Virtualized Data Lakes

Federated Big Data Analytics of Disparate Data Stores

Steven G. King
Regional Portfolio Manager
The MITRE Corporation
McLean, Virginia, USA
Email: sking [AT] mitre.org

Abstract— **Analogous to the move away from comprehensive data warehouses in favor of federated searches to access content from host sites prior to the advent of cloud computing, a federated data lake offers the benefits of applying advanced data analytics and data science to business data without requiring a monolithic data lake, whether in a cloud environment or in proprietary data warehouses. Using two case studies, this article illustrates the benefits of this novel approach to big data analytics. Federated searching costs significantly less than data warehousing, improves data quality, speeds refresh rates, and improves database stability. The advent of artificial intelligence and machine learning facilitate big data analytics on structured and unstructured datasets across multiple databases.**

Keywords- Data lake, Data warehousing, Federated search, Cloud, Government, Artificial intelligence, Machine learning, Deep learning, Cognitive computing, Data science.

I. Introduction

Government and industry spend considerable time and money creating cloud-based data warehouses. As a result, many have encountered significant problems maintaining data quality, which sometimes leads to corporate liability. While many information technology (IT) professionals debate the top-down (advocated by Bill Inmon) versus bottom-up (advocated by Ralph Kimball) approach to data warehousing, many new data consolidation efforts are skipping the entire debate by going virtual via federated searching. This paper details two big data analytics use cases that have successfully implemented federated searching in favor of traditional data lakes.

II. Background

"A data lake is a storage repository that holds a vast amount of raw data in its native format, including structured, semi-structured, and unstructured data. The data structure and requirements are not defined until the data is needed," according to Anne Buff of SAS Best Practices team [14]. A variety of business intelligence tools are applied to data lakes. A data lake requires the periodic extraction, transformation, and loading of data into the data warehouse. This inevitably leads to a lag between the data in the transaction database(s) and the data lake. In 2014, O'Leary examined many of the risks of the then-emerging data lake concept, including the requirement for increased data governance [15].

A. Big Data Analytics

O'Leary investigated using different artificial intelligence and crowdsourcing (human intelligence) applications in a data lake in order to integrate disparate data sources, facilitate master data management and analyze data quality [15]. What wasn't considered, however, was the ability of federated searching to rapidly search multiple disparate databases and seamlessly apply big data analytics to disparate databases. Gartner defines federated search as "passing a user's search to one or more other search engines, retrieving the results and presenting them to the user" [5]. Cognitive computing capabilities are the game changer when analyzing unstructured data. For example, IBM's Watson Machine Learning allows users to connect more than 30 different types of data stores to create, train, and deploy machine learning and deep learning models [16].

III. Critical Infrastructure Case Study

The US Department of Homeland Security (DHS) is responsible for leading the national effort to identify and protect critical infrastructure in the United States [8,4,12]. This requirement is established by several federal laws and national strategy documents. For example, the *National Strategy for Homeland Security*, which created DHS, specifies that critical infrastructure protection is one of four significant DHS mission goals [1]. In a recent speech, DHS Under Secretary Christopher Krebs said, "As a nation, we must work together seamlessly to share information, plan, train and respond to cyber and physical threats." [4]. Critical infrastructure includes physical, cyber, and human assets necessary for the functioning of the country. Examples of critical infrastructure include large dams, telecommunications switching centers, iconic bridges, chemical manufacturing facilities, nuclear power plants, large drinking water and water treatment facilities [8].

In order to carry out this important mission, DHS was directed in Section 1001 of the *Implementing Recommendations of the 9/11 Commission Act of 2007* to maintain and use a database to catalog the nation's critical infrastructure [12]. Initially, DHS created a National Asset

Database (known as the NADB), which was designed as a data warehouse intended to serve as a repository for information about the nation's critical infrastructure. The DHS program manager for the National Asset Database, Tim Huddleston, embarked on a fairly straightforward approach – not unlike the approach many businesses and organizations take in similar situations – of data warehousing. Huddleston found a variety of good to excellent data sources throughout the federal government [10]. To fill the gaps of information about critical infrastructure assets, DHS reached out to the states and territories in the US to gather the additional information. All of this information about critical infrastructure, from states, territories, and other federal agencies, was placed into the National Asset Database.

As an example, one of the 16 critical infrastructure sectors encompass dams, locks, and levees [3]. Huddleston found that the US Army Corps of Engineers and the US Department of Interior's Bureau of Reclamation both own and operate dams, locks, and levees. Both of these federal agencies had considerable data about the critical infrastructure they owned and/or operated. Additionally, the Federal Emergency Management Agency (FEMA) Dam Safety Office and the Federal Energy Regulatory Commission (FERC) both had considerable information about many of those critical infrastructure assets and tens of thousands of other dams [10]. To complicate matters further, most states and territories in the US maintained lists of privately-owned dams and levees, which are most commonly used in agriculture. Huddleston created the National Asset Database using the top-down approach Inmon espoused and began collecting data for the newly created data warehouse [7,10,11]. This approach is not unlike what IT professionals have been doing for years when creating a data warehouse.

A. Data Lake Failure

Two years later, congressional auditors reported that the National Asset Database had "created confusion in Congress and the media" [1]. Worse than the public criticism, Huddleston found that the National Asset Database suffered from many of the same data quality problems of many other data warehouses. Information about the Hoover dam, for example, seemed perennially out of date [10]. As a major national landmark, various federal agencies conducted frequent assessments of the dam and updated their agency's database. When DHS protective security advisors conduct a security assessment at the Hoover dam, their reports, photographs, and geospatial coordinates are entered into their agency's database. Inspections of the Hoover dam electricity generating systems conducted by FERC are placed into their agency database. This process is repeated by federal, state, and local agencies across the country for tens of thousands of critical infrastructure assets nationwide. A noticeable lag occurs between the updates to the individual agency's database and the National Asset Database.

In addition to the lag time, the quality of the data in the National Asset Database also suffered with this standard data warehouse approach. Agencies felt that their database contained the most accurate information. It is likely this perception arose from two causes. First, when people in one agency were notified that their database had been updated, they frequently immediately searched the then-new National Asset Database. Because of the lag time those users found dated information in the National Asset Database. Second, people associated with one agency tend to believe in the accuracy of information from their own agency when it conflicts with information from another agency. This confirmation bias leads people to view the other agency's data as incorrect when conflicting information is presented [13].

B. The Federated Approach

To overcome these and other problems with this traditional approach to data warehousing, Huddleston decided to start over. Instead of copying data from one federal or state-owned database, Huddleston embraced federated searching. Congressional auditors described the new system in their follow up report. "DHS, through its Office of Infrastructure Protection (IP), established an Infrastructure Data Warehouse, which includes infrastructure data from a variety of federal, state, and local sources and other authoritative open source infrastructure databases" [7]. The new approach was to find sources (i.e., databases) of critical infrastructure assets that are maintained by other agencies, which could be searched by the new National Asset Database – now named the Infrastructure Data Warehouse (IDW) [7]. The IDW evolved from traditional data warehouse technologies to federated virtual data warehouse technologies.

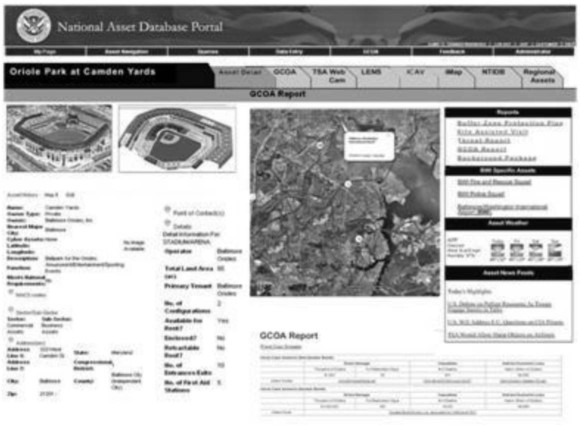


Figure 1: Image of Infrastructure Data Warehouse screen (fictional data)

Using a federated approach, the IDW presents users with the most current data available from all of the various databases that house critical infrastructure information at state and federal agencies. This federated search approach reduces duplication of efforts that would otherwise be required to maintain two separate data databases, as was required under the traditional approach used for the National Asset Database. Any inconsistencies in information about critical infrastructure were reported to the user, which allowed the user to view the conflict rather than allowing the system to choose one over the other. This approach actually improves data quality, although it seemed counterintuitive. The federated approach, used by the IDW, does not require the organization (DHS/IP) to maintain and update a second database from which they previously had to extract, transform, and load data from an external agency's database. This reduces the effort and maintenance of the data in the system and requires less disk space to store duplicated data.

IV. Nationwide Law Enforcement Case Study

Every day, law enforcement officers observe behaviors that are suspicious or receive such reports from concerned citizens. What might not seem significant (for instance, taking the picture of a ferry during loading), when combined with other actions and activity, may become a composite indicating the possibility of criminal or even terrorist activity. Traditionally, street officers have had little to do with counterterrorism; but, after the terrorist attacks of September 11, 2001, it became obvious that al Qaeda members had prepared not only in far-off Afghan training camps but also in Minnesota and flight schools in Florida. In today's policing, "connecting the dots" of suspicious activity before an incident occurs is an integral and imperative job for America's law enforcement.

The challenge lies in the connection process: spans of time and space, plus the multiple levels of agencies that gather intelligence and suspicious activity data have made linkages extremely difficult. But with the creation of state and local fusion centers, law enforcement officers at all levels of government have a means to share information, analyze data for clues, and run-down reports of suspicious packages, all in an effort to detect and prevent terrorism and other criminal activity. Although gathering, storing, and sharing intelligence information has had stringent oversight, resources, and legislation, suspicious activity reports (SARs) and their exchanges have previously lacked the same level of guidance [12]. To address this gap, the *2007 National Strategy for Information Sharing* calls for the establishment of a "unified process for reporting, tracking, and accessing [SARs]," in a manner that rigorously protects the privacy and civil liberties of Americans — what is now the Nationwide SAR Initiative (NSI).

The NSI is an outgrowth of a number of separate but related activities over the last several years that respond directly to the mandate in the *2007 National Strategy for Information Sharing*. The long-term goal is that most Federal, State, local, and tribal law enforcement organizations participate in this standardized, integrated approach to gathering, documenting, processing, analyzing, and sharing information about suspicious activity that is potentially terrorism-related. In addition to government agencies, the private sector and foreign partners are also potential sources for terrorism-related SARs.

A. The Federated Approach

The NSI is a coordinated effort that leverages and integrates all SAR-related activities into a unified nationwide manner. The NSI Program Management Office initially considered creating a data warehouse to store all of the SARs in a single secure data center. With so many strong counter-terrorism, intelligence centers, and law enforcement agencies creating SAR data repositories, the NSI Program Management Office opted to implement a federated approach rather than forcing organizations to transition to a new system [9].

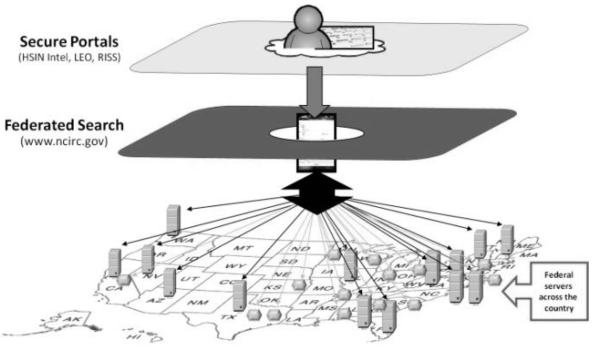
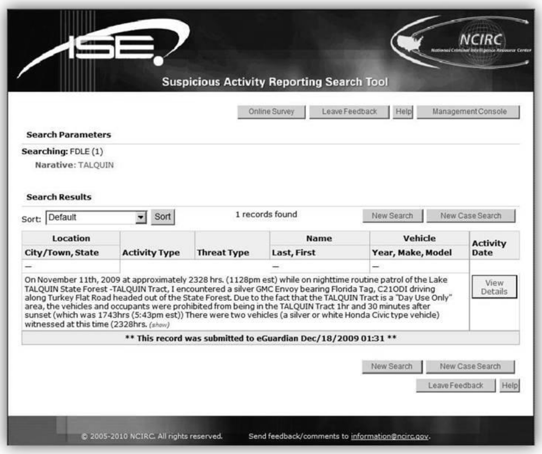


Figure 2: NSI Federated Environment

There are currently ten sites online, with twelve more currently in process. Once each site installs and tests the necessary equipment, receives appropriate training, and has a privacy policy, it begins sharing information in the NSI Federated Environment. The NSI Federated Environment has become so successful that the FBI has mandated the sharing of all unclassified SARs with local, state, federal, and tribal partners via the federated environment.

The benefits of this federated approach are numerous. The ability to allow existing fusion centers and law enforcement agencies who have already built their own data marts to continue to operate as they have in the past while gaining all of the benefits of having access to all of the other SAR databases across the country. Intelligence analysts and investigators can search all other organizations' SAR databases via a single web portal (see figure 3) with the same number of keystrokes necessary to search their own in-house SAR database.



Suspicious Activity Reporting Search Tool

Search Parameters

Searching: FDLE (1)
Narative: TALQUIN

Search Results

Sort: Default

1 records found

Location			Name	Vehicle	Activity Date
City/Town, State	Activity Type	Threat Type	Last, First	Year, Make, Model	
—			—	—	
On November 11th, 2009 at approximately 2328 hrs. (1128pm est) while on nighttime routine patrol of the Lake TALQUIN State Forest -TALQUIN Tract, I encountered a silver GMC Envoy bearing Florida Tag, C21ODI driving along Turkey Flat Road headed out of the State Forest. Due to the fact that the TALQUIN Tract is a "Day Use Only" area, the vehicles and occupants were prohibited from being in the TALQUIN Tract 1hr and 30 minutes after sunset (which was 1743hrs (5:43pm est)) There were two vehicles (a silver or white Honda Civic type vehicle) witnessed at this time (2328hrs. *(show)*					
** This record was submitted to eGuardian Dec/18/2009 01:31 **					

© 2005-2010 NCIRC. All rights reserved. Send feedback/comments to information@ncirc.gov.

Figure 3: NSI SAR web search page

Because the NSI Program Management Office implemented a federated approach, organizations across the country physically house their own server hardware, which provides a tremendous degree of comfort and control over the SAR data they collect. Given the diverse organizations involved in this effort at the federal, state, local, and tribal level, the security of controlling databases at each individual organization provides tremendous confidence for each organization to work together on this effort.

CONCLUSION

These two big data analytics projects purposely and successfully implemented federated searching in favor of traditional data lakes. The question for academics is whether these examples represent a trend toward federated searching of disparate data marts and away from comprehensive data lakes. For IT professionals this paper provides several reasons to consider implementing a federated approach in favor of traditional data lakes, which would not have been effective for big data analytics before the availability of commercially available cognitive computing capabilities and deep learning models.

**Disclaimer:* The author's affiliation with The MITRE Corporation is provided for identification purposes only, and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions, or viewpoints expressed by the author. This article is ‘Approved for Public Release; Distribution Unlimited. Case Number 18-2735.'

REFERENCES

[1] Congressional Research Service, Critical Infrastructures: Background, Policy, and Implementation, updated October 10, 2008, p. 28–29.

[2] Department of Homeland Security Office of Inspector General (2009). Efforts to identify critical infrastructure assets and systems, OIG-09-86. Government Printing Office: Washington, DC. Retrieved from http://www.dhs.gov/xoig/assets/mgmtrpts/OIG_09-86_Jun09.pdf

[3] Department of Homeland Security (2009). National Infrastructure Protection Plan, Government Printing Office: Washington, DC

[4] Krebs, Christopher (2018). Secretary Nielsen Addresses the 2018 Critical Infrastructure Summit, US Department of Homeland Security: Washington, DC. Retrieved from https://www.dhs.gov/blog/2018/03/02/secretary-nielsen-addresses-2018-critical-infrastructure-summit

[5] Gartner (2009). Case study: Federated Search. ID Number G00165211

[6] Gartner (2009). *Taxonomy and Definitions for the Multienterprise/B2B Infrastructure Model*. ID Number: G00166095.

[7] General Accountability Office (2009). Transportation security: Comprehensive risk assessments and stronger internal controls needed to help inform TSA resource allocation. GAO-09-492, p. 13. Government Printing Office: Washington, DC. Retrieved from http://www.gao.gov/new.items/d09492.pdf

[8] Homeland Security Council (2007). *National Strategy for Homeland Security*, Government Printing Office: Washington, DC

[9] Hong, David, personal communication, February 23, 2011.

[10] Huddleston, Timothy, personal communication, February 12, 2011.

[11] Inmon, W.H. (2005). Building the Data Warehouse, 4th ed. Wiley & Sons.

[12] Public Law 110–53 (2007), Implementing Recommendations of the 9/11 Commission Act of 2007, Government Printing Office: Washington, DC

[13] Trope, Yaacov; Bassok, Miriam (1982), "Confirmatory and diagnosing strategies in social information gathering", Journal of Personality and Social Psychology (American Psychological Association) 43 (1): 22–34, doi:10.1037/0022-3514.43.1.22.

[14] Dull, Tamara (2015). Data Lake vs Data Warehouse: Key Differences. KD Nuggets. Retrieved from https://www.kdnuggets.com/2015/09/data-lake-vs-data-warehouse-key-differences.html

[15] O’Leary, Daniel E. (2014, Nov 7). “Embedding AI and Crowdsourcing in the Big Data Lake,” *IEEE Intelligent Systems*. DOI: 10.1109/MIS.2014.82

[16] IBM (2018). Watson Machine Learning. Retrieved from https://www.ibm.com/cloud/machine-learning.