PySpark Tutorial

edureka!

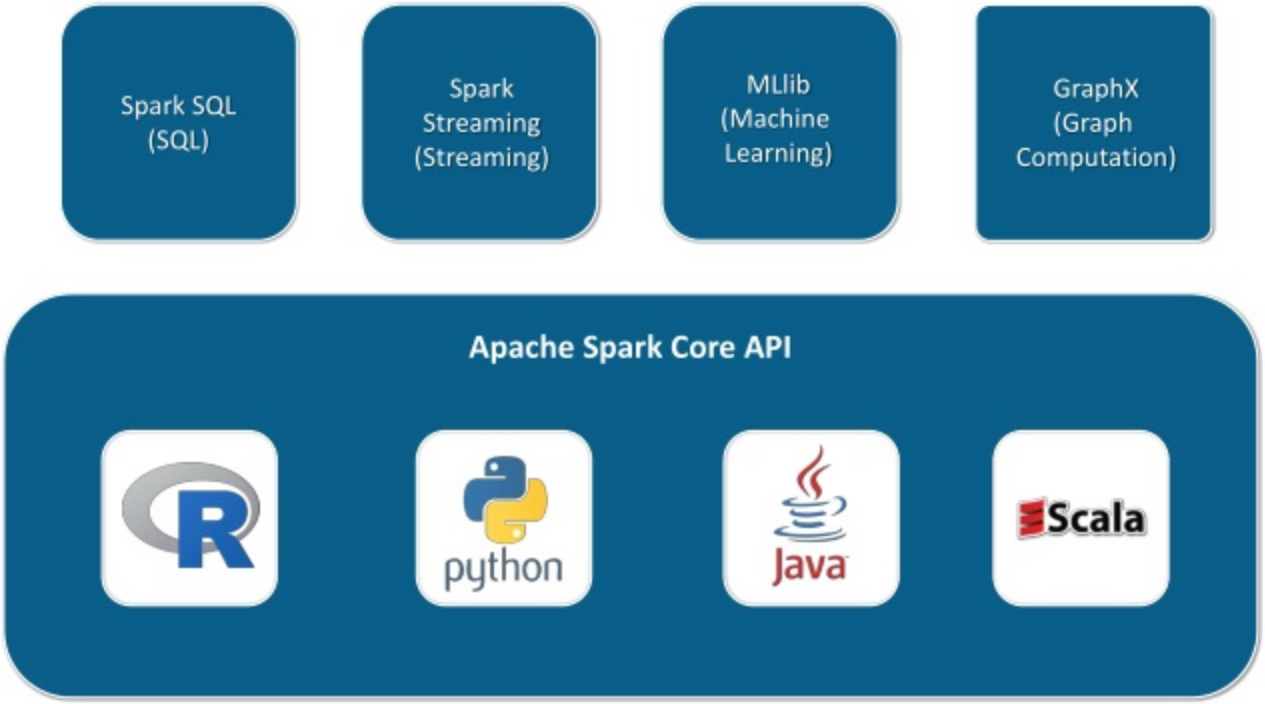
Objectives of Today's Training

1. PySpark
2. Advantages of PySpark
3. PySpark Installation
4. PySpark Fundamentals
5. Demo

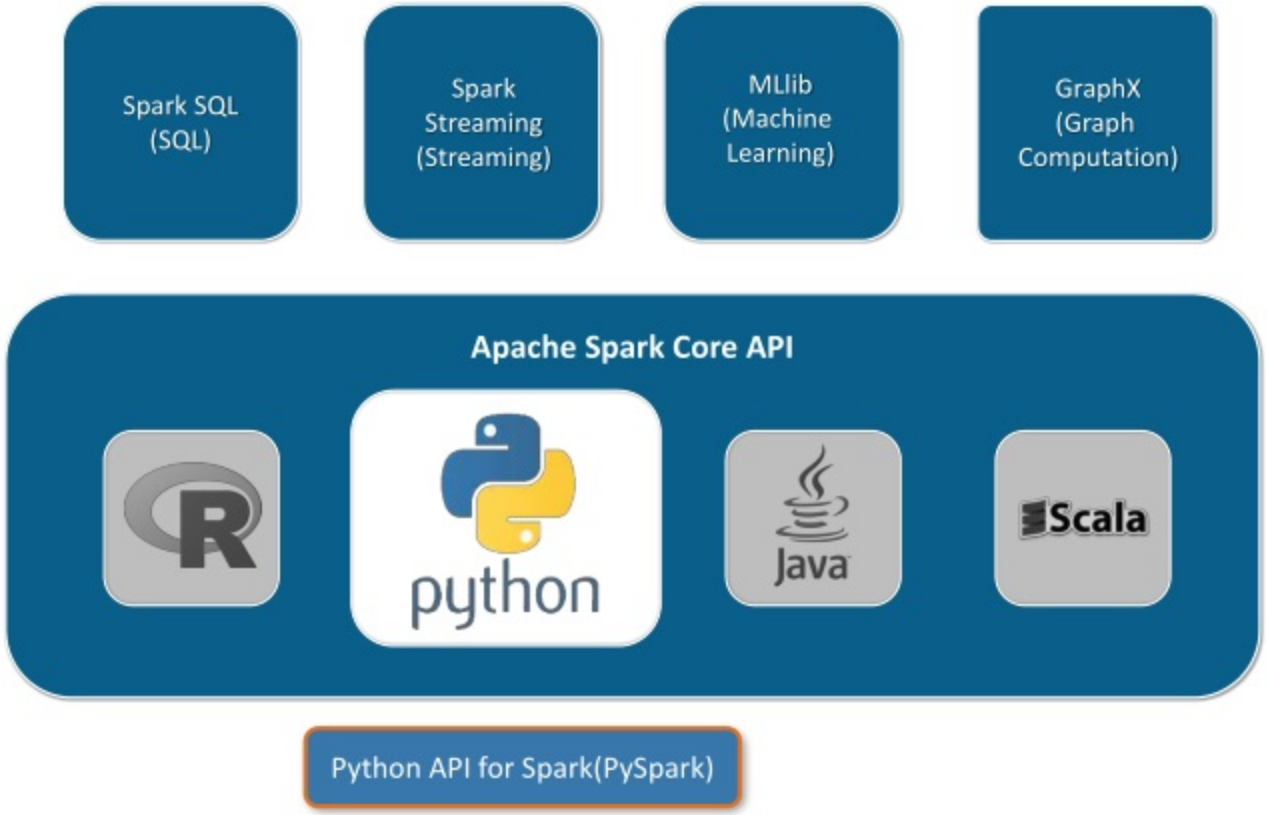
PySpark

edureka!

Spark Ecosystem



Python in Spark Ecosystem



PySpark

Spark is an open-source cluster-computing framework which is built around speed, ease of use, and streaming analytics

Python is general purpose high level programming language. It provides wide range of libraries and is majorly used for Machine Learning and Data Science

- It is a Python API for Spark majorly used for Data Science and Analysis
- Using PySpark, you can work with Spark **RDDs** in Python

Advantages Spark with Python

edureka!

Advantages

Advantages

EASY TO LEARN

SIMPLE & COMPREHENSIVE API

Advantages

EASY TO LEARN

SIMPLE & COMPREHENSIVE API

BETTER CODE READABILITY & MAINTENANCE

Advantages

EASY TO LEARN

SIMPLE & COMPREHENSIVE API

BETTER CODE READABILITY & MAINTENANCE

AVAILABITLITY OF VISUALIZATION

Advantages


EASY TO LEARN
SIMPLE & COMPREHENSIVE API
API
WIDE RANGE OF LIBRARIES
BETTER CODE READABILITY & MAINTENANCE
AVAILABITLITY OF VISUALIZATION

Advantages


EASY TO LEARN
ACTIVE COMMUNITY
SIMPLE & COMPREHENSIVE API
API
WIDE RANGE OF LIBRARIES
BETTER CODE READABILITY & MAINTENANCE
AVAILABITLITY OF VISUALIZATION

PySpark Installation

edureka!

PySpark Installation

1. Go to: https://spark.apache.org/downloads.html
2. Select the Spark version from the drop down list
3. Click on the link to download the file.

PySpark Installation

Install pip (version 10 or more)

Install jupyter notebook

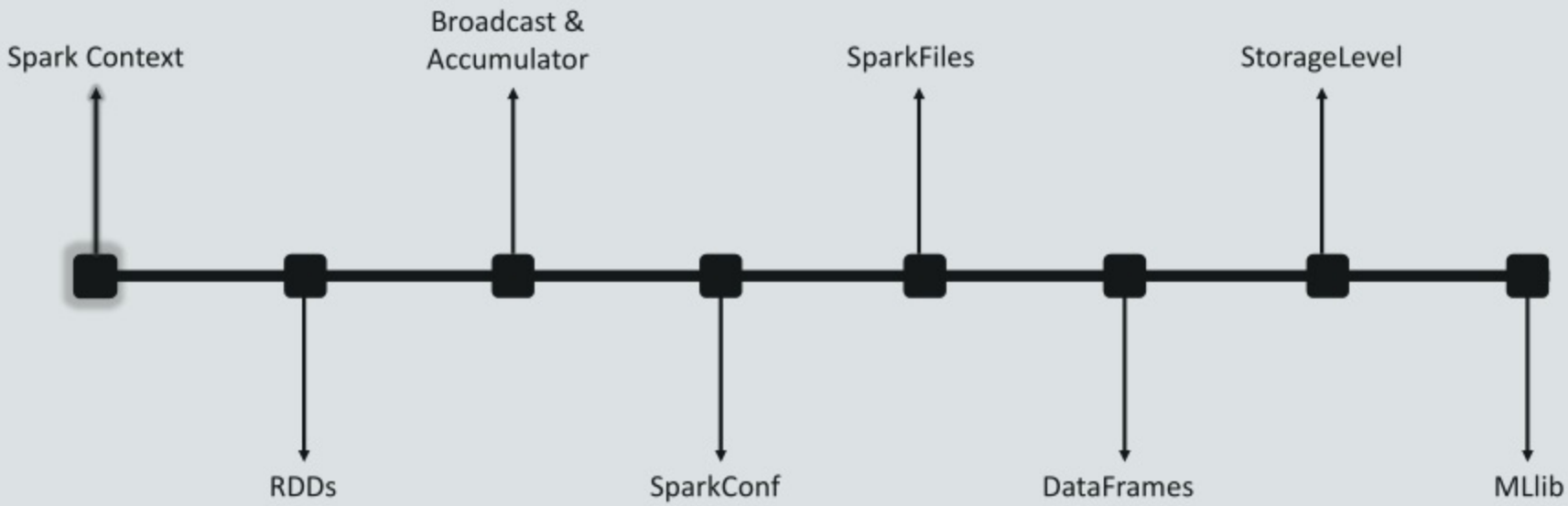
PySpark Installation

Add the Spark and PySpark in the bashrc file

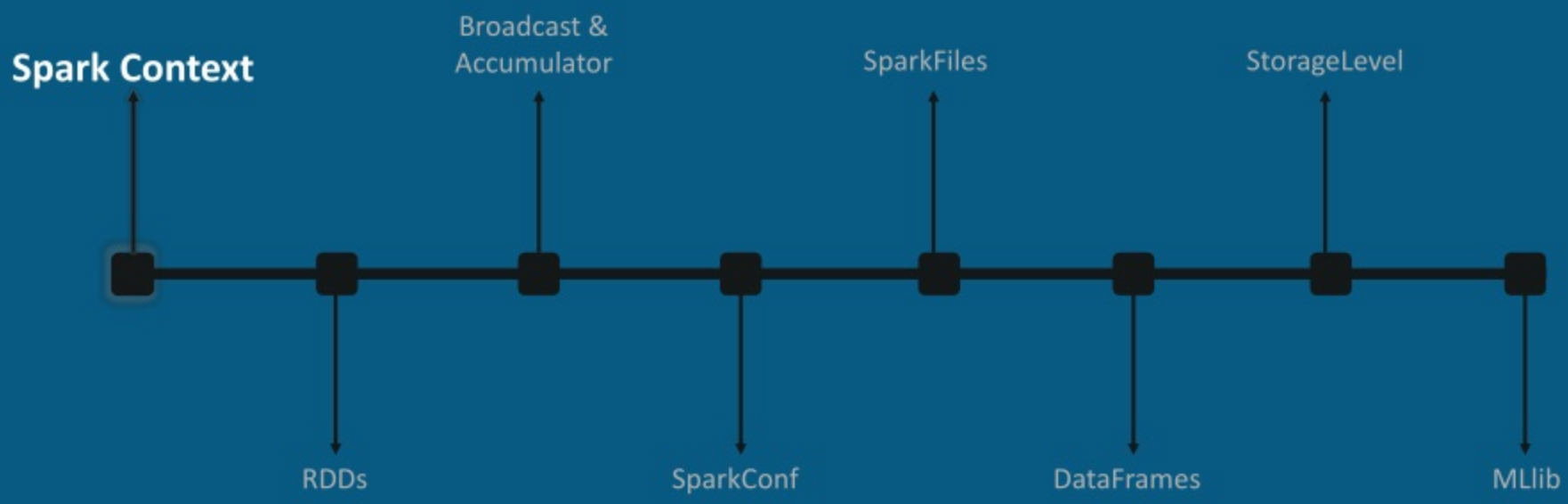
.bashrc
/home/edureka

```
# User specific aliases and functions

export HADOOP_HOME=/usr/lib/hadoop-2.7.3
export HADOOP_CONF_DIR=/usr/lib/hadoop-2.7.3/etc/hadoop
export HADOOP_MAPRED_HOME=/usr/lib/hadoop-2.7.3
export HADOOP_COMMON_HOME=/usr/lib/hadoop-2.7.3
export HADOOP_HDFS_HOME=/usr/lib/hadoop-2.7.3
export YARN_HOME=/usr/lib/hadoop-2.7.3
export PATH=$PATH:/usr/lib/hadoop-2.7.3/bin

export SPARK_HOME=/usr/lib/spark-2.2.1-bin-hadoop2.7
export PATH=$PATH:/usr/lib/spark-2.2.1-bin-hadoop2.7/bin

export PYSPARK_DRIVER_PYTHON=jupyter
export PYSPARK_DRIVER_PYTHON_OPTS='notebook'

export JAVA_HOME=/usr/lib/jvm/jdk1.8.0_151
export PATH=$JAVA_HOME/bin:$PATH

export KAFKA_HOME=/usr/lib/kafka_2.11-1.0.0
```

PySpark Fundamentals

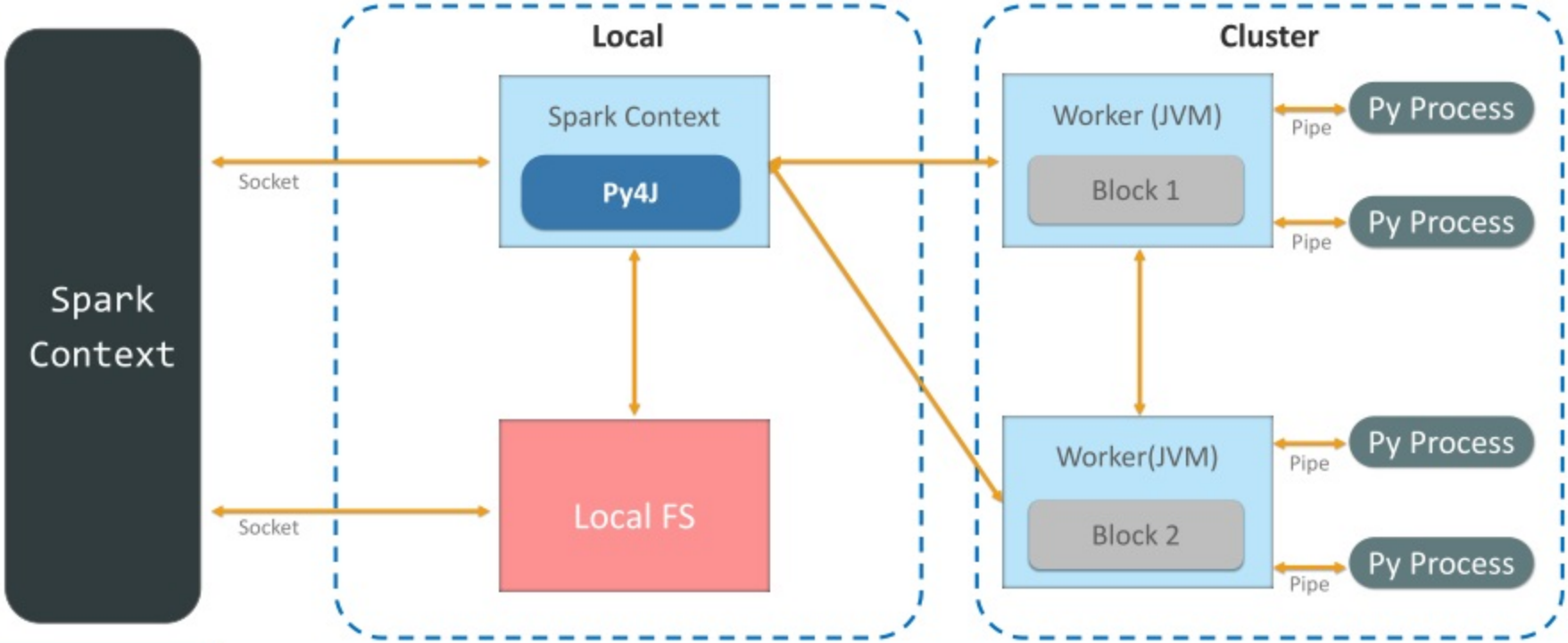


edureka!


Spark Context
Broadcast &
Accumulator
SparkFiles
StorageLevel
RDDs
SparkConf
DataFrames
MLlib

Spark Context

SparkContext is the entry point to any spark functionality



Spark Context

SparkContext parameters

Master | appName | sparkHome | pyFiles

Environment | batchSize | Serializer | conf

Gateaway | JSC | Profiler_cls

Spark Context

SparkContext parameters

Master

appName

sparkHome

pyFiles

Environment

batchSize

Serializer

conf

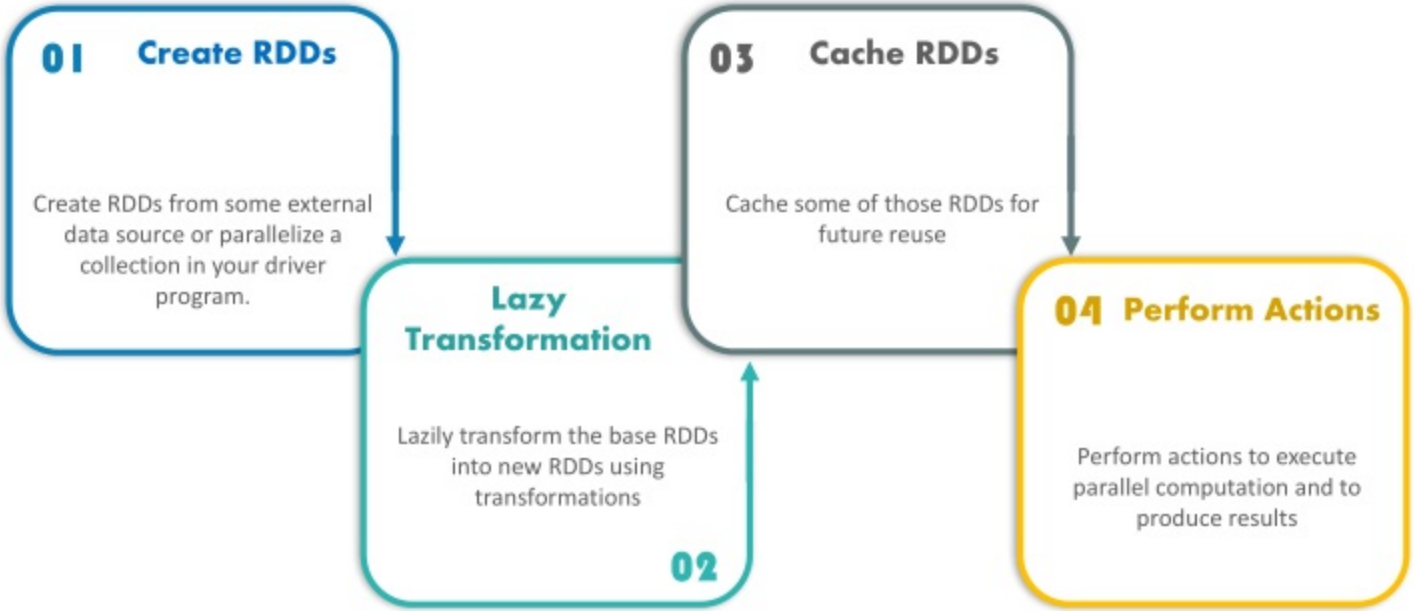
Gateaway

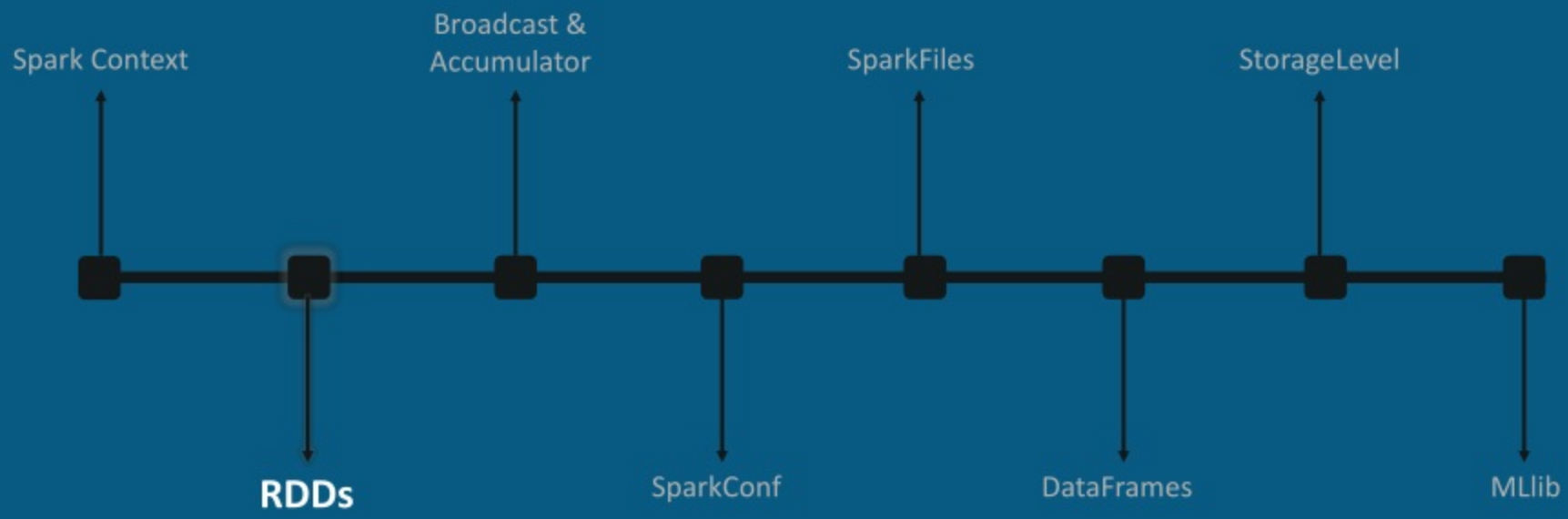
JSC

Profiler_cls

PySpark

Basic life cycle of a PySpark program




Spark Context
Broadcast &
Accumulator
SparkFiles
StorageLevel
RDDs
SparkConf
DataFrames
MLlib

Resilient Distributed Dataset (RDDs)

RDDs is the building block of every Spark application and is immutable

Resilient	Fault tolerant and is capable of rebuilding data on failure
Distributed	Data is distributed among the multiple nodes in a cluster
Dataset	Collection of partitioned data with primitive values or values of value

Transformations & Actions in RDDs

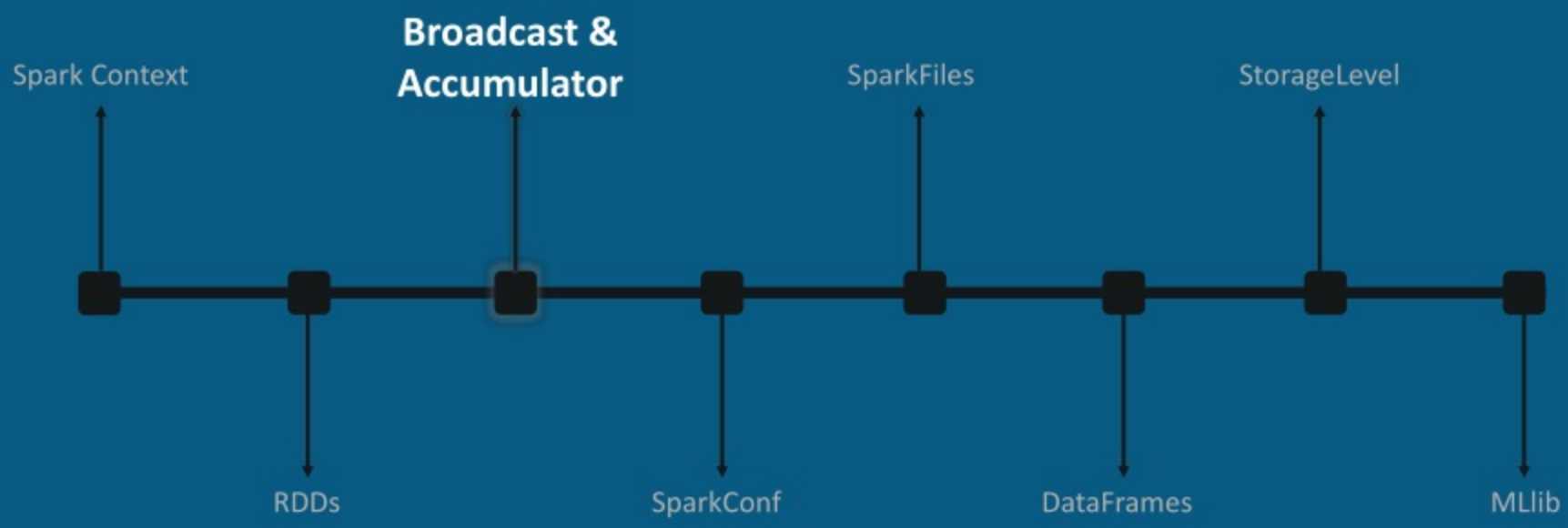
To work on this immutable data, you need to create a new one via Transformations and Actions

Transformations

- map
- flatMap
- filter
- distinct
- reduceByKey
- mapPartitions
- sortBy

Actions

- collect
- collectAsMap
- reduce
- countByKey/countByValue
- take
- first


Spark Context
Broadcast &
Accumulator
SparkFiles
StorageLevel
RDDs
SparkConf
DataFrames
MLlib

Broadcast & Accumulator

Parallel processing is achieved in Spark by using shared variables

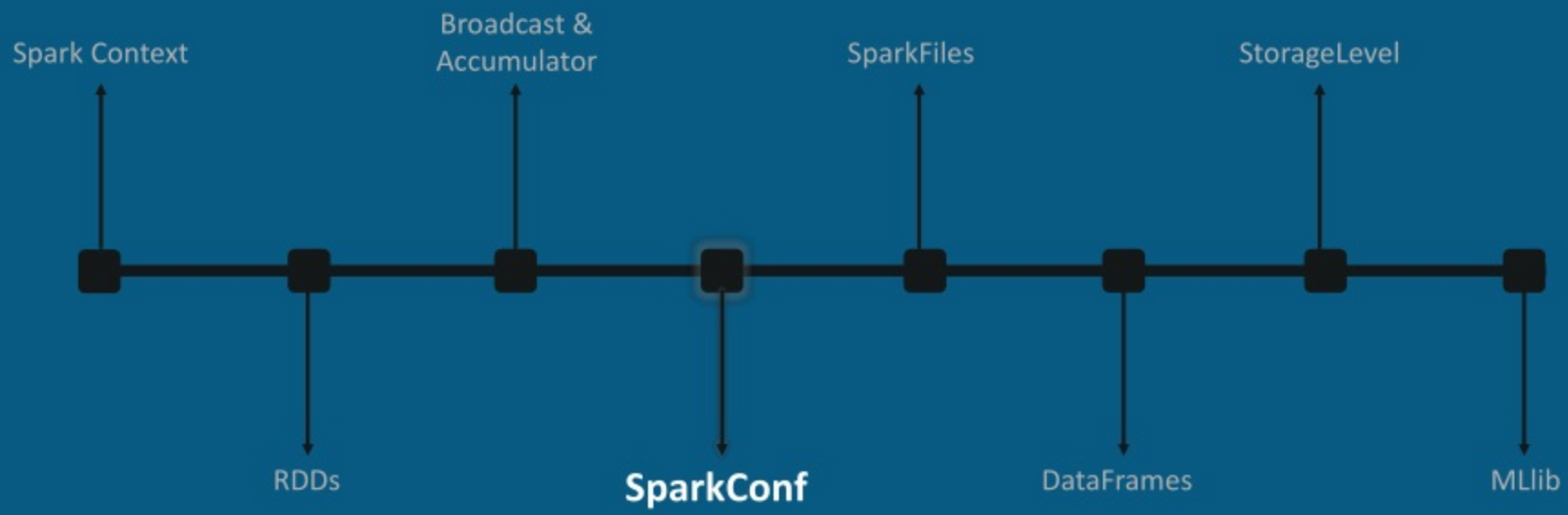
Shared Variables

Broadcast

These variables are used to save the copy of data across all nodes

Accumulator

These variables are used to aggregate the information through associative and commutative operations


Spark Context
Broadcast &
Accumulator
SparkFiles
StorageLevel
RDDs
SparkConf
DataFrames
MLlib

SparkConf

SparkConf provides the configurations to run a Spark application on a local system or a cluster

```
class SparkConf (
    loadDefaults = True,
    _jvm = None,
    _jconf = None
)
```

SparkConf object is used to set different parameters which takes priority over the system properties

Once SparkConf object is passed to Spark, it becomes immutable

SparkConf

Attributes of SparkConf class

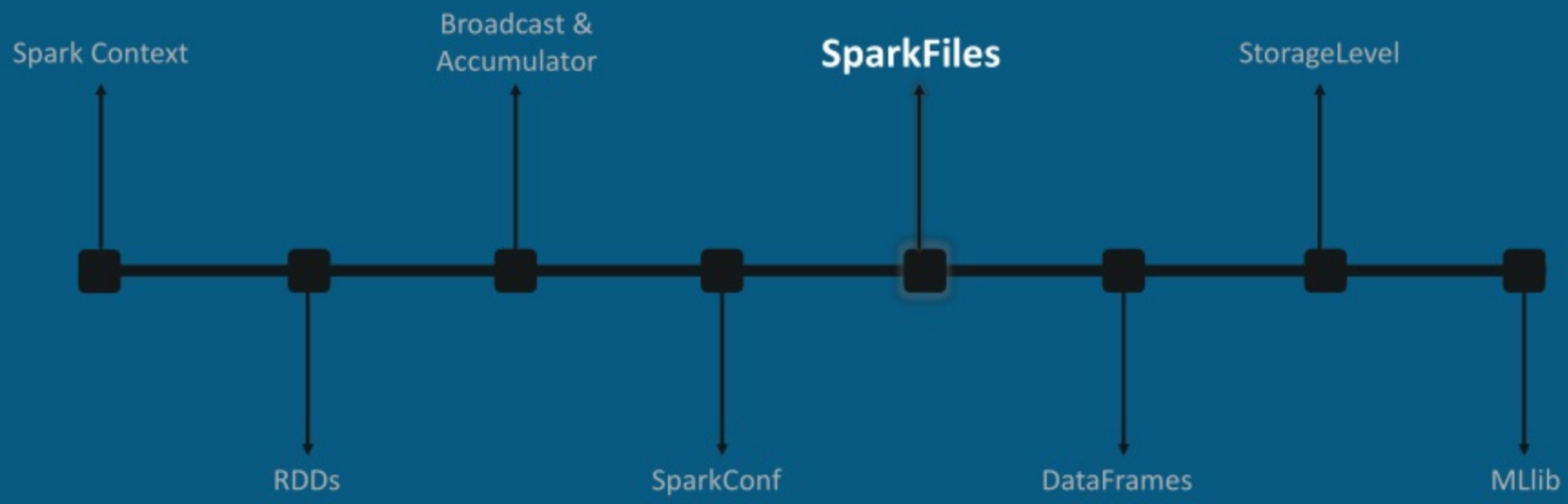
`set(key, value)` ..... Sets Config property

`setMaster(value)` ..... Sets the master URL

`setAppName(value)` ..... Sets an application's name

`get(key, defaultValue=None)` ..... Gets the configuration value of a key

`setSparkHome(value)` ..... Sets the Spark installation path on worker nodes

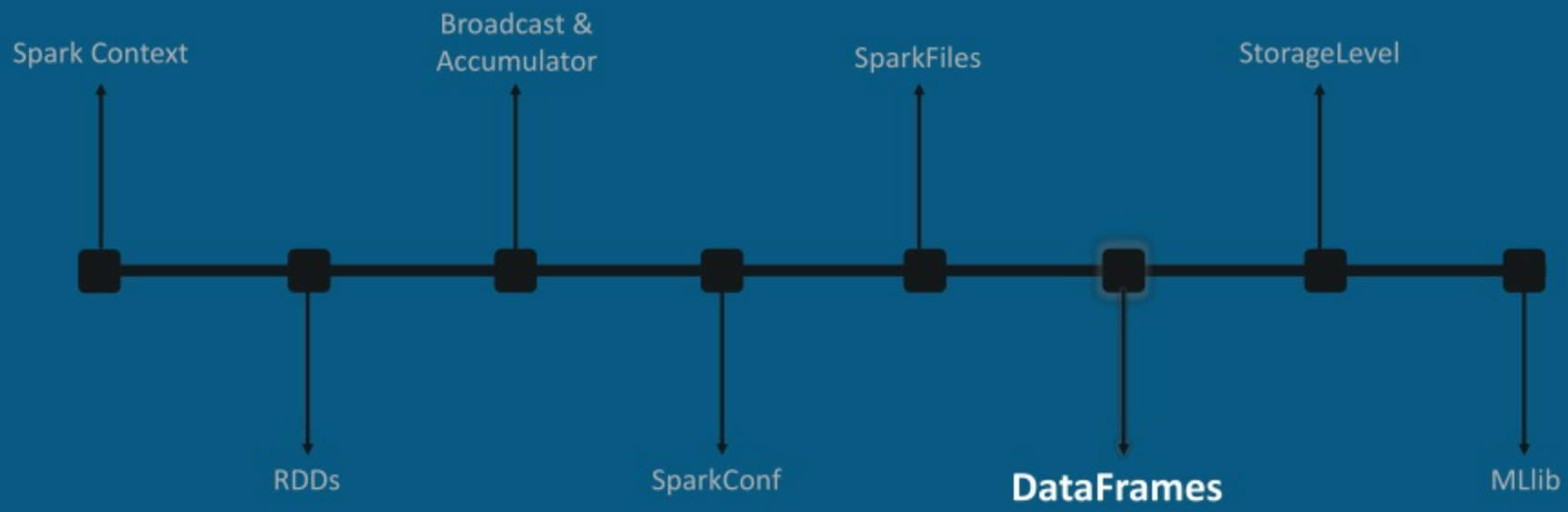

Spark Context
Broadcast &
Accumulator
SparkFiles
StorageLevel
RDDs
SparkConf
DataFrames
MLlib

SparkFiles

SparkFiles class helps in resolving the paths of files added to the Spark

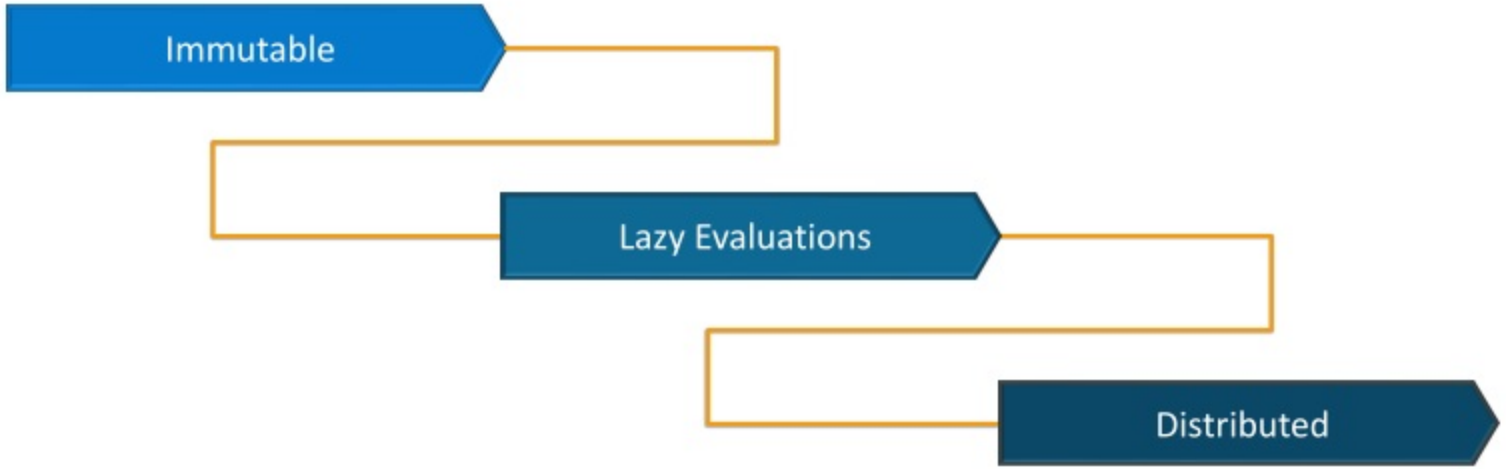
`get(filename)` ..... It specifies the path of the file that is added through sc.addFile()

`getrootdirectory()` ..... It specifies the path to the root directory of the file that is added through sc.addFile()

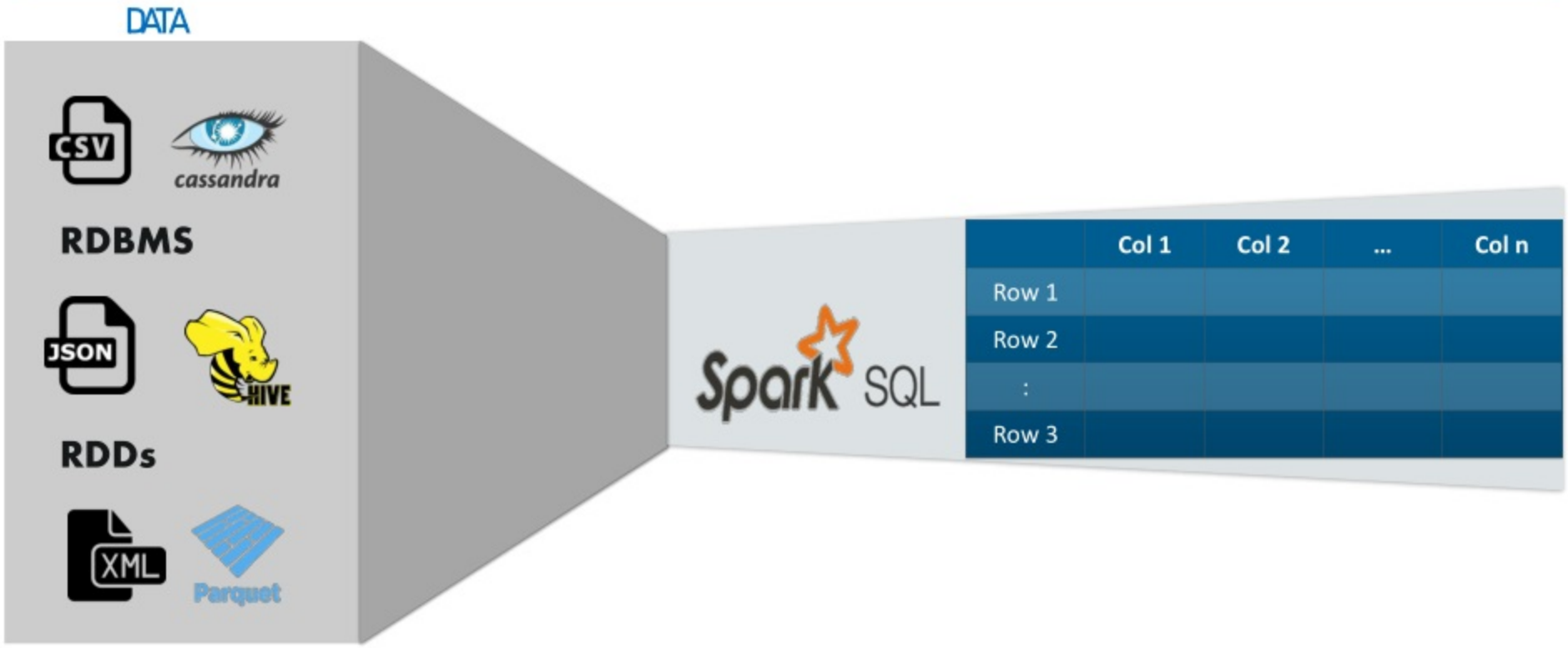

Spark Context
Broadcast &
Accumulator
SparkFiles
StorageLevel
RDDs
SparkConf
DataFrames
MLlib

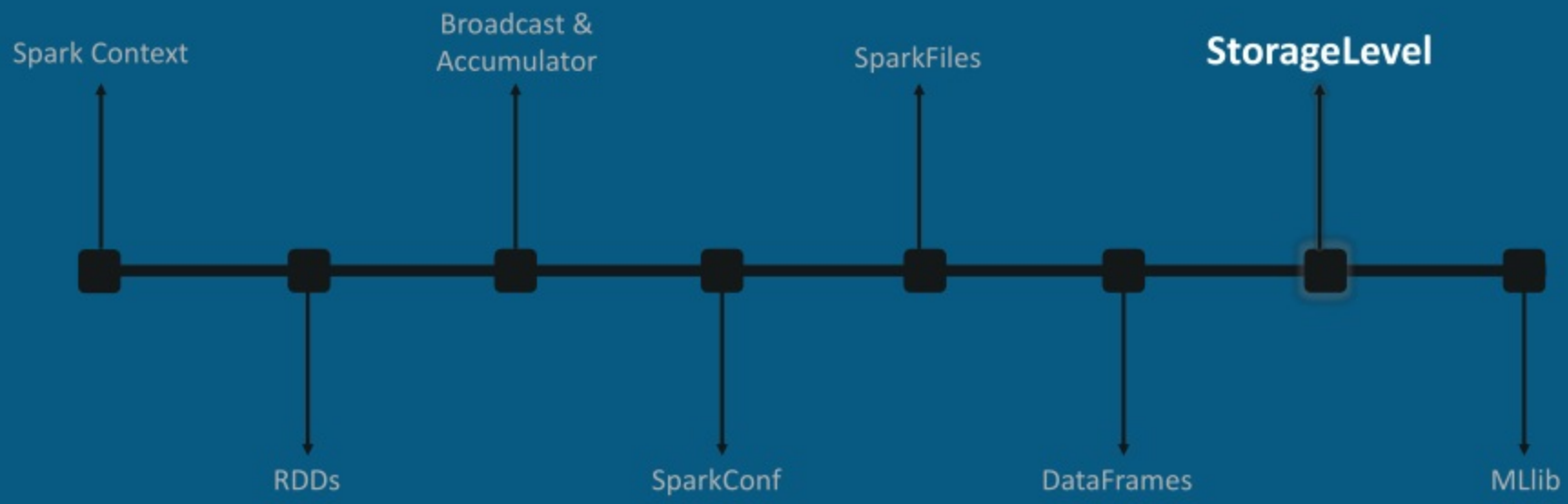
DataFrames

Dataframe is a distributed collection of rows under named columns

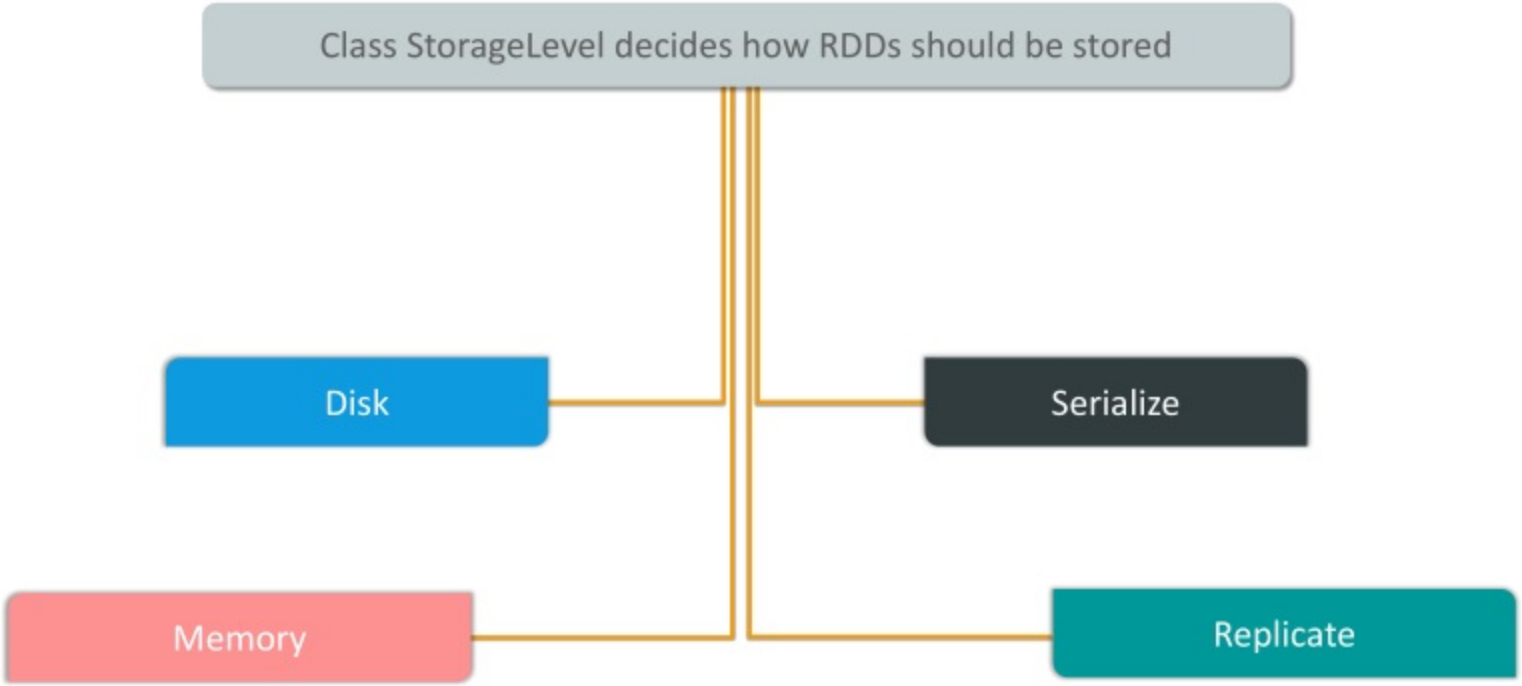


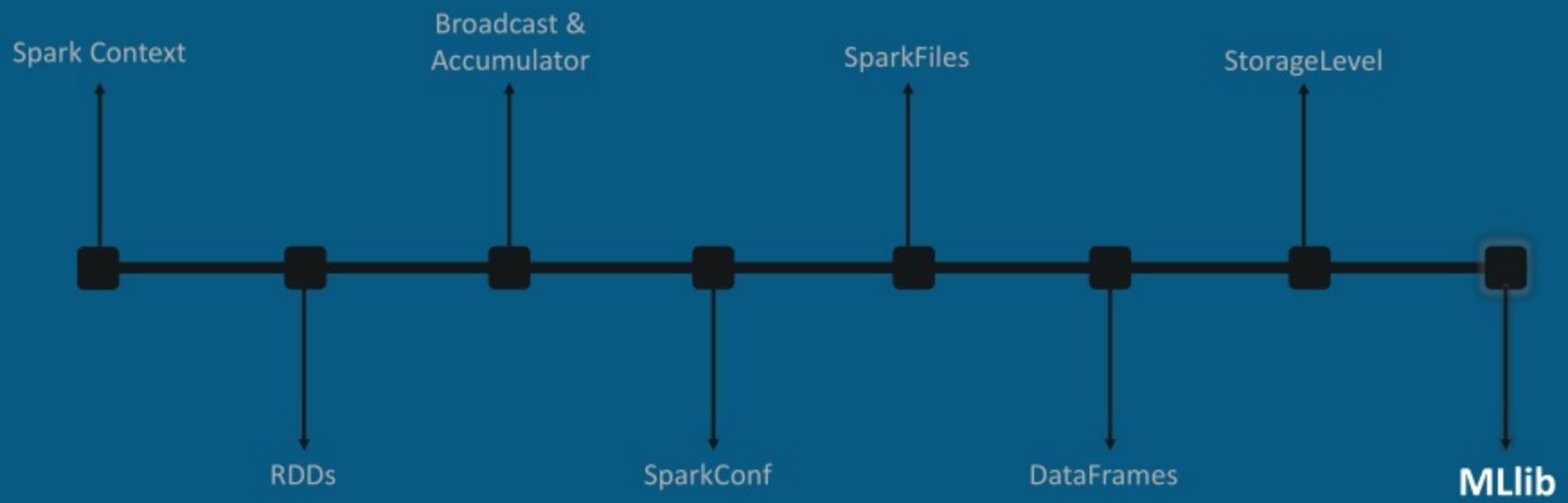
Dataframes


DATA
CSV
cassandra
RDBMS
JSON
HIVE
RDDs
XML
Parquet
Spark SQL
Col 1
Col 2
...
Col n
Row 1
Row 2
:
Row 3


Spark Context
Broadcast &
Accumulator
SparkFiles
StorageLevel
RDDs
SparkConf
DataFrames
MLlib

StorageLevels




Spark Context
Broadcast &
Accumulator
SparkFiles
StorageLevel
RDDs
SparkConf
DataFrames
MLlib

MLlib

Machine Learning API in Spark which interoperates with NumPy in Python is called **MLlib**

It provides an integrated Data Analysis workflow

Enhances speed and performance

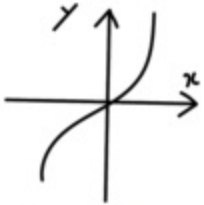
MLlib

Various algorithms supported by MLlib

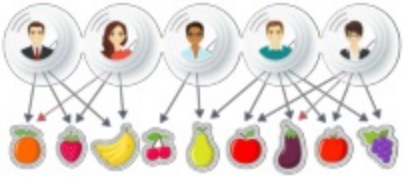
MLlib

Clustering

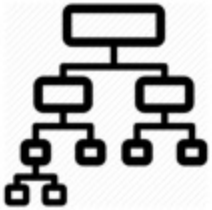
Frequent Pattern Matching



Linear Algebra



Collaborative Filtering

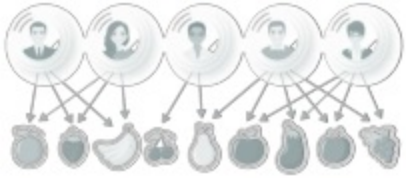


Classification

Linear Regression

MLlib

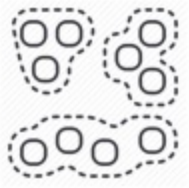
Various algorithms supported by MLlib



MLlib

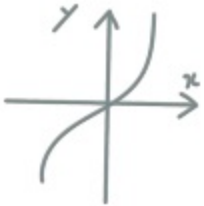
Various algorithms supported by MLlib

MLlib

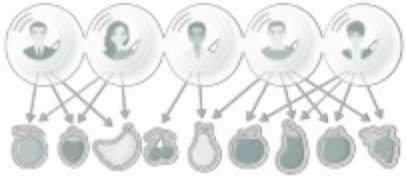


Clustering

Frequent Pattern Matching



Linear Algebra



Collaborative Filtering

Classification

Linear Regression

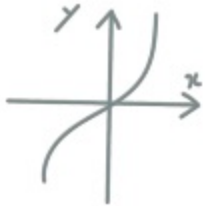
MLlib

Various algorithms supported by MLlib

MLlib

Clustering

Frequent Pattern Matching



Linear Algebra

Collaborative Filtering

Classification

Linear Regression

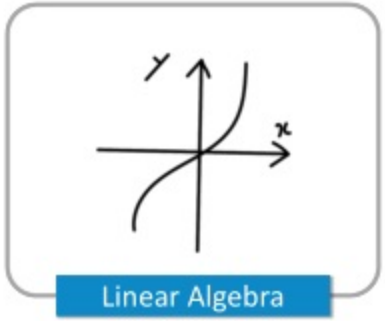
MLlib

Various algorithms supported by MLlib

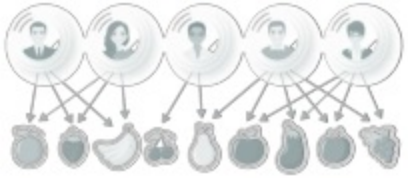
MLlib

Clustering

Frequent Pattern Matching



Linear Algebra



Collaborative Filtering

Classification

Linear Regression

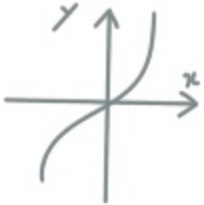
MLlib

Various algorithms supported by MLlib

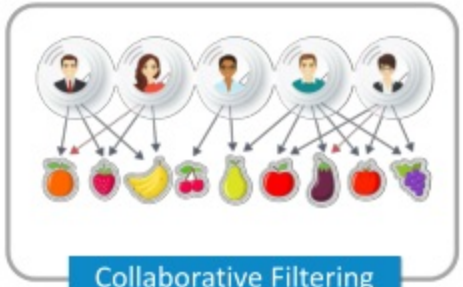
MLlib

Clustering

Frequent Pattern Matching



Linear Algebra



Collaborative Filtering

Classification

Linear Regression

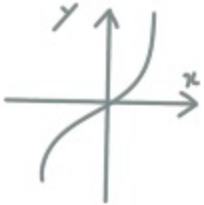
MLlib

Various algorithms supported by MLlib

MLlib

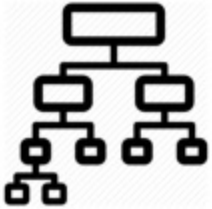
Clustering

Frequent Pattern Matching



Linear Algebra

Collaborative Filtering



Classification

Linear Regression

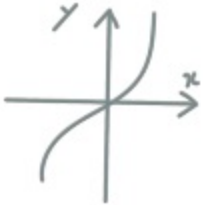
MLlib

Various algorithms supported by MLlib

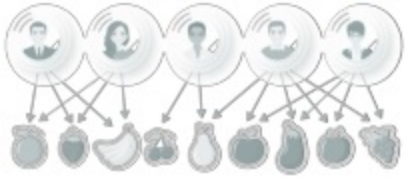
MLlib

Clustering

Frequent Pattern Matching

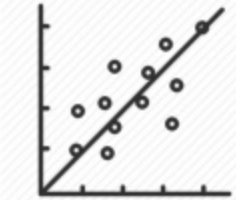


Linear Algebra



Collaborative Filtering

Classification



Linear Regression

Thank You

For more information please visit our website
www.edureka.co