Application of Predictive Analytics for Better Alignment of Business and IT

Boris Zibitsker, PhD
bzibitsker@beznext.com

July 25, 2014
Big Data Summit - Riga, Latvia

About the Presenter

- Boris Zibitsker started out as engineer at Computer Systems Research Institute working on modeling of computer systems and developing scheduling and storage management optimization algorithms
- Capacity management departments at Large Bank (FNBC) and Large Insurance company (CNA)
- Founder, CTO and Chairman of BEZ Systems (1983), acquired by Compuware in 2010
- At BEZ Systems Boris managed development of the commercial capacity management tools supporting multi-tier distributed parallel processing Teradata, Oracle, DB2 and SQL Server and J2EE Application servers
- Co-founder of Computer Systems Institute (1989)
- Founder of BEZNext (2011)
- Consulted many of Fortune 500 companies
- Current focus is on applying predictive analytics, machine learning and queueing theory for optimizing business and IT strategic, tactical and operational decisions
- MS and PhD research at BSUIR and NIIEVM
- Taught graduate courses on Modeling of Computer Systems, Queueing Theory with Computer Applications, Computer Communication Systems Design and Analysis at DePaul University in Chicago
- Taught capacity management courses for the Relational Institute founded by pioneers of relational technology Dr. Ted Codd and Chris Date
- Presented many papers on modeling, workload characterization, performance management, workload management and capacity planning

Abstract

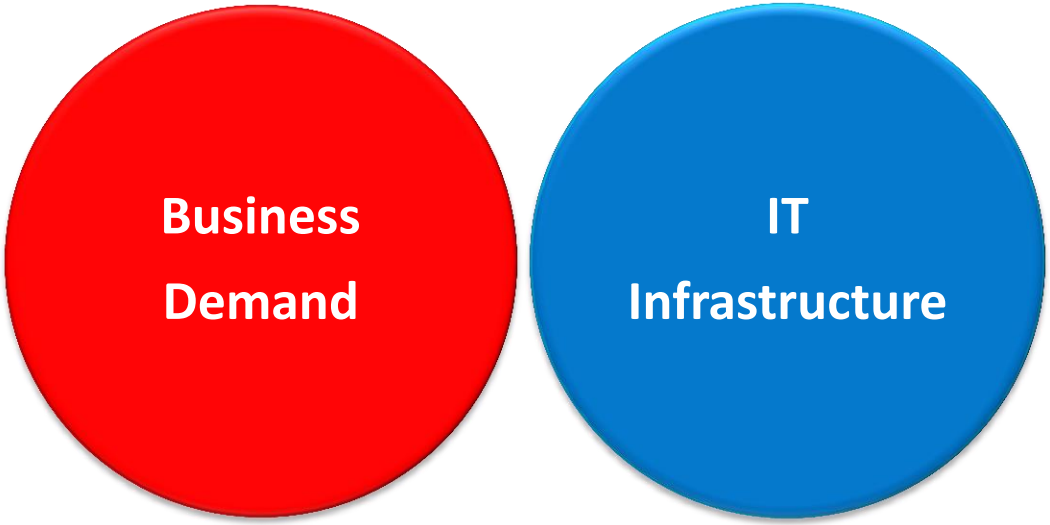
- Business decisions changing business processes often done without knowing how it will affect IT.
- On the other hand, it is difficult to make effective IT decisions with limited information about business processes and how they will change demand for IT resources.
- Luck of information about change of demand for resources and uncertainty how it will affect performance make it difficult to justify IT decisions related to architecture, hardware, software and DBMS platforms and configurations, implementation of Big Data and other technologies.
- As a result it is difficult to set realistic service level goals and performance expectations.
- It causes an uncertainty, risk of performance surprises and disappointment with high cost of IT.
- In this presentation, we will review several case studies illustrating how predictive analytics and optimization enables better alignment of business and IT through justification of strategic, tactical and operational proactive IT actions and verification of results.

Outline

- Introduction
- Methodology
- Data Collection
- Workload Characterization
- Workload Forecasting
- Performance Prediction
- Verification of Results

INTRODUCTION

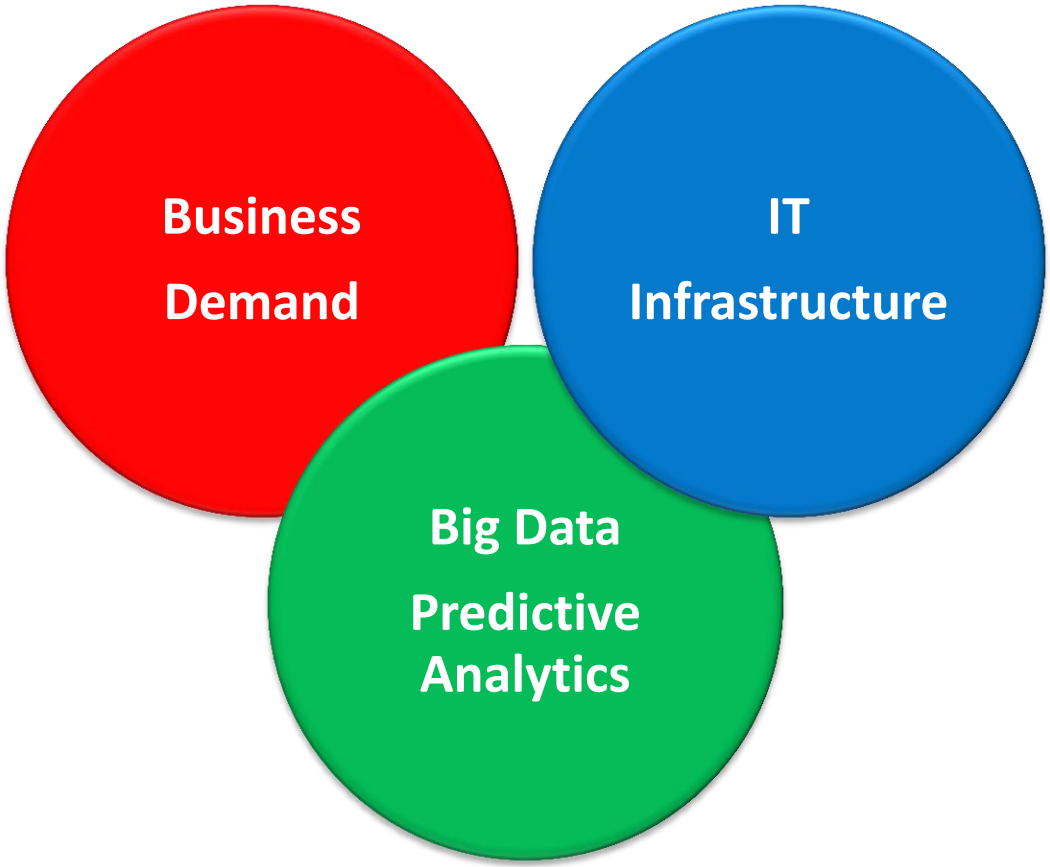
Complex relationships between business and IT



Business decisions often do not take into consideration how they will affect IT

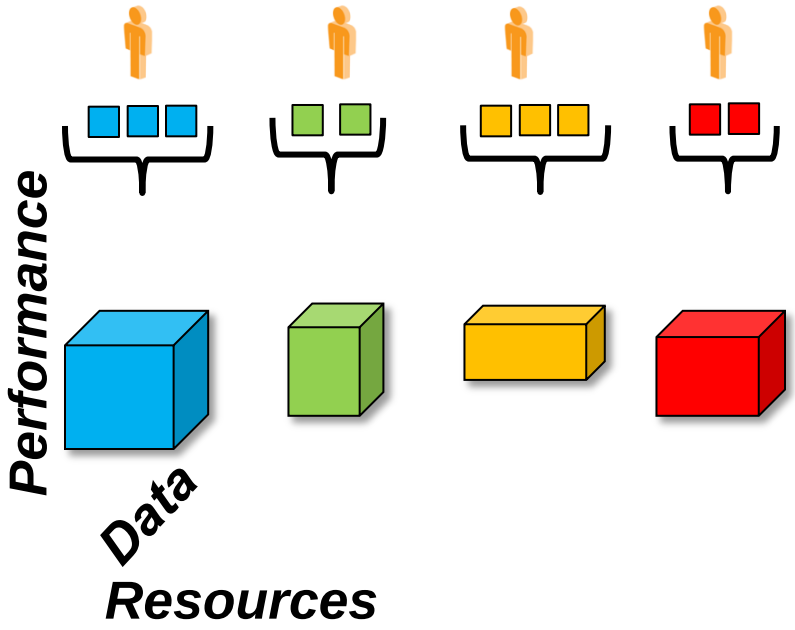
Role of Big Data

Reduce Cost, Load and Access Different Data Faster
Apply Analytics to Make Better Decisions



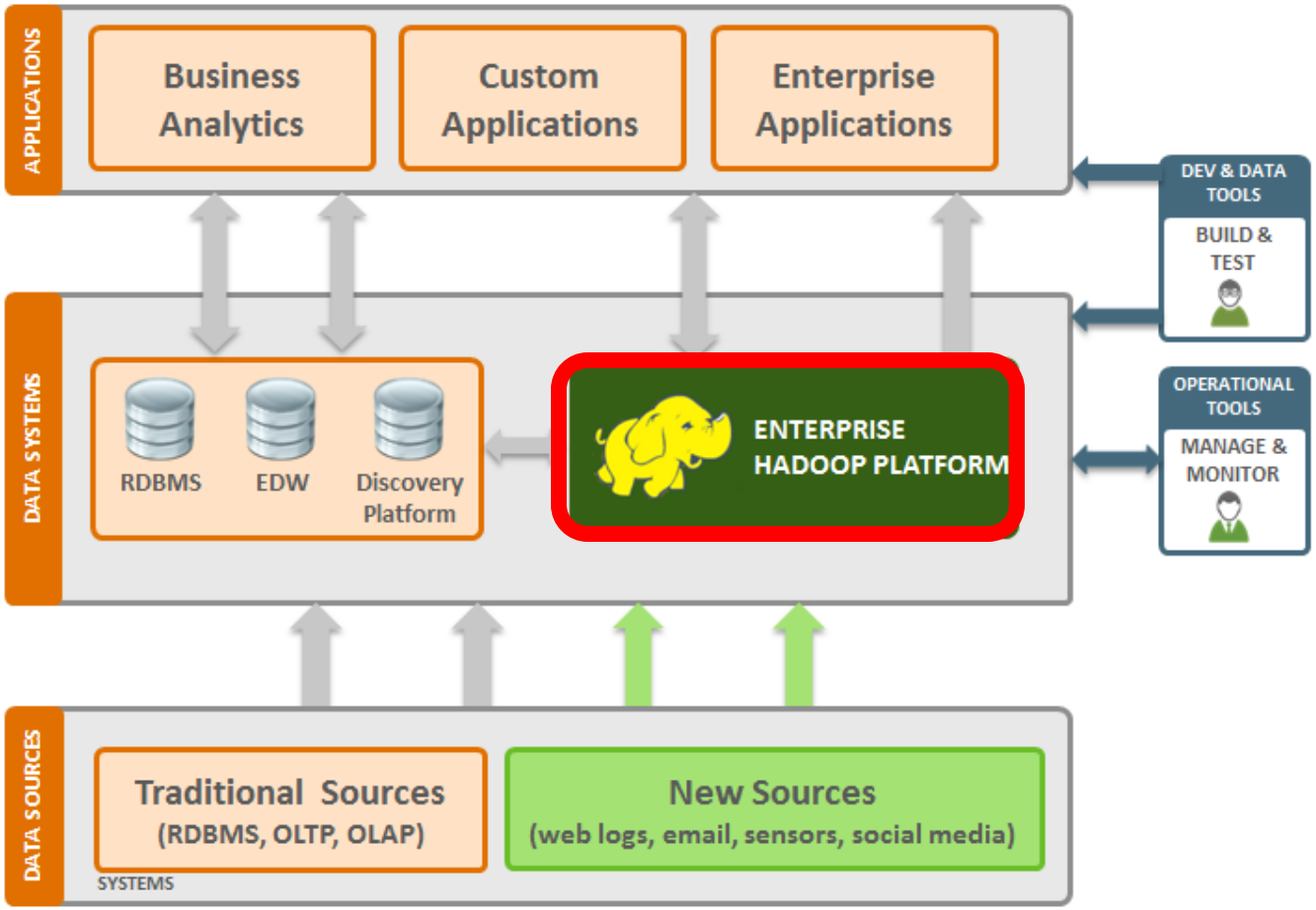
Big data predictive analytics enables better alignment of business and IT

Business Demand is Constantly Changing



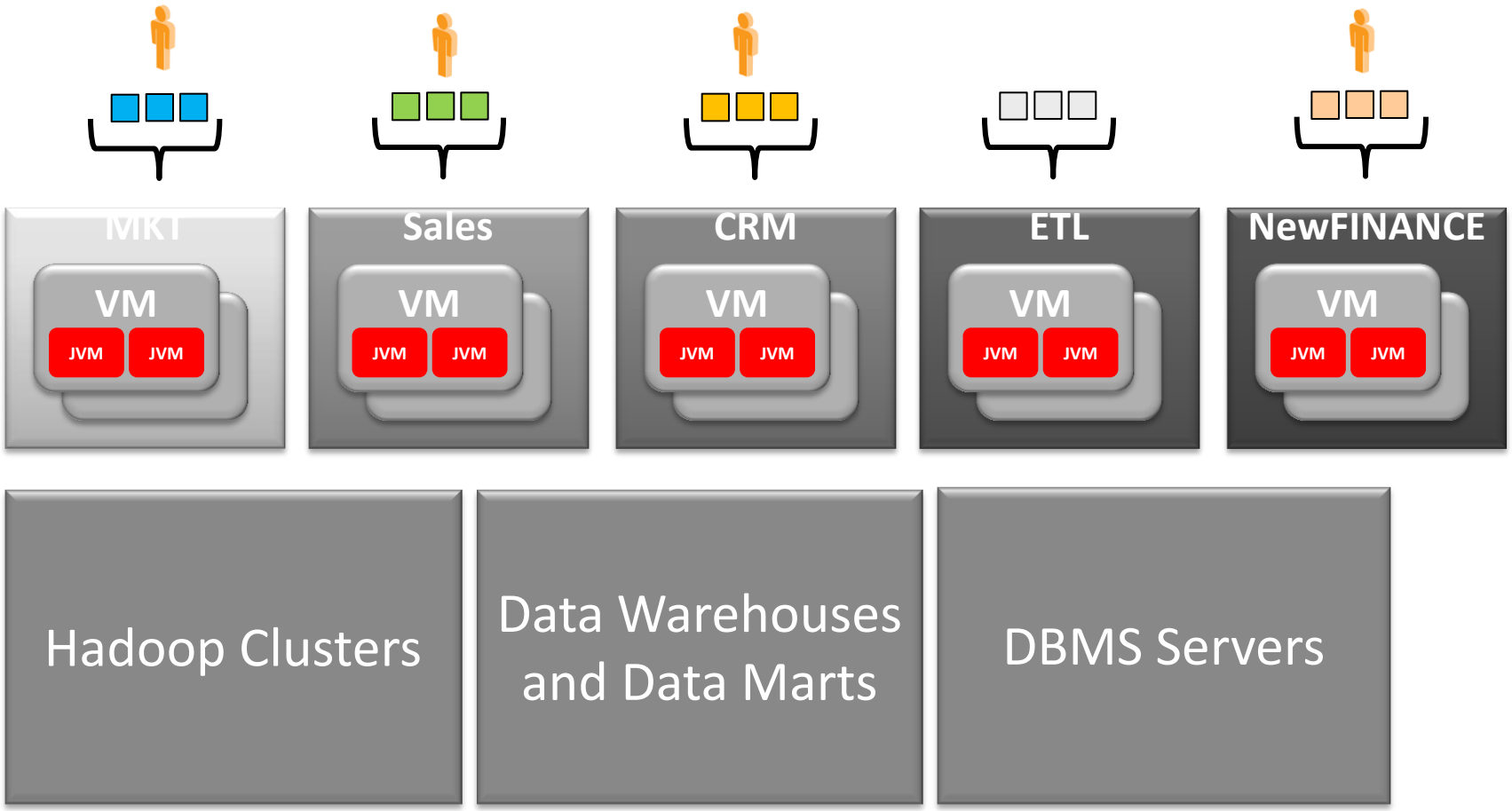
- Growth
- New Business Processes
- New Applications
- New Sources of Data
  - Business Process Models
  - Enterprise Data Model
  - Results of Testing
- Changing Business Plan
- Budget

Big Data Hadoop Cluster is a Part of the IT Infrastructure



Implementation of New Application

Complex multi-tier, distributed, virtualized infrastructure supporting mix load, batch, transaction processing and analytic workloads



it is difficult to make effective IT decisions with limited information about business processes and how they will change demand for IT resources.

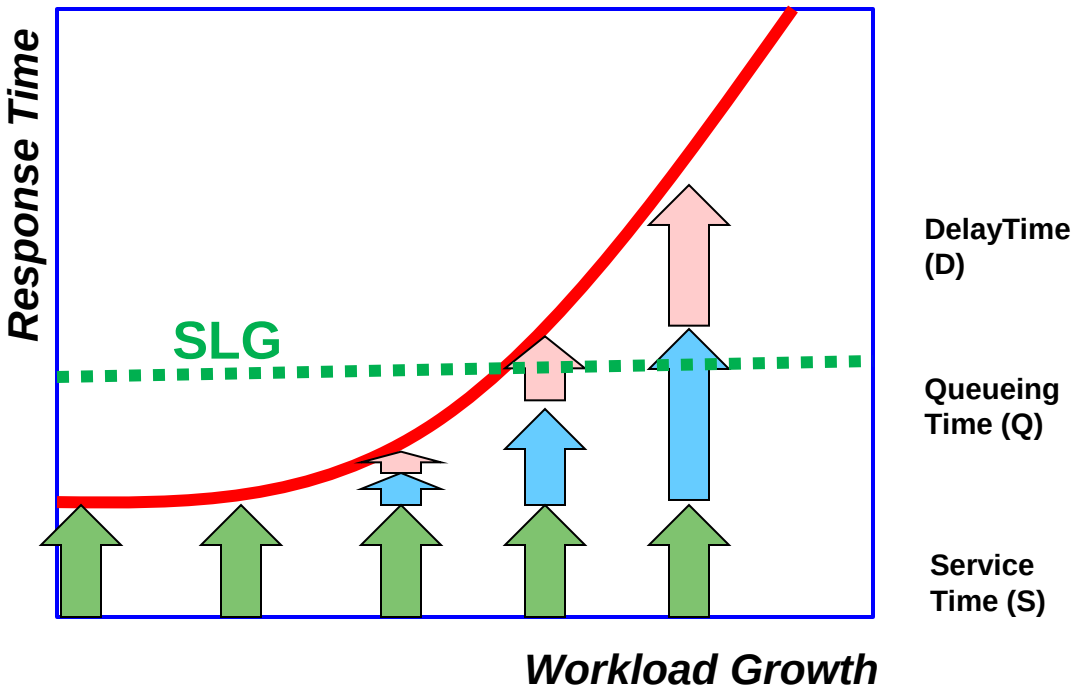
METHODOLOGY

Application of Predictive Analytics for Production and Test Environments

- Service Level Goals
- Data Collection
- Workload Characterization
- Workload Forecasting
- Building Models
- Predicting Impact of Moving New Application to Production Environment
- Justification of Architecture
- Capacity Management
  - Capacity Planning
  - Performance Management
  - Workload Management
- Setting expectations
- Verification

Major Factors Affecting Performance and Service Level Goals

- Workload profile
  - Performance
  - Usage of resources
  - Usage of data
- Expected growth
- Architecture
  - Hardware configuration
  - Software configuration
  - Virtual configuration
- Application design
- Design



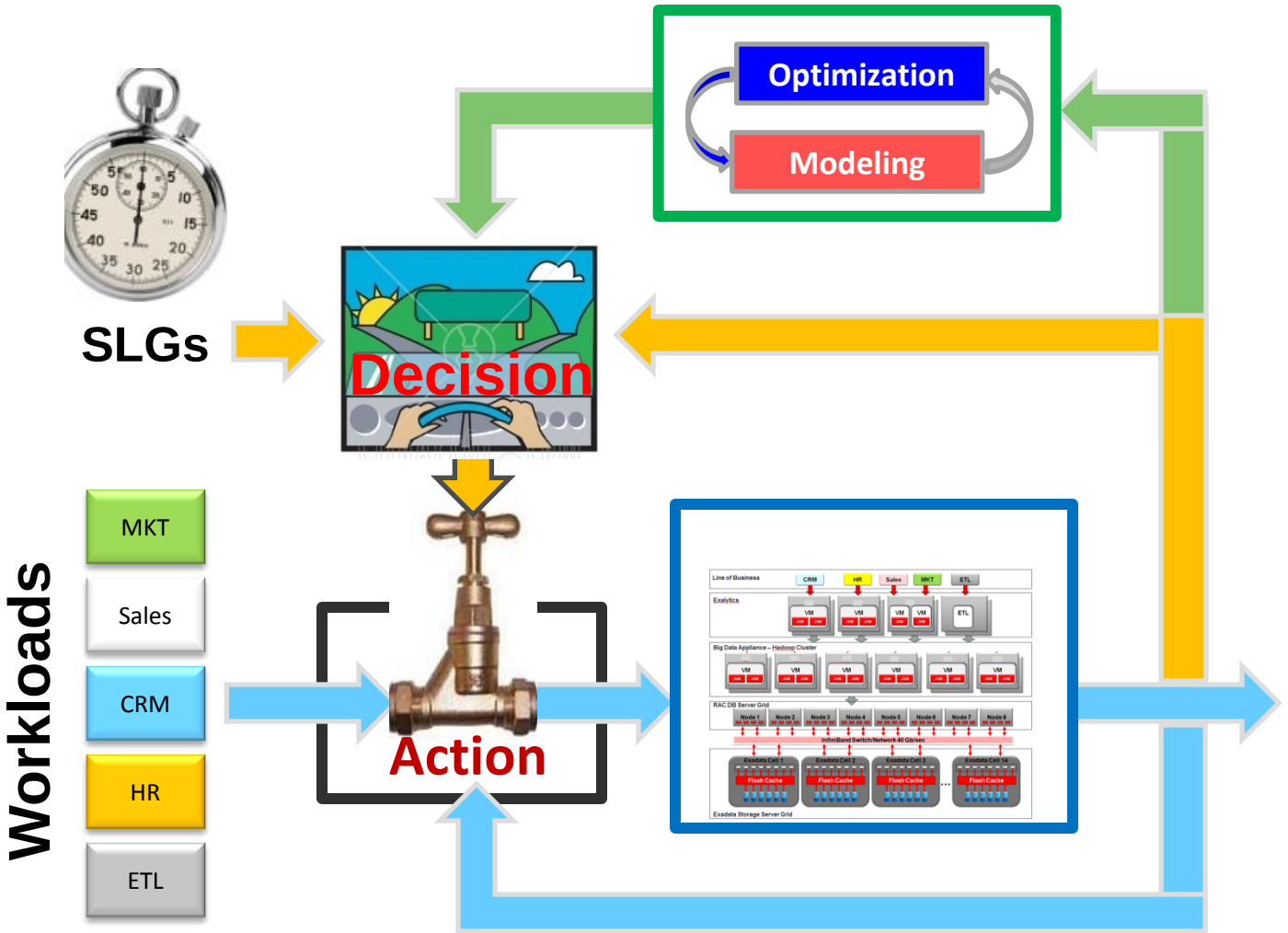
Types of Workloads and SLGs

- Business Process
  - Applications
  - Data
- Load – ETL/ELT
- Access
  - Transaction Processing
  - Analytics / DSS

- Cost / Performance

- Load data on time
- Response Time and Throughput

Role of Modeling and Optimization in Justification of Decisions



Collaboration Between Business and IT

- Current Business Demand
- Business Plan
- New Business Processes
- New Applications
- New Data
- Future Business Demand
- Budget
- Service Level Goals (SLG)

- Analysis of performance prediction results
- Analysis of changes and their cost necessary meeting SLGs
- Reevaluation of assumptions and SLGs
- Setting realistic SLGs

Capacity Management

Operational Workload Management

- How to change dynamically Priority, Concurrency and resource Allocation for the individual workloads

Tactical Performance Management

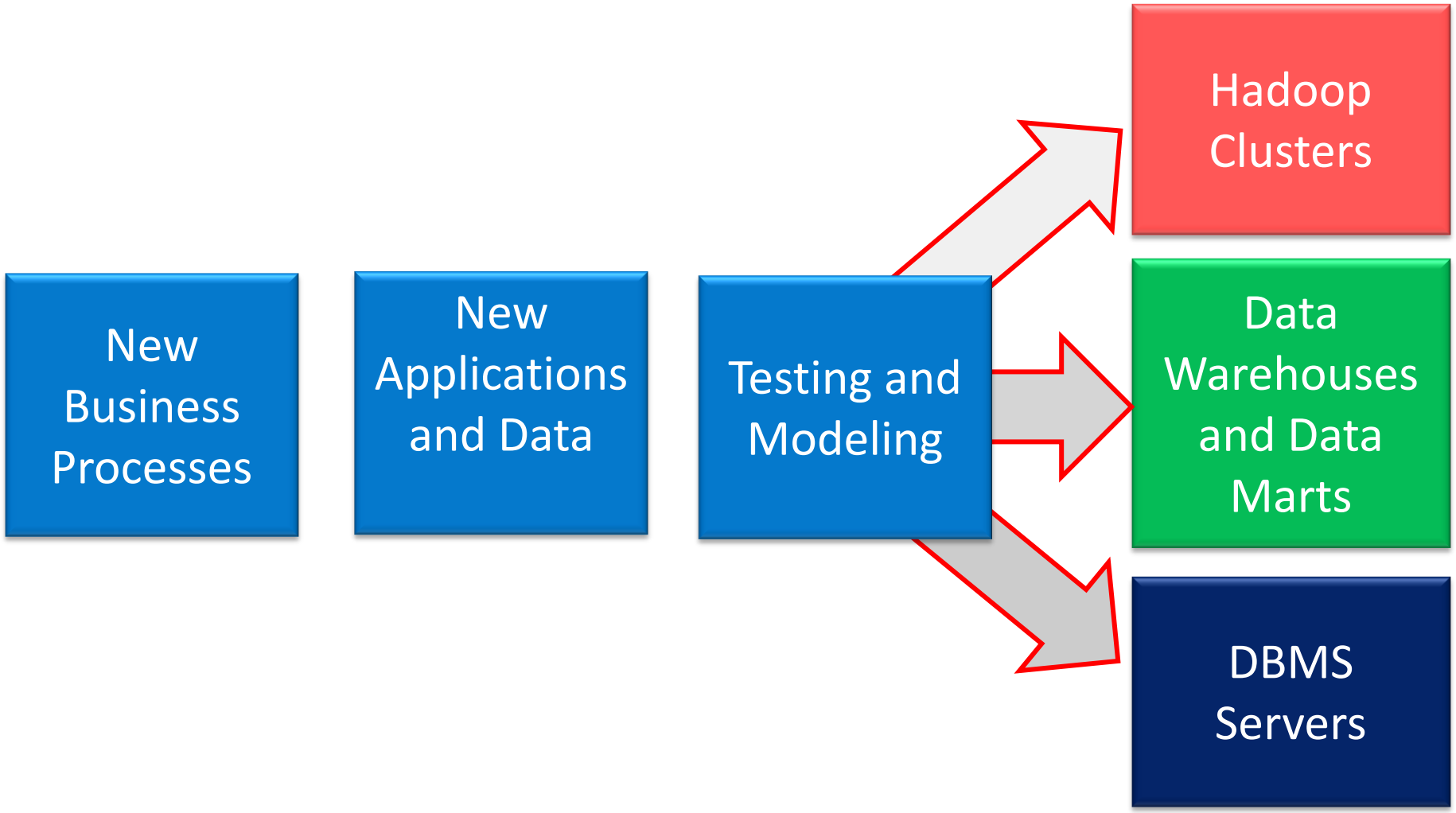
- What should be tuned proactively

Strategic Capacity Planning

- How to predict the impacts of the expected workload and volume of data growth, new applications implementation ?
- How to predict when system will be out capacity and Identify future performance bottlenecks?
- When and what type of changes will be required to continuously support service level goals for the individual workloads
- How to negotiate SLGs and set up realistic expectations

Testing and Predicting

ETL, Transaction and Analytic Processing Performance Prior to New Application Implementation



DATA COLLECTION

Data Collection

- Configuration
- OS
- Application

- Frequency
- Overhead
- Data repository

Workload aggregation

Workload characterization

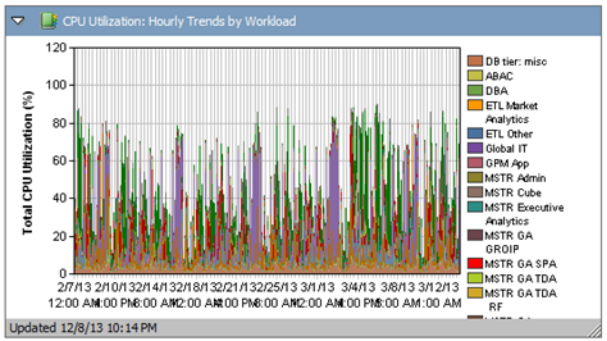
Workload profiles

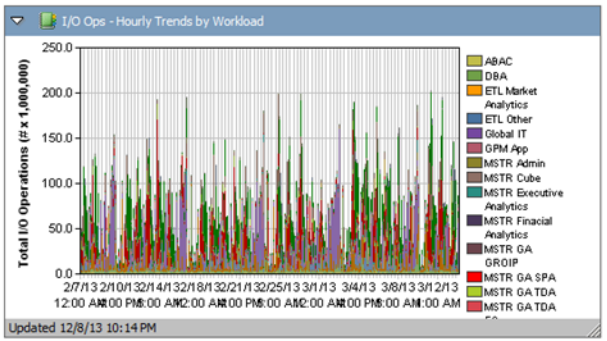
Performance analysis

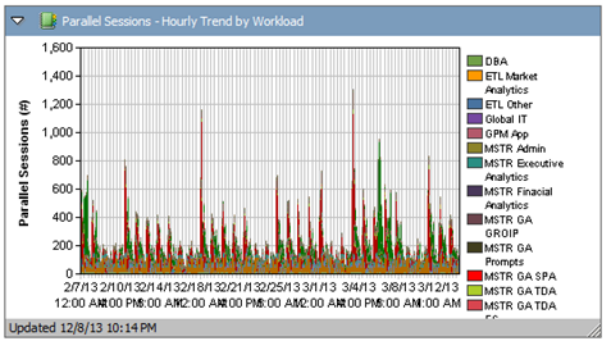
WORKLOAD CHARACTERIZATION

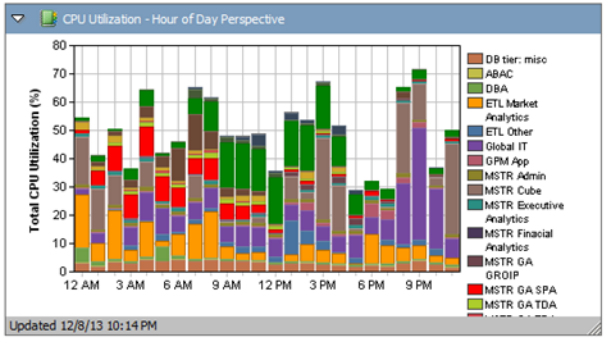
Workload Characterization

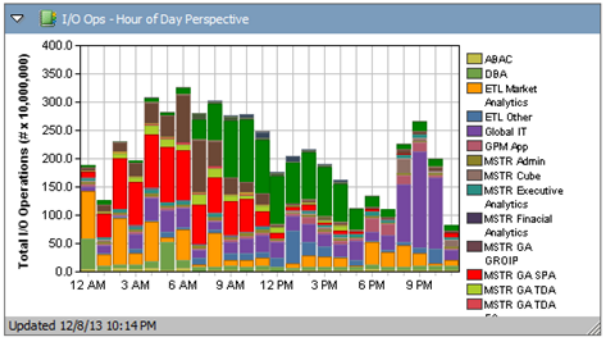
Workloads Profiles

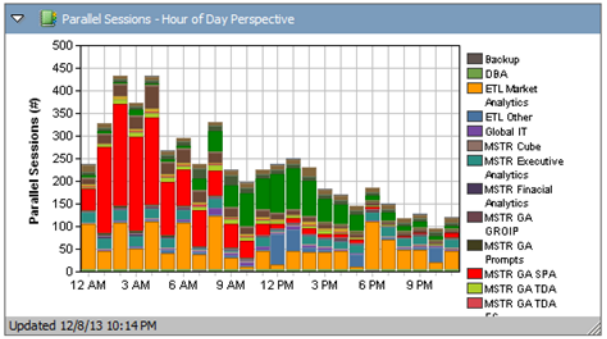


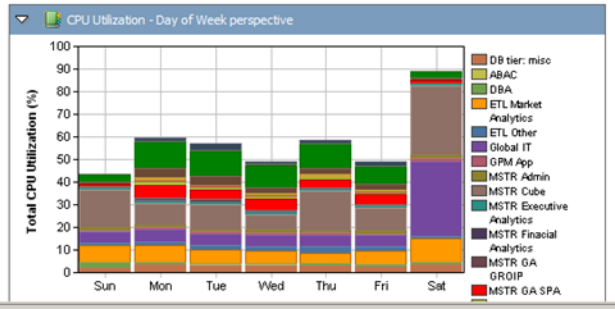


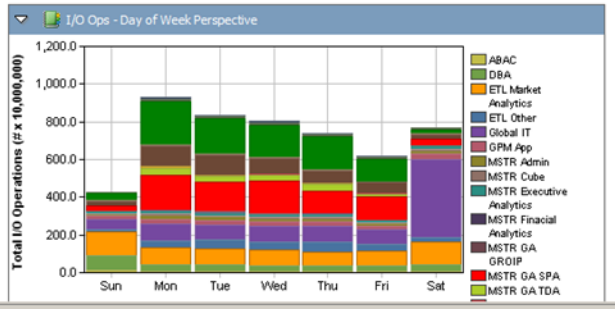


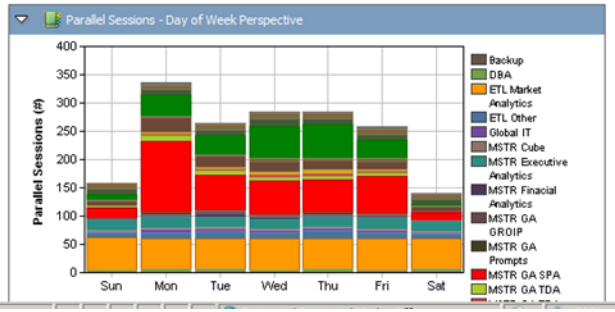




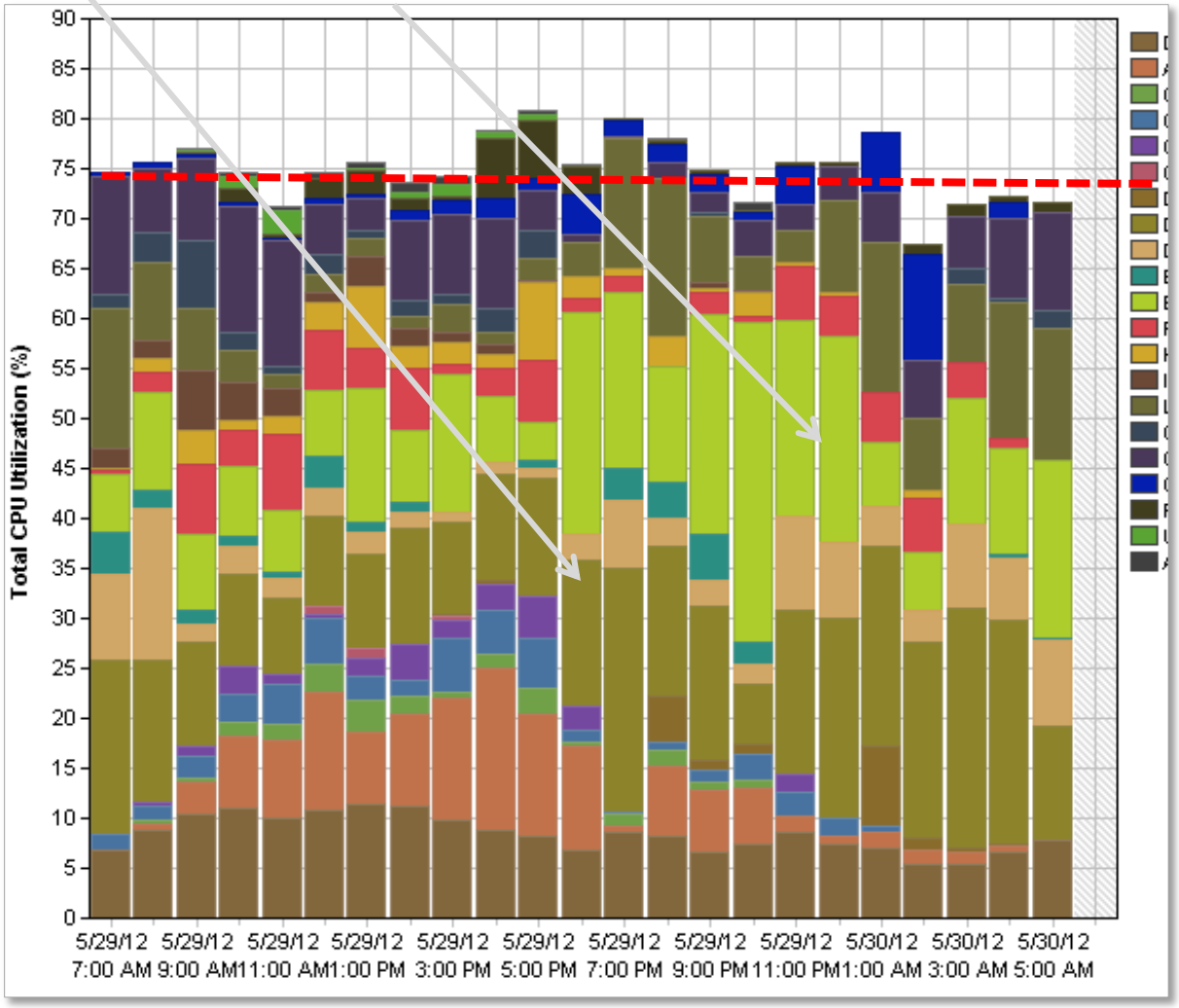




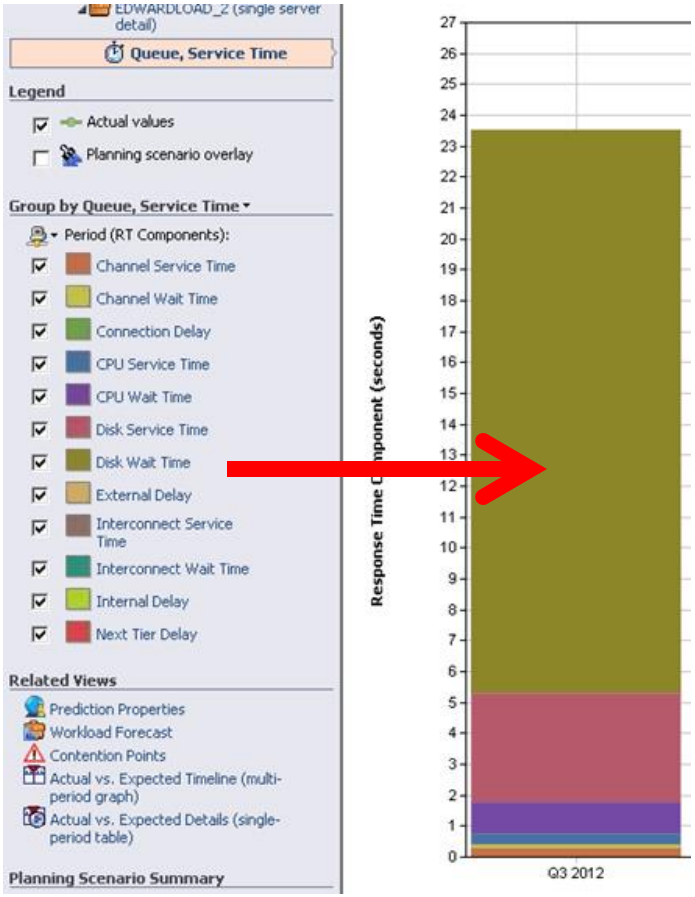




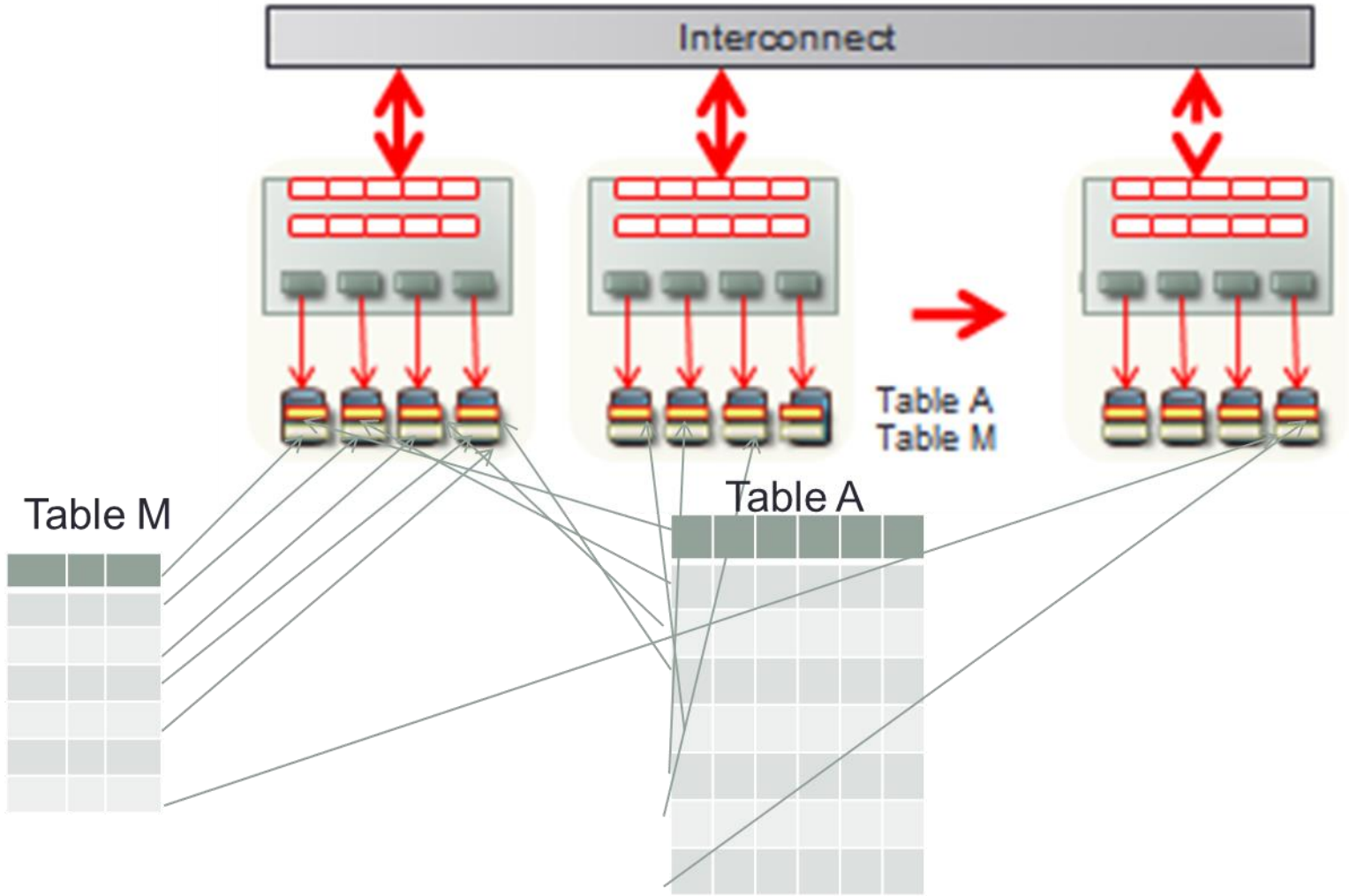
ETL is a Most Resource Consuming Workload



The Largest Component of the ETL is Disk Wait Time

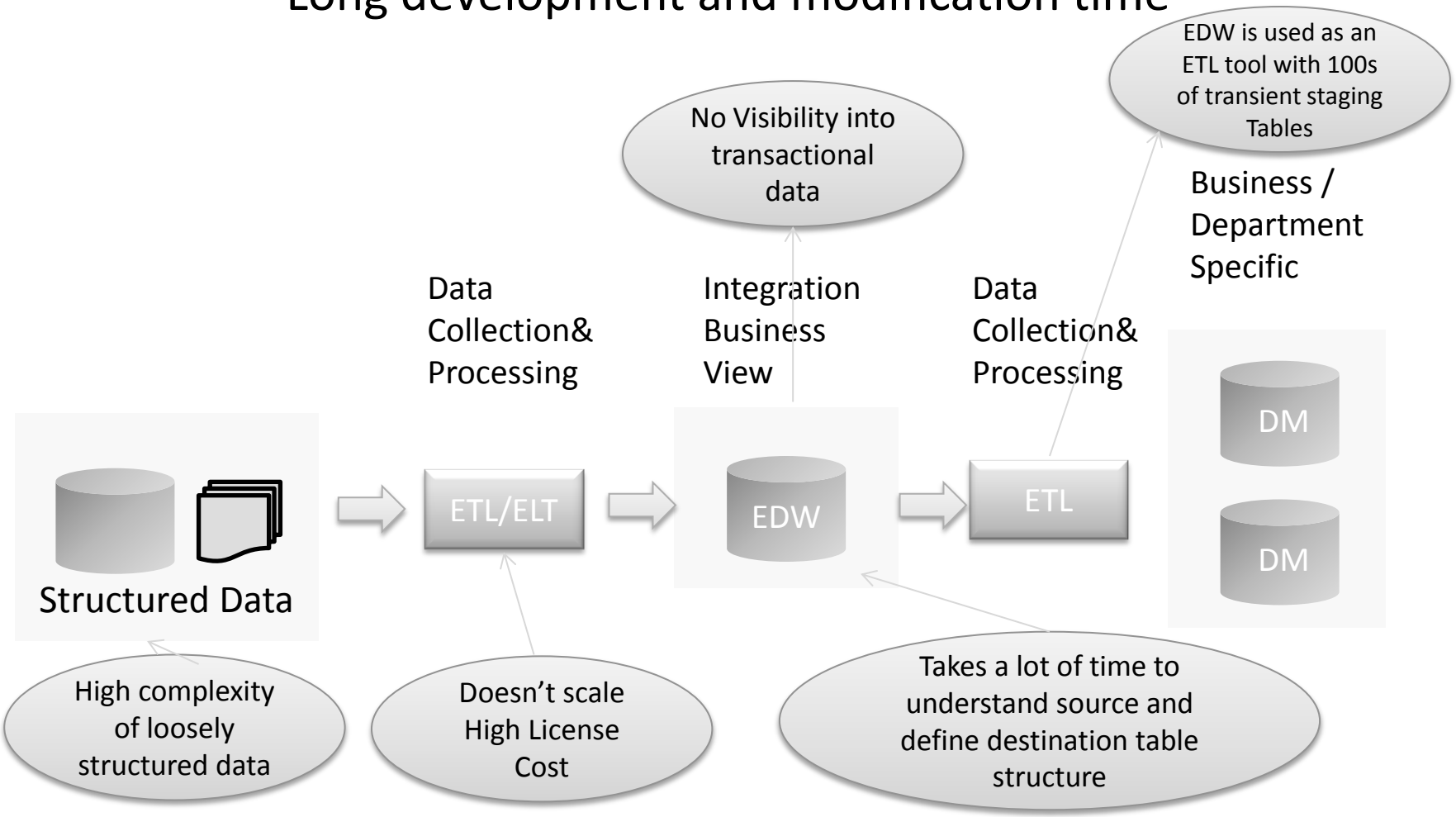


Massively Parallel Processing Systems Distribute Data Across all Nodes

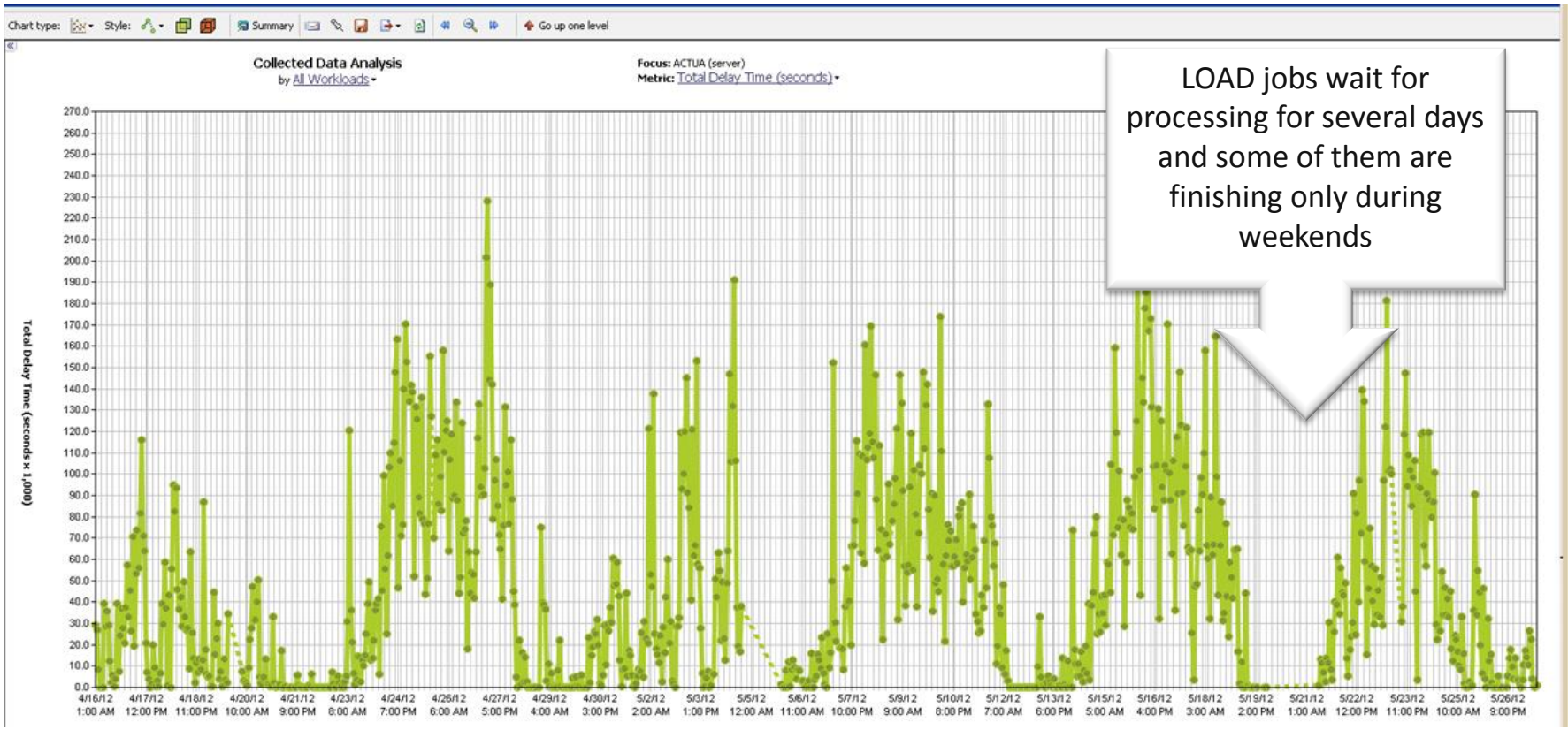


Implementation Traditional ETL Processes Challenges

Long development and modification time



Long ETL Time and Inconsistent Response Time in DW Environment



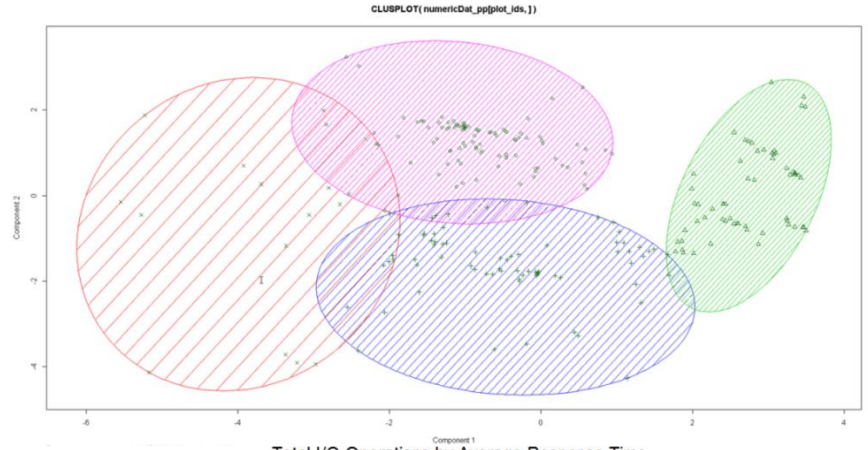
One of the Advantages of Big Data
Late Binding

- Do not map data before and during load time, instead map data during query time dynamically!

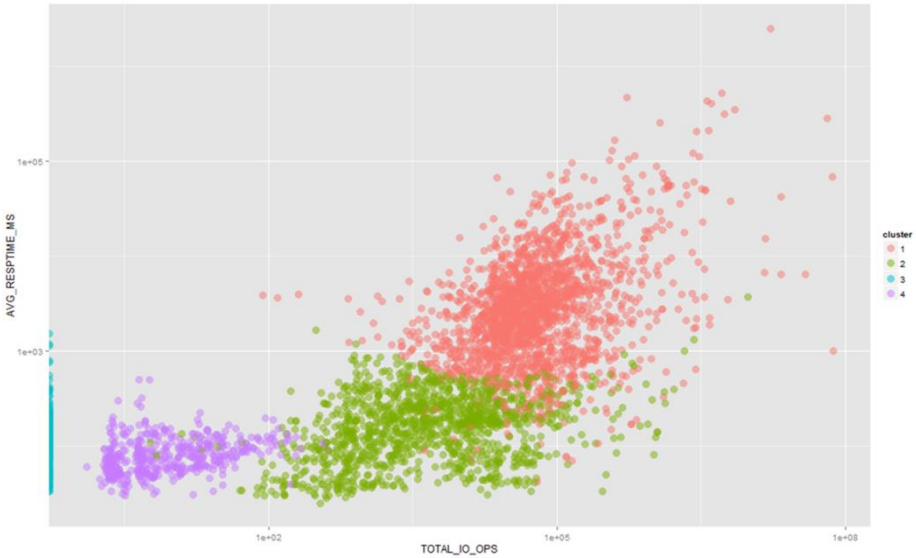
Data Exploration

Cluster Analysis - Unsupervised learning

- Discover structure in collection of data
- Multi Variable Cluster Analysis Identifies Groups of Requests with Similar Performance and Resource Utilization Characteristics
- Often used for exploratory analysis



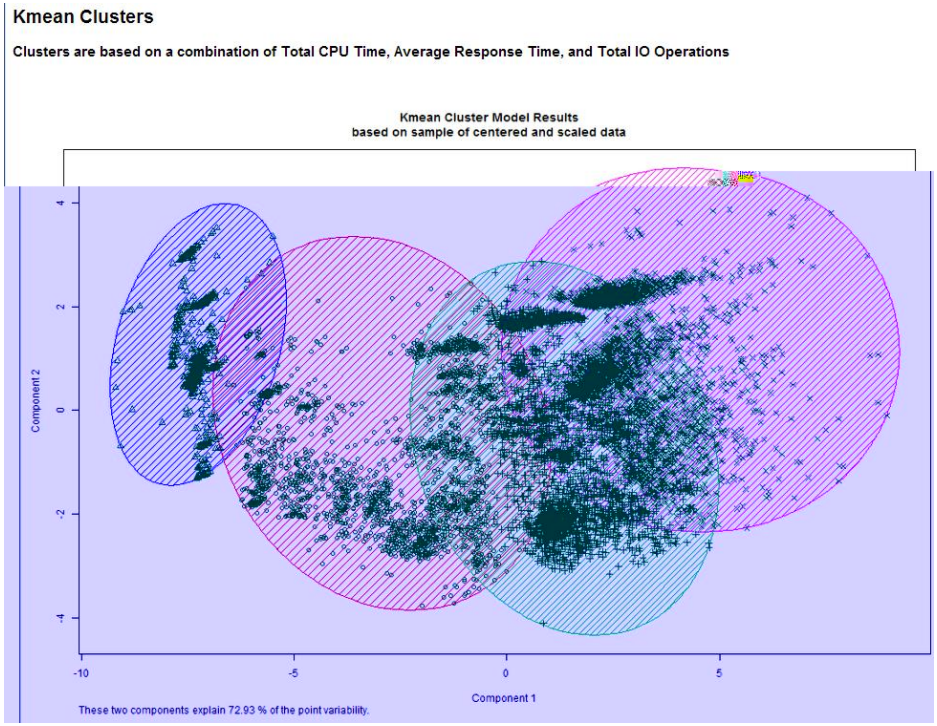
Total I/O Operations by Average Response Time
Clusters using 'Mclust'



Data Exploration

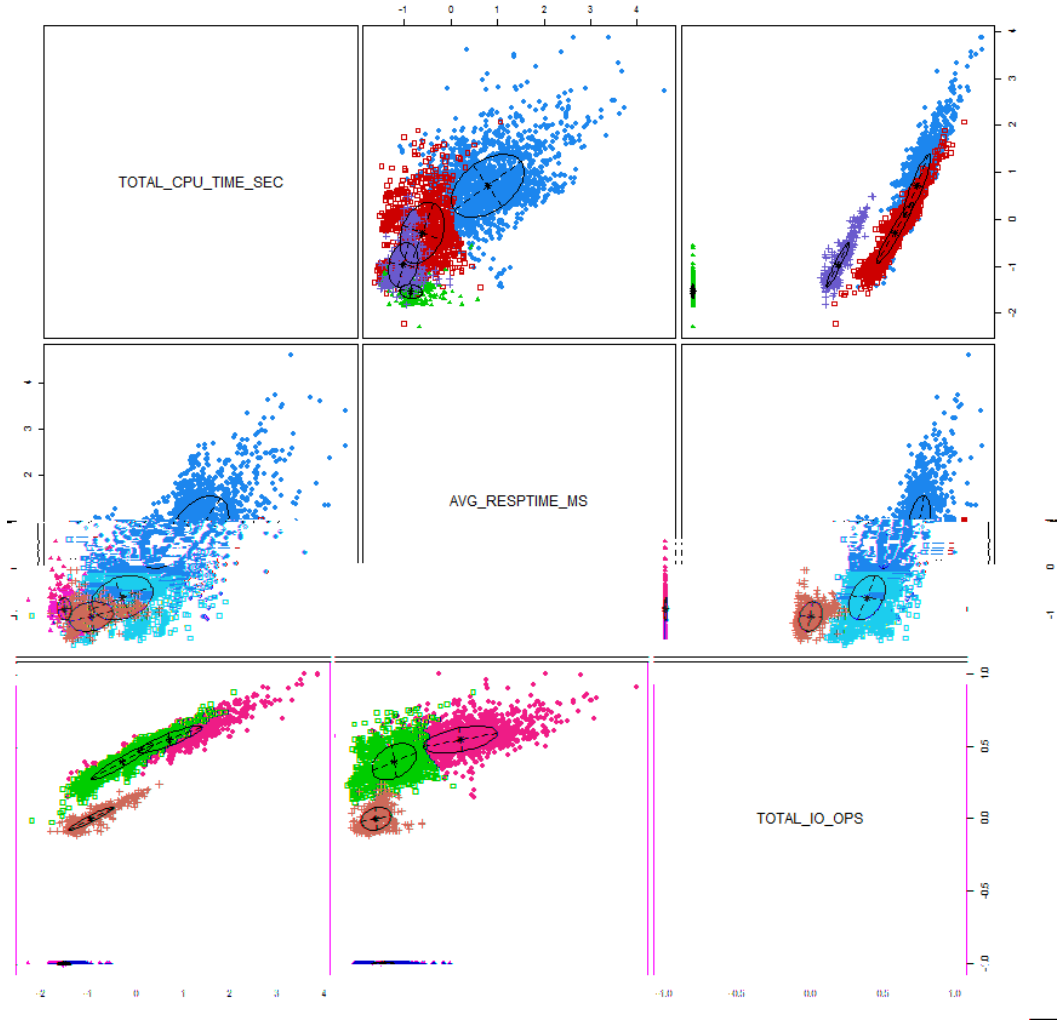
K-mean Clusters - Total CPU, Avg RT and I/O

- *k*-means clustering partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.
  - Input: Points in an *n*-dimensional space and a parameter *k*
  - Output: Grouping of the points into *k* clusters and a representative/central point for each cluster



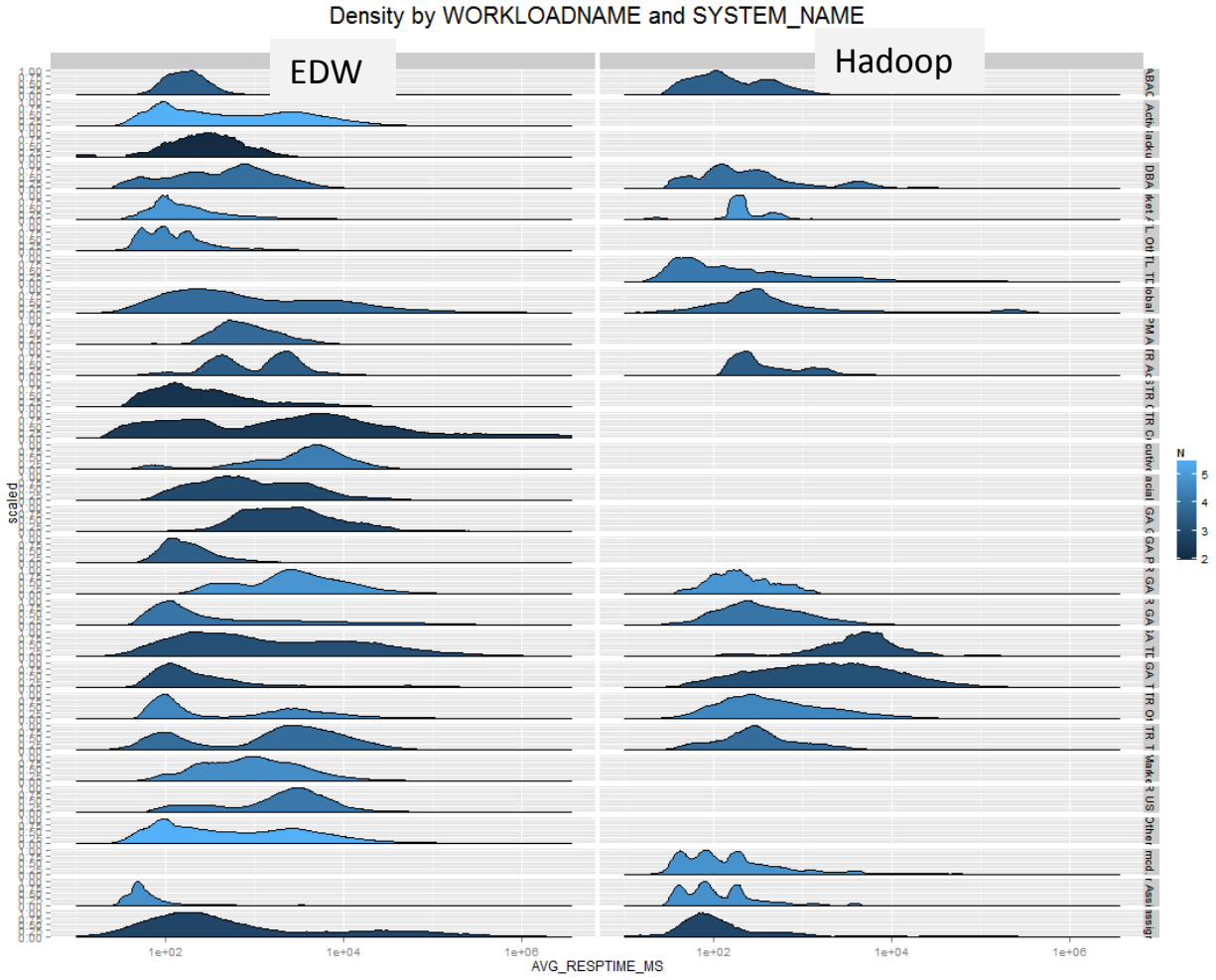
Data Exploration - MClust Results

Average Response Time vs Total CPU Time vs Total I/O



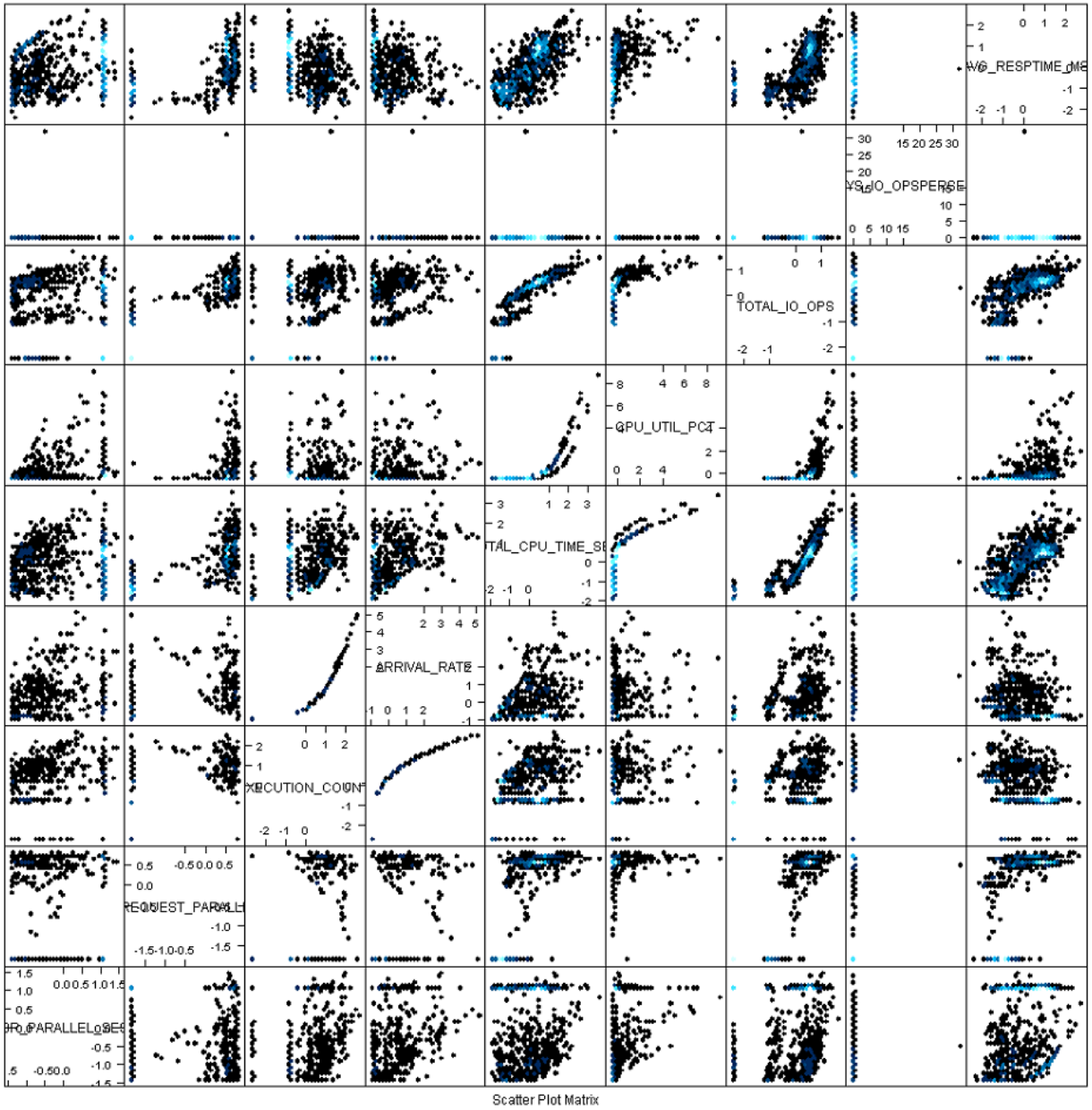
Data Exploration

Response Time Distribution by Workload and System



Data Exploration

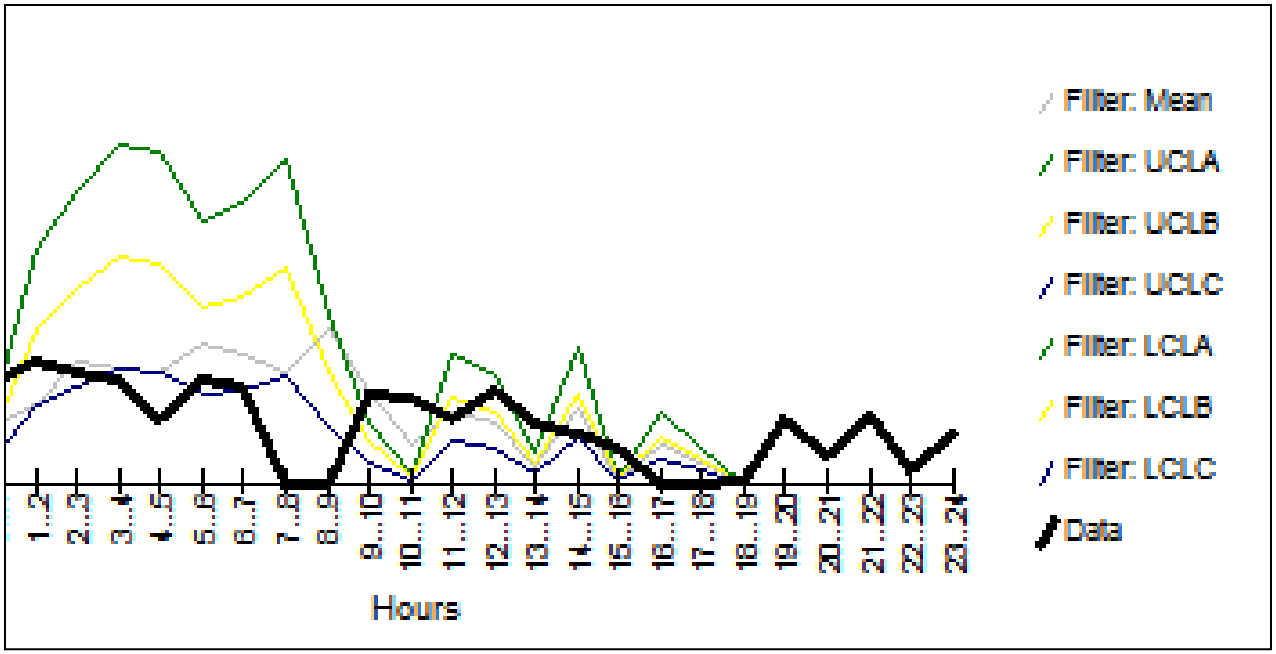
Scatter Plot Matrix Grouped / binned into hexagonal tiles



Determining Significant Changes

Statistical Process Control

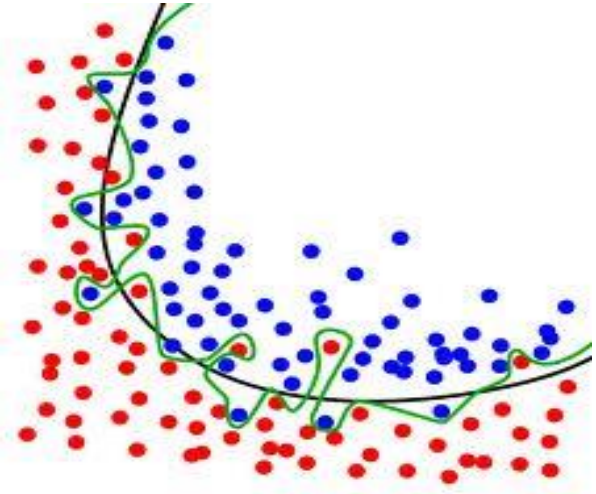
RT, Throughput and Resource Utilization

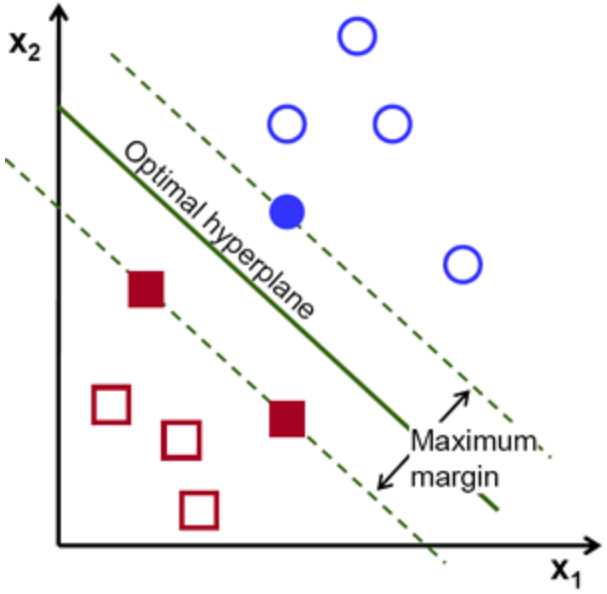


Support Vector Machines

Supervised Learning

- Linear and non linear classification segregates good tests from bad tests based on previous training
- Good performance vs performance anomaly

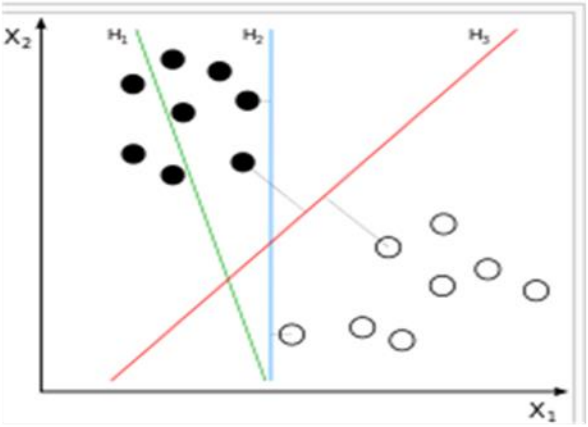


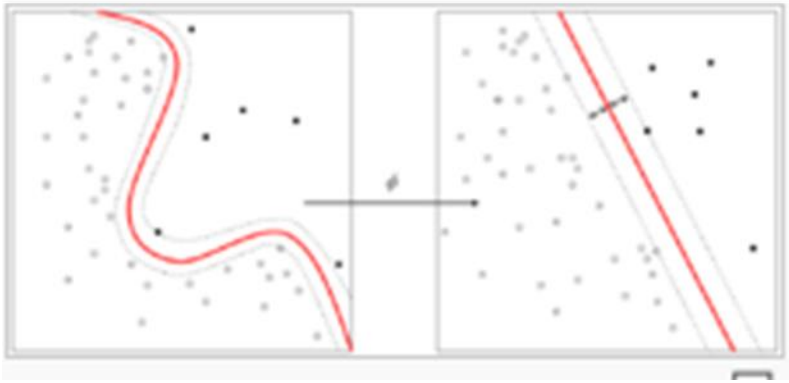


Root Cause Analysis

Logistic Regression - Semi-supervised learning

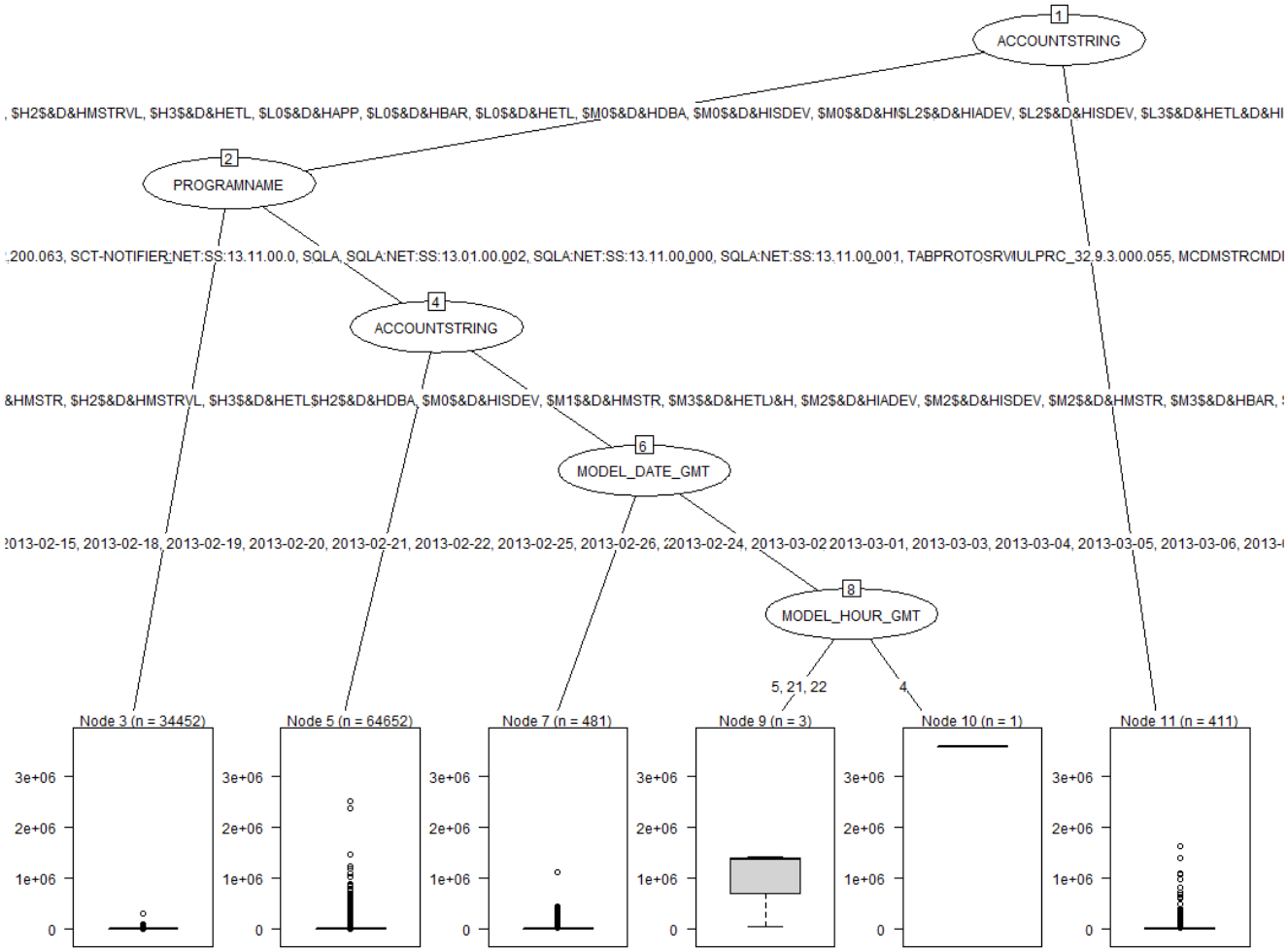
- Minimal value of PageRT in Bad Cluster gives a threshold segregating Bad tests from good tests





Root Cause Analysis

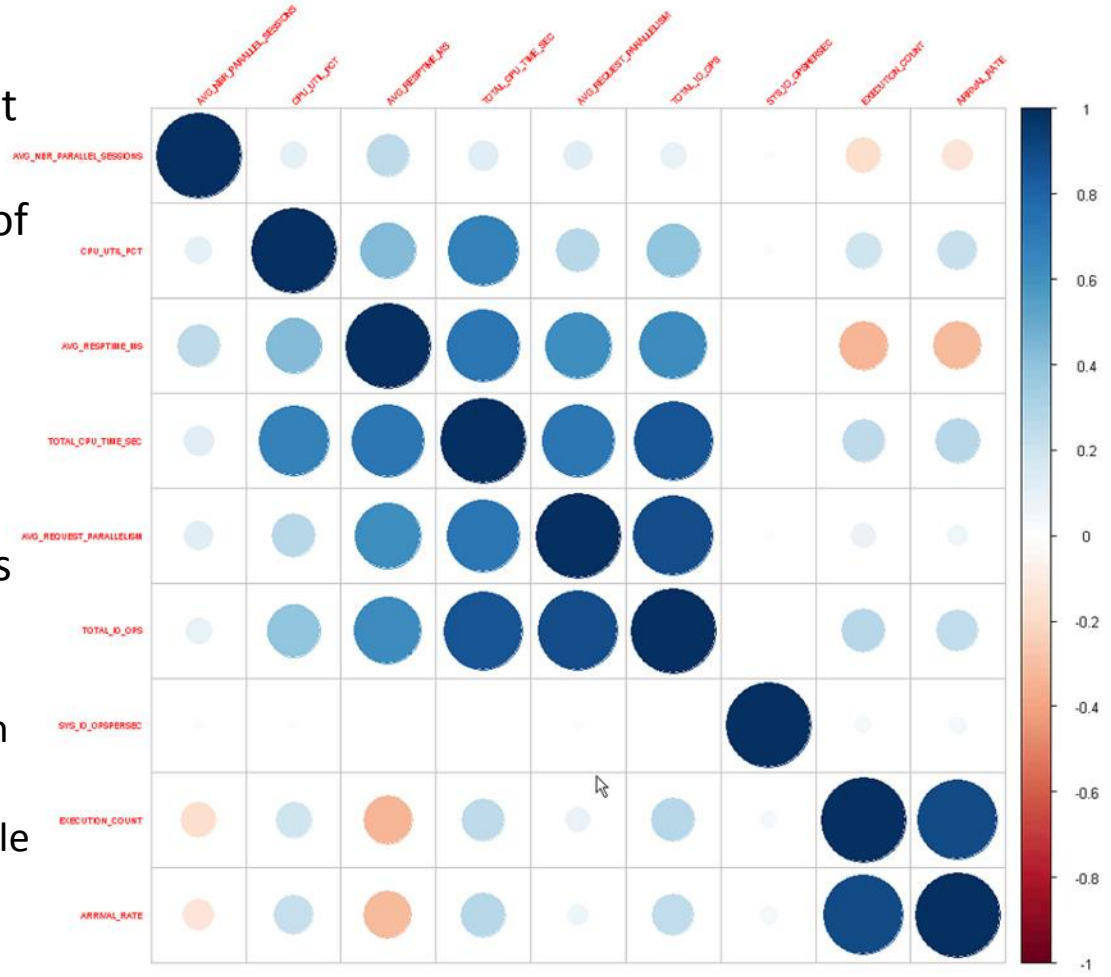
Decision Tree - Leaf page and branches identify the cause



Root Causes

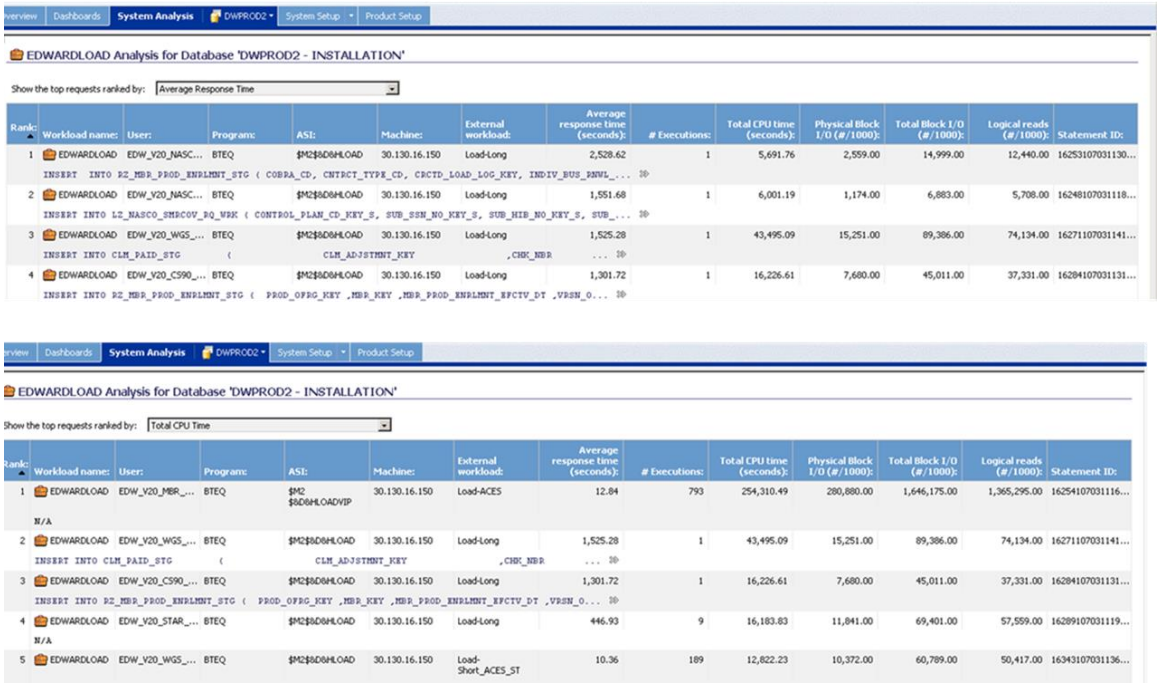
Correlation Analysis Identifies the Most Significant Variables Affecting RT

- We can eliminate variables that are highly correlated (for example an average and total of the same measure)
- Simplification has beneficial effects:
  - simpler data
  - more stable models
  - more interpretable results
- Collinearity
  - Pare of predictor variables have a substantial correlation
- Multicollinearity
  - Relationship between multiple predictors



What are the Options?

- Workload management optimization
- Performance tuning
- Hardware upgrade
- Moving ETL workload to Hadoop Cluster



Overview | Dashboards | System Analysis | DWPROD2 | System Setup | Product Setup

EDWARDLOAD Analysis for Database 'DWPROD2 - INSTALLATION'

Show the top requests ranked by: Average Response Time

Rank	Workload name:	User:	Program:	ASI:	Machine:	External workload:	Average response time (seconds):	# Executions:	Total CPU time (seconds):	Physical Block I/O (#/1000):	Total Block I/O (#/1000):	Logical reads (#/1000):	Statement ID:
1	EDWARDLOAD	EDW_V20_NASC...	BTEQ	$M2$8D8HLOAD	30.130.16.150	Load-Long	2,528.62	1	5,691.76	2,559.00	14,999.00	12,440.00	16253107031130...
	INSERT INTO RZ_MBR_PROD_ENRLMNT_STG (COBRA_CD, CNTRCT_TYPE_CD, CRCTD_LOAD_LOG_KEY, INDIV_BUS_ENWL_...												
2	EDWARDLOAD	EDW_V20_NASC...	BTEQ	$M2$8D8HLOAD	30.130.16.150	Load-Long	1,551.68	1	6,001.19	1,174.00	6,883.00	5,708.00	16248107031118...
	INSERT INTO LZ_NASCO_SMRCOV_RQ_WRK (CONTROL_PLAN_CD_KEY_S, SUB_SSN_NO_KEY_S, SUB_HIB_NO_KEY_S, SUB_...												
3	EDWARDLOAD	EDW_V20_WGS_...	BTEQ	$M2$8D8HLOAD	30.130.16.150	Load-Long	1,525.28	1	43,495.09	15,251.00	89,386.00	74,134.00	16271107031141...
	INSERT INTO CLM_PAID_STG (CLM_ADJSTMNT_KEY ,CHK_NBR ...												
4	EDWARDLOAD	EDW_V20_C590_...	BTEQ	$M2$8D8HLOAD	30.130.16.150	Load-Long	1,301.72	1	16,226.61	7,680.00	45,011.00	37,331.00	16284107031131...
	INSERT INTO RZ_MBR_PROD_ENRLMNT_STG (PROD_OFRG_KEY ,MBR_KEY ,MBR_PROD_ENRLMNT_EFCTV_DT ,VRSN_O...												

Overview | Dashboards | System Analysis | DWPROD2 | System Setup | Product Setup

EDWARDLOAD Analysis for Database 'DWPROD2 - INSTALLATION'

Show the top requests ranked by: Total CPU Time

Rank	Workload name:	User:	Program:	ASI:	Machine:	External workload:	Average response time (seconds):	# Executions:	Total CPU time (seconds):	Physical Block I/O (#/1000):	Total Block I/O (#/1000):	Logical reads (#/1000):	Statement ID:
1	EDWARDLOAD	EDW_V20_MBR_...	BTEQ	$M2 $8D8HLOADVIP	30.130.16.150	Load-ACES	12.84	793	254,310.49	280,880.00	1,646,175.00	1,365,295.00	16254107031116...
	N/A												
2	EDWARDLOAD	EDW_V20_WGS_...	BTEQ	$M2$8D8HLOAD	30.130.16.150	Load-Long	1,525.28	1	43,495.09	15,251.00	89,386.00	74,134.00	16271107031141...
	INSERT INTO CLM_PAID_STG (CLM_ADJSTMNT_KEY ,CHK_NBR ...												
3	EDWARDLOAD	EDW_V20_C590_...	BTEQ	$M2$8D8HLOAD	30.130.16.150	Load-Long	1,301.72	1	16,226.61	7,680.00	45,011.00	37,331.00	16284107031131...
	INSERT INTO RZ_MBR_PROD_ENRLMNT_STG (PROD_OFRG_KEY ,MBR_KEY ,MBR_PROD_ENRLMNT_EFCTV_DT ,VRSN_O...												
4	EDWARDLOAD	EDW_V20_STAR_...	BTEQ	$M2$8D8HLOAD	30.130.16.150	Load-Long	446.93	9	16,183.83	11,841.00	69,401.00	57,559.00	16289107031119...
	N/A												
5	EDWARDLOAD	EDW_V20_WGS_...	BTEQ	$M2$8D8HLOAD	30.130.16.150	Load-Short_ACES_ST	10.36	189	12,822.23	10,372.00	60,789.00	50,417.00	16343107031136...

Increase in number of users

Volume of data growth

New application implementation

WORKLOAD FORECASTING

Workload Forecasting

Growth Factors

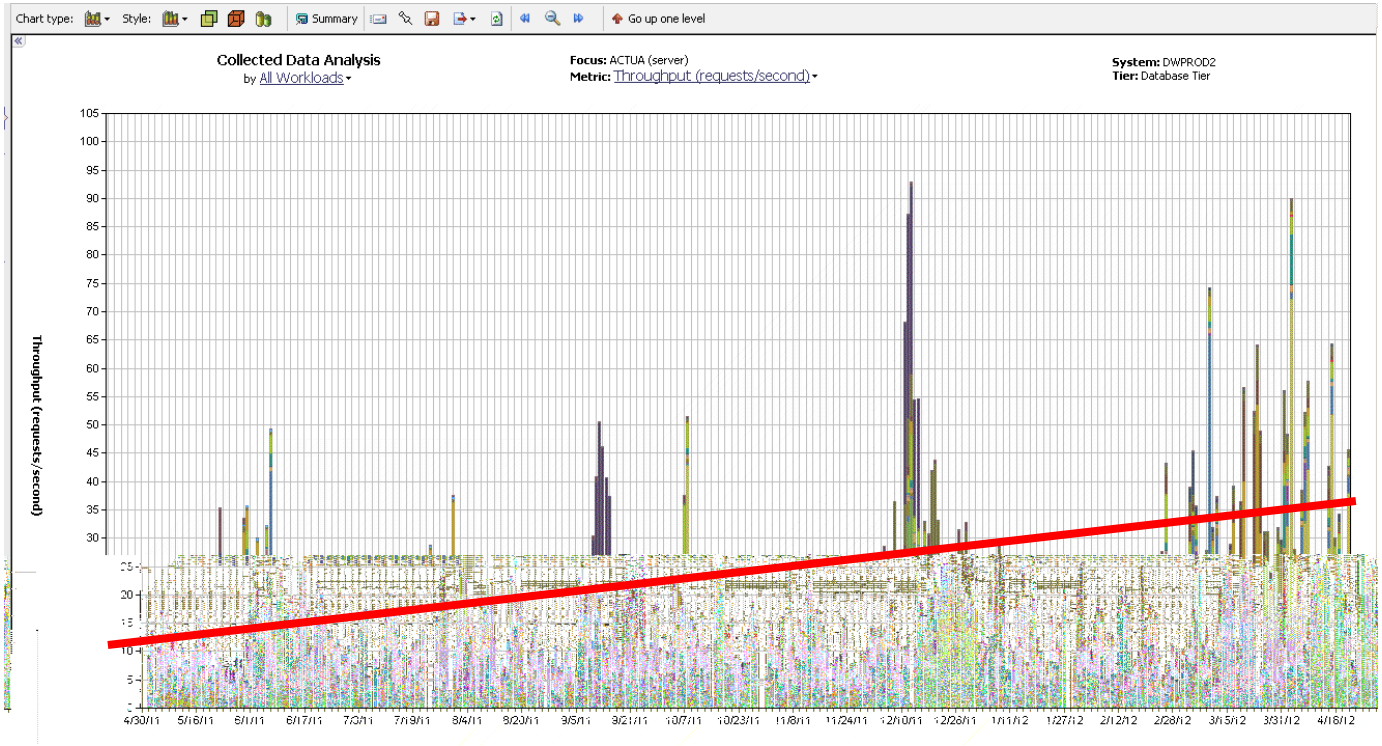
- Workload growth
- Volume of data growth
- Implementation of new applications
- Applications modification

Sources of Data

- Historical Trend - Regression analysis
- Business plan
- Application implementation plan
- Application modification plan
- Workload characterization on production and test platforms

Workload Forecasting

Trend Analysis



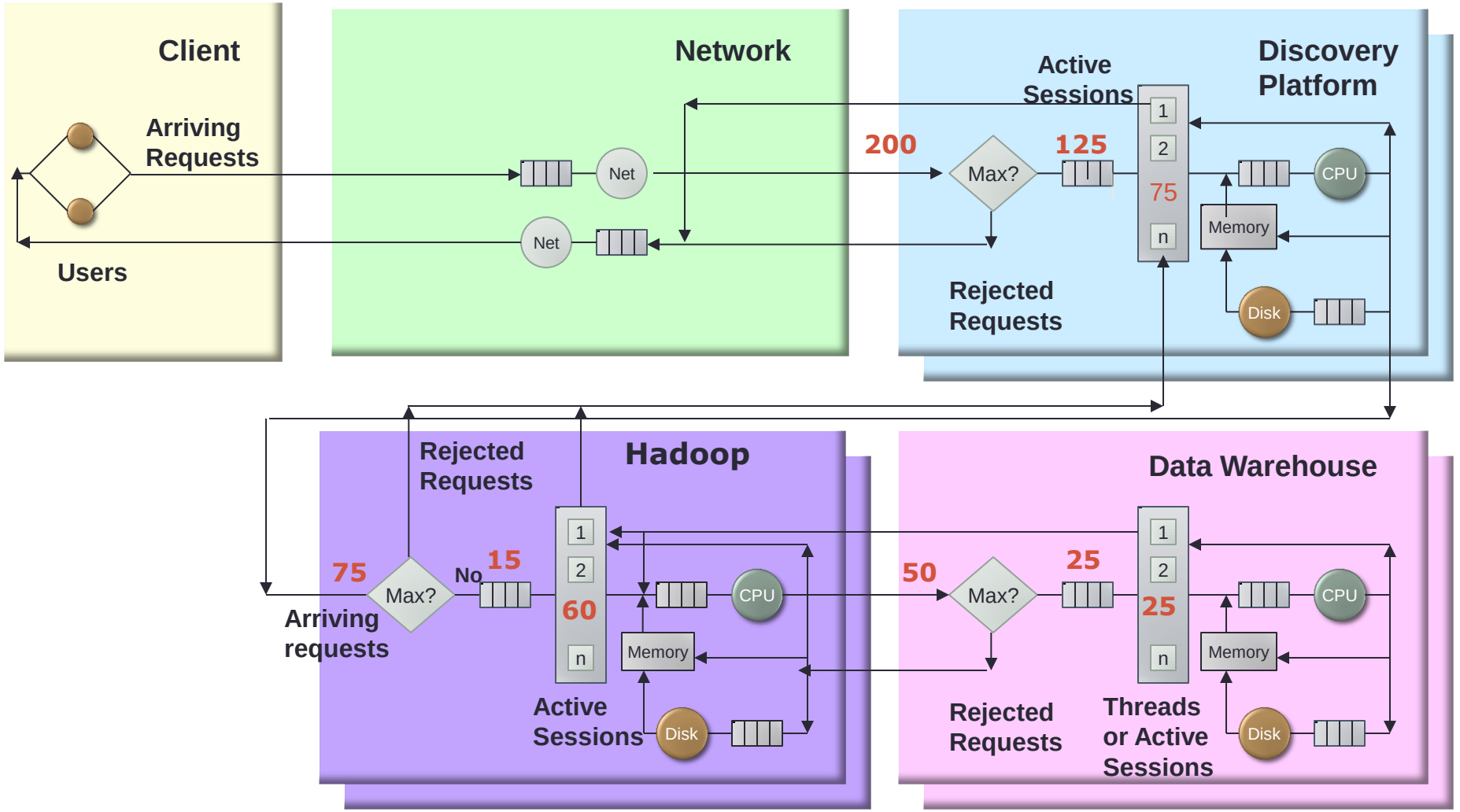
Increase in Number of Requests

- Workload and volume of data growth impact
- New application implementation impact
- Server consolidation impact
- How to predict how long will it take to run M/R job?
- How to change scheduling algorithm and adjust workload management parameters proactively?
- How to change dynamically virtual and physical configurations to meet SLGs for the individual workloads?
- Justification of decisions and Verification of results

PERFORMANCE PREDICTION

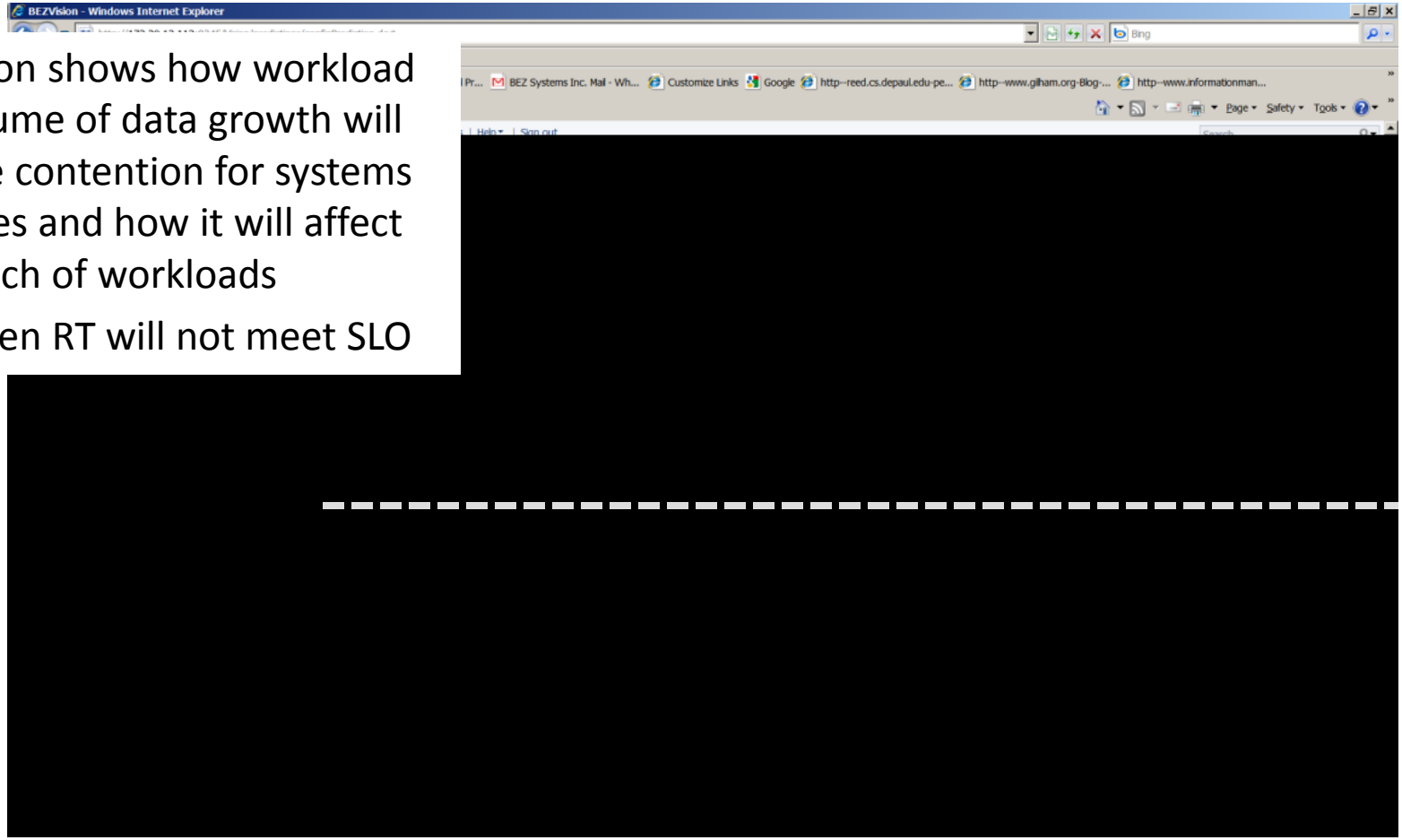
Performance Prediction

Simplified Queueing Network Model

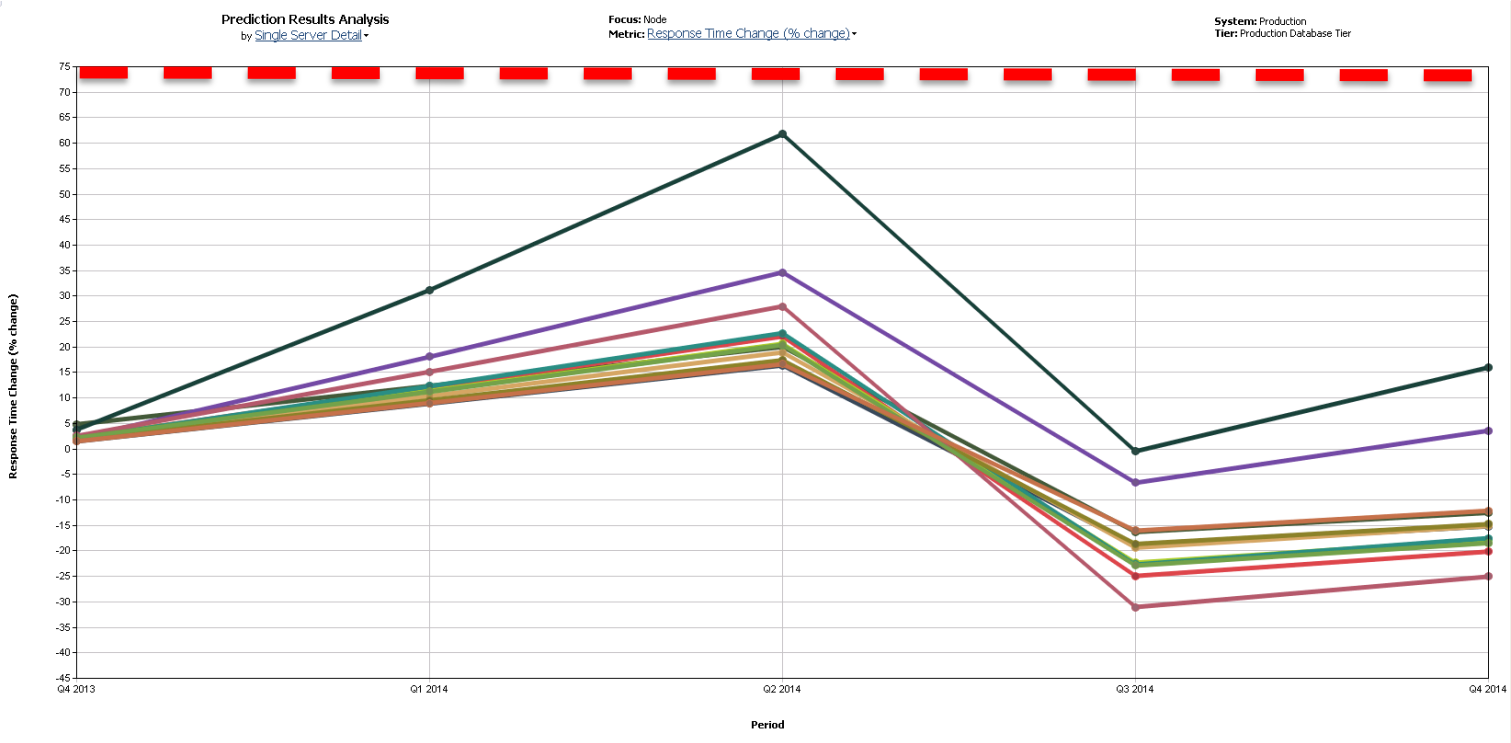


Predicting Workload and Volume of Data Growth Impact on Response Time

- Prediction shows how workload and volume of data growth will increase contention for systems resources and how it will affect RT of each of workloads
- Find when RT will not meet SLO

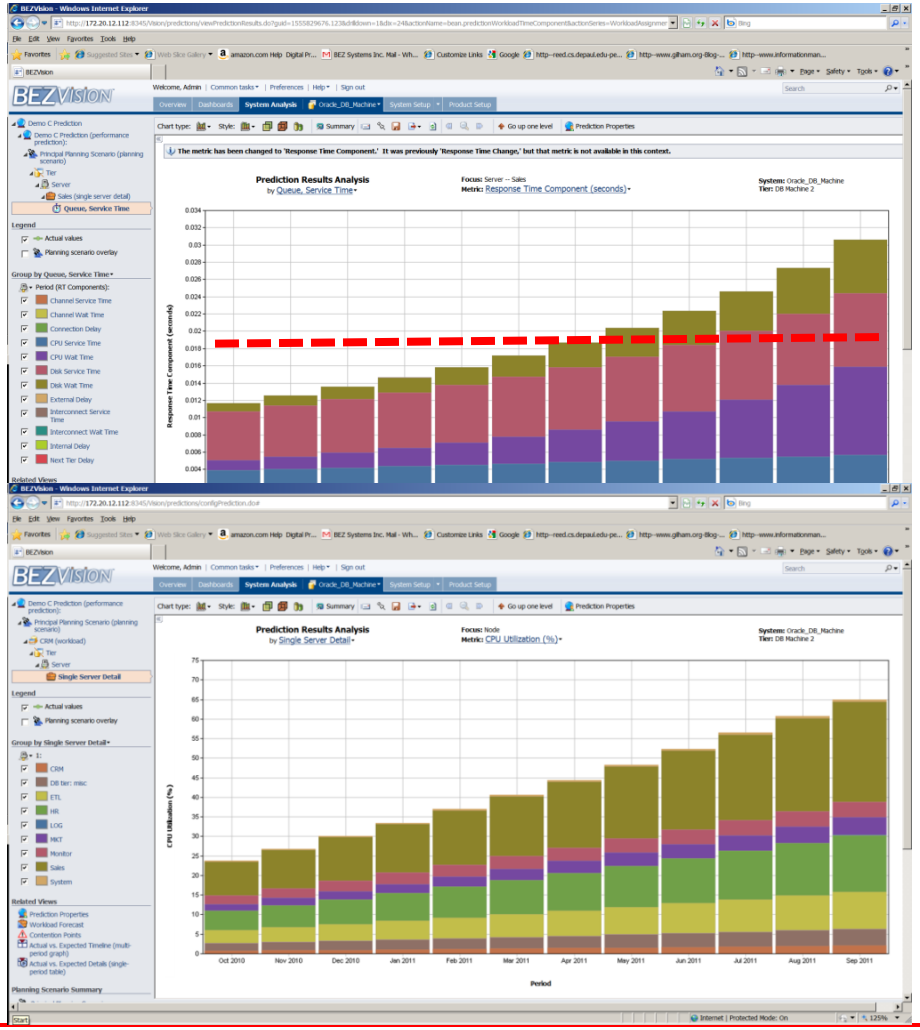


Justification of Required Changes to meet SLGs for all workloads



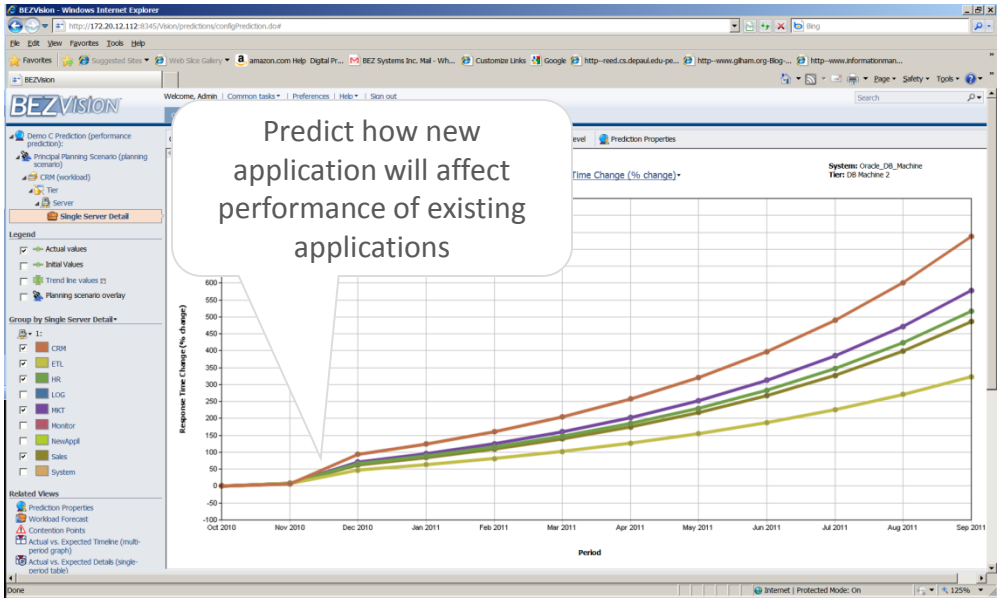
What is the Largest Component of the Response Time

- Find when SLO will not be met
- Find which workload will use most of CPU resources
- Identify options how to improve performance



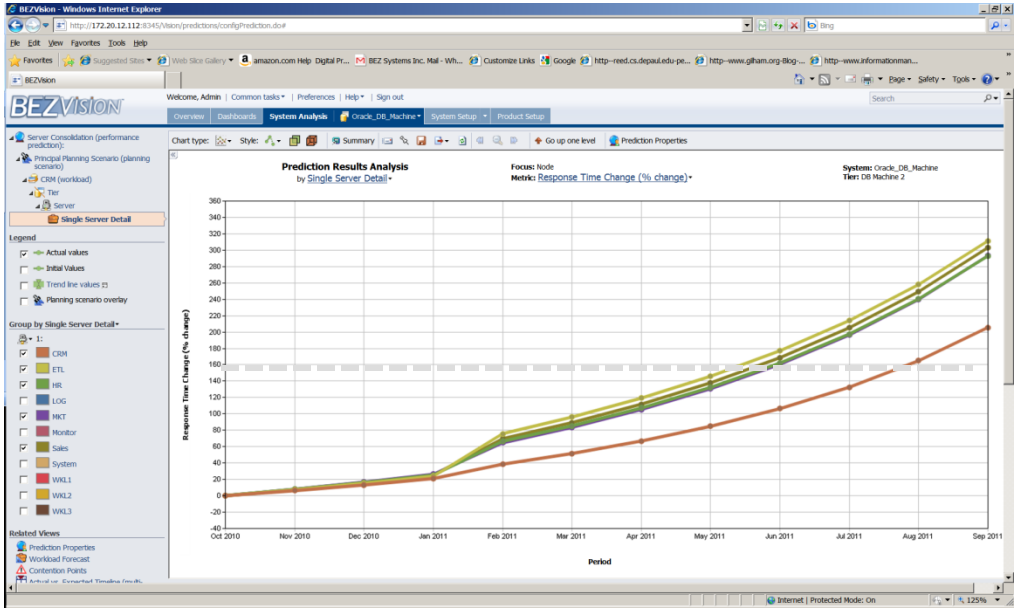
Predicting How New Application Will Impact Existing Workloads

- Simulation of moving workloads from test to production system predict how new workload will affect performance of the existing workloads
- Model take into consideration differences between hardware and software platforms, differences in volume of data, etc.
- Set realistic expectations and justify what should be changed proactively



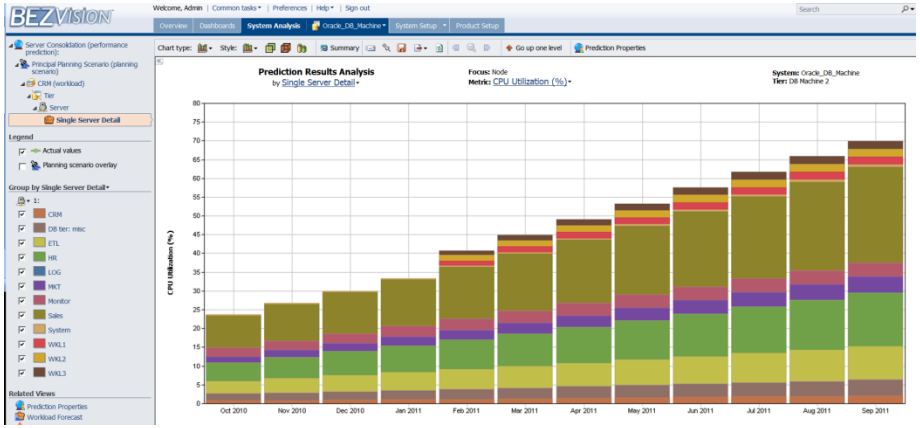
Predicting How Server Consolidation Will Affect Performance

- Prediction results evaluate the impact of planned server consolidation
- Shows when system will not meet SLOs
- Identify the minimum upgrade required to support SLOs



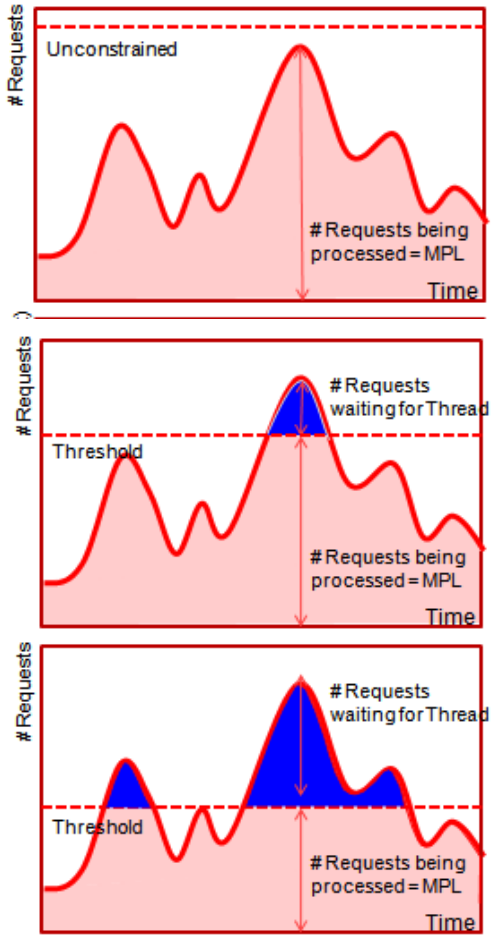
Predict Response Time and CPU Utilization After Consolidation

- Prediction results show how different workloads will perform after server consolidation and how it will affect CPU utilization



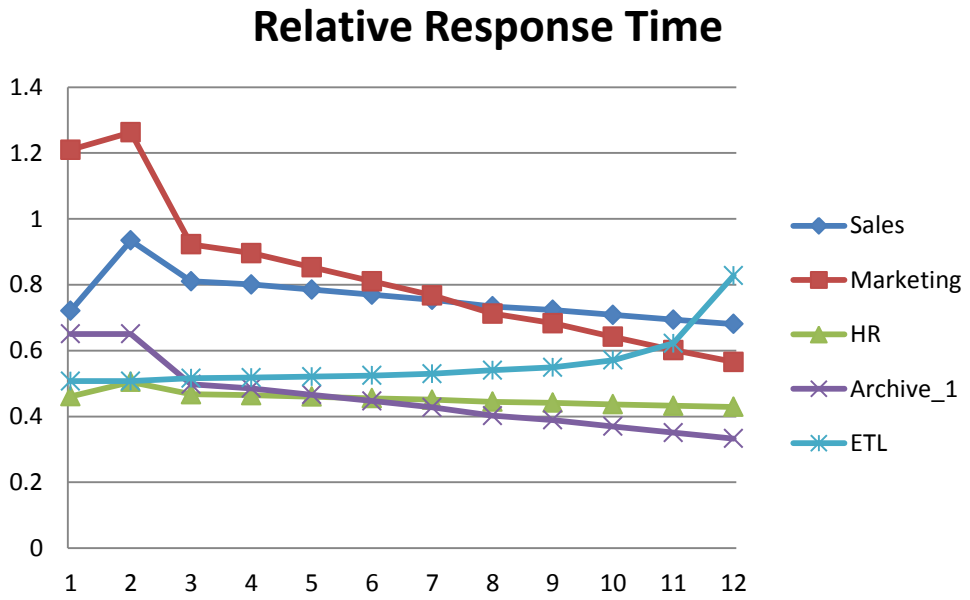
Limiting Concurrency Reduces Contention but Increases # of Requests Waiting for the Thread

- Limiting Concurrency for the workload can reduce contention for resources
- Requests of the workload with limited concurrency will spend less time waiting for resources, but spend more time waiting for the thread
- Performance of the workload with limited concurrency might suffer, but other workloads can have significant performance gain



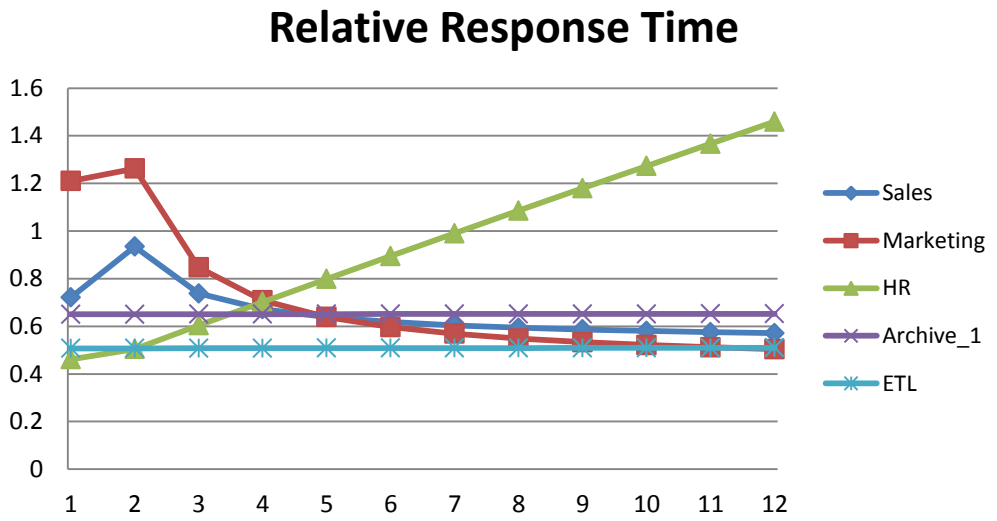
Predicting impact of lowering the level of concurrency for Load workload

- Data Load use a lot of resources, but satisfy SLO
- What if we limit Load concurrency starting Period #3?
- Load time will increase, but will be satisfactory
- Response time for other workloads will improve

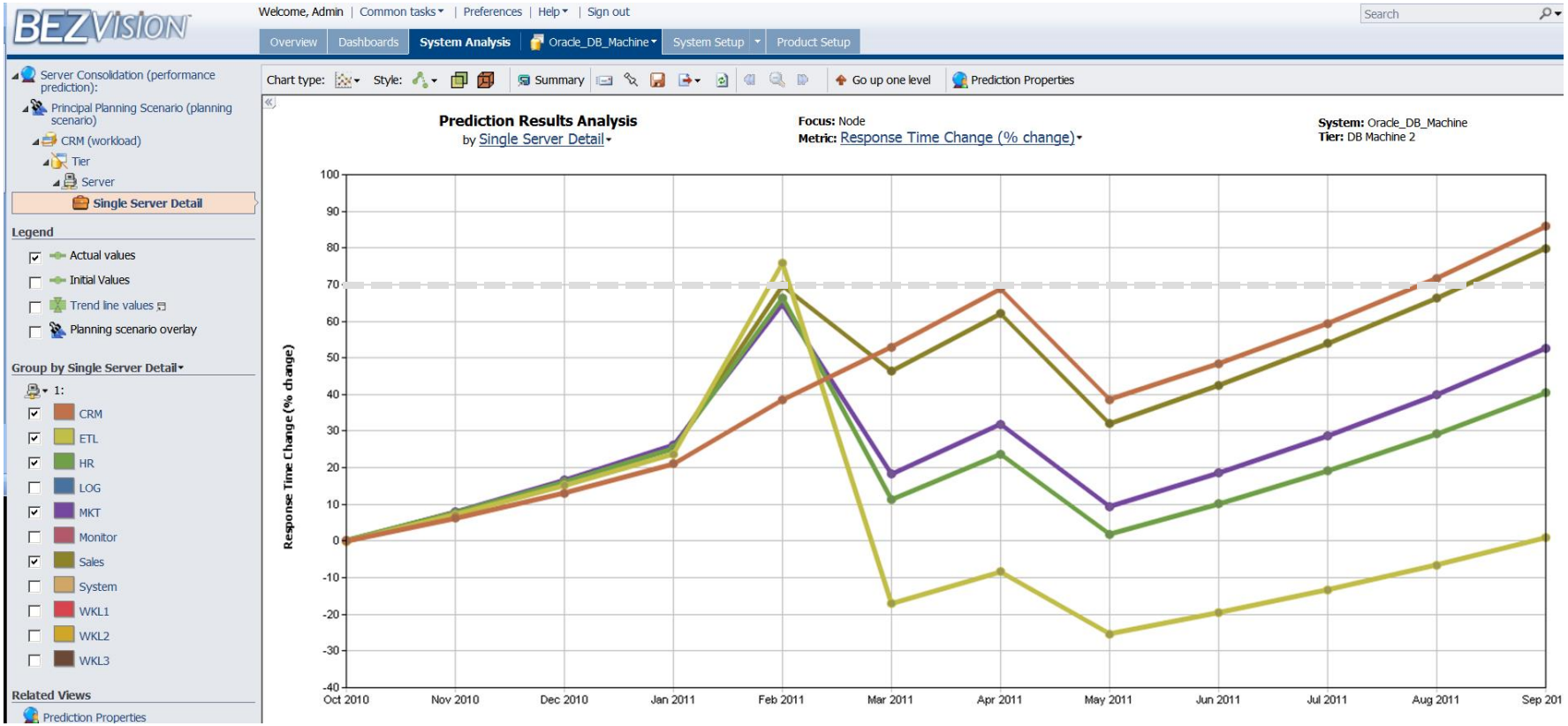


What will be an impact of workload priority change?

- Increase Priority for the critical Workloads will Improve their performance but negatively affect others
- Prediction results evaluate different alternatives and provide valuable information to justify proactive decisions

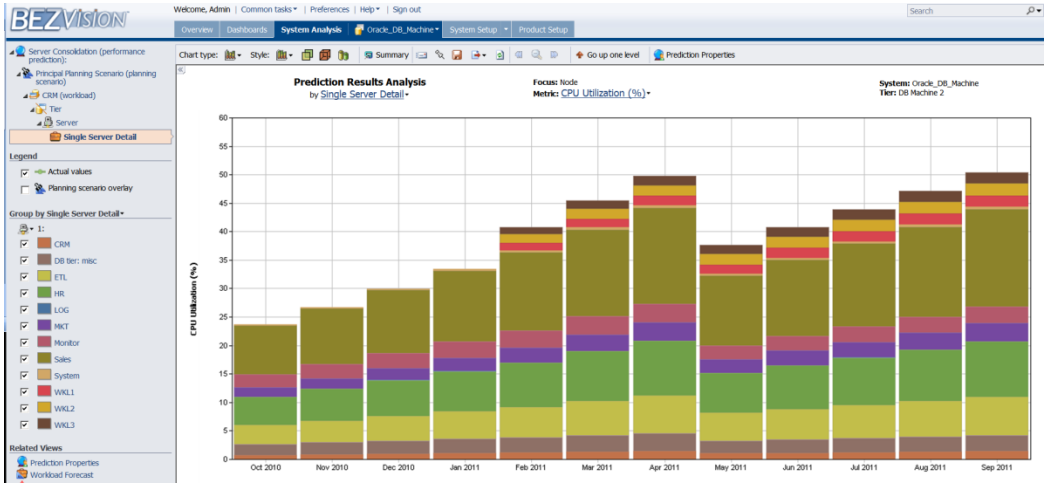


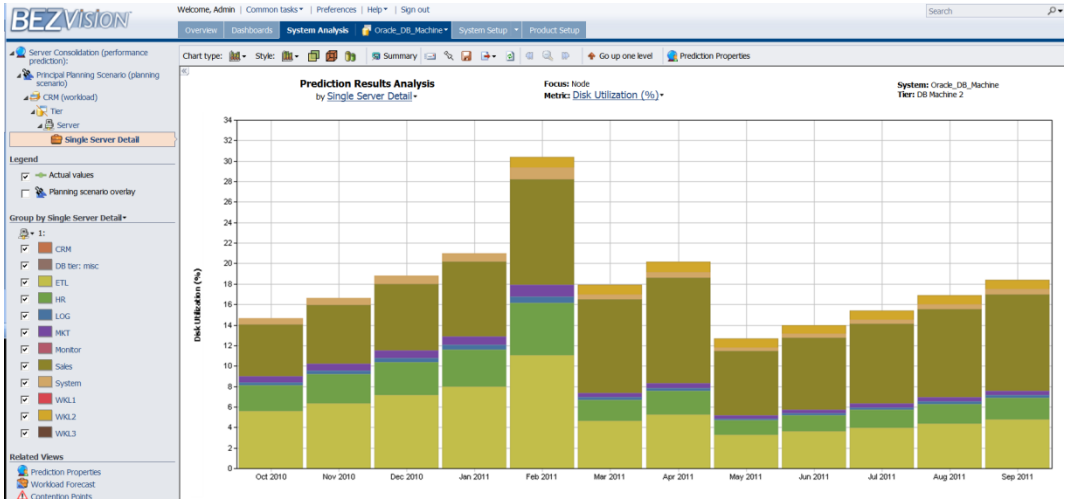
Predicted Impact of the Hardware Upgrade



Impact of Hardware Upgrade on CPU and Disk Utilization

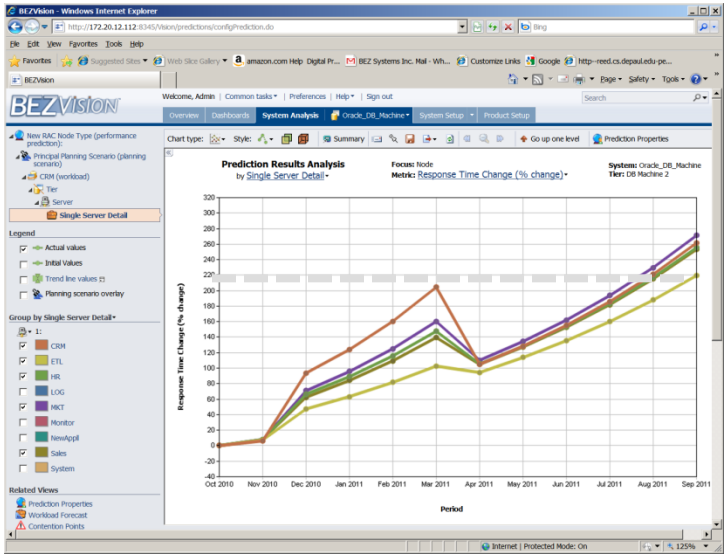
- Hardware upgrade reduce contention and improve performance

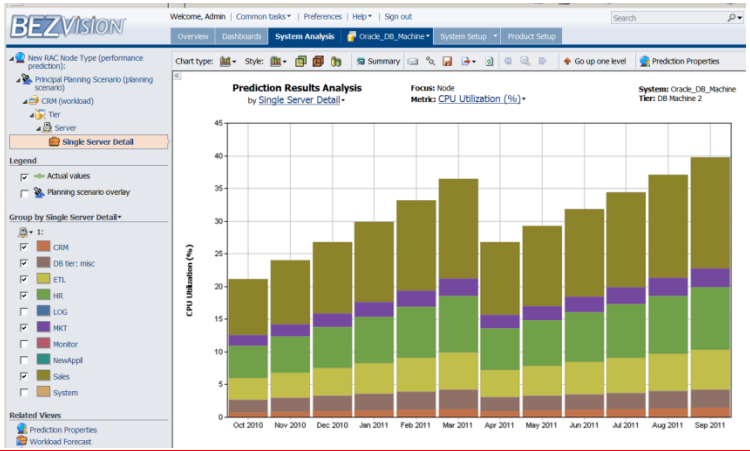




Predicted Impact of Changing Number of Processors per Node by 50%

- Increase node capacity will have positive impact on response time and reduce CPU utilization





Comparing Actual Results with Expected

VERIFICATION

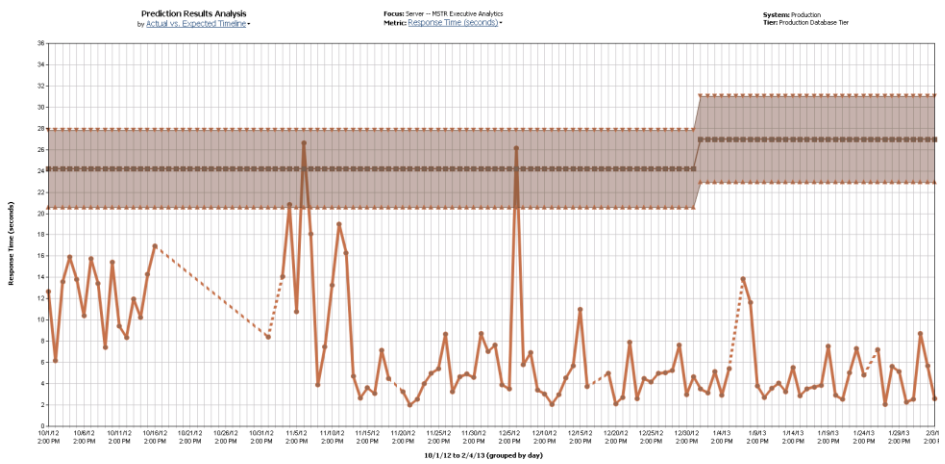
Comparing Actual Measurement Results with Expected (A2E)

- Performance prediction results provide
  - Justification of proactive actions
  - Performance (RT and Throughout) and Resource Utilization (CPU, I/O, Internode Communication, etc) expectations for each workload as a result of expected growth and implementation of recommended changes
- Comparison of the actual results with expected verifies accuracy of predictions and validity of growth assumptions
- If difference is significant Root cause analysis identify the reason
- Foundation for organizing continuous capacity management process
- How can you manage if you do not know what to expect

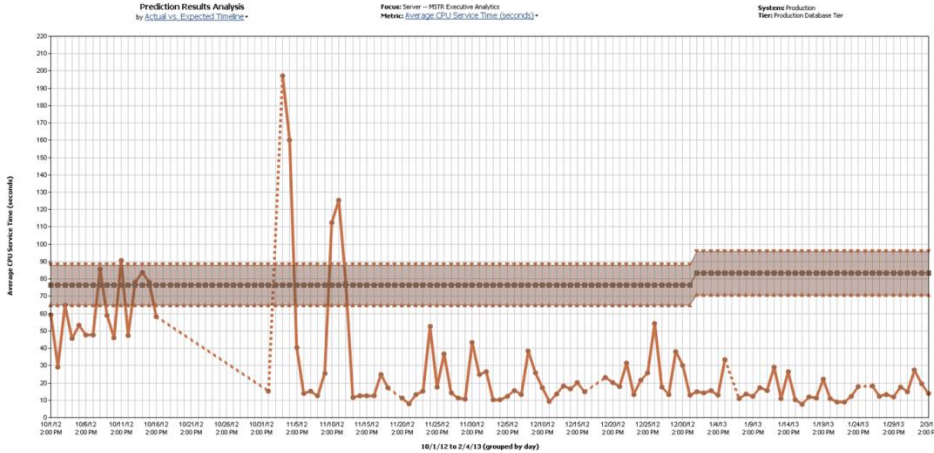
Verification of Results (cont)

increase in complexity caused increase in CPU utilization and RT A2E

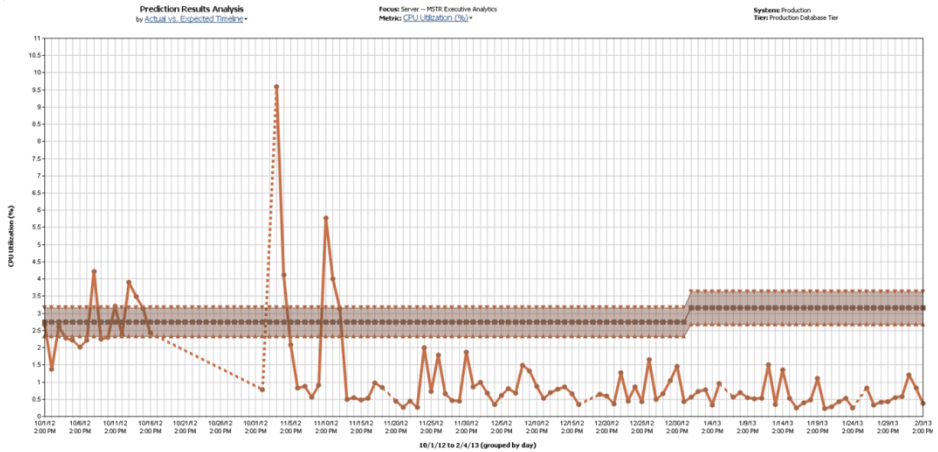
Response Time



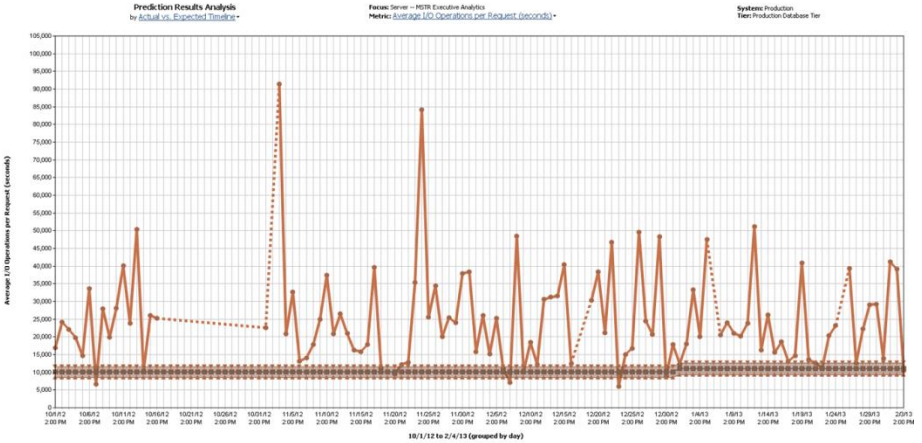
CPU Service Time



CPU Utilization



I/O Operations/Request



Summary

- We discussed application of Big Data and predictive analytics for organizing collaboration between business representatives and IT for
  - evaluation options, justification and verification of decisions
  - setting realistic service level goals and performance expectations
  - to review assumptions, modeling results, justify actions and review the comparison between actual results and expected
- We reviewed methodology and case study illustrating typical steps of applying big data predictive analytics, including,
  - Data Collection, Workload Characterization, Workload Forecasting, Performance Prediction and Verification of Results
- We reviewed how predictive analytics and optimization enables better alignment of business and IT
  - through justification of strategic, tactical and operational IT actions
- We discussed how ML and predictive analytics can be used for organizing a collaborative environment between business representative, IT and management

Thank you!

bzibitsker@beznext.com

www.beznext.com