Data Warehousing

So like a database but bigger?

Overview

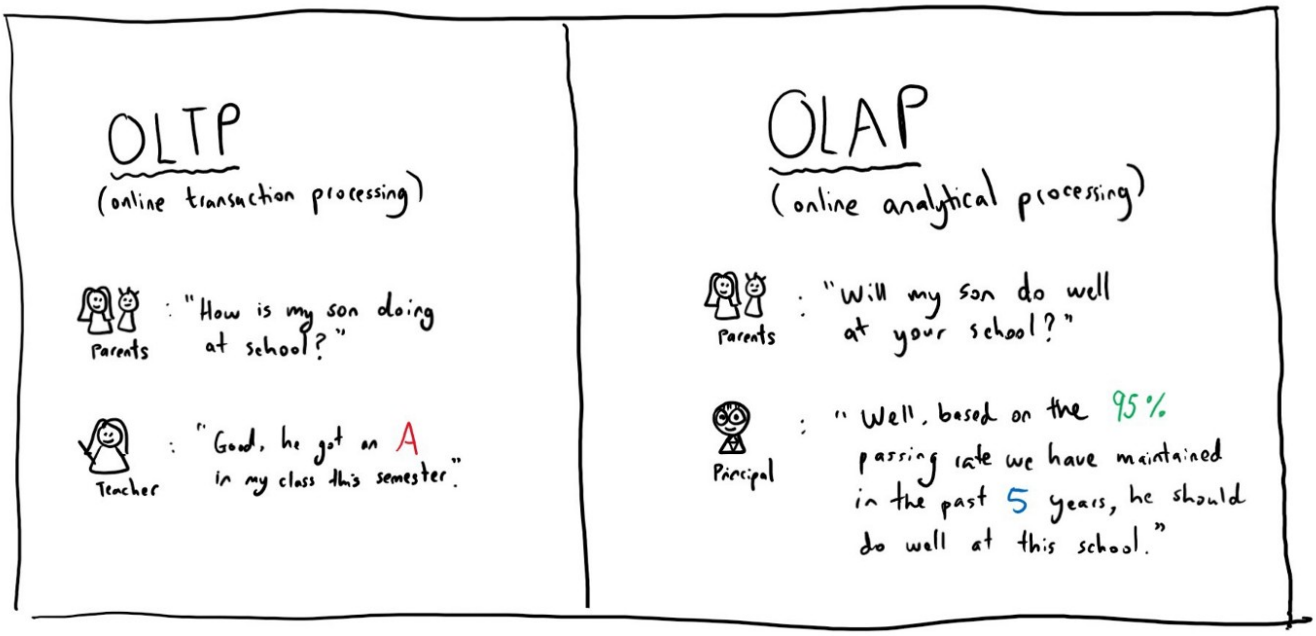
- History of data warehousing
- Data warehousing techniques
- AWS Redshift

Learning Objectives

- Clarify the difference between Databases and Data Warehouses
- Identify the different Data Warehouse schema types
- Explain how AWS Redshift works

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


OLTP
(online transaction processing)
Parents : "How is my son doing at school?"
Teacher : "Good, he got an A in my class this semester".
OLAP
(online analytical processing)
Parents : "Will my son do well at your school?"
Principal : "Well, based on the 95% passing rate we have maintained in the past 5 years, he should do well at this school."

Traditional Database

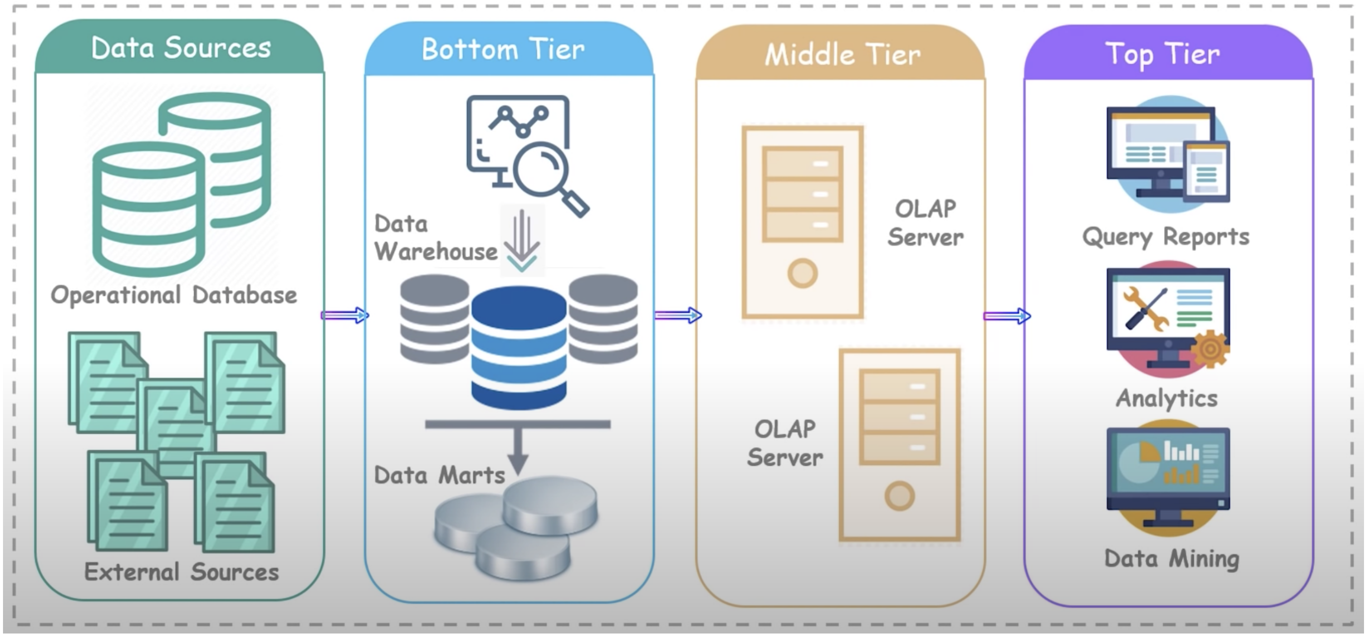
- Stores data in tables
- Uses Online Transactional Processing (OLTP)
- Helps perform the fundamental operations of a business

- Generally normalised - complex table joins

In OLTP, systems typically facilitate & manage (database) transaction-oriented applications. High throughput and are insert- or update-intensive.

```
Business operations such as payments, orders, customer data etc.
```

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
- 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**n**L**ine **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

OLAP is more about complex queries, smaller volumes, business intelligence or reporting. Optimized for read only queries.

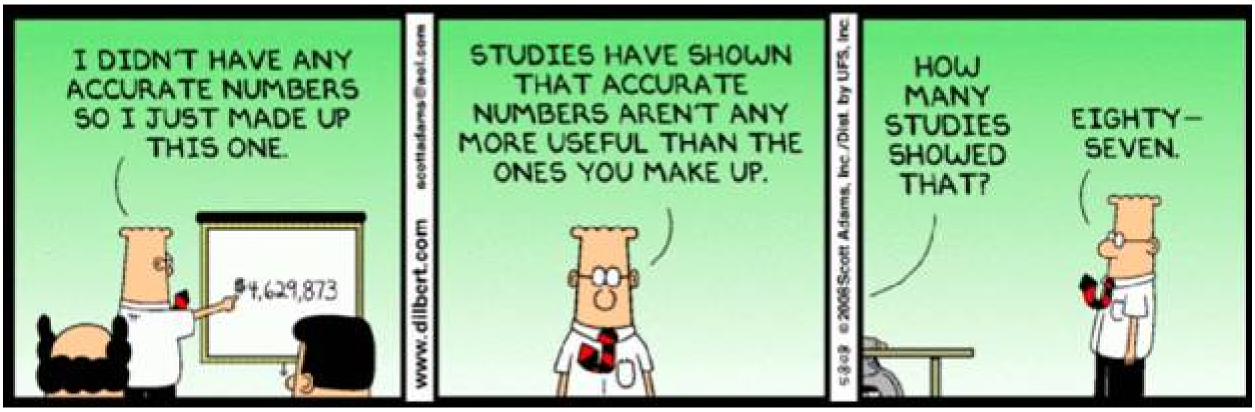
```
Denormalisation (leaving redundancy in) will serve to improve read
performance (at expense of write performance).

With data warehousing, data is united from many tables into one.
```

Business Intelligence

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

Can you think of any others and examples?


I DIDN'T HAVE ANY ACCURATE NUMBERS SO I JUST MADE UP THIS ONE.
$4,629,873
www.dilbert.com scottadams@aol.com
STUDIES HAVE SHOWN THAT ACCURATE NUMBERS AREN'T ANY MORE USEFUL THAN THE ONES YOU MAKE UP.
5-8-08 © 2008 Scott Adams, Inc./Dist. by UFS, Inc.
HOW MANY STUDIES SHOWED THAT?
EIGHTY-SEVEN.

e.g. Market Conditions, setting business objectives, identifying opportunities etc

---

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?

Chain of events: Ruth subscribes to a popular food delivery service.

```
Externally the service relies a food supply chain, the system needs to
forecast what will be available for her to choose from next week. Relying
on data from third parties - the food suppliers.

Internally, every week when her order is dispatched the system relies on
her address and payment details being available before the order is
dispatched.
```

Data Marts

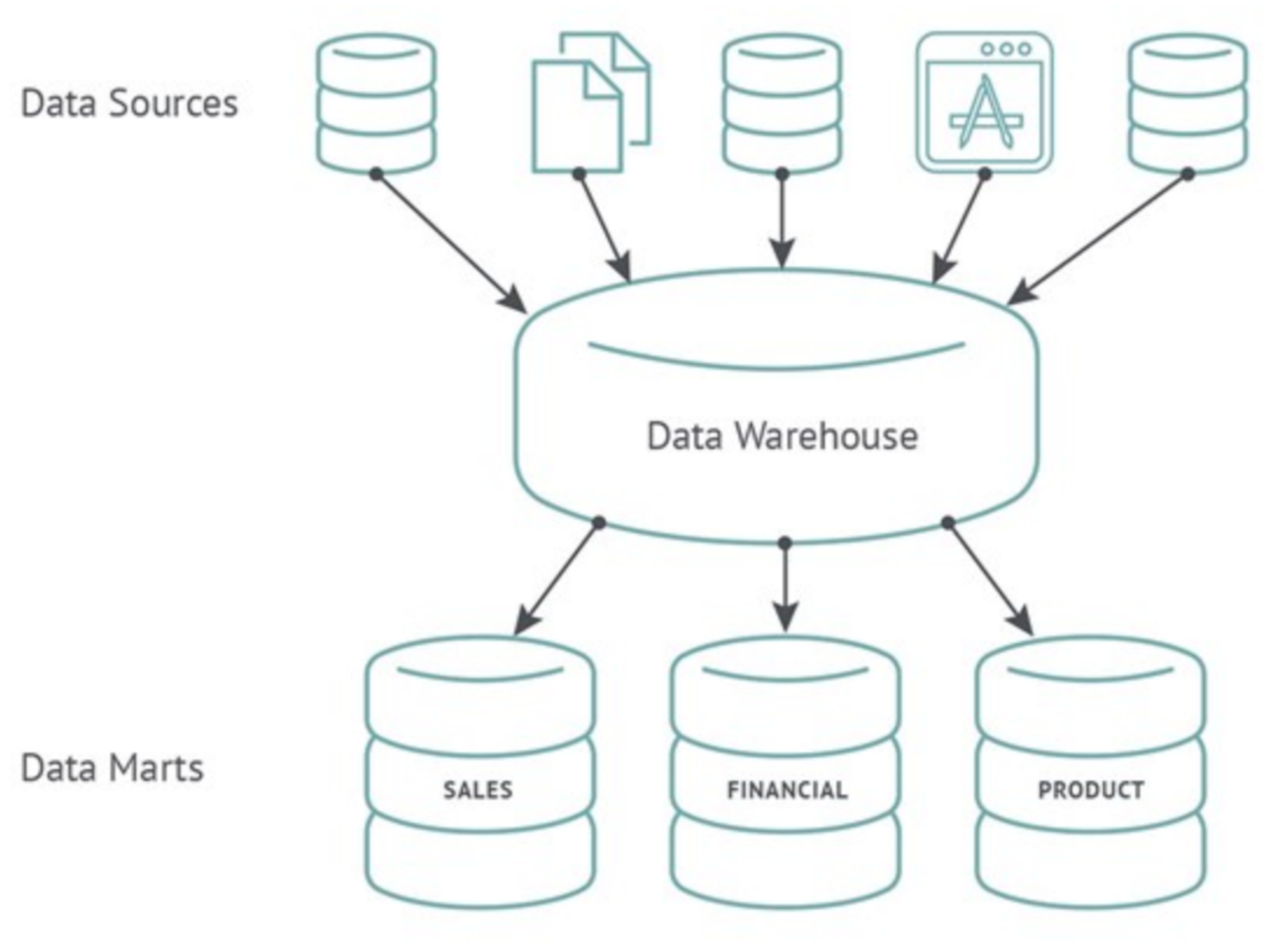
- Basically a condensed and more focused version of a data warehouse
- Each "Mart" contains a subset of the data warehouse, specifically oriented to a business sector or team
- They protect the data warehouse by decreasing the number of users
- Data Marts are intended to be **Read Only**

Data marts are basically more condensed and more versions of data warehouses that reflect the process specifications of each business unit (accounting, marketing, sales etc) within an organisation.

```
Data marts can also be dedicated to specific regions.

The subset of data may still span across or all of an enterprise's areas.
```

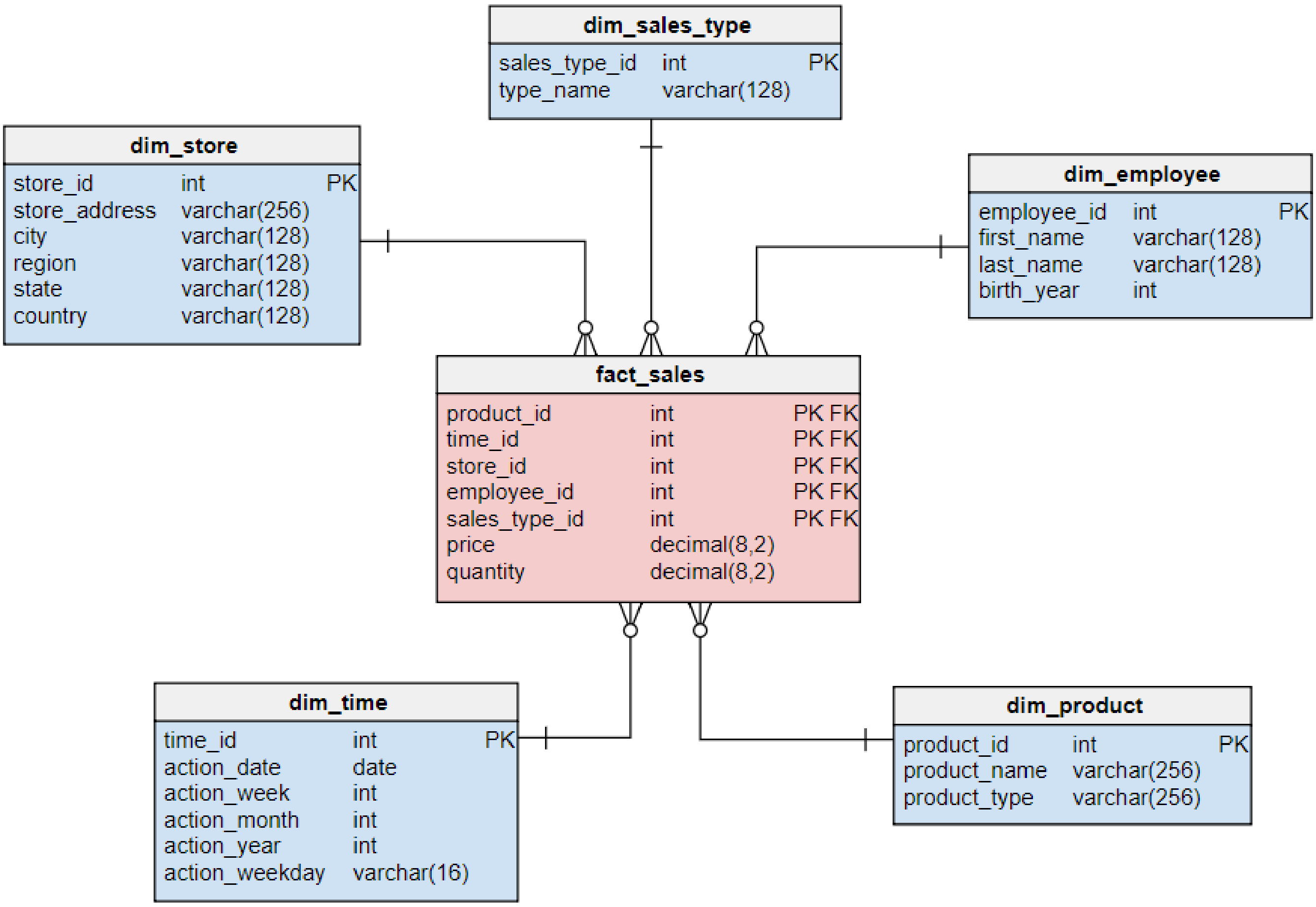
Data Marts



Organising Our Data

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

Star Schema



Read through: One or more 'fact tables' (measurements, metrics or facts of a business process) referencing any number of 'dimension tables' (categorize facts and measures in order to enable users to answer business questions - provide filtering, grouping & labelling) - Fact tables reference 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

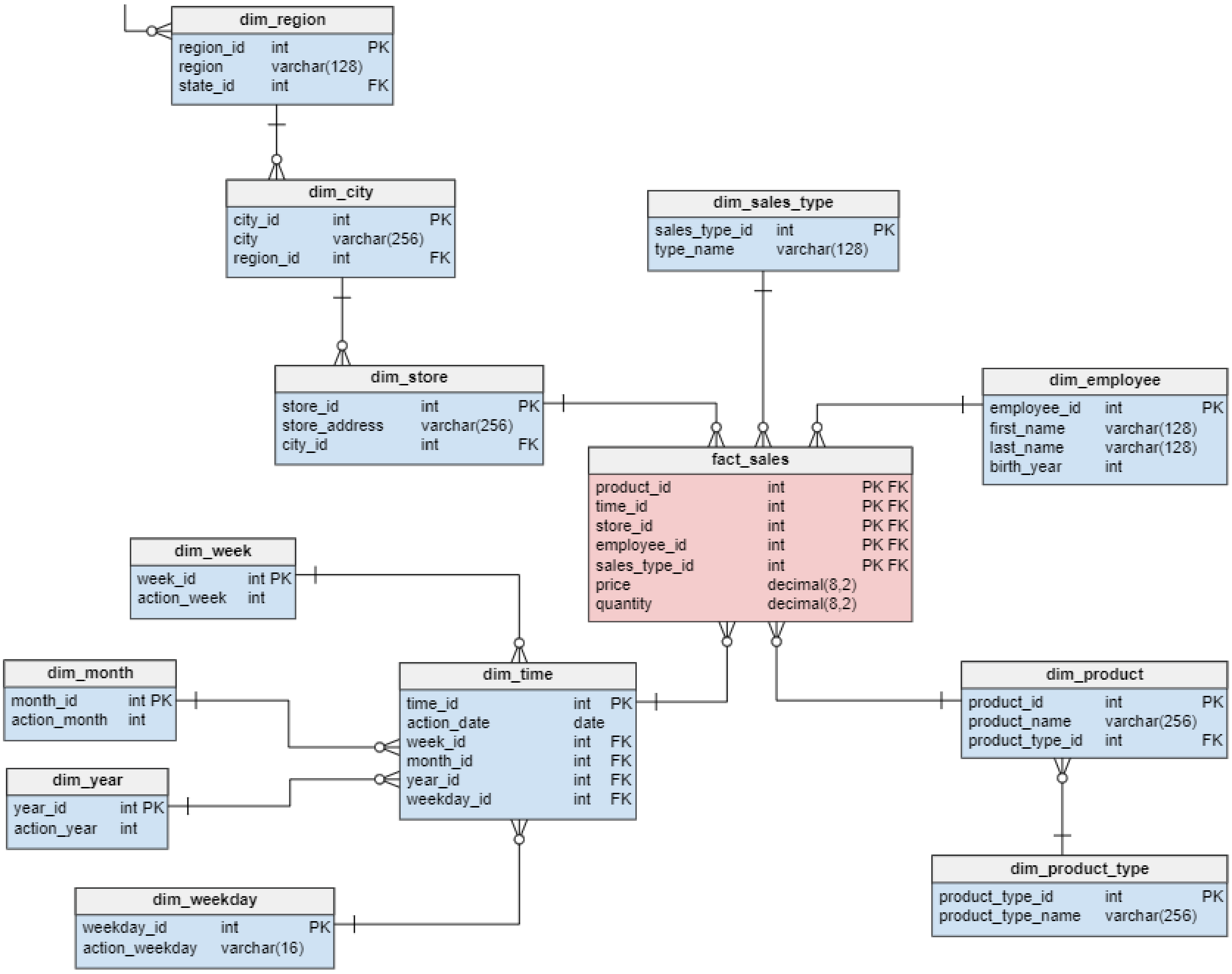
---

Star Schema

- Fact tables reference 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

---

Snowflake Schema



The principle behind snowflaking is normalization of the dimension tables by removing low cardinality attributes and forming separate tables.

```
- More complex approach based on Star Schema
- The fact tables are connected to multiple dimensions
- 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

- 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

---

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

The extraction/cleansing/transformation and loading of the data happens in the bottom tier

---

AWS Redshift

Before Redshift - Traditional Data Warehouses

- Time consuming to pull data from the large warehouses using traditional architecture
- Costly - hardware, setup, electricity, security, estate

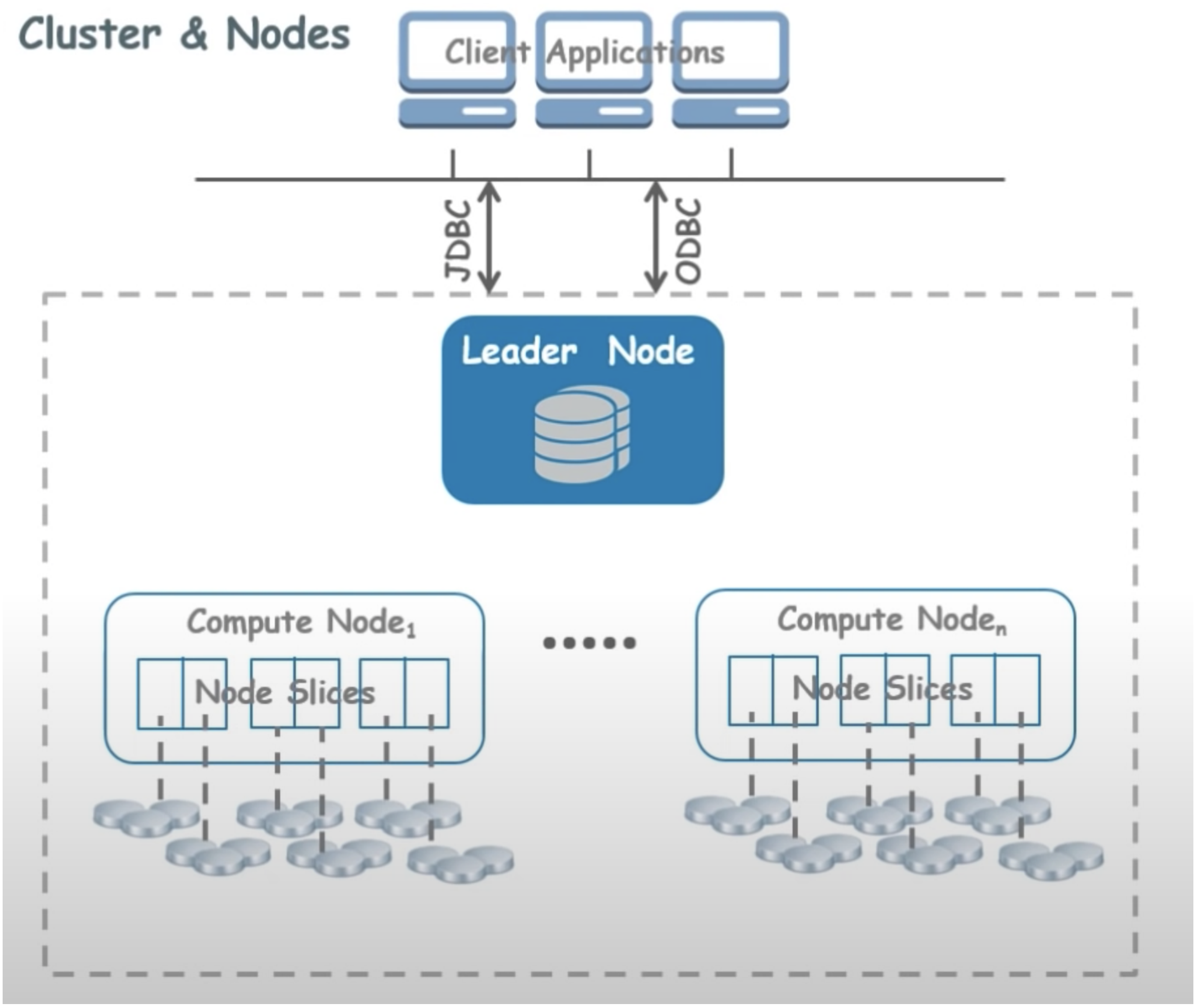
- Maintenance costs often outweighed the benefits (upgrading systems due to more data being added)
- Performance issues
- Auto-scaling is not an easy concept

Redshift

- Massively **parallel**, **column-oriented** database
- Simple and cost-effective to analyse your data
- Manages, monitors and scales your system
- Up to 10x better performance than traditional
- Collection of compute resources which are called nodes
- These nodes, when organised into groups, become **clusters**
- Each cluster runs a Redshift engine which contains one or more DBs

---

Architecture



Clusters

- A cluster has a leader node with one or more compute notes
- The leader node receives queries from client applications (BI, analytical software etc.)

It develops a suitable query execution plan and coordinates parallel executions of these plans with one or more compute notes

```
Once the compute nodes finish, the leader aggregates the results from the
nodes and sends back a response to the client application
```

Compute Nodes

- Compute resources which execute a query plan
- Transmits data among themselves to solve queries
- Nodes are further divided into (node) slices
- Each node slice receives an allocation of memory and performs operations in parallel

---

Node types

When you launch a (non-free tier) cluster, you need to specify the node types

Dense Storage Node: Storage optimized and used to handle huge data workloads (uses HDD).

Dense Compute Node: Compute optimized and used to handle high performance/intensive workloads (uses SSD).

Choosing the right node type:

Data Quantity: Be aware of the amount of data you want to import into your redshift DB

Complexity of queries: Different nodes support queries with differing levels of complexity

Downstream systems: That depends on the results of these queries

---

Columnar Data Storage

- The data is still represented with rows and columns as normal
- However, the data is physically stored by column, instead of rows
- Because the data stored is the same type, you can achieve better data compression
- Number of I/O operations decreases
- Also means you can query/perform data analysis on similar types of data far quicker than row storage

---

Quiz Time! 🤓

Which of the following best describes a Dense Storage Node?

1. Storage optimised and used to handle huge data workloads (uses HDD).
2. Storage optimised, used to handle high performance/intensive workloads (uses SSD).
3. Compute optimised, used to handle huge data workloads (uses HDD).
4. Compute optimised and used to handle high performance/intensive workloads (uses SSD).

Answer: 1

Which of the following best describes a Dense Compute Node?

1. Storage optimised and used to handle huge data workloads (uses HDD).
2. Storage optimised, used to handle high performance/intensive workloads (uses SSD).
3. Compute optimised, used to handle huge data workloads (uses HDD).
4. Compute optimised and used to handle high performance/intensive workloads (uses SSD).

Answer: 4

Learning Objectives Revisited

- Clarify the difference between Databases and Data Warehouses
- Identify the different Data Warehouse schema types
- Explain how AWS Redshift works

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:** Normalized data in multiple related tables, whereas the star schema's dimensions are denormalized
- **Redshift:** Store and analyse large quantities of data

References and Further Reading

- When Data Warehouse builds go wrong
- AWS Data Warehouse modernization
- Introduction to Data Vault Modelling
- Explain By Example: OLTP vs. OLAP