CTOlabs.com

White Paper:
Big Data Solutions For Law Enforcement

June 2012

A White Paper providing context and guidance you can use

Inside:

- **Big Data in Policing**
- **Use Cases Of Interest to Law Enforcement**
- **Lessons to learn for your enterprise deployment**

Big Data Solutions For Law Enforcement

Big Data, the data too large and complex for your current information infrastructure to store and analyze, has changed every sector in government and industry. Today's sensors and devices produce an overwhelming amount of information that is often unstructured, and solutions developed to handle Big Data now allowing us to track more information and run more complex analytics to gain a level of insight once thought impossible.

The dominant Big Data solution is the Apache Hadoop ecosystem which provides an open source platform for reliable, scalable, distributed computing on commodity hardware. Hadoop has exploded in the private sector and is the back end to many of the leading Web 2.0 companies and services. Hadoop also has a growing footprint in government, with numerous Hadoop clusters run by the Departments of Defense and Energy, as well as smaller deployments by other agencies.

One sector currently exploring Hadoop is law enforcement. Big Data analysis has already been highly effective in law enforcement and can make police departments more effective, accountable, efficient, and proactive. As Hadoop continues to spread through law enforcement agencies, it has the potential to permanently change the way policing is practiced and administered.

Apache Hadoop and Cloudera

Apache Hadoop is a project operating under the auspices of the Apache Software Foundation (ASF). The Hadoop project develops open source software for reliable, scalable, distributed computing. Hadoop is an exciting technology that can help analysts and agencies make the most of their data. Hadoop can inexpensively store any type of information from any source on commodity hardware and allow for fast, distributed analysis run in parallel on multiple servers in a Hadoop Cluster.

Hadoop is reliable, managing and healing itself; scales linearly, working as well with one terabyte of data across three nodes as it does with petabytes of data across thousands; affordable, costing much less per terabyte to store and process data compared to traditional alternatives; and agile, implementing evolving schemas for data as it is read into the system.

Cloudera is the leading provider of Hadoop-based software and services and provides Cloudera's Distribution including Apache Hadoop (CDH) which is the most popular and most efficient way to implement Hadoop. CDH is an open system assembled from the most useful projects in the Hadoop ecosystem bundled together and simplified for use. As CDH is available for free download, it's a great

place to start when implementing Hadoop, and Cloudera also offers support, management apps, and training to make the most of Hadoop.

Big Data in Policing

Law enforcement agencies generate and handle a tremendous amount of information. A department can get millions of calls for service a year, create thousands of police reports, and examine months of video and audio. When that information is analyzed, the data can grow exponentially.

The largest and best-funded agencies such as the LAPD handle that information with special departments, contracts with firms like IBM, and partnerships with universities. Smaller departments, however, cannot afford supercomputers, command centers, mathematicians and computer scientists on staff, proprietary software suites, or even numerous analysts and advanced analytic software. As a result, many departments only collect information structured and small enough to fit in a spreadsheet then do nothing more complicated than sums and averages with it to determine crime rates. These departments miss out on revolutionary changes in the way police work is conducted and, ironically, potential cost savings. Nationwide, agencies and departments have to reduce their resources and even their manpower but are expected to continue the trend of a decreasing crime rate. To do so requires better service with fewer resources. Big Data analytic solutions like Hadoop have been proven to deliver this and, as Hadoop is open source and runs on cheap, commodity hardware, launching a Hadoop cluster to store and analyze crime, equipment, and personnel data is within reach of most departments.

How Hadoop Can Help Law Enforcement Agencies

By keeping and collecting more data, long term trends and hidden patterns emerge that allow organizations to better target both waste and deficiencies. Private sector industries such as retail have used Hadoop and Big Data analysis to cut their costs dramatically while better serving their customers through smarter resource management, and law enforcement agencies are beginning to do the same.

Take, for example, dividing a city into policing districts and beats, a process every department has to conduct regularly due to crime trends and changing demographics. Patrol is the backbone of policing, and these divisions determine where patrol officers are allocated, with officers typically patrolling, answering calls for, and staying within their beat throughout their shift. Departments that do not use data analysis at all divide their city into equal portions, which is problematic as it assigns the same manpower to high crime and quiet areas. Others use crime statistics to draw up boundaries, but given

the many factors involved, patrol areas are rarely a great fit. For example, distribution of officers can have an effect on crime, so a new beat map may change the very data a department is analyzing. Also, if size is predicated only on crime, quiet beats may grow too large to effectively patrol. An officer's ability to receive backup should also be considered, as well as contingencies for reshuffling beats when officers have days off and get sick or injured. In a relatively large department, analysts may look at millions of calls for service over several years to best plan patrols, but type of crime is also a factor. Fairly distributing resources also means screening for biases and confounding variables, such as wealthier or more politically connected neighborhoods having a louder voice and getting more attention than those really in need.

Hadoop provides a platform to solve all of those problems. It makes storing historical records, even phone calls and videos, cheap, as they can be kept on commodity hardware. It lets you analyze them after the fact in any way you want, as Hadoop works with raw data and implements a schema-on-read, adding whatever structure you need whenever you need it. It can also run similar analysis for other resources to allow an agency to run more smoothly at a lower price, for example tracking gas consumption by cruisers, rates of ammunition used at the firing range, serial numbers on stolen goods, and paper and form usage at stations.

Predictive Policing

Police Departments nationwide have been using data and statistics to drive policing since the 90s in an approach founded by the NYPD named CompStat, credited with dramatic reductions in crime and increases in efficiency. CompStat, a process and philosophy rather than a single technology or software, uses databases and GIS to record and track criminal and police activity as well as to identify areas that are lagging or need more attention. While it provides much more information than "primal policing", CompStat has advanced little beyond simple spreadsheets and mapping software. Inspired by recent innovations in Big Data and Apache Hadoop as well as businesses like Walmart or Amazon using analytics to determine future demand, departments across the country and worldwide are looking to take this approach to the next level and go from tracking crime to predicting it.

The first department to adopt this strategy was Santa Cruz through their city-wide 6 month Predictive Policing Experiment. Forced by a declining budget and crime wave to do more with less, the department signed on to work with researchers at UCLA to test a new method of modeling crime.

UCLA mathematician George Mohler noticed that, over time, crime maps resemble other natural phenomenon and modified algorithms used to predict aftershocks to instead predict future property crimes from past data.

Past crimes can be predictive of future crimes because they indicate that an environment is target-rich, convenient to access for a criminal, vulnerable, or simply seems like a good place to strike due to a pattern of crime and poor control. To predict the most likely type, location, and time of future crimes, Mohler had to compare each past crime to the others and generate a massive amount of metadata. For the Santa Cruz Experiment, he went back to 2006, looking at roughly 5,000 crimes requiring 5,000! or 5,000 x 4,999 x 4,998… comparisons. When he compared his method to traditional CompStat maps for the LAPD's archives, he found that it predicted 20 to 95 percent more crimes. The experiment was recently concluded, and the department believes that its predictive policing program was a success. Despite having fewer officers on the force, SCPD reversed the crime wave and lowered crime by 11% from the first half of the year to 4% below historical averages for those months.

While predictive policing is showing promise and results, the practice is still in its infancy with plenty of room to grow. Much more metadata can be generated and factors included into the predictive algorithms. For example, Santa Cruz could only predict property crime, as violent crime depends less on targets and opportunities and more on events and interpersonal interactions. In business and counter-terrorism, however, tools like social network analysis and social media monitoring have been used successfully to get a better feel for social dynamics. Some departments such as Richmond, Virginia, have also plugged additional data into their analysis beyond other crimes, such as store, bar, housing, and ATM locations or recent events to find factors driving crime. As predictive policing gets more attention and is adopted more widely, we can expect to see these and other Big Data solutions applied to law enforcement more widely.

Better monitoring and accountability

Quality control and accountability is important for law enforcement to prevent abuse of power and insure that effective strategies and tactics are being implemented. Poor service can cost citizen and police lives, encourage crime, and hurt community relationships which are vital to effective policing, but it's difficult to get an objective picture of police and citizen interactions.

One example comes from a recent report on the Albuquerque Police Department (APD) by the Police Executive research Forum. On paper, all indications showed that police use of force should be decreasing. Crime and violent crime was declining, training had been implemented to reduce officer

involved shootings, a special unit developed to deal with unstable, disturbed, or mentally ill suspects, and even a risk monitoring program put in place to identify dangerous officer behavior, but police use of force continued to increase and was much higher than the national average.

APD had a wealth of information surrounding each use of force incident, which could include the criminal record of the suspect, biometric data, police cruiser dashboard video and audio of the encounter, surveillance from nearby cameras, the officer's report, proceedings from the numerous reviews that follow use of force, relevant newspaper articles and testimonies, the conversation with the police dispatcher, and court records. Only a small fraction of that information was being systematically analyzed to find red flags, determine who was at fault, or review policy and training surrounding officer use of force. Using Hadoop, Big Data analytics, and machine learning, all of this data could have been stored and analyzed, finding new patterns and unknown factors to explain the anomalous increase in violence. By throwing all available data at the problem, correlations that analysts would otherwise omit may come to light.

Another element of accountability is better data on officer performance. With CompStat, officers and detectives began to be evaluated by basic statistics such as number of arrests and percentage of cases closed. Using such crude metrics, however, can provide misleading results. The best patrol officers prevent crime in their beat, which makes for safer neighborhoods but less impressive arrest statistics. In investigations, different crimes will have different closure rates, with over 60% of homicides solved, sometimes much higher than that, but only about 10% of burglaries closed. By taking in all relevant data such as closure rates, crime rates, area averages, time of shift, past performance, and possibly other factors that the department had previously thought were insignificant, and analyzing it, the department can rate its effectiveness and the performance of individual officers much more accurately. In this way, law enforcement agencies can better reward performance and evaluate policies, initiatives, and strategies.

More Reading

For more federal technology and policy issues visit:

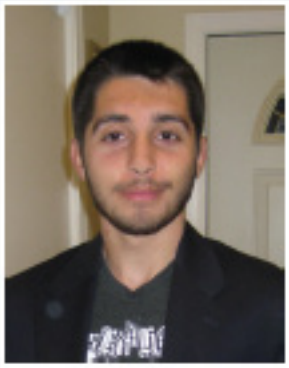
- **CTOvision.com-** A blog for enterprise technologists with a special focus on Big Data.
- **CTOlabs.com -** A reference for research and reporting on all IT issues.
- **J.mp/ctonews** - Sign up for the Government Technology Newsletters.
- **Cloudera.com** - Download Cloudera's Distribution featuring Apache Hadoop to start your Big Data project today.

Why Cloudera

Cloudera's Distribution Including Apache Hadoop is the best place for an organization to start exploring Hadoop. Not only is it the most popular and widely deployed Hadoop distribution, it is completely free and all open source so there is no vendor lock-in. CDH provides an integrated Hadoop stack with Hadoop Distributed File System and MapReduce, the heart of the Hadoop infrastructure along with Hadoop databases and programming languages, log collector and aggregator, machine learning algorithm library. All parts of Cloudera's Distribution are tested to work together and with popular software and operating systems.

If an organization is looking to scale up, Cloudera Enterprise offers a management tool, automated, wizard-based installation and configuration, monitoring, diagnosis, security, and optimization. Cloudera also offers support, training, and professional services such as use case discovery, deployment, and proof of concept. For agencies that already have extensive information infrastructures, Cloudera's Distribution is poised for poised for enterprise systems integration because of its inclusion of numerous hadoop oriented integration technologies. Cloudera is a good fit for law enforcement agencies of all sizes looking to better capture, store, and analyze Big Data from the first experimental cluster to an enterprise-wide deployment.

About the Author



Alex Olesker

Alexander Olesker is a technology research analyst at Crucial Point LLC, focusing on disruptive technologies of interest to enterprise technologists. He writes at http://ctovision.com.

Alex is a graduate of the Edmund A. Walsh School of Foreign Service at Georgetown University with a degree in Science, Technology, and International Affairs. He researches and writes on developments in technology and government best practices for CTOvision.com and CTOlabs.com, and has written numerous whitepapers on these subjects.

Contact Alex at AOlesker@crucialpointllc.com

For More Information

If you have questions or would like to discuss this report, please contact me. As an advocate for better IT in government, I am committed to keeping the dialogue open on technologies, processes and best practices that will keep us moving forward.

Contact:
Bob Gourley
bob@crucialpointllc.com
703-994-0549

CTOlabs.com