Chapter 12
Open Source Software Integrations

Gábor Bakos
Mind Eratosthenes Kft., Hungary

ABSTRACT

RapidMiner is a popular open source, Java-based data analytics software. This chapter shows a case studying how it can be integrated to other programs. R and Vega integration is also introduced briefly in connection to open source software integration. The authors cover some general practices regarding integrating software to a Java environment and collect some popular open source libraries that are found useful related to data analytics. They focus mostly on the Java platform, though some parts of the chapter are applicable to other platforms too.

INTRODUCTION

Open source software are already present in many business intelligence systems, though the open source tools on their own can be useful in many ways. Once you require a step to be performed by another tool, you should consider integrating it too. With open source software you have a not too hard task as we demonstrate in this chapter.

This chapter will introduce some of the challenges and options when we integrated RapidMiner to KNIME. Both are open source data analytics software (the former at the time of writing this chapter is "business source" (Mierswa, 2013)) with similar goals, but different strengths. This goal would be too specific for the chapter, so we will give a broader context, show how these bits and pieces can be used in similar situations and feature other products, like R and Vega.

These tools are usually multi-platform, but –because you have the sources– you can port to new platforms too with varying efforts.

By the end of this chapter you will have an understanding how should you start integrating an open source data analytics software to your business intelligence system. This chapter will also cover what useful/popular open source tools are available for the Java platform that can be used in this regard.

In this chapter first we will introduce the software –and their benefits– we had experience regarding integration (either by integrating, or examining the integrations), after that we will show a general plot how you can integrate a data analytics software and what you might expect.

DOI: 10.4018/978-1-4666-6477-7.ch012

We will go through the presented integrations and present the problems that were experienced with specific solutions to them later. Before the conclusion, we will present some future research directions, finally we recommend a few readings in the topic of open source software integrations.

BACKGROUND

KNIME and RapidMiner are both Java based applications that can load and analyze data. A bit market focused introduction to the companies behind KNIME, RapidMiner and R from Gartner: "Knime (www.knime.com), based in Zurich, Switzerland, offers a free, open-source, desktop-based advanced analytics platform. It also offers a commercial, server-based on-site or customer cloud solution providing additional enterprise functionality. Knime has a presence across a range of industries, but with particular experience in life science, government, education and communications." (Herschel, Linden, & Kart, 2014) "RapidMiner (www.rapidminer.com), formerly known as Rapid-I, is based in Cambridge, Massachusetts, U.S. RapidMiner is an open-source, client/server-based solution also available as a commercial solution with the ability to work on larger datasets and to connect to more data sources. The platform derives its extensibility via source-code availability and integration of other open-source solutions (for example, R and Weka)." (Herschel et al., 2014) "Revolution Analytics (www.revolutionanalytics.com) is based in Mountain View, California, U.S., and provides an enterprise-grade, multiplatform execution framework and an ecosystem of partnerships to the increasingly popular open-source R language." (Herschel et al., 2014) We should introduce Vega too: "Vega provides a higher-level visualization specification language on top of D3." (Heer, 2014) ("D3.js is a JavaScript library for manipulating documents based on data." (Bostock, 2013))

You might ask why you would want to use any of these tools. Because these already did the hard work of integrating the useful algorithms and provide them through an easier to use interface with building blocks that can be combined together. You can use them for:

- Agile data analysis (Snyder, 2014),
- Big data analysis
  - KNIME is used with Actian RushAccelerator for KNIME: "KNIME […] with thousands of data prep and analytics functions, is now tightly integrated with Actian DataRush, the world's fastest compute engine for commodity hardware and big data clusters" (Actian Corporation, 2014);
  - "Radoop is a fully graphical tool supporting the whole range of big data analytics from ETL and ad-hoc reporting to predictive analytics" (Radoop LLC, 2014) based on RapidMiner and Hadoop;
  - "The RevoScaleR package provides a mechanism for scaling the R language to handle very large data sets" (Rickert, 2011, p. 2);
  - Alternatively
    - You can use products of other vendors like Zementis to apply PMML models extracted from KNIME or R for big data
    - Or use JDBC drivers for Hive or other big data warehouse: "KNIME database nodes can already be used to perform Big ETL using Hadoop and HIVE" (KNIME.com AG, 2014)
- "Data access, data filtering and manipulation, predictive analytics, further advanced analytics, and analytical business" (Herschel et al., 2014)