StratosphereProposal

Stratosphere

Abstract

Stratosphere is an open source system for parallel data analysis. Stratosphere deeply integrates MapReduce and database technologies to provide expressive and optimizable programming interfaces and at the same time efficient and scalable execution.

Proposal

Stratosphere is an open source system for expressive, declarative, fast, and efficient data analysis. Stratosphere combines the scalability and programming flexibility of distributed MapReduce-like platforms with the efficiency, out-of-core execution, and query optimization capabilities found in parallel databases.

Background

There is currently a need for general-purpose cluster computing platforms that are compatible with the Hadoop ecosystem, are more efficient, easier to use, and can support more applications than Hadoop MapReduce, but are not restricted to a specific data model and language (such as the relational model and a variant of SQL). Stratosphere fulfils these needs.

Stratosphere exposes expressive APIs in Java and Scala (conceptually similar to Spark, Cascading, Scalding) that allow arbitrary user-defined functions in the same language and data model that the program is written in. Stratosphere programs pass through a cost-based optimizer that finds the best execution path for these programs depending on the data and cluster characteristics. The design and implementation of Stratosphere is based on research that generalizes query optimizers in relational databases. Stratosphere has a distributed runtime that is architected upon the principles of parallel databases, providing true pipelining (a basis for stream processing) and efficient out-of-core algorithms for grouping, sorting, joining, and aggregating data. Stratosphere provides first-class support for iterative algorithms via a built-in iterate operator, covering Machine Learning and graph analysis use cases. It achieves performance similar to Apache Giraph without being a specialized graph processing system.

Stratosphere has undergone three major releases (v0.1, v0.2, v0.4) and some minor ones.

Rationale

Stratosphere started out in 2008 as a research project by the Technical University of Berlin, the Humboldt University of Berlin, and the Hasso Plattner Institute, and has received subsequent funding from the German Research Council, the European Institute of Innovation and Technology, the European Commision (under EU FP7 DOPA-296448), and industry.

The traction of Stratosphere has by far exceeded our initial expectations, and we are therefore seeking an organizational long-term home for Stratosphere beyond the University walls that will house and further encourage contributors from companies and other organizations that are interested in Stratosphere. We believe that the Apache Software Foundation is the ideal home for Stratosphere. Stratosphere integrates with several existing Apache projects, such as HDFS, YARN, HBase, and Avro. The team is familiar with the Apache processes and fully subscribes to the Apache mission. One of the proposing members is a long-time Apache contributor and PMC member.

Initial Goals

- Move the existing codebase to Apache
- Integrate with the Apache development process
- Ensure all dependencies are compliant with Apache License version 2.0
- Incremental development and releases per Apache guidelines

Current Status

Meritocracy

Stratosphere operated on meritocratic principles from the get go. The initial project proposal submitted to the German Research Council in 2008 stated that all code developed in the project will be released as open source under the Apache 2 license. Currently, all the discussions pertaining to Stratosphere development are public on GitHub and our mailing list. The current incubation proposal includes the major code contributors to Stratosphere. Several additional people have worked on the Stratosphere codebase for research prototypes and 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

Currently, the core of Stratosphere is developed at TU Berlin, mainly by the committers listed in this proposal. Additional people from several Universities and companies in Europe are working with Stratosphere and are interested in becoming committers to the project.

During the years, Stratosphere has been adopted as a platform for research and teaching in several Universities (TU Berlin, HU Berlin, HPI, RWTH, Inria, KTH, U. Trento, UCSD, and others), and it is currently witnessing its first industrial installations. We are seeing a rapidly growing interest in Stratosphere by both startups and large companies, as well as a growing community (our first Stratosphere Summit in November 2013 attracted over 80 participants). Stratosphere was recently accepted as a mentoring organization in Google Summer of Code 2014.

We believe that acceptance in the Apache Software Foundation will consolidate the current community under one organizational umbrella, and most importantly accelerate the growth of the community.

Core developers

The core developers of the system are Stephan Ewen, Fabian Hueske, Daniel Warneke, Robert Metzger, Ufuk Celebi, and Aljoscha Krettek, who are all committers in the current proposal.

Alignment

Stratosphere is compatible with, and related to several Apache projects. Stratosphere re-uses parts of Apache Hadoop, in particular HDFS and YARN, as well as Apache HBase and Apache Avro. Stratosphere is a very good compilation target for query languages such as Apache Hive and Apache Pig.

Known Risks

Orphaned Products

There is strong interest in Stratosphere by several companies and organizations, and there is currently a long-term commitment to fund salaried developers for Stratosphere by public and private organizations in Europe.

Inexperience with Open Source

Sebastian Schelter is a committer and PMC member of Apache Mahout and Apache Giraph, member of the Apache Software Foundation, member of the Incubator PMC and project mentor for Apache Drill. Sebastian, along with our mentors, will guide the rest of the committers that have experience with releasing software as open source but little experience in participating in an open source project besides Stratosphere itself.

In mid-2013 Stratosphere transitioned from an “open source project with publicly accessible source code” to an open source project that puts the community first. We moved from a University-hosted git repository to GitHub, where we discuss all issues publicly. This also includes release planning (via GitHub’s milestone feature) and code reviews. We also moved our build system to the publicly available Travis-CI. The mailing lists are hosted with Google Groups, we use the public Maven repository infrastructure of Sonatype. The source code of the www.stratosphere.eu website is publicly available and is meant to be changed by external contributors (for example for documentation purposes).

Homogeneous Developers

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

Reliance on Salaried Developers

Currently, Stratosphere receives support from salaried developers, in particular from graduate students at TU Berlin that are funded by the German Research Council, the EIT ICT Labs, and the European Commission. These students work in their free time on Stratosphere in addition to their employment.

We expect that Stratosphere development will occur on both salaried and volunteer time. We will recruit additional committers, including non-salaried developers, and we will work to ensure that the project will move forward independently of salaried developers.

Relationship with Other Apache Products

Stratosphere interfaces with several existing Apache projects: Apache HBase for storage, Apache Hadoop (HDFS for storage, YARN for resource management, and Stratosphere contains a generic wrapper for Hadoop MapReduce input formats), and Apache Avro (for serialization). Stratosphere uses Apache Maven and Apache Commons libraries internally. Stratosphere can be a great compilation target for Apache Pig and Apache Hive, although such functionality is not yet implemented.

Stratosphere is also related with several projects undergoing incubation in the Apache Incubation project, such as Tez, Drill, and Spark (graduated). While all these projects target sufficiently different spaces and have different architectures, it would be interesting to explore code reuse possibilities. For example, we are currently basing our design for compiling SQL to Stratosphere on the Optiq library, also used by Apache Drill.

An Excessive Fascination with the Apache Brand

We believe that the Apache brand will help us attract contributors to Stratosphere, by giving us a well-defined, transparent development process under a known brand. At the same time, Stratosphere already has a healthy community and current funding guarantees the further codebase development and growth of the project for the next 3-5 years. The reason for this proposal is not to gain publicity, but to further strengthen the longevity of the project as explained in the Rationale section.

Documentation

- Project website
- Documentation
- Codebase
- Mailing list

Initial Source

Stratosphere is hosted on GitHub . This is the codebase that we will migrate to the Apache Foundation. The code was previously hosted on a TU Berlin's own git infrastructure. It has always been Apache 2.0 licensed.

Source and Intellectual Property Submission Plan

All initial and past committers will sign a CLA with the ASF while the incubator proposal for Stratosphere is being discussed. All organizations that have employed Stratosphere contributors in the past will sign a SGA. Current contributors will sign a CCLA. All major contributors are still active in the project.

External Dependencies

All critical dependencies are, to the extend of our knowledge, from other Apache projects. These include Apache Hadoop (for YARN and HDFS) and some libraries (log4j, commons codec, junit and more). Our web frontend uses some MIT-licensed JavaScript libraries.

Required Resources

Mailing list

We will migrate our mailing lists to the following:

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

Source control

We would like to use Git for source control and enable GitHib mirroring functionality, where code reviews on GitHub are automatically forwarded to the developer mailing list. (See also: https://blogs.apache.org/infra/entry/improved_integration_between_apache_and)

Issue tracking

We are currently using GitHub for issue tracking. We request an Apache-hosted JIRA, and we will import existing issues there.

Initial committers

- Stephan Ewen - stephan.ewen@tu-berlin.de
- Fabian Hueske - fabian.hueske@tu-berlin.de
- Daniel Warneke - warneke@posteo.de
- Robert Metzger - metrobert@gmail.com
- Ufuk Celebi - u.celebi@fu-berlin.de
- Aljoscha Krettek - aljoscha.krettek@gmail.com
- Kostas Tzoumas - kostas.tzoumas@tu-berlin.de
- Sebastian Schelter - ssc@apache.org

Affiliations

- Stephan Ewen (TU Berlin)
- Fabian Hueske (TU Berlin)
- Daniel Warneke (Amadeus IT Group)
- Robert Metzger (TU Berlin)
- Ufuk Celebi (FU Berlin)
- Aljoscha Krettek (TU Berlin)
- Kostas Tzoumas (TU Berlin)
- Sebastian Schelter (TU Berlin)

Sponsors

Champion

Alan Gates (gates@apache.org)

Nominated Mentors

- Sean Owen (srowen@apache.org) (Note: Sean is an Apache member but not currently on the IPC, he will need to request IPMC membership)
- Ted Dunning (tdunning@apache.org)
- Owen O'Malley (omalley@apache.org)
- Henry Saputra (hsaputra@apache.org)
- Ashutosh Chauhan (hashutosh@apache.org)

Sponsoring Entity

The Apache Incubator