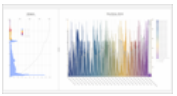


# Data Analysis Documentation

The Data Analysis Tool is a html+javascript based utility used by administrators, tuning experts, support engineers and consultants to create professional graphs based various types of input data. The tool is fully automatic and significantly faster than Excel at creating scatterplots, bar charts and line plots together with histogram analysis for almost any column separated input data such as **NMON** (*CPU_ALL, MEM, MEM_NEW, PROC, LPAR, PAGE*), **VMSTAT**, **SDFMON**, **TAANA**, **ABAPMETER**, **MHTML**, **NIPING** and **XLSX**. DataAnalysis.HTML will run in all modern web browsers like Chrome, Edge, Firefox, Opera, Safari (*Windows / Linux / MacOS / iPadOS*).

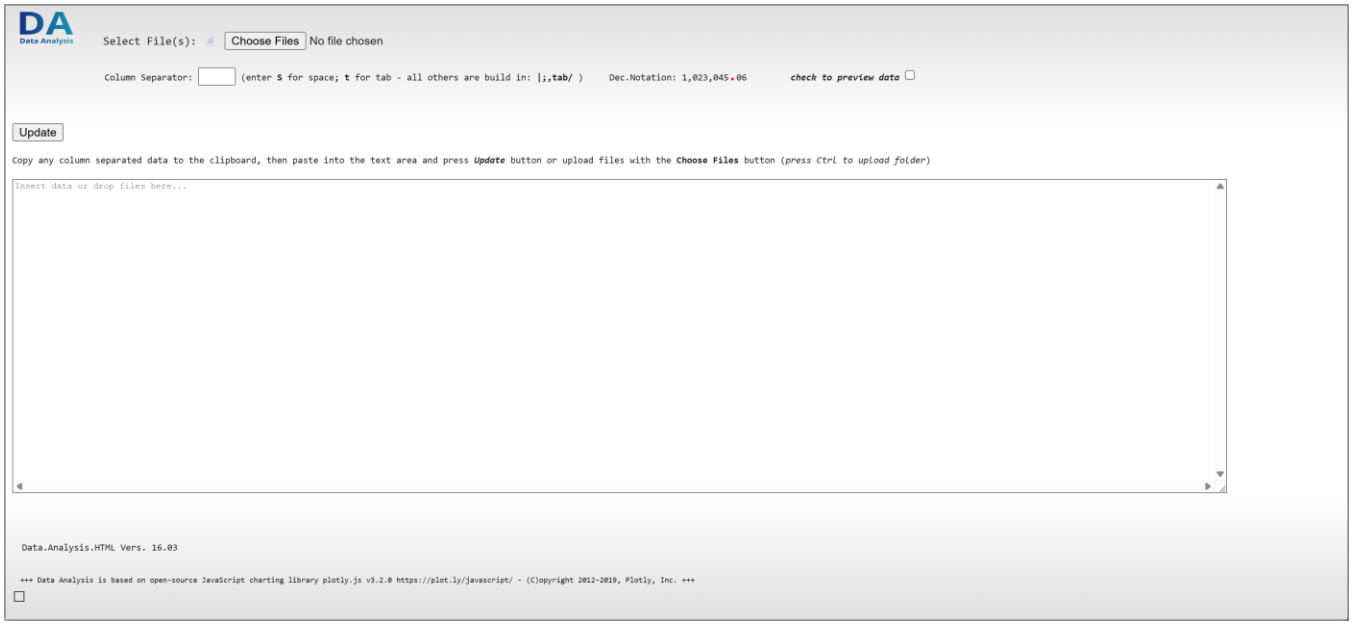
| | |
|---|---|
| SAP internal note | 3169320 - Data Analysis |
| SAP JAM | https://community.sap.com/t5/technology-blog-posts-by-sap/data-analysis-tool/ba-p/13548321 |
| Github | https://github.com/i813812/DA-Data-Analysis |

## Contents

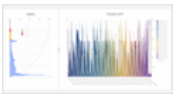
# 1 Select Input Data

DA Data Analysis allows to open and concatenate multiple input files, it will detect automatically the column separator (*tab, semicolon, comma..*) - any character not included in 0..9, A..Z, a..z or *space* will be detected. To load data into the Data Analysis tool you have three options:

- Use the "Choose Files" button and select one or multiple files *(press **Ctrl** Key to toggle between file and folder selection - if folder is active DA will select all files in a folder)*
- Drag and drop one or multiple files over the "Choose files" button
- Copy & paste any column separated data into the input box
- In some cases when many large numbers with thousand delimiter are included a wrong column separator, typically , or . might be selected - in that case we need to specify the separator manually (*use **S** for space, **T** for tab or specify any other character /|\,.;:-*)
- **Specify Decimal format** (*point or comma - if numbers are too small the program cannot determine the correct format automatically)*

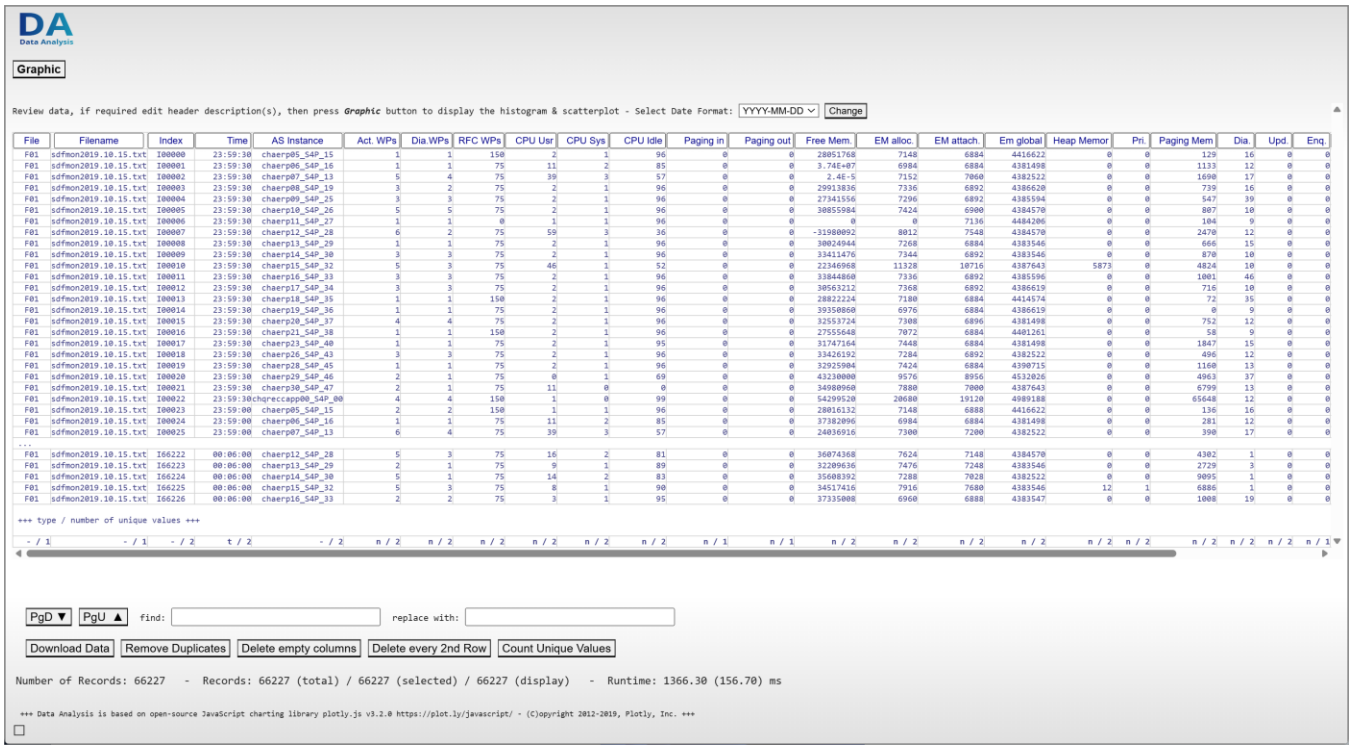



Paste content of SDFMON into text area – then use button **Update**



## 2 Review Input Data

On the second screen one can review the data (*if required specify the date format*) and modify the column headers or go directly to the graphic display.

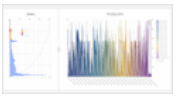



The last line in the table is indicating what type of data has been detected and how many distinct values exist in each column - the format is <type> / <number>

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| F01 | sdfmon2019.10.15.txt | I66225 | 00:06:00 | chaerp15_S4P_32 | 5 | 3 | 75 | 8 | 1 | 90 |
| F01 | sdfmon2019.10.15.txt | I66226 | 00:06:00 | chaerp16_S4P_33 | 2 | 2 | 75 | 3 | 1 | 95 |
| +++ type / number of unique values +++ | | | | | | | | | | |
| - / 1 | - / 1 | - / 66227 | t / 2891 | - / 23 | n / 69 | n / 70 | n / 32 | n / 100 | n / 19 | n / 101 |

The <type> indicates:

- **-** text value
- **d** date
- **t** time
- **n** numeric



The user has some options to manipulate the input data *(the buttons are only available if the user is putting the cursor into the header line of a column*)

- Find and/or Replace values
- Sort Data
- Delete Values
- Split or Delete Columns
- Perform basic calculations like

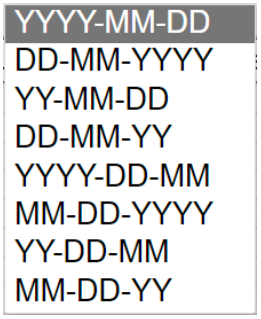
  = <value>|$nn [ # [<value>|$nn]] or

  # <value>|$nn

  where # can be any of the operators +-*/

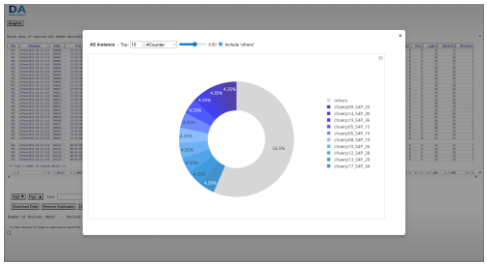
  $nn referes to the content of column nn

In some occasions the program might not be able to determine the date format automatically, in that case the user can choose from different date formats



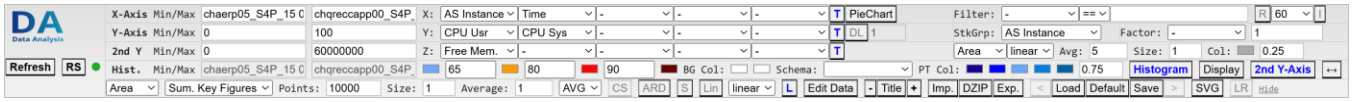

Additionally, the user can download the data (tab separated text file with extension .xls). To reduce the amount of data, columns which do not contain significant information (all values identical) can be removed. If files are too big then the user can also remove every 2nd data row.

When selecting the header line of a non-numeric column you can use button Pie Chart to generate a Pie Chart showing the data distribution of the Top nn entries



# 3 Customize Graphic Output

Below the various options to customize the graphic output




1. **Specify the fields for the X-Axis. Available are non-numeric fields (*date*, *time*, *server names*...) and all numerical fields**

2. **Specify the fields for the Y-Axis. Available are only numerical fields (*integer* or *float*)**

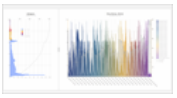
   *Important: Certain fields for the X-Axis like Date/Time are automatically proposed - for the Y Axis the DataAnalysis will default CPU User, CPU System, Resp.Time...*

   **Optional**

3. Activate 2nd Y-Axis and specify fields, chart type...
4. Select the chart type - available options are:
   - *Scatter Plot*
   - *Line Chart*
   - *Bar Chart*
   - *Area Chart*

5. If multiple key figures are select for the Y-Axis specify if those values should be summarized, stacked over each other, or shown independently
   - *Summarize Key Figures*
   - *Stack Key Figures*
   - *Independent Key Figures*

6. Select if any key figures should be used to group/stack the data. Each different value for the selected Stack variable will be shown in a different color. This option is only available if there are multiple different values available. A graph can only be stacked with this option if for each x-value there are an equal number of y-values available. If any x-value will have a different number of y-values this x-value will not be shown (*this might happen if x-values (time) differ by a few seconds - as workaround one can try to activate rounding*).



# 4 Additional Options

In the upper left you see the actual status of the Chart and you can Restart, Refresh or Reset the Graph

- **Restart**
  If you click on the Logo itself you can restart DA Data Analysis to load another data source.

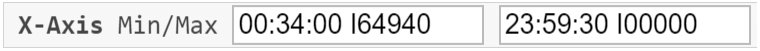
- Recalculate and Refresh the Graphic
  **Refresh**

- Reset Graphic and recalculate Axis scaling
  **RS**

- **Status**
  Ready ● Refresh Pending ● Busy ●

## 4.1 1st line - X-Axis


- X-Axis Min/Max – Specify the X-Axis Range

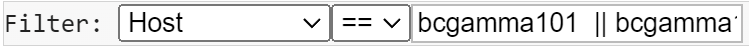
- Select fields for X-Axis (Show/Hide Data Labels)

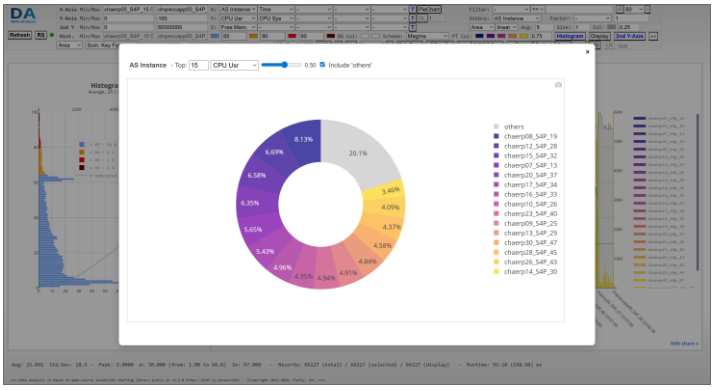
- Show/Hide X-Axis Labels (blue = Show, gray = hide)

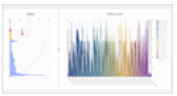
  
  

- Filter by selected x-value (**==** x-values equal selected values; **<>** x-values not equal selected values; multiple conditions can be combined with **||**, wildcard * allowed)

- If the selected field for the X-Axis is non-numeric a user can generate a Pie-Chart to show the Top nn values.





- Round time figures (select **R** to activate) to multiple of 1*, 5, 10, 15 … 60 … 3600* seconds

  R 60

  Available Rounding Values are (in seconds)

  1
  5
  10
  15
  20
  30
  60
  120
  300
  600
  900
  1200
  1800
  3600

- Specify if data points not present in all stack groups should be added or eliminated

  I A E

| | |
|---|---|
| I | Ignore data points which are not present in all stack groups |
| A | Add missing data points with zero value |
| E | Erase Data point from all stack groups if not present in all of them |

## 4.2 2nd line - Y-Axis


- Y-Axis Min/Max – Specify the Y-Axis Range

  **Y-Axis** Min/Max 0 50

- Select fields for Y-Axis (Show/Hide Data Labels)

  Y: CPU User% CPU Sys% - -

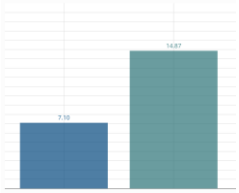
- Show/Hide Y-Axis Labels (blue = Show, gray = hide)

  T T

- Data Labels (Bar Charts only)

  DL 2

  Show data Labels above Bars with n decimals

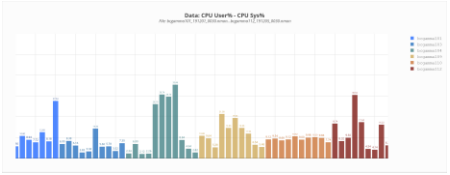


- Stack Group

  StkGrp: Host

  Data Points can be grouped by non-numeric fields as a separate data set, Each Stack Group will show in a separate color which will be interpolated from the four colors defined in the 4th line

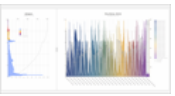
  **Example**
  X-Axis: Host + Time
  Y-Axis: CPU Usage; Stack Group: Host
  Colors:



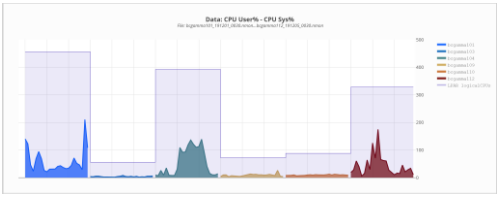
- Factor – we can multiply the Y-Values with the value of another column.

  Factor: LPAR logicalC 0.01

  A useful example is to multiply the CPU% with the total number of logical CPUs to display the logical CPUs used. We add here an additional factor of 0.01 to adjust for the percentage itself.



**Example**
X-Axis: Host + Time
Y-Axis: CPU Usage; Stack Group: Host
Factor: logical CPUs * 0.01
2nd Y-Axis: logical CPUs



## 4.3 3rd line - 2nd Y-Axis


- 2nd Y-Axis Min/Max – Specify the Y-Axis Range

- Select fields for 2nd Y-Axis (Show/Hide Data Labels)

- Show/Hide 2nd Y-Axis Labels (blue = Show, gray = hide)

- Details for 2nd Y-Axis

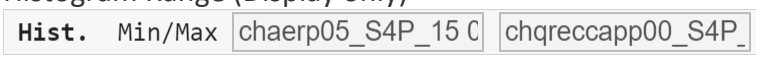
  
  

  Here we specify how the 2nd Y-Axis is drawn. From left to right

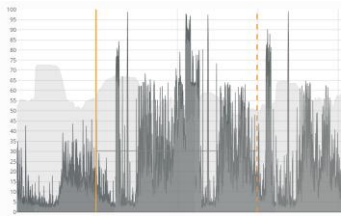
  - Chart Type (Area, Scatter, Line)
  - How to connect Data Points (line, spline, steps)
  - Average n data points
  - Point/Line Size
  - Point/Line Color
  - Transparency (from 0.05 to 1.00)

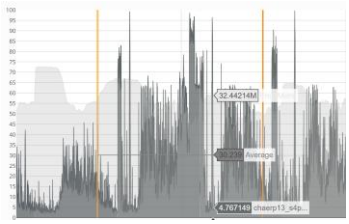
## 4.4 4th line – Histogram and Colors


- Histogram Range (Display only)

  Single click on the Chart to set first the left border (the right border will show in a dashed orange line), then click again to set the left border.

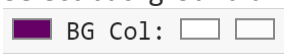
  

- Specify data range and color for the histogram (3rd line)

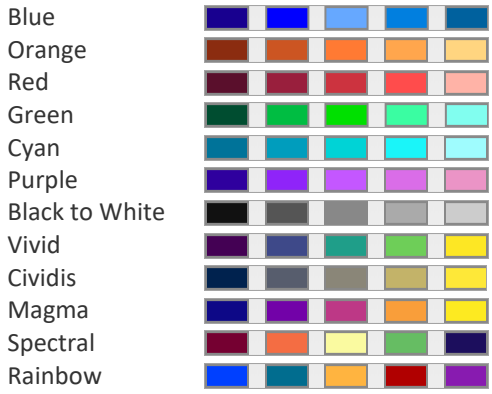
- Select background and chart color

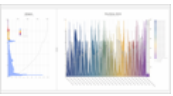
  
  

- Color Schema

  available color options are

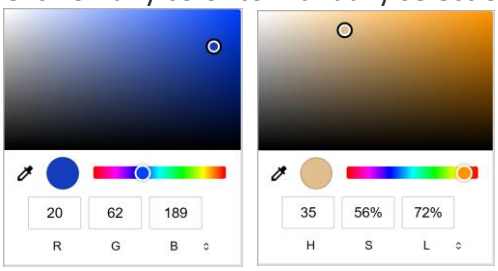
  
  



- Specify color for the 1st, 2nd, 3rd and 4th select Y-Value and Transparency. If a stack/group variable is used this for color will specify the color range used.

  Click on any color to manually select or choose the color picker

  The transparency ranges from 0.05 (almost invisible) to 1.00 (opaque)

- Show/Hide the histogram

- Display Data used in Chart

- Enable/Disable 2nd Y-Axis

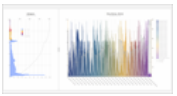
  
  

- Swap primary with 2nd Y-Axis

- Switch to 3D Display



## 4.5 5th line - Chart Options


- Select Chart Type - available Chart Types are:

  Bar
  Scatter
  Line
  **Area**

- Handling of Data Points. If multiple key figures are select for the Y-Axis specify if those values should be summarized, stacked over each other or shown independently

  **Sum. Key Figures**
  Stack Key Figures
  Indep. Key Figures

- Maximum number of points to be drawn - if the data contains after aggregation **m** points and only **x** point should be drawn then only every **n**th = **m/x** points will be drawn

  Points: 800

- Specify point size (*also line width in points, default = 3*)

  Size: 1

- Select how many data points should be aggregated (*calculate average of **n** points*)

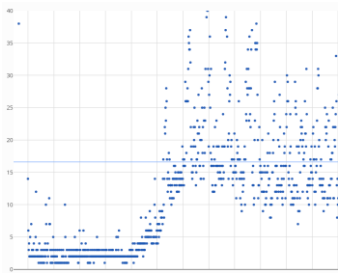
  Average: 1

- Specify calculation type if multiple data points are cumulated - default is AVG (average), available options are:

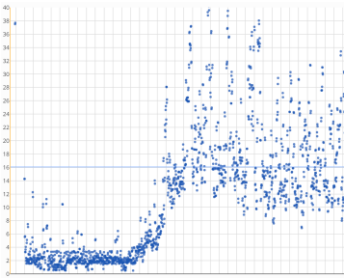
  SUM
  **AVG**
  MAX
  MIN

- Add random decimals to prevent banding.
  ARD

  If the measurements like CPU usage is only available with limited decimals values the graph will look unnatural because all values are on discrete lines. With button S one can add to each key figure a random value r between -d/2 < r < d/2 where d is the smallest detected difference between the y values. See below Example



*CPU values from SDFMON do not have decimals*



*Random decimals added (here between +/- 0.5)*

- Reduce Number of Data Points (Algorithm)
  S T
  **S** Skip every n'th data point
  **T** *Largest-Triangle-Three-Buckets*
  *(see* sveinn-steinarsson/flot-downsample: Downsample plugin for Flot charts. (github.com))

- Switch between linear or logarithmic Y-Axis
  Log Lin

- Specify how data points get connected in a line chart (linear, spline, or steps)
  linear
  spline
  steps

- Display or Hide Chart Legend
  L

- Edit/Preview Data
  Edit Data

- Set Chart & Axis Title (*supported are the modifiers:* ** Bold **, *<I> italic <I>*, <sub> subscript </sub>, <sup> superscript </sup>, and line breaks <br> ) and increase or decrease the font size
  - Title +

- Import/export Layout Profile or Export Chart

  Imp. DZIP Exp.

  The two buttons IMP. and Exp. allow to import or export a JSON which contains all chart settings. Those settings can be used when another data source is loaded.

  The button DZIP is generating a compressed JSON file which contains the chart settings and the data itself. Typically the DZIP file is around 20% of the original file size.

- Load/Save Settings set defaults. Undo/Redo last changes.

  < Load Default Save >

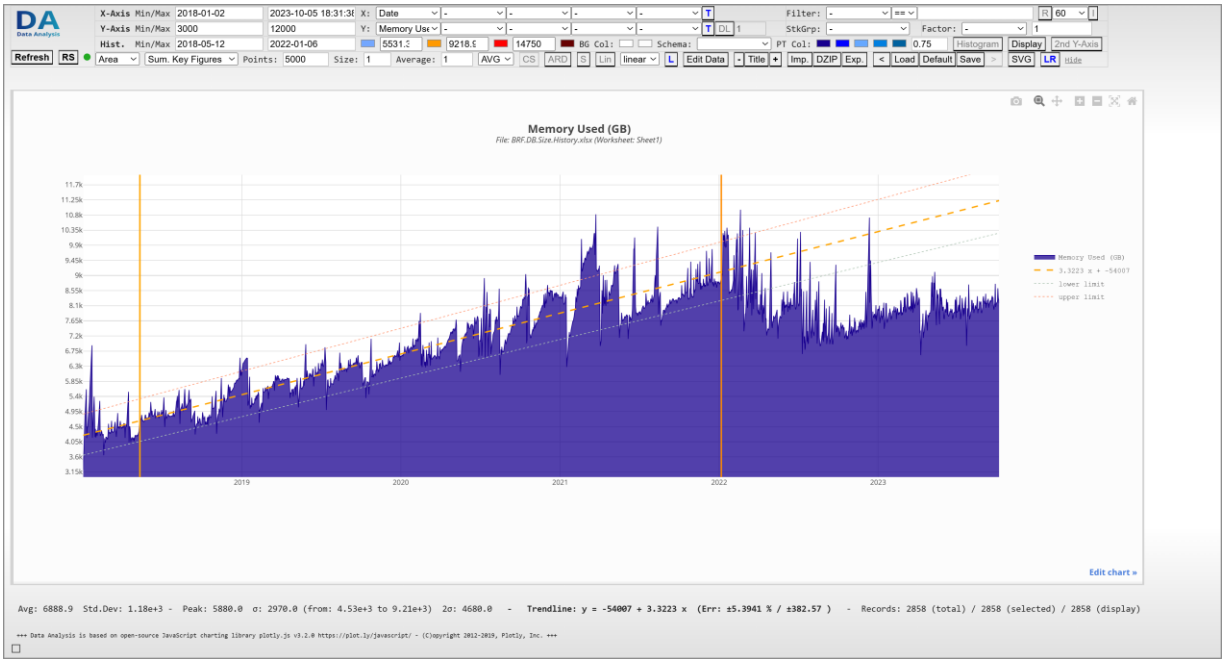
  The buttons Load and Save import/export the current chart settings into memory. They can be applied to the next file loaded. The two buttons < > allow to undo or redo the last changes.

- Save Chart as SVG Vector graphic

  SVG

- Create linear regression based trendline (only if X-Axis is Date/Time)

  LR disabled
  LR calculate trendline between delimiters
  LR calculate extended trendline

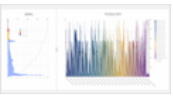



The trendline will indicate the forecast and upper/lower limit; the equation is shown below the graphic. The factor is always measured in xxx per day.

```
Trendline: y = -54007 + 3.3223 x  (Err: ±5.3941 % / ±382.57 )
```

- Autohide the chart options (that will make the charts bigger if the cursor is moved over them, the menu is shown again if the cursor is moved to the top)

  Hide Show



# 5 Plotly Options

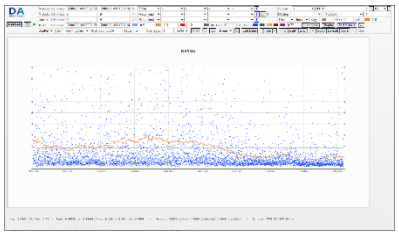
The below options are visible when you hover the mouse above the graph

- **Zoom** - use the mouse to select a rectangular area to be shown
- **Move** or **Re-Scale** axis (*move mouse to the end of the axis scales until an arrow is shown, then lets-click and move*)
- **Download** Graph as PNG file
- **Autoscale** / Reset Scale
- **Export** and open chart in Plotly Chart Studio
- Change scaling of X-Axis (double click on graph) – A popup allow to enter the tick distance (in hours, minutes, seconds…)

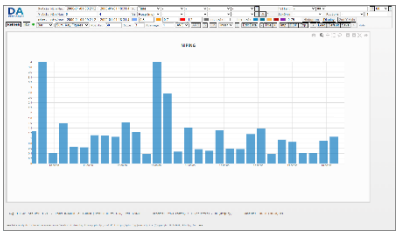

- If multiple values are shown, click on the legend to show/hide individual key figure or stack values

# 6 Examples

Available graph options are:

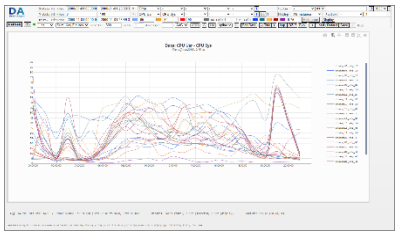


*Scatter plot*

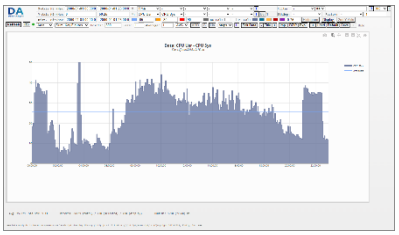


*Bar chart*

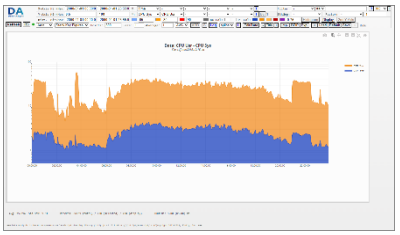
*Line chart*



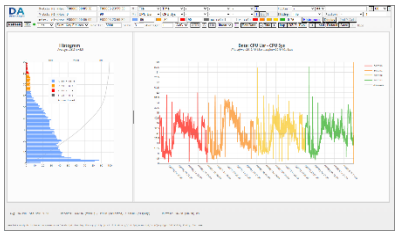
*Multi Line Charts*



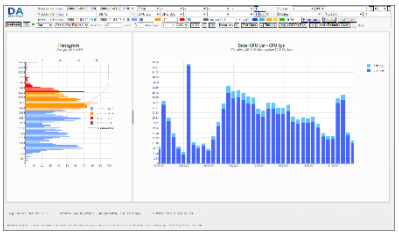
*Area Charts*



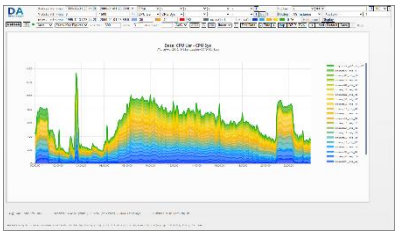
*Stacked Area Charts*



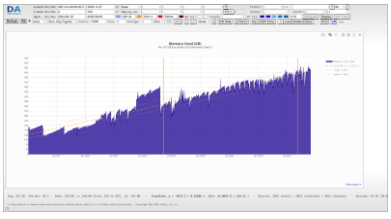
*Grouped Line Charts (SDFMON per Date)*



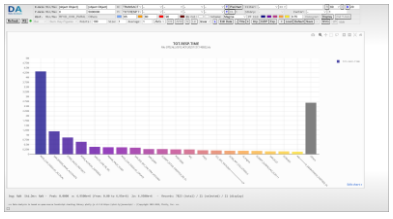
*Stacked Bar Charts (SDFMON per Appl.Server)*



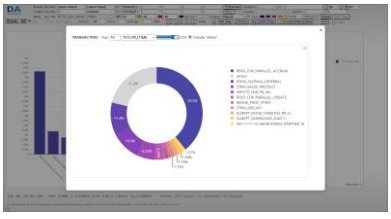
*Stacked Area Charts (SDFMON per Appl.Server)*



*Trendline*



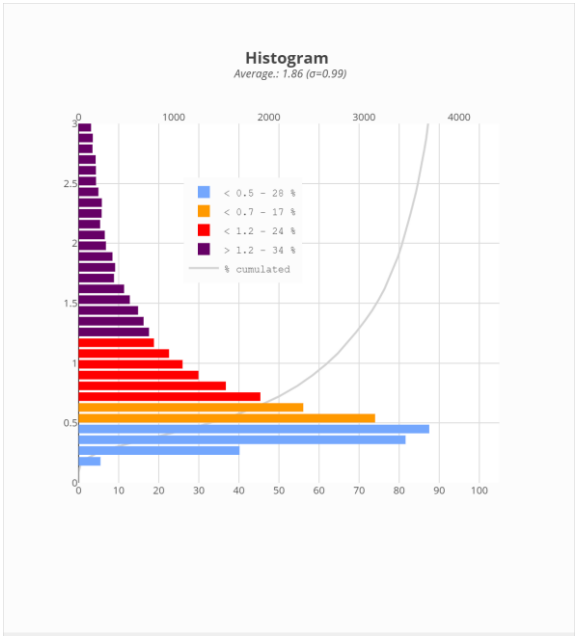
*Top n (descending) for Bar Charts only*



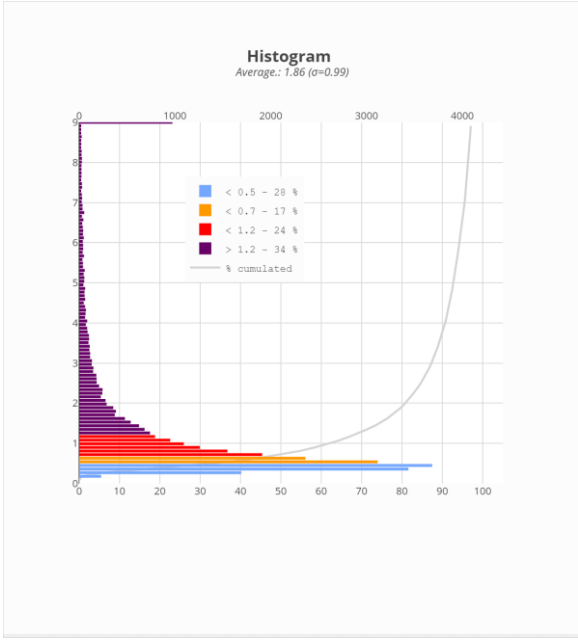
*Top n Pie Chart (only for non-numeric fields)*

# 7 Histogram

The histogram is calculated for the specified y-axis range divided in 100 equal intervals. For each of those intervals we count how many measurements are recorded and display the result as a horizontal bar chart. The color coding can be specified in the top of the screen:




*Histogram (Y-axis from 0 to 3 ms) - Peak at 0.45 ms with overall average of 1.86 ms and a sigma value of +/- 0.90 ms.*



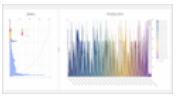

*The same data but with larger scale for the y-axis*

*The gray line is indicating the cumulative percentage value of the measurements - we see that 50% of the measurements are above 0.72 ms and 20% of the measurements are above 1.98 ms.*

Especially for highly fluctuating data (*NIPING results, CPU usage, number of active work processes*) a line/bar or scatterplot alone are difficult to interpret and often lead to wrong conclusions. A quantifiable metric to describe the difference in fluctuations is not available on those charts.

***Only a histogram analysis allows to exactly describe the random data distribution.*** *Especially for sizing purposes a statement like "95% (2 sigma) of all measurements are below 75 of CPU usage" is much more accurate than the standard deviation a line drawn above the highest observed values. Scatterplots combined with a histogram show the reality, how the individual measurements are fluctuating and at the same time to quantify the fluctuations, distribution and upper lower limits in percentage to the number of recorded measurements. Using maximum/minimum or averages do* ***NOT*** *show the reality and should* ***NOT*** *be used in the contact of capacity analysis. Assuming the histogram follows a gauss normal distribution then we can conclude that:*

- *In the interval of the deviation ±σ of the expected value, 68.27 % of all measured values can be found,*
- *In the interval of the deviation ±2σ of the expected value 95.45 % of all measured values are found,*



# 8 Performance

Depending on the file type, record length, number of columns the Data Analysis tool can load and parse approximately *140.000* records per second. The throughput of the graphic engine depends on the number of operations and fields but typically Data Analysis can calculate and draw around *15.000* measurement points per second. We can estimate the average total processing time ***T*** of the Data Analysis tool by ***T** (ms) = **R**/1000 * 7 + **P**/1000 * 60* where ***R*** is the total number of records loaded, ***P*** the number of data points in the graphic.

**Example(s):**

- We process 5 NIPING files (*with 1 sec frequency*) which contain a total of *432.000 records* - from those records we plot *20.000 data points*. The estimated total processing time will be at *432 * 5 + 20 * 60 = 4224 ms (4.2 seconds)*.
- A single NIPING file of 86.400 records with 10k data point will be ready in 0.60 + 0.60 = 1.2 seconds.
- We open 30 nmon files *(24h capture, snapshot very 5min, 57 CPU cores, each file with 150.000 records*) with the nmon analyzer to extract CPU_ALL, PROC, LPAR, MEM, MEM_NEW then each file takes about 1 minute of processing time. If we merge all files together with the nmon analyzer the process takes around 30 minutes. The Data Analysis tool can load and display the same data in less than 10 seconds.

# 9 Memory Usage:

The memory usage of Data Analysis tool depends on the file size and number of data points to be displayed. When we open a SDFMON file with 1.586.920 records the initial memory consumption reached 5GB of memory. During the creation of the graphic the total memory consumption of the browser (opera) increased to 11.5 GB of memory drawing all 1.586.920 data points (runtime 75 seconds). When we limit the number of data points to 200.000 the memory consumption reached only 2.1GB. The memory consumption scales with the number of data points drawn in the graphic and can be estimated at about 5-10 kB per data point.

# 10 Errors / Bugs

In case of any errors, please send an email to c.weyd@sap.com and include a copy (*or link*) to the used data source and a screenshot of the browser window (*including the used settings in the top*). Please also describe which browser and operating system was used.