2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)

DFS: A Dataset File System for Data Discovering Users

Yasith Jayawardana, and Sampath Jayarathna
(yasith,sampath)@cs.odu.edu
Old Dominion University
Norfolk, VA 23529

Abstract

Many research questions can be answered quickly and efficiently using data already collected for previous research. This practice is called secondary data analysis, and has gained popularity due to lower costs and improved research efficiency. In this paper we propose DFS, a file system to standardize the metadata representation of datasets, and DDU, a scalable architecture based on DFS for semi-automated metadata generation and data recommendation on the cloud, and explores their implications on datasets stored in digital libraries.

CCS Concepts

• **Information systems** → **Data management systems**; • **Applied computing**; • **Computing methodologies** → *Machine learning*;

Keywords

data recommendation, data discovering users, meta data

1 INTRODUCTION

With the advancements in digital technology, researchers have access to a vast amount of data collected during past research. This data is utilized by many research communities to fuel entirely new research or to expand on the original study. This practice, termed secondary data analysis, enables conducting non-experimental research with minimal cost. However, selecting a dataset for secondary data analysis is a complex process that involves searching for datasets, analyzing candidate datasets for applicability, and data wrangling [4]. The required pre-processing varies across different file types and data, and cannot be pre-determined without understanding the nature of data. Under such constraints, secondary data analysis could become overly complex, which is detrimental to the quality and efficiency of research.

We hypothesize that a standardized metadata format would compensate for this by streamlining information management in datasets and laying the groundwork for rule-based and machine learning algorithms to generate metadata. Though dataset versioning could be challenging, studies have shown efficient methods for versioning data and metadata [2]. Thus, standardizing metadata would enable semi-automated metadata management which, in turn, would drastically simplify secondary data analysis.

2 DATA DISCOVERING USERS

Figure 1 shows an overview of how different components are interconnected in DDU. Datasets are stored in repositories, which are responsible for handling data replication and version control. The DDU servers maintain metafiles for each dataset, and indexes datasets by their fields. The metadata is generated using machine learning and improved through crowd-sourcing. The ISDs expose APIs for users to query, discover and fetch datasets. Crowd-sourced metadata is fed into the DDU Servers through ISDs, to improve existing metadata. DDU Profilers maintain interest profile models (IPMs) [3] for tracking user interests and providing intelligent matches. The user-facing component, which is DDU SaaS, utilizes all aforementioned components to provide an ecosystem for collaborative research.

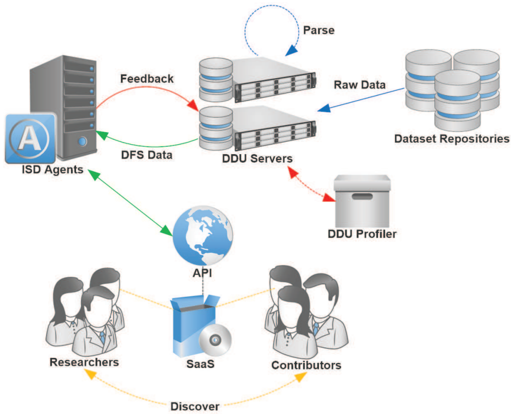

Figure 1: Architecture of DDU including DDU Profilers[3]

3 DATASET FILE SYSTEM

The main component of DFS is the metadata file (or metafile), which serves as the entry point to a dataset. Each metafile stores information about the dataset, data files, and data fields. This enables multiple data files to behave as one coherent set of data. Figure 2 shows a sample metafile (shortened for brevity) in DFS. Each metafile contains the fields "id" and "version" to uniquely identify a dataset. The "version" increments as the data files are subjected to revision. The "meta-version" field auto-increments as the metadata

```
"$schema": "dataset",
"$id": "ISSN-000-0000-0001", "version": "1.1", "meta-version": "3",
"name": "EEG Readings..", "description": "...",
"domain": "MEDICINE",
"tags": ["EEG", "ADOS", "ASD"],
"author": "...", "author_id": "...", "copyright": ".....",
"signature": "43278947328957439805847390257439205874390258473590",
"created": "12-20-2018", "modified": "01-27-2019", "published": "01-28-2019",
"files": [{
  "$id": "data_1", "path": "./test.json", "encoding": "JSON",
  "fields": [{"name": "child", "type": "ID", "description": "..."}],
  "description": "", "measured_variables": "", "measured_devices": [],
  "md5": "0123015035783941274895378"
}],
"links": [{
  "type": "ID", "description": "...",
  "fields": ["data_1.child", "data_2.id"]
}]
```

Figure 2: Sample Metafile in DFS

978-1-7281-1547-4/19/$31.00 ©2019 IEEE
DOI 10.1109/JCDL.2019.00068

matures over time. The "name" and "description" fields describe the dataset in a human-readable format. The "domain" field indicates the primary research domain that the dataset was created for (e.g. MEDICINE). The "tags" field indicates the research sub-domains, and is updated dynamically through crowd-sourcing and user profiling. The "author", "author_id", and "copyright" and fields provide authenticity while the "signature" field, which stores a digital signature of the metafile and data files, provides integrity. The "created", "modified" and "published" fields provide a timeline for dataset activity. The metafile points to data files through "files" field. It captures the path, file format, recording conditions, and fields in the data file. It also maintains a hash of the data file for integrity. The "links" section defines the semantic relationships between the fields in data files (e.g. identity).

The objective of the metafile is to capture as much information as possible about the underlying dataset, such that it eliminates the need to rely on external documentation to understand the dataset semantics. As an added benefit, the "id" and "version" fields provide version control capability and reference immutability, making them an ideal candidate for citation. Citing datasets using DFS would resolve ambiguities caused by evolving datasets.

3.1 APPLICATIONS OF DFS

The objective of DFS is to provide the infrastructure for representing datasets in rich detail, effectively eliminating user dependence on external sources to comprehend them. The most obvious application of DFS is for DDU. Here, DFS lays the groundwork for semi-automated metadata management in DDU. Apart from DDU, we identified several applications of DFS that are described in the sections below.

3.1.1 Integrating with existing tools: Data Wrangler [5] is a tool that supports creation and execution of pre-processing scripts for data files. Using the information present in DFS metafiles, these scripts could be automatically generated to simplify the effort needed for pre-processing. This information could also be leveraged to generate automatic machine learning pipelines using tools such as Auto Keras. A sample integration of DFS and DDU with these tools is shown in Figure 3.

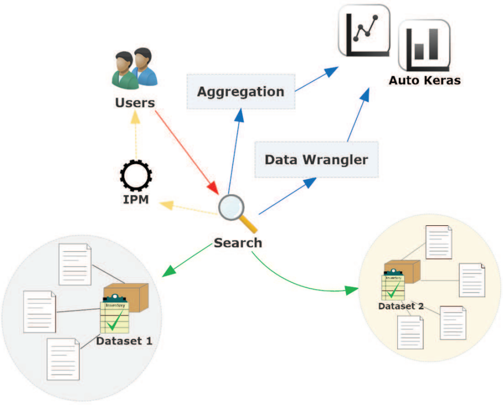

Figure 3: Integration with Data Wrangler and Auto Keras

3.1.2 Dataset aggregation: Dataset aggregation is the process of comparing datasets using their field information to determine if they could be merged. Studies have proposed methods for calculating dataset similarity using schema overlap [1], and for merging datasets using scalable algorithms [6]. Since DFS metafiles provide information about the data fields and how they are related to each other, datasets could be compared using their metafiles to determine if they could be merged.

Algorithm 1 provides a pseudo-code for dataset aggregation using DFS. For each metafile, the fields and their links are represented as a graph. Next, the two graphs are compared using graph similarity algorithms to determine if the datasets are comparable. If so, a join operation is performed on the fields and links of the two metafiles based on the schema overlap. This results in a connected graph which represents links among both datasets. This information is then used to create a new metafile which represents data from both datasets.

Algorithm 1: Dataset Aggregation using Metafiles

```
function aggregate $(\alpha, \beta)$ :
    if $similarity(graph(\alpha), graph(\beta)) \leq \epsilon$ then
        throw error;
    forall $\gamma \leftarrow fields(\alpha)$ do
        forall $\delta \leftarrow fields(\beta)$ do
            if $overlap(\gamma, \delta) \geq \sigma$ then
                $\alpha \leftarrow metajoin(\alpha, \beta, \gamma, \delta)$;
    return $\alpha$;
```

4 CONCLUSIONS AND FUTURE OUTLOOK

DFS and DDU provide a fresh outlook to how data is discovered, wrangled, and used for data analytics and machine learning. With DFS bringing new techniques for dataset aggregation and DDU enabling semi-automated metadata management and user interest profiling, research communities could collaborate efficiently on research and accelerate workflows.As a future work, we plan to evaluate the compatibility of DFS across multiple domains and file types to evaluate the cross-domain coverage of DFS.

References

[1] Mohamed Ben Ellefi, Zohra Bellahsene, Stefan Dietze, and Konstantin Todorov. 2016. Dataset Recommendation for Data Linking: An Intensional Approach. In *The Semantic Web. Latest Advances and New Domains*, Harald Sack, Eva Blomqvist, Mathieu d'Aquin, Chiara Ghidini, Simone Paolo Ponzetto, and Christoph Lange (Eds.). Springer International Publishing, Cham, 36–51.

[2] Souvik Bhattacherjee, Amit Chavan, Silu Huang, Amol Deshpande, and Aditya Parameswaran. 2015. Principles of dataset versioning: Exploring the recreation/storage tradeoff. *Proceedings of the VLDB Endowment* 8, 12 (2015), 1346–1357.

[3] Sampath Jayarathna and Frank Shipman. 2017. Analysis and Modeling of Unified User Interest. In *Information Reuse and Integration (IRI), 2017 IEEE International Conference on.* IEEE, 298–307.

[4] Sean Kandel, Jeffrey Heer, Catherine Plaisant, Jessie Kennedy, Frank van Ham, Nathalie Henry Riche, Chris Weaver, Bongshin Lee, Dominique Brodbeck, and Paolo Buono. 2011. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization* 10, 4 (2011), 271–288.

[5] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. 2011. Wrangler: Interactive Visual Specification of Data Transformation Scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 3363–3372. https://doi.org/10.1145/1978942.1979444

[6] Hung-Chih Yang, Ali Dasdan, and Ruey-Lung Hsiao. 2009. Map-reduce with merge to process multiple relational datasets. https://patents.google.com/patent/US7523123B2/en US Patent 7,523,123.