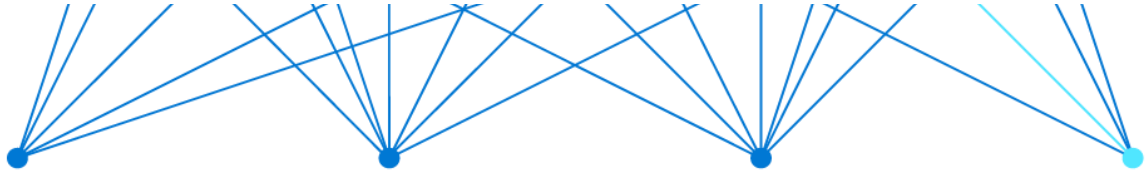
Microsoft Azure

Les Principes de base sur les données Microsoft Azure (Microsoft DP-900)

Dr.Ing Semeh Ben Salem
MCT, Azure Data Technologies
05/06/2023



© Copyright Microsoft Corporation. Tous droits réservés.

La certification DP-900

- Connaitre les principaux concepts des données et les services de données Microsoft Azure associés.
- Débuter avec les données sur le cloud: les charges de travail, les services, etc.
- Vous devez être familiarisé avec les points suivants :
  - Concepts des données relationnelles et non relationnelles.
  - Différents types de charges de travail de données, par exemple transactionnelles ou analytiques.
- Préparer d'autres certifications Azure: **Azure Database Administrator Associate (DP-300)** ou **Azure Data Engineer Associate (DP-203)**.
- La version anglaise de cet examen sera mise à jour le 4 août 2023.
- **Les compétences mesurées:**
  - Décrire les principaux concepts de données (25 à 30 %)
  - Identifier les points à prendre en compte pour les données relationnelles sur Azure (20 à 25 %)
  - Décrire les points à prendre en compte pour l'utilisation de données non relationnelles sur Azure (15 à 20 %)
  - Décrire une charge de travail analytique sur Azure (25-30%)

© Copyright Microsoft Corporation. All rights reserved.

Emploi du temps du cours

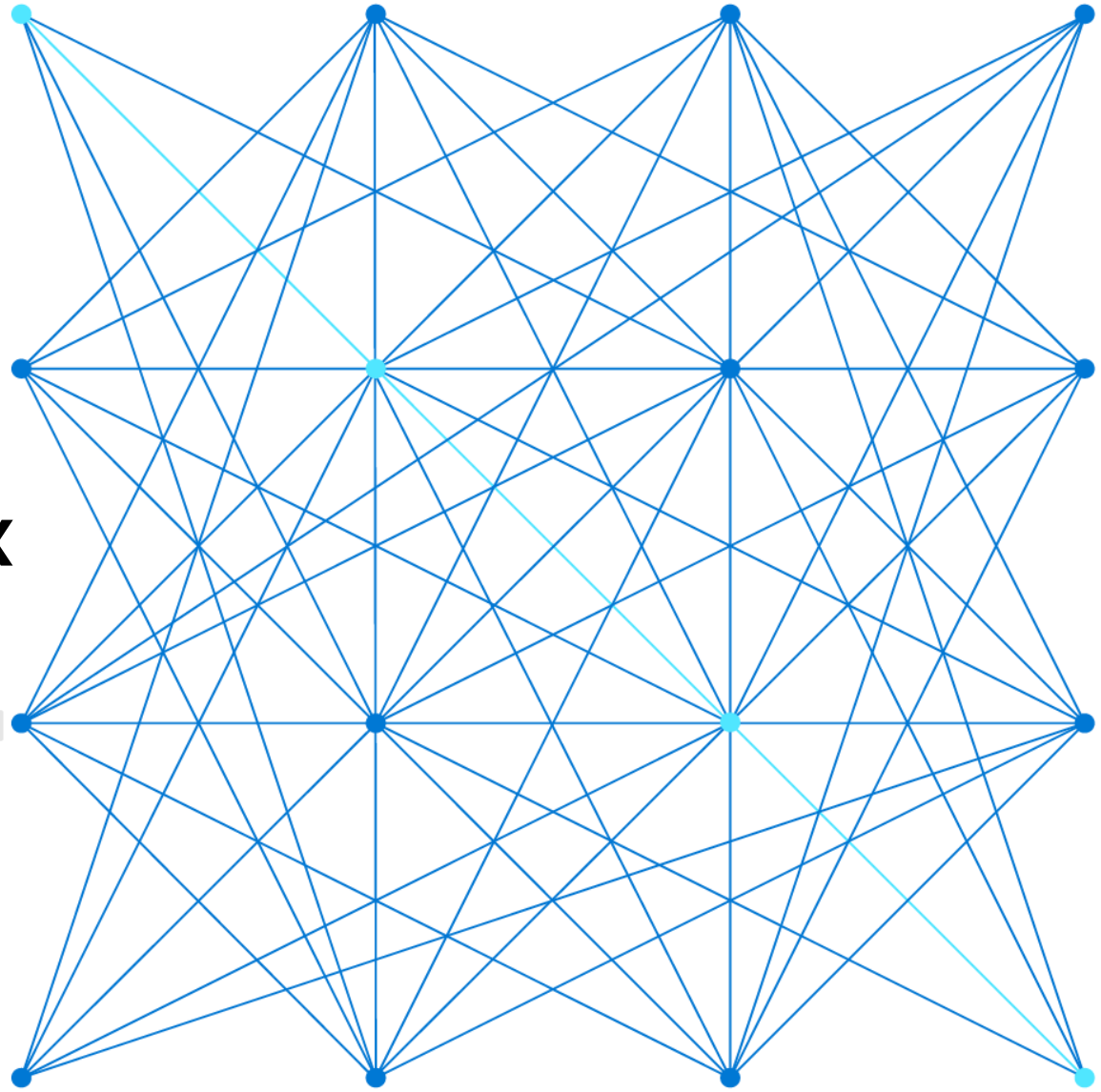
Contenu	Unités
Explorer les concepts de base des données	• Concepts clés des données • Rôles et services de données
Explorer les notions fondamentales des données relationnelles dans Azure	• Explorer les concepts des données relationnelles • Explorer les services Azure pour les données relationnelles
Explorer les notions fondamentales des données non relationnelles dans Azure	• Bases du Stockage Azure • Bases d'Azure Cosmos DB
Explorer les notions fondamentales de l'analytique données	• Entreposage de données à grande échelle • Streaming et analytique en temps réel • Visualisation des données

© Copyright Microsoft Corporation. All rights reserved.

Microsoft Azure

Module 1 : Explorer les principaux concepts de données

Débutant | Analyste de données | Ingénieur Data | Administrateur de base de données | Développeur | Architecte Solutions | Étudiant | Azure



© Copyright Microsoft Corporation. Tous droits réservés.

Objectifs d’apprentissage

Dans ce module, vous allez découvrir comment :

- Identifier les formats de données courants
- Décrire les options de stockage des données dans les fichiers
- Décrire les options de stockage des données dans les bases de données
- Décrire les caractéristiques des solutions de traitement des données transactionnelles
- Décrire les caractéristiques des solutions de traitement des données analytiques

© Copyright Microsoft Corporation. All rights reserved.

Que sont les données ?

Des valeurs utilisées pour enregistrer des informations, représentant souvent des *entités* qui ont un ou plusieurs *attributs*

Données structurées

Client

ID	Prénom	Nom	E-mail	Adresse
1	Joe	Jones	joe@litware.com	1 Main St.
2	Samir	Nadoy	samir@northwind.com	123 Elm Pl.

Produit

ID	Nom	Prix
123	Marteau	2.99
162	Tournevis	3.49
201	Clé	4.25

Données semi-structurées

```
{
    "firstName": "Joe",
    "lastName": "Jones",
    "address":
    {
        "streetAddress": "1 Main
St.",
        "city": "New York",
        "state": "NY",
        "postalCode": "10099"
    },
    "contact":
    [
        {
          "type": "home",
          "number": "555 123-1234"
        },
        {
          "type": "email",
          "address":
"joe@litware.com"
        }
    ]
}
```

```
{
    "firstName": "Samir",
    "lastName": "Nadoy",
    "address":
    {
        "streetAddress": "123 Elm
Pl.",
        "unit": "500",
        "city": "Seattle",
        "state": "WA",
        "postalCode": "98999"
    },
    "contact":
    [
        {
          "type": "email",
          "address":
"samir@northwind.com"
        }
    ]
}
```

Données non structurées

© Copyright Microsoft Corporation. All rights reserved.

Stockage des données: les magasins de données

Magasins de fichiers

Texte délimité

```
FirstName,LastName,Email
Joe,Jones,joe@litware.com
Samir,Nadoy,samir@northwind.com
```

JSON (JavaScript Object Notation)

```
{
  "customers":
  [
    { "firstName": "Joe", "lastName": "Jones"},
    { "firstName": "Samir", "lastName": "Nadoy"}
  ]
}
```

XML (Extensible Markup Language)

```
<Customer firstName="Joe" lastName="Jones"/>
```

BLOB (Binary Large Object)

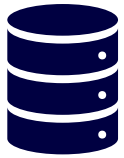
```
101101011010101100101...
```

Formats optimisés :

- Avro, ORC, Parquet

Bases de données

Relationnel



Client

ID	Courrier	Adresse
1	joe@litware.com	1 Main St.
2	samir@northwind.com	123 Elm Pl.

Produit

ID	Nom	Prix
123	Marteau	2.99
162	Tournevis	3.49
201	Clé	4.25

Commande

N°Commande	DateCommande	Client
1000	01/01/2022	1
1001	01/01/2022	2

LineItem

N°Commande	N°Article	ID Produit	Quantité
1000	1	123	1
1000	2	201	2
1001	1	123	2

Clé primaire – clé étrangère - normalisation – SQL -

Non relationnelle

Produits

Clé	Valeur
123	"Hammer ($2.99)"
162	"Screwdriver ($3.49)"
201	"Wrench ($4.25)"

Clé-Valeur

Clients

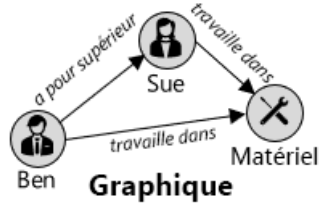
Clé	Document
1	{ "name": "Joe Jones" }
2	{ "name": "Samir Nadoy" }

Document

Commandes

Clé	Client		Produit	
	Nom	Adresse	Nom	Prix
1000	Joe Jones	1 Main St.	Hammer	2.99
1001	Samir Nadoy	123 Elm Pl.	Wrench	4.25

Famille de colonnes



Graphique

© Copyright Microsoft Corporation. Tous droits réservés.

Les formats de fichiers optimisés

- Les formats traditionnels ne sont pas optimisés pour le Big Data (stockage & traitement).
- Il faut des formats de fichiers spécialisés: compression, indexation, stockage et traitement rapide.
- ***Avro***: format en ligne créé par Apache. Chaque enregistrement contient un en-tête qui décrit la structure des données de l'enregistrement stocké au format JSON binaire. Une application utilise les informations de l'en-tête pour analyser les données binaires et extraire les champs qu'elles contiennent. Avro est un bon format pour la compression des données, réduisant au minimum les besoins en stockage et en bande passante réseau.
- ***ORC*** (format Optimized Row Columnar) organise les données en colonnes développé par HortonWorks pour optimiser les opérations de lecture et d'écriture dans Apache Hive DW. Un fichier ORC contient des *bandes* de données. Chaque bande contient les données d'une colonne ou d'un ensemble de colonnes. Une bande contient un index dans les lignes de la bande, les données pour chaque ligne et un pied de page qui contient des informations statistiques (nombre de lignes, somme, max, min, etc.) pour chaque colonne.
- ***Parquet*** est un autre format de données en colonnes créé par Cloudera et Twitter. Un fichier Parquet contient des groupes de lignes. Les données pour chaque colonne sont stockées ensemble dans le même groupe de lignes. Chaque groupe de lignes contient un ou plusieurs blocs de données. Un fichier Parquet contient des métadonnées qui décrivent l'ensemble de lignes qui se trouvent dans chaque bloc. Une application peut utiliser ces métadonnées pour localiser rapidement le bloc approprié pour un ensemble donné de lignes et récupérer les données dans les colonnes spécifiées pour ces lignes. Parquet est spécialisé dans le stockage et le traitement efficaces des types de données imbriqués. Il prend en charge des schémas de compression et d'encodage très efficaces.

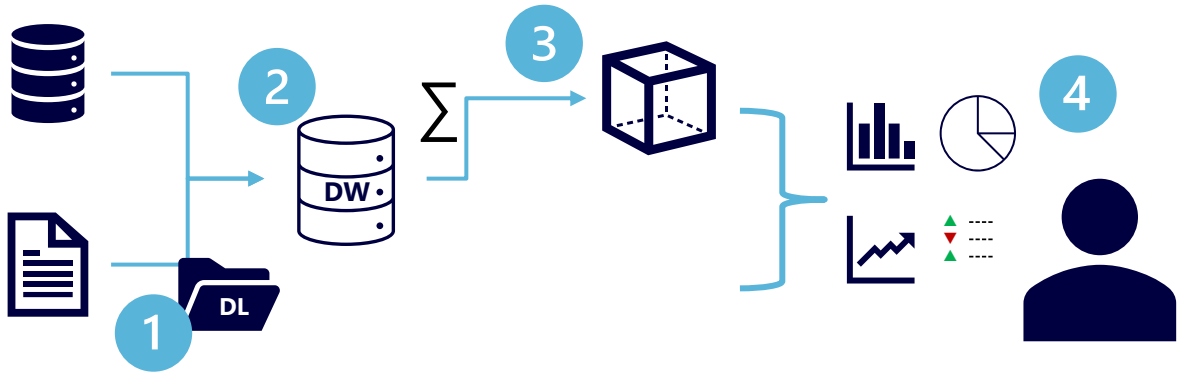
© Copyright Microsoft Corporation. Tous droits réservés.

Charges de travail de données transactionnelles

- Un système transactionnel d'entreprise enregistre des événements spécifiques appelés *transactions*: financière, vente, etc.
- Les systèmes transactionnels traitent souvent de gros volumes et peuvent parfois gérer plusieurs millions de transactions en une seule journée.
- Les données en cours de traitement doivent être accessibles très rapidement.
- Le travail effectué par les systèmes transactionnels est souvent appelé traitement transactionnel en ligne (OLTP, Online Transactional Processing).
- Les solutions OLTP s'appuient sur un système de BD dans lequel le stockage de données est optimisé pour les opérations de lecture, écriture et modification (*CRUD*).

© Copyright Microsoft Corporation. Tous droits réservés.

Charges de travail de données analytiques



1. Les fichiers de données peuvent être stockés dans un *lac de données* central pour analyse
2. Un processus d'extraction, de transformation et de chargement (ETL) copie des données depuis des fichiers et des bases de données OLTP dans un *entrepôt de données* optimisé pour l'activité de *lecture*
3. Les données de l'entrepôt de données peuvent être agrégées et chargées dans un modèle de traitement analytique en ligne (OLAP), appelé aussi *cube* à explorer avec des *hiérarchies*. On parle aussi *d'agrégations*.
4. Les données du lac de données, de l'entrepôt de données et du modèle analytique peuvent être interrogées pour produire des rapports et des tableaux de bord.

© Copyright Microsoft Corporation. Tous droits réservés.

Contrôle des connaissances

Comment sont organisées les données dans une table relationnelle ?

- [x] Lignes et colonnes
- [ ] En-tête et pied de page
- [ ] Pages et paragraphes

Parmi les exemples suivants, lequel représente des données non structurées ?

- [ ] Un fichier texte délimité par des virgules avec des champs *EmployeeID*, *EmployeeName* et *EmployeeDesignation*
- [x] Des fichiers audio et vidéo
- [ ] Une table dans une base de données relationnelle

Qu'est-ce qu'un entrepôt de données ?

- [ ] Base de données non relationnelle optimisée pour les opérations de lecture et d'écriture
- [x] Base de données relationnelle optimisée pour les opérations de lecture
- [ ] Emplacement de stockage pour les fichiers de données non structurées

© Copyright Microsoft Corporation. All rights reserved.

Rôles de professionnel des données

Administrateur de base de données (DBA)

- Provisionnement, configuration et gestion de bases de données
- Sécurité de la base de données et accès utilisateur
- Sauvegardes et résilience de base de données
- Monitoring et optimisation des performances de base de données

Ingénieur Data (Data Engineer)

- Pipelines d'intégration de données et processus ETL
- Nettoyage et transformation des données
- Schémas et chargements de données de magasin de données analytiques

Analyste Data (Data Analyst)

- Transformer et nettoyer les données
- Modélisation analytique
- Production de rapports et synthèse des données
- Visualisation des données

© Copyright Microsoft Corporation. All rights reserved.

Services cloud Microsoft pour les données

Magasins de données

Azure SQL

- Famille de services de base de données relationnelle basés sur SQL Server complètement managée PaaS.

Azure Database pour open source

- Maria DB, MySQL, PostgreSQL

Azure Cosmos DB

- Système de base de données non relationnelle à haute scalabilité

Stockage Azure

- Stockage et partage de fichiers, d'objets blob et de tables
- Espace de noms hiérarchique pour Data Lake Storage

Ingénierie et analytique des données

Azure Data Factory

- Définir et planifier des pipelines de données, ETL

Azure Synapse Analytics

- Analytique intégrée de bout en bout
- Pipelines, SQL, Apache Spark, Data Explorer, streaming, KQL ...

Azure Databricks

- Analytique et traitement des données Apache Spark, Notebooks, Pool Spark

Azure HDInsight

- Plateforme open source Apache, Spark, Hadoop, Kafka, Hbase (NoSQL)

Azure Stream Analytics

- Traitement des données en temps réel pour les solutions IoT

Explorateur de données Azure

- Analyse des données en temps réel pour les journaux et la télémétrie

Microsoft Purview

- Gouvernance des données d'entreprise
- Mappage et détectabilité des données

Microsoft Power BI

- Modélisation des données analytiques
- Visualisation interactive des données

© Copyright Microsoft Corporation. All rights reserved.

Contrôle des connaissances

Parmi les tâches suivantes, laquelle relève de la responsabilité d'un administrateur de base de données ?

- ☑ Sauvegarde et restauration de bases de données
- ☐ Création de tableaux de bord et de rapports
- ☐ Création de pipelines pour traiter les données dans un lac de données

Quel rôle est le plus susceptible d'utiliser Azure Data Factory pour définir un pipeline de données pour un processus ETL ?

- ☐ Administrateur de base de données
- ☑ Ingénieur Data
- ☐ Analyste Data

Quel service utiliseriez-vous pour implémenter des pipelines de données, une analytique SQL et une analytique Spark ?

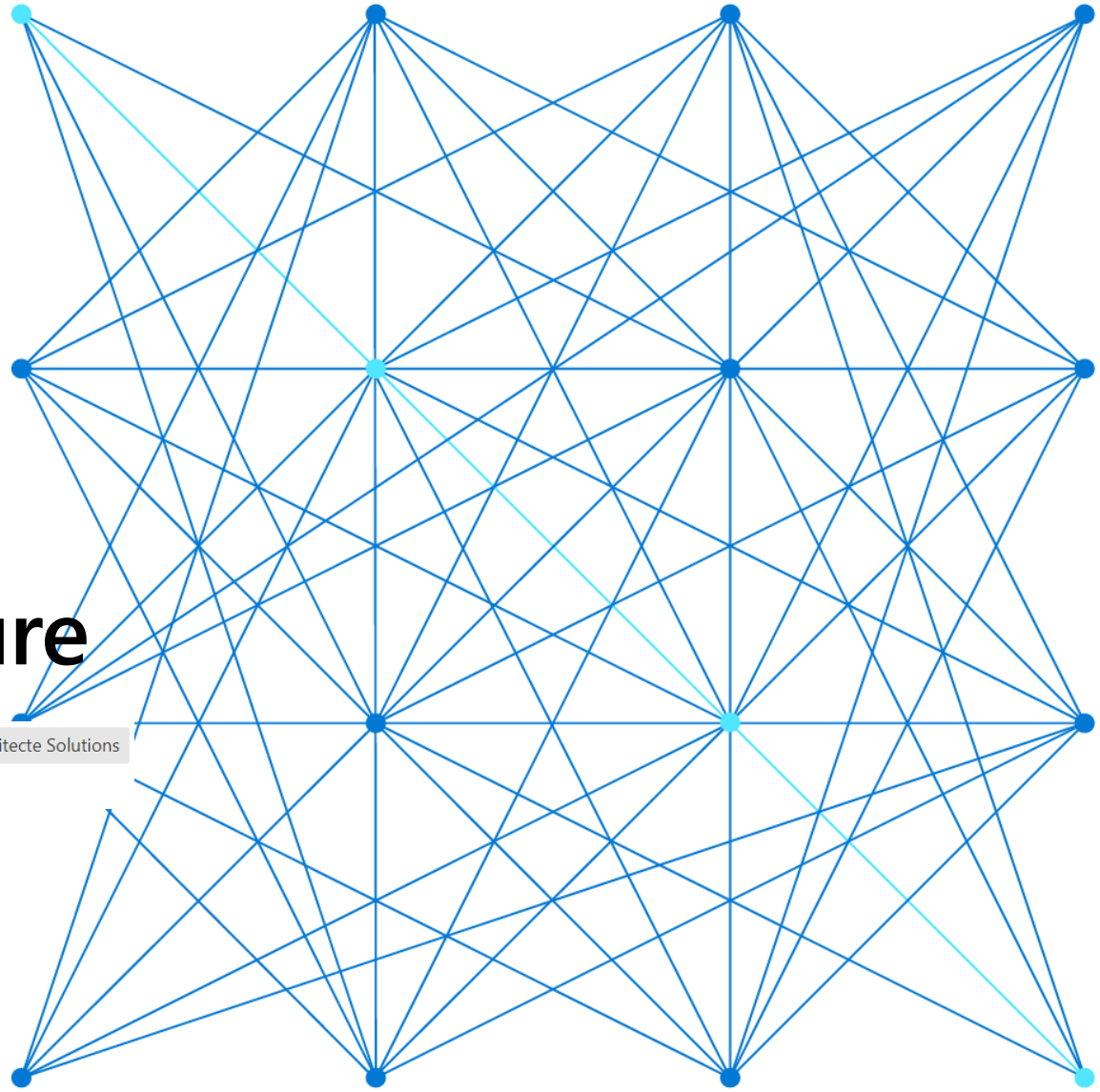
- ☐ Azure SQL Database
- ☐ Microsoft Power BI
- ☑ Azure Synapse Analytics

© Copyright Microsoft Corporation. All rights reserved.

Microsoft Azure

Module 2 : Explorer les bases des données relationnelles dans Azure

Débutant | Analyste de données | Ingénieur Data | Administrateur de base de données | Développeur | Architecte Solutions | Étudiant | Azure



© Copyright Microsoft Corporation. Tous droits réservés.

Objectifs d’apprentissage

Dans ce module, vous allez découvrir comment :

- Identifier les caractéristiques des données relationnelles
- Définir la normalisation
- Identifier les types d’instruction SQL
- Identifier les objets de base de données relationnelle communs

© Copyright Microsoft Corporation. All rights reserved.

Tables relationnelles

Manipuler des données structurées.

Les données sont stockées dans des *tables*

Les tables sont constituées de *lignes* et de *colonnes*

Parfois des valeurs vides (2ème prénom)

Toutes les lignes ont les mêmes colonnes

Un type de données est affecté à chaque colonne

Client

ID	Prénom	Deuxième prénom	Nom	E-mail	Adresse	Ville
1	Joe	David	Jones	joe@litware.com	1 Main St.	Seattle
2	Samir		Nadoy	samir@northwind.com	123 Elm Pl.	New York

Produit

ID	Nom	Prix
123	Marteau	2.99
162	Tournevis	3.49
201	Clé	4.25

Commande

N°Commande	DateCommande	Client
1000	01/01/2022	1
1001	01/01/2022	2

LineItem

N°Commande	N°Article	ID Produit	Quantité
1000	1	123	1
1000	2	201	2
1001	1	123	2

© Copyright Microsoft Corporation. Tous droits réservés.

Normalisation: minimiser la duplication

Données des ventes

N°Commande	DateCommande	Client	Produit	Quantité
1000	01/01/2022	Joe Jones, 1 Main St, Seattle	Marteau ($2.99)	1
1000	01/01/2022	Joe Jones- 1 Main St, Seattle	Tournevis ($3.49)	2
1001	01/01/2022	Samir Nadoy, 123 Elm Pl, New York	Marteau ($2.99)	2
…	…	…	…	…

1. Séparer chaque *entité* dans sa propre table
2. Séparer chaque *attribut* discret dans sa propre colonne
3. Identifier de façon univoque chaque instance d'entité (ligne) avec une *clé primaire*
4. Utilisez des colonnes de *clé étrangère* pour lier les entités associées

Client

ID	Prénom	Nom	Adresse	Ville
1	Joe	Jones	1 Main St.	Seattle
2	Samir	Nadoy	123 Elm Pl.	New York

Produit

ID	Nom	Prix
123	Marteau	2.99
162	Tournevis	3.49
201	Clé	4.25

Commande

N°Commande	DateCommande	Client
1000	01/01/2022	1
1001	01/01/2022	2

LineItem

N°Commande	N°Article	ID Produit	Quantité
1000	1	123	1
1000	2	201	2
1001	1	123	2

© Copyright Microsoft Corporation. Tous droits réservés.

SQL (Structured Query Language)

- SQL est un langage standard à utiliser avec les bases de données relationnelles
- Les standards sont gérés par les organisations ANSI et ISO
- La plupart des systèmes SGBDR prennent en charge les extensions propriétaires du SQL standard

Langage de définition de données (DDL)

CREATE, ALTER, DROP, RENAME

```
CREATE TABLE Product
(
  ProductID INT PRIMARY KEY,
  Name VARCHAR(20) NOT NULL,
  Price DECIMAL NULL
);
```

Produit

ID	Nom	Prix

Langage de contrôle de données (DCL)

GRANT, DENY, REVOKE

```
GRANT SELECT, INSERT, UPDATE
ON Product
TO user1;
```

Produit

ID	Nom	Prix
123	Marteau	2.99
162	Tournevis	3.49
201	Clé	4.25

Langage de manipulation de données (DML)

INSERT, UPDATE, DELETE, SELECT

```
SELECT Name, Price
FROM Product
WHERE Price > 2.50
ORDER BY Price;
```

Résultats

Nom	Prix
Marteau	2.99
Tournevis	3.49
Clé	4.25

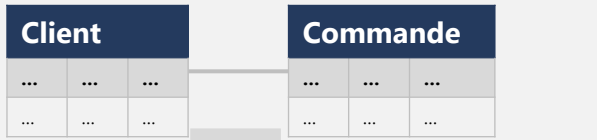
© Copyright Microsoft Corporation. Tous droits réservés.

Autres objets de base de données courants

Vues

- Des tables virtuelles basées sur le résultat d'un SELECT.
- Contient une partie des données.
- Interroger la vue et la filtrer comme une table.

```
CREATE VIEW Deliveries
AS
SELECT o.OrderNo, o.OrderDate,
       c.Address, c.City
FROM Order AS o JOIN Customer AS c
ON o.Customer = c.ID;
```



Livraisons

N°Commande	DateCommande	Adresse	Ville
1000	01/01/2022	1 Main St.	Seattle
1001	01/01/2022	123 Elm Pl.	New York

Procédures stockées

- Exécution à la demande.
- Des instructions SQL prédéfinies qui peuvent inclure des paramètres
- Renommer un produit

```
CREATE PROCEDURE RenameProduct
          @ProductID INT,
          @NewName VARCHAR(20)
AS
UPDATE Product
SET Name = @NewName
WHERE ID = @ProductID;
...
EXEC RenameProduct 201, 'Spanner';
```

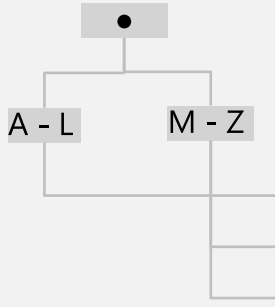
Produit

ID	Nom	Prix
201	~~Wrench~~ **Clé**	4.25

Index

- Rechercher des éléments dans une table.
- Créer des arborescences pour améliorer les performances.

```
CREATE INDEX idx_ProductName
ON Product(Name);
```



Produit

ID	Nom	Prix
123	Marteau	2.99
162	Tournevis	3.49
201	Clé	4.25

© Copyright Microsoft Corporation. Tous droits réservés.

Contrôle des connaissances

Parmi les affirmations suivantes, laquelle caractérise une base de données relationnelle ?

- ☐ Toutes les colonnes d'une table doivent être du même type de données
- ☑ Une ligne d'une table représente une instance unique d'une entité
- ☐ Les lignes d'une même table peuvent contenir des colonnes différentes

Quelle instruction SQL est utilisée pour interroger des tables et retourner des données ?

- ☐ QUERY
- ☐ READ
- ☑ SELECT

Qu'est-ce qu'un index ?

- ☑ Structure qui permet aux requêtes de localiser rapidement des lignes dans une table
- ☐ Une table virtuelle basée sur les résultats d'une requête
- ☐ Une instruction SQL prédéfinie qui modifie les données

© Copyright Microsoft Corporation. All rights reserved.

Azure SQL

Famille de services de base de données cloud basés sur SQL Server

SQL Server sur machines virtuelles Azure

- Compatibilité garantie avec SQL Server local
- Le client gère tout : les mises à niveau du système d'exploitation, les mises à niveau logicielles, les sauvegardes, la réplication
- Paiement pour les coûts d'exécution de la VM et les licences logicielles du serveur, et non pas par base de données
- Idéal pour le cloud hybride ou la migration de configurations de bases de données locales complexes

Azure SQL Managed Instance

- Compatibilité à près de 100 % avec SQL Server local
- Sauvegardes automatiques, mises à jour correctives logicielles, supervision des bases de données et autres tâches de maintenance
- Utiliser une seule instance avec plusieurs bases de données ou plusieurs instances dans un pool avec des ressources partagées
- Idéal pour migrer la plupart des bases de données locales vers le cloud

Azure SQL Database

- Compatibilité des fonctionnalités principales de base de données avec SQL Server
- Sauvegardes automatiques, mises à jour correctives logicielles, supervision des bases de données et autres tâches de maintenance
- *Base de données unique* ou *pool élastique* pour partager dynamiquement des ressources entre plusieurs bases de données
- Idéal pour les nouvelles applications cloud

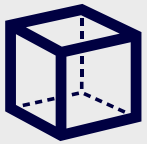
IaaS

PaaS

© Copyright Microsoft Corporation. Tous droits réservés.

Services Azure Database pour open source

Solutions managées Azure pour les SGBDR open source courants



Azure Database pour MySQL

- Implémentation PaaS de MySQL dans le cloud Azure, basée sur MySQL Community Edition
- Couramment utilisé dans les architectures d'application Linux, Apache, MySQL, PHP (LAMP)

Azure Database for MariaDB

- Implémentation du système de gestion de base de données MariaDB Community Edition adaptée pour s'exécuter dans Azure
- Compatibilité avec Oracle Database

Azure Database pour PostgreSQL

- Service de base de données relationnelle dans le cloud Microsoft basé sur le moteur de base de données PostgreSQL Community Edition
- Stockage relationnel et objet hybride

PaaS

© Copyright Microsoft Corporation. Tous droits réservés.

Contrôle des connaissances

Quelle option de déploiement offre la meilleure compatibilité lors de la migration d'une solution SQL Server locale existante ?

- ❑ Azure SQL Database (base de données unique)
- ❑ Azure SQL Database (pool élastique)
- ☑ Azure SQL Managed Instance

Parmi les affirmations suivantes, laquelle est vraie concernant Azure SQL Database ?

- ☑ La plupart des tâches de maintenance de base de données sont automatisées
- ❑ Vous devez acheter une licence SQL Server
- ❑ Il peut prendre en charge une seule base de données

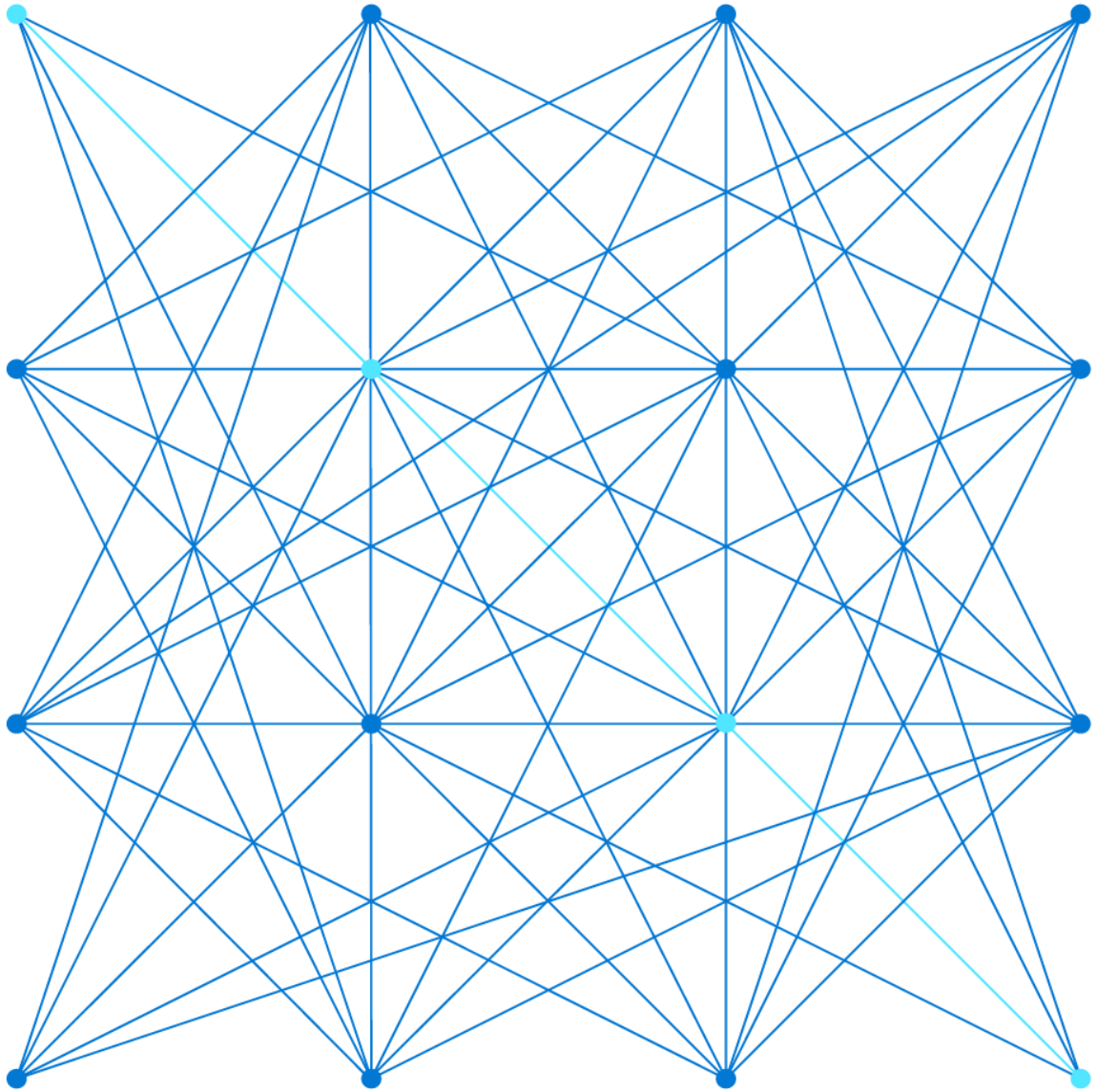
Quel service de base de données est l'option la plus simple pour migrer une application LAMP vers Azure ?

- ❑ Azure SQL Managed Instance
- ☑ Azure Database pour MySQL
- ❑ Azure Database pour PostgreSQL

© Copyright Microsoft Corporation. Tous droits réservés.

Microsoft Azure

Module 3 : Explorer les notions fondamentales des données non relationnelles dans Azure



© Copyright Microsoft Corporation. Tous droits réservés.

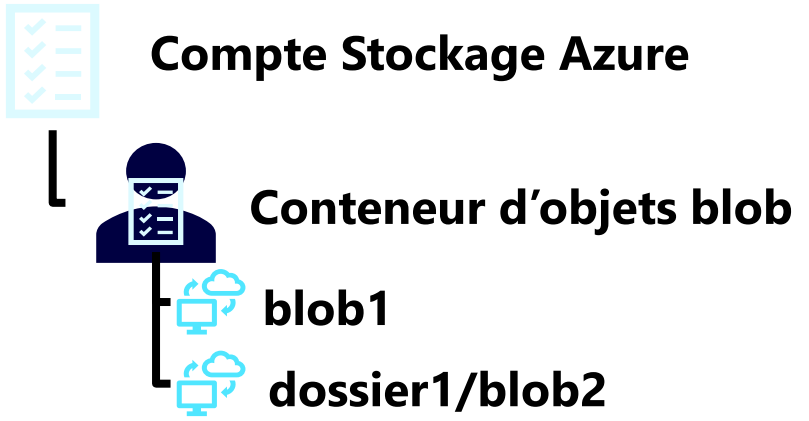
Stockage Blob Azure

Stockage pour les données en tant qu'objets blob dans des conteneurs

- Objets blob de blocs
  - Objets binaires volumineux et discrets, qui changent rarement
  - Les objets blob peuvent atteindre 4,7 To, composés de blocs pouvant atteindre 100 Mo et un objet blob peut contenir jusqu'à 50 000 blocs
- Objets blob de pages
  - Utilisés comme stockage de disque virtuel pour les machines virtuelles
  - Les objets blob peuvent atteindre 8 To, composés de pages d'octets de taille fixe de 512 octets
  - Optimisé pour les opérations de lecture et d'écriture aléatoires.
- Objets blob d'ajout
  - Objets blob de blocs utilisés pour optimiser les opérations d'ajout
  - Taille maximale légèrement supérieure à 195 Go : chaque bloc peut atteindre 4 Mo

Niveaux de stockage par objet blob

- **Chaud** : coût le plus élevé, latence la plus faible, pour les blobs fréquemment consultés, supports haute performance
- **Froid/ cool** : coût inférieur, latence plus élevée, données rarement consultées
- **Archive** : coût le plus bas, latence la plus élevée, les données ne doivent pas être perdues mais rarement consultables



© Copyright Microsoft Corporation. All rights reserved.

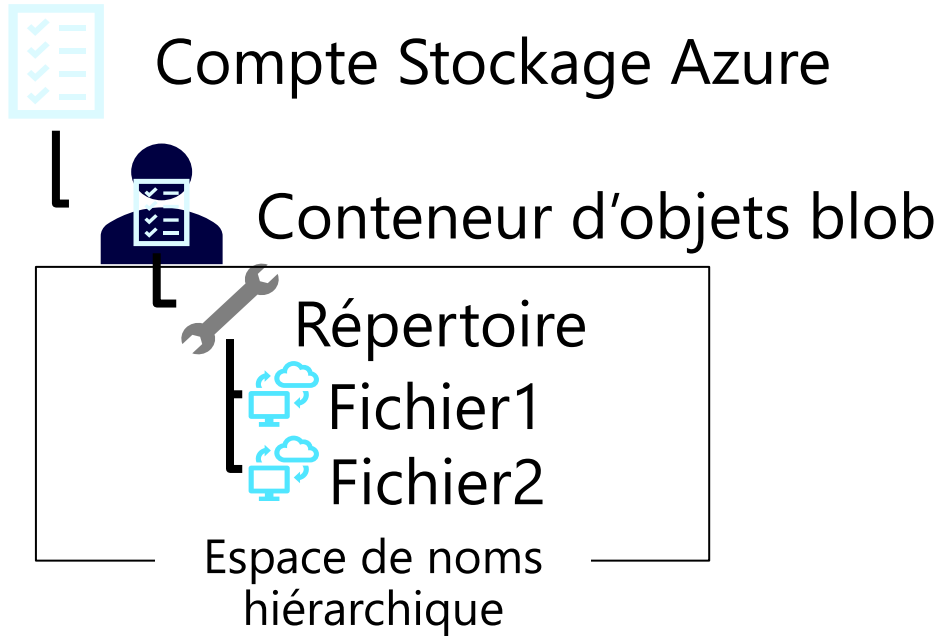
Azure Data Lake Store Gén. 2

Système de fichiers distribué basé sur le Stockage Blob

- Combine Azure Data Lake Store Gen 1 avec Stockage Blob Azure pour le stockage et l'analytique de fichiers à grande échelle
- Active la gestion et le contrôle d'accès au niveau du fichier et du répertoire
- Compatible avec les systèmes analytiques à grande échelle courants

Activé dans un compte de stockage Azure via l'option *Espace de noms hiérarchique*

- Définir lors de la création du compte
- Mettre à niveau un compte de stockage existant
  - Processus de mise à niveau unidirectionnel



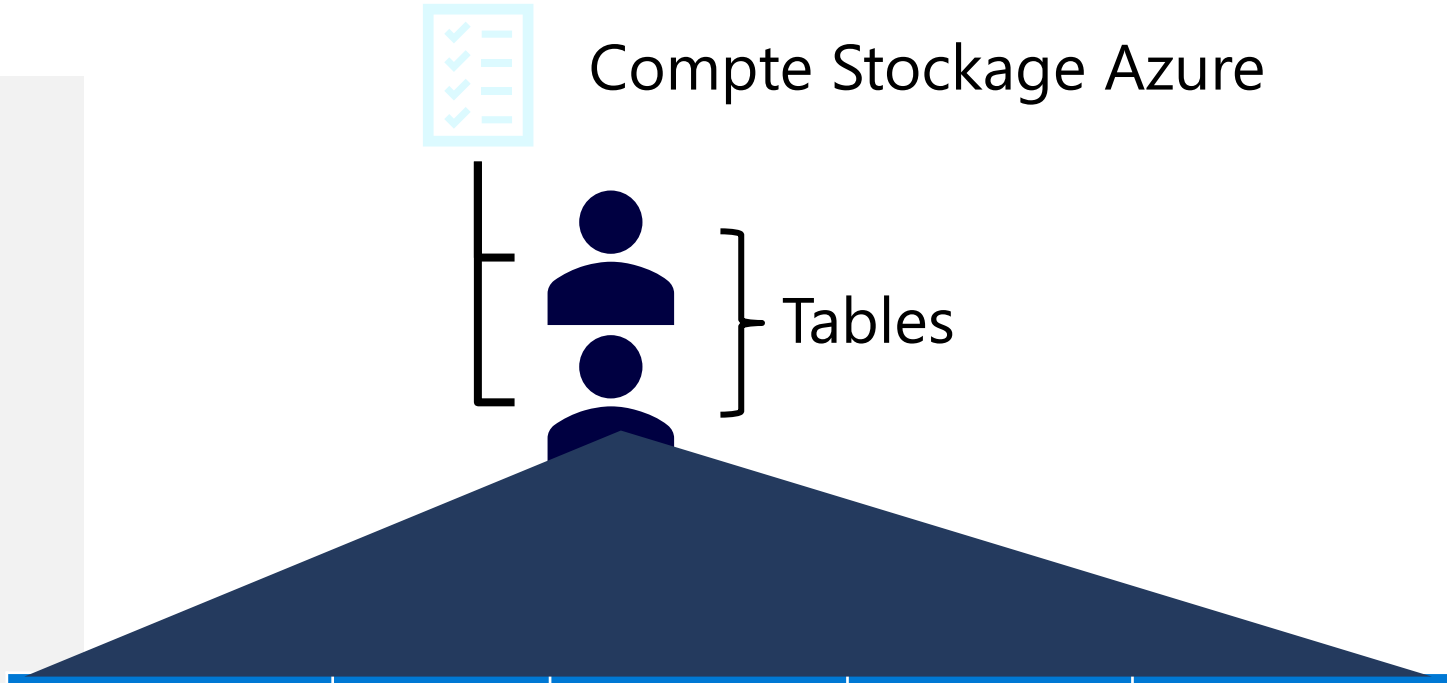
Le système de fichiers inclut des répertoires et des fichiers, et est compatible avec des systèmes d'analytique données à grande échelle tels que Hadoop, Databricks et Azure Synapse Analytics

© Copyright Microsoft Corporation. Tous droits réservés.

Stockage de table Azure

Stockage *clé-valeur* pour les données d'application

- Les tables se composent de colonnes *clé* et *valeur*
  - Clés de partition et de ligne
  - Colonnes de propriétés personnalisées pour les valeurs de données
    - Une *colonneTimestamp* est ajoutée automatiquement aux modifications de données de journal
- Les lignes sont regroupées en partitions pour améliorer les performances
- Les colonnes de propriété se voient affecter un type de données et peuvent contenir n'importe quelle valeur de ce type
- Les lignes n'ont pas besoin d'inclure les mêmes colonnes de propriété



PartitionKey	RowKey	Timestamp	Property1	Property2
1	123	2022-01-01	Une valeur	Une autre valeur
1	124	2022-01-01	Cette valeur	
2	125	2022-01-01		Cette valeur

© Copyright Microsoft Corporation. Tous droits réservés.

Contrôle des connaissances

Quels sont les éléments d'une clé de stockage Table Azure ?

- ❑ Nom de la table et nom de la colonne
- ☑ Clé de partition et clé de ligne
- ❑ Numéro de ligne

Que devez-vous faire d'un compte de stockage Azure existant afin de prendre en charge un lac de données pour Azure Synapse Analytics ?

- ❑ Ajouter un partage Azure Files
- ❑ Créer des tables de stockage Azure pour les données que vous souhaitez analyser
- ☑ Mettre à niveau le compte pour activer *l'espace de noms hiérarchique* et créer un conteneur de blobs

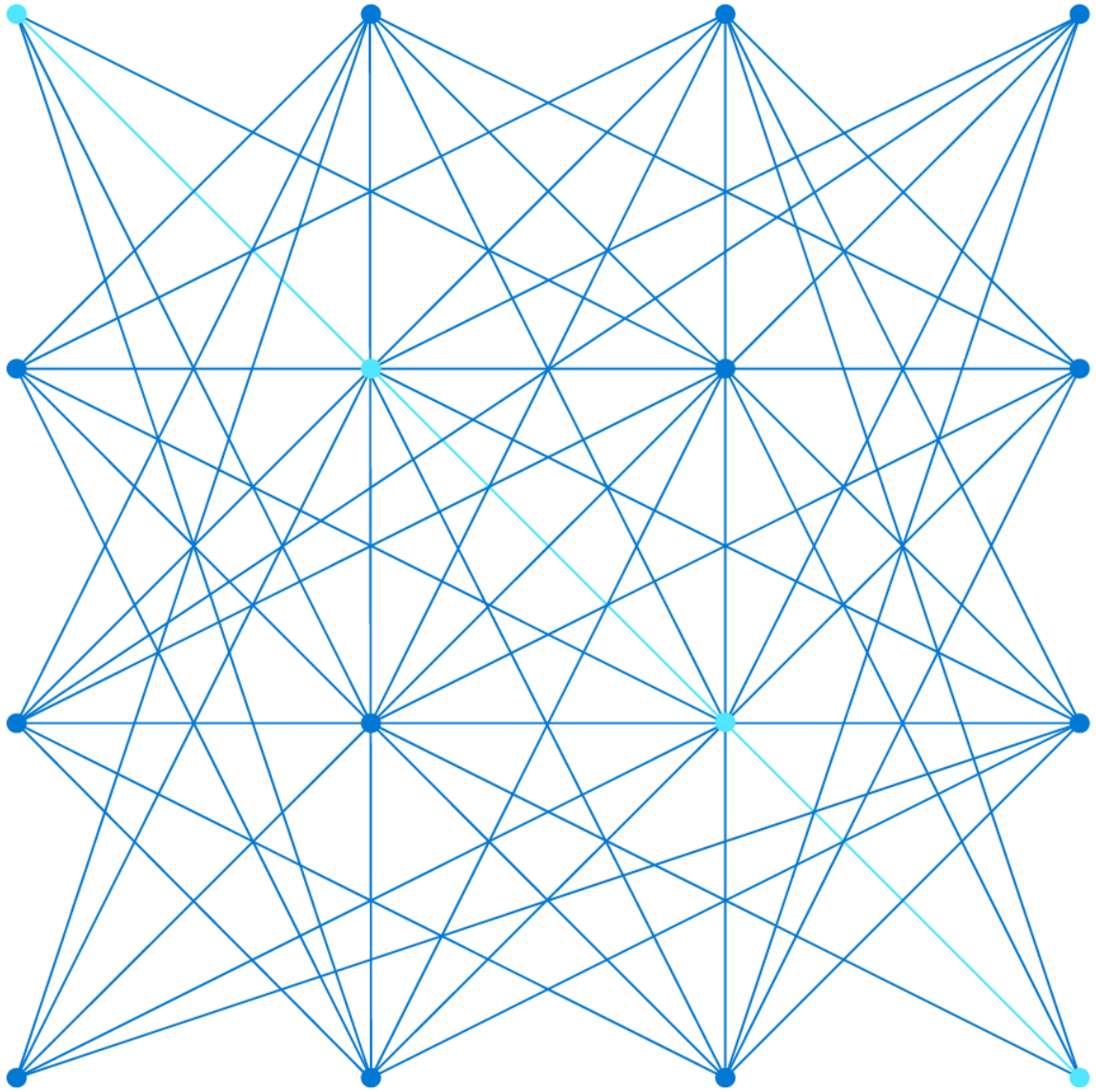
Quelle option Stockage Azure utiliser pour créer des partages de fichiers réseau basés sur le cloud ?

- ❑ Stockage Blob Azure
- ❑ Tables Azure
- ☑ Azure Files

© Copyright Microsoft Corporation. Tous droits réservés.

Microsoft Azure

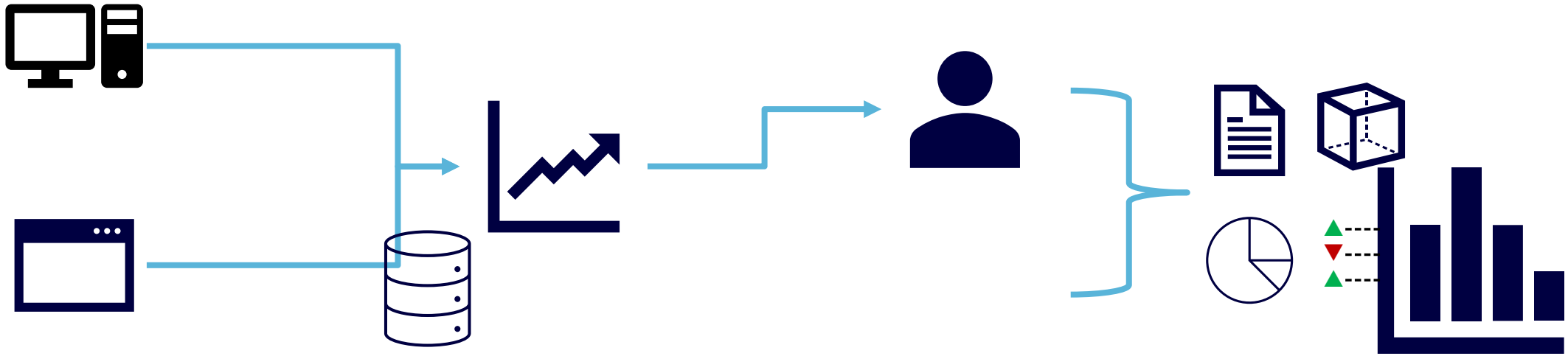
Module 4 : Explorer les bases de l'analytique données



© Copyright Microsoft Corporation. Tous droits réservés.

Qu’est-ce que l’entreposage de données à grande échelle ?

L’ingestion et le traitement des données	Magasin de données analytiques	Modèle de données analytiques	Visualisation des données



- *Orchestration ETL (extraction, transformation et chargement)* ou ELT (*extraction, chargement et transformation)*
- Traitement distribué pour nettoyer et restructurer les données à grande échelle
- Traitement des données par lots et en temps réel

- Stockage de données relationnelles dénormalisé dans un *entrepôt de données*
- Stockage de fichiers semi-structuré dans un *lac de données*

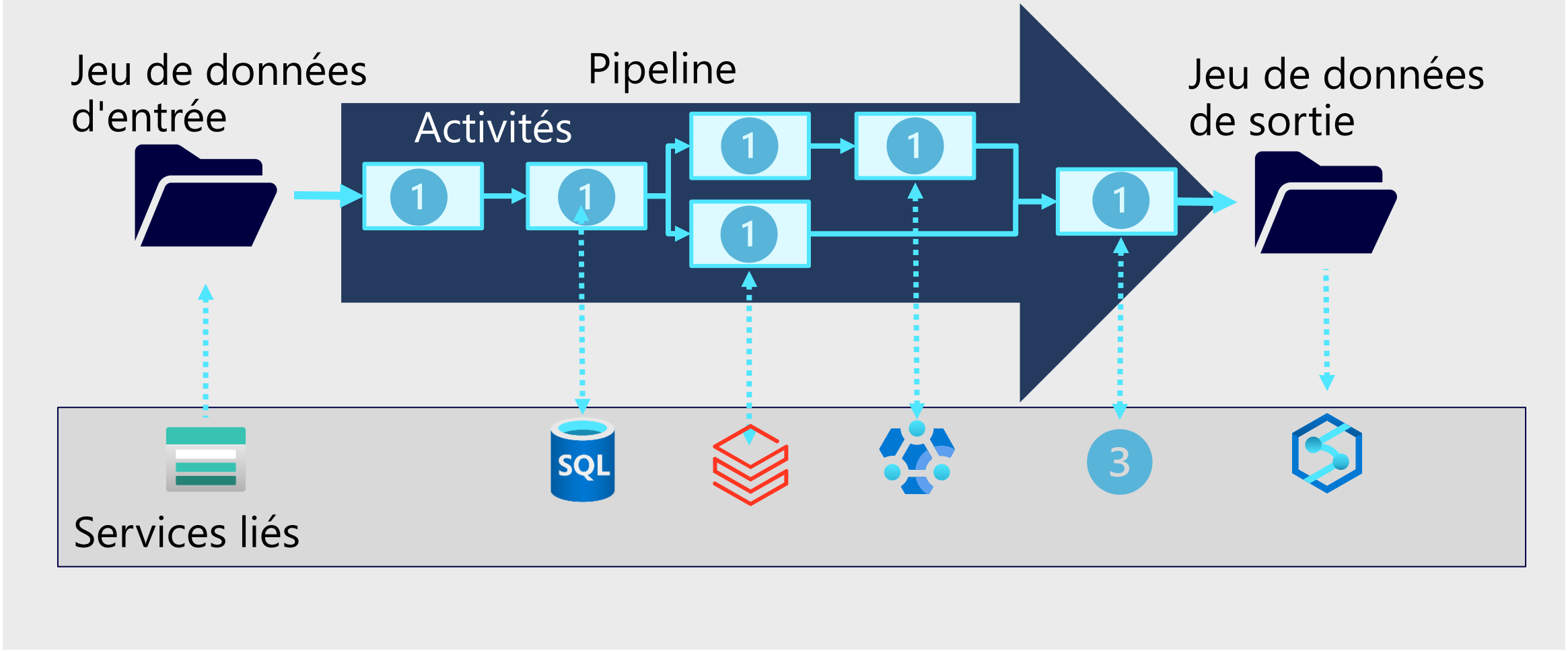
- Les agrégations, les niveaux de granularité
- Souvent sous la forme de *cubes* agrégés qui synthétisent les valeurs numériques sur une ou plusieurs *dimensions*

- Rapports
- Graphiques
- Tableaux de bord

© Copyright Microsoft Corporation. Tous droits réservés.

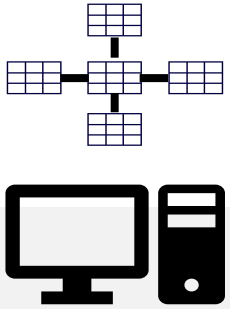
Pipelines d’ingestion et de traitement des données

Créer des pipelines dans **Azure Data Factory** ou **Azure Synapse Analytics**



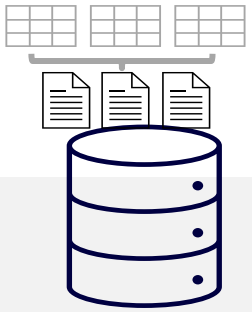
© Copyright Microsoft Corporation. Tous droits réservés.

Magasins de données analytiques



entrepôt de données

- Moteur de requête et magasin de bases de données relationnelles à grande échelle
- Les données sont *dénormalisées* pour l'optimisation des requêtes
  - Généralement sous forme de schéma en *étoile* ou en *flocon* des *faits* numériques pouvant être agrégés par *dimensions*



Data Lake

- Les fichiers de données sont stockés dans un système de fichiers distribué
- Des couches de stockage tabulaire peuvent être utilisées pour extraire des fichiers et fournir une interface relationnelle.
  - Utiliser des tables externes *PolyBase* ou créer une *base de données lake* dans Azure Synapse Analytics
  - Utiliser des tables de base de données et des points de terminaison SQL dans Azure Databricks
  - Utiliser Spark *Delta Lake* pour ajouter une sémantique de stockage relationnel et créer un *lakehouse de données* dans Azure Synapse Analytics, Azure Databricks et Azure HDInsight

© Copyright Microsoft Corporation. Tous droits réservés.

Choisir un service de magasin de données analytiques

Azure Synapse Analytics

- Solution unifiée pour l'entrepôt de données relationnelles et l'analytique du lac de données
- Traitement et interrogation évolutifs par le biais de plusieurs runtimes analytiques
  - SQL Synapse
  - Apache Spark
  - Synapse Data Explorer
- Expérience interactive dans Azure Synapse Studio
- Intégration de pipelines prédéfinis pour l'ingestion et le traitement des données

Utilisation pour une seule solution analytique à grande échelle unifiée sur Azure

Azure Databricks

- Implémentation basée sur Azure de la plateforme d'analytique cloud Databricks
- Interrogation Spark et SQL évolutives pour l'analytique de lac de données
- Expérience interactive dans l'espace de travail Azure Databricks
- Utiliser Azure Data Factory pour implémenter des pipelines d'ingestion et de traitement des données

Utilisation pour tirer parti des compétences Databricks et pour la portabilité cloud

Azure HDInsight

- Implémentation basée sur Azure de frameworks « Big Data » Apache courants basés sur un lac de données
  - Hadoop – Interroger des fichiers de lac de données à l'aide de tables Hive
  - Spark – Utiliser des API Spark pour interroger des données et extraire le stockage de fichiers sous-jacent sous forme de tables
  - Kafka – Traitement des événements en temps réel
  - Storm – Traitement des flux
  - HBase – Magasin de données NoSQL

Utilisation pour prendre en charge plusieurs plateformes open source

© Copyright Microsoft Corporation. Tous droits réservés.

Contrôle des connaissances

Quels services Azure pouvez-vous utiliser pour créer un pipeline visant à ingérer et traiter des données ?

- ☐ Azure SQL Database et Azure Cosmos DB
- ☑ Azure Synapse Analytics et Azure Data Factory
- ☐ Azure HDInsight et Azure Databricks

Que devez-vous définir pour implémenter un pipeline qui lit les données à partir du Stockage Blob Azure ?

- ☑ Un service lié pour votre compte Stockage Blob Azure
- ☐ Un pool SQL dédié dans votre espace de travail Azure Synapse Analytics
- ☐ Un cluster Azure HDInsight dans votre abonnement

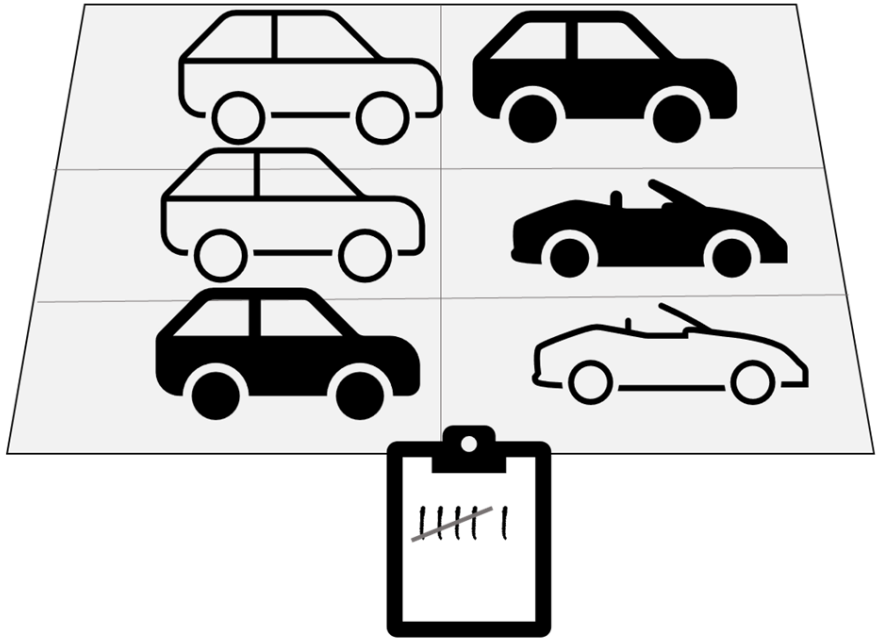
Quel moteur de traitement distribué open source Azure Synapse Analytics comprend-il ?

- ☐ Apache Hadoop
- ☑ Apache Spark
- ☐ Apache Storm

© Copyright Microsoft Corporation. All rights reserved.

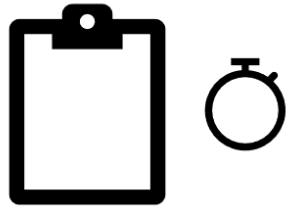
Traitement par lots et par flux

Traitement par lots



Les données sont collectées et traitées à intervalles réguliers

Traitement des flux de données

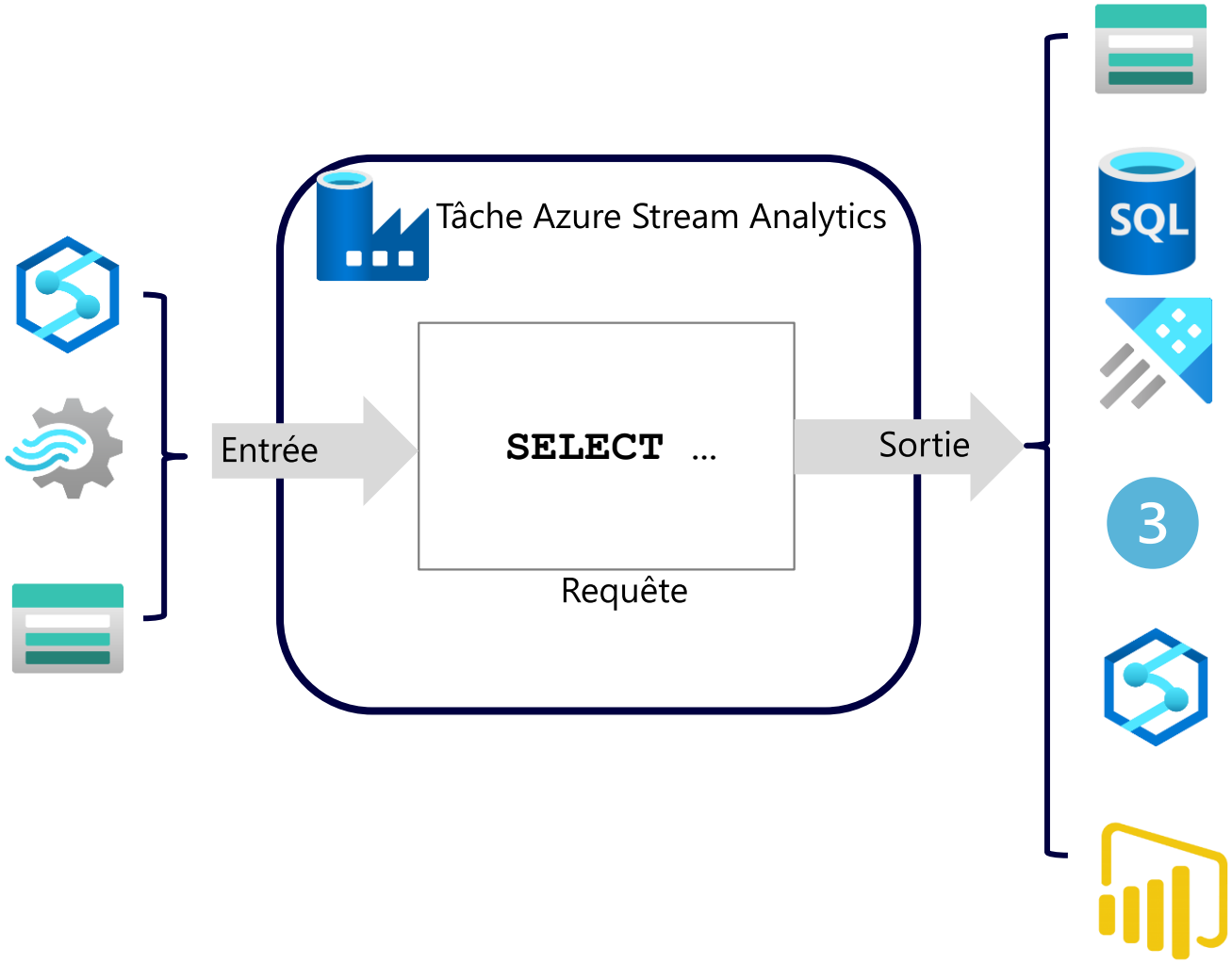


Les données sont traitées en temps (quasi) réel dès qu'elles arrivent

© Copyright Microsoft Corporation. Tous droits réservés.

Traitement des données en temps réel avec Azure Stream Analytics

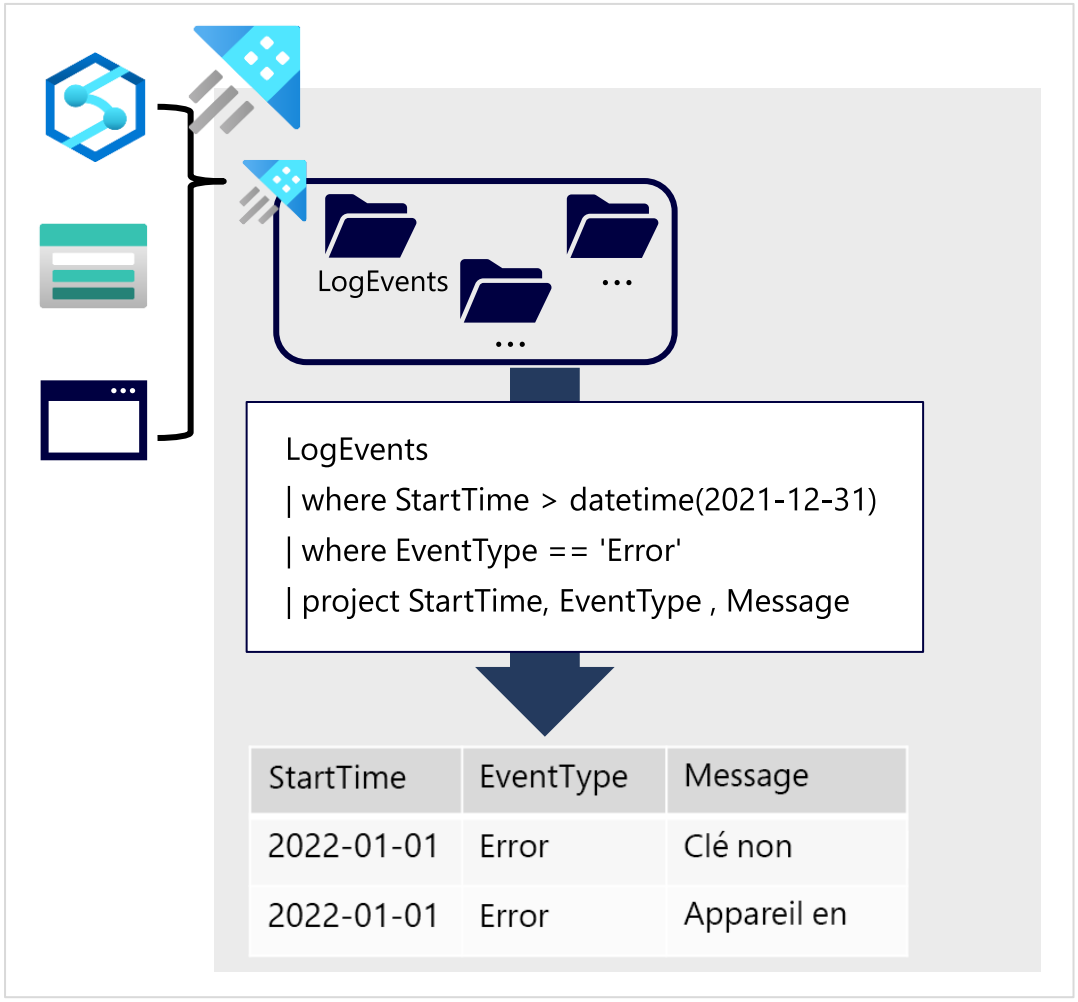
- Créer un *travail* Azure Stream Analytics individuel ou un *cluster* Azure Stream Analytics
- Ingérer des données à partir d'une *entrée*, comme :
  - Hubs d'événements Azure
  - Azure IoT Hub
  - Stockage Blob Azure
  - ...
- Traiter des données avec une *requête* perpétuelle
- Envoyer des résultats à une*sortie*, comme :
  - Stockage Blob Azure
  - Azure SQL Database
  - Azure Synapse Analytics
  - Fonction Azure
  - Azure Event Hubs
  - Power BI
  - ...



© Copyright Microsoft Corporation. Tous droits réservés.

Analyse des journaux et de la télémétrie en temps réel avec Azure Data Explorer

- Service évolutif à débit élevé pour les données de traitement par lots et de streaming
  - **Service dédié** Azure Data Explorer
  - **Runtime Azure Synapse Data Explorer** dans Azure Synapse Analytics
- Les données sont ingérées à partir de sources de traitement par lots et de streaming dans les tables d'une base de données
- Les tables peuvent être interrogées à l'aide du *Langage de requête Kusto* (KQL) :
  - Syntaxe intuitive pour les requêtes en lecture seule
  - Optimisé pour les données de télémétrie brutes et de série chronologique



© Copyright Microsoft Corporation. Tous droits réservés.

Contrôle des connaissances

Quelle est la définition correcte du *traitement du flux* ?

- [x] Les données sont traitées continuellement à mesure que de nouveaux enregistrements de données arrivent
- [ ] Les données sont collectées dans un magasin temporaire et tous les enregistrements sont traités ensemble en tant que lot
- [ ] Les données sont incomplètes et ne peuvent pas être analysées

Quel service utiliseriez-vous pour capturer continuellement des données à partir d'un IoT Hub, les agréger sur des périodes temporelles et stocker les résultats dans Azure SQL Database ?

- [ ] Azure Cosmos DB
- [x] Azure Stream Analytics
- [ ] Stockage Azure

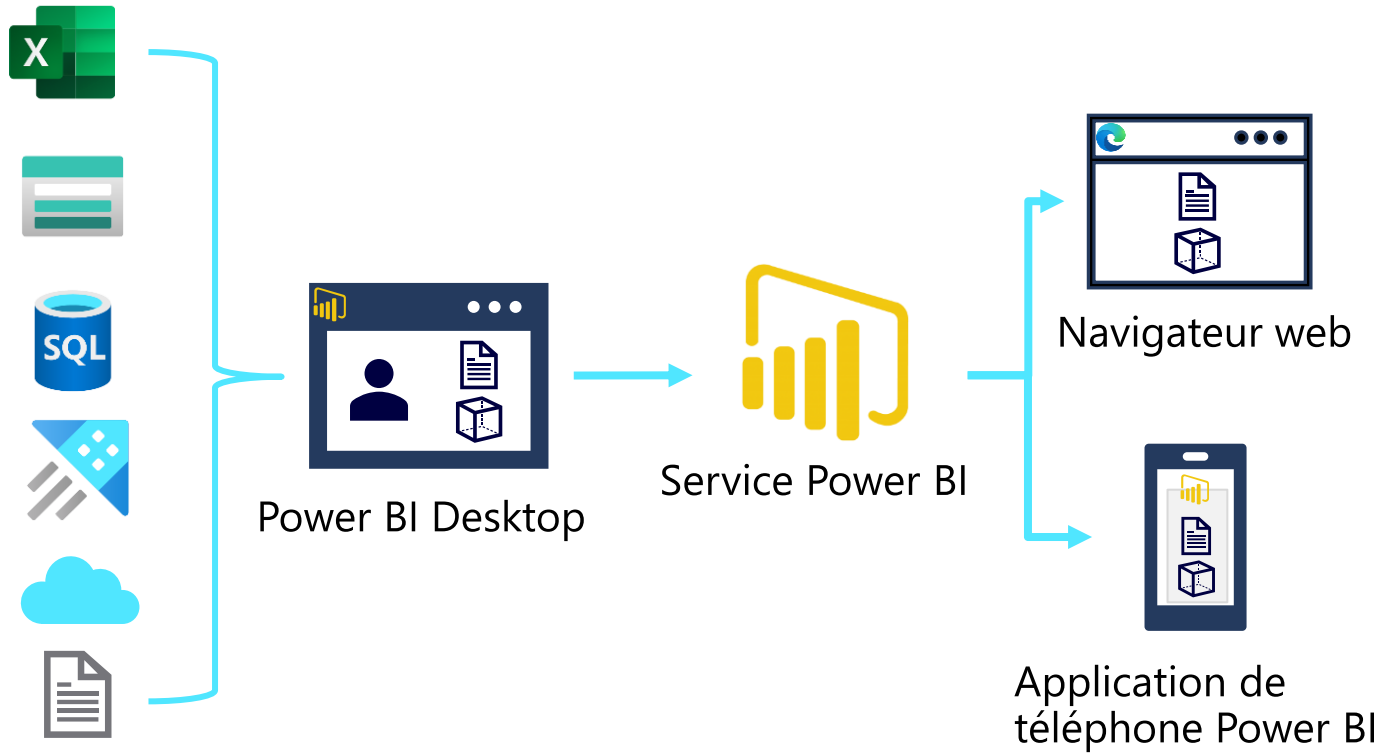
Quelle langue utiliseriez-vous pour interroger les données du journal en temps réel dans Azure Synapse Data Explorer ?

- [ ] SQL
- [ ] Python
- [x] KQL

© Copyright Microsoft Corporation. All rights reserved.

Présentation de la visualisation des données avec Power BI

- Commencer par Power BI Desktop
  - Importer des données depuis une ou plusieurs sources
  - Définir un modèle de données
  - Créer des visualisations dans un rapport
- Publier sur le service Power BI
  - Planifier l'actualisation des données
  - Créer des tableaux de bord et des applications
  - Partager avec d'autres utilisateurs
- Interagir avec des rapports publiés
  - un navigateur Web
  - Application de téléphone Power BI



© Copyright Microsoft Corporation. Tous droits réservés.

Modélisation des données analytiques

Client *(dimension)*

Clé	Nom	Adresse	City
1	Joe	1 Main St.	Seattle
2	Samir	123 Elm Pl.	New York
3	Alice	2 High St.	Seattle

Produit *(dimension)*

Clé	Nom	Category
1	Marteau	Outils
2	Tournevis	Outils
3	Clé	Outils
4	Bolts	Matériel

Ventes *(fait)*

Clé	TimeKey	ProductKey	CustomerKey	Quantité	Chiffre d'affaires
1	01012022	1	1	1	2.99
2	01012022	2	1	2	6,98
3	02012022	1	2	2	5,98

Mesures

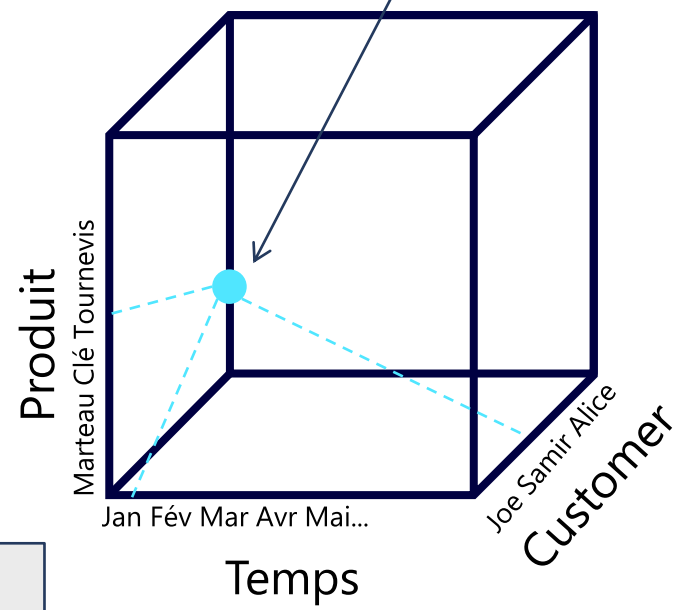
Temps *(dimension)*

Clé	Year	Month	Jour	WeekDay
01012022	2022	Jan	1	Sam
02012022	2022	Jan	2	Dim

Hierarchy

∑

Chiffre d'affaires total des ventes de clés à Samir en janvier



Le modèle agrège les mesures au niveau de chaque hiérarchie

Year	Month	Jour	Chiffre d'affaires
2022			8221,48
	Jan		574,86
		1	9,97
		2	5,98
		...	...

© Copyright Microsoft Corporation. Tous droits réservés.

Visualisations de données courantes dans les rapports

Tableaux et texte

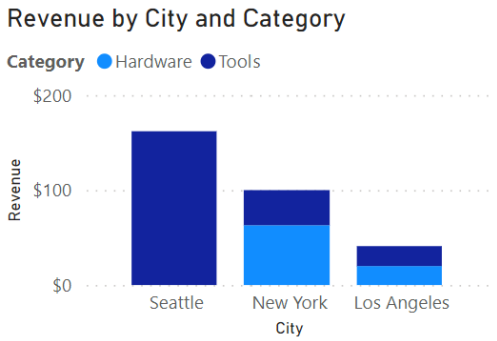
Product Sales

Name	Quantity
Bolts	2
Hammer	4
Nails	1
Screwdriver	2
Screws	2
Wrench	4
Total	**15**

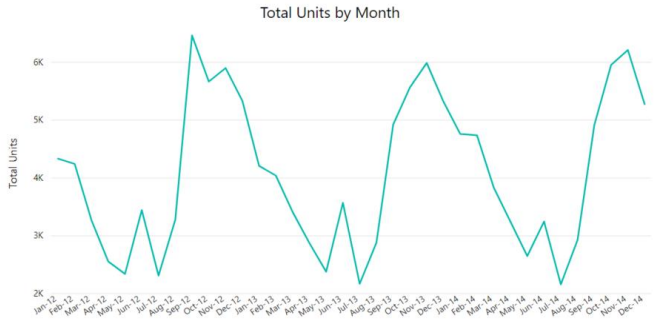
$302.91

Revenue

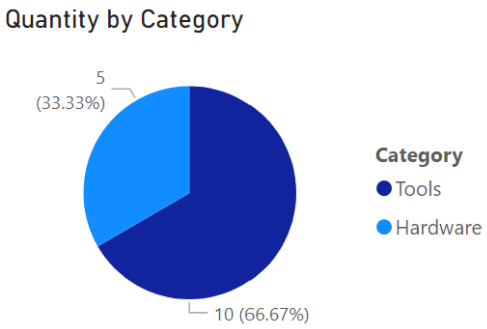
Graphique à barres ou histogramme



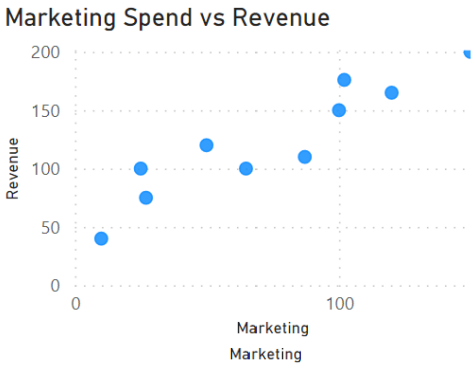
Graphique en courbes



Graphique en secteurs



nuage de points



Mappage

© Copyright Microsoft Corporation. Tous droits réservés.

Contrôle des connaissances

Quel outil devez-vous utiliser pour importer des données de plusieurs sources de données et créer un rapport ?

- [x] Power BI Desktop
- [ ] Application de téléphone Power BI
- [ ] Azure Data Factory

Que devez-vous définir dans votre modèle de données pour permettre une analyse ascendante/descendante ?

- [ ] Une mesure
- [x] Une hiérarchie
- [ ] Une relation

Quel type de visualisation devriez-vous utiliser pour analyser les taux de réussite de plusieurs examens dans le temps ?

- [ ] Un graphique à secteurs
- [ ] Un nuage de points
- [x] Un graphique en courbes

© Copyright Microsoft Corporation. All rights reserved.