

NoSQL Databases for Data Scientists

Lecture 3 : MongoDB and Elasticsearch

(Data Storage and Processing at Scale)

Hassan Ahmadkhani

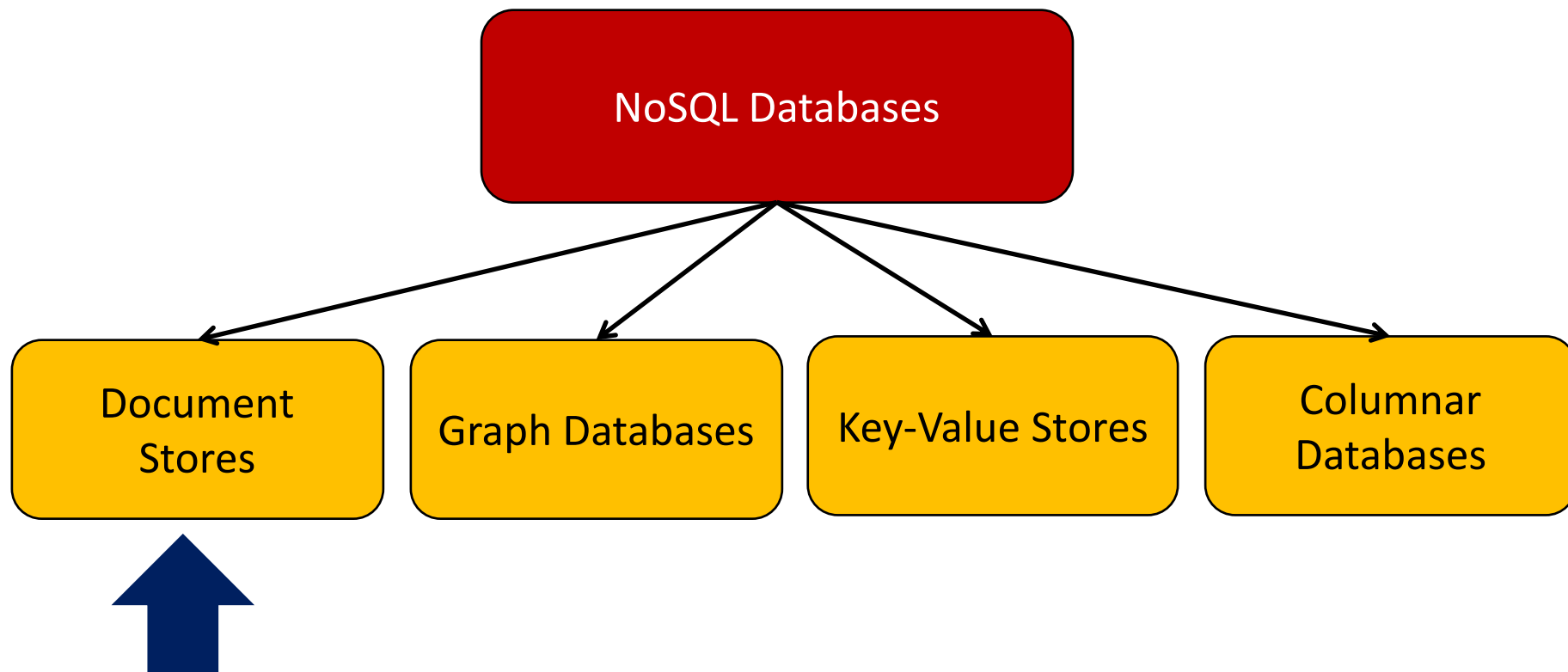
گروه دایکه . dayche.com





Types of NoSQL Databases

□ Taxonomy of NoSQL databases:



تولید محتوا: حسن احمدخانی

daychegroup

daychegroup

dayche.com | گروه دایکه

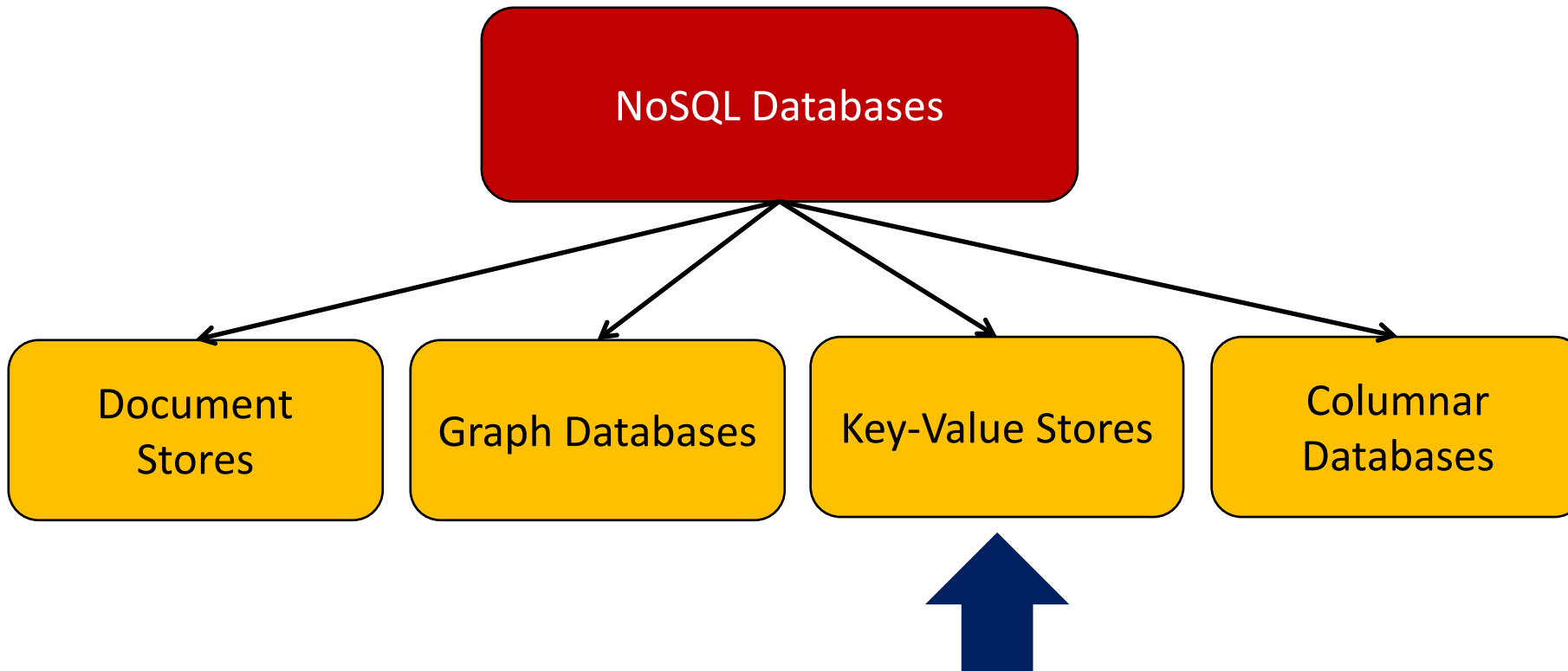


Document Stores

- ❑ Documents are stored in some standard format or encoding (e.g., XML, JSON, ...)
 - These are typically referred to as Binary Large Objects (BLOBs)
- ❑ Documents can be indexed
 - This allows document stores to outperform traditional file systems
- ❑ E.g., MongoDB and CouchDB, Elastic, ...

Types of NoSQL Databases

□ Taxonomy of NoSQL databases:



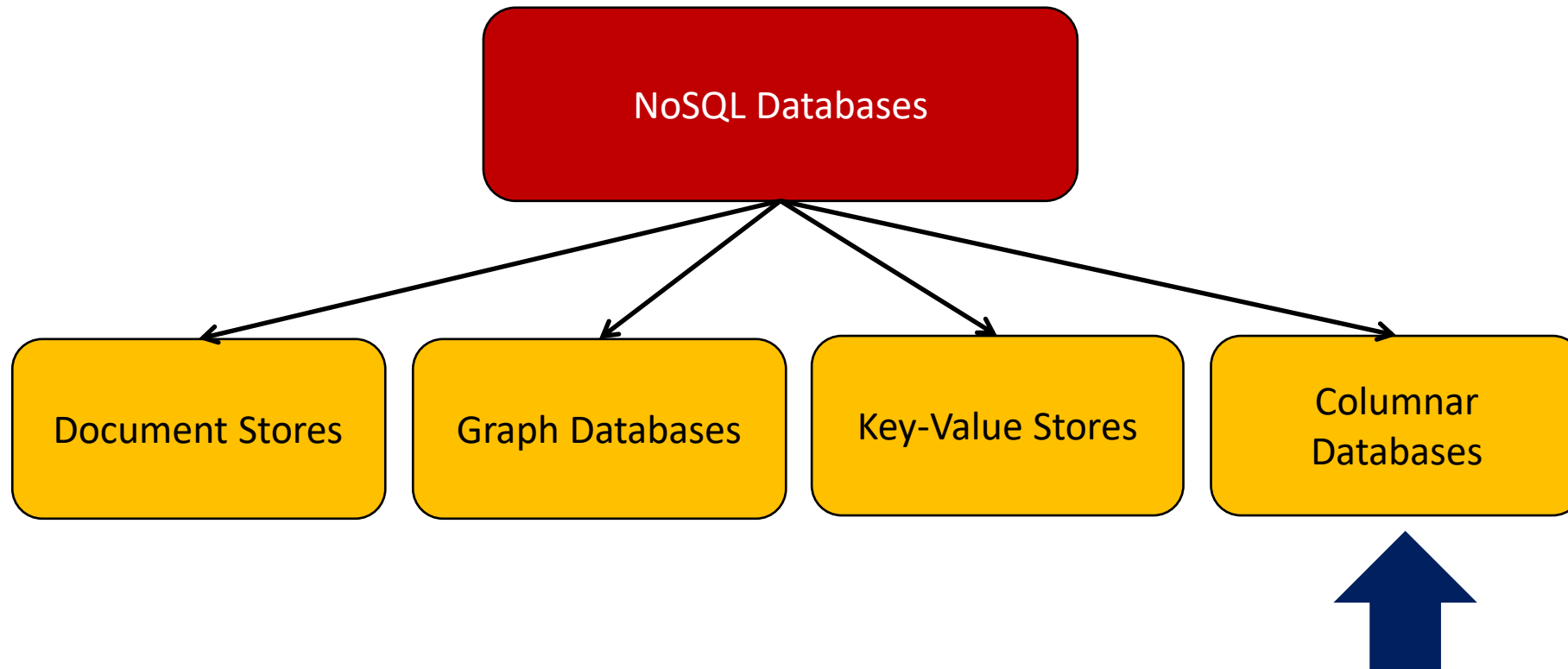


Key-Value Stores

- ❑ Keys are mapped to (possibly) more complex value (e.g., lists)
- ❑ Keys can be stored in a hash table and can be distributed easily
- ❑ Such stores typically support regular CRUD (create, read, update, and delete) operations
- ❑ E.g., Amazon **DynamoDB** (key value + document) , **Redis**, **Ignite** and etc.

Types of NoSQL Databases

□ Taxonomy of NoSQL databases:



تولید محتوا: حسن احمدخانی

daychegroup 

daychegroup 

dayche.com | گروه دایکه 



Columnar Databases

- ❑ Columnar databases are a hybrid of RDBMSs and Key-Value stores
 - Values are stored in groups of zero or more columns, but in Column-Order (as opposed to Row-Order)

Record 1

Alice	3	25	Bob
4	19	Carol	0
45			

Row-Order

Column A

Alice	Bob	Carol
3	4	0
19	45	

Columnar (or Column-Order)

Column A = Group A

Alice	Bob	Carol
3	25	4
0	45	19

Column Family {B, C}
Columnar with Locality Groups

- Values are queried by matching keys

- ❑ E.g., Apache **HBase**, Apache **Cassandra** , HP **Vertica** and etc.

تولید محتوا: حسن احمدخانی

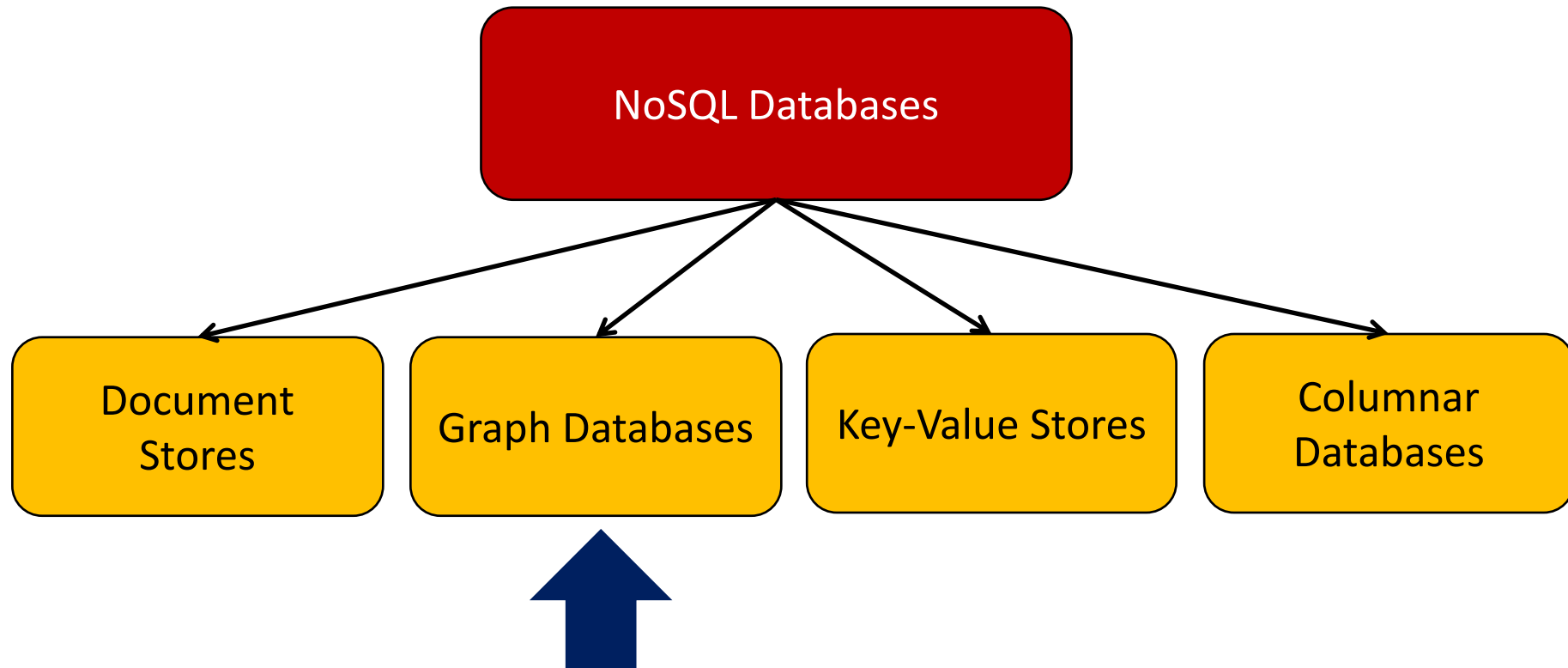
daychegroup

daychegroup

dayche.com | گروه دایکه

Types of NoSQL Databases

□ Taxonomy of NoSQL databases:



تولید محتوا: حسن احمدخانی

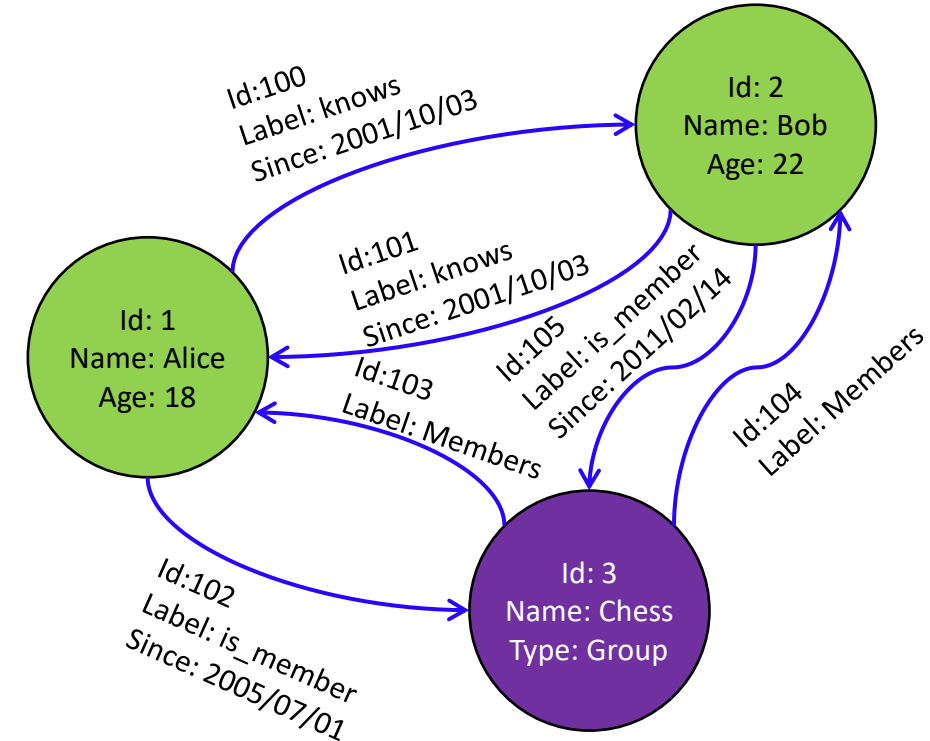
daychegroup 

daychegroup 

dayche.com | گروه دایکه 

Graph Databases

- ❑ Data are represented as vertices and edges



- ❑ Graph databases are powerful for graph-like queries (e.g., find the shortest path between two elements)

- ❑ E.g., **Neo4j** and **Vertex** DB



RDBMS , NoSQL , NewSQL and BigData

❑ RDBMS

- OLTP-ACID on Small and Mid Scale (DW if scale is Mid !?)

❑ NoSQL (Distributed)

- OLTP-Base (ACID !?) on Small , Mid and Large Scale with Scalability (DW ?)

❑ NewSQL (Distributed)

- OLTP-ACID on Small and Mid Scale with Scalability (DW ?)
- NuoDB , H-Store , ...

❑ Text Search – Text Indexing (Distributed)

- Solr – Splunk – ElasticSearch

❑ Big Data / Large Scale Data Warehouse, Data Lake and Lakehouse (Distributed)

- Hadoop Ecosystem

تولید محتوا: حسن احمدخانی

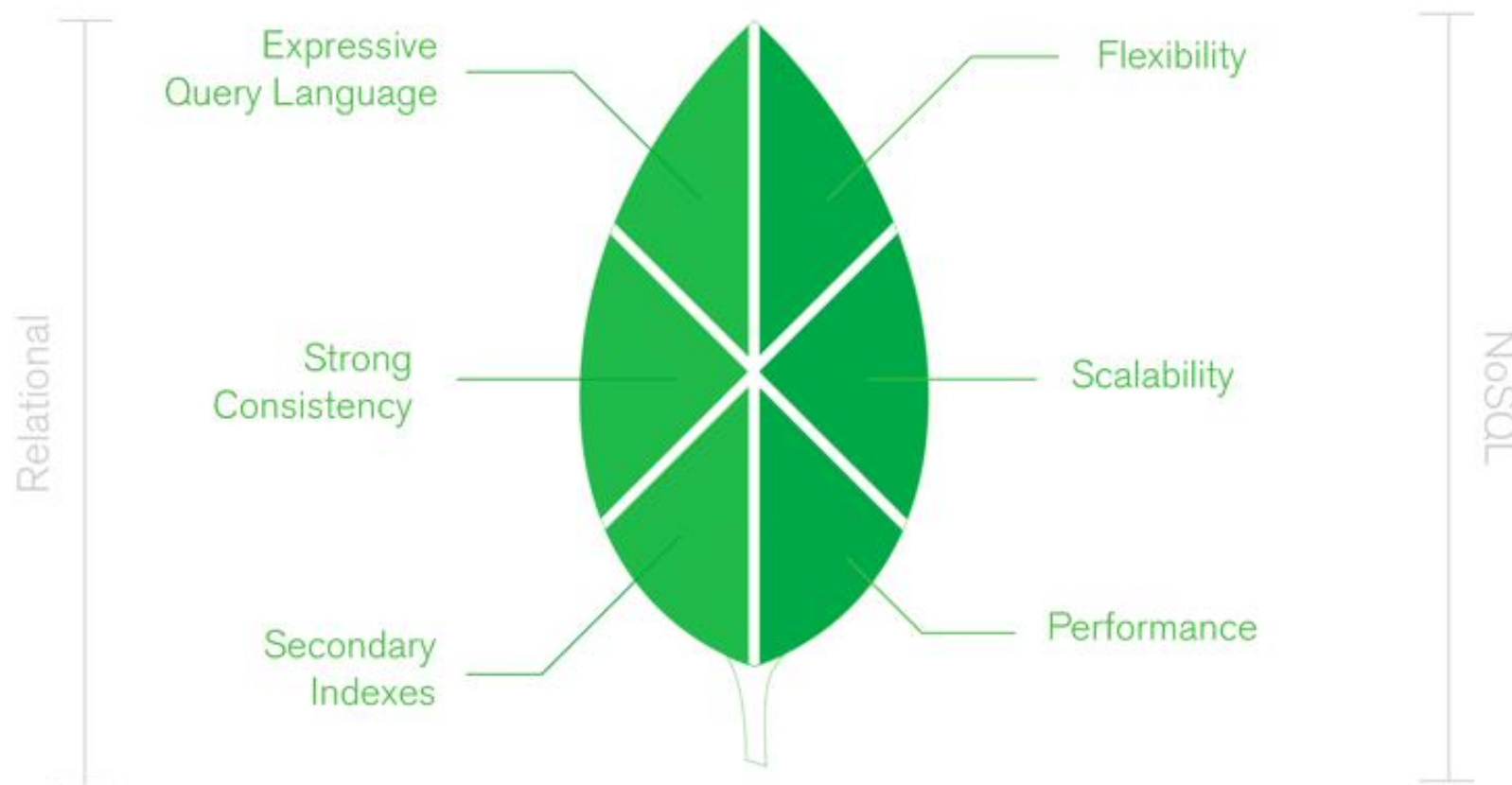
daychegroup 

daychegroup 

dayche.com | گروه دایکه 

Lecture 3 : NoSQL Databases

MongoDB



تولید محتوا: حسن احمدخانی

daychegroup 

daychegroup 

dayche.com | گروه دایکه 



JSON Data Model - JSON (JavaScript Object Notation)

```
{ _id: 1,  
  name: { first: "John", last: "Backus" },  
  birthyear: 1924,  
  contribs: [ "Fortran", "ALGOL", "Backus-Naur Form", "FP" ],  
  awards: [ { award_id: "NMSoo1",  
             year: 1975 },  
            { award_id: "TA99",  
              year: 1977 } ]  
}
```

MongoDB: Always indexed, automatically assigned unless provided

Array of documents

تولید محتوا: حسن احمدخانی

daychegroup

daychegroup

گروه دایکه | dayche.com

Lecture 3 : Mongo DB CRUD, Schema Design and Data Modeling



MongoDB Database

- ◆ **Database** = a number of “collections”
- ◆ **Collection** = a list of “documents”
- ◆ **Document** = a JSON object
- ◆ Must have an **_id** attribute whose value can uniquely identify a document within the collection
 - In other words, a database has collections of similarly structured “documents”

SQL Terms/Concepts	MongoDB Terms/Concepts
Database	Database
Table	Collection
Line	Document
Column	Field
Index	Index
Join	Embedded or linked document
Primary key	Primary key (start by « _id »)

تولید محتوا: حسن احمدخانی

daychegroup 

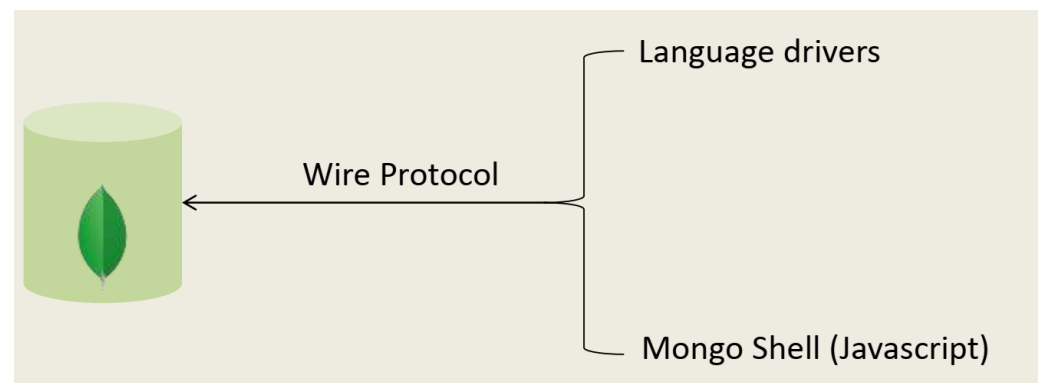
daychegroup 

dayche.com | گروه دایکه 

Lecture 3 : Mongo DB CRUD, Schema Design and Data Modeling



Operation	MongoDB	SQL
Create	Insert	Insert
Read	Find	Select
Update	Update	Update
Delete	Remove	Delete



تولید محتوا: حسن احمدخانی

daychegroup

daychegroup

dayche.com | گروه دایکه



Document Model – Normalization and Denormalization

PERSON RDBMS

Pers_ID	Surname	First_Name	City
0	Miller	Paul	London
1	Ortega	Alvaro	Valencia
2	Huber	Urs	Zurich
3	Blanc	Gaston	Paris
4	Bertolini	Fabrizio	Rome

CAR

Car_ID	Model	Year	Value	Pers_ID
101	Bently	1973	100000	0
102	Rolls Royce	1965	330000	0
103	Peugot	1993	500	3
104	Ferrari	2005	150000	4
105	Renault	1998	2000	3
106	Renault	2001	7000	3
107	Smart	1999	2000	2

Mongo DB

```
{
  first_name: 'Paul',
  surname: 'Miller',
  city: 'London',
  location: {
    type: 'Point',
    coordinates: [45.123, 47.232]
  },
  cars: [
    { model: 'Bentley',
      year: 1973,
      value: 100000, ... },
    { model: 'Rolls Royce',
      year: 1965,
      value: 330000, ... }
  ]
}
```

تولید محتوا: حسن احمدخانی

daychegroup

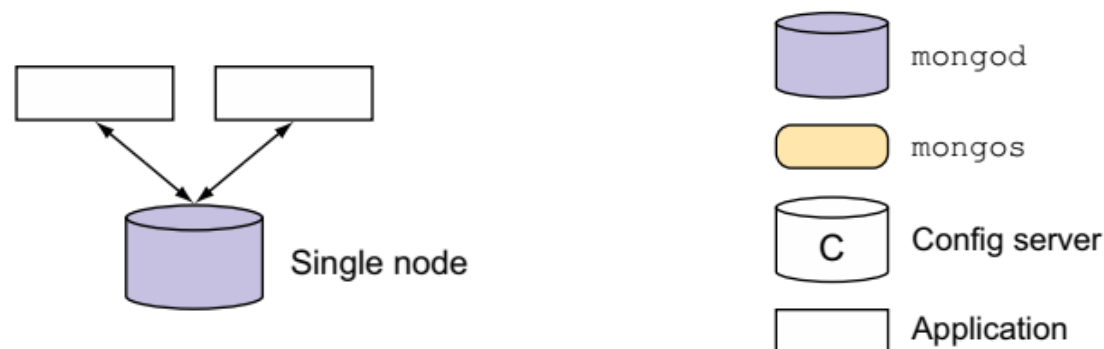
daychegroup

گروه دایکه | dayche.com

Lecture 3 : MongoDB Architecture and Replication (Replica Set)

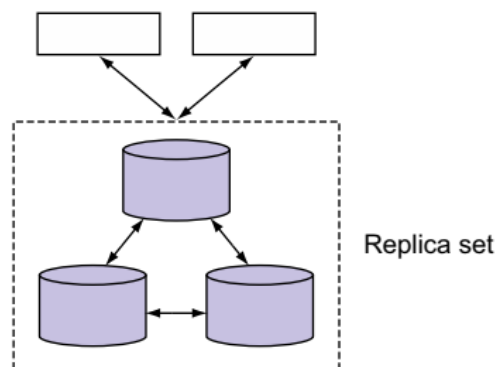
Cluster Topology

◆ Single node



◆ Replica Set

- minimum recommended deployment topology for a replica set is three nodes



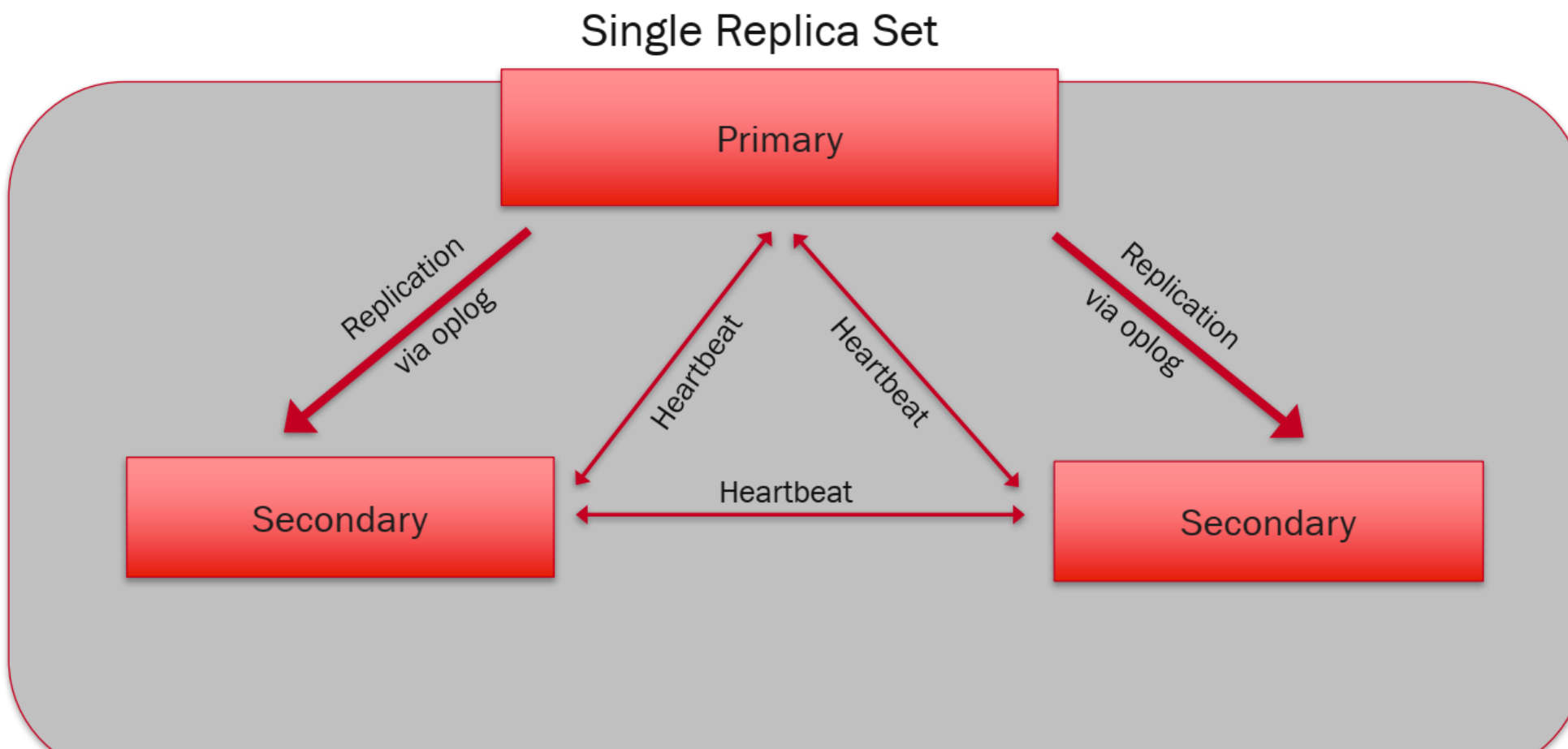
تولید محتوا: حسن احمدخانی

daychegroup 

daychegroup 

گروه دایکه | dayche.com 

Lecture 3 : MongoDB Architecture and Replication (Replica Set)



تولید محتوا: حسن احمدخانی

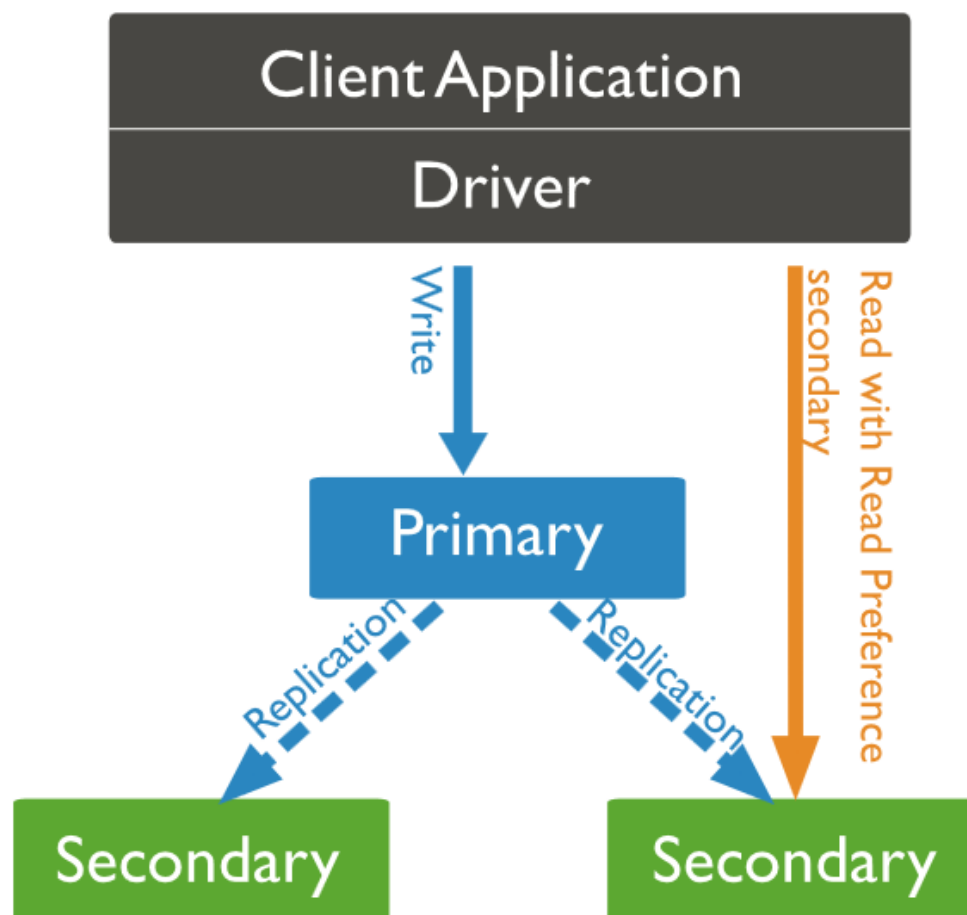
daychegroup 

daychegroup 

dayche.com | گروه دایکه 

Lecture 3 : MongoDB Architecture and Replication (Replica Set)

◆ Read and Write Operations



تولید محتوا: حسن احمدخانی

daychegroup 

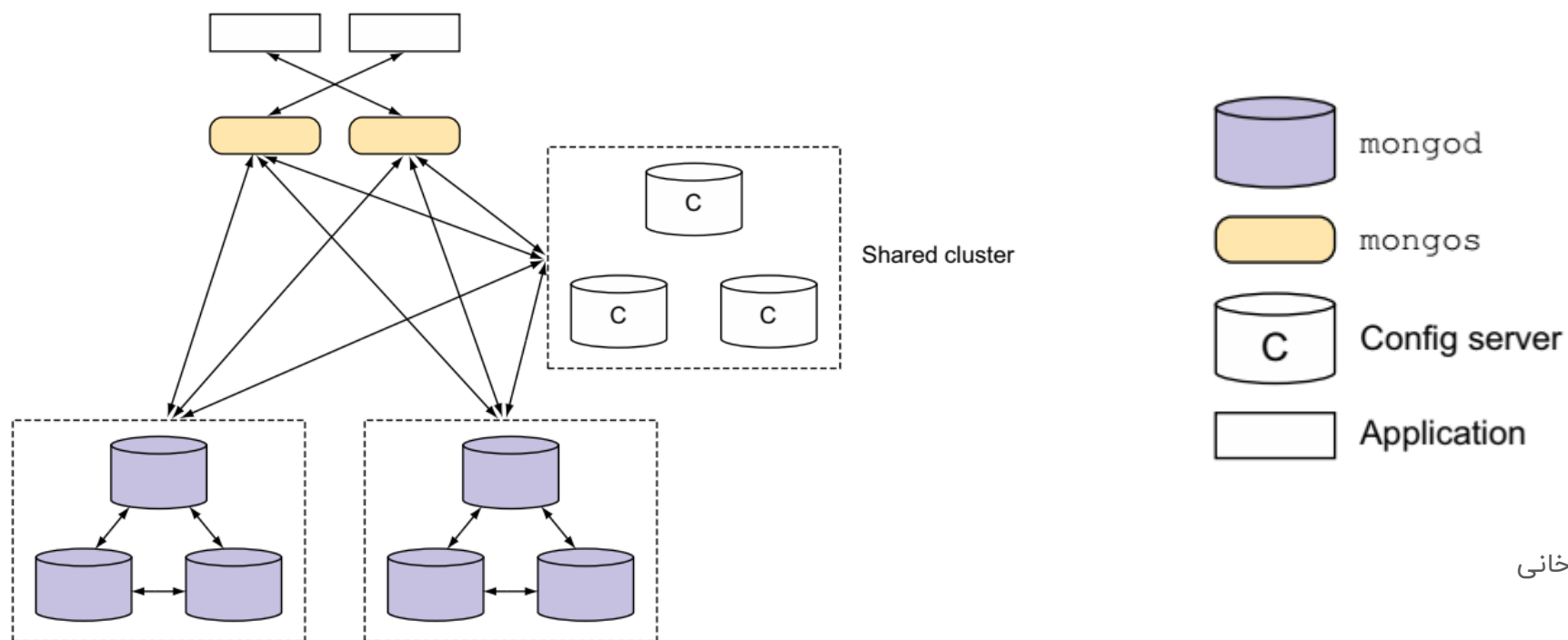
daychegroup 

dayche.com | گروه دایکه 

Cluster Topology

◆ Sharded Cluster

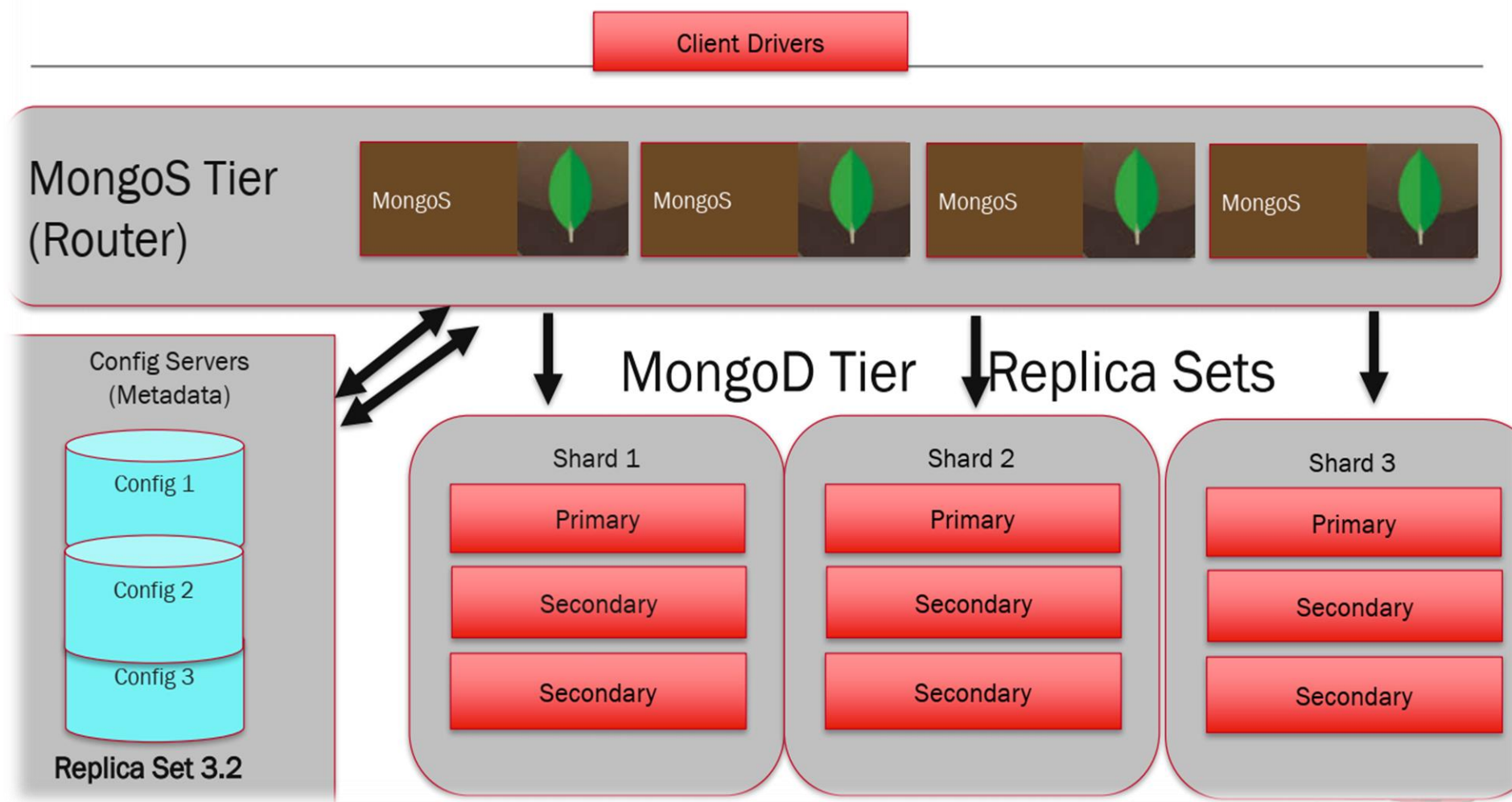
- minimum recommended deployment for a sharded cluster has two shards



Lecture 3 : MongoDB Architecture and Sharding



Sharded Cluster



تولید محتوا: حسن احمدخانی

daychegroup

daychegroup

dayche.com | گروه دایکه



Elasticsearch - Distributed Text Indexing and Search Engine

Relation Databases

- Database
- Table
- Row
- Column
- Schema



Elasticsearch

Index

Type

Document

Fields

Mapping

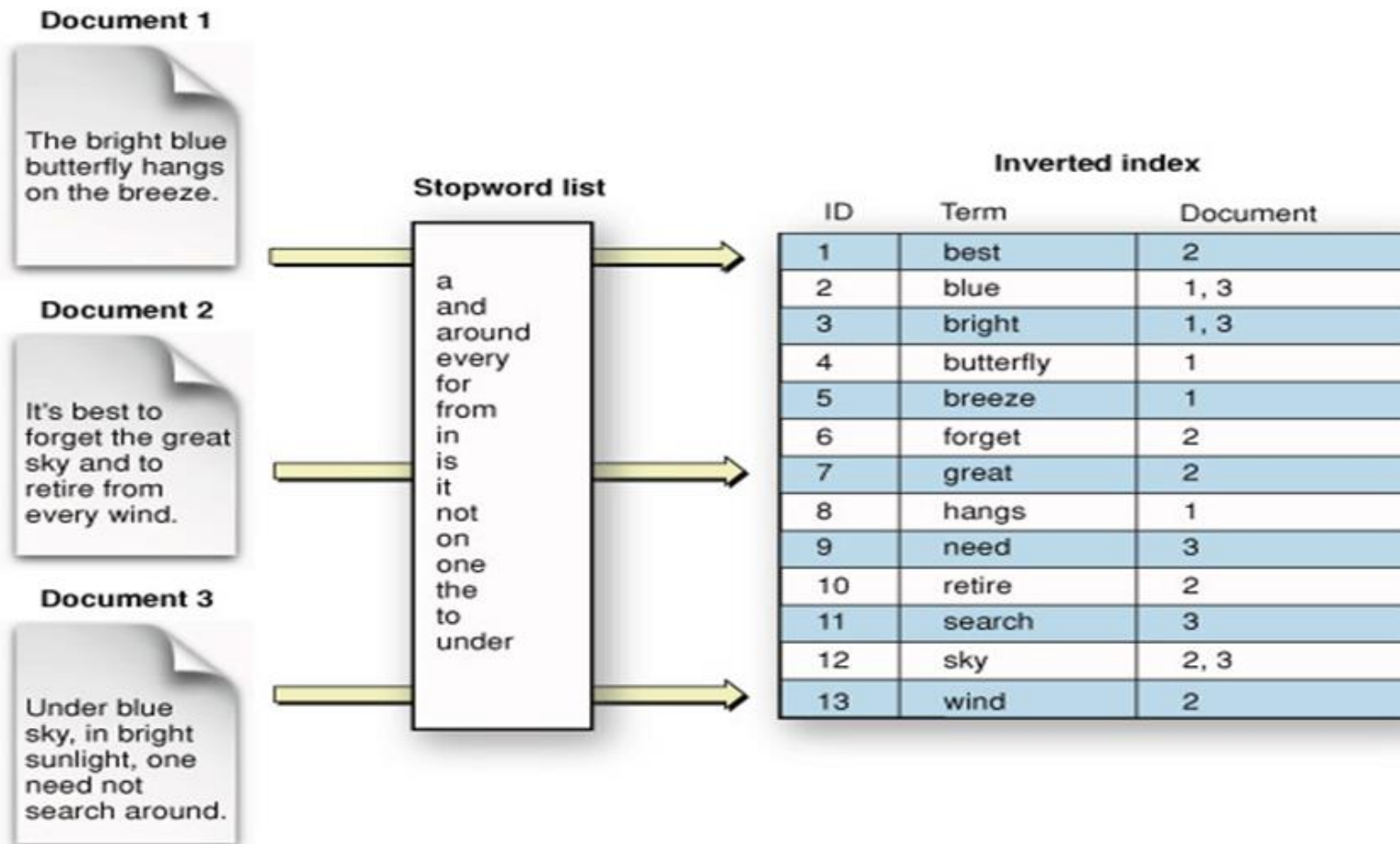
تولید محتوا: حسن احمدخانی

daychegroup

daychegroup

dayche.com | گروه دایکه

Elasticsearch - Inverted Index



تولید محتوا: حسن احمدخانی

daychegroup 

daychegroup 

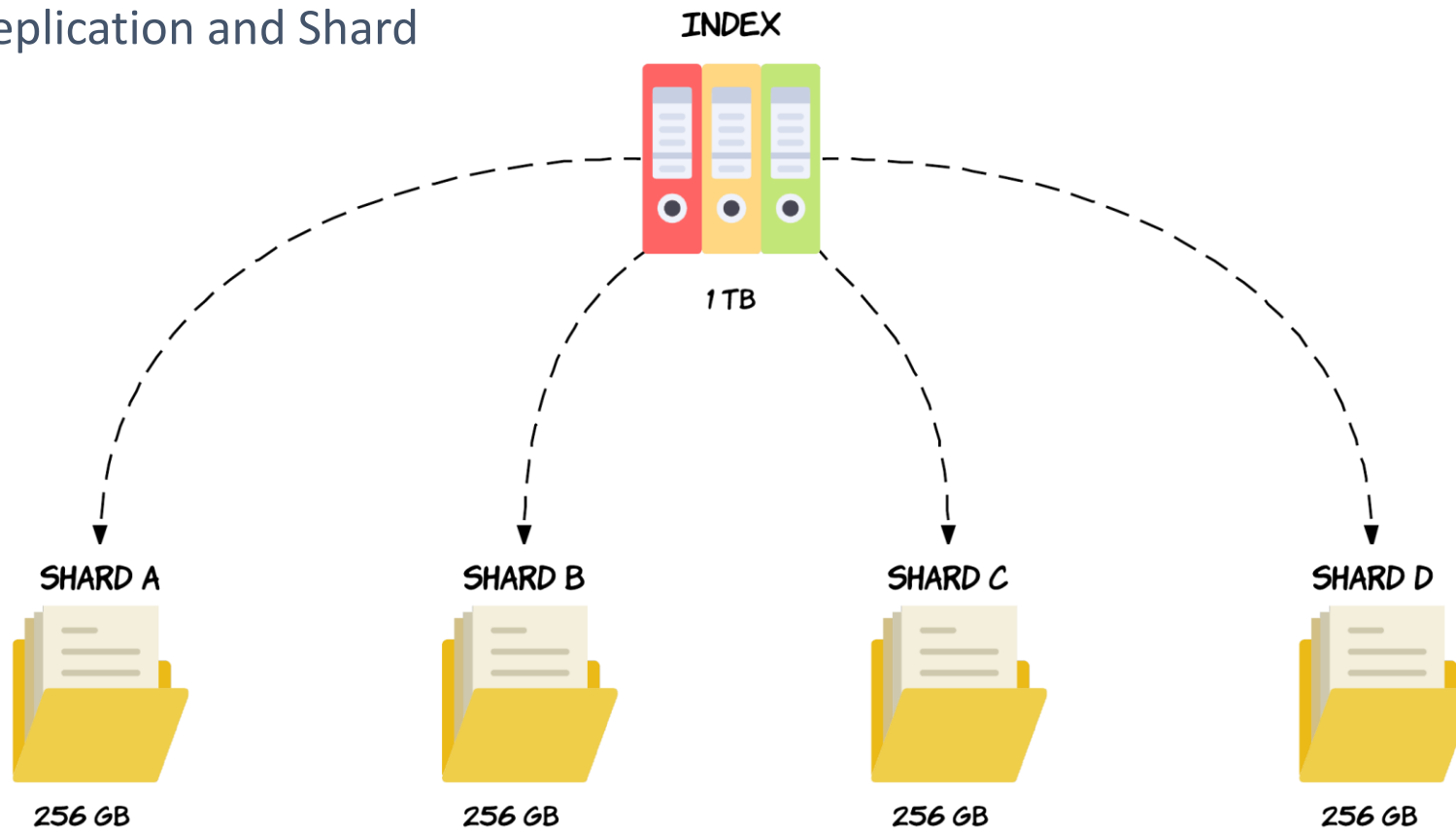
dayche.com | گروه دایکه 

Lecture 3 : Elasticsearch – Sharding and Replication



Elasticsearch

- Replication and Shard



تولید محتوا: حسن احمدخانی

daychegroup

daychegroup

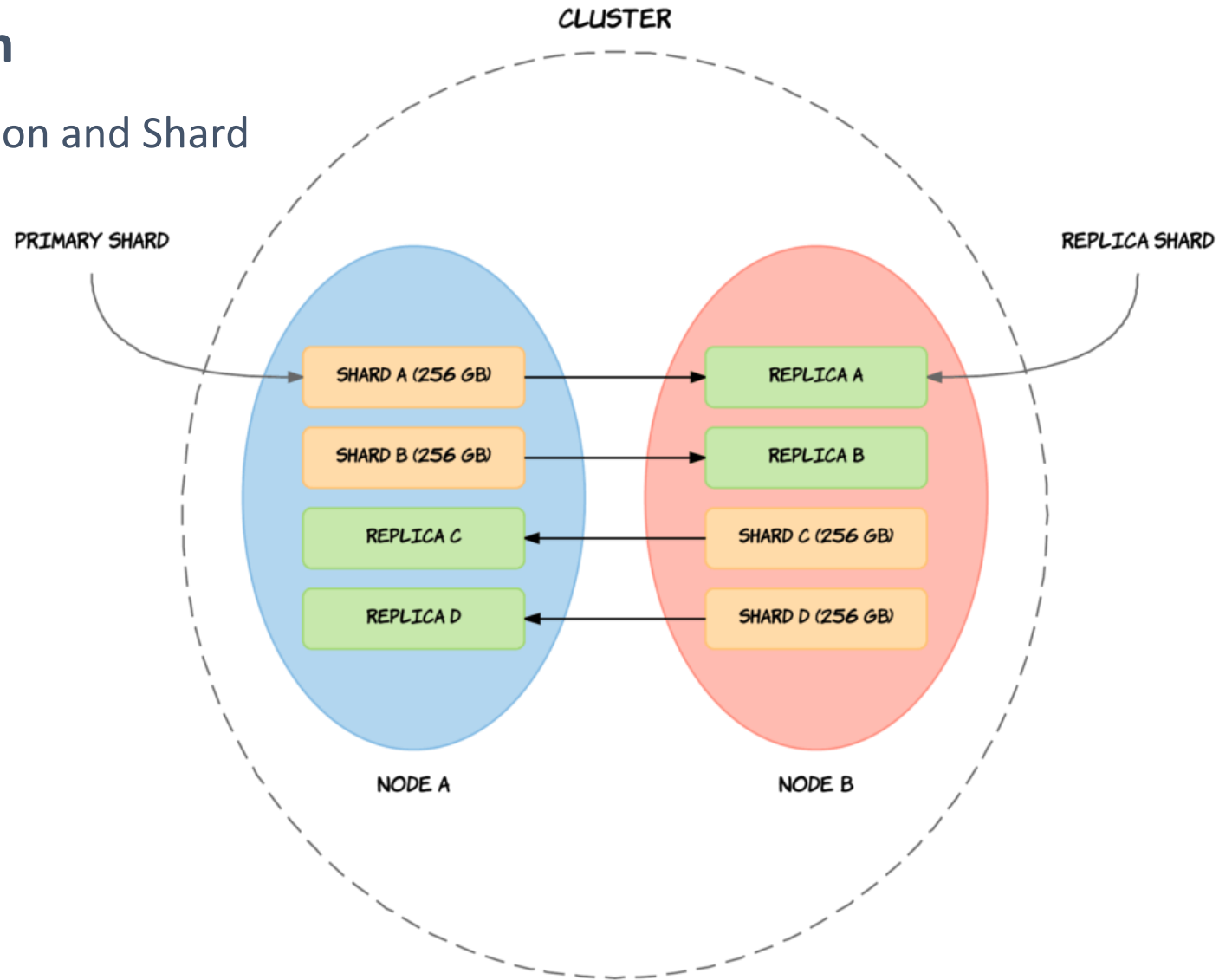
گروه دایکه | dayche.com

Lecture 3 : Elasticsearch – Sharding and Replication



Elasticsearch

- Replication and Shard





Elasticsearch

For Data Scientists

- JDBC
- Tabular return format

```
1 POST /_xpack/sql
2 {
3   "query": "SELECT agency_abbr,
4             COUNT(*) FROM mortgages WHERE
5             state_name = 'California' GROUP BY
6             agency_abbr"
7 }
```

1	agency_abbr	COUNT(1)
2	-----+-----	
3	HUD	1227028
4	CFPB	1690817
5	NCUA	195283
6	FDIC	161711
7	OCC	147797
8	FRS	114972
9		

تولید محتوا: حسن احمدخانی

daychegroup

daychegroup

گروه دایکه | dayche.com