目前互联网的各种领域的应用有很多已经变成了数据密集型应用，而不是计算密集型应用。其整体数据由【12】所描述的：体积、速度、多样性以及准确性。这些V中的每一个V都要求对爆炸式增长的数据尽心管理。这种爆炸式增长的数据来源于各种领域：像移动、传感器、视频、音频和社交网络等领域。

【伴随着5G、AI、大数据技术的发展，数据量迎来了爆发式增长；】现代应用中的数据爆炸现象给存储系统带来了巨大的压力。此前国际数据公司（IDC）估计，全球生成的数据量将每2年翻一番。如图一所示，到2025年，全球数据将增长到175ZB【1】。在数据爆发增长的背景下，存储系统面临的重大挑战之一是基础架构的可持续存储能力和效率问题。**目前存储领域更偏向于在大容量、低成本与高性能相结合的存储系统的研究。**为了应对这些问题，学者们提出了很多缓解这种压力的技术：重复数据删除、纠删码和数据压缩等数据缩减技术都能够被用来最大限度的减少数据的存储占用空间和IO成本。目前在许多应用程序中，数据压缩已经逐渐成为一种主流的数据缩减方式。

数据压缩技术能够在减少实际存储和io成本方面起到很好的作用。从压缩效果来看，压缩技术主要分为两类：重量级压缩算法和轻量级压缩算法【20】两种压缩技术在对数据的压缩率和压缩速度的表现上相差较大。**如图2所示。**不同级别的压缩算法在压缩率和压缩速度上各有所长。重量级压缩算法具有良好的通用性和较高的压缩率，能够有效的减少存储占用，适合对存储占用空间有较高要求的场景。比如存档、冷备份这种场景。但其缺点是解压时间较高且开销较大，并不适合对访问延迟有所要求的场景。在这样的背景下，为了减少压缩/解压缩所带来的开销，轻量级压缩方法在存储系统中也得到了普遍的应用。轻量级压缩算法对数据的压缩率较低，但是压缩/解压缩速度较快，在减少 I/O 量和数据传输量以及减少开销上有较大优势，可以有效缓解重量级压缩带来的开销问题。适合对访问速度有要求的数据场景。

我们发现不同的压缩算法有着不同的压缩效果，在不同的数据类型和数据分布的前提下，相同的压缩算法也会出现不同的压缩率和压缩效率，因此不同的数据在相同场景下所适合的最佳压缩算法有可能是不同的，因此在不同数据类型的场景下，仅使用一个固定的压缩算法是不能够获得最佳性能的。另外根据ZIPF定律【】的概念，应用程序中的数据往往会有冷热之分，互联网中大约80%的访问操作会落在20%的数据中，通常将这20%经常被访问的数据称为热数据，另外80%的一段时间没有被访问或者一直没有被访问的数据被称为冷数据。对于不同热度类型的数据来说，不同热度的数据有着不同的压缩需求。《同一个trace使用不同压缩算法的访问延迟和存储占用的对比图》例如：对于需要存档的冷数据来说，它几乎不会被访问到，因此该数据需要的是高压缩率来减少其占用的存储空间。而对于经常会被访问到的热数据来说，它需要的则是低访问延迟，那么就可以考虑不压缩或者使用具有高解压缩速度的压缩算法。综合上述所言，没有任何压缩算法能够为所有类型和所有场景下的数据来提供最佳性能。因此基于上述问题，我们需要一个能够自适应匹配数据类型和数据场景的智能压缩分配和调整框架。它可以动态的为数据选择最佳的存储方式并且可以自适应的根据数据的访问热度变化在线的调整其存储方式。

数据压缩技术作为数据缩减技术已经存在了很长时间，并且现在仍然是较为主流的数据缩减技术。目前存在很多数据压缩算法，从压缩性能上来看主要分为重量级和轻量级的压缩算法。前者注重存储空间的减少，但是会带来过多的开销；后者则更注重访问开销的减少，但是会减少其在压缩率上的表现。【不可压缩的数据】虽然数据压缩可以减少存储占用空间。但是对于有些格式的数据来说，在实际中是不可压缩的或者压缩效果极差；那么对这种数据使用压缩算法并不会得到预想的效果。反而会因为执行了无效的压缩算法，带来了不必要的系统开销。这样不仅浪费系统资源还会增加I/O的响应时间。【对比图的解释】因为不同的压缩算法有不同的压缩性能(压缩率和压缩/解压缩速度)。为了更好的了解压缩效率，我们挑选几种压缩算法分别应用在两种不同的数据集上进行实验。具体效果**如图3所示。图2**显示了三种压缩算法对两种类型(XML/二进制文件)的文件的压缩效果。定义压缩率为原始大小/压缩后的大小(CR=OriginSize/CompressSize)。压缩/解压缩速度的单位为MB/S；因此压缩率和压缩/解压缩速度越高越好。从图中可以明显看出两种类型的文件在压缩性能上是有所差异的，并且不同的压缩算法在不同压缩速度和解压缩速度下可以实现不同的压缩率。从图中可以了解到，一般来说有较高压缩率的算法都会有较低的压缩/解压缩速度。反之亦然。GZ和BZ2具有较高的压缩率，但同时其压缩/解压缩速度也会明显减少；而LZ4虽然没有较高的压缩率，但是在压缩/解压缩速度上明显高出很多。

了解实际的工作负载特征是非常重要的。它可以帮助我们根据实际的负载特征来设计模型，以便能够优化系统性能。我们挑选了两种访问跟踪数据集来研究实际访问过程中的冷热数据访问分布。数据集分别为遵循二八定律的ZIPF(alpha=0.8)分布和从剑桥微软研究院的IO访问跟踪中选出的访问数据集。图3绘制了两种数据集的数据访问模式。图中显示了整体 访问情况有冷热数据访问之分。从文件访问排名和访问次数的关系可以了解到：实际访问中存在一部分数据处于频繁访问的状态，剩余其他的数据仅有少量的访问次数甚至一次都没有被访问。总体呈现符合二八定律中所描述的：百分之八十的访问是落在百分之二十的数据上的。【这里再说一下动机？结合压缩和冷热】如果不考虑工作复杂特征，仅仅考虑数据压缩性对数据进行压缩，是以牺牲压缩和解压缩的额外的CPU计算开销为代价的。使用固定算法，不考虑数据访问情况，仅仅只能减少数据占用，但是带来的却是更高的访问延迟。换句话说，像目前这种仅考虑数据压缩性，减少数据量的利用数据压缩技术的方式，仅仅在系统空闲的时候是较为理想的，只有在没有访问或者访问很少的时候可以获得更高的空间节约，而不会明显影响系统性能。如果从结合数据访问情况的角度出发，将压缩算法的多样性和工作负载特征相结合，用来优化基于压缩的存储系统的整体性能，是进一步优化存储系统的潜在机会。并且在以往的研究中，该方向的研究还没有探索过，因此更激励我们从新的角度出发将压缩和数据热度相结合来进一步优化存储系统基于这些观察，从数据热度与压缩算法特性的角度出发，并根据访问情况动态调整其算法，很有可能在性能和空间效率之间取得良好的平衡。这就促使我们提出基于数据热度的动态自适应压缩存储方案。