

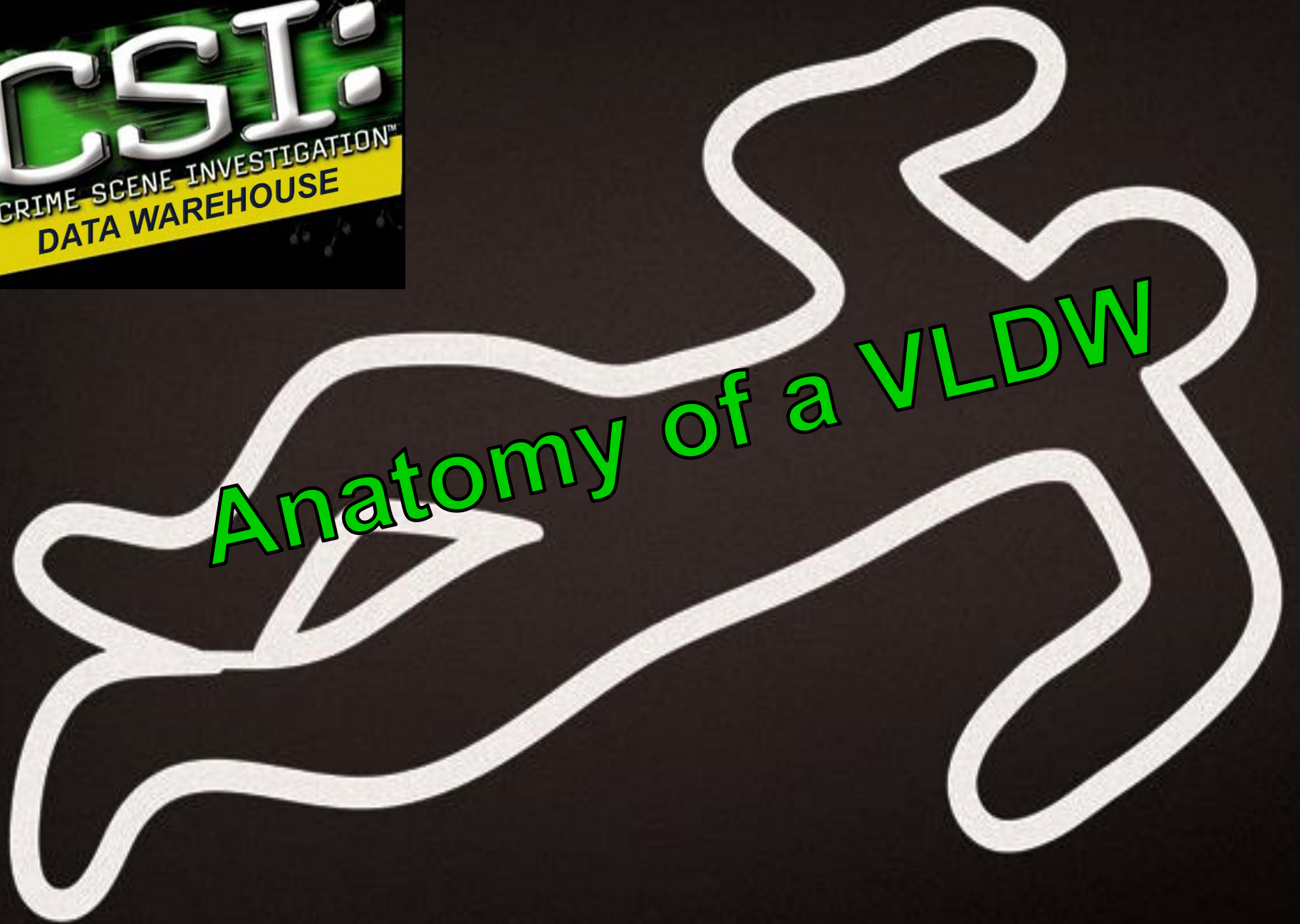
CSI:

CRIME SCENE INVESTIGATION™

DATA WAREHOUSE



Anatomy of a VLDW



CSI:DW

Anatomy of a VLDW

Dave Fackler
Business Intelligence Architect
davef@rollinghillsky.com



Agenda

- The Crime Scene
 - VA's DW and BI Landscape
 - DW Model and Metadata Infrastructure
- The Evidence
 - Database Infrastructure and Table Versioning
 - ETL and Data Distribution Architecture
 - Data Mart and Project Infrastructure
 - OLAP and Reporting Architecture
- Who Done it?
 - The Teams Behind it All

The Crime Scene



VA's DW and BI Landscape



- VA organized into VISN's
 - VISN = Veteran Integrated Systems Network (21 total)
 - Geographically close medical centers and clinics
- Each VISN manages 1 or more VistA systems
 - VistA = Electronic Health Record system for VHA
 - Clinical functions, financial functions, administrative functions, etc.
 - Basically, ***THE*** OLTP system for the VHA
 - Each VistA system is independent of others (130 total)
 - Can exchange data, staff can view data from multiple systems
 - VistA based on MUMPS and Intersystems Caché

VA's DW and BI Landscape



- Data Warehouse environment
 - Uses a “tiered” architecture
 - Corporate Data Warehouse (CDW) includes all national data (3+)
 - Regional Data Warehouse (RDW) includes regional data (4)
 - VISN Data Warehouse (VDW) includes VISN data (21)
 - Users connect to environment with scope of data they need
 - Same data model, same surrogate keys, same data
- BI (reporting) environment
 - Lots of tools (SSRS, Excel/PowerPivot, Pyramid Analytics)
 - Lots of teams developing data marts, cubes, reports
 - Wide range of what we call “projects” based on the DW

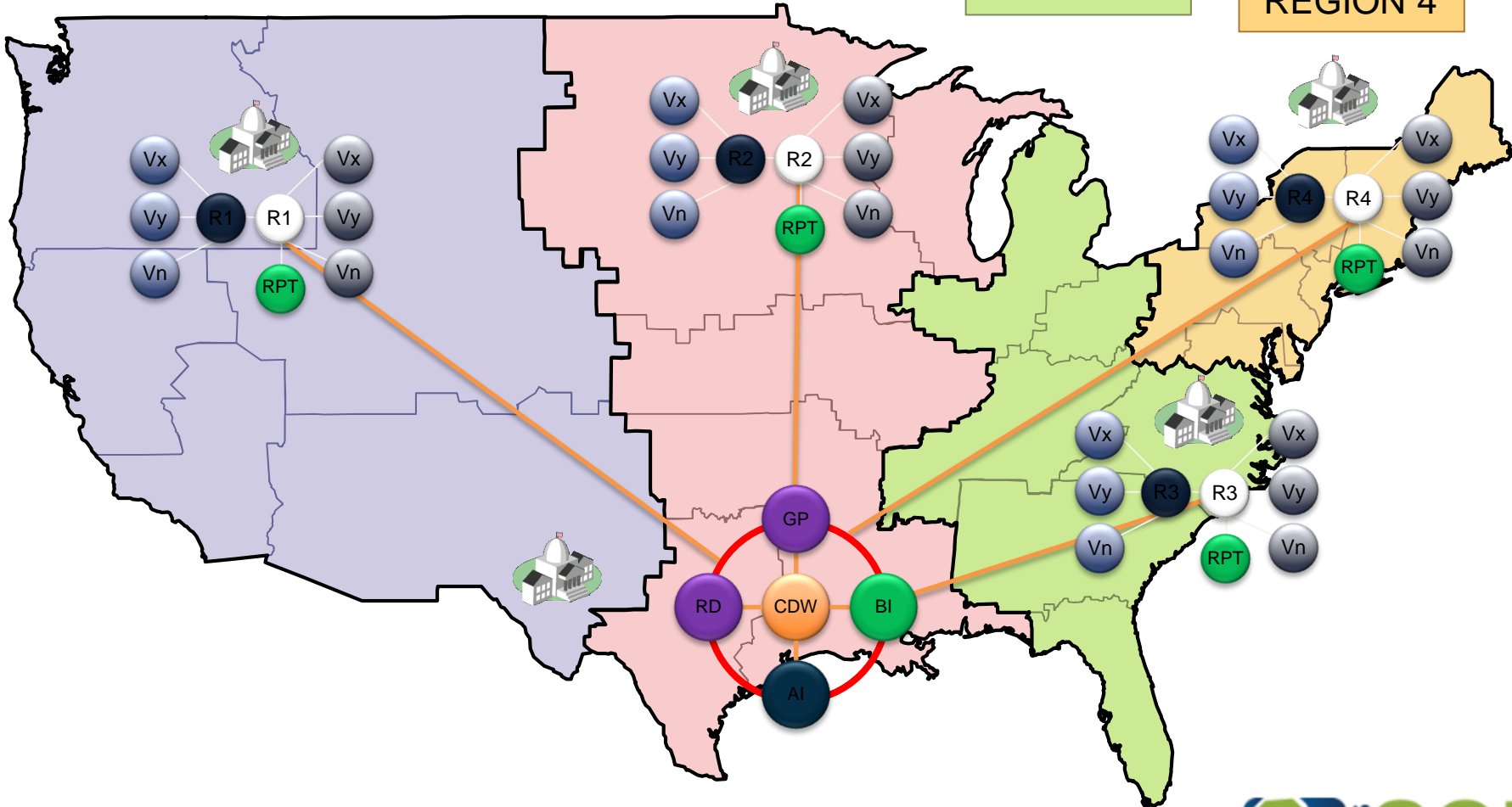


REGION 1

REGION 2

REGION 3

REGION 4



VA's DW and BI Landscape



- Explosive growth since 2009
 - Overall data volumes have surpassed a trillion rows
 - Over 4,700+ users authorized for CDW/RDW access
 - Over 30,000 users authorized for VDW access
- But continuing to grow (and grow and grow)
 - Continuously adding more data domains
 - Scaling up and out
 - Adding more users every month
 - Supporting more reporting solutions daily

Data Warehouse Model



- Star/snowflake schema
 - Dimension tables (274 dimension tables and growing)
 - Fact tables (25 domains, 175 fact tables and growing)
 - Every table has surrogate key (even fact tables)
 - Every table has consistent ETL auditing columns
 - Every table has business keys with standard names
 - Every table follows standards for naming, data types, etc.
- All users access data view two sets of views
 - First set filters out deleted records (all deletes are “soft”)
 - Second set includes deleted records

Metadata Infrastructure



- Data modeling team manages metadata database
 - Includes all table/view information
 - Published from ER/Studio via custom code and scripts
 - Also includes “mapping” information from VistA
 - VistA data model catalog imported via ETL process
 - So, data modeling team has that data available
- Metadata information made available to users
 - As report using SSRS
 - Shows metadata for tables/views plus mapping information
 - Also shows E/R diagrams published from ER/Studio
 - As data via views in the DW data model

The Evidence



Database Infrastructure



- Data model split into multiple databases
 - One database for dimension tables
 - Several databases for fact tables
 - Mostly based on legacy database management decisions
 - Databases consistent in CDW and RDW environments
 - In VDW environments, a single database is used
- All user access views kept in two databases
 - One filters deleted records, the other does not
 - Views use three-part names and database chaining

Table Versioning



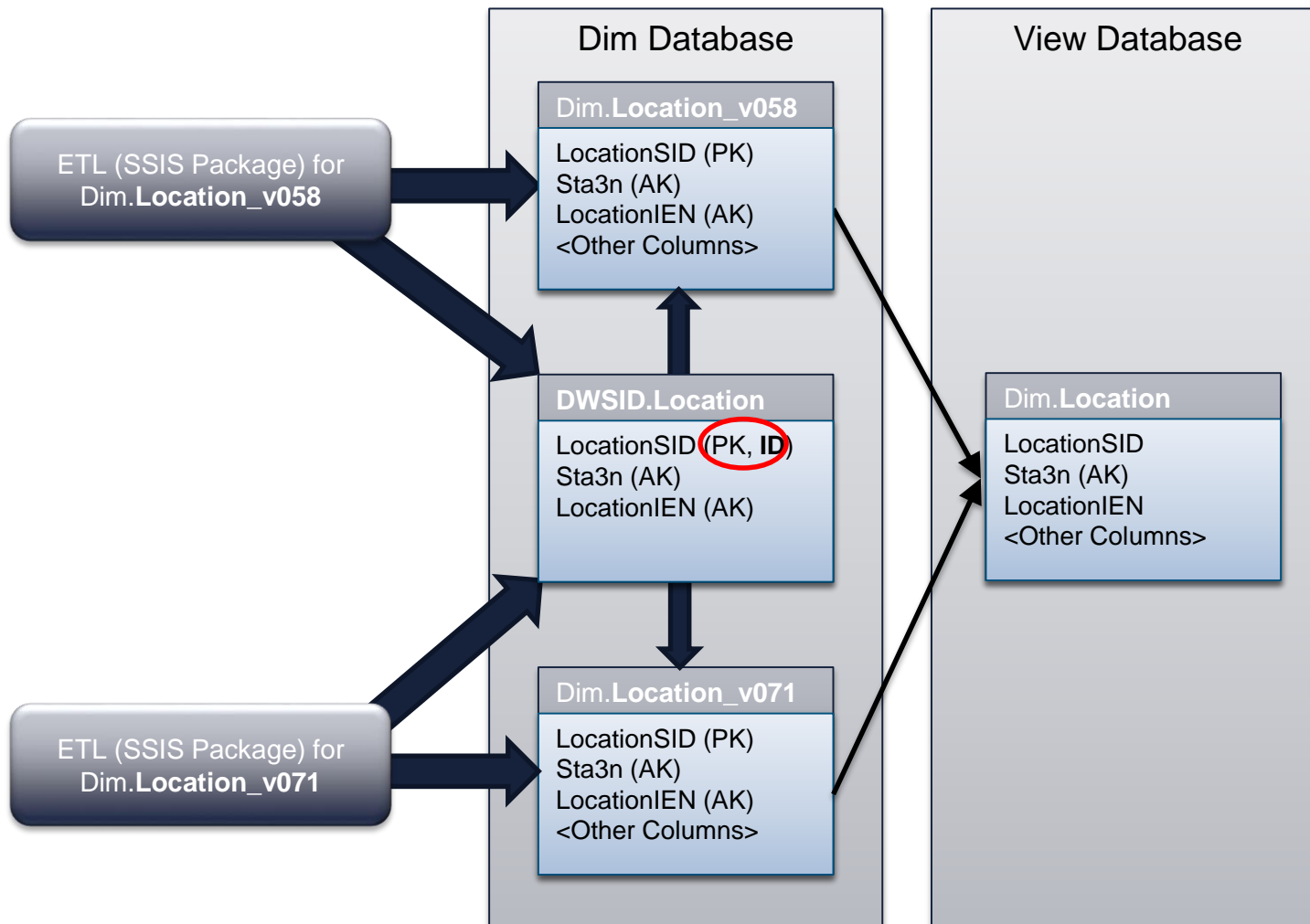
- Schema changes were hard to make early on
 - How do you manage adding a new column to a table with 2B rows? Particularly across 28+ environments?
 - How can you make schema changes when users will not accept table “down time”? Reports must run!!
- Decided on table versioning strategy
 - Every table is named with a version number at the end
 - User access views drop version numbers from names
 - Schema changes become new versions
 - User access views can then be changed when ready

Table Versioning



- Surrogate keys assigned using special table
 - Surrogate key implemented as identity() column
 - Table has business keys but nothing else
- Versioned tables get surrogate keys from there
 - Ensures business keys assigned the same surrogate key
 - Allows for moving from one version to the next
- As new table version is adopted, old one dropped
- System developed for orchestrating table development and deployment

Table Versioning



ETL Process



- Data pulled from VistA systems on a NRT basis
 - Data landed in a special set of tables (FDW)
 - Data is not modeled except in special circumstances
 - Deletes in VistA are marked as “soft” deletes
- ETL process then updates RDW environments
 - Runs on a nightly basis
 - Uses incremental extracts from FDW, loads into RDW
- Data Distribution System then updates CDW, VDW
 - Batches of data loaded into RDW are BCP'd to flat files
 - Files are then copied and loaded on other systems

ETL Process (Step 1)



- VistA systems use “journaling” for disaster recovery
 - Allow for “mirrors” of production systems, typically kept up-to-date with 5-30 seconds latency (asynchronous)
 - DW team leverages this and “mirrors” all 130 systems
- DW team developed “Journal Reader” application
 - Monitors journals going to mirrors looking for transactions
 - If transaction involves data entity we care about (based on metadata information), copies transaction to SQL Server
 - Feeds data into Feeder Data Warehouse (FDW) tables
 - Very simple transformations (dates, for example)

ETL Process (Step 2)



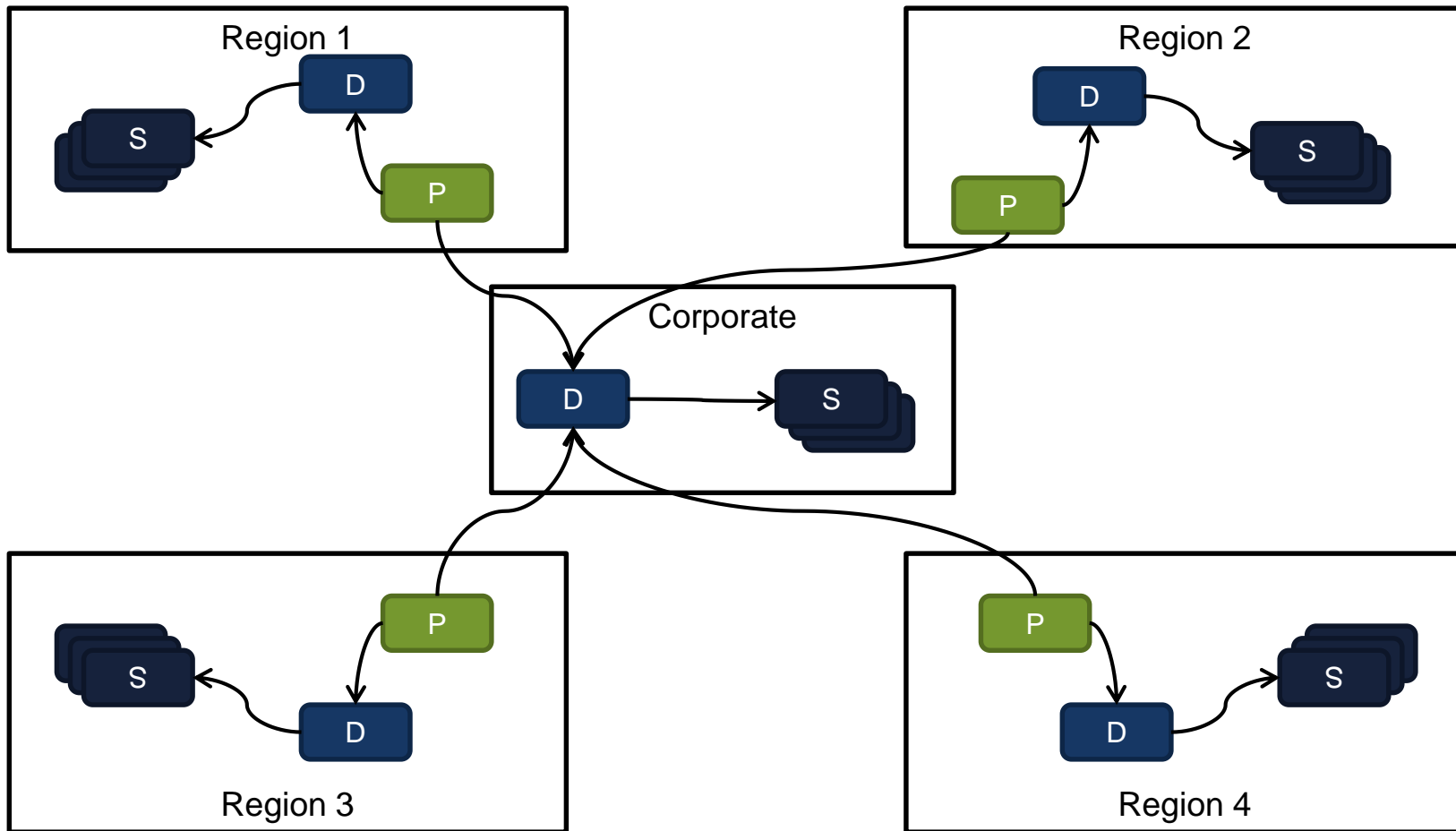
- ETL process then runs nightly
 - SSIS packages
 - One per versioned table
 - Master package dynamically orchestrates others (metadata)
 - Each package
 - Uses standard auditing logic and auditing table structures
 - Incrementally extracts from FDW table
 - Determines inserts versus updates
 - For smaller tables (10M and lower), just uses MERGE
 - For larger tables, uses INSERT (fast load) and UPDATE
 - Uses inferred member logic as needed
 - Uses special surrogate key tables for surrogate key generation

Data Distribution System



- Developed to handle loading CDW and VDW
 - Initially used linked servers back to RDW environments
 - CDW environments connected to each RDW environment
 - VDW environments connected to local RDW environment
 - Loading via MERGE commands via linked servers
 - Did not like linked server dependencies
 - Hard to troubleshoot and reload something
- DDS uses publisher/distributor/subscriber model
 - RDW environments “publish” flat-files of batches of data
 - Distributors move flat files from data center to data center
 - Subscribers load data from flat files

Data Distribution System



Data Mart and Project Infrastructure



- Lots of different teams want to report off DW data
 - Often too much data for what they want
 - Sometimes not structured how they want
 - Classic data warehouse versus data mart scenario
- Implemented “project” model and infrastructure
 - Teams now request “projects” which can include
 - SQL database(s), ETL server access and proxy
 - SSAS databases(s)
 - SharePoint (SSRS) site, reporting account
 - TFS support
 - Now supporting over 500 different projects

OLAP and Reporting Infrastructure



- Moving toward SSAS “farms” for projects
 - Processing servers and query servers
 - Developing “sync” framework and application now
- SharePoint with SSRS in integrated mode
 - Several different SharePoint farms for different groups
 - SSRS used for a lot of projects and a lot of reports
 - SSRS usage logs consolidated and made available
- Pyramid Analytics
 - Newest reporting tool implemented to replace ProClarity
 - Seeing very quick adoption (8,000+ users after 6 months)

Who Done It?



The Teams Behind it All



- Business Intelligence Service Line
 - Created to manage DW/BI infrastructure, solution, tools
 - Manager servers, data flow, “projects” workflow, etc.
 - Divided into “core” teams and “regional reporting” teams
- Core Teams
 - Architect Team (data modelers) 5
 - DBA Team (manage all aspects of DB servers) 7
 - VistA Extract Team (develop/manage extracts) 4
 - ETL Team (develop/manage all ETL procedures) 7
 - SSAS Team (manage SSAS infrastructure) 3

The Teams Behind it All



- Regional reporting teams
 - Four teams, one for each VA region
 - Develop solutions focused on healthcare operations
 - Morning reports, hospital operations, pharmacy costs, etc.
 - Partner with various business groups at various levels throughout their regions
 - Each team has 6-10 resources
- Ancillary resources
 - Training coordinator, technical writer, geospatial team
 - SA Team (responsible for hardware, OS of all servers)

Questions?



CSI:

CRIME SCENE INVESTIGATION™

DATA WAREHOUSE