

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:

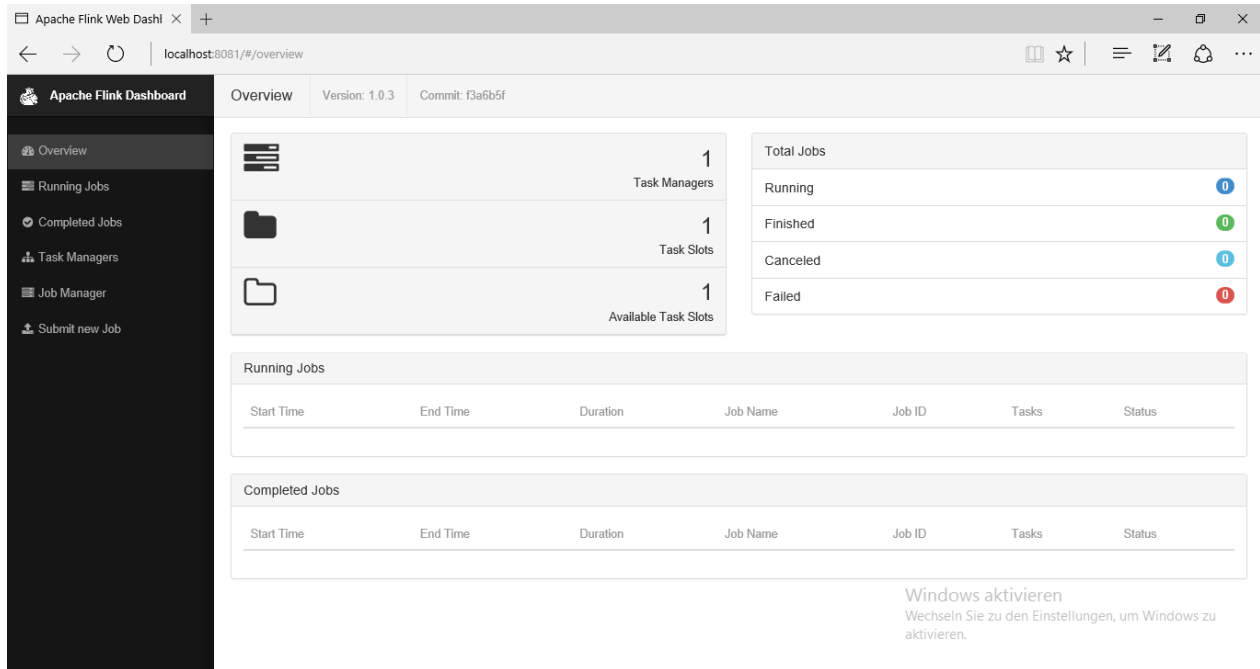
- Daily Build
- Backups

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



Apache Flink Web Dashboard

Overview | Version: 1.0.3 | Commit: f3a6b5f

Task Managers 1

Task Slots 1

Available Task Slots 1

Total Jobs

Job Status	Count
Running	0
Finished	0
Canceled	0
Failed	0

Running Jobs

Start Time	End Time	Duration	Job Name	Job ID	Tasks	Status
------------	----------	----------	----------	--------	-------	--------

Completed Jobs

Start Time	End Time	Duration	Job Name	Job ID	Tasks	Status
------------	----------	----------	----------	--------	-------	--------

Windows aktivieren
Wechseln Sie zu den Einstellungen, um Windows zu aktivieren.

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

```
D:\Angie\FH-Projekte\BLD>java -jar helloWorld.jar
```

- Folgender Output sollte am Ende zu sehen sein:

```
(sample,1)
(text,1)
(apache,1)
(flink,1)
(to,1)
(a,1)
(is,1)
(represent,1)
(streaming,1)
(this,1)
(hello,4)
```

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:

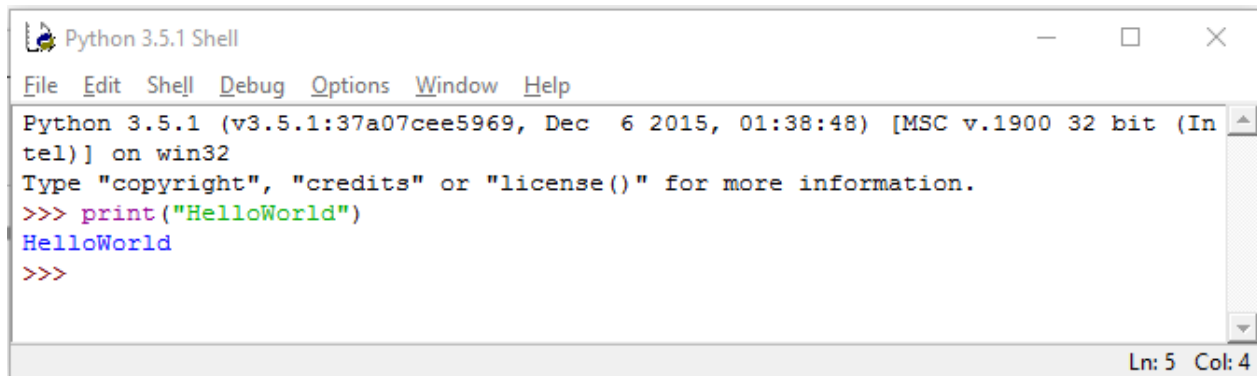
- <http://www.datasciencecentral.com/profiles/blogs/what-technology-tool-skills-do-data-scientists-jobs-require>

Assignment 2: Technologien

2.1

Siehe 1.2 ;).

2.2

A screenshot of a Python 3.5.1 Shell window. The window has a title bar 'Python 3.5.1 Shell' and standard Windows window controls. Below the title bar is a menu bar with 'File', 'Edit', 'Shell', 'Debug', 'Options', 'Window', and 'Help'. The main text area shows the following text:

```
Python 3.5.1 (v3.5.1:37a07cee5969, Dec 6 2015, 01:38:48) [MSC v.1900 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>> print("HelloWorld")
HelloWorld
>>>
```

The status bar at the bottom right indicates 'Ln: 5 Col: 4'.

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