BIG DATA ANALYTICS CHALLENGES AND OPPORTUNITIES

Dr. Boris Zibitsker, CEO BEZNext
Dr. Dominique Heger, CEO Data Analytica

4th International Conference and Expo on Big Data Analytics
Minsk 2018

Outline

1. History of Big Data Analytics Conferences in Belarus
2. Current Challenges and Problems
3. Enterprise IT Performance Assurance Solutions
   - Performance Engineering
   - Dynamic Performance Management
   - Long Term Capacity Planning
4. Summary

1. HISTORY

BIG DATA ANALYTICS CONFERENCES IN BELARUS

Big Data Analytics Conferences in Belarus

- 2013 - Start Big Data Training at BGUIR
- 2014 - Decision to organize Big Data Analytics conference
- 2015 - 1st Conference / Map Reduce Batch Processing
- 2016 - 2nd Conference / Real Time processing using Spark and Kafka
- 2017 - 3rd Conference / Deep Learning and Internet of Things

2. BIG DATA CHALLENGES

4TH CONFERENCE

2018 – Big Data Analytics

Challenges

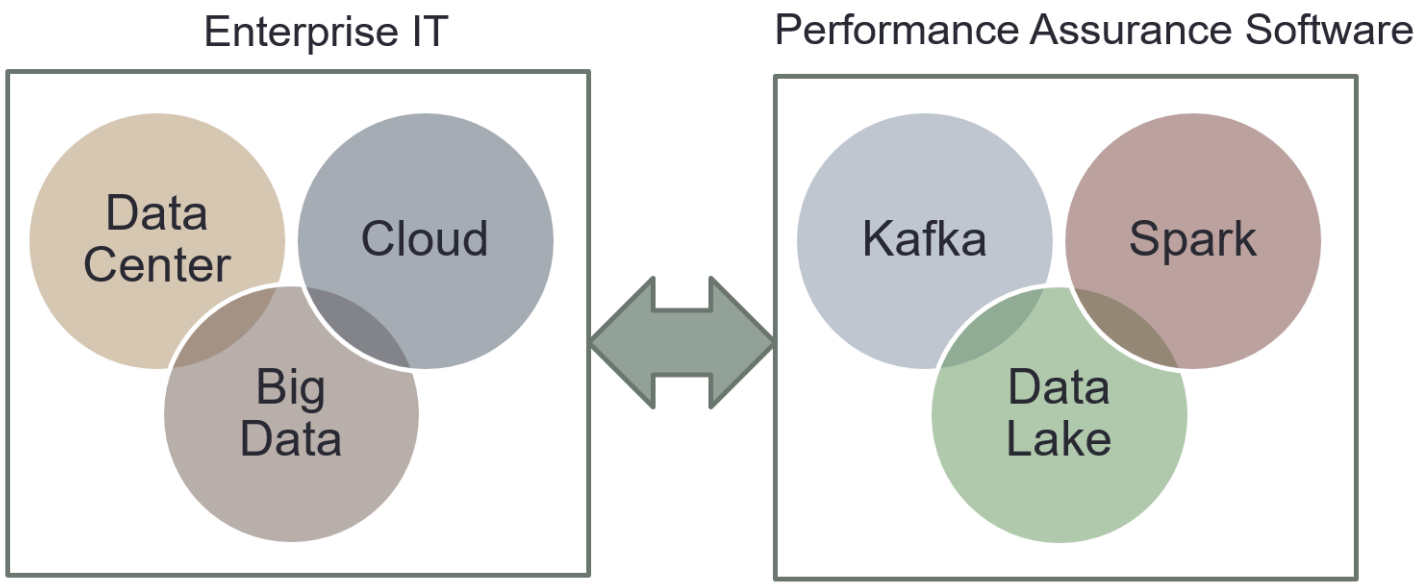
- Unrealistic expectations
- Complexity
- Performance, scalability and cost
- Manageability
- New technology / Blockchain
  - Supply Chain, FinTech, Identity Management, IOT/IOE, Crypto Commerce, Agri-Tech

Solutions and Conference

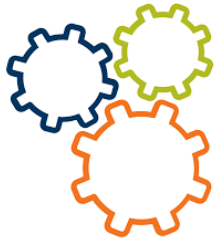
- Artificial Intelligence
- Automation
- Data Streaming
- Operationalization
- Big Data and Cloud
- Unsupervised Deep Learning
- Practical Education

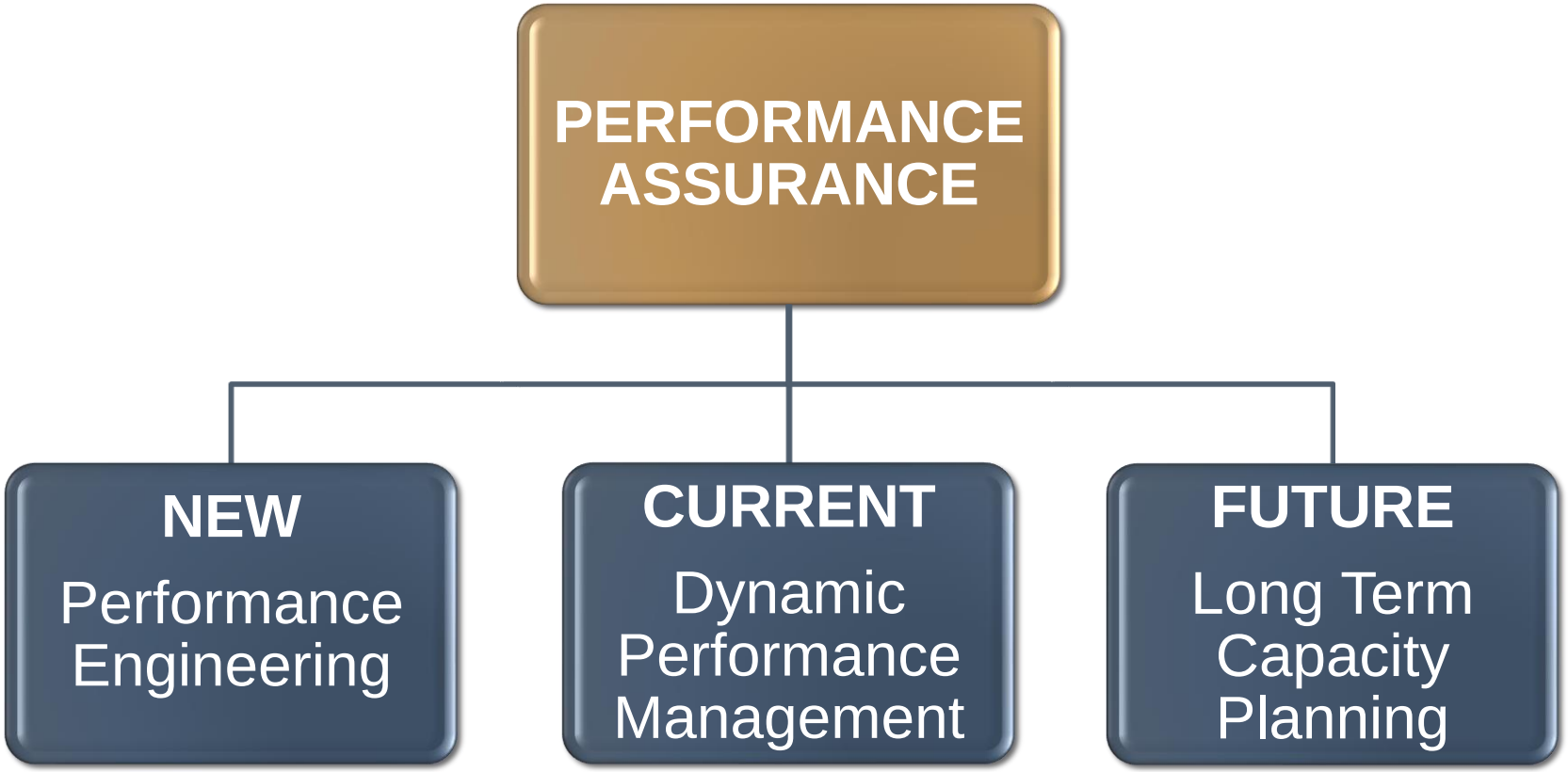
3. ENTERPRISE IT PERFORMANCE ASSURANCE

OVERVIEW OF BEZNEXT SOLUTIONS



Performance Assurance

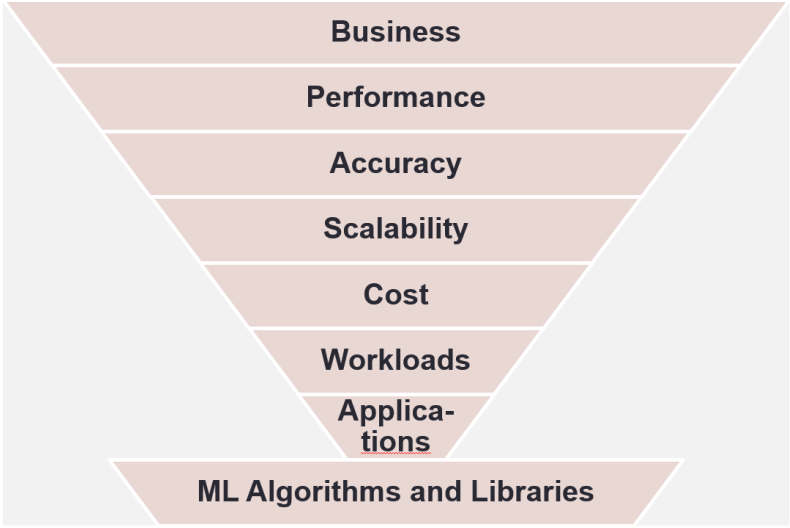




3.1. PERFORMANCE ENGINEERING

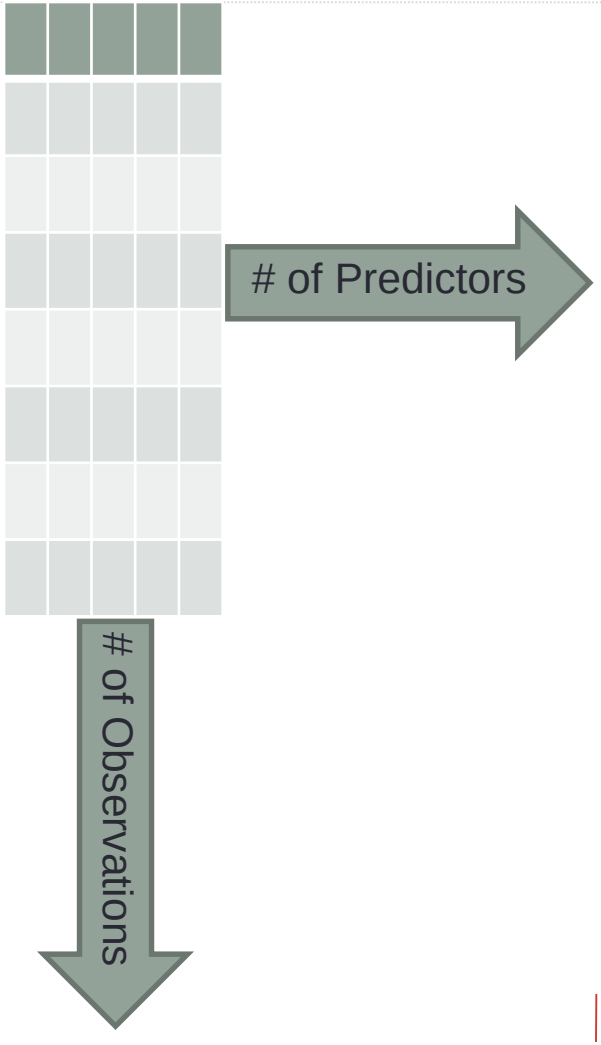
NEW APPLICATION DESIGN AND DEVELOPMENT

- ML Algorithms and ML Libraries Selection

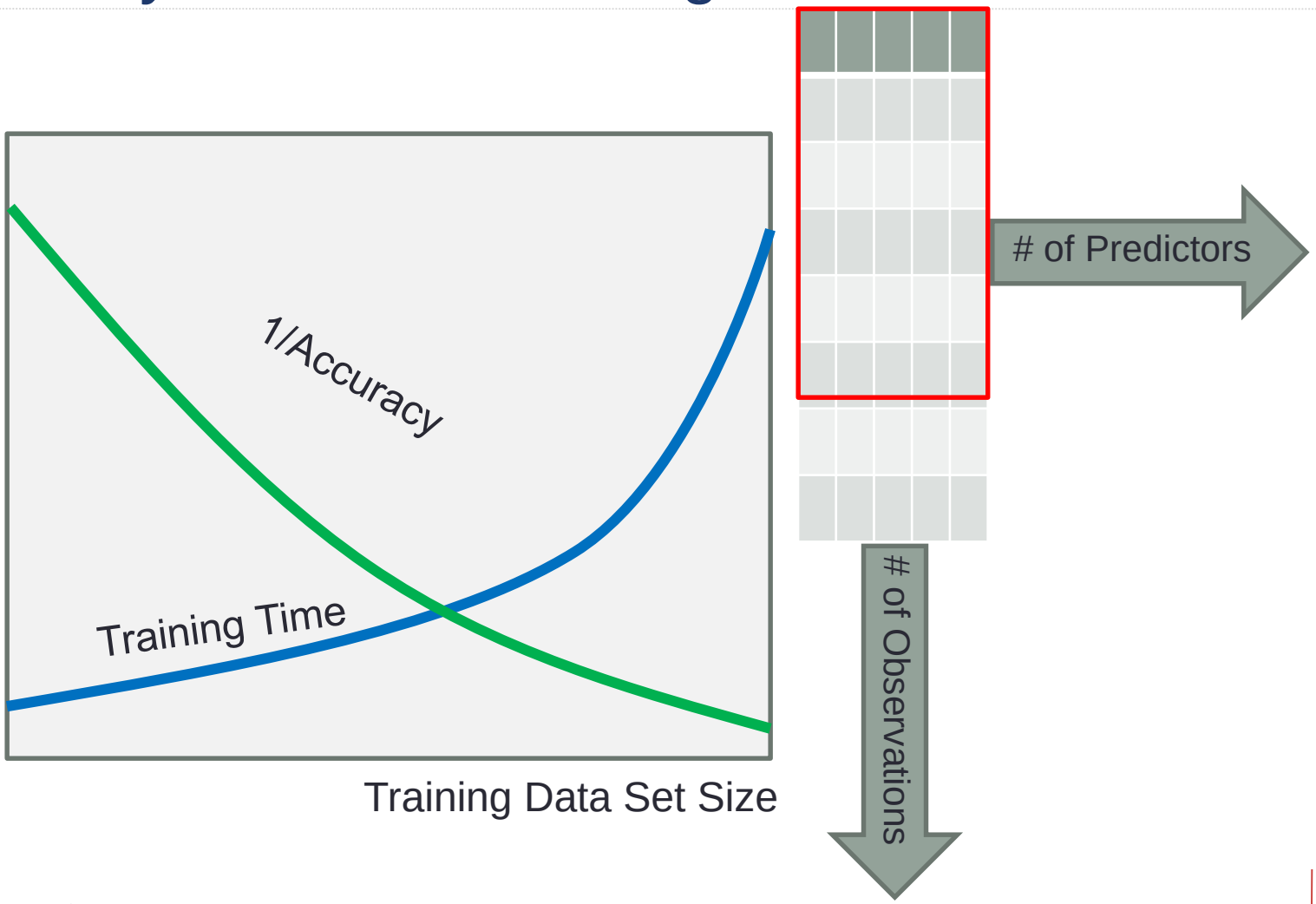


Benchmarking and Modeling

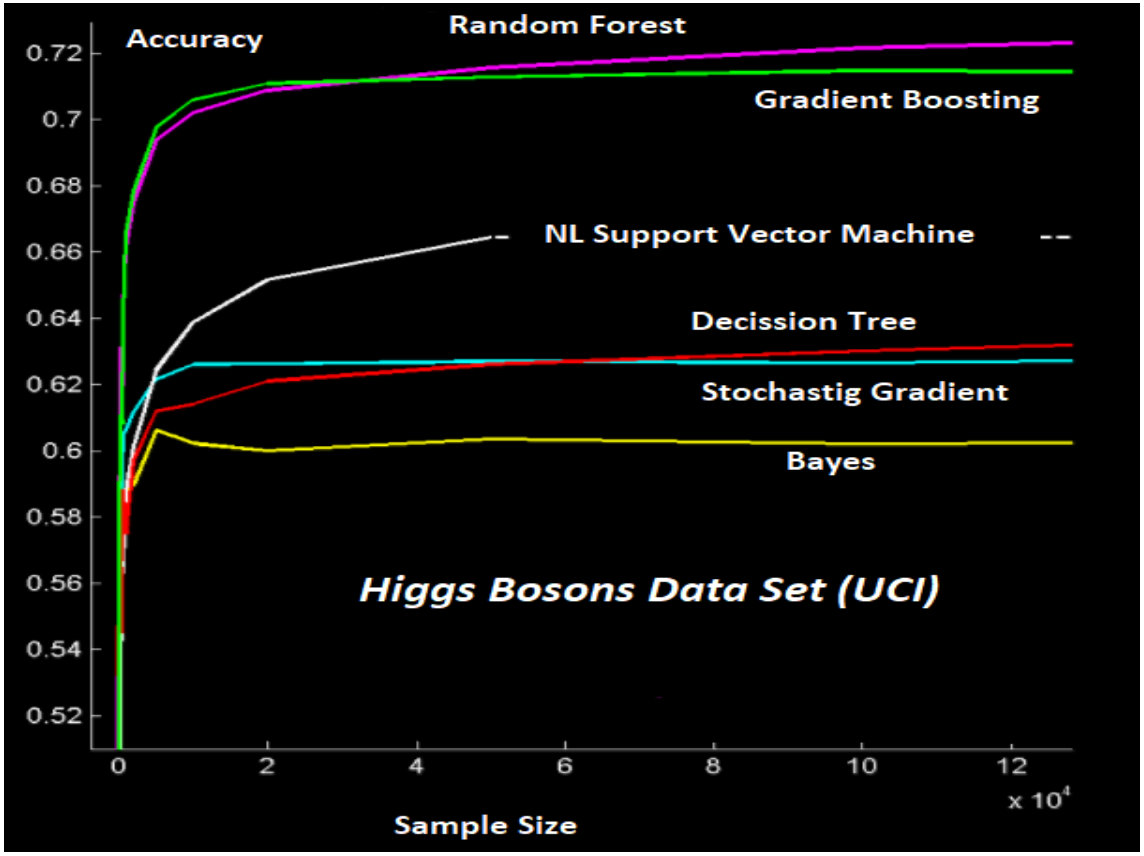
- Preparing Benchmark
- Running Benchmark
- Data Collection
- Modeling
- Building Knowledge Base
- Recommender



Selection of the Training Data Set Size affects Accuracy and Model Training Time

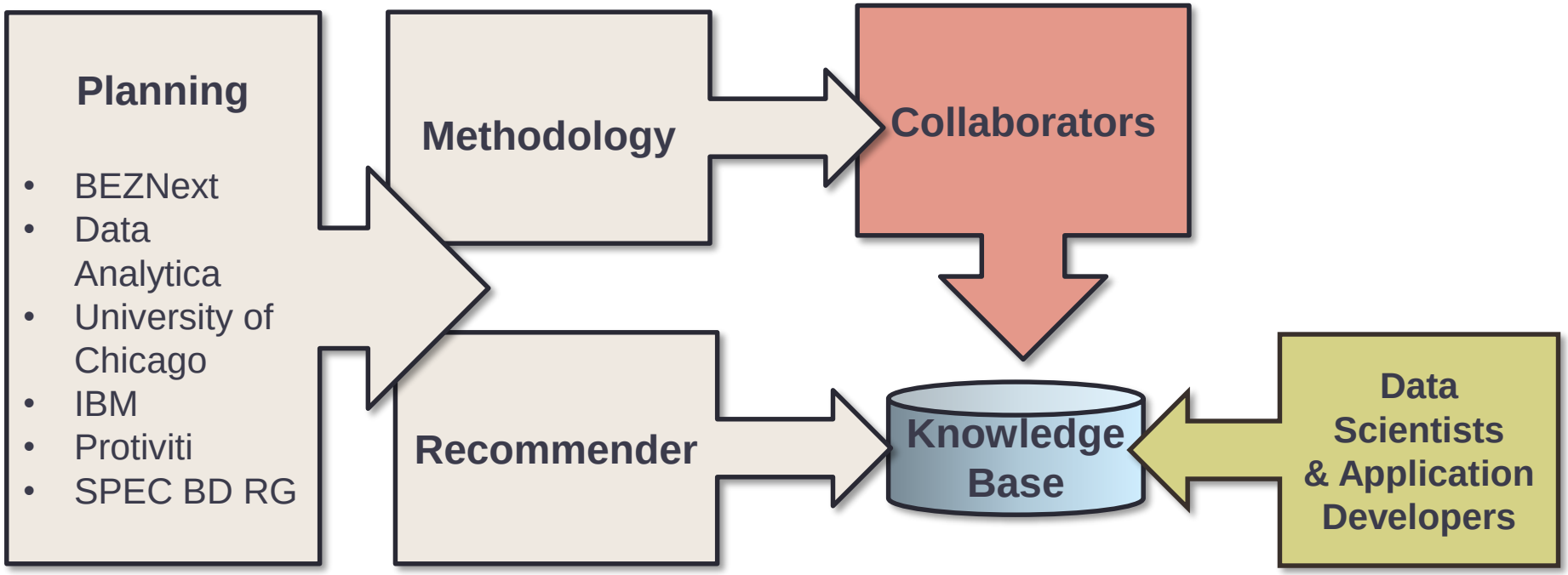


Benchmarking Results Shows the Impact of changing training Data Set size on Accuracy

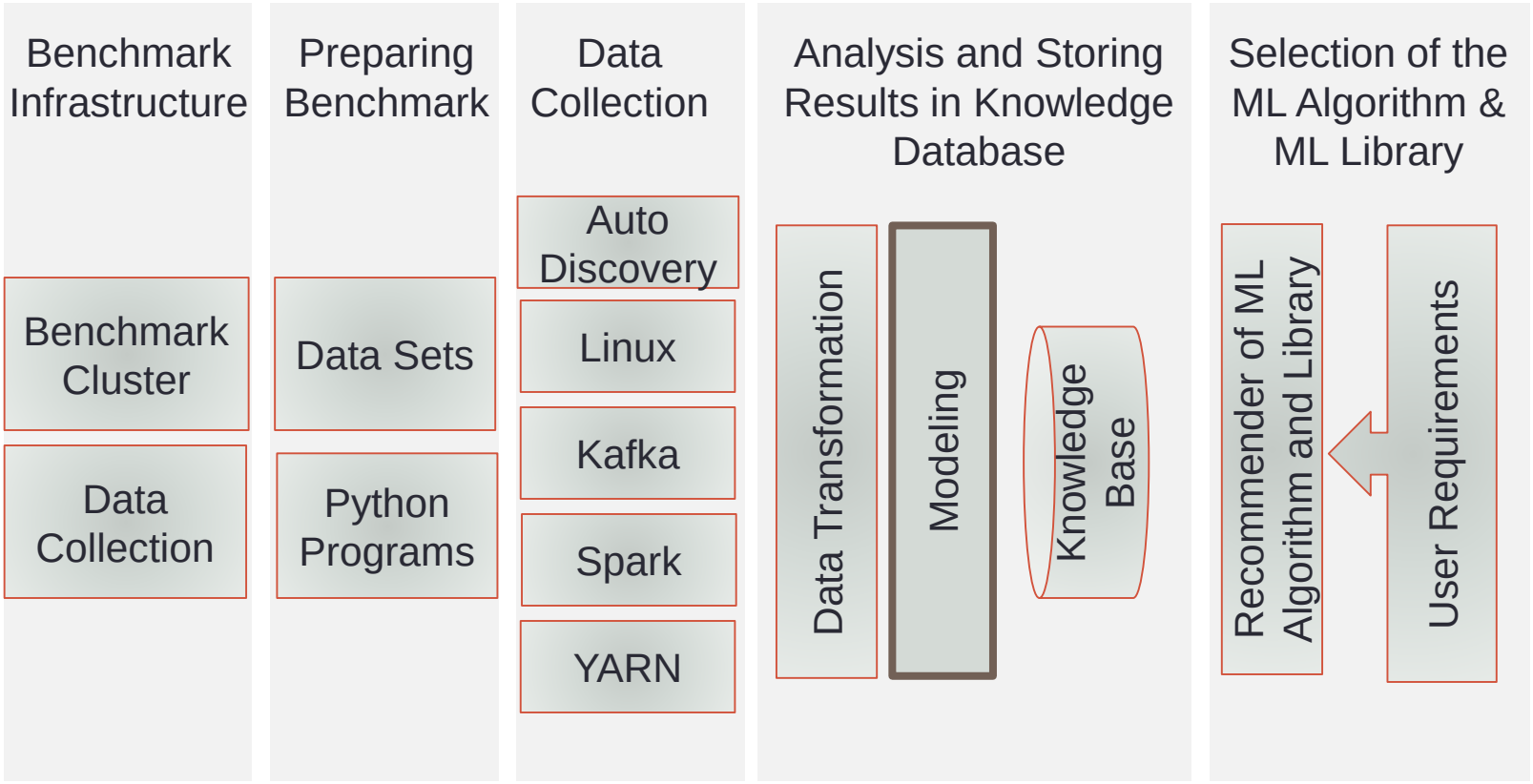


- Higgs Boson Data Set (Cern Benchmark) with 11m rows and 28 predictors https://archive.ics.uci.edu/ml/datasets/HIGGS

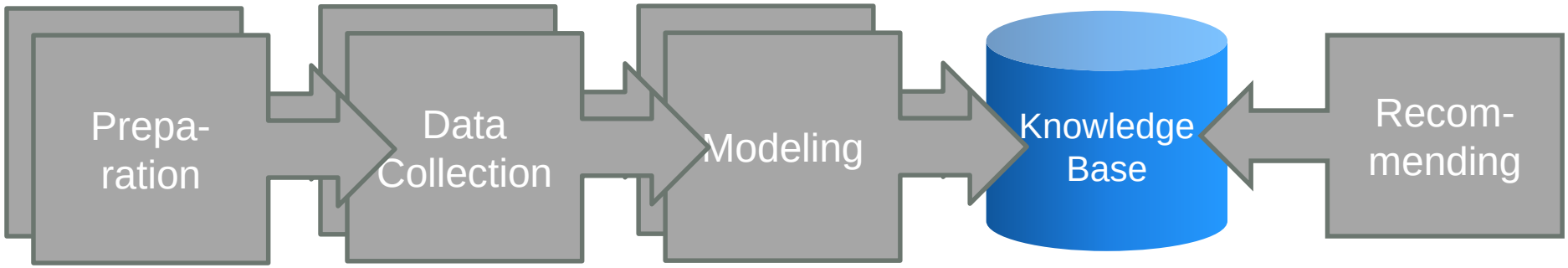
Collaborative Project



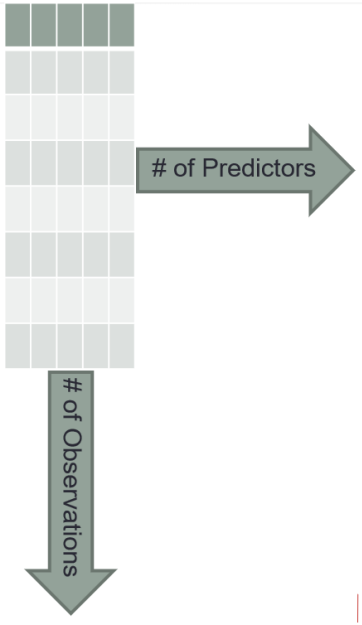
Benchmarking Process



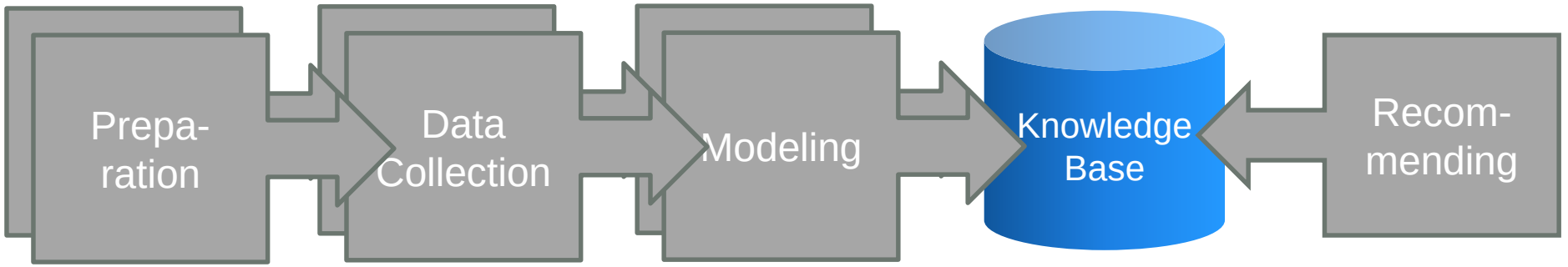
Benchmarking and Modeling



- Prepare Data Set
- Write Python Program
- Incorporate ML Algorithms and ML Libraries
- Hardware & Software



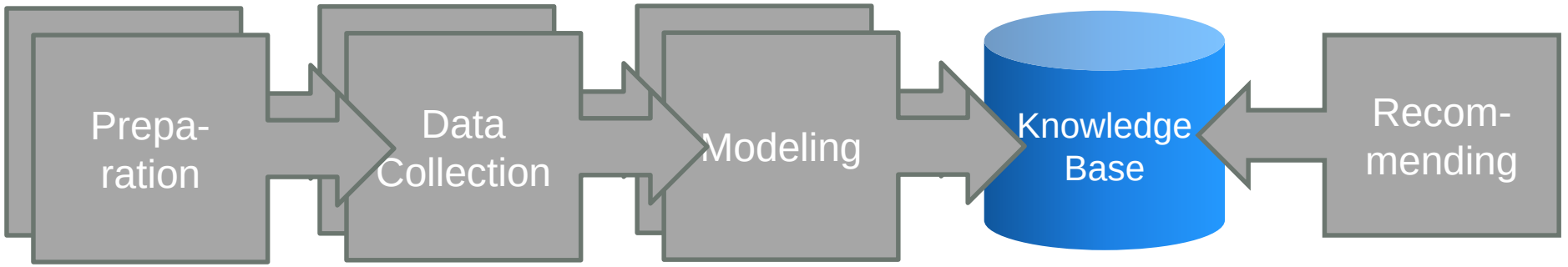
Benchmarking and Modeling



- Prepare Data Set
- Write Python Program
- Incorporate ML Algorithms and ML Libraries
- Hardware & Software

- RT
- Throughput
- Resource Utilization
- Accuracy

Benchmarking and Modeling

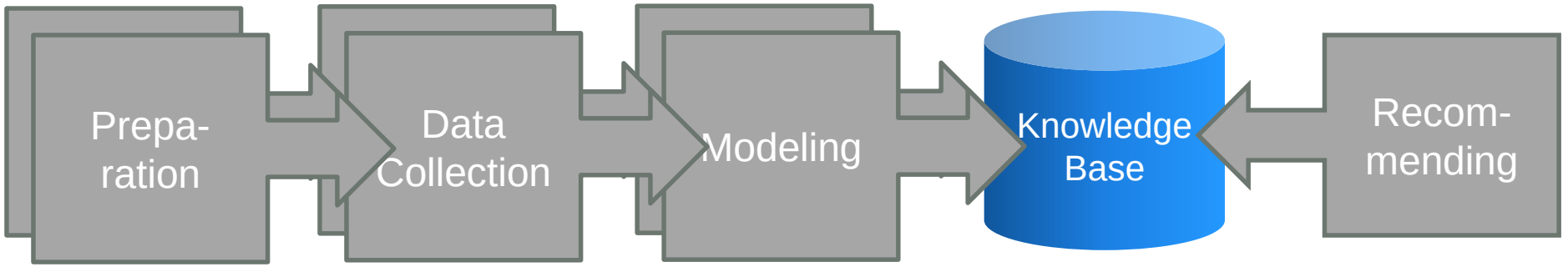


- Prepare Data Set
- Write Python Program
- Incorporate ML Algorithms and ML Libraries
- Hardware & Software

- RT
- Throughput
- Resource Utilization
- Accuracy

- Expand Benchmark
- Differences in Hrdw and Software

Benchmarking and Modeling



- Prepare Data Set
- Write Python Program
- Incorporate ML Algorithms and ML Libraries
- Hardware & Software

- RT
- Throughput
- Resource Utilization
- Accuracy

- Expand Benchmark
- Differences in Hardware & Software

- Define Business Needs
- Define Weight for each criteria
- Calculate Score Select ML algorithms & ML Libraries

Example of Score Calculation Considering Business Requirements

- The Score takes into consideration the type of ML algorithm, Number of Observations and Features / Predictors in Data Set, the relative importance of the different criteria, like response time, Accuracy, CPU Utilization, Memory utilization, Number of I/O operations, and other parameters:

Score = w1 * Accuracy + w2 * Response Time + w3 * CPU Utilization + w4 * Memory Utilization +w5 * Scalability , etc

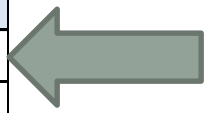
Where the weighting coefficients wi represent business priorities between 0 and 1.

Type of Requirement	Relative Weight
Accuracy	0.2
Response Time	0.4
CPU Utilization	0.2
Memory Utilization	0.2
Total	1

Score is Used to Recommend Appropriate ML Algorithm

- Response Time can vary between 0 and infinity. We transform the response time as 1 / (1 + RT) to make it as a number between 0 and 1, where 1 is better. In addition to calculating the score we check if predicted CPU Utilization and Memory Usage are less than 1. -
- Value of score is used to recommend the appropriate ML algorithm and ML Library.

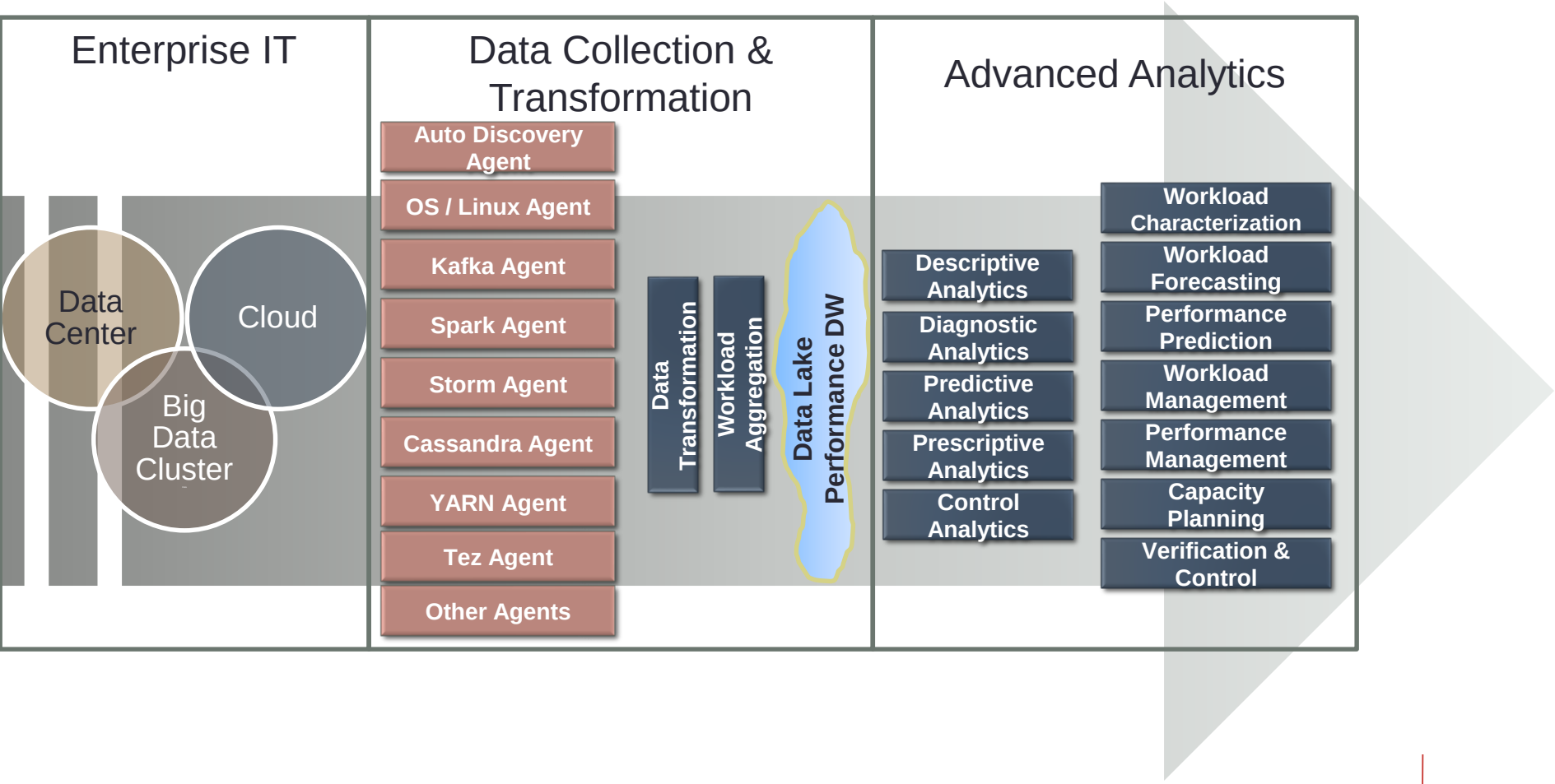
Algorithm	library	pred_score	pred_rank	true_score	true_rank
OLS	Python Sklearn	0.962057911	1	0.936165261	1
OLS	Pyspark ML	0.876712666	2	0.753752225	2
Ridge	Python Sklearn	0.781980143	3	0.725268522	3
Ridge	Pyspark ML	0.722426161	4	0.659234146	4
RF	Python Sklearn	0.476284999	5	0.429752013	5
RF	Pyspark ML	0.465422159	6	0.415271967	6



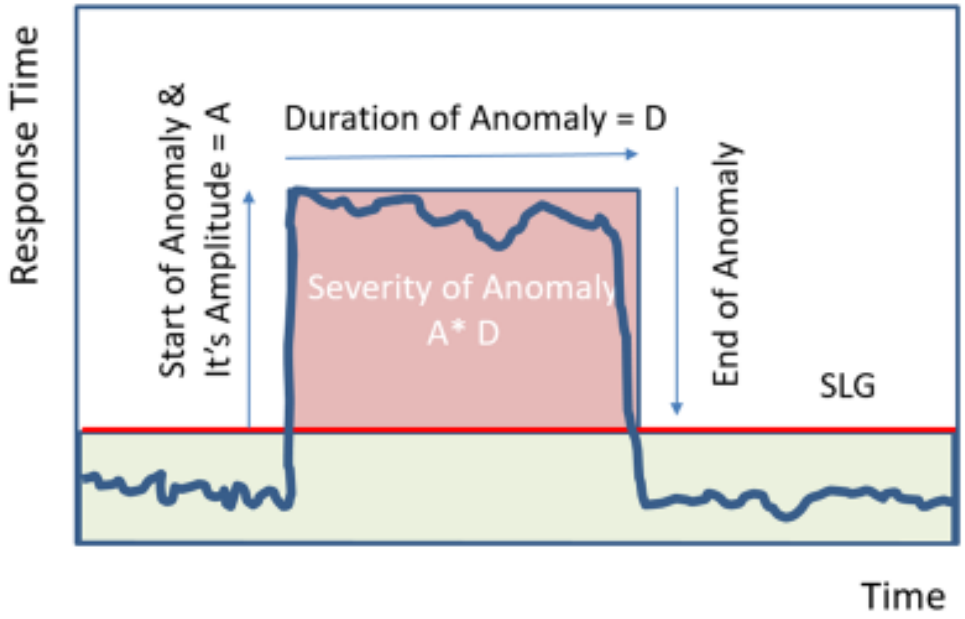
- ML OLS Algorithm using Python Sklearn ML library is the most appropriate algorithm to satisfy business requirements presented in example above.

3.2. DYNAMIC PERFORMANCE MANAGEMENT

Enterprise IT Performance Management and Capacity Planning



Anomalies Detection and Root Cause Determination



Anomaly Amplitude - A, Duration - D and Severity S = A * D.

Anomaly Detection (Single day overview)

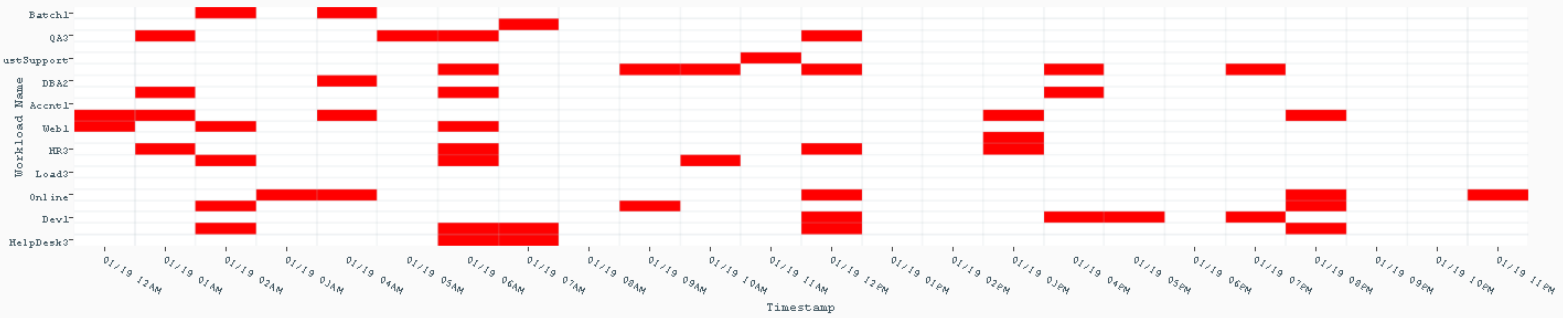
Anomalies Summary

Show 10 entries Search: 01/19/2016

Date	Time	Workload Name	Parameter Name	Expected Value	Actual Value	Severity
01/19/2016	01AM	HR3	# Req/Hour	19,085.15	86,505.00	3.53
01/19/2016	01AM	DBA1	# Req/Hour	661.60	3,265.00	3.93
01/19/2016	01AM	Sales2	# Req/Hour	518.55	2,279.00	3.39
01/19/2016	01AM	QA3	# Req/Hour	11.75	46.00	2.91
01/19/2016	02AM	Dev2	# Req/Hour	3,123.89	19,660.00	5.29
01/19/2016	02AM	Stream2	# Req/Hour	20,495.67	171,400.00	7.36
01/19/2016	02AM	Load2	# Req/Hour	24,818.62	57,438.00	1.31
01/19/2016	02AM	Web1	# Req/Hour	804.70	2,748.00	2.41
01/19/2016	02AM	Batch1	# Req/Hour	5,940.67	15,052.00	1.53
01/19/2016	03AM	Online	# Req/Hour	2,278.99	51,087.00	21.42

Showing 1 to 10 of 53 entries (filtered from 516 total entries) Previous 1 2 3 4 5 6 Next

Anomalies Hourly Distribution



Seasonal Peaks

Seasonal Peaks Determination Parameters

System ID: 1
Ruleset Name: 1
Parameter name: Total Executions Count
Workload Name(s): Select options
Threshold Value: 0.55
Start Date 2016/01/02 00:00
End Date 2016/01/20 23:00
Run

Seasonal Peaks Determination Results

Seasonal Peaks Summary

Show 10 entries Search:

Workload Name	Parameter Name	Peak Type	Peak Start Date	Duration	Avg Amplitude	Standard Deviation	Min Value	Max Value	95 Percentile	Peak Length STD
Accnt1	# Req/Hour	Daily	01/02 04AM	6	451.67	171.56	21.00	656.00	656.00	1.85
Admin1	# Req/Hour	Daily	01/02 05AM	1	8,740.43	3,256.4	1,225.00	11,909.00	11,909.00	0.00
BusDev1	# Req/Hour	Daily	01/02 02AM	1	10,393.75	6,801.83	16.00	23,335.00	20,568.60	0.00
CustSupport	# Req/Hour	Daily	01/02 03AM	1	6,687.75	2,974.55	27.00	15,504.00	7,235.60	0.00
Design	# Req/Hour	Daily	01/02 05AM	1	1,783.20	396.7	420.00	2,461.00	1,693.00	0.00
Dev1	# Req/Hour	Daily	01/02 04PM	1	90.75	30.49	27.00	129.00	129.00	0.00
HelpDesk3	# Req/Hour	Daily	01/02 11PM	1	91.00	30.62	9.00	108.00	108.00	0.50
HR3	# Req/Hour	Daily	01/02 05PM	1	66,391.88	47,305.03	64.00	205,571.00	146,633.30	0.50
HR4	# Req/Hour	Daily	01/02 11AM	1	711.33	387.09	14.00	1,095.00	1,095.00	0.00
Legal2	# Req/Hour	Daily	01/02 12AM	1	1,347.75	538.28	211.00	2,137.00	2,137.00	0.00

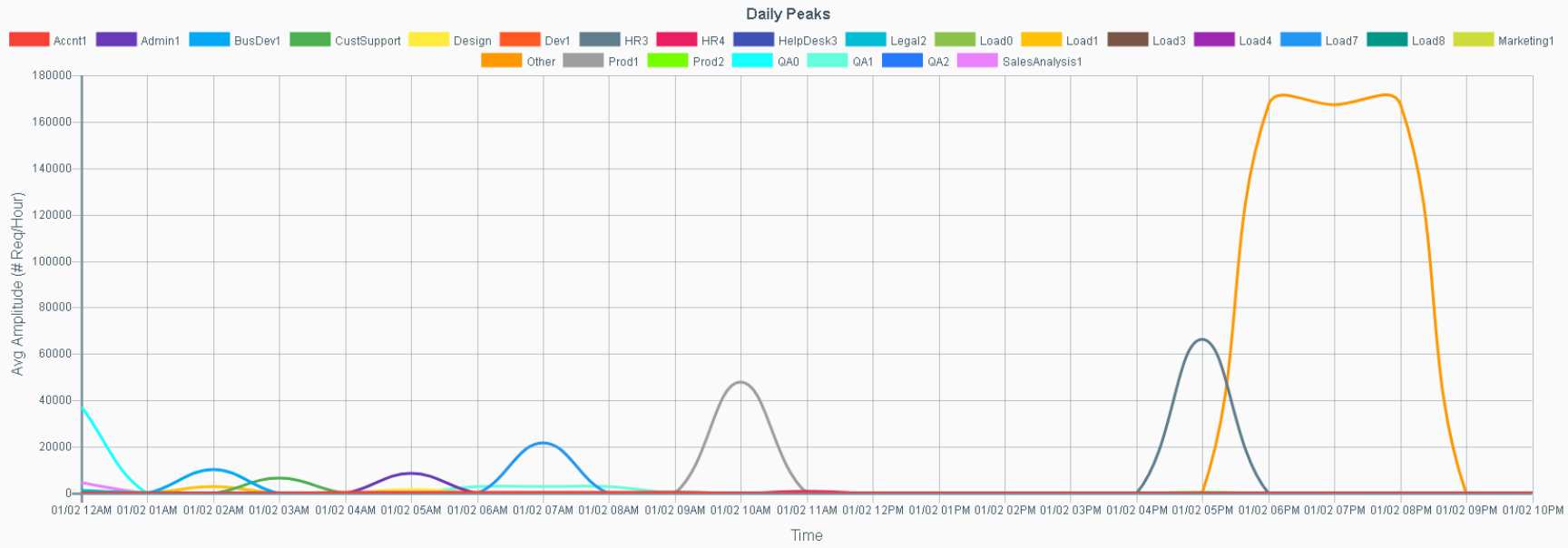
Showing 1 to 10 of 24 entries Previous 1 2 3 Next

Seasonal Peaks

HR3	# Req/Hour	Daily	01/02 05PM	1	66,391.88	47,305.03	64.00	205,571.00	146,633.30	0.50
HR4	# Req/Hour	Daily	01/02 11AM	1	711.33	387.09	14.00	1,095.00	1,095.00	0.00
Legal2	# Req/Hour	Daily	01/02 12AM	1	1,347.75	538.28	211.00	2,137.00	2,137.00	0.00

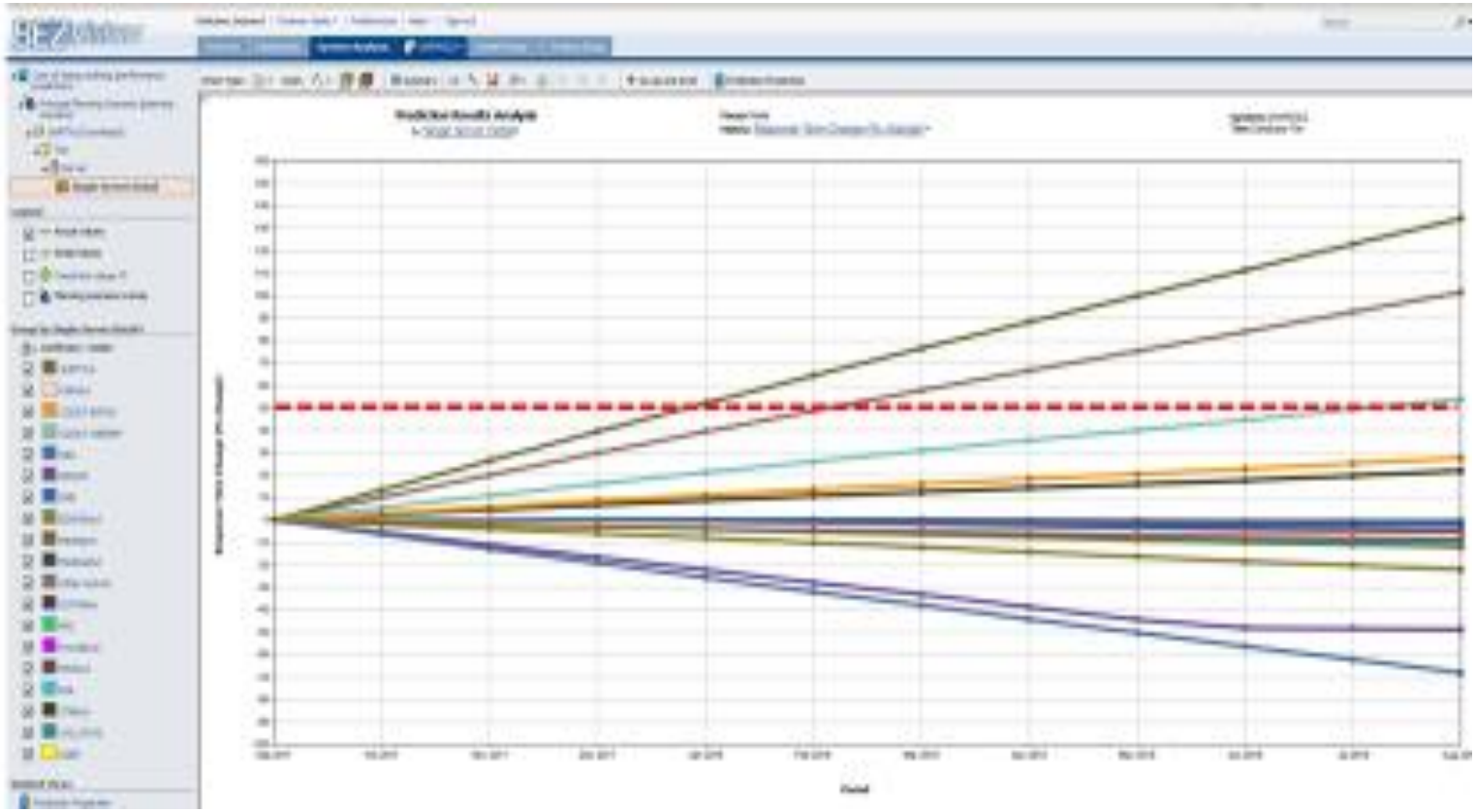
Showing 1 to 10 of 24 entries

Previous 1 2 3 Next



Prediction when SLGs will not be met

What should be changed to meet SLGs?

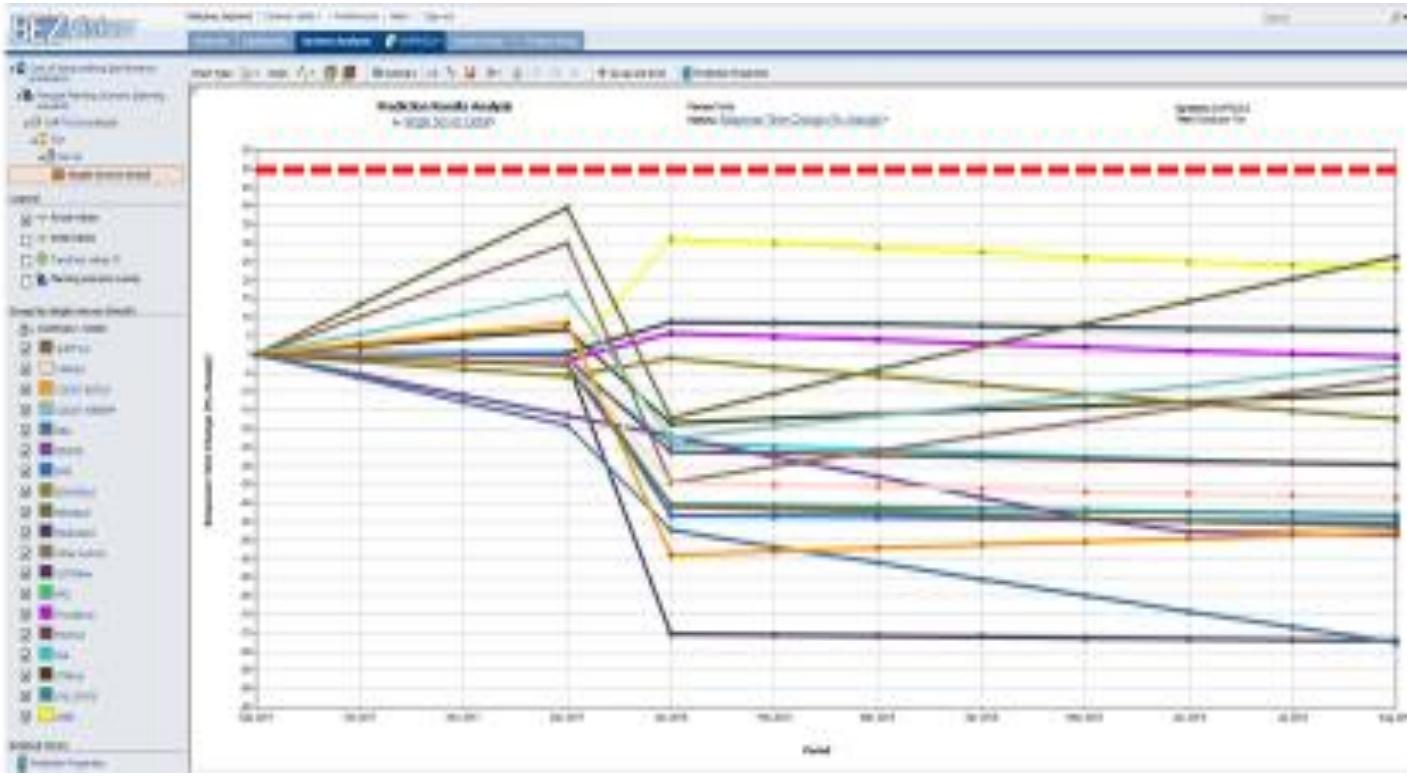


Option 1: Reduce Priority for “Yellow” workload

It will allow to meet SLG for “Blue”, but “Green” and “Brown” workloads will not meet SLGs

Option 2: Reduce Priority for “Yellow” workload and increase for “Green” and “Brown”

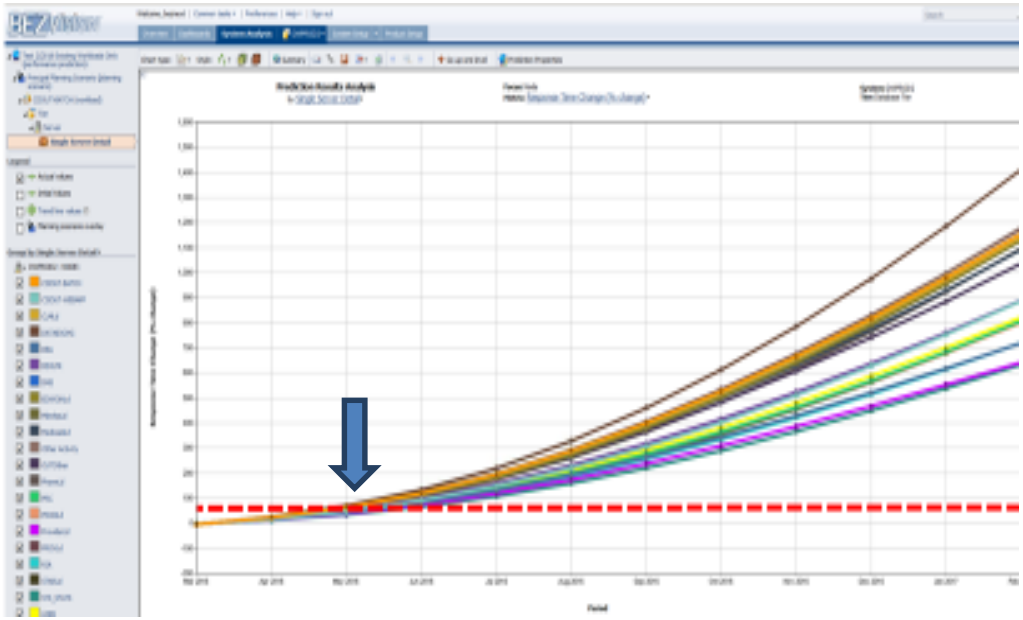
It will be sufficient to meet SLG for all workloads



3.3. LONG TERM CAPACITY PLANNING

According to Model SLGs will not be met in 2 months

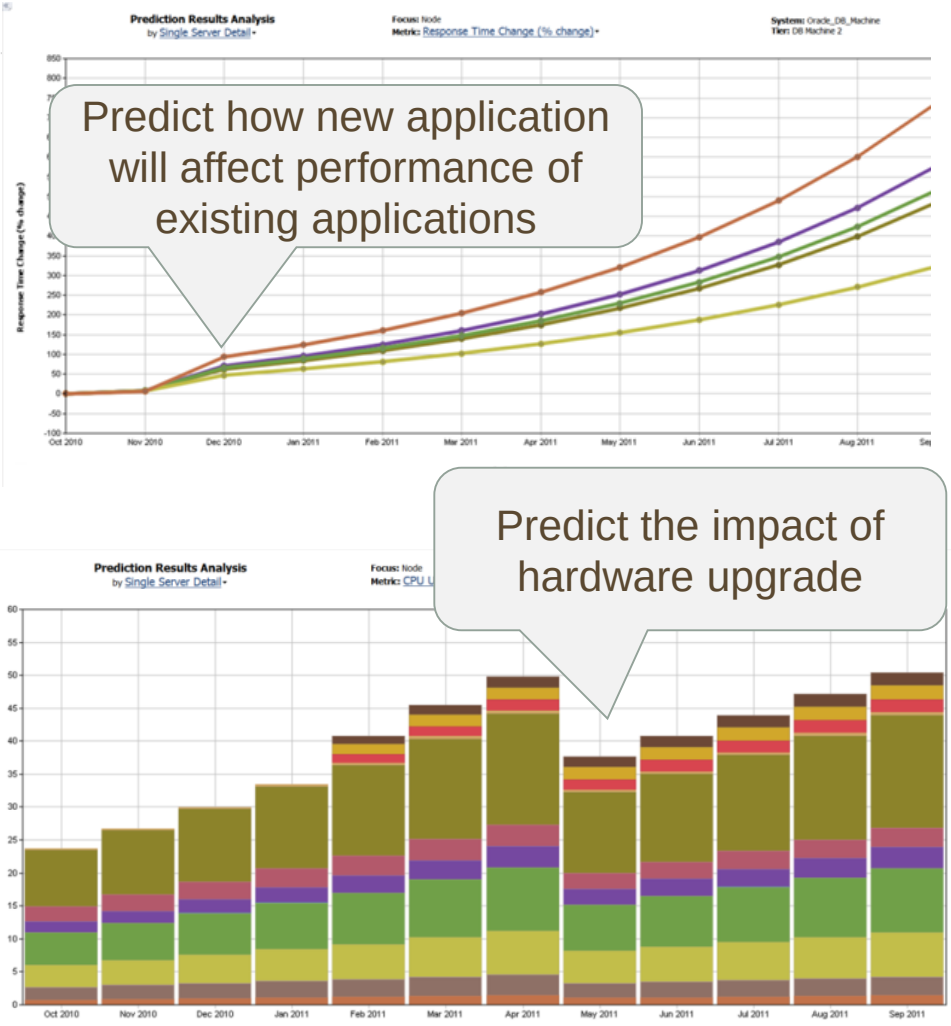
How much additional capacity will be required to meet SLGs for all workloads



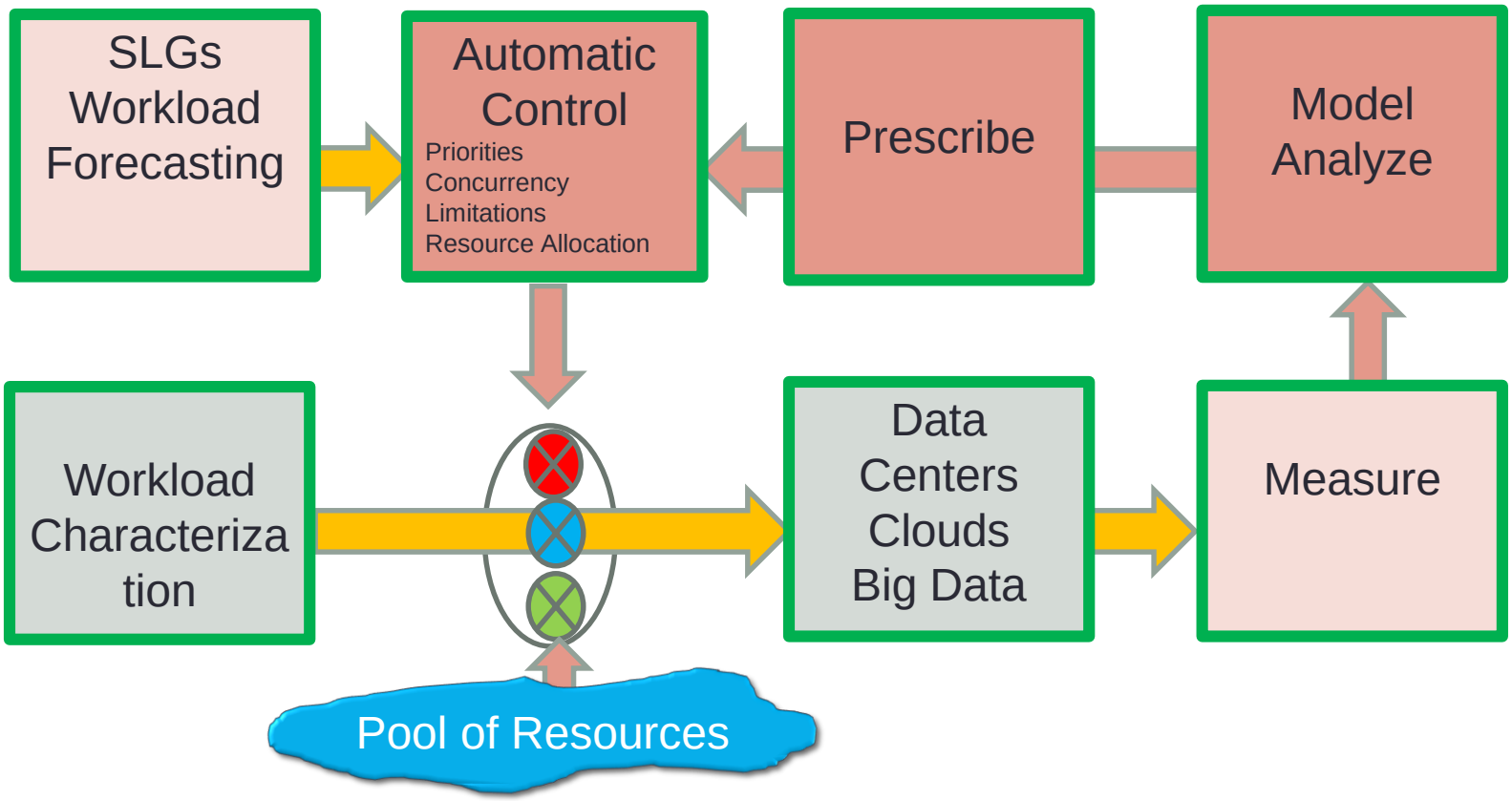
Predicting Impact of New Application Implementation and Development Recommendations

Predicting Analytics

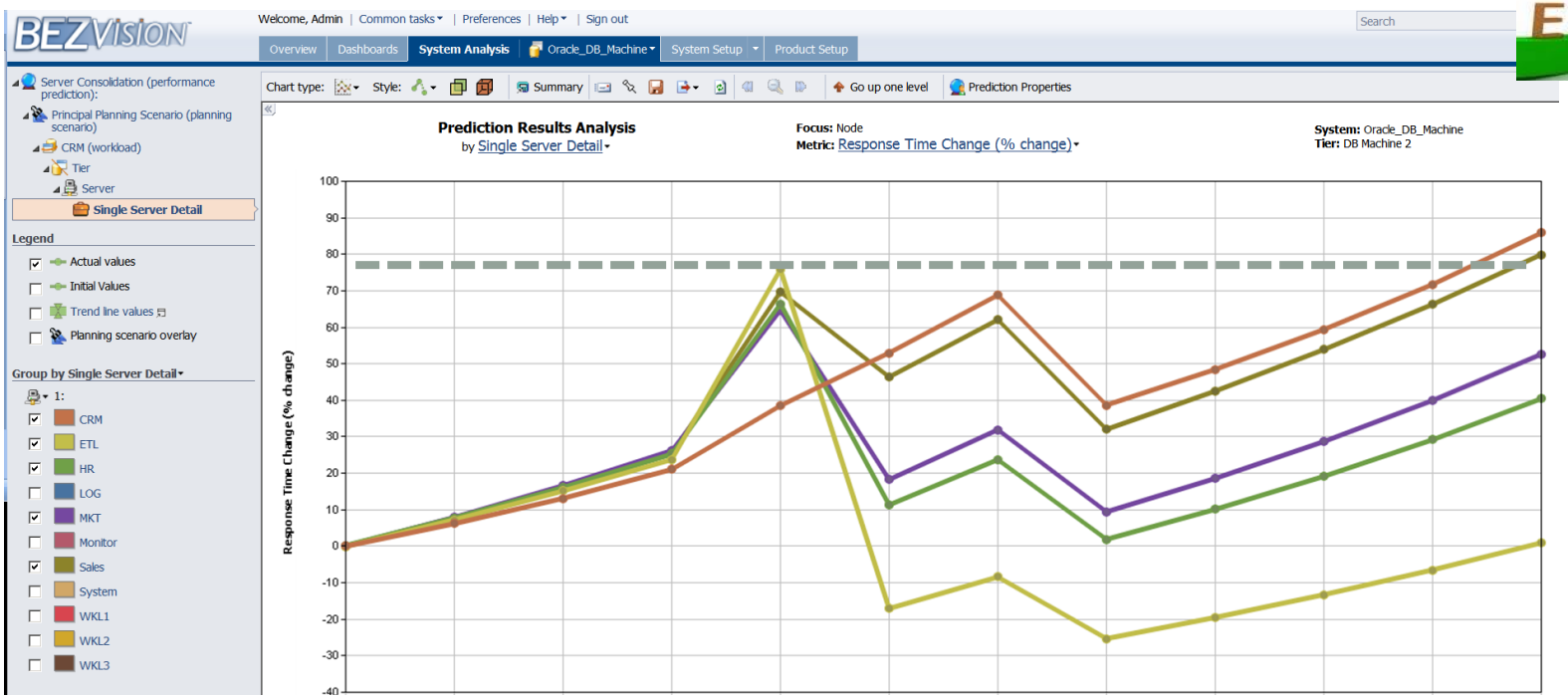
- Long Term
  - Queueing Network Models
- Short Term
  - Machine learning algorithms



Automation

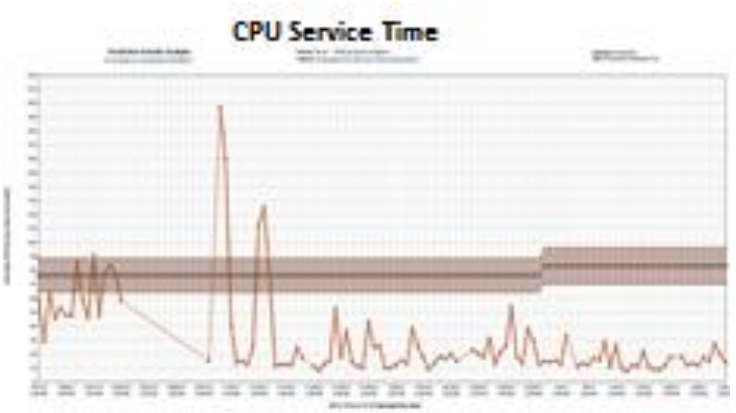


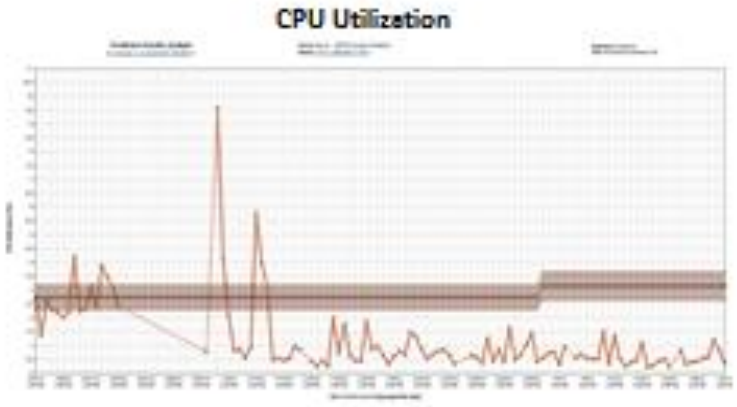
Plan of Actions and Predicted Expectations

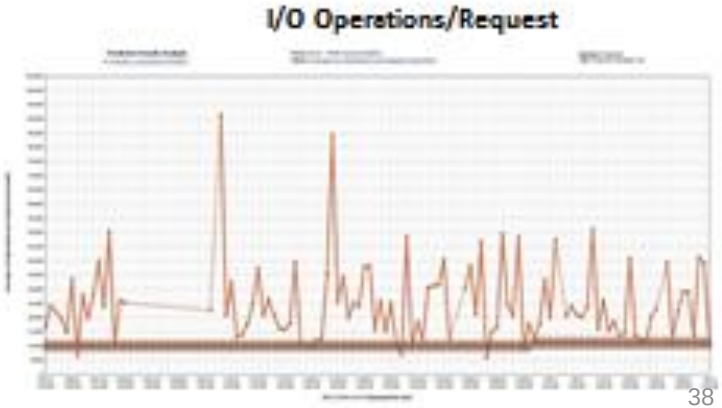


Automatic Verification and Feedback Control

Comparing Actual Results vs Expected (A2E)







4. SUMMARY

Summary

- Big Data industry is going through rapid changes during last 4 years and current challenges are on implementing AI, Data Streaming, Cloud Computing, Operationalization and New Technologies
- We reviewed examples of applying advanced analytics for Enterprise IT Performance Assurance, Performance Engineering, Dynamic Performance Management and Capacity Planning for IT
- You can join Collaborative project on ML Algorithms and ML Libraries selection

THANK YOU!

References

- Daniel A. Menasce “Software, Performance, or Engineering?” WOSP '02 Proceedings of the 3rd international workshop on Software and performance Pages 239-242
- Max Kuhn, Kjell Johnson, Applied Predictive Modeling, Springer, 2013
- B. Zibitsker, IEEE Conference, Delft, Netherlands, March 2016, Big Data Performance Assurance
- B. Zibitsker, Key note presentation “Role of Big Data Predictive Analytics” Big Data Predictive Analytics Conference, Minsk 2015
- B. Zibitsker, Key Note Presentation on “Big Data Advanced Analytics”, Big Data Advanced Analytics Conference, Minsk 2016,
- Dominique A. Heger “Big Data Predictive Analytics, Applications, Algorithms and Cluster Systems” ISBN:978-1-61422-951-3
- Dr. Joseph Hellerstein, *Microsoft Corp,* Yixin Diao, *IBM* Engineering Performance Using Control Theory
- B. Zibitsker, Proactive Performance Management for Data Warehouses with Mixed Workload, Teradata Partners, 2008, 2009
- J. Buzen, B. Zibitsker, CMG 2006, “Challenges of Performance Prediction in Multi-tier Parallel Processing Environments”
- B. Zibitsker, CMG 2008, 2009 “Hands on Workshops on Performance Prediction for Virtualized Multi-tier Distributed Environments”
- Heger, D., “Introduction to Apache YARN Schedulers & Queues”, Data Analytica, www.mlanalytica.com, 2016
- Heger, D. "Machine Learning in the Realm of Big Data Analytics", Fundcraft Publication, ISBN 978-0-578-19095-2, March 2017.