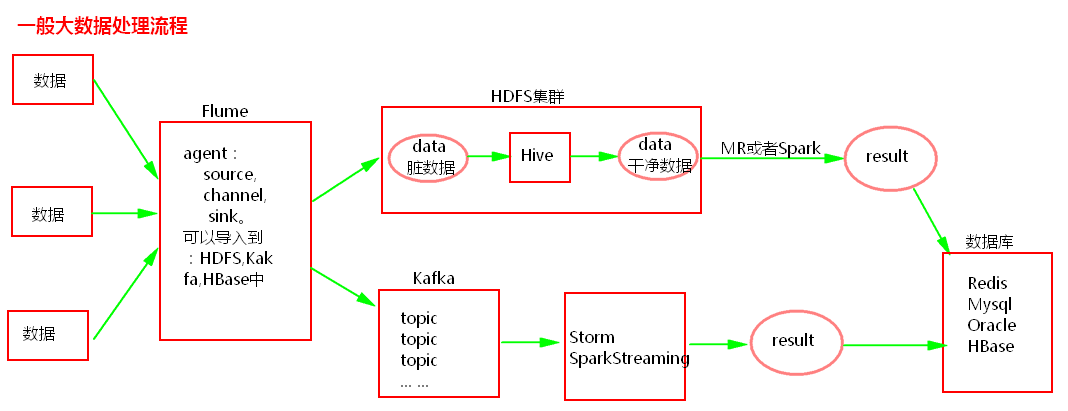
1. 项目介绍

详细见项目介绍文档。

1. 一般企业大数据架构

数据处理流程：



1. 公司有集群没有数据

分布式爬取数据，多节点爬取数据，一般将数据爬取到flume中，或者将数据直接爬取放入HDFS中。

1. 公司有集群有数据

每天每时每刻在产生数据，数据直接清洗放在HBase或者HDFS中。或者日志数据直接使用flume导入分布式文件中。

一般有了数据之后又分为两个大的方向处理数据：

1. 假设数据放在了HDFS集群中之后，一般下一步就要清洗数据，可以将数据通过Hive清洗，当然这里Hive一般使用外表，这样做的目的是可以将相同的数据只在HDFS中存入一份，避免过多的重复数据。清洗完成的数据一般又会放入Hive表中或者以结构化的数据放在HDFS上。得到清洗后的数据后一般会使用MR+Mahout或者使用Spark+Mllib来对数据进行分析处理，也可以对清洗后的数据使用SparkSQL来进行处理分析。之后，将分析完成的数据放入数据库中，如Redis,Mysql，Oracle中，供前端查询展示。
2. 如果数据放入了flume中，一般将数据sink到kafka中，不同数据的种类放入不同的topic中。然后对打入kafka中的数据进行流式处理，一般可以使用storm或者SparkStreaming对数据进行清洗，分析处理，然后将结果放到数据库中，如Redis,Mysql，Oracle中，以供前端页面来查询展示。
3. 关于运行Spark任务

* 如何将Spark任务提交到集群运行？

1. 最次也是脚本化执行Spark任务。
2. 平台化提交Spark任务。

* 平台化提交Spark任务流程图

submit后先将数据存入Mysql中，task当做唯一主键，这样做是为了简化任务执行失败时，可以直接在数据库中查询之前的提交的业务参数，当任务失败后，下次retry时方便执行。

submit后可以使用java调用liunx系统脚本，通过taskId得到系统中的业务参数数据。

脚本的内容：spark --master url --class jarPath taskId

注意： 假如使用tomcat实现平台化，那么tomcat应该部署在客户端。

* java代码中如何执行liunx脚本？

|  |
| --- |
| Process proc = Runtime.*getRuntime*().exec(“sh 脚本”);  proc.waitFor(); |
|  |
|  |

* sh脚本如何传参？

test.sh (chmod 755 test.sh)

|  |
| --- |
| 执行传参，多个参数空格隔开： |

1. 项目数据
2. 数据表

**monitor\_flow\_action表 ，监控车流量的数据**

date 日期 天

monitor\_id 卡口号

camera\_id 摄像头编号

car 车牌

action\_time 某个摄像头拍摄时间 s

speed 通过卡扣的速度

road\_id 道路id

area\_id 区域ID

**monitor\_camera\_info表 某一个卡扣对应的摄像头编号**

monitor\_id：卡扣编号

camera\_id：摄像头编号

1. 模拟数据类

com.spark.spark.test.MockData

1. 工具类

com.bjsxt.spark.util下

包含日期工具类，数字工具类，参数工具类，Spark工具类，字符串工具类。