|  |
| --- |
| OpenHack – Building and Operationalizing a Modern Data Warehouse |

# Overview

Microsoft’s OpenHack series is a three-day immersive, hands-on, challenge-driven hack that brings together developers from across the ecosystem and Microsoft to tackle scenarios influenced by common, real-world problems using Microsoft platform capabilities and other industry leading technologies.

Data is the lifeblood of any large organization. Many organizations are building enterprise data lakes to bring together the data from a variety of disparate sources into a single location that can be used to feed a wide variety of downstream systems – including data warehouses and machine learning systems. Most organizations don’t build these solutions out of thin air; these solutions are the product of extensive work with pre-existing systems. Building out these architectures is only half the story; modern cloud solutions must be built with an eye toward reliability, scalability, and maintainability. This requires a focus on the tenets of DevOps and the building of a solution that supports continuous integration and deployment – including testing and telemetry.

# Technologies

The “Modern Data Warehousing” OpenHack allows developers to learn how to develop, implement, and **operationalize** a multi-source data warehouse solution on Microsoft Azure, leveraging technologies such as Azure Data Lake Storage, Azure Data Factory, Azure Databricks, Azure SQL DW, and Azure DevOps. This is done by putting participants on teams and tasking them with challenges of incrementally increasing difficulty. In these challenges they will work with multiple and varied data sources and design, build, and deploy the architectural components necessary to empower the types of data consumption required by today’s organizations.

# Challenges

**Challenge 1: Select and provision storage for an enterprise data lake**

* Compare and contrast Azure storage offerings
* Provision the selected Azure storage service

**Challenge 2: Ingest data from cloud sources**

* Orchestrate the ingestion of data from multiple cloud-based sources to a single cloud-based store
* Ensure the protection of specific customer data at all times leveraging the current technology set and solution architecture

**Challenge 3: Pull data from on-premises and establish source control**

* Orchestrate the ingestion of data specifically from maintained “on-premises” solutions
* Implement a cloud-based source control repository for the developed solution

**Challenge 4: Transform and normalize data within the lake and establish branch policies**

* Transform data into a normalized schema for downstream consumption
* Create new policies to make certain all future changes leverage an appropriate review process

**Challenge 5: Populate a data warehouse and implement unit tests**

* Transform the data from the various source systems into a common data warehouse schema to support the generation of specific reports mandated by the business
* Orchestrate the dataflow into the data warehouse in an automated manner
* Build out unit tests across core components of the data pipeline
* Integrate automated testing into the code review process

**Challenge 6: Differential data loads and telemetry**

* Modify the solution to include doing differential data loads as well as the original bulk load
* Automate data load and processing to run daily
* Implement rich telemetry into the dataflow and deployment pipelines
* Add error handling to raise pipeline issues in real-time

**Challenge 7: Automated deployment with validation and approval**

* Operationalize the solution deployment process through automation
* Create and implement a testing environment
* Implement automated deployment processes and policies