CMPS 203 Project Proposal: A Unified File System Query Language with Provenance Support

Yan Li
yanli@ucsc.edu

Yunfei Chen
ychen@soe.ucsc.edu

***Abstract*—We will design a new file system query language because no previous query language can cover all these three query categories: free text, structural metadata and provenance.**

I. Clearly state the question to be solved

The question to be answered: how to design a query language that can gracefully cover three categories of information from one or more file systems: free text, structural metadata and provenance. The target audience is scientific communities who has a large sum of data to query on a daily base.

What will be covered:

- Why we need a new query language for file systems.
- It should work for both hierarchical and non-hierarchical file systems.
- We will start with a formal and strict query language, and may later add a free-form layer upon it for better usability, because the scientific community we were targeting may not have good training in computer science, and the prevailing Google has spoiled the mass.
- How the result should be presented in a way that is suitable for a very large amount of files.
- File system permission and security, i.e., you won't be able to query files that you have no permission to access, also, a system administrator should be able to query files that are accessible to a specific user or a group of users.

What will not be covered:

- How to efficiently implement the query language. If we have enough time we may be able to come up with a simple prototype, which should be able to show the essence of the language but we won't have time to implement an optimized version.

Who will benefit when the problem is solved (e.g. CSP/DataMesh, HPL, IBM)?

HPC communities that maintain large amount of files in a file system.

II. Why it is important? Why is it a problem?

Some of the file systems used in scientific HPC systems are reaching several peta bytes, and by that trend they will reach exascale in the following decade. In these file systems there are very large amount of files, sometimes cluttered in one single directory. Traditional hierarchical style of organizing files is inadequate for harnessing files of this large amount, and searching is becoming the major method for accessing these large file systems. Also the introduction of provenance into the scientific computation communities is pushing the need for a general query language that supports provenance query.

III. Why are all previous approaches insufficient?

Current file system query languages are designed before the peta- and exa-scale HPC era, and they failed in the following ways:

- handling very large amount of files
- handling non-hierarchical file systems
- returning results in a way that is meaningful for handling files with similar or identical names
- query free form text, structural metadata and provenance data in a uniformed way

IV. What is my approach?

What is the basic approach, method, idea or tool thats being suggested to solve the problem? (E.g. dynamic disk shuffling, stainless-steel mousetrap

springs, an AI tool for writing monthly progress reports.)

Design a new query language, give a description of the grammar, run it over test file sets to verify its effectiveness.

V. Hypotheses

What exactly are the expected effects of the proposed solution? (E.g. disk I/O time will increase to 2 seconds per request.) Why is this?

The proposed new language should be able to handling query of the three categories of information in a unified way. The grammar should be as natural and as general as possible for better usability.

VI. Collaborator

List all collaborator in the order of their importance.

Yan Li (from Storage Systems Research Center) and Yunfei Chen (from Information Retrieval and Knowledge Management Lab) will collaborate on this project.

VII. Experiments

What will be done to test out the hypotheses? (E.g. measurements, simulations, constructing code, thinking beautiful thoughts, hard vacationing). How will this confirm (or deny) the hypotheses? Why will the conclusions be believable?

A good query language need to be both expressive and effective. Being expressive means it has enough ability to express complex enough queries. Being effective means the query should reflect the user's intention: with high recall and precision, as well as low false-positive.

We will test these two aspects on various file sets.

Device/Equipment Needed

No special device needed.

VIII. Results

What will be the outcome of the work (papers, a working system, a graph of ...)? When?

We will write a paper to discuss the new query language and the experiments.

IX. Plan

The Winter 2012 ends on Mar 16. We have 8 weeks from now on (Jan 21st). Week 1 is Jan 22 to Jan 28. All milestones are to be checked at the Friday of that week.

Week	Milestones
1	Find personal data test file sets. Find scientific data test file sets.
2	Compose sample queries. Find sample provenance data.
4	Language definition.
5	Test on sample file sets.
7	Paper and poster.
8	Buffer

X. Prior Art

This work will be based upon the following literature.

Free text searching tools:

- keywords
- field-restricted search
- boolean queries
- phrase search
- proximity search
- regular expression
- wildcard search

XQuery is a query and functional programming language that is designed to query collections of XML data.

PQL[1] is a recent provenance query language.

We will explore more related works in the research project.

References

[1] David A. Holland, Uri Braun, Diana Maclean, Kiran-Kumar Muniswamy-Reddy, and Margo I. Seltzer. Choosing a data model and query language for provenance. In *Second International Provenance and Annotation Workshop*, 2008.