



KAISER PERMANENTE NATIONAL RESEARCH DATABASE

FUNDED IN PART BY THE NATIONAL LIBRARY OF MEDICINE
AND CHONDROGENE

Joseph Terdiman, MD, PhD, Director of IT, DOR
Larry Walter, MS, Director of Strategic Programming, DOR
Morris Collen, MD, Director Emeritus, DOR



KP RESEARCH DATABASE OVERVIEW

STATISTICS

RATIONALE

- WHY DOES KP NEED A NATIONAL RESEARCH DATABASE (NRDB)?

STRATEGIC OBJECTIVES

DESIGN

- DATA MODEL
- DATABASE ATTRIBUTES
- ARCHITECTURE
- CATEGORIES OF DATA
- DATA SUBJECT AREAS

DEVELOPMENT PHASES

KAISER PERMANENTE STATISTICS

- Founded 1945
- Start of EDC 1968
- Members 8.5 million
- Employees 152,814
- Physicians 12,879
- 8 Regions (NCal, SCal, NW, CO, GA, HI, MAS, OH)
- 32 Hospitals
- 413 Medical office buildings

WHY DO WE NEED A KP RESEARCH DATABASE?

RESEARCHERS PERSPECTIVE

- CURRENT TECHNOLOGY
 - MULTIPLE HETEROGENEOUS DATA SOURCES
 - POOR DOCUMENTATION
 - INDIVIDUAL DATA SILOS
 - 50-90% OF PROGRAMMING EFFORT TO CLEAN AND VALIDATE DATA
 - DUPLICATION OF EFFORT IN EACH REGION AND BETWEEN REGIONS
- RESEARCH DATABASE TECHNOLOGY
 - SINGLE DATA SOURCE – RELATIONAL DATABASE
 - CONSISTENT DATA DEFINITIONS
 - DOCUMENTATION THROUGH METADATA DATABASE

GOVERNMENT PERSPECTIVE

- UNIQUE RESOURCE
- RESEARCH AREAS OF INTEREST TO GOVERNMENT AGENCIES
- MODEL FOR NATIONAL HEALTH INFORMATION NETWORK
- PROMOTE UMLS CONCEPTS



STRATEGIC OBJECTIVES

- Create a unique research data warehouse for KP researchers
- Create a single source for most KP data used for research
- Support many different types of research
- Meet HIPAA and KP security requirements
- Facilitate collaborations between researchers in different KP regions and in other institutions
- Be consistent with U.S. health care data goals



UNIQUE RESEARCH DATA RESOURCE

- Records for over 8.5 million current KP members (3.3 million members NCAL)
- Records for over 20 million past KP members
- Data sources from 8 geographical regions
- Maximum time span of data collected – 40 years
- Full range of clinical data

TYPES OF RESEARCH

- Epidemiological studies
- Prognosis and survival studies
- Health care effectiveness research
- Etiology and prevention research
- Informatics research
- Clinical trials
- Syndromic surveillance
- Bioterrorism surveillance
- Adverse event monitoring
- Postmarketing drug surveillance
- Disease registries
- Genetic and genomic studies
- Geographical mapping of morbidity and mortality rates



PILOT RESEARCH DATABASE IN NORTHERN CALIFORNIA (NCRDB)

- Records for over 3.3 million current KP members
- Records for over 10 million past KP members
- One geographic region with a population having socioeconomic and ethnic diversity

NCRDB ATTRIBUTES

- Relational database (Oracle)
- Multi-terabyte storage requirement (>10 TB/yr.)
- Single virtual database
- Multiple physically distributed databases (federated design)
- Preserve data indefinitely
- Enable database security features

REQUIREMENTS OF NCRDB

DATABASE DESIGN NEEDS

- Research data warehouse separate from patient care systems
- Receive updates from patient care systems (at least daily)
- Aggregate legacy and current clinical information from multiple KP data sources
- Use standards and naming conventions for research database that may differ from patient care systems (e.g., UMLS concepts)
- Optimize database schema for research retrievals (e.g., longitudinal and cross-sectional studies)
- Use partitioning and indexing to improve performance
- Add tables and variables as needed by researchers (e.g., genomic data, survey data)
- Create metadata database

ACCESS NEEDS

- Separate research queries and applications from patient care and business functions
- Support multiple simultaneous queries
- Create multiple disease registries (e.g., via data marts)
- IRB approval required for access to NCRDB

DATA SOURCES

- Patient care data
 - Legacy systems
 - Current systems (EpicCare -- KP HealthConnect)
 - Ancillary data
- Research data
 - Project-specific data (e.g., survey instruments)
 - Legacy research databases
 - Non-KP member data (e.g., pedigrees)
 - Genomic data
 - Validation data
- Public use data sets (e.g., census, birth, mortality)

SECURITY

- Security levels (user, group, table, column, row)
- Confidentiality of retrieved data
 - Fully identified
 - Limited data set (identifiers removed, dates OK)
 - Anonymized data set
 - Protected Health Information (PHI) demographics in separate table
 - Internal ID number for each patient
 - Increment dates by random number
- Audit trail

KP HEALTHCONNECT INTERFACE

■ EpicCare databases

□ Chronicles (Cache) -- OLTP

- AIX platform
- Mumps database

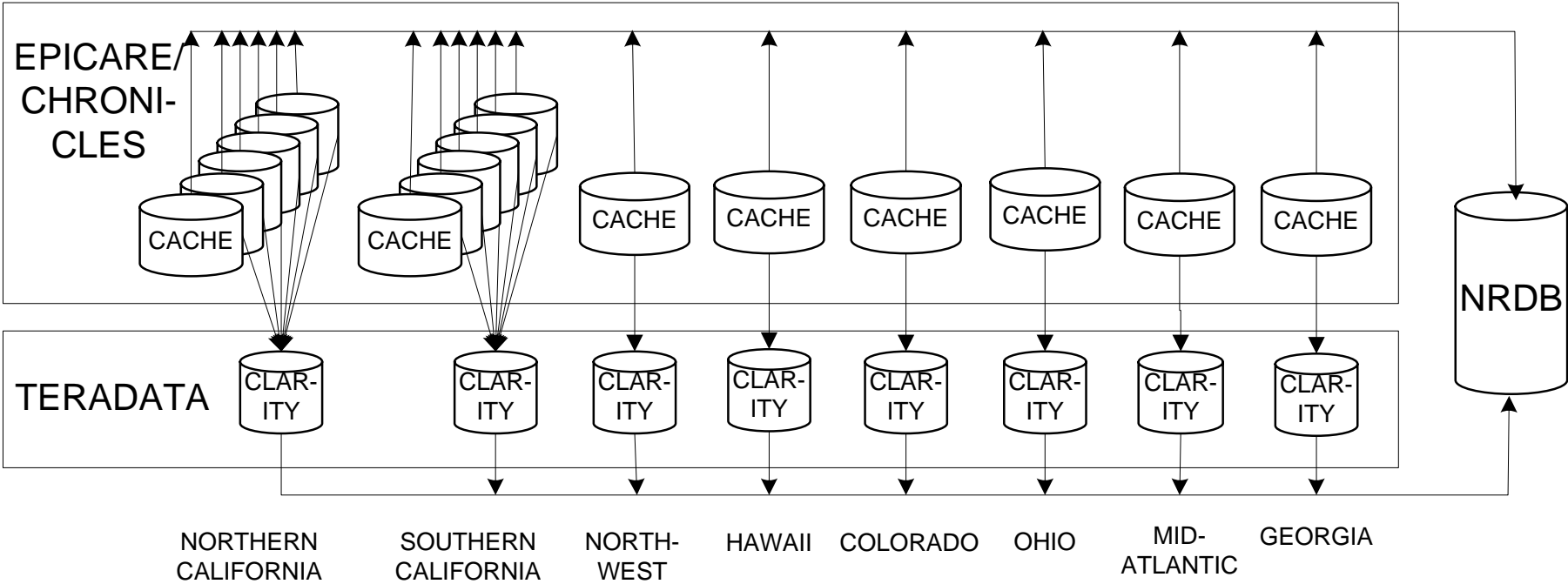
□ Clarity -- OLAP

- Teradata platform
- Relational database

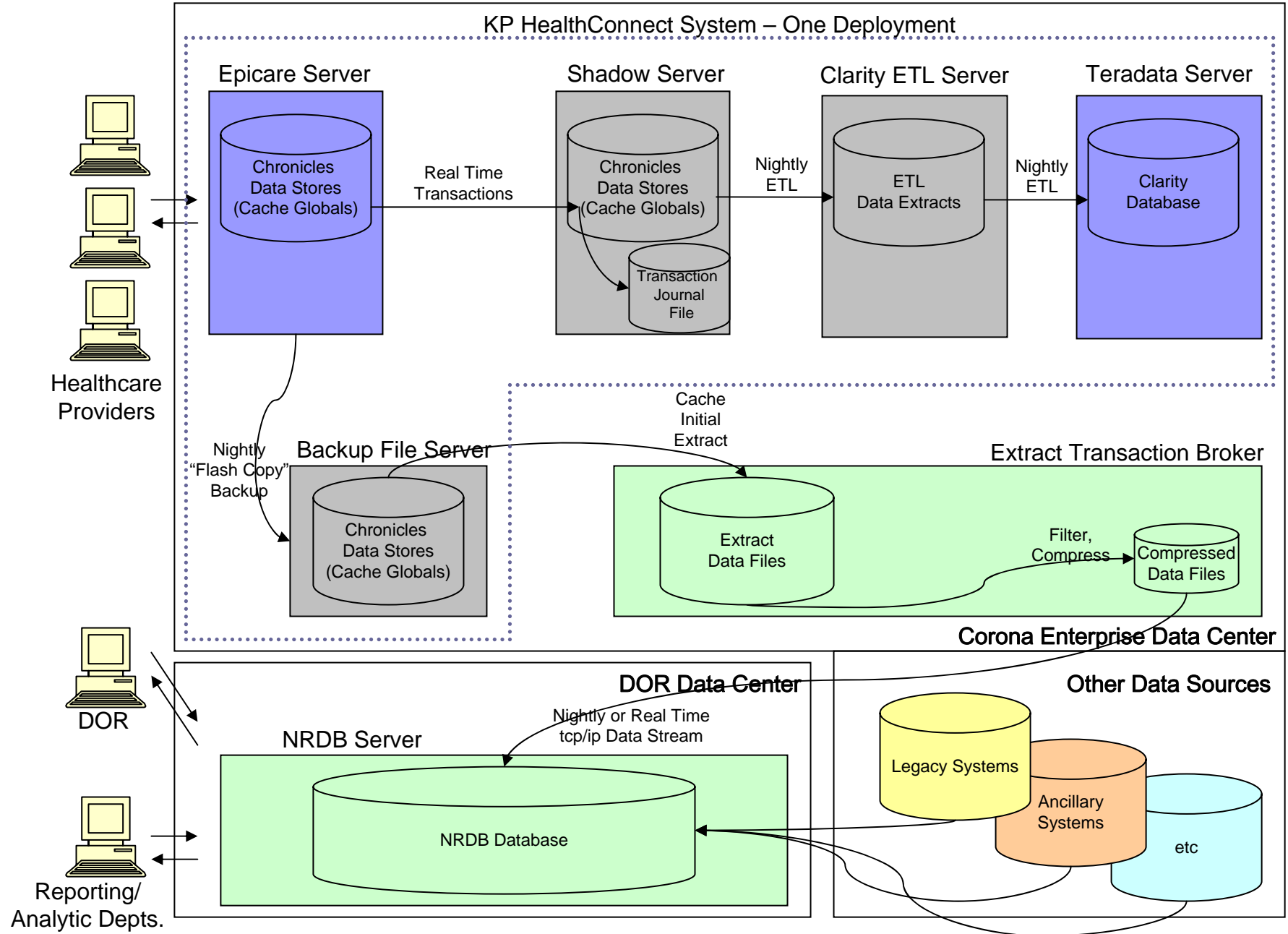
■ Problems

- <50% of Chronicles variables transferred to Clarity
- ETL process inefficient (may take 24 hours)
- Errors in data mapping Cache → Clarity

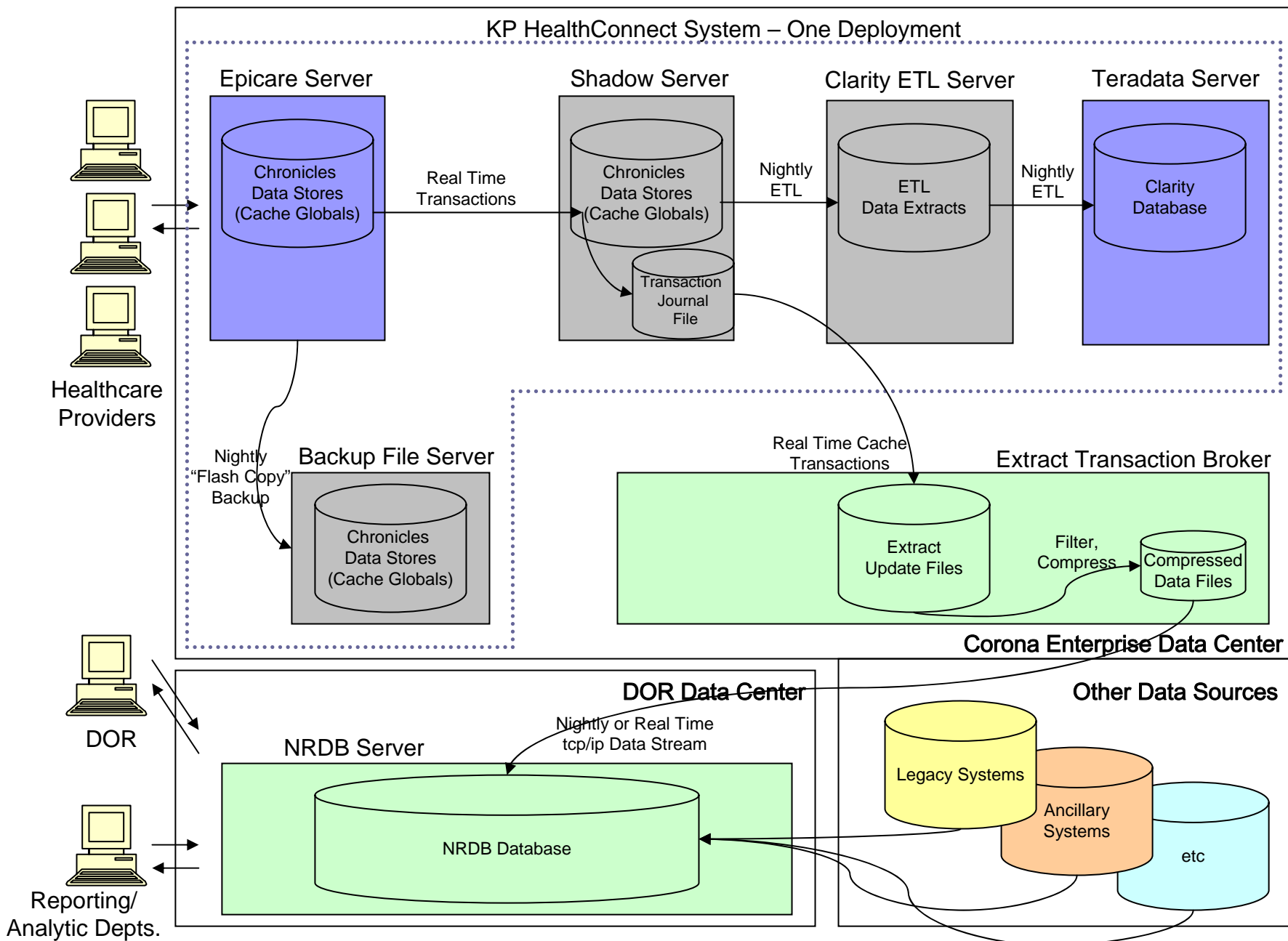
RELATIONSHIP BETWEEN NRDB AND EPICCARE AND EPICCARE



CACHE/CHRONICLES EXTRACT – INITIAL ETL



CACHE/CHRONICLES EXTRACT -- UPDATES



PLANNING PHASE OF NCRDB

- Identify sources of patient care data for NCRDB
 - Legacy data systems
 - KP HealthConnect
 - Ancillary systems
- Identify external data sources
 - Public use data sets (census, birth, mortality)
- Determine data elements to be stored in NCRDB from each source
- Determine data requirements of researchers
- Develop design specifications for each stage of NCRDB
 - Release 1 – mirror legacy, HealthConnect and ancillary data sources
 - Release 1.1 – enhance Release 1 with additional data and improved performance
 - Release 2 – create new schema with consistent variable definitions, coding, business logic and security

IMPLEMENTATION PHASE OF NCRDB

■ Release 1

- Consolidate legacy, HealthConnect and ancillary data into Oracle database, using legacy schema, variable names and formats
- Design ETL processes (one-time and daily)
- Configure data retrieval processes

■ Release 1.1

- Enhance database by loading archival legacy data into database
- Create table indexes and partitioning for improved performance

■ Release 2

- Design database schema for NCRDB using standards-based naming conventions, descriptors and coding (e.g., UMLS)
- Define metadata for each variable
- Design data marts to support specialized views of the data and disease registries
- Enhance interface to HealthConnect
- Configure data retrieval, text searching and reporting tools



MAJOR CHALLENGES

- Data update and validation strategy
- Realistic functional specifications
- Standardization of disparate data
- Scalability & performance
- Security and compliance
- Programmer re-training
- Resources