Visualization and Analysis of Large Cloud Platform Workload using Map-Reduce

Tulio Alberton Ribeiro

Abstract

The project consists of analyze the behavior of the cluster regarding on the hours and days of week usage from Azure dataset and plot it. The data set consists of 30 consecutive days and has 117GB.

1 Introduction

Basically I will run analytics through Hadoop and produce visualizations using R-Graphs after Map-Reduce executions. The dataset provided by Azure it is described at paper [2] and could be downloaded at [1].

2 Trace characteristics

The trace contains a representative subset of the first-party Azure VM workload in one geographical region [1]. The main trace characteristics and schema are:

- Dataset characteristics:
  Dataset size: 117GB composed of 128 files, representing 30 consecutive days; Total number of VMs: 2,013,767 totalizing 5,958 Azure subscriptions; Time series data: 5-minute VM CPU utilization readings, VM information table and subscription table (with main fields encrypted); Total VM hours: 104,371,713; Total number of VM CPU utilization readings: 1,246,539,221; Total virtual core hours: 237,815,104.
- Schema characteristics:
  Encrypted subscription id and deployment id; Timestamp in seconds (starting from 0) when first VM created; Count VMs created; Deployment size; Encrypted VM id; Timestamp VM created; Timestamp VM deleted; Max CPU utilization; Avg CPU utilization; P95 of Max CPU utilization; VM category; VM virtual core count; VM memory (GBs); Timestamp in seconds (every 5 minutes); Min, Max and Avg CPU utilization during the 5 minutes;

3 Panning

Description	week 1	week 2	week 3	week 4
Data acquisition and treatment	•			
Cluster installation and configuration	•	•		
Example tests	•	•		
Run real experiments		•	•	
Write short report			•	•
Final presentation				•

4 Technologies

It will be used a Hadoop cluster composed of three machines and some specific Map-Reduce functions to treat the data. After the data treatment the R-graph tool will be used to plot the results.

5 Topic chosen

Big Data Systems and data analytics.

References

[1] Azure public dataset. `https://github.com/Azure/AzurePublicDataset/`. Accessed: 2017-12-21.

[2] Cortez, E., Bonde, A., Muzio, A., Russinovich, M., Fontoura, M., and Bianchini, R. Resource central: Understanding and predicting workloads for improved resource management in large cloud platforms. In *Proceedings of the 26th Symposium on Operating Systems Principles* (2017), ACM, pp. 153–167.