Streamlining Data Migration from PostgreSQL to SQL Server

Christopher P. Boone

Western Governors University

**Table of Contents**

[Summary 3](#_Toc61349471)

[Review of Other Work 4](#_Toc61349472)

[Changes to the Project Environment 6](#_Toc61349473)

[Methodology 7](#_Toc61349474)

[Project Goals and Objectives 9](#_Toc61349475)

[Project Timeline 11](#_Toc61349476)

[Unanticipated Requirements 13](#_Toc61349477)

[Conclusions 14](#_Toc61349478)

[Project Deliverables 14](#_Toc61349479)

[References 16](#_Toc61349480)

[Appendix A 17](#_Toc61349481)

[Appendix B 18](#_Toc61349482)

[Appendix C 19](#_Toc61349483)

# Summary

Over the past two decades HS Advantage has been growing at an increasingly fast rate. Each year has shown a student growth of at least ten percent which has caused the development and deployment of substandard systems. The reporting aspect for their field team has reached a point that the consistency of the reports and their data is no longer reliable as well as run times of reports is taking a large amount of time. This time lag and unreliable reporting causes the field team to spend more time fact checking reports instead of using them to make decisions. To resolve this problem, HS Advantage decided to build a data warehouse and implement an ETL process that helps to clean up the data and ensure accurate and timely reporting.

They have built a preliminary process to move and clean data into the data warehouse, but it is taking close to eight hours to load data. The initial goal of the project was to have data refreshing at least once an hour to give the field team up to date information that they can depend on to make quick decisions. Another goal of the project was to free up the time of the field team so that they can focus more on growing the company and supporting their clients.

Realizing that they, as a company, lack the expertise and experience to implement a complete ETL and data warehouse solution that meets requirements in the required amount of time, HS Advantage has reached out to SQL Solutions Group to aid in completing the project in a timely manner. The amount of time that was saved by bringing in an outside company ended up saving HS Advantage money in the long term. It also enabled their field teams to grow the company more efficiently and increased the satisfaction of the clients and families they support.

SQL Solution Group’s solution began the project with a review of all documentation of the original system. They then took any questions they had and held a review meeting with HS Advantage Inc. to establish the scope of the project as well as ensured there was understanding as to what the project would not include. This was done to help minimize any potential scope creep. Once done, the development teams from both companies met to orient SQL Solutions Group to the current infrastructure and systems. The development team was then able to move forward by conducting a detailed analysis to identify the cause of the slow process as well as established base line numbers to measure improvement. From this, documentation was done to outline the proposed SSIS flows which were then used to create the new solution. Once created, testing and timing was done to ensure accuracy of data as well as making sure the run time was under one hour. It was discovered during this phase, that the original dimensional flow for SSIS was not fast enough so a redesign was done and implemented. This added four days to the projected timeline. Once the solution was completed, it was deployed to a test environment then to production for final testing. This process went without error, so the team was able to recover lost time to keep the project on time. Finally, training on the new solution was done and a project closeout meeting was held to conclude the project.

# Review of Other Work

This article covers the pros and cons of three different methods to load dimension tables in a data warehouse utilizing SSIS. It starts off with the first and most common type, the SCD transformation. This transformation is known to cause many performance issues though. The cause of this impact to performance is due to, “A lot going on under the hood of the SCD component, and it can slow down the flow of your data considerably” (Moore, 2013). Keeping this in mind, the author goes on to describe two other methods that produce a more performance-oriented solution. The first of which is by using the SSIS Dimension Merge component, and the second, utilizing Lookup transformations. SQL Solutions group decided to focus on the Lookup method due to, “Lookups are fast” (Moore, 2013) as well as “Lookups are also easy to maintain” (Moore, 2013). The combination of speed and ease of use makes this the ideal solution for HS Advantage Inc. It helps ensure the runtime of one hour or less is met while also making it easy for their development team to work with.

When developing an ETL solution it is paramount that the whole team has a firm understanding of all the steps and what each one entails. That is why an article that gives a detailed explanation on all aspects as well as their importance is extremely helpful. This article is also helpful when trying to explain to the client how you plan to achieve what they request. It gives transparency and understanding to what is involved in the development of a solution. Starting with an introduction, the author gives a quick overview of what ETL is. A few common mistakes when developing a process are also underlined. “The most common mistake and misjudgment made when designing and building an ETL solution is jumping into buying new tools and writing code before having a comprehensive understanding” (Pathak, 2019). This statement reinforces the need for both the client and developers to have a firm foundation in the ETL process. By understanding the ins and outs, a lot of money and time can be saved if due diligence is given before beginning development. The author goes on to talk about some of the reasons as to why ETL is necessary and then focuses on the main elements of it. These elements consist of data extraction, transforming the data (cleaning), and loading it into the destination system. The author puts an emphasis on loading the data by pointing out some key areas to keep in mind such as the frequency at which data is extracted (Pathak, 2019). After an in-depth review of things to keep in mind as well as common problems, the author concludes with a hope that we have a better understanding (Pathak, 2019).

When finalizing a project, having a clear for closing it out helps to ensure all resources and project requirements have been met. This is built upon multiple groups which this article gives detail to how and what each one is about. What is project closeout? A client, such as HS Advantage Inc. that is a growing company may not have extensive knowledge or understanding as to its importance so articles like this can help to explain. The ultimate purpose of this is, “to formally complete the project, phase, or contractual obligations” (Aziz, 2015). The author gives a quick description of what project close out is followed by examples and why it is so important. This process “serves an important purpose for the organization and helps it avoid unfavorable and adverse scenarios” (Aziz, 2015). Moving on, more examples are given as to what can happen if this isn’t done correctly as well as what there is to be gained by this process. He goes on to cover some of the different methods that can be used along with some of the current best practices. Finally, Aziz closed his article with a method to help in deciding how to close a project and how to determine the best way to do this.

# Changes to the Project Environment

There were no major changes to the existing infrastructure of HS Advantage Inc. The environment still consists of four servers, two for testing, and two for production, and the hardware they each had did not change. The majority of changes occurred in the SSIS solution that was designed. It now utilizes the execution of multiple data flows that occur at the same time. This allows the solution to establish multiple connections to the PostgreSQL database which circumvents the limitation of it supporting parallelization. At the server level, SQL Solutions Group altered memory allocation to give precedence to certain processes at certain times. This increased the amount of hardware for the ETL process which in turn, helped improve runtime.

# Methodology

To complete this project SQL Solutions Group used a SDLC (Software Development Lifecycle) methodology, specifically the waterfall model. Due to the clarity of this model, it made the process easy to follow for the client as well as defined the goals clearly to help avoid any ambiguity in the requirements.

For the requirement analysis phase, SQL Solutions group reviewed the client’s current design as well as became familiar with their data and infrastructure. They also developed a base line of the current process as well as analyzed timings to identify the bottleneck in the current solution.

For the system design phase, the development team began by developing a plan to resolve the issues uncovered in the analysis phase. Once they had a design put together, they moved forward in creating the new solution and prepared it to be deployed to a test environment.

During the third and fourth phase, Implementation and testing, the team took the newly created solution and deployed it into the test environment. Extensive testing occurred in which the team verified data, compared timings to baselines, and ensured the solution met all project requirements.

Once all testing was completed and all fixes had been implemented, the development team deployed the solution to the HS Advantage Inc. production environment. They also oversaw the initial data load of the data warehouse and re-verified that the project met all requirements.

The final phase consisted of the maintenance phase. This is where a project hand off occurred and all training was conducted. Upon completion of training, all documentation and files were handed over to HS Advantage Inc. and one final meeting between the companies was conducted to close out the project.

# Project Goals and Objectives

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Goal | Supporting objectives | Deliverables enabling the project objectives. | **Met/Unmet** |
| 1 | Improve runtime of current data migration process. | 1.a. Understand current client process. | 1.a.i. Review current design document. | Met |
| 1.a.ii. Conduct system orientation with client. | Met |
| 1.a.iii. Complete project scope statement and have signed by client. | Met |
| 1.b. Identify bottleneck in current process. | 1.b.i. Meet with client for Q&A on current process. | Unmet |
| 1.b.ii. Research key differences. | Met |
|  |  |
| 1.b.iii. Time individual process components and create baseline. | Met |  |
| 1.b.iv. Build flow chart detailing steps and changes to the current process. | Met |  |
| 1.c. Design solution to bottleneck. | 1.c.i. Develop design document. | Met |  |
| 1.c.ii. Create SSIS packages to migrate data from PostgreSQL to SQL Server. | Met |  |
| 1.c.iii. Deploy Solution to test environment | Met |  |
| 1.d. Deploy solution to production. | 1.e.i. Deploy complete solution to production server | Met |  |
| 1.e.ii. Run initial load of production database. | Met |  |
| 1.e.iii. Verify solution meets specifications and ensure data integrity in the production database. | Met |  |
| 1.f. Project handover | 1.f.i. Provide documentation and review with client. | Met |  |
| 1.f.ii. Train client. | Met |  |
| 1.f.iii. Conduct project close-out with client. | Met |  |

Goal 1: Improve runtime of current data migration process. The goal to decrease runtime of the ETL process from eight hours down to one hour or less was the main goal of the project. The goal was completed and declared successful when SQL Solutions Group deployed a complete solution to a production environment and timed its completion in less than one hour consistently across a 24-hour period. The goal was built of five objectives:

Objective 1.a: To figure out the best solution to improve performance it was key to have a firm understanding of the clients’ current process and systems as well as establishing a scope statement. This objective helped to ensure all client needs were met. This objective was completed after SQL Solutions Group completed review of the original design documents as well as conducted a system orientation with the client. The final step that signaled success was the signing of a complete project scope statement.

Objective 1.b: Identify current process bottleneck. For this objective, it was important to hold meetings with the client to allow the development team of SQL Solutions Group to ask any questions about the existing solution as well as being brought up to speed as to what the HS Advantage Inc. development team thought what the issue may be. It was also important to help build a strong relationship between the two teams to help with future training. Having the two teams work together allowed faster identification of the original issue while gathering baseline data. This objective was considered a success when the baseline statistics of the original solution were gathered as well as having the major bottleneck in the current system identified and documented.

Objective 1.c: Design solution to bottleneck. Using the baseline statistics and the identified bottlenecks in the original solution, the development team designed, documented, built, and implemented a new solution to the test environment. This objective was considered successful when the test environment had an implemented version of the new solution that successfully ran and passed all testing. In this stage, an implementation and rollback plan were also created.

Objective 1.d: Deploy solution to Production. Utilizing the implementation and testing plans developed, the final version of the solution was implemented on the production server. Once implemented, an initial data load was conducted, and the testing plan was executed to verify the production environment. This was marked ‘successful’ when the deployed solution executed in under one hour and the accuracy of all data was verified.

Objective 1.e: Project handover. For this objective, all documentation was handed off to the client and training to maintain and administer the new solution was given. The successful completion of this objective was marked by a signed document that listed all project requirements and how they were met. This document was signed by all parties to formally close the project.

# Project Timeline

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Milestone or deliverable | Planned Duration | Actual Duration | Projected start date | Anticipated end date | Actual start date | Actual end date | Explanation of deviation |
| Review current design document. | 1 day | 1 day | 5/4/2020 | 5/4/2020 | 5/4/2020 | 5/4/2020 | N/A |
| Conduct system orientation with client. | 5 days | 4 days | 5/5/2020 | 5/11/2020 | 5/5/2020 | 5/8/2020 | System was less complex than expected. |
| Complete project scope statement and have signed by client. | 2 days | 2 days | 5/12/2020 | 5/13/2020 | 5/11/2020 | 5/12/2020 | N/A |
| Meet with client for Q&A on current process. | 2 hours | 5 hours | 5/13/2020 | 5/13/2020 | 5/13/2020 | 5/13/2020 | Extra discussion required to clarify project requirements. |
| Research key differences. | 1 day | 1 day | 5/14/2020 | 5/14/2020 | 5/14/2020 | 5/14/2020 | N/A |
| Time individual process components and create baseline. | 2 days | 3 days | 5/15/2020 | 5/18/2020 | 5/15/2020 | 5/19/2020 | Unexpected server downtime due to internet outage at home office. |
| Build flow chart detailing steps and changes to the current process. | 5 days | 4 days | 5/20/2020 | 5/26/2020 | 5/20/2020 | 5/25/2020 | Team was able to repurpose existing flow chart which saved one day of work. |
| Develop design document. | 5 days | 5 days | 5/26/2020 | 6/1/2020 | 5/26/2020 | 6/1/2020 | N/A |
| Create SSIS packages to migrate data from PostgreSQL to SQL Server. | 10 days | 14 days | 6/2/20200 | 6/15/2020 | 6/2/20200 | 6/19/2020 | Team encountered unexpected errors due to getting PostgreSQL and SQL server to communicate. Fix required a more complex solution which added to development time. |
| Deploy Solution to test environment | 3 days | 1 day | 6/16/2020 | 6/17/2020 | 6/22/2020 | 6/22/2020 | Original estimate built in time for handling errors but none occurred. |
| Deploy complete solution to production server | 1 days | 1 day | 6/18/2020 | 6/18/2020 | 6/23/2020 | 6/23/2020 | N/A |
| Run initial load of production database. | 1 day | 1 day | 6/19/2020 | 6/21/2020 | 6/24/2020 | 6/24/2020 | N/A |
| Verify solution meets specifications and ensure data integrity in the production database. | 3 days | 3 days | 6/25/2020 | 6/29/2020 | 6/25/2020 | 6/29/2020 | N/A |
| Provide documentation and review with client. | 2 days | 2 days | 6/30/2020 | 7/1/2020 | 6/30/2020 | 7/1/2020 | N/A |
| Train client. | 15 days | 12 days | 7/2/2020 | 7/22/2020 | 7/2/2020 | 7/17/2020 | Due to client being familiar with the tools, the basics included in the original training plan were able to be omitted. |
| Conduct project close-out with client. | 4 hours | 4 hours | 7/20/2020 | 7/20/2020 | 7/20/2020 | 7/20/2020 | N/A |

Due to the diligence of both companies’ development teams the project was able to be completed on schedule. The project did not, however, complete with zero issues. During the start, system orientation was completed one day ahead of schedule. This was due to effective communication amongst the teams as well as adhering to a strict itinerary. Another attribute that contributed to this reduction in time was because the HS Advantage Inc. systems were less complex than originally anticipated. The project continued to progress on time until the meeting to discuss and agree upon project requirements ran long. It took longer than anticipated to work with the client in limiting the scope of the project to a manageable level that could be completed on time. This added 3 hours to the projected time. The next issue that could not have been anticipated was an outage from the internet service provider. This caused work to stall as many of the existing systems exist in a cloud environment and were inaccessible. This added one day to the project, but the team was able to recover the lost time during the design phase by the slight alteration of already existing flow charts. The issue that had the largest impact on the project, which set it back 4 days, was the occurrence of unanticipated connection issues during development. The team struggled in finding a solution which resolved issues with the data moving from a PostgreSQL environment to a SQL Server environment. The team was ultimately able to find a solution which required a more complex solution. With this setback, the development teams worked hard in trying to save as much time as possible to get back on track. Due to no errors being encountered in the deployment process, the team recovered two of the days lost. Entering the training phase, the project was at risk to meet the project due date. Since the HS Advantage Inc. development team had been working closely with SQL Solutions Group throughout the project, three days were able to be saved during the in-depth training. The project was then finalized on time with the project closeout meeting and all project requirements were completed on time.

# Unanticipated Requirements

Throughout the project, the main unexpected issue that arose was the need to re-design the SSIS flow for loading dimensions. This decision was based on the need for a more efficient run time as without it, the goal of the ETL process running in one hour or less could not be met. This change added four days to the creation of SSIS packages. Due to this delay, the project was feared to miss its target date of July 20th, 2020 but thanks to the diligent effort of both companies’ teams, they were able to avoid any delays in other aspects of the project as well as gain back the days lost to the increased design time.

# Conclusions

Streamlining the ETL process for HS Advantage Inc. has shown to create a significant improvement in the overall system. Not only does the client now have a system that helps to increase their reporting capabilities, but they now also have a trained team that is capable of expanding their new solution as the company grows. The project was considered a success as it allowed HS Advantage Inc. to complete their full data warehouse project ahead of schedule along with provided them with an ETL solution that executes in less than one hour.

# Project Deliverables

Appendix A shows the SSIS workflow for the overall ETL process that the SQL Solutions Group created. The flow shows a high-level view of all steps involved to move HS Advantage Inc. data from PostgreSQL database to the new data warehouse environment. It is built of seven steps which includes automated auditing, loading of staging tables, populating dimensions, facts, and users. It also gives a high-level view of the error handling and notification system that is part of the complete solution.

Appendix B shows the SSIS flow that SQL Solutions Group is using to populate the dimension tables of the data warehouse. It starts with the retrieval of data from the staging tables then progresses to the utilization of slowly changing dimensions. Instead of using the built-in task for SSIS, they have used a more efficient method to sort and load the data. This method is more complex to build, which added to the development time, but is much faster. It utilizes a series of look ups and updates to complete the loading of the dimensions.

Appendix C shows the SSIS flow template that SQL Solutions Group has modeled all packages that load the fact tables of the data warehouse. It shows how data is retrieve from the staging tables and compared to what already exists in the fact table. If the record is a complete match, the record is ignored. If it is an existing record with a slight change, it expires the record in the fact and creates a new one based on the most recent record. If the record does not exist, it inserts the record as a brand new one.

# References

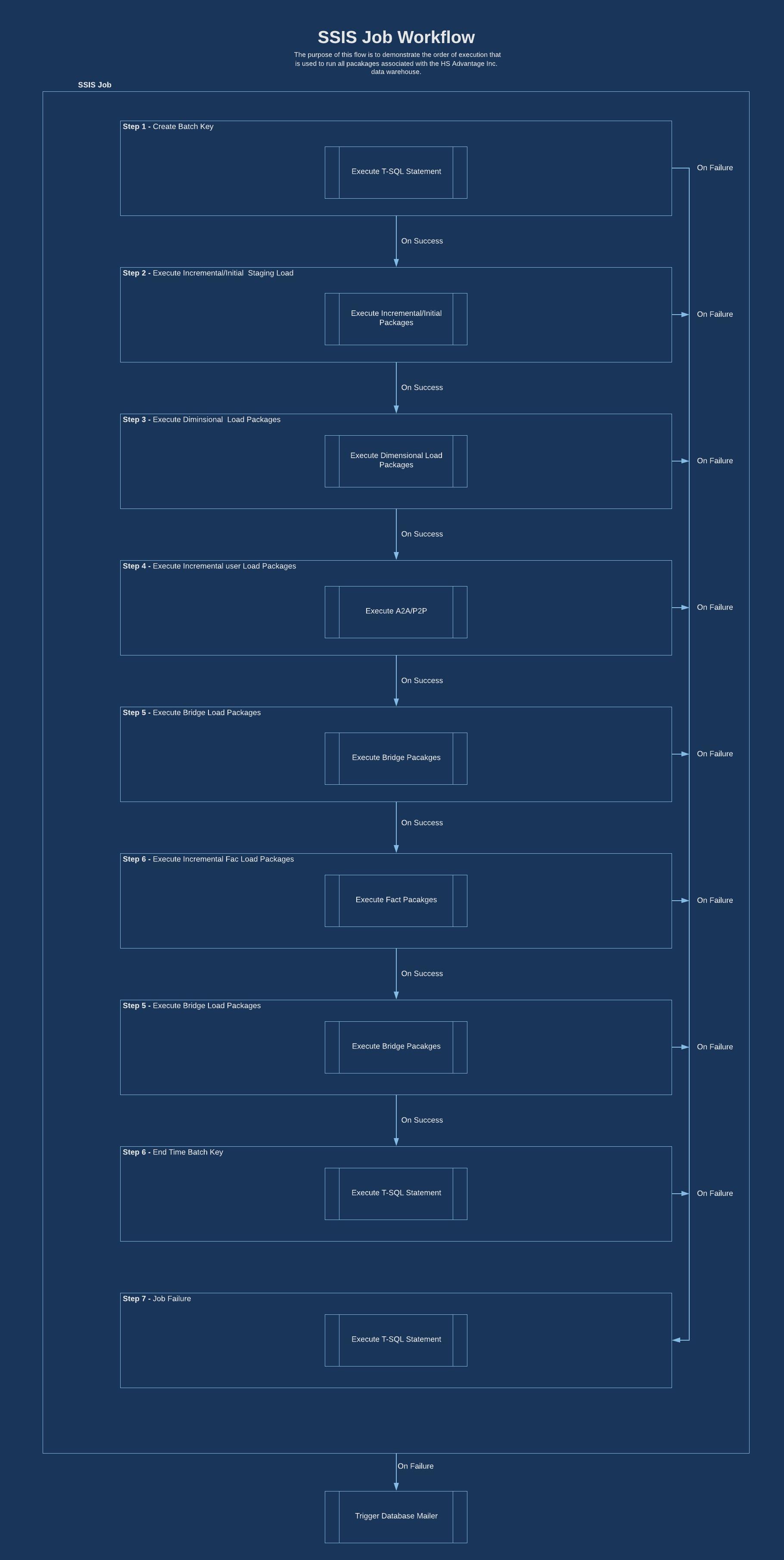
Aziz, E. E. (2015, October 10). Project closing the small process group with big impact. Retrieved January 11, 2021, from https://www.pmi.org/learning/library/importance-of-closing-process-group-9949

Moore, B. (2013, July 08). SSIS - Faster, Simpler Alternatives to the SCD Transform. Retrieved January 11, 2021, from https://benefic.net/ssis-alternatives-to-the-scd-transform/

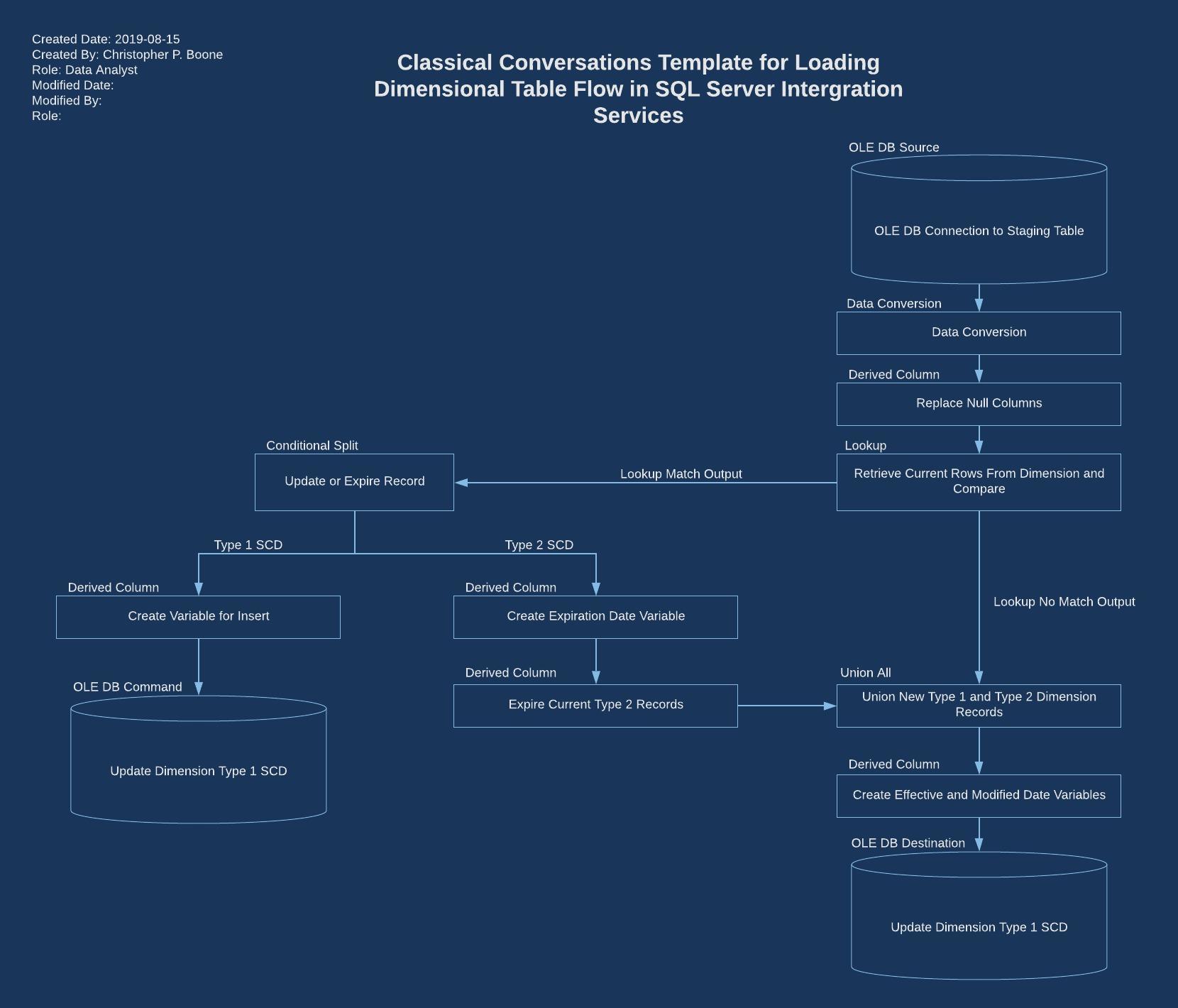
Pathak, P. (2019, January 07). ETL - Understanding It and Effectively Using It. Retrieved January 11, 2021, from https://medium.com/hashmapinc/etl-understanding-it-and-effectively-using-it-f827a5b3e54d

# Appendix A

SSIS Job Workflow



# Appendix B

SSIS Dimensional Load Workflow****

# Appendix C

SSIS Fact Load Workflow