OPEN SOURCE AND IBM SPSS MODELER

The best of both worlds

IBM **Analytics**

OVERVIEW

A successful data-driven organization has to provide the right tools for its data analysts, developers, and business end-users. Increasingly, this means leveraging open-source software. But for all their benefits, popular open-source data programs also come with potentially time-consuming drawbacks. This paper will discuss how to handle these drawbacks, and get the right data tools to the right people, by leveraging popular open source tools with IBM® SPSS® Modeler software.

The R or Python programming languages, along with theApache Spark and Hadoop data frameworks, are preferred open-source tools for many modern analysts and developers. Their strengths include sheer speed and broad support from highly active contributor communities.

However, these tools introduce configuration and security challenges, and can require an extensive learning curve. Open-source tools can feel intimidating to business users seeking a point-and-click interface and polished graphical presentation tools.

IBM SPSS Modeler allows users to collaborate using the tools they're comfortable with—whether it's raw code or an intuitive GUI. In addition, SPSS delivers superior data management and scalability. And SPSS users get access to a rich, ever-expanding collection of text analysis, machine learning algorithms and graphing libraries to help them gain deeper insights from their data.

By coupling SPSS with open source tools, a data-driven organization can get the best of all worlds.

BENEFITS
of open source languages

R, a favorite of math and science communities, and Python, a high-level language with wide adoption, are common open-source languages used for data science.

R

R is an open source programming language and software environment for statistical computing and graphics. The R language is widely used by statisticians and data scientists. In addition, R is widely used for advanced data analysis. R provides a wide variety of statistical and graphical techniques such as linear and nonlinear modeling, classical statistical tests, time-series analysis, classification and clustering. R has more than 11,000 packages available from multiple repositories.

Python

Python is a popular object-oriented high-level programming language, with widespread usage in math, science, engineering, and web applications. Python's high-level built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development. It's also commonly used as a scripting language to connect existing components together. One of Python's strengths is its large standard library, as well as over 100,000 third-party packages offering expanded functionality. Common Python libraries used for machine learning or other data science activities include scikit-learn, NumPy, SciPy and pandas.

BENEFITS
of open source data frameworks

The Apache Software Foundation supports two modern frameworks that are widely used for data analytics, Apache Hadoop and Apache Spark. Both Hadoop and Spark are supported by rich ecosystems of advanced tools for data science, visualization, machine learning, and more, thanks to extensive support from open-source contributors, such as IBM.

Apache Hadoop

Apache Hadoop is best known as a way to handle big data. It's a highly versatile framework for distributed data storage and processing available at hadoop.apache.org. It enables processing of large structured, semi-structured, and unstructured data sets. It can absorb and aggregate data from many disparate sources. And it's built for scale—from a single server up to thousands of machines. In general, it's a popular choice for organizations that need rapid and accurate answers from very large data sets.

Apache Spark

Apache Spark takes Hadoop's big data abilities and adds even more speed and analysis capabilities. Spark is an extremely fast cluster computing engine designed for high-performance data analysis, and available at spark.apache.org. Spark is compatible with Hadoop data, but its advanced in-memory computing engine can deliver 100x the performance of Hadoop alone for certain workloads. Apache Spark includes powerful built-in developer libraries, including Spark Streaming for streaming data applications, Spark SQL for structured data queries, MLlib for machine learning, and GraphX for graphing. These tools make it easy to build powerful apps for graphics and analysis.

It may seem appealing to use open-source software exclusively—especially given these programs are available at no charge.

BUT THERE ARE DRAWBACKS AND LIMITATIONS.

Learning curve

People with coding skills might be comfortable with R and Python, and data scientists and developers might appreciate the granular control of Hadoop and Spark. But nontechnical stakeholders can find these tools intimidating to learn and use.

Data connections

Getting R and Python to properly connect to data stores can pose a configuration challenge.

Data management

Keeping all your data organized and accessible to the right people can be time consuming with these tools.

Analysis

To compare different algorithms and techniques, the user has to juggle a suite of different open-source tools that all work differently. Without a unified platform to compare methods, it's much harder to develop precise models.

Deployment

It takes time and effort to get all of these programs running optimally in an operational environment.

Performance

If not properly configured and monitored, open-source data applications can quickly become memory hogs that consume resources.

Collaboration

A team using R will hit some speed bumps when it's time to share work with a team using Python, and so on.

Enterprise security

Your organization may demand greater levels of security and quality control than an open-source software release.

These limitations can add to the costs of running R, Python, Apache Hadoop and Apache Spark.

BENEFITS
of integrating with IBM SPSS Modeler

Recognizing that open-source data analytics have drawbacks, they have too many benefits to be disregarded. A smart choice is to harness the many advantages of R, Python, Apache Hadoop and Apache Spark by integrating them together using IBM SPSS Modeler software.

SPSS is a powerful data science platform that gives end-users the abilities to handle large data sets and get high quality graphs and other forms of output. SPSS software is intuitive to use. It can read and merge data from a variety of sources. And it offers the ability to distribute integrated open-source packages to a wide range of users, even those who aren't coders.

Here's a look at how SPSS can overcome the deficiencies of open-source software alone.

LEARNING CURVE

IBM SPSS Modeler's intuitive interface means users can jump in and start generating insights faster, even if they aren't coding experts. IBM SPSS Modeler provides a simple graphical user interface that supports a variety of data preparation, statistical analysis, machine learning and predictive modeling algorithms. R and Python code runs in the same GUI front-end, alongside all of the other functions and features that are already provided. Modeler also provides out-of- the-box access to select Python algorithms without having to write code.

DATA CONNECTIONS

Using open-source analytics software on your existing database can require a great deal of time and effort. Plan to write volumes of code, implement packages and maybe even employ Java. IBM SPSS Modeler simplifies data access. It can be used with SQL, Oracle, SAP, IBM Netezza®, DB2® commercial databases and more. IBM SPSS Modeler can read text input, spreadsheets, SAS files and other formats. IBM SPSS Modeler can also access data from Hive, BigSQL installed on Hortonworks HDP and Cloudera Impala. With the connection to relational databases, Hive and BigSQL, Modeler can push back data preparation steps back to the data source resulting in less data movement and quicker results. Wizards with prebuilt connectors access the data, which removes the excessive time-consuming burden of extracting, transforming and manipulating data before analysis. IBM SPSS Modeler provides powerful data manipulation techniques, encapsulated in point and click interfaces. Users can transpose, check and reformat data. IBM SPSS Modeler also provides automatic data preparation, which optimizes data for predictive modeling with a single click.

DATA MANAGEMENT

There are many advantages to using IBM SPSS Modeler and IBM SPSS Analytic Server in combination with open source databases. With processing distributed in the Hadoop environment, there's no need to move data, which leads to optimal performance on large volumes of varied data. Your organization can analyze enormous amounts of data by pushing the analytics to the data rather than taking the data to the analytics.

The combined IBM SPSS Modeler and IBM SPSS Analytic Server solution also enables you to access data from Hadoop and combine it with external data from other sources. In addition, you can use the IBM SPSS Modeler interface to add data sources to a Hadoop distribution. This and the accessibility to data wherever it is stored can help your business users get a complete picture and analyze all the available relevant data. And by integrating IBM SPSS Modeler with Spark, you can better extract value from big data, conduct deeper analyses and deliver results faster, all while reducing the time and effort required for coding. This capability is particularly important for real-time stream processing and machine learning, which requires iterative computation, a task that is normally prohibitively time consuming with massive data.

ANALYSIS

IBM SPSS Modeler contains more than 40 algorithms that were developed by IBM. These should be sufficient for most problems. However there are many more open source algorithms that many data scientists find to be very useful. Some are widely used in many situations while others may be fairly specialized. IBM SPSS Modeler provides two different ways to incorporate open source analysis in the analysis.

First, the software includes a visual access to a number of Python and Spark algorithms that are available "out of the box". A notable entry is XGBoost which is a highly effective algorithm that is frequently a winner in data science competitions. XGboost is available both via Python and Spark – where the Spark implementation can be pushed to Hadoop via IBM SPSS Analytic Server. Other Python and Spark algorithms available natively in Modeler include:

- SMOTE which is an advanced balancing technique
- One-Class SVM which detects if new patterns are present in data
- T-SNE which is a data reduction method that allows easy visualization of clusters
- Random forest in Spark – a widely used decision tree method that works with lower quality data
- K-means in Spark – a very commonly used clustering method

Second, IBM SPSS Modeler allows data scientists to run R or Python code where the Python code can invoke both Python and Spark machine learning libraries. This code can be run by itself – or the data scientist can create an extension. An extension is a custom node that allows non-coders to access open source algorithms on the Modeler GUI. We have developed a number of extensions already – these extensions can be easily installed into IBM SPSS Modeler.

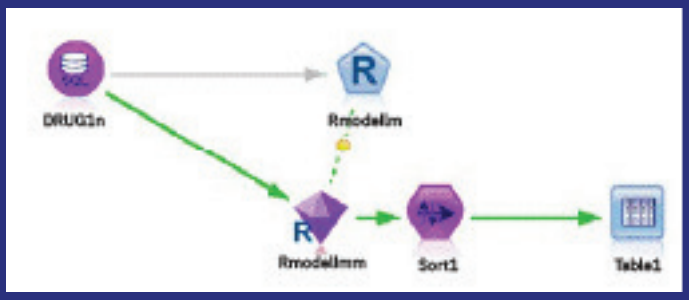

Figure 3: IBM SPSS Modeler is used to natively access data in an SQL database with an R model used to score records. Purple indicates that the analytical step is using SQL pushback, which improves performance by executing memory-intensive tasks in the database itself.

PERFORMANCE

Memory management for analytics applications can be a time-consuming configuration challenge. To make this sort of optimization easier, IBM SPSS Modeler enables you to partition or sample the data passed to your applications (Figure 3). In addition, IBM SPSS Modeler Server is a memory-exploiting technology that can spill analysis over to disk so memory remains available. You can run commands and create objects without a major impact on the overall performance of your data mining and modeling. Also, IBM SPSS Modeler can scale your applications in-database for IBM Netezza and SAP Hana environments, among others. And, with IBM SPSS Analytic Server, it can scale applications in Hadoop as well.

DEPLOYMENT

IBM SPSS Modeler software is a cost-effective way to deploy data models, including big data from Apache Hadoop, and models built in R or Python. IBM SPSS Modeler Gold allows for straightforward deployment of both IBM SPSS Modeler programs accessing any data source as well as open source programs embedded in IBM SPSS Modeler. This deployment does not require recoding and can be added to existing operational processes. Users can integrate predictive intelligence into their interactive, mobile dashboards alongside historical and real-time data views without the need to create or purchase additional software.

COLLABORATION

Most analytics work is a collaborative effort with a number of people contributing to models and statistical analysis. When you use IBM SPSS Modeler software and open-source software together, you'll gain IBM's world-class collaboration capabilities. IBM SPSS Modeler software provides a complete framework for centralizing, securing and automating analytical assets developed with IBM SPSS Modeler. This ensures that predictive models and statistical analysis developed can be shared securely and corporate governance can be applied. With extensions, the data science coder can effectively collaborate with a non-coding colleague. For example, programmers working in Python can create new IBM SPSS Modeler nodes through the SPSS extensions library that exploit algorithms from MLlib and other PySpark processes, then share those nodes with non-programmers. The visual users can then leverage the work of programmers by developing models using R and Python straight from the GUI without having to understand the code underneath.

ENTERPRISE SECURITY

Open-source is for use at your own risk. With every new software release, you're relying on an outside community to ensure each new update delivers the functions it claims to and is free from malicious code. IBM SPSS Modeler is tested rigorously as part of IBM's software QA process. Because the software is from IBM, you do not have to use risky practices that can threaten the security of your environment.

CONCLUSION

Data scientists, business analysts and developers all have different needs from analytics software, but share the need to collaboration and the desire for fast, accurate insights. IBM SPSS software integrates with open source software in a variety of ways that help data-driven organizations work more efficiently.

- Users can run R and Python code, including Python using Spark libraries and code pushed to Hadoop, directly in IBM SPSS Modeler.
- Developers can create a custom GUI node to enable non-coding users to run R and Python code.
- Users gain the ability to push supported functions (including most data preparation functions and some algorithms) into Spark installed in a supported Hadoop cluster when connecting to Hadoop via the IBM SPSS Analytic Server.
- Modeler ships with Python 2.7 and has Python (Jython) scripting to allow users to manipulate stream objects. Modeler also has three algorithms that run on Python directly: XGBoost, SMOTE and one-class SVM.
- An organization can leverage the many benefits of open-source data analysis tools—and surmount many of the drawbacks—by using IBM® SPSS® Modeler software.

For more INFORMATION

Access:
https://www.ibm.com/us-en/marketplace/spss-modeler

ABOUT Data Science

IBM Data Science Platform software delivers data-driven insights that help organizations work smarter and outperform their peers. This comprehensive portfolio includes solutions descriptive and predictive analysis, machine learning and decision optimization. Data Science solutions enable companies to identify and visualize trends and patterns in such areas as customer analytics and operational management that can have a profound effect on business performance. They can compare scenarios; anticipate potential threats and opportunities; better plan, budget and forecast resources; balance risks against expected returns and work to meet regulatory requirements.

By making data science and analytics widely available, organizations can align tactical and strategic decision making to achieve business goals. For more information, see ibm.com/datascience.