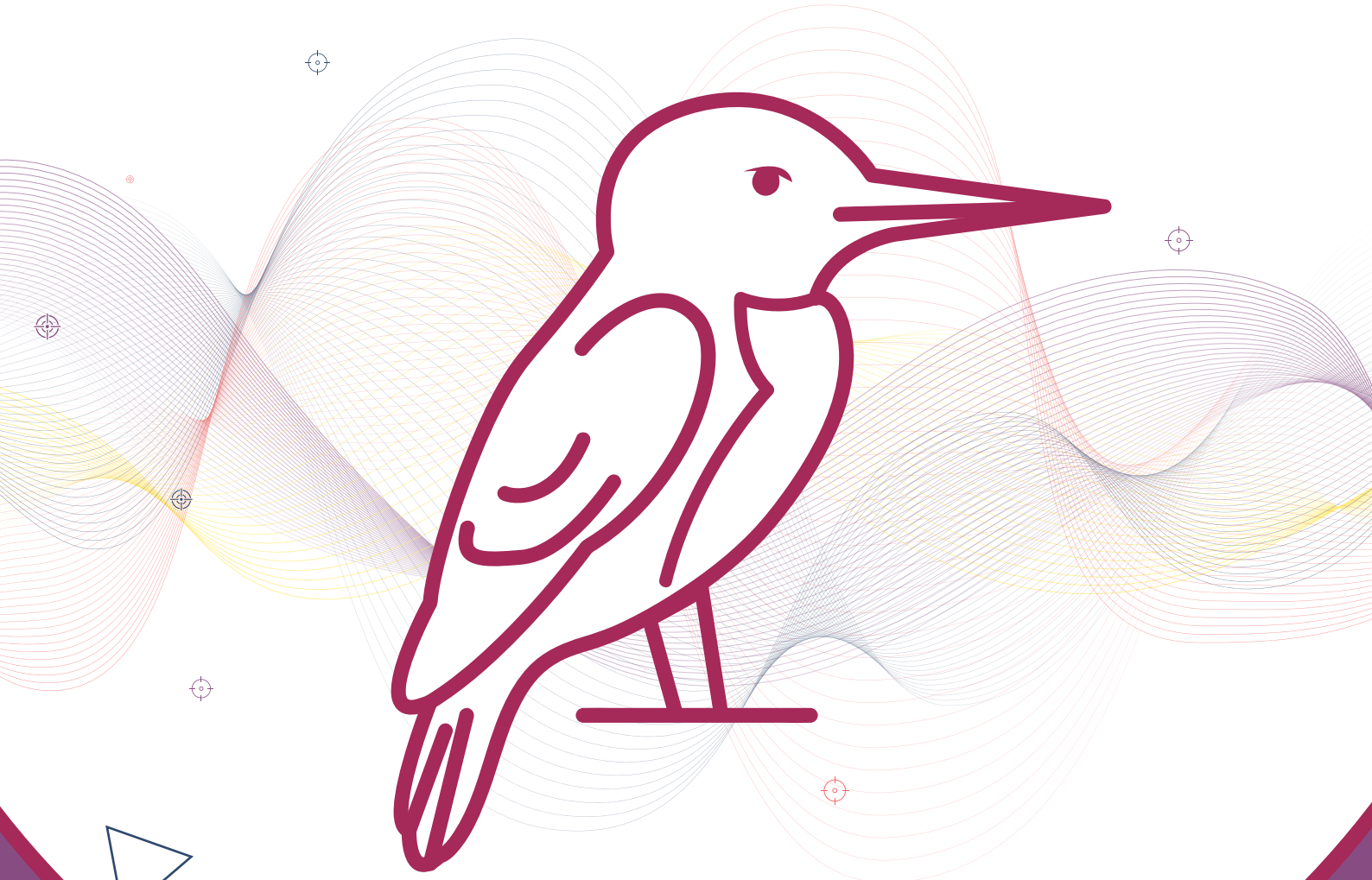


onix

Kingfisher User Guide

Table of Contents

1.Introduction

Kingfisher is an AI-powered data generator that creates highly accurate synthetic relational data. It learns patterns from real data and automatically detects Business Types to generate realistic data that can be exported for Data Science and Engineering needs.

2.Accessing Kingfisher UI

Use the provided hostname and port number. For example:
http://<hostname>:<port>

3.Functional Usage

Kingfisher's User Interface (UI) simplifies the process of generating, managing, and interacting with synthetic data. Here's what you can do:

- **Generate Synthetic Data:** Create customized datasets that mimic real-world data. Define data schemas, specify data types, and control the number of records. Kingfisher understands SQL, automatically extracts table relationships, and captures literals.
- **Manage Jobs:** Create and manage data generation jobs easily. Monitor their progress from setup to execution.
- **Download Data:** Download generated data in various formats (CSV, JSON, Fixed-Width, AVRO, Parquet).
- **View Job Canonical Information:** View detailed insights into generated data on the Job Details page.
- **Preview and Edit Canonical Data:** Preview and edit canonical information to fine-tune datasets. Regenerate jobs with updated settings.

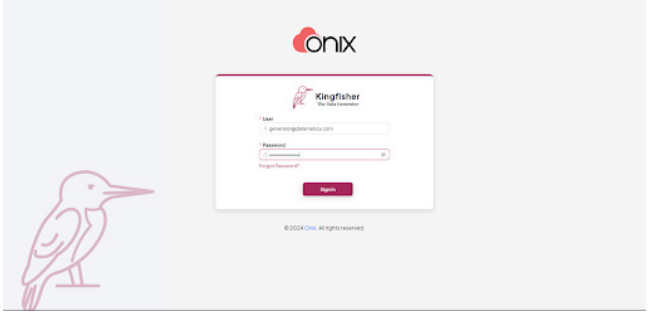
4.Step-by-Step Guide

This section provides a detailed walkthrough of the key processes involved in using the Kingfisher UI.

4.1. Login to the UI

- **Open the Kingfisher UI:** Launch your web browser and go to: http://<hostname>:<port>

  Load the Kingfisher login page.



- **Login Page:** Enter your username and password, then click "Sign in."
- **Access the Dashboard:** You'll be redirected to the Kingfisher dashboard, your central hub, where you can create jobs, monitor their progress, and download data.
- **Accessing Admin-Level Features:** Log in with admin credentials to access admin features (managing user roles, licenses, LLM Provider configurations).

Note: Contact the Kingfisher team for the login credentials.

4.2. Home/Dashboard Page

After login, you'll see the Home/Dashboard Page. The Home/Dashboard page gives you an overview of Kingfisher's activity, including:

4.2.1. Statistics Overview

Shows key metrics:

- Data processed (in KB or larger).
- Number of tables processed.

- Number of records handled.
- Total jobs executed.

4.2.2. Time-Series Chart

Visualizes daily, weekly, monthly, and yearly metrics:

- Tables processed daily.
- Records processed daily.
- Jobs executed daily.

4.2.3. Jobs Section

Table of recent jobs with the following details:

- Created date/time.
- Job Name.
- User (or email).
- Description (optional).
- Status (e.g., Analysis Completed, Data Generation Completed).

Each job has a **Details** button

4.2.4. Job Complexity Overview

Table of recent jobs with the following details:

- Created date/time.
- Job Name.
- User (or email).
- Description (optional).

Additional Actions in the Jobs Section:

- New Job: To create a new job.
- Search: Find jobs by name.
- Sort: Sort jobs by creation time.

4.3. Create Job

Navigate to the Job Create Page, where you will complete the following sections:

1. Basic Information
2. Analyze
3. Generator
4. Data Export

4.3.1. Basic Information

Fill out these fields:

- **Job Name:** Enter a unique name to identify the job.
- **Job Description:** Briefly describe the job's purpose.

4.3.2. Analyze

The Analyze section is where you configure all the details about your input data (e.g. DB Schema, Data Profile, SQLs) so that Kingfisher can generate the synthetic data you need.

4.3.3. Generator

The Generator section allows you to specify how many records Kingfisher should generate, based on your previous inputs.

- **Number of Records:** Input the desired number of records. Kingfisher can handle both small and large datasets.
- **Locale:** Set the regional format and language for the generated data to match specific cultural conventions.
- **Introduce Noise:** This feature allows you to add intentional variations, or "noise," to the data Kingfisher generates. You set a percentage to control the amount of noise added. This helps make the data more like real-world data, test how well your system handles irregularities, and create more diverse data for analysis. The noise is added based on the types of information detected or that you provide

4.3.4. Data Export

This section lets you configure how generated data is exported.

- **Supported File Types:** Choose one or more: CSV, JSON, AVRO, PARQUET, DAT, Fixed Width. You can select multiple types.

Note: Users can select multiple file types simultaneously for their export requirements.

4.4. Job Details Page

The Job Details page provides an overview of the job execution, allowing users to interact with generated data and metadata. It displays a list of tables, job management options, and detailed table information.

4.4.1. Job Hierarchy

- A list of tables involved in the job.
- Clicking a table displays its details

The download option lets you download generated data, canonical data, and reports.

- Generated Data: Synthetically generated data.
- Canonical: Metadata essential for data generation.
- Reports: A zip file of generator and analyzer reports.

4.4.2 Job Overview

Detailed job input information for tracking, debugging, and validation:

- Dialect
- Schema Type
- Mode of Operation
- File Format
- Number of Records
- Data Distribution Percentages (Associated, Default, InfoType, Literal, Modified, Nullable, Profiled).

4.4.3. ReRun Job

The ReRun Job feature lets you re-execute a job with modified parameters (including edits to statements) and a new number of records. You can also use these modes:

- Introduce Noise: Add variations to the generated data by setting a percentage.
- Modes: Choose how to handle existing data:
    - Clean: Removes all previous data and restarts.
    - Append: Adds new data to existing data.
    - Resume: Continues from the previous execution.

4.4.4. Preview Data

The **Preview Data** button lets you view the first 100 records of generated data for a selected table. This helps you verify if the data generation worked correctly and if the relationships and values are as expected. You can also preview data for individual tables and download data files for a specific table as a zip file.

4.4.5.Report Tab

This section displays the status of each analyzed and generated table. You can monitor table statuses, track records, identify errors, and review key metrics. Get overviews of column count, generation time, and generated records for each table.

Key Metrics Displayed:

- Table Qualification: Schema info alongside table name.
- Status: Table's current state (e.g., Analysis Completed). Filter by status using the filter icon.
- Column Count: Number of columns per table.
- Generation Time: Time taken to generate data.
- Generated Records: Number of successfully generated records.
- Exception Message: Error message for failed tables.
- Comment: Additional notes on table processing.

4.4.6. Summary Tab

The Summary Tab provides an overview of job execution, enabling you to track progress, identify failures, and assess job performance.

Key metrics displayed:

- Associations Count
- Literals Constraints Count
- Total Records Generated

Analysis Summary:

- Analysis Time: How long analysis took.
- Total Tables: Number of tables processed.
- Successful Tables: Number of successfully analyzed tables.
- Failed Tables: Number of tables with errors.
- Total Parsed SQL: Total SQL queries parsed.
- SQL Success/Failure: Breakdown of successful and failed SQL queries.

Generation Summary: (Insights into data generation)

- Generation Time: The duration of the data generation process.
- Input File Count: Number of files used for generation.
- Successful Tables Generated: Number of successfully generated tables.

5.Data Profile

A Data Profile is a structured summary of the statistical and structural characteristics of data. It includes key metrics such as minimum, maximum, mean, median, percentiles, standard deviation, value distributions, and the percentages of null/empty fields (table level as well as column level). These insights are critical for assessing the quality, consistency, and distribution of data.

6.Admin Page

The Kingfisher Admin page manages Licensing and LLM Providers. User Management and Database Configurations will be added in the future.

6.1. Licensing Section

Upload a license file to activate or update the license. The status and details will be displayed below (Active/Expired, start/end dates, days remaining, features).

6.2. LLM Providers Section

Manages Large Language Model (LLM) providers for AI features.

Configuration Form:

- Configuration Name: Unique name for the provider.
- Provider Type: Hosted or On-Prem.
- Provider Name: Select LLM (OpenAI for Hosted, Llama/Gemma for On-Prem).
- Authentication Token: Enter the token.
- Instance URL: URL for On-Prem providers.
- Submit/Cancel: Save or discard changes.

Note: Currently, Kingfisher supports OpenAI.

Provider Details: Lists configured providers:

- Configuration Name, Provider Type, Provider Name, Authentication Token (masked), Instance URL.
- IsActive: Activate/deactivate by checkbox.
- Actions: Edit provider (Authentication Token, Instance URL).

6.3. Database Configurations

This section lets you configure and manage database connections for Kingfisher in the Admin Portal. This allows Kingfisher to connect to various databases for data operations.

Key Features:

- Create, edit, and manage database

connections.

- Enables licensing-based data access control.

Configuring a Database Connection:

You need to provide:

- **Configuration Name:** A unique name for the connection.
- **Database:** Select database type.
- **Purpose:** Data Loading, Data Profiling, Read Profile.
- **Connection Details:** Hostname, Username, Password, Port, Service Account Key File.

6.4. User Management:

Admins can create and manage user accounts and roles to control access. Users log in with unique credentials and perform tasks based on their role

User Creation requires:

- **User Name:** Unique login ID.
- **Email:** User's email.
- **Password & Confirm Password.**
- **First/Last Name:** (Optional).
- **Role:** Generator (limited data access) or Admin (full access).

This provides a secure and efficient way to manage users and access control.

7. Supporting Documentation

7.1. Supported Infotypes

Please find below the list of info types along with their descriptions:

Info Type	Description	Generated data (example)
age	The age of a person.	80
bloodgroup	Blood group type (A, B, AB, O ,A+/A-).	A-
completionstatus	Status (complete, incomplete)	complete
yesno	Binary yes or no response.	yes
validstatus	Validity status (valid ,Invalid)	valid
passfail	Result indicating (pass , fail)	fail
gender	Gender identity of a person (male, female, other)	female
hiringstatus	Status of the hiring process (hired , rejected, on-	rejected
relationship	Type of relationship.(single, married, unmarried,	married
phonenumber	phone number.	4.4575E+11
timezone	Time zone information.	Europe/Malta
timezoneabbreviation	Abbreviated form of time zone.	UTC
timezoneoffset	Offset from Coordinated Universal Time (UTC).	07:00
timestampwithlocaltime zone	Timestamp with time zone.	27-02-2024 00:32

Info Type	Description	Generated data (example)
json	JSON formatted data.	[{"name": "Megan Callahan", "residency": "087 Eric Gardens\nMosstown, RI 27024"}]
address	Physical address.	545 Holly Lakes Apt. 541
postcode	Postal code.	99127
middleinitial	Middle initial of a name.	N
letter	Single alphabet letter.	D
uppercaseletter	Uppercase alphabet letter.	I
lowercaseletter	Lowercase alphabet letter.	v
double	Double-precision floating-point number.	49.20021329
licenseplate	Vehicle license plate number.	AIW-939
swiftcode	SWIFT code for bank identification.	ZGOTGBIFOU0
swiftvalue	Value containing SWIFT code.	GB56IVSX06273204570308
model	Model information.	Model S
rgbcolor	RGB color value.	2,52,81,156
hexcolor	Hexadecimal color code.	#e1755e
color	color information.	DarkBlue
colorname	Name of a color.	Snow
csscolor	CSS color value.	#a5b449
csscolorname	CSS color name.	Olive
creditcard	Credit card /debit card information.	2.26517E+15
cardexpire	Credit card / debit card expiration date.	Apr-30
creditcarddetail	Details of a credit card/ debit card.	Maestro Amanda Myers 60459795666 2/13 CVV: 534
creditcardprovider	Credit card /debit card provider.	VISA 16 digit
creditcardsecuritycode	Security code of a credit card/debit card.	980
md5	MD5 hash value.	529de6f86cde9ed7ed11651561a260a3

Info Type	Description	Generated data (example)
sha1	SHA-1 hash value.	6589b81bd027516288e4559cc9cfb8f1a92b9d64
sha256	SHA-256 hash value.	e4a120316545824ace9073413d80ac8e3f0a0b80d4e60cc59eb164608ad8ed9a
uuid	Universally unique identifier.	52bdabe0-0cbc-44c6-9784-66977846c18d
price	Price amount.	$506.71
currency	Currency symbol information.	MAD
currencyname	Name of a currency.	Singapore dollar
currencycode	Code of a currency.	PAB
cryptocurrency	Cryptocurrency information.	MazaCoin
cryptocode	Code of a cryptocurrency.	FTH
filepath	Path to a file.	/team/color.doc
folder	Folder name.	design
file	File name.	must.gif
fileextension	Extension of a file.	flac
company	Company name.	Davis Inc
companysuffix	Suffix of a company name.	and Sons
job	Job title.	Accountant, chartered certified
designation	Designation or title.	Fast food restaurant manager
portnumber	Number of a port.	5862
url	Uniform Resource Identifier.	https://www.vazquez.com/tag/tagsauthor.php
url	Uniform Resource Locator.	https://gomez-moore.com/
domainname	Name of a domain.	graham.com
httpmethod	HTTP method (GET, POST).	DELETE

Info Type	Description	Generated data (example)
imageurl	URL of an image.	https://placekitten.com/544/658
ipv4	IPv4 address.	168.48.121.245
ipv6	IPv6 address.	4eba:f591:5a08:32f5:768:2fe3:9712:1c89
macaddress	MAC address.	60:99:f5:a8:ce:f7
hostname	Name of a host.	db-40.santana.com
username	Username.	michellesheppard
password	Password.	e$LoQdh3%7
email	Email address.	fullerjonathan@example.org
word	Single word.	gas
description	Description text.	Maybe ball property. Suddenly ok visit it even.
sentence	Sentence of text.	Federal husband floor perform claim near cultural.
text	Text content.	Eight list capital main trial. Question close quickly discover its.
country	Country name.	Qatar
countrycode	Country code.	TH
zip	Zip code.	40118
streetaddress	Street address.	0286 Foster Circles Suite 969
city	City name.	Kota
state	State name.	Uttar Pradesh
streetname	Name of a street.	Sami Road
socialsecuritynumber	Social Security Number.	200-94-4979
name	Full name.	
firstname	First name.	
lastname	Last name.	
malename	Male name.	
femalename	Female name.	

Info Type	Description	Generated data (example)
malefirstname	Male first name.	
malelastname	Male last name.	
femalefirstname	Female first name.	
femalelastname	Female last name.	
nameprefix	Prefix to a name.	
femaleprefix	Prefix for a female.	
maleprefix	Prefix for a male.	
countrycallingcode	Country calling code.	809
num	provide random number	90
year	Year value (YYYY).	2006
month_name	Name of a month.	September

Custom Info Types			
Info Type	**Description**	**Example**	**Generated data (example)**
custom_list	Provide list of values NOTE - values should be pipe separated	custom_list(java|python)	python
custom_regex	provide custom pattern, Number signs ('#') are replaced with a random digit (0 to 9) , Question marks ('?') are replaced with a random character from letters. All other characters remain unchanged	custom_regex(##??)	52dD

Custom Info Types			
Info Type	Description	Example	Generated data (example)
date_between	provide range between dates , date_between(start_date|end_date |date format) NOTE - date_format specifies the output format and should be a valid date format., start date and end date should strictly follow date format -%Y-%m-%d %H:%M:%S	date_between(2023-08-01 15:06:25|2023-11-01 15:06:25|%Y-%m-%d %H:%M:%S)	23-10-2023 16:50
num_range	provide numbers within specified range	num_range(0|10))	3
date_time_format	provide date time format	date_time_format(%y-%m-%d)	00-10-06
str_len	provide specific length to string column	str_len(10)	zlrbklCZGK
decimal_with_scale_precision	provide specific precision and scale to decimal	decimal_with_scale_precision(10|2)	7840.93
date(%d-%b-%Y)	Provide date in format as 01-Jan-2024	date(%d-%b-%Y)	01-Jan-24
date(%d-%b-%y)	Provide date in format as 01-Jan-24	date(%d-%b-%y)	01-Jan-24