**МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ**

**УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ**

**ГОМЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ П. О. СУХОГО**

Факультет автоматизированных и информационных систем

Кафедра «Информатика»

ОТЧЕТ ПО ЛАБОРАТОРНОЙ РАБОТЕ

по дисциплине **«ИТРОД**»

на тему: «**Анализ проведенной работы»**

Выполнил: студенты гр. ИП-41

Сачек Д.Н

Маланичева А.С.

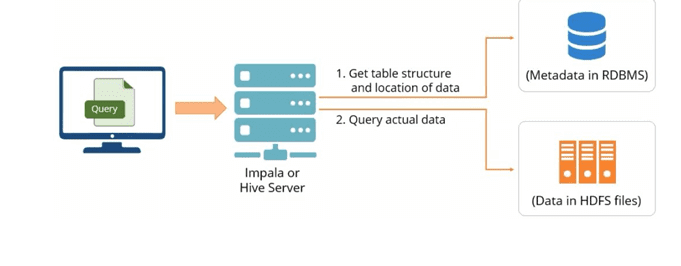
Принял: преподаватель

Шибеко В.Н.

Гомель 2020

1. Обеспечение масштабируемости и живучести, согласованности.

В Hive масштабируемость обеспечивается благодаря кластеру, состоящему из множества Data node, и курирующего их Name node.

Сам Hive Server использует отдельный Node для хранения метаданным о структуре БД. А сами данные хранятся на Data Node. ****

Живучесть обеспечивается механизмом репликации. Name Node постоянно следит за состоянием Data Node и в случае выхода из строя некоторой машины запускает механизм репликации данных.

Из вышесказанного следует и что за согласованность данных отвечает также Name Node.

1. Менеджмент CRUD операций.

Hive определяет простой [SQL](https://ru.bmstu.wiki/SQL_(Structured_Query_Language))-подобный язык запросов, называемый HiveQL (Hive Query Language), что позволяет пользователям, знакомым с [SQL](https://ru.bmstu.wiki/SQL_(Structured_Query_Language)) легко работать с запросами к данным. HiveQL имеет функции для работы с форматами XML и JSON, поддержку нескалярных типов данных, таких как массивы, структуры, ассоциативные массивы, поддерживает широкий набор агрегирующих функций, определяемые пользователем функции (User Defined Functions), блокировки. В то же время, этот язык позволяет программистам, которые знакомы со структурой MapReduce, подключить свои пользовательских карты для выполнения более сложного анализа, который может быть не поддерживаемым встроенными возможностями языка. HiveQL может быть расширен с пользовательскими скалярных функций (UDF), агрегаций (UDAF годов), и табличных функций (UDTF).

Однако, по-умолчанию, для поддержки операций обновления и удаления необходима дополнительная настройка и определенная структура таблиц.

1. Применение hadoop для big data

Используется для реализации поисковых и контекстных механизмов многих высоконагруженных веб-сайтов, в том числе, для [Yahoo!](https://ru.wikipedia.org/wiki/Yahoo!) и [Facebook](https://ru.wikipedia.org/wiki/Facebook)[[3]](https://ru.wikipedia.org/wiki/Hadoop#cite_note-_a27d67d55a679a1c-3). Разработан на [Java](https://ru.wikipedia.org/wiki/Java) в рамках вычислительной парадигмы [MapReduce](https://ru.wikipedia.org/wiki/MapReduce), согласно которой приложение разделяется на большое количество одинаковых элементарных заданий, выполнимых на узлах кластера и естественным образом сводимых в конечный результат.

По состоянию на 2014 год, проект состоит из четырёх модулей — Hadoop Common[[⇨]](https://ru.wikipedia.org/wiki/Hadoop#Hadoop_Common) ([связующее программное обеспечение](https://ru.wikipedia.org/wiki/%D0%A1%D0%B2%D1%8F%D0%B7%D1%83%D1%8E%D1%89%D0%B5%D0%B5_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D0%B5%D1%81%D0%BF%D0%B5%D1%87%D0%B5%D0%BD%D0%B8%D0%B5) — набор инфраструктурных программных библиотек и утилит, используемых для других модулей и родственных проектов), HDFS[[⇨]](https://ru.wikipedia.org/wiki/Hadoop#HDFS) ([распределённая файловая система](https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D1%91%D0%BD%D0%BD%D0%B0%D1%8F_%D1%84%D0%B0%D0%B9%D0%BB%D0%BE%D0%B2%D0%B0%D1%8F_%D1%81%D0%B8%D1%81%D1%82%D0%B5%D0%BC%D0%B0)), YARN[[⇨]](https://ru.wikipedia.org/wiki/Hadoop#YARN) (система для планирования заданий и управления кластером) и Hadoop MapReduce[[⇨]](https://ru.wikipedia.org/wiki/Hadoop#Hadoop_MapReduce) (платформа программирования и выполнения распределённых MapReduce-вычислений), ранее в Hadoop входил целый ряд других проектов, ставших самостоятельными в рамках системы проектов Apache Software Foundation.

1. Размещение в Docker

Кластер развертывается с помощью docker-compose, начальные настройки хранятся в hadoop-hive.env. После развертывания расширение кластера возможно с помощью изменения docker-compose файла.

Достоинства:

Достоинства этого подхода, этой платформы. Очень хорошая и гладкая масштабируемость. Т.е. если нам нужно обработать в два раза больше данных, или хранить в два раза больше данных, нам достаточно добавить в два раза больше машин в кластер. Т.е. не совсем ровно в два раза, но почти в два раза.

Нулевая стоимость software. Можно взять кластер и использовать open-source Hadoop, как систему анализа и хранения данных.

Недостатки:

Во-первых, это довольно высокая стоимость поддержки. Если у вас есть Hadoop-кластер, из многих машин, вам нужно найти умного системного администратора, который разберется в архитектуре Hadoop-а, в том, как это работает, и будет все это поддерживать. Т.е. это действительно непросто, это действительно занимает много времени.

Вам нужны будут люди, которые будут придумывать бизнес-часть, какие именно данные им нужны, и вам будет нужна команда Java-разработчиков, которые будут писать эти map-reduce job.

Проблема с real-time-ом. Hadoop это не real-time система. Если вы хотите получать какие-то данные, у вас не получится так, что вы будете запускать map-reduce job-ы, когда пользователь заходит на сайт. Вам нужно обновлять данные, хоть раз час, в background-е, а пользователю показывать уже рассчитанные данные.