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

УДК 004.75

ЖАКСЫЛЫК
Куаныш

АРХИТЕКТУРА ВЫСОКОНАГРУЖЕННОЙ РАСПРЕДЕЛЕННОЙ СИСТЕМЫ ПОТОКОВОЙ ОБРАБОТКИ ДАННЫХ ДЛЯ ЗАДАЧ РАСПОЗНАВАНИЯ РЕЧИ

Автореферат
на соискание степени магистра технических наук
по специальности 1-40 80 02 Системный анализ, управление и обработка информации (по отраслям)

_____

Научный руководитель
кандидат технических наук,
доцент кафедры СУ
ЗАХАРЬЕВ Вадим Анатольевич

_____

Минск 2024

ВВЕДЕНИЕ

В эпоху повсеместного распространения цифровых коммуникаций и распространения интеллектуальных устройств спрос на эффективные и точные системы распознавания речи достиг беспрецедентного роста. Поскольку человечество ориентируется в мире, в котором все больше используются речевые технологии, производительность и масштабируемость алгоритмов распознавания речи становятся ключевым фактором. Традиционные подходы к распознаванию речи часто сталкиваются с проблемами, возникающими из-за огромного объема и динамического характера аудиоданных в реальном времени. В ответ на эти проблемы интеграция распределенных вычислений и обработки потоков данных стала многообещающей парадигмой, предлагающей потенциал для повышения скорости, точности и масштабируемости систем распознавания речи.

Актуальность данной темы обусловлена стремительным развитием речевых технологий в современном информационном обществе. Распознавания речи активно применяется в различных областях ИКТ, таких как автоматизированные системы управления, виртуальные ассистенты и т.п.

Современные речевые системы и сервисы, обладающие хорошими показателями качества, из-за высокой степени сложности и ресурсоемкости используемых моделей, чаще всего имеют распределенную архитектуру. Следовательно, возникает необходимость в развитии методов проектирования, а также, в создании архитектур и алгоритмов распределённых систем, позволяющих обеспечить: повышение отказоустойчивости подобных систем, увеличение скорости доступа к моделям распознавания речи, находящимся на удалённых ресурсах, возможность их масштабирования в зависимости от нагрузки.

В магистерской диссертации исследуется синергия между распределенными вычислительными системами и методами обработки потоков данных в контексте распознавания речи. Конвергенция этих двух областей обещает устранить присущие моделям централизованной обработки ограничения и открыть новые возможности для крупномасштабного анализа речи в реальном времени. Используя присущую распределенным системам параллелизм и адаптируемость обработки потоков данных, можно преодолеть вычислительные узкие места, связанные с традиционными подходами пакетной обработки.

Целью данной работы является исследование существующих архитектур высоконагруженных распределенных систем и разработка инновационной архитектуры с целью оптимизации времени обработки запросов. Это включает в себя уменьшение времени простоя и ожидания

обработки в очереди, снижение числа отказов, повышение отказоустойчивости системы, а также улучшение характеристик масштабируемости и обслуживания.

В соответствии с указанной целью в работе поставлены и решены следующие задачи:

1. Анализ существующих методов распознавания речи и их применимости к распределенным системам.
2. Разработка архитектуры системы потоковой обработки данных, оптимизированной для эффективного распознавания речи.
3. Разработка механизмов обеспечения масштабируемости системы.
4. Разработка механизмов обеспечения отказоустойчивости и управления ресурсами в распределенной среде.

Объектом исследования является высоконагруженная распределенная система потоковой обработки данных.

Предмет исследования - архитектура системы, специализированная для задач распознавания речи.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Связь работы с крупными научными программами

Тема диссертационной работы соответствует пункту 1 приоритетных направлений научной, научно-технической и инновационной деятельности Республики Беларусь на 2021 –2025 гг., утвержденных Указом Президента Республики Беларусь №156 от 7 мая 2020 г. «Цифровые информационно-коммуникационные и междисциплинарные технологии, основанные на них производства», а также, статье 61 «Государственная поддержка развития отрасли информационно-коммуникационных технологий» главы 11 «Развитие отрасли информационно-коммуникационных технологий» раздела 1 «Основы регулирования отношений в сфере информатизации Республики Казахстан», как описано в Законе Республики Казахстан от 24 ноября 2015 года № 418-V «Об информатизации». Работа выполнялась в учреждении образования «Белорусский государственный университет информатики и радиоэлектроники».

Цель и задачи исследования

Целями данной диссертационной работы являются исследование и разработка архитектуры высоконагруженной распределенной системы, а также сравнение разработанной архитектуры с существующими методами для выявления преимуществ и особенностей.

Для достижения поставленной целей в диссертации решены следующие задачи:

1. Анализ существующих методов распознавания речи и их применимости к высоконагруженным системам.
2. Разработка архитектуры системы потоковой обработки данных, оптимизированной для эффективного распознавания речи.
3. Исследование механизмов обеспечения масштабируемости системы.
4. Разработка механизмов обеспечения отказоустойчивости и управления ресурсами в распределенной среде.

Личный вклад соискателя

Соискателем выполнены все изложенные в работе разработки и исследования. Постановка задач и обсуждение результатов проводились совместно с научным руководителем и сотрудниками кафедры информационных технологии автоматизированных систем и кафедры систем управления Белорусского государственного университета информатики и радиоэлектроники. Соавторы опубликованных работ принимали участие в

обсуждении промежуточных и конечных результатов. Обработка, интерпретация данных, а также выводы сделаны автором самостоятельно.

Апробация результатов диссертации

Основные положения диссертационной работы докладывались на следующих научных конференциях:

– «12-я международная научно-техническая конференция OSTIS» (Минск, 2022)

– «13-я международная научно-техническая конференция OSTIS» (Минск, 2023)

– «59-я научная конференция аспирантов, магистрантов и студентов: системы управления» (Минск, 2023)

– «10-я международная научно-практическая конференция BIGDATA and Advanced Analytics» (Минск, 2024).

Структура и объем диссертации

Диссертационная работа состоит из введения, общей характеристики работы, четырех глав с выводами по каждой главе, заключения, библиографического списка, двух приложений. Общий объем диссертационной работы составляет 134 страницы, из них 51 рисунок, 12 таблиц, список использованных библиографических источников (31 наименование), список публикаций автора по теме диссертации (4 наименования), 2 приложения.

Проверка на уникальность

Проведена экспертиза диссертации Жаксылык Куаныша «Архитектура высоконагруженной распределенной системы потоковой обработки данных для задач распознавания речи» на корректность использования заимствованных материалов с применением сетевого ресурса «Антиплагиат» (адрес доступа: https://antiplagiat.ru) в on-line режиме 17.04.2024 г. В результате проверки установлена корректность использования заимствованных материалов (оригинальность диссертационной работы составляет 94,79 %).

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении рассмотрены проблемы необходимости эффективных и точных систем распознавания речи в эпоху повсеместного распространения цифровых коммуникаций и интеллектуальных устройств. Обсуждается рост спроса на такие системы в связи с расширением использования речевых технологий не только в области автоматизированных систем управления, но и в сфере виртуальных ассистентов, медицинских информационных систем и многих других областях информационно-коммуникационных технологий. В то время как речевые технологии становятся неотъемлемой частью нашего повседневного опыта, традиционные методы распознавания речи часто оказываются недостаточно эффективными для обработки огромного объема и динамического характера аудиоданных, поступающих в реальном времени.

Этот рост спроса на точные системы распознавания речи подчеркивает необходимость поиска новых подходов к их разработке. Именно здесь на передний план выходит интеграция распределенных вычислений и обработки потоков данных. Эта парадигма предоставляет потенциал для повышения скорости, точности и масштабируемости систем распознавания речи. С помощью параллелизма и адаптируемости, характерных для распределенных систем, можно преодолеть вычислительные узкие места, связанные с традиционными подходами пакетной обработки.

В этом контексте магистерская диссертация посвящена исследованию синергии между распределенными вычислительными системами и методами потоковой обработки данных в контексте распознавания речи.

В **общей характеристике работы** показана связь работы с приоритетными направлениями научных исследований, цель и задачи исследования, личный вклад соискателя ученой степени, апробация результатов диссертации.

В первой главе представлен обзор существующих решений в области технологии автоматического распознавания речи (ASR) в распределенных системах. Далее рассматривается прогресс в этой области, включая создание систем, способных распознавать все большее количество слов, таких как «Google Assistant», «Apple Siri» и «Amazon Alexa».

Рассматривается значимость облачных решений в области распознавания речи. Облачные технологии, такие как «Google Cloud Speech-to-Text», «Microsoft Azure Speech Services API», «Amazon Transcribe» и «IBM Watson Speech to Text», предоставляют гибкие инструменты для интеграции распознавания речи в собственные проекты и продукты. Они обладают преимуществами, такими как простой доступ, масштабируемость и высокая точность обработки аудиоданных.

Далее представлен сравнительный обзор трех открытых моделей распознавания речи: «Kaldi», «wav2vec 2.0» и «Whisper». Освещены особенности каждой модели, их преимущества и ограничения. «Whisper» выделяется высокой точностью и удобством использования. Обзор также включает анализ показателей точности, где «Whisper» показывает лучшие результаты.

В конце представляется обзор использования распределенных систем в обработке речевых данных, с фокусом на различиях между пакетной и потоковой обработкой данных. Рассматриваются основные элементы системы потоковой обработки данных, такие как сообщения, генераторы, потоки данных, потребители, операторы и схемы сообщений. Обсуждаются как преимущества данных методов, так и недостатки.

Во второй главе представлен обзор аналитических и имитационных моделей для анализа распределенных систем потоковой обработки данных. Рассматривается суть аналитического моделирования как математического описания структуры и функционирования системы для определения ее эффективности. Особое внимание уделяется моделированию систем массового обслуживания (СМО), которые часто используют аналитические модели для описания поведения системы обслуживания запросов. В контексте распределенных систем выделяется важность сквозной задержки пакета, определяющей время от отправки до получения пакета. Далее описывается модель работы узла в распределенной системе, представленная как очередь M/PH/1, с вычислением среднего количества заданий в очереди и среднего времени обработки. В заключение обсуждается влияние режима репликации на среднюю сквозную задержку пакета.

В конце данной главы описываются имитационные модели, которые используются для создания моделей, наиболее точно отражающих реальные системы. Имитационное моделирование позволяет наблюдать за поведением системы в течение времени, особенно в случае сложных систем. Рассматривается применение имитационных моделей для определения потоков работ, процесса извлечения знаний из данных. Потоки работ представляют собой набор взаимосвязанных вычислительных задач, направленных на достижение конкретной цели, такой как проведение вычислительного эксперимента. Каждое действие в потоке работы представляет собой вычислительную задачу. Рассматривается использование различных моделей для формализации потоков работ, таких как модель ориентированного ациклического графа (DAG) и модель сетей Петри.

В третьей главе обсуждаются типичные две основные модели построения распределённой системы для обработки данных: модель "точка-

точка" и микросервисная модель. Представлен сравнительный анализ этих моделей с целью выявления их характеристик, преимуществ и недостатков.

Также представлен обзор на различные архитектурные подходы к потоковой обработке данных, начиная с двухслойной архитектуры, состоящей из слоя хранения и слоя обработки, и переходя к более расширенной модели, включающей пять слоев: слой потребления данных, слой обработки данных, слой хранения данных, слой управления ресурсами и слой генерации данных. Эта модель предлагает решение для проблем масштабирования, управления ресурсами и отказоустойчивости, с которыми может столкнуться двухслойная архитектура.

Далее спроектирована функциональная схема системы, которая включает в себя ключевые подсистемы, такие как мониторинг и логирование событий, вычислительный кластер для обработки данных и хранения моделей машинного обучения, а также конвейер непрерывной интеграции и непрерывного развертывания системы. В качестве главной платформы для развертывания кластера был использован инструмент оркестрации контейнеризированных приложений «Kubernetes», для обработки потоков данных «Apache Kafka», а для управления жизненным циклом модели «mlflow». В конце главы описаны алгоритмы работы системы, которые включат себя алгоритм распознавания речи в реальном времени и алгоритм конвейера непрерывной интеграции и непрерывного развертывания (CI/CD) системы.

В четвертой главе данной диссертации был представлен разработанный программный прототип и описывался процесс развёртывания тестового вычислительного кластера на облачном провайдере «Hetzner Cloud». Акцентировалось внимание на конфигурации всех необходимых механизмов и параметров, обеспечивающих отказоустойчивость и автомасштабирование кластера. Процесс установки необходимых программных компонентов осуществлялся при помощи «Helm» манифестов, выраженных на языке «YAML».

Далее, былы разработаны программные модули для серверной части системы и алгоритма распознавания речи, реализованных на языке программирования «Python». В рамках экспериментального исследования было проведено пять основных экспериментов. Один из них был посвящён тестированию работоспособности кластера, а три других - проведены непосредственно на брокере сообщений «Apache Kafka». Последний эксперимент был связан с оценкой производительности алгоритма распознавания речи.

ЗАКЛЮЧЕНИЕ

В данной диссертационной работе была разработана архитектура высоконагруженной системы потоковой обработки данных для задач распознавания речи. В процессе исследования предметной области была обнаружена высокая актуальность, что поток информации постоянно растет, а взаимодействие с технологиями голосового управления становится все более распространенным.

Сделан обзор на основные аналитические и имитационные модели в распределенных системах. Аналитическая модель охватывает три типа потоков запросов с соответствующими временами обслуживания, представленными распределениями фазового типа. Имитационная модель интерпретируется через модели потоков работ, которые представляют собой ориентированный ациклический граф.

Представлена типичная архитектура потоковой обработки данных, выявлены ее недостатки и предложена модифицированная архитектура на основе микросервисного подхода. Разработана обобщенная структурная схема. В качестве главного инструмента для управления распределенной системой был предложен вариант использования «Kubernetes» как платформы для оркестрации контейнеризированных приложений. Сделан обзор на основные компоненты «Kubernetes», а также на механизмы масштабируемости и отказоустойчивости. Масштабируемость достигается за счет вертикального, горизонтального и кластерного масштабирования. Отказоустойчивость достигается за счет применения консенсусного алгоритма «Raft». Также были рассмотрены компоненты системного мониторинга, а именно «Prometheus», «Grafana», «Tempo», «Loki». Предложен вариант использования инструмента «MLOps mlflow» для управления жизненным циклом модели. Были описаны алгоритмы работы системы, включая алгоритмы «Kubernetes» и «Apache Kafka», основанные на алгоритме распределенного консенсуса «Raft», а также алгоритмы обработки речи от получения исходного сигнала до передачи спектрограммы на вход модели «Whisper». Была представлена соответствующая схема реализации.

Тестирование архитектуры осуществлялось в кластере «Kubernetes» на вычислительных машинах облачного провайдера «Hetzner Cloud». Были развернуты соответствующие программные пакеты при помощи манифестов «Helm». Произведена отладка кластера и его тестирование при помощи утилиты «Sonobuoy». Было проведено 125 тестов на 3 узлах кластера. Для кластера «Apache Kafka» было произведено 3 эксперимента: анализ партиционирования и пропускной способности «Apache Kafka», анализ производительности создания партиций, анализ производительности

переназначения партиций. Анализ производительности создания партиций показал, что создание числа партиций происходит более эффективно с использованием «KRaft» по сравнению с «ZooKeeper». Время создания партиций увеличивается линейно для «ZooKeeper» (O(n)), в то время как для «KRaft» оно остается почти постоянным (O(1)). Существует ограничение примерно в 80 000 партиций как для кластеров «ZooKeeper», так и для «KRaft». Анализ производительности переназначения партиций показал значительную разницу во времени переназначения партиций между кластерами «ZooKeeper» и «KRaft». Время переназначения составило 600 секунд для «ZooKeeper» и всего 42 секунды для «KRaft».

Эксперимент по исследованию распознавания речи в реальном времени проводился на корпусе речевых аудиоматериалов ESIC, содержащий 179 документов с ручными транскриптами на английском и немецком языках. Эксперимент показал, что при минимальном размере блока 1 секунда транскрибация блока занимает в среднем 500 мс для английского и 700 мс для немецкого, а средняя задержка от произнесенного слова до правильной транскрибации составляет 3.3 секунды для английского, 4.4 секунды для немецкого.

В целом, данная диссертационная работая демонстрирует, что использование модели «ASR Whisper» в реальном времени в распределенной среде значительно улучшает качество распознавания и продолжительность распознанной речи. Распределенная архитектура не только оптимизирует обработку потоков и масштабирует систему, но и предотвращает простои в очереди и отказы системы.

В качестве направлений дальнейших исследований предлагается рассмотреть интеграцию с аналитической платформой для обработки больших объемов данных, такой как «Apache Spark», а также интеграцию с системой управления рабочими процессами «Apache Airflow» для сбора данных для обучения.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

1–А. Жаксылык, К. Аудиоинтерфейс интеллектуальных компьютерных систем нового поколения / К. Жаксылык, В. А. Захарьев [и др.] // 12-я международная научно-техническая конференция OSTIS: материалы международной научной конференции, Минск, 25 ноября 2022 г. / Белорусский государственный университет информатики и радиоэлектроники. редкол.: В. А. Богуш [и др.]. – Минск, 2022. – С. 239–250.

2–А. Жаксылык, К. Разработка системы распознавания звуков птиц с использованием онтологического подхода / К. Жаксылык, В. А. Захарьев [и др.] // 13-я международная научно-техническая конференция OSTIS : материалы международной научной конференции, Минск, 21 апреля 2023 г. / Белорусский государственный университет информатики и радиоэлектроники. редкол. : В. В. Голенков [и др.]. – Минск, 2023. – С. 165–170.

3–А. Жаксылык, К. Автоматическая классификация звуков окружающей среды / К. Жаксылык // Информационные технологии и управление : материалы 59-ой научной конференции аспирантов, магистрантов и студентов, Минск, 17–21 апреля 2023 года / Белорусский государственный университет информатики и радиоэлектроники ; редкол.: Л. Ю. Шилин [и др.]. – Минск, 2023. – С. 16–19.

4–А. Жаксылык, К. Распределенная система потоковой обработки данных для задач распознавания речи / К. Жаксылык, В. А. Захарьев // 10-я Международная научно-техническая конференция «BIG DATA and Advanced Analytics»: материалы международной научной конференции, Минск, 13 марта 2024 г. / Белорусский государственный университет информатики и радиоэлектроники. редкол. : В. А. Богуш [и др.]. – Минск, 2024. – С. 358–371.