Zadanie 2

Porównanie Delta Lake i Apache Iceberg

Cecha	Delta Lake	Apache Iceberg
Wsparcie w Databricks	Natywne wsparcie, pełna integracja	Eksperymentalne, mniej zintegrowane
Format plików	Parquet + _delta_log	Parquet/ORC + Manifests
Kompatybilność z narzędziami BI	Power BI, Tableau itp.	Zależne od silnika
Skalowalność metadanych	Ograniczenia przy miliardach plików	Skaluje się do bardzo dużych zbiorów
Wsparcie wielu silników	Głównie Spark, Databricks	Spark, Trino, Flink, Hive, Presto
Time travel	Bardzo łatwe	Wymaga więcej konfiguracji

Delta Lake wybieramy, gdy używamy Databricks jako głównej platformy analitycznej, zespół pracuje głównie w Spark i gdy potrzebujemy prostej obsługi transakcyjnej (ACID) i czasowego podróżowania po danych (time travel).

Apache Iceberg wybieramy, gdy architektura zakłada wielosilnikowość, zarządzamy bardzo dużymi zbiorami danych (setki TB–PB) i gdy jesteśmy bardziej zaawansowani technicznie.

Zadanie 3

Krytyka architektury medallion:

1. **Złożoność** – wiele warstw to więcej kodu i zarządzania.
2. **Redundancja** – te same dane kopiowane między warstwami.
3. **Opóźnienie** – dane dostępne dopiero po kilku transformacjach.
4. **Słaba skalowalność** – trudne do utrzymania przy dużych zbiorach.
5. **Brak standardu** – każdy zespół może rozumieć warstwy inaczej.
6. **Niska elastyczność** – nie wszystkie przypadki wymagają 3 warstw.
7. **Koszt utrzymania** – więcej pipeline'ów to więcej problemów.
8. **Słaby lineage** – trudno śledzić przepływ danych.
9. **Nie dla biznesu** – analitycy chcą szybko dane, nie struktur.
10. **Overengineering** – zbyt rozbudowane dla prostych projektów.
11. **Większe koszty chmury** – wiele zapisów i odczytów danych.
12. **Problemy z real-time** – medallion nie jest zoptymalizowany pod streaming.
13. **Brak automatyzacji** – ręczne zarządzanie przepływem warstw.
14. **Trudny rollback** – cofanie zmian wymaga wielu kroków.
15. **Backfill trudny** – trzeba aktualizować wszystkie warstwy.
16. **Zarządzanie katalogiem** – więcej tabel to większy chaos.
17. **Powielanie logiki** – logika biznesowa w kilku miejscach.
18. **Niejasne granice** – gdzie kończy się Silver, a zaczyna Gold?
19. **Ciężko o jakość** – trudniej walidować dane w wielu warstwach.
20. **Trudna eksploracja** – dane Bronze często są nieczytelne bez obróbki.