# Data Engineering

## Assignment 1: Big Data in Ihrem Umfeld

### 1.1

Schematisch:

* Zeitaufzeichnungen
* Rechnungen
* Daten Webservices (SOAP, REST)

Schemalos:

* Email Verkehr
* Kommunikation mittels Skype for Business

### 1.2

Gestreamt:

* Word Online Dokumente beim Kollaborativen arbeiten
* Reports

Batchverarbeitung:

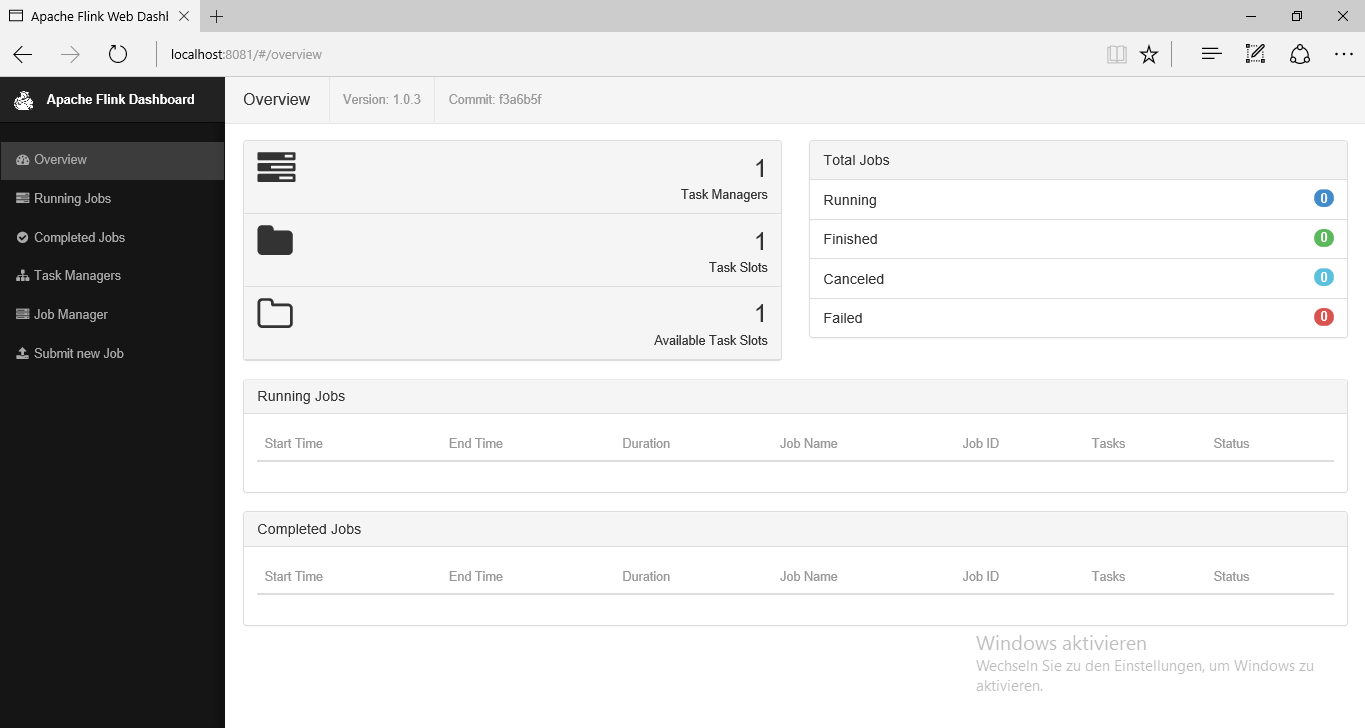
* Daily Build
* Backups

## Assignment 2: Big Data in Ihrem Umfeld

### 2.1

Ich habe mit beiden Plattformen keine Erfahrungen. Nach ein wenig Recherche habe ich mich Apache Flink entschieden. Apache Flink kann mit Streaming umgehen und weil ich das interessant finde, habe ich mich mal dafür entschieden.

### 2.2



### 2.3

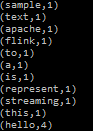
Als Framework würde ich Java nehmen in der IDE IntelliJ, da Flink Apache einfach mit einer Maven Dependency eingebunden werden kann.

## Assignment 3: Big Data in Ihrem Umfeld

* Sie finden das Beispiel im Ordner „HelloWorld“
* Ausführbar mittels der „helloWorld.jar“ Datei



* Folgender Output sollte am Ende zu sehen sein:



# Data Science

## Assignment 1: Technologien

### 1.1

* Matlab
* Apache Zeppelin
* SPSS

### 1.2

Ich habe noch nie mit einer dieser Technologien (außer Matlab) gearbeitet. Aber wenn ich mich entscheiden müsste, würde ich Python nehmen, da diese Sprache den Ruf hat leicht erlernbar zu sein und weil ich in einer Statistik es vorgeschlagen bekommen habe ;).

Quelle:

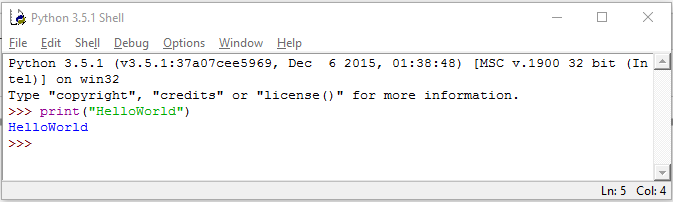
* <http://www.datasciencecentral.com/profiles/blogs/what-technology-tool-skills-do-data-scientists-jobs-require>

## Assignment 2: Technologien

### 2.1

Siehe 1.2 ;).

### 2.2



### 2.3

Als IDE würde ich PyCharm von JetBrains verwenden.

## Assignment 3: Big Science

Classification:

* Einteilung in vordefinierten Kategorien bzw. Klassen
* Z.B: Spamfilter

Regression:

* Numerische Werte ermittelt/vorhersagen
* Z.B: Einkaufsverlauf um Kaufverhalten des Kunden vorherzusagen (Amazon)

Clustering:

* Daten mit gemeinsamen Merkmalen werden zu Gruppen zusammengefasst
* Z.B: Analyse Kundenkarte => angepasste Angebote

Dimensional Reduction:

* Reduktion von Komplexen Daten um Performance zu erhöhen und Speicherplatz zu sparen
* Nur Relevante Daten werden berücksichtigt