HamaProposal

Abstract

Hama will develop a parallel matrix computational package based on Hadoop Map/Reduce.

Proposal

Hama will develop a parallel matrix computational package, which provides an library of matrix operations for the large-scale processing development environment and Map/Reduce framework for the large-scale Numerical Analysis and Data Mining, which need the intensive computation power of matrix inversion, e.g. linear regression, PCA, SVM and etc. It will be also useful for many scientific applications, e.g. physics computations, linear algebra, computational fluid dynamics, statistics, graphic rendering and many more.

Background

Currently, several shared-memory based parallel matrix solutions can provide a scalable and high performance matrix operations, but matrix resources can not be scalable in the term of complexity. And, Hadoop HDFS Files and Map/Reduce can only used by 1D blocked algorithm.

Rationale

Hama approach proposes the use of 3-dimensional Row and Column (Qualifier), Time space and multi-dimensional Columnfamilies of Hbase, which is able to store large sparse and various type of matrices (e.g. Triangular Matrix, 3D Matrix, and etc.) and utilize the 2D blocked algorithm. its auto-partitioned sparsity sub-structure will be efficiently managed and serviced by Hbase. Row and Column operations can be done in linear-time, where several algorithms, such as *structured Gaussian elimination* or *iterative methods*, run in O(the number of non-zero elements in the matrix / number of mappers) time on Hadoop Map/Reduce.

Current Status

In its current state, the 'hama' is buggy and needs filling out, but generalized matrix interface and basic linear algebra operations was implemented within a large prototype system. In the future, We need new parallel algorithms based on Map/Reduce for performance of heavy decompositions and factorizations. It also needs tools to compose an arbitrary matrix only with certain data filtered from hbase array structure.

Meritocracy

The initial developers are very familiar with meritocratic open source development, both at Apache and elsewhere. Apache was chosen specifically because the initial developers want to encourage this style of development for the project.

Community

Hama seeks to develop developer and user communities during incubation.

Core Developers

The initial set of Hama committers includes folks from the Hadoop & Hbase communities. We have varying degrees of experience with Apache-style open source development.

Alignment

The developers of Hama want to work with the Apache Software Foundation specifically because Apache has proven to provide a strong foundation and set of practices for developing standards-based infrastructure and server components.

Known Risks

Orphaned products

Most of the active developers would like to become Hama Committers or PMC Members and have long term interest to develop/maintain and **use** the code.

Inexperience with Open Source

We has already a good experience with Apache open source development process.

Homogenous Developers

The current list of Hama committers includes developers from several different companies (NHN, corp, TMAX software, Korea Research Institute of Bioscience and Biotechnology, Students) plus many independent volunteers. The committers are geographically distributed across the Europe, and Asia. They are experienced with working in a distributed environment.

Reliance on Salaried Developers

It is expected that Hama development will occur on both salaried time and on volunteer time, after hours. While there is reliance on salaried developers (currently from NHN, corp, but it's expected that other company's salaried developers will also be involved), the Hama Community is very active and things should balance out fairly quickly. In the meantime, NHN, corp might support the project in the future by dedicating 'work time' to Hama, so that there is a smooth transition.

Relationships with Other Apache Products

Hama has a strong relationship with Apache Hadoop, Hbase and Mahout. Being part of Apache could help for a closer collaboration between the three projects.

A Excessive Fascination with the Apache Brand

We believe in the processes, systems, and framework Apache has put in place. The brand is nice, but is not why we wish to come to Apache.

Documentation

- http://code.google.com/p/hama/w/list

Initial Source

- http://code.google.com/p/hama/source/checkout

External Dependencies

- Hadoop (HDFS, Map/Reduce) License: Apache License, 2.0
- Hbase (Sparse Matrix Table) License: Apache License, 2.0

Required Resources

- Developer and user mailing lists
  - hama-private@incubator.apache.org
  - hama-commits@incubator.apache.org
  - hama-dev@incubator.apache.org
  - hama-user@incubator.apache.org
- A subversion repository
  - https://svn.apache.org/repos/asf/incubator/hama
- A JIRA issue tracker

Initial Committers

- Edward J. Yoon, NHN (edward AT udanax DOT org)
- Chanwit Kaewkasi, Univ. of Manchester (chanwit AT gmail DOT com)
- Suh ChangHee, NHN (bluesvm AT gmail DOT com)
- Ha Yongho, TmaxSoft (yongho.ha AT gmail DOT com)
- Hong Taehui, KRIBB (hongtebari AT gmail DOT com)
- Yoon JooSun, NHN (ologist0 AT gmail DOT com)

Sponsors

Mentors

- Ian Holsman, (ianh AT apache DOT org)
- Jeff Eastman, (jeastman AT apache DOT org)
- Brett Porter, (brett AT apache DOT org)

Sponsoring Entity

The Apache Incubator.