# Introduction to Spark

Lambda World

# Who are we?

## Juan Pedro Moreno

Scala Software Engineer at 47Degrees
@juanpedromoreno

## Fran Pérez

Scala Software Engineer at 47Degrees
@FPerezP

Workshop repo: https://github.com/47deg/spark-workshop

# Roadmap

- Intro Big Data and Spark
- Spark Architecture
- Resilient Distributed Datasets (RDDs)
- Transformations and Actions on Data using RDDs
- Overview Spark SQL and DataFrames
- Overview Spark Streaming
- Spark Architecture and Cluster Deployment

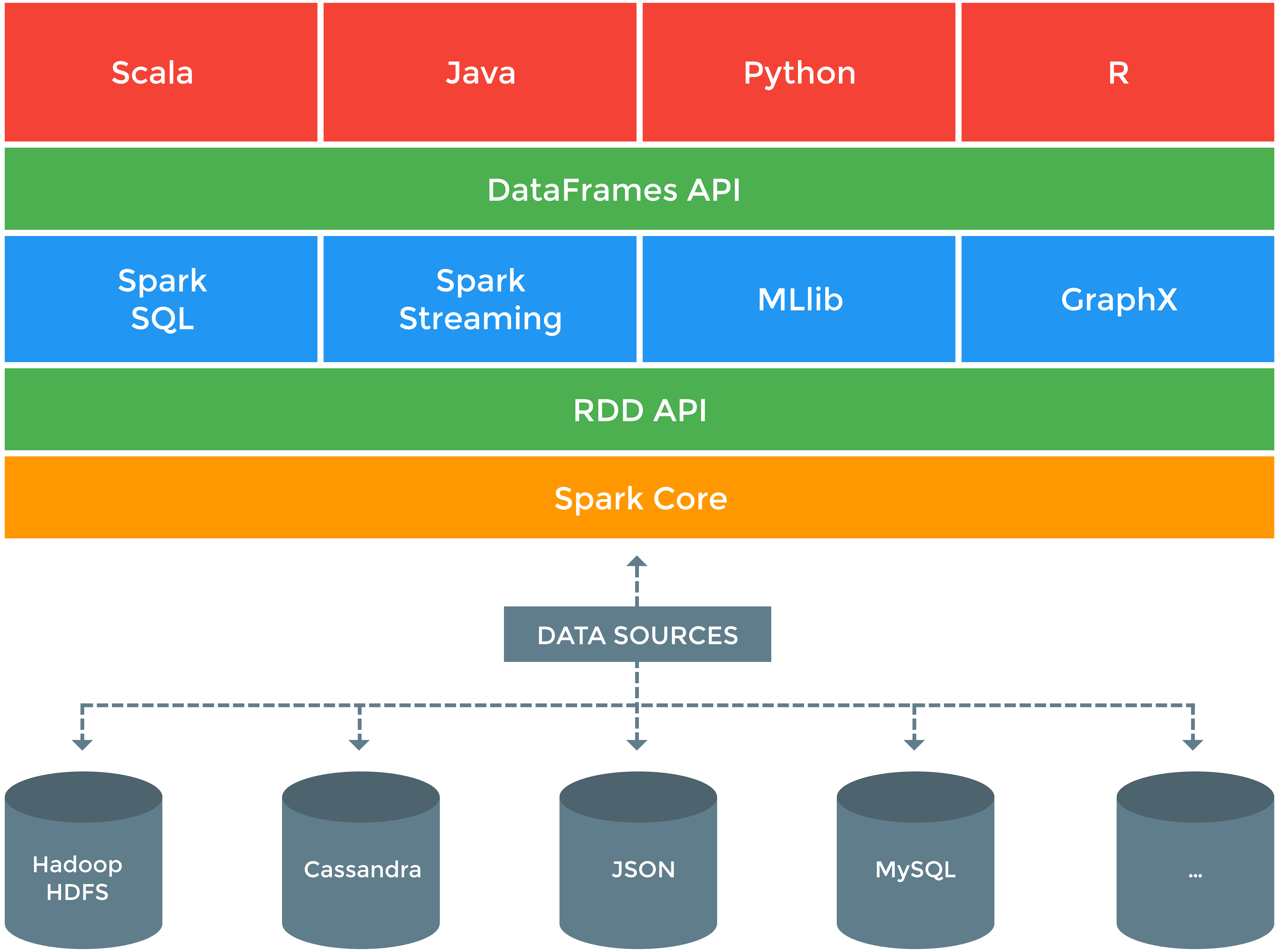
# Apache Spark Overview

- Fast and general engine for large-scale data processing
- Speed
- Ease of Use
- Generality
- Runs Everywhere

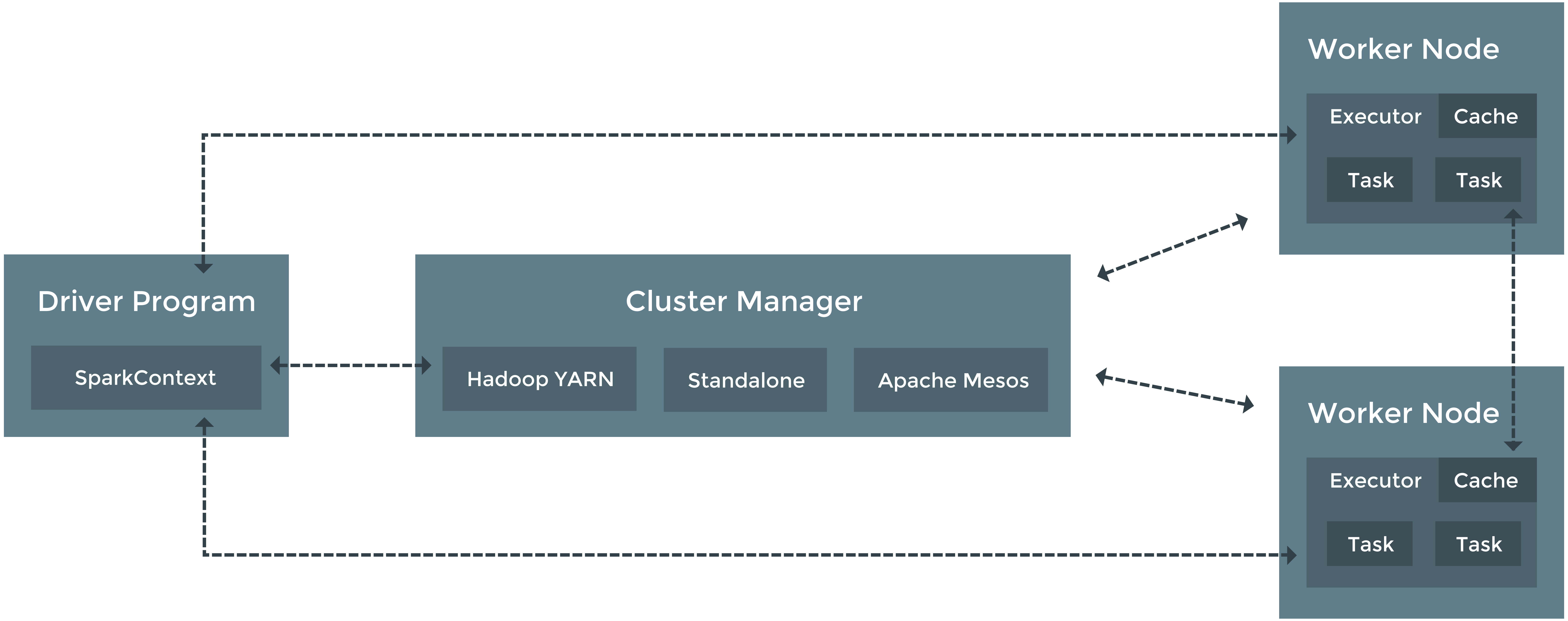
http://spark.apache.org

https://github.com/apache/spark

# Spark Architecture



# Spark Core Concepts



# Spark Core Concepts

- **Executor:** A process launched for an application on a worker node. Each application has its own executors.
- **Jobs:** A parallel computation consisting of one or multiple stages that gets spawned in response to a Spark action.
- **Stages:** Smaller set of tasks that each job is divided into.
- **Tasks:** A unit of work that will be sent to one executor.

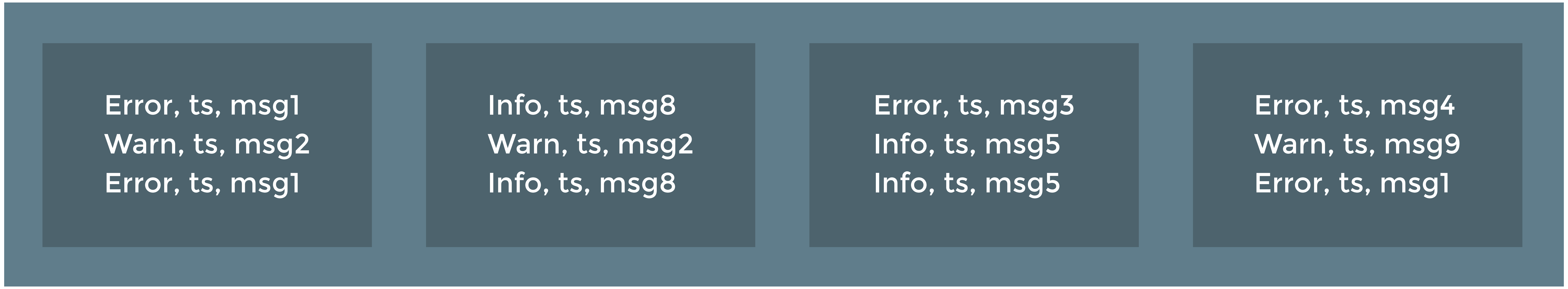
# Resilient Distributed Datasets

- Immutable.
- Partitioned collection.
- Operates in parallel.
- Customizable.

# RDDs - Partitions

- A **Partition** is one of the different chunks that a RDD is splitted on and that is sent to a node
- The more partitions we have, the more **parallelism** we get
- Each partition is candidate to be spread out to different **worker nodes**

RDD with 4 partitions



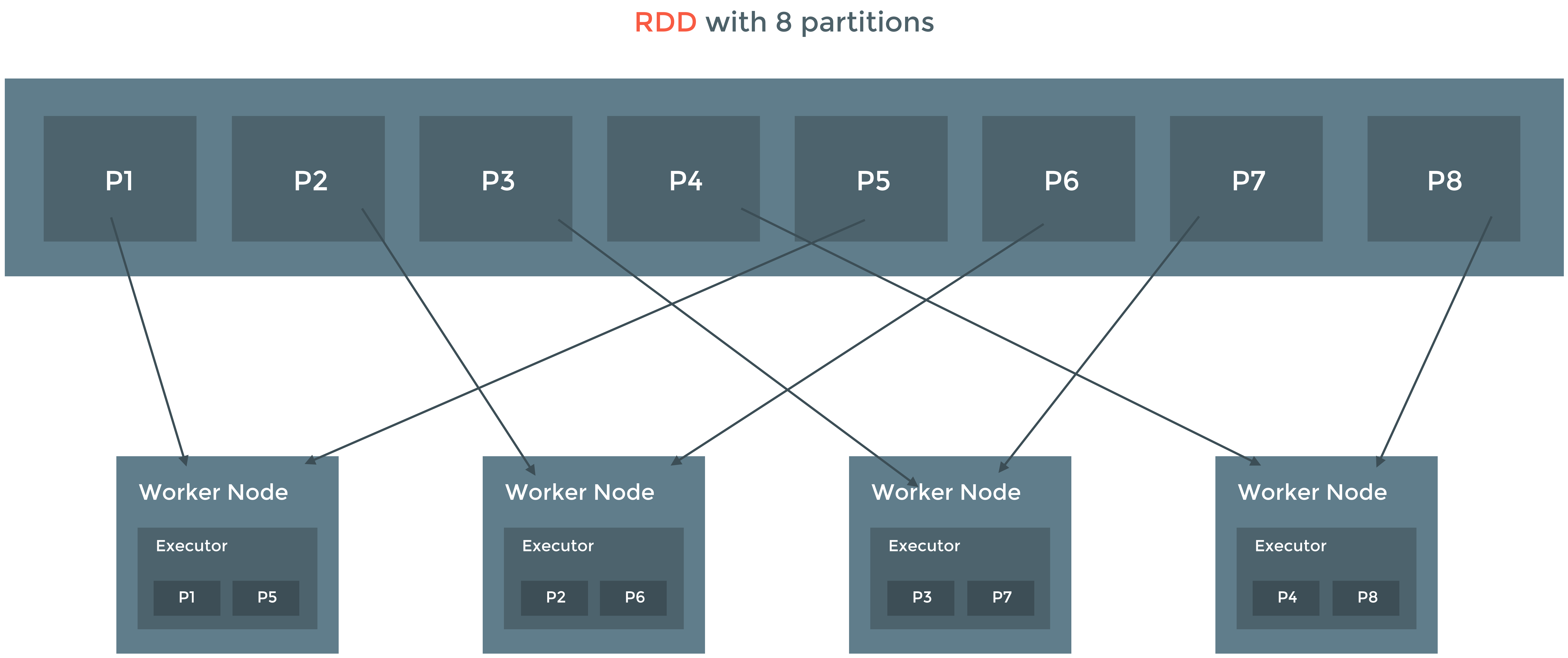
Error, ts, msg1
Warn, ts, msg2
Error, ts, msg1

Info, ts, msg8
Warn, ts, msg2
Info, ts, msg8

Error, ts, msg3
Info, ts, msg5
Info, ts, msg5

Error, ts, msg4
Warn, ts, msg9
Error, ts, msg1

# RDDs - Partitions



# RDDs - Operations

**Transformations**

- **Lazy** operations. They don't return a value, but a pointer to a new RDD.

**Actions**

- **Non-lazy** operations. They apply an operation to a RDD and return a value or write data to an external storage system.

# RDDs - Transformations

A set of some of the most popular Spark transformations:

- map
- flatMap
- filter
- groupByKey
- reduceByKey

# RDDs - Actions

A set of some of the most popular Spark actions:

- reduce
- collect
- foreach
- saveAsTextFile

# Transformations and Actions

With Visual Mnemonics, better.

Thanks to **Jeffrey Thompson**

- http://data-frack.blogspot.com.es/2015/01/visual-mnemonics-for-pyspark-api.html
- https://github.com/jkthompson/pyspark-pictures
- http://nbviewer.ipython.org/github/jkthompson/pyspark-pictures/blob/master/pyspark-pictures.ipynb

# Practice - Part 1 && Part 2

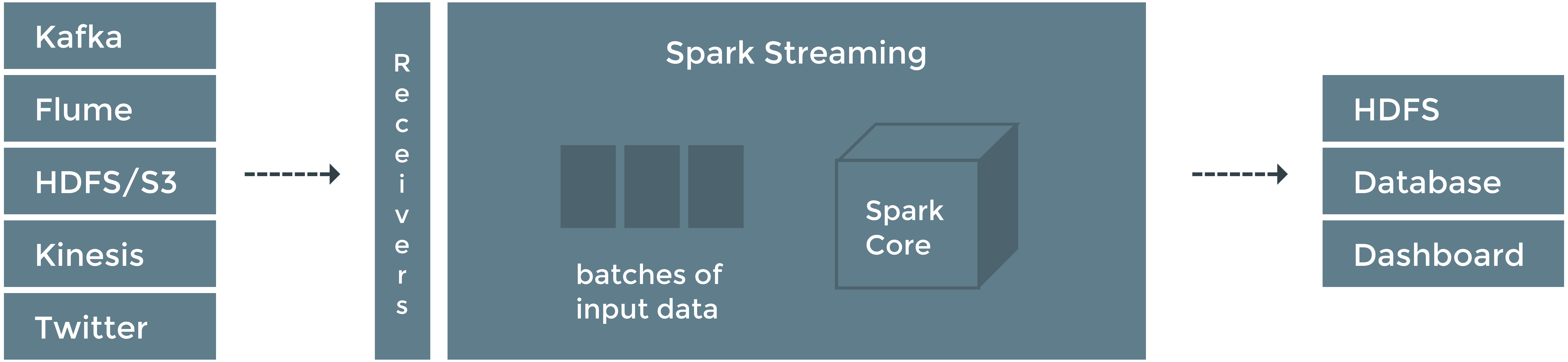
# Overview Spark SQL and DataFrames

- Works with structured and semistructured data
- DataFrame simplifies working with structured data
- Read/Write from structure data like JSON, Hive tables, Parquet, etc.
- SQL inside your Spark App
- Best Performance and more powerful operations API

# Practice - Part 3

# Overview Spark Streaming

- Streaming Applications
- DStreams or Discretized Streams
- Continuous Series of RDDs, grouped by batches



# Resources

- Official docs - http://spark.apache.org/docs/latest
- Learning Spark - http://shop.oreilly.com/product/0636920028512.do
- Databricks Spark Knowledge Base - https://goo.gl/wMy7Se
- Community packages for Spark - http://spark-packages.org/
- Apache Spark Youtube channel - https://www.youtube.com/user/TheApacheSpark
- API through pictures - https://goo.gl/JMDeqJ
- 47 Degrees Blog - http://www.47deg.com/blog/tags/spark

# Thanks!

Q&A