SPARK

АНАЛИЗ ДАННЫХ С APACHE SPARK

ШКОЛА БОЛЬШИХ ДАННЫХ

Введение в Spark

МОДУЛЬ 1

ШКОЛА БОЛЬШИХ ДАННЫХ

Что такое Spark

Apache Spark is a unified computing engine and a set of libraries for parallel data processing on computer clusters.

- способ параллельного программирования (как **Map Reduce**)
- набор библиотек (**Scala**) и выполняемых файлов (**spark-submit**)
- гораздо более функциональный, чем Map Reduce
  - ✓ похож на **pandas**
  - ✓ активно развивается (недавно вышла версия 3.0)
- нет привязки к Hadoop, но мы будем рассматривать Spark именно в связке с Hadoop

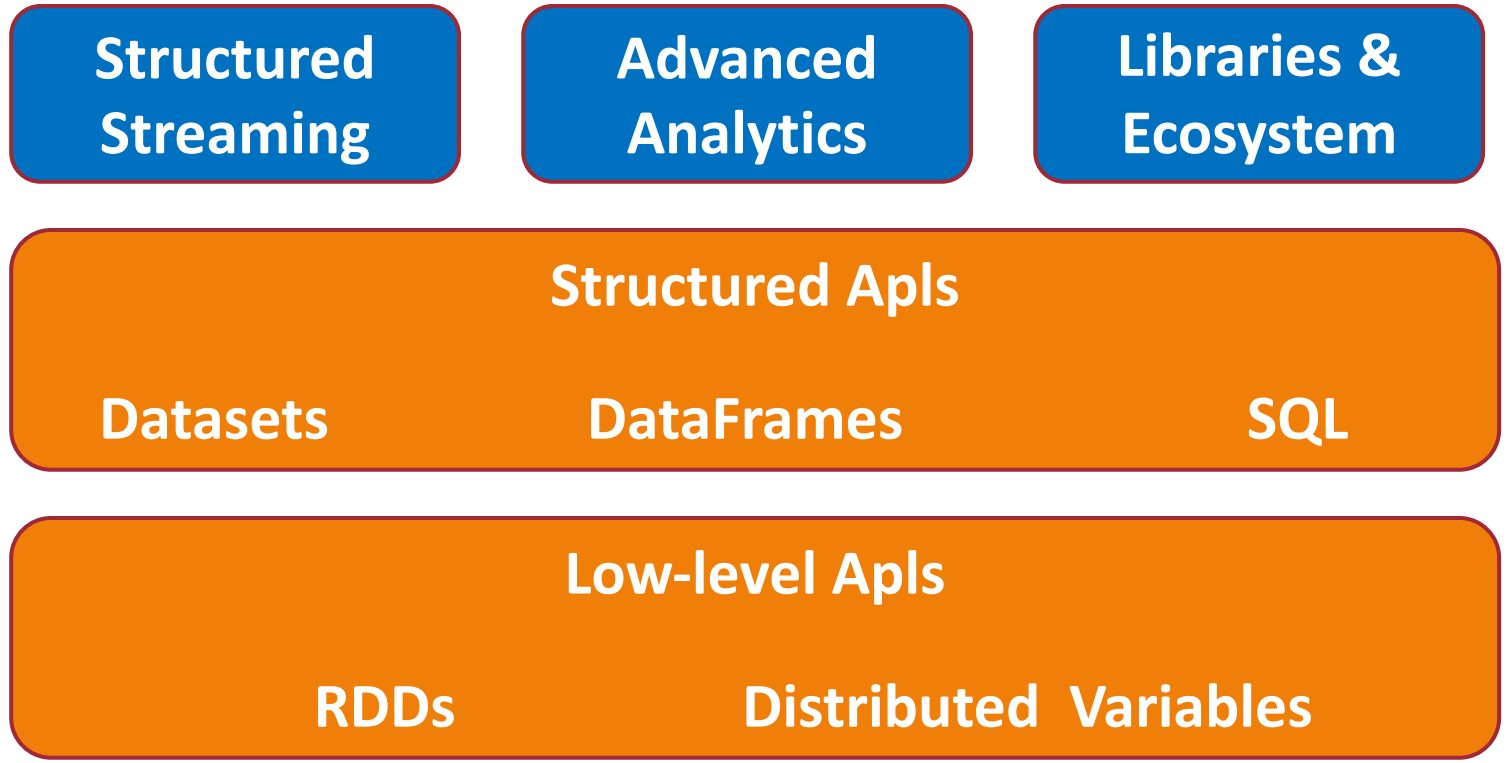
История Spark

- **2009**: проект в UC Berkley (Matei Zaharia)
- **2010**: open source
- **2013**: Apache Software Foundation

Версии:

- 1.0: **2014** (Spark SQL, spark-submit)
- 1.3: **2015** (Dataframe API)
- 2.0: **2016** (native SQL parsing, high leve' "structured" APIs)
- 2.4: **2018** (сейчас 2.4.6)
- 3.0: **2019** (preview)

Что входит в Spark



Если технически

- Spark SQL + Dataframes
- Spark Streaming
- Spark MLib
- Spark GraphX

Spark и python

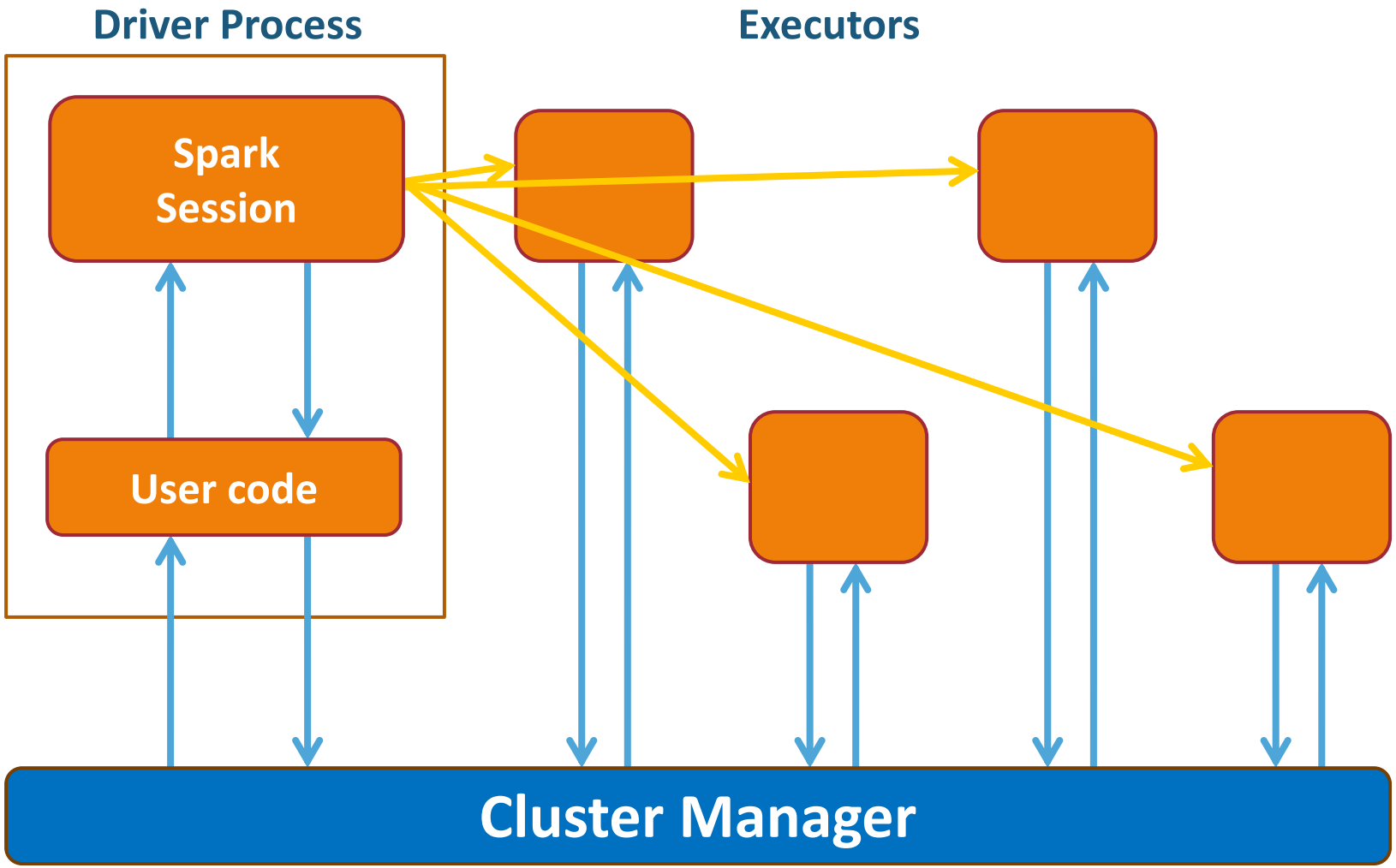
If you use just the Structured APIs, you can expect all languages to have similar performance chaaåracteristics

- Spark предоставляет API на нескольких языках
- наиболее популярное python API - **pyspark**
  - ✓ Scala, Java, SQL, R
- производительность не страдает

Как работают Spark приложения

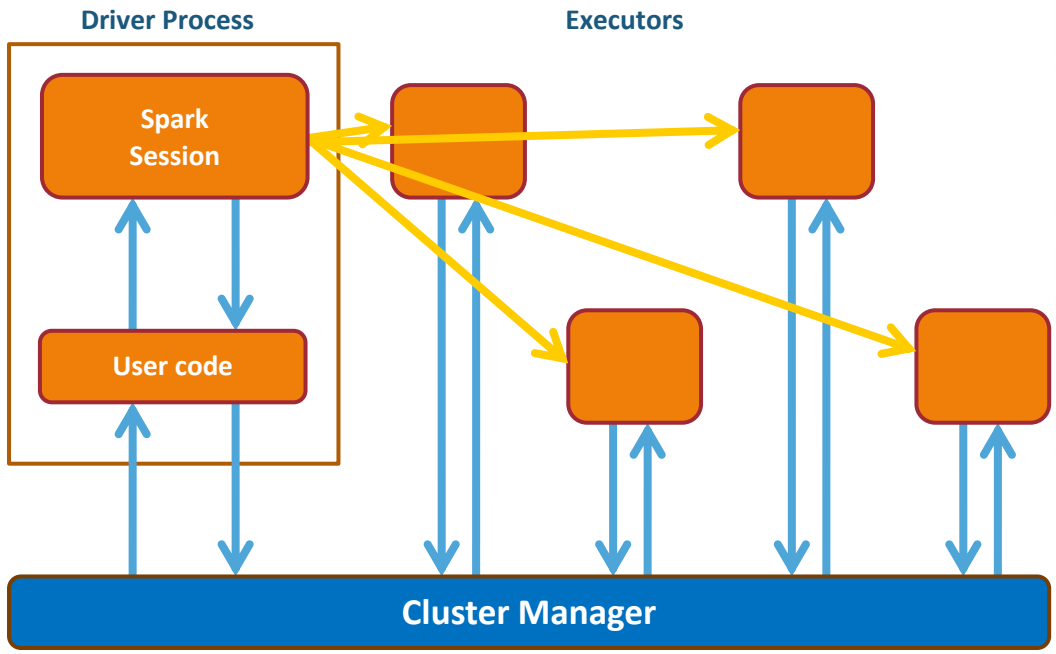
- **driver**: исполняет программу
  - ✓ поддерживает информацию о приложении
  - ✓ исполняет код программы
  - ✓ планирует и запускает работу executor
- **executor**: выполняет вычисления
  - ✓ исполняет код, переданный driver (JAR)
  - ✓ отчитывается перед driver о состоянии процесса вычисления
- **cluster manager**: управляет физическими машинами и выделяет ресурсы spark приложениям
  - ✓ standalone, YARN, Mesos, Kubernetes
- **spark session**: объект, через который происходит работа со spark (driver)

Как работают Spark приложения



Запуск приложения и режимы исполнения

- **приложения** могут быть
  - ✓ интерактивными (**shell**) - python, scala, SQL
  - ✓ "обычными" - через **spark-submit** (включая python)
- **execution mode**: возможность управления размещением процессов (**driver**, **executors**)
  - ✓ **cluster mode**: все в кластере
  - ✓ **client mode**: driver работает вне кластера, executor-ы - в кластере
  - ✓ **local mode**: все работает локально (потоки)



Вопросы?