“天镜”全流程指标计算功能优化

徐　达,曾　乐,王英杰
(国家气象信息中心,北京 100081)

摘　要:气象综合业务实时监控系统-“天镜”为全国气象部门针对基础观测数据和产品在收集、分发、入库、同步各个环节提供数据全流程监视服务。“天镜”系统中气象资料全流程的到报率和及时率指标作为对全国气象站上行资料的考核依据,是一线值班人员重点关注对象。为保障“天镜”系统的可靠性和高效性,解决目前“天镜”系统全流程指标计算慢的问题,基于 Spark 大数据计算引擎优化了全流程计算策略。文中将原本复杂的计算任务按气象资料类型和是否考核进行拆解,并重构了 Spark 任务生成的模板。结果表明,在相同的物理条件下,将单个庞大的计算任务拆解成多个子任务并行处理可以有效地提升 Spark 集群的计算效率,优化“天镜”系统中地面区域站气象资料的实时展示效果。

关键词:气象;“天镜”;Spark 计算;全流程;大数据
中图分类号:TP319　　**文献标识码**:A　　**文章编号**:1673-629X(2023)07-0020-07
doi:10.3969/j.issn.1673-629X.2023.07.003

Optimization of Calculation Function of “The Mirror” Whole Process Index

XU Da,ZENG Le,WANG Ying-jie
(National Meteorological Information Center,Beijing 100081,China)

Abstract:The meteorological integrated business real-time monitoring system,"The Mirror",provides the whole process monitoring service for the national meteorological departments in the collection,distribution,storage,and synchronization of basic observation data and products. The indicators of the arrival rate and timeliness rate are served as the assessment basis for the national meteorological stations´ uplink data,which are focused by operators on duty. To ensure the reliability and efficiency of the "The Mirror" system and solve the problem of the calculation speed,the whole process calculation strategy is optimized based on Spark big data calculation engine. In this paper,the original complex calculation task is disassembled according to the type of meteorological data and whether it is evaluated. What's more,we reconstruct the Spark task template. The result shows that under the same physical conditions,disassembling a single huge computing task into multiple subtasks for parallel processing can improve the computing efficiency of Spark cluster and optimize the real-time display effect of meteorological data of ground regional stations in "The Mirror" system.

Key words:meteorology;“The Mirror”;Spark calculation;whole process;big data

0　引　言

气象综合业务实时监控系统—“天镜”[1]是国家气象信息中心为建设统一数据环境、整合分散独立的监视业务建立的通用、综合、高效的集约化监视平台。“天镜”能够为全国气象部门在收集、分发、入库、数据同步各个环节提供实时观测数据和产品的数据全流程监视服务。目前“天镜”每小时接收处理气象业务监视全流程[2]数据记录达到 3 千万条,累计接入的数据资料超过 400 种,为了使目前的数据全流程监视业务可以更高效地在大数据计算和分布式存储架构上运行,需要对目前海量监视数据的处理中加大对计算策略和存储策略的研究力度。

Spark[3]是对海量数据计算处理的重要工具和手段,是基于弹性分布式数据集(RDD)的数据结构,具有数据流模型特点。RDD 将数据保留在内存中,且允许用户程序多次查询,降低了对磁盘和网络的开销,适用于在线计算和迭代计算。“天镜”系统使用 Spark 计算全流程数据,并按全国、省、市县维度的统计指标进

收稿日期:2022-04-24　　**修回日期**:2022-08-26
基金项目:国家发展改革委工程建设项目(发改投资〔2021〕231 号);国家气象信息中心“气象综合业务智能监控”创新团队攻关任务(NMIC-202011-05)
作者简介:徐　达(1992-),男,硕士,工程师,研究方向为气象监控技术;通讯作者:曾　乐(1977-),女,博士,高级工程师,研究方向为分布式数据存储与信息分析。

行汇聚。气象资料的接入和监视环节的扩展使得需要计算和处理的监视信息激增,使得 Spark 运行作业时间变长,这对于满足时效性要求而言需要缩短计算任务的运行时间,一种方式是从 Spark 集群框架和配置参数进行修改和优化,另一种方式则是通过对程序代码进行改动,采用最优的计算策略来提升计算效率。

1　研究环境简介

1.1　国内外监控系统分析

2017 年 10 月,中国气象局批复了由国家气象信息中心牵头,国家级各业务单位共同参与建设的气象综合业务实时监控系统(一期)项目。该项目旨在建立技术先进的监控系统技术框架,实现综合监视和告警运维核心功能,建立规范的监控信息采集接口,监视范围横向覆盖气象资料现有数据流程各环节,纵向覆盖信息系统从网络及安全、服务器、存储、中间件、应用软件运行状态。气象综合业务实时监控系统(一期)计划 2018 年底建成后,将完成气象综合业务实时监控的基础框架,建立系统的硬件平台和技术平台,从技术上解决了原 MCP 系统面临的性能瓶颈问题,建立规范化的监视信息采集接口,实现监视告警的核心功能,实现国家级基于 CIMISS 数据环境的资料数据流程的收集、分发、解码入库、接口服务等环节的监视,以及 CMACast 卫星广播系统、部际系统等系统的监视。但是随着监视信息不断增长,现有的运行环境在处理计算上会有延迟,尤其是在中国地面分钟级资料的实时监视上会出现页面为 0 的情况[4]。

国外气象行业的监视系统也是主要围绕着数据传输网络、数据收集生成、数据质量、观测设备状态进行监控,如美国国家海洋和大气管理局(NOAA)建设了观测系统监控中心(OSMC)实时监测全球海洋观测系统的性能[5],欧洲中期天气预报中心(ECMWF)通过常规观测告警系统检测数据可用性和质量问题[6],美国国家环境预报中心(NCEP)的实时数据监测系统(RTDMS)主要监测数据的数量和时效性[7]。国外的数据监视系统是基于传统的数据资料文件入库,并对该文件资料进行质量评估后,绘制该类观测资料的打点时序图,对资料进行分类监视。ECMWF 和 NOAA 更加侧重资料到报后的质量情况,通过设计测试的数值预报模式来校验到报的观测资料是否合格,通过地图打点的方式提供数据服务,并用颜色来区分该类资料的数据质量情况。

1.2　“天镜”系统

围绕《全国气象发展“十三五”规划》提出的“智慧气象”发展目标,气象业务在实施现代化、信息化、集约化、标准化的进程中,都需要监控系统来保障业务的高效稳定运行。但是,各气象业务的现有监控系统都是独立开发和运维,监控系统分散且数量庞大,运行维护人力成本高;各监控系统仅监控业务流程中的独立环节,上下游监控信息无法共享,缺乏对业务全流程的总体监控,出现故障时准确定位故障位置困难、分析故障原因不及时,导致业务监控运维效率低。因此,急需实现对观测、信息、预报预测、公共服务及政务的全流程、全要素、全过程的一体化监控和运维,以提升气象业务运行管理的质量和效率。2016 年底,按照中国气象局统一部署,由预报司牵头组织与协调,观测司配合,信息中心作为实施技术组组长单位,协同各成员单位上下一心,通力合作,共同推动气象综合业务实时监控系统建设,树立和打造气象综合业务监控品牌——“天镜”[8]。

1.3　气象全流程指标

气象全流程监控实现对数据从收集、分发、入库、数据同步到应用的全流程、全生命周期监控。在收集环节由国内气象传输系统(CTS)收到气象资料后,经过文件打包处理后,把文件分发给业务系统和用户。在入库环节中解码入库程序按照气象要素、时次等条件进行拆解,按照存储规则录入不同的数据库中。为了提供气象资料查询服务,需要将解码后的数据在不同类型库中进行同步。在气象资料全流程监视设计中需要对收集、分发、入库、同步环节进行监视。全流程实时指标见表 1,计算依赖于节目表信息和总控配置信息,节目表信息用来指定该类气象资料资料是否为考核资料,总控配置信息主要包含:资料业务时次配置信息、单站的单环节的单时次及时配置信息、统计规则(时次、时次截日、时次截小时、小时、日)、各个环节之间的关联关系、文件级资料的应收数、检测告警开始时间、需要告警指标、告警持续时间等相关配置。

表 1　全流程实时计算收集环节核心指标

序号	属性名称	英文标识	属性含义	参数类型
1	国家	COUNTRY	国家	string
2	国家编码	COUNTRY_ID	国家编码	string
3	省份(缺失为全国标识)	PROVINCE	省份(缺失为全国标识)	string
4	时次	DATA_TIME	时次	string

续表 1

序号	属性名称	英文标识	属性含义	参数类型
5	CTS 编码	CTS	CTS 四级编码	string
6	SOD 编码	SOD	SOD 四级编码	string
7	考核应收	CO_CHECK_TD	考核应收	long
8	考核及时收	CO_CHECK_INTIME_ACTUAL	考核及时收	long
9	考核逾期收	CO_CHECK_LATETIME_ACTUAL	考核逾期收	long
10	考核实收	CO_CHECK_ACTUAL	考核实收	long
11	考核缺收	CO_CHECK_LOC	考核缺收	long
12	考核及时率	CO_CHECK_INTIME_RATE	考核及时率	float
13	考核到报率	CO_CHECK_RATE	考核到报率	float
14	应收	CO_TD	应收	long
15	及时收	CO_INTIME_ACTUAL	及时收	long
16	逾期收	CO_LATETIME_ACTUAL	逾期收	long
17	实收	CO_ACTUAL	实收	long
18	缺收	CO_LOC	缺收	long
19	及时率	CO_INTIME_RATE	及时率	float
20	到报率	CO_RATE	到报率	float
21	统计时间	COUNT_TIME	统计时间	long

1.4 Spark 计算引擎

Spark 是一种快速、通用、可扩展的大数据分析引擎,2009 年诞生于加州大学伯克利分校 AMPLab,2010 年开源,2013 年 6 月成为 Apache 孵化项目。目前 Spark 生态系统已经发展成为一个包含多个子项目的集合,包含 SparkSQL、Spark Streaming、GraphX、MLlib、等子项目。

Spark 是基于内存计算的大数据并行计算框架,与 Hadoop 的 MapReduce 相比,Spark 基于内存的运算速度更快,同时保证了高容错性和高可伸缩性,Spark 实现了高效的 DAG 执行引擎,从而可以通过内存来高效处理数据流[9-10]。

在“天镜”中,Spark 的体系架构如图 1 所示。“天镜”采用 Standlone 模式部署 Spark 集群,通过 Zookeeper,一个开源的分布式应用程序协调服务软件进行集群管理,在 Spark 集群上创建常驻的 SparkSession 即常驻的 Driver 进程用于交互 Spark 程序,SparkSession 中包含开源的 ActorSystem,一套开源的用于设计跨处理器和网络的可扩展弹性系统。服务端的 ActorSystem 向 Zookeeper 注册自身的地址。在

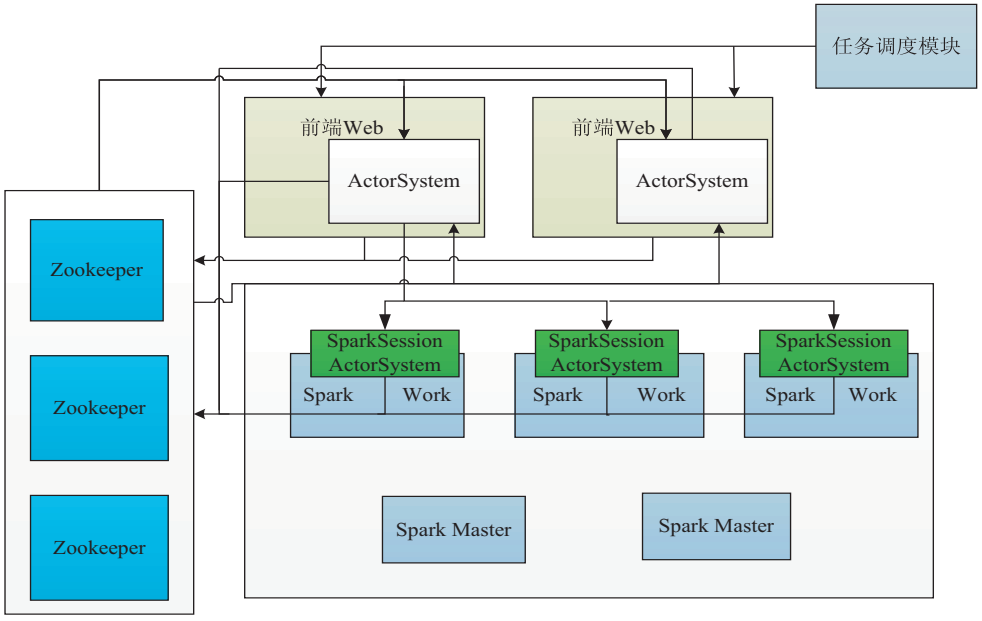

图 1 “天镜”中 Spark 的体系架构

外部调度任务模块的驱动下,将获取服务端的 Actor-System 地址,随机选择其中一个地址,提交 SprakSQL 任务,SparkSQL 任务提交成功后,会把任务和接收提交的 ActorSystem 信息注册到 Zookeeper,用于后续查看 SparkSQL 任务状态和取消任务。

2　气象全流程计算过程

全流程各环节监视信息通过接口网关进入后至高速缓冲通道,一路数据直接入库进行持久化,一路数据进行标准化构建和数据清洗形成中间结果表(见表 2)。

表 2　台站级资料预处理后中间结果表

CTS	SOD	台站号	资料 CTS 名称	资料 SOD 名称	原始时次	计算时次	规整时次	是否考核	在节目表标识	省	市	县	台站名称	应收数	应分发数	应入库数
A.0001.0032.R001	A.0010.0001.S001	50873	质控后地面气象要素资料(一体化新 Z 文件)	中国地面分钟降水资料	2018-06-24 02:00	2018-06-24 02:00	2018-06-24 02:00	1	1	黑龙江省	佳木斯市		佳木斯	1	1	1
A.0001.0032.R001	A.0012.0001.S001	50873	质控后地面气象要素资料(一体化新 Z 文件)	中国地面逐小时资料-要素存储	2018-06-24 02:00	2018-06-24 02:00	2018-06-24 02:00	1	1	黑龙江省	佳木斯市		佳木斯	1	1	1

根据总控配置表的业务频次(cron 表达式[0 0 0/1 * * ?]、统计规则[时次、时次截小时、时次截日、小时、日])信息计算出业务时次,并生成一个 sparkSQL 文件存入到 HDFS 中,提交给 spark 计算,计算考核指标的 SparkSQL 语句如下:

```
1. --考核应收
2. sum(coalesce(CO_CHECK_TD,0)) AS CO_CHECK_TD,
3. --考核及时收
4. sum(casewhen CHECK = '1' then coalesce(CO_INTIME_ACTUAL,0) ELSE 0 END) AS CO_CHECK_INTIME_ACTUAL,
5. --考核逾期收
6. sum(casewhen CHECK = '1' then coalesce(CO_LATETIME_ACTUAL,0) ELSE 0 END) AS CO_CHECK_LATETIME_ACTUAL,
7. --考核实收数
8. (sum(casewhen CHECK = '1' then coalesce(CO_LATETIME_ACTUAL,0) ELSE 0 END) + sum(case when CHECK = '1' then coalesce(CO_INTIME_ACTUAL,0) ELSE 0 END)) AS CO_CHECK_ACTUAL,
9. --考核缺收数
10. (sum(coalesce(CO_CHECK_TD,0)) - (sum(casewhen CHECK = '1' then coalesce(CO_LATETIME_ACTUAL,0) ELSE 0 END) + sum(case when CHECK = '1' then coalesce(CO_INTIME_ACTUAL,0) ELSE 0 END))) AS CO_CHECK_LOC,
11. --考核及时率
12. (sum(casewhen CHECK = '1' then coalesce(CO_INTIME_ACTUAL,0) ELSE 0 END)/sum(coalesce(CO_CHECK_TD,2147483646))) AS CO_CHECK_INTIME_RATE,
13. --考核到报率
14. ((sum(casewhen CHECK = '1' then coalesce(CO_LATETIME_ACTUAL,0) ELSE 0 END) + sum(case when CHECK = '1' then coalesce(CO_INTIME_ACTUAL,0) ELSE 0 END))/sum(coalesce(CO_CHECK_TD,2147483646))) AS CO_CHECK_RATE.
```

3　气象全流程计算优化

3.1　运行环境与系统架构

“天镜”系统部署在 36 台 IntelX86 物理服务器上(见图 2),其中 5 台服务器用于部署网关模块(gateway),数据预处理模块(standardizer)主要负责接收监视信息的收集和全流程中间结果(指标详情)的处理,3 台服务器用于部署消息中间件(kafka)集群,用于数据的高速缓存,避免因数据量过大导致后端数据库写入压力过大。18 台服务器部署分布式日志数据库用于对监视信息的原始指标,中间结果,最终计算指标进行存储。用于计算的 Spark 集群(版本 2.3.1)[11]部署在 5 台 CPU 24 核,内存 256G,3.2TSAS 磁盘,操作系统为 Centos7.3 服务器上。

3.2　优化技术原理

基于 Spark 计算引擎对气象全流程监视信息进行实时处理,作业调度任务每分钟执行一次,按照台站级气象资料(StationDiStaticJob)和文件级气象资料(FileDiStaticJob)分为两个计算任务。随着接入的气象资料种类越来越多,每分钟处理的监视信息也呈几何级增长,执行的 Spark 任务的耗时在 20 分钟以上,导致气象全流程监视界面中气象区域站资料无法及时显示。与此同时,运维人员发现 Spark 集群中有个别节点的负载特别高,这种情况是因为数据源单个 spark input read 数据量过大,或者单个 task 相对于其他 task spark input read 较大的情况,导致的读取数据源明显不均匀[12]。因此尽量使用可切割的文本存储,生成尽量多的 task 进行并行计算,可以从数据源避免倾斜,并从源头增大并行度[13]。通过观察 Spark 任务管理页面可以看到已完成的计算任务资源使用和耗时情况,如表 3 所示,正常计算任务需要分配计算资源 10 核,

内存 5 GB。

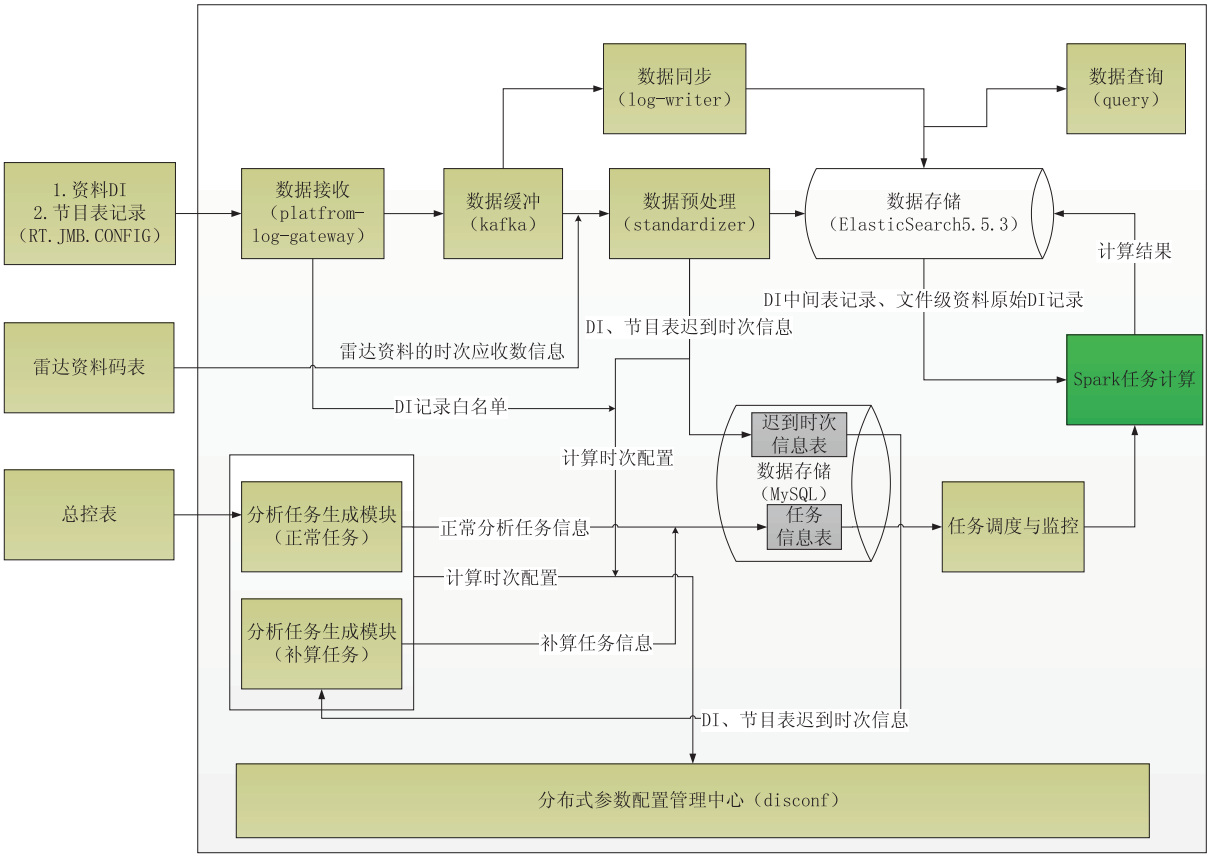

图 2 “天镜”-气象数据全流程系统架构

表 3 优化前 Spark 任务运行监视结果

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
app-20220112002105-201088	StationDiStatisJob_8b9019c87835400892e3deb03bd06938	10	5.0 GB	2022/01/12 00:21:05	root	FINISHED	24 min
app-20220112002210-201095	FileDiStatisJob_5e6f61cf4c69464bbaed9f6cd356bd7b	0	5.0 GB	2022/01/12 00:22:10	root	FINISHED	23 min
app-20220112002411-201105	StationDiStatisJob_01d5b974f69841c59f988a50e4794376	10	5.0 GB	2022/01/12 00:24:11	root	FINISHED	21 min
app-20220112002312-201100	StationDiStatisJob_dbfab47362854bee9c6ecf5e5f87653b	10	5.0 GB	2022/01/12 00:23:12	root	FINISHED	22 min
app-20220112002304-201097	FileDiStatisJob_8ba76bc94c4043cf8c0c2bc7ff63a9f8	0	5.0 GB	2022/01/12 00:23:04	root	FINISHED	22 min
app-20220112002205-201094	StationDiStatisJob_b6cdd06d73b046dfba5158ee438e9469	10	5.0 GB	2022/01/12 00:22:05	root	FINISHED	23 min

3.3 Spark 任务优化过程

进行 Spark 计算任务的优化的目的,是为了充分利用硬件本身的性能,最大限度地提升 Spark 中 Executor 的执行效率[14-17]。依据气象全流程监视界面资料展示情况,拆分为地面资料、海洋资料、高空资料、辐射资料、农业与生态资料、大气成分、雷达数据、卫星数据、气象服务产品、数值预报产品共 10 类资料,每类资料又分为考核资料和非考核资料。相较于优化前,虽然增加了 SparkSQL 模板的复杂度,但是提升了气象考核资料的计算效率,该文以传输环节考核资料为例,新增的 SparkSQL 模板如下:

1. base. sql. co. checks = sum(casewhen CHECK = '1' then coalesce(CO_TD,0) ELSE 0 END) AS CO_CHECK_TD, sum(case when CHECK = '1' then coalesce(CO_INTIME_ACTUAL,

0) ELSE 0 END) AS CO_CHECK_INTIME_ACTUAL, sum(case when CHECK = '1' then coalesce(CO_LATETIME_ACTUAL,0) ELSE 0 END) AS CO_CHECK_LATETIME_ACTUAL,(sum(case when CHECK = '1' then coalesce(CO_LATETIME_ACTUAL,0) ELSE 0 END) + sum(case when CHECK = '1' then coalesce(CO_INTIME_ACTUAL,0) ELSE 0 END)) AS CO_CHECK_ACTUAL,(sum(case when CHECK = '1' then coalesce(CO_TD,0) ELSE 0 END) - (sum(case when CHECK = '1' then coalesce(CO_LATETIME_ACTUAL,0) ELSE 0 END) + sum(case when CHECK = '1' then coalesce(CO_INTIME_ACTUAL,0) ELSE 0 END))) AS CO_CHECK_LOC,(sum(case when CHECK = '1' then coalesce(CO_INTIME_ACTUAL,0) ELSE 0 END)/coalesce(sum(case when CHECK = '1' then coalesce(CO_TD,0) ELSE 0 END), 2147483646)) AS CO_CHECK_INTIME_RATE,((sum(case when CHECK = '1' then coalesce(CO_LATETIME_ACTUAL,0) ELSE 0 END) + sum(case when CHECK = '1' then coalesce(CO_INTIME_ACTUAL,0) ELSE 0 END))/coalesce(sum(case when CHECK = '1' then coalesce(CO_TD,0) ELSE 0 END), 2147483646)) AS CO_CHECK_RATE

在向 Spark 进行任务提交时,客户端处理程序需要将气象资料按照上述分类进行拆解,核心代码如下:

```
1. …
2. for (TabMcmConfig config : tabOminCmCcSubsystem Allconfigs) {
3. // 1、资料编码
4. String ctsCode = config. getcCtsCode();
5. String sodCode = config. getcSodCode();
6. //资料大类
7. String dataClass = config. getDataClass();
8. String ctsSodCode = ctsCode. concat(":"). concat(sodCode);
9. //文件级还是站点级
10. String dataSourceType = config. getcDataSource();
11. if ("1". equals(dataSourceType)) {
12. if(! fileDiComputeEnabled) {
13. continue;
14. }
15. dataSourceType = "file";
16. }else {
17. dataSourceType = "station";
18. }
19. …
```

此段代码通过获取总控配置后对每类气象资料进行分类,分类后生成的计算任务与生成的 SparkSQL 模板匹配,从而完成计算任务拆解,单个 SparkSQL 只计算一类考核资料或者一类非考核资料。

3.4　全流程计算结果正确性验证及性能优化分析

该文采用自动化测试的方法,由于对程序代码结构进行了修改和微调,因此需要对优化后的全流程指标计算结果正确性进行验证。正确性可以根据监视页面中资料的统计指标和系统告警进行判断,如图 3 所示,可以通过查看 Spark 作业任务日志进行验证,如表 4 所示。该文展示的全流程监视界面与优化前资料监视统计指标计算结果一致,并且中国地面分钟降水数据在一级界面中可以显示正常。优化后单个计算任务的计算时间控制 2 分钟以内。

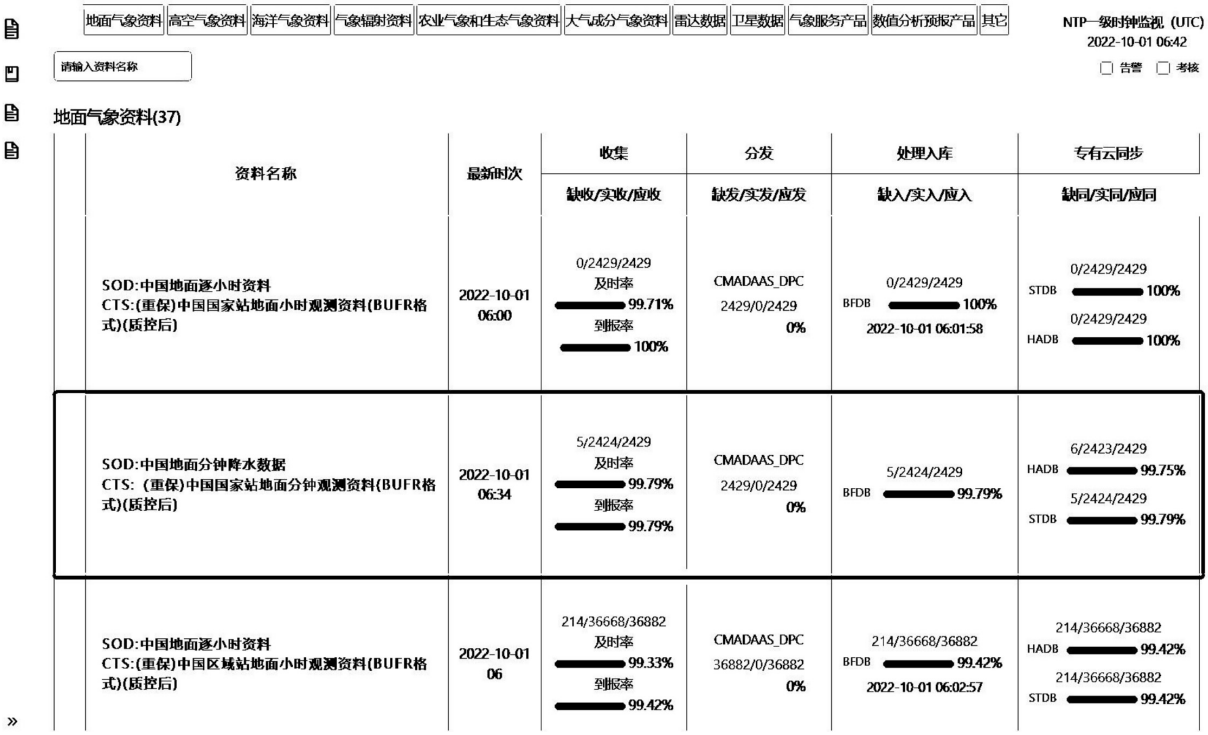

图 3　气象综合业务实时监控系统—“天镜”全流程监视界面

表 4 优化后 Spark 任务运行监视结果

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
app-20220125152853-392504	StationDiStatisJob_d0d6716d835a4244b4e261122f12b227 0	0	3.0 GB	2022/1/25 15:28	root	FINISHED	1.8 min
app-20220125152902-392524	FileDiStatisJob_138ae6a1e1e64225b016fde1758ecf4c	10	3.0 GB	2022/1/25 15:29	root	FINISHED	1.6 min
app-20220125152853-392511	StationDiStatisJob_e2f7aa94680942c8be1d4533da836019	10	3.0 GB	2022/1/25 15:28	root	FINISHED	1.7 min
app-20220125152902-392523	StationDiStatisJob_4e52dae03dea4e73bb3b02b424603f1a	10	3.0 GB	2022/1/25 15:29	root	FINISHED	1.6 min
app-20220125152853-392509	FileDiStatisJob_9b352476f25c463ca900f5645786611c	10	3.0 GB	2022/1/25 15:28	root	FINISHED	1.7 min
app-20220125152901-392521	StationDiStatisJob_069842c823c24f1b8c4c2128d02f4cff 0	0	3.0 GB	2022/1/25 15:29	root	FINISHED	1.5 min

4 结束语

通过拆分计算任务,生成尽可能多的 task 增加 Spark 计算并行度,成功将气象全流程计算框架优化并业务运行,如表 5 所示,获得了 10 倍的加速效果,提高了程序的运行效率。但是“天镜”系统在处理大数据计算时还是有瓶颈,原因是地面区域站气象资料会产生大量重复数据,要能够高效处理海量的监视数据,除了对计算任务拆分,还需要对计算任务设置优先级,针对核心资料优先分配计算资源计算,这就需要业务人员对资料的监视等级进行配置,同时要熟悉 Spark 资源分配机制,在此基础上来做系统优化,能够较好地提升优化效果。

表 5 “天镜”全流程 Spark 计算任务优化前后运行时间

	内存使用	CPU 核数	任务文件大小	运行耗时/min
优化前(单个计算任务)	5 GB	10	10 MB	22
优化后(单个计算任务)	3 GB	10	64 KB	1.7

参考文献:

[1] 孙 超,肖文明,陈永涛,等. 气象综合业务实时监控系统的设计[J]. 气象科技进展,2018,8(1):153-157.

[2] 曾 乐,孙 超,张来恩,等. 基于大数据技术的气象业务监视数据采集处理[J]. 计算机仿真,2021,38(7):181-188.

[3] 冯兴杰,王文超. Hadoop 与 Spark 应用场景研究[J]. 计算机应用研究,2018,35(9):2561-2566.

[4] 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]. 气象应用学报,2015,26(4):500-512.

[5] BOUCHARD R,KERN K,HANKIN S,et al. Observing system monitoring center[C]//IUGG XXV general assembly. Melbourne:The International Union of Geodesy and Geophysics,2011.

[6] DAHOUI M, ISAKSEN L, BORMANN N. Monitoring for conventional observation systems at ECMWF[C]//Observation monitoring meeting,Bruxelles:EUMETNET,2013.

[7] KUMAR K V,BALLISH B,STOUDT J. Real time data monitoring at NCEP[C]//22nd international conference on interactive information processing systems for meteorology,oceanography,and hydrology. Atlanta:American Meteorological Society,2006.

[8] 沈文海,何文春,孙 超. 从两个典型应用看气象信息业务的数据工作[J]. 中国信息化,2017(9):70-76.

[9] 吴黎兵,邱 鑫,叶璐瑶,等. 基于 Hadoop 的 SQL 查询引擎性能研究[J]. 华中师范大学学报:自然科学版,2016,50(2):149-155.

[10] 廖湖声,黄珊珊,徐俊刚,等. Spark 性能优化技术研究综述[J]. 计算机科学,2018,45(7):7-15.

[11] Marcelo Vanzin. ApacheSpark[EB/OL]. 2018. https://spark.apache.org/docs/2.3.1/(2018).

[12] 陈侨安,李 峰,曹 越,等. 基于运行数据分析的 Spark 任务参数优化[J]. 计算机工程与科学,2016,38(1):11-19.

[13] ZAHARIA M,CHOWDHURY M,FRANKLIN M J,et al. Spark:cluster computing with working sets[C]//Usenix conference on hot topics in cloud computing. Boston:USENIX Association,2010:10-17.

[14] 徐 计,王国胤,于 洪. 基于粒计算的大数据处理[J]. 计算机学报,2015,38(8):1497-1517.

[15] KARAN H,KONWINSKI A,WENDELL P,et al. Spark 快速大数据分析[M]. 北京:人民邮电出版社,2015.

[16] KAUR R,CHADHA D R. Performance comparison between Hive,Spark-sql&Flink-sql through IVR data analysis[J]. IOSR Journal of Computer Engineering,2017,19(3):6-11.

[17] LI X,ZHOU W. Performance comparison of Hive,Impala and Spark SQL[C]//2015 7th international conference on intelligent human-machine systems and cybernetics. Hangzhou:IEEE,2015:418-423.