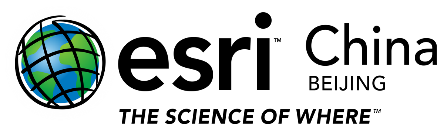
GeoAnalytics Server

成都市出租车大数据分析



易智瑞中国信息技术有限公司

2017 年 4 月

**——制定及修订记录——**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **版本** | **完成日期** | **编写/修订纪要** | **编写者** | **备注** |
| **V1.0** | **2017-4-27** | **初稿** | **胡源** |  |
| **V1.1** | **2017-4-28** | **修订** | **胡源** | **内容完善** |
|  |  |  |  |  |
|  |  |  |  |  |

目录

[GeoAnalytics Server 1](#_Toc481672922)

[成都市出租车大数据分析 1](#_Toc481672923)

[1. 本文简介 4](#_Toc481672924)

[2. 出租车数据介绍 4](#_Toc481672925)

[3. 分析环境介绍 4](#_Toc481672926)

[4. 数据准备过程 4](#_Toc481672927)

[5. 执行分析 6](#_Toc481672928)

[6. 性能分析 7](#_Toc481672929)

1. **本文简介**

本文介绍了如何结合HDFS和Hive将txt格式的出租车大数据注册到GA中的大数据文件共享，并在GA中执行分析，最后对性能进行了评估。

***注****：HDFS和Hive的搭建请参阅《Esri大数据平台搭建指南之HDFS》和《Esri大数据平台搭建指南之Hive》*

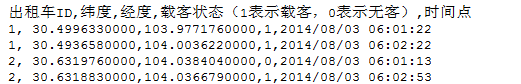
1. **出租车数据介绍**

**数据简介**：成都市1.4万多辆出租车的超过12亿条GPS记录，时间从2014年08月03日到08月30日（不包含7号、13号、17号）

**数据后缀**：txt

**下载链接**：http://pan.baidu.com/s/1cozJ90 密码：fjzn

**数据格式**：



1. **分析环境介绍**

Enterprise环境：Enterprise基础环境 + 4个GA节点

GA节点配置：8个vCPU（1.9GHz） 16G内存

HDFS环境：一个名字节点 + 三个存储节点

Hive环境：搭建在HDFS环境名字节点所在机器

1. **数据准备过程**

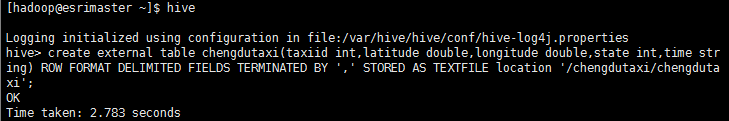
4.1将解压后的数据上传到HDFS（请删除掉“成都市出租车数据下”解压后的predPaths\_test.txt文件再进行上传）



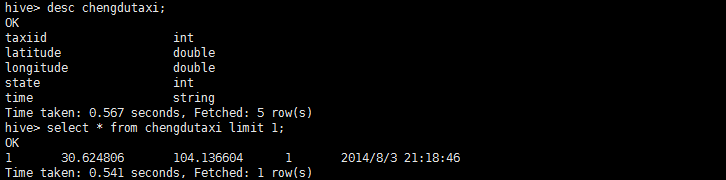
***提示****：上传过程可能有点儿漫长*

4.2在Hive中创建外部表，来映射到上一步存储到HDFS的数据

***注****：指定外部表的location会将location文件夹中的所有文件数据全部映射到表中*



4.3测试Hive表是否正确创建



4.4注册大数据文件共享或者更新大数据文件共享

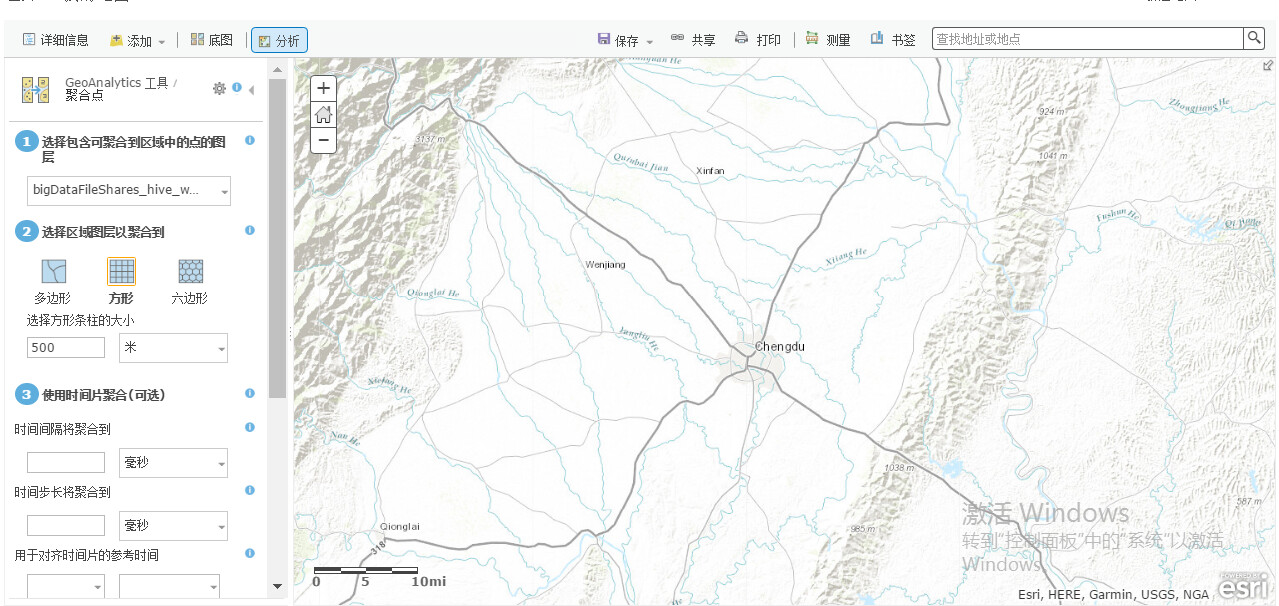


4.5查看注册数据的信息

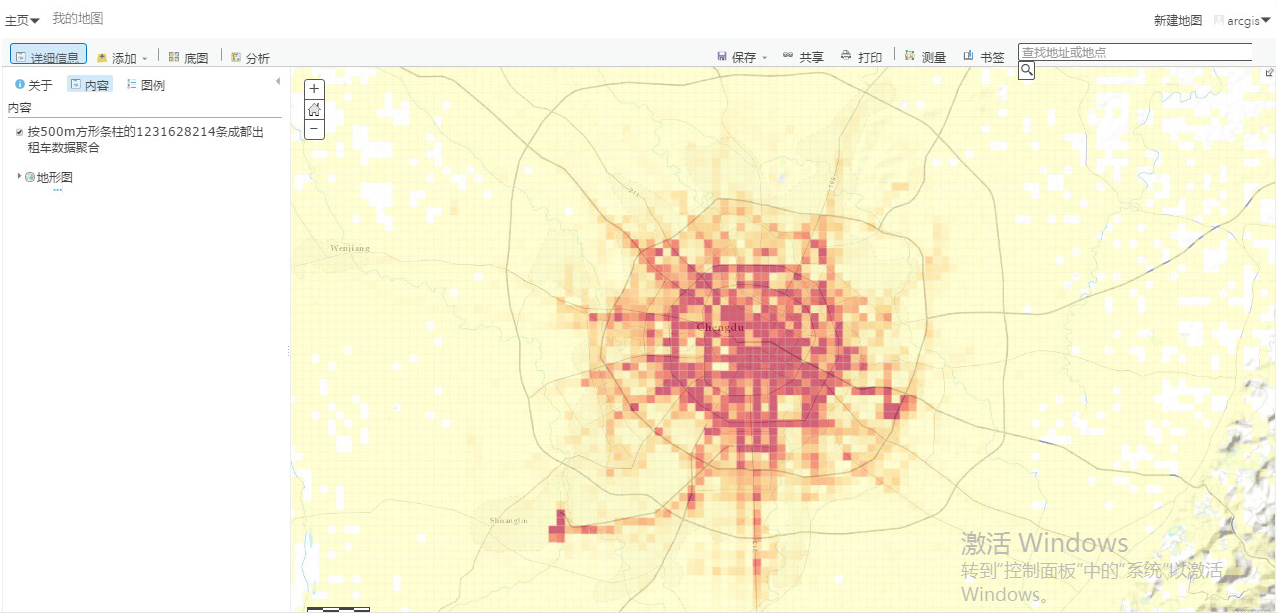


1. **执行分析**

将分析环境中空间参考设置为102100，点聚合工具设置参数如下：



分析后可视化结果：



1. **性能分析**

55G数据使用四台8个虚拟核、16G内存的 GA节点执行点聚合工具分析，50分钟分析完成。由于Spark是基于内存的计算框架，当内存不足时会把数据存入磁盘中或根据最近最少使用页面置换算法（Least Recently Used，LRU）将内存中使用频率较低的文件空间收回，该过程会产生大量的磁盘读写时间。提高CPU性能和内存容量可明显提高计算性能。