

基于 Spark 的电信运营商数据可视化应用研究

张桂花^{1*} 高攀¹

四川大学锦城学院 计算机与软件学院 四川 成都 611731

DOI: 10.18686/jsjxt.v1i3.1263

【摘要】数据的爆炸式增长催生了大数据技术的快速发展,大数据技术的深度应用在各企业得以体现。其中,Spark 大数据计算框架的设计遵循“一个软件栈满足不同应用场景”的理念,逐渐形成了一套完整的生态系统。电信运营商在线服务需对大量数据进行快速、高效、准确的处理,鉴于此,本文将 Spark 部署在 Hadoop 的 YARN 模式下,并采用 HDFS 进行数据存储,通过 Spark 及其提供的 SQL 查询能力进行数据处理,通过 ECharts 前端可视化组件进行数据展示,提供更直观和可靠的数据分析。

【关键词】Spark 大数据可视化 ECharts 电信运营商

当今社会已经处于数据爆炸式增长时代,如何从巨大数据量中挖掘出有价值的信息是数据处理的关键所在,也就是大数据处理的有效性。大数据的特征可总结为 4V,即数据规模、处理速度快、结构复杂以及数据真实有效性^[1]。大数据的这些特点使得无法在一定时间内用常规软件工具对其内容进行抓取、管理和处理^[2]。现在,Hadoop 技术的成熟和 Spark 技术新秀的出现给大数据技术应用带来了保障。

1 spark 介绍

Spark 是一个快速、分布式、可扩展、容错的大数据计算框架^[3],它主要基于内存进行计算,具有低延迟的复杂分析,采用 SCALA 进行代码编写,简洁高效。其框架结构如图 1 所示:

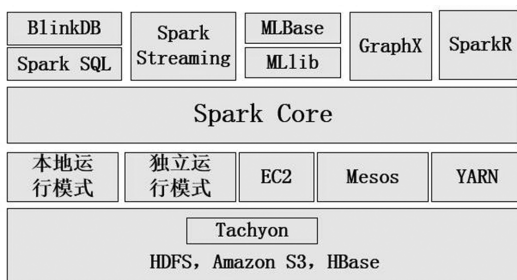


图 1 park 框架结构图

Fig.1 Spark framework diagram

Spark 的设计遵循“一个软件栈满足不同应用场景”的理念,逐渐形成了一套完整的生态系统,既能够提供内存计算框架,也可以支持 SQL 即席查询、实时流式计算、机器学习和图计算等^[4]。Spark 可以部署在资源管理器 YARN 之上,提供一站式的大数据解决方案。因此,Spark 所提供的生态系统足以应对上述三种场景,即同时支持批处理、交互式查询

和流数据处理。Spark 运行基本流程如图 2 所示,

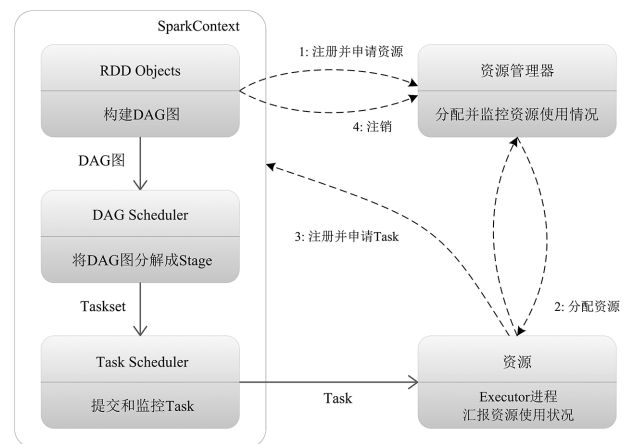


图 2 Spark 运行基本流程

Fig.2 The flow chart of the Spark

从图中可以看到 Spark 运行架构具有以下特点:

- (1)每个 Application 都有自己专属的 Executor 进程,并且该进程在 Application 运行期间一直驻留。Executor 进程以多线程的方式运行 Task
- (2)Spark 运行过程与资源管理器无关,只要能获取 Executor 进程并保持通信即可
- (3)Task 采用了数据本地性和推测执行等优化机制

由此,可以看到 Spark 有以下两个优点:一是利用多线程来执行具体的任务,减少任务的启动开销;二是 Executor 中有一个 BlockManager 存储模块,会将内存和磁盘共同作为存储设备,有效减少 IO 开销。

Spark 如今已吸引了国内外各大公司的注意,如腾讯、淘宝、百度、亚马逊等公司均不同程度地使用

了 Spark 来构建大数据分析应用,并应用到实际的生产环境中。

2 电信运营商数据可视化

2.1 背景

随着通信行业的普及,以及人们对网络的需求越来越大,因此运营商的一些在线服务需求也越来越大。虽然目前线下渠道占据主要地位,线上渠道作为业务办理的支撑和辅助。运营商主打便捷性卖点,向客户宣传推广电子渠道。对于客户体验来说,电子渠道提供了一个足不出户办理业务的便捷方式,对于运营商来说,电子渠道低成本分流了实体渠道的业务压力,将线下渠道的人力资源从低价值的业务办理中释放出来。对服务商给出的历史数据进行分析,以图形化的方式直观展现各服务区域服务指标达标情况,以及展现各服务区域的累积服务量。通过数据分析,对各类用户做多维画像,降低用户的投诉和流失^[5]。

2.2 技术选型

结合电信运营商对数据处理需求,为了更高效更充分地处理数据,拟采用大数据相关技术进行整合,分别是 Hadoop(分布式系统的基础架构)、Flume(一个高可用的,高可靠的,分布式的海量日志采集、聚合和传输的系统)、Spark(基于内存计算的开源的集群计算系统,为速度和通用目标所设计的集群计算平台)、ETL(数据抽取(Extract)、清洗(Cleaning)、转换(Transform)、装载(Load)的过程)、数据可视化(用 SpringMVC+Ajax+Echarts+MySQL 技术,把统计出的数据展示出来)

2.3 可视化实现组件

ECharts^[6],一个使用 JavaScript 实现的开源可视化库,可以流畅的运行在 PC 和移动设备上,兼容当前绝大部分浏览器(IE8/9/10/11,Chrome,Firefox,Safari 等),底层依赖轻量级的矢量图形库 ZRender,提供直观,交互丰富,可高度个性化定制的数据可视化图表。ECharts 提供了常规的折线图、柱状图、散点图、饼图、K 线图,用于统计的盒形图,

用于地理数据可视化的地图、热力图、线图,用于关系数据可视化的关系图、treemap、旭日图,多维数据可视化的平行坐标,还有用于 BI 的漏斗图,仪表盘,并且支持图与图之间的混搭。

电信运营商服务数据巨大,关系复杂,采用可视化组件进行展示是最佳选择。

2.4 效果展示

根据电信运营商具体业务需求,针对其数据量大、实时性要求高、数据之间关系复杂等特点,主要采用 Hadoop 的 HDFS 进行数据存储,Spark 进行数据处理,ECharts 进行可视化展示,最终实现效果图如图 3 所示:



图 3 效果图

Fig. 3 The effect chart

3 总结

大数据应用的深入和普及得益于大数据处理技术的快速发展,目前市场上比较成熟的就是 Hadoop 大数据处理框架,随着其应用的深入,逐渐暴露一些低性能问题。相应的解决方案应运而生,Spark 计算框架就是这样产生的,它具备很多优秀的特性,对大数据的处理有非常好的性能。本文研究的是电信行业运营商在线服务问题,每天所处理的数据量非常巨大,如何高效的对数据进行处理,从而抽取有价值数据,是一个非常重要的问题,那么,技术选型就非常重要。本文采用 Spark 技术对数据进行高效分析和处理,然后分析结果通过 Echarts 组件进行可视化展示,能更直观和有效地分析数据并进行决策,实际应用效果良好,希望在下一步工作中进一步提高其处理速度和准确度。

【参考文献】

- [1]Tom White. 华东师范大学数据科学与工程学院译. Hadoop 权威指南(第 3 版)(修订版). 北京:清华大学出版社,2015.
- [2]涂兰敬. 大数据与海量数据的区别[J]. 网络与信息. 2011. 12(1):22-27. PHam
- [3]肖芳,张良均主编 著. Spark 大数据技术与应用[M]. 北京:人民邮电出版社,2017.
- [4]林子雨. Spark 编程基础(Scala 版)[M]. 北京:人民邮电出版社,2018.
- [5]毛国君. 数据挖掘原理与算法[M]. 北京:清华大学出版社有限公司,2005. PH
- [6]<https://echarts.baidu.com/>