Data Engineering

Assignment 1: Big Data in Ihrem Umfeld (4 Punkte)

1.1 (2 Punkte)

Schauen Sie sich in Ihrem Umfeld um. FH Technikum oder Ihr Job. Nennen Sie mindestens ein Beispiel für Daten, die schemalos (unstrukturiert) sind und mindestens ein Bespiel für Daten, die strukturiert (schematisch) sind.

Unstrukturierte Daten (Schemalose Daten) sind zum Beispiel Nachrichten die mit einem Messenger versendet werden oder Emails. In dem Unternehmen in dem ich tätig bin, ist es auch wichtig diverse Emailverläufe zu sichern und gegebenenfalls wiederzufinden. Aber die Mails werden einfach irgendwo irgendwie abgelegt ohne jegliche Struktur.

Strukturierte Daten sind zum Beispiel Userdaten die in einer Datenbank verwaltet werden. In der Firma in der ich tätig bin gibt es eine Projektdatenbank bzw. ein Softwareverwaltungstool. Wo die ausgelieferten Softwarestände sowie Teststände eingecheckt werden und jederzeit abrufbar sind. Diese Daten werden schematisch abgelegt.

1.2 (2 Punkte)

Nennen Sie ein Beispiel für Daten in Ihrem Umfeld, die gestreamt verarbeitet werden, nennen Sie ein Beispiel für Daten in Ihrem Umfeld, die über Batchverarbeitung verarbeitet werden.

Batchverarbeitung:
Unser Unternehmen entwickelt Software für die Logistikbranche. Ein großer Teil besteht darin Logfiles und Reports zu sammeln um diese dann zur gegebenen Zeit auswerten zu können. Diese Art von Daten wird über Batchverarbeitung verarbeitet. Da sie nur zu bestimmten Zeiten (ca einmal pro Woche von allen verfügbaren Anwendungen eingesammelt werden und dann zur Analyse weiterverarbeitet werden. (große Datenmenge)

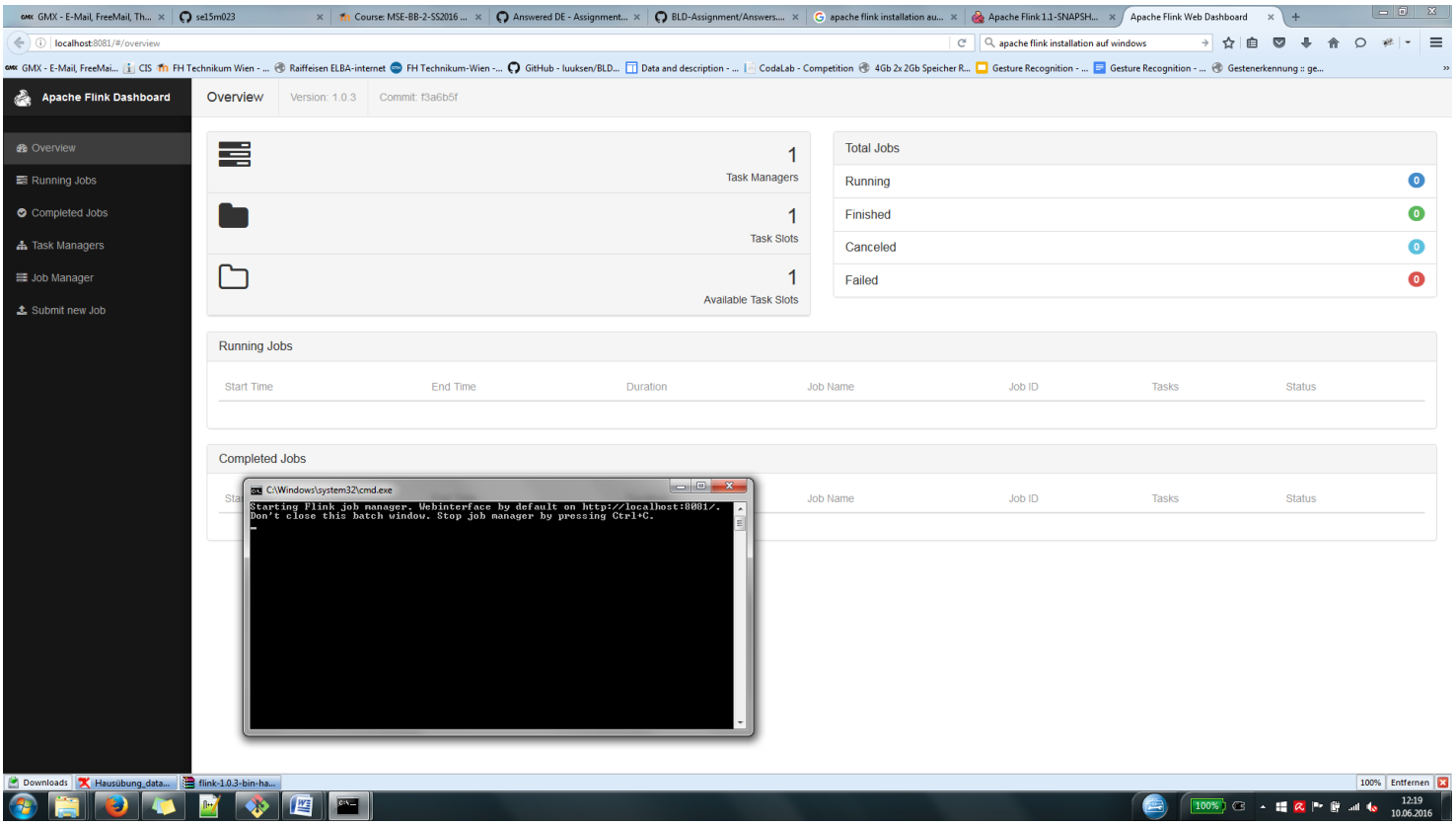
Gestreamt verarbeitete Daten:
Unsere Software berechnet ständig die Positionen der verfügbaren Stapler-Systeme die gerade in einem Lagerhaus unterwegs sind. Diese Postionen werden dann herangezogen um den Staplerfahrern den nächsten bestgeeignetsten Auftrag zuzuweisen um so effizient und schnell wie möglich die LKWs beladen zu können. Diese Positionsdaten werden von einer zentralen Software erfasst und muss quasi in Echtzeit verarbeitet werden um den Staplerfahrern den nächsten Auftrag zuweisen zu können. Diese Postionsdaten müssen somit gestreamt verarbeitet werden.

Assignment 2: Big Data in Ihrem Umfeld (4 Punkte)

Entscheiden Sie sich für eine Data Engineering Plattform. Apache Flink oder Apache Spark. Installieren Sie die auf Ihrem Arbeitsgerät.

- *Punkt: Erklären Sie ihre Entscheidung*
  Ich habe mich für Apache Flink entschieden weil es für Batch und Streaming Verarbeitung ausgelegt ist. Ich habe noch keine Erfahrung mit Flink und Spark und habe mich für die ersten Testversuche für Apache Flink entschieden, weil das System schnell aufgesetzt war und mir für erste Versuche einfacher erschien um Batchverarbeitung und Streamverarbeitung zu testen.

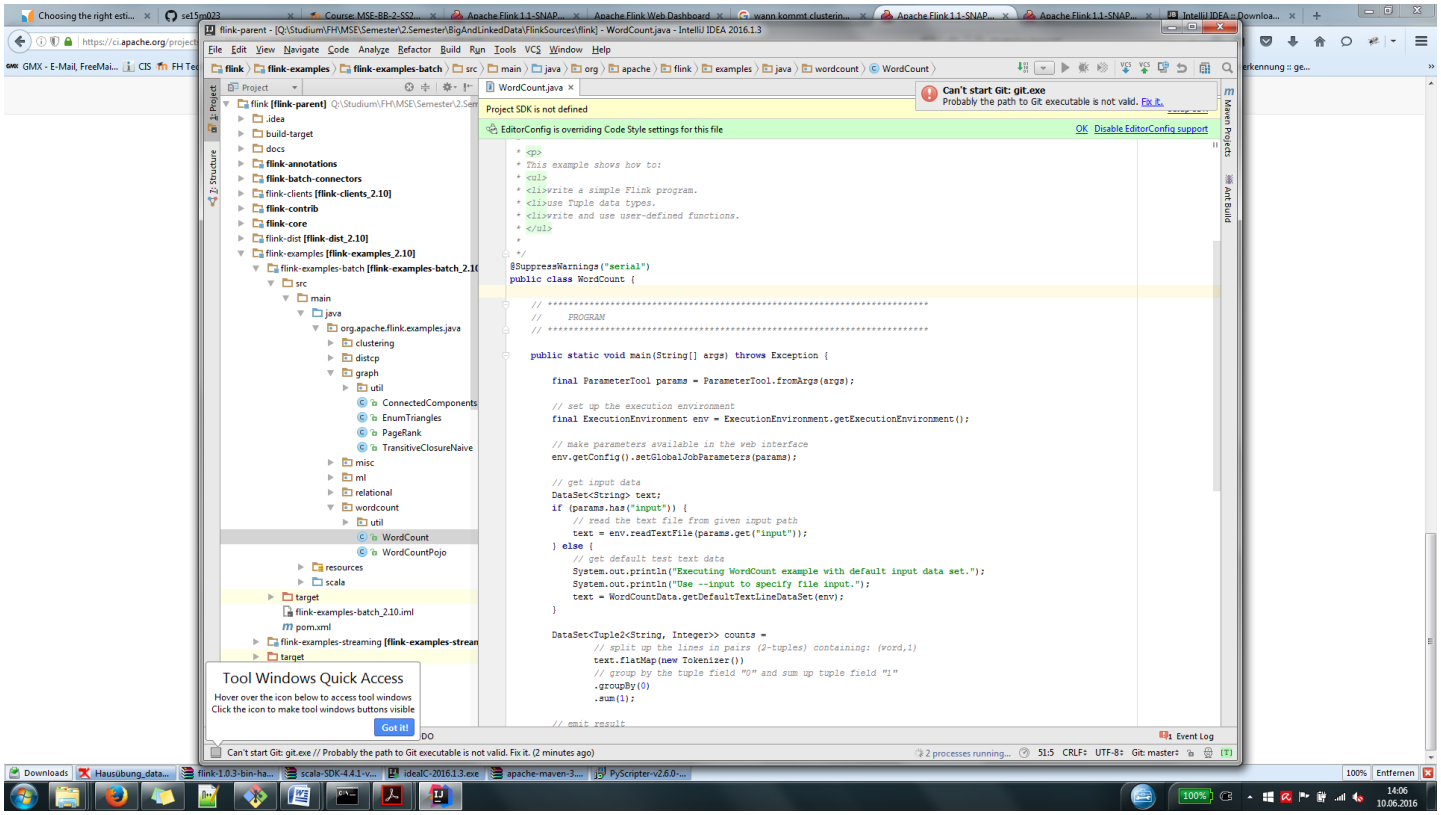
- *2. Punkte: Schicken Sie einen Screenshot der installierten Umgebung mit*



- *Punkt: Beschreiben Sie Ihre Toolchain, die Sie mit dem Framework nutzen würden (z.B: IDE)*
  Ich würde mich für “IntelliJ IDEA“ entscheiden. Ich habe zwar mehr Erfahrung mit Eclipse interessiere mich aber immer für neue IDEs. Außerdem habe ich gelesen dass es mit Eclipse und Scala Probleme gibt.

Assignment 3: Big Data in Ihrem Umfeld (4 Punkte)

Schreiben Sie ein simples Program mit dem Framework (z.B. Helloworld) und laden Sie es hoch.



- 2 Punkte für Programm

Ich habe versucht ein Beispielprogramm auszuführen. (WordCount)

- 2 Punkte, wenn das Programm auch ausführbar ist.

Leider habe ich es nicht geschafft das Programm auszuführen.