2019

企业级数据仓库实战

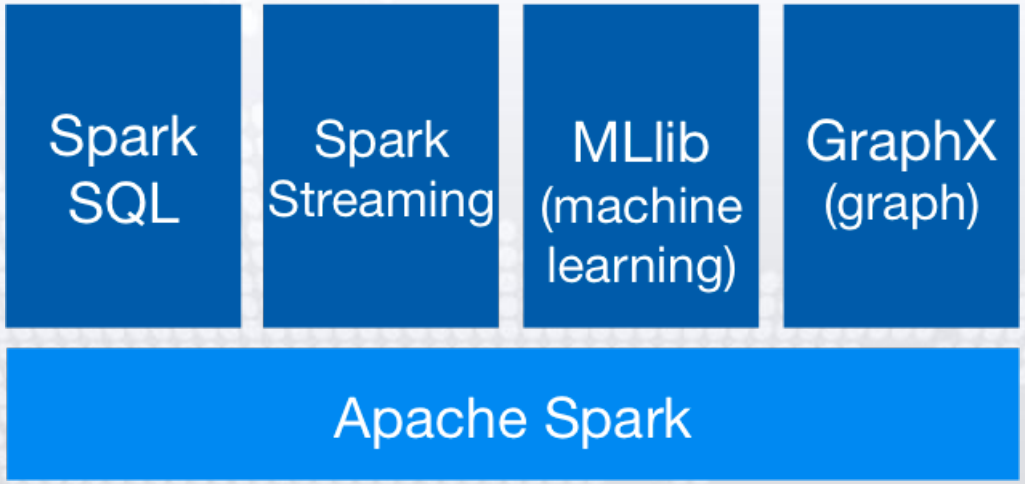
Spark基础

官方网站：https://spark.apache.org/

主要学习内容：

1、Spark SQL And DataFrames

2、Spark Streaming



为什么要学习Spark

Spark VS Hadoop

1、速度

基于内存计算，理论上比Hadoop快100倍

2、数据处理

Spark 同时支持批处理、流处理、数据挖掘和图计算，而Hadoop只能批处理

3、易用性

Spark支持多种算子，而Hadoop MR只有两种算子

4、语言

Spark主要使用Scala，比较好的支持链式编程，而Hadoop MR主要使用Java开发

相对来讲对链式编程支持比较少

5、延时

Spark Streaming 延迟比较低，支持到秒级别的微批处理，而Hadoop MR延迟比较高，一般支持到小时粒度

6、硬件

Spark需要中高级硬件，Hadoop对于硬件配置要求比较低

Spark SQL VS Hive VS Hive On Spark

1、版本发行

Spark SQL 跟随Spark 版本,而Hive与Hive On Spark 均跟随Hive版本发布

2、语法

均支持标准的SQL语法，即关系型数据库的SQL语法，在分析函数上，Hive比Spark SQL要丰富

3、开发者

Hive早先是由Facebook发布，后捐献给Apache 基金会，Spark SQL直接是Apache基金会下的项目

4、使用方法

Hive主要使用SQL或部分UDF进行开发

Spark SQL可以读取Hive中的数据，也可以配合Spark生态其他组件使用

5、交互式查询

Spark SQL支持交互式查询，时延较低，Hive时延较高

THANK YOU FOR YOUR GUIDANCE.

谢谢