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

Chinpei Tang

Introduction

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 trends of the survival period of hard drives of variety of hard drives available in the market. These trends along with some physical data, its major causes that results in hard drive failures can possibly be deduced.

Value

The project will explore and analyze the trends 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).

Goals

The main goals of the projects are:

- Understanding the failure rates of the hard drives based on its different manufacturer, models, and sizes. These may lead to the list of "high risk" hard drives that may be avoided for critical operations.
- Estimate and predict the survival rates of the hard drives based on its different manufacturer, models, and sizes. The data center can then proactively plan for the hard drive procurement and replacement to minimize inventory and maximize uptime.

Data 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.

Data Wrangling and Exploration

The major data wrangling work is the import of the huge amount of data into R or RStudio for analysis.

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 (starting 2013-04-10), 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.

Each csv file is about can range between: 3 MB to 4 MB per file for 2013 and 2014, and 10 to 11 MB for 2015. The total sizes are 738 MB for 2013, 2.81 GB for 2014 and 4.19 GB for 2015.

Key Data

Due to large amount of data logged in the dataset, we can probably only import certain key columns in the data. As the first step, we will look into only the date, serial number, model number, capacity and if the harddrive failed, and ignore the SMART data. We believe that by looking at the above data, we can deduce some failure conclusion.

Also, the study will first focus on a small set of csv files, i.e. only a few days, a week, or a month. Then it will extend to a year, or even full 3 years. The scope of study will need to be careful due to the amount of data involved, even with only 5 columns of data.

In the later stage if SMART data study is involved, good strategy during the data import stage will be required. It will need to import only relevant information (or just enough details) for study as the data size can quickly grows in size. Some of the information may need to be processed as soon as a csv file is imported to build the data frame. For instance, if one of the SMART data is required to study with respect to brand, the serial number may not need to be imported, and the model number may need to be converted immediately to brand, and the model number is then disgarded since the model numbers are strings that can take up much of the data size. Then the failure data and the corresponding SMART data of interest will be imported.

One of the validations that we will need to do is to ensure that there is no duplicated serial numbers in each row of the observation. This simple check can be done by counting the distinct number of serial numbers of each csv, and see if it is equal to the number of rows.

Preliminary Study

During the preliminary study, we look into only importing 15 csv data and creating a data frame, then export to a Rdata file to load and analyze later. 15 files were managable. We found that RStudio was unable to import and create a data frame of larger than 90 files, since it will be larger than the memory that can hold the data (about 8 GB RAM).

We then look into only loading the first 5 columns of data - date, serial number, model number, capacity and failure. The corresponding Rdata file could then be created in each year: hd2013.RData (~34 MB), hd2014.RData (~83 MB) and hd2015.RData (~110 MB). The file harddrive_readdata.R is used to import and create the data frame, and export to the RData files.

Data Exploration

The data exploration that can be done is to plot the total number of hard drives and the number of failed hard drives over the time. Different brand and capacity can be plotted as well, which we will do.

Data Manipulation

The project will use dplyr extensively to manipulate the data. The data will be grouped and sorted properly to better reveal information.

We also use grep function to identify the following expressions in model to assign to a new column brand:

- "HGST"" -> "HGST"
- "Hitachi" -> "Hitachi"
- "SAMSUNG" -> "Samsung"
- "TOSHIBA" -> "Toshiba"
- "WDC" -> "Western Digital"
- "^ST" -> "Seagate"

The only tricky part is all Seagate hard drive has prefix "ST", while "ST" is part of "HGST" that requires special distinction. Fortunately in "HGST" hard drives, "ST" won't be prefixes.

Data Modeling

Survival Analysis

The suitable analysis to be studied on the data is the survival analysis. Survival analysis is to analyze the expected duration of time until an event occurs. In the case of hard drive, our interest is to understand the expected duration of the hard drive can "survive" on the shelf before it fails. R has a package "survival" to perform the analysis.

This is important since the datacenter like BackBlaze can: (a) favorbly procure the hard drives that have high "survival" period, (b) optimize the time for procuring the replacement, i.e. not procuring too early to store too much inventory, or too late that affect the uptime.

A few covariates (basically properties that affect the survival) can be looked into. First is more on the hard drive physical properties itself like manufacturer (brand) and capacity (and may be rpm it operates at, since we can map model number to rpm). The reason why this is physical is that it is something that doesn't change over time.

Second is on the properties changed over time such as using the SMART data. This is more difficult since typically covariates are discrete variables, while SMART data are time series integers. Dependings on the progress of the project, this will probably be looked into very briefly.

Calculations

In order to perform survival analysis, the data needs to be coverted such that: * each of the rows is the serial number of the hard drive - which serves as the identifier of the "subjects" to be studied. * the number of days it survives over a period. For instance, initially we may look into a month worth of data (the period of study), then count the number of days it survives (number of rows with 0 for each serial number). * if the event (failure) occurs. If it fails within the period, put 1 in the event column. We may consider censorship, i.e. it doesn't fail within the period of study.

dplyr will be used to manipulate the data from the date - serial number - model - capacity - failure format into serial number - survival period - event - covariate 1 (model) - covariate 2 (capacity).

Validation

By looking at the survival analysis in the first case, we can hopefully see some trends like which manufacturer/brand has higher reliability, or which capacity as higher reliability. For covariate like capacity, we can potentially see if there is a "linear" trend between the capacility and the survival rate. The validation can be done by using one of the years as the base study, and validate over year 2014.

Data Story and Result

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.

If time-permitting, a Shiny application can be written so that the user can select different brand and capacity to display the survival curve over time on the visualization. This way it may serve as a decision tool for the next hard drive purchase for consumer.