Alluxio 2.9分享

Alluxio 孙守拙

Alluxio 2.9.0

- Page cache for Presto (Experimental)
- 针对 AI/ML场景元数据缓存研究
- Master 监控

Page cache

Columar 文件例如Parquet/Orc有时只读取文件的一小部分 e.g. 1mb

- 如果将文件全部以block方式缓存，无意义数据占据大量空间
- 如果将block size设为1mb，会造成block数量过多，给master端带来过大压力

只对最经常访问的部分做data locality的缓存，来提升cache hit

Page cache

支持读，写，cache eviction

用户环境中平均presto query latency 12 sec -> 8 sec, P90 19 sec -> 11 sec

Blog:

https://www.alluxio.io/blog/avoid-data-silos-in-presto-in-meta-the-journey-from-raptor-to-raptorx/

Documentation:

https://docs.alluxio.io/os/user/stable/en/core-services/Caching.html?q=page%20cache#experimental-paging-worker-storage

AI/ML场景元数据缓存

当训练文件数量达到百万级以上，元数据操作成为瓶颈。例如，

假如我们有文件/a/b/c/d/file。当我们需要该文件的元数据时，kernel fuse会issue calls：

1. getStatus(/a)
2. getStatus(/a/b)
3. getStatus(/a/b/c)
4. getStatus(/a/b/c/d)
5. getStatus(/a/b/c/d/file)

如果每次getStatus都需要向master发送PRC请求，会严重影响效率

AI/ML场景元数据缓存

Kernel Space Cache (Fuse)	User Space Cache (Alluxio)
每个文件占用300bytes - 1KB （100万文件3GB-10GB）	每个文件占用约2KB
缓存效果更好	缓存效果不如Kernel Space Cache
可设置缓存过期时间	可设置缓存过期时间
不可设置缓存数量	可设置缓存数量
不可对元数据缓存进行操作	可以清除部分或全部元数据 缓存/查看大小
Better resource utilization	Finer granularity

*缓存无法得知元数据的更改（文件大小，最后修改 时间等）

Alluxio Fuse

1. 支持Libfuse3。alluxio.fuse.jnifuse.lihbfuse.version=3。
2. 推荐使用Java 11
    a. JRE SIGSEGV: https://github.com/Alluxio/alluxio/issues/15015
    b. Java 11镜像已发布于https://hub.docker.com/r/alluxio/alluxio-jdk11

Master监控

可以通过metrics监控master端的压力。由CPU/内存使用以及master端的一些内部data structure决定

- Idle
- Active
- Stressed
- Overloaded

Doc: https://docs.alluxio.io/os/user/edge/en/kubernetes/Metrics-On-Kubernetes.html

Thank you!