HivemallProposal

Abstract

Hivemall is a library for machine learning implemented as Hive UDFs/UDAFs/UDTFs.

Hivemall runs on Hadoop-based data processing frameworks, specifically on Apache Hive, Apache Spark, and Apache Pig, that support Hive UDFs as an extension mechanism.

Proposal

Hivemall is a collection of machine learning algorithms and versatile data analytics functions. It provides a number of ease of use machine learning functionalities through user-defined function (UDF), user-defined aggregate function (UDAFs), and/or user-defined table generating functions (UDTFs) of Apache Hive. It offers a variety of functionalities: regression, classification, recommendation, anomaly detection, k-nearest neighbor, and feature engineering. Hivemall supports state-of-the-art machine learning algorithms such as Soft Confidence Weighted, Adaptive Regularization of Weight Vectors, Factorization Machines, and AdaDelta. Hivemall is mainly designed to run on Apache Hive but it also supports Apache Pig and Apache Spark for the runtime.

Background

Hivemall started as a research project of the main developer at National Institute of Advanced Industrial Science and Technology (AIST) in 2013 and the initial version was released on 2 Oct, 2013 on Github: https://github.com/myui/hivemall.

After the main developer moving to Treasure Data in 2015, the project has been actively developed as an open source product and changed the license from GNU LGPL v2.1 to Apache License v2 on Mar 16, 2015. The project copyright holders agreed to change the license then.

The community is growing incrementally and the project has 15 contributors, 431 stars, and 131 forks on Github as of Aug 15, 2016. The project was awarded for the InfoWorld Bossie Awards (the best open source big data tools) in 2014.

Past main contributions by external contributors includes Apache Pig supports from Daniel Dai (Hortonworks), Apache Spark porting and an integration to Apache YARN from Takeshi Yamamuro (NTT). Hivemall was originally designed for Apache Hive but it now supports Apache Spark and Apache Pig.

Rationale

User-defined function is a powerful mechanism to enrich the expressive power of declarative query languages like SQL, HiveQL, PigLatin, Spark SQL. Hive UDF interface is now becoming the de-facto standard for SQL-on-Hadoop platforms; Apache Spark and Apache Pig have full supports for Hive UDFs /UDAFs/UDTFs, and Apache Impala, Apache Drill, and Apache Tajo also have limited supports for Hive UDFs/UDAFs.

Hivemall can be considered as a cross platform library for machine learning as Hivemall is implemented as cross platform Hive UDFs/UDAFs/UDTFs; prediction models built by a batch query of Apache Hive can be used on Apache Spark/Pig, and conversely, prediction models build by Apache Spark can be used from Apache Hive/Pig.

Several database vendors are trying to offer machine learning functionality in relational databases, so that the costs of moving data can be eliminated. Apache MADlib, a machine learning library for HAWQ and PostgreSQL, is accepted as an Apache Incubator project. MADlib is implemented using PostgreSQL UDF interface.

Apache Hive has a JIRA ticket in HIVE-7940 to support machine learning functionalities. So, we consider this proposal is useful for the community. We consider that Hivemall is better to be a separated project to the Apache Hive because 1) we target other data processing frameworks such as Apache Spark as well for the runtime of Hivemall, and 2) the current codebase is large enough to be separated. Separation of concerns is good for project governance (e.g., release management). For example, Apache Datafu is data mining and statistics library for Apache Pig and a separated project to Apache Pig.

We consider that Hivemall would be a similar position to Apache Datafu but there are large differences in features and target runtimes. The target runtime of Apache Datafu is Apache Pig but Hivemall targets Apache Hive, Apache Spark, and Apache Pig for the target runtime. Apache Datafu is more likely to be statistics library and does not support machine learning features such as classification and regression but Hivemall is a machine learning library supporting them.

Initial Goals

The initial goals are as follows:

- Establish the project governance in the Apache way and broaden the community
- Improve documentations.
- Adding more unit/scenario tests.
- Handover of code and copyrights
    - get I-CLA from the initial committers
    - get SGA from other individuals not listed in the initial committers and AIST.
    - list all copyrights and licenses in NOTICE file and LICENSE file, respectively.

Current Status

Hivemall has several on-going WIP features.

Making a parameter server (a kind of distributed key-value store) as Apache YARN application is a major issue. Hivemall's parameter server is currently a standalone application. Parameter servers on Apache YARN enables to use Hadoop cluster resource efficiently and makes management of parameter servers easier.

Another major WIP issue is integrating XGBoost into Hivemall. We need more works and tests, e.g., supporting cross compilation of native JNI objects of XGBoost.

Meritocracy

The project members understand the importance of letting motivated individuals contribute to the project. Since Hivemall was initially released in 2014, it has received contributions from 14 contributors.

Our intent of this incubator proposal is building a diverse developer community following the Apache meritocracy model. We welcome external contributions and plan to elect committers from those who contribute significantly to the project.

Community

While there are 15 contributors in total, there are 3-4 active developers continuously involved for the major feature development at the moment. We hope to extend our contributor base and encourages suggestions and contributions from any potential user.

Core Developers

The current main developers are from employees of Treasure Data, NTT and Hortonworks. Some of them are Hadoop/Pig PMCs and/or Hive committers.

Alignment

Incubating at ASF is the natural choice for the Hivemall project because the Hivemall is targeting to run on Apache Hive, Apache Spark, and Apache Pig. We encourage integrations with other ASF data processing frameworks like Apache Impala and Apache Drill.

Known Risks

The contributions of the main developer is significant at the moment but the dependencies would decrease as the community grows.

Orphaned products

While the main developer is developing Hivemall as a full-time job at TreasureData, the company is well being aware of the open source philosophy and the importance of open governance of open source products. Orphanining ASF product can be considered itself as a risk. Hence, we think the the risks of it being orphaned are minimal.

Inexperience with Open Source

Hivemall also has been developed as an open source project since 2013. The majority of the project member have jobs developing open source products and some of them are working on other ASF projects like Apache Hadoop and Apache Pig. We thus considered that the project members have enough experiences for open source development.

Homogenous Developers

The current list of committers consists of developers from three different companies. The committers are geographically distributed across the U.S. and Asia. They are experienced with working in a distributed environment.

While not included in the initial committer, there are other external contributors to the project. So, we hope to establish a developer community that includes those contributors from several other corporations during the incubation process.

Reliance on Salaried Developers

The major developer is paid by his employer to contribute to this project and the other developers are payed by their employers for Hadoop-related open source development. While they might change their affiliations over time, they are willing to have their expertise for the open source development. So, the project would continue regardless their affiliations.

Relationships with Other Apache Products

Hivemall is a collection for machine learning functions on Apache Hive, Apache Spark, and Apache Pig. Apache MADlib is a collection of machine learning functions for relational databases, i.e., Apache HAWQ and PostgreSQL. There is no conflict in their target runtimes.

A Excessive Fascination with the Apache Brand

Our interest for this incubation is attracting more contributors, building a strong community with open governance, and increasing the visibility of Hivemall in the market/community. We will be sensitive to inadvertent abuse of the Apache brand for any commercial use and will work with the Incubator PMC and project mentors to ensure the brand policies are respected.

Documentation

Information on Hivemall can be found at: https://github.com/myui/hivemall/wiki

Initial Source

We released the initial version of Hivemall in 2013 at https://github.com/myui/hivemall and introduced Hivemall at the Hadoop Summit 2014.

Source and Intellectual Property Submission Plan

We know no legal encumberment to transfer of the source to Apache. We are going to get Contributor License Agreement (CLA) for all property of Hivemall.

Also, we plan to get a sign from AIST for Software Grant Agreement (SGA).

External Dependencies

Hivemall depends on the following third party libraries:

Core module:

- netty (The MIT License)
- smile (Apache License v2.0)
- org.takuaani.xz (Public Domain)
- xgboost (Apache License v2.0)
- hadoop (Apache License v2.0)
- hive (Apache License v2.0)
- log4j (Apache License v2.0)
- guava (Apache License v2.0)
- lucene-analyzers-kuromoji (Apache License v2.0)
- junit (Eclipse Public License v1.0)
- mockito (The MIT License)
- powermock (Apache License v2.0)
- kryo (BSD License)

Hivemall on Spark:

- spark (Apache License v2.0)
- commons-cli (Apache License v2.0)
- commons-logging (Apache License v2.0)
- commons-compress (Apache License v2.0)
- scala-library (BSD License)
- scalatest (Apache License v2.0)
- xerial-core (Apache License v2.0)

The dependencies all have Apache compatible licenses.

Cryptography

N/A

Required resources

Mailing lists

- private@hivemall.incubator.apache.org (with moderated subscriptions)
- commits@hivemall.incubator.apache.org
- dev@hivemall.incubator.apache.org
- user@hivemall.incubator.apache.org

Git Repository

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

JIRA assistance

JIRA project Hivemall (HIVEMALL)

Initial Committers

- Makoto Yui (myui@treasure-data.com)
- Takeshi Yamamuro (yamamuro.takshi@lab.ntt.co.jp)
- Daniel Dai (daijy@hortonworks.com)
- Tsuyoshi Ozawa (ozawa.tsuyoshi@lab.ntt.co.jp)
- Kai Sasaki (sasaki@treasure-data.com)

Affiliations

Treasure Data

- Makoto Yui
- Kai Sasaki

NTT

- Takeshi Yamamuro
- Tsuyoshi Ozawa Apache Hadoop PMC member

Hortonworks

- Daniel Dai (ASF member) Apache Pig PMC member

Sponsors

Champion

- Roman Shaposhnik (Pivotal, ASF member, IPMC member) Apache Bigtop/Incubator PMC member

Nominated Mentors

- Reynold Xin (Datariсks, ASF member) Apache Spark PMC member
- Markus Weimer (Microsoft, ASF member) Apache REEF PMC member
- Xiangrui Meng (Databricks, ASF member) Apache Spark PMC member

Sponsoring Entity

We are requesting the Incubator to sponsor this project.