pennsylvania
OFFICE OF ADMINISTRATION

# Information Technology Policy

## *Data Modeling Best Practices*

**Number**
BPD-INF003C

**Effective Date**
August 2, 2005

**Category**
Information

**Supersedes**
All Prior Versions

**Contact**
RA-ITCentral@pa.gov

**Scheduled Review**
September 2023

## 1. Data Modeling Best Practice Standards

| Best Practice | Rationale |
|---|---|
| Model access should be controlled with security features built into the software. | Limits access to authorized users to protect content of sensitive information. |
| Merge models and reconcile overlapping duplicate objects. Identify differences between model versions. | Modeling tools can compare and merge information between models, and produce the DBMS-specific alteration SQL to modify structures while allowing the model to be simultaneously updated to reflect changes on the server. |
| Projects can save time and money by using a previously used and approved logical and physical model as a starting point. | This leads to physical data sharing and less storage of redundant data. It also helps the organization recognize that information is an organization-wide resource and that data models are important information assets. |
| Existence of a single, accurate, and reliable source for a particular data model. | Multiple disparate sources create confusion during the development cycle. |
| A comprehensive change management process must be put in place to manage modifications to the model and ensure their accuracy.<br><br>The ability to bi-directionally compare and merge model and database structures will reduce additional work on the part of the Data administrators to manage change. | Time and money will be wasted if project teams cannot rely on the integrity of the data model. |
| A good model management process is essential. This is the method by which data and process models are | Without a model management infrastructure:<br>• Model content becomes inconsistent<br>• Redundant data structures are created |

| **Best Practice** | **Rationale** |
|---|---|
| developed, maintained, used, and reused within a model development life cycle. A rule of thumb is if you have 5 or more data modelers, you need a model management infrastructure.<br><br>A model-management infrastructure is critical for managing the impact of change and resolving model conflicts during and after the development process. Standards and procedures around model management must be put in place. | • Miscommunication between modelers, analysts, and users occurs<br>• There is an inability to share knowledge between projects and agencies.<br><br>Multi-user modeling environments built to handle real-world situations makes coordinated, large-scale modeling possible. |
| Data models should be created with the intention to share data structures across projects and agencies. | Opportunities to reduce data redundancy and share data will improve data quality and reduce costs. Information silos within and among agencies will be eliminated. |
| Models must be layered, with Master Data (Dimensional), Conditional Data (Dimensional), Transaction Data and KPI (Facts) Data, allowing for effective reuse and integration of models. | Reduce the overall number of models (model redundancy) across the enterprise, saving cost of creation, and maintenance (Create once and use many approach). |
| Models must have traceability (data lineage and audit) to track the source of the data. | Helps understand the data lifecycle, confirms business rules exist as expected, and provides auditability |
| Overall Governance Process is necessary, for all three areas: Model, Data and Security. | Ensure governance compliance and segregation of duties in Model definition and implementation processes, Data access and Security implementation for Model and Data.<br>Functionally separate models (ex: Engineering Models vs. Finance Models) should be housed in separate areas, with users from each functional domain accessing only what is allowed for them.<br>Access to Integrated models (ex: Plant Performance, which may include Employee, Plant and Finance data sets) should be allowed only to select group of people who has authorization to see ALL those data sets. |
| Data Privacy: Models must have built in security capabilities for both Row and Column Level.<br><br>Row Level Security – ex: Agency1 vs. Agency2<br><br>Column Level Security – Across agencies or within an Agency, mask column level data (ex: PII) | Data Privacy and Security should be designed in the model build, so multiple agencies can access the same model, but retrieve only data that is securely allowed for their preview.<br>Similarly, identify and design column level security (ex: Personally Identifiable Information – PII), so even within the same agency, column level information can be masked (Ex: XXX-XXX-XXXX for SSN) at agency level or user level. |
| Model level Performance consideration based on data consumption. | Model should have enough parameters that allows for efficient consumption of data, without |

| **Best Practice** | **Rationale** |
|---|---|
| | degrading the performance of the model. Factors like variables, calculation logic, and consumption mechanism should all be considered for optimal model consumption. |
| Models should be componentized, that additional transformation needs for models are minimal, | Model build should be componentized and layered, in order to be treated as independent building blocks, that can be enhanced or continued to be built upon as another layer, effectively with minimal change. This will reduce cost of ground-up build for new requirements, especially in scenarios where new transformations are requested. |
| Standard Naming Convention and standardization of Models for Enterprise | For effective Model Governance, Maintenance, and Enhancements, Standard Naming Convention must be followed. This reduces the Total Cost of Ownership and help retain/transfer knowledge among Model Governing teams. |
| Raw data vs. Calculated fields (with logic applied) ex: Age | Models should expose raw data from the system as much as possible and refrain from applying calculation logic to any field. The exception to this rule would be calculation logic applied to Enterprise-Wide fields (ex: Age Calculation or Profit Calculation), which has commonly accepted and agreed upon calculation logic, across the Enterprise, unambiguously.<br><br>Additional calculation logic can be applied by the consumers of the model data, as per their needs. This will allow the enterprise models to grow and mature in identifying and incorporating KPI and Calculation that are agreed upon by the entire organization instead of building multiple versions of the truth at the model level, which leads to confusion. |
| Avoid Data Staging (materialization) and always allow for real-time direct access to source. | Data Staging is the practice of storing the data, as of certain date and time for others to access. Staging leads to multiple data silos, which leads to multiple versions of the truth, which will demand significant time, effort, and resources to prove which version of truth is accurate.<br><br>The use of Data APIs to pull data in real time is strongly encouraged as a primary means of data access.<br><br>An efficient enterprise will avoid Data Staging at all costs, allowing for the raw data from source to flow real-time, and prevent stale data scenarios. This is a significant paradigm shift, from batch-oriented model design to real-time based model design.<br><br>In some cases, staging is unavoidable (depending on the consumer of the model, |

| Best Practice | Rationale |
|---|---|
| | technology etc.) and in those scenarios, careful vetting is needed for justifying the staging process. Data Volume, refresh frequency, and impact to the models should be considered carefully and consumers should be educated about the staleness of data before Staging is allowed in an organization. Model Governance plays an important role in approving these staging requirements. |
| For any new model to be created, there should be a clear use case and cost-benefit justification. Also, it should be proven that the model does not already exist in other forms, prior to requesting or enhancing a model. | Through the Model Governance process, any new model requests should have clear use case. It should also have justification on cost-benefit for new model build.<br><br>Multiple requests from variety of agencies or partners may come in as a new model request. These should be carefully considered, both functionally and technically, before creating new models or enhancing existing ones. |
| Models should be reviewed though the Model Governance process and approved prior to being implemented for use. | Ensure standardized modeling entities usage, compliance to enterprise architecture, and security practices. |

This chart contains a history of this publication's revisions.

| Version | Date | Purpose of Revision |
|---|---|---|
| Original | 08/2/2005 | Base Document |
| Revision | 11/18/2010 | ITP Refresh |
| Revision | 05/13/2021 | ITP Refresh |
| Revision | 09/09/2021 | Clarified that the use of Data APIs to pull data in real time is strongly encouraged as a primary means of data access. |