DFS: A Dataset File System for Data Discovering Users

Yasith Jayawardana and Sampath Jayarathna

Computer Science, Old Dominion University, Norfolk, VA

ABSTRACT

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.

BACKGROUND

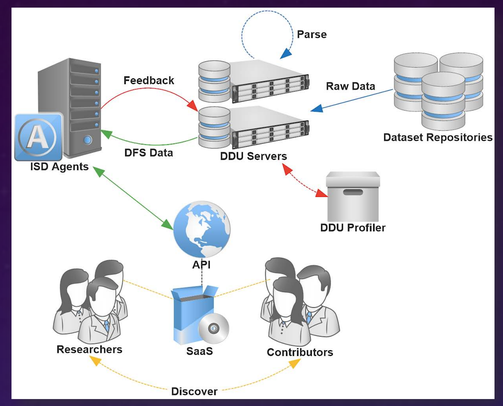
With the advancements in digital technology, researchers have access to a vast amount of data collected during past research. They are utilized by many research communities to fuel entirely new research or to expand on the original study. This practice, termed secondary data analysis, enables conducting 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. 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.

HYPOTHESIS

We hypothesize that a standardized metadata format would compensate for the tedious preprocessing and knowledge gathering steps required to understand and interpret datasets with inadequate documentation, by streamlining information management in datasets, introducing dataset versioning, and laying the groundwork for rule-based and machine learning algorithms to generate metadata.

DATA DISCOVERING USERS



1. Architecture of DDU including DDU Profilers [1]

DATASET FILE SYSTEM

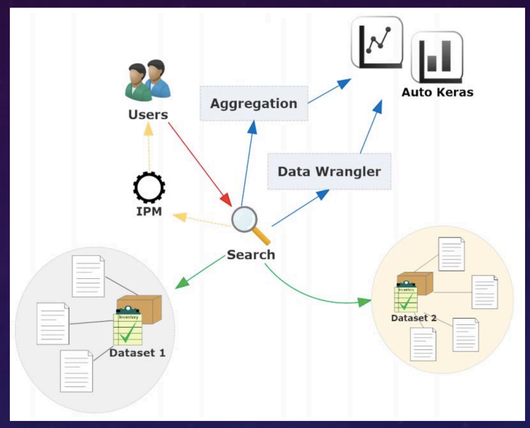
The main component of DFS is the metadata file (or metafile), which serves as the entry point to a dataset. Each metafile stores information about the dataset, data files, and data fields. This enables multiple data files to behave as one coherent set of data.

```
"$schema": "dataset",
"$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": "432789473289574398058473902574392058743902584735 90",
"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": "012301503578394127 4895378"
}],
"links": [{
  "type": "ID", "description": "...",
  "fields": ["data_1.child", "data_2.id"]
}]
```

2. Sample Metafile in DFS

The objective of the metafile is to capture as much information as possible about the underlying dataset, effectively eliminating the need for external documentation to understand the dataset semantics. In addition, the "id" and "version" fields provide version control capability and reference immutability, making it an ideal candidate for citation.

APPLICATIONS



3. Integration with Data Wrangler and Auto Keras

Algorithm 1: Dataset Aggregation using Metafiles

function ***aggregate*** (α, β) **:**
  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 α;

FUTURE WORK

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. In the future, we plan to evaluate the compatibility of DFS across multiple domains and file types to evaluate the cross-domain coverage of DFS.

REFERENCES

1. S. Jayarathna and F. Shipman. “Analysis and Modeling of Unified User Interest.” IEEE IRI, 2017.