Ingénieur en analytique Microsoft Fabric

Parcours 1: Bien démarrer avec Microsoft Fabric

Module 1: Présentation de l’analytique de bout en bout à l’aide de Microsoft Fabric

Unité 1: Présentation

Type: Introduction

Microsoft Fabric est une plateforme d’analytique de bout en bout qui fournit un environnement unique et intégré permettant aux professionnels des données et à l’entreprise de collaborer sur des projets de données. Fabric fournit un ensemble de services intégrés qui vous permettent d’ingérer, de stocker, de traiter et d’analyser des données dans un environnement unique.

Microsoft Fabric fournit des outils pour les praticiens des données citoyens et professionnels, et s’intègre aux outils dont l’entreprise a besoin pour prendre des décisions. Fabric inclut les services suivants :

Ingénierie des données

Intégration de données

Entrepôt de données

Informations en temps réel

Science des données

Informatique décisionnelle

Ce module présente la plateforme Fabric, décrit pour qui Fabric est destiné et explore les services Fabric.

Unité 2: Explorer l’analytique de bout en bout avec Microsoft Fabric

Type: Contenu

Explorer l’analytique de bout en bout avec Microsoft Fabric

L’analytique scalable peut être complexe, fragmentée et coûteuse. Microsoft Fabric simplifie les solutions d’analytique en fournissant un produit unique et facile à utiliser qui intègre différents outils et

services dans une plateforme.

Fabric est une plateforme SaaS (software-as-a-service) unifiée où toutes les données sont stockées dans un format ouvert unique dans OneLake. OneLake est accessible par tous les moteurs d'analyse de la plateforme, garantissant l'extensibilité, l'efficacité et l'accessibilité depuis n'importe où avec une connexion Internet.

OneLake est l'architecture de stockage de données centralisée de Fabric qui permet la collaboration en éliminant la nécessité de déplacer ou de copier des données entre les systèmes. OneLake unifie vos données entre les régions et les clouds dans un lac logique unique sans déplacer ni dupliquer des données.

OneLake est basé sur Azure Data Lake Storage (ADLS) et prend en charge différents formats, notamment Delta, Parquet, CSV et JSON. Tous les moteurs de calcul dans Fabric stockent automatiquement leurs données dans OneLake, ce qui le rend directement accessible sans avoir besoin de déplacement ou de duplication. Pour les données tabulaires, les moteurs analytiques dans Fabric écrivent les données au format delta-parquet et tous les moteurs interagissent avec le format de manière fluide.

Les raccourcis sont des références aux fichiers ou aux emplacements de stockage externes à OneLake, ce qui vous permet d'accéder aux données cloud existantes sans les copier. Les raccourcis garantissent la cohérence des données et permettent à Fabric de rester synchronisés avec la source.

Espaces de travail

Dans Microsoft Fabric, les espaces de travail servent de conteneurs logiques qui vous aident à organiser et gérer vos données, rapports et autres ressources. Ils permettent de séparer clairement les ressources, ce qui facilite le contrôle de l'accès et la maintenance de la sécurité.

Chaque espace de travail dispose de son propre ensemble d'autorisations, ce qui garantit que seuls les utilisateurs autorisés peuvent afficher ou modifier son contenu. Cette structure prend en charge la collaboration d'équipe tout en conservant un contrôle d'accès strict pour les utilisateurs professionnels et informatiques.

Les espaces de travail vous permettent de gérer les ressources de calcul et d'intégrer Git pour le contrôle de version. Vous pouvez optimiser les performances et les coûts en configurant les paramètres de calcul, tandis que l'intégration git permet de suivre les modifications, de collaborer sur le code et de gérer un historique de votre travail.

Administration et gouvernance

OneLake de Fabric est régi de manière centralisée et ouvert à la collaboration. Les données sont sécurisées et régies au même endroit, ce qui permet aux utilisateurs de trouver et d'accéder facilement aux données dont ils ont besoin. L'administration du réseau est centralisée dans le portail d'administration.

Dans le portail d'administration, vous pouvez gérer des groupes et des autorisations, configurer des sources de données et des passerelles, et surveiller l'utilisation et les performances. Vous pouvez également accéder aux API et kits sdk d'administration fabric dans le portail d'administration, ce qui peut automatiser les tâches courantes et intégrer Fabric à d'autres systèmes.

Le catalogue OneLake vous aide à analyser, surveiller et gérer la gouvernance des données. Il fournit des conseils sur les étiquettes de confidentialité, les métadonnées d'élément et l'état d'actualisation des données, offrant des insights sur l'état de gouvernance et les actions à améliorer.

Pour plus d'informations, consultez la documentation d'administration de Microsoft Fabric .

Unité 3: Explorer les équipes de données et Microsoft Fabric

Type: Contenu

Explorer les équipes de données et Microsoft Fabric

La plateforme d'analytique des données unifiée de Microsoft Fabric facilite la collaboration des professionnels des données sur des projets. Fabric augmente la collaboration entre les professionnels des données en supprimant les silos de données et en éliminant le besoin de plusieurs systèmes.

Rôles et défis traditionnels

Dans un processus de développement d'analytique traditionnel, les équipes de données sont souvent confrontées à plusieurs défis en raison de la division des tâches et workflows de données.

Les ingénieurs données traitent et organisent des données pour les analystes, qui l'utilisent ensuite pour créer des rapports métier. Ce processus nécessite une coordination approfondie, ce qui entraîne souvent des retards et des interprétations erronées.

Les analystes de données doivent souvent effectuer des transformations de données en aval avant de créer des rapports Power BI. Ce processus prend beaucoup de temps et peut manquer de contexte nécessaire, ce qui rend plus difficile pour les analystes de se connecter directement aux données.

Les scientifiques des données rencontrent des difficultés à intégrer des techniques de science des données natives à des systèmes existants, qui sont souvent complexes et rend difficile de fournir efficacement des insights pilotés par les données.

Évolution des flux de travail collaboratifs

Microsoft Fabric simplifie le processus de développement d'analytique en unifiant les outils dans une plateforme SaaS. Fabric permet à différents rôles de collaborer efficacement sans dupliquer les efforts.

Les ingénieurs données peuvent ingérer, transformer et charger des données directement dans OneLake à l'aide de Pipelines, qui automatisent les flux de travail et prennent en charge la planification. Ils peuvent stocker des données dans des lakehouses en utilisant le format Delta-Parquet pour bénéficier d'un stockage et d'un contrôle de version efficaces. Les notebooks fournissent des fonctionnalités de script avancées pour les transformations complexes.

Les analystes de données peuvent transformer des données en amont à l'aide de dataflows et se connecter directement à OneLake en mode Direct Lake, ce qui réduit la nécessité de transformations en aval. Ils peuvent créer des rapports interactifs plus efficacement à l'aide de Power BI.

Les scientifiques des données peuvent utiliser des notebooks intégrés avec prise en charge de Python et Spark pour créer et tester des modèles Machine Learning. Ils peuvent stocker et accéder aux données dans lakehouses et s'intégrer à Azure Machine Learning pour opérationnaliser et déployer des modèles.

Les ingénieurs en analyse servent de pont entre l'ingénierie des données et l'analyse en organisant des ressources de données dans les entrepôts de données, en garantissant la qualité des données et en facilitant l'analytique en libre-service. Ils peuvent créer des modèles sémantiques dans Power BI pour organiser et présenter efficacement des données.

Les utilisateurs de low-code/no-code et les développeurs citoyens peuvent découvrir des jeux de données organisés via OneLake Hub et utiliser des modèles Power BI pour créer rapidement des rapports et des tableaux de bord. Ils peuvent également utiliser des dataflows pour effectuer des tâches ETL simples sans compter sur les ingénieurs données.

Unité 4: Activer et utiliser Microsoft Fabric

Type: Contenu

Activer et utiliser Microsoft Fabric

Avant de pouvoir explorer les fonctionnalités de bout en bout de Microsoft Fabric, vous devez l'activer pour votre organisation. Vous devrez peut-être travailler avec votre service informatique pour activer Fabric pour votre organisation, y compris l'un des rôles suivants :

Administrateur d'infrastructure (anciennement administrateur Power BI) : gère les paramètres et les configurations fabric.

Administrateur Power Platform : supervise les services Power Platform, y compris Fabric.

Administrateur Microsoft 365 : gère les services Microsoft à l'échelle de l'organisation, notamment Fabric.

Activer Microsoft Fabric

Les administrateurs peuvent activer Fabric dans les paramètres du locataire du portail d'administration > dans le service Power BI. Fabric peut être activé pour l'ensemble de l'organisation ou pour des groupes de sécurité Microsoft 365 ou Microsoft Entra spécifiques. Les administrateurs peuvent également déléguer cette possibilité à d'autres utilisateurs au niveau de la capacité.

Si votre organisation n'utilise pas Fabric ou Power BI aujourd'hui, vous pouvez vous inscrire à une version d'évaluation gratuite de Fabric pour explorer ses fonctionnalités.

Créer des espaces de travail

Les espaces de travail sont des environnements collaboratifs dans lesquels vous pouvez créer et gérer des éléments tels que des lakehouses, des entrepôts et des rapports. Toutes les données sont stockées dans OneLake et accessibles via des espaces de travail. Les espaces de travail prennent également en charge la vue de traçabilité des données, fournissant une vue visuelle du flux de données et des dépendances pour améliorer la transparence et la prise de décision.

Dans les paramètres de l'espace de travail, vous pouvez configurer :

Type de licence permettant d'utiliser des fonctionnalités Fabric.

Accès OneDrive pour l'espace de travail.

Connexion de stockage Azure Data Lake Gen2.

Intégration Git pour le contrôle de version.

Paramètres de charge de travail Spark pour l'optimisation des performances.

Vous pouvez gérer l'accès à l'espace de travail via quatre rôles : administrateur, contributeur, membre et visionneuse. Ces rôles s'appliquent à tous les éléments d'un espace de travail et doivent être réservés à la collaboration. Pour un contrôle d'accès plus précis, utilisez des autorisations au niveau des éléments en fonction des besoins de l'entreprise.

En savoir plus sur les espaces de travail dans la documentation Fabric.

Découvrir des données avec le catalogue OneLake

Le catalogue OneLake dans Microsoft Fabric permet aux utilisateurs de trouver et d'accéder facilement à diverses sources de données au sein de leur organisation. Les utilisateurs explorent et se connectent aux sources de données, s'assurant ainsi qu'ils disposent des données appropriées pour leurs besoins.

Les utilisateurs voient uniquement les éléments partagés avec eux. Voici quelques points à prendre en compte lors de l'utilisation du hub OneLake :

Affiner les résultats par espaces de travail ou domaines (s'ils sont implémentés).

Explorez les catégories par défaut pour localiser rapidement les données pertinentes.

Filtrer par mot clé ou type d'élément.

Créer des éléments avec des charges de travail Fabric

Une fois que vous avez créé votre espace de travail Fabric, vous pouvez commencer à créer des éléments dans Fabric. Chaque charge de travail dans Fabric offre différents types d'éléments pour le stockage, le traitement et l'analyse des données. Les charges de travail Fabric sont les suivantes :

Ingénierie des données : créez des lakehouses et opérationnalisez des flux de travail pour créer, transformer et partager votre patrimoine de données.

Data Factory : ingérer, transformer et orchestrer des données.

Science des données : détecter les tendances, identifier les valeurs hors norme et prédire des valeurs à l'aide du Machine Learning.

Entrepôt de données : combinez plusieurs sources dans un entrepôt traditionnel pour l'analytique.

Bases de données : créez et gérez des bases de données avec des outils pour insérer, interroger et extraire des données.

Solutions industrielles : utilisez des solutions de données industrielles prêtes à l'emploi.

Real-Time Intelligence : traiter, surveiller et analyser les données de diffusion en continu.

Power BI : Créez des rapports et des tableaux de bord pour prendre des décisions pilotées par les données.

Fabric intègre des fonctionnalités d'outils Microsoft existants tels que Power BI, Azure Synapse Analytics et Azure Data Factory dans une plateforme unifiée. Fabric prend également en charge une architecture de maillage de données, ce qui permet la propriété décentralisée des données tout en conservant la gouvernance centralisée. Cela élimine la nécessité d'un accès direct aux ressources Azure, ce qui simplifie les flux de travail de données.

Unité 5: Évaluation du module

Type: Évaluation

Évaluation du module

Quel est l'avantage clé de l'utilisation de Microsoft Fabric dans les projets de données ?

Il permet aux professionnels des données de travailler indépendamment, sans collaboration.

Il nécessite la duplication des données entre les systèmes pour garantir la disponibilité.

Il fournit un environnement unique et intégré pour la collaboration sur les projets de données.

Quel est le format de stockage par défaut pour OneLake de Fabric ?

Quelle expérience Fabric est utilisée pour déplacer et transformer des données ?

Science des données

Entrepôt de données

Usine de Données

Vous devez répondre à toutes les questions avant de vérifier votre travail.

Unité 6: Résumé

Type: Résumé

Les professionnels des données sont de plus en plus censés être en mesure de travailler avec des données à grande échelle, et ceci de manière sécurisée, conforme et économique. En même temps, l'entreprise souhaite utiliser ces données plus efficacement et plus rapidement pour prendre de meilleures décisions.

Microsoft Fabric est une collection d'outils et de services qui répondent précisément à ces exigences de l'organisation. Dans ce module, vous avez découvert le stockage OneLake de Fabric, les charges de travail incluses dans Fabric, et comment activer et utiliser Fabric dans votre organisation.

Module 2: Bien démarrer avec les lakehouses dans Microsoft Fabric

Unité 1: Présentation

Type: Introduction

La base de Microsoft Fabric est un lakehouse, qui repose sur la couche de stockage scalable OneLake et utilise des moteurs de calcul Apache Spark et SQL pour le traitement du Big Data. Un lakehouse est une plateforme unifiée qui combine :

Le stockage flexible et scalable d'un lac de données.

La capacité à interroger et à analyser les données d'un entrepôt de données.

Imaginez que votre entreprise utilise un entrepôt de données pour stocker des données structurées à partir de ses systèmes transactionnels, telles que l'historique des commandes, les niveaux d'inventaire et les informations client. Vous collectez des données non structurées auprès de réseaux sociaux, de journaux de sites web et de sources tierces qui sont difficiles à gérer et à analyser en utilisant l'infrastructure d'entrepôt de données existante. La nouvelle directive de votre entreprise consiste à améliorer ses capacités de prise de décision en analysant les données dans différents formats et plusieurs sources ; l'entreprise choisit pour cela Microsoft Fabric.

Dans ce module, nous expliquons comment un lakehouse dans Microsoft Fabric fournit un magasin de données évolutif et flexible pour des fichiers et des tables que vous pouvez interroger en utilisant SQL.

Unité 2: Explorer le lakehouse Microsoft Fabric

Type: Contenu

Explorer le lakehouse Microsoft Fabric

Un lakehouse se présente sous la forme d'une base de données et s'appuie sur un lac de données à l'aide de tables au format Delta. Les lakehouses combinent les fonctionnalités analytiques basées sur SQL d'un entrepôt de données relationnelles à la flexibilité et la scalabilité d'un lac de données. Les lakehouses stockent tous les formats de données et peuvent être utilisés avec différents outils d'analytique et langages de programmation. En tant que solutions basées sur le cloud, les lakehouses peut être mis à l'échelle automatiquement et fournir une haute disponibilité et une reprise d'activité après sinistre.

Voici quelques-uns des avantages des lakehouses :

Les lakehouses utilisent des moteurs Spark et SQL pour traiter des données à grande échelle, et prennent en charge le Machine Learning ou l'analytique de modélisation prédictive.

Les données Lakehouse sont organisées dans un format de schéma en lecture, ce qui signifie que vous définissez le schéma en fonction des besoins plutôt que d'avoir un schéma prédéfini.

Les lakehouses prennent en charge les transactions ACID (Atomicité, Cohérence, Isolation, Durabilité) par le biais de tables au format Delta Lake à des fins de cohérence et d'intégrité des données.

Les lakehouses constituent un emplacement unique où les ingénieurs données, les scientifiques des données et les analystes de données peuvent accéder aux données et les utiliser.

Si vous souhaitez une solution analytique évolutive qui préserve la cohérence des données, vous pouvez opter pour un lakehouse. Il est important d'évaluer vos besoins spécifiques afin de déterminer

la solution la mieux adaptée.

Charger des données dans un lakehouse

Les lakehouses Fabric sont un élément central de votre solution d'analyse. Vous pouvez suivre le processus ETL (extraction, transformation et chargement) pour ingérer et transformer des données avant de les charger dans le lakehouse.

Vous pouvez ingérer des données dans de nombreux formats courants à partir de diverses sources, notamment des fichiers locaux, des bases de données ou des API. Vous pouvez également créer des raccourcis Fabric vers des données dans des sources externes, telles qu'Azure Data Lake Store Gen2 ou OneLake. Utilisez l'Explorateur de lakehouses pour parcourir les fichiers, les dossiers, les raccourcis et les tables et afficher leur contenu au sein de la plateforme Fabric.

Vous pouvez transformer les données ingérées, puis les charger en utilisant Apache Spark avec des notebooks ou des flux de données Gen2. Utilisez des pipelines Data Factory pour orchestrer vos différentes activités ETL et placer les données préparées dans votre lakehouse.

Les flux de données Gen2 sont basés sur Power Query, un outil familier des analystes de données utilisant Excel ou Power BI qui fournit une représentation visuelle des transformations comme alternative à la programmation traditionnelle.

Vous pouvez utiliser votre lakehouse pour de nombreuses raisons, notamment :

Analyser des données avec SQL.

Entraîner des modèles Machine Learning

Effectuer des analyses sur des données en temps réel.

Développer des rapports dans Power BI.

Sélectionner un lakehouse

L'accès au lakehouse est géré via l'espace de travail ou le partage au niveau des éléments. Vous devez utiliser des rôles d'espace de travail pour les collaborateurs, car ces rôles permettent d'accéder à tous les éléments de l'espace de travail. Le partage au niveau des éléments est idéal pour accorder un accès en lecture seule, par exemple à des fins d'analyse ou de développement de rapports Power BI.

Les lakehouses Fabric prennent également en charge les fonctionnalités de gouvernance des données, notamment les étiquettes de confidentialité, et peuvent être étendus à l'aide de Microsoft Purview avec votre locataire Fabric.

Pour plus d'informations, consultez la documentation sécurité dans Microsoft Fabric .

Unité 3: Travailler avec des lacs de données Microsoft Fabric

Type: Contenu

Travailler avec des lacs de données Microsoft Fabric

Maintenant que vous comprenez les principales fonctionnalités des lakehouses Microsoft Fabric, nous allons apprendre à les utiliser.

Créer et explorer un lakehouse

Lorsque vous créez un lakehouse, vous avez trois éléments de données différents créés automatiquement dans votre espace de travail.

Le lakehouse contient des raccourcis, des dossiers, des fichiers et des tables.

Le modèle sémantique (par défaut) fournit une source de données simple pour les développeurs de rapports Power BI.

Le point de terminaison d'analytique SQL autorise l'accès en lecture seule aux données d'interrogation avec SQL.

Vous pouvez utiliser les données du lakehouse dans deux modes :

lakehouse vous permet d' et d'interagir avec des tables, des fichiers et des dossiers dans la lakehouse.

Le point de terminaison d'analytique SQL vous permet d'utiliser SQL pour interroger les tables dans lakehouse et gérer son modèle sémantique relationnel.

Ingérer des données dans un lakehouse

L'ingestion de données dans votre lakehouse est la première étape de votre processus ETL. Utilisez l'une des méthodes suivantes pour importer des données dans votre lakehouse.

Charger : charger des fichiers locaux.

Dataflows Gen2 : Importer et transformer des données à l'aide de Power Query.

Notebooks : Utilisez Apache Spark pour ingérer, transformer et charger des données.

Pipelines Data Factory : utilisez l'activité Copier les données.

Ces données peuvent ensuite être chargées directement dans des fichiers ou des tables. Prenez en compte votre modèle de chargement des données lors de l'ingestion de données pour déterminer si vous devez charger toutes les données brutes sous forme de fichiers avant de traiter ou d'utiliser des tables intermédiaires.

Les définitions de travaux Spark peuvent également être utilisées pour envoyer des travaux de traitement par lots/streaming à des clusters Spark. En téléchargeant les fichiers binaires à partir de la sortie de compilation de différents langages (par exemple, .jar de Java), vous pouvez appliquer différentes logiques de transformation aux données hébergées sur un Lakehouse. Outre le fichier binaire, vous pouvez personnaliser davantage le comportement du travail en téléchargeant davantage de bibliothèques et d'arguments de ligne de commande.

Pour plus d'informations, consultez la documentation Créer une définition de travail Apache Spark .

Accéder aux données à l'aide de raccourcis

Une autre façon d'accéder aux données et d'utiliser des données dans Fabric consiste à utiliser des raccourcis. Les raccourcis vous permettent d'intégrer des données dans votre lakehouse tout en les conservant dans un stockage externe.

Les raccourcis sont utiles lorsque vous devez sourcer des données qui sont dans un autre compte de stockage ou même dans un autre fournisseur de cloud. Dans votre lakehouse, vous pouvez créer des raccourcis qui pointent vers différents comptes de stockage et d'autres éléments Fabric tels que des entrepôts de données, des bases de données KQL et autres lakehouses.

Les autorisations et les informations d'identification des données sources sont toutes gérées par OneLake. Lors de l'accès aux données via un raccourci vers un autre emplacement OneLake, l'identité de l'utilisateur appelant est utilisée pour autoriser l'accès aux données dans le chemin cible du raccourci. L'utilisateur doit disposer d'autorisations à l'emplacement cible pour lire les données.

Des raccourcis peuvent être créés dans des lakehouses et dans des bases de données KQL, et apparaître sous la forme d'un dossier dans le lac. Cela permet à Spark, SQL, intelligence en temps réel et Analysis Services d'utiliser tous les raccourcis lors de l'interrogation des données.

Pour plus d'informations sur l'utilisation des raccourcis, consultez la documentation sur les raccourcis OneLake dans la documentation Microsoft Fabric.

Unité 4: Explorer et transformer des données dans un lakehouse

Type: Contenu

Explorer et transformer des données dans un lakehouse

Transformez et chargez des données

La plupart des données nécessitent des transformations avant d'être chargées dans des tables. Vous pouvez ingérer des données brutes directement dans un lakehouse, puis les transformer et les charger dans des tables. Quelle que soit votre conception ETL, vous pouvez transformer et charger des données simplement en utilisant les mêmes outils pour ingérer des données. Les données transformées peuvent ensuite être chargées sous forme de fichier ou de table Delta.

Les notebooks sont privilégiés par les ingénieurs de données familiarisés avec différents langages de programmation, notamment PySpark, SQL et Scala.

Les Dataflows Gen2 sont excellents pour les développeurs familiarisés avec Power BI ou Excel car ils utilisent l'interface PowerQuery.

Les pipelines fournissent une interface visuelle pour exécuter et orchestrer les processus ETL. Les pipelines peuvent être aussi simples ou complexes que vous le souhaitez.

Analyser et visualiser des données dans un lakehouse

Une fois les données ingérées, transformées et chargées, elles sont prêtes à être utilisées par d'autres. Les articles en tissu offrent la flexibilité nécessaire à chaque organisation afin que vous puissiez utiliser les outils qui vous conviennent.

Les scientifiques des données peuvent utiliser des notebooks ou des Data Wrangler pour explorer et former des modèles d'apprentissage automatique pour l'IA.

Les développeurs de rapports peuvent utiliser le modèle sémantique pour créer des rapports Power BI.

Les analystes peuvent utiliser le point de terminaison d'analyse SQL pour interroger, filtrer, agréger et explorer les données dans les tables Lakehouse.

En combinant les fonctionnalités de visualisation des données de Power BI avec le stockage centralisé et le schéma tabulaire d'un data lakehouse, vous pouvez implémenter une solution d'analytique de bout en bout sur une plateforme unique.

Unité 5: Exercice – Créer un lakehouse Microsoft Fabric

Type: Exercice

Exercice – Créer un lakehouse Microsoft Fabric

Dans cet exercice, explorez les tâches de lakehouse Microsoft Fabric telles que la création d'un lakehouse, l'importation de données, l'interrogation de tables avec SQL et la génération de rapports. L'exercice met l'accent sur l'importance du lakehouse en tant que composant central dans l'engineering données, l'entreposage et l'analytique, permettant aux utilisateurs de gérer et d'analyser efficacement leurs données dans l'environnement de lakehouse.

Vous avez besoin d'une licence d'évaluation Microsoft Fabric avec la préversion Fabric activée dans votre locataire. Consultez Bien démarrer avec Fabric pour activer votre licence d'évaluation Fabric.

Lancez l'exercice et suivez les instructions.

Unité 6: Évaluation du module

Type: Évaluation

Évaluation du module

Qu'est-ce qu'un lakehouse Microsoft Fabric ?

Une base de données relationnelle basée sur le moteur de base de données Microsoft SQL Server.

Une hiérarchie de dossiers et de fichiers dans Azure Data Lake Store Gen2.

Un magasin analytique qui combine la flexibilité de stockage de fichiers d'un lac de données avec les fonctionnalités de requête SQL d'un entrepôt de données.

Vous souhaitez inclure des données dans un emplacement Azure Data Lake Store Gen2 externe dans votre lakehouse, sans être obligé de copier les données. Que faire ?

Créer un pipeline de données qui utilise une activité Copier des données pour charger les données externes dans un fichier.

Créer un raccourci.

Créer un flux de données Gen2 qui extrait les données et les charge dans une table.

Vous souhaitez utiliser Apache Spark pour explorer de manière interactive les données d'un fichier dans le lakehouse. Que faire ?

Créez un bloc-notes.

Basculer vers le mode Point de terminaison d'analyse SQL.

Créer un Flux de données Gen2.

Vous devez répondre à toutes les questions avant de vérifier votre travail.

Unité 7: Résumé

Type: Résumé

Dans ce module, nous avons appris comment les lakehouses s'intègrent dans une solution d'analytique des données à l'aide de Microsoft Fabric. Les lakehouses fournissent aux ingénieurs et aux analystes de données les avantages combinés du stockage de lac de données et d'un entrepôt de données relationnelles. Vous pouvez utiliser un lakehouse comme base d'une solution d'analytique

données de bout en bout qui comprend l'ingestion, la transformation, la modélisation et la visualisation des données.

Les lakehouses Fabric offre une valeur en tant que magasin de données Logiciel en tant que service qui offre tous les avantages avec moins d'administration.

Pour plus d'informations, consultez la documentation sur l'Ingénierie des données dans Microsoft Fabric.

Module 3: Utiliser Apache Spark dans Microsoft Fabric

Unité 1: Présentation

Type: Introduction

Apache Spark est un framework de traitement parallèle open source pour le traitement et l'analytique à grande échelle des données. Spark est devenu populaire dans les scénarios de traitement de « Big Data » et est disponible dans plusieurs implémentations de plateforme, notamment Azure HDInsight, Azure Synapse Analytics et Microsoft Fabric.

Ce module explore comment utiliser Spark dans Microsoft Fabric pour ingérer, traiter et analyser les données d'un lakehouse. Bien que les techniques de base et le code décrits dans ce module soient communs à toutes les implémentations Spark, les outils intégrés et la capacité à travailler avec Spark dans le même environnement que d'autres services de données dans Microsoft Fabric facilitent l'intégration du traitement des données basé sur Spark dans votre solution d'analytique données globale.

Unité 2: Se préparer à l'utilisation d'Apache Spark

Type: Contenu

Se préparer à l'utilisation d'Apache Spark

Apache Spark est un framework de traitement de données distribué qui prend en charge l'analytique données à grande échelle en coordonnant le travail sur plusieurs nœuds de traitement dans un cluster, ce que nous appelons un pool Spark dans Microsoft Fabric. En termes plus simples, Spark utilise une approche « diviser et conquérir » pour traiter rapidement de grands volumes de données en répartissant le travail sur plusieurs ordinateurs. Le processus de distribution des tâches et de regroupement des résultats est géré pour vous par Spark.

Spark peut exécuter du code écrit dans un large éventail de langages, notamment Java, Scala (un langage de script basé sur Java), Spark R, Spark SQL et PySpark (une variante de Python propre à Spark). En pratique, la plupart des charges de travail d'engineering et d'analytique données sont es à l'aide d'une combinaison de PySpark et de Spark SQL.

Un pool Spark se compose de nœuds de calcul qui distribuent les tâches de traitement de données. L'architecture générale est illustrée dans le diagramme suivant.

Comme illustré dans le diagramme, un pool Spark contient deux types de nœuds :

Un nœud principal dans un pool Spark coordonne les processus distribués par le biais d'un programme pilote.

Le pool comprend plusieurs nœuds Worker sur lesquels les processus Exécuteur exécutent les tâches de traitement de données réelles.

Le pool Spark utilise cette architecture de calcul distribuée pour accéder aux données et les traiter dans un magasin de données compatible, tel qu'un data lakehouse basé dans OneLake.

Pools Spark dans Microsoft Fabric

Microsoft Fabric fournit un pool de démarrage dans chaque espace de travail, ce qui permet de démarrer et d'exécuter rapidement des travaux Spark avec une configuration minimale. Vous pouvez

configurer le pool de démarrage pour optimiser les nœuds qu'il contient conformément aux besoins spécifiques de votre charge de travail ou à vos contraintes de coût.

En outre, vous pouvez créer des pools Spark personnalisés avec des configurations de nœud spécifiques qui prennent en charge vos besoins particuliers en matière de traitement de données.

La personnalisation des paramètres d'un pool Spark peut être désactivée par les administrateurs Fabric au niveau de la capacité Fabric. Pour plus d'informations, consultez Paramètres d'administration de la capacité pour l'engineering données et la science des données dans la documentation Fabric.

Vous pouvez gérer les paramètres du pool de démarrage et créer des pools Spark dans la section Portail d'administration des paramètres de l'espace de travail, sous Paramètres de capacité, puis Paramètres d'ingénierie/science des données.

Les paramètres de configuration spécifiques pour les pools Spark incluent :

Famille de nœuds : type de machines virtuelles utilisées pour les nœuds de cluster Spark. Dans la plupart des cas, les nœuds à mémoire optimisée fournissent des performances optimales.

Mise à l'échelle automatique: indique s'il faut ou non approvisionner automatiquement les nœuds selon les besoins et, le cas échéant, le nombre initial et le nombre maximal de nœuds à allouer au pool.

Allocation dynamique : indique s'il faut allouer dynamiquement les processus Exécuteur sur les nœuds Worker en fonction des volumes de données.

Si vous créez un ou plusieurs pools Spark personnalisés dans un espace de travail, vous pouvez définir l'un d'entre eux (ou le pool de démarrage) comme pool par défaut. Ce pool est utilisé si un pool spécifique n'est pas spécifié pour un travail Spark donné.

Pour plus d'informations sur la gestion des pools Spark dans Microsoft Fabric, consultez Configuration des pools de démarrage dans Microsoft Fabric et Comment créer des pools Spark personnalisés dans Microsoft Fabric dans la documentation Microsoft Fabric.

Runtimes et environnements

L'écosystème open source Spark comprend plusieurs versions du runtime Spark, qui détermine la version d'Apache Spark, de Delta Lake, de Python et d'autres composants logiciels de base installés. En outre, dans un runtime, vous pouvez installer et utiliser une large sélection de bibliothèques de code pour des tâches courantes (et parfois très spécialisées). Étant donné qu'une grande partie du traitement Spark est e à l'aide de PySpark, la vaste gamme de bibliothèques Python garantit que, quelle que soit la tâche que vous devez effectuer, il existe probablement une bibliothèque pour vous aider.

Dans certains cas, les organisations peuvent avoir besoin de définir plusieurs environnements pour prendre en charge diverses tâches de traitement de données. Chaque environnement définit une version de runtime spécifique ainsi que les bibliothèques à installer pour effectuer des opérations spécifiques. Les ingénieurs et les scientifiques des données peuvent ensuite sélectionner l'environnement qu'ils souhaitent utiliser avec un pool Spark pour une tâche particulière.

Runtimes Spark dans Microsoft Fabric

Microsoft Fabric prend en charge plusieurs runtimes Spark et ajoute en continu de nouveaux runtimes à mesure qu'ils sont publiés. Vous pouvez utiliser l'interface des paramètres de l'espace de travail pour spécifier le runtime Spark utilisé par l'environnement par défaut lors du démarrage d'un pool Spark.

Pour plus d'informations sur les runtimes Spark dans Microsoft Fabric, consultez Runtimes Apache Spark dans Fabric dans la documentation Microsoft Fabric.

Environnements dans Microsoft Fabric

Vous pouvez créer des environnements personnalisés dans un espace de travail Fabric, ce qui vous permet d'utiliser des runtimes, des bibliothèques et des paramètres de configuration Spark spécifiques pour différentes opérations de traitement de données.

Lors de la création d'un environnement, vous pouvez :

Spécifier le runtime Spark qu'il doit utiliser.

Afficher les bibliothèques intégrées installées dans chaque environnement.

Installer des bibliothèques publiques spécifiques à partir de PyPI (Python Package Index).

Installer des bibliothèques personnalisées en chargeant un fichier de package.

Spécifier le pool Spark que l'environnement doit utiliser.

Spécifier les propriétés de configuration Spark pour remplacer le comportement par défaut.

Charger des fichiers de ressources qui doivent être disponibles dans l'environnement.

Après avoir créé au moins un environnement personnalisé, vous pouvez le spécifier comme environnement par défaut dans les paramètres de l'espace de travail.

Pour plus d'informations sur l'utilisation d'environnements personnalisés dans Microsoft Fabric, consultez Créer, configurer et utiliser un environnement dans Microsoft Fabric dans la documentation Microsoft Fabric.

Options de configuration Spark supplémentaires

La gestion des pools et des environnements Spark est la principale façon de gérer le traitement Spark dans un espace de travail Fabric. Toutefois, des options supplémentaires vous permettent d'effectuer d'autres optimisations.

Moteur d'exécution natif

Le moteur d'exécution natif dans Microsoft Fabric est un moteur de traitement vectorisé qui exécute des opérations Spark directement sur l'infrastructure lakehouse. L'utilisation du moteur d'exécution natif peut améliorer considérablement les performances des requêtes lors de l'utilisation de jeux de données volumineux dans les formats de fichiers Parquet ou Delta.

Pour utiliser le moteur d'exécution natif, vous pouvez l'activer au niveau de l'environnement ou dans un notebook individuel. Pour activer le moteur d'exécution natif au niveau de l'environnement, définissez les propriétés Spark suivantes dans la configuration de l'environnement :

spark.native.enabled : true

spark.shuffle.manager : org.apache.spark.shuffle.sort.ColumnarShuffleManager

Pour activer le moteur d'exécution natif pour un script ou un notebook spécifique, vous pouvez définir ces propriétés de configuration au début de votre code, comme ceci :

Pour plus d'informations sur le moteur d'exécution natif, consultez Moteur d'exécution natif pour Fabric Spark dans la documentation Microsoft Fabric.

Mode de concurrence élevée

Quand vous exécutez du code Spark dans Microsoft Fabric, une session Spark est lancée. Vous pouvez optimiser l'efficacité de l'utilisation des ressources Spark en utilisant le mode de concurrence élevée pour partager les sessions Spark entre plusieurs utilisateurs ou processus simultanés. Un carnet utilise une session Spark pour son exécution. Lorsque le mode d'accès concurrentiel élevé est activé, plusieurs utilisateurs peuvent, par exemple, exécuter du code dans des notebooks qui utilisent

la même session Spark, tout en garantissant l'isolation du code pour éviter les variables d'un bloc-notes affecté par le code d'un autre bloc-notes. Vous pouvez également activer le mode de concurrence élevée pour les tâches Spark, ce qui permet d'obtenir une efficacité similaire pour l'exécution simultanée de scripts Spark non interactifs.

Pour activer le mode de concurrence élevée, utilisez la section Engineering/Science des données de l'interface des paramètres de l'espace de travail.

Pour plus d'informations sur le mode de concurrence élevée, consultez Mode de concurrence élevée dans Apache Spark pour Fabric dans la documentation Microsoft Fabric.

Journalisation MLFlow automatique

MLFlow est une bibliothèque open source utilisée dans les charges de travail de science des données pour gérer l'entraînement et le déploiement de modèles Machine Learning. Une fonctionnalité clé de MLFlow est la possibilité de journaliser les opérations d'entraînement et de gestion des modèles. Par défaut, Microsoft Fabric utilise MLFlow pour journaliser implicitement l'activité d'expérimentation du Machine Learning sans que le scientifique des données ait à inclure de code explicite pour le faire. Vous pouvez désactiver cette fonctionnalité dans les paramètres de l'espace de travail.

Administration Spark pour une capacité Fabric

Les administrateurs peuvent gérer les paramètres Spark au niveau de la capacité Fabric, ce qui leur permet de restreindre et de remplacer les paramètres Spark dans les espaces de travail au sein d'une organisation.

Pour plus d'informations sur la gestion de la configuration Spark au niveau de la capacité Fabric, consultez Configurer et gérer les paramètres d'engineering données et de science des données pour des capacités Fabric dans la documentation Microsoft Fabric.

Unité 3: Exécuter du code Spark

Type: Contenu

Exécuter du code Spark

Pour modifier et exécuter du code Spark dans Microsoft Fabric, vous pouvez utiliser des notebooks ou définir un travail Spark.

Lorsque vous souhaitez utiliser Spark pour explorer et analyser des données de manière interactive, utilisez un notebook. Les blocs-notes vous permettent de combiner du texte, des images et du code écrits dans plusieurs langues pour créer un élément interactif que vous pouvez partager avec d'autres personnes et collaborer.

Les notebooks se composent d'une ou plusieurs cellules, chacune pouvant avoir du contenu au format markdown ou du code exécutable. Vous pouvez exécuter le code de manière interactive dans le notebook et voir les résultats immédiatement.

Définition d'un travail Spark

Si vous souhaitez utiliser Spark pour ingérer et transformer des données dans le cadre d'un processus automatisé, vous pouvez définir un travail Spark afin d'exécuter un script à la demande ou en fonction d'une planification.

Pour configurer un travail Spark, créez une définition de travail Spark dans votre espace de travail et spécifiez le script qu'il doit exécuter. Vous pouvez également spécifier un fichier de référence (par

exemple un fichier de code Python contenant des définitions de fonctions utilisées dans votre script) et une référence à un lakehouse spécifique contenant des données que le script traite.

Unité 4: Utiliser des données dans un dataframe Spark

Type: Contenu

Utiliser des données dans un dataframe Spark

En mode natif, Spark utilise une structure de données appelée jeu de données distribué résilient (RDD, resilient distributed dataset). Toutefois, même si vous pouvez écrire du code qui fonctionne directement avec des jeux RDD, la structure de données la plus couramment utilisée pour utiliser des données structurées dans Spark est le dataframe, qui est fourni dans le cadre de la bibliothèque Spark SQL. Les dataframes dans Spark sont similaires à ceux de la bibliothèque Pandas Python omniprésente, mais sont optimisés pour fonctionner dans l'environnement de traitement distribué de Spark.

En plus de l'API Dataframe, Spark SQL fournit une API Dataset fortement typée qui est prise en charge dans Java et Scala. Dans ce module, nous allons nous concentrer sur l'API Dataframe.

Chargement des données dans un dataframe

Explorons un exemple hypothétique afin de voir comment utiliser un dataframe pour travailler avec des données. Supposez que vous disposez des données suivantes dans un fichier texte délimité par des virgules appelé products.csv dans le dossier Files/data de votre lakehouse :

Inférence d'un schéma

Dans un notebook Spark, vous pouvez utiliser le code PySpark suivant pour charger les données du fichier dans un dataframe et afficher les 10 premières lignes :

La ligne %%pyspark au début est appelée magic et indique à Spark que le langage utilisé dans cette cellule est PySpark. Vous pouvez sélectionner le langage que vous souhaitez utiliser comme valeur par défaut dans la barre d'outils de l'interface Notebook, puis utiliser une commande magic pour remplacer ce choix pour une cellule spécifique. Par exemple, voici le code Scala équivalent pour l'exemple de données des produits :

La commande magic %%spark est utilisée pour spécifier Scala.

Ces deux exemples de code produisent une sortie comme suit :

Spécification d'un schéma explicite

Dans l'exemple précédent, la première ligne du fichier CSV contenait les noms de colonne, et Spark pouvait déduire le type de données de chaque colonne en se basant sur les données qu'elle contenait. Vous pouvez également spécifier un schéma explicite pour les données, ce qui est utile lorsque les noms de colonne ne sont pas inclus dans le fichier de données, comme cet exemple CSV :

L'exemple PySpark suivant montre comment spécifier un schéma pour que le dataframe soit chargé à partir d'un fichier appelé product-data.csv dans ce format :

Les résultats seraient une fois de plus similaires à :

La spécification d'un schéma explicite améliore également les performances !

Filtrage et regroupement des dataframes

Vous pouvez utiliser les méthodes de la classe Dataframe pour filtrer, trier, regrouper et manipuler les données qu'elle contient. Par exemple, l'exemple de code suivant utilise la méthode select pour récupérer les colonnes ProductID et ListPrice à partir du dataframe df contenant les données de produit de l'exemple précédent :

Les résultats de cet exemple de code devraient ressembler à ceci :

Comme la plupart des méthodes de manipulation de données, select retourne un nouvel objet de dataframe.

La sélection d'une partie des colonnes d'un dataframe est une opération courante, qui peut également être réalisée à l'aide de la syntaxe plus courte suivante :

pricelist_df = df["ProductID", "ListPrice"]

Vous pouvez « chaîner » les méthodes ensemble pour effectuer une série de manipulations qui entraînent un dataframe transformé. Par exemple, cet exemple de code chaîne les méthodes select et where pour créer un dataframe contenant les colonnes ProductName et ListPrice des produits avec la catégorie Vélos VTT ou Vélos de route :

Pour regrouper et agréger des données, vous pouvez utiliser la méthode groupBy et les fonctions d'agrégation. Par exemple, le code PySpark suivant compte le nombre de produits de chaque catégorie :

Enregistrement d'un dataframe

Vous souhaiterez souvent utiliser Spark pour transformer des données brutes et enregistrer les résultats en vue de procéder à une analyse plus approfondie ou un traitement en aval. L'exemple de code suivant enregistre le DataFrame dans un fichier parquet dans le lac de données, en remplaçant tout fichier existant portant le même nom.

Le format Parquet est généralement préféré pour les fichiers de données que vous allez utiliser pour une analyse plus approfondie ou une ingestion dans un magasin analytique. Parquet est un format très efficace qui est pris en charge par la plupart des systèmes d'analytique de données à grande échelle. Parfois, votre besoin de transformation de données peut en fait simplement consister à convertir des données d'un autre format (comme CSV) vers Parquet !

Partitionnement du fichier de sortie

Le partitionnement est une technique d'optimisation qui permet à Spark d'optimiser les performances sur les nœuds Worker. Des gains de performances supplémentaires peuvent être obtenus lors du filtrage des données dans les requêtes en éliminant les E/S disque non nécessaires.

Pour enregistrer un dataframe en tant que jeu de fichiers partitionné, utilisez la méthode partitionBy lors de l'écriture des données. L'exemple suivant enregistre le dataframe bikes_df (qui contient les données de produit pour les catégories mountain bikes et road bikes) et partitionne les données par catégorie :

Les noms de dossiers générés lors du partitionnement d'un dataframe incluent le nom et la valeur de colonne de partitionnement au format column=value, de sorte que l'exemple de code crée un dossier nommé bike_data qui contient les sous-dossiers suivants :

Catégorie=Vélos de route

Chaque sous-dossier contient un ou plusieurs fichiers Parquet avec les données de produit pour la catégorie appropriée.

Vous pouvez partitionner les données selon plusieurs colonnes, ce qui aboutit à une hiérarchie de dossiers pour chaque clé de partitionnement. Par exemple, vous pourriez partitionner des données de commande par année et par mois, afin que la hiérarchie de dossiers inclue un dossier pour chaque

valeur des années, qui à son tour contient un sous-dossier pour chaque valeur des mois.

Charger des données partitionnées

Lors de la lecture de données partitionnées dans un dataframe, vous pouvez charger des données à partir de n'importe quel dossier de la hiérarchie en spécifiant des valeurs explicites ou des caractères génériques pour les champs partitionnés. L'exemple suivant charge des données pour des produits de la catégorie Road Bikes :

Les colonnes de partitionnement spécifiées dans le chemin de fichier sont omises dans le dataframe résultant. Les résultats produits par l'exemple de requête ne vont pas inclure la colonne Category ; la catégorie de toutes les lignes sera Road Bikes.

Unité 5: Utiliser des données à l'aide de Spark SQL

Type: Contenu

Utiliser des données à l'aide de Spark SQL

L'API Dataframe fait partie d'une bibliothèque Spark appelée Spark SQL, qui permet aux analystes Données d'utiliser des expressions SQL pour interroger et manipuler des données.

Création d'objets de base de données dans le catalogue Spark

Le catalogue Spark est un metastore pour les objets de données relationnelles tels que les vues et les tables. Le runtime Spark peut utiliser le catalogue pour intégrer de façon fluide le code écrit dans n'importe quel langage pris en charge par Spark avec des expressions SQL qui peuvent être plus naturelles pour certains analystes Données ou développeurs.

L'une des méthodes les plus simples pour rendre les données d'un dataframe disponibles pour pouvoir les interroger dans le catalogue Spark consiste à créer une vue temporaire, comme illustré dans l'exemple de code suivant :

Une vue est temporaire, ce qui signifie qu'elle est automatiquement supprimée à la fin de la session active. Vous pouvez également créer des tables persistantes dans le catalogue pour définir une base de données pouvant être interrogée à l'aide de Spark SQL.

Les tables sont des structures de métadonnées qui stockent leurs données sous-jacentes dans l'emplacement de stockage associé au catalogue. Dans Microsoft Fabric, les données des tables managées sont stockées dans l'emplacement de stockage Tables indiqué dans votre lac de données, et toutes les tables créées à l'aide de Spark sont répertoriées ici.

Vous pouvez créer une table vide à l'aide de la méthode spark.catalog.createTable, ou vous pouvez enregistrer un dataframe en tant que table à l'aide de sa méthode saveAsTable. La suppression d'une table managée supprime également ses données sous-jacentes.

Par exemple, le code suivant enregistre un dataframe sous la forme d'une nouvelle table nommée produits :

Le catalogue Spark prend en charge les tables basées sur des fichiers dans différents formats. Le format préféré dans Microsoft Fabric est delta, qui est le format d'une technologie de données relationnelles sur Spark nommé Delta Lake. Les tables Delta prennent en charge des fonctionnalités couramment présentes dans les systèmes de base de données relationnelles, notamment les transactions, le versioning et la prise en charge des données de streaming.

En outre, vous pouvez créer des tables externes à l'aide de la spark.catalog.createExternalTable méthode. Les tables externes définissent des métadonnées dans le catalogue, mais obtiennent leurs données sous-jacentes à partir d'un emplacement de stockage externe ; généralement un dossier dans la zone de stockage Fichiers d'un lakehouse. La suppression d'une table externe ne supprime pas les données sous-jacentes.

Vous pouvez appliquer la même technique de partitionnement aux tables Delta Lake que celle décrite pour les fichiers Parquet dans l'unité précédente. Le partitionnement des tables peut améliorer les performances lors de leur interrogation.

Utilisation de l'API Spark SQL pour interroger des données

Vous pouvez utiliser l'API Spark SQL dans le code écrit dans n'importe quel langage pour interroger les données du catalogue. Par exemple, le code PySpark suivant utilise une requête SQL pour retourner des données de la table products en tant que dataframe.

Les résultats de l'exemple de code ressembleraient au tableau suivant :

Utilisation du code SQL

L'exemple précédent a montré comment utiliser l'API Spark SQL pour incorporer des expressions SQL dans le code Spark. Dans un notebook, vous pouvez également utiliser la commande magic %%sql pour exécuter le code SQL qui interroge les objets du catalogue, comme suit :

L'exemple de code SQL retourne un jeu de résultats qui s'affiche automatiquement dans le notebook sous forme de table :

Unité 6: Visualiser des données dans un notebook Spark.

Type: Contenu

Visualiser des données dans un notebook Spark.

L'une des méthodes les plus intuitives pour analyser les résultats des requêtes de données consiste à les visualiser sous forme de graphiques. Les notebooks de Microsoft Fabric offrent des capacités de création de graphiques simples dans l'interface utilisateur, et si cette fonctionnalité ne répond pas à vos besoins, vous pouvez utiliser l'une des nombreuses bibliothèques de graphiques Python pour créer et afficher des visualisations de données dans ce notebook.

Utilisation de graphiques de notebooks intégrés

Lorsque vous affichez un dataframe ou exécutez une requête SQL dans un notebook Spark, les résultats s'affichent sous la cellule de code. Par défaut, les résultats sont restitués sous forme de tableau, mais vous pouvez également remplacer l'affichage des résultats par un graphique et utiliser les propriétés du graphique pour personnaliser la façon dont le graphique visualise les données, comme illustré ici :

La fonctionnalité de graphique intégrée dans les notebooks est utile lorsque vous souhaitez synthétiser rapidement les données visuellement. Lorsque vous souhaitez avoir davantage de contrôle sur la mise en forme des données, pensez à utiliser un package de graphiques pour créer vos propres visualisations.

Utilisation de packages de graphiques dans le code

Il existe de nombreux packages de graphiques que vous pouvez utiliser pour créer des visualisations de données dans le code. En particulier, Python prend en charge une grande sélection de packages ;

la plupart d'entre eux reposent sur la bibliothèque Matplotlib de base. La sortie d'une bibliothèque graphique peut être restituée dans un notebook, ce qui facilite la combinaison du code pour ingérer et manipuler des données avec des visualisations de données inline et des cellules Markdown pour fournir des commentaires.

Par exemple, vous pouvez utiliser le code PySpark suivant pour agréger des données de produits hypothétiques explorées précédemment dans ce module et utiliser Matplotlib pour créer un graphique à partir des données agrégées.

La bibliothèque Matplotlib nécessite que les données se situent dans un dataframe Pandas plutôt qu'un dataframe Spark, de sorte que la méthode toPandas est utilisée pour la convertir. Le code crée ensuite une figure avec une taille spécifiée et trace un graphique à barres avec une configuration de propriété personnalisée avant de montrer le tracé obtenu.

Le graphique produit par le code devrait ressembler à l'image suivante :

Vous pouvez utiliser la bibliothèque Matplotlib pour créer de nombreux types de graphiques ; ou si vous préférez, vous pouvez utiliser d'autres bibliothèques telles que Seaborn pour créer des graphiques hautement personnalisés.

Unité 7: Exercice - Analyser des données avec Apache Spark

Type: Exercice

Exercice - Analyser des données avec Apache Spark

Vous avez maintenant la possibilité de travailler avec Apache Spark dans Microsoft Fabric. Dans cet exercice, vous allez utiliser un notebook Spark pour analyser et visualiser des données à partir de fichiers dans un lakehouse.

Vous avez besoin d'une licence d'évaluation Microsoft Fabric avec la préversion Fabric activée dans votre locataire. Consultez Bien démarrer avec Fabric pour activer votre licence d'évaluation Fabric.

Lancez l'exercice et suivez les instructions.

Unité 8: Évaluation du module

Type: Évaluation

Évaluation du module

Vérifiez vos connaissances

Vous souhaitez utiliser Apache Spark pour explorer des données de manière interactive dans Microsoft Fabric. Que devez-vous créer ?

Une définition de travail Spark.

Un pipeline Data Factory.

Vous devez utiliser Spark pour analyser les données dans un fichier CSV. Quelle est la méthode la plus efficace pour atteindre cet objectif ?

Chargez le fichier dans un dataframe.

Importer les données dans une table d’un entrepôt.

Convertir les données au format Parquet.

Quelle méthode est utilisée pour diviser les données entre des dossiers lors de l’enregistrement d’un dataframe ?

Vous devez répondre à toutes les questions avant de vérifier votre travail.

Unité 9: Résumé

Type: Résumé

Apache Spark est une technologie clé utilisée dans l’analytique Big Data. La prise en charge de Spark dans Microsoft Fabric vous permet d’intégrer le traitement du Big Data dans Spark avec les autres fonctionnalités d’analytique et de visualisation des données de la plateforme.

Pour plus d’informations sur l’utilisation des données dans Spark, consultez le Guide Spark SQL, DataFrames et jeux de données dans la documentation Apache Spark.

Résumé Apache Spark est une technologie clé utilisée dans l’analytique Big Data. La prise en charge de Spark dans Microsoft Fabric vous permet d’intégrer le traitement du Big Data dans Spark avec les autres fonctionnalités d’analytique et de visualisation des données de la plateforme. Conseil Pour plus d’informations sur l’utilisation des données dans Spark, consultez le Guide Spark SQL, DataFrames et jeux de données dans la documentation Apache Spark. Commentaires Yes No

Module 4: Utiliser des tables Delta Lake dans Microsoft Fabric

Unité 1: Présentation

Type: Introduction

Les tables d'un entrepôt de données Microsoft Fabric sont basées sur le format de table Delta Lake, lui-même basé sur Linux et couramment utilisé dans Apache Spark. Delta Lake est une couche de stockage open source pour Spark qui permet des fonctionnalités de base de données relationnelles pour les données de traitement par lots et de diffusion en continu. En utilisant Delta Lake, vous pouvez implémenter une architecture lakehouse pour prendre en charge la sémantique de manipulation des données basée sur SQL dans Spark avec prise en charge des transactions et de l'application du schéma. Le résultat est un magasin de données analytique qui offre de nombreux avantages d'un système de base de données relationnelle avec la flexibilité du stockage de fichiers de données dans un lac de données.

Bien que vous n'ayez pas besoin de travailler directement avec les API Delta Lake pour utiliser des tables dans un lakehouse Fabric, comprendre l'architecture du metastore Delta Lake et être familier avec certaines opérations de table Delta plus spécialisées peut considérablement renforcer votre capacité de créer des solutions d'analyse avancées sur Microsoft Fabric.

Unité 2: Comprendre Delta Lake

Type: Contenu

Comprendre Delta Lake

Delta Lake est une couche de stockage open source qui ajoute la sémantique d'une base de données relationnelle au traitement du lac de données basé sur Spark. Les tables dans les lakehouses Microsoft Fabric sont des tables Delta qui sont indiquées par l'icône Delta triangulaire (■) sur les tables dans l'interface utilisateur des lakehouses.

Les tables Delta sont des abstractions de schéma sur des fichiers de données stockés au format Delta. Pour chaque table, le lakehouse stocke un dossier contenant des fichiers de données Parquet et un dossier _delta_Log dans lequel les détails de transaction sont journalisés au format JSON.

Les avantages d'utiliser des tables Delta sont les suivants :

Des tables relationnelles qui prennent en charge l'interrogation et la modification des données. Avec Apache Spark, vous pouvez stocker des données dans des tables Delta qui prennent en charge les opérations CRUD (créer, lire, mettre à jour et supprimer). En d'autres termes, vous pouvez sélectionner, insérer, mettre à jour et supprimer des lignes de données de la même façon que dans un système de base de données relationnelle.

Prise en charge des transactions ACID. Les bases de données relationnelles sont conçues pour prendre en charge les modifications de données transactionnelles qui fournissent l'atomicité (transactions terminées sous forme d'une unité de travail unique), la cohérence (les transactions laissent la base de données dans un état cohérent), l'isolation (les transactions in-process ne peuvent pas interférer entre elles) et la durabilité (lorsqu'une transaction se termine, les modifications apportées sont persistantes). Delta Lake apporte cette même prise en charge transactionnelle à Spark en implémentant un journal des transactions et en appliquant une isolation sérialisable pour les opérations simultanées.

Contrôle de version des données et voyage dans le temps. Étant donné que toutes les transactions sont enregistrées dans le journal des transactions, vous pouvez suivre plusieurs versions de chaque ligne de table et même utiliser la fonctionnalité de voyage dans le temps pour récupérer une version précédente d'une ligne dans une requête.

Prise en charge des données de traitement par lots et de diffusion en continu. Bien que la plupart des bases de données relationnelles incluent des tables qui stockent des données statiques, Spark inclut la prise en charge native des données de diffusion en continu via l'API Spark Structured Streaming. Les tables Delta Lake peuvent être utilisées en tant que récepteurs (destinations) et sources pour la diffusion en continu de données.

Formats standard et interopérabilité. Les données sous-jacentes des tables Delta sont stockées au format Parquet, qui est couramment utilisé dans les pipelines d'ingestion de lac de données. En outre, vous pouvez utiliser le point de terminaison d'analytique SQL pour le lakehouse Microsoft Fabric afin d'interroger des tables Delta en SQL.

Unité 3: Créer des tables delta

Type: Contenu

Créer des tables delta

Quand vous créez une table dans un lakehouse Microsoft Fabric, une table delta est définie dans le metastore pour le lakehouse et les données de la table sont stockées dans les fichiers Parquet sous-jacents de la table.

Avec la plupart des outils interactifs de l'environnement Microsoft Fabric, les détails du mappage de la définition de la table dans le metastore aux fichiers sous-jacents sont abstraits. Toutefois, quand vous utilisez Apache Spark dans un lakehouse, vous disposez d'un contrôle accru pour créer et gérer des tables delta.

Création d'une table delta à partir d'un dataframe

L'une des méthodes les plus simples pour créer une table delta dans Spark consiste à enregistrer un dataframe au format delta . Par exemple, le code PySpark suivant charge un dataframe avec des données à partir d'un fichier existant, puis enregistre ce dataframe sous la forme d'une table delta :

Le code spécifie que la table doit être enregistrée au format delta avec un nom de table spécifié. Les données de la table sont enregistrées dans des fichiers Parquet (quel que soit le format du fichier source que vous avez chargé dans le dataframe) dans la zone de stockage Tables du lakehouse ainsi qu'un dossier _delta_log contenant les journaux des transactions de la table. La table est répertoriée dans le dossier Tables du lakehouse dans le volet Explorateur de données .

Tables managées et externes

Dans l'exemple précédent, le dataframe a été enregistré en tant que table managée ; cela signifie que la définition de table dans le metastore et les fichiers de données sous-jacents sont tous deux gérés par le runtime Spark pour fabric lakehouse. La suppression de la table supprime également les fichiers sous-jacents de l'emplacement de stockage Tables pour le lakehouse.

Vous pouvez également créer des tables en tant que tables externes , dans lesquelles la définition de table relationnelle dans le metastore est mappée à un autre emplacement de stockage de fichiers. Par exemple, le code suivant crée une table externe pour laquelle les données sont stockées dans le dossier dans l'emplacement de stockage Fichiers pour le lakehouse :

Dans cet exemple, la définition de table est créée dans le metastore (la table est donc répertoriée dans l'interface utilisateur tables pour le lakehouse), mais les fichiers journaux Parquet et les fichiers journaux JSON pour la table sont stockés dans l'emplacement de stockage Fichiers (et sont affichés dans le nœud Fichiers dans le volet Explorateur Lakehouse).

Vous pouvez également spécifier le chemin complet d'un emplacement de stockage, comme suit :

La suppression d'une table externe du metastore lakehouse ne supprime pas les fichiers de données associés.

Création de métadonnées de table

Même s'il est courant de créer une table à partir de données existantes dans un dataframe, il existe souvent des scénarios où vous souhaitez créer une définition de table dans le metastore qui est remplie avec des données d'autres façons. Il existe plusieurs façons d'atteindre cet objectif.

Utiliser l'API DeltaTableBuilder

L'API DeltaTableBuilder vous permet d'écrire du code Spark pour créer une table en fonction de vos spécifications. Par exemple, le code suivant crée une table avec un nom et des colonnes spécifiés.

Utiliser Spark SQL

Vous pouvez également créer des tables delta avec l'instruction Spark SQL CREATE TABLE, comme illustré dans cet exemple :

L'exemple précédent crée une table managée. Vous pouvez également créer une table externe en spécifiant un paramètre LOCATION, comme illustré ici :

Lors de la création d'une table externe, le schéma de la table est déterminé par les fichiers Parquet contenant les données à l'emplacement spécifié. Cette approche peut être utile quand vous souhaitez créer une définition de table qui référence des données qui ont déjà été enregistrées au format delta ou basée sur un dossier dans lequel vous prévoyez d'ingérer des données au format delta.

Enregistrement des données au format delta

Jusqu'à présent, vous avez vu comment enregistrer un dataframe en tant que table delta (création de la définition de schéma de table dans le metastore et des fichiers de données au format delta) et comment créer la définition de table (qui crée le schéma de table dans le metastore sans enregistrer de fichiers de données). Une troisième possibilité consiste à enregistrer des données au format delta sans créer de définition de table dans le metastore. Cette approche peut être utile quand vous souhaitez conserver les résultats des transformations de données es dans Spark dans un format de fichier sur lequel vous pouvez ensuite « superposer » une définition de table ou un processus directement avec l'API Delta Lake.

Par exemple, le code PySpark suivant enregistre un dataframe dans un nouvel emplacement de dossier au format delta :

Les fichiers Delta sont enregistrés au format Parquet dans le chemin d'accès spécifié et incluent un dossier _delta_log contenant des fichiers journaux des transactions. Les journaux des transactions enregistrent les modifications apportées aux données, telles que les mises à jour apportées aux tables externes ou via l'API Delta Lake.

Vous pouvez remplacer le contenu d'un dossier existant par les données d'un dataframe en utilisant le mode overwrite, comme illustré ici :

Vous pouvez également des lignes d'un dataframe à un dossier existant à l'aide du mode d'ajout :

Si vous utilisez la technique décrite ici pour enregistrer un dataframe à l'emplacement tables dans le lakehouse, Microsoft Fabric utilise une fonctionnalité de découverte automatique de tables pour créer

les métadonnées de table correspondantes dans le metastore.

Unité 4: Utiliser des tables delta dans Spark

Type: Contenu

Utiliser des tables delta dans Spark

Vous pouvez utiliser des tables delta (ou des fichiers de format delta) pour récupérer et modifier des données de plusieurs façons.

Utilisation de Spark SQL

La façon la plus courante d'utiliser des données dans des tables delta dans Spark consiste à utiliser Spark SQL. Vous pouvez incorporer des instructions SQL dans d'autres langages (tels que PySpark ou Scala) à l'aide de la bibliothèque spark.sql . Par exemple, le code suivant insère une ligne dans la table products .

Vous pouvez également utiliser la fonctionnalité magique %%sql dans un notebook pour exécuter des instructions SQL.

Utiliser l'API Delta

Lorsque vous souhaitez utiliser des fichiers delta plutôt que des tables de catalogue, il peut être plus simple d'utiliser l'API Delta Lake. Vous pouvez créer une instance d'un DeltaTable à partir d'un emplacement de dossier contenant des fichiers au format delta, puis utiliser l'API pour modifier les données de la table.

Recourir au voyage dans le temps pour utiliser le contrôle de versions de table

Les modifications apportées aux tables delta sont journalisées dans le journal des transactions de la table. Vous pouvez utiliser les transactions journalisées pour afficher l'historique des modifications apportées à la table et récupérer les versions antérieures des données (appelées voyages temporels)

Pour afficher l'historique d'une table, vous pouvez utiliser la DESCRIBE commande SQL comme indiqué ici.

Le résultat de cette instruction indique les transactions qui ont été appliquées à la table, comme illustré ici (certaines colonnes ont été omises) :

Pour afficher l'historique d'une table externe, vous pouvez spécifier l'emplacement du dossier au lieu du nom de la table.

Vous pouvez récupérer des données à partir d'une version spécifique des données en lisant l'emplacement du fichier delta dans un dataframe, en spécifiant la version requise comme versionAsOf option :

Vous pouvez également spécifier un timestamp à l'aide de l'option timestampAsOf :

Unité 5: Utiliser des tables delta avec des données de streaming

Type: Contenu

Utiliser des tables delta avec des données de streaming

Toutes les données que nous avons explorées jusqu'ici étaient des données statiques dans des fichiers. Toutefois, de nombreux scénarios d'analytique des données impliquent le traitement en continu de données qui doivent être traitées en temps quasi réel. Par exemple, vous devrez peut-être capturer les lectures émises par les appareils IoT (Internet des objets) et les stocker dans une table à mesure qu'elles se produisent. Spark traite les données par lots et les données de diffusion en continu de la même façon, ce qui permet de traiter les données de diffusion en continu en temps réel à l'aide de la même API.

Streaming Structuré de Spark

Une solution de traitement de flux typique implique :

Lecture constante d'un flux de données à partir d'une source.

Éventuellement, traitement des données pour sélectionner des champs, des agrégats et des valeurs de groupe spécifiques, ou manipulation des données.

Écriture des résultats dans un récepteur.

Spark inclut la prise en charge native des données de diffusion en continu via Spark Structured Streaming, une API basée sur un dataframe sans limite dans lequel les données de streaming sont capturées pour le traitement. Un DataFrame Spark Structured Streaming peut lire des données à partir de nombreux types de sources de diffusion en continu, notamment :

Services de répartiteur de messages en temps réel tels qu'Azure Event Hubs ou Kafka

Emplacements du système de fichiers.

Pour plus d'informations sur Spark Structured Streaming, consultez le Guide de programmation structured streaming dans la documentation Spark.

Diffusion en continu avec des tables Delta

Vous pouvez utiliser une table Delta comme source ou récepteur pour Spark Structured Streaming. Par exemple, vous pouvez capturer un flux de données en temps réel à partir d'un appareil IoT et écrire le flux directement dans une table Delta en tant que récepteur. Vous pouvez ensuite interroger la table pour afficher les données diffusées en continu les plus récentes. Vous pouvez également lire une table Delta en tant que source de diffusion en continu, ce qui permet de créer des rapports en temps quasi réel lorsque de nouvelles données sont ajoutées à la table.

Utilisation d'une table delta comme source de diffusion en continu

Dans l'exemple PySpark suivant, une table delta est créée pour stocker les détails des commandes commerciales sur Internet :

Un flux de données hypothétique de commandes Internet est inséré dans la table orders_in :

Pour vérifier, vous pouvez lire et afficher des données à partir de la table d'entrée :

Les données sont ensuite chargées dans un DataFrame de diffusion en continu à partir de la table Delta :

Lorsque vous utilisez une table Delta comme source de diffusion en continu, seules les opérations d'ajout peuvent être incluses dans le flux. Les modifications de données peuvent causer une erreur, sauf si vous spécifiez l'option ignoreChanges ou ignoreDeletes.

Vous pouvez vérifier que le flux est en diffusion en continu à l'aide de la propriété isStreaming qui doit retourner True :

Transformer le flux de données

Après avoir lu les données de la table Delta dans un DataFrame de diffusion en continu, vous pouvez utiliser l'API Spark Structured Streaming pour les traiter. Par exemple, vous pouvez compter le nombre de commandes passées toutes les minutes et envoyer les résultats agrégés à un processus en aval pour une visualisation en temps quasi réel.

Dans cet exemple, toutes les lignes avec NULL dans la colonne Price sont filtrées et de nouvelles colonnes sont ajoutées pour IsBike et Total.

Utilisation d'une table Delta comme récepteur de diffusion en continu

Le flux de données est ensuite écrit dans une table Delta :

L'option checkpointLocation est utilisée pour écrire un fichier de point de contrôle qui suit l'état du traitement de flux. Ce fichier vous permet de récupérer à partir d'une défaillance au point où le traitement de flux a été arrêté.

Une fois le processus de diffusion en continu démarré, vous pouvez interroger la table Delta Lake pour voir ce qui se trouve dans la table de sortie. Il peut y avoir un court délai avant de pouvoir interroger la table.

Dans les résultats de cette requête, la commande 3005 est exclue, car elle a la valeur NULL dans la colonne Price. Et les deux colonnes qui ont été ajoutées pendant la transformation sont affichées : IsBike et Total.

Lorsque vous avez terminé, arrêtez les données de diffusion en continu pour éviter les coûts de traitement inutiles à l'aide de la méthode stop :

Pour plus d'informations sur l'utilisation des tables Delta pour le streaming des données, consultez Lectures et écritures de tables en streaming dans la documentation Delta Lake.

Unité 6: Exercice - Utiliser des tables delta dans Apache Spark

Type: Exercice

Exercice - Utiliser des tables delta dans Apache Spark

Maintenant, c'est à votre tour d'utiliser Apache Spark pour manipuler des tables delta dans un lakehouse Microsoft Fabric.

Pour terminer cet exercice, vous avez besoin d'un client Microsoft Fabric. Consultez Prise en main de Fabric pour savoir comment activer une licence d'évaluation Fabric.

Lancez l'exercice et suivez les instructions.

Exercice - Utiliser des tables delta dans Apache Spark Maintenant, c'est à votre tour d'utiliser Apache Spark pour manipuler des tables delta dans un lakehouse Microsoft Fabric. Remarque Pour terminer cet exercice, vous avez besoin d'un client Microsoft Fabric. Consultez Prise en main de Fabric pour savoir comment activer une licence d'évaluation Fabric. Lancez l'exercice et suivez les instructions. Commentaires Yes No

Unité 7: Évaluation du module

Type: Évaluation

Évaluation du module

Vérifier vos connaissances

Parmi les descriptions suivantes, laquelle convient le mieux à Delta Lake ?

API Spark pour l'exportation de données à partir d'une base de données relationnelle dans des fichiers CSV.

Couche de stockage relationnelle pour Spark qui prend en charge les tables basées sur des fichiers Parquet.

Solution de synchronisation qui réplique les données entre SQL Server et SPark.

Vous avez chargé un dataframe Spark avec des données que vous souhaitez maintenant utiliser dans une table delta. Quel format devez-vous utiliser pour écrire le dataframe dans le stockage ?

Vous disposez d'une table managée basée sur un dossier qui contient des fichiers de données au format delta. Si vous supprimez la table, que se passe-t-il ?

Les métadonnées de table et les fichiers de données sont supprimés.

La définition de table est supprimée du metastore, mais les fichiers de données restent intacts.

La définition de table reste dans le metastore, mais les fichiers de données sont supprimés.

Vous devez répondre à toutes les questions avant de vérifier votre travail.

Unité 8: Résumé

Type: Résumé

Delta Lake est une technologie qui ajoute la sémantique de base de données relationnelle à Apache Spark. Les tables d'un lakehouse Microsoft Fabric sont basées sur Delta Lake, ce qui vous permet de tirer parti de nombreuses fonctionnalités et techniques avancées par le biais de l'API Delta Lake.

Conseil / Astuce

Pour plus d'informations sur Delta Lake, consultez la documentation Delta Lake.

Résumé Delta Lake est une technologie qui ajoute la sémantique de base de données relationnelle à Apache Spark. Les tables d'un lakehouse Microsoft Fabric sont basées sur Delta Lake, ce qui vous permet de tirer parti de nombreuses fonctionnalités et techniques avancées par le biais de l'API Delta Lake. Conseil / Astuce Pour plus d'informations sur Delta Lake, consultez la documentation Delta Lake. Commentaires Yes No

Module 5: Orchestrer des processus et le déplacement des données avec Microsoft Fabric

Unité 1: Présentation

Type: Introduction

Les pipelines de données définissent une séquence d'activités qui orchestrent un processus global, généralement en extrayant des données d'une ou plusieurs sources et en les chargeant dans une destination, souvent en les transformant en cours de route. Les pipelines sont couramment utilisés pour automatiser les processus d'extraction, de transformation et de chargement (ETL) qui ingèrent des données transactionnelles à partir de magasins de données opérationnels dans un magasin de données analytique, tel qu'un lachouse, un entrepôt de données ou une base de données SQL.

Si vous êtes déjà familiarisé avec Azure Data Factory, les pipelines de données dans Microsoft Fabric vous seront immédiatement familiers. Ils utilisent la même architecture d'activités connectées pour définir un processus qui peut inclure plusieurs types de tâches de traitement des données et la logique de flux de contrôle. Vous pouvez exécuter des pipelines de manière interactive dans l'interface utilisateur de Microsoft Fabric ou planifier leur exécution automatique.

Unité 2: Comprendre les pipelines

Type: Contenu

Comprendre les pipelines

Les pipelines dans Microsoft Fabric encapsulent une séquence d'activités qui effectuent des tâches de déplacement et de traitement des données. Vous pouvez utiliser un pipeline pour définir des activités de transfert et de transformation de données, et orchestrer ces activités via des activités de flux de contrôle qui gèrent le branchement, le bouclage et d'autres logiques de traitement standard. Le canevas de pipeline graphique dans l'interface utilisateur Fabric vous permet de créer des pipelines complexes avec un développement minimal ou nul.

Concepts principaux des pipelines

Avant de créer des pipelines dans Microsoft Fabric, vous devez comprendre quelques concepts de base.

Les activités sont les tâches exécutables dans un pipeline. Vous pouvez définir un flux d'activités en les connectant dans une séquence. Le résultat d'une activité particulière (réussite, échec ou achèvement) peut être utilisé pour diriger le flux vers l'activité suivante dans la séquence.

Il existe deux grandes catégories d'activités dans un pipeline.

Activités de transformation de données : activités qui encapsulent les opérations de transfert de données, notamment les activités de copie simples de données qui extraient des données d'une source et les chargent vers une destination, et les activités de flux de données plus complexes qui encapsulent des flux de données (Gen2) qui appliquent des transformations aux données lors du transfert. D'autres activités de transformation de données incluent des activités notebook pour exécuter un notebook Spark, des activités de procédure stockée pour exécuter du code SQL, des activités supprimer des données pour supprimer des données existantes, et d'autres. Dans OneLake, vous pouvez configurer la destination sur un lakehouse, un entrepôt, une base de données SQL ou d'autres options.

Activités de flux de contrôle : activités que vous pouvez utiliser pour implémenter des boucles, des branchements conditionnels ou gérer des valeurs de variables et de paramètres. Le large éventail d'activités de flux de contrôle vous permet d'implémenter une logique de pipeline complexe pour orchestrer l'ingestion et le flux de transformation des données.

Pour plus d'informations sur l'ensemble complet d'activités de pipeline disponibles dans Microsoft Fabric, consultez la vue d'ensemble de l'activité dans la documentation de Microsoft Fabric.

Les pipelines peuvent être paramétrisés, ce qui vous permet de fournir des valeurs spécifiques à utiliser chaque fois qu'un pipeline est exécuté. Par exemple, vous pouvez utiliser un pipeline pour enregistrer les données ingérées dans un dossier, mais avoir la possibilité de spécifier un nom de dossier chaque fois que le pipeline est exécuté.

L'utilisation de paramètres augmente la réutilisabilité de vos pipelines, ce qui vous permet de créer des processus flexibles d'ingestion et de transformation des données.

Exécutions de pipeline

Chaque fois qu'un pipeline est exécuté, une exécution de pipeline de données est lancée. Les exécutions peuvent être lancées à la demande dans l'interface utilisateur de Fabric ou planifiées pour démarrer à une fréquence spécifique. Utilisez l'ID d'exécution unique pour passer en revue les détails de l'exécution afin de vérifier qu'elle s'est déroulée correctement et d'examiner les paramètres spécifiques utilisés pour chaque exécution.

Unité 3: Utiliser l'activité Copier des données

Type: Contenu

Utiliser l'activité Copier des données

L'activité Copier des données est l'une des utilisations les plus courantes d'un pipeline de données. De nombreux pipelines se composent d'une seule activité Copier des données qui est utilisée pour ingérer des données à partir d'une source externe dans un fichier ou une table de lakehouse.

Vous pouvez également combiner l'activité Copier des données avec d'autres activités pour créer un processus d'ingestion de données reproductible, par exemple en utilisant une activité Supprimer des données pour supprimer des données existantes, une activité Copier des données pour remplacer les données supprimées par un fichier contenant des données provenant d'une source externe et une activité Notebook pour exécuter du code Spark qui transforme les données dans le fichier et les charge dans une table.

L'outil Copier des données

Quand vous ajoutez une activité Copier des données à un pipeline, un outil graphique vous guide tout au long des étapes requises pour configurer la source de données et la destination de l'opération de copie. Un large éventail de connexions sources est pris en charge, ce qui permet d'ingérer des données à partir des sources les plus courantes. Dans OneLake, les destinations prises en charge sont un lakehouse, un entrepôt, SQL Database, etc.

Paramètres de l'activité Copier des données

Une fois que vous avez ajouté une activité Copier des données à un pipeline, vous pouvez la sélectionner dans le canevas du pipeline et modifier ses paramètres dans le volet en dessous.

Quand utiliser l'activité Copier des données

Utilisez l’activité Copier des données quand vous devez copier des données directement entre une source et une destination prises en charge sans appliquer de transformations, ou quand vous souhaitez importer les données brutes et appliquer des transformations dans des activités de pipeline ultérieures.

Si vous devez appliquer des transformations aux données à mesure qu’elles sont ingérées ou fusionner des données provenant de plusieurs sources, envisagez d’utiliser une activité de flux de données pour exécuter un flux de données (Gen2). Vous pouvez utiliser l’interface utilisateur de Power Query pour définir un flux de données (Gen2) qui inclut plusieurs étapes de transformation et l’inclure dans un pipeline.

Pour en découvrir plus sur les flux de données (Gen2) dans Microsoft Fabric pour ingérer des données, envisagez d’effectuer le module Ingérer des données avec des flux de données Gen2 dans Microsoft Fabric.

Unité 4: Utiliser des modèles de pipeline

Type: Contenu

Utiliser des modèles de pipeline

Vous pouvez définir des pipelines à partir de n’importe quelle combinaison d’activités que vous choisissez, ce qui vous permet de créer des processus d’ingestion et de transformation de données personnalisés pour répondre à vos besoins spécifiques. Toutefois, il existe de nombreux scénarios de pipeline courants pour lesquels Microsoft Fabric inclut des modèles de pipeline prédéfinis que vous pouvez utiliser et personnaliser en fonction des besoins.

Pour créer un pipeline basé sur un modèle, sélectionnez la vignette Modèles dans un nouveau pipeline, comme illustré ici.

La sélection de cette option affiche une sélection de modèles de pipeline, comme illustré ici.

Vous pouvez sélectionner le modèle le plus adapté à vos besoins, puis modifier le pipeline dans le canevas de pipeline pour le personnaliser en fonction de vos besoins.

Unité 5: Exécuter et superviser des pipelines

Type: Contenu

Exécuter et superviser des pipelines

Une fois que vous avez terminé un pipeline, vous pouvez utiliser l’option Valider pour vérifier que la configuration est valide, puis l’exécuter de manière interactive ou spécifier une planification.

Capture d’écran du menu d'exécution pour un pipeline dans Microsoft Fabric.

Afficher l’historique des exécutions

Vous pouvez afficher l’historique des exécutions d’un pipeline pour afficher les détails de chaque exécution, à partir du canevas du pipeline ou de l’élément de pipeline listé dans la page de l’espace de travail.

Unité 6: Exercice - Ingérer des données avec un pipeline

Type: Exercice

Exercice - Ingérer des données avec un pipeline

Vous avez maintenant la possibilité d'implémenter un pipeline dans Microsoft Fabric. Dans cet exercice, vous créez un pipeline qui copie les données d'une source externe dans un lakehouse. Ensuite, vous améliorez le pipeline en ajoutant des activités pour transformer les données ingérées.

Vous avez besoin d'une licence d'évaluation Microsoft Fabric avec la préversion Fabric activée dans votre locataire. Consultez Bien démarrer avec Fabric pour activer votre licence d'évaluation Fabric.

Lancez l'exercice et suivez les instructions.

Unité 7: Évaluation du module

Type: Évaluation

Évaluation du module

Vérifier vos connaissances

Qu'est-ce qu'un pipeline de données ?

Dossier spécial dans le stockage OneLake où les données peuvent être exportées à partir d'un lakehouse

Séquence d'activités permettant d'orchestrer un processus d'ingestion ou de transformation de données

Une requête Power Query enregistrée

Vous souhaitez utiliser un pipeline pour copier des données dans un dossier avec un nom spécifié pour chaque exécution. Que devez-vous faire ?

Créer plusieurs pipelines : un pour chaque nom de dossier

Utiliser un dataflow (Gen2)

un paramètre au pipeline et l'utiliser pour spécifier le nom du dossier pour chaque exécution

Vous avez précédemment exécuté un pipeline contenant plusieurs activités. Quelle est la meilleure façon de vérifier combien de temps chaque activité individuelle a pris pour se terminer ?

Réexécutez le pipeline et observez la sortie, minutez chaque activité.

Affichez les détails de l'exécution dans l'historique des exécutions.

Afficher la valeur actualisée pour le jeu de données par défaut de votre lakehouse

Vous devez répondre à toutes les questions avant de vérifier votre travail.

Unité 8: Résumé

Type: Résumé

Avec Microsoft Fabric, vous pouvez créer des pipelines qui encapsulent des processus complexes d'ingestion et de transformation des données. Les pipelines offrent un moyen efficace d'orchestrer les tâches de traitement des données qui peuvent être exécutées à la demande ou à intervalles planifiés.

Conseil / Astuce

Pour plus d'informations sur les pipelines dans Microsoft Fabric, consultez les pipelines de données dans la documentation Microsoft Fabric.

Résumé Avec Microsoft Fabric, vous pouvez créer des pipelines qui encapsulent des processus complexes d'ingestion et de transformation des données. Les pipelines offrent un moyen efficace d'orchestrer les tâches de traitement des données qui peuvent être exécutées à la demande ou à intervalles planifiés. Conseil / Astuce Pour plus d'informations sur les pipelines dans Microsoft Fabric, consultez les pipelines de données dans la documentation Microsoft Fabric. Commentaires Yes No

Module 6: Ingérer des données avec des flux de données Gen2 dans Microsoft Fabric

Unité 1: Présentation

Type: Introduction

Microsoft Fabric offre une solution unifiée pour l'engineering, l'intégration et l'analytique des données. L'ingestion des données constitue une étape cruciale de l'analytique de bout en bout. Les flux de données Gen2 sont utilisés pour ingérer et transformer des données à partir de plusieurs sources, puis pour déplacer les données nettoyées vers une autre destination. Il est possible de les intégrer à des pipelines de données pour une orchestration plus complexe des activités, et de les utiliser comme source de données dans Power BI.

Imaginez que vous travaillez pour une entreprise de vente au détail avec des magasins dans le monde entier. En tant qu'ingénieur Données, vous devez préparer et transformer des données provenant de différentes sources dans un format adapté à l'analyse et à la création de rapports de données. L'entreprise demande un modèle sémantique qui regroupe des sources de données disparates à partir des différents magasins. Les flux de données Gen2 vous permettent de préparer les données pour garantir la cohérence, puis de les placer dans la destination de prédilection. Ils permettent également la réutilisation des données et facilitent leur mise à jour. Sans flux de données, vous devriez extraire et transformer manuellement les données de chaque source, ce qui est fastidieux et sujet aux erreurs.

Dans ce module, nous expliquons comment utiliser des flux de données Gen2 dans Microsoft Fabric pour répondre à vos besoins d'ingestion de données.

Unité 2: Comprendre les flux de données Gen2 dans Microsoft Fabric

Type: Contenu

Comprendre les flux de données Gen2 dans Microsoft Fabric

Dans notre scénario, vous devez développer un modèle sémantique capable de normaliser les données et de fournir un accès à l'entreprise. En utilisant des flux de données Gen2, vous pouvez vous connecter aux différentes sources de données, puis préparer et transformer les données. Vous pouvez placer les données directement dans votre lakehouse ou utiliser un pipeline de données pour d'autres destinations.

Qu'est-ce qu'un flux de données ?

Les flux de données sont un type d'outil ETL (Extract, Transform, Load) cloud qui permet de créer et d'exécuter des processus de transformation de données scalables.

Les flux de données Gen2 vous permettent d'extraire des données de différentes sources, de les transformer avec un large éventail d'opérations de transformation et de les charger dans une destination. Power Query Online met également à votre disposition une interface visuelle pour effectuer ces tâches.

Par essence, un flux de données inclut toutes les transformations nécessaires pour réduire le temps de préparation des données, puis peut être chargé dans une nouvelle table, inclus dans un pipeline de données ou utilisé comme source de données par les analystes Données.

Utilisation des flux de données Gen2

Traditionnellement, les ingénieurs Données consacrent beaucoup de temps à extraire, transformer et charger des données dans un format consommable pour l'analytique en aval. L'objectif des flux de données Gen2 est de fournir un moyen simple et réutilisable d'effectuer des tâches ETL avec Power Query Online.

Si vous choisissez uniquement d'utiliser un pipeline de données, vous copiez des données, puis utilisez votre langage de programmation favori pour extraire, transformer et charger les données. Vous pouvez également créer au préalable un flux de données Gen2 pour extraire et transformer les données. Vous pouvez également charger les données dans un lakehouse et d'autres destinations. Désormais, l'entreprise peut facilement consommer le modèle sémantique organisé.

L'ajout d'une destination de données à votre flux de données est facultatif et le flux de données conserve toutes les étapes de transformation. Pour effectuer d'autres tâches ou charger des données dans une autre destination après la transformation, créez un pipeline de données et ajoutez l'activité de flux de données Gen2 à votre orchestration.

Une autre option peut consister à utiliser un pipeline de données et un flux de données Gen2 pour le processus ELT (Extract, Load, Transform). Pour cet ordre, vous utiliseriez un pipeline afin d'extraire les données et de les charger dans votre destination préférée telle que le lakehouse. Ensuite, vous créeriez un flux de données Gen2 pour vous connecter aux données de Lakehouse afin de nettoyer et de transformer des données. Dans ce cas, vous proposeriez le flux de données en tant que modèle sémantique organisé pour permettre aux Analystes Données de développer des rapports.

Les flux de données peuvent également être partitionnés horizontalement. Une fois que vous avez créé un flux de données global, les Analystes Données peuvent utiliser des flux de données afin de créer des modèles sémantiques spécialisés pour des besoins spécifiques.

Les flux de données vous permettent de promouvoir une logique ETL réutilisable qui évite la nécessité de créer plus de connexions à votre source de données. Les flux de données offrent une grande variété de transformations et peuvent être exécutés manuellement, selon une planification d'actualisation ou dans le cadre d'une orchestration de pipeline de données.

Rendez votre flux de données découvrable afin que les analystes Données puissent également s'y connecter via Power BI Desktop. Cela réduit la préparation des données pour le développement de rapports.

Avantages et limitations

Il existe plusieurs façons d'obtenir des données ETL ou ELT dans Microsoft Fabric. Tenez compte des avantages et des limitations de l'utilisation des flux de données Gen2.

Étendez les données avec des données cohérentes, telles qu'une table de dimension de date standard.

Autoriser les utilisateurs en libre-service à accéder à un sous-ensemble de l'entrepôt de données séparément.

Optimisez les performances avec des flux de données, qui permettent d'extraire les données une fois en vue de les réutiliser, ce qui réduit le temps d'actualisation des données pour les sources plus lentes.

Simplifiez la complexité des sources de données en exposant uniquement les flux de données à des groupes d'analystes plus importants.

Assurez la cohérence et la qualité des données en permettant aux utilisateurs de nettoyer et de transformer des données avant de les charger dans une destination.

Simplifiez l'intégration des données en fournissant une interface low-code qui ingère les données de différentes sources.

Les flux de données ne remplacent pas un entrepôt de données.

Ne prend pas en charge la sécurité au niveau des lignes.

Un espace de travail à capacité structurelle est requis.

Unité 3: Explorer un flux de données Gen2 dans Microsoft Fabric

Type: Contenu

Explorer un flux de données Gen2 dans Microsoft Fabric

Dans Microsoft Fabric, vous pouvez créer un flux de données Gen2 dans la charge de travail Data Factory ou l'espace de travail Power BI, ou directement dans le lakehouse. Étant donné que notre scénario est axé sur l'ingestion des données, examinons l'expérience de la charge de travail Data Factory . Les flux de données Gen2 utilisent Power Query Online pour visualiser les transformations. Consultez une vue d'ensemble de l'interface :

1. Ruban Power Query

Les flux de données (Gen2) prennent en charge un large éventail de connecteurs de source de données. Les sources courantes incluent les bases de données relationnelles cloud et locales, les fichiers plats ou Excel, SharePoint, SalesForce, Spark et les lakehouses Fabric. Il existe ensuite de nombreuses transformations de données possibles, telles que :

Filtrer et trier les lignes

un tableau croisé dynamique et supprimer le tableau croisé dynamique

Fusionner et des requêtes

Fractionnement et fractionnement conditionnel

Remplacer les valeurs et supprimer les doublons

, renommer, réorganiser ou supprimer des colonnes

Calculatrice de classement et de pourcentage

Choisir les N premières et N dernières valeurs

Vous pouvez également créer et gérer des connexions de sources de données, gérer des paramètres et configurer la destination des données par défaut dans ce ruban.

2. Volet Requêtes

Le volet Requêtes affiche les différentes sources de données , désormais appelées requêtes. Ces requêtes sont appelées tables lorsqu'elles sont chargées dans votre magasin de données. Vous pouvez dupliquer ou référencer une requête si vous avez besoin de plusieurs copies des mêmes données, telles que la création d'un schéma en étoile et le fractionnement de données en tables distinctes et plus petites. Vous pouvez également désactiver la charge d'une requête, au cas où vous n'avez besoin que de l'importation ponctuelle.

3. Affichage Diagramme

L'affichage Diagramme vous permet de voir comment les sources de données sont connectées et les différentes transformations appliquées. Par exemple, votre flux de données se connecte à une source de données, duplique la requête, supprime les colonnes de la requête source, puis désactive la requête

dupliquée. Chaque requête est représentée par une forme avec toutes les transformations appliquées et reliée par une ligne pour la requête dupliquée. Vous pouvez activer ou désactiver cette vue.

4. Volet Aperçu des données

Le volet Aperçu des données affiche uniquement un sous-ensemble de données pour vous permettre de voir quelles transformations vous devez effectuer et comment elles affectent les données. Vous pouvez également interagir avec le volet d'aperçu en faisant glisser des colonnes et en les déposant pour changer l'ordre ou en cliquant avec le bouton droit sur les colonnes pour filtrer ou apporter des modifications. L'aperçu des données affiche toutes vos transformations pour la requête sélectionnée.

5. Volet Paramètres de la requête

Le volet Paramètres de requête inclut les étapes appliquées. Chaque transformation est représentée par une étape, dont certaines sont automatiquement appliquées lorsque vous connectez la source de données. Selon la complexité des transformations, plusieurs étapes peuvent être appliquées pour chaque requête. La plupart des étapes ont une icône d'engrenage qui vous permet de modifier l'étape, sinon vous devez supprimer et répéter la transformation.

Chaque étape dispose également d'un menu contextuel lorsque vous cliquez avec le bouton droit pour pouvoir renommer, réorganiser ou supprimer les étapes. Vous pouvez également afficher la requête de la source de données lorsque vous vous connectez à une source de données qui prend en charge le pliage des requêtes.

Bien que cette interface visuelle soit utile, vous pouvez également afficher le code M via l'éditeur avancé.

Dans le volet Paramètres de requête, vous pouvez voir une option de destination de données pour atterrir vos données à l'un des emplacements suivants dans votre environnement Fabric :

Base de données SQL

Vous pouvez également charger votre flux de données dans Azure SQL Database, Azure Data Explorer ou Azure Synapse Analytics.

Les flux de données Gen2 fournissent une solution à faible niveau de code ou sans code pour ingérer, transformer et charger des données dans vos magasins de données Fabric. Les développeurs Power BI sont familiarisés et peuvent rapidement commencer à effectuer des transformations en amont pour améliorer la performance de leurs rapports.

Pour plus d'informations, consultez la documentation Power Query pour optimiser vos flux de données.

Unité 4: Intégrer des flux de données Gen2 et des pipelines dans Microsoft Fabric

Type: Contenu

Intégrer des flux de données Gen2 et des pipelines dans Microsoft Fabric

Les flux de données Gen2 constituent une excellente option pour les transformations de données dans Microsoft Fabric. La combinaison de flux de données et de pipelines est utile quand vous devez effectuer des opérations supplémentaires sur les données transformées.

Les pipelines de données sont un concept courant dans l'engineering données et offrent une grande variété d'activités à orchestrer. Voici certaines activités courantes :

Copier des données

Incorporer un flux de données

Obtenir les métadonnées

Exécuter un script ou une procédure stockée

Les pipelines fournissent un moyen visuel d'effectuer des activités dans un ordre spécifique. Vous pouvez utiliser un flux de données pour l'ingestion, la transformation et l'arrivée des données dans un magasin de données Fabric. Incorporez ensuite le flux de données dans un pipeline pour orchestrer des activités supplémentaires, telles que l'exécution de scripts ou de procédures stockées une fois le flux de données terminé.

Les pipelines peuvent également être planifiés ou activés par un déclencheur pour exécuter votre flux de données. À l'aide d'un pipeline pour exécuter votre flux de données, vous pouvez avoir les données actualisées lorsque vous en avez besoin au lieu de devoir exécuter manuellement le flux de données. Lorsque vous travaillez avec des données d'entreprise ou fréquemment modifiées, l'automatisation vous permet de vous concentrer sur d'autres responsabilités.

Unité 5: Exercice – Créer et utiliser un flux de données Gen2 dans Microsoft Fabric

Type: Exercice

Exercice – Créer et utiliser un flux de données Gen2 dans Microsoft Fabric

Dans cet exercice, vous allez utiliser un flux de données Gen2 pour charger des données transformées dans un lakehouse et un flux de données à un pipeline.

Vous avez besoin d'une licence d'évaluation Microsoft Fabric avec la préversion Fabric activée dans votre locataire. Consultez Bien démarrer avec Fabric pour activer votre licence d'évaluation Fabric.

Lancez l'exercice et suivez les instructions.

Exercice – Créer et utiliser un flux de données Gen2 dans Microsoft Fabric Dans cet exercice, vous allez utiliser un flux de données Gen2 pour charger des données transformées dans un lakehouse et un flux de données à un pipeline. Remarque Vous avez besoin d'une licence d'évaluation Microsoft Fabric avec la préversion Fabric activée dans votre locataire. Consultez Bien démarrer avec Fabric pour activer votre licence d'évaluation Fabric. Lancez l'exercice et suivez les instructions.

Unité 6: Évaluation du module

Type: Évaluation

Évaluation du module

Qu'est-ce qu'un flux de données Gen2 ?

Base de données hybride qui prend en charge les transactions ACID.

Un moyen d'exporter des données vers Power BI Desktop.

Un moyen d'importer et de transformer des données avec Power Query Online.

Quelle expérience de charge de travail vous permet de créer un flux de données Gen2 ?

Real-Time Intelligence.

Entrepôt de données.

Fabrique de données.

Vous devez vous connecter et transformer les données à charger dans un lakehouse Fabric. Comme vous n'êtes pas à l'aise avec les notebooks Spark, vous décidez d'utiliser des flux de données Gen2. Comment procéderiez-vous ?

Connectez-vous à la charge de travail Data Factory > Créez un flux de données Gen2 pour transformer des données > Ajoutez votre lakehouse comme destination de données.

Se connecter à la charge de travail Real-time Intelligence > Créer un pipeline pour copier des données > Transformer des données avec un Eventstream.

Connectez-vous à la charge de travail Data Factory > Créez un pipeline pour copier des données et les charger dans un lakehouse > Transformez les données directement dans le lakehouse.

Vous devez répondre à toutes les questions avant de vérifier votre travail.

Unité 7: Résumé

Type: Résumé

Dans ce module, nous avons parcouru un scénario dans lequel les deux ingénieurs données doivent ingérer, transformer et charger des données dans un magasin de données Fabric tel qu'un lakehouse. Nous avons également identifié que les analystes données doivent effectuer des transformations plus proches de la source de données pour prendre en charge le développement de rapports Power BI.

Avec Microsoft Fabric, vous pouvez créer des flux de données Gen2 pour effectuer l'intégration des données pour votre lakehouse, et éventuellement inclure le flux de données dans un pipeline de données. Vous avez découvert les flux de données Gen2 et comment les utiliser dans le cadre de votre processus d'intégration des données. Power Query Online offre une interface visuelle permettant d'effectuer des transformations de données complexes sans écrire de code.

Pour en savoir plus sur l'intégration des données, consultez la documentation sur Data Factory dans Microsoft Fabric.

Module 7: Bien démarrer avec les entrepôts de données dans Microsoft Fabric

Unité 1: Présentation

Type: Introduction

Les entrepôts de données relationnelles sont au centre de la plupart des solutions décisionnels d'entreprise . Bien que les détails spécifiques puissent varier entre les implémentations de l'entrepôt de données, un modèle courant basé sur un schéma dénormalisé, le schéma multidimensionnel a émergé comme la conception standard d'un entrepôt de données relationnelle.

L'entrepôt de données de Microsoft Fabric est une version moderne de l'entrepôt de données traditionnel. Il centralise et organise les données de différents services, systèmes et bases de données en une seule vue unifiée à des fins d'analyse et de création de rapports. L'entrepôt de données de Fabric fournit une sémantique SQL complète, notamment la possibilité d'insérer, de mettre à jour et de supprimer des données dans les tables. L'entrepôt de données de Fabric est unique, car il est basé sur lakehouse, qui est stocké au format Delta et peut être interrogé à l'aide de SQL. Il est conçu pour être utilisé par toute l'équipe de données, pas seulement pour les ingénieurs données.

L'expérience de l'entrepôt de données de Fabric est conçue pour relever ces défis. Fabric permet aux ingénieurs de données, aux analystes et aux scientifiques des données de travailler ensemble pour créer et interroger un entrepôt de données optimisé pour leurs besoins spécifiques.

Dans ce module, vous allez découvrir les entrepôts de données dans Fabric, créer un entrepôt de données, charger, interroger et visualiser des données.

Unité 2: Comprendre les principes de base d'un entrepôt de données

Type: Contenu

Comprendre les principes de base d'un entrepôt de données

Le processus de création d'un entrepôt de données moderne se compose généralement des tâches suivantes :

Ingestion des données : déplacement des données de systèmes sources vers un entrepôt de données.

Stockage des données : stockage des données dans un format optimisé pour l'analytique.

Traitement des données : transformation des données dans un format consommable par les outils analytiques.

Analyse et remise des données : analyse des données pour obtenir des insights et remise de ces insights à l'entreprise.

Microsoft Fabric permet aux ingénieurs et aux analystes de données d'ingérer, de stocker, de transformer et de visualiser des données dans un seul outil en combinant expérience traditionnelle et « low-code ».

Comprendre l'expérience d'entrepôt de données de Fabric

L'entrepôt de données de Fabric est un entrepôt de données relationnelle qui prend en charge les fonctionnalités T-SQL transactionnelles complètes que vous attendez d'un entrepôt de données

d’entreprise. Complètement managé, scalable et hautement disponible, il peut être utilisé pour stocker et interroger des données dans le lakehouse. L’entrepôt de données vous permet de contrôler entièrement la création de tables ainsi que le chargement, la transformation et l’interrogation des données à l’aide du portail Fabric ou de commandes T-SQL. Vous pouvez utiliser soit SQL pour interroger et analyser les données, soit Spark pour traiter les données et créer des modèles Machine Learning.

Les entrepôts de données dans Fabric facilitent la collaboration entre les ingénieurs données et les analystes de données, qui partagent alors la même expérience. Les ingénieurs données créent, au-dessus des données dans le lakehouse, une couche relationnelle dans laquelle les analystes peuvent utiliser T-SQL et Power BI pour explorer les données.

Concevoir un entrepôt de données

Comme toutes les bases de données relationnelles, l’entrepôt de données de Fabric contient des tables pour stocker les données à des fins d’analytique. Le plus souvent, ces tables sont organisées dans un schéma optimisé pour la modélisation multidimensionnelle. Dans cette approche, les données numériques liées aux événements (par exemple, les commandes des clients) sont regroupées selon différents attributs (date, client, magasin, etc.). Par exemple, vous pouvez analyser le montant total payé pour les commandes passées à une date spécifique ou dans un magasin particulier.

Tables d’un entrepôt de données

Les tables d’un entrepôt de données sont généralement organisées de manière à analyser efficacement de grandes quantités de données. Cette organisation, souvent appelée « modélisation dimensionnelle », implique de structurer les tables en tables de faits et tables de dimension.

Les tables de faits contiennent les données numériques que vous souhaitez analyser. Les tables de faits comprennent généralement un grand nombre de lignes et constituent la principale source de données pour l’analyse. Par exemple, une table de faits peut contenir le montant total payé pour des commandes passées à une date spécifique ou dans un magasin particulier.

Les tables de dimension contiennent des informations descriptives sur les données des tables de faits. Les tables de dimension comprennent généralement un petit nombre de lignes et fournissent le contexte des données des tables de faits. Par exemple, une table de dimension peut contenir des informations sur les clients qui ont passé des commandes.

En plus des colonnes d’attribut, une table de dimension contient une colonne clé unique qui identifie de manière unique chaque ligne de la table. En fait, il est courant pour une table de dimension d’inclure deux colonnes clés :

Une clé de substitution est un identificateur unique pour chaque ligne de la table de dimension. Il s’agit souvent d’une valeur entière générée automatiquement par le système de gestion de base de données quand une nouvelle ligne est insérée dans la table.

Une autre clé est souvent une clé naturelle ou métier qui identifie une instance spécifique d’une entité dans le système source transactionnel, comme un code produit ou un ID client.

Dans un entrepôt de données, les clés de substitution et les clés secondaires ont des finalités différentes. Vous avez donc besoin des deux. Les clés de substitution sont spécifiques à l’entrepôt de données et contribuent au maintien de la cohérence et de l’exactitude des données. Quant aux clés alternatives, elles sont spécifiques au système source et contribuent au maintien de la traçabilité entre l’entrepôt de données et le système source.

Tables de dimension de type spécial

Les dimensions de type spécial offrent un contexte supplémentaire et permettent une analyse des données plus complète.

Les dimensions de temps fournissent des informations sur la période pendant laquelle un événement s'est produit. Cette table permet aux analystes de données d'agréger des données sur des intervalles temporels. Par exemple, une dimension de temps peut inclure les colonnes « année », « trimestre », « mois » et « jour » pour indiquer quand une commande a été passée.

Les dimensions à variation lente sont des tables de dimension qui suivent les modifications apportées aux attributs de dimension au fil du temps, telles que les modifications apportées à l'adresse d'un client ou au prix d'un produit. Elles occupent une place importante dans un entrepôt de données, car elles permettent aux utilisateurs d'analyser et de comprendre les modifications apportées aux données dans le temps. Les dimensions à variation lente garantissent que les données sont à jour et exactes, ce qui est primordial pour prendre de bonnes décisions commerciales.

Conceptions de schémas d'entrepôts de données

Dans la plupart des bases de données transactionnelles utilisées dans les applications métier, les données sont normalisées pour réduire la duplication. Toutefois, dans un entrepôt de données, les données de dimension sont généralement dénormalisées pour réduire le nombre de jointures requises pour interroger les données.

Souvent, un entrepôt de données est organisé en tant que schéma en étoile, dans lequel une table de faits est directement liée aux tables de dimension, comme illustré dans cet exemple :

Vous pouvez utiliser les attributs d'un élément pour regrouper des nombres dans la table de faits à différents niveaux. Par exemple, vous pouvez trouver le chiffre d'affaires total d'une région entière ou d'un seul client. Les informations de chaque niveau peuvent être stockées dans la même table de dimension.

Voir Qu'est-ce qu'un schéma en étoile ? pour plus d'informations sur la conception de schémas en étoile pour Fabric.

S'il existe un grand nombre de niveaux ou si certaines informations sont partagées par des éléments différents, il peut être judicieux d'utiliser un schéma flocon à la place. Voici un exemple :

Dans ce cas, la table DimProduct a été divisée (normalisée) pour créer des tables de dimension distinctes pour les catégories de produits et les fournisseurs.

Chaque ligne de la table DimProduct contient des valeurs clés pour les lignes correspondantes dans les tables DimCategory et DimSupplier.

Une table DimGeography a été ajoutée contenant des informations sur l'emplacement des clients et des magasins.

Chaque ligne des tables DimCustomer et DimStore contient une valeur clé pour la ligne correspondante dans la table DimGeography .

Unité 3: Comprendre les entrepôts de données dans Fabric

Type: Contenu

Comprendre les entrepôts de données dans Fabric

Le lakehouse de Fabric est une collection de fichiers, dossiers, tables et raccourcis qui agissent comme une base de données sur un lac de données. Utilisé par le moteur Spark et le moteur SQL pour le traitement du Big Data, il propose des fonctionnalités pour les transactions ACID lors de l'utilisation de tables au format Delta open source.

L’expérience d’entrepôt de données de Fabric vous permet de passer de la vue du lac du lakehouse (qui prend en charge l’engineering données et Apache Spark) aux expériences SQL d’un entrepôt de données traditionnel. Le Lakehouse vous permet de lire des tables et d’utiliser le point de terminaison d’analyse SQL, tandis que l’entrepôt de données vous permet de manipuler les données.

Dans l’expérience d’entrepôt de données, vous pouvez modéliser les données à l’aide de tables et de vues, exécuter des commandes T-SQL pour interroger les données dans l’entrepôt de données et le lakehouse, utiliser T-SQL pour effectuer des opérations DML sur les données à l’intérieur de l’entrepôt de données et remettre des données à des couches de création de rapports comme Power BI.

Maintenant que vous comprenez les principes architecturaux de base d’un schéma d’entrepôt de données relationnel, nous allons voir comment créer un entrepôt de données.

Décrire un entrepôt de données dans Fabric

Dans l’expérience d’entrepôt de données dans Fabric, vous pouvez créer une couche relationnelle au-dessus des données physiques dans le lakehouse et l’exposer à des outils d’analyse et de création de rapports. Vous pouvez créer votre entrepôt de données directement dans Fabric à partir du hub de création ou dans un espace de travail. Après avoir créé un entrepôt vide, vous pouvez y des objets.

Une fois votre entrepôt créé, vous pouvez créer des tables en utilisant T-SQL directement dans l’interface de Fabric.

Ingérer des données dans votre entrepôt de données

Pour ingérer des données dans un entrepôt de données Fabric, plusieurs méthodes s’offrent à vous. Vous pouvez utiliser des pipelines, des flux de données, l’interrogation entre bases de données ou encore la commande COPY INTO. Après ingestion, les données peuvent être analysées par plusieurs groupes d’entreprise qui peuvent utiliser des fonctionnalités telles que le partage et l’interrogation entre bases de données pour y accéder.

Créer des tables

Pour créer une table dans l’entrepôt de données, vous pouvez utiliser SQL Server Management Studio (SSMS) ou un autre client SQL pour vous connecter à l’entrepôt de données et exécuter une instruction CREATE TABLE. Vous pouvez également créer des tables directement dans l’interface utilisateur de Fabric.

Vous pouvez copier des données à partir d’un emplacement externe dans une table de l’entrepôt de données à l’aide de la syntaxe COPY INTO. Par exemple :

Cette requête SQL charge les données d’un fichier CSV situé dans Stockage Blob Azure dans une table appelée « Region » dans l’entrepôt de données Fabric.

Cloner des tableaux

Vous pouvez créer des clones de table sans duplication avec des coûts de stockage minimes dans un entrepôt de données. Ces clones sont essentiellement des réplicas de tables créés en copiant les métadonnées tout en référençant les mêmes fichiers de données dans OneLake. Cela signifie que les données sous-jacentes stockées sous forme de fichiers Parquet ne sont pas dupliquées, ce qui permet d’économiser des coûts de stockage.

Les clones de table sont particulièrement utiles dans plusieurs scénarios.

Développement et test : Les clones permettent aux développeurs et aux testeurs de créer des copies de tables dans des environnements inférieurs, ce qui facilite les processus de développement, débogage, test et validation.

Récupération des données : En cas d'échec d'une mise en production ou d'une altération des données, les clones de table peuvent conserver l'état précédent des données, ce qui permet la récupération des données.

Rapports historiques : Ils aident à créer des rapports historiques qui reflètent l'état des données à des moments précis et à conserver les données à des étapes clés de l'activité.

Vous pouvez créer un clone de table avec la commande T-SQL CREATE TABLE AS CLONE OF.

Pour en savoir plus sur les clones de table, consultez Tutoriel : Cloner une table avec T-SQL dans Microsoft Fabric.

Considérations relatives aux tables

Au terme de la création de tables dans un entrepôt de données, il est important de prendre en compte le processus de chargement des données dans ces tables. Une approche courante consiste à utiliser des tables de mise en lots. Dans Fabric, vous pouvez utiliser des commandes T-SQL pour charger des données à partir de fichiers dans des tables de mise en lots dans l'entrepôt de données.

Les tables de mise en lots sont des tables temporaires qui peuvent être utilisées pour nettoyer, transformer et valider des données. Vous pouvez également utiliser des tables de mise en lots pour charger des données de plusieurs sources dans une table de destination unique.

En général, les données sont chargées dans le cadre d'un processus de traitement par lots périodique dans lequel les insertions et mises à jour de l'entrepôt de données sont coordonnées pour se produire à un intervalle régulier (par exemple quotidien, hebdomadaire ou mensuel).

Dans la plupart des cas, vous devez implémenter un processus de chargement d'entrepôt de données, qui effectue les tâches dans l'ordre suivant :

Ingérez les nouvelles données à charger dans un lac de données, en appliquant un nettoyage ou des transformations avant le chargement, selon les besoins.

Chargez les données à partir de fichiers dans des tables de mise en lots au sein de l'entrepôt de données relationnel.

Chargez les tables de dimension à partir des données de dimension dans les tables de mise en lots, en mettant à jour les lignes existantes ou en insérant de nouvelles lignes, et en générant des valeurs de clé de substitution le cas échéant.

Chargez les tables de faits à partir des données de faits dans les tables de mise en lots, en recherchant les clés de substitution appropriées pour les dimensions associées.

Effectuez une optimisation postchargement en mettant à jour les index et les statistiques de distribution des tables.

Si vous avez des tables dans le lac et que vous voulez pouvoir les interroger dans votre entrepôt - sans les modifier – avec un entrepôt de données Fabric, vous n'avez pas besoin de copier les données du lac vers l'entrepôt de données. Vous pouvez interroger les données dans le lakehouse directement à partir de l'entrepôt de données en utilisant l'interrogation entre bases de données.

L'utilisation de tables dans l'entrepôt de données Fabric présente actuellement certaines limitations. Pour plus d'informations, consultez Tables dans l'entreposage de données dans Microsoft Fabric.

Unité 4: Interroger et transformer des données

Type: Contenu

Interroger et transformer des données

Maintenant que vous savez comment implémenter un entrepôt de données dans Fabric, préparons les données pour l'analytique.

Il existe deux façons d'interroger des données à partir de votre entrepôt de données. L'éditeur de requête visual fournit une expérience sans code, glisser-déplacer pour créer vos requêtes. Si vous êtes à l'aise avec T-SQL, vous préférez peut-être utiliser l'éditeur de requête SQL pour écrire vos requêtes. Dans les deux cas, vous pouvez créer des tables, des vues et des procédures stockées pour interroger des données dans l'entrepôt de données et Lakehouse.

Il existe également un point de terminaison d'analytique SQL, où vous pouvez vous connecter à partir de n'importe quel outil.

Interroger des données à l'aide de l'éditeur de requête SQL

L'éditeur de requête SQL fournit une expérience de requête qui inclut IntelliSense, la saisie semi-automatique du code, la mise en surbrillance de la syntaxe, l'analyse côté client et la validation. Si vous avez écrit T-SQL dans SQL Server Management Studio (SSMS) ou Azure Data Studio (ADS), vous le trouverez familier.

Pour créer une requête, utilisez le bouton Nouvelle requête SQL dans le menu. Vous pouvez créer et exécuter vos requêtes T-SQL ici. Dans l'exemple ci-dessous, nous créons un nouvel affichage pour les analystes à utiliser pour la création de rapports dans Power BI.

Interroger des données à l'aide de l'éditeur de requête Visual

L'éditeur de requête visuelle offre une expérience similaire à la vue de diagramme en ligne Power Query. Utilisez le bouton Nouvelle requête visuelle pour créer une requête.

Faites glisser une table de votre entrepôt de données vers le canevas pour commencer. Vous pouvez ensuite utiliser le menu Transformer en haut de l'écran pour des colonnes, des filtres et d'autres transformations à votre requête. Vous pouvez utiliser le bouton (+) sur le visuel lui-même pour effectuer des transformations similaires.

Unité 5: Préparer des données pour l'analyse et la création de rapports

Type: Contenu

Préparer des données pour l'analyse et la création de rapports

Un modèle de données sémantique définit les relations entre les différentes tables du modèle sémantique, les règles d'agrégation et de synthèse des données ainsi que les calculs ou mesures utilisés pour générer des insights à partir des données. Ces relations et mesures sont incluses dans le modèle sémantique, qui est ensuite exploité pour créer des rapports dans Power BI.

Vous pouvez facilement basculer entre les vues Données, Requête et Modèle de Fabric à l'aide du menu situé en bas à gauche de l'écran. La vue Données présente les tables du modèle sémantique, la vue Requête les requêtes SQL utilisées pour créer le modèle sémantique et la vue Modèle le modèle sémantique.

Pour en savoir plus sur les modèles de données et le schéma de l'entrepôt de données, consultez Analyser les données dans un entrepôt de données relationnel.

Créer des relations

Les relations vous permettent de connecter les tables dans le modèle sémantique. Créez des relations entre les tables de votre entrepôt de données en utilisant l'interface « cliquer-glisser » dans la vue Modèle de Fabric.

Pour plus d'informations sur la création de relations, consultez Créer et gérer des relations.

Créer des mesures

Les mesures sont les métriques que vous souhaitez analyser dans votre entrepôt de données. Vous pouvez créer des mesures dans Fabric à l'aide du bouton Nouvelle mesure dans la vue Modèle.

Les mesures sont des champs calculés qui sont basés sur les données des tables de votre entrepôt de données et écrits en langage de formule DAX (Data Analysis Expressions).

Fabric propose de nombreux outils pour créer des transformations de données. La création de mesures avec DAX est l'une des nombreuses façons de créer des transformations de données. Pour en savoir plus sur DAX, consultez Utiliser DAX dans Power BI.

Masquer les champs

La création du modèle sémantique est une phase critique de la préparation de vos données en vue de leur utilisation pour créer des rapports en aval. Pour simplifier les choses pour vos générateurs de rapports, vous pouvez masquer des éléments, comme une table ou une colonne, de la vue. Cliquez avec le bouton droit sur la table ou la colonne, puis sélectionnez Masquer. Le masquage de champs supprime la table ou la colonne de la vue du modèle, mais elle est toujours disponible et exploitable dans le modèle sémantique.

Comprendre les modèles sémantiques

Chaque fois qu'un entrepôt de données est créé, Fabric crée un modèle sémantique auquel les analystes et/ou les utilisateurs professionnels peuvent se connecter pour générer des rapports.

Le modèle sémantique comprend des métriques utilisées pour créer des rapports. Dit simplement, les analystes utilisent le modèle sémantique que vous avez créé dans votre entrepôt et qui est stocké dans un modèle sémantique. Si vous connaissez Power BI, l'utilisation de modèles sémantiques créés par l'expérience de l'entrepôt de données ne vous posera aucune difficulté.

Les modèles sémantiques étant automatiquement synchronisés avec l'entrepôt de données, vous n'avez pas à vous soucier de leur maintenance. Vous pouvez également créer des modèles sémantiques personnalisés pour répondre à vos besoins spécifiques.

Comprendre le modèle sémantique par défaut

Un modèle sémantique par défaut est aussi créé automatiquement pour vous dans Fabric. Il hérite de la logique métier du lakehouse ou de l'entrepôt parent, qui lance l'expérience d'analytique en aval pour l'analyse et le décisionnel. Ce modèle sémantique est géré, optimisé et synchronisé pour vous.

Les nouvelles tables du lakehouse sont automatiquement ajoutées au modèle sémantique par défaut. Pour plus de flexibilité, les utilisateurs peuvent aussi sélectionner manuellement les tables ou vues de l'entrepôt qu'ils souhaitent inclure dans le modèle. Les objets qui se trouvent dans le modèle sémantique par défaut sont créés en tant que disposition dans la vue du modèle.

Les modèles sémantiques par défaut suivent les limitations actuelles des modèles sémantiques dans Power BI. Pour plus d'informations, consultez Modèles sémantiques Power BI par défaut.

Visualiser les données

Fabric vous permet de visualiser les résultats d'une requête unique ou de l'ensemble de votre entrepôt de données sans quitter l'expérience d'entrepôt de données. L'exploration des données pendant que vous travaillez pour vérifier que vous disposez de toutes les données et transformations nécessaires à

votre analyse est particulièrement utile.

Utilisez le bouton Nouveau rapport pour créer un rapport Power BI basé sur le contenu de l'ensemble de votre entrepôt de données. Le bouton Nouveau rapport ouvre l'expérience de service Power BI où vous pouvez créer et enregistrer votre rapport en vue d'une utilisation par l'entreprise.

Unité 6: Sécuriser et surveiller votre entrepôt de données

Type: Contenu

Sécuriser et surveiller votre entrepôt de données

La sécurité et la surveillance sont des aspects essentiels de la gestion de votre entrepôt de données.

La sécurité de l'entrepôt de données est importante pour protéger vos données contre l'accès non autorisé. Fabric fournit un certain nombre de fonctionnalités de sécurité pour vous aider à sécuriser votre entrepôt de données. Voici quelques-uns des éléments suivants :

Contrôle d'accès en fonction du rôle (RBAC) pour contrôler l'accès à l'entrepôt et à ses données.

Chiffrement TLS pour sécuriser la communication entre l'entrepôt et les applications clientes.

Azure Storage Service Encryption pour protéger les données en transit et au repos.

Azure Monitor et Azure Log Analytics pour surveiller l'activité de l'entrepôt et auditer l'accès aux données.

Authentification multifacteur (MFA) pour une couche supplémentaire de sécurité aux comptes d'utilisateur.

Intégration d'ID Microsoft Entra pour gérer les identités utilisateur et l'accès à l'entrepôt.

Autorisations d'espace de travail

Les données dans Fabric sont organisées en espaces de travail, qui sont utilisés pour contrôler l'accès et gérer le cycle de vie des données et des services. Les rôles d'espace de travail appropriés sont la première ligne de défense dans la sécurisation de votre entrepôt de données.

Outre les rôles d'espace de travail, vous pouvez accorder des autorisations d'élément et un accès via SQL.

Conseil / Astuce

Pour plus d'informations sur les rôles d'espace de travail , consultez Les espaces de travail dans Power BI .

Autorisations d'élément

Contrairement aux rôles d'espace de travail, qui s'appliquent à tous les éléments d'un espace de travail, vous pouvez utiliser des autorisations d'élément pour accorder l'accès à des entrepôts individuels. Cela vous permet de partager un entrepôt de données unique pour la consommation en aval.

Vous pouvez accorder des autorisations aux utilisateurs via T-SQL ou dans le portail Fabric. Accordez les autorisations suivantes aux utilisateurs qui doivent accéder à votre entrepôt de données :

Lecture : permet à l'utilisateur de se connecter à l'aide de la chaîne de connexion SQL.

ReadData : permet à l’utilisateur de lire des données à partir de n’importe quelle table/vue dans l’entrepôt.

ReadAll : permet à l’utilisateur de lire les données des fichiers parquet bruts dans OneLake qui peuvent être consommés par Spark.

Une connexion utilisateur au point de terminaison d'analytique SQL va échouer si la permission de lecture, au minimum, n'est pas accordée.

La surveillance des activités dans votre entrepôt de données est essentielle pour garantir des performances optimales, une utilisation efficace des ressources et une sécurité. Il vous aide à identifier les problèmes, à détecter les anomalies et à prendre des mesures pour que l’entrepôt de données s’exécute de manière fluide et sécurisée.

Vous pouvez utiliser des vues de gestion dynamique (DMV) pour surveiller la connexion, la session et l’état de la demande pour afficher les insights de cycle de vie des requêtes SQL en direct. Les DMV vous permettent d’obtenir des détails tels que le nombre de requêtes actives et d’identifier les requêtes qui s’exécutent pendant une longue période et qui doivent être arrêtées.

Il existe actuellement trois DMV disponibles pour l’utilisation dans Fabric :

sys.dm_exec_connections : retourne des informations sur chaque connexion établie entre l’entrepôt et le moteur.

sys.dm_exec_sessions : retourne des informations sur chaque session authentifiée entre l’élément et le moteur.

sys.dm_exec_requests : retourne des informations sur chaque requête active dans une session.

Analyse des requêtes

Utilisez « sys.dm_exec_requests » pour identifier les requêtes longues qui peuvent avoir un impact sur les performances globales de la base de données et prendre les mesures appropriées pour optimiser ou arrêter ces requêtes.

Commencez par identifier les requêtes qui s’exécutent depuis longtemps. Utilisez la requête suivante pour identifier les requêtes qui ont été exécutées le plus longtemps, dans l’ordre décroissant :

Vous pouvez continuer à examiner pour comprendre l’utilisateur qui a exécuté la session avec la requête de longue durée, en exécutant :

Enfin, vous pouvez utiliser la KILL commande pour terminer la session avec la requête longue :

Vous devez être administrateur de l’espace de travail pour exécuter la KILL commande. Les administrateurs d’espace de travail peuvent exécuter les trois DMV. Les rôles Membre, Contributeur et Visionneuse peuvent voir leurs propres résultats dans l’entrepôt, mais ne peuvent pas voir les résultats d’autres utilisateurs.

Unité 7: Exercice - Analyser des données dans un entrepôt de données

Type: Exercice

Exercice - Analyser des données dans un entrepôt de données

Vous devez maintenant créer un entrepôt de données dans Fabric et analyser vos données.

Vous avez besoin d’une licence d’évaluation Microsoft Fabric avec la version préliminaire Fabric activée dans votre compte client. Consultez Bien démarrer avec Fabric pour activer votre licence d’évaluation Fabric.

Lancez l’exercice et suivez les instructions.

Exercice - Analyser des données dans un entrepôt de données Vous devez maintenant créer un entrepôt de données dans Fabric et analyser vos données. Remarque Vous avez besoin d’une licence d’évaluation Microsoft Fabric avec la version préliminaire Fabric activée dans votre compte client. Consultez Bien démarrer avec Fabric pour activer votre licence d’évaluation Fabric. Lancez l’exercice et suivez les instructions. Commentaires Yes No

Unité 8: Évaluation du module

Type: Évaluation

Évaluation du module

Vérifier vos connaissances

Quel type de tableau une compagnie d’assurance doit-elle utiliser pour stocker les détails de l’attribut fournisseur pour l’agrégation des réclamations ?

Table de dimension.

Table intermédiaire.

Qu’est-ce qu’un modèle sémantique dans l’expérience de l’entrepôt de données ?

Un modèle sémantique est un modèle de données orienté entreprise qui fournit une représentation cohérente et réutilisable des données au sein de l’organisation.

Un modèle sémantique est un modèle de données physique qui décrit la structure des données stockées dans l’entrepôt de données.

Un modèle sémantique est un modèle Machine Learning utilisé pour effectuer des prédictions basées sur les données de l’entrepôt de données.

Quel est l’objectif des autorisations d’élément dans un espace de travail ?

Pour accorder l’accès à tous les éléments d’un espace de travail.

Pour accorder l’accès à des colonnes spécifiques dans une table.

Pour accorder l’accès à des entrepôts individuels pour la consommation en aval.

Vous devez répondre à toutes les questions avant de vérifier votre travail.

Unité 9: Résumé

Type: Résumé

Dans ce module, vous avez découvert les entrepôts de données et la modélisation dimensionnelle, créé un entrepôt de données, chargé, interrogé et visualisé des données, et décrit des modèles sémantiques et comment ils sont utilisés pour la création de rapports en aval.

Résumé Dans ce module, vous avez découvert les entrepôts de données et la modélisation dimensionnelle, créé un entrepôt de données, chargé, interrogé et visualisé des données, et décrit des modèles sémantiques et comment ils sont utilisés pour la création de rapports en aval. Commentaires Yes No

Module 8: Bien démarrer avec Real-Time Intelligence dans Microsoft Fabric

Unité 1: Présentation

Type: Introduction

La plupart des solutions d'analytique données modernes permettent deux modèles courants pour l'analyse des données :

Analytique des données batch, dans laquelle les données sont chargées dans un magasin de données analytique à intervalles réguliers en tant qu'opération de traitement par lots ; activation de l'analyse historique des données provenant d'événements passés.

Analytique des données en temps réel, dans laquelle les données d'événements sont ingérées en temps réel (ou en quasi temps réel) lorsque des événements se produisent dans un flux de données qui peut être analysé, visualisé et utilisé pour déclencher des réponses automatisées.

L'analytique données par lots est généralement bien comprise et est couramment mise en œuvre à l'aide d'architectures de type entrepôt de données ou lakehouse. L'analytique en temps réel peut être considérée comme plus spécialisée, mais de plus en plus elle est incorporée dans des solutions d'analyse de données à grande échelle sous la forme d'une architecture lambda qui combine le chargement périodique de données par lots pour l'analyse historique avec l'ingestion de flux de données pour l'analyse en temps réel.

Microsoft Fabric offre des fonctionnalités d'analytique par lots et en temps réel. Dans ce module, nous allons nous concentrer sur les fonctionnalités Real-Time Intelligence de Microsoft Fabric pour explorer comment vous pouvez créer des solutions d'analyse de données en temps réel avec un codage minimal qui s'adapte à d'énormes volumes de données provenant d'un large éventail de sources.

Les rubriques abordées dans ce module incluent :

Comprendre les concepts fondamentaux liés à l'analytique données en temps réel.

Comprendre les fonctionnalités d'intelligence en temps réel de Microsoft Fabric.

Exploration des principaux composants de Real-Time Intelligence dans Microsoft Fabric.

Ingestion de données en temps réel à l'aide d'un flux d'événements.

Utilisation d'un eventhouse et d'une base de données KQL pour l'analyse des données en temps réel dans Microsoft Fabric.

Visualisation des données dans des tableaux de bord en temps réel.

Utilisation de l'activateur dans Microsoft Fabric pour définir des alertes qui déclenchent des actions automatisées.

À la fin de ce module, vous serez en mesure de comprendre les capacités des fonctionnalités d'intelligence en temps réel de Microsoft Fabric. Vous acquérez également une expérience pratique via un exercice pratique.

Unité 2: Qu'est-ce que l'analytique données en temps réel ?

Type: Contenu

Qu’est-ce que l’analytique données en temps réel ?

L’analytique des données en temps réel est généralement basée sur l’ingestion et le traitement d’un flux de données qui se compose d’une série perpétuelle de données, généralement liée à des événements à un point dans le temps spécifiques. Par exemple, un flux de données peut contenir des détails de messages envoyés à un site de microblogs de réseau social, ou une série de mesures environnementales enregistrées par un capteur météo connecté à Internet.

Les données du flux peuvent être utilisées pour créer des visualisations en temps réel des données à des fins de surveillance ou pour déclencher des actions automatisées si certaines conditions se produisent. Par exemple, un flux de données provenant d’un capteur de contrôle environnemental dans un bâtiment de bureau peut permettre à des systèmes de chauffage et de climatisation d’être contrôlés dynamiquement pour optimiser le confort et le coût. Les données peuvent également être conservées dans un magasin de données et interrogées ultérieurement, ce qui permet aux analystes de mieux comprendre les changements au fil du temps. Par exemple, une organisation marketing peut effectuer une analyse des sentiments sur des messages de réseau social pour voir si une campagne publicitaire génère des commentaires plus positifs sur l’entreprise ou ses produits, ou une entreprise agricole peut monitorer les tendances de température et de pluie pour optimiser l’irrigation et les récoltes.

Les objectifs courants pour l’analytique en temps réel incluent

Analyse continue des données pour signaler des problèmes ou des tendances.

Compréhension du comportement des composants ou du système dans différentes conditions pour planifier des améliorations futures.

Déclenchement d’actions ou d’alertes spécifiques quand certains événements se produisent ou que des seuils sont dépassés.

Caractéristiques des solutions d’analytique des données en temps réel

Les solutions de traitement de flux pour l’analytique des données en temps réel présentent généralement les caractéristiques suivantes :

Un flux de données n’est pas lié : les données sont ajoutées au flux de façon perpétuelle.

Les enregistrements de données du flux incluent généralement des données temporelles (basées sur le temps) indiquant quand l’événement auquel l’enregistrement se rapporte (ou a été enregistré).

L’agrégation des données de diffusion en continu est souvent e sur des fenêtres temporelles , par exemple, l’enregistrement du nombre de publications de médias sociaux par minute ou la pluviosité moyenne par heure.

Les résultats du traitement des données de streaming peuvent être utilisés pour prendre en charge l’automatisation ou la visualisation en temps réel (ou quasiment en temps réel), ou conservés dans un magasin analytique à combiner avec d’autres données pour l’analyse historique. De nombreuses solutions combinent ces approches pour prendre en charge l’analytique en temps réel et historique.

Les fonctionnalités d’intelligence en temps réel de Microsoft Fabric vous permettent d’implémenter des solutions d’analytique en temps réel qui incluent les fonctionnalités décrites ici avec un effort de codage minimal (ou non) et une intégration dans le reste de l’écosystème Microsoft Fabric.

Unité 3: Informations en temps réel dans Microsoft Fabric

Type: Contenu

Informations en temps réel dans Microsoft Fabric

La solution Real-Time Intelligence de Microsoft Fabric fournit une solution de streaming de bout en bout pour l'analyse de données en temps réel à l'échelle du service Fabric.

Real-Time Intelligence assure des performances élevées pour les données dont la taille peut aller de quelques gigaoctets à plusieurs pétaoctets. Il peut gérer des données de différentes sources et de formats divers. La charge de travail Real-Time Intelligence de Fabric peut être utilisée pour des solutions d'IoT et d'analytique des journaux d'activité dans de nombreux secteurs d'activité : fabrication, pétrole et gaz, automobile, etc.

À l'aide de Microsoft Fabric Real-Time Intelligence, vous pouvez :

Créez un flux d'événements pour capturer, transformer et ingérer des données en temps réel à partir de différentes sources de streaming.

Stockez les données en temps réel capturées dans un eventhouse, qui inclut une ou plusieurs bases de données KQL.

Interrogez et analysez des données dans la maison d'événements à l'aide de requêtes KQL, organisées dans un ensemble de requêtes KQL.

Visualisez les données en temps réel dans un tableau de bord en temps réel ou à l'aide de Power BI.

Configurez des alertes qui utilisent Activateor pour déclencher des actions automatisées.

Hub en temps réel de Microsoft Fabric

Le hub en temps réel Microsoft Fabric fournit un emplacement centralisé pour la gestion des sources de données en temps réel.

Pour afficher le hub en temps réel, sélectionnez l'icône Temps réel dans la barre de menus principale de Fabric.

Dans le hub en temps réel, vous pouvez :

Recherchez et connectez-vous à des sources de données en temps réel et créez des flux d'événements.

Abonnez-vous aux événements Fabric et Azure, puis créez des flux d'événements et des alertes d'activateur .

Affichez un aperçu et gérez vos connexions de données en temps réel, notamment la navigation vers les données de flux capturées dans un eventhouse.

Créez des tableaux de bord en temps réel à partir de flux d'événements.

Approuvez et partagez des ressources de données en temps réel au sein de votre organisation.

Unité 4: Stocker et interroger des données en temps réel

Type: Contenu

Stocker et interroger des données en temps réel

Les eventhouses sont des lieux où vous stockez des données en temps réel, souvent ingérées par un flux d'événements et chargés dans des tables pour un traitement et une analyse supplémentaires.

Dans un eventhouse, vous pouvez créer :

Bases de données KQL : magasins de données optimisés en temps réel qui hébergent une collection de tables, de fonctions stockées, de vues matérialisées et de raccourcis.

Ensembles de requêtes KQL : collections de requêtes KQL que vous pouvez utiliser pour utiliser des données dans des tables de base de données KQL. Un jeu de requêtes KQL prend en charge les requêtes écrites à l'aide du langage KQL (Kusto Query Language) et d'un sous-ensemble du langage Transact-SQL.

Interroger les données

Pour interroger des données à partir d'une table dans une base de données KQL, vous pouvez utiliser le langage de requête Kusto (KQL) utilisé pour écrire des requêtes dans Azure Data Explorer, Azure Monitor Log Analytics, Microsoft Sentinel et Microsoft Fabric. Une requête KQL est une demande en lecture seule de traitement de données et de retour de résultats. Les requêtes KQL sont constituées d'une ou plusieurs instructions de requête.

Instructions de requête KQL

Une instruction de requête se compose d'un nom de table suivi d'un ou plusieurs opérateurs qui agissent sur les données (take, filter, transform, aggregate ou join). Par exemple, la requête suivante récupère 10 lignes d'une table nommée stock :

Un exemple plus complexe peut agréger les données pour trouver le prix moyen des actions par symbole boursier au cours des cinq dernières minutes :

Pour en savoir plus sur KQL, consultez la vue d'ensemble du langage de requête Kusto (KQL).

Utilisation de SQL

KQL est optimisé pour interroger de grands volumes de données, en particulier avec un élément basé sur le temps ; il convient donc parfaitement à l'analyse des données en temps réel. Toutefois, de nombreux professionnels des données connaissent déjà la syntaxe SQL. Ainsi, les bases de données KQL dans les eventhouses prennent en charge un sous-ensemble d'expressions SQL courantes.

Par exemple, la requête SQL équivalente à la requête 10 KQL décrite précédemment serait :

Utilisation de Copilot comme aide avec les requêtes

Microsoft Fabric inclut Copilo pour l'intelligence en temps réel, qui peut vous aider à écrire les requêtes dont vous avez besoin pour extraire des insights de vos données eventhouse. Copilot utilise l'IA pour comprendre les informations que vous recherchez et peut générer le code de requête requis pour vous.

Pour en savoir plus sur Copilot pour Real-Time Intelligence, consultez Copilot pour Real-Time Intelligence.

Unité 5: Exercice – Explorer l'intelligence en temps réel dans Fabric

Type: Exercice

Exercice – Explorer l'intelligence en temps réel dans Fabric

Il est maintenant temps d'essayer l'intelligence en temps réel de Fabric par vous-même.

Dans cet exercice, vous allez ingérer des données en temps réel dans Microsoft Fabric. Vous allez ensuite interroger et visualiser les données avant de définir une alerte pour automatiser une action en

fonction d’une valeur seuil dans le flux de données en temps réel.

Vous devez disposer d'un client Microsoft Fabric pour réaliser cet exercice. Pour plus d’informations sur l’accès à Microsoft Fabric, consultez Prise en main de Fabric.

Lancez l’exercice et suivez les instructions.

Unité 6: Évaluation du module

Type: Évaluation

Évaluation du module

Quel composant Microsoft Fabric Real-Time Intelligence devez-vous utiliser pour ingérer et transformer un flux de données en temps réel ?

Flux d’événements

Quel langage est optimisé pour interroger des données en temps réel dans un eventhouse ?

Quel composant Microsoft Fabric Real-Time Intelligence est utilisé pour visualiser et explorer les données en temps réel dans les vignettes ?

Tableaux de bord en temps réel

Vous devez répondre à toutes les questions avant de vérifier votre travail.

Unité 7: Résumé

Type: Résumé

Dans ce module, vous avez découvert Real-Time Intelligence de Microsoft Fabric. Le module a présenté les composants de base de Real-Time Intelligence et fourni la possibilité de le tester par vous-même.

Pour découvrir plus d’informations sur Real-Time Intelligence dans Microsoft Fabric, consultez la documentation Real-Time Intelligence dans Microsoft Fabric.

Résumé Dans ce module, vous avez découvert Real-Time Intelligence de Microsoft Fabric. Le module a présenté les composants de base de Real-Time Intelligence et fourni la possibilité de le tester par vous-même. Conseil Pour découvrir plus d’informations sur Real-Time Intelligence dans Microsoft Fabric, consultez la documentation Real-Time Intelligence dans Microsoft Fabric. Commentaires Yes No

Module 9: Bien démarrer avec la science des données dans Microsoft Fabric

Unité 1: Présentation

Type: Introduction

Imaginez que vous travaillez pour un supermarché et que vous voulez savoir combien de pains vous devez avoir en stock chaque semaine pour répondre aux demandes des clients tout en évitant le gaspillage alimentaire.

Ou peut-être souhaitez-vous analyser vos clients pour comprendre comment les cibler au mieux avec des offres personnalisées.

Chaque fois que vous souhaitez prendre des décisions éclairées au sein d'une organisation, vous pouvez utiliser la science des données pour obtenir des informations à partir des données dont vous disposez. La science des données est une combinaison de mathématiques, de statistiques et d'ingénierie informatique.

Lorsque vous effectuez une science des données, vous pouvez analyser vos données et identifier des modèles complexes qui peuvent vous fournir des insights significatifs pour votre organisation. Vous pouvez utiliser la science des données pour créer des modèles d'intelligence artificielle (IA) qui englobent les modèles complexes que vous trouvez dans vos données. Une approche courante consiste à utiliser la science des données pour entraîner des modèles Machine Learning à l'aide de bibliothèques comme scikit-learn dans Python pour obtenir l'IA.

Il peut être laborieux de s'occuper d'un projet de science des données du début à la fin. Microsoft Fabric offre un espace de travail pour gérer un projet de science des données de bout en bout.

Dans ce module, vous allez découvrir un projet de science des données classique. En outre, vous découvrirez les fonctionnalités de Microsoft Fabric que vous pouvez utiliser pour chaque partie du processus de science des données.

Unité 2: Comprendre le processus de science des données

Type: Contenu

Comprendre le processus de science des données

Une façon courante d'extraire des insights à partir de données consiste à visualiser les données. Chaque fois que vous avez des jeux de données complexes, vous pouvez approfondir et essayer de trouver des modèles complexes dans les données.

En tant que scientifique des données, vous pouvez entraîner des modèles de Machine Learning pour trouver des modèles dans vos données. Vous pouvez utiliser ces modèles pour générer de nouveaux insights ou de nouvelles prédictions. Par exemple, vous pouvez prédire le nombre attendu de produits que vous prévoyez de vendre au cours de la prochaine semaine.

Bien que l'apprentissage du modèle soit important, ce n'est pas la seule tâche d'un projet de science des données. Avant d'explorer le processus de science des données classique, nous allons explorer les modèles de Machine Learning courants que vous pouvez entraîner.

Explorer les modèles Machine Learning courants

L’objectif du Machine Learning est d’entraîner des modèles capables d’identifier des modèles dans de grandes quantités de données. Vous pouvez ensuite utiliser les modèles pour effectuer des prédictions qui vous fournissent de nouveaux insights pour lesquels vous pouvez prendre des mesures.

Les possibilités du Machine Learning peuvent sembler infinies. Commençons donc par comprendre les quatre types courants de modèles Machine Learning :

Classification : prédire une valeur catégorielle, indiquant, par exemple, s’il a un risque de perdre un client.

Régression : prédire une valeur numérique comme le prix d’un produit.

Clustering : regroupez des points de données similaires dans des clusters ou des groupes.

Prévision : prédire des valeurs numériques futures basées sur des données de séries chronologiques, comme les ventes attendues pour le mois à venir.

Pour déterminer le type de modèle Machine Learning que vous devez entraîner, vous devez d’abord comprendre le problème métier et les données disponibles.

Pour entraîner un modèle Machine Learning, le processus implique généralement les étapes suivantes :

Définir le problème : avec les utilisateurs métier et les analystes, déterminez ce que le modèle doit prédire et quand il réussit.

Obtenir les données : recherchez des sources de données et obtenez l’accès en stockant vos données dans un lakehouse.

Préparer les données : explorez les données en les lisant à partir d’un lakehouse dans un notebook. Nettoyez et transformez les données en fonction des exigences du modèle.

Entraîner le modèle : choisissez un algorithme et des valeurs d’hyperparamètres par tâtonnements en suivant vos expériences avec MLflow.

Générer des insights : utilisez le scoring par lots du modèle pour générer les prédictions demandées.

En tant que scientifique des données, vous consacrez la plupart de votre temps à la préparation des données et à l’entraînement du modèle. La façon dont vous préparez les données et l’algorithme que vous choisissez pour entraîner un modèle peuvent influencer la réussite de votre modèle.

Vous pouvez préparer et entraîner un modèle à l’aide de bibliothèques open source disponibles pour le langage de votre choix. Par exemple, si vous utilisez Python, vous pouvez préparer les données avec Pandas et Numpy, et entraîner un modèle avec des bibliothèques telles que Scikit-Learn, PyTorch ou SynapseML.

Lors de l’expérimentation, vous souhaitez conserver une vue d’ensemble de tous les différents modèles que vous avez entraînés. Vous souhaitez comprendre comment vos choix influencent la réussite du modèle. En suivant vos expériences avec MLflow dans Microsoft Fabric, vous pouvez facilement gérer et déployer les modèles que vous avez entraînés.

Unité 3: Explorer et traiter des données avec Microsoft Fabric

Type: Contenu

Explorer et traiter des données avec Microsoft Fabric

Les données sont la pierre angulaire de la science des données, en particulier lorsqu'il s'agit d'entraîner un modèle de Machine Learning pour atteindre l'intelligence artificielle. En règle générale, les modèles présentent des performances améliorées à mesure que la taille du jeu de données d'entraînement augmente. Outre la quantité de données, la qualité de celle-ci est tout aussi importante.

Pour garantir la qualité et la quantité de vos données, l'utilisation de moteurs robustes d'ingestion et de traitement des données de Microsoft Fabric est très utile. Vous avez la possibilité d'opter pour une approche à faible code ou orientée code lors de l'établissement des pipelines essentiels d'ingestion, d'exploration et de transformation de données.

Ingérer vos données dans Microsoft Fabric

Pour utiliser des données dans Microsoft Fabric, vous devez d'abord ingérer des données. Vous pouvez ingérer des données provenant de plusieurs sources, locales et cloud à la fois. Par exemple, vous pouvez ingérer des données à partir d'un fichier CSV stocké sur votre ordinateur local ou dans un Azure Data Lake Storage (Gen2).

Découvrez-en davantage sur la façon d'ingérer et d'orchestrer des données à partir de différentes sources avec Microsoft Fabric.

Après vous être connecté à une source de données, vous pouvez enregistrer les données dans un lakehouse Microsoft Fabric. Vous pouvez utiliser le lakehouse comme emplacement central pour stocker tous les fichiers structurés, semi-structurés et non structurés. Vous pouvez ensuite vous connecter facilement au lakehouse chaque fois que vous souhaitez accéder à vos données à des fins d'exploration ou de transformation.

Explorer et transformer vos données

En tant que scientifique des données, vous êtes peut-être plus à l'aise avec l'écriture et l'exécution de code dans les notebooks. Microsoft Fabric offre une expérience de notebook familière, optimisée par le calcul Spark.

Apache Spark est un framework de traitement parallèle open source pour le traitement et l'analytique à grande échelle des données.

Les notebooks sont automatiquement attachés à un calcul Spark. Lorsque vous exécutez une cellule dans un notebook pour la première fois, une nouvelle session Spark démarre. La session persiste lorsque vous exécutez les cellules suivantes. La session Spark s'arrête automatiquement après un certain temps d'inactivité pour réduire les coûts. Vous pouvez également arrêter manuellement la session.

Lorsque vous travaillez dans un notebook, vous pouvez choisir le langage que vous souhaitez utiliser. Pour les charges de travail de science des données, vous êtes susceptible d'utiliser PySpark (Python) ou SparkR (R).

Dans le notebook, vous pouvez explorer vos données à l'aide de votre bibliothèque préférée ou avec l'une des options de visualisation intégrées. Si nécessaire, vous pouvez transformer vos données et enregistrer les données traitées en les réécrivant dans le lakehouse.

Préparer vos données avec Data Wrangler

Pour vous aider à explorer et à transformer vos données plus rapidement, Microsoft Fabric propose Data Wrangler, un outil facile à utiliser.

Après avoir lancé Data Wrangler, vous obtiendrez une vue d'ensemble descriptive des données que vous utilisez. Vous pouvez afficher les statistiques récapitulatives de vos données pour trouver des problèmes tels que des valeurs manquantes.

Pour nettoyer vos données, vous pouvez choisir l'une des opérations de nettoyage de données intégrées. Lorsque vous sélectionnez une opération, un aperçu du résultat et du code associé est généré automatiquement pour vous. Une fois que vous avez sélectionné toutes les opérations nécessaires, vous pouvez exporter les transformations vers du code et les exécuter sur vos données.

Unité 4: Entraîner et évaluer des modèles avec Microsoft Fabric

Type: Contenu

Entraîner et évaluer des modèles avec Microsoft Fabric

Une fois que vous avez ingéré, exploré et prétraité vos données, vous pouvez les utiliser pour entraîner un modèle. L'entraînement d'un modèle est un processus itératif, vous devez pouvoir suivre votre travail.

Microsoft Fabric s'intègre à MLflow pour suivre et journaliser facilement votre travail, ce qui vous permet de l'examiner à tout moment afin de décider de la meilleure approche pour entraîner le modèle final. Lorsque vous suivez votre travail, vos résultats sont facilement reproductibles.

Tout travail que vous souhaitez suivre peut être suivi en tant qu'expériences.

Comprendre les expériences

Chaque fois que vous entraînez un modèle dans un notebook que vous voulez suivre, vous créez une expérience dans Microsoft Fabric.

Une expérience peut comporter plusieurs exécutions. Chaque exécution représente une tâche que vous avez exécutée dans un notebook, comme l'entraînement d'un modèle Machine Learning.

Par exemple, pour entraîner un modèle Machine Learning pour la prévision des ventes, vous pouvez essayer différents jeux de données d'entraînement avec le même algorithme. Chaque fois que vous entraînez un modèle avec un jeu de données différent, vous créez une nouvelle exécution d'expérience. Ensuite, vous pouvez comparer les exécutions d'expérience pour déterminer le modèle le plus performant.

Commencer à suivre des métriques

Pour comparer des exécutions d'expérience, vous pouvez suivre les paramètres, les métriques et les artefacts pour chaque exécution.

L'ensemble des paramètres, métriques et artefacts que vous suivez dans une exécution d'expérience sont affichés dans la vue d'ensemble des expériences. Vous pouvez afficher les exécutions d'expérience individuellement sous l'onglet Détails de l'exécution ou comparer les exécutions à l'aide de la liste d'exécutions :

En suivant votre travail avec MLflow, vous pouvez comparer les itérations d'entraînement du modèle et déterminer la configuration qui a donné lieu au modèle le plus adapté à votre cas d'usage.

Comprendre les modèles

Après avoir entraîné un modèle, vous souhaitez l'utiliser pour le scoring. Avec le scoring, vous utilisez le modèle sur de nouvelles données pour générer des prédictions ou des aperçus. Lorsque vous entraînez et suivez un modèle avec MLflow, les artefacts sont stockés dans l'exécution de l'expérience pour représenter votre modèle et ses métadonnées. Vous pouvez enregistrer ces artefacts dans Microsoft Fabric en tant que modèle.

En enregistrant vos artefacts de modèle comme modèle inscrit dans Microsoft Fabric, vous pouvez facilement gérer vos modèles. Chaque fois que vous entraînez un nouveau modèle et que vous l'enregistrez sous le même nom, vous ajoutez une nouvelle version au modèle.

Utiliser un modèle pour générer des aperçus

Pour utiliser un modèle pour générer des prédictions, vous pouvez utiliser la fonction PREDICT dans Microsoft Fabric. La fonction PREDICT est conçue pour s'intégrer facilement aux modèles MLflow et vous permet d'utiliser le modèle pour générer des prédictions par lots.

Par exemple, chaque semaine, vous recevez des données de ventes de plusieurs magasins. Sur une base d'historique de données, vous avez formé un modèle qui peut prédire les ventes pour la semaine suivante, en fonction des ventes des dernières semaines. Vous avez suivi le modèle avec MLflow et l'avez enregistré dans Microsoft Fabric. Chaque fois que les nouvelles données de ventes hebdomadaires arrivent, vous utilisez la fonction PREDICT pour permettre au modèle de générer les prévisions pour la semaine suivante. Les données de ventes prévues sont stockées sous forme de table dans un lakehouse, qui est visible dans un rapport Power BI que les utilisateurs professionnels peuvent consulter.

Unité 5: Exercice - Explorer la science des données dans Microsoft Fabric

Type: Exercice

Exercice - Explorer la science des données dans Microsoft Fabric

Vous pouvez maintenant explorer les fonctionnalités de science des données dans Microsoft Fabric.

Vous avez besoin d'une licence d'évaluation Microsoft Fabric avec la version préliminaire Fabric activée dans votre compte client. Consultez Bien démarrer avec Fabric pour activer votre licence d'évaluation Fabric.

Lancez l'exercice et suivez les instructions.

Exercice - Explorer la science des données dans Microsoft Fabric Vous pouvez maintenant explorer les fonctionnalités de science des données dans Microsoft Fabric. Remarque Vous avez besoin d'une licence d'évaluation Microsoft Fabric avec la version préliminaire Fabric activée dans votre compte client. Consultez Bien démarrer avec Fabric pour activer votre licence d'évaluation Fabric. Lancez l'exercice et suivez les instructions. Commentaires Yes No

Unité 6: Évaluation du module

Type: Évaluation

Évaluation du module

Vous avez accès à un jeu de données historique qui contient les dépenses mensuelles du service marketing. Vous voulez générer des prédictions des dépenses pour le mois à venir. Quelle tâche devez-vous effectuer pour prédire les dépenses du mois à venir ?

Quelle fonctionnalité de Microsoft Fabric devez-vous utiliser pour passer en revue les résultats du suivi de MLflow via une interface utilisateur ?

Quelle fonctionnalité de Microsoft Fabric devez-vous utiliser pour accélérer l’exploration et le nettoyage des données ?

Vous devez répondre à toutes les questions avant de vérifier votre travail.

Unité 7: Résumé

Type: Résumé

Microsoft Fabric met à disposition un espace de travail central pour effectuer une science des données du début à la fin.

Pour effectuer une science des données, vous devez d’abord définir le problème. Vous pouvez ensuite identifier les données nécessaires et les ingérer à Microsoft Fabric. Une fois que vous avez ingéré vos données, vous pouvez les explorer et les préparer à l’aide de notebooks ou du Data Wrangler.

Pour entraîner des modèles Machine Learning, dans le cadre de votre projet de science des données, vous pouvez suivre votre travail avec des expériences. Pour utiliser un modèle afin de générer des insights, vous pouvez utiliser la fonction PREDICT intégrée.

Module 10: Administrer un environnement Microsoft Fabric

Unité 1: Présentation

Type: Introduction

L'administration d'un environnement Microsoft Fabric implique des tâches essentielles pour garantir l'utilisation efficace et efficace de la plateforme Fabric au sein d'une organisation.

En tant qu'administrateur (admin) Fabric, vous devez connaître :

Architecture de Fabric

Fonctionnalités de sécurité et de gouvernance

Fonctionnalités d'analytique

Options de déploiement et de licence

Vous devez également connaître le portail d'administration Fabric et d'autres outils d'administration, et être en mesure de configurer et de gérer l'environnement Fabric pour répondre aux besoins de votre organisation.

Les administrateurs Fabric travaillent avec les utilisateurs professionnels, les analystes Données et les professionnels informatiques pour déployer et utiliser Fabric en vue de répondre aux objectifs métier et de respecter les stratégies et normes organisationnelles.

À la fin de ce module, vous aurez une bonne compréhension du rôle d'administrateur Fabric et des tâches et outils impliqués dans l'administration de Fabric.

Unité 2: Comprendre l'architecture Fabric

Type: Contenu

Comprendre l'architecture Fabric

Microsoft Fabric est une plateforme Software-as-a-Service qui offre une approche simple et intégrée tout en réduisant la charge administrative. Fabric fournit une solution d'analytique tout-en-un pour les entreprises qui couvre tout, du déplacement des données à la science des données, à l'analytique en temps réel et au décisionnel. Il offre une suite complète de services, notamment les suivants :

Entrepôt de données

Engineering données

Intégration des données

Science des données

Informations en temps réel

Toutes les données de Fabric sont stockées dans OneLake, qui repose sur l'architecture Azure Data Lake Storage (ADLS) gen2. OneLake est hiérarchique par nature pour simplifier la gestion dans toute votre organisation. Il n'y a qu'un seul OneLake par locataire et il fournit un espace de noms de système de fichiers qui s'étend aux utilisateurs, aux régions et même aux clouds dans une seule et même vue.

Comprendre les concepts de Fabric

Un locataire est un espace dédié permettant aux organisations de créer, de stocker et de gérer des éléments Fabric. Il y a souvent une seule instance de Fabric pour une organisation, qui elle est alignée sur Microsoft Entra ID. Le locataire Fabric est mappé à la racine de OneLake et se trouve au niveau supérieur de la hiérarchie.

La capacité est un ensemble dédié de ressources qui sont disponibles à un moment donné pour être utilisées. Une ou plusieurs capacités peuvent être associées à un locataire. La capacité définit la possibilité d'une ressource à effectuer une activité ou à produire une sortie. Les besoins en capacité varient selon l'élément et la durée d'utilisation. Fabric offre une capacité via la référence SKU et les évaluations Fabric.

Un domaine est un regroupement logique d'espaces de travail. Les domaines sont utilisés pour organiser des éléments d'une manière logique pour votre organisation. Vous pouvez regrouper des éléments de manière à faciliter l'accès des groupes de personnes aux espaces de travail. Par exemple, vous pouvez avoir un domaine pour les ventes, un autre pour le marketing et un autre pour les finances.

Un espace de travail est une collection d'éléments qui regroupe différentes fonctionnalités dans un seul locataire. Il agit comme un conteneur qui utilise la capacité disponible pour le travail exécuté, et fournit des contrôles pour les personnes qui peuvent accéder aux éléments qu'il contient. Par exemple, dans un espace de travail des ventes, les utilisateurs associés à l'organisation des ventes peuvent créer un entrepôt de données, exécuter des notebooks, créer des jeux de données, créer des rapports et plus encore.

Les éléments sont les composantes de la plateforme Fabric. Il s'agit des objets que vous créez et gérez dans Fabric. Il existe différents types d'éléments, tels que les entrepôts de données, les pipelines de données, les jeux de données, les rapports et les tableaux de bord.

En tant qu'administrateur, il est important de comprendre les concepts Fabric, car cela vous permet de comprendre comment gérer l'environnement Fabric.

Pour plus d'informations, consultez la documentation Démarrer un essai Fabric.

Unité 3: Comprendre le rôle Administrateur Fabric

Type: Contenu

Comprendre le rôle Administrateur Fabric

Maintenant que vous comprenez l'architecture fabric et ce que vous et votre équipe pouvez utiliser Fabric, examinons le rôle d'administrateur et les outils utilisés pour gérer la plateforme.

Il y a plusieurs rôles qui coopèrent pour administrer Microsoft Fabric pour votre organisation. Si vous êtes administrateur Microsoft 365, administrateur Power Platform ou administrateur de capacité Fabric, vous êtes impliqué dans l'administration de Fabric. Le rôle d'administrateur Fabric était anciennement l'administrateur Power BI.

En tant qu'administrateur Fabric, vous travaillez principalement dans le portail d'administration Fabric. Vous devrez peut-être également vous familiariser avec les outils suivants :

Centre d'administration Microsoft 365

Sécurité Microsoft 365 & Portail de conformité Microsoft Purview

Ouvrez Microsoft Entra ID dans le portail Azure

Cmdlets PowerShell

API et SDK d’administration

Pour obtenir des détails spécifiques sur les différents rôles d’administrateur et leurs responsabilités, consultez la documentation Qu’est-ce que l’administration Microsoft Fabric ?.

Décrire les tâches d’administration

En tant qu’administrateur, vous pourriez être responsable d’un large éventail de tâches pour garantir la bonne exécution de la plateforme Fabric. Il s’agit notamment des tâches suivantes :

Sécurité et contrôle d’accès : l’un des aspects les plus importants de l’administration de Fabric est la gestion de la sécurité et du contrôle d’accès pour garantir que seuls les utilisateurs autorisés peuvent accéder aux données sensibles. Vous pouvez utiliser le contrôle d’accès en fonction du rôle (RBAC) pour :

Définir qui peut afficher et modifier du contenu.

Configurer des passerelles de données pour vous connecter en toute sécurité à des sources de données locales.

Gérer les accès des utilisateurs avec Microsoft Entra ID.

Gouvernance des données : l’administration efficace de Fabric nécessite une bonne compréhension des principes de gouvernance des données. Vous devez savoir comment sécuriser la connectivité entrante et sortante dans votre locataire et comment superviser les métriques d’utilisation et de performances. Vous devez également savoir comment appliquer des stratégies de gouvernance des données pour garantir que les données au sein de votre locataire ne sont accessibles qu’aux utilisateurs autorisés.

Personnalisation et configuration : l’administration de Fabric implique également la personnalisation et la configuration de la plateforme pour répondre aux besoins de votre organisation. Vous pouvez configurer des liens privés pour sécuriser votre locataire, définir des stratégies de classification des données ou ajuster l’apparence des rapports et des tableaux de bord.

Supervision et optimisation : en tant qu’administrateur Fabric, vous devez savoir comment superviser les performances et l’utilisation de la plateforme, optimiser les ressources et résoudre les problèmes. Les exemples incluent la configuration des paramètres de supervision et d’alerte, l’optimisation des performances des requêtes, la gestion de la capacité et de la mise à l’échelle, et la résolution des problèmes d’actualisation des données et de connectivité.

Des tâches spécifiques varient en fonction des besoins de votre organisation et de la complexité de votre implémentation Fabric.

Décrire les outils d’administration

Il est important de vous familiariser avec quelques outils pour implémenter efficacement les tâches décrites précédemment. Les administrateurs Fabric peuvent effectuer la plupart des tâches d’administration à l’aide d’un ou de plusieurs des outils suivants : le portail d’administration Fabric, les applets de commande PowerShell, les SDK et API d’administration, et l’espace de travail de supervision de l’administration.

Portail d’administration Fabric

Le portail d’administration de Fabric est un portail web où vous pouvez gérer tous les aspects de la plateforme. Vous pouvez gérer, examiner et appliquer les paramètres de manière centralisée pour l’ensemble du locataire ou par capacité dans le portail d’administration. Vous pouvez également gérer

les utilisateurs, les administrateurs et les groupes, accéder aux journaux d'audit, et superviser l'utilisation et les performances.

Le portail d'administration vous permet d'activer et de désactiver les paramètres. Il existe de nombreux paramètres situés dans le portail d'administration. Un paramètre remarquable est le commutateur Fabric, situé dans les paramètres du locataire, qui permet aux organisations utilisant Power BI de choisir d'activer Fabric. Ici, vous pouvez activer Fabric pour votre locataire ou autoriser les administrateurs de capacité à activer Fabric.

Applets de commande PowerShell

Fabric fournit un ensemble d'applets de commande PowerShell que vous pouvez utiliser pour automatiser les tâches d'administration courantes. Une applet de commande PowerShell est une commande simple qui peut être exécutée dans PowerShell.

Par exemple, vous pouvez utiliser des applets de commande dans Fabric pour créer et gérer des groupes, configurer des sources de données et des passerelles, et superviser l'utilisation et les performances de manière systématique. Vous pouvez également utiliser les applets de commande pour gérer les SDK et API d'administration Fabric.

Pour obtenir plus de ressources sur les applets de commande PowerShell qui fonctionnent avec Fabric, consultez Applets de commande Microsoft Power BI pour Windows PowerShell et PowerShell Core.

SDK et API d'administration

Un SDK et une API d'administration sont des outils qui permettent aux développeurs d'interagir avec un système logiciel par programmation. Une interface de programmation d'applications (API, Application Programming Interface) est un ensemble de protocoles et d'outils qui permettent la communication entre différentes applications logicielles. Un Kit de développement logiciel (SDK, Software Development Kit) est un ensemble d'outils et de bibliothèques qui permettent aux développeurs de créer des applications logicielles pouvant interagir avec un système ou une plateforme spécifique. Vous pouvez utiliser des API et des SDK pour automatiser des tâches d'administration courantes et intégrer Fabric à d'autres systèmes.

Par exemple, vous pouvez utiliser des API et des sdk pour créer et gérer des groupes, configurer des sources de données et des passerelles, et surveiller l'utilisation et les performances. Vous pouvez également utiliser les API et les SDK pour gérer les SDK et API d'administration Fabric.

Vous pouvez effectuer ces demandes à l'aide de n'importe quelle bibliothèque de client HTTP qui prend en charge l'authentification OAuth 2.0, comme Postman, ou vous pouvez utiliser des scripts PowerShell pour automatiser le processus.

Espace de travail de supervision de l'administration

Les administrateurs de locataires Fabric ont accès à l'espace de travail de surveillance de l'administration. Vous pouvez choisir de partager l'accès à l'espace de travail ou à des éléments spécifiques qu'il contient avec d'autres utilisateurs de votre organisation. L'espace de travail de supervision de l'administration comprend le jeu de données et le rapport Adoption et Utilisation des fonctionnalités qui, ensemble, fournissent des insights sur l'utilisation et les performances de votre environnement Fabric. Vous pouvez utiliser ces informations pour identifier les tendances et les modèles, et résoudre les problèmes.

Pour plus d'informations sur ce qui est inclus dans l'espace de travail d'analyse administrateur, consultez Qu'est-ce que l'espace de travail de supervision de l'administrateur Fabric.

Unité 4: Gérer la sécurité Fabric

Type: Contenu

Gérer la sécurité Fabric

En tant qu'administrateur d'infrastructure, une partie de votre rôle consiste à gérer la sécurité de l'environnement Fabric, notamment la gestion des utilisateurs et des groupes, ainsi que la façon dont les utilisateurs partagent et distribuent du contenu dans Fabric.

Gérer les utilisateurs : attribuer et gérer des licences

Les licences utilisateur contrôlent le niveau d'accès et de fonctionnalité utilisateur dans l'environnement Fabric. Les administrateurs garantissent que les utilisateurs avec une licence disposent de l'accès dont ils ont besoin pour effectuer leurs tâches efficacement. Ils limitent également l'accès aux données sensibles et garantissent la conformité aux lois et réglementations relatives à la protection des données.

La gestion des licences permet aux administrateurs de superviser et de contrôler les coûts en veillant à ce que les licences soient allouées efficacement et uniquement aux utilisateurs qui en ont besoin. Cela permet d'éviter des dépenses inutiles et à garantir que l'organisation utilise efficacement ses ressources.

Le fait de disposer des procédures appropriées pour attribuer et gérer des licences permet de contrôler l'accès aux données et à l'analytique, de garantir la conformité aux réglementations et d'optimiser les coûts.

La gestion des licences pour Fabric est gérée dans le Centre d'administration Microsoft 365. Pour plus d'informations sur la gestion des licences, consultez Affecter des licences aux utilisateurs.

Le type de licence dans les paramètres de l'espace de travail est lié aux licences utilisateur répertoriées ici. Les utilisateurs peuvent voir des rapports en fonction de la licence utilisateur et de la licence de l'espace de travail. Pour plus d'informations, consultez la documentation sur les licences Microsoft Fabric .

Gérer les éléments et le partage

En tant qu'administrateur, vous pouvez gérer la façon dont les utilisateurs partagent et distribuent du contenu. Vous pouvez gérer la façon dont les utilisateurs partagent du contenu avec d'autres utilisateurs et la façon dont ils distribuent du contenu à d'autres utilisateurs. Vous pouvez également gérer la façon dont les utilisateurs interagissent avec les éléments, tels que les entrepôts de données, les pipelines de données, les jeux de données, les rapports et les tableaux de bord.

Les éléments des espaces de travail sont mieux distribués via une application d'espace de travail ou directement via l'espace de travail. L'octroi des droits les moins permissifs est la première étape de la sécurisation des données. Partagez l'application en lecture seule pour l'accès aux rapports ou accordez l'accès aux espaces de travail pour la collaboration et le développement. Un autre aspect de la gestion et de la distribution des éléments consiste à appliquer ces types de bonnes pratiques.

Vous pouvez gérer le partage et la distribution à la fois en interne et en dehors de votre organisation, conformément à ses stratégies et ses procédures.

Pour plus d'informations, consultez la documentation sécurité dans Microsoft Fabric .

Unité 5: Gouverner des données dans Fabric

Type: Contenu

Gouverner des données dans Fabric

Fabric inclut des fonctionnalités de gouvernance intégrées pour vous aider à gérer et à contrôler vos données. L'approbation est un moyen vous permettant, en tant qu'administrateur, de désigner des éléments Fabric spécifiques comme fiables et approuvés pour une utilisation dans toute l'organisation.

Les administrateurs peuvent également utiliser l'API scanneur pour analyser des éléments Fabric pour rechercher des données sensibles et la fonctionnalité de traçabilité des données pour suivre le flux de données via Fabric.

Approuver le contenu Fabric

L'approbation est une fonctionnalité de gouvernance clé qui génère une confiance dans vos ressources de données en marquant les éléments Fabric comme étant examinés et approuvés. Les éléments approuvés affichent un badge qui indique aux utilisateurs que ces ressources sont fiables. L'approbation permet aux utilisateurs de faire confiance aux données et vous aide également, en tant qu'administrateur, à gérer la croissance globale des éléments dans votre environnement.

Le contenu Fabric promu s'affiche avec un badge Promu dans le portail Fabric. Les membres de l'espace de travail ayant le rôle Contributeur ou Administrateur peuvent promouvoir du contenu au sein d'un espace de travail. L'administrateur Fabric peut promouvoir du contenu au sein de l'organisation.

Le contenu certifié nécessite un processus plus formel qui implique une révision du contenu par un réviseur désigné. Le contenu certifié s'affiche avec un badge Certifié dans le portail Fabric. Les administrateurs gèrent le processus de certification et peuvent le personnaliser pour répondre aux besoins de votre organisation.

Si vous n'êtes pas administrateur, vous devez demander la certification d'élément auprès d'un administrateur. Vous pouvez effectuer une certification de demande en sélectionnant l'élément dans le portail Fabric, puis en sélectionnant Demander la certification dans le menu Plus .

Pour plus d'informations sur le processus d'approbation de contenu, consultez Promouvoir ou certifier du contenu.

Rechercher des données sensibles

L'analyse des métadonnées facilite la gouvernance des données en activant le catalogage et la création de rapports sur toutes les métadonnées des éléments Fabric de votre organisation. L'API scanneur est un ensemble d'API REST d'administration qui vous permet d'analyser les éléments Fabric pour les données sensibles. Utilisez l'API de scanneur pour analyser les entrepôts de données, les pipelines de données, les jeux de données, les rapports et les tableaux de bord afin de rechercher des données sensibles. L'API de scanneur peut être utilisée pour analyser à la fois des données structurées et des données non structurées.

Avant que l'analyse des métadonnées puisse être exécutée, elle doit être configurée dans votre organisation par un administrateur. Pour plus d'informations, consultez la vue d'ensemble de l'analyse des métadonnées.

Suivre la traçabilité des données

La traçabilité des données est la possibilité de suivre le flux de données via Fabric, également appelée analyse d'impact. La traçabilité des données vous permet de voir d'où proviennent les données, comment elles sont transformées et leur destination. La vue de traçabilité dans les espaces de travail vous permet de comprendre les données disponibles dans Fabric et la façon dont elles sont utilisées.

Rapport sur les données sensibles

Avec le hub Microsoft Purview (préversion) dans Fabric, vous pouvez gérer et gouverner le paysage de données de données Fabric de votre organisation. Il contient des rapports qui fournissent des informations sur les données sensibles, l'approbation d'élément et les domaines, et sert également de passerelle vers des fonctionnalités plus avancées dans le portail Microsoft Purview, telles que Data Catalog, Information Protection, Data Loss Prevention et Audit.

Unité 6: Évaluation du module

Type: Évaluation

Évaluation du module

Parmi les affirmations suivantes, laquelle décrit le mieux le concept de capacité dans Fabric ?

La capacité fait référence à un espace dédié permettant aux organisations de créer, de stocker et de gérer des éléments Fabric.

La capacité définit la possibilité d'une ressource à effectuer une activité ou à produire une sortie.

La capacité est une collection d'éléments qui sont regroupés de manière logique.

Parmi les affirmations suivantes, laquelle est vraie concernant la différence entre la promotion et la certification dans Fabric ?

La promotion et la certification permettent toutes les deux à tout membre de l'espace de travail d'approuver du contenu.

La promotion nécessite un niveau d'autorisations plus élevé que celui de la certification.

La certification doit être activée dans le locataire par l'administrateur, tandis que la promotion peut être e par un membre de l'espace de travail.

Vous devez répondre à toutes les questions avant de vérifier votre travail.

Unité 7: Résumé

Type: Résumé

Dans ce module, vous avez découvert l'architecture Fabric et le rôle d'un administrateur pour la plateforme Fabric. Vous avez également exploré les différents outils disponibles pour la gestion de la sécurité et du partage, ainsi que les fonctionnalités de gouvernance qui peuvent être utilisées pour appliquer des normes et garantir la conformité. Si vous comprenez bien comment gérer un environnement Fabric, vous garantissez sa sécurité, sa conformité et sa gouvernance. Fort de ces connaissances, vous êtes bien équipé pour aider votre organisation à tirer le meilleur parti de Fabric et à retirer de précieux insights de toutes vos données.

Pour plus d'informations sur la gouvernance des données, complétez les données de gouvernance dans Microsoft Fabric avec le module Purview .

Parcours 2: Implémenter un entrepôt de données avec Microsoft Fabric

Module 1: Bien démarrer avec les entrepôts de données dans Microsoft Fabric

Unité 1: Présentation

Type: Introduction

Les entrepôts de données relationnelles sont au centre de la plupart des solutions décisionnels d'entreprise . Bien que les détails spécifiques puissent varier entre les implémentations de l'entrepôt de données, un modèle courant basé sur un schéma dénormalisé, le schéma multidimensionnel a émergé comme la conception standard d'un entrepôt de données relationnelle.

L'entrepôt de données de Microsoft Fabric est une version moderne de l'entrepôt de données traditionnel. Il centralise et organise les données de différents services, systèmes et bases de données en une seule vue unifiée à des fins d'analyse et de création de rapports. L'entrepôt de données de Fabric fournit une sémantique SQL complète, notamment la possibilité d'insérer, de mettre à jour et de supprimer des données dans les tables. L'entrepôt de données de Fabric est unique, car il est basé sur lakehouse, qui est stocké au format Delta et peut être interrogé à l'aide de SQL. Il est conçu pour être utilisé par toute l'équipe de données, pas seulement pour les ingénieurs données.

L'expérience de l'entrepôt de données de Fabric est conçue pour relever ces défis. Fabric permet aux ingénieurs de données, aux analystes et aux scientifiques des données de travailler ensemble pour créer et interroger un entrepôt de données optimisé pour leurs besoins spécifiques.

Dans ce module, vous allez découvrir les entrepôts de données dans Fabric, créer un entrepôt de données, charger, interroger et visualiser des données.

Unité 2: Comprendre les principes de base d'un entrepôt de données

Type: Contenu

Comprendre les principes de base d'un entrepôt de données

Le processus de création d'un entrepôt de données moderne se compose généralement des tâches suivantes :

Ingestion des données : déplacement des données de systèmes sources vers un entrepôt de données.

Stockage des données : stockage des données dans un format optimisé pour l'analytique.

Traitement des données : transformation des données dans un format consommable par les outils analytiques.

Analyse et remise des données : analyse des données pour obtenir des insights et remise de ces insights à l'entreprise.

Microsoft Fabric permet aux ingénieurs et aux analystes de données d'ingérer, de stocker, de transformer et de visualiser des données dans un seul outil en combinant expérience traditionnelle et « low-code ».

Comprendre l'expérience d'entrepôt de données de Fabric

L'entrepôt de données de Fabric est un entrepôt de données relationnelle qui prend en charge les fonctionnalités T-SQL transactionnelles complètes que vous attendez d'un entrepôt de données d'entreprise. Complètement managé, scalable et hautement disponible, il peut être utilisé pour stocker et interroger des données dans le lakehouse. L'entrepôt de données vous permet de contrôler entièrement la création de tables ainsi que le chargement, la transformation et l'interrogation des données à l'aide du portail Fabric ou de commandes T-SQL. Vous pouvez utiliser soit SQL pour interroger et analyser les données, soit Spark pour traiter les données et créer des modèles Machine Learning.

Les entrepôts de données dans Fabric facilitent la collaboration entre les ingénieurs données et les analystes de données, qui partagent alors la même expérience. Les ingénieurs données créent, au-dessus des données dans le lakehouse, une couche relationnelle dans laquelle les analystes peuvent utiliser T-SQL et Power BI pour explorer les données.

Concevoir un entrepôt de données

Comme toutes les bases de données relationnelles, l'entrepôt de données de Fabric contient des tables pour stocker les données à des fins d'analytique. Le plus souvent, ces tables sont organisées dans un schéma optimisé pour la modélisation multidimensionnelle. Dans cette approche, les données numériques liées aux événements (par exemple, les commandes des clients) sont regroupées selon différents attributs (date, client, magasin, etc.). Par exemple, vous pouvez analyser le montant total payé pour les commandes passées à une date spécifique ou dans un magasin particulier.

Tables d'un entrepôt de données

Les tables d'un entrepôt de données sont généralement organisées de manière à analyser efficacement de grandes quantités de données. Cette organisation, souvent appelée « modélisation dimensionnelle », implique de structurer les tables en tables de faits et tables de dimension.

Les tables de faits contiennent les données numériques que vous souhaitez analyser. Les tables de faits comprennent généralement un grand nombre de lignes et constituent la principale source de données pour l'analyse. Par exemple, une table de faits peut contenir le montant total payé pour des commandes passées à une date spécifique ou dans un magasin particulier.

Les tables de dimension contiennent des informations descriptives sur les données des tables de faits. Les tables de dimension comprennent généralement un petit nombre de lignes et fournissent le contexte des données des tables de faits. Par exemple, une table de dimension peut contenir des informations sur les clients qui ont passé des commandes.

En plus des colonnes d'attribut, une table de dimension contient une colonne clé unique qui identifie de manière unique chaque ligne de la table. En fait, il est courant pour une table de dimension d'inclure deux colonnes clés :

Une clé de substitution est un identificateur unique pour chaque ligne de la table de dimension. Il s'agit souvent d'une valeur entière générée automatiquement par le système de gestion de base de données quand une nouvelle ligne est insérée dans la table.

Une autre clé est souvent une clé naturelle ou métier qui identifie une instance spécifique d'une entité dans le système source transactionnel, comme un code produit ou un ID client.

Dans un entrepôt de données, les clés de substitution et les clés secondaires ont des finalités différentes. Vous avez donc besoin des deux. Les clés de substitution sont spécifiques à l'entrepôt de

données et contribuent au maintien de la cohérence et de l'exactitude des données. Quant aux clés alternatives, elles sont spécifiques au système source et contribuent au maintien de la traçabilité entre l'entrepôt de données et le système source.

Tables de dimension de type spécial

Les dimensions de type spécial offrent un contexte supplémentaire et permettent une analyse des données plus complète.

Les dimensions de temps fournissent des informations sur la période pendant laquelle un événement s'est produit. Cette table permet aux analystes de données d'agréger des données sur des intervalles temporels. Par exemple, une dimension de temps peut inclure les colonnes « année », « trimestre », « mois » et « jour » pour indiquer quand une commande a été passée.

Les dimensions à variation lente sont des tables de dimension qui suivent les modifications apportées aux attributs de dimension au fil du temps, telles que les modifications apportées à l'adresse d'un client ou au prix d'un produit. Elles occupent une place importante dans un entrepôt de données, car elles permettent aux utilisateurs d'analyser et de comprendre les modifications apportées aux données dans le temps. Les dimensions à variation lente garantissent que les données sont à jour et exactes, ce qui est primordial pour prendre de bonnes décisions commerciales.

Conceptions de schémas d'entrepôts de données

Dans la plupart des bases de données transactionnelles utilisées dans les applications métier, les données sont normalisées pour réduire la duplication. Toutefois, dans un entrepôt de données, les données de dimension sont généralement dénormalisées pour réduire le nombre de jointures requises pour interroger les données.

Souvent, un entrepôt de données est organisé en tant que schéma en étoile, dans lequel une table de faits est directement liée aux tables de dimension, comme illustré dans cet exemple :

Vous pouvez utiliser les attributs d'un élément pour regrouper des nombres dans la table de faits à différents niveaux. Par exemple, vous pouvez trouver le chiffre d'affaires total d'une région entière ou d'un seul client. Les informations de chaque niveau peuvent être stockées dans la même table de dimension.

Voir Qu'est-ce qu'un schéma en étoile ? pour plus d'informations sur la conception de schémas en étoile pour Fabric.

S'il existe un grand nombre de niveaux ou si certaines informations sont partagées par des éléments différents, il peut être judicieux d'utiliser un schéma flocon à la place. Voici un exemple :

Dans ce cas, la table DimProduct a été divisée (normalisée) pour créer des tables de dimension distinctes pour les catégories de produits et les fournisseurs.

Chaque ligne de la table DimProduct contient des valeurs clés pour les lignes correspondantes dans les tables DimCategory et DimSupplier.

Une table DimGeography a été ajoutée contenant des informations sur l'emplacement des clients et des magasins.

Chaque ligne des tables DimCustomer et DimStore contient une valeur clé pour la ligne correspondante dans la table DimGeography .

Unité 3: Comprendre les entrepôts de données dans Fabric

Type: Contenu

Comprendre les entrepôts de données dans Fabric

Le lakehouse de Fabric est une collection de fichiers, dossiers, tables et raccourcis qui agissent comme une base de données sur un lac de données. Utilisé par le moteur Spark et le moteur SQL pour le traitement du Big Data, il propose des fonctionnalités pour les transactions ACID lors de l'utilisation de tables au format Delta open source.

L'expérience d'entrepôt de données de Fabric vous permet de passer de la vue du lac du lakehouse (qui prend en charge l'engineering données et Apache Spark) aux expériences SQL d'un entrepôt de données traditionnel. Le Lakehouse vous permet de lire des tables et d'utiliser le point de terminaison d'analyse SQL, tandis que l'entrepôt de données vous permet de manipuler les données.

Dans l'expérience d'entrepôt de données, vous pouvez modéliser les données à l'aide de tables et de vues, exécuter des commandes T-SQL pour interroger les données dans l'entrepôt de données et le lakehouse, utiliser T-SQL pour effectuer des opérations DML sur les données à l'intérieur de l'entrepôt de données et remettre des données à des couches de création de rapports comme Power BI.

Maintenant que vous comprenez les principes architecturaux de base d'un schéma d'entrepôt de données relationnel, nous allons voir comment créer un entrepôt de données.

Décrire un entrepôt de données dans Fabric

Dans l'expérience d'entrepôt de données dans Fabric, vous pouvez créer une couche relationnelle au-dessus des données physiques dans le lakehouse et l'exposer à des outils d'analyse et de création de rapports. Vous pouvez créer votre entrepôt de données directement dans Fabric à partir du hub de création ou dans un espace de travail. Après avoir créé un entrepôt vide, vous pouvez y des objets.

Une fois votre entrepôt créé, vous pouvez créer des tables en utilisant T-SQL directement dans l'interface de Fabric.

Ingérer des données dans votre entrepôt de données

Pour ingérer des données dans un entrepôt de données Fabric, plusieurs méthodes s'offrent à vous. Vous pouvez utiliser des pipelines, des flux de données, l'interrogation entre bases de données ou encore la commande COPY INTO. Après ingestion, les données peuvent être analysées par plusieurs groupes d'entreprise qui peuvent utiliser des fonctionnalités telles que le partage et l'interrogation entre bases de données pour y accéder.

Créer des tables

Pour créer une table dans l'entrepôt de données, vous pouvez utiliser SQL Server Management Studio (SSMS) ou un autre client SQL pour vous connecter à l'entrepôt de données et exécuter une instruction CREATE TABLE. Vous pouvez également créer des tables directement dans l'interface utilisateur de Fabric.

Vous pouvez copier des données à partir d'un emplacement externe dans une table de l'entrepôt de données à l'aide de la syntaxe COPY INTO. Par exemple :

Cette requête SQL charge les données d'un fichier CSV situé dans Stockage Blob Azure dans une table appelée « Region » dans l'entrepôt de données Fabric.

Cloner des tableaux

Vous pouvez créer des clones de table sans duplication avec des coûts de stockage minimes dans un entrepôt de données. Ces clones sont essentiellement des réplicas de tables créés en copiant les métadonnées tout en référençant les mêmes fichiers de données dans OneLake. Cela signifie que les données sous-jacentes stockées sous forme de fichiers Parquet ne sont pas dupliquées, ce qui permet d'économiser des coûts de stockage.

Les clones de table sont particulièrement utiles dans plusieurs scénarios.

Développement et test : Les clones permettent aux développeurs et aux testeurs de créer des copies de tables dans des environnements inférieurs, ce qui facilite les processus de développement, débogage, test et validation.

Récupération des données : En cas d'échec d'une mise en production ou d'une altération des données, les clones de table peuvent conserver l'état précédent des données, ce qui permet la récupération des données.

Rapports historiques : Ils aident à créer des rapports historiques qui reflètent l'état des données à des moments précis et à conserver les données à des étapes clés de l'activité.

Vous pouvez créer un clone de table avec la commande T-SQL CREATE TABLE AS CLONE OF.

Pour en savoir plus sur les clones de table, consultez Tutoriel : Cloner une table avec T-SQL dans Microsoft Fabric.

Considérations relatives aux tables

Au terme de la création de tables dans un entrepôt de données, il est important de prendre en compte le processus de chargement des données dans ces tables. Une approche courante consiste à utiliser des tables de mise en lots. Dans Fabric, vous pouvez utiliser des commandes T-SQL pour charger des données à partir de fichiers dans des tables de mise en lots dans l'entrepôt de données.

Les tables de mise en lots sont des tables temporaires qui peuvent être utilisées pour nettoyer, transformer et valider des données. Vous pouvez également utiliser des tables de mise en lots pour charger des données de plusieurs sources dans une table de destination unique.

En général, les données sont chargées dans le cadre d'un processus de traitement par lots périodique dans lequel les insertions et mises à jour de l'entrepôt de données sont coordonnées pour se produire à un intervalle régulier (par exemple quotidien, hebdomadaire ou mensuel).

Dans la plupart des cas, vous devez implémenter un processus de chargement d'entrepôt de données, qui effectue les tâches dans l'ordre suivant :

Ingérez les nouvelles données à charger dans un lac de données, en appliquant un nettoyage ou des transformations avant le chargement, selon les besoins.

Chargez les données à partir de fichiers dans des tables de mise en lots au sein de l'entrepôt de données relationnel.

Chargez les tables de dimension à partir des données de dimension dans les tables de mise en lots, en mettant à jour les lignes existantes ou en insérant de nouvelles lignes, et en générant des valeurs de clé de substitution le cas échéant.

Chargez les tables de faits à partir des données de faits dans les tables de mise en lots, en recherchant les clés de substitution appropriées pour les dimensions associées.

Effectuez une optimisation postchargement en mettant à jour les index et les statistiques de distribution des tables.

Si vous avez des tables dans le lac et que vous voulez pouvoir les interroger dans votre entrepôt - sans les modifier – avec un entrepôt de données Fabric, vous n'avez pas besoin de copier les données du lac vers l'entrepôt de données. Vous pouvez interroger les données dans le lakehouse directement à partir de l'entrepôt de données en utilisant l'interrogation entre bases de données.

L'utilisation de tables dans l'entrepôt de données Fabric présente actuellement certaines limitations. Pour plus d'informations, consultez Tables dans l'entreposage de données dans Microsoft Fabric.

Unité 4: Interroger et transformer des données

Type: Contenu

Interroger et transformer des données

Maintenant que vous savez comment implémenter un entrepôt de données dans Fabric, préparons les données pour l'analytique.

Il existe deux façons d'interroger des données à partir de votre entrepôt de données. L'éditeur de requête visual fournit une expérience sans code, glisser-déplacer pour créer vos requêtes. Si vous êtes à l'aise avec T-SQL, vous préférez peut-être utiliser l'éditeur de requête SQL pour écrire vos requêtes. Dans les deux cas, vous pouvez créer des tables, des vues et des procédures stockées pour interroger des données dans l'entrepôt de données et Lakehouse.

Il existe également un point de terminaison d'analytique SQL, où vous pouvez vous connecter à partir de n'importe quel outil.

Interroger des données à l'aide de l'éditeur de requête SQL

L'éditeur de requête SQL fournit une expérience de requête qui inclut IntelliSense, la saisie semi-automatique du code, la mise en surbrillance de la syntaxe, l'analyse côté client et la validation. Si vous avez écrit T-SQL dans SQL Server Management Studio (SSMS) ou Azure Data Studio (ADS), vous le trouverez familier.

Pour créer une requête, utilisez le bouton Nouvelle requête SQL dans le menu. Vous pouvez créer et exécuter vos requêtes T-SQL ici. Dans l'exemple ci-dessous, nous créons un nouvel affichage pour les analystes à utiliser pour la création de rapports dans Power BI.

Interroger des données à l'aide de l'éditeur de requête Visual

L'éditeur de requête visuelle offre une expérience similaire à la vue de diagramme en ligne Power Query. Utilisez le bouton Nouvelle requête visuelle pour créer une requête.

Faites glisser une table de votre entrepôt de données vers le canevas pour commencer. Vous pouvez ensuite utiliser le menu Transformer en haut de l'écran pour des colonnes, des filtres et d'autres transformations à votre requête. Vous pouvez utiliser le bouton (+) sur le visuel lui-même pour effectuer des transformations similaires.

Unité 5: Préparer des données pour l'analyse et la création de rapports

Type: Contenu

Préparer des données pour l'analyse et la création de rapports

Un modèle de données sémantique définit les relations entre les différentes tables du modèle sémantique, les règles d'agrégation et de synthèse des données ainsi que les calculs ou mesures utilisés pour générer des insights à partir des données. Ces relations et mesures sont incluses dans le modèle sémantique, qui est ensuite exploité pour créer des rapports dans Power BI.

Vous pouvez facilement basculer entre les vues Données, Requête et Modèle de Fabric à l'aide du menu situé en bas à gauche de l'écran. La vue Données présente les tables du modèle sémantique, la vue Requête les requêtes SQL utilisées pour créer le modèle sémantique et la vue Modèle le modèle

sémantique.

Pour en savoir plus sur les modèles de données et le schéma de l'entrepôt de données, consultez Analyser les données dans un entrepôt de données relationnel.

Créer des relations

Les relations vous permettent de connecter les tables dans le modèle sémantique. Créez des relations entre les tables de votre entrepôt de données en utilisant l'interface « cliquer-glisser » dans la vue Modèle de Fabric.

Pour plus d'informations sur la création de relations, consultez Créer et gérer des relations.

Créer des mesures

Les mesures sont les métriques que vous souhaitez analyser dans votre entrepôt de données. Vous pouvez créer des mesures dans Fabric à l'aide du bouton Nouvelle mesure dans la vue Modèle.

Les mesures sont des champs calculés qui sont basés sur les données des tables de votre entrepôt de données et écrits en langage de formule DAX (Data Analysis Expressions).

Fabric propose de nombreux outils pour créer des transformations de données. La création de mesures avec DAX est l'une des nombreuses façons de créer des transformations de données. Pour en savoir plus sur DAX, consultez Utiliser DAX dans Power BI.

Masquer les champs

La création du modèle sémantique est une phase critique de la préparation de vos données en vue de leur utilisation pour créer des rapports en aval. Pour simplifier les choses pour vos générateurs de rapports, vous pouvez masquer des éléments, comme une table ou une colonne, de la vue. Cliquez avec le bouton droit sur la table ou la colonne, puis sélectionnez Masquer. Le masquage de champs supprime la table ou la colonne de la vue du modèle, mais elle est toujours disponible et exploitable dans le modèle sémantique.

Comprendre les modèles sémantiques

Chaque fois qu'un entrepôt de données est créé, Fabric crée un modèle sémantique auquel les analystes et/ou les utilisateurs professionnels peuvent se connecter pour générer des rapports.

Le modèle sémantique comprend des métriques utilisées pour créer des rapports. Dit simplement, les analystes utilisent le modèle sémantique que vous avez créé dans votre entrepôt et qui est stocké dans un modèle sémantique. Si vous connaissez Power BI, l'utilisation de modèles sémantiques créés par l'expérience de l'entrepôt de données ne vous posera aucune difficulté.

Les modèles sémantiques étant automatiquement synchronisés avec l'entrepôt de données, vous n'avez pas à vous soucier de leur maintenance. Vous pouvez également créer des modèles sémantiques personnalisés pour répondre à vos besoins spécifiques.

Comprendre le modèle sémantique par défaut

Un modèle sémantique par défaut est aussi créé automatiquement pour vous dans Fabric. Il hérite de la logique métier du lakehouse ou de l'entrepôt parent, qui lance l'expérience d'analytique en aval pour l'analyse et le décisionnel. Ce modèle sémantique est géré, optimisé et synchronisé pour vous.

Les nouvelles tables du lakehouse sont automatiquement ajoutées au modèle sémantique par défaut. Pour plus de flexibilité, les utilisateurs peuvent aussi sélectionner manuellement les tables ou vues de l'entrepôt qu'ils souhaitent inclure dans le modèle. Les objets qui se trouvent dans le modèle sémantique par défaut sont créés en tant que disposition dans la vue du modèle.

Les modèles sémantiques par défaut suivent les limitations actuelles des modèles sémantiques dans Power BI. Pour plus d'informations, consultez Modèles sémantiques Power BI par défaut.

Visualiser les données

Fabric vous permet de visualiser les résultats d'une requête unique ou de l'ensemble de votre entrepôt de données sans quitter l'expérience d'entrepôt de données. L'exploration des données pendant que vous travaillez pour vérifier que vous disposez de toutes les données et transformations nécessaires à votre analyse est particulièrement utile.

Utilisez le bouton Nouveau rapport pour créer un rapport Power BI basé sur le contenu de l'ensemble de votre entrepôt de données. Le bouton Nouveau rapport ouvre l'expérience de service Power BI où vous pouvez créer et enregistrer votre rapport en vue d'une utilisation par l'entreprise.

Unité 6: Sécuriser et surveiller votre entrepôt de données

Type: Contenu

Sécuriser et surveiller votre entrepôt de données

La sécurité et la surveillance sont des aspects essentiels de la gestion de votre entrepôt de données.

La sécurité de l'entrepôt de données est importante pour protéger vos données contre l'accès non autorisé. Fabric fournit un certain nombre de fonctionnalités de sécurité pour vous aider à sécuriser votre entrepôt de données. Voici quelques-uns des éléments suivants :

Contrôle d'accès en fonction du rôle (RBAC) pour contrôler l'accès à l'entrepôt et à ses données.

Chiffrement TLS pour sécuriser la communication entre l'entrepôt et les applications clientes.

Azure Storage Service Encryption pour protéger les données en transit et au repos.

Azure Monitor et Azure Log Analytics pour surveiller l'activité de l'entrepôt et auditer l'accès aux données.

Authentification multifacteur (MFA) pour une couche supplémentaire de sécurité aux comptes d'utilisateur.

Intégration d'ID Microsoft Entra pour gérer les identités utilisateur et l'accès à l'entrepôt.

Autorisations d'espace de travail

Les données dans Fabric sont organisées en espaces de travail, qui sont utilisés pour contrôler l'accès et gérer le cycle de vie des données et des services. Les rôles d'espace de travail appropriés sont la première ligne de défense dans la sécurisation de votre entrepôt de données.

Outre les rôles d'espace de travail, vous pouvez accorder des autorisations d'élément et un accès via SQL.

Conseil / Astuce

Pour plus d'informations sur les rôles d'espace de travail , consultez Les espaces de travail dans Power BI .

Autorisations d'élément

Contrairement aux rôles d'espace de travail, qui s'appliquent à tous les éléments d'un espace de travail, vous pouvez utiliser des autorisations d'élément pour accorder l'accès à des entrepôts

individuels. Cela vous permet de partager un entrepôt de données unique pour la consommation en aval.

Vous pouvez accorder des autorisations aux utilisateurs via T-SQL ou dans le portail Fabric. Accordez les autorisations suivantes aux utilisateurs qui doivent accéder à votre entrepôt de données :

Lecture : permet à l'utilisateur de se connecter à l'aide de la chaîne de connexion SQL.

ReadData : permet à l'utilisateur de lire des données à partir de n'importe quelle table/vue dans l'entrepôt.

ReadAll : permet à l'utilisateur de lire les données des fichiers parquet bruts dans OneLake qui peuvent être consommés par Spark.

Une connexion utilisateur au point de terminaison d'analytique SQL va échouer si la permission de lecture, au minimum, n'est pas accordée.

La surveillance des activités dans votre entrepôt de données est essentielle pour garantir des performances optimales, une utilisation efficace des ressources et une sécurité. Il vous aide à identifier les problèmes, à détecter les anomalies et à prendre des mesures pour que l'entrepôt de données s'exécute de manière fluide et sécurisée.

Vous pouvez utiliser des vues de gestion dynamique (DMV) pour surveiller la connexion, la session et l'état de la demande pour afficher les insights de cycle de vie des requêtes SQL en direct. Les DMV vous permettent d'obtenir des détails tels que le nombre de requêtes actives et d'identifier les requêtes qui s'exécutent pendant une longue période et qui doivent être arrêtées.

Il existe actuellement trois DMV disponibles pour l'utilisation dans Fabric :

sys.dm_exec_connections : retourne des informations sur chaque connexion établie entre l'entrepôt et le moteur.

sys.dm_exec_sessions : retourne des informations sur chaque session authentifiée entre l'élément et le moteur.

sys.dm_exec_requests : retourne des informations sur chaque requête active dans une session.

Analyse des requêtes

Utilisez « sys.dm_exec_requests » pour identifier les requêtes longues qui peuvent avoir un impact sur les performances globales de la base de données et prendre les mesures appropriées pour optimiser ou arrêter ces requêtes.

Commencez par identifier les requêtes qui s'exécutent depuis longtemps. Utilisez la requête suivante pour identifier les requêtes qui ont été exécutées le plus longtemps, dans l'ordre décroissant :

Vous pouvez continuer à examiner pour comprendre l'utilisateur qui a exécuté la session avec la requête de longue durée, en exécutant :

Enfin, vous pouvez utiliser la KILL commande pour terminer la session avec la requête longue :

Vous devez être administrateur de l'espace de travail pour exécuter la KILL commande. Les administrateurs d'espace de travail peuvent exécuter les trois DMV. Les rôles Membre, Contributeur et Visionneuse peuvent voir leurs propres résultats dans l'entrepôt, mais ne peuvent pas voir les résultats d'autres utilisateurs.

Unité 7: Exercice - Analyser des données dans un entrepôt de données

Type: Exercice

Exercice - Analyser des données dans un entrepôt de données

Vous devez maintenant créer un entrepôt de données dans Fabric et analyser vos données.

Vous avez besoin d'une licence d'évaluation Microsoft Fabric avec la version préliminaire Fabric activée dans votre compte client. Consultez Bien démarrer avec Fabric pour activer votre licence d'évaluation Fabric.

Lancez l'exercice et suivez les instructions.

Exercice - Analyser des données dans un entrepôt de données Vous devez maintenant créer un entrepôt de données dans Fabric et analyser vos données. Remarque Vous avez besoin d'une licence d'évaluation Microsoft Fabric avec la version préliminaire Fabric activée dans votre compte client. Consultez Bien démarrer avec Fabric pour activer votre licence d'évaluation Fabric. Lancez l'exercice et suivez les instructions. Commentaires Yes No

Unité 8: Évaluation du module

Type: Évaluation

Évaluation du module

Vérifier vos connaissances

Quel type de tableau une compagnie d'assurance doit-elle utiliser pour stocker les détails de l'attribut fournisseur pour l'agrégation des réclamations ?

Table de dimension.

Table intermédiaire.

Qu'est-ce qu'un modèle sémantique dans l'expérience de l'entrepôt de données ?

Un modèle sémantique est un modèle de données orienté entreprise qui fournit une représentation cohérente et réutilisable des données au sein de l'organisation.

Un modèle sémantique est un modèle de données physique qui décrit la structure des données stockées dans l'entrepôt de données.

Un modèle sémantique est un modèle Machine Learning utilisé pour effectuer des prédictions basées sur les données de l'entrepôt de données.

Quel est l'objectif des autorisations d'élément dans un espace de travail ?

Pour accorder l'accès à tous les éléments d'un espace de travail.

Pour accorder l'accès à des colonnes spécifiques dans une table.

Pour accorder l'accès à des entrepôts individuels pour la consommation en aval.

Vous devez répondre à toutes les questions avant de vérifier votre travail.

Unité 9: Résumé

Type: Résumé

Dans ce module, vous avez découvert les entrepôts de données et la modélisation dimensionnelle, créé un entrepôt de données, chargé, interrogé et visualisé des données, et décrit des modèles sémantiques et comment ils sont utilisés pour la création de rapports en aval.

Résumé Dans ce module, vous avez découvert les entrepôts de données et la modélisation dimensionnelle, créé un entrepôt de données, chargé, interrogé et visualisé des données, et décrit des modèles sémantiques et comment ils sont utilisés pour la création de rapports en aval. Commentaires Yes No

Module 2: Charger des données dans un entrepôt de données Microsoft Fabric

Unité 1: Présentation

Type: Introduction

Microsoft Fabric Data Warehouse est une plateforme complète pour les données, l'analytique et l'IA (Intelligence artificielle). Elle fait référence au processus de stockage, d'organisation et de gestion de grands volumes de données structurées et semi-structurées.

L'entrepôt de données dans Microsoft Fabric est optimisé avec Synapse Analytics. Cela lui confère un ensemble diversifié de fonctionnalités qui facilitent la gestion et l'analyse des données. Il inclut des fonctionnalités avancées de traitement des requêtes et prend en charge les fonctionnalités T-SQL transactionnelles complètes comme un entrepôt de données d'entreprise.

Contrairement à un pool SQL dédié dans Synapse Analytics, un entrepôt dans Microsoft Fabric est axé sur un seul lac de données. Les données de l'entrepôt Microsoft Fabric sont stockées au format de fichier Parquet. Cette configuration permet aux utilisateurs de se concentrer sur des tâches comme la préparation des données, l'analyse et la création de rapports. Il tire parti des fonctionnalités étendues du moteur SQL, où une copie unique de leurs données est stockée dans Microsoft OneLake.

Comprendre le processus ETL (Extraire, Transformer et Charger)

L'ETL constitue la base des flux de travail d'analytique données et de l'entrepôt de données. Examinons certains aspects de la manipulation des données dans un processus d'ETL.

Toutes ces étapes du processus d'ETL peuvent s'exécuter en parallèle, selon le scénario. Dès que certaines données sont prêtes, elles sont chargées sans attendre que les étapes précédentes prennent fin.

Dans les unités suivantes, nous explorons les différentes façons de charger des données dans un entrepôt et comment ces méthodes peuvent faciliter les tâches de création d'une charge de travail d'entrepôt de données.

Unité 2: Explorer les stratégies de chargement de données

Type: Contenu

Explorer les stratégies de chargement de données

Il existe de nombreuses façons de charger des données dans un entrepôt dans Microsoft Fabric. Cette étape est fondamentale, car elle garantit l'intégration dans un référentiel unique de données de haute qualité, transformées ou traitées.

L'efficacité du chargement des données a également un impact direct sur la ponctualité et la précision de l'analyse, ce qui est essentiel pour les processus décisionnels en temps réel. Il est essentiel d'investir du temps et des ressources dans la conception et l'implémentation d'une stratégie de chargement de données robuste afin d'assurer le succès du projet d'entrepôt de données.

Comprendre les opérations d'ingestion et de chargement des données

Bien que les deux processus font partie du pipeline d'extraction, de transformation et de chargement (ETL : extract, transform, load) dans un scénario d'entrepôt de données, ils sont généralement utilisés

à des fins différentes. L'ingestion/l'extraction de données consiste à déplacer des données brutes de différentes sources vers un référentiel central. En revanche, le chargement des données implique de prendre les données transformées ou traitées et de les charger dans la destination de stockage finale pour l'analyse et la création de rapports.

Les entrepôts de données Fabric et les lakehouses stockent automatiquement leurs données dans OneLake à l'aide du format Delta Parquet.

Indexer vos données

Il se peut que vous deviez générer et utiliser des objets auxiliaires impliqués dans une opération de chargement, comme des tables, des procédures stockées et des fonctions. Ces objets auxiliaires sont couramment appelés mise en scène. Les indexations agissent comme des zones de stockage et de transformation temporaires. Ils peuvent partager des ressources avec un entrepôt de données, ou résider dans leur propre zone de stockage.

L'indexation sert de couche d'abstraction, ce qui simplifie et facilite l'opération de chargement dans les tables finales de l'entrepôt de données.

En outre, la zone d'indexation fournit une mémoire tampon qui peut aider à réduire l'impact de l'opération de chargement sur les performances de l'entrepôt de données. Cela est important dans les environnements où l'entrepôt de données doit rester fonctionnel et réactif pendant le processus de chargement des données.

Examiner les types de chargements de données

Il existe deux types de chargements de données à prendre en compte lors du chargement d'un entrepôt de données.

Un processus d'ETL pour un entrepôt de données n'a pas systématiquement besoin à la fois du chargement complet et du chargement incrémentiel. Dans certains cas, une combinaison des deux méthodes peut être utilisée. Le choix entre un chargement complet et un chargement incrémentiel dépend de nombreux facteurs, dont la quantité de données, les caractéristiques des données et les exigences de l'entrepôt de données.

Pour en savoir plus sur l'exécution d'une charge incrémentielle, consultez Chargement incrémentiel.

Présentation de la clé d'entreprise et de la clé de substitution

Dans un entrepôt de données, les clés de substitution et les clés d'entreprise sont essentielles pour bénéficier d'un entreposage de données et d'une intégration des données efficaces, mais elles répondent à des objectifs différents.

Clé de substitution : Une clé de substitution est un identificateur généré par le système utilisé pour identifier de manière unique un enregistrement dans une table dans l'entrepôt de données. Elle n'a aucune signification métier, et est généralement un entier ou un identificateur unique. Les clés de substitution servent à maintenir la cohérence et la justesse dans l'entrepôt de données, en particulier lors de l'intégration de données à partir de plusieurs sources. Elles permettent d'éviter les problèmes qui peuvent provenir des modifications apportées aux systèmes sources, tels que la réutilisation ou la modification des clés d'entreprise.

Clé d'entreprise : Une clé métier, également appelée clé naturelle, est un identificateur qui provient du système source et qui a une signification métier. Elle sert à identifier de manière unique un enregistrement dans le système source. Parmi les exemples de clés d'entreprise, citons les codes de produit, les ID client et les numéros d'employés. Les clés d'entreprise sont importantes pour maintenir la traçabilité entre l'entrepôt de données et les systèmes sources. Elles permettent de s'assurer que les données de l'entrepôt peuvent être correctement mises en correspondance avec les enregistrements correspondants dans les systèmes sources.

Chargement d’une table de dimension

Considérez une table de dimension comme « qui, où, quand, pourquoi » de votre entrepôt de données. C’est l’arrière-plan descriptif qui donne du contexte aux chiffres bruts que l’on retrouve dans les tables de faits.

Par exemple, si vous exécutez un magasin en ligne, il se peut que votre table de faits contienne les données de vente brutes (c’est-à-dire le nombre d’unités vendues pour chaque produit). Mais sans table de dimension, vous ne pouvez pas savoir qui a acheté ces produits, quand ils ont été achetés, ni où se trouve l’acheteur.

Dimensions à variation lente

Les dimensions à variation lente évoluent au fil du temps, mais à un rythme lent et imprévisible. Prenez par exemple l’adresse d’un client dans une entreprise de vente au détail. Lorsqu’un client déménage, son adresse change. Si vous remplacez l’ancienne adresse par la nouvelle, vous perdez les données historiques. Toutefois, si vous souhaitez analyser les données de vente historiques, il est essentiel de savoir où résidait le client au moment de chaque vente. C’est là que les dimensions à variation lente deviennent essentielles.

Il existe plusieurs types de dimensions à variation lente dans un entrepôt de données. Les types 1 et 2 sont les plus fréquemment utilisés.

Type 0 SCD : Les attributs de dimension ne changent jamais.

Type 1 SCD : remplace les données existantes, ne conserve pas l’historique.

Type 2 SCD : ajoute de nouveaux enregistrements pour les modifications, conserve l’historique complet d’une clé naturelle donnée.

Type 3 SCD : L’historique est ajouté en tant que nouvelle colonne.

Type 4 SCD : une nouvelle dimension est ajoutée.

Type 5 SCD : quand certains attributs d’une grande dimension changent au fil du temps, mais l’utilisation du type 2 n’est pas réalisable en raison de la grande taille de la dimension.

Type 6 SCD : Combinaison de type 2 et de type 3.

Lorsqu’une nouvelle version du même élément est apportée à l’entrepôt de données dans la dimension à variation lente de type 2, l’ancienne version est présumée expirée et la nouvelle devient active.

Le code suivant montre un exemple simple sur la façon de gérer la clé métier dans un scD de type 2 pour la table Dim_Products à l’aide de T-SQL.

Le mécanisme de détection des modifications dans les systèmes sources est essentiel pour déceler quand les enregistrements sont insérés, mis à jour ou supprimés. Capture de données modifiées (CDC),suivi des modifications et déclencheurs sont toutes les fonctionnalités disponibles pour la gestion du suivi des données dans les systèmes sources tels que SQL Server.

Charger une table de faits

En règle générale, une opération de chargement standard dans un entrepôt de données implique de gérer les tables de faits après les tables de dimension. Cela garantit que les dimensions, que les faits référencent, sont déjà présentes dans l’entrepôt de données.

Les données de faits intermédiaires incluent généralement les clés d’entreprise des dimensions associées. Par conséquent, votre logique de chargement doit rechercher les clés de substitution correspondantes. Lors de l’utilisation de dimensions à variation lente dans l’entrepôt de données, il est essentiel d’identifier la version appropriée de l’enregistrement de dimension afin de garantir l’utilisation

de la clé de substitution correcte. Celle-ci correspond à l'événement enregistré dans la table de faits avec l'état de la dimension au moment où le fait s'est produit.

Dans de nombreux cas, vous pouvez récupérer la dernière version « actuelle » de la dimension. Toutefois, il peut arriver que vous deviez trouver l'enregistrement de dimension correct en fonction des colonnes DateTime qui indiquent la période de validité de chaque version de la dimension.

L'exemple suivant part du principe que les enregistrements de dimension ont des clés de substitution incrémentielles, et que la dernière version ajoutée d'une instance de dimension spécifique (qui aura la valeur de clé la plus élevée) est susceptible d'être utilisée.

Unité 3: Utiliser des pipelines de données pour charger un entrepôt

Type: Contenu

Utiliser des pipelines de données pour charger un entrepôt

L'entrepôt de Microsoft Fabric fournit des outils d'ingestion des données intégrés, afin de permettre aux utilisateurs de charger et d'ingérer des données dans des entrepôts à grande échelle via des expériences de codage ou de non-codage.

Le pipeline de données est le service basé sur le cloud pour l'intégration des données, qui permet la création de workflows pour le déplacement et la transformation des données à grande échelle. Vous pouvez créer et planifier des pipelines de données qui peuvent ingérer et charger des données provenant de différents magasins de données. Vous pouvez créer des ETL complexes ou des processus ELT qui transforment les données visuellement avec des flux de données.

La plupart des fonctionnalités des pipelines de données dans Microsoft Fabric proviennent d'Azure Data Factory, ce qui permet une intégration et une utilisation fluides de ses fonctionnalités dans l'écosystème Microsoft Fabric.

Toutes les données d'un entrepôt sont automatiquement stockées au format Delta Parquet dans OneLake.

Créer un pipeline de données

Il y a plusieurs façons permettant de lancer l'éditeur de pipeline de données.

Depuis l'espace de travail : Sélectionnez + Nouveau, puis sélectionnez Pipeline de données. Si ce n'est pas visible dans la liste, sélectionnez Autres options, puis recherchez Pipeline de données dans la section Obtenir des données.

Depuis la ressource de l'entrepôt : sélectionnez Obtenir des données, puis Nouveau pipeline de données.

Trois options sont disponibles lors de la création d'un pipeline.

Configurer l'assistant Copier des données

L'assistant de copie de données fournit une interface pas à pas qui facilite la configuration d'une tâcheCopier des données.

Choisir la source de données : Sélectionnez un connecteur et fournissez les informations de connexion.

Se connecter à une source de données : Sélectionnez, affichez un aperçu et choisissez les données. Ceci peut être e à partir de tables ou de vues, ou vous pouvez personnaliser votre sélection en

fournissant votre propre requête.

Choisir la destination des données : Sélectionnez le magasin de données en tant que destination.

Se connecter à la destination des données : Sélectionnez et mappez les colonnes de la source à la destination. Vous pouvez charger vers une nouvelle table ou une table existante.

Paramètres : Configurez d'autres paramètres tels que la mise en lots et les valeurs par défaut.

Une fois que vous avez copié les données, vous pouvez utiliser d'autres tâches pour les transformer et les analyser ultérieurement. Vous pouvez également utiliser la tâche Copier des données, afin de publier les résultats de transformation et d'analyse pour l'aide à la décision (BI) et l'utilisation d'application.

Planifier un pipeline de données

Vous pouvez planifier votre pipeline de données en sélectionnant Planifier dans l'éditeur de pipeline de données.

Vous pouvez également configurer la planification en sélectionnant Paramètres dans le menu Accueil de l'éditeur de pipeline de données.

Nous recommandons les pipelines de données pour une expérience sans code ou à code faible en raison de l'interface graphique utilisateur. Ils sont parfaits pour les workflows de données qui s'exécutent selon une planification ou qui se connectent à différentes sources de données.

Pour en savoir plus sur les pipelines de données, consultez Ingérer des données dans votre entrepôt à l'aide de pipelines de données.

Unité 4: Charger des données avec T-SQL

Type: Contenu

Charger des données avec T-SQL

Les développeurs SQL ou les développeurs citoyens, qui souvent connaissent bien le moteur SQL et sont habiles dans l'utilisation de T-SQL, seront satisfaits de l'entrepôt dans Microsoft Fabric.

Cela est dû au fait que l'entrepôt est alimenté par le même moteur SQL que celui qu'ils connaissent déjà. Il leur permet donc d'effectuer des requêtes et manipulations de données complexes. Ces opérations comprennent le filtrage, le tri, l'agrégation et la jointure de données à partir de différentes tables. De plus, le grand choix de fonctions et d'opérateurs du moteur SQL permet une analyse et des transformations de données sophistiquées au niveau de la base de données.

Utiliser l'instruction COPY

L'instruction COPY sert de méthode principale pour importer des données dans l'entrepôt. Elle facilite l'ingestion efficace de données à partir d'un compte de stockage Azure externe.

Elle propose plus de flexibilité, ce qui vous permet de spécifier le format du fichier source, de désigner un emplacement pour stocker les lignes rejetées pendant le processus d'importation et d'ignorer les lignes d'en-tête, parmi d'autres options configurables.

L'option permettant de stocker séparément les lignes rejetées est utile pour le nettoyage de données et le contrôle de qualité. Elle vous permet d'identifier et d'examiner facilement tous les problèmes liés aux données qui n'ont pas été importées avec succès.

Pour vous connecter à un compte de stockage Azure, vous devez utiliser la signature d'accès partagé (SAP) ou la clé de compte de stockage (SAK, storage account key).

Erreur de descripteur

L'option permettant d'utiliser un autre compte de stockage pour l'emplacement ERRORFILE (REJECTED_ROW_LOCATION) permet une meilleure gestion et débogage des erreurs. Elle facilite l'isolation et l'évaluation des problèmes qui surviennent pendant le processus de chargement des données. ERRORFILE s'applique uniquement au fichier CSV.

Charger plusieurs fichiers

La possibilité de spécifier des caractères génériques et plusieurs fichiers dans le chemin d'accès de l'emplacement de stockage permet à l'instruction COPY de gérer efficacement le chargement en bloc des données. Cela est utile lorsque vous traitez des jeux de données volumineux distribués sur plusieurs fichiers.

Vous pouvez spécifier plusieurs emplacements de fichiers uniquement à partir du même compte de stockage et du même conteneur en utilisant une liste avec des éléments séparés par virgules.

L'exemple suivant montre comment charger un fichier PARQUET.

Vérifiez que tous les fichiers ont la même structure (c'est-à-dire les mêmes colonnes dans le même ordre) et que cette structure correspond à celle de la table cible.

Charger la table à partir d'autres entrepôts et lakehouses

Vous pouvez charger des données à partir de différentes ressources de données dans un espace de travail, telles que d'autres entrepôts et lakehouses.

Pour référencer la ressource de données, veillez à utiliser un nommage en trois parties pour combiner sur ces ressources d'espace de travail des données qui proviennent de tables. Vous pouvez ensuite utiliser CREATE TABLE AS SELECT (CTAS) et INSERT...SELECT pour charger les données dans l'entrepôt.

Dans le cas où l'analyste a besoin de données à partir à la fois d'un entrepôt et d'un lakehouse, il peut utiliser cette fonctionnalité pour combiner ces données. Il peut ensuite charger ces données combinées dans l'entrepôt à des fins d'analyse. Cette fonctionnalité est utile lorsque les données sont distribuées à travers de nombreuses ressources dans un espace de travail.

La requête suivante crée une nouvelle table dans la analysis_warehouse qui combine les données du sales_warehouse et du social_lakehouse en utilisant product_id comme clé commune. La nouvelle table peut alors être utilisée pour des analyses complémentaires.

Tous les entrepôts qui partagent le même espace de travail sont intégrés au même serveur SQL logique. Si vous utilisez des outils clients SQL tels que SQL Server Management Studio, vous pouvez facilement effectuer une requête inter-bases de données comme dans n'importe quelle instance SQL Server.

MyWarehouse et Sales sont des ressources d'entrepôt qui partagent le même espace de travail.

Si vous utilisez l'explorateur d'objets dans l'espace de travail pour interroger vos entrepôts, vous devez les explicitement. Les entrepôts ajoutés seront également visibles à partir de l'éditeur de requête Visual.

Des données peuvent être chargées efficacement dans un entrepôt de Microsoft Fabric via l'instruction COPY, ou à partir d'autres entrepôts et lakehouses au sein du même espace de travail. Cela permet une gestion et une analyse harmonieuses des données.

Unité 5: Charger et transformer des données avec Dataflow Gen2

Type: Contenu

Charger et transformer des données avec Dataflow Gen2

Dataflow Gen2 est la nouvelle génération de flux de données. Il offre une expérience Power Query complète, qui vous guide tout au long de chaque étape de l'importation de données dans votre flux de données. Le processus de création de flux de données a été simplifié, avec une réduction du nombre d'étapes nécessaires.

Vous pouvez utiliser des flux de données dans des pipelines de données pour ingérer des données dans un lakehouse ou un entrepôt, ou pour définir un jeu de données pour un rapport Power BI.

Créer un flux de données

Pour créer un flux de données, accédez à votre espace de travail, puis sélectionnez + Nouveau. Si Dataflow Gen2 n'est pas visible dans la liste, sélectionnez Autres options, puis recherchez Dataflow Gen2 dans la section Data Factory.

Importer des données

Une fois Dataflow Gen2 lancé, de nombreuses options pour charger vos données sont disponibles.

Vous pouvez charger différents types de fichiers en quelques étapes. Vous pouvez par exemple charger un fichier texte ou CSV à partir de votre ordinateur local.

Une fois les données importées, vous pouvez commencer à créer votre flux de données, décider de nettoyer vos données, de les remodeler, de supprimer des colonnes et d'en créer de nouvelles. Toutes les étapes que vous effectuez sont enregistrées.

Transformer des données avec Copilot

Copilot peut être un précieux outil d'aide aux transformations de flux de données. Supposez que vous avez une colonne Gender qui contient « Male » et « Female » et que vous voulez la transformer.

La première étape consiste à activer Copilot dans votre flux de données. Après cela, vous pouvez fournir des instructions spécifiques sur la transformation que vous souhaitez effectuer.

Par exemple, vous pourriez entrer la commande suivante : « Transformer la colonne Gender. Si Male 0, si Female 1. Ensuite la convertir en entier. »

Copilot ajoute automatiquement une nouvelle étape, et vous pouvez toujours l'annuler si vous le souhaitez, ou poursuivre la génération afin d' des transformations supplémentaires.

une destination de données

Avec la fonctionnalité une destination de données, vous pouvez séparer votre logique ETL et votre stockage de destination. Grâce à cette séparation, votre code sera plus propre et plus simple à tenir à jour, et il sera plus facile de modifier le processus ETL ou la configuration de stockage sans affecter l'autre.

Une fois les données transformées, l'étape suivante consiste à une étape de destination. Sous l'onglet Paramètres de requête, sélectionnez + pour une étape de destination dans votre flux de données.

Les options de destination suivantes sont disponibles.

Azure SQL Database

Azure Data Explorer (Kusto)

Azure Synapse Analytics (SQL DW)

Les données chargées dans une destination tel qu'un entrepôt sont facilement accessibles et analysables à l'aide de différents outils. Cela améliore l'accessibilité de vos données, et permet une analyse des données plus flexible et plus complète.

Lorsque vous sélectionnez un entrepôt comme destination, vous pouvez choisir les méthodes de mise à jour suivantes.

: de nouvelles lignes à une table existante.

Remplacer : remplacer tout le contenu d'une table par un nouvel ensemble de données.

Publier un flux de données

Une fois que vous avez choisi votre méthode de mise à jour, la dernière étape consiste à publier votre flux de données.

La publication rend vos transformations et opérations de chargement de données actives, ce qui permet d'exécuter le flux de données soit manuellement, soit d'après une planification. Ce processus encapsule vos opérations ETL dans une unité unique et réutilisable, ce qui simplifie votre workflow de gestion des données.

Toutes les modifications apportées au flux de données prennent effet lors de sa publication. Par conséquent, veillez toujours à publier votre flux de données après avoir apporté des modifications pertinentes.

Unité 6: Exercice : Charger des données dans un entrepôt de données dans Microsoft Fabric

Type: Exercice

Exercice : Charger des données dans un entrepôt de données dans Microsoft Fabric

Vous pouvez maintenant charger des données dans un entrepôt dans Microsoft Fabric.

Dans cet exercice, vous allez apprendre à charger des données dans un entrepôt à l'aide de T-SQL.

Vous avez besoin d'une licence d'évaluation Microsoft Fabric avec la version préliminaire Fabric activée dans votre compte client. Consultez Bien démarrer avec Fabric pour activer votre licence d'évaluation Fabric.

Lancez l'exercice et suivez les instructions.

Unité 7: Évaluation du module

Type: Évaluation

Évaluation du module

Quelles sont les quatre options d’ingestion de données disponibles dans Microsoft Fabric pour le chargement de données dans un entrepôt de données ?

Instruction COPY (Transact-SQL), pipelines de données, flux de données et inter-entrepôts.

Instruction COPY (Transact-SQL), pipelines de données, flux de données et Data Wrangler.

Instruction COPY (Transact-SQL), pipelines de données, flux de données et ingestion multiplateforme.

Quelles sont les sources de données et les formats de fichiers pris en charge pour l’instruction COPY (Transact-SQL) dans Warehouse ?

Azure Data Lake Storage (ADLS) Gen2 et Stockage Blob Azure, avec des formats de fichiers PARQUET et CSV.

Azure Data Lake Storage (ADLS) Gen1 et Azure Blob Storage, avec des formats de fichiers PARQUET et CSV.

Azure Data Lake Storage (ADLS) Gen2 et Azure Blob Storage, avec des formats de fichiers ORC et CSV.

Quelle est la taille de fichier minimale recommandée lors de l’utilisation de données externes sur des fichiers dans Microsoft Fabric ?

Vous devez répondre à toutes les questions avant de vérifier votre travail.

Unité 8: Résumé

Type: Résumé

Il n’existe aucune solution unique pour le chargement de vos données. La meilleure approche dépend des spécificités de vos besoins métier et de la question que vous essayez de répondre.

Lorsqu’il s’agit de charger des données dans un entrepôt de données, il existe plusieurs considérations à prendre en compte.

Pour obtenir une lecture supplémentaire, vous pouvez consulter les URL suivantes :

Créer un entrepôt dans Microsoft Fabric

Ingérer des données dans l’entrepôt

Comparer l’entrepôt et le point de terminaison d’analyse SQL de Lakehouse

Module 3: Interroger un entrepôt de données dans Microsoft Fabric

Unité 1: Présentation

Type: Introduction

Microsoft Fabric Data Warehouse est une plateforme complète pour les données, l'analyse et l'IA (intelligence artificielle). Elle fait référence au processus de stockage, d'organisation et de gestion de grands volumes de données structurées et semi-structurées.

L'entrepôt de données dans Microsoft Fabric est optimisé avec Synapse Analytics. Cela lui confère un ensemble diversifié de fonctionnalités qui facilitent la gestion et l'analyse des données. Il inclut des fonctionnalités avancées de traitement des requêtes et prend en charge les fonctionnalités T-SQL transactionnelles complètes comme un entrepôt de données d'entreprise.

Le processus d'interrogation d'un entrepôt de données est un composant clé du décisionnel (business intelligence). Il implique l'extraction et la manipulation des données stockées dans un entrepôt de données, ce qui permet aux utilisateurs d'extraire des insights précieux à partir de grands volumes de données.

Conception de schémas en étoile

Dans un entrepôt de données classique, les données sont organisées à l'aide d'un schéma, souvent un schéma en étoile ou un schéma en flocon. Le schéma en étoile et le schéma en flocon sont des approches de modélisation matures largement adoptées par les entrepôts de données relationnelles. Il vous oblige à classer les tables en tant que tables de dimensions ou de faits.

Les tables de faits stockent les données mesurables et quantitatives sur une activité, tandis que les tables de dimensions contiennent des attributs descriptifs liés aux données de faits.

Une table de dimension représente le « qui, quoi, où, quand et pourquoi » de votre entrepôt de données. C'est l'arrière-plan descriptif qui donne du contexte aux chiffres bruts que l'on retrouve dans les tables de faits.

Par exemple, si vous exécutez un magasin en ligne, il se peut que votre table de faits contienne les données de vente brutes (c'est-à-dire le nombre d'unités vendues pour chaque produit). Mais sans table de dimension, vous ne pouvez pas savoir qui a acheté ces produits, quand ils ont été achetés, ni où se trouve l'acheteur.

Nous allons explorer différentes façons de connecter et d'interroger un entrepôt de données, et comment il permet de réaliser des tâches d'extraction efficace des informations.

Pour plus d'informations, consultez Comprendre le schéma en étoile et son importance pour Power BI.

Unité 2: Rechercher des données

Type: Contenu

Rechercher des données

Une fois les tables de dimension et de faits d'un entrepôt de données remplies avec des données, vous pouvez utiliser T-SQL pour interroger ces tables et effectuer une analyse des données. La syntaxe Transact-SQL (T-SQL) utilisée pour interroger des tables dans un entrepôt de Fabric ressemble beaucoup à la syntaxe SQL utilisée dans SQL Server ou Azure SQL Database. Cette familiarité permet

une transition facile celles déjà utilisées pour travailler avec ces plateformes.

Agréger des mesures par attributs de dimension

Dans la plupart des scénarios d'analytique données impliquant un entrepôt de données, le processus concerne généralement l'agrégation de mesures numériques à partir de tables de faits basées sur les attributs dans des tables de dimensions. En raison de la structure d'un schéma en flocon ou en étoile, ces requêtes d'agrégation dépendent des clauses JOIN pour lier des tables de faits à des tables de dimensions. Ils utilisent également des clauses et des fonctions d'agrégation GROUP BY pour définir des hiérarchies d'agrégation.

La requête SQL suivante agrège les montants des ventes par année et trimestre à partir des tables FactSales et DimDate dans un entrepôt de données hypothétique :

Les résultats de cette requête ressembleront à la table suivante.

Vous pouvez joindre autant de tables de dimension que nécessaire pour calculer les agrégations dont vous avez besoin. Par exemple, le code suivant étend l'exemple précédent pour décomposer les totaux des ventes trimestrielles par ville en fonction de l'adresse du client indiquée dans la table DimCustomer.

Cette fois, les résultats incluent un total des ventes trimestrielles pour chaque ville.

Jointure dans un schéma en flocon

Dans un schéma en flocon, les dimensions peuvent être partiellement normalisée. Plusieurs jointures sont parfois nécessaires pour lier des tables de faits aux dimensions en flocon. Par exemple, supposons que votre entrepôt de données inclut une table de dimension DimProduct à partir de laquelle les catégories de produits ont été normalisées dans une table DimCategory distincte. Voici un exemple de requête permettant d'agréger des éléments vendus par catégorie de produit :

Les résultats de cette requête incluent le nombre d'articles vendus pour chaque catégorie de produit :

Les clauses JOIN sur FactSales et DimProduct et sur DimProduct et DimCategory sont requises, même si aucun des champs de DimProduct n'est retourné par la requête.

Utilisation de fonctions de classement

Un autre type commun de requête analytique consiste à partitionner les résultats en fonction d'un attribut de dimension et à les classer dans chaque partition. Par exemple, vous pouvez établir un classement annuel des magasins en fonction de leur chiffre d'affaires. Pour atteindre cet objectif, vous avez la possibilité d'utiliser des fonctions de classement Transact-SQL : ROW_NUMBER, RANK, DENSE_RANK, NTILE, etc. Elles vous permettent de partitionner les données en catégories, chacune retournant une valeur spécifique qui indique la position relative de chaque ligne dans la partition :

ROW_NUMBER retourne la position ordinale de la ligne dans la partition. Par exemple, la première ligne porte le numéro 1, la deuxième le numéro 2, etc.

RANK retourne le rang de chaque ligne dans les résultats triés. Par exemple, dans une partition de magasins triés par volume de ventes, le magasin affichant le volume de ventes le plus élevé obtient le rang 1. Si plusieurs magasins présentent les mêmes volumes de ventes, ils reçoivent le même rang ; le rang attribué aux magasins suivants reflète alors le nombre de magasins dont les volumes de ventes sont plus élevés, égalités comprises.

DENSE_RANK classe les lignes d'une partition de la même façon que RANK, à une différence près : lorsque plusieurs lignes possèdent le même rang, le rang des suivantes est établi en ignorant les égalités.

NTILE retourne le centile spécifié dans lequel se situe la ligne. Dans une partition de magasins triés par volume de ventes, NTILE(4) retourne le quartile dans lequel le volume de ventes d'un magasin le place.

Examinons, par exemple, la requête suivante :

La requête partitionne les produits en groupes en fonction de leur catégorie. Dans chaque partition de catégorie, la position relative de chaque produit est déterminée en fonction de son prix catalogue. Voici les résultats possibles de cette requête :

Les résultats de l'exemple illustrent la différence entre RANK et DENSE_RANK. Notez que, dans la catégorie Accessoires, les produits Sprocket et Doodah possèdent le même prix catalogue et sont tous deux classés comme le troisième produit le plus cher. Le produit suivant possède une valeur RANK de 5 (il existe quatre produits plus chers que lui) et une valeur DENSE_RANK de 4 (il existe trois prix plus élevés).

Pour découvrir plus d'informations sur les fonctions de classement, consultez le module Utiliser des fonctions intégrées et GROUP BY dans Transact-SQL.

Récupération d'un nombre approximatif

Bien que l'objectif d'un entrepôt de données consiste principalement à prendre en charge les modèles et rapports de données analytiques pour l'entreprise, les analystes et scientifiques des données doivent souvent effectuer une exploration initiale des données, juste pour déterminer l'échelle et la distribution de base des données.

Par exemple, la requête suivante utilise la fonction COUNT pour récupérer le nombre de ventes de chaque année dans un entrepôt de données hypothétique :

Voici les résultats possibles de cette requête :

En raison du volume de données présentes dans un entrepôt de données, même les requêtes simples visant à compter le nombre d'enregistrements qui répondent à des critères spécifiés peuvent prendre beaucoup de temps. Dans de nombreux cas, un nombre précis n'est pas nécessaire : une estimation approximative suffit. Vous pouvez alors utiliser la fonction APPROX_COUNT_DISTINCT, comme dans l'exemple suivant :

La fonction APPROX_COUNT_DISTINCT utilise un algorithme HyperLogLog pour récupérer un nombre approximatif. Le résultat est garanti présenter un taux d'erreur maximal de 2 % avec une probabilité de 97 %. Voici donc les résultats possibles de cette requête avec les mêmes données hypothétiques qu'auparavant :

Les nombres sont moins précis, mais toujours suffisants pour une comparaison approximative des ventes annuelles. Avec un grand volume de données, la requête utilisant la fonction APPROX_COUNT_DISTINCT est exécutée plus rapidement. La précision réduite peut représenter un compromis acceptable pendant l'exploration des données de base.

Pour plus d'informations, consultez la documentation de la fonction APPROX_COUNT_DISTINCT.

Unité 3: Utiliser l'éditeur de requête SQL

Type: Contenu

Utiliser l'éditeur de requête SQL

L’éditeur de requête SQL dans Microsoft Fabric est un outil polyvalent qui prend en charge Transact-SQL (T-SQL), ce qui vous permet de créer et d’exécuter des scripts pour interroger votre entrepôt de données. Il fournit également des fonctionnalités comme IntelliSense et le débogage pour faciliter le processus de développement.

T-SQL permet aux utilisateurs, aux analystes et aux développeurs de manipuler les données stockées dans un entrepôt.

Lancer un éditeur de requête SQL

L’éditeur de requête SQL ne vous permet pas seulement d’interroger vos données : il permet aussi de les gérer efficacement. Qu’il s’agisse de créer une nouvelle table, d’insérer des lignes dans une table, d’accorder des autorisations sur des objets ou d’exécuter des requêtes complexes, l’éditeur de requêtes SQL est conçu pour faciliter ces tâches.

Pour lancer l’éditeur de requête SQL, vous devez sélectionner votre ressource d’entrepôt dans Mon espace de travail. Cette étape permet de se connecter automatiquement à l’entrepôt de données, sans qu’il soit nécessaire de demander des informations de connexion.

Dans l’Explorateur d’entrepôts, il existe plusieurs façons de créer un éditeur de requête SQL.

Menu Accueil : Sélectionnez Nouvelle requête SQL ou sélectionnez un des modèles disponibles.

Nœud Requêtes : Sélectionnez ... dans Mes requêtes, puis sélectionnez Nouvelle requête SQL.

Onglet Requête : Cette option ouvre un nouvel éditeur de requête.

Quelle que soit la façon dont vous lancez un nouvel éditeur de requête, un nouvel élément de requête est automatiquement créé dans le dossier Mes requêtes de l’Explorateur.

Tout nouvel élément de requête mis à jour est automatiquement enregistré.

Exécuter des requêtes et visualiser les résultats

Pour exécuter une requête, tapez-la dans un nouvel éditeur de requête, puis sélectionnez Exécuter en haut de l’éditeur.

La section Résultats affiche un aperçu, limité à 10 000 lignes.

Exporter les résultats

Les résultats de votre requête peuvent être exportés en tant que fichier Excel. Pour l’exporter, sélectionnez Télécharger le fichier Excel.

Vous devez sélectionner le texte d’une instruction SELECT dans votre requête pour exporter les résultats vers Excel.

De même, l’éditeur de requête offre la possibilité d’enregistrer votre requête mise en évidence sous forme de vue ou de table, chacune ayant des fonctionnalités distinctes :

Enregistrer en tant que vue : Vous permet de créer une vue dans l’entrepôt en utilisant l’instruction SELECT mise en évidence dans l’éditeur de requête.

Enregistrer en tant que table : Crée une table avec les résultats de votre requête.

Pour les deux options, vous devez fournir le schéma et le nom avant de confirmer la création.

Pour en savoir plus sur l’éditeur de requête SQL, consultez Interroger en utilisant l’éditeur de requête SQL.

Unité 4: Explorer l’éditeur de requête visuel

Type: Contenu

Explorer l’éditeur de requête visuel

L’éditeur de requête visuel dans l’entrepôt de données Microsoft Fabric est un outil qui offre une interface intuitive pour la création de requêtes SQL.

Il simplifie le processus d’écriture et de gestion des requêtes, en particulier pour ceux qui ne sont pas à l’aise avec la syntaxe SQL.

Interface graphique : l’éditeur fournit une interface graphique dans laquelle vous pouvez faire glisser et déplacer des tables, et concevoir visuellement vos requêtes. Ceci facilite la compréhension de la structure de votre requête et des relations entre les tables.

Génération automatique de requêtes : lorsque vous concevez votre requête à l’aide de l’interface visuelle, la requête SQL correspondante est générée automatiquement. Ceci vous permet de vous concentrer sur la logique de votre requête plutôt que sur sa syntaxe. Tous les utilisateurs de l’espace de travail peuvent enregistrer leurs requêtes dans le dossier Mes requêtes .

Créer visuellement votre requête

L’éditeur de requête visuel améliore la compréhension intuitive des relations des données en permettant aux utilisateurs d’organiser visuellement des tables et des champs.

En outre, l’éditeur de requête visuel est convivial, même pour ceux qui n’ont peu ou pas d’expérience SQL. Les membres d’une l’équipe non technique peuvent l’utiliser pour extraire des insights précieux de vos données, démocratisant ainsi l’accès aux données au sein de votre organisation et accélérant le processus décisionnel.

De même que l’éditeur de requête SQL, les options Enregistrer sous forme de table et Enregistrer en tant que vue sont également disponibles. Ces fonctionnalités peuvent être utiles pour réutiliser vos requêtes ou pour créer des requêtes plus complexes en fonction des résultats des requêtes précédentes.

Pour en savoir plus sur l’éditeur de requête SQL, consultez Requête à l’aide de l’éditeur de requête visuelle.

Unité 5: Utiliser des outils clients pour interroger un entrepôt

Type: Contenu

Utiliser des outils clients pour interroger un entrepôt

L’utilisation de SQL Server Management Studio (SSMS) pour vous connecter à un entrepôt de données dans Fabric peut faciliter votre workflow, en particulier si vous êtes déjà familiarisé avec l’outil.

Connexion à votre entrepôt de données

SQL Server Management Studio fournit une interface familière à ceux qui utilisent régulièrement SQL Server.

Suivez ces étapes pour vous connecter à un entrepôt de données dans Fabric à partir de SSMS :

Accédez à votre espace de travail Microsoft Fabric.

Sur votre ressource d'entrepôt, sélectionnez plus d'options, puis sélectionnez Copier la chaîne de connexion SQL.

Lancez SQL Server Management Studio et collez la chaîne de connexion SQL copiée dans la zone de nom Serveur, et fournissez les informations d'identification appropriées pour l'authentification.

Après avoir établi une connexion, SSMS affiche l'entrepôt connecté, ainsi que ses tables et vues correspondantes, prêtes pour l'interrogation.

Options d'authentification

Dans Microsoft Fabric, deux types d'utilisateurs authentifiés sont pris en charge via la chaîne de connexion SQL :

Principaux d'utilisateur Microsoft Entra ID (anciennement Azure Active Directory) ou identités d'utilisateur

Principaux de service Microsoft Entra ID (anciennement Azure Active Directory)

L'authentification SQL n'est pas prise en charge.

Tout outil tiers peut utiliser la chaîne de connexion SQL avec des pilotes ODBC ou OLE DB pour se connecter à un point de terminaison d'entrepôt Microsoft Fabric ou SQL Analytics, en utilisant l'authentification Microsoft Entra ID (anciennement Azure Active Directory).

Unité 6: Exercice : Interroger un entrepôt de données dans Microsoft Fabric

Type: Exercice

Exercice : Interroger un entrepôt de données dans Microsoft Fabric

Maintenant, c'est à vous d'interroger les données d'un entrepôt de données dans Microsoft Fabric.

Dans cet exercice, vous allez apprendre à interroger les données à l'aide de l'éditeur SQL dans Microsoft Fabric.

Vous avez besoin d'une licence d'évaluation Microsoft Fabric avec la préversion Fabric activée dans votre locataire. Consultez Bien démarrer avec Fabric pour activer votre licence d'évaluation Fabric.

Lancez l'exercice et suivez les instructions.

Unité 7: Évaluation du module

Type: Évaluation

Évaluation du module

Quel est le langage principal utilisé pour interroger un entrepôt de données ?

Pourquoi l'indexation est-elle importante dans un entrepôt de données ?

Il rend l'entrepôt de données plus organisé.

Il accélère les temps d'extracvtion de données.

Il aide à nettoyer les données.

Quel est l’objectif d’une table de faits dans un entrepôt de données ?

Pour stocker des données brutes.

Pour stocker les résultats des calculs.

Pour stocker les métadonnées relatives à l’entrepôt de données.

Vous devez répondre à toutes les questions avant de vérifier votre travail.

Unité 8: Résumé

Type: Résumé

Découvrir comment interroger un entrepôt de données permet aux utilisateurs d’extraire, d’analyser et d’interpréter de grands volumes de données stockées, en transformant des données brutes en insights significatifs. Ces insights peuvent favoriser la prise de décision stratégique, optimiser les opérations, et découvrir des modèles cachés et des tendances.

Il n’existe pas de solution unique pour interroger vos données. La meilleure approche dépend des spécificités de votre entrepôt de données et de la question à laquelle vous essayez de répondre.

Pour obtenir une lecture supplémentaire, vous pouvez consulter les URL suivantes :

Connectivité aux entrepôts de données dans Microsoft Fabric

Interroger le point de terminaison d’analytique SQL ou l’entrepôt dans Microsoft Fabric

Visualiser des données dans l’aperçu des données de Microsoft Fabric

Module 4: Surveiller un entrepôt de données Microsoft Fabric

Unité 01: Présentation

Type: Introduction

Un entrepôt de données est souvent central pour l'analyse et la création de rapports dans l'entreprise, et il s'agit donc d'une ressource métier critique. Il est important de surveiller un entrepôt de données pour suivre et gérer les coûts, identifier et résoudre les problèmes de performances des requêtes, et obtenir des insights sur la façon dont vos données sont utilisées.

Dans ce module, nous allons explorer certains outils et techniques que vous pouvez utiliser pour surveiller votre entrepôt de données Microsoft Fabric.

Unité 02: Surveiller les métriques de capacité

Type: Contenu

Surveiller les métriques de capacité

Lorsque votre organisation utilise Microsoft Fabric, la licence utilisée pour acheter le service détermine la capacité disponible. Une capacité est un ensemble de ressources que vous pouvez utiliser pour mettre en œuvre les capacités de la Fabric.

Le coût de l'utilisation de Fabric est basé sur des unités de capacité (CU). Chaque action que vous effectuez dans une ressource Fabric peut consommer des CU, pour lesquelles votre organisation est facturée. Il est donc important de pouvoir contrôler l'utilisation des capacités pour planifier et gérer les coûts. Dans les charges de travail des entrepôts de données, les CU sont consommées par les activités de lecture et d'écriture des données, de sorte que les requêtes dans votre entrepôt de données et les opérations de fichiers sous-jacentes vers le stockage OneLake sont un facteur significatif dans le coût de votre solution analytique Fabric.

Utiliser l'application de mesure de la capacité Microsoft Fabric

L'application de mesure de la capacité Microsoft Fabric est une application qu'un administrateur peut installer dans un environnement Fabric et utiliser pour surveiller l'utilisation de la capacité. Pour surveiller l'utilisation de la capacité liée à l'entreposage de données, vous pouvez filtrer l'interface pour afficher uniquement l'activité de l'entrepôt, comme suit :

À l'aide de l'application de mesure de la capacité de l'infrastructure, vous pouvez observer les tendances d'utilisation de la capacité pour déterminer quels processus consomment des CU dans votre environnement Fabric et si une limitation se produit (ce qui indique que vos processus nécessitent plus de capacité que ce qui est disponible dans les contraintes de votre licence de capacité achetée). Grâce à ces informations, vous pouvez optimiser votre licence de capacité en fonction de vos besoins.

Pour plus d'informations sur l'application de mesure de la capacité Microsoft Fabric, reportez-vous aux rapports de facturation et d'utilisation dans Synapse Data Warehouse dans la documentation Microsoft Fabric.

Unité 03: Surveiller l'activité actuelle

Type: Contenu

Surveiller l'activité actuelle

Vous pouvez utiliser des vues de gestion dynamique (DMV) pour récupérer des informations sur l'état actuel de l'entrepôt de données. Plus précisément, les entrepôts de données Microsoft Fabric incluent les DMV suivantes :

sys.dm_exec_connections : Retourne des informations sur une connexion d'entrepôt de données.

sys.dm_exec_sessions : Retourne des informations sur les sessions authentifiées.

sys.dm_exec_requests : Retourne des informations sur les requêtes actives.

Le schéma de ces tableaux est présenté ici :

Interrogation des DMV

Vous pouvez récupérer des informations détaillées sur les activités actuelles dans l'entrepôt de données en interrogeant les DMV dm_exec-* . Par exemple, considérez la requête suivante :

Cette requête retourne des détails sur les demandes actives dans la base de données actuelle, classées selon leur durée actuelle d'exécution. Cela peut servir à identifier les requêtes longues qui pourraient bénéficient d'une optimisation. Un exemple de jeu de résultats à partir de la requête est illustré ici :

Pour plus d'informations sur l'utilisation de DMV, reportez-vous à Surveiller les connexions, les sessions et les demandes à l'aide de DMV dans la documentation Microsoft Fabric.

Unité 04: Surveiller les requêtes

Type: Contenu

Surveiller les requêtes

Les entrepôts de données Microsoft Fabric incluent la fonctionnalité d'aperçus sur les requêtes qui fournit des informations historiques et agrégées sur les requêtes qui ont été exécutées ; vous permettant d'identifier les requêtes fréquemment utilisées ou longues et de vous aider à analyser et à optimiser les performances des requêtes.

Les vues suivantes permettent d'obtenir des aperçus sur les requêtes :

queryinsights.exec_requests_history : Détails de chaque requête SQL complétée.

queryinsights.long_running_queries : Détails du temps d'exécution de requête.

queryinsights.frequently_run_queries : Détails des requêtes les plus fréquentes.

Le schéma de ces tableaux est présenté ici :

Récupération des aperçus sur les requêtes

Les vues d'aperçu des requêtes sont une source utile d'informations sur les requêtes exécutées dans votre entrepôt de données.

Par exemple, considérez la requête suivante :

Cette requête utilise la vuequeryinsights.exec_requests_history pour identifier les requêtes qui ont été exécutées au cours de l'heure précédente.

En fonction des charges de travail simultanées, les requêtes peuvent prendre jusqu'à 15 minutes avant d'être reflétées dans les vues d'ensemble des requêtes.

Vous pouvez obtenir des détails sur les requêtes longues à partir de l'affichage queryinsights.long_running_queries comme suit :

Cette requête identifie les commandes SQL de longue durée qui ont été utilisées plus d'une fois et les renvoie par ordre décroissant de leur temps médian d'exécution.

Pour permettre aux vues de fournir des mesures agrégées, les requêtes avec prédicats sont paramétrées et considérées comme une seule et même requête si les déclarations paramétrées correspondent exactement. Par exemple, les requêtes suivantes seraient considérées comme la même commande :

SELECT * FROM sales WHERE orderdate > '01/01/2023'

SELECT * FROM sales WHERE orderdate > '12/31/2021'

Pour rechercher des requêtes couramment utilisées, vous pouvez utiliser la vue queryinsights.frequently_run_queries, comme suit :

Cette requête renvoie les détails des exécutions réussies et des échecs pour les commandes fréquemment exécutées.

Pour plus d'informations sur l'utilisation des aperçus de requête, reportez-vous à l'entreposage de données Query Insights dans la documentation de Microsoft Fabric.

Unité 05: Exercice : surveillance d'un entrepôt de données dans Microsoft Fabric

Type: Exercice

Exercice : surveillance d'un entrepôt de données dans Microsoft Fabric

Il est maintenant temps d'essayer de surveiller un entrepôt de données Microsoft Fabric.

Vous avez besoin d'une licence d'évaluation Microsoft Fabric avec la préversion Fabric activée dans votre locataire. Consultez Bien démarrer avec Fabric pour activer votre licence d'évaluation Fabric. Lancez l'exercice et suivez les instructions.

Exercice : surveillance d'un entrepôt de données dans Microsoft Fabric Il est maintenant temps d'essayer de surveiller un entrepôt de données Microsoft Fabric. Remarque Vous avez besoin d'une licence d'évaluation Microsoft Fabric avec la préversion Fabric activée dans votre locataire. Consultez Bien démarrer avec Fabric pour activer votre licence d'évaluation Fabric. Lancez l'exercice et suivez les instructions. Commentaires Yes No

Unité 06: Évaluation du module

Type: Évaluation

Évaluation du module

Vérifier vos connaissances

Vous souhaitez surveiller la consommation d'unités de capacité dans votre entrepôt de données Fabric. Quels outils devez-vous utiliser ?

Microsoft Azure Monitor

Application Métriques de capacité Microsoft Fabric

Microsoft Azure Data Studio.

Quelle vue de gestion dynamique fournit des détails sur les commandes SQL en cours d'exécution dans l'entrepôt de données ?

sys.dm_exec_requests

sys.dm_exec_connections

sys.dm_exec_sessions

Quelle vue devez-vous utiliser pour identifier les commandes couramment exécutées dans votre entrepôt de données ?

queryinsights.exec_requests_history

queryinsights.long_running_queries

queryinsights.frequently_run_queries

Vous devez répondre à toutes les questions avant de vérifier votre travail.

Unité 07: Résumé

Type: Résumé

Les entrepôts de données Microsoft Fabric constituent un actif important pour toute organisation, et il est important de les surveiller pour suivre et gérer les coûts, identifier et résoudre les problèmes de performance des requêtes, et obtenir des informations sur la façon dont les données sont utilisées.

Ce module a exploré les moyens suivants pour surveiller un entrepôt de données :

Application de mesure de la capacité Microsoft Fabric

Vues de gestion dynamique

Vue d'ensemble des requêtes

Module 5: Sécuriser un entrepôt de données Microsoft Fabric

Unité 1: Présentation

Type: Introduction

Microsoft Fabric Data Warehouse est une plateforme complète pour les données, l'analyse et l'IA (intelligence artificielle). Elle fait référence au processus de stockage, d'organisation et de gestion de grands volumes de données structurées et semi-structurées.

Dans un entrepôt, les administrateurs ont accès à une suite de technologies destinées à protéger les informations sensibles. Ces mesures de sécurité sont capables de sécuriser ou de masquer les données des utilisateurs ou des rôles sans autorisation appropriée, ce qui garantit la protection des données sur les points de terminaison de l'analytique SQL et de l'entrepôt. Cela garantit une expérience utilisateur fluide et sécurisée, sans qu'il soit nécessaire de modifier les applications existantes.

Comprendre les fonctionnalités de sécurité

Les ingénieurs Données, qui souvent connaissent bien le moteur SQL et sont experts dans l'utilisation de T-SQL, apprécient la facilité d'utilisation des entrepôts dans Microsoft Fabric.

C'est parce que les entrepôts sont alimentés par le même moteur SQL que celui qu'ils connaissent déjà, ce qui leur permet d'effectuer des requêtes et manipulations de données complexes. La vaste gamme de fonctionnalités de sécurité du moteur SQL permet d'avoir un mécanisme de sécurité sophistiqué au niveau de l'entrepôt.

Rôles d'espaces de travail : conçus pour fournir différents niveaux d'accès et de contrôle dans l'espace de travail. Vous pouvez attribuer des utilisateurs aux différents rôles d'espace de travail comme Administrateur, Membre, Contributeur et Viewer. Ces rôles sont essentiels pour maintenir la sécurité et l'efficacité des opérations d'entreposage de données au sein d'une organisation.

Autorisations d'élément : les entrepôts individuels peuvent avoir des autorisations d'élément attribuées directement. L'objectif principal de l'attribution de ces autorisations est de faciliter le partage de l'entrepôt pour une utilisation en aval.

Sécurité de protection des données : pour un contrôle plus précis, vous pouvez utiliser T-SQL pour accorder des autorisations spécifiques aux utilisateurs. L'entrepôt prend en charge une gamme de fonctionnalités de protection des données qui permettent aux administrateurs de protéger les données sensibles contre les accès non autorisés. Cela inclut la sécurité au niveau objet pour les objets de base de données, la sécurité au niveau colonne pour les colonnes de table, la sécurité au niveau ligne pour les lignes de table en utilisant des filtres de clause WHERE et Dynamic Data Masking pour masquer les données sensibles comme les adresses e-mail. Ces fonctionnalités garantissent la protection des données sur les points de terminaison des entrepôts et de l'analytique SQL sans avoir besoin de modifier les applications.

Dans les unités suivantes, nous explorons les différentes façons d'activer la sécurité dans un entrepôt et comment ces méthodes peuvent faciliter les tâches liées à la protection de la charge de travail de votre entrepôt de données.

Unité 2: Découverte du masquage dynamique des données

Type: Contenu

Découverte du masquage dynamique des données

Dynamic Data Masking (DDM) est une fonctionnalité de sécurité qui limite l'exposition des données aux utilisateurs non privilégiés en masquant les informations sensibles.

Le masquage dynamique des données offre plusieurs avantages clés qui améliorent la sécurité et la facilité de gestion de vos données. L'un des principaux avantages est sa fonctionnalité de masquage en temps réel. Lors de l'interrogation de données sensibles, DDM applique un masquage dynamique en temps réel. Ce processus signifie que les données réelles ne sont jamais exposées à des utilisateurs non autorisés, ce qui améliore la sécurité de vos données. En outre, DDM est simple à implémenter. Il ne nécessite pas de codage complexe, ce qui le rend accessible aux utilisateurs de tous les niveaux de compétence.

Un autre avantage de DDM est que les données de la base de données ne sont pas modifiées lorsque DDM est appliqué. Au lieu de cela, les règles de masquage sont appliquées aux résultats de la requête. Cet avantage signifie que les données réelles restent intactes et sécurisées, tandis que les utilisateurs non privilégiés voient uniquement une version masquée des données.

Définir une règle de masquage

Le masquage dynamique des données, qui est configuré au niveau de la colonne, offre une suite de fonctionnalités, notamment des fonctionnalités de masquage complètes et partielles, ainsi qu'une fonction de masquage aléatoire conçue pour les données numériques.

Les paramètres prefix_padding et suffix_padding de la fonction partial() spécifient le nombre de caractères à exposer au début et à la fin de la chaîne, et le paramètre padding_string spécifie la chaîne à utiliser pour masquer les caractères restants.

Les paramètres low et high de la fonction random() spécifient la plage de nombres aléatoires à générer.

Ces types de masquage permettent d'empêcher l'affichage non autorisé de données sensibles en permettant aux administrateurs de spécifier la quantité de données sensibles à révéler, avec un effet minimal sur la couche application. Elles sont appliquées aux résultats des requêtes, de sorte que les données de la base de données ne sont pas modifiées. Cette approche permet à de nombreuses applications de masquer les données sensibles sans modifier les requêtes existantes.

Configurer le masquage des données

Prenons un exemple d'entrepôt qui stocke des informations client. L'entrepôt contient une table Customer avec des champs tels que CustomerName, Email, PhoneNumber et CreditCardNumber.

Pour appliquer le masquage des données sur les colonnes CustomerName, Email, PhoneNumber et CreditCardNumber, exécutez la commande suivante :

Afficher les résultats masqués

Sans Dynamic Data Masking, si un utilisateur non privilégié exécute une requête pour récupérer les détails du client, il peut voir quelque chose comme suit :

Toutefois, avec DDM appliqué aux champs Email, PhoneNumber et CreditCardNumber, la même requête retourne :

Comme vous pouvez le voir, les données sensibles sont masquées pour l'utilisateur non privilégié, ce qui améliore la sécurité de vos données. Ce scénario est un exemple de base du fonctionnement de Dynamic Data Masking. Il permet de s'assurer que les données sensibles ne sont pas exposées aux utilisateurs qui n'ont pas les privilèges nécessaires pour les afficher.

Les utilisateurs non privilégiés disposant d'autorisations de requête peuvent déduire les données réelles, car les données ne sont pas physiquement masquées.

DDM doit être utilisé dans le cadre d'une stratégie complète de sécurité des données qui inclut une gestion appropriée de la sécurité au niveau objet avec des autorisations granulaires SQL et l'adhésion au principe des autorisations minimales requises.

Unité 3: Implémenter la sécurité au niveau des lignes

Type: Contenu

Implémenter la sécurité au niveau des lignes

La Sécurité au niveau des lignes (RLS) est une fonctionnalité qui fournit un contrôle granulaire sur l'accès aux lignes d'une table en fonction de l'appartenance à un groupe ou du contexte d'exécution.

Par exemple, dans une plateforme d'e-commerce, vous pouvez faire en sorte que les vendeurs aient accès seulement aux lignes de commande associées à leurs propres produits. De cette façon, chaque vendeur peut gérer ses commandes indépendamment, tout en conservant la confidentialité des informations sur les commandes d'autres vendeurs.

Si vous avez de l'expérience avec SQL Server, vous pouvez constater que la sécurité au niveau des lignes partage avec celui-ci des caractéristiques et des fonctionnalités similaires.

Protéger vos données

La sécurité au niveau des lignes (RLS) fonctionne en associant une fonction, appelée prédicat de sécurité, à une table. Cette fonction est définie pour retourner true ou false en fonction de certaines conditions, impliquant généralement les valeurs d'une ou plusieurs colonnes de la table. Quand un utilisateur tente d'accéder aux données de la table, la fonction de prédicat de sécurité est appelée. Si la fonction retourne true, la ligne est accessible à l'utilisateur ; si elle retourne false, la ligne n'apparaît pas dans les résultats de la requête.

Selon les besoins de l'entreprise, la sécurité au niveau des lignes peut être aussi simple que WHERE CustomerId = 29 ou plus complexe si nécessaire.

Ce processus est transparent pour l'utilisateur et est appliqué automatiquement par SQL Server, ce qui garantit une application cohérente des règles de sécurité.

La sécurité au niveau des lignes est implémentée en deux étapes principales :

Prédicats de filtrage – c'est une fonction table qui filtre les résultats en fonction du prédicat défini. Accès Définition CHOISIR Impossible d'afficher les lignes filtrées. MISE À JOUR L'utilisateur ne peut pas mettre à jour des lignes qui sont filtrées. SUPPRIMER Impossible de supprimer des lignes filtrées. INSÉRER Non applicable.

Prédicats de filtrage – c'est une fonction table qui filtre les résultats en fonction du prédicat défini.

Stratégie de sécurité – c'est une stratégie de sécurité qui appelle une fonction table pour protéger l'accès aux lignes d'une table.

Comme le contrôle d'accès est configuré et appliqué au niveau de l'entrepôt, les modifications à apporter à l'application, si elles sont nécessaires, sont minimales. En outre, les utilisateurs peuvent accéder directement aux tables et interroger leurs propres données.

Configurer la sécurité au niveau des lignes

Les commandes T-SQL ci-dessous montrent comment utiliser la sécurité au niveau des lignes dans un scénario où l'accès utilisateur est séparé par le locataire :

Ensuite, nous créons un schéma et une fonction table, et nous accordons à l'utilisateur l'accès à la nouvelle fonction. Le prédicat WHERE @TenantName = USER_NAME() OR USER_NAME() = 'TenantAdmin' évalue si le nom d'utilisateur qui exécute la requête correspond aux valeurs de colonne TenantName.

L'utilisateur tenantAdmin@contoso.com doit voir toutes les lignes. Les utilisateurs tenant1@contoso.com à tenant5@contoso.com ne doivent voir que leurs propres lignes.

Si vous modifiez la stratégie de sécurité avec WITH (STATE = OFF);, vous constatez que les utilisateurs peuvent voir toutes les lignes.

Il existe un risque de fuite d'informations si un attaquant écrit une requête avec une clause WHERE spécialement conçue et, par exemple, une erreur de division par zéro, pour forcer une exception si la condition WHERE est vraie. Il s'agit d'une attaque par canal auxiliaire.

Explorer des cas d'usage

La sécurité au niveau des lignes est idéale pour de nombreux scénarios, notamment :

Lorsque vous devez isoler l'accès départemental au niveau de la ligne.

Lorsque vous devez restreindre l'accès aux données de clients aux seules données relatives à leur entreprise.

Lorsque vous devez restreindre l'accès à des fins de conformité.

Appliquer les meilleures pratiques

Voici quelques bonnes pratiques à prendre en compte lors de l'implémentation de la sécurité au niveau des lignes :

Il est recommandé de créer un schéma distinct pour les fonctions de prédicat et les stratégies de sécurité.

Dans la mesure du possible, évitez les conversions de types dans les fonctions de prédicat.

Pour optimiser les performances, évitez d'utiliser des jointures de table excessives et une récursivité dans les fonctions de prédicat.

Unité 4: Implémenter la sécurité au niveau des colonnes

Type: Contenu

Implémenter la sécurité au niveau des colonnes

La sécurité au niveau des colonnes (CLS) vous permet de restreindre l'accès aux colonnes afin de protéger les données sensibles. Elle fournit un contrôle précis de l'accès aux éléments de données spécifiques, ce qui améliore la sécurité globale de votre entrepôt de données.

Sécuriser les données sensibles

Prenons un exemple pratique de sécurité au niveau des colonnes (CLS) dans le secteur de la santé. Supposons que nous avons une table nommée Patients avec les colonnes suivantes : PatientID, Name, Address, DateOfBirth et MedicalHistory.

La colonne MedicalHistory contient des informations sensibles sur la santé des patients. Conformément aux réglementations en matière de soins de santé et aux lois sur la protection des données personnelles, seul le personnel médical autorisé, dont les médecins et les infirmières, doit pouvoir accéder à ces informations.

Voici une façon implémenter la sécurité au niveau des colonnes dans ce scénario :

Identifiez les colonnes sensibles : dans ce cas, la MedicalHistory colonne est identifiée comme contenant des données sensibles.

Définir des rôles d'accès : définissez des rôles tels que Doctor et Nurse qui sont autorisés à accéder à la MedicalHistory colonne. L'accès à cette colonne peut être restreint pour d'autres rôles, tels que Receptionist ou Patient.

Attribuer des rôles aux utilisateurs : attribuez les rôles appropriés à chaque utilisateur de l'entrepôt. Par exemple, le rôle DrSmith peut être affecté à l'utilisateur Doctor, tandis que le rôle JohnDoe peut être affecté à l'utilisateur Patient.

Implémenter le contrôle d'accès : restreindre l'accès à la MedicalHistory colonne en fonction du rôle de l'utilisateur.

La sécurité au niveau des colonnes peut vous aider à garantir que les informations sensibles sur la santé peuvent uniquement être consultées par des personnes autorisées grâce à la protection des données personnelles des patients et le respect des réglementations en matière de soins de santé.

Configurer la sécurité au niveau de la colonne

Dans le scénario que nous venons de découvrir, la syntaxe pour implémenter la sécurité au niveau des colonnes peut revêtir la forme suivante :

Dans cet exemple, nous créons d'abord les rôles Doctor, Nurse, Receptionist et Patient. Nous accordons ensuite à tous les rôles les autorisations SELECT sur toutes les colonnes de la table Patients. Enfin, nous refusons les autorisations SELECT sur la colonne MedicalHistory pour les rôles Receptionist et Patient. Cela garantit que seuls les utilisateurs dotés du rôle Doctor ou Nurse peuvent accéder à la colonne MedicalHistory.

Comprendre les avantages

Dans la sécurité d'entrepôt, deux des techniques les plus utilisées sont la sécurité et les vues au niveau des colonnes. Ces deux méthodes permettent de restreindre l'accès aux données sensibles, mais elles le font de différentes manières et proposent différents avantages. Le tableau suivant établit une analyse comparative de ces deux techniques sur différents aspects, dont la granularité du contrôle d'accès, la maintenance, les performances, la transparence et la flexibilité.

Cette comparaison peut vous aider à comprendre les points forts et les faiblesses de chaque méthode. Elle peut également vous aider à choisir l'approche la plus appropriée pour vos exigences d'application spécifiques.

Le choix entre la sécurité ou les vues au niveau des colonnes dépend des exigences spécifiques de votre application. Veillez à tester systématiquement les modifications de sécurité dans un environnement sécurisé avant de les appliquer à un entrepôt de production.

Unité 5: Configurer des autorisations granulaires SQL à l'aide de T-SQL

Type: Contenu

Configurer des autorisations granulaires SQL à l'aide de T-SQL

Si vous êtes familiarisé avec les bases de données relationnelles et les entrepôts d'entreprise, vous savez qu'il existe quatre autorisations fondamentales régissant les opérations de langage de manipulation de données (DML). Ces autorisations, à savoir SELECT, INSERT, UPDATEet DELETE, sont universellement applicables sur toutes les plateformes de base de données.

Toutes ces autorisations peuvent être accordées, révoquées ou refusées sur les tables et les vues. Si une autorisation est accordée avec l'instruction GRANT, l'autorisation est donnée à l'utilisateur ou au rôle référencé dans l'instruction GRANT. Les utilisateurs peuvent également se voir refuser des autorisations avec la commande DENY. Si un utilisateur se voit accorder puis refuser une même autorisation, DENY l'emporte toujours sur l'octroi, donc l'accès à l'objet spécifique est refusé à l'utilisateur.

Autorisations des tables et des vues

Les tables et les vues représentent les objets sur lesquels des autorisations peuvent être accordées au sein d'un entrepôt. Dans ces tables et ces vues, vous pouvez aussi limiter les colonnes qui sont accessibles à un principal de sécurité donné.

Autorisations des fonctions et des procédures stockées

Comme les tables et les vues, les fonctions et les procédures stockées disposent de plusieurs autorisations qui peuvent être accordées ou refusées.

Principe du privilège minimum

L'idée de base du principe des privilèges minimum est que les utilisateurs et les applications doivent se voir accorder uniquement les autorisations nécessaires pour effectuer la tâche. Les applications doivent disposer uniquement des autorisations nécessaires pour effectuer la tâche en cours.

Par exemple, si une application accède à toutes les données via des procédures stockées, elle doit uniquement avoir l'autorisation d'exécuter les procédures stockées, sans accès aux tables.

Le SQL dynamique est un concept selon lequel une requête est créée programmatiquement. Le SQL dynamique permet de générer des instructions T-SQL dans une procédure stockée ou dans une requête elle-même. Un exemple simple est illustré ci-dessous.

Dans cet exemple, @tableName est le paramètre que vous pouvez remplacer par le nom de la table à inspecter. La fonction QUOTENAME est utilisée pour citer en toute sécurité le nom de la table, ce qui empêche les attaques par injection SQL. La procédure stockée sp_executesql est ensuite utilisée pour exécuter la requête générée dynamiquement.

Notez qu'il s'agit d'un exemple simple et que les tâches d'entrepôt de données réelles peuvent nécessiter des requêtes SQL dynamiques plus complexes. Soyez toujours prudent lors de l'utilisation de SQL dynamique en raison du risque d'attaques par injection SQL. Utilisez toujours des méthodes de paramétrisation comme sp_executesql ou QUOTENAME pour nettoyer les entrées.

Unité 6: Exercice : Sécuriser un entrepôt dans Microsoft Fabric

Type: Exercice

Exercice : Sécuriser un entrepôt dans Microsoft Fabric

À présent, vous allez pouvoir sécuriser un entrepôt dans Microsoft Fabric.

Dans cet exercice, vous allez apprendre à sécuriser un entrepôt en utilisant les concepts découverts dans ce module.

Vous avez besoin d'une licence d'évaluation Microsoft Fabric avec la préversion Fabric activée dans votre locataire. Consultez Bien démarrer avec Fabric pour activer votre licence d'évaluation Fabric.

Pour effectuer les exercices de ce labo, vous avez besoin de deux identités d'utilisateur. Si vous ne parvenez pas à créer un deuxième utilisateur, vous pouvez toujours effectuer l'exercice en utilisant votre compte d'utilisateur, mais vous ne pourrez pas faire l'expérience de ce qu'un utilisateur moins privilégié voit lorsque l'accès à des fonctionnalités spécifiques lui est accordé ou restreint.

Comment créer un deuxième utilisateur pour cet exercice

Si vous faites partie d'une organisation disposant d'un locataire Entra ou Microsoft 365 :

Collaborez avec votre administrateur d'identité pour créer le deuxième utilisateur dans Entra ou le Centre d'administration Microsoft 365.

Si vous n'êtes pas membre d'une organisation avec un locataire Entra ou Microsoft 365 :

Vous ne pouvez pas vous inscrire à un essai de Fabric avec votre adresse e-mail personnelle. Vous pouvez vous inscrire à un essai de Microsoft 365, et un locataire d'organisation est créé. Vous devenez alors administrateur d'utilisateurs et de facturation du locataire et pourrez créer des utilisateurs.

Créez le deuxième utilisateur dans le Centre d'administration Microsoft 365. Consultez : des utilisateurs

Activez l'essai de Fabric en vous reportant à l'article Bien démarrer avec Fabric.

Lancez l'exercice et suivez les instructions.

Unité 7: Évaluation du module

Type: Évaluation

Évaluation du module

Quel est le principal avantage de DDM (Dynamic Data Masking) ?

Il limite l'exposition des données en masquant les informations sensibles en temps réel.

Il change les données réelles de la base de données.

Son implémentation nécessite l'écriture d'un code complexe.

À quoi sert une fonction de prédicat de sécurité dans le cadre de la sécurité SNL (sécurité au niveau des lignes) ?

Elle détermine si une ligne est accessible à un utilisateur en fonction de certaines conditions.

Elle permet les conversions de type dans les fonctions de prédicat.

Elle permet aux utilisateurs d'exécuter des requêtes ad hoc.

Que se passe-t-il quand un utilisateur se voit octroyer une autorisation, puis se voit refuser cette même autorisation dans un entrepôt ?

L'instruction GRANT remplace l'instruction DENY, et l'utilisateur a accès à l'objet spécifique.

L’instruction DENY remplace toujours l’instruction GRANT, l’utilisateur se voit donc refuser l’accès à l’objet spécifique.

L’utilisateur dispose des deux autorisations, ce qui provoque un conflit.

Vous devez répondre à toutes les questions avant de vérifier votre travail.

Unité 8: Résumé

Type: Résumé

Dans ce module, vous avez découvert Dynamic Data Masking (DDM), la sécurité au niveau ligne (RLS), la sécurité au niveau colonne (CLS) et les autorisations SQL granulaires dans les entrepôts Fabric.

Les principaux points de départ de ce module incluent la compréhension de la façon dont DDM, RLS et CLS fonctionnent et leurs cas d’usage. DDM fonctionne au niveau colonne, offrant des types de masquage d’e-mail, de texte personnalisé et aléatoire complets. RLS fonctionne en associant une fonction de prédicat de sécurité à une table. CLS peut être implémenté en identifiant les colonnes sensibles, en définissant des rôles d’accès, en affectant des rôles aux utilisateurs et en implémentant le contrôle d’accès. De plus, vous avez appris le principe des privilèges minimum qui suggère que les utilisateurs et les applications doivent se voir accorder uniquement les autorisations nécessaires pour effectuer leurs tâches.

Pour obtenir une lecture supplémentaire, vous pouvez consulter les URL suivantes :

Créer un entrepôt dans Microsoft Fabric

Sécurité de l’entreposage de données dans Microsoft Fabric

Partager votre entrepôt et gérer les autorisations

Parcours 3: Utiliser des modèles sémantiques dans Microsoft Fabric

Module 1: Create DAX calculations in semantic models

Unité 1: Introduction

Type: Introduction

When you first create a semantic model by applying Power Query queries, the model consists of tables and columns. At this point, the model probably isn't ready for use. That's because you might need to create or adjust model relationships, create other tables or columns, add hierarchies, or set model properties. You might also identify the need for calculations to summarize the model data, especially when the requirement can't be achieved by summarizing a single column. For example, you might want to calculate year-to-date (YTD) sales revenue, which requires special time filters.

You can use Data Analysis Expressions (DAX) to add three types of calculations to your semantic model to complete your model design:

Calculated tables, which can be useful to create date tables or role-playing dimensions, or to enable what-if analysis.

Calculated columns, which can be added to tables in your model.

Measures, which achieve summarization over model data.

In this module, you learn why and how to create each of these calculation types.

Unité 2: Create calculated tables

Type: Contenu

Create calculated tables

A calculated table is created with a DAX formula that returns a table object, allowing you to duplicate or transform existing model data to produce a new table.

Duplicate a table

A common design challenge in data modeling is handling multiple relationships between tables. For example, the Sales table might have three relationships to the Date table, as shown in the following diagram:

The Sales table stores sales data by order date, ship date, and due date, resulting in three relationships to the Date table. However, only one relationship can be active at a time, as indicated by a solid line in the diagram. The other relationships are inactive and shown as dashed lines.

In our example, the active relationship filters the OrderDateKey column in the Sales table, so filters applied to the Date table affect sales by order date only.

To enable filtering sales by ship date, a new table can be created by duplicating the Date table with the following formula:

This formula creates a new table named Ship Date with the same columns and rows as the original Date table. When the Date table is refreshed, the Ship Date table is also recalculated to remain synchronized. Now you can create an active relationship between Ship Date and Sales tables to allow filtering by ship dates too.

Configure duplicate tables

When you create a calculated table, you need to apply any custom configurations to the new duplicate table you want carried over. For instance, it's a good idea to rename columns so that they better describe their purpose.

In our example, the Fiscal Year column in the Ship Date table can be renamed as Ship Fiscal Year. The Ship Full Date column should be sorted by the Ship Date column and the MonthKey column can be hidden to improve sorting and reporting.

Calculated tables are useful to work in scenarios when multiple relationships between two tables exist, as previously described. However, calculated tables increase the model's storage size and can prolong data refresh times, especially when they depend on other tables.

While duplicate tables solve this challenge, there are other more performant solutions. We cover this concept again when discussing measures later in this module.

Create a date table

A good use case for calculated tables is to add a date table to your model. Date tables are required to apply special time filters known as time intelligence.

You can create a calculated date table using the CALENDARAUTO function. The CALENDARAUTO function scans all date and date/time columns in the model to determine the earliest and latest dates, then generates a complete set of dates that span all years in the data. The argument specifies the last month of the financial year; for example, passing 6 sets June as the year end.

The resulting Due Date table contains a single column of dates. The following image shows the Due Date table in data view, with dates sorted from earliest to latest:

The CALENDAR function can also be used to create a date table by specifying a start and end date, either as static values or as expressions based on model data.

If the earliest date in the model is October 15, 2021, and the latest is June 15, 2022, the function returns dates from July 1, 2021, to June 30, 2022. This ensures the table includes complete years, which is required for marking a date table.

Mark as a date table

Once you create a date table, you need to Mark it as a date table in Power BI Desktop. This setting allows you to use time intelligence functions in DAX calculations. When you specify your own date table, Power BI Desktop performs the following validations of that column and its data, to ensure that the data contains:

Contiguous date values (from beginning to end).

Same timestamp across each value for Date/Time data types.

This setting applies to any date tables, either imported or created in Power Query or calculated tables.

You must either use a custom date table or use the built-in auto/date time feature in Power BI to use time intelligence. The auto/date time feature has limited values and can't be customized, which is one

reason to consider using a custom date table.

Unité 3: Create calculated columns

Type: Contenu

Create calculated columns

Occasionally, you need more columns than exist in your data. Ideally, you add these columns directly to the data source, which supports easier maintenance and allows the column logic to be reused in other models or reports. However, if you need to add the column once connected to the data in Power BI, you can either create a custom column in Power Query Editor or add a calculated column to the semantic model. No matter how you add the column, the result is the same from a report user's perspective.

In general, custom columns in Power Query are preferred because they're loaded into the model in a more compact and optimal way. Calculated columns are recommended only when adding columns to a calculated table or when the formula:

Depends on summarized model data.

Requires specialized modeling functions available only in DAX, such as RELATED, RELATEDTABLE, or functions for parent-child hierarchies.

A DAX formula can be used to add a calculated column to any table in the model. The formula must return a single value (scalar) for each row.

Calculated columns in import models increase storage size and can prolong data refresh times, especially when they depend on other tables that are refreshed. Therefore, be cautious when using too many calculated columns and consider if a measure can be used instead.

In the following examples, several calculated columns are created to support fiscal analysis to demonstrate the versatility of calculated columns.

This column determines the fiscal year for each date. The fiscal year starts in July, so dates from July to December are assigned to the next calendar year. The formula concatenates "FY" with the year, incrementing the year by one for dates in the second half of the year.

The following steps describe how Microsoft Power BI evaluates the calculated column formula:

The addition operator (+) is evaluated before the text concatenation operator (&).

The YEAR function returns the whole number value of the due date year.

The IF function returns the value when the due date month number is 7-12 (July to December); otherwise, it returns BLANK. (For example, because the Adventure Works financial year is July-June, the last six months of the calendar year will use the next calendar year as their financial year.)

The year value is added to the value that is returned by the IF function, which is the value one or BLANK. If the value is BLANK, it's implicitly converted to zero (0) to allow the addition to produce the fiscal year value.

The literal text value "FY" concatenated with the fiscal year value, which is implicitly converted to text.

This column assigns a fiscal quarter to each date, based on the fiscal year structure where Quarter 1 is July–September. The formula appends Q and the quarter number to the fiscal year label.

The MonthKey formula multiplies the due date year by the value 100 and then adds the month number of the due date. It produces a numeric value that can be used to sort the Due Month text values in chronological order.

The FORMAT function converts the Due Date column value to text by using a format string. In this case, the format string produces a label that describes the year, abbreviated month name, and day:

Many user-defined date/time formats exist. For more information, see Custom date and time formats for the FORMAT function.

After these additions, the Due Date table contains six columns: the original date column and five calculated columns. These columns support both time intelligence functions and readability for report consumers.

Understand row context

Power BI evaluates a calculated column's formula for each row in the table, which is called row context. Row context just means "the current row." For example, here's the Due Fiscal Year calculated column:

When Power BI runs this formula, 'Due Date'[Due Date] gives the value from the current row. If you've used Excel tables, this idea might feel familiar.

Row context only applies to the current table. If you need values from another table, you have two main options:

If a relationship exists between the two tables, use the RELATED or RELATEDTABLE function. RELATED gets a value from the one-side of a relationship. RELATEDTABLE gets a table of values from the many-side.

If there's no relationship, use the LOOKUPVALUE function.

Try to use RELATED when you can. It usually works faster than LOOKUPVALUE because of how Power BI stores and indexes data.

Here's an example. This formula adds a Discount Amount column to the Sales table:

Power BI calculates this formula for each row in the Sales table. It gets Order Quantity and Sales Amount from the current row. To get List Price from the Product table, it uses the RELATED function to find the right value for each sale.

Row context always applies when Power BI evaluates calculated column formulas. It also comes into play with iterator functions, which let you create more advanced summaries. You learn about iterator functions later in this module.

Unité 4: Understand implicit measures

Type: Contenu

Understand implicit measures

Measures in Power BI models are either implicit or explicit. Implicit measures are automatic behaviors that let visuals summarize column data. Explicit measures, often called measures, are calculations you add to your model. This section explains how implicit measures work and how you can use them.

Identify implicit measures

As a data modeler, you control how a column summarizes by setting the Summarization property. You can choose Don't summarize or select a specific aggregation function. If you set a column to Don't summarize, the sigma symbol disappears in the Data pane.

In the Data pane, a column with the sigma symbol (∑) shows two facts:

The column is numeric.

Values are summarized when used in a visual that supports summarization.

The following image shows the Sales table with implicit measures, a calculated column, and one column that can't be summarized.

Notice in the example with the Sales table, if you add the Sales Amount field from the Sales table to a matrix visual that groups by fiscal year and month, Power BI summarizes the values implicitly. The Sum aggregation function is selected by default.

If you add the Unit Price field to the matrix visual, Power BI uses Average as the default summarization, because summing unit prices doesn’t make sense since they’re rates, not totals.

The default summarization is now set to Average (the modeler knows that it's inappropriate to sum unit price values together because they're rates, which are non-additive).

Implicit measures allow the report author to start with a default summarization technique and lets them modify it to suit their visual requirements. Numeric columns support the widest range of aggregation functions to choose from, including:

Count (Distinct)

Standard deviation

Summarize non-numeric columns

Non-numeric columns like text, dates, and boolean (true/false) values can also be summarized in your visuals. While the sigma symbol (∑) only appears next to numeric fields in the Data pane, these columns are still being aggregated.

Text columns: First, Last, Count, Count (Distinct)

Date columns: Earliest, Latest, Count, Count

This flexibility is useful when you want to answer questions such as:

"How many unique products are there?" (Count distinct on a text column)

"What was the earliest order date?" (Earliest on a date column)

"How many orders were marked as complete?" (Count on a boolean column)

You can choose the aggregation option that best fits your analysis when you add a non-numeric field to your visual.

Considerations for implicit measures

Implicit measures are easy to use and flexible. They let report authors start with a default summarization and change it to fit their needs. Implicit measures let report authors quickly visualize data without needing to write calculations. As a data modeler, you spend less time creating explicit measures.

However, even if you set a default summarization, report authors can change it to something that might not make sense. For example, they could set Unit Price to Sum, which produces misleading results, as

shown in the following image. The Unit Price values are large because they're the sum of unit prices instead of the static unit price per product.

The biggest limitation is that implicit measures only work for simple scenarios. They can summarize column values using a single aggregation function, but they can’t handle more complex calculations. For example, if you need to calculate the ratio of each month’s sales amount over the yearly sales amount, you must create an explicit measure with a DAX formula.

Unité 5: Create explicit measures

Type: Contenu

Create explicit measures

You can add a measure to any table in your model by writing a DAX formula. A measure formula must return a single value. This section explains how to create explicit measures.

Measures don’t store values in the model. Instead, Power BI calculates them at query time to summarize model data. Measures can’t reference a table or column directly, so you must use a function to summarize the data.

A simple measure aggregates the values of a single column, just like an implicit measure.

For example, you can add a measure to the Sales table. In the Data pane, select the Sales table. On the Table Tools ribbon, select New measure.

The following formula creates a measure called Revenue. This measure uses the SUM function to total the values in the Sales Amount column. If you add this measure to a table alongside the Sales Amount implicit measure, the results are the same.

Adding measures and hiding columns helps report authors use the explicit measures instead.

In the Measure tools contextual ribbon, you can format the measure, set data type, and change the home table. The home table refers to where the measure shows when looking at the data pane.

It's a good practice to set the formatting options right after you create a measure. This ensures your values look consistent in all report visuals.

Consider you might need a measure to calculate Profit as shown in the following code:

In this example, the Profit Amount column is a calculated column. This approach isn’t optimal because you don’t need that column. In the next section, you see how to create a measure that calculates profit directly, which reduces model size and improves refresh times.

The following code creates two different measures to return Order Line Count and Order Count. The COUNT function counts non-BLANK values in a column. The DISTINCTCOUNT function counts unique values. Since an order can have multiple order lines, the Sales Order column has duplicates. Using DISTINCTCOUNT gives you the correct order count.

You can also write the Order Line Count measure using COUNTROWS, which counts the number of rows in a table:

All the measures referenced are considered simple measures because they aggregate a single column or single table.

Create compound measures

A compound measure references one or more other measures. For example, you can redefine the Profit measure by referencing other measures. This measure can be used instead of the calculated column previously referenced.

This change to the model presents an important lesson: By removing the calculated column, you optimize the semantic model because it results in a decreased semantic model size and shorter data refresh times. The Profit Amount calculated column wasn't required after all because the Profit measure can directly produce the required result by using existing measures.

Sometimes, it makes sense to define measures that depend on other measures. Always test changes carefully, because updates can affect all dependent measures.

Use Quick measures

Quick measures let you perform common calculations without writing DAX yourself. Power BI generates the DAX expression for you, which helps you learn and build your DAX skills.

For example, you can use a Quick measure to create a Profit Margin measure through the following steps:

Select Quick measure in the Table tools ribbon.

Choose Mathematical operations > Division.

Add the Profit measure into Numerator and Revenue into Denominator.

The new measure appears in the Data pane, and you can review its DAX formula:

Compare calculated columns and measures

Many DAX beginners find calculated columns and measures confusing at first. Both are created in the semantic model using DAX formulas, however, calculated columns and measures behave differently.

Recognizing these differences helps you choose the right approach for your modeling and reporting needs.

Unité 6: Use iterator functions

Type: Contenu

Use iterator functions

Iterator functions evaluate an expression for each row in a table. They give you flexibility and control over how your model summarizes data.

Single-column summarization functions, such as SUM, COUNT, MIN, and MAX, have equivalent iterator functions with an "X" suffix, like SUMX, COUNTX, MINX, and MAXX. Specialized iterator functions also exist for filtering, ranking, and semi-additive calculations over time.

Every iterator function requires a table and an expression. The table can be a model table or any expression that returns a table. The expression must return a single value for each row.

Single-column summarization functions, like SUM, act as shorthand. Power BI internally converts SUM to SUMX. For example, both of the following measures return the same result and have the same performance:

Iterator functions evaluate the expression for each row in a table, using row context—meaning they process one row at a time to compute the final result. Then the table is evaluated in filter context. For example, if a report visual filters by fiscal year FY2020, the Sales table contains only sales rows from that year.

Using iterator functions with large tables and complex expressions can slow performance. Functions like SEARCH and LOOKUPVALUE can be expensive. When possible, use RELATED for better performance.

Iterator functions for complex summarization

Iterator functions let you aggregate more than a single column. For example, a revenue measure can multiply order quantity, unit price, and a discount factor for each row, then sum the results.

Iterator functions can also reference related tables. The discount measure can use the RELATED function to access the list price from the product table:

The following image shows a table visual with the Month, Revenue, and Discount columns. Revenue and Discount are the measures previously created.

Iterator functions for higher grain summarization

Iterator functions can also summarize data at different levels of detail (grain). For example, you might want to calculate an average at the line item level or at the sales order level.

In this example, the Sales table contains one row for each line item in a sales order. Each row includes details such as the sales order number, product, quantity sold, unit price, and discount. Multiple rows can have the same sales order number, representing different items within the same order.

To calculate the average revenue per order line (line item), you can use the AVERAGEX function to iterate over each row in the Sales table. The formula calculates revenue for each line item, then averages the result across all line items in the current filter context:

If you want to calculate the average revenue per sales order (rather than per line item), you can use the VALUES function to get a list of unique sales order numbers first. Then, AVERAGEX iterates over each sales order and averages the total revenue for each order:

The VALUES function returns the unique sales orders based on the current filter context, so AVERAGEX iterates over each sales order for each month.

Ranking with iterator functions

The RANKX function calculates ranks by iterating over a table and evaluating an expression for each row.

Order direction can be ascending or descending. Ranking revenue usually uses descending order, so the highest value ranks first. Ranking something like complaints might use ascending order, so the lowest value ranks first. By default, RANKX uses descending order and skips ranks for ties.

For example, a product quantity rank measure can use RANKX and the ALL function to rank products by quantity:

The ALL function removes filters, so RANKX ranks all products. In the following image, two products tie for tenth place, so the next product is ranked twelfth and rank 11 is skipped.

You can also use dense ranking, which assigns the next rank after a tie without skipping numbers. To use dense ranking, the measure can include the DENSE argument:

Now, after two products tie for tenth place, the next product is ranked eleventh and numbering continues sequentially without skipping rank 11.

In this visual, the total row for the Product Quantity Rank measure shows one, because the total for all products is also ranked and there's only one value.

To avoid ranking the total, the measure can use the HASONEVALUE function to return BLANK unless a single product is filtered:

Now, the total for Product Quantity Rank is blank.

The HASONEVALUE function checks if the product column has a single value in filter context. This is true for each product group, but not for the total, which represents all products.

Iterator functions provide powerful ways to summarize, aggregate, and rank data in Power BI models. They support complex calculations and let you control the level of detail in your reports.

Unité 7: Exercise - Create DAX calculations

Type: Exercice

Exercise - Create DAX calculations

In this exercise, you learn how to use DAX to:

Create calculated tables.

Create calculated columns.

Create measures.

This lab takes approximately 45 minutes to complete.

A virtual machine containing the client tools you need is provided, along with the exercise instructions. Use the "Launch lab" button to launch the virtual machine.

A limited number of concurrent sessions are available. If the hosted environment is unavailable, please try again later.

Alternatively, you can open the instructions in a separate window.

Access your environment

Before you start this lab (unless you are continuing from a previous lab), select Launch lab above.

You are automatically logged in to your lab environment as data-ai\student.

You can now begin your work on this lab.

To dock the lab environment so that it fills the window, select the PC icon at the top and then select Fit Window to Machine.

Unité 8: Check your knowledge

Type: Contenu

Check your knowledge

Which statement about calculated tables is true?

Calculated tables increase the size of the semantic model.

Calculated tables are evaluated by using row context.

Calculated tables are created in Power Query.

Calculated tables can't include calculated columns.

Which statement about calculated columns is true?

Calculated columns are created in the Power Query Editor window.

Calculated column formulas are evaluated by using row context.

Calculated column formulas can only reference columns from within their table.

Calculated columns can't be related to noncalculated columns.

Which statement about measures is correct?

Measures store values in the semantic model.

Measures must be added to the semantic model to achieve summarization.

Measures can reference columns directly.

Measures can reference other measures directly.

You must answer all questions before checking your work.

Unité 9: Summary

Type: Résumé

This module covered using DAX calculations to extend your semantic model. You learned the differences between calculated columns and measures, including when and how Power BI evaluates them and how they store data. Explicit measures are important because they allow you to create complex DAX formulas to achieve the precise calculations that your report visuals need.

You also learned how calculated columns are evaluated within row context, and iterator functions are available in measures for advanced row-by-row calculations.

Understanding how to create and use DAX calculations is fundamental for building effective, flexible, and maintainable semantic models in Power BI. These concepts help you design reports that deliver accurate insights and support a wide range of analytical requirements.

For more information, see Introduction to DAX in Power BI.

Module 2: Concevoir des modèles sémantiques évolutifs

Unité 1: Présentation

Type: Introduction

Ce module couvre les meilleures pratiques de modélisation des données et les fonctionnalités de Microsoft Fabric pour concevoir des modèles sémantiques évolutifs. Les données à grande échelle ou à l'échelle de l'entreprise font référence à des tailles de table allant de centaines de milliers à des millions de lignes.

La scalabilité fait référence à la capacité d'un système, d'un réseau ou d'un processus à gérer une quantité croissante de travail, ou représente le potentiel de prendre en charge la croissance du volume de données et de la complexité sans compromettre les performances ou l'efficacité. Concevez vos modèles pour gérer cette croissance en tenant compte des points suivants :

Flexibilité : adaptation aux modifications du volume de données tout en conservant les performances acceptables des rapports.

Complexité réduite : garantir que les modèles sont moins complexes et gérables.

Les modèles sémantiques évolutifs permettent aux organisations d'analyser et de signaler facilement les sources de données volumineuses et complexes. Microsoft Fabric permet de travailler avec un volume élevé et des données à grande échelle avec le bon travail de base en place. Un modèle sémantique évolutif permet une expérience utilisateur optimale dans les rapports Power BI.

Imaginez que vous êtes sur l'équipe d'analytique d'une grande entreprise de commerce électronique, en préparant le plus grand événement annuel de ventes. Les solutions de création de rapports précédentes étaient manuelles et n'ont pas été mises à l'échelle. Vous êtes maintenant chargé d'améliorer les performances à l'aide de modèles sémantiques dans Microsoft Fabric pour l'analytique en aval et les rapports Power BI.

À la fin, vous pourrez choisir une infrastructure de modèle, concevoir un schéma en étoile et appliquer les meilleures pratiques pour créer un modèle sémantique optimisé pour l'analytique des données à grande échelle avec Power BI.

Unité 2: Choisissez le meilleur mode de stockage

Type: Contenu

Choisissez le meilleur mode de stockage

Lorsque vous concevez un modèle sémantique évolutif, la sélection du mode de stockage approprié est cruciale. Selon votre source de données, quatre options s'offrent à vous : Importation, DirectQuery, Direct Lake et Modèle composite. Chaque mode présente ses propres avantages et considérations pour garantir des performances et une scalabilité optimales.

Mode Importation

Importation implique l'importation et le stockage de données dans Power BI, offrant la meilleure flexibilité et les performances les plus rapides. Toutefois, il nécessite des actualisations périodiques pour maintenir les données à jour, ce qui signifie que les données ne sont à jour que lors de la dernière actualisation.

Les conseils relatifs à l'importation de modèles s'appliquent également aux autres modes de stockage. Vous devez toujours choisir Importer, si votre modèle l'autorise. L'objectif principal est de réduire la quantité de données que vous apportez dans le modèle sémantique et le nombre de transformations es dans Power BI. Les conseils d'optimisation sont les suivants :

Connectez à des affichages plutôt qu'à des tables lors de l'utilisation de bases de données relationnelles.

Incluez uniquement les tables, lignes et colonnes nécessaires.

Envisagez un partitionnement et une actualisation incrémentielle pour éviter de charger des données dont vous n'avez pas besoin.

Utilisez les types de données appropriés, tels que l'entier pour les colonnes d'ID au lieu de chaîne.

Assurez-vous d'effectuer un Query Folding, qui réduit le travail dans le moteur Power BI.

Découvrez-en plus sur les techniques permettant de réduire les données chargées dans des modèles d'importation.

Mode DirectQuery

Le mode DirectQuery vous permet d'interroger des données directement à partir de la source sans la stocker dans Power BI, ce qui le rend idéal pour gérer de grands volumes de données et la livraison de données en temps quasi réel. Toutefois, il présente des performances plus lentes par rapport au mode Importation et offre des fonctionnalités de modélisation limitées.

Conseils d'optimisation :

Évitez les calculs complexes à la source en simplifiant les expressions d'analyse de données (DAX).

Appliquez le mode de stockage double pour les dimensions liées aux tables de faits.

Utilisez la propriété Intégrité référentielle supposée sur les relations.

Évitez les relations sur les colonnes Identificateurs uniques et calculés.

Consultez la documentation pour davantage de conseils sur le modèle DirectQuery.

Mode Direct Lake

Direct Lake permet d'interroger des données directement à partir d'un entrepôt ou d'un lac Microsoft Fabric sans le stocker dans un modèle sémantique, ce qui le rend idéal pour gérer de grands volumes de données. Ce mode est optimisé pour charger rapidement des données dans la mémoire à partir de tables Delta dans Microsoft Fabric. Bien qu'il offre l'avantage d'un accès rapide à de grands volumes de données, il nécessite la configuration d'un lakehouse ou d'un entrepôt, et les tables Delta peuvent devoir être réglées pour des performances optimales.

Configurez Direct Lake pour accéder au stockage ADLS (Azure Data Lake Storage) via des raccourcis.

Configurez le comportement de secours sur DirectQuery pour les requêtes DAX complexes.

Le mode Modèle Composite peut combiner des modes Importer et DirectQuery, ou intégrer plusieurs sources de données DirectQuery. Ce mode prend en charge les relations plusieurs-à-plusieurs sans avoir besoin de tables de pont. Il offre la flexibilité des fonctionnalités d'interrogation et de création de rapports en combinant les avantages des modes Importation et DirectQuery. Toutefois, il nécessite des actualisations périodiques pour les tables en mode Importation et peut avoir des répercussions potentielles sur les performances lors de la combinaison de données provenant de différentes sources.

Vérifiez que la source principale dispose de ressources suffisantes.

Réduisez le nombre de valeurs littérales dans les requêtes sources.

Conservez la cardinalité des colonnes utilisées dans les relations basses.

Consultez la documentation pour obtenir des conseils supplémentaires sur les modèles composites.

Unité 3: Configurez des modèles sémantiques pour les données volumineuses

Type: Contenu

Configurez des modèles sémantiques pour les données volumineuses

Lorsque vous utilisez des sources de données très volumineuses, vous devez changer de stratégie de chargement des données. Dans cette unité, nous expliquons comment activer la mise en forme du stockage de modèles sémantiques volumineux et comment configurer l'actualisation incrémentielle. Ces fonctionnalités permettent aux modèles sémantiques volumineux d'être actualisés à un taux et une taille plus gérables.

Formats de stockage de modèles sémantiques volumineux

Le format de stockage de modèle sémantique volumineux dans Microsoft Fabric vous permet de gérer efficacement des sources de données plus volumineuses. Ce format est idéal pour les modèles analytiques complexes qui nécessitent davantage de mémoire et de puissance de traitement. Cette fonctionnalité permet à votre modèle sémantique de croître au-delà de la limite de 10 Go. Votre limite peut varier, car la capacité ou l'administrateur Fabric détermine la limite du modèle sémantique volumineux. Activer cette fonctionnalité est aussi simple que sélectionner une zone dans les paramètres de l'espace de travail.

Si votre modèle sémantique continue de croître et consomme progressivement plus de mémoire, veillez à configurer l'actualisation incrémentielle.

Consultez la documentation pour en savoir plus sur les modèles sémantiques volumineux.

Actualisation incrémentielle

L'actualisation incrémentielle vous permet d'actualiser uniquement les données qui ont changé ou qui ont été ajoutées depuis la dernière actualisation. Cette fonctionnalité réduit le temps et les ressources nécessaires pour les actualisations des données, ce qui est idéal pour les sources de données fréquemment mises à jour.

Supposons que votre table de ventes soit incroyablement volumineuse et que l'actualisation du modèle sémantique prenne beaucoup de temps. Si vous devez conserver les données à jour, vous pouvez configurer l'actualisation incrémentielle pour actualiser uniquement les transactions nouvelles et mises à jour.

Pour activer l'actualisation incrémentielle, vous devez d'abord définir des paramètres et filtrer la ligne à l'aide des paramètres dans Power Query. Une fois appliqué, vous configurez la stratégie d'actualisation incrémentielle sur la table dans Power BI Desktop. Vous pouvez éventuellement configurer la table pour obtenir les données les plus récentes avec DirectQuery. Pour tirer parti de cette fonctionnalité, votre modèle sémantique doit être publié dans un espace de travail Fabric pris en charge.

Consultez la documentation pour en savoir plus sur l'actualisation incrémentielle pour les modèles sémantiques.

L'actualisation incrémentielle se concentre uniquement sur l'actualisation des données modifiées ou nouvelles, tandis que le partitionnement implique de diviser la table entière en segments plus petits. Le

partitionnement implique de diviser une table volumineuse en morceaux plus petits et plus gérables appelés partitions. Chaque partition peut être traitée indépendamment, ce qui peut améliorer les performances des requêtes et la facilité de gestion.

Par exemple, vous pouvez créer des partitions pour chaque mois ou trimestre sur votre table de ventes, ce qui permet aux requêtes de cibler des partitions spécifiques plutôt que l'ensemble du jeu de données.

Pour activer le partitionnement, vous pouvez utiliser des outils tels que SQL Server Management Studio (SSMS) ou l'éditeur tabulaire.

Consultez la documentation pour en savoir plus sur le partitionnement pour l'actualisation incrémentielle avancée avec le point de terminaison XMLA.

Unité 4: Utiliser les relations

Type: Contenu

Utiliser les relations

Après avoir choisi l'infrastructure de modèle et transformé des données, vous devez créer des relations pour permettre aux données d'être filtrées et agrégées en fonction des données d'une autre table. Il est courant de créer un schéma en étoile et d'appliquer des règles qui filtrent les tables de dimension, ce qui permet aux relations de modèle d'appliquer efficacement ces filtres aux tables de faits.

Créer un schéma en étoile

Dans notre scénario, nous avons importé une table de dimensions à partir de .csv qui doit se connecter au lakehouse. Vous pouvez créer des relations pour mettre en forme un schéma en étoile une fois la table importée transformée.

Dans un schéma en étoile, les tables de faits et les tables de dimensions fonctionnent ensemble pour organiser et analyser des données. La table de faits stocke les données principales sur les activités commerciales, telles que les ventes ou les événements, tandis que les tables de dimensions fournissent des informations contextuelles et descriptives sur ces activités.

Par exemple, si vous avez une table de faits de vente, elle peut stocker des données telles que la quantité vendue et la date de chaque vente. Les tables de dimensions fournissent ensuite des détails supplémentaires, tels que les informations du client, les détails du produit et des périodes. En liant la table de faits aux tables de dimension, vous pouvez facilement filtrer, regrouper et analyser les données.

Créer des relations

La création de relations entre les tables de faits et de dimensions est simple, elle se fait en identifiant les colonnes associées dans chaque table et en créant la relation. Dans les sources relationnelles telles que les bases de données ou les entrepôts de données, ces colonnes sont souvent appelées clés. Si les colonnes clés sont absentes, examinez les tables pour déterminer les colonnes à utiliser pour les relations.

Les relations se présentent sous différents types :

Un-à-plusieurs est le type le plus courant, où un enregistrement d'une table se rapporte à plusieurs enregistrements d'une autre. En règle générale, une relation de type un-à-plusieurs provient d'une table de dimensions possédant une valeur unique vers une table de faits avec de nombreuses lignes associées à cette unique valeur.

Plusieurs-à-un est plus ou moins identique à Un-à-plusieurs, selon la façon dont le filtre est configuré entre les tables.

Un-à-un est moins courant, car les deux tables ont des données uniques. Réfléchissez si vous avez besoin de deux tables ou si vous pouvez les combiner en une seule table.

La relation Plusieurs-à-plusieurs est moins courante, mais nécessaire pour les données complexes. Ce type de relation permet à plusieurs enregistrements d'une table de se rapporter à plusieurs enregistrements d'une autre.

Pour les relations Plusieurs-à-plusieurs, vous devrez peut-être utiliser une table de pont. Une table de pont permet de gérer ces relations en liant les tables via des clés intermédiaires. Les modèles composites permettent également d'utiliser des relations plusieurs-à-plusieurs en permettant de combiner des données provenant de différentes sources.

Direction du filtre

Lorsque vous créez une relation, vous configurez la direction dans laquelle les données sont filtrées d'une table à une autre. Dans un schéma en étoile, la direction passe généralement de la table de dimensions à la table de faits, ce qui permet à la table de dimensions de filtrer les résultats de la table de faits.

Les filtres bidirectionnels sont également possibles et parfois utilisés dans des relations un-à-un ou plusieurs-à-plusieurs. Avant d'utiliser un filtre bidirectionnel, vérifiez que vos données et relations sont correctement configurées. Soyez prudent, car les filtres bidirectionnels peuvent affecter les performances des requêtes de modèle et éventuellement perturber les utilisateurs du rapport.

L'intégrité référentielle garantit que les relations entre les tables restent cohérentes. Cela signifie que chaque valeur d'une colonne de clé étrangère doit avoir une valeur correspondante dans la colonne de clé primaire de la table associée. Cette fonctionnalité améliore également les performances des requêtes en utilisant les joins INNER par opposition aux joins LEFT OUTER.

Comprendre la direction du filtre et de l'intégrité référentielle est essentiel pour la modélisation précise des données. Les filtres se propagent uniquement si le chemin de relation est intact et suit la direction définie. Ces configurations garantissent la cohérence et l'intégrité des données dans votre modèle.

Relations dormantes

Les relations peuvent être désactivées et leur contexte de filtre est modifié par les fonctions DAX. Parfois, vous avez besoin de plusieurs relations entre les tables, mais une seule peut être active à la fois entre deux tables. Dans ce cas, utilisez la fonction USERELATIONSHIP dans DAX pour référencer les relations inactives et obtenir le même comportement de filtrage. Voici un exemple :

Utiliser des fonctions DAX pour les relations est important, car elles vous permettent de créer des calculs dynamiques et flexibles qui peuvent s'adapter à différents scénarios de données. Cette flexibilité vous permet de gérer des modèles de données complexes et d'effectuer des analyses avancées qui seraient difficiles à réaliser avec des relations statiques seules.

Utiliser des tables déconnectées

Il est rare qu'une table de modèle ne soit pas associée à une autre table de modèle. Dans une conception de modèle valide, une telle table est décrite comme étant une table déconnectée. Une table déconnectée n'est pas destinée à propager des filtres à d'autres tables de modèle. Au lieu de cela, elle accepte des « entrées utilisateur » (peut-être avec un visuel de segment), ce qui permet aux calculs de modèle d'utiliser les valeurs d'entrée de manière significative. Prenons l'exemple d'une table déconnectée avec une plage de valeurs de taux de change. Lors du filtrage par une valeur à taux unique, une expression de mesure peut utiliser cette valeur pour convertir des valeurs de ventes.

Le paramètre de scénario (« what-if ») Power BI Desktop est une fonctionnalité qui crée une table déconnectée. Pour plus d'informations, consultez Créer et utiliser un paramètre de scénario pour visualiser des variables dans Power BI Desktop.

Unité 5: Écrivez DAX pour une meilleure lisibilité avec des calculs complexes

Type: Contenu

Écrivez DAX pour une meilleure lisibilité avec des calculs complexes

DAX fournit un ensemble puissant de fonctions qui vous permettent de créer des calculs de base à complexes pour mesurer les données de votre modèle sémantique dans Power BI. En utilisant des conditions de filtrage, des variables et des fonctions d'itérateur, de fenêtre et d'information, vous pouvez créer des formules efficaces et lisibles.

Les variables dans DAX vous aident à simplifier vos expressions et à améliorer les performances en stockant le résultat d'une expression et en le réutilisant plusieurs fois. Cela réduit le besoin de calculs répétés et rend votre code plus facile à lire et à maintenir.

Considérez le scénario dans lequel vous devez calculer le montant moyen des ventes par client. Vous pouvez utiliser une variable pour stocker le montant total des ventes, puis l'utiliser dans le calcul :

Bien que cet exemple soit toujours lisible sans la variable, les formules complexes peuvent utiliser plusieurs variables et augmenter considérablement la lisibilité de votre code.

Les fonctions itératives dans DAX, telles que SUMX, AVERAGEX et MAXX, effectuent des calculs ligne par ligne sur une table et renvoient une valeur unique. Ces fonctions sont utiles pour effectuer des calculs qui dépendent du contexte de chaque ligne.

Par exemple, pour calculer le bénéfice total de chaque produit, vous pouvez utiliser la fonction SUMX :

Soyez prudent lorsque vous utilisez des fonctions itératives pour de grandes quantités de données en raison du traitement ligne par ligne, ce qui peut affecter les performances.

Filtrage des tableaux

Les fonctions de filtrage de table, telles que FILTER, ALL et CALCULATETABLE, vous permettent de créer des tables filtrées en fonction de conditions spécifiques. Ces fonctions sont utiles pour créer des calculs dynamiques qui dépendent du contexte filtré.

Par exemple, pour calculer les ventes totales d'une catégorie de produits spécifique, vous pouvez utiliser la fonction CALCULATETABLE :

La fonction CALCULATETABLE modifie le contexte du filtre pour inclure uniquement les lignes dont la catégorie de produit est « Électronique ». La fonction SUMMARIZE répertorie ensuite les montants des ventes dans ce contexte filtré. Par conséquent, lorsque vous utilisez cette mesure dans un visuel présentant les ventes totales par catégorie, chaque ligne affichera uniquement les ventes totales pour « Électronique », quelle que soit la catégorie réelle dans le visuel.

Ces fonctions de filtrage de tableau vous aident à créer des calculs dynamiques et contextuels, permettant une analyse des données plus précise et plus approfondie.

Fonctions de fenêtrage

Les fonctions de fenêtrage dans DAX, telles que INDEX, OFFSET et WINDOW, vous permettent d'effectuer des calculs sur une fenêtre de données spécifiée. Ces fonctions sont utiles pour créer des

classements, des totaux courants et d’autres calculs qui dépendent de l’ordre des données.

Par exemple, pour comparer les ventes du produit actuel avec le produit précédent, vous pouvez utiliser la fonction OFFSET :

Dans cet exemple, la fonction OFFSET est utilisée pour déplacer le contexte vers la ligne précédente en fonction de l'ordre de la colonne Ventes[Date]. Cela vous permet de comparer le montant des ventes du produit actuel avec le précédent, ce qui peut être utile pour l'analyse des tendances et d'autres calculs comparatifs.

Fonctions d'information

Les fonctions d'information, telles que ISBLANK, ISNUMBER et CONTAINS, vous permettent d'effectuer des vérifications et de renvoyer des informations sur les données. Ces fonctions sont utiles pour créer des calculs conditionnels et gérer des cas particuliers.

Par exemple, vous pouvez utiliser la fonction HASONEVALUE pour vérifier si une colonne possède une seule valeur distincte. Cela est utile dans les scénarios où vous souhaitez effectuer des calculs uniquement lorsqu'une seule valeur est sélectionnée dans un contexte de filtre.

Considérez le scénario dans lequel vous souhaitez calculer le montant total des ventes uniquement si une seule catégorie de produits est sélectionnée :

La fonction HASONEVALUE vérifie s'il n'y a qu'une seule valeur distincte dans la colonne Produits[Catégorie]. Si une seule catégorie est sélectionnée, la formule calcule le montant total des ventes, et si plusieurs catégories sont sélectionnées, la formule renvoie BLANK().

En utilisant des variables et des fonctions DAX, vous pouvez créer des calculs plus efficaces, lisibles et puissants qui améliorent votre modèle sémantique dans Power BI.

Assurez-vous de mettre en favoris la référence Data Analysis Expressions (DAX) pour obtenir tous les détails sur la syntaxe, les différentes fonctions, les instructions et plus encore.

Unité 6: Créez des éléments de calcul dynamiques

Type: Contenu

Créez des éléments de calcul dynamiques

Les calculs statiques et les visuels peuvent présenter des problèmes pour les performances des rapports, mais vous pouvez de la flexibilité à votre modèle sémantique à l’aide de groupes de calcul, de chaînes de format dynamique et de paramètres de champ. Ces fonctionnalités rendent vos rapports évolutifs et conviviaux en simplifiant les calculs et en réduisant les visualisations de rapports.

Groupes de calcul

Les groupes de calcul vous permettent de définir des calculs réutilisables qui s’appliquent à plusieurs mesures, ce qui réduit la redondance et simplifie la maintenance de vos modèles sémantiques. Vous pouvez utiliser des groupes de calcul pour simplifier les calculs complexes, tels que les fonctions d’intelligence temporelle, sur l’ensemble de votre modèle.

Envisagez de calculer les données de ventes année à date (YTD), trimestre à date (QTD) et mois à date (MTD). Au lieu de créer des mesures distinctes pour chaque fonction d’intelligence temporelle, vous décidez d’utiliser des groupes de calcul pour simplifier ces calculs.

Pour cet exemple, nous avons un groupe de calcul appelé Time Intelligence avec les éléments de calcul suivants.

Vous pouvez maintenant utiliser le groupe de calcul dans le volet de filtre, un segment, un visuel et même en référence dans une mesure. Les différents éléments de calcul (YTD, QTD, MTD) apparaissent automatiquement pour filtrer ou développer dans le visuel.

Dans l'image suivante, nous avons une matrice avec les trois années fiscales avec le total des ventes entre YTD, QTD et MTD. Il existe également un segment pour le groupe de calcul pour permettre aux utilisateurs de basculer entre les différents choix. Le visuel est configuré comme suit :

Lignes : Champ Date[Year]

Colonnes : Groupe de calcul Time Calc

Valeurs : Mesure Total Sales

Sans groupe de calcul, vous devez créer des mesures YTD, QTD et MTD pour chaque calcul dont vous avez besoin, telles que Total des ventes, Bénéfice, Cible, etc. Au lieu de cela, créez vos visuels et ajoutez le groupe de calcul et les autres mesures.

La nature dynamique et réutilisable des groupes de calcul les rend incroyablement puissants pour mettre à l'échelle vos modèles sémantiques.

Consultez la documentation pour en savoir plus sur la Création de groupes de calculs dans Power BI.

Paramètres de champ

Les paramètres de champ vous permettent de créer des rapports interactifs en permettant aux utilisateurs de sélectionner dynamiquement différents champs ou mesures. Cette fonctionnalité est utile pour créer des rapports personnalisables dans lesquels les utilisateurs peuvent choisir les données qu'ils souhaitent voir.

Dans notre scénario, nous avons créé un nouveau paramètre pour inclure les champs Produit, Catégorie et Couleur. Maintenant, nous utilisons notre mesure Total des ventes et ajoutons le paramètre dans un visuel au lieu de ces champs individuels. Nous ajoutons également un segment avec le paramètre afin que les utilisateurs puissent basculer entre les champs sélectionnés. L'image suivante montre un histogramme pour Total des ventes par catégorie configuré avec le paramètre dans l'axe X et Total des ventes dans l'axe Y. Un segment est également présent pour basculer dynamiquement entre le Total des ventes par Produit, Catégorie et Couleur.

Avant les paramètres de champ, les développeurs de rapports peuvent créer un visuel pour Total des ventes par Produit et répéter pour la Catégorie et la Couleur. Les utilisateurs peuvent basculer entre les différents visuels d'une expérience similaire en superposant les visuels les uns sur les autres et en ajoutant des signets et des boutons. Toutefois, les visuels que vous ajoutez à une page de rapport peuvent avoir un impact sur les performances.

Consultez la documentation pour en savoir plus sur la façon de Permettre aux lecteurs de rapports d'utiliser des paramètres de champ pour modifier les visuels.

Chaînes de format dynamique

Les chaînes de format dynamique vous permettent d'ajuster le format d'une mesure en fonction des conditions, ce qui améliore la lisibilité et la présentation des données.

Envisagez d'afficher les chiffres des ventes dans différents formats en fonction de leur valeur :

Millions (M) pour les ventes de plus de 1 000 000.

Milliers (K) pour les ventes comprises entre 1 000 et 1 000 000.

Valeur exacte des ventes inférieures à 1 000.

Le code suivant utilise la fonction SWITCH pour appliquer les différents formats en fonction du montant des ventes :

Les chaînes de format dynamique simplifient votre présentation de données, ce qui réduit la nécessité de plusieurs calculs ou visuels, et peuvent être utilisées avec des groupes de calcul pour réduire la complexité et la maintenance de votre code.

Consultez la documentation pour en savoir plus sur la Création de chaînes de format dynamique pour les mesures.

Unité 7: Exercice - Concevoir un modèle sémantique évolutif

Type: Exercice

Exercice - Concevoir un modèle sémantique évolutif

Dans cet exercice, vous allez concevoir un modèle sémantique évolutif en effectuant les tâches suivantes :

Choisissez un mode de stockage.

Configurer l'actualisation incrémentielle.

Créez un schéma en étoile avec des relations.

Créez des formules DAX avec des éléments dynamiques.

Ce labo prend environ 30 minutes.

Vous aurez besoin d'une licence Microsoft Fabric et de Power BI Desktop pour réaliser cet exercice. Pour plus d'informations sur l'activation d'une licence d'essai Fabric gratuite, consultez Bien démarrer avec Fabric.

Lancez l'exercice et suivez les instructions.

Unité 8: Évaluation du module

Type: Évaluation

Évaluation du module

Quel est l'avantage principal de l'utilisation du mode Importation dans un modèle sémantique évolutif ?

Il prend en charge les relations plusieurs-à-plusieurs sans avoir besoin de tables de pont.

Il permet d'interroger des données directement à partir de la source sans les stocker dans Power BI.

Il offre la meilleure flexibilité et les performances les plus rapides, mais nécessite des actualisations périodiques pour maintenir les données à jour.

Quel est l'objectif des fonctions d'itérateur dans DAX ?

Stocker le résultat d'une expression et le réutiliser plusieurs fois

Effectuer des calculs ligne par ligne sur une table et retourner une valeur unique

Effectuer des vérifications et retourner des informations sur les données

Quel est l’avantage principal de l’utilisation des groupes de calcul dans un modèle sémantique ?

Ils permettent aux utilisateurs de sélectionner dynamiquement différents champs ou mesures.

Ils permettent des calculs réutilisables qui s’appliquent à plusieurs mesures, ce qui réduit la redondance et simplifie la maintenance.

Ils permettent une mise en forme dynamique basée sur certaines conditions.

Vous devez répondre à toutes les questions avant de vérifier votre travail.

Unité 9: Résumé

Type: Résumé

Dans ce module, nous avons exploré comment les modèles sémantiques évolutifs permettent aux organisations d’analyser et de signaler facilement des sources de données complexes et volumineuses. Microsoft Fabric permet aux utilisateurs de bénéficier d’une expérience optimale dans les rapports Power BI en gérant efficacement les données à grande échelle et en grand volume. Pour ce faire, nous avons abordé les points suivants :

Optimisation basée sur le mode de stockage.

Techniques de gestion des données volumineuses.

Conception d’un schéma en étoile et de relations.

Optimisation des calculs à l’aide de variables, de fonctions DAX et d’éléments dynamiques.

Sans Microsoft Fabric, l’équipe chargée de l’analyse s’appuierait sur des solutions de création de rapports manuelles et non évolutives, ce qui entraînerait un traitement des données fastidieux, une complexité accrue et une capacité limitée à traiter efficacement des jeux de données volumineux.

En concevant dès le départ des modèles sémantiques évolutifs, vous pouvez améliorer les performances et l’efficacité de l’analytique. Ces techniques permettent d’établir des rapports précis en temps voulu, de prendre de meilleures décisions, d’améliorer l’efficacité opérationnelle et d’accroître la satisfaction générale.

Pour aller plus loin :

Guide d’optimisation pour Power BI

Présentation de DAX

Module 3: Optimiser un modèle pour améliorer les performances dans Power BI

Unité 1: Présentation de l'optimisation des performances

Type: Introduction

Présentation de l'optimisation des performances

L'optimisation des niveaux de performance, également appelée réglage des performances, consiste à apporter des modifications à l'état actuel du modèle sémantique afin qu'il s'exécute plus efficacement. En gros, quand votre modèle sémantique est optimisé, il fonctionne mieux.

Vous constaterez peut-être que votre rapport s'exécute correctement dans les environnements de test et de développement, mais que des problèmes de performances se produisent quand il est déployé en production en vue d'une consommation élargie. Du point de vue d'un utilisateur du rapport, les performances médiocres se caractérisent par un chargement des pages de rapport et une mise à jour des visuels qui prennent plus de temps. Cette dégradation des performances entraîne une expérience utilisateur négative.

En tant qu'analyste de données, vous passez environ 90 % de votre temps à travailler avec vos données et, neuf fois sur dix, les performances médiocres sont le résultat direct d'un modèle sémantique incorrect et/ou d'expressions DAX (Data Analysis Expressions) incorrectes. Le processus de conception d'un modèle sémantique prenant en compte les performances peut être fastidieux et est souvent sous-estimé. Toutefois, si vous résolvez les problèmes de performances pendant le développement, vous obtenez un modèle sémantique Power BI robuste qui affiche de meilleures performances de création de rapports et procure une expérience utilisateur plus positive. Au final, vous serez aussi en mesure de maintenir des performances optimisées. À mesure que votre organisation se développe, la taille de ses données croît et son modèle sémantique devient plus complexe. En optimisant votre modèle sémantique de façon anticipée, vous pouvez atténuer l'impact négatif que cette croissance peut avoir sur les performances de votre modèle sémantique.

Un modèle sémantique de taille inférieure utilise moins de ressources (mémoire) et permet d'accélérer l'actualisation des données, les calculs et le rendu des visuels dans les rapports. Ainsi, le processus d'optimisation des performances implique de réduire la taille du modèle sémantique et de tirer pleinement parti des données dans le modèle, notamment en :

S'assurant que les types de données corrects sont utilisés

Supprimant les colonnes et les lignes inutiles

Évitant les valeurs répétées

Remplaçant les colonnes numériques par des mesures

Réduisant les cardinalités

Analysant les métadonnées du modèle

Résumant les données dans la mesure du possible

Dans ce module, vous allez découvrir les étapes, les processus et les concepts nécessaires à l'optimisation d'un modèle sémantique pour améliorer les performances au niveau de l'entreprise. Toutefois, gardez à l'esprit que, même si vous pouvez la plupart du temps vous appuyer sur les recommandations de base en matière de performances et de bonnes pratiques dans Power BI, pour optimiser un modèle sémantique afin d'améliorer les performances des requêtes, vous devrez probablement vous associer à un ingénieur Données pour conduire l'optimisation du modèle de

données dans les sources de données sources.

Par exemple, supposons que vous travaillez en tant que développeur Microsoft Power BI pour Tailwind Traders. Il vous a été demandé d'examiner un modèle sémantique créé il y a quelques années par un autre développeur, qui ne fait plus partie de l'organisation.

Le modèle sémantique produit un rapport qui a fait l'objet de commentaires négatifs de la part des utilisateurs. Les utilisateurs sont satisfaits des résultats qu'ils voient dans le rapport, mais pas des performances de ce dernier. Le chargement des pages dans le rapport prend trop de temps et les tables ne sont pas actualisées suffisamment rapidement quand certaines sélections sont es. En plus de ces commentaires, l'équipe informatique a mis en évidence que la taille de fichier de ce modèle sémantique particulier est trop grande, au point de peser sur les ressources de l'organisation.

Vous devez examiner le modèle sémantique pour identifier la cause racine des problèmes de performances et procéder à des changements pour optimiser les performances.

À la fin de ce module, vous serez en mesure de :

Passer en revue les performances des mesures, des relations et des visuels

Utiliser des variables pour améliorer les performances et la résolution des problèmes

Améliorer les performances en réduisant les niveaux de cardinalité

Optimiser les modèles DirectQuery avec un stockage au niveau de la table

Créer et gérer des agrégations

Unité 2: Passer en revue les performances des mesures, des relations et des visuels

Type: Contenu

Passer en revue les performances des mesures, des relations et des visuels

Si votre modèle sémantique contient plusieurs tables, des relations complexes, des calculs compliqués, plusieurs visuels ou des données redondantes, les performances des états risquent d'être médiocres, au point d'engendrer une expérience utilisateur négative.

Pour optimiser les performances, vous devez d'abord identifier l'origine du problème, en d'autres termes, repérer les éléments de l'état et du modèle sémantique à l'origine des problèmes de performance. Ensuite, vous pouvez agir pour résoudre ces problèmes et, ainsi, améliorer les performances.

Identifier les goulots d'étranglement des performances des états

Pour obtenir des performances optimales dans vos états, vous devez créer un modèle sémantique efficace doté de requêtes et de mesures à exécution rapide. Quand vous avez une bonne base, vous pouvez améliorer le modèle en analysant les plans de requête et les dépendances, puis en apportant des modifications pour optimiser les performances.

Vous devez examiner les mesures et les requêtes dans votre modèle sémantique afin de vous assurer que vous utilisez la méthode la plus efficace pour obtenir les résultats souhaités. Votre point de départ doit consister à identifier les goulots d'étranglement qui existent dans le code. Lorsque vous identifiez la requête la plus lente dans le modèle sémantique, vous pouvez vous concentrer d'abord sur le goulot d'étranglement le plus important et établir une liste de priorités pour traiter les autres problèmes.

Analyser les performances

Vous pouvez utiliser l'Analyseur de performances dans Power BI Desktop pour vous aider à découvrir les performances de chacun des éléments de votre état lorsque les utilisateurs interagissent avec eux. Par exemple, vous pouvez déterminer le temps nécessaire à l'actualisation d'un visuel particulier quand l'opération est démarrée par une interaction avec l'utilisateur. L'Analyseur de performances vous aide à identifier les éléments qui contribuent à vos problèmes de performance, ce qui peut être utile pour leur résolution.

Avant d'exécuter l'Analyseur de performances, et afin d'obtenir les résultats les plus précis dans votre analyse (test), veillez à ce que les caches de visuels et de moteur de données soient vides.

Cache de visuels : lorsque vous chargez un visuel, vous ne pouvez pas effacer ce cache sans fermer Power BI Desktop et le rouvrir. Pour éviter toute mise en cache en opération, vous devez démarrer votre analyse avec un cache de visuels vide. Pour vous assurer d'avoir un cache de visuels vide, ajoutez une page vierge à votre fichier Power BI Desktop (.pbix), puis, une fois cette page sélectionnée, enregistrez et fermez le fichier. Rouvrez le fichier Power BI Desktop (.pbix) que vous souhaitez analyser. Il s'ouvre sur la page vierge.

Cache de visuels : lorsque vous chargez un visuel, vous ne pouvez pas effacer ce cache sans fermer Power BI Desktop et le rouvrir. Pour éviter toute mise en cache en opération, vous devez démarrer votre analyse avec un cache de visuels vide.

Pour vous assurer d'avoir un cache de visuels vide, ajoutez une page vierge à votre fichier Power BI Desktop (.pbix), puis, une fois cette page sélectionnée, enregistrez et fermez le fichier. Rouvrez le fichier Power BI Desktop (.pbix) que vous souhaitez analyser. Il s'ouvre sur la page vierge.

Cache du moteur de données : lorsqu'une requête est exécutée, les résultats sont mis en cache, de sorte que les résultats de votre analyse sont trompeurs. Vous devez vider le cache de données avant de réexécuter le visuel. Pour vider le cache de données, vous pouvez soit redémarrer Power BI Desktop, soit connecter DAX Studio au modèle sémantique, puis appeler la commande Clear Cache (Vider le cache).

Cache du moteur de données : lorsqu'une requête est exécutée, les résultats sont mis en cache, de sorte que les résultats de votre analyse sont trompeurs. Vous devez vider le cache de données avant de réexécuter le visuel.

Pour vider le cache de données, vous pouvez soit redémarrer Power BI Desktop, soit connecter DAX Studio au modèle sémantique, puis appeler la commande Clear Cache (Vider le cache).

Une fois que vous avez vidé les caches et ouvert le fichier Power BI Desktop sur la page vierge, accédez à l'onglet Affichage et sélectionnez l'option Analyseur de performances.

Pour commencer le processus d'analyse, sélectionnez Démarrer l'enregistrement, sélectionnez la page de l'état que vous souhaitez analyser et interagissez avec les éléments de l'état que vous souhaitez mesurer. Vous verrez les résultats de vos interactions s'afficher progressivement dans le volet Analyseur de performances. Ensuite, cliquez sur le bouton Arrêter.

Pour en savoir plus, consultez Examiner les performances des éléments d'état à l'aide de l'Analyseur de performances.

Examiner les résultats

Vous pouvez examiner les résultats de votre test de performances dans le volet Analyseur de performances. Pour passer en revue les tâches par ordre de durée, de la plus longue à la plus courte, cliquez avec le bouton droit sur l'icône Trier en regard de l'en-tête de colonne Durée (ms), puis sélectionnez Durée totale en guise d'ordre Décroissant.

Les informations de journalisation de chaque visuel indiquent le temps nécessaire (durée) pour effectuer les catégories de tâches suivantes :

Requête DAX : temps qu'il a fallu au visuel pour envoyer la requête et à Analysis Services pour renvoyer les résultats.

Affichage de visuel : temps nécessaire pour afficher le visuel à l'écran, y compris le temps nécessaire pour récupérer les images web ou le géocodage.

Autre : temps qu'il a fallu au visuel pour préparer les requêtes, attendre la fin d'autres visuels ou effectuer d'autres tâches de traitement en arrière-plan. Si cette catégorie affiche une longue durée, le seul moyen réel de réduire celle-ci consiste à optimiser les requêtes DAX pour les autres visuels ou à réduire le nombre de visuels dans l'état.

Les résultats du test d'analyse vous aident à comprendre le comportement de votre modèle sémantique et à identifier les éléments que vous devez optimiser. Vous pouvez comparer la durée de chaque élément dans l'état et identifier les éléments qui ont une longue durée. Vous devez vous concentrer sur ces éléments et rechercher la raison pour laquelle ils sont longs à se charger sur la page d'état.

Pour analyser vos requêtes plus en détail, vous pouvez utiliser DAX Studio, outil open source gratuit fourni par un autre service.

Résoudre les problèmes et optimiser les performances

Les résultats de votre analyse identifient les domaines à améliorer et les opportunités d'optimisation des performances. Vous constaterez peut-être que vous devez apporter des améliorations aux visuels, à la requête DAX ou à d'autres éléments de votre modèle sémantique. Les informations suivantes fournissent des conseils sur les éléments à rechercher et les modifications que vous pouvez apporter.

Si vous identifiez des visuels comme goulot d'étranglement entraînant des performances médiocres, vous devez trouver un moyen d'améliorer les performances avec un impact minimal sur l'expérience utilisateur.

Tenez compte du nombre de visuels sur la page d'état : moins de visuels est synonyme de meilleures performances. Demandez-vous si un visuel est vraiment nécessaire et s'il ajoute de la valeur pour l'utilisateur final. Si la réponse est non, vous devez supprimer ce visuel. Plutôt que d'utiliser plusieurs visuels dans la page, envisagez d'autres moyens de fournir des détails supplémentaires, tels que des pages d'extraction et des info-bulles de page d'état.

Examinez le nombre de champs dans chaque visuel. Plus le nombre de visuels dans l'état est grand, plus les risques de problèmes de performance sont élevés. En outre, plus le nombre de visuels est grand, plus l'état peut sembler encombré et perdre en clarté. La limite supérieure pour les visuels étant de 100 champs (mesures ou colonnes), un visuel avec plus de 100 champs est lent à charger. Demandez-vous si vous avez vraiment besoin de toutes ces données dans un visuel. Vous constaterez peut-être que vous pouvez réduire le nombre de champs que vous utilisez actuellement.

Quand vous examinez les résultats dans le volet Analyseur de performances, vous pouvez voir combien de temps il a fallu au moteur Power BI Desktop pour évaluer chaque requête (en millisecondes). Une requête DAX qui prend plus de 120 millisecondes est un bon point de départ. Dans cet exemple, vous identifiez une requête qui a une longue durée.

L'Analyseur de performances met en évidence les problèmes potentiels, mais ne vous indique pas ce qui doit être fait pour y remédier. Vous souhaiterez peut-être effectuer des investigations supplémentaires afin de déterminer la raison pour laquelle le traitement de cette mesure prend tant de temps. Vous pouvez utiliser DAX Studio pour étudier vos requêtes plus en détail.

Par exemple, sélectionnez Copier la requête pour copier la formule de calcul dans le presse-papiers, puis collez-la dans DAX Studio. Vous pouvez ensuite examiner l'étape de calcul. Dans cet exemple, vous essayez de compter le nombre total de produits dont la quantité commandée est supérieure ou égale à cinq.

Après l'analyse de la requête, vous pouvez utiliser vos propres connaissances et expériences pour identifier où se situent les problèmes de performance. Vous pouvez également essayer d'utiliser différentes fonctions DAX pour déterminer si elles améliorent les performances. Dans l'exemple suivant, la fonction FILTER a été remplacée par la fonction KEEPFILTER. Quand le test a été réexécuté dans l'Analyseur de performances, la durée a été plus courte en raison de la fonction KEEPFILTER.

Dans ce cas, vous pouvez remplacer la fonction FILTER par la fonction KEEPFILTER pour réduire considérablement la durée d'évaluation de cette requête. Quand vous apportez cette modification, pour vérifier si la durée a été améliorée ou non, videz le cache de données, puis réexécutez le processus de l'Analyseur de performances.

Modèle sémantique

Si la durée des mesures et visuels affiche des valeurs basses (en d'autres termes, ils ont une durée courte), ils ne sont pas à l'origine des problèmes de performance. En revanche, si la requête DAX affiche une valeur de durée élevée, il est probable qu'une mesure est mal écrite ou qu'un problème a affecté le modèle sémantique. Le problème peut être dû à des relations, des colonnes ou des métadonnées dans votre modèle, ou bien à l'état de l'option Date/heure automatique, comme expliqué dans la section suivante.

Vous devez examiner les relations entre vos tables pour vérifier que vous avez établi les bonnes relations. Vérifiez que les propriétés de cardinalité des relations sont correctement configurées. Par exemple, une colonne « un » contenant des valeurs uniques peut être configurée de manière incorrecte en tant que colonne « plusieurs ». Vous en apprendrez davantage sur la façon dont la cardinalité affecte les performances plus loin dans ce module.

Il est conseillé de ne pas importer de colonnes de données dont vous n'avez pas besoin. Pour éviter de supprimer des colonnes dans l'Éditeur Power Query, vous devez essayer de les traiter à la source lors du chargement de données dans Power BI Desktop. Toutefois, s'il est impossible de supprimer les colonnes redondantes de la requête source ou que les données ont déjà été importées dans leur état brut, vous pouvez toujours utiliser l'Éditeur Power Query pour examiner chaque colonne. Demandez-vous si vous avez vraiment besoin de chaque colonne et essayez d'identifier l'avantage que chacune d'elles apporte à votre modèle sémantique. Si vous constatez qu'une colonne n'ajoute aucune valeur, vous devez la supprimer de votre modèle sémantique. Supposons, par exemple, que vous disposiez d'une colonne ID avec des milliers de lignes uniques. Comme vous n'utiliserez pas cette colonne dans une relation, elle ne sera utilisée dans aucun état. Vous devez donc considérer cette colonne comme inutile et admettre qu'elle gaspille de l'espace dans votre modèle sémantique.

Quand vous supprimez une colonne inutile, vous réduisez la taille du modèle sémantique et, par là même, la taille de fichier et la durée d'actualisation. En outre, étant donné que le modèle sémantique contient uniquement des données pertinentes, les performances globales de l'état sont améliorées.

Pour en savoir plus, consultez Techniques de réduction des données pour la modélisation des importations.

Les métadonnées sont des informations sur d'autres données. Les métadonnées Power BI comportent des informations sur votre modèle sémantique, telles que le nom, le type de données et le format de chacune des colonnes, le schéma de la base de données, la conception des états, la date de dernière modification du fichier, la fréquence d'actualisation des données, et bien plus encore.

Lorsque vous chargez des données dans Power BI Desktop, il est recommandé d'analyser les métadonnées correspondantes afin de pouvoir identifier les incohérences avec votre modèle sémantique et normaliser les données avant de commencer à créer des états. L'exécution d'une analyse sur vos métadonnées améliore les performances du modèle sémantique, car, au cours de cette opération, vous identifiez les colonnes inutiles, les erreurs au sein de vos données, les types de données incorrects, le volume de données chargées (le chargement des modèles sémantiques volumineux, notamment les données transactionnelles ou historiques, prend plus de temps) et bien plus encore.

Vous pouvez utiliser l'Éditeur Power Query dans Power BI Desktop pour examiner les colonnes, les lignes et les valeurs des données brutes. Vous pouvez ensuite utiliser les outils disponibles, tels que ceux mis en évidence dans la capture d'écran suivante, pour apporter les modifications nécessaires.

Les options Power Query sont les suivantes :

Colonnes inutiles : évalue la nécessité de chaque colonne. Si vous n'envisagez pas d'utiliser une ou plusieurs colonnes dans l'état, vous devez les supprimer à l'aide de l'option Supprimer les colonnes sous l'onglet Accueil, car elles ne sont pas nécessaires.

Lignes inutiles : vérifie les premières lignes du modèle sémantique pour voir si elles sont vides ou si elles comportent des données dont vous n'avez pas besoin dans vos états. Si tel est le cas, vous pouvez supprimer ces lignes à l'aide de l'option Supprimer les lignes de l'onglet Accueil.

Type de données : évalue les types de données de colonne pour s'assurer que chacun d'eux est correct. Si vous identifiez un type de données incorrect, changez-le en sélectionnant successivement la colonne, l'option Type de données sous l'onglet Transformer et le type de données approprié dans la liste.

Noms des requêtes : examine les noms des requêtes (tables) dans le volet Requêtes. À l'image des noms d'en-tête de colonne, vous devez changer les noms de requête inhabituels ou qui n'aident pas pour des noms qui sont plus évidents ou qui sont plus familiers à l'utilisateur. Vous pouvez renommer une requête en cliquant dessus avec le bouton droit, en sélectionnant Renommer, en modifiant le nom comme il se doit, puis en appuyant sur Entrée.

Détails de la colonne : l'éditeur Power Query dispose des trois options d'aperçu des données suivantes, qui vous permettent d'analyser les métadonnées associées à vos colonnes. Ces options se trouvent sous l'onglet Affichage, comme l'illustre la capture d'écran suivante. Qualité de la colonne : détermine le pourcentage d'éléments de la colonne qui sont valides, comportent des erreurs ou sont vides. Si le pourcentage de validité n'est pas 100, vous devez en rechercher la raison, corriger les erreurs et remplir les valeurs vides. Distribution en colonnes : affiche la fréquence et la distribution des valeurs dans chacune des colonnes. Vous étudierez cela plus loin dans ce module. Profil en colonnes : affiche un graphique de statistiques en colonnes et un graphique de distribution en colonnes.

Détails de la colonne : l'éditeur Power Query dispose des trois options d'aperçu des données suivantes, qui vous permettent d'analyser les métadonnées associées à vos colonnes. Ces options se trouvent sous l'onglet Affichage, comme l'illustre la capture d'écran suivante.

Qualité de la colonne : détermine le pourcentage d'éléments de la colonne qui sont valides, comportent des erreurs ou sont vides. Si le pourcentage de validité n'est pas 100, vous devez en rechercher la raison, corriger les erreurs et remplir les valeurs vides.

Distribution en colonnes : affiche la fréquence et la distribution des valeurs dans chacune des colonnes. Vous étudierez cela plus loin dans ce module.

Profil en colonnes : affiche un graphique de statistiques en colonnes et un graphique de distribution en colonnes.

Si vous examinez un grand modèle sémantique contenant plus de 1 000 lignes et que vous souhaitez analyser l'ensemble du modèle sémantique, vous devez changer l'option par défaut en bas de la fenêtre. Sélectionnez Profilage de la colonne en fonction des 1000 premières lignes>Profilage de colonne basé sur l'ensemble du jeu de données.

Les autres métadonnées que vous devez prendre en compte sont les informations sur le modèle sémantique dans son ensemble, telles que la taille du fichier et la fréquence d'actualisation des données. Vous pouvez trouver ces métadonnées dans le fichier Power BI Desktop (.pbix) associé. Les données que vous chargez dans Power BI Desktop sont compressées et stockées sur le disque par le moteur de stockage VertiPaq. La taille de votre modèle sémantique a un impact direct sur ses performances ; un modèle sémantique de taille inférieure utilise moins de ressources (mémoire) et permet d'accélérer l'actualisation des données, les calculs et le rendu des visuels dans les états.

Fonctionnalité Date/heure automatique

L'autre élément que vous devez prendre en compte lors de l'optimisation des performances est l'option Date/heure automatique dans Power BI Desktop. Par défaut, cette fonctionnalité est activée globalement, ce qui signifie que Power BI Desktop crée automatiquement une table calculée masquée pour chaque colonne de date, à condition que certaines conditions soient remplies. Les nouvelles tables masquées s'ajoutent aux tables que vous avez déjà dans votre modèle sémantique.

L'option Date/heure automatique vous permet d'utiliser l'Assistant Time Intelligence pour le filtrage, le regroupement et l'exploration des périodes calendaires au niveau du détail. Nous vous recommandons de garder l'option Date/heure automatique activée uniquement si vous utilisez des périodes calendaires et quand le modèle présente des exigences simples par rapport à l'heure.

Si votre source de données définit déjà une table de dimension de date, cette table doit être utilisée pour définir de manière cohérente l'heure au sein de votre organisation, et vous devez désactiver l'option Date/heure automatique globale. La désactivation de cette option peut réduire la taille de votre modèle sémantique et réduire le temps d'actualisation.

Vous pouvez activer/désactiver cette option Date/heure automatique de manière globale afin qu'elle s'applique à tous vos fichiers Power BI Desktop, ou bien l'activer ou la désactiver pour le fichier actuel afin qu'elle ne s'applique qu'à un fichier spécifique.

Pour activer/désactiver cette option Date/heure automatique, accédez à Fichier>Options et paramètres>Options, puis sélectionnez la page Global ou Fichier actuel. Sur l'une ou l'autre des pages, sélectionnez Chargement des données puis, dans la section Time Intelligence, cochez ou décochez la case en fonction des besoins.

Pour obtenir une vue d'ensemble et une présentation générale de la fonctionnalité Date/heure automatique, consultez Appliquer l'option de date/heure automatique dans Power BI Desktop.

Unité 3: Utiliser des variables pour améliorer les performances et la résolution des problèmes

Type: Contenu

Utiliser des variables pour améliorer les performances et la résolution des problèmes

Vous pouvez utiliser des variables dans vos formules DAX pour essayer d'écrire des calculs moins complexes et plus efficaces. Les variables sont sous-utilisées par les développeurs qui débutent dans Power BI Desktop, mais elles sont efficaces et vous devez les utiliser par défaut quand vous créez des mesures.

Certaines expressions impliquent l'utilisation de nombreuses fonctions imbriquées et la réutilisation de la logique d'expression. Le traitement de ces expressions prend plus de temps et celles-ci sont difficiles à lire, ce qui complique la résolution de leurs problèmes. Si vous utilisez des variables, vous pouvez économiser du temps de traitement des requêtes. Ce changement est une étape dans la bonne direction pour optimiser les performances d'un modèle sémantique.

L'utilisation de variables dans votre modèle de sémantique offre les avantages suivants :

Amélioration des performances : les variables peuvent rendre les mesures plus efficaces, car Power BI n'a plus besoin d'évaluer plusieurs fois la même expression. Vous pouvez obtenir les mêmes résultats dans une requête en environ moitié moins de temps par rapport au traitement d'origine.

Lisibilité améliorée : les variables ont des noms courts et autodescriptifs et sont utilisées à la place d'une expression à plusieurs mots ambiguë. Il peut s'avérer plus facile de lire et de comprendre les formules quand des variables sont utilisées.

Simplification du débogage : vous pouvez utiliser des variables pour déboguer une formule et des expressions de test, ce qui peut être utile lors de la résolution des problèmes.

Réduction de la complexité : les variables ne nécessitent pas l'utilisation des fonctions DAX EARLIER ou EARLIEST, qui sont difficiles à comprendre. Ces fonctions étaient nécessaires avant l'introduction des variables et étaient écrites dans des expressions complexes qui introduisaient de nouveaux contextes de filtre. Comme vous pouvez désormais utiliser des variables au lieu de ces fonctions, vous pouvez écrire moins de formules complexes.

Utiliser des variables pour améliorer les performances

Pour illustrer la façon dont vous pouvez utiliser une variable afin de rendre une mesure plus efficace, le tableau suivant présente une définition de mesure de deux manières différentes. Notez que la formule répète l'expression qui calcule « même période l'année précédente », mais de deux façons différentes : la première instance utilise la méthode de calcul DAX normale, tandis que la seconde utilise des variables dans le calcul.

La seconde ligne du tableau montre la définition de la mesure dans une version améliorée. Cette définition utilise le mot clé VAR pour introduire une variable nommée SalesPriorYear et utilise une expression pour attribuer le résultat « même période l'année précédente » à cette nouvelle variable. Elle utilise ensuite la variable à deux reprises dans l'expression DIVIDE.

Sans la variable

Avec la variable

Dans la première définition de la mesure dans le tableau, la formule n'est pas efficace, car elle nécessite que Power BI évalue deux fois la même expression. La seconde définition est plus efficace car, grâce à la variable, Power BI ne doit évaluer l'expression PARALLELPERIOD qu'une seule fois.

Si votre modèle sémantique a plusieurs requêtes avec plusieurs mesures, l'utilisation de variables peut réduire la durée globale de traitement des requêtes de moitié et améliorer les performances globales du modèle de données. En outre, cette solution est simple ; imaginez les économies au fur et à mesure que les formules deviennent plus compliquées, par exemple, quand vous manipulez des pourcentages et des totaux cumulés.

Utiliser des variables pour améliorer la lisibilité

En plus d'améliorer les performances, l'utilisation de variables peut s'avérer efficace pour simplifier la lecture du code.

Quand des variables sont utilisées, il est recommandé de leur attribuer des noms descriptifs. Dans l'exemple précédent, la variable est appelée SalesPriorYear (Ventes de l'année antérieure), indiquant

clairement ce qu'elle calcule. En revanche, l'utilisation d'une variable appelée X, temp ou variable1 n'aurait pas du tout mis en évidence l'objectif de la variable.

L'utilisation de noms clairs, concis et explicites vous permet de mieux comprendre ce que vous essayez de calculer, et il sera beaucoup plus simple pour d'autres développeurs de gérer le rapport par la suite.

Utiliser des variables pour résoudre plusieurs étapes

Vous pouvez utiliser des variables pour essayer de déboguer une formule et identifier le problème. Vous pouvez simplifier la résolution des problèmes de votre calcul DAX à l'aide de variables en évaluant chacune d'elles séparément et en les rappelant après l'expression RETURN.

Dans l'exemple suivant, vous testez une expression qui est attribuée à une variable. À des fins de débogage, vous réécrivez provisoirement l'expression RETURN à écrire dans la variable. La définition de la mesure retourne uniquement la variable SalesPriorYear, car c'est ce qui vient après l'expression RETURN.

L'expression RETURN affiche uniquement la valeur SalesPriorYear%. Cette technique vous permet de rétablir l'expression quand vous avez terminé le débogage. En outre, elle simplifie la compréhension des calculs en raison de la réduction de la complexité du code DAX.

Unité 4: Réduire la cardinalité

Type: Contenu

Réduire la cardinalité

La cardinalité est un terme utilisé pour décrire l'unicité des valeurs d'une colonne. La cardinalité est également utilisée dans le contexte des relations entre deux tables, où elle décrit la direction de la relation.

Identifier les niveaux de cardinalité dans les colonnes

Auparavant, quand vous utilisiez l'Éditeur Power Query pour analyser les métadonnées, l'option Distribution des colonnes de l'onglet Affichage présentait des statistiques sur le nombre d'éléments distincts et uniques présents dans chaque colonne des données.

Nombre de valeurs distinctes : nombre total de valeurs différentes trouvées dans une colonne donnée.

Nombre de valeurs uniques : nombre total de valeurs qui n'apparaissent qu'une seule fois dans une colonne donnée.

Une colonne dont la plage contient de nombreuses valeurs répétées (le nombre de valeurs uniques est faible) présente un niveau de cardinalité bas. À l'inverse, une colonne dont la plage contient de nombreuses valeurs uniques (le nombre de valeurs uniques est élevé) présente un niveau de cardinalité élevé.

Une cardinalité plus faible occasionne un niveau de performance plus optimisé, vous allez peut-être devoir réduire le nombre de colonnes présentant un niveau de cardinalité élevé dans votre modèle sémantique.

Réduire la cardinalité des relations

Quand vous importez plusieurs tables, il est possible que vous soyez amené à effectuer des analyses impliquant les données de ces tables. Les relations entre ces tables sont nécessaires pour obtenir des

résultats précis et afficher les informations correctes dans vos rapports. Power BI Desktop contribue à faciliter la création de ces relations. De fait, dans la plupart des cas, vous n'avez rien à faire ; la fonctionnalité de détection automatique se charge de tout. Toutefois, vous pouvez parfois être amené à créer des relations ou à apporter des modifications à une relation. Indépendamment du cas de figure, il est important de comprendre le fonctionnement des relations dans Power BI Desktop et comment créer et modifier des relations.

Quand vous créez ou modifiez une relation, vous pouvez configurer des options supplémentaires. Par défaut, Power BI Desktop configure automatiquement des options supplémentaires en fonction de sa meilleure estimation, qui peut varier d'une relation à l'autre suivant les données contenues dans les colonnes.

Les relations peuvent avoir une cardinalité différente. La cardinalité est la direction de la relation, et chaque relation de modèle doit être définie avec un type de cardinalité. Les options de cardinalité dans Power BI sont les suivantes :

Plusieurs à un (*:1) : cette relation est le type par défaut le plus courant. La colonne d'une table peut avoir plusieurs instances d'une valeur, tandis que la table connexe, souvent appelée table de recherche, n'a qu'une seule instance d'une valeur donnée.

Un à un (1:1) : dans ce type de relation, la colonne d'une table n'a qu'une seule instance d'une valeur donnée, tandis que la table connexe n'a qu'une seule instance d'une valeur donnée.

Un à plusieurs (1:) : dans ce type de relation, la colonne d'une table n'a qu'une seule instance d'une valeur donnée, tandis que la table connexe peut avoir plusieurs instances d'une valeur.

Plusieurs-à-plusieurs (:) avec les modèles composites, vous pouvez établir une relation plusieurs-à-plusieurs entre des tables, qui peuvent dès lors contenir des valeurs non uniques. Cette option supprime également les solutions de contournement précédentes, telles que l'introduction de nouvelles tables uniquement pour établir des relations.

Pendant le développement, vous créez et modifiez des relations dans votre modèle. Ainsi, quand vous générez de nouvelles relations dans votre modèle, quelle que soit la cardinalité choisie, vous devez toujours vous assurer que les deux colonnes impliquées dans une relation partagent le même type de données. Votre modèle ne fonctionnera jamais si vous essayez de créer une relation entre deux colonnes, dont une colonne a un type de données texte et l'autre un type de données entier.

Dans l'exemple suivant, le champ ProductID a le type de données Nombre entier dans les tables Product et Sales. Les colonnes avec le type de données Entier offrent de meilleures performances que les colonnes avec le type de données Texte.

Améliorer les performances en réduisant les niveaux de cardinalité

Power BI Desktop offre différentes techniques que vous pouvez utiliser pour réduire les données chargées dans des modèles sémantiques, tels que le résumé. En réduisant les données chargées dans votre modèle, vous améliorez la cardinalité des relations du rapport. Il est donc important que vous cherchiez à réduire les données à charger dans vos modèles. Cela est particulièrement vrai pour les modèles de grande taille ou les modèles appelés à croître au fil du temps.

La technique la plus efficace pour réduire la taille d'un modèle consiste peut-être à utiliser une table de résumé à partir de la source de données. Alors qu'une table de détails peut contenir chaque transaction, une table de résumé peut contenir un enregistrement par jour, par semaine ou par mois. Il peut s'agir d'une moyenne de toutes les transactions par jour, par exemple.

Ainsi, une table source de faits de ventes stocke une ligne pour chaque ligne de commande. Pour réduire significativement les données, vous pouvez résumer toutes les métriques de ventes en effectuant des regroupements par date, client et produit, si les détails des transactions individuelles ne sont pas nécessaires.

Vous pouvez ensuite obtenir une réduction encore plus significative des données en effectuant un regroupement par date au niveau du mois. Vous pourriez ainsi atteindre une réduction de 99 % de la taille du modèle. Toutefois, la création de rapports au niveau du jour ou d'une commande spécifique n'est alors plus possible. Le fait de résumer les données de type fait implique toujours un compromis avec les détails de vos données. Un inconvénient est que vous risquez de perdre la possibilité d'effectuer des recherches dans les données, car les détails n'existent plus. Ce compromis peut être atténué à l'aide d'une conception de type mode mixte.

Dans Power BI Desktop, une conception de type mode mixte produit un modèle composite. Pour l'essentiel, il vous permet de déterminer un mode de stockage pour chaque table. Ainsi, chaque table peut avoir sa propriété Mode de stockage définie sur Importer ou DirectQuery.

Une technique efficace pour réduire la taille du modèle consiste à définir sur DirectQuery la propriété Mode de stockage pour les tables de type de faits plus grandes. Vous pouvez combiner cette approche de conception avec les techniques utilisées pour résumer vos données. Par exemple, les données de ventes résumées peuvent être utilisées pour la création de rapports « résumés » hautes performances. Vous pouvez créer une page d'extraction afin d'afficher les ventes précises pour un contexte de filtre spécifique (et étroit), dans laquelle sont reprises toutes les commandes client propres à ce contexte. La page d'extraction inclut des visuels basés sur une table DirectQuery qui permettent de récupérer les données des commandes client (détails des commandes client).

Pour plus d'informations, consultez Techniques de réduction des données pour la modélisation des importations.

Unité 5: Optimiser les modèles DirectQuery avec un stockage au niveau de la table

Type: Contenu

Optimiser les modèles DirectQuery avec un stockage au niveau de la table

DirectQuery est un moyen d'obtenir des données dans Power BI Desktop. La méthode DirectQuery implique une connexion directe aux données de son référentiel source depuis Power BI Desktop. Il s'agit d'une alternative à l'importation de données dans Power BI Desktop.

Quand vous utilisez la méthode DirectQuery, l'expérience utilisateur globale dépend fortement des performances de la source de données sous-jacente. La lenteur des temps de réponse aux requêtes peut nuire à l'expérience utilisateur et, dans le pire des cas, les requêtes peuvent expirer. De plus, le nombre d'utilisateurs qui ouvrent les rapports en même temps alourdira la charge pesant sur la source de données. Par exemple, si votre rapport contient 20 visuels et que 10 personnes utilisent le rapport, 200 requêtes ou plus existent sur la source de données, car chaque visuel émet une ou plusieurs requêtes.

Malheureusement, les performances de votre modèle Power BI ne seront pas seulement affectées par les performances de la source de données sous-jacente, mais également par d'autres facteurs incontrôlables, tels que les suivants :

Latence du réseau : les réseaux plus rapides retournent les données plus rapidement.

Les performances du serveur de la source de données et la quantité d'autres charges de travail sur ce serveur. Par exemple, considérez les implications de l'actualisation d'un serveur pendant que des centaines de personnes utilisent ce même serveur pour différentes raisons.

Ainsi, l'utilisation de DirectQuery représente un risque pour la qualité des performances de votre modèle. Pour optimiser les performances dans ce cas, vous devez contrôler la base de données

source ou y accéder.

Pour en savoir plus, consultez le Guide du modèle DirectQuery dans Power BI Desktop.

Implications de l'utilisation de DirectQuery

Il est recommandé d'importer des données dans Power BI Desktop, mais votre organisation peut avoir besoin d'utiliser le mode de connectivité des données DirectQuery pour l'une des raisons suivantes (avantages de DirectQuery) :

Il est approprié dans les cas où les données changent fréquemment et que la création de rapports en quasi-temps réel est nécessaire.

Il peut gérer des données volumineuses sans qu'il soit nécessaire d'effectuer une pré-agrégation.

Il applique des restrictions de souveraineté des données pour se conformer aux exigences légales.

Il peut être utilisé avec une source de données multidimensionnelle contenant des mesures telles que SAP Business Warehouse (BW).

Si votre organisation doit utiliser DirectQuery, vous devez clairement comprendre son comportement dans Power BI Desktop et bien connaître ses limites. Vous pourrez alors prendre des mesures pour optimiser le modèle DirectQuery autant que possible.

Comportement des connexions DirectQuery

Lorsque vous utilisez DirectQuery pour accéder aux données dans Power BI Desktop, cette connexion se comporte de la manière suivante :

Lorsque vous utilisez initialement la fonctionnalité Obtenir les données dans Power BI Desktop, vous sélectionnez la source. Si vous vous connectez à une source relationnelle, vous pouvez sélectionner un ensemble de tables, et chacune d'entre elles définit une requête qui retourne un jeu de données de façon logique. Si vous sélectionnez une source multidimensionnelle, par exemple SAP BW, vous ne pouvez sélectionner que celle-ci.

Lorsque vous chargez les données, aucune donnée n'est importée dans Power BI Desktop, seul le schéma est chargé. Lorsque vous créez un visuel dans Power BI Desktop, des requêtes sont envoyées à la source sous-jacente pour récupérer les données nécessaires. Le temps nécessaire à l'actualisation du visuel dépend des performances de la source de données sous-jacente.

Si des modifications sont apportées aux données sous-jacentes, elles ne s'afficheront pas immédiatement dans les visuels existants dans Power BI en raison de la mise en cache. Vous devez effectuer une actualisation pour voir ces modifications. Les requêtes nécessaires sont présentes pour chaque visuel, et les visuels sont mis à jour en conséquence.

Lorsque vous publiez le rapport dans le service Power BI, cela génère un modèle sémantique dans le service Power BI, identique à celui de l'importation. Cependant, aucune donnée n'est comprise dans ce modèle sémantique.

Lorsque vous ouvrez un rapport existant dans le service Power BI ou que vous en créez un, la source sous-jacente est à nouveau interrogée pour récupérer les données nécessaires. Selon l'endroit où se trouve la source d'origine, vous devrez peut-être configurer une passerelle de données locale.

Vous pouvez épingler des visuels ou des pages de rapport entières sous forme de vignettes de tableau de bord. Les vignettes sont actualisées automatiquement selon une planification, par exemple toutes les heures. Vous pouvez contrôler la fréquence de cette actualisation en fonction de vos besoins. Quand vous ouvrez un tableau de bord, les vignettes reflètent les données au moment de la dernière actualisation et peuvent ne pas inclure les dernières modifications apportées à la source de données sous-jacente. Vous pouvez toujours actualiser un tableau de bord ouvert pour vous assurer qu'il est à

jour.

Limitations des connexions DirectQuery

L'utilisation de DirectQuery peut avoir des conséquences négatives. Les limitations varient selon la source de données utilisée. Tenez compte des points suivants :

Performances : comme indiqué plus haut, votre expérience utilisateur globale dépend fortement des performances de la source de données sous-jacente.

Sécurité : si vous utilisez plusieurs sources de données dans un modèle DirectQuery, il est important de comprendre comment les données se déplacent entre les sources de données sous-jacentes et les implications de sécurité associées. Vous devez également identifier si les règles de sécurité s'appliquent aux données de votre source sous-jacente car, dans Power BI, chaque utilisateur peut voir ces données.

Transformation des données : par rapport aux données importées, les données provenant de DirectQuery présentent des limites lors de l'application des techniques de transformation de données dans l'Éditeur Power Query. Par exemple, si vous vous connectez à une source OLAP, telle que SAP BW, vous ne pouvez pas effectuer de transformations ; l'intégralité du modèle externe est extraite de la source de données. Si vous souhaitez appliquer des transformations aux données, vous devez le faire dans la source de données sous-jacente.

Modélisation : certaines fonctionnalités de modélisation disponibles avec des données importées ne sont pas utilisables ou sont limitées dans le cadre de DirectQuery.

Reporting : la quasi-totalité des fonctionnalités de reporting disponibles avec des données importées sont également prises en charge pour les modèles DirectQuery, à condition que la source sous-jacente offre un niveau de performance approprié. Cependant, lorsque le rapport est publié dans le service Power BI, les fonctionnalités Quick Insights et Q&R ne sont pas prises en charge. En outre, l'utilisation de la fonctionnalité Explorer dans Excel entraîne probablement une baisse des performances.

Pour en savoir plus sur les limitations liées à l'utilisation de DirectQuery, consultez Implications de l'utilisation de DirectQuery.

Le fonctionnement de base de DirectQuery et les limitations qu'il présente n'ayant plus de secret pour vous, vous pouvez prendre des mesures pour améliorer les performances.

Optimiser les performances

Pour reprendre le scénario de Tailwind Traders, lors de votre examen du modèle sémantique, vous découvrez que la requête a utilisé DirectQuery pour connecter Power BI Desktop aux données sources. Cette utilisation de DirectQuery est la raison pour laquelle les utilisateurs sont confrontés à des performances de rapport médiocres. Le chargement des pages dans le rapport prend trop de temps et les tables ne sont pas actualisées suffisamment rapidement quand certaines sélections sont es. Vous devez prendre des mesures pour optimiser les performances du modèle DirectQuery.

Vous pouvez examiner les requêtes qui sont envoyées à la source sous-jacente afin d'essayer d'identifier la raison pour laquelle les performances de requête sont médiocres. Vous pouvez ensuite effectuer des modifications dans Power BI Desktop et la source de données sous-jacente pour optimiser les performances globales.

Optimiser les données dans Power BI Desktop

Quand vous avez optimisé la source de données autant que possible, vous pouvez prendre d'autres mesures dans Power BI Desktop en utilisant l'Analyseur de performances, dans lequel vous pouvez isoler les requêtes pour valider les plans de requête.

Vous pouvez analyser la durée des requêtes qui sont envoyées à la source sous-jacente pour identifier les requêtes dont le chargement prend beaucoup de temps. En d'autres termes, vous pouvez identifier où se trouvent les goulots d'étranglement.

Vous n'avez pas besoin d'utiliser une approche spéciale pour optimiser un modèle DirectQuery ; vous pouvez appliquer les mêmes techniques d'optimisation que celles que vous avez utilisées sur les données importées pour ajuster les données provenant de la source DirectQuery. Par exemple, vous pouvez réduire le nombre de visuels dans la page de rapport ou réduire le nombre de champs utilisés dans un visuel. Vous pouvez également supprimer les colonnes et les lignes inutiles.

Pour obtenir des conseils plus détaillés sur la façon d'optimiser une requête DirectQuery, consultez : Guide du modèle DirectQuery dans Power BI Desktop et Conseils pour utiliser DirectQuery avec succès.

Optimiser la source de données sous-jacente (base de données connectée)

Vous devez d'abord vous pencher sur la source de données. Vous devez optimiser la base de données source autant que possible, car tout ce qui permet d'améliorer ses performances améliorera également DirectQuery Power BI. Les actions que vous effectuez dans la base de données sont les plus efficaces.

Envisagez l'utilisation des pratiques de base de données standard ci-dessous, qui s'appliquent à la plupart des situations :

Évitez l'utilisation de colonnes calculées complexes, car l'expression de calcul est incorporée dans les requêtes source. Il est plus efficace de renvoyer (push) l'expression à la source, car cela évite l'envoi (push down). Vous pouvez également envisager d' des colonnes clés de substitution à des tables de type dimension.

Examinez les index et vérifiez que l'indexation actuelle est correcte. Si vous avez besoin de créer des index, assurez-vous qu'ils sont appropriés.

Reportez-vous aux documents d'aide de votre source de données afin d'implémenter leurs recommandations en matière de performances.

Personnaliser les options de réduction de requête

Power BI Desktop vous donne la possibilité d'envoyer moins de requêtes et de désactiver certaines interactions qui nuiront à l'expérience si les requêtes qui en résultent mettent longtemps à s'exécuter. L'application de ces options empêche les requêtes d'atteindre continuellement la source de données, ce qui devrait améliorer les performances.

Dans cet exemple, vous modifiez les paramètres par défaut pour appliquer les options de réduction de données disponibles à votre modèle. Pour accéder aux paramètres, sélectionnez Fichier>Options et paramètres>Options, puis faites défiler la page vers le bas et sélectionnez l'option Réduction des requêtes.

Les options de requête suivantes sont disponibles :

Réduire le nombre de requêtes envoyées par : par défaut, chaque visuel interagit avec tous les autres visuels. Si cette case est cochée, l'interaction par défaut est désactivée. Vous pouvez ensuite choisir les visuels qui interagissent les uns avec les autres à l'aide de la fonctionnalité Modifier les interactions.

Segments : par défaut, l'option Appliquer instantanément les changements de segment est sélectionnée. Pour forcer les utilisateurs du rapport à appliquer manuellement les changements de segment, sélectionnez l'option un bouton Appliquer à chaque segment pour apporter les changements quand vous êtes prêt.

Filtres : par défaut, l'option Appliquer instantanément les changements de filtre de base est sélectionnée. Pour forcer les utilisateurs du rapport à appliquer manuellement les changements de

filtre, sélectionnez l’une des options suivantes : des boutons Appliquer à chaque filtre de base pour appliquer les changements au besoin un seul bouton Appliquer au volet Filtre pour appliquer tous les changements à la fois (version préliminaire)

Filtres : par défaut, l’option Appliquer instantanément les changements de filtre de base est sélectionnée. Pour forcer les utilisateurs du rapport à appliquer manuellement les changements de filtre, sélectionnez l’une des options suivantes :

des boutons Appliquer à chaque filtre de base pour appliquer les changements au besoin

un seul bouton Appliquer au volet Filtre pour appliquer tous les changements à la fois (version préliminaire)

Unité 6: Créer et gérer des agrégations

Type: Contenu

Créer et gérer des agrégations

L’agrégation de données consiste à résumer ces dernières et à les présenter à un niveau plus général. Par exemple, vous pouvez résumer toutes les données de ventes et les regrouper par date, client, produit, etc. Le processus d'agrégation réduit la taille des tables dans le modèle sémantique, ce qui vous permet de vous concentrer sur les données importantes et vous aide à améliorer les performances des requêtes.

Votre organisation peut décider d'utiliser des agrégations dans ses modèles sémantiques pour les raisons suivantes :

Si vous utilisez une grande quantité de données (Big Data), les agrégations fournissent de meilleures performances de requête et vous aident à analyser et à révéler les insights de ces données volumineuses. Les données agrégées étant mises en cache, elles utilisent beaucoup moins de ressources que celles nécessaires pour les données détaillées.

En cas d’actualisation lente, les agrégations vous aident à accélérer le processus d’actualisation. La taille de cache étant plus petite, le temps d’actualisation est réduit et les utilisateurs disposent des données plus rapidement. Au lieu d’actualiser ce qui pourrait être des millions de lignes, vous devez actualiser une plus petite quantité de données.

Si vous avez un grand modèle sémantiques, les agrégations peuvent vous aider à réduire et à maintenir sa taille.

Si vous pensez que la taille de votre modèle sémantiques est appelée à augmenter, vous pouvez utiliser des agrégations afin de protéger de manière proactive votre modèle de données contre les risques éventuels de problèmes de performances, d'actualisation ainsi que de problèmes de requête globaux.

Dans le cadre du scénario Tailwind Traders, vous avez pris plusieurs mesures pour optimiser les performances du modèle sémantique, mais l'équipe informatique vous a informé que la taille du fichier est toujours trop grande. La taille du fichier est actuellement de 1 gigaoctet (Go). Vous devez donc la réduire à environ 50 mégaoctets (Mo). Pendant votre examen des performances, vous avez identifié que le développeur précédent n'a pas utilisé d'agrégations dans le modèle sémantique. Vous devez donc créer des agrégations pour les données de ventes afin de réduire la taille de fichier et optimiser les performances.

Créer des agrégations

Avant de commencer à créer des agrégations, vous devez choisir le degré de précision (niveau) avec lequel vous souhaitez les créer. Dans cet exemple, vous souhaitez agréger les données de ventes par jour.

Une fois le degré de précision choisi, l'étape suivante consiste à décider de la façon dont vous souhaitez créer les agrégations. Vous pouvez créer les agrégations de différentes façons et chaque méthode produit les mêmes résultats, par exemple :

Si vous avez accès à la base de données, vous pouvez créer une table avec l'agrégation, puis importer cette table dans Power BI Desktop.

Si vous avez accès à la base de données, vous pouvez créer une vue pour l'agrégation, puis importer cette vue dans Power BI Desktop.

Dans Power BI Desktop, vous pouvez utiliser l'Éditeur Power Query pour créer les agrégations pas à pas.

Dans cet exemple, vous ouvrez une requête dans l'Éditeur Power Query et remarquez que les données n'ont pas été agrégées ; il y a plus de 999 lignes, comme le montre la capture d'écran suivante.

Vous souhaitez agréger les données en fonction de la colonne OrderDate et afficher les colonnes OrderQuantity et SalesAmount. Commencez par sélectionner Choisir des colonnes sous l'onglet Accueil. Dans la fenêtre qui s'affiche, sélectionnez les colonnes que vous souhaitez dans l'agrégation, puis sélectionnez OK.

Quand les colonnes sélectionnées s'affichent dans la page, sélectionnez l'option Regrouper par sous l'onglet Accueil. Dans la fenêtre qui s'affiche, sélectionnez la colonne en fonction de laquelle vous souhaitez regrouper (OrderDate) et entrez un nom pour la nouvelle colonne (OnlineOrdersCount).

Sélectionnez l'option Avancé, puis sélectionnez le bouton une agrégation pour afficher une autre ligne de colonne. Entrez un nom pour la colonne d'agrégation, sélectionnez l'opération de la colonne, puis sélectionnez la colonne à laquelle vous souhaitez lier l'agrégation. Répétez ces étapes jusqu'à ce que vous ayez ajouté toutes les agrégations, puis sélectionnez OK.

Une fois l'agrégation affichée (au terme d'un délai pouvant prendre quelques minutes), vous voyez comment les données ont été transformées. Les données sont agrégées pour chaque date, avec pour chacune le nombre de commandes ainsi que les sommes respectives du montant des ventes et de la quantité commandée.

Sélectionnez le bouton Fermer et appliquer pour fermer l'Éditeur Power Query et appliquer les modifications apportées à votre modèle sémantique. Revenez à la page Power BI Desktop, puis sélectionnez le bouton Actualiser pour voir les résultats. Observez l'écran, car un bref message affiche le nombre de lignes que contient désormais votre modèle sémantique. Ce nombre de lignes doit être considérablement inférieur au nombre initial. Vous pouvez également voir ce nombre quand vous rouvrez l'Éditeur Power Query, comme l'illustre la capture d'écran suivante. Dans cet exemple, le nombre de lignes a été réduit à 30.

N'oubliez pas que vous avez commencé avec plus de 999 lignes. L'utilisation de l'agrégation a considérablement réduit le nombre de lignes dans votre modèle sémantique, ce qui signifie que Power BI a moins de données à actualiser et que votre modèle doit être plus performant.

Gérer les agrégations

Quand vous avez créé des agrégations, vous pouvez les gérer dans Power BI Desktop et apporter des modifications à leur comportement, si nécessaire.

Vous pouvez ouvrir la fenêtre Gérer les agrégations à partir de n'importe quelle vue dans Power BI Desktop. Dans le volet Champs, cliquez avec le bouton droit sur la table, puis sélectionnez Gérer les agrégations.

Pour chaque colonne d’agrégation, vous pouvez sélectionner une option dans la liste déroulante Résumé et apporter des modifications à la table et à la colonne de détails sélectionnées. Quand vous avez terminé de gérer les agrégations, sélectionnez Appliquer tout.

Pour plus d’informations sur la création et la gestion des agrégations, consultez Utiliser des agrégations dans Power BI Desktop.

Unité 7: Contrôle de vos connaissances

Type: Contenu

Contrôle de vos connaissances

Répondez aux questions suivantes pour tester les connaissances que vous avez acquises.

Quel avantage procure l’analyse des métadonnées ?

L’analyse des métadonnées vous permet d’identifier clairement les incohérences dans les données avec votre modèle sémantique.

L’analyse des métadonnées vous permet de vous familiariser avec vos données.

L’analyse des métadonnées vous permet de connaître le nombre de lignes, de colonnes et de tables chargées dans votre modèle.

Quel est l’intérêt de supprimer les lignes et colonnes inutiles ?

Il n’est pas nécessaire de supprimer les lignes et colonnes inutiles ; il est recommandé de conserver toutes les métadonnées intactes.

La suppression des lignes et colonnes inutiles réduit la taille d’un modèle sémantique ; il est recommandé de charger uniquement les données nécessaires dans votre modèle sémantique.

La suppression des lignes et colonnes inutiles peut endommager la structure du modèle sémantique.

Laquelle des affirmations suivantes sur les relations dans Power BI Desktop est vraie ?

Des relations ne peuvent être créées qu’entre des colonnes comportant le même type de données.

Des relations ne peuvent être créées qu’entre des tables comportant le même nombre de lignes.

Des relations peuvent être créées entre des tables comportant différents types de données.

Vous devez répondre à toutes les questions avant de vérifier votre travail.

Unité 8: Résumé

Type: Résumé

Dans le scénario de ce module, l’un des modèles sémantiques Power BI Desktop de votre organisation était inefficace et causait des problèmes. Les utilisateurs n’étaient pas satisfaits des performances de rapport et la taille du fichier du modèle était trop grande, au point de peser sur les ressources de l’organisation.

Vous avez été invité à examiner le modèle sémantique pour identifier la cause des problèmes de performances et apporter des modifications pour optimiser celles-ci et réduire la taille du modèle.

Power BI Desktop fournit une gamme d'outils et de fonctionnalités qui vous permettent d'analyser et d'optimiser les performances de ses modèles sémantiques. Vous avez démarré le processus d'optimisation à l'aide de l'Analyseur de performances et d'autres outils pour examiner les performances des mesures, des relations et des visuels, puis avez apporté des améliorations en fonction des résultats de l'analyse. Ensuite, vous avez utilisé des variables pour écrire des calculs moins complexes et plus efficaces. Vous avez ensuite examiné de plus près la distribution des colonnes et réduit la cardinalité de vos relations. À ce stade, le modèle sémantique était plus optimisé. Vous avez pris en compte dans quelle mesure la situation serait différente si votre organisation utilisait un modèle DirectQuery, puis vous avez identifié comment optimiser les performances de Power BI Desktop et de la base de données source. Enfin, vous avez utilisé des agrégations pour réduire considérablement la taille du modèle sémantique.

Si Power BI Desktop ne vous permettait pas d'optimiser des modèles sémantiques inefficaces, vous devriez consacrer beaucoup de temps à vos sources de données pour y améliorer les données. En particulier, sans l'Analyseur de performances, vous n'auriez pas identifié les causes des problèmes de performances dans vos rapports et les goulots d'étranglement dans les requêtes qui doivent être supprimés. Les utilisateurs peuvent s'en trouver frustrés et démotivés, au point éventuellement d'éviter d'utiliser les rapports.

Le rapport étant optimisé, les utilisateurs peuvent accéder aux données dont ils ont besoin dans un délai plus court ; ainsi, ils sont plus productifs et affichent une plus grande satisfaction au travail. La réduction que vous avez apportée à la taille de fichier du modèle soulage les ressources, procurant une série d'avantages à votre organisation. Vous avez correctement accompli la tâche qui vous a été confiée.

Utiliser l'analyseur de performances pour examiner les performances des éléments de rapport

Appliquer l'option de date/heure automatique dans Power BI Desktop

Techniques de réduction des données pour la modélisation des importations

Guide du modèle DirectQuery dans Power BI Desktop

Utiliser des agrégations dans Power BI Desktop

Module 4: Créer et gérer des ressources Power BI

Unité 1: Présentation

Type: Introduction

L'élaboration de rapports Power BI peut être simple ou complexe en fonction de vos besoins. Vous pouvez créer des ressources que vous utilisez à différentes fins et réduire les tâches répétitives ou prenant beaucoup de temps.

Imaginons que vous assurez le support du département financier d'une entreprise mondiale, et que toutes les données nécessaires sont stockées dans un seul entrepôt de données. Lorsque vous vous connectez à cette source de données avec Power BI Desktop, le modèle de données est volumineux et différentes régions ont des besoins différents. Vous vous retrouvez à générer un rapport volumineux avec de nombreuses pages et des performances poussives.

Power BI vous permet de diviser les données en plusieurs modèles sémantiques spécialisés, tous « chaînés » à un modèle sémantique de base. Cela a pour effet d'offrir une certaine flexibilité aux différentes régions et de réduire l'impact sur les performances de l'ensemble des rapports. Maintenant que vous avez des modèles sémantiques faisant référence à d'autres modèles sémantiques, vous pouvez utiliser un affichage de traçabilité dans le service Power BI pour explorer les relations. Si vous travaillez sur vos modèles sémantiques avec des outils externes, vous pouvez également utiliser le point de terminaison XMLA pour connecter et gérer des modèles sémantiques.

Dans ce module, nous allons vous montrer comment :

Créez des modèles sémantiques principaux et spécialisés et connectez-vous-y.

Créez des fichiers .pbit de modèle Power BI.

Suivez et gérez l'élaboration de rapports avec des fichiers Power BI Project.

Découvrez les relations entre les sources de données, les modèles sémantiques et les rapports.

Connectez-vous à des modèles sémantiques et gérez-les via un point de terminaison XMLA.

À la fin de ce module, vous devriez être en mesure de créer des modèles sémantiques principaux et spécialisés, d'explorer avec le service Power BI un affichage de traçabilité et d'utiliser un point de terminaison XMLA pour vous connecter à des modèles sémantiques.

Unité 2: Créer des ressources Power BI réutilisables

Type: Contenu

Créer des ressources Power BI réutilisables

Lors de la présentation initiale de Power BI, vous avez probablement appris à récupérer des données à partir de diverses sources, telles que des bases de données SQL Server, des fichiers Excel, voire des fichiers texte. La création d'un modèle sémantique identique ou similaire pour chaque rapport est fastidieuse. Par ailleurs, si d'autres personnes souhaitent également développer des rapports sur ce même modèle sémantique, quel est votre processus de partage ?

Créer des ressources de base réutilisables

Les ressources peuvent signifier des modèles sémantiques, des flux de données, des rapports et des tableaux de bord, et nous entendons généralement les modèles sémantiques dans ce module. Dans notre exemple précédent, vous avez passé un temps précieux à organiser le modèle sémantique parfait, et maintenant vous pouvez le partager et le réutiliser. Avant de créer des objets visuels, publiez le fichier sur le service Power BI et créez effectivement un modèle sémantique principal.

La prochaine fois qu'une personne doit créer un rapport sur ce modèle sémantique, elle peut se connecter à un modèle sémantique Power BI à partir de Power BI Desktop. Si vous le publiez dans un emplacement partagé, des pairs pourront utiliser le même modèle sémantique. Un modèle sémantique partagé unique protège l'intégrité des données. Ce modèle sémantique digne de confiance empêche également les modèles sémantiques orphelins lorsqu'une personne recrée un rapport sans supprimer les copies antérieures. En tant qu'analyste de données d'entreprise, il est de votre responsabilité d'être un bon gestionnaire de données pour promouvoir la démocratisation des données plutôt que la prolifération de copies disparates de données de qualité douteuse.

Créer des modèles sémantiques spécialisés

Dans notre exemple introductif, nous avons reconnu la nécessité de réduire la taille de rapport et de fournir des rapports ciblés dans les régions. Pour atteindre ces objectifs, nous pouvons créer des modèles sémantiques spécialisés. Envisagez de créer un modèle sémantique avec les données régionales spécifiques, au lieu de limiter l'accès des utilisateurs régionaux à un seul rapport global.

Pour créer un modèle sémantique spécialisé, ouvrez une nouvelle instance de Power BI Desktop et connectez-vous au modèle sémantique principal. Vous êtes maintenant connecté en direct à ce modèle sémantique, qui ne permet pas de modifier le modèle, mais vous permet de filtrer et visualiser des données. Utilisez l'option Apporter des modifications au modèle pour passer d'une connexion active à DirectQuery.

Dans la capture d'écran suivante, il existe trois tables du modèle sémantique connecté et deux nouvelles tables ajoutées. Vous pouvez des mesures, des groupes de calcul et bien plus encore.

Vous avez besoin d'une passerelle de données configurée pour prendre en charge le modèle DirectQuery si vous souhaitez apporter des modifications au modèle sémantique sous-jacent. Pour plus d'informations, consultez la documentation de Présentation de la passerelle de données locale.

Fichiers modèles Power BI

Les fichiers .pbit de modèle Power BI vous permettent d'enregistrer un rapport à réutiliser ou à partager à des fins différentes. Ces fichiers de modèle sont hautement personnalisables, en fonction de vos besoins. Lorsque vous avez enregistré un fichier de modèle, il contient les mêmes informations que si vous aviez enregistré un fichier .pbix traditionnel.

Vous pouvez créer un modèle de conception de rapport standardisé. Considérez que votre organisation commence à utiliser Power BI et que les utilisateurs apprennent toujours à interagir avec les rapports. Par exemple, vous pouvez créer un modèle avec un en-tête pour le titre, la description et l'emplacement des segments applicables. Vous pouvez maintenant enregistrer et fermer le modèle, puis ouvrir le modèle pour commencer à créer votre rapport. Si vous avez un rapport qui a déjà votre format souhaité, vous pouvez également supprimer toutes les sources de données et l'enregistrer afin que seuls le thème et les configurations visuelles soient stockés.

Vous pouvez également utiliser un modèle comme point de départ pour les rapports futurs ou comme copie étalon pour conserver l'état avant de modifier le rapport. Dans ce scénario, vous pouvez créer l'intégralité du rapport, puis enregistrer en tant que modèle sous forme de copie supplémentaire du fichier. Bien qu'il ne soit pas recommandé d'utiliser un fichier de modèle pour le contrôle de version, il peut être utile lorsque vous êtes toujours à l'étape de développement du rapport ou que vous collaborez sur le rapport et souhaitez une sauvegarde. Étant donné que les modèles Power BI s'ouvrent en tant que fichier non enregistré, les modifications apportées n'affectent pas le fichier de

modèle, sauf si vous enregistrez et remplacez spécifiquement le fichier de modèle.

Pour plus d'informations, consultez la documentation Créer et utiliser des modèles de rapport dans Power BI Desktop.

Unité 3: Utiliser la vue de la lignée et approuver les actifs de données

Type: Contenu

Utiliser la vue de la lignée et approuver les actifs de données

Nous avons abordé comment créer des modèles sémantiques Power BI de base et spécialisés à distribuer au sein de l'organisation. À présent, les développeurs de rapports peuvent se connecter directement à ces modèles sémantiques organisés, mais ils peuvent ne pas savoir quels modèles sémantiques ou autres éléments doivent être utilisés. Les modèles sémantiques chaînés peuvent également poser un problème si les développeurs de rapports ne savent pas quelle source de données ou le modèle sémantique de base est lié à un rapport. Vous pouvez résoudre ces deux problèmes dans le service Power BI avec la vue Approbation et Traçabilité.

Approuver des modèles sémantiques

Les utilisateurs peuvent découvrir d'autres modèles sémantiques dans le service Power BI et créer leurs propres rapports. Vous pouvez utiliser cette fonctionnalité et faciliter la collaboration via l'approbation. L'approbation aide les utilisateurs à trouver du contenu de haute qualité dont ils ont besoin en identifiant le contenu comme vérifié et fiable.

Les éléments peuvent recevoir l'un des trois badges d'approbation Promu, Certifié ou Données de base :

Promu signifie que les créateurs d'éléments pensent que l'élément est prêt pour le partage et la réutilisation. Tout utilisateur disposant d'autorisations d'écriture sur un élément peut le promouvoir.

Certifié signifie qu'un réviseur autorisé par l'organisation a certifié que l'élément répond aux normes de qualité de l'organisation, peut être considéré comme fiable et faisant autorité et est prêt à être utilisé dans l'ensemble de l'organisation.

Données principales signifie que les données de l'élément sont une source principale de données organisationnelles. La désignation de données principales peut être utilisée pour indiquer une source unique de vérité pour certains types de données, comme les codes de produit ou les listes de clients.

Consultez la capture d'écran suivante de la section Approbation et découverte dans les paramètres du modèle sémantique. Pour Certifié et Données principales, il existe des liens pour vous aider à obtenir vos données certifiées ou approuvées en tant que données principales. Ces liens sont configurés par un administrateur pour votre organisation. Notez également l'option permettant de rendre l'élément détectable ou non.

Power BI est également une expérience dans Microsoft Fabric et l'approbation peut être appliquée à tous les éléments Power BI et Fabric, à l'exception des tableaux de bord Power BI. Pour plus d'informations, consultez la documentation de Vue d'ensemble de l'approbation.

Explorer la vue Traçabilité

La vue Traçabilité dans un espace de travail Power BI vous permet d'afficher les dépendances entre les différents éléments de l'espace de travail. Par exemple, vous pourriez voir les sources de données sous-jacentes pour un modèle sémantique et le rapport final connectés par des lignes.

Dans la capture d'écran suivante, nous explorons un espace de travail unique dans la vue de traçabilité. De gauche à droite, vous voyez différentes sources de données, le modèle sémantique et enfin le rapport.

Les sources de données incluent un fichier HTML, deux fichiers Text/CSV différents et une instance SQL Server.

Le modèle sémantique Analyse des ventes comprend toutes les sources de données.

Le rapport Analyse des ventes est généré à partir du modèle sémantique Analyse des ventes.

Notez que le rapport Analyse des ventes a un badge avec une coche dans celui-ci, ce qui indique que le rapport est promu.

La vue Traçabilité vous permet d'identifier rapidement les dépendances entre un espace de travail pour vos ressources. Cette vue facilite votre administration, en particulier lorsque vous avez de nombreux éléments dans un seul espace de travail. L'examen de la vue de traçabilité est également parfois appelé analyse d'impact et est une forme de gouvernance des données.

Unité 4: Gérer un modèle sémantique Power BI à l'aide du point de terminaison XMLA

Type: Contenu

Gérer un modèle sémantique Power BI à l'aide du point de terminaison XMLA

Si vous avez déjà utilisé Microsoft Analysis Services, vous savez peut-être ce qu'est un point de terminaison XMLA. Bien que l'utilisation d'une API puisse être intimidante pour certains, elle est simple à utiliser. En bref, le point de terminaison XMLA est un autre moyen de vous connecter à des espaces de travail et des modèles sémantiques Power BI à partir d'applications externes. Il s'agit simplement d'une API avec une URL pointant vers un espace de travail ou un modèle sémantique.

Paramètres de point de terminaison XMLA

Par défaut, la connectivité de point de terminaison est en lecture seule pour la charge de travail des modèles sémantiques. Ce paramètre permet aux outils de visualisation des données d'accéder aux détails suivants du modèle sémantique :

Modéliser des données

Les outils de visualisation des données sont Microsoft Excel, Power BI Report Builder, Tabular Editor ou ALM Toolkit. Vous pouvez accéder à l'URL de connexion de l'espace de travail dans les paramètres de celui-ci.

La connectivité en lecture/écriture peut être activée pour fournir plus d'opérations, à savoir :

Modélisation sémantique avancée

Avec l'activation en lecture-écriture, les modèles sémantiques ont plus de parité avec les outils et processus de modélisation tabulaire de niveau entreprise d'Azure Analysis Services et SQL Server Analysis Services. Pour utiliser le point de terminaison XMLA pour les opérations en lecture-écriture, le modèle sémantique doit résider dans un espace de travail Premium ou Fabric.

Voici quelques-unes des utilisations courantes du point de terminaison XMLA dans Power BI :

Actualisation de composants individuels d'un modèle de données.

Exportation systématique de données à partir du modèle de données.

Automatisation de l'utilisation de Best Practices Analyzer.

Pour voir l'ensemble des fonctionnalités et limitations, consultez l'article Connectivité du modèle sémantique avec le point de terminaison XMLA.

Unité 5: Exercice : Créer des ressources Power BI réutilisables

Type: Exercice

Exercice : Créer des ressources Power BI réutilisables

Dans cet exercice, vous allez créer des ressources réutilisables pour prendre en charge le développement de modèles sémantiques et de rapports. Ces ressources incluent les fichiers de projets et de modèles Power BI et les modèles sémantiques partagés. À la fin, vous explorerez la vue de traçabilité pour observer la façon dont ces éléments se rapportent les uns aux autres dans le service Power BI.

Cet exercice peut être avec une licence Power BI ou Fabric.

Lancez l'exercice et suivez les instructions.

Unité 6: Évaluation du module

Type: Évaluation

Évaluation du module

Quelle fonctionnalité doit être activée pour gérer des modèles sémantiques Power BI à partir d'outils externes ?

API Service Power BI

Point de terminaison XMLA en lecture-écriture.

Point de terminaison XMLA en lecture seule.

Quel type de fichier Power BI permet de suivre et de gérer les changements avec VS Code et Git ?

Fichier .pbip du projet Power BI.

Fichier de modèle Power BI .pbit.

Fichier .pbids de source de données Power BI.

L'analyste Données d'une entreprise est chargé de créer un rapport à partir d'un modèle sémantique publié. Que doit faire l'analyste pour accomplir cette tâche ?

L'analyste doit copier le modèle sémantique existant et y apporter des modifications.

L'analyste doit se connecter au modèle sémantique Power BI existant à partir de Power BI Desktop.

L'analyste doit recréer le modèle sémantique à partir de zéro pour le nouveau rapport.

Un développeur de rapport dans une organisation tente de se connecter à des modèles sémantiques organisés, mais ne sait pas lesquels utiliser. Il trouve un modèle avec un badge « Promu ». Que cela implique-t-il ?

Le modèle a été vérifié et certifié par un réviseur autorisé par l'organisation comme étant conforme aux standards de qualité de l'organisation.

Les créateurs du modèle pensent qu'il est prêt pour être partagé et réutilisé dans l'organisation.

Les données du modèle sont une source principale de données organisationnelles et sont considérées comme la source unique de vérité faisant autorité pour certains types de données organisationnelles ou métier.

Vous devez répondre à toutes les questions avant de vérifier votre travail.

Unité 7: Résumé

Type: Résumé

Dans ce module, vous avez appris la valeur et l'importance de la création de ressources réutilisables qui réduisent le temps et la complexité lors du développement de rapports Power BI.

Nous avons abordé plusieurs fonctionnalités conçues pour une réutilisation, notamment les fichiers de modèles Power BI, et nous avons vu comment créer des modèles sémantiques principaux, puis des modèles sémantiques spécialisés pour d'autres besoins. Nous avons étudié comment utiliser des fichiers de projet Power BI, des pipelines de déploiement, l'intégration Git et le point de terminaison XMLA dans le cadre de la gestion du cycle de vie, même lors de l'utilisation d'outils externes. Finalement, nous avons démontré comment utiliser l'affichage de traçabilité dans le service Power BI pour comprendre les relations entre des sources de données, des modèles sémantiques, des rapports et des tableaux de bord.

Sans ces outils Power BI, les développeurs de rapports doivent créer des rapports et des modèles sémantiques à usage unique, et n'ont pratiquement aucune possibilité de suivre ou de gérer les modifications apportées aux modèles sémantiques. Microsoft Fabric permet aux développeurs d'utiliser ces outils en combinaison avec le contrôle de version de l'espace de travail et les pipelines de déploiement pour implémenter des pratiques d'intégration continue/livraison continue (CI/CD).

Module 5: Appliquer la sécurité du modèle Power BI

Unité 1: Présentation

Type: Introduction

Vous pouvez appliquer la sécurité du modèle en limitant l'accès à un sous-ensemble de données et en limitant l'accès à des tables et colonnes de modèle spécifiques. Vous pouvez restreindre l'accès, car certains consommateurs de rapports ne sont pas autorisés à afficher des données spécifiques, telles que les résultats des ventes d'autres régions de vente. L'obtention de cette exigence implique généralement la configuration de la sécurité au niveau des lignes (RLS), qui implique la définition de rôles et de règles dans ce modèle de données de filtre. Vous pouvez également configurer la sécurité au niveau de l'objet (OLS) pour restreindre l'accès à des tables ou colonnes entières.

Objectifs d'apprentissage

À la fin de ce module, vous pourrez :

Restreindre l'accès aux données de modèle Power BI avec RLS.

Restreindre l'accès aux objets de modèle Power BI avec OLS.

Appliquez de bonnes pratiques de développement pour appliquer la sécurité du modèle Power BI.

Unité 2: Restreindre l'accès aux données de modèle Power BI

Type: Contenu

Restreindre l'accès aux données de modèle Power BI

En tant que modélisateur de données, vous configurez RLS en créant un ou plusieurs rôles. Un rôle a un nom unique dans le modèle et inclut généralement une ou plusieurs règles. Les règles appliquent des filtres sur des tables de modèles à l'aide d'expressions de filtre DAX (Data Analysis Expressions).

Par défaut, un modèle de données n'a aucun rôle. Un modèle de données sans rôles signifie que les utilisateurs (autorisés à interroger le modèle de données) ont accès à toutes les données du modèle.

Conseil / Astuce

Il est possible de définir un rôle qui n'inclut aucune règle. Dans ce cas, le rôle fournit l'accès à toutes les lignes de toutes les tables de modèle. Ce rôle est adapté à un utilisateur administrateur autorisé à afficher toutes les données.

Vous pouvez créer, valider et gérer des rôles dans Power BI Desktop. Pour les modèles Azure Analysis Services ou SQL Server Analysis Services, vous pouvez créer, valider et gérer des rôles à l'aide de SQL Server Data Tools (SSDT).

Vous pouvez également créer et gérer des rôles à l'aide de SQL Server Management Studio (SSMS) ou à l'aide d'un outil tiers, tel que l'éditeur tabulaire.

Pour mieux comprendre comment RLS limite l'accès aux données, regardez l'image animée suivante.

Appliquer les principes de conception des schémas en étoile

Nous vous recommandons d'appliquer des principes de conception de schéma en étoile pour produire un modèle comprenant des tables de dimension et une table de faits. Il est courant de configurer Power

BI pour appliquer des règles qui filtrent les tables de dimension, ce qui permet aux relations de modèle de propager efficacement ces filtres aux tables de faits.

L'image suivante est le diagramme de modèle du modèle de données d'analyse des ventes Adventure Works. Il montre une conception de schéma en étoile comprenant une table de faits unique nommée Sales. Les autres tables sont des tables de dimension qui prennent en charge l'analyse des mesures de vente par date, territoire de vente, client, revendeur, commande, produit et vendeur. Notez que les relations de modèle connectent toutes les tables. Ces relations propagent des filtres (directement ou indirectement) à la table Sales .

Cette conception de modèle prend en charge les exemples présentés dans cette unité.

Définir des règles

Les expressions des règles sont évaluées dans le contexte de ligne. Le contexte de ligne signifie que l'expression est évaluée pour chaque ligne à l'aide des valeurs de colonne de cette ligne. Lorsque l'expression retourne TRUE, l'utilisateur peut « voir » la ligne.

Pour en savoir plus sur le contexte de ligne, parcourez le module des tables et des colonnes calculées à des modèles Power BI Desktop. Bien que ce module porte essentiellement sur l'ajout de calculs de modèle, il comporte une unité qui présente et décrit le contexte de ligne.

Vous pouvez définir des règles statiques ou dynamiques.

Règles statiques

Les règles statiques utilisent des expressions DAX qui font référence à des constantes.

Considérez la règle suivante appliquée à la table Region qui limite l'accès aux données aux ventes du Midwest.

Les étapes suivantes expliquent comment Power BI applique la règle. Elle effectue les actions suivantes :

Filtre la table Region , ce qui entraîne une ligne visible (pour le Midwest).

Utilise la relation de modèle pour propager le filtre de table Region à la table State , ce qui entraîne 14 lignes visibles (la région du Midwest comprend 14 états).

Utilise la relation de modèle pour propager le filtre de table State à la table Sales , ce qui entraîne des milliers de lignes visibles (les faits des ventes pour les états appartenant à la région du Midwest).

La règle statique la plus simple que vous pouvez créer limite l'accès à toutes les lignes de table. Considérez la règle suivante appliquée à la table Sales Details (non représentée dans le diagramme de modèle).

Cette règle garantit que les utilisateurs ne peuvent pas accéder aux données de table. Il peut être utile lorsque les vendeurs sont autorisés à voir les résultats agrégés des ventes (à partir de la table Sales), mais pas les détails au niveau des commandes.

La définition de règles statiques est simple et efficace. Envisagez de les utiliser lorsque vous devez créer seulement quelques-uns d'entre eux, comme cela peut être le cas dans Adventure Works où il n'y a que six régions américaines. Toutefois, tenez compte des inconvénients : la configuration de règles statiques peut impliquer des efforts importants pour créer et configurer. Il vous faudrait également mettre à jour et republier le jeu de données lorsque de nouvelles régions sont intégrées.

S'il existe de nombreuses règles à configurer et que vous prévoyez d' de nouvelles règles à l'avenir, envisagez plutôt de créer des règles dynamiques.

Règles dynamiques

Les règles dynamiques utilisent des fonctions DAX spécifiques qui retournent des valeurs environnementales (par opposition aux constantes). Les valeurs environnementales sont retournées à partir de trois fonctions DAX spécifiques :

USERNAME ou USERPRINCIPALNAME : renvoie l'utilisateur authentifié Power BI en tant que valeur de texte.

CUSTOMDATA : renvoie la propriété CustomData transmise dans la chaîne de connexion. Les outils de création de rapports non Power BI qui se connectent au jeu de données à l'aide d'une chaîne de connexion peuvent définir cette propriété, comme Microsoft Excel.

N'oubliez pas que la fonction USERNAME retourne l'utilisateur au format DOMAIN\username lorsqu'elle est utilisée dans Power BI Desktop. Toutefois, lorsqu'il est utilisé dans le service Power BI, il retourne le format du nom d'utilisateur principal de l'utilisateur (UPN), comme username@adventureworks.com. Vous pouvez également utiliser la fonction USERPRINCIPALNAME, qui retourne toujours l'utilisateur au format de nom d'utilisateur principal.

Envisagez une conception de modèle révisée qui inclut désormais la table AppUser (masquée). Chaque ligne de la table AppUser décrit un nom d'utilisateur et une région. Une relation de modèle à la table Region propage les filtres de la table AppUser .

La règle suivante appliquée à la table AppUser limite l'accès aux données aux régions de l'utilisateur authentifié.

La définition de règles dynamiques est simple et efficace lorsqu'une table de modèles stocke les valeurs de nom d'utilisateur. Ils vous permettent d'appliquer une conception RLS pilotée par les données. Par exemple, lorsque des vendeurs sont ajoutés ou supprimés de la table AppUser (ou sont affectés à différentes régions), cette approche de conception fonctionne simplement.

Valider les rôles

Lorsque vous créez des rôles, il est important de les tester pour s'assurer qu'ils appliquent les filtres appropriés. Pour les modèles de données créés dans Power BI Desktop, il existe la vue en tant que fonction qui vous permet de voir le rapport lorsque différents rôles sont appliqués et que différentes valeurs de nom d'utilisateur sont passées.

Configurer des mappages de rôles

Les mappages de rôles doivent être configurés avant que les utilisateurs accèdent au contenu Power BI. Le mappage de rôles implique l'affectation d'objets de sécurité Microsoft Entra aux rôles. Les objets de sécurité peuvent être des comptes d'utilisateur ou des groupes de sécurité.

Si possible, il est bon d'associer des rôles à des groupes de sécurité. De cette façon, il y aura moins de mappages, et vous pouvez déléguer la gestion des appartenances au groupe aux administrateurs réseau.

Pour les modèles développés par Power BI Desktop, le mappage de rôles est généralement dans le service Power BI. Pour les modèles Azure Analysis Services ou SQL Server Analysis Services, le mappage de rôles est généralement dans SSMS.

Pour plus d'informations sur la configuration de RLS, consultez :

Sécurité au niveau des lignes (RLS) avec Power BI

Conseils de sécurité au niveau des lignes (RLS) dans Power BI Desktop

Utiliser l'authentification unique (SSO) pour les sources DirectQuery

Lorsque votre modèle de données a des tables DirectQuery et que leur source de données prend en charge l'authentification unique, la source de données peut appliquer des autorisations de données.

Ainsi, la base de données applique RLS, et les jeux de données et rapports Power BI respectent la sécurité de la source de données.

Sachez qu'Adventure Works dispose d'une base de données Azure SQL pour ses opérations de vente qui résident dans le même espace que Power BI. La base de données applique RLS pour contrôler l'accès aux lignes de différentes tables de base de données. Vous pouvez créer un modèle DirectQuery qui se connecte à cette base de données sans rôles et le publier sur le service Power BI. Lorsque vous définissez les informations d'identification de la source de données dans le service Power BI, vous activez le SSO. Lorsque les consommateurs de rapports ouvrent des rapports Power BI, Power BI transmet leur identité à la source de données. La source de données applique ensuite RLS en fonction de l'identité du consommateur de rapports.

Pour plus d'informations sur la sécurité au niveau des lignes d'Azure SQL Database, consultez sécurité au niveau des lignes.

Les tables calculées et les colonnes calculées qui font référence à une table DirectQuery provenant d'une source de données avec authentification SSO ne sont pas prises en charge dans le service Power BI.

Pour plus d'informations sur les sources de données qui prennent en charge l'authentification unique, consultez l'authentification unique (SSO) pour les sources DirectQuery.

Unité 3: Restreindre l'accès aux objets de modèle Power BI

Type: Contenu

Restreindre l'accès aux objets de modèle Power BI

En tant que modeleur de données, vous pouvez envisager de restreindre l'accès utilisateur aux objets de modèle Power BI. Une sécurité au niveau de l'objet (OLS) peut restreindre l'accès à des tables et colonnes spécifiques, ainsi qu'à leurs métadonnées. En règle générale, vous appliquez OLS pour sécuriser des objets qui stockent des données sensibles, comme les données personnelles des employés.

Quand Power BI applique OLS, non seulement il restreint l'accès aux tables et aux colonnes, mais il peut également sécuriser des métadonnées. Lorsque vous sécurisez des métadonnées, il n'est pas possible de récupérer d'informations sur les tables et colonnes sécurisées à l'aide de Vues de gestion dynamique (DMV).

Des modèles tabulaires peuvent masquer des tables et des colonnes (et d'autres objets) à l'aide d'une perspective. Une perspective définit des sous-ensembles visibles d'objets de modèle pour vous aider à fournir un focus spécifique pour des auteurs de rapports. Les perspectives sont destinées à réduire la complexité d'un modèle, ce qui aide les auteurs de rapports à trouver des ressources intéressantes. Toutefois, les perspectives ne sont pas une fonctionnalité de sécurité, car elles ne sécurisent pas les objets. Un utilisateur peut toujours interroger une table ou une colonne même si elles ne pas visibles de lui.

Considérons un exemple chez Adventure Works. Cette organisation dispose d'une table de dimensions d'entrepôt de données nommée DimEmployee. La table inclut des colonnes qui stockent le nom, le téléphone, l'adresse e-mail et le salaire de l'employé. Bien que les consommateurs de rapports généraux puissent voir le nom et les coordonnées des employés, ils ne doivent pas être en mesure de voir les valeurs de salaires. Seul le personnel senior des ressources humaines est autorisé à voir les valeurs de salaires. Par conséquent, le modeleur de données a utilisé OLS pour accorder l'accès à la colonne des salaires uniquement à du personnel spécifique des ressources humaines.

OLS est une fonctionnalité héritée d'Azure Analysis Services (AAS) et de SQL Server Analysis Services (SSAS). La fonctionnalité est disponible dans Power BI Premium pour assurer la compatibilité descendante des modèles migrés vers Power BI. Pour cette raison, il n'est pas possible de configurer complètement OLS dans Power BI Desktop.

Pour configurer OLS, vous commencez par créer des rôles. Vous pouvez créer des rôles dans Power BI Desktop de la même façon que lors de la configuration de RLS. Ensuite, vous devez des règles OLS aux rôles. Cette fonctionnalité n'étant pas prise en charge par Power BI Desktop, vous devez adopter une approche différente.

Vous ajoutez des règles OLS à un modèle Power BI Desktop à l'aide d'un point de terminaison XML for Analysis (XMLA). Les points de terminaison XMLA sont disponibles avec Power BI Premium, et fournissent l'accès au moteur Analysis Services dans le service Power BI. Le point de terminaison en lecture/écriture prend en charge la gestion du jeu de données, la gestion du cycle de vie des applications, la modélisation avancée des données, etc. Vous pouvez utiliser des API avec point de terminaison XMLA pour les scripts, tels que Tabular Model Scripting Language (TMSL) ou le module PowerShell SqlServer. Vous pouvez également utiliser un outil client, comme SSMS. Il existe également des options d'outils tiers, comme l'éditeur tabulaire qui est un outil open source pour créer, maintenir et gérer des modèles.

Par défaut, toutes les tables et colonnes de modèle ne sont pas limitées. Vous pouvez les définir sur None ou Read. Quand elles sont définies sur None, les utilisateurs associés au rôle ne peuvent pas accéder à l'objet. Quand elles sont définies sur Read, les utilisateurs associés au rôle peuvent accéder à l'objet. Lorsque vous limitez des colonnes spécifiques, vérifiez que la table n'est pas définie sur None.

Une fois que vous avez ajouté les règles OLS, vous pouvez publier le modèle sur le service Power BI. Utilisez le même processus pour RLS afin de mapper des comptes et des groupes de sécurité aux rôles.

Dans un rapport Power BI, quand un utilisateur n'a pas l'autorisation d'accéder à une table ou une colonne, il reçoit un message d'erreur. Le message les informera que l'objet n'existe pas.

Considérez soigneusement si OLS est la bonne solution pour votre projet. Quand un utilisateur ouvre un rapport Power BI qui interroge un objet restreint (pour eux), le message d'erreur pourrait être déroutant et entraînera une expérience négative. Le rapport lui semble corrompu. Une meilleure approche pourrait être de créer un ensemble distinct de modèles ou de rapports pour les différentes exigences du consommateur de rapports.

Il existe quelques restrictions à connaître lors de l'implémentation d'OLS.

Vous ne pouvez pas combiner RLS et OLS dans le même rôle. Si vous devez appliquer RLS et OLS dans le même modèle, créez des rôles distincts dédiés à chaque type. Vous ne pouvez pas non plus définir de sécurité au niveau de la table si elle interrompt une chaîne de relation. Par exemple, s'il existe des relations entre les tables A et B, et B et C, vous ne pouvez pas sécuriser la table B. Si la table B est sécurisée, une requête sur la table A ne peut pas transiter les relations entre les tables A et B, et B et C. Dans ce cas, vous pourriez configurer une relation distincte entre les tables A et C.

Cependant, les relations de modèle qui font référence à une colonne sécurisée fonctionneront, à condition que la table de la colonne ne soit pas sécurisée.

Enfin, s'il n'est pas possible de sécuriser des mesures, une mesure qui référence des objets sécurisés est automatiquement restreinte.

Pour plus d'informations, consultez Sécurité au niveau de l'objet.

Unité 4: Appliquer de bonnes pratiques de modélisation

Type: Contenu

Appliquer de bonnes pratiques de modélisation

Il est essentiel que votre modèle applique des autorisations de données correctement et efficacement. La liste suivante vous fournit des bonnes pratiques de développement à appliquer.

Essayez de définir moins de jeux de données (modèles de données) avec des rôles bien conçus.

Essayez de créer moins de rôles à l'aide de règles dynamiques. Une solution pilotée par les données est plus facile à gérer, car vous n'avez pas besoin d' de nouveaux rôles.

Si possible, créez des règles qui filtrent les tables de dimension au lieu de tables de faits. Il aidera à fournir des performances de requête plus rapides.

Vérifiez que la conception du modèle, y compris ses relations et ses propriétés de relation, est correctement configurée.

Utilisez la fonction USERPRINCIPALNAME au lieu de la fonction USERNAME. Il fournit une cohérence lors de la validation des rôles dans Power BI Desktop et le service Power BI.

Validez rigoureusement RLS et OLS en testant tous les rôles.

Vérifiez que la connexion de source de données Power BI Desktop utilise les mêmes informations d'identification que celles qui seront appliquées lors de la configuration dans le service Power BI.

Unité 5: Évaluation du module

Type: Évaluation

Évaluation du module

Choisissez la meilleure réponse à chacune des questions ci-dessous.

Vérifiez vos connaissances

Joshua est modélisateur de données chez Adventure Works. Il développe un modèle pour un entrepôt de données volumineux. Le modèle doit appliquer RLS, et les rapports Power BI qui se connectent au modèle doivent offrir les meilleures performances possibles. Qu'est-ce que Joshua devrait faire ?

Appliquer des règles aux tables de dimension.

Appliquez des règles aux hiérarchies.

Appliquer des règles aux tables de faits.

Rupali est modélisatrice de données chez Adventure Works. Elle développe un modèle d'importation pour analyser les données de feuille de temps des employés. La table des employés stocke les numéros de sécurité sociale (SSN) des employés dans une colonne. Si le modèle sera disponible pour tous les chefs d'entreprise, il le sera également pour les employés du service Paie. Toutefois, les rapports ne doivent révéler les numéros de sécurité sociale des employés qu'aux employés du service de la paie. Quelle fonctionnalité Rupali devrait-elle utiliser pour restreindre l'accès à la colonne SSN ?

SSO (Authentification unique)

Kasper est modélisateur de données chez Adventure Works. Il développe qui doit appliquer RLS. Il doit restreindre l'accès aux seules régions de ventes affectées au consommateur du rapport. La base de données source inclut une table qui stocke les noms d'utilisateur des employés et leurs régions affectées. Qu'est-ce que Kasper devrait faire ?

Créer un rôle OLS et utiliser une règle dynamique.

Créer un rôle RLS et utiliser une règle dynamique.

Créer un rôle RLS et utiliser une règle statique.

Vous devez répondre à toutes les questions avant de vérifier votre travail.

Unité 6: Résumé

Type: Résumé

Lorsque les consommateurs de rapports peuvent voir toutes les données de modèle, il n'est pas nécessaire de prendre des mesures spéciales. Toutefois, lorsqu'ils doivent voir des sous-ensembles de données de modèle ou être limités à partir de certaines tables ou colonnes, il est essentiel que votre modèle applique une sécurité appropriée.

En règle générale, vous limitez l'accès aux données de modèle avec des règles RLS. Toutefois, lorsque vous développez un modèle DirectQuery pour les sources de données qui prennent en charge l'authentification unique, Power BI peut tirer parti de la source de données RLS. Dans ce cas, en tant que modélisateur de données, vous n'avez pas besoin de créer de rôles de modèle.

Il existe de nombreuses bonnes pratiques de développement que vous devez appliquer pour vous assurer que les autorisations de données sont appliquées avec précision et efficacement.

Parcours 4: Administration et gouvernance de Microsoft Fabric

Module 1: Administrer un environnement Microsoft Fabric

Unité 1: Présentation

Type: Introduction

L'administration d'un environnement Microsoft Fabric implique des tâches essentielles pour garantir l'utilisation efficace et efficace de la plateforme Fabric au sein d'une organisation.

En tant qu'administrateur (admin) Fabric, vous devez connaître :

Architecture de Fabric

Fonctionnalités de sécurité et de gouvernance

Fonctionnalités d'analytique

Options de déploiement et de licence

Vous devez également connaître le portail d'administration Fabric et d'autres outils d'administration, et être en mesure de configurer et de gérer l'environnement Fabric pour répondre aux besoins de votre organisation.

Les administrateurs Fabric travaillent avec les utilisateurs professionnels, les analystes Données et les professionnels informatiques pour déployer et utiliser Fabric en vue de répondre aux objectifs métier et de respecter les stratégies et normes organisationnelles.

À la fin de ce module, vous aurez une bonne compréhension du rôle d'administrateur Fabric et des tâches et outils impliqués dans l'administration de Fabric.

Unité 2: Comprendre l'architecture Fabric

Type: Contenu

Comprendre l'architecture Fabric

Microsoft Fabric est une plateforme Software-as-a-Service qui offre une approche simple et intégrée tout en réduisant la charge administrative. Fabric fournit une solution d'analytique tout-en-un pour les entreprises qui couvre tout, du déplacement des données à la science des données, à l'analytique en temps réel et au décisionnel. Il offre une suite complète de services, notamment les suivants :

Entrepôt de données

Engineering données

Intégration des données

Science des données

Informations en temps réel

Toutes les données de Fabric sont stockées dans OneLake, qui repose sur l'architecture Azure Data Lake Storage (ADLS) gen2. OneLake est hiérarchique par nature pour simplifier la gestion dans toute votre organisation. Il n'y a qu'un seul OneLake par locataire et il fournit un espace de noms de système de fichiers qui s'étend aux utilisateurs, aux régions et même aux clouds dans une seule et même vue.

Comprendre les concepts de Fabric

Un locataire est un espace dédié permettant aux organisations de créer, de stocker et de gérer des éléments Fabric. Il y a souvent une seule instance de Fabric pour une organisation, qui elle est alignée sur Microsoft Entra ID. Le locataire Fabric est mappé à la racine de OneLake et se trouve au niveau supérieur de la hiérarchie.

La capacité est un ensemble dédié de ressources qui sont disponibles à un moment donné pour être utilisées. Une ou plusieurs capacités peuvent être associées à un locataire. La capacité définit la possibilité d'une ressource à effectuer une activité ou à produire une sortie. Les besoins en capacité varient selon l'élément et la durée d'utilisation. Fabric offre une capacité via la référence SKU et les évaluations Fabric.

Un domaine est un regroupement logique d'espaces de travail. Les domaines sont utilisés pour organiser des éléments d'une manière logique pour votre organisation. Vous pouvez regrouper des éléments de manière à faciliter l'accès des groupes de personnes aux espaces de travail. Par exemple, vous pouvez avoir un domaine pour les ventes, un autre pour le marketing et un autre pour les finances.

Un espace de travail est une collection d'éléments qui regroupe différentes fonctionnalités dans un seul locataire. Il agit comme un conteneur qui utilise la capacité disponible pour le travail exécuté, et fournit des contrôles pour les personnes qui peuvent accéder aux éléments qu'il contient. Par exemple, dans un espace de travail des ventes, les utilisateurs associés à l'organisation des ventes peuvent créer un entrepôt de données, exécuter des notebooks, créer des jeux de données, créer des rapports et plus encore.

Les éléments sont les composantes de la plateforme Fabric. Il s'agit des objets que vous créez et gérez dans Fabric. Il existe différents types d'éléments, tels que les entrepôts de données, les pipelines de données, les jeux de données, les rapports et les tableaux de bord.

En tant qu'administrateur, il est important de comprendre les concepts Fabric, car cela vous permet de comprendre comment gérer l'environnement Fabric.

Pour plus d'informations, consultez la documentation Démarrer un essai Fabric.

Unité 3: Comprendre le rôle Administrateur Fabric

Type: Contenu

Comprendre le rôle Administrateur Fabric

Maintenant que vous comprenez l'architecture fabric et ce que vous et votre équipe pouvez utiliser Fabric, examinons le rôle d'administrateur et les outils utilisés pour gérer la plateforme.

Il y a plusieurs rôles qui coopèrent pour administrer Microsoft Fabric pour votre organisation. Si vous êtes administrateur Microsoft 365, administrateur Power Platform ou administrateur de capacité Fabric, vous êtes impliqué dans l'administration de Fabric. Le rôle d'administrateur Fabric était anciennement

l’administrateur Power BI.

En tant qu’administrateur Fabric, vous travaillez principalement dans le portail d’administration Fabric. Vous devrez peut-être également vous familiariser avec les outils suivants :

Centre d’administration Microsoft 365

Sécurité Microsoft 365 & Portail de conformité Microsoft Purview

Ouvrez Microsoft Entra ID dans le portail Azure

Cmdlets PowerShell

API et SDK d’administration

Pour obtenir des détails spécifiques sur les différents rôles d’administrateur et leurs responsabilités, consultez la documentation Qu’est-ce que l’administration Microsoft Fabric ?.

Décrire les tâches d’administration

En tant qu’administrateur, vous pourriez être responsable d’un large éventail de tâches pour garantir la bonne exécution de la plateforme Fabric. Il s’agit notamment des tâches suivantes :

Sécurité et contrôle d’accès : l’un des aspects les plus importants de l’administration de Fabric est la gestion de la sécurité et du contrôle d’accès pour garantir que seuls les utilisateurs autorisés peuvent accéder aux données sensibles. Vous pouvez utiliser le contrôle d’accès en fonction du rôle (RBAC) pour :

Définir qui peut afficher et modifier du contenu.

Configurer des passerelles de données pour vous connecter en toute sécurité à des sources de données locales.

Gérer les accès des utilisateurs avec Microsoft Entra ID.

Gouvernance des données : l’administration efficace de Fabric nécessite une bonne compréhension des principes de gouvernance des données. Vous devez savoir comment sécuriser la connectivité entrante et sortante dans votre locataire et comment superviser les métriques d’utilisation et de performances. Vous devez également savoir comment appliquer des stratégies de gouvernance des données pour garantir que les données au sein de votre locataire ne sont accessibles qu’aux utilisateurs autorisés.

Personnalisation et configuration : l’administration de Fabric implique également la personnalisation et la configuration de la plateforme pour répondre aux besoins de votre organisation. Vous pouvez configurer des liens privés pour sécuriser votre locataire, définir des stratégies de classification des données ou ajuster l’apparence des rapports et des tableaux de bord.

Supervision et optimisation : en tant qu’administrateur Fabric, vous devez savoir comment superviser les performances et l’utilisation de la plateforme, optimiser les ressources et résoudre les problèmes. Les exemples incluent la configuration des paramètres de supervision et d’alerte, l’optimisation des performances des requêtes, la gestion de la capacité et de la mise à l’échelle, et la résolution des problèmes d’actualisation des données et de connectivité.

Des tâches spécifiques varient en fonction des besoins de votre organisation et de la complexité de votre implémentation Fabric.

Décrire les outils d’administration

Il est important de vous familiariser avec quelques outils pour implémenter efficacement les tâches décrites précédemment. Les administrateurs Fabric peuvent effectuer la plupart des tâches

d’administration à l’aide d’un ou de plusieurs des outils suivants : le portail d’administration Fabric, les applets de commande PowerShell, les SDK et API d’administration, et l’espace de travail de supervision de l’administration.

Portail d’administration Fabric

Le portail d’administration de Fabric est un portail web où vous pouvez gérer tous les aspects de la plateforme. Vous pouvez gérer, examiner et appliquer les paramètres de manière centralisée pour l’ensemble du locataire ou par capacité dans le portail d’administration. Vous pouvez également gérer les utilisateurs, les administrateurs et les groupes, accéder aux journaux d’audit, et superviser l’utilisation et les performances.

Le portail d’administration vous permet d’activer et de désactiver les paramètres. Il existe de nombreux paramètres situés dans le portail d’administration. Un paramètre remarquable est le commutateur Fabric, situé dans les paramètres du locataire, qui permet aux organisations utilisant Power BI de choisir d'activer Fabric. Ici, vous pouvez activer Fabric pour votre locataire ou autoriser les administrateurs de capacité à activer Fabric.

Applets de commande PowerShell

Fabric fournit un ensemble d’applets de commande PowerShell que vous pouvez utiliser pour automatiser les tâches d’administration courantes. Une applet de commande PowerShell est une commande simple qui peut être exécutée dans PowerShell.

Par exemple, vous pouvez utiliser des applets de commande dans Fabric pour créer et gérer des groupes, configurer des sources de données et des passerelles, et superviser l’utilisation et les performances de manière systématique. Vous pouvez également utiliser les applets de commande pour gérer les SDK et API d’administration Fabric.

Pour obtenir plus de ressources sur les applets de commande PowerShell qui fonctionnent avec Fabric, consultez Applets de commande Microsoft Power BI pour Windows PowerShell et PowerShell Core.

SDK et API d’administration

Un SDK et une API d’administration sont des outils qui permettent aux développeurs d’interagir avec un système logiciel par programmation. Une interface de programmation d’applications (API, Application Programming Interface) est un ensemble de protocoles et d’outils qui permettent la communication entre différentes applications logicielles. Un Kit de développement logiciel (SDK, Software Development Kit) est un ensemble d’outils et de bibliothèques qui permettent aux développeurs de créer des applications logicielles pouvant interagir avec un système ou une plateforme spécifique. Vous pouvez utiliser des API et des SDK pour automatiser des tâches d’administration courantes et intégrer Fabric à d’autres systèmes.

Par exemple, vous pouvez utiliser des API et des sdk pour créer et gérer des groupes, configurer des sources de données et des passerelles, et surveiller l’utilisation et les performances. Vous pouvez également utiliser les API et les SDK pour gérer les SDK et API d’administration Fabric.

Vous pouvez effectuer ces demandes à l’aide de n’importe quelle bibliothèque de client HTTP qui prend en charge l’authentification OAuth 2.0, comme Postman, ou vous pouvez utiliser des scripts PowerShell pour automatiser le processus.

Espace de travail de supervision de l’administration

Les administrateurs de locataires Fabric ont accès à l’espace de travail de surveillance de l’administration. Vous pouvez choisir de partager l’accès à l’espace de travail ou à des éléments spécifiques qu’il contient avec d’autres utilisateurs de votre organisation. L’espace de travail de supervision de l’administration comprend le jeu de données et le rapport Adoption et Utilisation des

fonctionnalités qui, ensemble, fournissent des insights sur l'utilisation et les performances de votre environnement Fabric. Vous pouvez utiliser ces informations pour identifier les tendances et les modèles, et résoudre les problèmes.

Pour plus d'informations sur ce qui est inclus dans l'espace de travail d'analyse administrateur, consultez Qu'est-ce que l'espace de travail de supervision de l'administrateur Fabric.

Unité 4: Gérer la sécurité Fabric

Type: Contenu

Gérer la sécurité Fabric

En tant qu'administrateur d'infrastructure, une partie de votre rôle consiste à gérer la sécurité de l'environnement Fabric, notamment la gestion des utilisateurs et des groupes, ainsi que la façon dont les utilisateurs partagent et distribuent du contenu dans Fabric.

Gérer les utilisateurs : attribuer et gérer des licences

Les licences utilisateur contrôlent le niveau d'accès et de fonctionnalité utilisateur dans l'environnement Fabric. Les administrateurs garantissent que les utilisateurs avec une licence disposent de l'accès dont ils ont besoin pour effectuer leurs tâches efficacement. Ils limitent également l'accès aux données sensibles et garantissent la conformité aux lois et réglementations relatives à la protection des données.

La gestion des licences permet aux administrateurs de superviser et de contrôler les coûts en veillant à ce que les licences soient allouées efficacement et uniquement aux utilisateurs qui en ont besoin. Cela permet d'éviter des dépenses inutiles et à garantir que l'organisation utilise efficacement ses ressources.

Le fait de disposer des procédures appropriées pour attribuer et gérer des licences permet de contrôler l'accès aux données et à l'analytique, de garantir la conformité aux réglementations et d'optimiser les coûts.

La gestion des licences pour Fabric est gérée dans le Centre d'administration Microsoft 365. Pour plus d'informations sur la gestion des licences, consultez Affecter des licences aux utilisateurs.

Le type de licence dans les paramètres de l'espace de travail est lié aux licences utilisateur répertoriées ici. Les utilisateurs peuvent voir des rapports en fonction de la licence utilisateur et de la licence de l'espace de travail. Pour plus d'informations, consultez la documentation sur les licences Microsoft Fabric .

Gérer les éléments et le partage

En tant qu'administrateur, vous pouvez gérer la façon dont les utilisateurs partagent et distribuent du contenu. Vous pouvez gérer la façon dont les utilisateurs partagent du contenu avec d'autres utilisateurs et la façon dont ils distribuent du contenu à d'autres utilisateurs. Vous pouvez également gérer la façon dont les utilisateurs interagissent avec les éléments, tels que les entrepôts de données, les pipelines de données, les jeux de données, les rapports et les tableaux de bord.

Les éléments des espaces de travail sont mieux distribués via une application d'espace de travail ou directement via l'espace de travail. L'octroi des droits les moins permissifs est la première étape de la sécurisation des données. Partagez l'application en lecture seule pour l'accès aux rapports ou accordez l'accès aux espaces de travail pour la collaboration et le développement. Un autre aspect de la gestion et de la distribution des éléments consiste à appliquer ces types de bonnes pratiques.

Vous pouvez gérer le partage et la distribution à la fois en interne et en dehors de votre organisation, conformément à ses stratégies et ses procédures.

Pour plus d'informations, consultez la documentation sécurité dans Microsoft Fabric .

Unité 5: Gouverner des données dans Fabric

Type: Contenu

Gouverner des données dans Fabric

Fabric inclut des fonctionnalités de gouvernance intégrées pour vous aider à gérer et à contrôler vos données. L'approbation est un moyen vous permettant, en tant qu'administrateur, de désigner des éléments Fabric spécifiques comme fiables et approuvés pour une utilisation dans toute l'organisation.

Les administrateurs peuvent également utiliser l'API scanneur pour analyser des éléments Fabric pour rechercher des données sensibles et la fonctionnalité de traçabilité des données pour suivre le flux de données via Fabric.

Approuver le contenu Fabric

L'approbation est une fonctionnalité de gouvernance clé qui génère une confiance dans vos ressources de données en marquant les éléments Fabric comme étant examinés et approuvés. Les éléments approuvés affichent un badge qui indique aux utilisateurs que ces ressources sont fiables. L'approbation permet aux utilisateurs de faire confiance aux données et vous aide également, en tant qu'administrateur, à gérer la croissance globale des éléments dans votre environnement.

Le contenu Fabric promu s'affiche avec un badge Promu dans le portail Fabric. Les membres de l'espace de travail ayant le rôle Contributeur ou Administrateur peuvent promouvoir du contenu au sein d'un espace de travail. L'administrateur Fabric peut promouvoir du contenu au sein de l'organisation.

Le contenu certifié nécessite un processus plus formel qui implique une révision du contenu par un réviseur désigné. Le contenu certifié s'affiche avec un badge Certifié dans le portail Fabric. Les administrateurs gèrent le processus de certification et peuvent le personnaliser pour répondre aux besoins de votre organisation.

Si vous n'êtes pas administrateur, vous devez demander la certification d'élément auprès d'un administrateur. Vous pouvez effectuer une certification de demande en sélectionnant l'élément dans le portail Fabric, puis en sélectionnant Demander la certification dans le menu Plus .

Pour plus d'informations sur le processus d'approbation de contenu, consultez Promouvoir ou certifier du contenu.

Rechercher des données sensibles

L'analyse des métadonnées facilite la gouvernance des données en activant le catalogage et la création de rapports sur toutes les métadonnées des éléments Fabric de votre organisation. L'API scanneur est un ensemble d'API REST d'administration qui vous permet d'analyser les éléments Fabric pour les données sensibles. Utilisez l'API de scanneur pour analyser les entrepôts de données, les pipelines de données, les jeux de données, les rapports et les tableaux de bord afin de rechercher des données sensibles. L'API de scanneur peut être utilisée pour analyser à la fois des données structurées et des données non structurées.

Avant que l'analyse des métadonnées puisse être exécutée, elle doit être configurée dans votre organisation par un administrateur. Pour plus d'informations, consultez la vue d'ensemble de l'analyse des métadonnées.

Suivre la traçabilité des données

La traçabilité des données est la possibilité de suivre le flux de données via Fabric, également appelée analyse d'impact. La traçabilité des données vous permet de voir d'où proviennent les données, comment elles sont transformées et leur destination. La vue de traçabilité dans les espaces de travail vous permet de comprendre les données disponibles dans Fabric et la façon dont elles sont utilisées.

Rapport sur les données sensibles

Avec le hub Microsoft Purview (préversion) dans Fabric, vous pouvez gérer et gouverner le paysage de données de données Fabric de votre organisation. Il contient des rapports qui fournissent des informations sur les données sensibles, l'approbation d'élément et les domaines, et sert également de passerelle vers des fonctionnalités plus avancées dans le portail Microsoft Purview, telles que Data Catalog, Information Protection, Data Loss Prevention et Audit.

Unité 6: Évaluation du module

Type: Évaluation

Évaluation du module

Parmi les affirmations suivantes, laquelle décrit le mieux le concept de capacité dans Fabric ?

La capacité fait référence à un espace dédié permettant aux organisations de créer, de stocker et de gérer des éléments Fabric.

La capacité définit la possibilité d'une ressource à effectuer une activité ou à produire une sortie.

La capacité est une collection d'éléments qui sont regroupés de manière logique.

Parmi les affirmations suivantes, laquelle est vraie concernant la différence entre la promotion et la certification dans Fabric ?

La promotion et la certification permettent toutes les deux à tout membre de l'espace de travail d'approuver du contenu.

La promotion nécessite un niveau d'autorisations plus élevé que celui de la certification.

La certification doit être activée dans le locataire par l'administrateur, tandis que la promotion peut être e par un membre de l'espace de travail.

Vous devez répondre à toutes les questions avant de vérifier votre travail.

Unité 7: Résumé

Type: Résumé

Dans ce module, vous avez découvert l'architecture Fabric et le rôle d'un administrateur pour la plateforme Fabric. Vous avez également exploré les différents outils disponibles pour la gestion de la sécurité et du partage, ainsi que les fonctionnalités de gouvernance qui peuvent être utilisées pour appliquer des normes et garantir la conformité. Si vous comprenez bien comment gérer un environnement Fabric, vous garantissez sa sécurité, sa conformité et sa gouvernance. Fort de ces connaissances, vous êtes bien équipé pour aider votre organisation à tirer le meilleur parti de Fabric et à retirer de précieux insights de toutes vos données.

Pour plus d’informations sur la gouvernance des données, complétez les données de gouvernance dans Microsoft Fabric avec le module Purview .

Module 2: Sécuriser l’accès aux données dans Microsoft Fabric

Unité 1: Présentation

Type: Introduction

La sécurité dans Microsoft Fabric est optimisée pour sécuriser les données dans des cas d’utilisation spécifiques. Dans Fabric, les différents utilisateurs ont besoin de pouvoir effectuer diverses actions afin de remplir leurs responsabilités professionnelles. Fabric facilite cela en vous permettant d’accorder aux utilisateurs l’accès à des charges de travail de données spécifiques par le biais de permissions d’espace de travail et d’éléments, de permissions de calcul et de rôles d’accès aux données OneLake (préversion).

Sécuriser les données selon les cas d’utilisation

Un cas d’utilisation de la sécurité dans Fabric fait référence à un ensemble d’utilisateurs qui ont besoin d’accéder à des données et qui y accèdent d’une manière spécifique. Une fois qu’un cas d’utilisation a été identifié, les autorisations de Fabric associées à ce cas d’utilisation peuvent être configurées. Supposons que vous travailliez dans une entreprise de soins de santé disposant de plusieurs systèmes qui stockent des données, telles que des dossiers médicaux électroniques (EHR), des données sur les demandes d’assurance, des données sur les essais cliniques, des registres de patients et de maladies, et des données administratives. Au sein de votre entreprise ou d’organisations partenaires, différents utilisateurs doivent visualiser, transformer, analyser, agréger et utiliser ces données pour en tirer des informations métier. Les utilisateurs ont besoin d’accéder à différents moteurs de calcul, éléments et espaces de travail Fabric pour effectuer leur travail de manière efficace :

Les ingénieurs données ont besoin d’accéder aux données d’un lakehouse pour développer des produits de données en aval.

Les analystes d’entreprise ou de données doivent interroger les données pour répondre aux questions de l’entreprise.

Les scientifiques des données ont besoin d’accéder aux données dans un lakehouse et de les consommer via Apache Spark pour créer des modèles et des expériences.

Les créateurs de rapports doivent rédiger des rapports à partager avec les consommateurs de rapports.

Les consommateurs de rapports ont besoin de visualiser les données dans les rapports Power BI pour prendre des décisions.

Dans ce module, vous apprendrez à utiliser les fonctionnalités de contrôle d’accès de Fabric disponibles pour sécuriser vos données et fournir à votre équipe l’accès nécessaire au sein de Fabric pour effectuer leurs missions. Vous allez découvrir le modèle de sécurité multicouche de Fabric et comment l’utiliser pour gérer l’accès aux données.

À la fin de ce module, vous saurez comment configurer la sécurité pour l’intégralité d’un espace de travail, pour des éléments de données Fabric individuels et comment appliquer des autorisations granulaires au sein des éléments de données Fabric.

Unité 2: Comprendre le modèle de sécurité Fabric

Type: Contenu

Comprendre le modèle de sécurité Fabric

L’accès aux données dans les organisations est souvent restreint par les responsabilités et les rôles des utilisateurs, par l’architecture des données et les modèles de déploiement Fabric d’une organisation. Fabric a un modèle de sécurité flexible et multicouche qui vous permet de configurer la sécurité pour tenir compte des différentes exigences d’accès aux données. Le fait d’avoir la possibilité de contrôler les autorisations sur différentes couches signifie que vous pouvez adhérer au principe des privilèges minimum, limiter les autorisations utilisateur au strict nécessaire pour effectuer des tâches de travail.

Fabric a trois niveaux de sécurité qui sont évalués de manière séquentielle pour déterminer si un utilisateur a un accès aux données. L’ordre d’évaluation pour l’accès est le suivant :

Authentification Microsoft Entra ID : confirme que l’utilisateur peut s’authentifier au service de gestion des identités et des accès Azure, Microsoft Entra ID.

Accès à Fabric : confirme que l’utilisateur peut accéder à Fabric.

Sécurité des données : confirme que l’utilisateur peut effectuer l’action demandée sur une table ou un fichier.

Le troisième niveau, la sécurité des données, a plusieurs blocs de construction que vous pouvez configurer individuellement ou ensemble pour qu’ils s’alignent sur différentes exigences d’accès. Les contrôles d’accès primaires dans Fabric sont les suivants :

Rôles d’espace de travail

Autorisations d’élément

Autorisations granulaires ou de calcul

Contrôles d’accès aux données OneLake (préversion)

Il est utile de concevoir ces blocs de construction dans une hiérarchie pour comprendre comment les contrôles d’accès peuvent être appliqués individuellement ou ensemble.

Un espace de travail dans Fabric vous permet de distribuer des stratégies de propriété et d’accès à l’aide de rôles d’espace de travail. Dans un espace de travail, vous pouvez créer des éléments de données Fabric tels que lakehouses, entrepôts de données et modèles sémantiques. Les autorisations d’élément peuvent être héritées d’un rôle d’espace de travail ou définies individuellement en partageant un élément. Lorsque les rôles d’espace de travail fournissent trop d’accès, les éléments peuvent être partagés à l’aide d’autorisations d’élément pour garantir une sécurité appropriée.

Dans chaque élément de données, des autorisations de moteur granulaires telles que Read, ReadData ou ReadAll peuvent être appliquées.

Les autorisations granulaires ou de calcul peuvent être appliquées au sein d’un moteur de calcul spécifique dans Fabric, comme le modèle sémantique ou point de terminaison SQL.

Les éléments de données Fabric stockent leurs données dans OneLake. L’accès aux données dans le lakehouse peut être limité à des dossiers ou fichiers spécifiques en utilisant la fonctionnalité de contrôle d’accès en fonction du rôle (RBAC) appelé Contrôle d’accès aux données OneLake (préversion).

Unité 3: Configurer des autorisations d’espace de travail et d’élément

Type: Contenu

Configurer des autorisations d’espace de travail et d’élément

Les espaces de travail sont des environnements où les utilisateurs peuvent collaborer pour créer des groupes d’éléments. Les éléments sont les ressources avec lesquelles vous pouvez travailler dans Fabric, telles que les lakehouses, les entrepôts et les rapports. Les rôles d’espace de travail sont des ensembles d’autorisations préconfigurés qui vous permet de gérer ce que les utilisateurs peuvent faire et d’accéder à un espace de travail Fabric.

Les autorisations d’élément contrôlent l’accès à des éléments Fabric individuels au sein d’un espace de travail. Les autorisations d’élément vous permettent d’ajuster les autorisations définies par un rôle d’espace de travail ou d’accorder l’accès à un utilisateur à un ou plusieurs éléments dans un espace de travail sans l’utilisateur à un rôle d’espace de travail.

Examinons des scénarios où vous devez configurer l’accès aux données en utilisant des rôles d’espace de travail et des autorisations d’élément.

Comprendre les rôles d’espace de travail

Supposons que vous travaillez dans une entreprise de santé en tant qu’administrateur de la sécurité Fabric. Vous devez configurer un accès pour un nouvel ingénieur Données. L’ingénieur Données doit pouvoir effectuer ce qui suit :

Créer des éléments Fabric dans un espace de travail existant

Lire toutes les données d’un lakehouse existant qui se trouve dans le même espace de travail dans lequel il peut créer des éléments Fabric

Les rôles d’espace de travail contrôlent ce que les utilisateurs peuvent faire et accéder au sein d’un espace de travail Fabric. Il existe quatre rôles d’espace de travail qui s’appliquent à tous les éléments dans un espace de travail. Vous pouvez attribuer des rôles d’espace de travail à des individus, des groupes de sécurité, des groupes Microsoft 365 et des listes de distribution. Les utilisateurs peuvent être affectés aux rôles suivants :

Administrateur : peut afficher, modifier, partager et gérer tout le contenu et les données de l’espace de travail, y compris la gestion des autorisations.

Membre : peut afficher, modifier et partager tout le contenu et les données dans l’espace de travail.

Contributeur : peut afficher et modifier tout le contenu et les données dans l’espace de travail.

Visionneuse : peut afficher tout le contenu et les données dans l’espace de travail, mais ne peut pas les modifier.

Pour découvrir la liste complète des autorisations associées aux rôles d’espace de travail, consultez : Rôles dans les espaces de travail

Afin de répondre aux exigences d’accès pour le nouvel ingénieur Données, vous pouvez lui attribuer le rôle Contributeur. Cela lui accorde un accès pour modifier du contenu dans l’espace de travail, notamment la création d’éléments Fabric comme les lakehouses. Le rôle Contributeur lui permet également de lire des données dans le lakehouse existant.

Affecter des rôles d’espace de travail

Des utilisateurs peuvent être ajoutés à des rôles d’espace de travail à partir du bouton Gérer l’accès dans un espace de travail. Ajoutez un utilisateur en entrant son nom et en sélectionnant le rôle d’espace de travail à lui attribuer dans la boîte de dialogue des personnes.

Configurer des autorisations d’élément

Les autorisations d'élément contrôlent l'accès à des éléments Fabric individuels au sein d'un espace de travail. Les autorisations d'élément peuvent être utilisées avec des rôles d'espace de travail ou pour accorder l'accès à un utilisateur à un ou plusieurs éléments dans un espace de travail sans l'utilisateur à un rôle d'espace de travail.

Supposons que quelques mois après avoir lui voir accordé l'accès Contributeur sur un espace de travail, un ingénieur Données n'a plus besoin de créer des éléments Fabric et doit maintenant afficher uniquement un seul lakehouse et y lire ses données.

Étant donné que l'ingénieur n'a plus besoin d'afficher tous les éléments dans l'espace de travail, le rôle d'espace de travail Contributeur peut être supprimé et les autorisations d'élément sur le lakehouse peuvent être configurées pour que l'ingénieur puisse uniquement voir les données et métadonnées du lakehouse dans l'espace de travail, et rien de plus. La configuration d'accès à cet élément vous aide à adhérer au principe des privilèges minimum dans lequel l'ingénieur a uniquement accès à ce qui lui est nécessaire pour effectuer ses obligations professionnelles.

Un élément peut être partagé et des autorisations d'élément peuvent être configurées en sélectionnant les points de suspension (...) à côté d'un élément Fabric dans un espace de travail, puis Gérer les autorisations.

Dans la fenêtre Accès l'accès en lecture qui apparaît après la sélection de Gérer les autorisations, si vous ajoutez l'utilisateur sans sélectionner l'une des cases à cocher sous Autorisations supplémentaires, l'utilisateur a accès en lecture aux métadonnées du lakehouse. L'utilisateur n'a pas accès aux données sous-jacentes dans le lakehouse. Pour donner à l'ingénieur la capacité de lire des données, et pas seulement des métadonnées, l'option Lire toutes les données de point de terminaison SQL ou l'option Lire tout Apache Spark peuvent être sélectionnées.

Chaque élément de données Fabric a son propre modèle de sécurité. Pour découvrir plus d'informations sur les autorisations qui peuvent être accordées quand un lakehouse ou un autre élément de données Fabric est partagé, voir :

Modèle sémantique

Unité 4: Appliquer des autorisations granulaires

Type: Contenu

Appliquer des autorisations granulaires

Quand les autorisations fournies par les rôles d'espace de travail ou les autorisations d'élément sont insuffisantes, vous pouvez définir des autorisations granulaires telles que la sécurité au niveau des tables et la sécurité au niveau des lignes ainsi que l'accès aux fichiers et aux dossiers via :

Point de terminaison d'analytique SQL

Rôles d'accès aux données OneLake (préversion)

Modèle sémantique

Configurer l'accès aux données via le point de terminaison d'analytique SQL dans un lakehouse

Les données d'un lakehouse peuvent être lues via le point de terminaison d'analytique SQL. Chaque lakehouse dispose d'un point de terminaison d'analytique SQL généré automatiquement, qui peut être utilisé pour effectuer la transition entre la vue lac du lakehouse et la vue SQL du lakehouse. La vue lac prend en charge l'engineering données et Apache Spark. La vue SQL du même lakehouse vous permet de créer des vues, des fonctions, des procédures stockées ainsi que d'appliquer la sécurité

SQL et des autorisations au niveau des objets.

Les données d'un lakehouse Fabric sont stockées dans la structure de dossiers suivante :

Afficher la vue du point de terminaison d'analytique SQL du lakehouse

Le point de terminaison d'analytique SQL est utilisé pour lire les données dans le dossier /Tables du lakehouse en T-SQL.

Appliquer des autorisations granulaires au lakehouse en T-SQL

À l'aide du point de terminaison d'analytique SQL, vous pouvez appliquer des autorisations T-SQL granulaires aux objets SQL en utilisant des commandes DCL (langage de contrôle de données) par exemple :

Vous pouvez également appliquer la sécurité au niveau des lignes, la sécurité au niveau des colonnes et le Dynamic Data Masking à l'aide du point de terminaison d'analytique SQL. Consultez l'article :

Sécurité au niveau des lignes

Sécurité au niveau des colonnes

Masquage des données dynamiques

Configurer l'accès aux données via la vue lac du lakehouse

La vue lac du lakehouse est utilisée pour lire les données dans les dossiers /Tables et /Files du lakehouse.

Utiliser les rôles d'accès aux données OneLake pour sécuriser les données

Les autorisations relatives aux espaces de travail et aux éléments fournissent un accès grossier aux données d'un lakehouse. Pour affiner davantage l'accès aux données, vous pouvez sécuriser les dossiers de la vue lac du lakehouse à l'aide des rôles d'accès aux données OneLake (préversion). Vous pouvez créer des rôles personnalisés au sein d'un lakehouse, et octroyer des autorisations d'accès en lecture uniquement à des dossiers spécifiques dans OneLake. La sécurité des dossiers peut être héritée par tous les sous-dossiers. Pour créer un rôle d'accès aux données OneLake personnalisé :

Sélectionnez Gérer l'accès aux données OneLake (préversion) dans le menu de la vue lac du lakehouse.

Dans la fenêtre Nouveau rôle, créez un nom de rôle, puis sélectionnez les dossiers auxquels octroyer l'accès.

Une fois le rôle créé, attribuez-le à un utilisateur ou un groupe, puis sélectionnez les autorisations à octroyer.

Pour plus d'informations sur la façon dont les autorisations RBAC OneLake sont évaluées avec les autorisations relatives aux espaces de travail et aux éléments, consultez : Comment les autorisations RBAC OneLake sont évaluées avec des autorisations Fabric

Configurer des autorisations d'entrepôt granulaires

Vous pouvez appliquer des autorisations granulaires aux entrepôts à l'aide du point de terminaison d'analytique SQL, de manière similaire à l'utilisation du point de terminaison pour le lakehouse. Les mêmes autorisations peuvent être appliquées : GRANT, REVOKE et DENY ainsi que la sécurité au niveau des lignes, la sécurité au niveau des colonnes et le Dynamic Data Masking.

Configurer des autorisations de modèle sémantique

Le rôle d’un utilisateur dans un espace de travail lui octroie implicitement une autorisation d’accès aux modèles sémantiques d’un espace de travail. Les modèles sémantiques permettent de définir la sécurité à l’aide d’expressions DAX. Vous pouvez appliquer des autorisations plus granulaires à l’aide de la sécurité SNL (sécurité au niveau des lignes). Pour en savoir plus sur la gestion de la sécurité SNL ou des autorisations du modèle sémantique, consultez :

Autorisations de modèle sémantique

Sécurité au niveau des lignes (RLS) avec Power BI

Unité 5: Exercice : Sécuriser l’accès aux données dans Microsoft Fabric

Type: Exercice

Exercice : Sécuriser l’accès aux données dans Microsoft Fabric

À présent, vous allez pouvoir sécuriser un accès aux données dans Microsoft Fabric.

Dans cet exercice, vous apprenez à sécuriser l'accès aux données dans Fabric en utilisant les concepts explorés dans ce module.

Vous avez besoin d’une licence d’évaluation Microsoft Fabric avec la préversion Fabric activée dans votre locataire. Consultez Bien démarrer avec Fabric pour activer votre licence d’évaluation Fabric.

Pour effectuer les exercices de ce labo, vous aurez besoin de deux utilisateurs : un utilisateur doit être affecté au rôle Administrateur de l’espace de travail et l’autre sera affecté aux autorisations tout au long de ce labo. Pour attribuer des rôles à des espaces de travail, consultez Accorder l’accès à votre espace de travail .

Lancez l’exercice et suivez les instructions.

Unité 6: Évaluation du module

Type: Évaluation

Évaluation du module

Dans quel ordre l’évaluation de l’accès se fait-elle dans Fabric ?

Sécurité des données, accès à Fabric, authentification Microsoft Entra ID

Authentification Microsoft Entra ID, accès à Fabric, sécurité des données

Accès Fabric, authentification Microsoft Entra ID, sécurité des données

Quel rôle d’espace de travail doit être attribué à un ingénieur de données qui doit créer des éléments Fabric et lire toutes les données d’un lakehouse existant ?

Parmi les outils suivants, lequel peut être utilisé pour appliquer des autorisations granulaires d’accès aux données dans Fabric ?

Rôles d'accès aux données OneLake

Langage de manipulation de données (DML)

Sécurité au niveau des colonnes dans le modèle sémantique

Vous devez répondre à toutes les questions avant de vérifier votre travail.

Unité 7: Résumé

Type: Résumé

Dans ce module, vous avez appris à utiliser les fonctions de contrôle d'accès disponibles dans les différents moteurs Fabric pour sécuriser vos données et fournir à votre équipe l'accès nécessaire au sein de Fabric pour qu'elle puisse s'acquitter de ses tâches. Vous avez découvert le modèle de sécurité multicouche de Fabric et comment l’utiliser pour gérer l’accès aux données.

Ce module permet notamment de comprendre comment configurer la sécurité pour l'ensemble d'un espace de travail et pour des éléments de données Fabric individuels, et comment appliquer des autorisations granulaires.

Pour plus d’informations, vous pouvez consulter les URL suivants :

Livre blanc de sécurité Microsoft Fabric

Modèle d’autorisation Microsoft Fabric

Créer des architectures de données communes avec OneLake dans Microsoft Fabric

Comment sécuriser les données pour les architectures courantes de données Fabric

Scénario de sécurité de bout en bout

Module 3: Sécuriser un entrepôt de données Microsoft Fabric

Unité 1: Présentation

Type: Introduction

Microsoft Fabric Data Warehouse est une plateforme complète pour les données, l'analyse et l'IA (intelligence artificielle). Elle fait référence au processus de stockage, d'organisation et de gestion de grands volumes de données structurées et semi-structurées.

Dans un entrepôt, les administrateurs ont accès à une suite de technologies destinées à protéger les informations sensibles. Ces mesures de sécurité sont capables de sécuriser ou de masquer les données des utilisateurs ou des rôles sans autorisation appropriée, ce qui garantit la protection des données sur les points de terminaison de l'analytique SQL et de l'entrepôt. Cela garantit une expérience utilisateur fluide et sécurisée, sans qu'il soit nécessaire de modifier les applications existantes.

Comprendre les fonctionnalités de sécurité

Les ingénieurs Données, qui souvent connaissent bien le moteur SQL et sont experts dans l'utilisation de T-SQL, apprécient la facilité d'utilisation des entrepôts dans Microsoft Fabric.

C'est parce que les entrepôts sont alimentés par le même moteur SQL que celui qu'ils connaissent déjà, ce qui leur permet d'effectuer des requêtes et manipulations de données complexes. La vaste gamme de fonctionnalités de sécurité du moteur SQL permet d'avoir un mécanisme de sécurité sophistiqué au niveau de l'entrepôt.

Rôles d'espaces de travail : conçus pour fournir différents niveaux d'accès et de contrôle dans l'espace de travail. Vous pouvez attribuer des utilisateurs aux différents rôles d'espace de travail comme Administrateur, Membre, Contributeur et Viewer. Ces rôles sont essentiels pour maintenir la sécurité et l'efficacité des opérations d'entreposage de données au sein d'une organisation.

Autorisations d'élément : les entrepôts individuels peuvent avoir des autorisations d'élément attribuées directement. L'objectif principal de l'attribution de ces autorisations est de faciliter le partage de l'entrepôt pour une utilisation en aval.

Sécurité de protection des données : pour un contrôle plus précis, vous pouvez utiliser T-SQL pour accorder des autorisations spécifiques aux utilisateurs. L'entrepôt prend en charge une gamme de fonctionnalités de protection des données qui permettent aux administrateurs de protéger les données sensibles contre les accès non autorisés. Cela inclut la sécurité au niveau objet pour les objets de base de données, la sécurité au niveau colonne pour les colonnes de table, la sécurité au niveau ligne pour les lignes de table en utilisant des filtres de clause WHERE et Dynamic Data Masking pour masquer les données sensibles comme les adresses e-mail. Ces fonctionnalités garantissent la protection des données sur les points de terminaison des entrepôts et de l'analytique SQL sans avoir besoin de modifier les applications.

Dans les unités suivantes, nous explorons les différentes façons d'activer la sécurité dans un entrepôt et comment ces méthodes peuvent faciliter les tâches liées à la protection de la charge de travail de votre entrepôt de données.

Unité 2: Découverte du masquage dynamique des données

Type: Contenu

Découverte du masquage dynamique des données

Dynamic Data Masking (DDM) est une fonctionnalité de sécurité qui limite l'exposition des données aux utilisateurs non privilégiés en masquant les informations sensibles.

Le masquage dynamique des données offre plusieurs avantages clés qui améliorent la sécurité et la facilité de gestion de vos données. L'un des principaux avantages est sa fonctionnalité de masquage en temps réel. Lors de l'interrogation de données sensibles, DDM applique un masquage dynamique en temps réel. Ce processus signifie que les données réelles ne sont jamais exposées à des utilisateurs non autorisés, ce qui améliore la sécurité de vos données. En outre, DDM est simple à implémenter. Il ne nécessite pas de codage complexe, ce qui le rend accessible aux utilisateurs de tous les niveaux de compétence.

Un autre avantage de DDM est que les données de la base de données ne sont pas modifiées lorsque DDM est appliqué. Au lieu de cela, les règles de masquage sont appliquées aux résultats de la requête. Cet avantage signifie que les données réelles restent intactes et sécurisées, tandis que les utilisateurs non privilégiés voient uniquement une version masquée des données.

Définir une règle de masquage

Le masquage dynamique des données, qui est configuré au niveau de la colonne, offre une suite de fonctionnalités, notamment des fonctionnalités de masquage complètes et partielles, ainsi qu'une fonction de masquage aléatoire conçue pour les données numériques.

Les paramètres prefix_padding et suffix_padding de la fonction partial() spécifient le nombre de caractères à exposer au début et à la fin de la chaîne, et le paramètre padding_string spécifie la chaîne à utiliser pour masquer les caractères restants.

Les paramètres low et high de la fonction random() spécifient la plage de nombres aléatoires à générer.

Ces types de masquage permettent d'empêcher l'affichage non autorisé de données sensibles en permettant aux administrateurs de spécifier la quantité de données sensibles à révéler, avec un effet minimal sur la couche application. Elles sont appliquées aux résultats des requêtes, de sorte que les données de la base de données ne sont pas modifiées. Cette approche permet à de nombreuses applications de masquer les données sensibles sans modifier les requêtes existantes.

Configurer le masquage des données

Prenons un exemple d'entrepôt qui stocke des informations client. L'entrepôt contient une table Customer avec des champs tels que CustomerName, Email, PhoneNumber et CreditCardNumber.

Pour appliquer le masquage des données sur les colonnes CustomerName, Email, PhoneNumber et CreditCardNumber, exécutez la commande suivante :

Afficher les résultats masqués

Sans Dynamic Data Masking, si un utilisateur non privilégié exécute une requête pour récupérer les détails du client, il peut voir quelque chose comme suit :

Toutefois, avec DDM appliqué aux champs Email, PhoneNumber et CreditCardNumber, la même requête retourne :

Comme vous pouvez le voir, les données sensibles sont masquées pour l'utilisateur non privilégié, ce qui améliore la sécurité de vos données. Ce scénario est un exemple de base du fonctionnement de Dynamic Data Masking. Il permet de s'assurer que les données sensibles ne sont pas exposées aux utilisateurs qui n'ont pas les privilèges nécessaires pour les afficher.

Les utilisateurs non privilégiés disposant d'autorisations de requête peuvent déduire les données réelles, car les données ne sont pas physiquement masquées.

DDM doit être utilisé dans le cadre d'une stratégie complète de sécurité des données qui inclut une gestion appropriée de la sécurité au niveau objet avec des autorisations granulaires SQL et l'adhésion au principe des autorisations minimales requises.

Unité 3: Implémenter la sécurité au niveau des lignes

Type: Contenu

Implémenter la sécurité au niveau des lignes

La Sécurité au niveau des lignes (RLS) est une fonctionnalité qui fournit un contrôle granulaire sur l'accès aux lignes d'une table en fonction de l'appartenance à un groupe ou du contexte d'exécution.

Par exemple, dans une plateforme d'e-commerce, vous pouvez faire en sorte que les vendeurs aient accès seulement aux lignes de commande associées à leurs propres produits. De cette façon, chaque vendeur peut gérer ses commandes indépendamment, tout en conservant la confidentialité des informations sur les commandes d'autres vendeurs.

Si vous avez de l'expérience avec SQL Server, vous pouvez constater que la sécurité au niveau des lignes partage avec celui-ci des caractéristiques et des fonctionnalités similaires.

Protéger vos données

La sécurité au niveau des lignes (RLS) fonctionne en associant une fonction, appelée prédicat de sécurité, à une table. Cette fonction est définie pour retourner true ou false en fonction de certaines conditions, impliquant généralement les valeurs d'une ou plusieurs colonnes de la table. Quand un utilisateur tente d'accéder aux données de la table, la fonction de prédicat de sécurité est appelée. Si la fonction retourne true, la ligne est accessible à l'utilisateur ; si elle retourne false, la ligne n'apparaît pas dans les résultats de la requête.

Selon les besoins de l'entreprise, la sécurité au niveau des lignes peut être aussi simple que WHERE CustomerId = 29 ou plus complexe si nécessaire.

Ce processus est transparent pour l'utilisateur et est appliqué automatiquement par SQL Server, ce qui garantit une application cohérente des règles de sécurité.

La sécurité au niveau des lignes est implémentée en deux étapes principales :

Prédicats de filtrage – c'est une fonction table qui filtre les résultats en fonction du prédicat défini. Accès Définition CHOISIR Impossible d'afficher les lignes filtrées. MISE À JOUR L'utilisateur ne peut pas mettre à jour des lignes qui sont filtrées. SUPPRIMER Impossible de supprimer des lignes filtrées. INSÉRER Non applicable.

Prédicats de filtrage – c'est une fonction table qui filtre les résultats en fonction du prédicat défini.

Stratégie de sécurité – c'est une stratégie de sécurité qui appelle une fonction table pour protéger l'accès aux lignes d'une table.

Comme le contrôle d'accès est configuré et appliqué au niveau de l'entrepôt, les modifications à apporter à l'application, si elles sont nécessaires, sont minimales. En outre, les utilisateurs peuvent accéder directement aux tables et interroger leurs propres données.

Configurer la sécurité au niveau des lignes

Les commandes T-SQL ci-dessous montrent comment utiliser la sécurité au niveau des lignes dans un scénario où l'accès utilisateur est séparé par le locataire :

Ensuite, nous créons un schéma et une fonction table, et nous accordons à l'utilisateur l'accès à la nouvelle fonction. Le prédicat WHERE @TenantName = USER_NAME() OR USER_NAME() = 'TenantAdmin' évalue si le nom d'utilisateur qui exécute la requête correspond aux valeurs de colonne TenantName.

L'utilisateur tenantAdmin@contoso.com doit voir toutes les lignes. Les utilisateurs tenant1@contoso.com à tenant5@contoso.com ne doivent voir que leurs propres lignes.

Si vous modifiez la stratégie de sécurité avec WITH (STATE = OFF);, vous constatez que les utilisateurs peuvent voir toutes les lignes.

Il existe un risque de fuite d'informations si un attaquant écrit une requête avec une clause WHERE spécialement conçue et, par exemple, une erreur de division par zéro, pour forcer une exception si la condition WHERE est vraie. Il s'agit d'une attaque par canal auxiliaire.

Explorer des cas d'usage

La sécurité au niveau des lignes est idéale pour de nombreux scénarios, notamment :

Lorsque vous devez isoler l'accès départemental au niveau de la ligne.

Lorsque vous devez restreindre l'accès aux données de clients aux seules données relatives à leur entreprise.

Lorsque vous devez restreindre l'accès à des fins de conformité.

Appliquer les meilleures pratiques

Voici quelques bonnes pratiques à prendre en compte lors de l'implémentation de la sécurité au niveau des lignes :

Il est recommandé de créer un schéma distinct pour les fonctions de prédicat et les stratégies de sécurité.

Dans la mesure du possible, évitez les conversions de types dans les fonctions de prédicat.

Pour optimiser les performances, évitez d'utiliser des jointures de table excessives et une récursivité dans les fonctions de prédicat.

Unité 4: Implémenter la sécurité au niveau des colonnes

Type: Contenu

Implémenter la sécurité au niveau des colonnes

La sécurité au niveau des colonnes (CLS) vous permet de restreindre l'accès aux colonnes afin de protéger les données sensibles. Elle fournit un contrôle précis de l'accès aux éléments de données spécifiques, ce qui améliore la sécurité globale de votre entrepôt de données.

Sécuriser les données sensibles

Prenons un exemple pratique de sécurité au niveau des colonnes (CLS) dans le secteur de la santé. Supposons que nous avons une table nommée Patients avec les colonnes suivantes : PatientID, Name, Address, DateOfBirth et MedicalHistory.

La colonne MedicalHistory contient des informations sensibles sur la santé des patients. Conformément aux réglementations en matière de soins de santé et aux lois sur la protection des données personnelles, seul le personnel médical autorisé, dont les médecins et les infirmières, doit pouvoir accéder à ces informations.

Voici une façon implémenter la sécurité au niveau des colonnes dans ce scénario :

Identifiez les colonnes sensibles : dans ce cas, la MedicalHistory colonne est identifiée comme contenant des données sensibles.

Définir des rôles d'accès : définissez des rôles tels que Doctor et Nurse qui sont autorisés à accéder à la MedicalHistory colonne. L'accès à cette colonne peut être restreint pour d'autres rôles, tels que Receptionist ou Patient.

Attribuer des rôles aux utilisateurs : attribuez les rôles appropriés à chaque utilisateur de l'entrepôt. Par exemple, le rôle DrSmith peut être affecté à l'utilisateur Doctor, tandis que le rôle JohnDoe peut être affecté à l'utilisateur Patient.

Implémenter le contrôle d'accès : restreindre l'accès à la MedicalHistory colonne en fonction du rôle de l'utilisateur.

La sécurité au niveau des colonnes peut vous aider à garantir que les informations sensibles sur la santé peuvent uniquement être consultées par des personnes autorisées grâce à la protection des données personnelles des patients et le respect des réglementations en matière de soins de santé.

Configurer la sécurité au niveau de la colonne

Dans le scénario que nous venons de découvrir, la syntaxe pour implémenter la sécurité au niveau des colonnes peut revêtir la forme suivante :

Dans cet exemple, nous créons d'abord les rôles Doctor, Nurse, Receptionist et Patient. Nous accordons ensuite à tous les rôles les autorisations SELECT sur toutes les colonnes de la table Patients. Enfin, nous refusons les autorisations SELECT sur la colonne MedicalHistory pour les rôles Receptionist et Patient. Cela garantit que seuls les utilisateurs dotés du rôle Doctor ou Nurse peuvent accéder à la colonne MedicalHistory.

Comprendre les avantages

Dans la sécurité d'entrepôt, deux des techniques les plus utilisées sont la sécurité et les vues au niveau des colonnes. Ces deux méthodes permettent de restreindre l'accès aux données sensibles, mais elles le font de différentes manières et proposent différents avantages. Le tableau suivant établit une analyse comparative de ces deux techniques sur différents aspects, dont la granularité du contrôle d'accès, la maintenance, les performances, la transparence et la flexibilité.

Cette comparaison peut vous aider à comprendre les points forts et les faiblesses de chaque méthode. Elle peut également vous aider à choisir l'approche la plus appropriée pour vos exigences d'application spécifiques.

Le choix entre la sécurité ou les vues au niveau des colonnes dépend des exigences spécifiques de votre application. Veillez à tester systématiquement les modifications de sécurité dans un environnement sécurisé avant de les appliquer à un entrepôt de production.

Unité 5: Configurer des autorisations granulaires SQL à l'aide de T-SQL

Type: Contenu

Configurer des autorisations granulaires SQL à l'aide de T-SQL

Si vous êtes familiarisé avec les bases de données relationnelles et les entrepôts d'entreprise, vous savez qu'il existe quatre autorisations fondamentales régissant les opérations de langage de manipulation de données (DML). Ces autorisations, à savoir SELECT, INSERT, UPDATEet DELETE, sont universellement applicables sur toutes les plateformes de base de données.

Toutes ces autorisations peuvent être accordées, révoquées ou refusées sur les tables et les vues. Si une autorisation est accordée avec l'instruction GRANT, l'autorisation est donnée à l'utilisateur ou au rôle référencé dans l'instruction GRANT. Les utilisateurs peuvent également se voir refuser des autorisations avec la commande DENY. Si un utilisateur se voit accorder puis refuser une même autorisation, DENY l'emporte toujours sur l'octroi, donc l'accès à l'objet spécifique est refusé à l'utilisateur.

Autorisations des tables et des vues

Les tables et les vues représentent les objets sur lesquels des autorisations peuvent être accordées au sein d'un entrepôt. Dans ces tables et ces vues, vous pouvez aussi limiter les colonnes qui sont accessibles à un principal de sécurité donné.

Autorisations des fonctions et des procédures stockées

Comme les tables et les vues, les fonctions et les procédures stockées disposent de plusieurs autorisations qui peuvent être accordées ou refusées.

Principe du privilège minimum

L'idée de base du principe des privilèges minimum est que les utilisateurs et les applications doivent se voir accorder uniquement les autorisations nécessaires pour effectuer la tâche. Les applications doivent disposer uniquement des autorisations nécessaires pour effectuer la tâche en cours.

Par exemple, si une application accède à toutes les données via des procédures stockées, elle doit uniquement avoir l'autorisation d'exécuter les procédures stockées, sans accès aux tables.

Le SQL dynamique est un concept selon lequel une requête est créée programmatiquement. Le SQL dynamique permet de générer des instructions T-SQL dans une procédure stockée ou dans une requête elle-même. Un exemple simple est illustré ci-dessous.

Dans cet exemple, @tableName est le paramètre que vous pouvez remplacer par le nom de la table à inspecter. La fonction QUOTENAME est utilisée pour citer en toute sécurité le nom de la table, ce qui empêche les attaques par injection SQL. La procédure stockée sp_executesql est ensuite utilisée pour exécuter la requête générée dynamiquement.

Notez qu'il s'agit d'un exemple simple et que les tâches d'entrepôt de données réelles peuvent nécessiter des requêtes SQL dynamiques plus complexes. Soyez toujours prudent lors de l'utilisation de SQL dynamique en raison du risque d'attaques par injection SQL. Utilisez toujours des méthodes de paramétrisation comme sp_executesql ou QUOTENAME pour nettoyer les entrées.

Unité 6: Exercice : Sécuriser un entrepôt dans Microsoft Fabric

Type: Exercice

Exercice : Sécuriser un entrepôt dans Microsoft Fabric

À présent, vous allez pouvoir sécuriser un entrepôt dans Microsoft Fabric.

Dans cet exercice, vous allez apprendre à sécuriser un entrepôt en utilisant les concepts découverts dans ce module.

Vous avez besoin d'une licence d'évaluation Microsoft Fabric avec la préversion Fabric activée dans votre locataire. Consultez Bien démarrer avec Fabric pour activer votre licence d'évaluation Fabric.

Pour effectuer les exercices de ce labo, vous avez besoin de deux identités d'utilisateur. Si vous ne parvenez pas à créer un deuxième utilisateur, vous pouvez toujours effectuer l'exercice en utilisant votre compte d'utilisateur, mais vous ne pourrez pas faire l'expérience de ce qu'un utilisateur moins privilégié voit lorsque l'accès à des fonctionnalités spécifiques lui est accordé ou restreint.

Comment créer un deuxième utilisateur pour cet exercice

Si vous faites partie d'une organisation disposant d'un locataire Entra ou Microsoft 365 :

Collaborez avec votre administrateur d'identité pour créer le deuxième utilisateur dans Entra ou le Centre d'administration Microsoft 365.

Si vous n'êtes pas membre d'une organisation avec un locataire Entra ou Microsoft 365 :

Vous ne pouvez pas vous inscrire à un essai de Fabric avec votre adresse e-mail personnelle. Vous pouvez vous inscrire à un essai de Microsoft 365, et un locataire d'organisation est créé. Vous devenez alors administrateur d'utilisateurs et de facturation du locataire et pourrez créer des utilisateurs.

Créez le deuxième utilisateur dans le Centre d'administration Microsoft 365. Consultez : des utilisateurs

Activez l'essai de Fabric en vous reportant à l'article Bien démarrer avec Fabric.

Lancez l'exercice et suivez les instructions.

Unité 7: Évaluation du module

Type: Évaluation

Évaluation du module

Quel est le principal avantage de DDM (Dynamic Data Masking) ?

Il limite l'exposition des données en masquant les informations sensibles en temps réel.

Il change les données réelles de la base de données.

Son implémentation nécessite l'écriture d'un code complexe.

À quoi sert une fonction de prédicat de sécurité dans le cadre de la sécurité SNL (sécurité au niveau des lignes) ?

Elle détermine si une ligne est accessible à un utilisateur en fonction de certaines conditions.

Elle permet les conversions de type dans les fonctions de prédicat.

Elle permet aux utilisateurs d'exécuter des requêtes ad hoc.

Que se passe-t-il quand un utilisateur se voit octroyer une autorisation, puis se voit refuser cette même autorisation dans un entrepôt ?

L'instruction GRANT remplace l'instruction DENY, et l'utilisateur a accès à l'objet spécifique.

L’instruction DENY remplace toujours l’instruction GRANT, l’utilisateur se voit donc refuser l’accès à l’objet spécifique.

L’utilisateur dispose des deux autorisations, ce qui provoque un conflit.

Vous devez répondre à toutes les questions avant de vérifier votre travail.

Unité 8: Résumé

Type: Résumé

Dans ce module, vous avez découvert Dynamic Data Masking (DDM), la sécurité au niveau ligne (RLS), la sécurité au niveau colonne (CLS) et les autorisations SQL granulaires dans les entrepôts Fabric.

Les principaux points de départ de ce module incluent la compréhension de la façon dont DDM, RLS et CLS fonctionnent et leurs cas d’usage. DDM fonctionne au niveau colonne, offrant des types de masquage d’e-mail, de texte personnalisé et aléatoire complets. RLS fonctionne en associant une fonction de prédicat de sécurité à une table. CLS peut être implémenté en identifiant les colonnes sensibles, en définissant des rôles d’accès, en affectant des rôles aux utilisateurs et en implémentant le contrôle d’accès. De plus, vous avez appris le principe des privilèges minimum qui suggère que les utilisateurs et les applications doivent se voir accorder uniquement les autorisations nécessaires pour effectuer leurs tâches.

Pour obtenir une lecture supplémentaire, vous pouvez consulter les URL suivantes :

Créer un entrepôt dans Microsoft Fabric

Sécurité de l’entreposage de données dans Microsoft Fabric

Partager votre entrepôt et gérer les autorisations

Module 4: Gouverner les données dans Microsoft Fabric avec Purview