I. ABOUT

Author name: Maciej Tomaszewski

Create date: 05.03.2024

Description: my second Python & SQL project, this documentation describes its overview

Tech-stack: Python, Microsoft SQL Server, Oracle, draw.io, Git, GitHub, Microsoft Excel, Microsoft Word

Attachments:

1. **List of steps to perform before testing the solution:** 1_before_testing
2. **Ms SQL - tables with data:** HR, HR_addr
3. **Ms SQL – scripts preparing data from point 2:** 3_1_pyp2_mssql1, 3_1_pyp2_mssql2, 3_1_pyp2_mssql3
4. **External files:** 3_2_hr_avro, 3_2_hr_excel, 3_2_hr_json, 3_2_hr_parquet, 3_2_hr_xml
5. **Python script with main content:** 4_pyp2_main

At first, user should perform all steps described in the file "1_before_testing"!

II. OVERVIEW

The purpose of the project is to process data:

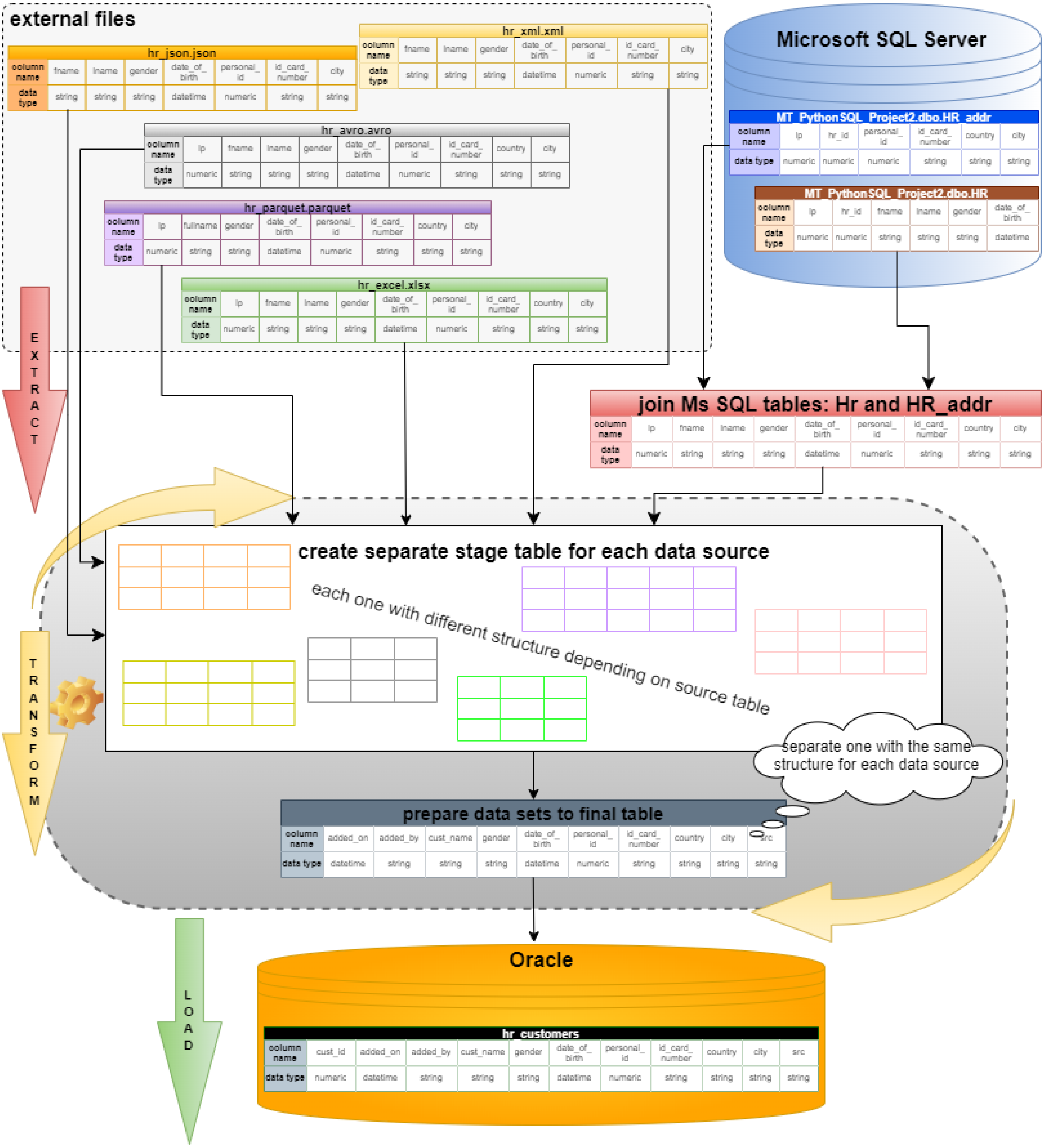
1. **Extract** from 5 external files and 2 tables in database management system – Ms SQL
2. **Transform** and standardize data structure
3. **Load** into target table in another database management system – Oracle

It consists of a few main steps:

- **0 import libraries and set variables**: prepare necessary components before further data processing
- **1 create stage tables**: prepare temporary tables to store extracted data
- **2 prepare and insert data to stage tables**: extract data and load them to stage tables
- **3 create target table**: prepare table to store final, transformed data
- **4 prepare data for target table**: transform data and create the same structure for each data source
- **5 insert data to target table**: load data to dedicated table
- **6 drop stage tables**: remove stage tables once no longer needed
- Also, the code returns information on successful or failed execution. In case of error, additional details are displayed.

This is the second project I prepared using Python and SQL. Its main purpose was to boost my knowledge in field of using Python to process data. Few dozens of hours were spent for completing it, I also faced plenty of problems and errors when doing so. It definitely boosted my knowledge in field in using Python to work with databases.

A chart with accurate data flow is attached on the next page.



external files

hr_json.json

column name	fname	lname	gender	date_of_birth	personal_id	id_card_number	city
data type	string	string	string	datetime	numeric	string	string

hr_xml.xml

column name	fname	lname	gender	date_of_birth	personal_id	id_card_number	city
data type	string	string	string	datetime	numeric	string	string

hr_avro.avro

column name	ip	fname	lname	gender	date_of_birth	personal_id	id_card_number	country	city
data type	numeric	string	string	string	datetime	numeric	string	string	string

hr_parquet.parquet

column name	ip	fullname	gender	date_of_birth	personal_id	id_card_number	country	city
data type	numeric	string	string	datetime	numeric	string	string	string

hr_excel.xlsx

column name	ip	fname	lname	gender	date_of_birth	personal_id	id_card_number	country	city
data type	numeric	string	string	string	datetime	numeric	string	string	string

Microsoft SQL Server

MT_PythonSQL_Project2.dbo.HR_addr

column name	ip	hr_id	personal_id	id_card_number	country	city
data type	numeric	numeric	numeric	string	string	string

MT_PythonSQL_Project2.dbo.HR

column name	ip	hr_id	fname	lname	gender	date_of_birth
data type	numeric	numeric	string	string	string	datetime

join Ms SQL tables: Hr and HR_addr

column name	ip	fname	lname	gender	date_of_birth	personal_id	id_card_number	country	city
data type	numeric	string	string	string	datetime	numeric	string	string	string

create separate stage table for each data source

each one with different structure depending on source table

separate one with the same structure for each data source

prepare data sets to final table

column name	added_on	added_by	cust_name	gender	date_of_birth	personal_id	id_card_number	country	city	src
data type	datetime	string	string	string	datetime	numeric	string	string	string	string

Oracle

hr_customers

column name	cust_id	added_on	added_by	cust_name	gender	date_of_birth	personal_id	id_card_number	country	city	src
data type	numeric	datetime	string	string	string	datetime	numeric	string	string	string	string