Numpy / Pandas

Im Schnelldurchlauf

Numpy

Geschichte

- 2005 werden Numarray und Numeric von Travis Oliphant zu Numpy zusammengeführt

- Numpy ist in C geschrieben, basiert aber auf BLAS/ LAPACK wie Matlab

- Ist Teil von Scipy

np.array

- Man könnte das Ding auch Tensor nennen
- Beliebig viele Achsen
- Ein Index pro Achse

Vorteile gegenüber reinem Python

- Schnell
- Speicherschonend - int in einer Liste hat in python 24 Bytes
- Mächtigere Indizierungsmöglichkeiten

Nachteile

- Fixe vorgegebene Größe (workaround mit array Modul möglich)

- Alle Elemente eines numpy arrays müssen den gleichen Typ haben

Pandas

Geschichte

- Entwickelt 2008 von Wes McKinney
- Inspiriert von DataFrames in r
- Basiert auf Numpy
- Excel-Sheet ohne GUI

Warum noch einen Layer über Numpy?

- In Pandas DataFrames können Spalten unterschiedliche Typen haben
- Spalten haben Namen
- Unterschiedliche Index-Typen (Datetime, Kategorien..)
- Zeitreihen, gruppieren von Zeilen, etc..

„ pandas rule of thumb: have 5 to 10 times as much RAM as the size of your dataset “

–Wes McKinney

Probleme

10 Things I Hate About Pandas

1. Pandas ist intern nicht maschinennah genug
2. Keine Unterstützung für memory mapped datasets (numpy kann das)
3. Schlechte Performance beim Datenbank-Import/Export
4. „Missing Data“ nur unzureichend unterstützt (hängt an numpy, kein np.nan für Integer)
5. Intransparent was Hauptspeicherverbrauch/management angeht
6. Schlechte Unterstützung kategorialer Datentypen (ohja :(..)
7. Bei komplexen Gruppierungen umständlich und langsam
8. Man kann nicht wirklich Daten zu DataFrames hinzufügen (numpy)
9. Beschränkte und nicht erweiterbare Unterstützung für Type-Metadaten (stimmt nicht mehr ganz)
10. Kein query-planning
11. Schlechter multicore support für größere Datenmengen (dask)

Ausblick

- Pandas2

- Arrow