

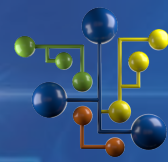
Data Science
Academy

Data Science Academy leandrobologna@hotmail.com 581c5bb75e4cde1fb58b4581

Machine Learning e IA em Ambientes Distribuídos



Data Science Academy



Data Science
Academy

Data Science Academy leandrobologna@hotmail.com 581c5bb75e4cde1fb58b4581

Machine Learning em Larga Escala

Parte 2

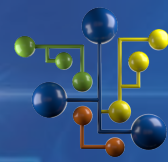


Data Science Academy



Machine Learning em Larga Escala – Parte 2





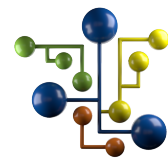
Data Science
Academy

Data Science Academy leandrobologna@hotmail.com 581c5bb75e4cde1fb58b4581

Componentes do Apache Spark



Data Science Academy



Componentes do Apache Spark

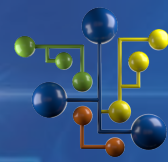
Resilient Distributed Datasets – API que oferece a funcionalidade para realizar operações em um ambiente distribuído.

SQL, DataFrames, Datasets – Interfaces para operação de dados estruturados.

Streaming (Dstreams) – Processamento de dados em tempo real.

Machine Learning Library (MLlib) – Coleção de modelos de Machine Learning.

GraphX – Computação integrada de grafos em Paralelo.



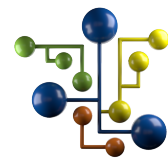
Data Science
Academy

Data Science Academy leandrobologna@hotmail.com 581c5bb75e4cde1fb58b4581

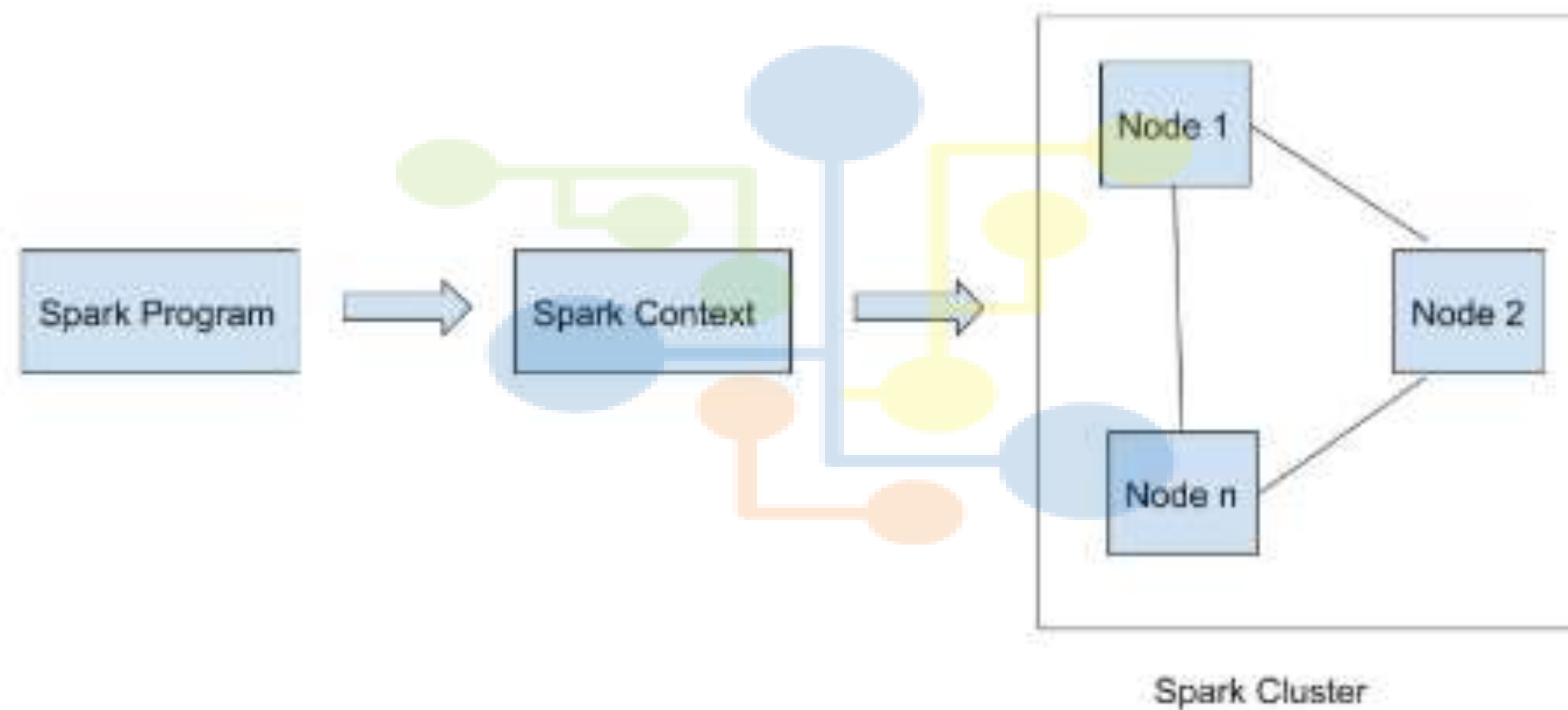
Modos de Execução do Spark: Standalone, Apache Mesos, Apache YARN e Kubernetes

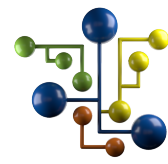


Data Science Academy



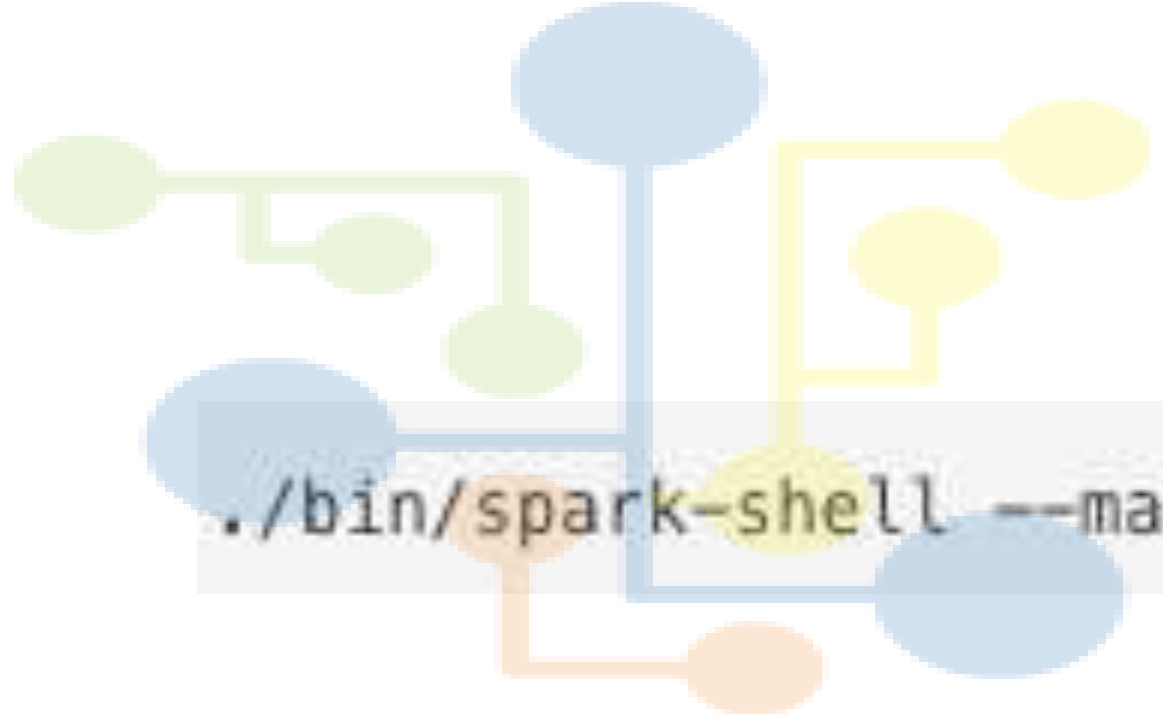
Modos de Execução do Spark: Standalone, Apache Mesos, Apache YARN e Kubernetes



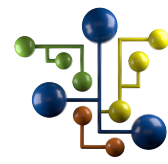


Modos de Execução do Spark: Standalone, Apache Mesos, Apache YARN e Kubernetes

Standalone

A diagram showing a network of nodes connected by lines, representing a distributed system. The nodes are colored in blue, green, yellow, and orange, and are connected in a complex, branching structure.

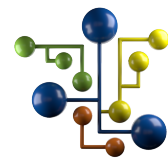
```
./bin/spark-shell --master spark://IP:PORT
```

Modos de Execução do Spark: Standalone, Apache Mesos, Apache YARN e Kubernetes

Apache Mesos

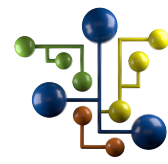
```
./bin/spark-submit \  
--class org.apache.spark.examples.SparkPi \  
--master mesos://207.184.161.138:7077 \  
--deploy-mode cluster \  
--supervise \  
--executor-memory 20G \  
--total-executor-cores 100 \  
http://path/to/examples.jar \  
1000
```



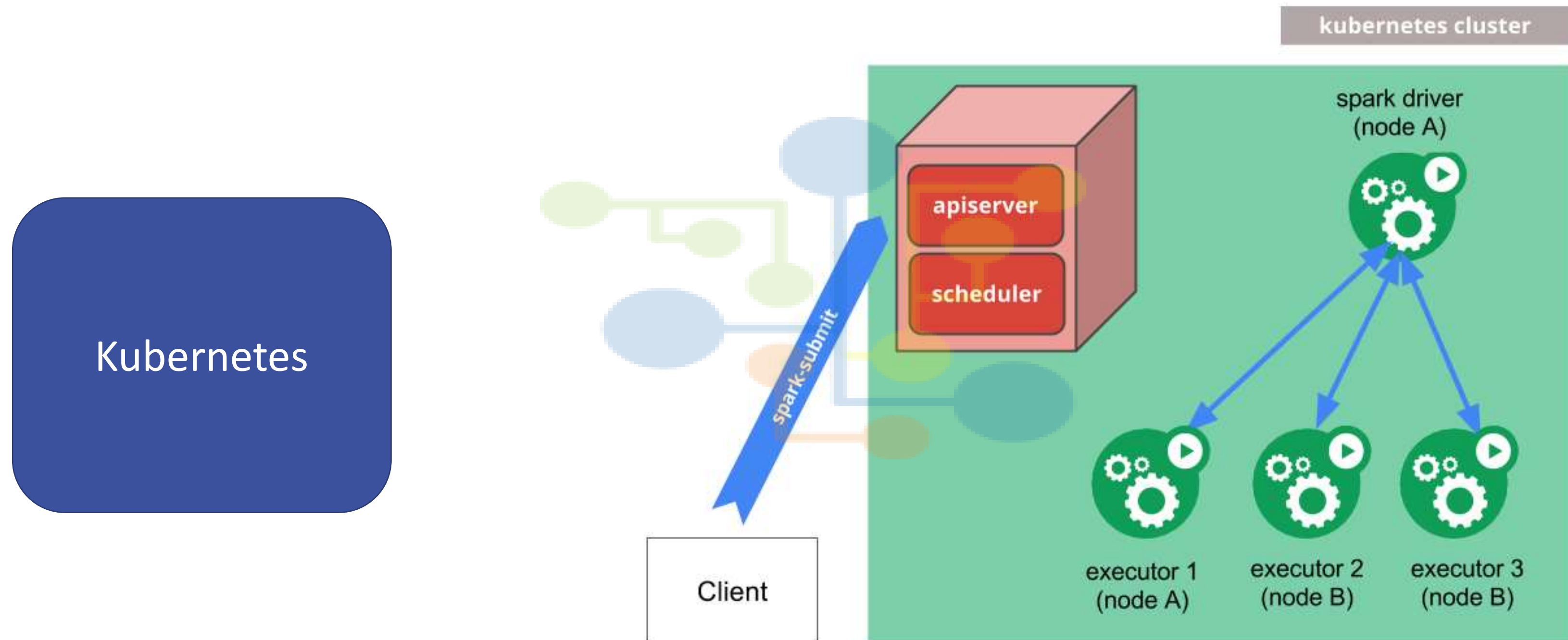
Modos de Execução do Spark: Standalone, Apache Mesos, Apache YARN e Kubernetes

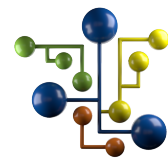
Apache YARN

```
$ ./bin/spark-submit --class org.apache.spark.examples.SparkPi \  
--master yarn \  
--deploy-mode cluster \  
--driver-memory 4g \  
--executor-memory 2g \  
--executor-cores 1 \  
--queue thequeue \  
examples/jars/spark-examples*.jar \  
10
```

Modos de Execução do Spark: Standalone, Apache Mesos, Apache YARN e Kubernetes

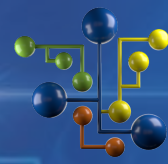




Modos de Execução do Spark: Standalone, Apache Mesos, Apache YARN e Kubernetes

Kubernetes

```
$ bin/spark-submit \  
  --master k8s://https://<k8s-apiserver-host>:<k8s-apiserver-port> \  
  --deploy-mode cluster \  
  --name spark-pi \  
  --class org.apache.spark.examples.SparkPi \  
  --conf spark.executor.instances=5 \  
  --conf spark.kubernetes.container.image=<spark-image> \  
  local:///path/to/examples.jar
```

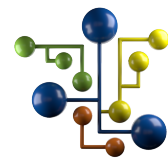
Data Science
Academy

Data Science Academy leandrobologna@hotmail.com 581c5bb75e4cde1fb58b4581

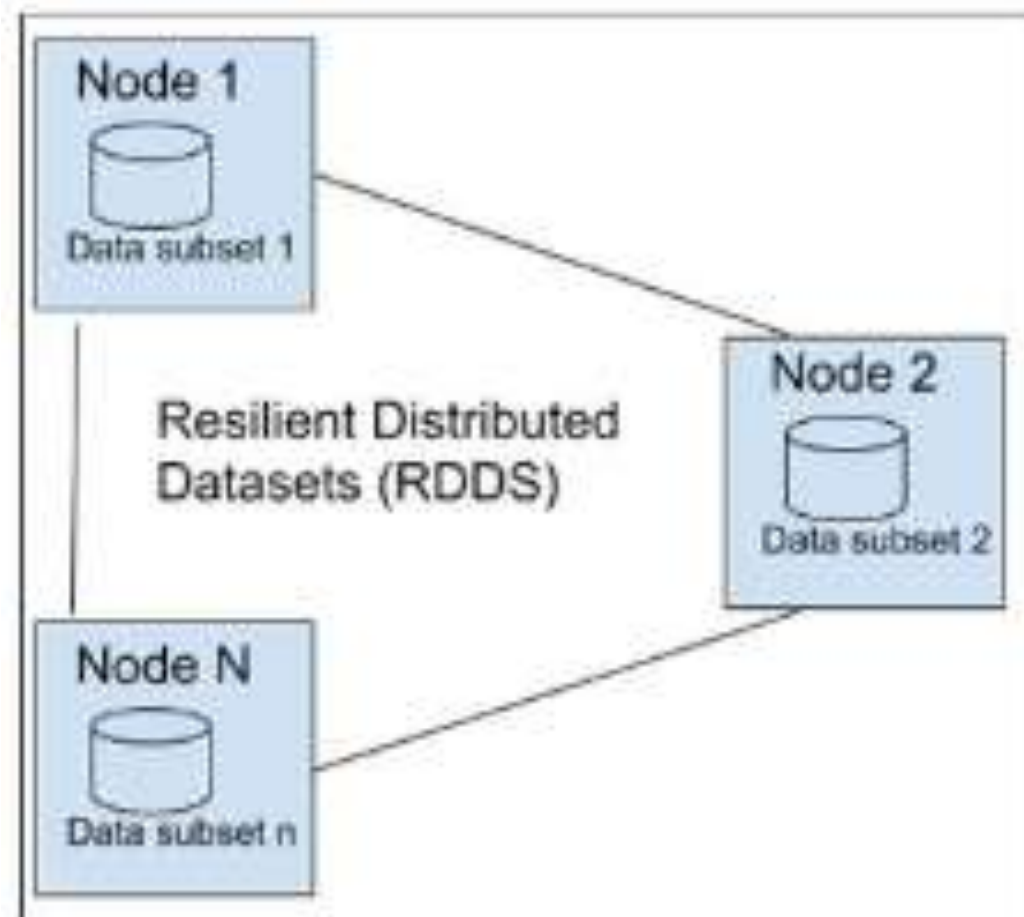
Resilient Distributed Datasets



Data Science Academy



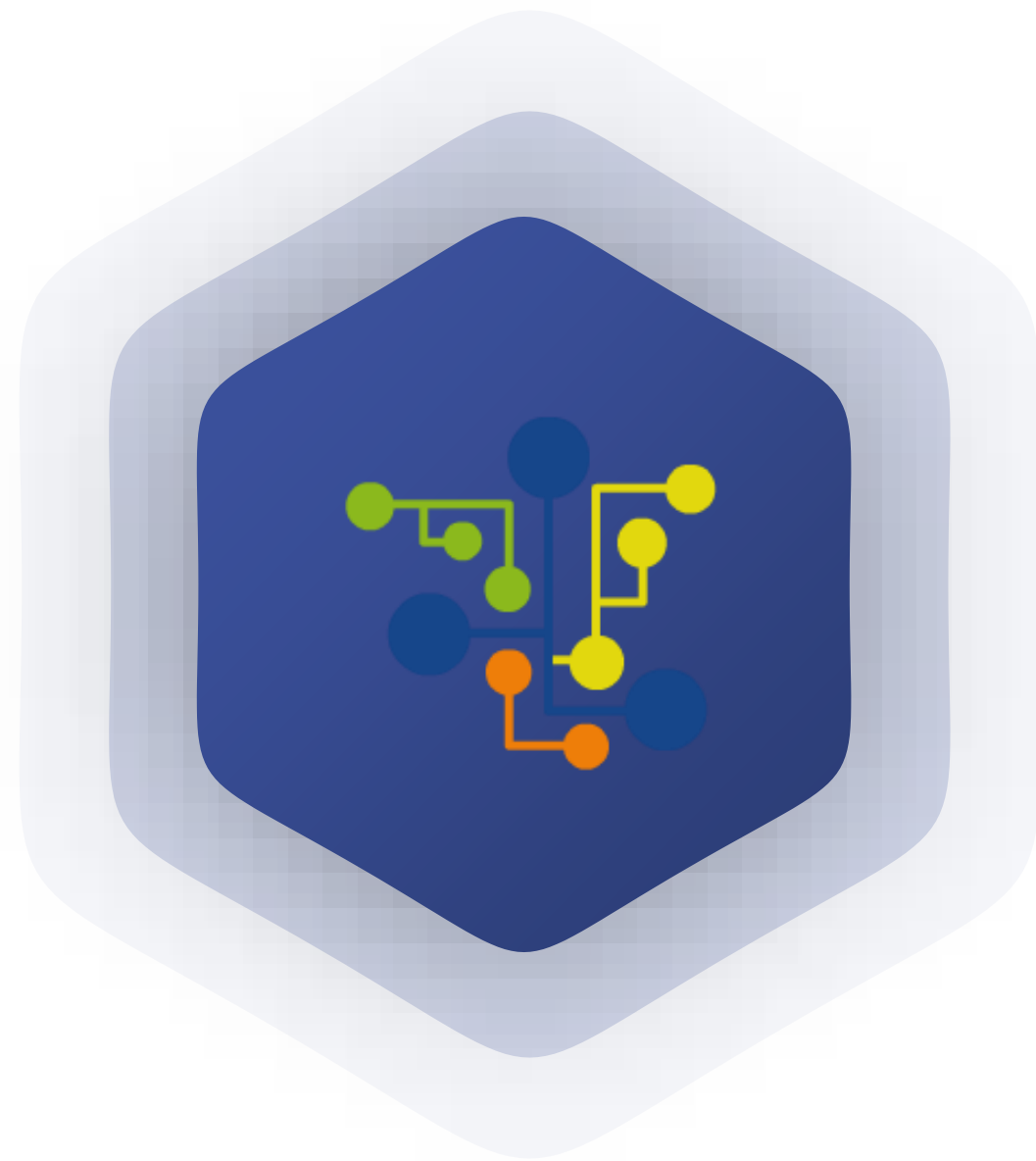
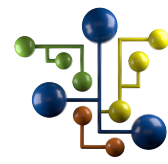
Resilient Distributed Datasets



Aplicações Spark executam código do analista e realizam operações paralelas em um ambiente distribuído.

As RDDs são uma abstração que oculta a complexidade por trás da computação distribuída, como consistência de estados e recuperação de falha dos nós do cluster.

As RDDs representam coleções de elementos que são distribuídos através de diferentes nós de um cluster.



Muito Obrigado.

É um prazer ter você aqui.

Tenha uma excelente jornada de aprendizagem.