### 用于数据分析的异构云存储配置

摘要

数据分析是公有云服务上一类重要的数据密集型工作负载。然而，为这些应用选择正确的计算与存储配置是很困难的，因为可用选项空间很大和选项之间的交互非常复杂。此外，分析型工作负载访问的不同数据流具有不同的特征，这些特征可以被不同的存储设备更好地服务。

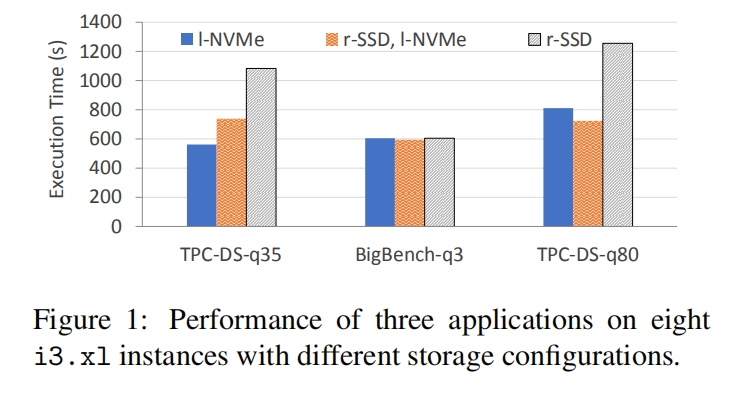
我们介绍selecta，这是一个为数据分析工作负载推荐近最优的云计算和存储资源配置的工具。Selecta使用潜在因素协同过滤来预测应用在不同配置下的性能，基于分析训练工作负载收集的稀疏数据。我们使用上百个SQL和ML应用对selecta进行评估，结果显示selecta选择近最优的性能配置（最优10%内）的概率为94%，近最优的成本配置的概率为80%。我们还使用selecta得出关于云存储系统的重要见解，包括NVME闪存设备的性能-成本效率，云存储关于支持细粒度容量和带宽分配的需求，端到端存储优化的动机。

#### 介绍

随着亚马逊AWS、谷歌云平台和微软AZure把它们的工作负载转移到平台上，公共云市场正经历着前所未有的增长。除了高弹性之外，公有云还承诺减少所有租户的总成本，因为资源是被租户共享的。然而，实现性能和成本效率要求为每一个给定的应用选择合适的配置。不幸的是，大量可用的实力类型和配置选项使为应用选择合适的资源非常困难。

存储选择通常是必要的，特别是对数据密集型分析的云部署。云厂商提供了各种各样的存储选项，包括对象、文件和块存储。块存储可以由硬盘、固态驱动器和高带宽、低延迟的NVMe组成。上述设备可能在运行应用程序的云实例本地或者远程。仅这些选项就会导致存储配置选项在吞吐量、延迟和位成本方面出现数量级的差异。随着基于3D-X poind 的新兴技术的出现，云存储领域变得越来越多样化。

选择正确的云存储配置对性能和成本来说都很重要。考虑SPARK SQL在两张128GB的表上等价查询的结果，我们发现8个EC2节点集群中的实例访问r-HDD比l-NVME查询花费时间要长8.7倍。这与最近的研究形成对比。与之前的spark相比，发现更快的存储只能将任务执行时间中位数最多提高19%。尽管NVMe单位时间成本更高，但NVMe的性能优势使该查询的执行成本降低8倍。如果我们还考虑核数和内存实例的一些选项，性能最好和性能最差VM存储配置之间的性能差距超过30倍。



为分析型应用确定云配置具有挑战性。即使我们将自己限制为单个实例类型并专注于优化性能，为特定应用程序选择存储配置仍然很重要。图一比较了三个spark应用程序的性能，这三个程序分别运行在存储设备分别为I-NVMe、r-SSD和混合(r-SSD用于输入/输出数据，l-NVMe用于中间数据)的i3.8xlarge EC2实例。第一个应用程序是I / O绑定的，并且受益于NVMe Flash的高吞吐量。第二个应用程序有CPU瓶颈，因此对三个存储选项配置表现性能一样。第三个应用程序是I/O绑定的，并且在混合存储选项表现性能最好，因为它最大限度的减少了读写I/O之间的干扰，读写I/Os在Flash上的性能是不对称的。这个结果不应该令人惊讶。分析工作负载访问多个数据流，包括输入和输出文件、日志和中间数据（例如混洗和广播）。每个数据流在访问频率、访问方式和数据生存时间方面都有不同的特征，这些特征使得不同的数据流更适合不同的存储设备。例如表1中的TPC-DS query 80应用程序，输入数据存储在r-SDD和中间数据存储在NVMe闪存比所有数据存储在NVMe表现得更好，因为它隔离了流并且消除了干扰。

我们介绍selecta，一种学习近最优VM和存储配置的工具，用于用户指定的性能成本目标的分析型应用程序。Selecta的目标是对新到达的数据进行频繁或定期重新运行的分析工作[1,25,55]。配置由云实例类型（核心计数和内存容量）以及用户输入/输出数据和中间数据的存储类型和容量定义。Selecta使用稀疏性能数据训练在各种云配置上配置的应用程序，以及仅在两个配置上配置的目标应用程序的性能度量。为了预测不同配置的应用性能，selecta使用潜因子协同过滤，这是一种普遍应用在推荐系统中的机器学习方法[10,57,11,22,23]。Selecta将稀疏性能数据用于配置在各种云配置上的训练应用程序，以及仅在两种配置上配置的目标应用程序的性能测量。与穷举搜索相比，selecta利用稀疏的训练数据来显著更快、更经济高效地学习。该方法还对CherryPick和Ernest等最新的系统进行了改进，这些系统的性能预测模型需要更多关于目标应用程序的信息，因此需要更多应用程序才能收敛[3,69]。更多的是，过去的工作没有考虑异构云存储选项或者每个应用程序中不同数据流的不同偏好[71]。

我们使用超过一百个Spark SQL和ML工作负载来评估selecta，每个工作负载都有两种不同的数据集缩放因子。我们证明selecta选择近最优性能配置的概率为94%（在最优的10%以内）和近最优的成本配置的概率为80%。我们还分析了selecta对各种参数的影响性，例如训练工作负载或者目标应用程序的可用信息量。

我们工作的关键贡献是我们通过分析型工作负载对云存储系统及其使用情况进行了分析，这带来了一些重要的见解。我们发现除了提供最佳性能之外，基于NVMe的的配置还为各种应用程序提供较低的执行成本。我们观察支持容量和带宽的细粒度分配的云存储选项的需求，类似于无服务器云服务提供的计算和内存资源的细粒度分配[7]。分解的NVMe闪存可以为这种灵活的云存储选项提供基础。最后，我们展示了对端到端云存储优化的需求，包括应用程序框架、操作系统和云服务，由于存储堆栈效率低下，几种存储配置没有发挥它们的潜力。

#### 动机与背景

我们讨论用于选择云存储配置的当前方法，并解释了所涉及的挑战。

##### 2.1当前方法

常规方法：用于数据分析作业的输入/输出传统上存在分布式文件系统，例如HDFS或者亚马逊的S3对象存储系统[62.6]。中间数据通常从每个节点上的专用本地块存储卷（例如l-ssd或者l-nvme）并在需要额外容量情况下溢出到r-HDD。在典型的Spark即服务云部署中，每个实例默认提供两个远程存储卷：一个用户实例根卷，一个用于日志[19]。

存在工具：最近的工作聚焦在在云中自动地选择一个最优的虚拟机配置[71,69,3]。然而这些工具往往忽略云存储选项的异构性。在下一节中，我们将讨论云存储配置空间的范围。

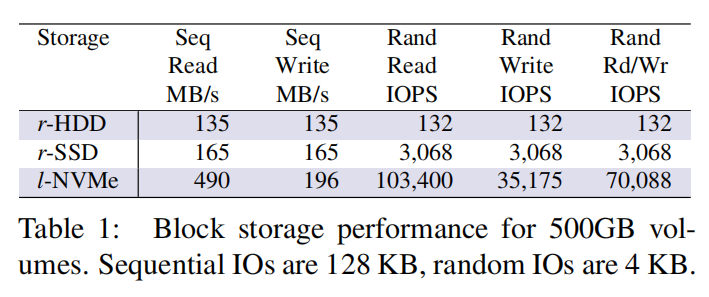
##### 2.2挑战

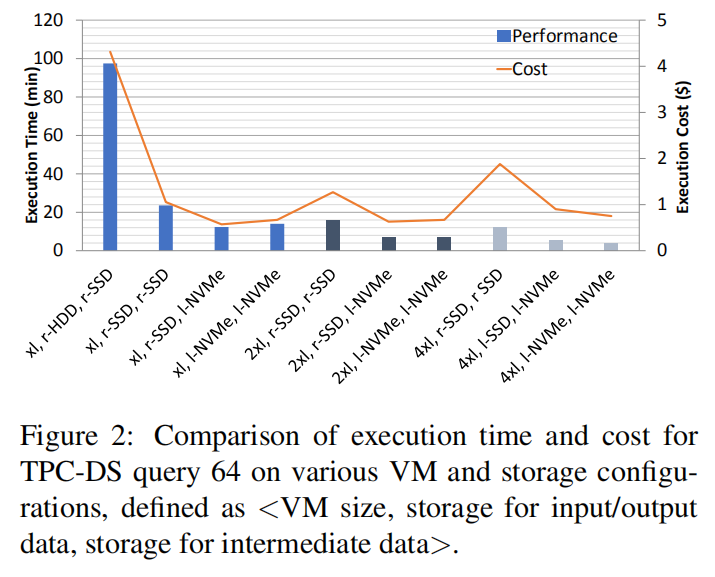
复杂配置空间：云存储有多种形式：对象存储（例如亚马孙S3[6]），文件存储（例如Azure Files[45]）和块存储（例如谷歌计算引擎持久化磁盘[29]）。块和对象存储最常用于数据分析。块存储进一步被细分为硬件选项：冷或吞吐量优化的硬盘，SAS SSD,或者Nvme闪存。块存储对实例来说可以是本地的（直接依附）或者远程的（通过网络）。本地块存储是临时的：数据存在当且仅当实例在运行。远程卷会一直存在直到用户明确删除。

表1比较了在AWS中三个可用的块存储选项。每个存储选项提供一个不同性能、成本和灵活性的权衡。例如，l-NVMe存储以较高的单位成本提供高吞吐量和低延迟。当前，云提供商通常以固定单位容量提供Nvme，该容量直接连接到选择的实例中，按每秒或每小时收费。与没有Nvme的实例相比较，对于有475GBNvme闪存的实例，AWS每小时多收费0.023美元。相比之下，S3的费用是基于容量（每月每GB0.023美元）和带宽（每10KGET请求）使用。除了存储配置外，用户还必须从各种各样 的VM类型中进行选择，以确定正确的CPU和内存数量，VM数量极其网络带宽。这些选择往往影响存储，必须一起考虑。例如，在拥有1GB/S网络带宽的实例中，网络限制了表1中的r-HDD和r-SSD存储卷可实现的序列吞吐量。

性能-成本目标：虽然拥有最多CPU核心、最多内存和最快存储的配置通常提供最高性能，但是其优化运行时间成本也要困难得多。设计用于优化特定目标（例如预测是性能最大化或者成本最小化的配置）的系统并不足以对复杂目标（例如预测在一定预算内使运行时间最小化的配置）提出建议。通过在候选配置上预测应用程序执行时间，我们的方法仍然通用。除非另有说明，否则我们把成本成为执行应用程序的成本。

异构应用数据：我们将分布式数据分析框架管理的数据主要分为两类：通常长期存储的输入输出数据和存活在作业执行期间的中间数据。





中间数据的样例包括在mapper和reduce之间交换的混洗数据、广播变量和从内存溢出的缓存数据集分区。这些数据流通常有不同的访问频率、数据生存时间、访问方式（随机访问、有序访问）和I/O大小。例如，输入/输出数据通常是长期存在的并且是顺序访问的，相反中间数据是短期存在的和大多数访问是随机的。

复杂的存储决策：为作业选择正确的配置可以显著地减少执行时间和成本，如图2所示，该图比较了在8个节点的集群中各种VM存储配置上的Spark SQL查询（TPC-DS query 64）。我们考虑在EC2(X1,2X1,4X1)中3种VM实例大小以及用于输入/输出数据和中间数据的存储选项。最低性能配置的执行时间是最好性能配置的24倍。在r-SSD存储输入/输出数据和在l-nvme存储中间数据（最低成本配置）的成本比在r-HDD存储输入/输出数据和在r-SSD存储中间数据低7.5倍。

#### selecta 设计

##### 3.1概述

Selecta是一个可以自动预测目标应用在一系列候选配置集合中的性能的工具。如图3所示，selecta使用以下作为输入：i）一组应用程序在几种配置上的执行时间；ii）目标应用在两个参考配置上的执行时间；iii）目标应用的性能-成本目标。配置由节点（VM实例）数量，CPU核心和每个节点内存以及用于输入/输出数据和中间数据的存储类型和容量定义。Selecta使用潜因子协同过滤（参阅3.2节）来预测目标应用在其余配置（非参考）上的性能。通过这些性能预测以及各种VM实例和存储选项的单位时间成本，selecta可以为用户的性能-成本目标推荐正确的配置。例如，selecta可以推荐最小化执行时间，最小化成本或者在一定预算内最小化执行时间的配置。

随着新应用的不断启动，这些性能测量会成为selecta不断增长的训练数据集的一部分并且精确度得到提高（见4.4节）。目标应用在selecta推荐的配置上运行结束之后，我们还会反馈性能测量结果-这有助于降低测量噪声和提高精度。由于selecta要花费约1分钟的时间来生成新的预测集合（确切的时间取决于训练矩阵大小），用户可以在使用新的数据集重新启动目标应用的时候来重新启动selecta去获取更精确的推荐。在我们的实验，对每个目标应用的推荐结果在两次反馈迭代后收敛。训练集不断增长的能力还为selecta提供了一种扩展其考虑的配置集的机制。最初，selecta评估的配置空间是原始训练集中的一组配置。当新的配置可用时，selecta会接受在此配置上的应用的分析数据，该工具将会开始预测此配置上所有的应用程序的性能。

##### 3.2性能预测

预测方法：