Техническое описание DATAGRAM

DATAGRAM – это программная платформа предназначенная для создания приложений по обработке данных (приложений класса ETL) с использованием технологий BIG DATA. DATAGRAM использует библиотеку **Apache Spark** и поддерживает как пакетный, так и потоковый режим обработки данных.

Ядром платформы является **сервер метаданных**. Сервер метаданных содержит информацию об источниках и приемниках данных, исполняющих средах, трансформациях данных, последовательности выполнения заданий и т.д. Кроме того, предоставляются средства для развёртывания, планирования и мониторинга программ обработки.

DATAGRAM поддерживает **полный цикл** разработки приложений:

- Визуальное проектирование процессов преобразования данных;
- Визуальное проектирование последовательности преобразований;
- Генерация исходного кода на языке Scala с использованием библиотеки Apache Spark;
- Компиляция и генерация приложения;
- Развертывание приложения на исполняющей среде и планирование исполнения;
- Мониторинг исполнения приложения;
- Инструменты для остановки и перезапуска приложений.

Исполняющие среды
Исполняющие среды DATAGRAM базируются на Apache Spark.
Datagram может выполнять запуск приложений на серверах Apache Oozie или Apache Livy. Сервер Apache Oozie используется для запуска последовательности преобразований. Сервер Apache Livy используется для запуска приложений из среды разработки.

Дизайнер трансформаций
Дизайнер трансформаций – web-интерфейс для визуальной разработки процессов преобразования данных. Интуитивный интерфейс с поддержкой drag-n-drop позволяет создавать преобразования данных произвольной сложности.

В дизайнере трансформаций поддерживается широкий спектр **источников/приемников данных**:

- RDBMS источники/приемники данных использующие соединение JDBC (включая хранимые процедуры);
- Файловые источники/приемники со сложной структурой: CSV, XLS, XML, AVRO и JSON;

- Источники/приемники данных файловой системы HDFS поддерживают форматы файлов: ORC, PARQUET;
- BigData источники/приемники Apache Hive, Apache HBase, Apache Kafka.

Типы преобразований данных:

- Широкий набор операций реляционной алгебры: join, sort, aggregation, union, selection, projections, pivot, explode arrays, sequence generation;
- Spark SQL;
- Анализ на основе машинного обучения с использованием Spark MLLib (decision trees, SVM, logistic regression и т.д.);
- Jboss Rules (Drools).

Основные возможности:

- Просмотр содержимого и структуры реляционных и файловых источников и приемников данных;
- Просмотр структуры потока данных, поступающего на вход элемента схемы трансформации, отслеживание происхождения отдельных полей потока данных и проверка структуры потока данных на выходе элемента (lineage);
- Частичное выполнение преобразования с просмотром промежуточных результатов;
- Просмотр сгенерированного кода приложения, его редактирование и запуск на исполнение;
- Валидация трансформации на основе данных о наиболее частых ошибках;
- Поддержка Spark Catalyst Optimizer.
- Поддержка типов данных struct и array;

Дизайнер Workflow

Дизайнер Workflow – интерфейс для визуальной разработки потоков управления последовательностями преобразований данных.
Основные возможности:

- Создание схем параллельного, последовательного или зависящего от заданных условий исполнения преобразований данных;
- Поддержка shell scripts и java программ;
- Возможность создания схем управления последовательностями преобразований с использованием вложенных объектов Workflow;
- Запланированное исполнение Workflow по времени или событиям файловой системы.

Интеграция метаданных

- Импорт метаданных из систем источников/приёмников
- Экспорт метаданных в Apache Atlas

Безопасность

- Централизованная аутентификация пользователей с использованием корпоративного сервера каталогов (LDAP).
- Ролевая авторизация. Возможные роли: developer, operator, viewer;
- Шифрование паролей для доступа к внешним системам;
- Использование алгоритма аутентификации Kerberos для подключения к исполняющим средам.

Версионность и teamwork

- Блокировка параллельных обновлений метаданных (optimistic locking);
- Интеграция с GIT и SVN;

Поддержка Gitflow и CI/CD

- параллельная работа в разных branches
- сохранение сгенерированного кода в GIT
- интеграция с Apache Maven
- Unit тестирование

Поддержка работы с несколькими инсталляциями

- Импорт/экспорт метаданных в виде zip-файлов;
- Перенос метаданных (полный или основанный на проекте) в новую среду;
- Перезапись URL-адресов, паролей и т.д. при переносе в новую среду.

Дополнительные инструменты

- **Консоль HDFS**: просмотр, сохранение файлов из/в файловой системы HDFS;
- **Консоль Livy**: просмотр задач на сервере Livy, просмотр журналов, запуск и отмена задач;
- **Консоль Oozie**: обзор workflow и задач координатора на сервере Oozie, просмотр журналов, отмена или перезапуск задач;
- **Обозреватель объектов**: просмотр дерева объектов метаданных, поиск объектов.