时序图描述

1.1 receive kafka message[json] spark从爬虫kafka服务端中接受数据作为RDD（数据类型为json格式）
1.2 data handle 开始执行数据处理（需要调用datahandleUtil中的方法）
1.3 parse json(scheduler 对接收到的kafka RDD进行处理，转化为数据类型）
1.4 scheduler的parserAndSave方法（对数据进行解析存储发送）
1.4-1 根据rowkey从hbase中取数据（调用hbaseUtil中的getDataByKey方法）
1.4-2 返回hbaseutil中获得的数据
1.4-3 根据task id执行html信息的解析
1.4-3-1 对html内容进行解析
1.4-3-2 将二级页面的url，taskid 等信息传给爬虫kafka客户端
1.4-3-3 parse程序返回解析的结果（结果类型为item的集合）
1.4-4-1 判断jedis中是否存在此数据
1.4-4-2 将正文的task信息发送给另一kafka队列（罗旭东的kafka队列）由他传递给爬虫端
1.4-4-3 更新jedis数据库中文章标题信息，同时进行类别，版块等信息的分类
1.4-4-3-1 将更新后的数据写入redis中
1.4-5 将更新后的数据存入hbase数据库中