**OpenHack - Entreposage de données moderne**

**Vue d’ensemble**

**Cet OpenHack permet aux participants** de développer, implémenter et rendre opérationnels des pipelines ETL pour une solution d’entrepôt de données multi-sources sur Microsoft Azure.

**Cet OpenHack simule un scénario du mondé réel** où une société de vente de DVD en ligne dont les données proviennent d’une multitude de sources disparates, mais doivent être stockées en un seul endroit, interprétées, puis utilisées pour alimenter une grande variété de systèmes en aval.

**Pendant le hack, les participants se concentreront sur :**

1. Ingérer et sécuriser systématiquement les données provenant de sources multiples.
2. Transformer les données pour les adapter au schéma requis par l’entreprise et surveiller le flux de données avec des niveaux de tests DevOps.

**À la fin de l’OpenHack, les participants auront élaboré une solution technique** qui est un entrepôt de données moderne entièrement opérationnel avec le pipeline CI/CD correspondant qui prend en compte la gestion des données répondant aux exigences de consommation de données de qualité supérieure comme la fiabilité, l’évolutivité et la maintenabilité.

**Technologies**

[Azure Data Lake Storage](https://docs.microsoft.com/fr-fr/azure/storage/blobs/data-lake-storage-introduction), [Azure Data Factory](https://docs.microsoft.com/fr-fr/azure/data-factory/), [Azure Databricks](https://docs.microsoft.com/fr-fr/azure/databricks/scenarios/what-is-azure-databricks), [Azure DevOps](https://azure.microsoft.com/fr-fr/services/devops/), [Azure Synapse Analytics](https://azure.microsoft.com/fr-fr/services/synapse-analytics/)

**Prérequis**

**Connaissances prérequises**

Pour réussir et tirer le meilleur parti de cet OpenHack, les participants doivent avoir une connaissance des structures et des concepts des bases de données relationnelles (par exemple, les tables, les jointures, SQL) et une expérience de SSIS ou de langages de programmation comme Scala ou Python. Une expérience préalable de la création de pipelines ETL, de la gestion du contrôle des sources, des tests automatisés et de l’automatisation des builds et des mises en production vous aidera à progresser plus rapidement.

Connaissance requise des [Principes de base d’Azure](https://docs.microsoft.com/fr-fr/learn/paths/azure-fundamentals/).

**Conditions préalables spécifiques au langage**

* Langages de programmation comme Scala ou Python.

**Outils prérequis**

Pour éviter tout retard dans le téléchargement ou l’installation des outils, préparez les éléments suivants avant le OpenHack :

* Un ordinateur portable moderne sous Windows 10 (1703 ou version ultérieure), macOS X (10.12 ou version ultérieure) ou l’une de [ces variantes Ubuntu](https://github.com/Azure/azure-functions-core-tools#linux)
* Installez l’environnement de développement intégré (IDE) de votre choix, par exemple [Visual Studio](https://visualstudio.microsoft.com/vs/community/)/[Visual Studio Code](https://code.visualstudio.com/docs#download)/[Eclipse](https://www.eclipse.org/)/[IntelliJ](https://www.jetbrains.com/idea/)
* Téléchargez la dernière version d’[Azure CLI](https://docs.microsoft.com/fr-fr/cli/azure/install-azure-cli?view=azure-cli-latest).
* [Outils de base de données SQL Server](https://docs.microsoft.com/fr-fr/sql/azure-data-studio/download-azure-data-studio?view=sql-server-ver15) (Azure Data Studio/SSMS)
* [SQL Server Data Tools](https://docs.microsoft.com/fr-fr/sql/ssdt/download-sql-server-data-tools-ssdt?view=sql-server-ver15) (y compris les outils de BI) - Si vous utilisez Visual Studio comme IDE

**Configuration de l’environnement de développement**

Aucun.

**Liens et ressources**

Comme vous travaillerez en équipe, de bonnes connaissances en programmation en binôme sont utiles. Nous vous recommandons de lire ce qui suit :

* [Programmation en binôme](https://www.agilealliance.org/glossary/pairing/)

**Recommandations post-apprentissage**

* [Implémenter un entrepôt de données avec Azure Synapse Analytics](https://docs.microsoft.com/fr-fr/learn/paths/implement-sql-data-warehouse/)
* [Traitement des données à grande échelle avec Azure Data Lake Storage Gen2](https://docs.microsoft.com/fr-fr/learn/paths/data-processing-with-azure-adls/)
* [Services cloud de base - Options de stockage de données Azure](https://docs.microsoft.com/fr-fr/learn/modules/intro-to-data-in-azure/)
* [Azure pour un ingénieur de données](https://docs.microsoft.com/fr-fr/learn/paths/azure-for-the-data-engineer/)
* [Appliquer l’ingénierie des données avec Azure Databrick](https://docs.microsoft.com/fr-fr/learn/paths/data-engineer-azure-databricks/)
* [Concevoir une plateforme de données dans Azure](https://docs.microsoft.com/fr-fr/learn/paths/architect-data-platform/)

**Problématiques**

**Défi 1 : Sélectionner et approvisionner le stockage pour un lac de données d’entreprise**

Dans ce défi, vous allez établir un lac de données d’entreprise.

Objectifs d’apprentissage :

* Comparer et opposer les offres de stockage Azure
* Approvisionner le service de stockage Azure sélectionné

**Défi 2 : Réception de données à partir de sources cloud**

Le deuxième défi consiste à extraire les premières données Southridge des bases de données Azure SQL et d’une collection Cosmos DB.

Objectifs d’apprentissage :

* Orchestrer l’ingestion des données de plusieurs sources cloud dans un seul magasin cloud
* Garantir la protection des données client spécifiques à tout moment en tirant parti du jeu de technologies actuel et de l’architecture de la solution

**Défi 3 : Extraire des données d’un emplacement local et établir un contrôle de code source**

Dans ce défi, vous allez incorporer des sources de données supplémentaires dans un nouveau lac de données. Alors que les premières données ont été extraites de sources basées sur le cloud, les données de ce défi viennent de magasins de données locaux.

Objectifs d’apprentissage :

* Orchestrer l’ingestion des données, en particulier des solutions « locales » gérées
* Implémenter un référentiel de contrôle de code source basé sur le cloud pour la solution développée

**Défi 4 : Transformer et normaliser les données dans le lac et établir des stratégies de branche**

Dans ce défi, vous découvrirez (si ce n’est pas déjà le cas !) que les données provenant de ces sources sont de types et de formats différents ; c’est à l’équipe de les prétraiter pour qu’elles soient cohérentes pour les consommateurs en aval.

Objectifs d’apprentissage :

* Transformer des données en schéma normalisé pour la consommation en aval
* Créer des stratégies pour faire en sorte que toutes les modifications ultérieures tirent parti d’un processus de révision approprié

**Défi 5 : Remplir un entrepôt de données et mettre en œuvre des tests unitaires**

Dans ce défi, vous devrez répondre aux besoins de l’entreprise en matière de rapports ! Un rapport Power BI sera fourni à l’équipe, mais c’est à elle de créer et d’alimenter le schéma en étoile auquel il se connecte. La définition du schéma cible est également fournie. Du point de vue de DevOps, l’équipe devra établir son approche de test unitaire.

Objectifs d’apprentissage :

* Transformez les données des différents systèmes sources en un schéma commun d’entrepôt de données pour permettre la génération de rapports spécifiques demandés par l’entreprise
* Orchestrez le flux de données dans l’entrepôt de données de manière automatisée
* Créer des tests unitaires sur des composants principaux du pipeline de données
* Intégrer des tests automatisés dans le processus de révision du code

**Défi 6 : Charges de données différentielles et télémétrie**

Dans ce défi, vous devrez répondre aux besoins récurrents de l’entreprise. Les données de chaque nouvelle journée d’activité doivent être ajoutées au lac de données, mais il serait inefficace de traiter de manière répétée toutes les données historiques. Vous devrez mettre en œuvre une charge incrémentielle et établir une solution de journalisation et de télémétrie grâce à laquelle vous pourrez surveiller la quantité de données nouvellement incorporées.

Objectifs d’apprentissage :

* Modifier la solution pour inclure des chargements de données différentiels en plus du chargement en masse original
* Automatiser le chargement et le traitement des données pour une exécution quotidienne
* Implémenter une télémétrie riche dans les pipelines de flux de données et de déploiement
* Ajouter la gestion des erreurs pour déclencher des problèmes de pipeline en temps réel

**Défi 7 : Déploiement automatisé avec validation et approbation**

Dans ce défi, vous devrez automatiser les déploiements automatiques de solutions nouvelles et mises à jour.

Objectifs d’apprentissage :

* Rendre le processus de déploiement de la solution opérationnel via l’automatisation
* Créer et implémenter un environnement de test
* Implémenter des stratégies et des processus de déploiement automatisés

**Proposition de valeur**

* Une solution cloud moderne qui augmente la fiabilité, l’évolutivité et la facilité de maintenance de grandes quantités de données.
* Introduction de nouveaux services de stockage de données pour répondre aux besoins de flux de données uniques et multiples

**Scénarios techniques**

* Sources de données disparates : intégration de données provenant de plusieurs sources de données différentes dans un seul emplacement avec un schéma normalisé pour une utilisation standardisée en aval
* Sécurité des données : protéger les données à tout moment lors de l’utilisation des pipelines ETL
* DevOps : apprenez à utiliser un pipeline de production pour gérer la couche de données

**Audience**

* Public cible :
  + Microsoft – CSE, CSA, GBB, ATT, SE, TPM
  + Client - Développeurs PHP
* Secteurs verticaux cibles : Intersectoriel
* Profil client :
  + Llients qui essaient de traiter et de stocker des données provenant de sources multiples
  + Les clients qui ont besoin d’une solution DevOps qui prend en compte la gestion des données

**Questions d’inscription**

|  |  |  |
| --- | --- | --- |
| **Obligatoire** | **Champ** | **Options de réponse** |
| Oui | Quel est votre niveau de compréhension de l’utilisation d’Azure aujourd’hui ? | Aucune Une certaine compréhension J’ai une expérience de travail de base sur Azure Je compte sur Azure aujourd’hui pour le cloud |
| Oui | Combien de temps avez-vous passé à créer des solutions de données complexes sur Azure ? | [Liste déroulante ou boutons radio]  < 3 mois  3-6 mois  6 mois - 1 an  > 1 an |
| Oui | Est-ce que vous ou un client avec lequel vous travaillez avez une solution de données complexe actuellement en production sur Azure ? | [Liste déroulante ou boutons radio]  Oui  Non |
| Oui | Parmi les langages ou technologies suivants, quels sont ceux que vous avez utilisés dans un scénario d’entreposage de données ? | [Boutons radio ou case]  Databricks  Python  Scala  SSIS  Azure Data Factory  Autres |
| Oui | Parmi les langages ou technologies suivants, quels sont ceux que vous souhaitez le plus explorer pendant le OpenHack ? | [Boutons radio ou case]  Databricks  Python  Scala  SSIS  Azure Data Factory  Autres |
| Oui | Parmi les technologies suivantes, quelles sont celles que vous avez utilisées pour tirer parti du contrôle de version du code source ? | [Liste déroulante ou boutons radio]  Aucun  Git  TFVC  Subversion  Autre (veuillez préciser) |

**Questions de recherche de coach**

Les questions suivantes peuvent être utilisées pour trouver des coachs ayant les connaissances appropriées. Consultez Sélection du coach dans le [playbook OpenHack](https://aka.ms/openhackplaybook) pour plus d’informations.

|  |  |  |
| --- | --- | --- |
| Question # | Question | Ce qu’il faut rechercher dans la réponse |
| 1 | Quel est votre niveau d’expérience avec Data Factory ? | Expérience dans la construction de pipelines dans Data Factory et l’extraction à partir de serveurs SQL locaux |
| 2 | Quel est votre niveau d’expérience avec Data Lake ? | Expérience de l’intégration de Data Lake avec Data Bricks, de la gestion de la sécurité dans un lac de données |
| 3 | Quel est votre niveau d’expérience en matière de modélisation dimensionnelle d’entrepôt de données ? | Bonne compréhension de ce que sont les faits et les dimensions dans un modèle dimensionnel |
| 4 | Quel est votre niveau d’expérience avec Data Bricks ? | Expérience des transformations de données à l’aide de Python ou Scala (bon à avoir, mais pas obligatoire) avec Data Bricks, traitement des flux de données en temps réel avec Data Bricks |

**FAQ**

Q : Y a-t-il une liste de OpenHacks auxquels un participant devrait assister d’abord, avant de se rendre au vôtre ?

R : Non, un participant peut commencer son expérience OpenHack avec MDW. Une expérience préalable de DevOps pourrait être bénéfique pour un participant, mais les défis supposent une expérience préalable nulle.

Q : Les sujets liés à la migration sont-ils couverts ?

R : Non, ils ne le sont pas.

Q : Mon client souhaite migrer ses packages SSIS en utilisant l’ETL pour le faire localement. Cela sera-t-il évoqué ?

R : Si les participants ont un intérêt particulier dans les packages SSIS, ils peuvent choisir d’utiliser SSIS Integration Runtime pour lever et déplacer ces packages vers Azure.

Q : La gestion des flux de données en temps réel sera-t-elle couverte ?

R : Non, les participants travailleront à partir de données historiques.