РАЗРАБОТКА ГОРИЗОНТАЛЬНО МАСШТАБИРУЕМОЙ РАСПРЕДЕЛЕННОЙ ВЫЧИСЛИТЕЛЬНОЙ СИСТЕМЫ

Н.С. Хитеш, И.А. Ботыгин
Томский политехнический университет
bia@tpu.ru

Введение

Одним из современных трендов ИТ-отрасли в последнее время является проблема «больших данных» – BIG DATA. «Большие данные» можно увидеть в области финансов и бизнеса, во многих научных исследованиях, в клинических данных, в астрономии, в океанологии, в ряде инженерных расчетов и многих других областях. Большие данные формируют новую информационную культуру, в которой и бизнес и ИТ-специалисты должны объединить свои силы, чтобы создаваемые новые технологии и инструменты помогали перерабатывать терабайты входных данных [1-4].

Для примера, на рис.1. представлена динамика использования «больших данных» по различным электронным СМИ для коммуникации. Электронная почта генерирует наибольшее количество данных (72%) по всему спектру электронных сообщений СМИ.

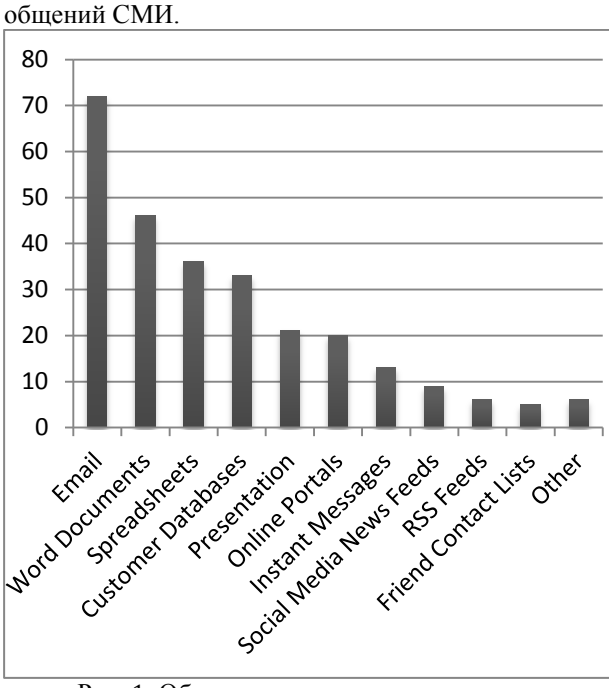

Рис. 1. Объемы генерируемых данных в электронных СМИ

Даже такой аспект рассмотрения больших данных может дать специалистам и пользователям основание принимать оптимальные решения для привлечения клиентов, оптимизации операций, предотвращения угроз и мошенничества [3].

Большие данные генерируются из многих источников с угрожающей скоростью, объемом и разнообразием. Такие объемы данных создают ряд проблем, в частности, необходимо большое количество времени, чтобы обработать данные, и достаточно большие ресурсы, чтобы сохранить их и поддерживать в актуальном состоянии. Несмотря на то, что вычислительная мощность компьютеров с каждым годом увеличивается, а стоимость хранения данных уменьшается – это не решает проблему, так как с развитием цифровых технологий потоки данных увеличиваются в большем масштабе.

Чтобы фирмам получить конкурентное преимущество в хранении больших данных и при этом извлечь выгоду от обработки больших данных, от них требуется инфраструктура, которая может управлять и обрабатывать этот взрывной объем структурированных и неструктурированных данных и обеспечивать их безопасность.

Функциональные требования к инфраструктуре

Для того, чтобы инфраструктура аппаратно-программной системы для больших данных обеспечивала ее использование в эффективной и гибкой форме, она должна удовлетворять определенным требованиям архитектуры.

Масштабируемость. Система должна быть горизонтально и вертикально масштабируемой, т.е. основываться на том факте, что объем данных, который должен храниться и обрабатываться, заранее неизвестен. Поэтому система должна иметь возможность увеличения вычислительной мощности узлов путем замены или обновления аппаратного обеспечения без изменения функционального программного обеспечения.

Кросс-платформность. Есть много программных платформ, работающих на компьютерах, и каждая из них имеет свои преимущества и недостатки. Система кросс-платформности избавит от зависимости от конкретной платформы, для которой написано программный код. Эта особенность системы может значительно упростить механизм ее масштабирования – за счет увеличения количества поддерживаемых платформ.

Простота в использовании. Для облегчения доступа и независимости пользователей или администраторов от наличия на компьютере специального клиентского приложения имеет смысл создать графический веб-интерфейс в виде веб-портала .

Гибкость. Необходима для анализа неструктурированных данных, который включает в себя реализацию различных алгоритмов. Неизвестно, какие могут возникнуть проблемы при обработке данных, и, следовательно, невозможно описать заранее алгоритмы и разработать программы, с

помощью которых система будет решать эти проблемы.

Архитектура системы

В настоящей работе представлена обобщенная функциональная структура системы, по мнению авторов, способная достаточно эффективно с использованием недорогой вычислительной техники решить проблему создания горизонтально масштабируемого вычислительного кластера с требуемыми информационными ресурсами (рис.2.).

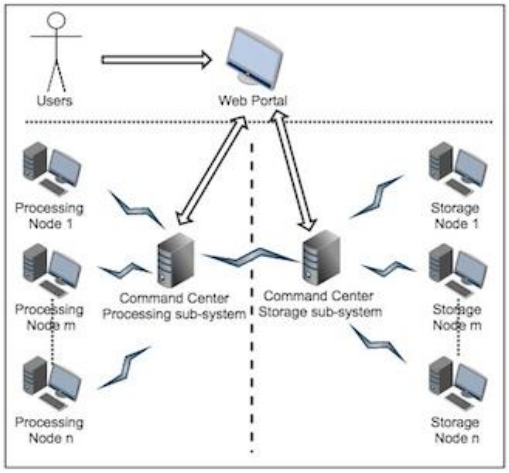

Рис. 2. Архитектура системы

Архитектура системы делится на две подсистемы: подсистему хранения и подсистему обработки. Подсистема обработки состоят из двух основных компонентов: командный центр и вычислительные узлы.

Командный центр обеспечивает функциональность параллельной обработки на вычислительных узлах. Система поддерживает веб-портал для сбора информации о доступных вычислительных ресурсах и предоставляет программный интерфейс для хранения данных и обработки.

Распределение нагрузки по узлам системы осуществляется пользователем.

Безопасность

По своей сути, распределенные системы являются более сложными, чем централизованные системы. Дополнительная сложность может увеличить потенциал для системных сбоев. Поскольку данные обрабатываются и хранятся в распределенной системе, может быть несанкционированный доступ к информации. Поэтому необходимо рассмотрение всех аспектов информационной безопасности и особенно аспекты, связанные с безопасностью на различных уровнях распределенных систем.

• Для входа в систему пользователь должен пройти авторизацию (идентификацию и аутентификацию).

• При передаче через Интернет, необходимо обеспечить безопасность канала передачи и защиту от перехвата. Для этого все соединения – между веб-порталом и командным центром, между командным центром и узлам (узлами хранения и обработки) должны иметь возможность шифровать информацию.

• Для защиты от подслушивания связи между командным центром и вычислительными узлами сообщения могут быть на разных портах. Командный центр постоянно слушает определенный порт, но данные по этому порту не передаются. Порты для передачи выбираются случайным образом. Узлы получают из командного центра эту информацию, а также шифруют передаваемую информацию.

Заметим, что изменять «портовые» данные можно регулярно через определенные промежутки времени или по указанию системного администратора. В этом случае, командный центр создает новый порт и отправляет всем вычислительным узлам команду для изменения канала передачи. Вычислительные узлы по командам от центра осуществляют закрытие старой связи и подключаются к постоянному порту, где они получают новый порт, открывают новое подключение к порту и продолжают работать в обычном режиме.

• При необходимости передачи конфиденциальных данных или данных для большей безопасности, можно использовать безопасный канал (Virtual Private Network).

• Необходимо также защищать данные их от повреждения. В подсистеме хранения все данные дублируются на разных узлах. При сбое дискового накопителя или потери связи с одним из узлов, командный центр будет просто запрашивать те же данные от других узлов.

Заключение

Практическая апробация системы осуществлялась на примере задачи систематизации, хранения и обработки данных наземных метеорологических наблюдений, полученных из сети гидрометеорологических станций Российской Федерации.

Литература

1. What is Big Data? [Электронный ресурс]. - Режим доступа: http://www.ibm.com/big-data/us/en/
2. Research Trends available for Big Data. [Электронный ресурс]. - Режим доступа: http://www.researchtrends.com/wp-content/uploads/2012/09/Research_Trends_Issue30.pdf
3. Research article published by Avanade Group, USA. [Электронный ресурс]. - Режим доступа: http://www.avanade.com/Documents/Research%20and%20Insights/Big%20Data%20Executive%20Summary%20FINAL%20SEOv.pdf
4. Big Data. [Электронный ресурс]. - Режим доступа: http://aws.amazon.com/ru/big-data/