# GoraProposal

## Gora Proposal for Apache Incubation

### Abstract

Gora is an ORM framework for column stores such as Apache HBase and Apache Cassandra with a specific focus on Hadoop.

### Proposal

Although there are various excellent ORM frameworks for relational databases, data modeling in NoSQL data stores differ profoundly from their relational cousins. Moreover, data-model agnostic frameworks such as JDO are not sufficient for use cases, where one needs to use the full power of the data models in column stores. Gora fills this gap by giving the user an easy-to-use ORM framework with data store specific mappings and built in Apache Hadoop support.

The overall goal for Gora is to become the standard data representation and persistence framework for big data. The roadmap of Gora can be grouped as follows.

- Data Persistence : Persisting objects to Column stores such as HBase, Cassandra, Hypertable; key-value stores such as Voldermort, Redis, etc; SQL databases, such as MySQL, HSQLDB, flat files in local file system of Hadoop HDFS.
- Data Access : An easy to use Java-friendly common API for accessing the data regardless of its location.
- Indexing : Persisting objects to Lucene and Solr indexes, accessing/querying the data with Gora API.
- Analysis : Accesing the data and making analysis through adapters for Apache Pig, Apache Hive and Cascading
- MapReduce support : Out-of-the-box and extensive MapReduce (Apache Hadoop) support for data in the data store.

### Background

ORM stands for Object Relation Mapping. It is a technology which abstacts the persistency layer (mostly Relational Databases) so that plain domain level objects can be used, without the cumbersome effort to save/load the data to and from the database. Gora differs from current solutions in that:

- Gora is specially focussed at NoSQL data stores, but also has limited support for SQL databases
- The main use case for Gora is to access/analyze big data using Hadoop.
- Gora uses Avro for bean definition, not byte code enhancement or annotations
- Object-to-data store mappings are backend specific, so that full data model can be utilized.
- Gora is simple since it ignores complex SQL mappings
- Gora will support persistence, indexing and anaysis of data, using Pig, Lucene, Hive, etc

### Rationale

ORM frameworks are nothing new. But with the explosion of data generated in Terabytes and even Petabytes, NoSQL data stores are gaining ever-increasing popularity. Coupled with limited support to already-proven Apache Hadoop support in current ORM frameworks, there was a need for a new project.

Gora is currently hosted at Github. However, Gora has ties to ASF in many ways. As detailed in the proposal section, Gora will be a high level client for many Apache projects and subprojects including Hadoop(common, hdfs, and mapreduce), HBase, Cassandra, Avro, Lucene, Solr, Pig, and Hive. Gora already uses Hadoop, HBase, Cassandra and Avro. Moreover, Gora started its life inside Apache Nutch project, and now Nutch trunk uses Gora as a library. Even more, the initial set of committers are all ASF members. Therefore, we think that Apache will be an excellent home for Gora.

### Initial Goals

Initial goals for Gora can be summarized as:

- Iron out the remaining issues with HBase, Cassandra and SQL support.
- Make the first release before the end of the year.
- Improve documentation
- Support for Cascading

### Current Status

#### Meritocracy

Current commit rights belong to the initial list of committers four of who are also ASF members. All the developers have extensive experience with Apache projects. We honor the meritocracy policy of ASF foundation.

#### Community

Gora’s community mostly overlap with that of Nutch, Hadoop, HBase, Avro and Cassandra. We have a small community for now (5 initial committers, 18 people tracking the project at Github), but have been piggybacking the Nutch community for a while. If Gora is accepted to Apache Incubator, we expect more traction. Moreover, with the increasing popularity of NoSQL databases, we expect more users.

### Core Developers

Gora was started by the initial code base inside Apache Nutch by Doacan Güney. Then Enis Söztutar has refactored and re-architected the project out of Nutch. Later Julien Nioche, Andrzej Bialecki and Doacan has ported Nutch to use the newly formed project. Later, Sertan Alkan has joined. Doacan and Julien are Nutch PMC members, Andrzej is the Nutch PMC chair. Enis is an Apache Hadoop PMC member.

### Alignment

As discusssed in the second paragraph of Rationale Section, all of the current developers are Apache people, and four of them are PMC members, which shows that we have some experience with the Apache way. Moreover, Gora is tightly related with lots of Apache projects, Nutch, Hadoop, HBase, Cassandra, Avro, Pig, Hive, Lucene to name a few. Gora has started its life inside Nutch, and now nutch trunk uses Gora to persist web crawl data to HBase, Cassandra and MySQL, which means that Gora is a very critical component in Nutch.

## Known Risks

### Orphaned Products

Most of the development depends on Enis and Doacan for now. Both of them intent to continue Gora development. However, we also acknowledge that more core developers are needed for the project to be truly successful. The general strategy to acquire more developers will be to acquire more users, and encourage users to be active in the community and develop patches. Moreover, the next release of Nutch planned before the end of 2010 has extensive Gora support. We expect more interest from Nutch community, and we will continue to announce Gora notifications at Hadoop,HBase and Cassandra mailing lists.

### Inexperience with Open Source

We believe that all of the developers have extensive open source experience. Four of the initial committers are apache members. The codebase is also open source since April 2010. We also have some documentation, wiki pages, issue tracker and dev mailing list.

### Homogeneous Developers

We have a semi-distributed development environment where Doacan, Enis and Sertan share the same office, but Andrzej and Julien are independent. With the aim of acquiring more developers, we expect more heterogeneous development.

### Reliance on Salaried Developers

Gora development have been supported by ant.com search engine as contract work. It is expected that this contract will continue in the future. However, even without sponsors, we are commited to continue on Gora development, since we believe in the technology it brings and it’s vital role in Nutch, and our other closed sourced projects.

### Relationships with Other Apache Products

Gora will be tightly related to lots of Apache projects:

- **Nutch** : Apache nutch was to home to Gora’s initial code base. Now, Nutch trunk uses Gora as a library. The next relase of Nutch, planned before the end of 2010 will be using Gora’s first release.
- **Hadoop** : Gora has extensive support for Hadoop MapReduce Gora defines all the necessary data structures for working with Hadoop .Data stored in column oriented data stores can be analyzed with Gora using Hadoop.
- **Avro** : Gora uses and extends Avro. Data beans in Gora are defined using Avro schemas ,and compiled into Java code with the extended version of the Avro compiler. Avro is also used in data serialization.
- **HBase** : Gora supports HBase as a persistency backend.
- **Cassandra** : Gora support Cassandra as a persistency backend.
- **Lucene/Solr** : Gora intends to support Lucene/Solr as a persistency and indexing backend.
- **Pig** : Gora intends to support Pig for data anaysis
- **Hive** : Gora intends to support Hive for data analysis

### An Excessive Fascination with the Apache Brand

Gora is a natural fit for Apache due to it's current commiters and depending projects.

## Documentation

- The project is currently hosted at http://github.com/enis/gora/.
- Wiki pages can be found at http://wiki.github.com/enis/gora/.
- List of issues can be found at http://github.com/enis/gora/issues/.
- Current web address: http://groups.google.com/group/gora-dev.

- Current email address: gora-dev@googlegroups.com.

# Initial Source

The initial source was developed as a patch to the Apache Nutch project. But the storage abstraction layer was orthogonal to the web crawler, and we decided to extract it to a separate project with much wider goals. Thus Gora, as a project, was born. The initial code is developed by Enis and Dogacan with ant.com’s sponsorship.

The code can be found at http://github.com/enis/gora/.

# External Dependencies

External dependencies excluding Apache projects are as follows

- JDOM - http://jdom.org/ - Apache-style license
- SQL Builder - http://openhms.sourceforge.net/sqlbuilder/ - Artistic License, LGPL. SQL Builder is intended to be removed from the source due to technical reasons anyway.
- HSQLDB - http://hsqldb.org/ - BSD-style license
- JUnit - http://junit.org - Common Public License 1.0
- SLF4J - http://www.slf4j.org/ - MIT License
- Google Guava Libraries - http://code.google.com/p/guava-libraries/ - Apache License 2.0

# Required Resources

## Mailing Lists

- gora-private (with moderated subscriptions)
- gora-dev
- gora-commits

## Subversion Directory

- http://svn.apache.org/repos/asf/incubator/gora

## Issue Tracking

- JIRA (GORA)

## Other Resources

We need a wiki at http://wiki.apache.org. Currently, we have a wiki at Github, Since there is not a lot of pages there, we can manually move the pages to the wiki at wiki.apache.org.

# Initial Committers

| Name | email | Affiliation | Timezone | | |
|---|---|---|---|---|---|
| <ac:structured-macro ac:name="unmigrated-wiki-markup" ac:schema-version="1" ac:macro-id="8916d453-42da-4145-b579-07a8ef803b3d"><ac:plain-text-body><![CDATA[ | Enis Söztutar | enis [at] apache. org | Konneka | +3 | ]]></ac:plain-text-body></ac: structured-macro> |
| <ac:structured-macro ac:name="unmigrated-wiki-markup" ac:schema-version="1" ac:macro-id="9fb9b1a6-dbc7-48dd-a6b8-7c453afd8c09"><ac:plain-text-body><![CDATA[ | Doacan Güney | dogacan [at] apache.org | Konneka | +3 | ]]></ac:plain-text-body></ac: structured-macro> |
| <ac:structured-macro ac:name="unmigrated-wiki-markup" ac:schema-version="1" ac:macro-id="89fb6e7d-6f0d-4540-afb8-485f5f7ca0d8"><ac:plain-text-body><![CDATA[ | Sertan Alkan | sertanalkan [at] gmail.com | Konneka | +3 | ]]></ac:plain-text-body></ac: structured-macro> |
| <ac:structured-macro ac:name="unmigrated-wiki-markup" ac:schema-version="1" ac:macro-id="9e216d96-e8d3-40dc-96d4-823ec3680d54"><ac:plain-text-body><![CDATA[ | Julien Nioche | jnioche [at] apache. org | [DigitalPe bble] | +1 | ]]></ac:plain-text-body></ac: structured-macro> |
| <ac:structured-macro ac:name="unmigrated-wiki-markup" ac:schema-version="1" ac:macro-id="336b57b8-6a86-441f-bc45-af83179a4263"><ac:plain-text-body><![CDATA[ | Andrzej Bialecki | ab [at] apache.org | Sigram | | ]]></ac:plain-text-body></ac: structured-macro> |
| <ac:structured-macro ac:name="unmigrated-wiki-markup" ac:schema-version="1" ac:macro-id="37462f77-2efa-4bf5-bf1e-6a57a3d8e16a"><ac:plain-text-body><![CDATA[ | Andrew Hart | ahart [at] apache. org | NASA JPL | -8 | ]]></ac:plain-text-body></ac: structured-macro> |
| <ac:structured-macro ac:name="unmigrated-wiki-markup" ac:schema-version="1" ac:macro-id="ad6a10f0-e53b-4271-97f8-5a69dec7eba8"><ac:plain-text-body><![CDATA[ | Dave Woollard | woollard [at] apache.org | NASA JPL | -8 | ]]></ac:plain-text-body></ac: structured-macro> |
| <ac:structured-macro ac:name="unmigrated-wiki-markup" ac:schema-version="1" ac:macro-id="6803686a-2765-4e76-aab6-776da7e22863"><ac:plain-text-body><![CDATA[ | Henry Saputra | hsaputra [at] apache.org | Yahoo! | -8 | ]]></ac:plain-text-body></ac: structured-macro> |

## Affiliations

All of the parties are affiliated with companies and organizations that are familiar with the development of open source . Most of the original Gora development was sponsored by ant.com, however we expect that the amount of volunteer work will increase, and more developers will come on board.

# Sponsors

## Champion

- Chris Mattmann (mattmann AT apache DOT org)

## Nominated Mentors

- Chris Mattmann (mattmann AT apache DOT org)
- Andrzej Bialecki (ab AT apache DOT org )
- Tom White (tomwhite AT apache DOT org)

## Sponsoring Entity

Apache Incubator. Successful graduation can result in either being a TLP, or a subproject of Hadoop, since most of the community is projected to overlap.