**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ (НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)»**

**Журнал практики**

Студент  **Бобряков Александр Сергеевич**

Институт №8 Информационные технологии и прикладная математика

Кафедра 810Б «Информационные технологии в моделировании и управлении»

Учебная группа М8О-203М-19

Направление подготовки (специальность)

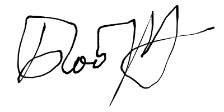
**02.04.02 «Фундаментальная информатика и информационные технологии»**

Вид практики**Преддипломная**

Руководитель практики от МАИ

**д.ф.-м.н. Абгарян К.К.** **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Подпись руководителя практики

Студент-практикант

**Бобряков А.С.**  « 03 » февраля 2020г.

Подпись

**Москва 2021**

1. **Место и сроки проведения практики**

Сроки проведения практики:

Дата начала практики: 03.02.2021

Дата окончания практики: 20.05.2021

Структурное подразделение: **МАИ, кафедра 810Б «Информационные технологии**

**в моделировании и управлении»**

1. **Инструктаж по технике безопасности**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ /Ганичева А.К.** «03» февраля 2020г.

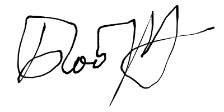
Подпись проводившего инструктаж Дата проведения

1. **Индивидуальное задание студенту**

* Разработать map-reduce задачу для возможности ее запуска с изменением ключевых параметров в несколько потоков
* Подготовить таблицы с данными в Hbase для возможности их обработки задачей
* Собрать обучающий набор данных к магистерской диссертации на основе запусков задачи и сбора статистики

1. **План выполнения индивидуального задания**

|  |  |  |
| --- | --- | --- |
| № пп | Наименование раздела или этапа | Срок выполнения |
| 1 | Разработка map-reduce задачи | 01.04.2021 |
| 2 | Подготовка окружения и базы данных, наполнение их информацией | 18.12.2021 |
| 3 | Запуск задачи и сбор метрик | 20.04.2021 |
| 4 | Фильтрация полученного набора данных для дальнейшего обучения | 11.05.2021 |

Cоруководитель практики от МАИ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ / **д.ф.-м.н. Абгарян К.К.**

Студент-практикант \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ **/Бобряков А.С.**/

« 03 » февраля 2020г.

1. **Отзыв руководителя практики**

Студент Бобряков А.С. хорошо справился со своими обязанностями в ходе преддипломной практики. Грамотно разбил задачу на составляющие, придерживался сроков плана и постоянно отчитывался о проведённых занятиях.

Безусловно, студент проделал большую и качественную работу, а также провёл достаточную подготовку информации к магистерской диссертации.

В целом, практика выполнена на очень хорошем уровне и соответствует всем предъявляемым требованиям. Информация, изложенная в отчёте студента, полностью соответствует индивидуальному заданию.

Соруководитель практики от МАИ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ / **д.ф.-м.н. Абгарян К.К,**

«03» февраля 2021г.

1. **Отчет студента о практике**

В начале преддипломной практики была определена ее цель, которая включает разработку map-reduce задачи, наполнение базы данных, а также многократный запуск задачи для сбора статистики и подготовки обучающего набора данных. Были выполнены следующие этапы:

1. Разработка map-reduce задачи

Задача, запускаемая в системе Hadoop, состоит из этапов, отображенных на рисунке 1. Основными этапами являются этап отображения (map) и этап свертки (reduce).

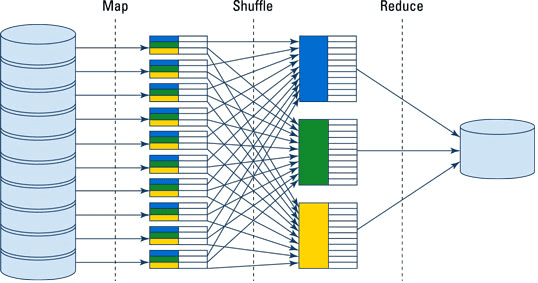


Рисунок 1 – Общая последовательность этапов запуска map-reduce задачи

В качестве map-reduce задачи для генерирования данных была разработана программа на языке Java. Задача выполняет сверку данных между двумя таблицами в базе данных Hbase.

Для ускорения сбора данных осуществлена возможность запуска задачи в несколько потоков на разных таблицах. При этом запускаемые задачи не влияют на скорость исполнения друг друга. При каждом запуске изменяются следующие характеристики: maxMappers, maxReducers, mapreduceTaskIoSortFactor, mapreduceTaskIoSortMb, reducejavaOpts, mapreduceMapMemoryMb, mapreduceReduceMemoryMb, mapreduceMapCpuVcores, mapreduceReduceCpuVcores, mapjavaOpts.

На вход задаче поступают все строки из двух таблиц базы данных Hbase. На этапе отображения каждой подзадаче поступает на вход строка таблицы, а также название самой таблицы. Происходит считывание идентификатора записи для дальнейшего определения на какой процесс reduce отправить текущую запись. На этап reduce поступают все строки по соответствующим идентификаторам. Таким образом, записей будет либо 1, либо 2. Далее происходит рекурсивное сравнение всех полей между записями с одинаковыми идентификаторами. Все различные поля записываются в файловую систему HDFS. Изначальная задача заканчивается, когда будут сравнены все строки двух таблиц. Для простоты анализа на вход каждой запускаемой задачи поступала одна таблица, которая сверялась сама с собой.

После окончания сверки каждой таблицы снимается метрика о длительности исполнения задачи, которое сохраняется в файл со всеми остальными изменяемыми параметрами,

1. Наполнение базы данных

В базе данных Hbase было создано тестовое пространство имен. В нем было создано 200 таблиц, которые были разделены на различное количество регионов. Количество регионов варьируется в пределах от 1 до 1000. Таким образом количество регионов стало еще одним отслеживаемым параметров, который может влиять на скорость сверки данных. В итоге было собрано 15000 запусков с изменением параметров, приведенных на рисунке 2.

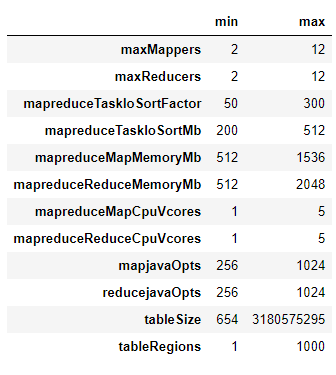


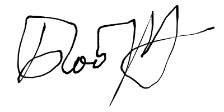
Рисунок 2 – Перечень изменяемых параметров с их min/max значениям при проведении экспериментальных запусков

1. Сбор статистик и подготовка обучающего набора данных

После запусков задачи и получения статистик была проведена подготовка и очистка данных для дальнейшего обучения модели (нейронной сети). Таким образом были убраны “всплески” данных, убраны дубликаты и очищены строки, содержащие нерелевантные данные, которые могли появиться в следствие возникновения ошибок при запуске задачи. Такие ошибки возникают при сбоях в работе кластера и зависят от инфраструктуры, а не от логики исполняемой задачи.

**Выводы**

Проделанная работа в ходе преддипломной практики позволила собрать набор обучающих данных, который является достаточным для обучения модели машинного обучения для предсказания времени выполнения задачи. Во время многократных запусков задачи варьировались 12 основных параметров в приемлемых диапазонах, которые влияют на снимаемую целевую метрику. Это позволяет провести дальнейший анализ оптимальной настройки параметров для запускаемых задач в системе Hadoop.

Студент-практикант

**Бобряков А.С.**  « 04» января 2020г.