

[](http://www.google.com/url?sa=i&rct=j&q=&esrc=s&frm=1&source=images&cd=&cad=rja&docid=owCr1QpR6RHqlM&tbnid=Nzjz_YVq0miSxM:&ved=0CAUQjRw&url=http://blogs.archives.gov/aotus/?attachment_id=1236&ei=51H6Ub_GHdO64APu34H4Aw&bvm=bv.50165853,d.dmg&psig=AFQjCNEMljwv59HAB11EckpVc6CXxxOUkA&ust=1375445858498456)

***National Archives & Records Administration***

***Online Public Access (OPA) System Administration Guide***

**Version 1.2**

***Contract Number GS-35F-0541U***

***Order Number NAMA-13-F-0120***

***November 14, 2014***

Contents

[1 Introduction 3](#_Toc403051189)

[1.1 Other Relevant Documents 3](#_Toc403051190)

[1.2 Prerequisites 4](#_Toc403051191)

[2 OPA System Overview 6](#_Toc403051192)

[2.1 Systems Organization 8](#_Toc403051193)

[2.2 Systems Inventory 9](#_Toc403051194)

[2.2.1 Hardware Inventory 9](#_Toc403051195)

[2.2.2 Software Inventory 10](#_Toc403051196)

[2.2.3 Information Inventory 14](#_Toc403051197)

[2.2.4 Amazon Web Services (AWS) 15](#_Toc403051198)

[3 User Administration 17](#_Toc403051199)

[3.1 User Classifications 17](#_Toc403051200)

[3.2 Adding and Disabling Users 19](#_Toc403051201)

[3.3 Setting User Permissions 21](#_Toc403051202)

[4 Server Administration 22](#_Toc403051203)

[4.1 Naming Conventions and Standards 22](#_Toc403051204)

[4.2 Access Control 23](#_Toc403051205)

[4.2.1 Squid 23](#_Toc403051206)

[4.2.2 Sudo 25](#_Toc403051207)

[4.3 Server Instance Startup / Shutdown Instructions 25](#_Toc403051208)

[5 Data Management Procedures 31](#_Toc403051209)

[5.1 Data Source Administration 31](#_Toc403051210)

[5.1.1 Data Sources 31](#_Toc403051211)

[5.1.2 Loading New Data 31](#_Toc403051212)

[5.1.3 Deleting Records 33](#_Toc403051213)

[5.1.4 Deleting Objects 33](#_Toc403051214)

[5.1.5 Storage Management 34](#_Toc403051215)

[5.1.6 Quarantined Objects 34](#_Toc403051216)

[5.2 OPA Database Administration 35](#_Toc403051217)

[5.3 ParcdB Administration 35](#_Toc403051218)

[5.4 CrawlDB Administration 35](#_Toc403051219)

[6 System Backup Procedures 36](#_Toc403051220)

[6.1 Instructions on How to Backup Using AWS EBS Snapshots 37](#_Toc403051221)

[6.1.1 Manual 39](#_Toc403051222)

[6.1.2 Automatic 45](#_Toc403051223)

[6.1.3 Cancelling Backups 45](#_Toc403051224)

[6.1.4 Permissions for Cancelling Backups 45](#_Toc403051225)

[6.1.5 Backup Hints 45](#_Toc403051226)

[6.2 Schedules 46](#_Toc403051227)

[6.2.1 MySQL Databases 46](#_Toc403051228)

[6.2.2 Search Engine 47](#_Toc403051229)

[6.2.3 Content Processing / Ingestion Servers 48](#_Toc403051230)

[6.2.4 Application Servers 48](#_Toc403051231)

[6.2.5 Reporting, Monitoring, and Admin Control 48](#_Toc403051232)

[6.3 Backing Up OPA System Configuration Files 48](#_Toc403051233)

[6.3.1 Content Processing / Ingestion 48](#_Toc403051234)

[6.3.2 OPA Storage 48](#_Toc403051235)

[6.3.3 Search Engine 49](#_Toc403051236)

[6.3.4 Application Server/Annotations Database 49](#_Toc403051237)

[6.3.5 User Interface 49](#_Toc403051238)

[7 System Recovery Procedures 50](#_Toc403051239)

[7.1 To copy a snapshot using the Amazon EC2 console 50](#_Toc403051240)

[7.2 To copy a snapshot using the command line 51](#_Toc403051241)

[7.3 Individual System Component Recovery 56](#_Toc403051242)

[7.3.1 Database Servers 56](#_Toc403051243)

[7.3.2 Content Processing / Ingestion Servers 56](#_Toc403051244)

[7.3.3 Search Engine Servers 58](#_Toc403051245)

[7.3.4 Application Servers 62](#_Toc403051246)

[7.3.5 Reporting, Monitoring and Admin Control 62](#_Toc403051247)

[7.4 Restoring OPA System Configuration Files 62](#_Toc403051248)

[7.4.1 Content Processing / Ingestion 62](#_Toc403051249)

[7.4.2 OPA Storage 62](#_Toc403051250)

[7.4.3 Search Engine 62](#_Toc403051251)

[7.4.4 Application Server/Annotations Database 62](#_Toc403051252)

[7.4.5 User Interface 63](#_Toc403051253)

[8 System Maintenance 64](#_Toc403051254)

[8.1 Periodic Procedures 64](#_Toc403051255)

[8.1.1 Optimizing the search index in Solr 64](#_Toc403051256)

[8.1.2 Recompiling the spell check dictionary in Solr 64](#_Toc403051257)

[8.1.3 Updating Stop Words (Stopwords.txt) in Solr 64](#_Toc403051258)

[8.1.4 Updating Featured Results (Elevate.xml) Solr 64](#_Toc403051259)

[8.1.5 Solr Best Practices 65](#_Toc403051260)

[8.1.6 ZooKeeper 65](#_Toc403051261)

[8.1.7 Frequency 65](#_Toc403051262)

[8.2 System Configuration Changes 66](#_Toc403051263)

[8.2.1 Config.xml 66](#_Toc403051264)

[8.2.2 User Interface Configuration Properties File 66](#_Toc403051265)

[8.3 Monitoring System Performance 66](#_Toc403051266)

[8.4 Monitoring and Reporting System Activity 66](#_Toc403051267)

[8.4.1 Data Sources 67](#_Toc403051268)

[8.4.2 Users 67](#_Toc403051269)

[8.4.3 Splunk Administration 69](#_Toc403051270)

[8.5 Maintaining Audit Records of System Operation 72](#_Toc403051271)

[8.6 Software Updates and Patches 72](#_Toc403051272)

[9 Security 74](#_Toc403051273)

[9.1 Issuing IDs and Passwords 74](#_Toc403051274)

[9.2 Maintaining Privileged User Accounts 74](#_Toc403051275)

[9.3 Installing and Maintaining Certificates 74](#_Toc403051276)

[9.4 License Agreements 74](#_Toc403051277)

[10 Scalability 75](#_Toc403051278)

[10.1 Hardware 75](#_Toc403051279)

[10.2 Large Ingestion of Data 75](#_Toc403051280)

[10.2.1 Downloading of DAS Exports 75](#_Toc403051281)

[10.2.2 Full Ingestion 76](#_Toc403051282)

[11 Escalation Issues 77](#_Toc403051283)

[11.1 Technical Support 77](#_Toc403051284)

[12 Troubleshooting Issues 78](#_Toc403051285)

[12.1 AR/AD Ingestion Failure (Quarantine) / Digital Object Ingestion Failure 78](#_Toc403051286)

[12.2 Rebuilding a Deleted Search Index 79](#_Toc403051287)

[12.3 Deleting Data 80](#_Toc403051288)

[12.3.1 Deleting Records 80](#_Toc403051289)

[12.3.2 Deleting Objects 80](#_Toc403051290)

[12.4 Resetting User Passwords 81](#_Toc403051291)

[13 Appendices 82](#_Toc403051292)

[13.1 Application Server Config.xml 82](#_Toc403051293)

[13.2 User Interface Configuration Properties File 85](#_Toc403051294)

[13.3 Aspire 86](#_Toc403051295)

[13.4 Solr Cloud 87](#_Toc403051296)

[13.5 Acronyms 88](#_Toc403051297)

[13.6 Reports 89](#_Toc403051298)

[13.6.1 System Access Reports 89](#_Toc403051299)

[13.6.2 Contributions Reports 90](#_Toc403051300)

[13.6.3 API Reports 93](#_Toc403051301)

[13.6.4 Export Reports 93](#_Toc403051302)

[13.6.5 Saved Lists – Reports 93](#_Toc403051303)

[13.6.6 Server Space Reports 93](#_Toc403051304)

[13.6.7 Shares – Reports 93](#_Toc403051305)

[13.6.8 Digital Analytics – Report 94](#_Toc403051306)

[13.6.9 Statistics – Report 94](#_Toc403051307)

Version Control

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Reviewer** | **Remark** |
| 0.1 | 2014-07-23 | Brandon Stahl | Initial Draft |
| 0.2 | 2014-07-31 | Brandon Stahl | Incorporated comments from ST internal team review |
| 0.3 | 2014-08-12 | Brandon Stahl | Added items to troubleshooting section |
| 0.4 | 2014-08-21 | Brandon Stahl | Updated backup and recovery procedures  Updated Section 5.1.2  Added Section 5.1.5  Updated hardware and software inventory sections |
| 0.5 | 2014-09-05 | Brandon Stahl | Updated Section 5.1.4  Added Section 13.2 |
| 1.0 | 2014-09-19 | Brandon Stahl | Incorporated comments from NARA |
| 1.1 | 2014-10-31 | Brandon Stahl | Updated Config.xml with new bulk export settings  Added Section 8.1 to add Solr best practices  Updated Section 3.2  Added Splunk DB Connect to software inventory section |
| 1.2 | 2014-11-14 | Brandon Stahl | Updated process for making OPA configuration changes  Updated UI Configuration Properties File  Removed “Confidential to Search Technologies” text from footer |

**List of Figures**

[Figure 2.1: High-Level System Architecture Diagram 6](#_Toc398905172)

[Figure 3.1: OPA Home Page – Login Button 19](#_Toc398905173)

[Figure 3.2: OPA Registered User Login Page 19](#_Toc398905174)

[Figure 3.3: OPA User Registration Page 20](#_Toc398905175)

[Figure 4.1: AWS Management Console Main Page 22](#_Toc398905176)

[Figure 4.2: AWS Instance Start / Stop 26](#_Toc398905177)

[Figure 6.1: AWS EBS Volume Backup 36](#_Toc398905178)

[Figure 7.1 AWS EBS Volume Restoration 51](#_Toc398905179)

[Figure 8.1: OPA Reporting System (Splunk) Interface 67](#_Toc398905180)

[Figure 8.2: Splunk Home Page 69](#_Toc398905181)

[Figure 8.3: Splunk User Administration 70](#_Toc398905182)

[Figure 8.4: Splunk Search Page 70](#_Toc398905183)

[Figure 8.5: Splunk Reports Tab 71](#_Toc398905184)

[Figure 8.6: Splunk Alerts Page 71](#_Toc398905185)

[Figure 8.7: Splunk Dashboards Page 72](#_Toc398905186)

[Figure 13.1: Aspire System Administration User Interface 86](#_Toc398905187)

[Figure 13.2: Solr Administration User Interface 87](#_Toc398905188)

# Introduction

The System Administrator's Guide documents relevant information regarding the deployment, configuration and administration of the production environment of the Online Public Access (OPA) system developed for the National Archives and Records Administration (NARA). It is oriented towards system administrators with a basic understanding of the system environment. This includes:

* User Administration
* Data Management Procedures
* Backup Procedures
* Recovery Procedures
* Periodic Maintenance Procedures
* System Monitoring
* System Configuration Modifications

## Other Relevant Documents

Other documents relevant to this document include:

* OPA Architecture Design v1-3
  + Identifies all major systems within OPA
  + Identifies all inputs and outputs to OPA
  + Identifies hardware and network layout overview
* OPA System Design v1-3
  + Identifies system requirements for OPA Production
  + Identifies requirements for other OPA Systems
* OPA Ingestion Design v1-3
  + Identifies overall ingestion system design
  + Identifies all other support functions required to make the ingestion work
* OPA Authorized User Interface Design v2-2
  + Identifies Account Administration API
  + Identifies Moderators API
* OPA User Interface Design v2-4
  + Identifies user interface presentation and controls
  + Identifies user interface operations
  + Identifies user interface technology and software/files layout
* OPA Application Server Design v1-3
  + Identifies how end-user and authorized user requests are handled
  + Identifies how registered user information is stored
* OPA Search Design v1-3
  + Identifies the design of the search engine
  + Identifies the index structure
  + Identifies how queries are performed
* OPA Reporting Design v1-4
  + Identifies sources used to generate reports
  + Identifies SQL/Pseudo queries that can be used to extract the information required for the reports
* OPA Public API Reference Guide v1-4
  + Identifies APIs to access records, digital objects, and public contributions
  + Identifies APIs for updating public contributions
* Amazon Web Services (AWS) Elastic Compute Cloud (EC2) User Guide
  + Identifies how AWS accounts are created and managed
  + Identifies how server instances are managed
  + Identifies how backup/recovery procedures are performed

## Prerequisites

The procedures captured in this document assume the following:

* Access to Amazon Web Services (AWS)
* OPA servers are installed and configured properly
  + Java JRE/JDK
  + Aspire and related components
  + Solr Cloud and ZooKeeper
  + Tomcat
  + MySQL Enterprise (OPA DB)
  + Splunk Enterprise
* Required external access has been properly configured
  + HTTP access to obtain DAS metadata
  + HTTP access to obtain NARA websites
  + SFTP for content storage

# OPA System Overview

The following diagram provides an overview of all OPA systems:



Figure 2.1: High-Level System Architecture Diagram

The purpose of each system is as follows:

* *Content Processing* – Is the back-end system responsible for ingestion, maintaining OPA storage, and keeping the search engine indexes up-to-date.

Content processing is also responsible for mapping NARA content into search engine indexes so that it can provide the features and deliver the results required by the end-user interfaces.

* *Search Array* – Is the search engine, structured as a series of independent search nodes, each one responsible for searching a portion of the entire index (index portions are called “partitions” and should hold around 10-50 million records depending on record size). Each search node has a redundant copy to increase query capacity and for failover.

New documents will be sent to the appropriate search node for indexing. Index updates will be automatically replicated to the failover node. Queries will be sent to all shards which will search their portion of the index in parallel, and their results will be merged before being sent back to the application server.

* *OPA Storage* – Is the long-term content storage for all publicly available NARA data. OPA storage will contain a copy of NARA data so that it can be delivered quickly and efficiently to the public.

OPA Storage will be structured based on NAID (NARA ID) into a series of “OPA-IPs” (OPA Information Packages – along the lines of the OAIS model) and will contain:

* + Digital objects
  + Alternative renditions of digital objects (thumbnails, image tiles, alternative resolutions)
  + Sub-documents, tables, and record-structured data
* *Annotations and Registration Database (OPA-DB)* – Contains tags, comments, transcriptions, and translations as well as bookkeeping information for all annotations such as lists of recently created or modified annotations, annotations per user, etc.

Also holds information on all registered users, including profile information, encrypted passwords, etc.

* *Application Server* – This is the system (made up of multiple servers) which handles end-user, API, and authorized user requests. The application server will have sub-systems for:
  + Delivering web pages (HTML) and client components (JavaScript)
  + Handling all API requests for end-user or authorized user functions
  + Sending searches to the search engine and formatting responses
  + Managing users (registering new users, profile updates, user authentication and authorization)
  + Managing the annotations (edit, update, versions, deletes, etc.)
  + Delivering and downloading content (called the “Access API” in the diagram)
  + All bulk-download and export operations
  + Requests from authorized users for account management and moderation and deletion of annotations
* *Bulk Exports Server* – Processes requests for large bulk exports of content and metadata in the background. It runs in “batch” mode and gathers up content and metadata from the system into large “tar.gz” files for download and off-line processing by end-users.
* *Client* – Client software will be written in JavaScript and HTML-5 and will run inside the user’s web browser. Client software will handle all of the:
  + Page presentation – Formatting JSON data from the OPA APIs into panes, text, controls, tables, etc.
  + Styling – All CSS for all styles will be rendered on the client
  + Events and control flow – All clicks and control flow from page to page will be handled in the browser
  + Interactive controls – All control interactivity which does not require a full-page refresh will be handled by the JavaScript in the client. For example, page previews, fly-outs, tabs, etc.

## Systems Organization

The following chart lists the OPA System Components that require administration to ensure proper operation.

|  |  |
| --- | --- |
| **OPA System Component** | **Administration Required**  **(Y/N)** |
| Content Processing / Ingestion | Y |
| Search Array | Y |
| OPA Storage | Y |
| OPA-DB | Y |
| Application Server | Y |
| User Interface | N |
| Reporting | Y |

## Systems Inventory

### Hardware Inventory

The following lists depict the hardware necessary in each OPA environment. The hardware is hosted by AWS.

* Production

|  |  |  |
| --- | --- | --- |
| **Server Configuration** | **Purpose** | **Quantity** |
| Server [122GB RAM, 16 CORES, 500GB HD, High IOPS] | Database | 2 |
| Server [30GB RAM, 8 CORES, 2TB HD, Medium IOPS] | Content Processing | 2 |
| Server [60.5GB RAM, 16 CORES, 1TB HD, High IOPS] | Search Engine | 4 |
| Server [30GB RAM, 16 CORES, 1TB HD, Low IOPS] | App Servers | 3 |
| Server [30GB RAM, 16 CORES, 1TB HD, Low IOPS] | Web Servers | 3 |
| Server [30GB RAM, 16 CORES, 2TB HD, Low IOPS] | Reporting | 1 |
| Server [1.7GB RAM, 1 CORE, 160GB HD, Low IOPS] | Jump NAT | 1 |
| Elastic Load Balancer | Distributes User Requests | 1 |

* User Acceptance Test (UAT)

|  |  |  |
| --- | --- | --- |
| **Server Configuration** | **Purpose** | **Quantity** |
| Server [122GB RAM, 16 CORES, 500GB HD, High IOPS] | Database | 2 |
| Server [30GB RAM, 8 CORES, 2TB HD, Medium IOPS] | Content Processing | 2 |
| Server [60.5GB RAM, 16 CORES, 1TB HD, High IOPS] | Search Engine | 4 |
| Server [30GB RAM, 16 CORES, 1TB HD, Low IOPS] | App Servers | 4 |
| Server [30GB RAM, 16 CORES, 1TB HD, Low IOPS] | Web Servers | 2 |
| Server [30GB RAM, 16 CORES, 2TB HD, Low IOPS] | Reporting | 1 |
| Elastic Load Balancer | Distributes User Requests | 1 |

* Sandbox

|  |  |  |
| --- | --- | --- |
| **Server Configuration** | **Purpose** | **Quantity** |
| Server [30GB RAM, 8 CORES, 2TB HD, Medium IOPS] | Content Processing | 1 |
| Server [60.5GB RAM, 16 CORES, 1TB HD, High IOPS] | Search Engine | 2 |
| Server [30GB RAM, 16 CORES, 100GB HD, Low IOPS] | App Server | 1 |
| Server [30GB RAM, 16 CORES, 100GB HD, Low IOPS] | Web Server | 1 |

* Development

|  |  |  |
| --- | --- | --- |
| **Server Configuration** | **Purpose** | **Quantity** |
| Server [122GB RAM, 16 CORES, 250GB HD, Medium IOPS] | Database | 1 |
| Server [30GB RAM, 8 CORES, 2TB HD, Medium IOPS] | Content Processing | 1 |
| Server [60.5GB RAM, 16 CORES, 1TB HD, Medium IOPS] | Search Engine | 4 |
| Server [30GB RAM, 16 CORES, 500GB HD, Low IOPS] | App Servers | 2 |
| Server [30GB RAM, 16 CORES, 500GB HD, Low IOPS] | Web Servers | 2 |
| Elastic Load Balancer | Distributes User Requests | 1 |

### Software Inventory

The following is a list of third party software required by each instance of OPA:

* AngularJS (Version 1.2.6)

An open-source web application that assists with creating one page web applications which require only HTML, CSS, and JavaScript on the client side. Its goal is to augment web applications with MVC capability. This application will run on the OPA User Interface component.

* Apache pdfbox (Version 1.8.6)

An open-source Java tool for working with PDF documents. It allows the creation of new PDF documents, manipulation of existing documents and the ability to extract content from documents. It will implemented in the Content Processing component of OPA.

* Apache Tika (Version 1.5)

An open-source toolkit that detects and extracts metadata, and text content from various documents. OPA’s Content Processing will use it for text and technical metadata extraction.

* Aspire Enterprise (Version 2.0.2)

A framework and libraries of extensible components designed to enable creation of solutions to acquire data from one or more content repositories (such as file systems, relational databases, cloud storage, or content management systems). It then extracts metadata and text from the documents, analyzes, modifies, and enhances the content and metadata (if needed), and then publishes each document, together with its metadata, to a search engine or other target application. It will be the framework implemented to handle the Content Processing role in OPA.

* Bootstrap (Version 3.0.2)

A free collection of tools for creating websites and web applications. It contains HTML and CSS-based design templates for typography, forms, buttons, navigation and other interface components, as well as optional JavaScript extensions. The User Interface component of OPA will utilize this toolset.

* Font Awesome (Version 4.1.0)

It gives you scalable vector icons that can instantly be customized — size, color, drop shadow, and anything that can be done with the power of CSS. This application is utilized by the OPA User Interface component.

* Glyphicons

A library of precisely prepared monochromatic icons and symbols, created with an emphasis on simplicity and easy orientation. This library will be utilized by OPA’s User Interface component.

* Heritrix Crawler (Version 3.1.1)

An open-source, extensible, web-scale, archival-quality web crawler project. It will be installed as a sub-component of the Aspire framework and utilized by OPA’s Content Processing component.

* ImageMagick (Version 6.5.4.7-7)

A free software to create, edit, and compose bitmap images in many formats from the command line or via programming interfaces. Images can be cropped, resized, and combined, effects can be applied, and text, polygons, and curves can be added. It will be utilized by OPA’s Content Processing component.

* iText (Version 5.5.2)

It is used to create PDF documents programmatically. Developers write Java or .NET code to produce PDFs directly from software applications, without—or with minimal—human intervention. This application is utilized by the export functionality within OPA.

* Image Manipulation [vips] (Version 7.38.5)

An open-source image processing software package which works particularly well with large images and on machines with multi-core processors. It will be used by the Content Processing component of OPA.

* jQuery (Version 1.10.2)

It is a fast, small, and feature-rich JavaScript library. It makes things like HTML document traversal and manipulation, event handling, animation, and Ajax much simpler with an easy-to-use API that works across a multitude of browsers. This library is used by the OPA User Interface component.

* MapDB (Version 1.0.5)

It provides concurrent Maps, Sets and Queues backed by disk storage or off-heap memory. It is a fast and easy to use embedded Java database engine. It will be installed as a sub-component of the Aspire framework and utilized by OPA’s Content Processing component.

* Maven (Version 3.2.1)

A software project management and comprehension tool. Based on the concept of project object model, it can manage a project’s build, reporting and documentation from a central piece of information. Maven will be used during the pre-production phase to aid in OPA development.

* MySQL (Version 5.1.73-3.el6\_5.x86\_64)

It enables you to deliver high-performance and scalable Online Transaction Processing (OLTP) applications. It provides the ease of use that has made MySQL famous along with industrial strength performance and reliability. It also includes InnoDB, making it a fully integrated transaction-safe, ACID compliant database. In addition, MySQL Replication allows you to deliver high performance and scalable applications.

* Open Seadragon (Version 1.0.0)

An open-source, web-based viewer for zoomable images, implemented in pure JavaScript. It will be used by OPA’s Content Processing for tiling.

* Pdf.js (Version 1.0.21)

It is an HTML5 technology experiment that explores building a faithful and efficient Portable Document Format (PDF) renderer without native code assistance. This application will be utilized by the OPA User Interface component.

* Quartz (Version 2.1.1)

A richly featured, open-source job scheduling library that can be integrated within virtually any Java application. It will be installed as a sub-component of the Aspire framework and utilized by OPA’s Content Processing component.

* Red Hat Enterprise Linux Server (Version 6.5)

It is a [Linux distribution](http://en.wikipedia.org/wiki/Linux_distribution) developed by [Red Hat](http://en.wikipedia.org/wiki/Red_Hat) and targeted toward the commercial market. Red Hat Enterprise Linux is released in server versions for x86, x86-64, Itanium, PowerPC and [IBM System z](http://en.wikipedia.org/wiki/IBM_System_z), and desktop versions for x86 and x86-64. It will serve as the operating system for OPA hosts.

* Splunk Enterprise (Version 6.1.2-213098)

A platform for collecting and indexing any machine data from virtually any source in real time. It can be used to create ad hoc reports and interactive dashboards to monitor key metrics. OPA will utilize Splunk for reporting.

* Splunk Forwarder (Version 6.1.1-207789)

Forwarders are lightweight Splunk instances, whose main purpose is to consume data and forward it on to Splunk Enterprise indexers for further processing. They require minimal resources and have little impact on performance, so they can usually reside on the machines where the data orginates. OPA will utilize Splunk forwarders for reporting.

* Splunk Add-on for Unix and Linux (Version 5.0.3)

The Splunk Technology Add-on for Unix and Linux works with the Splunk App for Unix and Linux to provide rapid insights and operational visibility into large-scale Unix and Linux environments. With its new pre-packaged alerting capability, flexible service-based hosts grouping, and easy management of many data sources, it arms administrators with a powerful ability to quickly identify performance and capacity bottlenecks and outliers in Unix and Linux environment. OPA will utilize the Splunk Add-on for Unix and Linux for reporting.

* Splunk DB Connect (Version 1.1.4)

Splunk DB Connect delivers reliable, scalable, real-time integration between Splunk Enterprise and relational databases. Integrate structured data from relational databases with data in Splunk Enterprise to drive deeper levels of analysis and operational intelligence.

* Spring Model-View-Controller (MVC) (Version 3.2.0)

A framework providing model-view-controller architecture and ready components that can be used to develop flexible and loosely coupled web applications. The MVC pattern results in separating the different aspects of the application (input logic, business logic, and UI logic), while providing a loose coupling between these elements. This framework will be implemented on the OPA Application Server.

* + The **Model** encapsulates the application data and in general they will consist of Plain Old Java Object.
  + The **View** is responsible for rendering the model data and in general it generates HTML output that the client's browser can interpret.
  + The **Controller** is responsible for processing user requests and building appropriate model and passes it to the view for rendering.
* Solr Cloud (Version 4.7.2)

It provides distributed index and search. Solr is an open-source enterprise search platform, featuring powerful full-text search, hit highlighting, faceted search, near real-time indexing, dynamic clustering, database integration, rich document handling, and geospatial search. It is implemented as the Search Engine for OPA.

* Tomcat (Version 7.0.54)

An open-source web server and servlet container which implements the Java Servlet and JavaServer Pages specifications. It provides a “pure Java” HTTP web server environment for Java code to run in. It will run on the OPA Application Server.

* Zookeeper (Version 3.4.5)

A server that manages all the search servers and provides for routing to leaders and shards. It can implemented as an ensemble, which is a collection of redundant Zookeepers, to provide a fail-over. It will implemented as part of the OPA Search Engine component

### Information Inventory

* Data Sources
  + Metadata from Description & Authority Service (DAS)
  + OPA Storage
  + NARA Websites
  + Annotations Database

Data from external sources (if it contains on-line content such as digital objects) will be stored in an OPA Information Package (OPA-IP).

* System Generated Reports
  + Access Reports – Related to various access statistics
  + Contributions Reports - Related to user contributions
  + API reports – Related to all the API requests from public users through systems other than OPA
  + Server Space Reports – Related to Server space information
  + Export Reports – Related to statistics on downloads/exports of OPA documents/objects/contributions etc.
  + Saved Lists Reports – Related to the “My Lists” created by users
  + Shares Reports – Related to shares on public social networking sites like Twitter, Facebook, etc.
  + Digital Analytics Reports – Reports which can be obtained using Google Analytics
  + Statistics Reports – Related to the statistics of OPA descriptions

### Amazon Web Services (AWS)

* Amazon Route 53

A highly available and scalable Domain Name System (DNS) web service.

* Elastic Load Balancing

It automatically distributes incoming application traffic across multiple Amazon EC2 instances to provide the required amount of load balancing capacity needed to distribute application traffic.

* Amazon Elastic Cloud Compute (EC2)

A web service that provides resizable compute capacity in the cloud, designed to make web-scale computing easier for developers.

* Amazon Elastic Block Store (EBS)

It provides persistent block level storage volumes for use with Amazon EC2 instances in the AWS Cloud.  Each Amazon EBS volume is automatically replicated within its Availability Zone to protect you from component failure, offering high availability and durability.

* Amazon S3

S3 provides a simple web-services interface that can be used to store and retrieve any amount of data, at any time, from anywhere on the web.

* Amazon CloudWatch (Basic)

CloudWatch is a monitoring service for AWS cloud resources and the applications you run on AWS. CloudWatch can be used to collect and track metrics, collect and monitor log files, and set alarms.

* AWS Identity and Access Management (IAM)

It enables you to securely control access to AWS services and resources for your users. Using IAM, you can create and manage AWS users and groups and use permissions to allow and deny their access to AWS resources.

* AWS Management Console

It is a single destination for managing all your AWS resources, from EC2 instances to DynamoDB tables. Use the Console to perform any number of tasks, from deploying new applications to monitoring the health of your application.

* Amazon Command Line Interface (CLI)

A unified tool to manage your AWS services. With just one tool to download and configure, you can control multiple AWS services from the command line and automate them through scripts.

* Amazon CloudFront

CloudFront is a content delivery web service. It integrates with other Amazon Web Services to give developers and businesses an easy way to distribute content to end users with low latency and high data transfer speeds.

# User Administration

In order to make access distinctions and track user activity, information systems must know who is making each request. The primary purpose of user administration is to provide information that helps a system make this determination.

The users of OPA include the following categories. Of course, a single person can be in all of these roles (the categories are not exclusive):

* **Non-Professional Users** – These are members of the public who are not professional or academic researchers. These users fall into various categories:
  + *Occasional searchers* – Log into NARA on occasion for occasional searching, for example to find family members or fellow soldiers.
  + *Contributors* – These are users who help contribute to the archive with comments, tags, transcriptions, or translations.
* **Researchers** – Researchers are looking for specific source materials for specific research goals. For example, to research a biography of a famous politician.
* **Third Party API users** – These are third party organizations that wish to interact programmatically with the OPA system, for example to bulk export images (The Digital Public Library of America, or Wikipedia) or to create new custom interfaces for searching OPA content.
* **NARA Research Support Services** – These are NARA employees who help researchers. It is expected they will be users of OPA to help the public find information.
* **NARA Contribution Moderators** – These are NARA employees who review contributions from the public. Content which is spam or vandalism will be removed (with a comment).
* **NARA Authorized Users** – These are NARA employees responsible for managing the user account database. They can deactivate and re-activate registered users and respond to support call requests (e.g. change my password, etc.).

## User Classifications

OPA users can be classified in three categories: Public (Non-Registered) Users, Registered Users and Authorized Users. Each category is further broken down into user types. Below is description of the each user type.

* Public/Non-Registered Users
  + Public User
    - No account or user not logged into system
    - Ability to print and export the top 500 search results
* Registered Users
  + Registered User
    - Established through OPA User Interface
    - Ability to print, export and save to a list the top 2,000 search results
    - Can deactivate their own account
  + Power User
    - Considered a Registered user
    - “Power” ability is granted by Account Administrator or Account Administrator/Moderator roles
    - Ability to print, export and save to a list the top 20,000 search results
    - Can deactivate their own account
* Authorized Users
  + Moderator
    - Access to the Moderator Workbench
    - Considered a Registered and Authorized User
    - Ability to view and moderate contributions made by other Registered Users, including other Account Administrators and Moderators
    - Can deactivate their own account
  + Account Administrator
    - Access to the Account Administrator Workbench
    - Can deactivate User accounts
    - Considered a Registered and Authorized User
    - Ability to create and manage all types of user accounts, including other Account Administrators and Moderators
    - Can deactivate their own account
  + Account Administrator/Moderator
    - Access to Account Administrator and Moderator Workbenches
    - Considered a Registered and Authorized User
    - Ability to view and moderate contributions made by other Registered Users, including other Account Administrators and Moderators
    - Ability to create and manage all types of user accounts, including other Account Administrators and Moderators
    - Can deactivate User accounts
    - Can deactivate their own account

## Adding and Disabling Users

* Adding Users

Users can be added to OPA in two ways:

* + Through the OPA User Interface (<https://research.archives.gov>)

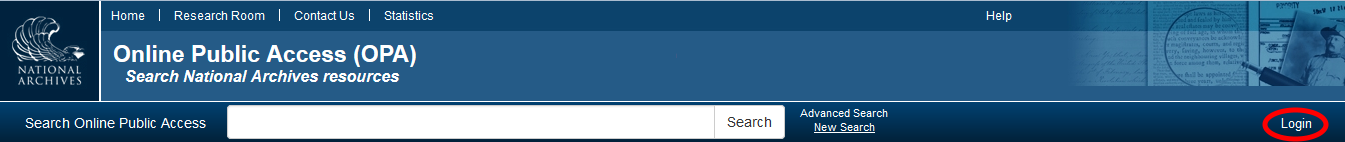


Figure 3.1: OPA Home Page – Login Button

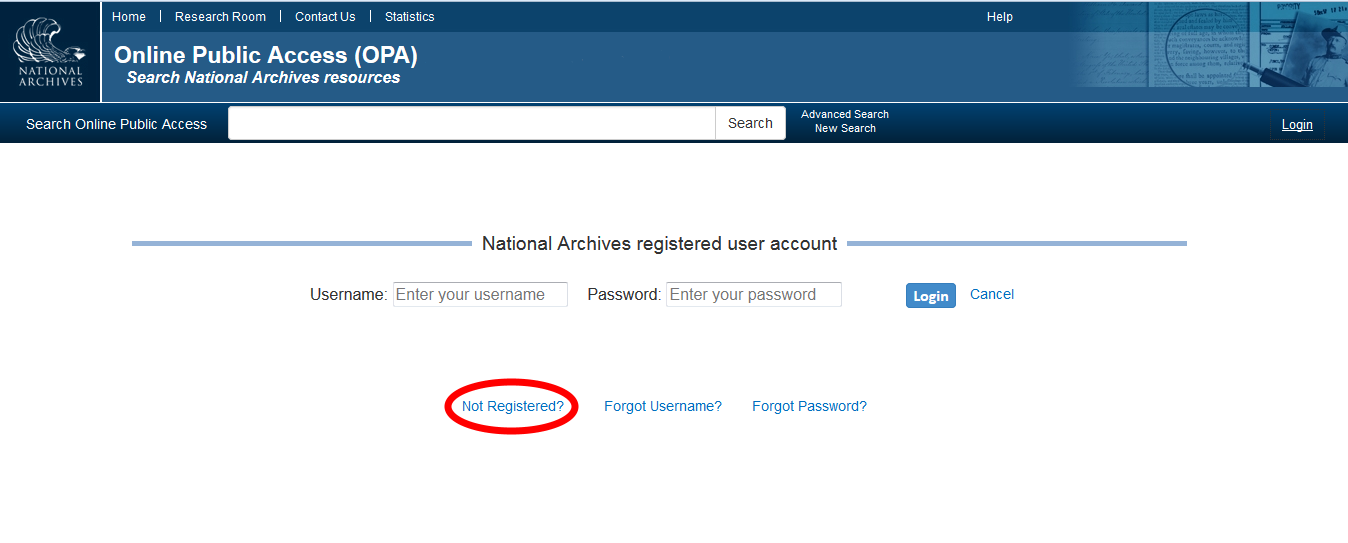


Figure 3.2: OPA Registered User Login Page

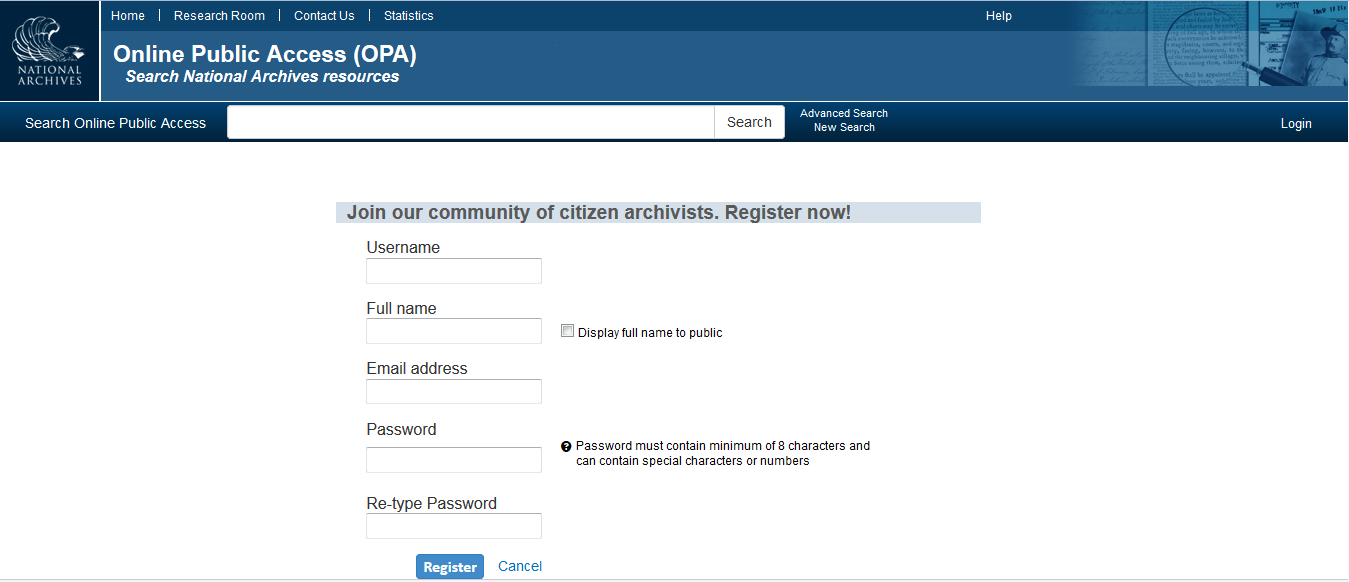


Figure 3.3: OPA User Registration Page

* + By the Account Administrator or Account Administrator/Moderator roles
* Disabling Users

A user account is automatically disabled after a year of inactivity. Prior to the user account being disabled, email notifications are sent to the user. The first email notification is sent two weeks prior to the account being disabled, and the second email notification is sent three days prior to the account being disabled. As part of the automated process, scripts have been developed for the disabling of inactive accounts .

Once the scripts are developed, execute the chmod 755 command. When you perform the **chmod 755 filename** command you allow everyone to read and execute the file, however, the owner is allowed to write to the file as well. Then add the following lines to the cron job on the jump.narasearch.us server.

#

# disable inactive admin accounts

#

30 6 \* \* \* root /opt/scripts/automation/disableinactiveuat

1 1 \* \* \* root /opt/scripts/automation/disableinactiveprod

#!/bin/bash

curl -k -c /tmp/mydevcookies -X POST 'https://dev.research.archives.gov/api/v1/login?user=accountAdminMod&password=P@ssword!'

sleep 5

curl -b /tmp/mydevcookies -X POST <https://dev.research.archives.gov/OpaAPI/iapi/v1/administrator/accounts/auto-disable>

rm -k -f /tmp/mydevcookies

#!/bin/bash

curl -k -c /tmp/myuatcookies -X POST 'https://uat.research.archives.gov/api/v1/login?user=accountAdminMod&password=P@ssword!'

sleep 5

curl -k -b /tmp/myuatcookies -X POST <https://uat.research.archives.gov/OpaAPI/iapi/v1/administrator/accounts/auto-disable>

rm -f /tmp/myuatcookies

#!/bin/bash

curl -k -c /tmp/myprodcookies -X POST 'https://prod.narasearch.us/api/v1/login?user=accountAdminMod&password=P@ssword!'

sleep 5

curl -k -b /tmp/myprodcookies -X POST <https://prod.narasearch.us/OpaAPI/iapi/v1/administrator/accounts/auto-disable>

rm -f /tmp/myprodcookies​

All Registered and Authorized Users have the ability to disable their own account. Upon disabling the user account, any saved search results lists associated with the user will be lost. If a user would like to have their account reactivated, then they will need to contact an Account Administrator or Account Administrator/Moderator to do so.

If a user account is disabled by an Account Administrator or Account Administrator/Moderator, then they will be prompted by OPA to provide a reason. In addition, there is a field where they can provide more detailed comments.

## Setting User Permissions

Permissions are determined by the user type. Once a Registered User account is created, the user type can only be edited/modified by an Account Administrator or an Account Administrator/Moderator.

# Server Administration

In order to ensure servers are optimized and functioning properly, the servers are monitored by the System Administrator through the AWS Management Console.

Below is the URL to user documentation on setting up an account on AWS, and how to get started using the services.

<http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/get-set-up-for-amazon-ec2.html>

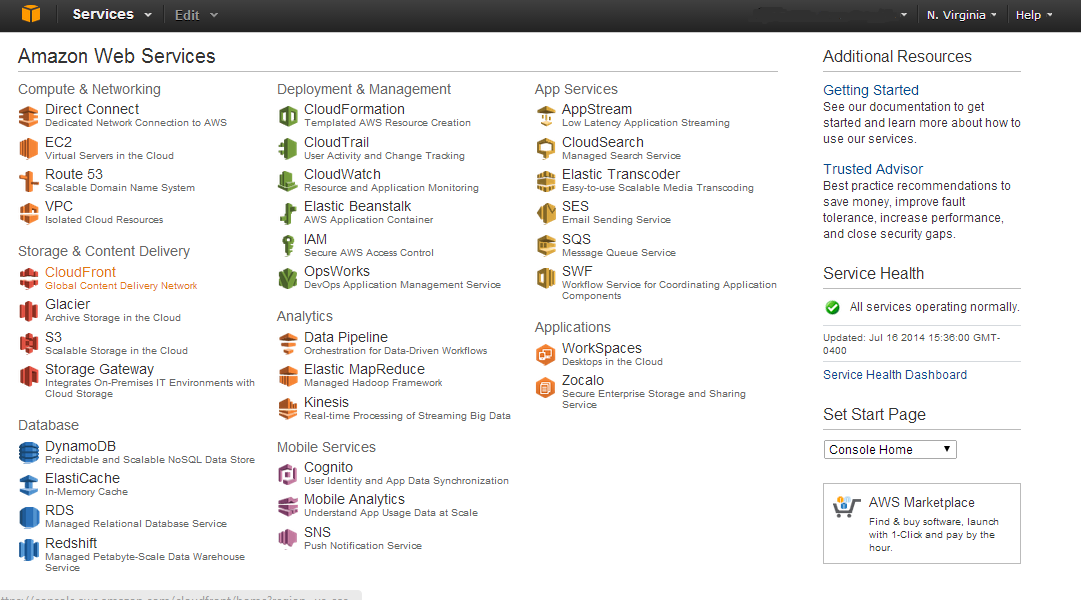


Figure 4.1: AWS Management Console Main Page

Below is the URL to user documentation on getting started with the AWS Management Console.

<http://docs.aws.amazon.com/awsconsolehelpdocs/latest/gsg/getting-started.html>

## Naming Conventions and Standards

The external DNS is currently using route53 domain narasearch.us. This is planned to be moved to a .gov namespace.

The internal DNS domain is aws.narasearch.us, and the internal DNS master server is on dw01 and slave server is on dw02.

Server hostnames are determined using the following criteria:

First character represents the environment

* d (dev)
* s (sandbox)
* u (uat)
* p (production)

Second alpha characters are for used to indicate server’s purpose

* w (web/httpd)
* a (application/java)
* db (dataBase)
* cp (content processing)
* s (search)

Followed by two-digit, sequential numbers (e.g. 01, 02, 03, 04, 05, etc.)

Examples:

dw01 is a development web/httpd server

pa02 is a production application/java server

All Amazon ec2 instances are tagged with

* Name=dnshostname
* Name=dw01 for instance

Service Groups are similar in that they contain the server’s environment and purpose in the name.

Example: DevAppserversSG

## Access Control

### Squid

OPA employs a password-protected internal proxy server utilizing Squid. Squid is a fully-featured HTTP/1.0 proxy. It offers a rich access control, authorization and logging environment to develop web proxy and content serving applications. Squid also offers a rich set of traffic optimization options.

Below are the steps to install and setup Squid (this resides on jump):

* yum -y install squid
* mv /etc/squid/squid.conf /etc/squid/squid.conf.orig
* create a new squid.conf
  + auth\_param digest program /usr/lib/squid/digest\_pw\_auth -c /etc/squid/passwd
  + auth\_param digest realm narasearch
  + auth\_param digest children 2
* acl auth\_users proxy\_auth REQUIRED
* http\_access allow auth\_users
* http\_access deny all
* http\_port 3128
* touch /etc/squid/passwd
* chmod 640 /etc/squid/passwd
* chown squid:squid /etc/squid/passwd

Setting a password for all the users you want you will need to set a password like this:

htdigest /etc/squid/passwd narasearch *joebob*

(It will prompt you for the password twice then it is set.)

service squid start

chkconfig squid on

Users will need to configure a browser. The proxy server name is jump.narasearch.us port 3128

Once users are in, they can hit internal http servers after entering username and password.

<http://ps01.aws.narasearch.us:8983/solr/>

<http://ps02.aws.narasearch.us:8983/solr/>

<http://ps03.aws.narasearch.us:8983/solr/>

<http://ps04.aws.narasearch.us:8983/solr/>

<http://pa01.aws.narasearch.us:8080/>

<http://pa02.aws.narasearch.us:8080/>

<http://pw01.aws.narasearch.us/>

<http://pw02.aws.narasearch.us/>

### Sudo

Only SSH (Secure Shell) will be allowed into OPA servers for system administration.

Sudo will handle System Administrator account and user privileges. Sudo allows a System Administrator to give certain users (or groups of users) the ability to run some (or all) commands as root while logging all commands and arguments. Sudo operates on a per-command basis, it is not a replacement for the shell. Its features include:

* The ability to restrict what commands a user may run on a per-host basis.
* Sudo does copious logging of each command, providing a clear audit trail of who did what. When used in tandem with syslogd, the system log daemon, sudocan log all commands to a central host (as well as on the local host).
* Sudo uses timestamp files to implement a "ticketing" system. When a user invokes sudo and enters their password, they are granted a ticket for 5 minutes (this timeout is configurable at compile-time). Each subsequent sudo command updates the ticket for another 5 minutes. This avoids the problem of leaving a root shell where others can physically get to your keyboard. There is also an easy way for a user to remove their ticket file, useful for placing in a .logout file.
* Sudo's configuration file, the **sudoers** file, is setup in such a way that the same sudoers file may be used on many machines. This allows for central administration while keeping the flexibility to define a user's privileges on a per-host basis.

## Server Instance Startup / Shutdown Instructions

Instances are started and stopped using the AWS Management Console.

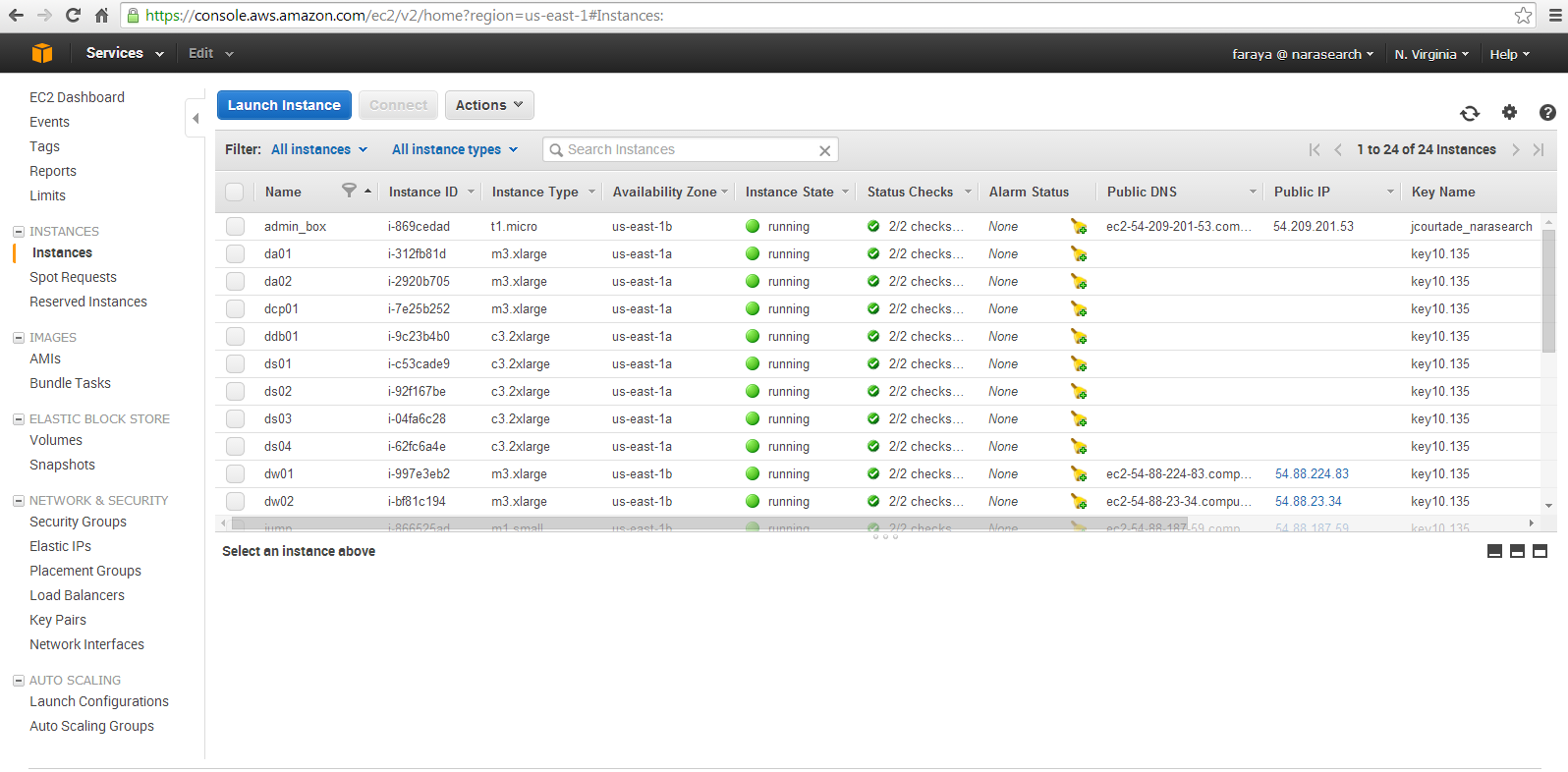


Figure 4.2: AWS Instance Start / Stop

<http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/Stop_Start.html>

Individual sudo scripts have been developed for Ingestion, Solr, Tomcat, and Zookeeper to start/stop/restart the applications. They are located in the following location: /etc/init.d/{scriptname}.

Any other services are handled by scripts contained in the operating system.

**Solr Sudo Script Example**

/etc/init.d/solr

#!/bin/bash

# description: solr Start Stop Restart

# processname: solr

# chkconfig: 234 20 80

export JAVA\_HOME=/usr/java/jdk1.7.0\_60

PATH=$JAVA\_HOME/bin:$PATH

export PATH

CATALINA\_HOME=/opt/apache-tomcat-7.0.54/

case $1 in

start)

/usr/bin/sudo -u tomcat /usr/bin/nohup /opt/solr/solr-4.7.2/example/startit &

;;

stop)

kill `ps -ef |grep java |grep start.jar |awk '{print $2}'`

;;

restart)

kill `ps -ef |grep java |grep start.jar |awk '{print $2}'`

sleep 2

/usr/bin/sudo -u tomcat /usr/bin/nohup /opt/solr/solr-4.7.2/example/startit &

;;

esac

exit 0

**Tomcat Sudo Script Example**

/etc/init.d/tomcat

#!/bin/bash

# description: Tomcat Start Stop Restart

# processname: tomcat

# chkconfig: 234 20 80

export JAVA\_HOME=/usr/java/jdk1.7.0\_60

PATH=$JAVA\_HOME/bin:$PATH

export PATH

CATALINA\_HOME=/opt/apache-tomcat-7.0.54/

case $1 in

start)

/usr/bin/sudo -u tomcat /bin/sh $CATALINA\_HOME/bin/startup.sh

;;

stop)

/usr/bin/sudo -u tomcat /bin/sh $CATALINA\_HOME/bin/shutdown.sh

;;

restart)

/usr/bin/sudo -u tomcat /bin/sh $CATALINA\_HOME/bin/shutdown.sh

/usr/bin/sudo -u tomcat /bin/sh $CATALINA\_HOME/bin/startup.sh

;;

esac

exit 0

**Ingestion Sudo Script Example**

/etc/init.d/ingestion

#!/bin/bash

# description: ingestion Start Stop Restart # processname: ingestion # chkconfig: 234 20 80

case $1 in

start)

/usr/bin/sudo -u tomcat /usr/bin/nohup /opt/scripts/init/ingestionstart & ;;

stop)

/usr/bin/sudo -u tomcat /usr/bin/nohup /opt/scripts/init/ingestionstop & ;;

restart)

stop

sleep 2

start

;;

esac

exit 0

**Zookeeper Sudo Script Example**

/etc/init.d/zookeeper

#!/bin/bash

# description: sorl Start Stop Restart

# processname: zookeeker

# chkconfig: 234 20 80

export JAVA\_HOME=/usr/java/jdk1.7.0\_60

PATH=$JAVA\_HOME/bin:$PATH

export PATH

case $1 in

start)

/usr/bin/sudo -u tomcat /opt/zookeeper/zookeeper-3.4.5/bin/startit

;;

stop)

/usr/bin/sudo -u tomcat /opt/zookeeper/zookeeper-3.4.5/bin/stopit

;;

restart)

/usr/bin/sudo -u tomcat /opt/zookeeper/zookeeper-3.4.5/bin/stopit

/usr/bin/sudo -u tomcat /opt/zookeeper/zookeeper-3.4.5/bin/startit

;;

esac

exit 0

# Data Management Procedures

OPA ingests, processes, and prepares system data for indexing by the search engine. The end user has the ability to input annotation data (e.g. transcriptions and tags) for system processing, which is recorded in a relational database. All of this requires sound data management practices throughout OPA.

## Data Source Administration

### Data Sources

* Description & Authority Service (DAS)

DAS is the source for Archival Descriptions (ADs) and Authority Records (ARs). The OPA System will read these data records from DAS for further processing and indexing. OPA receives this as a file downloaded via the Internet.

* OPA Storage

OPA Storage holds OPA Information Packages (OPA-IPs) that contain and structure digital objects (media files) associated with some ADs. OPA receives the digital objects into the pre-ingestion directory in OPA Storage via SFTP from NARA’s Content Storage.

* Annotations Database

The Annotations Database holds the data contributed by OPA users. Users can contribute comments, tags, transcriptions, and translations that are associated with records or objects. The Ingestion System will read these contributions from the Annotations Database and include their content in the Index Entries submitted to the Search Array.

* NARA Web Sites

The Archives.gov web site and the presidential library web sites are the sources of web pages that the OPA Ingestion System will process and index.

### Loading New Data

The data that comes out of DAS is only part of the data needed by OPA. The processing that happens during OPA’s ingestion of ADs and ARs creates additional data and puts data from several sources together to create Index Entries (IEs). The Web Crawler scans web sites starting with a list of seed URLs. URLs in those seed pages are then followed to other pages on the site, and the URLs in those pages are followed until essentially the entire web site has been retrieved.

#### Downloading of DAS Exports

Every Friday evening the DAS XML Export is kicked off at 8:30 pm EDT time. When the export is complete, an email will be auto-generated and sent to email addresses which will be provided by NARA.

The export file will always be placed on [http://opa.nara.ppc-cloud.com/dasexport/](http://opa.nara.ppc-cloud.com/dasexport/dasexport_7-27-14.tar.gz)

The file naming convention will be **dasexport\_mm/dd/yyyy.tar.gz** where the date is the date that the export was initiated.

The System Administrator should download the appropriate file into Xmlstore/transfer. Once the download is complete, and the System Administrator wants the Ingestion System to pick up the export, they will need to move the file to Xmlstore/incremental, so the Ingestion system will recognize and process the file (within a minute).

#### Initial (Full) Ingestion

The first ingestion, populating an empty system, will be a bulk load.

##### DAS to Xmlstore

OPA Ingestion System will write successfully processed ADs/ARs to Xmlstore/full and problematic ADs/ARs to Xmlstore/quarantine. Currently, DAS metadata exports will be done weekly and OPA will process it right away.

During the processing of ADs and ARs, the parcDB will automatically be built by the Ingestion system.

When DAS metadata is processed by OPA, it will be looking for any associated digital objects in their expected directory path inside the opastorage/pre-ingestion directory for processing. Currently, NARA is organizing the objects into the appropriate directories prior to the DAS bulk export being sent to OPA.

##### Websites

The WebCrawler is crawling a list of seed URLs, specifically Archives.gov and the presidential library websites. The WebCrawler downloads and extracts content for each link it encounters. It records its status (e.g. what web pages have been crawled and when) in its database, and caches the latest version of each website crawled in case a repopulation of the Search Array is necessary.

##### Content Storage

Content storage (e.g. digital media objects) is transmitted to the OPA pre-ingestion staging area via SFTP. Currently, objects are sent with a directory path that is replicated on pre-ingestion. For example,

*Media.archives.gov/presidential-libraries/truman/letter.jpg*

The web address is stripped off and remainder of the path is kept. The objects are then organized into this directory. For example,

*presidential-libraries/truman/letter.jpg*

These paths or “links” are added to the DAS description. If these links are added to DAS before the digital objects are available, this is considered an error. Any errors are logged and fed to Splunk (OPA’s Reporting Tool). Splunk can be set up to generate an alert, which notifies System Administrators.

#### Incremental Ingestion

Once a DAS XML Export is successfully downloaded, the System Administrator will move it to Xmlstore/incremental to be processed by the Ingestion system. The DAS export contains the full set of records. The Ingestion system checks the hash of each record against the previously stored hash, and updates accordingly.

### Deleting Records

When records need to be removed from OPA, NARA will provide a delete file in a DAS XML Export. The file contains a list of NAIDs to be removed. The act of deletion will remove:

* xml from xmlstore
* opastorage directory for that record
* Indexed Entry for search

### Deleting Objects

In order for data to be removed from OPA, NARA will provide the applicable data (e.g. ADs/ARs with their digital object tag removed) to be deleted. OPA will automatically reprocess the data with the objects removed by comparing the new list of objects in record to the list of objects previously in the record.

Deleted digital object files are automatically moved by OPA to opastorage/deleted/.

The “deleted” section can be used to recover from errors in case an object was inadvertently removed from the DAS description by NARA staff.

In addition, a command is sent from Content Processing/Ingestion to the Search Engine for the specified index entries (directly related to the deleted files) to be removed.

### Storage Management

The OPA system is required to generate a Server Space Report on the first Monday of a month (Req. 14.4.2). The findings in this report, coupled with DAS data in the pre-ingestion phase of Content Processing / Ingestion, will be used to make decisions concerning storage management in OPA.

#### Responsibility

Storage management will be handled by OPA’s System Administrator(s).

#### OPA Storage

OPA Storage is implemented with AWS’ Elastic Block Storage, and is mounted via Network File System (NFS) on the content processing (ingestion) servers. There is currently no external access to the storage. In order to gain access to it, a user would need to SSH into the jump NAT server and then access it via another server (e.g., ingestion, application, or reporting).

#### Server Access to OPA Storage

OPA Storage will be NFS mounted to all of the servers that require access to it. This includes:

* + Content Processing (Ingestion) servers (read/write)
  + Application servers (read only)
  + Reporting and server management servers (read only access)
  + Bulk export server (read/write)

Note: The search engine servers and database servers will *not* require access to OPA Storage.

### Quarantined Objects

The quarantine area contains OPA-IPs that have had something wrong detected in them and need to be separated out so they are not served. Packages in this area will need to be inspected, repaired, and moved to another area by the OPA System Administrator.

## OPA Database Administration

The OPA Annotations system is designed using a MySQL relational database. The annotation and account information is separated into four groups of tables: Annotations, Accounts, Administration and Moderator.

The following components of the OPA system will be accessing the database system.

* Content Processing/Ingestion Server – Will have read/write permissions
* Application Server – Will have read/write permissions
* Administrator Access – System administrator will have read/write/delete permissions
* Reporting Tool – Reporting System will have read permissions

The database will be administered via ssh and command line parameters.

## ParcdB Administration

The Parents, Authority Records, and Counts Database (ParcDB) is a database used by the OPA Ingestion System to store information needed during the loading of DAS XML. It is implemented using MapDB, a light-weight, local “cache DB”, and is intended only for internal processing to improve performance and reduce the impact on DAS.

Reference documentation for MapDB can be found at <http://www.mapdb.org>

Minimal administration is needed as primary duty is to backup ParcDB, so the records of the previous ingestions are not lost if the database is corrupted.

## CrawlDB Administration

The Heritrix Crawler keeps an internal database of its crawling status in the CrawlDB. The database is intended for internal processing as storage of what URLs have been crawled and the associated timestamps.

Minimal administration is needed as primary duty is to backup CrawlDB, so the records of crawled web pages are not lost if the database is corrupted.

# System Backup Procedures

*Note: This information source for these backup procedures is from the User Guide on the AWS EC2 website. Please the link below for the most current procedures.*

<http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EBSSnapshots.html>

System level backups will be developed from Amazon Elastic Block Store (EBS) snapshots. An Amazon EBS snapshot is a point-in-time backup copy of an Amazon EBS volume that is stored in Amazon Simple Storage Service (S3). Snapshots are incremental backups, which means that only the blocks on the device that have changed after your most recent snapshot are saved. When you delete a snapshot, only the data exclusive to that snapshot is removed. Active snapshots contain all of the information needed to restore your data (from the time the snapshot was taken) to a new Amazon EBS volume.

Configuration file data will be uploaded via cron jobs to S3 buckets.

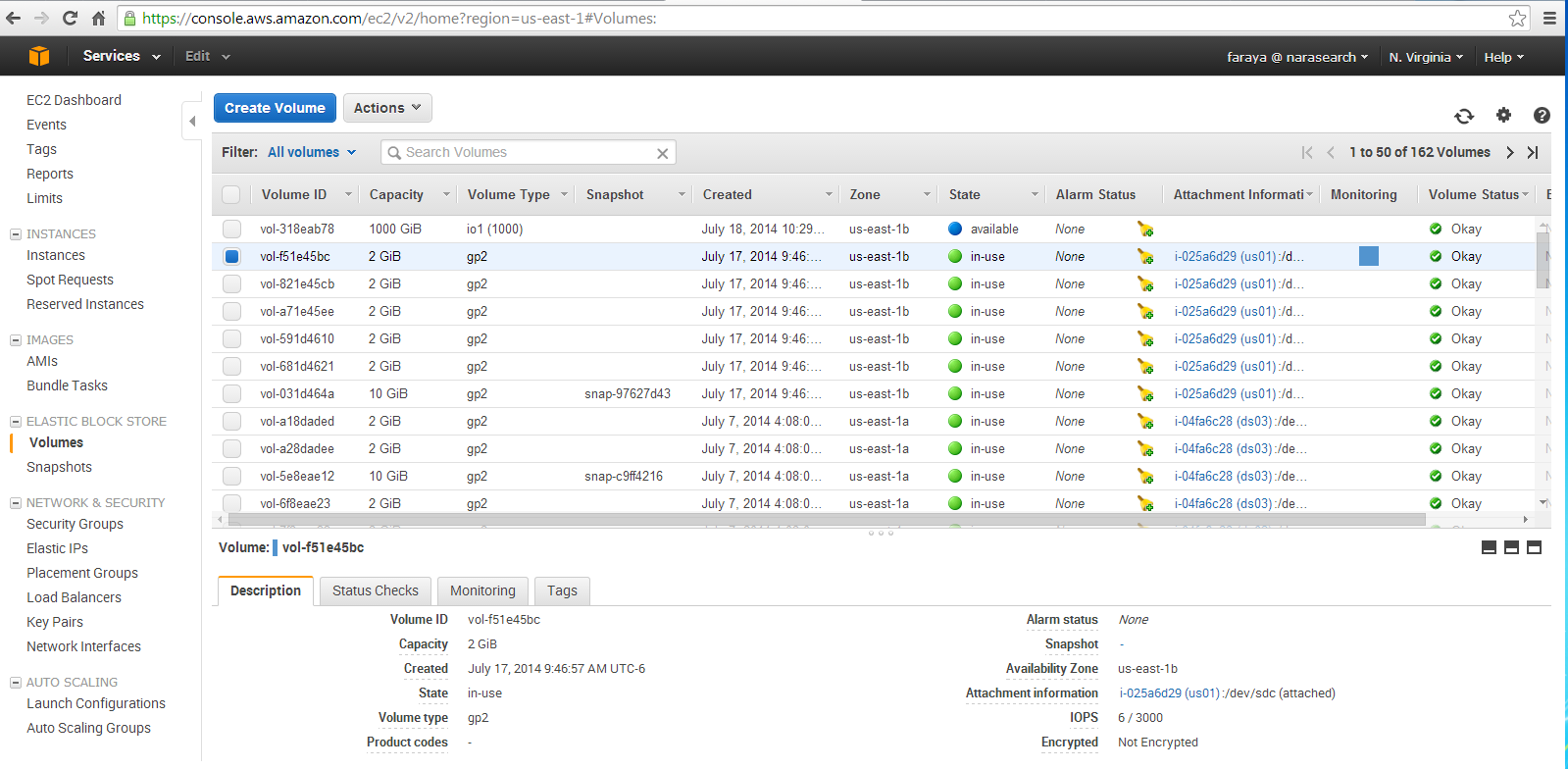


Figure 6.1: AWS EBS Volume Backup

Note: Backups are required for production environment only.

## Instructions on How to Backup Using AWS EBS Snapshots

In order to use AWS EBS Snapshots, the host must be configured with EC2 tools and the S3CMD. OPA hosts are configured with these tools by default. The tool installation procedures via the Amazon CLI are included for future reference.

Install amazon ec2 tools and s3cmd

http://s3tools.org/repositories

Install s3cmd and the ec2 tools

Amazon CLI automation (Link to documentation provided below)

http://docs.aws.amazon.com/AWSEC2/latest/CommandLineReference/set-up-ec2-cli- linux.html#setting\_up\_ec2\_command\_linux

*wget http://s3.amazonaws.com/ec2-downloads/ec2-api-tools.zip*

*mkdir /usr/local/ec2*

*unzip ec2-api-tools.zip -d /usr/local/ec2*

s3cmd

*wget http://dl.fedoraproject.org/pub/epel/6/x86\_64/epel-release-6-8.noarch.rpm*

*rpm -Uvh epel-release-6-8.noarch.rpm*

*cd /etc/yum.repos.d*

*wget http://s3tools.org/repo/RHEL\_6/s3tools.repo*

*yum -y s3cmd*

The s3cmd will work at this point.

The root user will have access to the backup user.

Create a backup user in the IAM.

The user is backups the key is here:

* Access Key ID: AKIAISMJSLEX6OKSURZA
* Secret Access Key: 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

This user is going to need pretty decent access.

It must be able to:

* Manage snapshots
* Stop and start instances
* Use s3storage

Create new group backups to add full admin access.

This will allow the backup user to create vms and destroy vms.

Modify the root user .bash\_profile to include the path for the Amazon tools,

and the keys for this user.

.bash\_profile

Get the aliases and functions

*if [ -f ~/.bashrc ]; then*

*. ~/.bashrc*

*fi*

User specific environment and startup programs

*export M2\_HOME=/usr/local/apache-maven/apache-maven-3.2.1*

*export M2=$M2\_HOME/bin*

*export MAVEN\_OPTS="-Xms256m -Xmx512m"*

*export JAVA\_HOME="/usr/java/default/"*

*export EC2\_HOME=/usr/local/ec2/ec2-api-tools-1.7.1.0*

*export AWS\_ACCESS\_KEY=AKIAISMJSLEX6OKSURZA*

*export AWS\_SECRET\_KEY=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*

*PATH=$PATH:$HOME/bin:$M2:$EC2\_HOME/bin*

*export PATH*

*chmod 600 ~/.bash\_profile*

Configure the s3cmd

*s3cmd --configure*

This will prompt you for the user and access key, and this is the resulting file

*/root/.s3cfg*

Removing this will allow you to reconfigure and you can also edit it if needed like for c.

*cat >> ~/.s3cfg<<EOD;*

*[default]*

*access\_key = AKIAISMJSLEX6OKSURZA*

*bucket\_location = US*

*cloudfront\_host = cloudfront.amazonaws.com*

*cloudfront\_resource = /2010-07-15/distribution*

*default\_mime\_type = binary/octet-stream*

*delete\_removed = False*

*dry\_run = False*

*encoding = UTF-8*

*encrypt = False*

*follow\_symlinks = False*

*force = False*

*get\_continue = False*

*gpg\_command = /usr/bin/gpg*

*gpg\_decrypt = %(gpg\_command)s -d --verbose --no-use-agent --batch --yes --passphrase-fd %(passphrase\_fd)s -o %(output\_file)s %(input\_file)s*

*gpg\_encrypt = %(gpg\_command)s -c --verbose --no-use-agent --batch --yes --passphrase-fd %(passphrase\_fd)s -o %(output\_file)s %(input\_file)s*

*gpg\_passphrase =*

*guess\_mime\_type = True*

*host\_base = s3.amazonaws.com*

*host\_bucket = %(bucket)s.s3.amazonaws.com*

*human\_readable\_sizes = False*

*list\_md5 = False*

*log\_target\_prefix =*

*preserve\_attrs = True*

*progress\_meter = True*

*proxy\_host =*

*proxy\_port = 0*

*recursive = False*

*recv\_chunk = 4096*

*reduced\_redundancy = False*

*secret\_key = lIzLQe3vKgUEv4CIr3Ym7PCA4IlvDA03rZMb4mJe*

*send\_chunk = 4096*

*simpledb\_host = sdb.amazonaws.com*

*skip\_existing = False*

*socket\_timeout = 300*

*urlencoding\_mode = normal*

*use\_https = True*

*verbosity = WARNING*

*hanging access keys for instance*

*EOD*

### Manual

#### Create a Snapshot Using the Console

* Open the Amazon EC2 console at <https://console.aws.amazon.com/ec2/>.
* Click **Snapshots** in the navigation pane.

The console displays a list of current snapshots.

* Click **Create Snapshot**.

The **Create Snapshot** dialog box appears.

* Select the volume to create a snapshot for and click **Create**.

Amazon EC2 begins creating the snapshot.

#### Create a Snapshot Using the Command Line

At this point we should be able to do things with the ec2 commands and the s3cmd.

ec2-describe-regions

REGION eu-west-1 ec2.eu-west-1.amazonaws.com

REGION sa-east-1 ec2.sa-east-1.amazonaws.com

REGION us-east-1 ec2.us-east-1.amazonaws.com

REGION ap-northeast-1 ec2.ap-northeast-1.amazonaws.com

REGION us-west-2 ec2.us-west-2.amazonaws.com

REGION us-west-1 ec2.us-west-1.amazonaws.com

REGION ap-southeast-1 ec2.ap-southeast-1.amazonaws.com

REGION ap-southeast-2 ec2.ap-southeast-2.amazonaws.com

<http://s3tools.org/s3cmd-howto> (reference)

[root@ua04 ~]# s3cmd ls

**Constraints**

System can support up to 100 buckets.

Bucket names must be globally unique.

Buckets cannot contain buckets.

There is no limit to the number of files in a bucket.

Naming of files is very important.

Convention will be hostname\_mm.dd.yyyy\_HH:MM:SS\_purpose.tar

We are only going to have one bucket for now

NARA\_08.11.2014\_12.55\_config\_backups

s3cmd mb s3://NARA\_08.11.2014\_12.55\_config\_backups

Bucket 's3://NARA\_08.11.2014\_12.55\_config\_backups/' created

Get the hostname

S3cmd ls s3://NARA\_08.11.2014\_12.55\_config\_backups/ |grep hostname

################################################

#

# jump nat scripts

#

################################################

mkdir -p /opt/scripts/backup

#

# put your scripts in here

#

/opt/scripts/backup/backitup

Script example

#!/bin/sh

export HOST=`hostname`

export DATSTR=`date +%m%d%Y\_%H%M%S`

export FILENAME=`hostname`\_"$DATSTR"\_configs.tar

BUCKET=s3://NARA\_08.11.2014\_12.55\_config\_backups

FILELIST="/etc/passwd /etc/shadow /etc/squid/squid.conf \

/etc/squid/passwd /etc/group /etc/sudoers"

#

# back the thing up

#

cd /

tar cfv /var/tmp/$FILENAME $FILELIST

s3cmd put /var/tmp/$FILENAME $BUCKET

#

# remove the local archive

#

unalias rm

rm /var/tmp/$FILENAME

#

# clean up the buckets

#

#

# remove all but the most recent 30

#

OLDFILES="`s3cmd ls $BUCKET/|grep $HOST|head -n-30|awk '{print $4}'`"

echo $OLDFILES

for m in $OLDFILES

do

echo deleting $m

s3cmd del $m

done

So that will back up the files specified in FILELIST

Run this out of cron daily. This will only store 30 days for a given host.

If there are over 30 files, it deletes all the older ones.

sudo vi /etc/crontab

30 0 \* \* \* root /opt/scripts/backup/backitup

It will run at 00:30 daily

########################################################

Then to get a particular day file is like this

s3cmd get s3://NARA\_08.11.2014\_12.55\_config\_backups/ip-10-135-1-199\_.tar

delete a file from storage is like this

s3cmd del s3://NARA\_08.11.2014\_12.55\_config\_backups/ip-10-135-1-199\_.tar

Probably want to limit the number of backups in the bucket per host

Let’s make the limit 30

Then we have 30 days of config file backups.

################################################################

#

# pw01 backup

#

#################################################################

#!/bin/sh

export HOST=`hostname`

export DATSTR=`date +%m%d%Y\_%H%M%S`

export FILENAME=`hostname`\_"$DATSTR"\_configs.tar

BUCKET=s3://NARA\_08.11.2014\_12.55\_config\_backups

FILELIST="/etc/passwd /etc/shadow /etc/crontab\

/etc/group /etc/sudoers \

/var/named /etc/named.conf /opt/rh/httpd24/ "

#

# back the thing up

#

cd /

mkdir /opt/backitup/

tar cfv /opt/backitup/$FILENAME $FILELIST --exclude='opt/rh/httpd24/root/etc/httpd/logs'

s3cmd put /opt/backitup/$FILENAME $BUCKET

#

# remove the local archive

#

#unalias rm

rm /opt/backitup/$FILENAME

#

# clean up the buckets

#

#

# remove all but the most recent 30

#

OLDFILES="`s3cmd ls $BUCKET/|grep $HOST|head -n-30|awk '{print $4}'`"

echo $OLDFILES

for m in $OLDFILES

do

echo deleting $m

s3cmd del $m

done

####################################################################

#

# pw02

#

####################################################################

mkdir -p /opt/scripts/backup

vi /opt/scripts/backup/backitup

#!/bin/sh

export HOST=`hostname`

export DATSTR=`date +%m%d%Y\_%H%M%S`

export FILENAME=`hostname`\_"$DATSTR"\_configs.tar

BUCKET=s3://NARA\_08.11.2014\_12.55\_config\_backups

FILELIST="/etc/passwd /etc/shadow /etc/crontab\

/etc/group /etc/sudoers \

/var/named /etc/named.conf /opt/rh/httpd24/ "

#

# back the thing up

#

cd /

mkdir /opt/backitup/

tar cfv /opt/backitup/$FILENAME $FILELIST --exclude='opt/rh/httpd24/root/etc/httpd/logs'

s3cmd put /opt/backitup/$FILENAME $BUCKET

#

# remove the local archive

#

#unalias rm

rm /opt/backitup/$FILENAME

#

# clean up the buckets

#

#

# remove all but the most recent 30

#

OLDFILES="`s3cmd ls $BUCKET/|grep $HOST|head -n-30|awk '{print $4}'`"

echo $OLDFILES

for m in $OLDFILES

do

echo deleting $m

s3cmd del $m

done

chmod 700 /opt/scripts/backup/backitup

test it

/opt/scripts/backup/backitup

This will do a bunch of stuff

make sure the file is up in s3

s3cmd ls s3://NARA\_08.11.2014\_12.55\_config\_backups/ |grep dw02

2014-08-13 11:45 149555200 s3://NARA\_08.11.2014\_12.55\_config\_backups/dw02\_08132014\_074626\_configs.tar

get the file and test to see if you can see the files you expect in it using list

s3cmd get s3://NARA\_08.11.2014\_12.55\_config\_backups/dw02\_08132014\_074626\_configs.tar

s3cmd ls s3://NARA\_08.11.2014\_12.55\_config\_backups/ |grep hostname

s3://NARA\_08.11.2014\_12.55\_config\_backups/

tar tfv dw02\_08132014\_074626\_configs.tar

Create a cronjob to do this nightly.

vi /etc/crontab

10 0 \* \* \* root /opt/scripts/backup/backitup

Restore files by getting the archive of the day you want and then extract the files.

### Automatic

The EBS Snapshots can be set up to be taken nightly. Please refer to AWS User Guide for the procedure.

<http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/AmazonEBS.html>

### Cancelling Backups

The scheduled EBS Snapshots can be cancelled via the AWS Management Console. Please refer to the AWS User Guide for the procedure.

### Permissions for Cancelling Backups

Only the OPA System Administrator or their designee have permission to cancel system backups in AWS. Additional users may be permitted to cancel backups only by written approval from NARA.

### Backup Hints

This section provides some system commands that can be utilized by a System Administrator to verify that the backups are being created and stored in S3.

These commands may be performed as root on any host (backups are performed as root user).

Check yesterday’s backups in total

aws s3 ls NARA\_08.11.2014\_12.55\_config\_backups |grep `date --date="yesterday" +%d`

Command output should look something like below:

2014-08-26 18:45:24   22743040 sw01\_08262014\_184522\_configs.tar

2014-08-26 09:21:05 2190632960 udb01\_08262014\_092101\_configs.tar

2014-08-26 09:22:05 2190632960 udb01\_08262014\_092201\_configs.tar

2014-08-26 09:46:04 2190643200 udb01\_08262014\_094601\_configs.tar

2014-08-26 12:02:04 2196951040 udb01\_08262014\_120201\_configs.tar

2014-08-26 09:47:09 6718627840 udb02\_08262014\_094701\_configs.tar

2014-08-26 09:57:20 2908016640 udb02\_08262014\_095701\_configs.tar

2014-08-26 10:11:21 3671490560 udb02\_08262014\_101101\_configs.tar

Check for backups of current host (e.g., host where user is performing these commands).

aws s3 ls NARA\_08.11.2014\_12.55\_config\_backups |grep `hostname`

Check for backups of a particular host.

aws s3 ls NARA\_08.11.2014\_12.55\_config\_backups |grep `hostname` |grep ps01

Check for yesterday’s backup in S3 for the current host.

aws s3 ls NARA\_08.11.2014\_12.55\_config\_backups |grep `hostname` |grep `date --date="yesterday" +%m%d`

Check for yesterday’s back up in S3 for any host by specifying the name.

aws s3 ls NARA\_08.11.2014\_12.55\_config\_backups |grep ps01 |grep `date --date="yesterday" +%m%d`

## Schedules

### MySQL Databases

Requires a daily full backup and a weekly full backup.

The OPA system will implement MySQL replication using a master server and a slave server. The replication allows us to easily maintain multiple copies of the data by automatically copying the data from the master to the slave. This will also allow for an easy backup procedure, using the mysqldump command, as the slave can be paused for a while and the data backed up without interrupting the master.

Here is the process for backing up MySQL in Production:

# Master and slave systems are done like this

# SSH to slave

# The slave does not take read requests and so is the perfect place to do this

# Verify slave status. The output from this command should be 0; anything else and replication is broken.

mkdir -p /opt/scripts/automation

cat >>/opt/scripts/automation/dumpthedb<<EOD;

#!/bin/bash

. /root/.bash\_profile

CHECK=\`mysql -u root -p@Wrk\\!N0w -e "show slave status \G" |grep Seconds\_Behind\_Master |awk '{print \$2}'\`

if [ "\$CHECK" = 0 ];

then

echo "slave is ok dumping the dbs"

mkdir -p /var/lib/mysql/backup/

mysqldump -u root -p@Wrk\!N0w --single-transaction --all-databases > /var/lib/mysql/backup/backup\_\`hostname\`\_\`date +%Y%m%d%H%M%S\`.sql

fi

# Keep 3 days of data

cd /var/lib/mysql/backup/

unalias rm

rm \`ls -t | awk 'NR>3'\`

EOD

chmod 755 /opt/scripts/automation/dumpthedb

Run it to test it and check the file (should only take a few minutes).

Put this dump into a cron job that runs nightly.

echo "1 1 \* \* \* root /opt/scripts/automation/dumpthedb &" >>/etc/crontab

The host based backup script follows behind this and back this up to s3. To recover from catastrophe create or clear a mysql master and slave as specified in the mysql installation instructions.

The master restore the most recent backup

shell> mysql < backup\_sunday\_1\_PM.sql

The backups are full systems images and are retained until they are deleted by the next full system image.

### Search Engine

* Full Backup – Monthly
  + Each monthly backup is stored indefinitely as a full machine image.
* Incremental Backup – Daily
  + Copies of all new index files
  + Files are to be copied to an offline storage volume (e.g. un-mounted)

### Content Processing / Ingestion Servers

The backups are full systems images and are retained until they are deleted by the next full system image. The log files are sent to the reporting server.

The path for the ParcDB database can be found at:

/opt/vol/ingestion/ingestion-%VERSIONNUMBER%-distribution/

data/Ingestion/parcDb/

The paths for the CrawlDB database can be found at:

/opt/vol/ingestion/ingestion-%VERSIONNUMBER%-distribution/data/Archives/

/opt/vol/ingestion/ingestion-%VERSIONNUMBER%-distribution/data/PresLibs/

The backups are full systems images and are retained until they are deleted by the next full system image.

### Application Servers

The backups are full systems images and are retained until they are deleted by the next full system image. The log files sent to the reporting server.

### Reporting, Monitoring, and Admin Control

The backups are full systems images and are retained until they are deleted by the next full system image.

## Backing Up OPA System Configuration Files

The configuration files for each system component are included in the backup strategy for OPA. Listed below are configuration files and their system paths by component.

### Content Processing / Ingestion

/opt/vol/ingestion/ingestion-%VERSIONNUMBER%-distribution/config/settings.xml

/opt/vol/ingestion/ingestion-%VERSIONNUMBER%-distribution/config/workflow-libraries/Opa.xml

/opt/vol/ingestion/ingestion-%VERSIONNUMBER%-distribution/config/content-sources/Annotations/content-source.xml

### OPA Storage

/opt/vol/ingestion/ingestion-%VERSIONNUMBER%-distribution/config/settings.xml

### Search Engine

/opt/solr/solr-4.7.2/example/solr/collection1/conf/elevate.xml

/opt/solr/solr-4.7.2/example/solr/collection1/conf/stopwords.txt

### Application Server/Annotations Database

/opt/vol/config/config.xml

(See Appendix 13.1 for a detailed copy of config.xml)

### User Interface

/www/html/scripts/<random number>config.js

# System Recovery Procedures

Recovery from site failure will require the following steps:

* Launch new copies of all server instances.
* Restore all databases from the latest backups.
* Reprocess updates since the latest backups were required.
* Re-index records as required.

It is conceivable that a complete re-index of all OPA content will be required to recover from a site-failure. If this is the case, then multiple content ingestion servers may need to be launched to reprocess records in parallel, to perform a complete re-index within 7 days. System Configuration file backups will just need to be pulled down from the S3 bucket and loaded into the proper location.

*Note: This information source for these backup procedures is from the User Guide on the AWS EC2 website. Please the link below for the most current procedures.*

<http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EBSSnapshots.html>

Copying an Amazon EBS Snapshot

Amazon Elastic Block Store (Amazon EBS) provides block level storage volumes for use with Amazon Elastic Compute Cloud (Amazon EC2) instances. With Amazon EBS, you can create point-in-time snapshots of volumes and store them on Amazon Simple Storage Service (Amazon S3). After you've stored a snapshot in Amazon S3, you can copy it from one AWS region to another, or within the same region, using the Amazon EC2 console, Amazon EC2 CLI, or the API. You can use snapshot copies to create backups of data, to create new Amazon EBS volumes, or to create Amazon Machine Images (AMIs).

## To copy a snapshot using the Amazon EC2 console

You can create a copy of an Amazon EBS snapshot using the Amazon EC2 console.

1. Open the Amazon EC2 console at <https://console.aws.amazon.com/ec2/>.
2. In the navigation pane, click **Snapshots**.
3. Select the snapshot to copy, and then select **Copy** from the **Actions** list.
4. In the **Copy Snapshot** dialog box, update the following as necessary:
   * **Destination region**: Select the region where you want to write the copy of the snapshot.
   * **Description**: By default, the description includes information about the source snapshot so that you can identify a copy from the original. You can change this description as necessary.
5. Click **Yes, Copy**.
6. In the **Copy Snapshot** confirmation dialog box, you can click **Snapshots** to go to the **Snapshots** page in the region specified, or click **Close**.

To view the progress of the copy process later, switch the Amazon EC2 console to the destination region, and then refresh the **Snapshots** page. Copies in progress are listed at the top of the page.

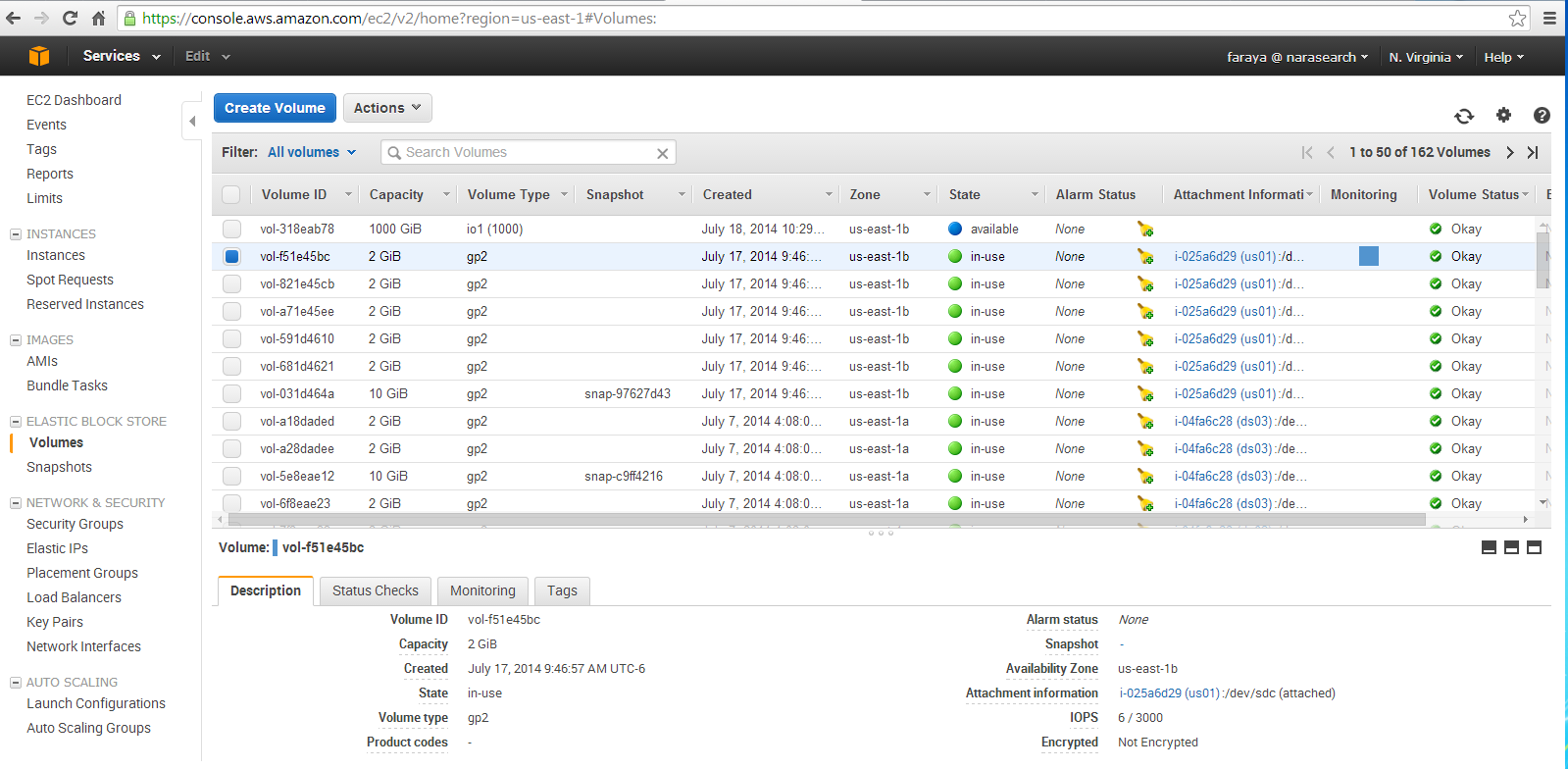


Figure 7.1 AWS EBS Volume Restoration

## To copy a snapshot using the command line

Take pw03 out of the Elastic Load Balancer

Must be DHCP (Configured this way to reboot as an AMI image)

on pa03 modify

/etc/sysconfig/network-scripts/ifcfg-eth0

It should just contain this when done

Reboot

DEVICE="eth0"

BOOTPROTO="dhcp"

ONBOOT="yes"

TYPE="Ethernet"

USERCTL="yes"

PEERDNS="no"

IPV6INIT="no"

PERSISTENT\_DHCLIENT="1"

Host is ready to have an image made of it. Stop the instance via sudo halt. Once it is stopped, make an ami of it.

Get the instance id

ec2-describe-instances --region us-east-1 |grep "pa03"

TAG instance i-8f3f09a4 Name pa03

Make an ami of it.

ec2-create-image --region us-east-1 i-8f3f09a4 -n pa03recovery -d ps03recovery

IMAGE ami-4832e020 (automatically named)

Check the ami into the console and verify it is done before you make a new instance from it

https://console.aws.amazon.com/ec2/v2/home?region=us-east-1#Images:

If it says pending, it is not ready. Wait for it to complete. This may take some time.

###################################

#

# create the new system from the ami

#

####################################

Obtain the security group for the current system.

C: ec2-describe-instances --region us-east-1 i-8f3f09a4 |grep sg

Output: INSTANCE i-8f3f09a4 ami-b06a98d8 ip-10-135-4-59.ec2.internal stopped key10.135 0 m3.xlarge 2014-08-13T10:06:47+0000 us-east-1b aki-919dcaf8 monitoring-disabled 10.135.4.59 vpc-b768b6d2 subnet-63c63d14 ebs paravirtual xen VWVGq1405631197513 sg-522f5037 default false

GROUP sg-522f5037 PROD\_APP\_SERV\_SG

Obtain the subnet.

C: ec2-describe-instances --region us-east-1 i-8f3f09a4 |grep subnet

INSTANCE i-8f3f09a4 ami-b06a98d8 ip-10-135-4-59.ec2.internal stopped key10.135 0 m3.xlarge 2014-08-13T10:06:47+0000 us-east-1b aki-919dcaf8 monitoring-disabled 10.135.4.59 vpc-b768b6d2 subnet-63c63d14 ebs paravirtual xen VWVGq1405631197513 sg-522f5037 default false

NIC eni-fdccd78a subnet-63c63d14 vpc-b768b6d2 309403645338 in-use 10.135.4.59 ip-10-135-4-59.ec2.internal true

ec2-run-instances ami-4832e020 -n 1 -k key10.135 -g sg-522f5037 --instance-type m3.large --subnet subnet-63c63d14

See URL below for information on instance types.

<http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/instance-types.html>

# It will come up on a new IP address that you will show up in the console.

RESERVATION r-da36e1f1 309403645338

INSTANCE i-532ef579 ami-4832e020 ip-10-135-4-63.ec2.internal pending key10.135 0 m3.large 2014-08-13T12:13:33+0000 us-east-1b aki-919dcaf8 monitoring-disabled 10.135.4.63 vpc-b768b6d2 subnet-63c63d14 ebs paravirtual xen sg-522f5037 default false

NIC eni-c061a3b6 subnet-63c63d14 vpc-b768b6d2 309403645338 in-use 10.135.4.63 ip-10-135-4-63.ec2.internal true

NICATTACHMENT eni-attach-1aa42379 0 attaching 2014-08-13T12:13:33+0000 true

GROUP sg-522f5037 PROD\_APP\_SERV\_SG

PRIVATEIPADDRESS 10.135.4.63 ip-10-135-4-63.ec2.internal

# ...10.135.4.63

# Check the console make sure it is ok

#

# This system took a good 10 minutes to be available. Via ssh, you can check when it is running.

while [ 0 -eq 0 ]; do nc -z 10.135.4.63 22; sleep 15;date; done

ssh -i mykey ec2-user@10.135.4.63

Change the hostname

vi /etc/sysconfig/network

NETWORKING=yes

**HOSTNAME=pa03**

NETWORKING\_IPV6=no

NOZEROCONF=yes

Reboot to make sure the hostname change sticks.

Change the sshd config to allow users with passwords to log in.

vi /etc/ssh/sshd\_config

PasswordAuthentication yes

sudo service sshd restart

ssh jcourtade@localhost

Ensure you can log in.

Also, log in from jump as your user, ssh to new host from jump, verify you can get in.

[ec2-user@ip-10-135-1-199 ~]$

[ec2-user@ipa03 ~]$

# Change the hostname to the new IP address of the restored system in DNS.

ssh to pw01

sudo su -

cd /var/named

vi master.aws.narasearch.us

Change the serial number

20140807145800 (yearmonthdayhourminsecond)

Change the forward entry to match the new IP address

pa05 IN A 10.135.5.230

pa05 IN A 10.135.5.245

Change the proper reverse record by editing appropriate zone file

reverse.1.135.10 reverse.2.135.10 reverse.3.135.10 reverse.4.135.10 reverse.5.135.10

vi reverse.4.135.10

Change the serial number

20140807145800 yearmonthdayhourminsecond

Change the entry for the host so the last octet number matches. The new host is 10.135.4.63

63 IN PTR pa03.aws.narasearch.us. (reverse) [need period on end]

63 IN PTR pa03.aws.narasearch.us.

10.135.4.63

10.135.4.245

245 IN PTR hostyadyayda.

Restart named and check the entries.

sudo service named restart

Check both name servers for both addresses.

Name servers are pw01 and pw02

Below is an example:

[root@pw01 named]# host 10.135.4.190 pw01

Using domain server:

Name: pw01

Address: 10.135.1.50#53

Aliases:

[root@pw01 named]# host pa04 localhost

Using domain server:

Name: localhost

Address: 127.0.0.1#53

Aliases:

pa04.aws.narasearch.us has address 10.135.4.190

[root@pw01 named]# host pa04 pw02

Using domain server:

Name: pw02

Address: 10.135.1.51#53

Aliases:

pa04.aws.narasearch.us has address 10.135.4.190

check reverse as well

host 10.135.4.190 localhost

host 10.135.4.190 pw02

190.4.135.10.in-addr.arpa domain name pointer pa04.aws.narasearch.us.

[root@pw01 named]# host 10.135.4.190 pw02

Using domain server:

Name: pw02

Address: 10.135.1.51#53

Aliases:

190.4.135.10.in-addr.arpa domain name pointer pa04.aws.narasearch.us.

If both servers don’t have the correct info, try restarting the name servers on both hosts.

Tag the instance in the AWS console with the hostname: pa03new

Tag the old instance: pa03old

Once the new instance is running well (couple of days), delete the old host in the AWS console.

Change the tag to pa03 (DNS host name).

## Individual System Component Recovery

The recovery method will depend on the type of server.

### Database Servers

A primary and a failover exist for the MySQL database servers.

A failure of either server will mean operating with a single server until its sibling server can be restored and the database mirrored.

### Content Processing / Ingestion Servers

* Recovery times for content processing servers are longer (2 days) than for other servers (90 minutes). Therefore, the recovery process for content processing will be:

###############################################

# process to swap out cp servers

###############################################

# schedule maintenance if possible do one appserver at a time.

# this will allow for minimized to 0 downtime.

# On ALL appservers stop the exportserver as this will prevent possible write issues.

#

sudo su -

service exportserver stop

#

# then do one appserver at a time

# stop tomcat

# swap out mount points

# start tomcat and exportserver

#

sudo su -

service tomcat stop

# umount the NFS mounts

umount /opt/vol/opastorage

umount /opt/vol/xmlstore

umount /opt/vol/bulk

umount /opt/vol/nonbulk

#

# modify /etc/fstab

# the server name needs to be changed to the new cp server

# the mounts stay the same

# as an example change pcp01 to pcp02

#

# this is the original

#

pcp01:/opt/vol/opastorage /opt/vol/opastorage nfs rsize=8192,wsize=8192,timeo=14,intr

pcp01:/opt/vol/xmlstore /opt/vol/xmlstore nfs rsize=8192,wsize=8192,timeo=14,intr

pcp01:/opt/vol/bulk /opt/vol/bulk nfs rsize=8192,wsize=8192,timeo=14,intr

pcp01:/opt/vol/nonbulk /opt/vol/nonbulk nfs rsize=8192,wsize=8192,timeo=14,intr

#

# this is what the new entries should be

#

pcp02:/opt/vol/opastorage /opt/vol/opastorage nfs rsize=8192,wsize=8192,timeo=14,intr

pcp02:/opt/vol/xmlstore /opt/vol/xmlstore nfs rsize=8192,wsize=8192,timeo=14,intr

pcp02:/opt/vol/bulk /opt/vol/bulk nfs rsize=8192,wsize=8192,timeo=14,intr

pcp02:/opt/vol/nonbulk /opt/vol/nonbulk nfs rsize=8192,wsize=8192,timeo=14,intr

#

# mount all the new mounts and check them

#

mount -a

df -h

#

# this should show the new mounts

#

pcp02:/opt/vol/opastorage

                       14T  6.9T  6.5T  52% /opt/vol/opastorage

pcp02:/opt/vol/xmlstore

                       14T  6.9T  6.5T  52% /opt/vol/xmlstore

pcp02:/opt/vol/bulk   493G  251G  237G  52% /opt/vol/bulk

pcp02:/opt/vol/nonbulk

                      493G  200M  487G   1% /opt/vol/nonbulk

#

# start the appserver

#

service tomcat start

#

# start the exportserver

#

service exportserver start

#

# comment out the sync in the pcp02 croontab.

# and save it

#

vi /etc/crontab

# 10 0 \* \* \* root nohup /opt/scripts/automation/syncit &

#

# enable the sync on the NEW? Slave server by uncommenting that line from the # # #/etc/crontab

### Search Engine Servers

A primary and failover server exists for each search engine server.

Therefore, a failure of either server will mean operating with a single server for the specified index partition until its sibling server can be restored and the index copied.

#################################################################

# search two different scripts

# one for the primary and one for the secondary

#################################################################

sudo su -

# add this to the /etc/crontab

# with the other variables at the top

vi /etc/crontab

AWS\_CONFIG\_FILE="/root/.aws/config"

AWS\_ACCESS\_KEY\_ID=AKIAISMJSLEX6OKSURZA

AWS\_SECRET\_ACCESS\_KEY=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

mkdir -p /opt/scripts/backup

vi //opt/scripts/backup/backitup

##############################################

# search primary zookeeper

##############################################

#!/bin/bash

export HOST=`hostname`

export DATSTR=`date +%m%d%Y\_%H%M%S`

export FILENAME=`hostname`\_"$DATSTR"\_configs.tar

BUCKET=s3://NARA\_08.11.2014\_12.55\_config\_backups

FILELIST="/etc/sysconfig/network /etc/resolv.conf \

/etc/passwd /etc/shadow /etc/crontab \

/etc/group /etc/sudoers \

/opt/fitnesse/ /opt/solr/ \

/etc/init.d/fitnesse /etc/init.d/solr \

/etc/init.d/zookeeper"

# back the thing up

cd /

mkdir /opt/backitup/

tar cfv /opt/backitup/$FILENAME $FILELIST --exclude="opt/fitnesse/logfile" \

--exclude="\*/solr/solr\*/example/logs/\*" \

--exclude="opt/zookeeper/zookeeper-3.4.5/zookeeper.out"

aws s3 cp /opt/backitup/$FILENAME $BUCKET

# remove the local archive

rm -f /opt/backitup/$FILENAME

# clean up the buckets

# remove all but the most recent 30

OLDFILES="`aws s3 ls $BUCKET/|grep $HOST|head -n-30|awk '{print $4}'`"

echo $OLDFILES

for m in $OLDFILES

do

echo delteing $m

aws s3 rm $m

done

chmod 755 /opt/scripts/backup/backitup

echo "5 \* \* \* \* root cp /dev/null /opt/solr/solr-4.7.2/example/nohup.out">>/etc/crontab

# on prod dev and sandbox systems

echo "25 2 \* \* \* root /opt/scripts/backup/backitup" >>/etc/crontab

# UAT times have to be different so they get backed up during the day

echo "10 7 \* \* \* root /opt/scripts/backup/backitup" >>/etc/crontab

##############################################

# search NOT primary zookeeper host

##############################################

sudo su -

mkdir -p /opt/scripts/backup

vi //opt/scripts/backup/backitup

##############################################

# search NOT primary zookeeper host

##############################################

#!/bin/bash

export HOST=`hostname`

export DATSTR=`date +%m%d%Y\_%H%M%S`

export FILENAME=`hostname`\_"$DATSTR"\_configs.tar

BUCKET=s3://NARA\_08.11.2014\_12.55\_config\_backups

FILELIST="/etc/sysconfig/network /etc/resolv.conf \

/etc/passwd /etc/shadow /etc/crontab \

/etc/group /etc/sudoers \

/opt/solr/ /etc/init.d/solr "

#

# /opt/solr/ /etc/init.d/solr "

#

# this is only needed on the primary in each env

# /etc/init.d/fitnesse /etc/init.d/solr \

# /etc/init.d/zookeeper /opt/fitnesse/ \

# /opt/zookeeper/ /opt/solr/"

# back the thing up

cd /

mkdir /opt/backitup/

tar cfv /opt/backitup/$FILENAME $FILELIST --exclude="opt/fitnesse/logfile" \

--exclude="\*/solr/solr\*/example/logs/\*" \

--exclude="opt/zookeeper/zookeeper-3.4.5/zookeeper.out"

aws s3 cp /opt/backitup/$FILENAME $BUCKET

# remove the local archive

rm -f /opt/backitup/$FILENAME

# clean up the buckets

# remove all but the most recent 30

OLDFILES="`aws s3 ls $BUCKET/|grep $HOST|head -n-30|awk '{print $4}'`"

echo $OLDFILES

for m in $OLDFILES

do

echo delteing $m

aws s3 rm $m

done

chmod 755 /opt/scripts/backup/backitup

echo "5 \* \* \* \* root cp /dev/null /opt/solr/solr-4.7.2/example/nohup.out">>/etc/crontab

# on prod dev and sandbox systems

echo "25 2 \* \* \* root /opt/scripts/backup/backitup" >>/etc/crontab

# UAT times have to be different so they get backed up during the day

echo "10 7 \* \* \* root /opt/scripts/backup/backitup" >>/etc/crontab

### Application Servers

Application servers can be recovered at any time by simply launching a new instance of the server and adding it to the server farm. No recovery of backups is required.

### Reporting, Monitoring and Admin Control

Reporting servers can be restoring the applicable system component log files, restoring Splunk configuration files, and launching a new instance of Splunk Enterprise.

ZooKeeper servers can be recovered at any time by simply launching a new instance of the server and adding it to the server farm. See Section 8.1 concerning removing old snapshots and log files from the data directory.

## Restoring OPA System Configuration Files

The configuration files for each system component are used to provide recovery for OPA. Listed below are the applicable configuration files and their correct system paths by component.

### Content Processing / Ingestion

/opt/vol/ingestion/ingestion-%VERSIONNUMBER%-distribution/config/settings.xml

/opt/vol/ingestion/ingestion-%VERSIONNUMBER%-distribution/config/workflow-libraries/Opa.xml

/opt/vol/ingestion/ingestion-%VERSIONNUMBER%-distribution/config/content-sources/Annotations/content-source.xml

### OPA Storage

/opt/vol/ingestion/ingestion-%VERSIONNUMBER%-distribution/config/settings.xml

### Search Engine

/opt/solr/solr-4.7.2/example/solr/collection1/conf/elevate.xml

/opt/solr/solr-4.7.2/example/solr/collection1/conf/stopwords.txt

### Application Server/Annotations Database

/opt/vol/config/config.xml

(See Appendix 13.1 for a detailed copy of config.xml)

### User Interface

/www/html/scripts/<random number>config.js

# System Maintenance

## Periodic Procedures

### Optimizing the search index in Solr

The index will be need to be optimized whenever there is a new ingestion of data. The optimize command impacts Solr performance (e.g. slows down), so it will need to performed during an off-peak time as part of scheduled maintenance.

### Recompiling the spell check dictionary in Solr

The spell check dictionary is built during the Solr commit process after the first big ingestion of data. The dictionary is index-based, meaning the suggestion are based upon index entries.

### Updating Stop Words (Stopwords.txt) in Solr

If you are looking to edge out a bit of performance from Solr Cloud, one of the many things you can do is optimize your stop words file. The more entries you have in this file means there are fewer terms that end up in the Solr index. The stop words configuration file is a text file. Solr should be re-indexed and restarted for any updates to this file to be effective. For production release 1, this file will be empty.

See Section 7.4.3 for the file location in the system directory.

### Updating Featured Results (Elevate.xml) Solr

Solr has a component that can configure the top results for a given query regardless of the normal Lucene scoring. It is able to match the user query text to a configured map of top results. The text can be any string or non-string IDs, as long as it's indexed.

For each term XYZ for which you want to configure featured results

* Add the term XYZ in the query text
* Add the corresponding document ID’s.

These will be OPA ID’s. These can be obtained via the Search Engine. The following example shows Featured Results Configuration for term “harry”.

|  |
| --- |
| <elevate>  <query text="harry">  <doc id="person-10655016" />  <doc id="desc-1721336" />  </query>  </elevate> |

The documents will be returned in the order they are listed in the elevate.xml file.

Following these featured documents, the remainder of the results will follow as per relevancy or sort order.

See Section 7.4.3 for the file location in the system directory.

### Solr Best Practices

Whenever a SoLR restart is required, it is important to make sure the Ingestion feed is paused or SoLR will enter a corrupted state, and will produce the error "No leader found". This is due to the fact that the Ingestion feed always looks for the leader.

### ZooKeeper

* + Ongoing Data Directory Cleanup

The ZooKeeper Data Directory contains files which are a persistent copy of the znodes stored by a particular serving ensemble. These are the snapshot and transactional log files. As changes are made to the znodes, these changes are appended to a transaction log. Occasionally, when a log grows large, a snapshot of the current state of all znodes will be written to the file system. This snapshot supersedes all previous logs.

A ZooKeeper server **will not remove old snapshots and log files** when using the default configuration, this is the responsibility of the operator. Zookeeper's configuration file will have the setting to control where the Zookeeper data is stored. This is currently found in dataDir=/opt/zookeeper/data. System Administrators can remove the old files as necessary to ensure Zookeeper uses the proper files.

Please reference the ZooKeeper documentation for maintenance details at the following URL: <http://zookeeper.apache.org/doc/trunk/>

### Frequency

|  |  |
| --- | --- |
| **Maintenance Item** | **Frequency** |
| Optimizing the search index | After each data ingestion |
| Recompiling the spell check dictionary | After several data ingestions |
| Updating Stop Words (Stopwords.txt) | As necessary |
| Updating Featured Results (Elevate.xml) | As necessary |
| ZooKeeper Ongoing Data Directory Cleanup | As necessary |
| ZooKeeper Debug Log Cleanup (log4j) | As necessary |

## System Configuration Changes

### Config.xml

If any changes need to be made to the config.xml file (see Section 13.1), then Tomcat must be stopped, changes made to config.xml, and restarted. The commands below provide guidance:

----------------------------

ssh ua01,02,03,04:

----------------------------​

sudo service tomcat stop

ps -ef | grep apache  (make sure it was killed)

edit /opt/vol/config/config.xml

ls -al (if not tomcat:tomcat --> chown tomcat:tomcat config.xml)

sudo service tomcat start

*Note: This operation must be performed on all OPA Application Servers.*

### User Interface Configuration Properties File

If any changes need to be made to the UI Configuration Properties File (see Section 13.2), the user’s browser will need to be refreshed for change to become effective in OPA.

## Monitoring System Performance

Performance monitoring metrics will be collected using Amazon Cloud Watch. This service will include basic monitoring of CPU utilization, data transfer, and disk usage activity.

## Monitoring and Reporting System Activity

Splunk Enterprise is the tool being utilized by OPA for monitoring and reporting purposes. The data for reports will be obtained from various server log files and from the OPA MySQL database. The following diagram depicts the how the reporting system obtains the system log files and database table information needed to generate the NARA-specified reports. User documentation for Splunk Enterprise can be found at: <http://docs.splunk.com/Documentation/Splunk>.

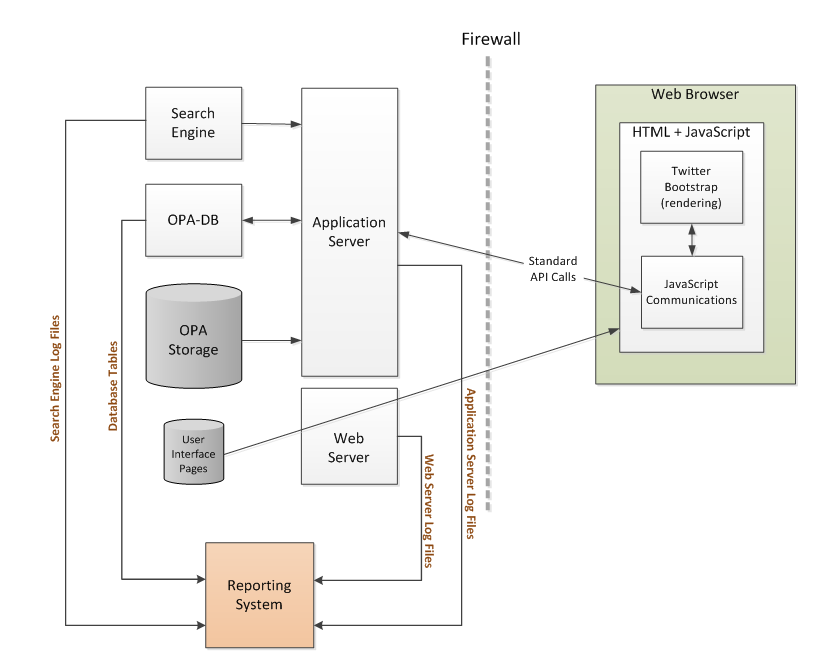


Figure 8.1: OPA Reporting System (Splunk) Interface

### Data Sources

* System Logs
* Error Logs

### Users

System administrators have the ability to create user accounts in Splunk. There are two type of authorized users permitted by OPA in Splunk.

* System Administrators
* Reporters

Both authorized users have the following permissions in Splunk:

* Able to generate, save and print reports
* Able to modify reporting period for all OPA system reports

In addition to the permissions listed above, System Administrators are able configure the settings on the reports listed below:

* Configure Top N Count
  + Top 10 statistics specific to a report
  + Top 100 statistics in specific to a report
  + Top 1,000 statistics in specific to a report
* Configure Time Based Reports
  + Weekly
  + Monthly
  + Yearly
  + Since report inception

See Appendix 13.6 for a complete listing of Reports.

### Splunk Administration

Splunk Enterprise User Documentation can be found at:

<http://docs.splunk.com/Documentation/Splunk>

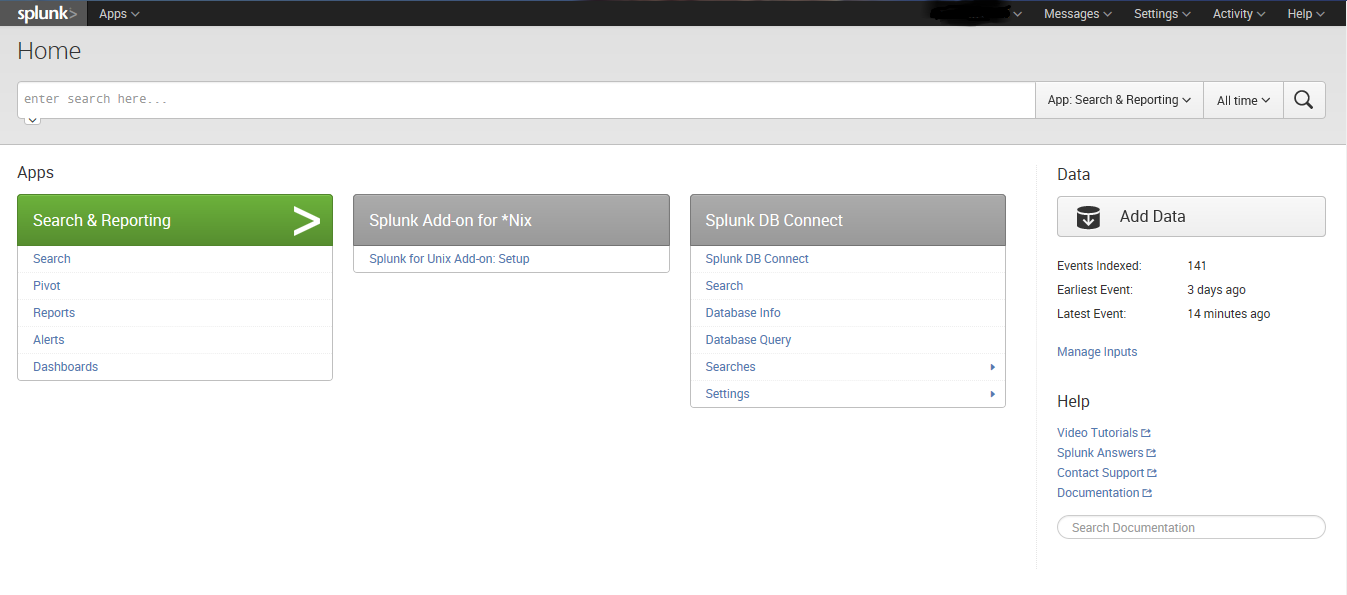


Figure 8.2: Splunk Home Page

#### Users

On the Home page, select settings, then select access controls to add/edit/delete/clone users. Modifying user roles can also be done on this page.

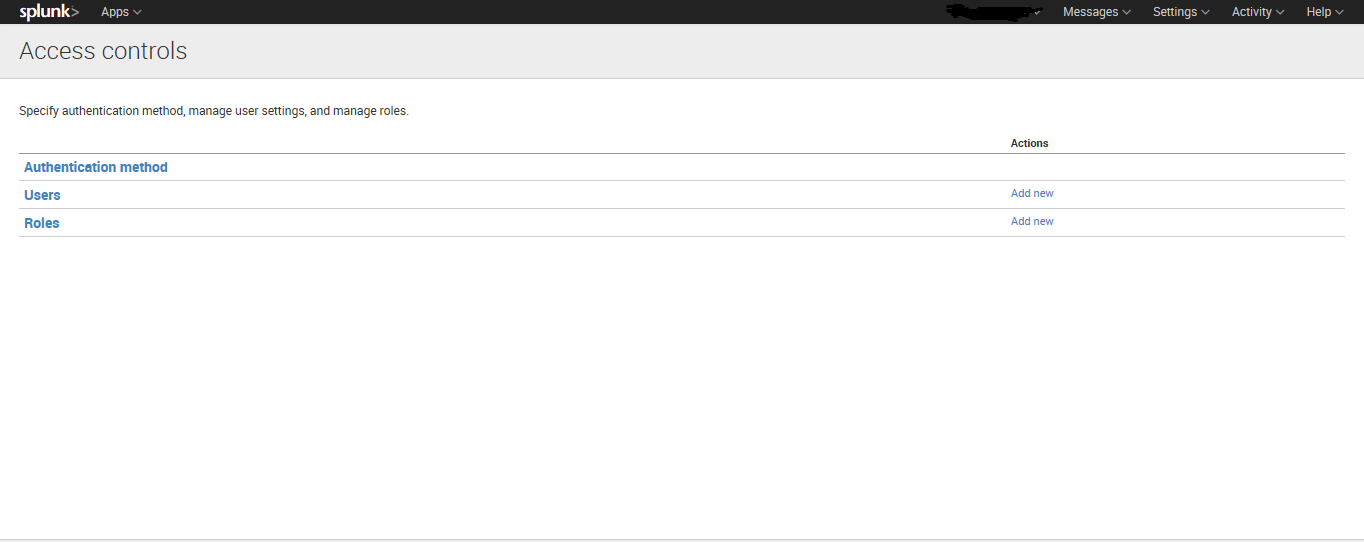


Figure 8.3: Splunk User Administration

#### Search

Search is the primary way users navigate data in Splunk Enterprise. You can write a search to retrieve events from an index, use statistical commands to calculate metrics and generate reports, search for specific conditions within a rolling time window, identify patterns in your data, predict future trends, and so on. Searches can be saved as reports and used to power dashboards.

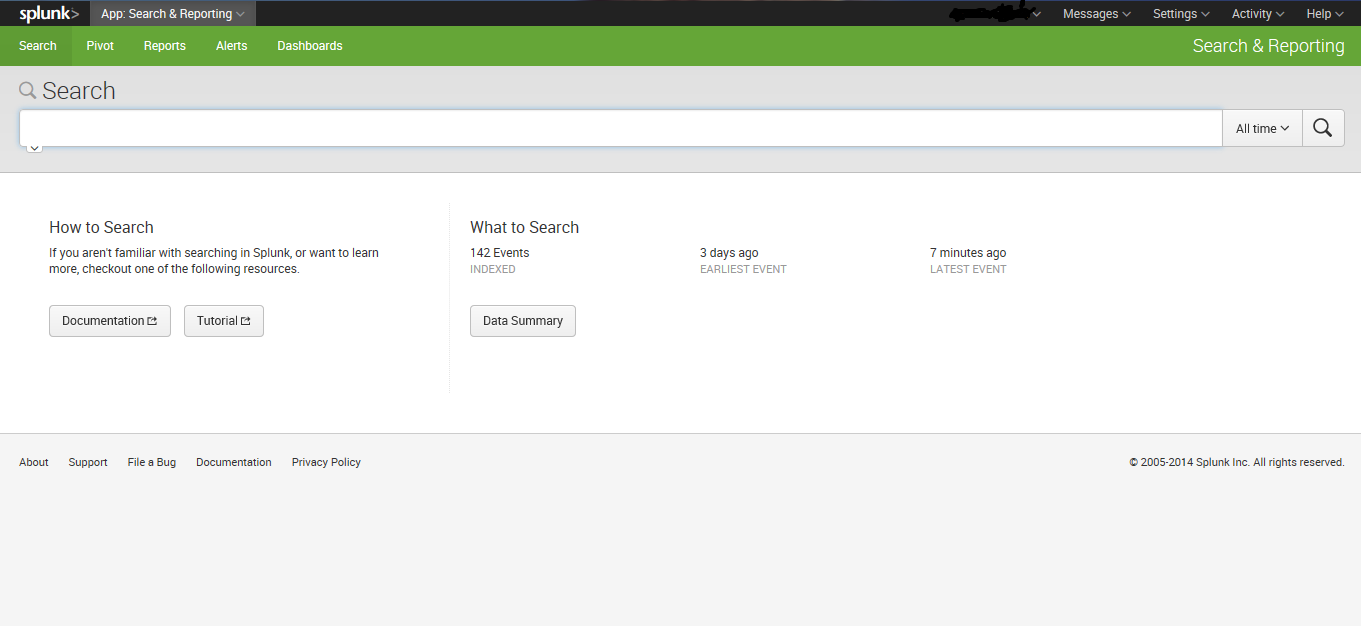


Figure 8.4: Splunk Search Page

#### Reports

Reports are saved searches and pivots. You can run reports on an adhoc basis, schedule them to run on a regular interval, set a scheduled report to generate alerts when the results of their runs meet particular conditions. Reports can be added to dashboards as dashboard panels.

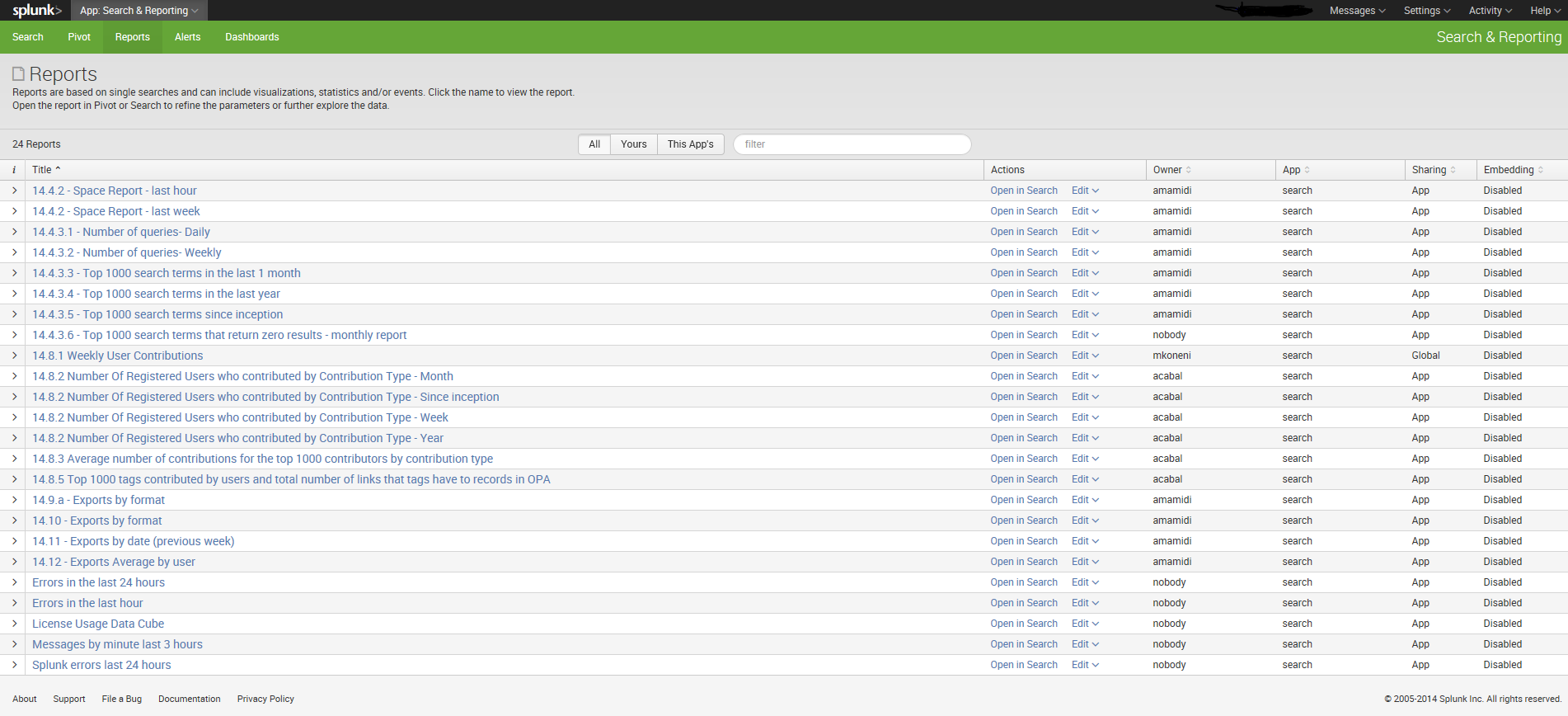


Figure 8.5: Splunk Reports Tab

#### Alerts

Alerts are triggered when conditions are met by search results for both historical and real-time searches. Alerts can be configured to trigger actions such as sending alert information to designated email addresses, post alert information to an RSS feed, and run a custom script, such as one that posts an "alert event" to syslog.

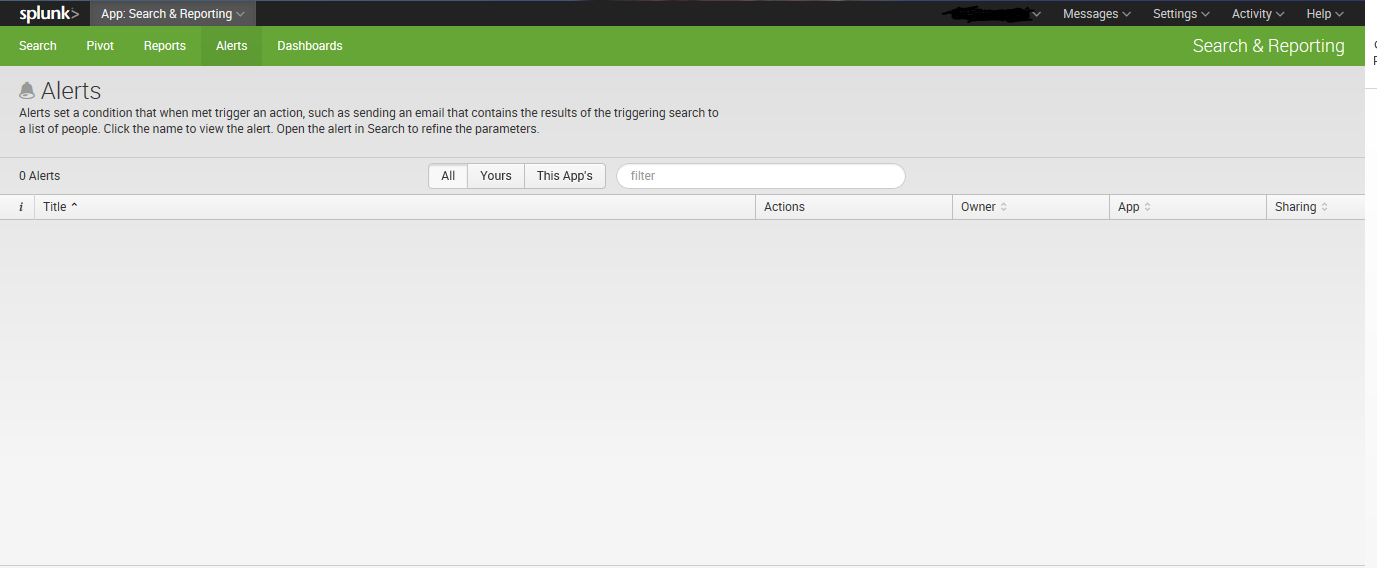


Figure 8.6: Splunk Alerts Page

#### Dashboards

Dashboards are made up of panels that contain modules such as search boxes, fields, charts, tables, forms, and so on. Dashboard panels are usually hooked up to saved searches or pivots. They can display the results of completed searches as well as data from background real-time searches.

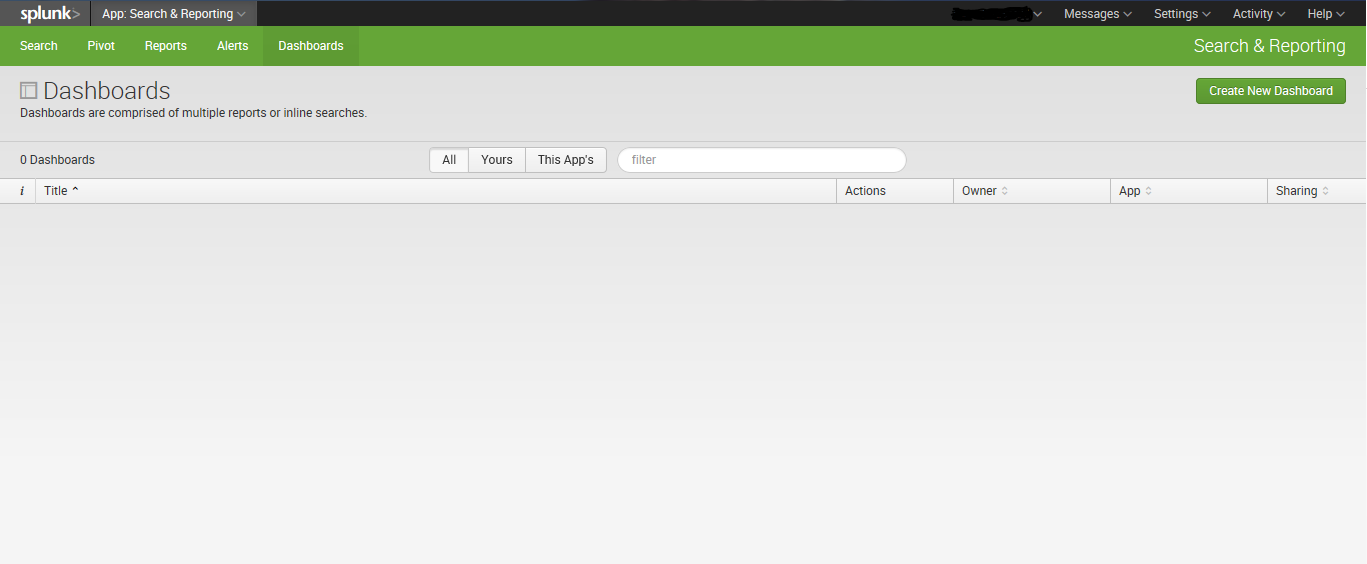


Figure 8.7: Splunk Dashboards Page

## Maintaining Audit Records of System Operation

Auditing will be done with the help of the NARA security team, and based on NARA Linux recommended configurations.

We anticipate that this will include:

* Administrator auditing with the “psacct” package and perhaps other packages (such as rootsh logging)
* Appropriate configuration of syslogd, including auditing of successful and unsuccessful account events
* Verification of logs generated to /var/log/security and /var/log/audit/audit.log
* Protection of logs from unauthorized modification
* Capturing operating system errors

## Software Updates and Patches

The operating system will be maintained and patched using the cloud-recommended procedures. All Red Hat Linux operating system updates and patches will be obtained and run manually from Amazon’s Red Hat servers.

This may involve:

* Halt updates to the system.
  + For example, turn of index updates by the ingestion servers.
  + This will limit the amount of data which is changing. Most servers will now have “idle” systems with files that do not change.
* Taking the server to be patched off-line.
* Patching the operating system as necessary.
* Bringing the server back on-line.
* Re-synchronize database files as necessary.
  + If the ingestion servers are idle, then search engine indexes, application servers, and ingestion servers will not require re-synchronization.
  + Therefore, only the RDBMS may still be receiving updates that require synchronization, when a database server is taken off-line, patched, and then brought back on-line.

All critical servers have fail-over siblings which will allow for either one or the other to be ought off-line for patching as necessary.

# Security

The OPA system will adhere to NARA Information System security policies and guidelines. In addition, the AWS EC2 service provides addition security controls via its AWS Management Console.

## Issuing IDs and Passwords

Initially, this will be handled by Search Technologies. Once system handoff occurs, it will be up to NARA’s O&M team. Search Technologies will comply with NARA’s Information System Security Policy concerning User IDs and Passwords.

3 different kinds of users

* AWS Identity and Access Management (IAM) for the Amazon Management Console. This user will be able to view CloudWatch metrics, start/stop instances, make snapshots, etc. <http://docs.aws.amazon.com/IAM/latest/UserGuide/IAMBestPractices.html>
* Linux users have a local password on every machine for which they have access.
* Main users will also need an account on the jump proxy server.

## Maintaining Privileged User Accounts

Root has no password and no keys anywhere. The sysadmin account and people needing elevated privileges is handled with sudo.

## Installing and Maintaining Certificates

The only certificates in use are on the Amazon Elastic Load Balancers.

<http://docs.aws.amazon.com/ElasticLoadBalancing/latest/DeveloperGuide/US_UpdatingLoadBalancerSSL.html>

## License Agreements

There are several software applications utilized in OPA which require licensing (e.g. annual or one-time). For reference, they are listed below:

* Aspire Enterprise
* MySQL Enterprise
* Splunk Enterprise
* Glyphicons

# Scalability

## Hardware

* Dependent upon Amazon Web Services
* Adding search server rows for additional Queries per Second (QPS) will require about 2 weeks.
  + Machine instances can be created to launch servers quickly.
  + Servers can be added as new “slave replicas” with simple configuration.
  + Two weeks would be required for initial index replication, testing, and to account for possible roll-backs and re-attempts should something go wrong.
* Adding additional index partitions for additional content will require about six weeks
  + Machine instances can be created to launch servers quickly.
  + Servers can be added as new “partitions” with simple configuration
  + Six weeks would be required to re-balance the documents across the partitions
* Adding new application servers for additional end-user capacity will require about three days
  + New application servers can be added at any time
  + No complex data replications are required (they all share a master database)
  + All API and UI transactions are stateless (state is carried in cookies and on the client)

Note that these times (2 weeks for a search row, 6 weeks for additional index partitions, and 3 days for additional application servers) could be reduced with additional testing, scripting, and process documentation.

## Large Ingestion of Data

### Downloading of DAS Exports

Every Friday evening the DAS XML Export is kicked off at 8:30 pm EDT time. When the export is complete, an email will be auto-generated and sent to email addresses which will be provided by NARA.

The export file will always be placed on [http://opa.nara.ppc-cloud.com/dasexport/](http://opa.nara.ppc-cloud.com/dasexport/dasexport_7-27-14.tar.gz)

The file naming convention will be **dasexport\_mm/dd/yyyy.tar.gz** where the date is the date that the export was initiated.

The System Administrator should download the appropriate file into Xmlstore/transfer. Once the download is complete, and the System Administrator wants the Ingestion System to pick up the export, they will need to move the file to Xmlstore/incremental, so the Ingestion system will recognize and process the file (within a minute).

### Full Ingestion

#### DAS to Xmlstore

OPA Ingestion System will write successfully processed ADs/ARs to Xmlstore/full and problematic ADs/ARs to Xmlstore/quarantine. Currently, DAS metadata exports will be done weekly and OPA will process it right away.

During the processing of ADs and ARs, the parcDB will automatically be built by the Ingestion system.

When DAS metadata is processed by OPA, it will be looking for any associated digital objects in their expected directory path inside the opastorage/pre-ingestion directory for processing. Currently, NARA is organizing the objects into the appropriate directories prior to the DAS bulk export being sent to OPA.

#### Websites

The WebCrawler is crawling a list of seed URLs, specifically Archives.gov and the presidential library websites. The WebCrawler downloads and extracts content for each link it encounters. It records its status (e.g. what web pages have been crawled and when) in its database, and caches the latest version of each website crawled in case a repopulation of the Search Array is necessary.

#### Content Storage

Content storage (e.g. digital media objects) is transmitted to the OPA pre-ingestion staging area via SFTP. Currently, objects are sent with a directory path that is replicated on pre-ingestion. For example,

*Media.archives.gov/presidential-libraries/truman/letter.jpg*

The web address is stripped off and remainder of the path is kept. The objects are then organized into this directory. For example,

*presidential-libraries/truman/letter.jpg*

These paths or “links” are added to the DAS description. If these links are added to DAS before the digital objects are available, this is considered an error. Any errors are logged and fed to Splunk (OPA’s Reporting Tool). Splunk can be set up to generate an alert, which notifies System Administrators.

# Escalation Issues

## Technical Support

Search Technologies will handle Tier Level 1, 2, and 3 support issues during a 45-day transition period. It is expected that NARA’s Operation and Maintenance will take over Tier 1 and 2 duties after this transition period, while Search Technologies will continue to provide Tier 3 support.

# Troubleshooting Issues

## AR/AD Ingestion Failure (Quarantine) / Digital Object Ingestion Failure

During the Ingestion process, problematic data (e.g., records, descriptions, objects) that fail to process properly are placed in the quarantine folder. The system will generate an error message and log it in two places:

* A XML file with the term “-result” added
* System error log

Below is an example of a sample error file (7603109-result):

<?xml version="1.0"?>

[<result outcome="**error**" jobId="**pdfset/7603109.xml**"><error factory="**aspire-application**" errorType="**other**" component="**/Ingestion/ingestion-main**" code="**invalid uri**">](file:///C:\Users\bstahl\Documents\NARA\QA\7603109-result.xml)<message>java.net.URISyntaxException: Illegal character in path at index 57: http://media.nara.gov/dc-metro/rg-237/7419198-9-11-faa/5. Miscellaneous/(e) Other/5 DCC 474 FLIGHT LIST JFK EWR LGA LIST.pdf</message><trace>AspireException(invalid uri): com.searchtechnologies.aspire.services.AspireException: java.net.URISyntaxException: Illegal character in path at index 57: http://media.nara.gov/dc-metro/rg-237/7419198-9-11-faa/5. Miscellaneous/(e) Other/5 DCC 474 FLIGHT LIST JFK EWR LGA LIST.pdf at gov.nara.opa.ingestion.ObjectsXml.getPathAndFilename(ObjectsXml.java:265) at gov.nara.opa.ingestion.ObjectsXml.createObjectsXmlObject(ObjectsXml.java:86) at gov.nara.opa.ingestion.ObjectsXml.addObjectArray(ObjectsXml.java:136) at gov.nara.opa.ingestion.WriteObjectsXmlStage$ObjectsXmlFileWriter.createObjectsXml(WriteObjectsXmlStage.java:58) at gov.nara.opa.ingestion.WriteObjectsXmlStage$ObjectsXmlFileWriter.writeFile(WriteObjectsXmlStage.java:41) at gov.nara.opa.ingestion.WriteObjectsXmlStage$ObjectsXmlFileWriter.execute(WriteObjectsXmlStage.java:31) at gov.nara.opa.ingestion.WriteObjectsXmlStage.process(WriteObjectsXmlStage.java:80) at com.searchtechnologies.aspire.application.JobHandlerImpl.runNested(JobHandlerImpl.java:158) at com.searchtechnologies.aspire.application.PipelineManagerImpl.process(PipelineManagerImpl.java:268) at com.searchtechnologies.aspire.application.BranchInfoImpl.processJob(BranchInfoImpl.java:580) at com.searchtechnologies.aspire.application.JobHandlerImpl.processBranch(JobHandlerImpl.java:326) at com.searchtechnologies.aspire.application.JobHandlerImpl.runNested(JobHandlerImpl.java:276) at com.searchtechnologies.aspire.application.PipelineManagerImpl.process(PipelineManagerImpl.java:268) at com.searchtechnologies.aspire.application.BranchInfoImpl.processJob(BranchInfoImpl.java:580) at com.searchtechnologies.aspire.application.JobHandlerImpl.processBranch(JobHandlerImpl.java:326) at com.searchtechnologies.aspire.application.JobHandlerImpl.runNested(JobHandlerImpl.java:253) at com.searchtechnologies.aspire.application.PipelineManagerImpl.process(PipelineManagerImpl.java:268) at com.searchtechnologies.aspire.application.BranchInfoImpl.processJob(BranchInfoImpl.java:580) at com.searchtechnologies.aspire.application.JobHandlerImpl.processBranch(JobHandlerImpl.java:326) at com.searchtechnologies.aspire.application.JobHandlerImpl.runNested(JobHandlerImpl.java:276) at com.searchtechnologies.aspire.application.JobHandlerImpl.run(JobHandlerImpl.java:80) at java.util.concurrent.ThreadPoolExecutor.runWorker(Unknown Source) at java.util.concurrent.ThreadPoolExecutor$Worker.run(Unknown Source) at java.lang.Thread.run(Unknown Source) Caused by: java.net.URISyntaxException: Illegal character in path at index 57: http://media.nara.gov/dc-metro/rg-237/7419198-9-11-faa/5. Miscellaneous/(e) Other/5 DCC 474 FLIGHT LIST JFK EWR LGA LIST.pdf at java.net.URI$Parser.fail(Unknown Source) at java.net.URI$Parser.checkChars(Unknown Source) at java.net.URI$Parser.parseHierarchical(Unknown Source) at java.net.URI$Parser.parse(Unknown Source) at java.net.URI.<init>(Unknown Source) at gov.nara.opa.ingestion.ObjectsXml.getPathAndFilename(ObjectsXml.java:260) ... 23 more </trace></error></result>

All the error logs are being forwarded to Splunk, where an ad hoc report can be created to display all the records that failed during a given ingestion event.

## Rebuilding a Deleted Search Index

* Turn off the scheduled annotations indexing and web crawls.
* Stop ingestion. bin/shutdown.sh
* Move data/Ingestion/ to a backup location so ingestion will not skip records it thinks it's seen before.
* Configure config/settings.xml:
  + downloadDigitalObjectsEnabled to false
  + convertJPEG2000Enabled to false
  + createThumbnailsEnabled to false
  + createZoomImagesEnabled to false
  + bulkLoad to true
* Comment out db titles stage in config/ingestion.xml:

  <!-- <stage component="PopulateOpaTitlesTable" /> -->

* Make sure xmlstore/incremental/ has only the latest DAS export in it.
* Startup ingestion. bin/startup.sh
* Let ingestion run all the way through the full export.
* When feeding is complete, shut down ingestion. bin/shutdown.sh
* Remove data/Ingestion/.
* Put the previously backed up data/Ingestion/ in place.
* Uncomment the db titles stage in config/ingestion.xml:

  <stage component="PopulateOpaTitlesTable" />

* Configure config/settings.xml:
  + downloadDigitalObjectsEnabled to true
  + convertJPEG2000Enabled to true
  + createThumbnailsEnabled to true
  + createZoomImagesEnabled to true
  + bulkLoad to false
* Start ingestion.
* Turn on scheduled annotations updates and web crawls.

## Deleting Data

### Deleting Records

When records need to be removed from OPA, NARA will provide a delete file in a DAS XML Export. The file contains a list of NAIDs to be removed. The act of deletion will remove:

* xml from xmlstore
* opastorage directory for that record
* Indexed Entry for search

### Deleting Objects

In order for data to be removed from OPA, NARA will provide the applicable data (e.g. ADs/ARs with their digital object tag removed) to be deleted. OPA will automatically reprocess the data with the objects removed by comparing the new list of objects in record to the list of objects previously in the record.

Deleted digital object files are automatically moved by OPA to opastorage/deleted/.

The “deleted” section can be used to recover from errors in case an object was inadvertently removed from the DAS description by NARA staff.

In addition, a command is sent from Content Processing/Ingestion to the Search Engine for the specified index entries (directly related to the deleted files) to be removed.

## Resetting User Passwords

* OPA
  + Users are able to reset their own passwords
* AWS Management Console / IAM users
  + Performed by System Administrator(s)
* Splunk
  + Performed by Splunk Administrator(s)

# Appendices

## Application Server Config.xml

/opt/vol/config/config.xml

<?xml version="1.0" encoding="utf-8"?>

<!DOCTYPE project>

<configuration>

**<!-- UI SETTINGS -->**

<searchRunTime>10000</searchRunTime>

<!-- # OF MILLISECONDS BEFORE A SEARCH TIMES OUT (10 seconds) -->

***Note: Use numerical values only (e.g. integer value > 0). Be advised that making a large increase to this value may adversely impact system performance.***

**<!-- SESSION SETTINGS -->**

<sessionTimeout>1800</sessionTimeout>

<!-- # OF SECONDS BEFORE A SESSION TIMES OUT (30 minutes) -->

***Note: Use numerical values only (e.g. integer value > 0). Be advised that making a large increase to this value may adversely impact system performance.***

**<!-- FAILED LOGIN ATTEMPT SETTINGS -->**

<loginAttempts>5</loginAttempts>

<!-- MAXIMUM # OF FAILED LOGIN ATTEMPTS -->

<timeLocked>15</timeLocked>

<!-- # OF MINUTES AN ACCOUNT IS LOCKED FOR AFTER MAXIMUM FAILED LOGIN ATTEMPTS IS HIT -->

<attemptsTime>15</attemptsTime>

<!-- # OF MINUTES AN ACCOUNT IS LOCKED FOR BEFORE RESETTING THE FAILED LOGIN COUNT BACK TO 0 -->

***Note: Use numerical values only (e.g. integer value > 0).***

**<!-- SEARCH SETTINGS -->**

<maxSearchResultsPublic>500</maxSearchResultsPublic>

<!-- MAXIMUM # OF SEARCH RESULTS FOR A PUBLIC USER -->

<maxSearchResultsStandard>2000</maxSearchResultsStandard>

<!-- MAXIMUM # OF SEARCH RESULTS FOR A STANDARD USER -->

<maxSearchResultsPower>20000</maxSearchResultsPower>

<!-- MAXIMUM # OF SEARCH RESULTS FOR A POWER USER -->

***Note: Use numerical values only (e.g. integer value > 0). Be advised that making a large increase to this value may adversely impact system performance.***

**<!-- PUBLIC API SEARCH SETTINGS -->**

<maxApiSearchResults>200</maxApiSearchResults>

<!-- MAXIMUM # OF SEARCH RESULTS FOR THE PUBLIC API -->

***Note: Use numerical values only (e.g. integer value > 0). Be advised that making a large increase to this value may adversely impact system performance.***

**<!-- USER LIST SETTINGS -->**

<maxResultsPerListPublic>500</maxResultsPerListPublic>

<!-- MAXIMUM # OF ITEMS THAT CAN BE ADDED TO A LIST FOR A PUBLIC USER -->

<maxResultsPerListStandard>2000</maxResultsPerListStandard>

<!-- MAXIMUM # OF ITEMS THAT CAN BE ADDED TO A LIST FOR A STANDARD USER -->

<maxResultsPerListPower>20000</maxResultsPerListPower>

<!-- MAXIMUM # OF ITEMS THAT CAN BE ADDED TO A LIST FOR A POWER USER -->

<maxListsPerUser>100</maxListsPerUser>

<!-- MAXIMUM # LISTS PERMITTED PER USER -->

***Note: Use numerical values only (e.g. integer value > 0). Be advised that making a large increase to this value may adversely impact system performance.***

**<!-- PRINT SETTINGS -->**

<maxPrintResults>200</maxPrintResults>

<!-- MAXIMUM # OF RESULTS A USER IS ALLOWED TO PRINT -->

***Note: Use numerical values only (e.g. integer value > 0). Be advised that making a large increase to this value may adversely impact system performance.***

**<!-- BULK EXPORT SETTINGS -->**

<maxNonBulkTimer>30000</maxNonBulkTimer>

<!-- MAXIMUM # OF MILLISECONDS A NON-BULK EXPORT CAN TAKE TO COMPLETE -->

<maxNonBulkFileSizeLimit>100000</maxNonBulkFileSizeLimit>

<!-- MAXIMUM # OF MEGABYTES A NON-BULK EXPORT FILE SIZE CAN BE -->

<maxBulkExportFileSize>10000000</maxBulkExportFileSize>

<!-- MAXIMUM # OF MEGABYTES A BULK EXPORT FILE SIZE CAN BE -->

<bulkExpDays>60</bulkExpDays> <!-- TOTAL # OF DAYS BEFORE A BULK EXPORT REQUEST EXPIRES -->

***Note: Use numerical values only (e.g. integer value > 0).***

**<!-- ROW RETRIEVAL SETTINGS -->**

<maxNotificationRows>200</maxNotificationRows>

<!-- MAXIMUM # OF ROWS RETRIEVED IN A NOTIFICATIONS QUERY -->

<maxOpaTitlesRows>200</maxOpaTitlesRows>

<!-- MAXIMUM # OF ROWS RETRIEVED IN AN OPA TITLES QUERY -->

<maxSummaryRows>200</maxSummaryRows>

<!-- MAXIMUM # OF ROWS RETRIEVED IN A CONTRIBUTION SUMMARY QUERY -->

<maxContributionRows>200</maxContributionRows>

<!-- MAXIMUM # OF ROWS RETRIEVED IN A CONTRIBUTIONS QUERY -->

***Note: Use numerical values only (e.g. integer value > 0).***

**<!-- ANNOTATION SETTINGS -->**

<tagsLength>100</tagsLength>

<!-- MAXIMUM # OF CHARACTERS ALLOWED FOR A TAG -->

**<commentsLength>250</commentsLength> [Not in R1]**

**<!-- MAXIMUM # OF CHARACTERS ALLOWED FOR A COMMENT -->**

**<commentsFormat>collapsed</commentsFormat> [Not in R1]**

**<!-- DISPLAY FORMAT FOR COMMENTS -->**

***Note: Use numerical values only (e.g. integer value > 0). Be advised that making a large increase to this value may adversely impact system performance.***

**<!-- TRANSCRIPTION INACTIVITY TIME SETTINGS -->**

<transcriptionInactivityTime>30</transcriptionInactivityTime>

<!-- # OF MINUTES BEFORE A TRANSCRIPTION LOCK IS RELEASED -->

***Note: Use numerical values only (e.g. integer value > 0).***

**<!-- CONTRIBUTION DISPLAY TIME SETTINGS -->**

<transcriptionsDisplayTime>180</transcriptionsDisplayTime>

<!-- # OF DAYS A TRANSCRIPTION WILL APPEAR IN THE MODERATOR STREAM -->

<tagsDisplayTime>180</tagsDisplayTime>

<!-- # OF DAYS A TAG WILL APPEAR IN THE MODERATOR STREAM -->

***Note: Use numerical values only (e.g. integer value > 0).***

**<!-- GENERAL SETTINGS -->**

<naraEmail>@nara.gov</naraEmail>

**<!-- BANNER LINK SETTINGS -->**

<bannerLinkHomeDisplay>Home</bannerLinkHomeDisplay>

<bannerLinkHome>http://www.archives.gov/</bannerLinkHome>

<bannerLinkResearchRoomDisplay>Research Room</bannerLinkResearchRoomDisplay>

<bannerLinkResearchRoom>http://www.archives.gov/research/</bannerLinkResearchRoom>

<bannerLinkContactUsDisplay>Contact Us</bannerLinkContactUsDisplay>

<bannerLinkContactUs>http://www.archives.gov/contact/</bannerLinkContactUs>

<bannerLinkStatisticsDisplay>Statistics</bannerLinkStatisticsDisplay>

<bannerLinkStatistics>http://research.archives.gov/#/statistics/</bannerLinkStatistics>

<bannerLinkHelpDisplay>Help</bannerLinkHelpDisplay>

<bannerLinkHelp><http://www.archives.gov/research/search/help/using-opa.html>

</bannerLinkHelp>

***Note: URLs must be valid.***

</configuration>

## User Interface Configuration Properties File

/www/html/scripts/<random number>config.js

var opaConfig = angular.module('opaAppConfig', []);

var API\_LOCATION = 'https://prod.narasearch.us’;

opaConfig.value('configServices', {

'API\_LOCATION': API\_LOCATION,

'API\_END\_POINT': API\_LOCATION + '/OpaAPI/api/v1',

'IAPI\_END\_POINT': API\_LOCATION + '/OpaAPI/iapi/v1',

'OPASTORAGE\_URL': API\_LOCATION + '/OpaAPI/media/',

'TIMEOUT\_TRANSCRIPTION': 1800000,

'RESULTS\_PER\_PAGE': [20,50,75,100],

'PRINT\_LIMIT': 500,

'CONTRIB\_RETRIEVAL\_LIMIT': 200,

'REVISION': 'revNumberPH'

});

**Configuration Descriptions**

*var API\_LOCATION = 'https://prod.narasearch.us';*

*(App server that runs the API. It needs to be a public URL.)*

*'API\_END\_POINT': API\_LOCATION + '/OpaAPI/api/v1',*

*'IAPI\_END\_POINT': API\_LOCATION + '/OpaAPI/iapi/v1',*

*(Those are API endpoints, for public and internal API.)*

*'OPASTORAGE\_URL': API\_LOCATION + '/OpaAPI/media/',*

*(This is the URL of the OPA storage used to retrieve thumbnails, content details, etc.)*

*'TIMEOUT\_TRANSCRIPTION': 1800000,*

*(The time that the browser will take to automatically lock a transcription being edited in msec.)*

*'RESULTS\_PER\_PAGE': [20,50,75,100],*

*(That fills a dropdown with the number of brief results that can be retrieved in each search. The user can selection one of those options.)*

*'PRINT\_LIMIT': 500,*

*(This is the maximun number of records a not logged in user can print.)*

'CONTRIB\_RETRIEVAL\_LIMIT': 200,

*(This the maximum number of contributions to be requested for the summary page of "my account" for the my account contributions section.)*

*'REVISION': 'revNumberPH'*

*(This is the number that shows up in the header and footer.)*

## Aspire



Figure 13.1: Aspire System Administration User Interface

Please SSH into the Ingestion server and follow the path below to access the Aspire Administration User Interface.

pcp01@aws.narasearch.us:50505/aspire/files/home.html

## Solr Cloud

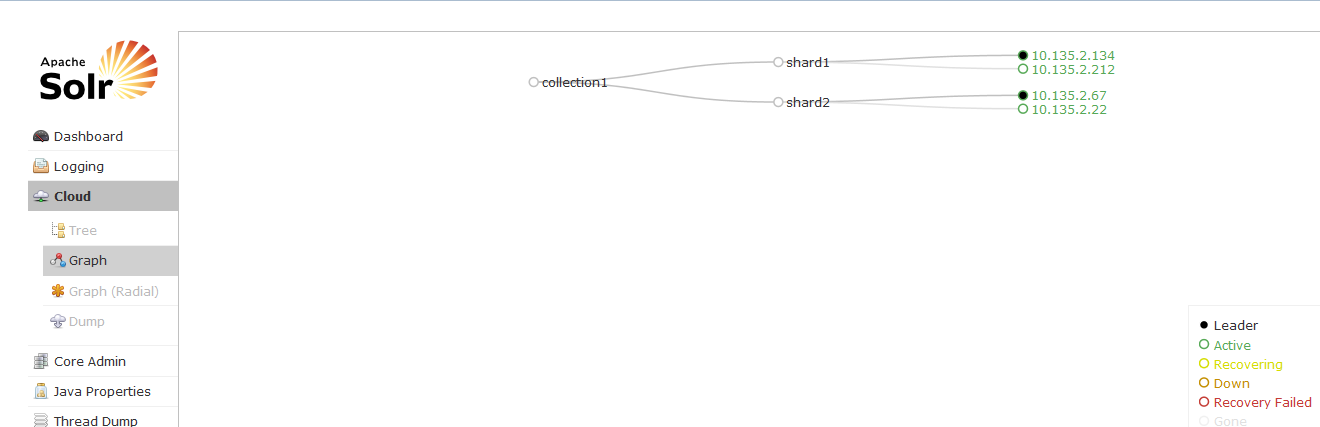


Figure 13.2: Solr Administration User Interface

Please reference the following URL for more information on the Solr Administration User Interface.

<https://cwiki.apache.org/confluence/display/solr/Using+the+Solr+Administration+User+Interface>

## Acronyms

AD: Archival Description

API: Application Program Interface

AR: Authority Record

AWS: Amazon Web Services

CLI: Command Line Interface

DAS: Description and Authority Services

DNS: Domain Name System

EBS: Elastic Block Storage

EC2: Elastic Compute Cloud

IAM: Identity and Access Management

IE: Index Entry

NARA: National Archives and Records Administration

OPA: Online Public Access

OPA-IP: Online Public Access Information Package

QPS: Queries Per Second

SSH: Secure Shell

UAT: User Acceptance Testing

UI: User Interface

## Reports

The logs used to create reports in Splunk are backed and saved as full system images. They can be retained as long as needed, or until the next full system image is created.

### System Access Reports

The System Administrator will receive an email notification when the following system access reports have been generated by the system:

#### Number of Queries

* Daily
* Weekly for Last Quarter

#### Top 1,000 Search Terms

* Monthly
* Yearly
* Since inception – Monthly
* Zero results – Monthly

#### Top 1,000 Most Accessed Files

* + - Weekly
      * Descriptions with Digital Objects in Online Holdings
      * Descriptions without Digital Objects
      * Authority Records
      * Webpages from Archives.gov and Presidential Libraries
    - Since inception
      * Descriptions with Digital Objects in Online Holdings
      * Descriptions without Digital Objects
      * Authority Records
      * Webpages from Archives.gov and Presidential Libraries

#### Simple Search Access Reports

* Number of times filter options were accessed

#### Advanced Search Access Reports

* Number of times filter options were accessed

### Contributions Reports

The System Administrator will receive an email notification when the following system access reports have been generated by the system:

#### Number of User Contributions

* Type of Contribution
  + - Translations
    - Transcriptions
    - Comments
    - Tags
    - Cumulative

#### Number of Registered Users

* Weekly
  + - Type of Contribution
      * Translations
      * Transcriptions
      * Comments
      * Tags
    - Cumulative
    - Monthly
    - Type of Contribution
      * Translations
      * Transcriptions
      * Comments
      * Tags
    - Cumulative
    - Yearly
    - Type of Contribution
      * Translations
      * Transcriptions
      * Comments
      * Tags
    - Cumulative
    - Since inception
    - Type of Contribution
      * Translations
      * Transcriptions
      * Comments
      * Tags
    - Cumulative

#### Top 100 Records with Most Comments

* Descriptions
* Authorities
* Digital Objects

#### Top 100 Records with Most Translations

* Digital Objects

#### Top 100 Records with Translations with Most Edits

* Digital Objects
* Cumulative Edits to Translations

#### Top 100 Records with Transcriptions with Most Edits

* Digital Objects
* Average Edits to Transcriptions
* Cumulative Edits per Transcription

#### Reasons for Removing Contributions

* Contribution Type (Translations, Transcriptions, Comments, Tags)
  + - Reason
    - Cumulative for Each Type

#### Reasons for Restoring Contributions

#### Average Number of Contributions

* Contribution
  + - Top 1,000 Contributors

#### Top 100 Contributors

* + - Username
    - Excluding NARA users
    - Total Number of Contributions

#### Top 1,000 Tags Contributed by Users

* Tags
* Total Number of OPA Records Tagged

### API Reports

#### Number of Queries Generated through the API

#### Characteristics of Queries made through API

#### Number of Times Exports are Generated through API

#### Total Number of Bytes Served via API

#### Number of User Contributions via the API

#### Number of Unique Contributors via the API

### Export Reports

#### Number of Times Records are Exported

#### Average Number of Exports per User

### Saved Lists – Reports

#### Number of Saved Lists by Registered Users

* Weekly
* Monthly
* Yearly
* Since inception

### Server Space Reports

#### OPA Server Space Report

### Shares – Reports

#### Number of Shares Made by Users

* Last 1 week
* Last 1 month
* Last 1 year
* Since inception

#### Number of Shares per User

#### Records with Most Number of Shares

#### Top 100 Shares by Destination

#### Number of Times Copy URL is Accessed

### Digital Analytics – Report

#### Total Number of Times Tabbed Groupings are Accessed

* Last 1 week
* Last 1 month
* Last 1 year

#### Total Number of Times Advanced Search Screen is Accessed

#### Total Number of Times the Policy and Help Link are Accessed

* Weekly
* Monthly
* Yearly

### Statistics – Report

#### View Amount of Data per Level of Description (e.g. Collection, Record Group, Series, File Unit and Item)