Spark训练营第一期问题汇总

1. 读mysql这样的外部数据源的默认分区数怎么判断？

答：具体可以参考DataFrameReader.jdbc相关接口，默认是1个task去读取，也可以指定partition个数或者是分区条件

2. sparksql 读取其它存储的数据吗？

答：基于DataSource api可以接入多种数据源，比如hdfs、kudu、oss等等。根据场景可以选择，有些数据源的connector需要自己实现，社区暂时没有提供。

3. 取ceph数据场景多吗？

答：目前从业界实践来看，ceph的访问不是很常见。当然如上一个问题，这个也可以支持。ceph的特性这个有机会也可以聊聊。

4. Spark的DataFrames 和R语言的区别大吗？pandas里面的dataframe和spark的dataframe是一个概念吗？

答：三者的dataframe本质上都是一样的。不过spark层面的dataframe语义没有r和pandas支持的完备，缺乏矩阵的特性，主要用了类关系型数据表的特性。不过spark的实现是分布式的，在大数据场景下更有优势。感兴趣可以查阅一下论文：https://arxiv.org/abs/2001.00888

5. 请问ss怎么写hive？

答：有相关的sink，可以参考一下。这个可以搜索下

6. spark与Hadoop的关系是？

答：都是大数据生态的。spark应该比较的是hadoop mapreduce。hadoop还包含yarn hdfs等。

7. shuffle 重写了吧？

答：emr spark是重写了shuffle。包括我们也在做自己的remote shuffle service。除了shuffle还重写了code gen相关逻辑。性能大幅度提升

本次分享时间较短，很多内容没法覆盖，建议关注“Apache Spark技术交流社区”和查看官方相关文档。当然，后续也多参与一下活动。