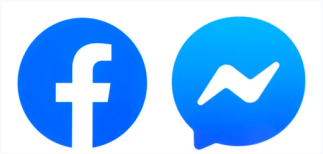
ManyChat

Этапы построения платформы аналитических данных в облаках

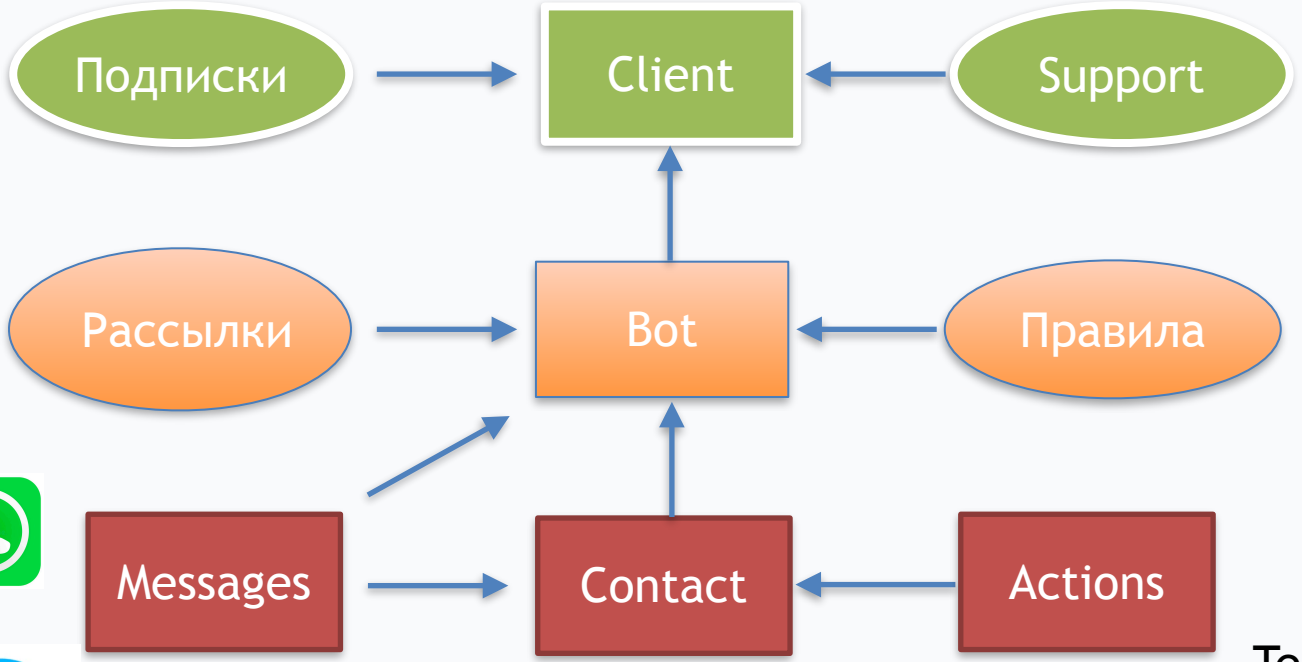
Голов Николай

July, 2021



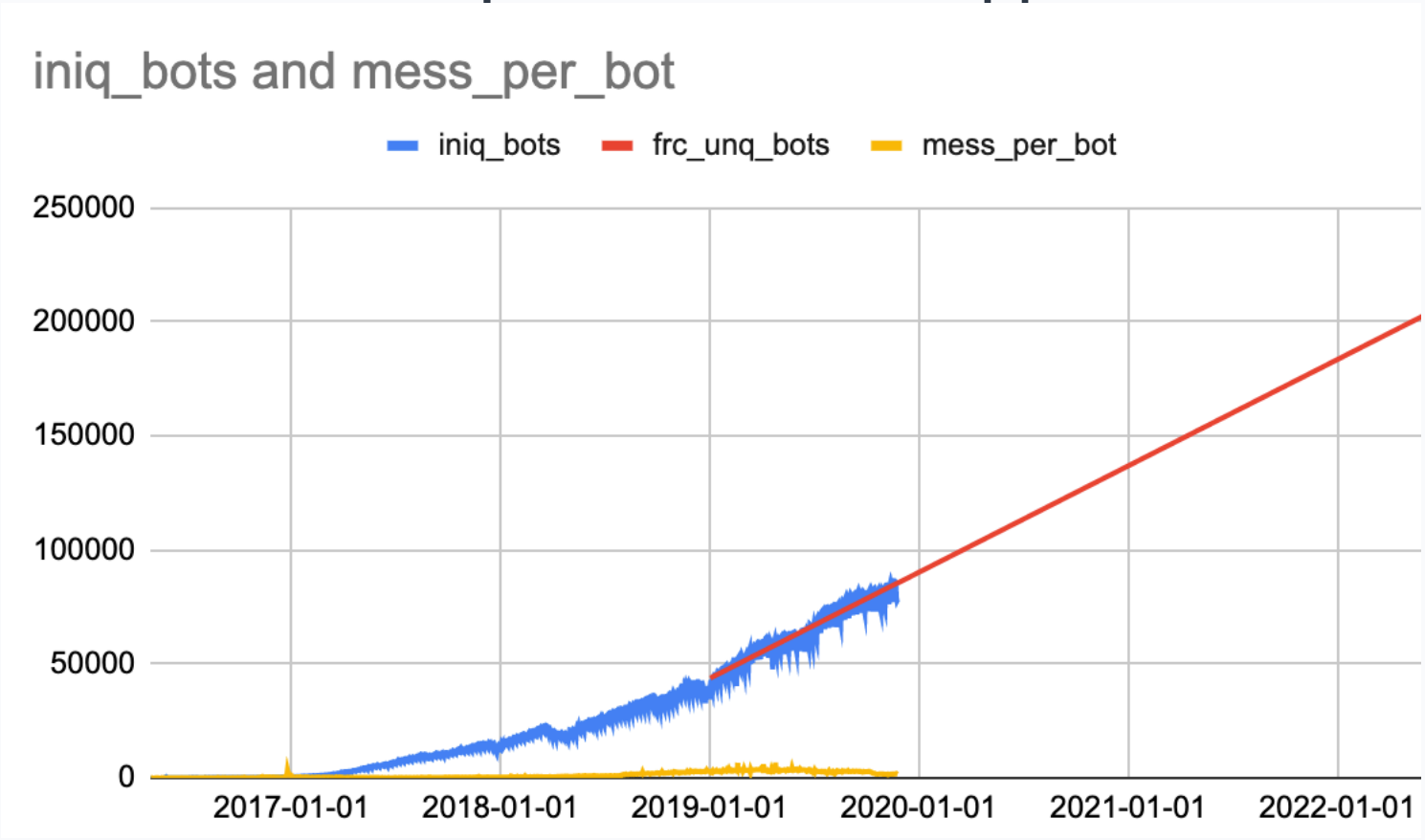
ManyChat

Basic Entities of ManyChat



ManyChat

Прогноз на 3 года



ManyChat

Нужна Analytical Data Platform

3. Data Platform: Аналитические задачи

- **Dashboards**
  - ANSI SQL
  - BI tools
- **Ad-hocs**
  - ANSI SQL
  - Python
- **ML/AI*****

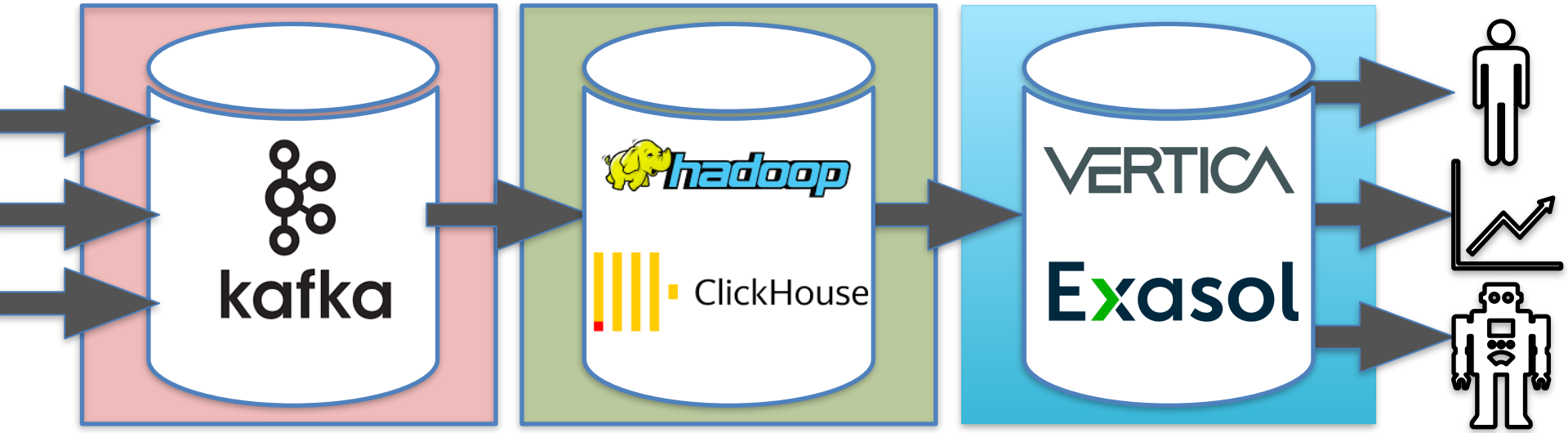
2. Data Platform: Data Consolidation

- Data Cleansing
- Data Matching
- Deduplication
- Historicity
- Normalization
- JSON parsing

1. Data Platform: Data Ingestion

- High-volume …
- High-velocity data inserts
- High volatility

Лучший инструмент для каждого блока?



1. Ingestion
2. Consolidation
3. Analytics

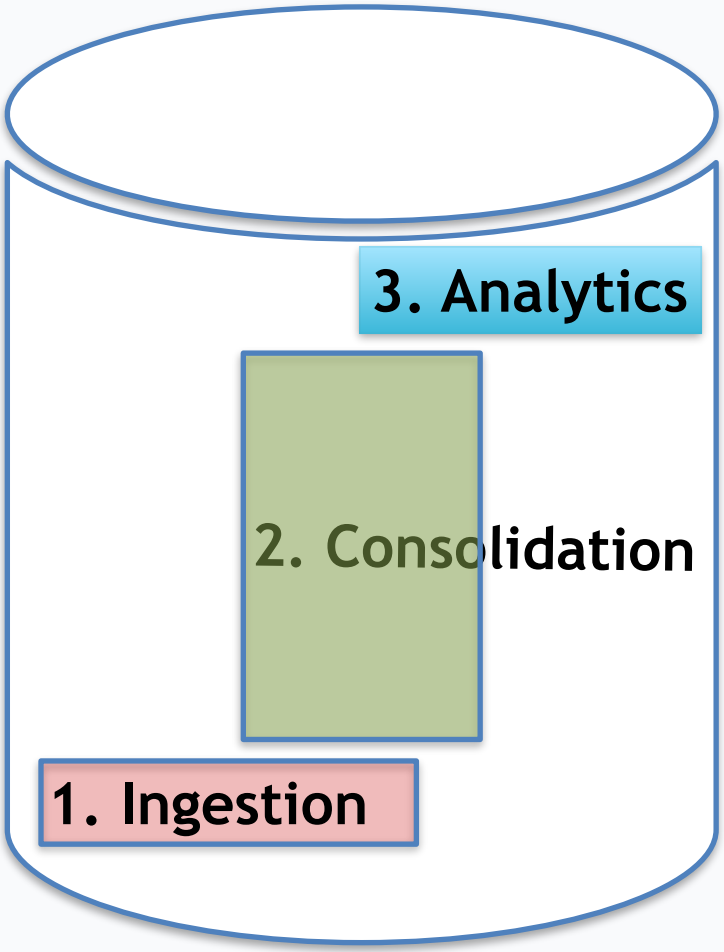
Все 3 блока - в одном инструменте?

- Agile:
  - легко расследовать 
  - легко исправлять 
  - Data Engineer ~ Аналитик ~ DS 
- Малые объемы - MS SQL, Oracle, Postgre SQL

Все 3 блока - в одном инструменте?

- Каждый блок - нагружает весь инструмент
- Аналитика - требует ANSI SQL + OLAP
- Consolidation - требует объемов хранения
- Ingestion - требует высокочастотных вставок

Все 3 блока:
в одном
инструменте?



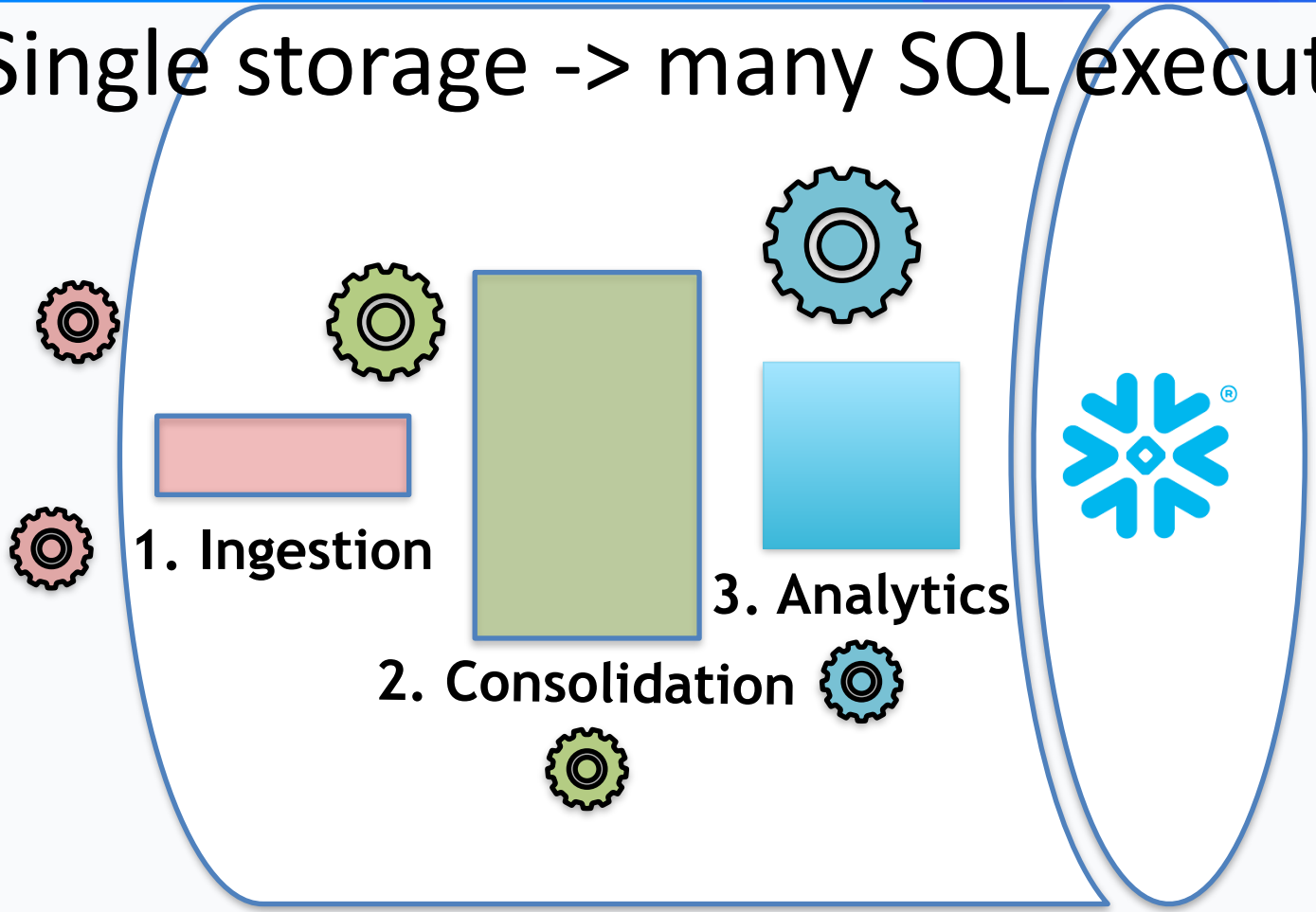
Postgre SQL/...?

Vertica/...?

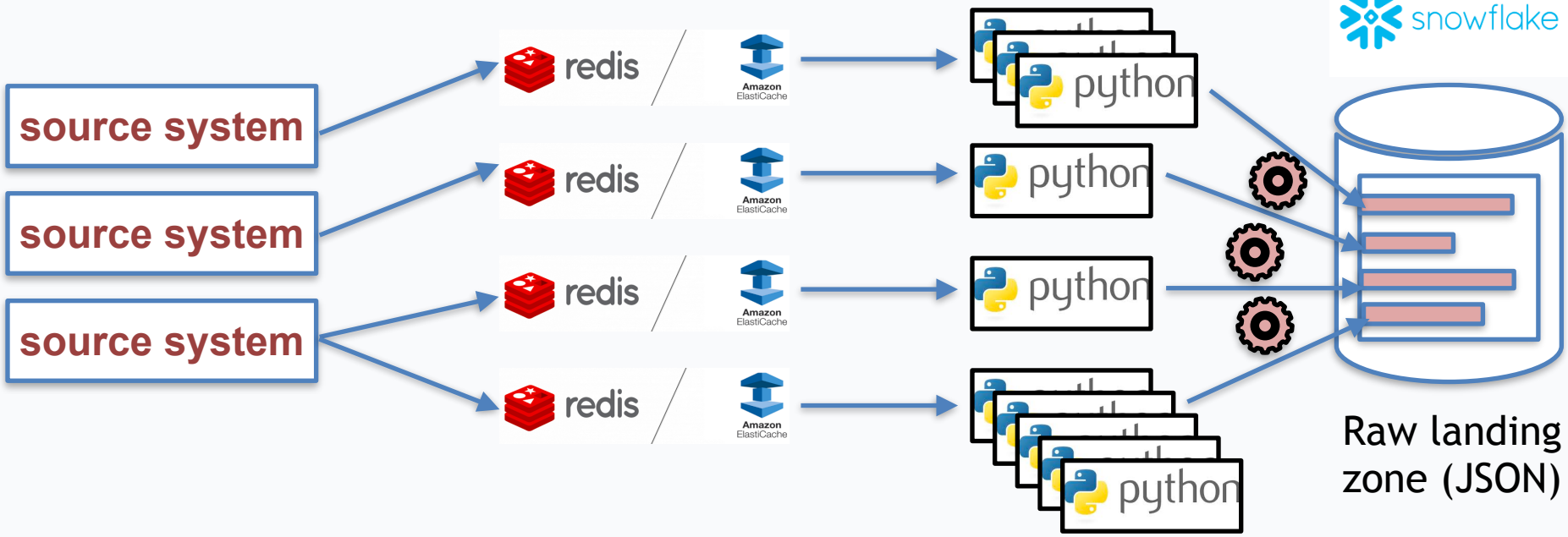
ClickHouse/...?

Kafka/...?

Single storage -> many SQL executors

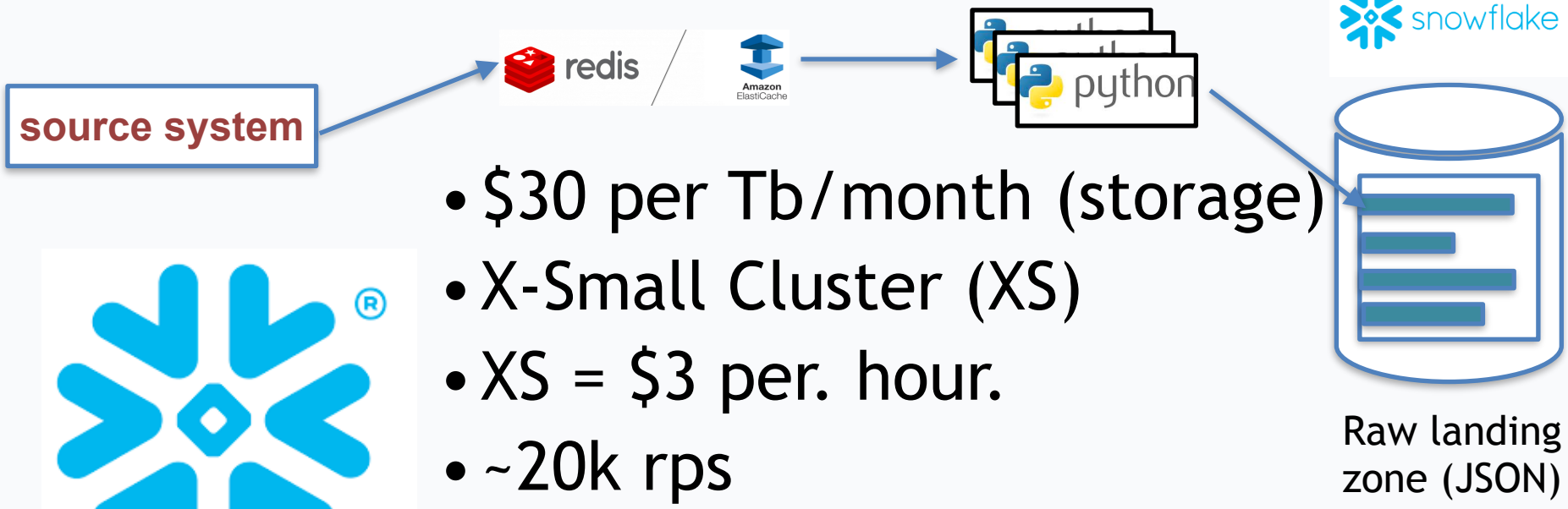


Snowflake: ingestion



- “Keep redis empty!”

Snowflake: ingestion



- $30 per Tb/month (storage)
- X-Small Cluster (XS)
- XS = $3 per. hour.
- ~20k rps
- Snow down -> files on disc
- … Data Lake!

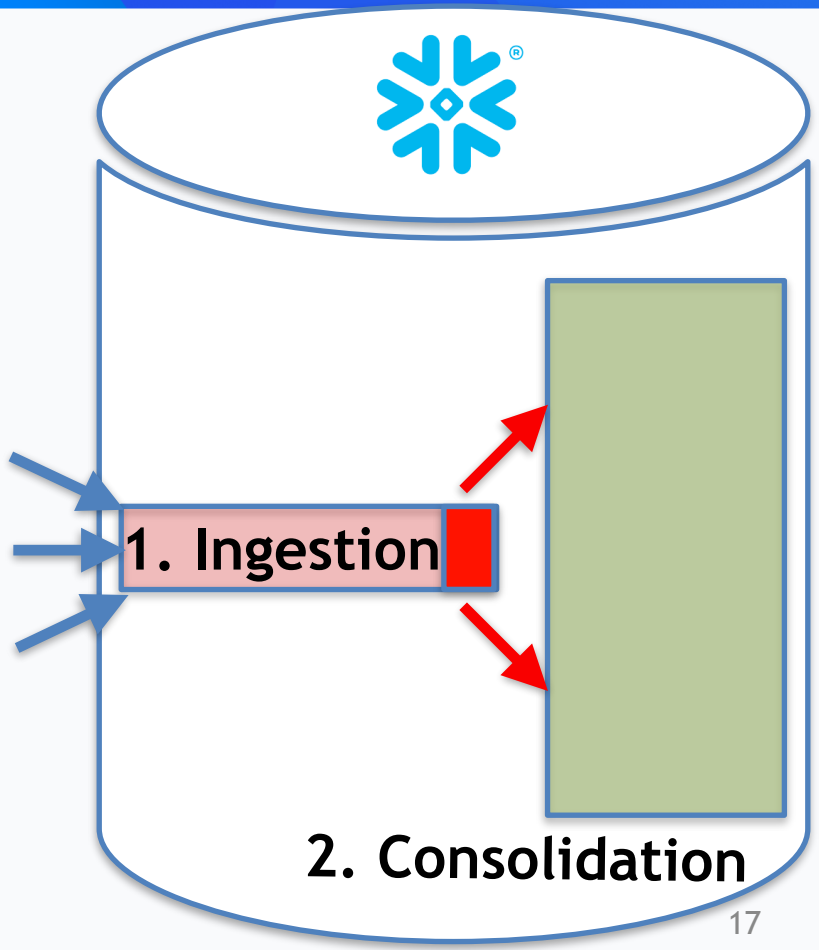
Ingestion data lake: game over?

- Snowflake SQL - super native JSON
- JSON = forget columnar DB
- All analysis - touches all data
- More data -> more time -> more money
- Schema on read….

=Super expensive Data Swamp

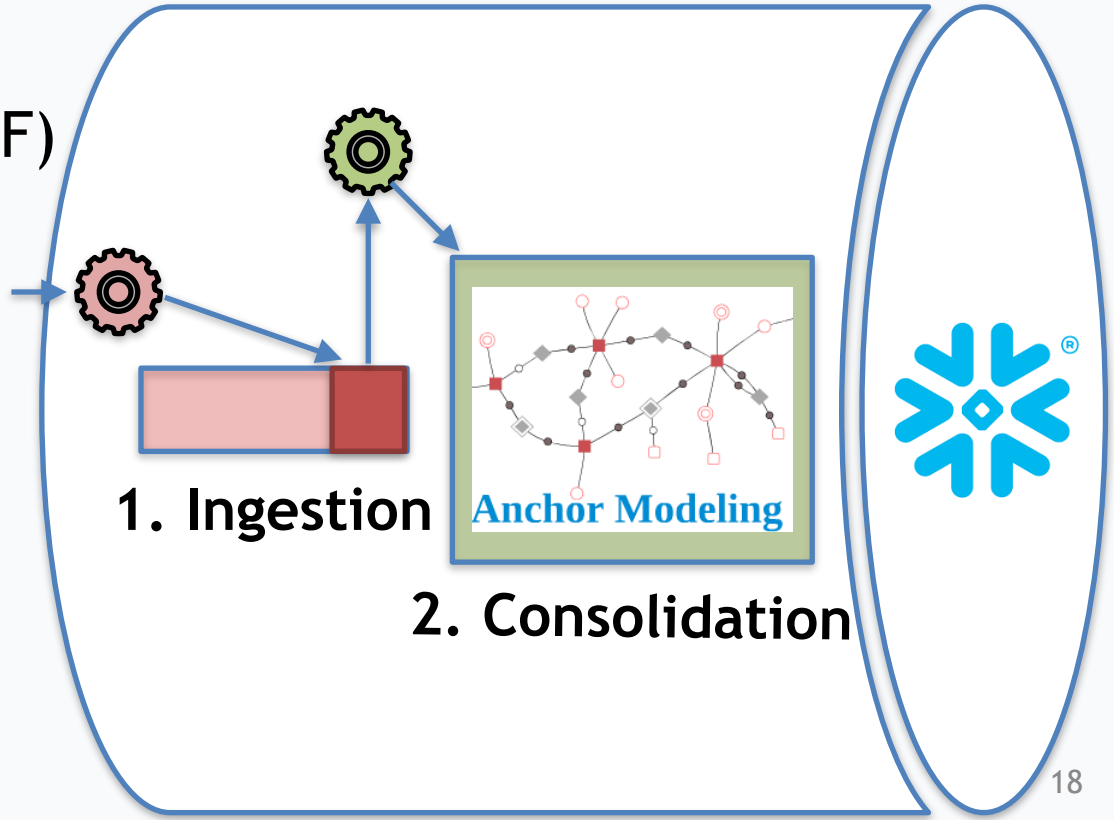
Snowflake: data consolidation

- Separate processing to ingest (blue)
- Separate processing to incrementally upsert to consolidation (red)
- Take only increment from Raw



Consolidation zone

- Anchor Modelling (6NF)
- Incremental load from ingestion
- Schema
- Historicity

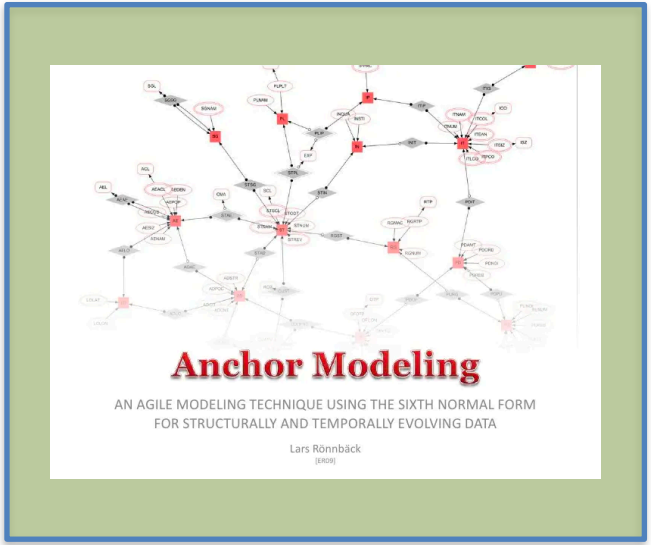


Consolidation layer: Anchor Modeling

- High normalisation - 6NF

- Table per Entity - Anchor
- Table per Attribute - Attribute
- Table per Relation - Tie

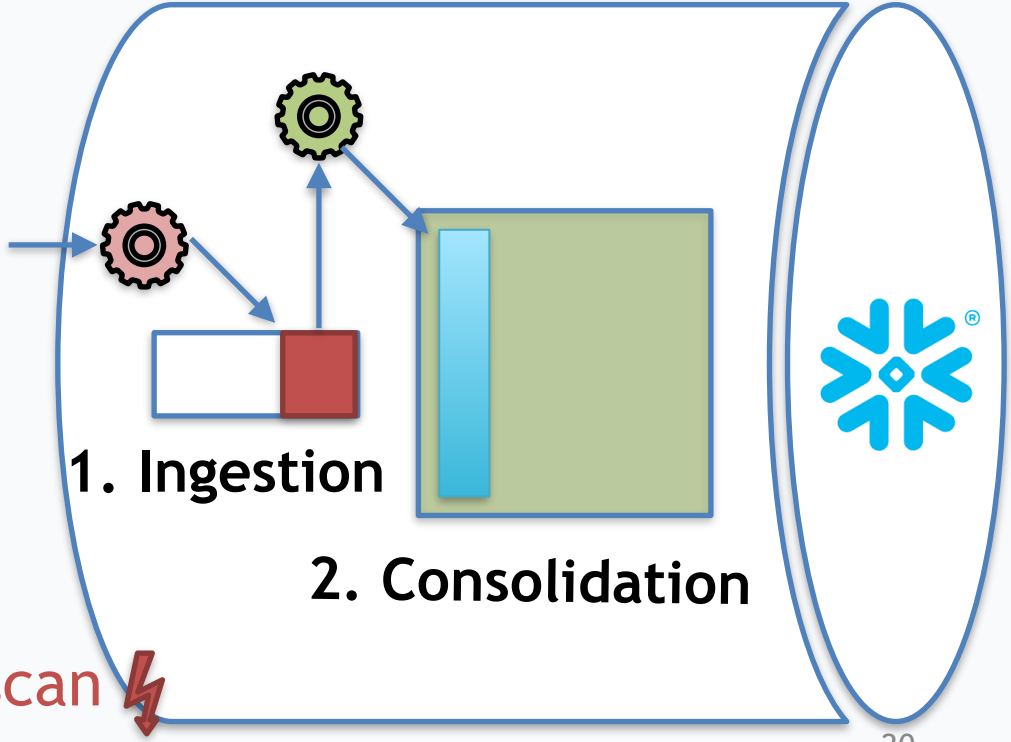
- Each table - as narrow, as light as possible.



2. Consolidation

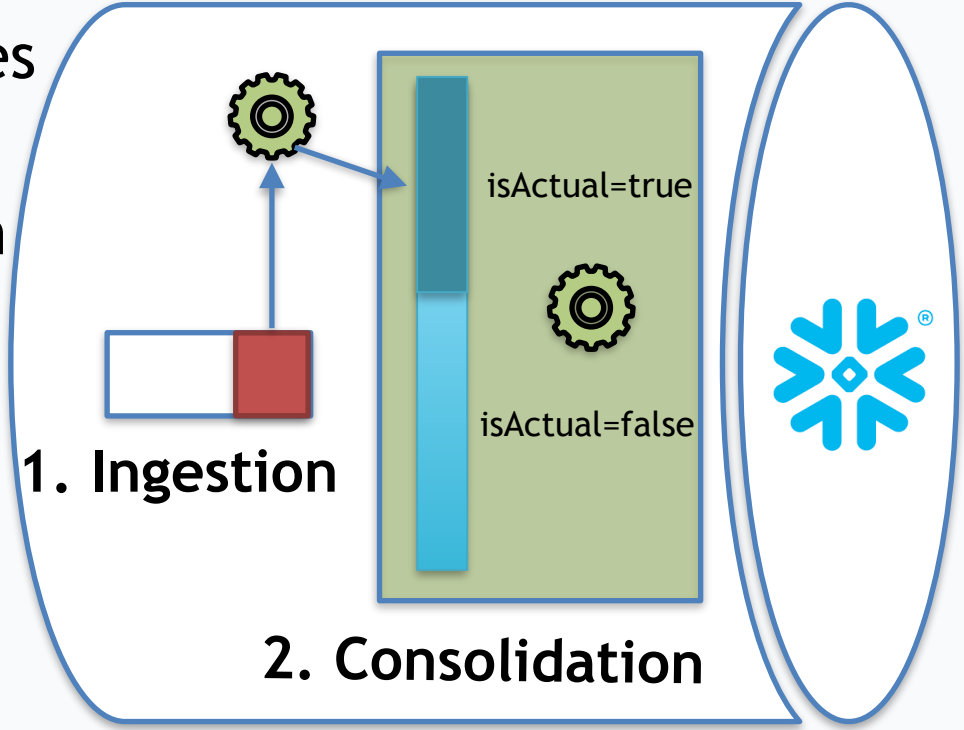
Anchor modelling in Snowflake

- Super-narrow tables
- Efficiently no index
- Only new values shall be inserted
- Each table, each load = full scan



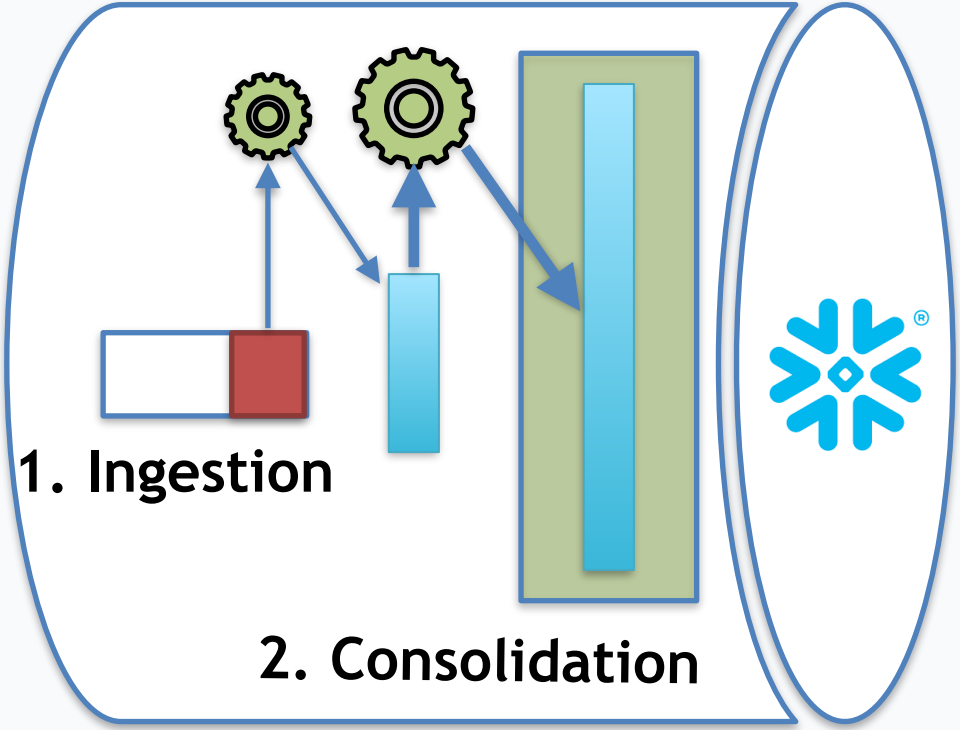
Optimising Snowflake: own pruning

- isActual field in biggest tables
- Table partitioned by isActual
- Ingested data compared with isActual=true part
- isActual regularly updated
- **for Anchors**



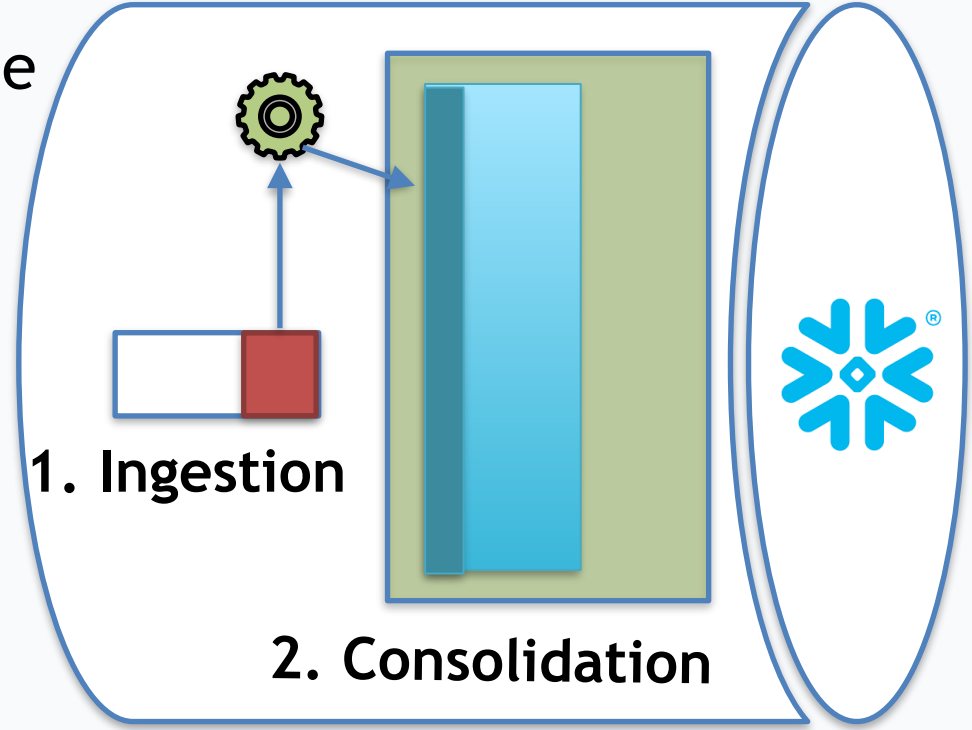
Optimising Snowflake: delayed loading

- Ingestion -> buffer table
- Buffer table small
- ~10 loads -> buffer merged into big target table
- truncate buffer
- 10 small reads + 1 big read
- for **Attributes, Relations**



Optimising Snowflake: hashing trick

- Attributes with huge text value
- Store value, hash(value)
- Ingested data compared with table data by hash…
- just 8 bytes per row…
- **for Attributes**



Snowflake: data preparation: DDS

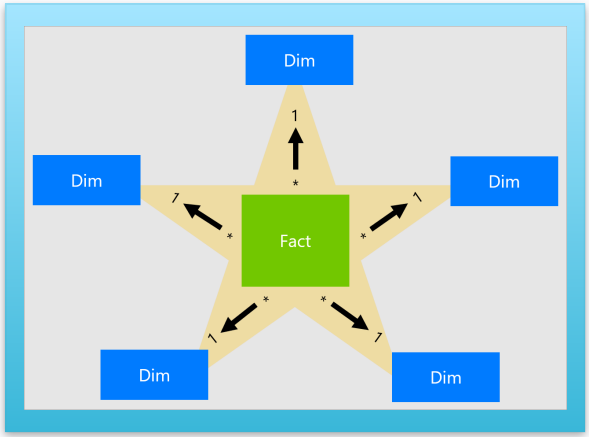
- Anchor Modeling (6NF) in S3
- $30 per Tb/month
- Few XS and S clusters
- XS = $3 per. hour.
- S = $6 per. hour.
- … **Data Warehouse?**

6NF Anchor Modeling DDS: game over?

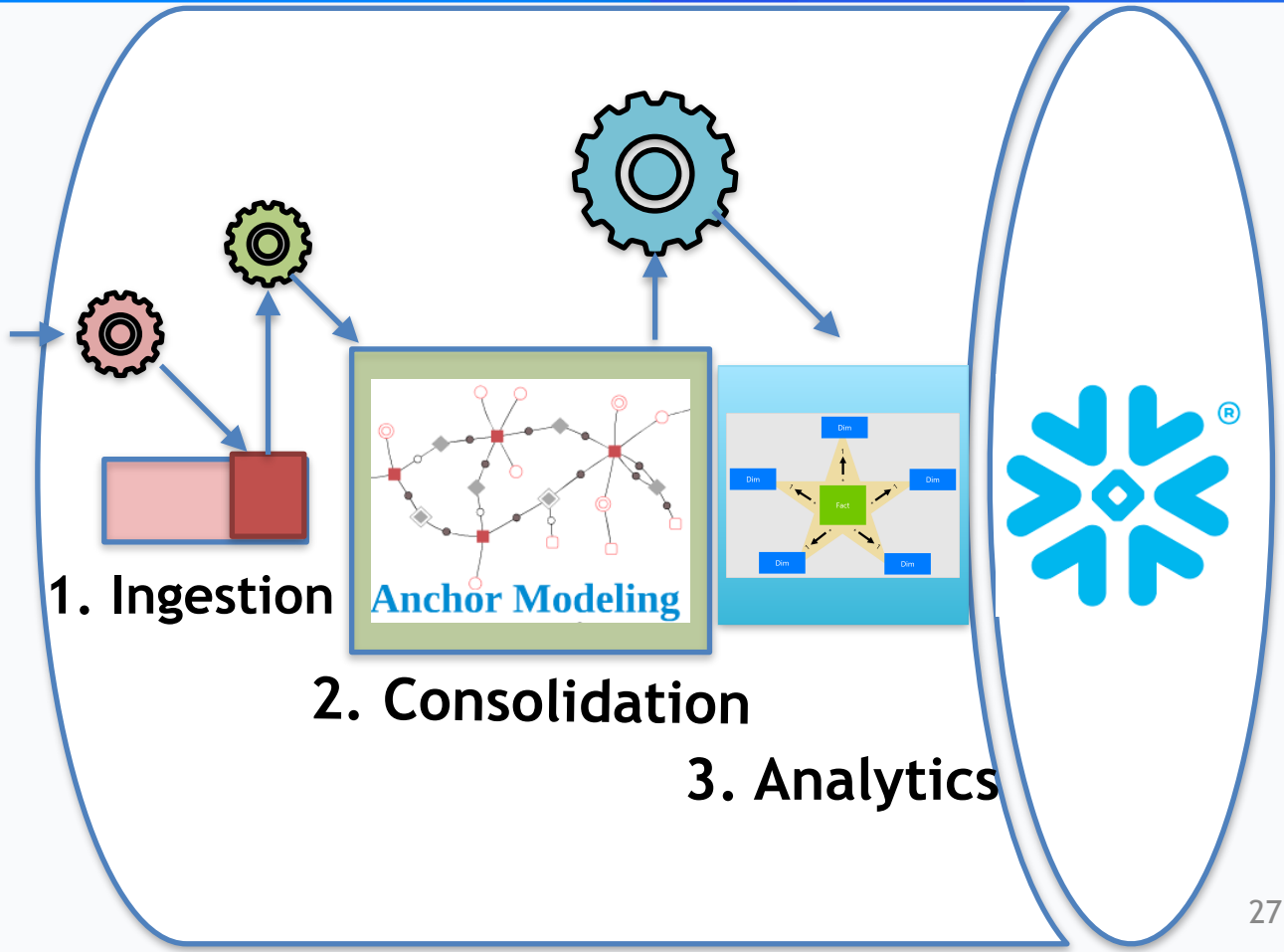
- Columnar DB, ANSI SQL Joins
- Many Joins needed
- Huge tables (~100 bln. rows)
- **Super expensive Data Warehouse**

Snowflake: data analysis: data marts

- Denormalised data marts
- Business oriented data marts
- Big marts - incremental ones



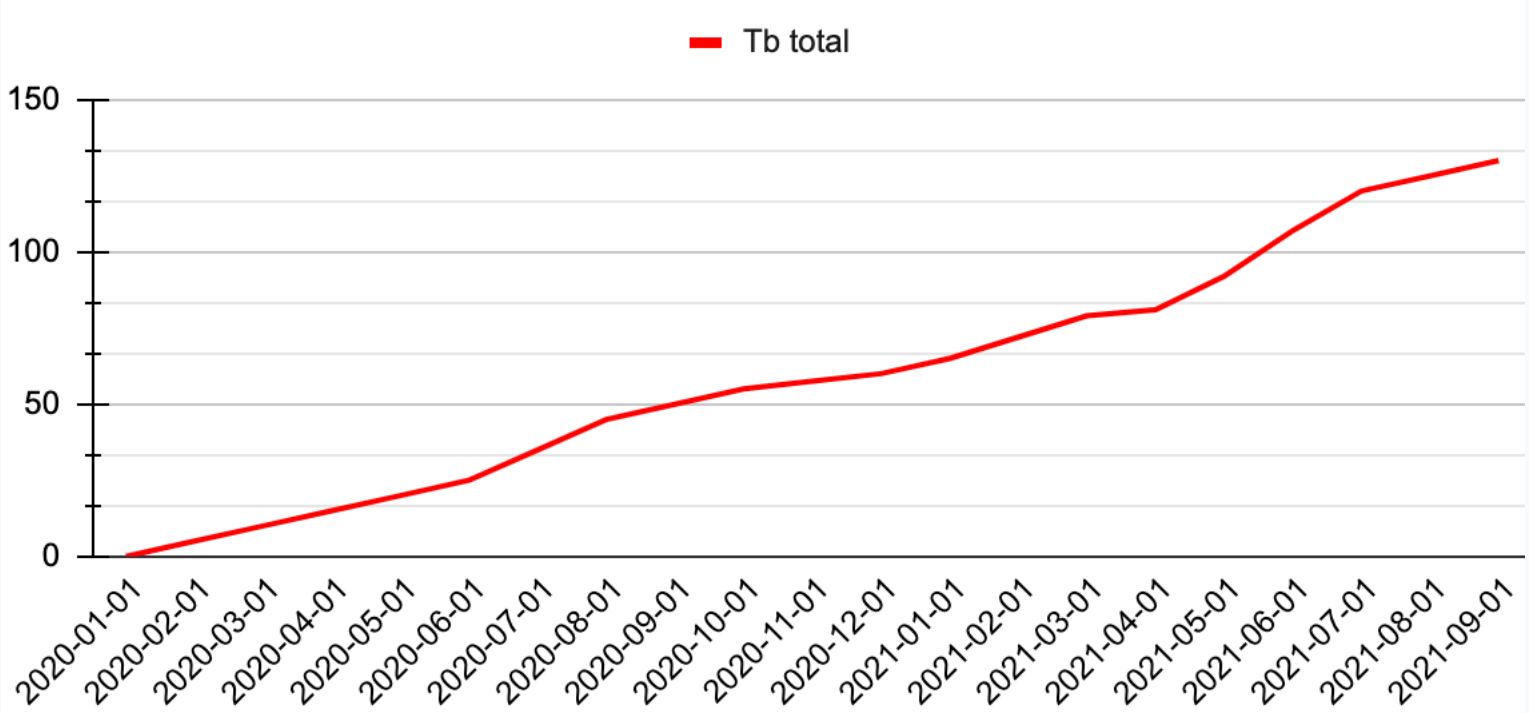
3. Analytics


1. Ingestion
Anchor Modeling
Dim
Dim
Fact
Dim
Dim
Dim
2. Consolidation
3. Analytics

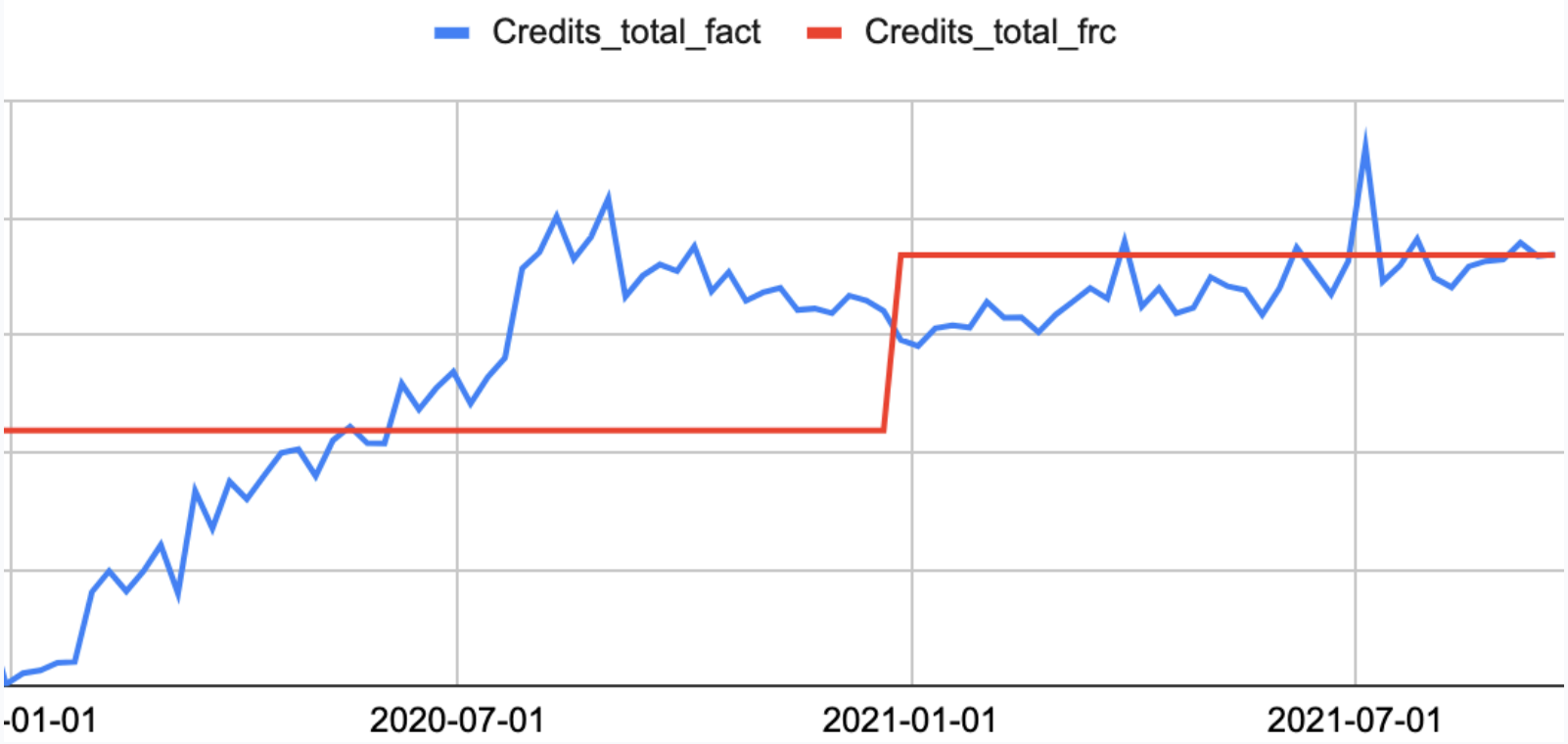
Snowflake: data analysis: Data Marts

- Denormalised data marts
- S + M clusters for daily materialisation
- S cluster for Tableau + analytical ad-hocs
- S = $6 per. hour., M = $12 per. hour

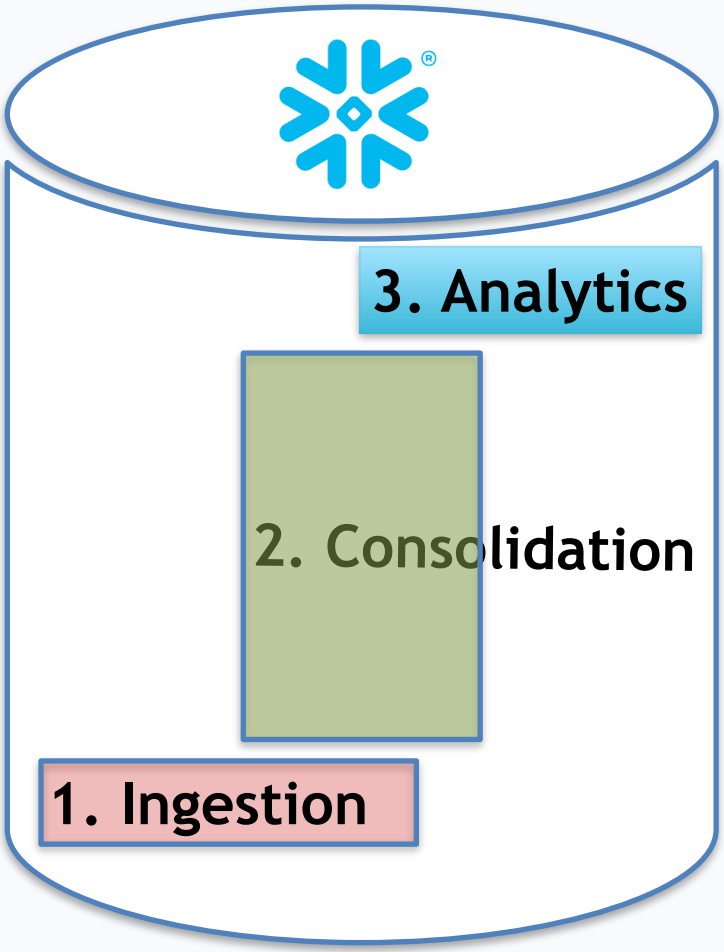
ManyChat data growth



ManyChat Snowflake credit usage graph



Все 3 блока:
в одном
инструменте!
(независимы)



Thanks!

Всем спасибо!

Голов Николай
azathot.mail@gmail.com
nikolay@manychat.com

Tack!

谢谢