Hard Drive Reliability and Failures

Chinpei Tang

May 21, 2016

Problem Statement

Hard drive reliability details are important not only for data centers to provide high uptime service, but also to hard drive manufacturer to ensure the production of high quality hard drives to support the competitive and growing business. The project aims at analyze the major causes that results in hard drive failures.

Value

The project will explore and analyze the major causes that result in hard drive failures. The hard drive manufacturers can use such information to focus on improving the key weaknesses. The data centers can also use the information to predict the risk and decide which types of hard drives they should choose to run data centers to ensure high uptime and low customer dissatisfaction.

Dataset Available

Backblaze is an online personal/business backup and cloud storage service provider, which consumes about 1000 hard drive per month. The company wrote scripts to track the hard drive health information in from 2013 to 2015. See more information here. The dataset is made open-source here.

The data contains key properties of the hard drives, whether or not it failed, and 80 to 90 SMART (Self-Monitoring, Analysis and Reporting Technology) parameters (or 40 to 45 normalized values). This can leads to failure mode identification.

However, there has been some known issues with the dataset, so some work will need to be done to extract the valid data.

Solution Approach

- First perform some exploration of the data to look for major trends that lead to hard drive failure.
- Get some familiarity to the monitored parameters to see if they relate to each other, then come out with hypothesis.
- Test the hypothesis with some analysis methods.
- May come out with scoring system to evaluate hard drive reliability.

Deliverables

The analysis result will be delivered in a PDF report constructed using R Markdown. A slide deck will be created to present to the interested audience.