CS554 Project Ideas

XSearch: Benchmark the state-of-the-art distributed search platforms (on disk)

Overview

Computer clusters, supercomputers, grid systems and cloud systems have become popular platforms for performing scientific research, pushing forward not only the boundaries of scientific discovery but also leading the advancement of computer technology. Applications that generally run on these platforms can take hours or days to complete and commonly access and process data sets that reach sizes of hundreds of terabytes. In order to store and access large quantities of data computing centers and data centers employ parallel and distributed file systems at large scales, such as: IBM Spectrum Scale, Lustre, OrangeFS, the Hadoop File System and Ceph. While in literature the focus is on developing techniques and strategies for systems and applications to store, access and process data in a serene manner and in a reasonable amount of time, searching and aggregating information over the large storage systems are challenges that have not been thoroughly explored.

The larger objective of the project is to design and build a distributed indexing platform that would allow scientists and engineers to do search and extract relevant information from large scale storage systems in a short period of time, without interfering with other applications, allowing the submission of complex queries and giving control over the data back to the users. Modern personal computer desktops and laptops have search capabilities build with the operating systems that allows users to quickly retrieve relevant documents from the local file system. A lot of effort has been put into search engines, such as: Google, Bing and Duck Duck Go; that allow users to query the Internet for relevant web pages in a matter of seconds and with high accuracy. When moving to the computing center and data center side, modern parallel and distributed systems lack any king of search capabilities over an ocean of generally unstructured data that is dominated by scientific formats, including numerical data and image files. In order to discover files either by name or by content scientists most often rely on classical Linux tools, such as: *ls* and *grep* or *find*; or implement static programs that can use the intrinsic interfaces of the file system to perform parallel search, but still in a exhaustive manner. Previous work shows the limitation of classical Linux tools [2]. Other communities choose to build catalogs and databases in which data can be structure, stored and then queried, but this requires specialized intervention from the user for structuring and storing the information in the required format and usually not reusable between scientific fields.

In this project you will benchmark several popular distributed search platforms from the perspective of how performance is defined in parallel and distributed systems. You will focus on Apache Solr, Elasticsearch and a third solution which can be either a proposed custom implementation (a naive one is acceptable) or another platform. You can choose as a third option GUFI (Grand Unified File-Index), that you can find at this link: https://github.com/mar-file-system/GUFI.You will focus on identifying the bottlenecks of both the functional components of the library and the built search engine as a whole. You will vary through your experiments the location where data and indexes are stored, by setting up and exploring different parallel distributed systems, such as Luster, OrangeFS, HDFS and Ceph. You will benchmark the libraries on both typical text data (i.e. the wikipedia dumps) and file system metadata from real supercomputing centers.

CS554 Project Ideas

Relevant Systems and Reading Material

Please read the following papers (and their references) before submitting your proposal:

[1] Itua Ijagbone, “Scalable indexing and searching on distributed file systems”. Master thesis. Illinois Institute of Technology, 2016

Available online: http://datasys.cs.iit.edu/publications/2016_IIT_MS-thesis_Itua-Ijagbone.pdf

[2] Alexandru Iulian Orhean, Itua Ijagbone, Dongfang Zhao, Kyle Chard, Ioan Raicu. “Toward Scalable Indexing and Search on Distributed and Unstructured Data”, IEEE Big Data Congress 2017

Available online: https://ieeexplore.ieee.org/abstract/document/8029306/

[3] Alexandru Iulian Orhean, Kyle Chard, Ioan Raicu. “XSearch: Distributed Information Retrieval in Large-Scale Storage Systems”. Oral qualifying exam. Illinois Institute of Technology, 2018.

Posted on Piazza.

Prefered/Required Skills

- Principles: operating systems, distributed systems, file systems, information retrieval, databases;
- Programming: C, C++, Java (only for Lucene) Bash, multi-threaded/multi-process programming;
- Operating System: Linux/UNIX;

Evaluation and Metrics:

You need to run multi-node scalability experiments and evaluate the indexing throughput, index size and query throughput, as a function of increasing number of files, increasing file size, increasing number of threads, and other factors that could have an impact on performance. Conduct experiments on Chameleon Cloud bare-metal instances.

Project Mentor

Alexandru Iulian Orhean -- aorhean@hawk.iit.edu