Metaadat alapú integrációs platform (HST)

Szécsi Péter, IT Üzleti Intelligencia Architekt

- SQL,
- Python,
- BI,
- Cloud,
- Open-source

https://github.com/szecsip

https://linkedin.com/in/peter-szecsi/

Posta

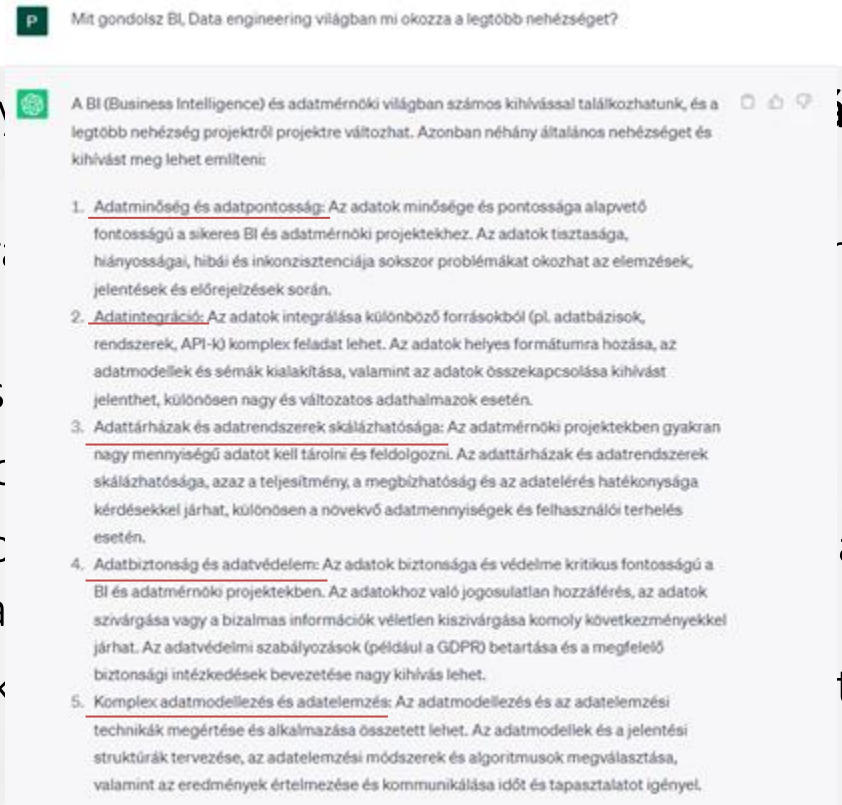
- Éves belföldi csomagforgalom: ~ **27,5 millió db**
- Éves levélforgalom: **~ 500 millió db**
- Közel **5000** járműből álló flotta,
- Éves szinten ~ **90 millió km** futásteljesítmény
- **~ 450 db** csomagautomata

https://linkedin.com/company/magyar-posta/

Magyar Posta

Problémafelvetés

- Projektjeink túlny **ációk** átfutási ideje a legnagyobb
- Egyes adatintegrá nás üzemeltetői egységeknél van
- Egyedi megoldás
- Performancia pro
- Forrásoldali válto az adatintegrációba
- Forrásrendszerek tások



Magyar Posta

Igény

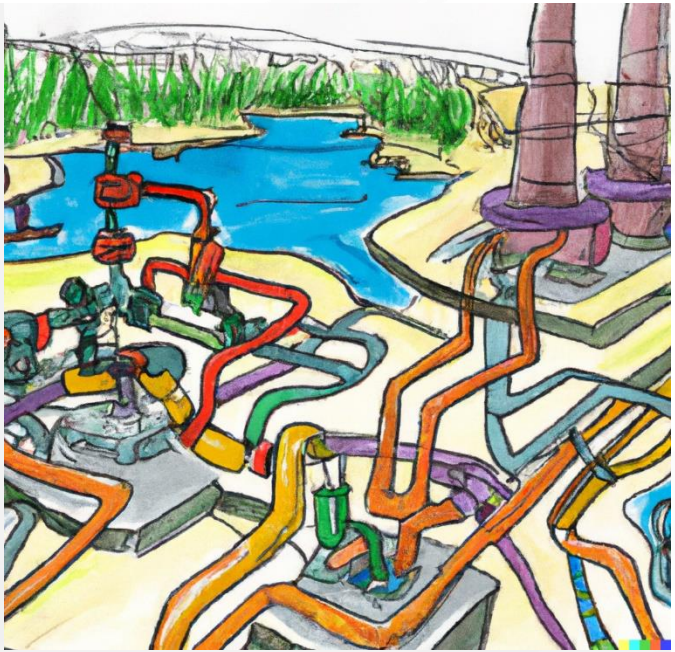
Szükség van egy olyan rendszerre, amely:

- Minimalizálja a humán erőforrás igényét és a hibázás lehetőségét
- Onpremise és felhős környezetben egyaránt használható
- Automatikusan felismeri az adatbázis megváltozását
- Használatával könnyen (automatikusan generálható módon) létre tudunk hozni adatintegrációkat
- Minimális forrásoldali terhelést okozzon
- Terhelés függvényétől könnyen skálázható
- Auditálható, részletes logolással rendelkezik
- Anonimizálásokat támogatja
- Könnyen üzemeltethető
- SQL tudással könnyen használható data productok álljanak elő

Alkotó komponensek – avagy a főszereplők

HST rendszer

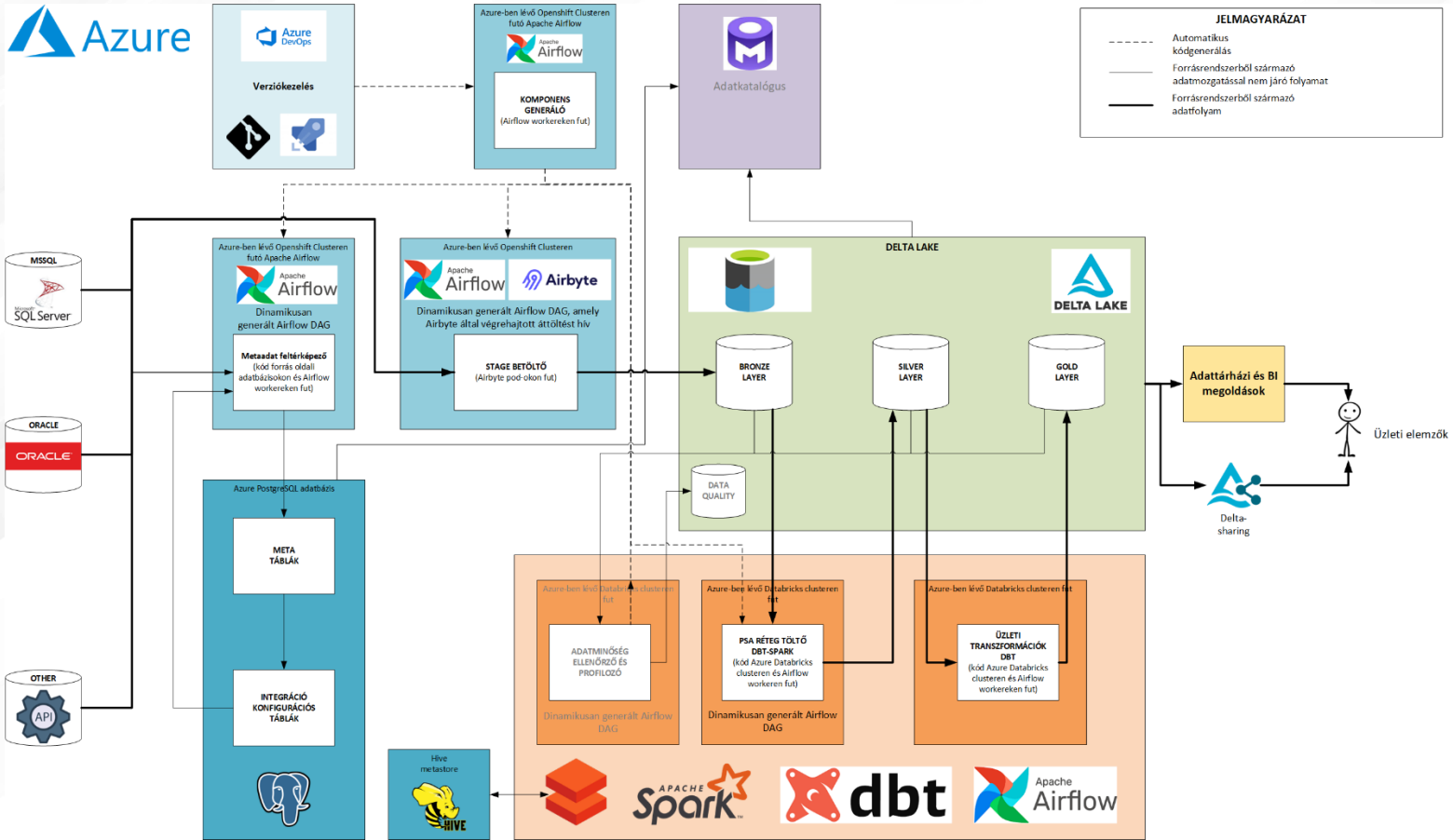
- Integráció konfigurátor
- Komponens generátor
- Metaadat feltérképező
- Stage betöltő
- Historizáló
- Adattranszformátor
- Profilozó
- Adatminőség ellenőrző
- Adatkatalógus



DALL·E 2 megértése a Lakehouseról és a pipelineokról

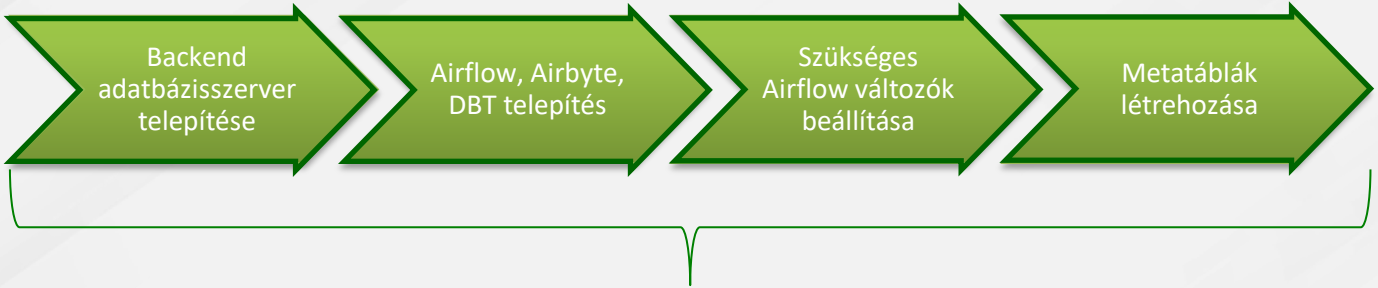
Magyar Posta

HST rendszer

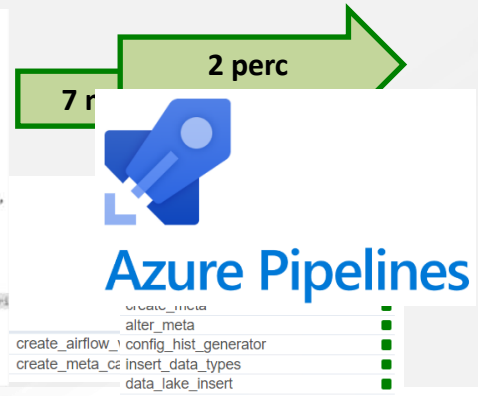


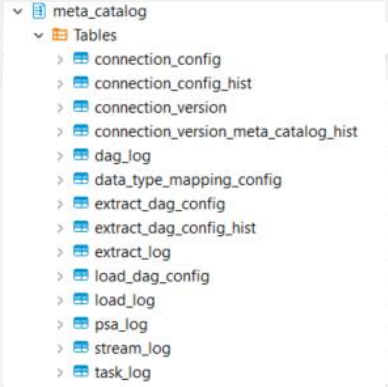
Magyar Posta

HST telepítése



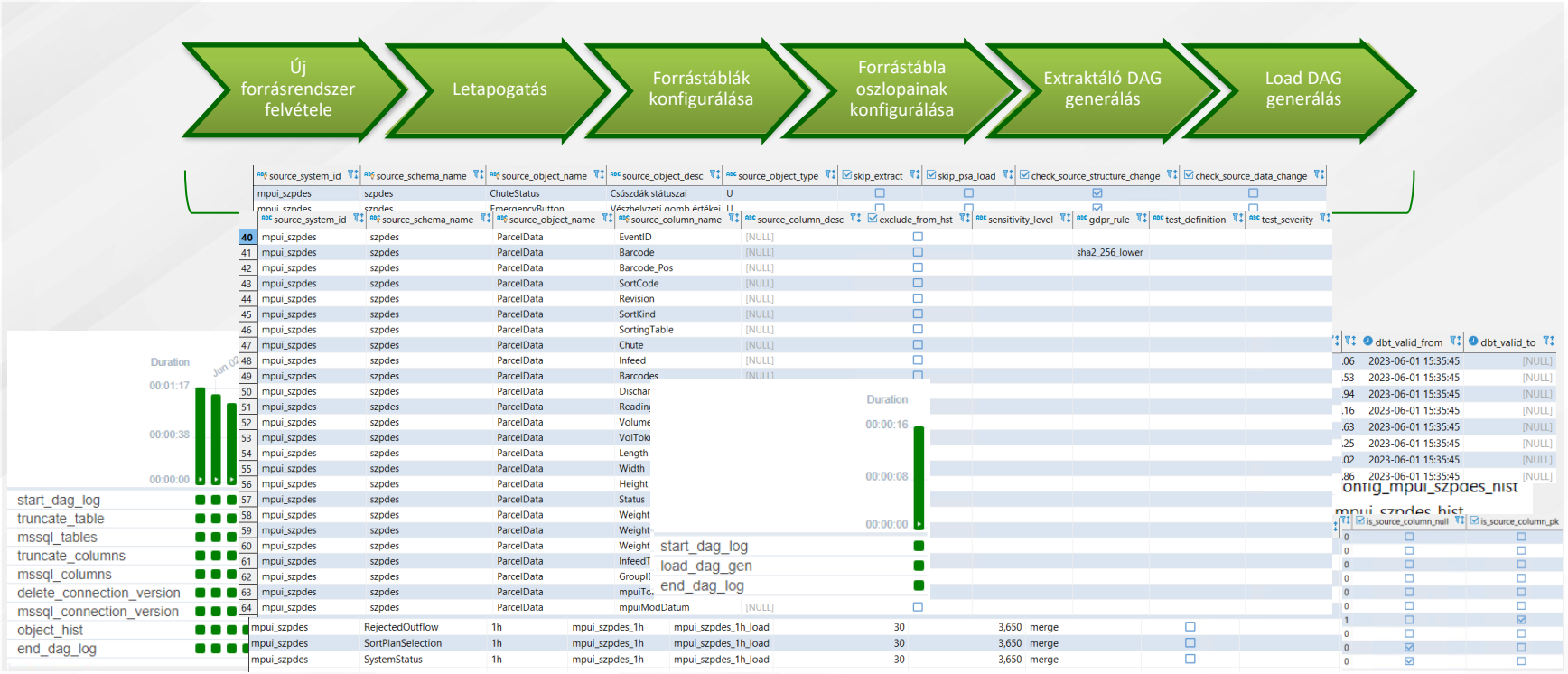
Automatikusan végrehajtható folyamat, időigénye kb **15 perc**





Magyar Posta

HST integráció konfigurációja



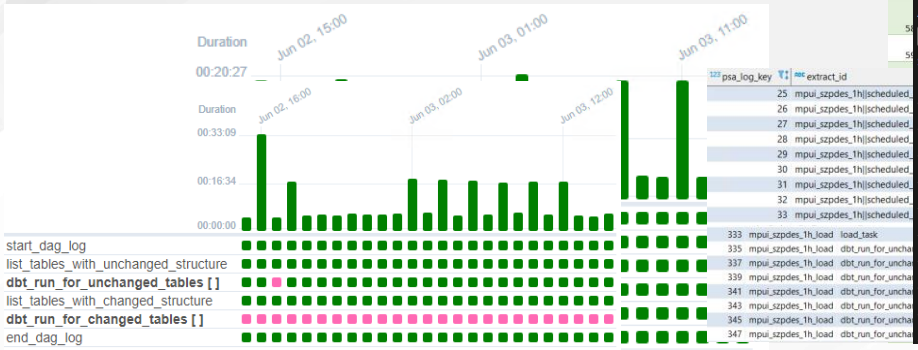
Magyar Posta

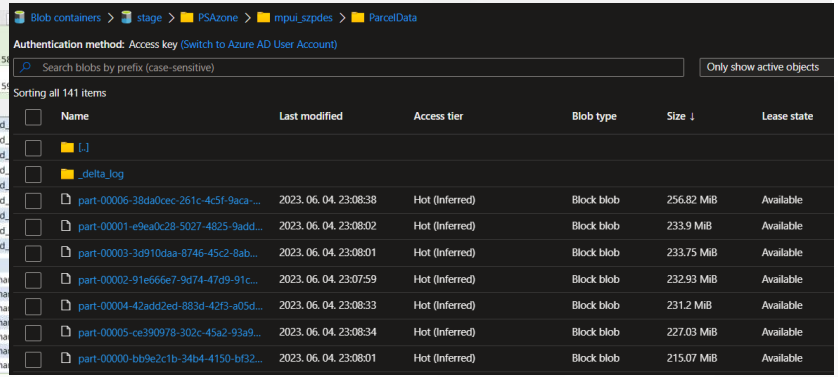
Lakehouse töltése HST használatával

Extraktáló DAG futása

Load DAG futása

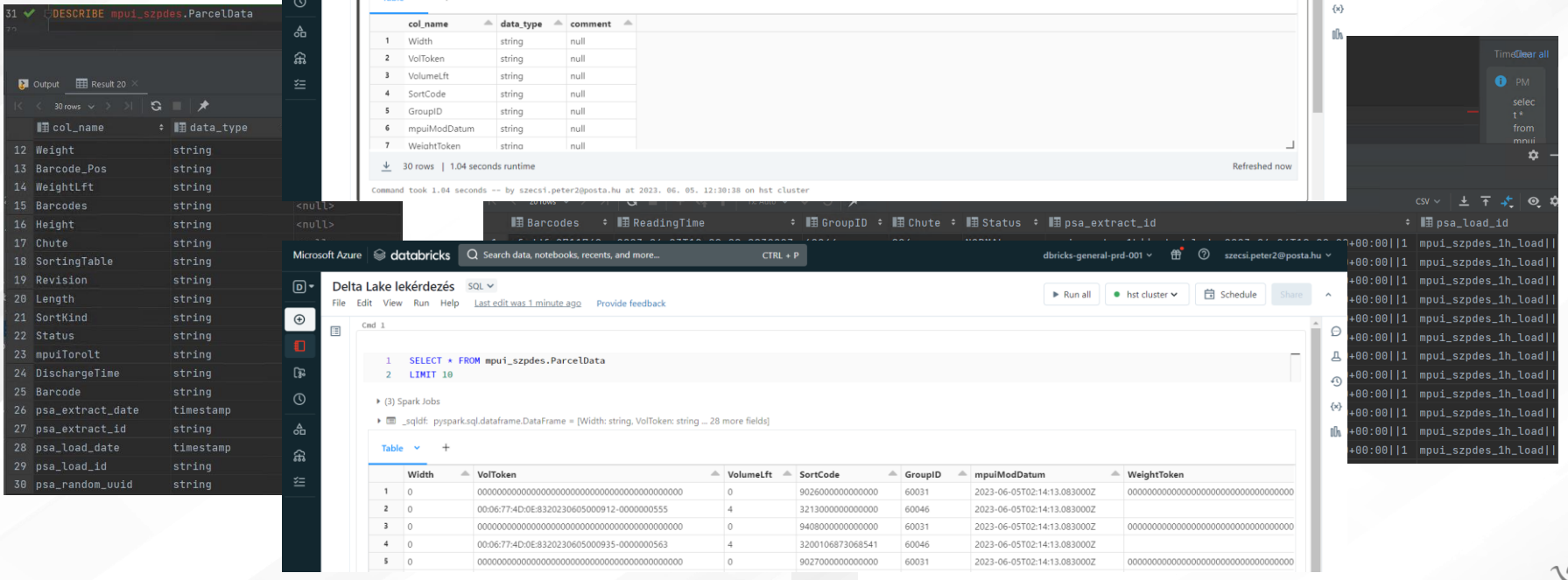
End-to-end automatizált és monitorozott folyamatok





Magyar Posta

HST eredményének használata (fejlesztőkörnyezetekből)



Magyar Posta

HST eredményének használata (Time travel alapú lekérdezés)

```
DESCRIBE HISTORY mpui_szpdes.ParcelData
```

	version	timestamp	op...	operationMetrics	userMetadata
39	23	2023-06-03 06:05:07.000000000	MERGE	{"executionTimeMs":"8845","numOutputRows":"342","numS…	mpui_szpdes_1h||scheduled__2023-06-03T04:00:00+0
40	22	2023-06-03 05:05:45.000000000	MERGE	{"executionTimeMs":"12061","numOutputRows":"1128","nu…	mpui_szpdes_1h||scheduled__2023-06-03T03:00:00+0
41	21	2023-06-03 04:05:55.000000000	MERGE	{"executionTimeMs":"7727","numOutputRows":"2541","num…	mpui_szpdes_1h||scheduled__2023-06-03T02:00:00+0
42	20	2023-06-03 03:05:31.000000000	MERGE	{"executionTimeMs":"10342","numOutputRows":"5774","nu…	mpui_szpdes_1h||scheduled__2023-06-03T01:00:00+0
43	19	2023-06-03 02:05:28.000000000	MERGE	{"executionTimeMs":"12188","numOutputRows":"4167","nu…	mpui_szpdes_1h||scheduled__2023-06-03T00:00:00+0
44	18	2023-06-03 01:05:27.000000000	MERGE	{"executionTimeMs":"12433","numOutputRows":"4144","nu…	mpui_szpdes_1h||scheduled__2023-06-02T23:00:00+0
45	17	2023-06-03 00:05:09.000000000	MERGE	{"executionTimeMs":"12214","numOutputRows":"2689","nu…	mpui_szpdes_1h||scheduled__2023-06-02T22:00:00+0
46	16	2023-06-02 23:05:41.000000000	MERGE	{"executionTimeMs":"10691","numOutputRows":"4302","nu…	mpui_szpdes_1h||scheduled__2023-06-02T21:00:00+0
47	15	2023-06-02 22:04:53.000000000	MERGE	{"executionTimeMs":"8624","numOutputRows":"2049","num…	mpui_szpdes_1h||scheduled__2023-06-02T20:00:00+0
48	14	2023-06-02 21:05:28.000000000	MERGE	{"executionTimeMs":"12475","numOutputRows":"2271","nu…	mpui_szpdes_1h||scheduled__2023-06-02T19:00:00+0
49	13	2023-06-02 20:04:47.000000000	MERGE	{"executionTimeMs":"11621","numOutputRows":"2435","nu…	mpui_szpdes_1h||scheduled__2023-06-02T18:00:00+0
50	12	2023-06-02 19:04:41.000000000	MERGE	{"executionTimeMs":"11572","numOutputRows":"2526","nu…	mpui_szpdes_1h||scheduled__2023-06-02T17:00:00+0
51	11	2023-06-02 17:21:13.000000000	MERGE	{"executionTimeMs":"5052","numOutputRows":"1950","num…	mpui_szpdes_1h||scheduled__2023-06-02T16:00:00+0
52	10	2023-06-02 17:20:26.000000000	MERGE	{"executionTimeMs":"6720","numOutputRows":"1156","num…	mpui_szpdes_1h||scheduled__2023-06-02T15:00:00+0
53	9	2023-06-02 16:04:25.000000000	MERGE	{"executionTimeMs":"9808","numOutputRows":"2671","num…	mpui_szpdes_1h||scheduled__2023-06-02T14:00:00+0
54	8	2023-06-02 15:05:08.000000000	MERGE	{"executionTimeMs":"9708","numOutputRows":"581","numS…	mpui_szpdes_1h||scheduled__2023-06-02T13:00:00+0
55	7	2023-06-02 14:05:13.000000000	MERGE	{"executionTimeMs":"8954","numOutputRows":"137","numS…	mpui_szpdes_1h||scheduled__2023-06-02T12:00:00+0

```
select count(*)
from mpui_szpdes.ParcelData TIMESTAMP AS OF "2023-06-02 10:30:00.000";
```

Output | count(*):decimal

1 row

	count(1)
1	10284521

```
select count(*)
from mpui_szpdes.ParcelData TIMESTAMP AS OF "2023-06-02 12:30:00.000";
```

Output | count(*):decimal

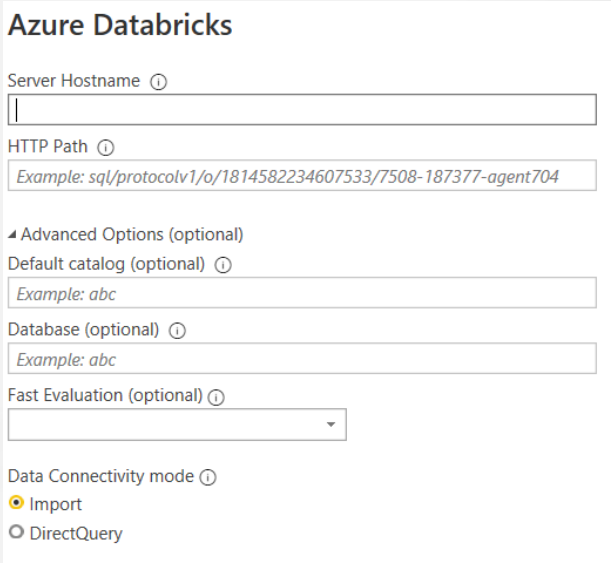
1 row

	count(1)
1	10284687

Magyar Posta

HST eredményének használata (Power BI)

- Közvetlenül Azure Databricks connectoron keresztül
  - Custom queryt is támogat
  - Time travelt támogat



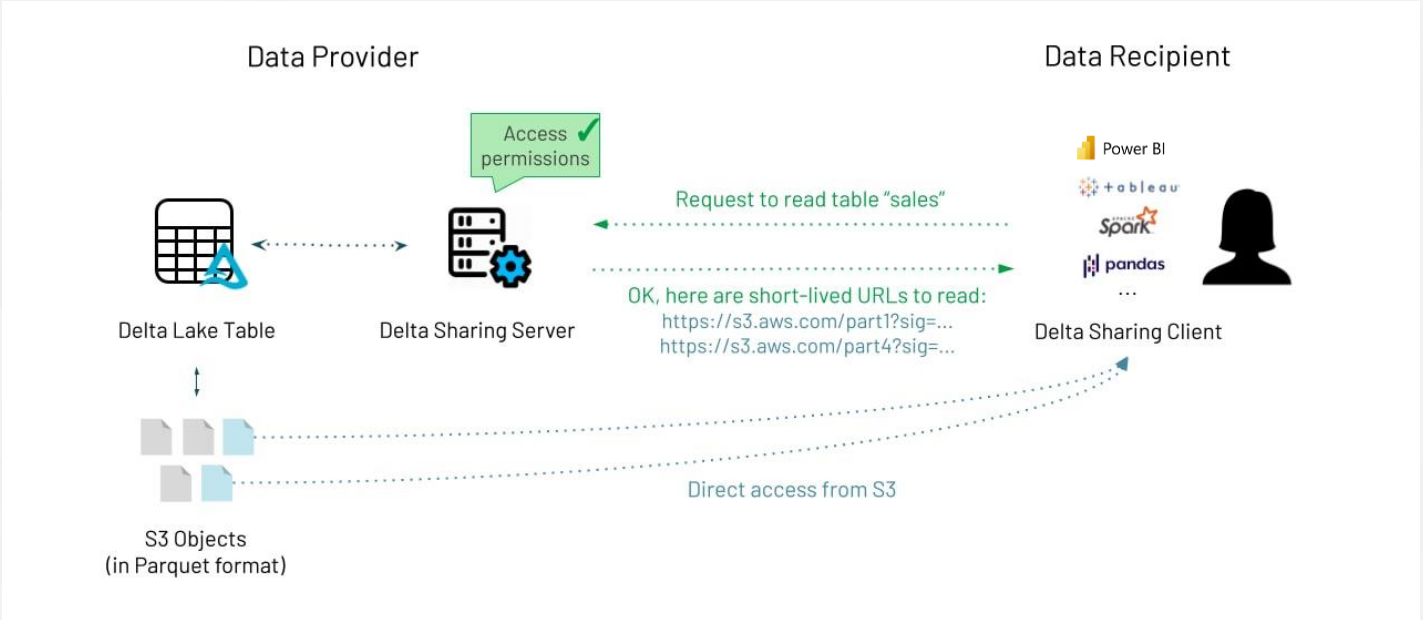
Magyar Posta

HST eredményének használata (Power BI)

- Power Query használatával
  - Spark clustert (Databrickset nem igényel a betöltéshez)
  - Custom queryt nem támogat, de számos betöltési adatoptimalizációt igen

HST eredményének használata (Power BI)

- Delta share használatával



Magyar Posta

Monitorozás

- Forrástáblák változásai
- Konfiguráció változások
- Futások száma
- Hibás futások aránya
- Extraktált és emittált rekordok száma és adatmennyisége
- Aktuálisan futó extraktálások állapota
- Extraktálás időszükséglete
- Extraktálás során hálózati átvitel
- Előző extraktálás óta eltelt idő
- Következő tervezett extraktálásig hátralévő idő
- Extraktálás során jelentkező késés a tervezett ütemezéshez képest

Magyar Posta

Továbbfejlesztési tervek

- UI felület fejlesztése a konfigurátor számára
- Adatkatalógus keretrendszerrel történő integrálás
- Delta sharing integrálása
- Great Expectations-el történő integrálás
- Új forrásrendszer típusok támogatása:
  - Valósidejű streaming: Apache Kafka
  - Fájlok: .csv, .xlsx, .json
  - Egyéb adatbázis típusok
  - API-k támogatása

Kérdések és válaszok

Magyar Posta