ПРИНЦИПЫ ФУНКЦИОНИРОВАНИЯ И ОБЛАСТИ ПРИМЕНЕНИЯ СИСТЕМ ОБРАБОТКИ БОЛЬШИХ МАССИВОВ ДАННЫХ

Кротов Д.А.

Институт информационных технологий БГУИР,
г. Минск, Республика Беларусь

Скудняков Ю.А.– доцент каф. ПЭ, к.т.н,, доцент

В данной работе рассматриваются различные способы обработки больших данных. Описаны основные особенности способов обработки больших данных. Рассмотрена технология MapReduce, значительно ускоряющая обработку данных.

Большие данные (BigData) – серия подходов и методов обработки структурированных и неструктурированных данных огромных объёмов.

Ежегодно общий мировой объем данных увеличивается в 2 раза. Недорогие диски и онлайн-хранилища позволяют с легкостью отложить принятие решений о том, что делать со всеми этими данными. В этом случае в распоряжении разработчиков и пользователей имеется достаточно памяти для сохранения данных. С помощью современных информационно-вычислительных систем имеется возможность сохранения решений в различных сферах человеческой деятельности. В связи с этим возникает вопрос: можно ли обрабатывать эти данные и как эффективно их использовать [1]?

Большие данные предполагают нечто большее, чем просто анализ больших объемов информации. Проблема не в том, что каждый день организации создают большое количество данных, а в том, что большая часть данных представлена в неструктурированном формате — это веб-журналы, видеозаписи, текстовые документы, машинный код или, например, геопространственные данные. Все эти данные хранятся в различных местах. В результате организации имеют доступ к большому количеству своих данных, но не имеют нужных инструментов, которые могли бы их правильно обработать. К тому же в настоящее время данные обновляются все чаще и чаще, и в итоге организация приходит к ситуации, когда традиционные методы анализа информации не могут быть применены на имеющиеся данные.

В качестве определяющих характеристик для больших данных традиционно выделяют «VVV»: объём (volume), скорость (velocity), многообразие (variety). Под объемом понимается величина физического объема, под скоростью понимается не только рост самого количества данных, но и

скорость их обработки и, наконец, под многообразием понимается возможность одновременной обработки различных типов данных.

Для корректного функционирования система больших данных должна быть основана на следующих принципах:

1.Горизонтальная масштабируемость. Если объем данных увеличивается, то вместе с ним должно и быть увеличено количество серверов для поддержания производительности. В этом и состоит смысл принципа горизонтальной масштабируемости.

2.Отказоустойчивость. При наличии большого количества серверов с ними рано или поздно будут возникать проблемы. Без кластеризации сбой сервера приводит к тому, что поддерживаемые им приложения или сетевые сервисы оказываются недоступны до восстановления его работоспособности. Отказоустойчивая кластеризация исправляет эту ситуацию, перезапуская приложения на других узлах кластера без вмешательства администратора в случае обнаружения сбоев.

3.Локальность данных. Для снижения издержек и соответственно повышения производительности данные необходимо обрабатывать на том же сервере, на котором они изначально находятся.

BigData чаще всего находит себе применение в таких областях как:

1. Торговля. Розничные торговцы прекрасно знают, как использовать BigData в своем бизнесе. Они используют собранную информацию для построения долгосрочных и дружественных отношений с клиентами. Модели поведения прежних покупателей могут быть проанализированы для определения отличительных характеристик тех, кто скорее всего сменит продукт, и тех, кто останется лояльным.

2. Образование. Педагоги смогут модернизировать систему образования, мотивировать учеников и студентов ВУЗов к более успешной работе. Также преподавателю будет проще выявить отстающих, убедиться в понимании темы аудиторией и реализовать более эффективную систему оценки, либо же сформировать общий рейтинг обучающихся, в котором все оценки будут иметь свой вес и на их основе будет формироваться рейтинг, отличный от просто усредненного значения.

3. Банкинг. Каждый день банкиры сталкиваются с колоссальным объемом информации, которая поступает из бесчисленных источников. Грамотная обработка имеющихся инфопотоков позволит повысить удовлетворенность клиентов, минимизировать кредитные риски и предотвратить мошенничество (детектирование аномального поведения). Поэтому финансисты заинтересованы в поиске новых инновационных способов применения BigData, как никто другой.

4. Здравоохранение. Истории болезни, планы лечения, клинические анализы, генетические исследования и рецепты врачей— все это можно объединить в одной базе данных. Аналитика собранных сведений поможет сделать новые выводы о применяемых методах терапии и улучшить уход за пациентами.

5. Производство. Среди всех применений BigData особенно хочется отметить производство. В условиях жесткой рыночной конкуренции важно минимизировать расходы сырья и повысить качество продукции. Решение этих задач подскажет прогнозная аналитика. Например, компьютер может определить, сколько еще проработает та или иная деталь в автомобиле на основе уже собранных данных о деталях автомобилей.

В заключение необходимо отметить, что в настоящее время обработка больших массивов данных является весьма актуальной задачей. Технологии BigData имеют ряд преимуществ, в сравнении с традиционными методами обработки, однако внедрение такого рода технологий является весьма затратным для большинства компаний. Внедряя BigData, компания получает преимущество в скорости обработки данных, в сравнении с конкурентами, что, безусловно, положительно скажется на доходе компании.

Список использованных источников:

1.Ian H. Witten, Eibe Frank, Mark A. Hall Data Mining: Practical Machine Learning Tools and Techniques, Third Edition.

2. Docplayer [Электронный ресурс]. – Электронные данные. – Режим доступа: https://docplayer.ru/34904991-Big-data-aktualnost-i-perspektivy-ispolzovaniya.html.– Дата доступа 12.01.2018.

3.TAdviser [Электронный ресурс]. – Электронные данные. – Режим доступа: http://www.tadviser.ru/_index.php/ Статья: Большие_данные_(Big_Data).– Дата доступа 12.01.2018.