01 марта 2005

Пастух больших данных: как мы используем Azure Data Factory в качестве единого сервиса для задачи миграции

КАК ЭТО РАБОТАЕТ НА ПРАКТИКЕ

Итак, у нас есть задача регулярно обновлять отчеты в BI-системе на основании информации, которая приходит по почте или выкладывается в облачное хранилище, например, Google Drive.

1. Первый шаг — вызов Azure Logic Apps, сервиса, позволяющего производить запуск Azure Data Factory Piplines для интеграции с различными источниками данных.

2. Azure Logic Apps в соответствии с настройками отправляет запрос в Google Drive.

3. Google Drive возвращает информацию о том, что в нем хранится: список файлов, дату модификации, ссылки на скачивание и пр.

4. Эта информация передается из Logic Apps в Data Factory.

5. Data Factory в соответствии с этими данными производит копирование файлов. Сервис сохраняет их в Data Lake либо в виде файла в формате Parquet, или как есть (например, zip).

Здесь надо сказать пару слов об Apache Parquet, столбцовом формате хранения данных. Он позволяет эффективно сжимать информацию и считывать файлы частично, по мере необходимых столбцов. Parquet предоставляет возможность задавать схемы сжатия на уровне столбцов и добавлять новые кодировки по мере их изобретения и реализации. Именно поэтому этот формат файлов часто используют при работе с Big Data.

6. Мы получили файлы и теперь читаем их с помощью Data Bricks, сервиса аналитики данных, в котором реализована поддержка нескольких языков: Python, Scala, SQL. Это позволяет решать множество различных задач по обработке, анализу и инжинирингу данных.

7. После работы Data Bricks поля приведены в нужный вид, и новые преобразованные данные записываются в Data Lake.

8. Дальше при необходимости мы записываем информацию в базу данных в Azure Synapse.

9. Обновляем витрины и отображаем эти данные в отчете BI.

В итоге мы сделали так, что в дашбордах наши пользователи видят актуальные данные, вне зависимости от того, в каком виде и куда пришла к нам первоначальная информация. Также отформатированные данные доступны для дальнейшей работы аналитиков и data scientist-ов.

Профиты подхода

Я сформулировал несколько преимуществ Azure Data Factory при работе с преобразованием больших данных:

- Единый инструмент для работы с интеграциями. Можно настроить большое количество интеграций, которые будут работать параллельно или последовательно в одном или нескольких Azure Data Factory Pipelines.
- Масштабируемость. Производительность решения зависит от мощностей, которые у вас есть, и от пропускной способности сети. Масштабировать их можно автоматически, благодаря облачной среде и ее инструментам.
- Мониторинг процессов. При использовании других инструментов нужно постоянно следить за каждым, чтобы интеграция и миграция данных шли нормально. С Data Factory можно мониторить все процессы в одном месте.
- Экономия ресурсов. Не думайте, что я хочу отобрать хлеб у специалистов по другим областям, но при миграции данных вам не потребуется искать десяток экспертов по различным технологическим продуктам.
- Развитие решения. Microsoft постоянно добавляет функциональность, чтобы Azure становился производительнее и функциональнее.

Напоследок пару слов о том, кому подходит Azure. Применять его можно в любой организации, которая работает с данными. Вот примеры отраслей, в которых инструмент может быть наиболее удобным:

1.

Энергетические сбытовые компании, в которые стекается информация с различных измерительных устройств.

2.

Розничные сети, которым требуется оперативно следить за продажами в магазинах во всей стране

3.

Логистические операторы, в которых надо следить передвижением товаров.

4.

Фармацевтические сети, собирающие статистические данные по всей страны.

Количество компаний, которые работают с большими данными, постоянно растет: их используют в ecommerce, в производстве, в агросфере. Технологии для работы с ними переходят из разряда конкурентных преимуществ в базовую необходимость, и этот тренд будет только усиливаться.