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 and actually act on that impulse.

When Bob had his internal team start the move into security data science it was difficult to resist the urge to spin up a giant 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 tech, they stopped and focused on defining what single question they would like answered if they had the data. That question ended up being transformed into a statement: “We want to be able to search for an indicator of compromise (in this case, an IP address) across **all** our perimeter devices in less than five minutes.” Said devices were in excess of 200 distinct systems, but even with that scope the total volume of data stood well within “medium-sized” (i.e. not “big”) proportions, even for a large enterprise.

His team focused on using a mix traditional SQL (MariaDB), NoSQL (MongoDB & Redis), R, Python and JavaScript and Bob had one of his staff attain the University of Washington’s Data Science certificate to ensure there was a healthy mix of skillsets on the team. For six long months, they iterated from failure to failure, trying different ways to acquire data, structure schemas and formulate queries to meet the five minute challenge. 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. Second, follow the mantra of “no single tool; no single database; and, no single approach to solving a problem”. Third, failure is expected, but you must learn from each journey down the wrong path.

Ultimately, Bob’s team met the five-minute challenge and has moved on to other questions. Your team—and, 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.