Data Engineering

Assignment 1: Big Data in Ihrem Umfeld

Schemalose Daten:

Ein Kunde für den ich arbeite verkauft Personendaten, diese sind in XML Dokumente gespeichert und können beliebig erweitert werden. Zum Beispiel durch neue Lifestyle Attribute usw. Als DB wird Marklogic verwendet.

Strukturierte Daten:

Parallel zur Marklogic DB gibt es noch eine Postgres Datenbank in der alle Kunden Informationen, Job Status usw gespeichert werden. Hier gibt es so gut wie keine Änderungen.

Batch Processing:

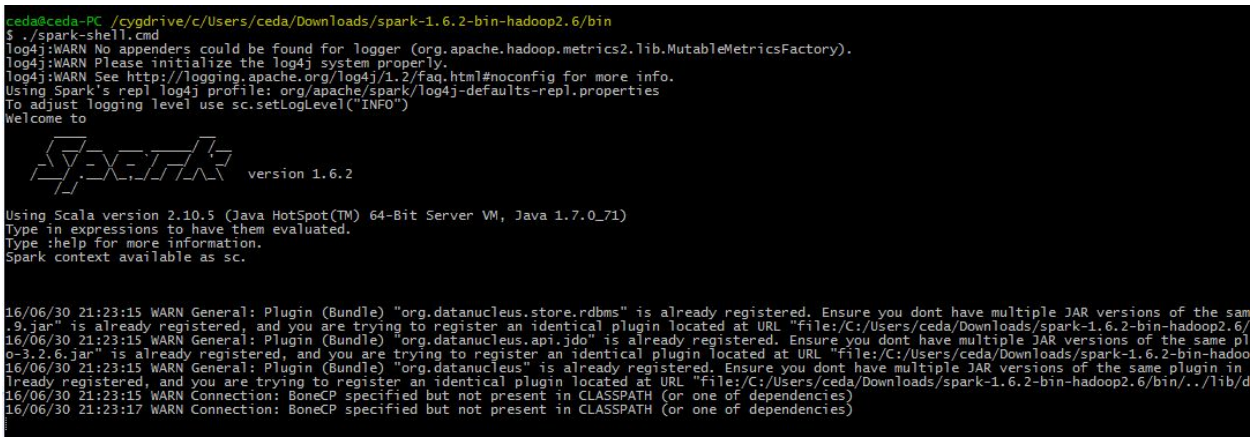
In den oben genannten Big Data Projekt werden die Daten durch Talend ETL Jobs aufbereitet und danach werden die Dokumente erstellt bzw aktualisiert. Der Verarbeitung passiert in Batches.

Streaming:

Am anderen Ende sitzt der Kunde der sich über eine Weboberfläche (Java + HTML + JS) diverse Selektionen zusammenbaut und diese gegen die Datenbank ausführt und sofort ein Ergebnis erhält.

Assignment 2: Big Data in Ihrem Umfeld

Ich habe mich für Apache Spark entschieden, da wir in der Firma mehr mit Spark arbeiten als mit Flink. Weiters ist in Spark Batch und Streaming möglich.



Da ich ursprünglich aus dem Java Bereich komme, ist in meinem Toolstack immer Eclipse dabei. Und um auch auf Windows alle Vorteile einer Command Line nutzen zu können ist das wichtigste Tool Cygwin, da ich mir dadurch einige Eclipse Plugins erspare (GIT, MAVEN, …) und andere Tools z.B. Putty erspare.

Data Science

Assignment 1: Technologien

Alternativen zu R und Python:

- Java
- SQL
- SAS

Data Science Tools:

- SQL
- Java
- Excel
- Shell
- Notepad ++
- XML → XPath, XQuery

Assignment 2: Technologien

Ich habe mich für Python entschieden, da es eine größere Community gibt, das heißt besserer Support, mehr Libs usw. Weiters kann man Python gut mit Webapplikation und Datenbank verwenden. R ist aus meiner sicht doch mehr für Standalone Anwendungen zu gebrauchen.

```
C:\Users\ChristophEder>python
Python 3.5.2 (v3.5.2:4def2a2901a5, Jun 25 2016, 22:01:18) [MSC v.1900 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> print "Python is fun!"
  File "<stdin>", line 1
    print "Python is fun!"
                         ^
SyntaxError: Missing parentheses in call to 'print'
>>> print "Python is fun!"
  File "<stdin>", line 1
    print "Python is fun!"
                         ^
SyntaxError: Missing parentheses in call to 'print'
>>> print "Python is fun!"
  File "<stdin>", line 1
    print "Python is fun!"
                         ^
SyntaxError: Missing parentheses in call to 'print'
>>> print("Python is fun!")
Python is fun!
>>>
```

Toolchain:
Ich würde PyCharm verwenden, da ich gute Erfahrungen mit den JetBetrains Produkten bis jetzt gemacht habe(PHPStorm, Intellij, AndroidStudio). Ich find es sehr praktisch das man von Haus aus bei den JetBrains Tools die Eclipse Shortcuts einstellen kann. Somit hat man in jeder IDE die gleichen Shortcuts.