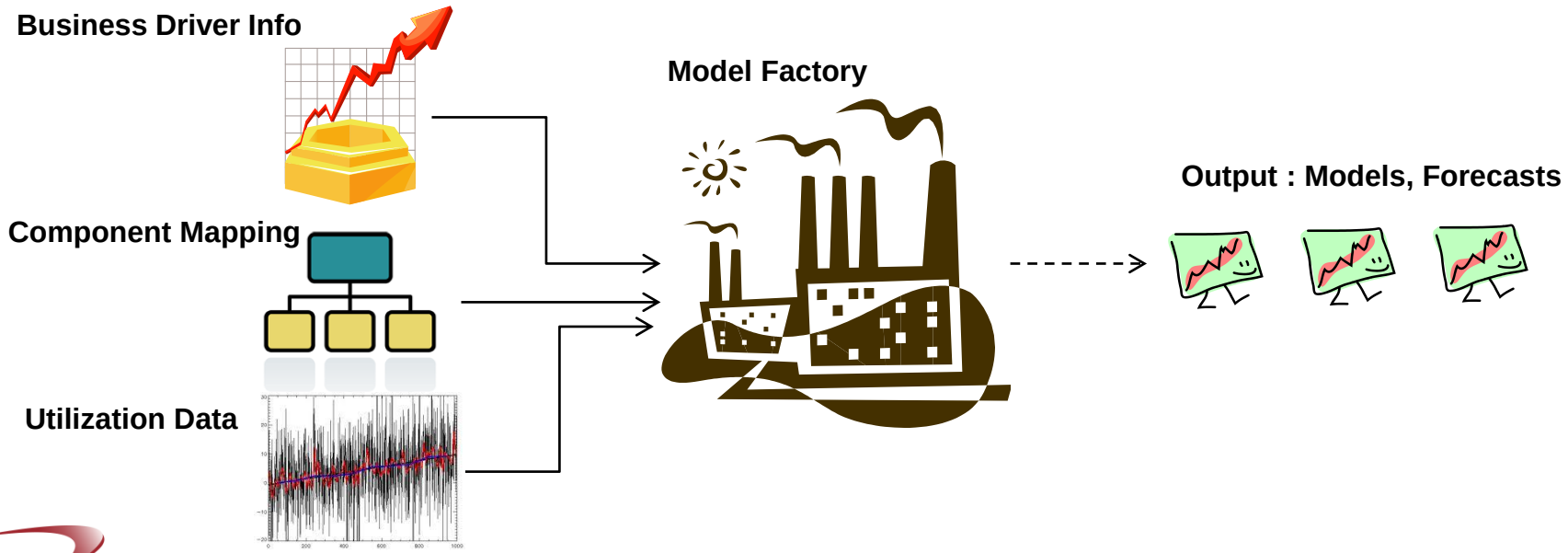
The Model Factory – Correlating Server and Database Utilization with Customer Activity

Patent application:
SYSTEMS AND METHODS FOR MODELING COMPUTER RESOURCE METRICS (US Patent Application 2016/0379143)

Igor Trubin – igor.trubin@capitalone.com
Kevin McLaughlin - Kevin.McLaughlin@capitalone.com
Mark Schutt – mark.schutt@capitalone.com

Modeling Factory Concept (Patent Application ABSTRACT)

"This disclosure relates generally to system modeling, and more particularly to systems and methods for modeling computer resource metrics. In one embodiment, a processor-implemented computer resource metric modeling method is disclosed. The method may include detecting one or more statistical trends in aggregated interaction data for one or more interaction types, and mapping each interaction type to one or more devices facilitating the transactions. The method may further include generating one or more linear regression models of a relationship between device utilization and interaction volume, and calculating one or more diagnostic statistics for the one or more linear regression models. A subset of the linear regression models may be filtered out based on the one or more diagnostic statistics. One or more forecasts may be generated using the remaining linear regression models, using which a report may be generated and provided."



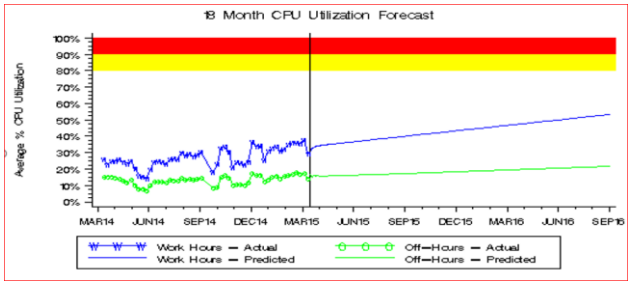
Capacity Planning 1.0

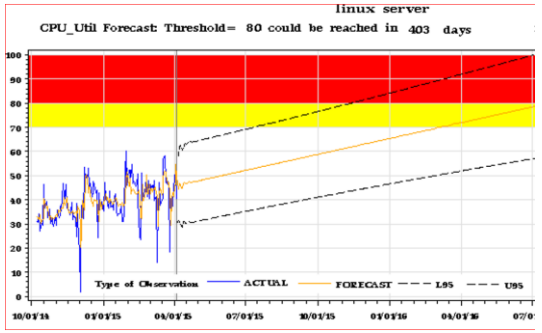
Technology Capacity Planning has traditionally relied on monitoring the infrastructure, simple trending, and reacting to trends to prevent service outages. Before the Model Factory, autoregressive models were the primary method for identifying capacity bottlenecks.

Technology Capacity Data

```
proc forecast
    data=<actual_dataset>
    interval=day lead=30
    method=<STEPAR|EXPO|WINTERS>
    trend=<1|2|3>
    out=<actual&predicted_dataset>;
run;
```

Capacity Trends





This has two notable shortcomings:

- **Technology cannot articulate bottlenecks in terms our business partners can understand.**
- **The current process is not scalable to an IT infrastructure the size of large bank.**

Business Service Modeling

Which of these statements provides our customers more value?

Q: You are running at 40-60% of CPU capacity on your servers…OR…Your systems are currently handling peaks of 4K-6K transactions per hour. They are built to handle over 10K.

A: Although they both measure the same thing, the second statement provides business context for the system's user or customer. It identifies what is driving utilization and allows business customers to make better decisions in marketing campaigns and infrastructure investments in the future.

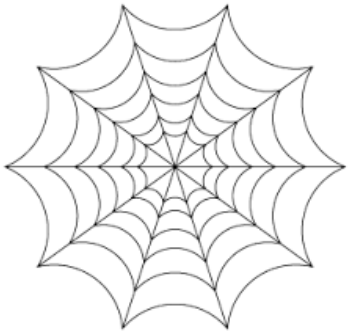
To create business service models, we need:

Independent

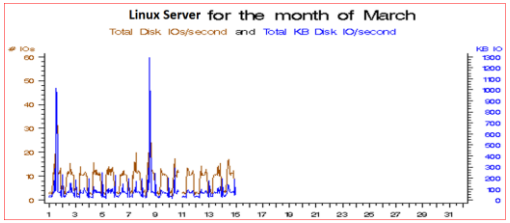
Business Demand Metrics

Over 100 customer transaction data sets in IT data warehouse.

Some way to relate the variables



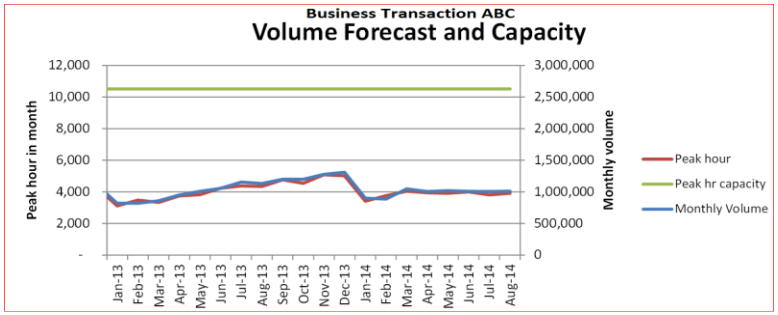
Infrastructure Utilization Metrics



Data for 18K servers and 1K databases in the IT Capacity Management system.

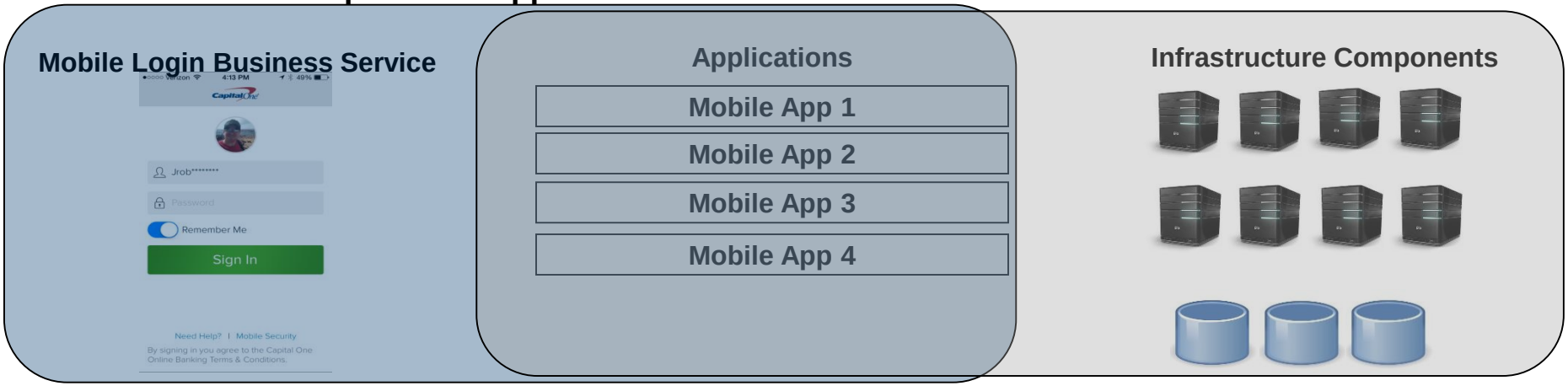
Dependent

A model with usage and threshold in business terms!



Scalability Challenge #1: Relationships and Grouping Data Sets

The first challenge in creating models for critical business services is understanding what infrastructure components support critical business services.

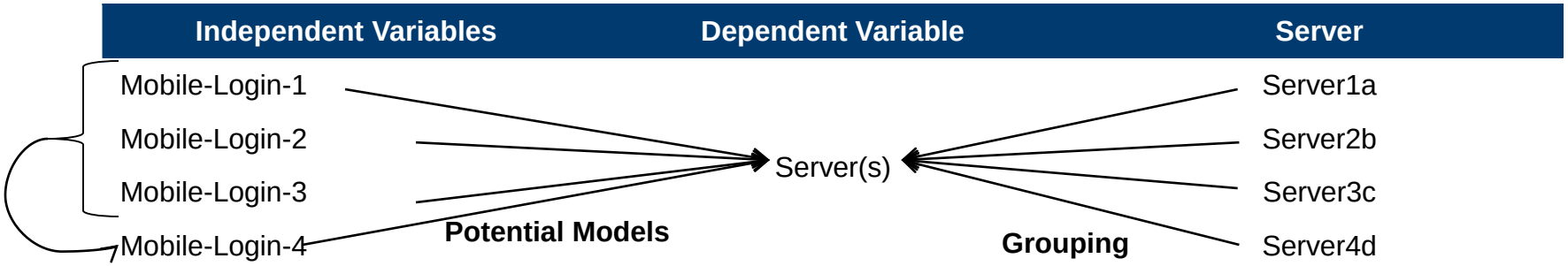


Critical business service to application mapping is captured in the IT Data Warehouse

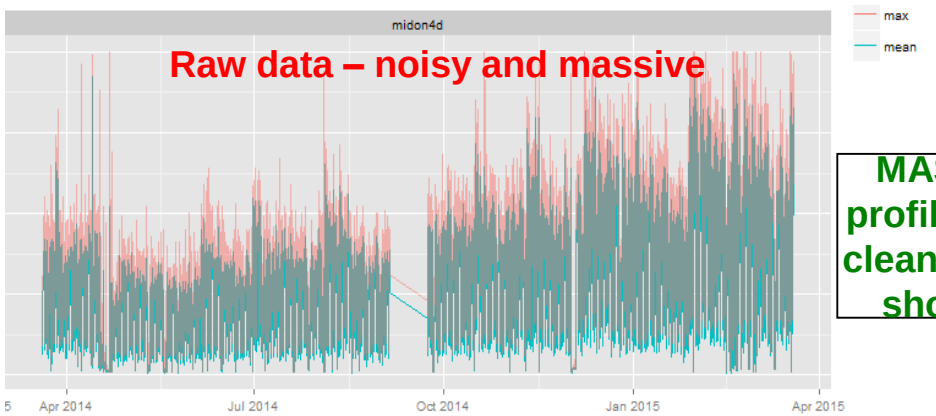
Infrastructure to application mapping is captured in the Configuration Management Database

- **We were still left with thousands of possible combinations to model. However, we know from experience that:**
  - **Business facing servers rarely operate in isolation. There are typically configured in groups of 2, 4, or 8 for redundancy and service resiliency purposes.**
  - **There are many business facing services that leverage the same applications and infrastructure**

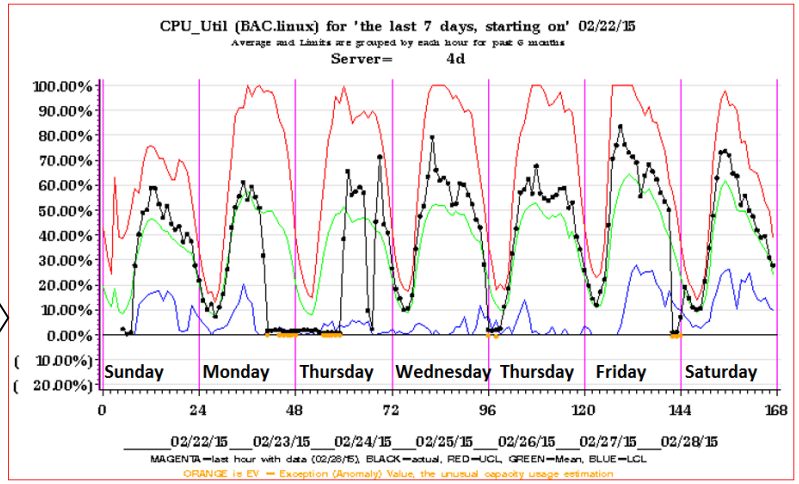
- **The Model Factory ran the process to find servers with similar names, indicating they perform a similar function, and pooled their capacity. The process correlated their utilization to confirm that they behave similarly.**
- **Business services were also grouped together to understand cumulative impact of transactions upon the supporting infrastructure.**
- **The result is a flattened table: the set of critical relationships to model for capacity constraints.**



Scalability Challenge #2: Multivariate Adaptive Statistical Filtering profiles vs. Raw Data



MASF profiles – clean and short



R-Script to build MASF Profiles against raw hourly stamped time-series data:
http://itrubin.blogspot.com/2012/03/r-script-to-aggregate-etl-to-mysql.html

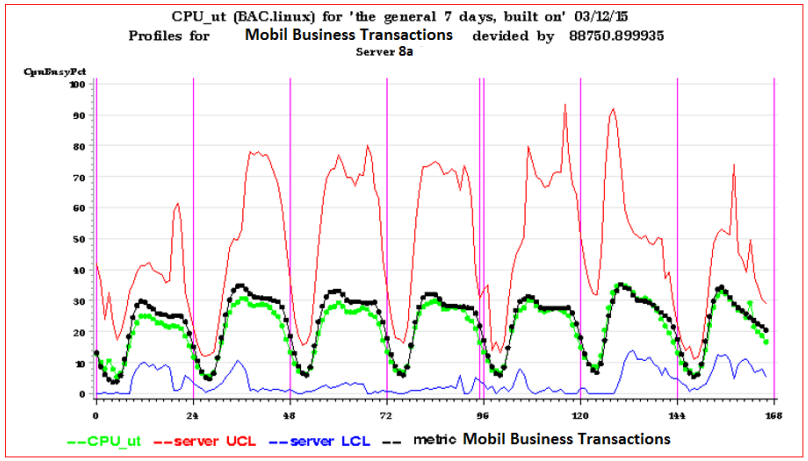
C:\Documents and Settings\Administrator\My Documents\SEDS\SETDS\R\SEDS_ETL.R - R Editor

```
# R(ODBC)-script to transform raw date-hour data (input is CpuUtil table)
# to data for building IT Control Chart (output is ActualVsHistorical table)- iTrubin 2011,2012

library(RODBC)                      # Start by requesting the RODBC library
myConn <- odbcConnect("SETDStest")  # Now create an ODBC connection object
sqlQuery(myConn,                    # Actual data
 "create temporary table Actual as select date,     (DAYOFWEEK(date)-1)*24+hour as HoursInWeek, CPUused
      from servermetrics.CPUutil where date > '2011-04-02'")
sqlQuery(myConn,                    # Baseline data
 "create temporary table Historical as select date, (DAYOFWEEK(date)-1)*24+hour as HoursInWeek, CPUused
      from servermetrics.CPUutil where date <='2011-04-02'")
sqlQuery(myConn,                    # Control Limits Calculation
 "create temporary table AggregateHistorical as select avg(CPUUsed) as AVGCPUUsed ,STDDEV(CPUUsed) as STDDEVCPUUsed
         ,CASE WHEN avg(CPUUsed) + 3*STDDEV(CPUUsed)>100 THEN 100 ELSE avg(CPUUsed) + 3*STDDEV(CPUUsed) END as UCL
         ,CASE WHEN avg(CPUUsed) - 3*STDDEV(CPUUsed)<0   THEN 0   ELSE avg(CPUUsed) - 3*STDDEV(CPUUsed) END as LCL
         ,HoursInWeek
      from Historical
      group by HoursInWeek order by HoursInWeek")
sqlQuery(myConn,                    # joining actual with baseline data Control Limits and Avarage
 "create table servermetrics.ActualVsHistorical as select a.HoursInWeek ,FORMAT(a.AVGCPUUsed,2) as AVGCPUUsed
         ,FORMAT(a.UCL,2) as UCL ,FORMAT(a.LCL,2) as LCL ,b.CPUUsed as CPUUsedActual
      from AggregateHistorical a
      inner join Actual b on a.HoursInWeek = b.HoursInWeek")
cchrt <-sqlQuery(myConn,"SELECT HoursInWeek, CPUUsedActual, UCL, AVGCPUUsed,LCL FROM actualvshistorical")
```

MASF weekly profiles are data cubes* with the two following dimensions: 30 weeks (~6 month worth) historical baseline and 168 (24*7) week hours.

In good models, the business Driver MASF profile should be consistent with server's:

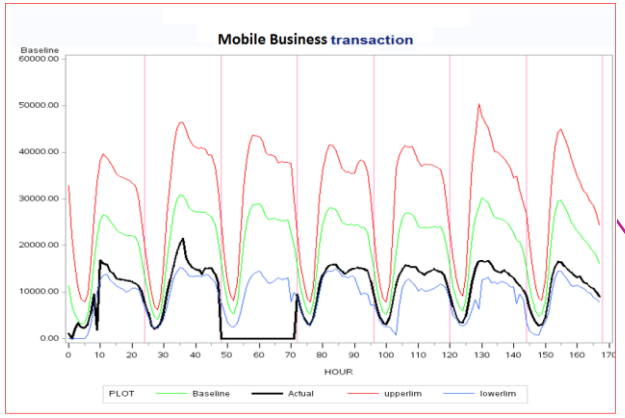


*There are two time dimensions in the MASF profiles. The hour of the week and the thirty data point for each of those hours. The third dimension is system utilization.

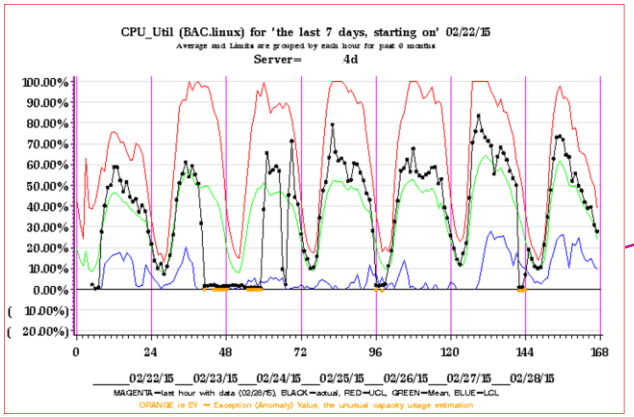
Creating Capacity Models – Mobile App Example

Related MASF Profiles

Independent Variable: MASF profile



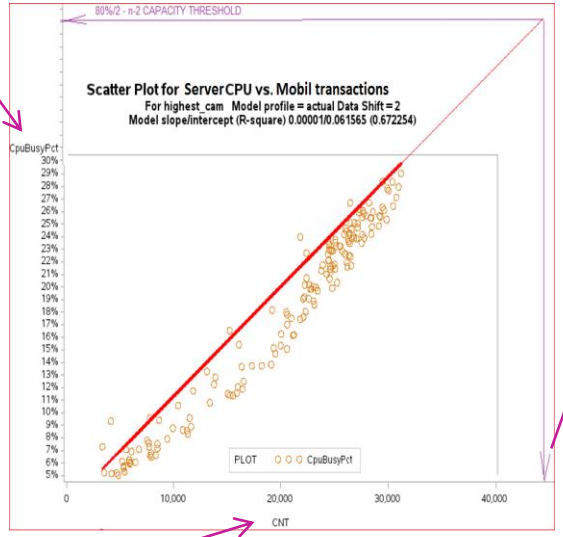
Dependent Variable: Capacity Resource (CPU) MASF profile



Model the Variables

Linear Regression

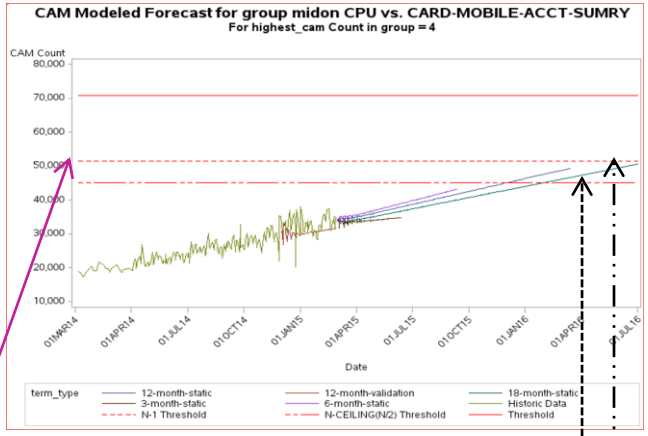
Extrapolate the line to determine at what transaction volume the capacity threshold is reached.



Capacity in Business Terms

Business Driver Capacity usage forecast

Actuals | **Forecasts***



There are multiple thresholds calculated depending on redundancy in server configuration:
N-1 – assumes one server is down/passive.
N/2 – assumes half the environment is down/passive.

Model Data Store (MDS) and Forecast Data Store

Once monthly, for each matching pair of independent and dependent variables, the Model Factory creates a linear regression model to estimate a capacity threshold. All modeling statistics and relevant variable data are stored in a SAS data set (MDS), ready for further analysis and filtering.

MDS Fields (Sub-sample)

- BUSINESS_SERVICE
- CHANNEL
- LOB
- Driver metrics rates
- server
- datasource
- subsystem
- profiletype
- shift
- OS
- Memory Util. %
- CPU Util. %
- Disk I/O rate
- Slope
- Intercept
- RSQ
- Validation
- Doubling
- Driver_Max
- Driver_capacity
- Model Score
- Date

Independent
Dependent
Model

Sample Values for Mobile APP

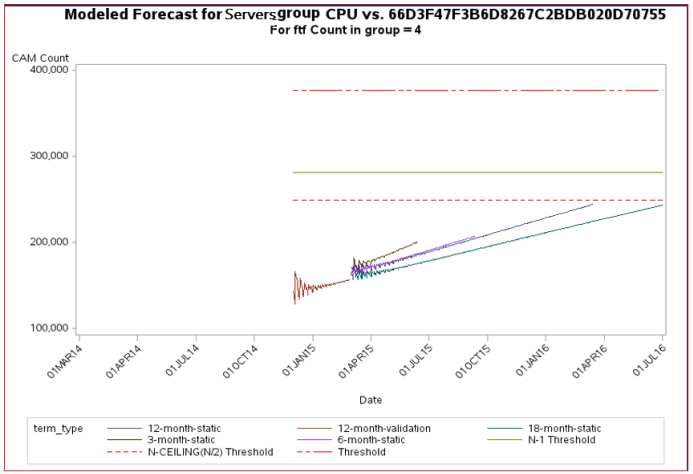
LOB	Banking
CHANNEL	Mobile
date	4/14/2015
Model_Score	(Multiple Items)

Row Labels	Max of Drv_capacity	Min of Drv_max
Account Login	64.2%	125,224
MOBILE-LOGIN	64.2%	125,224
Rewards	54.9%	1,454
MOBILE-RED	54.9%	1,454
Cards	71.9%	67,053
CARD-MOBILE	71.9%	67,053

What this means…

The Mobile traffic is currently between 55%-72% of what its IT infrastructure is built to handle. There are many models for Mobile business transactions and their infrastructure components. Model selection is an important filtering mechanism for the Model Factory.

Forecasts are critical for Capacity Management to plan ahead for business demand.



- With no reliable single source for business transaction forecasts, the Model Factory employs the Proc Reg function in SAS to create time-series forecasts for business transactions.
- We use 3 month's history for a 3-month forecast, 6 month's history for a 6-month forecast, etc. and forecast up to 18 months in the future.

Model Scoring and Selection

To filter out poor models, the Model Factory evaluates the model statistics generated in the process. R^2, slope, intercept, and root mean squared error of the models are used to score and filter models for further analysis.

Poor	Model	Good
Low	R^2	High
Negative	Slope	Positive
Negative or High	Intercept	Between 0-5%
High	RMSE	Low

Numerical Score Assigned

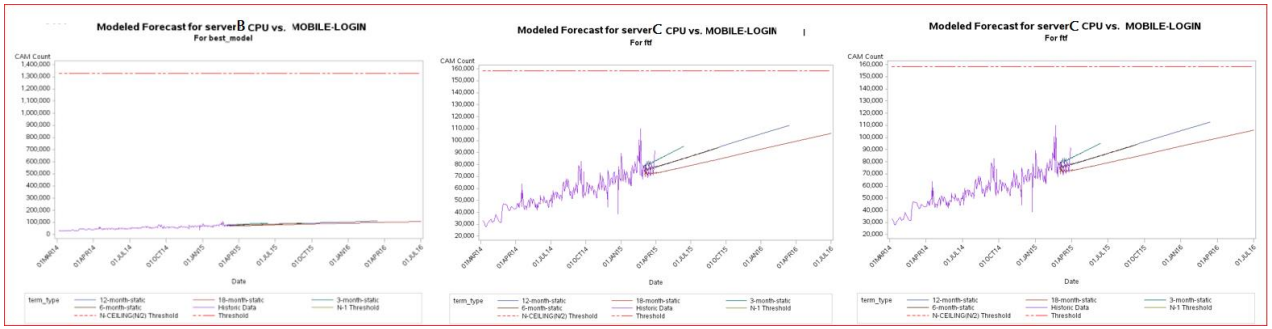
60-70K Models

508 Models

Which models do we report? For each business transaction, we publish:

- Best Model – this is the model that scores highest in the evaluation.
- Highest Utilized – from a capacity perspective, this highlights the infrastructure that is currently the most heavily utilized.
- First-to-fail – this shows the infrastructure projected to reach its threshold first based on independent variable forecasts.

Best Model | **Highest Utilized** | **First-to-Fail**



Note: The best model, highest utilized, and first-to-fail criteria can select the same model or three different models. In Mobile Banking, the first-to-fail and highest utilized models are the same. The best model is different.

Sample Output and Diagnostics

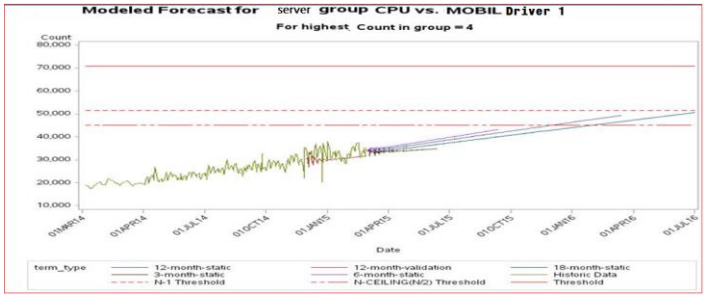
The Model Factory portal to share its models with Technology and business customers.

Capacity and Risk

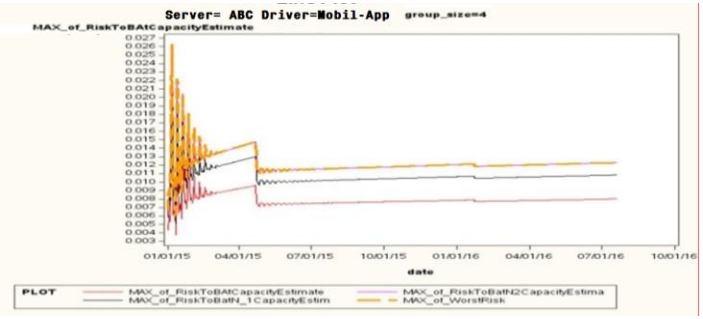
Left: shows the number of business transactions, autoregressive forecasts, and calculated thresholds.

Right: projects the risk of a capacity threshold being reached on a certain date in the future.

Business Transactions Volume with Forecast



Risk Index

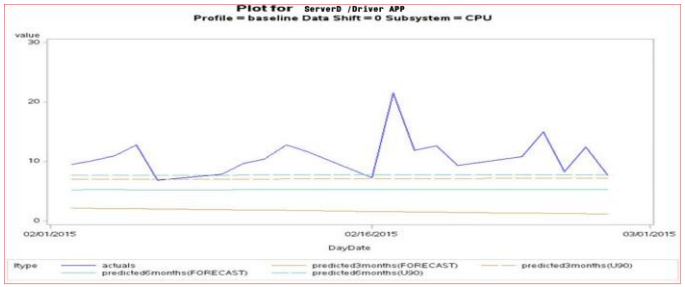


Accuracy

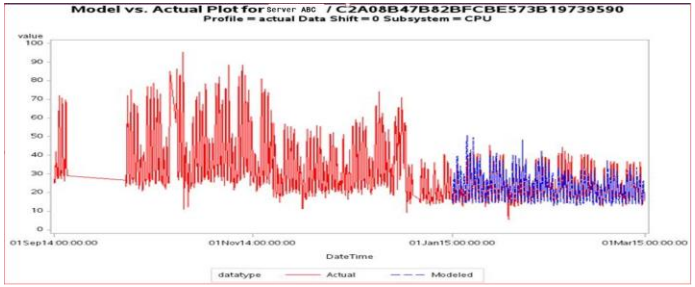
Left: for the past month, actual component utilization and forecasted component utilization.

Right: Modeled utilization (running the dependent variable through the regression formula) and actual utilization for the past six months.

Server Utilization Prediction Validation



Model Accuracy

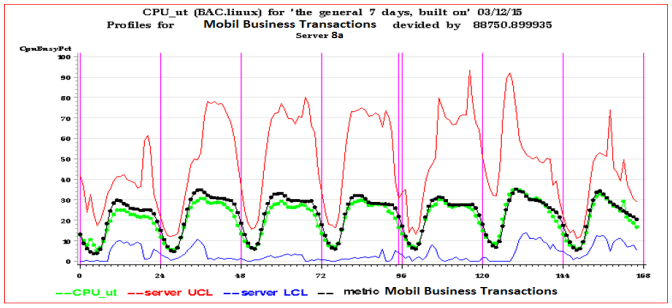


Model Output

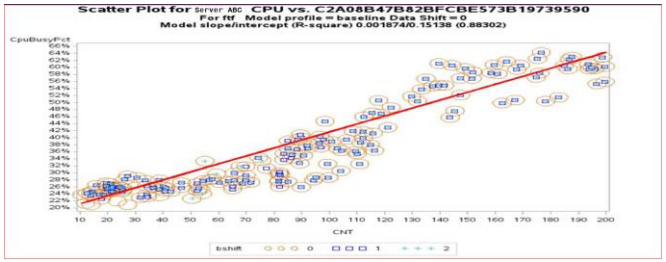
Left: side by side comparison of 168 hour MASF profiles for business transactions and component utilization.

Right: The scatterplot for the model charting component utilization against transaction volume. The regression line is shown.

Business Transactions Profile vs. Utilization Profile



Business Transaction/Utilization Scatter Plot



Business Value: Model Factory work led to a proactive finding on upcoming capacity implications for the Mobile App rollout.

Inputs

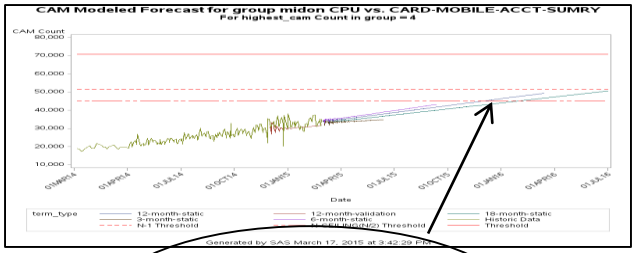
- Business Service Name and Channel
- Business Transactions
- Application Names
- Historical business transaction data
- Historical server performance data

– E.G.:

- Processing 182,674 transactions/peak hour
- >80% CPU utilization at peak hour

Output

- Capacity Models for Business Services
- Maximum capacity (number of transactions) for applications and servers
- First-to-fail servers and estimated timeline, including N-1, N/2 scenarios



Models show that the server cluster will be reaching its max capacity by end of 2015.

What else: In a shared infrastructure environment, the Model Factory looks into multivariate and non-linear models. This better reflects the reality of multiple business transactions sharing common infrastructure.

REFERENCES

1. I Trubin, M Schutt, J Robinson: "***SYSTEMS AND METHODS FOR MODELING COMPUTER RESOURCE METRICS***", US Patent 20,160,379,143, 2016 - pending.

2. Igor Trubin: ***"Exception Based Modeling and Forecasting"***, Proceedings of the Computer Measurement Group, 2008.

3. Igor Trubin: ***"AIX frame and LPAR level Capacity Planning. User Case for Online Banking Application"***, Proceedings of the Computer Measurement Group, 2012.

4. Jeffrey Buzen and Annie Shum: "***MASF - Multivariate Adaptive Statistical Filtering***", Proceedings of the Computer Measurement Group, 1995, pp. 1-10

5. Igor Trubin and Kevin McLaughlin: ***"An Exception Detection System Based on the Statistical Process Control Concept"***, Proceedings of the Computer Measurement Group, 2001