Assignments BLD Papp

# **Data Engineering**

# Assignment 1

## 1.1

Schematisch: Ich arbeite in der Firma an einem Projekt, wo ich Daten von Webservices (SOAP) abrufe. Diese sind in strukturierter Form zu verarbeiten.

Schemalos: Mailverkehr zwischen Kollegen, Kunden und mir. Kommunikationsdaten über Lync.

## 1.2

Gestreamt: Ich habe mich in der Firma mit Sitecore befasst. Sitecore ist ein CMS, welches auch Kundenaktionen auf der Webseite trackt und in eine NoSQL Datenbank speichert. Diese werden verarbeitet und die Ergebnisse in eine „Reporting“ Datenbank (MSSQL) gespeichert. Ich kann nicht sicher sagen, ob diese Daten tatsächlich gestreamt verarbeitet werden, aber möglich ist es.

Batchverarbeitung: Es werden regelmäßig Backups von Datenbanken gezogen. Dies wird einmal am Tag/Nacht gemacht.

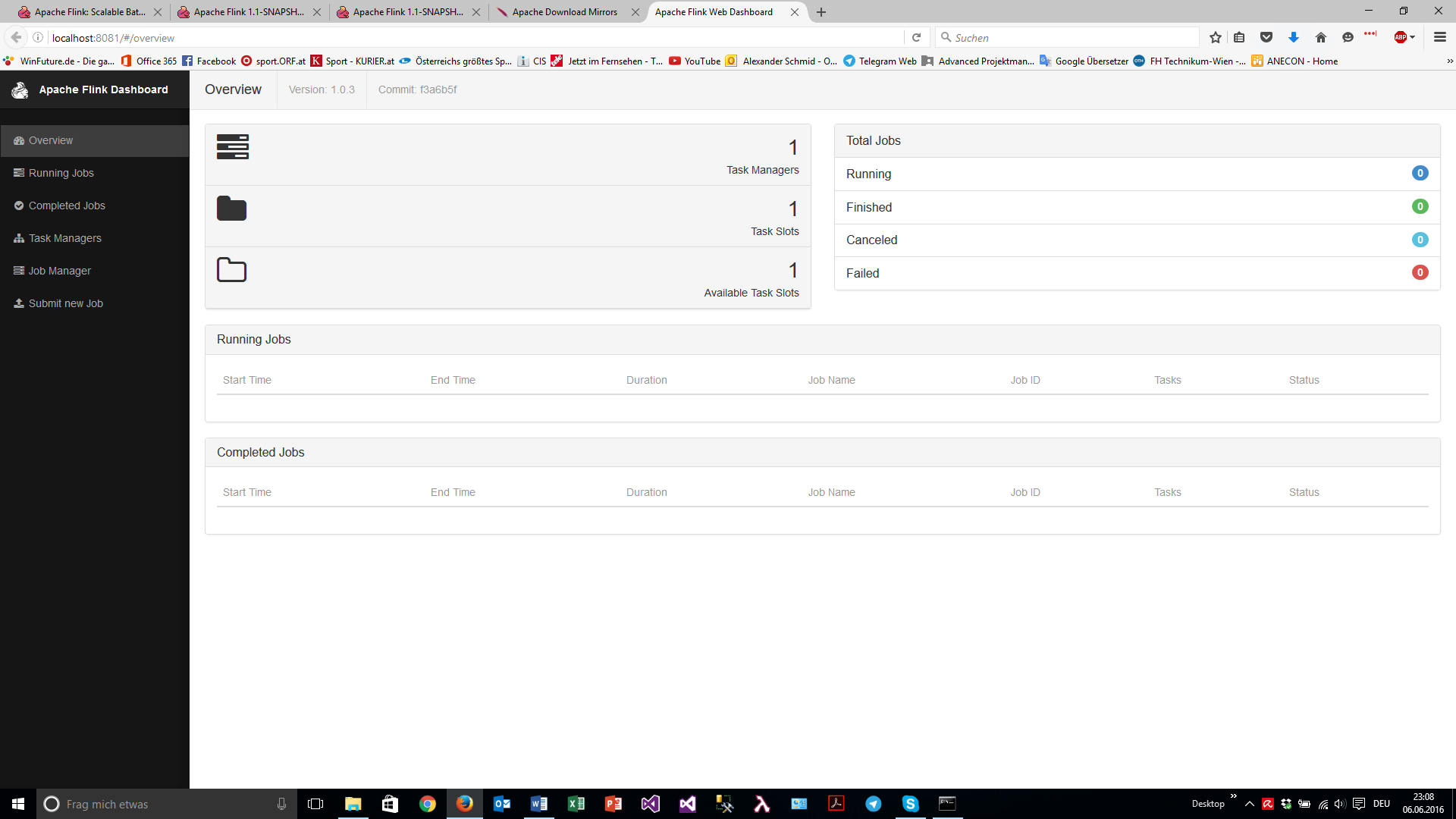
# Assignment 2

## 2.1

Ich habe mich für Apache Flink entschieden. Der Grund dafür ist, dass Apache Flink auch mit Streaming besser umgehen kann (hat Streaming Engine und nicht Batch Engine wie Spark) und somit eine relativ niedrige Latenzzeit erreicht. Das ist ein wesentliches Kriterium und hat mich überzeugt.

## 2.2

Für die Installation wurde folgende Anleitung verwendet: <https://ci.apache.org/projects/flink/flink-docs-master/setup/local_setup.html> Es wurde die Source Archiv Datei heruntergeladen, entpackt und über die CMD-Shell die start-local.bat ausgeführt und somit Flink gestartet. Unter folgender Url kann die Startseite nun aufgerufen werden: <http://localhost:8081/>. Der Screenshot der Startseite sieht wie folgt aus:



## 2.3

Da Flink mit Java verwendet werden kann und nur über Maven Dependencies eingebunden werden muss, würde ich Java verwenden. Meine bevorzugte IDE für Java ist IntelliJ IDEA. Die Abhängigkeiten zu Apache Flink würde ich über Maven einbinden.

# Assignment 3

Das Programm liegt im Ordner Flink\_Programm. Das Programm kann mit einer IDE (vorzugsweise IntelliJ) geöffnet werden. Der Unittest kann anschließend gestartet werden durch maven test.

Im Unittest wird ein einfacher Wordcount ausgeführt. Die Beispieleingabe besteht aus einem Satz in dem 5 mal das Wort „Alexander“ enthalten ist. Der Unittest überprüft dies. Ebenfalls werden die Ergebnisse ausgegeben.

# **Data Science**

# Assignment 1

## 1.1

## 1.2

# Assignment 2

## 2.1

## 2.2

## 2.3

# Assignment 3