Firmendaten: Kundendaten, Itemdaten, Orderlines, ExchangeRates

Preprocessing (Notebook 1)

- Read CSV (Zeit messen)
- Daten zeigen & verstehen
- Data Cleaning (Daten vor 05.2019 löschen)
- Save in Parquet

Spark statistics (Notebook 2)

- Messungen und vergleiche zwischen CSV-Parquet
- Zeit zum Laden
- CPUS, usw…
- Abhängigkeit von Anzahl CPUs

—> Ordner (Ursprüngliche Daten & Bearbeitete Daten (CSV+Parquet)

Analysis - Teil 1 (Notebook 3)

- Read Parquet & weiterarbeiten mit dem
- Analysis 1: Umsatz in für CHF-Kunden im Jahr 2020 (Join orders x customers)
- Analysis 2: Umsatz in für alle Kunden im Jahr 2020 (Berücksichtigung der Währung)
- Analysis 3: Umsatz in für alle Kunden im Jahr 2020 pro Item-Gruppe
- Analyse 4: Analyse 3 für alle im Datensatz vorhandenen Jahre
  (**Bar charts**)

Analysis - Teil 2 (Notebook 4)

Analyse 5 (**go beyond some aspects**): Vorhersage

- Filter 2019-2021
- Training
- Vorhersage für 2022 (pro Item Gruppe)

Zusammenfassung

- Notebook
- PPT (08.07.2024)