About this presentation

This presentation is designed to provide an overview of this course and data warehousing concepts.

Agenda

- About me
- About this class
- Data Warehousing Fundamentals

Fall 2024

About me

Who am I?

Dominic Sellitto, CISSP

Clinical Assistant Professor
Director, MS BA Program

About Me

Education:

- Bachelor of Science, Business Administration
- Master of Science, Management Information Systems

Experience:

- Consultant/Senior Consultant, Cyber Risk services, Deloitte
- Lead Cybersecurity Consultant, Loptr LLC

Professional affiliations:

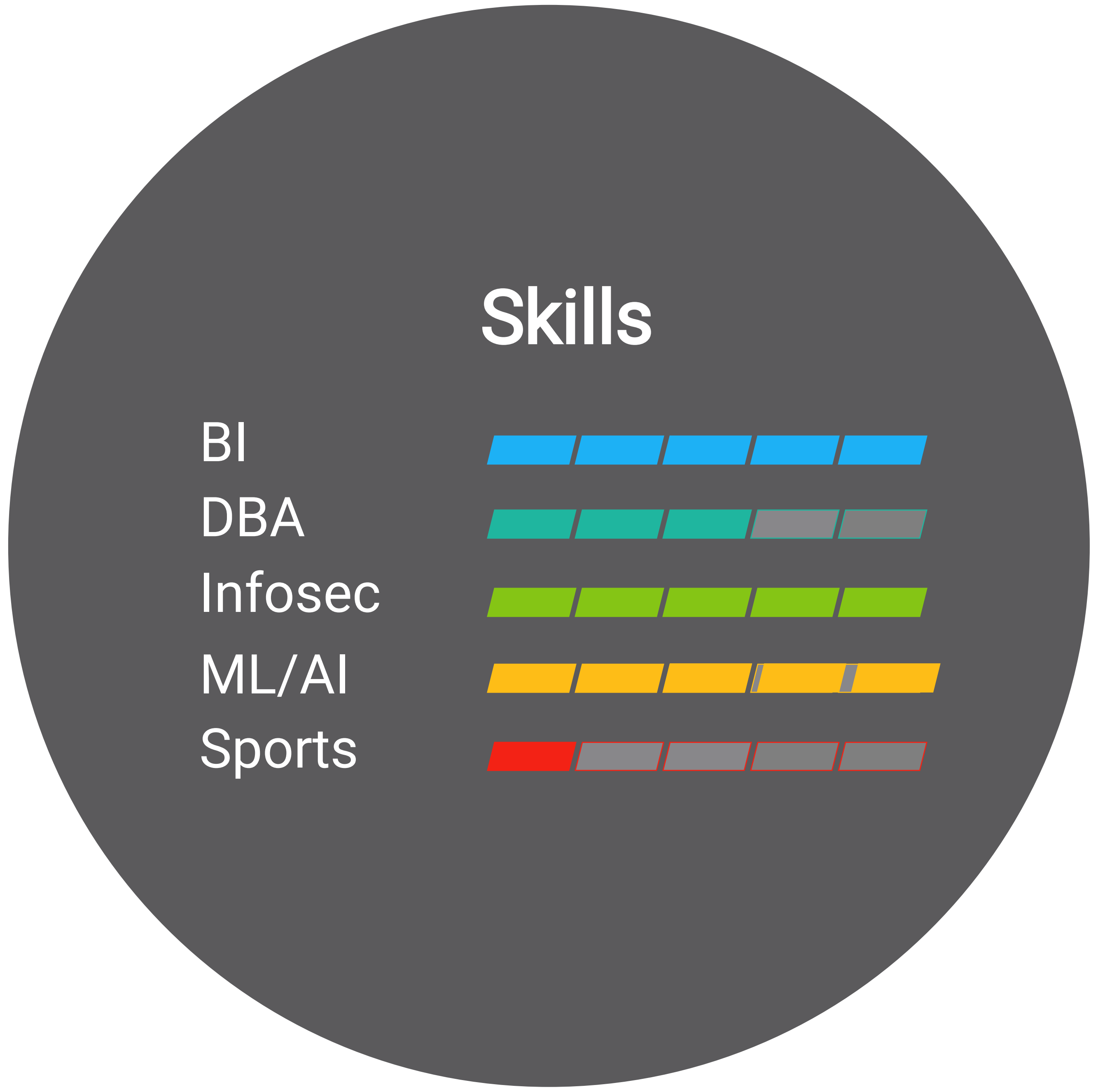
- ISC^2; Certified Information Systems Security Professional (CISSP)

Publications:

- An Investigation into the Role of Cybersecurity Professionals in Shaping AI Integration and Strategy (AMCIS, 2024)
- Invisible Threats: Accessing the Heart of Machine Learning (Hidden Layer, 2023)
- Vulnerability Assessment (ISACA, 2017)

Hats worn:

- CEO
- Chief Information Security Officer
- Project Manager
- Security Analyst
- Security Architect
- Consulting Manager
- Software Developer



About this class

My philosophy

You will get out of this class exactly what you put into it…

To get the most out of it:

- Come to class, stay for the whole thing
- Complete assignments on time
- Ask questions
- Contribute equally to groups
- Be respectful to guest speakers and classmates

About this class

Frequently asked questions

Here are some frequently asked questions and their respective answers:

1. There are exams!?
   - Yeah, there are exams. Bummer. They should be easy if you’ve been paying attention.
2. What if I can’t finish an assignment on time?
   - Late assignments won’t be accepted. Special circumstances will be handled per University policy.
3. Is there a TA?
   - Yes, our TA is Victoria Gonzalez and her contact information is in the syllabus
   - Victoria will have office hours, as will I
4. How can we contact you?
   - The best way to reach me and Victoria is by email. We’ll respond as quickly as we can (best to CC me on all emails).
   - Second best is to schedule an appointment with Victoria or me for office hours.
   - You can also hang out before/after class

About this class

Areas of focus

This class was developed around 4 high-level areas of focus…

Hands-on experience	We'll focus on **practical** knowledge and **hands-on** exercises
Understand the why	We'll explore **why** organizations implement Data Warehouses
Hone soft-skills	We'll learn that **communicating** is just as important as doing
Connect with experts	We'll hear **different viewpoints** from industry professionals

About this class

Assignments and exams

Throughout the semester, there are various assignments and exams that you will be expected to complete…

Critical responses (3)	Write-ups on a specific topic, formatted for a specific audience
Lab assignments (5)	Hands-on activities to teach specific tools and techniques
Course readings (+/-)	There’s a book, and I’ll assign sections of it each class
Exams (2)	One for the first half of the course and the other for the second half

About this class

Things you will need

In order to follow the course content, complete the labs, and follow along with in-class demos, you will need:

- A computer (Windows or Mac is fine, but you must be able to download and run software using Java)
- The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, 3rd Edition (Ralph Kimball ISBN-13: 978-1118530801, ISBN-10: 1118530802)
    - **Protip:** UB’s library has an e-copy of this available online for free.

What is a Data Warehouse?

“A *data warehouse* is a type of data management system that is designed to enable and support business intelligence (BI) activities, especially analytics. Data warehouses are solely intended to perform queries and analysis and often contain large amounts of historical data.”

-Oracle (https://www.oracle.com/database/what-is-a-data-warehouse)

Okay… but what does that **mean?**

A data warehouse is a specialized database designed to provide data analysis capabilities to **the business**..

It’s all about the business.

What *isn't* a data warehouse?

Transactional Systems, or OLTP (Online Transactional Processing Systems)

OLTP systems should not be considered data warehouses.

Why?

- OLTP systems, like sales/invoice databases, are often primary operation systems, and analytical queries can cause performance issues
- OLTP systems are normalized (more on this later) to reduce redundancy and increase consistency, which makes them highly un-friendly for ad-hoc queries, and makes it difficult for business users to interact with data
- OLTP systems typically don't store data for long periods of time (e.g., years), and only keep what is necessary for operations.

What *isn't* a data warehouse?

Data Lakes

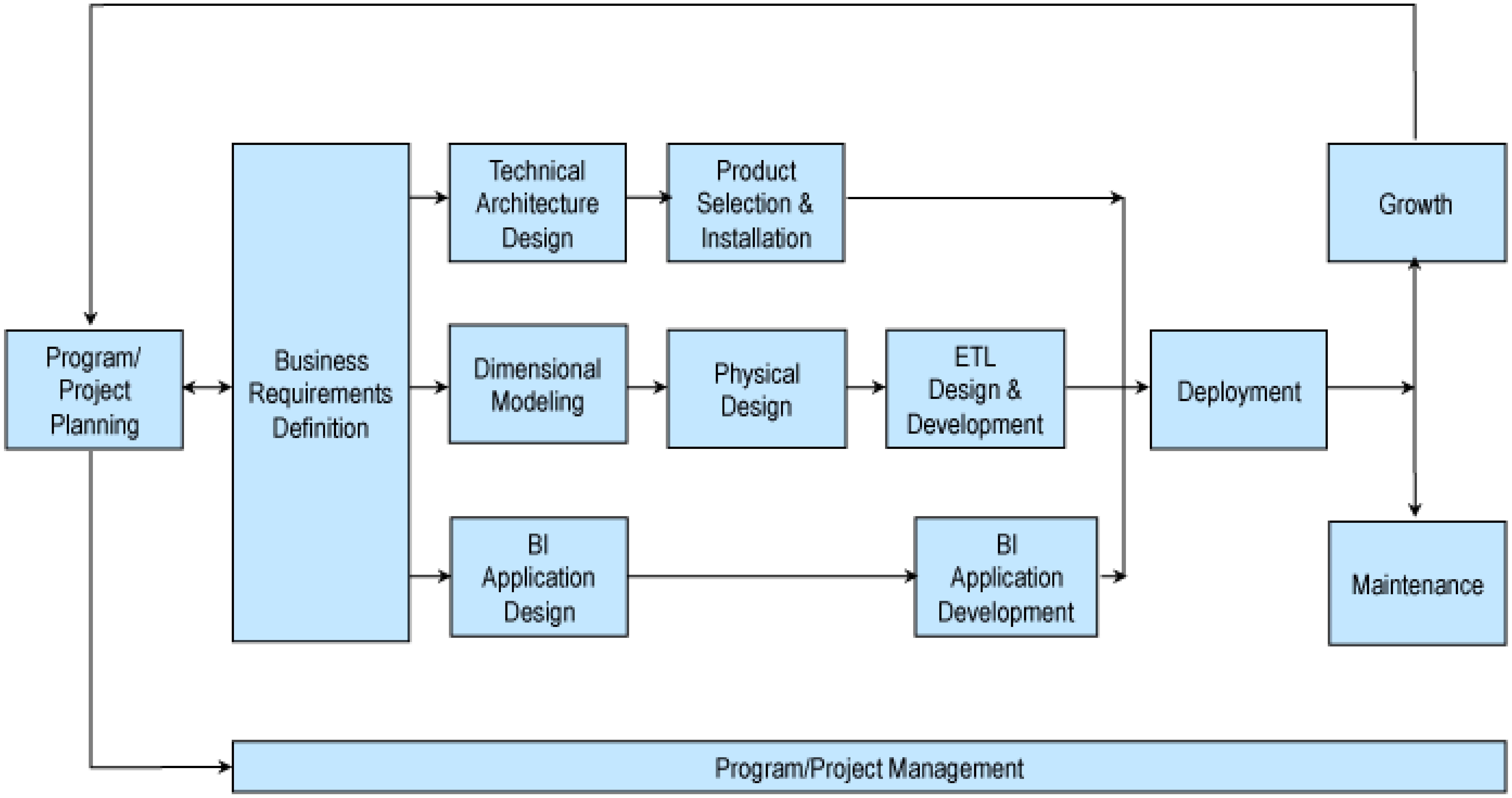
Data lakes are becoming more popular and are often confused with data warehouses.

Why isn't a data lake a data warehouse?

A data lake is designed to incorporate raw data from a variety of sources. It is unformatted and unstructured and can be searched and readied for analysis when needed.

A data warehouse contains data that is structured and formatted– put another way, data that is already ready for analysis.

The data warehouse lifecycle


Program/ Project Planning
Business Requirements Definition
Technical Architecture Design
Product Selection & Installation
Dimensional Modeling
Physical Design
ETL Design & Development
BI Application Design
BI Application Development
Deployment
Growth
Maintenance
Program/Project Management

Reference: https://www.kimballgroup.com

Success criteria

A successful data warehouse environment should meet each of the following requirements:

Accessibility	Data in the warehouse must be intuitive and understandable..
Consistency	Data in the warehouse must utilize a common language.
Adaptability	The warehouse must be resilient to changing business requirements.
Timeliness	New data must be added to the warehouse at regular intervals.
Security	Data must be secured from unauthorized access and modification.
Trustworthiness	Data in the warehouse must have the right data for the business.
Acceptance	The warehouse and its outputs must be accepted by the business.

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013.

Translating into tasks…

A data warehouse manager must:

1. Understand the business users
2. Deliver high-quality, relevant, and accessible information and analytics to those users
3. Maintain the environment

Reference: Kimball, Ralph, and Margy Ross. *The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling*, John Wiley & Sons, Incorporated, 2013.

Some other terms…

Don’t worry, we’ll come back to each of these…

Data Mart: A data warehouse scoped for a specific business unit or function.

Dimensional Modeling: Widely accepted and popular methodology for developing data warehouse environments.

Star Schema: The application of dimensional modeling to create a database schema that resembles a star (or, hub and spoke), typically implemented on a standard RDBMS platform

OLAP Cube: The application of dimensional modeling to create a multidimensional cube for data analysis purposes.

Normalization: A relational database design methodology that reduces data redundancy and improves data consistency.

Reference: https://www.talend.com/resources/what-is-data-mart/

Data Warehouse Methodologies

The Kimball Method: Leverages a bottom-up approach for data warehouse design characterized by denormalization and the concepts of facts and dimensions– typically implemented using a Star Schema approach.

The Inmon Method: Leverages a top-down approach for design characterized by normalized database structures and avoids data redundancy considerations of de-normalized databases.

In this course, we're going to focus on the Kimball method. ***Why?*** It is widely accepted due to its performance benefits and simplicity of use by the business. While the Inmon Method might be more familiar to those with DBA experience, many Inmon warehouses are complimented by dimensional data marts to provide simplicity and efficiency for end users.

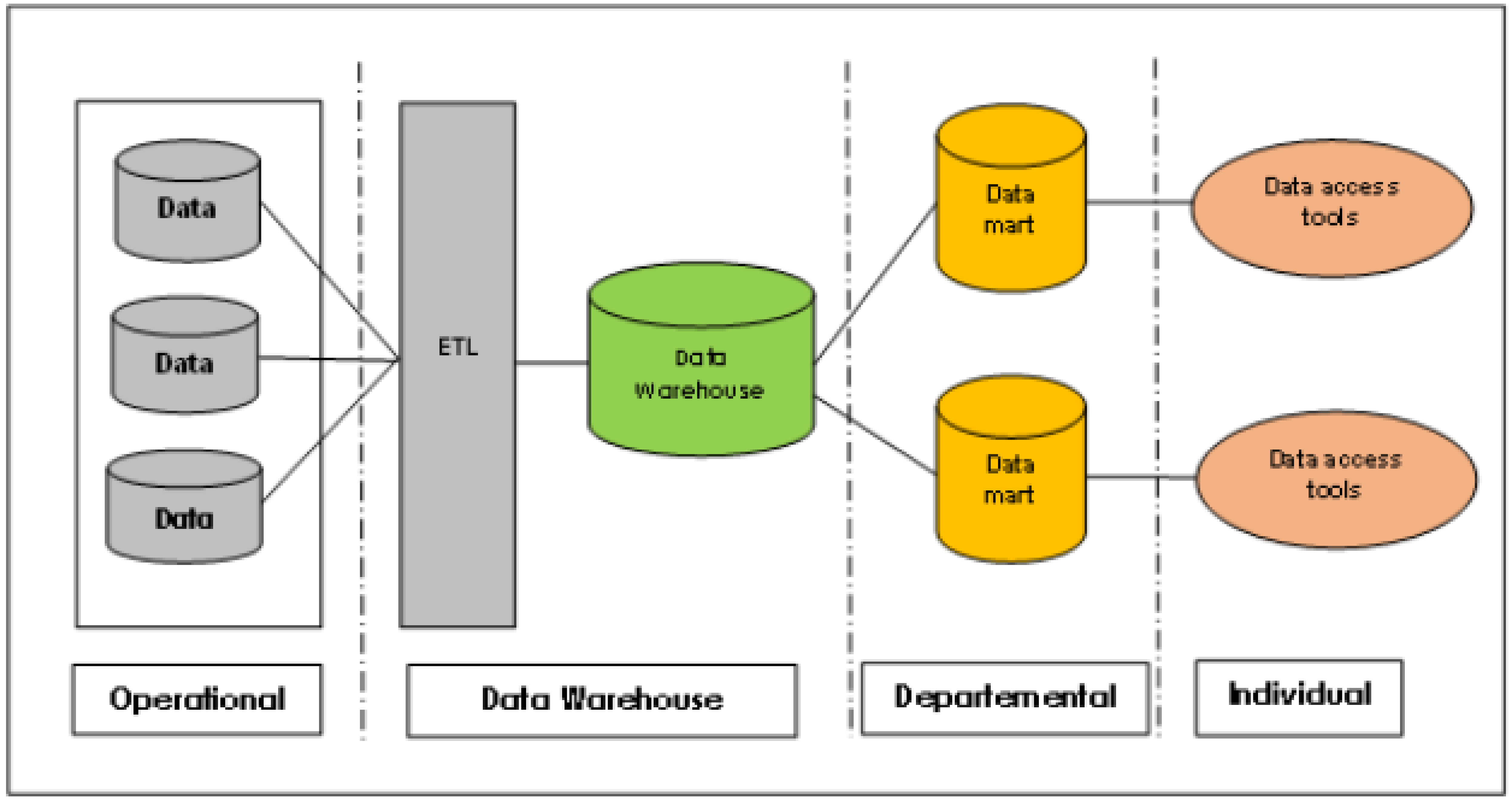
Reference: https://www.astera.com/type/blog/data-warehouse-concepts/

Data Warehouse Methodologies

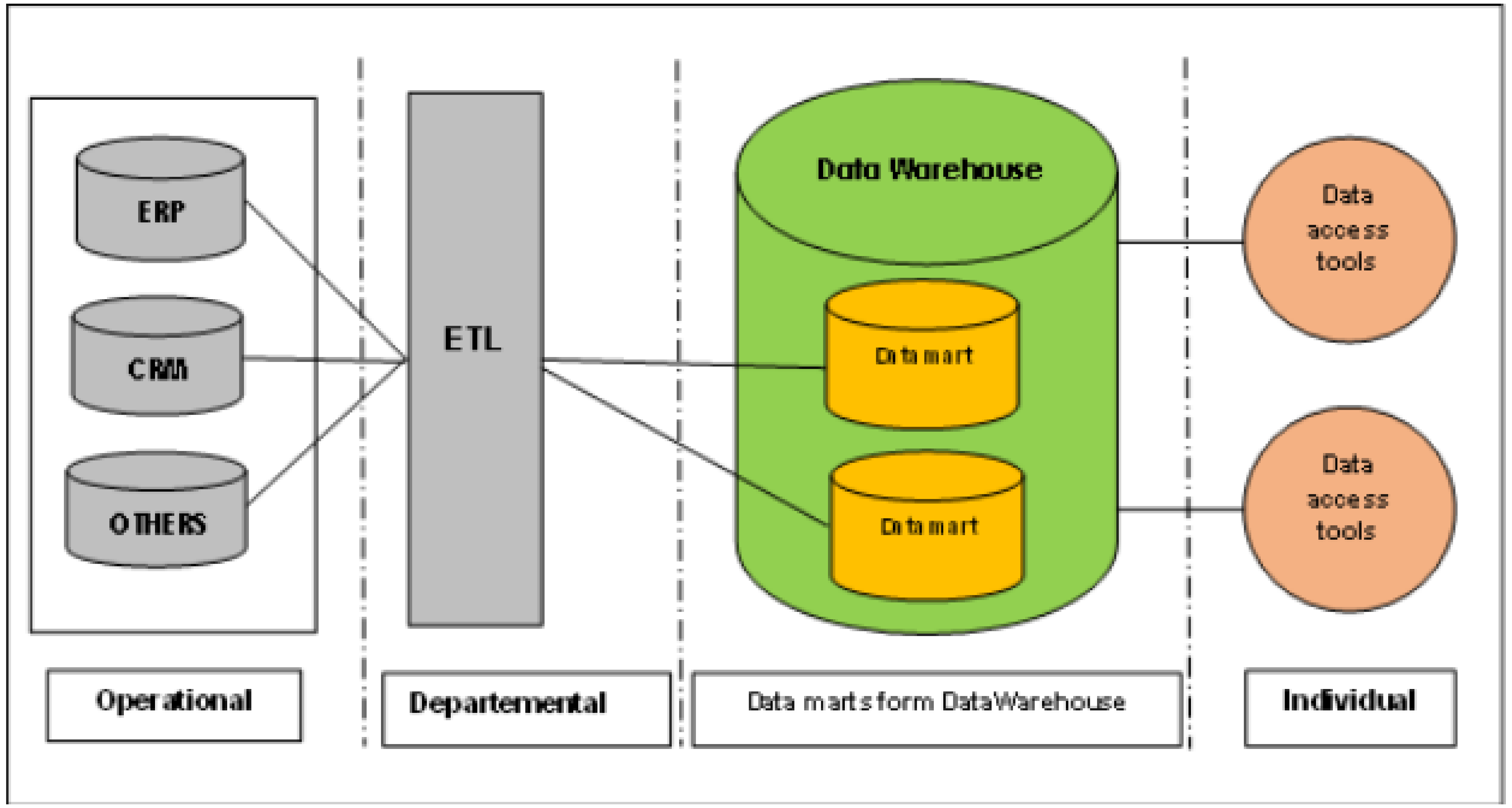
Characteristics	Favours Kimball	Favours Inmon
Business decision support requirements	Tactical	Strategic
Data integration requirements	Individual business requirements	Enterprise-wide integration
The structure of data	KPI, business performance measures, scorecards…	Data that meet multiple and varied information needs and non-metric data
Persistence of data in source systems	Source systems are quite stable	Source systems have a high rate of change
Skill sets	A small team of generalists	Bigger team of specialists
Time constraint	Urgent needs for the first data warehouse	Longer time is allowed to meet business needs.
Cost to build	Low start-up cost	High start-up costs

https://www.zentut.com/data-warehouse/kimball-and-inmon-data-warehouse-architectures/

Data Warehouse Methodologies



Inmon



Kimball

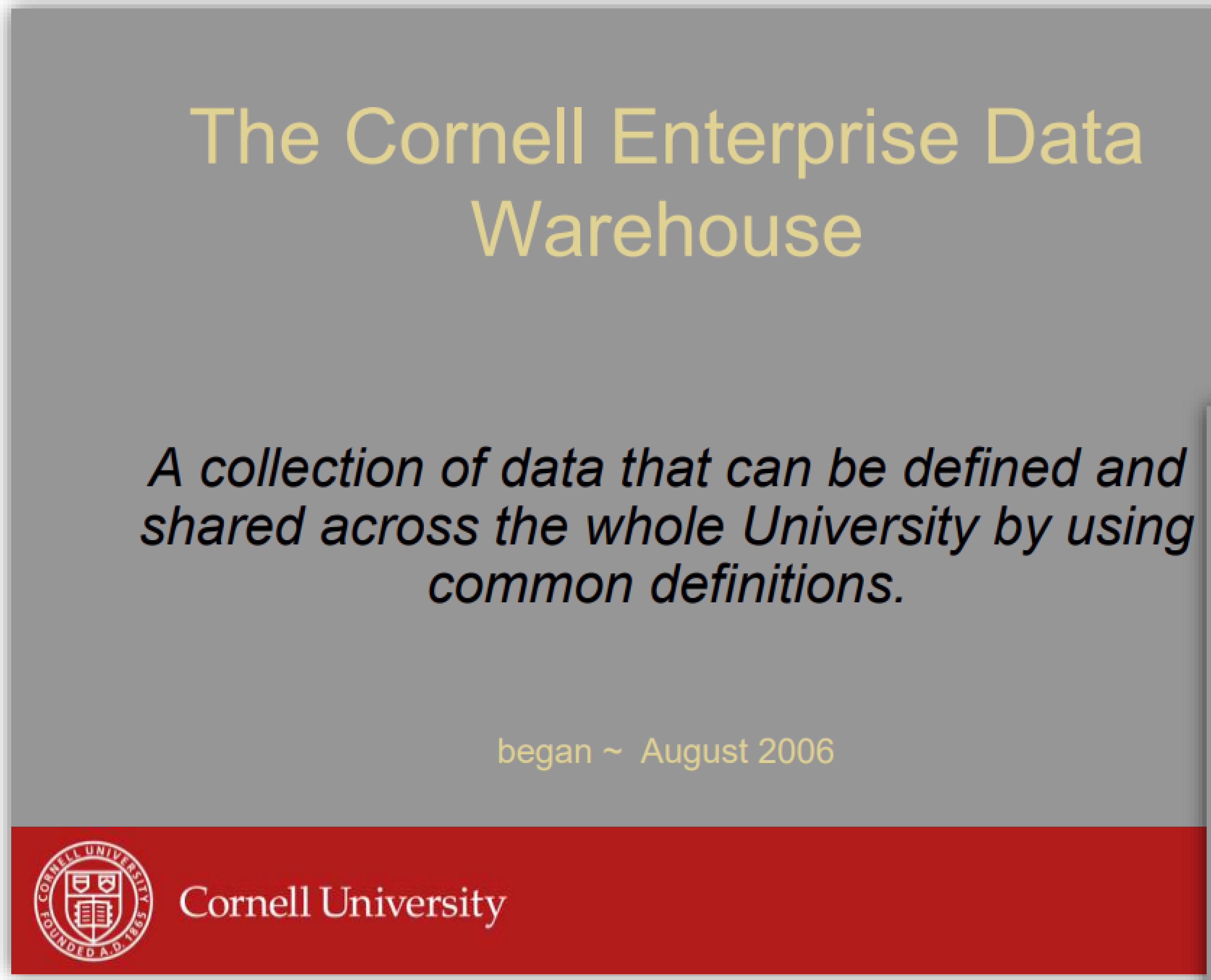
https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7815845

Data Warehouse Example #1: Walmart

“Wal-Mart’s data warehouse, the first commercial EDW to reach 1 terabyte of data in 1992, began, like many good things, as an accident. One of the retailer’s computer operators, tired of retrieving archival tapes for historical sales data, secretly ‘borrowed’ excess storage space on a company server, where he downloaded and stored the data from the most-requested tapes… Soon, every transaction in 6,000 Wal-Mart stores was available for analysis in the data warehouse within seven minutes. This treasure trove of data enabled Wal-Mart to react in near real-time to sales and supply data.”

Reference: https://www.healthcatalyst.com/wal-mart-birth-of-data-warehouse/

Data Warehouse Example #2: Cornell University

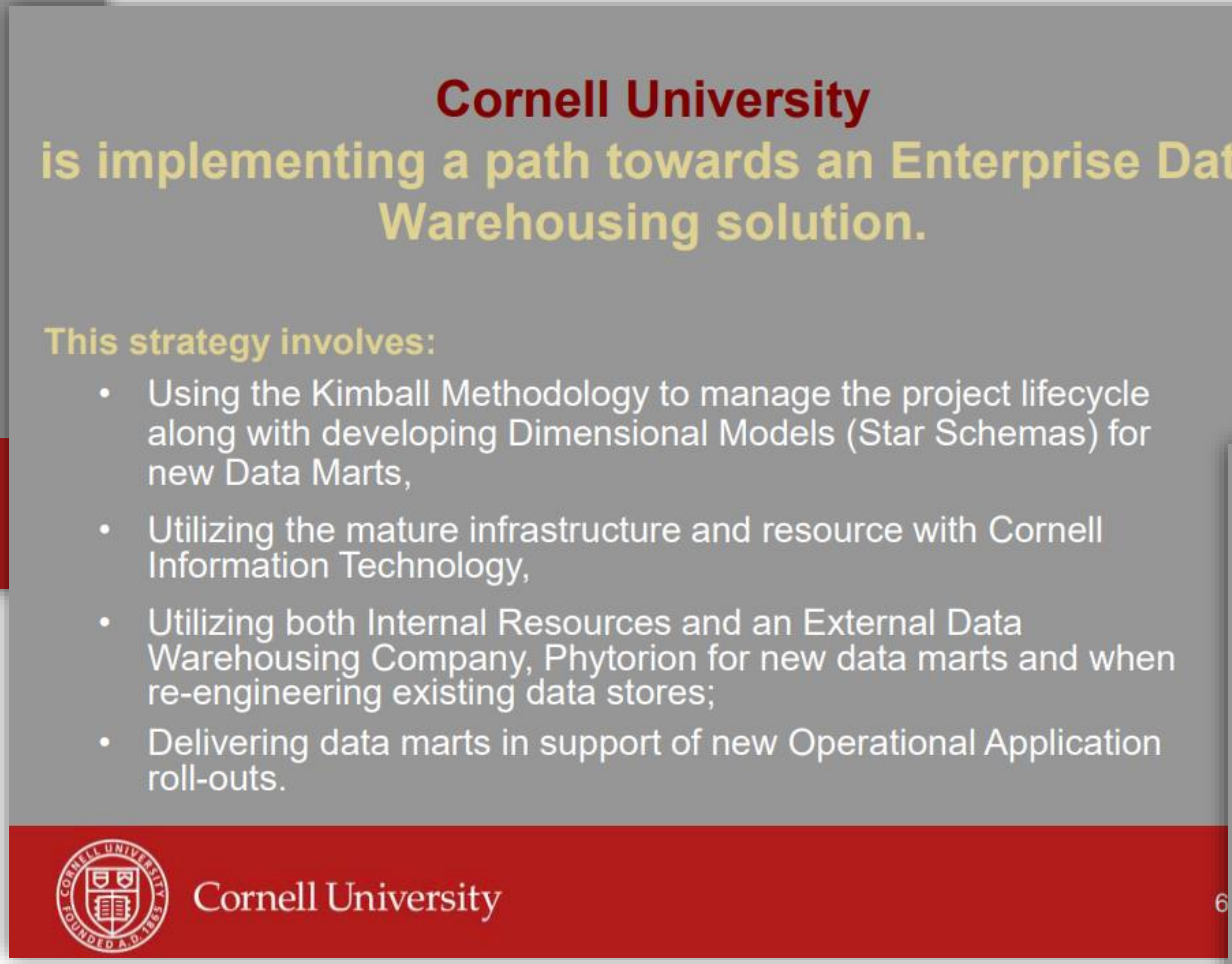


The Cornell Enterprise Data Warehouse

A collection of data that can be defined and shared across the whole University by using common definitions.

began ~ August 2006

Cornell University

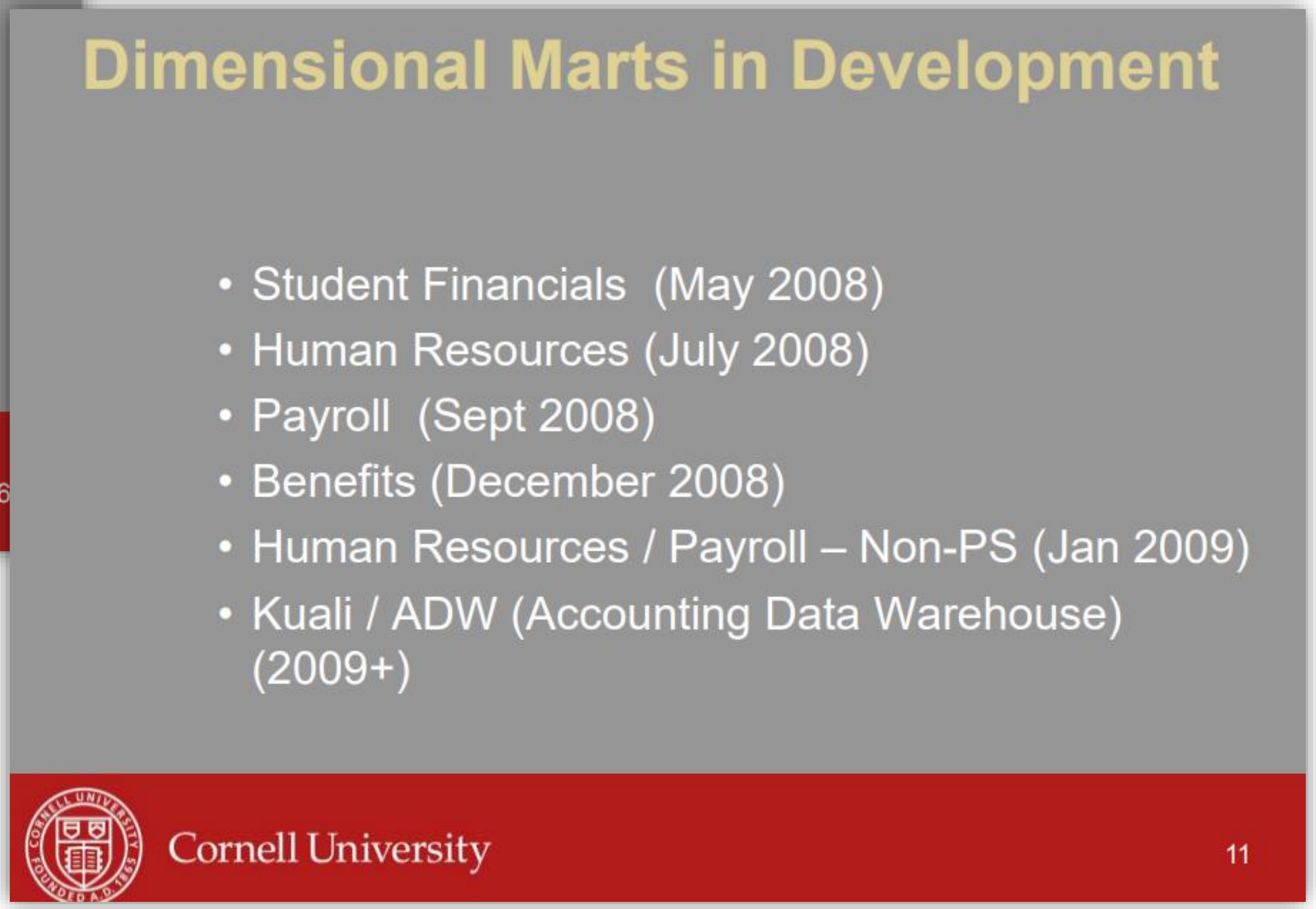


Cornell University is implementing a path towards an Enterprise Data Warehousing solution.

This strategy involves:

- Using the Kimball Methodology to manage the project lifecycle along with developing Dimensional Models (Star Schemas) for new Data Marts,
- Utilizing the mature infrastructure and resource with Cornell Information Technology,
- Utilizing both Internal Resources and an External Data Warehousing Company, Phytorion for new data marts and when re-engineering existing data stores;
- Delivering data marts in support of new Operational Application roll-outs.

Cornell University



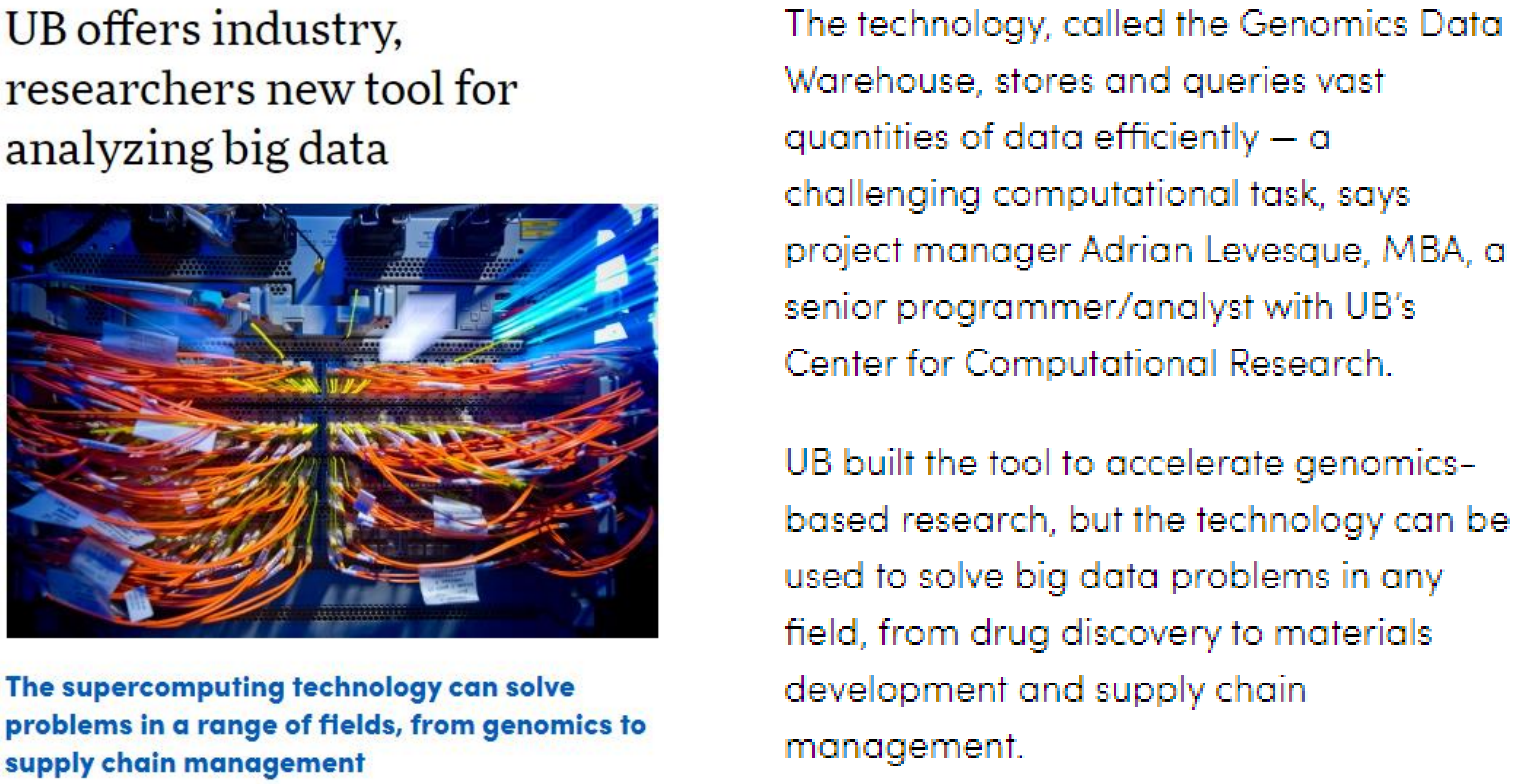
Dimensional Marts in Development

- Student Financials (May 2008)
- Human Resources (July 2008)
- Payroll (Sept 2008)
- Benefits (December 2008)
- Human Resources / Payroll – Non-PS (Jan 2009)
- Kuali / ADW (Accounting Data Warehouse) (2009+)

Cornell University

Reference: https://phytorion.com/_media/pdf/Cornell%20Presentation,%20HEDW%20Conference%202008.pdf

Data Warehouse Example #3: UB



UB offers industry, researchers new tool for analyzing big data

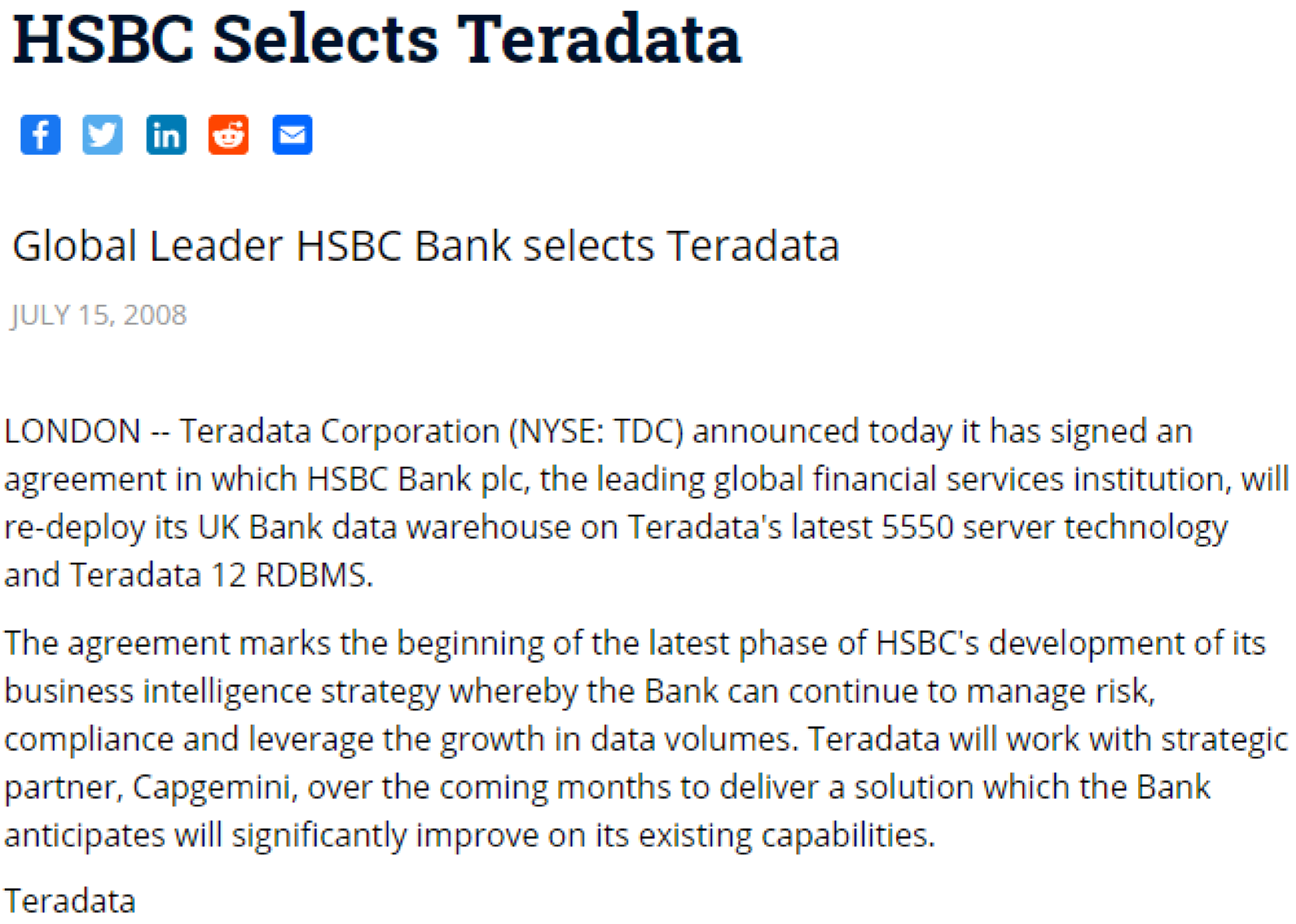
The supercomputing technology can solve problems in a range of fields, from genomics to supply chain management

The technology, called the Genomics Data Warehouse, stores and queries vast quantities of data efficiently — a challenging computational task, says project manager Adrian Levesque, MBA, a senior programmer/analyst with UB's Center for Computational Research.

UB built the tool to accelerate genomics-based research, but the technology can be used to solve big data problems in any field, from drug discovery to materials development and supply chain management.

Reference: https://www.buffalo.edu/ccr/about-us/news-events/latest_news.host.html/content/shared/www/ccr/ccr-news/ccrbigreleasegenomicsdatawarehouse.detail.html

Data Warehouse Example #4: HSBC



HSBC Selects Teradata

Global Leader HSBC Bank selects Teradata

JULY 15, 2008

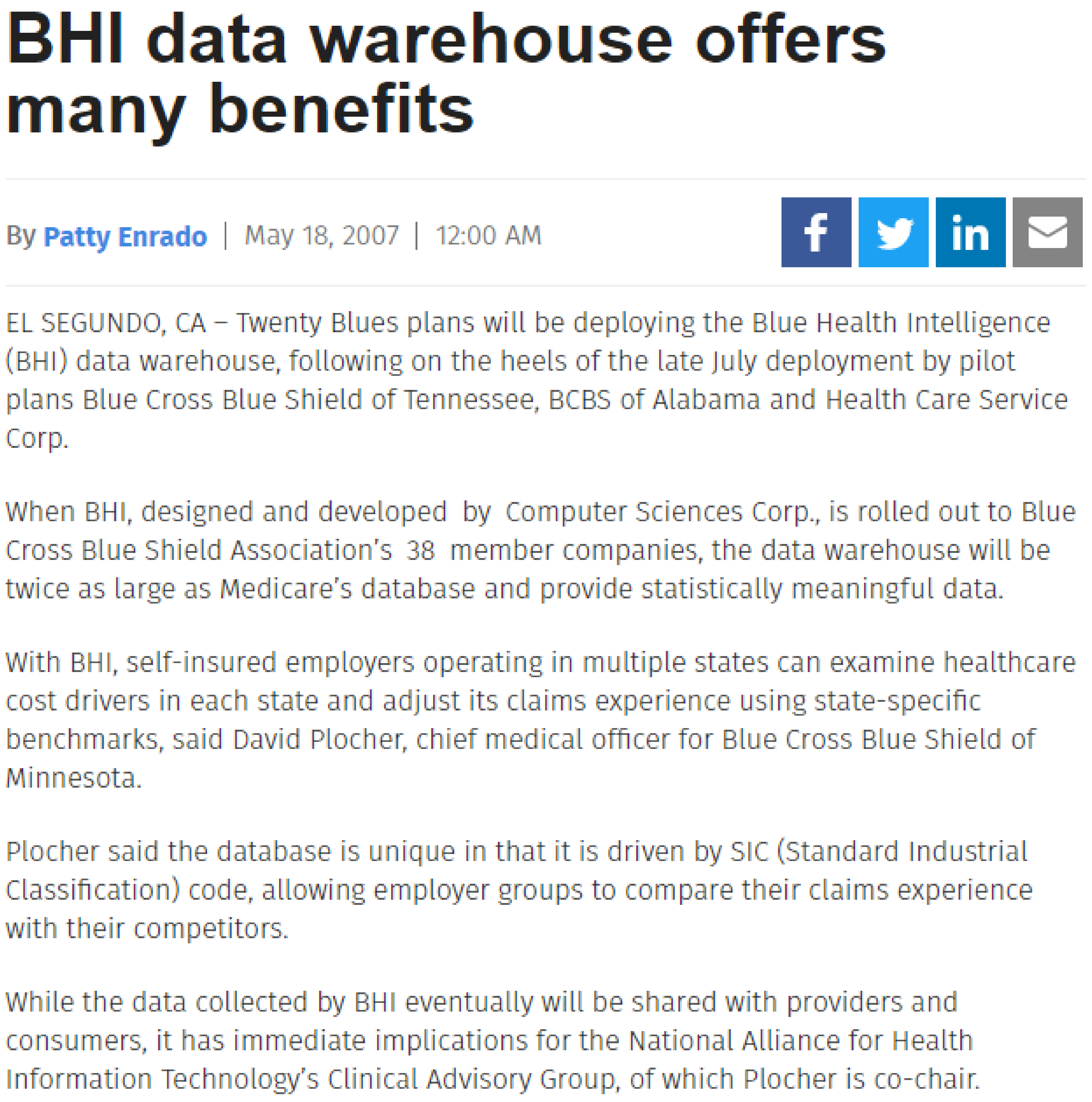
LONDON -- Teradata Corporation (NYSE: TDC) announced today it has signed an agreement in which HSBC Bank plc, the leading global financial services institution, will re-deploy its UK Bank data warehouse on Teradata's latest 5550 server technology and Teradata 12 RDBMS.

The agreement marks the beginning of the latest phase of HSBC's development of its business intelligence strategy whereby the Bank can continue to manage risk, compliance and leverage the growth in data volumes. Teradata will work with strategic partner, Capgemini, over the coming months to deliver a solution which the Bank anticipates will significantly improve on its existing capabilities.

Teradata

Reference: https://www.networkcomputing.com/careers-and-certifications/hsbc-selects-teradata

Data Warehouse Example #5: Blue Cross Blue Shield



BHI data warehouse offers many benefits

By Patty Enrado | May 18, 2007 | 12:00 AM

EL SEGUNDO, CA – Twenty Blues plans will be deploying the Blue Health Intelligence (BHI) data warehouse, following on the heels of the late July deployment by pilot plans Blue Cross Blue Shield of Tennessee, BCBS of Alabama and Health Care Service Corp.

When BHI, designed and developed by Computer Sciences Corp., is rolled out to Blue Cross Blue Shield Association's 38 member companies, the data warehouse will be twice as large as Medicare's database and provide statistically meaningful data.

With BHI, self-insured employers operating in multiple states can examine healthcare cost drivers in each state and adjust its claims experience using state-specific benchmarks, said David Plocher, chief medical officer for Blue Cross Blue Shield of Minnesota.

Plocher said the database is unique in that it is driven by SIC (Standard Industrial Classification) code, allowing employer groups to compare their claims experience with their competitors.

While the data collected by BHI eventually will be shared with providers and consumers, it has immediate implications for the National Alliance for Health Information Technology's Clinical Advisory Group, of which Plocher is co-chair.

Reference: https://www.healthcareitnews.com/news/bhi-data-warehouse-offers-many-benefits