FalconProposal

Falcon Proposal

Abstract

Falcon is a data processing and management solution for Hadoop designed for data motion, coordination of data pipelines, lifecycle management, and data discovery. Falcon enables end consumers to quickly onboard their data and its associated processing and management tasks on Hadoop clusters.

Proposal

Falcon will enable easy data management via declarative mechanism for Hadoop. Users of Falcon platform simply define infrastructure endpoints, data sets and processing rules declaratively. These declarative configurations are expressed in such a way that the dependencies between these configured entities are explicitly described. This information about inter-dependencies between various entities allows Falcon to orchestrate and manage various data management functions.

The key use cases that Falcon addresses are:

- Data Motion
- Process orchestration and scheduling
- Policy-based Lifecycle Management
- Data Discovery
- Operability/Usability

With these features it is possible for users to onboard their data sets with a comprehensive and holistic understanding of how, when and where their data is managed across its lifecycle. Complex functions such as retrying failures, identifying possible SLA breaches or automated handling of input data changes are now simple directives. All the administrative functions and user level functions are available via RESTful APIs. CLI is simply a wrapper over the RESTful APIs.

Background

Hadoop and its ecosystem of products have made storing and processing massive amounts of data commonplace. This has enabled numerous organizations to gain valuable insights that they never could have achieved in the past. While it is easy to leverage Hadoop for crunching large volumes of data, organizing data, managing life cycle of data and processing data is fairly involved. This is solved adequately well in a classic data platform involving data warehouses and standard ETL (extract-transform-load) tools, but remains largely unsolved today. In addition to data processing complexities, Hadoop presents new sets of challenges and opportunities relating to management of data.

Data Management on Hadoop encompasses data motion, process orchestration, lifecycle management, data discovery, etc. among other concerns that are beyond ETL. Falcon is a new data processing and management platform for Hadoop that solves this problem and creates additional opportunities by building on existing components within the Hadoop ecosystem (ex. Apache Oozie, Apache Hadoop DistCp etc.) without reinventing the wheel. Falcon has been in production at InMobi, going on its second year and has been managing hundreds of feeds and processes.

Falcon is being developed by engineers employed with InMobi and Hortonworks. This platform addition will increase the adoption of Apache Hadoop by driving data management tractable for end users. We are therefore proposing to make Falcon an Apache open source project.

Rationale

The Falcon project aims to improve the usability of Apache Hadoop. As a result Apache Hadoop will grow its community of users by increasing the places Hadoop can be utilized and the use cases it will solve. By developing Falcon in Apache we hope to gather a diverse community of contributors, helping to ensure that Falcon is deployable for a broad range of scenarios. Members of the Hadoop development community will be able to influence Falcon's roadmap, and contribute to it. We believe having Falcon as part of the Apache Hadoop ecosystem will be a great benefit to all of Hadoop's users.

Current Status

Falcon is widely deployed in production within InMobi and moving on to its second year. A version with a valuable set of features is developed by the list of initial committers and is hosted on github.

Meritocracy

Our intent with this incubator proposal is to start building a diverse developer community around Falcon 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

We are happy to report that the initial team already represents multiple organizations. We hope to extend the user and developer base further in the future and build a solid open source community around Falcon.

Core Developers

Falcon is currently being developed by three engineers from InMobi – Srikanth Sunderrajan, Shwetha G S, and Shaik Idris, two Hortonworks employees – Sanjay Radia and Venkatesh Seetharam. In addition, Rohini Palaniswamy and Thiruvel Thirumoolan, were also involved in the initial design discussions. Srikanth, Shwetha and Shaik are the original developers. All the engineers have built two generations of Data Management on Hadoop, having deep expertise in Hadoop and are quite familiar with the Hadoop Ecosystem. Samarth Gupta & Rishu Mehrothra, both from InMobi have build the QA automation for Falcon.

Alignment

The ASF is a natural host for Falcon given that it is already the home of Hadoop, Pig, Knox, HCatalog, and other emerging “big data” software projects. Falcon has been designed to solve the data management challenges and opportunities of the Hadoop ecosystem family of products. Falcon fills the gap that Hadoop ecosystem has been lacking in the areas of data processing and data lifecycle management.

Known Risks

Orphaned products & Reliance on Salaried Developers

The core developers plan to work full time on the project. There is very little risk of Falcon getting orphaned. Falcon is in use by companies we work for so the companies have an interest in its continued vitality.

Inexperience with Open Source

All of the core developers are active users and followers of open source. Srikanth Sundarrajan has been contributing patches to Apache Hadoop and Apache Oozie, Shwetha GS has been contributing patches to Apache Oozie. Seetharam Venkatesh is a committer on Apache Knox. Sharad Agarwal, Amareshwari SR (also a Apache Hive PMC member) and Sanjay Radia are PMC members on Apache Hadoop.

Homogeneous Developers

The current core developers are from diverse set of organizations such as InMobi and Hortonworks. We expect to quickly establish a developer community that includes contributors from several corporations post incubation.

Reliance on Salaried Developers

Currently, most developers are paid to do work on Falcon but few are contributing in their spare time. However, once the project has a community built around it post incubation, we expect to get committers and developers from outside the current core developers.

Relationships with Other Apache Products

Falcon is going to be used by the users of Hadoop and the Hadoop ecosystem in general.

A 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 Falcon 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

http://wiki.apache.org/incubator/FalconProposal

Initial Source

The source is currently in github repository at: https://github.com/sriksun/ivory

Source and Intellectual Property Submission Plan

The complete Falcon code is under Apache Software License 2.

External Dependencies

The dependencies all have Apache compatible licenses. These include BSD, MIT licensed dependencies.

Cryptography

None

Required Resources

Mailing lists

- falcon-dev AT incubator DOT apache DOT org
- falcon-commits AT incubator DOT apache DOT org
- falcon-user AT incubator apache DOT org
- falcon-private AT incubator DOT apache DOT org

Subversion Directory

Git is the preferred source control system: git://git.apache.org/falcon

Issue Tracking

JIRA FALCON

Initial Committers

- Srikanth Sundarrajan (Srikanth.Sundarrajan AT inmobi DOT com)
- Shwetha GS (shwetha.gs AT inmobi DOT com)
- Shaik Idris (shaik.idris AT inmobi DOT com)
- Venkatesh Seetharam (Venkatesh AT apache DOT org)
- Sanjay Radia (sanjay AT apache DOT org)
- Sharad Agarwal (sharad AT apache DOT org)
- Amareshwari SR (amareshwari AT apache DOT org)
- Samarth Gupta (samarth.gupta AT inmobi DOT com)
- Rishu Mehrothra (rishu.mehrothra AT inmobi DOT com)

Affiliations

- Srikanth Sundarrajan (InMobi)
- Shwetha GS (InMobi)
- Shaik Idris (InMobi)
- Venkatesh Seetharam (Hortonworks Inc.)
- Sanjay Radia (Hortonworks Inc.)
- Sharad Agarwal (InMobi)
- Amareshwari SR (InMobi)
- Samarth Gupta (InMobi)
- Rishu Mehrothra (InMobi)

Sponsors

Champion

- Arun C Murthy (acmurthy at apache dot org)

Nominated Mentors

- Alan Gates (gates AT apache DOT org)
- Chris Douglas (cdouglas AT apache DOT org)
- Devaraj Das (ddas AT apache DOT org)
- Owen O'Malley (omalley AT apache DOT org)

Sponsoring Entity

Incubator PMC