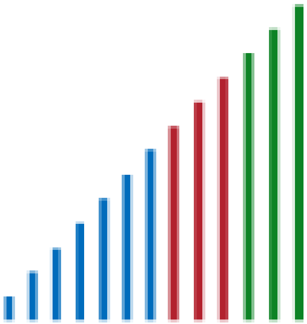

Data Day Seattle 2016

SOFTWARE COMPILATION LEADS TO BUSINESS SUCCESS

By Susan D. Baird-Joshi

September 13, 2016

Introduction

Data handling methods ranging from data analysis and machine learning to real-time streaming and natural language processing occupied the technology focus of Data Day Seattle 2016, held on July 23. But the offerings in open-source software made one theme clear. A compilation of software technologies and business goals can lead to reproducible success.

Abstraction and Flexibility Leads to Reproducible Success

For Dr. Michael Berthold, visual business workflows with reusable components placed the data analysis emphasis where it should be--on the business process. Berthold, co-founder and President of KNIME.com AG, said the concept behind the open-source KNIME analytics platform is to provide reproducible visual workflows that bring business process and data governance to the data points regardless of programming language, data format or technology.

Berthold said there are four key analytics personas: programmer/statistician, data scientist, citizen data scientist and business analyst. Programmers and statisticians create new mathematical algorithms in the form of software libraries and packages, while data scientists use programming languages to combine those algorithms with data. On a more basic level citizen data scientists re-use “analytical best practices” that data scientists codify. When these tools and best practices are in KNIME, the result is a repository of abstracted and encapsulated business subject matter knowledge and tools. Business analysts then use this tool repository to apply their business knowledge to their company's data stores and data dictionary, creating results such as graphs, charts or reports.

Traditional business intelligence (BI) includes descriptive and diagnostic functionality, said Berthold. BI addresses questions about past activity—what happened and why. While advanced analytics also addresses the why, it primarily focuses on two predictive and prescriptive questions--“What is going to happen?” and “How can we make it happen?” In Berthold’s words, KNIME is a platform that “make[s] programmers work with data scientists” to address all those questions.

According to Berthold, KNIME users have several options when completing their data handling workflows: using proprietary workflows and code incorporated into KNIME; writing their own custom code from another language such as R, Python, Java or SQL; and accessing and managing data on a variety of data platforms. Berthold said the product's flexibility enables customers to choose the data stores, visual workflow components and data science tools that are right for them and empowers them to complete a variety of jobs, from running nightly batch jobs for cleaning data to implementing a data analysis and report generator pipeline.

Open-Source and Proprietary Software Combine to Achieve Goals

Dr. Steve Kramer used a variety of proprietary and open source software to accomplish his primary strategy of detecting anomalies in the relationship of social media data. Kramer, founder and Chief Scientist of Paragon Science, said the big question in social networking data was who talked to whom about what. The largest number of anomalies indicated "the most viral topics." He used a process that took the data based on timestamp, broke it into clusters and performed feature vector encoding. From there he detected outliers and performed the analysis at intervals. Initially he used Python for pre-processing and his own proprietary Java code to perform the machine learning. But he said he plans to move to the open source application Apache Spark supported by the Apache Software Foundation.

Kramer used visualization and natural language software to answer key questions, such as what the network of Twitter feeds looked like, what was the tone of written comments and how the dynamic Twitter network changed over time. The Cytoscape open source platform proved valuable in creating a visual network map of the Twitter data feeds regarding the Islamic State of Iraq and Syria (ISIS), the United States 2016 elections and the British exit (BREXIT) vote. By leveraging LIWC software, or Linguistic Inquiry and Word Count, he calculated and assigned an anxiety score for comments based upon a library of words. The KeyLines software developed by Cambridge Intelligence enabled him to incorporate time series animation and visualization into his solutions.

Open Source Workflow Addresses Business and Data Needs

Business outcomes were important factors in choosing how to address data analytics problems, said John Akred, founder and Chief Technology Officer of Silicon Valley Data Science. Akred and his speaking partners, Stephen O'Sullivan and Mark Mims, Vice President of Engineering and Principal Engineer of SVDS, respectively, talked about developing data pipelines with open-source Apache Kafka and Spark. Akred said a roadmap helped developers reach the business destination. During the process, integration was often the most challenging piece and had to be addressed first.

SVDS made technical choices based upon business and project goals. Mims said for automation and testing, writing great requirements is critical. SVDS purposefully chooses to focus on business case integration and utilization and be platform and software agnostic. An example is the data acquisition and ingestion process. O'Sullivan said one key consideration is whether the source pushes the data, such as log or Internet of Things (IoT) data, or the application pulls the data, such as API calls.

The complexity of logic also factored into their decisions about what software to use. O'Sullivan said Apache Storm handles event-based decisions, while Spark handles batch processes. The frequency and size of data packets impacts bandwidth and throughput, driving another decision tree. In summary, the result of the integration of multiple Apache Software Foundation applications was a series of decisions and data handoffs to applications based upon business and performance priorities and the ability of those software pieces to help SVDS accomplish the final goal.

Conclusion

In the past, businesses had to commit to one platform and software house regardless of their business goals. Now, the evolution in intercommunication between proprietary and open-source software and the exponential growth of technical offerings has shifted the focus back to the business use cases, challenges and questions. An emphasis on reusability, interoperability and workflows has created a resurgence in the need for understanding business vertical markets and their unique use cases and data variables. In short, a role reversal has taken place. The business cases are the masters, and the software is the servant.