MOODY'S ANALYTICS

WHITEPAPER

JULY 2013

Author
Brian Heale,
Senior Director

Contact Us
For further information please contact:
Brian.Heale@moodys.com

Alternatively, you may contact our customer service team:

Americas	+1.212.553.1653
Europe	+44.20.7772.5454
Asia-Pacific	+85.2.3551.3077
Japan	+81.3.5408.4100

Analytical Data: How Insurers Can Improve Quality

In the second in a series of papers, which focus on key data topics, Brian Heale gives his view on the types of analytical data required for Solvency II and capital/risk decision making with a particular focus on the techniques for improving quality.

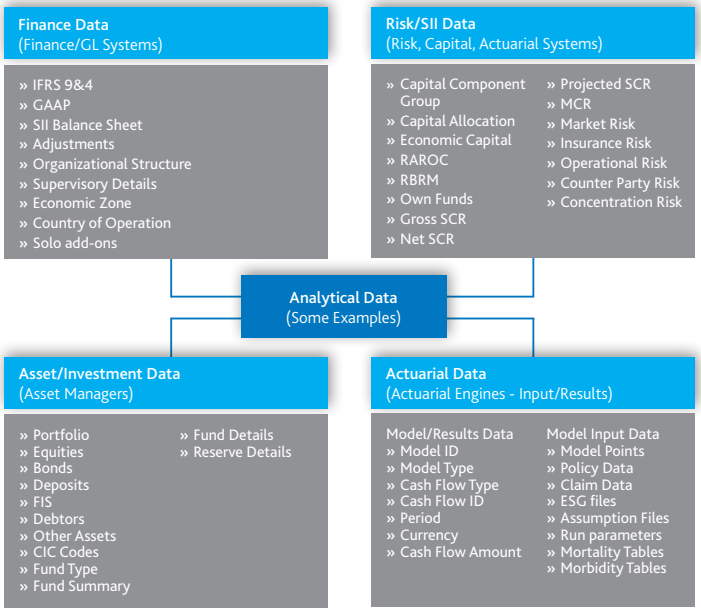
This paper has been written for risk, capital and finance practitioners within an insurance organization and aims to demystify how IT techniques can be used to improve data quality.

Introduction

In the previous paper[1] we examined the overall data requirements for key regulatory initiatives such as Solvency II (SII), the new International Financial Reporting Standards (IFRS) and for capital/risk decision making in the business. In this paper we will delve more deeply into the topic of ***analytical data***: what it is and how various tools, techniques and approaches can be employed to improve the quality of that data. The subject of spread sheets is also examined because these are, and will continue to be, an important source of analytical data. Regulators, such as the Prudential Regulation Authority in the UK, have acknowledged that such sources of analytical data will not go away.

What is Analytical Data?

Analytical data can be defined as the actuarial, finance, investment and risk data required for SII/IFRS and multi-faceted management and business reporting. The following table illustrates, with some examples the types of and provenance of data that can be classified as analytical.



1 Data Governance Best Practice: Smoothing The Way To Solvency II Compliance: http://www.moodysanalytics.com/~/media/Insight/Regulatory/Solvency-II/Thought-Leadership/2013/2013-25-04-Data-governance-best-practice-smoothing-the-way-for-Solvency-II.ashx

Analytical Data is Different

Analytical data by its very nature is different from the transactional or operational data that insurers are traditionally used to handling and storing. The key differences are summarized below:

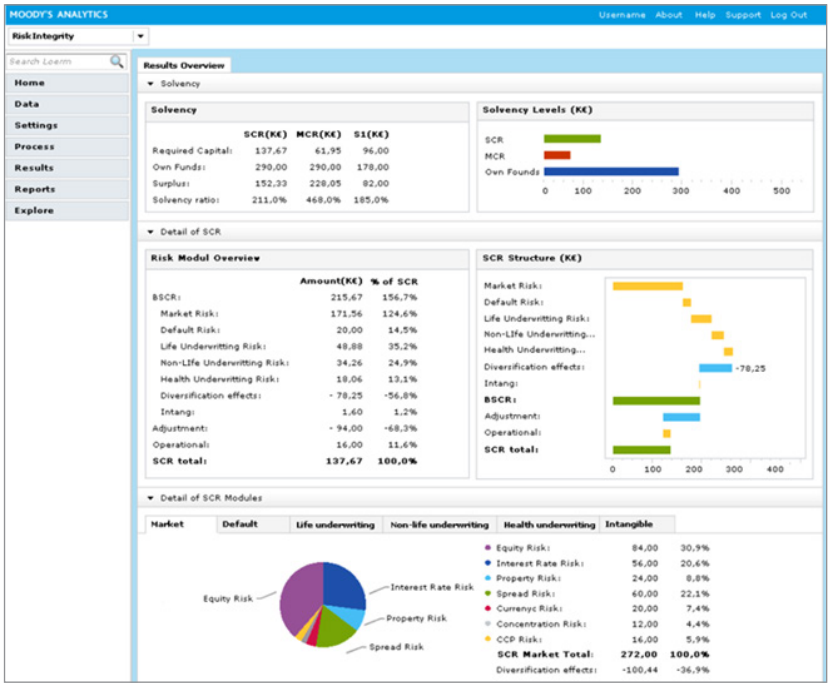
- Analytical data comes from a range of *different sources*—primarily finance and actuarial systems, but also from, for example, external fund management systems. Some of these sources are desktop based systems and spreadsheets, and because these are not linked in to the main IT infrastructure, problems with extraction and standardization often occur.
- Analytical data typically requires a much higher level of granularity in the data than is required for regulatory and compliance reporting purposes. For example, the asset and technical provisions Quantitative Reporting Templates (QRTs) require very granular data, and the Asset D1 template can potentially require millions of lines of asset transaction details.
- Analytical data may well have to undergo complex aggregations and transformations (e.g. the generation of model points). Equally, the data has to be carefully reconciled into a *single source of truth*. This is because the same data may come from different sources, for example premium data may come from both administration systems and general ledgers and must be reconciled.

Solvency II is acting as a catalyst driving insurers to consider how they are going to handle analytical data: are they going to adapt their existing transactional data repository, or build a new one? Practice varies but we are seeing a trend towards insurers building an *Integrated Risk and Finance Data Repository*.

Perhaps a good example of the different characteristics of analytical data is in the actuarial arena. There are potentially hundreds of inputs to an actuarial model: mortality tables, Economic Scenario Generation (ESG) files, assumptions sets, run parameters, and policy data. There is significant value in storing these in a repository to enable enterprise access, management and audit controls. Equally, there is value in storing the ***output*** *of actuarial models (e.g. cash flows, loss triangles, P&L simulation and capital allocation) in a sufficiently high level of granularity. For example, the TP - E3 Claims template requires over 10,000 separate figures. Storing such granular data is particularly useful for sensitivity analysis and populating the technical provision QRT templates, but most modelling technology doesn't have this capability.*

Business Intelligence and Reporting

The whole purpose of extracting and storing analytical data is to be able to use it to good effect. For insurers that primarily means relying on it for regulatory and management reporting and capital/risk decision making. Ease of analysis and interpretation can be best achieved when information is presented in a graphical format – a dashboard interface is therefore to be regarded as the ideal – as illustrated on page 3.



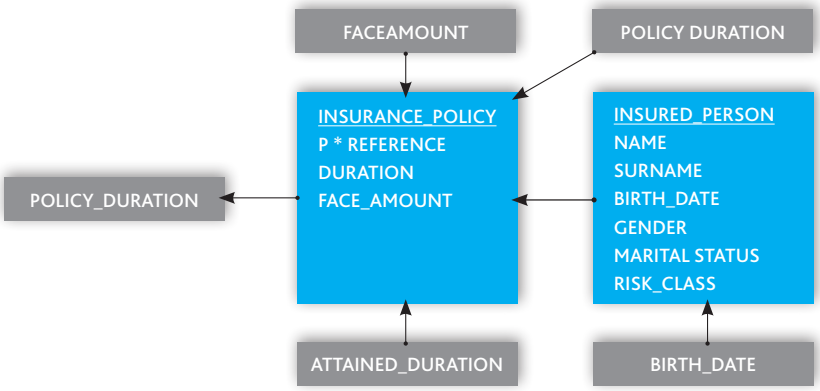
For such a dashboard to have real value, users must be able to drill down to the underlying data. In order to achieve this, the underlying data has to be carefully structured and stored in a star schema* or snowflake schema in a relational data warehouse or in a special-purpose data management system.

However, in order to actually generate the dashboards, users need to build On-Line Analytical Programming (OLAP) Cubes to structure and present the data in the right way.

*An **OLAP cube** is required in order to present the data in a userfriendly, graphical manner as this will facilitate ease of assessment and interpretation. An OLAP cube is an array of data within which the cells comprising the cube hold a number that represents some measure which represents an element of the business such as premium, claim, capital, expenses, budget or forecast. The OLAP cube effectively provides the drill down capability and granularity required. While business users need to carefully define the data required and the level of granularity it is often an IT function to construct the OLAP cubes.*

The diagram below represents an extract from a typical database structure:

Database Structure Extract



* A star schema is a database structure comprising a number of primary (or master) data tables referencing any number of other dimensional tables. The star schema is an important special case of the snowflake schema, and is more effective for handling simpler queries. A snowflake schema is simply a logical arrangement of tables in a multidimensional database such that the relationship diagram of the tables resembles a snowflake in shape.

Examples of Regulatory and Management reports that extensively utilize analytical data are listed below. Clearly these vary considerably by the type of insurer and environment in which they operate, but the volume of reports is staggering: typically hundreds of reports have to be produced at both local entity and group level.

Regulatory Reports	Business Reports
» SII QRT Templates	» Balanced Scorecards
» Solvency II Risk Margin	» Key Performance Indicators
» SFCR	» Market Risk Dashboard
» RSR	» Profitability
» CRSA	» Costing
» Use Test	» Budgeting
» Regulatory Returns - PRA (UK) Forms and equivalent across Europe - BAFIN, DNV	» Variation Reporting
	» Exception Reporting
» Tax Reports	» Analysis and Trending
» VAR Reports	» Forecasting
» EV/EEV/MC EV	» What if/scenario planning
» Analysis of Change in EV/EEV/MC EV	» Reconciliation process
» IFRS 4 Phase II returns	» Sensitivity analysis
» IFRS plus sensitivities and stresses	» Experience analysis
» ICA (UK)	» Group Capital Adequacy
	» Capital Allocation
	» Risk Adjusted Return Measures
	» Valuation
	» Analysis of Profits/losses
	» Historical Asset Shares
	» Exposures

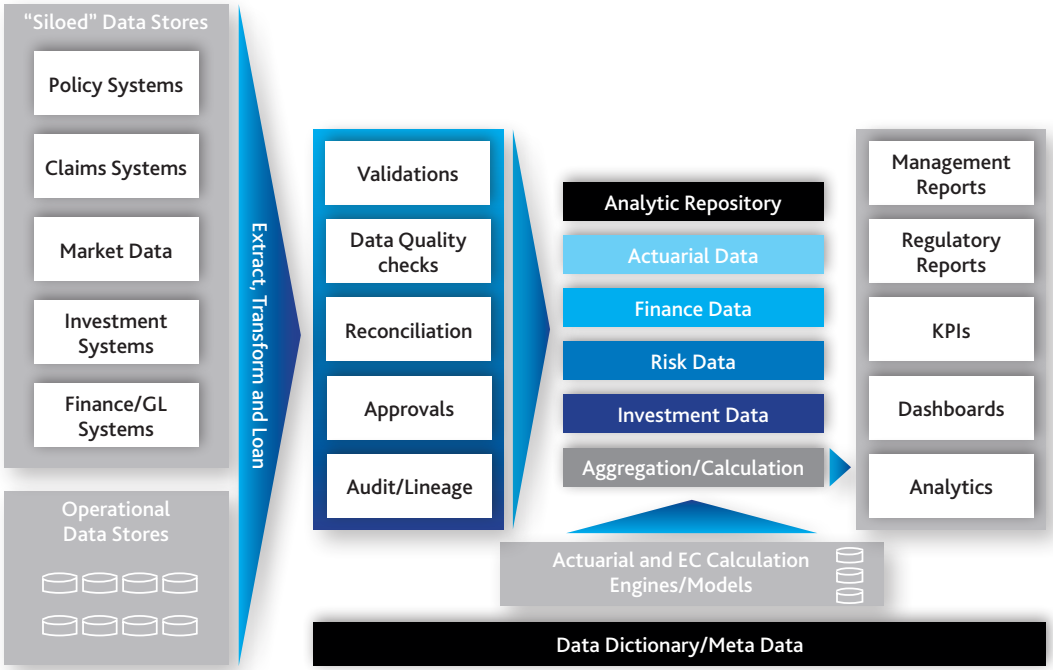
Data Quality Processes Improvements

Data Silos

The numerous sources of raw analytical data within an organization give rise to questions regarding its quality, consistency and reliability, and this is particularly the case as the volume of data increases. To compound the problem, risk data is often organized in to separate silos. This sometimes means that there is duplication of data and inconsistent values. Risk data therefore comes disaggregated in multiple silos according to different dimensions, such as legal entity, line of business, risk category, etc. The silo approach produces a general tendency towards low quality data, mainly due to the proliferation of data duplication and multiple data quality approaches from one silo to the next.

An example of data quality issues can be found in the multiple policy administration systems maintained by insurers—each of which may store a policy holder's age and birth date in different formats. Similarly, Premium data may come from both a policy administration system and the general ledger, and they rarely match.

One answer to the problem of data quality is to establish a central analytical repository and use technology to improve the quality of data and reconcile it. This approach gives you ***a single source of truth*** to provide consistency in reporting and decision making. The diagram below illustrates how data can be drawn from multiple silos/source systems and effectively improved and centralized.



Data Quality/Accuracy

Data can be considered of high quality if it is fit for purpose in terms of its intended use (e.g. statutory reports, business planning and decision making). What makes up data quality? Here are the major factors, again bearing in mind the purpose:

» *Accuracy*
» *Completeness*
» *Appropriateness*
» *Relevance*
» *Consistency*
» *Reliability*

The quality of data in most insurance organizations is often quite poor so improvement is essential. Improving the quality of data, is however a multi-faceted process that takes raw data and subjects it to a set of algorithms and business rules. This, coupled with expert judgment, enables the data to be validated or corrected. The data quality tools used to do this have inbuilt data "logic" in terms of patterns, trends and rules built up over a number of years. Data is tested against this logic. Simple errors can thus be automatically corrected and flags raised for data that requires expert judgment. The end result may not always produce perfect data (no process can do that) but the data should at least be fit for purpose.

The following table looks at a typical process that raw data may go though to improve the quality. The steps may not necessarily follow this order.

Checklist for Data Improvement

Process	Description
Data Extraction – the "ETL" process	» Extract data from the various source systems – policy and claims systems, general ledgers, actuarial systems, investment system, etc. » Store in a staging area of the repository. » There are specific Extract, Transform and Load (ETL) tools to control this process. Alternatively, data may be captured in Spreadsheets, CSV files or even input manually. The data then needs to be loaded into the staging area of the repository.
Data Profiling	Utilize data profiling techniques to make an initial assessment of the data to understand its overall quality challenges and anomalies. This primarily utilizes patterns, trends and algorithms (both general and specific to the insurance industry) to produce a picture of the overall data quality – typically expressed as a percentage of the data that appears accurate.
Data Accuracy	Execute a series of data quality checks/rules against the data. There are a number of data quality tools that include many thousands of pre-built data quality rules (both general and industry specific) and these are then supplemented with a number of user-defined rules. Best practice is to execute the data quality rules within the repository. Alternatively this can be undertaken in the extraction (ETL process).
Generalized "cleansing and deduping"	This is a dual process: **Data cleansing** » Identify and modify corrupt or inaccurate data from the repository. » Remove or modify incomplete, incorrect, inaccurate, irrelevant data. This process may involve removing typographical errors or validating and correcting values against a known list of entities. The validation may be strict (such as rejecting any address that does not have a valid postal code) or fuzzy (such as correcting records that partially match existing, known records). After cleansing, a data set will be consistent with other similar data sets in the repository. **Data de-duplication** » Reduce storage needs by eliminating redundant data. » Retain only one unique instance of the data within the repository. Redundant data is replaced with a pointer to the unique data instance.
Quality Monitoring	Keep track of data quality over time. Use software to auto-correct the variations based on pre-defined business rules. The process should only be repeated on values that have changed; this means that a cleansing lineage would need to be kept, which would require efficient data collection and management techniques. These processes can be in real time or batch oriented.
Enrichment	Enhance the value of data held in the repository by appending related attributes from external sources (for example, consumer demographic attributes or geographic data). This may be valuable for the underwriting and pricing of household policies (e.g. flood risk data) or marketing certain types of products to certain customers who have certain socio-economic attributes, e.g. high disposable income.
Load	Store validated data in Results and Repository.

Making Use of Data Profiling Tools

We can examine the data available in a data repository and make an assessment of its quality and consistency, uniqueness and logic using data profiling techniques. This is one of the most effective ways of improving data accuracy in an analytical repository. A number of proprietary data profiling tools are available from leading vendors.

Data profiling utilizes different kinds of descriptive techniques and statistics such as minimum, maximum, mean, mode, percentile, standard deviation, frequency, and variation as well as other aggregates such as count and sum to analyze data according to known patterns. Using these it is possible for an expert to find values that are unexpected and therefore potentially incorrect. Profiling can help insurers identify missing values which can then be replaced by more logical values generated by data augmentation algorithms.

The key benefits of data profiling are summarized below:

Data profiling will help you to:	
1	Understand any inherent anomalies in the raw data.
2	Determine whether the data can be used for multiple purposes and assess the impact of usage in new/other applications The insight gained by data profiling can be used to determine how difficult it will be to use existing data for other purposes.
3	Generate initial metrics on data quality.
4	Improve search capabilities by tagging or using key words.
5	Assess whether the metadata accurately describes the actual values in the source database.

Introducing Data Quality Rules

In order to further improve ***data accuracy***, firms must execute a number of data quality rules against their data. Various vendors offer data quality tools which include thousands of rules. These comprise a number of generic rules together with some specific to the insurance industry. Additionally, such tools also enable insurers to define supplementary rules specific to their own lines of business or function. Some examples of types of data quality rules are set out below:

Type of Rule	Description	Insurance Example
Basic Business Rules	Basic business logic rules .	» The date of a claim cannot be earlier than the date of inception of the policy.
Data-Type Constraints	Values in a particular column must be of a particular data type, e.g., Boolean, numeric (integer or real), date, code, etc.	» CIC/ISO Code in asset data » Premium frequency – monthly, quarterly, annually.
Regulatory Constraints	Data validations or rules laid down by regulators such as EIOPA/PRA.	» The Validation rules contained in the QRT templates specified by EIOPA.
Range Constraints	Typically, numbers or dates should fall within a certain range. That is, they have minimum and/or maximum allowable values..	» Eligible ages for insurance contracts - eligible age must be between say 18-65. » Minimum premium amounts in GDP. » Reinsurance limits.
Mandatory Constraints	Certain columns cannot be empty, e.g. not null.	
Unique Constraints	A field, or a combination of fields, must be unique across a dataset.	No two policyholders can have the same national insurance number.

Type of Rule	Description	Insurance Example
Set-Membership constraints	The values for a column come from a set of discrete values or codes.	For example, a person's gender may be Female, Male or Unknown (not recorded).
Foreign-key constraints	This is the more general case of set membership. The set of values in a column is defined in a column of another table that contains unique values. terminology.	For example, in a "town" column it must be aligned to a county from a "County Table".
Regular expression patterns	Occasionally, text fields will have to be validated this way.	For example, phone numbers may be required to have the pattern (0044) 1234-5678 or DD/MM/YYYY format.

While it is the role of the IT department to execute data quality rules it is up to the practitioners in the business to provide the "logic" in conjunction with rules that are very specific to a particular set of data. When discussing this logic with IT, consider carefully the ultimate usage of the data. For example, for policy data the input required for actuarial modeling is primarily around the type of contract, benefits/ coverage, premiums, term, etc. These have to be correct as they impact on the accuracy of the cash flows. Other policy related data such as post code, phone number, etc. are not relevant for these purposes and if incorrect have no impact on accuracy.

Where to Execute the Rules?

Executing the data quality rules during the transformation phase of the ETL process is a logical choice when it is required to transfer data directly from one IT system to another. However, when applied to analytical data this has a number of problems:

1. ETL tools are complex and not readily accessible or comprehended by the business people who play an important part in the process.
2. Managing a large number of data quality rules in the ETL process can be complex and may result in duplication of some rules and inconsistencies.
3. Under Solvency II the users, actuaries, risk mangers etc, must be confident that the data is accurate—to do this they view the data, and this is best done via the repository rather than the ETL process.
4. Finally there must be reconciliation of data from multiple sources particularly accounting data and again this is best executed in the repository.

Our suggested approach is to execute the data quality rules in the repository after the data loading as this allows the business users to have direct access to the data quality checks (even if they are not directly responsible for the data quality). This means that the business users, in their business as usual activities such as modeling and auditing, can have a view of the data quality and address each single data check.

Furthermore, embedding the quality checks inside the repository enables the full traceability and auditability of the data and shows the regulator whether or not low quality data has been accepted too early in the process. Finally, as we have mentioned above, Solvency II requires reconciliation against accounting data. Executing data quality rules after data loading makes the reconciliation process both easier and more traceable.

What is Metadata?

IT departments often use the term "Metadata". In simple terms Metadata is just data about data – basically descriptive information about a particular data element. IT describes what a particular data element is, how and when and by whom a particular set of data was collected, and how the data is formatted.

Metadata is essential for understanding information stored in a data repository as it describes the structural components of underlying tables and their elements. For example, metadata about an element could include data types, name of data, size and many more characteristics such as length of fields, number of columns, where the tables are located and other relevant information.

One of the main uses for metadata is to provide a link between the people who created the data and the people who are actually going to use it. Metadata allows the users to speed up the search for individual data by allowing users to set parameters for searches, allowing the filtering of unwanted information.

There is often confusion between meta data, a data directory (specifically required for SII) and a data dictionary:

» **Metadata** – high level descriptive information about data as described above.

» **SII Data Directory** – non-technical description about data that ordinary business users can understand. The data directory (or data glossary if you prefer) is meant to be a documented repository which different users can use to understand which data is being used, where it comes from (source systems, how it's being used and what its specific dependencies and characteristics are. This is a requirement of the Solvency II Data Governance Framework. The traceability of data through the various layers is best represented diagrammatically to help understand where the data comes from.

» **Data Dictionary** – is more of a technical descriptor that typically lists all the tables, fields, primary keys, foreign keys, etc) that are available in the repository, the number of records in each table, and the information about the fields

The Role of Spreadsheets

No review of analytical data quality would be complete without considering the role of spreadsheets. Spreadsheets are now commonly considered as part of the wider group of technology assets called end-user computing (EUC)—that is any technology asset which may be created, updated or deleted outside the purview of the IT department or standard software development lifecycle management processes. Other assets in this class include MS Access databases, CSV files, MatLab scripts, etc. However, spreadsheets tend to be the most prolific and problematic of all EUCs because of their flexibility and familiarity to most users.

The ability of a spreadsheet to act as a data source (e.g. data connections/links, expert judgment), a data manipulator (e.g. policy data, assumption tables, run parameters) and as an application (e.g. formulas, macros) create a variety of data quality issues. When combined with the additional uncertainty of ownership and access rights over specific spreadsheets, it is not surprising that spreadsheet control issues have received specific mention in data thematic reviews conducted by the FSA (now PRA).

Spreadsheets pervade many financial and actuarial processes in insurance but the regulatory focus of Solvency II has been drawn particularly to spreadsheets that hold and manipulate data prior to its utilization in calculations and the internal model. It is very common to find extensive 'webs' of thousands of spreadsheets connected by spreadsheets 'links' that may have an impact on data quality. In practice many of these are dormant, but their presence and the possibility of erroneous updates creates uncertainty and risk in the modeling process.

Resolution of the problem can be seen in terms of three steps: ***discovery, triage and control***.

Discovery is the process by which businesses can evaluate their current dependence on spreadsheets. Such a 'health check' will consider the scope and complexity of spreadsheet usage. It can be done manually but may be accelerated using technology.

Once the spreadsheet landscape is mapped and analyzed, the future of identified spreadsheets and spreadsheet-supported processes can be '**triaged**' for different forms of improvement. This may simply be a matter of training users to further exploit existing solutions, or require the adoption of new software capability through integration or third-party vendor solutions. Triage perspectives include ranking spreadsheets by:

- Financial materiality – Sarbanes Oxley, Solvency II.
- Operational inefficiency – by costs and hours of rekeying.
- Operational losses due to errors, missed deadlines, revenue leakage, fraud, compliance fines. Also consider exposures created by MS Access and PC based databases, the three dimensional equivalents to spreadsheets.

It is likely that the spreadsheet triage process will produce a road map for process improvement that will take some years to complete, so that business-critical spreadsheets will continue to exist in the business for a lengthy period. Furthermore, constant business innovation will undoubtedly continue to create more spreadsheets. Both these factors mean that spreadsheet elimination should be seen as a continuous process, rather than a planned destination. Businesses are therefore increasingly turning to technology to provide on-going spreadsheet **control** in the form of enterprise spreadsheet management software. This provides the opportunity to detect and report user activity which is outside pre-determined tolerance levels across the key risk areas of data, functionality and security. However this does not mean that some spreadsheets cannot be eliminated by:

- Expanding the functionality of existing systems or by building new dedicated systems.
- Uploading unique data in the spreadsheet into user defined tables within the repository with enhanced reporting generated from the repository.
- Reporting from the spreadsheet can be migrated to a business intelligence system or OLAP dashboards.
- Some or most of the spreadsheet model and data is removed in several steps, reducing risk along the way until the spreadsheet is finally eliminated.

Conclusion

There is little doubt that analytical data is a foundation not only for Solvency II and IFRS programs but also better informed capital/risk decision making. Insurers need to master data management and can do so by leveraging the following ten principles:

1. Analytical data (actuarial, finance, risk and asset) is very different in character from the transactional data insurers are used to using and storing.
2. When considering analytical data, the business needs look to the ultimate usage of the data, usually reports and dashboards—and level of granularity and drill through required.
3. IT can utilize OLAP techniques to provide sophisticated multi-dimensional views of data but only if the required outputs are well defined.
4. Ensuring the quality of analytical data is absolutely critical—without this, the accuracy of all generated finance and actuarial numbers can be called into question, and insurers may fail to meet the EIOPA data quality requirements.
5. Data accuracy requirements are not the same for all purposes. It is important to articulate why accuracy is needed, otherwise valuable effort can be wasted on improving data quality that has little materiality.
6. There are tools which can help improve the quality of data which are combination of techniques and expert judgment. These should be utilized wherever possible.
7. Business rules are an important part of the data quality process and while there are many pre-built generic rules it will be critical to supplement these with user defined rules that reflect unique business considerations.
8. Improving data quality is an ongoing process – not just when the data is initially loaded. Data is constantly changing. An analytical repository is an essential element in improving data quality.
9. Spreadsheets remain an important element of analytical data and must be carefully managed and controlled.
10. Most data related projects don't fail because of the technology; they fail because practitioners cannot define precisely what data they want, what purpose they are going to use it for and the quality criteria necessary.

SP24283