Disk Usage Trend

Abstract

In this era, where we store huge amount of data on our servers, we would hate to get out of disk space unexpectedly without having any backup plan of how to proceed further. So to prevent this from happening we would love to receive a initial warning a few time prior of our disk reaching its storage limit. As the usage statistics are different for everyone and disk spaced gets filled at different rate for servers, therefore we needed a method of prediction of filling of the disk space based on the disk usage trends of the server.

This project deals with forecasting the amount of free disk space of the servers based on the usage statistics of disk space of the previous days.

Requirements

Softwares required

- Python 3.5
- Any Python IDE (Pycharm, Jupyter Notebook or Spyder)

Libraries required

- Pandas
- Scikit-learn
- Numpy

Dataset Used

Link – https://www.kaggle.com/wildgrok/disk-space-trend/data

Libraries Description/Purpose

Pandas- Pandas is probably the best tool aroung for reaading/writng data and processing it. It converts our datasets into dataframes on which we perform a lot of data manipulation operations.

Scikit-learn- Scikit-learn is a powerful Python library for machine learning. It contains function for regression, classification, clustering, model selection and dimensionality reduction. Here we intent it to used for building the regression model in which the target value is expected to be the linear combination of the input variables

Numpy- Adds python support for large, multi-demsional arrays and matrices, along with a high-level mathematical functions to operate on these arrays.

ML: Algorithm

Regression- As this is the case of predictive analysis we chose to apply regression model for this scenario.

Regression establishes a relationship between dependent variable and one or more independent variable using a best fit straight line (also known as regression line).

Line can be represented by the equation Y=a+b*X+e (where, a is intercept, b is slope of the line and e is the error term).This equation can be use to predict output. Best-fit line easily be obtained by using the method of least squares. It calculates the best fit line for the observed line by minimizing its distance from each data-point to the line

After obtaing the best fit line we can predict the output based on any given input variable. In this case, we will be able to predict the time and date at the which the disk-space might get filled completely(or 100%)

Input Data Description

This dataset contains the data about the disk-space captured for several monts for a set of Windows servers.

Column description of the dataset-

Machine- This column denotes the Id for each machine.

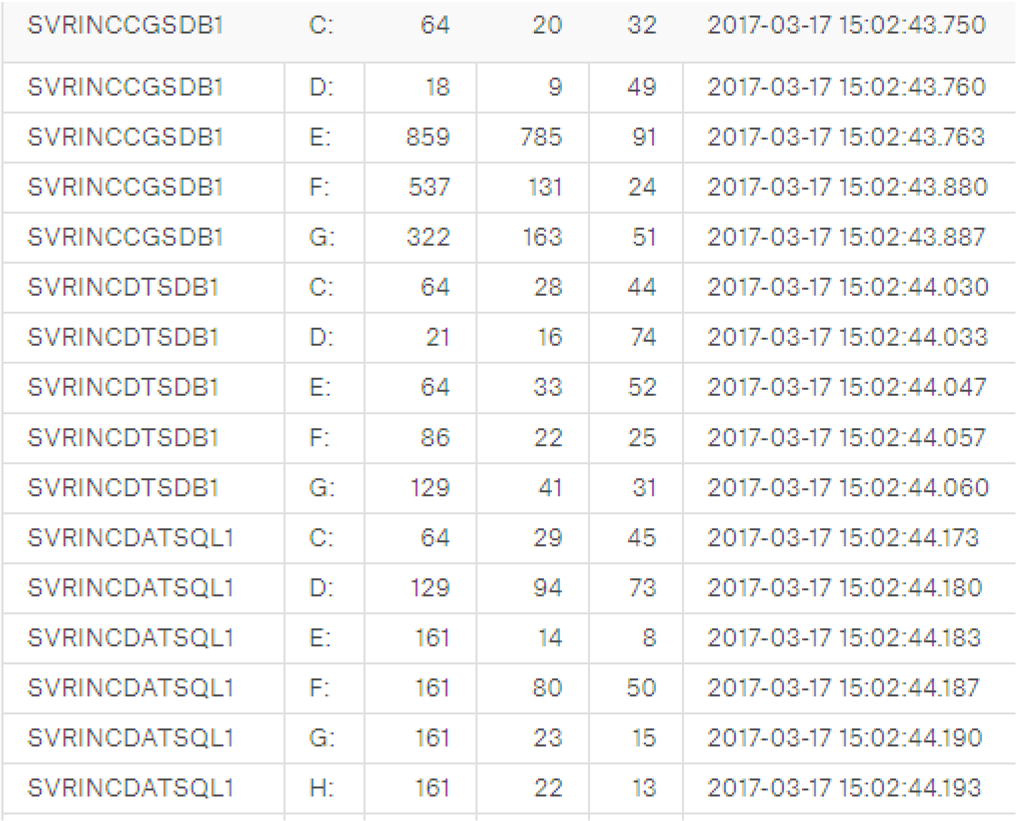
Drive- It contains the name of the drive data is about.

Total- It contains the total space of the particular drive

Used- It contains the space currently in use by the drive.

Free- It contains the amount of space left in the drive

Date- It contains the date and time stamp for the particular log entry about the disk-space usage.



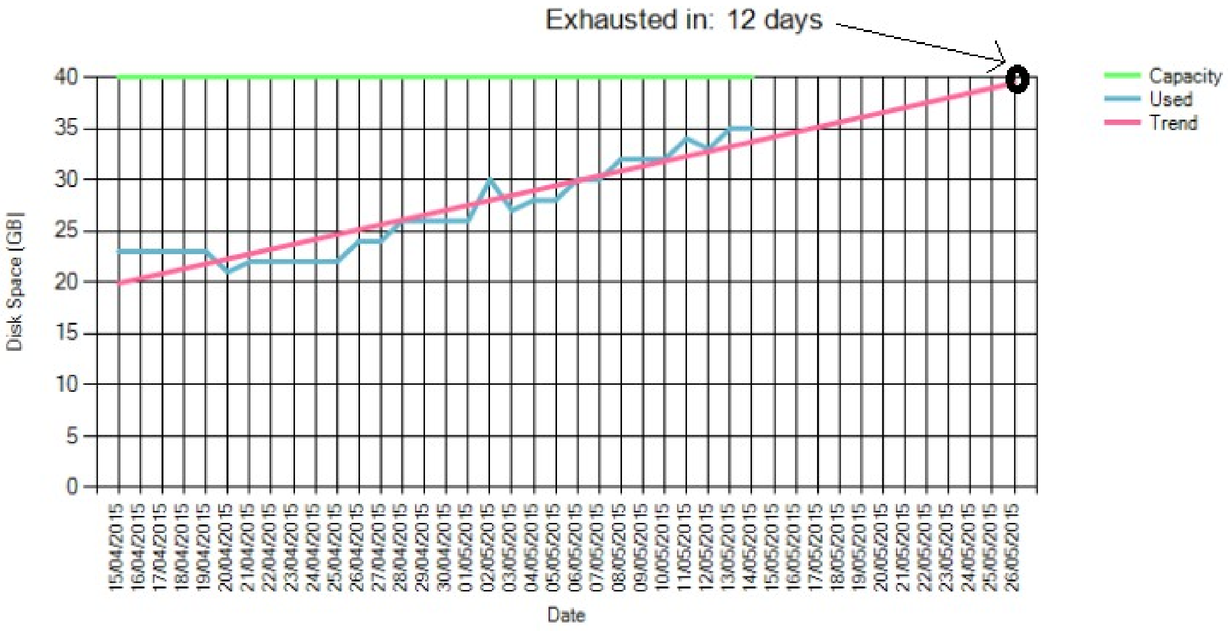
SVRINCCGSDB1	C:	64	20	32	2017-03-17 15:02:43.750
SVRINCCGSDB1	D:	18	9	49	2017-03-17 15:02:43.760
SVRINCCGSDB1	E:	859	785	91	2017-03-17 15:02:43.763
SVRINCCGSDB1	F:	537	131	24	2017-03-17 15:02:43.880
SVRINCCGSDB1	G:	322	163	51	2017-03-17 15:02:43.887
SVRINCDTSDB1	C:	64	28	44	2017-03-17 15:02:44.030
SVRINCDTSDB1	D:	21	16	74	2017-03-17 15:02:44.033
SVRINCDTSDB1	E:	64	33	52	2017-03-17 15:02:44.047
SVRINCDTSDB1	F:	86	22	25	2017-03-17 15:02:44.057
SVRINCDTSDB1	G:	129	41	31	2017-03-17 15:02:44.060
SVRINCDATSQL1	C:	64	29	45	2017-03-17 15:02:44.173
SVRINCDATSQL1	D:	129	94	73	2017-03-17 15:02:44.180
SVRINCDATSQL1	E:	161	14	8	2017-03-17 15:02:44.183
SVRINCDATSQL1	F:	161	80	50	2017-03-17 15:02:44.187
SVRINCDATSQL1	G:	161	23	15	2017-03-17 15:02:44.190
SVRINCDATSQL1	H:	161	22	13	2017-03-17 15:02:44.193

Output Data Description

Output of the project is to determine when the disk-space may be filled. Hence, using the generated model build previous usage statistics of the disk-space on server we can predict the approximate time and date at which the disk might get filled.

We can later also set a threshold on the percentage of disk-space used, which when crossed a warning is generated to warn the user beforehand that disk might be out of storage in this approximate amount of time period.

Model of the data and output will be similar to this:


Exhausted in: 12 days
Capacity
Used
Trend
Disk Space [GB]
Date
15/04/2015
16/04/2015
17/04/2015
18/04/2015
19/04/2015
20/04/2015
21/04/2015
22/04/2015
23/04/2015
24/04/2015
25/04/2015
26/04/2015
27/04/2015
28/04/2015
29/04/2015
30/04/2015
01/05/2015
02/05/2015
03/05/2015
04/05/2015
05/05/2015
06/05/2015
07/05/2015
08/05/2015
09/05/2015
10/05/2015
11/05/2015
12/05/2015
13/05/2015
14/05/2015
15/05/2015
16/05/2015
17/05/2015
18/05/2015
19/05/2015
20/05/2015
21/05/2015
22/05/2015
23/05/2015
24/05/2015
25/05/2015
26/05/2015