**VAST 2012 Mini-Challenge 1 Solution Description for Reviewers and Committee**

**Updated 7/2/2012**

## Introduction

The 2012 VAST Challenge consists of two Mini-Challenges that concern a fictitious world called BankWorld and a fictitious international bank called Bank of Money (BOM). Mini-Challenge 1 (MC1) focuses on issues involving large-scale situation awareness for the BOM enterprise. Contestants are provided with background information that provides the setting and context for the problem and solution, two questions that they must answer in their submission concerning the scenario, four datasets, and a standard answer form. The Challenge descriptions provided to contestants can be found at the VA Community Wiki Site, under “[VAST Challenge 2012: Challenge Descriptions](http://www.vacommunity.org/tiki-index.php?page=VAST%20Challenge%202012%3A%20Challenge%20Descriptions)”. Before reviewing solution packets, reviewers may want to familiarize themselves with the BankWorld and Bank of Money background material. Specific information about the geography of BankWorld and the physical environment for the Bank of Money enterprise can be found in the [information packet](http://www.vacommunity.org/tiki-index.php?page=VAST%20Challenge%202012%3A%20Submission%20Instructions%20and%20Downloads) provided with the Mini-Challenge 1 materials. The geography is interesting with respect to the setup of a synthetic world and banking enterprise and relevant to answers that contestants will be providing.

The primary goal of the contestants is stated in the Challenge Descriptions. To wit:

The Bank of Money (BOM) Corporate Information Officer (CIO) has assigned you to create a situation awareness visualization of the entire enterprise. This is a considerable challenge, considering that BOM operates from BankWorld's coast to coast. In addition to observing the global situation, he would also would like to be able to detect operational changes outside of the norm.

So, the main goal for the 2012 Mini-Challenges is very specifically large scale cyber situation awareness.

Summary of Mini-Challenge Questions and Answers

The two questions to be answered for MC1 are listed below, along with summary answers. Further explanation of the scenario and answers can be found in the following sections.

**Question MC 1.1:** Create a visualization of the health and policy status of the entire Bank of Money enterprise as of 2 pm BMT (BankWorld Mean Time) on February 2. What areas of concern do you observe? (Short Answer)

**Anticipated answer to MC 1.1:** The contestants may answer this question using only a snapshot of the data at the requested time or using the entire larger dataset containing trend information. The submission is expected to present a snapshot of the entire state of the Bank of Money enterprise at the selected day and time. The contestants may note that there are several Region 25 machines offline at this time. This effect would be related to a hurricane, although the participants would not be aware of this unless they considered the large dataset of trending data. The contestants may also note that there are a limited number of machines reporting from Data Center 5. There are no issues related to unexpected teller machine activities in this dataset. If other areas of concern are noted, there must be significant justification for them to be called out.

**Question** MC 1.2 Use your visualization tools to look at how the network’s status changes over time. Highlight up to five potential anomalies in the network and provide a visualization of each. When did each anomaly begin and end? What might be an explanation of each anomaly? (Detailed Answer)

**Anticipated Answer to MC 1.2**: The primary events inserted in the data are summarized as follows. The details of the scenario are explained in the following sections.

1. A series of machine outages corresponding to the path of a hurricane.
2. A data center comes on line.
3. Abnormal behavior teller machine traffic over time.
4. Decrease in general network health over time.

Other identified anomalies must be explained with well-reasoned arguments supported by data.

## Scenario Components

There is one general trend and three specific events whose effects are embedded in the data.



Figure Bounding boxes moving across Region 25

*Overall trend*

The overall status of the BOM network for the 2 day sample shows machines becoming less healthy in general with connection traffic rising during business hours. The health trend is reflected in the field **policystatus** increasing over time.

*Specific Event: a hurricane*

A hurricane rolls up the eastern sea coast of BankWorld causing a rolling blackout. BOM computers in Region 25 go offline starting at 10am and continuing throughout that evening.

Large numbers of computers along the southeast coast of BankWorld change from normal status to "off". Machines that are determined to be inside sets of bounding geo-coordinates are affected (Figure 1).

When a bounding box sits over an area that contains BOM offices, all machines in those offices stop functioning. Outages can range from 1 to 6 hours.

Note that the hurricane encounters a small regional office as it progresses across the coast.

A simple graph showing the impact of the hurricane on the reporting of machines in that region is shown in Figure 2.

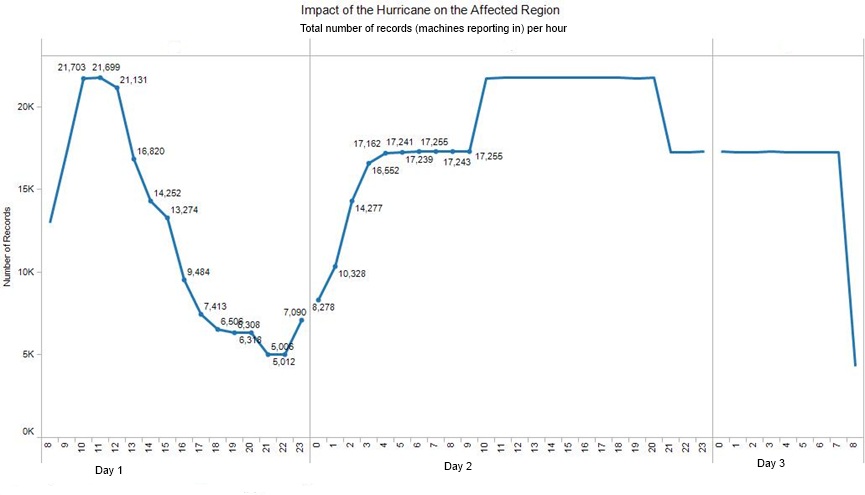


Figure Impact of the Hurricane on the affect region's computers

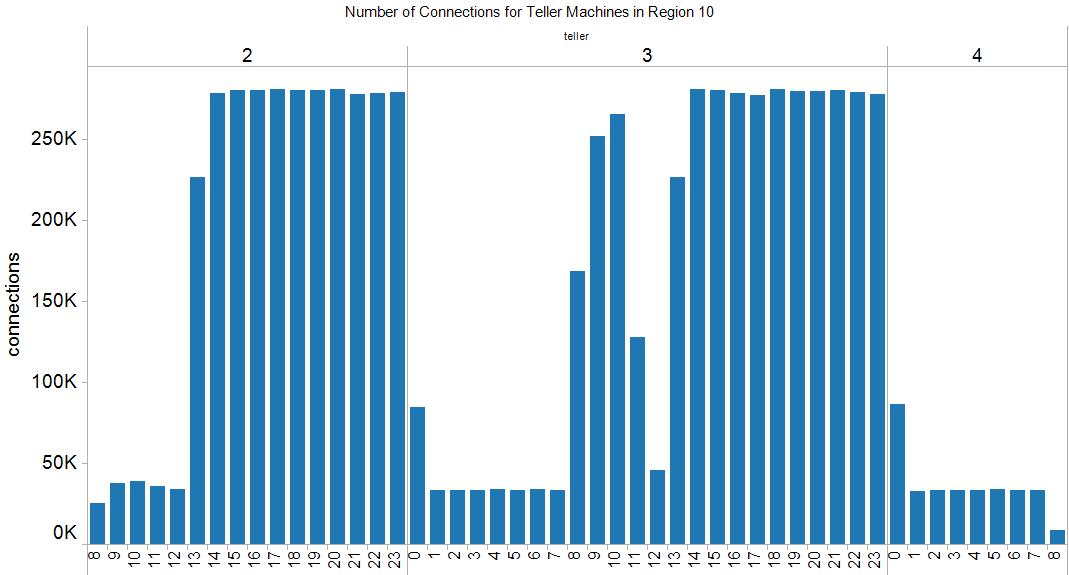
The major impact of the storm occurs at approximately 2200 hours during the first day.

Although this situation was based on a slow moving large storm, the contestants will not have any information pointing them explicitly to a hurricane. Any large spatially moving phenomenon could be a plausible explanation.

*Specific Event: Suspicious Use of Teller Workstations Off-Hours*

Teller machines are being used off-hours. The behavior starts at 2am in a single branch in Region 10, moves to 3 more branches in an hour, then 3 more the following hour. The next night all available machines in the region are involved. The behavior is expressed as a substantially higher number of connections on a larger percentage of the teller machines. If a machine is “off” at night, it is really off and would not be participating.

Figure 3 and 4 provide an indication of how these rogue teller machines are becoming active over time. Figure 3 displays the abnormal behavior in the region, while Figure 4 displays normal behavior over the same time period in a different region. The range of “number of connections” shifts from 1-10 up to 40-100 per machine in the affected region.

 Figure Indications of the increased use of teller machines during off hours in Region 10. The bars indicate number of connections in this affected region.

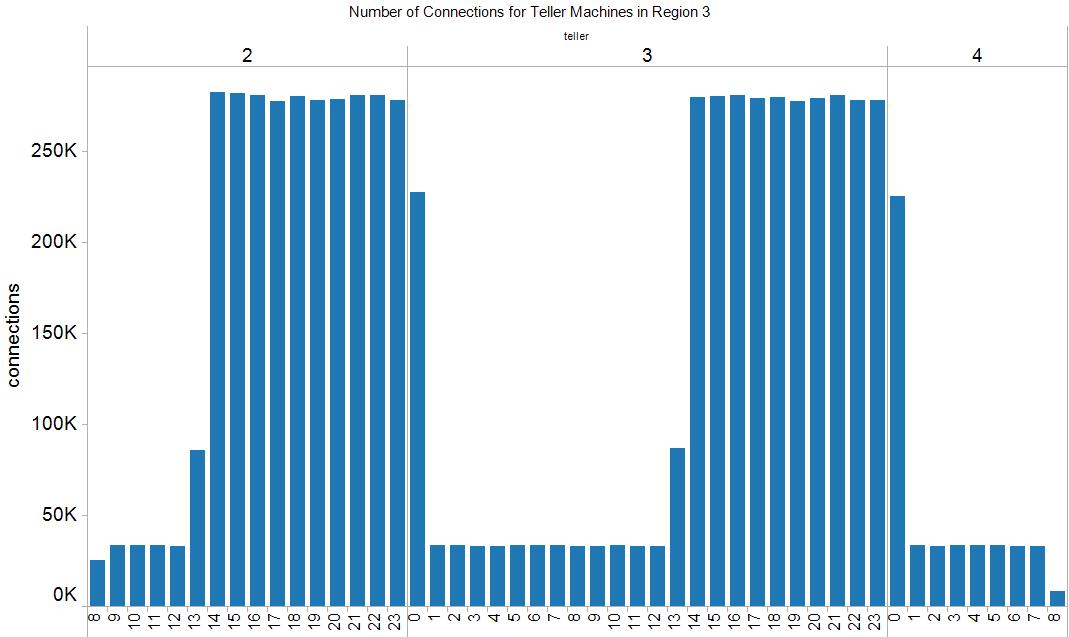


Figure Normal amount of communication for teller machines in a large region (this is taken from unaffected Region 3).

Specific Event: Server Farm Comes On Line

The “datacenter-5” server farm comes on line on February 2nd. Figure 5 shows the reporting behavior of machines from the data center coming on line.

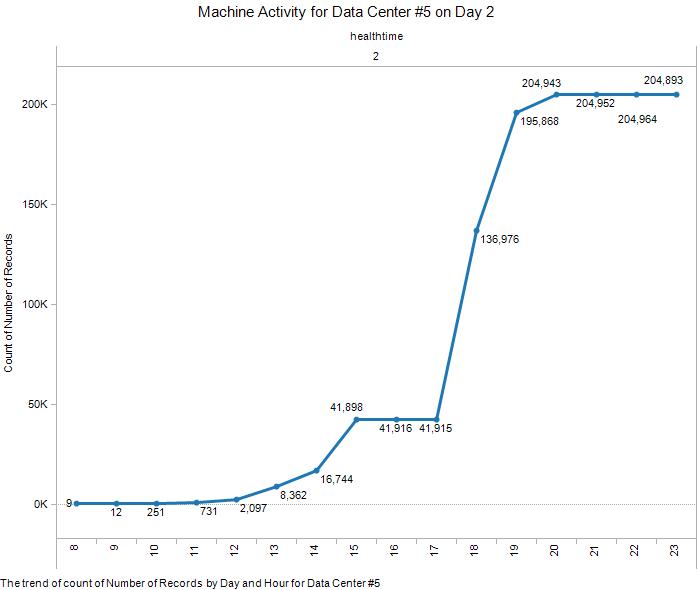


Figure Data Center #5 coming on line on Day 2

Machines tend to come on line and some fall off line as the Data Center is attempting to achieve full capacity.

## Details about the Challenge Data

There are four datasets provided to contestants for Mini-Challenge 1.

1. “VASTgeneralinformation.zip”, contains several files with background information on the Mini-Challenge.
2. “mc1postgresExport.out” in its Postgres form or “meta-3-7.csv” in CSV form, contains information describing the computing devices comprising the Bank of Money enterprise.
3. “metaStatus-3-7.csv” in CSV form, contains two days’ worth of status information from all of the computing devices in the enterprise.
4. “windowOneSingle.csv” provided only in CSV form, gives a single day, single hour snapshot of the Bank of Money status.

**Dataset Characteristics**

The following sections contain information on the characteristics of the datasets and the dataset generation, including some of the parameters and constraints that were used.

The Bank of Money has:

* 1 headquarters location
* 5 data centers, which are overseen by the headquarters
* 10 Large regions, which each consist of a headquarters facility and 200 branches
* 40 Small regions, which each consist of a headquarters facility and 50 branches

The breakdown of the number of computers within each of these components is shown in **Error! Reference source not found.**:



Table 1 Numbers of Machines in Bank of Money

Approximately 895,000 machines produced data for this challenge.

**Business/Base Data Generation Rules:**

1. Business hours are Monday-Friday, 7am – 6 pm in each time zone.
2. 40% of the workstations are turned off at night.
3. Planned Maintenance. For machines that are up and running, each machine class has a percent chance to shut down and miss 1-4 reports. Servers have a 15% chance, workstations have a 30% chance, and ATMs have a 5% chance.

**Connection Ranges**. Number of connections made by a machine will vary according to a random distribution over certain ranges.

Business hours

* Servers: 5-100 connections
* Workstations: 5-50 connections
* ATMs: 1-10 connections

Non Business Hours

* Servers: 1-50 connections
* Workstations: 1-10 connections
* ATMs: 0-4 connections

**Dataset Content**

The dataset, **meta-3-7.csv**, contains metadata on each of the machines in the challenge. The format of the data is as follows:

* Ipaddr: The IP address assigned to a machine.
* Machineclass: Either “workstation”, “server”, or “ATM”.
* Machinefunction: For workstations, either “teller”, “loan”, or “office”. For server, either “web”, “email”, “file server”, “compute”, or “multiple”. For ATM, just “atm”.
* Businessunit: Either “headquarters” or one of “region-1” to “region-50”
* Facility: For a “headquarters” businessunit, the value is either “headquarters” or one of “datacenter-1” to “datacenter-5”. For a region, the value is either “headquarters” or one of “branch1” to “branch200”. Regions 1-10 are large regions and have 200 branches each. Regions 11-50 are small regions and have 50 branches each.
* Latitude: in decimal form.
* Longitude: in decimal form.

The dataset, **metaStatus-3-7.csv**, contains status information for each machine reporting during a reporting period. The reporting periods are 15 minutes. The data reported includes number of connections (both incoming and outgoing combined), the policy status (or how well the machine is compliant with security standards), and activity status (or what activity is the machine performing in the period). The format of the data is as follows:

* Tkey: Not used for this mini-challenge.
* Ipaddr: The IP address of the machine reporting its status.
* Healthtime: Date and time of the report.
* Numconnections: Number of connections to and from the machine.
* Policystatus: How well the machine complies with security policy. The value is an integer from 1 to 5, with indications as follows:

1 – Healthy / normal

2 – Moderate policy deviation

3 – Serious policy deviation / non-critical patches failing

4 – Critical; patches failing

5 – Possible infection / questionable files found

When a machine indicates policy status code, it will never return to a lower value status, but it may move to a higher value status.

* Activityflag: An integer code for the current machine activity. The value is an integer from 1 to 5, with indication as follows:

1 – Normal

2 – Going down for maintenance

3 – More than 5 invalid login attempts

4 – CPU fully consumed

5 -- Device has been added / like an external device or thumb drive