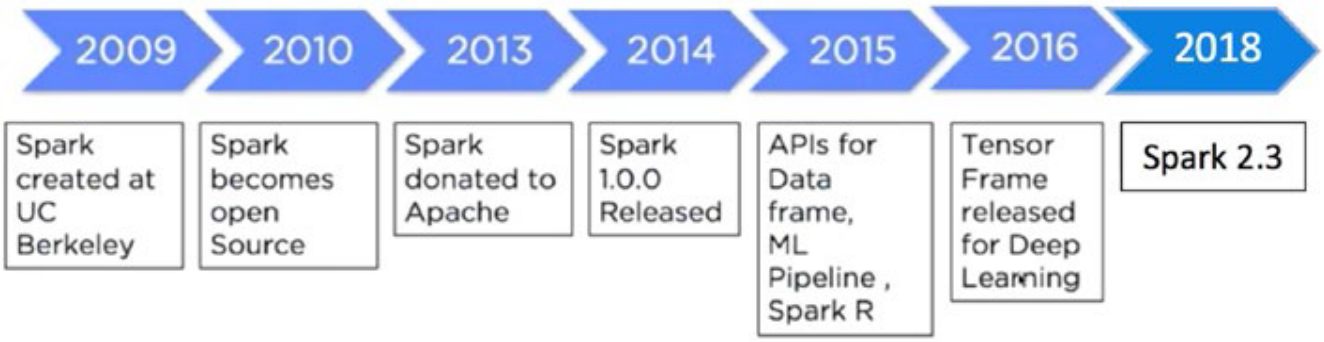
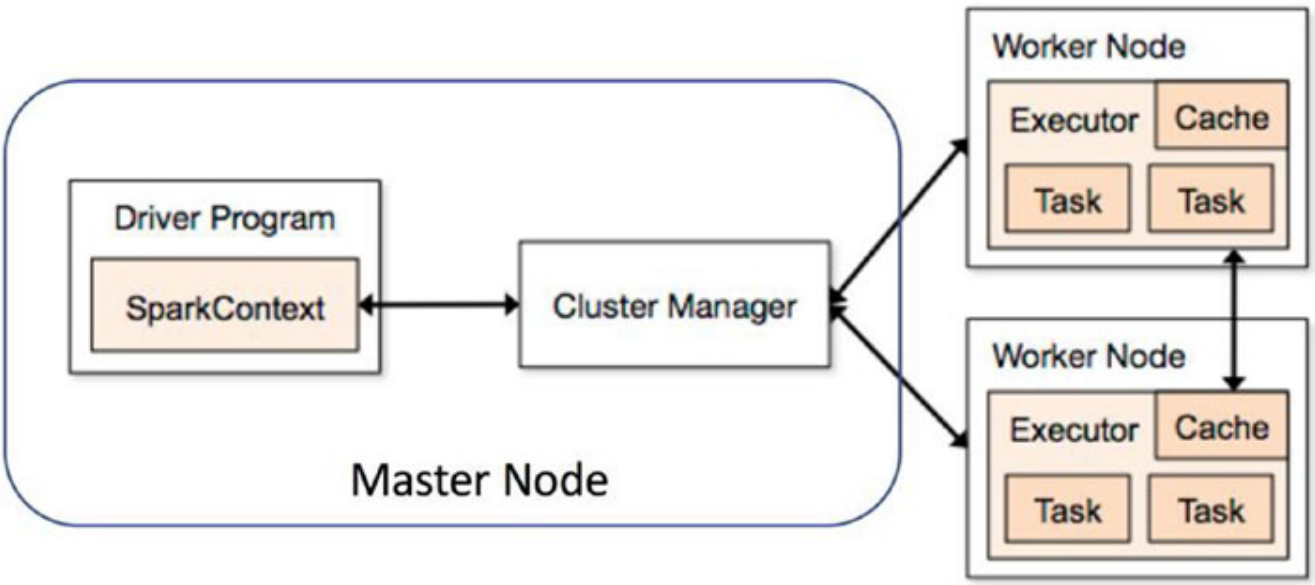
# 1. 数据的演变

## 1.1 数据的生成

## 1.2 Spark

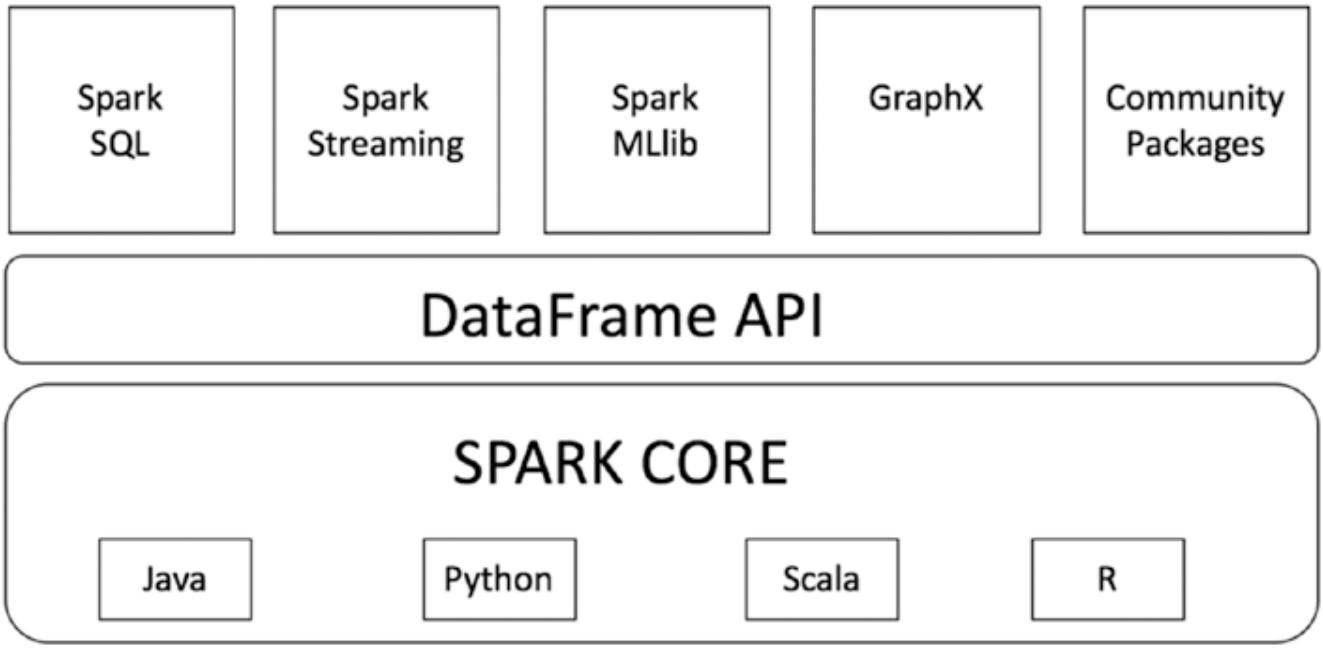



`Spark`发展史




`Spark`之所以广受欢迎，主要是因为它很容易用于数据处理、机器学习和流式数据；而且由于它可以进行所有内存计算，所以速度相对较快。由于Spark是一种通用的数据处理引擎，因此它可以很容易与各种数据源一起使用，例如`HBASE, Cassandra, Amazon S3, HDFS`等。`Spark`为用户提供了四种语言使用它：`Java, Python, Scala和R`。

### 1.2.1 spark core




## 1.2.2 spark 组件

### 1.2.2.1 spark sql

这个组件主要处理结构化数据处理。关键思想是获取更多关于数据结构的信息，以执行额外的优化。它可以被视为分布式 `SQL` 查询引擎。

### 1.2.2.2 spark streaming

该组件以可扩展和容错的方式处理实时流数据。它使用微批处理来读取和处理传入的数据流。它创建流数据的微型批处理，执行批处理，并将其传递到某个文件存储或实时仪表板。 `Spark Streaming` 可以接受从 `Kafka` 和 `Flume` 多个源接收数据。

### 1.2.2.3 spark mllib

此组件用于以分布式方式在大数据上构建机器学习模型。当数据量很大时，使用 `Python` 的 `scikit learn` 库构建 `ML` 模型的传统技术面临许多挑战，而 `Mllib` 的设计方式则提供了大规模的功能工程和机器学习。 `Mllib` 实现了分类、回归、聚类、推荐系统和自然语言处理的大部分算法。

### 1.2.2.4 spark graphX/Graphframe

该组件擅长图形分析和图形并行执行。图形框架可以用来理解底层关系，并从数据中可视化洞察。

# 1.3 设置环境

### 1.3.1 windows

1. Anaconda (Python 3.x)
2. Java (in case not installed)
3. Apache Spark latest version
4. Winutils.exe

### 1.3.2 anaconda

### 1.3.3 java安装

### 1.3.4 spark安装

选择 `Pre-built for Apache Hadoop 2.7 and later` 安装。下载得到 `.tgz` 文件。解压缩。

下载winutils.exe

https://github.com/steveloughran/winutils/blob/master/hadoop-2.7.1/bin/winutils.exe

设置环境变量

- HADOOP_HOME
- SPARK_HOME
- PYSPARK_DRIVER_PYTHON
- PYSPARK_DRIVER_PYTHON_OPTS
- JAVA_HOME

```
import pyspark
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()
```

### 1.3.5 IOS

### 1.3.6 docker

### 1.3.7 datatricks

## 1.4 结论

这一章，讲述了Spark结构，组件，不同环境的设置。