FACULTY OF ENGINEERING
PRINCE OF SONGKLA UNIVERSITY

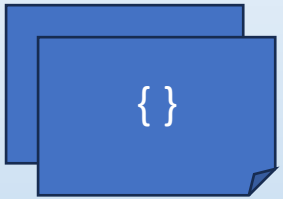
Basic Data Engineering

By Assoc. Prof. Dr. Pichaya Tandayya

Outline

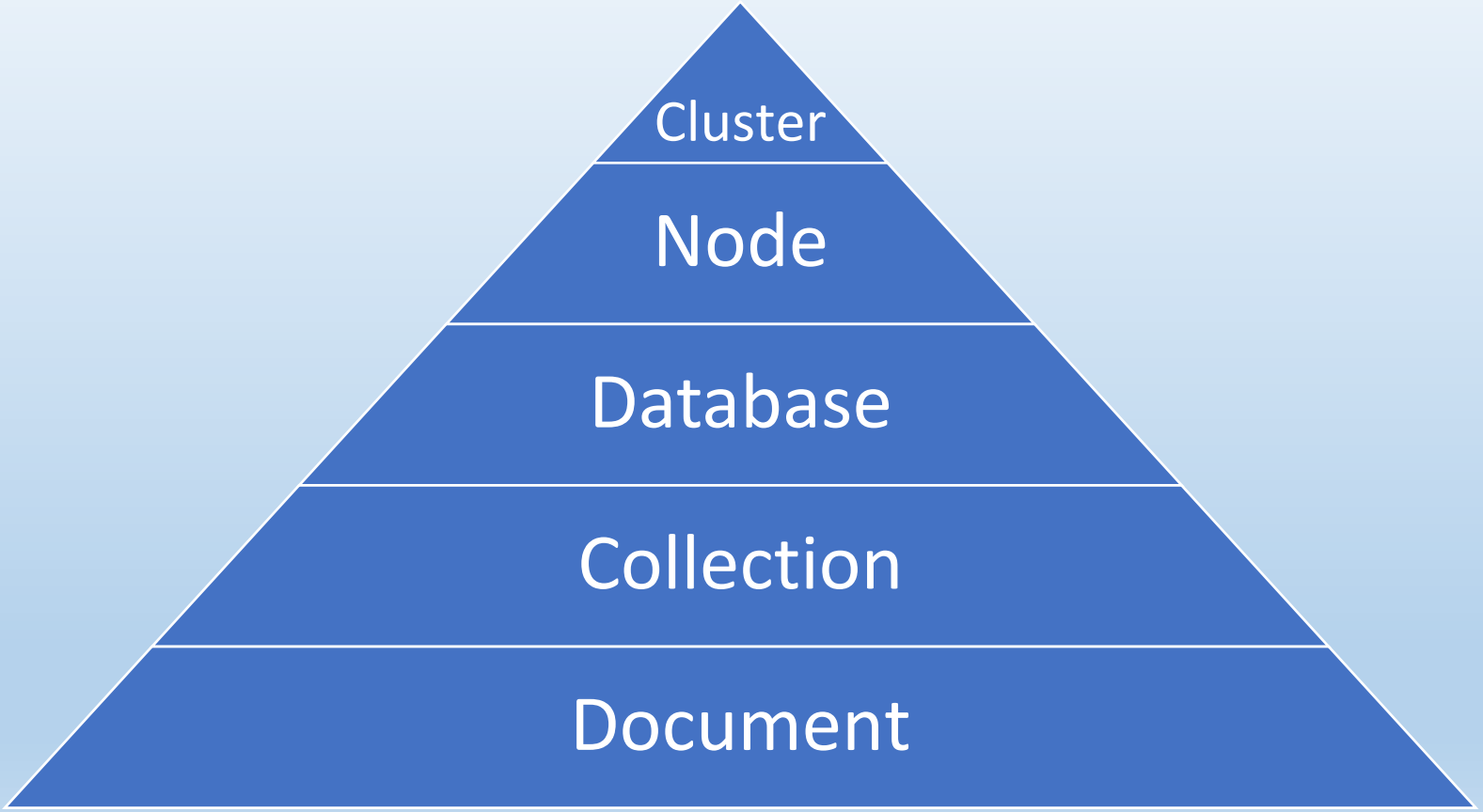
- Document structure
- JSON vs BSON
- Polymorphic data
- Limits

Document structure




```
{
    title: ‘Goldfinger’,
    year: 1968,
    plot: “Invertigating a gold magnate’s smuggling, James Bond
    uncovers a plot to contaminate the Fort Knox gold reserve.”
}
```

Document structure



Document structure

Field:

- Strings
- Unique
- Descriptive

☑ username: “pichaya”
☒ un: “pichaya”

Document Example

FACULTY OF ENGINEERING
PRINCE OF SONGKLA UNIVERSITY

MongoDB Document Example

```
{
  _id: ObjectId('573a1395f29313caabce2360'),
  plot: "Investigating a gold magnate's smuggling, James Bond uncovers a plot to
contaminate the Fort Knox gold reserve.",
  genres: [ 'Action', 'Adventure', 'Thriller' ],
  runtime: 110,
  rated: 'APPROVED',
  cast: [ 'Sean Connery', 'Honor Blackman', 'Gert Frèbe', 'Shirley Eaton' ],
  title: 'Goldfinger',
  languages: [ 'English', 'Chinese', 'Spanish' ],
  released: ISODate('1965-01-09T00:00:00.000Z'),
  directors: [ 'Guy Hamilton' ],
  writers: [ 'Richard Maibaum (screenplay)', 'Paul Dehn (screenplay)' ],
  lastupdated: '2015-09-06 00:04:52.777000000',
  year: 1964,
  imdb: { rating: 7.8, votes: 128247, id: 58150 }
  countries: [ 'UK' ],
  type: 'movie'
}
```

Document Example

MongoDB Document Example

```
{
  _id: ObjectId('573a1395f29313caabce2360'),
  plot: "Investigating a gold magnate's smuggling, James Bond uncovers a plot to
contaminate the Fort Knox gold reserve.",
  genres: [ 'Action', 'Adventure', 'Thriller' ],
  runtime: 110,
  rated: 'APPROVED',
  cast: [ 'Sean Connery', 'Honor Blackman', 'Gert Frèbe', 'Shirley Eaton' ],
  title: 'Goldfinger',
  languages: [ 'English', 'Chinese', 'Spanish' ],
  released: ISODate('1965-01-09T00:00:00.000Z'),
  directors: [ 'Guy Hamilton' ],
  writers: [ 'Richard Maibaum (screenplay)', 'Paul Dehn (screenplay)' ],
  lastupdated: '2015-09-06 00:04:52.777000000',
  year: 1964,
  imdb: { rating: 7.8, votes: 128247, id: 58150 }
  countries: [ 'UK' ],
  type: 'movie'
}
```

- array

Document Example

FACULTY OF ENGINEERING
PRINCE OF SONGKLA UNIVERSITY

MongoDB Document Example

```
{
  _id: ObjectId('573a1395f29313caabce2360'),
  plot: "Investigating a gold magnate's smuggling, James Bond uncovers a plot to
contaminate the Fort Knox gold reserve.",
  genres: [ 'Action', 'Adventure', 'Thriller' ],
  runtime: 110,
  rated: 'APPROVED',
  cast: [ 'Sean Connery', 'Honor Blackman', 'Gert Frèbe', 'Shirley Eaton' ],
  title: 'Goldfinger',
  languages: [ 'English', 'Chinese', 'Spanish' ],
  released: ISODate('1965-01-09T00:00:00.000Z'),
  directors: [ 'Guy Hamilton' ],
  writers: [ 'Richard Maibaum (screenplay)', 'Paul Dehn (screenplay)' ],
  lastupdated: '2015-09-06 00:04:52.777000000',
  year: 1964,
  imdb: { rating: 7.8, votes: 128247, id: 58150 },
  countries: [ 'UK' ],
  type: 'movie'
}
```

- nested

Document structure

Values:

- Strings
- Numbers
- Booleans
- Arrays
- Document Objects

JSON vs BSON

JSON

> { “hello” : “world”}

BSON

\x16\x00\x00\x00

\x02hello\x00

\x06\x00\x00\x00world\x00

\x00

BSON

Extension of JSON

Additional data types

Dates

ObjectId

Timestamps

Flexible Schema

```
{
    _id: ObjectId('573a1399f29313caabcec07b'),
    imdb: { rating: 6, votes: 99874, id: 99938 },
    year: 1990,
    genres: [ 'Action', 'Comedy', 'Crime' ],
    rated: 'PG-13',
    metacritic: 61,
    title: 'Kindergarten Cop',
    lastupdated: '2015-09-15 03:35:02.090000000',
    languages: [ 'English', 'Spanish' ],
    type: 'movie',
    released: ISODate('1990-12-21T00:00:00.000Z'),
    countries: [ 'USA' ],
    cast: [
      'Arnold Schwarzenegger',
      'Penelope Ann Miller',
      'Pamela Reed',
      'Linda Hunt'
    ],
    directors: [ 'Ivan Reitman' ],
    runtime: 111
}
```

```
{
  _id: ObjectId('573a1395f29313caabce2498'),
  imdb: { rating: 8.1, votes: 126585, id: 58461 },
  year: 1964,
  genres: [ 'Action', 'Drama', 'Western' ],
  rated: 'R',
  title: 'A Fistful of Dollars',
  lastupdated: '2015-09-02 00:17:22.303000000',
  languages: [ 'Italian', 'Spanish', 'English' ],
  type: 'movie',
  released: ISODate('1967-01-18T00:00:00.000Z'),
  countries: [ 'Italy', 'Spain', 'West Germany' ],
  cast: [
    'Clint Eastwood',
    'Marianne Koch',
    'Gian Maria Volontè',
    'Wolfgang Lukschy'
  ],
  directors: [ 'Sergio Leone' ],
  runtime: 99,
}
```

Flexible Schema

```
> db.movies.find()

{
    _id: ObjectId('573a1393f29313caabcdcb42'),
    genres: [ 'Comedy', 'Fantasy', 'Romance' ],
    runtime: 118,
    rated: 'PG-13',
    cast: [ 'Meg Ryan', 'Hugh Jackman', 'Liev Schreiber', 'Breckin Meyer' ],
    title: 'Kate & Leopold',
    directors: [ 'James Mangold' ],
    year: 2001,
    type: 'movie'
  },
  {
    _id: ObjectId('573a139af29313caabcf0f07'),
    year: 2001,
    genres: [ 'Adventure', 'Fantasy' ],
    rated: 'PG-13',
    title: 'The Lord of the Rings: The Fellowship of the Ring',
    type: 'movie',
    cast: [ 'Alan Howard', 'Noel Appleby', 'Sean Astin', 'Sala Baker' ],
    directors: [ 'Peter Jackson' ],
    runtime: 178
  }
```

- The order of fields can be diffrent

Flexible Schema

```
db.movies.insertOne({
  title: 'Snatch',
  year: 2000,
  genres: [ 'Comedy', 'Crime'],
  rated: 'R',
  runtime: 102,
  type: 'movie',
  cast: [ 'Jason Statham', 'Brad Pitt', 'Stephen Graham'],
  directors: [ 'Guy Ritchie' ],
  filming_locations: [ 'London, UK', 'Buckinghamshire, UK', 'Hertfordshire, UK' ]
})
```

- Fields in each document can be different

Document structure

Key-value pairs:

- Text
- Geospatial data
- Time-series
- Graph data

Document limits

Maximum document size: 16 MB

Maximum levels of nesting: 100

You have learnt about

Structure

Field-value pairs

Arrays

Embedded documents

JSON vs BSON

Flexible schema

Limits

Quiz

When naming fields in MongoDB documents … ?

A. Using short, abbreviated names to conserve space on disk

B. Re-using field names within a document to optimize indexes

C. Using descriptive, unique names

D. Using generic names

Quiz

What is the key characteristic of MongoDB's document model that allows for handling polymorphic data, or data of different shapes and types?

A. Maximum document size of 16 MB

B. Similarity to JSON objects

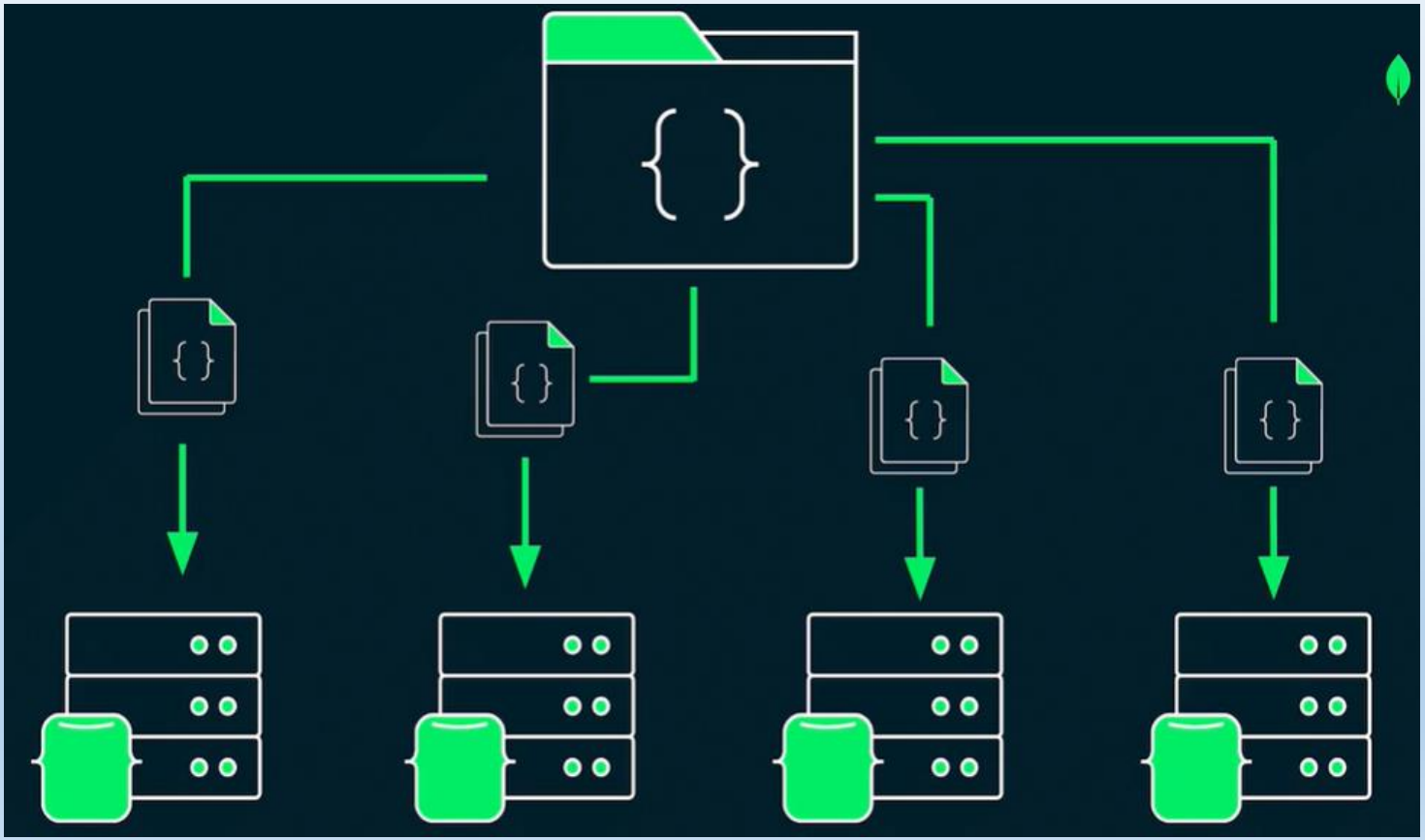
C. Flexible schema

D. Single data type storage

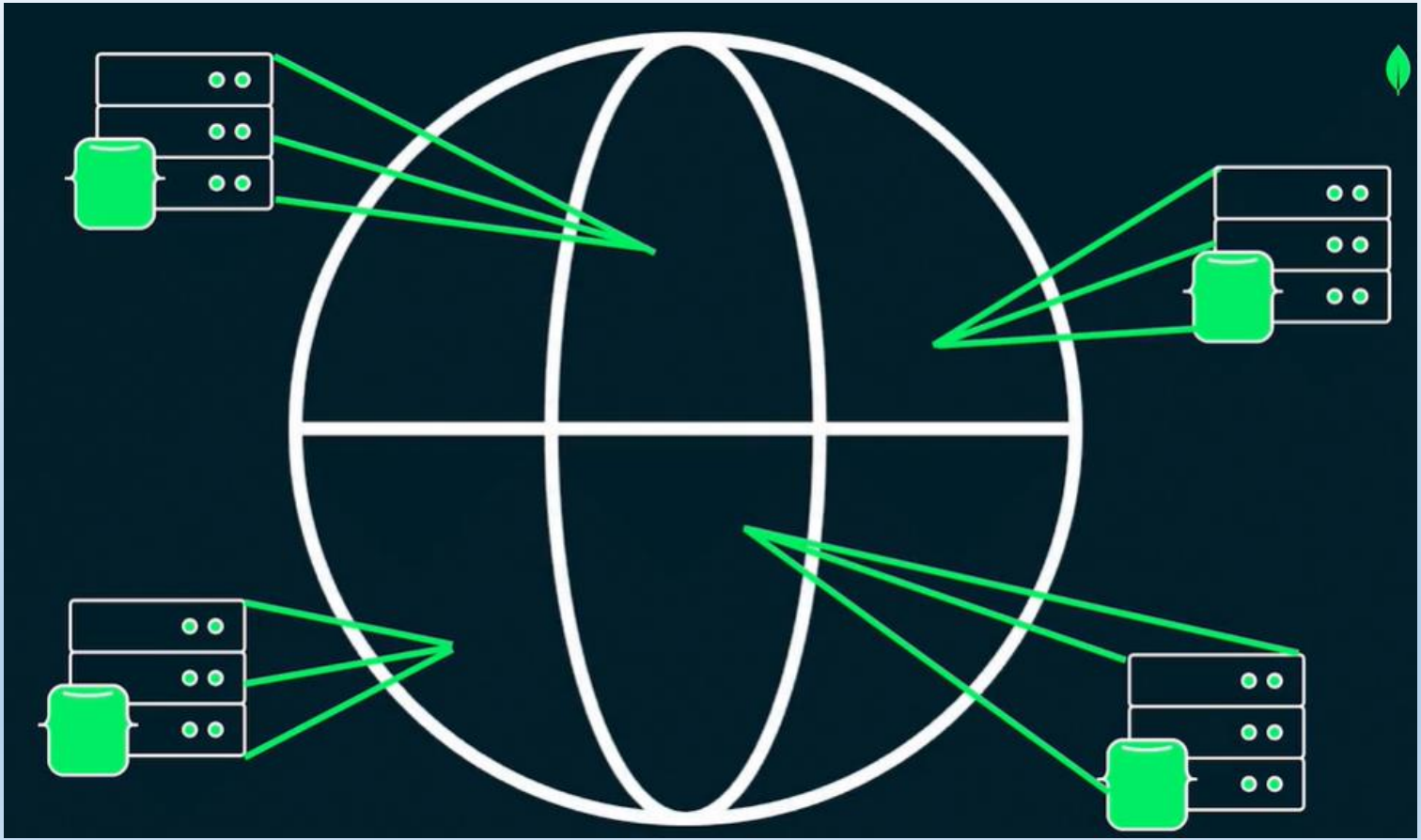
Advantages

- Distributed system architecture
- Distributed database
- Document model
- Flexible schema

Distributed System Architecture



Distributed System Architecture



Distributed database

- Data shared across machines
- If one machine fails, system can still function
- Consistent service and reliability

Distributed database

- Store data close to the source
- Reduce latency

Document Model

- Similar to Javascript Object Notation

```
{
  "_id": 1,
  "name": {
    "first": "Ada",
    "last": "Lovelace"
  },
  "title": "The First Programmer",
  "interests": ["mathematics", "programming"]
}
```

Document Model

```
{
  "_id": 1,
  "name": {
    "first": "Ada",
    "last": "Lovelace"
  },
  "title": "The First Programmer",
  "interests": ["mathematics", "programming"],

  "address": "100 St John Street London",
  "parents": ["Lord Byron","Lady Byron"]

}
```

Polymorphic data

- Data that can take on multiple types within the same structure

Flexible schema

- Each data record can have a unique structure
- Permit various data types
- Unstructured or semi-structure datatype

Social Media App

{

Common Fields:

user_id

timestamp

likes

}

Social Media App: Posts collection

FACULTY OF ENGINEERING
PRINCE OF SONGKLA UNIVERSITY

{

Text Posts:
- content
- user_id

Photo Post:
- photo_url
- caption

Video Post:
- video_url
- title
- duration

Live Stream Post:
- date_filmed

}

Other Applications

- IoT Applications
- Mobile Apps
- Content Management
- AI

MongoDB Architecture

- CAP Theorem
- High availability replication
- Consistency: read and write concerns
- Scaling: sharding

- Document
- Object and any related metadata
- Field-value pairs
- Data types: strings, numbers, dates, arrays, objects, and more

Collection

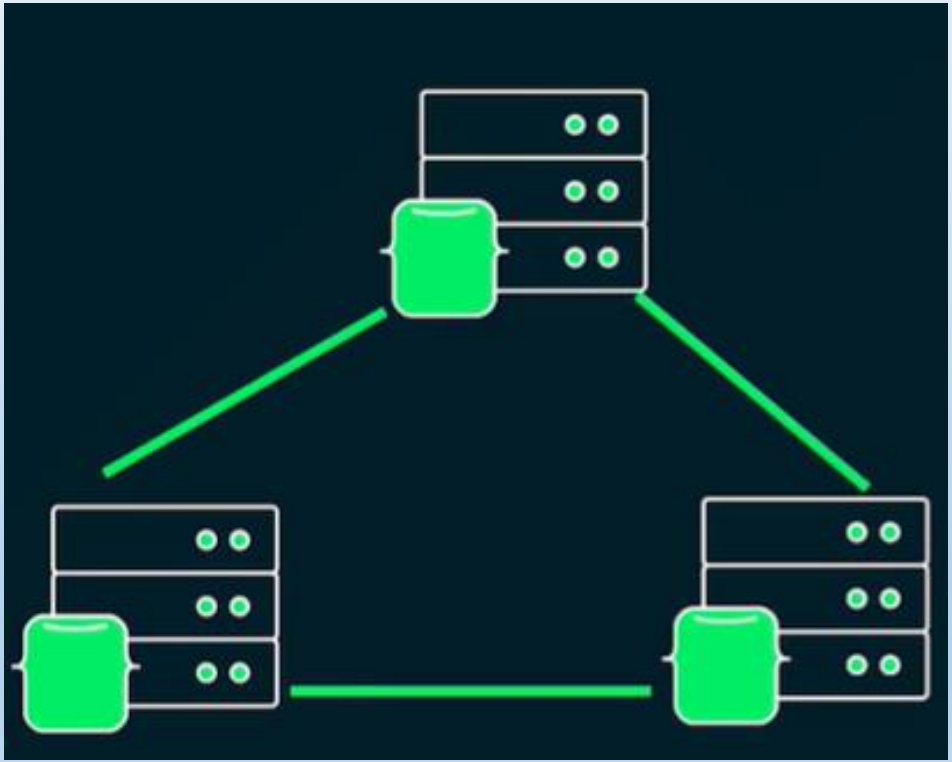
A collection is a group of documents that correspond to an entity

Can support multiple entities and different shapes

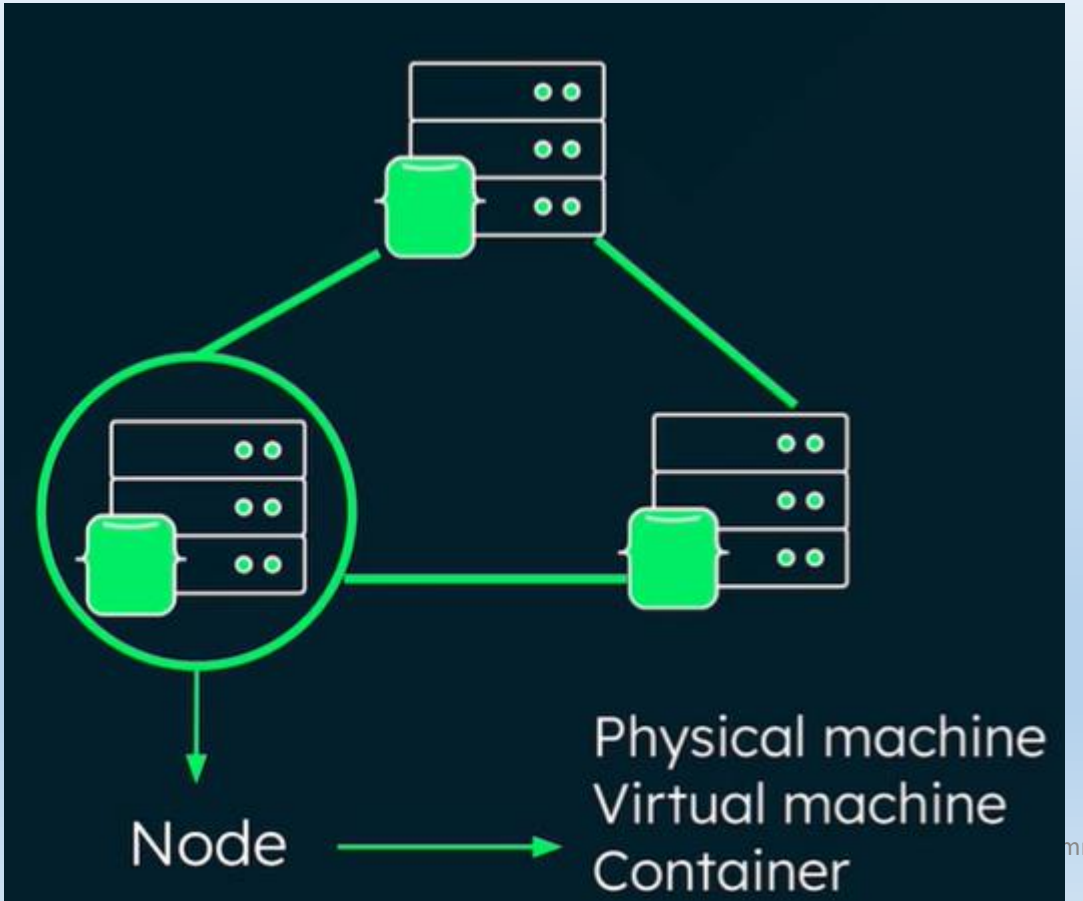
Database

A database is a group of collections

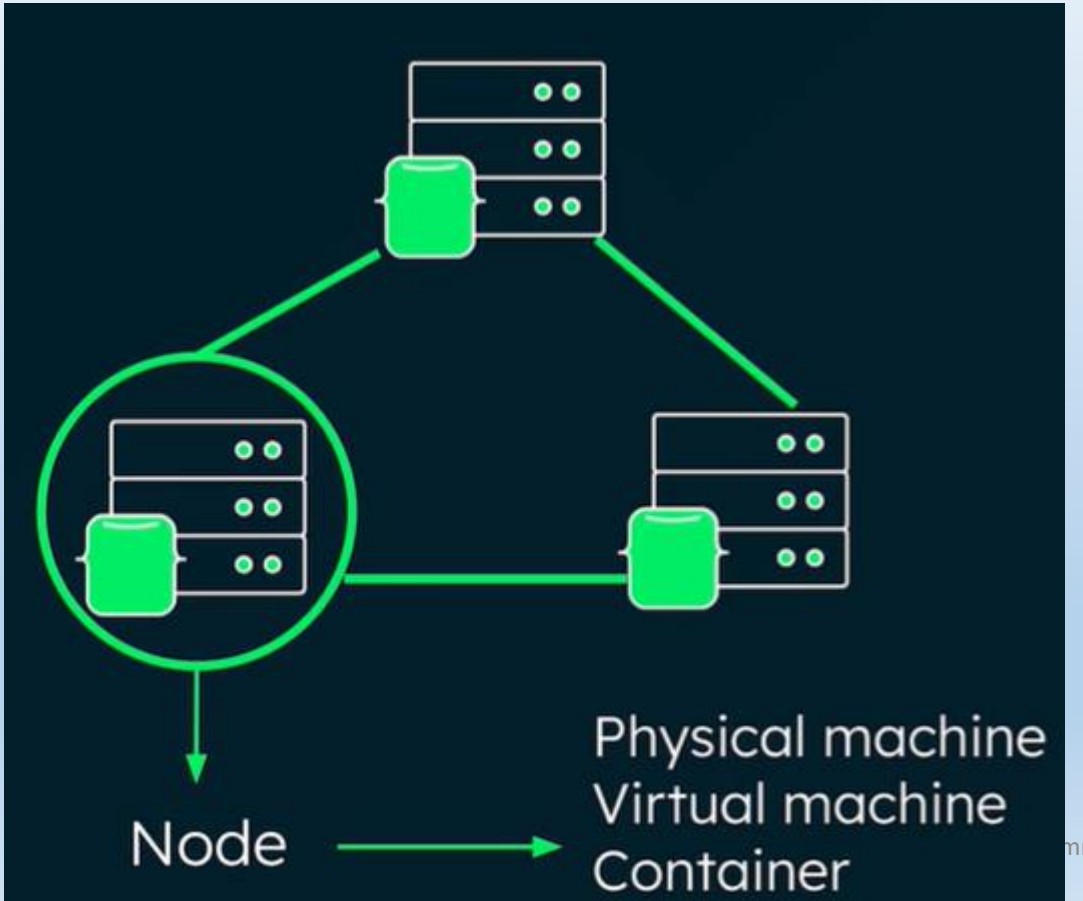
Node and Cluster



Node and Cluster

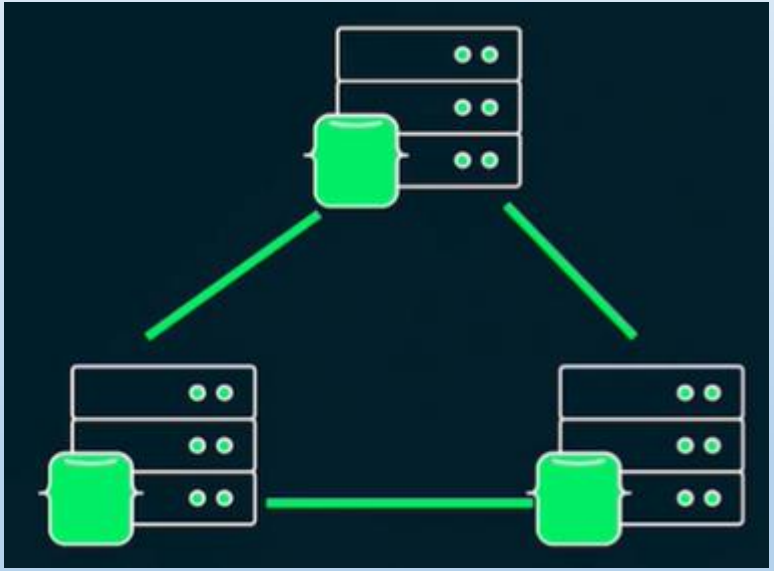


Node and Cluster

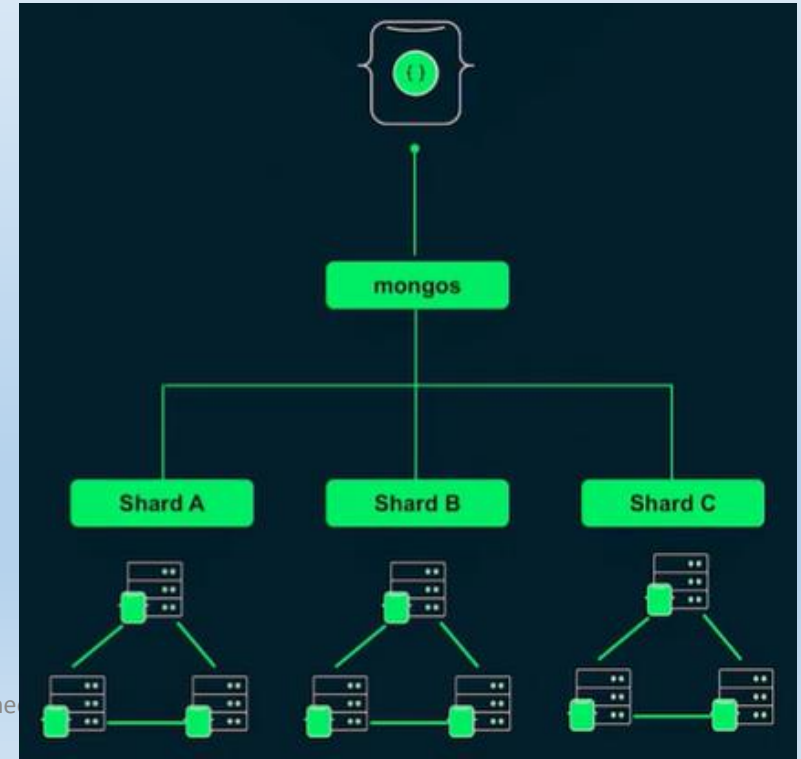


Cluster Types

Replica Set



Shared cluster



Cluster Types

FACULTY OF ENGINEERING
PRINCE OF SONGKLA UNIVERSITY

Replica Set

- High availability

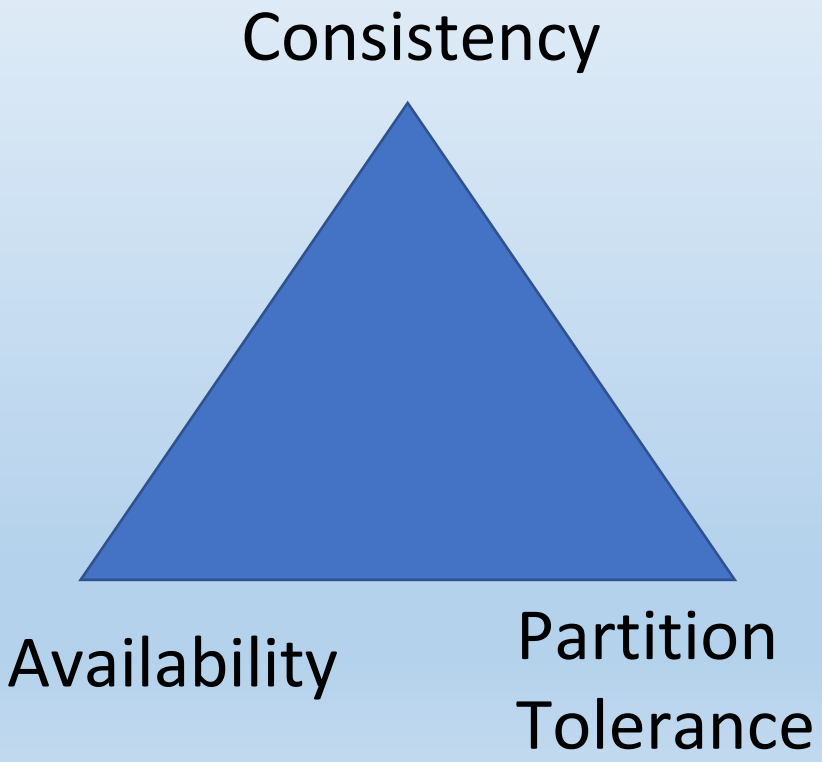
Shared cluster

- Scaling

CAP Theorem

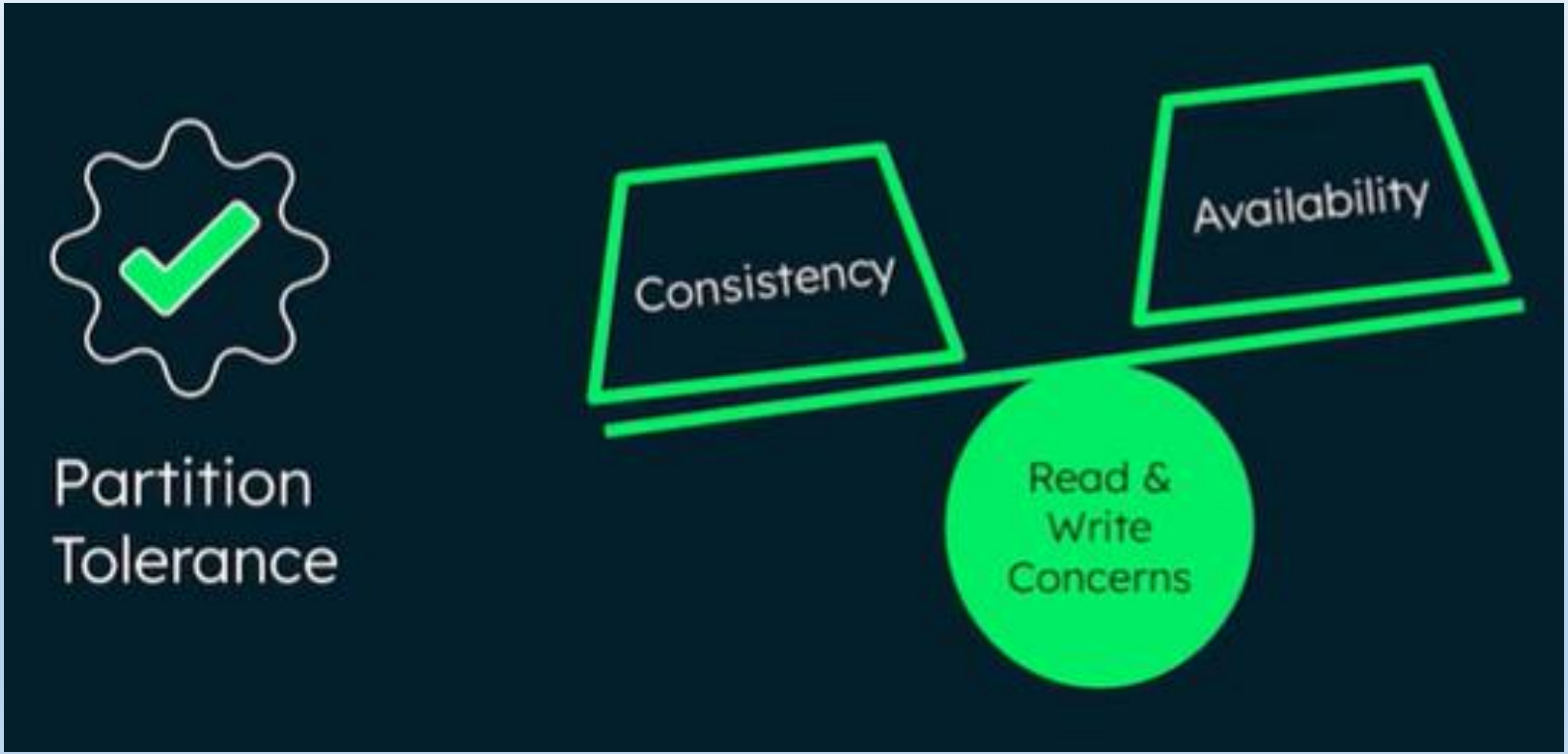
- It is possible to simultaneously guarantee consistency, availability, and partition tolerance.

CAP Theorem

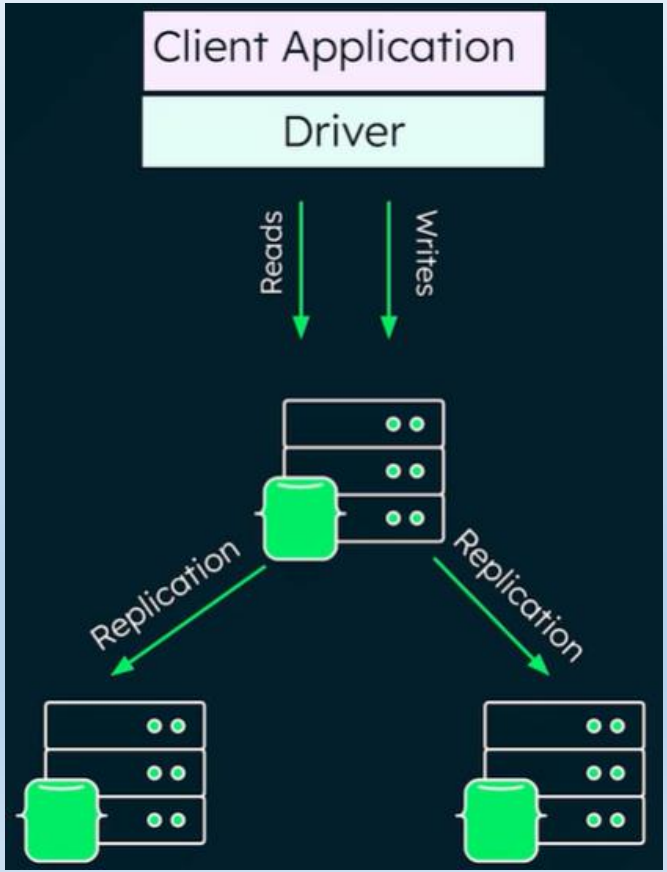


- Consistency: Every read receives the most recent write or an error
- Availability: Every request (read or write) receives a response
- Partition Tolerance: The system continues to function despite network partitions

MongoDB and the CAP Theorem



Replcation


Client Application
Driver
Reads
Writes
Replication
Replication


Client Application
Driver
Reads
Writes
Replication