S2GraphProposal

S2Graph Proposal

Abstract

S2Graph is a distributed and scalable OLTP graph database built on Apache HBase to support fast traversal of extremely large graphs.

Proposal

S2Graph provides a scalable distributed graph database engine over a key/value store such as HBase. S2Graph provides a fully asynchronous API to manipulate data as a property graph model and fast breadth-first-search queries over the graph. S2Graph is designed for OLTP-like workloads on graph data sets instead of batch processing. Also, S2Graph provides INSERT/UPDATE operations. Its name 'S2Graph' is an abbreviated word of *S*uper *S*imple **Graph** Database.

Here are additional materials to introduce S2Graph.

- HBaseCon 2015 - http://www.slideshare.net/HBaseCon/use-cases-session-5
- Apache: Big Data 2015 - http://schd.ws/hosted_files/apachebigdata2015/06/s2graph_apache_con.pdf

Background

S2Graph initially started as an internal project at Kakao.com to efficiently store user relations and user activities as one large graph and to provide a unified query interface to traverse the graph. It was open sourced on Github about a 3 months ago in June 2015.

Over time, S2Graph using HBase as the storage tier has begun by adapted into various applications, such as messaging, social feeds, and realtime recommendations at Kakao.

Users can benefit by using S2Graph`s generalized high level graph abstraction API instead of querying via low-level key/value APIs, just as Apache Phoenix provides a SQL layer over HBase.

Rationale

Graph data (highly interconnected data) is very abundant and important these days. When users have a multitude of relationships, each with complex properties associated with them, a graph model is more intuitive and efficient than tabular formats (RDBMS).

There are many ASF projects that provide SQL tiers, but there is no ASF projects that provide a scalable graph layer on top of the existing hadoop ecosystem. When graph data grows to the trillion edge scale, the process of traversing takes a long time and can be costly. However, with the benefit of HBase`s scalable architecture, S2Graph can traverse large graphs in a breadth-first-search manner efficiently.

S2Graph also interoperates with several existing Apache projects (HBase, Apache Spark) to provide means of merging real time events and batch processed data using the property graph data model.

Many developers run their own domain specific API servers to serve their data products, but a graph model is general and the S2Graph API fully supports traversal of the graph, so it can be used as a scalable general purpose API serving layer for various domains. As long as data can be modeled as graph, then users can avoid tedious work developing customized API servers if they use S2Graph.

Initial Goals

The initial goals will be to move the existing codebase to Apache and integrate with the Apache development process. Once this is accomplished, we plan for incremental development and releases that follow the Apache guidelines.

Current Status

Meritocracy

S2Graph operated on meritocratic principles from the get go. Currently, all the discussions pertaining to S2Graph development are public on Github. The current incubation proposal includes the major code contributors to S2Graph. Several additional people have worked on the S2graph codebase for industry use cases and would be interested in becoming committers. We are starting with a small committer group and we plan to add additional committers following an open merit-based decision process during the incubation phase.

Community

We have already begun building a community but at this time the community consists only of S2Graph developers – all Kakao employees – and prospective users. S2Graph seeks to develop developer and user communities during incubation.

Core Developers

S2Graph is currently being designed and developed by 2 engineers from Kakao. - Doyung Yoon, Deawon Jeong.

Alignment

Our proposed S2Graph effort aligns closely with Apache HBase. The HBase project perimeter is denoted by a simple byte-array based Create, Read, Update, Delete and Scan API with no current plans to extend beyond these bounds.

S2Graph complements this with a higher level API for a property graph model.

S2Graph was designed to offer a scalable distributed graph database skin over HBase from the beginning in order to provide a property graph model and breadth first search, and will continue to focus on providing the graph model.

Known Risks

Orphaned Products

The core developers of S2Graph team plan to work full time on this project. There is very little risk of S2Graph getting orphaned since at least one large company (Kakao) is extensively using it in their production HBase clusters. For example, currently there are 20+ use cases with more than 1+Trillion edges and 140 million breadth first search query requests per minute using S2Graph in production. We plan to extend and diversify this community further through Apache.

Inexperience with Open Source

The core developers are all active users and followers of open source. They are already committers and contributors to the S2Graph Github project. All have been involved with the source code that has been released under an open source license. Though the core set of Developers do not have Apache Open Source experience, there are plans to onboard individuals with Apache open source experience to the project.

Homogenous Developers

Most committers in this proposal belong to the same institution (Kakao). The engagement of these committers goes well beyond the necessary development to support research, and all committers work on S2Graph full time. Several people from other institutions are working on and are familiar with the S2Graph codebase. We will work to attract them as future committers during the incubation phase, following a merit-based approach.

Reliance on Salaried Developers

Kakao invested in S2Graph as the distributed graph database solution on top of HBase and some of its key engineers are working full time on the project. We look forward to other Apache developers and researchers contributing to the project. Also key to addressing the risk associated with relying on Salaried developers from a single entity is to increase the diversity of the contributors and actively lobby for Domain experts in the graph database space to contribute. Apache S2Graph intends to do this.

Relationships with Other Apache Products

S2Graph has a strong relationship and dependency with Apache HBase and Apache Spark. Being part of Apache's Incubation community, could help with a closer collaboration among these two projects and as well as others.

In terms of graph processing frameworks, S2Graph and Apache Giraph look similar. However, their goals are apparently different to each other. Giraph aims at analytical batch processing on immutable graph data sets. In contrast, S2Graph is designed for OLTP-like workloads on graph data sets, and S2Graph provides INSERT/UPDATE operations too.

An Excessive Fascination with the Apache Brand

S2Graph is proposing to enter incubation at Apache in order to help efforts to diversify the committer-base, not so much to capitalize on the Apache brand. The S2Graph project is in production use already inside Kakao, but is not expected to be a Kakao product for external customers. As such, the S2Graph project is not seeking to use the Apache brand as a marketing tool.

Documentation

Information about S2Graph can be found at https://github.com/kakao/s2graph. The following links provide more information about S2Graph in open source:

- S2Graph web site: https://steamshon.gitbooks.io/s2graph-book/content/
- Codebase at Github: https://github.com/kakao/s2graph
- Issue Tracking: https://github.com/kakao/s2graph/issues
- User community: https://groups.google.com/forum/#!forum/s2graph

Initial Source

The S2Graph codebase is currently hosted on Github: https://github.com/kakao/s2graph.

Source and Intellectual Property Submission Plan

Currently, the S2Graph codebase is distributed under the Apache 2.0 License.

External Dependencies

Beyond relying on Apache HBase, S2Graph has the following external dependencies:

- Asynchbase (BSD)
- Play Framework (Apache 2.0 license)
- Scala (http://www.scala-lang.org/license.html)
- Spark (Apache 2.0 license)
- Kafka (Apache 2.0 license)

Required Resources

Mailing list

We will migrate our mailing lists to the following:

- users@s2graph.incubator.apache.org
- dev@s2graph.incubator.apache.org
- private@s2graph.incubator.apache.org
- commits@s2graph.incubator.apache.org

Source control

The S2Graph team would like to use Git for source code control, due to our current use of Git. We request a writeable Git repo for S2Graph, and mirroring to be set up to Github through INFRA.

Issue Tracking

S2Graph currently uses the github issue tracking system associated with its github repo (https://github.com/kakao/s2graph/issues). We will migrate to the Apache JIRA (http://issues.apache.org/jira/browse/S2Graph).

Other Resources

- Jenkins/Hudson for builds and test running.
- Wiki for documentation purposes.
- Blog to improve project dissemination.

Initial Committers

- Doyung Yoon <shom83 at gmail dot com>
- Daewon Jeong <blueiur at gmail dot com>
- Jaesang Kim <honeysleep at gmail dot com>
- Hwansung Yu <deejayfwan at gmail dot com>
- Min-Seok Kim <mskim.org at gmail dot com>
- Chul Kang <miralchul at gmail dot com>
- Luke Han <lukehan at apache dot org>
- Alexander Bezzubov <bzz at apache dot org>

Affiliations

- Doyung Yoon, Kakao
- Daewon Jeong, Kakao
- Jaesang Kim, Kakao
- Hwansung Yu, Kakao
- Min-Seok Kim, Kakao
- Chul Kang, Kakao,
- Luke Han, Ebay Inc.
- Alexander Bezzubov, NFLabs

Sponsors

Champion

Hyunsik Choi

Nominated Mentors

- Andrew Purtell - Apache Member, Salesforce
- Sergio Fernández - Apache Member, Redlink
- Hyunsik Choi - Apache Member, Gruter Inc.
- Seetharam Venkatesh - IPMC, Hortonworks Inc.

Sponsoring Entity

- The Apache Incubator