Globus File Transfer for Common Crawl

Big Data Analytics

Introduction

The Common Crawl is an **open repository of web crawl data** collected over last 7 years which contains several petabytes of data.

It has web page data, meta data extracts and text extracts

The data is stored in **WET**, **WARC** and **WAT** file formats

Stored as chunks of data on **Amazon Web Services - Public Data Sets** across multiple academic cloud platforms.

We need this data for 2 Big Data Processing Projects:

LORELEI language classification
Biological Data classification

Globus File Transfer is needed to transfer the data to the **Pittsburgh Supercomputing Center (PSC)** for carrying out experiments in the above mentioned projects.

ABOUT - Globus File Transfer

The **Globus transfer** service provides high-performance, secure, file transfer and synchronization between endpoints. [Globus]

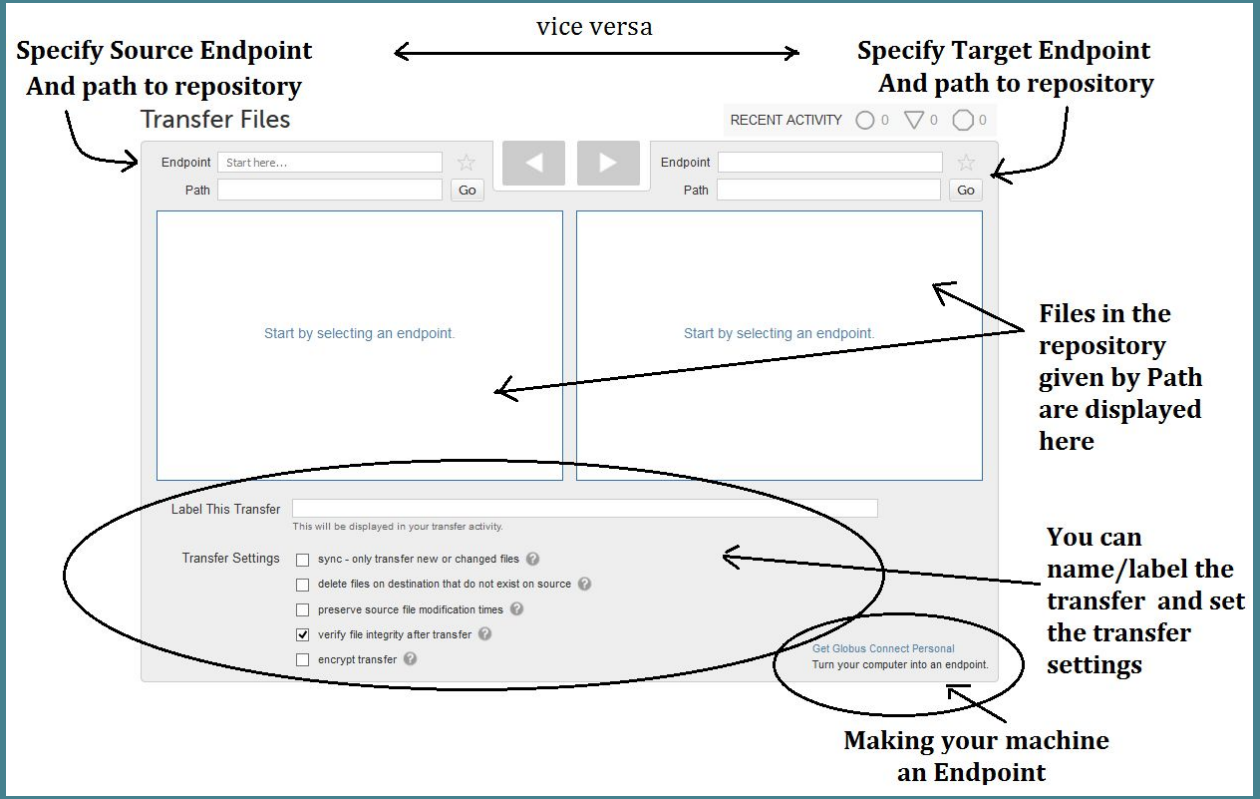
Features of Globus:

- Recover from transfer failures due to network unreliability
- Ensure integrity of transferred files
- Maintain required levels of security
- Can deal with different security configurations between resources
- In- place sharing of data
- Interfaces provided - web browser and command line
- Maximizes bandwidth usage
- Notifies about transfer completion and problems
- Software as a Services (SaaS)

Initial Experiments with Globus Web Interface

Transferred data between PSC Bridges endpoint and our own PC using the web browser.

Realised the need for an API to transfer larger chunks of data at a time. (Especially because the transfer would not be between PCs and dummy end points, but between PSC and Common crawl AWS s3 endpoints)



Experiments with Globus SDK for Python

- The Globus SDK is a Pythonic REST based API
- Provides **Transfer API** and **Globus Auth API** used to transfer files between two resources with different security configurations
- Endpoint IDs of the source and destination needed by Transfer Client.
- Authentication and transfer tokens needed to accomplish transfer.
- Tokens expire every 48 hours. Need to be refreshed.
- PSC requires certain configurations. Globus SDK was not able to transfer efficiently due to these constraints.
- Dropped the idea of using any form of APIs till we figured out the transfer protocols of PSC.

Common Crawl S3 Endpoint

- Globus transfer is done between **endpoints.**
  - **Common Crawl bucket S3 endpoint -> PSC Bridges endpoint**
- Creating endpoint requires a Globus subscription
  - Joined XSEDE Plus Sponsors group
- S3 Endpoint is a **managed endpoint**, one has to be a subscription manager to create an S3 endpoint.
  - Derek Simmel from PSC Bridges created one and gave us the permission to access it.
  - https://docs.globus.org/how-to/amazon-aws-s3-endpoints/
- Each month's data is atleast 50 TB and some cases it was higher than 100 TB. We can estimate the corpus to be around 2 Petabytes (can be much higher).
  - Requires more space on PSC Bridges to run the experiments.

How to create the endpoint?

- Can only be done from a subscription manager account (since S3 endpoint is a managed endpoint)
- https://docs.globus.org/how-to/amazon-aws-s3-endpoints/
  - In Part A, skip steps 2-6 (Creating a policy) - standard policy available works
  - In Part A, steps 7-13 (Creating a role): step 12 (Attach Policy Page): add AmazonS3ReadOnlyAccess - the ARN for this is arn:aws:iam::aws:policy/AmazonS3ReadOnlyAccess
  - In Part B, step 2 (specifying the region specific URL for the S3 bucket): Use the following link https://s3.amazonaws.com/aws-publicdatasets
  - Then, activate the endpoint and test it out from web interface or the CLI.

Data Transfer

- Current transfer speed is low : 25-50 Mb/s
- Need to setup parallel transfers from multiple globus accounts.
- August 2016, July 2016 data downloaded by Derek to pylon2.
  - We didn’t have the storage space to download that scale of data.
- Used web interface for the transfer.