System Design: Data Distribution

PDS System Design Review II
Greenbelt, Maryland
June 21-22, 2011

Sean Hardman

Topics

- Overview
- Data Distribution Related Components and Flow
- Search Scenarios
- Discovery and Retrieval
- Deployment and Plans
- Wrap Up

Overview

- The topic of Data Distribution was briefly discussed in the first review.
- The approach has not changed since then, but this presentation will offer more detail.
- Distribution involves discovery and retrieval of products.
  - The focus through builds 1 and 2 is on discovery.

Design Status

- Documents completed and reviewed:
  - Report Requirements and Design
- Documents in process:
  - Search and Operator Portal Requirements and Design
  - Search Scenarios
- Documents to be started:
  - Data Consumer Portal, Subscription and Monitor Requirements and Design
- Latest versions posted to Engineering Node site
  - http://pds-engineering.jpl.nasa.gov/index.cfm?pid=145&cid=134

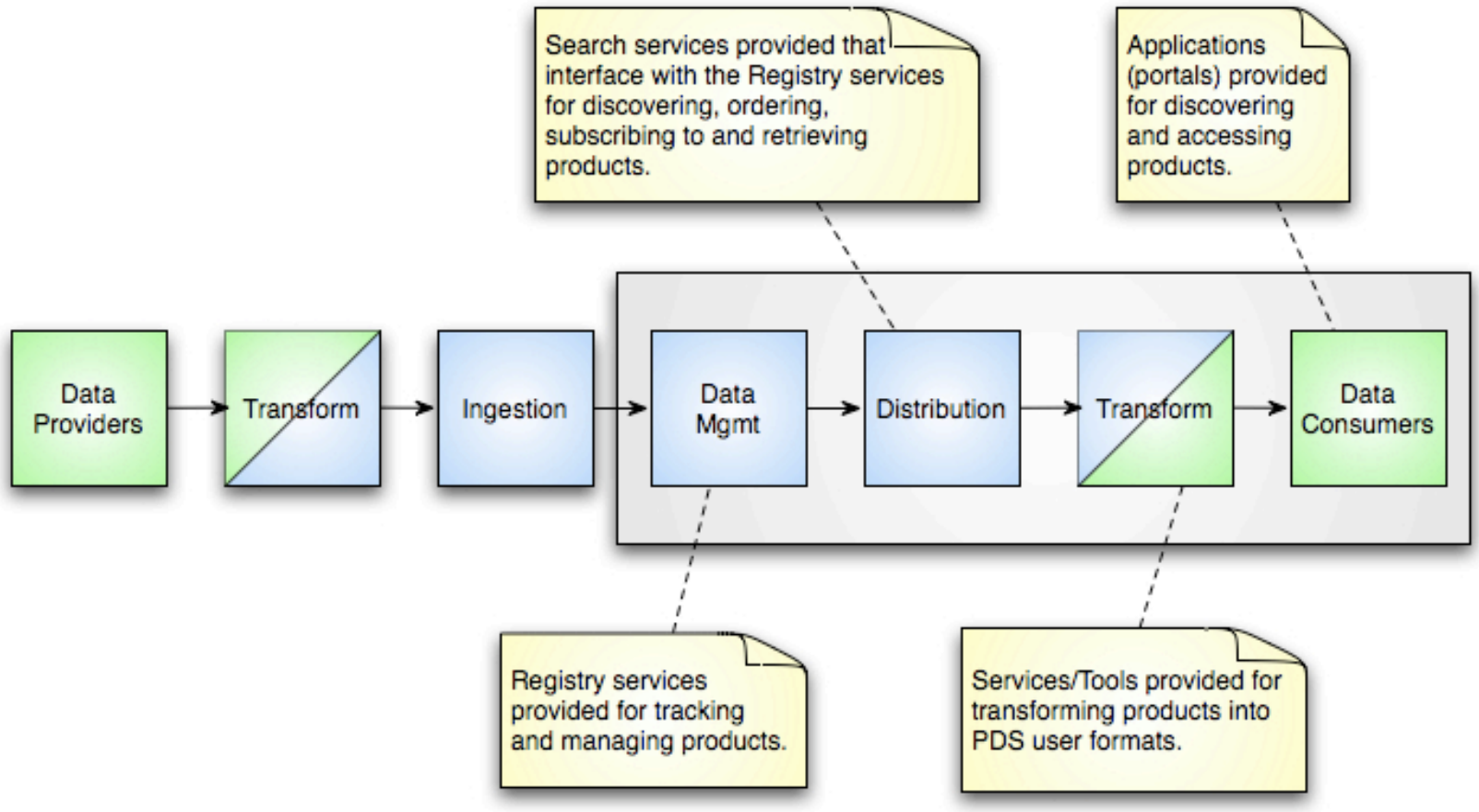
Key Level 3 Requirements

- **2.2.2** PDS will track the status of data deliveries from data providers through the PDS to the deep archive
- **2.8.3** PDS will provide standard protocols for locating, moving, and utilizing data, metadata and computing resources across the distributed archive, among PDS nodes, to and from missions, and to and from the deep archive
- **3.1.1** PDS will provide online interfaces allowing users to search the archive
- **3.1.2** PDS will provide online interfaces for discipline-specific searching
- **3.1.3** PDS will allow products identified within a search to be selected for Retrieval

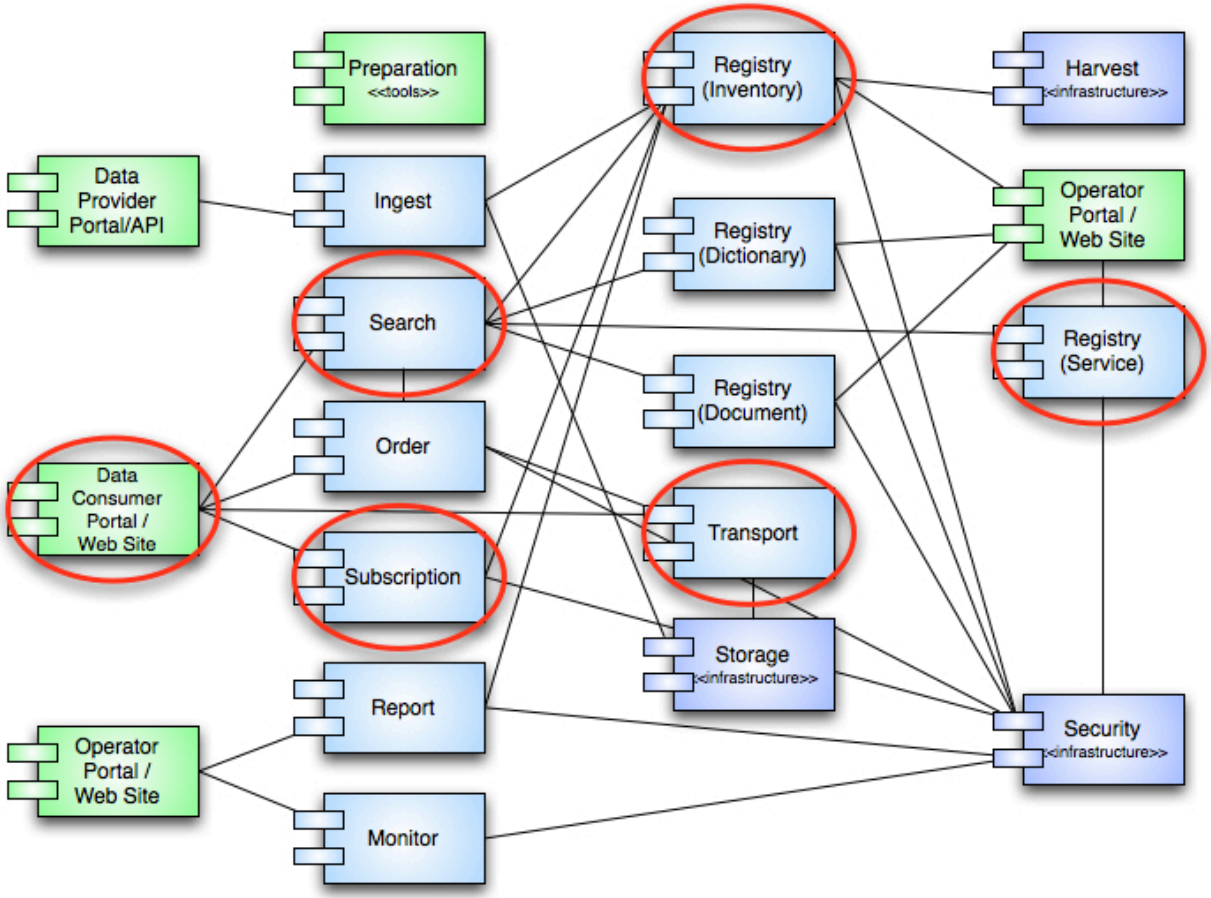
Topics

- Overview
- Data Distribution Related Components and Flow
- Search Scenarios
- Discovery and Retrieval
- Deployment and Plans
- Wrap Up

Data Distribution
(Discovery and Retrieval of Products)



Data Distribution Related Components



Data Distribution Related Components

Data Consumer Portal, Subscription and Transport Services

- Data Consumer Portal
  - Integrate the PDS-wide portal (http://pds.nasa.gov/) with the Search Service.
  - Includes applications for Dictionary and Phonebook viewing and the catalog-level search interface.
- Subscription Service
  - Provides subscription to data, document and software release announcements.
  - Replace the current implementation with one built on the Registry Service.
- Transport Service
  - Integrate existing delivery mechanisms (e.g., FTP, HTTP, etc.).

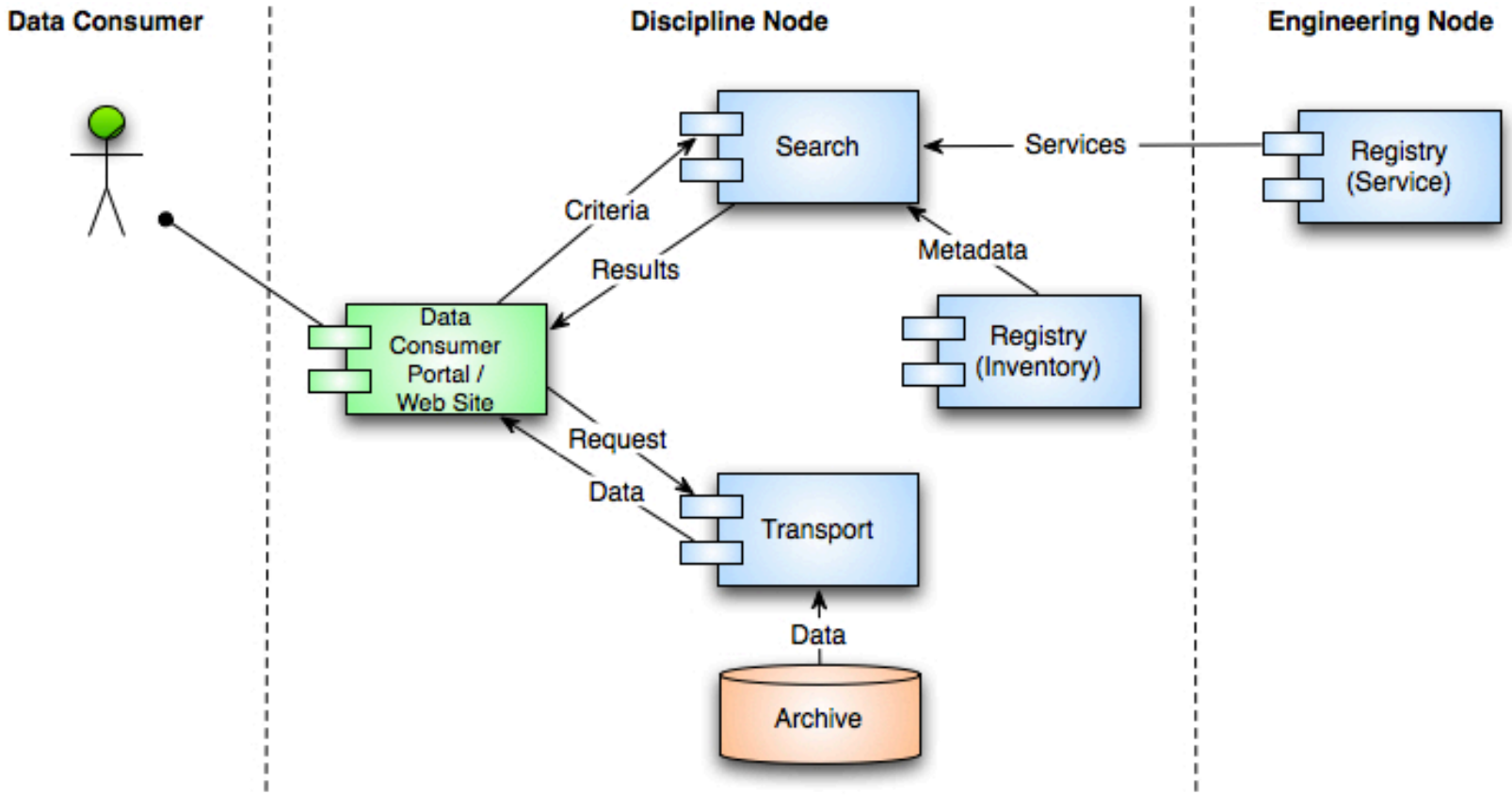
Data Distribution Related Components
Search Service

- This service is a deployable component that accepts queries for data and returns a set of matching results.
- Provides the public interface (REST-based API over HTTP) to the metadata contained in the federated registries.
- Provides the second line of metadata harvesting within the system in order to facilitate discovery of products.
- Generation of search indices from registry metadata supports multiple query formats and is tailor-able for customized search interfaces.

Data Distribution Related Components
Registry (Aggregate) Service

- The EN will host an aggregate Registry Service that periodically pulls registry entries from the distributed set of Node-hosted Registry Service instances.
- Allows the Nodes to maintain governance of their own registry contents.
- The aggregate registry instance facilitates PDS-wide tracking, reporting and catalog-level search.

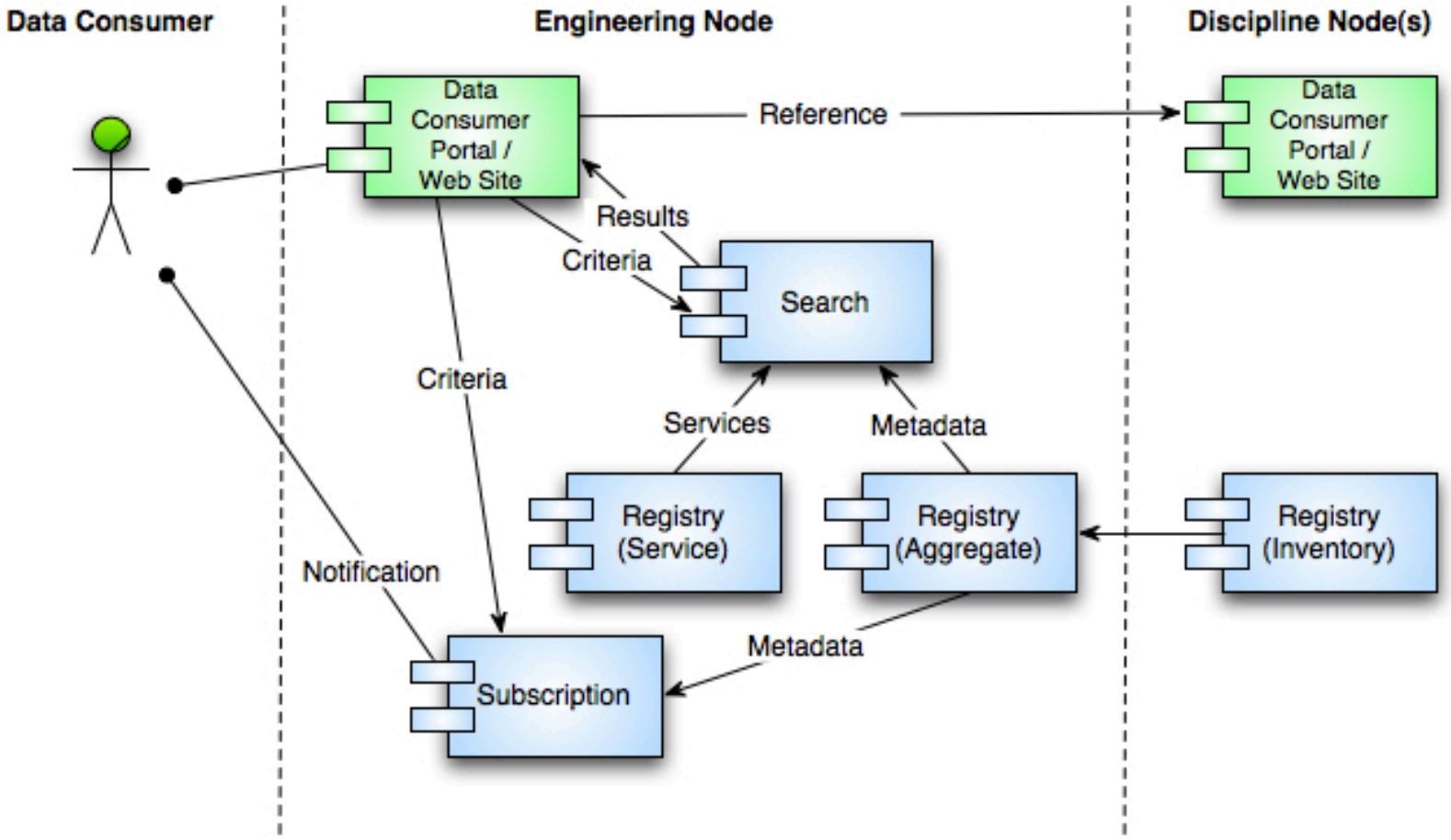
Data Distribution Flow
Request Initiated at the Discipline Node



Data Distribution Flow Details
Request Initiated at the Discipline Node

1. Search service generates a search index utilizing the Service registry to discover the appropriate Registry service(s) for obtaining the metadata for the index. Tailoring of the search index enables support for the Node-specific search tools.
2. Data Consumer submits a query for data through a portal / web site interface.
3. Portal / web site interface forwards the query to the local Search service.
4. Search service returns results to the portal / web site interface with options for retrieving product(s) that match the query criteria.
5. Data Consumer makes a request to the Transport service for delivery of the product(s).

Data Distribution Flow
Request Initiated at the Engineering Node



Data Distribution Flow Details
Request Initiated at the Engineering Node

1. Search service generates a search index utilizing the Service registry to discover the appropriate Registry service(s) for obtaining the metadata for the index.
2. Data Consumer submits a query for data through a portal / web site interface. The Data Consumer may also subscribe to release information via the Subscription service.
3. Portal / web site interface forwards the query to the Search service.
4. Search service returns results to the portal / web site interface with options for retrieving product(s) that match the query criteria.
5. Data Consumer makes a request to the Transport service from the appropriate Node for delivery of the product(s).

Topics

- Overview
- Data Distribution Related Components and Flow
- Search Scenarios
- Discovery and Retrieval
- Deployment and Plans
- Wrap Up

Search Scenarios

- PDS search scenarios have been gathered from a number of sources.
- The Search Service design focuses on the search infrastructure.
- Scenarios feed into query model development as well as metadata capture.
- Many of the scenarios have a common theme where they are looking for data:
  - From a particular instrument
  - Of a specific target or specific location on a target
  - For a designated time range

Scenario Example

A science user wants to select a time sequence of both **wide and narrow angle frames** obtained when an associated storm was visible on the daylight side of **Saturn**. Storms drift around the planet at various rates. What they want to do is to execute a sequence of searches. First, set the gate to include **images** within **+/- one month** and retrieve a time-ordered list of **start time**, max and min **latitude** (0-20 deg) min and max **longitude** (25-45 deg), **camera name**, **filter** and **exposure time**.

Hopefully, after they process selected frames from this data they can derive a longitudinal drift rate for the storm system and return with a series of **bracketed searches** to isolate the storm and study its evolution over a maximum length of time.

Catalog-Level Search

- Queries the catalog or context products to discover data sets and discipline-specific search tools.
- Results provide links to Node resources.
- Heavily dependent on PDS3 catalog migration.
  - Cross-references are key to defining the relationships, for example:
    - Instrument -> Instrument Host -> Investigation
    - This feeds the facet-based approach of the search interface.

Product-Level Search

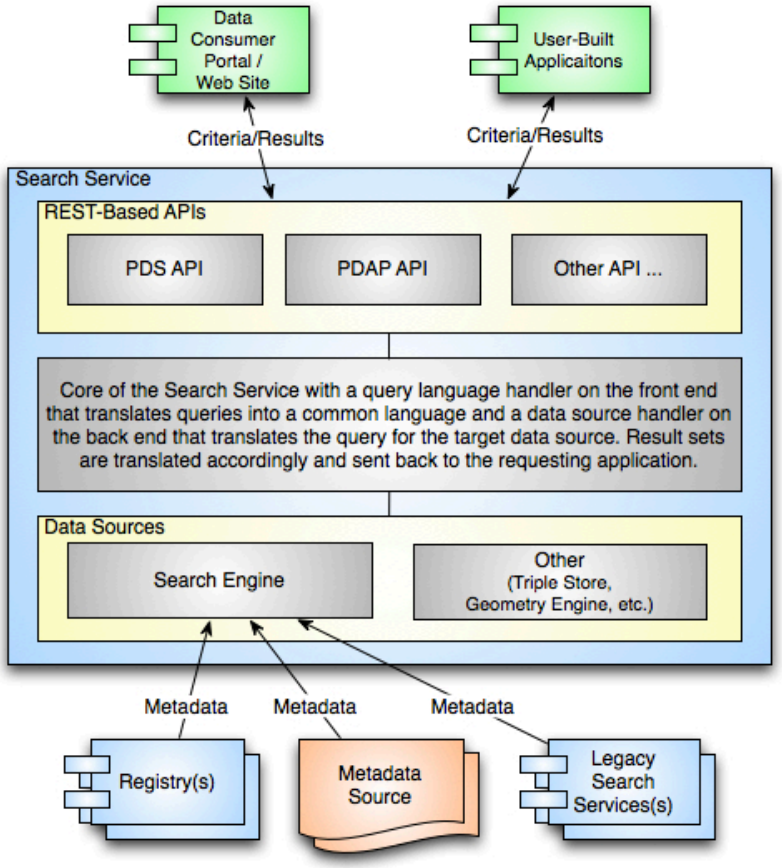
- Queries product-level metadata to discover products.
- Requires coordination among Discipline-Specific model development.
  - Reuse of classes and attributes
  - Promotion to a broader namespace where possible
  - This effort for PDS4 is in progress.
- Facilitates cross-Node search where appropriate.
  - Generic product-level queries across all of the Nodes is not a goal.
  - Targeted queries involving specific Nodes is a goal and is common in many of the search scenarios.

Topics

- Overview
- Data Distribution Related Components and Flow
- Search Scenarios
- Discovery and Retrieval
- Deployment and Plans
- Wrap Up

Search Architecture

- Search indexes built from multiple sources.
- Allows for annotation of archive metadata.
- Customizable for a discipline-specific search interface.



Search Engine

- Decided to use Apache's Solr for the search engine portion of the Search Service.
- Developers at the EN have experience with the software package.
  - It is currently utilized on the backend of the current catalog-level search application.
- Recent releases contain some useful features:
  - Improved geospatial support
  - Range faceting on numeric fields
- Although Solr offers a robust query language, we are designing a layer above it to allow for other search engine options in the future.

Index Generation

- The main source of metadata for the search index is the contents of the Registry Service.
  - At EN, that would be the aggregated registry.
  - At DNs, their local inventory registry.
- The index will make use of the associations in the registry to help users find related products, documents, etc.
- It will also make use of the classification schemes in the registry to help drive the faceted search capabilities.
- Indexes can be tailored for Node-specific search applications.

Metadata Annotation

- The data model and the Search Service architecture facilitate metadata annotation.
  - Allows search to be based on the latest and most accurate metadata.
- A defined product with an associated table allows for updated or additional metadata to be specified for a set of products.
  - This product will be associated in the registry with a data set or subset of products.
- The Search Service will support a similar structure for supplying metadata separate from registered content.

Search REST-Based API
Three Aspects

- Discovery
  - Focuses on discovering content, whether at the catalog or product level.
  - Facilitated by support for search parameters and paged result sets.
- Link
  - Focuses on passing search parameters from one service to another.
  - Deployment of the Search Service facilitates parameter passing and integration.
- Access
  - Focuses on retrieval of product files.

Search REST-Based API
Design Considerations

- Currently designing the PDS query language implemented as a REST-based API over HTTP.
- For discovery, supports return of paged results in a defined structure (e.g., XML or JSON).
- For retrieval, supports return of products and product packages.
- The architecture allows support for other query languages (e.g., IPDA's Planetary Data Access Protocol (PDAP)).
- The current plan is to consolidate PDS and PDAP into a single query language and propose the result back to the IPDA.

Query Language

- Specifies HTTP parameter passing to simplify interfaces and minimize encoding:
  - …?target=mars&investigation=mro&resource_class=dataset
- Provides a rich set of relational operators:
  - =, !=, <, <=, >, >=, LIKE, NOTLIKE, IS, ISNOT
- Supports logical operators (AND, OR, NOT) and ranges.
- Supports functions:
  - transform(JPG), package(ZIPPED_BUNDLE), etc.

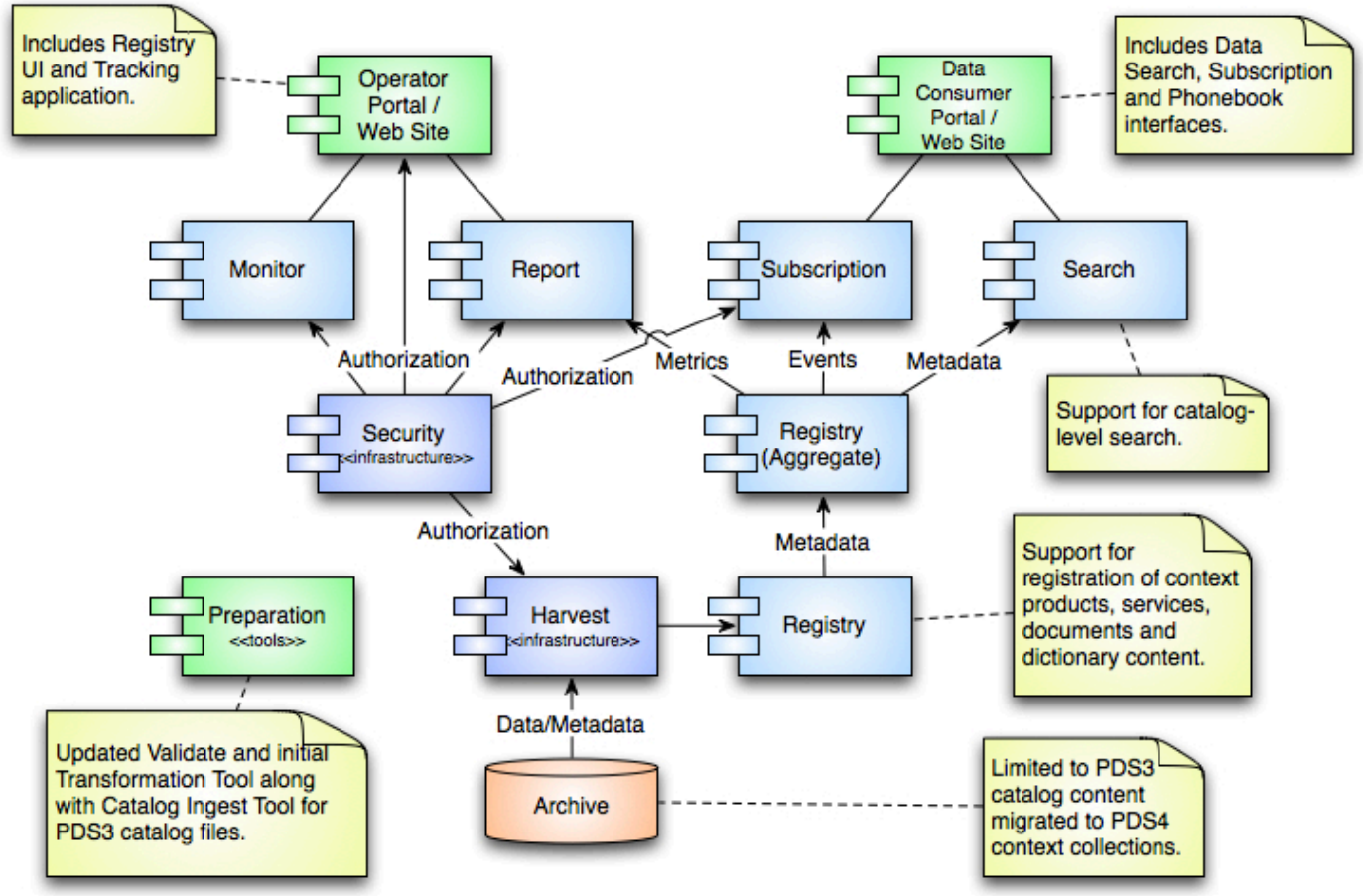
Transport Service

- Build 2 relies on existing Node distribution services (e.g., FTP, HTTP, etc.).
- Plan to design and develop a service that incorporates the following:
  - Transfer mechanisms based on current data movement evaluations.
  - Transformation functions initially developed for the tool by the same name.
  - Product packaging (i.e., zipped bundles, etc.).

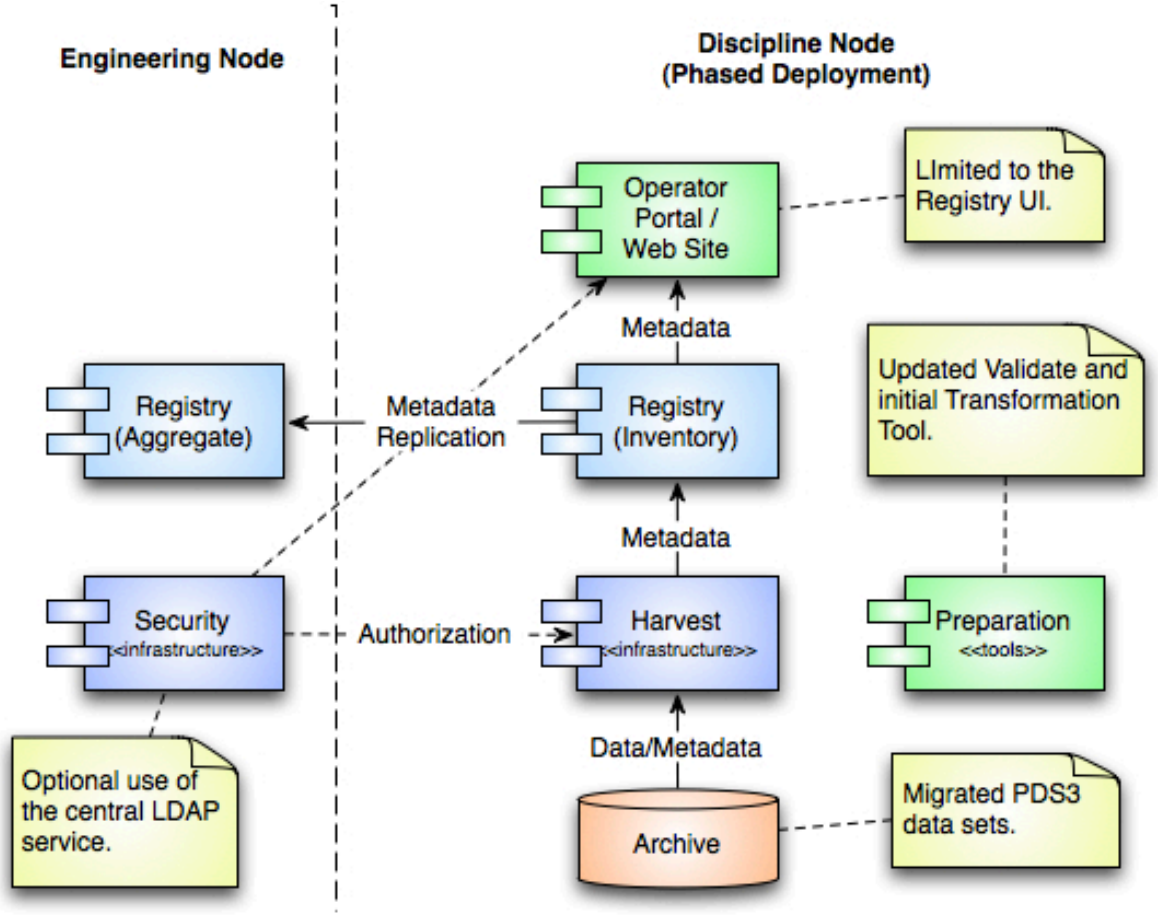
Topics

- Overview
- Data Distribution Related Components and Flow
- Search Scenarios
- Discovery and Retrieval
- Deployment and Plans
- Wrap Up

Build 2 Deployment
Engineering Node



Build 2 Deployment
Discipline Node



Build 2 Deployment
Additional Details

- Deployment at the Nodes will be phased throughout the year.
- Once deployment is complete, Node should begin registration of PDS3 products with the Harvest Tool.
  - This is not PDS3 product migration.
  - A generic proxy label is generated and registered for tracking purposes.
- In order to populate the Report Service, Nodes should make web and FTP logs available.

Preparation for Build 2

- Complete Search API design and development.
- Initiate design and development of the Subscription Service.
- Develop user interfaces for phonebook and dictionary viewing.
- Evaluate off-the-shelf products for satisfying the Monitor Service.

Plans for Build 3

- Tools for transformation and visualization.
  - A framework for data product transformation allowing contribution of transformations from others.
  - Replacement functionality for NASAView available as a desktop tool and a library to be integrated with other components.
- Design and develop the Transport Service.
- Incorporate findings of ongoing research into data movement and storage solutions.
- Focus on integration of new components with existing Node software and infrastructure.

Topics

- Overview
- Data Distribution Related Components and Flow
- Search Scenarios
- Discovery and Retrieval
- Deployment and Plans
- Wrap Up

Wrap Up

- Design and development of the Search Service has focused on an architecture that supports the varied PDS search scenarios.
- The search architecture currently employed for PDS, where catalog-level search directs users to available data and discipline-specific search interfaces, has been continued.

Questions/Comments