IRI RowGen
Smart Test Data Generation

Product Overview

Technical Summary, Samples, and Specifications

IRI
Total Data Management

Introduction

IRI RowGen *generates rows* of safe, intelligent test data values and puts them directly into database (DB), file, and report targets. RowGen was designed by data modeling, integration (ETL), and application development experts as a means of saving time in the creation of realistic, privacy-law-compliant test data. RowGen uses existing or auto-created metadata to randomly build and/or draw values that adhere to the structure and relationship of linked tables and files, but it also supports complex data customization to meet exacting business needs.

Because it supplies test data of the highest quality, safety, and utility, RowGen is an essential part of every Test Data Management (TDM) initiative. According to the InfoSys white paper, "Test Data Management, Enabling Reliable Testing through Realistic Test Data,"

> *Test teams not only have to follow exact test methodologies, but also ensure the accuracy of test data. They also need to ensure tests correctly reflect production situations, both functionally and technically.*

RowGen produces safe and realistic persistent or virtual test data for DevOps, and:

- Load and query verification through VLDB and EDW prototype tables
- Capacity (stress) testing and functional validation in software design
- Performance benchmarks for new hardware and software
- Product/service demonstration or outsourced development

RowGen enables development in parallel by providing data to those who can define their stages and the data they need before it exists (i.e., gets created in a prior stage). And by transforming and formatting data *as it is generated*, RowGen lets developers: validate the data being created, reduce their need to import test data into other programs, and share test tables or custom files.

RowGen allows speed- and regression-testing in these cases because the test data it builds is:

- loaded, in the right locations and formats
- safe, so it complies with data privacy laws
- complete, in terms of good, bad, and null data values
- real-looking, and manipulation-ready
- large, for "big data" application simulations
- range-distributed, to reflect natural frequencies/occurrences
- referentially correct, to maintain query (join) integrity

TDM recommendations -- and the role of RowGen in TDM -- is covered in these blog articles:

TDM Step 1: Goal Setting & Team Building

TDM Step 2: Test Data Needs Assessment

TDM Step 3: Test Data Generation & Provisioning

TDM Step 4: Test Data Sharing & Persistence

RowGen Product Features

RowGen uses the IRI CoSort data sorting and mapping engine to optimize the generation, transformation, formatting, and population of “big test data” targets. RowGen also allows you to:

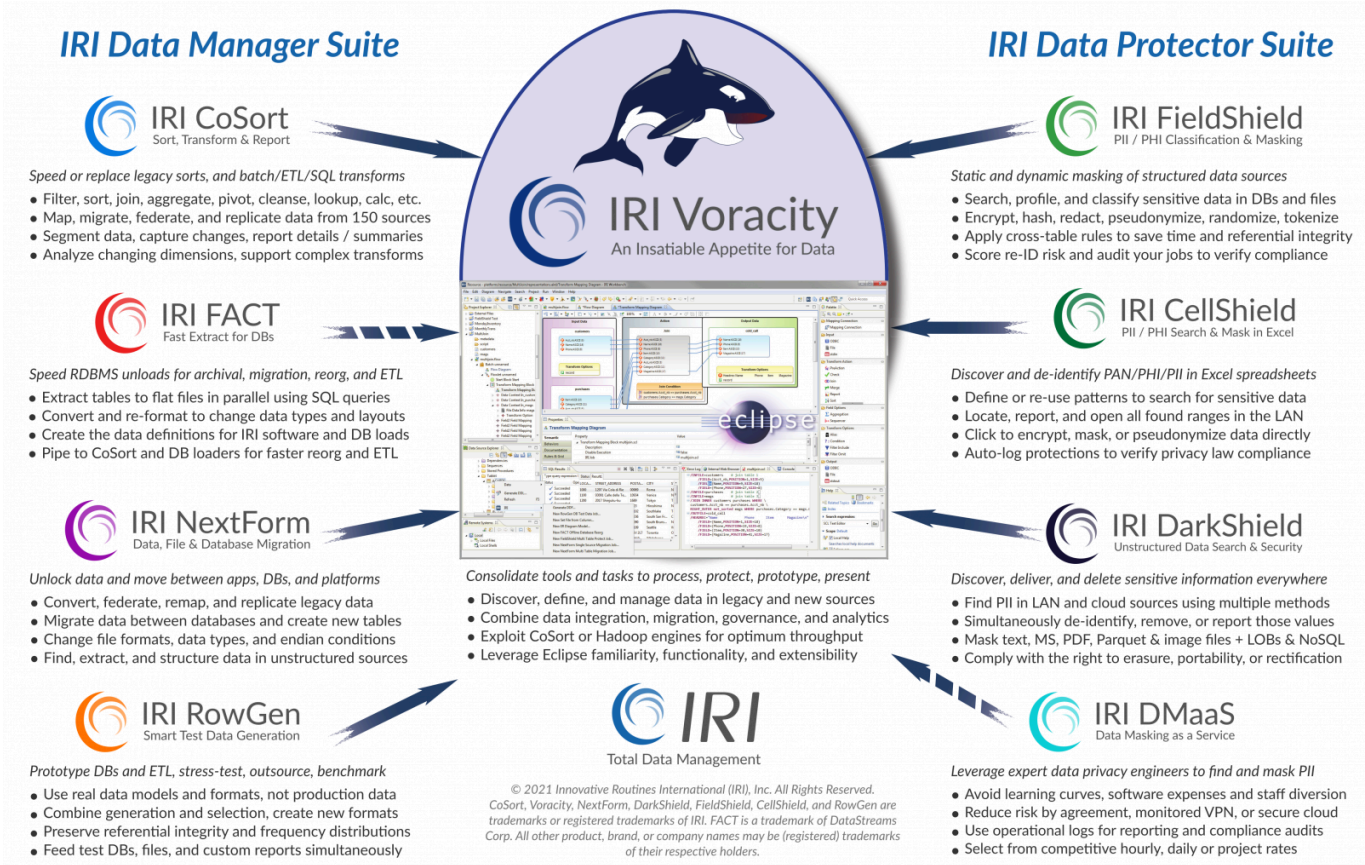
- Create and modify jobs in the IRI Workbench GUI for RowGen, or through 4GL scripts
- Connect to and profile files and RDBs, and define their metadata and ERDs in Eclipse
- Customize data values, formats, volumes, ranges, distributions, etc. on the fly
- Stylize the appearance (layout and labeling), content, distribution, and volume of data
- Write the synthetic test data to multiple targets and formats at once, as well as reports
- Maintain referential integrity, including compound, and self-referencing foreign, keys
- Filter, transform, and pre-sort the test data in the same job that generates it
- Specify and save rows-per-target, null percentages, and cross-table generation rules
- Use existing, or auto-create new, set files to randomly select values from real data
- Save, re-use, and share RowGen jobs using free version control repositories like Git
- Generate test data values in applications via APIs, CI/CD pipelines, or Hadoop clusters
- Run or schedule test data jobs from the GUI, command line, batch job, or tools like UAC
- Create masked database subsets (as an alternative to large scale EDW generation)
- Leverage the same metadata in other SortCL-driven jobs, like FieldShield and CoSort

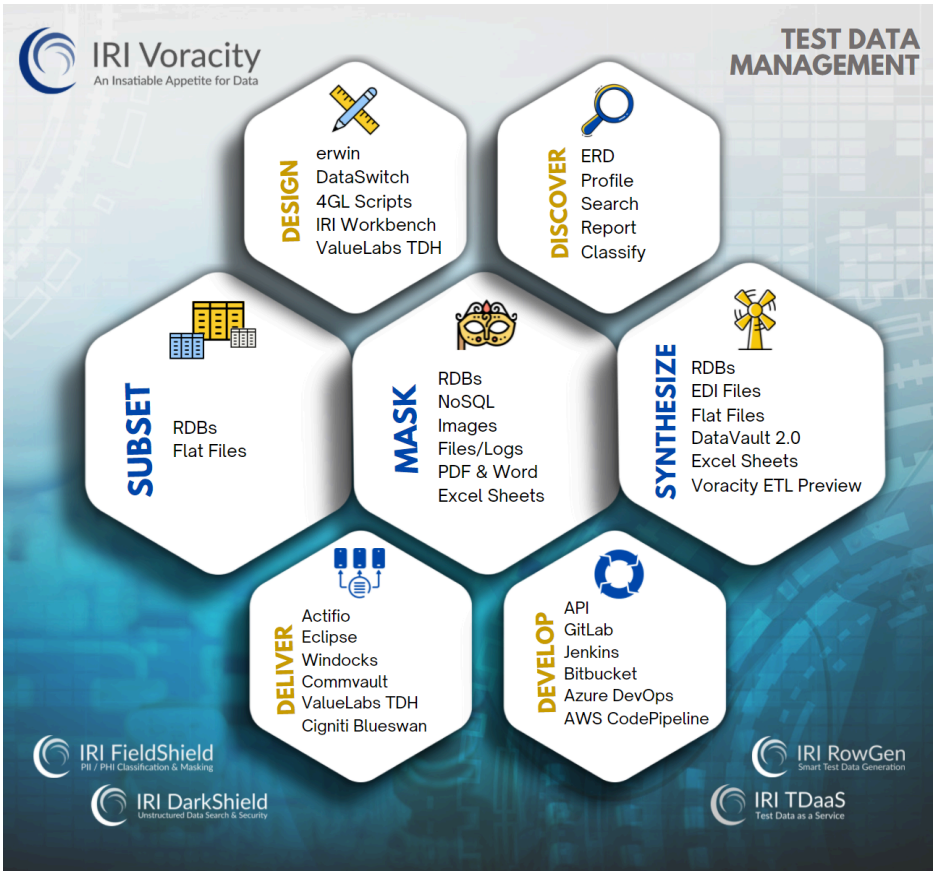
RowGen Technical Specifications

Test Generation Functions • Random data generation and selection • Parallel sorting (speeds loads & queries) • Conditional filtering (for value ranging) • Multi-targeting (for related tables and custom files) • Calculation, aggregation & sequencing (RowID) • Linear, normal & weighted frequency distributions • Compound, range, literal & DB column values • Custom flat-file and report formatting • Job Logging (in XML format, for compliance)	**Test File Formats** • ACUCOBOL-GT Vision • Fixed- or variable-position text • Line, Record, or Variable Sequential • LDAP Interchange Format (LDIF) • Micro Focus Variable Length & MF-ISAM • Mainframe Variable Blocked Format • MS Comma-Separated Values (CSV) • W3C Extended (Web) Logs • W3C XML (Flat)
Supported Database Classes • Columnar • Hbase • In-memory • Mainframe • NoSQL • Relational	**Supported Metadata** • JDBC and ODBC DDL • COBOL copybooks • IRI SortCL-based data definition files (.ddf) • CSV, flat XML, LDIF & W3C ELF headers • SQL*Loader, DB2 Load, BCP • Data Vault 2.0 DW models and business keys
Special Data Generation Functions • GUID & UUID • Korea & USA SSNs • Italy Codice Fiscale & Netherlands BSN • Singapore NRIC • Credit card #s and Email addresses	**Supported Hardware** • AIX (IBM i, p & z Series) • HP-UX (RISC and Itanium) • Linux (x86, Itanium, IBM i, p & z Series) • Solaris (SPARC and Intel) • Windows (Server 2000-2022, 7-11 macOS
Test Data Types (for Random Generation) • ASCII & EBCDIC Characters • Numeric, Whole, Currency, IP Address • Alpha & EBCDIC Digits • RM and MF COBOL Numerics • Other Binary Numerics • US, European, ISO, Japanese Timestamps • Unicode & Multibyte Characters	**Compatible Products** • DataSwitch, Erwin & MIMB metadata bridging frameworks • Data Vault 2.0 - Data Modeling (Voracity generation wizard) • IRI CoSort - Data Transformation & Reporting • IRI FieldShield - DB and File Data Masking • IRI NextForm - Data, File & DB Conversion • IRI Voracity - Total Data Management • Windocks, Value Labs Test Data Hub & Cigniti BlueSwan

RowGen in Context

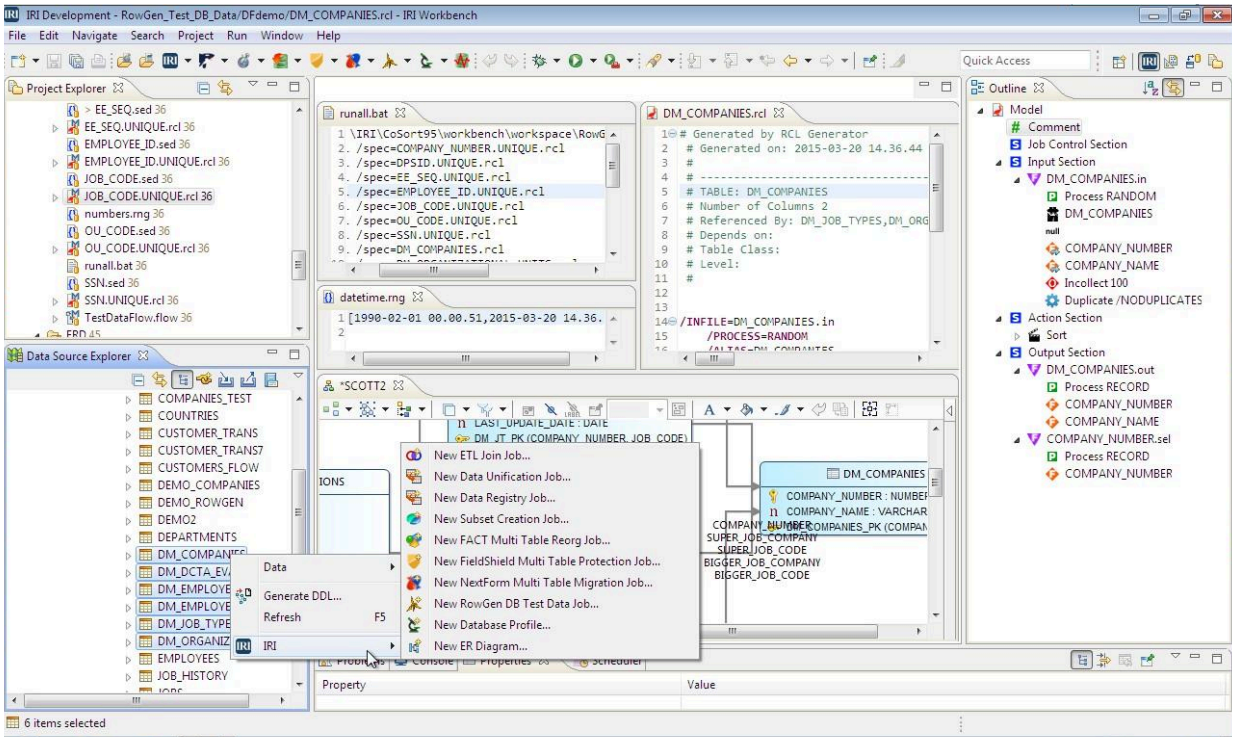
RowGen is a standalone product, and an IRI Voracity data management platform component





RowGen GUI

Companies licensing RowGen have unlimited access to a graphical user interface (GUI) built on Eclipse™ for job design, deployment, and management. The IRI Workbench GUI for RowGen has data profiling tools and new job wizards to locate and build metadata, and then use it to create, sort, and load new test data. Only the data generation box needs license keys.



The GUI contains several job wizards and functional dialogs, plus a syntax-aware script editor and dynamic outline for RowGen Control Language (RCL) jobs, and a form editor for metadata.

Several RowGen-specific wizards in the IRI Workbench provide end-to-end test data solutions:

1. *New Test Data Job* wizard, for building single test targets (files, DB tables, and reports)
2. *New DB Test Data Job* wizard, for parsing, generating, and populating full test DBs
3. *New Set File* wizard, for creating custom lookup sets for random selections from compound (bespoke) data values, range or literal values, and values in DB columns.
4. *New Subset Creation Job* wizard, for building masked, smaller replicas of DB tables
5. Business-key correct Data Vault 2.0 test data generation is also supported; see this.

Also available to RowGen users are Data Discovery wizards, including:

1. Data Class Discovery for RDB schema and for flat files in specified directories
2. Database Profiling, for searching, classifying, and reporting statistics about tables
3. Flat File Profiling, for doing the same on flat-files
4. ER Diagramming, for visualizing the content and relationship of any relational DB
5. Dark Data Discovery, for searching and extracting values in unstructured sources

RowGen users can get up to speed quickly from the Welcome screen in the IRI Workbench. First Steps include a getting started "cheat sheet" that links to the RowGen job wizards.

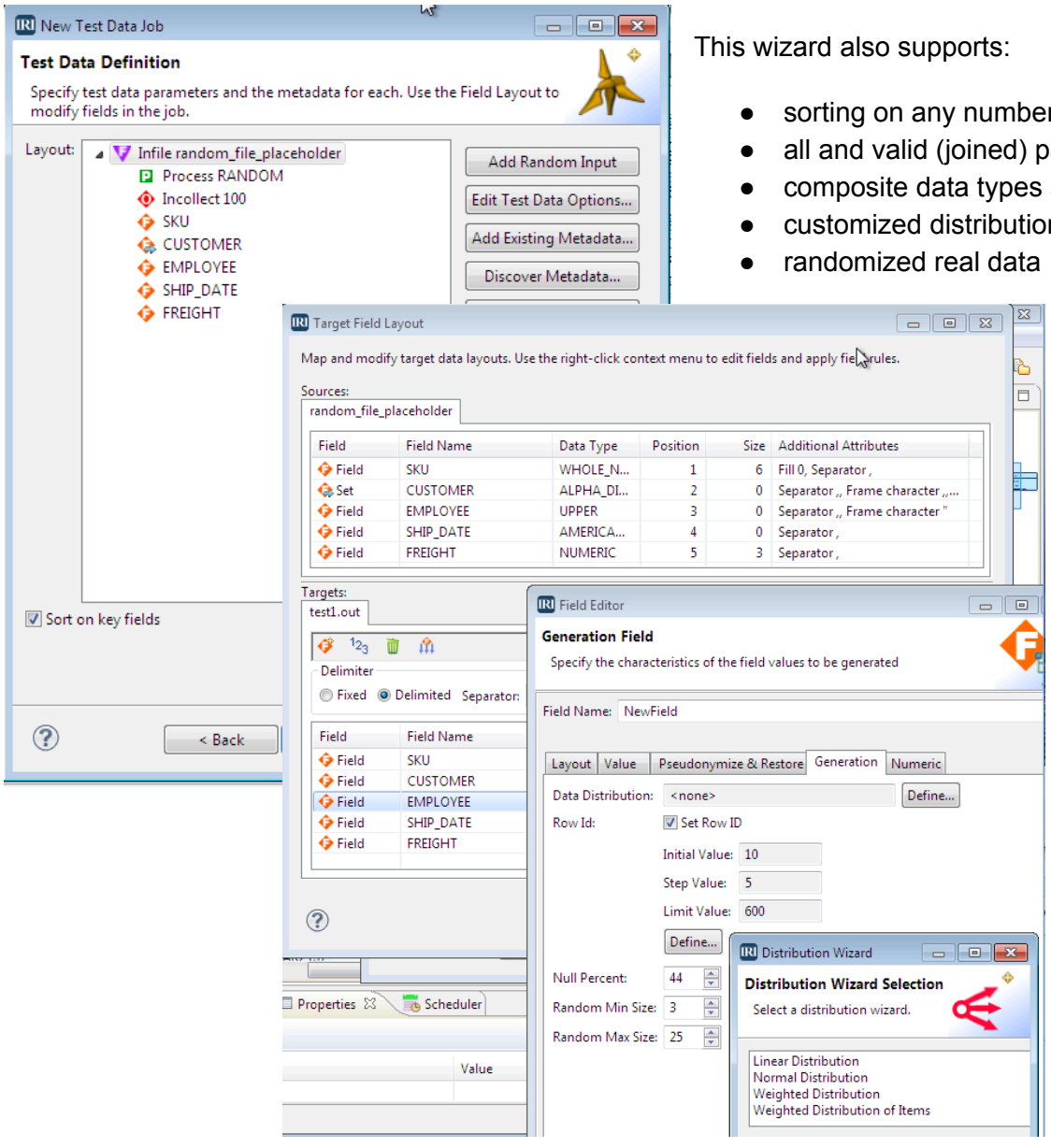
The .pdf reference manual for RCL conventions and job samples, as well as a Java SDK for embedding test data generation and selection functions into applications, is also provided with the GUI, even though those components run externally.

These dialogs show some of the initial steps in using an existing database structure (which need not contain any data) to generate the right test data. To create test data in flat-file formats, you can use similar metadata discovery and definition dialogs in the RowGen GUI.

New Test Data Job Wizard

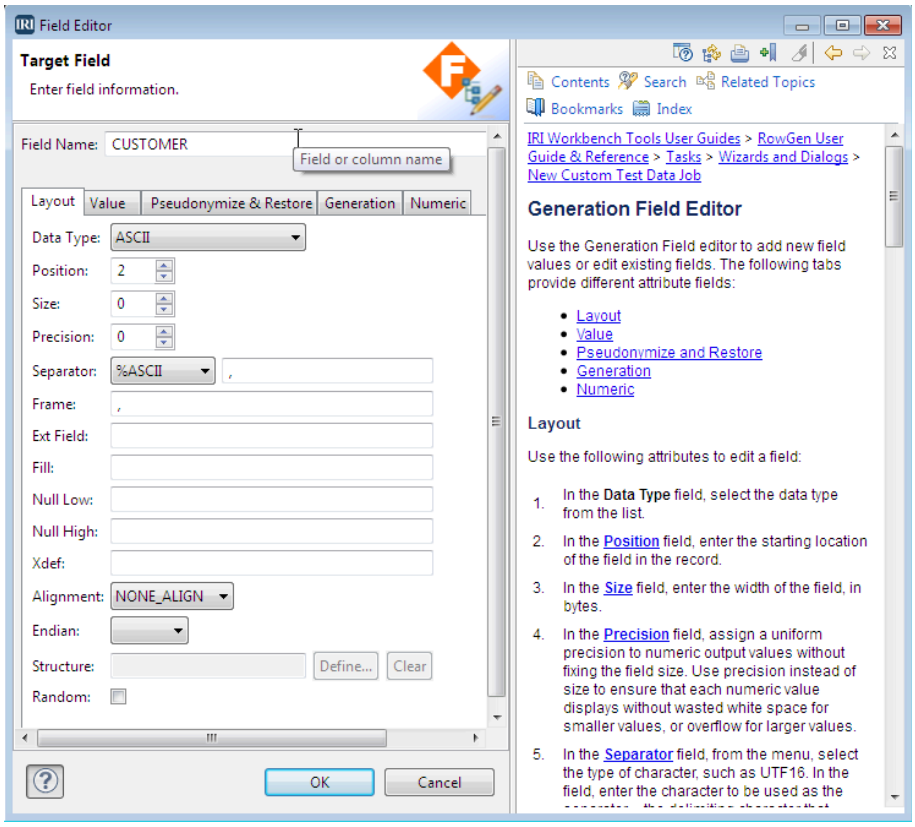
The first RowGen wizard creates single, ad hoc test data sets structured in custom flat-file, report, or ODBC target formats. For a complete list of platforms you can populate directly, see:

http://www.iri.com/products/workbench/data-sources



This wizard also supports:

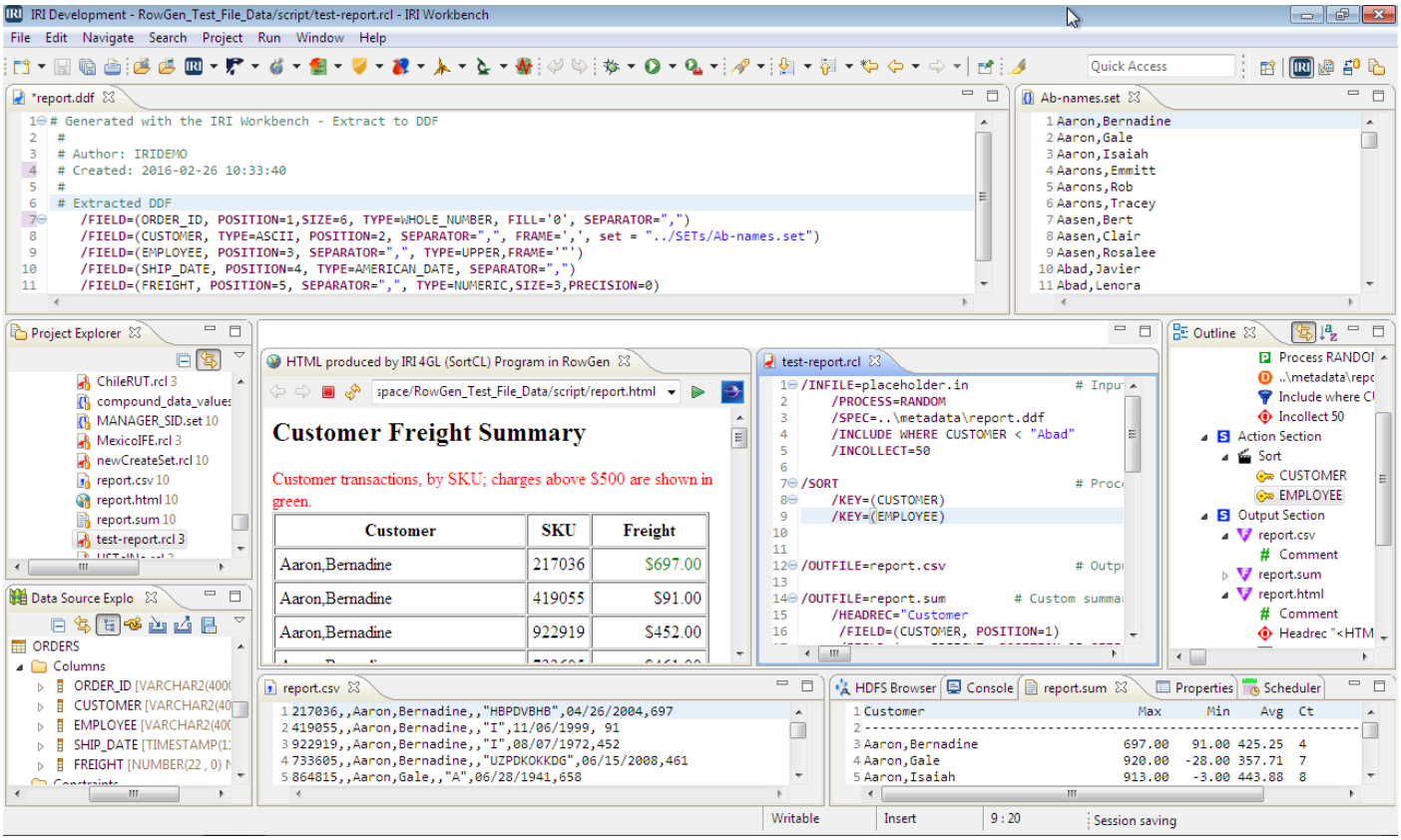
- sorting on any number of keys,
- all and valid (joined) pairs
- composite data types and values
- customized distributions
- randomized real data



Help-enabled dialogs in the wizard support customization of the generated values and their layouts in each target.

The dialogue can also open from the GUI later, so you do not have to re-run the wizard or learn the scripting language to modify test data or target attributes.

At the end of the wizard, an initial RCL job script is created with the test data you have defined. That script can then be expanded upon to create additional, and/or even more elaborate targets:

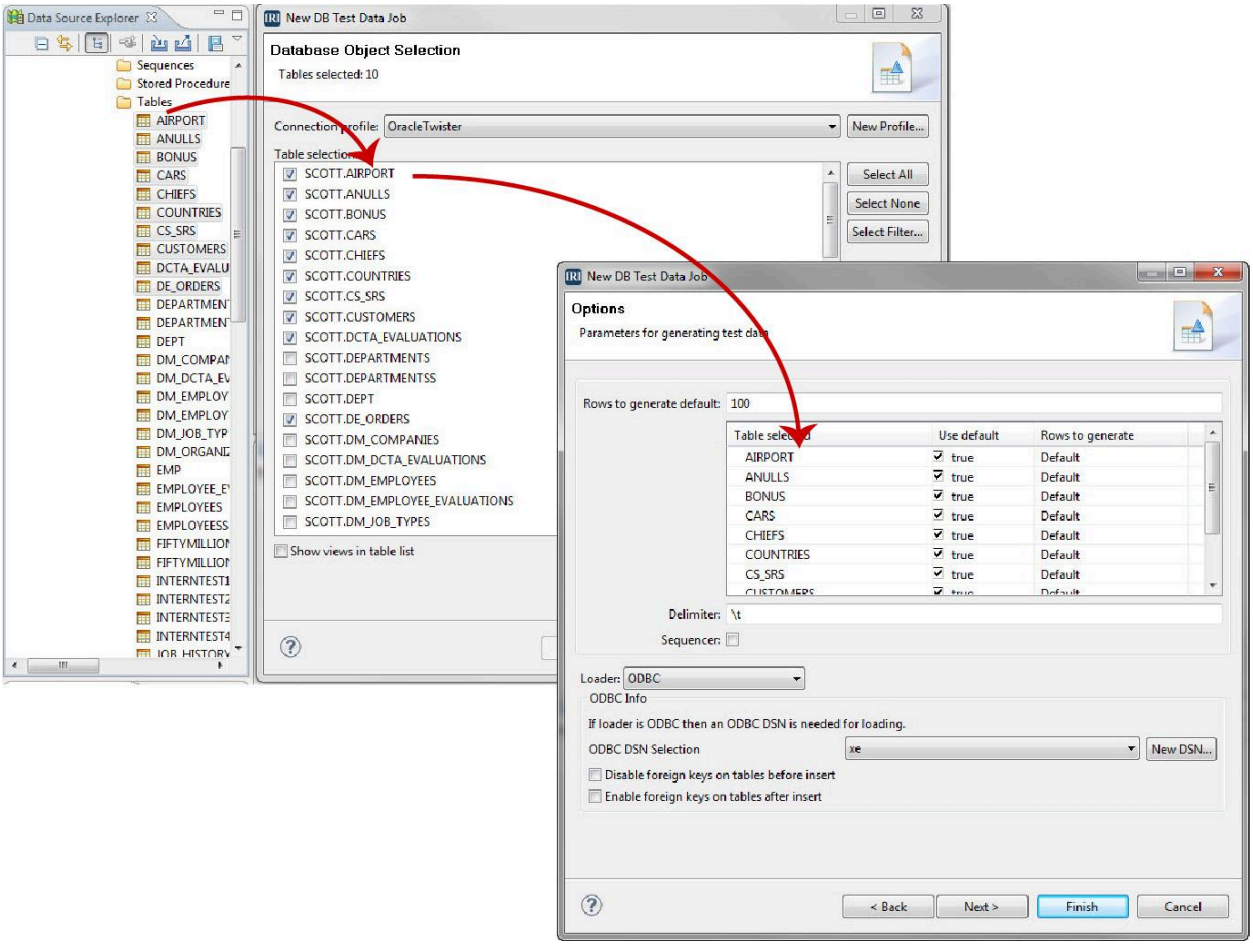


New *DB Test Data* Job Wizard

RowGen automates the creation and loading of synthetic, production-quality data sets for relational, star, and other database schema through the *New DB Test Data* wizard. It handles:

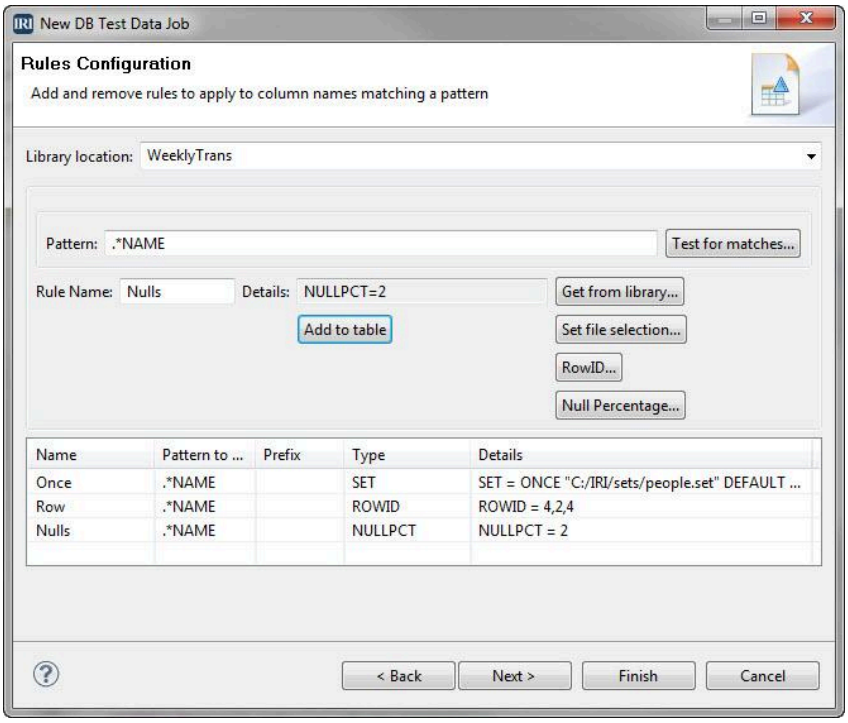
- *Parsing* - by selecting the schema and tables you want to populate, RowGen translates the database table descriptions and integrity constraints into RCL scripts that specify the source structure, dependent sets, and data creation.
- *Generation* - by building and running the RCL scripts to create one test file per table that can be bulk loaded, and/or saved for future use and/or modification.
- *Population* - by bulk loading the target tables in the order necessary to maintain all primary and foreign key relationships.

Begin with a JDBC-connected DB schema displayed in the Eclipse Data Tools Platform (DTP) Data Source Explorer in the GUI. Right-click on the (empty) tables you highlight for population and choose *New DB Test Data Job*, or launch the wizard from the RowGen menu in the toolbar.



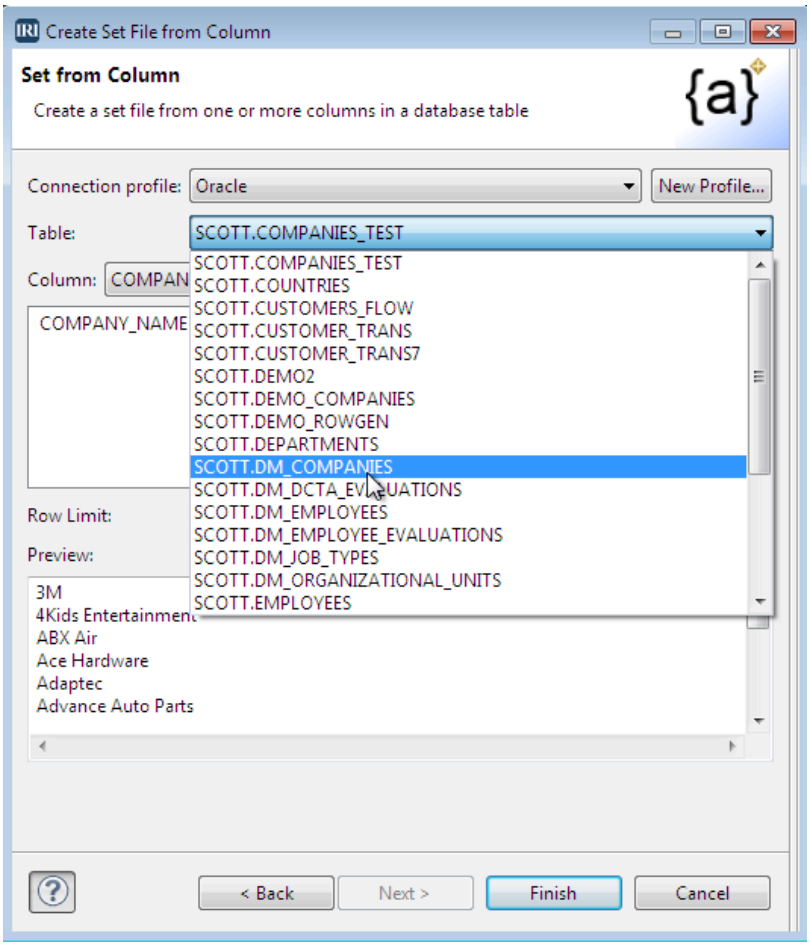
The wizard reads the DB structure and relationship information to build new, referentially correct test data and bulk-loading the tables automatically though individual, batched job scripts (or a full XML workflow you can run and visualize in IRI Voracity). When you run the final batch job from the GUI or command line, all of the table-specific, inter-dependent test data generation and DB load scripts run in the order necessary to maintain referential integrity.

Applying Rules



The '*New DB Test Data Job*' wizard in the IRI RowGen GUI supports the specification and re-use of cross-table data generation rules for columns whose names conform to regular expression patterns.

This allows you to populate target tables with the same kinds of randomly generated or randomly selected test values in those columns.



For example, you can specify that real company names randomly selected from a set file will appear in every company name column throughout the database. The 'Set file selection …' option runs another wizard to allow you build a (random draw) set file from real table data.

You can choose the random draw method for each column, too, as "any", "all", "once," etc.

RowGen has many options for instilling data realism into both key and non-key columns. The rules you apply are your own, and many customization options exist. The values that rules consistently populate across tables also preserve referential integrity among those tables.

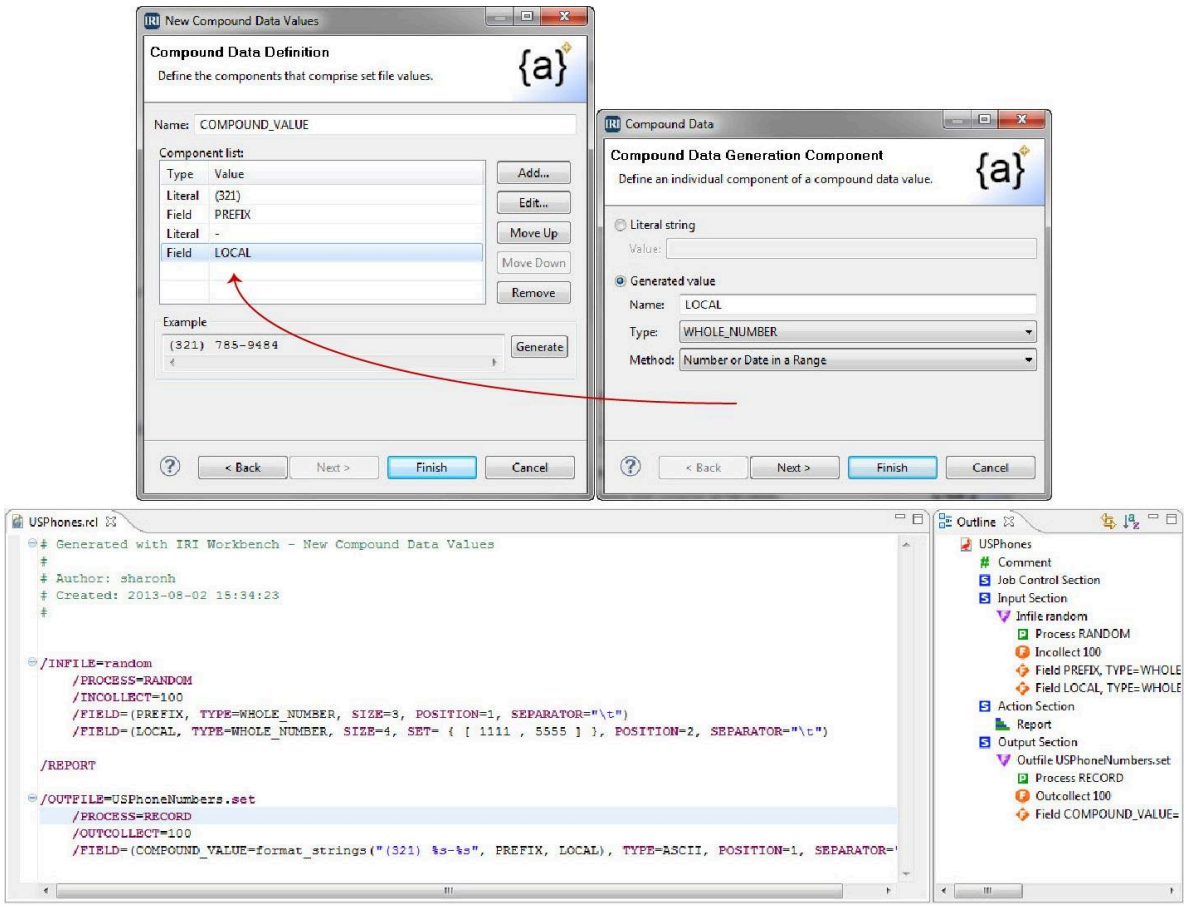
Using “New Set File” Wizards to Customize Data Values

RowGen allows you to fully customize the content, appearance, and frequency of generated data using the GUI’s point-and-click interfaces (wizards and dialogs), or its syntax-aware job script editor (where you can manually specify and change the same options). If you use the IRI Voracity platform with RowGen built-in, you can also access workflow and transform mapping diagram palettes, and manage individual target and batch flow properties even more visually.

RowGen wizards and functions let you create, and randomly draw, from ‘real data’ sets that:

- pre-exist, and you designate as single or multi-column flat files
- come from a connected DB column, or columns for joined (valid) pairs or "all pairs"
- consist of new types (like phone numbers) built on the fly as ‘compound data values’
- contain random values in specified sizes and ranges in the data type(s) you choose
- use custom field functions that create computationally valid (or invalid) data. IRI provides built-in field-level functions for: credit card numbers, US and Korean social security numbers, GUIDs and UUIDs, and national ID values for Italy, the Netherlands, and Singapore. You can write your own using the same C-compatible calling conventions.

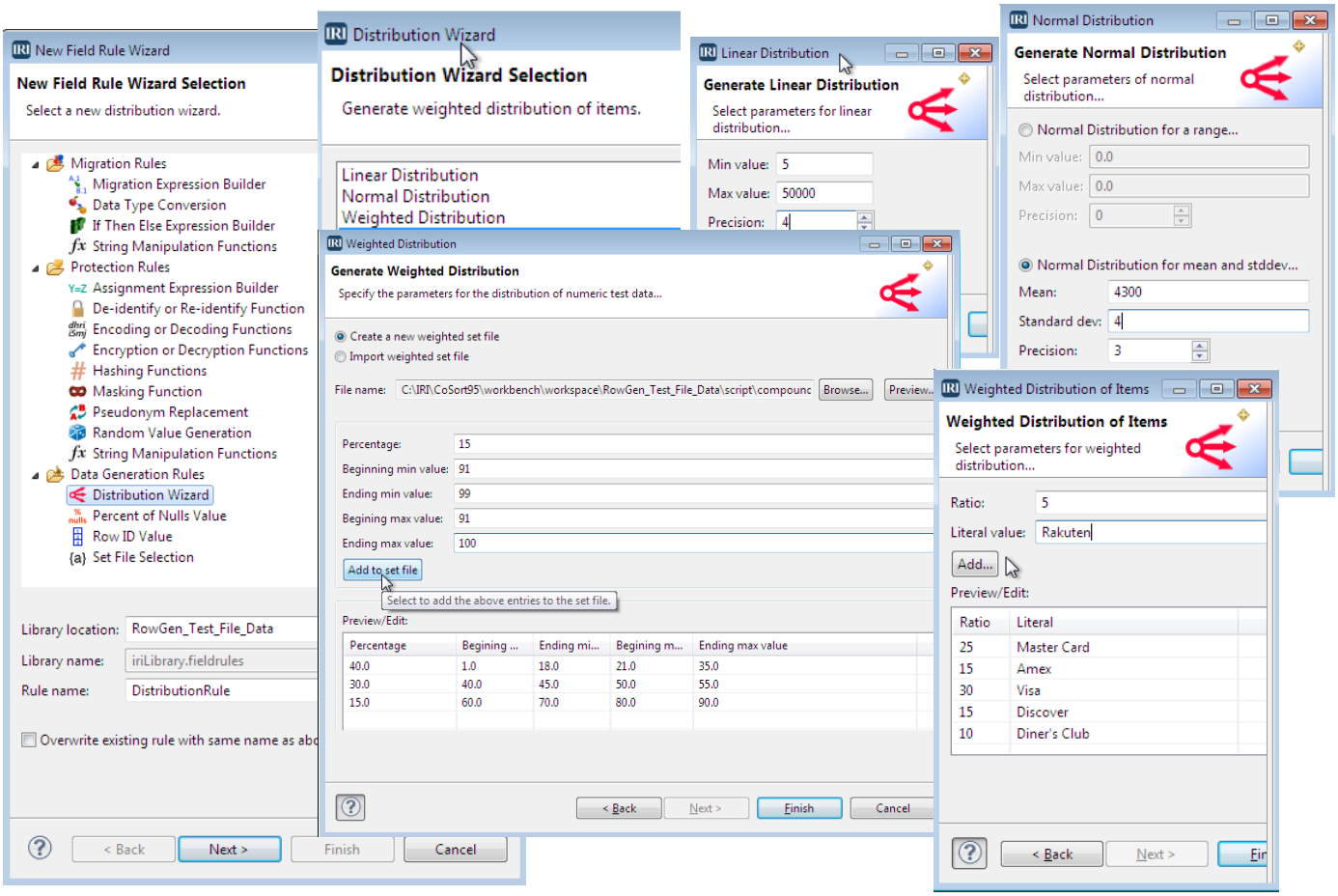
Field values not pulled from sets are randomly generated. Generation functions respect the field size and data type declared, and values display in the correct positions.



The 'Compound Data Value' wizard supports any combination of set-file or random-generated data elements and literals within new fields you construct (like master data definitions).

Using “Distribution” Wizards to Customize Value Occurrences

RowGen also allows you to specify various ways to mimic the natural frequency of values.



Modification Options

From within the RowGen GUI in IRI Workbench, or in any text editor, you can easily modify:

- test value attributes (field specifications in each RowGen script)
- set file values and random set-file selection methods
- value generation ranges and distribution (frequency) criteria
- field-level manipulations and functions applied to the test values
- target row layouts (e.g., field position and formatting commands)
- database load utility configurations files or other target details
- the layout, appearance, and other elements of a custom test report

Because the RowGen GUI is built on Eclipse, you can use embedded text searches and find/replace functions to search across projects and metadata repositories to facilitate job modification. Distributed metadata management systems like Git can control data definition (layout) and job control (script) file versions, user access rights, and test data or script lineage.

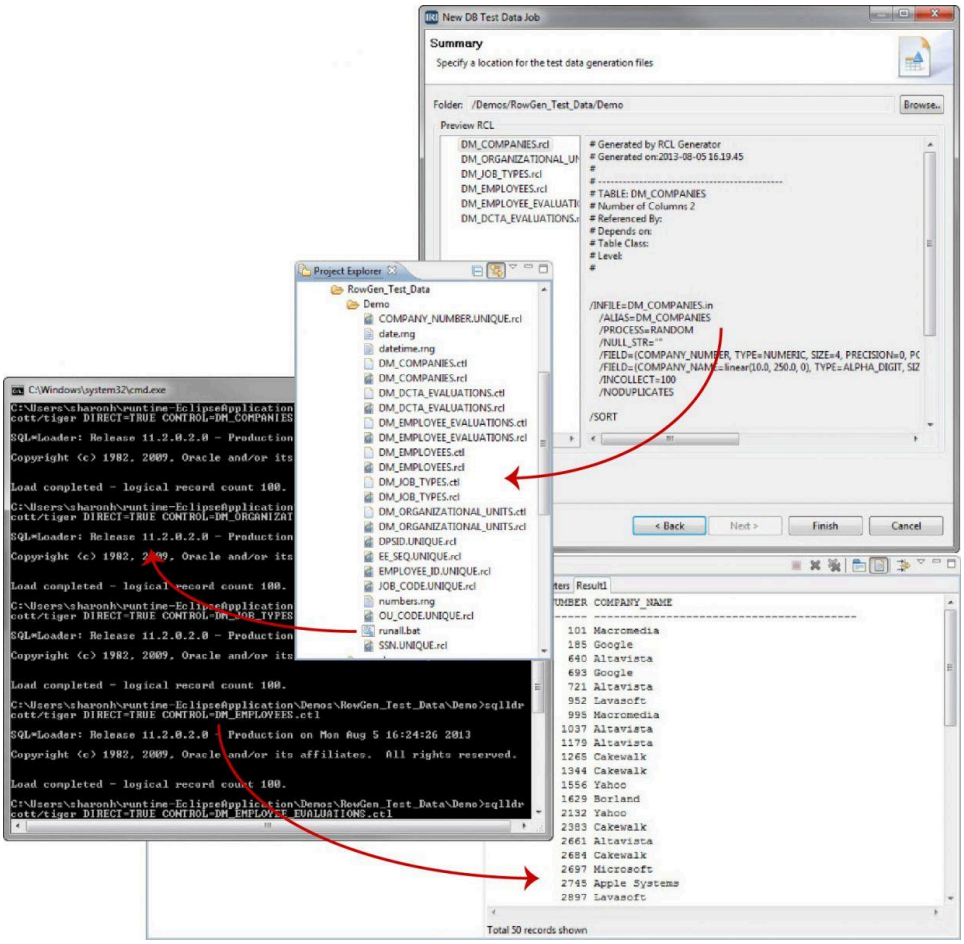
Target Population

RowGen Control Language (RCL) job scripts have three main phases: input, sort, and output. Test data actually gets created in the input section, with each field statement specifying random value generation or set file selection. Below possible pre-action reformatting and sort criteria, the output section is where one or more targets are simultaneously defined, each with optional:

- file name and format, DNS name and table structure, or named report layout
- specification and positioning of the generated column values
- row counts, and/or conditional value inclusion or omission
- transformations; e.g., cross-calculation, aggregation and type conversion
- report-oriented field labels, page headers, footers, and numbers
- markup tags for XML, HTML, or other formatting embellishments
- custom-incrementing “RowID” numbers for indexing

Create and modify these attributes in the GUI, or in any text editor, and re-run the jobs with low row counts until you are satisfied with the output.

Such custom test targets are less common in bulk database prototypes, where the layouts are pre-determined by the DDL information of the tables, and auto-defined in the RowGen *New DB Test Data Job* wizard. The wizard builds multiple RCL scripts, along with bulk DB load scripts, and a batch file. That batch job generates and loads all the structurally and referentially correct test data into the target tables simultaneously.



This dialog is part of RowGen's 'New DB Test Data Job' *wizard. Towards the end, it displays the test data generation scripts it built; each one is for a separate target table (where the metadata was gleaned from the original DDL).*

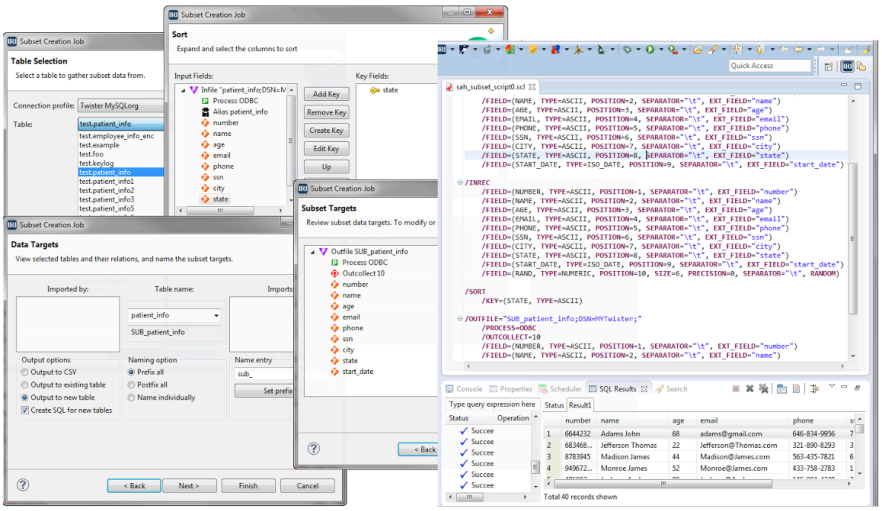
This script runs in a batch sequence with others, and bulk loads the data (pre-sorted) into target tables in order, for the sake of referential integrity.

By pre-sorting the test data in primary key order, and bypassing the loader's native sort, bulk loads are much faster, and the queries will be faster, too.

The GUI's 'Run' menu provides ad hoc, batch, or scheduled execution options for local Windows, or remote Windows or Unix, platforms. TDM web portals from Cigniti (BlueSwan) and ValueLabs (TDH) can also launch RowGen synthesis and FieldShield masking jobs in on-demand DevOps.

Whether you create one, or one thousand RowGen job scripts and test sets, both the metadata and data are accessible inside and outside the RowGen GUI (IRI Workbench) environment. Freely deploy, share, and modify the text materials on the command line or any Windows, Linux, or Unix platform. Eclipse file management, job execution, and version control tools like EGit, SVN, and CVS are at your disposal.

Database Subsetting



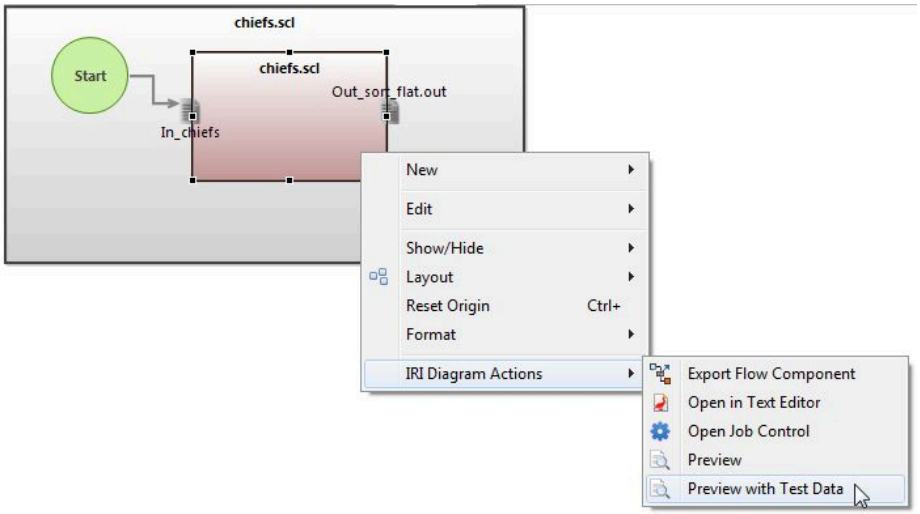
In addition to generating synthetic test data, RowGen users can also produce database test data by creating plain or masked *subsets* of tables with referential integrity.

A DB subsetting wizard allows you to define the size, maskings, and mappings of smaller copies of data for testing.

ETL Task Previews

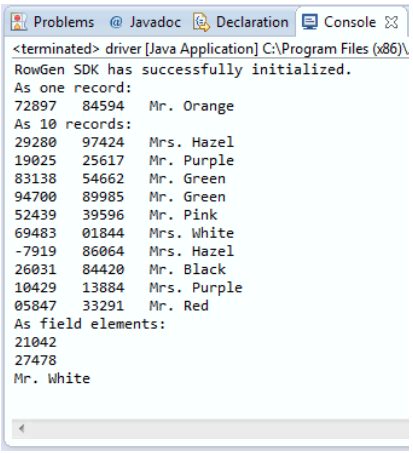
The Voracity GUI allows ETL architects to prototype their mappings using RowGen-built test data.

Use the workflow palette to specify one or more tasks (flowlets) within a larger project (flow), edit properties, and preview mapping targets using a subset of production data, or random test data.



RowGen SDK

IRI also has a Software Development Kit (SDK) for RowGen that Java programmers can call to generate test data dynamically in applications, or across Hadoop nodes. This article introduces the library, and how you can use it to generate custom records with either randomly generated synthetic, and/or randomly-selected real, field values.



Test Data as a Service (TDaaS) and Other Test Data Management (TDM) Options

1. Test Data as a Service (TDaaS), is a remotely provided professional engagement with IRI using RowGen or any data masking and subsetting features in Voracity to provide highly customized test data without licensing, learning, or maintaining new technology.

2. Run RowGen (or FieldShield) as a command task in a CI/CD pipeline. See these examples for:
   a. Amazon CodePipeline
   b. Azure DevOps
   c. GitLab
   d. Jenkins

3. Run RowGen (or FieldShield) jobs with these DB virtualization tools, which call the IRI job scripts at cloning time:
   a. Actifio
   b. Commvault
   c. Windocks

4. On-demand TDM web portals are tightly integrated with RowGen software too, including:
   a. Cigniti BlueSwan
   b. ValueLabs Test Data Hub (TDH)

INNOVATIVE ROUTINES INTERNATIONAL (IRI), INC.

2194 Highway A1A
Melbourne, FL 32937 USA
Phone +1 321-777-8889
iri.com/rowgen