BlurProposal

Blur Proposal

Abstract

Blur is a search platform capable of searching massive amounts of data in a cloud computing environment. Blur leverages several existing Apache projects, including Apache Lucene, Apache Hadoop, Apache ZooKeeper and Apache Thrift. Both bulk and near real time (NRT) updates are possible with Blur. Bulk updates are accomplished using Hadoop Map/Reduce and NRT are performed through direct Thrift calls.

Proposal

Blur is an open source search platform capable of querying massive amounts of data at incredible speeds. Rather than using the flat, document-like data model used by most search solutions, Blur allows you to build rich data models and search them in a semi-relational manner similar to joins while querying a relational database. Using Blur, you can get precise search results against terabytes of data at Google-like speeds. Blur leverages multiple open source projects including Hadoop, Lucene, Thrift and ZooKeeper to create an environment where structured data can be transformed into an index that runs on a Hadoop cluster. Blur uses the power of Map/Reduce for bulk indexing into Blur. Server failures are handled automatically by using ZooKeeper for cluster state and HDFS for index storage.

Background

Blur was created by Aaron McCurry in 2010. Blur was developed to solve the challenges in dealing with searching huge quantities of data that the traditional RDBMS solutions could not cope with while still providing JOIN-like capabilities to query the data. Several other open source projects have implemented aspects of this design including elasticsearch, Katta and Apache Solr.

Rationale

There is a need for a distributed search capability within the Hadoop ecosystem. Currently, there are no other search solutions that natively leverage HDFS and the failover features of Hadoop in the same manner as the Blur project. The communities we expect to be most interested in such a project are government, health care, and other industries where scalability is a concern. We have made much progress in developing this project over the past 2 years and believe both the project and the interested communities would benefit from this work being openly available and having open development. In future versions of Blur the API will more closely follow the API's provided in Lucene so that systems that already use Lucene can more easily scale with Blur. Blur can be viewed as a query execution engine that Lucene based solutions can utilize when scale becomes an issue.

Initial Goals

The initial goals of the project are:

- To migrate the Blur codebase, issue tracking and wiki from github.com and integrate the project with the ASF infrastructure.
- Add new committers to the project and grow the community in "The Apache Way".

Current Status

Meritocracy

Blur was initially developed by Aaron McCurry in June 2010. Since then Blur has continued to evolve with the support of a small development team at Near Infinity. As a part of the Apache Software Foundation, the Apache Blur team intends to strongly encourage the community to help with and contribute to the project. Apache Blur will actively seek potential committers and help them become familiar with the codebase.

Community

A small community has developed around Blur and several project teams are currently using Blur for their big data search capability. The source code is currently available on GitHub and there is a dedicated website (blur.io) that provides an overview of the project. Blur has been shared with several members of the Apache community and has been presented at the Bay Area HUG (see http://www.meetup.com/hadoop/events/20109471/).

Core Developers

The current developers are employed by Near Infinity Corporation, but we anticipate interest developing among other companies.

Alignment

Blur is built on top of a number of Apache projects; Hadoop, Lucene, ZooKeeper, and Thrift. It builds with Maven. During the course of Blur development, a couple of patches have been committed back to the Lucene project, including LUCENE-2205 and LUCENE-2215. Due to the strong relationship with the before mentioned Apache projects, the incubator is a good match for Blur.

Known Risks

Orphaned Products

There is only a small risk of being orphaned. The customers that currently use Blur are committed to improving the codebase of the project due to its fulfilling needs not addressed by any other software. In addition, one customer is providing financial support to further develop Blur given its importance on mission-critical projects.

Inexperience with Open Source

The codebase has been treated internally as an open source project since its beginning, and Near Infinity has extensive experience developing and releasing open source projects (http://www.nearinfinity.com/products/open_source). We do not anticipate difficulty in operating under the Apache Way.

Homogeneous Developers

Current developers are all employed by Near Infinity but we are actively seeking contributors from different companies and would welcome their participation.

Reliance on Salaried Developers

Blur was originally created by Aaron McCurry as a personal project and he remains the primary contributor. Currently, Aaron's employer (Near Infinity) fully supports his continued participation with paid, dedicated time to work on Blur. All other current developers are paid by Near Infinity to work on Blur as well.

Relationships with Other Apache Products

Blur dependencies:

- Apache Hadoop
- Apache Lucene
- Apache ZooKeeper
- Apache Thrift
- Apache log4j

Apache Brand

Our interest in releasing this code as an Apache project is due to its strong relationship with other Apache projects, i.e. Blur has dependencies on Hadoop, Lucene, ZooKeeper, and Thrift and its uniqueness within the Hadoop ecosystem.

Documentation

Current documentation can be found at http://blur.io and https://github.com/nearinfinity/blur.

Initial Source

Blur has been in development since summer 2010. The core codebase consists of about ~29,000 (~10,000 if the generated RPC code is not included) lines of code mainly Java.

Source and Intellectual Property Submission Plan

Blur core code, examples, documentation, and training materials will be submitted by Near Infinity Corporation.

External Dependencies

- concurrentlinkedhashmap - Apache 2.0 License - http://code.google.com/p/concurrentlinkedhashmap/

Cryptography

none

Required Resources

- Mailing Lists
  - blur-private
  - blur-dev
  - blur-commits
  - blur-user
- Subversion Directory

    - https://git-wip-us.apache.org/repos/asf/blur.git
- Issue Tracking
    - JIRA
- Continuous Integration
    - Jenkins
- Web
    - http://incubator.apache.org/blur/wiki at http://wiki.apache.org or http://cwiki.apache.org

Initial Committers

- Aaron McCurry (aaron.mccurry at nearinfinity dot com)
- Scott Leberknight (scott.leberknight at nearinfinity dot com)
- Ryan Gimmy (ryan.gimmy at nearinfinity dot com)
- Tim Williams (twilliams at apache dot org)
- Patrick Hunt (phunt at apache dot org)
- Doug Cutting (cutting at apache dot org)

Affiliations

- Aaron McCurry, Near Infinity
- Scott Leberknight, Near Infinity
- Ryan Gimmy, Near Infinity
- Patrick Hunt, Cloudera
- Doug Cutting, Cloudera

Sponsors

- Champion: Patrick Hunt

Nominated Mentors

- Tim Williams (twilliams at apache dot org)
- Doug Cutting (cutting at apache dot org)
- Patrick Hunt (phunt at apache dot org)

Sponsoring Entity

- Apache Incubator