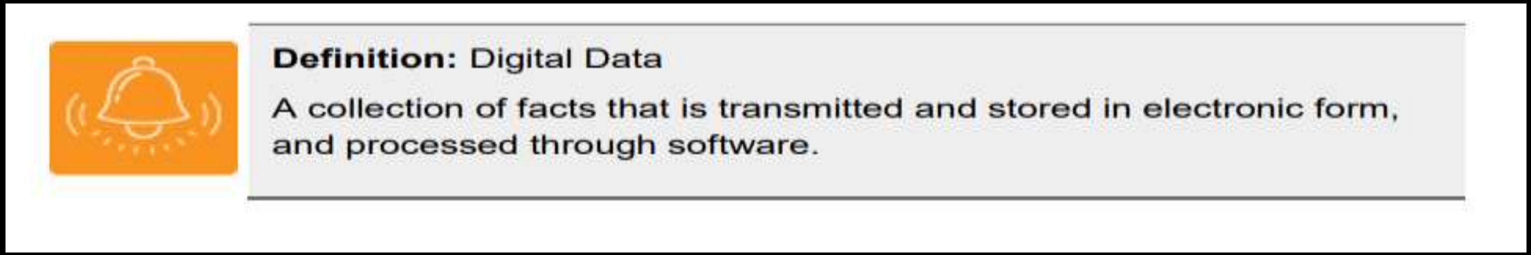
ANALYTICS WITH ANAND
LET THE DATA DO THE TALKING...

LETS TALK ABOUT DATA WAREHOUSING

@AnalyticsWithAnand

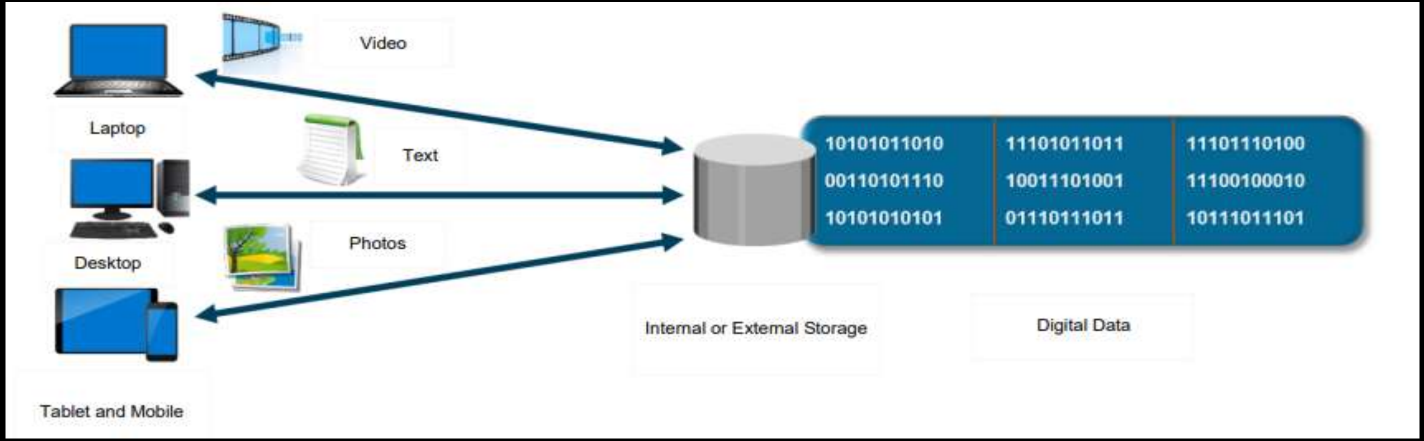
DIGITAL DATA



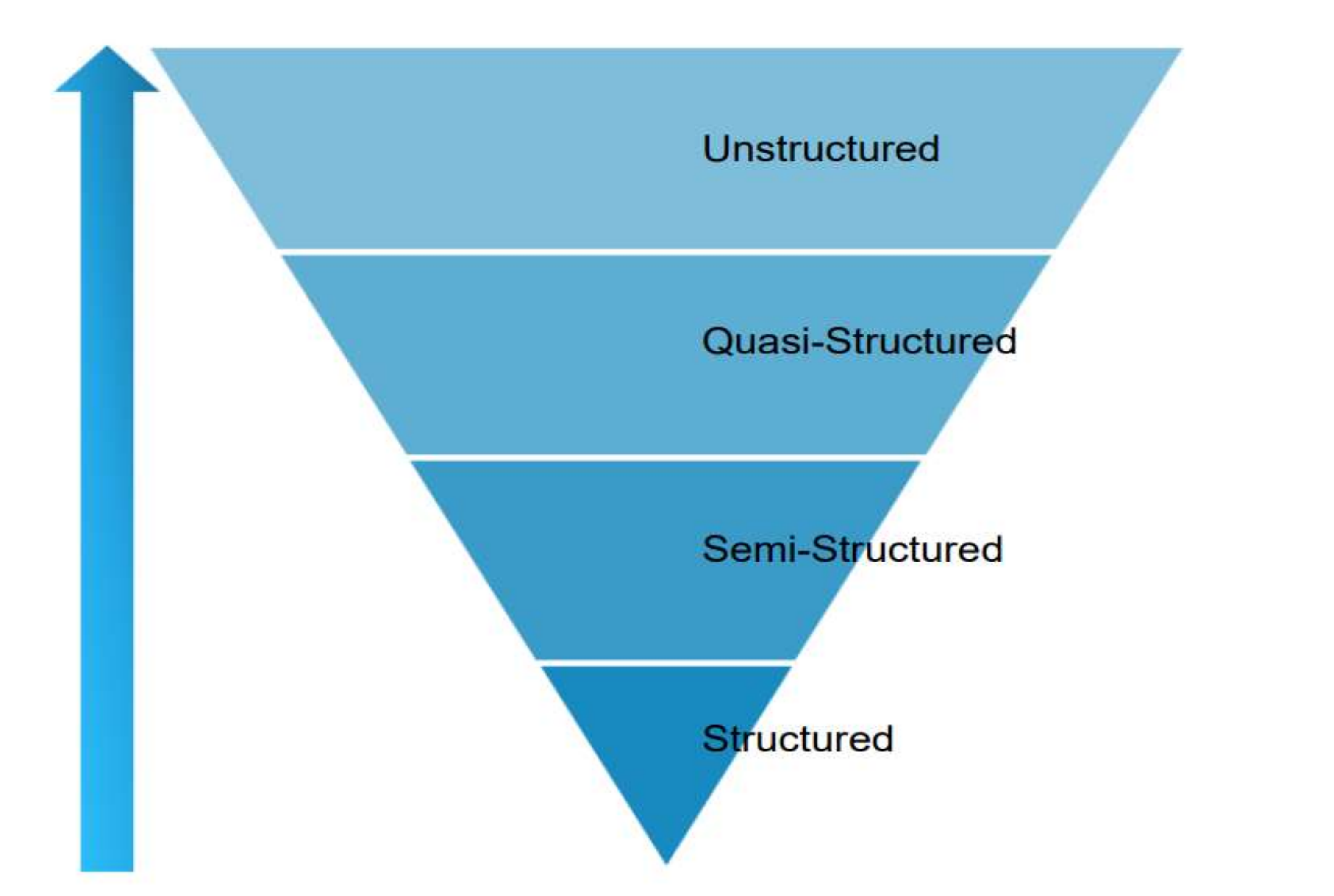
Definition: Digital Data

A collection of facts that is transmitted and stored in electronic form, and processed through software.

A generic definition of data is that it is a collection of facts, typically collected for analysis or reference. Data can exist in various forms such as facts stored in a person's mind, photographs and drawings, a bank ledger, and tabled results of a scientific survey. Digital data is a collection of facts that is transmitted and stored in electronic form, and processed through software. Devices such as desktops, laptops, tablets, mobile phones, and electronic sensors generate digital data



TYPES OF DATA



- **Structured data :** It is organized in fixed fields within a record or file. To structure the data, you require a data model. A data model specifies the format for organizing data, and also specifies how different data elements are related to each other. For example, in a relational database, data is organized in rows and columns within named tables.

- **Semi-structured data :** It does not have a formal data model but has an apparent, self-describing pattern and structure that enable its analysis. Examples of semistructured data include spreadsheets that have a row and column structure, and XML files that are defined by an XML schema.

- **Quasi-structured data:** It consists of textual data with erratic data formats, and can be formatted with effort, software tools, and time. An example of quasistructured data is a “clickstream” that includes data about which webpages a user visited and in what order – which is the result of the successive mouse clicks the user made. A clickstream shows when a user entered a website, the pages viewed, the time that is spent on each page, and when the user exited.

- **Unstructured data :** It does not have a data model and is not organized in any particular format. Some examples of unstructured data include text documents, PDF files, emails, presentations, images, and videos.

DATAWAREHOUSE

- A data warehouse is a central repository of integrated data that is gathered from multiple different sources. It stores current and historical data in a structured format. It is designed for query and analysis to support the decision-making process of an organization. For example, a data warehouse may contain current and historical sales data that is used for generating trend reports for sales comparisons.
- It is like a hub which contain all your database.
- Data Warehouses are required for queries, as well as all DML operations, including loading data into tables.
- Data Warehouses can be started and stopped at any time. They can also be resized at any time, even while running, to accommodate the need for more or less compute resources, based on the type of operations being performed by the warehouse.

DATA LAKE

A data lake is a centralized repository that allows organizations to store vast amounts of structured, semi-structured, and unstructured data at any scale. It is designed to handle massive volumes of data in its raw form, without requiring a predefined schema or structure. The concept of a data lake emerged as a response to the limitations of traditional data warehouses, which often struggle to handle the variety, volume, and velocity of modern data sources.

DATA MART

A Data Mart is a subset of a data warehouse that is focused on a specific business function, department, or topic area. Unlike a data warehouse, which stores comprehensive and integrated data from various sources across an organization, a data mart contains a more tailored and condensed set of data that is relevant to a particular group of users or a specific analytical purpose.

DATABASE

- A database is a collection of data that is organized, which is also called structured data. It can be accessed or stored in a computer system.
- Each database consists of one or more schemas, which are logical groupings of database objects, such as tables and views.

- There are different kinds of databases:

Relational Database:

A relational database is made up of a set of tables with data that fits into a predefined category.

Distributed Database:

A distributed database is a database in which portions of the database are stored in multiple physical locations, and in which processing is dispersed or replicated among different points in a network.

SCHEMA

- A schema is **a collection of database objects like tables, triggers, stored procedures, etc**. A schema is connected with a user which is known as the schema owner.

- A schema is a logical grouping of database objects (tables, views, etc.). Each schema belongs to a single database.

- Databases and schemas are used **to organize data stored in Snowflake.**

VIEW

- A view is **a subset of a database that is generated from a user query and gets stored as a permanent object.**

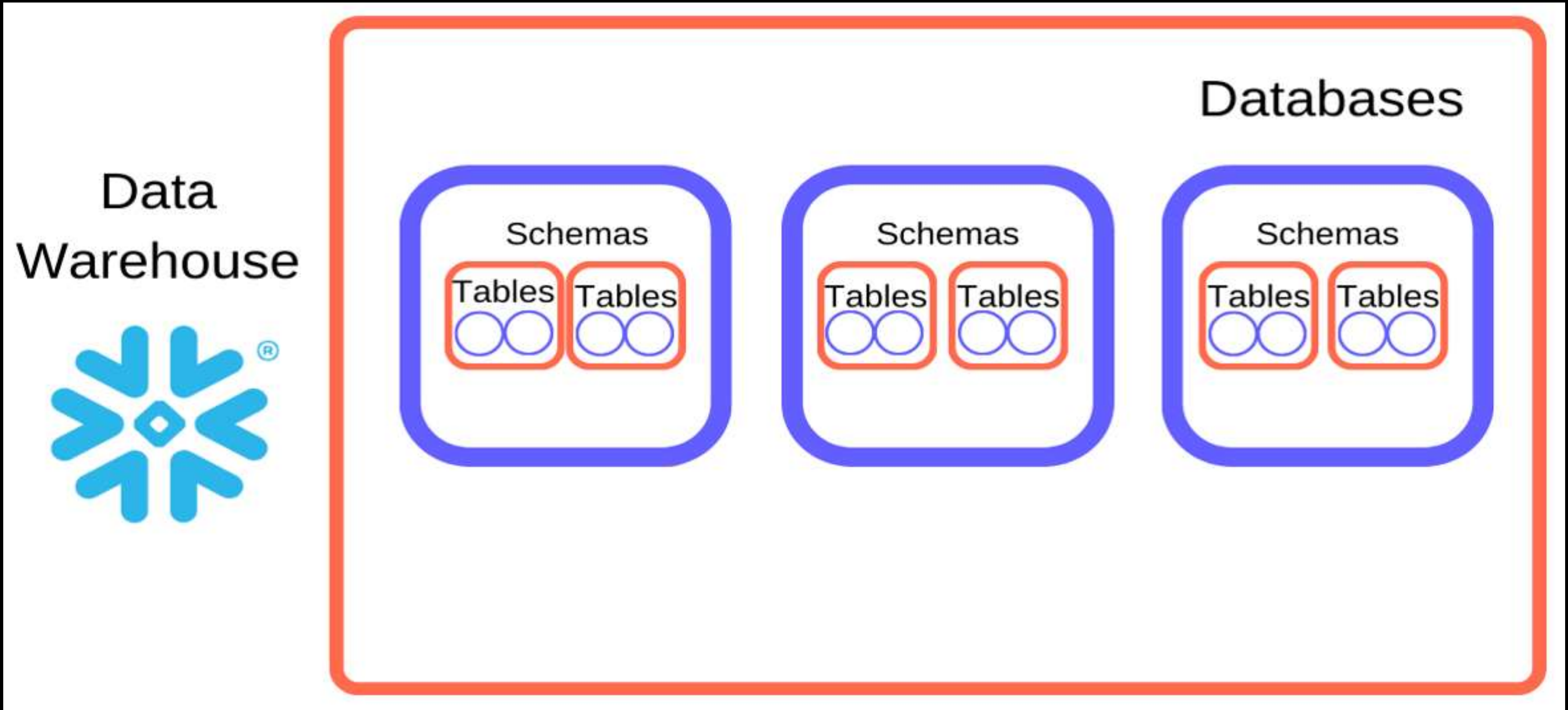
- Views serve a variety of purposes, including combining, segregating, and protecting data.

- A view is **a virtual table whose contents are defined by a query**.

TABLE

- Tables are **database objects that contain all the data in a database**.

- In tables, data is logically organized in a row-and-column format similar to a spreadsheet. Each row represents a unique record, and each column represents a field in the record.



Data Warehouse

Databases

Schemas

Tables

Tables

Schemas

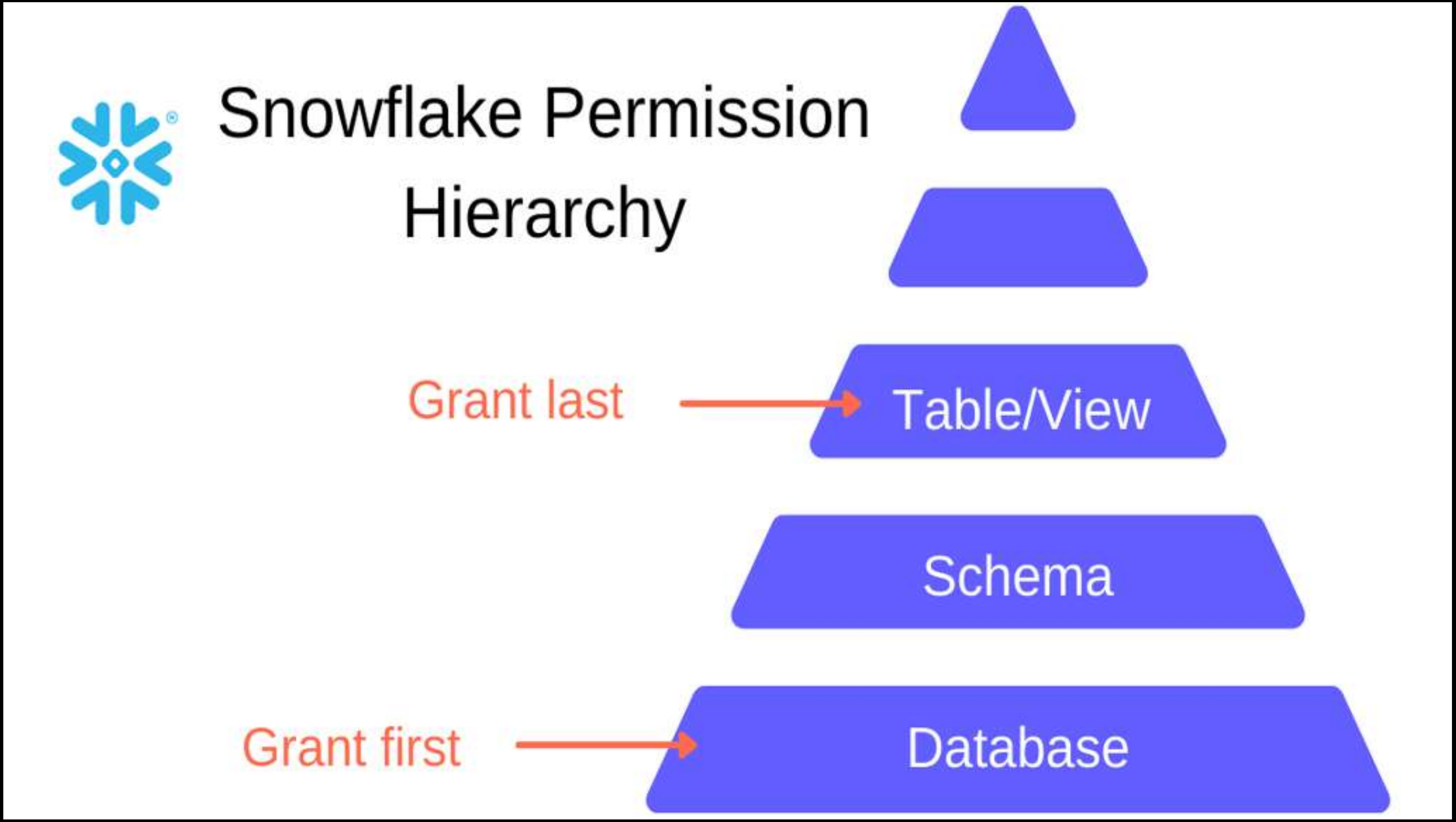
Tables

Tables

Schemas

Tables

Tables



Snowflake Permission Hierarchy

Grant last → Table/View

Schema

Grant first → Database

Thank You

@AnalyticsWithAnand