a place of mind

THE UNIVERSITY OF BRITISH COLUMBIA

Data Management, Workflow Processing, and Web Services for CyberSKA

University of British Columbia
Okanagan Campus

Venkat Mahadevan
Dr. Erik Rosolowsky
The CyberSKA Project Team

The Growing Demands on Connectivity and Information Processing in Radio Astronomy from VLBI to the SKA

Aveiro, Portugal, 24th - 25th May 2011

a place of mind

THE UNIVERSITY OF BRITISH COLUMBIA

Overview

- Very high data rates and volumes necessitate a need for cyber-infrastructure solutions for:
  - Data management, storage, and distribution.
  - Distributed data processing.
  - Tools for interacting with data storage and processing services.

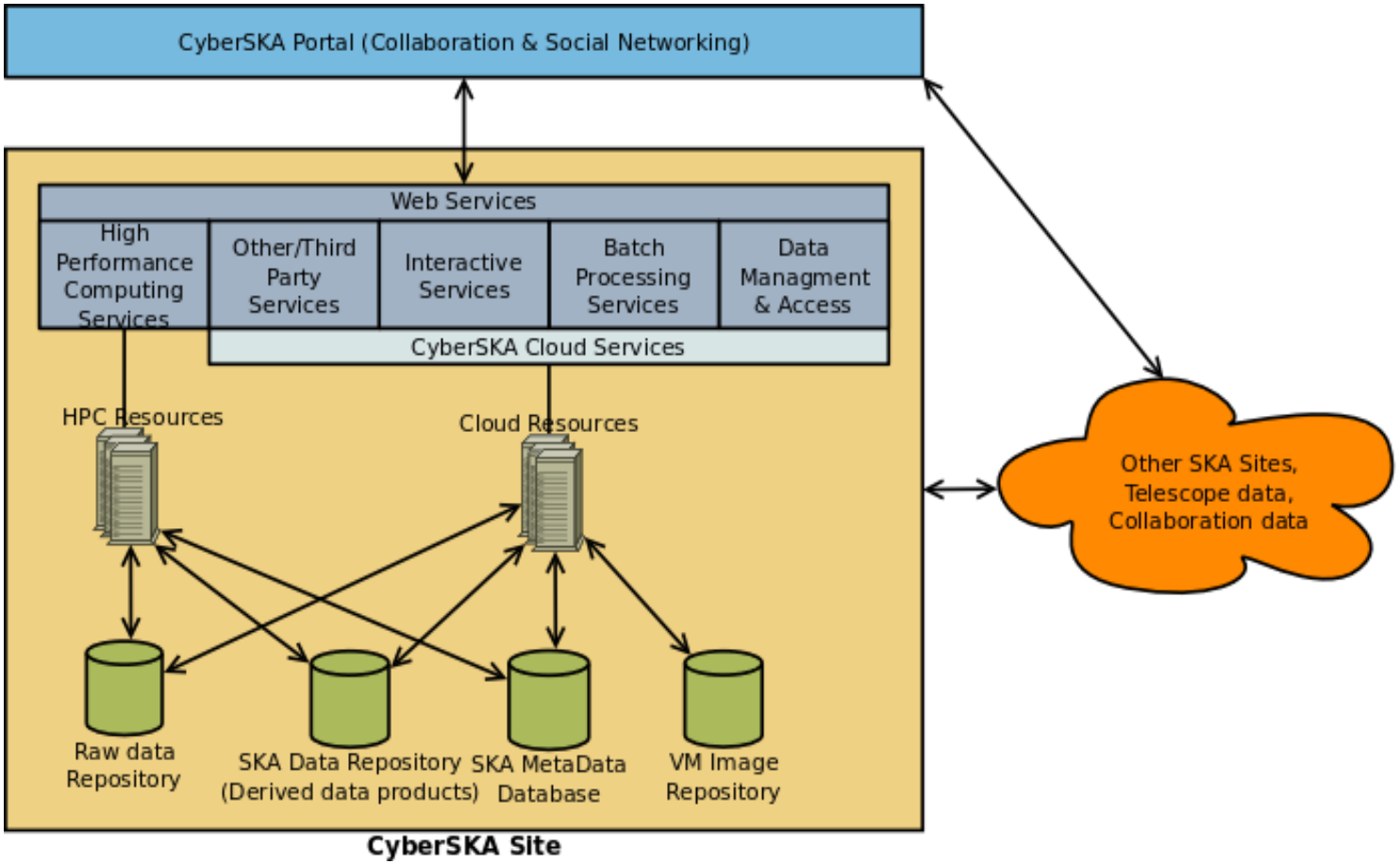
Overview (cont.)

- The goals for the data management and data processing services are:
  - Scalability and expandability/deploy-ability across multiple sites.
  - Accessibility via HTTP as a web service.
  - Security.
  - Transparency: provide access to data and processing services with a high level of abstraction.

Overview (cont.)

- Customizable: allow user defined scripts and plug-ins to run within the processing framework.
- Dynamic: allocate processing resources on demand.
- Interoperability: provide connectivity to data using VO standards.

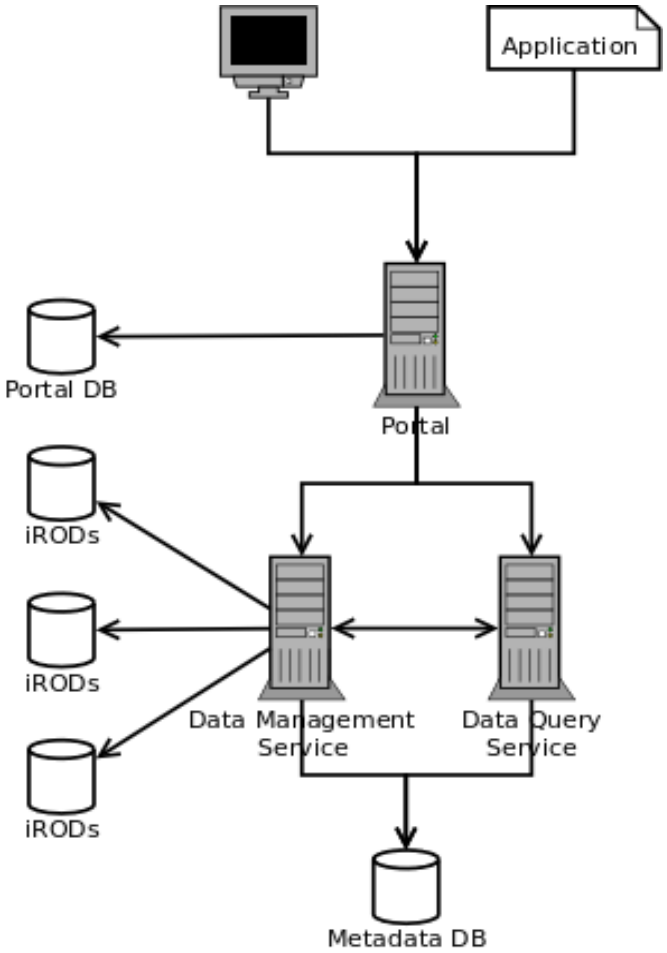
Overview (cont.)



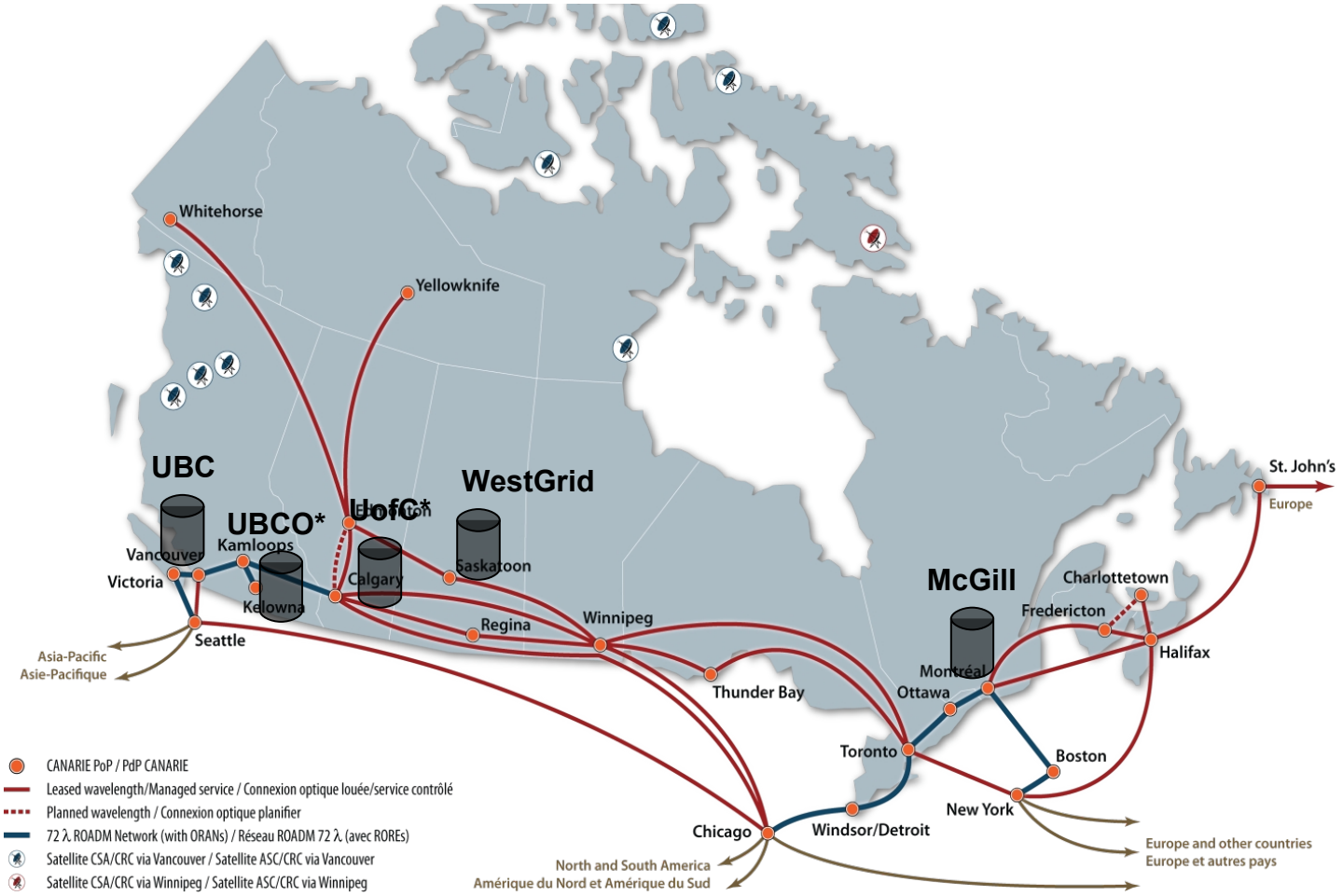
Data Management Service

- Built on the Integrated Rule-Oriented Data System (iRODS) and the iRODS Java API Jargon.
- Access to the Data Management Service (DMS) functions is provided via a HTTP accessible API using the Groovy/Grails framework.
- Basic file system type functions such as create file, copy file, move file, delete file, create collection, delete collection, etc. are provided via the HTTP accessible API.

Data Management Service (cont.)



Data Management Service (cont.)



Data Management Service (cont.)

- The location of data is abstracted from the user.
- Data can be replicated and backed up to multiple sites.
- Site to site data transfers using multiple TCP streams can be used to transfer large files quickly.

Data Management Service (cont.)

- Customized rules can be written to do specific operations on files when they are added to the system.
- Security is implemented using a combination of CyberSKA portal and OAuth (Open Authentication) mechanisms.
  - OAuth: open standard for authorization by service providers using key based signatures and access tokens instead of user names/passwords.

Data Management Service (cont.)

- Access permissions for files are stored in the portal's database on a user/group basis.
- The portal authenticates with the Data Management Service on behalf of the user to access files. The Data Management Service itself exposes a HTTP accessible API.

Data Management Service (cont.)

- The HTTP API is secured using a 2-legged OAuth scheme:
  - 2-legged OAuth involves signing each HTTP request with a pair of keys and other data in the request header. The web service verifies the signature before authorizing a method call.
  - Used by popular web services like Twitter, Google Apps Premier, etc.

Data Management Service (cont.)

- Various metadata is stored about different files in the CyberSKA portal.
- Special file types such as FITS images have another metadatabase stored separately that is populated when the FITS file is checked into iRODS.
- The files can then be queried using various parameters by the Data Query Service.

Data Query Service

- Data from FITS files is "ingested" into the metadatabase and can be queried using a combination of:
  - Spatial coordinate parameters.
  - Spectral frequency and stokes parameters.
  - Temporal date parameters.

Data Management Service Query Form

Usage: This query form supports between 1 and 4 query types linked together or separately (1 spatial, 1 spectral, 1 stokes, 1 temporal).

Spatial data in the metadatabase is based on Galactic longitude and Galactic latitude, which can be specified in a variety of formats as described in the comments to the side of each input box. After entering the appropriate data in the the query fields for the query types you want to process, select these query types in the "Set Query Types To Process" field set and hit the "Submit Query" button. Please note that queries which return many results may take several minutes to process.

Bounding Box Query

bbox swx	150d	bounding box south west x-coordinate i.e. Galactic longitude (in radians (e.g. 0.5), degrees (e.g. 70d), or degrees, minutes, seconds (e.g. 10d 12m 11.3s))
bbox swy	-2d	bounding box south west y-coordinate i.e. Galactic latitude (in radians (e.g. 0.5), degrees (e.g. 70d), or degrees, minutes, seconds (e.g. 10d 12m 11.3s))
bbox nex	155d	bounding box north east x-coordinate i.e. Galactic longitude (in radians (e.g. 0.5), degrees (e.g. 70d), or degrees, minutes, seconds (e.g. 10d 12m 11.3s))
bbox ney	2d	bounding box north east y-coordinate i.e. Galactic latitude (in radians (e.g. 0.5), degrees (e.g. 70d), or degrees, minutes, seconds (e.g. 10d 12m 11.3s))

Query Type ○ Contains ◉ Overlaps

Circle Radius Query

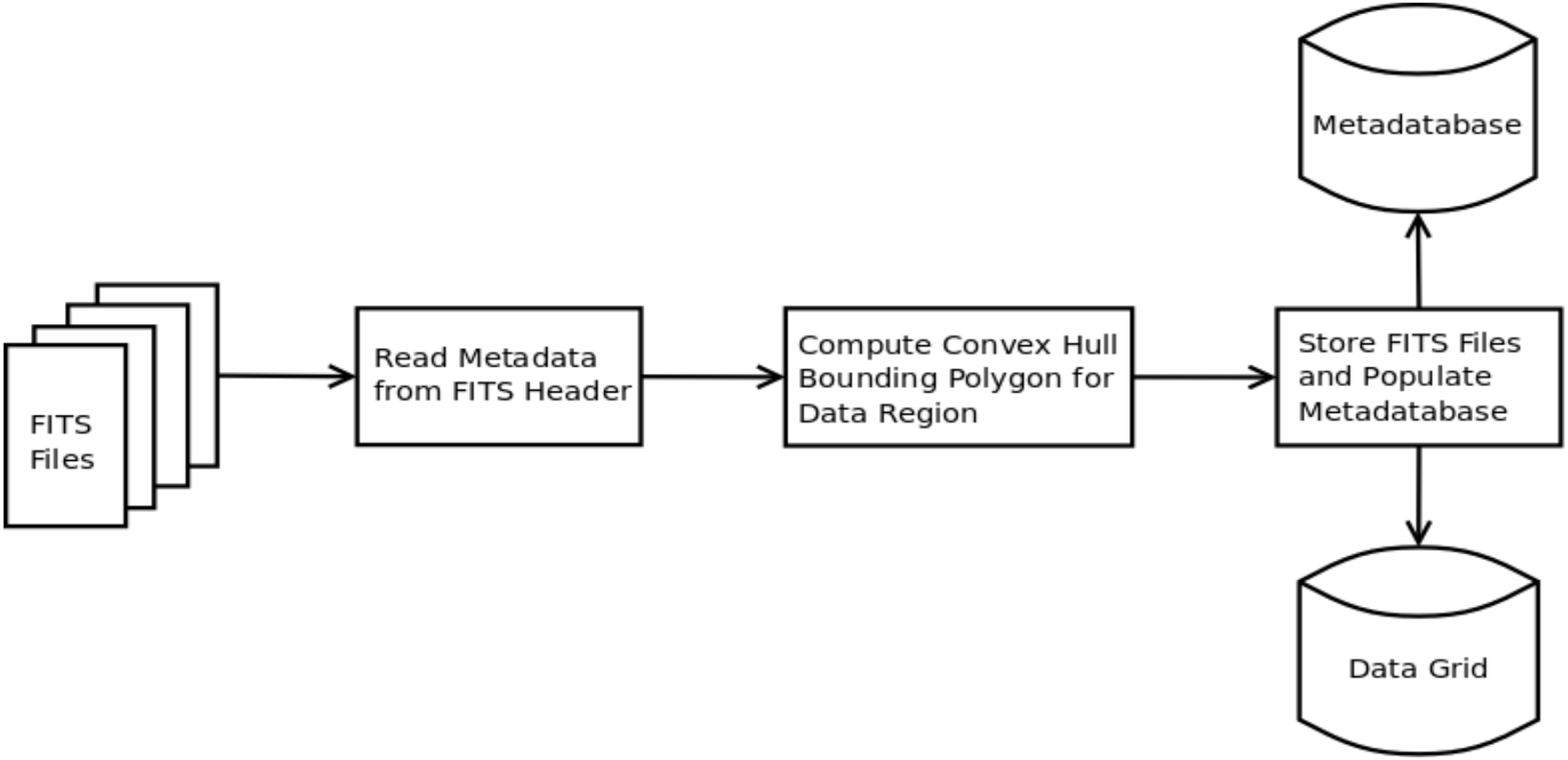
circle x		bounding circle centre x-coordinate i.e. Galactic longitude (in radians (e.g. 0.5), degrees (e.g. 70d), or degrees, minutes, seconds (e.g. 10d 12m 11.3s))
circle y		bounding circle centre y-coordinate i.e. Galactic latitude (in radians (e.g. 0.5), degrees (e.g. 70d), or degrees, minutes, seconds (e.g. 10d 12m 11.3s))
radius		bounding circle radius (in radians (e.g. 0.5), degrees (e.g. 70d), or degrees, minutes, seconds (e.g. 10d 12m 11.3s))

Query Type ◉ Contains ○ Overlaps

Contains Point Query

point x		point x-coordinate i.e. Galactic longitude (in radians (e.g. 0.5), degrees (e.g. 70d), or degrees, minutes, seconds(e.g. 10d 12m 11.3s))
point y		point y-coordinate i.e. Galactic latitude (in radians (e.g. 0.5), degrees (e.g. 70d), or degrees, minutes, seconds(e.g. 10d 12m 11.3s))

Data Query Service (cont.)



Data Query Service (cont.)

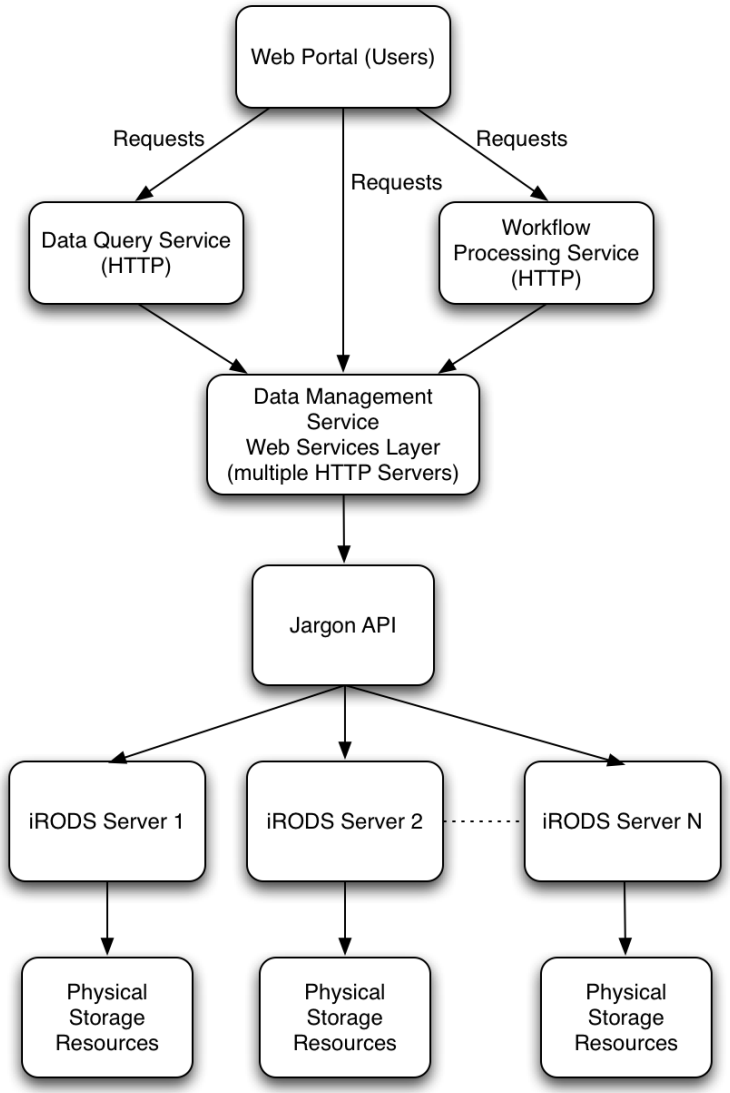
- A spatially enabled PostgreSQL/PgSphere database is used to maintain resource metadata.
- The schema is based on IVOA Resource Metadata recommendations.

Data Query Service (cont.)

- Using a spatially enabled database has key advantages when working with astronomical data:
  - Spatial data types and queries: e.g. polygon contains/overlaps, circle contains/overlaps, etc.
  - Ability to generate complex "astrospatial" queries for data using a more natural SQL syntax.
- GiST (Generalized Search Tree) indexes can be used to speedup spatial queries on large databases.

Workflow Processing/Web Services

- We have developed a web based workflow builder that currently supports image segmentation, image mosaicking (based on the excellent Montage package), spatial reprojection, and plane extraction from data cubes.
- While leveraging distributed data storage and data processing facilities in the background, the user's experience is abstracted away from these details.


Web Portal (Users)
Requests
Requests
Requests
Data Query Service (HTTP)
Workflow Processing Service (HTTP)
Data Management Service Web Services Layer (multiple HTTP Servers)
Jargon API
iRODS Server 1
iRODS Server 2
iRODS Server N
Physical Storage Resources
Physical Storage Resources
Physical Storage Resources

Workflow Processing (cont.)

- In the instance where a user needs to perform a data processing operation on data that is stored at multiple sites, we have implemented the concept of "intelligent" web services.
- Our intelligent web services for workflow (data) processing work as follows:
  - The user assembles a pipeline consisting of discrete data processing operations (for example, image smoothing or image subsetting).

Workflow Processing (cont.)

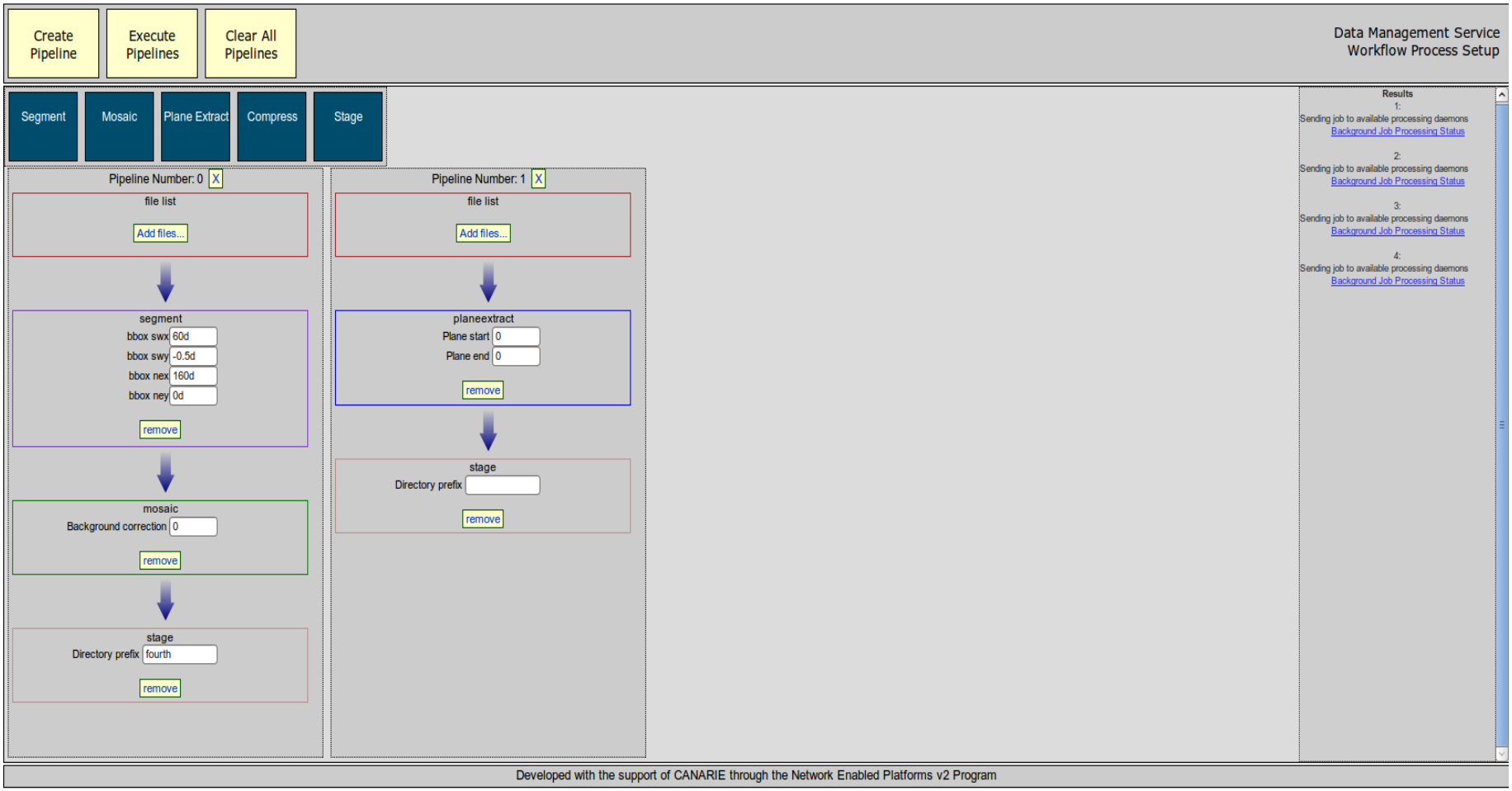
- The output of one component of the pipeline is the input to the next stage of the pipeline.
- The final result is a set of files that have gone through all the processing steps.

▪ Our intelligent web services automatically determine the most efficient course of action regarding where data is to be retrieved from and processed.

Workflow Processing (cont.)

- If a data processing operation can be handled locally at a site without initiating a data transfer within the data grid, the web service does this.
- If a data processing operation requires a data transfer from multiple sites within the data grid (for example, to assemble a mosaic of images), this is also handled automatically.
- The user never needs to manually retrieve any data or specify the locations of any data in order to process it.

Workflow Processing (cont.)



Future Work

- An API for community developed modules to run within the workflow builder is being developed.
- Addition of more data processing web services such as image statistics and Fourier transforms.
- Full integration with the CyberSKA cloud computing framework.

a place of mind

THE UNIVERSITY OF BRITISH COLUMBIA