CONFEXX
CONSULTING

Knowledge Assessment

Name of Candidate:

Date:

June 19, 2020

Useful Resources

Introduction to Talend Studio	https://www.youtube.com/watch?v=iXFAajRCAbw
What is ETL (Extract, Transform, Load)?	https://www.talend.com/resources/what-is-etl/

Technical Keywords

Talend, ETL, MySQL, SQL Server, Data Integration, Python, R

Required Time

4h + Presentation

Setting Up Your Local Machine (depending on your choice of tools / technologies)

https://www.anaconda.com/

Anaconda

https://www.apachefriends.org/de/index.html

Apache, MySQL, PHP

https://www.python.org/downloads/

Python3

https://de.talend.com/download/

Talend Open Studio

Use Case

Functional Description

Business has a large customer database, which has been maintained via a html web frontend. The web frontend application is some 20 years old. Along with the customer database, there is also a database, which contains credit card information. The credit card database is logically and technically separated from the customer database.

Business is aware that there they have a lot of issues with the quality of their data. For instance, there are no table keys defined so there may be a lot of redundant entries. The web frontend does not check any input and all entry fields are free text.

Business asks for your help with this. They would like you to analyze the data and give them recommendations on data quality and how to improve it. In a first step, they would like to know if there is a set of credit card information for each customer (which they assume should be the case). Additionally, they would like to see if their data can be merged and exported to other formats to distribute them company-wide.

Technical Description

Data Integration

Please implement the following requirements using a tool of your choice (being it a ETL tool like Talend or a programming language like Python, Java…). You are free to choose any implementation approach / tool if you think it's better suited for a particular task.

You are given two different data sources, customers.xml and creditcard_data.json. These files contain full snapshots, meaning they contain all the available data at one point in time, one, being customers and, two, credit card data.

In a first step, please read the contents of each of the full snapshots and output them to console. Then enhance your job and add database output. As a database backend you may use MSSQL Server Express or MySQL or any relational database that is available to you. Create a database table for each data source. Add meaningful primary and foreign keys. Then add an additional column to each table, INSERT_TS. INSERT_TS is the insert timestamp with the following sample pattern: "2009-10-02 16:52:30 (CET + 1)".

Add another output channel, json. Both data sources should be merged into one json file. Assign credit card data to the respective customer when available (if applicable).

Presentation

Present your results. Don't spend much time on your presentation's preparation. You are free to choose any type of presentation media that you like. You have max. 20 minutes for you presentation followed by an approx. 10 minute long discussion round.

Business wants you to touch on the following aspects during your presentation:

- Are there any issues with the data quality?
- Are there any credit card datasets that could not be connected to customers?
- What do you recommend business on how to improve their existing data model?