Денис Силаков

Twitter и Open Source

Активное использование свободного ПО в интернет-компаниях — не новость. При этом крупные игроки не просто используют готовые продукты, но и адаптируют их под свои нужды. Результаты такой адаптации нередко возвращаются назад в сообщество и их разработка приобретает открытый характер. Одним из примеров такого сотрудничества интернет-гигантов и мира FLOSS является сервис микроблогов Twitter.

Открытое ПО внутри Twitter

Полагаю, не вызовет большого удивления тот факт, что тысячи серверов Twitter работают под управлением ОС Linux. Правда, особым вкладом в развитие непосредственно ядра Linux или низкоуровневых библиотек компания похвастаться пока не может — ее интересы лежат в несколько другой области.

Так, для хранения данных компания использует MySQL, а точнее - собственный форк, исходный код которого выкладывает в открытый доступ. В открытом доступе находится и Twemcache - переработанный вариант инструментария кэширования данных Memcached, сопровождаемый утилитами наподобие измерителя производительности twemperf и прокси twemproxy. Есть и прототип собственной реализации клиента и сервера Memcached на языке Scala - schmemcached.

В компании активно используют наработки проекта фонда Apache (библиотеку полнотекстового поиска Lucene, фрейморк для разработки и выполнения распределённых программ Hadoop, распределенную СУБД Cassandra и другие продукты). Для Cassandra в Twitter разработали свой клиент Cassie на языке Scala. Для использования в Hadoop поддерживается hadoop-lzo (распараллеливаемая реализацию сжатия по алгоритму LZO), система генерации индексов Elephant-twin, а также scalding и pycascading - библиотеки-обертки (на Scala и Python соответственно) к Cascading - слою промежуточного ПО поверх Hadoop, предназначенного для выполнения сложных задач по анализу данных без погружения в дебри MapReduce. Полезна и утилита hdfs-du, занимающаяся визуализацией использования HDFS (Hadoop Distributed File System) - распределённой файловой системы, используемая в Hadoop. Наконец, небезынтересным продуктом является Ambrose - платформа для визуализации и мониторинга потоков данных в MapReduce.

В процессе разработки ПО в Twitter применяются системы контроля версий Subversion и Git, инструментарий непрерывной интеграции Jenkins, IDE Eclipse и прочие известные программы и утилиты. К слову, разработка (во всяком случае, видимая ее часть) ведется на C/C++, Objective-C, Java, JavaScript, Python и Ruby, но самое пристальное внимание уделяется Scala - схожему с Java языку программирования, программы на котором могут компилироваться в байт-код Java либо .NET. При разработке на этих языках используются свободные компиляторы, интерпретаторы, фреймворки и другой инстурментарий. Во многие из этих продуктов сотрудники Twitter время от времени передают свои патчи и улучшения. Многие из этих наработок находят применение и вне компании.

Инструментарий для разработчиков ПО

Даже при беглом взгляде на перечень открытых проектов от Twitter, бросается любовь компании

к языку Scala. Для программистов, использующих этот язык, в компании разработан ряд компонентов, доступных под открытыми лицензиями:

- twitterActors - улучшенная версия библиотеки Actors для Scala (используемой для распараллеливания процессов);
- scala-json - интсрументарий для работы с JSON (популярным текстовым форматом обмена данными, основанном на JavaScript), а также jerkson - обертку на Scala для Jackson - обработчика JSON на Java;
- finagle - собственная реализация механизма удаленного вызова процедур (Remote Procedure Call, RPC) для Scala, а также rpc-client - библиотека, инкапсулирующая различные аспекты RPC;
- joauth - библиотека поддержки авторизации с помощью Oauth;
- chainsaw - обертка на Scala для библиотеки логгирования SLF4J;
- cassovary - библиотека для работы с графами больших размеров;
- ... и другие относительно небольшие библиотеки и утилиты.

Помимо библиотек для самого языка, инженеры компании разработали ряд плагинов и надстройки для sbt (инструменту сборки для Scala и Java) и Thrift (языка описания интерфейсов программных компонентов). В частности, сотрудники Twitter поддерживают Scrooge - генератор Thrift-описаний для Scala.

Те, кто хочет познакомится со Scala, могут найти на сайте http://github.com/twitter учебник Effective Scala и набор уроков Scala Lessons по основам языка. Полезно ознакомиться и с интервью с разработчиками Twitter, в котором они рассказывают о преимуществах языка — http://goo.gl/wmlFW.

Впрочем, одной Scala активность Twitter на поприще инструментов для программистов не ограничивается, разработчики на других языках также не обделены вниманием. Ruby-программисты получили в свое распоряжение открытый инструментарий для тестирования Webrat и библиотеку scribe для работы с одноименным открытым сервером сбора журналов с большого количества систем в реальном времени (к слову, разработанном в Facebook). Разработчики на Java найдут полезными querulous (набор классов для работы с базами данных, избавляющий программистов от кучи рутины, возникающей при прямом использовании JDBC), библиотеку cloudhopper-commons-gsm для работы с такими аспектами телефонии, как SMS и MMS, а также Java-реализацию протокола SMPP (Short message peer-to-peer protocol, используемый для взаимодействия с SMS-сервером).

Веб-разработчики оценят фреймворк для создания сайтов Twitter Bootstrap и сопутствующие утилиты наподобие Recess, проверяющей качества кода на CSS, а также JavaScript-инструменты hogan.js и mustache.js для обработки шаблонов при верстке веб-сайтов.

Есть и инструменты, которые пригодятся многим разработчикам, независимо от используемого языка программирования. Например, Iago - инструментарий нагрузочного тестирования серверов (генерирующий поток данных из сети), или Snowflake - сервис для генерации большого количества уникальных числовых идентификаторов.

«Большие» программные продукты

Помимо различных фреймворков, библиотек и утилит для разработчиков, Twitter открыла код и

нескольких крупных программных продуктов, играющих важную роль в работе самого сервиса.

На заре появления Twitter, данные в нем хранились в MySQL. Однако со временем возможностей открытой СУБД стало не хватать, и не спасал даже Memcached. Для решения проблемы обработки лавинообразно растущего объема данных, в компании разработали Flockdb - распределенное отказоустойчивое хранилище для данных, представляемых в виде графов (фактически, разновидность сетевой СУБД). Именно FlockDB используется в компании для хранения информации о пользователях и их связях друг с другом. С 2010 года код FlockDB выложен в открытый доступ. Для работы с распределенными хранилищами FlockDB используется фреймворк Gizzard - также открытый.

Разрабатывается в Twitter и еще одно собственное хранилище данных — Haplocheirus, предназначенное для хранения твитов. В настоящее время Haplocheirus считается экспериментальным проектом, еще не готовым к промышленному использованию.

Еще одним заметным продуктом с открытым кодом является отказоустойивая система анализа данных Storm, предназначенная для обработки больших потоков информации в реальном времени (и называемый многими не иначе как «Hadoop реального времени»). Интересно, что Storm достался интернет-гиганту вместе с приобретением компании BackType, и это не единственный пример открытия кода приобретенных продуктов — например, на http://github.com/whispersystems/ можно найти код нескольких приложений WhisperSystems, приобретенной Twitter в 2011 году.

Несколько выделяется из общей группы TwUI — фреймворк для создания графического интерфейса для приложений на MacOS X. Для отрисовки GUI, созданного в TwUI, используются вычислительные мощности видеокарты.

Завершая разговор о программных продуктах Twitter, нельзя не отметить инструментарий Ostrich для сбора статистических данных с различных серверов и создания отчетов, распределенную систему мониторинга Zipkin, используемую в самом Twitter для отслеживания состояния всех сервисов компании, а также собственную систему очередей сообщений Kestrel, использующую протокол Memcached.

Twitter API

Важной особенностью Twitter, способствующей его популярности, является открытость API сервиса - функций, с помощью которых можно получать и отправлять сообщения-твиты (отмечу, что хотя API и открыты, для непосредственной работы приложения с Twitter это приложение необходимо сначала зарегистрировать на сайте сервиса и получить OAuth-токен, с помощью которого будет происходить идентификация при общении с сервисом).

Общение приложений с сервисом происходит по протоколу HTTP (предоставляются REST API, основанный на взаимодействии посредством отдельных POST- и GET-запросов, а также Streaming API, подразумевающий использование постоянного соединения по HTTP). Для многих языков существуют библиотеки-обертки, избавляющие программистов от низкоуровневой работы с протоколом HTTP; список открытых продуктов для работы с Twitter API можно найти здесь - https://dev.twitter.com/docs/twitter-libraries. Часть этих библиотек поддерживаются сообществом, в разработке некоторых принимают участие и сотрудники Twitter.

Помимо инструментария для непосредственной работы с API, компания предоставляет библиотеку twitter-text для анализа твитов (например, извлечения из них имен пользователей, тэгов, ссылок и другой интересной информации). Реализация библиотеки доступна для Ruby,

Java, Objective-C и JavaScript.

Организационная поддержка

Помимо написания кода, Twitter активно поддерживает различные организации, занимающиеся развитием открытых проектов, и связанные с миром FLOSS активности. Интернет-гигант поддерживает Ada Initiative (призванную увеличить долю представительниц прекрасного пола в мире FLOSS), фонды Apache Software Foundation и Eclipse Foundation, участвует в работе Java Community Process (в рамках которого разрабатываются спецификации будущих версий Java) и проекте OpenJDK. Не так давно Twitter стал серебряным членом консорциума The Linux Foundation.

Чтобы систематизировать взаимодействие с сообществом FLOSS, в 2011 году в Twitter создали отдельное подразделение - Open Source Office (https://dev.twitter.com/opensource), которое курирует открытые проекты самой компании. С момента создания этого подразделения в 2010 году (http://www.nixp.ru/news/10139.html), общее число открытых проектов от Twitter выросло с 22 до 84; в работе над ними принимают участие почти две сотни сотрудников компании (при том что всего в компании работает порядка 900 человек).

Заключение

Итак, за последние два с половиной года Twitter превратился в достаточно заметного участника мира открытого ПО. В отличие от многих крупных корпораций, о которых я рассказывал в предыдущих статьях в OS, и чей основной бизнес связан с аппаратным обеспечением, деятельность Twitter на арене FLOSS связана прежде всего с веб-разработкой и созданием стрессоустойчивых сервисов, способных выдерживать колоссальные объемы запросов от пользователей.

При этом компания не ограничивается выпуском различных вспомогательных библиотек и утилит; сообщество получает программные продукты, прошедшие серьезные испытания в самой компании и обеспечивающие успех ее деятельности. Приятно, что компания признает важную роль свободного ПО в своем бизнесе и с охотой идет на сотрудничество с сообществом. Остается пожелать ей поддерживать темпы развития этого сотрудничества — уверен, оно принесет выгоду всем участникам.