

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-szecsip/>



- 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/>

Problémafelvetés

P

Mit gondolsz BI, Data engineering világban mi okozza a legtöbb nehézséget?

- Projektjeink túlnyomórészt a legnagyobb
- Egyes adatintegrációs egységeknél van
- Egyedi megoldás
- Performancia probléma
- Forrásoldali váltás adatintegrációba
- Forrásrendszerek



A BI (Business Intelligence) és adatmérnöki világban számos kihívással találkozhatunk, és a legtöbb nehézség projektről projektre változhat. Azonban néhány általános nehézséget és kihívást meg lehet említeni:

1. Adatminőség és adatpontosság: Az adatok minősége és pontossága alapvető fontosságú a sikeres BI és adatmérnöki projektekhez. Az adatok tisztasága, hiányosságai, hibái és inkonzisztenciája sokszor problémákat okozhat az elemzések, jelentések és előrejelzések során.
2. Adatintegráció: Az adatok integrálása különböző forrásokból (pl. adatbázisok, rendszerek, API-k) komplex feladat lehet. Az adatok helyes formátumra hozása, az adatmodellek és sémák kialakítása, valamint az adatok összekapcsolása kihívást jelenthet, különösen nagy és változatos adathalmazok esetén.
3. Adattárházak és adatrendszerek skálázhatósága: Az adatmérnöki projektekben gyakran nagy mennyiségű adatot kell tárolni és feldolgozni. Az adattárházak és adatrendszerek skálázhatósága, azaz a teljesítmény, a megbízhatóság és az adatelérés hatékonysága kérdésekkel járhat, különösen a növekvő adatmennyiségek és felhasználói terhelés esetén.
4. Adatbiztonság és adatvédelem: Az adatok biztonsága és védelme kritikus fontosságú a BI és adatmérnöki projektekben. Az adatokhoz való jogosulatlan hozzáférés, az adatok szivárgása vagy a bizalmas információk véletlen kiszivárgása komoly következményekkel járhat. Az adatvédelmi szabályozások (például a GDPR) betartása és a megfelelő biztonsági intézkedések bevezetése nagy kihívás lehet.
5. Komplex adatmodellezés és adatelemzés: Az adatmodellezés és az adatelemzési technikák megértése és alkalmazása összetett lehet. Az adatmodellek és a jelentési struktúrák tervezése, az adatelemzési módszerek és algoritmusok megválasztása, valamint az eredmények értelmezése és kommunikálása időt és tapasztalatot igényel.



kezdő átvezető

és az üzemeltetői

az

szakértők

Szükség van egy olyan rendszerre, amely:

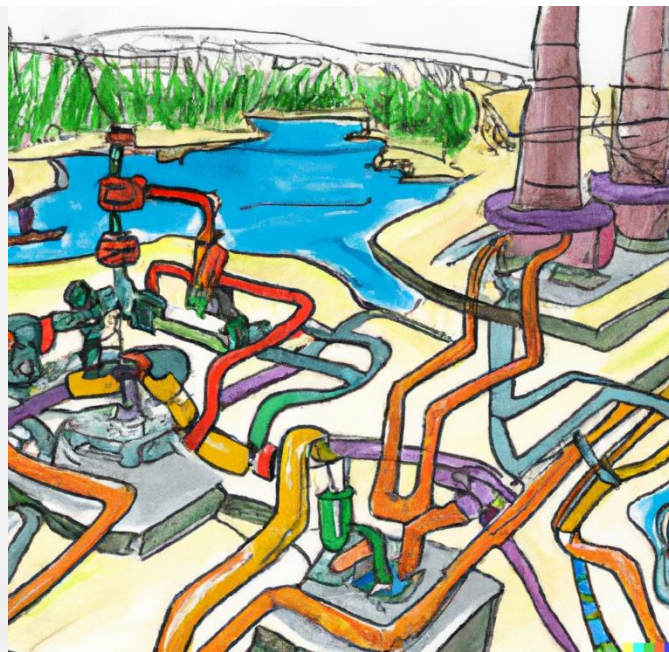
- Minimalizálja a humán erőforrás igényét és a hibázás lehetőségét
- Onpremise és felhős környezetben egyaránt használható
- Automatikusan felismeri az adatbázis megváltozását
- Használatával könnyen (automatikusan generálható módon) létre tudunk hozni adatintegrációkat
- Minimális forrásoldali terhelést 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



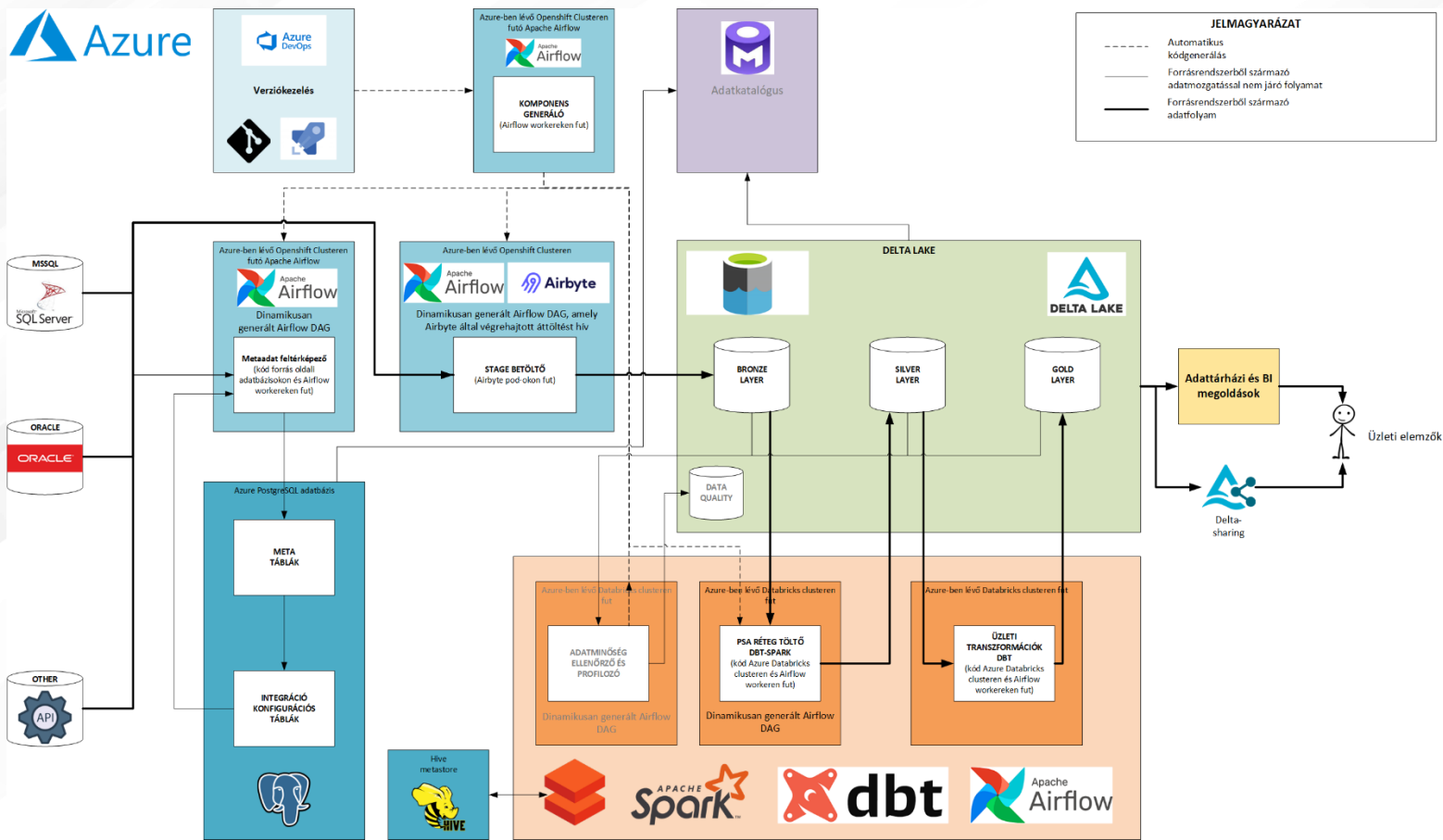
HST rendszer

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

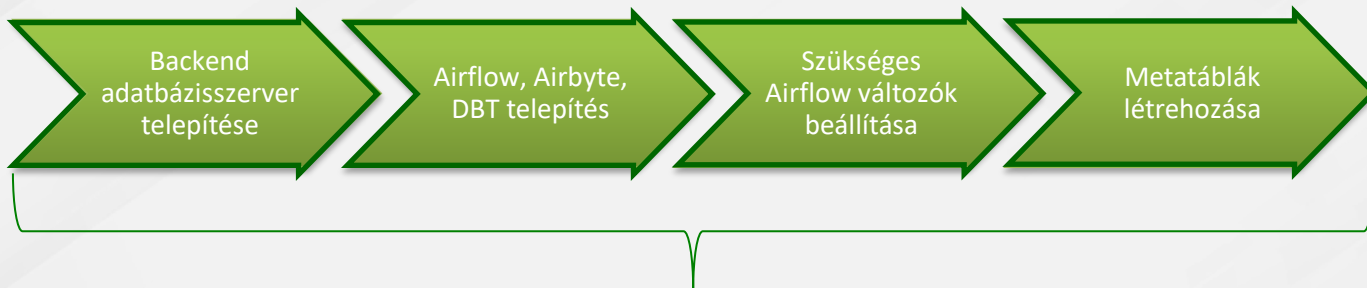


DALL-E 2 megértése a Lakehouseról és a pipelineokról

HST rendszer



HST telepítése



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

The image shows a screenshot of an Airflow DAG configuration interface on the left and an Azure Pipelines execution console on the right.

Airflow DAG Configuration:

- create_and_alter_meta_catalog** (highlighted with a red box)
- create_and_alter_script
- hst_airflow_vars_param
- hst_new_connection_and_dag_insert_param
- start_extract_dag_generator
- start_load_dag_generator
- start_meta_object_dag_generator

Azure Pipelines Execution:

- 7 r (7 minutes)
- 2 perc (2 minutes)
- Tasks: create_meta_alter_meta, create_airflow_config_hist_generator, create_meta_insert_data_types, data_lake_insert

meta_catalog Tables:

- connection_config
- connection_config_hist
- connection_version
- connection_version_meta_catalog_hist
- dag_log
- data_type_mapping_config
- extract_dag_config
- extract_dag_config_hist
- extract_log
- load_dag_config
- load_log
- psa_log
- stream_log
- task_log

Overlaid on the image are the logos for **HELM** and **BASH**, and a **Trigger** button in the Airflow interface.

HST integráció konfigurációja



mpui_szpdes	szpdes	ChuteStatus	Csúszdák státuszai	U	skip_extract	skip_psa_load	check_source_structure_change	check_source_data_change
mpui_szpdes	szpdes	ParcelData	EventID	[NULL]	<input type="checkbox"/>			
mpui_szpdes	szpdes	ParcelData	Barcode				sha2_256_lower	
mpui_szpdes	szpdes	ParcelData	Barcode_Pos	[NULL]	<input type="checkbox"/>			
mpui_szpdes	szpdes	ParcelData	SortCode	[NULL]	<input type="checkbox"/>			
mpui_szpdes	szpdes	ParcelData	Revision	[NULL]	<input type="checkbox"/>			
mpui_szpdes	szpdes	ParcelData	SortKind	[NULL]	<input type="checkbox"/>			
mpui_szpdes	szpdes	ParcelData	SortingTable	[NULL]	<input type="checkbox"/>			
mpui_szpdes	szpdes	ParcelData	Chute	[NULL]	<input type="checkbox"/>			
mpui_szpdes	szpdes	ParcelData	Infeed	[NULL]	<input type="checkbox"/>			
mpui_szpdes	szpdes	ParcelData	Barcodes	[NULL]	<input type="checkbox"/>			
mpui_szpdes	szpdes	ParcelData	Dischar					
mpui_szpdes	szpdes	ParcelData	Readin					
mpui_szpdes	szpdes	ParcelData	Volume					
mpui_szpdes	szpdes	ParcelData	VolTok					
mpui_szpdes	szpdes	ParcelData	Length					
mpui_szpdes	szpdes	ParcelData	Width					
mpui_szpdes	szpdes	ParcelData	Height					
mpui_szpdes	szpdes	ParcelData	Status					
mpui_szpdes	szpdes	ParcelData	Weight					
mpui_szpdes	szpdes	ParcelData	Weight					
mpui_szpdes	szpdes	ParcelData	Weight					
mpui_szpdes	szpdes	ParcelData	InfeedT					
mpui_szpdes	szpdes	ParcelData	GroupID					
mpui_szpdes	szpdes	ParcelData	mpuiTC					
mpui_szpdes	szpdes	ParcelData	mpuiModDatum	[NULL]	<input type="checkbox"/>			

Duration	start_dag_log	truncate_table	mssql_tables	truncate_columns	mssql_columns	delete_connection_version	mssql_connection_version	object_hist	end_dag_log
00:01:17	■	■	■	■	■	■	■	■	■
00:00:38	■	■	■	■	■	■	■	■	■
00:00:00	■	■	■	■	■	■	■	■	■

Duration	start_dag_log	load_dag_gen	end_dag_log
00:00:16	■	■	■
00:00:08	■	■	■
00:00:00	■	■	■

dbt_valid_from	dbt_valid_to	dbt_valid
.06	2023-06-01 15:35:45	[NULL]
.53	2023-06-01 15:35:45	[NULL]
.94	2023-06-01 15:35:45	[NULL]
.16	2023-06-01 15:35:45	[NULL]
.63	2023-06-01 15:35:45	[NULL]
.25	2023-06-01 15:35:45	[NULL]
.02	2023-06-01 15:35:45	[NULL]
.86	2023-06-01 15:35:45	[NULL]

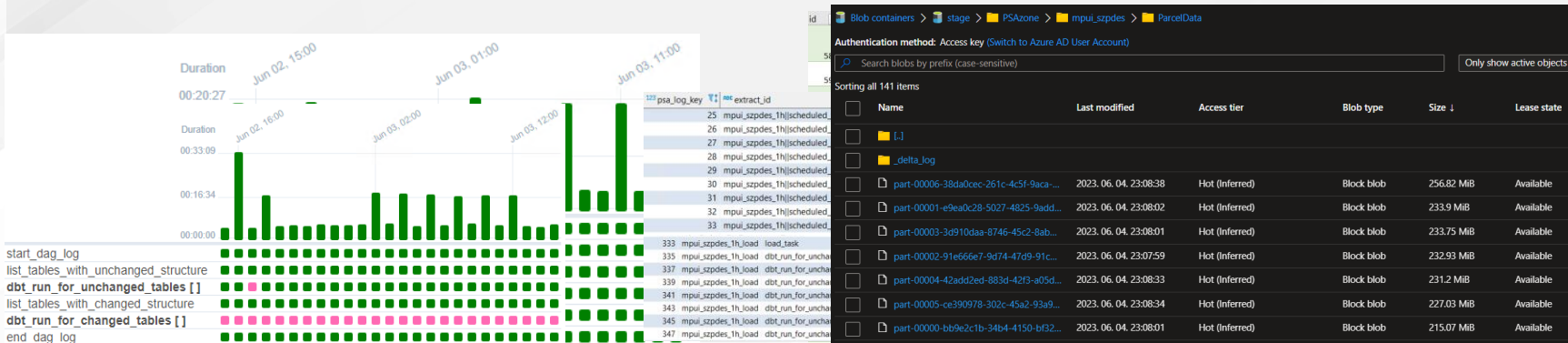
mpui_szpdes_hist	is_source_column_null	is_source_column_pk
0	<input type="checkbox"/>	<input type="checkbox"/>
0	<input type="checkbox"/>	<input type="checkbox"/>
0	<input type="checkbox"/>	<input type="checkbox"/>
0	<input type="checkbox"/>	<input type="checkbox"/>
0	<input type="checkbox"/>	<input type="checkbox"/>
0	<input type="checkbox"/>	<input type="checkbox"/>
0	<input checked="" type="checkbox"/>	<input type="checkbox"/>
0	<input type="checkbox"/>	<input type="checkbox"/>

Lakehouse töltése HST használatával

Extraktáló
DAG futása

Load DAG
futása

End-to-end automatizált és monitorozott folyamatok



HST eredményének használata (fejlesztőkörnyezetekből)

Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P dbricks-general-prd-001 szecci.peter2@posta.hu

Delta Lake lekérdezés SQL

File Edit View Run Help Last edit was now Provide feedback

Run all hst cluster Schedule Share

```
Cmd 1
```

```
1 DESCRIBE mpui_szpdes.ParcelData
```

	col_name	data_type	comment
1	Width	string	null
2	VolToken	string	null
3	VolumeLft	string	null
4	SortCode	string	null
5	GroupID	string	null
6	mpuiModDatum	string	null
7	WeightToken	string	null

30 rows | 1.04 seconds runtime Refreshed now

Command took 1.04 seconds -- by szecci.peter2@posta.hu at 2023. 06. 05. 12:30:38 on hst cluster

```
31 ✓ DESCRIBE mpui_szpdes.ParcelData
```

col_name	data_type
Weight	string
Barcode_Pos	string
WeightLft	string
Barcodes	string
Height	string
Chute	string
SortingTable	string
Revision	string
Length	string
SortKind	string
Status	string
mpuiTorolt	string
DischargeTime	string
Barcode	string
psa_extract_date	timestamp
psa_extract_id	string
psa_load_date	timestamp
psa_load_id	string
psa_random_uuid	string

Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P dbricks-general-prd-001 szecci.peter2@posta.hu

Delta Lake lekérdezés SQL

File Edit View Run Help Last edit was 1 minute ago Provide feedback

Run all hst cluster Schedule Share

```
Cmd 1
```

```
1 SELECT * FROM mpui_szpdes.ParcelData
```

```
2 LIMIT 10
```

(3) Spark Jobs

pspark.sql.dataframe.DataFrame = [Width: string, VolToken: string ... 28 more fields]

	Width	VolToken	VolumeLft	SortCode	GroupID	mpuiModDatum	WeightToken
1	0	00	0	9026000000000000	60031	2023-06-05T02:14:13.083000Z	00
2	0	00:06:77:4D:0E:8320230605000912-0000000555	4	3213000000000000	60046	2023-06-05T02:14:13.083000Z	00
3	0	00	0	9408000000000000	60031	2023-06-05T02:14:13.083000Z	00
4	0	00:06:77:4D:0E:8320230605000935-0000000563	4	3200106873068541	60046	2023-06-05T02:14:13.083000Z	00
5	0	00	0	9027000000000000	60031	2023-06-05T02:14:13.083000Z	00

TimedClear all

PM

```
select * from mpui
```

CSV ↓

+00:00		1		mpui_szpdes_1h_load	
+00:00		1		mpui_szpdes_1h_load	
+00:00		1		mpui_szpdes_1h_load	
+00:00		1		mpui_szpdes_1h_load	
+00:00		1		mpui_szpdes_1h_load	
+00:00		1		mpui_szpdes_1h_load	
+00:00		1		mpui_szpdes_1h_load	
+00:00		1		mpui_szpdes_1h_load	
+00:00		1		mpui_szpdes_1h_load	
+00:00		1		mpui_szpdes_1h_load	
+00:00		1		mpui_szpdes_1h_load	
+00:00		1		mpui_szpdes_1h_load	
+00:00		1		mpui_szpdes_1h_load	
+00:00		1		mpui_szpdes_1h_load	
+00:00		1		mpui_szpdes_1h_load	
+00:00		1		mpui_szpdes_1h_load	
+00:00		1		mpui_szpdes_1h_load	
+00:00		1		mpui_szpdes_1h_load	
+00:00		1		mpui_szpdes_1h_load	
+00:00		1		mpui_szpdes_1h_load	
+00:00		1		mpui_szpdes_1h_load	
+00:00		1		mpui_szpdes_1h_load	

HST eredményének használata (Time travel alapú lekérdezés)

version	timestamp	op...	operationMetrics	userMetadata
39	2023-06-03 06:05:07.000000000	MERGE	{\"executionTimeMs\":8845,\"numOutputRows\":342,\"numS...	mpui_szpdes_1h scheduled_2023-06-03T04:00:00+0
40	2023-06-03 05:05:45.000000000	MERGE	{\"executionTimeMs\":12061,\"numOutputRows\":1128,\"nu...	mpui_szpdes_1h scheduled_2023-06-03T03:00:00+0
41	2023-06-03 04:05:55.000000000	MERGE	{\"executionTimeMs\":7727,\"numOutputRows\":2541,\"num...	mpui_szpdes_1h scheduled_2023-06-03T02:00:00+0
42	2023-06-03 03:05:31.000000000	MERGE	{\"executionTimeMs\":10342,\"numOutputRows\":5774,\"num...	mpui_szpdes_1h scheduled_2023-06-03T01:00:00+0
43	2023-06-03 02:05:28.000000000	MERGE	{\"executionTimeMs\":12188,\"numOutputRows\":4167,\"nu...	mpui_szpdes_1h scheduled_2023-06-03T00:00:00+0
44	2023-06-03 01:05:27.000000000	MERGE	{\"executionTimeMs\":11433,\"numOutputRows\":4144,\"nu...	mpui_szpdes_1h scheduled_2023-06-02T23:00:00+0
45	2023-06-03 00:05:09.000000000	MERGE	{\"executionTimeMs\":12214,\"numOutputRows\":2689,\"nu...	mpui_szpdes_1h scheduled_2023-06-02T22:00:00+0
46	2023-06-02 23:05:41.000000000	MERGE	{\"executionTimeMs\":10691,\"numOutputRows\":4302,\"nu...	mpui_szpdes_1h scheduled_2023-06-02T21:00:00+0
47	2023-06-02 22:04:53.000000000	MERGE	{\"executionTimeMs\":8624,\"numOutputRows\":2049,\"num...	mpui_szpdes_1h scheduled_2023-06-02T20:00:00+0
48	2023-06-02 21:05:28.000000000	MERGE	{\"executionTimeMs\":12475,\"numOutputRows\":2271,\"nu...	mpui_szpdes_1h scheduled_2023-06-02T19:00:00+0
49	2023-06-02 20:04:47.000000000	MERGE	{\"executionTimeMs\":11621,\"numOutputRows\":2435,\"nu...	mpui_szpdes_1h scheduled_2023-06-02T18:00:00+0
50	2023-06-02 19:04:41.000000000	MERGE	{\"executionTimeMs\":11572,\"numOutputRows\":2526,\"nu...	mpui_szpdes_1h scheduled_2023-06-02T17:00:00+0
51	2023-06-02 17:21:13.000000000	MERGE	{\"executionTimeMs\":5052,\"numOutputRows\":1950,\"num...	mpui_szpdes_1h scheduled_2023-06-02T16:00:00+0
52	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\":9888,\"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	10284521
---	----------

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

Output count(*) decimal

1	10284687
---	----------

HST eredményének használata (Power BI)

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

Azure Databricks

Server Hostname ⓘ

HTTP Path ⓘ

Example: sql/protocolv1/o/1814582234607533/7508-187377-agent704

Advanced Options (optional)

Default catalog (optional) ⓘ

Example: abc

Database (optional) ⓘ

Example: abc

Fast Evaluation (optional) ⓘ

Data Connectivity mode ⓘ

Import

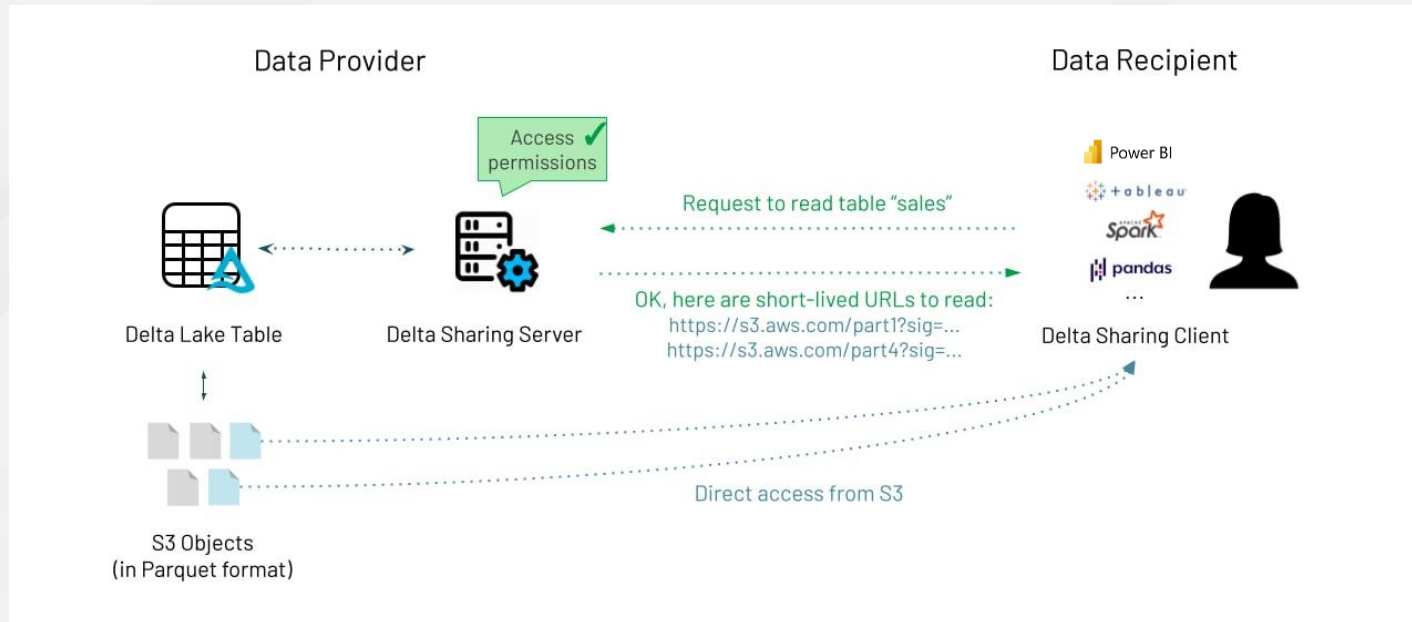
DirectQuery

HST eredményének használata (Power BI)

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

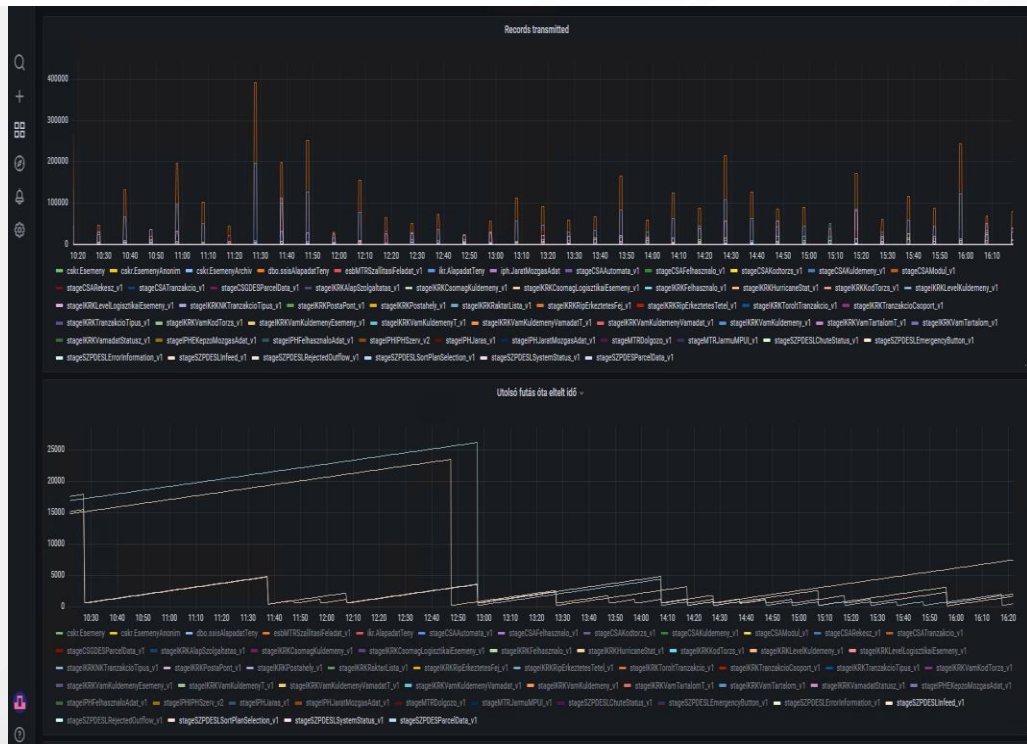
HST eredményének használata (Power BI)

- Delta share használatával



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



Továbbfejlesztési tervek

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

Kérdések és válaszok