Assignment 1: Big Data in Ihrem Umfeld

1.1

Schemalose Daten:

Unstrukturierte Daten im CIS z.B. Bilder, PDFs, Videos, Dokumente, etc.

Strukturierte Daten:

Schematische Daten sind z.B. die Studentendaten (Vorname, Nachname, Email-Adresse, Semester, Studiengang, Noten, etc.), Lehrveranstaltungen (Lektor, Zeiten, Studiengang, Raum), etc.

1.2

Stream:

Je nachdem welche Informationen gerade benötigt werden, verarbeitet man die dafür notwendigen Daten in Echtzeit, z.B. CIS-Login für Studentinnen, Abrufen der Leistungsbeurteilung, Abrufen der aktuellen Wochenplanung, etc.

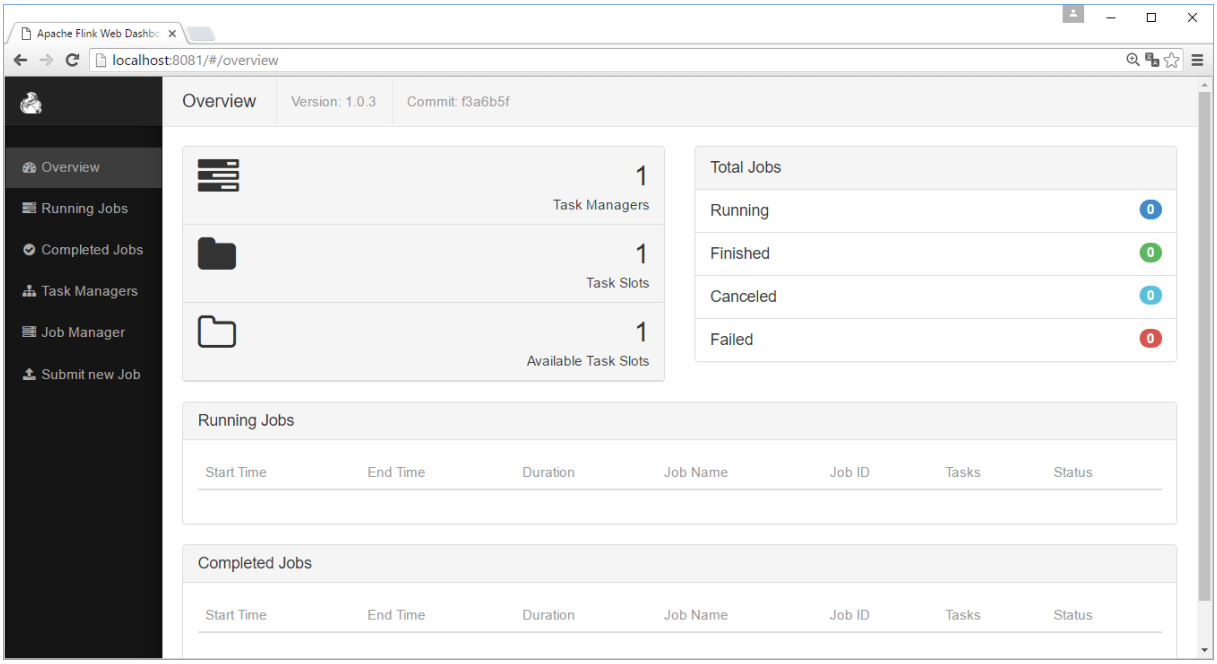
Batch:

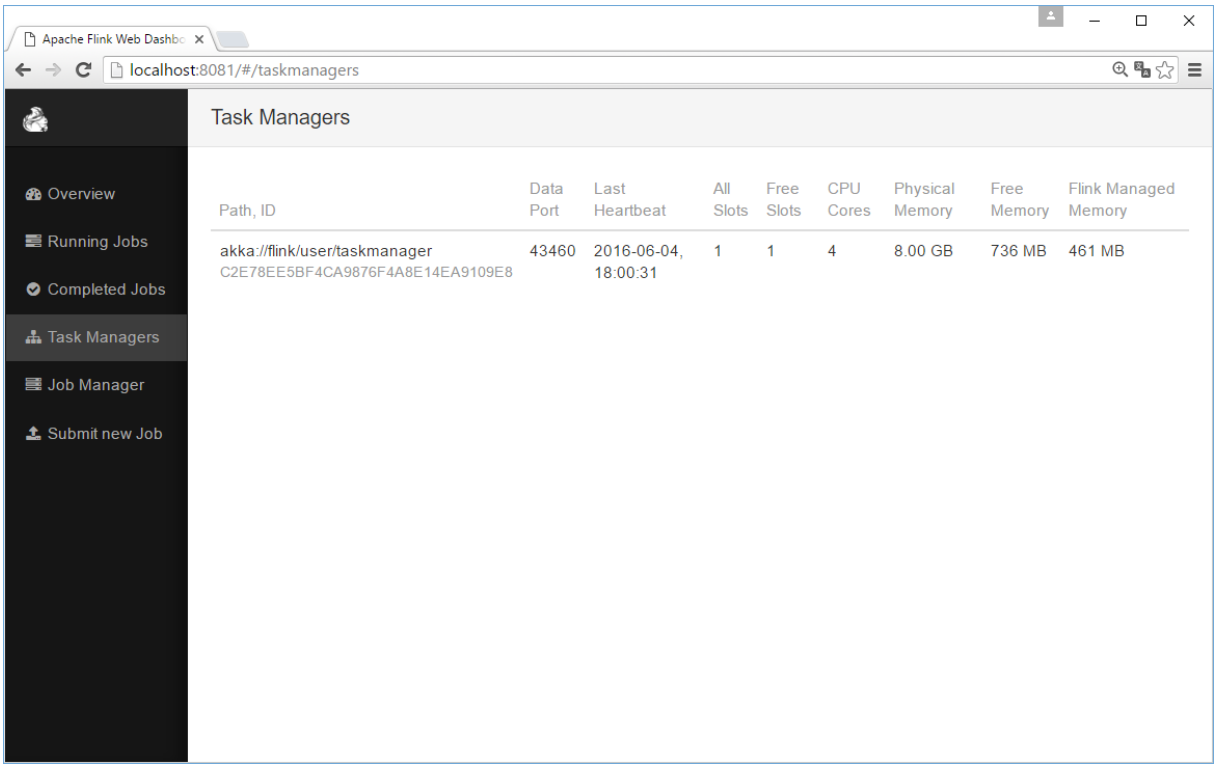
Bei Batch hingegen werden z.B. für eine Analyse größere Daten auf einmal abgerufen, z.B. Anzahl der Studenten-Logins in dieser Woche, Anzahl Besucher Master-Studiengang, Anzahl Besucher Bachelor-Studiengang über 22 Jahre, etc. Dieser Prozess wird in der Regel im Intervall (z.B. wöchentlich) ausgeführt, um die Daten miteinander vergleichen zu können.

Assignment 2: Big Data in Ihrem Umfeld

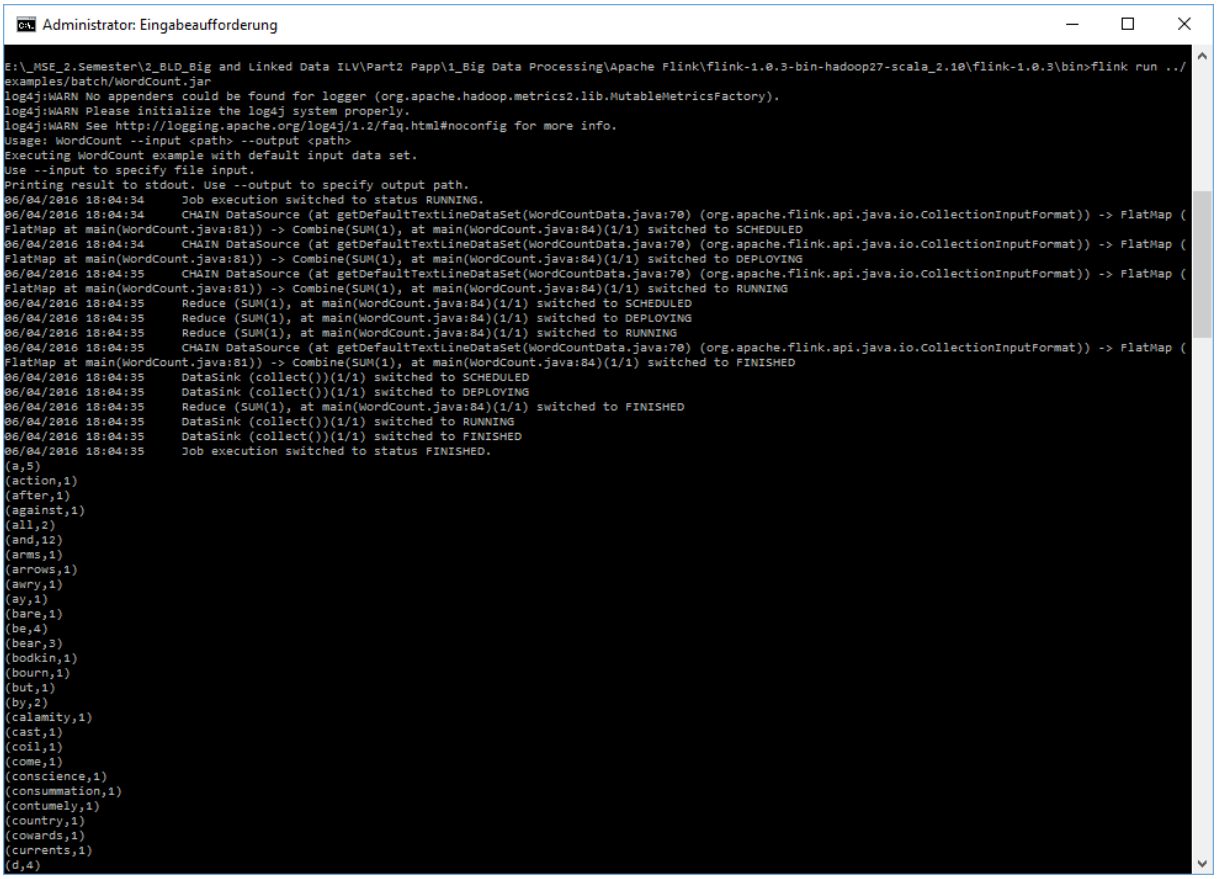
2.1

- Ich habe Apache Flink gewählt, weil es Open Source ist, zur neueren Generation gehört und wir dafür bereits eine Anleitung bekommen haben. Außerdem bietet Flink neben Streaming auch batch data processing an.
- Screenshots:





WordCounter Beispiel:



- Ich würde für Apache Flink die IDE IntelliJ verwenden, da ich mit der IDE einigermaßen Vertraut bin und es viele Dokumentationen dazu bereits im Internet zu finden sind. Flink ist auch einfach in IntelliJ integrierbar.

Assignment 3: Big Data in Ihrem Umfeld

Link zum Projekt:

https://github.com/EmreB51/BLDExercise

Data Science

Assignment 1: Technologien

1.1

Neben Python und R gibt es noch Jupyter, Appache Zeppelin sowie Beaker, Julia, etc.

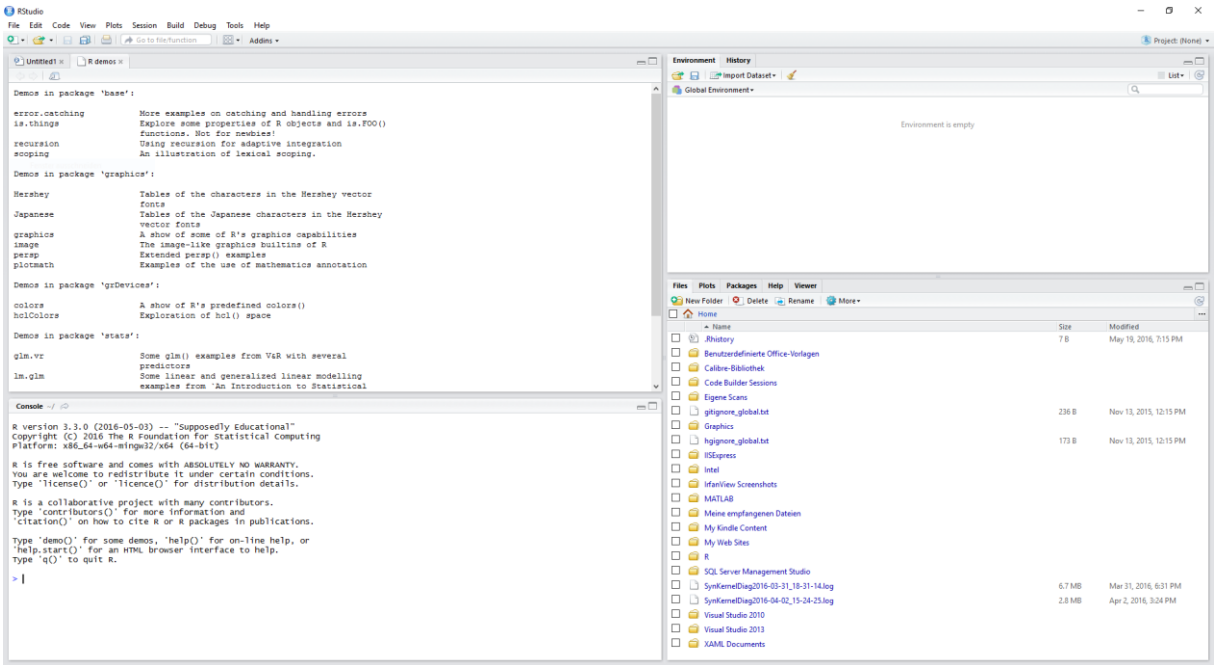
1.2

Ich würde zu R tendieren, da bei R auch die eigene Entwicklungsumgebung RStudio dabei ist. Auf den Ersten Blick scheint R Benutzerfreundlicher zu sein, und ist laut Recherchen auch beliebter bei dem Data Scientist. Es ist kostenlos, wird laufend weiter entwickelt und scheint für Data Science optimierter als Python zu sein.

…

Assignment 2: Technologien

- Wie bereits bei Punkt 1.2 beschrieben habe, würde ich mich eher für R entscheiden. Ich habe mir beide Technologien installiert und finde auch nach der Installation, dass R mir eher gefällt. Es gibt bei R mehrere Ansichten, Menü Punkte, mehr Beschreibung, die mir bei Python letztendlich fehlen. R bietet mehr Pakete und Bibliotheken für professionelle Data Science an.
- Screenshot:



- Da R bereits eine eigene IDE anbietet, kann man direkt mit RStudio arbeiten.

Assignment 3: Big Science

- Classification: …Daten nach Eigenschaften trennen, z.B. Gewicht, Größe, Alter, etc.

- Regression … wird verwendet, um die Stärke und Richtung von Beziehungen zwischen Variablen zu schätzen, die linear miteinander verbunden sind.
  Abhängigkeit von x und y z.B. y = 0.25x+20, y ist 0.25 Mal x +20 groß

- Clusting .. Struktur von gemessenen Werten durch Beobachtung erkennen, damit eine Gruppierung der Daten erfolgen kann, z.B. Besucher von Amazon nach Herkunft

- Dimensional reduction … Konvertieren von Datensätzen mit mehreren Dimensionen in Datensätze mit weniger Dimensionen, z.B. 2 Dimensionale Daten in ein 1-Dimensionale Daten konvertieren (Bildverarbeitung)

Beispiel:

Amazon: Bei der Produkte Suche oder vor/nach einem Kauf werden weitere Produkte, die miteinander Verknüpft sind oder auch gekauft bestellt werden, von Amazon empfohlen.