Создание системы распределенных вычислений на основе технологий Hadoop

Потапов В.П., Гиниятуллина О.Л., Харлампенков И.Е.

В работе рассматриваются вопросы построения системы распределенных вычислений. В настоящее время Кемеровский филиал Института вычислительных технологий СО РАН (КФ ИВТ СО РАН) принимает участие в проектах, генерирующих потоки данных различных предметных областей (данные экологического мониторинга горнодобывающих предприятий, спутниковые снимки и т.д.). Обработка подобного рода разнородных данных требует применения специальных алгоритмов и подходов к интеграции, что накладывает определенные ограничения на технологию и реализацию вычислений. В докладе предлагается применение концепции BigData [1] для решения задач, так как поступающая информация соответствует определяющим характеристикам для больших данных, известным как «три V»[1]:

* объём (volume) – физический размер данных;
* скорость (velocity) как поступления, так и обработки для получения результатов;
* многообразие (variety) источников, структур и форматов обрабатываемых данных.

Для построения системы предлагается использовать технологию MapReduce[1], которая позволяет выполнять отдельные операции обработки данных параллельно на нескольких узлах сети, а потом суммировать их для получения конечного результата. В качестве реализации выбрана система Apache Hadoop [2] с рядом инфраструктурных проектов.

Взаимодействие с пользователем осуществляется через web-интерфейс, содержащий ряд инструментов: средства поиска данных из всех доступных источников (базы данных, файловые архивы и внешние сервисы), их преобразования и загрузки в кэш на основе NoSQL систем; конструктор заданий обработки на основе комбинации готовых процедур; подсистема запуска и контроля выполнения задач; средства визуализации результатов и их анализа. Расширение возможностей информационной системы по обработке данных предлагается через написание новых модулей на основе Hadoop framework. Также возможно создание пользовательских алгоритмов на языке Python в рамках предлагаемого программным комплексом API.

В настоящее время система распределенных вычислений находится на стадии формирования окончательных требований, проектирования архитектуры, выбора библиотек и механизмов из взаимодействия.

ЛИТЕРАТУРА

1. Sawant N, Shah H. Big Data Application Architecture Q&A. New York: Apress. 2013. – 172 p.
2. Apache Hadoop [Электронный ресурс]: http://hadoop.apache.org/