type="general"

Building A Real Life Security Data Science Team

There is an allure surrounding data science. Those who wield its magical mathematical and technological incantations are treated with a reverence akin to the gurus of early days of UNIX. Yet, this mystique quickly wears off when you move from thinking about building a data science team to actually acting on that impulse.

When Bob had his internal team start their move into security data science it was difficult to resist the urge to spin up a giant, shiny Hadoop cluster and start importing every log from every system into a massive data store. In truth, his team did start down the Hadoop path and found it fraught with peril (and screenfulls warning messages).

Rather than focus on the technology, they stopped and focused on defining what single question (to improve security capability) they would like answered if they had the data. Not five. Not three. One. That single question ended up being a statement: “We want to be able to search for an indicator of compromise (in this case, an IP address or IP/port combination) across **all** our perimeter devices in less than five minutes.” For most organizations—including Bob’s—the total volume of such data would fit well within the category of “medium-sized” (i.e. not “big”) data.

His team focused on using a traditional SQL (MariaDB), NoSQL (MongoDB & Redis), R, Python and JavaScript. Bob had one of his staff attain the University of Washington’s Data Science certificate to ensure there was a healthy mix of formal analytical knowledge and domain expertise on the team. For six long months, they iterated from failure to failure, trying different ways to acquire, clean and store data (they call that *data curating*), and structure schemas and formulate queries. Along the way, they suffered setbacks when log file formats changed without warning, when data access issues cropped up, and when the absolute need for referential metadata reared its ugly head.

Three core principles focused the team. First, master the open source versions of tools before engaging vendors. If you don’t know how the sausage is being made, you really have no idea what’s being done and this is vital when working with real data. Second, follow the mantra of “no single tool; no single database; and, no single approach to solving a problem”. Do not put blinders on because you are either comfortable with certain technologies or have an affinity for a certain tool. Third, failure is expected, but you must learn from each journey down the wrong path, much like a mortar team must adjust angle and deflection after each errant shot.

Ultimately, Bob’s team met the five-minute challenge and has moved on to other questions. Your team—and, just like firing a mortar, it is a team effort—will also be successful if start with a question, be iterative and methodical in your approaches and never stop learning, both from your mistakes and the successes of others.