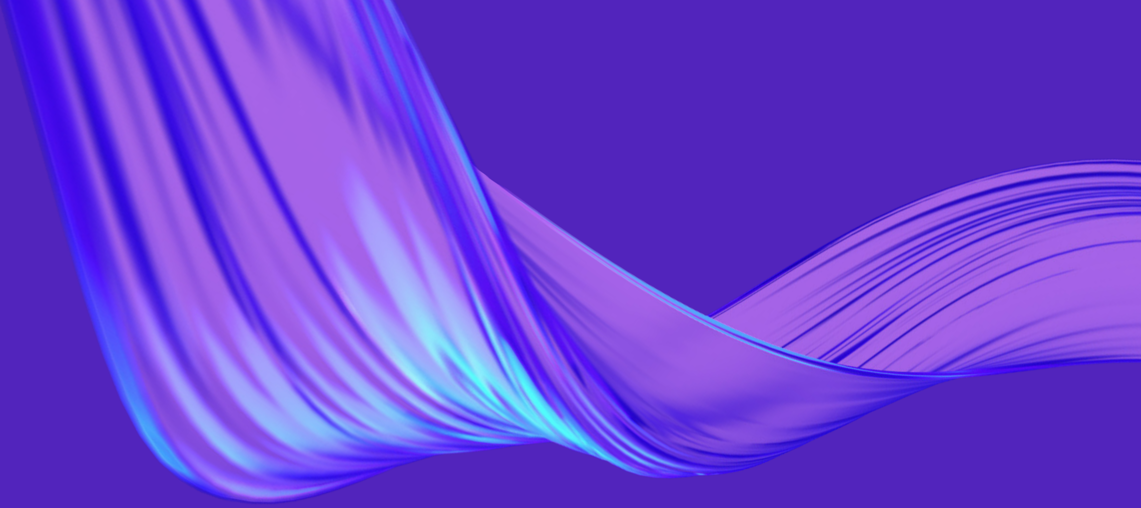
hazy



Synthetic Data Maturity Model

A framework to evaluate and advance your organisation's synthetic data strategy

Executive summary

- Synthetic data is reshaping how organisations democratise their data safely to maintain competitive advantage and supercharge growth.
- Advances in AI and machine learning, along with the rapid emergence of generative AI applications, have led to an increased focus on synthetic data as a solution to data protection as well as its positive impact on innovation and data mobility.
- Synthetic data has become a strategic imperative, and organisations are looking at accelerating the adoption and scaling of enterprise synthetic data platforms.
- As an enabling technology, synthetic data is at a critical inflection point in terms of real business impact.
- To unlock the wide-ranging benefits of synthetic data, technology leaders must first ask themselves the following questions:
  - How can synthetic data enable us to achieve our strategic growth objectives?
  - What do we need to do to realise these objectives?
  - Is there a framework for the adoption of synthetic data at scale across our business?
- This white paper introduces the first ever Synthetic Data Maturity Model – enabling technology leaders to accelerate their journey toward enterprise-wide synthetic data.

Synthetic data in the enterprise

Before we dive into the synthetic data maturity model, let's recap on synthetic data within the enterprise.

> Synthetic data is artificial data generated using AI techniques that can be used as a drop-in replacement for real data. When deployed securely and generated with sufficient privacy, quality and utility levels, synthetic data has a wide range of enterprise capabilities and benefits:

Enterprise capabilities

Accelerate development of analytics and AI – Generate high-quality data to train the algorithms that power AI applications and enable automation.

Transform digital infrastructure – Deploy realistic test data to validate new systems and technologies.

Unlock faster innovation – Quickly share data internally and externally to validate new products and vendors, and deliver more customer value.

Empower business intelligence – Enable teams with analytics to improve decision making.

Productise data – Identify new ways to productise data and generate new revenue streams.

Enterprise benefits

Mitigation of privacy concerns – Differentially private synthetic data can be used and shared safely with reduced compliance risk.

Time and cost savings – With fast access to data on demand, less time and fewer resources are spent procuring data and more on creating value.

Improved model performance – Synthetic data can augment production data and improve the performance of AI/ML models.

New revenue streams – Unlock incremental and new revenue streams by monetising synthetic data and insights.

Advantages of synthetic data over traditional masking and anonymisation

Enhanced privacy protection – provided the data is generated using the right techniques, it is less susceptible to attacks and there is sufficiently low re-identifiability risk.

Better quality – retains statistical properties of the source data and information is not lost.

Speed and cost – quicker and cheaper to provision as it is less resource intensive (provided you are using an appropriate enterprise tool).

Scalability – can be generated at scale to satisfy a range of requirements and use cases across the business.

The Synthetic Data Maturity Model

The Synthetic Data Maturity Model is a framework that describes the different stages of development and use of synthetic data in an enterprise organisation. It can be used to assess the current state of your organisation's synthetic data capabilities and identify areas to bridge the gap between where you are now and where you want to be.

Methodology

Working hand in hand with our customers over several years, we've observed the opportunities and challenges organisations encounter in successfully adopting and scaling synthetic data. We've collected these insights and experiences, validated these with our customers and industry experts, and distilled this into the synthetic data maturity model. The maturity model stages and behaviours are designed to inform and guide your organisation's synthetic data journey. The model is technology-agnostic and thus relevant to the field of enterprise synthetic data as a whole. It will continue to evolve as enabling technologies and organisations reach new heights.

1 Explore	2 Evaluate	3 Operationalise	4 Scale	5 Embed
Stages	→	→	→	→
Explore what synthetic data is, its benefits, and identify early adopters.	Evaluate synthetic data through initial projects to test its effectiveness and to build awareness.	Operationalise your synthetic data capability and lay the foundations to scale.	Scale the use of synthetic data across the organisation and expand to different teams and departments.	Embed synthetic data on-demand or via a marketplace and merge it fully with data and business strategies.
Behaviours	→	→	→	→
Complex and restrictive governance around secure data High risk of data breaches and fines Interest in alternatives to legacy anonymisation techniques	Identified specific needs for an alternative to the use of real data Limited to a handful of use cases Use cases driven by compliance requirements	Measurable proof points from initial use cases Gaining interest from other teams Building rapid delivery functions to service demand	Business-wide understanding and application of synthetic data Data governance describes appropriate use of synthetic data Producer and consumer model is well established	Synthetic data prioritised over real data and available on demand Data consumers are able to review, validate and select data through a marketplace Strategy is aligned to value and ROI

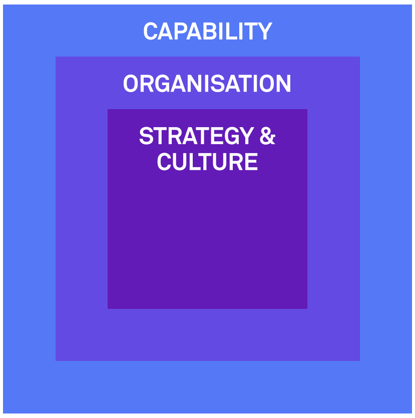
As with any new or emerging technology, organisations often struggle knowing where to start and understanding the parameters of what's possible. Once underway, it can be difficult to sustain momentum without clear focus and direction. In complex enterprise settings, this jeopardises the chances of achieving successful outcomes and can put months or even years of time and effort at risk. The below aims and key questions should be considered to keep you on track at each stage.

Stage	Aims	Key questions
1 **Explore**	Gain a better understanding of the benefits and limitations of synthetic data and identify areas where it can be used effectively.	• What is synthetic data? • What are the benefits and business value drivers? • What are the potential use cases? • Which areas of the organisation could be suitable early adopters?
2 **Evaluate**	Conduct formal evaluations and assessments to determine whether synthetic data solves your business challenges.	• How does the synthetic data compare to the real-world data in terms of accuracy, privacy and reliability? • How does the quality compare to legacy anonymisation techniques? • How should the outputs of experiments be presented to the wider organisation to build awareness and gain buy-in?
3 **Operationalise**	Incorporate synthetic data into new and existing workflows while continuing to demonstrate value to the organisation.	• Which team/department will own the synthetic data capability? • What is the production and consumption model (e.g. centralised or distributed)? • Have "producers" been trained on how to generate data that meets consumer requirements? Have consumers been trained on how to use synthetic data effectively?
4 **Scale**	Maximise the value of synthetic data by driving adoption and engagement among teams and stakeholders.	• Is there sufficient capacity to meet the demand for synthetic data? • Which metrics are being used to monitor and improve the quality and relevance of the data generated for a given use case? • Have data policies and governance been updated to account for the transition from real to synthetic data?
5 **Embed**	Leverage the full potential of synthetic data to drive innovation and business outcomes.	• Does the overall data strategy reflect the transition to synthetic data? • To what extent is synthetic data being consumed on demand? • How much business value has synthetic data created and are there opportunities to expand value further (e.g. privacy, compliance, time and costs, revenue optimisation, etc.)?

Maturity drivers

The maturity drivers are the key pillars that form the bedrock of a successful synthetic data transformation: strategy and culture, organisation and capability. Each pillar is made up of several components which should be considered by the business sponsor(s) and product owner(s) to advance through the maturity stages.

A successful synthetic data programme will consider the drivers holistically to maximise business value.



Strategy and culture

The strategic alignment and cultural readiness of your organisation for the adoption of synthetic data.

Components:

→ Executive support and sponsorship
→ Vision and roadmap
→ Culture and mindset

Organisation

Your organisation's structure, processes and governance frameworks for the adoption of synthetic data.

Components:

→ Data policies and governance
→ Delivery approach (e.g. producer & consumer model, collaboration, CoE)
→ Change management

Capability

Your organisation's technical and knowledge capabilities to achieve your synthetic data goals.

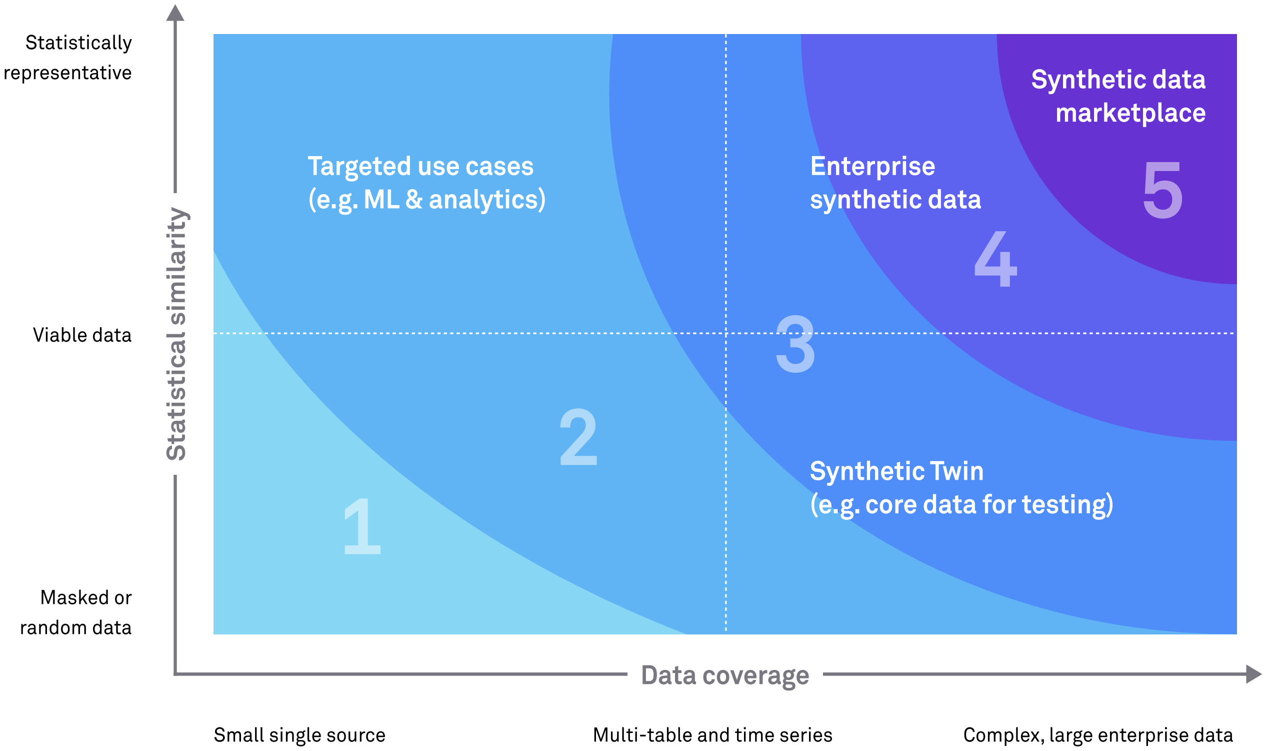
Components:

→ Tech stack and infrastructure
→ Data quality and coverage
→ Resource, knowledge and expertise

Moving towards enterprise synthetic data

Hazy customers and prospects often ask for guidance on how to start their synthetic data initiatives. Our response is always to consider “why synthetic data, and why now?”. Typically, there will be one or more push factors, for example, regulatory pressure, inaccessible data that is stifling innovation or lengthy time for onboarding third party vendors. In any case, it’s important that use cases address key business challenges for maximum impact.

We’ve created a maturity pathway to help leaders align synthetic data with their business needs and help navigate adoption activities through the maturity stages. It has two key aspects, as represented by the axes: statistical similarity and data coverage.


Statistically representative
Viable data
Masked or random data
Statistical similarity
Targeted use cases (e.g. ML & analytics)
Enterprise synthetic data
Synthetic data marketplace
Synthetic Twin (e.g. core data for testing)
1
2
3
4
5
Data coverage
Small single source
Multi-table and time series
Complex, large enterprise data

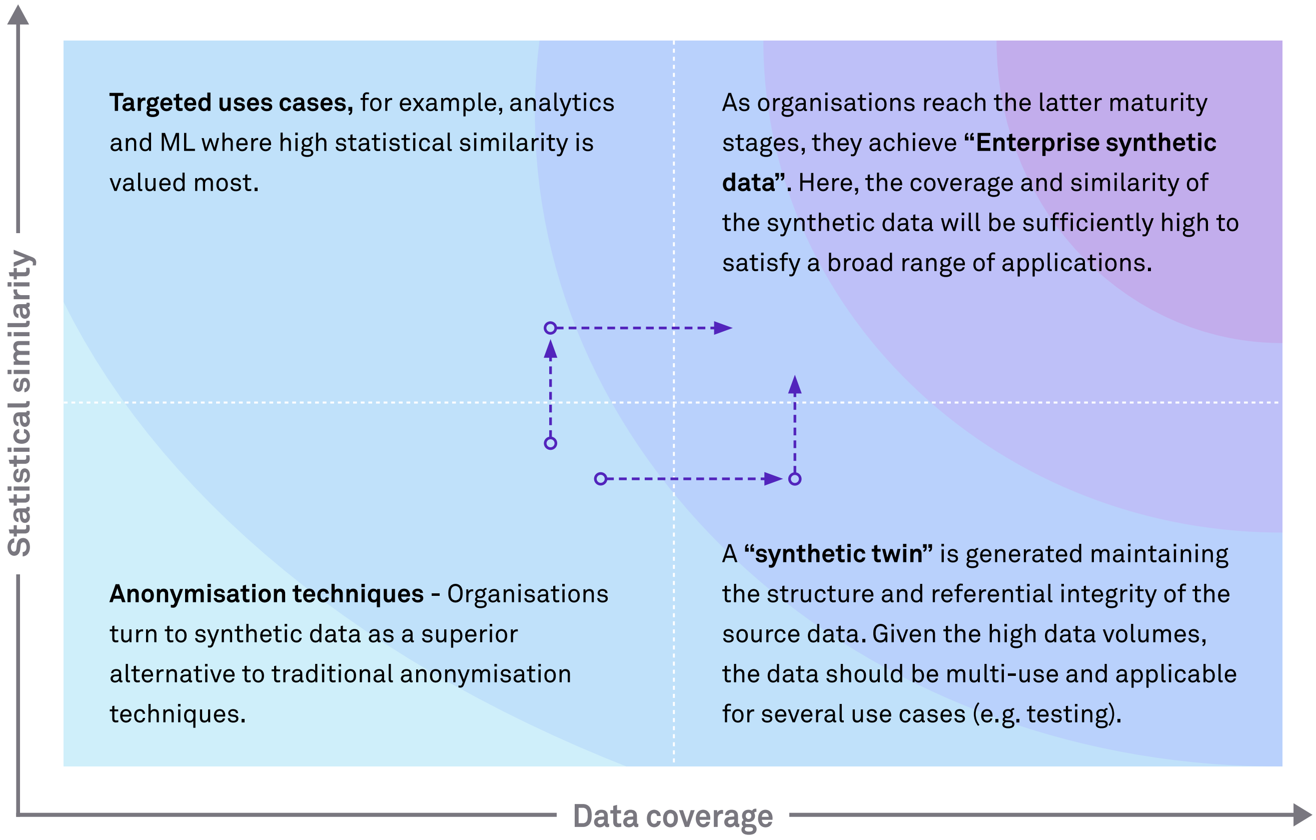
Statistical similarity and data coverage

Each axis represents a key consideration for any given use case and will be determined by the downstream use of the synthetic data. Let's explain each in turn:

- **Statistical similarity** - The concept of statistical similarity is essential to synthetic data as the source data has statistical properties, such as distributions of values. Depending on the use case, consumers may need these properties and distributions to be mirrored to a greater or lesser extent. For example, high statistical similarity is typically required for analytics/ML use cases, however, preserving structural properties and relationships is often the main priority for testing use cases.
- **Data coverage** - Refers to the scope, volume and complexity of the source data being synthesised. Data coverage could range from a single dataset from one source to complex databases housing inter-connected data across domains.

Transition through the quadrants to Enterprise Synthetic Data

As organisations advance their synthetic data maturity (and transition through the stages), they will land in one of these four quadrants:



Targeted uses cases, for example, analytics and ML where high statistical similarity is valued most.

As organisations reach the latter maturity stages, they achieve **"Enterprise synthetic data"**. Here, the coverage and similarity of the synthetic data will be sufficiently high to satisfy a broad range of applications.

Anonymisation techniques - Organisations turn to synthetic data as a superior alternative to traditional anonymisation techniques.

A **"synthetic twin"** is generated maintaining the structure and referential integrity of the source data. Given the high data volumes, the data should be multi-use and applicable for several use cases (e.g. testing).

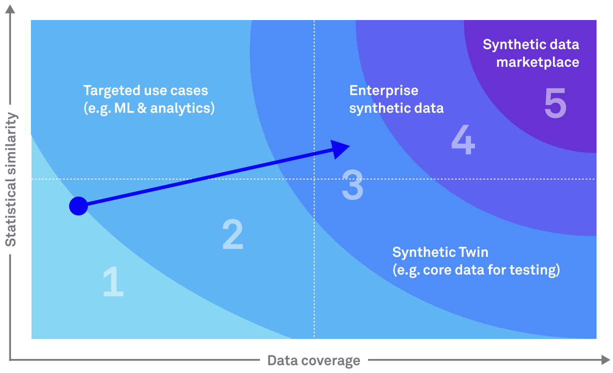
Synthetic data marketplace

As synthetic data replaces real data across a wide range of applications, the consumption of synthetic data is likely to move towards an on-demand or internal marketplace model (top right of the maturity pathway). Once embedded, data consumers will be able to seamlessly request or select the data they need and move it safely into their own environments.

The producer teams will be responsible for synthesising core datasets (high data coverage) and for enforcing standards and governance for the quality of the generated data (e.g. validating similarity, privacy and utility).

Making it real: An illustrative example

A leading financial services firm began its synthetic data journey by looking at an innovation use case to validate external vendors more quickly. Partnering with fintechs is a fast, safe way for incumbent firms to access technology that would take them years to build in-house. Yet for the firms to prove the partners meet the procurement and compliance regulations of the banks, their software must be evaluated using data. Using real data was a slow, resource-intensive, and costly approach.


Synthetic data marketplace
5
Targeted use cases (e.g. ML & analytics)
Enterprise synthetic data
4
3
2
1
Synthetic Twin (e.g. core data for testing)
Statistical similarity
Data coverage

They installed the Hazy platform next to the source data for the delivery of the initial use case. After the successful completion of the initial data sharing use case (run by a single team), the firm expanded Hazy and synthetic data across the organisation - using metrics to monitor the quality, privacy and utility of the synthetic data generated across teams.

The financial firm synthesised its customer data, which enabled them to evaluate more than 100 new vendors in a year. By expanding synthetic data production, the organisation now has a synthetic customer dataset that is available for the business to use across various use cases.

What next?

Now that you're equipped with the synthetic data maturity model, we recommend following these simple steps:

→ Pinpoint where you currently are based on the stages and behaviours

→ Set a target for which stage you'd like to get to by when

→ Consider the appropriate steps based on the maturity drivers and pathway

We're here to support you as synthetic data experts. We can advise you on how to begin or accelerate your synthetic data journey, drawing on experience from our customers and industry best practices.

Authors

Jonathan Hardy
Director of Customer Success

Gareth Rees
VP Customer

Lauren Arthur
Marketing Director

Marisa Teh
Chief Product Officer

Harry Keen
CEO & Co-Founder

About Hazy

Hazy is the world's leading synthetic data company, re-engineering enterprise data so that it's faster, easier and safer to use.

Our software replicates enterprise data, preserving the statistical properties needed for testing, analytics and innovation, while omitting all personal identifiable information to maintain privacy and compliance.

Hazy synthetic data can be repeatedly generated and safely shared under regulation such as the GDPR, freeing your businesses to transform faster, accelerate AI and innovate without restriction.

Find out more at hazy.com