Im Check: Microsoft Azure Synapse Analytics

Azure Synapse Analytics

Aufbau einer ETL Pipeline, Vorteile und Nutzen im Überblick

Mit einer Vielzahl von Data Warehousing- und Data Analytics-Lösungen positionieren sich Cloud Anbieter, um die Datengrundlage für BI und Analytics möglichst einfach zu implementieren. Wachsende Datenbestände erfordern eine immer leistungsstärkere, skalierbare Datenverarbeitung. Dabei wird der Bedarf an vereinfachten Integrationen mit verschiedenen On-Premise Lösungen, SQL vs. NOSQL-Technologien oder einer Multi-Cloud-Lösung immer größer. Um eine solche Nachfrage zu bedienen, positionieren sich die Big Player im Cloud Umfeld (AWS, Azure, GCP) mit ihren Tools.

In dieser Artikel Publikation zeigen wir Ihnen, wie Sie das Microsoft Tool Azure Synapse, welches einen solchen Bedarf adressiert, verwenden können. Dabei fokussieren wir uns auf den Nutzen und die Einsatzzwecke der Software und geben beispielhaft einen Einblick über die Erstellung einer Analytics Pipeline (ETL), welche u.a. als Datengrundlage für Reporting Software wie Power BI dienen kann. Abschließend ordnen wir die Software durch einen Vergleich mit dem Wettbewerber Databricks für Sie ein.

BUSINESS.
DATA.
TECHNOLOGY.

CINTELLIC
CONSULTING GROUP

Was ist Azure Synapse Analytics und was bietet es?

Neben einem Analyseservice bietet Microsoft mit seiner Fertiglösung Azure Synapse auch ein universelles Data Warehouse an. Azure Synapse vereint nicht nur SQL-Technologie und Data Warehousing, sondern bietet zudem auch Big Data Technologien wie Spark in der Integration an. Diese BI-soliden Datengrundlagen können an Microsoft Power BI oder ein anderes integriertes Reporting-Tool geliefert werden. Azure Synapse agiert hierbei als grenzenloser Analyseservice und als End-to-End-Lösung.

Die Azure Synapse Analytics Grundarchitektur besteht aus vier Komponenten:

1. **Synapse SQL**
   Ein Analysedienst, der eine Oberfläche für User bietet, mittels T-SQL-Abfragen die Möglichkeit Abfragen gegenüber den Azure-Datenbeständen innerhalb der eigenen Umgebung durchzuführen.
2. **Apache Spark**
   Spark als Cluster Framework wird in Azure Synapse durch Spark Pools unterstützt. Durch das Framework erhält man die Möglichkeit, die volle In-Memory Clustercompute Engine von Apache Spark zu nutzen.
3. **Synapse Pipeline**
   Ein No-Code ETL-Tool, welches Datenflüsse innerhalb Azure Synapse automatisiert.
4. **Synapse Studio**
   Synapse Studio agiert als Front End, mit allen Komponenten – welche auch als Hubs bezeichnet werden - die Synapse anbietet.

Abbildung 1 zeigt, wie die Daten per Ingest (Prozess des Imports großer, sortierter Datendateien) in einen Cloud Storage / in das Data Warehouse gelangen und zur Weiterverarbeitung an Azure Synapse Analytics übermittelt werden. Dabei besteht die Möglichkeit, diese Daten nochmals im Data Lake Storage permanent abzuspeichern und dadurch einen Zugriff für weitere Services zu ermöglichen. Neben einem Datenfeldkatalog in Azure Purview werden auch BI-Report-Lösungen (wie z.B. Microsoft Power BI) angeboten.

Was sind die größten Vorteile von Azure Synapse?

- ✓ **Einheitliche Datenplattform:** Azure Synapse Analytics bietet eine Vollintegration mit Azures besten Datenservices wie Azure Data Factory, Power BI, SQL-Pools und Azure Purview. Darüber hinaus können neben Data Engineers, die ihre Pipelines erstellen, Datenanalysten ihre Analysen in der ihnen beliebten Form wie z.B. Notebooks durchführen.

- ✓ **Spark Integration:** Azure Synapse besitzt eine selbst verwaltete Apache Spark Umgebung, wodurch Cluster Computing mit Spark Technologie möglich ist. Dies ist bei der Verarbeitung großer Datenquellen ressourceneffizient.

- ✓ **Azure Analytics Features:** Azure Synapse besitzt ein zentralisiertes Datenmanagement, welches auf Massively Parallel Processing (MPP) Technologie basiert und dadurch große Workloads schnell und effektiv bearbeitet. Mittels Workload Isolation können Ressourcen ausschließlich für eine Arbeitsauslastungsgruppe durch rollenbasierte Service Level Agreements (SLAs) reserviert werden.

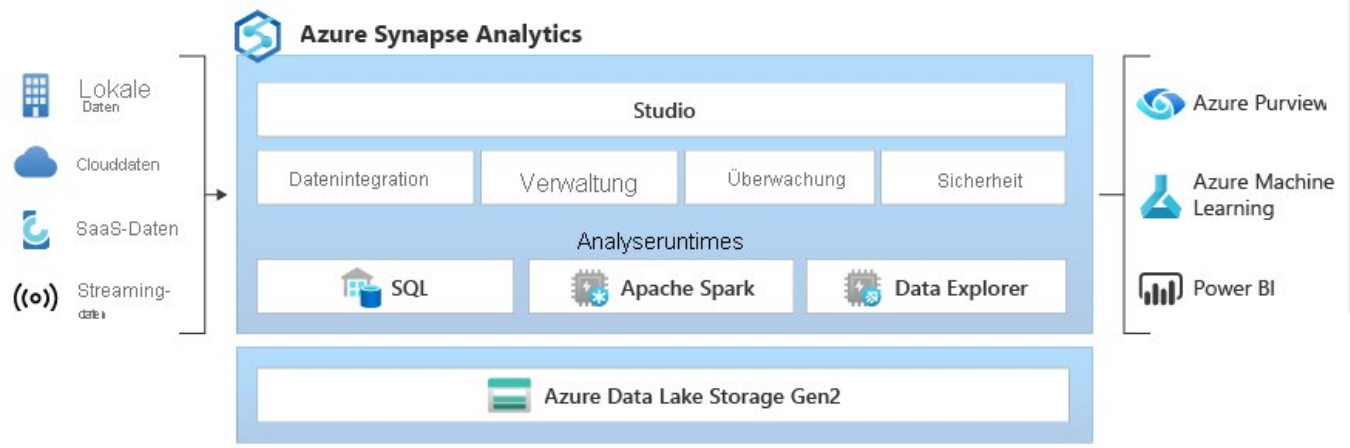

Abbildung 1: Architektur Azure Synapse und Integration weiterer Services

Leitfaden zur Erstellung eines Synapse Services und Erstellung einer ETL Pipeline

1. Erstellung eines Workspace

Ein Azure Synapse Workspace dient als übergeordnete Arbeitsfläche für Azure Synapse Analytics. Azure Synapse lässt sich im Azure Portal über die Suchleiste finden.
Nachdem der Workspace eingerichtet worden ist, haben Sie die Möglichkeit, Azure Synapse Studio zu öffnen. Synapse Studio dient als Navigator-Oberfläche, um Ihre Daten abzurufen, Ihre Pipelines zu erstellen und den gesamten Service zu überwachen und zu verwalten.

2. Erstellung von SQL-Skripten und Notebooks

Im sogenannten "HUB" lassen sich SQL-Skripte, Synapse-Notebooks und Datenflüsse erstellen.
SQL-Skripte basieren auf T-SQL (MSSQL mit proprietären mathematischen Erweiterungen) und verwenden bereitgestellte SQL-Pools oder bedarfsgesteuerte serverlose Pools. Neben diesen lassen sich die Notebooks auf dedizierten Spark-Pools erstellen – also Notebooks, die auf die Spark Cluster Ressourcen zurückgreifen, die zuvor erstellt wurden.
Zuletzt lassen sich Datenflüsse erstellen, die aufzeigen, aus welcher Quelle die Daten kommen und wie sie transformiert an das Ziel gelangen. Hierfür bietet Azure Synapse die Option an, integrierte oder Inline-Datensets[1] zu verwenden, und diese anhand von gängigen SQL-Operatoren wie Selects, Aggregation oder Pivotisierung zu transformieren.

Um nun eine Pipeline zu erstellen, wird das HUB „Integrieren" verwendet, welches die Oberfläche für Pipelines abbildet.

3. Erstellung einer Pipeline

Eine Pipeline verknüpft die bereits erstellten Elemente aus Schritt 2 wie SQL-Skripten entweder mit anderen Prozessschritten wie z.B. dem Kopieren oder Konditionierung/Filterung von Daten, oder mit Azure Functions, um einen separaten Schritt zu triggern. Zudem lassen sich diese Datenflüsse in einem DRY-Run Modus ausführen (Validate) sowie einem Debug Modus, der auf mögliche Fehler in der Prozesskette hinweist. Ein besonderer Vorteil ist, dass dies bereits als ETL-Tool einzelne Datenflüsse miteinander kombiniert und zudem einen Trigger verwendet.

4. Speichern in einen Exportbereich

Eine Senke ist ein in Azure Synapse verfügbarer Exportbereich. Dazu gehören beispielsweise Azure Data Lake Storage Gen2 oder Azure Cosmos DB (SQL-API). Die Einstellungen hierfür lassen sich innerhalb eines Datenflusses auswählen und bilden den letzten Schritt der Pipeline ab.

Einordnung des Tools: Was kann Azure Synapse, was die Konkurrenz nicht kann?

Azure Synapse bietet als eine einheitliche Data-Warehouse-Lösung ein Komplettpaket an, mit dem Datenintegration, Datenanalyse und Datenbereitstellung auf einer Plattform durchgeführt werden können. Die Verwendung des Tools ist jedoch mit einigen Risiken und Herausforderungen verknüpft:

! **Verarbeitung von unstrukturierten Daten:** Aktuell beschränkt sich die Verarbeitung auf strukturierte Daten – unstrukturierte Daten Beispielsweise können Mediadaten noch nicht genügend oder gar nicht in die Plattform eingelesen werden. Das primäre Datenbankmodel beschränkt sich hierbei auf ein relationales Datenbankmanagementsystem.

! **Eingeschränkte SQL-Fähigkeiten:** User von SQL-Server Admin Services wie SSMS können keine Admin Tasks oder Logins erstellen. Für diese werden T-SQL Befehle benötigt. Viele SQL-Befehle sind bisher noch nicht verfügbar, was eine Migration von anderen Datenbanken erschweren kann.

! **Längere Lernkurve:** Die Verwendung von Azure Synapse wird voraussichtlich länger als erwartet sein, da Azure Plattform Kenntnisse notwendig sind, um die Architektur und Tools zu verstehen, die innerhalb Synapse miteinander arbeiten.

Im Gegensatz hierzu bietet der Konkurrent ***Databricks***, welcher ebenfalls auf Azure, AWS und GCP verfügbar ist, die gleiche Funktionalität wie Azure Synapse an. Zudem besteht die Möglichkeit der Verarbeitung von NOSQL-Datenquellen (z.B. maschinelle Daten) und Versionierungserweiterungen durch die Integration mit Delta Lake, die es auf Azure Synapse so bisher nicht gibt. Durch die Verwendung von geclusterten und ungeclusterten Columnstore-Indexes erreicht Azure Synapse einen geringeren Bedarf an teuren Speicherressourcen. Diese Ressourcenoptimierung unterstützt Azure Synapse in seiner Massenparallelrechner-Architektur, was für die Geschwindigkeit der Datenverarbeitung von Vorteil ist.
Dabei gilt es, dass Azure Synapse sich vor allem als solides Cloud-basiertes Data Warehouse positioniert. Databricks tritt hingegen eher als Data Lake Analytics Plattform am Markt auf und die Anwendungsfelder befinden sich im Bereich Machine Learning, ELT und Data Science.

1 *Definition: Ein in Azure erstelltes Datenobjekt, welches zu den verknüpften Services von Azure eine Verbindung herstellt.*

Im Bereich Sicherheit kann Azure Synapse wie Databricks punkten. Wo Synapse mit Features aus der Netzwerk-Sicherheit, Sicherung von Authentifizierungsbarrieren oder Zugangskontrolllisten (ACLs) glänzt, besitzt Databricks eine rollenbasierte Zugangskontrolle (RBAC), automatische Verschlüsselung und weitere Sicherheitsfeatures.

Schlussendlich ist jedoch die Verwendung beider Software-Angebote mit der ersten Evaluierung der eigenen Architektur sowie mit der Ermittlung potenzieller Anwendungsfelder im Cloud Bereich verknüpft.

Beratung in Datenmanagement / Customer Analytics und Cloud

Falls Sie sich noch nicht sicher sind, welcher Cloud-DWH-Dienst für Ihre aktuellen Herausforderungen und Ziele der richtige ist, empfiehlt sich ein Assessment in Verbindung mit einer Toolauswahl.
CINTELLIC kann Ihnen helfen, das geeignete Datenmanagement für Ihre Anwendungsfälle bzw. Anforderungen aufzubauen sowie auch zu pflegen.

Falls Sie Unterstützung bei Microsoft Azure Synapse Analytics oder Databricks benötigen, kontaktieren Sie uns gerne für ein unverbindliches Beratungsgespräch.

von Fais Chalo & Zuzanna Kowalska
CINTELLIC Consulting Group

Ansprechpartner

Dr. Jörg Reinnarth
Geschäftsführer
CINTELLIC Consulting Group
joerg.reinnarth@cintellic.com

Stephan Klöckner
Senior Manager
CINTELLIC Consulting Group
stephan.kloeckner@cintellic.com

Über CINTELLIC

Die 2010 gegründete CINTELLIC Consulting Group ist eine auf digitales Kundenmanagement spezialisierte Unternehmensberatung, die ihre Klienten vom ersten Konzept bis zur Umsetzung in der Praxis ganzheitlich begleitet. An den Standorten in Bonn, Frankfurt am Main und München arbeiten über 70 Mitarbeiterinnen und Mitarbeiter.

Zu den Klienten zählen DAX-Konzerne, führende mittelständische Unternehmen und insbesondere zahlreiche sogenannte „Hidden Champions" mit den Branchenschwerpunkten Banken und Versicherungen, Telekommunikation, IT, Medien, Unterhaltung, Handel, E-Commerce, Versorger und Logistik.

www.cintellic.com

#jointheteam

CINTELLIC befindet sich auf Wachstumskurs. Vielleicht mit Ihnen? Jetzt Stellenanzeigen entdecken und bewerben!
https://www.cintellic.com/stellenangebote/

Cintellic im Social Web

Cintellic GmbH
Remigiusstraße 16
53111 Bonn
t +49 228 92 65 18 20
info@cintellic.com
www.cintellic.com

CINTELLIC
CONSULTING GROUP