SSC17-XI-05

Cloud-Based Ingest & Processing Framework (I&PF) - A Cloud-Hosted Framework for Satellite (and Other) Data Integration and Processing

Richard Baker
Solers, Inc.
7474 Greenway Center Dr., Suite 400, Greenbelt, MD 20770; (240) 790-3338
richard.baker@solers.com

Peter MacHarrie
Solers, Inc.
7474 Greenway Center Dr., Suite 400, Greenbelt, MD 20770; (240) 790-3394
peter.macharrie@solers.com

ABSTRACT

As part of an Internal Research & Development (IR&D) project, Solers created and demonstrated a Cloud-Based Ingest & Processing Framework (I&PF) hosted within the Amazon Web Services (AWS) cloud infrastructure, as a mechanism to enable fast/easy integration of satellite (and other) data sources, data processing/product generation algorithms, and data consumers within a cloud-hosted workflow (or "data pipeline") framework. This framework leverages AWS cloud services and open source software. It provides web-based user interfaces and RESTful web services for discovery and access of the ingested and processed data, as well as workflow monitoring and management.

During the IR&D project, Solers implemented 3 different NOAA-specific use cases to demonstrate the flexibility of the Cloud-Based I&PF for NOAA satellite and radar data integration and processing. Solers further extended the framework by applying it to a commercial project with OmniEarth, leveraging it to automate OmniEarth's previously manual satellite imagery ingest and land classification deep learning algorithm execution workflows within their AWS cloud infrastructure. OmniEarth leverages land classified satellite imagery to assist with water budget calculations for their commercial Water Resource Management product, and the automation provided by the Cloud-Based I&PF helped OmniEarth significantly reduce their ingest and algorithm execution timeline (from days to hours).

NOAA-FOCUSED CLOUD-BASED I&PF IR&D PROJECT

In 2015-2016, Solers conducted a NOAA-focused IR&D project to create a Cloud-Based I&PF capability hosted within AWS cloud infrastructure, to support ingest, inventory, product generation, discovery, access, and visualization of a variety of NOAA data and products. This NOAA-focused Cloud-Based I&PF IR&D project leveraged the following AWS managed cloud services and open source technologies:

- Amazon Elastic Compute Cloud (EC2) with the Apache NiFi Workflow Engine – Workflow Framework
- Amazon Simple Storage Service (S3) – Data Store
- Amazon Elasticsearch – Metadata Inventory/Catalog
- Amazon ElastiCache (Redis) – Used to provide an in-memory messaging/queuing service for passing data between different workflows
- Anaconda 2 Python Distribution – Used to execute metadata extraction and product generation scripts/routines as part of automated workflows
- Google Polymer web development framework and MapBox web-based mapping API – Used to create a cloud-hosted web-based User Portal to demonstrate discovery, access, and visualization of the ingested/processed data and metadata

Figure 1 depicts the architecture for the Solers NOAA-focused Cloud-Based I&PF IR&D project.

Figure 1: Solers' NOAA-Focused Cloud-Based I&PF IR&D Project Architecture

The Cloud-Based I&PF leverages a combination of readily available commercial managed cloud services and open source software. This creates an always-available environment hosted within the AWS cloud infrastructure for ingesting, processing, and making available (via ad-hoc search, subscription, access, and visualization) massive amounts of data sets and generated products. Such a framework can scale rapidly on-demand to meet growing data and processing needs. Through such an environment, NOAA could leverage the AWS cloud infrastructure to work with other scientists and engineers across the global community in order to rapidly and continuously develop, integrate, and improve upon product generation algorithms.

Use of AWS cloud infrastructure with on-demand auto-scaling capabilities eliminates the need for dedicated on-premise infrastructure. In addition to providing on-demand auto-scaling for compute and network resources, it also provides mechanisms to offload/archive older data sets and algorithm outputs into less expensive cloud-hosted "cold storage" services, while keeping more recent data sets and algorithm outputs that are of more demand and interest in always-available cloud-hosted storage services. The ability to perform all of these tasks without dedicated on-premise infrastructure, and only pay for what is actually used in terms of compute, network, and storage resources within the AWS cloud infrastructure, is expected to yield significant cost and time savings for developing, integrating, and testing new data sets and product generation algorithms.

Germaine to the Cloud-Based I&PF is ease of use by scientists, developers, engineers, and other users. The ability to easily create workflows (or "data pipelines") that automate the ingest and processing of data sets and algorithms in a scalable and massively parallel fashion is key to the Cloud-Based I&PF's ease of use and success. The Cloud-Based I&PF leverages readily-available open source software hosted on scalable compute resources within the AWS cloud infrastructure, and provides a workflow framework with a web-based graphical user interface that scales to meet growing data and processing demands. With the Cloud-Based I&PF's workflow framework, scientists, developers, and engineers can use their web browser to graphically create, share, and re-use workflows that represent the necessary business logic and performance characteristics for their data processing and analysis use cases. Workflows can be created, updated, started, stopped, and monitored in real-time, making the act of creating, updating, and executing data processing pipelines more like "molding clay" via real-time interaction with the system, instead of compiling code and deploying a new pre-compiled package into some kind of application server framework for each update that needs to be made. At their core, workflows are represented as XML configuration files that can be exported and imported as templates via the web-based user interface, which greatly eases the ability to configuration manage, share, and re-use them. Figure 2 depicts the CB I&PF Apache NiFi Workflow Engine's Web-Based User Interface.

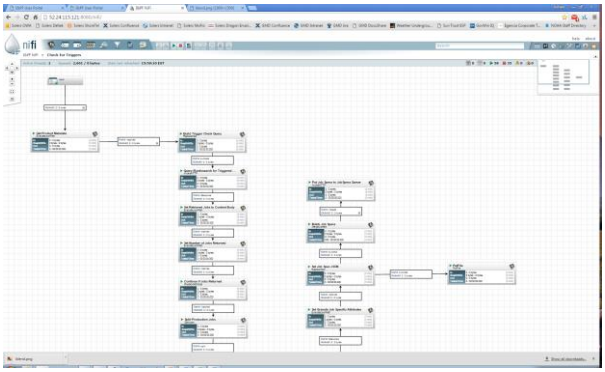

Figure 2: Cloud-Based I&PF Apache NiFi Workflow Engine's Web-Based User Interface

While the Cloud-Based I&PF's workflow framework provides the means to control and execute the primary business logic and data pipelines necessary to conduct the data ingest and processing, other scalable managed cloud services are also needed to provide metadata repositories/catalogs, data storage, messaging/notification, and other similar services. The open source software that comprises the workflow framework comes with many built-in components (called "processors") that enable out-of-the-box integration with a large variety of managed cloud services (e.g., AWS), including messaging/notification and queuing services; in-memory caching services; NoSQL and Relational database services; distributed file systems and object data stores; standard file transfer services such as FTP/FTPS/SFTP; High Performance Computing (HPC) cluster and job scheduling services; "Big Data" and Analytics platforms such as Hadoop,

Spark, Storm, and Flink; data transformation utilities such as eXtensible Stylesheet Language Transformation (XSLT); external command/script execution; and virtually any service that can be accessed via HTTP/HTTPS-based web services with XML or JSON-based inputs and outputs (i.e., RESTful or SOAP-based). The Cloud-Based I&PF's workflow framework also enables data to be compressed/decompressed during transfer and processing, and it supports common "Big Data" formats such as Avro and Parquet, allowing data to be stored in compressed formats at rest to reduce footprint, further reducing resource usage and costs within the AWS cloud infrastructure. Additionally, the Cloud-Based I&PF's workflow framework provides a well-documented Java API for creating additional custom processors that can be integrated into the workflow framework in order to implement any other service interfacing or data processing needs. All of these capabilities can be leveraged in order to efficiently execute and optimize satellite data processing and product generation algorithms.

The following use cases were completed under Solers' NOAA-focused Cloud-Based I&PF IR&D project.

NOAA S-NPP ATMS Data Ingest; MIRS Product Generation and Visualization Use Case

This use case demonstrated the ability to ingest and inventory Suomi National Polar Partnership (S-NPP) Advanced Technology Microwave Sounder (ATMS) records obtained from NOAA's Comprehensive Large Array-data Stewardship System (CLASS), and leverage those records to generate Level 2 Microwave Integrated Retrieval System (MIRS) products, through automated workflows hosted within cloud infrastructure. It enabled discovery, access, and visualization of NOAA S-NPP ATMS records and generated MIRS products through cloud-hosted web-based user interfaces.

Figure 3 depicts an overview of the data ingest and processing workflows that were implemented within the Cloud-Based I&PF for this use case.

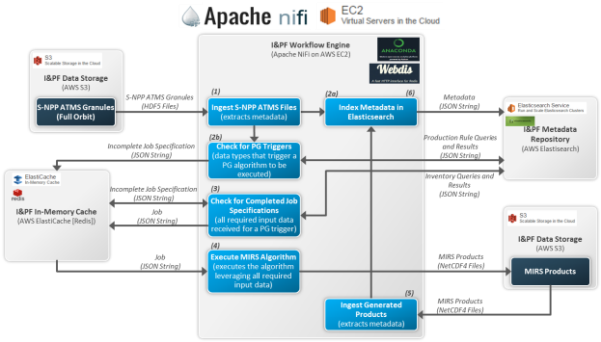

Figure 3: NOAA S-NPP ATMS Data Ingest; MIRS Product Generation Use Case Workflows

Figure 4 depicts the discovery and visualization of the ingested NOAA S-NPP ATMS records via the Cloud-Based I&PF User Portal.

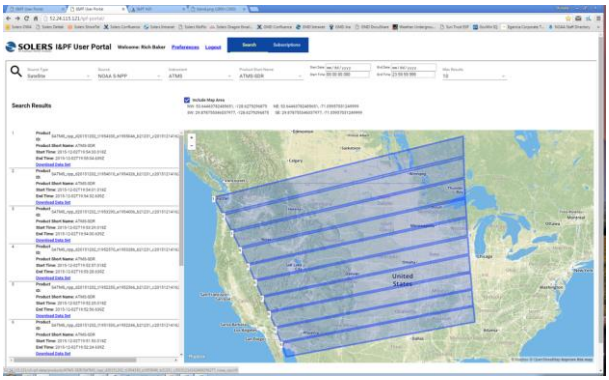

Figure 4: Discovery and Visualization of Ingested NOAA S-NPP ATMS Records

Figure 5 depicts the discovery and visualization of the generated MIRS products via the Cloud-Based I&PF User Portal.

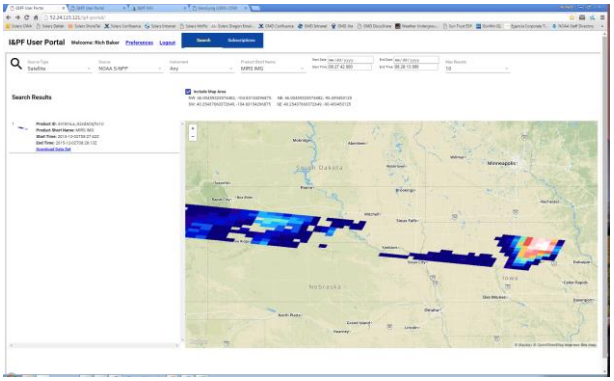

Figure 5: Discovery and Visualization of Generated MIRS Products

NOAA Nexrad II Weather Radar Data Ingest Use Case

This use case demonstrated the ability to ingest and inventory NOAA's Nexrad II weather radar data sets that have been published to Amazon S3 via the NOAA Big Data Project, through automated workflows hosted within cloud infrastructure. It enabled discovery, access, and visualization of NOAA Nexrad II weather radar data sets through cloud-hosted web-based user interfaces.

Figure 6 depicts an overview of the data ingest and processing workflows that were implemented within the Cloud-Based I&PF for this use case.

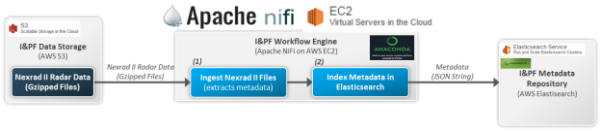

Figure 6: NOAA Nexrad II Weather Radar Data Ingest Use Case Workflows

Figure 7 depicts the discovery and visualization of the ingested NOAA Nexrad II weather radar data sets via the Cloud-Based I&PF User Portal.

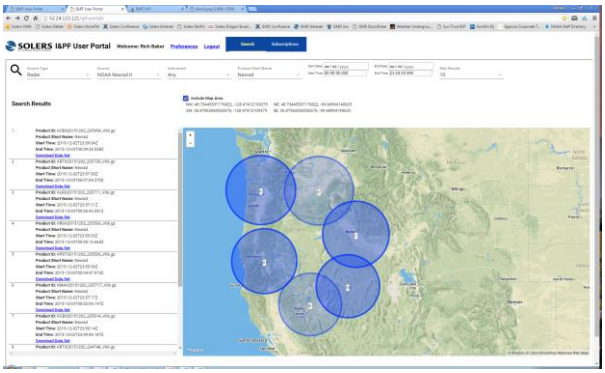

Figure 7: Discovery and Visualization of Ingested NOAA Nexrad II Weather Radar Data Sets

New MIRS / Nexrad II Blended Product Algorithm Development Use Case

This use case demonstrated the ability for a scientist/engineer to leverage the ingested NOAA Nexrad II weather radar data and the generated MIRS products within the Cloud-Based I&PF, in order to develop a new MIRS / Nexrad II blended product algorithm. This new algorithm combines the MIRS Snow/Water data with Nexrad II weather radar data, in order to provide enhanced snow/water coverage over mountainous regions.

Figure 8 depicts an overview of this use case, leveraging the ingested and generated data/products within the Cloud-Based I&PF.

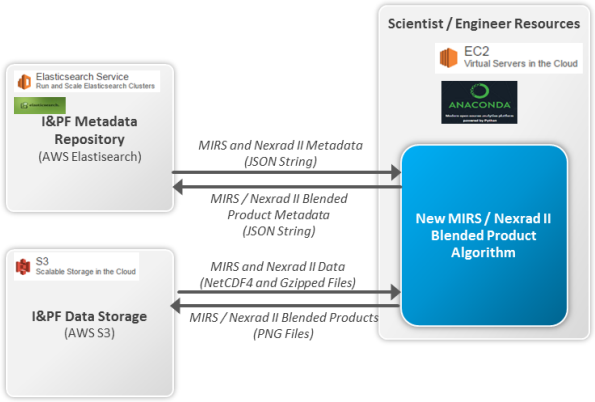

Figure 8: New MIRS / Nexrad II Blended Product Algorithm Development

Figure 9 depicts an instance of a MIRS / Nexrad II blended product that was generated via the Cloud-Based I&PF.

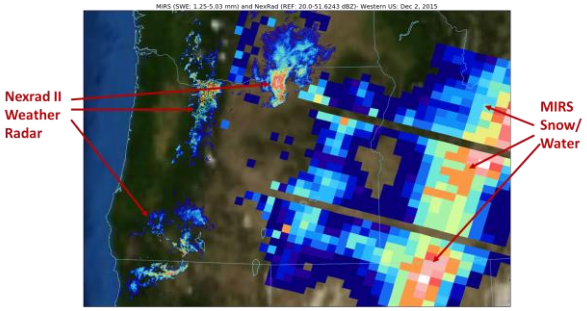

Figure 9: New MIRS / Nexrad II Blended Product Instance

OMNIEARTH ANALYTICS PLATFORM

OmniEarth is a data science analytics company, dedicated to streamlining the detection, quantification and change in a variety of parcel and property features using satellite and aerial imaging analytics. OmniEarth pioneered the application of artificial intelligence and machine learning to imagery and property data to automate the verification and identification of infrastructure and parcel features, currently a time-intensive, costly and manual process. Using proprietary algorithms, OmniEarth fuses imagery with other geospatial data to identify relevant parcel attributes on the ground.

In response to the drought in the western United States, OmniEarth pioneered the development of efficiency-based water budgets by combining satellite imagery, their land cover classification algorithms, and daily

weather information. The analysis results provide parcel-level details on how much water a resident or business should be using, and compares the budget to actual use data to identify targets for conservation measures for various water districts in California.

OmniEarth's analytics platform performs satellite imagery collection, deep learning model/algorithm development and training, metadata extraction, and land classified imagery generation activities in support of their commercial Water Resource Management application and customers, as well as other future targeted commercial applications and customers. In 2016, after the completion of the initial NOAA-focused I&PF IR&D project, Solers consulted with OmniEarth to help them expand their analytics platform to automate the satellite data ingestion and pre-processing within the AWS cloud infrastructure, leveraging the Cloud-Based I&PF architecture.

POTENTIAL FUTURE APPLICATIONS OF A CLOUD-BASED I&PF

Commercial Small Satellite Companies

Several commercial small satellite companies are either planning to launch or have already launched their own commercial satellite constellations to provide a variety of different remotely-sensed data types. A Cloud-Based I&PF would provide for an ideal data ingest, processing, and distribution platform to enable these companies to quickly process and make their data available to end users (including collaboration with NOAA, NASA, and other Government agencies), without requiring on-premise infrastructure.

NOAA Big Data Project

NOAA is making some of its data sets available to users via public cloud storage systems (such as AWS S3) as part of the NOAA Big Data Project. An example is the Nexrad II weather radar data, which was already one of the initial use cases for the Solers Cloud-Based I&PF IR&D project. A Cloud-Based I&PF would offer an ideal complementary capability to users leveraging AWS public cloud services to make the NOAA data sets searchable and accessible, and enable users to quickly define their own data processing and analytics pipelines using NOAA data. Workflows defined within a Cloud-Based I&PF could automate the data ingest process for making the data searchable and accessible, as well as support users' data processing and analytics pipelines leveraging the NOAA Big Data Project provided data.

NASA Earth Exchange (NEX)

NEX is a platform for scientific collaboration, knowledge sharing and research for the Earth science community. Three data sets from the NEX platform have been made available on Amazon S3 for users to access and process via AWS public cloud infrastructure. A Cloud-Based I&PF could provide a similar complementary add-on for data sets being made available via NASA NEX, as it would for the NOAA Big Data Project.

Other Government (and Commercial) Satellite Ground System Development, Integration, and Test Environments

A Cloud-Based I&PF could enable the performance of calibration and validation of new product algorithms that leverage multiple satellite (and other) data sets within a scalable cloud-hosted framework, prior to integrating them into operations, without requiring on-premise infrastructure. It would also enable the Government to collaborate with other U.S. and International scientists in order to continuously improve algorithms and products prior to integrating them into operational missions/systems. NOAA, NASA, USGS, and other Government agencies could benefit from such a capability for satellite ground systems such as those for the Geostationary Operational Environmental Satellite R-Series (GOES-R), Joint Polar Satellite System (JPSS), Earth Observing System Data and Information System (EOSDIS) and Landsat.