**БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**Факультет прикладной математики и информатики**

**Кафедра технологий программирования**

Аннотация к дипломной работе

# «Проектирование и реализация системы для потоковой обработки данных»

Венедиктов Никита Валерьевич

Научный руководитель – ст. преподаватель Зенько Т. А.

2024

# РЕФЕРАТ

Дипломная работа, 74 с., 36 рис., 1 табл. 15 источн., 9 прил.

ПОТОКОВАЯ ОБРАБОТКА ДАННЫХ, BIG DATA, ОЗЕРО ДАННЫХ, ХРАНИЛИЩЕ ДАННЫХ, APACHE KAFKA, APACHE AIRFLOW, APACHE SPARK, GOOGLE CLOUD PLATFORM, LOOKER STUDIO

Объектом исследования являются основные принципы и подходы к работе с большим объемом данных, их потоковой и пакетной обработке и хранению.

Цель работы – проектирование и разработка системы потоковой обработки и хранения пользовательских данных и последующей визуализации некоторых метрик и показателей. В качестве источника данных выступает условное музыкальное приложение.

За время работы были реализованы следующие задачи: изучены понятия и методы работы с большим объемом данных, спроектирована и реализована система потоковой обработки данных музыкального приложения на базе вычислительных и программных ресурсов облачного сервиса Google Cloud Platform с использованием брокера сообщений Apache Kafka, оркестратора процессов Apache Airflow, инструмента распределенной обработки данных Apache Spark, создан интерактивная информационная панель с визуализацией диаграмм и сводок на основе пользовательских данных с использованием инструмента Looker Studio.

Областью применения является отдел аналитики соответствующего музыкального приложения. Визуализированные метрики и диаграммы способны помочь отслеживать некоторые тенденции среди пользователей, более качественно планировать маркетинговую политику с учетом территории и характеристик пользователей и быстро реагировать на определенные события в обществе. Дополнительно, обработанные данные, загруженные в хранилище, могут быть использованы в машинном обучении, например, для реализации алгоритмов рекомендаций или прогнозирования потенциальных нагрузок.

# РЭФЕРАТ

Дыпломная праца, 74 с., 36 мал., 1 табл. 15 крын., 9 дад.

СТРУМЕНЕВАЯ АПРАЦОЎКА ДАДЗЕНЫХ, ВОЗЕРА ДАДЗЕНЫХ, СХОВІШЧА ДАДЗЕНЫХ, APACHE KAFKA, APACHE AIRFLOW, APACHE SPARK, GOOGLE CLOUD PLATFORM, LOOKER STUDIO

Аб'ектам даследавання з'яўляюцца асноўныя прынцыпы і падыходы ў працы з вялікім аб'ёмам дадзеных, іх струменевай і пакетнай апрацоўцы і захоўванні.

Мэта працы - праектаванне і распрацоўка сістэмы струменевай апрацоўкі і захоўванні карыстацкіх дадзеных і наступнай візуалізацыі некаторых метрык і паказчыкаў. У якасці крыніцы дадзеных выступае ўмоўнае музычнае прыкладанне.

За час працы былі рэалізаваны наступныя задачы: вывучаны паняцці і метады працы з вялікім аб'ёмам дадзеных, спраектавана і рэалізавана сістэма струменевай апрацоўкі дадзеных музычнага прыкладання на базе вылічальных і праграмных рэсурсаў воблачнага сэрвісу Google Cloud Platform з выкарыстаннем брокера паведамленняў Apache Kafka, аркестратара працэсаў Apache Airflow, прылады размеркаванай апрацоўкі дадзеных Apache Spark, створана інтэрактыўная інфармацыйная панэль з візуалізацыяй дыяграм і зводак на аснове карыстацкіх дадзеных з выкарыстаннем прылады Looker Studio.

Вобласцю ўжывання з'яўляецца аддзел аналітыкі адпаведнага музычнага дадатку. Візуалізаваныя метрыкі і дыяграмы здольныя дапамагчы адсочваць некаторыя тэндэнцыі сярод карыстальнікаў, больш якасна планаваць маркетынгавую палітыку з улікам тэрыторыі і характарыстык карыстальнікаў і хутка рэагаваць на пэўныя падзеі ў грамадстве. Дадаткова, апрацаваныя дадзеныя, загружаныя ў сховішча, могуць быць скарыстаны ў машынным навучанні, напрыклад, для стварэння алгарымтаў рэкамендацый ці прагназавання нагрузак.

# ABSTRACT

Diploma thesis, 74 pages, 36 illustrations, 1 table, 15 sources, 9 applications.

STREAMING DATA PROCESSING, BIG DATA, DATA LAKE, DATA WAREHOUSE, APACHE KAFKA, APACHE AIRFLOW, APACHE SPARK, GOOGLE CLOUD PLATFORM, LOOKER STUDIO

The object of research is the basic principles and approaches in working with large amounts of data, their streaming and batch processing and storage.

The purpose of the work is to design and develop a system of stream processing and storage of user data and subsequent visualization of some metrics and indicators. A conditional music application acts as a data source.

During the work the following tasks were realized: the concepts and methods of working with large amounts of data were studied, the service of stream processing of music application data was designed and implemented on the basis of computing and programming resources of Google Cloud Platform using Apache Kafka message broker, Apache Airflow process orchestrator, Apache Spark distributed data processing tool, an interactive dashboard with visualization of diagrams and summaries based on user data was created using Looker Studio tool.

The scope of the application is the analytics department of a relevant music application. Visualized metrics and diagrams can help to track certain trends among users, to better plan marketing policy taking into account the territory and characteristics of users and to react quickly to certain events in the society. Additionally, the processed data uploaded to the repository can be used in machine learning, for example, to create recommendation algorithms.