РЕАЛИЗАЦИЯ ПРОГРАММНОГО КОМПЛЕКСА ДЛЯ СОЗДАНИЯ СЕРВИСОВ СБОРА И ОБРАБОТКИ ДАННЫХ НА БАЗЕ ПЛАТФОРМЫ APACHE HADOOP

Афанасьев А.П., Волков С.Ю., Гринберг Я.Р., Джорджевич Т.С., Корх А.В., Смирнов С.А., Рубцов А.О., Сухорослов О.В.
Институт проблем передачи информации РАН

Целью проводимых исследований является разработка экспериментального образца программного комплекса (ЭО ПК) для создания масштабируемых облачных сервисов сбора и обработки данных (ССОД) на базе платформы Apache Hadoop. 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.

Разработанный ЭО ПК предоставляет набор модулей, решающих следующие задачи: сбор данных из территориально распределенных информационных источников; надежное хранение собранных данных в распределенной файловой системе; высокопроизводительная обработка собранных данных с использованием модели вычислений MapReduce; выбор и анализ собранных данных по требованию с использованием SQL-подобных языков запросов и других высокоуровневых средств.

Модуль сбора данных (МСД) предназначен для поддержки передачи данных из распределенных источников в ССОД в непрерывном режиме. Данный модуль реализован на базе технологии Apache Flume. Для поддержки передачи данных из Flume в ССОД создана специальная реализация приемника данных (data sink), который передает данные в ССОД по сети.

Модуль хранения данных (МХД) предназначен для обеспечения масштабируемого хранения данных ССОД. Данный модуль реализован на базе распределенной файловой системы HDFS, входящей в состав платформы Apache Hadoop. МХД выступает в роли посредника между HDFS и другими модулями ЭО ПК. МХД реализует программный интерфейс (API), позволяющий другим модулям производить чтение и запись данных, а также другие операции с файловой системой.

Модуль обработки данных (МОД) предназначен для поддержки распределенной обработки данных на основе модели вычислений MapReduce. Данный модуль реализован на базе платформы Apache Hadoop. МОД выступает в роли посредника между Hadoop MapReduce и другими модулями ЭО ПК. МОД реализует программный интерфейс (API), позволяющий другим модулям производить запуск MapReduce-заданий и контроль их выполнения. Результаты MapReduce-заданий размещаются в HDFS, откуда они могут быть получены с помощью МХД.

Модуль выборки и анализа данных (МВАД) предназначен для поддержки работы с данными, хранимыми ССОД, при помощи высокоуровневых языков запросов и средств доступа. В отличие от МОД, данный модуль позволяет описывать процедуру выборки или анализа данных не в виде низкоуровневого программного кода, а в виде выражений на SQL-подобном языке запросов. Это позволяет облегчить выбор и анализ данных по требованию. Для реализации МВАД была выбрана технология Apache Pig, как одна из наболее развитых и широко используемых. Для описания процедур выборки и анализа данных Pig предоставляет высокоуровневый язык программирования Pig Latin, сочетающий в себе SQL-операторы и процедурный стиль программирования. Также в состав Pig входит транслятор программ на PigLatin в MapReduce-задания, выполняемые на платформе Hadoop. МВАД реализует программный интерфейс (API), позволяющий другим модулям ЭО ПК производить запуск Pig-программ и контроль их выполнения. Результаты выполнения Pig-программ размещаются в HDFS, откуда они могут быть получены с помощью МХД.

Модуль безопасности (МБ) предназначен для поддержки аутентификации, авторизации и разграничения доступа к данным между пользователями ССОД. МБ самостоятельно осуществляет аутентификацию пользователей ССОД, а для авторизации использует возможности Hadoop. Модули ЭО ПК используют полученные от МБ имена пользователей при доступе к данным и запуске заданий от имени пользователей.

Модуль создания ССОД (МСС) предназначен для поддержки создания специализированных веб-сервисов для сбора и обработки данных. Данный модуль использует в своей работе функциональность всех других модулей ЭО ПК. МСС представляет собой набор готовых компонент для реализации ССОД: адаптер для реализации веб-сервисов обработки данных (АВСОД) на базе МОД; адаптер для реализации веб-сервисов выборки и анализа данных (АВСВАД) на базе МВАД.

Работа выполнена 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.