АНАЛИЗ БОЛЬШИХ ДАННЫХ С ПОМОЩЬЮ AZURE DATA LAKE И AZURE DATA FACTORY

И.О. Вахович
Аспирант БГУИР

Н.А. Волорова
Заведующая кафедрой информатики БГУИР, кандидат технических наук, доцент

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь

Abstract. Big Data, Azure Data Lake, Azure Data Factory, U-SQL.

В этом году предложения Microsoft Azure Big Data были пополнены сервисом Azure Data Lake(ADL) который наряду с возможностью создания комплексных контейнеров больших данных, использует ADL и Azure Data Factory(ADF). ADL упрощает и делает доступной обработку больших данных, предоставляя несколько ключевых технологий. Язык U-SQL является мощной комбинацией SQL и C# и поддерживает параллельное выполнение. ADF - это облачный сервис интеграции данных, который координирует и автоматизирует перемещение и преобразование данных. Интеграция ADF и ADL позволяет:

- перемещать данные из указанного источника в ADL Store;
- создавать конвейеры больших данных ADF, выполняющих скрипты U-SQL как стадию обработки в сервисе ADL Analytics.

Существует ряд распространенных сценариев Big Data, на которые ориентированы ADL и ADF(см. [1] примеры сценариев), но общий порядок работы с сервисами следующий.

Приступая к конкретной задаче анализа больших данных, мы начинаетм с создания учетных записей ADL Store, аналитики и подготовки фабрики данных. ADL Store — это сервис для хранения больших данных, к которому можно обращаться из HDFS-совместимых систем, включая инструменты бизнес-анализа (business intelligence, BI) и приложения на предприятии. Учетные записи ADL Store можно создавать раздельно и использовать с другими сервисами, но чаще всего учетная запись создается в тандеме с ADL Analytics. Далее следует подготовка фабрики данных. Фабрики данных являются комбинацией хранилищ данных, связанных сервисов и конвейеров. Хранилища данных и связанные сервисы представляют собой определения внешних сущностей, которые обычно

уже существуют вне ADF. Конвейеры – это логические группы действий в ADF. Они используются для группирования действий в единицу, которая выполняется какую-либо задачу.

Первый шаг в анализе данных – перемещение данных в ADL Store. Для этого можно использовать действие Copy в конвейере ADF. Чтобы выполнить операцию копирования, мы должны создать связанные сервисы, наборы данных и конвейеры ADF.

Нам также нужны два набора данных ADF. Наборы данных являются логическими ссылками на данные по учетной записи Azure Storage или ADL Store. В самой ADF никакие пользовательские данные не хранятся, поэтому ADF нужны определения наборов данных, чтобы идентифицировать структуру данных во внешних хранилищах, включая таблицы, файлы, папки и документы. Поскольку ADF не известна структура этих данных, мы должны определить ее здесь, чтобы система знала, какие столбцы и типы следует ожидать.

Наконец, чтобы произошло копирование данных, мы должны создать ADF-конвейер, содержащий действие Copy. ADF-конвейер — это логические группы действий, таких как копирование данных, которые можно выполнять через разные интервалы, а также действия скриптов Hive, Pig или U-SQL(см. [2] более подробно про использование U-SQL и ADL), которые можно выполнять регулярно — через каждые 15 минут, ежечасно, ежедневно или ежемесячно.

Когда действие Copy в ADF-конвейере завершается успешно, данные оказываются перемещенными из Azure Blob Storage в Azure Data Lake Store.

Теперь, поместив данные в ADL Store, можно запускать скрипты U-SQL в сервисе ADL Analytics для обработки и анализа данных. Мы можем создать конвейеры, которые будут потреблять данные из ADL Store, выполнять скрипты U-SQL в сервисе ADL Analytics как стадию обработки и отправлять вывод в ADL Store. Затем нижестоящие приложения могут потреблять обработанный вывод напрямую из ADL Store или, копировать данные из ADL Store в хранилище Azure SQL, если наши BI-приложения используют в качестве серверного хранилища базу данных SQL.

Сервис Data Factory предоставляет надежное, полное представление о хранении, обработке и перемещении данных. Он помогает быстро обращаться к показателям работоспособности комплексных конвейеров данных, выявлять проблемы и при необходимости предпринимать корректирующие меры. Кроме того, можно визуально отслеживать операционный журнал преобразований (operational lineage) и связи между данными в любом из источников, а также просматривать полные хронологические сведения о выполнении заданий, работоспособности системы и зависимостях с единой информационной панели.

В докладе описаны основные шаги, которые позволяют создать комплексный конвейер больших данных с помощью Azure Data Factory, который позволит перемещать данные в Azure Data Lake Store, и, используя скрипт U-SQL в сервисе Azure Data Lake Analytics, обрабатывать данные для получения необходимой аналитики. Полученная система будет динамической, и ее можно

расширить для выполнения на регулярной основе. Кроме того, можно выполнять дальнейшую обработку вывода и помещать его в другое серверное хранилище, чтобы результаты использовались Power BI или любым другим BI-приложением, применяемым в организации. Более того, при желании можно задействовать командлеты ADF PowerShell, C# SDK и плагин Visual Studio для создания E2E-конвейеров больших данных с помощью ADL. Azure Data Lake совместно с Azure Data Factory исключает сложности, обычно связанные с большими данными в облаке, обеспечивая текущие и будущие потребности бизнеса.

Литература

[1]. Gaurav Malhotra. Big Data - Create pipelines of large data using Azure Data Lake and Azure Data Factory. MSDN, 2016, № 2 (170), 23-30

[2]. Ed Macauley. Tutorial: Get started with Azure Data Lake Analytics U-SQL language [Electronic resource]. – 2016. - Mode of access: https://azure.microsoft.com/en-us/documentation/articles/data-lake-analytics-u-sql-get-started/