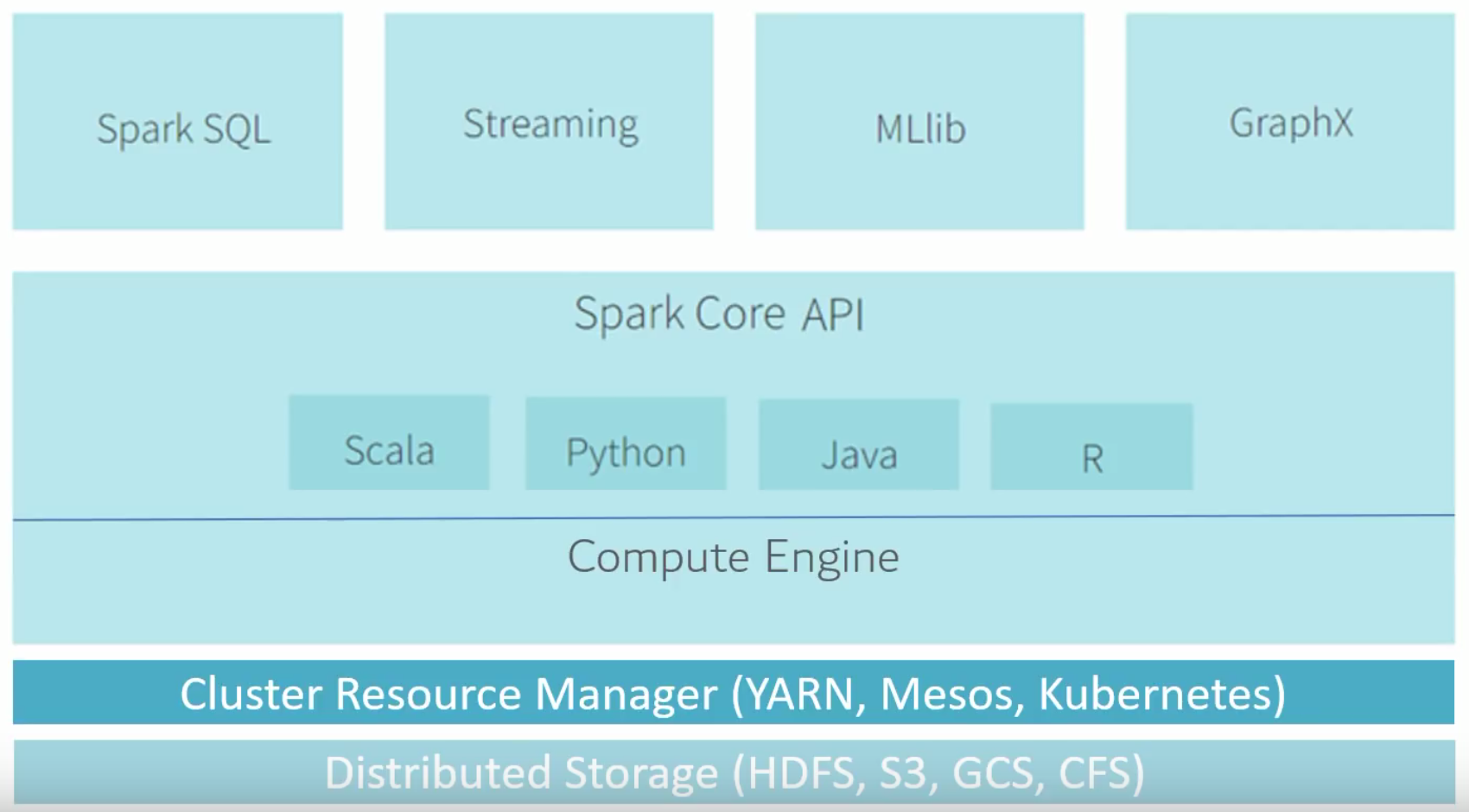
**Apache Spark —** это фреймворк с открытым исходным кодом для параллельной обработки и анализа слабоструктурированных данных в оперативной памяти.

Главные преимущества Spark — производительность, удобный программный интерфейс с неявной параллелизацией и отказоустойчивостью. Spark поддерживает четыре языка: Scala, Java, Python и R.

Главное отличие от Hadoop: способ обращения к данным. Hadoop сохраняет данные на жесткий диск на каждом шаге алгоритма MapReduce, а Spark производит все операции в оперативной памяти. Благодаря этому Spark выигрывает в производительности до 100 раз и позволяет обрабатывать данные в потоке.



Фреймворк состоит из пяти компонентов: ядра и четырех библиотек, каждая из которых решает определенную задачу:

1. **Spark Core** — основа фреймворка. Оно обеспечивает распределенную диспетчеризацию, планирование и базовые функции ввода-вывода.
2. **Spark SQL** — одна из четырех библиотек фреймворка для структурированной обработки данных. Она использует структуру данных, называемую DataFrames и может выступать в роли распределенного механизма запросов SQL. Это позволяет выполнять запросы Hadoop Hive до 100 раз быстрее.
3. **Spark Streaming** — простой в использовании инструмент для обработки потоковых данных. Несмотря на название, Spark Streaming не обрабатывает данные в реальном времени, а делает это в режиме micro-batch. Создатели Spark утверждают, что производительность от этого страдает несильно, поскольку минимальное время обработки каждого micro-batch 0,5 секунды. Библиотека позволяет использовать код приложений batch-анализа для потоковой аналитики, что облегчает реализацию λ-архитектуры.  Spark Streaming легко интегрируется с широким спектром популярных источников данных: HDFS, Flume, Kafka, ZeroMQ, Kinesis и Twitter.
4. **MLlib** — это распределенная система машинного обучения с высокой скоростью. Она в 9 быстрее своего конкурента — библиотеки Apache Mahout при тестировании бенчмарками на алгоритме чередующихся наименьших квадратов (ALS).

MLlib включает в себя популярные алгоритмы:

* классификация,
* регрессия,
* деревья принятия решений,
* рекомендация, кластеризация,
* тематическое моделирование.

     5. **GraphX** — это библиотека для масштабируемой обработки графовых данных. GraphX не подходит для графов, которые изменяются транзакционным методом: например, базы данных.

Spark работает:

* в среде кластеров Hadoop на YARN,
* под управлением Mesos,
* в облаке на AWS или других облачных сервисах,
* полностью автономно

Он же поддерживает несколько распределенных систем хранения:

* HDFS
* OpenStack Swift
* NoSQL-СУБД
* Cassandra
* Amazon S3
* Kudu
* MapR-FS

