RyaProposal

Rya Proposal

Abstract

Rya (pronounced "ree-uh" /r/) is a cloud-based RDF triple store that supports SPARQL queries.

Proposal

Rya is a scalable RDF data management system built on top of Accumulo. Rya uses novel storage methods, indexing schemes, and query processing techniques that scale to billions of triples across multiple nodes. Rya provides fast and easy access to the data through SPARQL, a conventional query mechanism for RDF data.

Background

RDF is a World Wide Web Consortium (W3C) standard used in describing resources on the Web. The smallest data unit is a triple consisting of subject, predicate, and object. Using this framework, it is very easy to describe any resource, not just Web related. For example, if you want to say that Alice is a professor, you can represent this as an RDF triple like (Alice, rdf:type, Professor). In general, RDF is an open world framework that allows anyone to make any statement about any resource, which makes it a popular choice for expressing a large variety of data.

RDF is used in conjunction with the Web Ontology Language (OWL). OWL is a framework for describing models or ontologies for RDF. It defines concepts, relationships, and/or structure of RDF documents. These models can be used to 'reason/infer' information about entities within a given domain. For example, you can express that a Professor is a sub class of Faculty, (Professor, rdfs:subClassOf, Faculty) and knowing that (Alice, rdf:type, Professor), it can be inferred that (Alice, rdf:type, Faculty).

SPARQL is an RDF query language. Similar with SQL, SPARQL has SELECT and WHERE clauses; however, it is based on querying and retrieving RDF triples.

Work on Rya, a large scale distributed system for storing and querying RDF data, started in 2010.

Rationale

With the increase in data size, there is a need for scalable systems for storing and retrieving RDF data in a cluster of nodes. We believe that Rya can fulfill that role. We expect that communities within government, health care, finance, and others who generate large amounts of RDF data will be most interested in this project.

From its inception, the project operated with an Apache-style license, but it was open to mostly US government-related projects only. We believe that having the project and the development open for all will benefit both the project and the interested communities.

Current Status

The project source code and documentation are currently hosted in a private repository on Github. New users are added to the repository upon request.

Meritocracy

Meritocracy is the model that we currently follow, and we want to build a larger and more diverse developer community by becoming an Apache project.

Community

Rya has being building a community of users and developers for the past 3 years. There is currently an active workgroup with monthly meetings and the number of participants in the meeting is increasing.

Core Developers

The core developers are a diverse group of people who are either government employees or former / current government contractors from different companies.

Alignment

Rya is built on top of Accumulo, an Apache project.

Known Risks

Orphaned Products

There is a very small risk of becoming orphaned. The current contributors are strongly committed to the project, there is a large enough number of developers interested in contributing to the project, and we believe that the support for the project will continue to grow from the interested communities.

Inexperience with Open Source

The initial committers have various degrees of experience with open source projects - from very new to experienced. This project was open source within government from the beginning. We are aware that it will be different and more difficult functioning in a real open source environment. We are enthusiastic and committed to learning the Apache way and being successful in operating under Apache's development process.

Homogenous Developers

The current list of developers form a heterogeneous group, with people for academia, government, and industry, collaborating from distributed geographic locations. We aim to expand the list of contributors with the help of the Apache incubation process.

Reliance on Salaried Developers

Many but not all of the developers working on the project are salaried employees, paid to work on this project. They will continue to contribute to the open source project. Some of the initial committers continued as volunteers even if no longer employed to work on this project and they plan to continue supporting the project.

Relationships with Other Apache Products

Rya uses Apache Accumulo, Hadoop, Zookeeper, Maven.

*Apache Jena API or Apache Commons RDF API could become the RDF API used by Rya, but such a decision was not made.
*Apache Clerezza is database/triple store agnostic, and as such could be complementary to Rya.
*Apache Stanbol focuses on providing semantic services, while Rya focuses on providing a distributed triple store solution, with support for SPARQL and OWL reasoning.
*Apache Marmotta provides an implementation of a Linked Data Platform, and overlaps in some of the goals and functionality with Rya (RDF triple store, SPARQL support among others). There are many opportunities for collaboration with these projects and we are looking forward to such a collaboration.

Apache Brand

Rya has generated interest in the government. It also generated interest within academia and industry. We believe that everyone could benefit from having Rya as an open source project. Due to its strong ties to Accumulo, an Apache project, and due to the values of the Apache Foundation, we believe that Apache incubator is the right place for Rya.

Documentation

Two peer-reviewed publications [1,2] about Rya were published in 2012 and 2015. More documentation is available in the code.

[1] Roshan Punnoose, Adina Crainiceanu, David Rapp. Rya: A Scalable RDF Triple Store for the Clouds. Proceedings of the 1st International Workshop on Cloud Intelligence, Pages 4:1-4:8, August 2012

[2] Roshan Punnoose, Adina Crainiceanu, David Rapp. SPARQL in the Clouds Using Rya. Information Systems, Volume 48, Pages 181-195, March 2015 (Available online 23 July 2013)

Initial Source

The code is currently in a private Github repository, due to security and IP review processes. We intend to open it up via transferring the code to an ASF repository.

Source and Intellectual Property Submission Plan

The source code has been released under the Apache License, Version 2. Software grant, and CCLAs have been submitted. ICLAs for initial committers have been submitted or are in progress.

External Dependencies

- OpenRDF Sesame (BSD license)
- GeoMesa (Apache License, Version 2.0)
- Accumulo (Apache License, Version 2.0)
- Hadoop (Apache License, Version 2.0)
- Pig (Apache License, Version 2.0)
- TinkerPop (Apache License, Version 2.0)

Cryptography

The proposal does not involve any cryptographic code.

Required Resources

Mailing lists

- private@rya.incubator.apache.org
- dev@rya.incubator.apache.org
- commits@rya.incubator.apache.org

Git Repository

https://git-wip-us.apache.org/repos/asf/incubator-rya.git

Issue Tracking

JIRA Rya

Initial Committers

- Roshan Punnoose, roshanp at gmail dot com
- David Rapp, dnrapp at ncsu dot edu
- Adina Crainiceanu, adinancr at gmail dot com
- Aaron Mihalik, aaron.mihalik at gmail dot com
- Puja Valiyil, pujav65 at gmail dot com
- Jennifer Brown, jennifer.brown at parsons dot com
- Steve Wagner, steve.r.wagner at gmail dot com

Affiliations

- Roshan Punnoose, Enlighten IT Consulting
- David Rapp, North Carolina State University
- Adina Crainiceanu, US Naval Academy
- Aaron Mihalik, Parsons
- Puja Valiyil, Parsons
- Jennifer Brown, Parsons
- Steve Wagner, Enlighten IT Consulting

Sponsors

Champion

- Adam Fuchs, ASF Member, afuchs at apache dot org

Nominated Mentors

- Josh Elser josh dot elser at gmail dot com
- Edward J. Yoon edwardyoon at apache dot org
- Sean Busbey busbey at cloudera dot com
- Venkatesh Seetharam venkatesh at innerzeal dot com

We are seeking additional mentors

Sponsoring Entity

Apache Incubator