Documentation

ETL Workflow

Zuerst werden die Daten von der API geholt. Im Anschluss werden in Hadoop die Raw und Final Ordner erzeugt. Im Anschluss sollten die JSON-Daten optimiert werden. Dies geschieht indem ein MapReduce-Job angewandt wird. Im Anschluss wird aus der JSON-Datei eine Hive-Tabelle erstellt sowie die Finale SQL-Tabelle. In der finalen SQL-Tabelle werden dann noch die doppelten Karten entfernt. Allerdings funktioniert (zumindest auf meiner VM) schon das Erstellen der Ordner nicht mehr (Fehler: hadoop not found). Außerdem gibt es Probleme dem MapReduce-Job. So funktionieren auch die Folge-Jobs nicht wirklich.

List of Jobs/Transformations

Jobs:

- 01_download_flow
- 02_hadoop_flow
- 03_create_tables
- Project_flow

Job/Transformation Descriptions

01: Holen der Daten von der MTG-API und abspeichern im Zielverzeichnis

02: Falls nicht vorhanden, werden die Raw- und Final-Ordner in Hadoop erstellt, sowie eine Überprüfung ob die Roh-Daten vorhanden sind

03: Anlegend er Hive-Tabelle sowie der SQL-Tabelle

External Tools

Hadoop, Hive

SerDe-Jar-File wurde eingebunden, um die JSON-Datei automatisch in die Tabelle einzufügen