Assignment : Understanding gradual changes in distribution in Cloud usage trace

Dataset: Overview of Google Cluster Usage Trace

The Google cluster dataset offers a detailed insight into the workload dynamics of one of the world's leading technology giants. The dataset provides a comprehensive overview of how tasks utilize computer resources within our clusters. This table presents a granular and detailed look into how tasks are functioning and consuming resources within our computer clusters. They routinely observe and monitor these tasks approximately every 5 minutes. Here's a rundown of what our table includes:

- Start Time & End Time: Pinpoints when we initiated and concluded our observation.
- Job IDs & Machine IDs: Unique identifiers for the tasks and machines.
- Mean CPU Usage Rate: Indicates the average CPU power exerted.
- Memory Usage: Details about canonical, assigned, unmapped page cache, and total page cache memory usage.
- Disk Usage and Efficiency: Information on mean disk I/O time, mean local disk space used, and maximum disk IO time.
- Special Metrics: Such as CPI (Cycles Per Instruction) and MAI (Memory Accesses per Instruction).
- Sampled CPU Usage: This metric displays the CPU's average utilization during a randomly selected second within our 5-minute observation window.
- Linux's Own Memory Consumption: Highlighting how much memory the Linux system is consuming on behalf of the task.
- Disk Space Usage: Excluding storage used by large systems like GFS and Colossus, we record the amount of space occupied by tasks.

Task is focused on the CPU ,disk I/O time & memory usage only

Assignment Tasks :

You have to mark the gradual change in the time series graphs by considering changes in the capacity groups. Time series flow shows you a mean CPU, memory and disk I/O time is given in as graphs the interface with the sliding window to select the change range and to mark it. Gradual change occurs over a longer period of time. Marking of gradual change in the time series needs some consideration points in mind. To effectively recognize drift patterns, the continuous values indicating usage need to be segmented into capacity groups for better understanding. These groupings rely on threshold values established in academic papers, which offer insights into the categorization of "high", "medium", and "low" usage on the basis of these thresholds. Capacity groups of each resource given in table 1& 2 below :

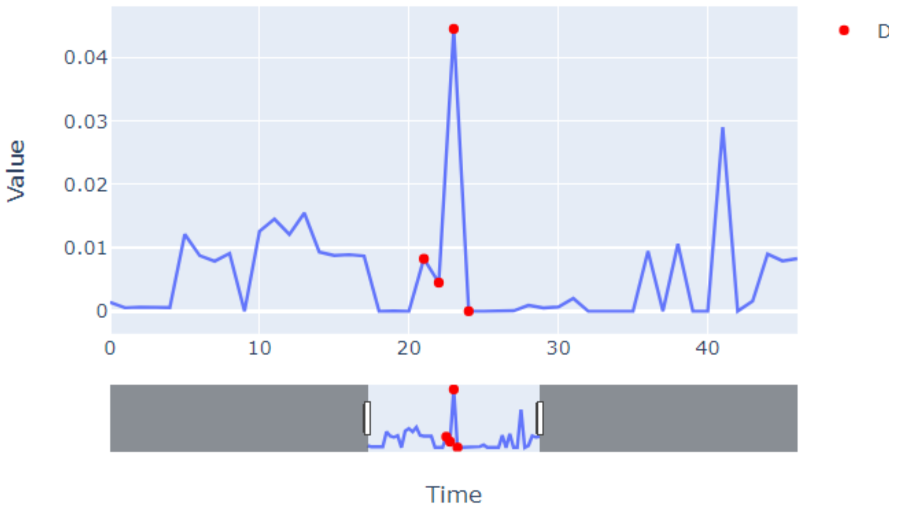
Table1: Capacity group for CPU & Memory

Low	Medium	High
Group 1: 0-0.0345 Group 2:0.0345- 0.0517 Group 3:0.0517-0.0689 Group 4: 0.0689-0.0759 Group 5:0.0759-0.0776 Group 6:0.0776- 0.12	Group 7:0.12-0.1362 Group 8:0.1362-0.1505 Group 9:0.1505-0.1638 Group 10:0.1638-0.18 Group 11:0.18- 0.6	Group 12:0.6- 0.8909 Group 13:0.8909-0.9679 Group 14:0.9679-0.9973 Group 15:0.9973- 1.0000

Table2: Capacity group for Disk IO Time

Low	Medium	High
Group 1: 0-0.034 Group 2:0.08- 0.12	Group 7:0.12-0.3 Group 8:0.3-0.6	Group 14:0.6-0.9 Group 15:0.9- 1.0

In the figure below you can see the ranges on the y axis ; these must be considered interim of capacity groups while considering the range. When a value jumps from one group to another group with a change of three and above, the class will be marked as gradual. This change can be started from even a small jump of two to three groups. In this case the red dots represent a gradual drift as it moves from Group 1 to 4. Remember the duration of change can vary in circumstances.



Figure

Tutorial

To open Dash Interface (figure1) the Collab (figure 2)is divided in four sections : Library , data loading , code & output.

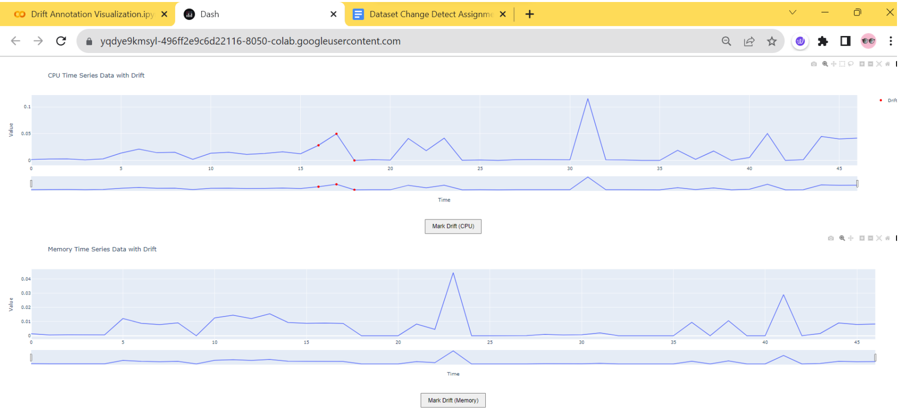
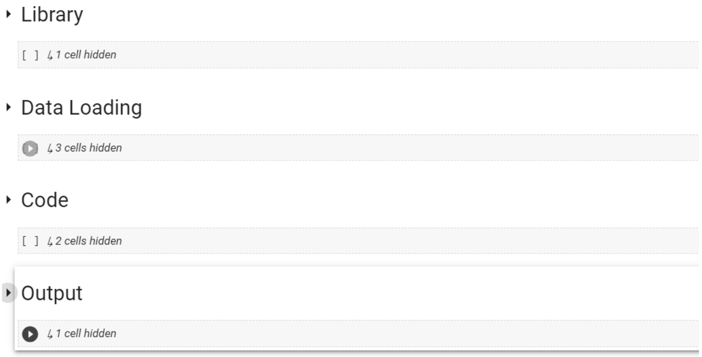

Drift Annotation Visualization.ipy
Dash
Dataset Change Detect Assignm
yqdye9kmsyl-496ff2e9c6d22116-8050-colab.googleusercontent.com
CPU Time Series Data with Drift
Value
Time
Drift
Mark Drift (CPU)
Memory Time Series Data with Drift
Value
Time
Mark Drift (Memory)

Figure 1



▸ Library

[ ] ↳ 1 cell hidden

▸ Data Loading

↳ 3 cells hidden

▸ Code

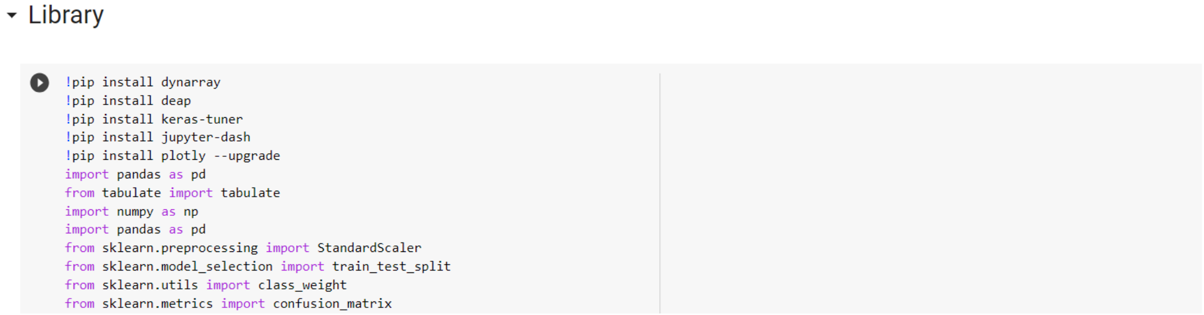
[ ] ↳ 2 cells hidden

▸ Output

↳ 1 cell hidden

Figure 2

1:Load necessary imports by running the “Library” section as shown in figure 3



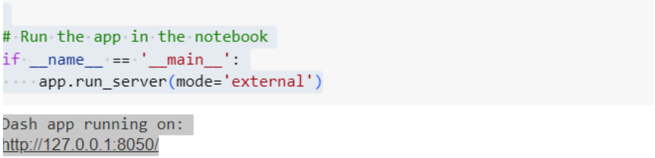
▾ Library

```
!pip install dynarray
!pip install deap
!pip install keras-tuner
!pip install jupyter-dash
!pip install plotly --upgrade
import pandas as pd
from tabulate import tabulate
import numpy as np
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.utils import class_weight
from sklearn.metrics import confusion_matrix
```

Figure 3

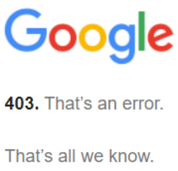
2: In “data loading” section you have to enter the file name you have to work (Note: don't enter file extension just name) and then run all the data loading cells.

3: Now run the “code” cell you will see a local host link e.g. http://127.0.0.1:8050/(figure 4) click on it you will see a graph on the dashboard on the new tab. Note: wait 20 sec to load it completely



```
# Run the app in the notebook
if __name__ == '__main__':
    app.run_server(mode='external')
```

```
Dash app running on:
http://127.0.0.1:8050/
```



403. That's an error.

That's all we know.

Figure 4

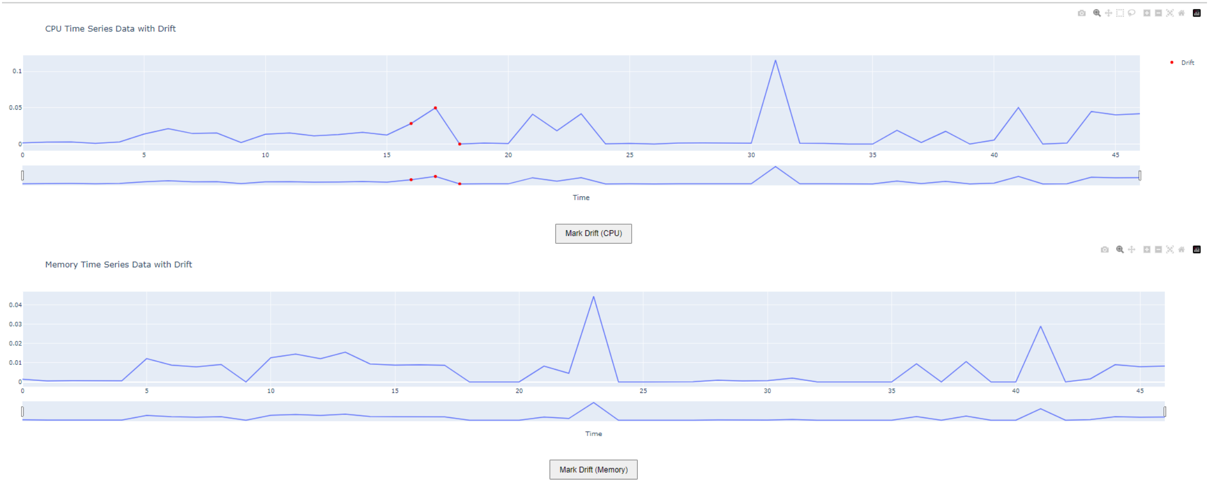


Figure 5

4: A sliding window approach is given in the interface (figure 6) to analyse the data over time. You can choose a window size based on the granularity of your data. Start from moving the slider at the bottom of the graph and select drift range.
(Note: Incase you of no sliding window Click that particular mark button to make it apper)

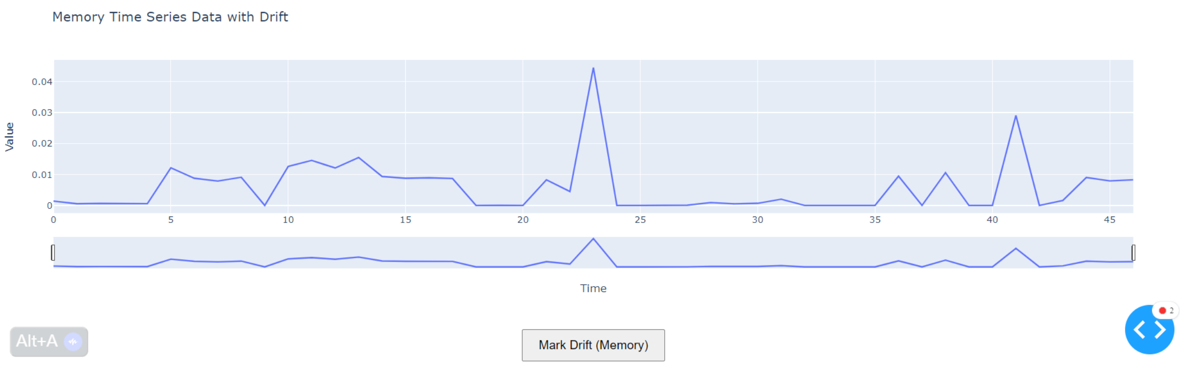


Figure 6

In order to mark drift, the select range from that particular graph as shown in figure 7 and click on the mark drift button.

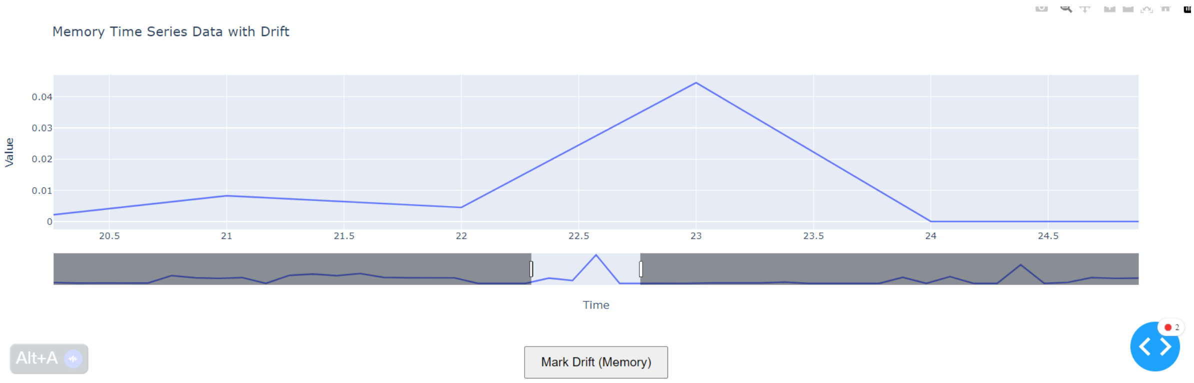


Figure 7

To ensure the correct marking red circle will appear on the change part once clicked as shown in figure 8.

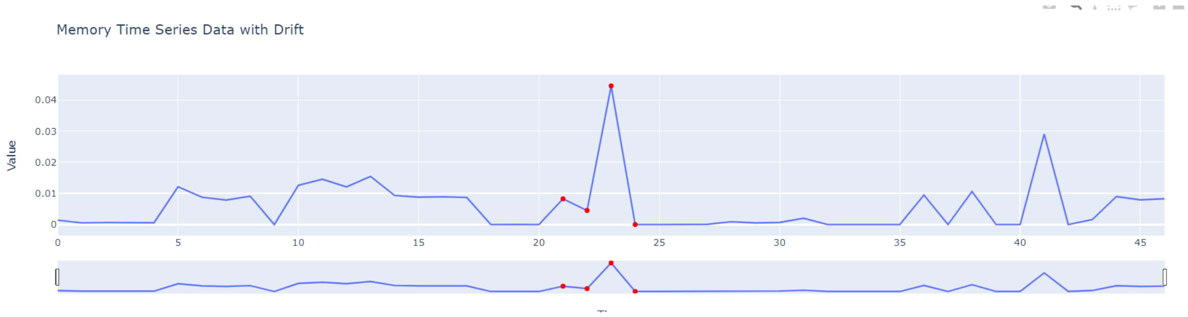


Figure 8

5: Heuristic for Gradual Change Detection:

For each sliding window, calculate the percentage of time spent in each resource usage level (low, medium, high). Compare the current distribution of resource usage levels with a historical baseline (e.g., a rolling average of the distribution over a previous period).Mark the changes in the distribution over time and how long the change has occurred . Gradual change may be indicated by a sustained increase in high resource usage over several windows. A decrease in low resource usage with a corresponding increase in high or medium resource usage. Just read the above capacity group in table 1& 2 to better understand the gradual change part. These capacity groups are marked through coloured lines on the graph to assist.

6: Now go back to your collab and run the “Output” section. You can see the result in the tabular form holding an instance with a red back that you have marked as change. The result will be saved as a csv file, download it and submit it.

Note:

In case you have marked wrong drift you have to load the dataset against the dash and then save the file.

Links:

Script & Dataset

https://github.com/Tajwarresearch/DataAnotation