DSB Production Root Cause Analysis

For Production Outage on 24th Jan 2018

Prepared by: Technical.Support@anna-dsb.com
Date: 6th February 2018

Table of Contents

Revision History

Version	Date	Reason
1.0	6th February 2018	Document reviewed and finalized

IMPACT ASSESSMENT & CATEGORIZATION

Critical/System Down (Severity One – S1)

First Occurrence
Start: 15:36
Resolved: 15:40
Second Occurrence
Start: 15:51
Resolved: 15:55
Third Occurrence
Start: 16:46
Resolved: 16:50
Total: 12 minutes
Production application down or major malfunction resulting in a product inoperative condition. Users unable to reasonably perform their normal functions.
The specific functionality is mission critical to the business and the situation is considered an emergency.

Condition 1- When a critical system, network component or key application is under outage (or imminent outage) with critical impact to all clients.
Condition 2 - Total loss of service to entire user base which includes total unavailability of critical applications for entire end users and in all locations.

Significant Impact (Severity Two - S2)
Start: 14:53
Resolved: 17:23
Total: 2 hours 26 minutes
Critical loss of application functionality or performance resulting in high number of users unable to perform their normal functions. Major feature/product failure; inconvenient workaround or no workaround exists. The program is usable but limited.

Condition 1: A key component of the solution, an application across all users, a set of users or intermittent network degradation or instability leading to performance and degradation of service.
Condition 2: An incident which is not yet S1, but might lead to a potential S1 incident.
Condition 3: Partial users at a particular location are affected but not all the users in all locations

INTRODUCTION

The purpose of this Root Cause Analysis (RCA) is to determine the cause that contributed to the recent loss of service and "Something went wrong" error message and response code "HTTP 500" encountered by clients in the DSB production environment on 24th January 2018 between the hours of 14:53 UTC and 17:23 UTC. This RCA determines what happened during the event, how it happened, and why it happened. To accomplish this, an investigation took place internally between the DSB support, Development teams and senior analysts to ascertain the primary root cause or a list of root causes that contributed to this issue.

EXECUTIVE SUMMARY - FINDINGS AND ROOT CAUSE

Wednesday 24th January 2018

This was due to performance issues on SOLR (http://lucene.apache.org/solr/) resulting in a lack of responsiveness as the DSB observed an increase in CPU load and memory consumption on the SOLR servers. After examination of log files and alerts received both pre- and post-event, the DSB technology team increased the available resources to the service and modified the application configuration to improve responsiveness. This was to deal with the large number of authentications taking place as a result of the requests being placed against the platform. During the period of the event no duplicate ISIN's were created.

A further capacity increase across all servers was planned for the weekend of the 27th to lower further the risk of these issues re-occurring. Changes are also planned to implement authentication caching to lower the resource requirements of the SOLR servers. These changes will be rolled out to production at the end of Q1 2018 following successful UAT testing.

CORRECTIVE ACTIONS TAKEN & PLANNED

- 24th Jan SOLR resource increases and GC configuration change
- 27th Jan Upgrade the capacity of all SOLR servers
- 10th Feb UAT Testing for Authentication caching
- Q1 2018 - Target production date for Authentication caching

EVENT DESCRIPTION

On 24th January 2018 at 14:53 pm UTC, the production environment experienced an issue with Solr services, causing some established FIX and ReST API connections to drop without successfully reconnecting. This took place between the hours of 14:53 and 17:23. Clients also had difficulty to login via the web GUI and therefore the web portal was put into maintenance mode at 15:21.

"Something went wrong" error messages were experienced by clients when searching or creating ISIN's due to the Cordra (https://cordra.org/) service timing out on their Solr services connections. This was due to the Solr service being unresponsive during this period because of extended Garbage Collection within the Java Virtual Machine.

Between the hours of 14:53 and 17:23, all users experienced reconnects and disconnections via FIX. ISIN creation and search services were unavailable between the S1 start and finish times specified on page 2.

As a result of the issues experienced, the DSB technology team planned the implementation of a configuration change to the Solr resources available and an adjustment to the SOLR Garbage Collection (GC) configuration on the 24th January to stabilize and restore services.

After these configuration changes were implemented all FIX and ReST servers were restarted. After this restart all services then returned to a healthy state.

On the weekend of the 27th the instance resources were increased again and the SOLR Garbage Collection configuration tuned further to reduce the risk of large GC pause times as a result of a large number of authentications during periods of high load.

Finally, the development team has placed an enhancement request with the CORDRA development team to optimize the authentication processing by the implementation of an internal cache, in order to reduce significantly the load on the SOLR service. PROD Deployment is expected towards the end of Q1 2018, after successful testing in UAT.

CHRONOLOGY OF EVENTS / TIMELINE

Wednesday 24th January 2018

14:53 PM UTC – Wednesday 24th January 2018
Alerts were triggered on Solr and Cordra services. Technical support start investigations

14:57 PM UTC – Wednesday 24th January 2018
Health checks on Solr services confirm Solr services are unresponsive to health check due to long GCs seen on the Solr GC logs

14:59 PM UTC – Wednesday 24th January 2018
Clients report issues with some FIX and ReST connections being dropped and unable to reconnect

15:11 PM UTC – Wednesday 24th January 2018
As part of troubleshooting efforts, one of three Solr servers restarted but services did not stabilize

15:18 PM UTC – Wednesday 24th January 2018
Steps to shutdown FIX endpoints begins, remaining two Solr servers restarted in sequence

15:21 PM UTC – Wednesday 24th January 2018
Placed the WEB GUI and ReST in maintenance

15:28 PM UTC – Wednesday 24th January 2018
Notification email was sent to the clients

15:29 PM UTC – Wednesday 24th January 2018
All Solr servers back to reporting active status

15:36 PM UTC – Wednesday 24th January 2018
All endpoints confirmed closed

15:40 PM UTC – Wednesday 24th January 2018
Started to remove ReST and GUI from maintenance. All FIX endpoints started

15:42 PM UTC – Wednesday 24th January 2018
Monitoring alerts seen with Solr Garbage Collection logs and "500 Internal Error". Ongoing work to remove Rest and GUI from maintenance and FIX endpoints halted and works start to revert the enablement of endpoints.

15:42 PM UTC – Wednesday 24th January 2018
Deployment of an adjustment to the garbage collection configuration to Solr on to one of three backend servers and restarted.

15:47 PM UTC – Wednesday 24th January 2018
One of three Solr Server checkouts are green and in an active state.

15:51 PM UTC – Wednesday 24th January 2018
Confirmed all endpoints closed and proceeded in updating and restarting remaining two backend severs.

15:55 PM UTC – Wednesday 24th January 2018
Started to remove ReST and GUI from maintenance and system monitored for stability

16:24 PM UTC – Wednesday 24th January 2018
FIX servers started

16:37 PM UTC – Wednesday 24th January 2018
Solr health checks shows similar GC log alerts shortly after all endpoints were enabled.

16:39 PM UTC – Wednesday 24th January 2018
As part of troubleshooting steps, GUI and ReST were put into maintenance and FIX endpoints begin to be shutdown

16:40 PM UTC – Wednesday 24th January 2018
A rolling restart was then performed on all the Solr services and checks were green

16:46 PM UTC – Wednesday 24th January 2018
Shutdown of FIX, ReST and GUI endpoints confirmed

16:50 PM UTC – Wednesday 24th January 2018
Started to remove ReST and GUI from maintenance and system monitored for stability

17:06 PM UTC – Wednesday 24th January 2018
Client updated notification email sent

17:23 PM UTC – Wednesday 24th January 2018
Started all FIX servers

17:29 PM UTC – Wednesday 24th January 2018
Confirmed FIX servers are good, confirmed successful login via GUI and removed from maintenance

17:38 PM UTC – Wednesday 24th January 2018
Resolved Notification email sent