GearpumpProposal

Gearpump Proposal

Abstract

Gearpump is a flexible, efficient and scalable micro-service based real-time big data streaming engine developed by Intel Corporation which has been licensed by Intel under the Apache License 2.0.

Proposal

Gearpump is a reactive real-time streaming engine; completely based on the micro-service Actor model. Gearpump provides extremely high performance stream processing while maintaining millisecond latency message delivery. It enables reusable, composable flows or partial graphs that can be remotely deployed and executed in a diverse set of environments, including IoT edge devices. These flows may be deployed and modified at runtime – a capability few real time streaming frameworks provide today.

The goal of this proposal is to incubate Gearpump as an Apache project in order to build a diverse, healthy, and self-governed open source community around this project.

Background

In past decade, there have been many advances within real-time streaming frameworks. Despite many advances, users of streaming frameworks often complain about flexibility, efficiency, and scalability. Gearpump endeavors to solve these challenges by adopting the micro-service Actor model. The Actor model was proposed by Carl Hewitt in 1973. In the Actor model, each actor is a message driven micro-service; actors are the basic building blocks of concurrent computation. By leveraging Actor Model's location transparency feature, Gearpump allows a graph to be composed of several partial graphs, where, for example, some parts may be deployed to remote IoT edge devices, and other parts to a data center. This division and deployment model can be changed at runtime to adapt to a changing physical environment, providing extreme flexibility and elasticity in solving various ingestion and analytics problems. We've found Actors to be a much smaller computation unit compared with threads, where smaller usually means better concurrency, and potentially better CPU utilization.

Rationale

Gearpump tightly integrates and enhances the big data community of Apache projects. Intel believes Gearpump can bring benefits to the Apache community in a number of ways:

1. Gearpump complements many existing Apache projects, in particular, those commonly found within the big data space. Users of this project are also users of other Apache projects, such as Hadoop ecosystem projects. It is beneficial to align these projects under the ASF umbrella. In real-time streaming, Gearpump offers some special features that are useful for Apache users, such as exactly-once processing with millisecond message level latency and dynamic DAGs that allow online topology modifications.

2. Gearpump tightly integrates with Apache big data projects. It supports for Apache HDFS, YARN, Kafka, and HBase. It uses Apache YARN for resource scheduling and Apache HDFS as the essential distributed storage system.

3. The micro-service model of reusable flows that Gearpump has adopted is very unique, and it may become common in the future. Gearpump sets a good example about how distributed software can be implemented within a micro-service model. An open project is of best interest to our users. By joining Apache, it will be a neutral infrastructure platform that will benefit everyone.

4. The process and development philosophy of Apache will help Gearpump grow, and build a diverse, healthy, and self-governed open source community.

Initial Goals

1. Migrate the existing codebase to Apache.

2. Setup Jira, website and other development tools by following Apache best practices.

3. Start the first release per Apache guidelines as soon as possible.

Current Status

Gearpump is hosted on Github. It has 1922 commits, 38284 line of code, and 31 major or minor releases, with release notes highlighting the changes for every release. It is licensed under Apache License Version 2. There is a documentation site at http://gearpump.io including a user guide, internal details, use cases and a roadmap. There is also an issue tracker where every code commit is tracked by a bug Id. Every pull request is reviewed by several reviewers and will only be merged based on consensus rule. These match Apache's development ideals.

Meritocracy

We think an open, fair, and renewing community culture is what we need and what our users require, that will protect everyone in the community. We would like the project to be free from potential undue influence from any single organization. We will invest in supporting a meritocratic model.

Community

Gearpump has a growing community with hundreds of stars on Github and an active WeChat group with hundreds of subscriptions. We organize regular offline meetup events. These efforts should help us to grow the community at Apache.

Core Developers

Most of the initial committers are Intel employees from China, the US, and Poland. We are committed to build a diverse community which involves more companies and individuals.

Alignment

Gearpump has good alignment with other Apache projects. Gearpump is tightly integrated with Apache Hadoop ecosystem. It uses Apache YARN for resource scheduling and Apache HDFS for storage. The unique streaming processing abilities Gearpump complements other Apache big data projects today. We believe there will be a synergistic effect by aligning Gearpump under the Apache umbrella.

Known Risks

Orphaned products

Intel has a long-term interest in big data and open source and a proven record of contributing to Apache projects. The risk of the Gearpump project being abandoned is very small. Besides, Intel is seeing an increasing interest in Gearpump from different organizations. We are committed to get more support, adoption, and code contribution from different companies.

Inexperience with Open Source

Gearpump is an existing project under the Apache License, Version 2.0 with a long history record of open development. Initial committers of this project have years of open sourcing contribution experiences, including code contribution to HDFS, HBase, Storm, YARN, Sqoop, and etc. Some of the initial committers are also committers to other Apache projects.

Homogeneous Developers

The current list of committers includes developers from different geographies and time zones; they are able to collaborate effectively in a geographically dispersed environment. We are committed to recruit more committers from different companies to get a more diverse mixture.

Reliance on Salaried Developers

Most of our current Gearpump developers are Intel employees who are contributing to this project. Our developers are passionate about this project and spend a lot of their own personal time on the project. We are confident that their interests will remain strong. We are committed to recruiting additional committers from the community as well.

Relationships with Other Apache Product

Gearpump codebase is closely integrated with Apache Hadoop, Apache HBase, and Apache Kafka. Gearpump also has some similarities with Apache Storm. Although Gearpump and Storm are both systems for real-time stream processing, they have fundamentally different architectures. In particular, Gearpump adopts the micro-service model, building on the Akka framework, for concurrency, isolation and error handling, which we believe is a future trend for building distributed software. We look forward to collaboration with other Apache communities.

An Excessive Fascination with the Apache Brand

The ASF has a strong brand; we appreciate that fact and will protect the brand. Gearpump is an existing open source project with many committers and years of effort. The reasons to join Apache are outlined in the Rationale section above.

Documentation

Information on Gearpump can be found at:
Gearpump website: http://gearpump.io Codebase: https://github.com/gearpump/gearpump

Initial Source and Intellectual Property Submission Plan

The Gearpump codebase is currently hosted on Github: https://github.com/gearpump/gearpump. We will use this codebase to migrate to the Apache foundation. The Gearpump source code is licensed under Apache License Version 2.0 and will be kept that way. All contributions on the project will be licensed directly to the Apache foundation through signed Individual Contributor License Agreements or Corporate Contributor License Agreements.

External Dependencies

All of Gearpump dependencies are distributed under Apache compatible licenses.

Gearpump leverages Akka which has Apache 2.0 licensing for current and planned versions http://doc.akka.io/docs/akka/2.3.12/project/licenses.html#Licenses_for_Dependency_Libraries

Cryptography

Gearpump does not include or utilize cryptographic code.

Required Resources

We request that following resources be created for the project to use

Mailing lists

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

Git repository

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

Documentation

https://gearpump.incubator.apache.org/docs/

JIRA instance

JIRA Gearpump (GEARPUMP) https://issues.apache.org/jira/browse/gearpump

Initial Committers

- Xiang Zhong <xiang dot zhong at intel dot com>
- Tianlun Zhang <tianlun dot zhang at intel dot com>
- Qian Xu <qian dot a dot xu at intel dot com>
- Huafeng Wang <huafeng dot wang at intel dot com>
- Kam Kasravi <kam dot d dot kasravi at intel dot com>
- Weihua Jiang <weihua dot jiang at intel dot com>
- Tomasz Targonski <tomasz dot targonski at intel dot com>
- Karol Brejna <karol dot brejna at intel dot com>
- Gang Wang <gang1 dot wang at intel dot com>
- Mark Chmarny <mark dot chmarny at intel dot com>
- Xinglang Wang <xingwang at ebay dot com >
- Lan Wang <lan dot wanglan at huawei dot com>
- Jianzhong Chen <jianzhong dot chen at cloudera dot com>
- Xuefu Zhang <xuefu at apache dot org>
- Rui Li <rui dot li at intel dot com>

Affiliations

- Xiang Zhong – Intel
- Tianlun Zhang – Intel
- Qian Xu – Intel
- Huafeng Wang – Intel
- Kam Kasravi – Intel
- Weihua Jiang – Intel
- Tomasz Targonski – Intel
- Karol Brejna – Intel
- Mark Chmarny – Intel
- Gang Wang – Intel
- Mark Chmarny – Intel
- Xinglang Wang – Ebay
- Lan Wang – Huawei
- Jianzhong Chen – Cloudera
- Xuefu Zhang – Cloudera
- Rui Li – Intel

Sponsors

Champion

Andrew Purtell <apurtell at apache dot org>

Nominated Mentors

- Andrew Purtell <apurtell at apache dot org>
- Jarek Jarcec Cecho <Jarcec at cloudera dot com>
- Todd Lipcon <todd at cloudera dot com>
- Xuefu Zhang <xuefu at apache dot org>
- Reynold Xin <rxin at databricks dot com>

Sponsoring Entity

Apache Incubator PMC