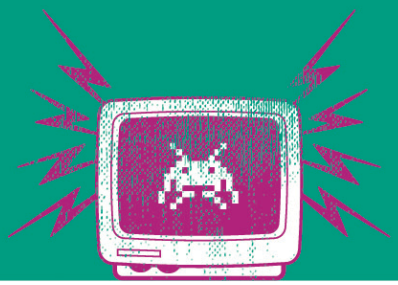
Softwareentwicklung



Big and Linked Data

Assignments
Data Engineering
Data Science

Christoph Edinger (se15m025)

Data Engineering – Big Data in Ihrem Umfeld

Assignment 1

1.1 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.

Im Arbeitsumfeld sind bspw. Emails, Word-Dokumente, PowerPoint Präsentationen, Textdateien, Inhalt von Messages etc. **unstrukturierte Daten**, also im Prinzip alle Daten, welche nicht in einer Datenbank oder einer speziellen Datenstruktur vorliegen.
Strukturierte Daten hingegen unterliegen, wie bereits erwähnt, einer speziellen Datenstruktur und sind meistens in einer Datenbank abgelegt und Beispiele sind: Userdaten von bspw. einer App, Kontodaten + Transaktionen etc.

1.2 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.

Streaming: Das beste Beispiel sind in der heutigen Zeit vermutlich Anbieter wie Netflix, Amazon Video, Maxdome, Newsfeeds. Ich arbeite als Softwaredienstleister für Privatbanken und hier wäre ein Beispiel Datenfeeds von diversen Kurslieferanten.

Batchverarbeitung: Daten werden von AS/400 nach Tagesendverarbeitung nach MS SQL Server übertragen => AS 400 bereitet Daten auf und werden durch Overnight-Batchjobs in SQL Datenbank gespielt , welche wiederum von einer Webanwendung verwendet wird.

Assignment 2

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

Entscheidung

Für dieses Assignment habe ich Flink gewählt aus dem Grund, da es verschiedenste APIs anbietet. Somit wäre es möglich die Flink Engine für verschiedene Applikationen zu benutzen. Flink ist außerdem optimiert für zyklische und iterative Prozesse.

Screenshot: Apache Flink Dashboard

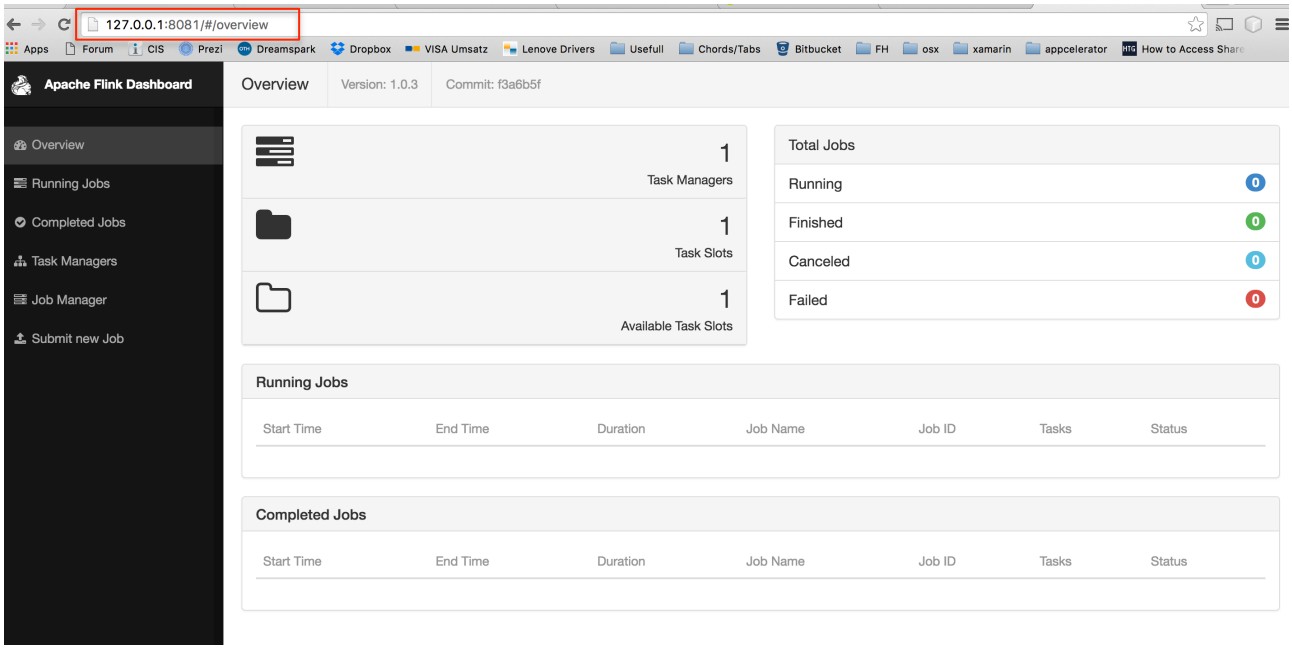


Abbildung 1: Apache Flink Dashboard - localhost

Toolchain

Für die Entwicklung von Java Source würde ich mich für die IDE von Jetbrains, nämlich IntelliJ IDEA entscheiden. Für das Builden würde ich Maven verwenden. Um ein Programm zu erstellen wären dafür folgende Schritte notwendig [1]:

1. Installation IntelliJ IDEA
2. Mit dem Befehl CURL HTTPS://FLINK.APACHE.ORG/Q/QUICKSTART.SH | BASH Template für Java Projekt anlegen
3. Projekt von Flink Directory importieren
4. Code schreiben
5. JAR File mit dem Befehl erzeugen: MVN CLEAN INSTALL -PBUILD-JAR
6. Ausführen via Terminal oder neuen Job via Flink Webinterface submitten

Assignment 3

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

Programm

Das Programm ist im Repository vorhanden (flink-java-project) und ist ausführbar. Des Weiteren wurde über das Flink Webinterface ein neuer Job submittet (JAR auswählen und Entry Class muss angegeben werden, in dem Fall HelloWorldFlink).

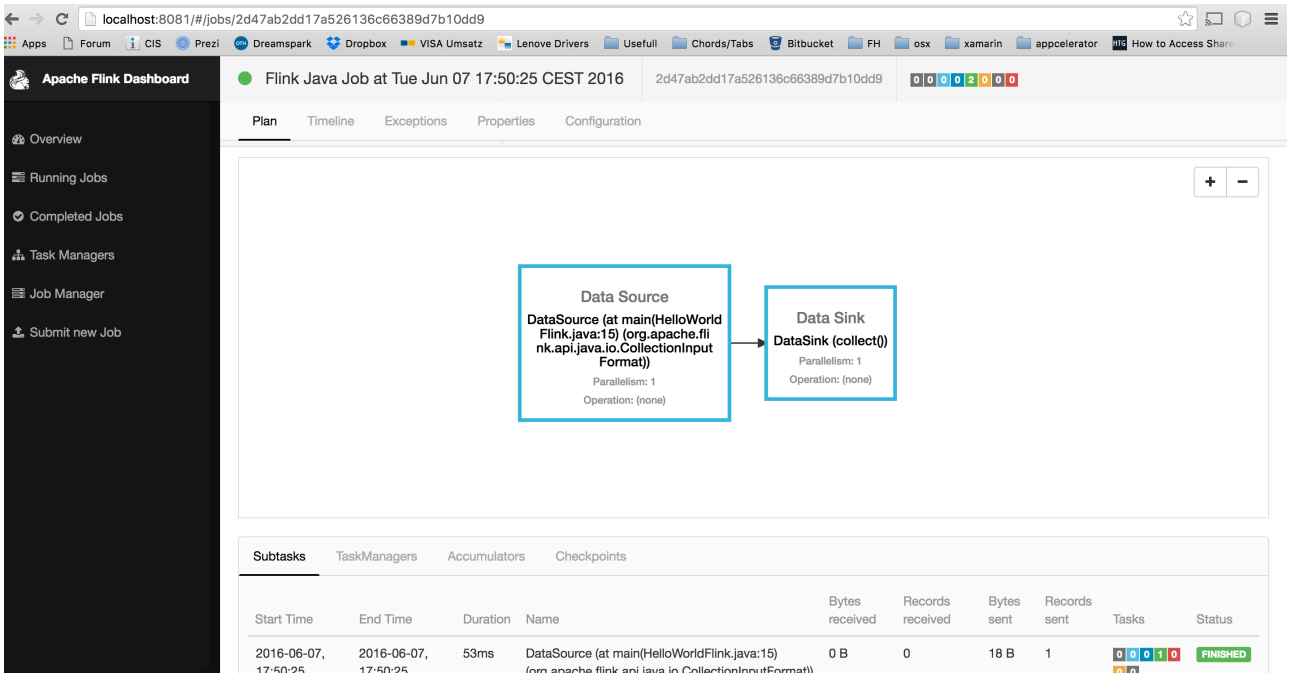


Abbildung 2: Apache Flink Job - HelloWorldFlink

Data Science

Assignment 1: Technologien

1.1 Sie haben in der LVA zwei Frameworks kennengelernt (R und Python). Nennen Sie zwei weitere Technologien, um Daten zu analysieren (müssen nicht open source sein)

Sage (Open Source), GNU Octave (Open Source)

1.2 Sie bekommen den Auftrag, sich mit einer Data Science Technologie zu arbeiten. Nennen Sie Technologie, die ihnen auf dem ersten Blick am besten für Sie erscheint und begründen Sie das!

Ich würde mich für R entscheiden, da diese Technologie sehr gut geeignet ist für explorative Arbeit. R ist auch sehr gut geeignet für jegliche Art von Datenanalyse [2]. Außerdem bietet R eine eigene IDE nämlich das RStudio.

Assignment 2: Technologien

Entscheidung

Ich würde mich für R entscheiden, da ich es einerseits noch nicht kenne und gerne neue Technologien ausprobiere und andererseits aus meiner Sicht sehr gut geeignet ist für Datenanalysen.

Screenshot - RStudio

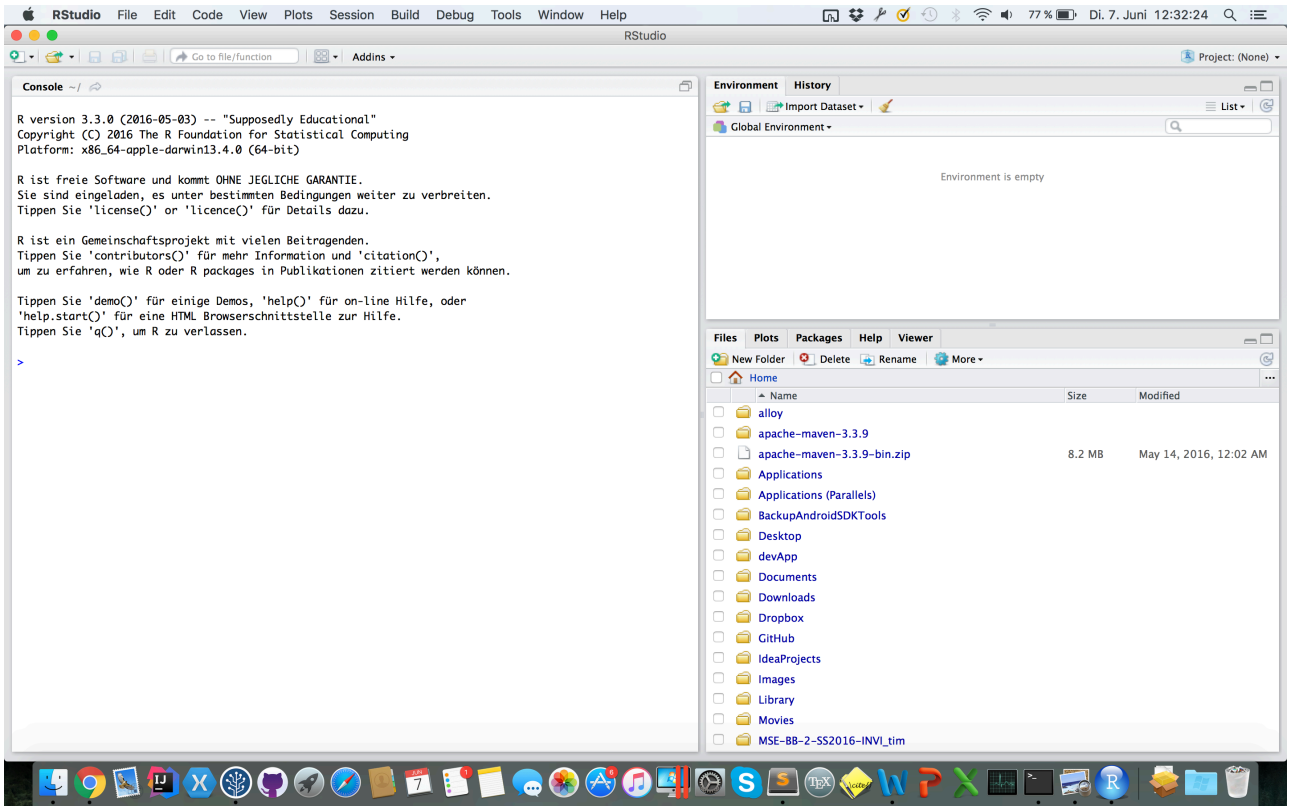


Abbildung 3: RStudio

Toolchain

Im Prinzip bietet das RStudio bereits alles, was benötigt wird.

Assignment 3: Big Science

Der Cheatsheet auf http://scikit-learn.org/stable/tutorial/machine_learning_map/ ist eine einfache Anleitung, wie man den richtigen Algorithmus zum richtigen Data Science Problem findet.
Schauen Sie in Google nach und lernen Sie classificatiom, regression, clustering und dimensional reduction unterscheiden.
Nennen Sie ein Beispiel aus ihrem Umfeld, wo Sie mit dem Algorithmus zu tun haben. Das kann ein Beispiel sein, wie: Wenn Sie bei Amazon einkaufen. Wenn Sie von einem Marketinginstitut angerufen werden, etc.

Classification: Wie der Name schon vermuten lässt, handelt es sich hierbei um die Klassifizierung von Daten, was bedeutet, dass die Daten in bestimmte Kategorien geteilt werden. Beispiel: Spamfilter versuchen Emails zu klassifizieren und entsprechen zu markieren [3]

Regression: Regression versucht einen reellen Wert für eine Variable aus die vorhandenen Daten vorauszusagen. Ein Beispiel wäre die Frage „Wie hoch werden die Kosten für ein bestimmtest Haus sein?“. Mit Regression könnte nun nach ähnlichen Häusern gesucht werden und so ein Model erstellt werden damit man diese Frage beantworten kann. Regression und Classification sind miteinander verwand jedoch unterschiedlich, da Regression Ergebnisse sozusagen beziffert. [3]

Clustering: Hierbei werden ebenfalls, wie bei Classification, Gruppierungen durchgeführt, jedoch müssen diese Gruppen nicht vorher bekannt sein. Ein Beispiel dafür wäre die Logging Daten einer Webseite zu analysieren und bestimmte Trends herauszufinden.

Dimensional Reduction: Daten werden reduziert damit die wichtigen und entscheidenden Daten von den anderen getrennt werden können. Ein Beispiel dafür wäre, aus den Loggingdaten einer Webseite, jene herauszufiltern, welche zur Verbesserung der Usability beitragen.

Quellen

[1] Apache Flink, *IDE Setup*, https://ci.apache.org/projects/flink/flink-docs-master/internals/ide_setup.html [online: 01.06.2016]

[2] KDnuggets, Data Mining, Analytics, Big Data and Data Science ,*R vs Python for Data Science: The Winner is...*, http://www.kdnuggets.com/2015/05/r-vs-python-data-science.html [online: 01.06.2016]

[3] KDnuggets, Data Mining, Analytics, Big Data and Data Science, *Fundamental methods of Data Science: Classification, Regression And Similarity Matching* http://www.kdnuggets.com/2015/01/fundamental-methods-data-science-classification-regression-similarity-matching.html [online: 01.06.2016]