РАСПРЕДЕЛЕННАЯ СИСТЕМА ОБРАБОТКИ ДАННЫХ

Дымникова Н. А., студентка кафедры системного программирования СПБГУ, natalia.dymnikova@gmail.com

Введение

Объемы информации, генерируемые современными вычислительными системами, требуют всё больше ресурсов для анализа и обработки. Вычислительных ресурсов одного устройства уже не хватает для достаточно быстрого решения поставленных задач

На сегодняшний день, разработка программного обеспечения для выполнения в пределах одного вычислительного устройства является не рациональной, потому что современные бизнес-задачи имеют строгие ограничения на время исполнения и любые задержки приводят к финансовым потерям.

Для этого создается распределенная среда, объединяющая десятки и сотни вычислительных устройств в единую сеть для параллельной обработки информации

Эта работа описывает решение проблемы обработки данных, хранящихся на различных устройствах в одном кластере при помощи построения запроса, состоящего из нескольких этапов: создание или чтение данных, их обработка и слияние нескольких потоков.

Применение

Разрабатываемая система предназначена для анализа диагностической информации, производимой аппаратно-программным комплексом EMC Elastic Cloud Storage (ECS), с целью поиска причин аномалий функционирования.

Основной особенностью продукта EMC ECS является высокая горизонтальная масштабируемость, что позволяет строить системы с большим числом узлов, а это влечет за собой создание большого объема диагностических данных, распределенных по кластеру. Обработку и анализ этих данных необходимо проводить в реальном или приближенном к реальному времени.

Цель

Основной целью является создание программного продукта с дружелюбным для человека интерфейсом для доступа к диагностической информации, распределенной на кластере, для упрощения анализа поведения EMC ECS.

Система позволяет получать в графическом представлении результаты запросов, задаваемых оператором системы, с распределением по времени.

Разработанное решение

Полученное решение принципиально отличается от существующих тем, что требует гораздо меньшего от существующей инфраструктуры. Оно не требует дублирования данных, так как остальным решениям необходимо, чтобы исходные данные были загружены на их сервера для обработки.

Также полученная система накладывает минимальные требования на существующую инфраструктуру: для функционирования не нужна установка специализированных средств управления конфигурацией кластера. Необходимо лишь обеспечить доступ к исходной системе по протоколу SSH и настроенная маршрутизация между всеми вовлеченными узлами сети.

И, наконец, еще одним существенным отличием является возможность работы с исходными данными без предварительно индексирования, что позволяет быстрее приступить к анализу. Это реализовано за счет существования возможности обработки неиндексированных данных и построения индекса «на лету».

Система строится на базе Java фреймворка Akka, позволяющего производить распределенные вычисления в кластере, состоящем из одноранговых узлов. Для ускорения поиска и обработки информации используются индексы под управлением Apache Lucene. Система предоставляет интерфейс, построенный на основе HTML5 с использованием React и D3.

Заключение

Разработка такой системы позволяет глубоко изучить подходы построения распределенных вычислительных сетей, а также испытать на практике уже существующие многочисленные наработки сообщества с открытым программным кодом. А наличие такой системы упростит разработку EMC ECS.