

# CDCR EA Data Warehouse / Business Intelligence / Reporting Strategy Overview



February 12, 2010



# Agenda

1. Purpose - Present a high-level Data Warehouse (DW) / Business Intelligence (BI) / Reporting Strategy
  2. Define Terms
  3. Key References
  4. Project Scope for Strategy Overview
  5. Concerns
  6. Context Diagram
  7. Mappings by Programs, Projects & Initiatives (As-Is)
  8. DW Maturity / Readiness Matrix
  9. Strategy
  10. 3 Phased Targets
  11. Dependencies
  12. EDW Best Practices
- Appendices



## Define Terms

**Data Warehouse (DW)** – Is a repository of an organization's electronically stored data. DWs are designed to facilitate reporting and analysis. DW provides a single 360 degree view of the business and a platform for business intelligence tasks ranging from predictive analysis to near real-time strategic and tactical decision support.

Three keys to DW to get correct are 1) the hardware configuration, 2) the data model, and 3) the data loading process. DW supports research by providing data organized and optimized for results.

**Business Intelligence (BI)** – Refers to skills, processes, technologies, applications and practices used to support decision making. BI technologies provide historical, current, and predictive views of business operations. Common functions of Business Intelligence technologies are reporting, analytics, data mining, business performance management, predictive analytics. BI supports research by providing triggers, outcomes, data mining and trending, and makes results available to decision makers.

**Reporting** – Ability to provide strategic appropriate authorized relevant accurate reports to enable design making, including portal services to control access and provide the intelligence to the correct users.



## Define Terms – cont.

### **DW Workload:**

**Research** – Ability to mine, analyze, report and provide analytics on available enterprise data to the executive staff concerning the impact of potential new policies

**Compliance** – Alignment to reporting bodies (legislature, courts, governor's office)

**Operations** – Day to day IT support to programs by mission critical application

**Outcomes** – Office of Research (OR) provides outcomes based on their analysis and analytics; also specified as outcomes of CDCR strategic plan goals except clinical focused outcomes

**Performance** – COMPSTAT provides performance based on institution reported data except clinical focused performance



## Define Terms – cont.

### Technology Stack:

**Governance** – The governance process across the different Technology Stack components

**Portal** – An Enterprise Portal is a Web software infrastructure providing access to, and interaction with, relevant information assets (information/content, applications and business processes), knowledge assets and human assets, by select targeted audiences, delivered in a highly personalized manner. There are two types of Portals: Vertical portals focus on accessing specific applications or business functions; Horizontal portals seek to integrate and aggregate information from multiple cross enterprise applications, as well as specific line-of-business tools and applications.

**Reporting** – Ad hoc, canned and variable reports available on-line and batch, hardcopy or electronic

**Business Intelligence (BI)** – described by the Technology Bricks as: Business intelligence (BI) refers to technologies, applications and practices for the collection, integration, analysis, and presentation of business information and sometimes to the information itself. The purpose of business intelligence is to support better business decision making. Thus, BI is also described as a decision support system (DSS).

**Data Mart (DM)** – A data mart is a subset of an organizational data store, usually oriented to a specific purpose or major data subject, that may be distributed to support business needs. Data marts are analytical data stores designed to focus on specific business functions for a specific community within an organization.

**Data Warehouse (DW)** – Data Warehouse specific software tools that provide specific DW functions; described by the Technology Bricks as: A data warehouse is a central repository for all or significant parts of the data that an enterprise's various business systems collect. Typically, a data warehouse is housed on an enterprise server. Data from various online transaction processing (OLTP) applications and other sources is selectively extracted and organized on the data warehouse database for use by OLAP or data mining analytical applications and user queries. Data warehousing emphasizes the capture of data from diverse sources for useful analysis and access, but does not generally start from the point-of-view of the end user or knowledge worker who may need access to specialized, sometimes local databases. (This definition is specific to the Technology Stack, different from the higher level DW strategy definition in the previous slide.)

**Data** – Element level data that describes CDCR's business which are organized into tables, schemas, views, and databases held in what can be considered a data staging server environment

**Software** – Layered software supporting the operations of the hardware such as operating systems, tools, middleware

**Hardware** – Server level hardware, in this case specifically supporting the software, data, DW and BI components; may include database servers, application servers, reporting servers and web servers if needed

**Enterprise Identity Management (EIdM)** – broad administrative area that deals with identifying individuals in a system (such as a country, a network or an organization) and controlling the access to the resources in that system by placing restrictions on the established identities.



## Key References

1. EDAC – Enterprise Data Architecture Committee Charter  
[http://portal.cdcr.ca.gov/Admin/eis/pmo/ea/Team/Shared%20Documents/Enterprise%20Data%](http://portal.cdcr.ca.gov/Admin/eis/pmo/ea/Team/Shared%20Documents/Enterprise%20Data%20Architecture/EnterpriseDataArchitectureCharter.pdf)
2. FEA Data Reference Model Version 2.0 <http://www.whitehouse.gov/omb/asset.aspx?AssetId=561>
3. DAMA International Foundation <http://www.dama.org/i4a/pages/index.cfm?pageid=1>
4. A Call to Action for State Government: Guidance for Opening the Doors to State Data  
<http://www.nascio.org/publications/index.cfm#118>
5. Data Governance Part I: Data Governance [www.nascio.org/publications/documents/NASCIO-DataGovernance-Part1.pdf](http://www.nascio.org/publications/documents/NASCIO-DataGovernance-Part1.pdf) - 2009-02-06
6. Data Governance Part II: Maturity [www.nascio.org/publications/researchBriefs.cfm](http://www.nascio.org/publications/researchBriefs.cfm)
7. Data Governance Part III: Frameworks  
[www.nascio.org/publications/researchBriefsSubject2.cfm?category=14](http://www.nascio.org/publications/researchBriefsSubject2.cfm?category=14)
8. The purpose of Data.gov is to increase public access to high value, machine readable datasets generated by the Executive Branch of the Federal Government [www.data.gov](http://www.data.gov)
9. California Logical Model [www.cdcr.ca.gov/News/reentry\\_slideshow/logic\\_model.pdf](http://www.cdcr.ca.gov/News/reentry_slideshow/logic_model.pdf)
10. Little Hoover Commission Report: “A New Legacy System: Using Technology to Driver Performance...”  
<http://www.lhc.ca.gov/studies/193/report193.html>



## Key References – cont.

### **Respected CA State Sources**

1. Governor's Reorganization Plan
2. CDCR Strategic Plan
3. CDCR IT Strategic Plan
4. CDCR IT Capital Plan
5. COMPSTAT Charter
6. Office of Research (OR) Charter
7. SOMS Contract
8. BIS Contract
9. OCIO Statewide Data Strategy Report



# Key References – cont.

## Technology Stack Mapped to Key References and Respected Sources

Supports	Technology Stack	Key References (from page 6)											Respected CA State Sources (from page 7)								
		1	2	3	4	5	6	7	8	9	10	11	1	2	3	4	5	6	7	8	9
	Governance	Green					Green	Green	Green								Green	Green	Green	Green	
	Portal																		Green	Green	
	Reporting										Green			Green	Green		Green	Green	Green	Green	
	BI				Green									Green	Green		Green	Green	Green	Green	
	DW		Green		Green									Green	Green		Green	Green	Green	Green	
	DM		Green		Green									Green	Green		Green	Green	Green	Green	
	Data			Green	Green	Green				Green			Green				Green	Green	Green	Green	Green
	Software											Green	Green			Green			Green	Green	
	Hardware											Green	Green			Green			Green	Green	

**Legend:** Green reflects that the reference supports the given technology stack component



# Project Scope for Strategy Overview

## Scope

1. Define the lifecycle of a data warehouse specific implementation
2. Define data warehouse industry "best practices" principles and frameworks of thinking
3. Leverage current production data capabilities "as-is" (including OR & COMPSTAT)
4. Leverage the technology opportunities and strategic needs of approved projects
5. Align to Enterprise Data Architecture Committee Charter (EDAC) Principles\*
6. Produce the overall strategy for ongoing maturity
7. Produce a plan for the first target step mapped to the lifecycle
8. Effort limited to 70 hours planned for 10/05/09 – 11/30/09
9. Focus was on CDCR's "as-is" environment. CPHCS environment details will be incorporated into the next planning cycle as CPHCS falls into the future state architecture. CDCR and CPHCS future state EDW, BI, and reporting will be integrated.

## Out-of-Scope

- Budget/Funding not included
- Required PYs not included
- Due to the limited scope of this engagement, follow-on tasks may be needed

## Meetings were conducted with

- COMPSTAT
- EA
- EdFIRST
- OISB
- SOMS
- BIS
- CPHCS

\* See Appendix - C for EDAC alignment



## Concerns

18 specific concerns resulted during meetings with the various teams which can be summarized as follows\*:

- Data quality / data cleansing needed / common vocabulary needed\*\*
- Lack of governance – confusion on who has authority to make decisions & policies
- Insufficient performance metrics (fidelity)
- Misaligned reports (not supportive to current business needs)
- Tactical, ad hoc, silo'ed, reactive, inconsistent applications
- Technical performance needs improvement (bandwidth)
- Cost effectiveness (new Executive order)
- Economics of scale (support model)

\*\* Does not include CPHCS targeted data quality requirements

\* Details are included in Appendix - A



# Context Diagram (As-Is)



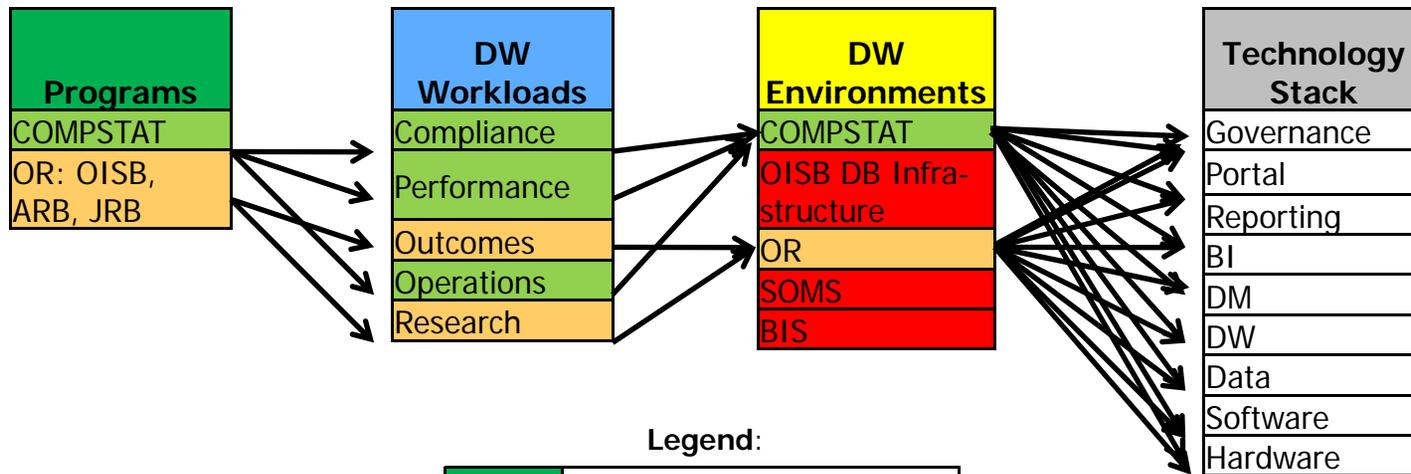
**Legend:**

Green	Programs, Projects, & Initiatives
Blue	DW Workloads
Yellow	DW / BI / Reporting Environments



# Mapping by Programs (As-Is)

Findings: Programs are silo'ed



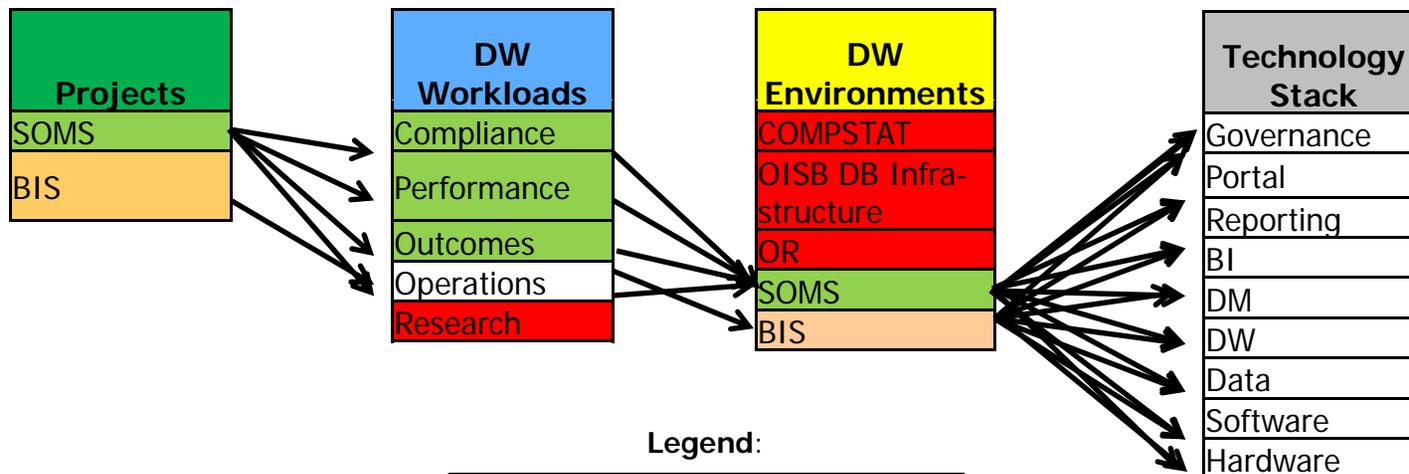
Legend:

Green	Programs
Blue	DW Workloads
Yellow	DW Environments
White	Technology Stack
Light Green	
Green	COMPSTAT Program
Tan	OR Program
Red	Gaps
White	Overlaps



# Mapping by Projects (As-Is) – cont.

Findings: Projects are silo'ed



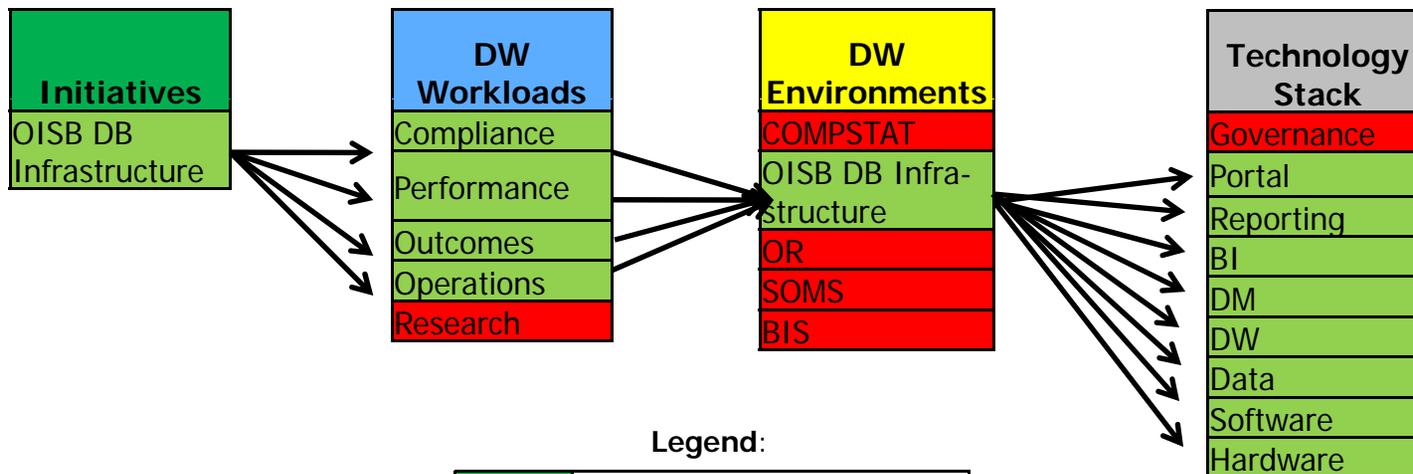
Legend:

Green	Projects
Blue	DW Workloads
Yellow	DW Environments
White	Technology Stack
Light Green	SOMS Project
Tan	BIS Project
Red	Gaps
White	Overlaps



# Mapping by Initiatives (As-Is) – cont.

Findings: Initiatives are silo'ed



Legend:

Green	Initiatives
Blue	DW Workloads
Yellow	DW Environments
White	Technology Stack
Light Green	OISB DB Infrastructure
Green	Initiative
Tan	
Red	Gaps
White	Overlaps



# Strategy

## Vision

Build a strategic consolidated Enterprise Data Warehouse (EDW) technical infrastructure with virtualized Data Marts (DM), Business Intelligence (BI) and Reporting functions.

As-Is

Technology Stack	Distributed	Consolidated
Governance	Yes	
Portal		Yes
Reporting	Yes	
BI	Yes	
DM	Yes	
DW	Yes	
Data	Yes	
Software	Yes	
Hardware	Yes	

To-Be

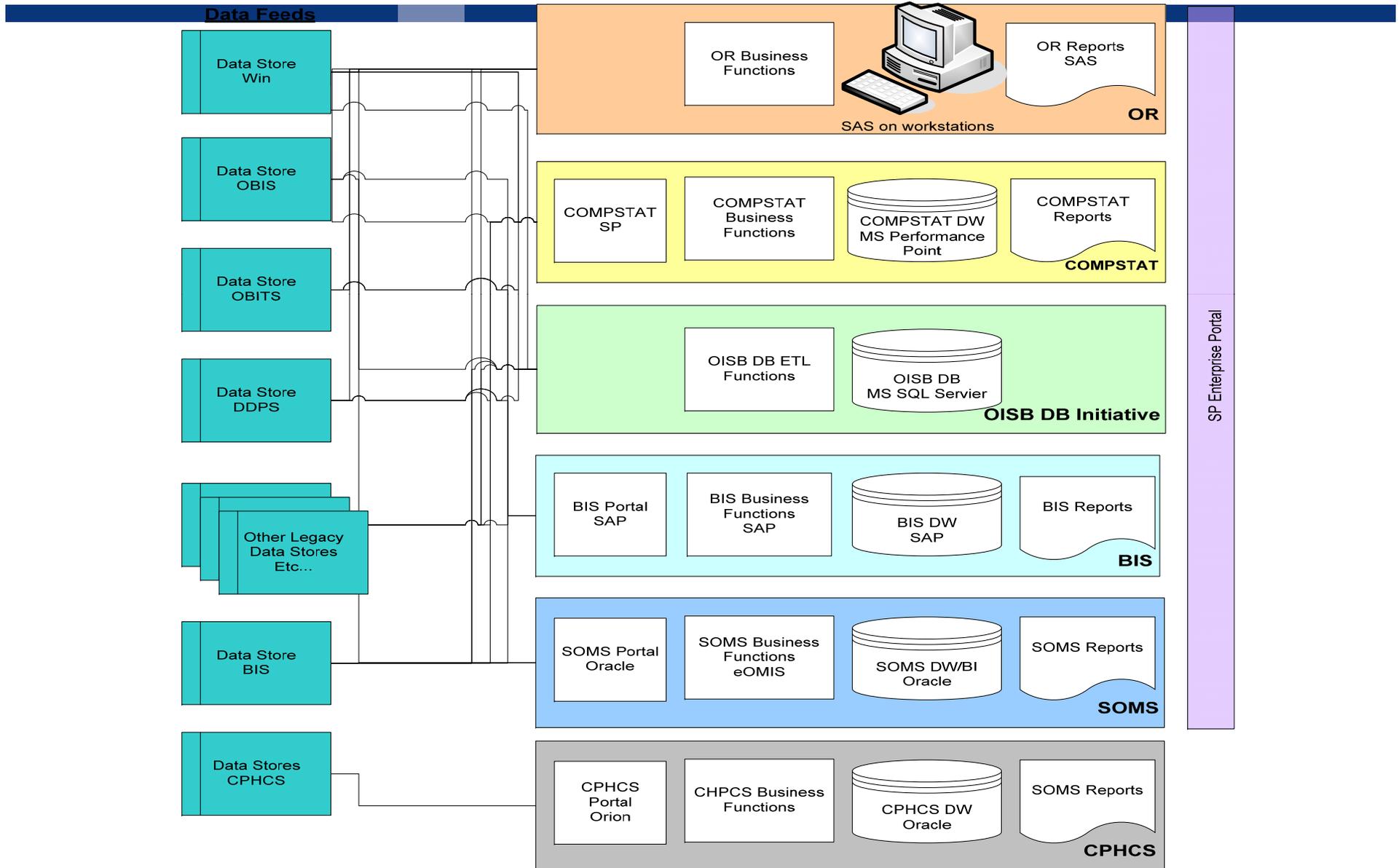
Technology Stack	Virtualized	Consolidated
Governance		Yes
Portal		Yes
Reporting*	Yes	
BI*	Yes	
DM*	Yes	
EDW		Yes
Data		Yes
Software		Yes
Hardware		Yes

\* The business will own and use their own tools from the datacenter.



# Strategy – cont. As-Is

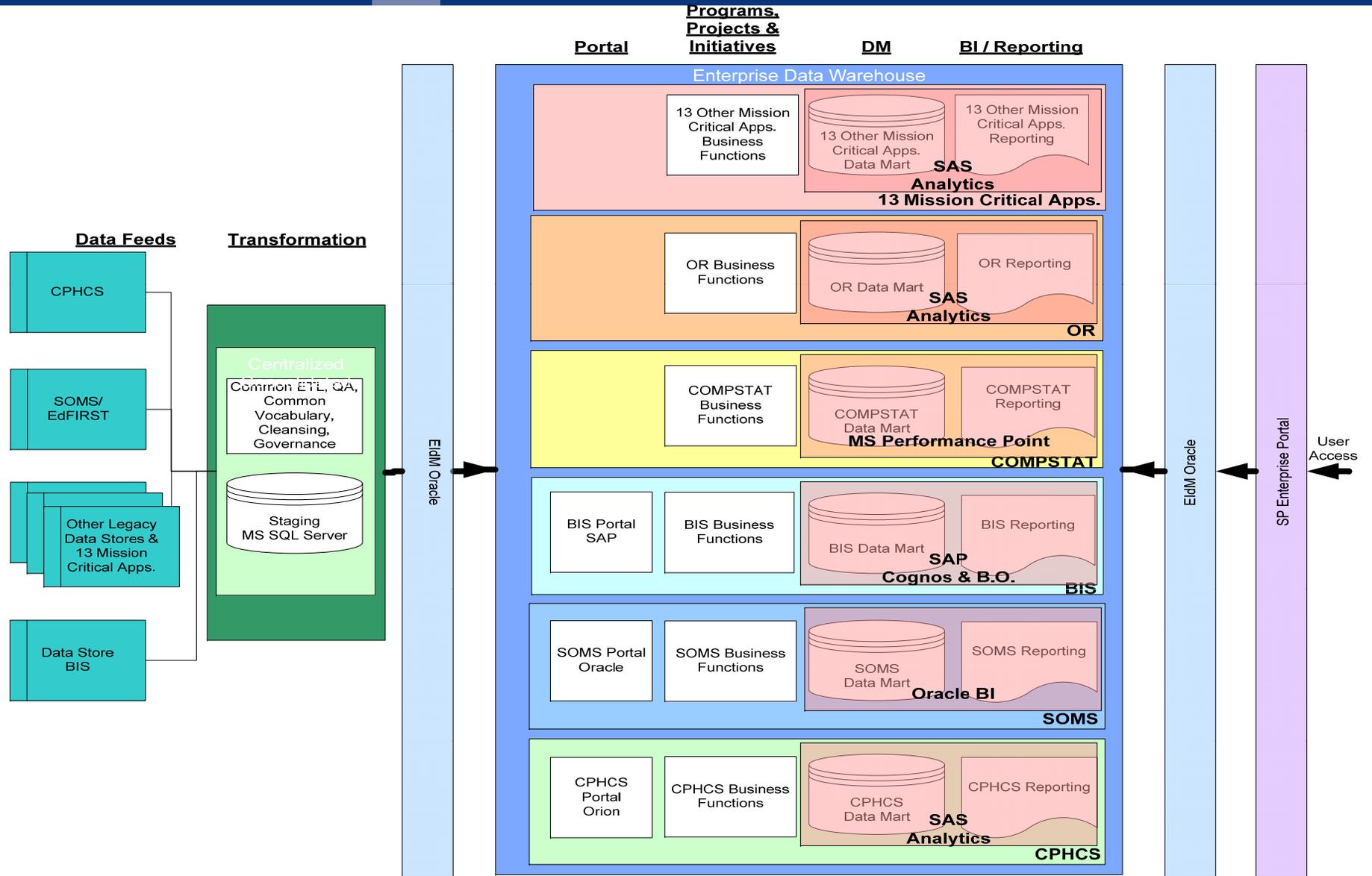
## Numerous Redundant Data Feeds To Silo'ed Environments Programs, Projects & Initiatives





# Strategy – cont. Phased To-Be

Consolidated, Integrated, Quality, Governed Data To EA Aligned TRM









## Strategy – cont. Authoritative Drivers

### Drivers (not prioritized, sorted alpha)

- CDCR Enterprise-wide Integration
- Executive (Budgeting)
- Outcome Measures
- Support the Business Values:
  - **Leverage** prior investments to implement new business functions – in a more timely and cost effective manner (SOMS, EdFIRST, DECATS, DORMS)
  - **Expose** CDCR business process gaps, allowing them to be resolved – BEFORE litigation occurs
  - Provide crucial planning information required to **integrate** CDCR and CPHCS
  - **Expedite** the integration and/or migration of legacy and new systems
- Support Court Requirements
- Support CPHCS
- Support Current Data Warehouses (COMPSTAT, OR, BIS, SOMS, others)
- Support EA Best Practices
- Support Enterprise Data Architecture Committee (EDAC)\*
- Support Legislative Requirements (CROB / AB900)
- Support Operational Needs
- Support Reportable Projects (PLM, CITIP, SOMS, RSTS, DECS, EdFIRST, ARNAT, WINPLO, BIS, RACS, CIPS)

\* See Appendix - C for EDAC alignment



## Strategy – cont.

### Strategies

- **Governance** - Clarify authoritative sources (who does what), organization authority, organization structure and funding mechanisms with consistent consolidated governance
- **Data** - Leverage CODB, OISB & COMPSTAT processes as a starting point to standardize and provide best practices. Consolidate data feeds – single centralized data source (quality, timeliness, accuracy)
- **Hardware/Software** - Leverage OISB DB Infrastructure initiative to create infrastructure and performance baseline and align to EA standards for HW / SW (Portal, DW, BI, reporting tools) consolidating on EDW / BI tools and leverage licenses from SOMS, OR, COMPSTAT
- **Strategic Reports** - Provide consistent integrated reports to provide a common look and feel



# Phased Targets

Phase I Target  
 Proof-of-Concept  
*(anticipated to be completed in Year 1)*

Distributed	Consolidated
	Governance
	Portal
Reporting	
BI	
DM	
DW	
	Data
	Software
	Hardware

Phase II Target  
*(anticipated to be completed in Year 1)*

Virtualized	Consolidated
	Governance
	Portal
Reporting*	
BI*	
DM*	
	EDW
	Data
	Software
	Hardware

Phase III Target

Virtualized	Consolidated
	Governance
	Portal
Reporting*	
BI*	
DM*	
	EDW
	Data
	Software
	Hardware

\* The business will own and use their own tools from the datacenter.



## Phased Targets\* - cont.

### As-Is – To-Be Target Steps Plan:

**Phase I Target** – Fund and execute OISB DB Infrastructure initiative as a proof-of-concept with one metric as the first step, focusing on consolidating Governance, Data, Software and Hardware

- Provide long-term funding and budget

**Phase II Target** – Focus on funding, data issues, scalability and metrics:

- Resolve key data issues for single central source, common vocabulary and data quality
- Build virtualized Reporting, BI and DM with consolidated EDW
- Plan scalability, capacity
- Build out infrastructure based on Phase I proof-of-concept
- Implement additional metrics through a consolidated governance process
- BIS integrated data with SOMS
- Focus on legacy systems migration

**Phase III Target** – Bring CPHCS and SOMS/EdFIRST, OR, and COMPSTAT feeds onto the EDW environment with Reporting

- Complete Reporting, BI, DM and EDW

\* See details in Troux



## Planning Considerations

- **SOMS** – SOMS is working on a roadmap for Release 1A scheduled Q1 2011, efforts can be dovetailed to maximize benefits
- **EIdM** – The EIdM project is planning to release with SOMS 1A for SOMS use, full functionality will be in Release 3 of SOMS
- **BIS** – BIS will be data integrated with SOMS 1A
- **SharePoint** – SP is the CDCR Enterprise portal, no dependencies have been found
- **CPHCS** – CPHCS is focused on their Medication Records project for 2011-2012 and does not have any direct dependencies; An opportunity may exist to use Oracle Health BI once SOMS is delivered to leverage the Oracle licensing.
- **SAS** – SAS analytics used at CPHCS and OR



## EDW Best Practices

- Start by defining your user population's requirements in terms of complexity and frequency of queries, both today and in the future
- Factor in how your data volume is expected to grow, the frequency of data loads, and the mix of workloads
- Consider scalability of data storage, network bandwidth, and processing capacity
- Establish your analytic requirements
- Look for an integrated analytics solution for best performance
- Design in acceptable performance criteria and SLAs
- Determine security requirements
- Map the skills and resources required to implement and manage your data warehouse today and in the future to what you already have.
- Determine your time-to-market requirements
- Analyze ongoing management costs and long-term data warehousing investment
- Identified a list of solution providers that meet your requirements
- Talk to reference customers
- Run benchmark tests or proof of concepts
- Provide reasonable timelines in your project plans
- Use real source data from operational systems

Reference: <http://www.tdwi.org/Publications/BIJournal/display.aspx?ID=8266>



## EDW Best Practice Links

- SAP - [http://searchsap.techtarget.com/tip/1,289483,sid21\\_gci1134359,00.html](http://searchsap.techtarget.com/tip/1,289483,sid21_gci1134359,00.html)
- Oracle - <http://www.oracle.com/technology/products/oracle-data-integrator/pdf/odi-bestpractices-datawarehouse-whitepaper.pdf>
- Microsoft - <http://msdn.microsoft.com/en-us/library/cc719165.aspx>
- Microsoft - [http://www.microsoft.com/presspass/press/2009/feb09/02-23SQLFastTrackPR.msp?rss\\_fdn=Press%20Releases](http://www.microsoft.com/presspass/press/2009/feb09/02-23SQLFastTrackPR.msp?rss_fdn=Press%20Releases)
- Gartner - <http://www.gartner.com/DisplayDocument?id=1221936>
- Gartner magic quadrant - <http://www.gartner.com/technology/media-products/reprints/ncr/article2/article2.html>
- Federal Enterprise Architecture - <http://www.whitehouse.gov/omb/e-gov/fea/>
- NASCIO Maturity Model - <http://www.nascio.org/publications/documents/NASCIO-EAMM.pdf>



Questions ???



# Appendix

- Appendix A – 18 Concerns Details
- Appendix B – DW Workloads Mapped to FEA PRM Framework
- Appendix C – EDAC Alignment



## Appendix A - 18 Concerns Details

1. Data quality / cleansing issues
2. Metrics not well established
3. Ad hoc tactical environment
4. Silo'ed data warehouses
5. Too many inconsistent portals
6. Lack of governance
7. Competition for operational data feeds is taxing providers
8. Security issues
9. Technical Performance
10. Requests are not well thought out causing miss-communication and confusion of reports
11. EIS services (SLAs) lacking
12. Data definition and semantics inconsistencies between data warehouses
13. Don't understand the lifecycle clearly which re-enforces a reactive mode
14. Don't reconcile conflicts and inconsistencies of demands, results in dissatisfied customers
15. DW / BI / reporting currently is tactical ad hoc
16. Too many projects in flight saturates ability to respond
17. Performance measurements are inconsistent, not correctly stated, does not provide a consistent story
18. Enterprise policies or direction not established resulting in tactical silo'ed decisions



# Appendix B - DW Workloads Mapped to FEA PRM Framework

## FEA PRM Framework

DW Workload	Processes & Activities	Mission & Business Results	Customer Results	Human Capital	Technology	Other Fixed Assets
Research						
Compliance						
Operations						
Outcomes						
Performance						

**Legend:** Green notes workloads that support the FEA PRM Framework



## Appendix C - EDAC Alignment

The DW / BI / Reporting Strategy is positioned to support EDAC alignment and is leveraged by the first strategy noted above (Clarify authoritative sources, organization authority, organization structure and funding mechanisms):

- ✓ Manage from a consolidated enterprise-wide perspective
- ✓ Collect and manage information as an asset in accordance with its business values
- ✓ Promote application consolidation, standardization, and integration
- ✓ Secure information assets from unauthorized access, use, modification, destruction, and / or disclosure
- ✓ Delegate security authorization authority regarding information classification and access
- ✓ Promote a high level of security awareness to employees and partners
- ✓ Assign a single steward or authoritative source for each individual data item with clearly defined locations and data accessibility
- ✓ Separate production databases from data warehouses used for decision support
- ✓ Maintain common and consistent information / data definitions across all business units (e.g., data dictionary)
- ✓ Ensure data accessibility to all authorized entities
- ✓ Store and transmit data electronically to avoid data transcription and manual re-entry
- ✓ Preserve and maintain data such that the information remains accessible and useable for the designated retention period
- ✓ Ensure system data is accessed through the use of business rules