

Lakehouse Analytics

mit Microsoft Fabric und Azure Databricks



Lakehouse Analytics

mit Microsoft Fabric und Azure Databricks

3

Erschließen Sie den Wert Ihrer Daten

13

Programmieren Sie schneller mit GitHub Copilot, Visual Studio Code und Azure Databricks

3

Was ist ein Lakehouse?

17

Integration von Azure Databricks mit OneLake

5

Datenverwaltung und -Analytics mit Microsoft Fabric und Azure Databricks

20

Fazit

7

Vereinheitlichen Sie Daten mit Microsoft Fabric

20

Nächste Schritte

Erschließen Sie den Wert Ihrer Daten

Heutzutage werden Daten oft mit Gold aufgewogen. Unternehmen auf der ganzen Welt versuchen, ihre riesigen Datenbestände zu extrahieren, zu verfeinern und daraus Kapital zu schlagen. Diese Transformationswelle wird durch eine Public-Cloud-Infrastruktur vorangetrieben, die Unternehmen enorme Rechenleistung und Speicherkapazität für die Verarbeitung ihrer wachsenden Datenbestände bietet. Microsoft Azure, die heute führende Cloud-Plattform, bietet Dienste, die Unternehmen in die Lage versetzen, diese modernen Datenherausforderungen zu meistern. An vorderster Front steht Microsoft Fabric, ein neues Analytics-Kraftpaket, das den Bereich des Datenmanagements neu gestaltet. In Verbindung mit Azure Databricks verfügen Unternehmen über eine breite Palette von Tools, mit denen sie die Leistungsfähigkeit der Cloud nutzen und ihre anspruchsvollsten Dateninitiativen in Angriff nehmen können.

Von Dateningenieur*innen bis zu Entscheidungsträger*innen stellt Fabric allen Benutzer*innen die notwendigen Tools und Insights zur Verfügung, um Unternehmen in die Zukunft zu führen. Durch die Bereitstellung einer zusammenhängenden Plattform, die verschiedene Analytics-Funktionen vereint, stellt Fabric sicher, dass Unternehmen das Potenzial ihrer Daten erschließen können, ohne die Komplexität der Verwaltung mehrerer Tools. Mit OneLake, einer bahnbrechenden Datenplattform, ermöglicht Fabric End-to-End-Analytics und bietet eine sichere, verwaltete Speicherschicht für den Aufbau und die Wartung eines modernen Data Lakehouse.

Was ist ein Lakehouse?

Die Lakehouse-Architektur ist der moderne Ansatz zum Aufbau einer skalierbaren Analytics-Plattform für den wachsenden Datenbestand eines Unternehmens. Es kombiniert die Präzision eines herkömmlichen Data Warehouse mit der enormen Größe und Flexibilität eines Data Lake. Durch den Einsatz moderner Tools wie Microsoft Fabric, Microsoft Power BI und Azure Databricks können Unternehmen ein Lakehouse aufbauen, das die Anforderungen von Dateningenieur*innen, Geschäftsanalyst*innen und Datenwissenschaftler*innen erfüllt – alle teilen eine einzige Kopie der Daten, die in einem offenen Format gespeichert und von einem einheitlichen Katalog verwaltet wird.

Die Lakehouse-Architektur auf Azure nutzt die Leistungsfähigkeit von Apache Spark, einer skalierbaren Analytics-Engine für Unternehmen, vollständig aus. Mit den KI-Assistenten von Microsoft Fabric und Azure Databricks können Teams effizienter arbeiten, Daten gemeinsam nutzen, neue Insights gewinnen und fortschrittliche KI-Modelle entwickeln.

Die traditionelle Data-Lake- und Warehouse-Architektur

Data Lakes sind, wie der Name schon sagt, riesige Reservoirs, die eine Vielzahl von Datentypen in ihrer rohen, unverarbeiteten Form speichern. Als zentralisierte Repositories sind Data Lakes darauf ausgelegt, große Datenmengen in ihrer authentischsten Form aufzunehmen und zu speichern. Diese Architektur zeichnet sich durch ihre Flexibilität aus, die es Unternehmen ermöglicht, strukturierte, halbstrukturierte und unstrukturierte Daten ohne die Einschränkungen eines vordefinierten Schemas zu speichern.

Data Warehouses sind strukturierte Repositories, die für effiziente Abfragen und Analysen konzipiert sind. Ein Data Warehouse speichert bereinigte, organisierte, verarbeitete und transformierte Daten und bereitet sie für Business Intelligence (BI)-Anwendungen vor.

Bereich	Data Lake	Data Warehouse
Datenspeicher	Es kann unstrukturierte, halbstrukturierte und strukturierte Daten im Rohformat erfassen und speichern. Ein Data Lake speichert alle Arten von Daten, unabhängig von Quelle und Struktur.	Es können nur strukturierte Daten erfasst und aufbewahrt werden. Ein Data Warehouse speichert Daten in quantitativen Metriken mit seinen Attributen. Daten werden transformiert und bereinigt.
Schema-Definition	Normalerweise wird das Schema definiert, nachdem die Daten gespeichert wurden. Dies bietet eine hohe Agilität und recht einfache Datenerfassung, erfordert jedoch Arbeit am Ende des Prozesses (Schema-on-Read).	Normalerweise wird ein Schema definiert, bevor Daten gespeichert werden. Es erfordert Arbeit zu Beginn des Prozesses, bietet jedoch Leistung, Sicherheit und Integration (Schema-on-Write).
Datenqualität	Alle Daten, die kuratiert sein können oder nicht (z. B. Rohdaten).	Hochgradig kuratierte Daten, die als "zentrale Version der Wahrheit" dienen.
Nutzende	Ein Data Lake ist ideal für Benutzer*innen, die sich intensiv mit Analysen beschäftigen, wie etwa Datenwissenschaftler*innen, Dateningenieur*innen und Datenanalyst*innen.	Ein Data Warehouse ist ideal für operative Benutzer*innen wie Geschäftsanalyst*innen, da es gut strukturiert und einfach zu verwenden und zu verstehen ist.
Preis und Leistung	Die Speicherkosten sind im Vergleich zu einem Data Warehouse relativ niedrig und die Abfrage von Ergebnissen erfolgt schneller.	Die Speicherkosten sind hoch und das Abfragen der Ergebnisse zeitaufwändig.
Barrierefreiheit	Ein Data Lake unterliegt wenigen Einschränkungen und ist leicht zugänglich. Daten können schnell geändert und aktualisiert werden.	Ein Data Warehouse ist von Natur aus strukturiert, was den Zugriff und die Manipulation erschwert.

Tabelle 1: Vergleich von Data Lake und Data Warehouse

Datenverwaltung und -Analytics mit Microsoft Fabric und Azure Databricks

Microsoft Fabric ist eine einheitliche Analytics-Plattform, in der alle von Unternehmen benötigten Daten- und Analytics-Tools gebündelt sind. Azure Databricks ist ein vollständig verwalteter Azure-Erstanbieterdienst, der Analytics und KI in einem Lakehouse in Azure ermöglicht.

Fabric und Azure Databricks bieten beide End-to-End-Lösungen, die es Dateningenieur*innen, Datenwissenschaftler*innen, Datenadministrator*innen, Datenanalyst*innen und Datenkonsument*innen ermöglichen, zusammenzuarbeiten, um wertvolle Insights zu gewinnen. Da beide Plattformen auf Delta Lake-Speicher basieren, können die beiden Dienste zusammenarbeiten und dieselben Kopien der Daten gemeinsam nutzen. In Kombination bieten Microsoft Fabric und Azure Databricks eine leistungsstarke Synergie, die Daten-Analytics, -verarbeitung und KI-Lösungen verbessert. Hier sind fünf konkrete Beispiele dafür, wie Microsoft Fabric besser mit Azure Databricks funktioniert:

- **Effiziente Datenpipelines:** Mit Azure Databricks können Datentechniker*innen die Leistung von Apache Spark mit Photon-Beschleunigung nutzen, um effiziente, skalierbare Datenpipelines zu erstellen. Diese Pipelines können Daten in OneLake einspeisen und so sicherstellen, dass Daten für Analytics und andere Vorgänge jederzeit verfügbar sind. So ist gewährleistet, dass Dateningenieur*innen und Datenwissenschaftler*innen mit Daten auf beiden Plattformen nahtlos arbeiten können.
- **Einheitliche Datenspeicherung mit OneLake:** Azure Databricks kann direkt mit den in OneLake gespeicherten Daten interagieren. Unabhängig davon, ob die Daten aus On-Premises-Systemen stammen oder aus Quellen wie Azure Databricks erfasst werden, bietet OneLake die Tools zur Konsolidierung dieser Daten. Diese Architektur minimiert Datenkopien, bietet konsolidierte Governance und ermöglicht Benutzer*innen die Verwendung bevorzugter Anwendungen wie Azure Databricks für Abfragen und Datenwissenschaft.
- **Synergie zwischen AI und Machine Learning:** Azure Databricks hilft beim Aufbau komplexer KI- und Machine Learning-Modellen. In Kombination mit KI-basierten Copilot-Funktionen in Fabric können Unternehmen effizienter Insights aus ihren Daten gewinnen. Diese Synergie stellt sicher, dass Datenwissenschaftler Modelle in Azure Databricks erstellen und bereitstellen und dann Fabric für BI-Analytics verwenden können, indem sie die KI-Funktionen beider Plattformen nutzen.

- **Flexibilität bei der Datenspeicherung:** Es gibt zwei mögliche Ansätze, Daten mit Azure Databricks in OneLake zu laden:
 - **ADLS-basiertes Lakehouse:** Daten werden in Delta Lake-Tabellen in einem ADLS-Konto (Azure Data Lake Storage) gespeichert. Anschließend kann in der Fabric Lakehouse-Datenbank eine Verknüpfung zu diesem Speicher erstellt werden, um sicherzustellen, dass Daten in Fabric einfach und dennoch sicher zugänglich sind.
 - **OneLake-basiertes Lakehouse:** Daten werden direkt am ADLS-Standort von OneLake gespeichert. Bei diesem Ansatz muss der Standardspeicherort für Fabric Lakehouse ermittelt werden. Sobald Daten an diesem Ort gespeichert sind, können sie innerhalb von Fabric abgerufen und verwendet werden, wodurch eine nahtlose Integration zwischen den beiden Plattformen gewährleistet wird.
- **Optimierte Datenvisualisierung:** Sobald die Daten gespeichert und verarbeitet sind, folgt ein entscheidender Schritt, die Daten müssen visualisiert werden. Durch die Integration von Azure Databricks und Fabric können Benutzer*innen problemlos Power BI-Datensätze auf den verarbeiteten Daten erstellen. Diese Integration stellt sicher, dass Datenwissenschaftler*innen und Analyst*innen ihre Daten visualisieren, Insights ableiten und fundierte Entscheidungen auf der Grundlage der von beiden Plattformen bereitgestellten Analytics treffen können.

Die Integration zwischen Fabric und Azure Databricks bietet eine umfassende Lösung für die Data-Analytics-Anforderungen eines Unternehmens. Mit der Veröffentlichung der öffentlichen Vorschau von OneLake besteht für diese beiden Kraftpakete das Potenzial, Analytics-Aufgaben zu vereinfachen.

OneLake: der Kern von Microsoft Fabric

OneLake, auch bekannt unter dem Namen Microsoft Fabric Lake, ist das grundlegende Element aller Fabric-Dienste. Es bietet einen einheitlichen Speicher-Hub für Unternehmensdaten, der auf dem robusten ADLS Gen2 basiert. OneLake richtet sich an eine vielfältige Basis von Benutzer*innen, von erfahrenen Profis bis hin zu angehenden Entwickler*innen, wobei das Hauptziel darin besteht, Datensilos aufzulösen. Es fördert die einfache Datenerkennung und -freigabe und gewährleistet eine zentralisierte Sicherheitskonformität.

Vereinheitlichen Sie Daten mit Microsoft Fabric

Microsoft Fabric ist weit mehr als ein herkömmliches Analytics-Tool. Die zentrale Plattform optimiert den Analytics-Prozess, von der Datenintegration bis hin zu Echtzeit-Insights. Als umfassende Analytics-Lösung für Unternehmen macht Fabric unterschiedliche anbieterspezifische Dienste überflüssig. Microsoft Fabric vereint neue und bestehende Komponenten von Power BI, Azure Synapse und Azure Data Factory in einer einzigen SaaS-Plattform (Software as a Service) mit einheitlicher User Experience.

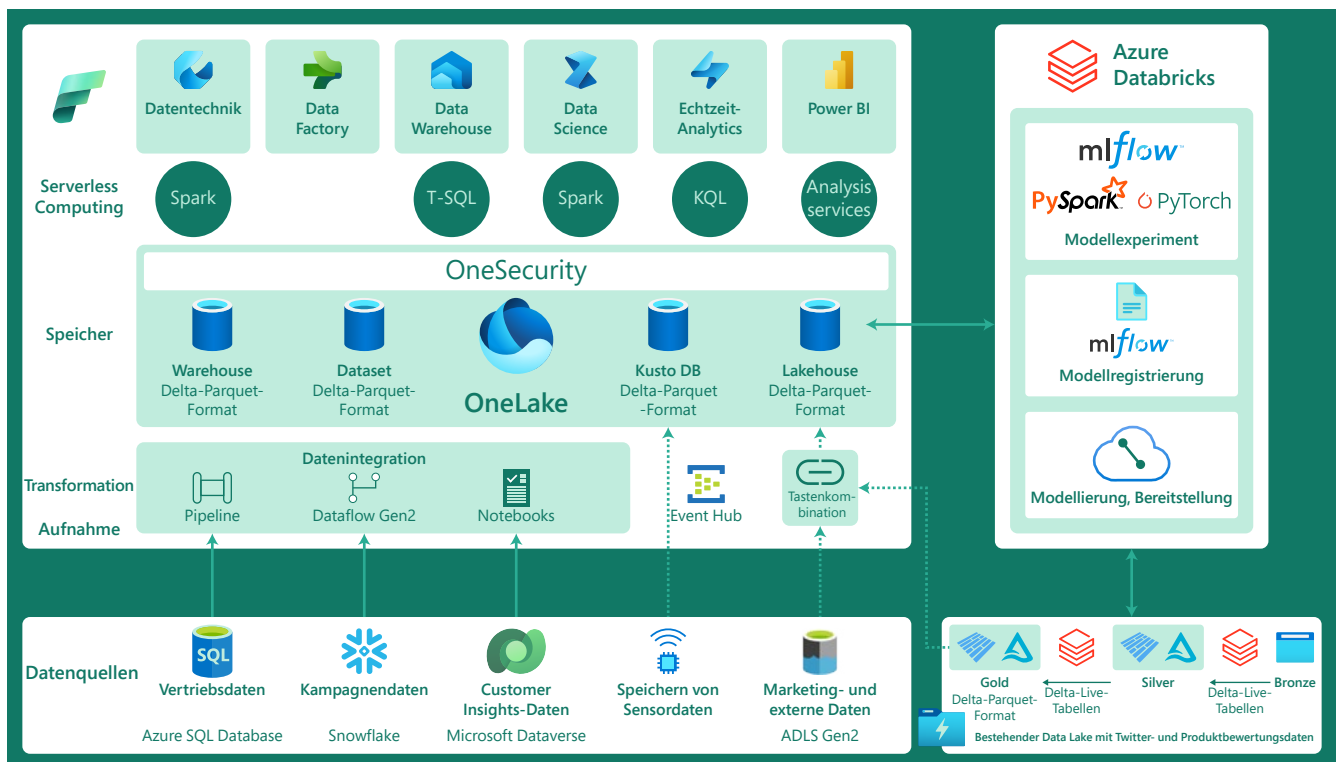


Abbildung 1: Architektur von Microsoft Fabric und Azure Databricks

Komponenten von Microsoft Fabric

Microsoft Fabric bietet End-to-End-Analytics-Lösungen. Es verbindet alle Datenquellen und Analytics-Dienste durch branchenführende Erfahrungen in einer Vielzahl von Kategorien.

- **Synapse Data Engineering:** Synapse Data Engineering ist eine Kernerfahrung von Microsoft Fabric und stellt Dateningenieur*innen eine erstklassige Spark-Plattform zur Verfügung, die groß angelegte Datentransformationen innerhalb des Lakehouse-Paradigmas erleichtert. Es optimiert den Prozess der Aufnahme, Transformation und Weitergabe von Unternehmensdaten in einem offenen Format.
- **Data Factory:** Data Factory ist für komplexe hybride Extract-Transform-Load (ETL)- und Datenintegrationsprojekte konzipiert. Sie hilft bei der Umwandlung unorganisierter Rohdaten in verwertbare Insights. Data Factory ermöglicht die Erstellung datengesteuerter Workflows, sogenannte Pipelines, um Daten aus verschiedenen Datenspeichern aufzunehmen. Diese Pipelines können Datenflüsse oder Compute-Dienste wie Azure HDInsight Hadoop, Azure Databricks und Azure SQL-Datenbank verwenden, um Daten zu transformieren.
- **Synapse Data Science:** Synapse Data Science erschließt den Wert von Daten im Analytics-Workflow eines Unternehmens. Es bietet Tools und Funktionen, die es Datenwissenschaftler*innen ermöglichen, mithilfe von Azure Machine Learning-Integrationen nahtlos Modelle für Machine Learning zu erstellen, bereitzustellen und zu verwalten.
- **Synapse Data Warehouse:** Durch die Integration mit Power BI bietet Fabric Data-Warehousing-Funktionen der nächsten Generation. Es unterstützt ein offenes Datenformat und stellt sicher, dass Dateningenieur*innen und IT-Teams nahtlos zusammenarbeiten können, indem es Daten im offenen Delta-Lake-Format speichert und so eine erstklassige SQL-Leistung gewährleistet.
- **Synapse Real-Time Analytics:** Synapse Real-Time Analytics ist für die Verarbeitung von Beobachtungsdaten aus verschiedenen Quellen konzipiert und verarbeitet halbstrukturierte Daten in großen Mengen. Es bietet Insights in Echtzeit und stellt sicher, dass Unternehmen fundierte Entscheidungen auf der Grundlage der neuesten Daten treffen können.
- **Power BI:** Power BI ist eine umfassende Sammlung von Softwarediensten, Apps und Konnektoren, die darauf ausgelegt sind, unabhängige Datenquellen in kohärente, visuell immersive und interaktive Insights umzuwandeln. Unabhängig davon, ob sich die Daten in einer Excel-Tabelle oder in einer Kombination aus Cloud-basierten und lokalen Data Warehouses befinden, ermöglicht Power BI eine einfache Verbindung, Visualisierung und den Austausch von Insights. Die Kompatibilität sowohl mit On-Premises- und Cloud-Lösungen stellt sicher, dass Unternehmen je nach ihren spezifischen Anforderungen wählen können, wo sie ihre Berichte speichern möchten.
- **Data Activator:** Data Activator bietet ein No-Code-Erlebnis und ermöglicht es Geschäftsanalyst*innen, Aktionen automatisch aus Daten zu steuern. Dies vereinfacht den Prozess der Datenaktivierung und erleichtert es Benutzer*innen, aus ihren Daten umsetzbare Insights abzuleiten.

Vorteile von Microsoft Fabric

Unternehmen sind ständig auf der Suche nach Tools, die große Datenmengen nahtlos integrieren und Insights liefern können, die Entscheidungsfindung und Innovation vorantreiben. Während viele Lösungen umfassende Analytics-Funktionen versprechen, bietet Microsoft Fabric ein wirklich integriertes Erlebnis, das auf die vielfältigen Anforderungen moderner Unternehmen zugeschnitten ist. Es ist auf mehr Einfachheit und Integration ausgelegt und bietet eine umfassende Suite von Analytics-Diensten. Im Vergleich zu herkömmlichen Data Lakes und Warehouses bietet Microsoft Fabric mehrere Vorteile:

- **Umfassende Analytics Suite:** Fabric dient als umfassende Analytics-Lösung, die Datenbewegung, Datenwissenschaft, Echtzeit-Analytics und BI abdeckt und den Analytics-Prozess rationalisiert.
- **Integrierte Infrastruktur:** Fabric bietet ein integriertes Ökosystem von Diensten, einschließlich Data Lake, Data Engineering und Datenintegration. Dies gewährleistet einen reibungslosen Datenfluss in der gesamten Analytics-Pipeline, reduziert Datensilos und fördert die Effizienz.
- **Vereinfachte Bereitstellung und Verwaltung:** Als SaaS-Plattform verwaltet Fabric die Einrichtung und Wartung der Infrastruktur, sodass sich Unternehmen auf Analytics statt auf Infrastrukturmanagement konzentrieren können.
- **Skalierbarkeit und Flexibilität:** Fabric lässt sich problemlos skalieren, kann unterschiedliche Workloads bewältigen und sorgt so für eine effiziente Datenverarbeitung ohne Leistungsprobleme.
- **Echtzeit-Analytics:** Fabric unterstützt Echtzeit-Analytics und ermöglicht es Unternehmen, Insights aus Live-Datenströmen abzuleiten und zeitnah fundierte Entscheidungen zu treffen.
- **Erweiterte Data-Science-Funktionen:** Die Plattform bietet Datenwissenschaftler*innen Tools, mit denen sie komplexe Datenanalysen durchführen, Modelle für Machine Learning entwickeln und vorausschauende Erkenntnisse gewinnen können.
- **End-to-End-Lösung:** Microsoft Fabric bietet eine ganzheitliche Analytics-Lösung, die den gesamten Workflow von der Datenerfassung bis zur Visualisierung optimiert.
- **Sicherheit und Compliance:** Als Teil des Microsoft-Ökosystems gewährleistet Fabric robuste Sicherheitsmaßnahmen und Compliance-Zertifizierungen und schützt so sensible Daten.
- **Benutzungsfreundliche Oberfläche:** Die auf Barrierefreiheit ausgelegte Benutzeroberfläche der Plattform richtet sich sowohl an technische als auch an nicht-technische Benutzer*innen und demokratisiert die Daten-Analytics unternehmensübergreifend.
- **Kosteneffizienz:** Durch die Integration einer Reihe von Diensten kann Fabric im Vergleich zur Verwaltung mehrerer separater Analytics-Lösungen Kosteneinsparungen erzielen.

Durch das Angebot dieser Dienste und Funktionen stellt Fabric sicher, dass Unternehmen über einen umfassenden Satz an Tools verfügen, mit denen sie alle ihre Analytics-Anforderungen erfüllen können, von der Datenintegration bis zur Ableitung von Insights in Echtzeit.

Integration von KI

Azure OpenAI Service basiert auf der speziell entwickelten, KI-optimierten Infrastruktur von Azure. Diese Infrastruktur unterstützt verschiedene Microsoft-Produkte. Die KI-Integration ist ein Eckpfeiler von Microsoft Fabric und unterscheidet es von anderen Analytics-Plattformen:

- **Einheitliche KI-gestützte Plattform:** Fabric ist nicht einfach nur eine Analytics-Plattform, sondern eine KI-gestützte Analytics-Plattform. Das bedeutet, dass jede Ebene von Microsoft Fabric mit KI-Funktionen integriert ist, was die Daten-Analytics intuitiver, effizienter und leistungsstärker macht. So können Unternehmen das volle Potenzial ihrer Daten ausschöpfen.
- **Generative KI und Sprachmodelle:** Microsoft Fabric verwendet generative KI und Sprachmodelldienste, wie beispielsweise Azure OpenAI Service. Diese Dienste ermöglichen die Schaffung alltäglicher KI-Erlebnisse, die die Art und Weise, wie Mitarbeitende ihre Zeit nutzen, neu definieren. Um beispielsweise organisationsspezifische KI-Erlebnisse zu ermöglichen, ist ein stetiger Strom sauberer Daten aus einem gut integrierten Analytics-System erforderlich. Microsoft Fabric sorgt dafür, dass diese Daten schnell verfügbar sind und effizient verarbeitet werden.
- **Copilot in Microsoft Fabric:** Eines der herausragenden Features von Microsoft Fabric ist Copilot. Da Copilot mit jedem Datenerlebnis von Microsoft Fabric integriert ist, können Benutzer*innen Konversationssprache verwenden, um eine Reihe von Aufgaben auszuführen. Dazu gehören das Erstellen von Datenflüssen und Pipelines, das Generieren von Code, das Erstellen von Modellen für Machine Learning und das Visualisieren von Ergebnissen. Dieser dialogorientierte KI-Ansatz vereinfacht komplexe Aufgaben und macht die Plattform benutzerfreundlicher. Darüber hinaus hält Copilot die von Microsoft eingegangenen Verpflichtungen in Bezug auf Datensicherheit und Datenschutz ein und stellt so sicher, dass die Unternehmensdaten geschützt bleiben.
- **KI-gesteuerte Einblicke:** Die Integration von Microsoft Fabric und Power BI ermöglicht KI-gesteuerte Analytics und ermöglicht es Geschäftsanalyst*innen und Benutzer*innen, wertvolle Insights aus Daten abzuleiten. Das Power BI-Erlebnis innerhalb von Microsoft Fabric ist außerdem tief mit Microsoft 365 integriert, sodass sichergestellt ist, dass Insights dort verfügbar sind, wo Geschäftsbenutzer*innen tätig sind.
- **Datenaktivator:** Obwohl diese Funktion bald verfügbar ist, bietet Data Activator in Fabric eine Echtzeiterkennung und -überwachung von Daten. Es löst Benachrichtigungen und Aktionen basierend auf festgelegten Mustern in den Daten aus, alles in einer No-Code-Umgebung. Diese KI-gesteuerte Funktion verbessert die Fähigkeiten der Plattform weiter, zeitnahe und relevante Erkenntnisse bereitzustellen.

Verantwortungsvoller KI-Ansatz

Microsoft erkennt die Bedeutung verantwortungsvoller KI an, insbesondere bei leistungsstarken Technologien wie generativen Modellen. Sie haben einen iterativen Ansatz für große Modelle gewählt und arbeiten eng mit OpenAI und ihren Kund*innen zusammen, um Anwendungsfälle zu bewerten und potenzielle Risiken anzugehen.

Azure OpenAI Service hat Leitplanken implementiert, die auf die [Grundsätze von Microsoft für eine verantwortungsvolle KI](#) ausgerichtet sind. Entwickler*innen müssen ihren beabsichtigten Anwendungsfall oder ihre beabsichtigte Anwendung beschreiben, bevor sie auf den Dienst zugreifen. Es sind Inhaltsfilter vorhanden, um sowohl die für den Dienst bereitgestellten Eingaben als auch die generierten Inhalte zu überwachen und sicherzustellen, dass sie den Richtlinienrichtlinien entsprechen.

Microsoft hat die folgenden sechs Prinzipien identifiziert, die die Entwicklung und Nutzung von KI leiten sollen:

- **Fairness:** Microsoft betont die Bedeutung von Fairness in KI-Systemen. Das Unternehmen möchte sicherstellen, dass KI-Systeme Chancen, Ressourcen oder Informationen auf eine Weise zuweisen, die allen Benutzer*innen gerecht wird.
- **Zuverlässigkeit und Sicherheit:** Microsoft legt großen Wert auf die Zuverlässigkeit und Sicherheit: von KI-Systemen, um sicherzustellen, dass sie konsistent funktionieren und potenzielle Schäden verhindern.
- **Privatsphäre und Sicherheit:** Der Schutz der Benutzerdaten steht an erster Stelle. Microsoft stellt sicher, dass KI-Systeme die höchsten Datenschutz- und Sicherheitsstandards einhalten und Benutzerinformationen vor unbefugtem Zugriff und Verstößen schützen.
- **Inklusion:** Microsoft glaubt an die Entwicklung von KI-Systemen, die allen gerecht werden und sicherstellen, dass alle Benutzer*innen, unabhängig von ihrem Hintergrund oder ihren Fähigkeiten, von der Technologie profitieren können.
- **Transparenz:** Microsoft setzt sich dafür ein, KI-Systeme transparent zu machen und sicherzustellen, dass Benutzer*innen verstehen, wie diese Systeme funktionieren und Entscheidungen treffen.
- **Rechenschaftspflicht:** Microsoft übernimmt die Verantwortung für die von ihm entwickelten KI-Systeme und stellt sicher, dass Mechanismen vorhanden sind, um alle Probleme oder Bedenken im Zusammenhang mit der Technologie anzugehen.

Im Wesentlichen priorisieren diese Grundsätze zur Steuerung einer verantwortungsvollen KI-Entwicklung Fairness, Sicherheit, Datenschutz, Inklusion, Transparenz und Verantwortlichkeit und stellen sicher, dass KI-Technologien vertrauenswürdig und für alle Benutzer*innen von Vorteil sind.

Vereinfachtes Preismodell

Microsoft Fabric hat ein optimiertes Preismodell entwickelt, das darauf ausgelegt ist, die Effizienz zu optimieren und sich an den finanziellen Zielen des Unternehmens auszurichten. Im Mittelpunkt dieser Strategie steht der Erwerb eines einzigen Satzes Rechenguthaben, der für die Unterstützung aller Vorgänge innerhalb des Microsoft Fabric Frameworks bestimmt ist.

Das Preismodell basiert auf einem einheitlichen Compute-Credits-System. Ob Datentechnik, Integration oder Analytics – Fabric verfügt über einen konsolidierten Kredit-Pool. Dieser Ansatz bietet betriebliche Flexibilität und potenzielle Kosteneinsparungen. Während es spezielle Tools gibt, die auf spezifische Bedürfnisse zugeschnitten sind, integriert Fabric verschiedene Funktionen in einer kosteneffizienten Plattform und erleichtert so die Finanzplanung. Es kann jedoch zu Herausforderungen kommen, wenn verschiedene Abteilungen, wie etwa Dateningenieur*innen und Geschäftsanalyst*innen, über unterschiedliche Budgets verfügen. Die kollaborative Entscheidungsfindung wird für eine effiziente Kreditvergabe unerlässlich.

Das umfassende Preismodell von Microsoft Fabric bietet:

- **Wirtschaftlichkeit:** Die Zusammenführung verschiedener Analytics-Aufgaben in einer Preisstruktur kann zu erheblichen Einsparungen führen. Dadurch entfällt die mühsame Nachverfolgung einzelner Servicekosten und sorgt für besser vorhersehbare Kosten.
- **Anpassungsfähigkeit:** Das Shared-Credit-System ist vielseitig. Gutschriften können dorthin geleitet werden, wo sie am meisten benötigt werden, und so an die sich ständig ändernden Geschäftsanforderungen angepasst werden.
- **Unkomplizierte Auswahl:** Ein einheitliches Preismodell durchbricht das Labyrinth der vielfältigen Servicekosten, rationalisiert Entscheidungen und reduziert administrative Aufgaben.
- **Steigerung der Plattformnutzung:** Microsoft Fabric bietet unkomplizierte Preise und umfassende Lösungen, die für Unternehmen ein Magnet sein können.
- **Zukunftsfähigkeit:** Um der Expansion von Unternehmen und ihren sich verändernden Datenanforderungen gerecht zu werden, hilft dieses Preismodell Unternehmen dabei, Wachstum vorherzusagen und den variablen Analytics-Bedarf zu verstehen.

Programmieren Sie schneller mit GitHub Copilot, Visual Studio Code und Azure Databricks

Tools wie GitHub Copilot, unterstützt von OpenAI Codex, verändern die Art und Weise, wie Unternehmen Systeme entwerfen und nutzen. In Kombination mit Anwendungsfunktionen in Visual Studio Code bietet Copilot Entwickler*innen ein erstklassiges Erlebnis. Doch um KI-Systeme wie GitHub Copilot zu erstellen, reicht ein einfacher Code-Editor nicht aus. Hier bietet sich die Databricks-Erweiterung für Visual Studio Code an, mit der Entwickler*innen in Visual Studio Code arbeiten, aber die Leistungsfähigkeit von Azure Databricks für die Bereitstellung ihrer Projekte nutzen können. Es ist eine Mischung aus einfacher Entwicklung und leistungsstarker Ausführung mit den Codierungstipps von Copilot.

Erstellen Sie einen KI-Codierungshelfer für Azure Databricks

Dieses Tutorial zeigt, wie Sie mit Visual Studio Code, GitHub Copilot und der Databricks-Erweiterung für Visual Studio Code schneller programmieren können. Mit der Databricks-Erweiterung für Visual Studio Code kann jeder in Visual Studio Code erstellte Code reibungslos auf Azure Databricks ausgeführt werden. Die Erweiterung hilft außerdem dabei, Modell zu protokollieren, was die Bereitstellung vereinfacht.

Um diesem Beispiel zu folgen, müssen Sie Folgendes tun:

- Visual Studio Code 1.69.1 oder höher muss installiert und für die Python-Codierung konfiguriert sein.
- Stellen Sie sicher, dass Visual Studio Code ausgeführt wird und ein lokales Projekt geöffnet ist.
- Generieren Sie ein persönliches Azure Databricks-Zugriffstoken für Ihren Azure Databricks-Zielarbeitsbereich.
- Fügen Sie Ihr persönliches Azure Databricks-Zugriffstoken als Token-Feld zusammen mit der URL Ihrer Arbeitsbereichsinstanz (z.B. `https://dbc-a7h3345c-d6e7.cloud.databricks.com`) dem `DEFAULT`-Konfigurationsprofil in der lokalen `.databrickscfg`-Datei hinzu.
- Stellen Sie sicher, dass Visual Studio Code für die Python-Codierung eingerichtet ist. Dies umfasst normalerweise die Installation der Python-Erweiterung vom Visual Studio Code Marketplace und die Konfiguration des Python-Interpreters.
- Generieren Sie ein persönliches Azure Databricks-Zugriffstoken für Ihren Azure Databricks-Zielarbeitsbereich. Dieses Token wird zur Authentifizierung Ihres Visual Studio Codes mit Azure Databricks verwendet.

- Zugriff auf GitHub Copilot. Weitere Informationen zu einer kostenlosen Testversion finden Sie hier: <https://github.com/features/copilot>.

Installieren Sie die Databricks-Erweiterung für Visual Studio Code

So installieren Sie die Databricks-Erweiterung für Visual Studio Code:

1. Öffnen Sie Visual Studio Code.
2. Gehen Sie zur Ansicht `Extensions`, indem Sie `Ctrl+Shift+X` drücken oder auf das quadratische Symbol in der Seitenleiste klicken.
3. Geben Sie in der Suchleiste `Databricks extension for Visual Studio Code` ein.
4. Suchen Sie in den Suchergebnissen die Erweiterung und klicken Sie auf die grüne Schaltfläche `Install`.
5. Warten Sie, bis die Installation abgeschlossen ist.

Konfigurieren Sie die Databricks-Erweiterung für Visual Studio Code

Führen Sie die folgenden Schritte aus, um die Databricks-Erweiterung für Visual Studio Code zu konfigurieren:

1. Klicken Sie in der Seitenleiste von Visual Studio Code auf das Azure Databricks-Logo.
2. Klicken Sie im angezeigten Bereich `Configuration` auf `Configure Databricks`.
3. Es öffnet sich eine Befehlspalette, in der Sie nach dem `Databricks Host` gefragt werden. Geben Sie die URL Ihres Azure Databricks-Arbeitsbereichs ein, und betätigen Sie die Eingabetaste.
4. Als Nächstes werden Sie zur Authentifizierung aufgefordert. Klicken Sie auf `DEFAULT: Authenticate using the DEFAULT profile`.
5. Kehren Sie zum Einrichten des Azure Databricks-Clusters zum Bereich `Configuration` zurück, klicken Sie auf `Cluster` und dann auf das angezeigte Zahnradsymbol. Dadurch können Sie den Cluster auswählen und konfigurieren, den Sie verwenden möchten.
6. Wenn der ausgewählte Cluster nicht ausgeführt wird, klicken Sie auf das Symbol `Start Cluster` neben `Cluster`.
7. Richten Sie nun das Synchronisierungsziel ein. Klicken Sie im Bereich `Configuration` auf `Sync Destination` und dann auf das Zahnradsymbol. Es erscheint eine Befehlspalette mit einem zufällig generierten Namen des Synchronisierungsziels. Klicken Sie darauf, um es auszuwählen.

Installieren der GitHub Copilot-Erweiterung

Um die GitHub Copilot-Erweiterung zu installieren, führen Sie die folgenden Schritte aus:

1. Öffnen Sie Visual Studio Code:
2. Drücken Sie `Ctrl+Shift+X` oder klicken Sie auf das quadratische Symbol in der Seitenleiste, um die Erweiterungsansicht zu öffnen.
3. Geben Sie `GitHub Copilot` in der Suchleiste ein und suchen Sie die GitHub Copilot-Erweiterung in den Suchergebnissen.
4. Klicken Sie auf die grüne Schaltfläche `Install` neben der GitHub Copilot-Erweiterung und warten Sie, bis die Installation abgeschlossen ist.

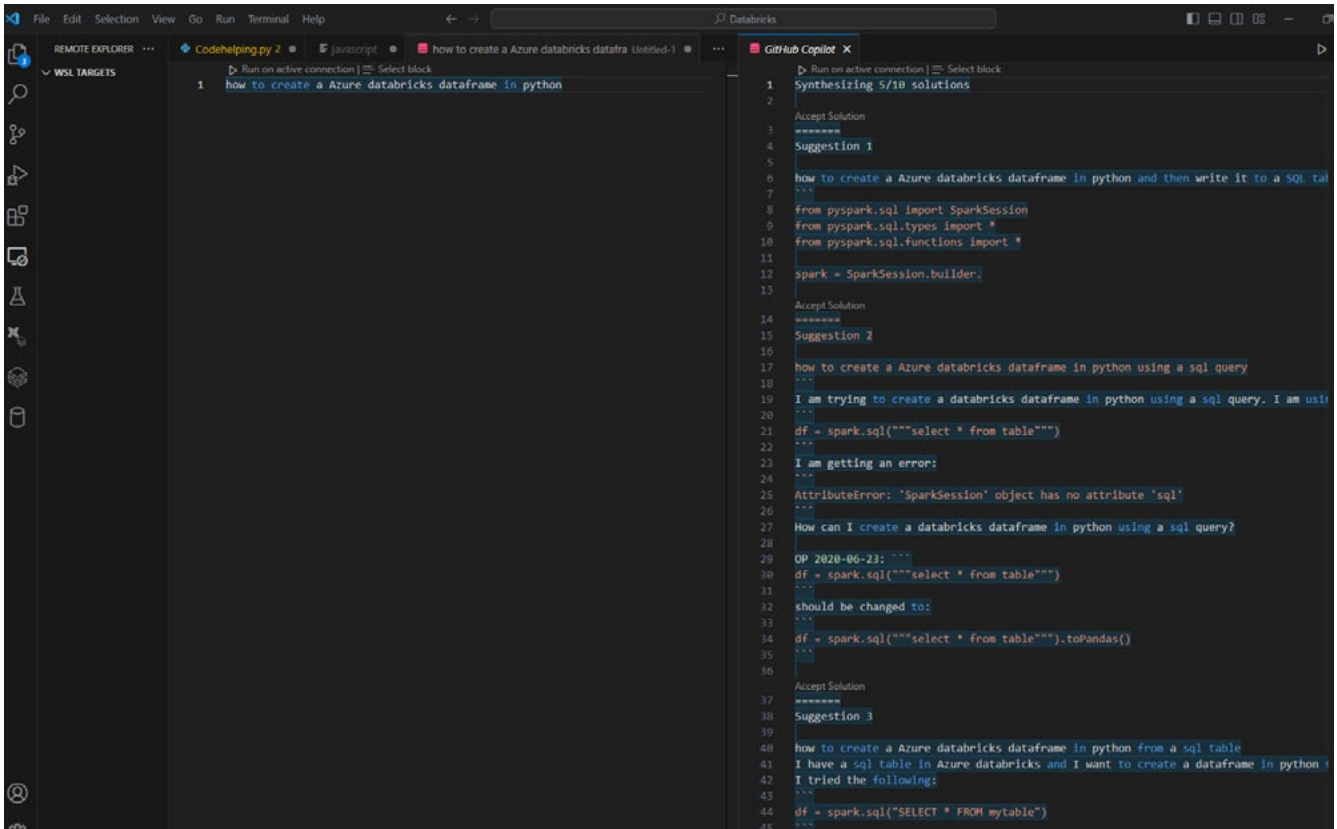
Nach der Installation werden Sie von der GitHub Copilot-Erweiterung aufgefordert, sich mit Ihren GitHub-Anmeldeinformationen zu authentifizieren. Befolgen Sie die Anweisungen auf dem Bildschirm, um sich anzumelden und die erforderlichen Berechtigungen zu erteilen. Nachdem beide Erweiterungen installiert und konfiguriert sind, können Sie Code in Visual Studio Code entwickeln und dabei die Funktionen von Azure Databricks und die KI-gestützten Vorschläge von GitHub Copilot nutzen.

Schreiben eines Python-Skripts mit KI-gestützten Vorschlägen

Befolgen Sie diese Schritte, um ein Python-Skript mit KI-gestützten Vorschlägen zu generieren:

1. Erstellen Sie eine neue Python-Datei.
2. Klicken Sie in Visual Studio Code auf `File > New File`.
3. Speichern Sie die Datei mit der Erweiterung `.py`, z. B. `script.py` oder `codehelping.py`.
4. Beginnen Sie mit der Eingabe Ihres Python-Codes. Wenn Sie mit der Eingabe beginnen, stellt GitHub Copilot automatisch Codevorschläge in Echtzeit bereit.
5. Geben Sie zum Beispiel `"How to create a Azure databricks dataframe in python"` ein.

6. Sobald Sie dazu aufgefordert werden, sehen Sie ein ähnliches Ergebnis wie in *Abbildung 2*.



```
1 how to create a Azure databricks dataframe in python
```

```
1 Synthesizing 5/10 solutions
2
3 Accept Solution
4 =====
5 Suggestion 1
6 how to create a Azure databricks dataframe in python and then write it to a sql table
7 """
8 from pyspark.sql import SparkSession
9 from pyspark.sql.types import *
10 from pyspark.sql.functions import *
11
12 spark = SparkSession.builder
13
14 Accept Solution
15 =====
16 Suggestion 2
17 how to create a Azure databricks dataframe in python using a sql query
18 """
19 I am trying to create a databricks dataframe in python using a sql query. I am using
20 """
21 df = spark.sql("""select * from table""")
22 """
23 I am getting an error:
24 """
25 AttributeError: 'SparkSession' object has no attribute 'sql'
26 """
27 How can I create a databricks dataframe in python using a sql query?
28
29 OP 2020-06-23: """
30 df = spark.sql("""select * from table""")
31 """
32 should be changed to:
33 """
34 df = spark.sql("""select * from table""").toPandas()
35 """
36
37 Accept Solution
38 =====
39 Suggestion 3
40 how to create a Azure databricks dataframe in python from a sql table
41 I have a sql table in Azure databricks and I want to create a dataframe in python
42 I tried the following:
43 """
44 df = spark.sql("SELECT * FROM mytable")
45 """
```

Abbildung 2: Von KI generiertes Python-Skript

Dies zeigt die Leistungsfähigkeit von Azure Databricks als führende Plattform für Daten und KI in Kombination mit Entwickler-Tools wie Visual Studio Code und GitHub Copilot.

Integration von Azure Databricks mit OneLake

OneLake dient innerhalb von Microsoft Fabric als einheitlicher Speicherort für die Speicherung der Daten einer Organisation. Die Stärke der Lakehouse-Architektur liegt darin, dass mehrere Analytics-Engines die in OneLake gespeicherten Daten verarbeiten können. Dies können Engines innerhalb von Fabric sein, beispielsweise Synapse Data Warehouse. Allerdings können auch andere Azure-Dienste, wie zum Beispiel Azure Databricks, in OneLake gespeicherte Daten verarbeiten. Dies ermöglicht es Unternehmen, die volle Leistungsfähigkeit von Azure innerhalb des Fabric-Datenbestands zu nutzen.

Das folgende Tutorial zeigt, wie Sie über Azure Databricks eine Verbindung zu OneLake herstellen. Die Verbindung mit OneLake ermöglicht das Lesen und Schreiben von Daten in Microsoft Fabric Lakehouse aus einem Azure Databricks-Arbeitsbereich.

Um diesem Beispiel zu folgen, benötigen Sie:

- Ein etablierter Fabric-Arbeitsbereich und Fabric Lakehouse
- Azure Databricks-Arbeitsbereich (Premium)

Hinweis:

Nur Premium-Azure Databricks-Arbeitsbereiche unterstützen das Passthrough von Microsoft Azure Active Directory-Anmeldeinformationen, was für diese Integration unerlässlich ist.

Schritt-für-Schritt-Anleitung

Befolgen Sie diese Schritte, um Azure Databricks mit OneLake zu integrieren:

1. Zunächst müssen Sie Ihre Azure Databricks-Cluster so konfigurieren, dass sie das `credential passthrough` verwenden. Dadurch kann Azure Databricks Ihre Benutzeridentität verwenden, wenn eine Verbindung zu einem Azure Storage-Konto hergestellt wird, in diesem Fall Ihrem OneLake-Konto.
2. Öffnen Sie Ihren Azure Databricks-Arbeitsbereich. Wählen Sie im linken Bereich die Option `Compute`. Klicken Sie dann auf die Schaltfläche `Create compute`.
3. Befolgen Sie Ihr übliches Verfahren zum Erstellen eines neuen Azure Databricks-Clusters. Stellen Sie jedoch sicher, dass Sie den Abschnitt `Advanced options` erweitern und die Option `Enable credential passthrough for user level data access` aktivieren. Dadurch können Sie Ihre Microsoft-Benutzungsidentität verwenden, um eine Verbindung zu OneLake herzustellen.
4. Wenn Sie alle gewünschten Konfigurationsoptionen festgelegt haben, klicken Sie auf `Create compute`.
5. Sobald Ihr neuer Cluster fertig ist, klicken Sie im Bedienfeld auf der linken Seite des Bildschirms auf `Workspace`. Klicken Sie auf die Schaltfläche `Add` in der oberen rechten Ecke des Bildschirms und wählen Sie `Notebook` aus dem Dropdownmenü aus. Dadurch wird ein neues Azure Databricks-Notebook erstellt.
6. Wenn das Notizbuch geöffnet wird, verwenden Sie das Dropdown-Menü oben im Notizbuch, um sicherzustellen, dass Sie mit dem neuen Cluster verbunden sind, den Sie gerade erstellt haben.
7. Um mit dem Tutorial fortzufahren, müssen Sie die Details Ihres OneLake-Speicherkontos abrufen. Um diese Informationen zu finden, gehen Sie zu Ihrem Fabric Lakehouse und suchen Sie den Azure Blob Filesystem (ABFS)-Pfad. Diese finden Sie im Bereich `Eigenschaften` der Arbeitsbereichsseite.

Hinweis:

Azure Databricks unterstützt ausschließlich den ABFS-Treiber für Interaktionen mit ADLS Gen2 und OneLake. Ein Pfad könnte beispielsweise so aussehen:
`abfss://myWorkspace@onelake.dfs.fabric.microsoft.com/`

8. Sie können jetzt Azure Databricks verwenden, um Daten in Ihr OneLake-Speicherkonto zu laden. Fügen Sie den folgenden Code in eine Zelle Ihres Notebooks ein. Dieser Code erstellt einen neuen Datensatz in Microsoft Fabric basierend auf Beispieldaten in Azure Databricks:

```
oneLakePath = `abfss://myWorkspace@onelake.dfs.fabric.microsoft.com/myLakehouse.lakehouse/Files/`

# Load data from a Databricks public dataset into a dataframe.
#Alternatively, you can source data from elsewhere in Fabric or
#from another ADLS Gen2 account you own. For example:

yellowTaxiDF = spark.read.format("csv").option("header", "true").
option("inferSchema", "true").load("/databricks-datasets/nyctaxi/
tripdata/yellow/yellow_tripdata_2019-12.csv.gz")

#Process your data as needed. This could involve filtering,
#joining with other datasets, or other transformations. For
#instance:

filteredTaxiDF = yellowTaxiDF.where(yellowTaxiDF.fare_amount<4).
where(yellowTaxiDF.passenger_count==4)
display(filteredTaxiDF)

#Write the processed dataframe to your Fabric Lakehouse using the
#OneLake path you saved earlier:

filteredTaxiDF.write.format("csv").option("header", "true").
mode("overwrite").csv(oneLakePath)

#To ensure your data was written successfully, read the newly
#loaded file and display a sample:

lakehouseRead = spark.read.format('csv').option("header", "true").
load(oneLakePath) display(lakehouseRead.limit(10))
```

Mit diesem Tutorial haben Sie Azure Databricks erfolgreich mit OneLake in Fabric integriert. Mit Azure Databricks können Sie jetzt nahtlos Daten in Fabric lesen und schreiben.

Fazit

Effektives Datenmanagement und -Analytics sind für Unternehmen von entscheidender Bedeutung, um sich einen Wettbewerbsvorteil zu verschaffen. Microsoft Fabric ist eine zentrale Lösung und bietet eine einheitliche, skalierbare und sichere Plattform, die speziell auf die Anforderungen moderner Unternehmen zugeschnitten ist. Diese Plattform gewährleistet nicht nur eine nahtlose Datenintegration und -verwaltung, sondern fördert auch die Zusammenarbeit, erhöht die Sicherheit und treibt Innovationen voran.



Nächste Schritte

- ⦿ Erfahren Sie mehr über Microsoft Fabric und starten Sie mit einer kostenlosen [Testversion](#).
- ⦿ Ermöglichen Sie Daten, Analytics und KI-Anwendungsfälle mit [Azure Databricks](#).
- ⦿ Erfahren Sie, wie Sie [mit OneLake in Fabric ein Lakehouse erstellen](#).