Cassandra

Abstract

Cassandra is a distributed storage system for managing structured/unstructured data while providing reliability at a massive scale.

Background

Development of Cassandra started in Facebook in June 2007. It started of a system to solve the Inbox Search problem and since then has matured to solve various storage problems associated with structured/unstructured data.

Rationale

Cassandra is a distributed storage system for managing structured data that is designed to scale to a very large size across many commodity servers, with no single point of failure. The philosophy behind the design of the storage portion of Cassandra is that it be able to satisfy the requirements of applications that demand storage of large amounts of structured data. Reliability at massive scale is a very big challenge. Outages in the service can have significant negative impact. Hence Cassandra aims to run on top of an infrastructure of hundreds of nodes (possibly spread across different datacenters). At this scale, small and large components fail continuously; the way Cassandra manages the persistent state in the face of these failures drives the reliability and scalability of the software systems relying on this service.

Initial Source

Intial Source can be obtained from the following site - http://the-cassandra-project.googlecode.com/svn/branches/development/. The mailing list is currently maintained at the same site. We will move it over to Apache once this proposal has been accepted.

Source and Intellectual Property Submission Plan

External Dependencies

- All dependencies have Apache compatible licenses. Dependencies are log4j, Thrift, Apache Commons.

Cryptography

- None

Committers

- Avinash Lakshman
- Prashant Malik
- Kannan Muthukkaruppan
- Jiansheng Huang
- Dan Dumitriu

Current Status

Meritocracy

- Though initial development was done at Facebook, Cassandra was intended to be released as an open source project from its inception. Environment will lend itself to support meritocracy at all times.

Community

- Folks who are actively considering deploying/prototyping Cassandra in their respective organizations.

Core Developers

- Avinash Lakshman
- Prashant Malik

- Kannan Muthukkaruppan

License

- The Cassandra codebase is Apache 2.0 licensed, and currently hosted at Google Code.

Known Risks/Avoiding the Warning Signs

Orphaned Products

- Cassandra is already deployed within Facebook and many other organizations are actively moving to deploy this in production. Original developers are and will actively stay involved and hence there is no realistic chance of it getting orphaned.

Homogenous Developers

- The current list of committers includes developers from different companies. The committers are geographically distributed across the U.S.

Reliance on Salaried Developers

- Yes. But don't expect this to be a risk of any nature.

Relationships with Other Apache Products

- The Cassandra project is 'similar' to hbase/HDFS in concept, but Cassandra is more geared for Online web site usage than batch. It also doesn't have a single point of failure, which makes it interesting as well.
- Cassandra makes use of the Thrift project.

An excessive fascination with the Apache brand

- Cassandra has already attracted a stable base of users. There are at least 3 companies who are planning to use Cassandra in production as far as we know. The reasons for joining Apache are not to advertise the project, but rather to demonstrate the commitment to open source by divorcing the trunk from any one corporation and pursuing further integration with other Apache projects.

Required Resources

Mailing lists

Once the project is approved, the following mailing lists will be used for discussion.

- cassandra-user@incubator.apache.org

Subversion Directory

*[WWW] https://svn.apache.org/repos/asf/incubator/cassandra

Issue Tracking

- JIRA Cassandra

Sponsors

Champion

- Brian McCallister

Mentors

- Torsten Curdt
- Brian McCallister

- Matthieu Riou
- Ian Holsman

Sponsoring Entity

- Incubator