Evaluating NextCloud as a File Storage for Apache Airavata

Sachin Kariyattin
Science Gateway
Research Center,
Indiana University,
skariyat@indiana.edu

Suresh Marru
Science Gateway
Research Center,
Indiana University,
smarru@iu.edu

Marlon Pierce
Science Gateway
Research Center,
Indiana University,
marpierc@iu.edu

ABSTRACT

Science gateways enable researchers from broad communities to access advanced computing and storage resources. The researchers analyze large amounts of data using the compute resources and the generated results, usually files are saved in the storage. Consider a scenario where a researcher has large output data files of historically run experiments on an external server. If the researcher wants to move the data to the gateway storage, then the only way to do it is through data transfer. This task would be cumbersome and time consuming. The paper discusses an approach through which historic or any data existing on a different server or in a cloud storage (Google Drive) or in an object storage (Amazon S3) can be ingested into the existing gateway without actually transferring it to the server. We discuss about a software called NextCloud and how it can be used as a gateway storage by integrating it with Apache Airavata. Airavata currently uses local file storage to store user related data files. On the client side, Airavata clients use different protocols like HTTP and SFTP for file transfer. NextCloud is an open source file share and communication platform that provides a common file access layer through its universal file access to different data sources. Integrating NextCloud with Airavata could solve the problem of providing unified file transfer API across all the Airavata clients. As NextCloud supports various external storages, its integration with Airavata would also enable the data ingestion and importing large data from different storage sources to Airavata.

CCS CONCEPTS

Software and its engineering → Open source model; • Computer systems organization → Cloud computing; • Information systems → Cloud based storage; • Information systems → Computing platforms

KEYWORDS

Apache Airavata, NextCloud, WebDAV, File Transfer, File Storage

ACM Reference format:

S. Kariyattin, S. Marru, and M. Pierce. 2018. Evaluating NextCloud as a File Storage for Apache Airavata. In *Proceedings of Practice and Experience in Advanced Research Computing (PEARC'18) ,July 2018, Pittsburgh, PA, USA*

1 INTRODUCTION

Science Gateways allow users to perform operations on high end resources. Researchers upload input data files on which the computation needs to be done and then analyze the results using the information present in the generated output files. All these files will be stored in the file storage of the deployed gateway environment [6].

Files related to the gateway users would be stored under a single file storage server. Now consider the scenario of integrating some gateway related data which exists in a different server (X) into an existing gateway file storage server (Y). One option would be to transfer all the files from X to Y. Usually, the size of data files is large thus the data transfer might take a long time to complete. Also, what if the external gateway related data is stored in a cloud storage like Dropbox or an object storage like Amazon S3 Ingesting data to the gateway file storage server would be difficult in this case.

In this paper, we try to discuss file storage server alternative for gateways. We will try to look at solutions that allow gateway file storage to register data from different kinds of storages without actually moving the data but instead just registering the storage reference.

2 FILE STORAGE ALTERNATIVES

To access data from different kind of storage references through one file server, we would need a tool that can register different kind of storages and then would allow users to access data from those storages through an API. The tool would act as a central storage with references to other external storages.

Globus solves a similar problem. In order to share data, users have to copy the data to an external storage system usually hosted in the cloud. With Globus users need not move the data in order to share it. Any storage system with Globus connection can be configured to allow secure data sharing directly by the users. Globus basically keeps track of the connected systems and shared files in a central database [1].

Globus enables users to share files without actually storing it. We need a storage server for gateway which can store data files as well as allow access to external storage connected to it. A free and open source solution that we think can be suitable for this approach would be NextCloud. In the further sections, we have tried to explain how NextCloud can be used as a file storage for a gateway by integrating it with Airavata.

PEARC '18, July 22–26, 2018, Pittsburgh, PA, USA
© 2018 Copyright is held by the owner/author(s).
ACM ISBN 978-1-4503-6446-1/18/07.
https://doi.org/10.1145/3219104.3229270

3 NEXT CLOUD OVERVIEW

NextCloud is an open source, self-hosted file share and communication platform. It provides a common file access layer through its Universal File Access, keeping data where it is and retaining the management and control mechanisms IT currently has in place to manage risk. NextCloud unifies data from cloud storage, Windows network drive and legacy data storage to users in a single, easy interface empowering them to access, sync and share files on any device, wherever they are, managed, secured and controlled by IT. The NextCloud Server is a PHP web application running on a Linux web server like Apache or NGINX. It stores file sharing information, user details, application data and configuration as well as file information in a database (MySQL). The files in the NextCloud server are stored in conventional directory structures and can be accessed using the WebDAV protocol. The storage layer can leverage any storage protocol that can be mounted on a server, like NFS, GFS2, Windows Network Drive, CIFS, Red Hat Storage, IBM Elastic Storage and object stores compatible with SWIFT and S3. It is also possible to mount sFTP and external cloud storage services like Google Drive and Dropbox in the user storage [2].

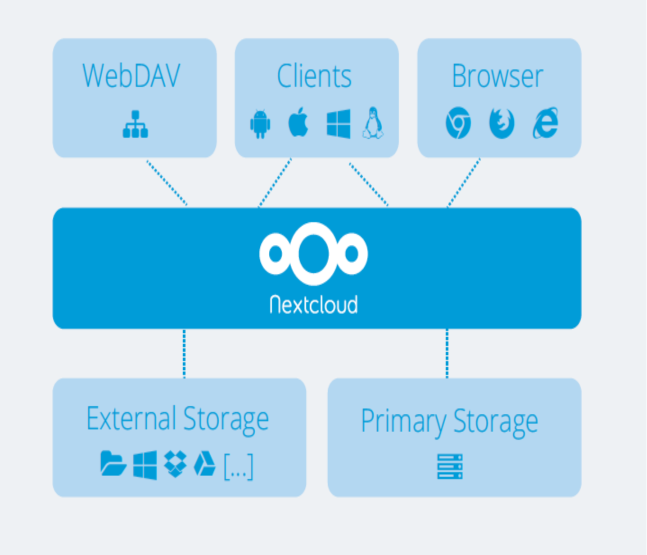


Figure 1: NextCloud unified file access supports various file storage systems

4 AIRAVATA FILE STORAGE

Currently Airavata uses local file storage to store data files related to a user. The Airavata clients use the HTTP or the SFTP protocols in order to upload the files to the Airavata file storage. Using NextCloud for Airavata file storage could solve two problems related to future implementations of Airavata viz. Data Ingestion and API for file uploads.

Data Ingestion. Consider the scenario of ingesting data into the existing Airavata file storage. One way of doing this would be to execute a file transfer task from the source to the Airavata server. However, this would work only when the files are stored in a conventional file system like the one Airavata uses currently. What if the data to be ingested is very large? What if the data is stored in an object storage like Amazon S3? NextCloud can solve these problems. If the data is too large to be transferred, then an SFTP mount of the external data can be created and added as an external storage to NexCloud. Thus, the data need not be moved to the Airavata file storage and can be accessed from where it is. Similarly, if the data to be ingested is in an S3 storage, then S3 location can be added as an external storage device to NextCloud.

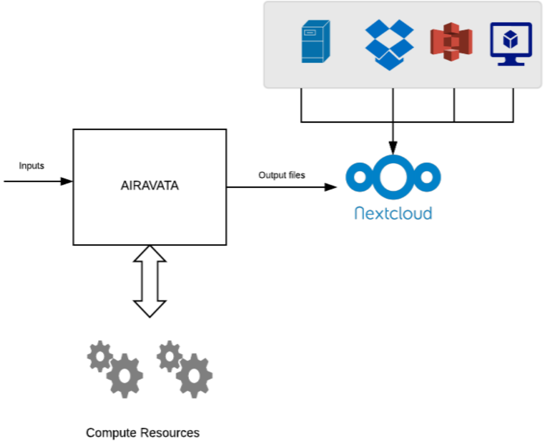


Figure 2: NextCloud supports various external storages

API for File Upload. The current clients of Airavata use different protocols to upload files to the storage. For instance, the Airavata Django and PHP clients use the HTTP protocols for file upload whereas the Java desktop client uses an SFTP implementation built over Apache Mina. As NextCloud provides WebDAV API for file related operations, NextCloud integration with Airavata would provide a unified API to upload files across all the clients.

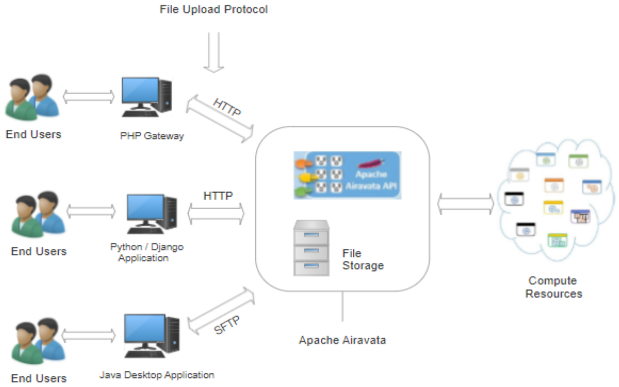


Figure 3: Current clients of Apache Airavata using the HTTP and SFTP file transfer protocols.

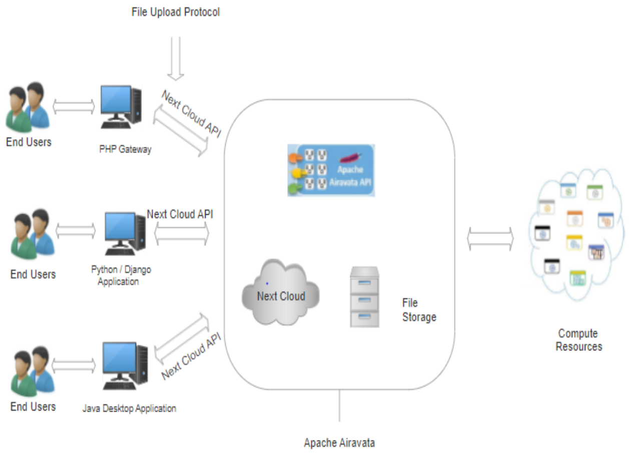


Figure 4: Proposed Architecture of the Apache Airavata using Nextcloud API for file uploads

5 INTEGRATING NEXT CLOUD WITH AIRAVATA

5.1 Keycloak Compatibility

Keycloak is an open source software product to allow single sign-on with Identity Management and Access Management aimed at modern applications and services [4]. Since Airavata uses Keycloak as an identity provider, the first step was to verify whether Nextcloud could be integrated with Keycloak. Using SSO and SAML app for Nextcloud it is easy to integrate existing single sign on solution with Nextcloud. I was able to successfully configure Nextcloud with Keycloak as the identity provider.

The SSO and SAML app needs to be enabled on NextCloud web console. On the Keycloak side, a client can be created and configured with all the required SAML attributes.

The Nextcloud v12 also supports OAuth2 authentication method and various other enterprise authentication methods.

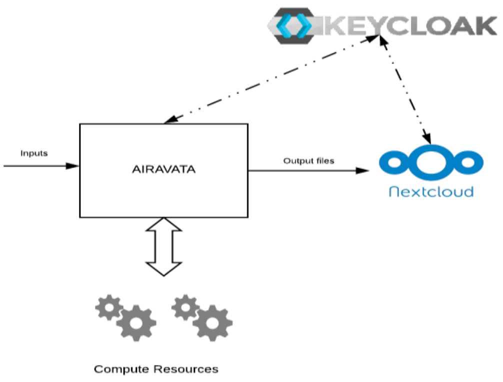
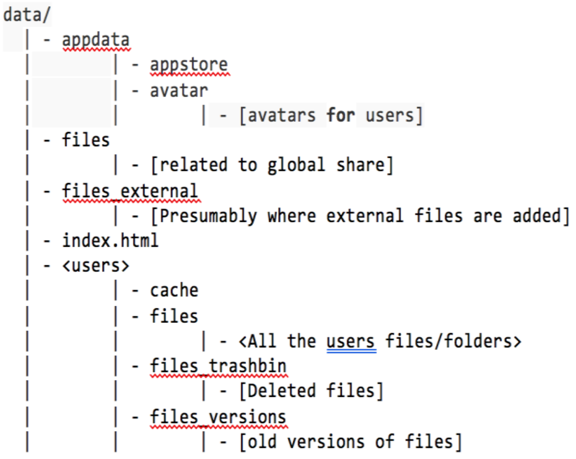


Figure 5: NextCloud authentication with Keycloak

5.2 Importing Large Data to NextCloud

Nextcloud stores all the user data in its data directory (nextcloud/data/<username>/files). Given below is the data directory hierarchy of Nextcloud.



```
data/
  | - appdata
  |         | - appstore
  |         | - avatar
  |         |         | - [avatars for users]
  | - files
  |         | - [related to global share]
  | - files_external
  |         | - [Presumably where external files are added]
  | - index.html
  | - <users>
  |         | - cache
  |         | - files
  |         |         | - <All the users files/folders>
  |         | - files_trashbin
  |         |         | - [Deleted files]
  |         | - files_versions
  |         |         | - [old versions of files]
```

Figure 6: Data directory of NextCloud

So, basically any files/folders can be imported to the <username>/files directory and then run a scan on all the user directories to map the imported files/folders in the NextCloud database.

In order to migrate large data, the data present in the external drive can be mounted (sFTP mount) or by configuring an FTP server where data is present or by directly copying the data into the Nextcloud data directory. Given below are the general steps to follow while importing data to Nextcloud

- Move/copy/mount the files/folders to nextcloud data directory.
- Run the scan command using the next cloud occ utility `sudo nextcloud.occ files:scan –all`

The Airavata data directory structure (gateway/user/projects/…) is very similar to the NextCloud data directory structure. Given the fact that NextCloud does not store the data files for a user in a distributed way makes it easy to import data into Airavata server.

5.3 NextCloud Client API

NextCloud provides WebDAV api for file related operations like listing directories, downloading and uploading files etc. [2] At present, all the Airavata APIs are developed in Java but there is no Java API implementation available for NextCloud. The goal is to develop Airavata API methods that expose file related operations by communicating with the NextCloud server. One way to do this would be to use an open source

WebDAV Java client like Sardine and wrap it inside Airavata API methods.

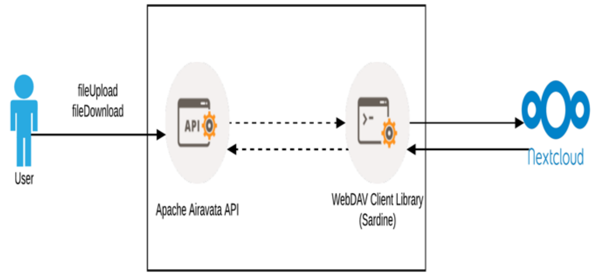


Figure 7: Airavata API using WebDAV client library for file related operations

5.4 Impact to Airavata Sharing Service

Airavata uses sharing service to share experiment data files with different users by registering it in the database [6]. The NextCloud UI also provides an option to the users to share files or directories by either generating a dynamic link or by sharing it with a particular user. However, if a user shares the file/folder using just the NextCloud UI, then the sharing information won't be registered in the database as it would bypass the Airavata API. One possible way of solving this problem would be to completely disable sharing from the NextCloud UI and allow sharing from only the Airavata sharing service. This can be done by using the File Access Control application. Using this app, admin can configure the NextCloud server by disabling the file sharing.

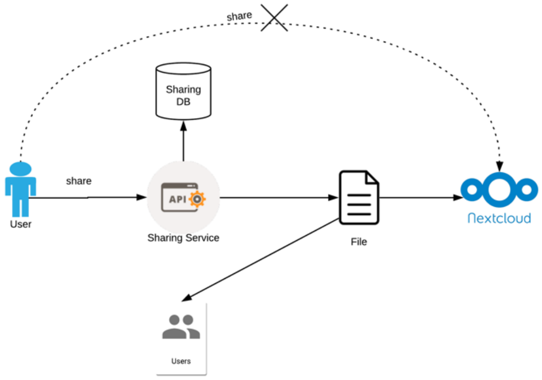


Figure 8: Using Airavata Sharing service instead of NextCloud direct sharing

6 FUTURE WORK

The paper discussed using NextCloud as a possible file storage for gateways with Apache Airavata. Even though this POC proves that NextCloud could be used as Airavata file storage, the performance and other related aspects can only be analyzed once it is integrated into the actual production gateway environment. The first step would be to develop API methods using the WebDAV client library which would act as a bridge between Airavata and NextCloud integration. Further, inking external data from different servers or object storage like S3 or Google drive need to be tested.

7 CONCLUSION

In this paper, We tried to present NextCloud as a file storage for a science gateway with Airavata and its advantages. The paper discusses how NextCloud can be configured to work with Airavata and also how it can be used enable data ingestion from different data sources without actually transferring the data from external storages. The integration of NextCloud with Airavata will also provide unified API to upload files to all the gateway clients. We have also discussed about how NextCloud can be configured with Airavata's existing Keycloak identity management server.

REFERENCES

[1] File Sharing with Globus. https://www.globus.org/data-sharing. Accessed 2018-04-30

[2] Nextcloud Solution Architecture. https://nextcloud.com.vn/wp-content/themes/nextcloud.vn/assets/files/architecture-whitepaper.pdf. Accessed 2018-04-30

[3] Science Gateway. https://en.wikipedia.org/wiki/Science_gateway. Accessed 2018-04-30

[4] Keycloak. https://en.wikipedia.org/wiki/Keycloak. Accessed 2018-04-30

[5] The Apache Airavata Application Programming Interface: Overview and Evaluation with the UltraScan Science Gateway. Marlon Pierce and Suresh Marru. https://goo.gl/bJm9bQ

[6] Apache Airavata: Design and Directions of a Science Gateway Framework. Marlon Pierce and Suresh Marru. https://ieeexplore.ieee.org/document/6882068/?reload=true

[7] Apache Airavata Sharing Service: A Tool for Enabling User Collaboration in Science Gateways. Supun Nakandla et. al. https://dl.acm.org/citation.cfm?id=3093359