QOSQO

Version 1.0

July 2019

QUIPU4 WHITEPAPER

QOSQO BV
De Loop 40
5688 EW Oirschot
The Netherlands

Phone: +31 (0)499 577 562
E-mail: info@QOSQO.nl
http://www.QOSQO-services.com

IBAN: NL72 ABNA 0421 6436 17
C.o.C.: Midden Brabant: 17226645
VAT: NL.81.94.65.847.B.01

GENERATING
DATA SOLUTIONS

Contents

INTRODUCTION 3

DATA IS KEY 3
EVOLUTION OF AUTOMATION 3
LOOKING AT PATTERNS 5
RISK REDUCTION AND SHORTER TIME TO MARKET IN DATA MANAGEMENT 5

QUIPU4 8

THE INCREDIBLE MACHINE 8
INTRODUCTION TO QUIPU4 9
MODEL-TO-MODEL TRANSFORMATION 10
STEPS IN THE GENERATION PROCESS 11
MODULAR APPROACH 12
OUR VISION 13
DEALING WITH NON-PATTERNS 15

Introduction

Data is key

Data becomes available in enormous quantities from very diverse sources. Put it to use to add value.

Data has always been key in organisations. Whatever activities a company engages in, all its processes result in data collection to record and smooth these same processes. And the same data is needed for reporting and analysis purposes.

Initially, most data came from in-house source systems or from partners in the value chain. However, over the course of time we've seen an exponential growth in data worldwide.

Enabled by technological advances in computing power, connectivity and cheap storage we have witnessed the rise of the worldwide web and social media participation. Billions of smartphones have driven down the cost of cameras and sensors, giving birth to the Internet of Things (IoT). All these devices generate vast amounts of data that can be used to tailor services to the individual and automate very complex processes (like self-driving cars).

These developments open a world of completely new opportunities for any organisation to explore if data can be made available on time, of good quality and in a repeatable way.

Today, turning data into information requires data processing at an unprecedented level. Organisations are seeking to extract more and more value from data from various sources and in various formats.

Evolution of automation

The emergence of best practice modelling techniques to capture and store data has given birth to data warehouse automation solutions.

Companies have been involved in defining and implementing solutions to turn data into information ever since the first databases. Initially, small datasets were transformed into what we now call data marts. Over time, the number of data sources grew, and more and more data needed integration. The building up of history in the database was required for time-series analysis and auditing. These needs gave rise to the data warehouse. During this period, organisations were focusing mainly on extracting structured data from database systems.

The multitude of hand-coded interfaces between source systems and analytical applications proved hard to manage and maintain. To get around all this coding we've seen software companies jump on the band wagon. Initially by delivering so-called ETL (Extract, Transform and Load) tools, to make the design and implementation partly graphical and hence easier to maintain and transfer.

In 1997, when Dan Linstedt documented and shared his idea known today as Data Vault, it boosted the development of a new kind of solution that falls into the category data warehouse automation. Data Vault as a modelling technique is very compelling and as a result delivers a well-defined but rather extended model structure in the data warehouse that screams for automation.

It was this moment in history that gave birth to the earlier versions of QUIPU, taking ETL to the next level by *generating* data model and data load code. By that time QUIPU distinguished itself because of its capability to also detect model changes and apply them to the models and previously generated code.

Today, however, as mentioned before, the data warehouse is no longer the leading architecture to cope with data integration and historisation for analytical and reporting purposes. As companies go through a digital transformation process, they strive to become more data-driven. Other architecture designs have been introduced – involving distributed file systems, for example – that are geared towards capturing data in large quantities and in various formats in hybrid environments often called data lakes. These developments call for a more open approach to automating data management processes, that we call 'model-driven development'.

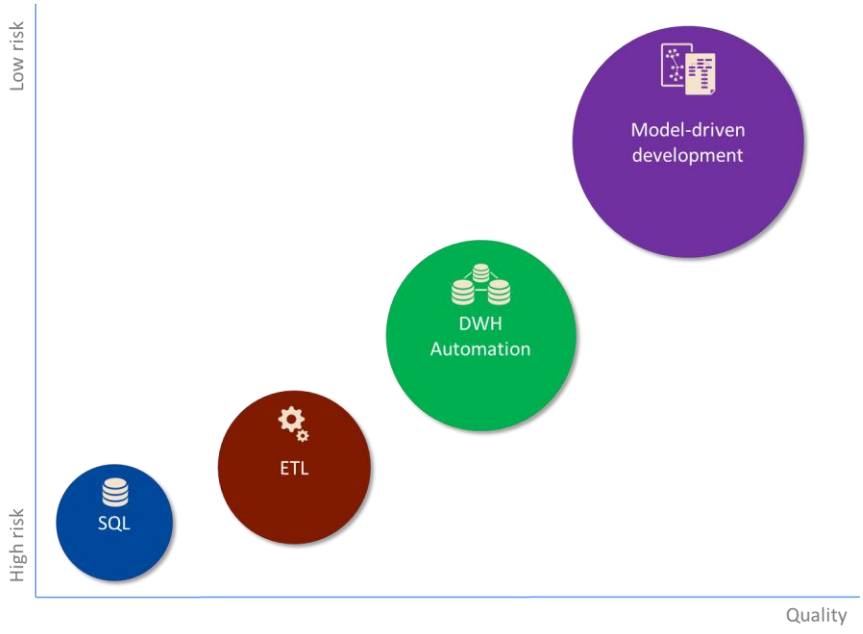

Figure 1. Evolution of automation

Looking at patterns

The focus on automating patterns using a generic approach is the foundation of QUIPU4.

Generating Data Vault structures and related load code is essentially applying a design pattern to a source model, transforming it into a Data Vault structure.

This same pattern-based concept can be used to transform data models to other types of data storage models and data models, a challenge that can be recognised in a variety of use cases.

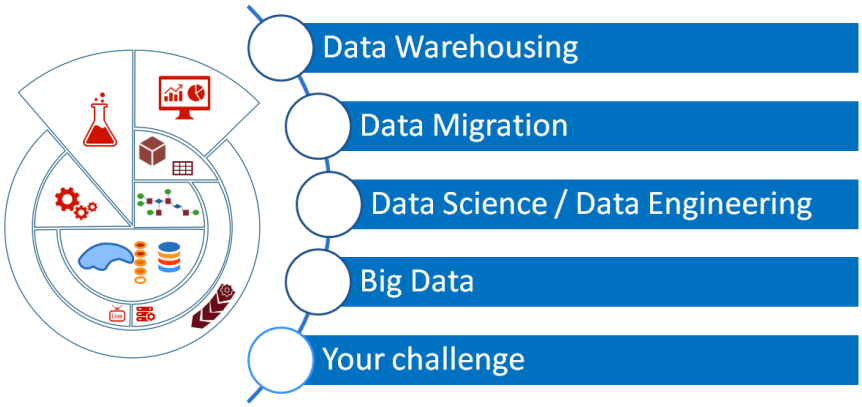

Figure 2. Patterns are everywhere

When looking at all this and considering QUIPU3, the team behind the solution concluded that a next step in automation was required. One that focuses on the underlying patterns rather than on specific architectures (like the layered data warehouse) we distinguished so far. This resulted in a complete redesign of QUIPU, leading to the release of QUIPU4, a model-driven automation tool.

Risk reduction and shorter time to market in data management

Current data management practices are rigid: they can't cope with changes in data quality and structures. This is a serious risk. We need adaptive systems to reduce this risk.

Apart from the fact that developers like QUIPU's capability to reduce repetitive tasks, in the end, risk reduction and a shorter time to market is what counts for an organisation.

Companies don't like complex software structures. Companies don't want to be dependent on individual developers or freelance consultants. Companies need to avoid tailor-made solutions they can't maintain or extend, and which are bound to collapse when the creator moves on.

In an era where traceability and accountability are key and privacy counts, data management solutions need to be solid, reliable and auditable. Data lineage needs to be fully in place. Investment in data management solutions should count. Companies are no longer willing to build a solution that only holds for a few years.

Nor can they afford solutions that take ages to build. Businesses need to be more and more agile on the market they operate on and data and information accessibility should facilitate this.

And yet, today's solutions tend to implement only the 'happy flow': the data will flow through the system if nothing goes wrong and nothing changes. When the flow is disrupted, it requires manual intervention from highly skilled and knowledgeable data engineers. When a disruptive pattern is detected, an automated fix is designed and implemented for only this situation.

That approach will not hold in the future. Not only is the number of data sources growing exponentially making the process more vulnerable and time consuming with every new source, but most of these sources are not (or not fully) under your control: they are created and maintained by third parties and tend to evolve over time. In most cases, you only find out your source has a different structure when you try to extract data from it.

Introducing a technology that captures any data without regard for content or structure (like Hadoop) is only kicking the can down the road: the data will have to be structured at some point to bring it into a form that allows for analysis. The chain will break at the point when the assumed structure is no longer valid.

We therefore need systems that examine the structure of the data (the data model) from a source each time data is extracted, detect any changes, analyse the changes and send alerts and reports to the people responsible. But we can take this one step further and use smart algorithms to determine the best approach (a pattern!) to deal with the detected changes and adapt all extraction, load and transformation code to reflect these changes. The system can thus minimise data flow disruptions and dramatically reduce the number and scope of manual interventions. Hence reduce the risk in data management and speed up time to market of new data solutions.

	No automation High risk				Ultimate Automation No/low risk
	None	Reusable objects	Metadata-driven	Change handling	Adaptive
Models	Manual design	Import existing models (reverse engineering)	Fully metadata-driven forward engineering	+ Model change detection	Self-learning autonomous model generation and adaptation
Load Routines	Manual coding, some code copy-paste	Template-based, manually adjusted	Template-based, parameter driven	+ Data migration	Self-learning autonomous load routine generation and adaptation
Transformations	Manual coding	Template-based technical transformations		Rule-based specifications in metadata	Self-learning autonomous rule specification and injection
Process Control	Simple flow, fixed execution order	More advanced scheduling	Load execution in metadata		Automated closed-loop
Metadata	No use of metadata	Import/export of basic metadata	Multiple metadata repositories with exchange	Integrated metadata	One single metadata repository
Deployment	Environment agnostic, tool specific	DTAP metadata in one repository		Assisted change release	Automated

Figure 3. Risk reduction in data management

QUIPU4

The Incredible Machine

Do you remember the Incredible Machine? It was a game that enslaved players worldwide from 1993 onwards.

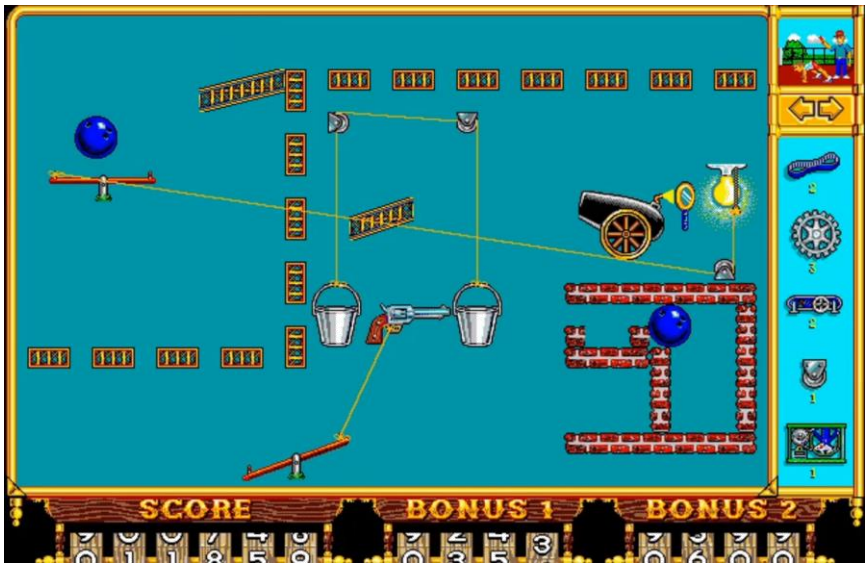

Figure 4. One level of the Incredible Machine game, 1993, Sierra Entertainment

The general goal of the game is to create a series of Rube Goldberg devices: arrange a given collection of objects in a seemingly needlessly complex fashion so as to perform some deceptively simple task (e.g. "put the ball into a box" or "start a mixer & turn on a fan").

The available objects ranged from simple ropes and pulleys to electrical generators, bowling balls, and even cats, mice and humans, most of which had specific interactions with or reactions to other objects (for example, mice will run towards nearby cheese). The levels usually have some fixed objects that cannot be moved by the player, and so the only way to solve the puzzle is carefully arrange the given objects around the fixed items.

The solution you thus try to create only *seems* to be overly complex. The restrictions that are presented in the arrangement and availability of the various components turn the task into a real challenge.

There is also a "freeform" option that allows the user to "play" with all the objects with no set goal or build their own puzzles with goals for other players to attempt to solve (source: Wikipedia).

The idea of building your own flow of objects – using specific components from a toolbox – to achieve a certain result and the fun in doing so inspired us to build QUIPU4. Of course, leaving out the 'needlessly complex fashion' and replacing that with risk reduction capabilities and efficiency in building data solutions.

Introduction to QUIPU4

QUIPU4 delivers the next level in data management automation using patterns as the guiding principle.

QUIPU4 takes things a step further by introducing the next level in data management automation using patterns as the guiding principle. This makes data warehouse automation, data migration, big data applications and similar projects much faster and easier.

We strongly believe in leveraging the knowledge of data engineers and data architects. By automating straightforward tasks and thus killing tedious, boring jobs we bring back the fun. Allowing them to focus on the complex tasks ahead when business rules need to be applied and to spend more time satisfying analytical business needs and truly participating in creating value for the company.

The previous version of QUIPU4 already offered software and services that generate data warehouse solutions based on Data Vault. QUIPU4 takes this to the next level and generates highly specific data management solutions based on standard patterns wherever possible. Including data warehouses, data migrations (e.g. to the cloud) and big data applications.

QUIPU4 also supports the specification of mutations where standard patterns do not suffice. Thus, the names of attributes may be changed manually to be in better alignment with business terminology. Moreover, specific business rules can be specified that are inserted into the QUIPU4 generated code. These exceptions are kept in the repository when models are regenerated (e.g. to accommodate changes in the source systems). In this way, regeneration will apply the same patterns again but will also re-apply the mutations, resulting in the expected outcome.

Upon its initial release, QUIPU4 only supports a few basic – but already very useful – scenarios. QUIPU4 allows you to read the structure of any source database that can be accessed via a JDBC driver, and generate a suitable staging area for your data warehouse and finally a straightforward source data warehouse layer – that we have named the Historical Data Archive or HDA – that will capture all the changes in the data of the source database. QUIPU4 also

generates all the ETL code needed to extract the data from the source database and feed it into the HDA, via the staging area.

If you require the Data Vault model for your source or business data warehouse layer, we currently refer you to QUIPU version 3.4 that is still available and supported while we transfer functionality to QUIPU4. You can run QUIPU 3.4 and QUIPU4 in parallel.

Model-to-model transformation

The QUIPU4 modules operate on data models (set of tables) instead of single tables.

QUIPU4 is primarily a model-to-model conversion tool. And this is precisely where it distinguishes itself from existing ETL tools and other data warehouse automation tools. These tools do a good job in transforming and manipulating data at the entity or attribute level but lack the support for defining and developing transformations at the model level. It is at this higher level that patterns can be found and applied and can thus be generated.

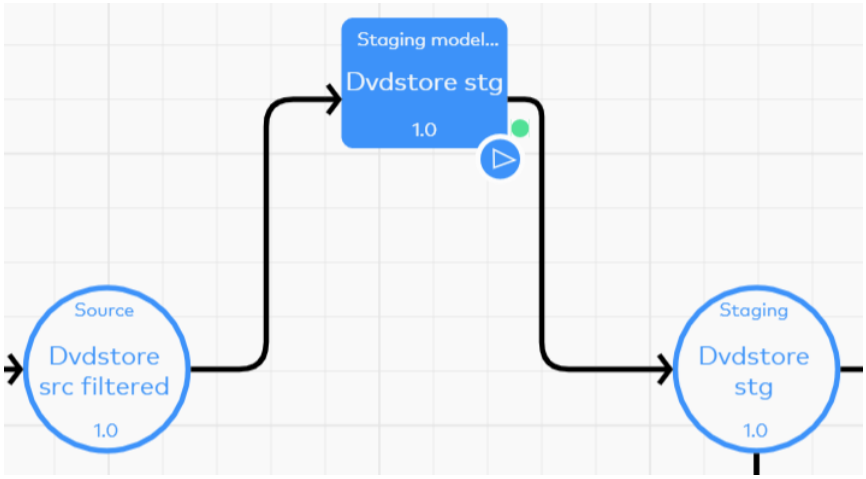

Figure 5. Model-to-model conversion using QUIPU4

This is a far more efficient way of developing data management and data warehousing solutions. QUIPU4 can, for instance, take an entire staging data model as input and generate a model that stores data historically.

QUIPU4 not only generates modified data models but also keeps track of the mappings between source and target models. These mappings allow QUIPU4 to generate the ETL (SQL) code – or views, if that is possible and preferred – to move the data in the production environment. At the same time, these mappings are the basis for lineage reporting and analysis.

QUIPU4 generates all output using open templates, so the output can be easily adjusted for specific database platforms or customer standards.

QUIPU4 generates code based on the design of a data flow. The code will need to be executed (and scheduled) to implement the solution in your environment.

Steps in the generation process

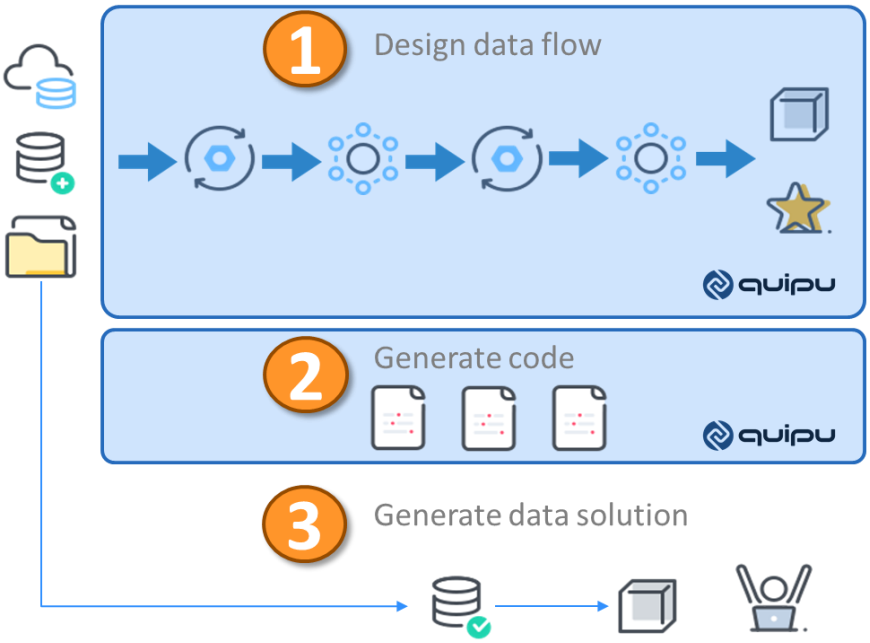

Figure 6. Steps in the generation process

Step 1

Using QUIPU4, an architect follows a certain path in designing and implementing data management solutions. Using the diagram in QUIPU4 a data flow can be designed. This data flow is a relevant combination of connected objects where modules are available that have been created to support a certain pattern. Examples of modules include data source model reverse engineering, staging layer modelling, staging layer ETL-code generation, HDA layer (Historic Data Archive) modelling, HDA layer ETL-code generation and table filtering. This list will grow in the future based on customer requests for new patterns to be automated.

Step 2

Once the data flow has been designed, the flow can be executed (step by step or in its entirety) resulting in model and code generation.

Step 3

This code is then implemented on the platform of choice. The data solution is ready to be used in production.

QUIPU4 offers you building blocks (modules) to design your solution. We will expand the set of modules to support broader and smarter scenarios.

Modular approach

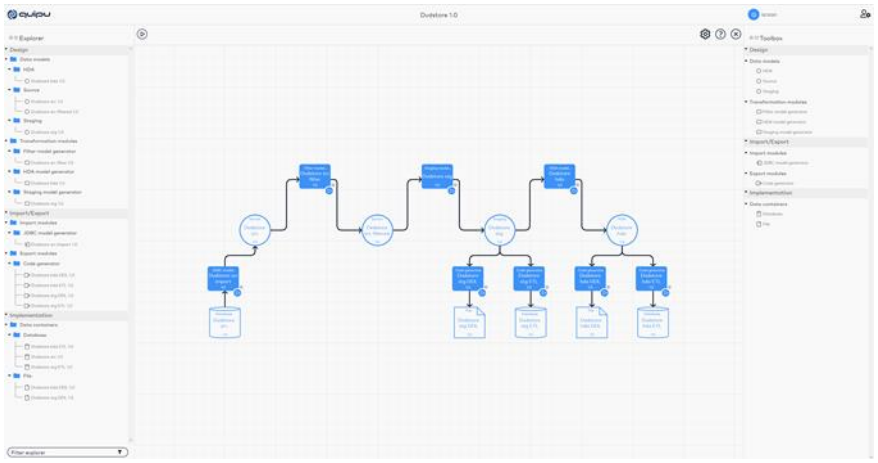

Figure 7. Sample data flow in QUIPU4

QUIPU4 provides maximum flexibility in designing and automating whatever flow of architectural layers is required. The new web-based graphical user interface provides a canvas where one can select and configure the desired modules to achieve this. We can develop any new modules required in a short period of time, leading to an expanding set of building blocks for you to choose from.

In the example in figure 7, in what we call a diagram, a typical data warehouse architecture is designed using the Dvdstore data sample case. From left to right a connection is established with the data source Dvdstore. The first module (blue block) is capable of reverse engineering the data source model properties (tables, attributes, data types, keys, relationships). This results in a description of the Dvdstore source model that can be examined.

Next, the list of tables and attributes can be filtered to exclude those that are not required for implementation in the data warehouse. This results in a limited (filtered) set of tables and attributes that can be examined again.

The next module performs staging model generation resulting in a Dvdstore staging model in the repository. From there, other modules can be used that can generate the data definition code (DDL) and data manipulation code (DML) or ETL code in the format required by the target platform (Oracle, SQL Server, etc.).

Similarly, a Historic Data Archive model (HDA) can be derived and generated and stored in the repository followed by modules that generate the accompanying DDL and DML.

This is just one example. Imagine what other data flows can be designed in this way. The team that develops QUIPU4 can create new modules in a matter of weeks. Whatever your data solution challenge is.

For a live demonstration and explanation of how to build and execute a data flow watch this video on YouTube (18 min 35 sec).
[https://youtu.be/LwBBADTiuRQ]

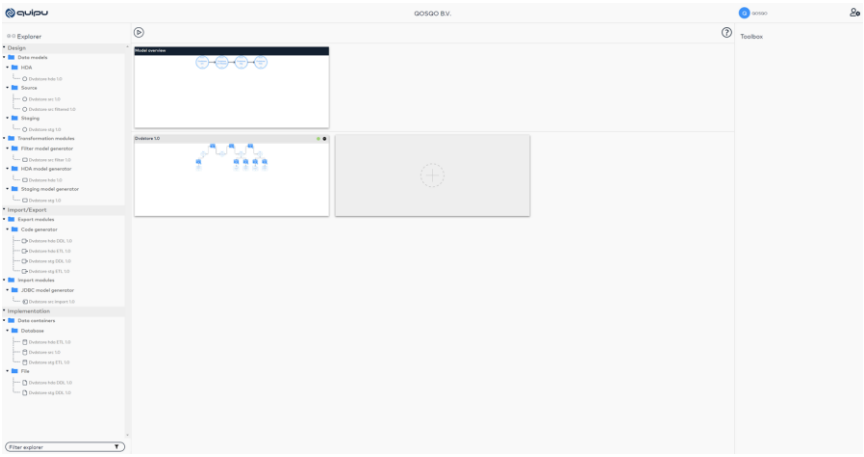

Figure 8. Sample diagrams containing different data flows

All configured objects in *any* diagram are stored in the repository and are shown on the left-hand side. By a simple drag and drop these flows can be reused in other new diagrams for the fast design of additional data flows.

Our vision

QUIPU4 is an added-value component to your existing technology stack and integrates with existing tools and technologies.

QUIPU4 aims at reducing the risks in data management by relying on one single metadata repository and by generating model and load code for whatever architecture is desired – by focusing on patterns. That is why in the QUIPU4 roadmap we are also looking at self-learning autonomous model generation and adaptation. The current focus is on creating the ability to report data lineage to comply with traceability and accountability demands.

QUIPU4 is an added-value component to your existing technology stack. It integrates with existing tools and technologies. There is no vendor lock-in. QUIPU4 is fully metadata-driven. Models and transformations are based on source and/or business models.

QUIPU4 relies on an open and documented REST API interface to the server components. All changes to the repository are recorded.

All models are reverse engineered or generated, but exceptions can be recorded as mutations. These mutations in turn are kept (and reapplied) after regeneration. Models can also be imported/exported using Excel giving even more control.

All code generation is performed by translating the contents of the repository to executable code via template sets. These template sets can be optimised per platform for database (appliance), data integration or documentation purposes.

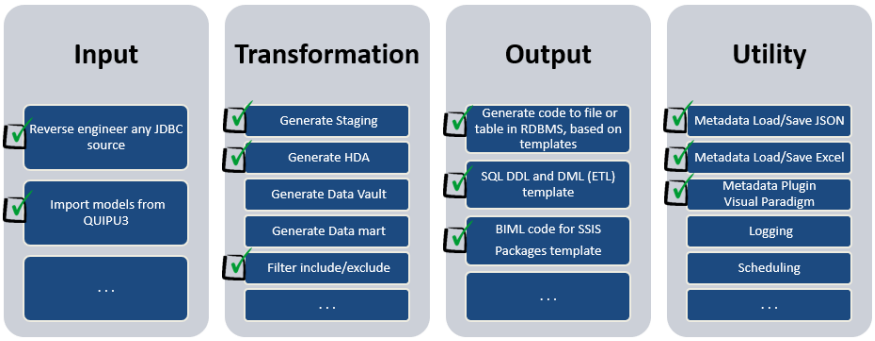

Figure 9. Modules available in QUIPU4 as of April 2019 (and growing)

The figure above shows what modules are available for building data solutions. Currently, existing customers drive the creation of new modules. In close cooperation with them, we thus add more value every day. Reducing risk at an ever-increasing speed. Giving architects the flexibility they need.

The phases of new module creation we provide to our customers are:

Phase 1: focus on functionality

- Develop a new module as soon as possible, focus on functionality
- Expose the configuration via REST API
- Front end automatically visualises configuration as JSON
- Validate input on back end
- Roll out solution to ***customer***

Phase 2: focus on user experience

- Adapt design GUI from phase 1
- Add validation in front end
- Add dedicated visualisations
- Roll out solution to ***market***

You can modify the metadata for individual tables, attributes or mappings to deal with situations that do not follow a pattern. You can also modify the templates to customise the implemented pattern.

Dealing with non-patterns

The unique approach of QUIPU4 saves tremendous amounts of time by applying patterns to transform models. But what about the situations when the pattern does not apply?

One could easily manually modify the code produced by QUIPU to deal with those exceptions. But that would reduce the benefit of having QUIPU to a single run (because QUIPU would overwrite all your changes on the next run).

Of course, one could save changes in a separate file and compare the newly generated QUIPU code with the manually changed code to find and reapply these changes. Obviously, that is a very tedious and error-prone process.

QUIPU has automated this approach. QUIPU always keeps generated models (and mappings) separate from any changes that were made (manually via the user interface or by uploading files). Every time a model is regenerated, a new version of the generated model is stored, and the previous modifications are re-applied automatically.

NOTE 1: It remains the customer's responsibility to validate that the manual changes made are still valid for the data model that has now changed due to the regeneration.

NOTE 2: At this point, there is no direct access to either the different versions of the generated models or the manual changes made. Providing this access is high on our priority list and will assist in managing the changing environment.

GENERATING
DATA SOLUTIONS

More information:
quipu.nl
datawarehousemanagement.com