基于hadoop生态圈的数据仓库设计

主要介绍基于hadoop生态圈的数仓设计，主要用于调优和解决数据存储、增量执行、数据查询等特性。

在一些业务场景中不只是业务流程和算法会影响到性能，数据仓库和设计和数据库的选取也会影响性能。

# 基于spark sql设计数据仓库

# 利用spark性能提高的数据仓库设计

# 关注查询性能

1. 方案1：Hive
2. 方案2：Spark缓存
3. 方案3：HBASE+ES

把全部的数据从HIVE数据库一次性读入HBASE，同时再存放elasticsearch全文索引，软件执行时，先通过elasticsearch 找到HABASE的ROWKEY，再到HBASE查询相关信息，加载到RDD进行处理。

# 增量处理

# 出发点和一些思考点

方便spark业务

提高业务性能

elastic search索引

spark sql和spark streaming

join是应该避免还是可以容忍？

小数据表是否应该基于一般的关系型数据库？