**Atelier 2 : Conception d’une architecture de données modernes avec Cloud Computing**

**Partie 1 : Recherche et Analyse des Technologies Cloud**

**Présentation des Principales Plateformes Cloud**

Les trois plateformes de Cloud Computing dominantes – **Amazon Web Services (AWS)**, **Microsoft Azure**, et **Google Cloud Platform (GCP)** – offrent des services complets pour répondre aux besoins de stockage, traitement des données, analyse en temps réel, et gestion des bases de données. Voici un aperçu de chaque plateforme et des services clés pertinents pour les scénarios proposés :

**1. Amazon Web Services (AWS)**

* **Stockage** :
  + **Amazon S3 (Simple Storage Service)** : Solution de stockage d’objets flexible et scalable, idéale pour les Data Lakes et les sauvegardes.
  + **Amazon RDS (Relational Database Service)** : Pour les bases de données relationnelles (ex : MySQL, PostgreSQL, SQL Server).
  + **Amazon DynamoDB** : Base de données NoSQL gérée, adaptée aux applications nécessitant une faible latence et un haut débit (par exemple, applications e-commerce ou IoT).
* **Traitement des Données** :
  + **AWS Glue** : Service de traitement ETL (Extract, Transform, Load) sans serveur, utilisé pour préparer et transformer les données.
  + **Amazon EMR (Elastic MapReduce)** : Permet d’exécuter des clusters Hadoop, Spark pour le traitement de grandes quantités de données.
  + **AWS Lambda** : Service de calcul sans serveur permettant de déclencher des traitements en réponse à des événements.
* **Analyse en Temps Réel** :
  + **Amazon Kinesis** : Plateforme de streaming qui permet la collecte et le traitement en temps réel des données de streaming (comme les transactions e-commerce ou les données de capteurs IoT).
  + **Amazon QuickSight** : Service de visualisation de données pour créer des tableaux de bord interactifs et des rapports.
* **Bases de Données** :
  + **Amazon Aurora** : Base de données relationnelle compatible avec MySQL et PostgreSQL, optimisée pour le Cloud.
  + **Amazon Neptune** : Base de données orientée graphe, idéale pour les applications de recommandation.

**2. Microsoft Azure**

* **Stockage** :
  + **Azure Blob Storage** : Solution de stockage d’objets, idéale pour les applications nécessitant un stockage de grandes quantités de données non structurées.
  + **Azure SQL Database** : Base de données relationnelle SQL entièrement managée pour les applications nécessitant des bases de données robustes et scalables.
  + **Azure Cosmos DB** : Base de données NoSQL multi-modèles, distribuée mondialement, idéale pour les applications nécessitant une faible latence et une scalabilité rapide.
* **Traitement des Données** :
  + **Azure Data Factory** : Service ETL qui permet de créer, planifier, et orchestrer des flux de données.
  + **Azure Synapse Analytics** : Plateforme d’analyse pour des workloads analytiques de grande envergure ; elle intègre des capacités de traitement en batch et en streaming.
  + **Azure Functions** : Permet d’exécuter des fonctions sans serveur pour des tâches spécifiques déclenchées par des événements (par exemple, traitement de données en temps réel).
* **Analyse en Temps Réel** :
  + **Azure Stream Analytics** : Service permettant le traitement en temps réel des flux de données de capteurs, des logs ou des événements.
  + **Power BI** : Outil de visualisation de données qui permet de créer des rapports interactifs et des tableaux de bord.
* **Bases de Données** :
  + **Azure SQL Database** : Service de bases de données relationnelles pour des applications de haute disponibilité.
  + **Azure Cosmos DB** : NoSQL avec support multi-modèle et multi-région, permettant une flexibilité pour des applications nécessitant une disponibilité mondiale.

**3. Google Cloud Platform (GCP)**

* **Stockage** :
  + **Google Cloud Storage** : Solution de stockage d’objets avec une grande capacité, idéale pour les Data Lakes.
  + **Cloud SQL** : Base de données relationnelle gérée, compatible avec MySQL, PostgreSQL, et SQL Server.
  + **Cloud Firestore** : Base de données NoSQL orientée document, adaptée aux applications nécessitant des bases de données flexibles et scalables.
* **Traitement des Données** :
  + **Google Dataflow** : Service de traitement de données sans serveur pour des flux ETL et des traitements de données en streaming (idéal pour les données IoT).
  + **Google BigQuery** : Entrepôt de données analytiques sans serveur pour les requêtes à grande échelle, avec des fonctionnalités ML intégrées.
  + **Google Cloud Functions** : Exécution de fonctions sans serveur, utile pour le déclenchement d’événements et de tâches spécifiques.
* **Analyse en Temps Réel** :
  + **Google Cloud Pub/Sub** : Service de messagerie pour la diffusion et la collecte en temps réel des données de streaming.
  + **Google Data Studio** : Outil de visualisation de données qui permet de créer des rapports personnalisés et des tableaux de bord interactifs.
* **Bases de Données** :
  + **Bigtable** : Base de données NoSQL gérée, idéale pour le stockage et l’analyse de données IoT à grande échelle.
  + **Spanner** : Base de données relationnelle globale offrant une scalabilité horizontale avec un support pour SQL, utilisée dans des applications nécessitant une faible latence à l’échelle mondiale.

**Analyse des Différentes Solutions**

* **Stockage des Données** :
  + Les solutions de stockage d'objets (S3, Blob Storage, Cloud Storage) sont idéales pour les Data Lakes, car elles permettent de stocker des données structurées et non structurées de manière flexible.
  + Les bases de données SQL (RDS, SQL Database, Cloud SQL) sont adaptées aux applications nécessitant une structure relationnelle stricte (comme la gestion des commandes).
  + Les bases de données NoSQL (DynamoDB, Cosmos DB, Firestore) sont utiles pour les données non structurées ou semi-structurées, offrant une scalabilité rapide.
* **Traitement des Données** :
  + Les services de traitement par lots (EMR, Data Factory, Dataflow) permettent de traiter des volumes importants de données à des intervalles réguliers.
  + Les services de traitement en temps réel (Kinesis, Stream Analytics, Pub/Sub) sont essentiels pour des applications nécessitant une analyse instantanée, par exemple, pour le suivi des transactions ou l'analyse de données IoT.
* **Analyse et Visualisation des Données** :
  + Les solutions BI (QuickSight, Power BI, Data Studio) permettent de créer des rapports visuels et d'analyser les données à l'aide de tableaux de bord.
  + Les plateformes d’analyse avancée comme BigQuery et Synapse Analytics sont idéales pour traiter de grandes quantités de données rapidement.
* **Bases de Données Relationnelles et Non-Relationnelles** :
  + Les bases de données relationnelles sont recommandées pour les applications nécessitant une intégrité transactionnelle (par exemple, la gestion des commandes e-commerce).
  + Les bases de données NoSQL, quant à elles, offrent une flexibilité pour les applications nécessitant une scalabilité rapide (comme les données d’utilisateurs sur une plateforme de streaming).

**2. Étude des Architectures de Données Modernes**

Les architectures de données modernes se composent de plusieurs couches, chacune jouant un rôle essentiel dans le traitement et l'analyse de données de masse. Voici un aperçu des principales architectures :

* **Data Lakes** :
  + **Définition** : Un Data Lake est un espace de stockage centralisé qui peut stocker des données structurées, semi-structurées et non structurées en grande quantité. Les Data Lakes permettent de stocker les données brutes dans leur format d'origine, offrant ainsi une flexibilité pour l'analyse future.
  + **Utilisation** : Utilisé pour des cas d'usage nécessitant de grandes quantités de données brutes provenant de diverses sources (ex. logs, capteurs, fichiers multimédias) qui peuvent être analysées ultérieurement.
  + **Avantages** : Évolutivité, stockage économique des données en vrac, compatibilité avec des outils d’analyse avancée (machine learning, big data).
  + **Exemples de services** : *Amazon S3*, *Azure Data Lake Storage*, *Google Cloud Storage*.
* **Data Warehouses** :
  + **Définition** : Contrairement aux Data Lakes, un Data Warehouse est structuré pour organiser et stocker des données traitées et prêtes pour l’analyse. Il est optimisé pour exécuter des requêtes analytiques rapides et est principalement utilisé pour des analyses commerciales et des rapports.
  + **Utilisation** : Idéal pour les données structurées et les analyses de tendances. Utilisé pour des rapports BI et des analyses en temps quasi réel.
  + **Avantages** : Performance pour des requêtes analytiques, structuration pour des données propres, cohérentes et prêtes pour l’analyse.
  + **Exemples de services** : *Amazon Redshift*, *Azure Synapse Analytics*, *BigQuery*.
* **Pipelines ETL/ELT modernes** :
  + **Définition** : Les pipelines ETL (Extract, Transform, Load) et ELT (Extract, Load, Transform) sont des processus de transfert de données depuis des sources variées vers un système cible (Data Lake ou Data Warehouse). Dans un environnement Cloud, les pipelines ETL/ELT sont utilisés pour automatiser le flux de données.
  + **Utilisation** : Essentiel pour préparer les données de manière structurée pour les analyses. ETL est plus adapté pour les Data Warehouses, tandis que ELT est souvent préféré dans les Data Lakes pour transformer les données après leur chargement.
  + **Avantages** : Automatisation des flux de données, transformations adaptables, meilleure gestion du traitement des données en temps réel.
  + **Exemples de services** : *AWS Glue*, *Azure Data Factory*, *Google Dataflow*.

Image   
Une image contenant texte, capture d’écran, Police, logiciel

Description générée automatiquement  
  
Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement