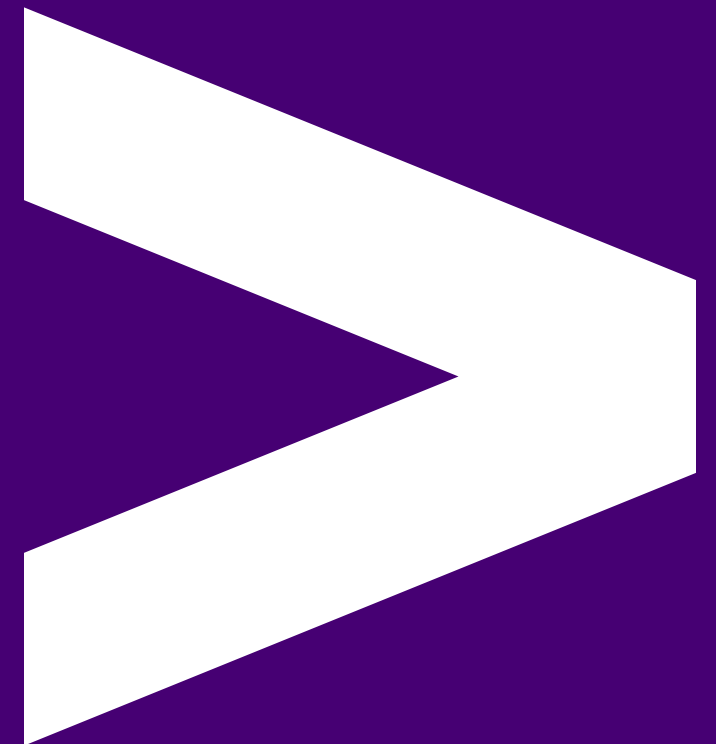


Data Warehousing

So like a database but bigger?



Overview

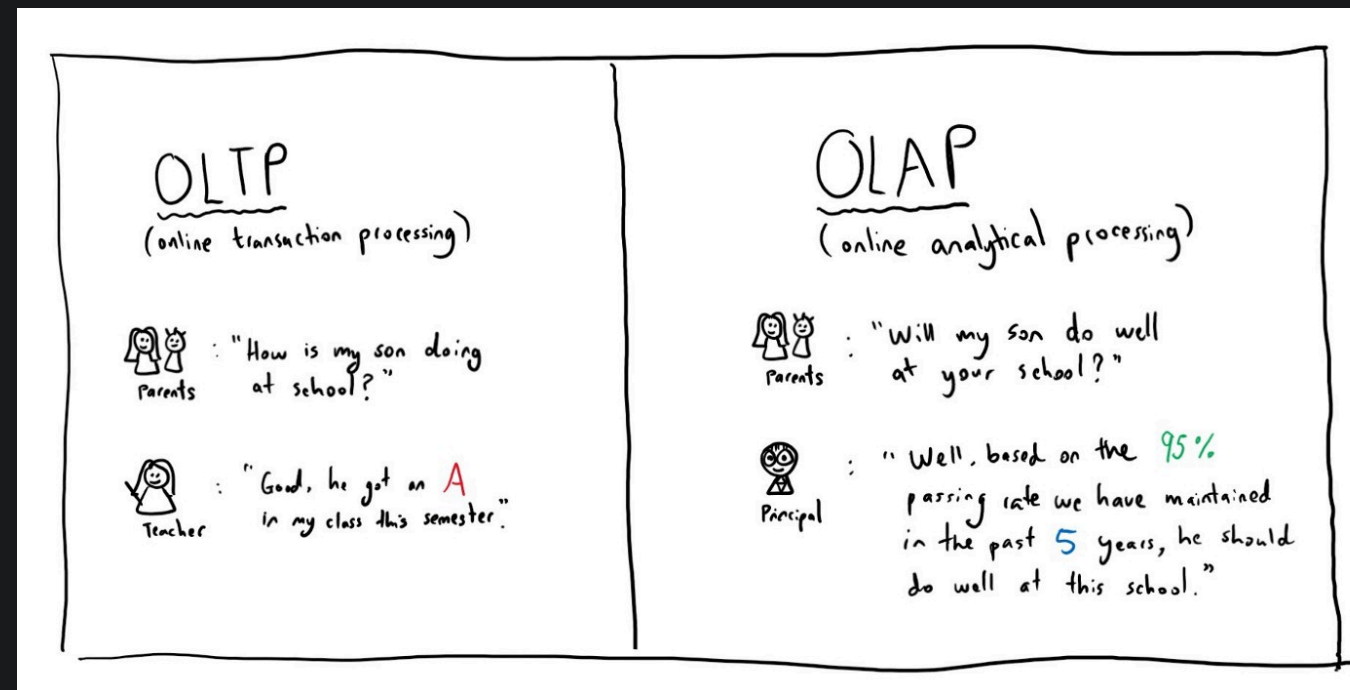
- History of data warehousing
- Data warehousing techniques

Learning Objectives

- Clarify the difference between Databases and Data Warehouses
- Identify the different Data Warehouse schema types

OLTP vs. OLAP

- **OLTP (Database):** Online Transaction Processing Information systems facilitates and manages transaction-oriented applications
- **OLAP (Data Warehouse):** Online Analytical Processing is an approach to answer multi-dimensional analytical queries swiftly



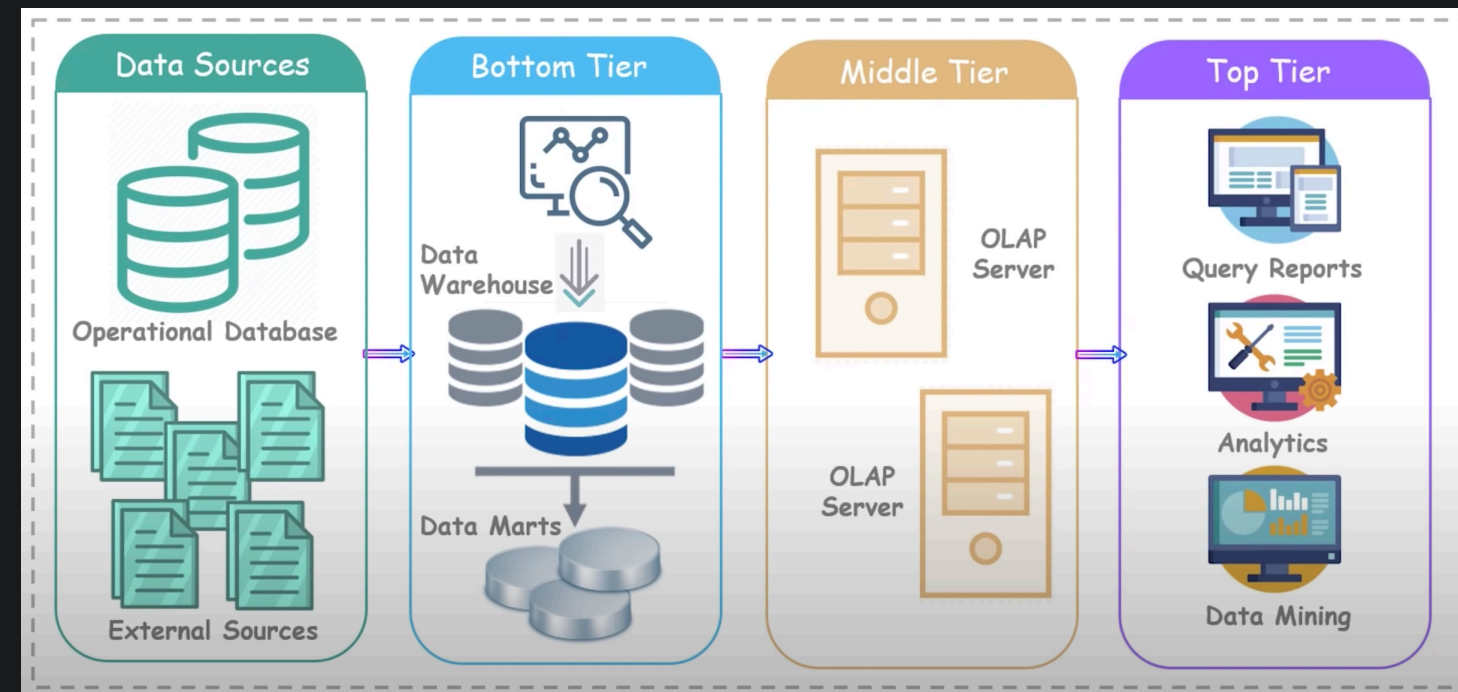
Traditional (Relational) Database

- Stores data in tables
- Uses Online Transactional Processing (OLTP)
- Row based data with high integrity
- Helps perform the fundamental operations of a business
- Generally normalised - with complex schema / table joins

Data Warehouse

- A system that aggregates data from multiple sources into a single, central and consistent data store
- Helps prepare data for data analytics, business intelligence (BI), data mining, visualisation tools, and other forms of advanced analytics

Data Warehouse



Data Warehouse Architecture

Data Sources:

- Internal sources such as wages, personnel, or maintenance databases
- External sources are not being generated from within the organisation like markets, competitors, or demographics

Bottom Tier:

- Warehouse Database Server
- Typically, *Column*-based storage
- Uses various backing tools to extract data from different sources
- Cleanses data and transforms it before loading into a Data Warehouse

Data Warehouse Architecture cont.

Middle Tier:

- OLAP Server (**O**nline **A**nalytical **P**rocessing)
- Performs multi-dimensional analysis of business data
- Transforms the data into a format that we can perform complex calculations and data modelling on

Top Tier:


- Like a front-end client layer
- Holds different types of querying and reporting tools for which client applications can perform data analysis

Emoji Check:

Do you feel you understand the difference between OLTP and OLAP? Say so if not!

1. 🥲 Haven't a clue, please help!
2. 😞 I'm starting to get it but need to go over some of it please
3. 😐 Ok. With a bit of help and practice, yes
4. 😊 Yes, with team collaboration could try it
5. 😄 Yes, enough to start working on it collaboratively

Business Intelligence (BI)



What do you think it means?

Business Intelligence

Business intelligence (BI) is software that ingests business data and presents it in user-friendly views such as reports, dashboards, charts and graphs. Analysing this data helps businesses gain actionable insights and inform decision-making.

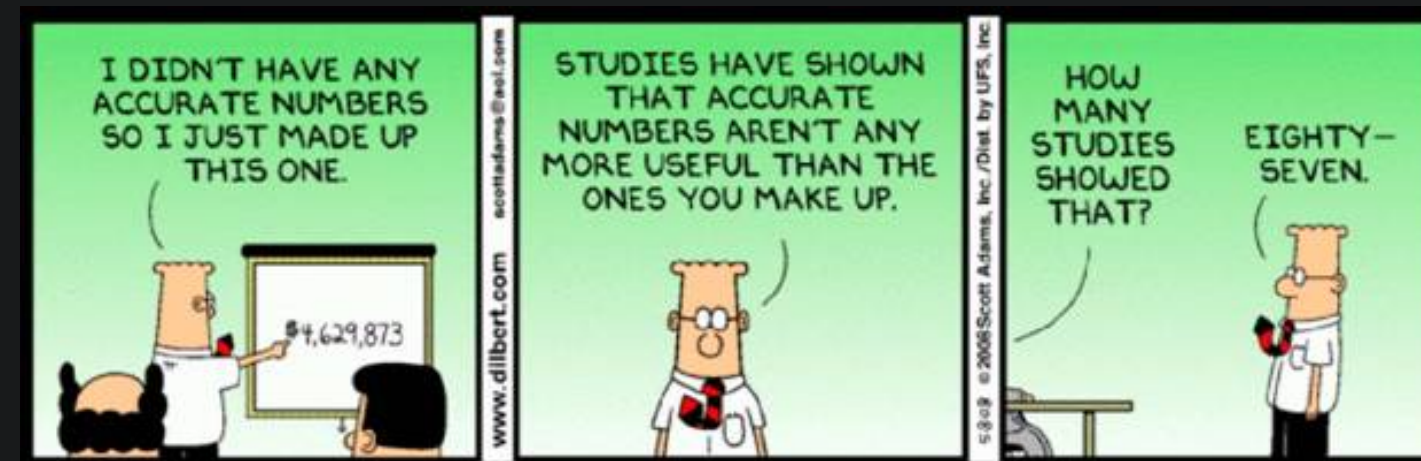
BI tools enable business users to access different types of data — historical and current, third-party and in-house, as well as semi-structured data and unstructured data like social media. Users can analyze this information to gain insights into how the business is performing.

Source: [ibm.com/topics/business-intelligence](https://www.ibm.com/topics/business-intelligence)

Why Business Intelligence?

- Marketing
- Commercial Strategy
- Development Metrics... i.e. A/B Testing

Can you think of any others and examples?



Observations and Trends

- Data sources can be pretty varied
- Data tends to be imported into staging tables as soon as possible for processing
 - **Staging tables:** Temporary tables containing data before it has been processed
- Often long chains of events that rely on previous stages completing exist
- Can you think of any potential issues occurring?

Emoji Check:

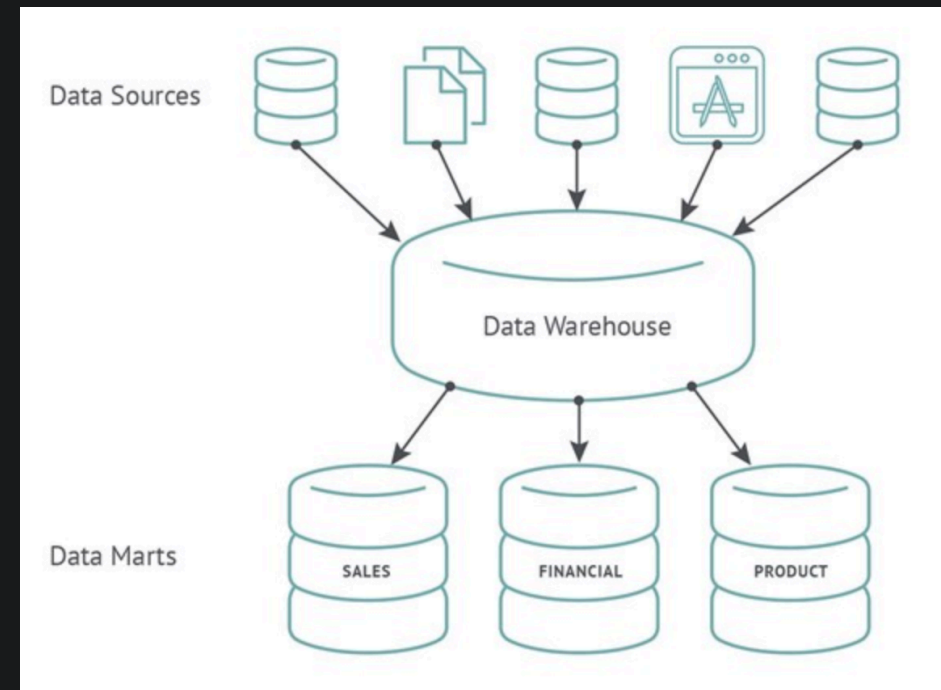
Do you feel you understand the basics of Business Intelligence? Say so if not!

1. 😓 Haven't a clue, please help!
2. 😞 I'm starting to get it but need to go over some of it please
3. 😐 Ok. With a bit of help and practice, yes
4. 😊 Yes, with team collaboration could try it
5. 😄 Yes, enough to start working on it collaboratively

Data Marts

- A condensed and more focused version of a data warehouse
- Each "Mart" contains a subset of the data warehouse, specifically oriented to a business sector or team (e.g. only Sales, or only Stock levels by week)
- They protect the data warehouse by decreasing the number of users directly accessing the main data
- Data Marts are intended to be **Read Only**

Data Marts



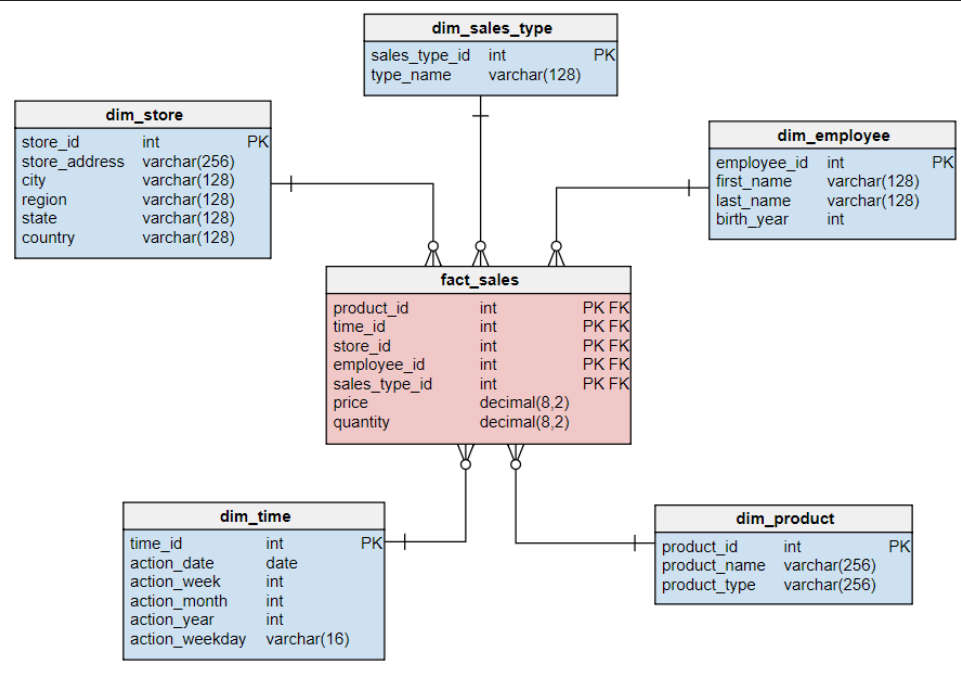
Organising Our Data

- There are two common schemas for storing data within Data Warehouses/Marts:
 - Star Schema
 - Snowflake Schema

Star Schema

- Introduced by Ralph Kimball in 1996
- Fact tables can refer to any number of Dimension Tables
- Tables are usually denormalised, allowing for writing simpler queries, involving less joins
- Because of this denormalisation, data integrity is relaxed, which may allow for data anomalies

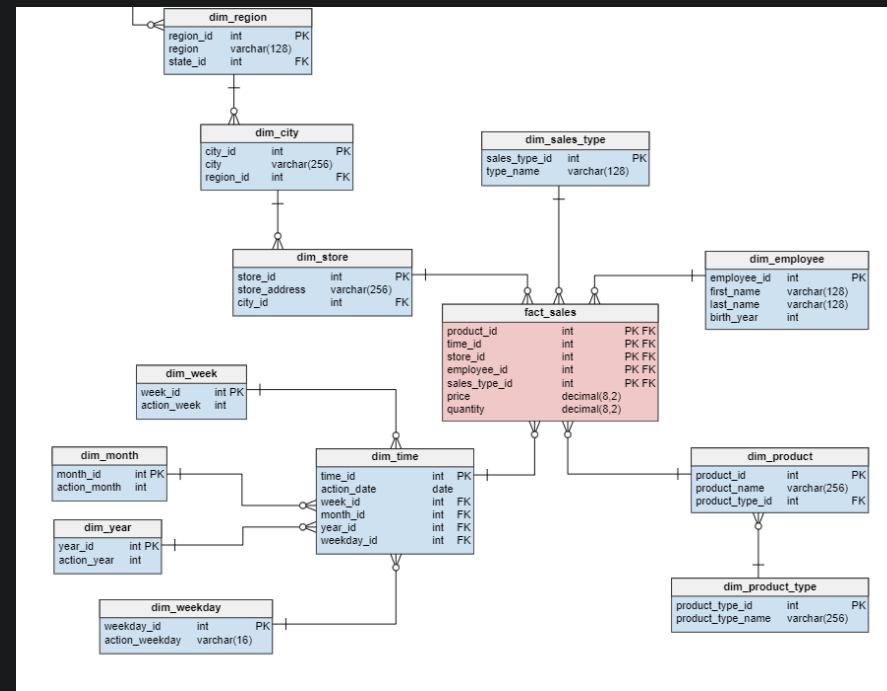
Star Schema



Snowflake Schema

- The fact tables are connected to multiple dimensions
- More complex approach based on Star Schema
- It strips out low cardinality attributes (unique values) and forms separate tables
- Dimensions are normalized into multiple related tables
- Queries can become complex with a number of joins needed to retrieve all data
- Stricter data integrity leads to less anomalies like duplication, or missing relation data

Snowflake Schema



Emoji Check:

Do you feel you understand the basics of Star and Snowflake schemas?
Say so if not!

1. 🥲 Haven't a clue, please help!
2. 😞 I'm starting to get it but need to go over some of it please
3. 😐 Ok. With a bit of help and practice, yes
4. 😊 Yes, with team collaboration could try it
5. 😄 Yes, enough to start working on it collaboratively

Quiz Time! 🧐

What data processing system does a traditional database use?

1. OLAP
2. OLTA
3. OLTP
4. OLAT

Answer: 3

Bonus point if you can remember what it stands for!

What data processing system does a data warehouse use?

1. OLAP
2. OLTA
3. OLTP
4. OLAT

Answer: **1**

Bonus point if you can remember what it stands for!

Which tier in a traditional data warehouse architecture would this be in?

Cleanse data and transform it before loading into the data warehouse.

1. Data Sources
2. Bottom Tier
3. Middle Tier
4. Top Tier

Answer: 2

Tips for your team projects (offline)

(This is repeated from the final AWS session, [../aws-08-cfn-ec2.](#))

For your project time, there is a file of a few pointers and gotchas to consider.

- See [../aws-08-cfn-ec2/handouts/README-team-project-considerations.md](#)

Terms and Definitions - recap

- **OLAP:** Answers multi-dimensional analytical queries swiftly
- **Business Intelligence:** Applying data analytics to business practice
- **Data Marts:** Condensed, more focused version of a data warehouse
- **Star Schema:** One or more 'fact tables', referencing any number of 'dimension tables'
- **Snowflake Schema:** Normalised data in multiple related tables, whereas the star schema's dimensions are denormalised

Overview - recap

- History of data warehousing
- Data warehousing techniques

Learning Objectives - recap

- Clarify the difference between Databases and Data Warehouses
- Identify the different Data Warehouse schema types

References and Further Reading

- [Explain By Example: OLTP vs. OLAP](#)
- [Dimensional Modelling by Kimball](#)
- [Introduction to Data Vault Modelling](#)

Emoji Check:

On a high level, do you think you understand the main concepts of this session? Say so if not!

1. 🥲 Haven't a clue, please help!
2. 😞 I'm starting to get it but need to go over some of it please
3. 😐 Ok. With a bit of help and practice, yes
4. 😊 Yes, with team collaboration could try it
5. 😄 Yes, enough to start working on it collaboratively