1. 冷数据处理流程

logstash -> HDFS -> Hive -> ES (采用Snappy压缩，压缩比20%左右)

1. 概述

1）通过logstash的webhdfs output插件将数据源抽取到HDFS存储；

2）通过Hive创建外部数据表，将需要处理的导入Hive；

此过程主要的操作过程：

首先，创建textfile格式的外部表，将数据导入；

示例：

a.创建表

*CREATE EXTERNAL TABLE IF NOT EXISTS snappytempjc (*

*Stamp string,*

*host string,*

*sourceId bigint,*

*callAdvTime string,*

*orderId bigint,*

*orderInputId bigint,*

*channelType string,*

*source string,*

*uuid string,*

*fromip string,*

*delMode string,*

*logtype string,*

*repeatTime bigint,*

*clientip string,*

*scid string,*

*responseMsg string,*

*orderSourceId bigint,*

*level string,*

*channelTime string,*

*advId bigint,*

*ideaId bigint,*

*port bigint,*

*threadname string,*

*levelvalue bigint,*

*appid string,*

*proId bigint,*

*status bigint,*

*cid string,*

*delPlat string,*

*landingPageId bigint,*

*clickTime string,*

*scname string,*

*channelMoney float,*

*inputMoney float,*

*notifyChannelUrl string,*

*callAdvUrl string,*

*callChannelUrl string*

*)*

*ROW FORMAT DELIMITED*

*FIELDS TERMINATED BY ' '*

*stored as TEXTFILE;*

*SET hive.exec.compress.output=true;*

*SET mapred.compress.map.output=true;*

*SET mapred.output.compress=true;*

*SET mapred.output.compression=org.apache.hadoop.io.compress.SnappyCodec;*

*SET mapred.output.compression.codec=org.apache.hadoop.io.compress.SnappyCodec;*

*set mapred.output.compression.type=BLOCK;*

*SET io.compression.codecs=org.apache.hadoop.io.compress.SnappyCodec;*

*alter table snappytempjc SET SERDEPROPERTIES('serialization.null.format' = 'null');*

b.写入数据

*LOAD DATA INPATH /online\_history\_backup/2018-03-26/jc-history-channelclick-2018-03-26.log.snappy' OVERWRITE INTO TABLE snappytempjc;*

*select count(\*) from snappytempjc;*

*insert into table snappyjc select \* from snappytempjc;*

其次，创建流格式的外部表，将textfile表中的数据复制到当前表，这样MR程序才可以进行多任务拆分，更加高效、快速的进行处理；

示例：

1. 创建表

*drop table snappyjc;*

*CREATE EXTERNAL TABLE IF NOT EXISTS snappyjc (*

*Stamp string,*

*host string,*

*sourceId bigint,*

*callAdvTime string,*

*orderId bigint,*

*orderInputId bigint,*

*channelType string,*

*source string,*

*uuid string,*

*fromip string,*

*delMode string,*

*logtype string,*

*repeatTime bigint,*

*clientip string,*

*scid string,*

*responseMsg string,*

*orderSourceId bigint,*

*level string,*

*channelTime string,*

*advId bigint,*

*ideaId bigint,*

*port bigint,*

*threadname string,*

*levelvalue bigint,*

*appid string,*

*proId bigint,*

*status bigint,*

*cid string,*

*delPlat string,*

*landingPageId bigint,*

*clickTime string,*

*scname string,*

*channelMoney float,*

*inputMoney float,*

*notifyChannelUrl string,*

*callAdvUrl string,*

*callChannelUrl string*

*)*

*COMMENT 'snappyjc log details'*

*ROW FORMAT DELIMITED*

*FIELDS TERMINATED BY ' '*

*stored as SEQUENCEFILE;*

*alter table snappyjc SET SERDEPROPERTIES('serialization.null.format' = 'null');*

1. 导入数据

*LOAD DATA INPATH /online\_history\_backup/2018-03-26/jc-history-channelclick-2018-03-26.log.snappy' OVERWRITE INTO TABLE snappytempjc;*

*select count(\*) from snappytempjc;*

*insert into table snappyjc select \* from snappytempjc;*

最后，可以通过创建外部ES表的形式，将数据写入ES。

1. 创建表

*drop table esjc1;*

*CREATE EXTERNAL TABLE IF NOT EXISTS esjc1 (*

*Stamp string,*

*host string,*

*sourceId bigint,*

*callAdvTime string,*

*orderId bigint,*

*orderInputId bigint,*

*channelType string,*

*source string,*

*uuid string,*

*fromip string,*

*delMode string,*

*logtype string,*

*repeatTime bigint,*

*clientip string,*

*scid string,*

*responseMsg string,*

*orderSourceId bigint,*

*level string,*

*channelTime string,*

*advId bigint,*

*ideaId bigint,*

*port bigint,*

*threadname string,*

*levelvalue bigint,*

*appid string,*

*proId bigint,*

*status bigint,*

*cid string,*

*delPlat string,*

*landingPageId bigint,*

*clickTime string,*

*scname string,*

*channelMoney float,*

*inputMoney float,*

*notifyChannelUrl string,*

*callAdvUrl string,*

*callChannelUrl string*

*)*

*STORED BY 'org.elasticsearch.hadoop.hive.EsStorageHandler'*

*TBLPROPERTIES('es.resource' = 'jc-es-{channelType}-{logtype}/logs', 'es.nodes'='10.61.8.15','es.read.metadata' = 'true','es.field.read.empty.as.null' ='true');*

1. 写入数据

*insert into table esjc1 select \* from snappyjc;*

3）通过Hive清洗、处理和计算原始数据；

4）HIve清洗处理后的结果，如果是需要经常查询场景的可存入ES；

数据应用从ES查询数据，如果是统计、处理后的最终的结果，可以存入MySQL；



