

MANAGING AN ARCHIVE WITH METRICS

Bill Anderson, Marc Genty, Erich Thanhardt
National Center for Atmospheric Research
andersnb@ucar.edu, mgenty@ucar.edu,
erich@ucar.edu

The Problem

- We have a 35 PB data archival system with disk cache and tape that we recently converted to HPSS.



- How many tape drives do we need for our archival system?
- What limits should we set on the number of concurrent transfers that a single user can run?
- If we increase the rate of internal data migration, how will user data transfers be affected?

Using Metrics

- Much of HPSS was new to us; to help manage it, we developed software to track various metrics.
- Have tried to build on what others have done and extend it in some ways: different time scales, comprehensiveness, statistical techniques.

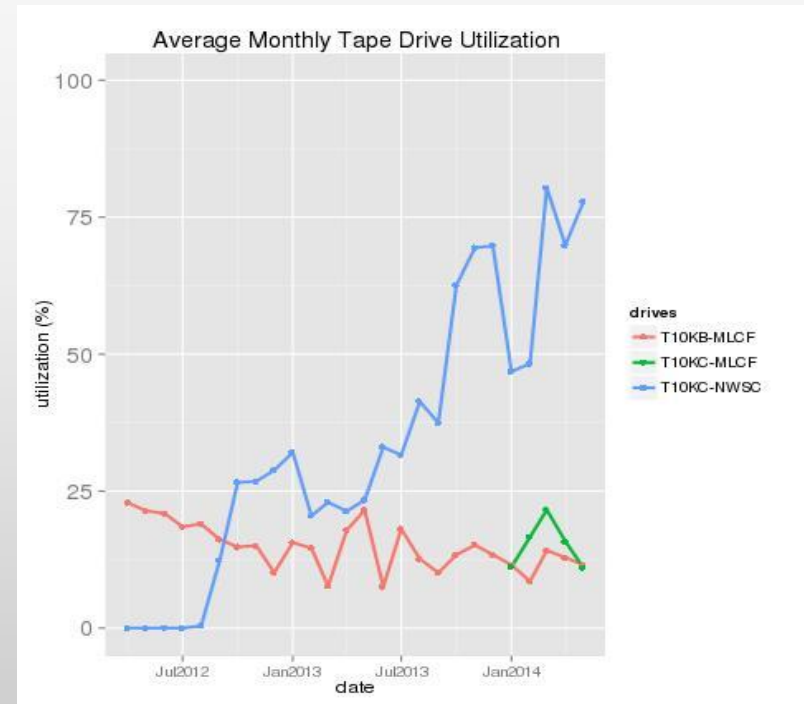
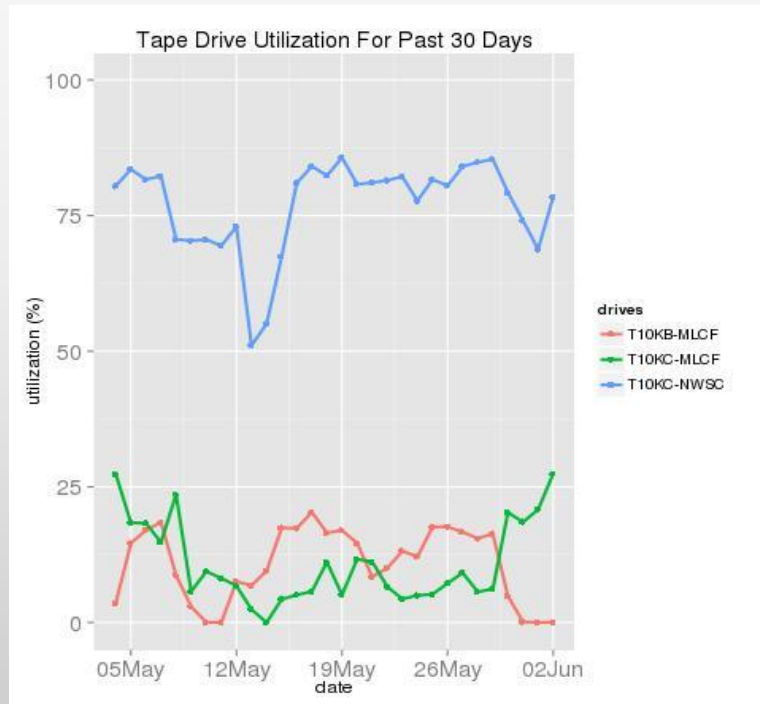
Our Approach

- Different time scales: short for operational management, long for capacity planning
- Didn't find a single existing tool that would do all that we need
- Modular approach: use *bash* and *python* to gather data, *mysql* to store the data, *R* to analyze and plot it, and *HTML* to display it

Our Approach

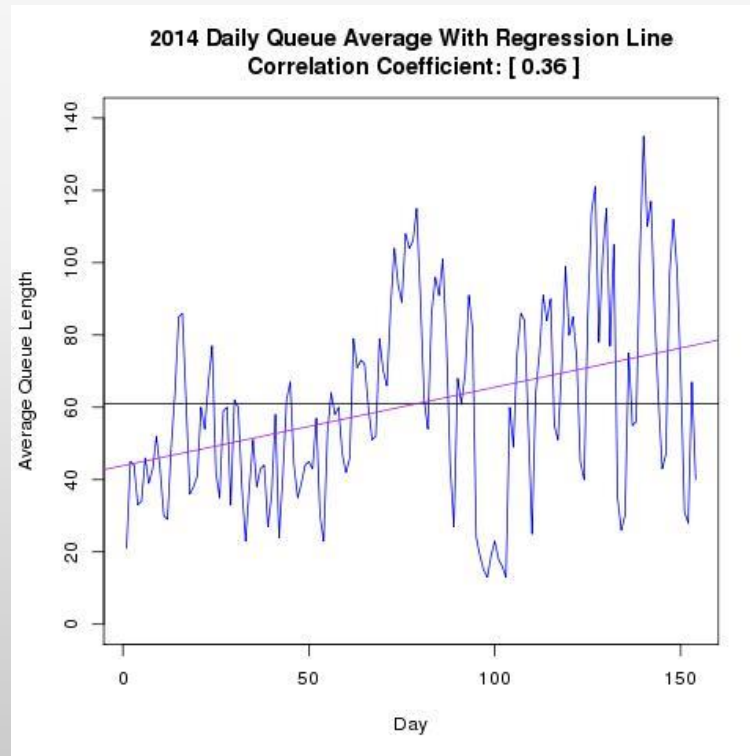
- Use basic statistical techniques and some more advanced ones (e.g., regression)
- Rely on a variety of data sources (e.g., logs, command output, other tools)

Example Metrics: Tape Drive Utilization



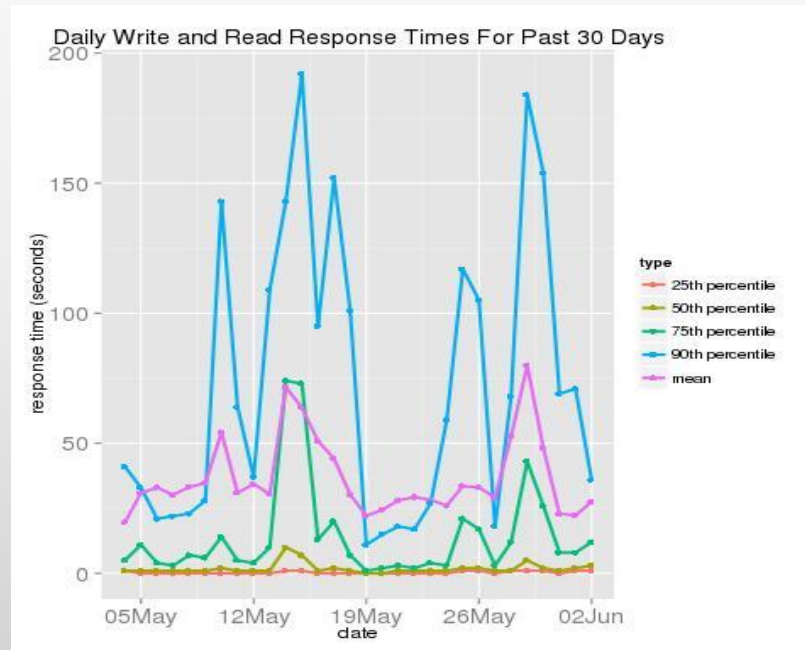
- Operational management
- How many tape drives do we need?
- Different time scales meet different needs

Example Metrics: Queue Length



- Statistical techniques quantify activity and trends

Example Metrics: response time



- Helps us quantify service to users
- Effect of system changes on users

Summary

- Developed a set of actionable metrics to help us manage and provision our system
- Different time scales and statistical approaches have been used
- Future work: new metrics, new statistical techniques
- Interested in hearing about others' uses of metrics