

需求分析

项目名称：出租车车辆 GPS 定位挖掘

项目类别：☐ 电子商务

☐ 移动终端应用

☒ 大数据分析

☐ 物联网应用

☐ 人机交互应用

☐ 其他()

命题企业：北京瑞德云网科技有限公司

咨询邮箱：zhangguangjun@ict.ac.cn

2017 年 12 月 1 日

项目需求分析

一、引言

1.1 项目背景

随着国民经济的发展，人民的生活水平有了很大的提高，交通工具也变得多样化，大小车辆变得越来越多。道路变得拥堵成为当前社会紧要解决的问题，铺设道路对于开辟新城市或许是有用的，但对于当前已有的道路也是无能为力。所以驱使科技的发展来解决这一难题，此时出现的车辆 GPS 位置信息应用可以缓解这一情况。通过实时获取车辆的位置信息，计算当前道路的通达性，对其他近时间路过这里的车辆友好的给予提示，这样方便了出行的用户，也相对缓解交通压力。不仅仅是对出租车，同时对私家车辆也是可以起到一样的作用。所以相关车辆位置信息的应用在当前交通情况下是非常重要的，也是迫不及待的。

1.2 项目目的

获取实时的出租车的位置定位信息，收集位置信息，经过消息队列的接收 Storm 的处理写入分布式 HDFS 文件系统中，可以使用数据分析工具，实时的显示车辆的位置信息，对照位置信息能够对道路的通行情况实时的掌握。本实验案例采用的一段时间出租车的位置信

息，通过 Kafka 流经 Storm 最后写入 HDFS 文件系统中，提供给数据分析使用。能帮学生快速构建一个流式处理的架构和流程。完善各类大数据应用场景的需要。

二、项目需求

2.1 功能需求

2.1.1 数据录入

本项目采用的是已经收集好的出租车位置信息数据文本，对整个大数据处理流程做分析的项目，数据已经准备好。上传到大数据集群的 HDFS 文件系统中即可完成数据的录入。

2.1.2 数据处理

对现有的出租车位置信息数据进行清洗处理，最终处理成结构化的数据储存到 HDFS 文件系统中，提供数据分析提取分析。

2.1.3 数据分析

对清洗过后的出租车位置信息数据进行分析，根据现有的数据字段属性情况，可以对出租车位置信息数据在地图展现预测道路通达性。

2.1.4 数据展现

使用地图显示车辆的位置信息。

2.2 性能需求

2.2.1 可扩展性

大数据集群可以快速无缝的横向扩展，数据达到一定的规模之后不需要担心数据承载出现问题，加存储或者服务器即可完成集群规模的横向扩展。数据会根据集群的情况合理调整调度，尽可能的合理使用资源。

2.2.2 稳定性

采用的是大数据领域主流的组件进行开发，每个技术都是非常领先。大数据处理平台采用的是高可用，数据副本机制保障数据的安全。并且在行业已经运用到各行各业中，数据的规模也是在 PB 级别，社区活跃度非常高，在运维和开发成本上相对较低，可持续性较强。

2.3 任务要求

2.3.1 大数据平台搭建

能通过安装文档个人完整的把 CRH 技术平台搭建完成，并且正常运行，为后续的案例开发实验做基础。

2.3.2 数据接入

独立把数据上传至 HDFS 分布式文件系统，提供给 Dataiku 数据分析工具进行数据分析使用。使用 Kafka 把文本数据加载到消息队列，调用 storm 程序对数据进行处理，最终写入 HDFS 文件系统。

2.3.3 数据分析

使用 Dataiku 直接读取 HDFS 文件系统中的数据，抽取数据，提取经纬度数据在地图上进行展现。

三、运行环境

3.1 软件环境

服务器操作系统：RedHat 或者 CentOS（英文版）

服务器操作系统版本：RedHat7 或者 CentOS7

JDK 版本：Oracle1.8

CRH 版本：CRH5.1

分析工具：Dataiku

3.2 硬件环境

推荐测试环境：

内存：8G

存储：100G

CPU：双核处理器

推荐生产环境：

内存：128G 或者 256G

存储：服务器满配每块盘 3T 或者 4T

CPU：48 核

3.2 网络环境

每台服务器或者操作系统之间能相互连通，有时间同步服务器，终端能连接上即可。

四、实现过程

4.1 实现思路

- 数据接入从文本中读取
- 读取进入 Kafka
- 调用 Kafka 处理数据
- 写入 HDFS 文件系统
- 数据分析工具抽取数据
- 对结果进行画图展现

4.2 实现技术

4.2.1 HDFS

大数据分布式文件存储，实现数据的存储保证数据的安全，HDFS 文件系统的容量可以横向的扩展。

4.2.2 Kafka

Kafka 是一种高吞吐量的分布式发布订阅消息系统，它可以处理消费者规模的网站中的所有动作流数据。这种动作（网页浏览，搜索和其他用户的行动）是在现代网络上的许多社会功能的一个关键因素。这些数据通常是由于吞吐量的要求而通过处理日志和日志聚合来解决。对于像 Hadoop 的一样的日志数据和离线分析系统，但又要求实时处理的限制，这是一个可行的解决方案。Kafka 的目的是通过 Hadoop 的并行加载机制来统一线上和离线的消息处理，也是为了通过集群来提供实时的消费。

4.2.3 Storm

Apache Storm 是一个分布式实时大数据处理系统。Storm 设计用于在容错和水平可扩展方法中处理大量数据。它是一个流数据框架，具有最高的摄取率。虽然 Storm 是无状态的，它通过 Apache ZooKeeper 管理分布式环境和集群状态。它很简单，您可以并行地对实时数据执行各种操作。

Apache Storm 继续成为实时数据分析的领导者。Storm 易于设置和操作，并且它保证每个消息将通过拓扑至少处理一次。

4.2.4 DATAIKU

企业级客户提供基于云技术的大数据服务分析平台，数据分析工程师可以很简单的完成数据的收集，分析，展现。

4.3 实现计划

4.3.1 CRH 环境搭建

使用 CRH5.1 搭建大数据基础平台，搭建过程详见 CRH 安装手册

4.3.2 数据录入

把本地测试数据通过 Kafka 读取进消息队列中。

4.3.3 数据处理

调用 Storm 程序从 Kafka 拉取数据，然后对数据进行处理，写入到 HDFS 文件系统中。

4.3.4 数据分析

使用 Dataiku 连接 CRH 大数据平台 HDFS 文件系统中，抽取数据，对出租车位置信息数据在地图上显示。