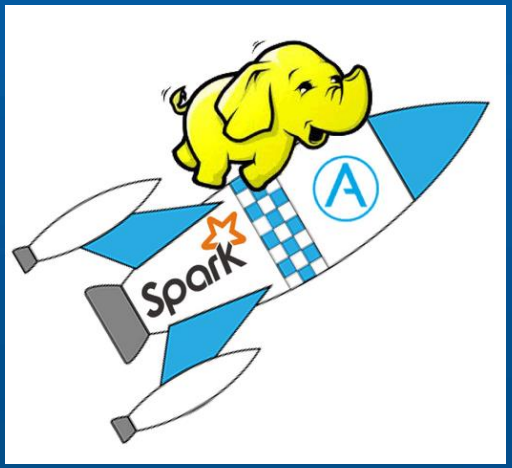
Spark SQL 模块　　20:30开始

高级互联网大数据架构师：Yasaka



QQ群： **172599077 , 156927834**

北京尚学堂大数据极限班课程官网地址：
http://www.bjsxt.com/html/cloud/

好消息！！！ 大数据登陆上海滩！！！

(北京)大数据线下班将于5月6日开班！火热报名中！！！

(上海)大数据线下班将于6月21日开班！火热报名中！！！

-- 老师面授课程！传统式教室教学已开班多期！学习完美就业！

(北京)大数据周末班将于5月7日再次开班！火热报名中！！！

线上班Hadoop阶段推出后，第二阶段Spark线上班5月22日正式推出！！！

--附送随堂讲课视频

贾老师：1786418286

何老师：1926106490

詹老师：2805048645

讨论技术可以加入以下QQ群： 172599077，156927834

第一阶段linux+搜索+hadoop体系

Linux基础 → shell编程 → 高并发架构 → lucene,solr搜索 → hadoop体系 →

HDFS → mapreduce → hbase → zookeeper → hive →

flume → sqoop → 项目实战一

第二阶段机器学习

R语言 → mahout → 项目实战二

第三阶段storm流式计算

kafka → storm → redis → 项目实战三

第四阶段spark内存计算

scala编程 → spark core → spark sql → spark streaming → spark mllib →

spark graphx → 项目实战四 → python机器学习 → spark python编程 → 项目实战五 →

项目实战六

第五阶段云计算平台

docker → kvm → openstack云计算 → 就业指导

教学多重保障：

1，贯彻实战教育理念

2，每节随堂笔记，有图有代码

3，提供服务器配置，搭建步骤说明

4，有问题老师一对一辅导

5，同学们良好的学习氛围

6，尚学堂科技有限公司是一个实体公司，已做教育多年，有着良好的口碑

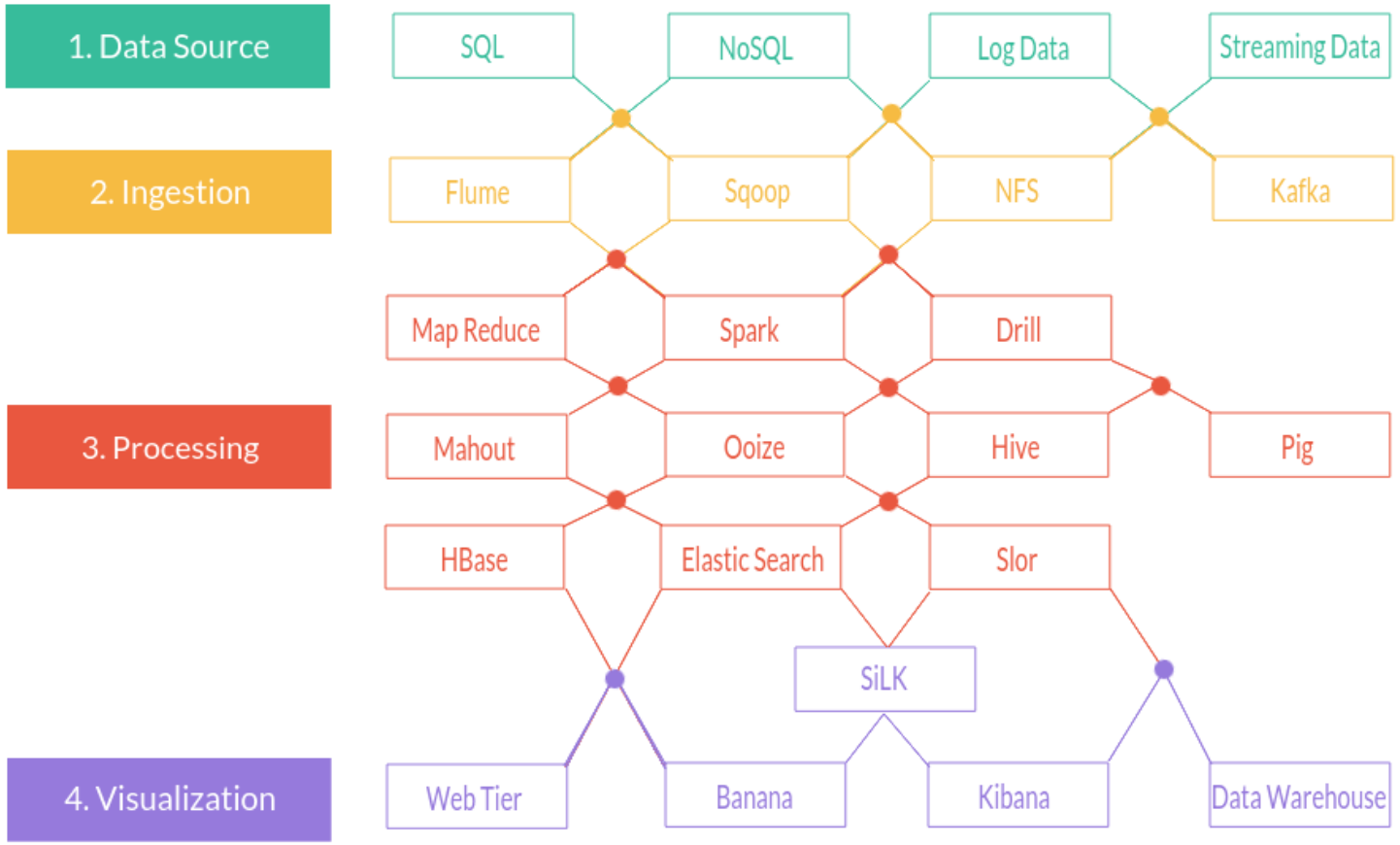
7，尚学堂大数据班老师均有丰富的授课经验，线下线上课都有

官网：

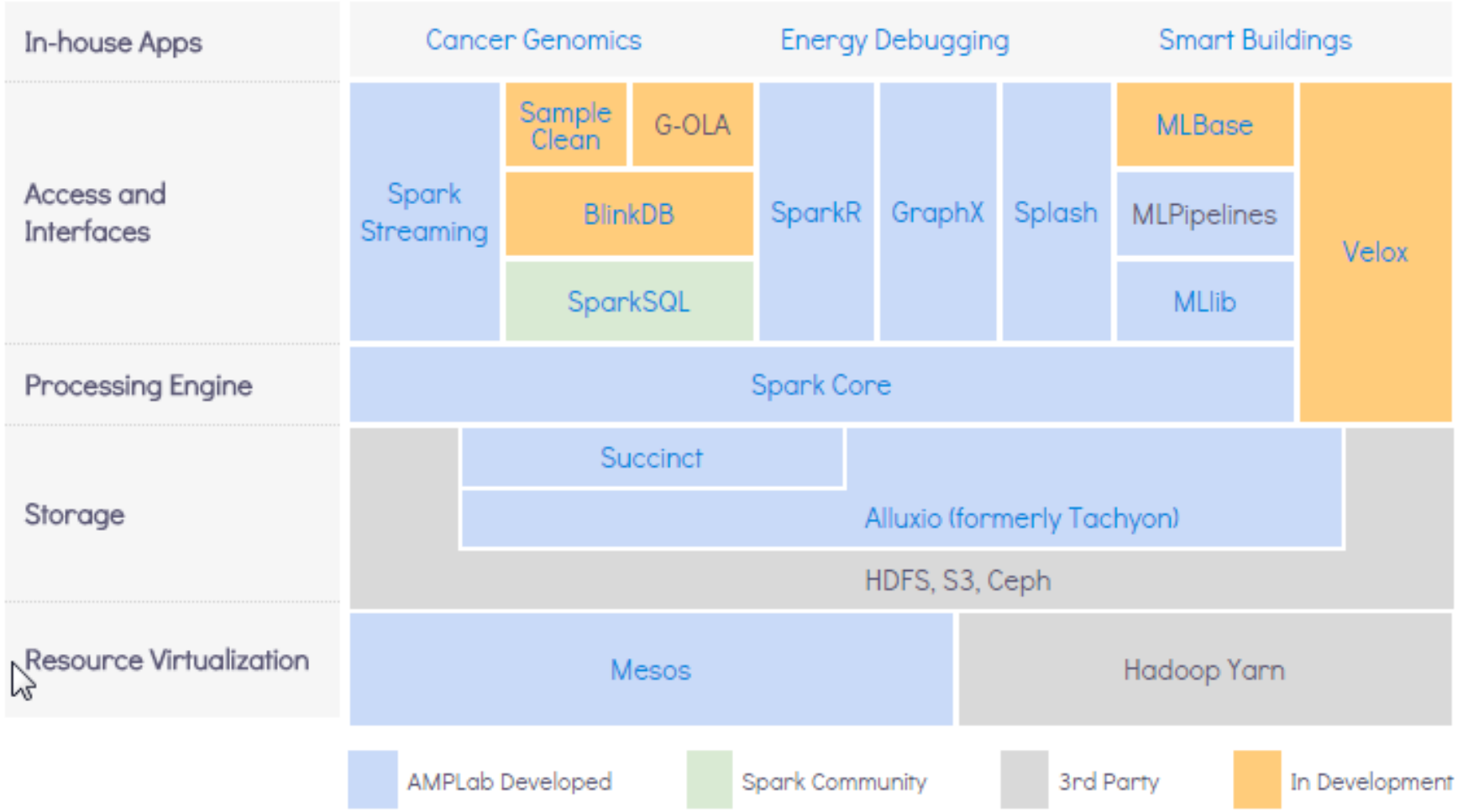
http://www.bjsxt.com/html/cloud/

- 分享主题内容
  - 1.Spark在大数据生态的位置
  - 2.Spark生态系统
  - 3.Spark SQL模块
  - 4.Spark SQL 与 DataFrame
  - 5.DataFrame操作
  - 6.Spark SQL底层架构
  - 7.Spark 为何如此令人着迷

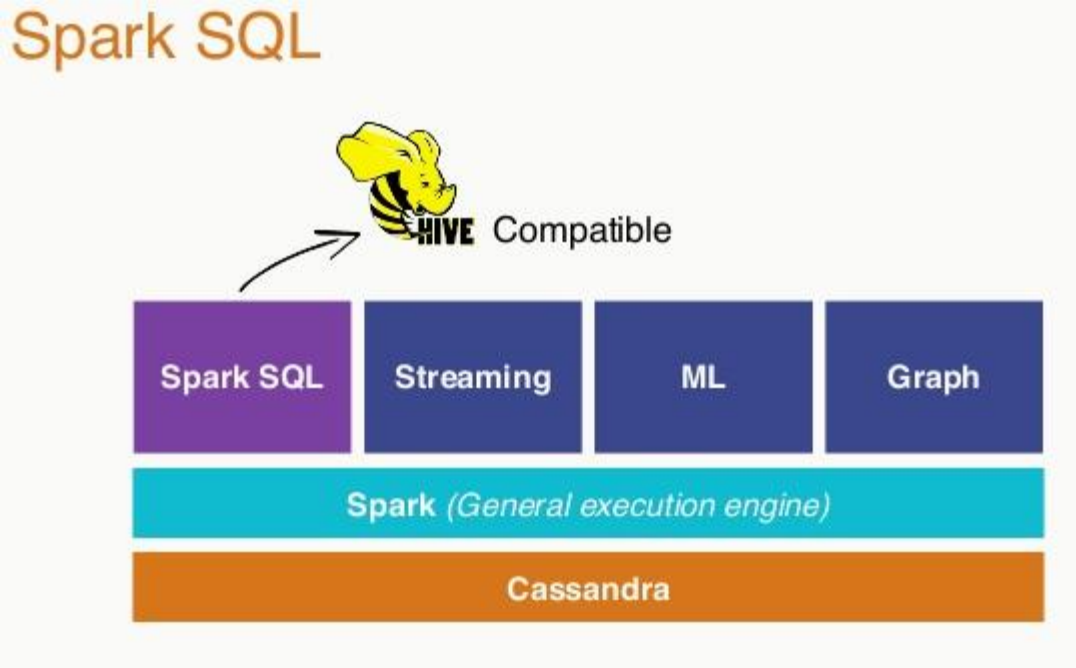
- Is Apache Spark going to replace Hadoop?

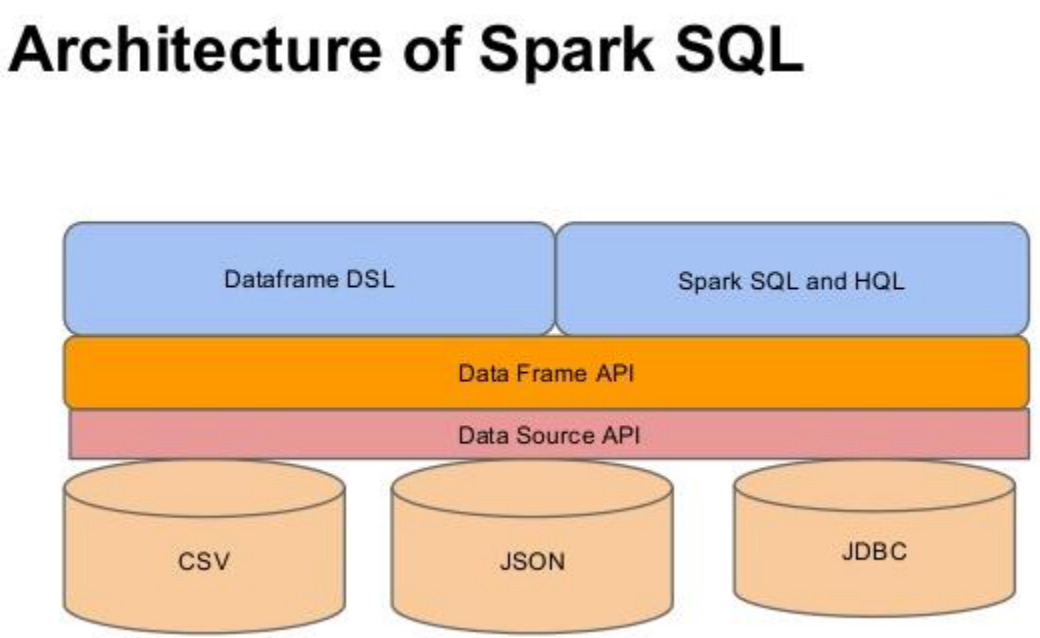


- http://spark.apache.org/



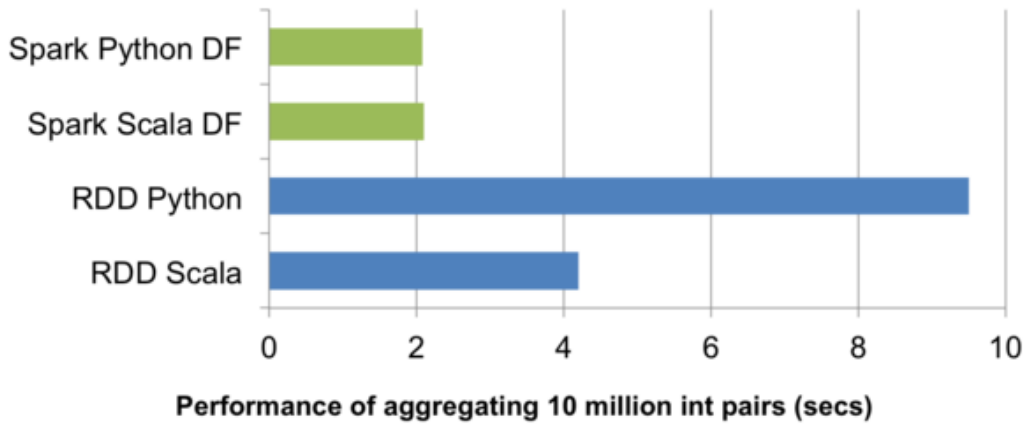
- Spark SQL模块　http://spark.apache.org/sql/





- Dataframe的最佳搭档——Spark SQL
  - Spark SQL是Spark的核心组件之一，于2014年4月随Spark 1.0版一同面世，在Spark 1.3当中，Spark SQL终于从alpha阶段毕业，除了部分developer API以外，所有的公共API都已经稳定，可以放心使用了。
  - Spark 1.3更加完整的表达了Spark SQL的愿景：让开发者用更精简的代码处理尽量少的数据，同时让Spark SQL自动优化执行过程，以达到降低开发成本，提升数据分析执行效率的目的。为此，在Spark 1.3中引入了与R和Python Pandas接口类似的DataFrame API
  - 与RDD类似，DataFrame也是一个分布式数据容器。然而DataFrame更像传统数据库的二维表格，除了数据以外，还掌握数据的结构信息，即schema。同时，与Hive类似，DataFrame也支持嵌套数据类型（struct、array和map）。从API易用性的角度上 看，DataFrame API提供的是一套高层的关系操作，比函数式的RDD API要更加友好，门槛更低。由于与R和Pandas的DataFrame类似，Spark DataFrame很好地继承了传统单机数据分析的开发体验。

- RDD vs DataFrame
- the Catalyst optimizer
- Tungsten execution engine



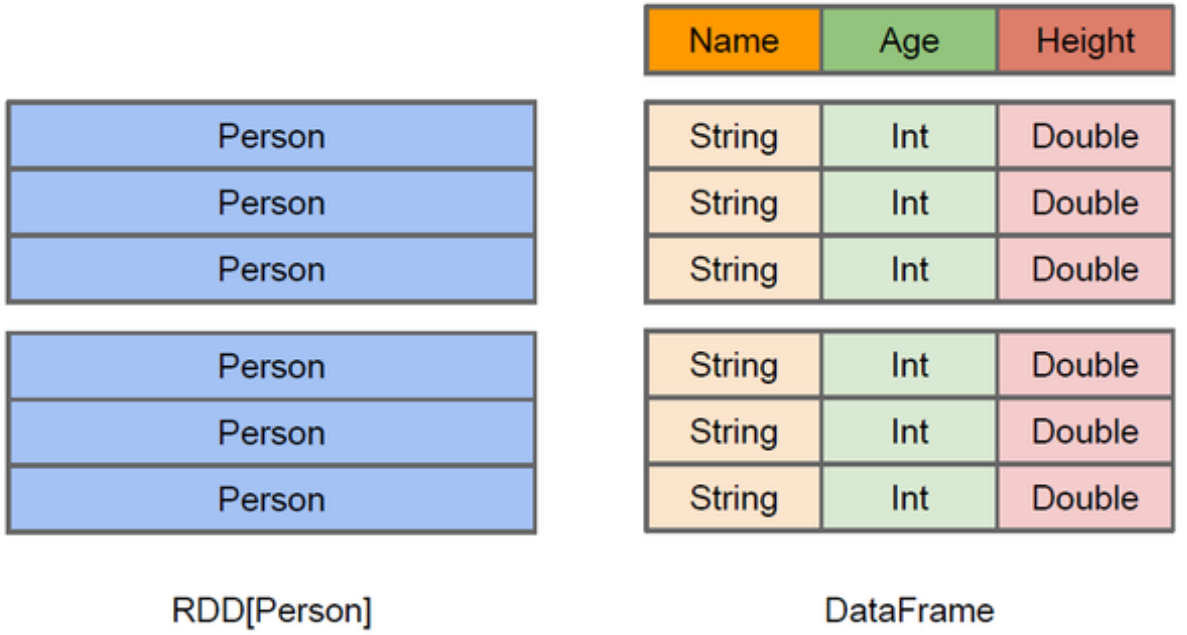


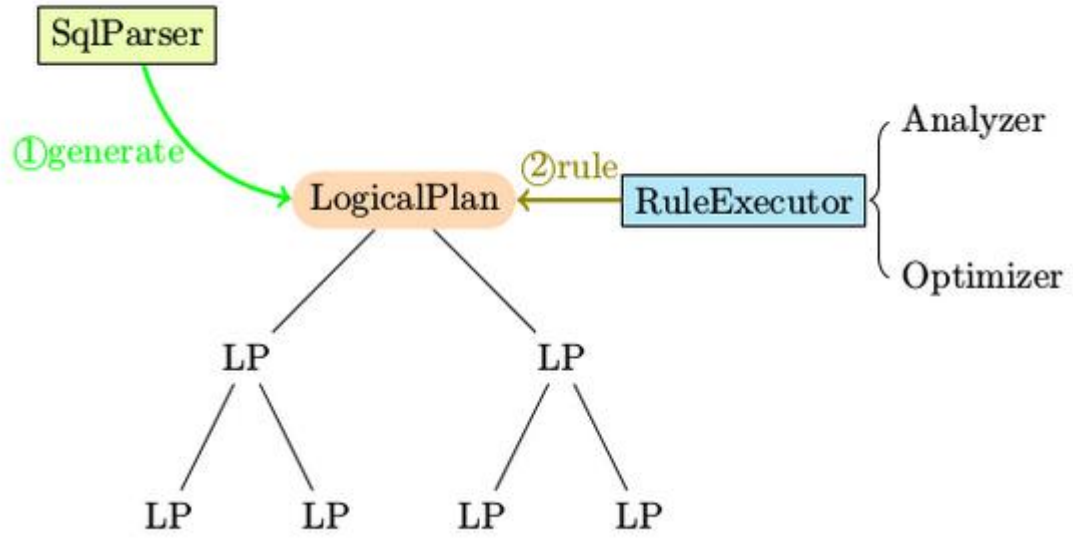
图2：DataFrame和 RDD的区别

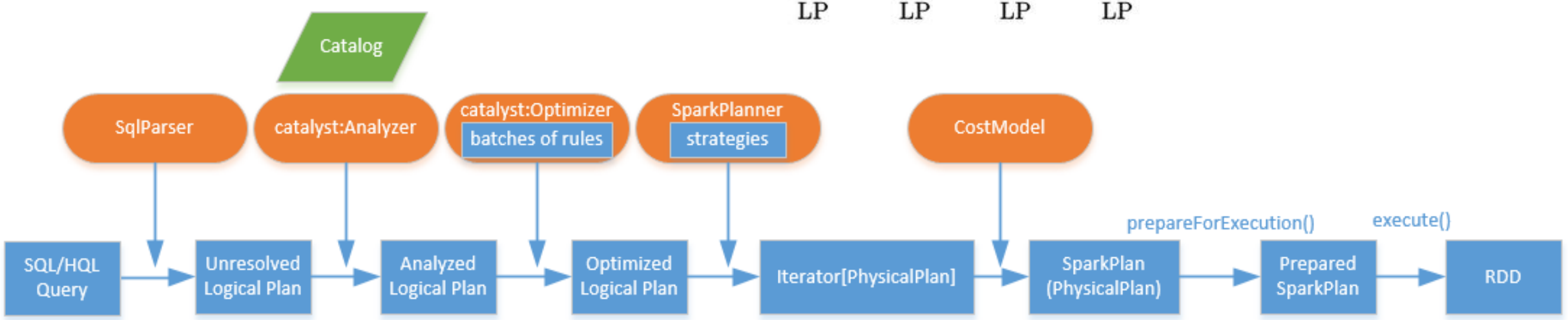
- http://spark.apache.org/docs/latest/sql-programming-guide.html

```scala
val sc: SparkContext // An existing SparkContext.
val sqlContext = new org.apache.spark.sql.SQLContext(sc)

// this is used to implicitly convert an RDD to a DataFrame.
import sqlContext.implicits._
```

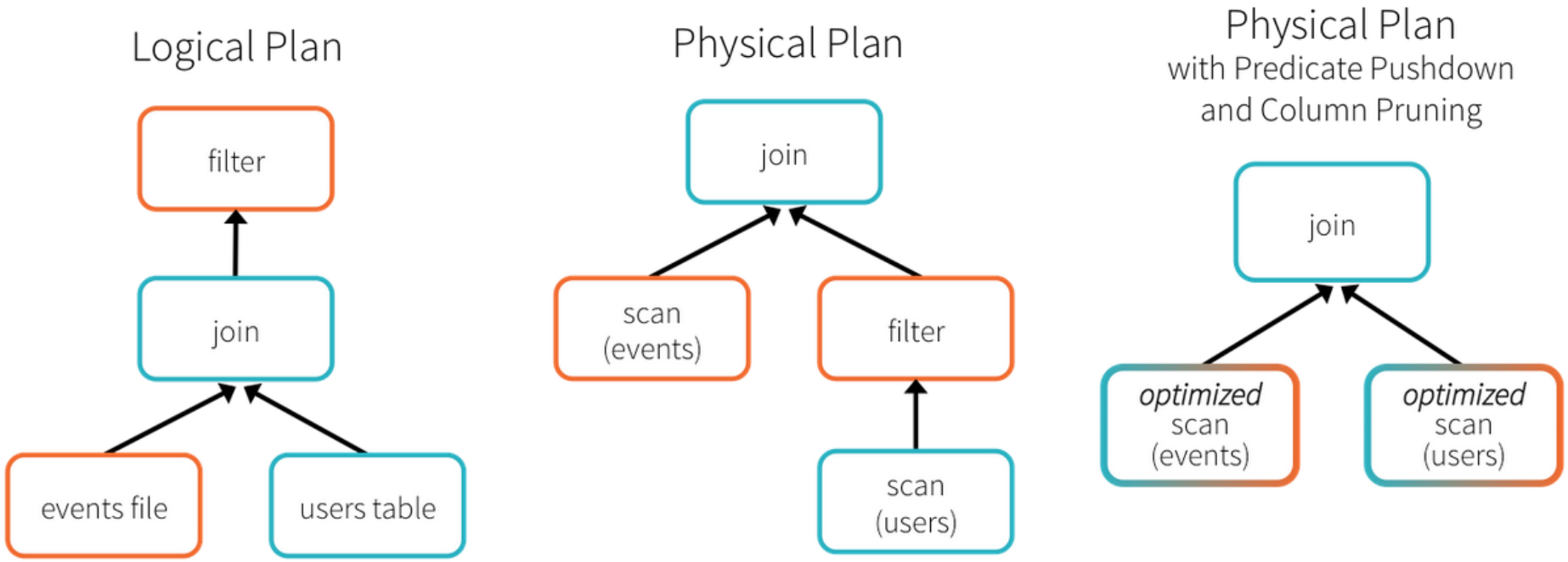
- Spark SQL 底层架构



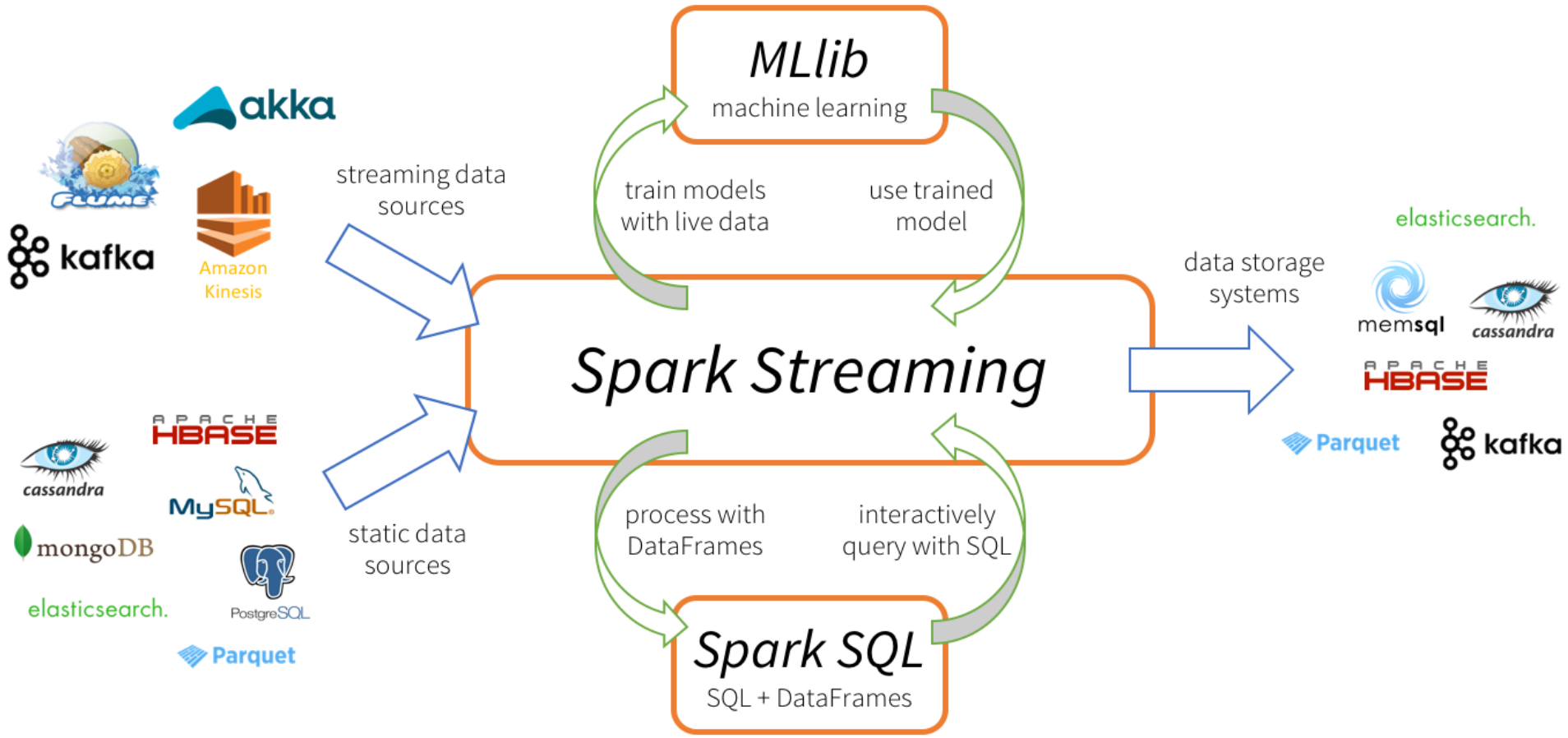


```
def add_demographics(events):
    u = sqlCtx.table("users")                    # Load partitioned Hive table
    events \
      .join(u, events.user_id == u.user_id) \    # Join on user_id
      .withColumn("city", zipToCity(df.zip))     # Run udf to add city column

events = add_demographics(sqlCtx.load("/data/events", "parquet"))
training_data = events.where(events.city == "New York").select(events.timestamp).collect()
```



- ^_^


MLlib
machine learning
train models with live data
use trained model
Spark Streaming
Spark SQL
SQL + DataFrames
process with DataFrames
interactively query with SQL
streaming data sources
static data sources
data storage systems
akka
kafka
Amazon Kinesis
cassandra
HBASE
MySQL
mongoDB
elasticsearch.
PostgreSQL
Parquet
memsql
kafka

- Come here Quick ! Do not hesitate !

- Come here Quick ! Do not hesitate !