

## ПРИНЦИПЫ РЕАЛИЗАЦИИ МАСШТАБИРУЕМЫХ ОБЛАЧНЫХ СЕРВИСОВ СБОРА И ОБРАБОТКИ ДАННЫХ НА БАЗЕ ПЛАТФОРМЫ APACHE HADOOP

Афанасьев А.П., Корх А.В., Смирнов С.А., Рубцов А.О., Сухорослов О.В.

*Институт системного анализа РАН*

В последнее время широкое распространение получила модель облачных вычислений “приложение как сервис” (Software as a Service, SaaS), заключающаяся в оформлении приложения в виде удаленно доступного сервиса. Одной из перспективных областей применения модели SaaS является реализация масштабируемых облачных сервисов сбора и обработки данных (ССОД). Актуальность данного направления исследований обусловлена растущей потребностью в организации сбора и обработки больших объемов данных с использованием технологий распределенных вычислений, как в науке, так и бизнесе.

Несмотря на появление развитых технологических решений в этой области, среди которых лидером является платформа Apache Hadoop, остается нерешенной проблема, связанная с развертыванием и настройкой соответствующей вычислительной инфраструктуры и программной платформы. Для решения данной проблемы предлагается использовать модель облачных вычислений, предоставив доступ к функциональности платформы Apache Hadoop через набор проблемно-ориентированных ССОД в различных предметных областях.

Технологический задел, сформированный в рамках платформы Hadoop и родственных проектов, может быть эффективно использован для решения задач по сбору, хранению и обработке больших объемов данных в рамках ССОД. При этом может быть обеспечено соблюдение важных нефункциональных требований, таких как производительность, масштабируемость, надежность и безопасность. Использование стандарта де-факто в области обработки больших объемов данных, каким является в настоящее время платформа Hadoop, также позволит обеспечить быстрый переход пользователей, знакомых с этой платформой, на ССОД.

Разработана архитектура программного комплекса, дополняющего базовую функциональность компонентов платформы Hadoop с целью поддержки создания масштабируемых ССОД. Данная архитектура включает следующие программные модули.

- 1) Модуль хранения данных предназначен для обеспечения масштабируемого хранения данных ССОД. Данный модуль планируется реализовать на базе распределенной файловой системы HDFS, входящей в состав платформы Apache Hadoop.
- 2) Модуль сбора данных предназначен для поддержки передачи данных из распределенных источников в ССОД в непрерывном режиме. Данный модуль планируется реализовать на базе технологии Apache Flume.
- 3) Модуль обработки данных предназначен для поддержки распределенной обработки данных на основе модели вычислений MapReduce. Данный модуль планируется реализовать на базе платформы Apache Hadoop.
- 4) Модуль выборки и анализа данных предназначен для поддержки работы с данными, хранимыми ССОД, при помощи высокоуровневых языков и средств доступа. В отличие от МОД, данный модуль должен позволять описывать процедуру выборки или анализа данных не в виде низкоуровневого программного кода, а в виде выражений на SQL-подобном языке запросов. Это позволит облегчить выбор и анализ данных по требованию.
- 5) Модуль безопасности предназначен для поддержки аутентификации, авторизации и разграничения доступа к данным между пользователями ССОД.
- 6) Модуль создания ССОД предназначен для поддержки создания специализированных веб-сервисов для сбора и обработки данных. Данный модуль использует в своей работе функциональность всех других модулей ЭО ПК. В качестве базовой платформы для организации удаленного доступа к сервисам предлагается использовать протоколы и технологии Web, основанные на архитектурном стиле REST.

Работа выполняется в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2013 годы» при финансовой поддержке Минобрнауки, государственный контракт № 14.514.11.4021.