HAWQProposal

Abstract

HAWQ is an advanced enterprise SQL on Hadoop analytic engine built around a robust and high-performance massively-parallel processing (MPP) SQL framework evolved from Pivotal Greenplum Database.

HAWQ runs natively on Apache Hadoop clusters by tightly integrating with HDFS and YARN. HAWQ supports multiple Hadoop file formats such as Apache Parquet, native HDFS, and Apache Avro. HAWQ is configured and managed as a Hadoop service in Apache Ambari. HAWQ is 100% ANSI SQL compliant (supporting ANSI SQL-92, SQL-99, and SQL-2003, plus OLAP extensions) and supports open database connectivity (ODBC) and Java database connectivity (JDBC), as well. Most business intelligence, data analysis and data visualization tools work with HAWQ out of the box without the need for specialized drivers.

A unique aspect of HAWQ is its integration of statistical and machine learning capabilities that can be natively invoked from SQL or (in the context of PL /Python, PL/Java or PL/R) in massively parallel modes and applied to large data sets across a Hadoop cluster. These capabilities are provided through MADlib – an existing open source, parallel machine-learning library. Given the close ties between the two development communities, the MADlib community has expressed interest in joining HAWQ on its journey into the ASF Incubator and will be submitting a separate, concurrent proposal.

HAWQ will provide more robust and higher performing options for Hadoop environments that demand best-in-class data analytics for business critical purposes. HAWQ is implemented in C and C++.

HAWQ has a few runtime dependencies licensed under the Cat X list:

- gperf (GPL Version 3)
- libgsasl (LGPL Version 2.1)
- libuuid-2.26 (LGPL Version 2)

However, given the runtime (dynamic linking) nature of these dependencies it doesn't represent a problem for HAWQ to be considered an ASF project.

Proposal

The goal of this proposal is to bring the core of Pivotal Software, Inc.'s (Pivotal) Pivotal HAWQ codebase into the Apache Software Foundation (ASF) in order to build a vibrant, diverse and self-governed open source community around the technology. Pivotal has agreed to transfer the brand name "HAWQ" to Apache Software Foundation and will stop using HAWQ to refer to this software if the project gets accepted into the ASF Incubator under the name of "Apache HAWQ (incubating)". Pivotal will continue to market and sell an analytic engine product that includes Apache HAWQ (incubating). While HAWQ is our primary choice for a name of the project, in anticipation of any potential issues with PODLINGNAMESEARCH we have come up with two alternative names: (1) Hornet; or (2) Grove.

Pivotal is submitting this proposal to donate the HAWQ source code and associated artifacts (documentation, web site content, wiki, etc.) to the Apache Software Foundation Incubator under the Apache License, Version 2.0 and is asking Incubator PMC to establish an open source community.

Background

While the ecosystem of open source SQL-on-Hadoop solutions is fairly developed by now, HAWQ has several unique features that will set it apart from existing ASF and non-ASF projects. HAWQ made its debut in 2013 as a closed source product leveraging a decade's worth of product development effort invested in Greenplum Database. Since then HAWQ has rapidly gained a solid customer base and became available on non-Pivotal distributions of Hadoop. In 2015 HAWQ still leverages the rock solid foundation of Greenplum Database, while at the same time embracing elasticity and resource management native to Hadoop applications. This allows HAWQ to provide superior SQL on Hadoop performance, scalability and coverage while also providing massively-parallel machine learning capabilities and support for native Hadoop file formats. In addition, HAWQ's advanced features include support for complex joins, rich and compliant SQL dialect and industry-differentiating data federation capabilities. Dynamic pipelining and pluggable query optimizer architecture enable HAWQ to perform queries on Hadoop with the speed and scalability required for enterprise data warehouse (EDW) workloads. HAWQ provides strong support for low-latency analytic SQL queries, coupled with massively parallel machine learning capabilities. This enables discovery-based analysis of large data sets and rapid, iterative development of data analytics applications that apply deep machine learning – significantly shortening data-driven innovation cycles for the enterprise.

Hundreds of companies and thousands of servers are running mission-critical applications today on HAWQ managing over PBs of data.

Rationale

Hadoop and HDFS-based data management architectures continue their expansion into the enterprise. As the amount of data stored on Hadoop clusters grows, unlocking the analytics capabilities and democratizing access to that treasure trove of data becomes one of the key concerns. While Hadoop has no shortage of purposefully designed analytical frameworks, the easiest and most cost-effective way to onboard the largest amount of data consumers is provided by offering SQL APIs for data retrieval at scale. Of course, given the high velocity of innovation happening in the underlying Hadoop ecosystem, any SQL-on-Hadoop solution has to keep up with the community. We strongly believe that in the Big Data space, this can be optimally achieved through a vibrant, diverse, self-governed community collectively innovating around a single codebase while at the same time cross-pollinating with various other data management communities. Apache Software Foundation is the ideal place to meet those ambitious goals. We also believe that our initial experience of bringing Pivotal Gemfire into ASF as Apache Geode (incubating) could be leveraged thus improving the chances of HAWQ becoming a vibrant Apache community.

Initial Goals

Our initial goals are to bring HAWQ into the ASF, transition internal engineering processes into the open, and foster a collaborative development model according to the "Apache Way." Pivotal and its partners plan to develop new functionality in an open, community-driven way. To get there, the existing internal build, test and release processes will be refactored to support open development.

Current Status

Currently, the project code base is commercially licensed and is not available to the general public. The documentation and wiki pages are available at FIXME. Although Pivotal HAWQ was developed as a proprietary, closed-source product, its roots are in the PostgreSQL community and the internal engineering practices adopted by the development team lend themselves well to an open, collaborative and meritocratic environment.

The Pivotal HAWQ team has always focused on building a robust end user community of paying and non-paying customers. The existing documentation along with StackOverflow and other similar forums are expected to facilitate conversions between our existing users so as to transform them into an active community of HAWQ members, stakeholders and developers.

Meritocracy

Our proposed list of initial committers include the current HAWQ R&D team, Pivotal Field Engineers, and several existing partners. This group will form a base for the broader community we will invite to collaborate on the codebase. We intend to radically expand the initial developer and user community by running the project in accordance with the "Apache Way". Users and new contributors will be treated with respect and welcomed. By participating in the community and providing quality patches/support that move the project forward, contributors will earn merit. They also will be encouraged to provide non-code contributions (documentation, events, community management, etc.) and will gain merit for doing so. Those with a proven support and quality track record will be encouraged to become committers.

Community

If HAWQ is accepted for incubation, the primary initial goal will be transitioning the core community towards embracing the Apache Way of project governance. We would solicit major existing contributors to become committers on the project from the start.

Core Developers

A few of HAWQ's core developers are skilled in working as part of openly governed Apache communities (mainly around Hadoop ecosystem). That said, most of the core developers are currently NOT affiliated with the ASF and would require new ICLAs before committing to the project.

Alignment

The following existing ASF projects can be considered when reviewing HAWQ proposal:

Apache Hadoop is a distributed storage and processing framework for very large datasets, focusing primarily on batch processing for analytic purposes. HAWQ builds on top of two key pieces of Hadoop: YARN and HDFS. HAWQ's community roadmap includes plans for contributing Hadoop around HDFS features and increasing support for C and C++ clients.

Apache Spark™ is a fast engine for processing large datasets, typically from a Hadoop cluster, and performing batch, streaming, interactive, or machine learning workloads. Recently, Apache Spark has embraced SQL-like APIs around DataFrames at its core. Because of that we would expect a level of collaboration between the two projects when it comes to query optimization and exposing HAWQ tables to Spark analytical pipelines.

Apache Hive™ is a data warehouse software that facilitates querying and managing large datasets residing in distributed storage. Hive provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL. Hive is also providing HCatalog capabilities as table and storage management layer for Hadoop, enabling users with different data processing tools to more easily define structure for the data on the grid. Currently the core Hive and HAWQ are viewed as complimentary solutions, but we expect close integration with HCatalog given its dominant position for metadata management on the Hadoop clusters.

Apache Phoenix is a high performance relational database layer over HBase for low latency applications. Given Phoenix's exclusive focus on HBase for its data management backend and its overall architecture around HBase's co-processors, it is unlikely that there will be much collaboration between the two projects.

Known Risks

Development has been sponsored mostly by a single company (or its predecessors) thus far and coordinated mainly by the core Pivotal HAWQ team.

For the project to fully transition to the Apache Way governance model, development must shift towards the meritocracy-centric model of growing a community of contributors balanced with the needs for extreme stability and core implementation coherency.

The tools and development practices in place for the Pivotal HAWQ product are compatible with the ASF infrastructure and thus we do not anticipate any on-boarding pains.

The project currently includes a modified version of PostgreSQL 8.3 source code. Given the ASF's position that the PostgreSQL License is compatible with the Apache License version 2.0, we do NOT anticipate any issues with licensing the code base. However, any new capabilities developed by the HAWQ team once part of the ASF would need to be consumed by the PostgreSQL community under the Apache License version 2.0.

Orphaned products

Pivotal is fully committed to maintaining its position as one of the leading providers of SQL-on-Hadoop solutions and the corresponding Pivotal commercial product will continue to be based on the HAWQ project. Moreover, Pivotal has a vested interest in making HAWQ successful by driving its close integration with both existing projects contributed by Pivotal including Apache Geode (incubating) and MADlib (which is requesting Incubation), and sister ASF projects. We expect this to further reduces the risk of orphaning the product.

Inexperience with Open Source

Pivotal has embraced open source software since its formation by employing contributors/committers and by shepherding open source projects like Cloud Foundry, Spring, RabbitMQ and MADlib. Individuals working at Pivotal have experience with the formation of vibrant communities around open technologies with the Cloud Foundry Foundation, and continuing with the creation of a community around Apache Geode (incubating). Although some of the initial committers have not had the experience of developing entirely open source, community-driven projects, we expect to bring to bear the open development practices that have proven successful on longstanding Pivotal open source projects to the HAWQ community. Additionally, several ASF veterans have agreed to mentor the project and are listed in this proposal. The project will rely on their collective guidance and wisdom to quickly transition the entire team of initial committers towards practicing the Apache Way.

Homogeneous Developers

While most of the initial committers are employed by Pivotal, we have already seen a healthy level of interest from existing customers and partners. We intend to convert that interest directly into participation and will be investing in activities to recruit additional committers from other companies.

Reliance on Salaried Developers

Most of the contributors are paid to work in the Big Data space. While they might wander from their current employers, they are unlikely to venture far from their core expertise and thus will continue to be engaged with the project regardless of their current employers.

Relationships with Other Apache Products

As mentioned in the Alignment section, HAWQ may consider various degrees of integration and code exchange with Apache Hadoop, Apache Spark and Apache Hive projects. We expect integration points to be inside and outside the project. We look forward to collaborating with these communities as well as other communities under the Apache umbrella.

An Excessive Fascination with the Apache Brand

While we intend to leverage the Apache ‘branding’ when talking to other projects as testament of our project’s ‘neutrality’, we have no plans for making use of Apache brand in press releases nor posting billboards advertising acceptance of HAWQ into Apache Incubator.

Documentation

The documentation is currently available at http://hawq.docs.pivotal.io/

Initial Source

Initial source code will be available immediately after Incubator PMC approves HAWQ joining the Incubator and will be licensed under the Apache License v2.

Source and Intellectual Property Submission Plan

As soon as HAWQ is approved to join the Incubator, the source code will be transitioned via an exhibit to Pivotal's current Software Grant Agreement onto ASF infrastructure and in turn made available under the Apache License, version 2.0. We know of no legal encumberments that would inhibit the transfer of source code to the ASF.

External Dependencies

Runtime dependencies:

- gimli (BSD)
- openldap (The OpenLDAP Public License)
- openssl (OpenSSL License and the Original SSLeay License, BSD style)
- proj (MIT)
- yaml (Creative Commons Attribution 2.0 License)
- python (Python Software Foundation License Version 2)
- apr-util (Apache Version 2.0)
- bzip2 (BSD-style License)
- curl (MIT/X Derivate License)
- gperf (GPL Version 3)
- protobuf (Google)
- libevent (BSD)
- json-c (https://github.com/json-c/json-c/blob/master/COPYING)
- krb5 (MIT)
- pcre (BSD)

- libedit (BSD)
- libxml2 (MIT)
- zlib (Permissive Free Software License)
- libgsasl (LGPL Version 2.1)
- thrift (Apache Version 2.0)
- snappy (Apache Version 2.0 (up to 1.0.1)/New BSD)
- libuuid-2.26 (LGPL Version 2)
- apache hadoop (Apache Version 2.0)
- apache avro (Apache Version 2.0)
- glog (BSD)
- googlemock (BSD)

Build only dependencies:

- ant (Apache Version 2.0)
- maven (Apache Version 2.0)
- cmake (BSD)

Test only dependencies:

- googletest (BSD)

Cryptography N/A

Required Resources

Mailing lists

- private@hawq.incubator.apache.org (moderated subscriptions)
- commits@hawq.incubator.apache.org
- dev@hawq.incubator.apache.org
- issues@hawq.incubator.apache.org
- user@hawq.incubator.apache.org

Git Repository

https://git-wip-us.apache.org/repos/asf/incubator-hawq.git

Issue Tracking

JIRA Project HAWQ (HAWQ)

Other Resources

Means of setting up regular builds for HAWQ on builds.apache.org will require integration with Docker support.

Initial Committers

- Lirong Jian
- Hubert Huan Zhang
- Radar Da Lei
- Ivan Yanqing Weng
- Zhanwei Wang
- Yi Jin
- Lili Ma
- Jiali Yao
- Zhenglin Tao
- Ruilong Huo
- Ming Li
- Wen Lin
- Lei Chang
- Alexander V Denissov
- Newton Alex
- Oleksandr Diachenko
- Jun Aoki
- Bhuvnesh Chaudhary
- Vineet Goel
- Shivram Mani
- Noa Horn
- Sujeet S Varakhedi
- Junwei (Jimmy) Da
- Ting (Goden) Yao
- Mohammad F (Foyzur) Rahman
- Entong Shen

- George C Caragea
- Amr El-Helw
- Mohamed F Soliman
- Venkatesh (Venky) Raghavan
- Carlos Garcia
- Zixi (Jesse) Zhang
- Michael P Schubert
- C.J. Jameson
- Jacob Frank
- Ben Calegari
- Shoabe Shariff
- Rob Day-Reynolds
- Mel S Kiyama
- Charles Alan Litzell
- David Yozie
- Ed Espino
- Caleb Welton
- Parham Parvizi
- Dan Baskette
- Christian Tzolov
- Tushar Pednekar
- Greg Chase
- Chloe Jackson
- Michael Nixon
- Roman Shaposhnik
- Alan Gates
- Owen O'Malley
- Thejas Nair
- Don Bosco Durai
- Konstantin Boudnik
- Sergey Soldatov
- Atri Sharma

Affiliations

- Barclays: Atri Sharma
- Bloomberg: Justin Erenkrantz
- Hortonworks: Alan Gates, Owen O'Malley, Thejas Nair, Don Bosco Durai
- WANDisco: Konstantin Boudnik, Sergey Soldatov
- Pivotal: everyone else on this proposal

Sponsors

Champion

Roman Shaposhnik

Nominated Mentors

The initial mentors are listed below:

- Alan Gates - Apache Member, Hortonworks
- Owen O'Malley - Apache Member, Hortonworks
- Thejas Nair - Apache Member, Hortonworks
- Konstantin Boudnik - Apache Member, WANDisco
- Roman Shaposhnik - Apache Member, Pivotal
- Justin Erenkrantz - Apache Member, Bloomberg

Sponsoring Entity

We would like to propose Apache incubator to sponsor this project.