МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра математического обеспечения вычислительных комплексов и информационных систем

РЕАЛИЗАЦИЯ ВЫСОКОЭФФЕКТИВНОЙ ETL-СИСТЕМЫ С ИСПОЛЬЗОВАНИЕМ СРЕДСТВ BIG DATA

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 273 группы
направления 01.04.02 Прикладная математика и информатика
факультета компьютерных наук и информационных технологий
Ванюкова Александра Николаевича

Научный руководитель:
Д. ф.-м. н. Д.К. Андрейченко
(подпись, дата)

Зав. кафедрой:
Д. ф.-м. н. Д.К. Андрейченко
подпись, дата

Саратов 2019

ВВЕДЕНИЕ

Актуальность темы. Современный мир с его многообразием умных устройств генерирует огромное количество данных, и объем этих данных будет только продолжать стремительно расти от года к году. Для сбора и хранения, а также обработки, такого массива данных требуются мощные кластеры и программные системы способные работать с ними. Кроме того накопленные данные имеют крайне разнородную структуру и не могут быть напрямую использованы для автоматизированной обработки. Это приводит к тому, что современные аналитические системы включают обширный комплекс подсистем, предназначенных для извлечения данных из разнотипных источников для дальнейшей обработки.

Как следствие, ранее описанная проблема обуславливает необходимость использования специального инструментария для извлечения данных из источников различного формата, их преобразования, очистки, обобщения и размещения в хранилище данных. Такой комплекс программных средств получил обобщенное название ETL (от англ. extraction, transformation, loading – «извлечение», «преобразование», «загрузка»). Сам процесс переноса данных и связанные с ним действия называются ETL-процессом, а соответствующие программные средства – ETL системами. Одним из таких комплексов подсистем является набор программных средств получившее обобщенное название ETL (от англ. extraction, transformation, loading – «извлечение», «преобразование», «загрузка»). Для реализации такой подсистемы, удовлетворяющей требованиям работы с большими потоками данных, на помощь приходит область Big Data, которая помогает эффективно собирать, хранить и обрабатывать большие наборы данных.

Более того Big Data стала играть ключевую роль в ведении современного бизнеса. Крупные компании стремятся к внедрению технологий Big Data для получения дополнительного преимущества над своими конкурентами, выявления новых направлений деятельности, и принятию наиболее верных стратегических решений.

Целью данной магистерской работы является построение высокоэффективной ETL-системы средствами Big Data. Для достижения данной цели будет решен следующий набор задач:

- Изучение и определение понятия ETL-системы;

- Изучение области Big Data;
- Разбор основных технологий применяемых при решении задач области Big Data:
  - Apache Hadoop и HDFS;
  - Apache Flume;
  - Apache Hive;
  - Apache Sqoop;
  - Apache Spark;
  - Apache Airflow;
- Проектирование ETL-системы;
- Реализация источника данных для тестирования спроектированной ETL-системы;
- Конфигурирование и исполнение Apache Flume;
- Реализация этапов ETL-процесса средствами Apache Hive;
- Реализация этапов ETL-процесса средствами Apache Spark;
- Интеграция ETL-процессов с оркестратором Apache Airflow;
- Выполнения сравнительного анализа MapReduce и Spark решений.

Целью магистерской работы – является построение высокоэффективной ETL-системы средствами Big Data.

Поставленная цель определила **следующие задачи**:

1. Изучение и определение понятия ETL-системы;
2. Изучение области Big Data;
3. Разбор основных технологий применяемых при решении задач области Big Data (Apache Hadoop и HDFS, Apache Flume, Apache Hive, Apache Sqoop, Apache Spark, Apache Airflow);
4. Проектирование ETL-системы;
5. Реализация источника данных для тестирования спроектированной ETL-системы;
6. Конфигурирование и исполнение Apache Flume;

7. Реализация этапов ETL-процесса средствами Apache Hive;
8. Реализация этапов ETL-процесса средствами Apache Spark;
9. Интеграция ETL-процессов с оркестратором Apache Airflow;
10. Выполнения сравнительного анализа MapReduce и Spark решений.

Методологические основы реализации ETL-систем и особенностей использования средств Big Data представлены в работах Thomas Erl, Wajid Khattak, Paul Buhler [1], Паклин Н.Б., Орешков В.И [18], Клеппман М. [24], Холден Карау, Рейчел Уорреу [25], Мартин Одерски, Лекс Спун и Билл Веннерс [27].

Теоретическая значимость бакалаврской работы. В процессе изучения области Big Data были представлены основные концепции современных технологий этой области. Проведенный сравнительный анализ MapReduce и Spark подходов раскрывает существенные различия этих подходов, а наличие цифр дает возможность определиться в выборе одного из подходов.

Практическая значимость бакалаврской работы. Практической значимостью данной работы является возможность пере использования применяемых техник и комбинаций технологий при построении коммерческих ETL-систем. А проведенный сравнительный анализ по способствует выбору правильного подхода.

Структура и объём работы. Магистерская работа состоит из введения, 3 разделов, заключения, списка использованных источников и 3 приложений. Общий объем работы – 98 страниц, из них 76 страниц – основное содержание, включая 25 рисунков и 5 таблиц, список использованных источников информации – 33 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «ETL-Системы и область Big Data**»** посвящен области Big Data, основным ее концепциям и определениям, а также определению понятия ETL-системы и все что с ним связано. Раздел включает в себя несколько подразделов.

Подраздел **«**Определение ETL-Системы**»** содержит определение понятия ETLсистемы и проблем, решаемых ею.

Подраздел **«**Цели и задачи ETL-Систем**»** показывает какие цели и задачи решает ETL-системы, а также из каких процессов состоит такая система.

Подраздел **«**Извлечение данных в ETL**»** описывает процесс извлечения данных ETL-системой.

Подраздел **«**Преобразование и загрузка данных в ETL**»** описывает процессы преобразования и загрузки данных, а именно, что из себя представляет каждый из этих процессов, какие операции могут быть в них включены. Довольно подробно описывается процедура по очистке данных и типичные проблемы с данными, которые необходимо устранить на этапе очистки.

Подраздел **«**Область BIG DATA**»** содержит определение области Big Data.

Подраздел **«**Концепции и терминология Big Data**»** включает в себя основные концепции и понятия, относящиеся к области Big Data,

Подраздел **«**Характеристики Big Data**»** описывает характеристики присущие данным обрабатываемым в области Big Data.

Второй раздел «Технологии Big Data**»** посвящен описанию основных технологий, применяемых сегодня при решении задач Big Data.

Подраздел **«**Apache Hadoop и HDFS**»** включает в себя описание такой технологии как Apache Hadoop и ее основных компонентов.

Подраздел **«**Apache Flume**»** представляет собой описание технологии Apache Flume, ее конфигурацию и способы запуска.

Подраздел «Apache Hive» описывает технологию Apache Hive, а именно ее возможности и способы работы с ней.

Подраздел «Apache Sqoop» представляет собой описание технологии Apache Sqoop, ее конфигурацию и способы запуска, а также задачи решаемы ею.

Подраздел «Apache Spark» содержит описание универсальной и высокопроизводительной кластерной вычислительной платформы Apache Spark, а именно ее архитектуры, основных компонентов и модели распределенных вычислений.

Подраздел «Apache Airflow» освещает вопрос о планировщике Apache Airflow, его основных абстракциях и задачах, решаемых им.

Третий раздел «Построение ETL-системы и сравнение MapReduce и Spark решений» посвящен реализации ETL-систем в двух вариациях с применением Apache Hive и Apache Spark и сравнительному анализу двух получившихся решений.

Подраздел «Описание архитектуры ETL-системы» содержит непосредственно описание архитектуры ETL-системы. Любая ETL система включает в себя три основных этапа это загрузка, трансформация и сохранения данных. Наша ETL-система будет строится с применением следующих технологий, а именно Apache Flume, Apache Airflow, Apache Hive, Apache Spark и Hadoop. В системе будет реализован следующий набор преобразований:

- Очистка данных (подход к очистке будет описан позднее);
- Создание новых данных;
- Агрегирование данных;

Подраздел «Реализация источника данных» включает в себя информацию по реализации собственного источника данных. Собственный source, имеет название EventProducer.

EventProducer должен будет генерировать информацию о покупках и отправлять ее по TCP в некоторый сервис, в нашем случае принимающим сервисом будет выступать Apache Flume.

Подраздел «Конфигурация и запуск Apache Flume» описывается настройка Apache Flume, а именно следующего набора компонентов source, channel и sink.

Подраздел «Реализация Hive ETL» описывает реализацию ETL средствами Apache Hive. Реализация состоит из следующих шагов: реализации собственной UDF функции, написании HQL скриптов и реализации Airflow DAG.

Подраздел «Реализация Spark ETL» описывает реализацию ETL средствами Apache Spark. Реализация состоит из следующих шагов: реализации собственной Spark jobs на Scala и реализации Airflow DAG.

Подраздел «Сравнение MapReduce и Spark решений» в данном подразделе происходит сравнение двух реализованных подходов на Apache Hive и Apache Spark. Приводится описание эти двух технологий, сравниваются их возможности и приводятся результаты тестирования.

ЗАКЛЮЧЕНИЕ

В рамках данной выпускной квалификационной работы были решены все поставленные задачи — изучена область Big Data и технологии применяемые в ней, также изучена концепция ETL-системы, описаны ее основные этапы работы, подходы к очистке данных и распространенные проблемы с ними.

Была спроектирована ETL-система. Спроектированная ETL-система включает в себя 8 этапов обработки информации, 7 этапов из которых могут выполняться параллельно и являют собой агрегацию и создание новых данных, а оставшийся этап представляет собой техническую очистку данных.

Был разработан источник данных для тестирования спроектированной ETL-системы и сконфигурирован Apache Flume для трансляции сгенерированных данных в HDFS.

Кроме того ETL-система была реализована в двух вариациях, используя SQL и JVM-ориентированные парадигмы [2], а именно реализации на основе Apache Hive и Apache Spark с интеграцией с Apache Airflow.

Также в результате был выполнен сравнительный анализ двух получившихся решений - основанных на MapReduce и Spark подходах. В результате было выполнено тестирование решений на двух наборах данных, объемом в 1Гб и 10Гб соответственно. MapReduce подход оказался значительно медленнее, чем подход с использованием Spark. В некоторых случаях разница составляла 10 кратный разрыв между этими двумя подходами. Такую разницу можно объяснить более современных подходом к обработке данных, использующимся в Spark, основанным на отложенной обработке данных с планированием и их обработкой непосредственно в памяти компьютера. Если говорить про MapReduce, то основной его проблемой является частая запись и считывание данных на диск при выполнении Map и Reduce задач [7].

Таким образом при разработке ETL-системы предпочтительно использовать технологию Spark.

СПИСОК ИСТОЧНИКОВ И ЛИТЕРАТУРЫ

1. Thomas Erl, Wajid Khattak, and Paul Buhler – Big Data Fundamentals Concepts, Drivers & Techniques, 2016, ISBN-13: 978-0-13-429107-9
2. Chao Wang – High Performance Computing for Big Data, 2018, ISBN: 978-1-4987-8399-6.
3. Arun K. Somani – Big Data Analytics: Tools and Technology for Effective Planning, 2018, ISBN: 978-1138032392.
4. Kuan-Ching Li – Smart Data: State-of-the-Art Perspectives in Computing and Applications, 2018, ISBN: 978-1138032392.
5. Tom White - Hadoop: The Definitive Guide, Fourth Edition, 2015, ISBN: 978-1-491-90163-2.
6. Hrishikesh Karambelkar – Apache Hadoop 3 Quick Start Guide: Learn about big data processing and analytics, 2018, ISBN: 978-1-788-99983-0.
7. Mark Grover, Ted Malaska, Jonathan Seidman, and Gwen Shapira - Hadoop Application Architectures, 2015, ISBN: 978-1-491-90008-6.
8. Deepak Vohra - Practical Hadoop Ecosystem, 2016, ISBN-13 (electronic): 978-1-4842-2199-0.
9. Hari Shreedharan - Using Flume, 2015, ISBN: 978-1-449-36830-2.
10. Flume 1.8.0 User Guide [Электронный ресурс]. – Режим доступа: https://flume.apache.org/FlumeUserGuide.html – Дата обращения: 23-05-2019.
11. Edward Capriolo, Dean Wampler, and Jason Rutherglen - Programming Hive, 2012, ISBN: 978-1-449-31933-5.
12. Wiki Apache Hive [Электронный ресурс]. – Режим доступа: https://cwiki.apache.org/confluence/display/Hive – Дата обращения: 23-05-2019.
13. Kathleen Ting and Jarek Jarcec Cecho - Apache Sqoop Cookbook, 2013, ISBN: 978-1-449-36462-5.

14. Sqoop User Guide [Электронный ресурс]. – Режим доступа: http://sqoop.apache.org/docs/1.4.7/SqoopUserGuide.html – Дата обращения: 23-05-2019.
15. Карау Х., Конвински Э., Венделл П., Захария М. — Изучаем Spark: молниеносный анализ данных. - М.: ДМК Пресс, 2015. - 304 с.: ил. ISBN: 978-5-97060-323-9.
16. Quick Start — Airflow Documentation [Электронный ресурс] – Режим доступа: https://airflow.apache.org/start.html – Дата обращения: 23-05-2019.
17. Mike Frampton - Mastering Apache Spark, 2015, ISBN: 978-1-78398-714-6.
18. Паклин Н.Б., Орешков В.И. — Бизнес-Аналитика: от данных к знаниям. Учебное пособие 2-е издание., испр. – СПб.: Питер, 2013. – 704 с.: ил. ISBN 978-5-459-00-717-6.
19. Консолидация данных – ключевые понятия [Электронный ресурс] – Режим доступа: https://www.cfin.ru/itm/olap/cons.shtml?printversion – Дата обращения: 23-05-2019.
20. Найдич А. Big Data: проблема, технология, рынок [Электронный ресурс]. Режим доступа: http://compress.ru/article.aspx?id=22725 – Дата обращения: 23-05-2019.
21. GeoIP2 Precision Web Services [Электронный ресурс]. – Режим доступа: https://dev.maxmind.com/geoip/geoip2/web-services/ – Дата обращения: 23-05-2019.
22. Russel Jurney - Agile Data Science 2.0, 2017, ISBN: 978-1-491-96011-0.
23. Силен Дэви, Мейсман Арно, Али Мохамед — Основы Data Science и Big Data. Python и наука о данных. - СПб.:Питер, 2017.- 336 с.: ил. - (Серия «Библиотека программиста») ISBN: 978-5-496-02517-1.

24. Клеппман М. — Высоконагруженные приложения. Программирование, масштабирование, поддержка. — СПб.: Питер, 2018. — 640 с.: ил. — (Серия «Бестселлеры O’Reilly»). ISBN: 978-5-4461-0512-0.
25. Холден Карау, Рейчел Уоррен — Эффективный Spark. Масштабирование и оптимизация. — СПб.: Питер, 2018. — 352 с.: ил. — (Серия «Бестселлеры O’Reilly»). ISBN: 978-5-4461-0705-6.
26. Sandy Ryza Advanced Analytics with Spark, 2017, ISBN: 978-1-491-97295-3.
27. Мартин Одерски, Лекс Спун, Билл Веннерс - Scala Профессиональное программирование. - СПб.: Питер, 2017. ISBN: 978-5-496-02951-3.
28. Хорстманн К. — Scala для нетерпеливых / пер. с англ. А. Н. Киселева – 2-е изд. – М.: ДМК Пресс, 2017. – 414 с.: ил. ISBN: 978-5-97060-536-3.
29. Pascal Bugnion — Scala for Data Science, 2016, ISBN: 978-1-78528-137-2.
30. Rishi Yadav - Spark Cookbook, 2015, ISBN: 978-1-78398-706-1.
31. Mohammed Guller – Big Data Analytics with Spark, 2015, ISBN-13 (electronic): 978-1-4842-0964-6.
32. Petar Zecevic, Marko Bonaci – Spark in Action, 2017, ISBN: 978-1-617-29260-6.
33. Apache Airflow — Национальная библиотека им. Н. Э. Баумана [Электронный ресурс] – Режим доступа: https://ru.bmstu.wiki/Apache_Airflow – Дата обращения: 23-05-2019.