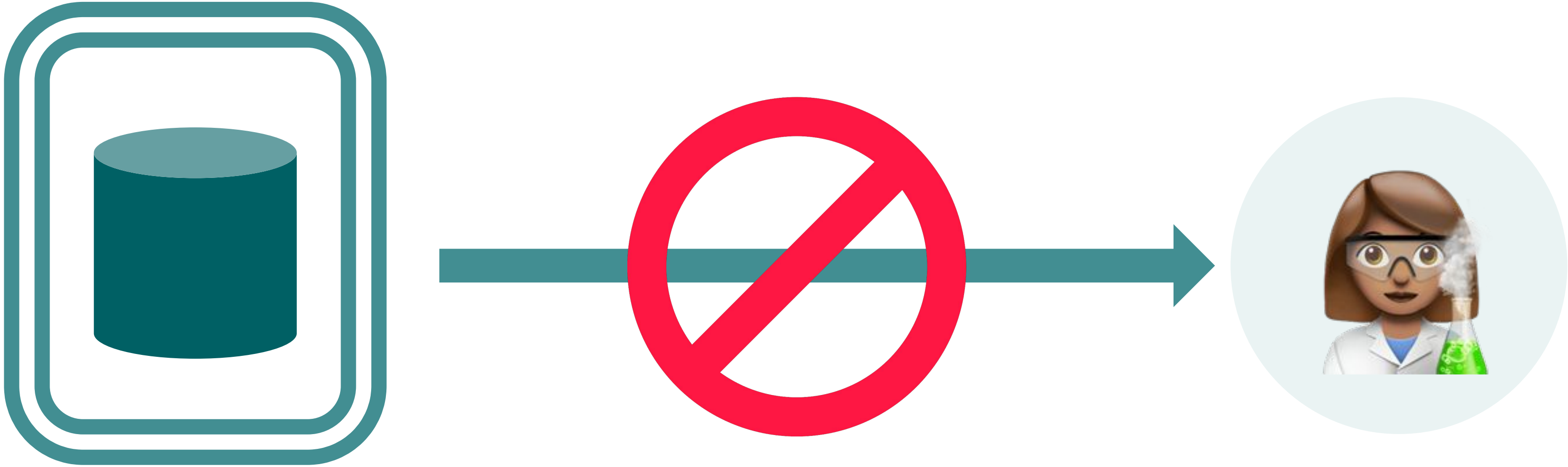
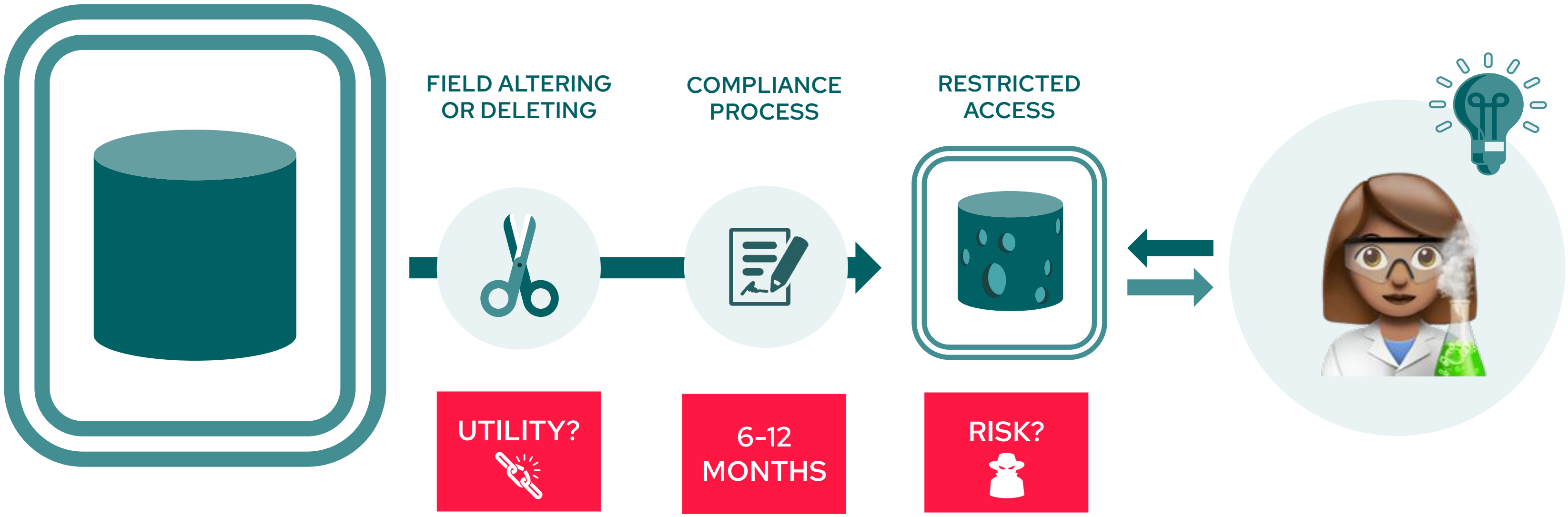
sarus
technologies

Unlock All Sensitive Data Assets with the Gold Standard of Data Protection

Sensitive data is hardly accessible by design



Accessing it means months-long processes



FIELD ALTERING OR DELETING

UTILITY?

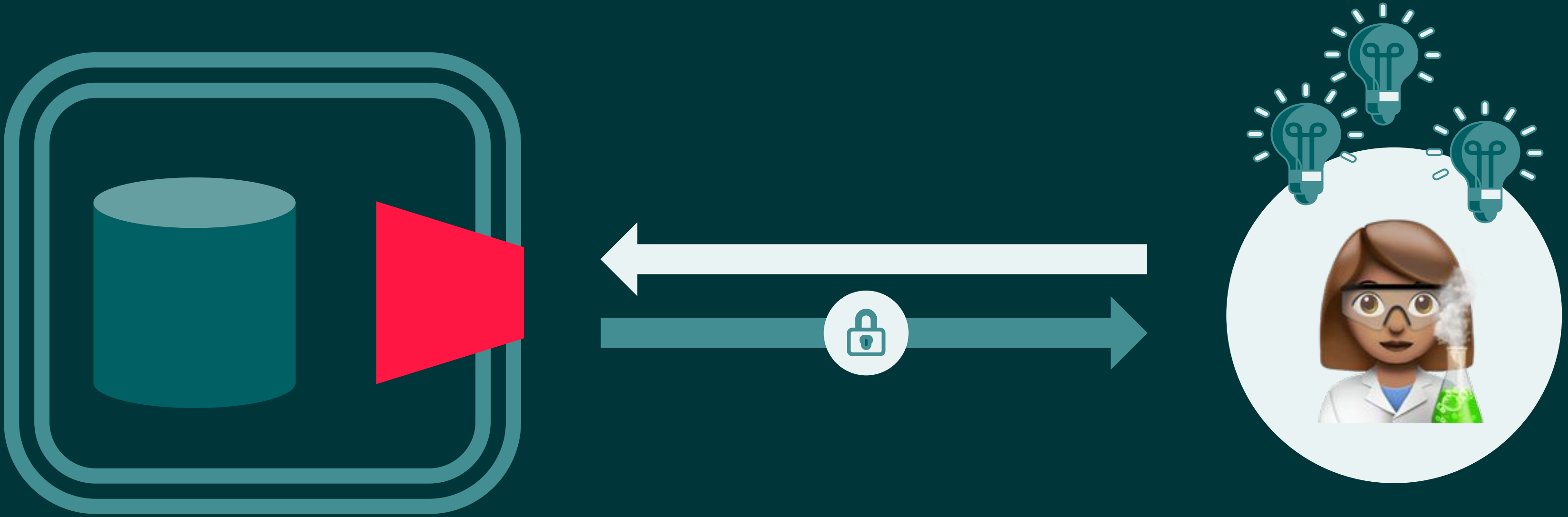
COMPLIANCE PROCESS

6-12 MONTHS

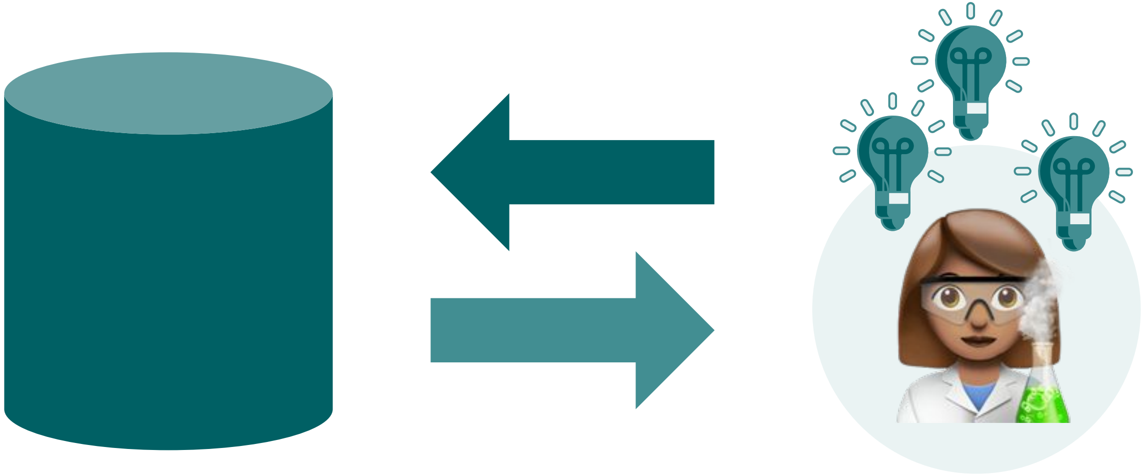
RESTRICTED ACCESS

RISK?

With Sarus, leverage original data directly with state-of-the-art data protection



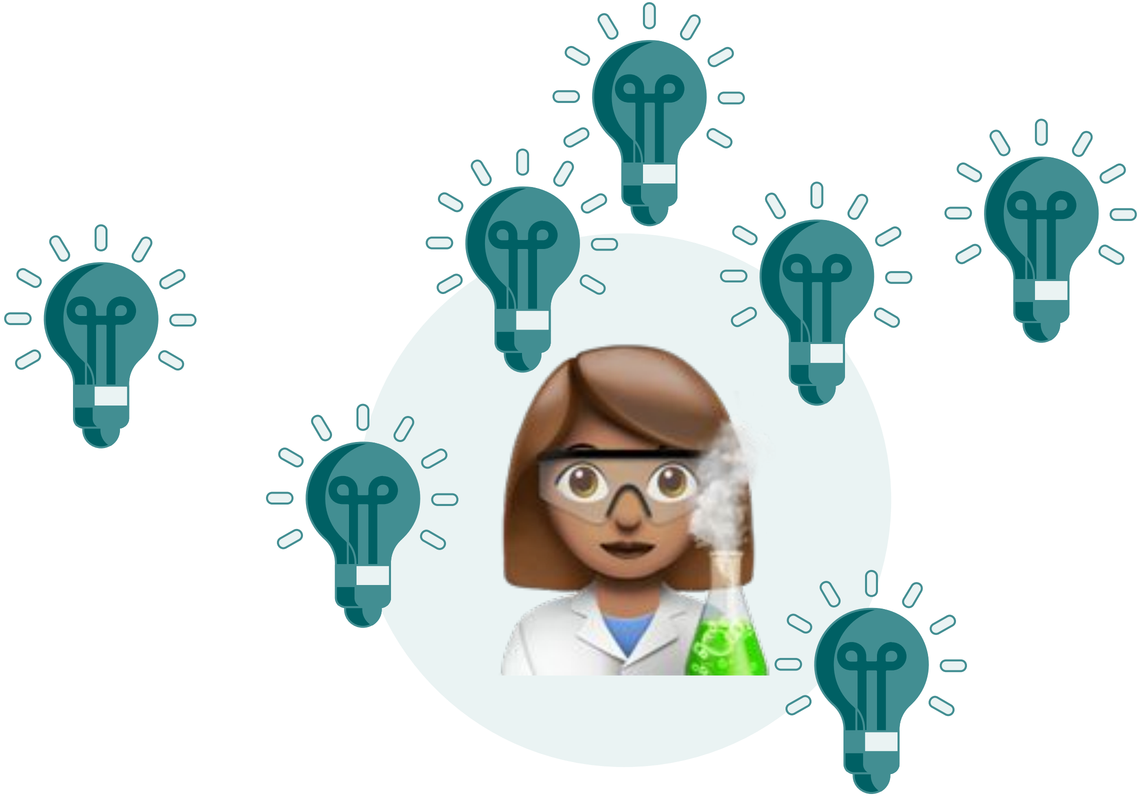
Instantly Empower Data Teams



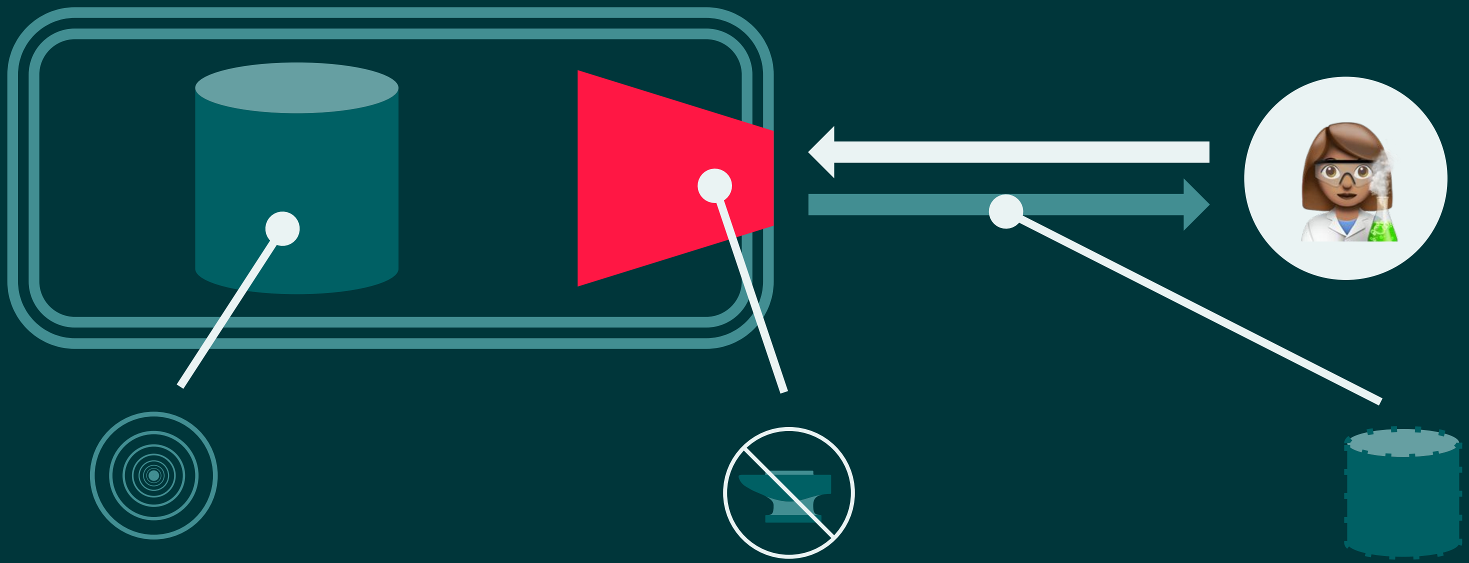
DATA ASSETS ACCESSIBLE INSTANTLY

Data Collaboration Made Easy

Powerful Results from High Fidelity Data



Privacy-first Gateway to Sensitive Data

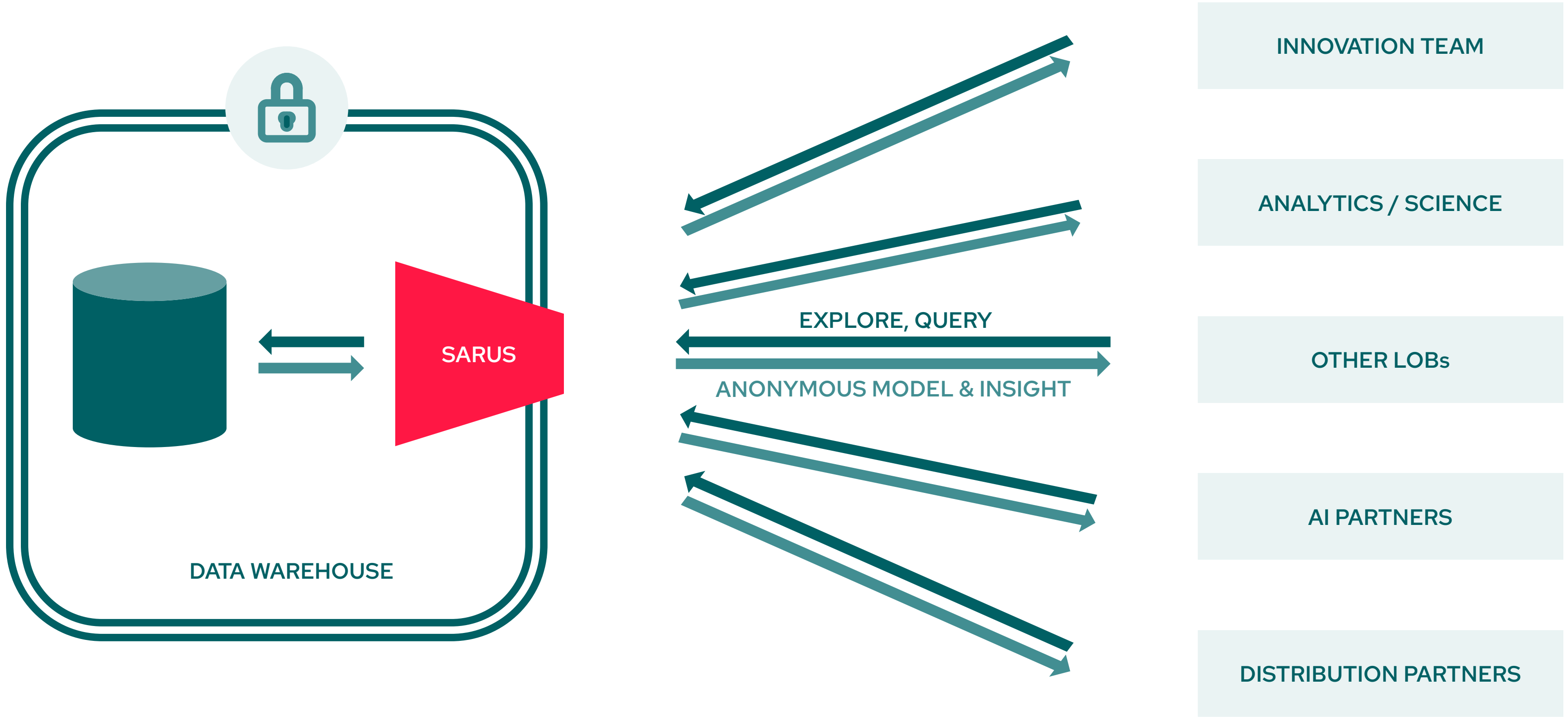


No code
No copy or custom engineering, use original data directly

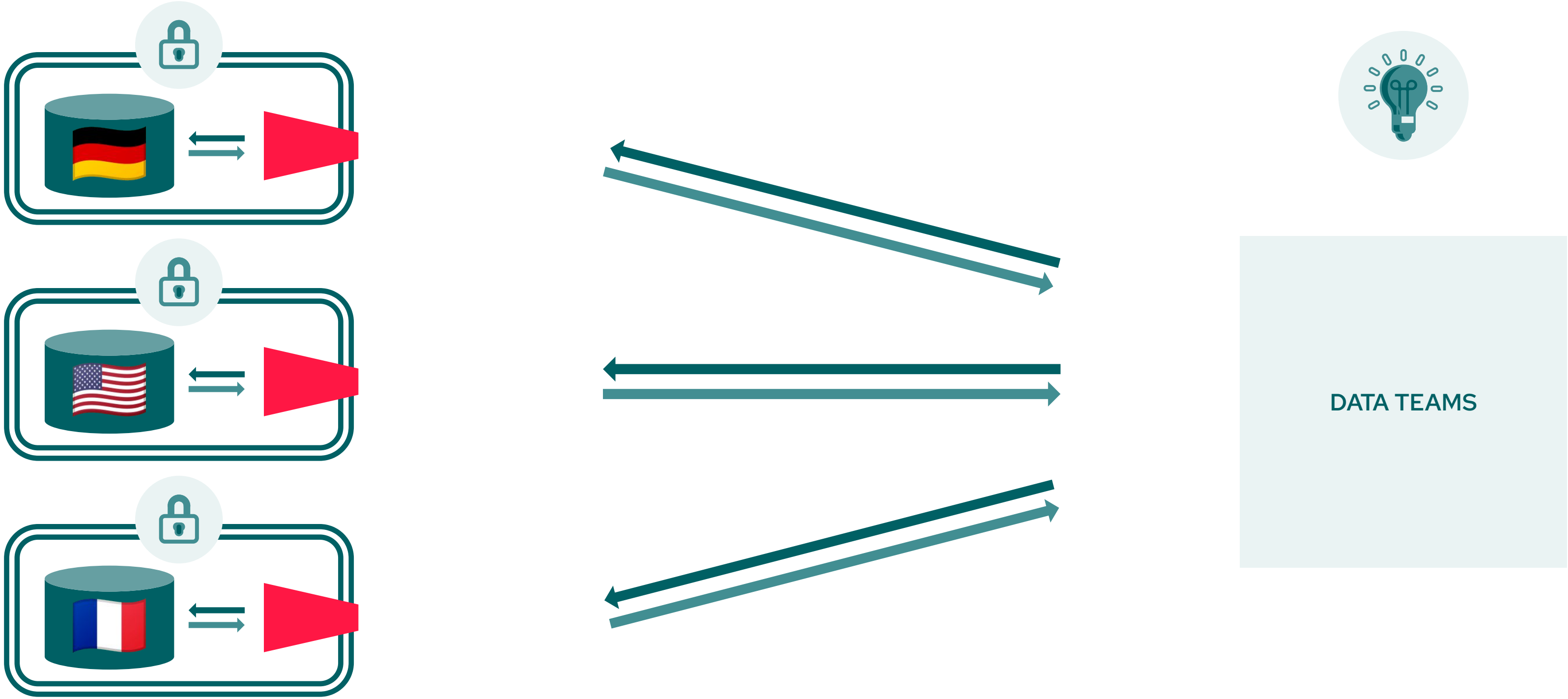
Scale compliance
Math-based privacy policy templates to speed up all compliance processes

Synthetic data
Available natively to ease all preparatory work and tests

The Foundations of a Data-centric Platform



Innovation across silos & regulatory borders



Appendix

Differential privacy: the gold standard used by Apple, Google, or the US Census

$$\Pr[\mathcal{A}(\mathcal{D}) \in \mathcal{S}] \leq e^{\epsilon} \Pr[\mathcal{A}(\mathcal{D}') \in \mathcal{S}] + \delta$$

Protection guarantees:

1. For all data types, no matter how sensitive it is
2. For all learning objectives
3. Whatever the receiver may already know

> “*2020 US Census results will be protected using* differential privacy, *the new gold standard in data privacy protection.*”

source: https://www.census.gov/

Sarus vs Synthetic data

Sarus main technological proposition
Sarus makes it easy for an analyst/data scientist to work on a data source that cannot be accessed directly. Synthetic data and the rich feature set that includes SQL/ML/pandas combined with native synthetic data samples makes the experience seamless. Sarus provide mathematical guarantees with differential privacy.

Attacker model: the data owner does not trust the data practitioner with personal data.

Comparison with Synthetic data
Synthetic data focuses on generating a fake dataset that mimics the statistical properties of the original data. It should have the same structure and distributions but not reveal individual records.

Key differences:

- There is no guarantee that insights derived from synthetic data is close to the same insights on the original data. It can happen by chance but is not suitable for decision making.
- Sarus does provide synthetic data but limits its usage to exploration and preliminary analyses. Insights are eventually derived from the original data.
- Synthetic data may still leak information on individual records if it is not produced with differential privacy

Sarus vs legacy anonymization methods

Comparison with legacy anonymization methods (aka data masking)
In traditional methods, some habilitated data engineer is responsible for altering and redacting the original data to make re-identification harder. It includes the deletion of obvious identifiers (names, social security numbers, ids, addresses, birth dates…) and the alteration of less obvious ones (precise dates, series of events, unique combination of features).
Each data comes with an ad hoc anonymization strategy and compliance will need to approve it each time.

Key differences:

- Strong anonymization requires deleting almost everything in the data, destroying utility
- Weaker anonymization means a lot of requirements on the attacker model (e.g.: the data practitioner should not try to connect the data with external sources)
- Either way it requires bespoke data engineering and compliance processes

Sarus vs Federated learning

Comparison with Federated learning

FL focuses on working from multiple sources whereas Sarus's main use case is a single source (though it can be applied to multiple sources but not seamlessly)

Attacker model:

- data owners do not trust one another or the data practitioner or a third party to host their data
- they do trust the data practitioner to only do analyses that do not leak personal information
- they also trust the other data owners to input their true data so that the output is a mix of everyone's data

Key differences:

- FL focuses specifically on machine learning whereas Sarus covers more use cases (analytics, classical machine learning)
- FL usually does not bring DP guarantees which requires significant trust in what the analyst does
- FL typically does not come with synthetic data so building models may be challenging
- The attacker model is a bit convoluted

Sarus vs SMPC

Comparison with SMPC
SMPC focuses on doing light computation on distributed datasets. It protects the contribution of each source during computation.

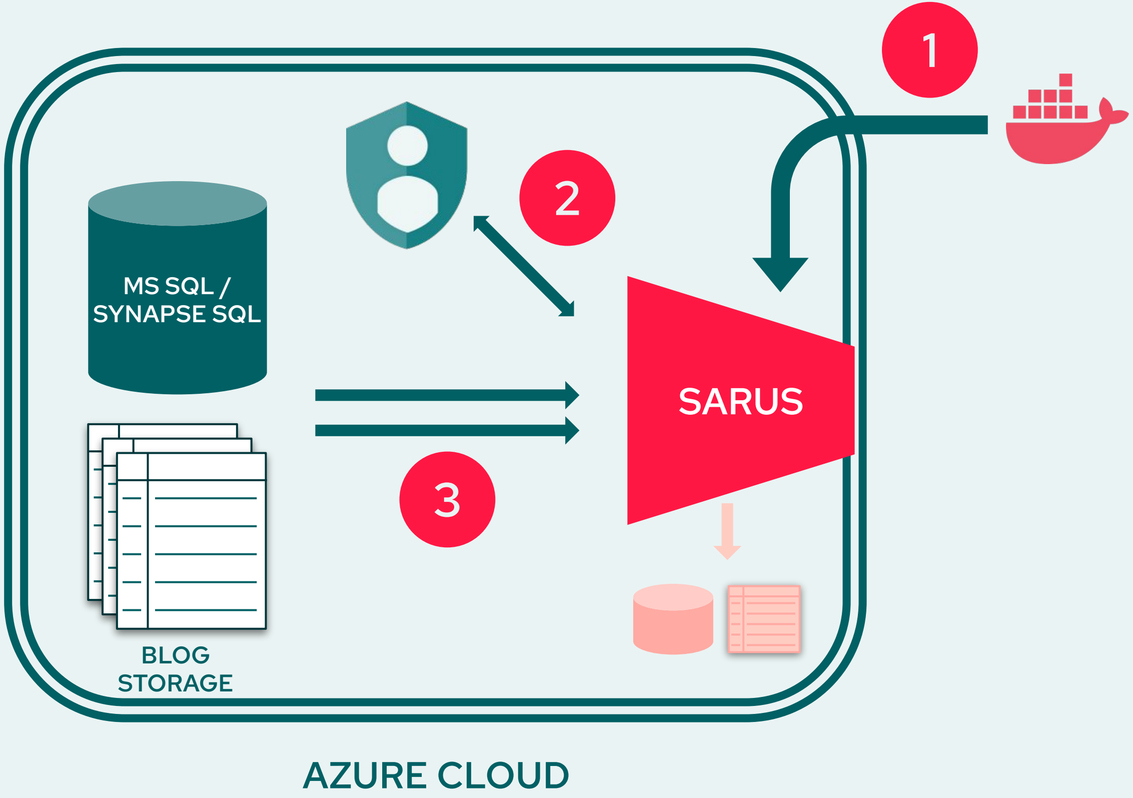
Attacker model:

- data owners do not trust one another or the data practitioner or a third party to host their data
- they do trust the data practitioner to only do analyses that do not leak personal information
- they also trust the other data owners to input their true data so that the output is a mix of everyone's data

Key differences:

- It is very narrow in terms of types of studies using bespoke libraries (simple statistics mostly)
- Computational needs grow exponentially with the size of the processing
- It imposes stringent constraints on data to be used whereas Sarus works natively with all common data sources and data types
- It does not guarantee that personal information does not leak into the output
- The attacker model is a bit convoluted

How it works: Installing Sarus



1. Install Sarus docker image onto current infrastructure

2. Sync with existing user roles and permissions (LDAP, SAML, OIDC)

3. Connect sources for use via Sarus (S3, GCS, Hadoop, SQL DB) creating Sarus-ready representation