Improving Data Analysis Through Diverse Data Source Integration

Jennifer Casper, jcasper@mitre.org
Jing Hu, jinghu@mitre.org

Motivation

- Overwhelmed users spend precious time managing logistics of accessing diverse data sources and manually integrating data for basic analysis.
  - The number and size of the data sources is continually growing, only making this problem worse.
  - Time spent accessing and manually integrating data reduces the amount of time and effort left for needed analysis.
- Improved data management techniques are needed to **reduce the analysis timeline** to enable more **rapid and flexible analysis of diverse live and historical data sources**.

This image is representative of an overwhelming environment in which one analyst is required to make sense of all the unique data sources given to him.

Image courtesy of The Denver Post.
http://www.denverpost.com/nationworld/ci_4103478

Background

- Advantages of Streaming Systems
  - Receive new data in real time
  - Process windows of data in main memory with minimal delay
  - Feed results to multiple applications
  - Focus on maximizing continuous output rate

- Integrated queries
  - Events such as abnormalities require both live and historical data
  - Data stream management systems have been employed with access to databases

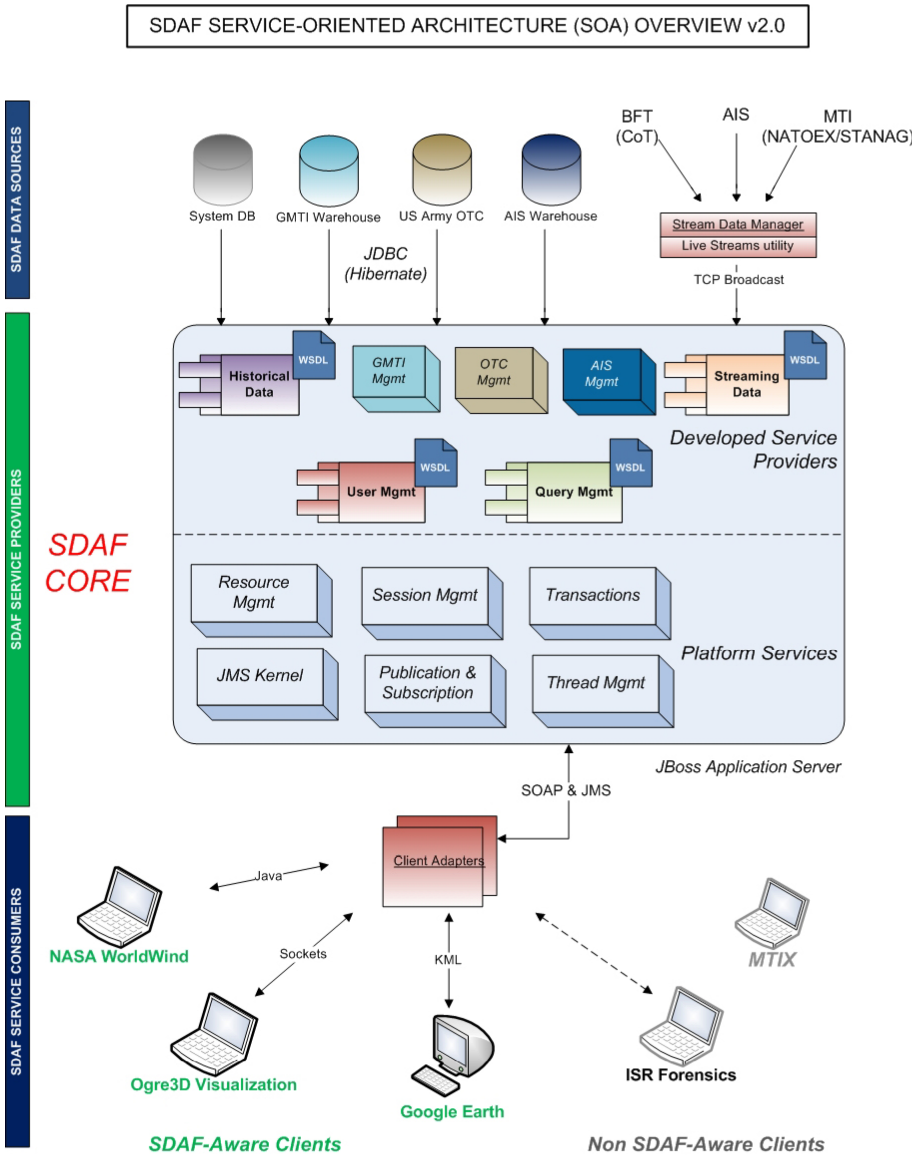
- The SOA Advantage
  - Inherent scalability and ease of service additions
  - Successful relevant applications include processing streaming weather data
  - Applied in Sensor Data & Analysis Framework (SDAF)

Refer to Publication *Improving Sensor Data Analysis Through Diverse Data Source Integration* for list of citations.

What is SDAF?

An event driven service oriented architecture (SOA) based framework.

- ❑ Features multiple SDAF *Service Providers*.
- ❑ Manages diverse, dynamic queries.
- ❑ Can add data algorithms as services.
- ❑ Scales to support different data sources & clients.



SDAF Software Architecture

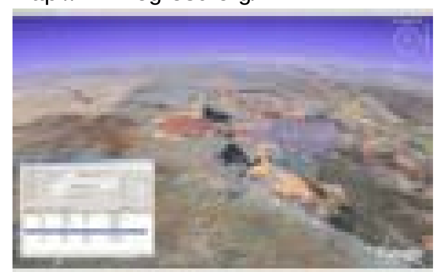
Command & Control Center

- Open Standards implementation
  - Java EE and Web Services: built on JBoss Application Server
  - Layered architecture; sources, providers, consumers
- Data Sources include Oracle, SQLServer and MySQL
- Service Providers comprise the core of SDAF
  - Historical & live data processing, internal messaging, queries, users
  - Streaming data management incorporates Esper (open source SPE)
    - Parse & filter large amounts of streaming sensor data in near real-time
    - Current operators include geo-temporal parameters, alerts, data filters
- Service Consumers interface via standards
  - Current SDAF-Aware clients are World Wind, Google Earth, Ogre3D
  - Not limited to these, can be any client.

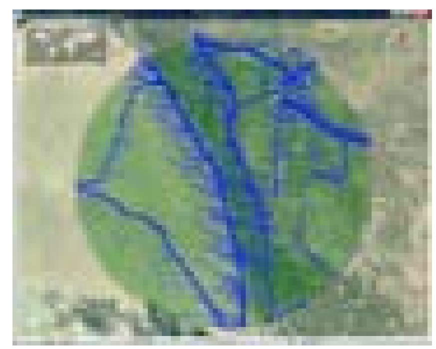
SDAF Clients

- Ogre 3D
  - C++, open source
- Google Earth
  - C#, popular mapping tool
- NASA World Wind
  - Java, high performance mapping tool
  - Used custom OpenGL render
  - Can handle in excess of 300,000 dots
- Other
  - Client adapters enable support of other clients

http://www.ogre3d.org/



http://earth.google.com/



http://www.worldwind.arc.nasa.gov/

Framework Performance

- Timely query submission
  - Reduced from about 6.0 seconds to about 0.08 seconds
  - Reduced SOAP transformations from 14 to 2 by removing the ESB
  - Improved hardware and software components
- Processing large amounts of live streaming data
  - Replaced Hypersonic with MySQL
  - 3.8 million events an hour
  - Peak bursts at 2000 events per second
- Further performance enhancements
  - Load balancing and cluster technologies

Field Applications

- Deploying SDAF with a group at Hanscom Air Force Base (HAFB) in Bedford, Massachusetts.
- Collaborating with another group at HAFB to add in specific features.
- Supporting Intelligence, Surveillance and Reconnaissance (ISR) programs.
- Supporting armed services operational test community.

Lessons Learned

- SOAs may be configured to handle intense data loads.
- ESBs are advantageous in some scenarios, but disadvantageous in others.
- Cosmetic attributes in clients may be unnecessary and costly when dealing with live data streams.
- Flexible and general-purpose client tools struggle with intense data loads.
- Historical data sources need to be constructed to support user queries.

Conclusions and Future Work

- SDAF has demonstrated numerous live and historical data sources can be simultaneously accessed, analyzed and efficiently delivered to multiple clients via a service-oriented architecture.

- Additional research to be done for risk-reduction:
  - Advanced stream mining for assistive analysis
  - Performance analysis
  - Architectural alternatives & specialized delivery methods
  - Assist in data source coverage redundancy
  - Database analytics for correlation and stream mining support
  - Advanced data pedigree for information confidence

Questions

Please contact:

Jennifer Casper ✉ jcasper@mitre.org

Jing Hu ✉ jinghu@mitre.org