Projektowanie Lakehouse w Azure

Zadanie 1 – Koszty roczne projektu Lakehouse w Azure

Dane wejściowe

- Dane historyczne: **10 TB**
- Przyrost miesięczny: **200 GB**
- Roczny przyrost: **2,4 TB**
- Łączna ilość danych: **12,4 TB**

Usługi i szacowane koszty

Usługa	**Szacowany koszt roczny**
Azure Storage Account	12,4 TB × 600 zł = 7 440 zł
Azure Databricks	10 000 – 15 000 zł
Azure Key Vault	ok. 500 zł
Azure Data Factory	2 000 – 3 000 zł
Azure SQL Server	4 000 – 6 000 zł
Suma	**24 000 – 32 000 zł / rok**

Zadanie 2 – Delta Lake vs Apache Iceberg

Delta Lake

- Silna integracja z Databricks.
- Obsługa ACID, schema evolution.
- Wydajna optymalizacja zapytań (Z-Ordering, Data Skipping).
- Łatwe wdrożenie w mniejszych projektach.
- **Scenariusze:** klasyczne BI, ETL, lakehouse w Databricks.

Apache Iceberg

- Niezależność od dostawcy chmury.
- Lepsze skalowanie i zarządzanie metadanymi.
- Obsługa wielu silników (Trino, Flink, Presto).
- **Scenariusze:** petabajtowe zbiory, multi-cloud, streaming.

Zadanie 3 – Krytyka architektury Medallion

Opis architektury

Architektura medalu dzieli dane na trzy warstwy:

- **Bronze** – dane surowe.
- **Silver** – dane oczyszczone, zjoinowane.
- **Gold** – dane gotowe do analizy biznesowej.

20 punktów krytyki architektury Medallion

1. Złożoność architektury.
2. Wysokie koszty przechowywania.
3. Dodatkowe opóźnienia w dostępie do danych.
4. Trudność śledzenia błędów w przepływie danych.
5. Wymaga zaawansowanego DevOps.
6. Nadmiarowość dla małych projektów.
7. Utrudniony dostęp dla użytkowników biznesowych.
8. Powielanie logiki transformacji.
9. Brak jednolitych standardów implementacji.
10. Trudność testowania regresji danych.
11. Wyższe koszty obliczeniowe.
12. Złożone pipeline'y.
13. Dłuższy czas wdrażania projektu.
14. Trudne zarządzanie wersjami danych.
15. Ograniczona elastyczność dla eksperymentów.
16. Problemy jakościowe nadal obecne mimo warstw.
17. Wysoka krzywa uczenia dla nowych członków.
18. Brak automatycznego śledzenia pochodzenia danych.
19. Dużo powielanego kodu w transformacjach.
20. Niekoniecznie potrzebna w każdym projekcie.