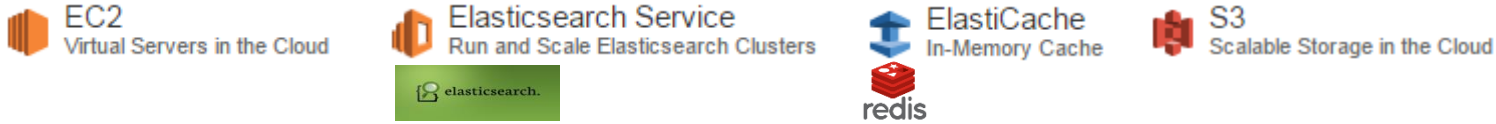
Next Generation Cloud Based Ingest & Processing Framework (I&PF) for Environmental Data

GSAW 2016

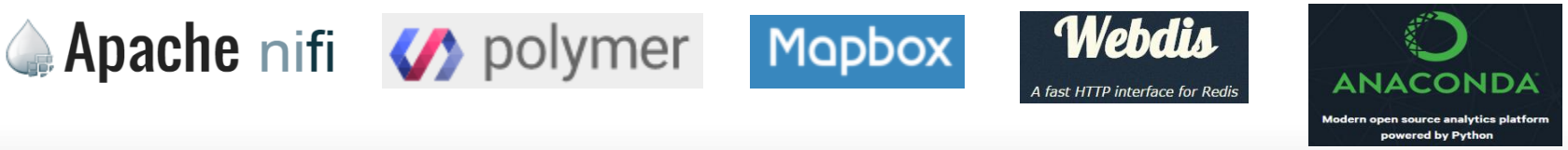
Rich Baker
Chief Architect
Solers, Inc.
Email: richard.baker@solers.com
Phone: 240-790-3338

Cloud Based Ingest & Processing Framework (I&PF)

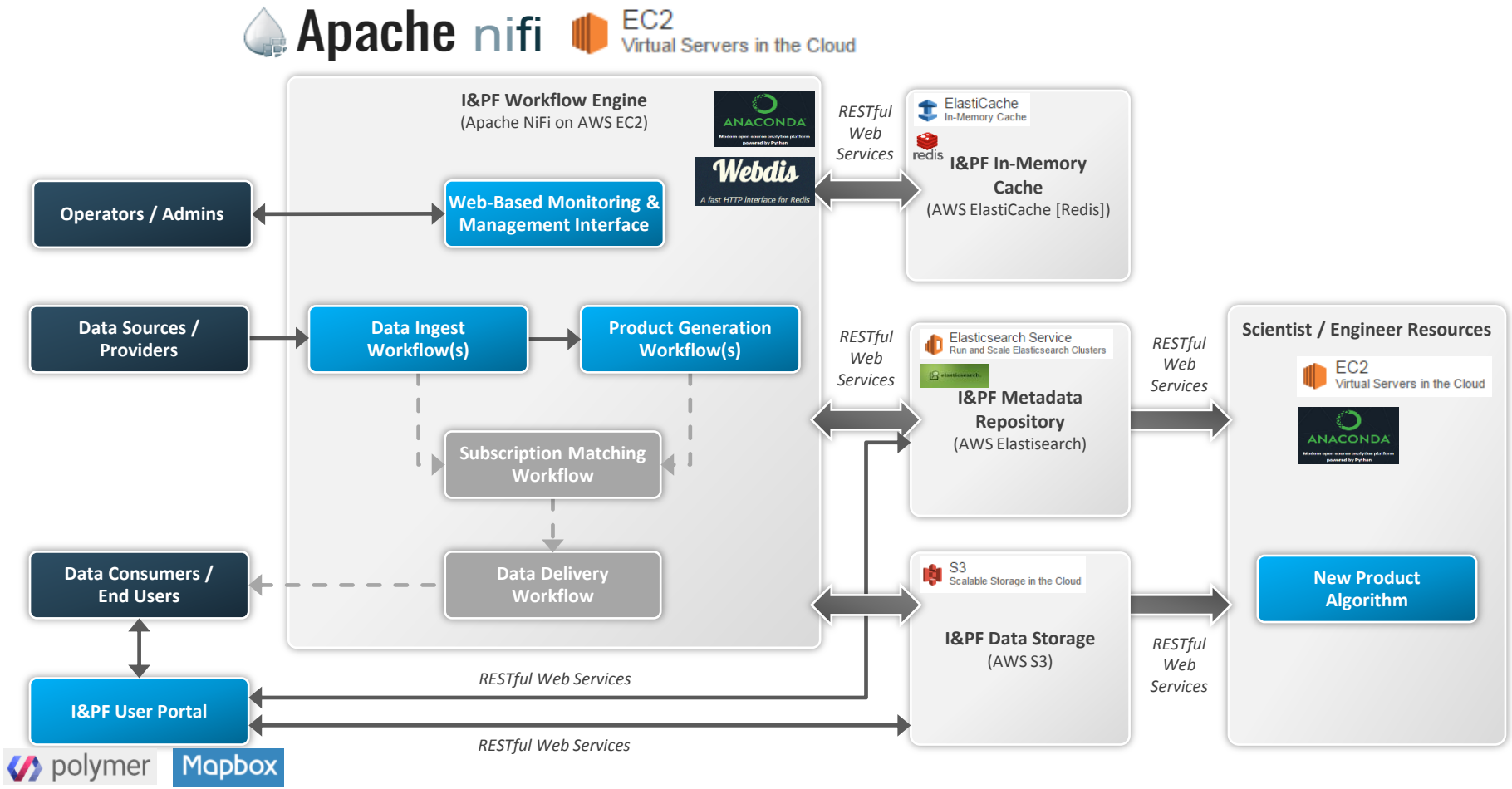
- **Currently a Solers Internal Research & Development (IR&D) project**
- **Primary Objectives:**
  - ***Enable fast/easy integration*** of data sources, product algorithms, and data consumers within a cloud based workflow (or "data pipeline") framework
  - Provide ***easy to use web-based user interfaces*** for search and access (for end users), as well as workflow monitoring and management (for system operators/admins)
  - Provide ***RESTful web services*** for other developers, scientists, etc. to ***discover and access*** the ingested/processed ***data and metadata***, for use in other research / engineering initiatives (e.g., developing a new product algorithm)
- **Leverages readily available commercial Amazon Cloud services:**



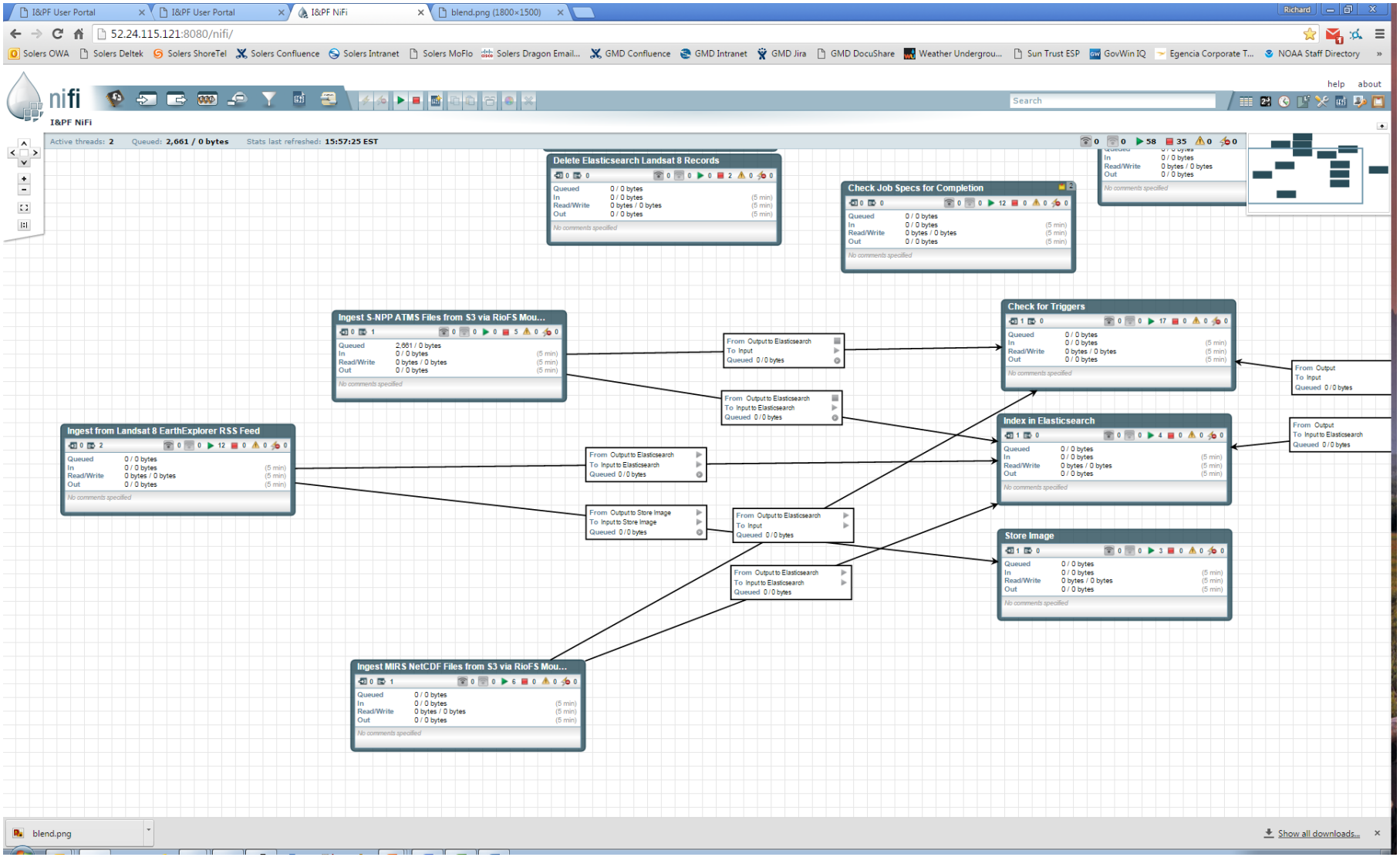
- **Leverages readily available open source technologies:**



Cloud Based I&PF Architecture



I&PF Apache NiFi Workflow Engine



I&PF Initial Demonstration Use Cases

NOAA S-NPP ATMS and MIRS

- Ingests and inventories Suomi National Polar Partnership (S-NPP) Advanced Technology Microwave Sounder (ATMS) granules
- Generates Microwave Integrated Retrieval System (MIRS) products from the ATMS granules
- Makes ATMS granules and MIRS products searchable and accessible

NOAA Nexrad II Weather Radar

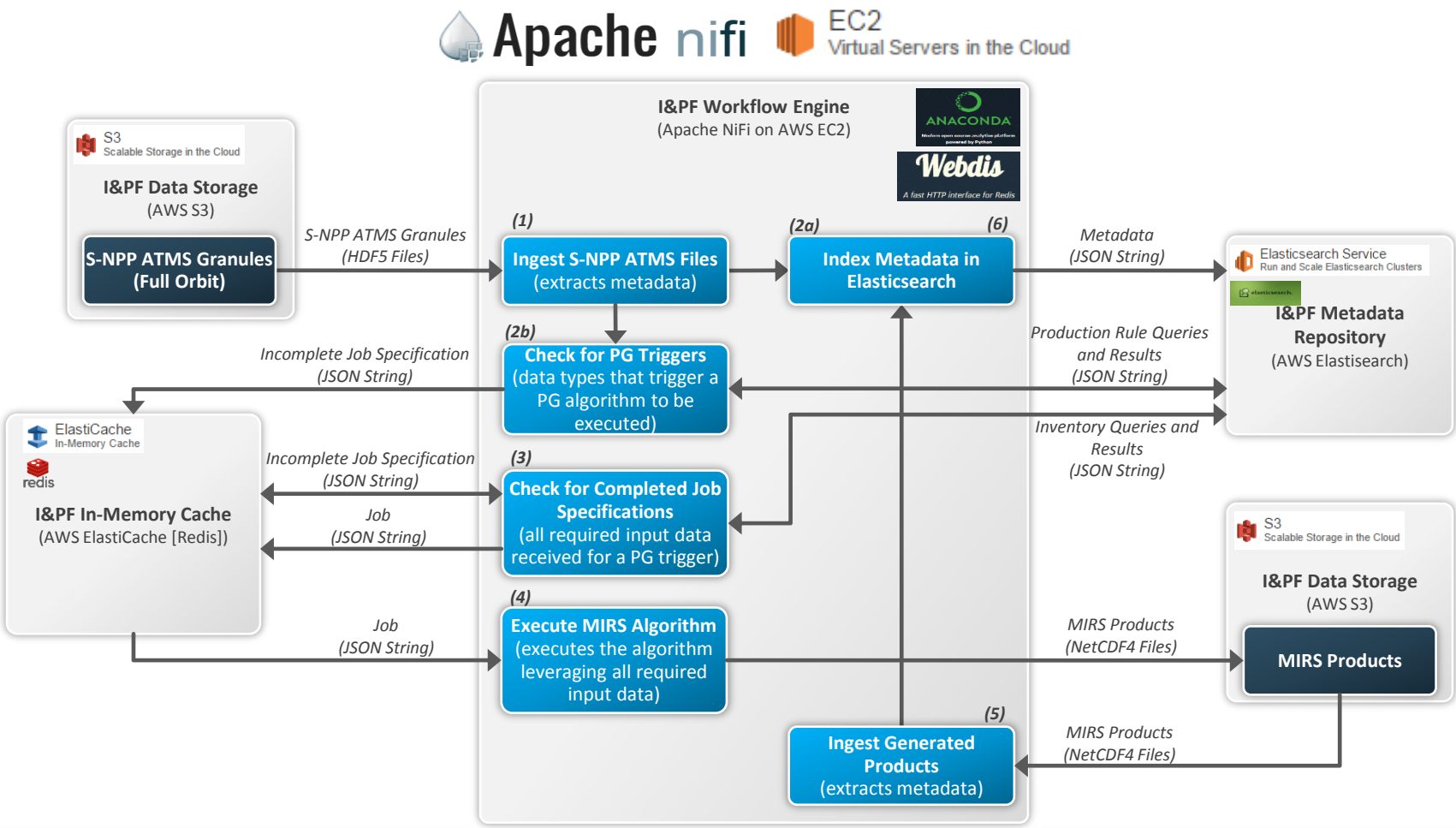
- Ingests and inventories NOAA Nexrad II Weather Radar data sets that were published on AWS S3 as part of the NOAA Big Data Project
- Makes NOAA Nexrad II Weather Radar data sets searchable and accessible

MIRS / Nexrad II Blended Product

- Leverages the available MIRS products and NOAA Nexrad II Weather Radar data sets to produce a new blended product that combines the MIRS snow/water data with the Nexrad II radar data over mountainous regions

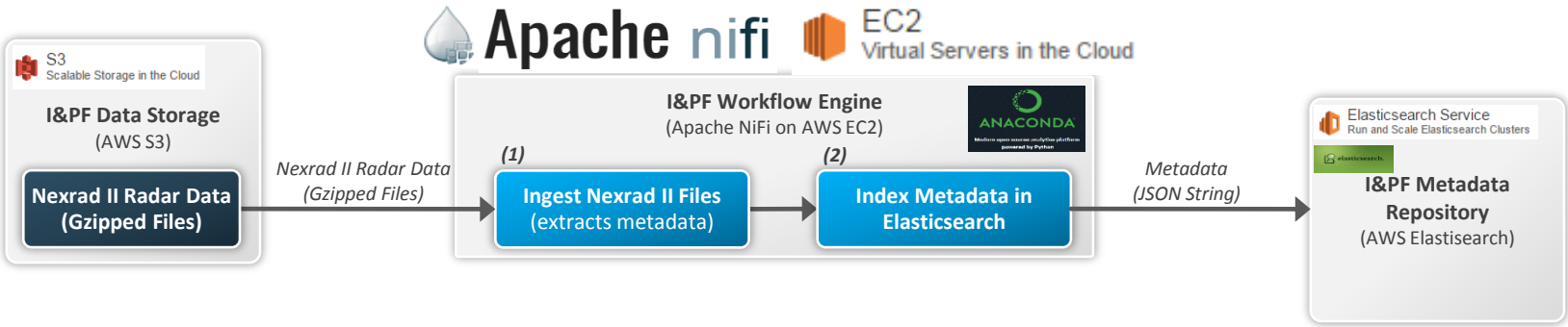
NOAA S-NPP ATMS and MIRS

Use Case: I&PF leverages automated workflows to ingest S-NPP ATMS granules, and generate MIRS products from them



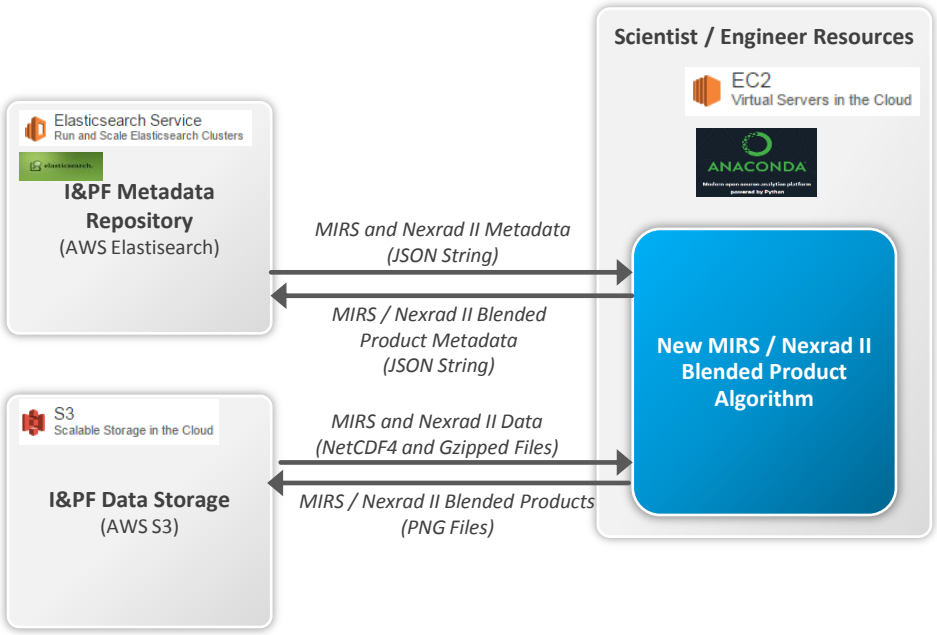
NOAA Nexrad II Weather Radar

Use Case: I&PF leverages automated workflows to ingest Nexrad II Weather Radar data

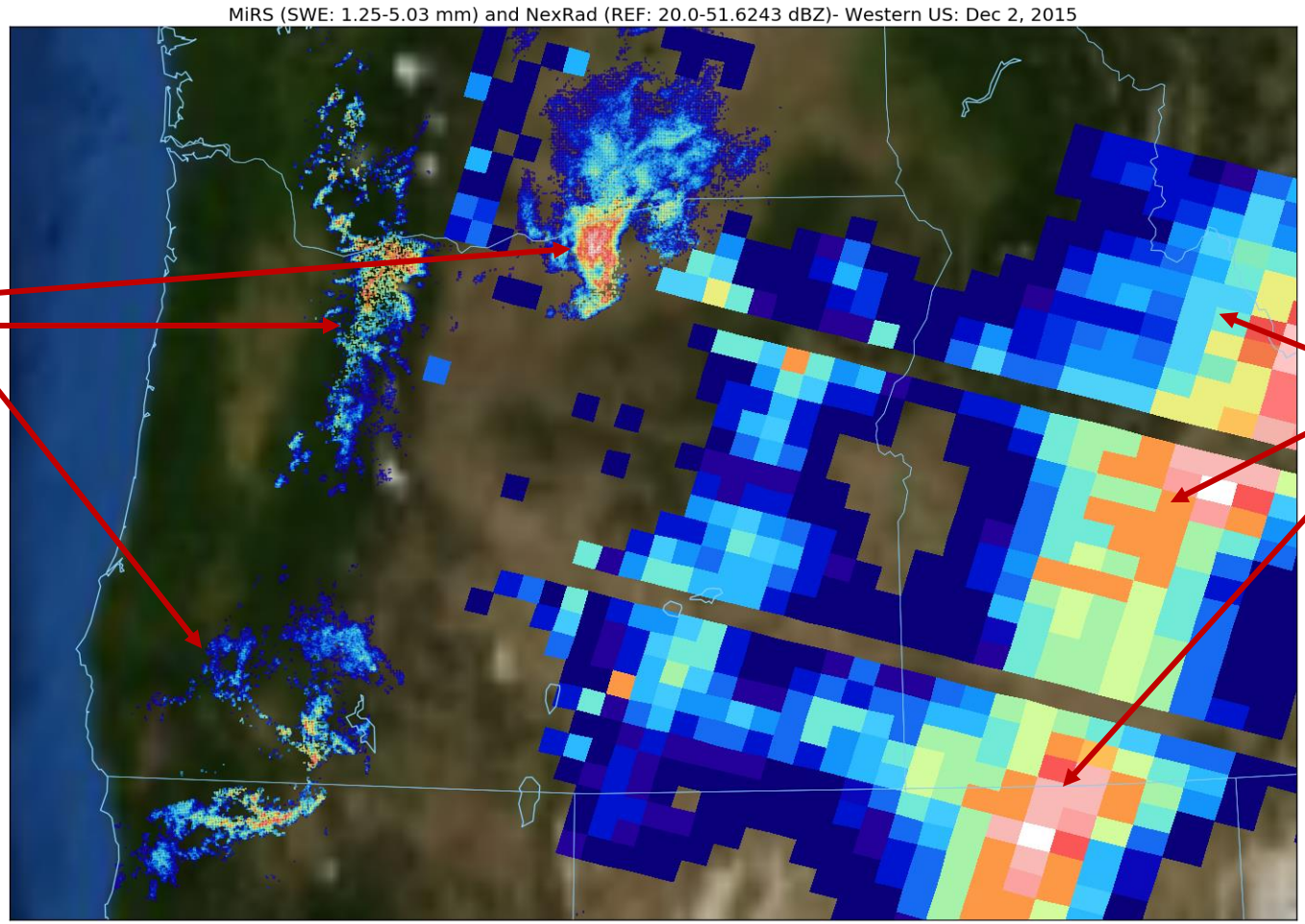


New MIRS / Nexrad II Blended Product Algorithm

Use Case: Scientist/Engineer leverages the ingested/processed MIRS products and Nexrad II Weather Radar data from the I&PF to develop a new MIRS / Nexrad II Blended Product algorithm

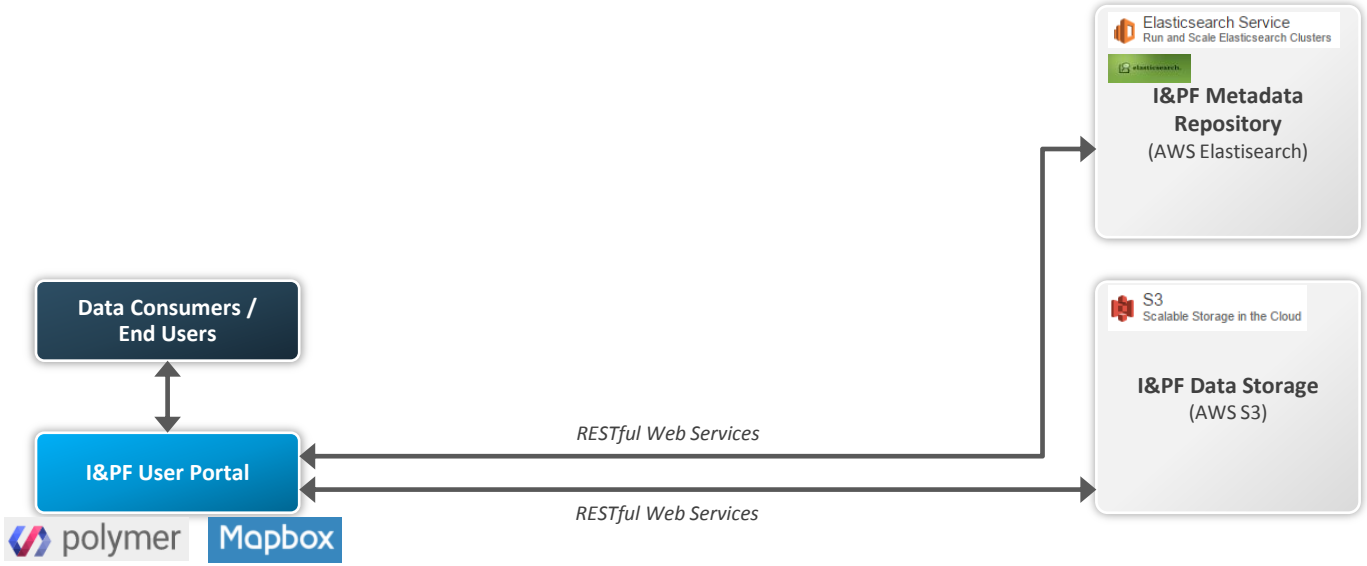


MIRS / Nexrad II Blended Product

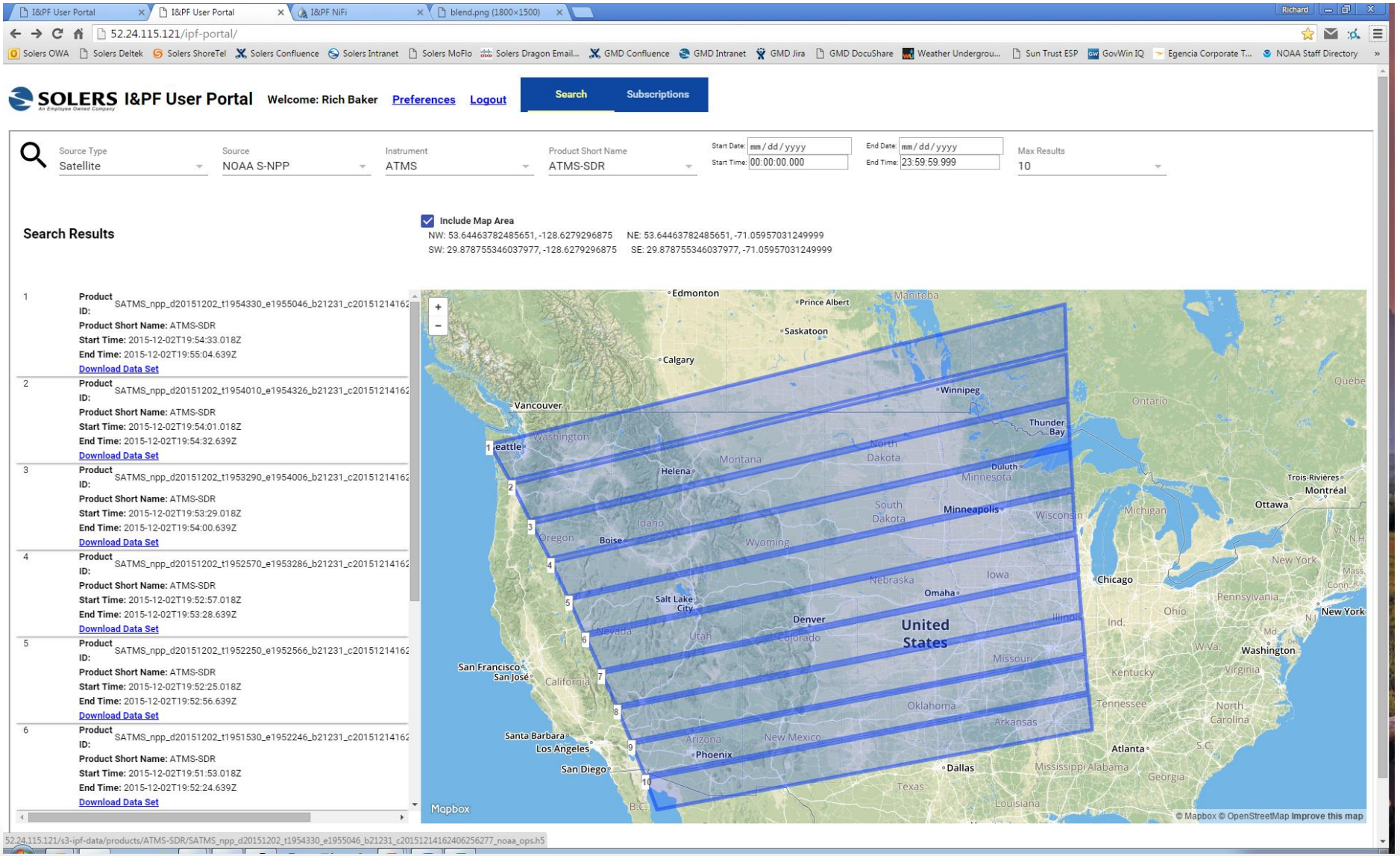


I&PF User Portal

Provides a web-based user interface for data consumers / end users to discover, access, and visualize the data and metadata that has been ingested and processed by the I&PF



I&PF User Portal: S-NPP ATMS Discovery and Access



I&PF Ongoing Use Case

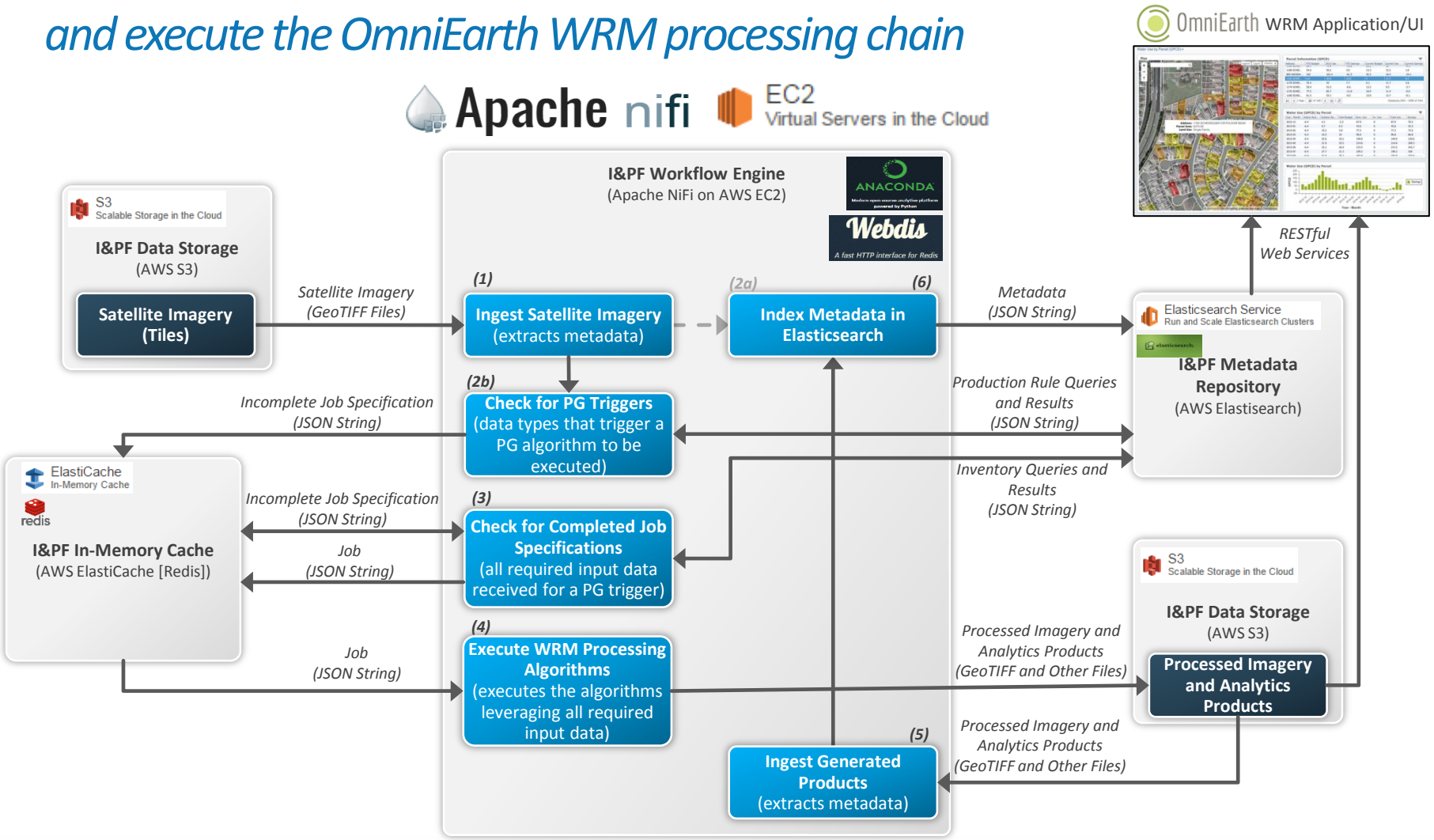
➢OmniEarth Water Resource Management (WRM)

- OmniEarth utilizes large satellite imagery sets combined with advanced machine learning algorithms to classify land cover for purposes of determining outdoor water budgets at the parcel level. These budgets aid water agencies in drought-ridden communities in the US to best target water over-users.
- This Use Case includes:
  - Ingesting satellite imagery required by OmniEarth's WRM processing chain
  - Creating workflows to automate their WRM processing chain that produces processed imagery and analytics products, which are utilized by their user-facing WRM Application / User Interface (UI)
  - Leveraging RESTful web services to automate the provisioning of the processed imagery and analytics products to their user-facing WRM Application/UI

WRM Information: http://water.omniearth.net

OmniEarth WRM (Planned)

Use Case: I&PF leverages automated workflows to ingest satellite imagery and execute the OmniEarth WRM processing chain



I&PF Future Considerations

➢Resource Management, Job Scheduling, and "Big Data" Analytics

- Provide load distribution and auto-scaling for concurrent data processing / algorithm execution tasks
- Provide a "Big Data" analytics platform that can leverage the ingested/processed data and metadata from the I&PF for large-scale analytics tasks
- Currently evaluating Hadoop YARN and Apache Spark via AWS Elastic MapReduce (EMR)

EMR Managed Hadoop Framework hadoop YARN Spark

➢Elastic's Found as a Replacement for AWS Elasticsearch

- AWS Elasticsearch is Amazon's hosted Elasticsearch service (currently only supports Elasticsearch v1.5.2, and has no involvement from Elastic)
- Found is Elastic's own hosted Elasticsearch service on AWS (latest version and features)

I&PF Potential Utility

- **Ingest and processing framework for commercial small satellite startup companies**
  - Enable them to quickly get their satellite data ingested, processed, and available to users via a scalable cloud-based workflow or “data pipeline” framework, without requiring on-premise infrastructure

- **Development, integration, and test environment for Government (and commercial) satellite ground systems**
  - Perform calibration and validation of new product algorithms that leverage multiple satellite (and other) data sets within a scalable cloud-based framework, prior to integrating them into operations, without requiring on-premise infrastructure

I&PF Technologies

Current Technologies:

- Amazon Web Services (Public Cloud Services): http://aws.amazon.com
  - AWS EC2 (Virtualized Computing): http://aws.amazon.com/ec2
  - AWS Elasticsearch (Metadata Repository): http://aws.amazon.com/elasticsearch-service
    http://www.elastic.co/products/elasticsearch
  - AWS ElastiCache (In-Memory Redis Cache): http://aws.amazon.com/elasticache
    http://redis.io
  - AWS S3 (Data Storage): http://aws.amazon.com/s3
- Apache NiFi (Workflow Engine): http://nifi.apache.org
- Continuum Anaconda (Python Framework): http://www.continuum.io
- Webdis (RESTful HTTP Interface for Redis): http://webd.is
- Google Polymer (Web Framework): http://polymer-project.org
- Mapbox (Web Mapping Toolkit): http://www.mapbox.com

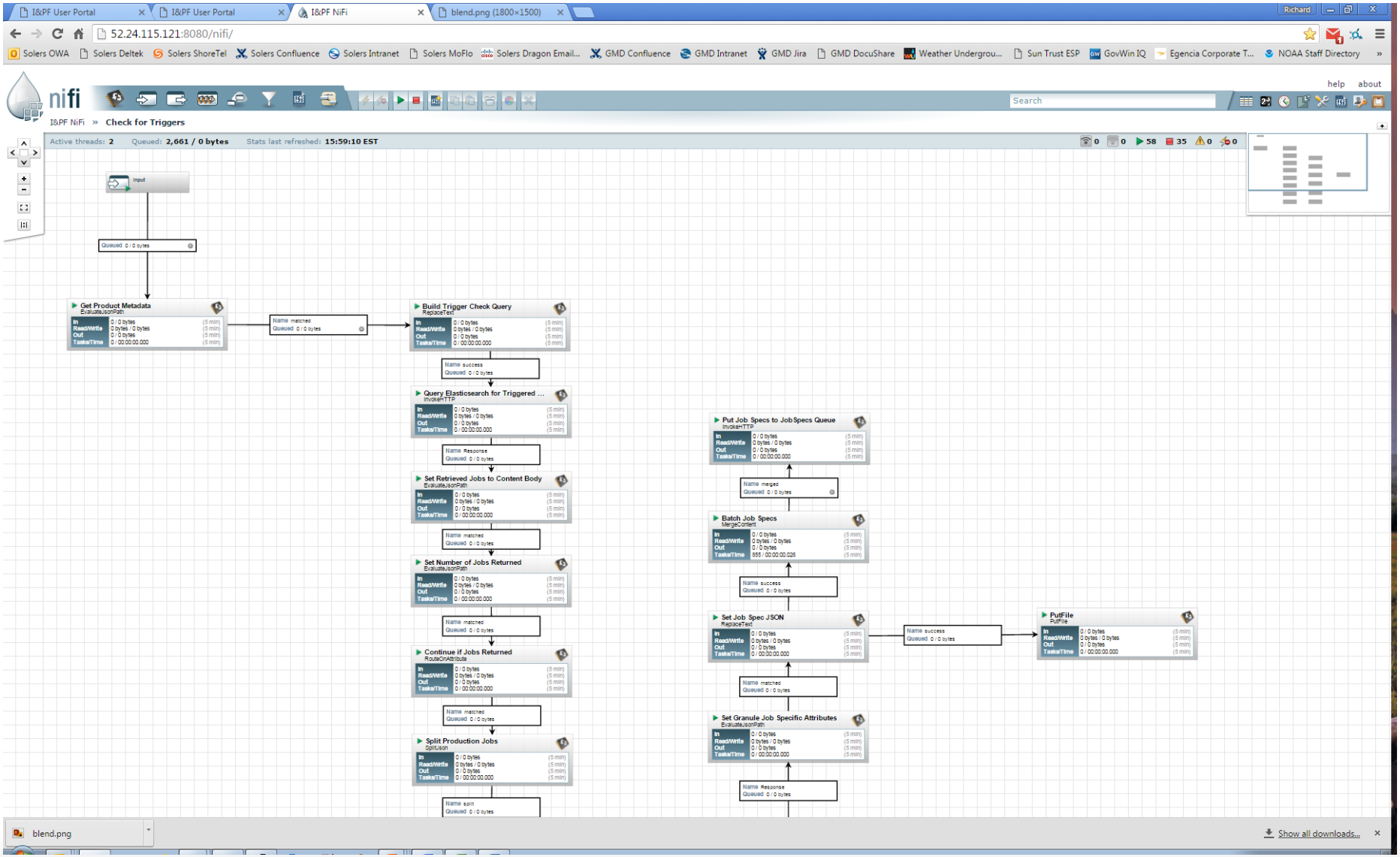
Future Technology Considerations:

- AWS EMR (Managed Hadoop/Spark): http://aws.amazon.com/elasticmapreduce
  http://hadoop.apache.org
  http://spark.apache.org
- Elastic Found (Elastic's Hosted Elasticsearch on AWS): http://www.elastic.co/found

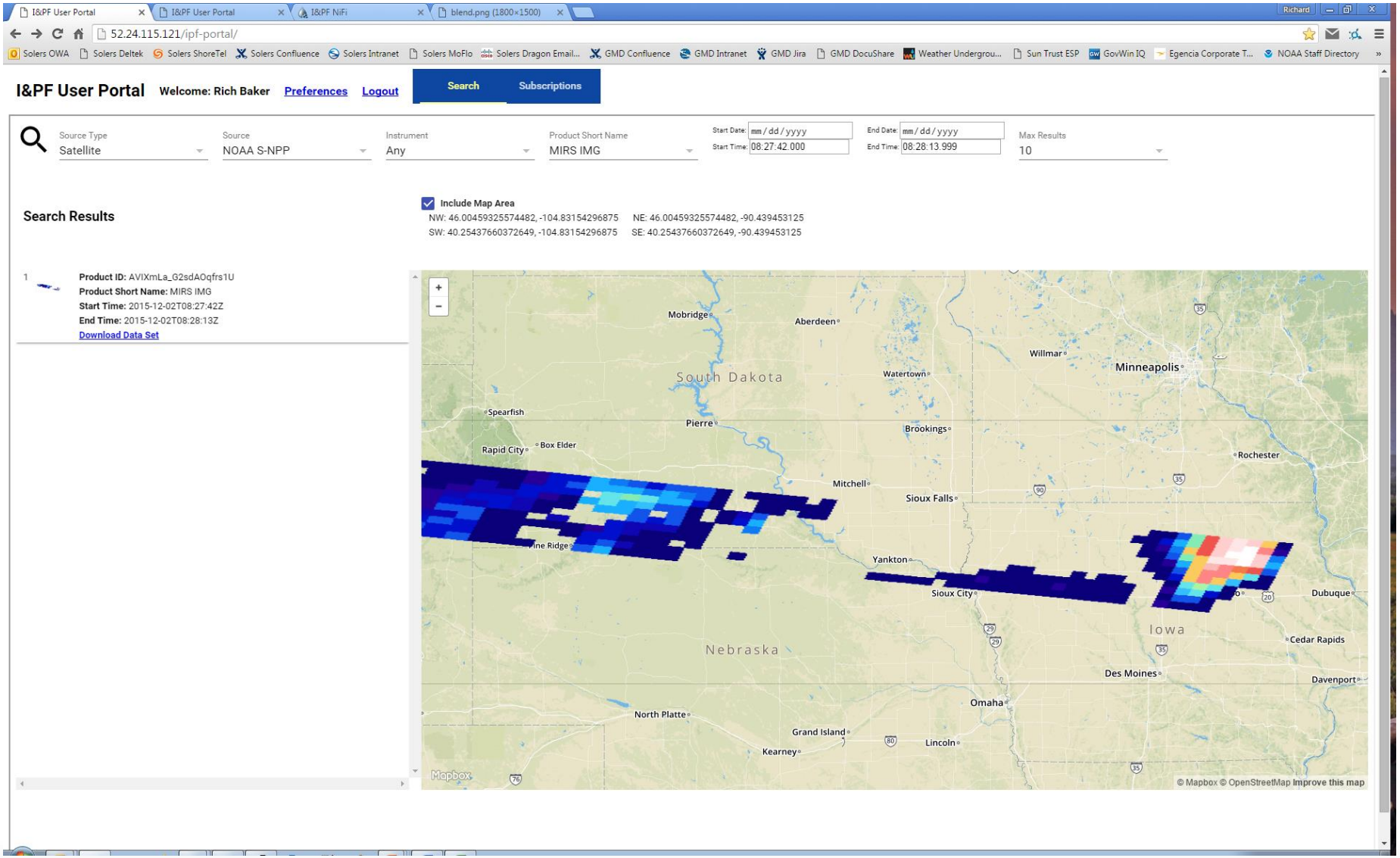
Questions

BACKUP

I&PF Apache NiFi Workflow Definition



I&PF User Portal: MIRS Discovery and Access



I&PF User Portal: Nexrad II Discovery and Access