doi:10.6041/j.issn.1000-1298.2019.01.019

基于 Hadoop 的气象大数据分析 GIS 平台设计与试验

李　涛[1]　冯仲科[1]　孙素芬[2]　程文生[1]

(1. 北京林业大学精准林业北京市重点实验室, 北京 100083; 2. 北京市农林科学院农业科技信息研究所, 北京 100097)

摘要: 针对海量气象数据在传统 WebGIS 平台下存储和分析计算受到限制的问题,提出基于 Hadoop 的分布式计算和存储框架,使用了 Hadoop 生态体系中的 HDFS 分布式文件存储框架来存储管理分析海量气象数据。在海量数据的并行计算分析方面,使用 MapReduce 作为分布式计算编程模型,该模型通过分析海量气候数据可对农业生产进行决策。最后,利用地理信息系统空间可视化技术,在前端页面以三维形式对分析结果进行展示,并分析比较数据量和集群中节点数对计算耗时的影响。试验结果表明,使用分布式多节点集群可以有效提高海量气象数据的存储和计算效率,解决了传统 WebGIS 平台数据存储与计算的局限性问题。

关键词: 气象数据; 分布式; Hadoop; MapReduce

中图分类号: S157.1; TP79　　**文献标识码**: A　　**文章编号**: 1000-1298(2019)01-0180-09

Design and Test of GIS Platform for Meteorological Data Analysis Based on Hadoop

LI Tao[1]　FENG Zhongke[1]　SUN Sufen[2]　CHENG Wensheng[1]

(1. *Beijing Key Laboratory of Precision Forestry*, *Beijing Forestry University*, *Beijing* 100083, *China*
2. *Institute of Agricultural Science and Technology Information*, *Beijing Academy of Agricultural and Forestry Sciences*, *Beijing* 100097, *China*)

Abstract: Massive meteorological data is limited in storage and analysis on the traditional WebGIS platform. A distributed computing and storage framework based on Hadoop to manage and analyze a large number of meteorological data was proposed. The HDFS distributed file storage framework was used in Hadoop ecosystem to store and manage massive meteorological data. In the aspect of parallel computing and analysis of massive data, MapReduce was used as the basis of distributed computing programming model. This model can make decision for agricultural production by analyzing massive climatic data. The application of regional large data decision analysis suitable for crop growth and the analysis of large data for meteorological disaster assessment were tried out. It had great application value for the research of climate change information extraction and analysis in agricultural production decision-making and other fields. Finally, the front-end pages displayed the analysis results in three-dimensional form by using the geographic information system spatial visualization technology, which made the analysis results more intuitive, and easier to analyze and decision-making, and then the impact of size of data and the number of nodes in the cluster on computing time-consuming was analyzed and compared, and the configuration was tuned the most efficient. Experiment results showed that using distributed multi-node cluster can effectively improve the storage and calculation efficiency of massive meteorological data, and solve the limitations of traditional WebGIS platform.

Key words: meteorological data; distributed; Hadoop; MapReduce

0　引言

海量的气象数据可以通过物联网天气传感器设备或者网络爬虫收集,这些数据生成源以连续的方式生成大量数据[1-3],这种气候数据大多是传统数据处理工具和技术无法处理的结构化、半结构化和

收稿日期: 2018-07-31　修回日期: 2018-10-06

基金项目: 国家自然科学基金项目(U1710123)、北京市自然科学基金项目(6161001)和北京林业大学青年教师科学研究中长期项目(2015ZCQ-LX-01)

作者简介: 李涛(1992—),男,博士生,主要从事林业 3S 技术集成与开发研究,E-mail: 472542625@qq.com

通信作者: 冯仲科(1962—),男,教授,博士生导师,主要从事森林计量学和精准林业研究,E-mail: fengzhongke@126.com

非结构化数据[4-6]。传统数据挖掘算法和统计方法难以存储并处理这类数据[7-9]，气候数据需要一个可扩展的分布式框架来存储和处理，并在季节性气候中获得更有意义的变化信息[10-11]。虽然国内外有许多气候气象 WebGIS 数据管理分析系统，但由于气象站和计量中心在不间断地产生新的实时数据，这些数据在传统 WebGIS 平台中是无法进行存储与计算的[12-15]。因此，需要可扩展的分布式地理空间 WebGIS 系统来分析和利用气象数据[16]。本文结合 Hadoop 的分布式计算和存储技术、地理信息系统相关技术、数据库技术，以实际需要的设计要求，实现气象地理信息的采集、筛选、储存、分析、显示应用等功能，采用浏览器端进行数据的展示与分析。

1　平台设计

1.1　气象大数据获取与分析方法

气象地理信息主要包括气象属性信息以及对应的地理空间信息。随着互联网技术的飞速发展，如今可以利用网络爬虫技术抓取相关的网络平台数据，通过筛选所需要的数据并转换为云计算可用的数据结构，便可以积累海量的可进行分布式计算的气象地理信息数据。

通过网络爬虫或者物联网天气传感器设备获取到的海量气象地理信息数据大多是非结构化的文本格式数据，也可以通过其他方式获取可以用于气象分析与展示的栅格影像。这些数据一般是 TB 级以上的数据量，由于硬件资源限制，在单机环境下是无法进行处理或存储的[17-21]。为了解决海量气象数据的存储问题，通常情况下是将数据分配到多个操作系统管理磁盘中，但是该种方式不便于工作人员的管理和维护，因此迫切需要一种能够同时管理多台机器上文件的分布式文件管理系统[22-25]。

分布式文件管理系统种类很多，但是所有的系统都是基于一次写入、多次查询的情况，不支持并发与写入情况，本文采用 Hadoop 体系下开源的分布式文件管理系统 HDFS，其采用主从架构来管理文件，即由一个名称节点和多个数据节点组成了一个分布式文件系统（Hadoop distributed file system，HDFS）集群。名称节点的作用为：负责客户端请求和响应；元数据的管理，包括查询和修改。数据节点的作用为：存储管理用户文件块数据；定期向名称节点汇报自身所持有的块信息，即通过心跳信息上报自身情况。

在海量气象数据的分析计算方面，传统单节点 WebGIS 系统通过扩展到集群来分布式运行，将极大地增加程序的复杂度和开发难度，因此本文引入一个分布式运算程序的 MapReduce 编程框架，其核心功能是将用户编写的业务逻辑代码和自带默认组件整合成一个完整的分布式运算程序，并发运行在服务器集群上。开发人员可以将绝大部分工作集中到业务逻辑开发上，而将分布式计算中的复杂度交由框架来处理[26-27]。

1.2　系统平台架构设计

本系统基于 Hadoop 生态体系进行搭建，包括数据获取层、云计算层、云存储层和前端显示层。数据获取层的数据来源于网络爬虫爬取的气象数据[28]或者物联网天气传感器设备中采集的数据。云存储层、云计算层分别是 Hadoop 的分布式存储框架 HDFS 和分布式计算框架 MapReduce，主要功能是将当前的海量空间数据进行统一格式化处理，并将其存入到分布式文件系统中，通过并行处理框架可以对包含空间属性的气象数据进行大数据量的快速分布式计算得到分析结果。前端显示层则是利用 Cesium 进行三维可视化展现。整个平台结构如图 1 所示。

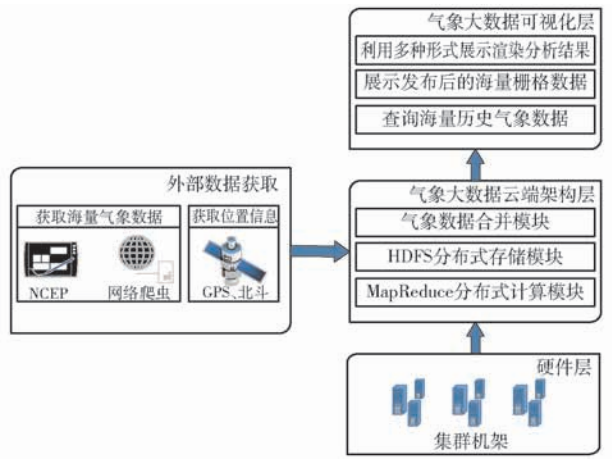

图 1　平台结构

Fig. 1　Platform structure

1.2.1　数据获取模块

从物联网天气传感器设备获取数据，或利用网络爬虫抓取气象相关网页来获取相关数据。利用爬虫 WebDriver 和 PhantomJs 技术相应接口对得到的网页内容进行解析，进而获取需要的气象地理信息。由于得到的信息很多杂乱无章，为了获取真实需要的信息数据，在获取元素信息后，还需要利用正则表达式对这些属性信息进行筛选判别，同时进行格式统一处理，最后将输出后的数据以 GeoJson 或者其他文本格式合并保存。

1.2.2　云存储模块

由于 HDFS 平台不适合管理小文件，所以首先

对采集到的大量小文件进行合并。

小文件合并有以下 3 种方式:①在采集数据时,将小文件合并为大文件再上传至 HDFS。②在业务处理之前,使用 HDFS 上的 MapReduce 程序对小文件进行合并。③在利用 MapReduce 处理小文件时,采用 conbineInputFormat 提高效率。

1.2.3 云计算分析模块

由 MapReduce 的工作流程可知,Hadoop 下的空间数据并行操作共需要 6 个步骤:

(1) MapReduce 程序启动时,最先启动的是 MRAppMaster(MapReduce 程序启动节点),MRAppMaster 启动后根据本次作业的描述信息,计算出需要的 Map 任务实例数量,然后向集群申请启动相应数量的 Map 任务进程。

(2) 利用客户指定的输入格式来获取 RecordReader 并读取数据,形成输入键值对。

(3)将输入键值对传递给客户定义的 Map 方法,做逻辑运算,并将 Map 方法输出的键值对收集到缓存中。

(4)将缓存中的键值对按照键值分区排序后不断溢写到磁盘文件中。

(5) MRAppMaster 监控到所有 Map 任务完成后,根据客户指定的参数启动相应数量的 Reduce 任务进程,并告知 Reduce 任务进程要处理的数据范围,进行数据分区。

(6) Reduce 任务进程启动之后,根据 MRAppMaster 告知的待处理数据所在位置,从若干台运行 Map 任务的机器上获取若干个输出结果文件,并在本地进行重新归并排序,然后按照相同键值的键值对为一个组,调用客户定义的 Reduce 方法进行逻辑运算,并收集运算输出的结果键值对,然后调用客户指定的输出个数将结果数据输出到外部存储,通过空间数据转换接口将结果保存成 GeoJson 类型数据并存储在各个 HDFS 节点中,整个并行操作过程就此结束。

气象数据计算流程图如图 2 所示。

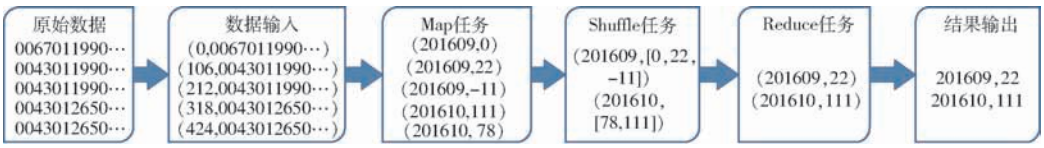

图 2 气象数据计算流程图

Fig. 2 Flow chart of weather data calculation

1.2.4 海量气象数据结构

由于云计算需要的数据类型应该是非格式化的数据,地理信息常用的格式 shp 或者 dbf 都是格式化数据,因此不满足云计算的数据要求。在地理空间非格式化数据中,GeoJson 基于 Json(Javascript 对象简谱),数据以键值对的形式进行存储,可以充分满足这种数据结构要求,也符合开放地理空间信息联盟(OGC)标准。另一方面,在前端进行三维可视化时,通过 AJAX(异步 Javascript 和 XML)也可以很方便地使用这种格式。GeoJson 数据的 geometry 属性中的 type 字段包含了点、线、面、多点、多线、多面等常用的地理信息系统几何类型,因此本研究采用 GeoJson 作为 Hadoop 中的分布式存储管理格式。示例数据中具有地理实体的唯一标识符号 id、地理实体坐标信息 coordinates、地理实体的气象属性信息数组 properties 等。其数据格式为:{

"type": "FeatureCollection",

"totalFeatures": 1376,//要素数量合计

"features": [{

"type": "Feature",//要素类型

"id": "china_air_quality20171216_0.1",//要素编号

"geometry": {//几何属性

"type": "Point",//几何类型为点状要素

"coordinates": [116.366, 39.8673]//要素的经纬度坐标},

"geometry_name": "the_geom",

"properties": {//要素的属性表

"监测点编码": "1001A","监测点名称": "YN001","经度": 116.366,"纬度": 39.8673,"日期": 20161216,"时间": 0,"AQI": 41,"$PM_{2.5}$": 14,"$PM_{2.5}$_24h": 46,"PM_{10}": 41,"PM_{10}_24h": 41,"SO_2": 1,"SO_2_24h": 6,"NO_2": 7,"NO_2_24h": 32,"O_3": 60,"O_3_24h": 63,"O_3_8h": 59,"O_3_8h_24h": 60,"CO": 0.2,"CO_24h": 0.7}

},...],

"crs": {"type": "name",

"properties": {"name":"urn:ogc:def:crs:EPSG::4326"}}}

本文使用网络爬虫爬取的气象数据包括逐日监测点编码、监测点名称、经纬度、获取日期、AQI、$PM_{2.5}$ 含量、PM_{10} 含量、SO_2 含量、NO_2 含量、O_3 含量、CO 含量、气温、降水量、相对湿度、日照时数等参数。使用的数据为从爬取的数据中筛选出的 2016 年云

南省气象台站历史数据。

2　平台试验

系统代码编写工具为 Ecplise，版本为 Mars。使用 Maven 作为项目管理构建工具。

2.1　海量气象数据网络爬虫

获取 PhantomJSDriver 的工具类实例代码如下

```
public static WebDriver getPhantomJs( ) {
String osname = System. getProperties( ). getProperty
("os. name");
if (osname. equals("Linux")) {
System. setProperty ( " phantomjs. binary. path ",
"/usr/bin/phantomjs");
} else {
System. setProperty ( " phantomjs. binary. path ",
". /phantomjs/win/phantomjs. exe");}
DesiredCapabilities = DesiredCapabilities. phantomjs( );
    desiredCapabilities. setCapability ( " phantomjs.
page. settings. userAgent", " Mozilla/5. 0 ( Windows
NT 6. 3; Win64; x64; rv:50. 0) Gecko/20100101
Firefox/50. 0");
    desiredCapabilities. setCapability ( " phantomjs.
page. customHeaders. User-Agent ", " Mozilla/5. 0
(Windows NT 6. 3; Win64; x64; rv: 50. 0) Gecko/
20100101 Firefox/50. 0");
if (Constant. isProxy) {
   org. openqa. selenium. Proxy proxy = new org.
openqa. selenium. Proxy( ); proxy. setProxyType( org.
openqa. selenium. Proxy. ProxyType. MANUAL);
proxy. setAutodetect(false);
String proxyStr = "";
do {
proxyStr = ProxyUtil. getProxy( );
} while (proxyStr. length( ) = =0);
proxy. setHttpProxy(proxyStr);
desiredCapabilities. setCapability ( CapabilityType.
PROXY, proxy);
}
return new PhantomJSDriver(desiredCapabilities);
}try{
WebDriver = PhantomJsUtil. getPhantomJs( );
webDriver. get(url);
SleepUtil. sleep(Constant. SEC_5);
PhantomJsUtil. screenshot(webDriver);
WebDriverWait wait = new WebDriverWait(webDriver,
10);
wait. until ( ExpectedConditions.
presenceOfElementLocated(By. id(inputId)));
    Document = Jsoup. parse ( webDriver.
getPageSource( ));
    }finally{
if (webDriver ! = null) {
webDriver. quit( );}}
```

2.2　气象大数据云计算

针对海量影像数据的存储，利用 HDFS 技术对 2TB 左右容量的全球栅格地图进行分节点管理。对于海量文本格式的气象地理信息数据，利用 MapReduce 框架实现分布式计算功能以及云南省空气质量监测站点空间位置及其气象参数的快速查询展示。

2.2.1　集群部署

Linux 环境下 Centos 版本为 7.4，Hadoop 版本为 Hadoop3.0，JDK 版本为 Java 1.8_161。本次试验采用 8 个服务器节点组成的集群。配置 8 个节点的 IP 地址、机器名称以及其代表的角色、网络配置，如表 1 所示。

表 1　集群各节点地址及其角色

Tab. 1　Cluster node addresses and their roles

主机名	IP 地址	角色名	网络速率/(MB·s^{-1})
Master	192.168.40.132	NameNode	100
Slave1	192.168.40.133	DataNode	100
Slave2	192.168.40.134	DataNode	100
Slave3	192.168.40.135	DataNode	100
Slave4	192.168.40.136	DataNode	100
Slave5	192.168.40.137	DataNode	100
Slave6	192.168.40.138	DataNode	100
Slave7	192.168.40.139	SecondaryNameNode	100

在 Master 节点上的 hosts 文件中添加集群中各节点的主机名和 IP 地址。安装 jdk1.8 环境、Hadoop3.0 环境到名称节点并远程复制到其余 7 个数据节点上。在数据节点上修改 Hadoop 目录下的 /etc/hadoop/workers 为数据节点的机器名称。最后配置 Hadoop 集群环境：①core-site.xml 是 Hadoop 的核心配置文件，这里需要配置两个属性，fs.default.name 配置 Hadoop 的 HDFS 系统的名称，位置为主机的 9000 端口。hadoop.tmp.dir 配置 Hadoop 的临时目录根位置。②hdfs-site.xml 是 HDFS 的配置文件，dfs.http.address 配置 HDFS 的 http 访问位置，dfs.replication 配置文件块的副本数，一般不大于从机的个数。③配置文件 mapred-site.xml 是 MapReduce 任务的配置，由于 hadoop2.x 使用了

Yarn 框架,所以要实现分布式部署,必须在 mapreduce.framework.name 属性下配置为 Yarn。其中 mapred.map.tasks 和 mapred.reduce.tasks 分别为 Map 和 Reduce 的任务数。④配置节点 yarn-site.xml,该文件为 Yarn 框架的配置,为一些任务的启动位置。

为了方便集群的维护,Hadoop 自带了一个历史服务器,可以通过历史服务器查看已经运行完的 MapReduce 作业记录,比如用了多少个 Map 或者 Reduce、作业提交时间、作业启动时间、作业完成时间等信息。默认情况下,Hadoop 历史服务器是没有启动的,可以通过/hadoop-3.0.0/sbin/mr-jobhistory-daemon.sh start historyserver 命令来启动 Hadoop 历史服务器。这样就可以访问主机的 19888 端口,查看已经运行完的气象数据分析作业情况。

2.2.2 MapReduce 分布式计算代码

(1)Map 阶段代码

```
public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
String line = value.toString();
String year = line.substring(15, 19);
int airTemperature;
if(line.charAt(14) == '+')
{airTemperature = Integer.parseInt(line.substring(11, 12));}
else{
airTemperature = Integer.parseInt(line.substring(15, 16));}
String temperature = line.substring(19, 21); if(airTemperature != MISSING && temperature.matches("[01459]"))
{context.write(new Text(year), new IntWritable(airTemperature));}}
```

(2)Reduce 阶段代码

```
public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException, InterruptedException {
int minValue = Integer.MAX_VALUE;
for(IntWritable value:values)
{minValue = Math.min(minValue, value.get());}
context.write(key, new IntWritable(minValue));}
```

MapReduce 程序的运行步骤为:启动 HDFS 和 Yarn,然后在集群中的任意一台服务器上启动执行程序[hadoop@lt mapreduce] $ hadoop jarhadoop-mapreduce-gis.jar geojson/ncep1979 - 2017.gz/geojson/out。

2.3 基于 Cesium 的三维可视化

采用 Cesium 框架的三维可视化进行前端结果展示,实现如下功能:①海量矢量格式数据的查询渲染展示,利用 Cesium 的 Entity 来实现任意渲染,有诸多点、线、面渲染形式可供选择,点击可查看管理其属性,也可以利用字段模糊查询得到结果,使用 Cesium 的 Infobox 模块可以实现,最终以三维可视化的方式展示给用户。②发布分布式文件系统中的影像地图为 Geoserver 地图服务,并加载显示。③利用 Ceisum 调用 OpenStreetMap 的开源兴趣点(POI)搜索库,实现气象站点的搜索定位功能。④利用 AJAX 技术调用 Geoserver 的 Web 地图要素服务(WFS)、Web 地图服务(WMS)、地理标记语言服务(GML),可以获取并发布通过云计算分析得到的矢量数据、元数据、图例等信息,用于动态展示。⑤空间分析功能,在查询输入框内输入需求可以获得最终的分析结果,并且叠加在三维地球上以图形的形式展示。⑥利用 Cesium 实现常用的测量、标绘等功能。⑦其他辅助性功能,如地图缩放功能支持底图显示 18 个级别,点击三维底图上的矢量数据实体可以提取出该实体的所有气象相关属性信息并展示在右上角的小窗口中。可视化界面如图 3 所示。

图 3 气象数据三维可视化界面

Fig. 3 3D visualization interface of meteorological data

Cesium 三维可视化渲染代码如下

```
function ShowAttribute(attribute) {
    viewer.entities.removeAll();
    var coordinate;
    var attributeSizeArray = [];
   for (var i = 0; i < shp_data.length; i++) {
attributeSizeArray.push(shp_data[i].properties[attribute]);
   }
   var minattributeSize = attributeSizeArray.min();
      var maxattributeSize = attributeSizeArray.max();
       for (var i = 0; i < shp_data.length; i++) {
          //判断该属性如果是负数,则改为正数,并且
```

渲染为柱状图

```
var size = ((shp_data[i].properties[attribute]-minattributeSize)/(maxattributeSize-minattributeSize)) *100;
// alert(maxattributeSize);
if (shp_data[i].geometry.type == "Point")
{
// alert(shp_data[i].geometry.type);
coordinate = shp_data[i].geometry.coordinates;
// alert(coordinate);
}
if (shp_data[i].geometry.type == "MultiPolygon") {
// alert(shp_data[i].geometry.type);
coordinate = shp_data[i].geometry.coordinates[0][0][0];
// alert(coordinate);
}
addentity(coordinate, size, shp_data[i].properties[attribute],shp_data[i].properties.NAME);
}
}
```

3　试验结果与分析

本文主要针对传统 WebGIS 服务器与 Hadoop 集群环境下海量气象数据的存储与计算进行性能对比。选择 8 台服务器节点作为集群运行环境，节点 CPU 为 i5 处理器，频率为 2.7 GHz，内存均为 8 GB，硬盘容量为 500 GB。试验数据为 1996—2016 年云南省气象信息，数据量约为 4.6 GB。为了对比集群中节点个数对气象数据存储及管理的影响，使用 4 种方案配置节点，集群节点个数分别为单节点、2 个节点、4 个节点、8 个节点，集群中部分节点启动后的页面如图 4 所示。

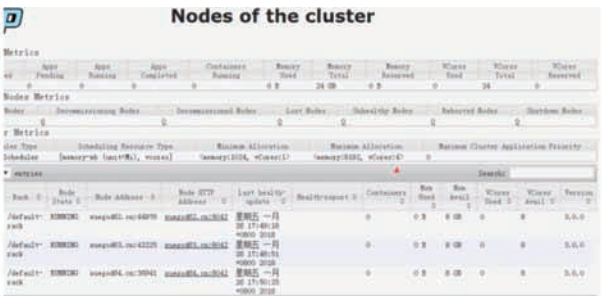

图 4　集群运行页面

Fig. 4　Cluster operation page

3.1　集群随节点变化的计算性能

随着节点数的变化，数据集中的气温最大值、最小值、平均值的计算消耗时间如图 5 所示。可以看出，集群随着节点数的增加，计算性能增加，但是节点越多，数据传输通信时间成本越大，因此计算性能随节点数的增大速率降低。

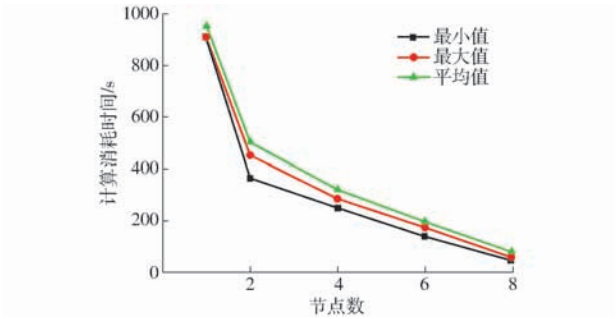

图 5　计算消耗时间与节点数的关系

Fig. 5　Relationship of calculating time with number of nodes

3.2　集群随 Map 任务并行度变化的计算性能

随 Map 任务并行度变化的集群计算性能试验结果如图 6 所示，通过试验发现，每个节点的最优并行度为 13 ~ 15 个 Map 任务，每个 Map 任务的执行时间至少 1 min。如果每个作业的 Map 任务或者 Reduce 任务的运行时间都只有 30 ~ 40 s，那么就减少该作业的 Map 任务或者 Reduce 任务数量。因为调度器在调度任务时，中间过程可能要花费几秒钟，如果每个任务都非常快就跑完了，则会浪费太多中转调度时间。

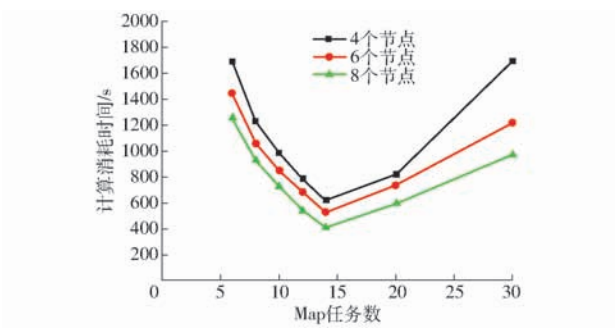

图 6　计算消耗时间与 Map 任务数的关系

Fig. 6　Relationship of calculating time with number of Map tasks

配置作业的 Java 虚拟机重用可以改善上述问题，Java 虚拟机重用技术不是指同一作业的两个或两个以上的任务可以同时运行于同一 Java 虚拟机上，而是排队按顺序执行。mapred. job. reuse. jvm. num. task，默认是 1，表示一个 Java 虚拟机上最多可以顺序执行的任务数目是 1，也就是说一个任务启用一个 Java 虚拟机。在 mapred-default. xml 文件中配置块容量，如果输入的文件非常大，比如 1TB，可以考虑将 HDFS 上的每个块容量设大，比如设成 256 MB 或者 512 MB。Reduce 任务的并行度同样影响整个作业的执行并发度和执行效率，但与 Map 任

务的并发数由切片数决定不同,Reduce 任务数量可以直接手动设置,默认值是 1,可以手动设置为 4,即 job. setNumReduceTasks(4)。如果数据分布不均匀,就有可能在 Reduce 任务阶段产生数据倾斜,因此要注意 Reduce 任务数量并不是任意设置,还要考虑业务逻辑需求,有些情况下,需要计算全局汇总结果,就只能有一个 Reduce 任务。

3.3 集群量随节点变化的存储性能

如图 7 所示,集群随着节点增加,存储性能变高,最重要的是如果数据量超出了单节点服务器硬盘容量,则无法进行存储,而集群多节点架构可以解决这一问题。

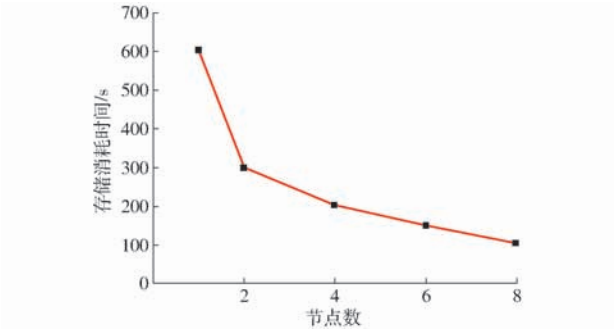

图 7 存储消耗时间与节点数的关系

Fig. 7 Relationship of storage time with number of nodes

4 GIS 平台在农业上的应用

农业数据类型包括农业生产数据、资源数据、技术数据、市场经济数据以及政策法规数据等,分为非结构化数据和结构化数据。在地理因素和季节等因素影响下,农业数据表现出了特殊的数据离散性和实效性。随着农业信息化程度不断提升,我国农业现代化的步伐也随之加快,转型升级在逐步进行,农业及其相关数据正在被大量收集、归纳、整理。随着农田中物联网设备的大量布署,农业数据源源不断地产生,形成农业大数据。该数据有以下特性:①数据量非常巨大,并且会连续产生。②因为作物生长在时间方面具有季节性,故农业大数据必须具有时效性,需及时处理数据并且反馈结果。③农业大数据的种类繁杂。④数据量巨大造成数据价值密度较低,但是价值量非常大。如果可以有效利用这些数据,将会大大加快农业信息化的进程。

4.1 农作物生长区域大数据决策分析

利用气象大数据分析昆明市种植适宜地,使用 ID3 决策树分类方法对气象大数据信息进行分析,决策树分类是常用的分类挖掘模型,本次试验在海量的 AQI、$PM_{2.5}$浓度、降水量、相对湿度、日照时数等大数据中挖掘、分析出具体种植适宜地,可以为农业生产适宜区域的选择提供决策支持。

ID3 分类方法是以信息增益来评判属性,选择属性分裂后信息属性增益最大的进行分裂,采用贪心思想遍历所有决策空间。使用云计算架构,分类计算步骤为:

(1)输入样本属性集 A,样本类别集 B,样本训练集 C。其中样本训练集 C 如表 2 所示。

(2)创建样本节点 R,如果训练集 C 为空,则返回父节点中多数类标记 R;如果训练集 C 中样本属于同一类别 B,则标记类 B 的节点 R 为该叶子节点;如果 A 为空则返回 C 中的多数类标记;如果计算得出了 A 中增益率最大的属性为 S,则用 S 标记节点 R。

(3)根据计算出的 S 的值$\{s_i \mid i=1,2,\cdots,m\}$将训练集 C 分成$\{C_i \mid i=1,2,\cdots,m\}$。

(4)递归执行 ID3TREE($R-S$, B, C_1), ID3TREE($R-S$, B, C_2), …, ID3TREE($R-S$, B, C_n),直至最终计算结果中的元组属于同一类,信息增益是原信息和新的需求信息的差,样本集信息熵的计算公式为

$$I(s_1,s_2,\cdots,s_n) = -\sum_{i=1}^{m} P_i \mathrm{lb} P_i \qquad (1)$$

式中 I——信息熵

P_i——训练集 C 中任意元组属于类 B_i 的概率

表 2 样本训练集

Tab. 2 Agricultural data set

序号	气温	天气	降水量		适宜种植
1	高	晴	高	…	是
2	高	雨	低	…	否
3	低	雨	高	…	否
4	低	多云	中	…	是
5	中	雨	高	…	否
6	低	雨	中	…	否
7	低	多云	低	…	是
8	高	晴	高	…	是

输入对应的键值,通过文中的云计算架构,对上述数据集进行分类属性的选择,建立 ID3 决策树,分别得出气温、天气、降水量、相对湿度等信息的信息增益,依据上述计算流程进行集群分布式运算,前端对运算结果进行可视化展示,该应用利用海量气象信息精准判断某种植区域是否适宜种植作物。

4.2 气象灾害评估大数据分析

在实际农业生产中,气象灾害评估的实时性和准确性面临极大考验,气象灾害的发生会导致农作物生长受到影响,产生巨大的经济损失。利用大数据分析则可以及时察觉即将到来的气象灾害,对灾害进行分类和灾害等级评估,提前采取预防措施,减

少经济损失。对此参考部分气象灾害的等级指标，建立其灾害等级指标，利用本平台大数据计算分析模块处理海量农业气象数据集，得到可用于气象灾害评估的信息。试验以影响作物生长的低温灾害指数为例，利用基于最近邻法（KNN）组合分类器分布式计算模块，将温度数据代入低温灾害指数公式中计算，将其转换为低温灾害指数，其低温灾害指数为

$$f=\begin{cases}0 & (t\geqslant 8℃)\\ 1-\dfrac{1}{\sigma\sqrt{2\pi}}\displaystyle\int_{-\infty}^{t} e^{\frac{-(t-u)^2}{2\sigma^2}}dt & (-30℃<t<8℃)\\ 1 & (t\leqslant -30℃)\end{cases} \tag{2}$$

式中　f——低温灾害指数

t——当前最小温度，℃

u、σ——正态分布参数

最后利用分类后的低温灾害风险指数进行灾害等级预测以及作物受损程度评估。利用云南省 1996—2016 年低温气象数据，通过计算得到昆明市的低温灾害风险指数，建立低温灾害评估等级，如表 3 所示，其中 1 级灾害作物受损较轻，还可正常生长，4 级灾害时，作物生长停止或者死亡，此时会发出预警，农户则可及时做相应预防措施减少损失。

表 3　低温灾害评估

Tab. 3　Low temperature disaster assessment

	低温灾害指数			
	0 ~ 0. 2	0. 2 ~ 0. 4	0. 4 ~ 0. 7	0. 7 ~ 1
灾害风险等级	1 级	2 级	3 级	4 级
作物受损程度	轻度	较严重	严重	极其严重
灾害次数	68 212	140 221	28 312	21 223

5　结束语

基于分布式气象大数据分析的 GIS 平台采用 Hadoop 体系架构，利用数据爬取技术在互联网上获取海量气象数据，并通过云存储技术进行分布式存储，解决了传统单节点服务器 WebGIS 系统硬件受限制的问题。在气象大数据的分析计算方面，试验结果表明，多节点集群下效率更高，在查询遍历性能方面也比传统 WebGIS 单节点服务器高；通过对海量气象数据采用云计算技术进行分析可以帮助有关部门进行决策；利用 Cesium 对计算得到的气象信息进行三维可视化，可以直观看到气象站点历史气象参数的变化情况和计算决策结果，但是在具体分析功能扩展方面还需要完善。

参　考　文　献

[1] 陈睿嘉，康志忠，张卫涛. 基于网络爬虫的导航深度服务信息自动采集[J]. 测绘工程，2015，24(1)：17 - 24.
CHEN Ruijia, KANG Zhizhong, ZHANG Weitao. Web based crawler navigation depth service information automatic acquisition [J]. Mapping Engineering, 2015, 24(1): 17 - 24. (in Chinese)

[2] 董日壮，郭曙超. 网络爬虫的设计与实现[J]. 电脑知识与技术，2014，10(17)：3986 - 3988.
DONG Rizhuang, GUO Shuchao. Design and implementation of web crawler [J]. Computer Knowledge and Technology, 2014, 10(17): 3986 - 3988. (in Chinese)

[3] 侯东阳，武昊，王军锋，等. 基于深层网络爬虫的 Web 地图服务发现方法[J]. 地理与地理信息科学，2015，31(5)：10 - 13.
HOU Dongyang, WU Hao, WANG Junfeng, et al. Web map service discovery method based on deep web crawler [J]. Geographic and Geographic Information Science, 2015, 31(5): 10 - 13. (in Chinese)

[4] MAYOR S. Internet crawler uses unconventional information sources to track infectious disease outbreaks[J]. British Medical Journal, 2008, 337: a742.

[5] XU S, YOON H J, TOURASSI G. A user-oriented web crawler for selectively acquiring online content in e-health research[J]. Bioinformatics, 2014, 30(1): 104 - 114.

[6] 张亮. 基于 HTMLParser 和 HttpClient 的网络爬虫原理与实现[J]. 电脑编程技巧与维护，2011，18(20)：94 - 103.
ZHANG Liang. The principle and implementation of web crawler based on HTMLParser and HttpClient [J]. Computer Programming Skills and Maintenance, 2011, 18(20): 94 - 103. (in Chinese)

[7] HOEHNDORF R, GKOUTOS G V, SCHOFIELD P N. Datamining with ontologies[J]. Methods in Molecular Biology, 2016, 1415: 385 - 397.

[8] SCHWEIGEL H, WICHT M, SCHWENDICKE F. Salivary and pellicle proteome: a datamining analysis [J]. Scientific Reports, 2016, 6: 38882.

[9] 邓仲华，刘伟伟，陆颖隽. 基于云计算的大数据挖掘内涵及解决方案研究[J]. 情报理论与实践，2015，38(7)：103 - 108.
DENG Zhonghua, LIU Weiwei, LU Yingjun. Research on the connotation and solution of big data mining based on cloud computing [J]. Intelligence Theory and Practice, 2015, 38(7): 103 - 108. (in Chinese)

[10] 李德仁，张良培，夏桂松. 遥感大数据自动分析与数据挖掘[J]. 测绘学报，2014，43(12)：1211 - 1216.
LI Deren, ZHANG Liangpei, XIA Guisong. Automatic analysis and data mining of remote sensing big data [J]. Journal of Surveying and Mapping, 2014, 43(12): 1211 - 1216. (in Chinese)

[11] 彭昱忠，王谦，元昌安，等. 数据挖掘技术在气象预报研究中的应用[J]. 干旱气象，2015，33(1)：19 - 27.

PENG Yuzhong, WANG Qian, YUAN Chang'an, et al. Application of data mining technology in meteorological forecast research [J]. Drought Meteorology, 2015,33(1): 19 - 27. (in Chinese)

[12] DURUZ S, FLURY C, MATASCI G, et al. A WebGIS platform for the monitoring of farm animal genetic resources (GENMON)[J]. PLoS One, 2017, 12(4): e176362.

[13] KHOSHABI M, TALEAI M, MOTLAGH A, et al. Developing a WebGIS for geo-visualization of cancer[J]. Iran J Cancer Prev, 2016, 9(2): e3910.

[14] KOLIOS S, STYLIOS C, PETUNIN A. A WebGIS platform to monitor environmental conditions in ports and their surroundings in South Eastern Europe[J]. Environ Monit Assess, 2015, 187(9): 574.

[15] 胡勇,刘奇峰. 基于 WebGIS 的分布式电动汽车充电桩运营管理系统设计与实现[J]. 电力建设, 2014,35(1): 98 - 103.
HU Yong, LIU Qifeng. Design and implementation of distributed electric vehicle charging pile operation management system based on WebGIS[J]. Electric Power Construction, 2014,35(1): 98 - 103. (in Chinese)

[16] 赵曦,姬建中,常俊,等. 基于 WebGIS 的地震数据服务系统建设及关键技术研究[J]. 灾害学, 2014,29(3): 224 - 228.
ZHAO Xi, JI Jianzhong, CHANG Jun, et al. Construction of WebGIS based seismic data service system and its key technologies [J]. Disaster Science, 2014,29(3): 224 - 228. (in Chinese)

[17] AJI A, SUN X, VO H, et al. Demonstration of Hadoop-GIS: a spatial data warehousing system over MapReduce[J]. Proc ACM Sigspatial Int Conf Adv Inf, 2013: 528 - 531.

[18] AJI A, WANG F, VO H, et al. Hadoop-GIS: a high performance spatial data warehousing system over MapReduce[J]. Proceedings VLDB Endowment, 2013, 6(11): 1009 - 1020.

[19] 何涛. 面向海量空间数据并行高效处理的存储模式设计与研究[D]. 成都:电子科技大学, 2014.
HE Tao. Design and research of storage mode for parallel and efficient processing of massive spatial data [D]. Chengdu: University of Electronic Science and Technology of China, 2014. (in Chinese)

[20] 刘磊,尹芳,冯敏,等. 基于开源 Hadoop 的栅格数据分布式处理[J]. 华中科技大学学报(自然科学版), 2013,41(7): 103 - 108.
LIU Lei, YIN Fang, FENG Min, et al. Distributed processing of raster data based on open source Hadoop [J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2013,41(7): 103 - 108. (in Chinese)

[21] 殷兵. 基于 Hadoop 的分布式遥感图像处理研究[D]. 上海:华东师范大学, 2015.
YIN Bing. Research on distributed remote sensing image processing based on Hadoop [D]. Shanghai : East China Normal University, 2015. (in Chinese)

[22] ABDIAN N, GHASEMID D P, HASHEMZADEH C M, et al. Comparison of human dermal fibroblasts (HDFs) growth rate in culture media supplemented with or without basic fibroblast growth factor (bFGF)[J]. Cell Tissue Bank, 2015, 16(4): 487 - 495.

[23] 尹颖,林庆,林涵阳. HDFS 中高效存储小文件的方法[J]. 计算机工程与设计, 2015,36(2): 406 - 409.
YIN Ying, LIN Qing, LIN Hanyang. Efficient storage of small files in HDFS [J]. Computer Engineering and Design, 2015, 36(2): 406 - 409. (in Chinese)

[24] 张建. 基于 Hadoop 的云计算模型研究及气象应用[D]. 南京:南京信息工程大学, 2012.
ZHANG Jian. Research on cloud computing model and meteorological application based on Hadoop [D]. Nanjing: Nanjing University of Information Science and Technology, 2012. (in Chinese)

[25] 周小平,刘祥磊. 海量铁路机车 GIS 定位数据分布式处理技术[J]. 中国科技论文, 2015,10(7): 812 - 816.
ZHOU Xiaoping, LIU Xianglei. Distributed processing technology of massive railway locomotive GIS positioning data [J]. China Science and Technology Thesis, 2015,10(7): 812 - 816. (in Chinese)

[26] 王凯,曹建成,王乃生,等. Hadoop 支持下的地理信息大数据处理技术初探[J]. 测绘通报, 2015(10): 114 - 117.
WANG Kai, CAO Jiancheng, WANG Naisheng, et al. Geographic information big data processing technology supported by Hadoop [J]. Surveying and Mapping Bulletin, 2015(10): 114 - 117. (in Chinese)

[27] 张海寰. 基于混合矢量结构的分布式道路选线方法与原型系统设计[D]. 成都:电子科技大学, 2014.
ZHANG Haihuan. Distributed road alignment method and prototype system design based on hybrid vector structure [D]. Chengdu:University of Electronic Science and Technology of China, 2014. (in Chinese)

[28] 胡戎,冯仲科,蒋君志伟. 基于 CheerIO 的 MEAN Stack 气象数据网络爬虫研究[J/OL]. 农业机械学报, 2016, 47(6): 275 - 282. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? flag=1&file_no=20160636&journal_id=jcsam. DOI: 10.6041/j.issn.1000-1298.2016.06.036.
HU Rong, FENG Zhongke, JIANG Junzhiwei. Research on MEAN Stack meteorological data crawler based on CheerIO [J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2016,47(6): 275 - 282. (in Chinese)