OlympianProposal

Apache Proposal Structure

Abstract

Olympian (formerly Titan) is software designed to support the processing of graphs so large that they require storage and computational capacities beyond what a single machine can provide. Scaling graph data processing for real time traversals and analytical queries is Olympian's main benefit.

Proposal

Olympian consists of about 75K of Java code under the Apache 2 license. It supports very large graphs, with many concurrent transactions and operational graph processing. Olympian graphs scale with the number of machines in the cluster. Olympian already integrates with a number of Apache projects:
Provides native support for the popular property graph data model exposed by Apache TinkerPop. Provides native support for the Gremlin graph traversal language defined by Apache TinkerPop for programming language agnostic connectivity. Provides graph persistence solutions with:
Apache Cassandra Apache HBase Provides advanced indexing with:
Apache Lucene Apache Solr Supports global graph analytics and batch graph processing through the Apache Hadoop framework with processors implemented with:
Apache Spark Apache Giraph

Other software Olympian interfaces with includes:
BerkeleyDB Elasticsearch

Background

Marko Rodriguez and Matthias Broecheler, cofounders of the Aurelius graph consulting rm, developed the Titan distributed graph database system and made it available under the Apache 2 license in 2012. Marko is also a cofounder of the Apache TinkerPop project and the primary developer of the Gremlin graph traversal language. Other developers of Titan include Dan LaRocque, Stephen Mallette, Daniel Kuppitz, and Pavel Yaskevich. Datastax acquired Aurelius in February 2015, prior to the Titan 1.0 release in September 2015.

Since Titan became available on GitHub, there have been 4434 commits, 38 branches, 23 releases, and 35 contributors. In 2016 there has been less activity as the original authors are busy with other software development, but there is significant interest from the community.

Rationale

1. There are a number of Apache projects that integrate with Titan. 2. Apache Atlas (incubating) packages and ships Titan as an essential component, yet Titan is not part of Apache. 3. There are a number of existing users of Titan who are keen to continue to develop the code. These users provide the basis of the community for the proposed project.

Initial Goals

The initial goals are as follows:
Establish the project governance in The Apache Way and broaden the community. Distribute an incubating release aligned with the latest Apache TinkerPop version and prepared in accordance with the Apache release process. Improve the documentation. Add more unit/scenario tests. Contribute functional and performance-related enhancements to the code.

Current Status

The project will be forked off the existing Titan code base. This code has been available under the Apache 2 License but has not been subject to the Apache governance. The proposed project will adhere to Apache's governance and processes. This is one of the key benefits and reasons for bringing the project forward as an incubator candidate.

There are 37 pull requests currently open against Titan, and the last pull request was merged in June 2016. During incubation, the community will adopt a voting-based approach to review and commit those changes into the code base in preparation for the first incubating release.

Meritocracy

The proposed project will adopt the familiar process of progression from submitter to contributor to PMC. The community includes active committers and PMC members on other Apache projects (e.g. Apache TinkerPop, Apache Atlas (incubating), Apache HBase).

Community

There is an active and passionate community of existing Titan users. It is believed that this community will continue to grow and to progress. Titan is well-designed to support different backends, and the community will naturally grow as more backends are written to fit into the Titan architecture. Since the Titan 1.0 release, 3 different storage providers have become available. Also once an incubation release is made available, the community will likely see quick adoption from the Apache TinkerPop user base.

Core Developers

The community includes developers from a number of vendors (e.g. Google, HortonWorks, IBM, Mindmaps, Classmethod) and users (both academic and commercial). It contains two active committers and PMC members from the Apache TinkerPop project, one active committer and PPMC member from Apache Atlas (incubating), and one committer from Apache HBase. The developers represent a good mixture of skills, including expertise with each of the supported providers.

Alignment

The proposed project will be used by or integrated with a number of other Apache components, including (probably) TinkerPop, Atlas, Hadoop, Spark, Cassandra, and HBase. It is logical that the project should also be homed within Apache and subject to the governance principles of Apache.

Known Risks

Orphaned products

All the companies and developers associated with academic institutions who are engaged or want to be engaged with Titan are well aware of the open source philosophy and the importance of open governance of open source products. Hence, we think the risks of Titan being orphaned are minimal.

Inexperience with Open Source

The project is based on an existing open source code base (Titan 1.0) and the community consists of developers and vendors who have a history and strategy of open development and governance. The initial committers include committers and PMC members from other Apache projects.

Homogenous Developers

The community consists of geographically-dispersed volunteers from academic and a range of commercial organisations. The geographic diversity includes North America, Europe, Asia, and Australia.

Reliance on Salaried Developers

Many of the developers are salaried by the vendors in the community, but the vendors have publicly stated their support for open systems and whilst we might expect to see some gradual replacement of members of the community, we believe that it will remain stable and viable into the future. All members of the community are passionate about the project and are likely to contribute outside of 'normal working hours'.

Relationships with Other Apache Products

The proposed project has dependencies on other Apache projects, including Cassandra and HBase, for example. There are Apache projects that depend upon the availability of an open, scalable graph database. Apache Atlas is an example of such a project. Apache S2Graph (incubating) is currently an incubator project at Apache, however it does not currently implement the Apache TinkerPop interfaces, although it has an open JIRA for that effort.

An Excessive Fascination with the Apache Brand

Whilst the Apache brand will help to attract developers and consumers to the project, it is not for this reason that the proposal is being made. It is to align the governance of the project with that of the other components with which it is commonly used and to benefit from the development principles adopted by Apache. In particular, TinkerPop is Titan's most critical component/dependency, one so tight that Titan releases are contemporaneous or follow TinkerPop releases.

Documentation

Information on the existing Titan code base can be found at: http://titan.thinkaurelius.com/

Initial Source

The initial source will be based off a fork of the Titan code base. The latter can be found at: https://github.com/thinkaurelius/titan. The fork to be used as the base is from: https://github.com/pluradj/titan

Source and Intellectual Property Submission Plan

Since Datastax owns the copyright and trademark for Titan, when the proposal is accepted to the ASF Incubator, the community will choose a different name. It is proposed that Titan will enter incubation with the name Olympian. The community will finalize and document the name research during incubation. Individuals in the community have discussed the possibility of a software grant from Datastax, but Datastax was not interested in donating code or brand to the ASF. When asked if they would block others taking it to Apache they did not respond.

External Dependencies

Titan has the following external dependencies:

- Java 1.8
- Apache Maven 3.0.5 (Apache 2.0 License)
- JUnit 4.12 (EPL)
- MRUnit 1.1.0 (Apache 2.0 License)
- Apache Cassandra (Apache 2.0 License)
- Jamm (Apache 2.0 License)
- Metrics 2.1.1 and 3.0.1 (Apache 2.0 License)
- Sesame 2.7.10 (Eclipse Public License Version 1.0)
- slf4j 1.7.5 (MIT)
- Apache HTTPComponents 4.4.1 (Apache 2.0 License)
- Apache Hadoop 1.2.1 & 2.7.1 (Apache 2.0 License)
- Apache HBase (Apache 2.0 License)
- Jackson 1.9.2 & 2.4.4 (Apache 2.0 License)
- Apache Lucene 4.10.4 (Apache 2.0 License)
- Elasticsearch 1.5.1 (Apache 2.0 License)
- Apache Commons Beanutils 1.7.0 (Apache 2.0 License)
- Joda Time 1.6.2 (Apache 2.0 License)
- Google ConcurrentLinkedHashMap (Apache 2.0 License)
- Antlr 2.7.7 And 3.2 (BSD License)
- ASM 3 & 4 (http://asm.ow2.org/license.html)
- Apache Zookeeper 3.4.6 (Apache 2.0 License)
- Jersey 1.9 (CDDL 1.1 and GPL v2)
- JNA 4.0.0 (LGPL 2.1 and Apache 2.0 License)
- Kuali Maven s3 Wagon 1.1.20 (Educational Community License, Version 2.0)
- Apache Tomcat Jasper 5.5.23 (Apache 2.0 License)
- Berkeley DB 5.0.73 (Sleepycat License)

Upon acceptance to the incubator, we would begin a thorough analysis of all transitive dependencies to verify this information and introduce license checking into the build and release process by integrating with Apache Rat. In the case where a dependency has an Apache incompatible license, such as Berkeley DB, we will remove or replace it with an appropriate alternative.

Cryptography

Titan will support encryption of client-server communication through its use of the Apache TinkerPop Gremlin Server. We do not expect Titan to be a controlled export due to its use of encryption.

Required resources

Mailing lists

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

Git Repository

The team would like to use git for source control. We request a writable git repo https://git-wip-us.apache.org/repos/asf/incubator-olympian.git, and mirroring to be set up to GitHub through INFRA. We also request configuration for continuous integration with Travis CI.

Issue Tracking

Titan currently uses the GitHub issue tracker and the team would like to migrate all of these issues to the Apache JIRA.

Initial Committers

- Dylan Bethune-Waddell - dylan.bethune.waddell@mail.utoronto.ca
- Mathias Bogaert - mathias.bogaert@gmail.com
- Misha Brukman - mbrukman@google.com
- Felix Chapman - felix@mindmaps.io
- Sheldon Hall - sheldon@mindmaps.io
- Jing Chen (Jerry) He - jerryjch@apache.org
- Madhan Neethiraj - mneethiraj@hortonworks.com
- Alexander Patrikalakis - amcp@me.com
- Jason Plurad - pluradj@apache.org
- Suma Shivaprasad - sumasai@apache.org
- Lindsay Smith - lindsaysmith@google.com
- Filipe Teixeira - fppintoteixeira@gmail.com
- Ted Wilmes - twilmes@apache.org

Affiliations

- Dylan Bethune-Waddell - Jurisica Lab, Princess Margaret Cancer Centre, UHN
- Mathias Bogaert - Independent Contractor
- Misha Brukman - Google
- Felix Chapman - Mindmaps
- Sheldon Hall - Mindmaps
- Jing Chen (Jerry) He - IBM
- Madhan Neethiraj - Hortonworks
- Alexander Patrikalakis - Classmethod, Inc.
- Jason Plurad - IBM
- Suma Shivaprasad - Hortonworks
- Lindsay Smith - Google
- Filipe Teixeira - Mindmaps
- Ted Wilmes - Expero Inc.

Sponsors

Champion

Henry Saputra - hsaputra@apache.org

Nominated Mentors

- Alan Gates - gates@apache.org
- P. Taylor Goetz - ptgoetz@apache.org
- Henry Saputra - hsaputra@apache.org
- Michael Stack - stack@apache.org

Sponsoring Entity

The Apache Incubator