TN03 008 : Enterprise Data Management
(การบริหารจัดการข้อมูลองค์กร)

Noppol Thangsupachai, Ph.D.
noppol@sut.ac.th

อบรมหลักสูตร Upskill-Reskill มหาวิทยาลัยเทคโนโลยีสุรนารี

Digital Arts & Science
DIGITECH

มหาวิทยาลัยเทคโนโลยีสุรนารี

Suranaree University of Technology

TN03 008 Enterprise Data Management (การบริหารจัดการองค์กร)

อบรมหลักสูตร Reskill-Upskill

โดย :

อาจารย์ ดร.นพพล ตั้งสุภาชัย

Noppol Thangsupachai, Ph.D.

noppol@sut.ac.th

กลุ่มวิชาศาสตร์และศิลป์ดิจิทัล (Digital Arts and Science)

โครงการจัดรูปการบริหารวิชาการด้านเทคโนโลยีดิจิทัลรูปใหม่

Digital Arts & Science DIGITECH

3

Data Management Component

การบริหารจัดการข้อมูล

Data lifecycle phases:

- Data capture - Create
- Data Integrate
- Data Trasform

  Store
- Data usage - Use
- Data publication -Share
- Data archival - Archive
- Data purging - Destroy

Infrastructure and Data Lifecycle, 2019

https://roaringelephant.org/2019/01/15/episode-123-infrastructure-and-data-lifecycle-part-2/

Components of Enterprise Information Management

- Data Sourcing
- Data Integration and Exchange
- Data Governance and Quality
- Data Architecture and Models
- Master Information Management
- Metadata Management

3.1

Data Sourcing

6

Data Sourcing

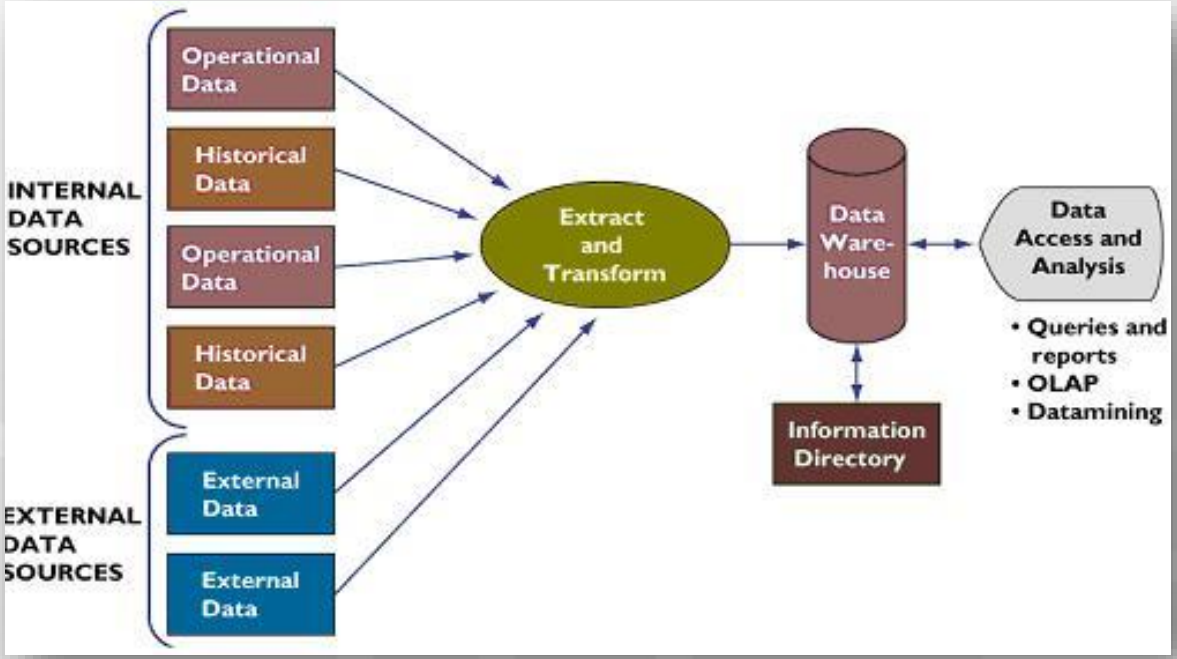
- กำหนดแหล่งข้อมูลที่มีความสำคัญต่อองค์กร
- รูปแบบ ลักษณะ ที่มาของข้อมูล
- กำหนดวิธีการสกัด (Extract) ข้อมูล เพื่อองค์กรสามารใช้งานเมื่อต้องการ
- กำหนดนโยบายในการนำเข้า แปลง (Transform) และจัดเก็บข้อมูล ในคลังข้อมูล (Load)

Building A Data Landscape, 2013
https://online-behavior.com/analytics/data-landscape

Data Sourcing (cont.)

แหล่งข้อมูลที่องค์กรสามารถนำมาใช้ในการดำเนินงานประกอบด้วย

- แหล่งข้อมูลภายในองค์กร
  - ระบบสารสนเทศในองค์กร
  - ข้อมูลการปฏิบัติงาน
  - ระบบบันทึกข้อมูลการทำงานอัตโนมัติ
- แหล่งข้อมูลภายนอกองค์กร
  - ข้อมูลผู้ที่เกี่ยวข้องกับองค์กร
  - ข้อมูลจากสื่อสังคมออนไลน์



DATA WAREHOUSE + DATA MINING, 2013
https://9chooknow.blogspot.com/2013/03/data-warehouse-data-mining.html

Data Sourcing (cont.)

Business requirement mapping to source systems

- KPI dimension matrix
  - Profile Source Systems for Relevant Datasets
  - Define Source Extract Mechanisms
  - Provide Source Extract Files for Information Integration:
    - structure data file
    - naming convention data heading
    - frequency of generate data
    - mode of delivery

	COMMON DIMENSIONS						
BUSINESS PROCESSES	Date	Product	Warehouse	Store	Promotion	Customer	Employee
Issue Purchase Orders	X	X	X				
Receive Warehouse Deliveries	X	X	X				X
Warehouse Inventory	X	X	X				
Receive Store Deliveries	X	X	X	X			X
Store Inventory	X	X		X			
Retail Sales	X	X		X	X	X	X
Retail Sales Forecast	X	X		X			
Retail Promotion Tracking	X	X		X	X		
Customer Returns	X	X		X	X	X	X
Returns to Vendor	X	X		X			X
Frequent Shopper Sign-Ups	X			X		X	X

The Matrix revisited, 2005

https://www.kimballgroup.com/2005/12/the-matrix-revisited/

Data Sourcing (cont.)

Table 3-2. *Key Differences Between Push and Pull Mechanisms*

Parameters	Push Mechanism	Pull Mechanism
Nature of extraction	Source system team provides the source data extracts in the interface formats provided by the information integration team.	The information integration team is provided read access to source tables to query and pick up the relevant data sets for further processing.
Source system knowledge	The source system team has extensive knowledge of the source systems and provides the source data extracts as per the interface formats agreed with the information integration team.	The information integration team has to build knowledge of the source system and **extract** the relevant data from the source tables based on the access provided by the source systems team.
Source system changes	In case of push mechanism, there is no impact of source system structure changes as the source system team generates the extract files. The information integration process is insulated from the source system changes.	In case of pull mechanism, the source system structure changes have to be understood by the information integration team and there will be changes to the information integration jobs that access the source systems to pull relevant data.

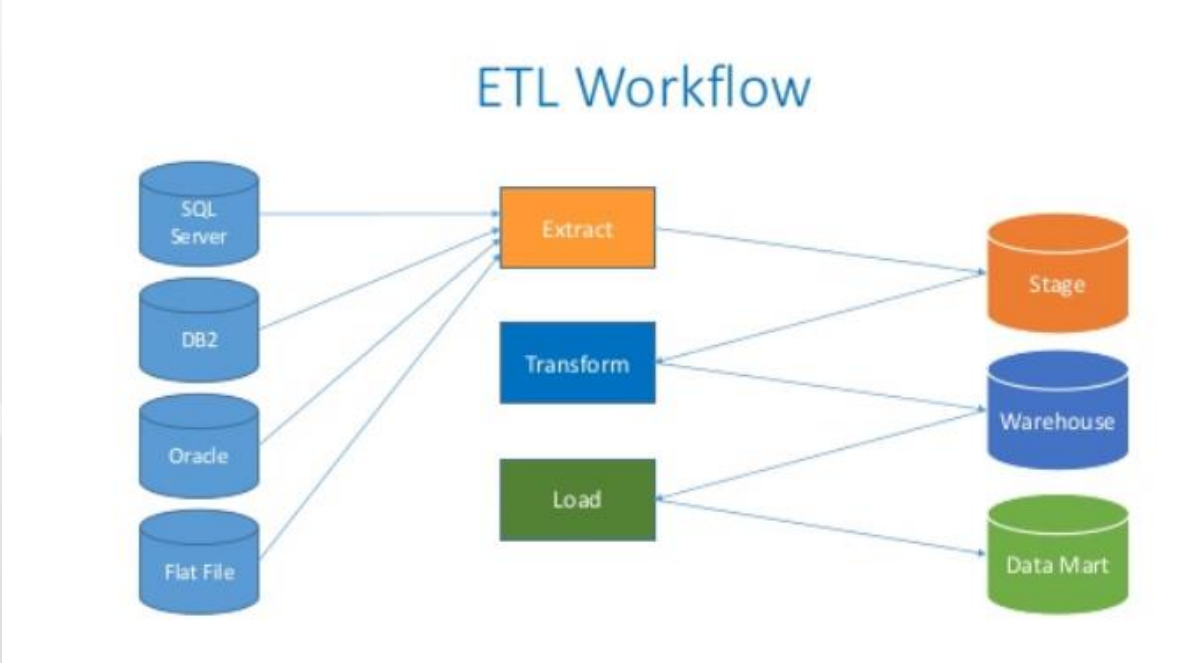
ที่มา:

Enterprise Information Management in Practice: Managing Data and Leveraging Profits in Today's Complex Business Environment, Saumya Chaki, 2015)

Data Sourcing (cont.)

Information Sourcing Patterns and Challenges

- Logical Data Extraction
  - Full extraction
  - Incremental extraction
  - Change data capture
- Physical Data Extraction
- Automated Data Extraction

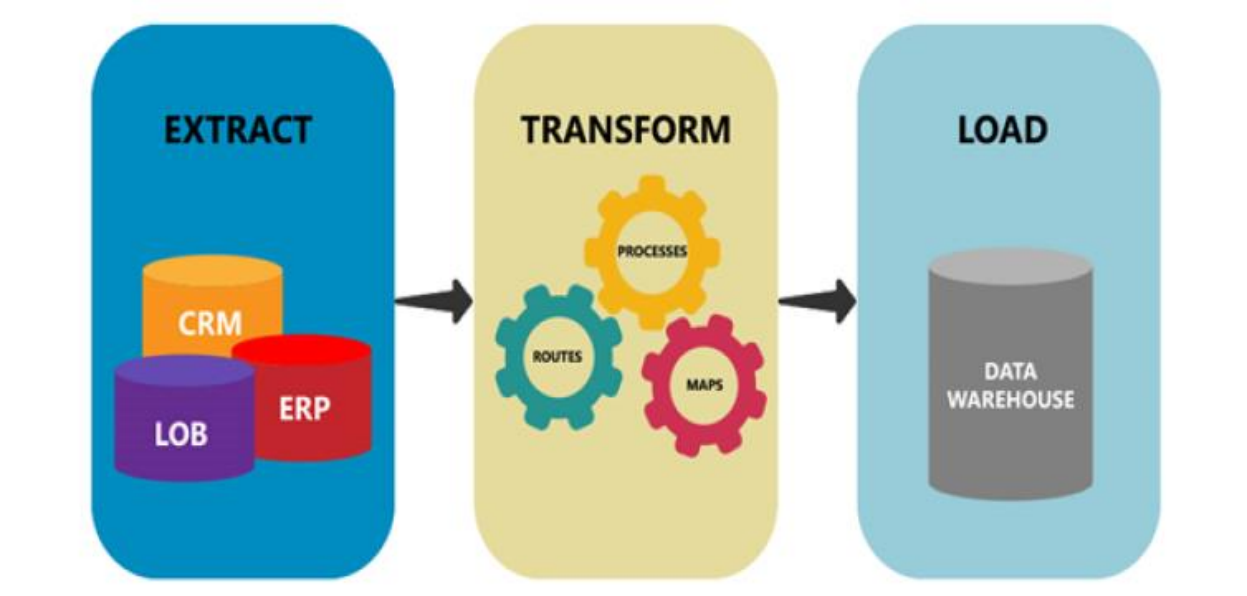


Which Data Extraction Approach is Best for Your Data Warehouse?, 2018
https://datawarehouseinfo.com/data-warehouse-data-extraction-models/

Data Sourcing (cont.)

Information Sourcing Patterns and Challenges

- Data conversion challenges
- Metadata gaps
- Mergers and acquisitions
- Manual data
- Real-time source data extraction

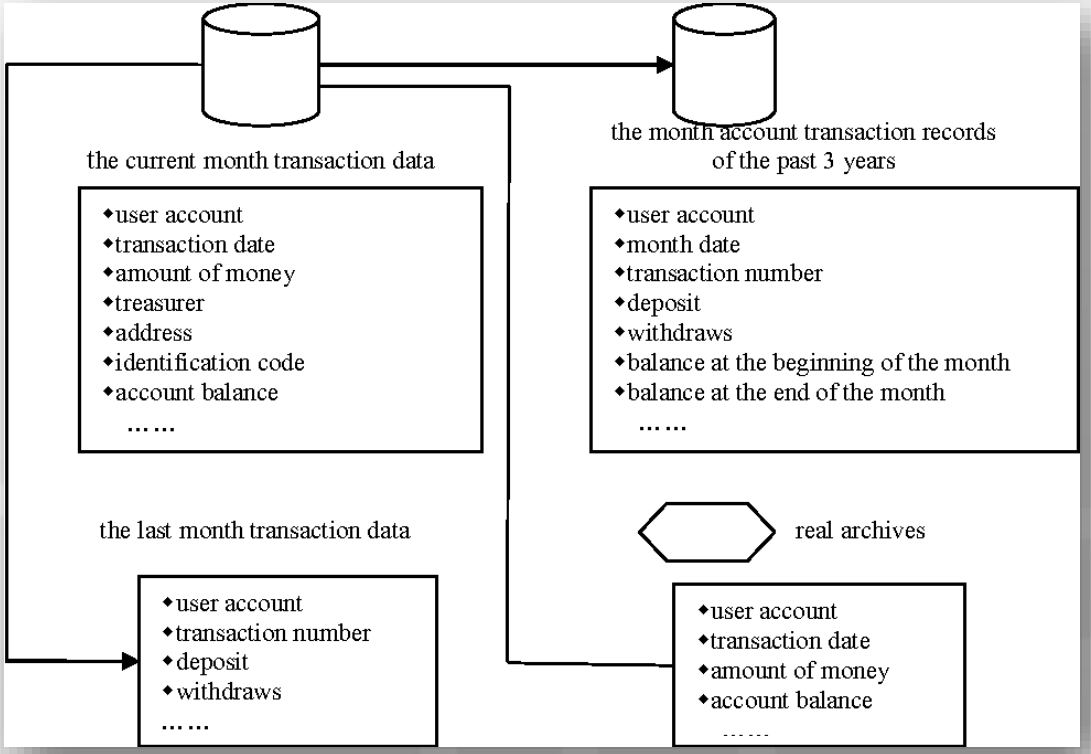


ETL vs. ELT: Transform First or Transform Later?, 2018
https://datawarehouseinfo.com/etl-vs-elt-transform-first-or-transform-later/

Data Sourcing (cont.)

Data Granularity

- Data volumes and storage costs
- Query performance
- Source data availability
- Batch performance impact



Classification of Data Granularity in Data Warehouse, 2017
https://www.semanticscholar.org/paper/Classification-of-Data-Granularity-in-Data-Lv-Zhou/aea746ba5fcdd504c51ace554dc343d55d2c024b

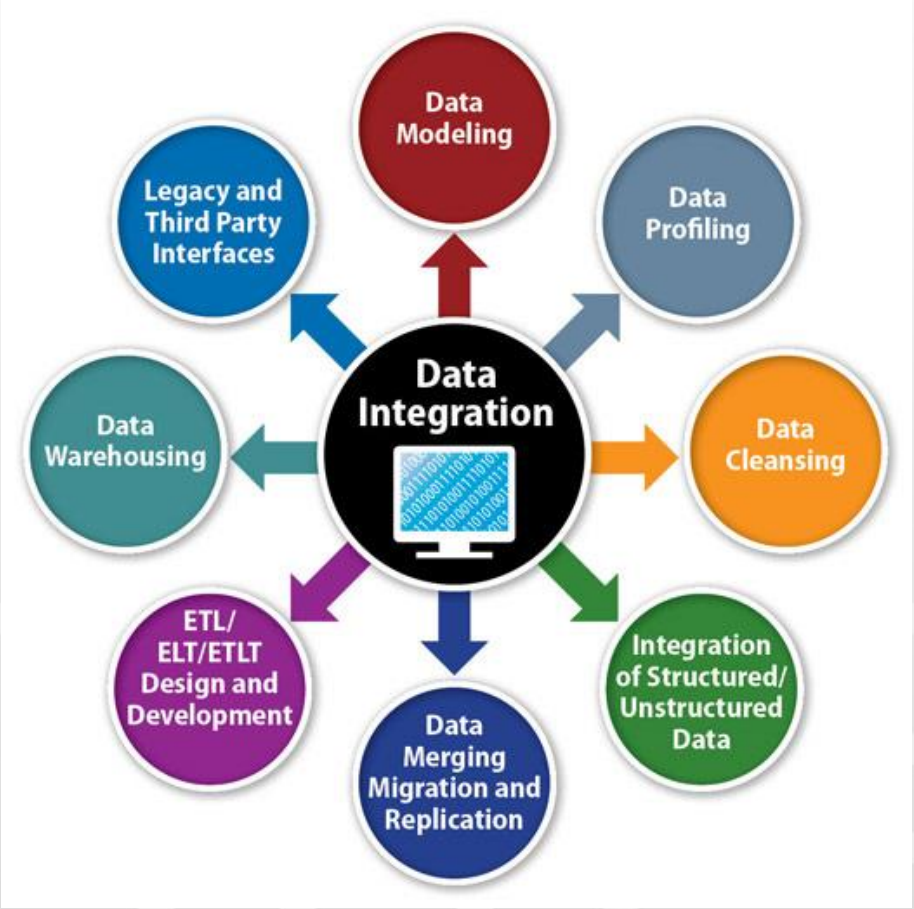
3.2

Data Integration and Exchange

Data Integration and Exchange

key role in determining in data integration strategy

- Nature of extraction
- Type of connecters
- Leverage data integration engine
- Data integration hubs
- Slowly changing dimensions
- Real-time data integration



5 Leading Data Integration Use Cases

https://www.datamation.com/big-data/data-integration-use-cases.html

Nature of extraction (Push/Pull)

Push based integration system

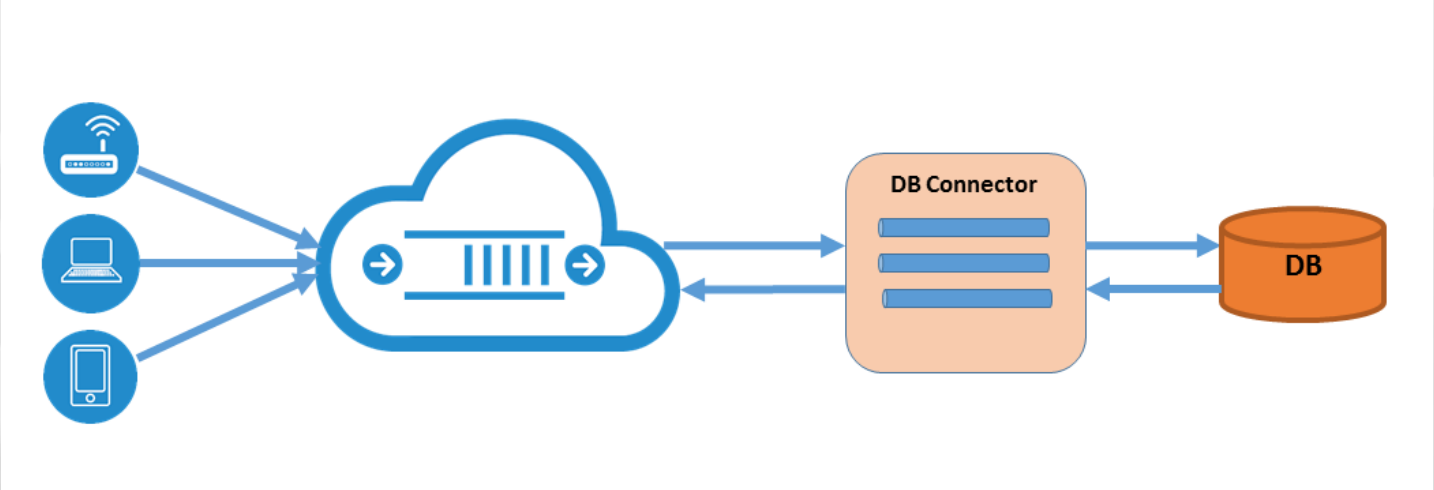
- Secure area where the file can be landed for further processing.
- Predefined basis and transfer file to landing area.
- Transformation and loading into target system.

Pull based integration system

- Integration process also has to write the extraction logic.
- Execute the query on the source system tables
- Then process the data.
- would be provided access to replica source database for enterprise data secured.

Type of Connectors for Source systems

- would involve the integration process accessing the source system tables.
- Some system or application specific connectors are needed.
- The nature of source data also determines whether a connector is needed for not.

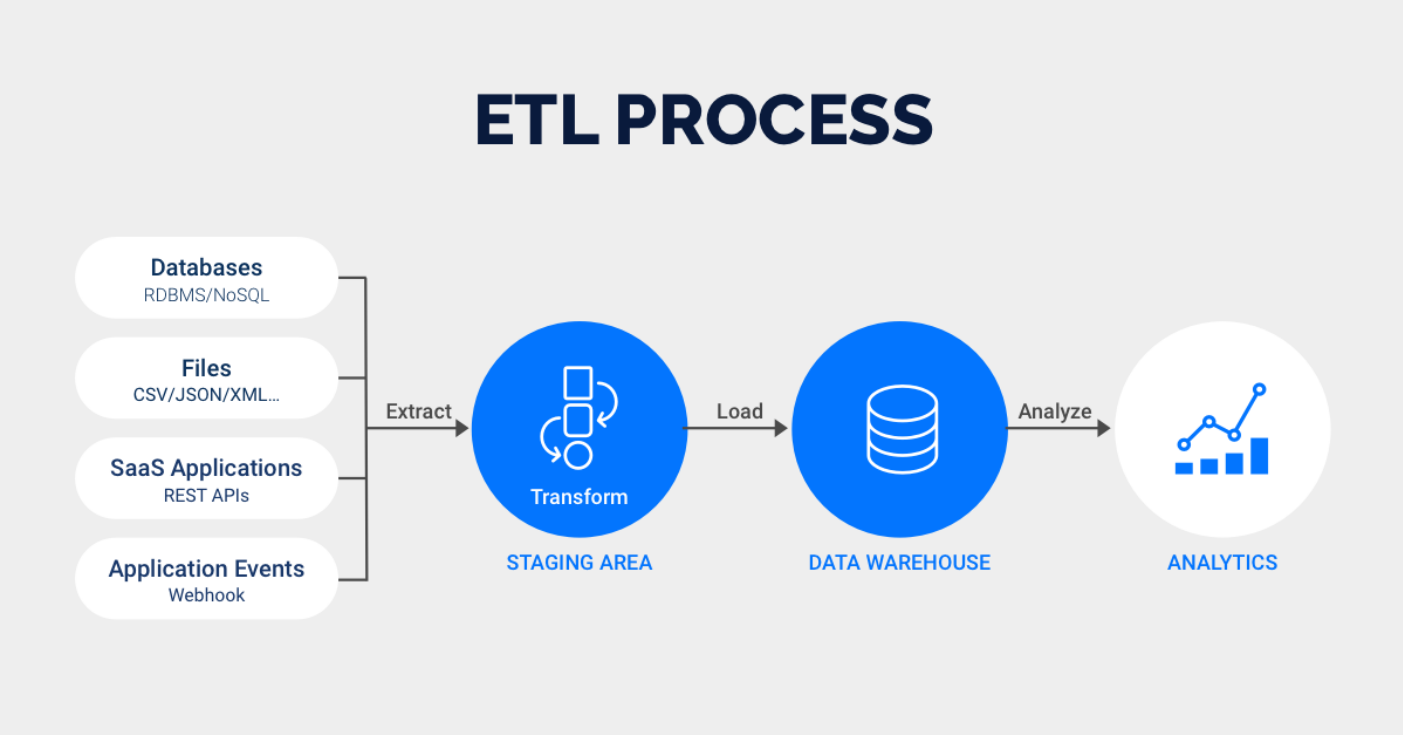


Salesforce Connector

https://robomq.readthedocs.io/en/latest/connectors/

Leverage Data Integration Engine

extract, transform, and load (ETL)

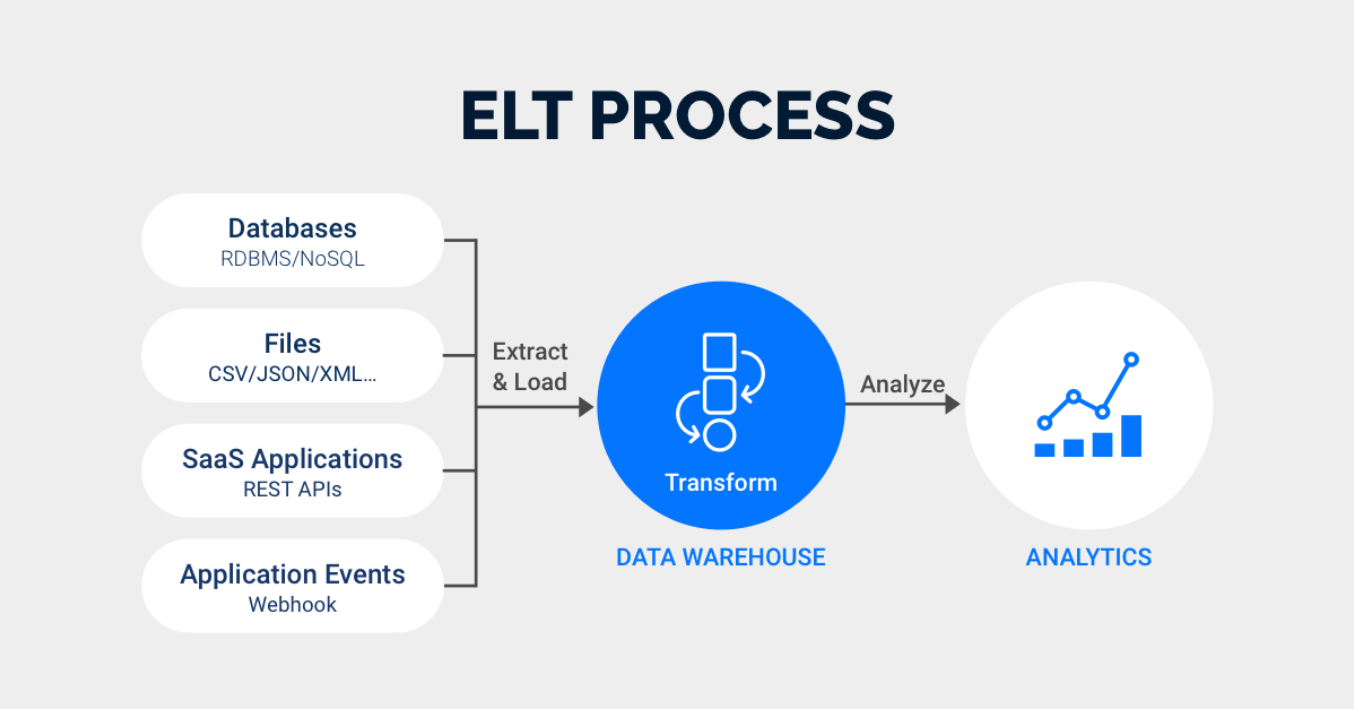


ETL vs. ELT: What's the Difference?, 2020

https://rivery.io/etl-vs-elt-whats-the-difference/

Leverage Data Integration Engine

extract, load, and transform (ELT)



ETL vs. ELT: What's the Difference?, 2020

https://rivery.io/etl-vs-elt-whats-the-difference/

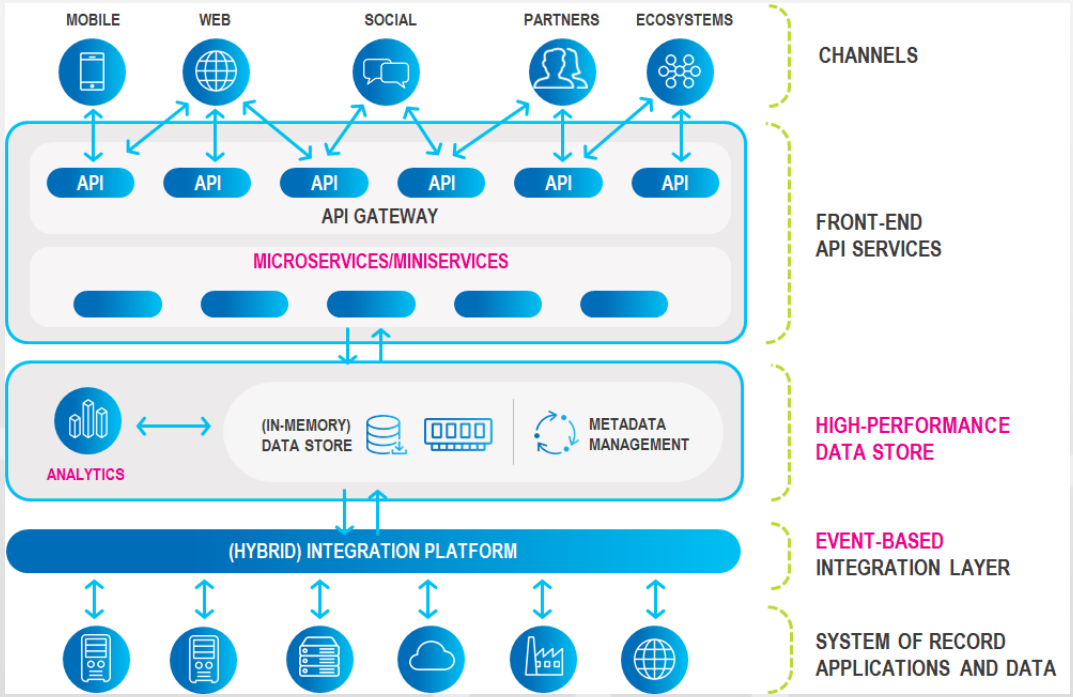
Data Integration Hubs

Key considerations for data integrations hubs are as follows:

- Persistence of data flowing through the hubs
- Canonical forms
- Data quality controls

Barriers to Adoption

- Funding issues
- More resistance to change within an organization



Turbocharge your Enterprise Application Strategies with an Intelligent Digital Integration Hub, 2019
https://www.gigaspaces.com/blog/turbocharge-your-enterprise-application-strategies-with-gigaspaces-and-informatica/

Slowly Changing Dimensions (tables)

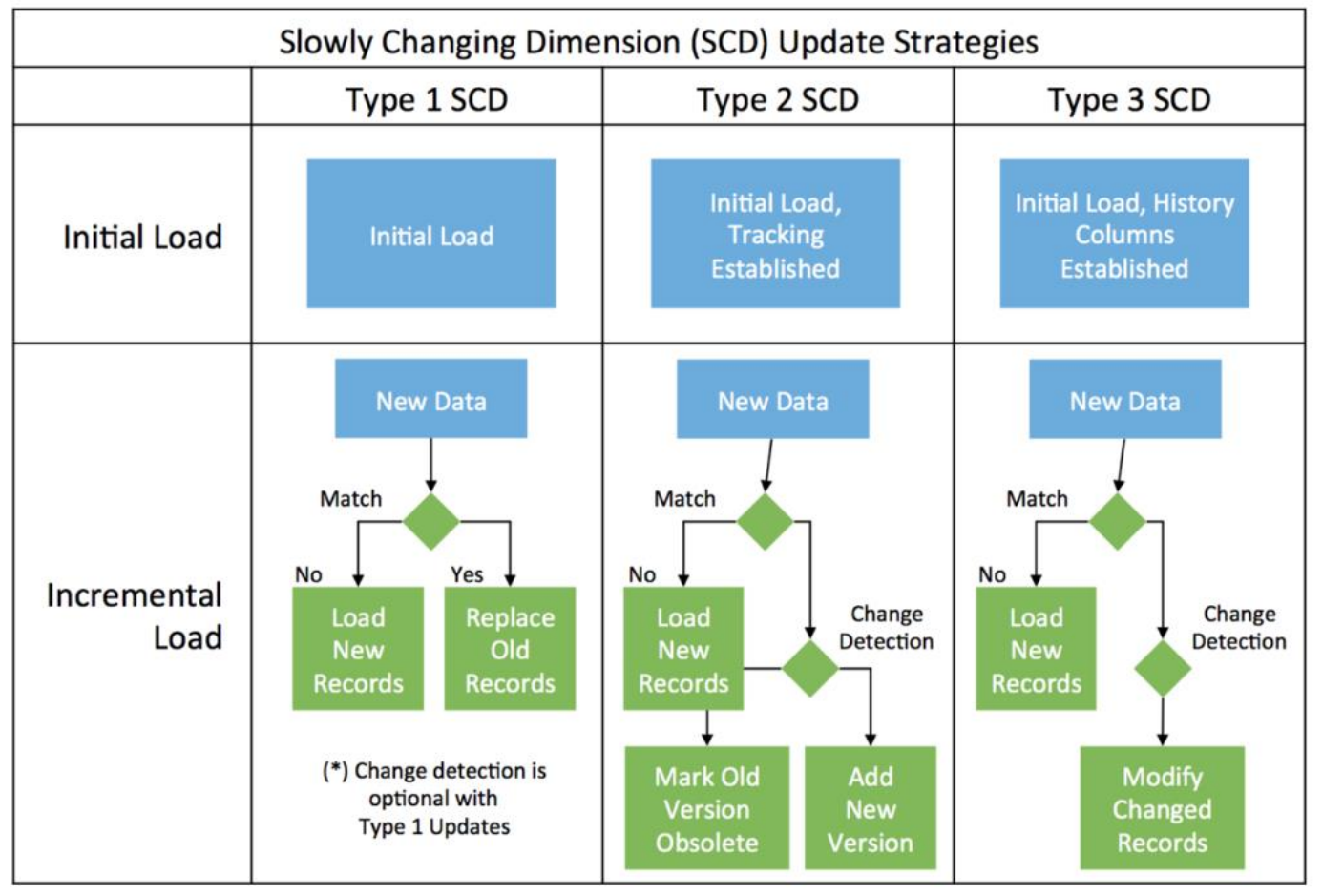
Often in decision support systems is a need to track the historical changes in dimension attributes over time. This requirement can be addressed by implementing slowly changing dimensions (SCD) in the dimension tables.

- Dimension สวนใหญ่มักจะคงทตลอดเวลา
- จะมีหลายๆ dimension ที่อาจจะไม่คงที่ แต่มีความเปลี่ยนแปลงเกิดขึ้นอย่างช้าๆ
- แอทริบิวต่างๆจะมีความเปลี่ยนแปลงเกิดขึ้นอย่างช้าๆ
- การเขียนค่าใหม่ทับค่าเดิมนั้น ไม่ใช่ทางเลือกทีเหมาะสมในการจัดทำคลังข้อมูล

Slowly Changing Dimensions (tables) (cont.)

- In general there are many ways to deal with SCD where the most used are probably the following:
- Type 0: keeping the original value
- Type 1: overwriting the old value with new value
- Type 2: adding a new record
- Type 3: adding a new column
- Type 4: adding a history table
- Type 6: combining 1, 2, and 3

Slowly Changing Dimensions (tables) (cont.)



Update Hive Tables the Easy Way Part 2

https://blog.cloudera.com/update-hive-tables-easy-way-2/

Slowly Changing Dimensions (tables) (cont.)

- SCD Type 2 example:

ID	Name	Email	State	ValidFrom	ValidTo
93	Tosha Parisian	arline72@hotmail.com	IL	2017-01-01	null

ID 93 Before Type 2 Merge

ID	Name	Email	State	ValidFrom	ValidTo
93	Tosha Parisian	arline72@hotmail.com	IL	2017-01-01	2017-07-24
93	Tosha Parisian	junie45@price.info	CA	2017-07-24	null

ID 93 After Type 2 Merge

We have simultaneously and atomically expired the first record while adding a new record with up-to-date details, allowing us to easily track full history for our dimension table.

Update Hive Tables the Easy Way Part 2

https://blog.cloudera.com/update-hive-tables-easy-way-2/

Slowly Changing Dimensions (tables) (cont.)

- SCD Type 3 example:

ID	Name	Email	LastEmail	State	LastState
1	Dr. Iza Gerhold	loragutkowski@yahoo.com	loragutkowski@yahoo.com	NJ	NJ
2	Katharyn Goyette DVM	sadiewunsch@hotmail.com	sadiewunsch@hotmail.com	VI	VI
3	Nikolas Tromp	danniekemmer@yahoo.com	danniekemmer@yahoo.com	VI	VI
4	Ms. Shawnna Gerlach DVM	reichelson@hotmail.com	reichelson@hotmail.com	MP	MP

Type 3 Initial Managed Table

Update Hive Tables the Easy Way Part 2

https://blog.cloudera.com/update-hive-tables-easy-way-2/

Slowly Changing Dimensions (tables) (cont.)

- SCD Type 4 example:

Product table

Product Key	Product Code	Product Name	Product Category
123	CLA12	Cola	Beverages
124	BK11	Baking Soda	Baking Products

Product History

Product Key	Product Code	Product Name	Product Category	Create Date
123	CLA12	Cola	Aerated Drinks	26-Jun-2011
123	CLA2	Cola	Carbonated Drinks	14-Apr-2013
124	BK11	Baking Soda	Bakery & Food	14-Aug-2013

Figure 5-3. *Type 4 slowly changing dimensions (SCD) illustration*

Update Hive Tables the Easy Way Part 2

https://blog.cloudera.com/update-hive-tables-easy-way-2/

Slowly Changing Dimensions (tables) (cont.)

- SCD Type 6 example:

Customer table

Customer Key	Customer Code	Customer Name	Current City	Historical City	Start Date	End Date	Current City Flag
100	G123	Ravi Gupta	Pune	Pune	02-Feb-2009	31-Dec-9999	Y

Figure 5-4. *Customer table*

Customer table with city changes

Customer Key	Customer Code	Customer Name	Current City	Historical City	Start Date	End Date	Current City Flag
100	G123	Ravi Gupta	Bangalore	Pune	02-Feb-2009	09-Aug-2012	N
101	G123	Ravi Gupta	Bangalore	Bangalore	10-Aug-2012	31-Dec-9999	Y

New record inserted for Type 2

Type 1 – Current city updated in previous record as well

Extra column to maintain previous value of city (Type 3)

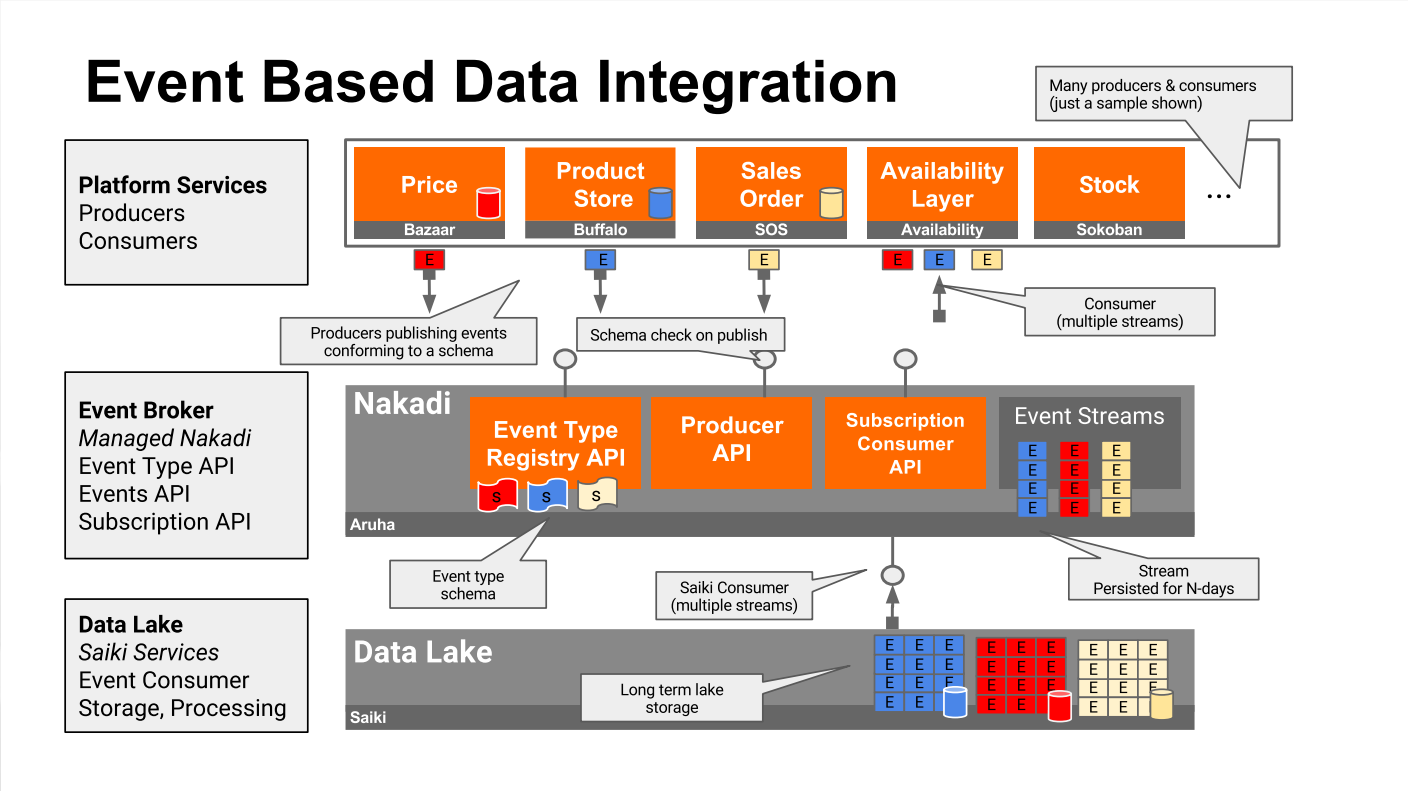
Figure 5-5. *Type 6 slowly changing dimensions (SCD) illustration with different surrogate key*

Update Hive Tables the Easy Way Part 2

https://blog.cloudera.com/update-hive-tables-easy-way-2/

Real-time data integration

- The principal approaches to real-time data integration include the following:
- Change data capture
- Events or streams-based

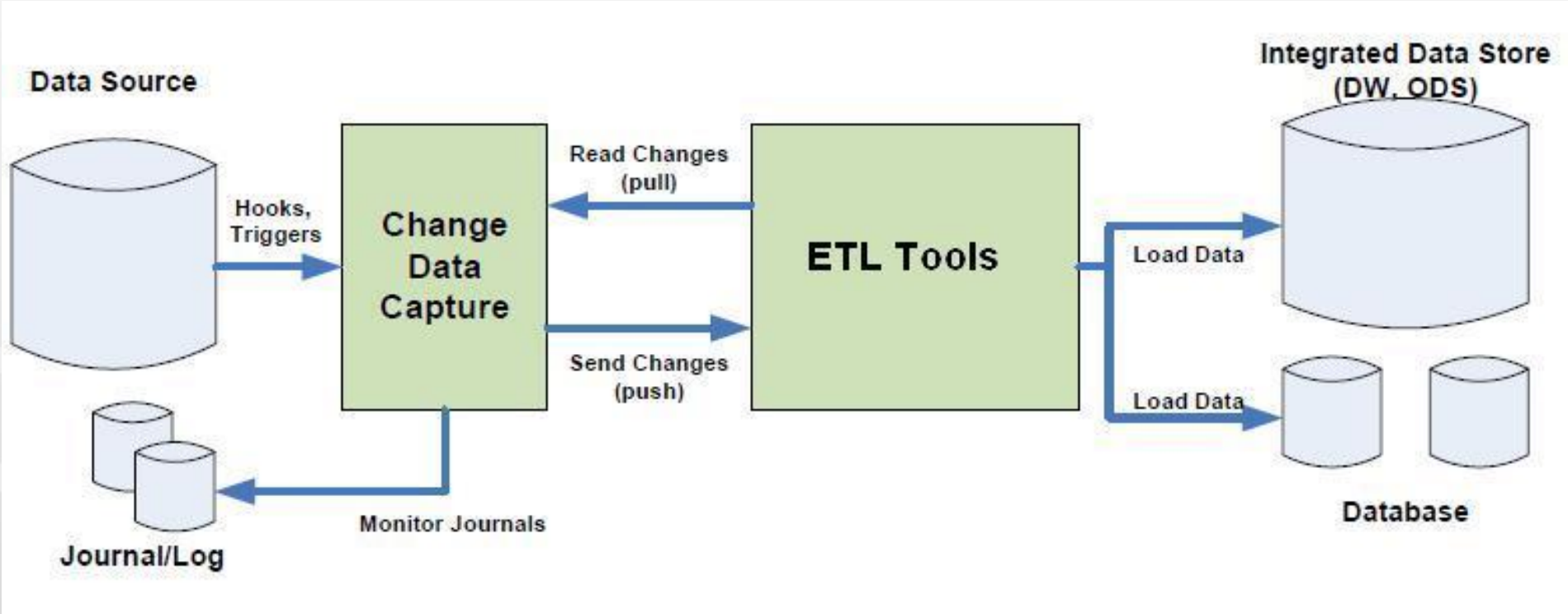


Nakadi Event Broker

https://nakadi.io/manual.html

Real-time data integration (cont.)

- Change data capture: implement with database, middleware, file or direct to target.



Speeding ETL Processing in Data Warehouses Using High-Performance Joins for Changed Data Capture (CDC)

https://www.researchgate.net/figure/Working-of-CDC-in-conjunction-with-ETL-tools_fig4_224202553

Real-time data integration (cont.)

Events or streams-based data integration:

- The objective is to detect events such as opportunities and threats and respond to them as quickly as possible.
- There are different types of events that can be detected where threshold values of a parameter is exceeded.

Use cases:

- Health care
- Transportation
- Telecommunications

Thank you

Week 3 : Data Management Component

Data sourcing

Data integration and exchange

Digital Arts & Science
DIGITECH