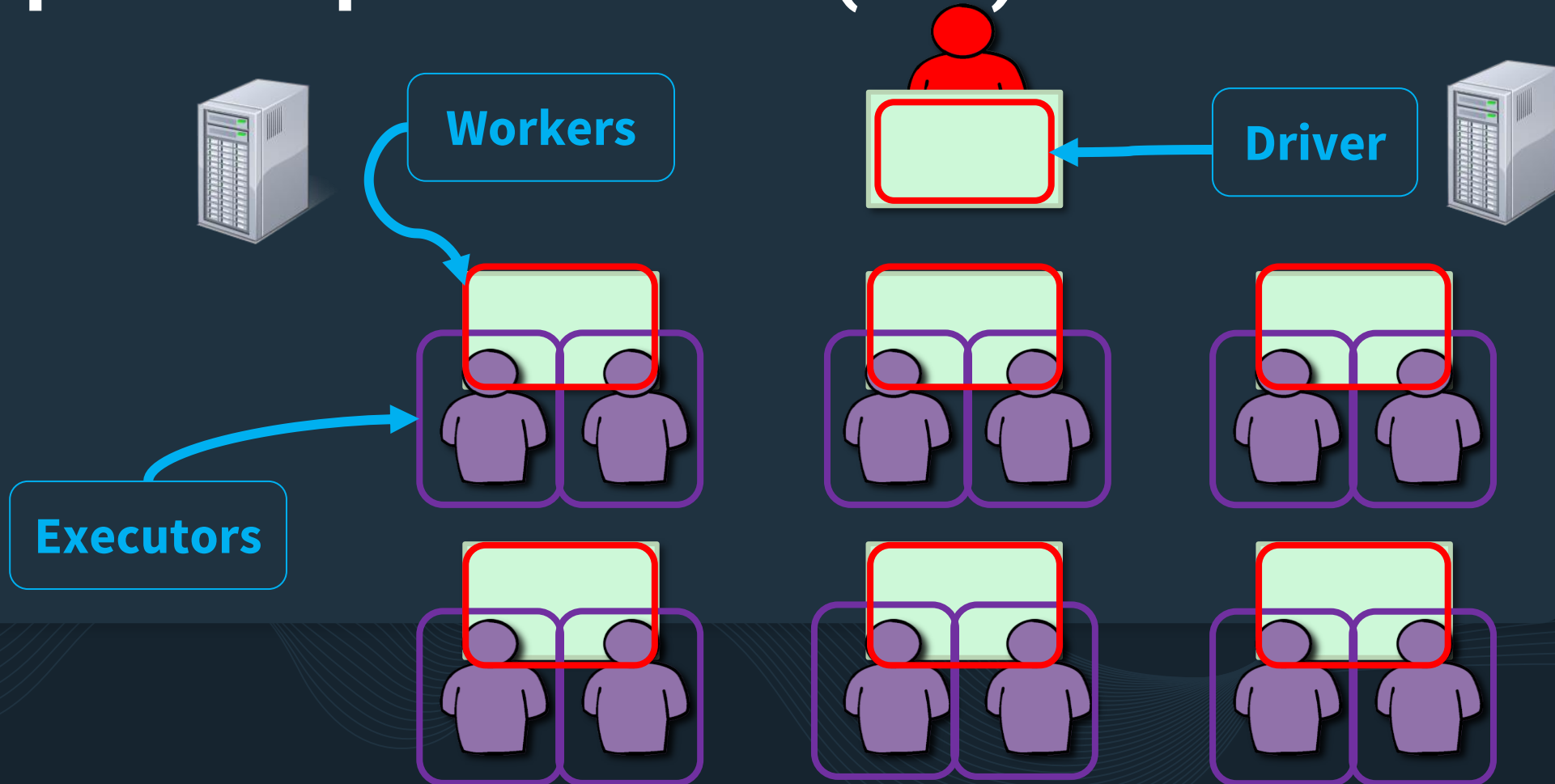
Why parallel processing? (5/5)

- Some individuals are faster
 - While some are slower

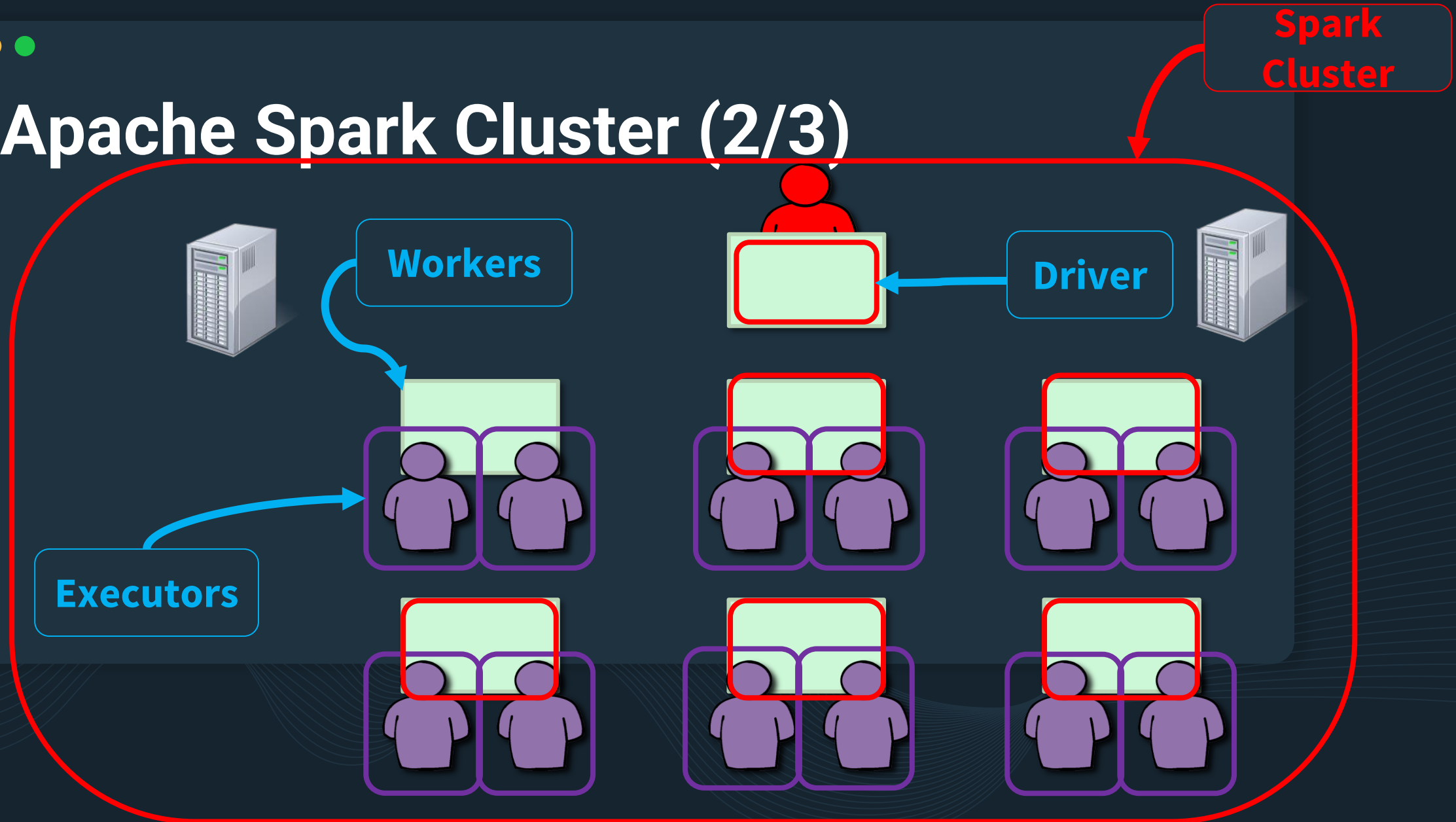
Fair distribution of the workload



Apache Spark Cluster (1/3)



Apache Spark Cluster (2/3)



Apache Spark Cluster (3/3)

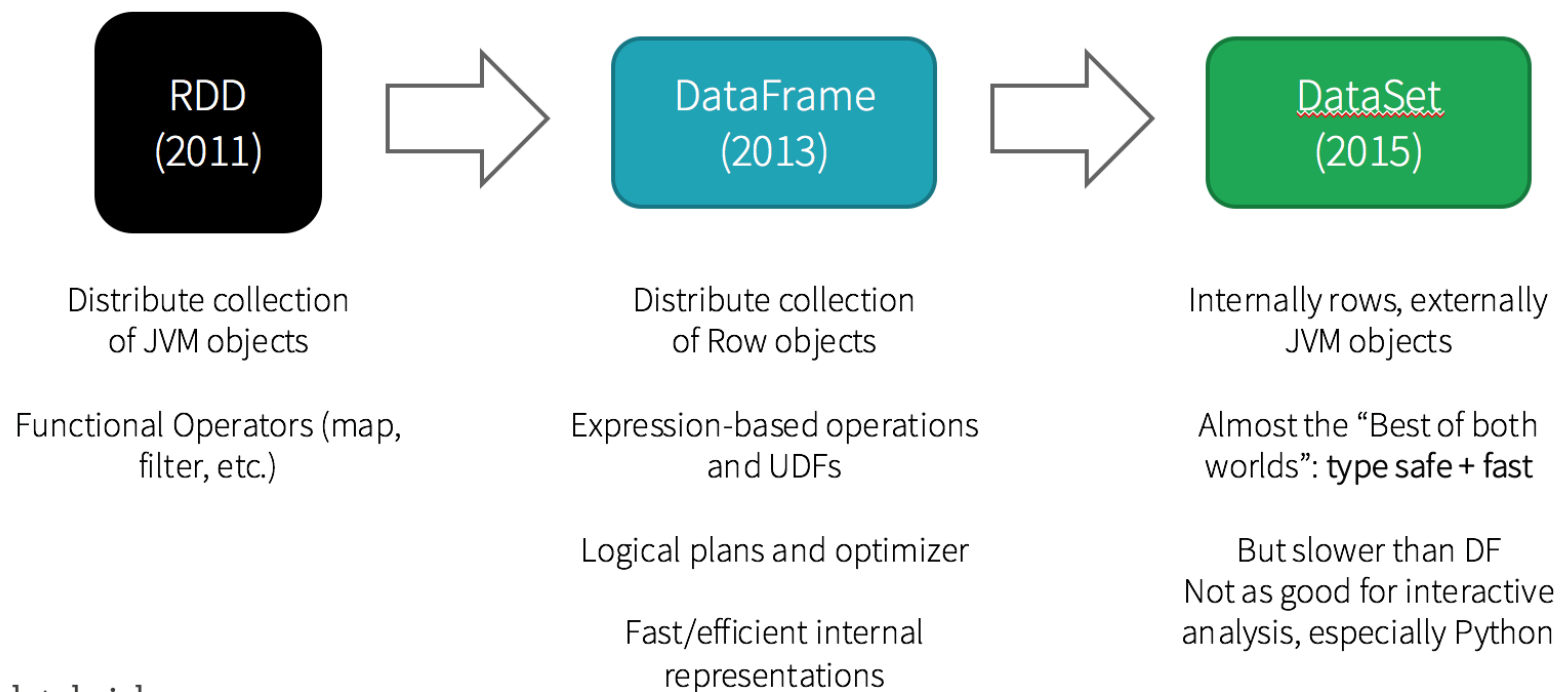
- Used with Hadoop Distributed File System – HDFS
- Hadoop is ideal for large-scale batch processing

hdfs://user/myfolder/...



Apache Spark APIs

History of Spark APIs



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/Kj8RBwX20MzC3SsKV6w0l2
```

```
https://pithos.okeanos.grnet.gr/public/ow3M82qHWgwGkbsX8KDA6
```

```
https://pithos.okeanos.grnet.gr/public/Kx8dj2N00JtmJWAUS0TS03
```


Visit Databricks Platform



<https://community.cloud.databricks.com>



databricks

Register to Databricks

To register:

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



Try Databricks free

Test-drive the full Databricks platform free for 14 days on your choice of AWS, Microsoft Azure or Google Cloud. Sign-up with your work email to elevate your trial experience.

- ☒ Create high quality Generative AI applications
Build production quality generative AI applications and ensure your output is accurate, current, aware of your enterprise context, and safe.
- ☒ Simplify data ingestion and automate ETL
Ingest data from hundreds of sources. Use a simple declarative approach to build data pipelines.
- ☒ Get \$400 in serverless compute credits to use during your trial
Access instant, elastic compute during your trial. Please note that serverless compute is not available on Google Cloud Platform or for Databricks Partners.



Mercedes-Benz



Create your Databricks account

1/2

Sign up with your work email to elevate your trial with expert assistance and more.

First name Last name

Email

Company Title

Phone (Optional)

Country

☐ Yes, I would like to receive marketing communications regarding Databricks services, events and open source products. I understand I can update my preferences at any time.

Continue



Try Databricks free

Test-drive the full Databricks platform free for 14 days on your choice of AWS, Microsoft Azure or Google Cloud. Sign-up with your work email to elevate your trial experience.

- ☒ Create high quality Generative AI applications
Build production quality generative AI applications and ensure your output is accurate, current, aware of your enterprise context, and safe.
- ☒ Simplify data ingestion and automate ETL
Ingest data from hundreds of sources. Use a simple declarative approach to build data pipelines.
- ☒ Get \$400 in serverless compute credits to use during your trial
Access instant, elastic compute during your trial. Please note that serverless compute is not available on Google Cloud Platform or for Databricks Partners.



Mercedes-Benz



Square



MICHELIN

How will you be using Databricks?

2/2

Professional use

Pick your cloud provider. You'll need admin access to your cloud account to get started.



Continue

By clicking "Continue," you agree to the [Privacy Policy](#) and [Terms of Service](#).

Personal use

Community Edition is a limited, single node version of Databricks for personal or educational use.

Get started with Community Edition

By selecting "Databricks Community Edition," you agree to the [Privacy Policy](#) and [Terms of Service](#).



THANK YOU!

Do you have any questions?



koutroumanis@unipi.gr