White Paper

HEALTH CARE DATA WAREHOUSE FOR BUSINESS INTELLIGENCE AND ANALYTICS

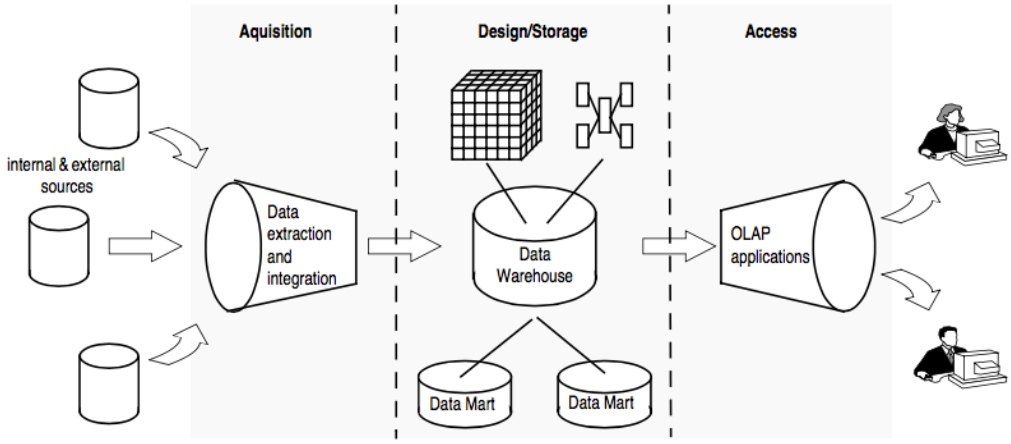
Data warehouse and its applications

A white paper understanding of data warehouse and its usage in business intelligence and analytics applications built for health care domain.

Prashant Verma | vprash360@gmail.com

Introduction

Data warehouse design process began with data modeling, including the definition of metadata. Major design decisions associated with developing our data model included establishing naming standards, operating parameters, dimensions, and keys, as well as the overall configuration.



A typical data warehouse system architecture

Data Model Design

A prerequisite to beginning the design process was to thoroughly under- stand the business processes and available data sources within our organization. Once this milestone was achieved, we used the following data model design process:

- Choose the subject areas to be modeled.
- Identify and design the conformed dimensions.
- Define the dimension granularity (the lowest level of information to be stored in the dimension tables).
- Design the facts.
- Determine the fact granularity (the lowest level of information to be stored in the fact tables).
- Transform conceptual design to physical design. – Naming standards and conventions. Data files/SQL files , Database , Application
- Character case. – Apply design considerations for database parameters. – Define fact and dimension tables as the physical data model.

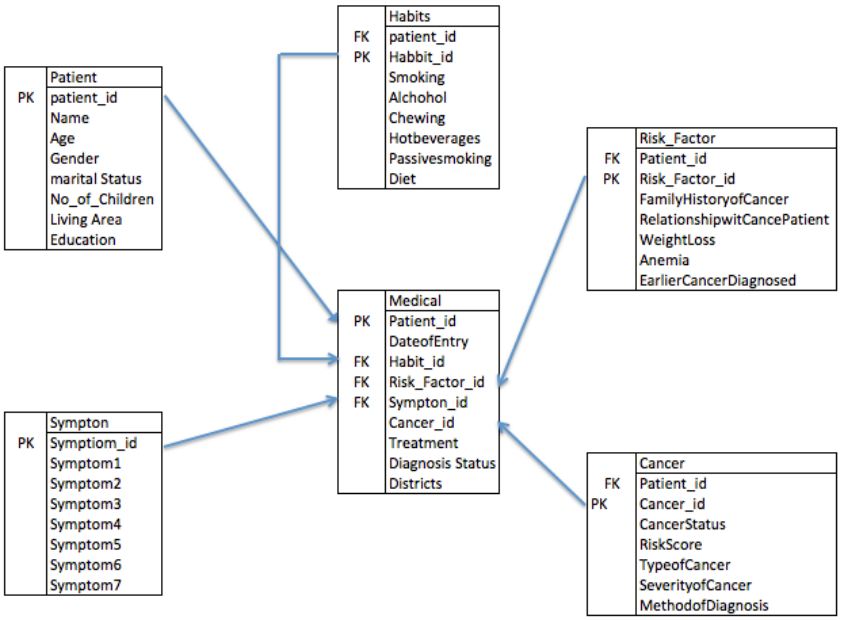
Datawarehouse Architecture Methodologies:

Inmon and Kimbol approach

	Inmon	Kimball
Methodology and architecture		
Overall approach	Top-down	Bottom-up
Architectural structure	Enterprisewide (atomic) data warehouse "feeds" departmental databases	Data marts model a single business process; enterprise consistency achieved through data bus and conformed dimensions
Complexity of the method	Quite complex	Fairly simple
Comparison with established development methodologies	Derived from the spiral methodology	Four-step process; a departure from RDBMS methods
Discussion of physical design	Fairly thorough	Fairly light
Data modeling		
Data orientation	Subject- or data-driven	Process oriented
Tools	Traditional (ERDs, DISs)	Dimensional modeling; a departure from relational modeling
End-user accessibility	Low	High
Philosophy		
Primary audience	IT professionals	End users
Place in the organization	Integral part of the Corporate Information Factory (CIF)	Transformer and retainer of operational data
Objective	Deliver a sound technical solution based on proven database methods and technologies	Deliver a solution that makes it easy for end users to directly query the data and still get reasonable response times

Source: Data Warehousing : Battle of the Giants: Comparing the Basics of the Kimball and Inmon Models by Mary Breslin.

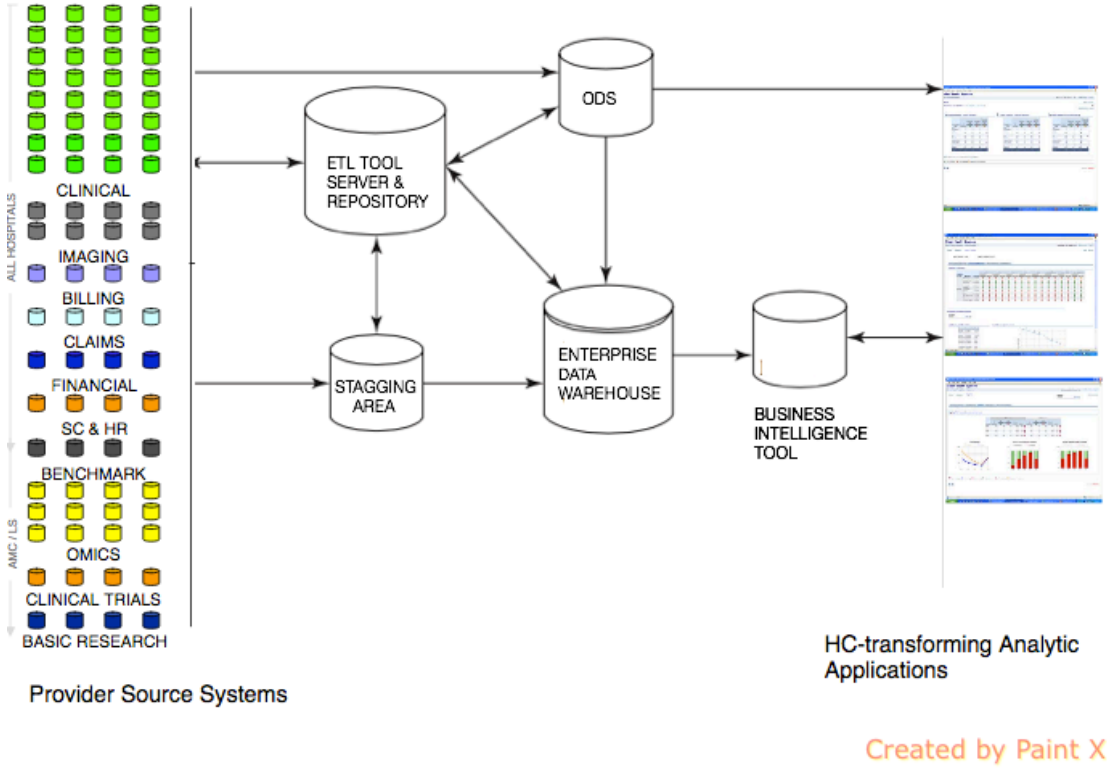
A typical data warehouse for Patient Behavioral Analytics



Architecture

This section outlines the design of the logical and physical architectures of our data warehouse. The logical architecture defines the integration of logical components, whereas the physical architecture describes hardware configuration, operating system, and file system.

Logical Architecture

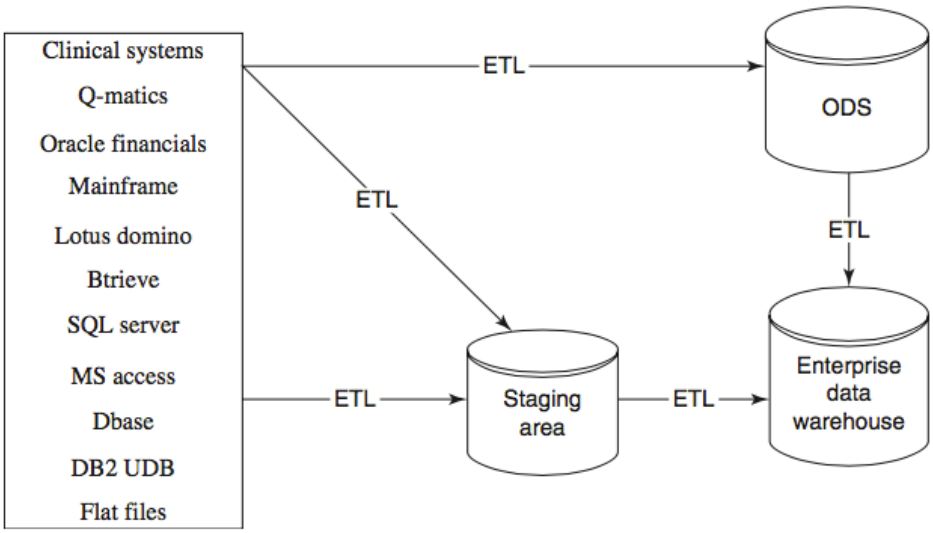


The logical architecture of our data warehouse, shown in Figure 7.3, incorporates both data and applications. The source applications use a variety of database and file formats, including Microsoft Access and flat files.

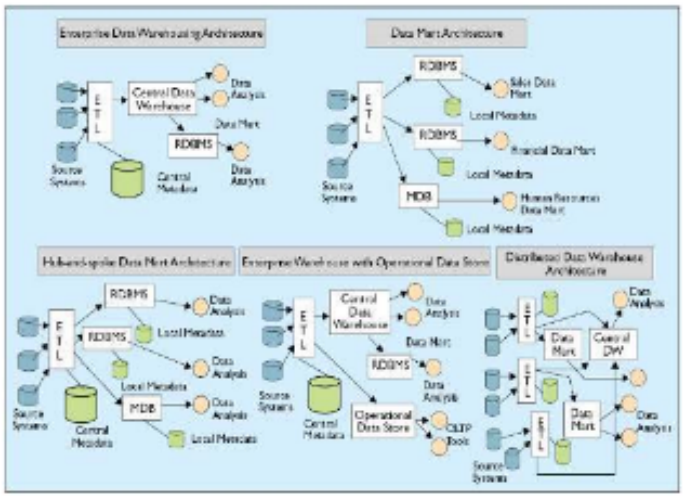
We have an operational data store (ODS) and enterprise data warehouse, both implemented in Oracle 10g. The ODS enables near real-time reporting of key performance indicators (KPIs). The enterprise data warehouse provides a buffer between the source systems and the query applications. This design protects the source systems from the impact of the intense, periodic ETL activity.

Extraction, Transformation and Loading

To populate the data warehouse, a well efficient, accurate and complete ETL operations are essentials for proper operation till the reporting. The data is extracted from sources and dumped at staging area 'as it is'. Why we call it 'as it is', because, data is in its original form. In the loading process, data from the staging area are loaded into the data warehouse databases after performing transformation activity on the data. The transformation process involves cleaning and transforming the data. Healthcare data can reside in many different data source like Cerner, Mainframe, Oracle Financials, SQL Server, Btrieve, Microsoft Access, Q-matics, DB2 UDB, Lotus Domino, dBase source systems, flat files etc. Every data source need be connect properly in order to prevent any data leakage.



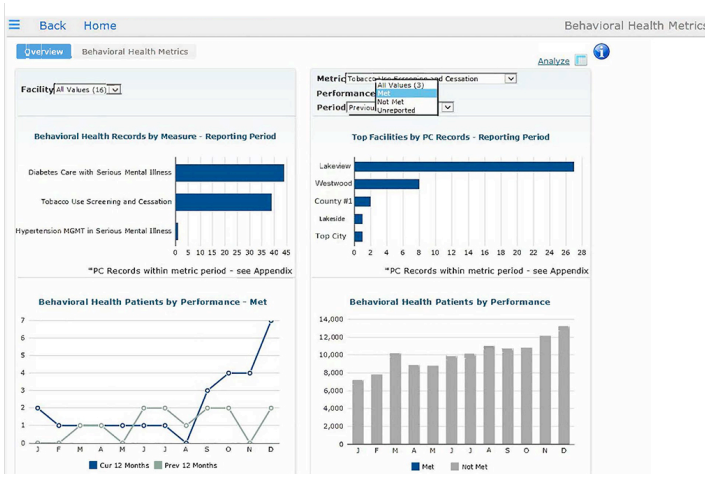
Below are different types of data warehouse architecture which can be used for different objectives.



Source: A Data Warehouse Architecture for Clinical Data Warehousing, by Tony R. Sahama and Peter R. Croll

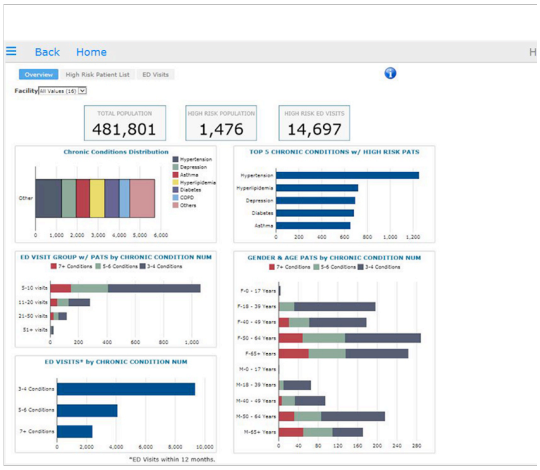
Business Intelligence Dashboard

Behavioral Health Dashboard
contains matrices that address early detection, treatment and management of patients with behavioral health and medical conditions.

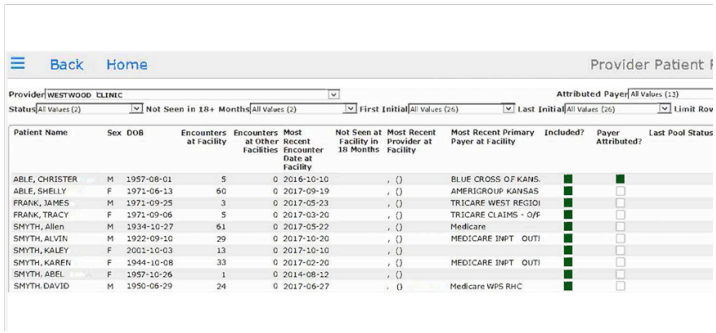


Quality Metrics dashboard
analysis of preventive care procedures commonly required and approved for quality reporting programs for clinic practices.

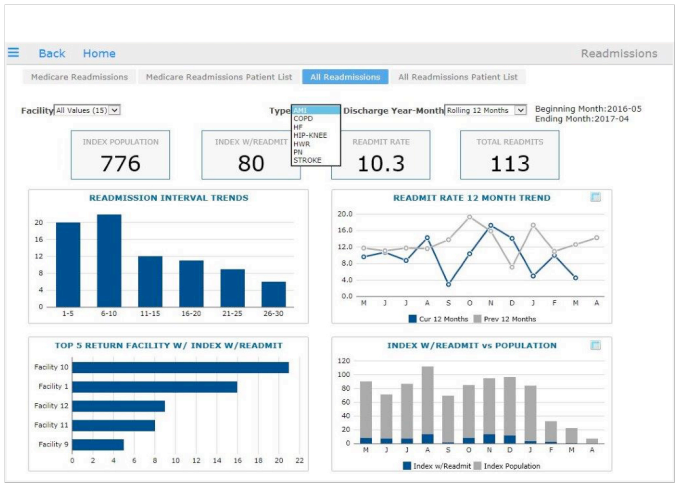
High Risk Patient dashboard
identifies patients considered most at risk for poor health outcomes. For this analysis, high risk patients are defined as patients with three or more chronic conditions and five or more emergency department visits in a 12-month period.



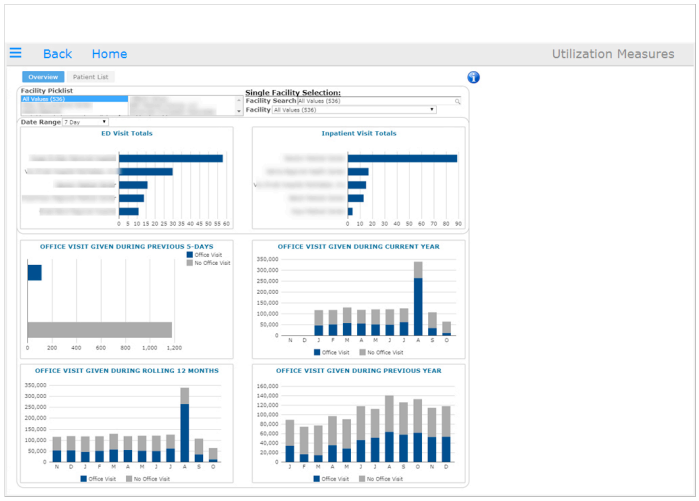
Patient Attribution
The summary view displays patient name, visit activity and most recent primary provider and payer. Patient level encounter detail is available.



Disease Registries display
specific patient populations with certain high or at risk conditions, and sets the stage for physicians to take steps .



Utilization dashboard
presents recent patient activity for inpatient admissions, emergency department and office visits. This dashboard displays all patients in the population with eligible service activity, un-restricted by age, disease condition or level of utilization. Additional charts display office visit activity



Analytics with Data Warehouse

Traditionally information hidden in data has been uncovered through the use of statistical software packages R or Python. This involves a user developing a theory about a possible relation in a database and converting that hypothesis into a query. The difference with data mining is that the interrogation of the data is done by the data mining algorithm, rather than by the user.

In other words, data mining is a data-driven, self-organizing, bottom-up approach to data analysis, whereas statistics are user or verification Driven The major functions of data mining and their roles in yielding different types of knowledge can be described as follows:

Characterization
– for the generation of characteristic rules. Characterization generalizes a set of task-relevant data into a generalized data cube. Different kinds of rules can then be extracted to form characteristic rules. Defining the symptoms of a specific disease is an example of characteristic rule.

Comparison
– for the generation of discriminate rules which summarize the features that distinguish the class being examined [i.e. the target class] from other classes [i.e. contrasting classes]. For example, to distinguish one disease from others, a discriminate rule summarizes the symptoms that discriminate the disease from others.

Classification
– for the generation of classification rules. Classification analyzes a set of objects whose class label is known and constructs a model for each class based on the features in the data. For example, the process of classification may classify diseases and provide the symptoms which describe each class or subclass.

Association
– for the generation of association rules. Association discovers traits or patterns in the data which apply to the set or subset of classes. For example, the process of association may discover a set of symptoms often occurring together with certain kinds of diseases and further study the reasons behind them.

Prediction
– for the prediction of possible values of some missing data or the value distribution of certain attributes in a set of objects. Prediction involves finding the set of attributes relevant to the attribute of interest, and predicting the value distribution based on the set of data similar to the selected objects. For example, a patients potential recovery time can be predicted given the recovery time distribution of similar patients.

Cluster analysis
– for the grouping of data in the database or data warehouse into a set of "clusters" to ensure that the interclass similarity is low and the intraclass similarity is high. For example, cluster analysis may cluster all the surgeons at Vienna General Hospital according to the department they mainly work for, surgery class they primarily operate and some sort of performance measure.

Time-series analysis
– for the analysis of time-related data in databases and data warehouses. Such analyses may include: similarity analysis, periodicity analysis, sequential pattern analysis, and trend and deviation analysis. For example, a time-series analysis may find the general characteristics of the treatment of patients whose recovery period has been 20% less than average given their respective diagnosis.

***

