



游戏大数据平台 workflow 引擎研究与实践

陈杰¹, 武娟², 苏洋¹, 唐勇¹

(1. 炫彩互动网络科技有限公司, 江苏 南京 210029;

2. 中国电信股份有限公司广州研究院, 广东 广州 510630)

摘要:随着业务人员对数据分析的需求不断增加, 基于开源 Hadoop 的游戏大数据平台运行着越来越多相互依赖的作业任务。为了更加高效地完成作业管理和调度, 升级了 Zeus 构建 workflow 引擎, 优化 Zeus 任务启动过程, 实现了多线程启动任务, 缩短了启动时间。通过扩展 Zeus 系统, 使其支持 Oracle 存储过程。同时, 利用管道技术, 实现了异构数据源的无缝高效共享和传输。

关键词:大数据作业; 工作流; 开源

中图分类号: TP311

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2016080

Research and practice on working flow engine of big data platform for gaming service

CHEN Jie¹, WU Juan², SU Yang¹, TANG Yong¹

1. Dazzle Interactive Network Technologies, Nanjing 210029, China

2. Guangzhou Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China

Abstract: The big data platform for gaming service is based on Hadoop open source technology. With the demand growth of the service operation staff, there are more and more jobs which depend on each other running on the big data platform. A new working flow engine was developed on Zeus to manage and schedule jobs efficiently. In order to shorten the start time, Zeus task starting process was optimized and multi-threaded start task was achieved. Zeus was also improved to support the Oracle stored procedure. Pipeline technology was used to seamless and efficient sharing and transmission of heterogeneous data sources.

Key words: big data job, work flow, open source

1 引言

基于 Hadoop 的开源大数据系统经历了多个版本的开发, 技术更加成熟、性能更加稳定, 众多企业和团队基于 Hadoop 系统, 辅以 Hive、HBase 等工具搭建满足自身业务数据分析需求的大数据平台。为了满足作业调度和执行的批量管理、自动处理、状态监控等需求, 业务复杂的大数据

分析平台需要一套 workflow 引擎工具, 将作业之间的依赖关系、执行顺序、状态管理等有效衔接, 从而提高作业任务的生产、运行、管理效率。

本文介绍了游戏大数据平台的现状和主流大数据 workflow 引擎, 重点研究游戏大数据平台引入开源的宙斯 (Zeus) workflow 引擎系统的过程和效果, 并针对自身业务需求在其基础上实现二次功能的开发和增强。

2 游戏大数据平台概述

目前游戏平台上支持手机游戏、智能电视游戏、PC 游戏及 HTML5 游戏等多种业务,日志数据源达数十种,按其存储类型可分为 Oracle、MySQL、SQLSever、FTP、Local FileSystem 等。为处理种类繁多的业务,游戏大数据平台的单集群节点数约为 50 台服务器,日新增数据量约为 0.5 TB,日均 Hadoop job(作业)数为 300 余个,日均计算量为 100 亿余行。

游戏大数据平台^[1]架构以 HDFS 为基础,其上部署 Hadoop、HBase、Hive、Storm 等子系统,平台结构如图 1 所示,各部分作用说明如下。

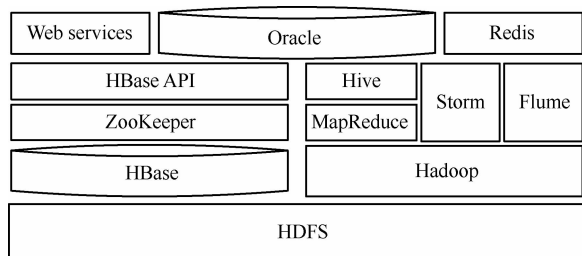


图 1 游戏大数据平台架构

- HDFS: 构建分布式文件系统,所有的数据都存储在这里,并对数据进行了分块。
- HBase: 针对特殊场景下对处理速度与处理效率的要求,实现快速数据检索。

- Storm: 降低实时处理的复杂性,并且支持多种语言,提升开发效率。
- Flume: 主要用于海量日志收集,采用分布式模式,对海量日志进行采集传输。
- ZooKeeper: 提供一致性服务,实现各节点高效、可靠的协同工作。
- MapReduce: 大数据分治处理方法,是一种编程模型。
- Hive: 基于 Hadoop 的一种结构化数据仓库,提供了简单的类 SQL 查询功能。

游戏大数据平台数据流转的关系如图 2 所示,游戏原始的业务数据通过文件系统进入大数据平台存储和处理,经过简单的清洗和分类转变为经营分析数据存放于数据仓库,然后由大数据平台运行成百上千个 Hadoop 作业,分析计算出所需结果后提交给各个业务系统接口,可以看出,大数据平台的数据工作流程是有序的,可以由程序化软件控制自动完成整个工作流程。

3 大数据平台工作流引擎

3.1 引擎选型

早期游戏平台采用的大数据任务调度方式为 Linux 自带的 crontab 方式,即使用 Linux 的定时执行指令,每分钟检查是否有预定的 cron job 需要执行。为规范化执行作业任务,统一将一个个 Hadoop job 拆分为数据准备、逻辑计算、结果输出 3 个步骤,然后封装成 Linux shell 脚本,配置

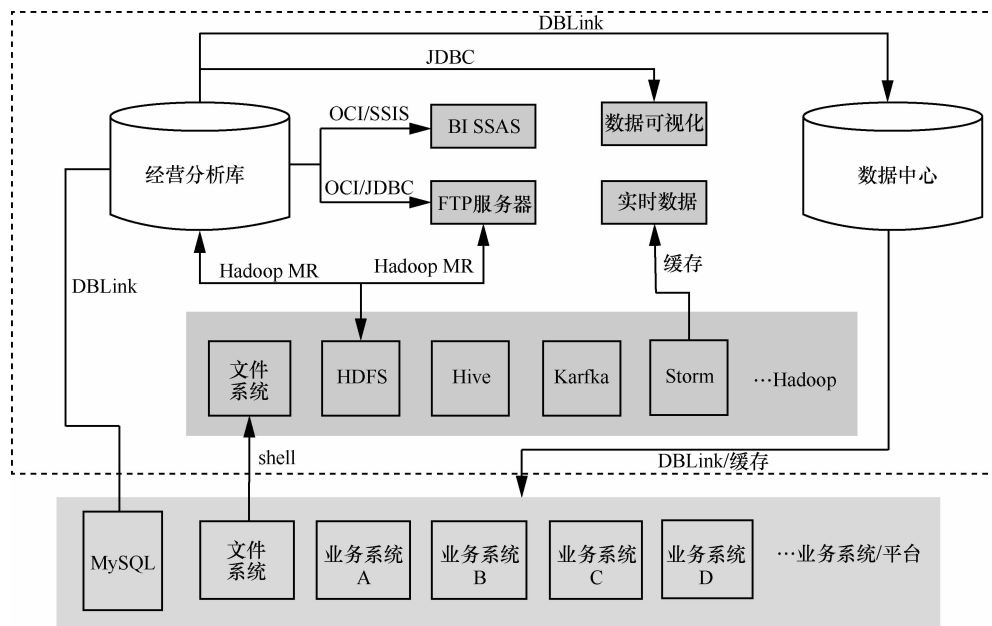


图 2 游戏大数据平台数据流转关系



至 cron job 中定时执行。这种调度方式的好处就是简单、稳定,在业务需求相对简单、作业任务较少的应用场景,此方法完全可以满足需求。

随着定时任务数量的增加,任务之间的关系越来越复杂,crontab 方法存在的不足也日益突出:首先是不易管理,所有任务均部署在配置文件中,增、删、改、查均不直观;二是安全性较差,团队成员均可独立修改 crontab 配置文件,容易相互影响;三是任务之间的依赖关系只能实现简单的定时执行,配置容易出错;四是任务监控、错误处理没有统一的框架级解决方案。这些问题使得游戏大数据平台迫切需要一套完整的用于管理和调度作业任务的工作流引擎系统。

Apache Oozie^[2]是用于 Hadoop 平台的开源工作流调度引擎。该框架使用 XML 定义工作流,由 Oozie 协调器管理作业之间的相互依赖关系,可以使用预定的时间或数据可用性来触发 Oozie。Oozie 支持多种 action,即工作流节点的执行操作,例如支持 Hadoop MapReduce (MR)、pig、Hive、shell 和 Java 等类型的作业,所有的 action 以有向无环图 (direct acyclic graph,DAG) 的模式运行。在 action 的运行步骤上是有方向的,只能等上一个 action 运行完成后才能运行下一个 action。

Azkaban^[3]是 LinkedIn 公司提出的另一个优秀的开源任务调度系统,操作比 Oozie 要简单^[4],它使用属性 (property) 文件定义工作流。Azkaban 以预定义的一个或多个可存在依赖关系的作业组成的工作流 (flow) 为执行单元进行定时调度。Azkaban 支持 pig、Java、shell、MapReduce 类型的作业,从版本 2.1 开始,Azkaban 也支持 Hive 类型的作业。

Zeus^[5]是由阿里巴巴集团开发的一个分布式设计水平扩展的开源作业工作流引擎,以可视化界面的方式管理与调度 Hadoop 上的任务,Zeus 2 支持 Hadoop 2.0 以上版本的系统。

如图 3 所示,Zeus 系统主要包含界面 (Web UI)、作业调度 (job scheduler) 和作业执行 (job worker) 3 个角色,其中 Web UI 提供可视化页面访问,作业调度负责作业任务的调度、恢复等,作业执行负责完成调度指派的任务。

Zeus 支持任务之间复杂依赖关系的设置,依赖任务零时延触发,有完备的报警系统,能够与其他系统通过 ZooKeeper 进行联动。Zeus 支持 Hadoop MapReduce、Hive、shell 等作业类型,相对于 Oozie 和 Azkaban,Zeus 支持的作

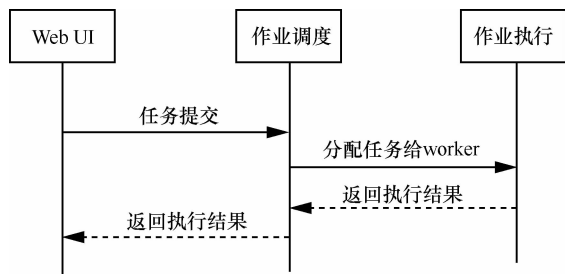


图 3 Zeus 任务调度流程

业类型较少,目前尚不支持 pig、flume、sqoop 等作业类型,但对 Hive 的支持比较完备,例如支持 Hive 元数据的可视化查询与数据预览,有完整的文档管理,因此对于以 Hive 为主要分析工具的游戏大数据平台来说,选择 Zeus 作为工作流引擎更加贴合国内业务平台的需求、更加易用,而且通过二次开发可以进一步优化 Zeus 系统。

3.2 对 Zeus 系统的改进

游戏大数据平台引入 Zeus 系统后,发现 Zeus 系统在执行任务调度时,任务启动所消耗的时间较长,在作业数较多的情况下,启动任务的时间比较长,有必要进一步改进优化任务启动过程。

另外,原生 Zeus 系统不支持 Oracle 存储过程,由于大数据平台目前部分报表生成工作仍使用 Oracle 存储过程执行,因此,需要对 Zeus 进行一定的扩展,使其能够支持 Oracle 存储过程。

最后,Zeus 系统在异构数据源之间的数据同步没有便捷的实现,运行的 Hive 查询结果也不方便导出成文件,这方面可通过第三方工具实现异构数据的同步。

3.2.1 优化任务启动方式

在 Zeus 任务调度过程中,根据业务拆分的作业比较小,但启动非常频繁,每个作业的执行时间短,通过执行的日志发现,有时会出现一个作业的启动时间很长,需要几十秒。由于业务上很看重增量的速度,几十秒的等待对业务系统来说不可接受。

游戏大数据平台使用的是 Hadoop 2.6.0 版本,仔细分析 TaskTracker 的日志发现,Child 子进程启动过程中,存在过长时间等待的问题。从 Zeus 源码可以看到,Child 子进程启动是在一个线程中串行完成,启动过程包括 DistributedCache 文件的获取。由于 Hadoop 集群同时可能有各种各样的任务提交,当某个任务的启动时间长时,会影响下一个任务的启动,严重时甚至会影响 TaskTracker 发送心跳信息。

因此,需要通过修改 TaskTracker 的工作方式,彻底优化启动时间。最简单的办法是让每个任务各自通过一个任务线程来启动,以避免各个任务之间的竞争,从根本上解决这个问题。

```
class StartNewTask extends Thread {
    默认将线程任务成员变量置空

    public StartNewTask(TaskInProgress tip) {
        构造函数,根据传参初始化线程任务对象
    }

    public void run() {
        try {
            调用 StartNewTask 方法,执行线程
        } catch (InterruptedException e) {
            // 返回异常原因
        } catch (Throwable th) {
            // 错误处理
        }
    }
}
```

以上方法需要在 TaskLauncher.run() 中增加 new StartNewTask(tip).start(), 这带入了新的问题,TaskTracker 中关于正在运行的任务有两种锁,一个是 runningJobs,用来锁住对 Map<JobID, RunningJob> runningJobs 的访问,这个 HashMap 里存在的是正在运行的所有任务;另一个是 rJob,用来存放单个任务的信息。原版的代码中,是先获得 runningJobs 锁,再获得 rJob 锁,最后再释放 runningJobs 锁,代码如下:

```
private RunningJob addTaskToJob(JobID jobId,
TaskInProgress tip) {
    同步代码块(runningJobs) {
        初始化 RunningJob 对象 rJob 为空值;
        if (runningJobs 中找不到 jobId 对象) {
            // 新建 RunningJob
        } else {
            // 获取 RunningJob 的 JobId
        }
    }
    同步代码块 (rJob) {
        将 tip 对象增加到 rJob 对象的 task 成员中
```

```
}
    返回 rJob 对象;
}
```

由于 setupCache 的操作是在 rJob 锁里完成,会间接导致 runningJobs 一直等待 rJob,这时即使任务启动的是多线程任务,也会由于别的任务在下载 DistributedCache,长期占用 rJob 的锁,导致其他线程的 runningJobs 等待 rJob,使得当前的任务启动后再无法获得 runningJobs。

有一种比较直接的方法,即强行把 runningJobs 和 rJob 分开,只有在更新 runningJobs 时,才需要获得锁,且锁住的范围不包括 rJob。这样做的改动范围大,而且容易出错,造成死锁。

参考 MAPREDUCE-2364 补丁包^[9]的解决办法,游戏大数据平台采取的是当下载 DistributedCache 时,去掉 rJob 锁,从而使 runningJobs 和 rJob 锁住的操作中,没有长时间的任务。通过增加以下两个变量:volatile boolean localized 和 boolean localizing(这两个变量控制是否需要下载 DistributedCache),从而去掉 rJob 锁。最终,改进的 Hadoop 版本,包括实现了多线程启动任务和 MAPREDUCE-2364 的 patch。

3.2.2 支持 Oracle 存储过程

由于历史原因,游戏业务平台仍然有一部分报表计算任务使用 Oracle 存储过程完成,但 Oracle 存储过程的调度仍然独立于 Hadoop、Hive、shell 任务。为了使 Oracle 存储过程任务与 Hadoop、Hive、shell 任务建立依赖关系,需要对 Zeus 系统进行改造升级。在 Zeus 宿主机上安装 Oracle 客户端,并修改 Zeus UI 使其支持 *.sql 脚本,然后利用 shell 脚本调用 Zeus 系统中对应的 *.sql 文件,最后将其加载至 Oracle 中执行,并返回执行状态码。

为使存储过程的开发更加简单,还需要将 Java 和 SQL 代码混合,定制解析器,利用解析器把混合在里面的 SQL 语句变成“字符串”加入 Java 代码中,生成 Java 类源码。然后由 Java 编译器把第一步生成的 Java 类源码编译为 class 文件。使用 Java 类加载器加载第二步生成的 class 文件并运行,运行后将生成纯 SQL 语句。最后再调用 Oracle 专用存储过程将 SQL 语句加载至 Oracle 中执行。

3.2.3 异构数据源同步

为了满足非结构化数据库 Hive 和结构化数据库 Oracle 之间异构数据的快速同步要求,弥补 Zeus 系统缺乏异构数据同步功能的缺憾,游戏大数据平台还引入



sqlldr2 数据导出工具, 基于文本文件格式+管道技术, 实现异构数据源的无缝高效共享和传输。

文本文件作为基础的文件格式, 数据库软件普遍支持, 并且能够快速加载; 管道技术, 实现基于不同异构数据库内存或缓存间的无缝对接。其实现原理为, 源数据库计算输出为文本文件, 通过内存实现直接导入目的数据库, 计算结束后自动清除缓存, 实现异构数据源间的数据无缝对接和交互。

4 改进效果

改进后的 Zeus 系统在调度作业时, 任务之间没有等待时间, 最大限度地利用了集群资源。在同一实验环境下 (任务部署节点: 8 核 64 GB、CentOS 6.4; Hadoop 集群主备节点 2 台: 8 核 64 GB、CentOS 6.4, 数据和任务节点公用 13 台: 8 核 32 GB、CentOS 6.4; 100 Mbit/s 网卡) 使用 Zeus 调度与 crontab 调度方式总执行时间对比的结果如图 4 所示 (选取 2015 年 3 月 1-5 日数据)。

在优化之前, 原有 Zeus 系统调度作业平均耗时为 43 min, 使用优化后的 Zeus 调度方法后, 作业平均耗时

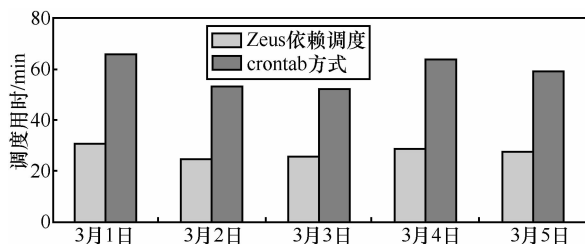


图4 Zeus 和 crontab 调度用时对比

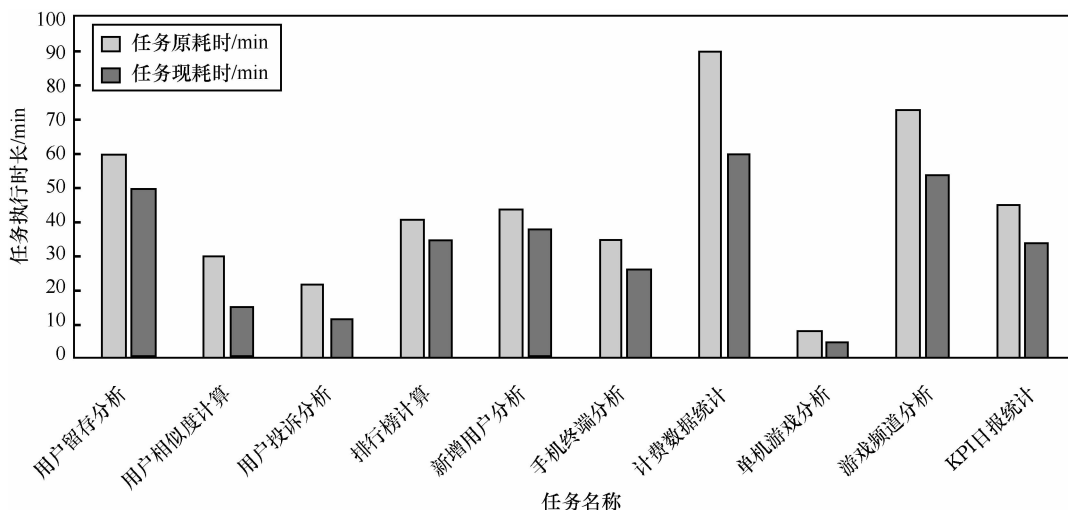


图5 Zeus 优化前后调度用时对比

为 31 min, 提升 27.6%, 示例任务的前后效果对比如图 5 所示。

通过改造, 实现在 Zeus 系统内支持 Oracle 存储过程, 使得原有 Oracle 计算任务与大数据平台计算任务无缝衔接, 同时也为 Zeus 支持其他关系型数据库存储过程提供了通用的解决方案。

相对于 Hadoop 生态中的 Sqoop 技术、Hive 自身的用户定义函数 (UDF) 接口技术等异构数据同步方式, 通过管道技术实现的异构数据源同步方法, 节约了数据格式转换、存储和对接同步时间, 效率提升近 70%。

5 结束语

从实际使用效果来看, 游戏大数据平台引入 workflow 引擎工具后, 极大提升了 Hadoop 作业的管理和调度效率, 通过 Zeus 的权限划分功能, 不仅可对作业的编制、修改、删除进行有序管理, 更可在流程和制度上规范业务人员和平台运维人员的操作, 使其协同度更高。

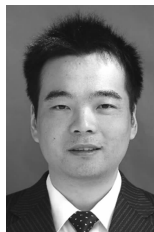
围绕 Hadoop 的开源生态环境, 存在很多诸如 Oozie、Azkaban 等优秀的开源 workflow 引擎, 在国外公司得到广泛应用, 相比之下 Zeus 由国人开发, Zeus 系统的功能强大、使用便捷, 界面交互习惯和流程设计更符合中国公司的需求, 二次开发更容易上手。

参考文献:

- [1] 苏洋, 刘晓军, 唐勇, 等. 游戏大数据平台研究与实践 [J]. 电信科学, 2014, 30(10): 21-26.

- SU Y, LIU X J, TANG Y, et al. Research and practice of big data platform of games [J]. Telecommunications Science, 2014, 30(10): 21-26.
- [2] Apache Oozie workflow scheduler for Hadoop[EB/OL]. [2015-05-30]. <http://oozie.apache.org/>.
- [3] Open-source workflow manager [EB/OL]. [2014-02-19]. <http://azkaban.github.io/>.
- [4] Workflow engines for Hadoop, Joe Crobak[EB/OL]. [2012-07-05]. <http://www.crobak.org/2012/07/workflow-engines-for-hadoop/>.
- [5] 淘宝 Hadoop 作业平台宙斯 Zeus[EB/OL]. [2015-10-16]. <https://github.com/alibaba/zeus>.
- Taobao Hadoop worker platform Zeus [EB/OL]. [2015-10-16]. <https://github.com/alibaba/zeus>.
- [6] MAPREDUCE-2364 [EB/OL]. [2011-03-07]. <https://issues.apache.org/jira/browse/MAPREDUCE-2364>.
- [7] WHITE T. Hadoop 权威指南(第3版)[M]. 曾大聃,周傲英,译. 北京:清华大学出版社, 2015: 125-130.
- WHITE T. Hadoop: The Definite Guide [M]. Translated by ZENG D R, ZHOU A Y. Beijing: Tsinghua University Press, 2015:125-130.

[作者简介]



陈杰(1974-),男,炫彩互动网络科技有限公司高级工程师,主要研究方向为互联网大型业务平台技术架构、大数据技术框架、通信技术和移动互联网等。

武娟(1977-),女,中国电信股份有限公司广州研究院高级工程师,主要研究方向为云计算、大数据、数据网络、移动互联网技术等。

苏洋(1988-),男,炫彩互动网络科技有限公司工程师,主要研究方向为大数据、移动互联网等。

唐勇(1973-),男,炫彩互动网络科技有限公司高级工程师,主要研究方向为大数据、手机游戏运营平台等。