**Практическое занятие № 12**

**Тема практического занятия:** Использование облачных сервисов копирования и транформации данных

**Цель практического занятия:** овладеть навыками использования облачных сервисов для эффективного копирования и трансформации данных, а также освоить инструменты, предоставляемые облачными платформами, для оптимизации процессов обработки информации.

**В результате выполнения данной работы обучающийся должен**

**уметь:**

* выбирать подходящие облачные сервисы для копирования и трансформации данных в соответствии с поставленными задачами.
* настраивать и организовывать процессы копирования данных, включая использование автоматизированных средств.
* применять инструменты обработки данных для их трансформации в облачной среде.
* обеспечивать безопасность данных при их обработке в облаке.
* проводить оценку эффективности процессов копирования и трансформации данных с использованием облачных сервисов.

**знать:**

* основы облачных вычислений и их понятия.
* принципы организации облачных сервисов.
* методы копирования данных в облачных средах.
* принципы работы облачных хранилищ данных.
* основные методы трансформации данных.
* процессы обработки данных в облачных окружениях.
* инструменты и сервисы для копирования данных в облаке.
* функционал инструментов трансформации данных.
* принципы обеспечения безопасности данных в облачных средах.
* методы шифрования и контроля доступа.

**Перечень оборудования, необходимого для выполнения задания:**

* Автоматизированные рабочие места по количеству обучающихся (процессор Intel Core i7 или аналогичный, БП 700 Вт, 16 Гб ОЗУ, SSD 256 Гб, HDD 1 ТБ SATA 7200 rpm);
* Автоматизированное рабочее место преподавателя (процессор Intel Core i7 или аналогичный, БП 700 Вт, 16 Гб ОЗУ, SSD 256 Гб, HDD 1 ТБ SATA 7200 rpm);
* Монитор 27” IPS, 1920x1080, 178°/178°, 250cd/m2, 1000:1, 5 ms, HDMI, DP (2 монитора на одно рабочее место)
* Интерактивная панель 86” с OPS ПК
* Программное обеспечение общего и профессионального назначения

**Общие теоретические сведения:**

Yandex Cloud предоставляет ряд облачных сервисов, которые можно использовать для копирования и трансформации данных. Ниже представлен примерный список шагов и инструментов, которые могут быть использованы на платформе Yandex Cloud:

1. Выбор хранилища данных:

Используйте объектное хранилище Yandex.Cloud Object Storage (COS) для сохранения исходных данных.

Сервис Yandex Object Storage — это универсальное масштабируемое решение для хранения данных. Оно подходит как для высоконагруженных сервисов, которым требуется надежный и быстрый доступ к данным, так и для проектов с невысокими требованиями к инфраструктуре хранения.

Используя Object Storage, вы можете:

* Хранить данные произвольного формата как объекты с доступом по текстовому идентификатору.
* Использовать различные классы хранилища для объектов и управлять их жизненным циклом.
* Хранить большие объекты размером в несколько терабайт.
* Опубликовать статический веб-сайт.

1. Копирование данных:

Для копирования данных в Yandex.Cloud COS можно использовать инструменты командной строки, такие как yc (Yandex.Cloud CLI) или SDK (Software Development Kit) для вашего языка программирования.

1. Трансформация данных:

Для трансформации данных можно использовать Yandex DataSphere, платформу для анализа данных, которая предоставляет инструменты для обработки данных в распределенной среде.

1. Выполнение вычислений:

Используйте Yandex.Cloud Functions для выполнения вычислений на основе событий. Также можно рассмотреть возможность использования Yandex Managed Service for Apache Spark для распределенной обработки данных.

1. Мониторинг и логирование:

Для мониторинга и логирования можно воспользоваться сервисами мониторинга Yandex.Cloud, такими как Yandex Monitoring и Yandex Log Management.

1. Распределенные системы хранения данных:

В зависимости от требований к обработке данных, вы можете использовать Yandex Managed Service for ClickHouse для аналитической обработки данных в реальном времени.

1. Безопасность:

Обеспечьте безопасность данных, используя возможности управления доступом, шифрования и другие меры безопасности, предоставляемые Yandex.Cloud.

1. Создание пайплайнов:

Для оптимизации процессов создания пайплайнов можно использовать Yandex Data Proc, предоставляющий управление распределенными вычислениями.

1. Интеграция с другими сервисами:

Интегрируйте Yandex.Cloud с другими сервисами, такими как базы данных, сервисы машинного обучения и т. д.

**Задание:**

Разработать и реализовать конвейер обработки данных с использованием облачного сервиса для копирования и трансформации данных. Оценить производительность и эффективность полученного конвейера.

**Технология выполнения работы:**

1. Выбор облачной платформы:

* Выберите одну из популярных облачных платформ.
* Создайте аккаунт на выбранной платформе, если у вас его еще нет.

1. Копирование данных в облако:

* Загрузите набор данных (можете воспользоваться общедоступными наборами или создать собственный).
* Используя инструменты облачной платформы, скопируйте данные в облако.

1. Трансформация данных в облаке:

* Выберите набор данных в облаке и определите необходимые трансформации (например, изменение формата, фильтрация данных).
* Примените выбранные трансформации с использованием сервисов обработки данных в облаке.

1. Безопасность данных:

* Обеспечьте безопасность данных в процессе копирования и трансформации. Используйте средства шифрования и механизмы управления доступом.

1. Оценка эффективности и мониторинг:

* Оцените производительность сервисов облачной платформы при копировании и трансформации данных.
* Используйте инструменты мониторинга для отслеживания процессов и оптимизации производительности.

**Указания по технике безопасности:**

В процессе выполнения работы необходимо соблюдать правила безопасности и поведения на рабочем месте.

**Чек-лист результатов выполнения работы:**

* Разработан и реализован конвейер обработки потоковых данных с использованием облачных сервисов копирования и трансформации.
* Проведена оценка собранных данных, включая анализ целостности и производительности конвейера.
* Подготовлен подробный отчет, описывающий каждый этап выполненной работы.

**Контрольные вопросы:**

1. Что представляют собой облачные сервисы копирования и трансформации данных?
2. Какие преимущества обладают облачные сервисы в сравнении с локальными решениями?
3. Какие критерии следует учитывать при выборе облачных сервисов для копирования и трансформации данных?
4. На какие популярные облачные платформы можно полагаться для подобных сервисов?
5. Каковы этапы обработки данных в конвейере, и почему важно определить последовательность операций?
6. Как происходит интеграция конвейера данных с Apache Hadoop, и почему важно использовать Hadoop Distributed File System (HDFS)?
7. Какие методы трансформации данных могут быть применены в облачных сервисах, и как обеспечивается согласованность данных?
8. Какие инструменты или метрики можно использовать для оценки производительности конвейера данных?
9. Почему важно обеспечить выгрузку обработанных данных в удобный формат, например, CSV, и как обеспечивается безопасная выгрузка данных из HDFS?

**Основные и дополнительные источники, электронные ресурсы:**

1. Радченко, И.А, Николаев, И.Н. Технологии и инфраструктура Big Data. – СПб: Университет ИТМО, 2020.
2. Сенько, А. Работа с BigData в облаках. Обработка и хранение данных с примерами из Microsoft Azure. — СПб.: Питер, 2019. — 448 с.: ил.
3. Тесленко, И. Б., Губернаторов, А. М., Дигилина, О. Б., Крылов, В. Е. Большие данные. 2021.
4. Шеннон, Брэдшоу, Йон, Брэзил, Кристина, Ходоров. MongoDB: полное руководство. Мощная и масштабируемая система управления базами данных / пер. с англ. Д. А. Беликова. – М.: ДМК Пресс, 2020.