CAPACITY FORECASTING IN A BACKUP STORAGE ENVIRONMENT

Mark Chamness
Principal Engineer
EMC
IT Behavior is Reactive

If it’s not broken (yet), don’t fix it.
Problem: 100% Disk Capacity

- Backups Fail
- Late night alerts
- Administrative short-cuts
  - Delete files
  - Decrease retention policy
  - Remove snapshots
- Fire drill for new equipment and $
Solution: Proactive Prevention

Predict & plan for future capacity needs
Prediction Simplified – Single model

1. Model subset of data (past 30 days)
2. Apply linear regression
3. Choose timeframe for notification: next 90 days
4. Run analysis and generate notifications
Is a single model generally effective?
Single model often performs poorly

- Fixed subset often results in poor predictions
- Why? Not adaptable to changes in behavior
Prediction – Two Models

Generate models for two periods & select best one

Both Wrong!
Prediction – Optimal

Generate all possible models & choose best

Select largest $R^2$ ("Regression Sum of Squares")
Prediction using optimal subset

- Maximum $R^2$ occurs at change in behavior
- Result – best model to fit the data
Prediction using optimal subset

Application of “optimal” model to data from 10,000+ Data Domain backup storage systems

Most of the regression models generated have $R^2$ close to 1.0, indicating good fit to data.
Prediction using optimal subset

Example: model adapts to recent behavior
Prediction using optimal subset

Example: a shelf was added, increasing capacity
Prediction using optimal subset

Example: “Roller-coaster”

\[ R^2 = 0.98 \]
Prediction using optimal subset

Example: Model too ambitions

Model not very good – over-fits recent data
Prediction using optimal subset

Example: Schizophrenic

Model does not work
Model Validation

Requirements for publishing forecasts:
• Goodness-of-fit: $R^2 > 0.90$
• Positive slope
• Forecast time frame < 10 years
• Sufficient statistics: 15 days data
• Space utilization: minimum 10%
• Last data point trumps all
Model Validation

- Last data point trumps all previous data
- Can no longer predict behavior
Analysis of model across all systems

Median time to 100% capacity is 6 months
(Note: Median system is 80% full)
Analysis of model across all systems

Possible explanations:
- Efficient use of capital
- Usage exceeded expectations

6 months