

Overview

QuickSilver is a light-weight, flexible and distributed policy engine to manage data on a tiered Lustre filesystem. QuickSilver is composed of single-purpose agents that handle tasks such as gathering file metadata, enforcing policy decisions, and executing policy actions like purging or data migration.

These agents are designed to communicate using the NATS message queue software, allowing the number of individual agents to be scaled up as needed. QuickSilver is designed to function while maintaining minimal state information.

Motivation

- Frontier Orion Lustre filesystem is composed of two tiers:
 - An SSD-based Performance Tier
 - An HDD-based Capacity Tier
- Need for a software solution to manage data across these tiers
- Solution must be scalable and reliable
- Scalability ensures it could operate from multiple nodes in a distributed fashion
- Reliability ensures continued Lustre operations in the face of increased IO volume and spikes

Design

- Comprised of fully independent and distributed components
- Scan Agents collect the filesystem state using Lustre’s lfs find
- Policy Agents process the collected filesystem metadata and interpret predefined policy definitions
- Purge Agents perform purge actions on files eligible for purging
- A last-minute check done per file before purge action
- Migration Agents perform migration operation on files eligible for migration

Tools and Techniques

- A decoupled agents-based publish-subscribe style implementation using the C++17
- Uses the NATS message queuing library [https://nats.io]
- NATS serves as a coordination and message queuing system
- Multiple NATS servers enable resiliency, high availability and load-balance in case of high traffic volume

Architecture and Implementation

```
graph TD
    Policies[Policies] --> PA1[Policy Agent]
    Policies --> PA2[Policy Agent]
    Policies --> PA3[Policy Agent]
    Policies --> PA4[Policy Agent]
    SA1[Scan Agent] --> MQ1[Message Queue]
    SA2[Scan Agent] --> MQ1
    SA3[Scan Agent] --> MQ1
    SA4[Scan Agent] --> MQ1
    MQ1 --> PA1
    MQ1 --> PA2
    MQ1 --> PA3
    MQ1 --> PA4
    PA1 --> MQ2[Message Queue]
    PA2 --> MQ2
    PA3 --> MQ2
    PA4 --> MQ2
    MQ2 --> Purge[Purge Agents]
    MQ2 --> Migration[Migration Agents]
```

scan agent

```
[root@rage6 15:58:42][build]# ./scan_agents/sc
an_agent_cmd /lustre/crius/demo/files/capacity
Scan agent(139798265416576): Running...
```

policy agent

```
Policy engine(140702675774336): Has decided that
/lustre/crius/demo/files/performance/work.1/dir
_Ptk1cBlu/file_0NA9C7zTneeds migration
Policy engine(140702675774336): Evaluating messag
e from scan agent { record: "metadata", type: "f"
, path: "/lustre/crius/demo/files/performance/wo
rk.1/dir_Ptk1cBlu/file_21UOX9fd", atime: "1650582
595", mtime: "1650582595", size: "4609472", uid:
"0", gid: "0", format: { filesystems: "lustre", ost_p
ool: "performance", stripe_count: "2", fid: "0x74
00013a0:0x7586:0x0" } }
Policy engine(140557313088384): Has decided that
/lustre/crius/demo/files/performance/work.1/dir
_Ptk1cBlu/file_cz4jHdpHn0pH4CA7fio
```

monitor purge

```
Every 5.0s... rage1: Fri Apr 22 16:31:34 2022
lfs find: warning: /lustre/crius/demo/files/c
apacity/proj.1/dir_ZNaeFtCh/file_8k62f6AQ does
not exist: No such file or directory (2)
138
```

purge agent

```
rius/demo/files/capacity/proj.1/dir_wTzHUSDi/file_e4WmVoNk"
Purge agent(140501452389248): Asked to remove /lustre/crius/demo/
files/capacity/proj.1/dir_wTzHUSDi/file_e4WmVoNk
Purge agent(140501452389248): Received message: {path: "/lustre/c
rius/demo/files/capacity/proj.1/dir_wTzHUSDi/file_pkSjSzC"}
Purge agent(140501452389248): Asked to remove /lustre/crius/demo/
files/capacity/proj.1/dir_wTzHUSDi/file_pkSjSzC
Purge agent(140501452389248): Received message: {path: "/lustre/c
rius/demo/files/capacity/proj.1/dir_wTzHUSDi/file_SmUNov6x"}
Purge agent(140501452389248): Asked to remove /lustre/crius/demo/
files/capacity/proj.1/dir_wTzHUSDi/file_SmUNov6x
```

migration agent

```
Migration agent(140140472490880): Received message: {path: "/lust
re/crius/demo/files/performance/work.2/dir_rMFjJjrb/file_kSIxBpp
E"}
Migration agent(140140472490880): Received message: {path: "/lust
re/crius/demo/files/performance/work.2/dir_rMFjJjrb/file_0cUvXdy
0"}
Migration agent(140140472490880): Received message: {path: "/lust
re/crius/demo/files/performance/work.2/dir_rMFjJjrb/file_14fynls
J"}
Migration agent(140125632920448): Received message: {path: "/lust
re/crius/demo/files/performance/work.2/dir_rMFjJjrb/file_kSIxBpp
E"}
Migration agent(140125632920448): Received message: {path: "/lust
re/crius/demo/files/performance/work.2/dir_rMFjJjrb/file_UrGuSxS
b"}
Migration agent(140125632920448): Received message: {path: "/lust
re/crius/demo/files/performance/work.2/dir_rMFjJjrb/file_kRnU50L
Q"}
Migration agent(140125632920448): Received message: {path: "/lust
re/crius/demo/files/performance/work.2/dir_rMFjJjrb/file_17YfJdS
L"}
Migration agent(140125632920448): Received message: {path: "/lust
re/crius/demo/files/performance/work.2/dir_rMFjJjrb/file_gfykZp6
0"}
```

migration agent

```
"testbed-mgmt2.ccs.ornl" 12:31 22-Apr-22
```

Summary

- A distributed and scalable policy engine to manage data on a tiered Lustre filesystem. Designed as independent agents working in coordination.
- Implemented using C++ platform using NATS message queue system.
- An on-going development with a working prototype. Performance of the current implementation measured over a 1 million files mock dataset.
- The system works in a fully distributed mode over distinct independent nodes

Acknowledgments

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. We would like to thank our collaborators Dustin B. Leverman, Jesse A. Hanley, Bran Radovanovic, Philip Curtis.