2016

GCSS Analytics Proof of Concept

Kendall, Tony; Belli, Greg; Schwamm, Riqui; Cote, Scott

Monterey, California. Naval Postgraduate School

MONTEREY, CALIFORNIA

GCSS Analytics Proof of Concept

Report Type: Final Report
Period of Performance: 10/01/2015-02/28/2017

Project PI: Tony Kendall, Lecturer, Information Sciences Department, Naval Postgraduate School
Additional Author/Authors:
Greg Belli, Researcher, Information Sciences Department, Naval Postgraduate School
Riqui Schwamm, Researcher, Information Sciences Department, Naval Postgraduate School
Scott Cote, , Senior Lecturer, Information Sciences Department, Naval Postgraduate School

Prepared for: Topic Sponsor: N2/N6

Commented [HSCD1]: Sponsor was HQMC I&L

Research POC: Maj Nic Martinez (LX, DC I&L)

Research POC Contact Information: nicolas.l.martinez@usmc.mil, 571-256-2739

EXECUTIVE SUMMARY

Project Summary

Our previous research (TRWG 13-01-016) demonstrated the feasibility and usefulness of ADF (Application Development Framework) to develop a suite of supply analytics to augment GCSS-MC. ADF allows developers to quickly develop data driven web based analytics. This study took the next step by installing the "Analtyics Suite (AS) on NPS virtual servers (using VMware Horizon View). The servers were installed in the .edu domain at NPS. Proper IA procedures were applied to the VMware physical machines. AS (csviewtb.nps.edu) have been made available to selected Marine personnel for evaluation. The goal of this research is to bring this proof of concept as close to a model of a production server as possible. Evaluation of the performance was studied as well.

In addition to the above, the research looked at how the GCSS data can be "cleaned up" and an evaluation of the technology needed to deal with dirty or missing data.

Background

We developed a supply analytics suite (AS) to overcome certain absence of supply analytics. This can be done at a much lower cost than modificatins done to the current Oracle E-Business Suite (EBS) –the foundation of GCSS-MC. Lower cost and time are because there are two common elements that can be leveraged by Oracle Application Development Framework, ADF which is the basis for our AS. Standard Oracle database and WebLogic which runs GCSS-MC also can run ADF web applications and be developed in a timely manner. This has been demonstrated both with logistics and supply data using ADF technology to develop quick applications. This was done two separate times by Marine thesis students under supervision of the NPS researchers.

As stated, previous research demonstrated through the feasibility and usefulness of ADF to develop a suite of analytics to augment GCSS-MC. ADF allows developers to quickly develop web applications used for analytics. We took the next step by successfully creating the analytic system on virtual servers (csviewtb.nps.edu). Our goal is to bring this proof of concept as close to a model of a production server as possible. Evaluation of the performance was studied as well. Previously the AS was just running on a stand alone laptop. What we did in the next step was to actually deploy the application to the full Oracle Enterprise Servers, for both the data base and the WebLogic Servers. These can be the basis for work needed to go into production. One item lacking in our AS is SSO (Single Sign On). However, OID (Oracle Internet Directory) or LDAP should be able to integrate seamlessly to the AS.

Findings and Conclusions

We successfully created a "proof of concept" for a pre-production GCSS-MC analytics platform for supply reports (for this report called, Analytics Suite or AS). In the previous research the GCSS-MC analytics suite ran only on a stand alone desktop and could not be accessible world wide. The new version is accessible through the WWW via a virtual desktop using passwords to access (GCSS-MC test data used). The virtual desktop accessed two "real world" servers: one for the database and one for the WebLogic application server. This would be the most likely configuration for a production machine. It is not recommened that the AS users access the GCSS-MC database directly for the following reasons:

- **Security**. Create another database instance separate from the enterprise one.
- **Performance**: Analytical tools should not be used at the expense of operational data on a transactional database so a separate database would help both the transactional side and the GCSS-MC analytics suite.
- **Dirty data or missing data**. Before these analytics can be deployed the database issues must be addressed and resolved. We found in several cases missing data that would make our analytics worthless. In some cases we "fixed" the database by adding test data or other methods to overcome any problems. This is documented in Appendix 3 of the previous study. One classic example is a closed service requests with several tasks and all with zero hours. Our informal investigations leads us to believe this may be due to poor training or not valuing the importance of inputing accurate data into GCSS
- **Normalization.** What is common with packaged software and suites such as E-business suite is that the databases may not be normalized or have the proper constraints needed to implement analytics and reports.

In the previous report we suggested possible actions to deal with the dirty or missing data problem:

1. Education and policy enforcement
2. Software enforcement of business rules

Our proof of concept assumes these actions have not been done although we still recommend those actions. Our solution then is to have two servers (WebLogic and the database) or one server for a separate database to be used for the GCSS-MC Analytics Suite. Extracts based on our previous research, would be taken from the source GCSS-MC database and then staged to this new database (as reflected in our GCSS-MC AS Proof of Concept). This is done through the ETL (Extraction Transformation Loading) to create a separate database or a "data mart" if historical data is of interest. ETL uses automated or manual software (scripts) to deal with the data:

1. Extraction: Using the roadmap in our previous deliverable pull out those tables and fields needed to drive the GCSS-MC AS.

2. Transformation: This is the most critical element of the process if the source database remains "dirty." A smart ETL process is required to clean up the dirty data and to deal with missing data. There are strategies to minimize the "damage" that is done by missing data needed for analytics. One advantage of analytics over transactional data is usually analytics are dealing with averages or trends and does not have to be exact if the decision maker still makes the right decision. An ETL process that can automate some of the cleanup will save time and cost. Contraint problems and other minor problems may be able to be resolved through the ADF framework so perfection in ETL may not be needed. Unification: Semantically disparate databases are likely to have tables and columns named inconsistently. Applying machine learning and necessary heuristics with human-in-the-loop to connect disparate silos of data event if the naming conventions are different is essential part of ETL overall process.

3. Loading. Since the requirement is not for streaming or near real time data the loading of the data shouldn't be a difficult task once it is transformed.

The ETL process therefor is essential for repairing anomalies. Several ETL products were investigated and one is given below as a suggestion of what capabilities are needed if none exist currently:

"The Tamr platform leverages machine learning algorithms to unify and prepare data across silos. Tamr's workflow combines automation with human collaboration across a range of data sources, including Hadoop, relational, cloud and so on."

"Tamr's automation has the capability to connect disparate silos of data at the entity and attribute level. It can identify common attributes and records even if the naming conventions are different. Continually leverages customer experts for validation of machine-learning matches to further automate future datasets and versions." Tamr (www.tamr.com)

In summary here are the basic elements needed to go into production and the progress so far:

- GCSS-MC Analytics Suite Proof of Concept: This can be ported to USMC servers with some network work required and of course IA and other security actions beyond the scope of our research.

- ETL: Must select ETL products are work with scripts to extract the data. The challenge is to select ETL that will aid in cleaning up the dirty or missing data. This is critical.

- Virtual or real?: The GCSS-MC Analytics Suite Proof of Concept run on VMWare Horizon. We would recommend setting up a virtual system which would minimize IA issues. If you keep the virtual desktop that would access the virtual servers you would reduce IA and security overhead. Some performance might

be gained by eliminating the virtual desktop and logging into the GCSS-MC AS server directly. It is recommend that you have at least 4 cores for each server and 8 GB of system memory (real or virtualized). If a virtual solution is selected and one doesn't exist in the network, we recommend VMWare Horizon over any Oracle product. We believe the management tools are superior.

- Training: Require at least one person to attend basic and advance Oracle ADF training and some knowledge of Java a big plus. Must have a WebLogic administrator who has taken both basic and advance WebLogic administrator course.

We are making the assumption that you have the human resources to deal with the data cleaning issue, ETL selection/implementation, network/security issues and basic database management administration (DBA). Without ETL we could not recommend implementing the Analytics Suite.

Also, LtCol Paul Ouellette (Head, LPC-3 HQMC I&L) showed the researchers additional analytics done manually on spreadsheets (offline) and asked: could this be done with the Analytics Suite, or in essence could ADF create a functionally equivalent? We believe yes, with Oracle Faces we could provide a functional equivalent. This could be a follow-on project to see how well ADF can create an online version that would meet the requirements. If the data is in GCSS-MC then we think this capability can be added to the AS.

References

Evaluating the Oracle Platform as a Decision Support System(TRWG 13-01-016), final report by Anthony Kendall, 2015.

"SAP HANA Defining Capabilities", SAP White Paper, Database Trends and Applications, November 2013, http://www.dbta.com/DBTA-Downloads/WhitePapers/SAP-HANAs-Defining-Capabilities-4805.aspx

TAMR (ETL software). (2017). Retrieved from http://tamr.com/