Analyzing backup metadata

Methods and Techniques

Rule-Based Methods

- Simple rule-based method using statistical metrics:
  - Examine metadata like backup frequency, time of day or size
  - Calculate mean and standard deviation
  - Find anomalies by checking for outliers (e.g. > 3σ)

- Can be improved by adding more complex rules

- Tools: NumPy, Pandas

Time Series Analysis

- Interpret the metadata as a time series
- Moving Average:
  - Use a moving average to smooth out noise
  - Anomalies deviate from the smoothed data
- Decomposition:
  - Decompose the time series into different components (trend, cyclical, seasonal, residual)
  - Anomalies have a high residual component
- Tools: Pandas, statsmodels

Density-Based Techniques

- Find anomalies by looking for outliers in the data:
  - k-nearest Neighbor: Outliers don’t have close neighbors
  - Local Outlier Factor: Outliers have a different density than their neighbors
  - Isolation Forest: Outliers can be isolated using few partitions
- One-Class Support Vector Machines
- Tools: Scikit-learn

Neural Networks

- Replicator Neural Networks:
  - Train a network to predict the metadata of the next backup from the previous backups
  - Anomalies are found when the prediction is bad
- Autoencoders:
  - Train an autoencoder on normal backup metadata
  - Find anomalies by looking for backups where autoencoding performs worse
- Tools: Tensorflow, Keras, PyTorch