**DMAS Demo-Showcase**

**Snowflake (Atlas of Economical Complexity)**

|  |  |
| --- | --- |
| Version: | 1.0 |
| Datum: | 06.02.2025 |
| Status: | **in Arbeit** / vorgelegt / freigegeben |
| Autor: | Matthias Balzer |
| Ablage: |  |

# Ziel

Der Showcase demonstriert die allgemeinen und Fähigkeiten des Cloud-DBMS Snowflake Daten zu laden. transformieren und zu präsentieren.

Der Showcase lehnt sich an den 'Atlas of Economical Complexity' (AoEC) aus dem 'Harvard Dataverse' an (<https://dataverse.harvard.edu/dataverse/atlas>). Die verwendeten Daten wurden aus diesem System bezogen

Technisch sollen die folgenden Aspekte demonstriert werden:

* Laden und Speichern großer Datenmengen in Snowflake mit relationalen Datenstrukturen
* Laden aus Cloud-Systemen
* Konvertierung von speziellen Datenformaten, hier Stata Binärdaten.
* Transformation der Daten in Strukturen, die für die Visualisierung optimiert sind.
* Visualisierung der Daten

Darüber hinaus wird eine Architektur vorgestellt, die sowohl

* Plattform unabhängig wie auch
* unabhängig vom Visualisierungstool

ist.

# Umsetzung

*Hier Verweis auf die zu Grunde liegende Confluence Epic*

# Beschreibung

## Fachliche Problemstellung

*Fachliche Beschreibung steht noch aus.*

## Generelles Systembeschreibung

Die Daten wurden aus dem AoEC-System kopiert und auf unterschiedlichen Plattformen abgelegt. Diese Plattformen sind:

* Windows-Filesystem (PC)
* Clous-System (AWS)

Von dort werden sie in das Snowflake-System übertragen und in einer relationalen Tabellenstruktur (Database) abgelegt.

Die Statistikdaten liegen in einem Binärformat vor, das für die Verarbeitung, insbes. Visualisierung, so nicht verwendet werden kann. Deshalb müssen diese Daten in ein ladbares Format konvertiert werden. Nachdem die Daten konvertiert und relational abgelegt sind, werden sie für die Visualisierung vorbereitet und dafür in einen Datamart geschrieben.

Aus diesem Datamart bezieht die Visualisierungskomponente ihre Eingangsdaten und benötigt dann keine weiteren Ergänzungen oder Transformationen mehr.

Ein Bild, das Text, Diagramm, Plan, technische Zeichnung enthält.

Automatisch generierte Beschreibung

Cloud AWS: Dies kann auch ein anderes Cloud-System sein, wie z.B. MS Azure oder Google Clous Service.

Stage (extern): Die Daten aus der Cloud werden nicht physisch abgelegt, sondern es handelt sich nur um Verweise, die das Kopieren aus der Cloud ermöglichen. Hier fallenkeine Kosten für Datenspeicher an.

Stage (intern): Daten, die nicht in einem Clous-System abgelegt sind, müssen zuerst physisch in das Snowflake-System übertragen werden (Upload) , damit sie dort verarbeitbar sind. Nach der Übertragung in die Datenbasis (Elementary Data) können die Daten aber wieder entfernt werden, um Kosten für Speicherplatz zu vermeiden.

Conversion & Load: Üblicherweise werden die binären Stata-Daten für allgemeine System (nicht-Stata) verfügbar gemacht, indem sie in CSV-Files konvertiert werden. In Snowflake kann diese Konvertierung aber ohne explizite physische Ablage erfolgen, d.h. die Daten werden transient konvertiert und danach sofort, ohne dauerhaft gespeichert zu werden, in die Datenbanktabellen geschrieben.

Datamart: Der Datamart stellt eine einheitliche Schnittstelle für unterschiedliche Visualisierungstools dar. Damit können diese ohne besondere Funktionalitäten für eigene Transformationen auskommen. Als Visualisierungstools sind angedacht:

* Streamlit (Snowflake-eigenes Dialog- und Darstellungstool)
* Tableau
* dbt

## Fachliche Systembeschreibung

Ein Bild, das Text, Diagramm, Plan, Screenshot enthält.

Automatisch generierte Beschreibung

## Architektur des Showcase

Ein Bild, das Diagramm, Text, Entwurf, Zeichnung enthält.

Automatisch generierte Beschreibung

# Aufbau

## Basisstruktur

Die Basisstruktur für die Datenbank des Showcase in Snowflake besteht aus:

* Einem Snowflake-Account
* in diesem Account eine Engine für Datentransporte (Snowflake Warehouse)
* in diesem Account eine Datenbank 'DMAS' (Snowflake Database)
* in dieser Datenbank das Schema 'AOEC'
* einem GitHub-Repository 'DMAS'

| ***#*** | ***Objekt-Name*** | ***Objekttyp*** | ***System*** | ***Erstellung mit*** | ***Anmerkung*** |
| --- | --- | --- | --- | --- | --- |
| 1 | DMAS\_2X | Warehouse | SN | dmas\_setup.DB.sql[[1]](#endnote-1) | WH speziell für Uploads |
| 2 | DMAS | Database | SN | dmas\_setup.DB.sql\* |  |
| 3 | AOEC | DB-Schema | SN | dmas\_setup.DB.sql\* |  |
| 4 | DMAS | Repository | GitHub | --- | Skripte und Notebooks |

## Datenstrukturen

### Tabellen der Stammdaten

### Tabellen der Bewegungsdaten

### Tabellen und Views des Datamart

# Laden der Daten

## Voraussetzungen

### Installationen

| ***#*** | ***Item*** | ***Typ*** | ***Version*** | ***Anmerkung*** |
| --- | --- | --- | --- | --- |
|  | snowSQL | Snowflake-Utility | aktuellste |  |
|  | AOEC\_FILES\_HS12 | Snowflake Stage | --- |  |
|  | Python |  | 3.9 |  |
|  | pandas | Python Package |  |  |
|  | sqlalchemy | Python Package |  |  |

## Laden der Stammdaten

### Tabellen der Stammdaten

|  |
| --- |
| * LOCATION\_COUNTRY |
| * LOCATION\_GROUP |
| * LOCATION\_GROUP\_MEMBER |
| * PRODUCT\_HS12 |
| * PRODUCT\_SERVICES\_UNILATERAL |
| * PRODUCT\_SITC |

Die Tabellen werden beim Ausführen des Lade-Programmes automatisch erstellt.

### Laden über Stage

Tabelle 1: Upload der Masterdaten (CSV-Dateien) nur mit SN-Mitteln

| ***#*** | ***Aktion*** | ***Umgebung*** | ***Skript/Programm*** | ***Anmerkung*** |
| --- | --- | --- | --- | --- |
| 1 | Erstellen der Zieltabellen | SN (Snowsight oder snowsql) | dmas\_setup.DB.sql |  |
| 2 | Upload der CSV-Dateien mit Masterdaten zu Stage | Windows-Kommandoeingabe | aoec.products.load.bat | Führt aus: aoec.products.load.sql |
| 3 | Übertragen der Masterdaten in Tabellen | SN (Snowsight oder snowsql) |  |  |

### Laden mittels Python-Programm

Tabelle 2: Upload der Masterdaten (CSV-Dateien) nach SN mit Python-Programm

| ***#*** | ***Aktion*** | ***Umgebung*** | ***Skript/Programm*** | ***Anmerkung*** |
| --- | --- | --- | --- | --- |
| 1 | Installation benötigter Python-Packages |  | --- | Packages:   * pandas * sqlalchemy |
| 2 | Konfiguration des Python-Programms import\_MasterData\_directly.py |  | --- | Zu ändernd:   * [Zeile 19] Parameter 'u' von sql.create\_engine: aktueller SN-User-Name * [Zeile 20] Parameter 'p' von sql.create\_engine: Passwort * [Zeile 30] Variable 'csv\_path': Folder der CSV-Dateien mit Master-Daten |
| 3 | Upload der CSV-Dateien mit Masterdaten in Tabellen | Windows-Kommandoeingabe | python import\_MasterData\_directly.py | Die Ziel-Tabellen werden beim Ausführen des Lade-Programmes automatisch erstellt. |

## Laden der Statistikdaten

Die Statistikdaten können aus dem Internet-Auftriff des AoEC als .zip-Datei heruntergeladen werden und liegen dann mit dem Namen vor:

datvers\_files.zip

Diese Datei enthält .dat-Files im STATA-Format. Dieses Format kann nicht direkt von Snowflake verarbeitet werde. Es muss entweder in ein CSV-Format oder in einen Dataframe umgewandelt werden.

Um zwei unterschiedliche Methoden des Ladens großer Datenmengen zu demonstrieren, wurde die Menge der Dateien mit Statistikdaten in zwei Gruppen aufgeteilt:

### Laden über Stage

#### File-Set

Dieser Ladevorgang wurde auf die folgende Teilmenge der HS12-Statistikdaten eingeschränkt (*Data-Set A*):

* hs12\_country\_country\_product\_year\_1.csv
* hs12\_country\_country\_product\_year\_2.csv
* hs12\_country\_country\_product\_year\_4\_2012\_2016.csv
* hs12\_country\_product\_year\_1.csv
* hs12\_country\_product\_year\_2.csv
* hs12\_country\_product\_year\_4.csv
* hs12\_country\_product\_year\_6.csv

#### Voraussetzungen

1. Interner Stage DMAS.AOEC.AOEC\_FILES\_HS12

#### Ablauf/Struktur

##### Vorbereitungen

Step Aktion Erläuterungen

1

##### Start

Das Skript:

aoec.hs12.statistics.datSet\_A.upload.bat

muss in einer Windows-Kommando-Box gestartet werden.

##### Ablauf

|  |  |  |  |
| --- | --- | --- | --- |
| ***Step*** | ***Umgebung*** | ***Aktion*** | ***Erläuterungen*** |
| 1 | Windows CMD | Starten von aoec.hs12.statistics.datSet\_A.upload.bat | Führt snowSQL-Skript aoec.hs12.statistics.datSet\_A.upload.sql aus. |

# Anhang

## Glossar

| ***Abkürzung/Begriff*** | ***Bedeutung*** |
| --- | --- |
| AoEC | Atlas of Economical Complexity |
| Data-Set A | Teilmenge der Statistik-Dateien des Daten-Set HS12 des AoEC. (Siehe dazu: [File-Set](#_File-Set)) |
| sog. | sogenannter, sogenannte, sogenanntes |

1. In GitHub-Repository DMAS abgelegt. [↑](#endnote-ref-1)