PIE (Predictive Insights Engine)

With the advent of cloud computing, IOT and mobile, big data is a thing of the present. Petabytes of data flowing through the pipelines and big surge in the demands for services and applications that have 99.9999% uptime, DevOps and IT teams usually have a response time of few seconds to minutes to resolve any issues. Service impact directly affects revenue in most cases. Besides this performance is equally important when it comes to user experience, as an unresponsive application will adversely affect user retention, which affects revenue.

With the solution I am proposing operations and infrastructure teams can take a proactive approach in addressing the problems, which might occur in the future. This would help them find problems faster and be prepared for actionable items even before they have occurred. This solution would also aid in efficient capacity planning, scheduling cluster maintenance, budget planning for necessary resources, etc.

Approach:

One of the main issues with this project was getting data for an operational datacenter (DC). For this project I had to generate my own data to be able to run analysis on it. I am currently running this experiment on a 50 node Mesos cluster (http://mesos.apache.org/) that has node-monitoring agents running. I have divided the project into 4 major milestones (Green = done, yellow= in progress, red = not started):

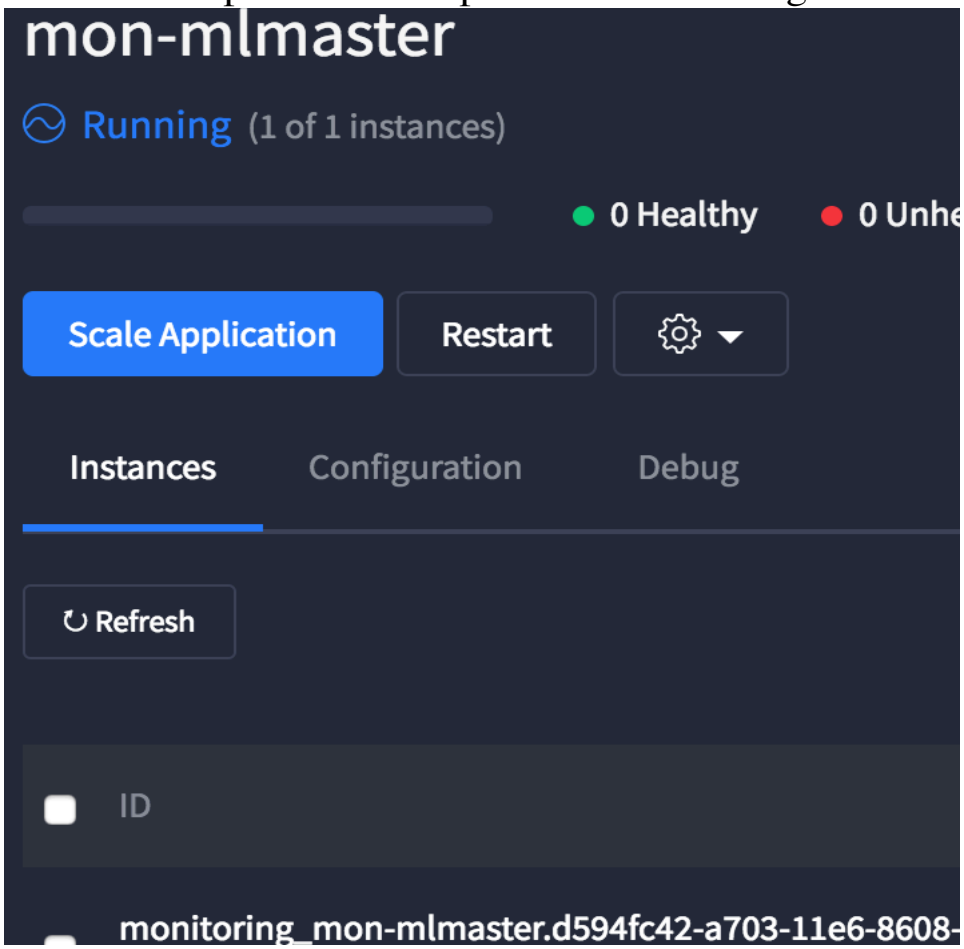
1. Node monitoring agents(data collection): I have implemented functionality to send disk used percentage data from monitoring agents running on all nodes of a cluster. This agent collects data points for disk used percentage every given polling interval

(currently set to 2 minutes), which is being written to a document based data store, ElasticSearch.

2. Deployment of the Apache Spark Cluster: When I started looking into this, portability and ease of setup were two of my major concerns. To address those I started with Vagrant to bring up a local Mesos cluster on your laptop. Code for this is checked into git.

However, for the time being I have put the Vagrant approach aside and brought up the spark cluster in the same DC I am running the agents for data collection. Deployment configuration files that were used to launch docker containers are checked into git under deployment folder.

Here is a snapshot of the Spark Master running on the cluster:



3. Application of Machine Learning (ML) techniques to do predictive analytics

Currently I am in this stage where I am exploring the data and ML techniques that could be used. I will be using Java for implementing my analytics engine.

4. Trend Visualization:

Once I have run the analysis, data will then be sent to a database and graphed. For this stage I will use a relational DB and Grafana (http://grafana.org/) or Kibana for visualization.