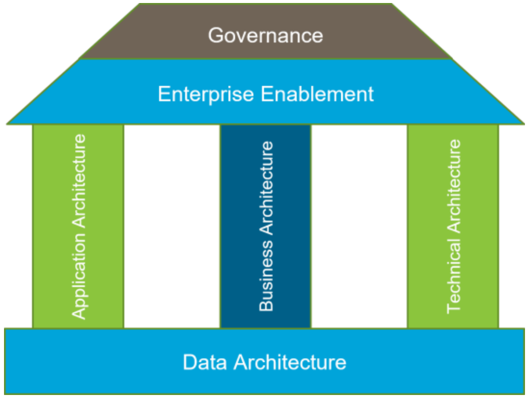
ENTERPRISE DATA ARCHITECTURE

IT4350 - Information Systems Architecture and Applications

Dr. Nguyen Binh Minh

1

Enable Data Governance With Enterprise Architecture



Governance

Enterprise Enablement

Application Architecture

Business Architecture

Technical Architecture

Data Architecture

2

ARCHITECTURE DEFINITIONS

- **Enterprise Architecture**
  - A defined practice (artifacts) for conducting enterprise analysis, design, planning, and implementation, using a comprehensive approach at all times, for the successful development and execution of strategy.
  - Enterprise architecture applies architecture principles and practices to guide organizations through the business, information, process, and technology changes necessary to execute their strategies.
- **Data Architecture**
  - Data architecture is composed of models, policies, rules or standards that govern which data is collected, and how it is stored, arranged, integrated, and put to use in data systems and in organizations.

ARCHITECTURE DEFINITIONS

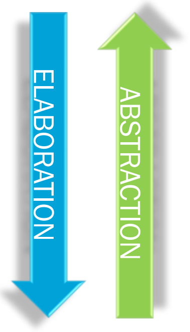
- **Business Architecture**
  - "A blueprint of the enterprise that provides a common understanding of the organization and is used to align strategic objectives and tactical demands."
- **Application Architecture**
  - Describes the behavior of applications used in a business, focused on how they interact with each other and with users. It is focused on the data consumed and produced by applications rather than their internal structure. Applications are mapped to business functions.
- **Technical Architecture**
  - Computer system architecture 'layer' which defines and specifies the interfaces, parameters, and protocols used by product architecture and system architecture layers.

Enterprise Data Serves Two Purposes

1. Running the Business
   - Tracking transactions such as sales, invoices, payments, deliveries, payroll, benefits, etc.
2. Managing the Business
   - Accounting, Budgeting and Planning
   - Marketing
   - Asset Management
   - Process Optimization
   - Strategy
     - Definition
     - Monitoring
     - Investment and resource allocation

DATA MODELS

- Conceptual
  - Technology-neutral, high-level layout of entities and their relationships
  - Used to establish contextual consensus among modeling domain stakeholders
- Logical
  - Adds detail to conceptual models in a technology-neutral rendering
  - More context on the entity relationships, including terms and definitions
- Physical
  - Tied to a particular database implementation
  - Includes implementation-level details such as indexing and federation



Why is Data Architecture So Hard?

- **Complex**
  - Data IS complex
  - The world changes rapidly, hard to keep up
- **Shared Data and Competing Stakeholder Interests**
  - Stakeholders have different definitions and uses for shared data
  - Conflicts are often resolved with incompatible models and duplication
- **Data Proliferation**
  - Useful data is replicated or transferred to other systems (e.g. accounting)
  - How and where data is used is often unknown
- **Data can be inaccurate**
  - Data entry is error prone and data duplication is common
  - Calculation errors are surprisingly common
- **Resource Intensive**
  - Databases can be very large
  - Managing data can consume a lot of staff resources
  - Tools and licenses can be pricey

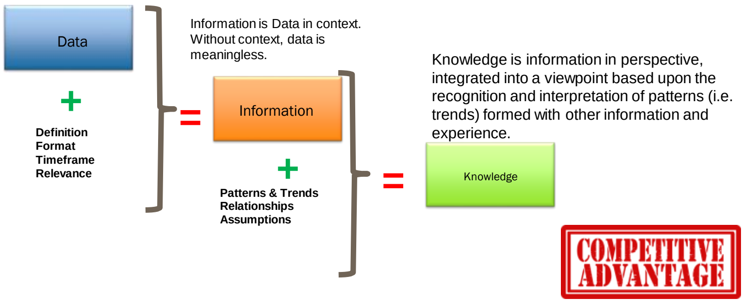
7

A Robust Data Architecture Should Be

- Trusted
- Accessible
- Timely
- Integrated
- Consistent
- Comprehensive (Complete)
- Verified
- Documented
- Discoverable
- Flexible
- Scalable
- Cost Effective
- Transparent and Visible
- Safe
- Governed
- Auditable (Lineage and eDiscovery)
- Secure

8

DATA VALUE CHAIN



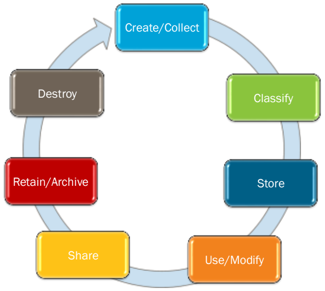
Information is Data in context. Without context, data is meaningless.

Knowledge is information in perspective, integrated into a viewpoint based upon the recognition and interpretation of patterns (i.e. trends) formed with other information and experience.

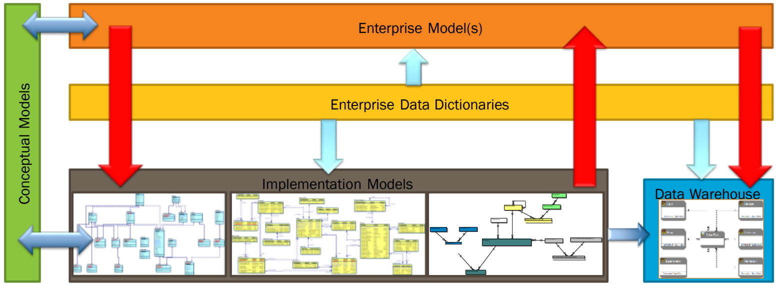
Data is the representation of facts as text, numbers, graphics, images sound or video

DATA - LIFECYCLE

- Describes how a data element is created, read, updated, deleted (CRUD)
- Many factors come into play
  - Business rules
  - Business processes
  - Applications
- There may be more than 1 way a particular data element is created
- Need to model:
  - Business process
  - Data lineage
    - Data flow
    - Integration
    - Include Extract Transform and Load (ETL) for data warehouse/data marts and staging areas



DATA MODELING CONTEXT



A Successful Data Architecture Must Deal With

1. Data Repositories
2. Data Capture
3. Data Definition and Design
4. Data Integration
5. Data Access and Distribution
6. Data Analysis

And More….

Data Repositories

- **Objective**:
  - Repositories must be extremely solid and reliable.
- **Issues**
  - Complexity of tools
  - Licensing, resource and skill costs can be high
  - Tolerance for failure is extremely low
- **Types of Data Repositories**
  - Application
  - Reference/Master
  - Meta Data
  - Data Warehouse
  - Discovery
  - Archive
  - Offsite/DR Repositories
- **Technology**
  - Traditional RDBMS: Oracle, DB2, SQL Server, Informix
  - Data Warehouse: Neteeza, Greenplum, Exadata, etc
  - Big Data: Hadoop, CouchDB, Redis, HBase, CouchDB

Data Capture

- Objective:
  - Capture all relevant business data
  - Ensure data is both accurate and complete at point of capture
- Issues
  - Great diversity (online entry, ftp, APIs, real time streams, etc., unstructured data)
  - Incoming data is often unreliable, incomplete
  - Big data explosion
- Types of Data Capture
  - Online
  - File transfers
  - Social media (real time)
  - Logs (web, security, o/s)
  - Data subscriptions (address databases, etc.)
- Technology
  - Online Applications (JEE, .NET, COBOL, etc.)
  - File transfer: FTP
  - APIs: Web services, REST, SOAP

Data Definition and Design

- **Objective**:
  - Organize the data to suit business requirements
  - Define data components, structures and processes to enable shared use and collaboration
  - Define rules for format, domain, and relationships
- **Issues**
  - Data modeling is complex
  - As data is distributed or shared, quality, structure and definition tend to worsen
  - Provide sufficient flexibility for graceful evolution
- **Types of Data Definition Artefacts**
  - Data Glossaries
  - Meta Data
  - Data Domains
  - Data Structures
  - Data Relationships
  - Validation Rules
  - Data Transformations
- **Technology**
  - Data modeling tools (Erwin, Power Designer, XMLSpy, etc.)
  - Metadata workbenches
  - Business glossaries

Data Integration

- **Objectives**
  - Data must usually be shared (copied or transferred) to realize its full value
  - Shared data requires transformation, filtering, verification, security controls, calculations, aggregations etc.
- **Issues**
  - In most cases, data integration is done in an inconsistent manner
  - Lack of enterprise standards and controls. Department silos
  - Data can become less unreliable each time it is handled
  - Rules for data integration poorly understood and documented
- **Types of Data Integration**
  - EAI, SOA
  - ETL
  - Federation
- **Technology**
  - ETL Tools: DataStage, Informatica, SSIS, Oracle ETL
  - ESB: Web services, REST, Websphere, WebLogic
  - Messaging servers: MQSeries, WebSphere, ActiveMQ, MessageBroker, Biztalk
  - Federated databases

Data Access and Distribution

- **Objectives**
  - Data is of no use unless it can be efficiently and reliably consumed by client applications
  - Data must be presented in a consistent, trusted and well understood way
- **Issues**
  - Access patterns are very diverse
  - Client requirements can cause conflicts that are difficult to resolve
- **Types of Access**
  - SQL
  - File APIs
  - Dashboards
  - Query engines
  - BI tools
  - Canned Reports
  - Discovery tools
  - APIs
  - Published views and queries
  - Publication and replication
- **Technology**
  - Reporting tools: Cognos, BO, Tableau
  - Data feeds: ETL
  - Query languages: SQL, SAS, R, etc.

17

Data Analysis

- **Objectives**
  - Provide insight into business operations
  - Assist in planning; provide predictions
- **Issues**
  - Complex and difficult to do right
  - Skills shortages
- **Types of Analysis**
  - Predictive analytics
  - Correlation
  - Pricing and underwriting
- **Technology**
  - R
  - Machine Learning algorithms

18

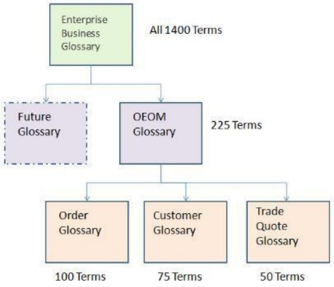
More Details

- There are many aspects to consider in good architecture, including:
  - Business Glossary
  - Metadata
  - Data Discovery and Profiling
  - Data Lineage
  - Data Glossary
  - Data Modeling
  - Master Data
  - Data Wrangling
  - Data Pipelines

Business Glossary

- Business Glossary is the place where all stakeholders can agree what the data actually means mean and what data rules apply
  - In the real world, lack of precision and conflict about terminology and meaning is common
- This information needs to be managed, coordinated and subject to quality controls
- An essential first step to a robust enterprise architecture
- Business stakeholders own the data definitions

BUSINESS GLOSSARIES



- Alignment to functional areas
- Child glossaries inherit a subset of parent terms
- No limit to hierarchy level

Meta Data

- For every data field, record and process its useful to provide detailed definitions, descriptions and rules that govern production, use and storage of any data item
- This information is managed, coordinated and subject to quality controls
- Targeted to technical stakeholders

Data Discovery and Profiling

- Data Discovery is about knowing where the data arrives from, where it goes to and where it is used
- Data profiling analyzes the content, structure, and relationships within data to uncover patterns and rules, inconsistencies, anomalies, and redundancies.

23

Data Lineage

- In many cases its important to know how data has reached a particular point.
  - Example: Why do reports A and B disagree about last month’s revenues?
  - Answer: Examine the data sources to ensure that both reports are using the same data sources.
- NOTE: Sometimes data can move a long way and through many steps before it ends up in a report

24

Data Synchronization and Reconciliation

- When data exists in multiple locations, there is always a danger that errors will occur
- Therefore as part of a robust control process, its useful to compare the data in source/target systems to find errors or omissions
- This can however become very complex as it may involve reconstructing the data lineage of every item

25

Data Modeling

- Data modeling and dimensional modeling is relatively mature, but the real world is complex, which can make some models hard to design, build and use
- NoSQL takes a different approach
  - Data is stored as documents
  - Data is flexible and dynamic

26

Master Data

- When data is generated and consumed in multiple locations and systems its very difficult to manage the data
- A Master Data Management (MDM) repository can:
  - Act as the corporate system of record
  - Define the shared canonical data model
  - Enforce data rules consistently across all systems
  - Resolve conflicts
  - Detect and remove duplicates

Data Wrangling

- Data is often less than perfect
  - Missing data
  - Incompatible fields
  - Semantic incompatibilities
- Data wrangling is the art of making merging diverse data sources into a consistent data store with consistent semantics, formats and values

Data Pipelines

- A data pipeline is the process which handles the transfer of data from one or more sources and delivers it to one or more targets
- A data pipeline may have the following stages:
  1. Extraction
  2. Cleansing
  3. Transformation
  4. Load

Types of Transfer:

1. Batch (ETL)
2. Continuous/Real Time
   - Push
   - Pull

Issues:

Completeness, Accuracy, Latency, Access method, Technology and tools

Data Pipelines – Data Extraction

- Data is extracted from a source system and landed in a staging area
- Data extracts can be:
  - Complete dumps of all data
  - Changes only
- Principles:
  - Capture Everything
  - Land everything in a staging area
  - Read once, write many

Data Pipelines - Data Cleansing

- Cleansing processes will fix or flag:
  - Invalid data
  - Missing data
  - Inaccurate data
- Principles:
  - Partition data into clean and rejected data.
  - Generate rejection reports
  - Land data into a zone for ‘clean’ data
  - Errors should be corrected in the source systems when detected

Data Pipelines - Data Transformation

- **Transformation** is a data integration function that modifies existing data or creates new data through functions such as calculations and aggregations.
- Common transformations include:
  - Calculations
  - Splits
  - Joins
  - Lookups
  - Aggregations
  - Filtering
- Principles
  - Stage transformed files to publish ready landing zones

Data Pipelines - Data Loading

- Data can be loaded:
  - As files
  - Via RDBMS utilities
  - SQL
  - Messages
- Principles
  - Group loading files by target repository

Data Pipelines – Design Practices

- For each stage of the data pipeline, its advisable to create:
  - Conceptual Models – what will happen
  - Logical Models – how will it happen (no technical details)
  - Physical Models – Technical details to realize the design