Enterprise GIS: Database Planning

John Alsup – Sr. Systems Engineer

Jeff DeWeese – Enterprise Solutions Architect

Agenda

- **Overview**
- **Database Design**
- **Data Maintenance**
- **Infrastructure Architecture**
- **Data Distribution**

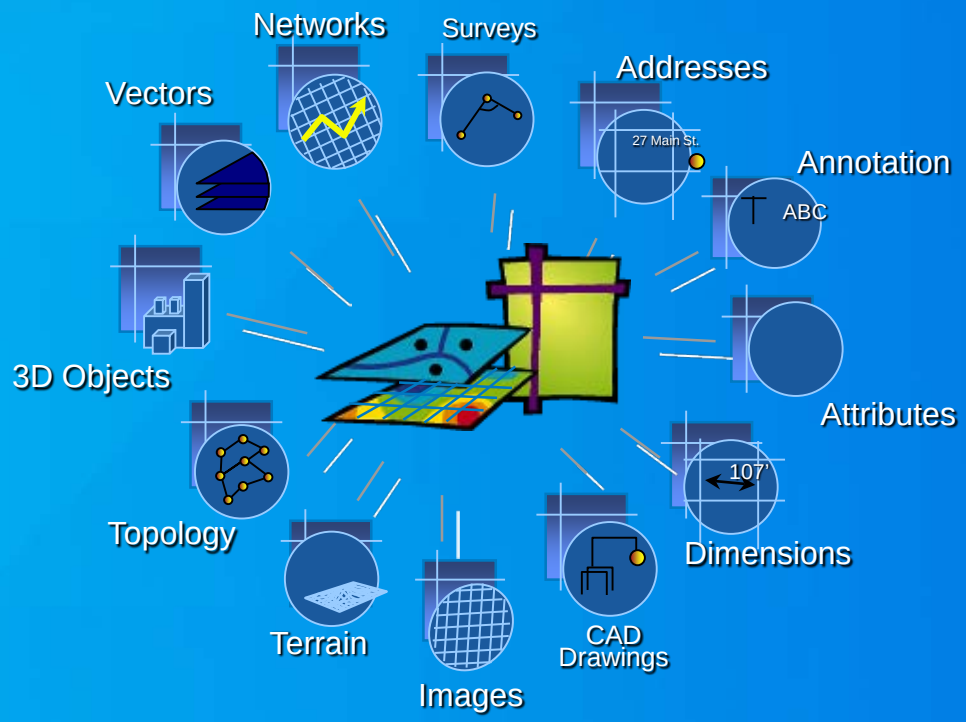
Learn the key planning phases and components of a geodatabase project. This session will outline Esri's experience and best practices involved in the design, development and delivery of an enterprise geodatabase.

Overview

John Alsup

What is a Geodatabase?

- A database or file structure used to store, query and manipulate spatial data.
- Data and functionality
- Three types:
  - File Geodatabase
  - Personal Geodatabase
  - ArcSDE Geodatabase
    - DB2
    - Informix
    - Oracle
    - PostgreSQL
    - SQL Server



Enterprise GIS

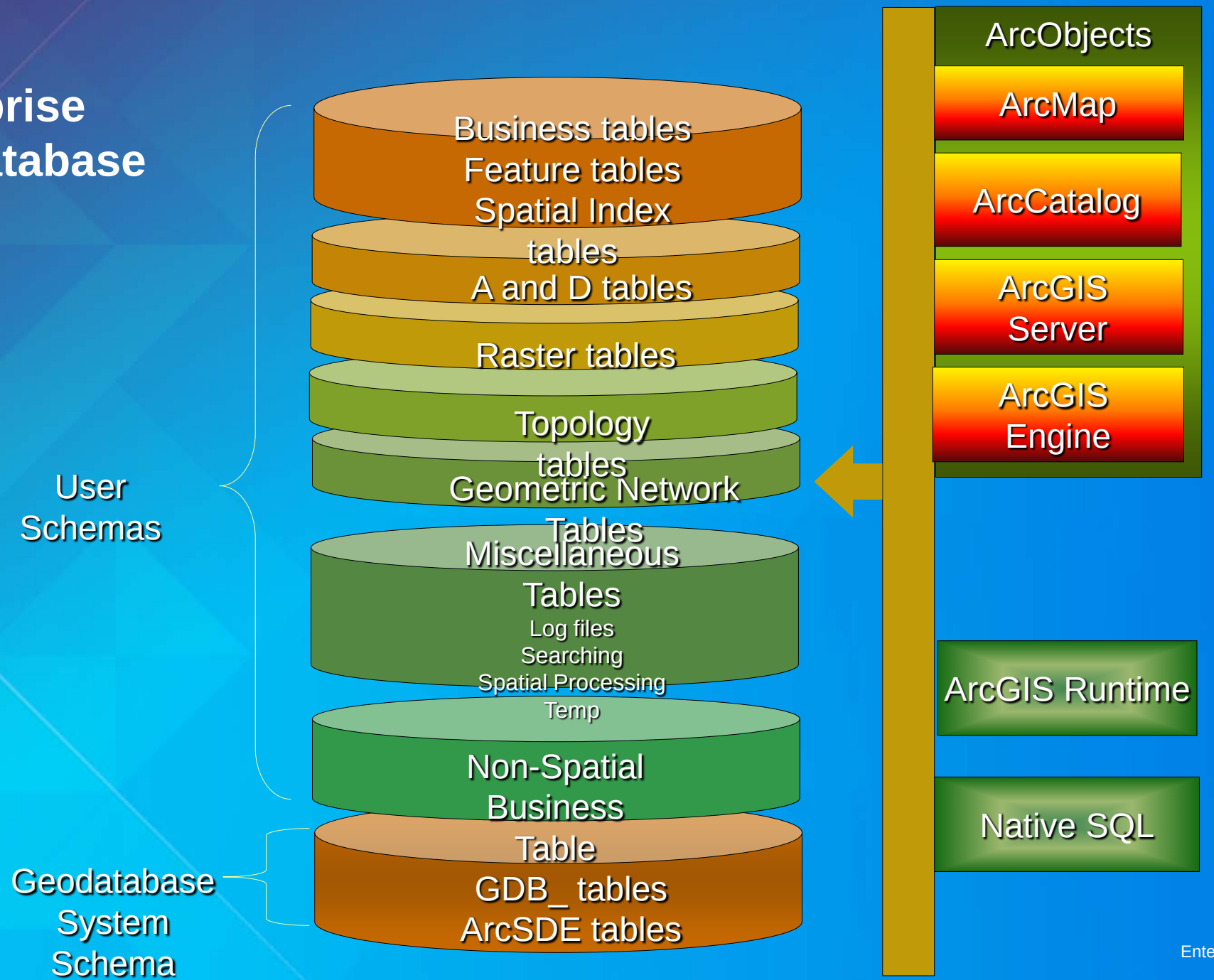
- GIS technology regarded by users and IT as key to business operations
  - May be considered mission critical
- Mainstream IT – deployed and managed like any other IT system
  - Architecture, Interfaces, Development tools, Deployment strategies, Standards
- Integrated with other enterprise systems
- Requires a higher level of planning, integration, testing and support

What is an Enterprise Geodatabase?

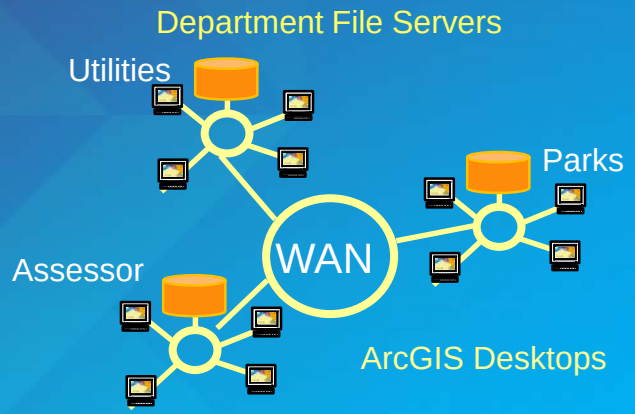
- Data
  - Serves data promptly and efficiently
  - Supports multiple users and departments concurrently
  - Provides seamless data
  - Centralized data management
  - Data integrity

- Functionality
  - SQL support
  - Collaborative editing, and long transactions
  - Quality control and quality assurance
  - Infrastructure for distributing and replicating data
  - Integrates spatial and business data with other systems
  - Leverages existing GIS and IT skills and resources

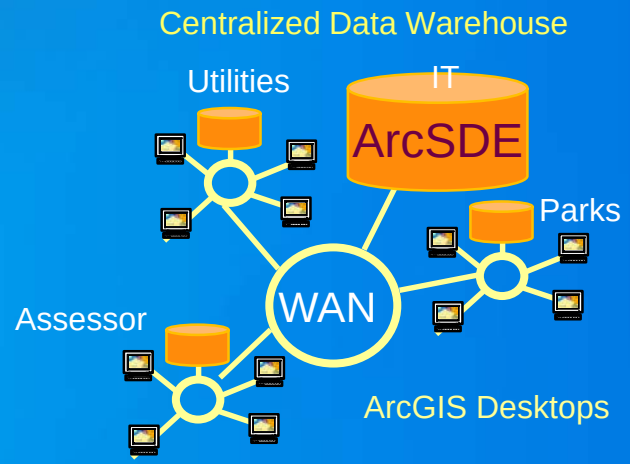
Enterprise Geodatabase



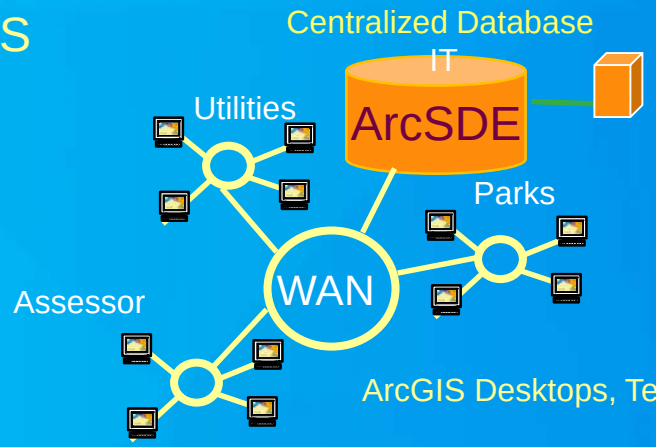
Organizational GIS Configurations



Distributed Client/Server
Departmental GIS



Data Warehouse
Departmental GIS Operations
Centralized Data Sharing



Centralized Database
Enterprise GIS Operations
Centralized Data Administration

Database Design

John Alsup

Why Plan an Enterprise Geodatabase?

- **Some key reasons:**
  - **Foundation for enterprise-wide use of GIS.**
  - **Geodatabase projects are complex**
  - **Enterprise Geodatabases and GIS application design requires diligent alignment**
  - **Large geodatabase projects span organizational groups and disciplines**
  - **Impacts almost every part of an enterprise GIS solution**

Spatial data is a key component of an enterprise GIS architecture . . .

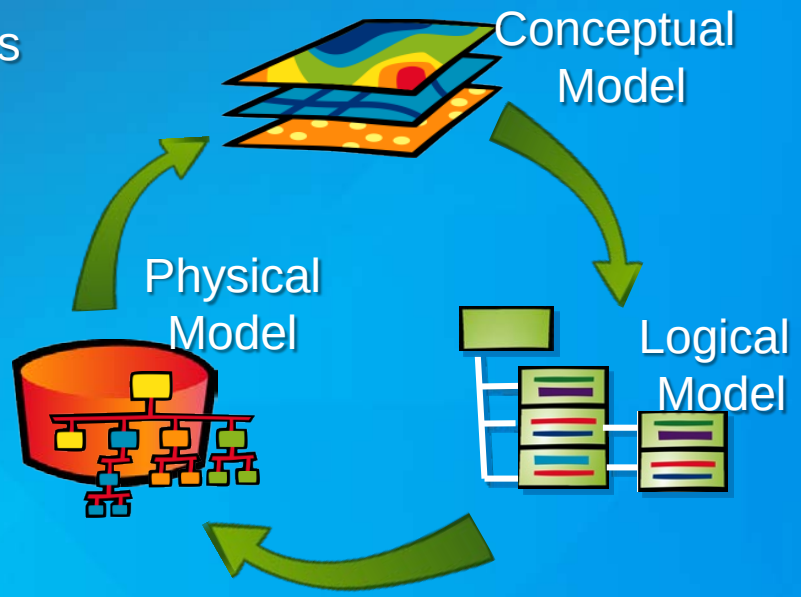
. . . delivery of spatial data must be fast, and this requires planning.

Geodatabase Project Scales

- **Larger Multi-phased Approach**
  - Elaborate, large databases
  - Custom applications
  - Large user base
  - Potentially outsourced, dedicated project management
- **Lighter Workgroup Approach**
  - Evolve the geodatabase, gradually upgrade data and applications
  - COTS application functionality where possible
  - Built in-house, part-time project management
- **All enterprise geodatabase projects require planning …**

Data Modeling Methodology

Three Stages



- **Conceptual Design Tasks:**
- **Identify business needs**
- **Identify thematic layers**
- **Identify required applications**
- **Leverage data model template**
- **Document**

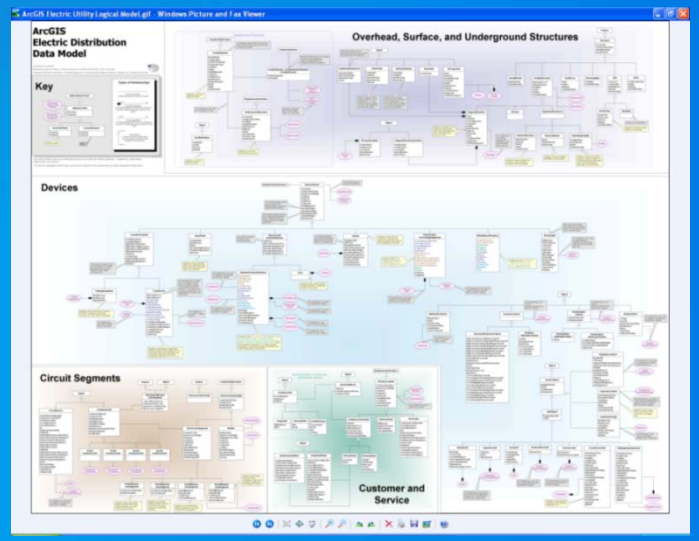
Physical Design Tasks:

- Create and implement model design
- Generate physical schema in the DBMS
- Testing and validation
- Document

Logical Design Tasks:

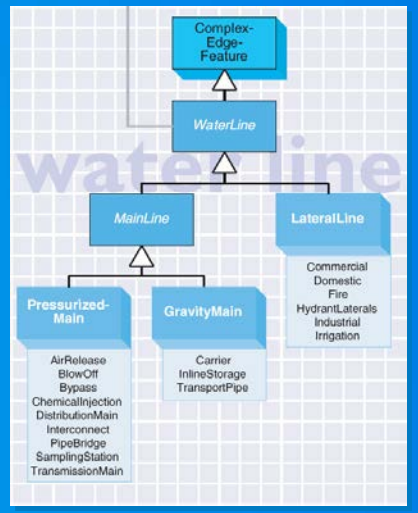
- Define tabular database structure
- Define relationships
- Determine spatial properties
- Document

Conceptual Model



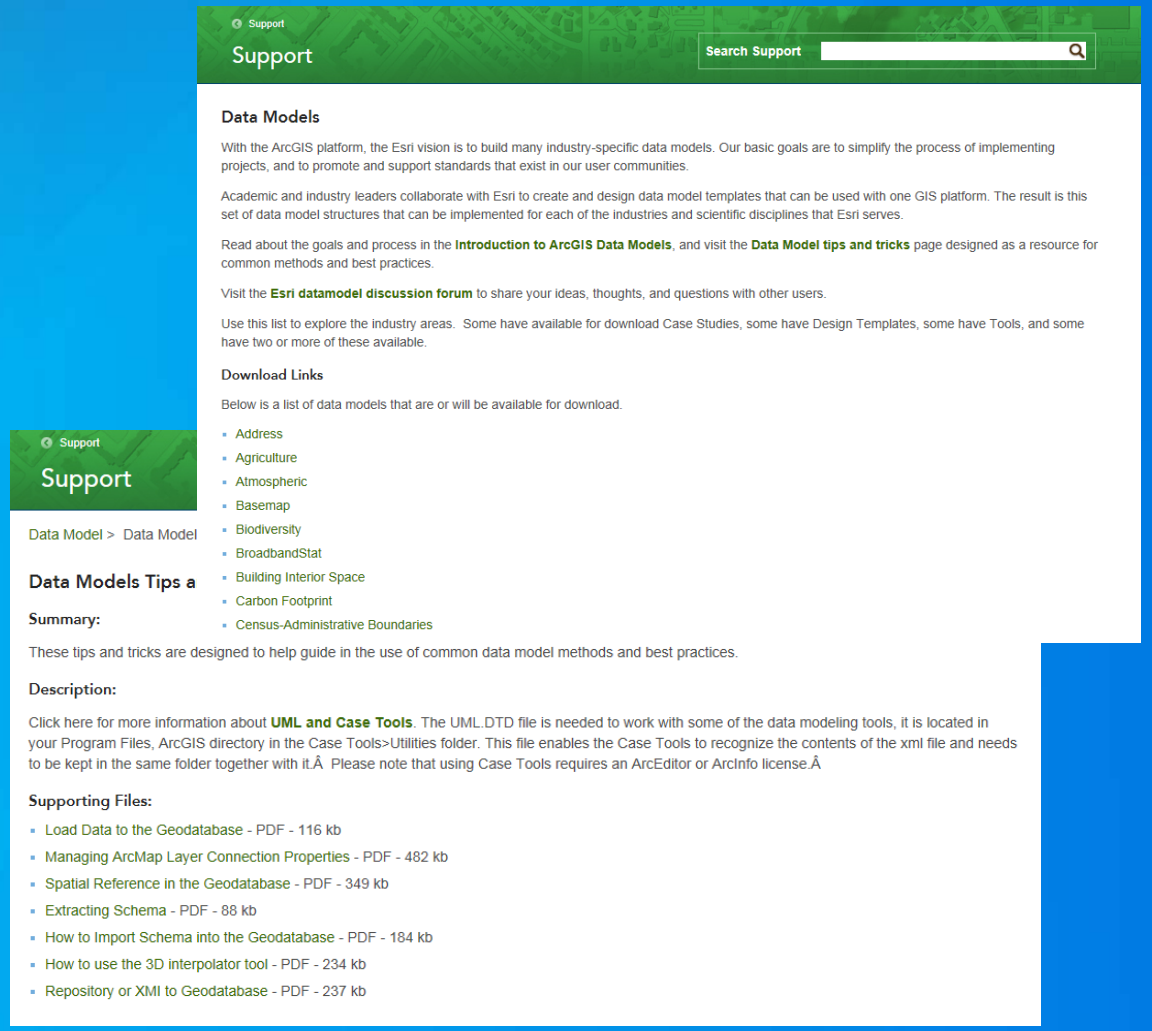
- **Identify and Document:**
  - Business needs - requirements
  - Thematic layers
  - Required applications and system interfaces

- **Leverage existing model templates**
  - Pre-designed schema of data objects
  - Best practices



ArcGIS Data Models Web site: http://support.esri.com/downloads/datamodel

- Over 25 industry-specific data models
- Conceptual and logical diagrams, sample Geodatabase schemas
- Case studies
- Tips and Tricks documents
- Developed and maintained by user and industry communities



Logical Model Design key topics

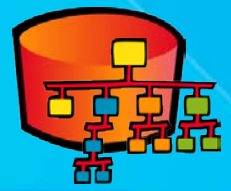
- Projection
  - Projection on the fly can be expensive
  - All feature classes in the same Geometric Network must use spatial reference
- Density of Features
  - High vertex count can be expensive
  - Any geometry with more than 4,000 vertices requires separate trips to the database
  - Complex symbology merged with complex features = slow
- Spatial placement vs. Logical placement
- Data update cycle
  - Replacement vs. editing
  - Change tracking/versioning
- Use of Feature Datasets
  - Designed for advanced functionality
  - Increased queries to the database when opening
- Relationships
  - Must be traversed during specific operations, like reconcile and replica synchronization

Important Considerations

- Field Names
- Geometry Storage Types
- RDBMS's used
- External systems and interfaces – key for enterprise GIS
  - CRM, WMS,SAP, other Financials, Reporting
  - Number of interfaces depends upon the organization
  - Consider data sharing - field data types, naming and length
- Symbology
  - Angle
  - Number of symbols
  - Scale suppression
- Geometry Storage

Physical Model Design

- Implementing the physical Geodatabase - prototype, test, review, and refine
- Documenting the design for distribution and efficient updating
- Test, refine and tune data model design for deployment



Field Name	Data Type
InstallationDate	Date
NominalDiameter	Text
NominalDiameterUnits	Text
ActualInternalDiameter	Double
OperatingPressure	Double
CoatingType	Text
CPSystemStatus	Long Integer
DesignPressureEntered	Double
DesignPressureStandard	Double
DesignPressureUnitOfMeasure	Text
ElectricSurveyIndicator	Text
LastLeakSurvey	Date
LeakSurveyFrequency	Text

Click any field to see its properties.

Field Properties

Alias	InstallationDate
Allow NULL values	Yes
Default Value	

- ControllableFitting
- ControlPoint
- CPAnode
- CPArea
- CPBondJunction
- CPBondWire
- CPRectifier
- CPRectifierCable
- CPTestPoint
- DistributionMain
- Drip
- ExposedPipeInspection
- GasGeomNetwork_Junctions
- GasLamp
- GasPipeCasing
- GasValve
- GatheringFieldPipe
- LeakReport
- LeakSurveyArea
- LineHeater
- MainJobSeparator
- MeterSetting
- NonControllableFitting
- Odorizer
- PipeExposure
- PipelineMarker
- PressureMonitoringDevice
- RegulatorStation
- ReliefValve
- RuralTap
- Scrubber
- Service

RDBMS Geometry Storage Format

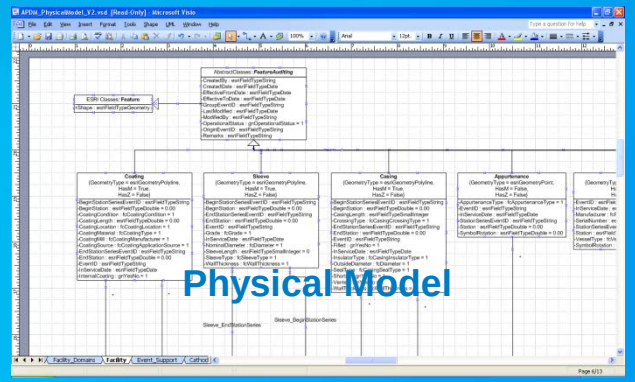
RDBMS	Geometry Storage
DB2	ST_Geometry, SDEBinary
Informix	ST_Geometry, SDEBinary
SQL Server	Geometry, Geography, , SDEBinary
Oracle	ST_Geometry, SDO, SDEBinary
PostgreSQL	ST_Geometry or Geometry
Netezza	VarChar (Shape)

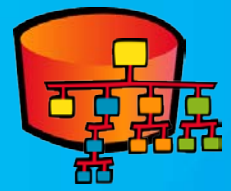
External System Interface

- ETL
- Database Level, duplicating data
  - Triggers
  - Update tables
- Database Views
  - Joins data from same or different databases

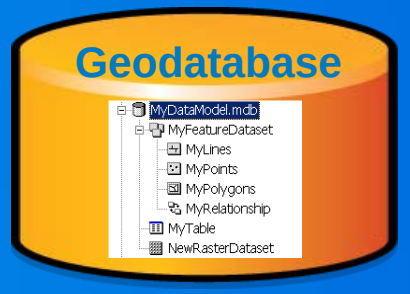
Creating Structure

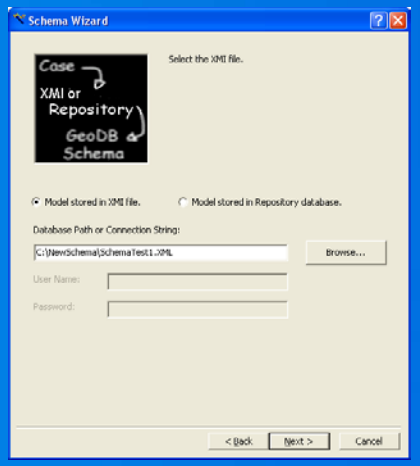
- Look to existing tools
  - CASE and UML tools – Visio, Rational Rose, etc.
  - Other tools (some free) and samples may work depending on approach
- Inheritance, re-use of objects through abstract and concrete classes





XMI (XML Design)

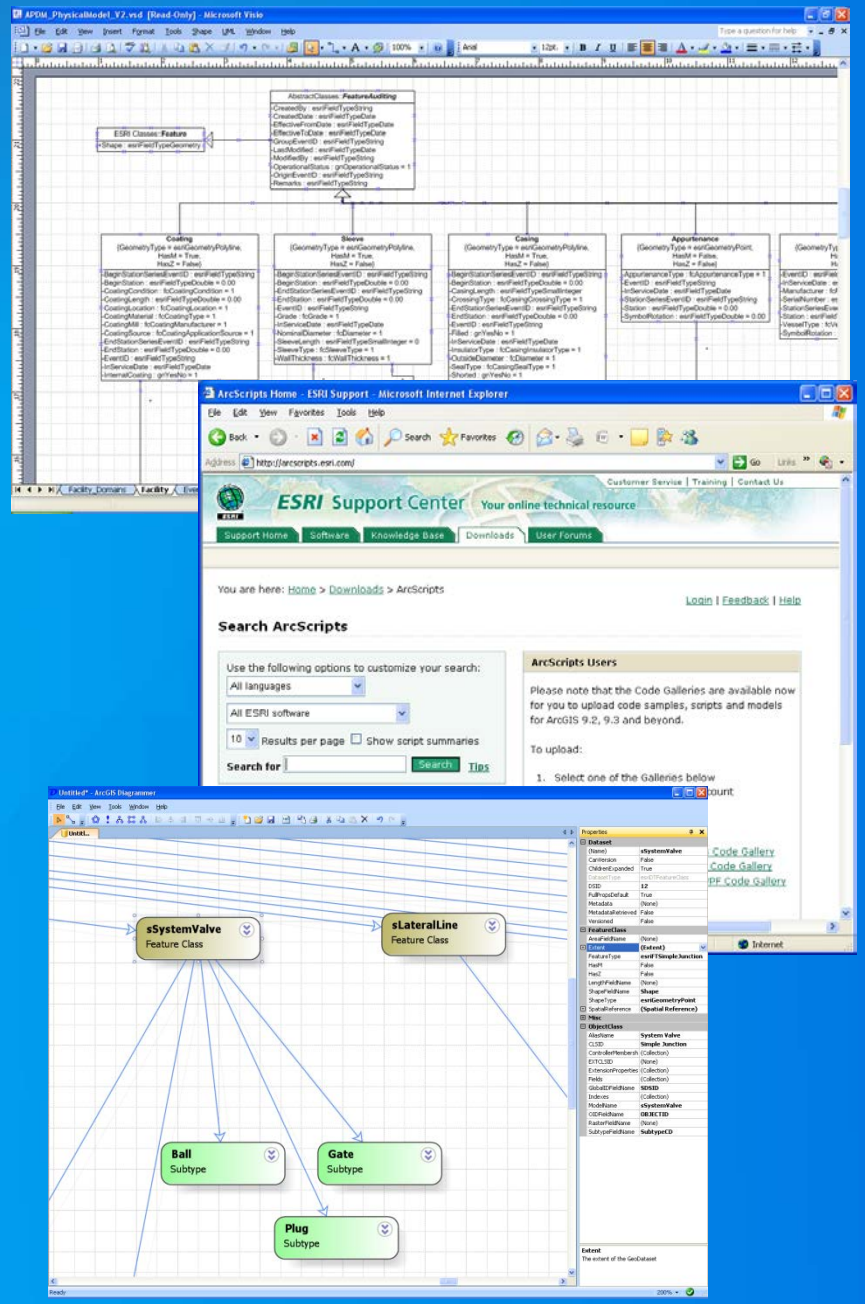




Data Modeling Tools

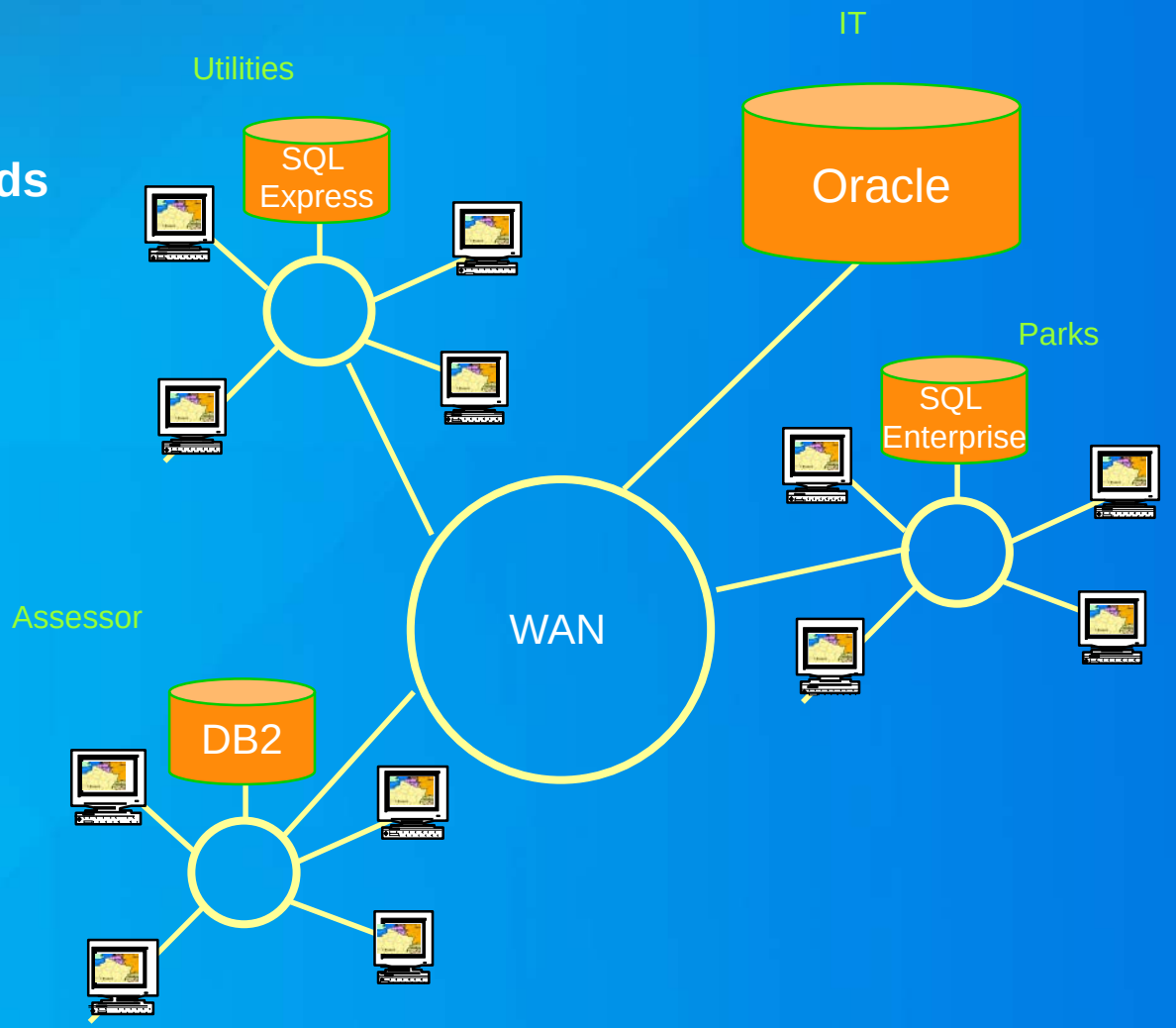
- Visio
- Rational Rose
- Enterprise Architect
- Free Esri Tools on ArcScripts:
  - ArcGIS Diagrammer
  - GDB Xray
  - Geodatabase Diagrammer
  - Geodatabase Designer

Free Tools are not supported....



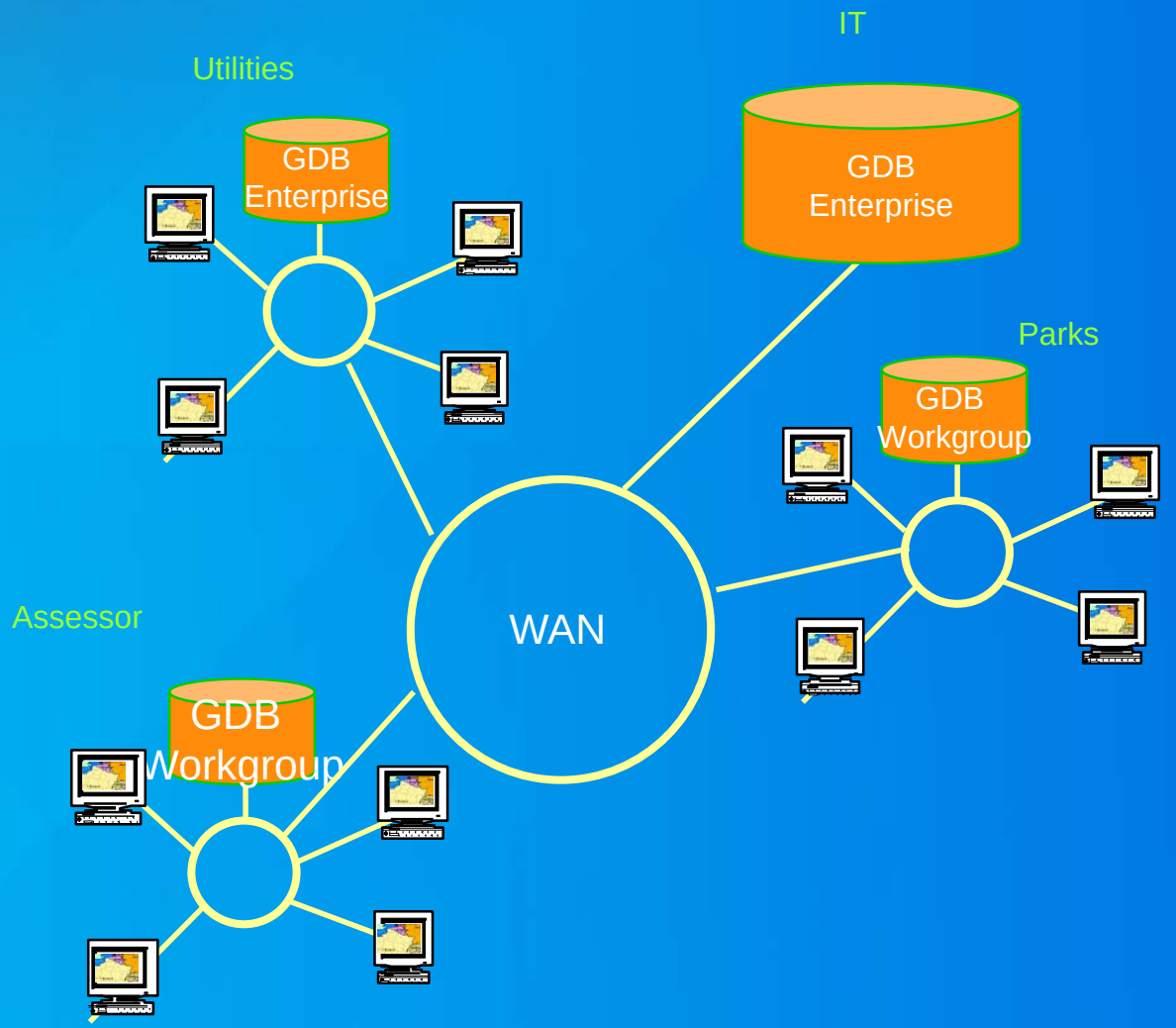
Mixed RDBMS Environments

- **For consideration:**
  - **Field Names, length and keywords**
  - **Field Data Types and Lengths**
  - **Database behaviors**



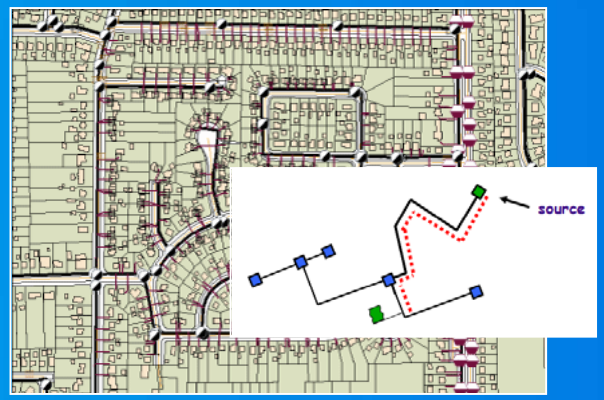
Mixed RDBMS GDB License Levels

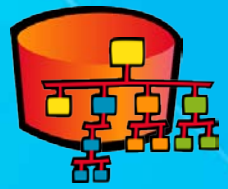
- **For consideration:**
  - **Domain authentication**
  - **Field Data Types and Lengths**
  - **Database behaviors**

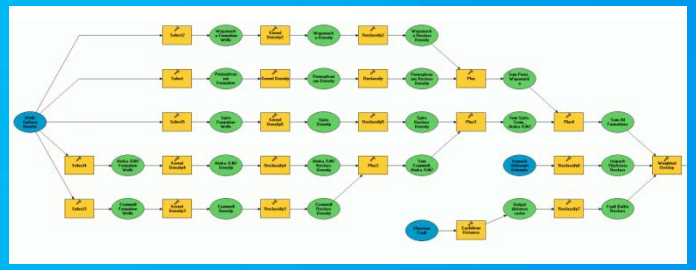


Testing and Refining

- Small pilot data migration with sample data
- Application testing – Test workflows
  - Functionality
  - Performance
  - Flexibility and consistency
- Team review and demonstration
  - Show how tasks are performed using GIS
  - Show maps, reports, online demos







Data Planning

- Migration and Conversion
  - Migration deals with moving existing geospatial data between different GIS environments or platforms
  - Conversion refers to development of new data by creating new digital geospatial data
  - Conversion is typically more significant and costly than migration
- Data procurement
  - Landbase
  - Imagery
- Data loading
  - Tools – In-house or outsourced
  - Procedures

Data Maintenance

John Alsup

Overview of Data Maintenance

- Plan and manage the maintenance workflow in the geodatabase

- Key Tasks
  - Analyze and build on business process requirements
  - QA/QC
  - Design your maintenance strategy
  - Plan for versioning
  - Define maintenance workflows

Consider QA / QC

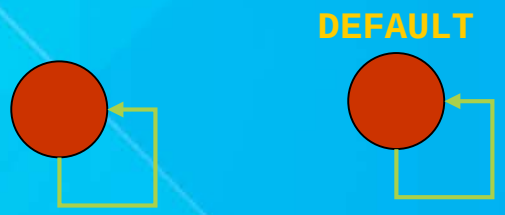
- Ensure data is captured, loaded and maintained accurately
- Quality Assurance
  - Review data to discover errors and perform data cleaning activities to improve quality.
- Quality Control
  - Ensure data products are designed to meet or exceed data requirements.
- QA/QC Plan
  - Versioning
  - Manual and automated procedures
  - Validations

Versioning and Multiuser Geodatabase

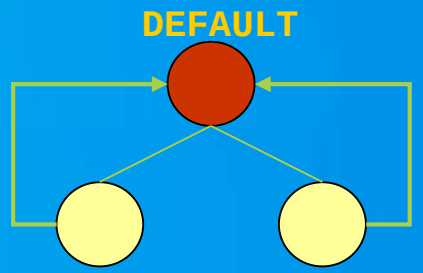
- **Defining versioning specifications and workflows:**
- **Versioning structure**
- **Reconcile, post, compress regimes**
- **Edit volumes, version durations**

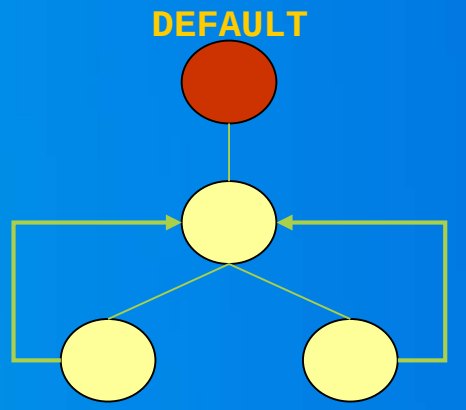
All impact performance…

Non-Versioned Editing



Versioned Editing





Considerations for Versions

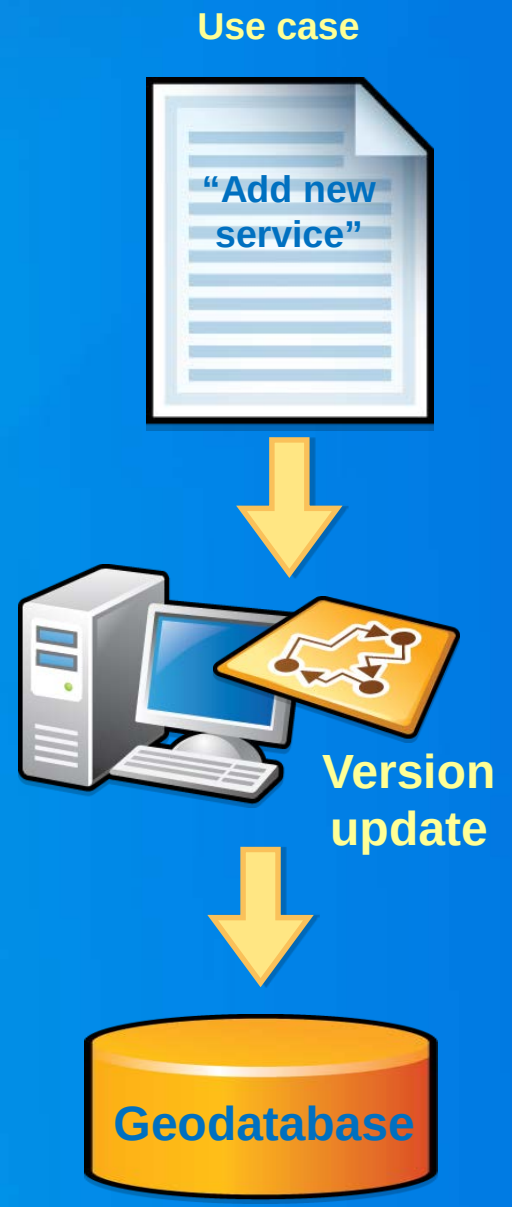
- Decide how versions will be handled:
  - Lifespan
  - Reconciling
  - Conflict management
  - Naming conventions
  - Structure
    - Staging or QC version between user versions and DEFAULT (Very risky!!!)
    - Security
    - Versions for groups or departments (Very risky!!!)
- Workflow Management Systems for Handling Versions
  - Can provide workflows and efficiencies , some examples:
    - Workflow Manager for ArcGIS
    - ArcFM and Network Engineer – In the Utility Area

Advanced GDB Functionality

- Relationship Classes
  - Persisted vs. temporal
- Geometric Network
  - Performance implications
- Topology

User Workflows

- Document with Use Cases
  - A description of the task you need to perform:
  - “Add new parcel”, “Update new asset”
- Evaluate business needs:
  - What data needs to be edited and in what order
  - Tracking of data changes
  - Conflict detection and resolution
- Security – user roles, etc.
- QA/QC steps – enforced through application or database



Data Performance and Scalability

- Essential Tasks
  - Review anticipated data loads
    - Volume (data file growth management)
    - Volatility (storage partitioning)
  - Identify key business transactions
    - Maintenance operations
    - Publication operations
  - Identify performance requirements for key business transactions
    - Response time
    - Initial and scheduled user loads
    - Throughput
    - Testing

Performance

- Geodatabase designs
  - Potential performance issues related to database design
    - Relationships
      - Both # and Type
      - Schema Cache can help reduce performance cost
    - Size of data stored in records
    - Projection on the fly
    - Number of records returned in a query
    - Density of data, both number of features and number of vertices
- Application design
  - Can have a significant affect on performance; e.g.,
    - Frequently opening a table
    - Retrieving features one at a time vs. bulk

Infrastructure Architecture

Jeff DeWeese

Infrastructure Design - Key Considerations

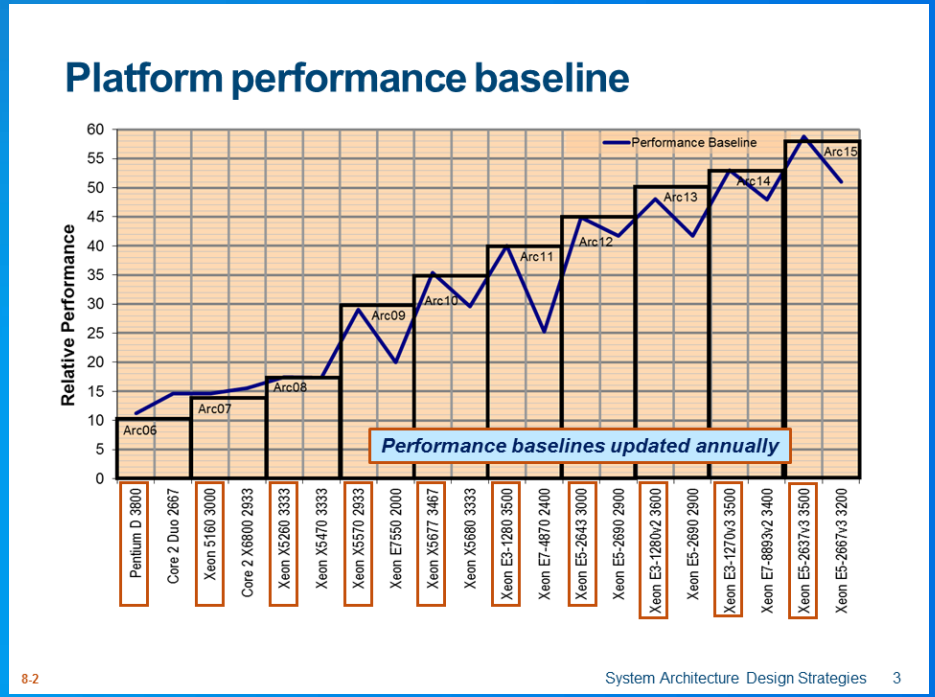
- Are actionable requirements available?
- Is the technology appropriate?
- Is it available enough?
- Is it continuous enough?
- Is there enough capacity?
- Is it meeting performance SLAs?

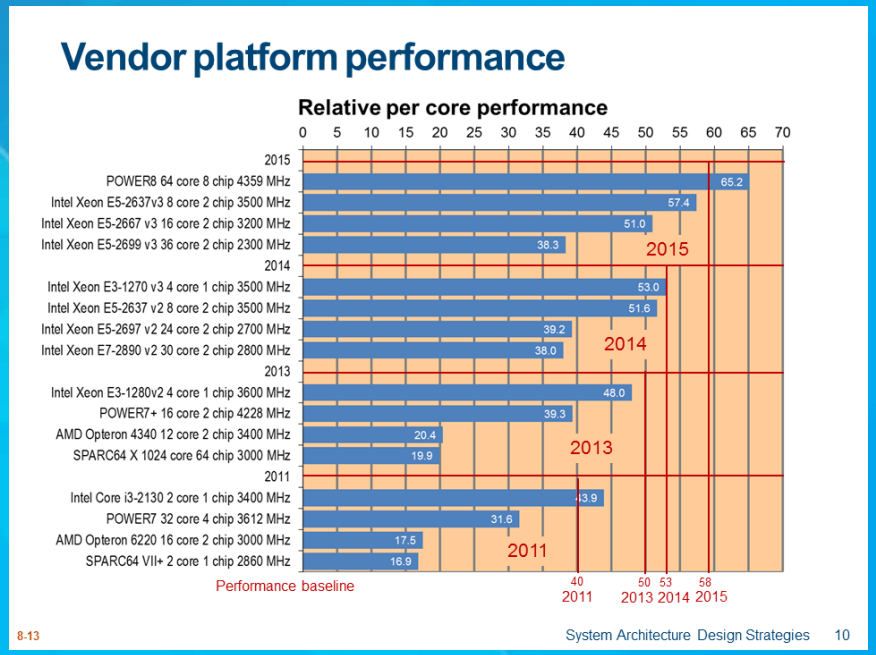
Requirements Driven Design

- Non-Functional Requirement (NFRs)
  - Functionality (architecturally significant)
    - Example: The system needs to support at-rest data encryption.
  - Usability
    - Example: The system needs to support dual monitors with at least 1920 x 1080 resolution.
  - Reliability
    - Example: The system should be available and online from 5am - 10 pm PST 7 days a week.
  - Performance
    - Example: Typical user interactions such as pans and zooms should be less than 5 seconds.
  - Supportability
    - Example: The system needs to scale easily by adding additional capacity with minimal reconfiguration.

Server Technology Selection

- **Why is Server Technology Selection Important?**
  - **Key to optimal scalability and performance**
  - **Keeps up with expected user productivity**
  - **Saves costs by reducing server footprint**





Case Study:

Customer was using 3 year old 32 core server. Recommended upgrading to available newer 24 core server.

- ***30% faster performance (DB processing)***
- ***Slightly increased server scalability***
- ***Oracle licensing savings of $100K+***

Virtualized Database Servers

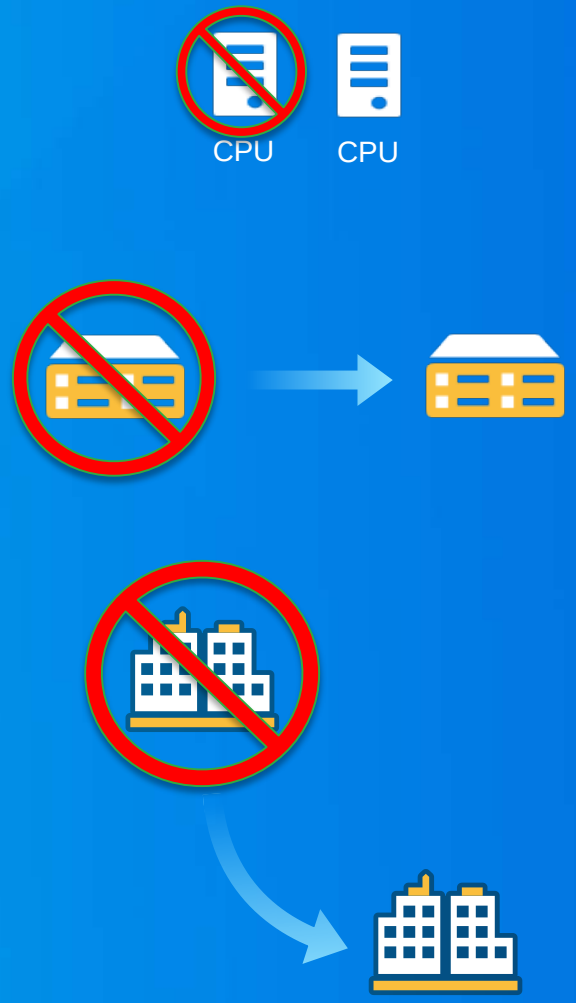
- Historical recommendation = “Avoid it!”
- Last year’s recommendation = “Test it!”
- Today’s recommendation = “Consider it!”

Physical vs. Virtual Server

- Reality Check
  - Virtualization is expanding and many IT organization desire to virtualize “everything”
  - Many database servers can be virtualized
  - Multiple factors to consider:
    - Architecture of the virtualized infrastructure
    - Accuracy of the access plan
    - Database complexity
    - Application complexity
    - User load

Fault Tolerance vs. High Availability vs. Disaster Recovery

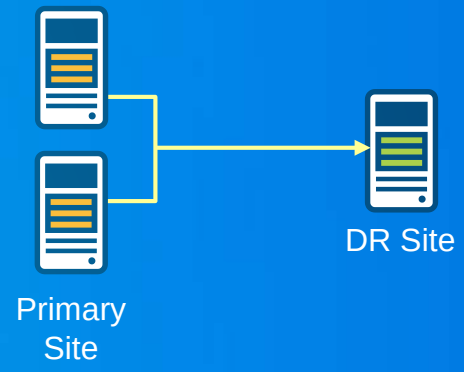
- **Fault-Tolerance (FT) addresses localized failures by the use of redundant components means.**
  - **CPU failure / Disk failure / Power Supply**
- **High-Availability (HA) addresses minor outages in a short time frame with largely automated means.**
  - **Server failure / SAN storage failure**
- **Disaster Recovery (DR) addresses major outages that are expected to last for a significant time period.**
  - **Flood / Fire / Earthquake**
  - **Core network failure / Major power outage**



High-Availability Design Considerations

- **Business requirements driven**
  - The Business defines their needs as Non-Functional Requirements (NFRs)
  - IT responds by providing a standards-based technical solution
  - System classification: Mission Critical, Business Critical, etc.

- **Balance between benefits and costs**
  - HA/DR configurations are more complex/costly
  - HA/DR configurations result in increased administration



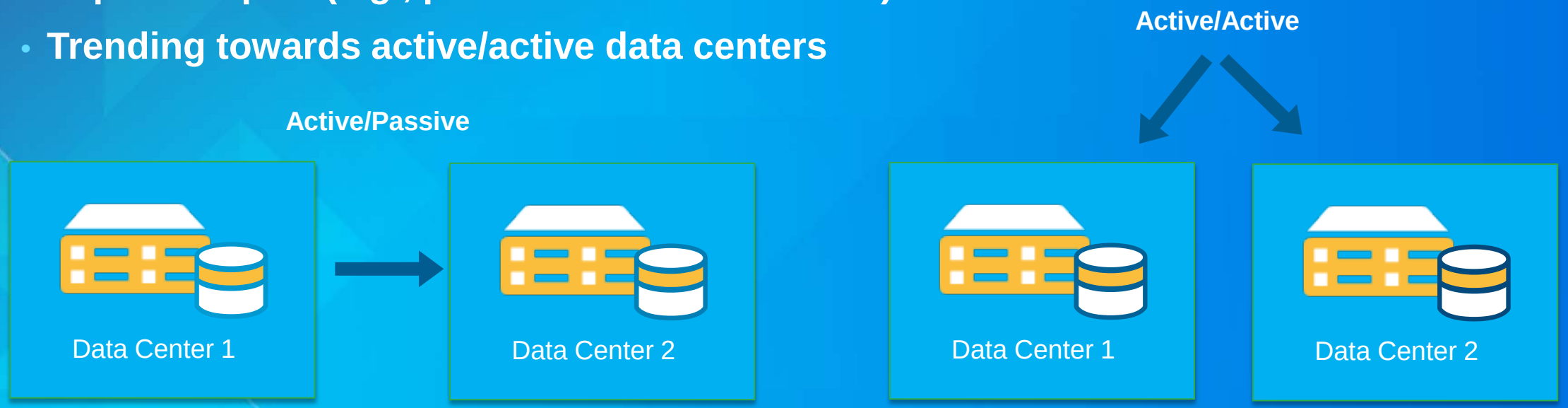
- **Consider maintenance windows**
  - Compress / DB statistics
  - Reconcile / Post services
  - Database schema changes / software patching
  - Database integrity checks post-restore

High-Availability Database Solutions

- **Fault Tolerant Servers**
  - Provides seamless failover – nearly zero down-time
    - Stratus FT Server
    - VMware vSphere Fault Tolerance
- **Active/Active Clusters**
  - User connections fail-over to remaining active node(s) – minimal down-time
    - Oracle RAC
- **Active/Passive Fail-Over Clusters**
  - Services fail-over to stand-by node – recovery time in minutes
    - Microsoft Failover Cluster
- **Virtual Server Clusters**
  - Provides “basic HA” – recovery time in minutes
    - VMware vSphere
    - Citrix XenServer
    - Microsoft Hyper-V

Database Disaster Recovery Options

- Virtual server replication solutions
- Synchronous database solutions (e.g., Oracle DataGuard, SQL Server AlwaysOn)
- Export / Import (e.g., publication environments)
- Trending towards active/active data centers



Tip: Proceed with caution when crossing data centers for active/active to a single database.

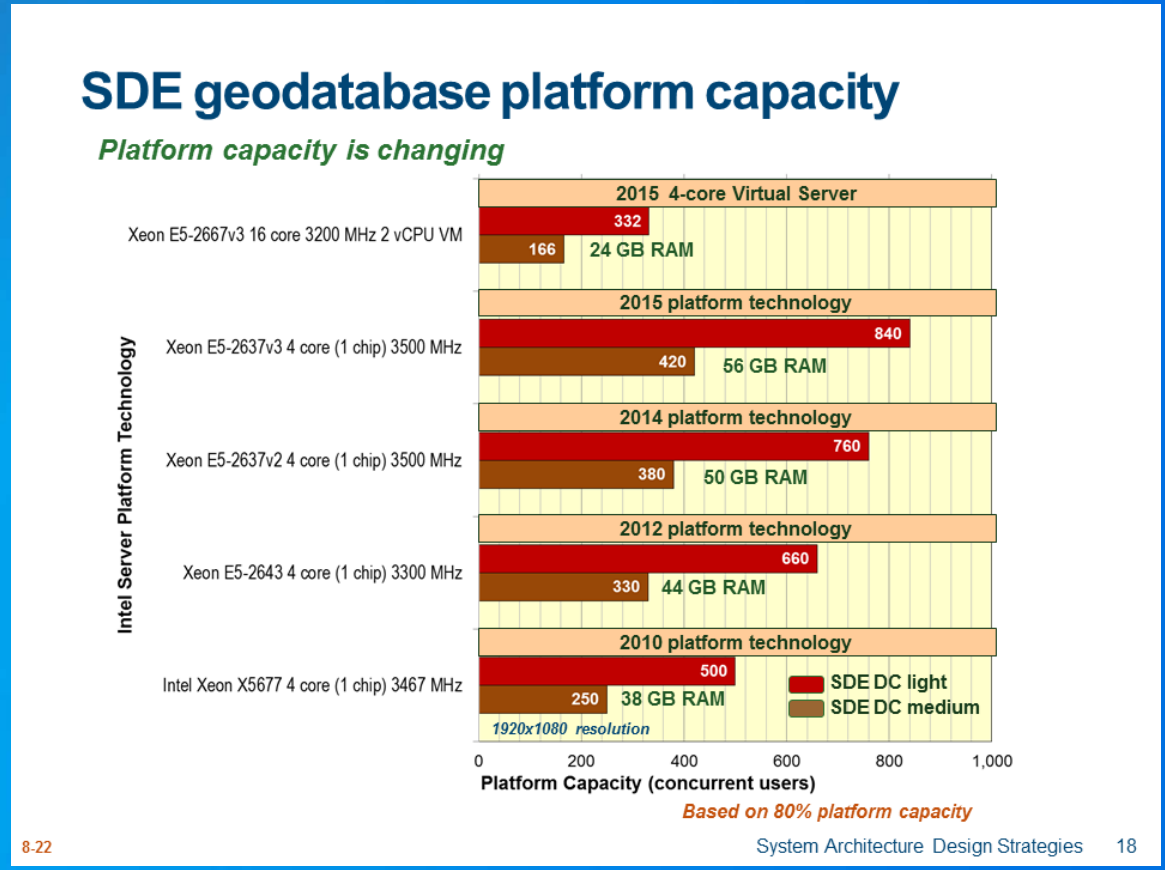
Database Disaster Recovery Options - Enhancements

- **Oracle DataGuard Stand-By Database**
  - Prior to 10.3.1, connections to a stand-by DataGuard stand-by database were not supported due to necessary database updates (e.g., writes to process information table)
  - At 10.3.1 and 10.2.1 GUP3 added functionality to detect if the database is in a read-only state, and modify how we connect (e.g. don't write to the process information table)
  - Some limitations such as performing selection >100 features which updates log files
- **SQL Server AlwaysOn Availability Groups**
  - At 10.2 added functionality to detect if the database is in a read-only state, and modify how we connect (e.g. don't write to the process information table)
  - At 10.3 added support for additional AlwaysOn/Failover Cluster connection parameters
    - APPLICATIONINTENT: Allows you to direct connections made to the Listener to either the READWRITE primary (this is the default) or to a READONLY secondary.
    - MULTISUBNETFAILOVER: Enables faster failover for all Availability Groups and or Failover Cluster Instance
    - https://desktop.arcgis.com/en/desktop/latest/manage-data/gdbs-in-sql-server/connections-highly-available-sqlserver.htm

Database Server CPU Capacity

- **Server CPU capacity is a function of:**
  - **CPU service time**
  - **Throughput**
  - **Max allowed CPU%**
  - **Relative performance of the hardware**

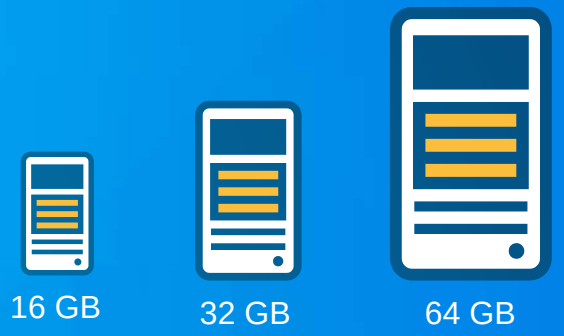
$$Required\ Cores = \frac{ST \times Tput(tph)}{3600 \times CPUmax} \times \frac{SPECbaseline}{SPECtarget}$$



Tip: Proper capacity is required to support expected peak user loads while maintaining acceptable performance.

DB Server Memory Capacity

- **Memory capacity is a function of:**
  - **Number of DB instances**
  - **Memory per connection**
  - **Number of MXD layers**
  - **Number of connections**
  - **Database size**
  - **Index size**



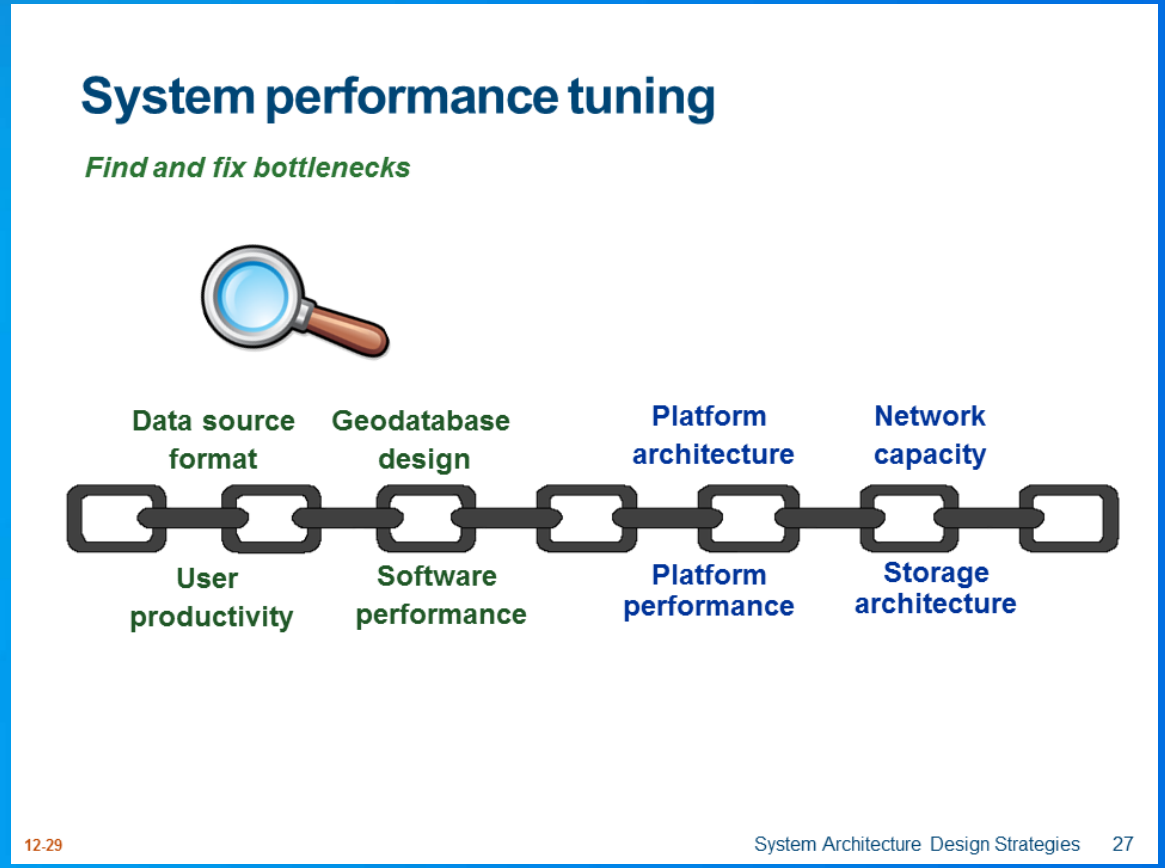
Tip: Providing adequate memory for the database server is critical for scalability and performance.

Single vs. Multiple Instances

- There are trade-offs….
- Single Instance
  - Main Driver: Reduced administration time and resources.
  - Larger databases result in scalability challenges (e.g., large data dictionary)
  - Increased chances of performance issues from poor queries, etc., impacting entire environment.
- Multiple Instances
  - Main Driver: Separation of data, department isolation, security, etc.
  - Need to carefully manage server resources
  - Increased database administration time and resources (e.g., more scripts)

Removing Constraints

- ***The infrastructure can only be as good as constraints allow…***
  - **Is the server hardware adequate?**
  - **Is the network adequate?**
  - **Has the database been tuned?**
  - **Are the workflows reasonable?**
  - **Is the storage architecture bottleneck free?**



Data Distribution

Jeff DeWeese

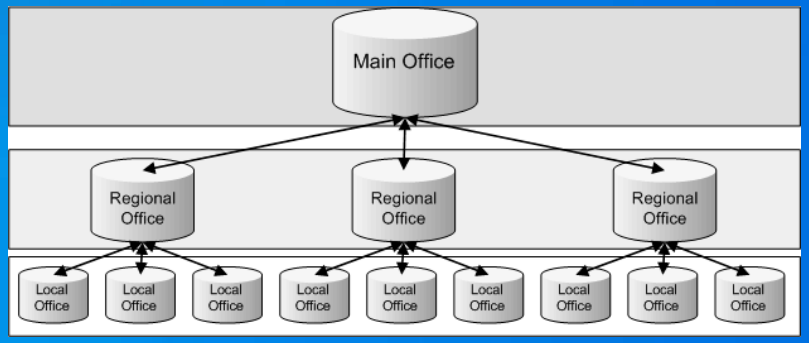
Data Distribution Options

- **Copy/Paste**
  - Export to FGDB / Import
  - Can be very time consuming
  - Potentially does not synchronize GUIDs and Object IDs
- **Database export/import**
  - Target database has to be stopped for the update
  - Can be very time consuming (entire DB export)
- **DBMS level replication**
  - Snapshot / Multi-master/ Merge / Transactional / Hybrid
  - Limited since NOT geodatabase or version aware!
  - Does not know how to properly replicate advanced geodatabase objects
  - Cannot edit DBMS replica using ArcGIS…only parent can be edited
  - Avoid using DB solutions (e.g., Oracle DataGuard) for replication as they need to be in a “passive” state.
- **Esri Geodatabase Replication**
  - Next slide….

Q. Which one should I use?

A. Depends….

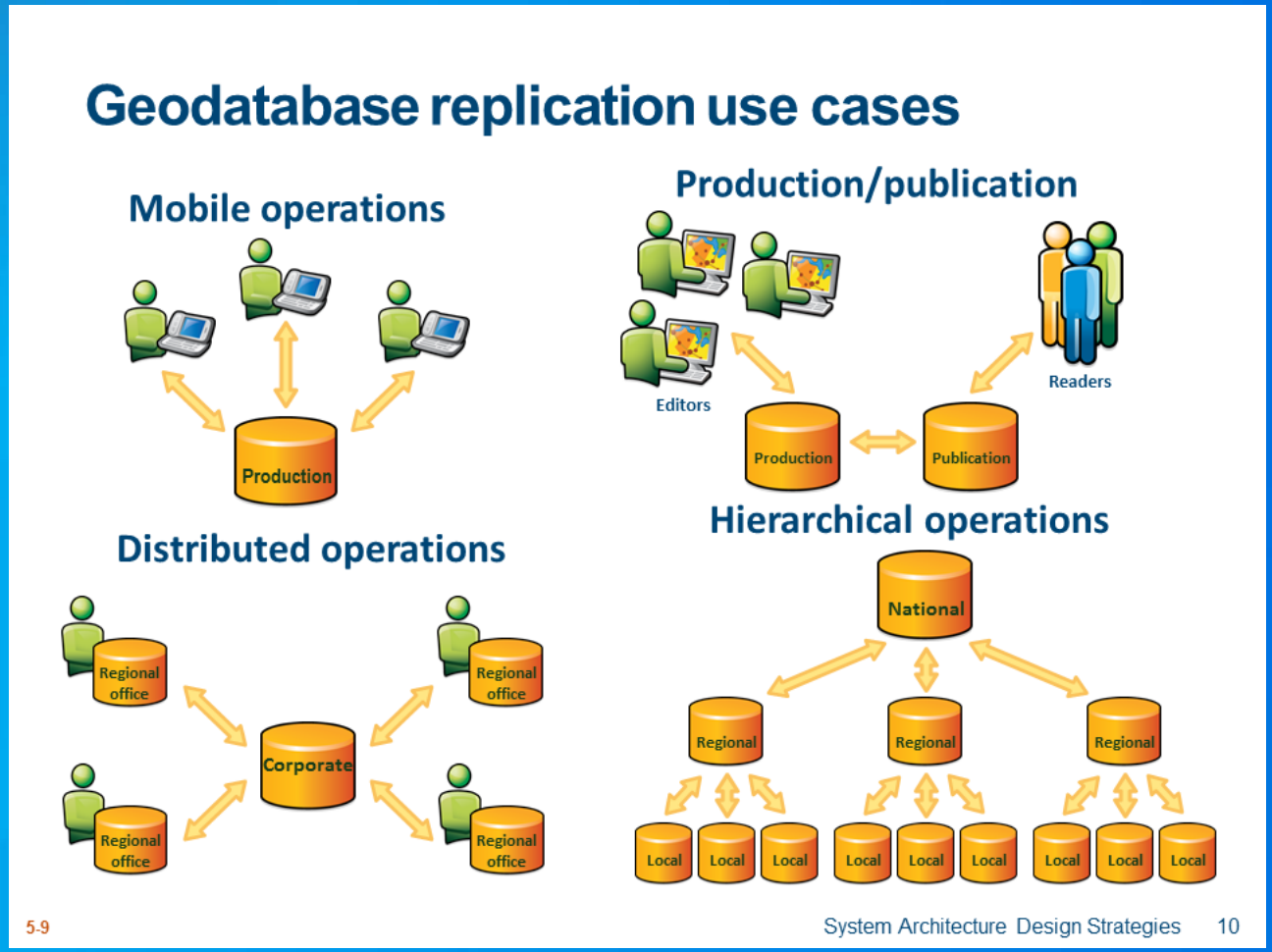
Geodatabase Replication



- Leverage Geodatabase Replication when possible
  - Built on top of geodatabase versioning
  - Supports the full geodatabase data model such as topologies, networks and relationships
  - Avoids limits or complexities associated with other methods
  - May not be the best solution in all circumstances (e.g., complex data models, complex geometric networks, large numbers of relationship classes).
  - Some data type limits (e.g., Network dataset time zone tables)
  - Much improved stability and reliability with latest software releases.

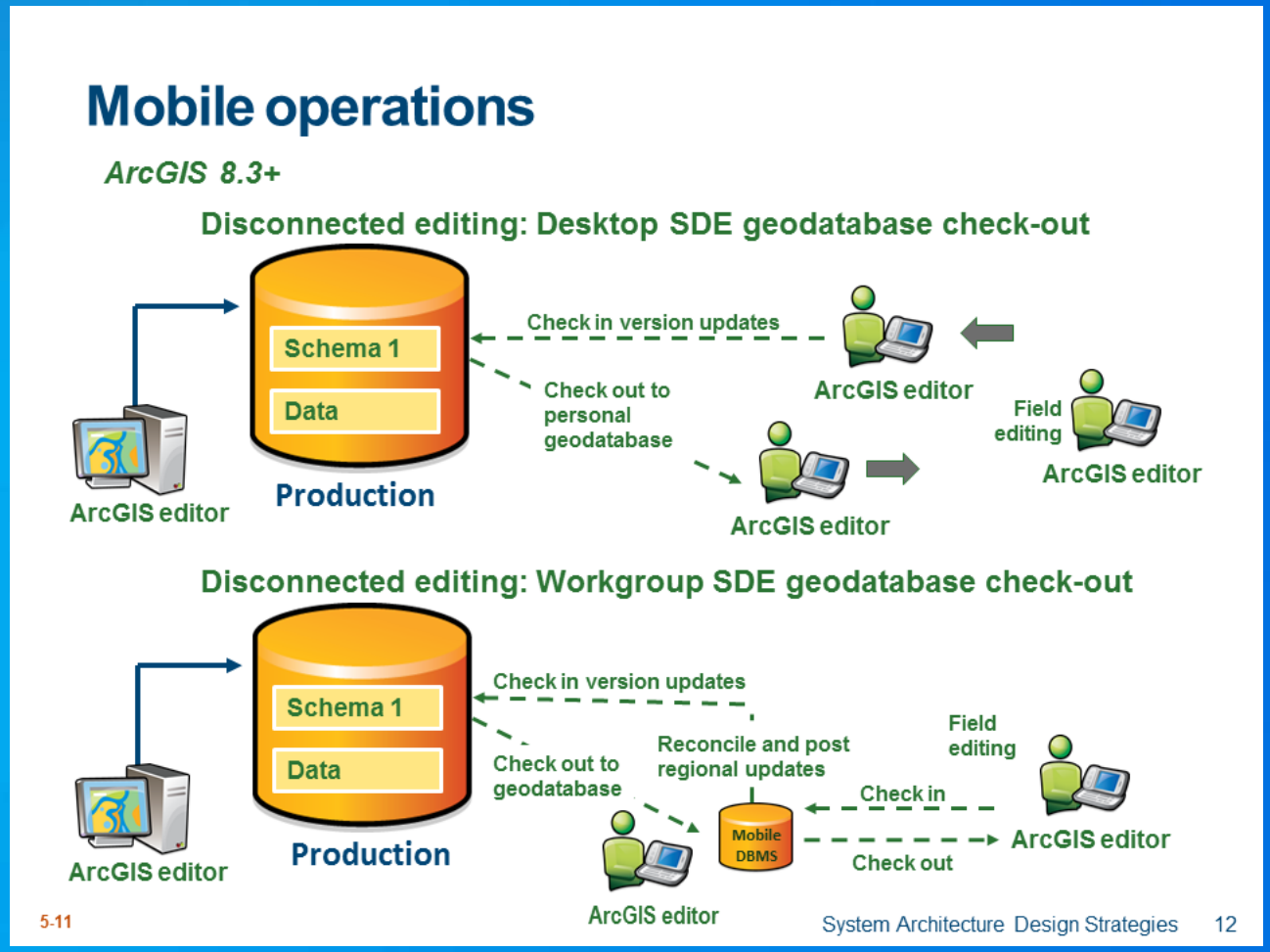
Geodatabase Replication Use Cases

- Mobile Operations
- Publication
- Distributed Operations
- Hierarchical / Federated



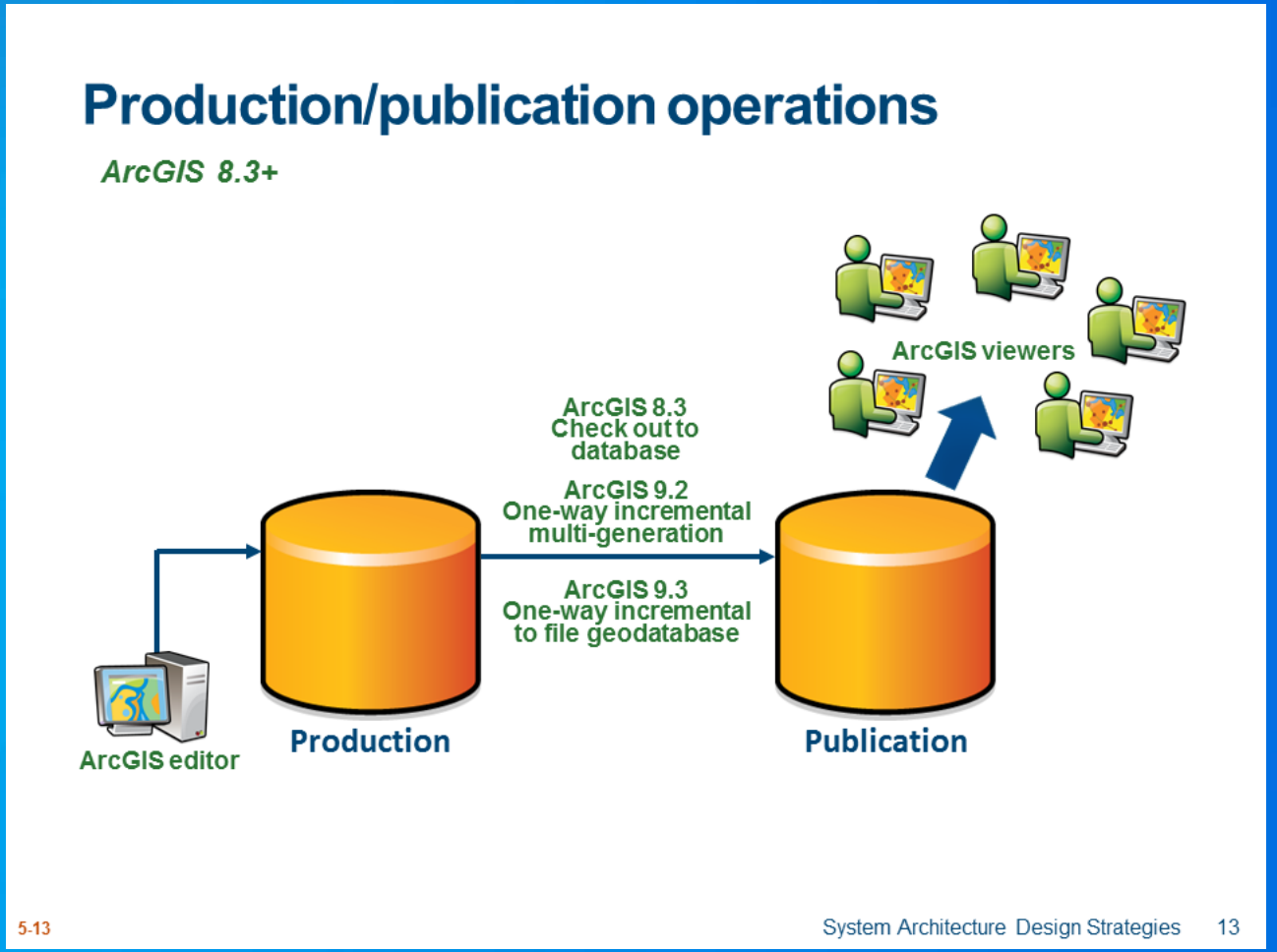
Geodatabase Replication – Mobile Operations

- Leverages check-out/check-in replication
- Enables field editing in a disconnected state



Geodatabase Replication – Publication Operations

- Leverages one-way replication
- Typical scenario is data publication for web operations
- May have practical limits with very complex data sets (i.e., hundreds of relationship classes, large geometric networks, etc.)

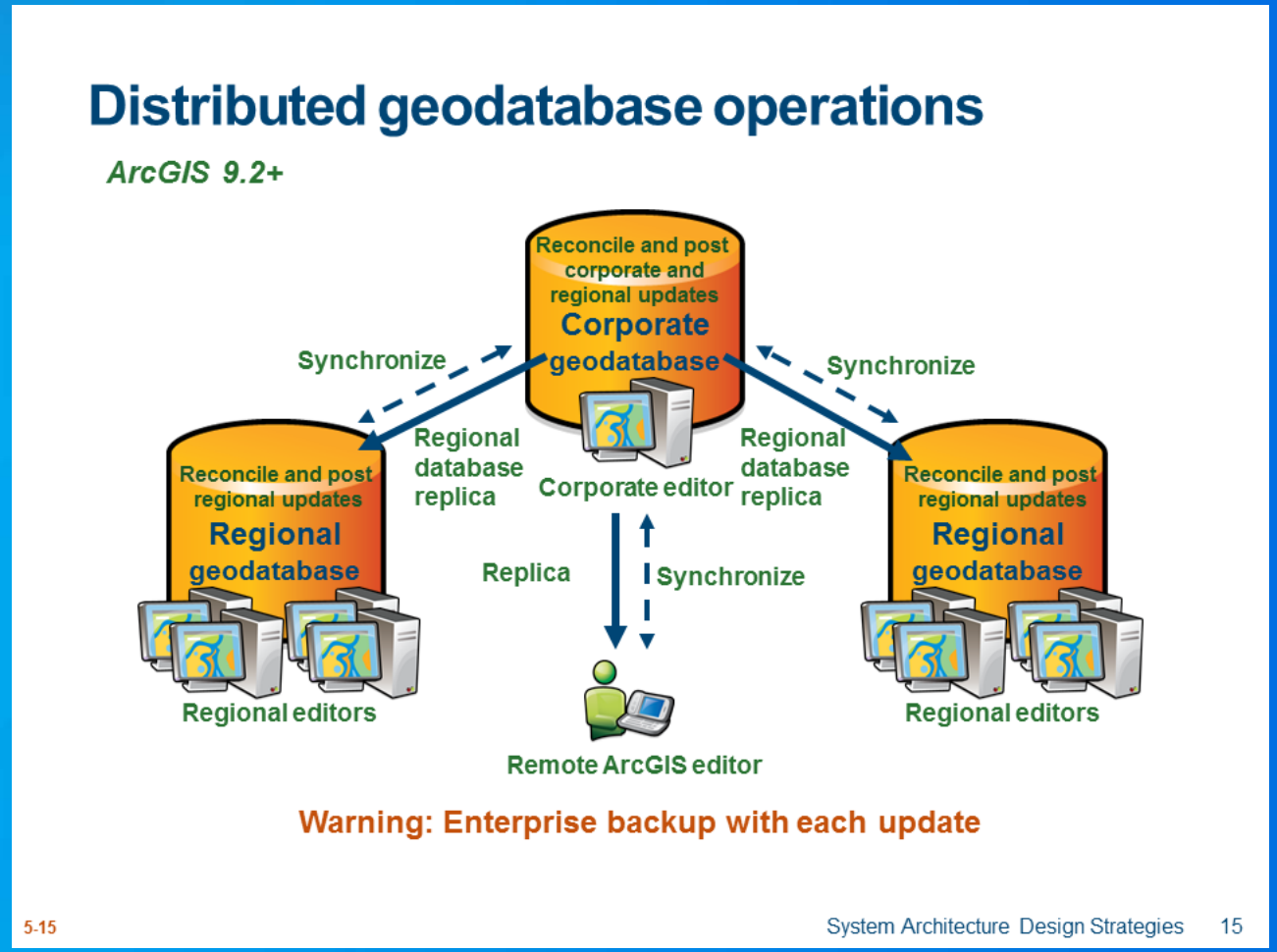


Publication Operations – Additional Considerations

- **Improves Performance / Scalability**
  - Publication is typically read-only non-versioned feature classes
  - Replication target can be FGDB for optimal performance
  - Hardware isolation of editors and viewers
- **Enhances Security**
  - Separation of duties – Data Maintenance and Data Publication
  - Allows for publication of subset of data and attributes
  - Supports data publication to the Cloud
- **Limitations / Cons**
  - Not a good solution if real-time data is required
  - Increases complexity and maintenance

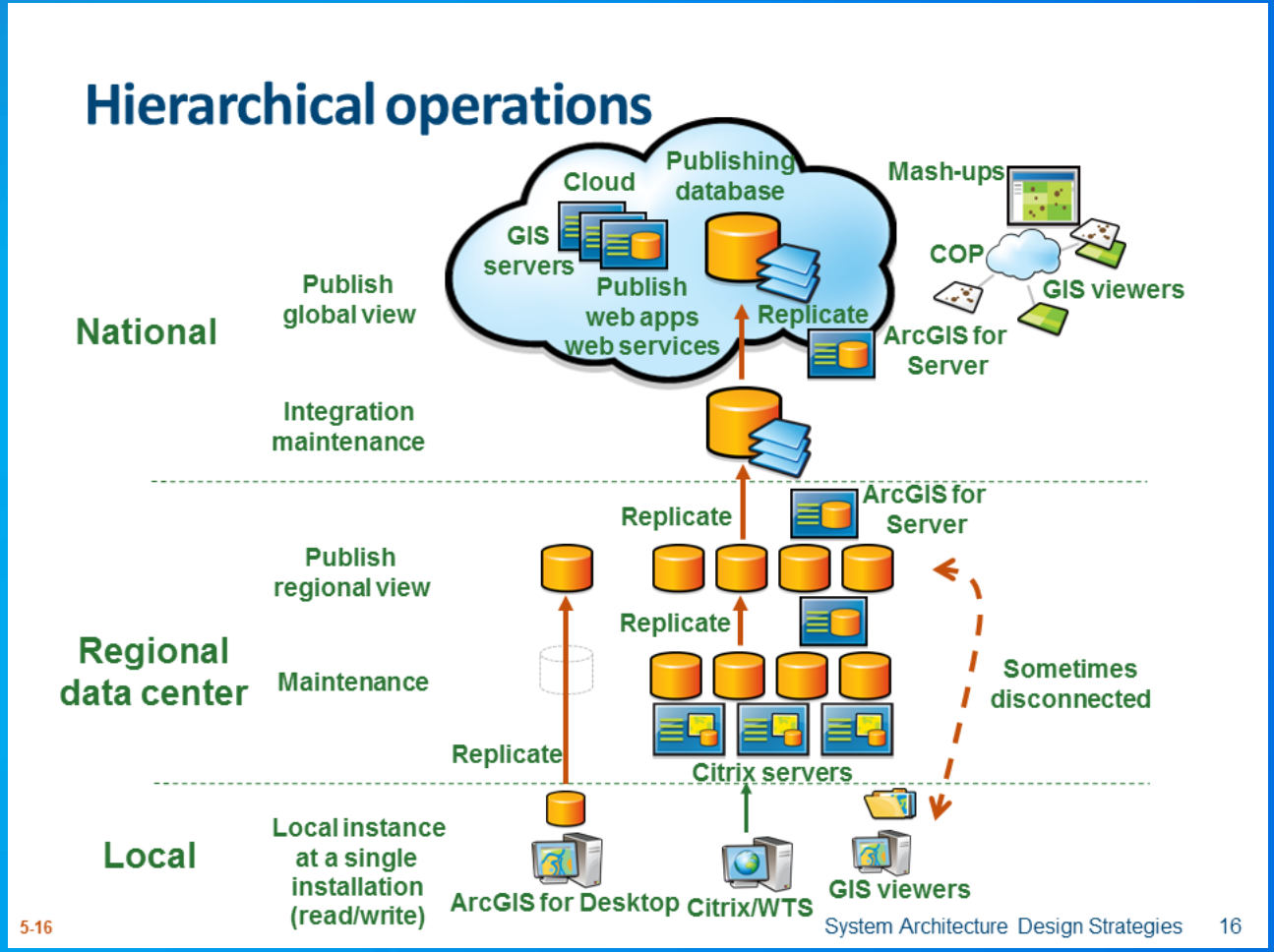
Geodatabase Replication – Distributed Operations

- Often based on two-way replication
- Typical scenario is data publication for distributed user sites
- Trend is moving away from distributed configurations towards centralized configurations



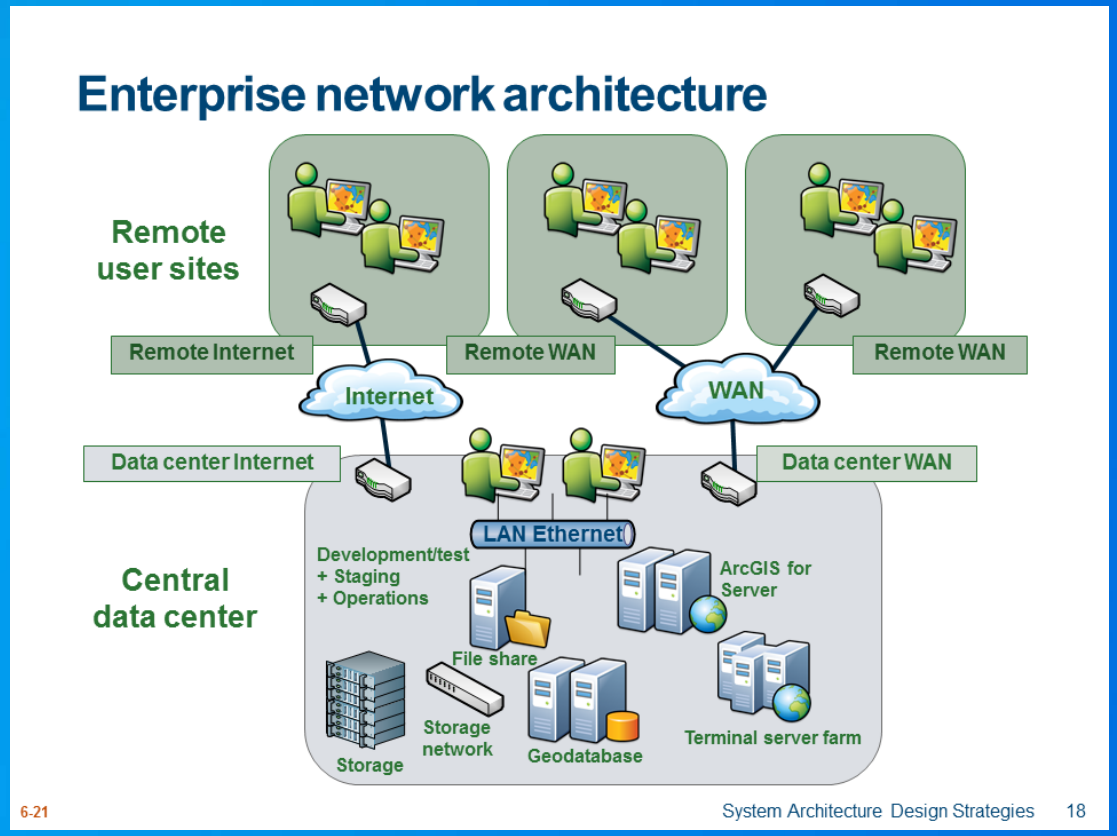
Geodatabase Replication – Hierarchical / Federated

- Leverages one-way replication
- Typical scenario: State agency level rolling-up data to national level



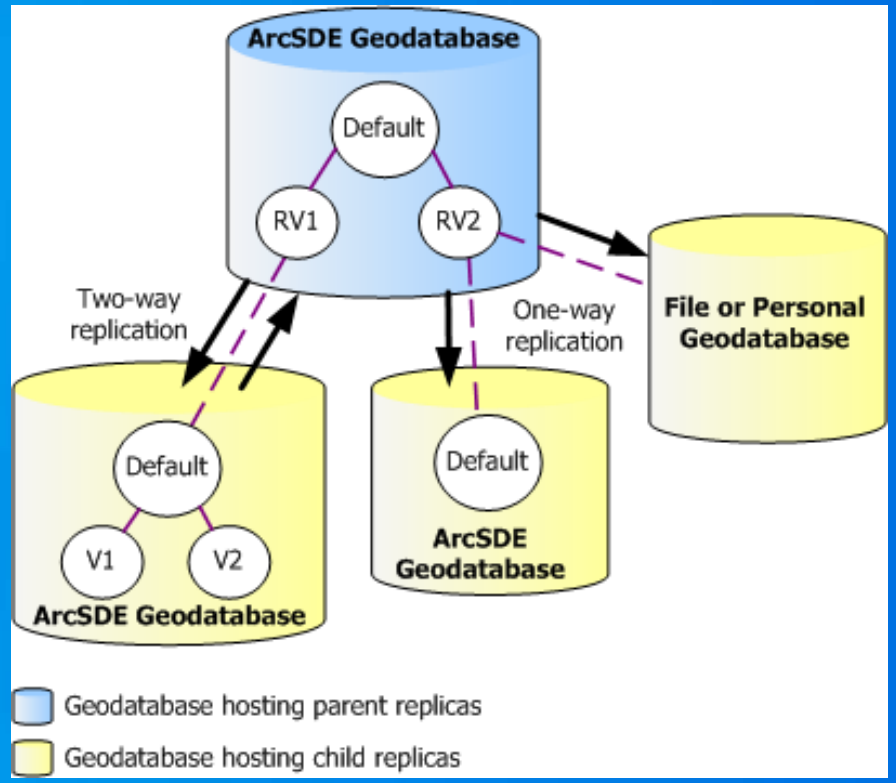
Geodatabase Replication – Network Strategies

- **Fast Network?**
  - Consider Connected Synchronization
    - Connects directly to the geodatabase
    - Low volume of changes
- **Internet or WAN?**
  - Consider Geodata Service Synchronization
    - Uses HTTP/S Web service
    - Low volume of changes
    - Supports check-out/check-in replication
- **Latent or No Network?**
  - Consider Disconnected Synchronization
    - Large volume of changes
    - Uses check-out/check-in replication
    - Uses delta XML file



Geodatabase Replication – Essential Tasks

- Identify data refresh requirements
  - Hourly? Nightly? Weekly?
  - Impacts architecture design decisions
- Understand replication constraints
  - Not all data types can be replicated (e.g., terrain data structures)
  - No imagery replication
  - Very complex data sets may be challenging (e.g., hundreds of relationship classes)
- Identify specific data to be replicated
  - May not need to replicate the entire database!
    - Options: Complete / By area / By attribute
- Select appropriate replication target
  - EGDB or FGDB



Questions & Answers

Contact info:

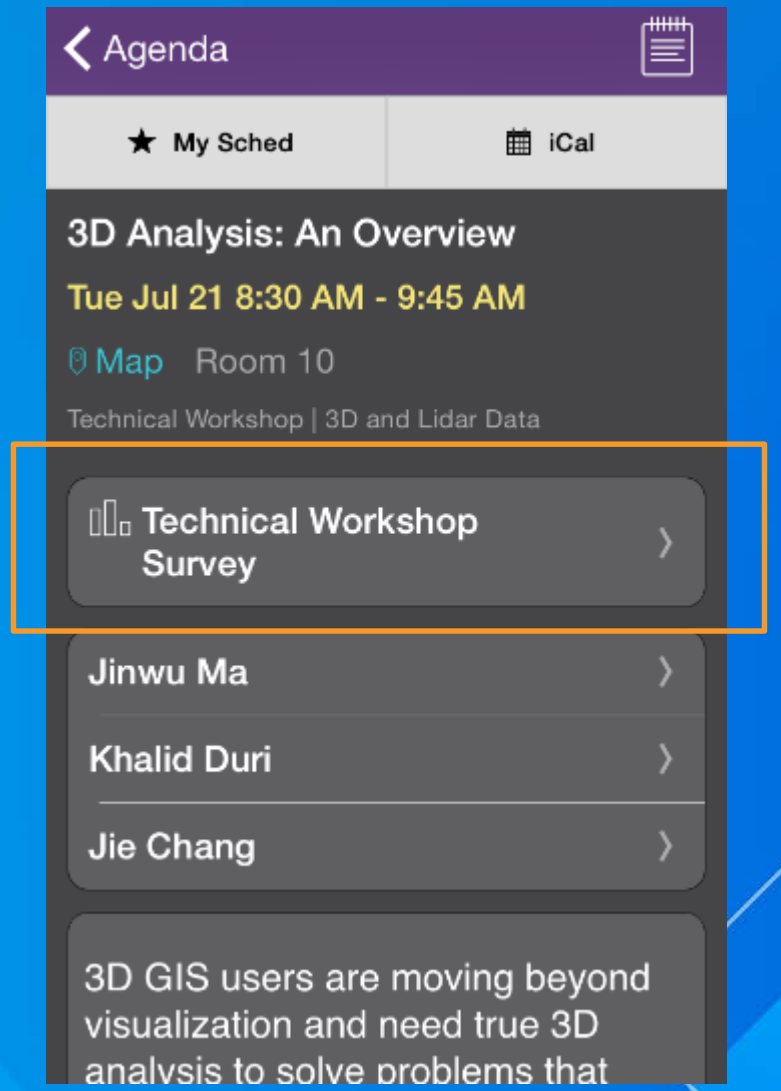
John Alsup
jalsup@esri.com

Jeff DeWeese
jdeweese@esri.com

Thank you…

Enterprise GIS: Database Planning

- Please fill out the session survey in your mobile app
- Select “Enterprise GIS: Database Planning” in the Mobile App
  - Use the Search Feature to quickly find this title
- Click “Technical Workshop Survey”
- Answer a few short questions and enter any comments



Understanding our world.