**SSIS Package Development Case Study for Duplicate Invoice Analytics**

**Objective**

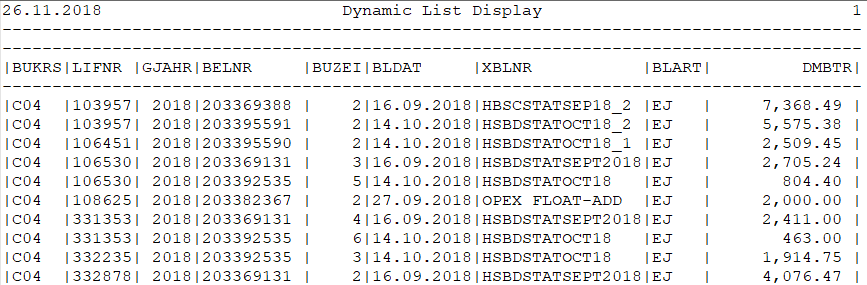
The objective of this case study is to evaluate the candidate’s knowledge and ability to translate data processing requirements into implementation using the Microsoft SQL Server data analytics tools provided by the SQL Server Integration Service (SSIS). This case study resembles one of the skills required for the role, albeit in a simplified format, and thus it is expected that the candidate is able to perform this task with minimal guidance and supervision.

**Case Overview**

One of the controls that are checked during an account payable audit is whether there have been any duplicate payments made to the same vendor for a particular service or product delivered to the company over the course of a period. Since the payment process in the organization is all automated in the ERP system, the way to check this is to download all the payment data from the system and perform data analysis in order to see whether there had been any cases of payment made to the same vendor for a particular invoice (based on invoice number, date and amount). In order to monitor these exceptions on continuous basis, the Internal Audit will automate this data analysis process using the Microsoft SQL Server Integration Service (SSIS) tool and let it run periodically on monthly basis.

**The Raw Data**

The raw data obtained from the ERP system is delimited text data (separated by pipes i.e. the “|” character) and has the header names on the top row. It has some rows that need to be ignored (could be due to erroneous line) when we process the data so this need to be taken care of when importing raw data using the SSIS package component. The raw data looks like the snapshot below:



The following are the details of the columns with their descriptions:

|  |  |
| --- | --- |
| Field | Description |
| BUKRS | Company Code. For the purpose of this case study, just analyze the duplicate invoices for company C04 |
| LIFNR | Vendor code (unique to each vendor) |
| GJAHR | Financial Year. Not useful for our analysis. |
| BUZEI | Line item (BELNR combined with BUZEI is a unique key for the record) |
| BELNR | Accounting document number (BELNR combined with BUZEI is a unique key for the record) |
| BLDAT | Invoice document date on the physical invoice |
| XBLNR | Invoice reference number on the physical invoice |
| BLART | Document type. Not useful for our analysis |
| DMBTR | The Invoice Amount |

The raw data is provided alongside this case study paper.

**The Analyses**

1. The first task for this case study is to find exact duplicate invoice record entries in the payment dataset provided. For this, you have to come up with the analytic logics that determine what a ‘duplicate payment’ record is i.e. what conditions would constitute a duplicate payment entered in the system.

If any exact duplicate payment records are found, the exceptions will then be exported into a CSV file for the business to analyze and investigate.

1. The second task is to produce near-duplicate invoice records in the system. Near-duplicate invoice is an invoice number that differs only by a few characters with another invoice (e.g. an invoice that differs only with an apostrophe which is a common technique used to circumvent system controls). There are various approaches that can be used to find near-duplicate records in the database. However, for the purpose of this case study, only the readily available SSIS component will be used.

If any near duplicate payment records are found, these will be exported into a CSV file for the business to analyze and investigate.

1. The third task is to provide an exploratory data visualization from the invoice records using Google Data Studio. The visualization report should aim to give an insightful view to the management on the dataset that we’re working with.

**The Steps**

You’re required to develop two SSIS packages that imports the raw data, processes it and produces all the results into CSV files (one package for exact duplicate and another for near-duplicates). These must be performed entirely within the SSIS packages. The data visualization dashboard meanwhile is to be developed on Google Data Studio.

Below are the recommended step-by-step to achieve the analytics objective. However, you may use any approach that might be suitable as long as the required deliverables are met.

Step 1: Installing the required tools

* Download and install the SQL Server 2017 Developer Edition (Free) if you don’t have a SQL database yet on your machine (<https://www.microsoft.com/en-us/sql-server/sql-server-downloads>).
* Download and install the Visual Studio Community Edition (The IDE is needed to develop the SSIS package). This tool is freely downloadable from <https://visualstudio.microsoft.com/vs/community/>
* Download and install the Microsoft SQL Server Data Tools for the Visual Studio IDE, which is crucial when developing SSIS on the Visual Studio (you have to install this after you’ve installed the Visual Studio IDE) (the installer can be obtained from <https://docs.microsoft.com/en-us/sql/ssdt/download-sql-server-data-tools-ssdt?view=sql-server-2017>)
* Download and install the SQL Server Management Studio (SSMS), which is an essential database management tool for your database for task like viewing your result tables, testing SQL scripts etc. (<https://docs.microsoft.com/en-us/sql/ssms/download-sql-server-management-studio-ssms?view=sql-server-2017>)

Step 2: Configure and create a database and working tables

* Configure your SQL server service to run locally on your machine
* Open SSMS and connect to your new SQL Service
* Create an empty database
* The SSIS package that you will develop will use the newly created database

Step 3: Create SSIS packages

* Create a simple SSIS package (using the standard SSIS components) to perform the following task:
  + Import raw data and put it in working table(s) in the database
  + Process the records in order to find exact duplicate invoices into a result table
  + Export the result table records into a CSV file
* Create another SSIS package (using the standard SSIS components) to perform the following task:
  + Import raw data and put it in working table(s) in the database
  + Decide what threshold parameter to be used for the near-duplicate component, if any, and justify why it is so.
  + Process the records in order to find near duplicate invoices into a result table
  + Export the result table records into a CSV file
* (Optional) Explore alternative ways of achieving the exact duplicate invoice record entries within SSIS (i.e. using SQL task vs. common data flow components). List down the pros and cons of using the two different approaches when developing SSIS packages. Discuss these during the interview.

Step 4: Create data visualization using Google Data Studio

* You may have to create a Google account in order to use Google Data Studio.
* Export the dataset into a data source type that is usable by Google Data Studio.
* Develop a simple dashboard report on the platform to provide an exploratory data visualization insights on the dataset provided.

**The Deliverables**

Please submit the following files to us prior to the interview:

1. A short explanation of the logics and/or justifications used to determine what constitutes a duplicate invoice is, and write them in a text file named Logics1.txt.
2. A short explanation of the logics and/or justifications used to determine what constitutes a near-duplicate invoice is, and the corresponding threshold parameter used if any, and write them in a text file named Logics2.txt.
3. The complete Visual Studio solution project files used to develop the two SSIS packages in Step 3
4. The CSV output file of all the exceptions generated from the raw data using the SSIS package
5. The link to the Google Data Studio report developed in Step 4 (please don’t make the report public, direct share the report with [faizula@maxis.com.my](mailto:faizula@maxis.com.my) instead)

Please zip all the files and information above and email to us, or alternatively share them via a cloud file sharing e.g. Google Drive, Dropbox or OneDrive.

You are required to present all the solutions during the interview session. Please bring your own laptop and ensure that all the deliverables above can be demonstrated during the session.

Many thanks and see you at the interview!