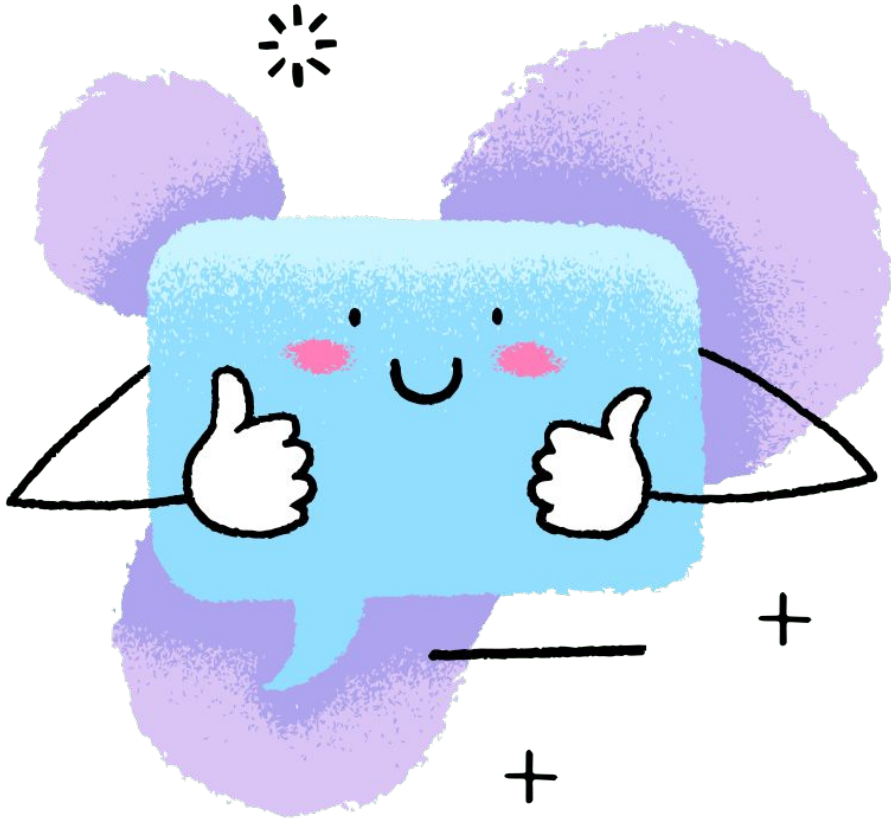
PySpark



В предыдущей серии...

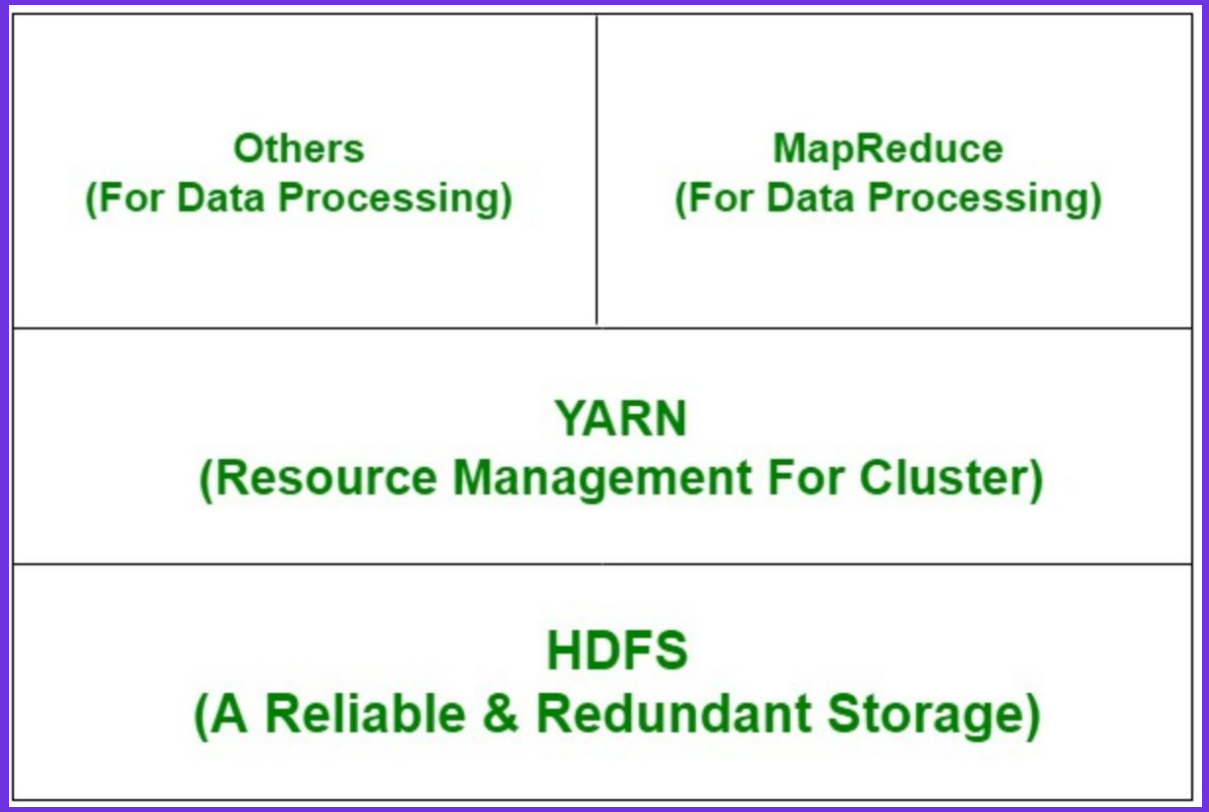
1. Горизонтальное vs вертикальное масштабирование
2. Проблемы MapReduce
3. Форматы хранения данных (текстовые и бинарные)

Почему нужен Spark?

Что будет на уроке

1. Spark, архитектура
2. YARN зачем и как работает
3. SparkSession. Параметры приложения: количество экзекьюторов, ядер, памяти.
4. Типы данных в Spark (в сравнении с питоном)

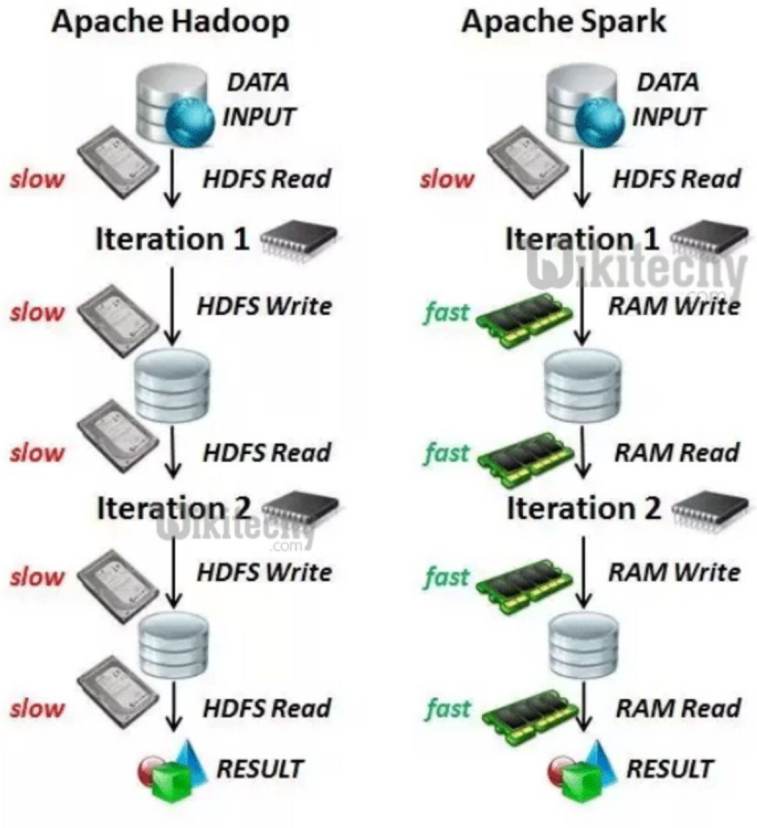
Hadoop Modules:



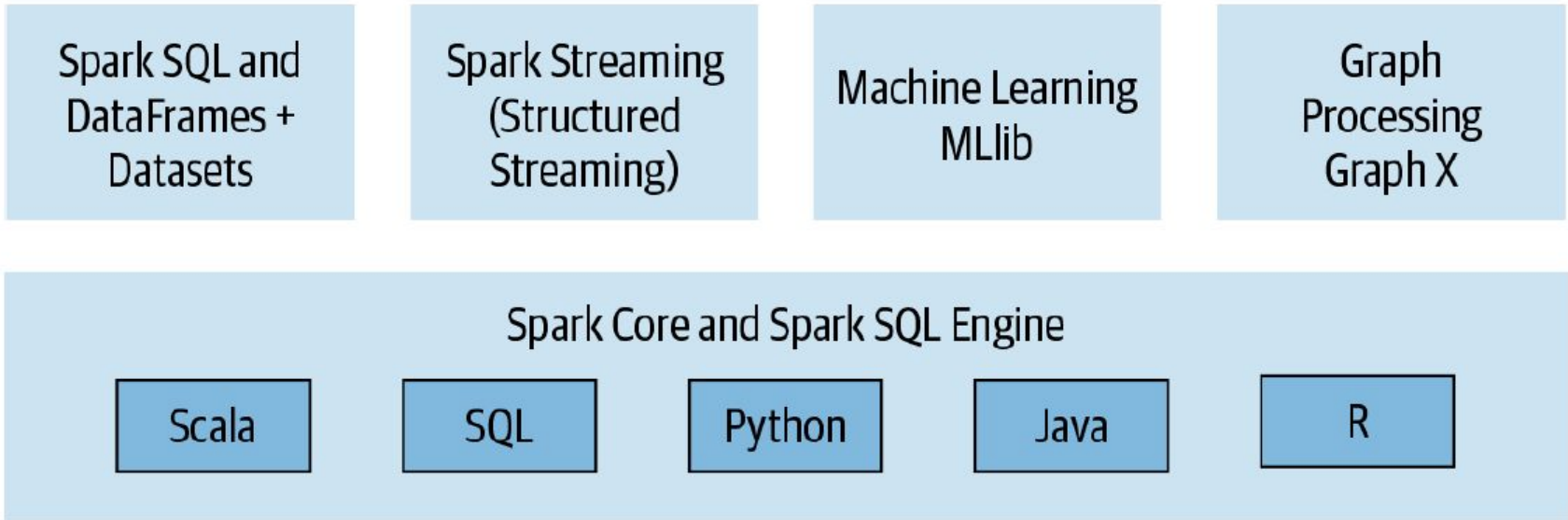
Spark (2009) vs Hadoop MapReduce:

1. in-memory storage for intermediate results between iterative and interactive map and reduce computations

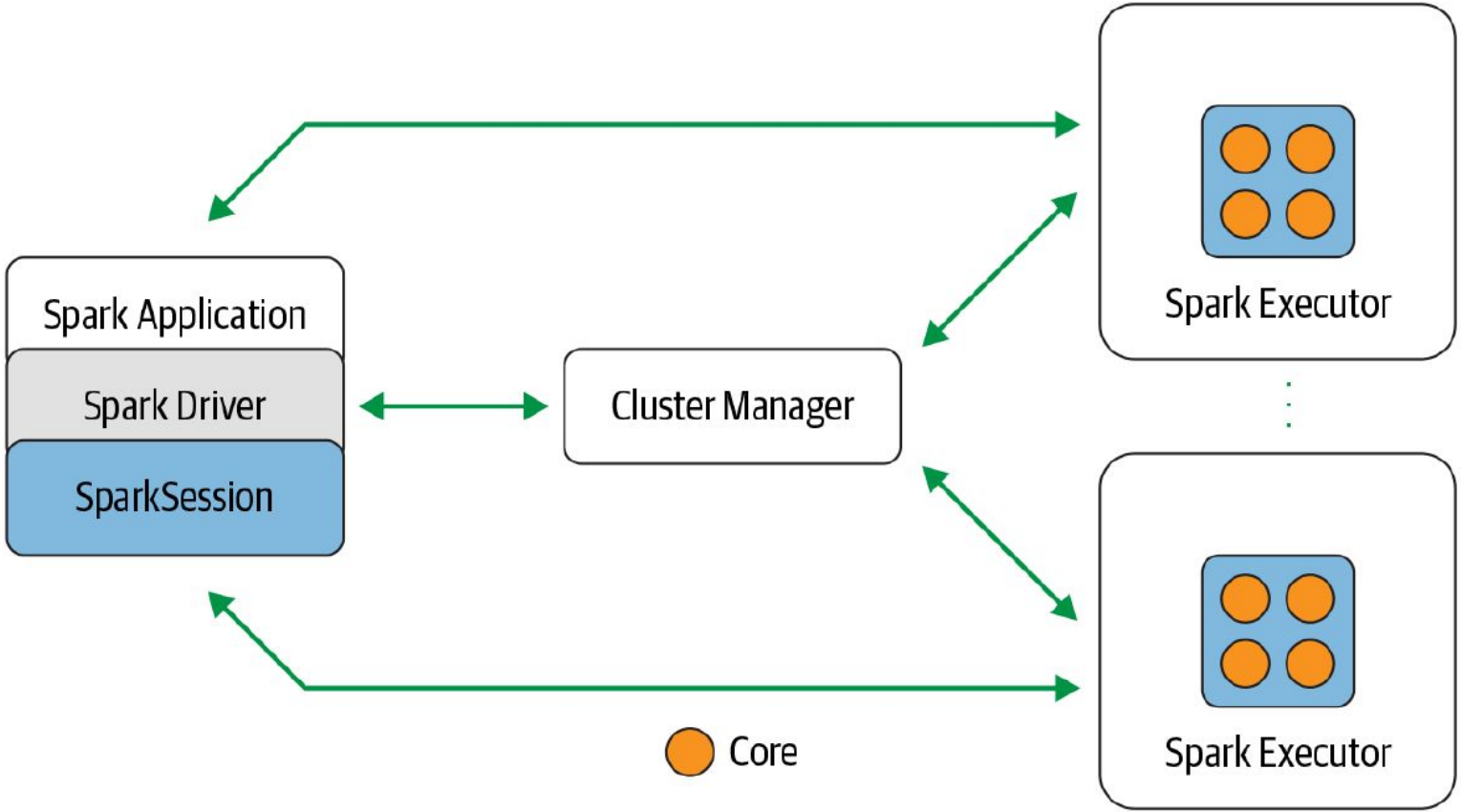
2. offer easy and composable APIs in multiple languages as a programming model



Spark -- Unified Engine for Big Data Processing -- объединил в себе возможности пакетной обработки, работы с графами, потоками и SQL



Архитектура приложения Spark



Cluster manager

Распределяет ресурсы между spark приложениями

standalone cluster manager

FIFO исполнение приложений

Apache Hadoop YARN

стандартное решение

распределяет, освобождает ресурсы между различными приложениями

Часть Hadoop

Apache Mesos

YARN-like, but большая изолированность процессов

поддерживает не-hadoop приложения

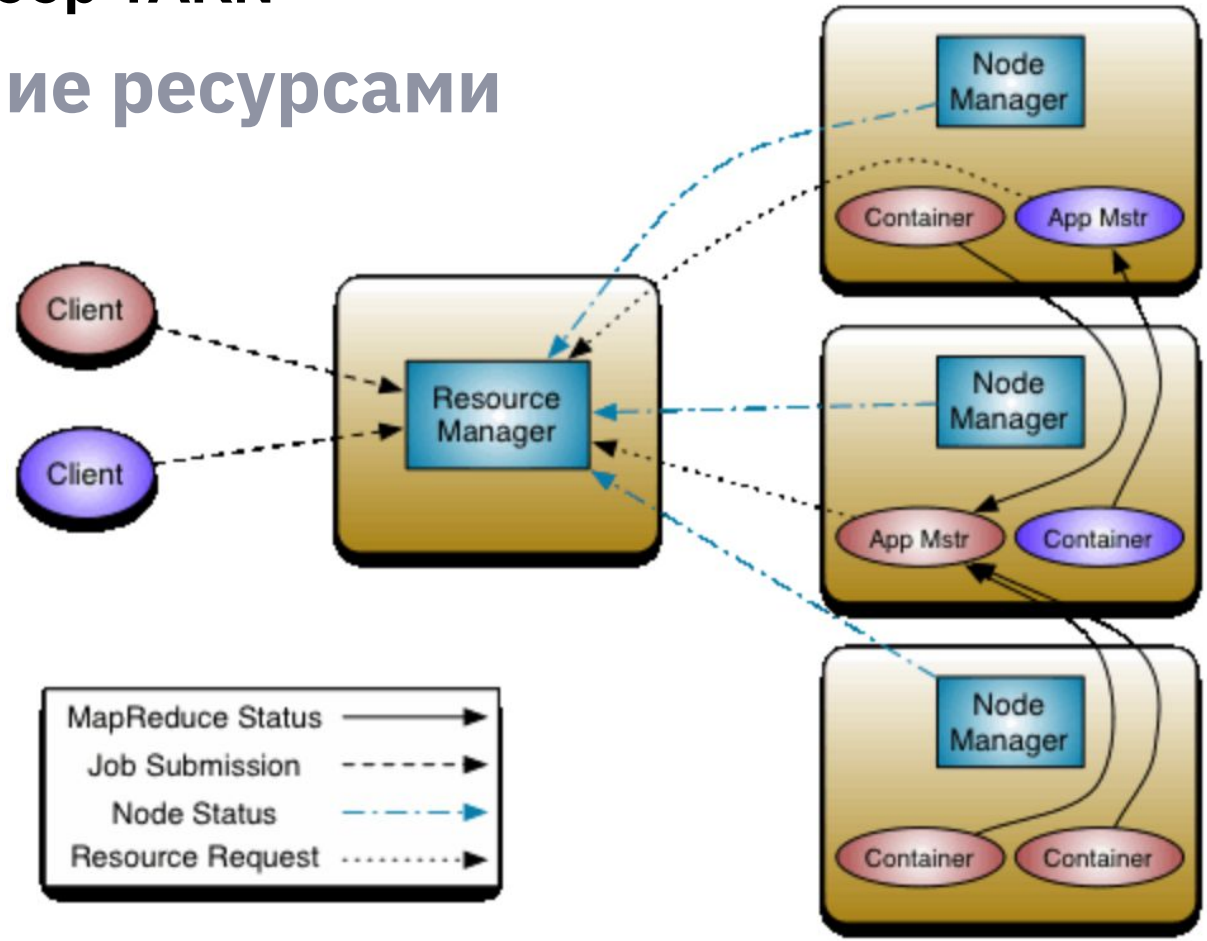
Kubernetes

запуск в контейнерах = абсолютная изолированность

существенно сложнее в поддержке

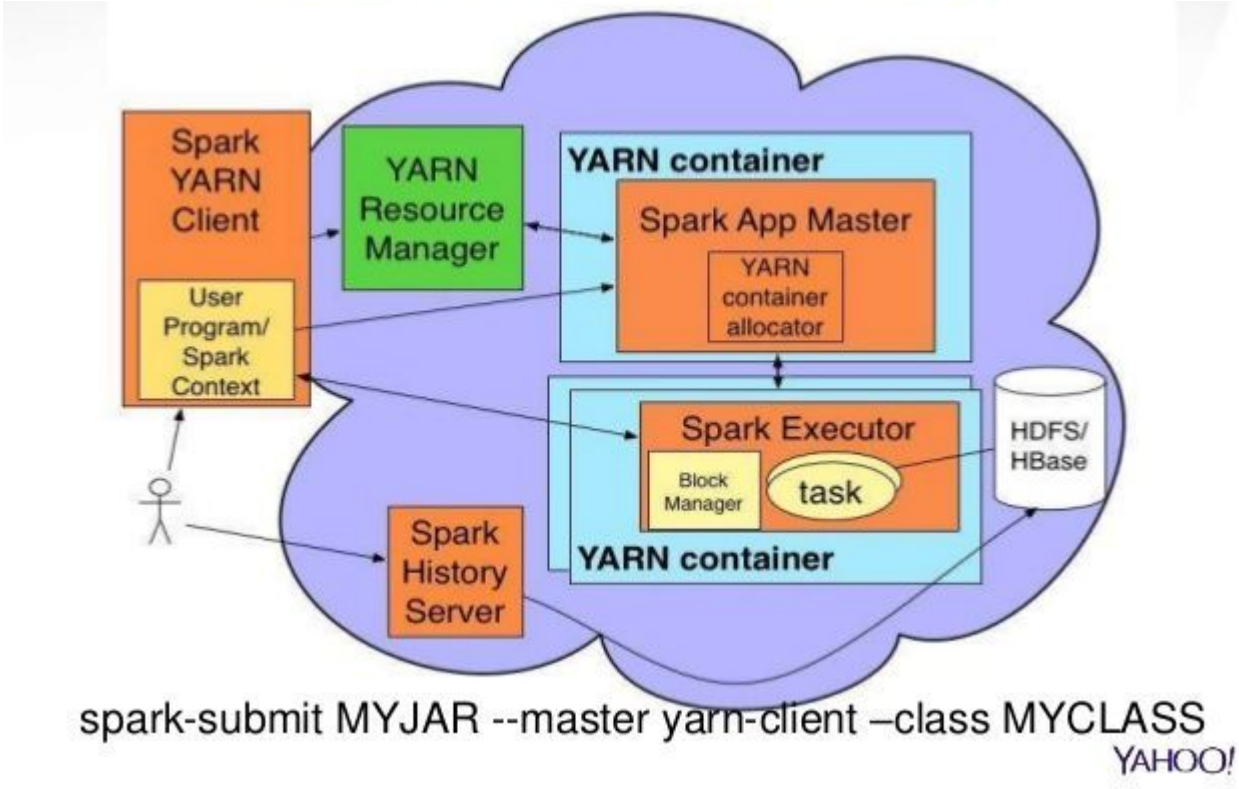
Apache Hadoop YARN

Управление ресурсами



Apache Hadoop YARN

Управление ресурсами



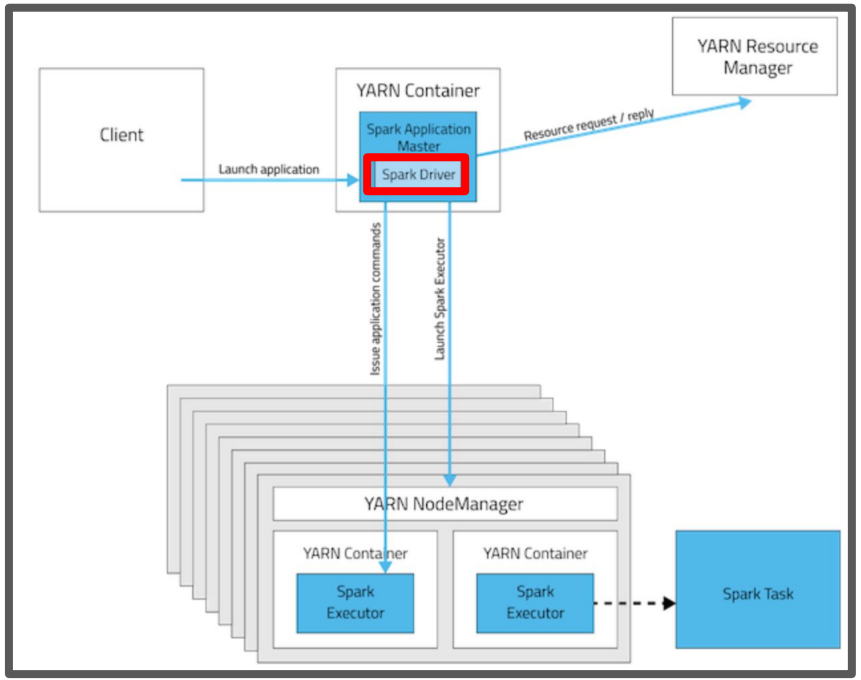
Spark driver

Может находиться вне кластера

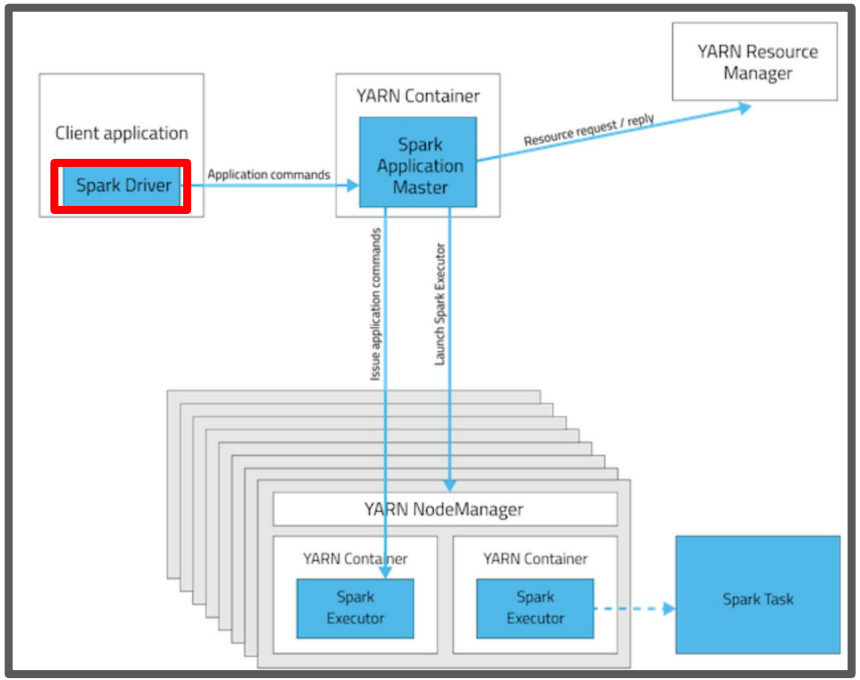
- it communicates with the cluster manager
- it requests resources (CPU, memory, etc.) from the cluster manager for Spark’s executors (JVMs)
- it transforms all the Spark operations into DAG computations
- it distributes their execution as tasks across the Spark executors

Spark on YARN deployment modes

cluster mode



client mode



SparkSession

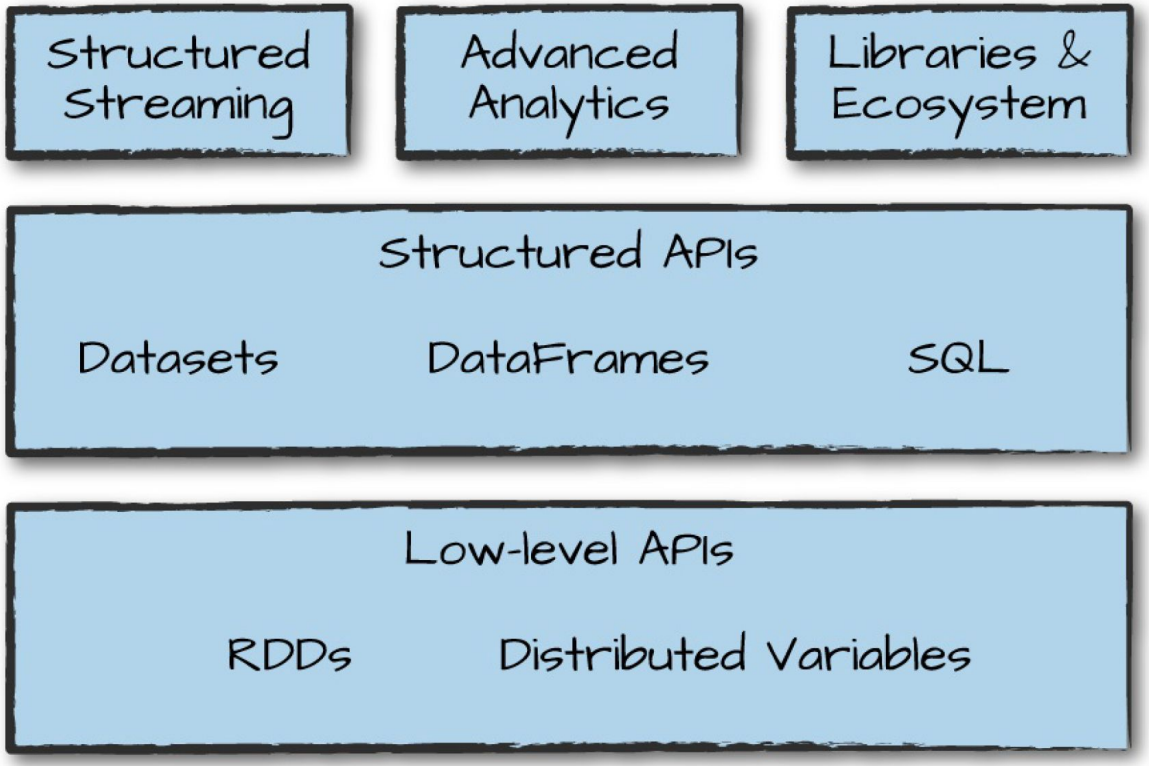
Точка входа в приложение

С его помощью можно:

- Управлять параметрами spark приложения
- читать и записывать данные в spark DataFrame
- исполнять SQL

```
1 from pyspark.sql import SparkSession
2
3 spark = SparkSession \
4     .builder \
5     .appName("Python Spark SQL basic example") \
6     .config("spark.some.config.option", "some-value") \
7     .getOrCreate()
```

Компоненты Spark



Типы данных в Spark

Data type	Value type	API to access or create data type
ByteType	int or long **Note:** Numbers are converted to 1-byte signed integer numbers at runtime. Make sure sure that numbers are within the range of -128 to 127.	ByteType()
ShortType	int or long **Note:** Numbers are converted to 2-byte signed integer numbers at runtime. Make sure sure that numbers are within the range of -32768 to 32767.	ShortType()
IntegerType	int or long	IntegerType()
LongType	long **Note:** Numbers are converted to 8-byte signed integer numbers at runtime. Make sure sure that numbers are within the range of -9223372036854775808 to 9223372036854775807.Otherwise, convert data to decimal.Decimal and use DecimalType.	LongType()
FloatType	float **Note:** Numbers are converted to 4-byte single-precision floating point numbers at runtime.	FloatType()
DoubleType	float	DoubleType()
DecimalType	decimal.Decimal	DecimalType()
StringType	string	StringType()
BinaryType	bytearray	BinaryType()
BooleanType	bool	BooleanType()
TimestampType	datetime.datetime	TimestampType()
DateType	datetime.date	DateType()
ArrayType	list, tuple, or array	ArrayType(elementType, [containsNull]) **Note:**The default value of containsNull is True.

Типы данных в Spark

ArrayType	list, tuple, or array	ArrayType(elementType, [containsNull]) **Note:**The default value of containsNull is True.
MapType	dict	MapType(keyType, valueType, [valueContainsNull]) **Note:**The default value of valueContainsNull is True.
StructType	list or tuple	StructType(fields) **Note:** fields is a Seq of StructFields. Also, two fields with the same name are not allowed.
StructField	The value type of the data type of this field (For example, Int for a StructField with the data type IntegerType)	StructField(name, dataType, [nullable]) **Note:** The default value of nullable is True.

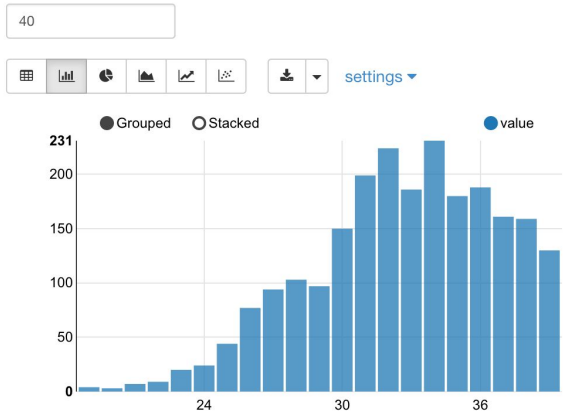
Домашнее задание

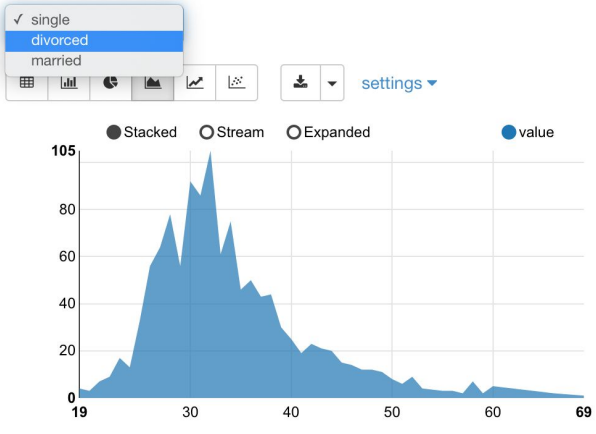
Визуализация

spark.table(“homework.bank”)

https://zeppelin.apache.org/docs/0.8.0/usage/dynamic_form/intro.html

1. Построить распределения клиентов по возрастам
2. Распределение по возрасту с динамическим численным параметром max_age
3. Распределение по возрасту с динамическим параметром “marital”





Домашнее задание

Преобразование типов

spark.table(“homework.bank”)

1. Вывести типы данных
2. Конвертировать возраст клиентов в String
3. Написать метод, который по типу данных в Python выводит тип данных в PySpark

Спасибо!

Каждый день вы становитесь лучше :)