Data Observability Monte Carlo-val – Mit, miért, hogyan?

Budapest Data Forum 2022

Nagy András István
Senior Data Solution Architect

Nagy András István

Senior Data Solution Architect

Data Observability

Mi az a "Data observability"?

Miért kell nekünk ilyen?

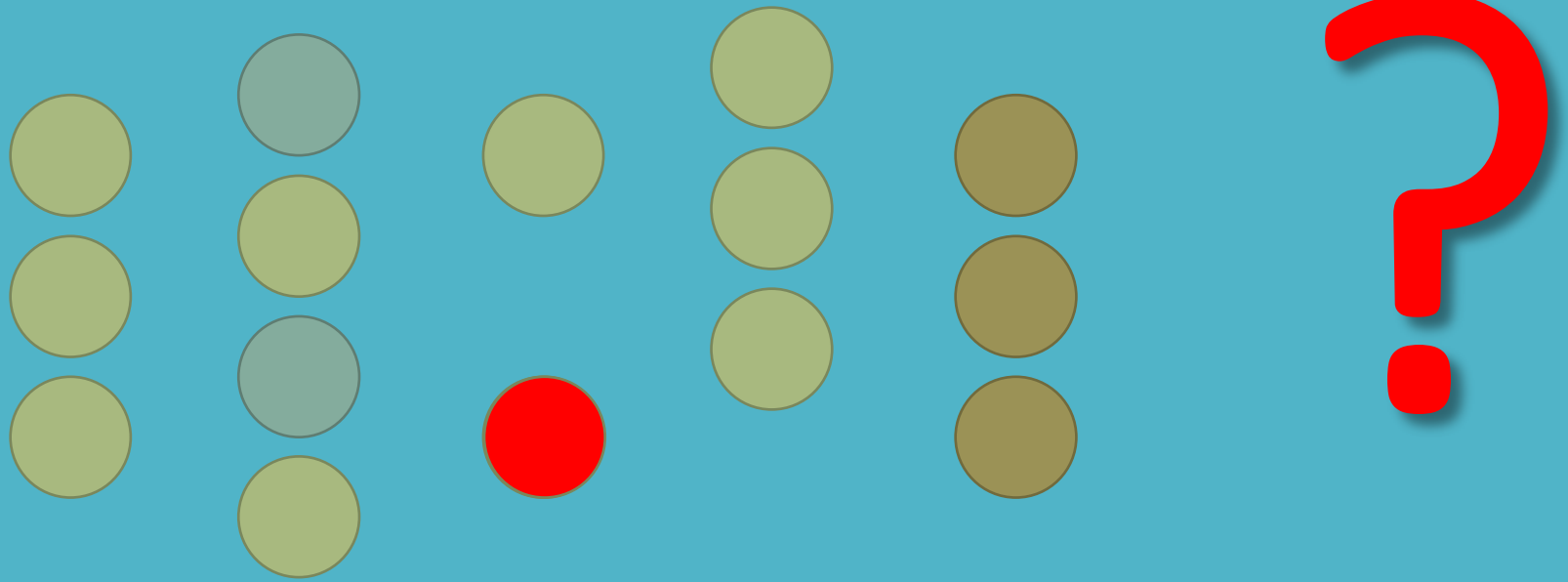
Hogyan tudjuk elérni?

Ügyfelünknél hogyan vezettük be:

- Eszköz kiválasztása
- Bevezetés
- Beillesztés a folyamatainkba

~~Data~~ Observability

Downtime



~~Data~~ Observability

Egy rendszer képessége arra, hogy megfigyelhető legyen a működése

1) Egyáltalán mit kell megfigyelni?
2) Mi a teendő, ha gond van?
3) Észrevenni előre nem definialt eseményekből is az esetleges hibát

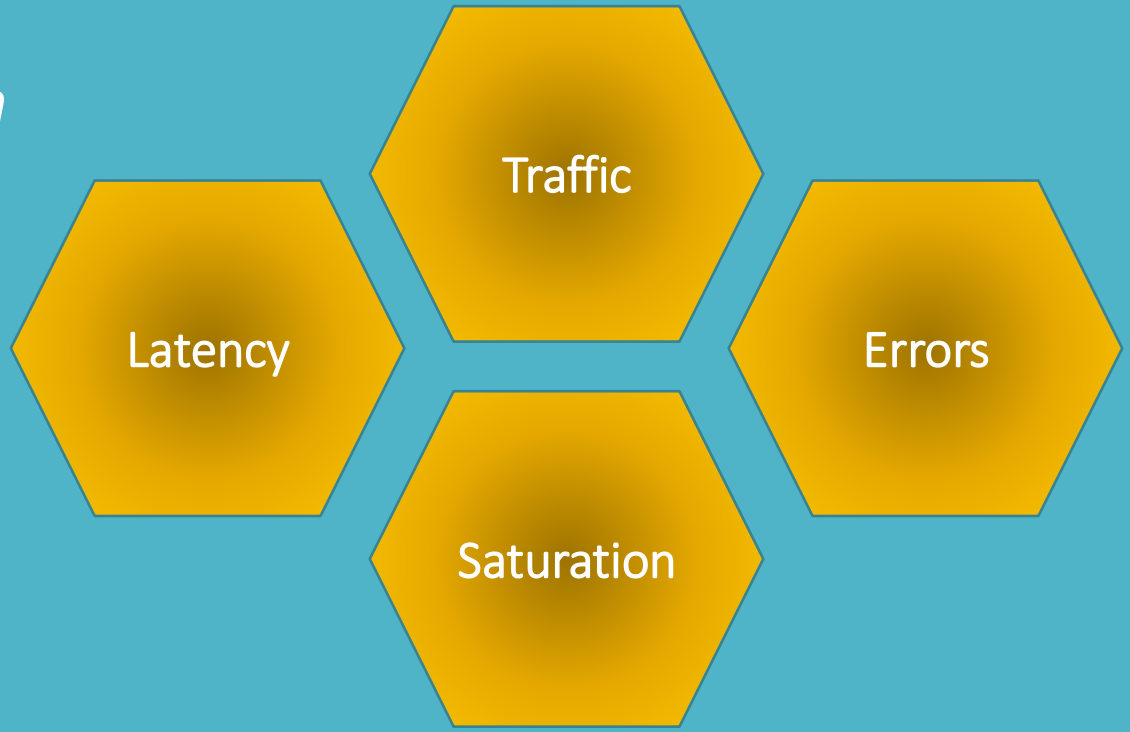
Eszközök:

- Metrikák
- Logok
- Trace-ek
- Health check

~~Data~~ Observability

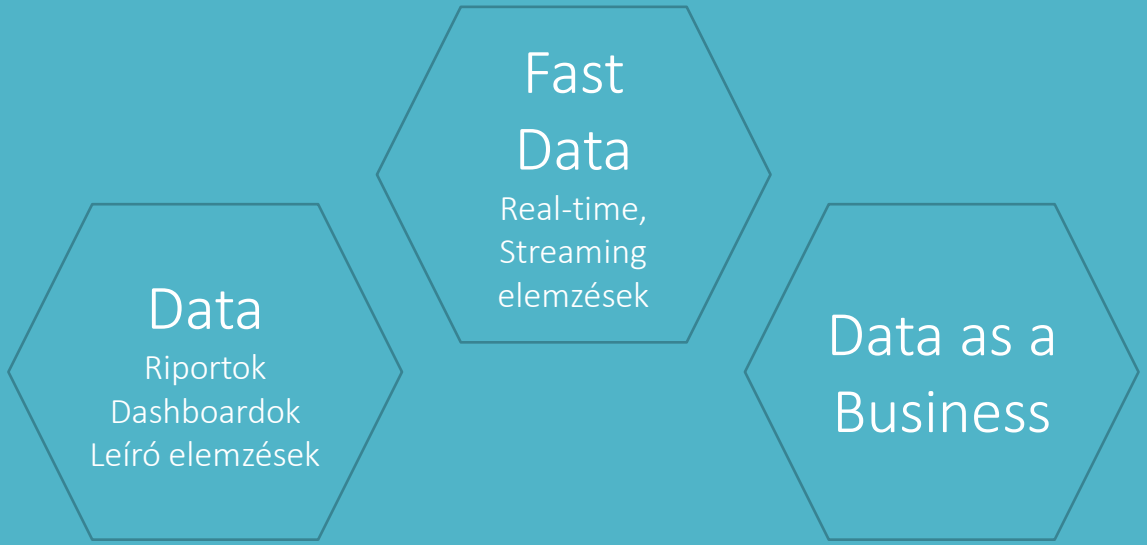
4 Golden Signals –
Site Reliability Engineering: How Google Runs Production Systems. Sebastopol, CA: O'Reilly, 2016.

Eszközök:
- White-box monitoring
- Black-box monitoring



Data Observability

Data Downtime

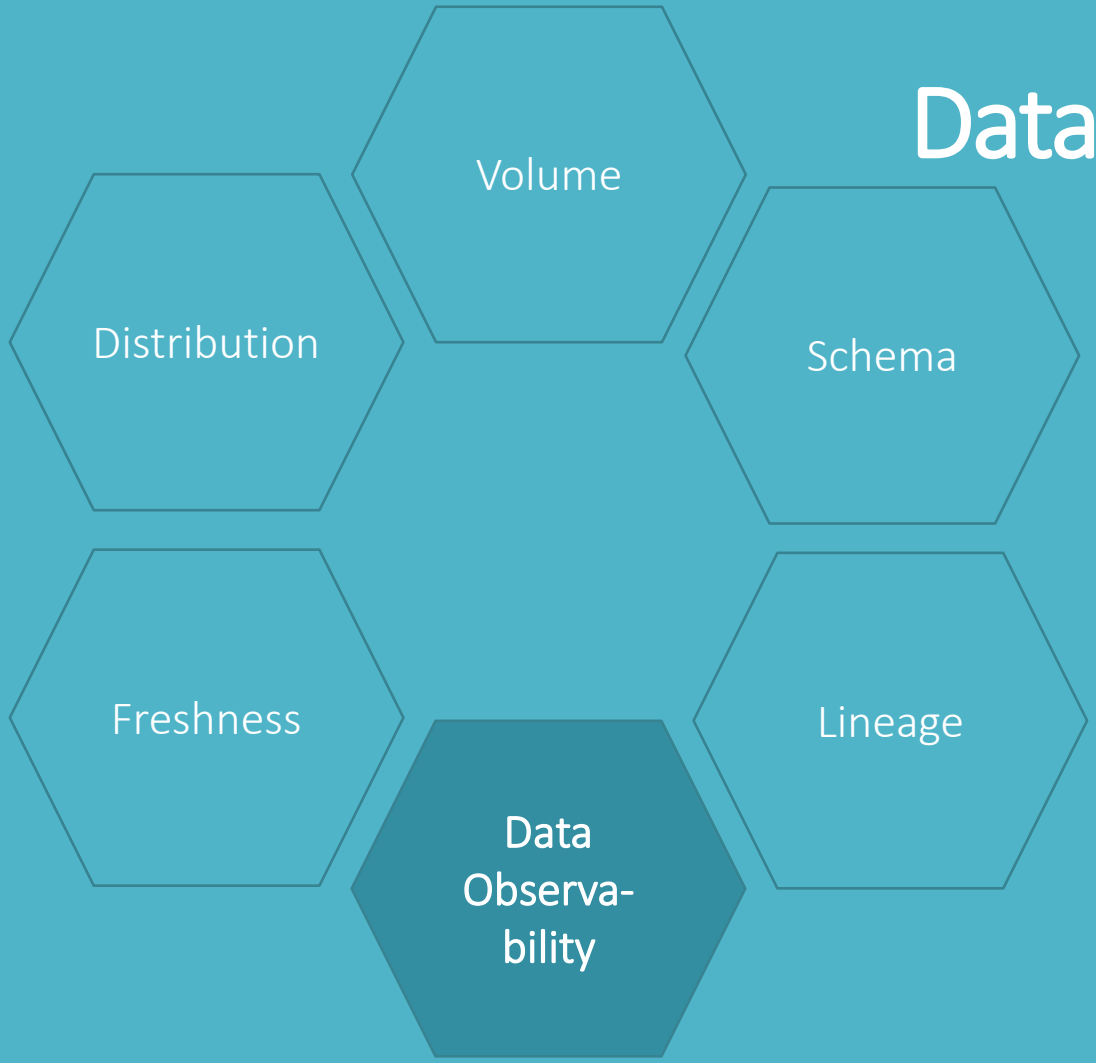


Data Observability

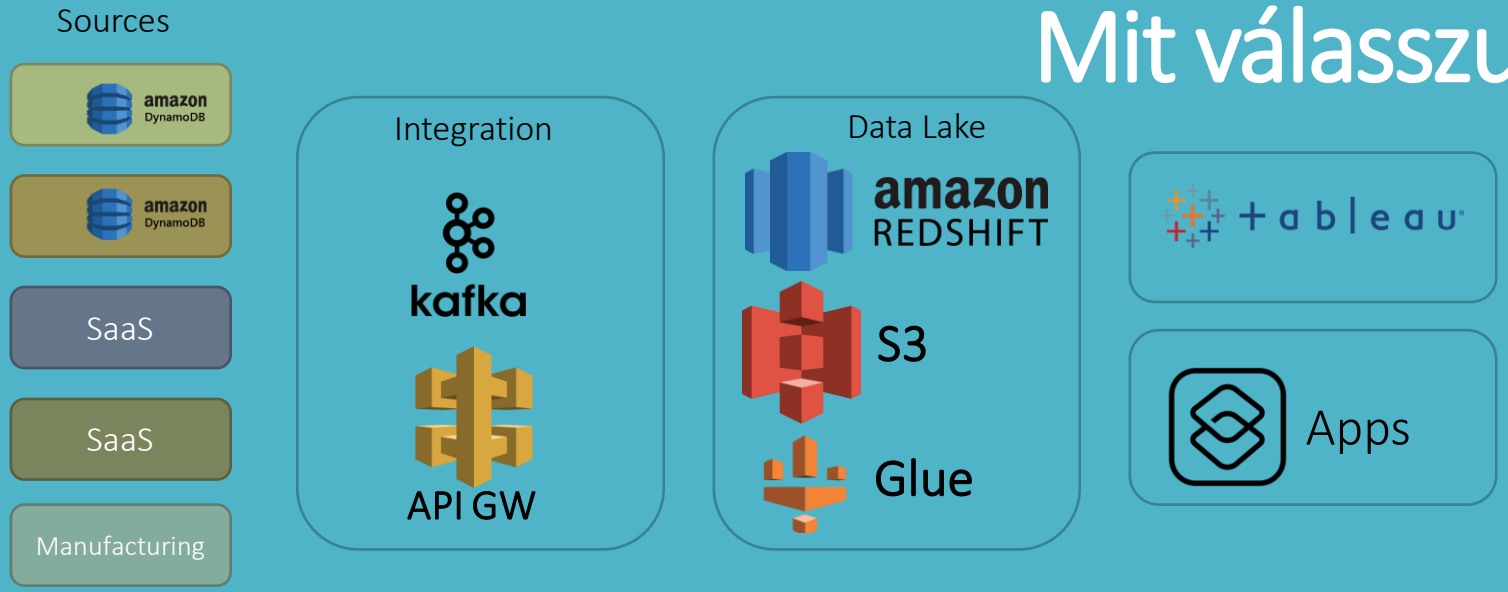
Data Downtime

- Megfigyelhető (Observable)
- Kontextus
- Reagálás

Data Observability



Mit válasszunk?



Forrásrendszerek:

- Egyedi alkalmazások (DynamoDB, D. Streams)
- SaaS (ServiceNow, Coupa, Jira,…)
- Egyéb rendszerek (API, CDC)

DWH / Data Lake:

- S3 – Raw, Transformed: ML training
- Redshift: staging, ODS, Reporting layer

Analitikus alkalmazások:

- Tableau
- Egyedi data app-ok

Mit válasszunk?

Fő szempontok:

- Data Observability képességek
- A bevezetés ráfordításigénye
- Illeszkedés a környezetünkbe

Mit válasszunk?

- Data Lineage feltérképezése, lineage alapján impact analízis
- Sémaváltozás detektálása, processzek
- Data Security and privacy
- Szabályok egyedi definiálása (DQ expectations)
- Szabályok automatikus felállítására, automatizált anomália detekció
- Adatok frissességének monitorozása
- Csatolás a pipeline-ok és az eszköz között - hozzá kell-e nyúlni a pipeline-jainkhoz?
- Bevezetés erőforrásigénye
- Üzemeltetés erőforrásigénye
- Teljesítmény, skálázhatóság

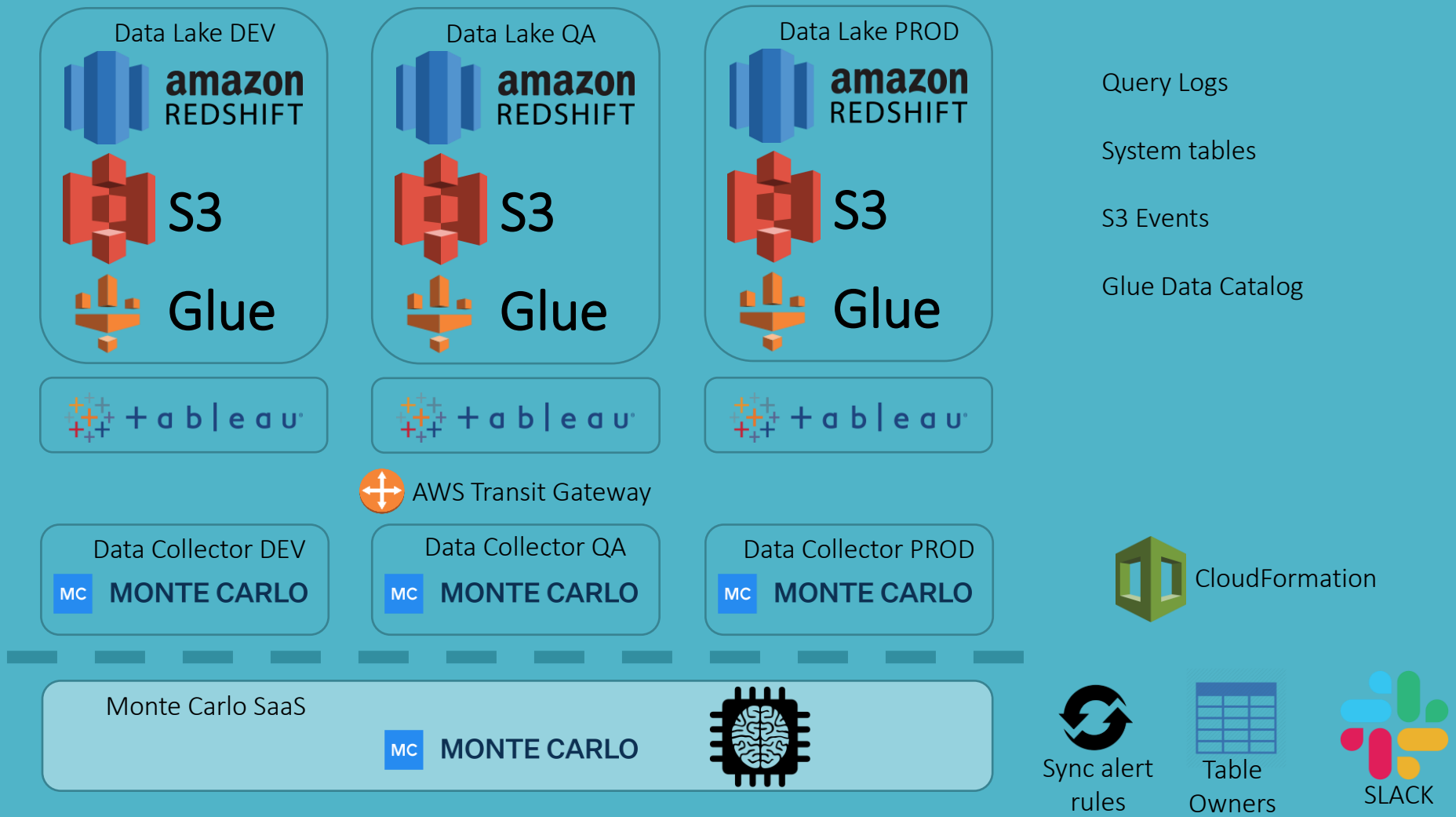
Mit válasszunk?

Interfészek:

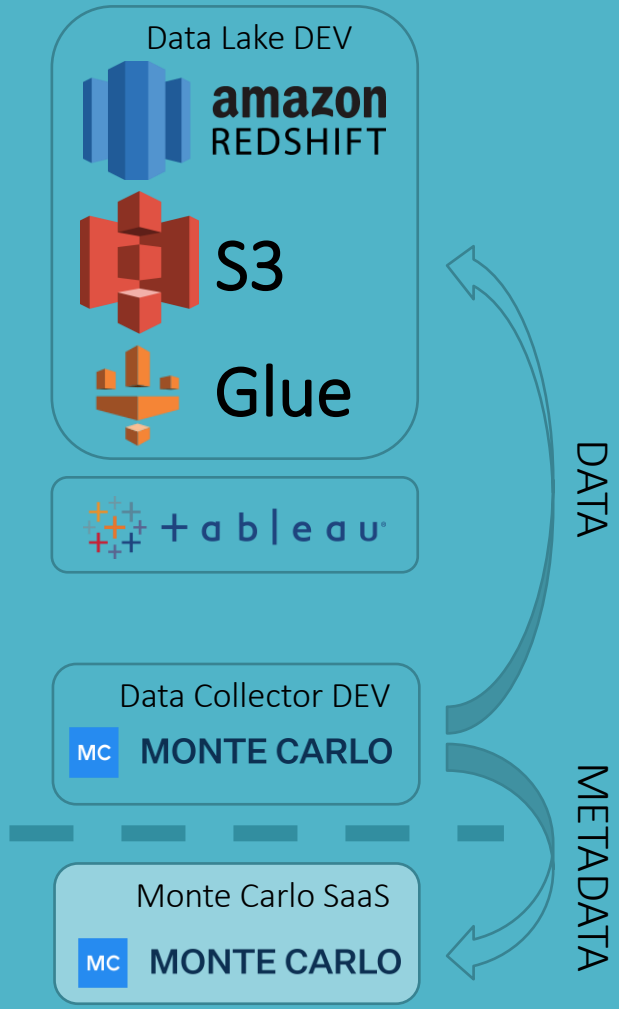
- Interaktív, vizuális felhasználói felület
- Incidensek esetében értesítések kiküldése, triaging
- API az integrációhoz
- Access Control
- Integráció: Redshift, Tableau, S3, Glue Data Catalog, Kafka
- Illeszkedés AWS-re épülő data infrastruktúránkhoz

Mit válasszunk?

- Célzott, custom megoldás Apache Deequ / AWS DQAF alapon
- BigEye (DO) + Alation (DC)
- Collibra DQ (DO/DQ) + Collibra (DC)
- Monte Carlo Data (DO)


Data Lake DEV
amazon REDSHIFT
S3
Glue
Data Lake QA
amazon REDSHIFT
S3
Glue
Data Lake PROD
amazon REDSHIFT
S3
Glue
Query Logs
System tables
S3 Events
Glue Data Catalog
+ableau
+ableau
+ableau
AWS Transit Gateway
Data Collector DEV
MC MONTE CARLO
Data Collector QA
MC MONTE CARLO
Data Collector PROD
MC MONTE CARLO
CloudFormation
Monte Carlo SaaS
MC MONTE CARLO
Sync alert rules
Table Owners
SLACK

Privacy


Data Lake DEV
amazon
REDSHIFT
S3
Glue
tableau
DATA
Data Collector DEV
MC
MONTE CARLO
METADATA
Monte Carlo SaaS
MC
MONTE CARLO

Következő lépések

- Lineage kiterjesztése
- Circuit Breakers: SQL Monitor integráció Airflow DAG-ből

Tapasztalatok

Közös slack csatorna, gyors reakció

Dedikált Customer Success Manager

Része a folyamatainknak, incidenseket mi vesszük észre, nem az ügyfeleink

Házon belüli elterjedtség nő