February 04, 2019

OPTIMIZATION OF CLOUD PLATFORMS

Johnson Manuel-Devadoss

David Stanford

Billie Ansell

Sunil John

Luis Bravo

See next page for additional authors

Inventor(s)
Johnson Manuel-Devadoss, David Stanford, Billie Ansell, Sunil John, Luis Bravo, Miguel Figueroa, and Larone Gardner

OPTIMIZATION OF CLOUD PLATFORMS

AUTHORS:
Johnson Manuel-Devadoss
David Stanford
Billie Ansell
Sunil John
Luis Bravo
Miguel Figueroa
Larone Gardner

ABSTRACT

Embodiments presented herein provide a mechanism for optimizing the cost of a cloud platform. A cognitive cloud platform is trained using performance counters to process unstructured data in order to identify underutilized cloud resources. The platform may also generate policies to shut down during idle times. Present embodiments may utilize techniques such as robotic process automation (RPA), straight through processing (STP), and/or natural language processing (NLP).

DETAILED DESCRIPTION

Enterprises that adopt a cloud platform seek to optimize costs by identifying inefficient uses of cloud resources. In order to identify the majority of wasted cloud resources, administrators may shut down idle or underutilized cloud instances, run instances only when needed, clean up unused storage, and operate in less-expensive regions. However, there is a need for a cognitive cloud management utility that can identify unused or underused cloud resources, evaluate the most cost-effective regions and instance sizes, identify optimal discounting options based on cloud resource usage levels, and generate policies to restrict cost usage based on budget controls without changes to the core systems.

The embodiments presented herein employ solutions such as robotic process automation (RPA), machine learning, cognitive computing models, and straight-through processing (STP) to reduce business process costs. In one embodiment, the present invention analyzes cognitive cloud activity logs to identify underused cloud resources, and

Published by Technical Disclosure Commons, 2019

initiates robotic process automation to shut down idle time and restart the resources when necessary. In some embodiments, a cognitive model is trained using unstructured log collection data from a cloud platform to perform data mining on cloud resource performance and usage against performance counters for the corresponding cloud resources.

Present embodiments feed performance counter measures to the cognitive model in order to train the model to determine the performance against unstructured data. For example, if a virtual machine or cloud instance is hosted with a hosting application, then present embodiments may feeds the measurements of performance counters, such as server counter measures like system uptime, percent processor time, percent processor user time, percent processor privileged time, processor interrupts per second, percent processor DPC time, percent processor interrupt time, processor DPCs queued/sec, processor DPC rate, percent processor idle time, percent processor C1 time, percent processor C2 time, percent processor C3 Time, processor C1 transitions/sec, processor C2 transitions/sec, processor C3 transitions/sec, and the like, as well as measure from application counter, such as request wait time, application running, requests queued, requests total, requests timed out, requests succeeded, etc.

Collected metrics are fed into a cognitive model to determine the cloud instance utilization. This monitoring process may take place over a period of time in order to measure the baseline metrics of a system and rank them based on the weight of underutilized metrics to identify percentages of utilization. Once the unused or underutilized resources are identified, a robotic process automation process may be initiated to shut down the identified resources during idle times. In other embodiments, machine learning techniques, such as natural language processing, may be leveraged to read and understand unstructured messages that are being presented in the cloud logs in non-English in order to make human-like judgmental decisions, based on training and historical data using cognitive techniques. As used herein, training may refer to providing performance counter metrics to a cognitive model to understand and process the unstructured log messages for a given cloud resource.

Present embodiments may provide a mechanism for automating end-to-end processing of utilization logs or activity logs of an underutilized cloud instances. A

cognitive model may process or control all elements of a work flow designed for cost optimization, including processing the activity logs of cloud resources, identifying the underused cloud resources by processing logs, and generating cloud policies to restrict the cost usage based on the budget controls in place. Thus, the overall operation can be defined as electronically capturing and processing activity logs, determining the cloud resource needs to be put under control by applying policies in one pass. A single pass may occur from the point of first monitoring cloud resources to the point at which policies are applied to restrict usage based on the budget controls without changes to the core systems. For example, present embodiments may generate a policy that allows only free-tier stock keeping units onto cloud resources in order to meet budgetary expectations. Similarly, a policy may be generated in which all cloud resources are to be deployed to the less-expensive regions.

Figure 1 depicts an example of a cloud system in accordance with present embodiments.

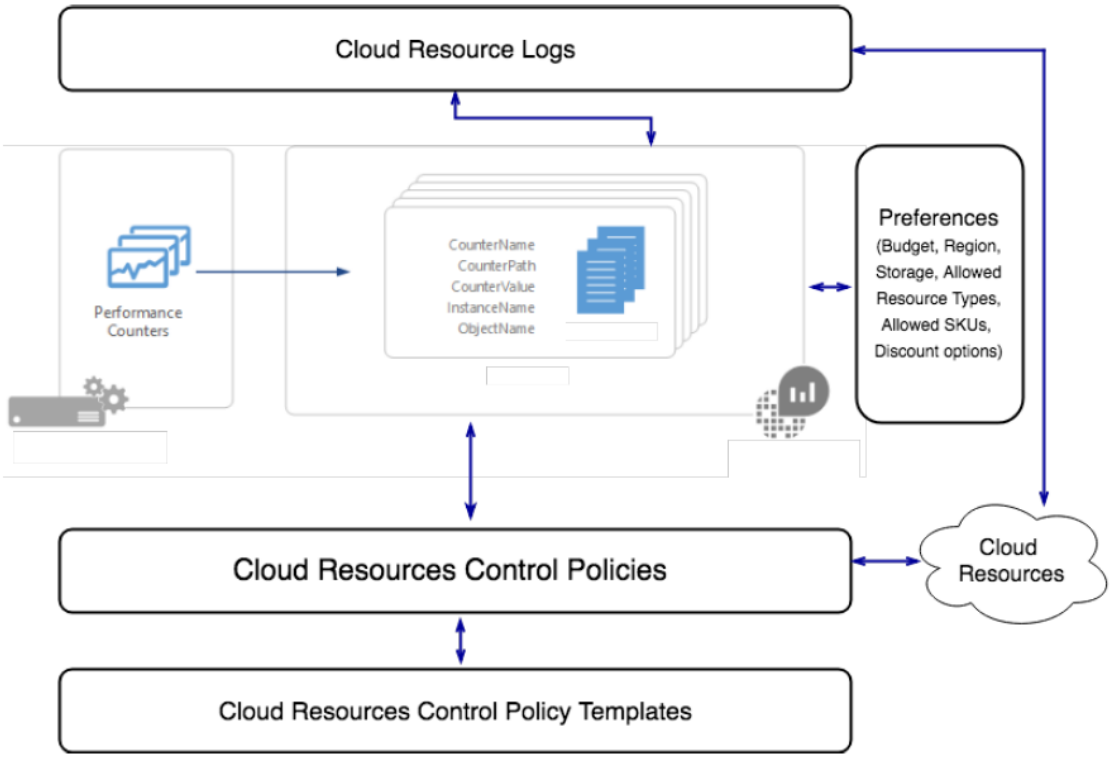

Figure 1

As shown in Figure 1, cloud users may provide preferences, such as monthly budget preferences, region preferences, allowed storage sizes, allowed resource types, allowed

stock keeping units, and discount options. Present embodiments leverage RPA as the first step for generating cloud resource control policies based the user preferences, producing cloud resources control policy templates to apply the cloud resources control policies at a cloud tenant level. Next, the cognitive models are trained using performance counters and unstructured cloud resources logs data to process and determine the utilization percentage on each cloud resource, ranking them according to utilization. Machine learning techniques, such as natural language processing, may be employed to read and process the unstructured and non-English language content to determine the uptime of cloud resources that are compared against provided performance counter metrics. Based on the utilization rankings, RPA may be initiated to generate cloud resource control policies to shut down each top ranking unused/underutilized resources during non-peak hours.