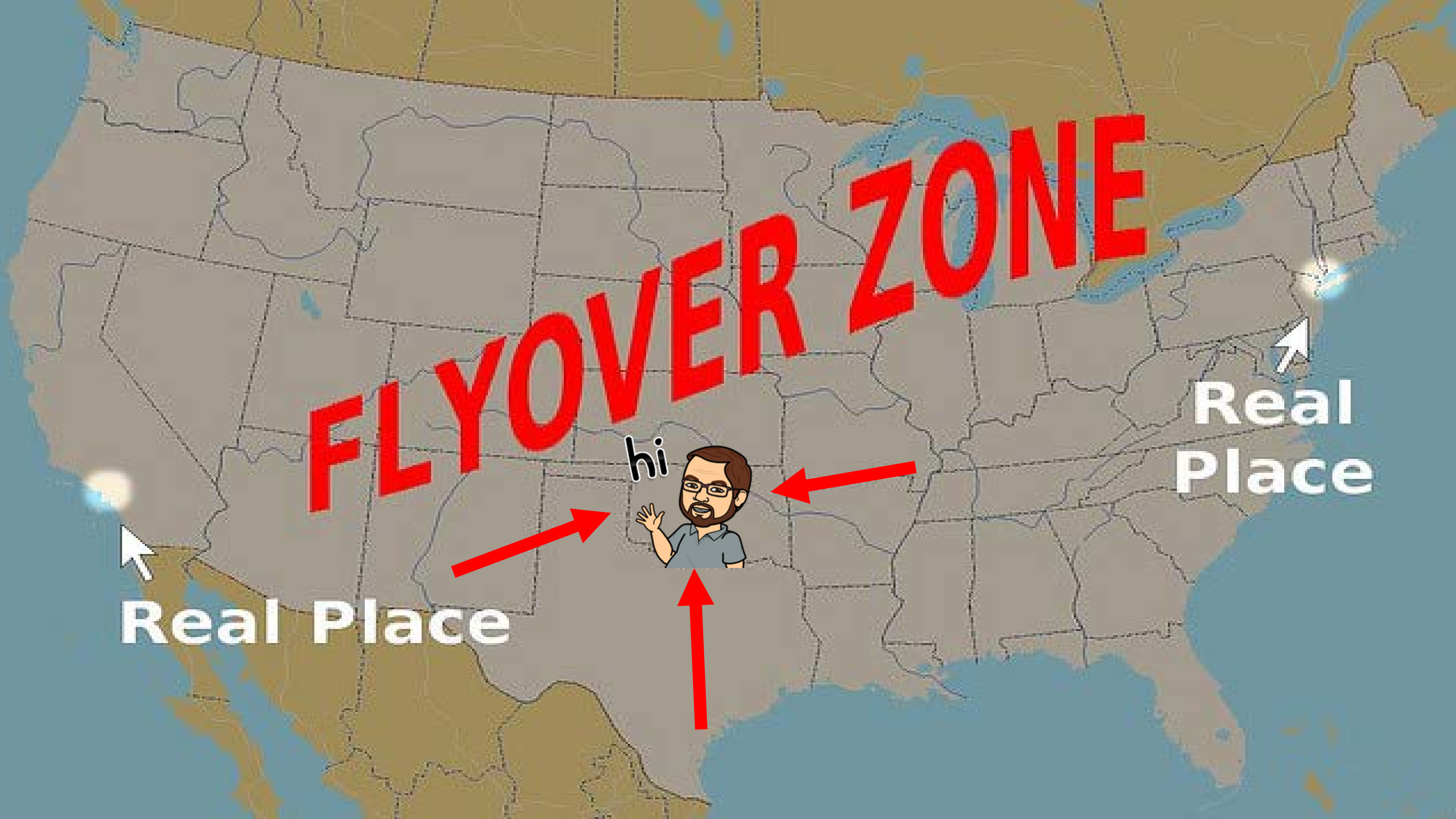
ETL and Data Warehouses

Chris Williams, MD
University of Oklahoma
Health Sciences Center


FLYOVER ZONE
hi
Real Place
Real Place
Real Place

https://en.wikipedia.org/wiki/Tornado

Disclosures

I have no relevant financial relationships with commercial interests to disclose in relation to the content of this presentation.

Learning Objectives

- Describe a Data Warehouse
- Define the ETL (extract-transform-load) process
- How does a Data Warehouse add value

Agenda

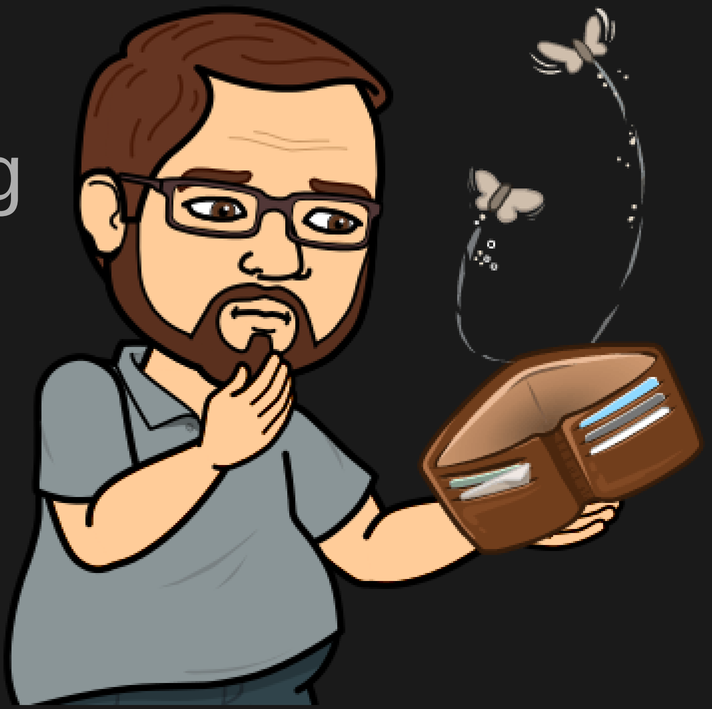
The Why

The economy, stupid.

-James Carville

The Why

Competitive (and non-competitive) forces are driving down reimbursements



The Why

What’s the *<u>VALUE</u>*?

The Why

What’s the *RISK*?

The Why

Quality

Occasional disconnect between Perception and Reality

The Why

High Performing Enterprises

are Innovators

The Why

Healthcare Systems are in the Information Business, some of them act like it

Clinical Use of an Enterprise Data Warehouse

R. Scott Evans, MS, PhD[1,2], James F. Lloyd, BS[1], Lee A. Pierce, BS, MIS[3]
[1]Medical Informatics, Intermountain Healthcare, [2]Department of Biomedical Informatics, University of Utah, [3]Enterprise Data Warehouse, Intermountain Healthcare
Salt Lake City, Utah

Abstract

The enormous amount of data being collected by electronic medical records (EMR) has found additional value when integrated and stored in data warehouses. The enterprise data warehouse (EDW) allows all data from an organization with numerous inpatient and outpatient facilities to be integrated and analyzed. We have found the EDW at Intermountain Healthcare to not only be an essential tool for management and strategic decision making, but also for patient specific clinical decision support. This paper presents the structure and two case studies of a framework that has provided us the ability to create a number of decision support applications that are dependent on the integration of previous enterprise-wide data in addition to a patient's current information in the EMR.

Introduction

Inpatient and outpatient electronic medical records (EMR) are accumulating enormous amounts of patient, provider, facility, financial and process information. During the early 1990s, this information began to be recognized as an extremely valuable and untapped resource for management and clinical research. However, EMR administrators were concerned about the impact of running large research queries on the clinical database. It was determined that healthcare needed to convert this data into aggregated and separate information systems that could support retrospective and population-based analysis[1]. Data warehouses had emerged in other industries; however, their adoption by healthcare was slow due to the complexity and heterogeneity of medical, operational, and clinical data[2].

As an effort to facilitate access to this wealth of medical information, data warehouses that contained clinical and administrative data from healthcare organization began to be developed[3]. Using network technologies, interfaces were developed to collect the data from the different databases and stored in a single large database. However, early on, it was recognized that the data from many sources not only needed to be integrated, but also cleansed, and formatted[4]. Data semantics were then used to regroup and merge patients' medical data from the autonomous and heterogeneous health information systems[5]. As expected, this raised concerns of data security and patient privacy. Solutions supporting U.S. and European laws for high level of security, retrieval audit, and user authentication needed to be incorporated to ensure privacy and confidentiality[2,5]. As further uses for data warehouses were identified, image data using the (digital imaging and communication in medicine) DICOM standard was used to integrate information from picture archiving and communication system (PACS)[6-8]. The advantage of sharing data owned by different organizations was identified and federated information models were developed[9-11]. While HL7 is often used as the interface standard for integrating the data from divergent data silos, other data standards including RxNorm[12], SNOMED-CT, ICD, CPT, LOINC, UMLS and DRG codes[3,13,14] are also often included within the data warehouses. The data stored within these data warehouses can be managed and accessed through direct Structured Query Language (SQL) calls or SQL that is imbedded inside of Application Programming Interfaces (APIs) that are programmed in C++, Java, Perl, etc. User interfaces have also been developed in Visual Basic and distributed as ActiveX objects embedded in an HTML page[15] or information retrieval can be performed using metadata-based semantic and full-text search methods[10]. Web front-ends using i2b2 and caGrid frameworks

Intermountain Healthcare

Established EDW ~20 yrs ago

Data Driven Enterprise

Business Intelligence

Clinical Operations

Research

Evans, R.S, et. al. (2012) Clinical Use of an Enterprise Data Warehouse. *AMIA Annu Symp Proc.* .p189-198 PMID: 23304288

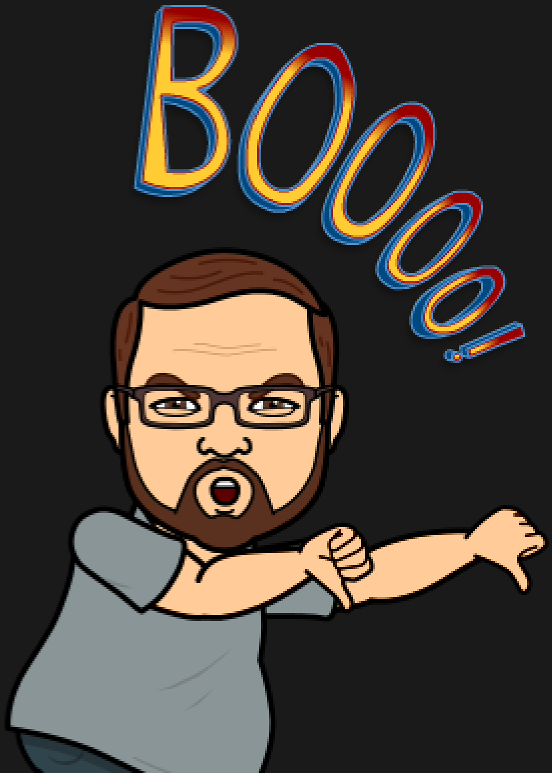
The What

How to define a Data Warehouse?

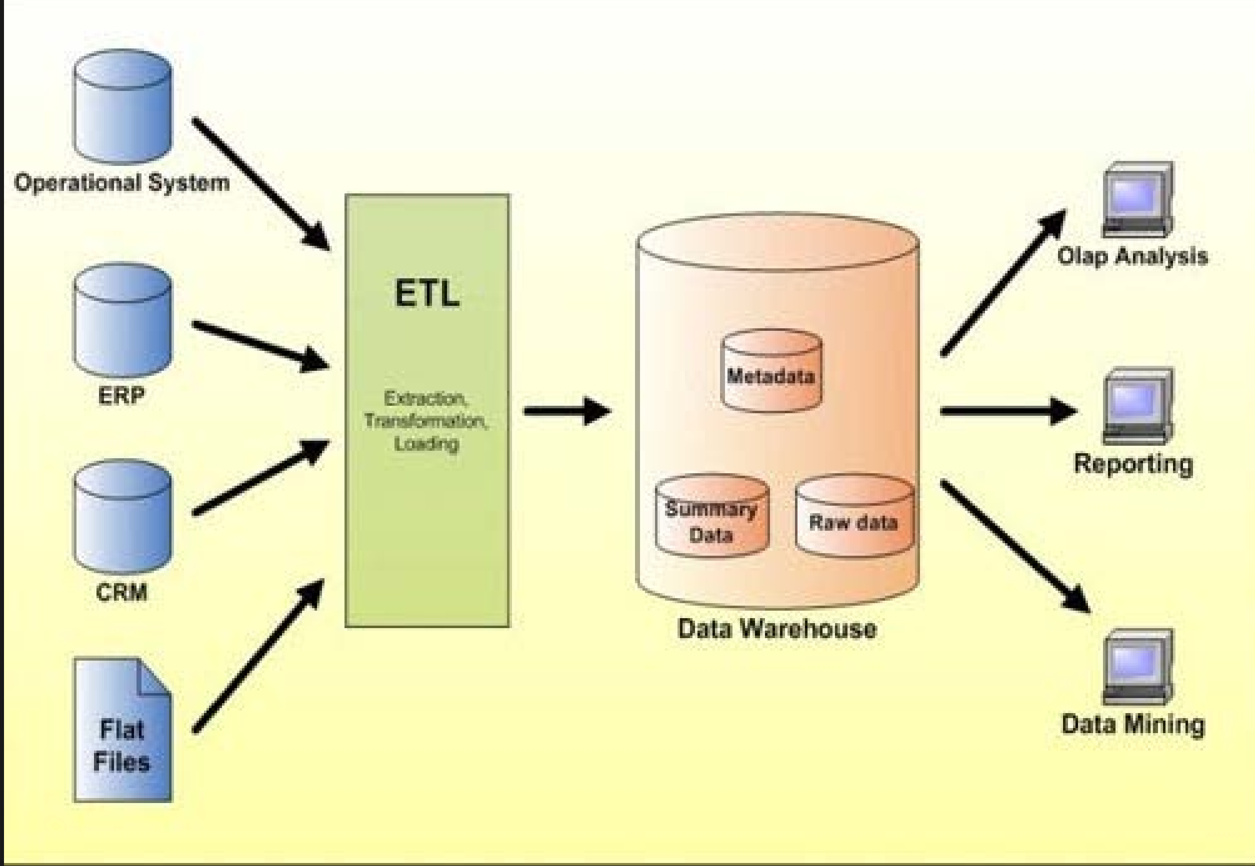
The What

How to define a Data Warehouse?

Wikipedia = No Help



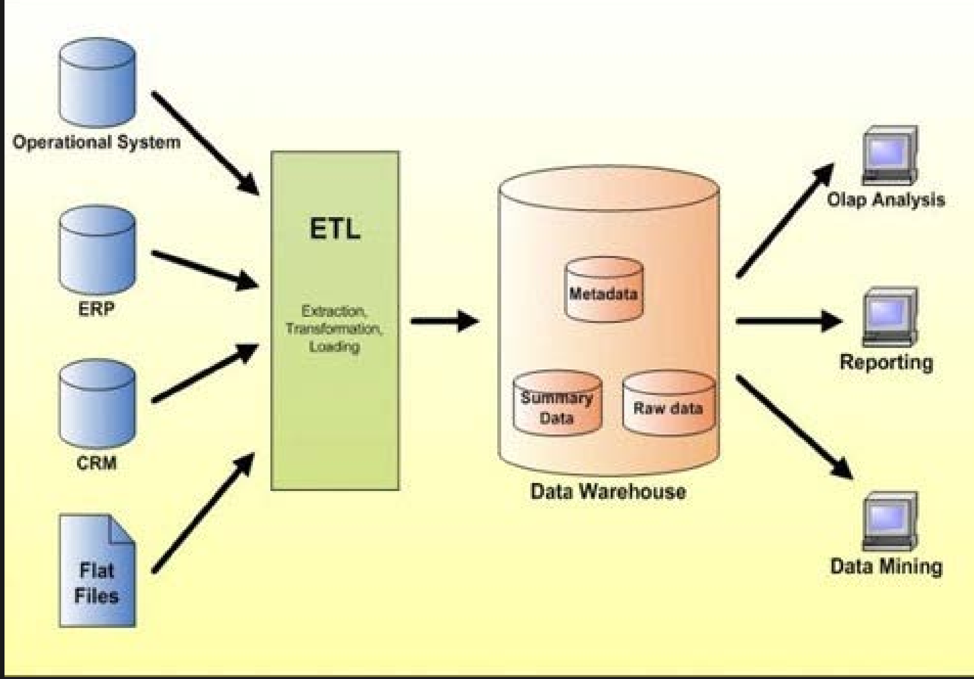
The What



Distillation of every Data Warehouse diagram

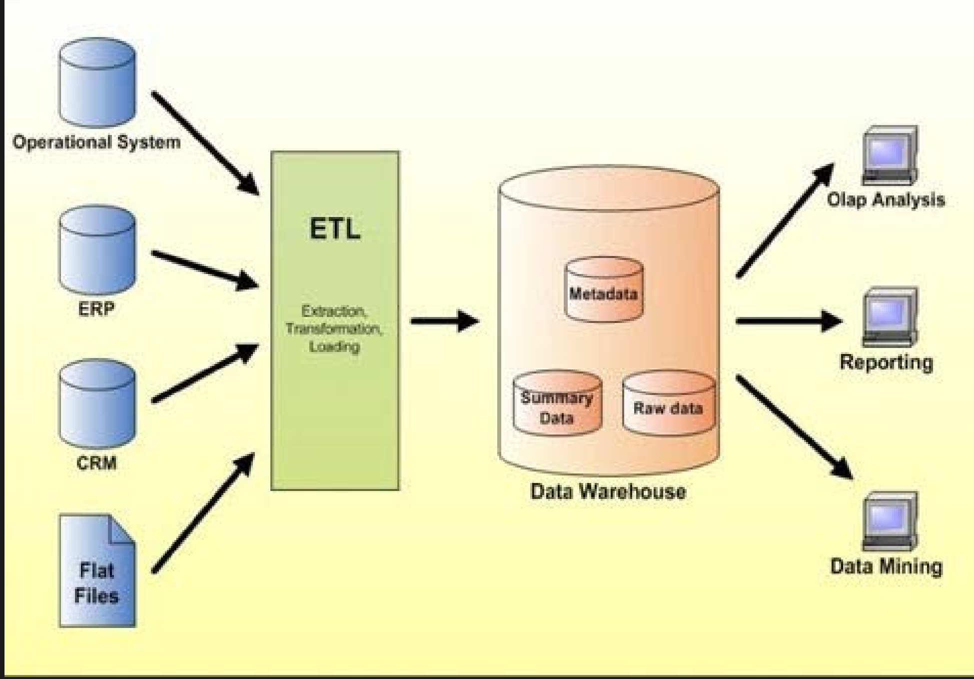
The What

1. Ambiguous cylinders



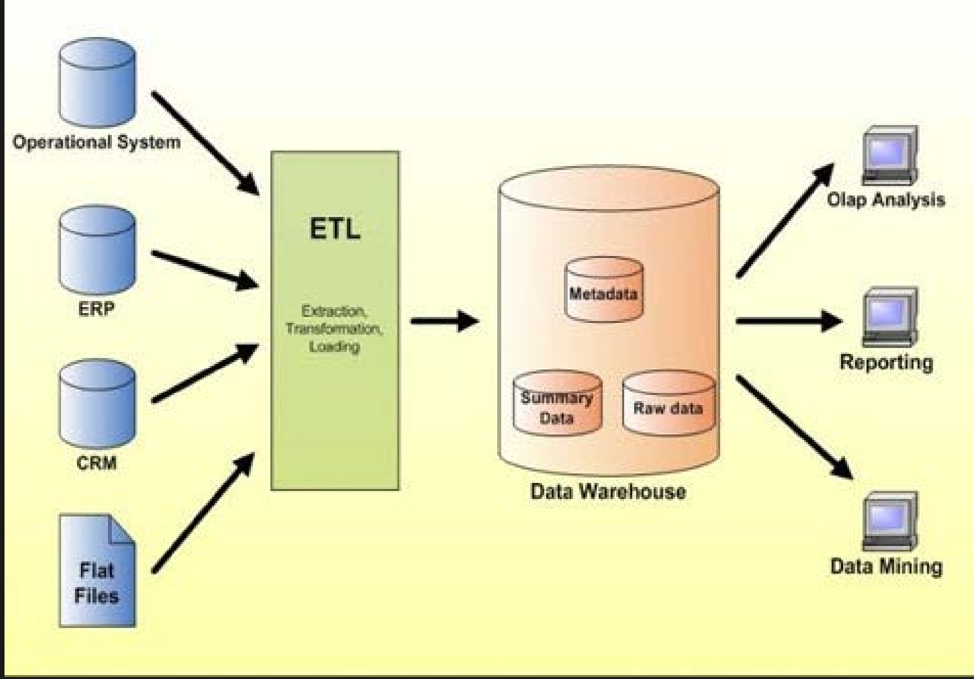
The What

1. Ambiguous cylinders
2. Magic Box



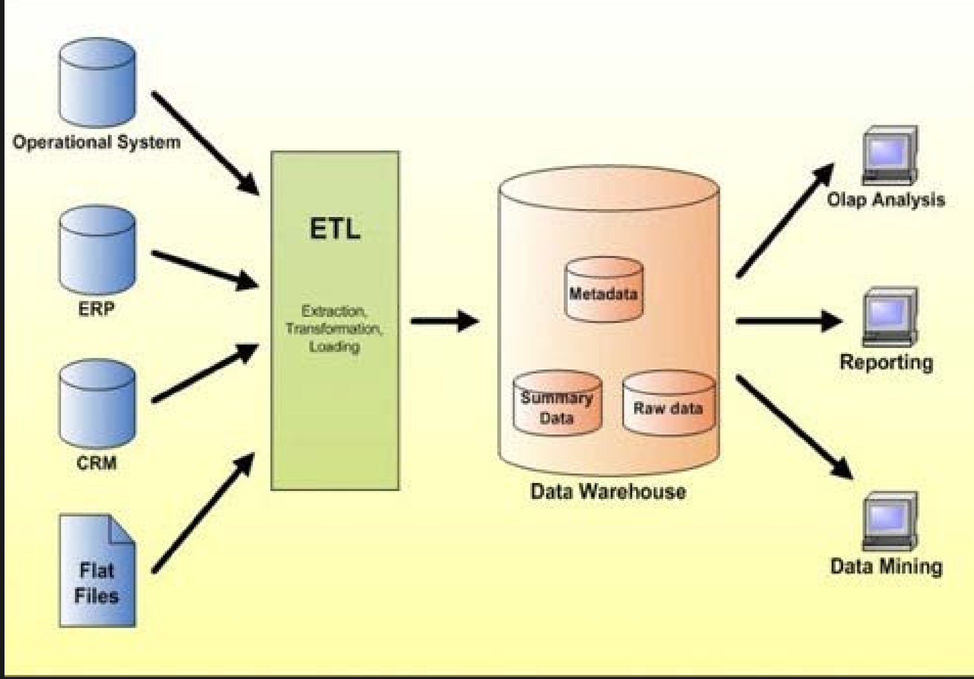
The What

1. Ambiguous cylinders
2. Magic Box
3. Larger cylinder



The What

1. Ambiguous cylinders
2. Magic Box
3. Larger cylinder
4. 80’s era terminals



The What

Common Characteristics of a Data Warehouse

1. Multiple sources of data

The What

Common Characteristics of a Data Warehouse

1. Multiple sources of data
2. Data Transformation/Normalization

The What

Common Characteristics of a Data Warehouse

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case

The What

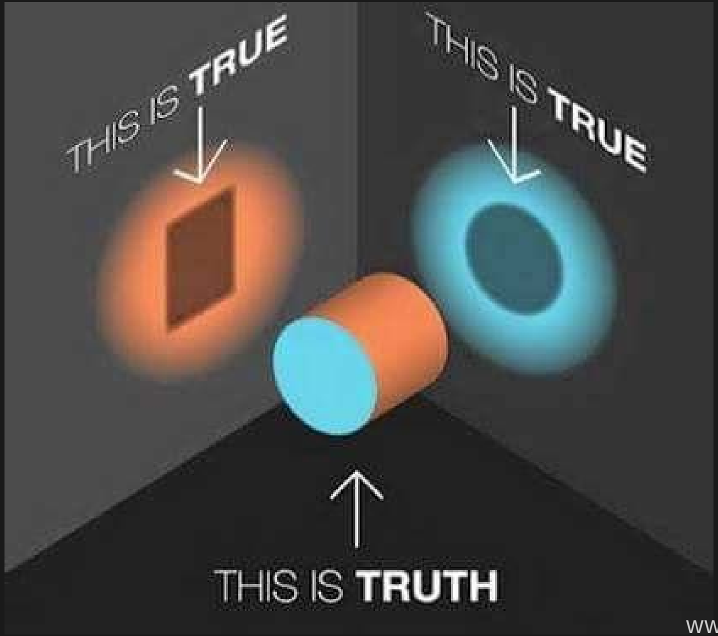
Common Characteristics of a Data Warehouse

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

Define the source of “Truth”



www.flickr.com/photos/leighblackall/19693376419

The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

Define the source of “Truth”

Source systems vs. Local Copy

The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

Federated Model

Upfront Cost

Recurring Effort

The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

Local Copy

Upfront Cost

Recurring Effort

The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

Data Model

Enterprise data model

Data Lake

The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

Enterprise Data Model

“A place for everything, and everything in its place.”

The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

Enterprise Data Model

Effort is theoretically front-loaded in setup and design, reporting should easier since data is “clean”

The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

ETL - **Extract**, Transform, Load

Import data from source system

The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

ETL - Extract, **Transform**, Load

Convert formatting to conform to destination data format

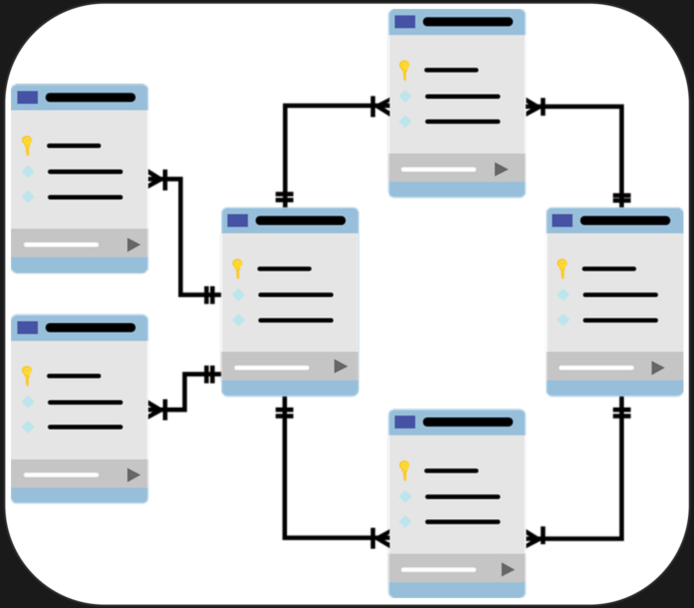
The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

ETL - Extract, **Transform**, Load

Star schema - Dimensions and Facts

Database Normalization



The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

ETL - Extract, **Transform**, **Load**

After conversion, store data in to the data warehouse

The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

ETL - Extract, **Transform**, Load

Data conversion is one way…

Potential for loss of meaning, may not be able to reconstruct original data

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

The How

ETL - Extract, **Transform**, Load

Trending values over time… have methods and reference ranges remained constant?

The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

ETL - Extract, **Transform**, Load

What version of AJCC or WHO were used for diagnosis or classification?

The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

ETL - Extract, **Transform**, Load

May need to consider Metadata

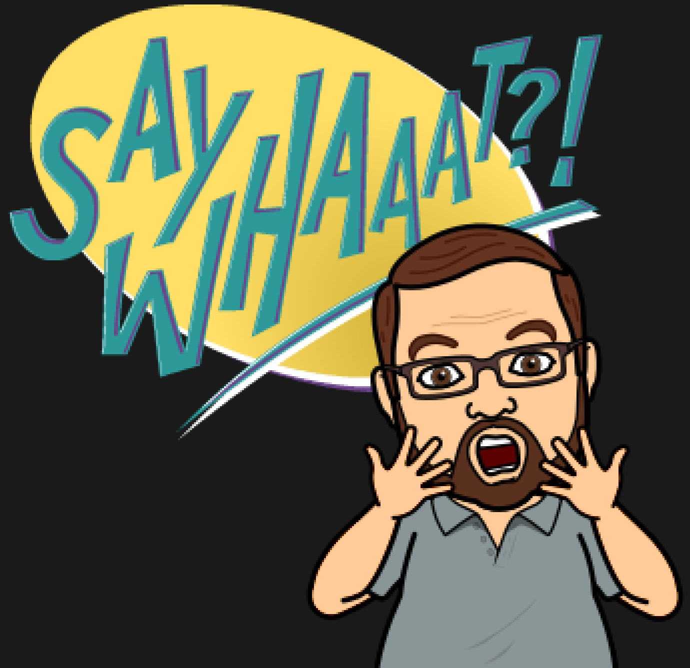
“Metadata is just data about data”

The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

Enterprise Data Model

Inability to recover the original data is not ideal



The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

Enterprise Data Model

Inability to recover the original data is not ideal

Falling storage costs have enabled the Data Lake concept

The How

Data Lake

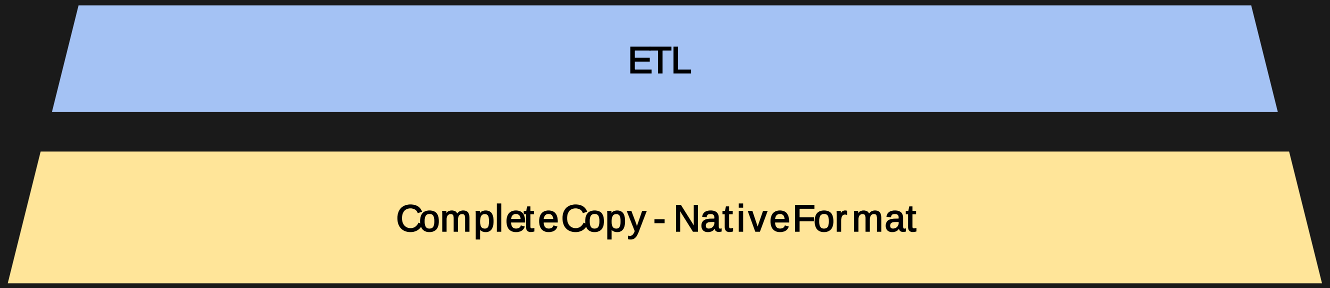
1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

Complete Copy - Native Format

The How

Data Lake

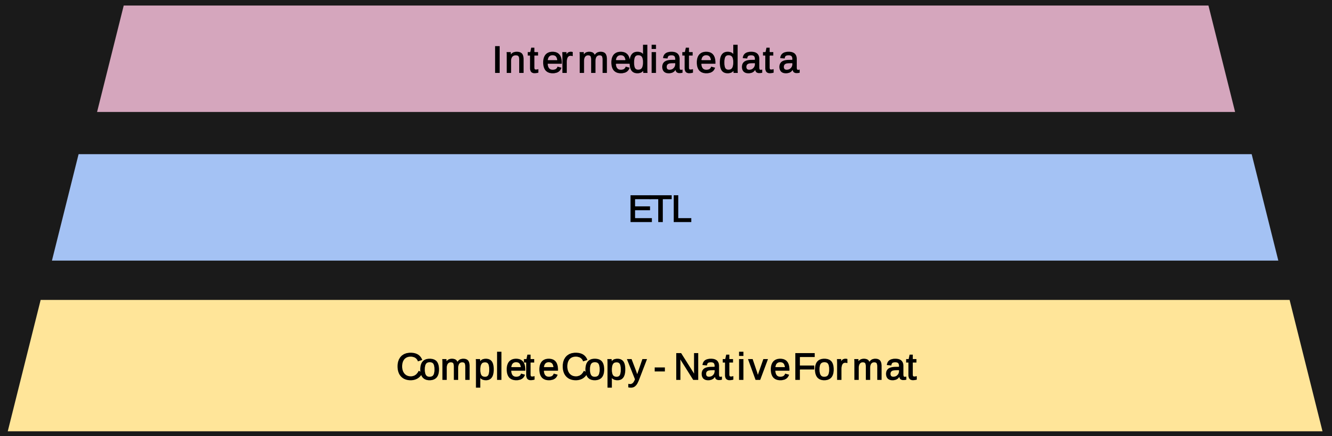
1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools



The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

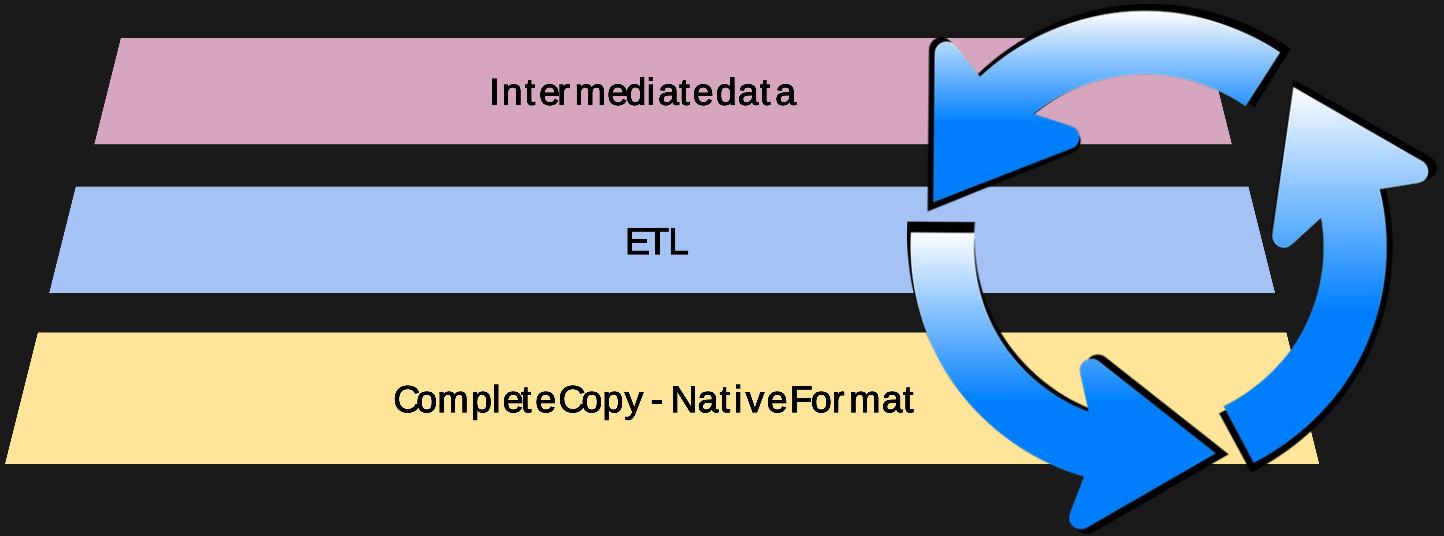
Data Lake



The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

Data Lake



The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

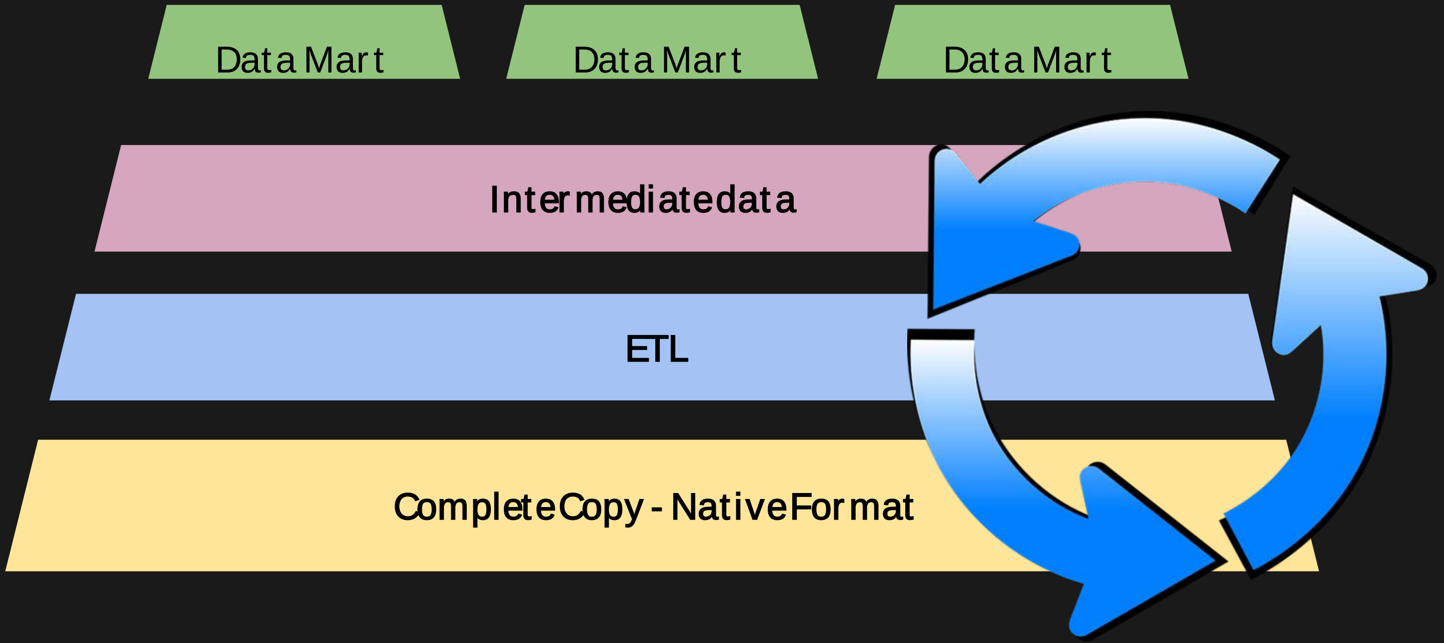
Data Marts

- Smaller DBs with limited scope
- Designed for a specific purpose

The How

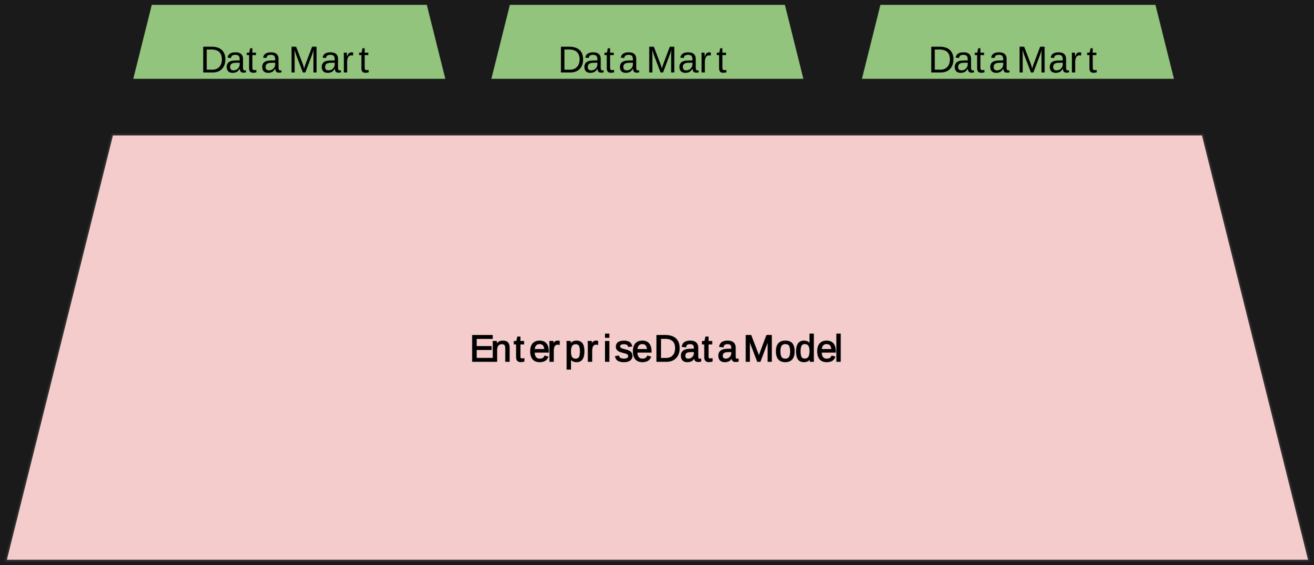
1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

Data Lake



The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools



The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

Data Marts

- Smaller DBs with limited scope
- Designed for a specific purpose
- Enhanced performance

The How

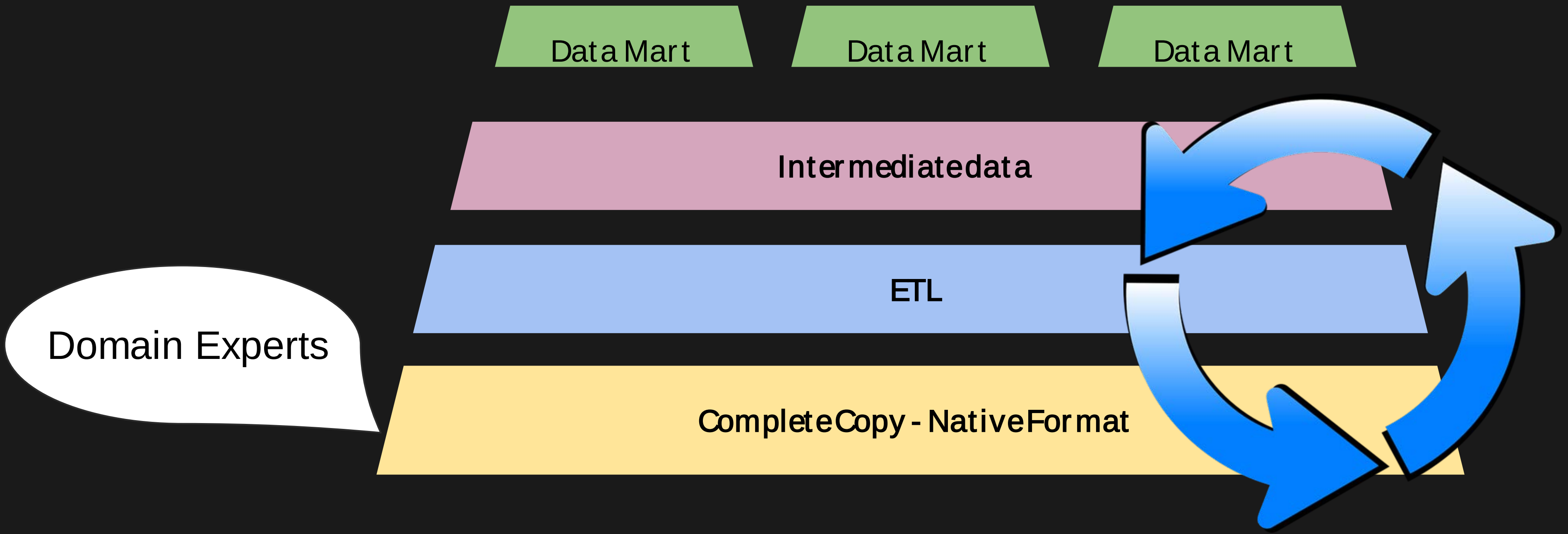
1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

Data Marts

- Smaller DBs with limited scope
- Designed for a specific purpose
- Enhanced performance
- Encapsulate Domain Specific Knowledge

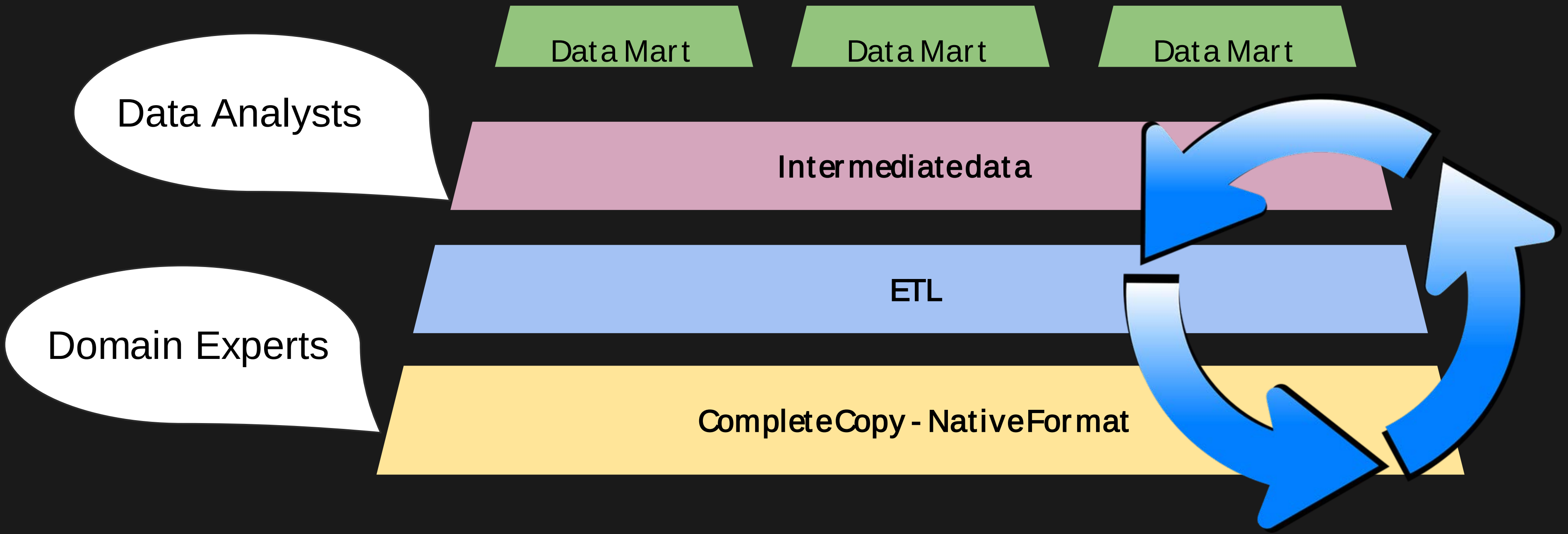
The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools



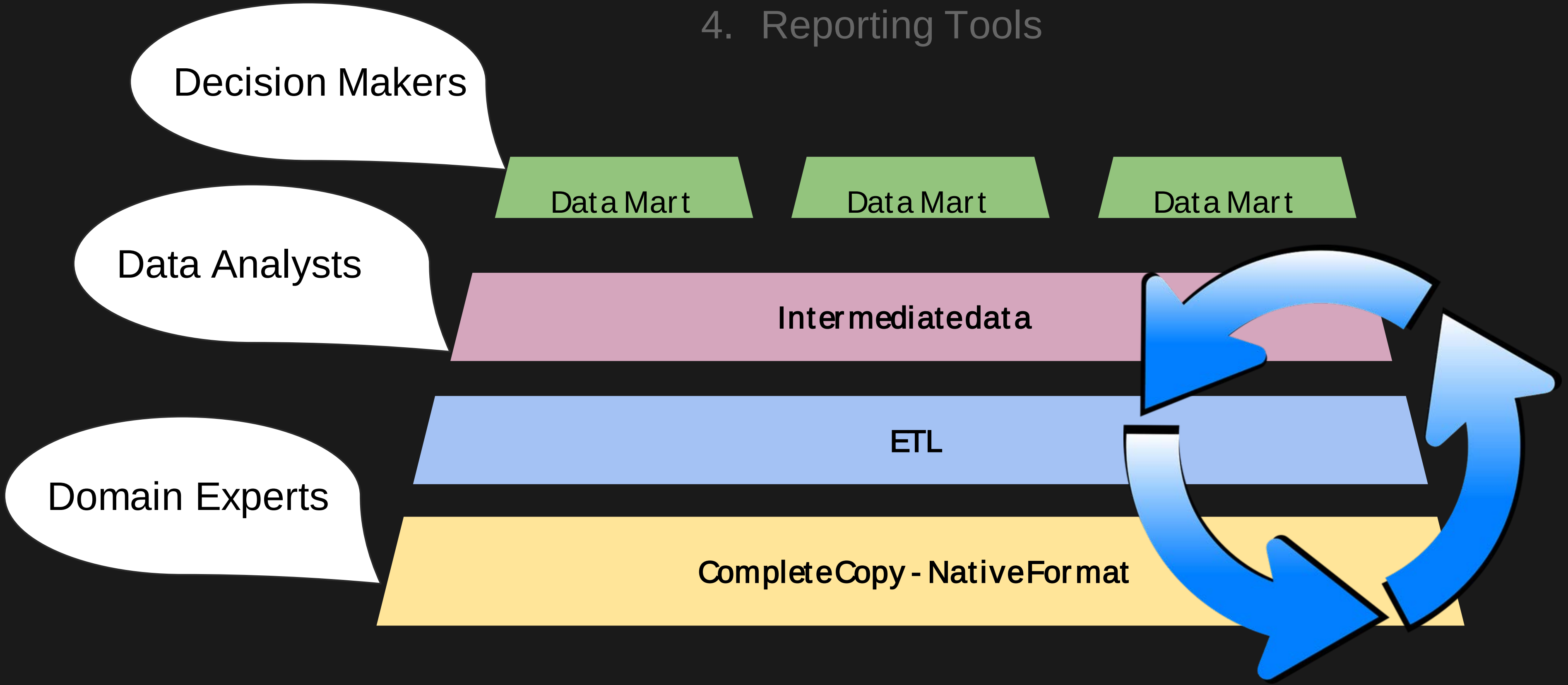
The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools



The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools



The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

Often overlooked but Critical to the success of a data warehouse

The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

Often overlooked but Critical to the success of a data warehouse

The purpose of the data warehouse is to empower end users

The How

1. Multiple sources of data
2. Data Transformation/Normalization
3. Sub-aggregate/interpret based on Use Case
4. Reporting Tools

Typically referred to as “Business Intelligence”

Applications:

Tableau, Sisense, MS Power BI, Crystal Reports, Excel, etc.

References

Evans, R. S., Lloyd, J. F., & Pierce, L. A. (2012). Clinical Use of an Enterprise Data Warehouse. AMIA Annual Symposium Proceedings, 2012, 189–198.

Foran, D. J., Chen, W., Chu, H., Sadimin, E., … DiPaola, R. S. (2017). Roadmap to a Comprehensive Clinical Data Warehouse for Precision Medicine Applications in Oncology. Cancer Informatics

Bae, C. J., Griffith, S., Fan, Y., Dunphy, C., Thompson, N., Urchek, J., … Katzan, I. L. (2015). The Challenges of Data Quality Evaluation in a Joint Data Warehouse. eGEMs, 3(1), 1125.

Elliott, T. E., Holmes, J. H., Davidson, A. J., … Steiner, J. F. (2013). Data Warehouse Governance Programs in Healthcare Settings: A Literature Review and a Call to Action. EGEMS, 1(1), 1010. http://doi.org/10.13063/2327-9214.1010

Krasowski, M. D., Schriever, A., Mathur, G., Blau, J. L., ... & Ford, B. A. (2015). Use of a data warehouse at an academic medical center for clinical pathology quality improvement, education, and research. Journal of Pathology Informatics, 6, 45.

Feldman, D. (2018, March 20). Data Lakes, Data Hubs, Federation: Which One Is Best? Retrieved May 8, 2018, from https://www.marklogic.com/blog/data-lakes-data-hubs-federation-one-best/

Questions?

Thank you

Chris Williams, MD

christopher-williams@ouhsc.edu

References

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3540441/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5392017/

https://www.ncbi.nlm.nih.gov/pubmed/26290882

https://www.ncbi.nlm.nih.gov/pubmed/25848561

https://www.ncbi.nlm.nih.gov/pubmed/26284156

https://www.marklogic.com/blog/data-lakes-data-hubs-federation-one-best/