Headroom projection of the database CPU

By Karl Arao

# Some background

I have provided some detailed supplemental content on the appendix section. The items show the previous work/research that has been done. Reading on the documents/scripts will provide the needed foundation to get familiar with the data set and why the need for such a capstone project.

To get to the point of why is there a need for “headroom calculation” read/view the following:

./capstone/04 TheArtOfScalability2ndEd/conversation with Marty.docx

* Marty’s clarification on how he does headroom calculation

./capstone/05 Capacity planning at Exa-Scale talk/capacity planning at exa-scale.mp4

* Rishi’s talk explaining how we did the headroom projection on the previous project

To get an end to end view of my research so far. Read on the appendix section in chronological order (01-05).

* The research started with [mining the AWR data](#_01_-_Mining) and using it for troubleshooting, visualization, and capacity planning. I’ve been using the collector scripts for a long time and have been pretty useful on a lot of customer engagements.
* Knowing the AWR data pretty well, I got really interested in Statistics and read this book called “[Statistics without tears](#_Statistics_without_tears)” and “[Forecasting Oracle Performance](#_Forecasting_Oracle_Performance)” which resulted to the tool called [r2toolkit](#_02_-_Linear).
* Then I started to do a lot of capacity planning and sizing. The [unpublished doc snippet](#_03_-_Expert) of the Exadata Book 2nd Ed explains the consolidation, sizing workflow, and headroom projection.
* The headroom projection concept I used for a while was based on [Marty Abbott’s work](#_04_-_The). Although the concept was pretty solid and useful I needed something more definitive and a headroom projection method that’s based on the raw workload numbers.
* Just last year I worked on this [project with Rishi Khan](#_05_-_Capacity). He made use of his previous work on the headroom projection requirements for this customer. He used Monte Carlo to run through the raw performance data.

The capstone project will just focus on the CPU data and specialize on the core functionality of the “headroom projection” and also explore the different and better ways of doing it using other models. The ultimate end goal is to put all the previous work together with the capstone project into one integrated tool or product.

# The capstone project

## Capstone questions

### What is the problem you want to solve?

The monitoring systems available in the market for database workloads mainly focus on visualization (presenting the raw metrics), troubleshooting, and alerting/adaptive thresholds. From what I see in the field on customer engagements there's also a need for a capacity modeling tool which presents the workload data in a **consolidated view** and on top of that being able to do **projection** for remediation (tuning), hardware buying (additional capacity) or planning purposes.

### Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn’t have otherwise?

The customer would be any IT shop that runs Oracle databases. They would use the capstone project to quantify the headroom projection of their database environments which they would most likely rely on guesswork or guesstimates without such a tool.

### What data are you going to use for this? How will you acquire this data?

The data that will be used is time series Oracle performance data coming from the AWR historical performance views (see [Mining the AWR](#_Mining_AWR_data)). A tool will be used to acquire the data set and it is called ESP which is available at this URL <https://github.com/carlos-sierra/esp_collect>.

The ESP will be executed on each database environment which will gather all the info for all the up and running databases. The tool is just a bunch of SQLs in one SQL file which will output a single CSV file ([example output here](#_Example_ESP_data)). The raw metrics in the CSV will have to be reshaped and data frames will be created for consolidated and individual views of the database components. The data frames will then be used for the analysis and projection.

### In brief, outline your approach to solving this problem (knowing that this might change later).

* Acquire and consolidate data from multiple environments
* Reshape data
* Analyze and project the workload growth of the CPU resource component (consolidated and non-consolidated) with percentile accuracy (risk allowance)
* Try other models to forecast the data and compare the quality of the forecast (pros and cons) and backtesting
  + The projection could be as short as 1-3 months or as long as 6-12 months in the future from the current date
* Produce a report (Visualizations)
* Additional features that can be done
  + Add an artificial multiplier, let's say if there's a business forecast of 2x growth (not seen on the raw data)
  + Store the daily forecast results to a database
  + Visualize the daily forecast results in a consolidated view that's useful for the customer
  + Automate the end to end workflow

### What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.

* Code
* Demo
* Paper
* Slides

# Appendix

## 01 - Mining AWR data

./capstone/01 Mining the AWR/Mining the AWR Repository for Capacity Planning and Visualization\_v2.pdf

* The research paper

./capstone/01 Mining the AWR/MiningTheAWR-KarlArao\_final.ppt

* The slides

./capstone/01 Mining the AWR/run\_awr-quickextract

* The AWR mining scripts

## 02 - Linear Regression

./capstone/02 Linear regression/r2project.docx

* Some initial documentation on how the Linear Regression tool works

./capstone/01 Mining the AWR/MiningTheAWR-KarlArao\_final.ppt

* The slides 69-97 shows the Linear Regression in action

./capstone/02 Linear regression/r2toolkit

* The Linear Regression tool

## 03 - Expert Oracle Exadata 2ndEd - unpublished Consolidation and RM - snippet

./capstone/03 Expert Oracle Exadata 2ndEd - unpublished Consolidation and RM - snippet/Expert Oracle Exadata 2nd Ed - Consolidation and RM - snippetv2.docx

* The doc that explains the consolidation, sizing workflow, and headroom projection

./capstone/03 Expert Oracle Exadata 2ndEd - unpublished Consolidation and RM - snippet/ExadataProvisioningWorksheet.xlsm

* The worksheet/tool used on the sizing and headroom section

## 04 - The Art of Scalability

./capstone/04 TheArtOfScalability2ndEd/TheArtOfScalability2ndEdition.pdf

* Read the following sections of the book:
  + Chapter 11 – Determining Headroom for Applications
  + Appendix B – Capacity Planning Calculations

./capstone/04 TheArtOfScalability2ndEd/conversation with Marty.docx

* Marty’s clarification on how he does headroom calculation

## 05 - Capacity planning at Exa-Scale talk

./capstone/05 Capacity planning at Exa-Scale talk/capacity planning at exa-scale.mp4

* Rishi’s talk explaining how we did the headroom projection on the previous project

./capstone/05 Capacity planning at Exa-Scale talk/cap\_management\_v1/headroom/analyze.R

./capstone/05 Capacity planning at Exa-Scale talk/cap\_management\_v1/headroom/functions.R

* The version 1 of the core headroom calculation functionality

./capstone/05 Capacity planning at Exa-Scale talk/cap\_management\_v2/capacity-management/capacityManager

* The version 2 with everything in it and packaged as a capacity planning tool

## Statistics without tears

./capstone/xx Statistics without tears/Statistics-Without-Tears-Derek-Rowntree.pdf

* The “Statistics without tears” book

## Forecasting Oracle Performance

./capstone/xx Forecasting Oracle Performance/(\_APRESS) Forecasting Oracle Performance/Forecasting Oracle Performance.pdf

* The “Forecasting Oracle Performance” book which also showed Linear Regression on Chapter 9
* I automated this concept which resulted to the r2toolkit (The Linear Regression tool)

## Example ESP data

./capstone/xx example ESP data/esp\_requirements.zip

* ESP collection from four databases