Streamlining Data Migration from PostgreSQL to SQL Server

Christopher P. Boone

Western Governors University

**Table of Contents**

[Proposal Overview 3](#_Toc61016922)

[Problem Summary 3](#_Toc61016923)

[IT Solution 4](#_Toc61016924)

[Implementation Plan 4](#_Toc61016925)

[Review of Other Work 5](#_Toc61016926)

[Project Rationale 8](#_Toc61016927)

[Current Project Environment 9](#_Toc61016928)

[Methodology 10](#_Toc61016929)

[Project Goals, Objectives, and Deliverables 12](#_Toc61016930)

[Goals, Objectives, and Deliverables Table 12](#_Toc61016931)

[Goals, Objectives, and Deliverables Descriptions 12](#_Toc61016932)

[Project Timeline with Milestones 18](#_Toc61016933)

[Outcome 19](#_Toc61016934)

[References 20](#_Toc61016935)

# Proposal Overview

## Problem Summary

HS Advantage Inc. Is a mid-sized company that focuses on creating and providing home school curriculum as well as training and community support for parents. Their home office is in Southern Pines, North Carolina and have enjoyed large growth over the past five years. The way they provide support for the parents and families is by having many representatives spread out across the United States who are responsible for training, support, admissions collection, and growing the number of families that are involved with HS Advantage Inc.

The representatives rely heavily on reports and dashboards that are created by the business intelligence team. Due to significant growth of the company, the influx of data has grown to a point where running a single report can take up to five minutes or more. This impacts the ability of the representatives to quickly perform their job responsibilities in a timely manner.

To resolve this, HS Advantage Inc. has decided to build a data warehouse for reporting purposes. They are at the phase in the data warehouse project where they are trying to move data from their PostgreSQL database to the new data warehouse which is Microsoft SQL Server. To migrate the data, they are utilizing a tool called SQL Server Integration Services to automate the transfer of data. The issue they are facing is that the amount of time it takes to move the data takes eight hours and they are unable to get it to run at the desired one hour or less target time. The business intelligence team has been trying to improve the process for close to two months and the project has been put at risk. To get the project back on track they have decided to outsource the work of improving this process to SQL Solutions Group to assess the current process, improve the runtime to one hour or less, and train the HS Advantage Inc. business intelligence team on how to maintain and troubleshoot the new process.

## IT Solution

SQL Solutions Group has been hired to analyze, improve, and train HS Advantage Inc. on a revised solution to migrating data into their new data warehouse. The goal of this solution will be a complete redesign of the current process that will leverage current technology to implement an SSIS process that utilizes parallelization to maximize efficiency in loading the data to the new environment. The client’s current version of PostgreSQL does not inherently support the utilization of multiple cores for a single process which forces the current process to use only one core when retrieving data. Microsoft SQL Server does not have this limitation which means the loading process will not need to be rebuilt. To circumvent this, SQL Solutions Group will modify the current package to section out the data into batches and run multiple processes with separate database connections. Establishing numerous connections will cause PostgreSQL to create separate sessions which in turn will cause it to use more that one core. The new process will break it into four simultaneous batches to decrease the runtime to less than one hour. Doing this will utilize four cores of the six available, instead of one, and leave two cores available for other operations on the server. Once the new process has been implemented and meets requirements all documentation will be handed over to the client and extensive training on how the system works as well as modifying and maintaining it will be conducted.

## Implementation Plan

SQL Solutions Group will conduct this project using a phased approach. This is because each phase is dependent on the one before it to be completed. Without the tasks of the previous phase being done, work cannot begin as it builds on the work done before it. The first phase will consist of getting SQL Solutions Group familiar with the current environment, systems, and data of HS Advantage Inc. The second phase will consist of an analysis of the current process will take place as well as establishing a base line to measure improvement. This will allow the team to identify where the problem in the current process is. Once the cause of the problem has been pinpointed SQL Solutions Group will begin the third phase. This phase will consist of designing, documenting, building, and testing. Once this phase is complete, the solution will be moved into production for the fourth phase and re-tested to ensure all requirements are met. The project will close with the final phase which consists of handing over all files and documentation to the client for final review. Extensive training on the new system and how to maintain and modify it will also be conducted.

# Review of Other Work

When developing a new SQL Server Integration Services (SSIS) solution it is always important to keep in mind best practices. Doing this helps to ensure the design of an efficient solution that does not require an extensive tuning effort or redesign to get maximum speed and accuracy from the solution. That is where reviewing an article that provides an overview of best practices is important. Koen Verbeeck (2017) gives a good summary of where to focus attention and what some of the best practices are to ensure a SSIS solution that runs efficiently. The article does touch on parallelism, which is key to SQL Solution groups proposal, by saying, “Going into parallel will surely improve performance, but this is heavily influenced by available memory and the number of processors” (Verbeeck, 2017) but that there are limits. Keeping these limits in mind during development will help to leverage the available environment to obtain maximum performance. Continuing, Verbeeck discusses common elements to avoid such as blocking transformations as well as a handful of other transformations that impact performance negatively. Removing any of these transformations from the current process as well as avoiding them in any new designs will be important to ensure the best performance to meet the one-hour requirement for the project. Lastly, he touches on how to ensure writing data to the destination database is performing as quickly as possible.

When beginning a tuning cycle as well as trying to identify any bottlenecks in a current process an analysis must occur. This is where the article by Marcel Frank comes in handy. In the article he starts of explaining how to use a utility called Performance monitor. After going in depth on different metrics you can collect and how to understand them, Frank (2012) goes into detail on how to gather and interpret wait stats, latches, locks, virtual file stats, and top query stats. The main area of importance in relation to this project is the section that explains Performance Monitor. This is because the main purpose of this tool is, “To analyze and identify potential bottlenecks” (Frank, 2012). On top of being able to monitor and record performance metrics specific to SQL Server Integration Services (SSIS), the collection of other key system resources such as memory, disk, and processor use can be done all with one tool as well as create a log of all the metrics that spans a designated period. Though the article covers many other monitoring and performance tips which are all very important regarding performance tuning for SQL Server. The scope of this project mainly deals with SSIS so therefore the in-depth explanation as to how and what to monitor with Performance Monitor is key in the SQL Solutions Group development team. This tool aids in determining what the main cause of poor performance is as well as areas that can be improved on to maximize performance.

<https://nexxtjump.com/2012/06/12/ssis-performance-tuningmonitoring-data-collection/>

Knowing where to begin your tuning effort when trying to increase the performance of an ETL process can be a daunting task. Through the aid of reliable articles though, this task can be made much easier and they can help to give direction to tuning efforts. Often, they will also contain bits of information that you may not have known about. In an article written by Anoop Kumar (2021), he highlights some of the top methods to improve ETL performance. One of these is the importance to “extract data in parallel” (Kumar, 2021). Kumar (2021) goes on to explain, while using SQL Server Integration Services (SSIS), how to “design a package in such a way that it can pull data from non-dependent tables or files in parallel, which will help to reduce overall ETL execution time” (Kumar 2021). This portion of his article directly applies to the project for HS Advantage Inc. The company is expressing an issue with poor performance with their current ETL process and upon review, SQL Solutions group discovered that one key improvement to the current solution would be to utilize the hardware’s ability to processes large amounts of data at the same time (in parallel). Kumar goes on to highlight some other key areas to consider when tuning such as component configuration in SSIS, the type of data connectors to use, and the proper way to type cast when moving data between a PostgreSQL database and a SQL Server database

Another key area of ensuring an efficient process is to ensure your database connections are configured properly. Also, understanding how these connections work can help in pinpointing potential performance issues. By knowing the best way to connect to a PostgreSQL database, SQL Solutions Group can limit the number of errors that may have to be handled in a more complicated way. In the article, the author found that, “after mapping the source columns to the target table, I received several error messages: “Cannot Convert between Unicode” and “Non-Unicode String Data Type” (Alston, 2016). Upon more investigation it was found that this was due to the ADO.net connection manager, “does an implicit conversion from varchar to nvarchar regardless of the source database encoding” (Alston, 2016). She further elaborates as to how handling this issue would drastically increase the development time of solutions which in turn could delay the project. The article goes on to explain the alternative Alston found to save development time as well as a few other errors encountered when trying to work with the PostgreSQL database. This method included utilizing the 32-bit version of the PostgreSQL ODBC driver as opposed to the 64-bit. This is due to SQL Server Integrations Services (SSIS) being a 32-bit application (Alston, 2016). In conclusion, the article ends with a summary of which ODBC driver to use when working with PostgreSQL and SSIS as well as pointing out the difficulty she had of, “finding immediate answers to many of these issues online” (Alston, 2016).

# Project Rationale

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 will be saved by bringing in an outside company will end up saving HS Advantage money in the long term. It will also enable their field teams to grow the company more efficiently and increase the satisfaction of the clients and families they support.

After reviewing multiple proposals. HS Advantage decided that SQL Solution Group’s solution required the fewest changes as well as providing a foundation that would continue to support their ever-growing company. By not having to completely redesign and start from scratch, SQL Solutions Group was able to propose a timeline that surpassed the other proposals as well as provide one of the most cost-effective options. With the agreement of provided training on the new system, HS Advantage also ensures that their company will be equipped to maintain and modify the system as they grow to meet future needs.

# Current Project Environment

HS Advantage Inc. currently has much of the infrastructure already in place for this project. They have four server environments consisting of two that host a test and production version of their PostgreSQL database as well as two more that consists of a test and production environment for their data warehouse project. Both production machines are built on Windows Server 2017 and have 2.5 GHz eight core processors with 64 GB of memory. The production SQL Server box hosts SQL Server Enterprise edition which has SQL Server and SQL Server Integration Services (SSIS) installed. The PostgreSQL server contains the database connected to HS Advantage Inc.’s user portal and currently supports all reporting needs for the company. These are hosted in an AWS cloud environment to ensure high availability as well as fail over support. The test environments consist of slightly smaller machine, also hosted on AWS, that are configured with 2.5 GHz quad core processors and 32 GB of memory. The PostgreSQL test server contains a weekly copy of the production PostgreSQL database and the SQL Server test box currently contains the ETL process that takes over eight hours to run.

The proposed solution will use the current infrastructure to help keep costs low. Once complete the new design and system tuning will increase the efficiency at which the ETL process runs allowing room for additional processes without negatively impacting performance of the current process. It will also ensure that, with a quick ETL process, the reporting capabilities and accuracy for HS Advantage Inc. will improve. This enhancement will allow the companies field team to be able to rely on data accuracy, spend less time verifying report accuracy, and allow them to contribute more time to growing the company’s student base.

The training and establishment of a strong initial ETL process will set the company up for success long term. Not only will HS Advantage Inc. be left with a finished ETL process that runs in less than one hour, but they will have a team that is trained and fully capable of maintaining the system as well as expanding it to accommodate future growth.

# Methodology

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

For the requirement analysis phase, SQL Solutions group will review the clients current design as well as become familiar with their data and infrastructure. They will also develop a base line of the current process as well as analyze timings to identify the bottleneck in the current solution.

For the system design phase, the development team will begin to develop a plan to resolve the issues uncovered in the analysis phase. Once they have a design put together, they will move forward in creating the new solution and prepare it to be deployed to a test environment.

The third and fourth phase, Implementation and testing, the team will take the newly created solution and deploy it into a test environment. Extensive testing will occur where the team will verify data, compare timings to baselines, and ensure the solution meets all project requirements.

Once all testing has completed and any fixes have been implemented the development team will deploy the solution to the HS Advantage production environment. They will also oversee an initial data load of the data warehouse and verify a second time that the project meets all requirements.

The final phase will consist of the maintenance phase. This is where a project hand off will occur and all training will be conducted. Upon completion of training, all documentation and files will be handed over to HS Advantage Inc. and one final meeting between the companies will be conducted to close out the project and obtain any final signatures.

# Project Goals, Objectives, and Deliverables

## Goals, Objectives, and Deliverables Table

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

## Goals, Objectives, and Deliverables Descriptions

Goal 1: Improve Runtime of Current Data Migration Process. This is the main goal for the project being done for HS Advantage Inc. to improve the runtime performance of an existing SQL Server Integration Services (SSIS) solution that migrates data from a PostgreSQL database to a Microsoft SQL Server database. This goal will be considered a success when a new SSIS solution is in place for a test and production environment. It is also important that the new solution can complete a full run within 1 hour. The whole project falls under this one goal and will be completed using five objectives:

Objective a: Understand current client process. To figure out the best solution to improve performance it will be key to have a firm understanding of the clients’ current process and systems as well as establishing a scope statement. This objective will help to ensure all client needs are met. This objective contains three deliverables:

Deliverable i: Review current design document. The existing design document for the original process was reviewed by team members individually. Our team then met together to build a list of questions and concerns to

present to the client.

Deliverable ii: Conduct system orientation with client. Meeting with the client, our two teams will conduct a review of current database structures as well

as process flows. We will work with the client to answer any questions

we have in regard to how they have configured and implemented their

database servers.

Deliverable iii: Complete project scope statement and have signed by client. Finalize all project requirements as well as what will not be in scope of the project. Once all parties agree, document, and have the business owner sign.

Objective b: Identify current process bottleneck. For this objective, it will be 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 thinks may be the issue. It will also be important to help build report between the two teams to help with future training. Having the two teams work together will allow faster identification of the current issue while gathering baseline data. This objective will be considered completed when baseline statistics of the current solution are gathered as well as having the major bottleneck in the current system identified and documented. This objective contains four deliverables:

Deliverable i: Meet with client for Q&A on current process. Conduct in person meeting with client’s development team to familiarize with the current

process and any issues they have encountered. Also use this time to gather insight into what the client suspects may be the issue as well as to build team cohesion between the organizations.

Deliverable ii: Research key differences. Conduct system comparison to identify key differences between SQL Server and PostgreSQL that may require extra processing time when migrating the data. Some key aspects to consider

will be the utilization of parallelism, data conversion requirements and

identifying known issues for both platforms at an individual level as well as when

trying to make the systems work together.

Deliverable iii: Time individual process components and create baseline.

Establish baseline execution times for each component of the current

process. Collection of data in respect to the volume of data being

transferred as well as CPU and memory usage will be necessary in establishing an

accurate baseline to measure performance increase for the new process and in

identifying any bottle necks.

Deliverable iv: Build flow chart detailing steps and changes to the current

Process. After analyzing all baseline data build detailed models describing how

each bottleneck will be improved. Include details on what steps are

needed for tuning and workarounds to better utilize modern technology

to improve execution. This may include adjusting hardware and server

settings, removing unnecessary steps, performing analysis of data being moved and limiting data sets to only what is needed.

Objective c: Design solution to bottleneck. Using the baseline statistics and the identified bottlenecks in the current solution, the development team will design, document, build, and implement a new solution to a test environment. This objective will be considered complete when the test environment has an implemented version of the new solution that successfully runs and passes all testing. In this stage, an implementation and rollback plan will also be created. This objective contains three deliverables:

Deliverable i: Develop design document. During this phase it will be important to create a formal document that details all changes and process

improvements that will be done. This document will contain process flows and detailed descriptions of all steps necessary to implement changes to the current process to ensure execution time requirements are met.

Deliverable ii. Create packages to migrate data from PostgreSQL to SQL Server. The development team will take the design document and build the

processes that extract, transform, and load the data from the source system to the destination system. Developers will conduct their own level of testing to ensure data is maintaining its integrity in the move. The measurable for this step will be a complete set of packages to migrate data from PostgreSQL to SQL Server.

Deliverable iii: Deploy Solution to test environment. Development team will create an implementation plan as well as a rollback plan in case of any failures during the deployment of the solution. Once the document is created and reviewed the team will then conduct the solution deployment following the plan and document any new steps that may have been missed in the original implementation plan. Once deployed, the team will perform end to end unit testing as well as data verification to ensure data integrity between the two systems. Lastly, the team will time the execution of the solution to ensure that it meets project requirements. The measurable for this step as well as a milestone will be a deployed solution that meets all project requirements in a test environment along with an implementation plan.

Objective d: Deploy solution to Production. Utilizing the implementation and testing plans developed, the final version of the solution will be implemented on the production server. Once implemented, an initial data load will be conducted and the testing plan

Deliverable i: Deploy complete solution to production server. Using the implementation plan that was tested in the previous phase, deploy the new solution to the production server.

Deliverable ii: Run initial data load of production database. Conduct a complete run of the deployed solution. During this test, no errors should occur. If the process fails for any reason, conducting a roll back according to the implementation plan will be done, and the error will be recreated and resolved in the test environment before re-deploying. The measurable for this step will be a complete solution deployed on the production server that is scheduled to run at a regular interval per the project requirements without errors.

Deliverable iii: Verify solution meets specifications and ensure data integrity in the production database. Using the baseline established at the beginning of the project, all steps of the new solution will be timed individually and as a whole. From this, the amount of improvement can be measured as well as allows to ensure all time requirements are met for the project. At the same time, the development team will use the testing plan developed to ensure all data between the source and destination systems are correct.

Objective f: Project handover. For this objective all documentation will be handed off to the client and any training to maintain and administer the new solution will be given. The successful completion of this objective will be marked by a signed document that lists all project requirements and how they were met. This document will lastly be signed by all parties to formally close the project.

Deliverable i: Provide documentation and review with client. All documentation that was created during the project will be handed over to the clients. This will ensure the client has all the tools available to train on the new system and maintain it moving forward.

Deliverable ii: Train client. Conduct hands on training with the client on how to maintain, monitor, administer, and troubleshoot the new solution. This will consist of work sessions where documentation and processes will be reviewed combines with hands on learning with the new system in the test environment.

Deliverable iii: Conduct project close-out with client. The final step in the process will entail all business owners as well as our team meeting to review the project statement to ensure all requirements were met. If there are any requirements that the client feels are not met, a discussion on how to resolve the discrepancy will take place. A lessons learned meeting will also be conducted before all parties agree the project is complete. Finally, a formal write-up documenting the results of the project as well as stating all the requirements were met will be created and signed by all parties.

# Project Timeline with Milestones

|  |  |  |  |
| --- | --- | --- | --- |
| Milestone or deliverable | Duration (hours or days) | Projected start date | Anticipated end date |
| Review current design document. | 1 day | 5/3/2021 | 5/3/2021 |
| Conduct system orientation with client. | 5 days | 5/4/2021 | 5/10/2021 |
| Complete project scope statement and have signed by client. | 2 days | 5/11/2021 | 5/12/2021 |
| Meet with client for Q&A on current process. | 2 hours | 5/13/2021 | 5/13/2021 |
| Research key differences. | 1 day | 5/14/2021 | 5/14/2021 |
| Time individual process components and create baseline. | 2 days | 5/17/2021 | 5/18/2021 |
| Build flow chart detailing steps and changes to the current process. | 5 days | 5/19/2021 | 5/25/2021 |
| Develop design document. | 5 days | 5/26/2021 | 6/1/2021 |
| Create SSIS packages to migrate data from PostgreSQL to SQL Server. | 10 days | 6/2/20201 | 6/15/2021 |
| Deploy Solution to test environment | 3 days | 6/16/2021 | 6/17/2021 |
| Deploy complete solution to production server | 1 days | 6/18/2021 | 6/18/2021 |
| Run initial load of production database. | 1 day | 6/19/2021 | 6/21/2021 |
| Verify solution meets specifications and ensure data integrity in the production database. | 3 days | 6/22/2021 | 2/24/2021 |
| Provide documentation and review with client. | 2 days | 6/25/2021 | 6/28/2021 |
| Train client. | 15 days | 6/29/2021 | 7/19/2021 |
| Conduct project close-out with client. | 4 hours | 7/20/2021 | 7/20/2021 |

# Outcome

Streamlining the ETL process for HS Advantage Inc. will create a strong environment for further company growth as well as increase employee productivity. Drastically reducing the ETL run time from eight hours to less than one will unlock a level of accurate reporting the company has never had. At the end of this project, HS Advantage Inc. will have a fully trained and proficient team to maintain and grow their new ETL process as well as increased reliability and efficiency in their new data warehouse for reporting needs. Success will be measured when the HS Advantage Inc. ETL process runs in less than one hour as well as having a team that is trained in maintaining and growing the new process within three months of the project kick off.

# References

Alston, A. (2016, March 14). SQL Server Integration Services Connecting to PostgreSQL Database. Retrieved January 08, 2021, from https://www.credera.com/insights/sql-server-integration-services-connecting-to-postgresql-database/

Franke, M. (2012, June 12). SSIS Performance Tuning–Monitoring & Data Collection. Retrieved January 08, 2021, from https://nexxtjump.com/2012/06/12/ssis-performance-tuningmonitoring-data-collection/

Kumar, A. (2013, August 15). Top 10 Methods to Improve ETL Performance Using SSIS. Retrieved January 08, 2021, from https://www.developer.com/db/top-10-methods-to-improve-etl-performance-using-ssis.html

Verbeeck, K. (2017, August 31). SQL Server Integration Services SSIS Performance Tuning Techniques. Retrieved January 08, 2021, from https://www.mssqltips.com/sqlservertutorial/9070/sql-server-integration-services-ssis-performance-tuning-techniques/