文章编号：2096-1618(2018)05-0536-04

气象大数据平台的数据采集与处理系统初探

李从英， 王 彪， 金石声， 郭 茜
(贵州省气象信息中心，贵州 贵阳 550002)

摘要：气象资料是典型的具有时间和空间属性的地球科学数据，其组成种类繁多，体量巨大，且实际应用中具有高时效性要求，因此快速获取和处理实时气象数据是开展气象数据应用的关键。通过阐述基于组件的不同种类数据可视化采集处理方法的设计和实现，完成大数据平台前端各类气象数据的快速采集，具有较强的借鉴意义。

关 键 词：气象大数据；采集；处理；高时效性；可视化
中图分类号：TP311.13 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2018.05.009

0 引言

面对日益增长的气象数据[1]，传统的数据存储和共享方式已经逐渐不能满足现代气象业务的应用发展。数据的存储和存取速度[2]严重影响了气象相关业务。目前大数据的发展[3]为气象服务[4]开阔了思路和方向[5]。

目前的大数据[6]处理技术各有自身优势特点与适合的场景，没有一种技术能够完全满足所有业务分析场景的数据处理要求[7]，本文构建的大数据平台从气象数据自身特点[8]以及实际业务应用出发，采用Hadoop[9]、Impala、RDB的混搭架构组建数据平台，有针对性地对各类气象数据进行存储。将结构化数据存储到Kudu分布式存储结构中，非结构数据存储在HDFS与HBase中，其中小于10 M的数据直接存储于HBase，大于10 M的数据直接存储于HDFS，非结构化数据的索引数据及相关专题应用数据存储到Oracle中。主要介绍气象数据同步到大数据平台中的处理流程。

1 采集过程

大数据平台中的数据包括CIMISS[10](China integrated meteorological information service system)中的结构化数据和非结构化数据，行业外部系统数据，互联网数据。针对各类数据的采集，通过不同的采集方式将数据写为标准csv文件，然后以文件新增的方式触发数据的采集入库流程。

CIMISS结构化数据和非结构化数据都是采用FTP/SFTP方式进行数据采集。CIMISS数据经过简约流程处理后，生成标准的CSV文件，其中增量数据(周期批量新增)与更新数据(包含单记录新增、更新、删除)单独分目录存放。大数据平台通过新增文件扫描方式进行数据的采集，将新增文件采集到大数据平台接口机进行预处理。其中增量数据采用load的方式进行数据批量加载到相应集群；更新数据生成相应的更新语句(Oracle的更新语句、Kudu的更新操作语句)到相应集群执行。

行业外部系统数据按照协商接口进行数据的采集。采集后将数据生成标准csv文件，按照文件方式进行后续处理流程。可以使用以下几种方式。第一可以使用与外部网络互通的接口机进行数据的采集，第二可以使用FTP/SFTP方式接口按照文件新增方式进行采集，第三可以使用其他方式接口(如socket)通过定制化程序实现数据的采集。

互联网数据是以爬虫的方式对互联网数据进行采集的，通过文件新增的方式触发后续采集入库流程。爬取后内容解析成的结构化数据，将其写成标准的csv文件，文件命名要求至少包含数据接口名称、时间信息，文件命名以英文、数字为主。爬取后的非结构化文件以原有数据文件格式进行存储。

2 处理过程

2.1 结构化数据处理

针对结构化增量数据处理，以CIMISS系统中的结构化数据为例，其典型的数据处理流程如图1所示。

收稿日期：2018-04-27
基金项目：贵州省科技技术资助项目(黔科合支撑[2017]2819)

根据配置的文件全路径规则表达式,实时监测有无数据从接口机服务器下载到 ETL[11] 服务器。接口机根据配置的清洗、转换规则对数据进行处理。对小文件进行合并处理,并将数据文件的压缩成 zip、gz、lzo 等压缩格式,压缩完成后自动删除原有文件。

如图 1 所示,压缩后的数据文件会通过支持 Hadoop 的加载接口和文件挂载方式加载到 HDFS(hadoop distributed file system)中。基础数据使用基于 Kudu 的 Spark 进行关联或汇总,最终将汇总数据同步到 Oracle 集群,并通过 load 方式加载到 Impala 集群。汇总完的基础数据、维度汇总数据及专题结论数据是从 HDFS 加载到 Oracle 中,并供上层应用使用。

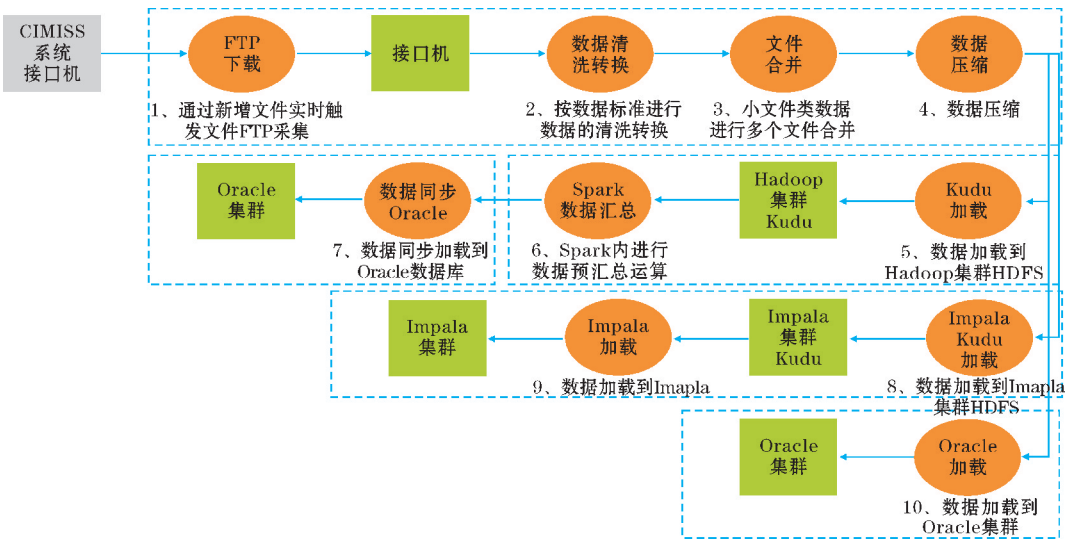

图 1　结构化增量数据处理流程

结构化更新数据处理流程,主要实现对已采集数据的更新及重新汇总,其典型处理流程见图 2(以 CIMISS 系统数据为例)。

根据配置的文件全路径规则表达式,实时监测有无新增文件,实现数据从接口机服务器下载到 ETL 服务器。根据更新数据及 Oracle 语法,生成 insert/update/delete 更新操作语句。连接 Oracle,执行生成的更新语句,对采集的原始数据进行更新。根据更新数据及 Kudu 语法,生成 insert/update/delete 更新操作语句。连接 Kudu,执行生成的更新语句,对采集的原始数据进行更新。对更新后原始数据根据相应汇总周期使用 Spark 进行数据的重新汇总。重新汇总的数据同步到 Oracle 临时表,并以 Merger 方式更新汇总数据。

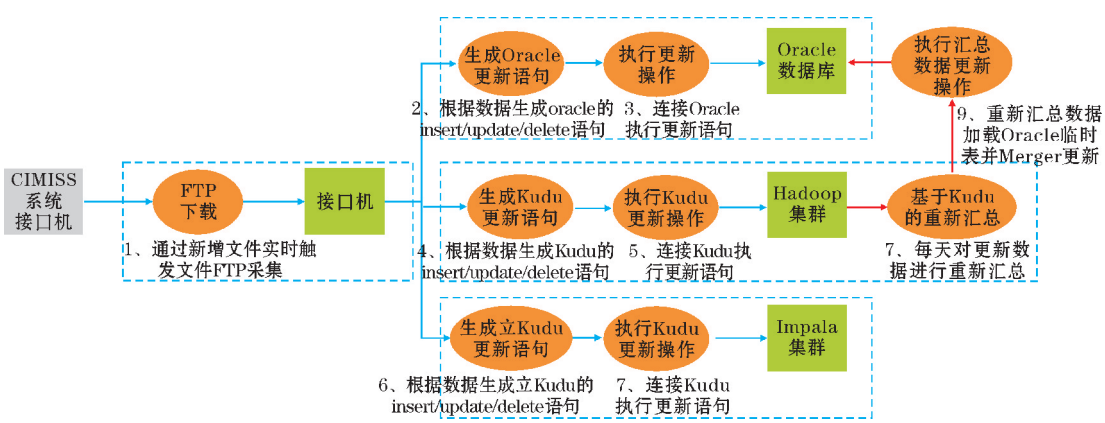

图 2　结构化更新数据处理流程

图 3 是一个具体的可视化的结构化数据处理流程,展示了中国地面逐小时资料的处理过程。在此流程中,首先设定下载后文件存放的规则,然后将 CSV 文件下载到接口机,CSV 文件经过规定的数据处理规则,存储在 HDFS 中 hive 设定的目录,通过 hive 将结构化的数据文件映射为一张临时数据库表,并将数据加载到 Kudu,并将临时表删除,最终对于所有基于原始数据的长序列查询通过 Impala 实现支撑。

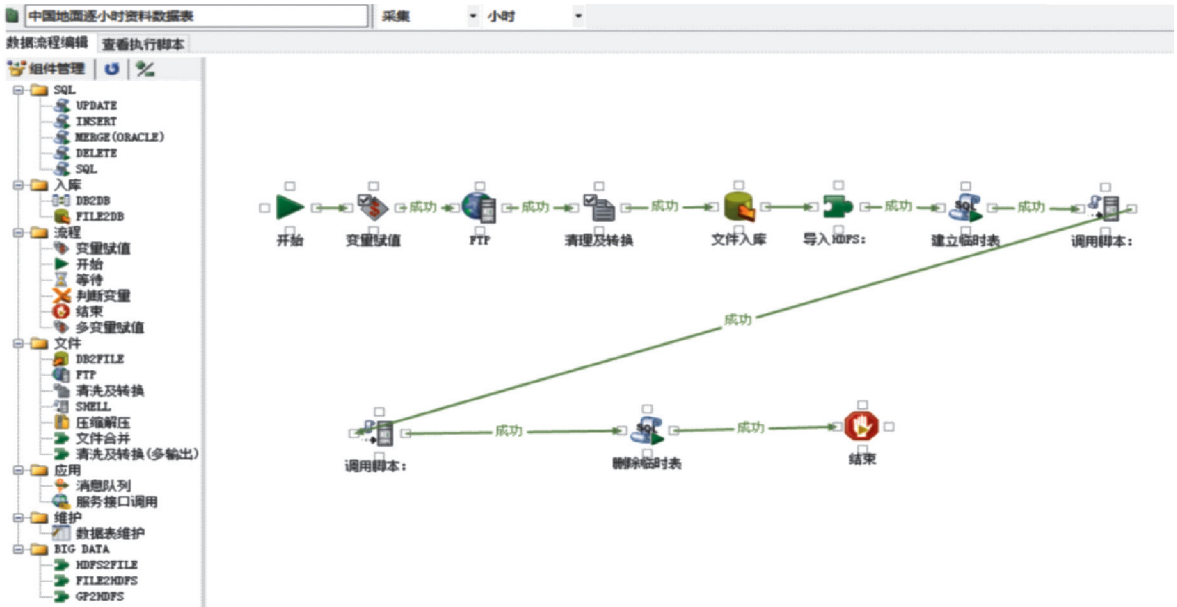

图 3 中国地面逐小时数据的处理流程

2.2 非结构化数据处理

非结构化数据处理流程如图 4 所示,主要是 CIMISS 系统中的非结构化数据和互联网采集后的非结构化数据。根据配置的文件全路径规则表达式,实时监测有无新增文件,实现数据从接口机服务器下载到 ETL 服务器。根据非结构化数据的文件名时间信息,构建 rowkey 加载至 Hbase,供应用查询。对大于 10 M的非结构化数据文件加载到 Hadoop 集群的 HDFS 中,支持两种加载方式:Hadoop 的加载接口、文件挂接方式。对于非结构化数据的索引数据则写入 Oracle 中。

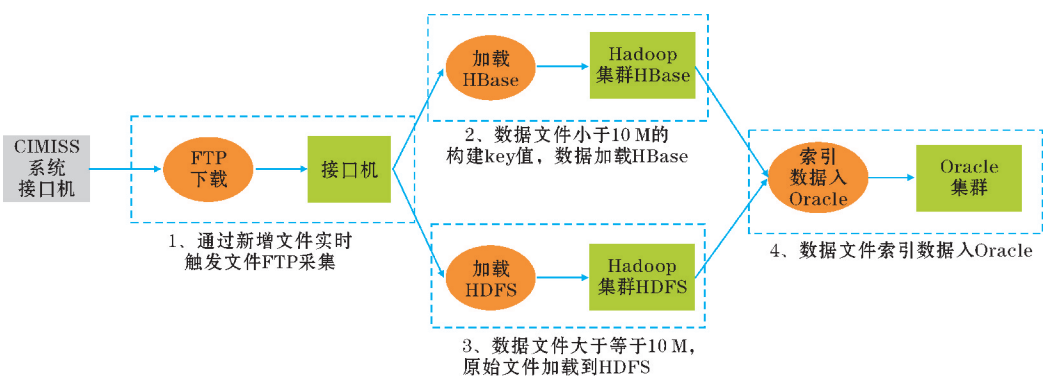

图 4 非结构化数据处理流程

3 结论

主要介绍了大数据平台中结构化数据和非结构化数据采用的不用的采集处理方式。通过可视化平台实现了数据的采集和处理。其中结构化基础数据存储在 Kudu 中,非结构化数据存储在 HDFS 或者 HBase 中,支撑非结构数据存储、查询及结构化数据的关联、处理。Oracle 集群作为应用库,存储日常统计数据和应用专题数据,支撑固定查询及专题应用,满足了数据需求的及时性,也充分发挥了气象数据的公众服务的作用[12],为特色专题服务提供了数据支撑。

参考文献:

[1] 沈文海. 气象数据的“大数据应用”浅析[J]. 中国信息化,2014(11):21-31.

[2] 李社宏. 大数据时代气象数据分析应用的新趋势[J]. 陕西气象,2014(2):41-44.

[3] 沈文海. 再析气象大数据及其应用[J]. 信息化研究,2016(1):84-96.

[4] 张蕾,杨勇,湛莹莹. 走进大数据时代的气象服务——气象“云”气象万千[N]. 中国气象报,2014-12-01.

[5] 汪惜金. 浅析气象大数据的未来应用服务趋势[J]. 信息通信,2017(4):290-291.
[6] 维克托·迈尔-舍恩伯格,肯尼斯·库克耶. 大数据时代[M]. 盛杨燕,周涛译. 杭州:浙江人民出版社,2013.
[7] 顾荣. 大数据处理技术与系统研究[D]. 南京:南京大学,2016.
[8] "气象大数据"以何种方式在气象领域蔓延[EB/OL]. http://www.chinawuliu.com.cn/zhxw/201402/18/278296.shtml.
[9] 周敏齐,王晓玲,钱卫宁,等. Hadoop 权威指南[M]. 2 版,北京:清华大学出版社,2010.
[10] 熊安元,赵芳,王颖,等. 全国综合气象信息共享系统的设计与实现[J]. 应用气象学报,2015,26(4):500-512.
[11] 张亮. 分布式数据仓库中 ETL 技术的研究[D]. 沈阳:沈阳航空工业学院,2009:1-17.
[12] 唐延婧,彭芳,罗喜平,等. 大数据在贵州专业气象服务的应用及展望[J]. 气象科技进展,2017,7(2):54-59.

Preliminary Study on Data Acquisition and Processing System of Meteorological Big Data Platform

LI Cong-ying, WANG Biao, JIN Shi-sheng, GUO Xi
(Meteorological Information Center of Guizhou Province, Guiyang 550002, China)

Abstract: Meteorological data is a typical of earth science data with the essentail property of time and space. There are many different kinds, with a wide range of properties. And it has a very high demand to the time limit efficiency. So, rapid acquiring and processing real-time meteorological data is the key to application of meteorological data. This paper illustrates the design and realization for the visual acquiring and processing of different kinds data. The role of the design is to collect data quickly for the front-end of big data platform. It has stronger reference meanings.

Keywords: meteorological big data; data acquisition; data processing; high efficiency; visualizations