White paper

GENERATING MODEL DATA GOVERNANCE IMPACT

MODEL DATA GOVERNANCE

Banking organizations across the globe have been focused on building strong risk models with little emphasis and focus on having effective data determination and governance mechanism of data inputs to run those models.

In addition to the objective of aligning with enterprise level data management and governance policies, regulatory push have increased focus on bank's risk data modeling and governance capabilities through SRC 11-7 guidelines on model risk management. This is particularly important as now banks possess huge volumes of data at their disposal, sourced from public domain, social media or from third party vendors. While large pool of data as model input would be critical and invaluable for better insight generation, forecasting and capital planning; their effective management would be an uphill task.

As G-SIBs would plan to re-look and revamp their existing data governance infrastructure extending it to model data areas, D-SIBs emerge as new players in contention to implement comprehensive model data governance and quality programs. Establishment of strong risk data governance practice is the first step toward building effective model risk management infrastructure.

Superior data governance demands to break-away from data and process silos with improved collaboration between cross-functional teams and establishment of standardized policies, processes and workflows.

This paper highlights major challenges a banking organization faces in practicing model data governance. It also proposes an approach on building focused model data governance mechanism by providing strategic solutions on data management functions encompassing policy formulation, meta data management, data validation, issue remediation, work flow management and documentation.

> “Quality of model outputs depends on the quality of input data and assumptions, and errors in inputs or incorrect assumptions will lead to inaccurate outputs”

Source: Office of the Comptroller of the Currency (OCC), SRC 11-7 regulation on model risk management

SRC 11-7 Guidelines are descriptive and not prescriptive in nature

Office of the Comptroller of the Currency (OCC), Federal Reserve Board and Federal Deposit Insurance Corporation, through SRC 11-7 guidelines placed significant emphasis on quality of data used to develop a risk model. Guidelines state that, rigorous assessment of quality of data and its relevance to the model should be subjected to validation. It requires model developers to be able to demonstrate that suitability of data for the model and that they are consistent with the theory behind the approach and with the chosen methodology. If data proxies are used, they should be carefully identified, justified and documented.

If data is not representative of the bank's portfolio or other characteristics, or if assumptions are made to adjust the data and information, these factors should be properly tracked and analyzed so that users are aware of potential limitations. Thorough review of the quality, credibility and applicability of such model data and other components supplied by third party vendors are to be validated by banks and supervisory authorities, using a mechanism to exchange views on the merits and limitations, challenges.

Although regulators insist banks to have strong model data governance mechanism in place, the guidelines are mostly idiosyncratic and would leave it to banks themselves to design and operate its own model on data governance

Demystify the SRC 11-7, to call out key drivers

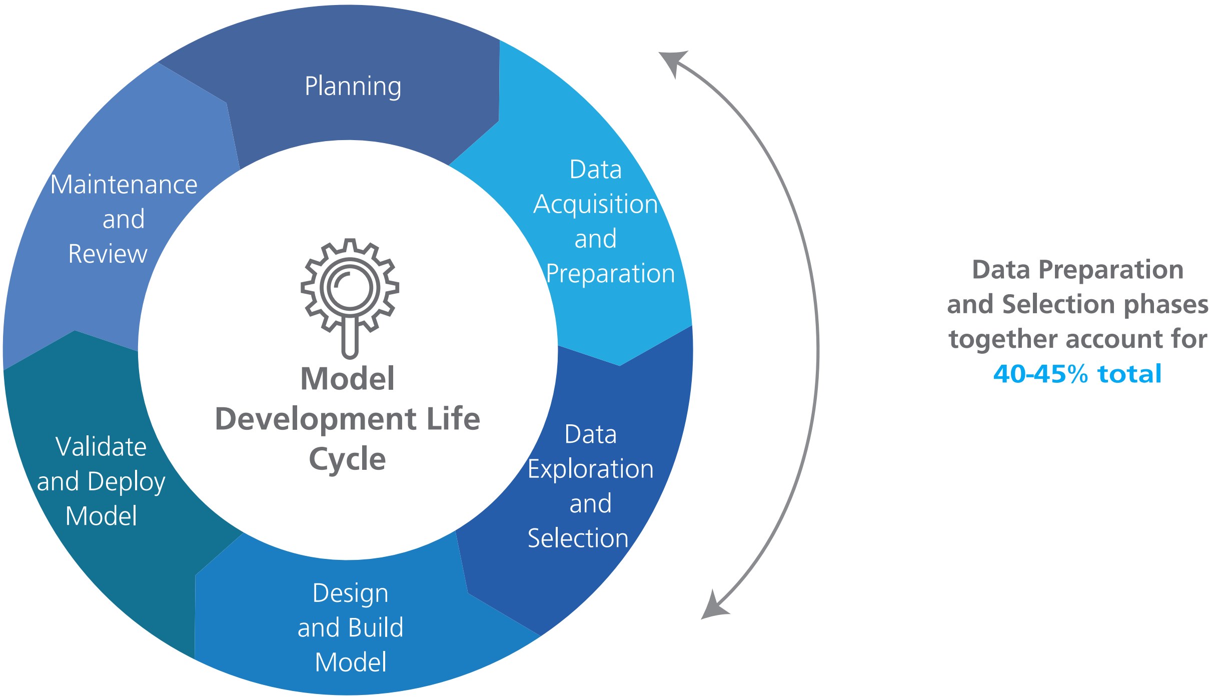
Although the SRC 11-7 directive does not explicitly refer to all drivers listed below, it greatly emphasizes the need for superior model data governance and management.

- Regulatory compliance
- Business relevance in determining Critical Data Elements(CDEs)
- Elimination of data and process silos
- Model data standardization
- Enhanced work flow management
- Increased insights and improved decision making
- Cost and resource optimization

Why is model data governance so important?

Banks and researchers put great importance on quality of input data and consider the data limitations as a key impediment to design and build strong risk models. Data preparation activities include acquisition, cleansing, exploration and selection, which accounts for nearly 40-45% of total time and effort in model development life cycle.

Exhibit 1: Model Development Life Cycle


Planning
Data Acquisition and Preparation
Data Exploration and Selection
Design and Build Model
Validate and Deploy Model
Maintenance and Review
Model Development Life Cycle
Data Preparation and Selection phases together account for 40-45% total

Current State Challenges

01

Data availability and consolidation

Availability of data mostly dictates the degree of complexity and the choice of methodology in modeling risk parameters. Credit models can use both internal and external data for risk estimation. While internal data would be complete representative of the portfolio, external data requires banks to demonstrate data are representative of underlying population of the bank. According to Basel regulatory norms, the observation data set needs to be at least for a period of five years. In case of short fall banks are allowed to use external data, but with conservatory margins. And, in some cases data sub-sets are identified using random draw and preferred when available data sets on a complete portfolio is too large to estimate a model.

Case 1: A commercial bank is sourcing external data on counterparty ratings from Moody's and Bloomberg for PD risk parameter estimation. However, coverage of such pooled external PD data against bank's total portfolio and consequent reliability is a key consideration, as it will affect the accuracy in PD estimation.

02

Lack of business relevance to CDE

Availability of data mostly dictates the degree of complexity and the choice of methodology in modeling risk parameters. Credit models can use both internal and external data for risk estimation. While internal data would be complete representative of the portfolio, external data requires banks to demonstrate data are representative of underlying population of the bank. According to Basel regulatory norms, the observation data set needs to be at least for a period of five years. In case of short fall banks are allowed to use external data, but with conservatory margins. And, in some cases data sub-sets are identified using random draw and preferred when available data sets on a complete portfolio is too large to estimate a model.

Case 2: LGD model developer requires data on charge-offs for his model. In the meantime, data experts have computed values for net charge offs (charge offs minus recovery) and stored in a field called NET_CHRG_OFF. With non-existence of communication channel with data experts and with his functional judgment, developer sources gross charge offs from a different field are called CHARGE_OFFS in risk data warehouse.

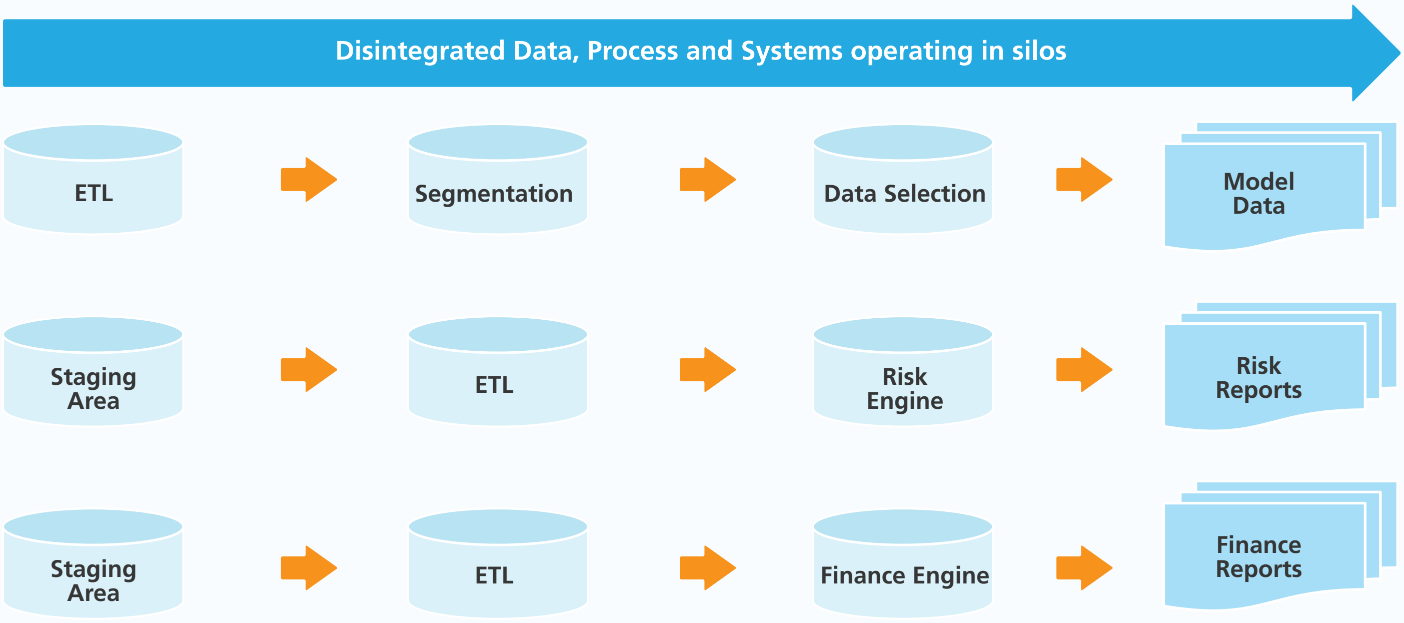
03

Data and process silos

Although, data governance and quality is established as business function within the banking organizations, these programs are mostly restricted and aimed at risk data ware house and data marts. The data being used for risk model development and validation purpose still manages to flow as input to risk models without effective data quality controls and not aligned to overarching data governance of the banking enterprise. Model development and validation team leverages stand-alone exception detection mechanism to identify and remove data exclusions. There exists a lack of co-ordination and collaboration between multiple teams involved, resulting in parallel efforts, resulting in data and process silos. This silo approach further extends to risk and finance areas, which requires additional efforts on reconciliation and adjustments. Data generators are far removed from the process of shortlisting critical data elements for models. Impact of bad data and its influence on model output are never contemplated and deliberated between data generators and consumers. Data creators such as branch bankers, loan officers and traders do not understand the impact of the data submitted by them, as there is little incentive for them to adopt a data governance model. Also, model developers seldom clarify all the issues and weakness of on boarded data being used with business.

Case 3: Risk management team in a bank identifies discrepancies between internal rating and external rating of counterparties. They observe that, it's a case of recent rating migration and which is yet get updated in bank's reference data systems. However, the banks do not log their observations. Without such logs or record, a risk modeler would perform rating review exercise again. This would result in duplication of effort.

Exhibit 2: Data and process silos



04 Absence of model data dictionary

Model developers use data sourced from multiple systems and layers. The practice of establishing formal model data dictionary and its maintenance is almost non-existent. Further to this, data elements aggregated or consolidated from many sources never validated for aspects such as data taxonomies, functional glossaries and functional/technical definitions. Rules configuration mostly limited to statistical and basic ETL logic. Data quality and business rules which could bring functional dimensions are seldom applied.

Case 4 : A risk modeller is required to consolidate data on subsequent lien at business entity level by aggregating values for Second Lien, Junior Lien and Subordinate Lien. However, in the absence of a dedicated model data dictionary and his limited domain knowledge, finding all aliases of subsequent lien seems to be a manual and time consuming effort.

05 Data pollution

Risk data before it reaches as model input, gets polluted at multiple stages and with high possibility of missing critical data elements. Some of the important reasons which can contaminate data are (I) unrestrained manual edits, (II) adjustments done external to source, (III) incorrect ETL logic and data loss.

Absence of work flow management which encompasses multiple levels in the model data governance model is the root cause of many problems listed on data pollution. Communication, co-ordination and collaboration between multiple teams and stake holders involved in data management rely heavily on dysfunctional manual activities and processes, which could lead to uncontrolled manual edits. Manual intervention and edits on large volume of inconsistent data investigation will prove to be time consuming and costly exercise.

Case 5 : An LGD risk modeler receives file with incomplete data on recovery for the last one year. He performs missing value treatment on data received to replace all the missing values with “0” (Zero) to make the data set complete for estimation. This would mean bank has no recovery from borrowers of the loan on the outstanding amount, during certain period. In true business context, this may not be true.

06

Issue management, remediation and documentation

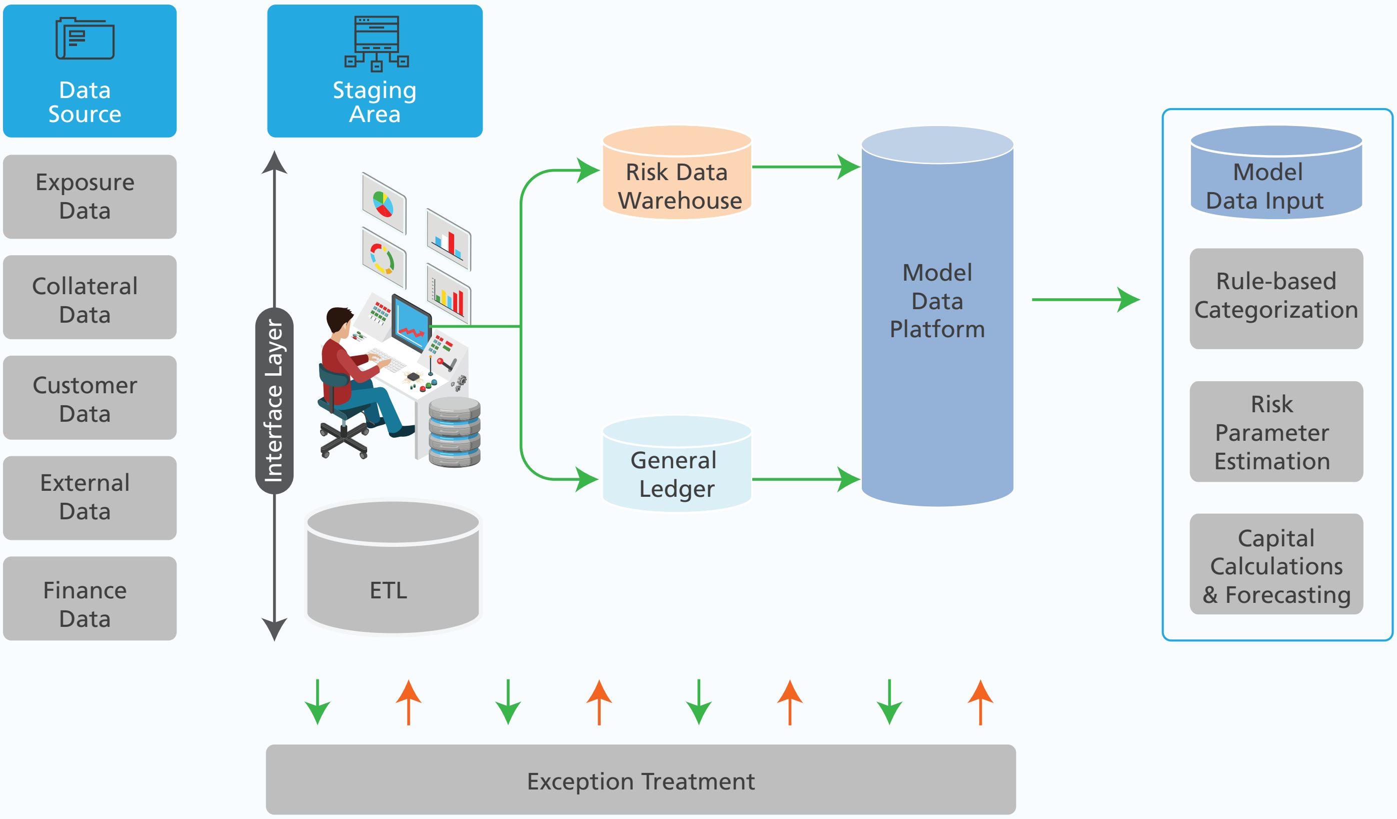
Accuracy of aggregate data and its comparability to a bank's actual portfolio is very critical. If these two standards are not met, the use of aggregate data can potentially disguise individual loan-specific effects to which a bank is exposed. This is important to consider as banks' overtly relies on aggregate data to estimate individual borrower parameters and in turn default events.

The entire job of outlier identification, investigation and rectification would be performed by model developers manually. This mode of operation without complete involvement of data domain experts and formal workflow would open up the problems on issue prioritization, classification, parties involved, tracking, remediation strategy and timely closure. Data that has not been carefully screened for invalid, out-of-range and missing values can produce highly misleading results. Also, exception remediation most often not prioritized based on criticality of a model input variable.

Regulators have also highlighted that, documentation of details on model development and validation is either completely absent or incomplete. Incompleteness and lack of clarity in documentation also extends to basic aspects of data such as, its source, nature and type.

Case 6 : Credit risk modeling team in a bank has the practice of prioritizing data anomalies to be fixed based on anomaly identification date, instead of computed criticality scores. Where, criticality scores designed to take amount of exposure attached to an anomaly as one of key consideration.

Exhibit 3: Current State



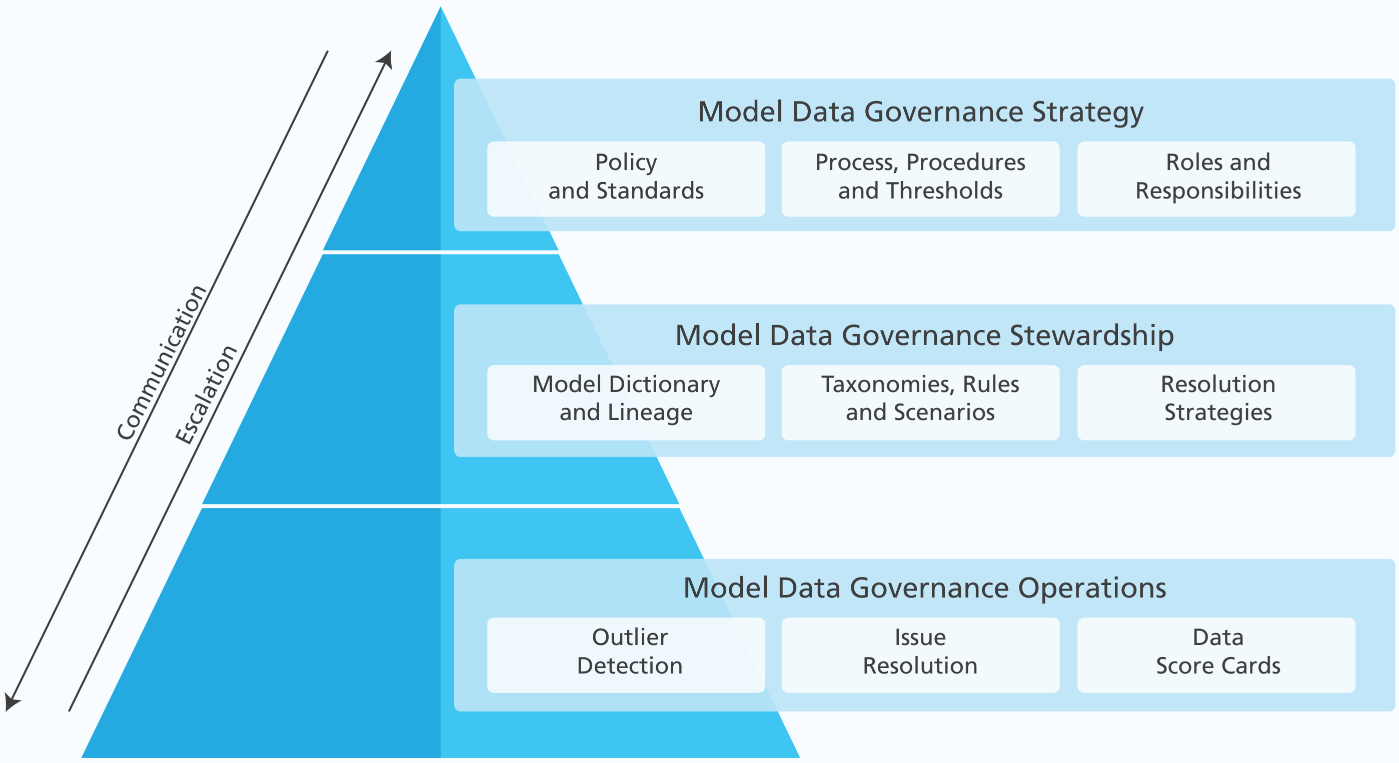
Proposed Approach

While banking organizations face uphill task to overcome model data governance implementation challenges, a focused and methodical approach can help to extract synergy from several existing capabilities and consolidating them

01 Establishment of model data governance operating model

Establishment of formal model data governance and compliance operating model would be the first step in addressing data challenges in risk modeling life cycle. Data governance council or steering committee would operate at the highest level and responsible to define governance policies, procedures, formulate strategy and establish standards, thresholds and workflows to assure how changes are authorized and audited. Data stewardship layer would include both business and technical data stewards responsible for defining and maintaining model data dictionary, perform data validation for outlier detection and formulate defect resolution strategies. At the bottom of the pyramid, operations team would involve in performing technical data lineage, defect resolution, critical data element score cards and report generation activities. While, escalation flow would be always bottom-up, communication would flow two ways.

Exhibit 4: Model Data Governance Operating Model



Model Data Governance Strategy

Policy and Standards | Process, Procedures and Thresholds | Roles and Responsibilities

Model Data Governance Stewardship

Model Dictionary and Lineage | Taxonomies, Rules and Scenarios | Resolution Strategies

Model Data Governance Operations

Outlier Detection | Issue Resolution | Data Score Cards

*Rules Includes Business, Data Quality and Statistical Rules

02

Theme-specific data dictionary

Establishment and maintenance of model data dictionary containing critical data elements with industry standard data definition, taxonomies, semantics and rules is one of the core functions of data stewardship group. The data dictionary would provide perspective from business and technology dimensions on CDEs and would serve as single source of truth for all model data requirements. Model data dictionary versions and hierarchies are also maintained to ensure that data remains in sync at all times.

03

Collaboration between cross-functional teams

In the new team structure, a risk consultant and data experts would work along with risk model developer in a cross-functional team for model input identification. This would not only bring transparency in the selection process, but will also bring multiple dimensions of business relevance and functional judgment to CDE identification.

04

Breaking away from data and process silos

Data integration at enterprise level helps overcome the challenges posed by data silos. Unified data platform helps to streamline the complexity of matching, cleaning and preparing all data for model development and validation tools. Tools and processes maintain a high quality, consistent set of master data to provide a common point of reference and act as golden source. Through unified data integration and data management, banks can successfully support the modeling and analytics systems that drive the business. In the proposed eco system risk, finance and model data infrastructure co-exist to interact and reconcile, seamlessly.

05

Reduced manual edits and increased automation

Manual data edits is the most resource intensive, time consuming and error prone process in model data cycle. It affects during outlier investigation, classification and corrections stages. Manual edits can be controlled in multiple ways. Selective editing is one such technique, which prioritizes those exceptions or outliers, which could have substantial impact on principal risk parameter outputs. This requires a mechanism to zero in on highly influential outliers, which can be short listed for selective manual editing. It is well established during research and surveys, that mostly minor or negligible percentage of outliers could play part as influential exceptions. Hence, the manual edits can be limited to only such outlier population. Another most effective method is to perform editing automatically. It includes series of steps such as (I) configuration of edit rules for automatic outlier detection, which includes both statistical and domain specific rules, (II) localization of errors, (III) Treatment of exceptions with corrections and imputations, (IV) amend values for consistency, (V) compute field scores and selection.

06

Well-defined work flow mechanism

Model data work flow mechanism helps in seamless communication and co-ordination between various stakeholders and teams to optimally engage them in well-defined process-oriented model data governance model.

(I) Governing Council: Governing council can leverage policy management module to formulate and manage policies and procedures on data security, sharing, SLAs, thresholds and approval mechanism. They supervise and review results to approve and communicate changes in policies and boundaries as necessary.

(II) Data Stewardship: Data stewardship group would work along with modelers and domain experts in identifying CDE for data dictionary. Data stewards would help user to review, edit, update and approve definitions and mappings. Data stewards map CDEs to information segments and domain areas with in the dictionary to help user to understand the business lineage. They would also identify gaps in data hops when data travels between systems, applications and layers. Data mapping to relevant systems, applications and layers serve to establish traceability to achieve technical lineage. Stewardship group would formulate exception resolution strategies and communicates to operations team. They play vital link between council and operations team for all communications.

(III) Data Operations: Data operations team would identify gap, outlier or breach with application of data quality, business, statistical rules and scenarios configured by data stewards working along with domain experts and statisticians. In addition to statistical techniques, data quality rules and business rules will bring multiple dimensions of locating an exception. Data stewards would then formulate remediation strategy, prioritize, assign to operations team and track the defect. Operations team would then analyze, remediate, fix, communicate and close the defect.

(IV) Data Scoring: The users can select, rate and rank data variables based on its criticality, domain relevance and functional judgment. The higher rated variables would be filtered, reviewed and approved by cross-functional team members and subject matter experts to finally determine as critical data elements to run models.

07

Detailed data and process documentation with version control

Model development and validation exercise documentation should cover vital aspects of data functions, processes and assets. It should provide detailed information on (I) sources and type of data, (II) data dictionary used, (III) data lineage performed, (IV) ETL logic and rules applied (V) test scenarios, (V) sampling techniques, (VI) data limitations, weakness and outliers, (VII) remediation strategy, (VIII) selection filters and score cards, (IX) assumptions, (X) data management and modeling resources, (XII) documentation version and hierarchies

Transformation from tactical state to ideal state

Proposed model data governance operating model in response to varies challenges outlined in this paper will ensure a banking entity to have smooth transition from 'As Is' to 'To Be', through tactical phases. It provides focused solution approach to overcome model data specific challenges encompassing the below:

- Industry standard data dictionary to act as single source of truth for all model data requirements
- Comprehensive outlier detection and resolution mechanism with business, data and statistical Rules
- Reconciliation between risk and finance data with unified data infrastructure
- Critical data element identification with domain filters and statistical score cards
- Seamless communication, co-ordination and collaboration between multiple teams

Exhibit 5: Tactical State

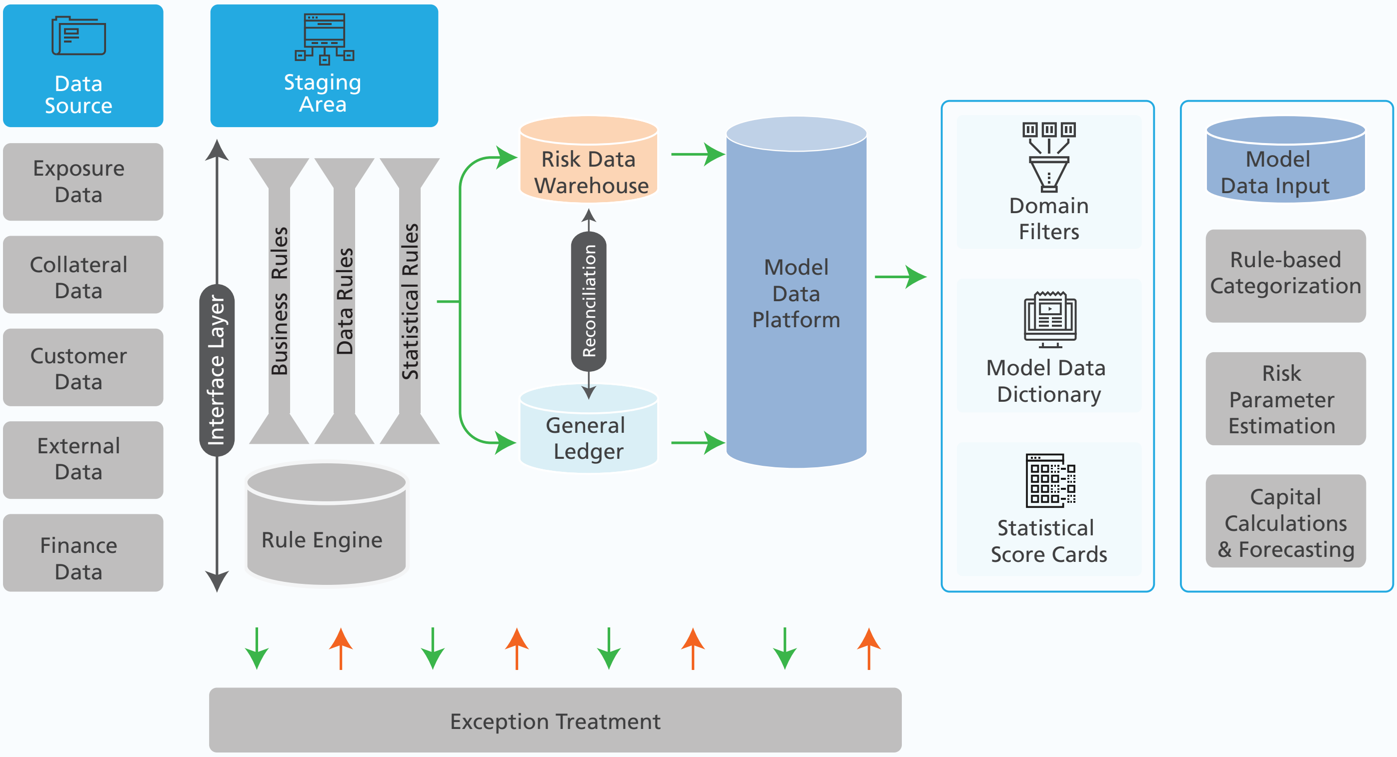
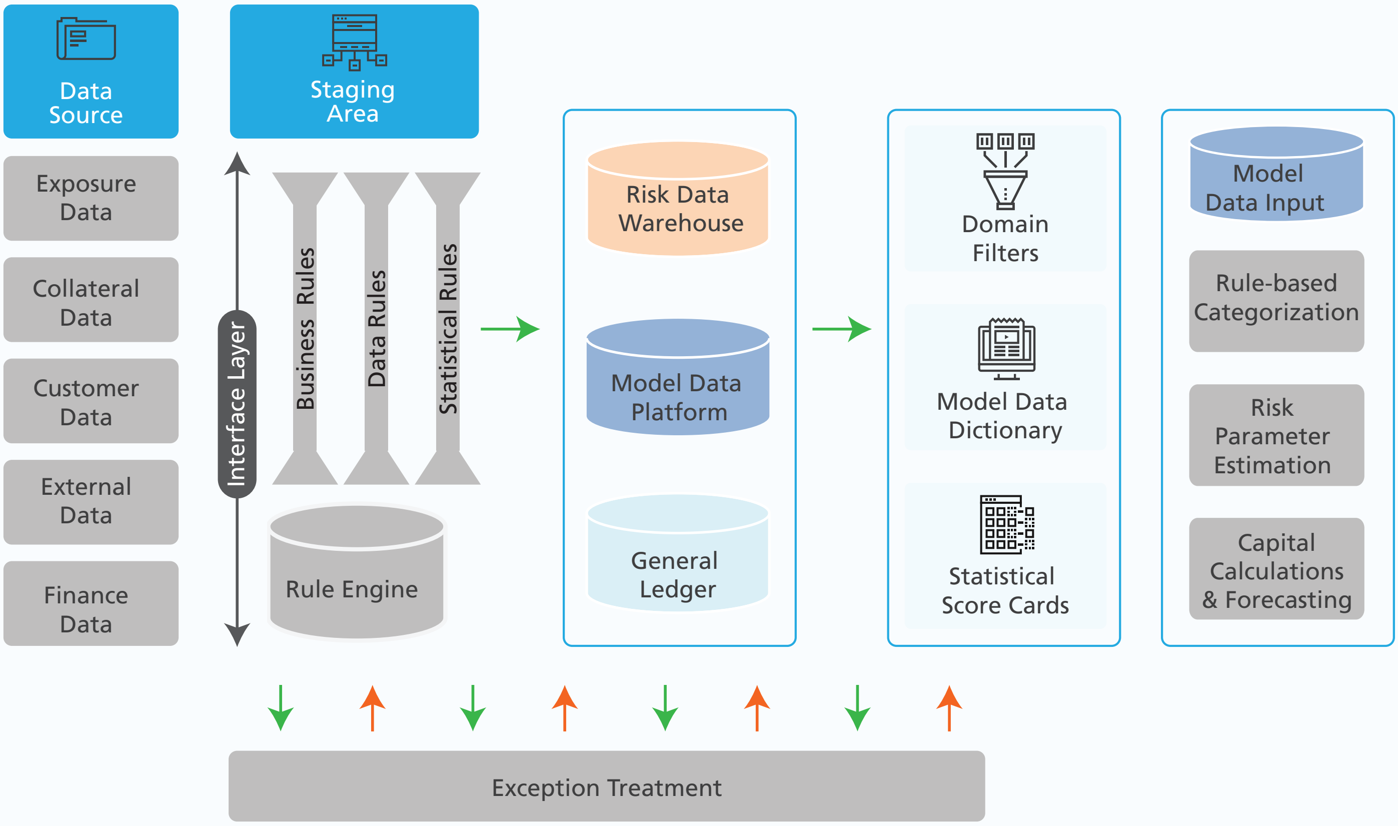


Exhibit 5: Ideal State



Conclusion

Risk model data is considered as vital component in accurate risk parameter estimation, capital calculations and forecasting. Further to mitigation of default risk, it has ability to predict potential systemic risks and economy fallout. Given the importance of risk model data, regulators have issued mandate through SRC 11-7 and in response financial institutions are in the process of formulating model data governance programs. Proposed model risk data governance operating model will help banks to roll out comprehensive model data governance program, which will help them to identify, integrate and transform all essential data management functions to effectively manage model risk data.

Authors

Narasimha Swamy is a Financial Risk Consultant with Data Management Practice, at Genpact Analytics. He has 10+ years of experience in the areas of financial services, IT and business consulting. He is currently the Risk Data Governance Lead Consultant for a leading North American banking client. He has been involved in multiple risk consulting engagements and implementation projects on both regulatory and strategic fronts in Basel II/III and CCAR areas for financial institutions in the US. He is also responsible for developing solution offerings and frameworks on Basel III, CCAR/DFAST, BCBS 239 and Liquidity Risk areas, for Genpact clients in the US banking sector.

Prakash Jagannathan is an operating leader and offshore data practice lead in Genpact, where he manages multiple accounts from a delivery and solutions perspective. He has 14+ years of experience in the areas of Risk Data Management & Reporting. He has played delivery, consulting and solution development roles in organizations including Citi, Credit Suisse and S&P India (CRISIL) with focus in the areas of risk data governance and management and risk solution development.

Please email us at narasimhaswamy.h.r@genpact.com or prakash.jagannathan@genpact.com for additional information