Projektarbeit
Business Analytics

Fritz Greulich

Abstract

Das vorliegende Projekt beschreibt den Aufbau und Entwicklungsprozess eines Data Warehouse Prototyps zur Nutzung durch Studierende des Lehrstuhls für Wirtschaftsinformatik an der Technischen Universität Freiberg. Anhand des Prototyps sollen grundlegende Schritte zur Bearbeitung eines Datensatzes mithilfe von ETL – Prozessen (Extract, Transform, Load) erläutert werden. Hierfür wird eine ETL – Pipeline gebaut, die den Datensatz vorverarbeitet, transformiert und für die Auswertung nutzbar macht. Die transformierten Daten werden anschließend in die Visualisierungssoftware Tableau Desktop eingebunden und weiter ausgewertet. Die Bearbeitungsschritte werden Schritt für Schritt anhand einer Dokumentation erklärt. Anschließend wird in der Diskussion darauf eingegangen, welchen Limitierungen der Prototyp unterliegt und wie dieser erweitert werden kann. Außerdem wird die Einbindung des Projektes in eine mögliche Online – Lehre diskutiert mit dem Ergebnis, dass durch Ausbau des Datensatzes und entsprechender Vorbereitung der Lehre die Vermittlung von ETL – Prozessen sehr gut behandelt werden kann. Diese Arbeit ist für Studierende als auch für Lehrende im Bereich der asynchronen Online – Lehre interessant und kann als Anreiz für entsprechende Lehrformen verstanden werden.

INHALTSVERZEICHNIS

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

ABKÜRZUNGSVERZEICHNIS

DWH	Data Warehouse
ETL	Extraction, Transformation, Load
PDI	Pentaho Data Integration
OLAP	Online Analytical Processing
ID	Identificator
DIM	Dimension Table
FT	Fact Table
ADAPT	Application Design for Analytical Processing Technologies
UML	Unified Modeling Language
COVID-19	Corona Virus Disease - 2019

1 Einleitung

Das Data Warehouse (DWH) ist heute integraler Bestandteil der „Data-Driven Company“. Diese für Analysezwecke optimierte Datenbank führt Daten aus meist heterogenen operativen Quellen des Unternehmens zusammen. Die daraus gewonnenen Erkenntnisse unterstützen das Management maßgeblich bei Entscheidungsprozessen. Jedoch erfordert die Extraktion, Transformation und das Laden (ETL) der Daten in die eigentliche Datenbank einen erheblichen zeitlichen Aufwand. Entsprechend werden knapp 80% der gesamten Analysearbeit im Data Mining auf die Vorverarbeitung der Daten, das heisst die Beschaffung und Vorbereitung verwendet [Press 2016]. Das Üben dieser Sachverhalte findet in der Praxis meist erst im Beruf statt, da der Privatanwender keinen Zugang zu Testframeworks hat oder diese erst umständlich in Eigenverantwortung gebaut werden müssen. Im Rahmen des Hochschulprojektes der Technischen Universität Freiberg im Fachbereich Wirtschaftsinformatik soll ein Data Warehouse erstellt werden, das dem Studierenden als Handreiche zur Bearbeitung von ETL-Prozessen und dem anschließenden Dashboardbau dienen soll. Hierfür wird prototypisch ein Datensatz der Importe/Exporte Indiens der Jahre 2014 – 2017 transformiert und in ein Tableau-Frontend eingegliedert. Im Folgenden wird der Entwicklungsprozess erläutert.

2 Vorstellung der Werkzeuge

Bei den zum Einsatz kommenden Werkzeugen handelt es sich um Pentaho Data Integration (PDI) als Tool zur Erarbeitung der ETL-Pipelines. Für die Dashboardentwicklung wurde Tableau gewählt. Bei beiden Werkzeugen handelt es sich um eine sogenannte No-Code Software. Daten werden mithilfe einer grafischen Benutzeroberfläche per Drag and Drop miteinander verbunden.

Pentaho ist in seiner Basisversion als Open Source Werkzeug frei verfügbar und wird innerhalb des Lehrstuhls für Wirtschaftsinformatik an der Technischen Universität Freiberg verwendet. Die Software ist vollständig in Java entwickelt worden und deckt die Bereiche ETL, Reporting, OLAP und Data Mining ab. Die Bearbeitung einer Transformation erfolgt über ein Bausteinprinzip, bei dem per Drag and Drop verschieden funktionelle Komponenten in die Transformation gezogen und miteinander verbunden werden.

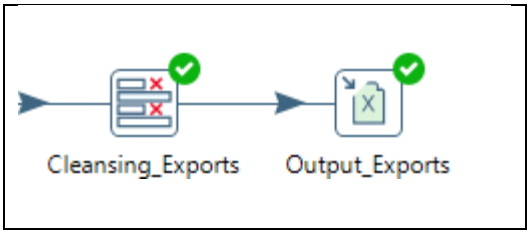


Abbildung 1: Komponentenprinzip Pentaho

Tableau Desktop ist eine Visualisierungssoftware des gleichnamigen Herstellers Tableau Software. Entsprechend liegt der Schwerpunkt des Programmes auf der Datenvisualisierung und dem Reporting bzw. Dashboardbau. Ebenfalls besitzt Tableau diverse Möglichkeiten, um eingegliederte Datensätze zu bearbeiten oder via Joins miteinander zu verknüpfen. Während Tableau Desktop primär für die Erstellung von Visualisierungen und Dashboards konzipiert ist, macht es Tableau Server möglich, Daten in ein bestehendes Data Warehouse zu integrieren, was wiederum die Kollaboration bezüglich Daten und Ergebnissen möglich macht.

3 Projektbeschreibung und Entwicklungsprozess

3.1 Projektbeschreibung

Ziel des Projektes ist die Dokumentation der Entwicklung eines DWH mit angegliedertem Dashboard. Bei den zur Verfügung gestellten Daten handelt es sich um Importe bzw. Exporte von Indien im Zeitraum von 2014 – 2017. Das Framework soll genutzt werden, um Studierenden des Lehrstuhls für Wirtschaftsinformatik ETL-Prozesse und das Erstellen von Dashboards näherzubringen, damit Diese wesentliche Erkenntnisse aus den Daten extrahieren können. Die Teilziele dieses Projektes sind:

- Der **Datensatz** soll entsprechend erweitert bzw. vorverarbeitet werden:
    - Eine zufällige Zuordnung von Monat und Jahr soll bei den einzelnen Transaktionen durchgeführt werden
    - Probleme bezüglich der Datenqualität sollen erhalten bleiben, bis auf jene, welche eine weitere Verarbeitung der Daten unmöglich machen
    - Anreicherung des Datensatzes um weitere Attribute
- Die Dokumentation soll ein **Datenmodell** liefern, das:
    - Zwei Faktentabellen (Import und Export) und drei Dimensionstabellen (Zeit, Region, Produkt) aufweist
    - Die Dimensionstabellen sollen entsprechend erweitert werden (zum Beispiel um Kontinente oder Produktkategorien)
- Nach Absprache mit dem Betreuer soll das **Tableau** Dashboard eine Timeline und eine Übersichtskarte des aufbereiteten Datensatzes beinhalten

Die fertige Arbeit soll entsprechend umfassen:

- Beschreibung und Erklärung des Entwicklungsprozesses
- Installationsanweisungen für die ETL-Prozesse
- Den vorläufigen Prototyp
- Eine Evaluation der Möglichkeiten einer Online-Lehre in diesem Bereich

Im Nachfolgenden wird der Entwicklungsprozess dieser Komponenten im Einzelnen beschrieben und erklärt.

3.2 Entwicklungsprozess

Datensatz

Der übergebene Datensatz besteht aus insgesamt sechs .csv – Dateien. Eine tiefere Recherche zum Datensatz hat ergeben, dass der Erfassungszeitraum der Transaktionen von April bis einschließlich März gewählt wurde [GOV 2020]. Entsprechend handelt es sich bei den Exporten von 2014 – 2015 um Exporte im Zeitraum April 2014 – März 2015.

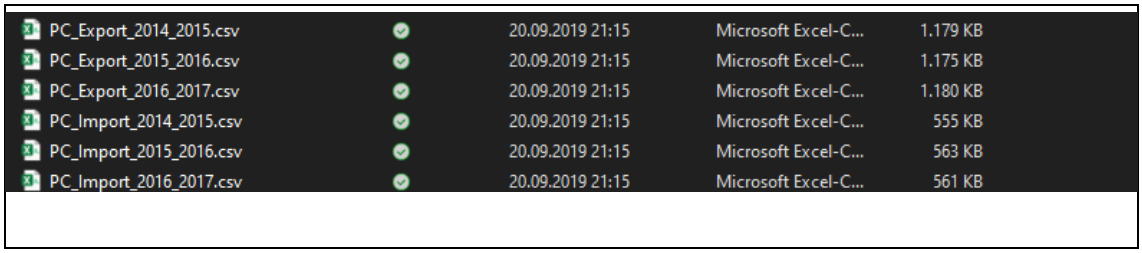


Abbildung 2: Übersicht Datensatz

Tabelle 1: Exemplarischer Aufbau der .csv - Dateien

Name	Beschreibung	Beispiel	Datentyp
pc_code	Produkt ID	A1	String
pc_description	Produktbeschreibung	Tea	String
unit	Mengeneinheit	Kgs	String
country_code	Länder ID	1213	Integer
country_name	Länderbeschreibung	Kenya	String
quantity	Anzahl der Güter	2277547	Integer
value	Umsatz in Mio. USD	4,276366	Number

Tabelle 1 zeigt die inhaltliche Gliederung der jeweiligen Datensätze. Nach anfänglicher Sichtung der Datensätze ist festzustellen, dass es auch unspezifizierte Transaktionen mit der Produkt ID „99“ gibt. Hierbei handelt es sich laut [GOV 2020] um:

- Handelstransaktionen, bei denen das Ursprungsland, die Lieferung, das Zielland nicht spezifiert und ungültige Länder IDs aufweisen
- Reimporte bzw. Reexporte, welche auf den ersten Punkt zurückzuführen sind

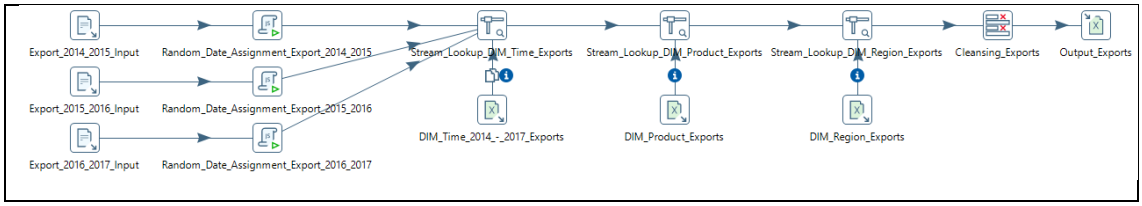


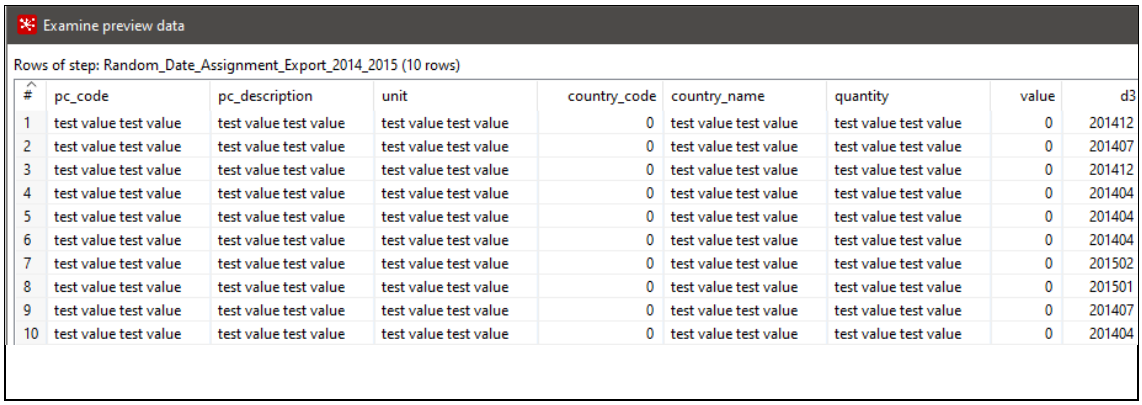
Abbildung 3: Aufbau der ETL - Pipeline für die Exporte

Im Folgenden wird der Aufbau der ETL – Pipeline via Pentaho näher beschrieben. Wie in Abbildung 3 dargestellt, erfolgt die Eingliederung über den „Microsoft Excel Input“ – Step. Die erkannten Datentypen wurden entsprechend der Tabelle 1 angepasst. Anschließend erfolgt die zufällige Zuweisung von Monat und Jahr bei jeder der Transaktionen. Diese wurde mit einem „Modified Javascript value“ – Step realisiert:

```
var d1 = new Date();                        //  Beispiel in Pseudocode:
d1.setDate(1);                              //  d1 = 03.08.2020;
d1.setMonth((Math.random()*12));            //  d1 = 01.03.2020;
var d2 = date2str(d1,"MM");                 //  d2 = "03";
if (str2num(d2) >= 4){                      //  if (d2 >= 4){
      d1.setFullYear(2014);                 //  d1 = 01.03.2014;
}else{                                      //  }else{
      d1.setFullYear(2015);                 //  d1 = 01.03.2015;
}                                           //  }
var d3 = date2str(d1,"yyyy")+d2;            //  d3 = "201503";
```

Abbildung 4: Code-Snippet "Random_Date_Assignment_Export_2014_2015"

Das Script in Abbildung 4 weist der Variable d1 das derzeitige Datum zu. Daraufhin wird der Tag des Datums auf den ersten Tag des Monats gesetzt. Dem Monat wird anschließend eine zufällige Zahl zwischen eins und zwölf zugeordnet. Dieses Datum wird mit der Methode „`date2str()`“ in einen String umgewandelt. Der Monat im Format „MM“ wird entnommen und der Variable d2 zugewiesen. Wenn der zugewiesene Wert von d2 größer gleich vier ist, wird das Jahr von d1 auf 2014, andernfalls auf 2015 gesetzt. Anschließend wird das Datum von d1 in einen String umgewandelt, das Jahr entnommen, mit dem Wert von d2 konkateniert und der Variable d3 zugewiesen.


Examine preview data

Rows of step: Random_Date_Assignment_Export_2014_2015 (10 rows)

#	pc_code	pc_description	unit	country_code	country_name	quantity	value	d3
1	test value test value	test value test value	test value test value	0	test value test value	test value test value	0	201412
2	test value test value	test value test value	test value test value	0	test value test value	test value test value	0	201407
3	test value test value	test value test value	test value test value	0	test value test value	test value test value	0	201412
4	test value test value	test value test value	test value test value	0	test value test value	test value test value	0	201404
5	test value test value	test value test value	test value test value	0	test value test value	test value test value	0	201404
6	test value test value	test value test value	test value test value	0	test value test value	test value test value	0	201404
7	test value test value	test value test value	test value test value	0	test value test value	test value test value	0	201502
8	test value test value	test value test value	test value test value	0	test value test value	test value test value	0	201501
9	test value test value	test value test value	test value test value	0	test value test value	test value test value	0	201407
10	test value test value	test value test value	test value test value	0	test value test value	test value test value	0	201404

Abbildung 5: Script-Test "Random_Date_Assignment_Export_2014_2015

Es entsteht in der Variable d3 ein String in Form „YYYYMM“, welcher als Zuweisungsschlüssel für die Dimensionstabelle Zeit (DIM_Time) genutzt werden kann. Analog hierzu wird bei den Jahren 2015 – 2016 und 2016 – 2017 verfahren. Nach Hinzufügen des zufälligen Datums findet ein Lookup mithilfe von „Stream Lookup“ – Steps statt. Hierfür wurden in Excel drei verschiedene Dimensionstabellen erstellt. Der erste „Stream Lookup“ – Step greift auf die Tabelle „DIM_Time“ zu. Enthalten sind die in Tabelle 2 angegebenen Dimensionen für den Zeitraum 2014 – 2017:

Tabelle 2: Beispielinhalt DIM_Time

Beschreibung	Wert	Datentyp
date	2014-01-01 00:00:00.000	Datum
day_id	20140101	Integer
month_id	201401	Integer
month_description	Jan 14	String
calender_month_id	1	Integer
calender_month_name	Jan	String
calender_month	Januar	String
month_start_id	20140101	Integer
month_end_id	20140131	Integer
quarter_id	20141	Integer
quarter_name	Q1 14	String
quarter_without_year_ID	1	Integer
year_id	2014	Integer

Entsprechend wird mit den Dimensionstabellen für die Produkte (DIM_Product) und Region (DIM_Region) verfahren. Die Inhalte dieser Dimensionstabellen werden in Tabelle 3 und 4 abgebildet.

Tabelle 3: Beispielinhalt DIM_Product

Beschreibung	Wert	Datentyp
pc_code	A1	String
pc_description	Tea	String
pc_category	agriculture_products	String

Tabelle 4: Beispielinhalt DIM_Region

Beschreibung	Wert	Datentyp
region_code	10001	Integer
region_country	Afghanistan	String
region_continent	Asia	String
region_subcontinent	Asia_Central	String

Nach dem Einfügen zusätzlicher Dimensionen wird der entstehende Datensatz über den „Select Values“ – Step sortiert und formatiert. Tabellennamen werden angepasst und vereinheitlicht. Eine anschließende Ausgabe sollte ursprünglich mit dem „Microsoft Excel Output“ – Step realisiert werden. Dieser erzeugt eine binäre XLS-Datei. Die Vorteile solcher binären Daten sind verkürzte Lade- und Speicherzeiten. Leider ließ sich der erzeugte Output der Exporte nicht in Tableau einfügen. Eine Umstellung auf den „Microsoft Excel Writer“ – Step löste das Problem jedoch.

Die abschließende Exceldatei wird bezüglich ihrer Inhalte nicht verändert, um somit mögliche Datenqualitätsprobleme dem Studierenden zu überlassen. Weiterhin findet eine Dimensionserweiterung hinsichtlich Zeit, Produkt und Region statt. Entsprechend analog zu den Exporten fand dieser ETL – Prozess auch für die Importe statt. Details hierzu finden sich in der Vorstellung des Prototyps. Die ausgegebenen Dateien werden dann in Tableau Desktop geöffnet und dort ausgewertet.

Datenmodell

Das Datenmodell wurde noch vor Entwicklung der ETL – Prozesse entwickelt und im Laufe der Entwicklung angepasst. Gefordert sind laut Aufgabenstellung zwei Faktentabellen (Import, Export) und drei Dimensionstabellen (Zeit, Produkt, Region) mit entsprechender Erweiterung der Dimensionen. Abbildung 6 beschreibt das fertige ADAPT Modell. Da sich Importe und Exporte lediglich inhaltlich unterscheiden, wurden diese in einem Cube zusammengefasst. Dieser bezieht Informationen aus den drei Dimensionen Product, Region und Time. In der Dimension Region werden Länder zu Subkontinenten und schließlich zu Kontinenten subsummiert. Zur Aufteilung der Kontinente gibt es keine allgemeingültigen Definitionen oder Richtlinien. Die hier aufgeführte Variante soll daher lediglich als Vorschlag dienen. Die Dimension Zeit hätte entsprechend Aufgabenstellung nur Monat und Jahr enthalten. Zur Vorbereitung möglicher anderer Aufgabenstellungen wurden jedoch auch Tage hinzugefügt. Die Produktdimension beinhaltet die eigentlichen Produkte und eine Kategorisierung dieser in allgemeine Güter (wie etwa Produkte aus der Agrarwirtschaft, weiterverarbeitete Güter). Ursprünglich wurde eine Zwischenkategorisierung eingefügt, jedoch wieder verworfen, weil diese in Teilen redundant war.

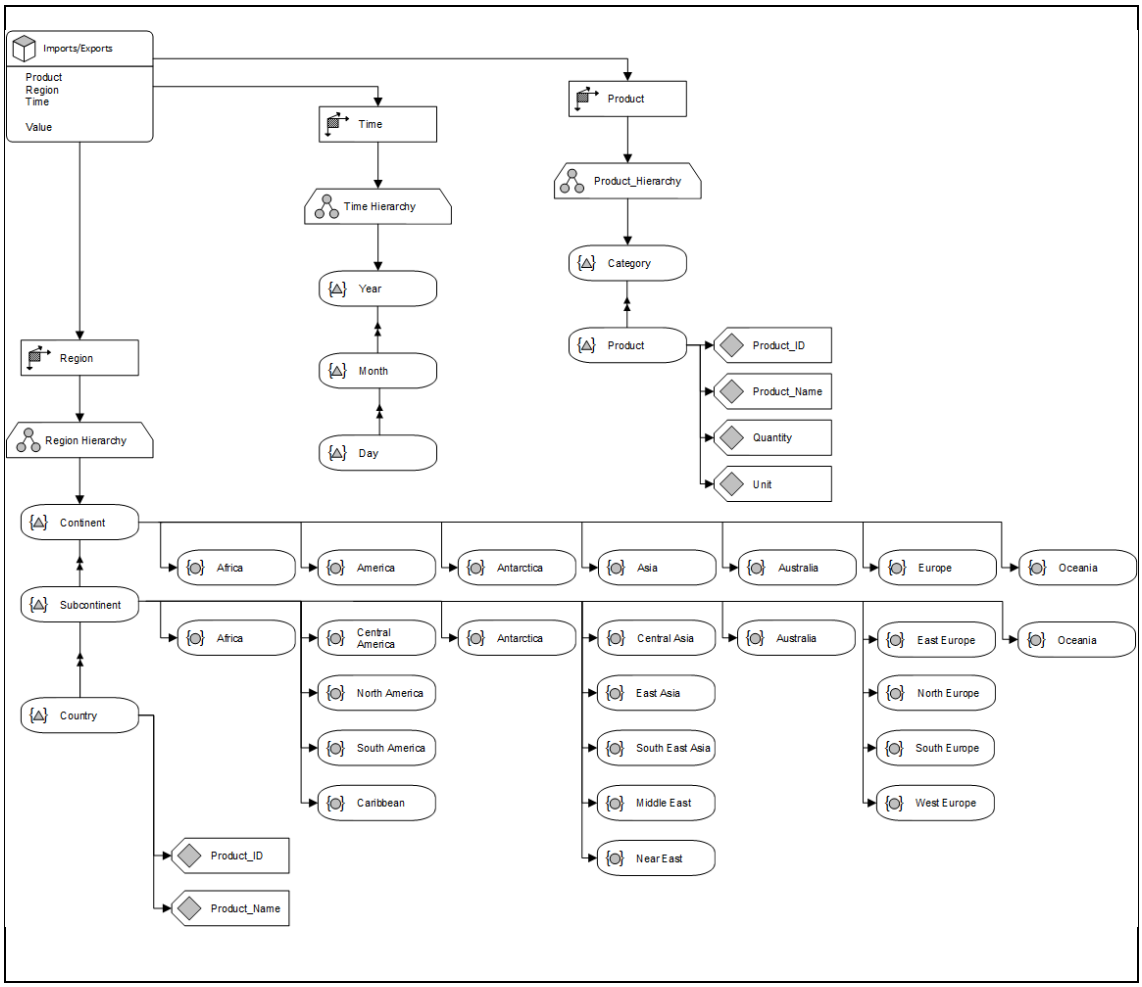


Abbildung 6: ADAPT - Modell

Nach Erstellung des ADAPT Modells wurde entsprechend auch ein physisches Modell für Importe wie Exporte entworfen. Auch diese unterscheiden sich lediglich in der Namensgebung der Faktentabelle und im Inhalt.

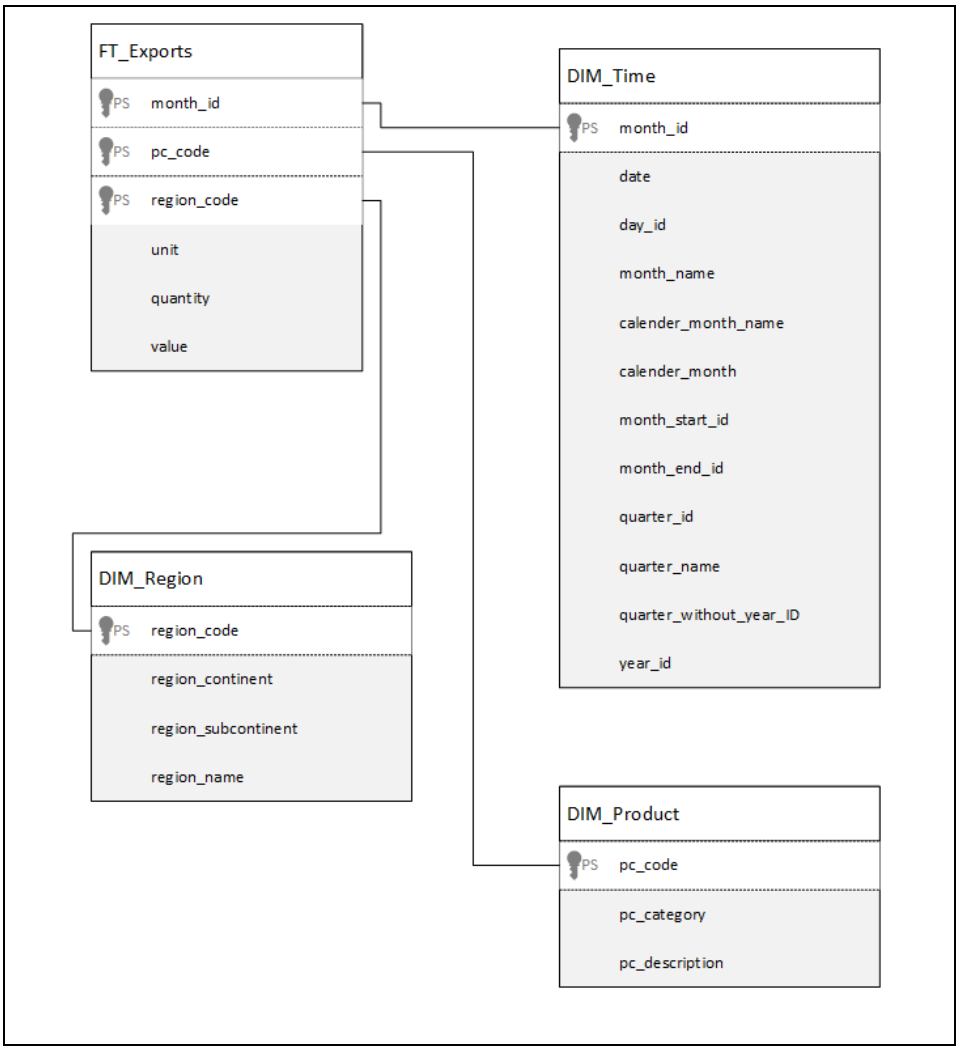


Abbildung 7: UML - Diagramm Exports

Das in Abbildung 7 gezeigte UML Diagramm beschreibt das physische Datenmodell der Exporte. Die Faktentabelle ist in der Lage, durch die Primärschlüssel month_id (DIM_Time), pc_code (DIM_Product), region_code (DIM_Region) auf die jeweiligen Dimensionstabellen zuzugreifen und Information entsprechend des Bedarfs zu extrahieren. Bei Betrachtung der Importe ändert sich in der UML Darstellung lediglich der Name der Faktentabelle.

Tableau

Die Ergebnisdatei des ETL – Prozesses wird in Tableau Desktop eingefügt und weiter bearbeitet. Nach Aufgabenstellung wird nun eine entsprechende Timeline und eine Map

View für den Datensatz entworfen. Beim Öffnen der Excel-Datei erkennt Tableau die meisten Felder bereits korrekt. Es lässt sich jedoch festellen, dass die month_id ungeeignet für eine Datumserstellung ist. Das Datum ist zur Erstellung der Timeline nötig. Entsprechend wird zusätzlich zur month_id das entsprechende Datum in der Dimensionstabelle ermittelt. Beim Einrichten der Datenquelle wird die Spalte date entsprechend auf den Dateityp „Datum“ eingestellt. Die Erstellung der Timeline erfolgt durch Drag and Drop der Kennzahl Value und des Datums. Für die Granularität des Datum wird Monat gewählt.

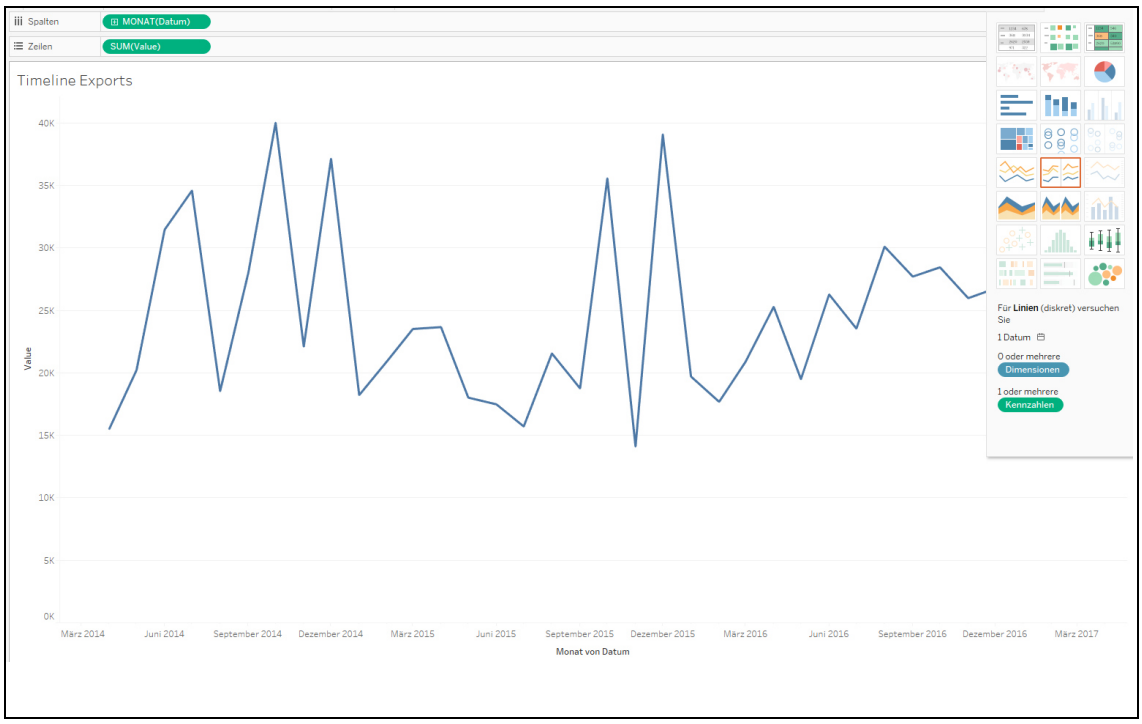


Abbildung 8: Umsätze im Export

Die Erstellung der Map View erfolgt durch Zuweisung einer geografischen Rolle der Spalten region_continent bzw. region_subcontinent. Diese Zuweisung kann aus der Spalte region_name erfolgen. Diese Zuweisung hat zur Folge, dass die in Spalte region_name aufgeführten Länder via Tableau semantisch entsprechend der in Spalte region_continent bzw. region_subcontinent aufgeführten Namen untergruppiert werden. Tableau ordnet zur Erkennung der Länder die Strings der Spalte region_name voreingestellten Ländern zu. Aufgrund undeutlicher Namensabkürzungen in der Spalte region_name können jedoch nicht alle Länder zweifelsfrei zugeordnet werden. Hier muss eine manuelle Anpassung der Zuordnung der Länder erfolgen. Abbildung 9 zeigt die fertige Zuordnung der einzelnen Länder zum entsprechenden Subkontinent. Die Farbgebung

Gelb – Grün zeigt die absoluten Umsätze im Export pro Subkontinent, wobei grün die höchsten Umsätze und gelb die niedrigeren Umsätze darstellen soll. Des Weiteren kann der Eintrag „Unspecified“ nicht zugeordnet werden und wird entsprechend nicht mitberücksichtigt. Hierbei handelt es sich um die oben angegebenen Reexporte bzw. nicht zuordenbaren Transaktionen. Obwohl die Texterkennung von Tableau im Erkennen der Länder sehr fortgeschritten ist, kann es hier auch zu Fehlklassifizierungen kommen. Eine Prüfung der automatisch zugeordneten Länder sollte dennoch erfolgen.

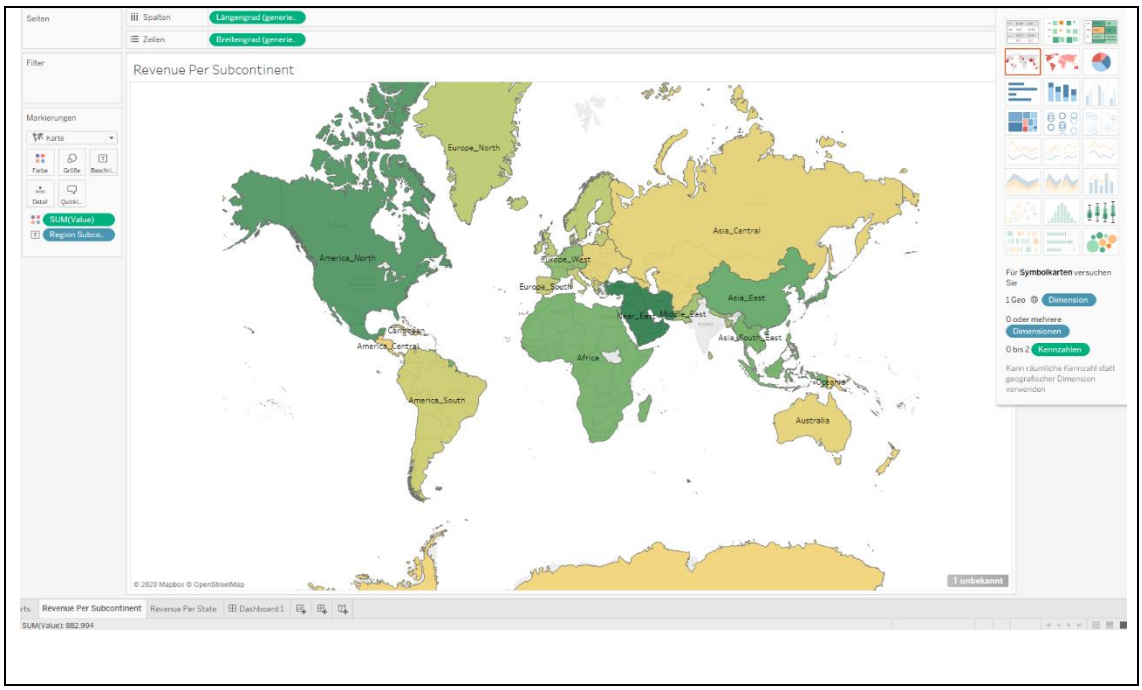


Abbildung 9: Umsätze pro Subkontinent

4 Prototyp

Der Aufbau des Prototyps im Einzelnen wurde in **3 Projektbeschreibung und Entwicklungsprozess** erklärt. Dieses Kapitel beschreibt den fertigen Prototyp in seiner Gesamtheit.

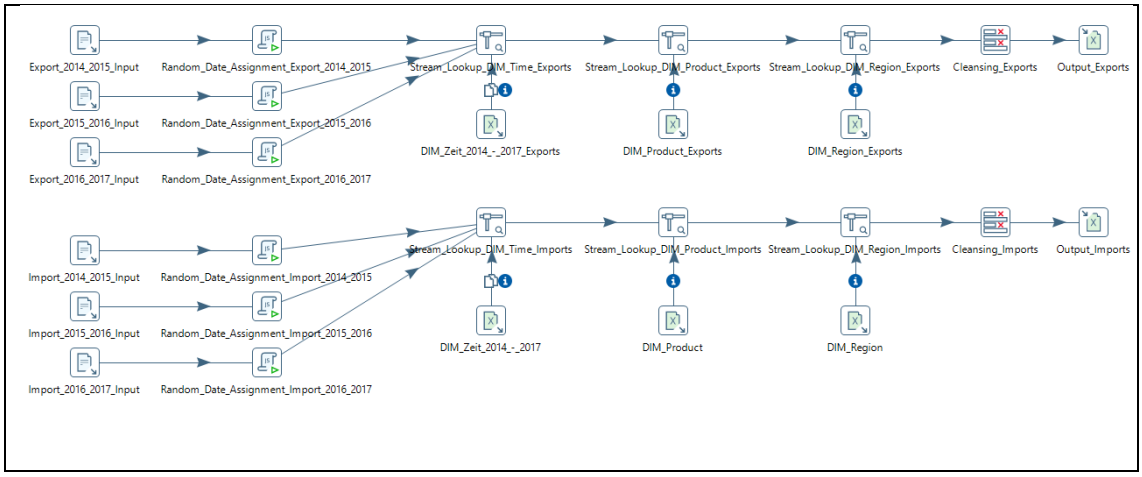


Abbildung 10: Kompletter Kettle-Prozess für Imports/Exports

Während im Entwicklungsprozess lediglich der Kettle – Prozess für die Exporte beleuchtet wird, zeigt Abbildung 10 den komplettierten Prozess. Da die Datensätze der Importe und Exporte gleich aufgebaut sind, ist der ETL – Prozess bei den Importen dem der Exporte sehr ähnlich. Der allgemeine Ablauf ist wie folgt:

- Laden der .csv – Dateien und Zuordnung der Datentypen
- Hinzufügen eines zufälligen Datums an jeder Transaktion im Format „yyyyMM“ als month_id
- Anreichern des Datensatzes durch Lookup der entsprechenden Dimensionstabellen:
    - Lookup der Dimensionstabelle DIM_Zeit über die vorher festgelegte month_id
    - Lookup der Dimensionstabelle DIM_Product über pc_code
    - Lookup der Dimensionstabelle DIM_Region über region_code
- Cleansing des angereicherten Datensatzes durch Selektion der erforderlichen Spalten
- Ausgabe des fertigen Datensatzes als Excel – Datei

Die ausgegebene Excel – Datei ist Grundlage der Datenanalyse via Tableau und wird dort entsprechend weiterverarbeitet. Die Excel – Datei weist folgendes Schema auf:

Tabelle 5: Tabellenschema Exports/Imports Ausgabedatei

Beschreibung	**Wert**	**Datentyp**
pc_code	A1	String
pc_description	Tea	String
pc_category	agriculture_products	String
region_code	10001	Integer
region_name	Afghanistan	String
region_continent	Asia	String
region_subcontinent	Asia_Central	String
unit	Kgs	String
quantity	2277547	Number
value	4,28	Number
date	2014-07-31 00:00:00.000	Date

Entsprechend der eingangs aufgeführten Teilziele kann zusammenfassend gesagt werden, dass der **Datensatz** vorverarbeitet und erweitert worden ist. Es findet im Laufe des ETL – Prozesses eine zufällige Zuordnung von Monat und Jahr statt. Probleme bezüglich der Datenqualität bleiben erhalten, da über die bereits vorhanden Schlüssel pc_code, region_code bzw. month_id auf Dimensionstabellen zugegriffen wird. Diese sind bereits von Haus aus eindeutig. Eine Anpassung von Beschreibungen oder Werten fand innerhalb des Datensatzes nicht statt.

Die Dokumentation liefert ein semantisches und ein physisches **Datenmodell** in Form einer ADAPT – Modellierung bzw. eines UML – Diagrammes zur Beschreibung des Datensatzes. Entsprechend wurde das Datenmodell so angepasst, dass mit zwei Faktentabellen (FT_Import, FT_Export) sowie mit drei Dimensionstabellen (DIM_Time, DIM_Product, DIM_Region) gearbeitet wird. Die Dimensionstabellen wurden entsprechend der Vorgaben erweitert und können bei Bedarf skaliert werden.

Der bearbeitete Datensatz wurde in **Tableau** Desktop weiterverarbeitet und angepasst, um die geforderte Timeline und eine exemplarische Übersichtskarte zu erstellen. Die erforderlichen Deliverables werden also umfassen:

- Den ETL – Prozess als Kettle-Datei (.ktr) zur Bearbeitung in PDI
- Die benötigten Dimensionstabellen in Form von Excel – Dateien
- Eine PDF – Datei als Anlage mit dem ADAPT – Modell und den UML – Diagrammen. Des Weiteren die entsprechenden MSVisio – Dateien
- Die Tableau – Projekte für Importe und Exporte
- Eine Schritt-für-Schritt Anleitung zur Erstellung der ETL – Pipeline

5 Diskussion

Das durchgeführte Projekt bietet eine solide Grundlage zur Einführung in den Bereich der ETL – Prozesse und der Grundlagen des Data Warehousings. Außerdem ist der Studierende in der Lage, den Datensatz über Tableau Desktop zu visualisieren. Sowohl der Dashboardbau als auch die Entwicklung eigener ETL – Pipelines sind wichtige Grundkenntnisse zukünftiger Business Analysten. Bei der Erstellung wurde darauf geachtet, dass möglichst skalierend gearbeitet werden kann. Die einzelnen Teilschritte sind so konzipiert, dass problemlos weitere Export- bzw. Import Dateien angefügt werden und verarbeitet werden können. Im Grunde muss nur die Anpassung der zufälligen Datumszuteilung erfolgen. Außerdem ist die Dimensionstabelle DIM_Time nur von 2014 – 2017 gepflegt.

Limitierungen

Angefangen bei der Vorverarbeitung des Datensatzes via PDI ist fraglich, ob es nicht noch andere Wege gibt, den Transaktionen ein zufälliges Datum zuzuordnen. Nicht jeder Studierende kennt sich im Scripting aus. Auch, wenn PDI hierzu durch Beispielfunktionen Hilfestellung bietet, kann dies ein potentieller Stolperstein werden. Eine mögliche Erweiterung des Datensatzes um zufällige Transaktionstage (nicht nur Monat und Jahr), kann kann nur schwerlich über zufällige Zuordnungen von 1 – 31 erfüllt werden. Nicht jeder Monat hat 31 Tage (Auch Schaltjahre müssen berücksichtigt werden). Auf diesen Sachverhalt müsste beim Scripting ebenfalls geachtet werden. Auch die Einbindung der Stream Lookups auf mögliche Dimensionstabellen stellt nur eine Möglichkeit der Bearbeitung dieses Projektes dar. So können beispielsweise Value Mapper bei entsprechender Vorsortierung der Daten eingesetzt werden, um zusätzliche Spalten zur Dimensionserweiterung zu ergänzen. Da der Fokus des Projektes auf der Erstellung der ETL – Pipeline liegt, wurden die Tableau Visualisierungen weniger umfangreich untersucht. Festzustellen war, dass in der Einbindung der Datensätze noch einmal nachgearbeitet werden musste bzw. Anpassungen an den Dimensionen vorgenommen werden müssen.

Das Projekt wurde von einer einzelnen Person bearbeitet. Entsprechend ist der Aufbau der ETL-Pipeline nach den subjektiven Präferenzen des Autors geschehen. Trotz mehrmaligem Prüfen können noch immer Daten fehlerhaft sein. Auch ist es durchaus möglich, dass sich im Laufe der zukünftigen Bearbeitung günstigere Prozesse ergeben,

mithilfe derer ein zufriedenstellendes Ergebnis erreicht werden kann. Die für die Entwicklung verwendete Version von Pentaho Data Integration ist v8.1. Diese Version ist von 2018. Es kann durchaus sein, dass im Laufe der Bearbeitung mit neueren Versionen nicht nur mehr Funktionalität hinzugekommen ist, sondern auch Prozesse angepasst werden müssen. Dies wurde nicht überprüft.

Möglichkeiten der Erweiterung

Der Datensatz besitzt keine größere Erkenntnistiefe. Hier gäbe es die Möglichkeit, ein künstliches Szenario zu schaffen, um diese Erkenntnistiefe auszuarbeiten. Die Sinnhaftigkeit der Bearbeitung dieser Datensätze schafft bei Studierenden zusätzliche Motivation und könnte zur Leistungssteigerung beitragen. Denkbar wäre eine Storyline, anhand derer sich der Student besser in die Bearbeitung bzw. Situation hineindenken kann. Zusätzlich können Bearbeitungsmeilensteine eingefügt werden. Diese Bearbeitungsmeilensteine könnten dann abgeprüft werden, um so den Fortschritt des Studierenden zu dokumentieren oder um mögliche Verbesserungen im Gesamtaufbau der Aufgabe vornehmen zu können.

In Zeiten von COVID-19 haben sich hinsichtlich des Zusammenarbeitens vor Ort deutliche Einschränkungen ergeben. Allerdings zeigen inzwischen viele Unternehmen und Lehreinrichtungen, dass es Mittel und Wege gibt, im Homeoffice produktiv zu bleiben. Im Falle einer synchronen Online-Lehre bietet sich die Zusammenarbeit über Plattformen wie Zoom, Slack, Discord etc. an. In der Praxis gestaltet sich die Arbeit mit Screen-Sharing Tools jedoch als schwierig. Der Dozent kann nur durch Kontrollfragen kontrollieren, ob Studierende folgen können. Lernende, die dem Unterricht so nur schwer folgen können, werden damit unnötigerweise abgehängt. Eine weitere Überlegung ist die Möglichkeit, das komplette Projekt in Form einer asynchronen Online-Lehre zu vermitteln. Die asynchrone Online-Lehre hat den Vorteil, dass sich Studierende mit eigenem Lerntempo dem Sachverhalt nähern und schlussendlich nur das Ergebnis der Leistung zählt. Hierfür könnte ein größeres Projekt in Teilzielen selbstständig von Studierenden innerhalb eines festgelegten Zeitabschnittes bearbeitet werden. Die Kontrolle der Arbeitsergebnisse erfolgt dann über ein wöchentliches/monatliches Meeting mit der lehrenden Person. Zur Bearbeitung können konkrete Anleitungen als Hilfestellung mitgegeben werden oder der Dozent richtet eine virtuelle Maschine (VM) zur Bearbeitung ein, die auf

den zu bearbeitenden Sachverhalt ausgelegt ist. Der Vorteil hierfür wäre die Bereitstellung einer einheitlichen Datenbasis und Entwicklungsumgebung für alle Studierenden. Im Idealfall beläuft sich der Installationsaufwand für den Studierenden lediglich auf die Virtualisierungsumgebung.

Das zu diesem Bericht mitgegebene Material darf gern als Anreiz für mögliche Lösungsszenarien angesehen werden. Es wurde darauf geachtet, Schritte sehr feingliedrig zu beschreiben, damit auch Beginner einen schnellen Einstieg finden. Dennoch wird in der Dokumentation nicht erklärt, warum die Schritte so ausgeführt werden, wie sie ausgeführt werden. Diesbezüglich könnte die Dokumentation durchaus noch erweitert werden.

6 Fazit

Das Projekt stellt eine gute Möglichkeit zur Übung wichtiger Grundkenntnisse im Bereich der Business Analytics dar. Der Studierende bekommt die Möglichkeit, die Anreicherung und Transformation der zu verarbeitenden Daten im Data Warehouse mitverfolgen zu können. Diese Arbeit beschreibt den Entwicklungsprozess eines Data Warehouse Prototyps in dem ETL-Prozesse und das Erstellen von Tableau Visualisierungen umgesetzt werden. Der Prototyp verarbeitet die Importe und Exporte Indiens aus den Jahren 2014 – 2017 und wurde so konzipiert, dass dieser möglichst skalierbar für Folgeprojekte als Grundlage genutzt werden kann. Wichtig ist, dass dieses Projekt sich noch immer in der Entwicklung befindet. Die vorgestellten Prozesse sind dabei als richtungsweisend und nicht als allgemeingültig zu verstehen. Dennoch kann dem Studierenden so die Möglichkeit geboten werden, ETL – Prozesse anhand praktischer Beispiele zu erfahren, um Vorgänge im Data Warehouse besser verstehen und anwenden zu können.

LITERATURVERZEICHNIS

[Press 2016] Press, Gil. Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/. Aufgerufen am: 22.07.2020

[GOV 2020] GOVERNMENT OF INDIAMinistry of Commerce. https://commerce-app.gov.in/eidb/default.asp. Aufgerufen am: 22.07.2020

Eidesstattliche Versicherung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig und ohne unerlaubte Hilfe angefertigt und andere als die in der Arbeit angegebenen Hilfsmittel nicht benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus anderen Schriften entnommen sind, habe ich als solche kenntlich gemacht.

Freiberg, 05.08.2020 Unterschrift: