DATA QUALITY

1 'Data Quality' meanings

~ Good pipelines start with good documentation ~

DISCOVERABILITY

Decision makers need to know if data is available & accessible (catalogued).

Data cataloguing is an infrastructure problem, not a data engineering one (but both closely related).

CLEAR/ COMPLETE DEFINITIONS

Even with good data, incomplete data understanding can lead to bad business decisions.

Black Swan events can expose how complete definitions might still miss crucial contextual factors.

ie. Zillow ~$500Mio loss due to lower valuation of property buy AI-powered models. Lack of clear definitions/ data silos lead to overvalued property purchase decisions.

DATA INTEGRITY

Consistency & reliability on your datasets. Prioritize:

- Absence of NULLs
- No duplicates
- Naming conventions (ie. Airbnb uses 'dim.', 'fct.', 'm.' in all its dimension, fact, measure tables)

BUSINESS VALUE GENERATION

Monetary impact: revenue generation, cost saving (ie. AWS spending optimization), pipeline efficiency measurements.

Indirect value: lead indicators for revenue; multi-step value chain.

ie. Google Ads ⟶ Facebook signups ⟶ Feed engagement ⟶ Revenue

USABILITY

Making data "easy to use" means anticipating how others will interact with it.

Set practices that lead to data democratization (ie. clear column names, sensible data types).

Foster self-documenting structures.

TIMELINESS

Ensure data arrives at a timely manner.

Data Engineering & Data Analytics face issues with agreed data refresh interval delays.

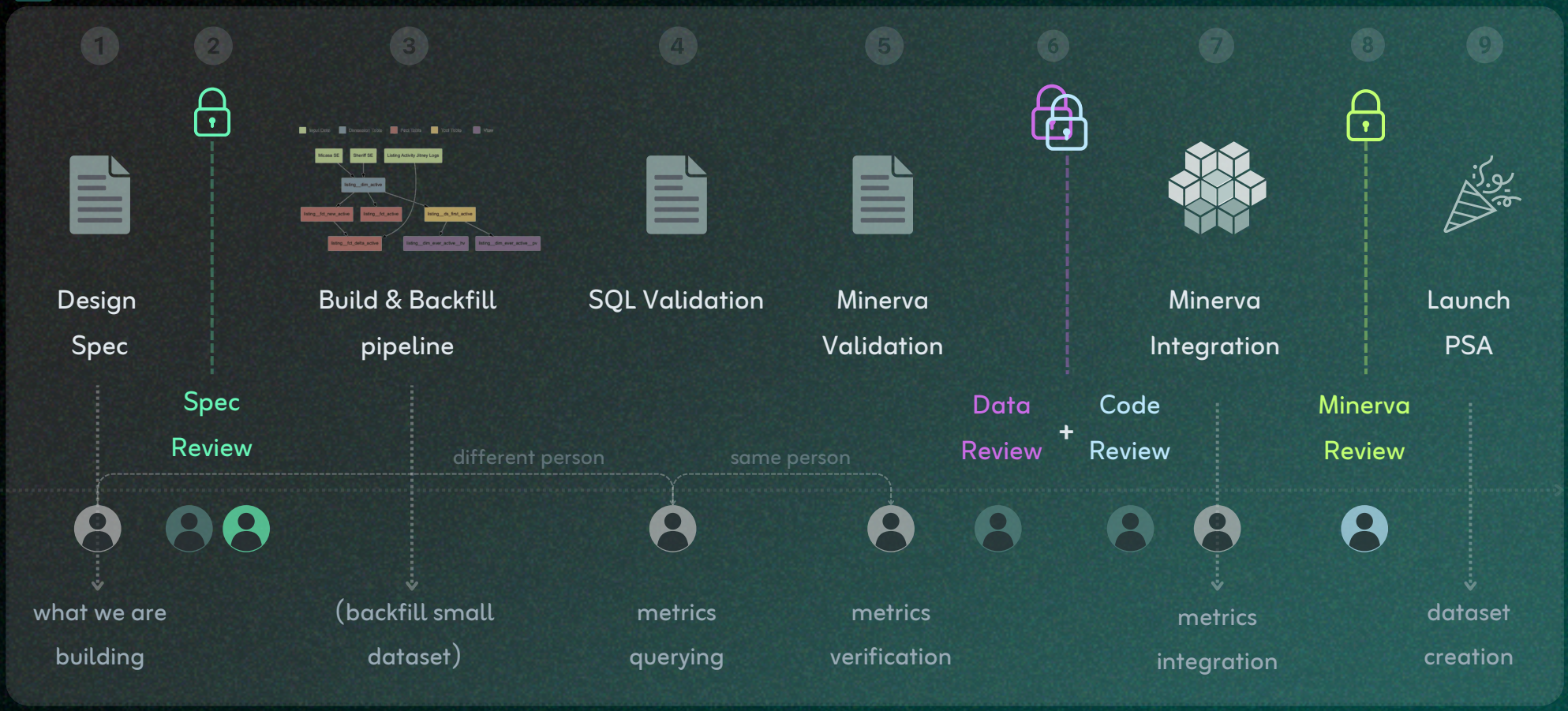
Process management: clear SLAs with analysts; streamlined request handling; set efficient troubleshooting procedures.

DATA QUALITY

2 The MIDAS Process (Airbnb ~ 2020)

Data quality (DQ) certification protocol that ensures quality & timeliness of critical data sets & metrics.

Process flow



Sequential review process:

- **Spec**
  - Evaluates data model proposed design specs prior to implementation. Ensures fit to standards & dev feasibility.
  - Co-led by a Data Architect (technical review) + a Business Stakeholder (scope & completeness review)
- **Data**
  - Ensures accurate, reliable data in the data pipeline through DQ checks & nothing was omitted in spec review.
  - Led by a Data Architect
- **Code**
  - Checks code quality, efficiency and adherence to standards (ie. check both unit & integration tests are captured).
  - Led by a Data Architect
- **Minerva**
  - Verifies the source of truth metric definitions implemented in Minerva (Airbnb's metric service)
  - Co-led by the Chief Data Scientist

Process benefits

- Pre-built stakeholder buy-in increases trust
- Prevents backfills when missed requirements

Process Challenges

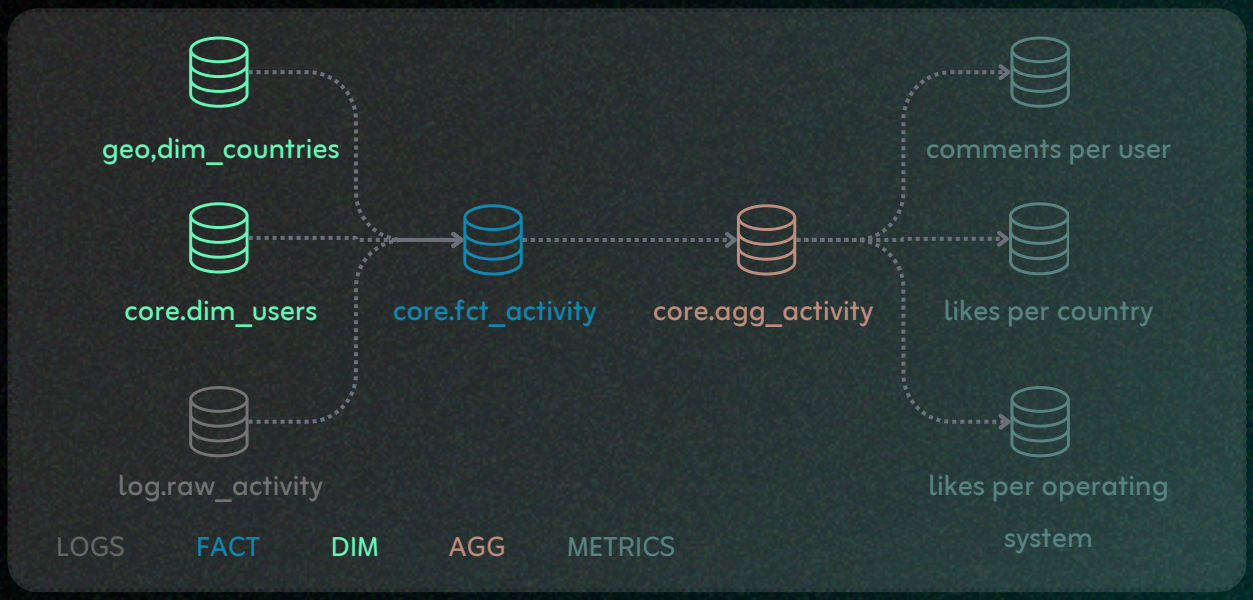
- **Non-scalability**: resource-intensive process, may not scale well as the volume of data and number of data assets increase
- **Frontline resistance**: significant investment required can led to resistance from the frontline data team, (process percieved as burdensome).
- **Resource investment**: meeting the full data quality criteria necessitates a considerable investment in design, development, validation, and maintenance (challenging to sustain).

MIDAS is a success on how to build a solid golden pipeline that sticks for long time, people understand/ know to to debug/ improve it

DATA QUALITY

3 Design Spec: elements to make it good

✓ Flow diagram



✓ Schema

DDL statements.

CREATE TABLE statements.

Other:

- Good names ('fct', 'dim', 'sdc', 'agg' in included in columns names)
- Comments (in every column)
- Follow company naming conventions
  In case no naming conventions exists, take the initiative & set them

✓ DQ check levels (3 levels)

Basic

Fundamental validations required in every data pipeline.

- Is there data? (basic volume checks)
- NULLs (sign of upstream data collection issues).
- Duplicates (critical to maintain data integrity).
- Enum values all valid? (more critical in Data Lake environments where contraints aren't enforced).

Intermediate

Row count analysis

- Week-on-week consistent trend.

Dimensional analysis

- Break down metrics by key business dimensions; helps isolate issues to specific segments (ie. order counts by country/ product category,..).

Business rule enforcement (validates domain-specific logic; ensures business constraints maintained).

Advanced

Seasonality adjusted counts

- Compare metrics accounting for seasonal patterns (more sophisticated than simple week-over-week; ie. this Xmas vs. last years' Xmas sales comparison).

Machine learning integration

- Use ML for anomaly detection.
- Higher signal-less noise checks but more costly to maintain.

✓ DQ check types (2 types)

Dimensional

- Growth patterns are either flat or steadily increasing.
- Simpler to implement than fact table checks.
- Should avoid sharp changes. Use percentage thresholds (ie. you don't expect customer dimension to double in a week).
- Generate fewer false positives.

Validation rules:

- Basic daily growth checks ("Is table growing?").
- Reasonable growth rate validation.
- Complex relationship verification (ie. Facebook friends count limits).

Fact

- Growth/shrink based on user behavior.
- More prone to duplicates in logs & NULL values.
- References to dimensional entities must exist (ie. no notifications sent to Facebook deactivated users).

Special considerations

- Seasonality (week-on-week best; holiday periods).
- Data Infra: more tolerant to NULLs; more flexible SLAs.
  In Presto, use APPROX_COUNT_DISTINCT over COUNT(DISTINCT) (99.9% accuracy but more efficient; key for high-volume data analysis; prevents query timeouts

✓ Metric definitions (what are you trying to measure?)

✓ Example queries (show how things get done)