DSB Production
Root Cause Analysis

For Production Outage on 23rd Feb 2018

Prepared by:	**Technical.Support@anna-dsb.com**
Date:	**5th March 2018**

Table of Contents

Revision History

Version	Date	Reason
1.0	5th March 2018	Document reviewed and finalized

IMPACT ASSESSMENT & CATEGORIZATION

Critical/System Down (Severity One – S1)

Start: 01:50
Resolved: 03:29
Total: 1 hour 39 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: 01:41
Resolved: 03:34
Total: 1 hour 53 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, “Something went wrong” error message and response code “HTTP 500” encountered by clients in the DSB production environment on 23rd February 2018 between the hours of 01:41 UTC and 03:34 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

Friday 23rd February 2018

This was due to performance issues on SOLR (http://lucene.apache.org/solr/) 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 stopped all API end points and restarted all SOLR servers in sequence to stabilize the service as initial health checks were unresponsive.

A configuration change was conducted across all servers on the weekend of the 24th February to lower further the risk of these issues re-occurring. Changes are also planned to amend the configuration for authentication caching and will be rolled out to production during Q2 2018 following successful UAT testing.

CORRECTIVE ACTIONS TAKEN & PLANNED

- All SOLR servers restarted in sequence to recover service
- 24th February upgrade to the storage capacity of all SOLR servers
- Target production date for Authentication caching configuration change - in Q2 2018

EVENT DESCRIPTION

On 23rd February 2018 at 01:41 am UTC, the production environment experienced an issue with all SOLR services, across all servers causing some established FIX and ReST API connections to drop without successfully reconnecting. This took place between the hours of 01:41 and 03:26. Clients were also unable to login via the web GUI and therefore the web portal was put into maintenance mode at 02:50.

"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 (GC) within the Java Virtual Machine.

"HTTP 500" error messages were encountered by clients utilizing the Production GUI during this time due to the SOLR service being unresponsive.

Between the hours of 01:41 and 03:26, all users experienced 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 initiated a restart of all SOLR servers in sequence after all end points were shut down in order to stabilize and restore the service.

After the SOLR restart, all SOLR service health checks reported green and therefore all FIX and ReST endpoints were opened. The service then returned to a healthy state.

No duplicate ISIN's were created as a result of this event.

On the weekend of the 24th February the instance resources were increased and the SOLR Garbage Collection configuration tuned further to reduce the risk of large GC pause times during periods of high load

An enhancement is currently pending deployment in UAT to optimize the GC log configuration in order to reduce significantly the load on the SOLR service. Deployment in production is expected in Q2 2018, after successful testing in UAT.

CHRONOLOGY OF EVENTS / TIMELINE

Friday 24th February 2018

01:41 AM UTC – Friday 23rd February 2018
Alerts were triggered on SOLR and Cordra services.

01:42 AM UTC – Friday 23rd February 2018
Technical support start investigations

01:42 AM UTC – Friday 23rd February 2018
Health checks on SOLR services confirm all SOLR services are down across all servers

02:24 AM UTC – 23rd February 2018
Notification email sent to all users

02:29 AM UTC – 23rd February 2018
Start shutting down all FIX servers

02:35 AM UTC – 23rd February 2018
All FIX endpoints shutdown

02:50 AM UTC – 23rd February 2018
Outage landing page invoked for Production GUI

02:53 AM UTC – 23rd February 2018
ReST Production endpoints put into maintenance

03:11 – 03:18 AM UTC – 23rd February 2018
Start of SOLR services restarted - in sequence

03:18 – 03:18 AM UTC – 23rd February 2018
Finish restart of SOLR services and all health checks are green

03:26 AM UTC – 23rd February 2018
Begin starting all FIX servers

03:34 AM UTC – 23rd February 2018
Maintenance lifted on ReST endpoints

03:43 AM UTC – 23rd February 2018
All FIX servers started and health checks are green

03:45 AM UTC – 23rd February 2018
Updated notification distributed to all users

04:12 AM UTC – 23rd February 2018
All health checks are green and system stabilized, outage landing page removed on Production GUI

05:30 AM UTC – 23rd February 2018
Update notification sent to users informing that the DSB Production services is now available