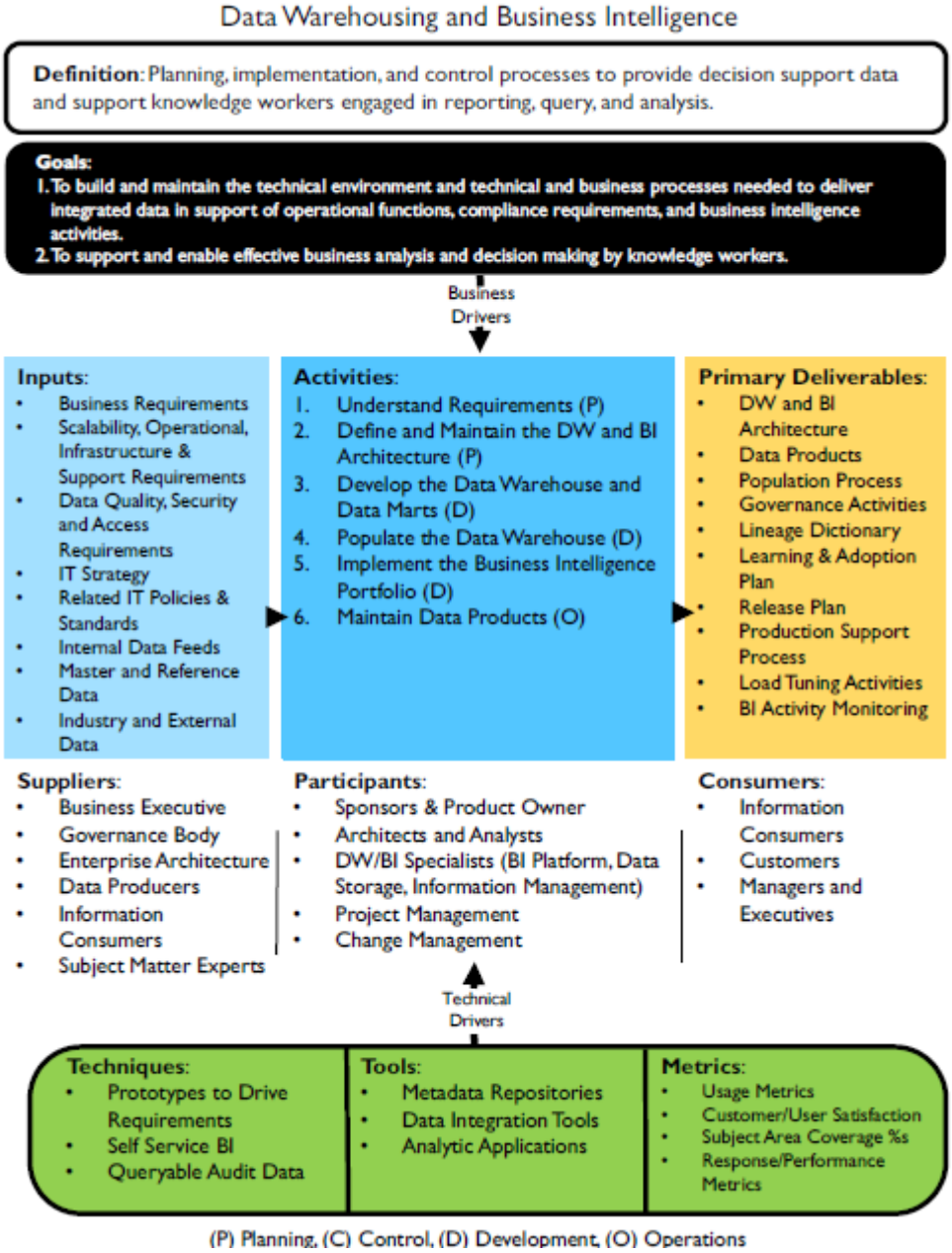
Data Warehousing and Business Intelligence

The data warehouse is meant to enable decision support systems which could share core enterprise data from a common data model. The enterprise warehouse reduces data redundancy, improves the consistency on information, and enables the enterprise to make better decisions.

1 Introduction



1.1 Business Drivers

- Business Intelligence support
- Compliance requirements
- Support operational activities

- Enables effective business analysis and decision-making
- Find ways to innovate based on insights from data

1.2 Goals and Principles

Goals:
1. To build and maintain the technical environment and technical and business processes needed to deliver integrated data in support of operational functions, compliance requirements, and business intelligence activities.
2. To support and enable effective business analysis and decision making by knowledge workers.

Organisations implement data warehouses to:

- Support BI activity
- Enable effective business analysis and decision making
- Find ways to innovate based on insights from data

Principles to implement a Data Warehouse:

- **Focus on business goals:** DW solves business problems
- **Start with the end in mind:** Business priority and scope end-data-delivery in BI space drives creation of DW content
- **Think and design globally; act and build locally:** Architecture guided by end-vision, but build and deliver in sprints to enable return on investment.
- **Summarise and optimise last, not first:** Build on atomic data
- **Promote transparency and self-service:** Provide more context (Metadata) and keep users informed of updates and changes
- **Build Metadata with the warehouse:** To be able to explain the data, capture as part of the development cycle and manage as an ongoing activity
- **Collaborate:** With other data initiatives especially DG, DQ and Metadata
- **One size does not fit all:** Different groups of data consumers need different tools

1.3 Essential Concepts

1.3.1 Business Intelligence

Two meanings:

- **Type of data analysis** aimed at understanding organisational activities and opportunities
- **A set of technologies** that enable decision support analysis (querying, data mining, statistical analysis, reporting, scenario modelling, data visualisation and dashboarding).

1.3.2 Data Warehouse

A Data Warehouse (DW) is a combination of two components:

- Integrated decision support database
- Software programs used to collect, cleanse, transform and store data from a variety of internal and external sources

An **Enterprise Data Warehouse** (EDW) is as centralised data warehouse designed to service the BI needs of the entire organisation. Adheres to the enterprise data model.

1.3.3 Data Warehousing

Data warehousing describes the operational extract, cleansing, transformation, control and load processes that maintain the data in a data warehouse. Integrated, historical, enforces business rules and maintains business relationships. Processes interact with Metadata repositories.

- **Structured data:** elements in defined fields as documented in data models
- **Semi-structured data:** Electronic elements organised as semantic entities with no required attribute affinity
- **Unstructured data:** Not predefined through a data model

1.3.4 Approaches to Data Warehousing

Two thought leaders – Bill Inmon and Ralph Kimball:

- **Inmon**:
  - A DW is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision-making processes.
  - Normalised, relational model
- **Kimball**:
  - A copy of transaction data specifically structured for query and analysis.
  - Dimensional model

Core ideas:

- Warehouses store data from other systems
- Storage includes organising data in ways that increases its value
- Warehouses make data accessible and useable for analysis
- Organisations build warehouses to make reliable, integrated data available to authorised stakeholders
- Warehouse data serves many purposes

1.3.5 Corporate Information Factory (Inmon)

CIF Illustrates the differences between warehouses and operational systems

- **Subject oriented:** Based on major business entities, not function or application
- **Integrated:** The warehouse becomes a system of record for the data.
- **Time variant:** Stores data as it exists at a set point in time
- **Non-volatile:** New records are appended not updated
- **Aggregate and detail data:** Details of atomic level transactions as well as summarised data
- **Historical:** The focus of operational systems is current data. Warehouses contain lots of historical data as well.

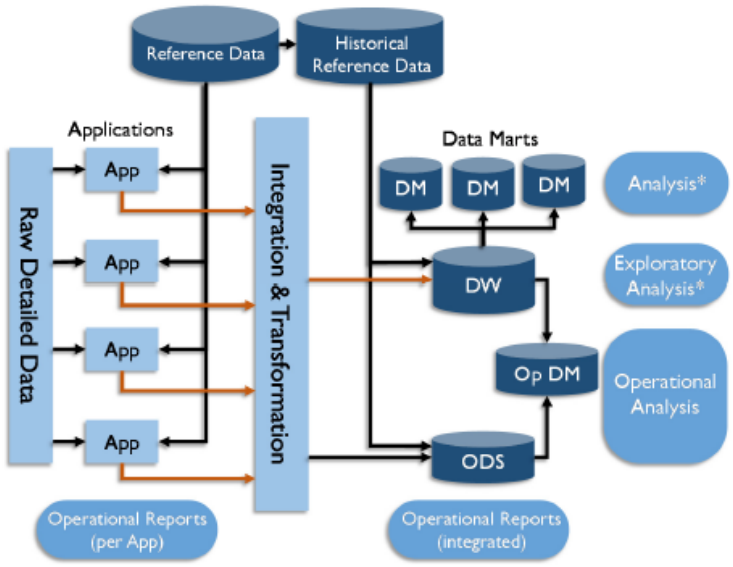


Figure 80 The Corporate Information Factory

CIF Components:

- **Applications:** Applications perform operational processes and bring data into the DW and operational data store (ODS) where it can be analysed
- **Staging area:** A database between the operational and target where the ETL takes place. Data is transient.
- **Integration and transformation:** Data from disparate sources is transformed in the integration layer into the standard corporate representation model in the DW and ODS
- **Operational data store (ODS):** an integrated relational database of operational data, containing current or near-term data, volatile. Meet the need for low latency data. Can be the primary source for the data warehouse.
- **Data Marts:** Provide data prepared for analysis. Sub-set of the warehouse designed for the needs of specific consumers. Data marts are designed using dimensional modelling and denormalisation.
- **Operational Data Mart (OpDM):** Sourced from ODS, volatile and focused on tactical decision support
- **Data Warehouse:** A single integration point for corporate data to support management decision making and strategic analysis and planning.
- **Operational reports:** The output from the data stores.
- **Reference, master and external data:** Data required to understand data from applications, and simplifies integration to DW.

Comparison between data stored in DW and Marts and Data on operational systems:

Data in DW and Marts	Data in Operational Systems
• Organised by subject	• Organised by function
• Integrated	• Siloed
• Data is time-variant	• Current value only
• High latency	• Lower latency
• Significantly more historical data	• Less historical data

1.3.6 Dimensional DW (Kimball)

Kimball: “a copy of transaction data specifically structured for query and analysis”.

Warehouse data is stored in a dimensional model, not normalised, enabling consumers to use the data and optimise query performance.

Star Schema or **dimensional** models:

- **Facts** contain quantitative data about business processes. (The fact table is also called a meter which contains measures – Hoberman)
- **Dimensions** store descriptive attributes related to fact data and allow consumers to answer questions about the facts

One Fact table is joined to many dimensions – looks like a star.

Conformed dimensions are common dimensions and are shared by multiple fact tables via a bus. Multiple data marts can also be integrated at enterprise level by plugging into the bus of conformed dimensions.

Table 27 DW-Bus Matrix Example

	Subject Areas				
Business Processes	**Date**	**Product**	**Store**	**Vendor**	**Warehouse**
Sales	X	X	X		
Inventory	X	X	X	X	X
Orders	X	X		X	
Conformed Dimension Candidate	*Yes*	*Yes*	*Yes*	*Yes*	*No*

Kimbal’s Data Warehouse Chess Pieces is more expansive than Inmon’s:

- **Operational source systems**: Equivalent to the applications in CIF
- **Data Staging area:** Processes needed to integrate and transform data for presentation. Kimball’s EDW
- **Data presentation area:** Like data marts in the CIF. The DW Bus conformed dimensions
- **Data access tools:** End users’ needs drive adoption of data access tools

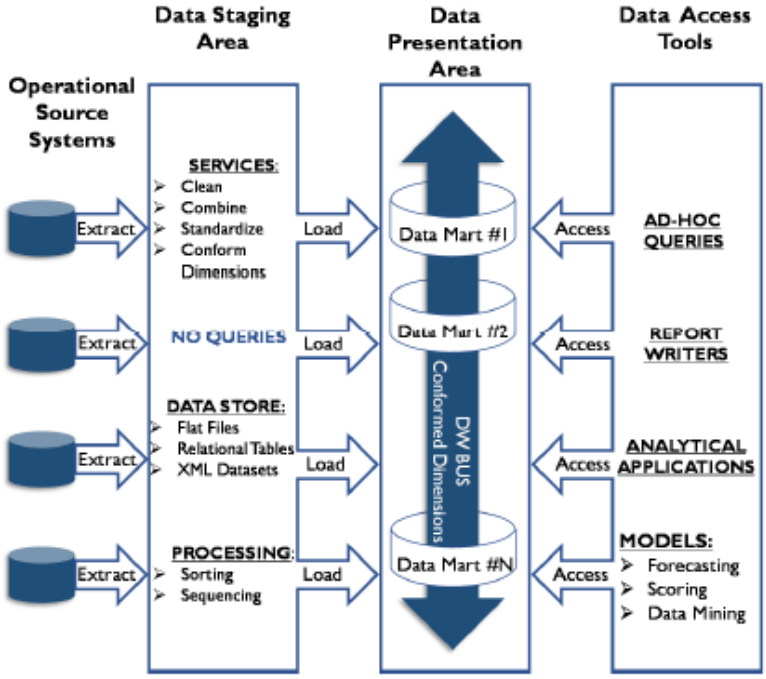
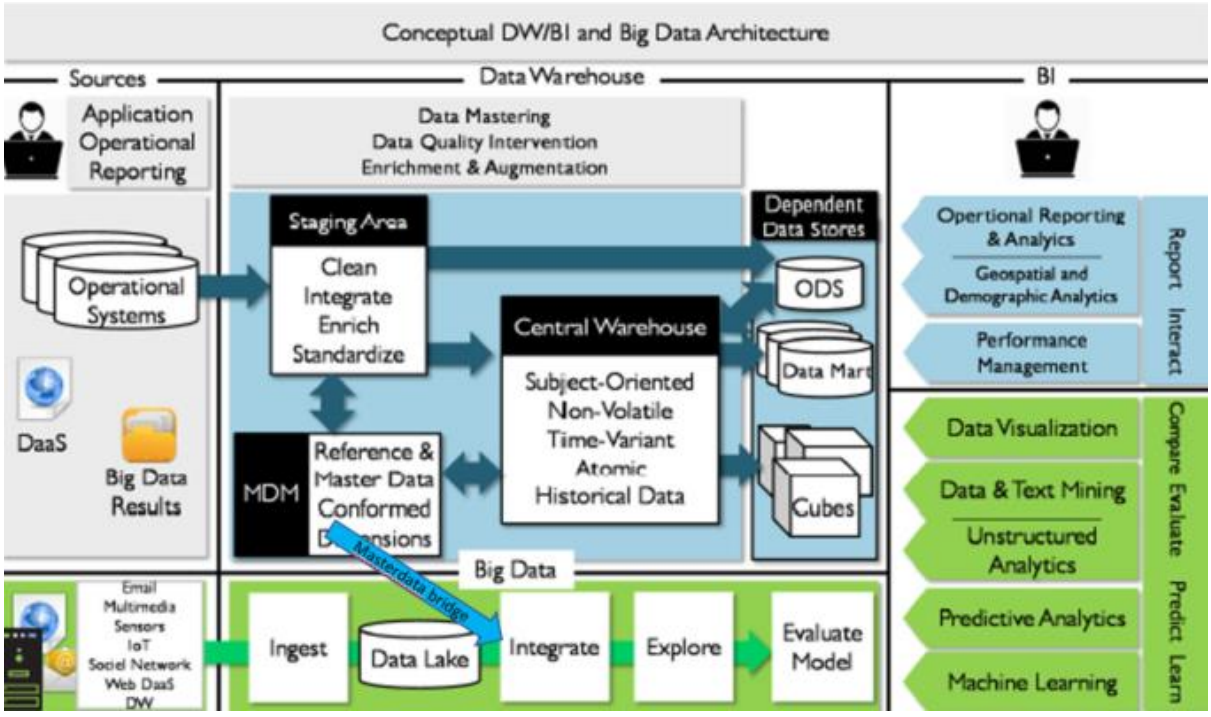


Figure 81 Kimball's Data Warehouse Chess Pieces[67]

1.3.7 DW Architecture Components



The evolution of Big Data has added to the DW/BI landscape by providing another data entry path. Lifecycle is depicted:

- **Source Systems:** Operational systems and external data to be brought into the DW/BI environment
- **Data Integration:** ETL, data virtualisation and other techniques to get data into a common form and location. (Arrows represent the integration process)
- **Data Storage Areas:** The warehouse has a set of storage areas

- **Staging area:** Intermediate storage for ETL and integration
- **Reference and Master data conformed dimensions:** May be separate repositories
- **Central warehouse:** DW Data usually persists in the central or atomic layer, which maintains all historic atomic data as well as the latest batch run. Data structure based on performance needs and use patterns:
  - Relationship between Business and surrogate keys for performance
  - Indexes and foreign keys to support dimensions
  - Change data capture (CDC) techniques that detect, maintain and store history
- **Operational data store (ODS):** Version of the central persisted store that supports operational use
- **Data Marts:** Type of data store often used to support presentation layers of the data warehouse environment
- **Cubes:** Three classic implementation approaches support Online Analytical Processing (OLAP). Their names relate to underlying database types, such as Relational (ROLAP), Multi-dimensional (MOLAP) and Hybrid (HOLAP).

1.3.8 Types of Load Processing

Historical (loaded once) and ongoing updates (consistently scheduled)

- **Historical Data:** Data warehouse captures detailed history of the data it stores
  - **Inmon:** All data stored in single DW layer with common integration and transformation layer. (Need Enterprise Model for success)
  - **Kimball:** DW is a combination of departmental data marts which store history at atomic level
  - **Data Vault:** Hybrid between 3NF and Star Schema (See DM & D Chapter 5). Hubs (PK), Links and Satellites. Facts persist as atomic structures.
- **Batch Change Data Capture:** Different change capture techniques:

Table 28 CDC Technique Comparison

Method	Source System Requirement	Complexity	Fact Load	Dimension Load	Overlap	Deletes
Time stamped Delta Load	Changes in the source system are stamped with the system date and time.	Low	Fast	Fast	Yes	No
Log Table Delta Load	Source system changes are captured and stored in log tables	Medium	Nominal	Nominal	Yes	Yes
Database Transaction Log	Database captures changes in the transaction log	High	Nominal	Nominal	No	Yes
Message Delta	Source system changes are published as [near] real-time messages	Extreme	Slow	Slow	No	Yes
Full Load	No change indicator, tables extracted in full and compared to identify change	Simple	Slow	Nominal	Yes	Yes

- **Near-real-time and Real-time:** Operational BI requires lower latency and the inclusion of volatile data in the warehouse.
  - **Trickle feeds (source accumulation):** Batch loads on a more frequent schedule, or when a threshold is reached

- **Messaging (Bus accumulation):** Small messages are published to a bus when they occur. Used by Data-as-a-Service (DaaS)
- **Streaming (Target accumulation):** A target system collects data as it is received into a buffer or queue.

2 Activities

2.1 Understand Requirements

Operational systems depend on precise, specific requirements. A data warehouse brings together data that will be used in many systems, used to explore and analyse.

2.2 Define and maintain the DW/BI Architecture

DW/BI architecture describes where the data comes from, where it goes, when it goes, why and how it goes into the warehouse. The technical requirements include performance, availability and timing needs.

2.2.1 Define DW/BI Technical Architecture

DW/BI architectures should have a mechanism to connect back to the transactional and operational reports in an atomic DW.

Conceptual model aligns with business needs. Test by prototyping. Validate with the Enterprise Data Model.

2.2.2 Define DW/BI Management Processes

Address production management with a coordinates and integrated maintenance process, delivering regular releases to the business community. Establish a release schedule to manage each update to the deployed data product as a software release.

2.3 Develop the Data Warehouse and Data Marts

Three concurrent development tracks:

- **Data:** The data business requires for the analysis it wants to do:
  - Identify best sources
  - Remediation, transformation, integration, storage and availability rules
  - How to handle data that does not fit
- **Technology**: Back end systems and processes supporting data storage and movement
- **Business Intelligence tools**

2.3.1.1 Map Sources to Targets

Source to target mapping establishes transformation rules for data elements from individual sources to the target system. A solid taxonomy is necessary. It is often the logical model.

2.3.1.2 Remediate and transform data

Cleansing activities enforce standards to correct and enhance the domain values of individual data elements. Should be done in source systems.

2.4 Populate the data warehouse

The DW/BI team must publish clear rules for what data detail the DW contains, and what will be available via only operational reporting.

Key factors to consider when defining the population approach:

- required latency
- availability of sources
- batch windows or upload intervals
- target databases
- dimensional aspects
- timeframe consistency of the data warehouse and data mart.
- Data quality processing
- Time to perform transformations and late-arriving dimensions and data rejects
- Change data capture processes

2.5 Implement the Business Intelligence Portfolio

Identify the right tools for business communities – find similarities through alignment of common business processes, analyses, management styles and requirements.

- **Group users according to needs:** Know the user groups, then match the tool.
    - IT Developers extracting data
    - Information consumers
    - Users' needs may change according to roles and skills
- Match tools to user requirements:
    - Imbedded analytics
    - Virtualisation
    - BI Suites

2.6 Maintain Data Products

An implemented data warehouse and its customer-facing BI tools is a data product. Enhancements and extensions should be implemented incrementally.

- **Release Management:** Release management is critical to an incremental development process that grow new capabilities. This process keeps the warehouse up-to-date, clean and performing at its best. Below a quarterly schedule is illustrated.

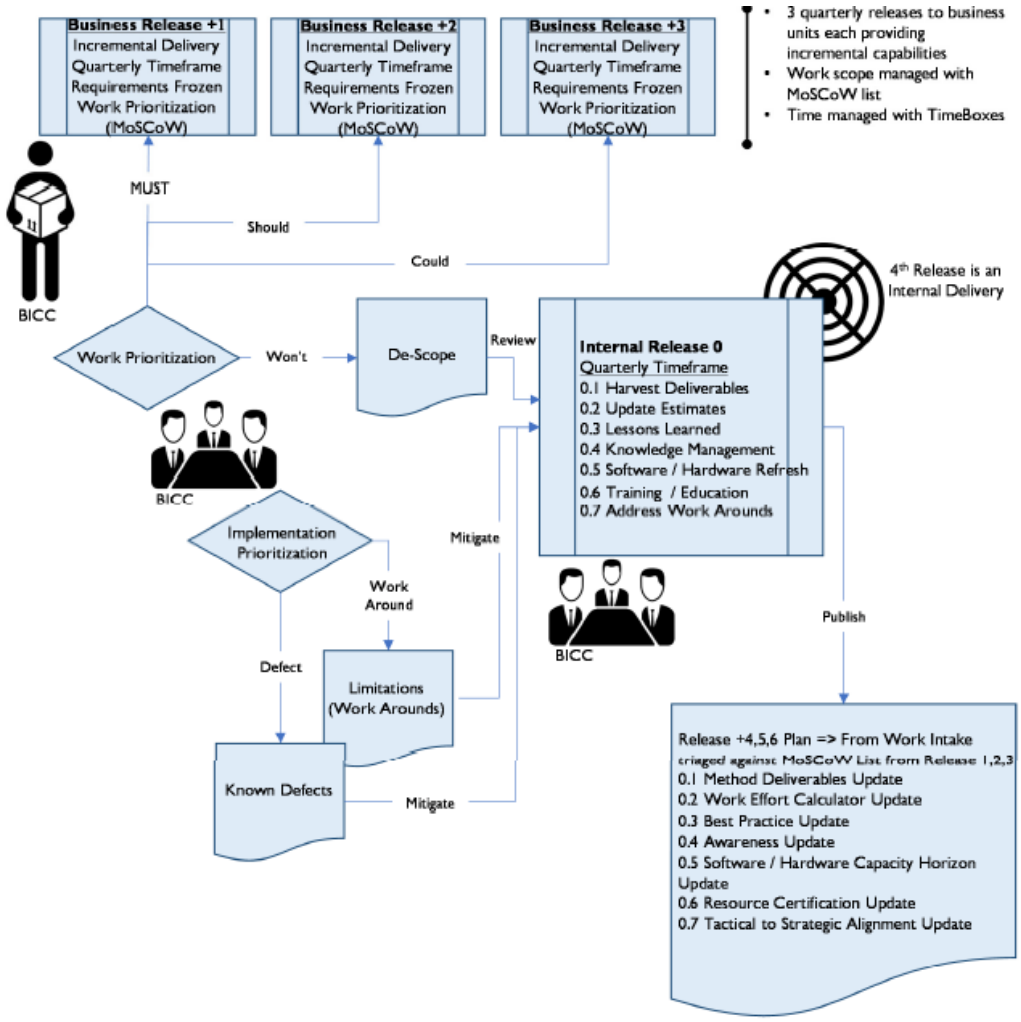


Figure 83 Release Process Example

- **Manage Data Product Development Lifecycle:** While the consumers are using the DW, the DW team is preparing the next iteration, which can extend an existing increment or add new functionality by onboarding a new business unit.
- **Monitor and tune load processes:** Monitor for bottlenecks and dependencies. Tune the database. Users consider the warehouse an active archive.
- **Monitor and tune BI activity and performance:** Customer-facing satisfaction metrics

3 Tools

- **Metadata repository:** Automate and integrate population
  - **Data dictionary / glossary:** Necessary for use of DW. Contains business terms and other information needed to use data.
  - **Data and Data Model Lineage:** Integration tools offer lineage analysis for the population code and the physical model and database. Uses for documented data lineage:
    - investigation of root causes of issues
    - impact analysis of changes
    - determine reliability of data based on origin
- **Data Integration Tools:** Used to populate the warehouse. Also schedule jobs that account for complex data delivery from many sources.
- **Business Intelligence Tools Types:**

- **Operational Reporting:** Analyse business trends to discover patterns. Tactical BI to support short term decisions
- **Business Performance Management:** Formal assessment of metrics aligned with organisational goals. Strategic BI to support long term corporate goals and objectives.
- **Descriptive, self-service analytics:** BI to the front line of business. Service-oriented Architecture (SOA) and Big Data.
- **Operational Analytic Applications:** Differ from OLAP and BI tools as they include processes to extract data from source systems, a data mart and pre-built reports and dashboards.
- **Multi-dimensional Analysis – OLAP:** Online Analytical processing
  - **Slice:** A subset of a multidimensional array corresponding to a single value for one or more members not in the subset
  - **Dice:** A slice on more than two dimensions of a data cube or more than two consecutive slices
  - **Drill down/up:** User navigates among levels of data from most summarised (up) to most detailed (down)
  - **Roll-up:** Compute all the data relationships for one or more dimensions.
  - **Pivot:** Changes the dimensional orientation of a report or page display
- **3 implementation processes:**
  - **Relational Online Analytical Processing (ROLAP):** Supports OLAP using multidimensionality in the two-dimensional tables of RDBMS
  - **Multi-dimensional Online Analytical Processing (MOLAP):** Supports OLAP by using specialised multi-dimensional database technology
  - **Hybrid Online Analytical Processing (HOLAP):** Combination of ROLAP and MOLAP

4 Techniques

- Prototypes to drive requirements
  - Profile the data:
    - Reduce risk of unexpected data
    - Disclose differences in sources that may be obstacles in integration
- Self-service BI
- Audit data that can be queried

5 Implementation Guidelines

- **Readiness assessment / Risk Assessment:** May be a gap between embracing venture and ability to sustain it
  - Pre-requisite checklist:
    - Has business support
    - Aligned with strategy
    - has defined architectural approach
  - Define data sensitivity and security constraints
  - Select tools
  - Secure resources
  - Create source data ingestion process
  - Identify sensitive or restricted data that may need masking
  - Ensure data governance processes for review and approval have been followed

- **Release Roadmap:** Suggested approach:
  - Incremental leveraging of DW Bus matrix as a marketing and communication tool
  - Use business-determined priorities tethered to exposure metrics to determine how much rigor and exposure to apply to each increment
  - Apply consistent needs and abilities processes to determine next business unit to onboard.
  - Maintain a back-order work item list to identify outstanding capabilities
  - Determine technical difficulties which may alter the order of delivery
  - Package work into a software release
  - Agree on pace of release (quarterly, monthly, weekly or faster if appropriate)
  - Roadmap of release dates to manage with business partners
- **Configuration Management:**
  - Aligns with release roadmap
  - Scripts to automate development, testing and transport to production
  - Provides version control as it brands the release at the database level
  - Automated and manually generated programs are tied to that brand
- **Organisation and Cultural Change:** Keep business focus throughout the DW/BI lifecycle. Align projects behind real business needs and assess necessary business support. Critical success factors are:
  - **Business sponsorship:** DW/BI projects require strong executive sponsorship, an engaged steering committee and commensurate funding
  - **Business goals and scope:** Clearly identified business need
  - **Business resources:** Management commitment to resource
  - **Business readiness:**
    - Business prepared for long term incremental delivery
    - Business committed to establishing centres of excellence to sustain product
    - Breadth of knowledge or skill gap within target community that can be crossed within a single increment
  - **Vision alignment:** How well does the IT strategy support business vision?
  - **Dedicated Team:** To manage ongoing operations of the production environment:
    - **Front office group** to notify maintenance team of deficiencies to be addressed
    - **Back office** support team ensures production configuration has executed as required.

6 DW/BI Governance

Governance activities should be completed and addressed during implementation. Specific governance deliverables can be added to the Software Development Lifecycle. Warehouse governance processes should be:

- Aligned with risk management
- Business driven as different business units have different needs
- Mitigate risks (not curtail execution)

Most critical functions:

- Those that govern the business operated discovery or refinement area
  - Handshaking
  - Instantiate data

    - Transfer data
    - Discard data
    - Data archival and time horizons for boundaries to prevent sprawl
- Those that ensure quality within the warehouse
    - Assign time, resources and programs to remediate data
- One-off events: part of lifecycle but curtail them in the pilot area
- Policies required for procedures in real-time environment
- Risk exposure mitigation matrix. Curtail risk with governance functions:

6.1 Enabling Business Acceptance

Sign-off by business is part of User Acceptance Testing, and is paramount for every DW/BI implementation. Critically important architectural sub-components:

- **Conceptual Data Model:** Key business concepts and how they are related to each other
- **Data quality feedback loop:**
    - How data issues are identified and remediated
    - Owners of the systems where they originated informed and held accountable for fixing them
    - How are issues caused by the DW data integration process remediated?
- **End-to-end Metadata:** How does the architecture support integrated end-to-end flow of Metadata? Is access to meaning and context part of the architecture?
- **End-to-end verifiable data lineage:** Is a system of record identified for all data?

6.2 Customer / User Satisfaction

Collecting, understanding and acting on customer feedback can be facilitated through regular meetings with user representatives.

6.3 Service Level Agreements

Specify business and technical requirements.

6.4 Reporting Strategy

Reporting strategy across the BI portfolio must address:

- Standards, processes, guidelines, best practices and procedures
- Security access to sensitive data elements to only entitled users
- Access mechanisms to describe how users want to interact, report, examine or view data
- User community type and appropriate tool to consume it with
- Nature of reports (summary, detailed, exception) and frequency, timing, distribution, storage formats
- Use of visualisation capabilities
- Trade-offs between timeliness and performance

Evaluate regularly to see if they are still providing value.

Data Source governance, monitoring and control are vital.

Centre of Excellence can empower business users to the self-service model.

6.5 Metrics

- **Usage Metrics:**
    - Number of registered, connected or concurrent users

    - Number of user accounts for each tool
    - Number of queries per user community per timeframe
- **Subject area coverage percentages**
    - How much of the warehouse is being accessed by each department?
    - Mapping operational sources to targets
- **Response and performance metrics**
    - Response times can be retrieved from query tools
    - Load times for each data product in raw format
    - Query records, data refresh and data extract times for objects provided to users
    - Use the metrics to validate or adjust service levels