AtlasProposal

Apache Atlas Proposal

Abstract

Apache Atlas is a scalable and extensible set of core foundational governance services that enables enterprises to effectively and efficiently meet their compliance requirements within Hadoop and allows integration with the complete enterprise data ecosystem.

Proposal

Apache Atlas allows agnostic governance visibility into Hadoop, these abilities are enabled through a set of core foundational services powered by a flexible metadata repository.

These services include:

- Search and Lineage for datasets
- Metadata driven data access control
- Indexed and Searchable Centralized Auditing operational Events
- Data lifecycle management – ingestion to disposition
- Metadata interchange with other metadata tools

Background

Hadoop is one of many platforms in the modern enterprise data ecosystem and requires governance controls commensurate with this reality.

Currently, there is no easy or complete way to provide comprehensive visibility and control into Hadoop audit, lineage, and security for workflows that require Hadoop and non-Hadoop processing.

Many solutions are usually point based, and require a monolithic application workflow. Multi-tenancy and concurrency are problematic as these offerings are not aware of activity outside of their narrow focus.

As Hadoop gains greater popularity, governance concerns will become increasingly vital to increasing maturity and furthering adoption. It is a particular barrier to expanding enterprise data under management.

Rationale

Atlas will address issues previously discussed by providing governance capabilities in Hadoop – using both a prescriptive and forensic model enriched by business taxonomical metadata. Atlas, at its core, is designed to exchange metadata with other tools and processes within and outside of the Hadoop stack – enable governance controls that are truly platform agnostic and effectively (and defensibly) address compliance concerns.

Initially working with a group of leading partners in several industries, Atlas is built to solve specific real world governance problems that accelerate product maturity and time to value.

Atlas aims to grow a community to help build a widely adopted pattern for governance, metadata modeling and exchange in Hadoop – which will advance the interests for the whole community.

Current Status

An initial 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 proposal is to start building a diverse developer community around Atlas 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 Atlas.

Core Developers

Atlas development is currently being led by engineers from Hortonworks – Harish Butani, Venkatesh Seetharam, Shwetha G S, and Jon Maron. All the engineers have deep expertise in Hadoop and are quite familiar with the Hadoop Ecosystem.

Alignment

The ASF is a natural host for Atlas given that it is already the home of Hadoop, Falcon, Hive, Pig, Oozie, Knox, Ranger, and other emerging “big data” software projects.

Atlas has been designed to solve the data governance challenges and opportunities of the Hadoop ecosystem family of products as well as integration to the tradition Enterprise Data ecosystem.

Atlas fills the gap that the Hadoop Ecosystem has been lacking in the areas of data governance and compliance 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 Atlas getting orphaned. A prototype of Atlas is in use and being actively developed by several companies and have vested interest in its continued vitality and adoption.

Inexperience with Open Source

Many of the core developers are PMC and committers of Apache. Harish Butani is PMC Apache Hive, Venkatesh Seetharam is PMC on Apache Falcon and Apache Knox, Shwetha GS is PMC on Apache Falcon and Apache Oozie committer.

Homogeneous Developers

The current core developers are from diverse set of organizations such as Hortonworks, Aetna, JPMC, Merck, SAS, Schlumberger and Target. We expect to quickly establish a developer community that includes contributors from additional organizations post incubation.

Reliance on Salaried Developers

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

Relationships with Other Apache Products

Atlas is going to be used by the users of Apache Hadoop and the Hadoop ecosystem in general – particularly with Apache Falcon and Apache Ranger for rationalizing data lifecycle and security policies respectively.

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 Atlas 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://people.apache.org/~venkatesh/atlas/

Initial Source

The source is currently hosted at: http://people.apache.org/~venkatesh/atlas/

Source and Intellectual Property Submission Plan

The complete Atlas 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

- atlas-dev AT incubator DOT apache DOT org
- atlas-commits AT incubator DOT apache DOT org
- atlas-private AT incubator DOT apache DOT org

Subversion Directory

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

Issue Tracking

JIRA Atlas

Initial Committers

- Venkatesh Seetharam (venkatesh AT apache DOT org)
- Harish Butani (rhbutani AT apache DOT org)
- Shwetha Shivalingamurthy (shwethags AT apache DOT org)
- Jon Maron (jmaron AT hortonworks DOT com)
- Andrew Ahn (aahn AT hortonworks DOT com)
- David Kaspar (david DOT kaspar AT merck DOT com)
- Ivo Lasek (ivo DOT lasek AT merck DOT com)
- Dennis Fusaro (fusarod AT aetna DOT com)
- Chris Hyzer (hyzerc AT aetna DOT com)
- Daniel Markwat (markwatd AT aetna DOT com)
- Greg Senia (seniag AT aetna DOT com)
- James Vollmer (james DOT vollmer AT target DOT com)
- Aaron Dossett (aaron DOT dossett AT target DOT com)
- Mitch Schussler(Mitch DOT Schussler AT jpmorgan DOT com)
- Viswanath Avasarala (VAvasarala AT SLB dot com)
- Anil Varma (AVarma AT SLB dot com)
- Barbara Stortz (Barbara DOT stortz AT sap DOT com)
- Srikanth Sundarrajan (sriksun AT apache DOT org)
- Suresh Srinivas (suresh AT hortonworks DOT org)
- Venkat Ranganathan (vranganathan AT hortonworks DOT com)

Affiliations

- Venkatesh Seetharam (Hortonworks)
- Harish Butani (Hortonworks)
- Swetha Shivalingamurthy (Hortonworks)
- Jon Maron (Hortonworks)
- Andrew Ahn (Hortonworks)
- David Kasper (Merck)
- Ivo Lasek (Merck)
- Dennis Fusaro (Aetna)
- Chris Hyzer (Aetna)
- Daniel Markwat (Aetna)
- Greg Senia (Aetna)
- James Vollmer (Target)
- Aaron Dossett (Target)
- Schussler, Mitch (JPMC)
- Viswanath Avasarala (Schlumberger)
- Anil Varma (Schlumberger)
- Barbara Stortz (SAP)
- Srikanth Sundarrajan (InMobi)
- Suresh Srinivas (Hortonworks)
- Venkat Ranganathan (Hortonworks)

Sponsors

Champion

- Jitendra Nath Pandey (jitendra AT apache DOT org)

Nominated Mentors

- Arun Murthy (acmurthy AT apache DOT org)
- Chris Douglas (cdouglas AT apache DOT org)
- Jakob Homan (jghoman AT apache DOT org)
- Vinod Kumar Vavilapalli (vinodkv AT apache DOT org)

Sponsoring Entity

Incubator PMC