

Министерство науки и высшего образования Российской Федерации

Национальный исследовательский университет ИТМО

Инфраструктура больших данных

Весна

2024

Лабораторная работа №7

ВИТРИНА ДАННЫХ НА SPARK

Цель работы:

Получить навыки разработки витрины данных и последующей её интеграции.

Ход работы:

1. Разработать витрину данных на языке Scala для реализации протокола, единого формата данных. Витрина предназначена для формирования запросов к источнику и отгрузки результатов работы модели. В данном случае модель не взаимодействует с источником данных напрямую, а лишь через витрину данных.
2. Реализовать предварительную обработку данных на стороне витрины, проверить работоспособность, удалить функционал по предобработке из сервиса модели, так как пред обработанные данные поступают из витрины данных (data mart).
3. Провести интеграцию витрины данных в имеющийся контур: модель (лабораторная 5) и источник данных (лабораторная 6).

Схема: модель-витрина-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

Полезные материалы:

1. <https://www.scala-lang.org/>
2. https://www.tutorialspoint.com/scala/scala_environment_setup.htm
3. <https://laxfed.medium.com/data-analysis-with-scala-and-spark-part-1-cda0ef2ac2cc>
4. <https://laxfed.medium.com/data-analysis-with-scala-and-spark-part-2-a48f76f2e865>
5. <https://alvinalexander.com/scala/scala-jdbc-connection-mysql-sql-select-example/>