Delta: An Open Source Data Service for Research

Lexington Whalen
Department of Computer Science
University of South Carolina
Columbia, United States
LAWHALEN@email.sc.edu

Homayoun Valafar
Department of Computer Science
University of South Carolina
Columbia, United States
homayoun@cec.sc.edu

***Abstract*—Every research project necessitates data, often requiring sharing and collaborative review within a team. Traditionally, services like Dropbox [2], Google Drive [1], or OneDrive [3] have been employed for storing research data. However, the limitations of these platforms become evident with the growing scale of collected data. Existing file-sharing services generally mandate paid subscriptions for increased storage or additional members, diverting research funds from addressing the core research problem that a lab is attempting to work on. Moreover, these services often lack direct features for reviewing or commenting on data quality, a vital part of ensuring high quality data generation. In response to these challenges, we present Delta, a specialized file transfer service crafted for specifically for researchers. Delta operates as an application hosted on a research lab server. This design ensures that, with access to a machine and an internet connection, teams can facilitate file storage, transfer, and review without incurring extra costs. Being an open-source project, Delta can be customized to suit the unique requirements of any research team, and is able to evolve to meet the needs of the research community. We open source the code here: https://github.com/lxaw/Delta.**

***Index Terms*—data transfer, data storage, data reviewing, data sharing**

I. Introduction

There is a noticeable dearth of quality open source research services with regards to data. Most data services are merely file sharing ones, that only allow file upload, download, and organization. While this does allow researchers to transmit their collected data among another, it lacks many critical features that most researchers would like to have, such as quality control, commenting, tagging, organization under groups, annotation, and more. Furthermore, these services are oftentimes prohibitively expensive to smaller research labs, and their free versions have limitations on both number of users and amount of storage. The current state of research is to use services such as Microsoft's OneDrive [3], Dropbox [2], or Google Drive [1]. Each of these services [something about service limitations]

Furthermore, the modern times have seen unprecedented growth in the amount of data collected. [something about deep learning data, internet of things, personal data growth]

II. The Problem

[write about the architecture]

III. Current Alternatives

Several cloud storage providers offer APIs and developer tools that enable programmatic access to their services. For example:

- Google Drive provides a REST API [1] for uploading, downloading, searching, and manipulating files stored on Google Drive. Client libraries are available in multiple programming languages.
- Dropbox offers a similar REST API [2] with SDKs for major platforms to integrate Dropbox capabilities into applications. Key features include file upload/download, sharing, search, and user management.
- Microsoft OneDrive also has a comprehensive REST API [3] for accessing OneDrive files, folders, and other data. Client libraries support integration with web, mobile, and desktop apps.
- Box provides a content management platform with a REST API [4] for building custom applications. It offers features like file preview, version control, and granular access permissions, catering to enterprise needs.
- Amazon S3 (Simple Storage Service) is a scalable object storage service with an API [5] for storing and retrieving data. While not primarily designed for end-user file management, it's often used as a foundation for building cloud storage applications.

While these APIs enable building custom applications with cloud storage, the underlying limitations of the services still apply - costs scale with storage and user needs, and specialized features for research data management are lacking. Leveraging the APIs still requires significant development effort to create a tailored solution. Furthermore, some services like Amazon S3 are more suited to developers and lack user-friendly interfaces out-of-the-box. Furthermore, these services are not open sourced, so any modifications (i.e. addition of service, design change, bug fix) cannot be done directly by the users. We make such issues more clear in *Issues with Current Approaches*.

IV. Issues with Current Approaches

While current cloud storage services offer APIs and enable building custom applications, they have several notable shortcomings, especially in the context of research data management:

1) **Cost:** Services like Google Drive, Dropbox, and OneDrive can become expensive as data storage needs grow, often requiring paid subscriptions for additional storage or users. This diverts funds from core research activities.
2) **Lack of Specialization:** General-purpose storage services lack features tailored for research, such as data quality control, peer review workflows, metadata management, and data provenance tracking.
3) **Limited Customization:** Although APIs enable custom app development, the underlying platforms cannot be easily modified or extended. Researchers cannot add new features or modify existing behavior to suit their specific needs.
4) **Vendor Lock-In:** By relying on proprietary services, researchers risk being locked into a particular vendor's ecosystem. Migration to alternative platforms can be difficult and costly.
5) **Data Ownership and Control:** With commercial services, there may be ambiguity around data ownership and control. Researchers may have concerns about intellectual property rights and the ability to access their data if a service is discontinued.
6) **Data Privacy and Security:** Storing sensitive research data on third-party servers raises privacy and security concerns. Researchers may be hesitant to entrust confidential data to external providers.
7) **Collaboration Barriers:** While cloud services facilitate file sharing, they often lack advanced collaboration features like real-time co-authoring, version control, and granular access controls that are vital for research teams.
8) **Integration Challenges:** Integrating cloud storage with existing research tools and workflows can be challenging. Researchers may need to develop custom glue code or rely on limited third-party integrations.
9) **Dependency on Internet Connectivity:** Cloud services require reliable internet access, which can be a constraint in field research settings or areas with limited connectivity.
10) **Long-Term Preservation:** Commercial services may not prioritize long-term data preservation, which is crucial for research reproducibility and data archiving. There may be uncertainties about data durability and accessibility over extended periods.

An open source solution like Delta addresses these issues by providing a specialized, customizable, and cost-effective platform for research data management. By hosting Delta on their own infrastructure, research teams have full control over their data, can tailor the platform to their specific needs, and avoid vendor lock-in. The open source nature ensures transparency, enables community-driven development, and allows for integration with a wide range of tools. Moreover, an open source solution can be deployed in local or offline environments, mitigating concerns about internet connectivity and data privacy. Delta empowers researchers to manage their data on their own terms, prioritizing the unique requirements of scientific research.

V. Delta Architecture

[write about delta architecture]

A. Backend

B. Frontend

VI. Current Product

VII. Future Goals

References

Please number citations consecutively within brackets [**?**]. The sentence punctuation follows the bracket [**?**]. Refer simply to the reference number, as in [**?**]—do not use "Ref. [**?**]" or "reference [**?**]" except at the beginning of a sentence: "Reference [**?**] was the first . . ."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [**?**]. Papers that have been accepted for publication should be cited as "in press" [**?**]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [**?**].

References

[1] Google, "Google Drive: Free Cloud Storage for Personal Use," Google.com, 2019. https://www.google.com/drive/

[2] Dropbox, "Dropbox," Dropbox, 2018. https://www.dropbox.com/

[3] "Personal Cloud Storage – Microsoft OneDrive," Microsoft.com, 2024. https://www.microsoft.com/en-us/microsoft-365/onedrive/

[4] Box, "Box Developer Documentation," Box, 2023. https://developer.box.com/

[5] Amazon Web Services, "Amazon S3 API Reference," Amazon.com, 2023. https://docs.aws.amazon.com/AmazonS3/latest/API/Welcome.html