Implement a Serverless Data Analytics Strategy with Microsoft Azure

Roberto Freato

orienteering

data governance

key principles

people

size

technology

data mesh principles

Domain-oriented decentralized data ownership and architecture

"responsibility to people closest to the data in order to support continuous change and scalability"

Data as a product

"the domain data product owner can be responsible for the objective measures that ensure data is delivered as a product"

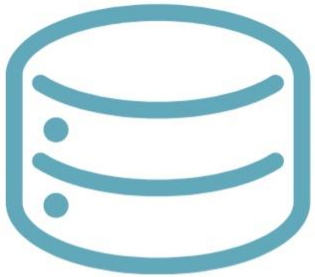
Self-serve data infrastructure as a platform

"the only way that teams can autonomously own their data products is to have access to a high-level abstraction of infrastructure that removes complexity and friction of provisioning and managing the lifecycle of data products"

Federated computational governance

"maintaining an equilibrium between centralization and decentralization [...] creating interoperability and a compounding network effect through discovery and composition of data products"

https://martinfowler.com/articles/data-mesh-principles.html



sourcing

- clicks report


```
{
    “createdAt”: “2021-01-01T06:38:10+00:00”,
    “country”: “NL”,
    “account”: “db65103e-ce9b-4022-bbf4-f3de44b280a8”,
    “campaign”: “b0dc149c-bc4d-401d-ab11-e0b03e799b0f”,
    “cost”: 14.50,
    “clicks”: 44416
}
```

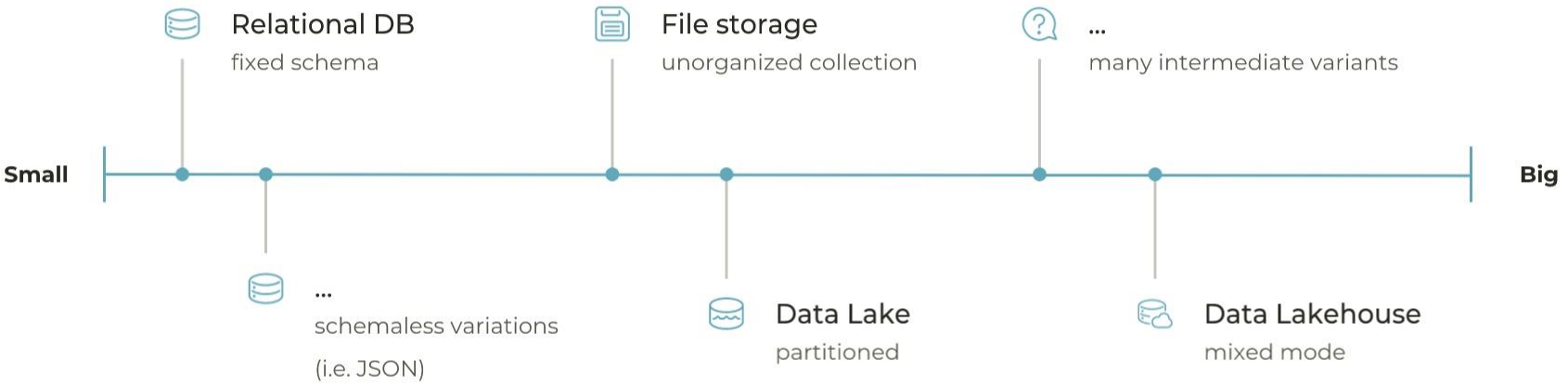

- impressions report


```
{
  “campaign_id”: “017a1af5-8852-46b3-b32e-2939811be7ed”,
  “creativity_id”: “e”,
  “impression_count”: 796300,
  “cost”: 70.37
}
```

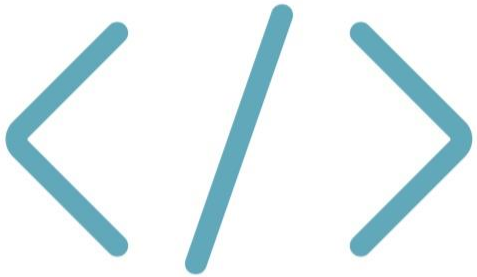

- conversions report


```
{
  “createdAt”: “2021-01-01T06:36:44+00:00”,
  “country”: “DE”,
  “account”: “98ae7c71-61c2-4197-849e-22da3cf9c771”,
  “campaign”: “d7750de6-91af-4937-ba4e-d14b9995a958”,
  “sellout”: 6186.76,
  “cost”: 61.03,
  “conversions”: 13
}
```


datasets



typical data strategy evolution



ingestion

code-based ingestion

synapse analytics

what it is - overview

consolidation alternatives

some of them, in the context on Synapse

Data Copy Activities

Data Flows

Spark Pools

Synapse Serverless

Synapse Dedicated (ex DWH)

consolidation alternatives

some of them, in the context on Synapse

Data Copy Activities

PRO: fast and intuitive

CONS: it's just an appender, issues with schema-bounded files (i.e. parquet)

Data Flows

PRO: fully managed and intuitive, spark-based

CONS: somehow rigid in configuration

Spark Pools

PRO: super flexible and fast

CONS: requires programming skills

Synapse Serverless

PRO: transformation made via SQL and billed per data scanned

CONS: not designed to be an ETL solution, just a Data Lake serverless bridge

Synapse Dedicated (ex DWH)

PRO: the PRO of a data warehouse

CONS: the CONS of a data warehouse

synapse analytics

consolidation - data copy

synapse analytics

schema drifting and aggregation - data flow

synapse analytics

spark pools - spark tables

pay attention

Stability of source

The degree of your source data can change over time

Fixed-Schema sources

The structure of your sources and related issues from the schema drifts

(de)Centrality

The risk your Master data is in more than one place

Updateability

The capability of your dataset to be updated or not after delivery


we have raw data
bronze level
transform for business
gold level
Now
Next
filter, clean, enrich
silver level
Data Lakehouse
medallion architecture
approach

NEXT STEPS

thanks