BLD Assignment

Data Engineering

Assignment 1

1.1

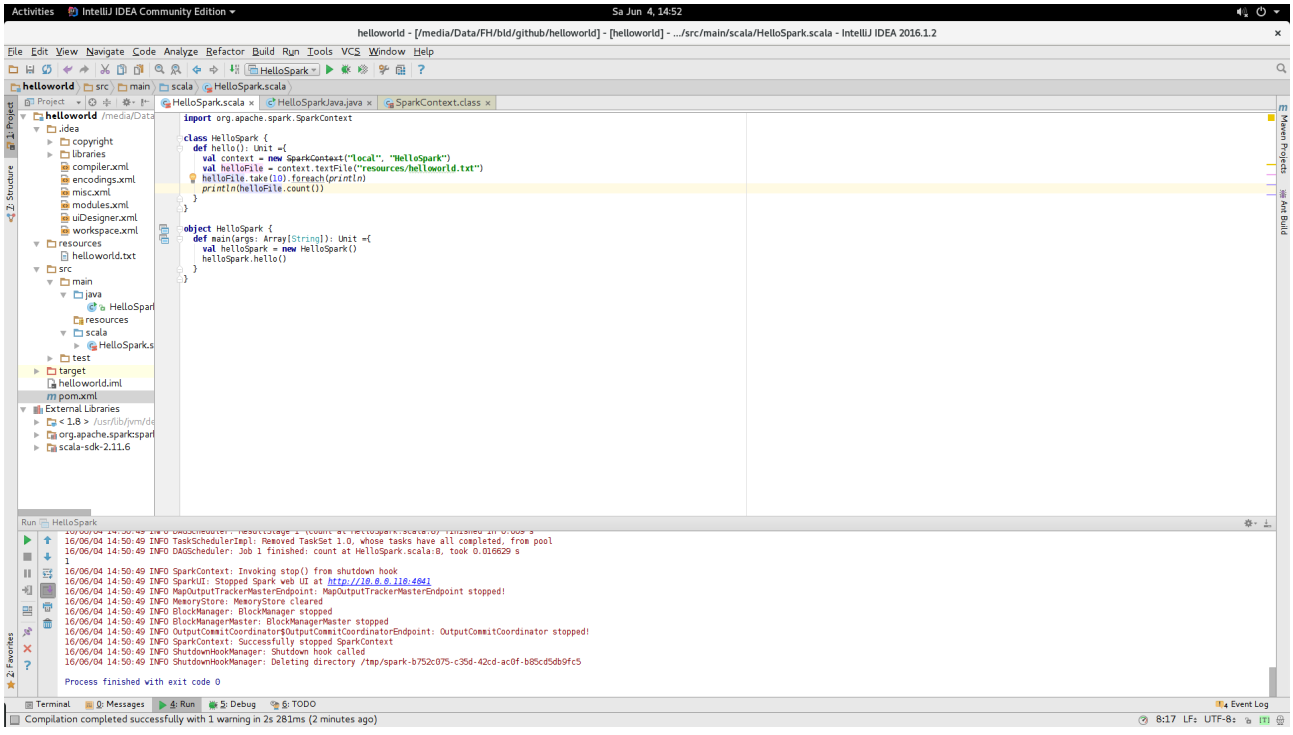
Die vom Buildserver erzeugten Metadaten sind strukturiert in einer SQL Datenbank gespeichert. Alle Daten die im Netzwerkshare gespeichert werden sind (grundsätzlich) schemalos. Einzelne Dateien können ein Schema haben, aber man kann nicht davon ausgehen, dass alle Dateien ein Schema haben oder dass Dateien mit Schema vorhanden sind.

1.2

Der automatische Daily Build ist ein Beispiel für Batchverarbeitung. Alle Änderungen werden einmal täglich erfasst und verarbeitet. WebServices sind ein Beispiel für Streamverarbeitung. Auf die übertragenen Daten wird sofort die Businesslogik der Anwendung ausgeführt (zB JIRA API).

Assignment 2

Ich verwende Apache Spark, weil es die Programmierung nach einem funktionalen Paradigma ermöglicht. Beim Umgang mit Datenmengen finde ich das besser. Außerdem bietet es RDD und arbeitet im RAM. Als Entwicklungsumgebung verwende ich IntelliJ mit Maven Dependencies. Programmiersprache ist Java oder Scala.



Assignment 3

Das Program befindet sich in der Datei helloworld/src/main/scala/HelloSpark.scala. Es gibt auch eine Java-Version in helloworld/src/main/java/HelloSparkJava.java. Die Anwendung ist mit IntelliJ ausführbar.

Data Science

Assignment 1

1.1

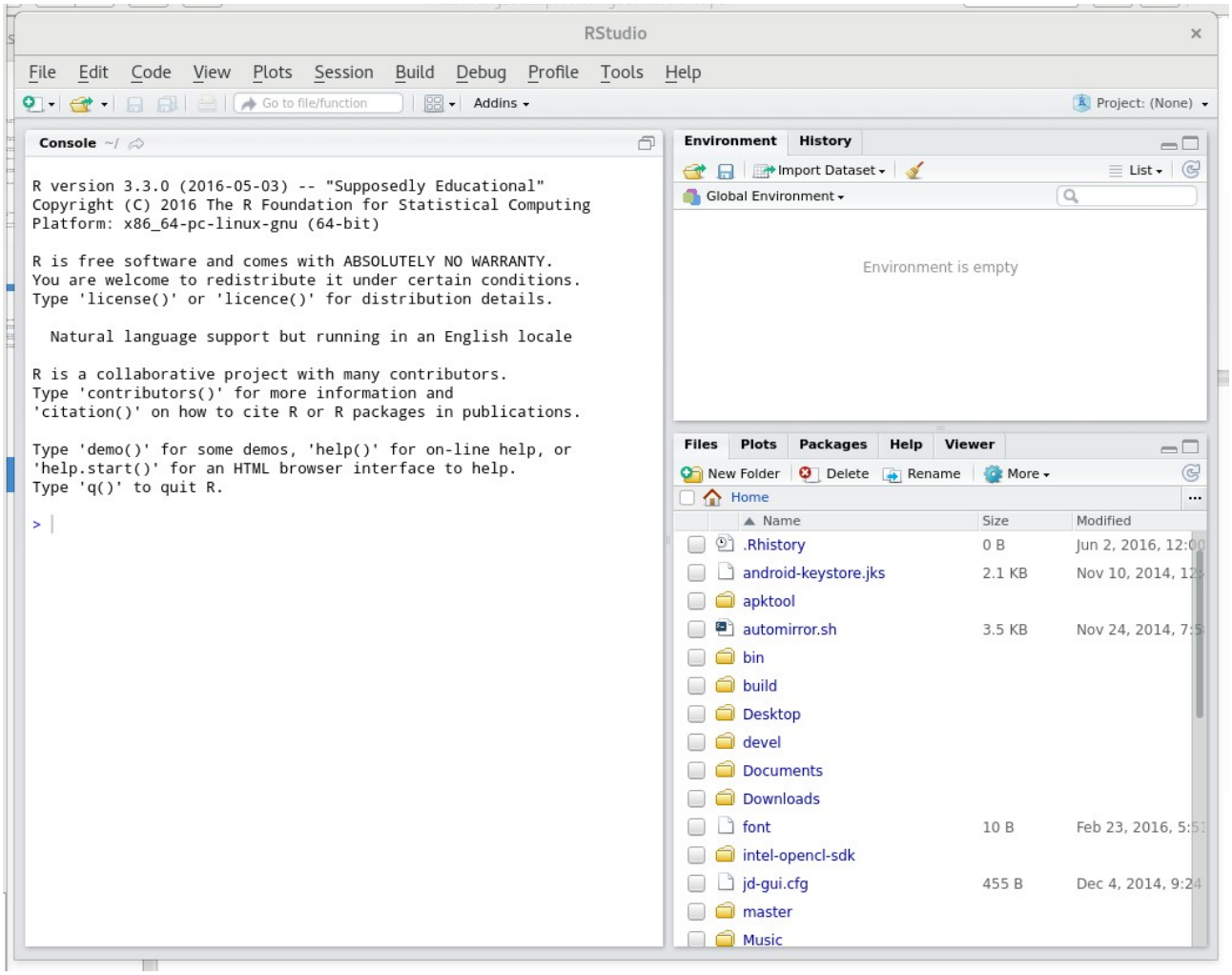
Java und in einer JVM laufende Sprachen wie Scala sind auch weit verbreitet. Sie kommen vor allem zusammen mit Hadoop zum Einsatz. Julia ist ein Newcomer und könnte in Zukunft R und Python ablösen. Die Sprache ist speziell für Big Data entwickelt worden.

1.2

Für einen Auftrag bevorzuge Python, weil ich die Sprache schon kenne und die Tools schon installiert habe. Python bietet eine umfassende Library für die Analyse von Daten.

Assignment 2

Ich benutze für diese Übung R, weil ich bisher noch keine Gelegenheit hatte, die Sprache kennenzulernen und weil sie im Big Data Bereich weit verbreitet ist.



Als Entwicklungsumgebung verwende ich RStudio.

Assignment 3

Classification ist die Aufteilung der Datensätze auf vordefinierte Kategorien. Für die automatische Klassifizierung muss das System mit Daten trainiert werden. Diese Technik wird auch als Supervised Learning bezeichnet.
Regression wird zur Findung eines Zusammenhangs zwischen Werten verwendet. Der ermittelte Zusammenhang kann zur Bildung einer Prognose verwendet werden.
Clustering ist eine Methode zur Findung von Kategorien. Es kann sein, dass die gefundenen Kategorien nicht offensichtlich sinnvoll sind. Diese Technik wird auch als Unsupervised Learning bezeichnet.
Dimension reduction reduziert die Komplexität der Daten um Speicherplatz und Rechenzeit zu sparen. Dabei werden meistens Daten mit möglichst hoher Varianz beibehalten.

Diese Techniken können zum Beispiel für die Befüllung der Startseite der Youtube-App verwendet werden. Dort werden Videos und Channels empfohlen die den Vorlieben des jeweiligen Benutzers entsprechen sollen. Aufgrund der bereits gesehen und positiv bewerteten Videos kann ein Benutzer einer Kategorie zugeordnet werden. Auf der Startseite werden dann Videos aus der selben Kategorie empfohlen.