IT Data Mining Tool Uses in Aerospace

Gilena A. Monroe, Kenneth Freeman, Kevin L. Jones
NASA Ames Research Center
Moffett Field, CA, 94035
Gilena.A.Monroe@nasa.gov, Kenneth.Freeman-1@nasa.gov, Kevin.L.Jones@nasa.gov

Abstract - *Data mining has a broad spectrum of uses throughout the realms of aerospace and information technology. Each of these areas has useful methods for processing, distributing, and storing its corresponding data. This paper focuses on ways to leverage the data mining tools and resources used in NASA's information technology area to meet the similar data mining needs of aviation and aerospace domains. This paper details the searching, alerting, reporting, and application functionalities of the Splunk system, used by NASA's Security Operations Center (SOC), and their potential shared solutions to address aircraft and spacecraft flight and ground systems' data mining requirements. This paper also touches on capacity and security requirements when addressing sizeable amounts of data across a large data infrastructure.*

Keywords: Data mining, flight systems, avionics, Splunk, aerospace.

1 Introduction

Processing voluminous data sets is a concern shared by many across the data mining population. This issue is especially true for aerospace entities that collect large amounts of data from numerous system sources. NASA, for example, has the task of harnessing and analyzing petabytes of Earth science data that it collects from satellites, sensors, and models [1]. As advances in flight technology systems are made, though, the need to efficiently handle those systems' data output increases. Managing constant streams of data and making large amounts of data useful, and beneficial, to an organization can prove to be a formidable task. In order to meet this task, effective data mining systems and practices, above algorithm-only development solutions [2], must be introduced.

While tools and systems exist for collecting and processing aerospace data, there are non-aerospace specific tools and systems that can offer additional data mining support to meet the ever-growing needs of data handling. Though, in some cases, utilizing commercial-off-the-shelf/government-off-the-shelf (COTS/GOTS) options can be more costly than in-house custom code products [2], sometimes, employing a combination of the two may provide great benefit.

NASA's Security Operations Center (SOC) uses the Splunk engine to assist in its goal of protecting NASA's data and information technology (IT) infrastructure. Splunk offers the SOC a web interface to its data log aggregation system, with enhanced capabilities such as command-line searching, monitoring and alerting, robust reporting with charts and dashboards, and application development for specific, targeted data environments, all of which can be beneficial to aerospace data mining requirements. Splunk also provides built-in security functions to help protect secure data and to enforce security policies that may already be in place.

The paper begins by highlighting some data mining tools and systems that are currently in use by both aerospace and IT organizations along with some of their limitations in mining large data sets. Next the paper describes particular Splunk uses for aerospace data collection and processing. Finally, the paper concludes with a summary and other proposed future uses for the Splunk system.

2 Overview of Data Mining Systems

The aerospace field uses various systems to warehouse and process data that are collected from aircraft and spacecraft systems. A few of these NASA systems include the Surface Operations Data Analysis and Adaptation (SODAA) tool and the Discovery in Aeronautics Systems Health (DASHlink) system, both of which provide many uses and benefits in mining data. They also have some limitations with regard to certain functionalities such as real time data processing, continuous monitoring and alerting, and application development and automated reporting. Some of these benefits and limitations are detailed below.

The Splunk system, used by NASA's SOC, serves many purposes to IT data handling, including distributed searching across multiple host locations, custom dashboard development, and security incident handling.

2.1 SODAA

The main purpose of the SODAA tool is two-fold, to support analysis of airport surface and terminal airspace operations and to develop and maintain airport adaptations for decision support automation such as the Surface Management System (SMS) [3]. The data analysis portion of the tool allows it to ingest aircraft position data, store it in a database, and compute additional derived data elements also stored in the database. SODAA uses an on-line data warehouse to simplify access to airport surface data [4].

The SODAA Client allows users to submit basic data queries, where the results of queries can be plotted on a map of the airport, graphed, displayed in a table, or exported to a file for additional analysis outside of the SODAA tool. Figure 1 shows an example of aircraft data graphed on a display map in SODAA. SODAA does not, however, permit users to search available data sets in real time. While the system does offer some alerting capabilities, it does not offer the ability to continuously monitor data and provide alerts.

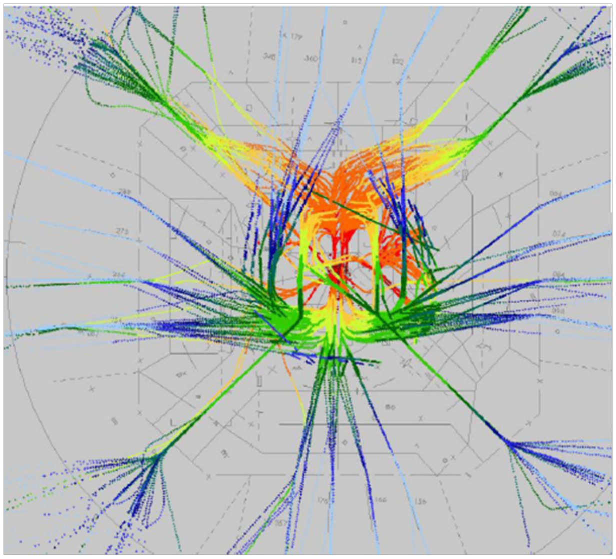


Figure 1. SODAA map display of plotted aircraft data.

2.2 DASHlink

DASHlink is a social networking site for persons/individuals with an interest in systems health management, data mining, or related fields. It serves as an online resource to collaboratively distribute information on systems health algorithms, data, and research. DASHlink is an interactive forum for users to share project information, data, and software applications quickly.

Currently, though, DASHlink use is restricted to NASA employees or NASA civil servant-sponsored public registrants. The site also does not offer any dynamic application development for individualized data needs, nor does it contain any automation elements for ease of use.

2.3 Splunk

Splunk is a dynamic, scalable IT data engine. Splunk collects and indexes both live and historical data and allows users to search, report, monitor, and analyze it in real time or not [5]. The tool reads data as a string and accepts data from any source or format, permitting users to search and locate information across various systems. Data can be given more significance by naming and tagging data fields and data points and fields. The monitoring and alerting functions allow direct communication, via e-mail message or web feed formats such as Really Simple Syndication, or Rich Site Summary, (RSS) notifications, to be sent to users based on conditions set on the data. Splunk can also be integrated with other tools, such as other NASA security-related systems, to increase data viability through correlation and relationship matching. The tool is used in a wide range of arenas, including private companies, government agencies, and individual customers.

Potential limiting factors may include long-term license maintenance costs, especially for extraordinarily large data sets, as well as ongoing interaction with the vendor for license increases, troubleshooting, or system upgrades.

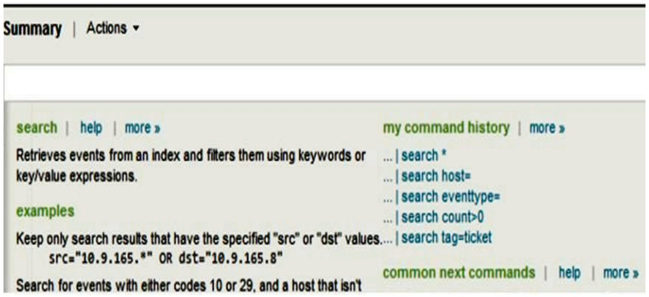
3 Splunk for Aerospace

Splunk has multiple uses that may help benefit the data processing, distribution, and storage requirements from aerospace flight and ground technology systems. This section details how Splunk can meet some of the needs of aerospace data mining that may be limited with tools similar to SODAA and DASHlink, including searching capabilities, alerting functions, reporting methods, and application development.

3.1 Searching

NASA’s SOC uses Splunk’s search feature in numerous ways, but is used primarily to investigate IT security incidents. This capability allows staff to use common search terms to access specific firewall data or even retrieve all data related to a specific IP address.

The search feature may prove helpful to searching flight system data. By typing a simple search command into the Splunk search assistant, shown in Figure 2, a user can quickly retrieve aircraft flight data, such as aircraft type, flight ID, arrival and departure information, and routing information. The search feature also uses Boolean expression to filter data and only return wanted information. For example, a search of aircraft types on a particular flight route from Los Angeles International Airport to New York LaGuardia International Airport can, at once, also include a subsearch to return the actual arrival runway assignments for those aircraft types on that particular route. Quick, easy access to this type of information can prove useful in the optimization process of both air and ground operations. This subsearching capability can also prove beneficial to non-commercial aviation operations as well.


Summary | Actions ▾
search | help | more »
examples
my command history | more »
common next commands | help | more »

Retrieves events from an index and filters them using keywords or key/value expressions.

Keep only search results that have the specified "src" or "dst" values.

```
src="10.9.165.*" OR dst="10.9.165.8"
```

Search for events with either codes 10 or 29, and a host that isn't

```
... | search *
... | search host=
... | search eventtype=
... | search count>0
... | search tag=ticket
```

Figure 2. Splunk search assistant.

3.2 Alerting

With a wide variety of sensors and software applications present on aircraft and spacecraft, substantial amounts of important data are produced, especially regarding vehicle health and conflict detection systems. Splunk's alerting functionality sends automatic notifications of any changes made to data, file systems, or any other devices that are being monitored. As displayed in Figure 3, the alert function allows a user to schedule a search, set conditions that activate an alert, and specify actions to take when the alert is triggered. This utility may prove to be valuable for processing critical data in real time. For instance, an alert can be scheduled every 5 minutes on a spacecraft's pressure system, that sends an e-mail if the pressure falls below an allowable level.

3.3 Reporting and Application Development

Reporting results is an integral part of data processing and analysis. Like many data mining tools, Splunk offers many common reporting methods such as pie charts and bar charts, as illustrated in Figure 3.

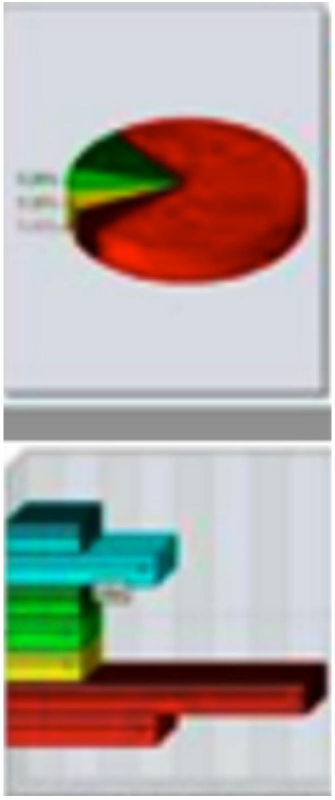
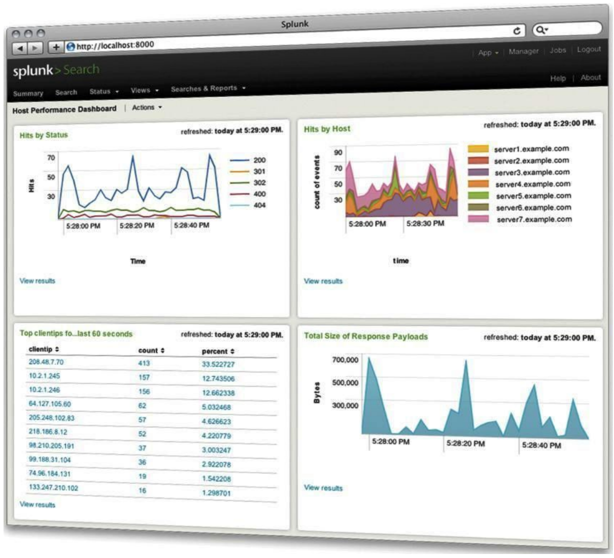


Figure 3. Sample Splunk pie and bar charts.

One of the primary differences with Splunk is the ability to search data with eval and stats functions in order to immediately return useful information. The eval command calculates an expression and places the resulting value into a useful field. For example, *split (X, "Y")* returns X as a multi-valued field, split by a delimiter Y. The stats function uses a basic statistical function to return information. The stat function *count (X)* returns the number of incidents of the field value X, that can be instantly added to chart reports. These built-in capabilities help reduce the need for individually developed algorithms.

Splunk also offers the option to schedule reports to be saved to Portable Document Format (PDF) files or integrate them into a dashboard. A daily PDF document of spacecraft vehicle health systems data can be saved and sent to management or other users for individual business use. As demonstrated in Figure 4, a dashboard offers a quick-view, in real time, of multiple data reports. The dashboard can be personalized dependent on user needs and also works in conjunction with the alerting capability. In the case of avionics flight systems, the dashboard can display control position data, vehicle fuel level data, and incident data all on one page. This function allows strategic decision making based on instantaneous data results.



Top clientips fo...last 60 seconds refreshed: today at 5:29:00 PM.

clientip ⇕	count ⇕	percent ⇕
208.48.7.70	413	33.522727
10.2.1.245	157	12.743506
10.2.1.246	156	12.662338
64.127.105.60	62	5.032468
205.248.102.83	57	4.626623
218.186.8.12	52	4.220779
98.210.205.191	37	3.003247
99.188.31.104	36	2.922078
74.96.184.131	19	1.542208
133.247.210.102	16	1.298701

View results

Figure 4. Sample Splunk dashboard.

4 Summary

NASA's IT data mining solutions traverse the aerospace sector. Splunk's searching, alerting, and reporting present practical alternatives to avionics flight system data mining requirements. Flight data, vehicle health data, sensor and software application data can be quickly and easily processed to produce useful, valuable information in real time.

With the ever-increasing quantity of data produced by flight systems, Splunk offers continuous monitoring across numerous data sources, applications, or devices to ensure a full perspective on system operations. This dynamic capability exceeds the more static resources currently available in aerospace data mining practices.

Already in line with Federal Information Security Management Act requirements, Splunk demonstrates a viable, robust choice for government data mining needs, as well as those of non-public entities.

5 Future Uses

The use of IT data mining tools, such as Splunk, may possibly move beyond both the IT security and aerospace domains, and also address the data mining needs of areas like Earth science, space mission, satellites, and ground stations as well.

Acknowledgment

The authors would like to thank John Ray, Helen Euler, and Kiku Murakami for their support of, and contribution to, this work.

References

[1] K. Bhaduri, K. Das and P. Votava, "Distributed anomaly detection using satellite data from multiple modalities," NASA Conference on Intelligent Data Understanding, Mountain View, CA, pp. 109-123, 2010.

[2] K. Bhaduri and A. Srivasta, "A local scalable distributed expectation maximization algorithm for large peer-to-peer networks," IEEE International Conference on Data Mining, Miami. FL, pp. 31-40, 2009.

[3] T. Pfarr and J. Reis, "The integration of COTS/GOTS within NASA's HST command and control system," Proceedings of the 1st International Conference on COTS-Based Software Systems, Orlando, FL, pp. 209-221, February 2002.

[4] SODAA, Surface Operations Data Analysis and Adaptation tool, Software Package, Ver. 1.8, Mosaic ATM, Leesburg, VA, 2008.

[5] Splunk, Inc., *Splunk Admin Manual*, Ver. 4.1.5, San Francisco, CA, 2010.