Azure Synapse Analytics - Eine Einführung

Kategorie
SQL Server

Anna
Mussenbrock

Das neue Update von SQL Server bringt einige Neuerungen mit in die Welt der Datenbankentwicklung. In der kommenden SQL Server 2022 Version gibt es erstmals eine starke Konnektivität zu Cloud Diensten, wie beispielsweise Azure Cloud. Wir stellen Ihnen heute den Azure Synapse Analysedienst - vorher SQL DW - etwas näher vor.

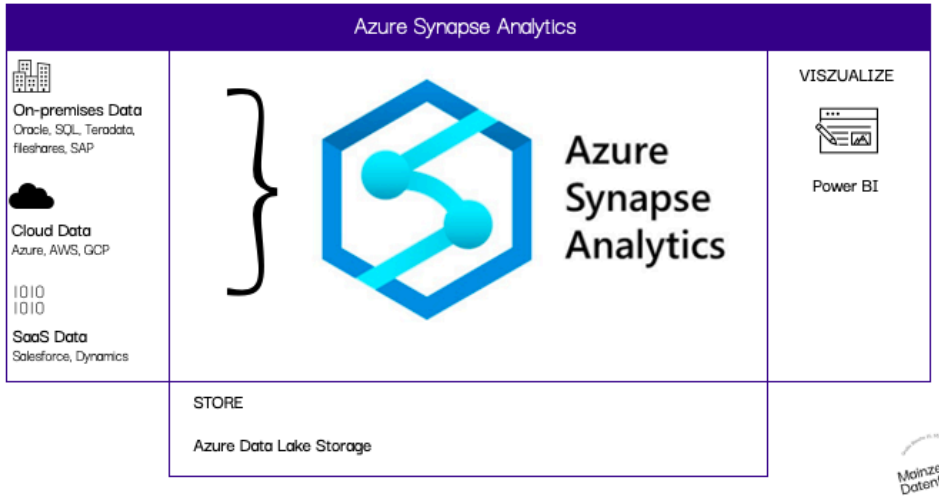
Was ist Azure Synapse Analytics?

Damals noch unter dem Namen "SQL DW" hat der Data Warehouse Dienst einige Iterationen im Hause Azure durchlaufen, sodass er Ende 2019 in Azure Synapse Analytics umbenannt wurde. Dieser Dienst dient hauptsächlich zu unbegrenzten Analysezwecken, der Datenintegration und Data Warehousing auf Unternehmensniveau, kombiniert mit Big-Data. Damit liefert Azure Ihnen eine flexible und individuell anpassbare Datenabfrage für Business-Intelligence und Machine-Learning Anwendungen an. Wichtig zu wissen an diesem Punkt: Für diesen Analysedienst ist kein dedizierter Server notwendig.

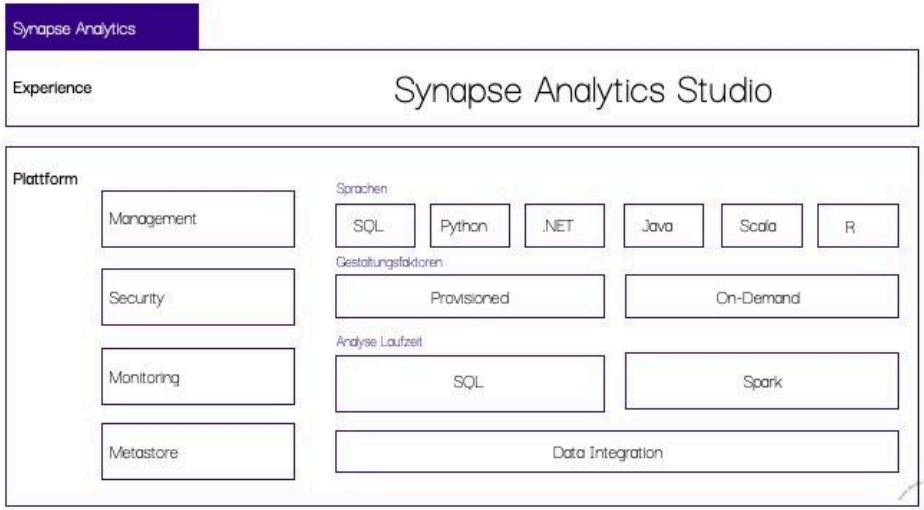
High-Level-Architektur

Azure Synapse Analytics umfasst eine High-Level Architektur mit OLTP- und OLAP-Anwendungen. Online Transaction Procession Workloads (OLTP) beinhalten Transaktionsdaten, die mit einer hohen Anzahl von Lese- & Schreibvorgängen gespickt sind. Dabei lässt sich ein Muster der Datenzugriffe erkennen. Viele skalare und tabellarische Datensätze sind dort zu finden. Weiterhin lässt sich erkennen, dass die Datenaufnahme in der Regel durch Benutzertransaktionen durchgeführt wird. OLAP-Anwendungen (Online Analytical Processing) hingegen speichern und verarbeiten diese Datenmengen aus verschiedenen Quellen. Anschließend werden aus diesen Datensätzen Ad-Hoc Berichte und analytische Anwendungsfälle erstellt. Die Anwendung Azure Data Lake Storage bildet das Fundament zur Big-Data Speicherung und die Visualisierungsebene übernimmt Power BI.

Mainzer Datenfabrik GmbH
Große Bleiche 15, 55116 Mainz
info@madafa.de
www.madafa.de
Tel 06131·3331612



Azure Synapse Komponenten- & Funktionen



Komponenten

1. Synapse Analytics ist im herkömmlichen Sinne eine Analyseplattform mit unbegrenzter Unterstützung von diversen Analyseworkloads.
2. Mit Synapse Workspaces bietet Azure eine integrierte Verwaltungs- & Steuerungskonsole zum Betreiben verschiedener Komponenten und Dienste von Azure Synapse Analytics.
3. Synapse Analytics Studio ist eine webbasierte Entwicklungsumgebung, die eine codefreie oder low-code Arbeit mit Synapse Analytics ermöglicht.

Funktionen

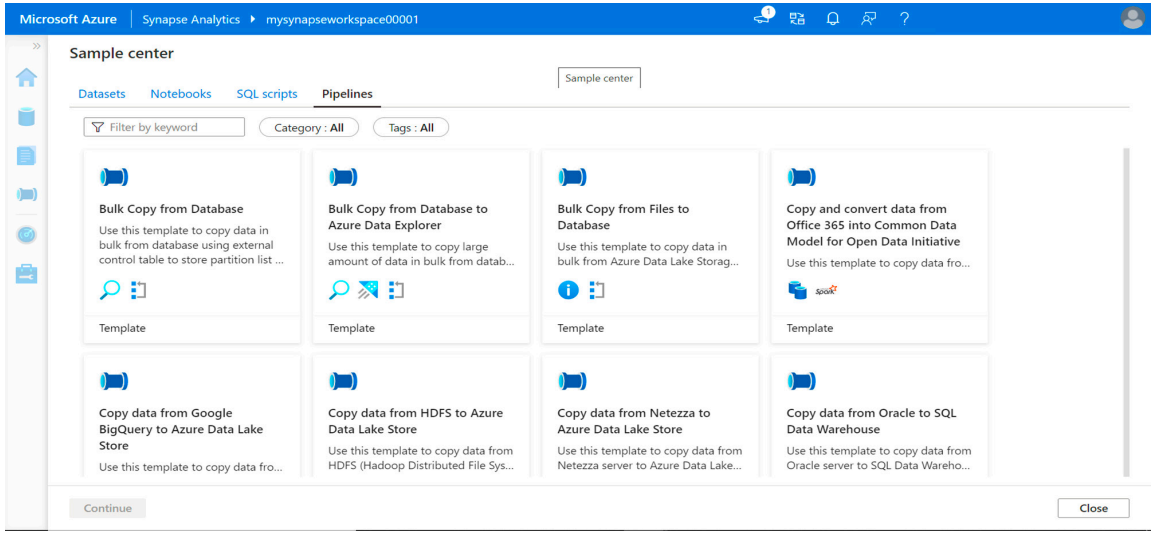
1. Synapse unterstützt eine Reihe von Programmiersprachen, die in der Regel von analytischen Workloads verwenden werden, wie z.B. SQL, Python, .NET, Java, Scala und R.
2. Es werden zwei Arten von Analyselaufzeiten unterstützt. Mit SQL und Spark können beispielsweise Daten im Batch-, Streaming- und interaktiven Modus verarbeitet werden.
3. In Synapse sind zahlreiche Azure-Datendienste integriert, wie z.B. Azure Data Catalog, Azure Lake Storage, Azure Databricks, Azure HDInsight, Azure Machine Learning und Power BI.
4. Zur Überwachung, Verwaltung und Sicherung der Daten bietet Synapse verschiedene integrierte Dienste zum Schutz an.
5. Synapse profitiert von Data Lake Storage Modellen, die als Speicherschicht und Datenquellenschicht fungieren. Daten werden in der Regel aus diesen Data Lake Storages für Analyseworkloads in Synapse eingespielt.

Mainzer Datenfabrik GmbH
Große Bleiche 15, 55116 Mainz
info@madafa.de
www.madafa.de
Tel 06131·3331612

Die Grundpfeiler von Azure Synapse Analytics

Azure Synapse Studio

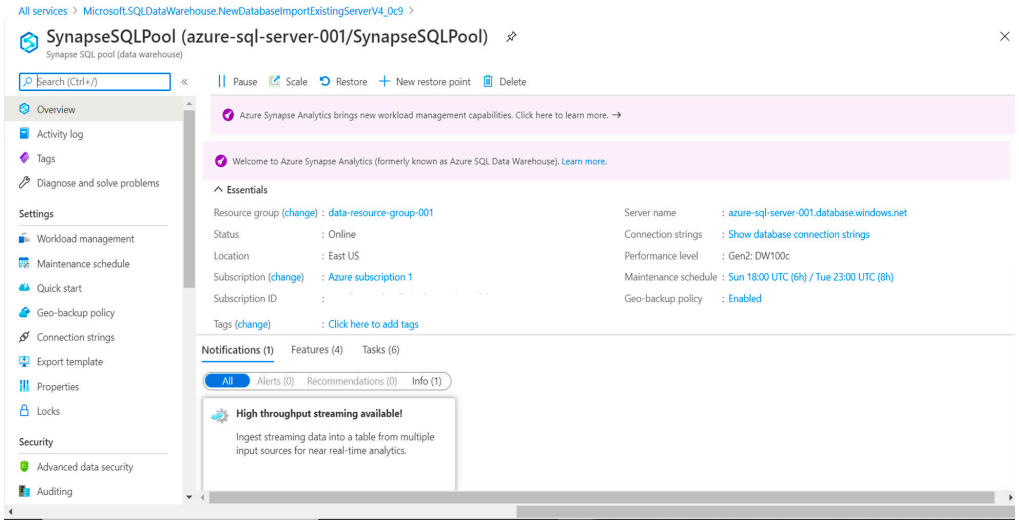
Azure Synapse Studio ist ein webbasiertes SaaS-Tool, das als Verwaltungs- & Steuerungsplattform fungiert und mit dem Entwickler innerhalb einer Konsole alle integrierten Dienste verwenden können. In der analytischen Lösungsentwicklung mit Synapse startet man in der Regel mit einer Erstellung eines Arbeitsbereichs inklusive Zugriff auf verschiedene Synapse Funktionen. Diese Funktionen umfassen zum Beispiel den Datenimport mithilfe verschiedener Mechanismen oder Datenpipelines, sowie das Erstellen von Datenflüssen, die Datensuche und die Datenanalyse mit Spark-Jobs oder SQL-Skripten. Für die Visualisierung der Daten für Reporting- & Dashboarding Zwecke ist die Integration von Power BI verantwortlich. Azure Synapse Studio bietet mithilfe einer CI/CD Integration weitere Funktionen zum Erstellen von Artefakten, Code-Debugging und Leitungsoptimierung.



Mainzer Datenfabrik GmbH
Große Bleiche 15, 55116 Mainz
info@madafa.de
www.madafa.de
Tel 06131·3331612

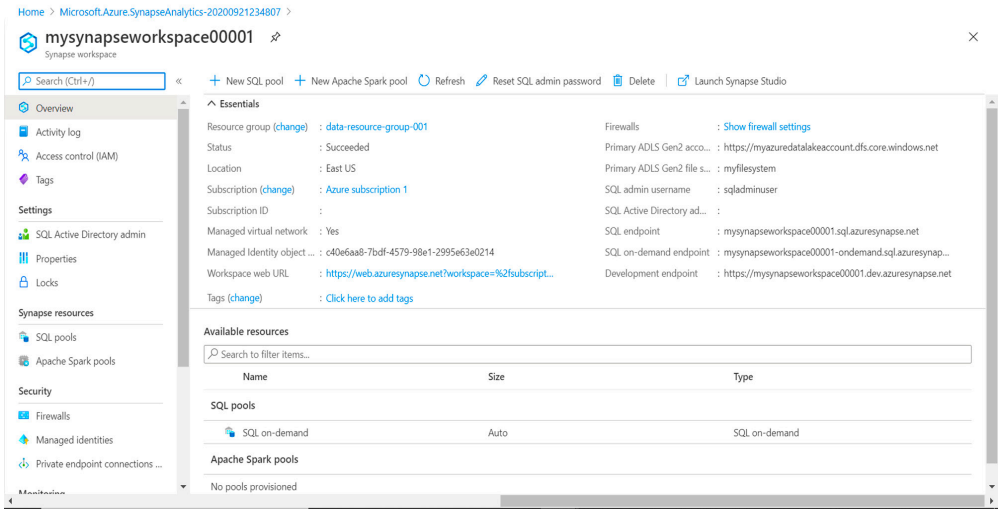
Synapse SQL-Pools

Der Synapse SQL Pool bietet die herkömmlichen Data Warehousing Funktionen, die Sie vermutlich auch noch aus SQL DW kennen. Merkmal des Dienstes ist, dass der Instanz eine feste Kapazität von DWU-Einheiten für die Datenverarbeitung zugewiesen wird. Der Datenimport funktioniert mithilfe verschiedenen Mechanismen, wie z.B. SSIS, Polybase, Azure Data Factory uvm. Synapse speichert die Daten in einem Spaltenformat und ermöglicht Abfragefunktionen, die OLAP-Workloads begünstigen. Weiterhin werden Datenstreaming , wie auch KI-Funktionen von Synapse unterstützt. Grundsätzlich ist der Synapse SQL Pool als Teil einer Azure SQL-Server Instanz zu verstehen und kann z.B. gleichermaßen mit SSMS genutzt werden. Wie auch bereits erwähnt, ist die Verwendung eines Servers in Kombination mit Azure Synapse nicht zwingend notwendig. Azure verwaltet an dieser Stelle die Infrastrukturkapazität eigenständig, um entsprechende Anforderungen der Workloads zu erfüllen. So gestaltet sich auch die Preisgestaltung von Synapse, denn diese richtet sich nach der Anzahl der verarbeiteten Datenmengen, anstatt nach der Anzahl der verwendeten Instanzen.



Apache Spark für Azure Synapse

Wie bereits erwähnt, werden in Azure Synapse zwei Analyselaufzeiten wie Spark und SQL verwendet. Diese sind dafür verantwortlich, dass das Laden von Daten, Datenverarbeitung, Datenvorbereitung, ETLs und andere Aufgaben, die mit dem Data Warehousing zusammenhängen, ausgeführt werden. Data Bricks wird zwar ebenfalls von Azure Synapse bereitgestellt und verfügt über vergleichbare Funktionen wie Spark. Der Vorteil der Nutzung von Spark ist jedoch, dass keine zusätzlichen Cluster zur Datenverarbeitung verwaltet werden müssen. Daten werden automatisiert skaliert und unterstützen weitere Funktionen wie z.B. .NET oder SparkML-Algorithmen, Delta Lake, Azure ML-Integration und Notebooks im Jupyter-Stil. Weiterhin ist eine multilinguale Unterstützung für Sprachen wie C#, Pyspark, Scala, Spark SQL und Java gegeben.



Azure Synapse-Sicherheit

Abgesehen von den ganzen Funktionen, die wir Ihnen in den oberen Absätzen bereits beschrieben haben, ist ein weiterer sehr wichtiger Aspekt, eine Reihe von Sicherheitsfunktionen, die in Synapse enthalten sind. Diese erfüllen bereits fast 30 branchenführende Konformitäten wie ISO, SOC, FedRAMP, DISA, HIPAA, FIPS usw. Sie unterstützen die Azure AD-Authentifizierung, SQL-basierte Authentifizierung sowie die Multifaktor Authentifizierung. Weiterhin wird die Datenverschlüsselung im Ruhe- & Aktivzustand aktiviert, sowie die Datenklassifizierung für sensible Daten. Die Sicherheit ist auf Zeilenebenen, Spaltenebene sowie auf Objektebene zusammen mit einer dynamischen Datenmaskierug geboten. Darüberhinaus unterstützen diese Sicherheitsfunktionen auch auf Netzwerkebene mit virtuellen Netzwerken und Firewalls.

Mainzer Datenfabrik GmbH
Große Bleiche 15, 55116 Mainz
info@madafa.de
www.madafa.de
Tel 06131·3331612

Konfiguration eines Azure Synapse Analytics Workspaces

Aufbauend auf die Einführung in Azure Synapse Analytics, zeigen wir Ihnen nachfolgend, wie Sie dieses Tool verwenden und auf Basis dessen, ein individuelles Workspace erstellen können. Ein Azure Synapse Workspace dient als zentrale Konsole für den Zugriff auf eine Vielzahl an Tools und Features im Zusammenhang mit Azure Analytics. Nach der Implementierung von Azure Synapse Analytics steht als erster Schritt die Konfiguration eines geeigneten Arbeitsbereichs an.

Vorgehen:

1. Navigieren Sie zunächst zu Azure Synapse Analytics Workspace. Die entsprechende Seite öffnet sich.
2. Erstellen Sie zum ersten Mal einen Arbeitsbereich, ist die Seite noch leer. Das wird sich jedoch gleich ändern.
3. Klicken Sie auf die Schaltfläche "Synapse Workspace erstellen". Ein Assistent öffnet sich und wird uns durch das Procedere führen.
4. Der erste Schritt ist die Angabe von grundlegenden Details, wie Azure Abonnement und der Ressourcengruppe. Es könnte bei der Einrichtung des ersten Worspace vorkommen, dass eine Warnung eintrifft, die Sie über eine nicht vorhandene Registrierung des Synapse Ressourcenanbieters informiert. Solange kein Ressourcenanbieter des Dienstes in Ihrem Abonnement registriert ist, kann der Dienst nicht verwendet werden.
5. Klicken Sie auf "Klicken Sie, um sich zu registrieren", um den Synapse Anbieter beim Abonnement zu registrieren.
6. Um Details für die Konfigurationsoptionen bereitzustellen, geben Sie nun den Namen der Ressourcengruppe an, in dem der Arbeitsbereich erstellt werden soll.
7. Geben Sie nachfolgend die Details zum Arbeitsbereich an, wie z.B. Namen des Bereichs und die entsprechende Region, in der der Arbeitsbereich erstellt werden soll.
8. Um über Synapse auf verschiedene Repositorys zugreifen und Daten aus Azure Data Lake Storage herauslesen zu können, benötigen Sie ein Konto und Dateisystem. Haben Sie bereits Azure Data Lake Konto, können Sie über Ihr Abonnement darauf zugreifen. Falls nicht, sollten Sie an dieser Stelle über das Dialogfeld ein neues Konto mit Zugriff auf Mitwirkender-Ebene erstellen.
9. Neben dem Azure Data Lake Konto, benötigen Sie auch ein entsprechendes Dateisystem, auf dem die Daten gespeichert werden können. Klicken Sie dafür auf "Neu erstellen" und folgen den Anweisungen.
10. Gewährleisten Sie den Zugriff auf Contributor-Ebene, um auf Daten aus Azure Data Lake zugreifen zu können. Klicken Sie auf "Weiter", um die Sicherheits- & Netzwerkkonfiguration vorzunehmen.
11. Im Abschnitt "Sicherheit & Netzwerk" müssen Sie im ersten Schritt die Administratorkennung zur Verbindungsherstellung mit den SQL Pools eingeben. Dieser Schritt ist essentiell zur Wahl der Runtimes über SQL oder Spark.
12. Abgesehen von den Anmeldeinformationen, können wir auch Datenpipelines in Azure Synapse integrieren. Damit diese Pipelines auf die SQL-Pools zugreifen können, muss das entsprechende Kontrollkästchen mit dem Titel "Pipelines" angeklickt werden.
13. Um den Datenverkehr zwischen dem Workspace und den Datenquellen nicht über das offene Internet laufen zu lassen, werden wir im nächsten Schritt ein von Synapse verwaltetes, virtuelles Netzwerk aktivieren. Die Aktivierung wird über "Enable managed virtual network" vorgenommen. Wichtig zu Wissen: Es könnten an dieser Stelle Zusatzkosten auf Sie zukommen.
14. Ein weiterer Netzwerkaspekt ist die Auswahl der IP-Adressen, die eine Verbindung zu diesem Workspace herstellen dürfen/können. Hier bestätigen Sie das Kontrollkästchen "Verbindungen von allen IP-Adressen zulassen". Dieser Schritt ist für die Verbindungsherstellung zu webbasierten Tools wie Azure Synapse Studio notwendig.
15. Mit einem Klick auf "Weiter" gelangen Sie in den Abschnitt, der es Ihnen ermöglicht, Tags und Metadaten hinzuzufügen.
16. Im nächsten Schritt gelangen Sie zur Zusammenfassung der konfigurierten Einstellung Ihres Arbeitsbereichs. Überprüfen Sie nun Ihre Einstellungen und bestätigen, sofern alles in Ordnung ist. Beachten Sie, dass beim Erstellen eines Arbeitsbereichs der SQL On-Demand Pool standardmäßig erstellt und bereitgestellt wird. Hierfür fallen Zusatzkosten von 5 USD/TB gescannter Daten an.
17. Bestätigen Sie die Angaben, wird der Arbeitsbereich gemäß Ihren Einstellungen erstellt.
18. Mit einem Klick auf "Gehe zu Ressourcen" gelangen Sie auf das Arbeitsbereich-Dashboard.
19. Das Dashboard dient als Zentrale Plattform und hier können Sie sich alle Eigenschaften und Endpunkte anzeigen lassen, sowie neue SQL Pools erstellen, Anmeldeinformationen zurücksetzen, Firewall Einstellungen ändern und die IP-Adressen Zugriffe verwalten.

Mit dieser Anleitung sollte die Konfiguration eines Azure Synapse Workspace spielend leicht von Statten gehen. Der Assistent leitet Sie durch die einzelnen Teilschritte. Sie müssen lediglich für ein aktives Administratoren Konto in Azure Data Lake verfügen.

Mainzer Datenfabrik GmbH
Große Bleiche 15, 55116 Mainz
info@madafa.de
www.madafa.de
Tel 06131·3331612

Fazit

Alles in Allem ist Azure Synapse ein vollumfassendes Datentool, welches eine integrierte Plattform zur Datenverwaltung- und Verarbeitung bietet, sowie diverse Aufgaben und Prozesse von Analyseworkloads abdeckt. Mit der Erstellung eines Azure Synapse Workspace zeigen wir Ihnen den ersten wichtigen Schritt in der Arbeit mit Azure Synapse Analytics. Das ganze Tool ist modular zusammenstellbar und bietet damit unheimlich viel Flexibilität.

Die Mainzer Datenfabrik ist bestens ausgebildet in Sachen Azure und den dazugehörigen Diensten und Komponenten. Wenn auch Sie mehr über die Möglichkeiten mit Azure erfahren wollen, kontaktieren Sie uns gerne über unser Kontaktformular und vereinbaren Sie ein unverbindliches Beratungsgespräch.

Mainzer Datenfabrik GmbH
Große Bleiche 15, 55116 Mainz
info@madafa.de
www.madafa.de
Tel 06131·3331612