УСТАНОВКА И ЗАПУСК BI.QUBE METASTAGING

Москва, 2023

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Компонент MetaStaging позволяет консолидировать в стейджинговом слое хранилища данные из гетерогенных источников с поддержанием целостности и унифицированности метаданных, также уменьшает нагрузку на операционные базы при выполнении запросов, а кроме того, обеспечивает надежное подключение различных БД из разнородных источников для помещения данных в единый слой стэйджинга (staging area) с поддержанием целостности метаданных в системе-назначения.

В документе приведено описание компонента и принципы работы с ним. Рассмотрены примеры загрузки данных с помощью компонента из разных источников.

Изучение данного документа позволит понять принцип работы компонента.

ГЛОССАРИЙ

1.	MetaStaging - BI.Qube	Инструмент, предназначенный для транспортировки данных.
2.	Хранимая процедура	Объект базы данных, представляющий собой набор SQL-инструкций, который компилируется один раз и хранится на сервере
3.	Представление	Виртуальная таблица, содержимое которой определяется запросом
4.	Бизнес-представление	Представление, в котором собраны Hub, Satellite и Link для сущности
5.	Материализация	Процесс сохранения результата запроса бизнес-представления в таблицу для ускорения выборки.
6.	Инкрементальная загрузка (загрузка с параметрами)	Регулярная загрузка данных в Greenplum. Извлекаются актуальные данные с даты последней загрузки. В таблице stg.session базы данных settings.db можно отследить историю всех загрузок.
7.	Полная загрузка (снэпшоты)	Загрузка данных в Greenplum без параметризации. Применяется, когда необходима полная перезагрузка всех данных в таблице на источнике (например, при отсутствии столбца, подходящего для секционриования).
8.	Полная загрузка с сохранением истории	Загрузка данных в Greenplum без параметризации. Но представления на Greenplum перенацеливаются на новые Parquet-файлы, а старые не удаляются из S3.
9.	Профиль	Добавляются в таблице stg.profile

10.	Экстрактор	Компонент системы для извлечения данных из источников в S3. Исполняемый файл находится в директории LoadingToS3. Вызывается в Airflow в соответствии с командами в настроечной БД (settings.db).
11.	External Table (ET)	Вид таблицы в Greenplum, обеспечивающий доступ к внешним источникам данных, как к объекту самой БД Greenplum. В системе используется для получения доступа к файлам в S3. Используется фреймворк PXF.
12.	Сервисные процедуры	Процедуры, вызываемые автоматически в процессе работы компонента.

1. ЦЕЛИ И НАЗНАЧЕНИЕ METASTAGING

Цель MetaStaging – обеспечить транспортировку данных из систем источников в файловое S3-совместмое хранилище данных (HDFS, ObjectStorage) с автоматической генерацией в СУБД Greenplum объектов типа «представление» на каждый полученный файл хранилищем.

Компонент MetaStaging, предназначен для передачи данных из различных источников, как правило, из учетных систем в целевое корпоративное хранилище данных (КХД) с поддержкой целостности метаданных систем-источников, при формировании промежуточного физического слоя хранения учитываются особенности целевой платформы.

Компонент MetaStaging входит в состав системы BI.Qube и может эксплуатироваться как отдельный компонент, так и в составе системы, так и под управлением компонента MetaOrchestrator, в такой конфигурации использование компонента является наиболее эффективной.

2. УСТАНОВКА И ЗАПУСК

Компонент MetaStaging для развертывания, функционирования и настройки использует различные программные инструменты и фреймворки. Обязательным условием является наличие у них открытого исходного кода.

Поддерживаемые операционные системы: Linux (различные дистрибутивы, такие как Ubuntu, Mint, РЕД ОС), другие Unix-подобные системы, а также есть возможность развернуть компонент под Windows.

Настроечные данные компонента могут храниться посредством СУБД: PostgreSQL (9.0 и позднее) / Postgres Pro (10.22 и позднее) / Arenadata Postgres (ADPG) (14.2.1) / Greenplum на выбор заказчика.

Для тестирования корректности загрузки данных в S3 хранилище с помощью файлов «.parquet» использовались инструменты s3-browser и parquet-viewer.

Инструменты разработки DBeaver, Visual Studio Code

Среды выполнения Python, «.Net Core».

В качестве библиотек для взаимодействия с системами источниками и назначениями, а также для обеспечения интеграции данных используются:

- AWSSDK.S3 - Amazon Simple Storage Service для Amazon S3 (nuget.org)
- CommandLineParser - Terse syntax C# command line parser for .NET.
- ExcelDataReader - Lightweight and fast library written in C# for reading Microsoft Excel files
- Google.Cloud.BigQuery.V2 - Recommended Google client library to access the BigQuery API.
- Microsoft.Data.SqlClient - Provides the data provider for SQL Server.
- MySql.Data
- Newtonsoft.Json - Json.NET high-performance JSON framework for .NET
- Npgsql - is the open source .NET data provider for PostgreSQL.
- ParquetSharp -.NET library for reading and writing Parquet files.
- YandexDisk.Client - .NET library wrapper of Yandex Desktop RestAPI.
- Встроенные модуле из стандартной библиотеки Python.
- Psycopg2 – PostgreSQL database adapter for the Python programming language.

Добавлено примечание ([DP1]): @Andrey Azarchenkov что-то тут с окончаниями

В связи с высокой сложностью развертывания компонента в среде целевой СУБД установку компонента осуществляет вендор.

2.1. Доступ к площадке

Доступ к площадке осуществляется через подключение к удаленному рабочему столу:

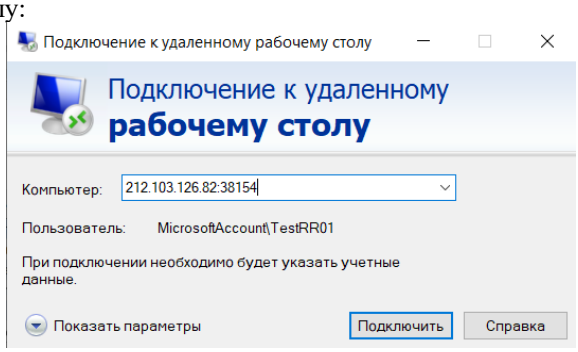

Рисунок 1. Подключение к удаленному рабочему столу

Компьютер: 212.103.126.82:38154

Пользователь: TestRR01

Пароль: ***

На машине установлена среда разработки DBeaver для подключения к настроечной БД компонента MetaStaging.

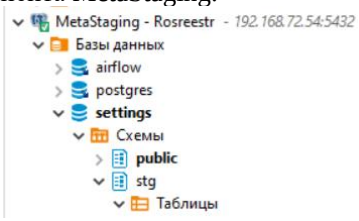

Рисунок 2. Среда разработки DBeaver

2.2. Запуск компонента на linux

Для вызова MetaStaging необходимо подключиться к машине через SSH (там находятся все исполняемые файлы), используя программу Putty () :

HostName: 192.168.72.54

User: itpro_admin

Password: ***

Для запуска компонента MetaStaging необходимо выполнить следующую команду, предварительно проверив корректность заполнения настроечной БД (*см. гл.* ***Ошибка! Источник ссылки не найден.***)

sudo python3 MetastagingExecutor/MetastagingExecutor.py

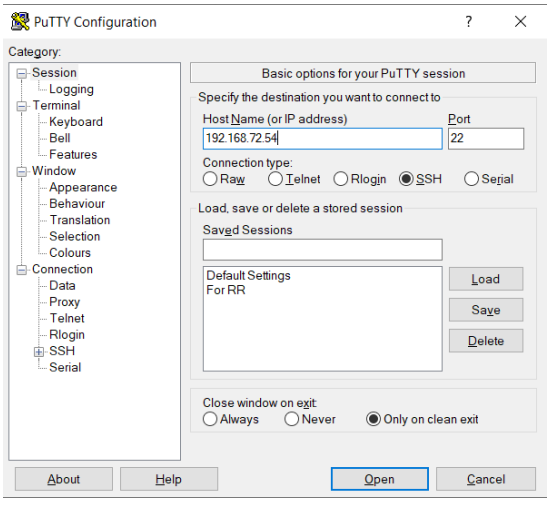

Рисунок 3. Putty

2.3. Запуск компонента на windows

Запуск MetaStaging с удаленного рабочего стола windows осуществляется при вызове в командной строке на той же рабочей машине, к которой подключились через RDP:

python3 "C:\MetaStaging\MetaStagingExecutor\MetastagingExecutor.py"