Data Warehouse Design and Implementation – Case Study

Data Warehouse Workshop

Helana Jaraiseh
Computing Science, University of the Fraser Valley
Supervisors: Dr. Youry Khmelevsky and Dr. Albert Wong
September 24, 2025

IEEE Okanagan College Student Branch
Okanagan College Computer Science Department
Langara. The College of Higher Learning.
UFV
IEEE Thompson - Okanagan Section

Land Acknowledgement

Okanagan College respectfully acknowledges that the Penticton, Kelowna and Vernon campuses are located on the lands of the Syilx Okanagan People and the Salmon Arm campus is located on the lands of the Secwépemc People. We honour our relations and hold up their knowledge, welcoming all to our house of learning.

Introduction to Data Warehouse Design

Introduction

- The Explosive Growth of Data!
- Terabytes to petabytes available and collected *[1]*
- Major sources of abundant data *[1]*
  - Business: Web, e-commerce, transactions, stocks, …
  - Science: Remote sensing, bioinformatics, systems, …
  - Society: News, searches, social media, …
- *"Data rich but information poor" [1]*
- Abundance of data, need powerful data analysis tools

[1] pp. 1 – 2, 5, Ch1: "Introduction" in Data Mining Concepts and Techniques

Introduction

Business Intelligence (BI) Applications provide:

- Solid platform of consolidated, historical data to support information processing for analysis.
- Decision support database separate from organization's operational databases.

What is a Data Warehouse?

"A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process." [1]

- Organized around major subjects
- Focused on modeling and data analysis for decision makers
- Provide simple and concise view around issues
- Exclude irrelevant data for decision-making

[1] pp. 126, Ch4: "Data Warehousing and Online Analytical Processing" in Data Mining Concepts and Techniques

What is a Data Warehouse?

- Dimension Tables
- Fact Tables

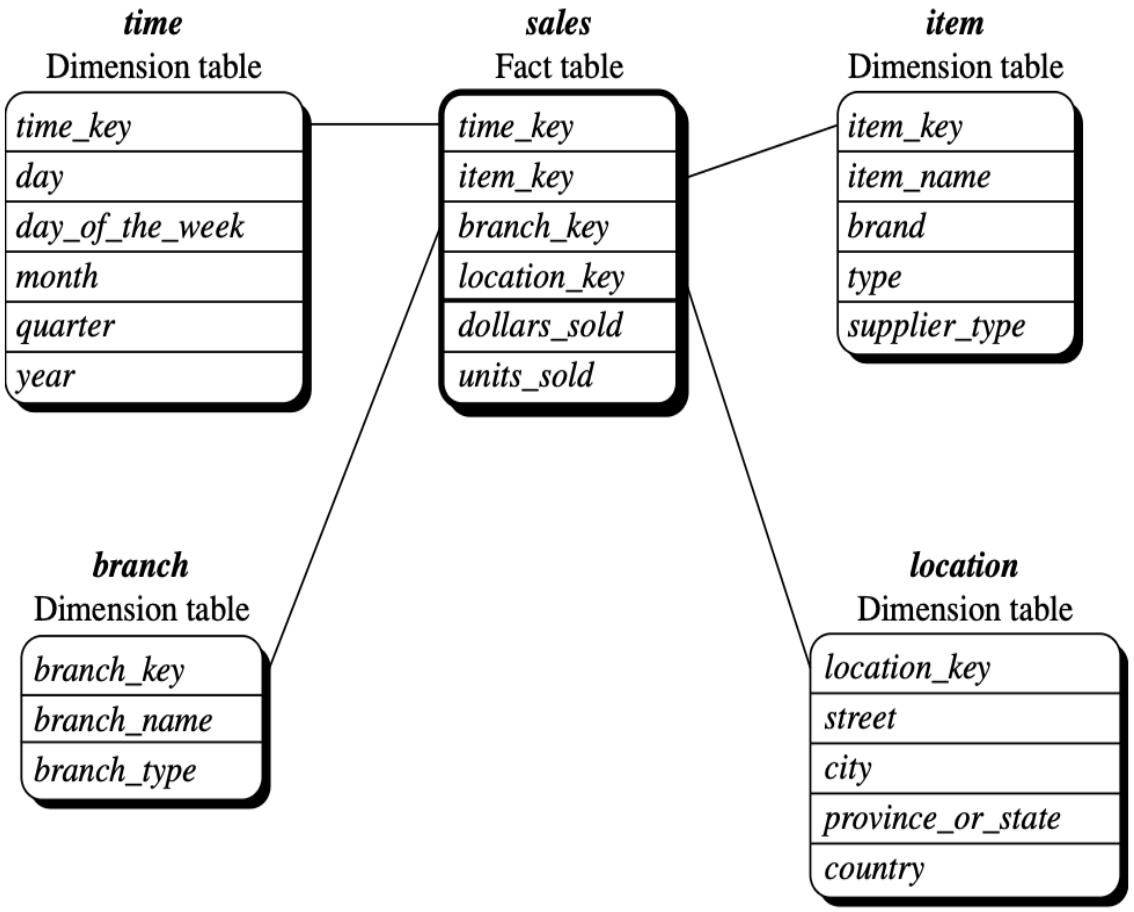


Figure 1 – Star Schema of Sales Data Warehouse (See Figure 4.6 from [1])

[1] pp. 140, Ch4: "Data Warehousing and Online Analytical Processing" in Data Mining Concepts and Techniques

Data Warehousing Architecture

- Data Sources
- Extract, Load, Transform (ELT)
- Extract, Transform, Load (ETL)
- Data Warehouse Server
- OLAP Server
- Front-end Tools

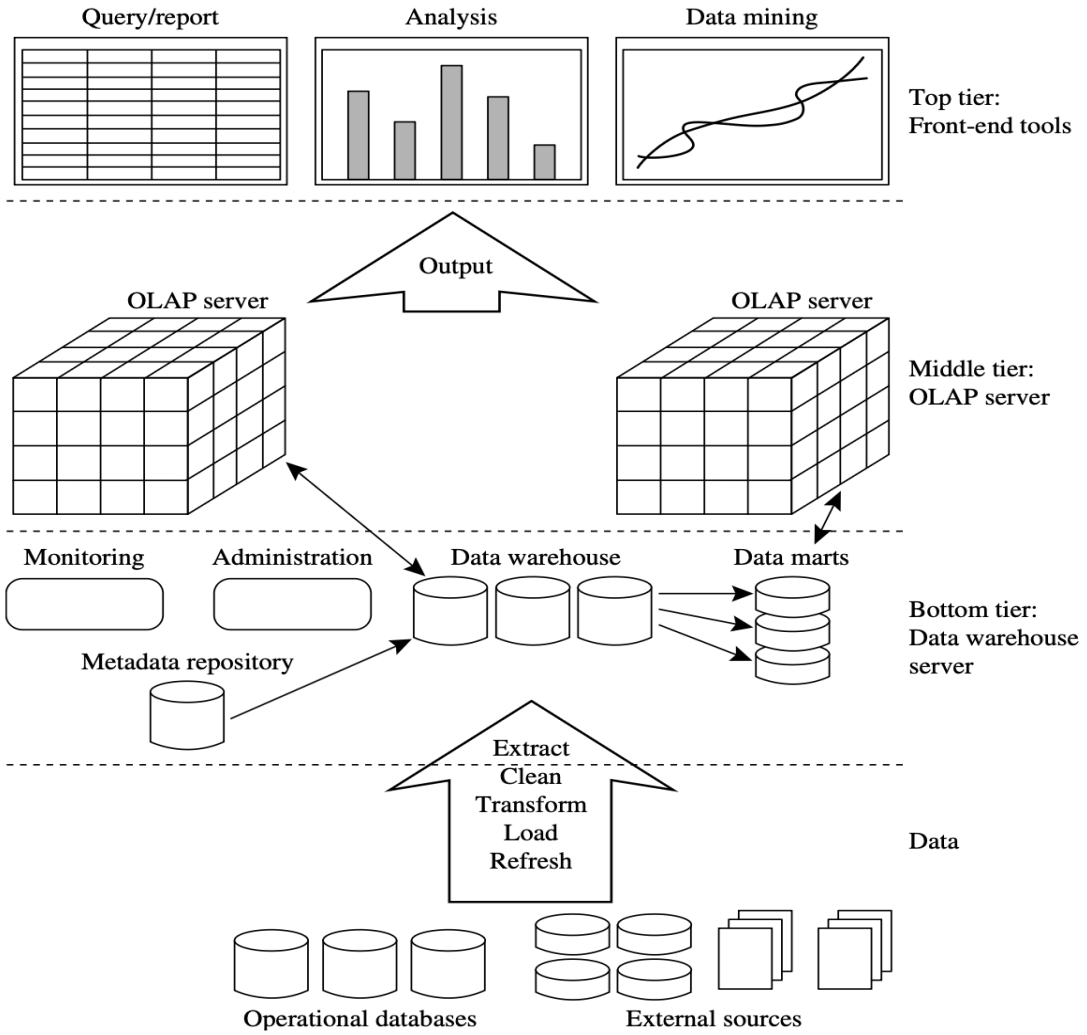


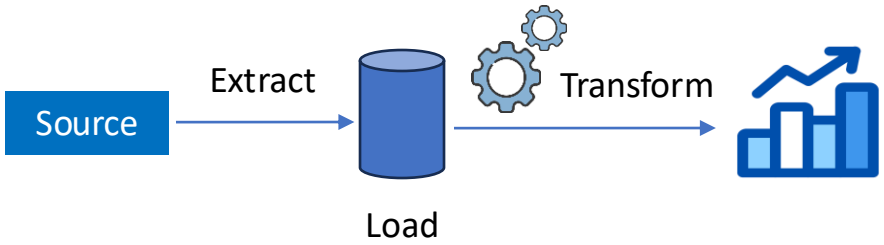
Figure 2 – Three-tier Data Warehousing Architecture (See Figure 4.1 from [1])

[1] pp. 131, Ch4: "Data Warehousing and Online Analytical Processing" in Data Mining Concepts and Techniques

ELT and ETL

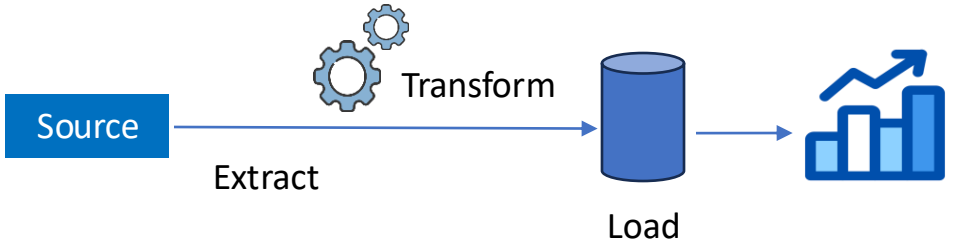
Two data-processing approaches that filter, sort, and clean large volumes of data.

Extract, Load, and Transform (ELT)



- Extracts and Loads data to storage as it is
- Transforms data after storing raw data
- Recommended for students and researchers
- Flexible, can handle raw data
- Raw data stored for recovery for years

Extract, Transform, and Load (ETL)



- Data is governed with a set of business rules
- Transforms data before loading to data warehouse
- Not recommended for students and researchers
- Must define architecture and data rules before ETL
- Hard to find raw data online years later

Algorithmic Trading, A Case Study

Data Warehouse Design

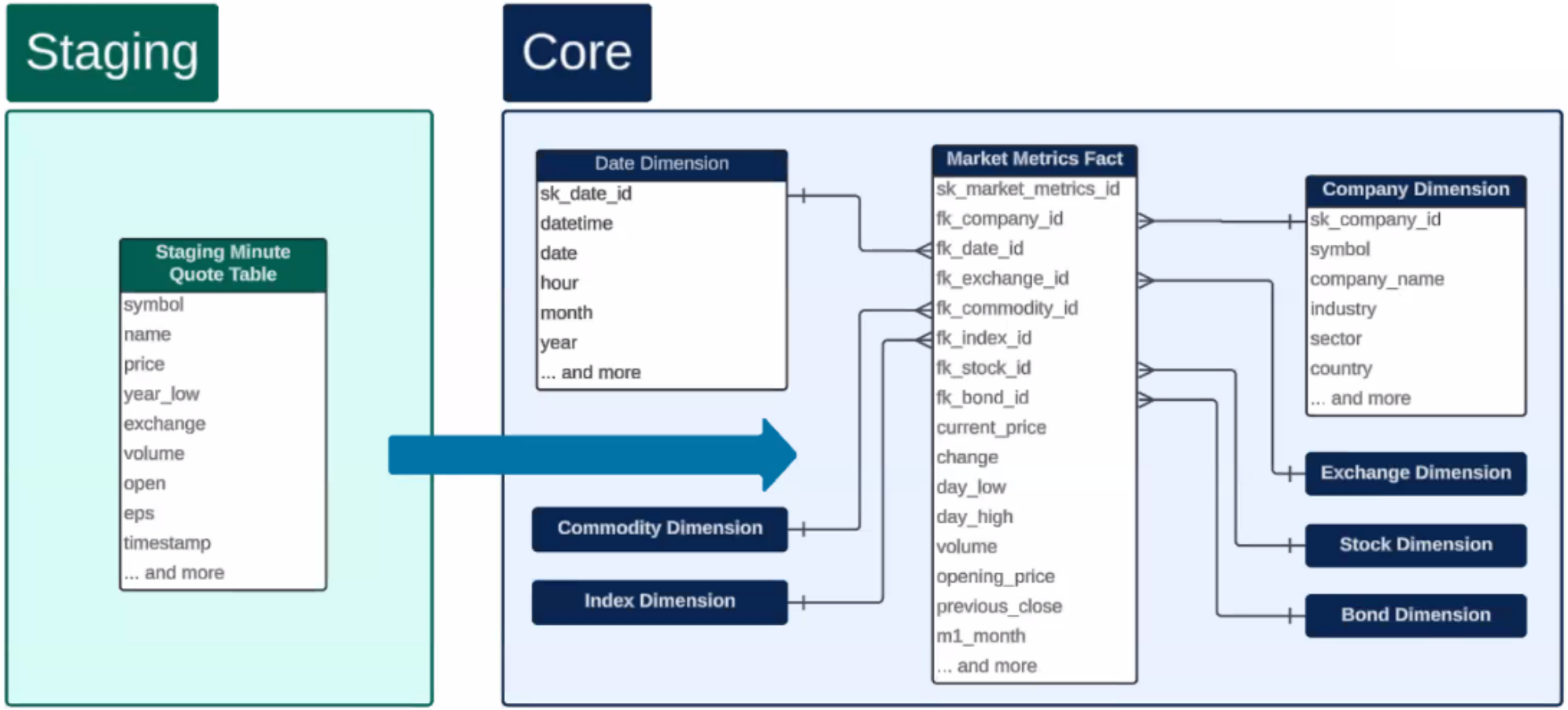


Figure 3 – Algorithmic Trading Data Warehouse Design (See Slide #10 in [2])

[2] Data Extraction, Transformation, and Loading (ETL) Process Automation and Data Warehouse Implementation for Algorithmic Trading Machine Learning Modelling, 2025 IEEE SysCon

Data Warehousing Architecture

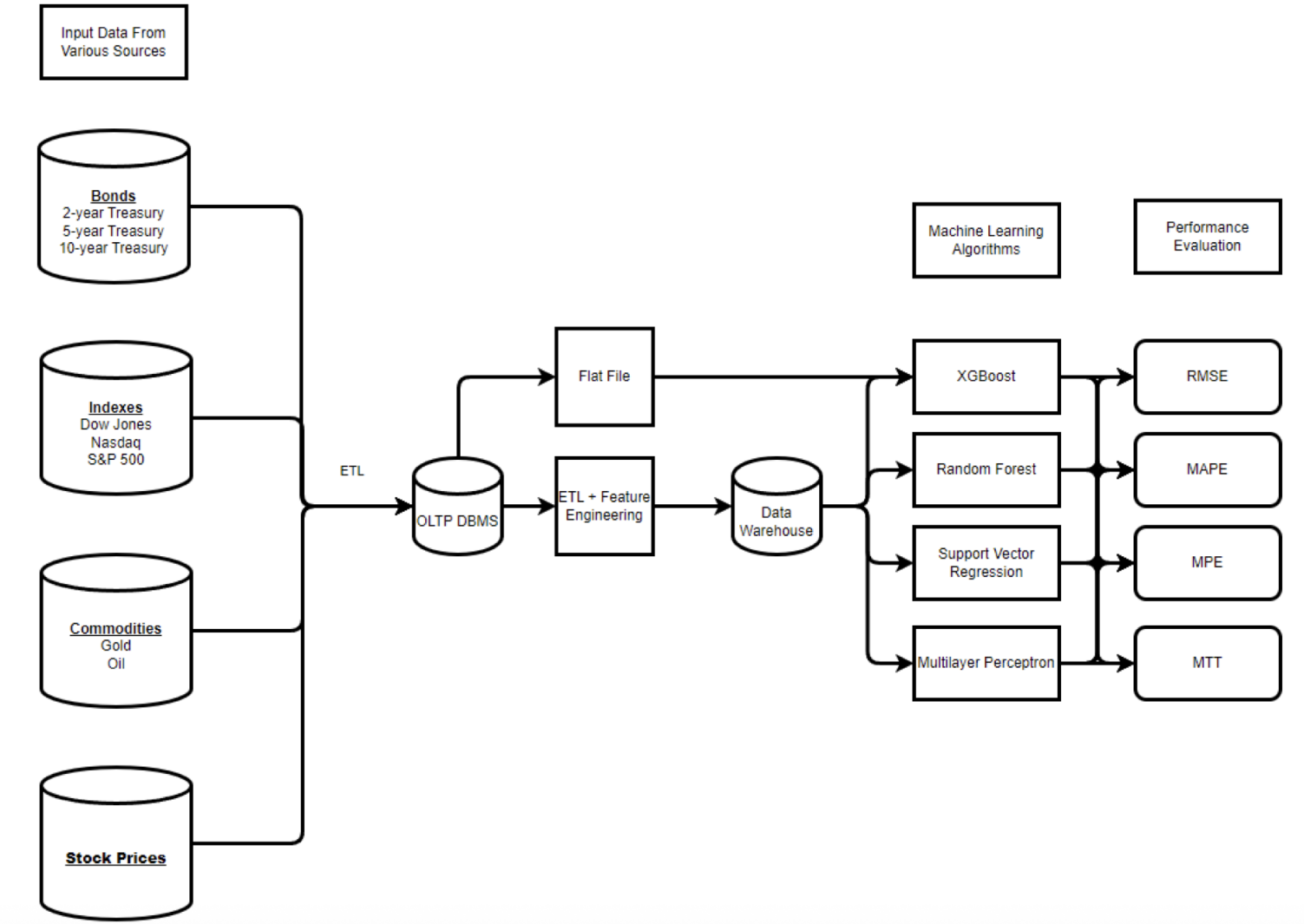


Figure 4 – Algorithmic Trading ETL-Automation-Architecture (See Fig. 4 from [3])

[3] Data Extraction, Transformation, and Loading Process Automation for Algorithmic Trading Machine Learning Modelling and Performance Optimization

Challenges and Lessons Learned

- BI Applications need well-built architecture
- BI Applications also need complete, reliable, and real-time data
- Algorithmic Trading Problem - *"Garbage in, Garbage out"*
  - *incomplete data may have produced price overestimation spikes*
- Data Collection was biggest challenge for this project
  - API datasets contain missing entries
  - Hard to find APIs with high granularity
  - Manually "patching" missing values
  - Historical data and real-time data can be expensive

Future Work

Data Warehousing Architecture

- ELT Scripts
- Staging Databases for Raw Data
- Data Warehouse for Model Features and Values

Data Collection and Modelling

- Working with CSVs in parallel
- Data Analysis and Engineering on Datasets
- Feature Engineering for Models

Questions?

References

[1] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, 3rd ed., Waltham, MA, USA: Elsevier Inc., 2012, total number of pages: 6.

[2] K. Cormier, K. Gagnier, J. Padron-Uy, D. Sareen, A. Parihar, and Y. Khmelevsky. (Apr. 2025). Data Extraction, Transformation, and Loading (ETL) Process Automation and Data Warehouse Implementation for Algorithmic Trading Machine Learning Modelling. Published in: 2025 IEEE International systems Conference (SysCon). [Google Slides]. Available: https://docs.google.com/presentation/d/1zuUUPy9MNaIdEwdmUDbYjhzezK1jjcOmeS4KkISQ1rU/edit.

[3] N. Ebadifard, A. Parihar, Y. Khmelevsky, G. Hains, A. Wong, and F. Zhang. (Dec. 2023). Data Extraction, Transformation, and Loading Process Automation for Algorithmic Trading Machine Learning Modelling and Performance Optimization. arXiv preprint arXiv:2312.12774. [Online]. Available: https://arxiv.org/pdf/2312.12774.