Springboard Foundations of Data Science Capstone Project Proposal: Hard Drive Reliability and Failures

Chinpei Tang

Introduction

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 analyzing the survival period of hard drives, and 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.

Hard Drive Failures

There are two types of hard drive failures:

- Predictable failures: resulting from slow processes such as mechanical wear and gradual degradation of storage surfaces. Monitoring can determine when such failures are becoming more likely.
- Unpredictable failures: happening without warning and ranging from electronic components becoming defective to a sudden mechanical failure (which may be related to improper handling).

SMART (Self-Monitoring, Analysis, and Reporting Technology)

The technical documentation for SMART is in the AT Attachment (ATA) standard. First introduced in 2004, it has undergone regular revisions, the latest being in 2008. Each drive manufacturer defines a set of attributes and sets threshold values beyond which attributes should not pass under normal operation.

Each attribute has:

- a raw value, whose meaning is entirely up to the drive manufacturer (but often corresponds to counts or a physical unit, such as degrees Celsius or seconds)
- a normalized value, which ranges from 1 to 253 (with 1 representing the worst case and 253 representing the best)
- a worst value, which represents the lowest recorded normalized value. Depending on the manufacturer, a value of 100 or 200 will often be chosen as the initial normalized value.

Manufacturers that have implemented at least one SMART attribute in various products include Samsung, Seagate, IBM (Hitachi), Fujitsu, Maxtor, Toshiba, Intel, sTec, Inc., Western Digital and ExcelStor Technology.

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.

Data Organization

The data is organized into three directories. They are named data_2013, data_2014 and data_2015. Each folder contains csv file of the hard drive details on the shelf on each day. The file names are organized in the format of yyyy-mm-dd.csv. There are 266 days of data in 2013, 365 days of data in both 2014 and 2015. Hence, in total there are 996 days of data.

In each of the csv file, each row of data is the details of each hard drive that is operational on the shelf each day. It is identified by the serial numbers of the individual hard drives. The model and capacity of the hard drive is reported. If the hard drive is failed on the particular day, it is labeled 1 in the failure column; otherwise it is labeled 0. The row with the failed hard drive will be removed in the csv file of the next day, and, potentially a new hard drive will be added. The SMART data of each of the hard drive is also reported in columns. On 2013-04-10, there are 21,195 hard drives on the shelf, while, on 2015-12-31, there are 57,544 hard drives on the shelf.

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.