Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК_____

Божко
Светлана Сергеевна

Обработка и анализ большого объема данных в высоконагруженных системах

АВТОРЕФЕРАТ
на соискание степени магистра технических наук
по специальности 1-40 80 04 «Математическое моделирование, численные методы и комплексы программ»

Научный руководитель
Пилецкий Иван Иванович
доцент, кандидат физ.-мат. наук

Минск 2016

КРАТКОЕ ВВЕДЕНИЕ

На протяжении последних нескольких лет данные, хранящиеся в мире, увеличивались экспоненциально. Этот феномен называется «Big Data». Эта тема очень трендовая на данный момент.

В настоящее время наиболее известным и используемым инструментом для обработки такого объема данных является Hadoop. Но он имеет ряд ограничений. И в первую очередь это касается возможности обработки только пакетных данных (batch). Это ограничение препятствует использованию Hadoop в качестве инструмента для обработки потоковых данных.

В потоковой обработке наряду с понятием объема данных понятие скорости является определяющим. Можно утверждать, что такие системы, как правило, относятся к классу высоконагруженных. Поэтому нужны новые инструменты для высокоскоростной обработки большого объема данных. Таким образом, возникает потребность в исследовании принципов построения и масштабирования высоконагруженных систем обработки и анализа потоков данных.

В качестве практического применения знаний, полученных в результате исследования систем обработки и анализа потоков больших данных, будет предложена и реализована архитектура новой процессинговой платформы на предприятии. Платформа ориентирована на высокие нагрузки. Она нацелена на решение следующих проблем в существующей на предприятии системе:

– пакетная природа обработки данных в ETL-процессе;
– неоптимальное использование ресурсов (CPU, RAM);
– проблема с хранением промежуточных данных и пакетов как таковых;
– ручное контролирование потока данных;
– сложность в масштабировании;
– сложно расширять модель данных;
– отчисления на лицензии.

В настоящее время на рынке появилось много инструментов для построения систем потоковой обработки данных. Одним из таких инструментов является Apache Storm. Strom будет глубоко проанализирован и использован в практической части данной работы.

Значительную сложность представляет собой горизонтальное масштабирование систем потоковой обработки из-за сложностей при построении распределенных систем с гарантиями согласованности данных в условиях нестабильной сети. Данная проблема сейчас представляет особый интерес, поскольку ведется активное внедрение современных масштабируемых баз данных, имеется тенденция перехода от NoSQL- к NewSQL-решениям. Это также подчеркивает актуальность исследований в данной области.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы диссертации. В настоящее время обработка большого объема данных является достаточно распространенным явлением благодаря довольно широкому использованию Hadoop и MapReduce. Для многих бизнес-задач очень важно знать, что происходит сейчас, и в связи с этим принимать решения так быстро, как это возможно. Но Hadoop имеет ряд ограничений. И в первую очередь это касается возможности обработки только пакетных данных (batch). Определенно, это ограничение препятствует использованию Hadoop в качестве инструмента для обработки данных по мере их поступления, данных последней версии. Поэтому обработка потоковых данных с низкой задержкой обновления, практически в реальном времени является вызовом на ближайшее будущее. Данная задача решается многими компаниями по-разному. Унифицированного подхода еще нет. В настоящее время на рынке появилось много инструментов для построения систем потоковой обработки данных. Предлагаются решения как open source, так с закрытым исходным кодом. Одним из таких инструментов является Apache Storm. Он становится все более и более популярным и является весьма многообещающим. Поэтому Strom будет глубоко проанализирован и использован в практической части данной работы.

Цели исследования. Главными целями данной работы являются:

– проанализировать предметную область Big Data с точки зрения существующих методов, технологий и походов к обработке потоков данных в высоконагруженных системах;

– предложить и реализовать новую архитектуру процессинговой платформы на базе Apache Storm;

– реализовать прикладную задачу на базе построенной платформы.

Связь работы с приоритетными направлениями научных исследований и запросами реального сектора экономики. Бизнес-задача, решаемая платформой обработки больших объемов потоковых данных (далее будет называться «процессинговая платформа»), относится к сфере онлайн-рекламы. Пользователь постоянно совершает какие-то действия с сайтом или мобильным приложением. Это может быть прокрутка страницы, просмотр баннера, наведение курсора, клик и пр. Все они являются событиями – транзакциями, описывающими взаимодействие конечного пользователя с контентом. Все транзакции поступают на сервер, где должны быть верным образом обработаны, чтобы показывать более релевантную рекламу. Релевантность и успешность рекламной кампании оценивается по отчетам. Так что построение отчетов и подготовка для них данных тоже входит в список задач процессинговой платформы. Идея процессинга заключается в обработке входящих транзакций.

Иногда приходится осуществлять эту обработку несколько раз. По своей сути, обработка – это подготовка данных для отчетов.

В рамках практической части была предложена и реализована новая архитектура процессинговой платформы для решения прикладных задач на предприятии. Процессинговая платформа представляет собой набор решений и серверов, целью которых является управление потоком данных для последующего построения по ним отчетов. Новая платформа должна заменить существующее на предприятии решение.

Научно-практическая новизна и значимость полученных результатов. По результатам практической части было дано заключение о возможности использования фреймворка Apache Storm для построения масштабируемых, отказоустойчивых систем потоковой обработки данных, определены его слабые и сильные стороны. Особое внимание было уделено таким характеристикам, как простота развертывания, надежность и производительность. Одной из приоритетных задач данной работы является разработка методики и общих принципов построения систем потоковой обработки данных на базе Storm.

Личный вклад магистранта. Работа представляет собой результат анализа современных подходов к построению высоконагруженных систем, а также преимуществ и недостатков различных систем хранения данных. С учетом этого была предложена новая архитектура, а также ее реализация на примере платформы потоковой обработки данных на коммерческом предприятии. Проведено системное проектирование прикладной платформы с тщательным выбором технологической базы на основе аналитического обзора научно-технической литературы. Разработанная система базируется на продуктах Apache Storm и Apache Mesos. Уделено особое внимание вопросам отказоустойчивости и масштабируемости при проектировании каждого компонента. Используемые подходы позволяют значительно уменьшить простой системы и максимально использовать современные многоядерные процессоры. Проведено тестирование процессинговой платформы, а также был подробно разобран пример прикладного использования разработанной платформы.

Апробация результатов диссертации. Результаты диссертации внедрены в производство на предприятии. Об этом свидетельствует справка о внедрении.

Опубликованность результатов. По результатам работы было сделано 8 докладов конференциях, в том числе и международных. Результаты работы были опубликованы (4 публикации).

Структура и объем диссертации. Работа состоит из введения, трех глав и заключения. Диссертация изложена на 88 страницах машинописного текста, библиография включает 62 наименования.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность исследования, показывается степень разработанности обозначенной проблемы, ставится цель, формулируются задачи, определяются методы исследования, раскрывается практическая значимость полученных результатов.

В первой главе дана детальная постановка задачи на диссертационное исследование. Определены требования, предъявляемые к практичекой части диссертации — реализации новой процессинговой платформы, которая должна заменить существующее на предприятии решение.

Во второй главе исследуется предметная область Big Data. Определено понятие «Big Data», его сущность и харакетистики. Рассмотрены различные подходы и системы хранения данных в контексте Big Data, в частности: реляционные СУБД, NoSQL базы данных, NewSQL базы данных, а также HDFS и платформа Hadoop. Особое внимание уделено вопросам горизонтальной масштабируемости различных хранилищ данных, а также проблеме построения распределенных систем с гарантиями согласованности данных в условиях нестабильной сети.

В третьей главе дано описание реализации исследовательского проекта: платформы обработки большого объема данных на базе Apache Storm. В данной главе есть детальное описание процессинговой платформы, а также актуальности замены предыдущего решения для потоковой обработки данных новой платформой. Рассмторны методы и технологии построения систем потоковой обработки Big Data, существующие на рынке. Особое внимание уделено лямбда-архитектуре и ее применимости для решения поставленных задач. Детально описана компонентная архитектура новой процессинговой платформы. Соответствие нового решения требованиям, предъявляемым к процессинговой платформе, проверено с помощью нагрузочного и системного тестирования. Спроектирована и реализована прикладная задача на процессинговой платформе. В последнем разделе главы было дано заключение о возможности использования фреймворка Apache Storm для построения масштабируемых, отказоустойчивых систем потоковой обработки данных, определены его слабые и сильные стороны.

В заключении определены теоретические и практические результаты диссертационного исследования.

ЗАКЛЮЧЕНИЕ

В рамках данной работы были рассмотрены следующие направления в Big Data и потоковой обработке данных:

– масштабирование различных систем хранения данных;

– влияние модели данных на масштабирование хранилища;

– распределенная обработка больших объемов данных;

– ограничения в распределенных системах, обусловленные CAP-теоремой;

– платформа Hadoop и ее ограничения для построения систем потоковой обработки данных;

– современные системы, предоставляющие возможности потоковой обработки;

– lambda-архитектура как способ объединения потоковой и пакетной обработки данных.

Была предложена и реализована архитектура новой платформы для потоковой обработки большого объема данных. Платформа ориентирована работу в режиме высоких нагрузок. Для реализации были использованы продукты Apache: Storm, Kafka, Mesos. В рамках реализации платформы было разработано следующее ПО:

– система управления потоком данных с наличием инфраструктуры для мониторинга метрик и логирования;

– компонент для получения данных из Apache Kafka;

– библиотека компонентов для решения наиболее частых сценариев использования;

– компонент для поддержки взаимодействия платформы с приложениями, написанными под .NET.

Проведено тестирование соответствия платформы заявленным требованиям по масштабируемости и отказоустойчивости. На базе платформы было разработано и внедрено прикладное пользовательское приложение.

Результаты исследования используются в производстве, о чем свидетельствует справка о внедрении. В рамках магистерской работы был сделан ряд докладов для конференций (Минск, Львов, Вильнюс, Лондон), опубликованы тезисы выступлений, а также проведен мастер-класс на международной конференции по Big Data в Минске.

Программа исследований была выполнена в полном объеме и в соответствии с установленными сроками.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Божко, С.С. Некоторые алгоритмы потоковой обработки данных / С.С. Божко, И.И. Пилецкий, К.Ю. Слисенко // BIG DATA and Predictive Analytics. Использование BIG DATA для оптимизации бизнеса и информационных технологий: сборник материалов междунар. науч.-практ. конф., Минск, 16–19 июня 2015 г. / редкол.: М.П. Батура [и др.]. – Минск: БГУИР, 2015. – С. 201–205.

2. Божко, С.С. Задача оптимизации работы кластера с помощью алгоритмов машинного обучения / К.Ю. Слисенко, С.С. Божко, С.И. Сиротко // BIG DATA and Predictive Analytics. Использование BIG DATA для оптимизации бизнеса и информационных технологий: сборник материалов междунар. науч.-практ. конф., Минск, 16–19 июня 2015 г. / редкол.: М.П. Батура [и др.]. – Минск: БГУИР, 2015. – С. 122–125.

3. Божко, С.С. Практика построения высоконагруженных систем / С.С. Божко, И.И. Пилецкий // 51-я научная конференция аспирантов, магистрантов и студентов по направлению 4: Компьютерные системы и сети: материалы конф., Минск, 13–17 апреля 2015 г. / редкол.: В.А. Прытков [и др.]. – Минск: БГУИР, 2015. – С. 154–156.

4. Божко, С.С. Высоконагруженный игровой сервер / С.С. Божко, К.Д. Яшин // 50-я научная конференция аспирантов, магистрантов и студентов по направлению: Компьютерное проектирование и технология производства электронных систем: материалы конф., Минск, 24–28 марта 2014 г. / редкол.: М.П. Батура [и др.]. – Минск: БГУИР, 2014. – С. 54.

СПИСОК КОНФЕРЕНЦИЙ, НА КОТОРЫХ БЫЛИ СДЕЛАНЫ ДОКЛАДЫ ВО ВРЕМЯ РАБОТЫ НАД ДИССЕРТАЦИЕЙ

1. TeamsPark TechWeek в БГУИР / «Высоконагруженные проекты: что нужно знать, чтобы этим заниматься» / 23–26 марта 2015 г. – Беларусь, Минск.

2. 15-ая встреча белорусских Scala-разработчиков «ScalaBy #15» / «Переход на Scala: полученный опыт» / 27 апреля 2015 г. – Беларусь, Минск.

3. Международная научно-практическая конференция «BIG DATA and Predictive Analytics» / «Некоторые алгоритмы потоковой обработки данных» / 6–19 июня 2015 г. – Беларусь, Минск.

4. ScalaUA #9 Conference / «Переход на Scala: полученный опыт» / 11 октября 2015 г. – Украина, Львов.

5. Specific Java Day – JVM technologies meetup / «Переход на Scala: полученный опыт» / 17 октября 2015 г. – Беларусь, Минск.

6. Functional Vilnius meetup #6 / «Golang from Scala developer's perspective» / 21 октября 2015 г. – Литва, Вильнюс (доклад на английском языке).

7. Первая белорусская встреча Golang-разработчиков / «Golang с точки зрения Scala разработчика» / 8 декабря 2015 г. – Беларусь, Минск.

8. Final Selection Day at «Entrepreneur First» / «Steam Data Processing Platform» / 16 января 2016 г. – Великобритания, Лондон (доклад на английском языке).