文章编号:1007-1423(2017)09-0120-04 DOI:10.3969/j.issn.1007-1423.2017.09.027

Spark 的图计算框架:GraphX

孙海

(四川大学计算机学院,成都 610065)

摘要:

Spark 是 UC Berkeley AMP Lab 所开源的类 Hadoop MapReduce 的通用并行框架,是专为大规模数据处理而设计的快速通用的计算引擎,在如今的大数据环境下,Spark 所发挥的作用正越来越大。介绍 Spark 的图计算框架 GraphX。

关键词:

Spark;并行;大数据;GraphX

1 Spark 生态圈介绍

Spark 生态圈即 BDAS(伯克利数据分析栈)包含了 Spark Core、Spark Streaming 、Spark SQL、MLLib 和 GraphX 等组件,其中 Spark Core 提供内存计算、Spark Streaming 主要处理实时应用、Spark SQL 提供及时查询、MLLib 的机器学习和 GraphX 的图处理,它们都是由 AMP 实验室提供,能够无缝地集成并提供一站式解决平台[1]。

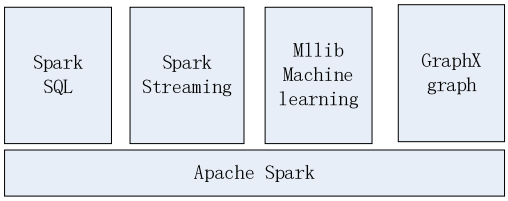

图 1

GraphX 是 Spark 中用于图的并行计算的模块,可以认为是 GraphLab 和 Pregel 在 Spark 上的重写及优化,与其他分布式图计算框架相比,GraphX 具有的优势在于,它是依附于 Spark 之上的,天然的具备了 Spark 的一些特点,并且它提供了数据的一栈式解决方案,可以非常快速且有效地完成一整套图计算的流水作业。GraphX 是 Spark 生态中的非常重要的组件,融合了图并行计算以及数据并行计算的优势,虽然在单纯的计算阶段的性能相比不如 GraphLab 等计算框架,但是如果从整个图处理流水线的视角(图构建,图合并,最终结果的查询)看,那么性能就非常具有优势了。流水作业处理过程如图 2 所示。

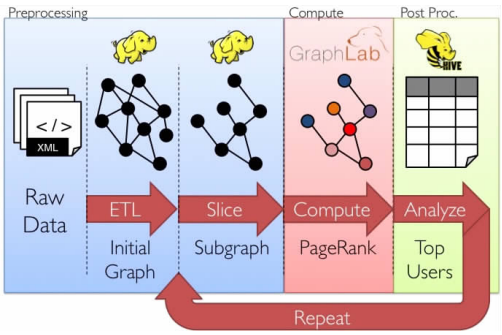

图 2

2 属性图

在 GraphX 中,图的点和边都带有属性,而且这种属性图拥有 Table 和 Graph 两种视图,但是只有一份物理存储。Table 视图将图看成 Vertex Property Table 和 Edge Property Table 等的组合,这些 Table 继承了 Spark RDD 的 API[2]。具体说明如图 3 所示:

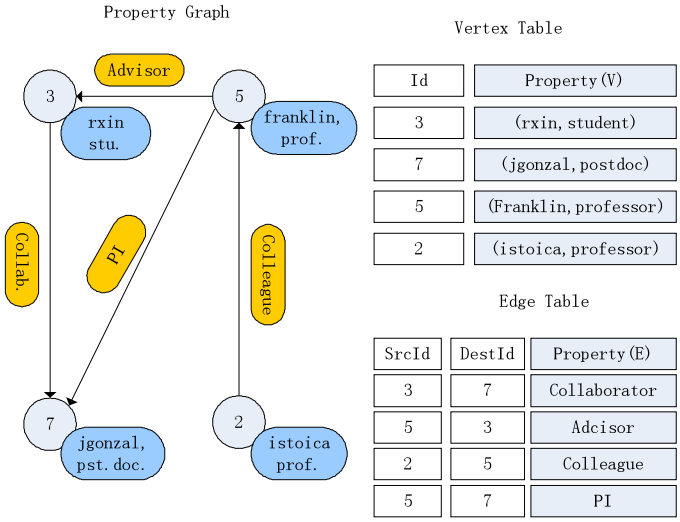

图 3

可知点和边都带有属性,在实际应用中,我们可以根据需求自定义点和边的属性,从而更好地解决我们的问题。Graph 视图上包括了 reverse/subgraph/mapV(E)/mrTriplets 等操作,这些操作可以使我们更加灵活地对目标图进行一系列的操作。

GraphX 的这种属性图的特质带来了如下的好处:

点分割:graphX 存储图的方式为点分割方式。这种存储图方式与传统图计算框架不同的是,不同的机器有可能存储相同的点,但是任何一条边只会出现在一台机器上。所以当点被分割到不同机器上时,会是相同的镜像,有一个点是作为主点(master),其他的点作为虚拟点(ghost)。之所以这样设计,是因为当出现点的数据发生变化时,首先要做的是更新该点的 master 的数据,然后该点的 ghost 所在的机器接受该点所有更新好的数据,更新该点的虚拟点。这样做的好处是对于某个点与它的邻居的交互操作,只要满足结合律和交换律,极大地节省了操作消耗,并且可以保证在边的存储上是没有冗余的。例如求邻居权重的和,求点的所有边的条数这样的操作,可以在不同的机器上并行进行,然后对其进行汇总,这样做可以减少网络开销。

Join Elimination:例如在 PageRank 计算中,一个点值的更新只跟邻居点的值有关,而跟该点本身的值无关,那么在 mrTriplets 计算中,就不需要 Vertex Table 和 Edge Table 的 3-way join,而只需要 2-way join。

Caching for Iterative mrTriplets & Indexing Active Edges:在算法迭代的后期,只有较少的点有更新,因此对没有更新的点使用 local cached 能够大幅度降低通信所耗。

3 GraphX 所支持的算法

在最新版本的 Spark2.1.0 中,GraphX 包含了一系列的图计算的算法来简化用户对图分析的工作[3],具体情况如图 4 所示,其中的数字表示该模块在源码中的代码量。

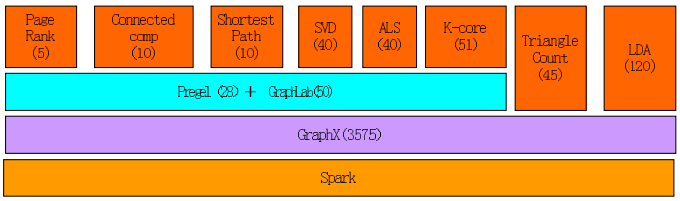

图 4

由上图可知,GraphX 对图计算的算法支持性较好,既包括了像 PageRank 这样的经典算法,也涵盖类 SVD、K-core 等主流的图计算算法,可以基本满足用户对图计算的要求。

如同 Spark 一样,GraphX 的 Graph 类提供了丰富的图运算符,大致结构如图 5 所示。

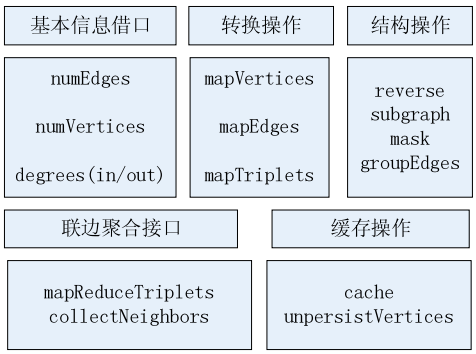

图 5

GraphX 在 Spark 1.3.1 改变了部分用户正在使用 API:

(1)为了改进性能,引入了一个新版的 mapReduceTriplets 称为 aggregateMessages,它取先前返回信息从 mapReduceTriplets 通过一个回调 EdgeContext 而不

是通过返回值。我们正在遗弃 mapReduceTriplets,鼓励用户查阅过度指南。

(2)在 spark1.0 和 1.1,EdgeRDD 的签名切换从 EdgeRDD[ED] 到 EdgeRDD[ED, VD]来进行一些缓存优化。我们已经发现了一个更加优雅的解决方案,恢复了签名到更加自然地 EdgeRDD[ED]类型。

4 GraphX 的应用

随着智能化手机的普遍应用,我们也处在了一个信息化的时代,人们每天的衣食住行都可以产生大量的数据。国内外一些互联网公司为了更好地了解消费者的消费习惯,从而提高经济效益,所以人类行为产生的海量数据就被应用到了广告、报表、推荐系统等这些业务上。这些应用场景的普遍特点是效率要求高、计算量大。其中涉及大量的图计算。GraphX 对于这种情况具有天然的优势,因为它依附于并行化处理框架 Spark 之上,自然地就拥有 Spark 处理大数据的优势;同时 GraphX 自己就包含了很多关于图计算的算法,其属性图的特质可以使用户自定义点和边的属性,极大地提高了图的效率。

阿里巴巴旗下的天猫和淘宝两大平台的广告和搜索业务最开始为了解决一些复杂的机器学习上的一些问题,使用的是机器学习框架 Mahout,但这样做的后果是代码不易维护且效率低下。在使用了 Spark 中的 GraphX 模块来挖掘用户商品关系的生产问题时,取得了较好的结果。同时在商品推荐、社区发现,关系挖掘、基于三角形计数的关系衡量、基于随机游走的用户属性传播等[4]方面表现良好。

优酷土豆在使用 Hadoop 的突出问题主要包括:第一是商业智能 BI 方面,分析师提交任务之后需要等待很长时间才得到结果;第二就是海量数据的计算,例如进行一些模拟广告投放之时,计算量非常大的同时对效率要求也比较高,最后就是机器学习和图计算的迭代运算也是需要耗费大量资源且运算速度很慢。

合一集团下的优酷土豆最开始在进行广告投放和商业智能 BI 方面的技术所用的是 Hadoop,最开始也能满足商业上的需求,但随着视频会员的不断增长以及广告类型的多样化,海量的数据运算就对 Hadoop 带来了很大的压力。在广告投放这一块,优酷土豆的广告技术部门使用了 GraphX 来进行图计算,实际生产中发现集群相较于使用 Hadoop 时压力陡然减小。在对比 Spark 和 Hadoop 的性能时可以发现,Spark 要比 Hadoop 高出 100 倍左右,同时 Spark 提供了诸多模块来适应不同的需求,可以很好地满足工业生产需求。

综上所述,Spark 在如今的互联网公司中应用较为广泛,而 GraphX 更是在公司的核心业务上发挥着举足轻重的作用。可以预见,在大数据时代下,GraphX 必将拥有更加辉煌的明天。

5 结语

本文详细介绍了 Spark 的图计算框架 GraphX,使我们对 GraphX 有了一个较为清楚的认识。同其他图计算框架相比,GraphX 在计算、图优化和性能方面都具有一定的优势。同时,GraphX 依附于 Spark 之上,其计算引擎提供了强大的计算接口,方便了编程,可以很容易地实现 PageRank 等图算法。其在一些互联网公司的应用也很广泛,是一种图计算比较好的计算框架。

参考文献:

[1]黎文阳. 大数据处理模型 Apache Spark 研究[J]. 现代计算机:专业版, 2015(3):55-60.

[2]Xin R S, Crankshaw D, Dave A, et al. GraphX: Unifying Data-Parallel and Graph-Parallel Analytics[J]. Computer Science, 2014.

[3]陈虹君. Spark 框架的 GraphX 算法研究[J]. 电脑知识与技术, 2015(1):75-77.

[4]黄明, 吴炜. 快刀初试: Spark GraphX 在淘宝的实践[J]. 程序员, 2014(8):98-103.

作者简介:

孙海(1991-),男,硕士,研究方向为大数据

收稿日期:2017-02-21　　修稿日期:2017-03-10

(下转第 127 页)

作者简介：
鲜茜(1992-),女,四川成都人,硕士,研究方向为软件质量保证与测试
收稿日期:2017-03-11　　修稿日期:2017-03-20

Application of Code Refactoring in Agile Development

XIAN Xi

(College of Computer Science, Sichuan University, Chengdu 610065)

Abstract:
Software products blow up the trend that continues to increase due to the vigorous development of the Internet. Many companies choose agile development models to save time costs. At the same time, the need for continuous improvement and improvement of software, which lead software become more and more complex as well as the maintenance costs increased. Therefore, the code refactoring is very important. Summarizes the improved methods of optimizing the code by referring to a large number of documents and relying on the project experience, and shows the improvement of performance with result makes the software easier to understand and maintain.
Keywords:
Refactoring; Agile Development; Improvement Methods

(上接第 122 页)

Spark´s Graph Calculation Framework: GraphX

SUN Hai

(College of Computer Science, Sichuan University,Chengdu 610065)

Abstract:
Spark is a generic parallel framework for the open source Hadoop MapReduce from UC Berkeley AMP Lab, a fast and versatile computing engine designed for large-scale data processing. In today´s big data environment, Spark´s role is growing. Introduces Spark´s graph calculation framework GraphX.
Keywords:
Spark; Parallel; Big Data; GraphX