BIDW Group Assignment 2

O-2 Group H

ETL – Data Integration

Members: Dagoberto Romer, Abhyudaya Choumal, Duarte Dias Costa, Areknaz Khaligian, Harpreet Perhar, Philippe El Hage, Conrad Lee

Instruction Manual:

Note: Attached to this manual, there will be a .rar file. An appropriate tool is needed to open this file into all the parts that will be needed to generate and populate the database. (In the case of using Blackboard, the instructions will be inside the rar file to upload 1 single file).

Tools Needed: .rar tool, Dataiku DSS (free community version is enough), Pentaho PDI (free community edition is enough), MySQL 8 + MySQL Workbench.

When opening the .rar file a folder and 2 files will be inside. H2AVisaDBV2 is the schema file that will generate the schema in MySql for the database.

The ETL Maps PPT file is documentation to help you understand the process better. And the Folder contains everything needed to run the integration tool, the datasets and the Backups in case of a problem.

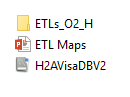


Figure 1. Contents of Rar file

The first thing we need to do is forward engineer the database and then create the dataset that we will use to work.

To forward engineer the database, Open the H2AVisaDBV2 file in MySQL workbench. Select forward engineer and make sure that the database doesn’t exist before. Else check the options to DROP Objects and create DROP Schema before forward engineering so we have a blank slate.

Next is generating the dataset in Dataiku.

In the Input folder inside ETLs\_O2\_H is a Dataiku Project called BIASSIGNMENT2. Import this in Dataiku DSS, open the flow and build the last dataset in the flow. Make sure to check Recursive and Forced Recursive in the options.

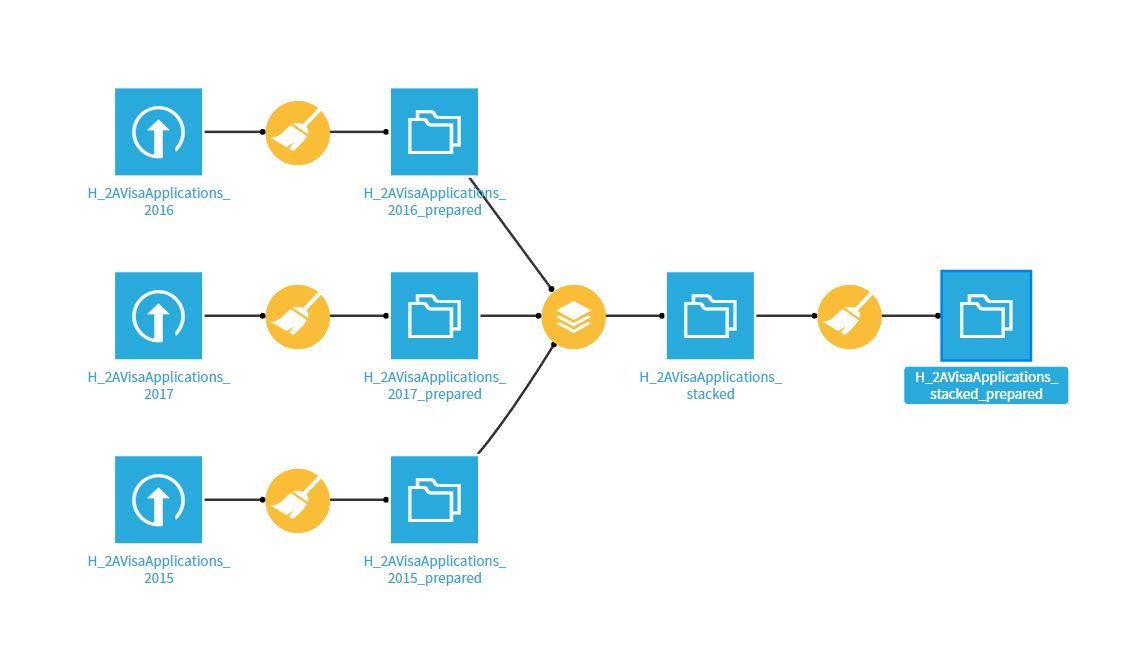


Figure 2. Dataiku project flow with needed dataset highlighted

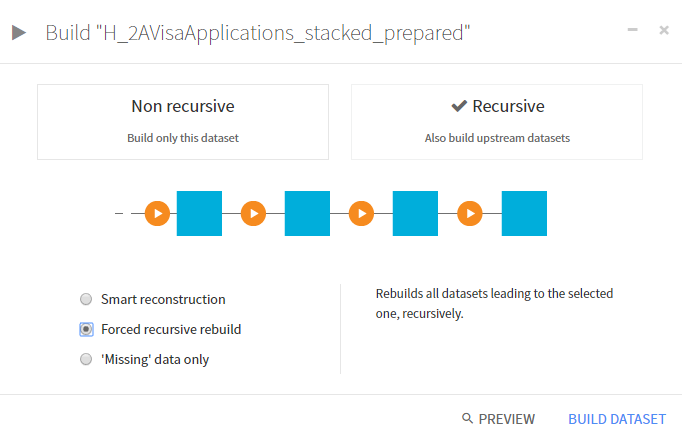


Figure 3. Build options and dataset name

Once the dataset is build click on “Export” on the top right and export it as a comma delimited CSV file.

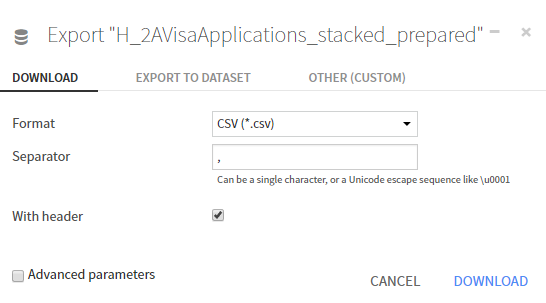


Figure 4. Export options for the Dataset

**IMPORTANT NOTE: for the Pentaho script to work, the resulting dataset needs to be put inside the /ETLs\_O2\_H/Input/ directory. Right next to where the Dataiku Project was. The name of the file needs to be “H\_2AVisaApplications\_stacked\_prepared” exactly. Any other file will mean Pentaho won’t be able to find the file.**

After putting having the MySQL database generated and the dataset inside the Input Folder. We are ready to run the Pentaho job.

**Note:** The job is set up to run automatically, however there are some requirements. First is that the database was forward engineered by MySQL so that the schema name matches the connection in PDI. Also, the database needs to be in localhost and the port needs to be at 3306. If any of these things change, the the connection string in every transformation needs to be updated.

Then we need to open the Job file in the ETL Folder called JOB\_H2AVISAAPPS. We should get the following screen.

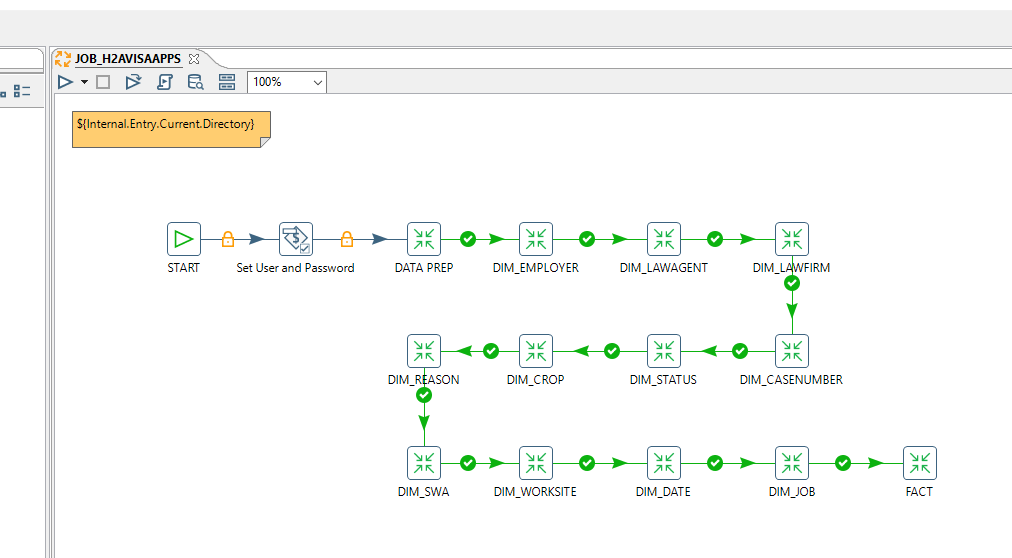


Figure 5. Main Job in Pentaho PDI

The important part here is to double click the “Set User and Password” node to introduce the database credentials.

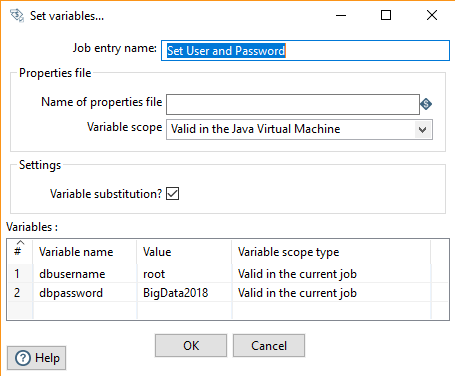


Figure 6. Setting root username and password

This 2 variables will be propagated to the other transformations as soon as you click “Start Job”

Save the Job file, and click on “Run” to start the Data Integration process.

The run time can vary from 15 minutes on a i5-4690k Desktop with 16GB RAM, to 35 minutes on a i5-6300U Laptop with 8 GB RAM.

When the Job is done, the database should be populated with the data from the dataset.

Troubleshooting:

In case some part of this process can’t run. We have provided backups for every step inside the Backups folder.

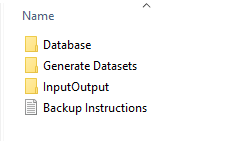


Figure 7. Backup Folder

On the Generate Dataset is another Dataiku Project and the original datasets to be uploaded, and run to get the input dataset. On the InputOutput folder we have provided the input dataset needed by Pentaho as well as the output dataset that Pentaho generates on the TR\_DATA\_PREP transformation.

On the Database folder there are SQL Scripts to generate and populate all the tables in the dataset in case Pentaho doesn’t work or is unavailable. There is also a SQL script called “h2achecktablesquery” with a simple query that joins all the tables to check that the relationships between tables work.