Spark企业级大数据项目实战 第1课

DATAGURU专业数据分析社区

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

http://edu.dataguru.cn

炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

课程介绍

课程大纲： http://www.dataguru.cn/article-12660-1.html

课程特色：基于Spark的一线生产项目

前置基础： 有一点Hadoop、Spark的基础

课程相关生态栈： Hadoop、Hbase、Kafka、ElasticSearch、Flume、Azkaban等

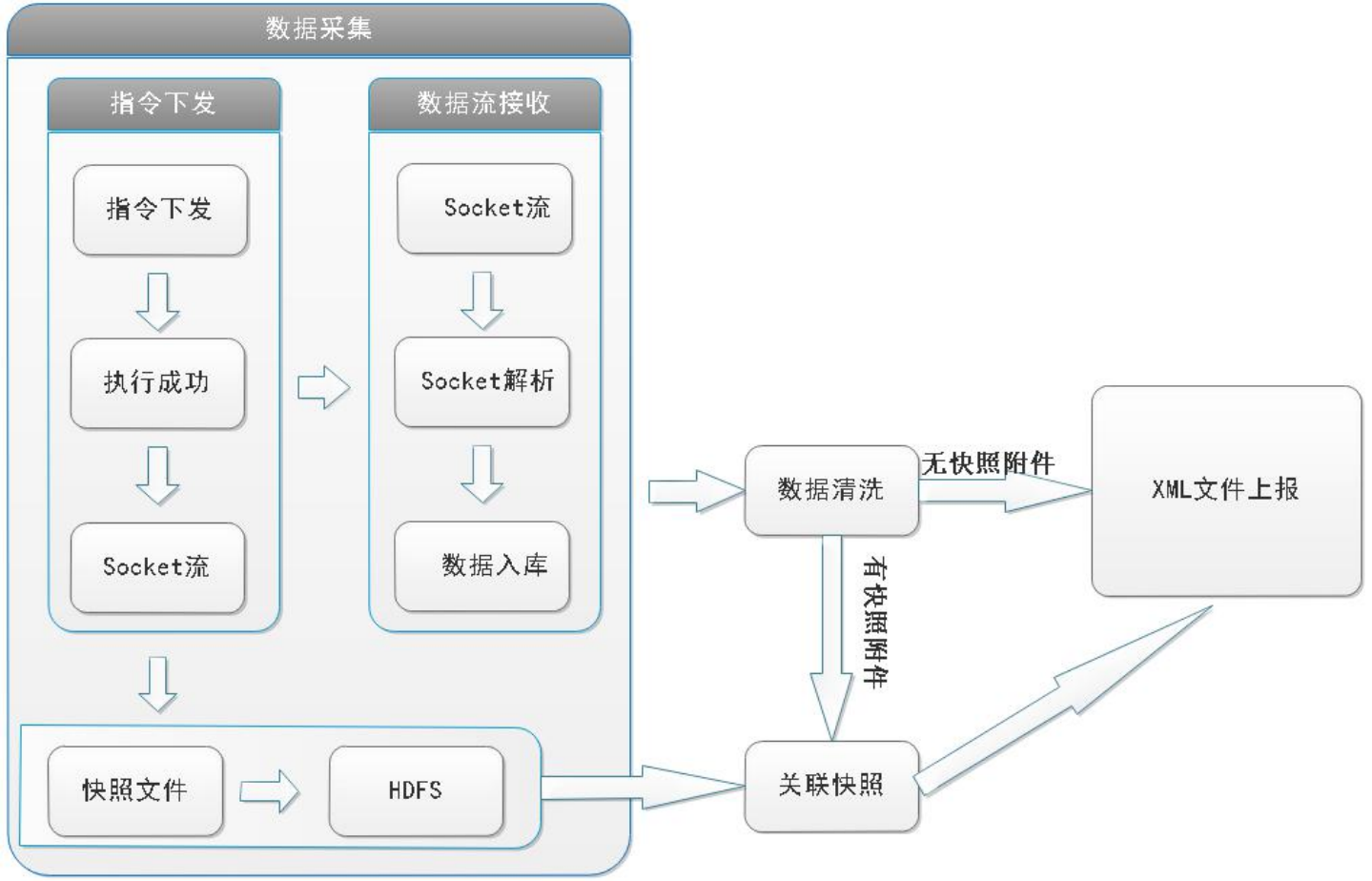
收获预期： 生产项目架构

Spark生产项目开发、优化、流程调度

大数据处理流程各环节（采集、清洗、分析调度等）的高可用

大数据周边生态圈（phoenix、Presto等）

项目1：Spark Streaming+Kafka保证数据零丢失（1）

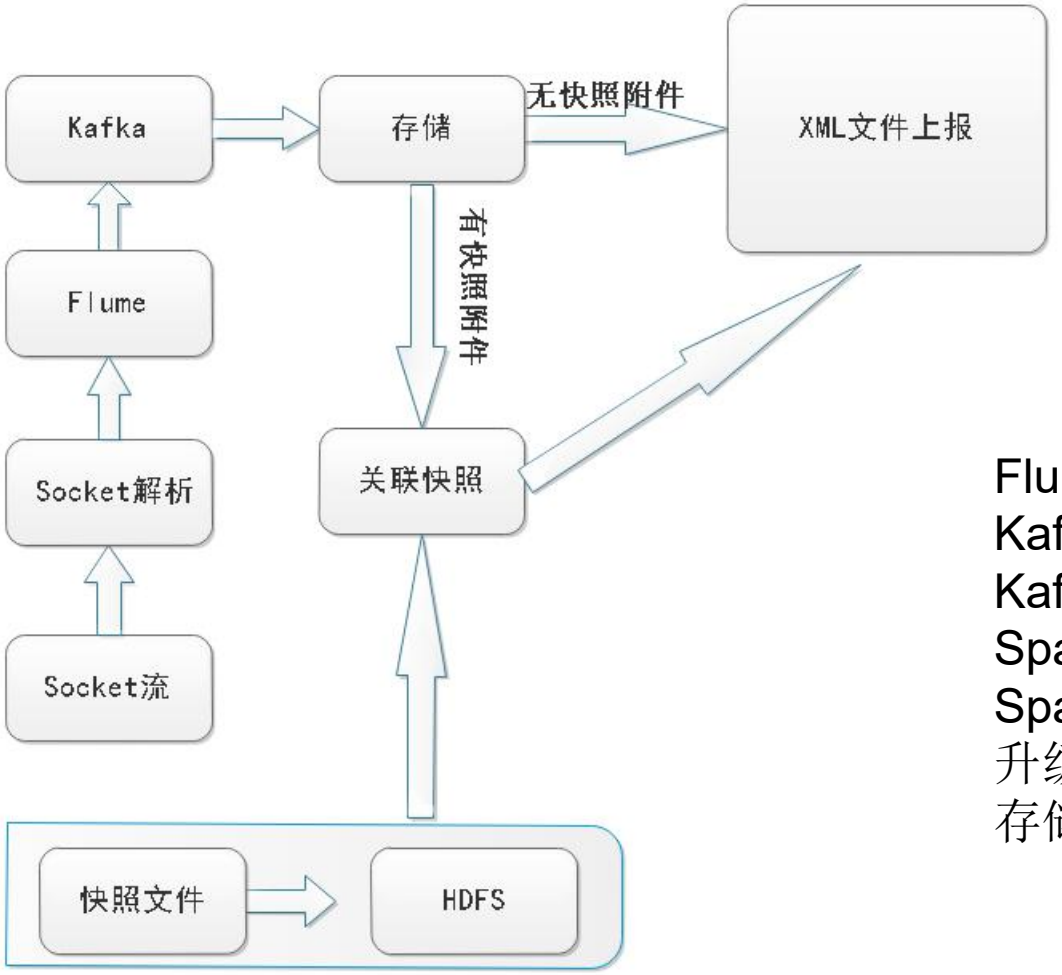


原数据处理流程

要求： 数据准确率99.6%
全流程10分钟
数据不达标后果很严重！！

性能瓶颈： Socket流解析
Oracle入库
数据清洗
XML文件上报

项目1：Spark Streaming+Kafka保证数据零丢失（2）



大数据处理流程

Flume如何保证高可用？
Kafka集群挂了？
Kafka的如何保证数据不丢失？
Spark Streaming程序挂了？
Spark Streming挂了很久，kafka积压大量数据， 性能？
升级？
存储如何选择？

项目1：Spark Streaming+Kafka保证数据零丢失（3）

1. Spark Streaming 整合Kafka的几种方式对比

2. Kafka的offset管理（Checkpoints、Hbase、Zookeeper等）

3. 三种计算语义（at most once、at least once、exactly once）

4. Spark Streaming + kafka整合Hbase、ElasticSearch、Oracle、Kafka（生产）等

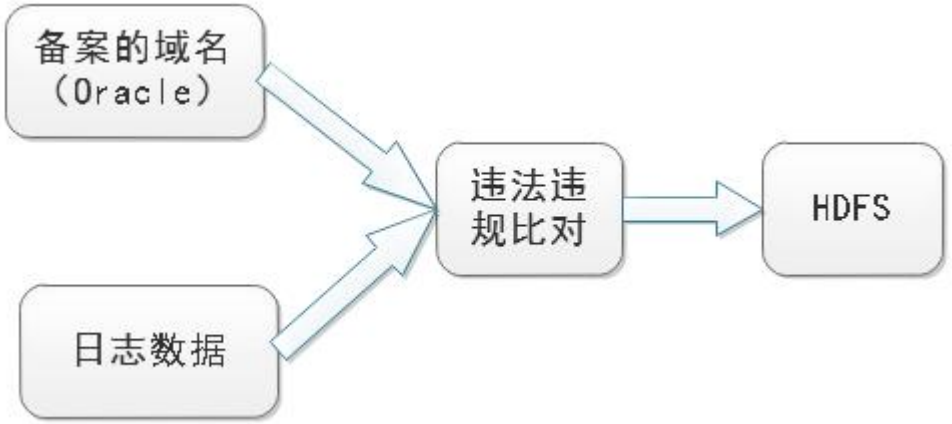
5. 如何实现exactly once语义

6. 四种大数据方案对比

7. 其他： 优化、坑等

项目2：离线日志分析（1）

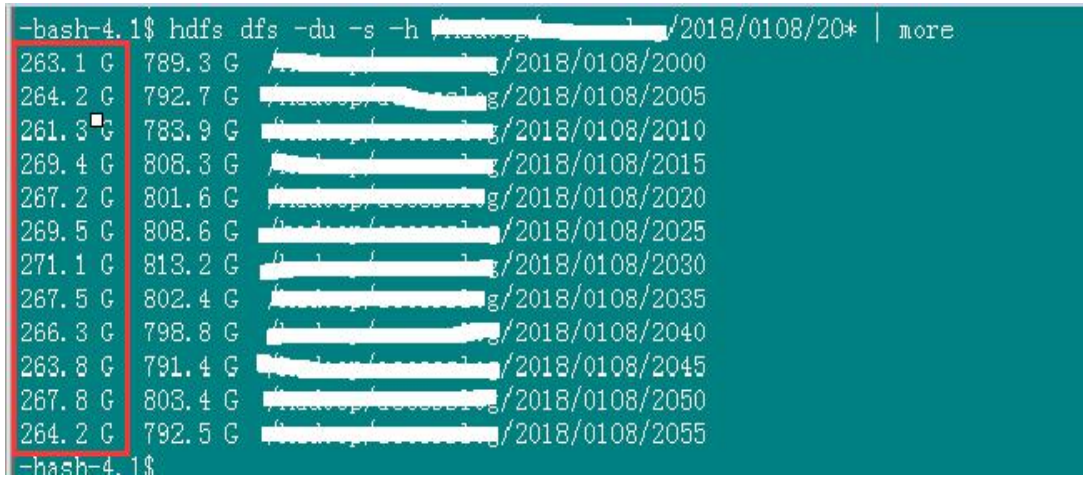
数据总体处理流程



准实时流程

问题分析：

1. 数据量非常大
2. 数据准确性
3. NameNode负载
4. 小文件过多
5. 文件存储格式
6. 数据处理效率



每5分钟的数据量

项目2：离线日志分析（2）

1. ETL流程分析

2. 文件存储格式对比、选择

3. Flume高可用（HDFS维护升级等如何保证数据不丢失）

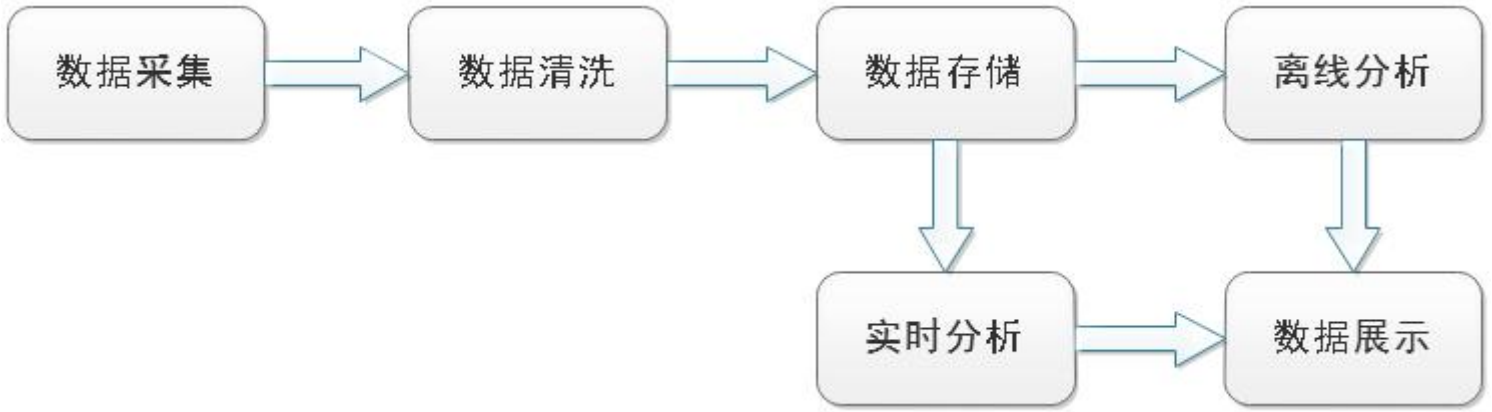
4. 解决小文件的几种方案

5. 解决数据准确性问题

6. Spark + Hive整合， 实现ETL流程调度

7. Tune Spark Jobs

项目3：企业预警实时监控（1）



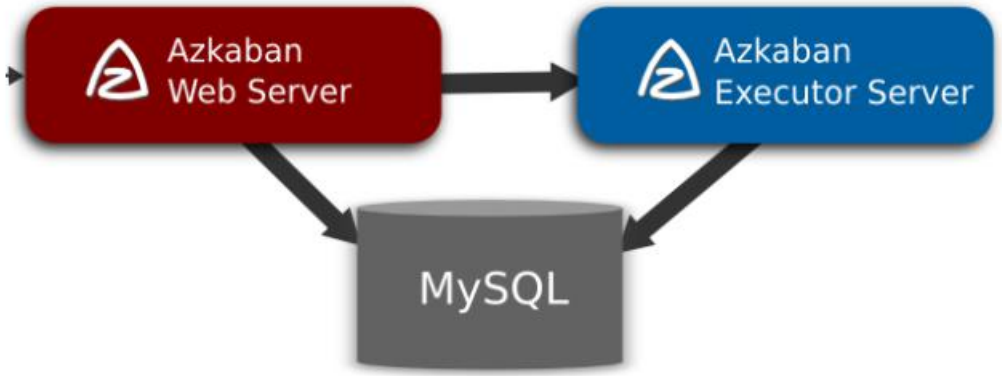
改造前：
1. 数据采集： 原始数据入HDFS目录
2. 数据清洗： Spark Streaming
3. 数据存储： Hive
4. 实时分析： 每5分钟启动离线分析任务

项目3：企业预警实时监控（2）

1. Spark Streaming监控文件目录开发、问题分析
2. 基于离线ETL取代Spark Streaming
3. Spark jobserver、Livy、Spark Thrift Server
4. 乱码处理
5. SQL on Hbase的几种方式

Azkaban：
- Server
- Executor
- MySQL

三个组件如何做HA？



其他

1. Presto
2. phoenix
3. Hbase二级索引
4. ElasticSearch + Hbase整合
5. Spark的Driver内存调优
6. 等等...

Thanks

FAQ时间