

Министерство науки и высшего образования Российской Федерации
Национальный исследовательский университет ИТМО
Инфраструктура больших данных

Весна

2024

Лабораторная работа №6

ИСТОЧНИКИ ДАННЫХ НА PYSPARK

Цель работы:

Получить навыки выгрузки исходных данных и отправки результатов модели с использованием различных источников данных согласно варианту задания.

Ход работы:

1. Обеспечить выгрузку данных при каждом запуске модели.
2. Обеспечить загрузку данных сразу по завершении работы модели.
3. Необходимо разработать протокол взаимодействия между моделью и источником данных.
4. Необходимо разработать формат хранения данных исходя из особенностей источника данных.
5. Рекомендуется использование docker контейнеров.

Схема: модель-источник данных.

Результаты работы:

1. Отчёт о проделанной работе;
2. Ссылка на репозиторий GitHub;
3. Актуальный дистрибутив модели в zip архиве.

Варианты задания

Номер	Источник данных
1	PostgreSQL
2	MySQL
3	Oracle
4	MS SQL Server
5	MongoDB (в случае недоступности использовать Neo4J)
6	Apache HBase
7	Redis
8	Cassandra
9	Greenplum
10	ClickHouse
11	HDFS