Transwarp Stream大数据流计算技术在智能交通分析性监控系统中的应用

孙元浩1　胡刚2
（1、2 星环信息科技（上海）有限公司，上海市 200030）

摘要：智能交通分析性监控系统需要在不断产生的、海量的、格式多样的过车数据中快速发现问题，最好在问题发生时就发出预警。交管部门传统使用的数据库能力有限，无法满足对实时性的需求。所幸，这个难题恰好是近年来热门的大数据技术的强项。大数据技术以其分布式的计算方式尤其擅长对海量数据的快速处理。大数据发展到现在已经有相对成熟的技术来处理以下三种问题：复杂的批量数据处理、基于历史数据的交互式查询和基于实时数据的流处理。其中的流处理，顾名思义，是在数据产生并流入系统时就进行处理并马上得出结果，非常适合分析型监控中过车数据不断产生的场景和对实时性的需求。本文首先分析了当前分析监控系统的不足，通过对业务逻辑关系的深入分析，采用Transwarp Stream技术实现大数据的实时处理，支撑实时显示和告警机动车违规违章活动热点，以及分析机动车活动轨迹并做预测等应用。

关键词：Transwarp Stream；流处理；大数据；智能交通分析性监控系统
中图分类号：P223.1, U371

Application of Bigdata Transwarp Stream Technology in the Intelligent traffic analysis monitoring System

Sun yuanhao1 HuGang2

（1, 2 Traffic Management Research Institute of the MPS，Jiangsu Wuxi, 214151, China）

Abstract：Intelligent traffic monitoring system needs to find the problem in the continuous generated, massive, variety formats of passing vehicle data, it is the best to alert a warning in the first time when the problem occurs. The traditional database capacity used by the traffic control department is limited, can not meet the real-time demand. Fortunately, this problem can been resolved by the popular big data technology in recent years. Big data technology is especially good at the rapid processing of massive data in the distributed computing enviroment. The big data technology now have relatively mature method to deal with the following three problems: complex batch data processing, historical data interactive queries and real-time stream data processing. stream data processing, as the name suggests, as data flow into the system, it will been processed simultaneously and get results immediately, It's very suitable for Intelligent traffic monitoring system. At first, this paper analyzes the deficiency of current system, through in-depth analysis of business logic, we use Transwarp Stream technology to realize real-time processing of large data and real-time display and alarm of vehicle violation activities for hot support, as well as the analysis of vehicle trajectory and prediction application.

Key words：Transwarp Stream; Stream data processing; Bigdata; Intelligent traffic analysis monitoring System

引言

道路交通作为“衣食住行”中的“行”和我们的日常生活息息相关。我们对道路交通的需求无非是安全和快捷。但作为全球第二大经济体，全国的机动车数量高速增长，同时随着公路建设的发展，机动车的流动性也大大增强，在机动车越来越多的今天，交通事故屡见不鲜，交通堵塞更是家常便饭，我们的需求显得有些奢侈。要改变道路交通的现状，我们不仅需要政府的宏观举措—增加道路建设、加强交通法规教育、发展公共交通等，更需要交管部门落到细节的管理。我国的机动车保有量极大（一个省的机动车数量在千万级别），道路交通还具有不受统一调度、行车轨迹复杂等特点，管理难度很高。为了提高管理能力，各地的交管部门纷纷部署了统一的智能交通监控系统，通过电子眼、传感器、测速器等设备对交通情况进行全天候的监控。

1、智能交通分析性监控系统概况

1.1 概念

智能交通道路监控可以分为两类—观察型监控和分析型监控。我国交管部门的观察型监控的使用已经相当成熟，对违章行为的捕捉率非常高，有效地降低了违章率。分析型监控就要复杂许多，常见的任务有套牌车分析、伴随车分析、碰撞车分析、黑名单车辆预警、旅行时间计算、道路流量统计等等。这些任务需要交通卡口不间断地记录所有经过车辆的过车数据，并且对这些数据进行查找、关联、比对等处理。因为记录条数多并且包含图像信息，过车数据的体量非常庞大，对监控系统的存储、查询和计算能力都提出了很高的要求。事实上，由于数据量过大，大多数交管部门采用离线分析进行分析型监控，也就是将一个周期内（比如一天内）全部的过车数据都存储起来后再对整个数据集进行计算。这种处理方式显然延时过高，监控系统在特殊状况发生很久以后才能将结果报告给交警。分析型监控的任务常常具有时效性，比如黑名单车辆通过某个卡口时，需要系统立刻捕捉到这一行为并通知卡口附近的交警前往拦截；再比如道路流量统计的目的是通知交警在某地交通流量过大时前去疏导。离线分析的高延时使得交警无法对这些状况进行及时响应。

1.2 发展概况

为进一步推进公路交通安全管理科技建设，提升动态化、信息化条件下的公路交通安全管控水平，2013年公安部交管局在全国组织推广了智能交通分析性监控系统联网并加大建设力度，智能交通分析性监控系统建设取得了很大进展。但作为全球第二大经济体，全国的机动车数量高速增长，同时随着公路建设的发展，机动车的流动性也大大增强，造成当前各省市部署的分析监控系统积聚了海量的过车数据等信息。传统的分析监控系统针对海量过车信息，主要通过预设的条件，例如按月或按周的形式形成报表。随着系统的逐步推广和深入应用，业务需求开始向多样化方向发展，原有的模式存在以下弊端：1）只能对实时数据进行预先设定的分析，不能进行ad-hoc分析；2）实时视图和历史视图保持一致比较困难；3）服务数据库响应速度慢，不能满足快速分析。传统技术由于无法及时准确的提供有效可靠的机动车违规违章信息给公安部门，刑侦抓捕工作由此变的异常艰难。因而，迫切的需要一种新型、快速的分析工具，能够实时的告警机动车分析监控的相关信息。

1.3 传统智能交通分析监控系统存在的不足

由于数据量过大，大多数交管部门采用离线分析进行分析型监控，也就是将一个周期内（比如一天内）全部的过车数据都存储起来后再对整个数据集进行计算。这种处理方式显然延时过高，监控系统在特殊状况发生很久以后才能将结果报告给交警。分析型监控的任务常常具有时效性，比如黑名单车辆通过某个卡口时，需要系统立刻捕捉到这一行为并通知卡口附近的交警前往拦截；再比如道路流量统计的目的是通知交警在某地交通流量过大时前去疏导。离线分析的高延时使得交警无法对这些状况进行及时响应。

分析型监控的技术难点在于监控系统需要在不断产生的、海量的、格式多样的过车数据中快速发现问题，最好可以在问题发生时就发出预警。交管部门传统使用的数据库能力有限，无法满足对实时性的需求。

3、Transwarp Stream大数据流计算技术在智能交通分析性监控系统中的应用

基于Transwarp Stream流技术的机动车分析监控方法及系统，将机动车位置事件数据和Transwarp Stream流处理技术结合，先从海量数据中筛选机动车违规违章等出现的关键信息，然后基于机动车、违规违章等事件、时间、空间进行多维分析，实现能像卫星气象云图一样准实时的显示和告警机动车违规违章活动热点、以及分析机动车违规违章活动的轨迹趋势，预测下一个犯罪活动区域，成为机动车分析监控的终极利器。整个系统处理过程中没有大量数据积攒的延迟，系统框架同时也是专门针对延迟做了优化，保证整个数据处理链条在极低的延迟内完成。从而保证对某省所属所有机动车进行实时监控。

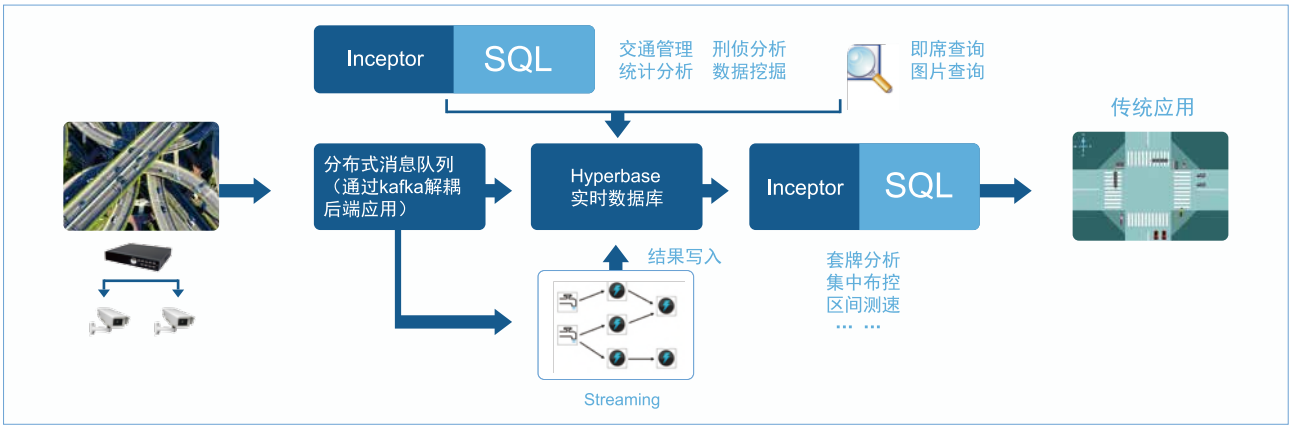

图1：软件逻辑架构

Fig 1 Transwarp Stream technology Architecture

3.1 Transwarp Stream大数据流计算技术实现原理

传统分布式计算一般首先拿到一个长时间积累后的智能交通数据，再进行数据拆分和聚合。流处理则主要通过事件机制，对动态产生的智能交通数据进行实时计算并及时反馈结果，类似流管道一样，立即处理消息并响应。流处理具有低延迟、高性能、分布式、可扩展、高容错等特点。目前主要的流计算技术包括：storm，Transwarp Stream等。

Transwarp Stream是一种构建在Spark上的实时计算框架，它扩展了Spark处理大规模流式数据的能力。Transwarp Stream将流式计算分解成一系列短小的批处理作业，经过操作变成中间结果保存在内存中。整个流式计算根据业务的需求可以对中间的结果进行叠加，或者存储到外部设备。

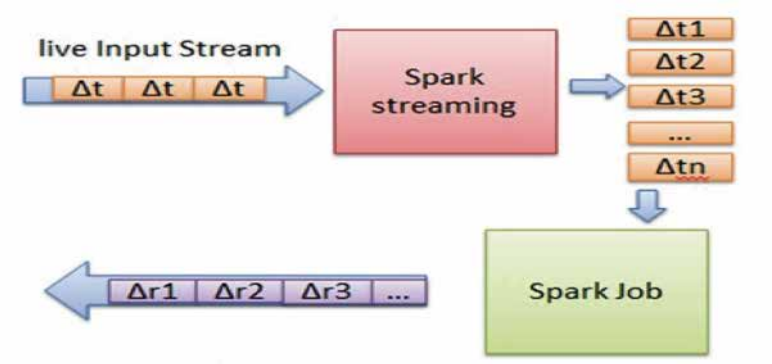

图2 Transwarp Stream大数据流计算技术
Fig 2 Transwarp Stream Bigdata stream data processing technology

Transwarp Stream把卡口过车数据按照时间切片Δt（200ms）为单位切分，在流处理平台中为后续的计算将数据切分为每200ms的一个数据块。对于Vehicle_Info Topic来说，每个时间切片（如200ms）中的数据为该时间段内完整的卡口过车数据，Topic的信息结构如下表所示。

序号	Topic名称	信息表结构
1	Vehicle_Info	{开始时间，结束时间，卡口编号、方向类型、车道号、号牌号码、号牌种类、过车时间、车辆速度、车辆限速、违法代码、车外廓长、号牌颜色、车辆类型、辅助号牌种类、辅助号牌号码、辅助号牌颜色、车辆品牌、车辆外形、车身颜色、通行图片路径、通行图片1、通行图片2、通行图片3、特征图片；}

表1 Topic信息结构表
Table 1 Topic information structure

将每一个时间切片的卡口过车数据进行实时的批处理，本质上本流处理平台是基于小作业的高速低延时批处理分析，优势在于批处理在状态维护、不丢不重的精准完整性语意完成上更加容易和自然。因为一致性状态的维护、完全不丢失不重复需要的元信息维护代价都非常大，传统的流处理系统因为面向单条数据，在出现错误恢复时无力完成完全的精准恢复，从而造成数据或者状态的丢失。业务应用逻辑以DAG（有向无环图）形式的服务常驻在集群内存中，生产系统的消息通过实时消息队列进入计算集群，在集群内以Pipeline方式被依次处理。

卡口过车数据经过流处理，根据获得的异常车牌地理位置信息，获得异常车牌和驾驶证的活动范围以及移动趋势，实时输出结果数据，触发告警，为进一步的卡口过车下一路口拦截以及分析监控应对提供可靠实时保障。

这样，流处理系统通过在软件层面通过冗余、重放、借助外部存储等方式实现容错，可以避免数台服务器故障、网络突发阻塞等问题造成的数据丢失的问题。通过定义弹性数据集RDD来实现容错。RDD是一种数据结构的抽象，它封装了计算和数据依赖，数据可以依赖于外部数据或者其他RDD, RDD本身不拥有数据集，它只记录数据衍变关系的谱系，通过这种谱系实现数据的复杂计算变换，在发生错误后通过追溯谱系重新计算完成容错，如果计算的衍变谱系比较复杂，系统支持checkpoint来避免高代价的重计算发生。

3.2 大数据流计算技术对比

目前主要的流计算技术包括：storm，Transwarp Stream等。

Storm是Twitter支持开发的一款分布式的、开源的、实时的、高容错大数据流式计算系统。Storm集群主要由一个主节点和一群工作节点（worker node）组成，通过 Zookeeper进行协调。

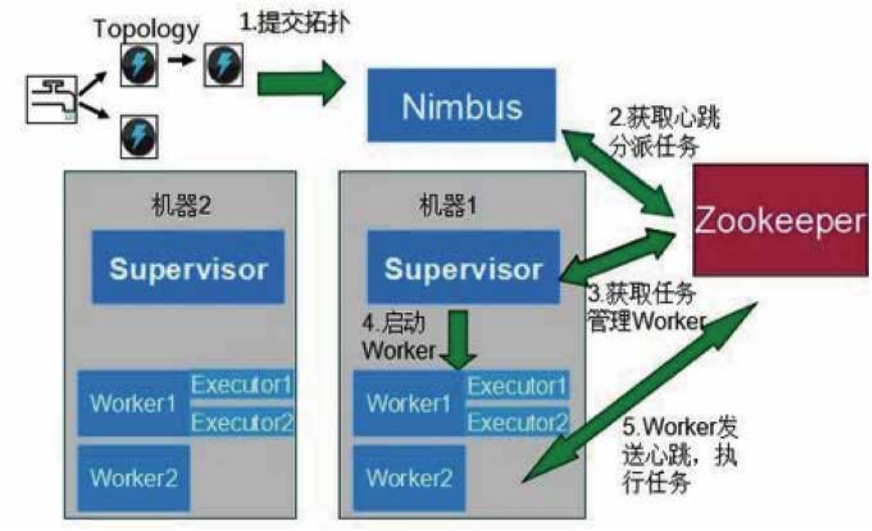

图3 Storm大数据流计算技术
Fig 3 Storm Bigdata stream data processing technology

下面给出storm，Transwarp Stream的功能，性能等的对比，基于下述对比，我们选择了Transwarp Stream流处理技术。

	Transwarp Stream	Storm
原理	基于时间切片(如200ms)的低延迟批处理	面向单条数据的事件驱动处理
容错性	Transwarp Stream的小批量处理系统的优势在于批处理在状态维护、不丢不重的精准完整性语意完成上更加容易和自然，同时借助新的计算容错思想，通过弹性数据集RDD (Resilient Distributed Dataset)实现容错。	一致性状态的维护、完全不丢失不重复需要的元信息维护代价都非常大，由于Storm面向单条数据，在出现错误恢复时无力完成完全的精准恢复，从而造成数据或者状态的丢失。Storm系统为了完成精准恢复和数据不丢失不重复开发了Trident组件，借助与外部存储实现，结果增加了额外的外部存储系统，Trident本身性能也十分低下。
吞吐量	基于低延时的小批量处理系统，吞吐性能大大高于面向单条数据的事件驱动处理系统	
复杂业务实现	基于批处理对于复杂业务场景的天然支持，支持实时异常检测，实时复杂的清洗转换，实时时间窗口统计，甚至实时在线模型训练	基于单条数据的事件驱动，比较容易支持实时规则比对等简单业务逻辑，较难实现实时统计分析等业务逻辑
处理延时	延时在1s~2s之间	简单业务场景下延时较低，在毫秒级延时

图4 Storm和Transwarp Stream大数据流计算技术对比
Fig 1 Transwarp Stream technology Architecture

3.2 流处理分析性监控系统采用的相关创新技术

3.2.1 分布式消息队列技术

分布式消息队列是基于Zookeeper协调管理的。卡口过车数据定制机动车的数据Topic进行数据发送。将完整的卡口过车数据发送至分布式消息队列。流处理平台根据卡口过车产生数据量速率，在分布式消息队列中将卡口过车数据近乎均匀的分散到各个服务器中多个Partition中。流处理引擎Transwarp Stream在分布式集群中开启多个并发数据流消费线程，组成针对于不同业务规则的多个消费组Consumer Group。在每个Consumer Group中，Partition的个数是数据流消费总线程数的倍数，每个计算线程消费相同个数卡口过车数据的Partition，以达到集群负载均衡的目的。

3.2.2 Transwarp Stream流处理技术

Transwarp Stream是建立在Spark（Berkeley的交互式实时计算系统）上的实时计算框架，Transwarp Stream的优势在于：能运行在100+的结点上，并达到秒级延迟。Transwarp Stream的基本原理是将输入数据流以时间片（秒级）为单位进行拆分，然后以类似批处理的方式处理每个时间片数据。 Transwarp Stream是将流式计算分解成一系列短小的批处理作业。这里的批处理引擎是Spark，也就是把Transwarp Stream的输入数据按照batch size（如1秒）分成多段数据（Discret-

ized Stream），每一段数据都转换成Spark中的RDD（Resilient Distributed Dataset），然后将Transwarp Stream中对DStream的Transformation操作变为针对Spark中对RDD的Transformation操作，将RDD经过操作变成中间结果保存在内存中。整个流式计算根据业务的需求可以对中间的结果进行叠加，或者存储到外部设备。

Transwarp Stream的优势主要表现在容错性、实时性、可扩展性。

•容错性表现在两个方面：一是使用HDFS作为文件系统。HDFS的备份机制保证了数据不易丢失；二是将采集的数据保存　到2个节点上，防止数据在源头丢失。

•实时性主要涉及流式处理框架的应用场景。Transwarp Stream将流式计算分解成多个Spark Job，对于每一段数据的处理都会经过Spark DAG图分解，以及Spark的任务集的调度过程。对于目前版本的Transwarp Stream而言，其最小的Batch Size的选取在0.5~2秒钟之间（Storm目前最小的延迟是100ms左右），所以Transwarp Stream能够满足除对实时性要求非常高（如高频实时交易）之外的所有流式准实时计算场景。

•扩展性与吞吐量表现在Spark的节点数量上。Spark目前在EC2上已能够线性扩展到100个节点（每个节点4Core），可以以数秒的延迟处理6GB/s的数据量（60M records/s），其吞吐量也比流行的Storm高2～5倍。在Berkeley利用Grep所做的测试中，Transwarp Stream中的每个节点的吞吐量是670k records/s，而Storm是115k records/s。

3.3 流处理分析性监控系统的优越性

与传统采用关系型数据库采集智能交通数据的方式进行分析性监控系统相比，采用Transwarp Stream大数据流计算技术的分析性监控系统具有以下几大优点：

一是数据查询效率高很多。在中国某大省使用扩充版Transwarp Stream大数据流计算技术的系统后，针对过车信息的精确查询、模糊查询均能达到秒级响应。

二是解决了支队数据上传严重积压的问题。在核心版系统中，支队和总队间的数据传输使用的Java传输通道，由于各地系统运维和系统可承受负载的问题，导致各地系统中普通积压量大量的待上传数据。使用扩充版系统后，各地过车信息能够在毫秒级实时上传总队扩充版系统。

三是能够监测到更多的路面车辆通行状况。在核心版中，仅能监测各地自有车辆的逾期未检验、逾期未报废等情况，使用扩充版系统后，能够监测全省的逾期未检验、逾期未报废、强制注销、车主驾照满分、暂扣等车辆、全国重点车辆在本省的通行情况等，为各业务部门的工作提供详细的数据支撑。

4、流处理分析性监控系统应用情况

在试运行期间，基于Transwarp Stream大数据流计算技术的扩充版系统合计分析出嫌疑车10.46万辆，发出预警85.97万次。

其中：

1、逾期未年检：发现车辆68,167辆，预警728,918次

2、逾期未报废：发现车辆2,396辆，预警48,776次

3、特殊省份车辆：发现车辆759辆，预警5,144次

4、凌晨2点至5点上路行驶的客运车辆：公路客运121辆、旅游客运45辆，合计发出预警信息475次

5、车主驾驶证无效车辆：其中驾驶证吊销注销撤销19辆，驾驶证暂扣33辆，驾驶证满分38辆，合计预警409次

6、在途行驶的已注销、强制注销车辆：其中注销车辆2442辆，强制注销车辆31672辆，合计预警77,652次

由此可见，基于Transwarp Stream大数据流计算技术在分析性监控系统中的应用已极为普及，基本替代了传统采用关

系型数据库等方式进行分析型监控的智能交通系统，已成为智能交通系统发展的主流。

5 结束语

交通拥堵和安全问题越发严重的今天，全国各省都在计划部署省级的智能交通监控系统。河北、山西、重庆、湖南、浙江、甘肃等地区已完成相关项目，四川，安徽、辽宁、吉林、山西、江西、山东、湖北、青海等省市地区也已开始建设工作。在省级智能交通监控实现的同时，智能交通监控的全国联网是大势所趋，届时，系统所面临的数据处理任务将更加艰巨。基于Hadoop的大数据流处理平台扩展性极强，存储和计算能力都可以无限提升，在全国智能交通监控系统中会发挥更大的威力。充分运用大数据流处理技术，使道路建设、法规制定和事件处理配合无间，让交通管理变得更加“智慧”，道路交通时时处处安全、快捷的实现也就指日可待。

参考文献:

[1] 徐晓东, 孔晨晨, 席正祺. 大数据云计算技术在全国机动车分析缉查布控系统中的应用[J]. 中国公共安全(学术版) , 2015, 38(1)：87-91.

[2] Spark, lightning-fast cluster computing. 2013. http://spark-project.org/

作者简介： 第1作者：孙元浩（1977-），男，江苏省无锡市人， 研究员， 硕士.

第2作者：胡刚（1976-），男，湖北省黄冈人， 副研究员， 学士.

1.联系人：秦东炜； 2.通讯地址：江苏省无锡市钱荣路88号（214151）； 3.电子信箱：qin_dw@aliyun.com；电话：13665180747，0510-85505161。