|  |
| --- |
| Vijf tips voor een succesvolle Proof of Concept - Bizztracker |



|  |
| --- |
| Data transformation POC  80K in five minutes?!  23-01-2022 – Thijmen Brand |

# Introduction

Contents

[Introduction 1](#_Toc125385784)

[Abstract 1](#_Toc125385785)

|  |
| --- |
|  |

# Abstract

<quick overview of whats done and quick overview of the findings>

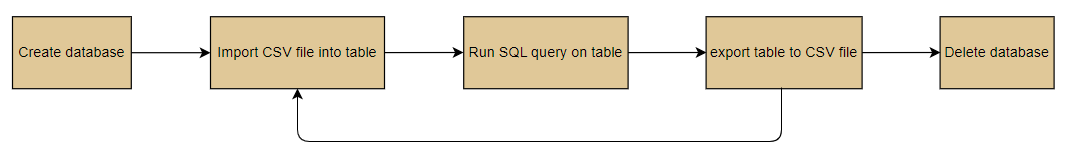
# Introduction

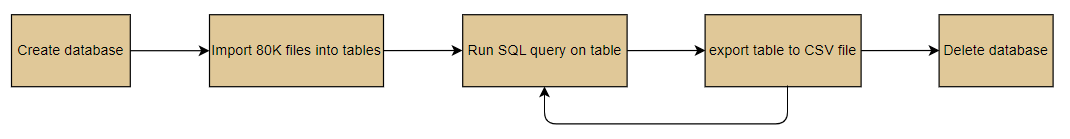
The objective of this POC is to transform CVS files from one format to the other. This itself isn’t that hard to do, but it becomes difficult when that process needs to be done with 80 to 100 thousand CSV files.

To easily perform the transformation we can run some SQL query on the data. This can rapidly transform data. But in order to be able to execute SQL queries on the data we need to somehow load the file into a database. Because we can’t directly query the file.

The best database to do this is an in memory database, because the transformations have to happened very quick we need to be able to read and write in micro seconds of speeds. And this is best achievable with an in memory database. There are two databases that will be used for this proof of concept. The first one is DuckDB and the other one will be Sqlite. These database provide an easy to use api to spin up a quick proof of concept.

In order to do the transformations there are a few possible methods which could work.





# 

These three flows showed above will be tested with the two database mentioned above.

# DuckDB

## Method 1

### NodeJS

Memory: 11.3 70%

CPI: 1.3 – 1.5 GHz 20 – 50% usage

Time: 15:16.041

### Python

Time: 08:08

## Method 1.2

### Nodejs

Out of memory

### Python

11:26

## Method 2.1

### Python

07:12

## Method 2.2

### Python

27:43