DATABASE WORKLOAD CAPACITY PLANNING USING TIME SERIES ANALYSIS AND MACHINE LEARNING

BY **ANTONY S. HIGGINSON, MIHAELA DEDIU, OCTAVIAN ARSENE**

GEORGE MICHOULIS 82

SCHOOL OF INFORMATICS
Data & Web Science MSc Program

MOTIVATION

- Cloud Service Providers (CSPs) problem is the stability for short-, medium- and long-term resource allocation.
- Nowadays work on cloud modelling is at building on modelling of virtual machines (VMs).
- Provisioning for DBaaS presents short- and long-term challenges:
  i. **Short term** - predicting when system will run out of resources (proactive monitoring).
  ii. **Long term** - capacity planning, whether that is on-premises, off-premises or non-cloud environments.

RELATED WORK ON WORKLOAD MODELLING AND PREDICTION IN CLOUDS.

- There is no prior work for making predictions that try to answer, “**how much resource do I need?/when will I require more resource?”** type questions.
- Current works focus on: 1) prediction on VM layer, 2) provisioning exercise 3) prediction as Quality of Service (QoS), 4) prediction as part of costing design of priorities.

Workload Modeling:

- Short term adaptations or longer resource requirements.
- identified VMs that could be a problem and then taking action to adapt.
- Novel work on workload of a PaaS (Database) using Concurrent Query Performance Prediction (CQPP).

Prediction in Clouds:

- Prediction on workloads in cloud is new and the study of their elasticity and pay-as-you-go nature.
- Much work on IaaS level, but whether in terms of placement, provisioning, monitoring or where the database resides, work on the PaaS layer is scarce
- e,.g., Research on a MySQL database cluster in an Amazon EC2 cloud and used ARIMA to predict the workload.

PROBLEM DEFINITION

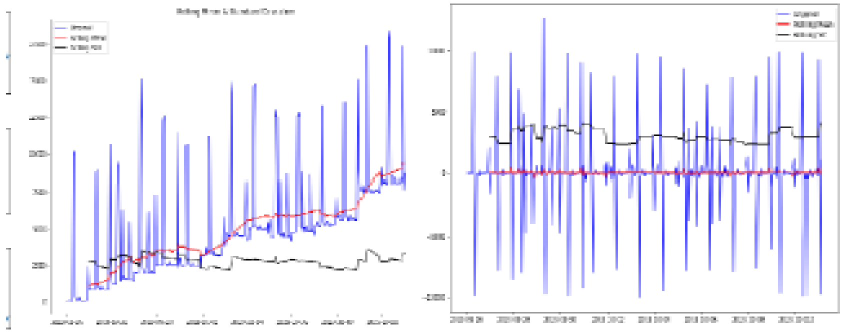
"Given a time series m that provides monitoring information about a workload w, generate a prediction z for a period following on from that of w. The prediction should account for external influences on w."

- The time series m is a trace or log of data in a time series format.
- The time series captures specific features of w (e.g., CPU, memory or logical IOs).
- The frequency of the prediction matches the frequency of m.
- The prediction z consists of the predicted values and associated error bars.

- The external influences on the prediction are events such as batch jobs and backups that routinely and sporadically occur in computational workloads, and which do not always follow a uniform pattern.

MODELS (1)

- Time Series Analysis is broken down into two main areas:
  - Time Domain - ARIMA uses techniques such as Box- Jenkins and Dicky-Fuller to detect if the data is stationary, trending or requires an element of differencing.
  - Frequency Domain - Techniques such as Fast Fourier Transform (FFT) to analyse data that is complex in a time domain.



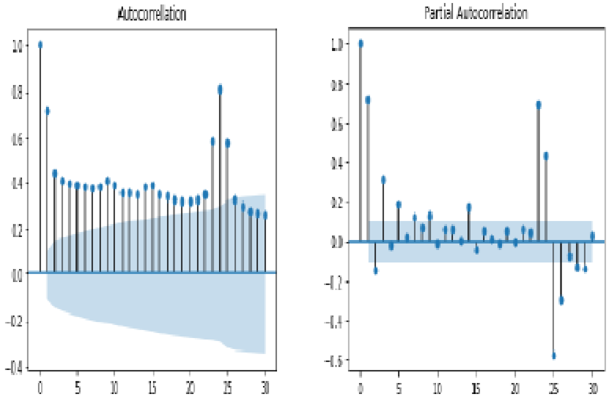
(c) Differencing Data (Trend)

Seasonal ARIMA (SARIMA), encapsulate seasonality and a class of models that capture the subtle structures of time series data. It depends on some variables:

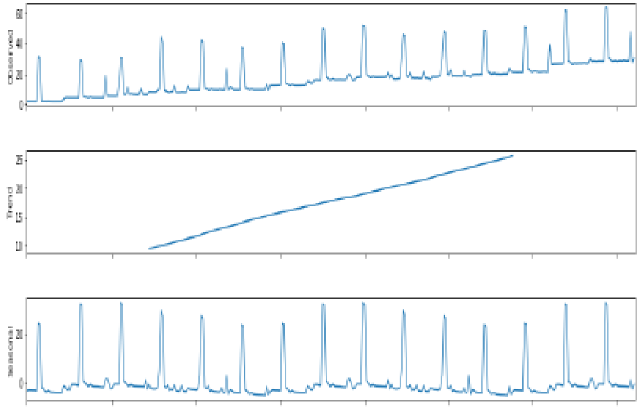
1. Autoregressive Terms (AR) - p
2. nonseasonal differences needed for stationarity utilizing Box-Jekins, for short-term monitoring. d
1. Lagged forecast errors for the prediction -q

The **Seasonality** nature depends on exogenous vars:

1. Order of the seasonal AR - P
2. Seasonal differences for stationarity. -D
3. Seasonal Lagged forecast errors -Q
4. Frequency -F



(a) Correlograms (ACF/PACF)



(b) SARIMAX Model Decomposition

MODELS (2)

Exponential Smoothing

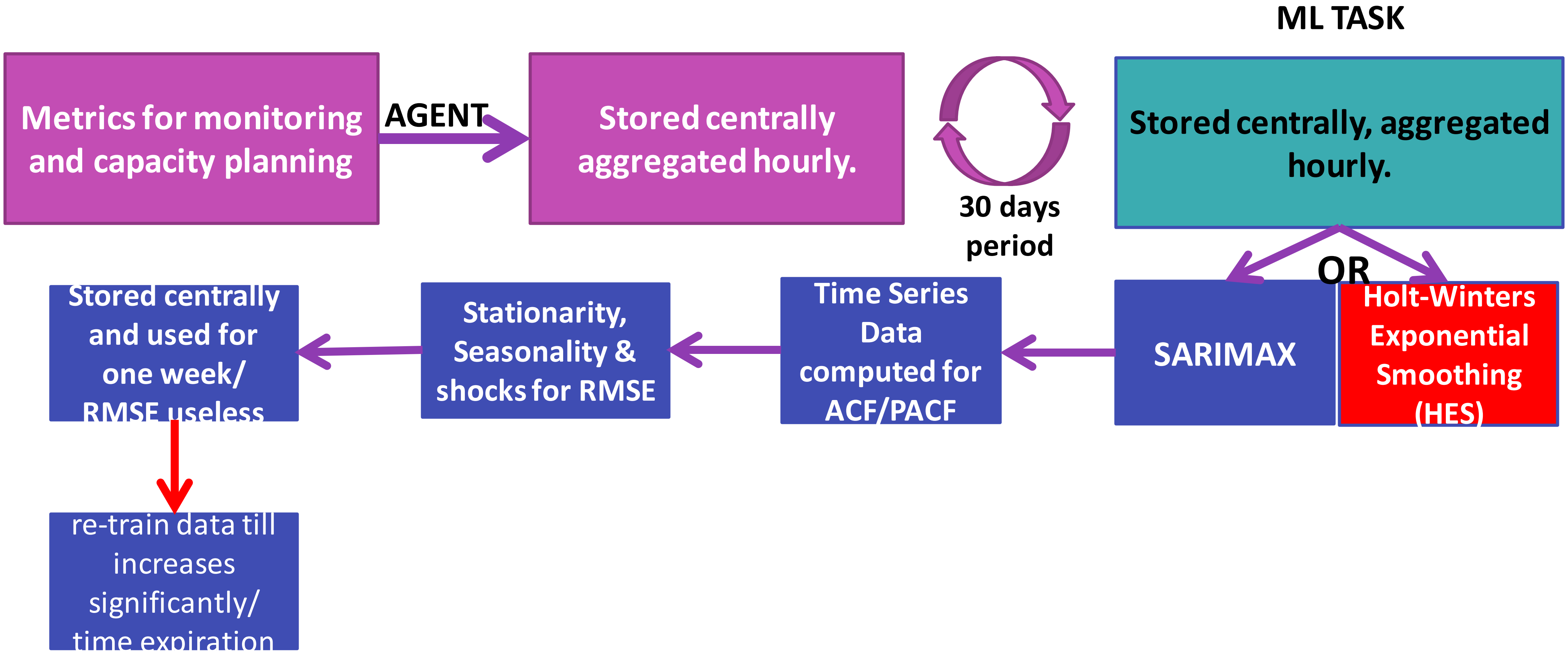
- Forecasting is based on weighted averages of past observations.
- Methods:
  - Holt's linear trend (HLT) , extended HES to allow data with trend, and
  - Holt-Winters seasonal method, extended to include seasonality. Complex seasonal patterns

Trigonometric seasonality **B**ox-Cox **A**RMA **T**rend **S**easonal, incorporating Box-Cox transformations, Fourier representations with time varying coefficients and ARMA error correction

Fourier Terms

- ❑ Fourier terms, which are used as external regressors.
- ❑ The number of Fourier terms (ki) are chosen to find the best SARIMAX parameters, and then ordered appropriately and selected based on which gives the best root mean squared error (RMSE).

APPROACH


ML TASK
Metrics for monitoring and capacity planning
AGENT
Stored centrally aggregated hourly.
30 days period
Stored centrally, aggregated hourly.
OR
SARIMAX
Holt-Winters Exponential Smoothing (HES)
Time Series Data computed for ACF/PACF
Stationarity, Seasonality & shocks for RMSE
Stored centrally and used for one week/ RMSE useless
re-train data till increases significantly/ time expiration

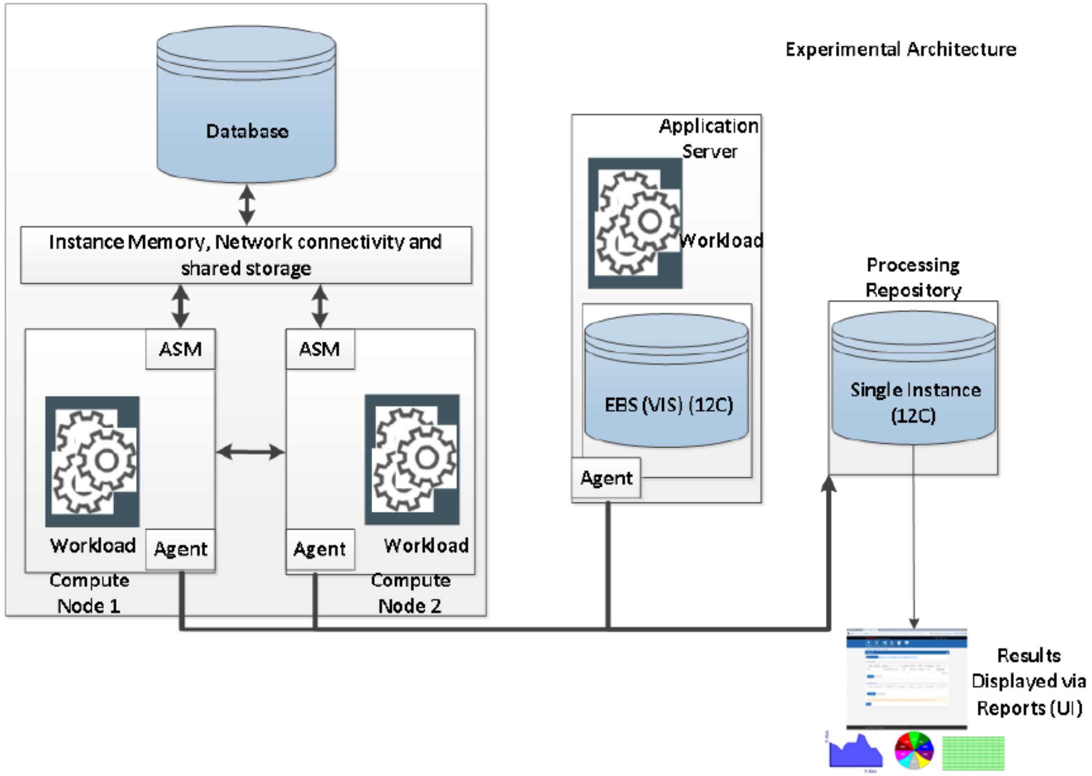


Figure 5: Experimental Architecture: Typical customer N-tier Architecture.

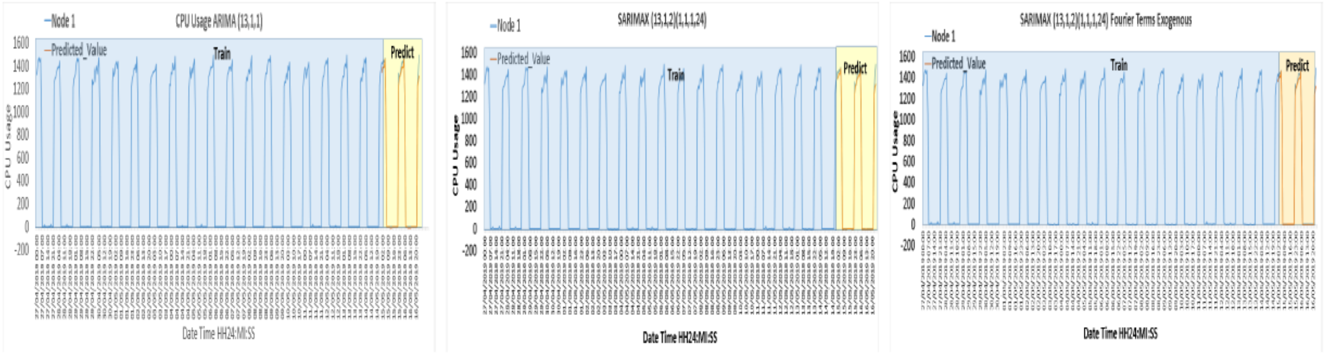
EXPERIMENTAL SETUP

- Oracle clustered database; web server application

The three techniques and the number of models are:

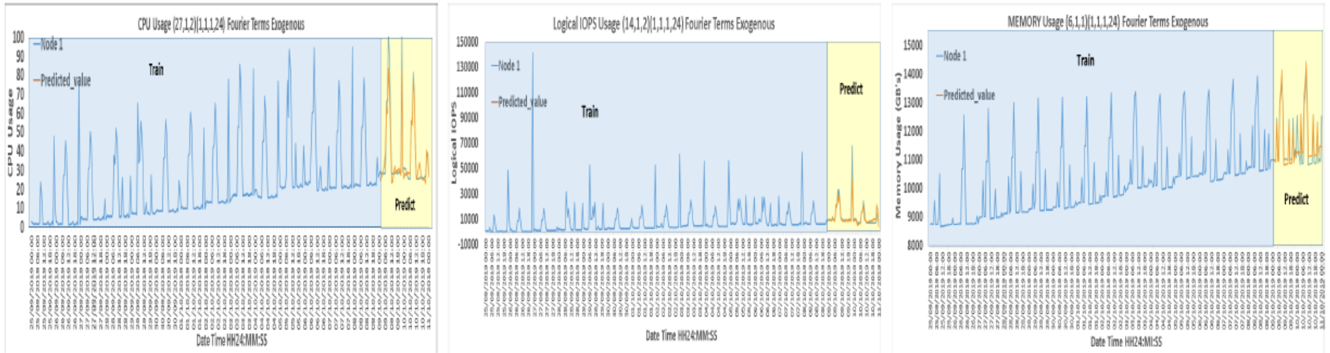
- ARIMA p,d,q = 180 models per instance (totalling 360 models)
- SARIMAX p,d,q,P,D,Q,F = 660 models per instance (totalling 1320 models)
- SARIMAX p,d,q,P,D,Q,F + Exogenous (4)+ Fourier Terms (2) = 666 models per instance (totalling 1332 models)

EXPERIMENTS



(a) Exp 1 - ARIMA (b) Exp 1 - SARIMAX (c) Exp 1 - SARIMAX With Exog and Fourier Terms

Figure 6: Experiment 1: Prediction charts Comparing Three ARIMA Techniques



(a) Exp 2 - CPU (b) Exp 2 - Logical IOPS (c) Exp 2 - Memory

Figure 7: Experiment 2: Prediction Charts Using SARIMAX with Exogenous and Fourier Terms

EXPERIMENT 1

- An OLAP workload with several users connecting across the cluster performing OLAP activities that are high in IO and execute for long periods of time.
- IO is generated via SQL activity and data manipulation language (DML).
- **RESULTS**: SARIMAX with Exogenous variables and FFT is more accurate and reduces the error across the metrics (CPU, Logical IO and Memory)

EXPERIMENT 2

- Like problem 1 using OLTP with Trends and Stationariness (monitoring per day) and also Multiple seasonality (users grow per day).
- **RESULTS**: SARIMAX with Exogenous variables and FFT is consistently more accurate

CONCLUSION

- Machine learning has greatly increased the ability of these models to be utilised on computational workloads that exhibit diverse patterns over time.
- By using this approach, now it is possible to advise users on what “may” happen to their workloads.
- Still faces challenges like a long period of fault system.
- Still needs the threshold-based monitoring.

THANK YOU!