

Datensammlung und Datenstruktur

Dokumentation Big Data Projekt

des Studienganges Informatik an der Dualen Hochschule Baden-Württemberg Stuttgart

von

Luca Holder

22.11.2025

Bearbeitungszeitraum 04.11.2024-22.11.2024

Matrikelnummern, Kurs 5234642, TINF22C

Gutachter*in der Dualen Hochschule Marcel Mittelstädt

Inhaltsverzeichnis

Inhaltsverzeichnis Abbildungsverzeichnis 1 Datensammlung		1 1 2			
			1.1	Datenstruktur	2
			1.2	Datenabfrage	3
1.3	Automatisierung der Datensammlung	4			
Literaturverzeichnis		6			
Abbi	Idungsverzeichnis				
Abbildung 1: eigene Abbildung, Datenstruktur der Karten		3			
Abbildung 2: Screenshot, Konsolenausgabe der Ausführungszeiten		4			

1 Datensammlung

Dieses Kapitel beschreibt die Herangehensweise der Datensammlung.

Die benötigten Daten lassen sich mit der API "Magic: The Gathering API" abrufen, die Informationen über die Karten und Sets von dem Fantasy-Kartenspiel "Magic: The Gathering". (vgl. [1])

1.1 Datenstruktur

Die Daten sind nach dem collection_job_mtg im JSON-Format verfügbar. Sie liegen somit als Sem-strukturierte Daten vor, da das JSON-Format zwar Struktur bietet, nicht aber auf der Ebene von strukturierten Daten wie Tabellen. Die genaue Struktur ist in der folgenden Abbildung sichtbar.

Abbildung 1: eigene Abbildung, Datenstruktur der Karten

1.2 Datenabfrage

Die Dokumentation der API weist auf SDKs hin, die die Datenabfragen vereinfachen sollen. Zudem ist die Datenabfrage über die API auf maximale 5000 Abfragen pro Stunde sowie auf maximal 100 Anfragen pro 100 Karten eingeschränkt wird. (vgl. [1]) Aus diesem Grund wird eine SDK zur Abfrage verwendet.

Erstellt wird die Abfrage in Python, dabei wird das Python SDK verwendet, dass mit dem Befehl "pip install mtgsdk" installiert werden kann.

```
All card data has been loaded in 384.38 seconds.

Time since last heartbeat(0.03 s) < heartrate(5.0 s), sleeping for 4.972679 s

Time since last heartbeat(0.05 s) < heartrate(5.0 s), sleeping for 4.952313 s

Time since last heartbeat(0.03 s) < heartrate(5.0 s), sleeping for 4.974349 s

Time since last heartbeat(0.02 s) < heartrate(5.0 s), sleeping for 4.984048 s

Time since last heartbeat(0.03 s) < heartrate(5.0 s), sleeping for 4.967826 s

Time since last heartbeat(0.02 s) < heartrate(5.0 s), sleeping for 4.98203 s

All card data has been saved as JSON to /home/airflow/raw_data/mtg_data.json in 29.71 seconds.

Time since last heartbeat(0.03 s) < heartrate(5.0 s), sleeping for 4.972811 s

Deleted /user/hadoop/data/raw/mtg_data.json

Old data deleted from HDFS raw directory.

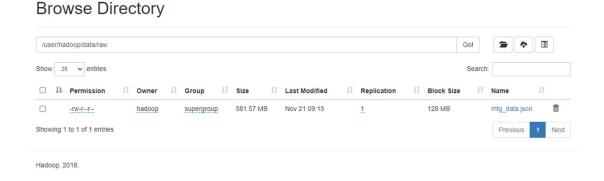
Time since last heartbeat(0.03 s) < heartrate(5.0 s), sleeping for 4.972612 s

Data successfully uploaded to HDFS at /user/hadoop/data/raw/mtg_data.json.

All card data has been saved as JSON to HDFS in 6.81 seconds.
```

Abbildung 2: Screenshot, Konsolenausgabe der Ausführungszeiten

Bei der Abfrage aller Karten (durchgeführt am 21. November 2024 auf einer Google VM \rightarrow n2-standard-2) betrug die Ausführungszeit für die Abfrage aller Karten 384,38 Sekunden, das ungefähr 6min und 40 Sekunden entspricht. Das Überführen der Daten in JSON dauerte ca. 30 Sekunden. Die Speicherung der JSON-Datei in den JDFS dauerte ca. 7 Sekunden.



Die Datei hatte am Ende eine Größe von Size: 581.57 MB und 93.643 Karten.

1.3 Automatisierung der Datensammlung

Die Datensammlung wird mit einem Python-Skript durchgeführt. Dieser Job läuft unter dem Namen collect_job_mtg und ruft alle Karten der API über die Python SDK auf und speichert diese als JSON-Datei ab. Anschließend wird über das Skript die Datei in den HDFS in das Verzeichnis "raw" als JSON-Datei gespeichert, wie die Prüfungsleistung erfordert.

Literaturverzeichnis

[1] "MTG API Docs", Magic: The Gathering API Documentation. Zugegriffen: 5. November 2024. [Online]. Verfügbar unter: https://docs.magicthegathering.io/