TATA CONSULTANCY SERVICES
Experience certainty.

ETL - Code Review Automation

20/05/2015

White Paper

Healthcare/CardinalHealth Inc.

Nikunj Kesharwani (535999)

Vidhi Agarwal (661400)

Data warehouse/ETL

nikunj.kesharwani@tcs.com
vidhi.agarwal@tcs.com

knowmax
Our Intelligence Hub

Confidentiality Statement
Include the confidentiality statement within the box provided. This has to be legally approved **Confidentiality and Non-Disclosure Notice** The information contained in this document is confidential and proprietary to TATA Consultancy Services. This information may not be disclosed, duplicated or used for any other purposes. The information contained in this document may not be released in whole or in part outside TCS for any purpose without the express written permission of TATA Consultancy Services.

Tata Code of Conduct
We, in our dealings, are self-regulated by a Code of Conduct as enshrined in the Tata Code of Conduct. We request your support in helping us adhere to the Code in letter and spirit. We request that any violation or potential violation of the Code by any person be promptly brought to the notice of the Local Ethics Counsellor or the Principal Ethics Counsellor or the CEO of TCS. All communication received in this regard will be treated and kept as confidential.

Table of Contents

Abstract

Paper proposes the Automation of Static Code Review process of IBM Infosphere Datastage ETL code. Currently this code review process is complete manual process in which developer, Integration lead and code governance team is involved.

Code governance team is a team within the organization who is responsible for performing the code review of Datastage ETL code submitted by developer for review.

Datastage ETL developer creates the code review submission document and sends an email to the code governance team. Code governance team performs the review manually by going through the code one by one against pre-defined rules sets manually. Code Review Automation, is to automate manual ETL Datastage code review process, which automatically validates the code against pre-defined rule sets to make sure code complies with standards for maintainability purpose.

About the Author

Nikunj Kesharwani
Around 4 years of experience with TCS as an ITA as ETL Data warehouse Architect\Designer and Developer. In-depth knowledge and experience of ETL Migration\Upgrade project along with the creating high level and low level ETL design and its development. Have handful healthcare domain knowledge and a good command on IQMS and IPMS processes. More than 1 year of experience of managing team of around 5 associates.

Vidhi Agarwal
Over 2.7 years of experience with TCS as a Systems Engineer with focus on Data warehouse/ETL. Have proficiency in developing and reviewing the ETL code. Also have in-depth knowledge of health care domain and exposure and understanding of various quality process & procedures laid in Integrated Quality Management System.

About the Domain

Data warehouse/ IBM Infosphere Datastage ETL

Data Warehousing is electronic storage of a large amount of information by a business. Warehoused data must be stored in a manner that is secure, reliable, easy to retrieve and easy to manage. There can be many type of Data Warehouses, most popular of them are Data Mart, Online Analytical Processing (OLAP), Online Transaction Processing (OLTP) and Predictive Analysis. In simple words we can say that Data Warehousing helps to maintain structured and unstructured data such that data is secure, reliable, easy to retrieve and easy to manage.

ETL comes from Data Warehousing and stands for Extract-Transform-Load. ETL covers a process of how the data are loaded from the source system to the data warehouse. Currently, the ETL encompasses a cleaning step as a separate step. The sequence is then Extract-Clean-Transform-Load.

There are many ETL tools in the market, that enterprises uses to maintain their Data Warehousing. One of the major such tools is IBM Infosphere Datastage 9.1(Datastage). Datastage can connect range of systems from flat files to complex databases. Datastage possess the ability to extract data from these systems, do necessary transformation (also cleansing) and load into the required target. Requirement in today's business world is to maintain large volume data in shortest time possible to make business decision on the data as early as possible to achieve business objective. Datastage helps organizations to meet their objective.

1. Problem Definitions

Current manual process of IBM Infosphere Datastage ETL code review takes lots of time and effort. This is performed by a dedicated Code governance team in organization. Because of the manual intervention and huge dependency on Code governance team, wait time for review completion is more and might also cause missed of any standard and will result into bad code in production.

1.1 Time

The first and the most important problem is that huge time involved in manual code review process. Time involved in creating the checklists as well as reviewing the code one by one manually and sending the approval/rejection emails. In case of code review failure, developer has to follow the same process again until the successful completion of the review. Along with this, wait time between submitting the code review and its completion is also more.

1.2 Efforts

This is 100% manual effort process. Developer has to create a code review checklist submission document, raising a code review request. Governance team has to go through the code one by one and require longer turnaround for approval or rejection.

1.3 Dependency

After raising a code review, dependency on Code Governance team and not able to migrate the code in further environments until code review is completed.

1.4 Budget

With big on-going projects, chances of budget overshoot occur due to code review efforts involved for more complex jobs due to time spent to perform their code review.

1.5 Quality

As the review process is a manual process, it can prone to some mistakes and there may be a chance that code is not standards compliance or as per expected quality.

1.6 Manage promise

Not able to deliver the deliverables on time due to waiting time of code review completion, delivering after deadline or delivering with poor quality will leave client with a dark impression.

1.7 Communication

Not updating progress of project and updating client about the concerns regarding delivery timelines due to waiting for code review to be done may lead to reduce trust factor.

2. Solution Details

In order to overcome the problems of manual process, paper proposes an idea of Self-Service ETL Automation tool using .NET code and validating the Datastage ETL code against a pre-defined rule set with the help of .Net XML parser created as a part of automation.
Below components developed as a part of this Self-Service Automation tool:

Windows batch script: A batch script to export the Datastage ETL code to be reviewed from IBM Infosphere using dsexport utility [1] of IBM Infosphere. This batch script is called within a desktop application built and installed in every developer's machine.
.NET XML Parser: A .NET code written to read the ported XML from above step and to parse against the pre-defined rule set to maintain standard and compliance.
Service now Integration: This is the ticketing tool [2] used in an organization to maintain the incidents and tasks. We have integrated the .NET XML parser in the mid layer of service now tool, where all the code resides. Once the user submits the task for performing code review, after all the appropriate approvals it calls the .NET XML parser and validates the ported XML against it. After validation, it sends an email for approval or rejection.

2.1 Time
Automation of the code review process will reduce the manual time spent and send the results very fast and quick.

2.2 Efforts
With the help of the automation tool, it will require minimal effort (including xml generation of the code, placing it on share path and raising a request to the Governance team) to review the code and generating the results.

2.3 Dependency
As XML Parser is a self-service tool, user needs to export the XML code of the ETL (through a static desktop application, installed on the VPC's of developers) and then put in a shared path. After this, user needs to submit a service now request, which will undergo the approval processes. Once all the approvals are done, it will trigger the XML parser to validate the ported XML's and send out an email to the requestor with the outcome (Approved/Rejected). In case of Code review rejection, an attachment will be sent out with the review comments.

2.4 Budget
After this automation, there will be no need of Code governance team and associates from that team performing the code review manually will not be charging the project bucket resulting in savings.

2.5 Quality
With reduction in time and manual effort and increase in speed with zero human errors the quality of the code will be maintained.

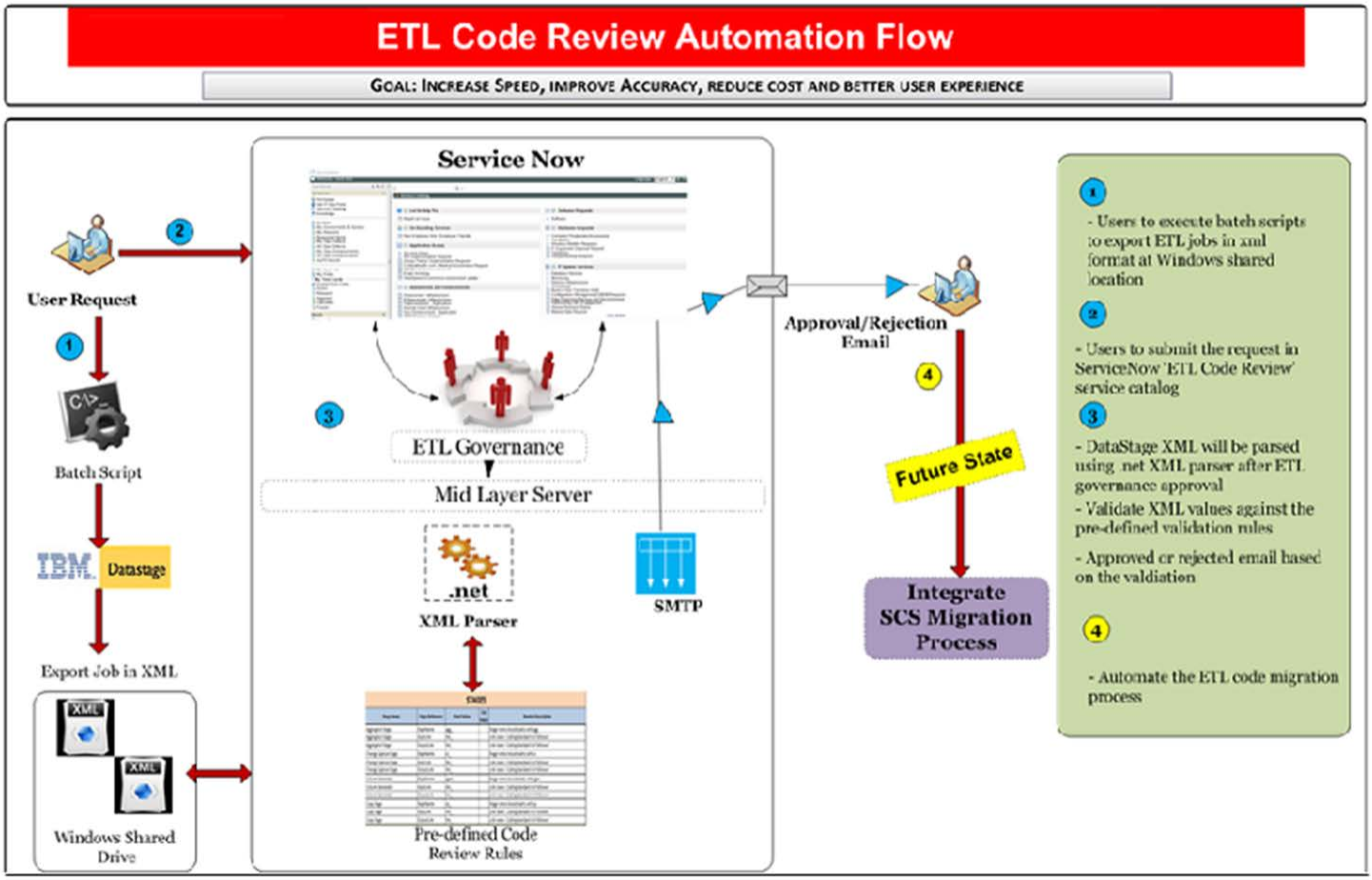
2.6 Manage promise
Developer will be able to complete the code review itself without waiting for anyone and hence they will be able to keep their promises to deliver the projects within given timelines.

2.7 Communication

User will be confidently stuck on what the project timelines have been set and will be able to communicate to the client about delivery of the projects.

2.8 Automation Flow

Below is the flow that will explain how the automation tool will work for automation of the Code Review process.



1. Developer needs to execute the batch script to generate the XML for Datastage ETL code for code review submission and need to keep the same at shared location from where service now will pick.
2. Service now is a ticketing tool, in which user need to submit a request for code review of ported Datastage ETL code.
3. Service Now will perform the code review with pre-defined rule set by parsing the ported XML with XML parser created in .NET and integrated with service now and send an email using SMTP server for approval or rejection.
4. This is the future state of this automation we are targeting. SCS team is the code migration team, who migrate the code from one environment to other environment. As of now this process is a manual process and 5 associates team is involved in same. By using istool utility [3] of IBM Infosphere Datastage, we are targeting to automate this code migration process.

3. Business Benefits

3.1 Self Service Tool

XML Parser serves as a self-service tool with no dependency on other team leading to on-time deliveries of the projects.

3.2 Increased Speed/Productivity

Automation of the process results into higher performance in terms of speed as less time will be required to complete the process.

3.3 Improved Quality

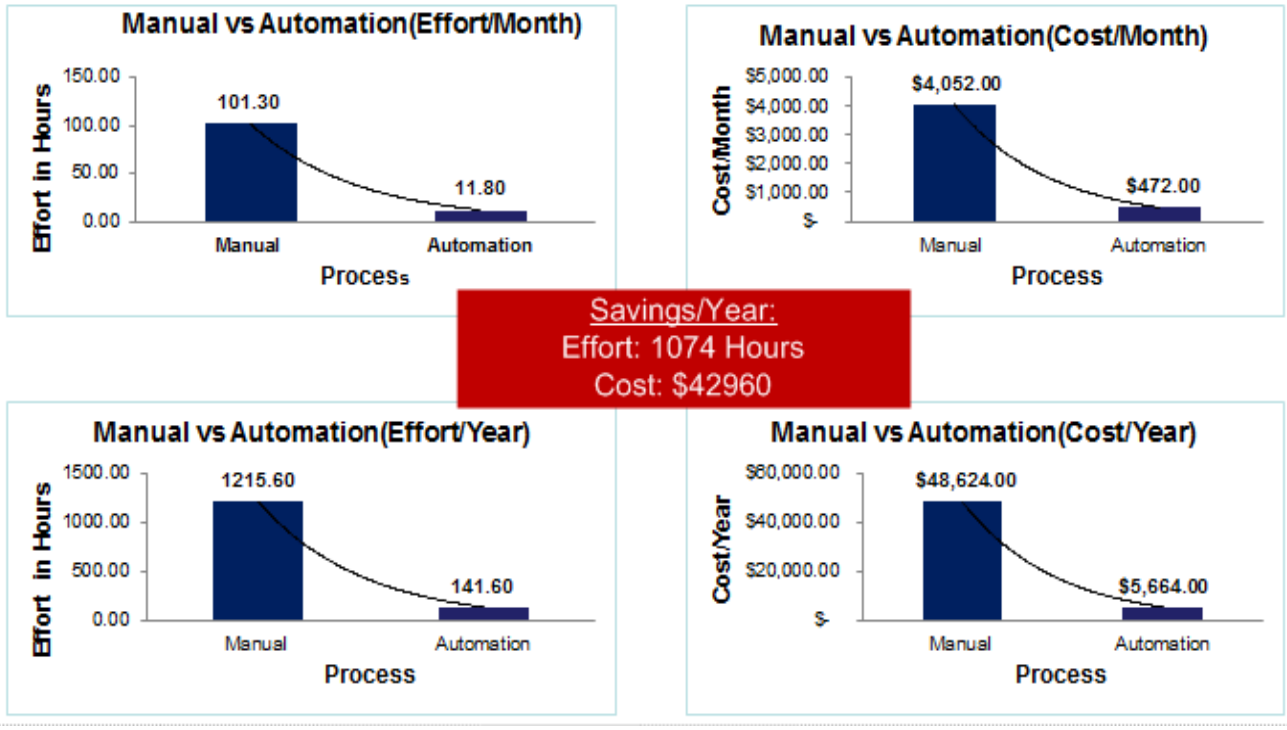
Manual efforts will be minimized and human errors will be undergrounded hence increasing the quality of the code to the mark.

3.4 Cost Reduction

As the tool will be installed on individual desktops and no extra team will be required to perform the tasks then it will reduce the cost of the entire process.

4. Metrics

This metrics is based on average 300 objects submitted per month for code review.



5. Conclusion

In order to minimize the manual efforts and increasing the productivity and improving the quality of the code, we will be able to achieve the same by automating the Code Review process and will be able to deliver the projects within specific time boundaries without any dependency and impact.

6. Acknowledgements

Thanks Subhankar Mohapatro and Parag Upadhye for their support in making this automation successful.

7. References

1. http://www-01.ibm.com/support/knowledgecenter/SSZJPZ_11.3.0/com.ibm.swg.im.iis.ds.design.doc/topics/dsexportcommand.html
2. http://www.servicenow.com/products/it-service-automation-applications/incident-management.html
3. http://www-01.ibm.com/support/docview.wss?uid=swg21437906

TATA CONSULTANCY SERVICES
Experience certainty.

Thank You

About Tata Consultancy Services (TCS)

Tata Consultancy Services is an IT services, consulting and business solutions organization that delivers real results to global business, ensuring a level of certainty no other firm can match. TCS offers a consulting-led, integrated portfolio of IT and IT-enabled infrastructure, engineering and assurance services. This is delivered through its unique Global Network Delivery Model™, recognized as the benchmark of excellence in software development. A part of the Tata Group, India's largest industrial conglomerate, TCS has a global footprint and is listed on the National Stock Exchange and Bombay Stock Exchange in India.

For more information, visit us at **www.tcs.com**.

IT Services
Business Solutions
Consulting