HowlProposal

Abstract

Howl is a table and storage management service for data created using Apache Hadoop.

Proposal

The vision of Howl is to provide table management and storage management layers for Apache Hadoop. This includes:

- Providing a shared schema and data type mechanism.
- Providing a table abstraction so that users need not be concerned with where or how their data is stored.
- Providing interoperability across data processing tools such as Pig, Map Reduce, Streaming, and Hive.

Background

Data processors using Apache Hadoop have a common need for table management services. The goal of a table management service is to track data that exists in a Hadoop grid and present that data to users in a tabular format. Such a table management service needs to provide a single input and output format to users so that individual users need not be concerned with the storage formats that are chosen for particular data sets. As part of having a single format, the data will need to be described by one type of schema and have a single datatype system.

Additionally, users should be free to choose the best tools for their use cases. The Hadoop project includes Map Reduce, Streaming, Pig, and Hive, and additional tools exist such as Cascading. Each of these tools has users who prefer it, and there are use cases best addressed by each of these tools. Two users on the same grid who need to share data should not be constrained to use the same tool but rather should be free to choose the best tool for their use case. A table management service that presents data in the same way to all of the tools can alleviate this problem by providing interfaces to each of the data processing tools.

There are also a few other features a table management service should provide, such as notification of when data arrives.

A couple of developers at Yahoo! started the project. It is based on the Hive MetaStore component. There is good amount of interest in such a service expressed from Yahoo!, Facebook, LinkedIn, and, others. We are therefore proposing to place Howl in the Apache incubator and to build an open source community around it.

Rationale

There is a strong need for a table management service, especially for large grids with petabytes of data, and where the data volume is increasing by the day. Hadoop users need to find data to read and have a place to store their data. Currently users must understand the location of data to read, the storage format, compression techniques used, etc. To write data they need to understand where on HDFS their data belongs, the best compression format to use, how their data should be serialized, etc.

Most users do not want to be concerned with these issues. They want these managed for them.

Having it as an Apache Open Source project will highly benefit Howl from the point of view of getting a large community that currently uses Hadoop and the other products built around Hadoop (like Pig, Hive, etc.). Users of the Hadoop ecosystem can influence Howl’s roadmap, and contribute to it. Looking at it in another way, we believe having Howl as part of the Hadoop ecosystem will be a great benefit to the current Hadoop/Pig/Hive community too.

Current Status

Meritocracy

Our intent with this incubator proposal is to start building a diverse developer community around Howl following the Apache meritocracy model. We have wanted to make the project open source and encourage contributors from multiple organizations from the start. We plan to provide plenty of support to new developers and to quickly recruit those who make solid contributions to committer status.

Community

Howl is currently being used by developers at Yahoo! and there has been an expressed interest from LinkedIn and Facebook. Yahoo! also plans to deploy the current version of Howl in production soon. We hope to extend the user and developer base further in the future. The current developers and users are all interested in building a solid open source community around Howl.

To work towards an open source community, we have started using the GitHub issue tracker and mailing lists at Yahoo! for development discussions within our group.

Core Developers

Howl is currently being developed by four engineers from Yahoo! - Devaraj Das, Ashutosh Chauhan, Sushanth Sowmyan, and Mac Yang. All the engineers have deep expertise in Hadoop and the Hadoop Ecosystem in general.

Alignment

The ASF is a natural host for Howl given that it is already the home of Hadoop, Pig, HBase, Cassandra, and other emerging cloud software projects. Howl was designed to support Hadoop from the beginning in order to solve data management challenges in Hadoop clusters. Howl complements the existing Apache cloud computing projects by providing a unified way to manage data.

Known Risks

Orphaned Products

The core developers plan to work full time on the project. There is very little risk of Howl getting orphaned since large companies like Yahoo! are planning to deploy this in their production Hadoop clusters. We believe we can build an active developer community around Howl (companies like Facebook and LinkedIn have also expressed interest).

Inexperience with Open Source

All of the core developers are active users and followers of open source. Devaraj Das is an Apache Hadoop committer and Apache Hadoop PMC member, and has experience with the Apache infrastructure and development process. Ashutosh Chauhan is an Apache Pig committer and Apache Pig PMC member. Sushanth Sowmyan and Mac Yang made contributions to the Apache Hive and the Apache Chukwa projects.

Homogeneous Developers

The current core developers are all from Yahoo! However, we hope to establish a developer community that includes contributors from several corporations, and we are starting to work towards this with Facebook and LinkedIn.

Reliance on Salaried Developers

Currently, the developers are paid to do work on Howl. However, once the project has a community built around it, we expect to get committers and developers from outside the current core developers. Companies like Yahoo! are invested in Howl being a solution to the data management problem in Hadoop clusters, and that is not likely to change.

Relationships with Other Apache Products

Howl is going to be used by users of Hadoop, Pig, and Hive. See section Initial Source below for more information about Howl's relationship to Hive.

An Excessive Fascination with the Apache Brand

While we respect the reputation of the Apache brand and have no doubts that it will attract contributors and users, our interest is primarily to give Howl a solid home as an open source project following an established development model. We have also given reasons in the Rationale and Alignment sections.

Documentation

Information about Howl can be found at http://wiki.apache.org/pig/Howl. The following sources may be useful to start with:

- The GitHub site: https://github.com/yahoo/howl
- The roadmap: http://wiki.apache.org/pig/HowlJournal

Initial Source

Howl has been under development since Summer 2010 by a team of engineers in Yahoo!. It is currently hosted on GitHub under an Apache license at https://github.com/yahoo/howl.

The initial development of Howl has consisted of:

- maintaining a branch of the entire Hive codebase
- getting Howl-related patches committed to Hive
- developing Howl-specific plugins and wrappers to customize Hive behavior

At runtime, Howl executes Hive code for metastore and CLI+DDL, disabling anything related to Hadoop map/reduce execution. It also makes use of the RCFile storage format contained in Hive.

This approach was taken as a first step in order to validate the required functionality and get a production version working. However, in the long-term, maintaining a clone of Hive is undesirable. One possible resolution is to factor the metastore+CLI+DDL components out of Hive and move them into Howl (making Hive dependent on Howl). Another possible resolution is to remove the copy of Hive from Howl and do the build/release engineering necessary to make Howl depend on Hive. As part of the incubation process, we plan to work towards resolution of these issues.

External Dependencies

The dependencies all have Apache compatible licenses.

Cryptography

Not applicable.

Required Resources

Mailing Lists

- howl-private for private PMC discussions (with moderated subscriptions)
- howl-dev
- howl-commits
- howl-user

Subversion Directory

https://svn.apache.org/repos/asf/incubator/howl

Issue Tracking

JIRA Howl (HOWL)

Other Resources

The existing code already has unit tests, so we would like a Hudson instance to run them whenever a new patch is submitted. This can be added after project creation.

Initial Committers

- Devaraj Das
- Ashutosh Chauhan
- Sushanth Sowmyan
- Mac Yang
- Paul Yang
- Alan Gates

A CLA is already on file for Sushanth.

Affiliations

- Devaraj Das (Yahoo!)
- Ashutosh Chauhan (Yahoo!)
- Sushanth Sowmyan (Yahoo!)
- Mac Yang (Yahoo!)
- Paul Yang (Facebook)
- Alan Gates (Yahoo!)

Sponsors

Champion

Owen O’Malley

Nominated Mentors

- Olga Natkovich (Pig PMC member and Apache VP for Pig)
- Alan Gates (Pig PMC member)
- John Sichi (Hive PMC member)

Sponsoring Entity

We are requesting the Incubator to sponsor this project.