# MESS:
*Meta-data Extraction System for Schemas*

**William Spoth & Oliver Kennedy**
University at Buffalo

University at Buffalo
The State University of New York

**Abstract**

Data is often collected first and cleaned later as an after thought. It is common for essential information to the curation process to be embedded into file names, time stamps, or other file system meta-data. A common workload will use some aspect of this meta-data to sort or select over, such as 'ORDER BY Date', 'WHERE StudyNumber = 1', etc. Through operator pushdown on meta-data columns, we can eliminate paths in our Relational Algebra tree, reduce the number of files accessed, and increase performance.

**Problem Statement**

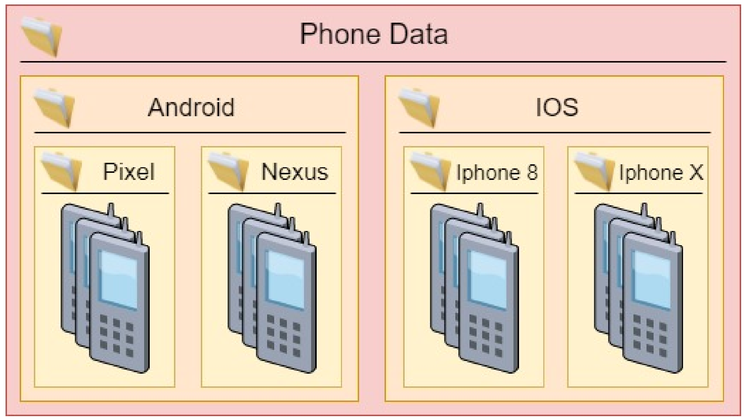
- Extracting and projecting meta-data is tedious and is given no priority over regular data and columns.
- Users should be able to pass 'hints' to their optimizers. These hints can reduce the time spent on heavy operations such as DISTINCT and ORDER BY, eliminate paths before any files are accessed, and partition files initially for Spark.
- Some example hints are, 'these files are already sorted', 'these files are all of the same type', 'only check these files for the information', etc.

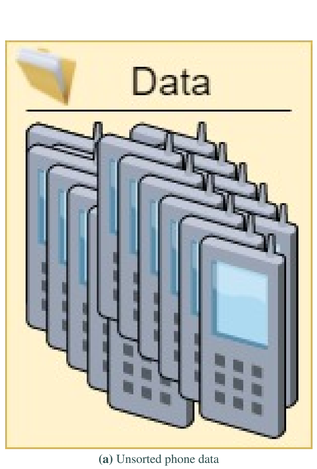
**Relevant Work**

- NoDB's [1] ability to reduce the number of file read operators.
- HDFS, Apache Spark [2] , and Apache Drill's ability to partition data based on key or in this case by file.
- Oracle's SQL*Loader has a similar goal for loading and partitioning batches of files but by using a control file.
- Data Wrangler explores an interactive and UI driven approach to curating data.

**Applications**

Consider a study where a researcher wishes to collect responses from participants and users' device information is to be added in post processing. Conveniently to distinguish these files apart, the file names contain this information to be recovered.



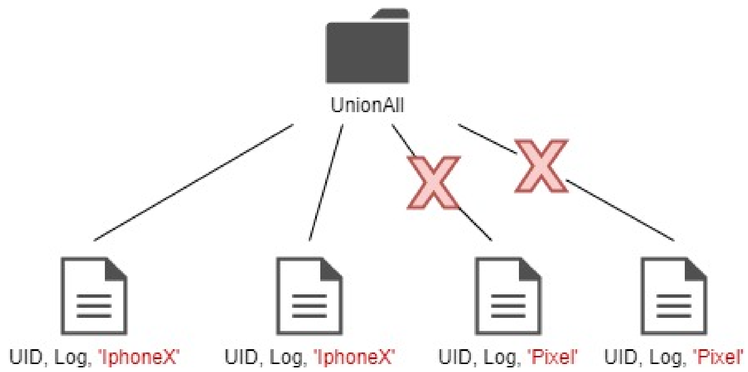

(a) Unsorted phone data


(b) Displaying typical file naming

UID, Name, LogTrace, PhoneBrand, OperatingSystem

Because meta-data is a constant value that is projected for each row, we can push down our selection operator and evaluate our predicate before the query is executed.

```
SELECT * FROM Data WHERE PhoneBrand = 'IphoneX'
```




**Challenges**

- Creating a system that allows for easy meta-data incorporation and code re-usability.
- Creating general rules for the optimizer to apply the hints to.

**References**

[1] Ioannis Alagiannis, Renata Borovica, Miguel Branco, Stratos Idreos, and Anastasia Ailamaki. Nodb: Efficient query execution on raw data files.

[2] Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster computing with working sets.