**OpenHack - 最新のデータ ウェアハウジング**

**概要**

**この OpenHack では、参加者が** Microsoft Azure 上のマルチソース データ ウェアハウス ソリューション用の ETL パイプラインを開発、実装、運用できます。

**この OpenHack では実際のシナリオをシミュレートします。**オンライン DVD 会社で、異なる大量のソースから入ってくるデータを、一つの場所に保存し、解釈してから、さまざまなダウンストリーム システムにフィードする必要があるというものです。

**OpenHack の期間中、参加者は次のことに重点的に取り組みます。**

1. 複数のソースから体系的にデータを取り込み、セキュリティ保護する。
2. ビジネスで必要なスキーマに合わせ、DevOps テストのレベルでデータフローをモニターするためにデータを変換する。

**OpenHack の終わりまでに、参加者は技術ソリューションを構築します。**それは、信頼性、スケーラビリティ、保守性などの最高品質のデータ使用要件を満たすデータ管理を考慮した対応する CI/CD パイプラインを備えた、完全に機能する最新のデータ ウェアハウスです。

**テクノロジ**

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

**前提条件**

**知識に関する前提条件**

この OpenHack を順調に進めて最大限に活用するには、参加者は、リレーショナル データベースの構造と概念 (テーブル、結合、SQL など) に関する既存の知識と、SSIS、または Scala や Python などのプログラミング言語の経験を持っている必要があります。ETL パイプラインの作成、ソース管理、自動テスト、ビルドとリリースの自動化の経験があると、より迅速に進むことができます。

[Azure の基礎の知識が必要です。](https://docs.microsoft.com/ja-jp/learn/paths/azure-fundamentals/)

**言語固有の前提条件**

* Scala や Python などのプログラミング言語。

**ツールの前提条件**

ツールのダウンロードやインストールの遅延を避けるため、OpenHack に先だって以下を準備してください。

* Windows 10 (1703 以降)、Mac OSX (10.12 以降)、または[これらの Ubuntu バージョン](https://github.com/Azure/azure-functions-core-tools#linux)のいずれかを実行する最新のノート PC
* 任意の統合開発環境 (IDE) ソフトウェア ([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/)) のインストール
* 最新バージョンの [Azure CLI](https://docs.microsoft.com/ja-jp/cli/azure/install-azure-cli?view=azure-cli-latest) のダウンロード
* [SQL Server Database Tooling](https://docs.microsoft.com/ja-jp/sql/azure-data-studio/download-azure-data-studio?view=sql-server-ver15) (Azure Data Studio/SSMS)
* [SQL Server Data Tools](https://docs.microsoft.com/ja-jp/sql/ssdt/download-sql-server-data-tools-ssdt?view=sql-server-ver15) (BI ツールを含む) – Visual Studio for IDE を使用している場合のみ

**開発環境の構成**

ありません。

**リンクとリソース**

チームで作業するため、ペア プログラミングの概要を把握しておくと役立ちます。次を読むことをお勧めします。

* [ペア プログラミング](https://www.agilealliance.org/glossary/pairing/)

**学習後の推奨事項**

* [Azure Synapse Analytics を使用して Data Warehouse を実装する](https://docs.microsoft.com/ja-jp/learn/paths/implement-sql-data-warehouse/)
* [Azure Data Lake Storage Gen2 を使用した大規模なデータ処理](https://docs.microsoft.com/ja-jp/learn/paths/data-processing-with-azure-adls/)
* [コア Cloud Services - Azure データ ストレージ オプション](https://docs.microsoft.com/ja-jp/learn/modules/intro-to-data-in-azure/)
* [データ エンジニア向け Azure](https://docs.microsoft.com/ja-jp/learn/paths/azure-for-the-data-engineer/)
* [Azure Databricks を使用して Data Engineering を実行する](https://docs.microsoft.com/ja-jp/learn/paths/data-engineer-azure-databricks/)
* [Azure でのデータ プラットフォームの設計](https://docs.microsoft.com/ja-jp/learn/paths/architect-data-platform/)

**課題**

**課題 1:エンタープライズ データ レイク用のストレージを選択してプロビジョニングする**

この課題では、エンタープライズ データ レイクを確立します。

学習目標:

* Azure Storage オファリングを比較対照する
* 選択した Azure Storage サービスをプロビジョニングする

**課題 2:クラウド ソースからデータを取り込む**

この課題では、Azure SQL データベースと Cosmos DB コレクションから Southridge の初期データを抽出します。

学習目標:

* 複数のクラウドベースのソースから単一のクラウドベース ストアへのデータ インジェストを調整する
* 現在のテクノロジ セットおよびソリューション アーキテクチャを活用して、特定の顧客データが常に保護されるようにする

**課題 3:オンプレミスからデータをプルしてソース管理を確立する**

この課題では、追加のデータ ソースを新しいデータ レイクに組み込みます。初期データはクラウド ベースの Azure SQL データベースから抽出されましたが、この課題のデータは、オンプレミスのデータ ストアからのものです。

学習目標:

* 特に、管理された "オンプレミス" ソリューションからのデータの取り込みを調整する
* 開発されたソリューション用にクラウドベースのソース管理リポジトリを実装する

**課題 4:レイク内のデータを変換および正規化してブランチ ポリシーを確立する**

この課題では、 (まだご存じではなかった場合に) これらのソースからのデータにはさまざまなデータの型と形式があることがわかるでしょう。ダウンストリームのコンシューマー用に一貫性があるように前処理するのはチームの責任です。

学習目標:

* ダウンストリームでの使用のために、データを正規化されたスキーマに変換する
* 将来のすべての変更で適切なレビュープロセスが活用されるようにするために、新しいポリシーを作成する

**課題 5:データ ウェアハウスにデータを取り込み、単体テストを実装する**

この課題では、ビジネスの報告ニーズを満たす必要があります。Power BI レポートがチームに提供されますが、接続先のスター スキーマの作成と取り込みはチームの責任です。ターゲット スキーマの定義も提供されます。DevOps の観点から、チームは単体テストへのアプローチを確立する必要があります。

学習目標:

* ビジネスによって必須とされている特定のレポートの生成をサポートするために、さまざまなデータ ソース システムからのデータを共通のデータ ウェアハウス スキーマに変換する
* 自動化された方法でデータ ウェアハウスへのデータフローを調整する
* データ パイプラインのコア コンポーネント全体で単体テストを構築する
* 自動テストをコード レビュー プロセスに統合する

**課題 6:差分データ読み込みとテレメトリ**

この課題では、ビジネスの進行中のニーズに対処する必要があります。ビジネスで毎日発生するデータはデータ レイクに追加される必要がありますが、これはすべての履歴データを繰り返し処理するには非効率的です。段階的な読み込みを実装し、新しく取り込まれたデータの量をモニターできるログとテレメトリ ソリューションを確立する必要があります。

学習目標:

* 元の一括読み込みと差分データ読み込みを行うことを含めるようにソリューションを修正する
* 毎日実行するようにデータ読み込みと処理を自動化する
* データフローとデプロイ パイプラインに豊富なテレメトリを実装する
* パイプラインの問題をリアルタイムで提起するためにエラー処理を追加する

**課題 7:検証と承認を使用した自動デプロイ**

この課題では、新規および更新されたソリューションのゲート デプロイを自動化する必要があります。

学習目標:

* 自動化によってソリューションのデプロイ プロセスを運用する
* テスト環境を作成して実装する
* 自動デプロイ プロセスとポリシーを実装する

**価値の提案**

* 大量のデータの信頼性、スケーラビリティ、保守性を向上させる最新のクラウド ソリューション。
* 独自の複数のデータ ストリーム ニーズを満たす新しいデータ ストレージ サービスの紹介。

**技術シナリオ**

* 異なるデータ ソース: 標準化されたダウンストリームで使用するために、1 つの正規化されたスキーマを使用して、複数の異なるデータ ソースから 1 つの場所にデータを取り込みます
* データのセキュリティ: ETL パイプラインを使用しながら、データを常に保護します
* DevOps: 生産パイプラインを使用してデータ レイヤーを処理する方法を学びます

**対象者**

* 対象者:
  + Microsoft – CSE、CSA、GBB、ATT、SE、TPM
  + 顧客 – アプリ開発者
* 対象業種:さまざまな業界
* 顧客プロファイル:
  + 複数のソースからのデータを処理して保存することを試みているお客様
  + データ管理を考慮する DevOps ソリューションが必要なお客様

**登録時の質問**

|  |  |  |
| --- | --- | --- |
| **必須** | **フィールド** | **回答オプション** |
| はい | 現時点で、Azure の使用についてどの程度理解していますか? | まったく理解していない/少し理解している Azure でパイロット作業を行ったことが少しある 現在、クラウドには Azure を利用している |
| はい | Azure で複雑なデータ ソリューションを構築することに どの程度の時間がかかりましたか? | [ドロップダウンまたはラジオ ボタン]  3 か月以上  3 - 6 か月  6 か月 - 1 年  1 年以上 |
| はい | お客様、または共同作業相手には、Azure の運用環境で実行されている複雑なデータソリューションがありますか? | [ドロップダウンまたはラジオ ボタン] はい いいえ |
| はい | 次のうち、データ ウェアハウス シナリオで活用した言語またはテクノロジはどれですか? | [ラジオ ボタンまたはチェック ボックス]  Databricks  Python  Scala  SSIS  Azure Data Factory  その他 |
| はい | 次の言語またはテクノロジのうち、OpenHack の期間中に最も探ってみたいものはどれですか? | [ラジオ ボタンまたはチェック ボックス]  Databricks  Python  Scala  SSIS  Azure Data Factory  その他 |
| はい | 次のテクノロジのうち、ソース コードのバージョン管理を活用するために使用したものはどれですか? | [チェックボックスまたはラジオ ボタン]  なし Git  TFVC  Subversion  その他 (具体的に記載してください) |

**コーチ調達のための質問**

次の質問は、適切な知識を持つコーチを調達するために使用できます。詳細情報については、[OpenHack プレイブック](https://aka.ms/openhackplaybook)のコーチの選定に関するページをご覧ください。

|  |  |  |
| --- | --- | --- |
| 質問 # | 質問 | 回答の要注目点 |
| 1 | Data Factory での経験レベルはどの程度ですか? | Data Factory でのパイプライン構築、オンプレミス SQL サーバーからの抽出の経験 |
| 2 | Data Lake での経験レベルはどの程度ですか? | Data Lake と Data Bricks の統合、データ レイクのセキュリティ管理の経験 |
| 3 | Data Warehouse ディメンショナル モデルでの経験レベルはどの程度ですか? | ディメンショナル モデルに含まれるファクトとディメンションについての十分な理解 |
| 4 | Data Bricks での経験レベルはどの程度ですか? | Data Bricks で Python または Scala を使用してデータ変換を行った経験 (あればよいが、必須ではない)、Data Bricks でリアルタイム データ ストリームを処理した経験 |

**FAQ**

Q: 参加者が事前に参加しておくべき OpenHacks の推奨フローはありますか?

A: いいえ。参加者は MDW で OpenHack の経験を開始できます。事前の DevOps の経験は参加者のためになるかもしれませんが、課題は事前の経験がゼロであることを前提としています。

Q: 移行に関するトピックは対象ですか?

A: いいえ、違います。

Q: 顧客が、オンプレミスで行うために ETL を使用して SSIS パッケージを移行することを望んでいます。これは議論されますか?

A: 参加者が SSIS パッケージに特定の投資をしている場合は、SSIS Integration Runtime を使用してそのパッケージを Azure にリフト アンド シフトすることを選択できます。

Q: データのリアルタイム ストリームは対象ですか?

A: いいえ。参加者は履歴データから作業します。