Module 01 : Design Goals, Architecture and Installation

Database Categories

Oracle
MySQL
MS SQL
DB2
Etc.

Netezza
SAP Hanna
Oracle Expre
Etc.

MongoDB
Hbase
Cassandra
CauchDB
Etc.

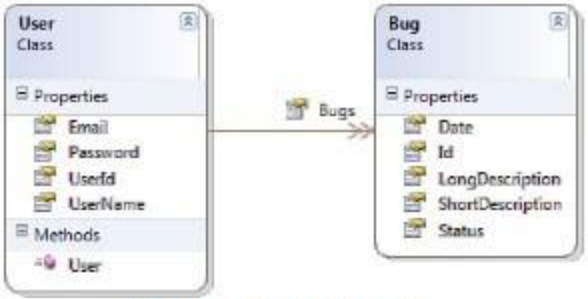
What is NoSQL?

edurek

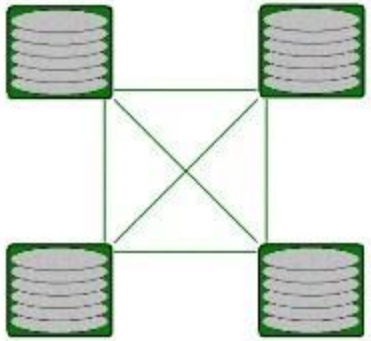
Next Generation Databases

Not Only SQL

Not Only SQL

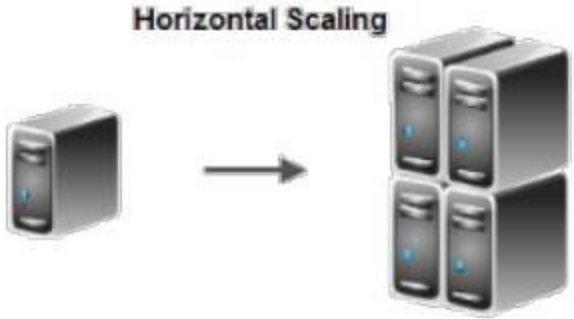


Non – Relational



Distributed Architecture

Open Source



Horizontally Scalable

What is NoSQL?

Schema - Free

Easy – Replication

Simple API

Can Manage Huge Amount of Data

Can be implement on Commodity Hardware's

~ 150 No SQL Database are there in Market

Categories of NoSQL Database

edurek

Document Base

- ✓ Document databases pair each key with a complex data structure known as a document.
- ✓ Documents can contain many different key-value pairs, or key-array pairs, or even nested documents.

Key – value Stores

- ✓ Key-value stores are the simplest NoSQL databases.
- ✓ Every single item in the database is stored as an attribute name (or "key"), together with its value.

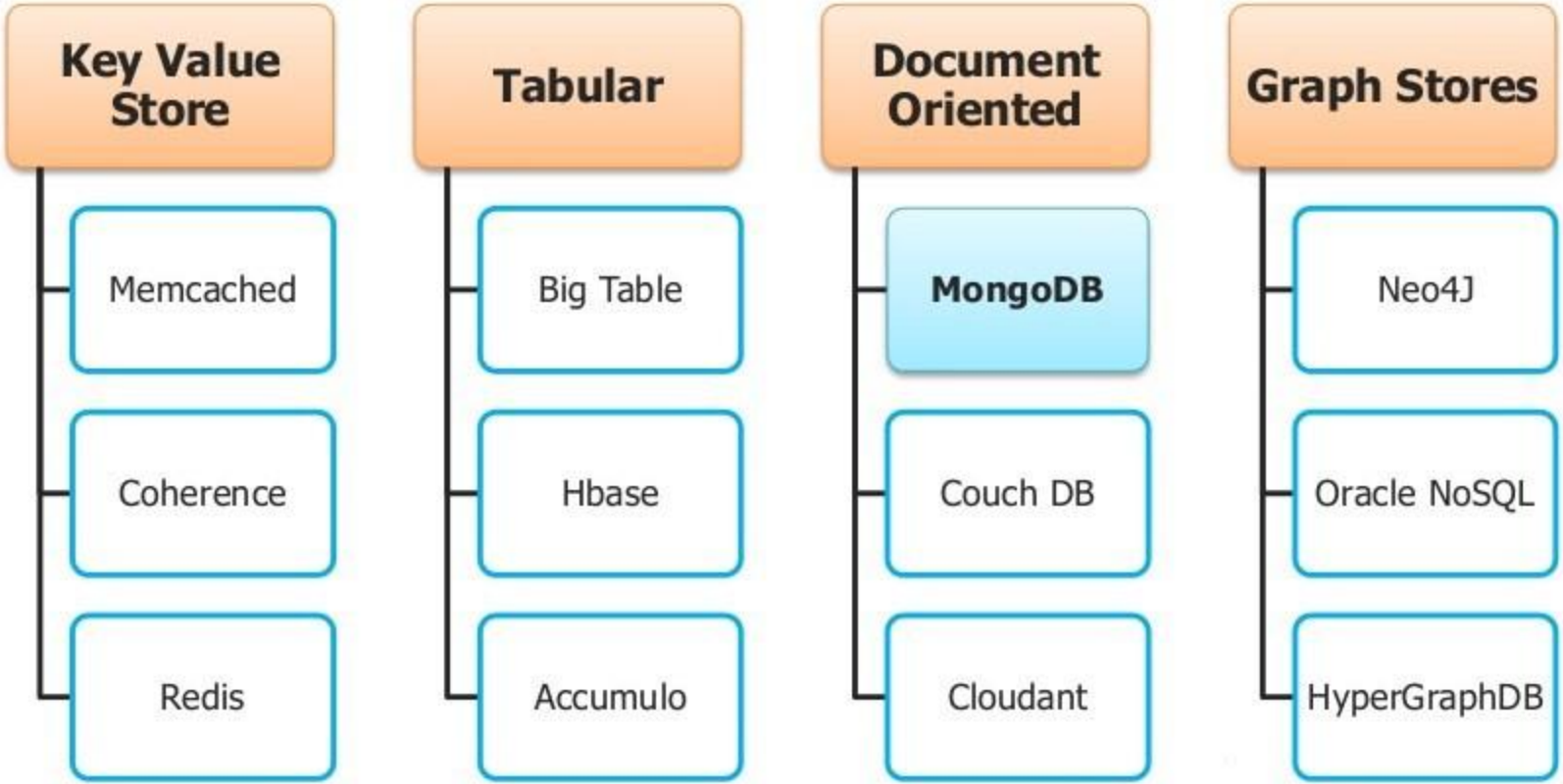
Graph Store

- ✓ Graph stores are used to store information about networks, such as social connections.
- ✓ Graph stores include Neo4J and HyperGraphDB.

Wide Column Stores%

- ✓ Wide-column stores such as Cassandra and HBase are optimized for queries over large datasets, and store columns of data together, instead of rows.

Type of No SQL Databases



SQL Databases (ACID Property)

edurek

Atomic

- ✓ A transaction is a logical unit of work which must be either completed with all of its data modifications, or none of them is performed.

Consistent

- ✓ At the end of the transaction, all data must be left in a consistent state.

ACID Property

Isolated

- ✓ Modifications of data performed by a transaction must be independent of another transaction. Unless this happens, the outcome of a transaction may be erroneous.

Durable

- ✓ When the transaction is completed, effects of the modifications performed by the transaction must be permanent in the system.

Cap Theorem

edurek

CAP theorem states that there are **3 basic requirements** which exist in a special relation when designing applications for a distributed architecture.

Consistency

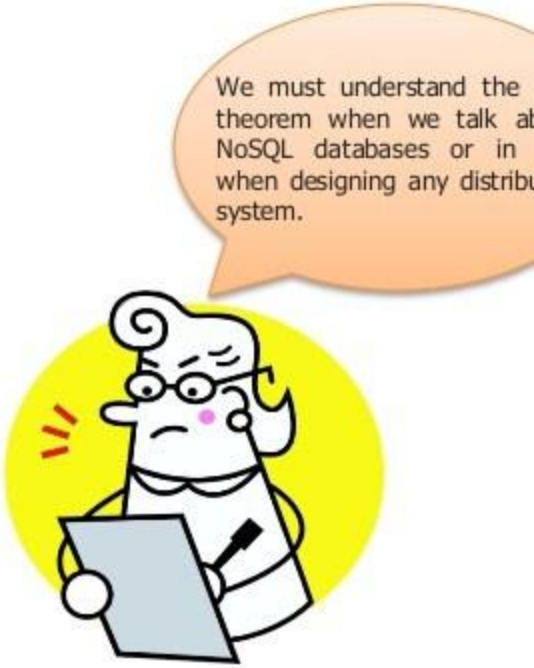
This means that the data in the database remains consistent after the execution of an operation. For example after an update operation all clients see the same data.

Availability

This means that the system is always on (service guarantee availability), no downtime.

Partition Tolerance

This means that the system continues to function even the communication among the servers is unreliable, i.e. the servers may be partitioned into multiple groups that cannot communicate with one another.



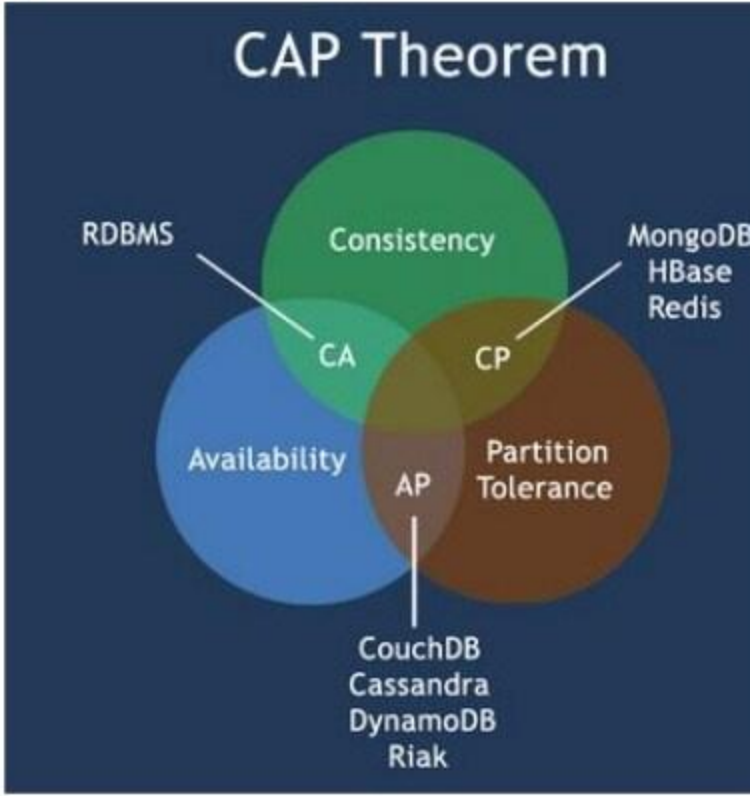
Cap Theorem

- ✓ In theoretically it is **impossible** to fulfill all 3 requirements.
- ✓ CAP provides the basic requirements for a distributed system to follow **2 of the 3 requirements.**
- ✓ Therefore all the current NoSQL database follow the different **combinations of the C, A, P** from the CAP theorem.

Cap Theorem

Here is the brief description of three combinations CA, CP, AP :

- ✓ **CA** - Single site cluster, therefore all nodes are always in contact. When a partition occurs, the system blocks.
- ✓ **CP** - Some data may not be accessible, but the rest is still consistent/accurate.
- ✓ **AP** - System is still available under partitioning, but some of the data returned may be inaccurate.



BASE

A BASE system gives up on consistency.

Basically Available

- ✓ **B**asically **A**vailable indicates that the system **does guarantee** availability, in terms of the CAP theorem.

Soft State

- ✓ **S**oft **S**tate indicates that the state of the system **may change over time**, even without input. This is because of the eventual consistency model.

Eventual Consistency

- ✓ **E**ventual **C**onsistency indicates that the system **will become consistent over time**, given that the system doesn't receive input during that time.

Annie's Question

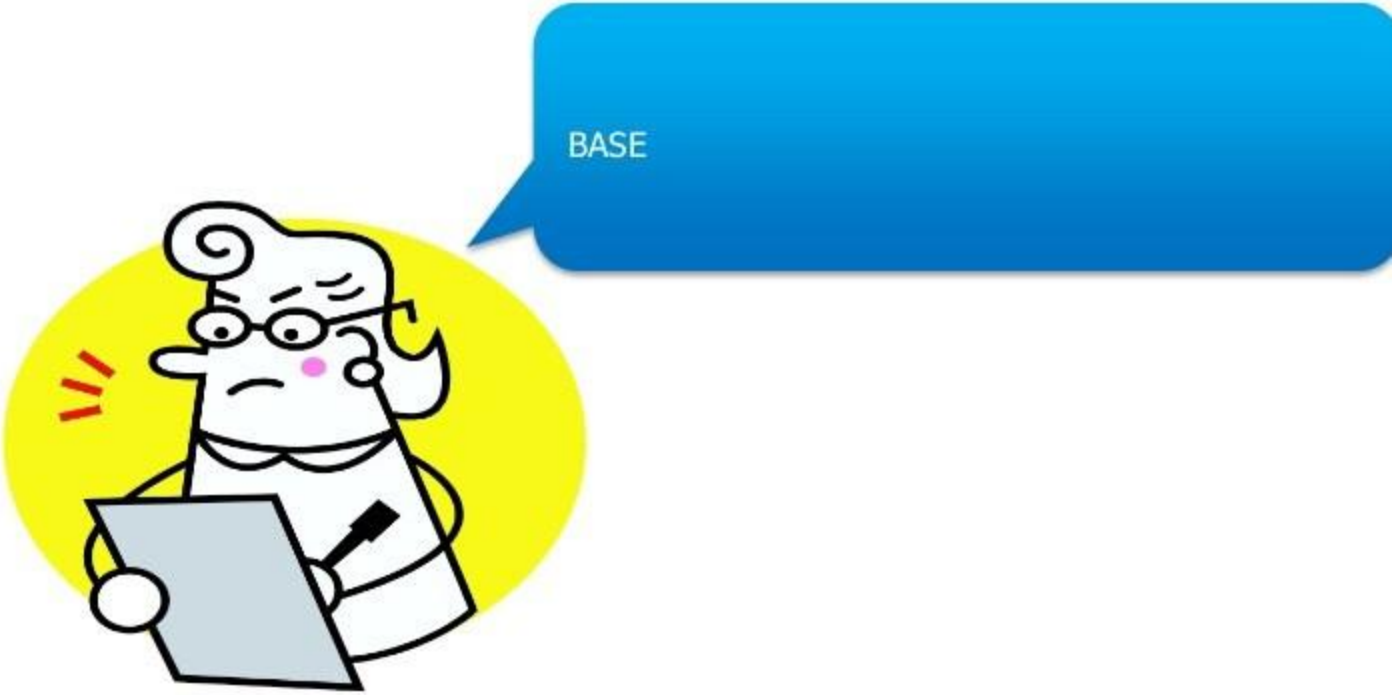
Annie's Answer

MongoDB → Document Oriented Database
Neo4J → Graph Database
Cassandra → Columnar Database
Hbase → Tabular Database

Annie's Question

Which concept is followed by NoSql, chose from below list
1→ ACIS
2→ CAP
3→ BASE

Annie's Answer



MongoDB Overview

MongoDB Overview

edurek

Mongo DB is an Open-source database.

Developed by 10gen, for a wide variety of applications.

It is an agile database that allows schemas to change quickly as applications evolve.

Overview

Scalability, High Performance and Availability.

By leveraging in-memory computing.

MongoDB's native replication and automated failover enable enterprise-grade reliability and operational flexibility.

New World

New Apps

New Development Methods

New Data Volumes

New Architectures

New Data Types

What is MongoDB?

edurek

Open Source

Document Oriented Storage

Object Oriented

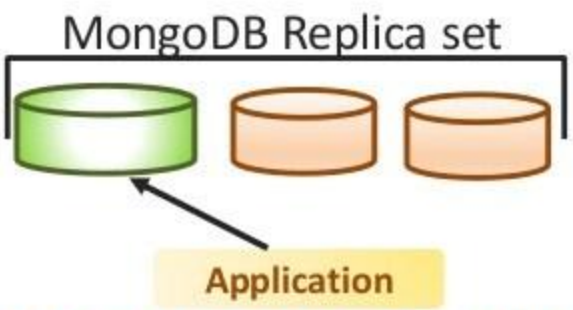
C++

Written in C++

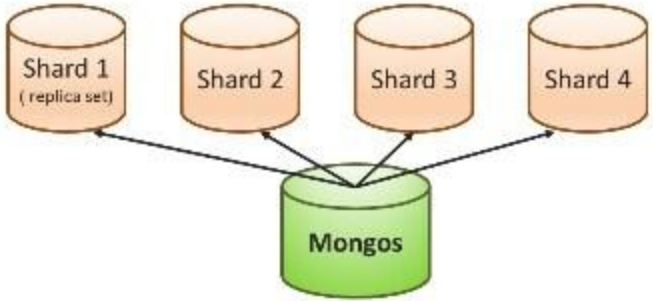
Easy to Use

Full Index Support

What is MongoDB?



Replication and High Availability



Auto Sharding

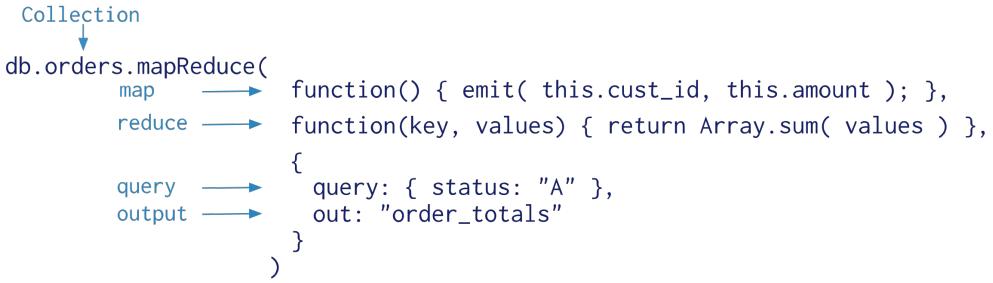
Easy Query

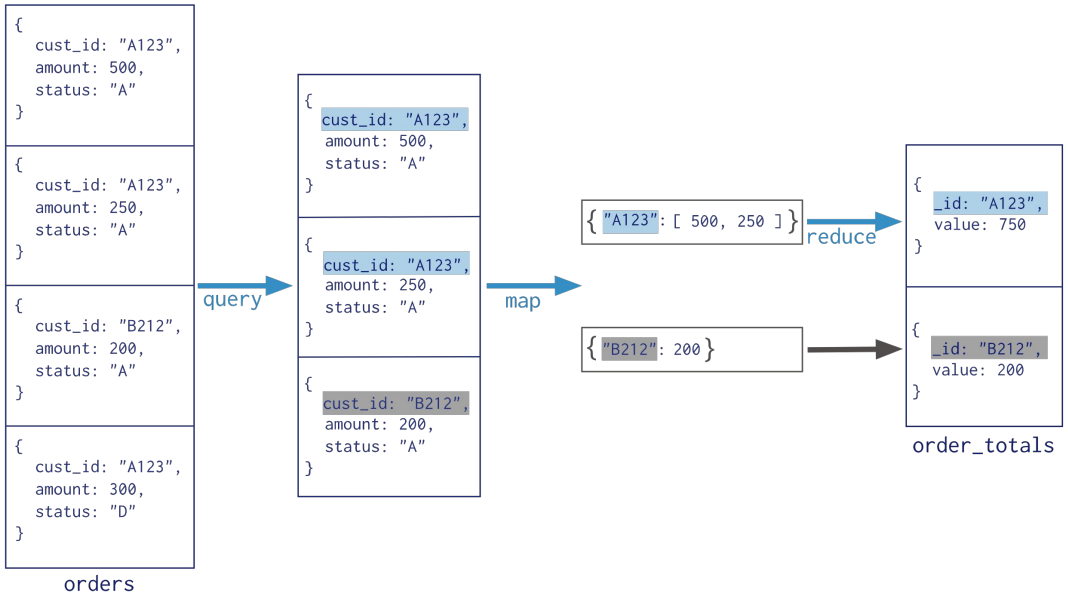
Map Reduce

Gird FS

Support from Expert

MapReduce support





Partition vs Sharding

Original Table

CUSTOMER ID	FIRST NAME	LAST NAME	FAVORITE COLOR
1	TAEKO	OHNUKI	BLUE
2	O.V.	WRIGHT	GREEN
3	SELDA	BAĞCAN	PURPLE
4	JIM	PEPPER	AUBERGINE

Vertical Partitions

VP1

CUSTOMER ID	FIRST NAME	LAST NAME
1	TAEKO	OHNUKI
2	O.V.	WRIGHT
3	SELDA	BAĞCAN
4	JIM	PEPPER

VP2

CUSTOMER ID	FAVORITE COLOR
1	BLUE
2	GREEN
3	PURPLE
4	AUBERGINE

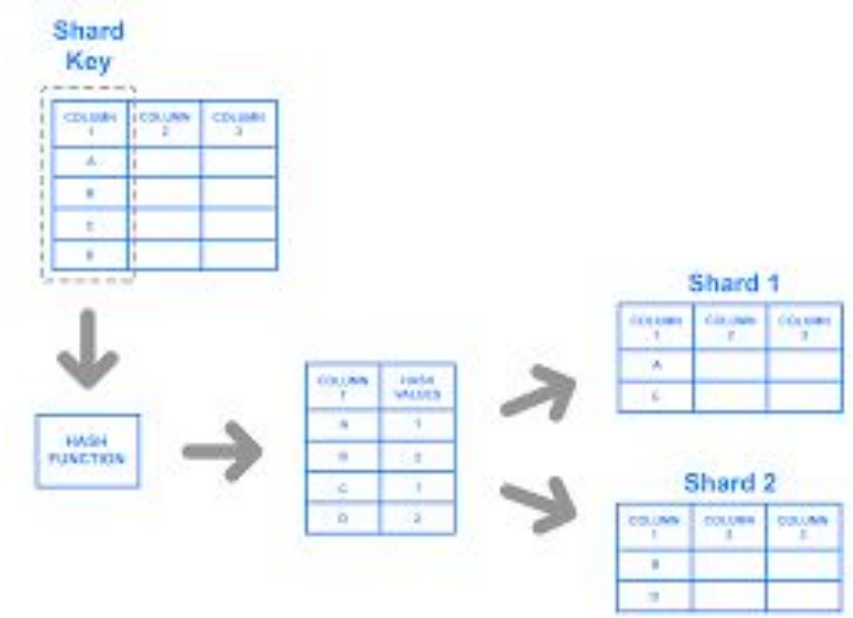
Horizontal Partitions

HP1

CUSTOMER ID	FIRST NAME	LAST NAME	FAVORITE COLOR
1	TAEKO	OHNUKI	BLUE
2	O.V.	WRIGHT	GREEN

HP2

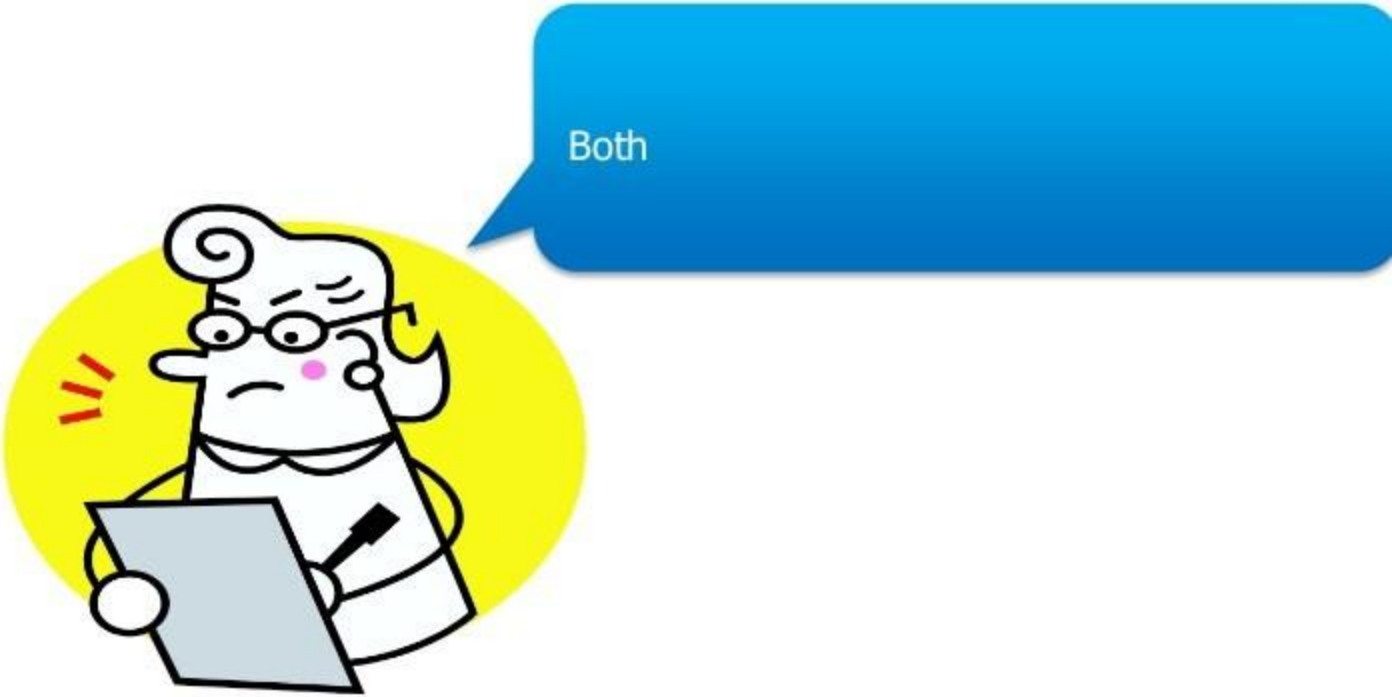
CUSTOMER ID	FIRST NAME	LAST NAME	FAVORITE COLOR
3	SELDA	BAĞCAN	PURPLE
4	JIM	PEPPER	AUBERGINE



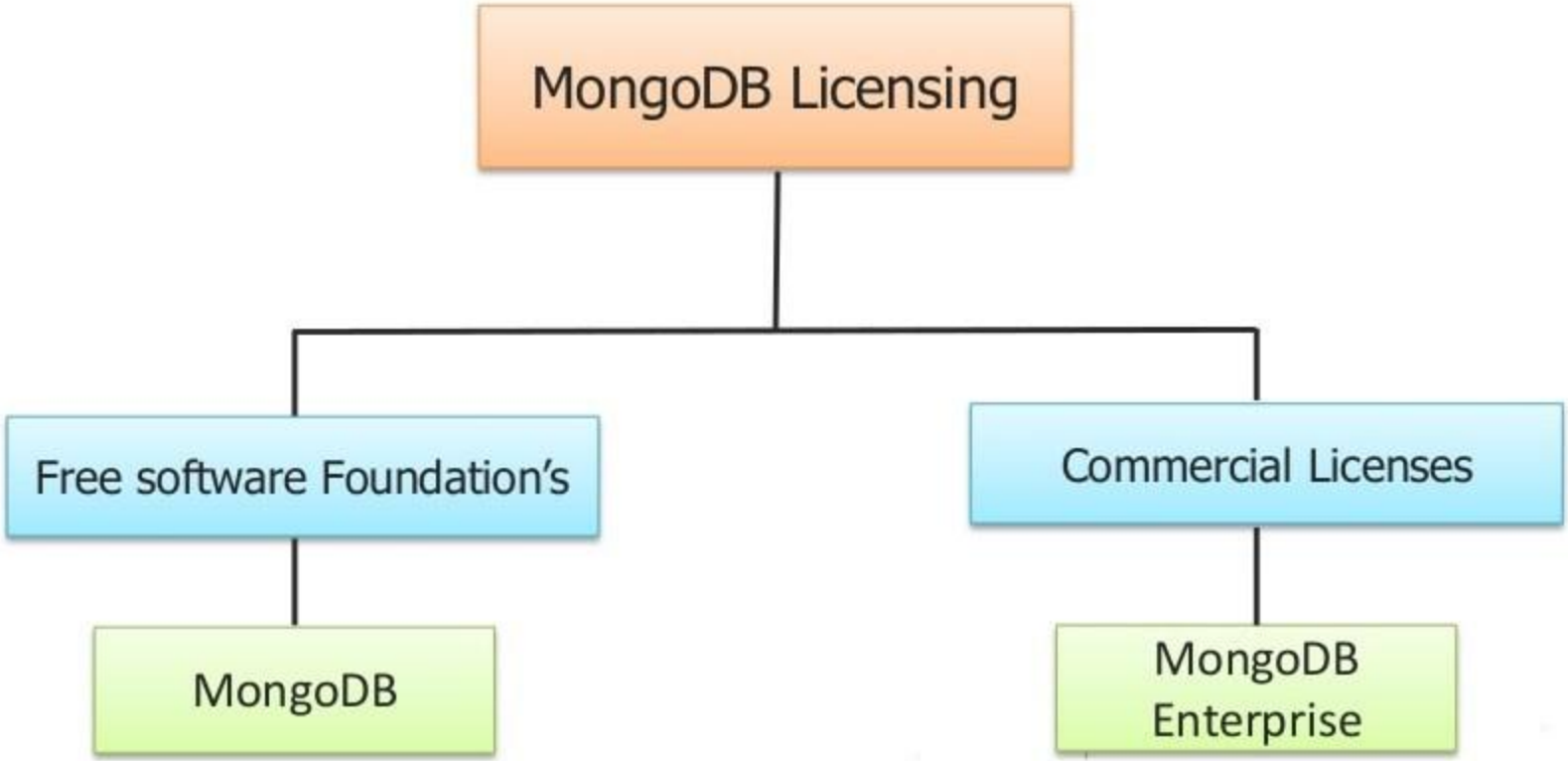
Annie's Question

Which kind of data can be processed with MongoDB, choose from below option
1→ Online Data
2→ Offline Data
3→ Both

Annie's Answer



MongoDB Licensing



Few MongoDB Clients

Few MongoDB Clients

edurek

- ✓ Metlife uses MongoDB for "The Wall" an innovative customer service application provides a 360-degree, consolidated view of MetLife customers, including policy details and transactions across lines of business.

- ✓ ebay has a number of projects running on MongoDB for search suggestions, metadata storage, cloud management and merchandizing categorization.

- ✓ MongoDB is the repository that powers MTV Networks' next-generation CMS, which is used to manage and distribute content for all of MTV Networks' major websites.

- ✓ MongoDB is used for back-end storage on the SourceForge front pages, project pages, and download pages for all projects.

- ✓ Craigslist uses MongoDB to archive billions of records.

- ✓ ADP uses MongoDB for its high performance, scalability, reliability and its ability to preserve the data manipulation capabilities of traditional relational databases.

Few MongoDB Clients

- ✓ CNN Turk uses MongoDB for its infrastructure and content management system, including the tv.cnnturk.com.

- ✓ Foursquare uses MongoDB to store venues and user 'check-ins' into venues, sharding the data over more than 25 machines on Amazon EC2.

- ✓ Justin.tv is the easy, fun, and fast way to share live video online. MongoDB powers Justin.tv's internal analytics tools for virality, user retention, and general usage stats that out-of-the-box solutions can't provide.

- ✓ ibibo ('I build, I bond') is a social network using MongoDB for its dashboard feeds. Each feed is represented as a single document containing an average of 1000 entries; the site currently stores over two million of these documents in MongoDB.

Industry /Domains Where MongoDB is Used

edurek

Government

Financial Services

Healthcare

Media and Entertainment

Tele-communications

Retail

Financial Services

- ✓ Risk Analytics and Reporting
- ✓ Reference Data Management
- ✓ Market Data Management
- ✓ Portfolio Management
- ✓ Order Capture
- ✓ Time Series Data

Government

- ✓ Surveillance Data Aggregation
- ✓ Crime Data Management and Analytics
- ✓ Citizen Engagement Platform
- ✓ Program Data Management
- ✓ Healthcare Record Management

Health Care

- ✓ 360-Degree Patient View
- ✓ Population Management for At-Risk Demographics
- ✓ Lab Data Management and Analytics
- ✓ Mobile Apps for Doctors and Nurses
- ✓ Electronic Healthcare Records (EHR)

Media and Entertainment

edurek

- ✓ Content Management and Delivery
- ✓ User Data Management
- ✓ Digital Asset Management
- ✓ Mobile and Social Apps
- ✓ Content Archiving

Retail

- ✓ Rich Product Catalogs
- ✓ Customer Data Management
- ✓ New Services
- ✓ Digital Coupons
- ✓ Real-Time Price Optimization

Telecommunication

edurek

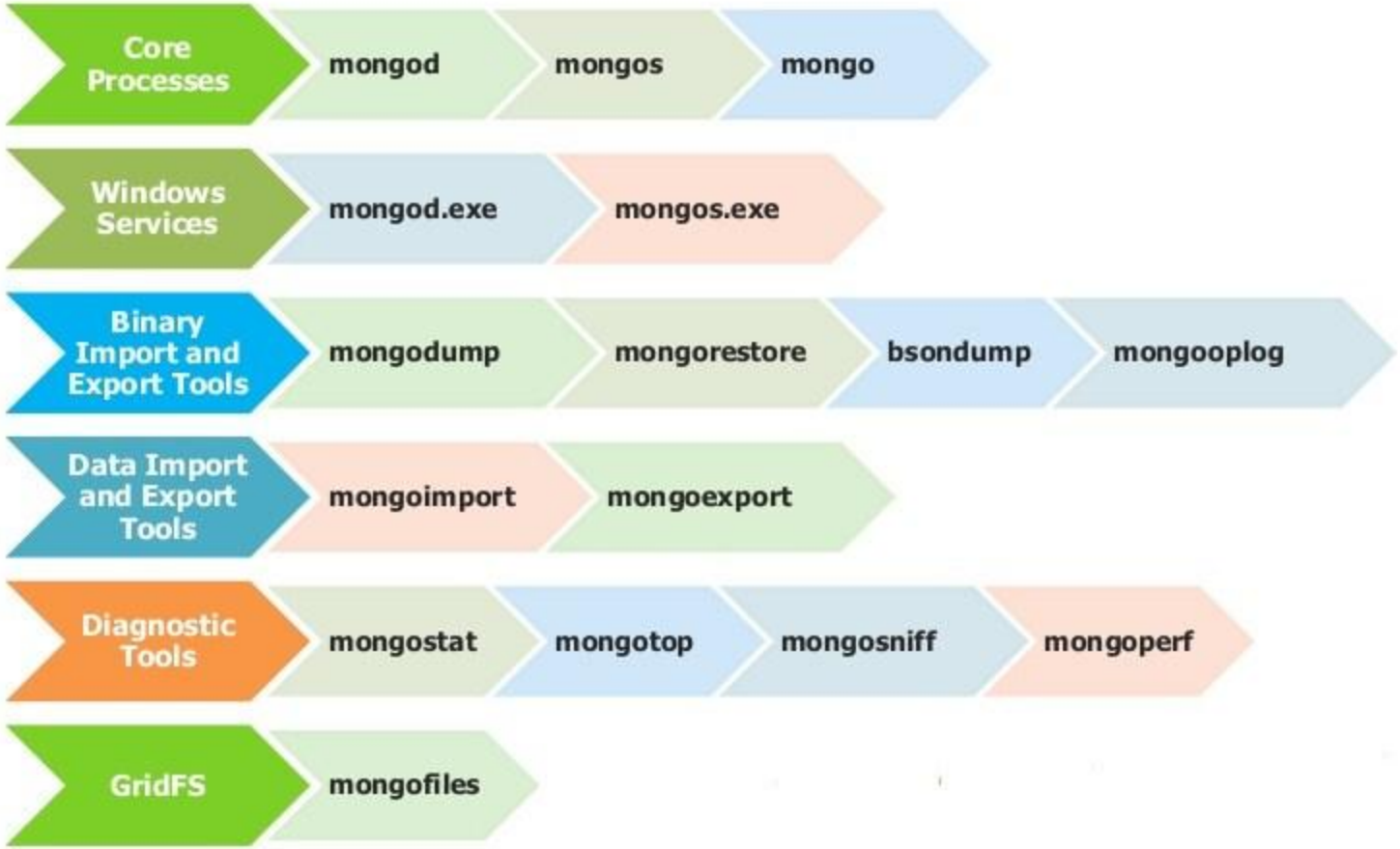
- ✓ Consumer Cloud
- ✓ Product Catalog
- ✓ Customer Service Improvement
- ✓ Machine-to-Machine (M2M) Platform
- ✓ Real-Time Network Analysis and Optimization

MongoDB Tools

MongoDB Package Components

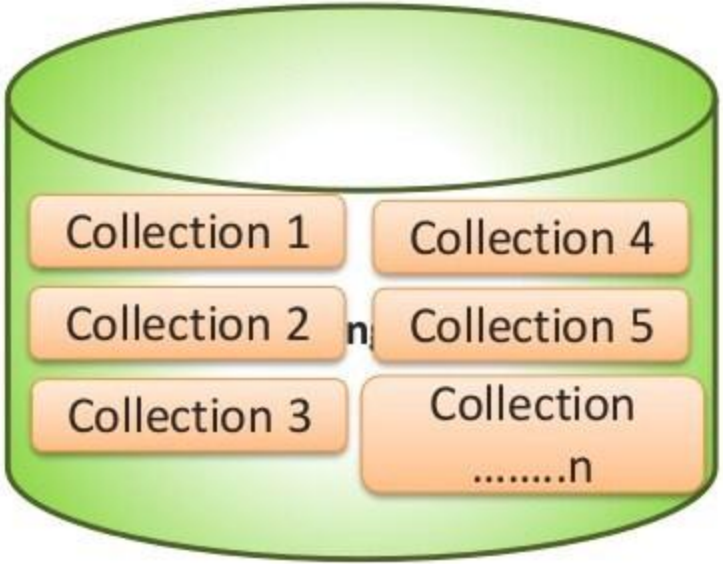
- ✓ mongod
- ✓ mongos
- ✓ mongo
- ✓ mongod.exe
- ✓ mongos.exe
- ✓ mongodump
- ✓ mongorestore
- ✓ bsondump
- ✓ mongooplog
- ✓ mongoimport
- ✓ mongoexport
- ✓ mongostat
- ✓ mongotop
- ✓ mongosniff
- ✓ mongoperf
- ✓ mongofiles

MongoDB Package Components (Tools)



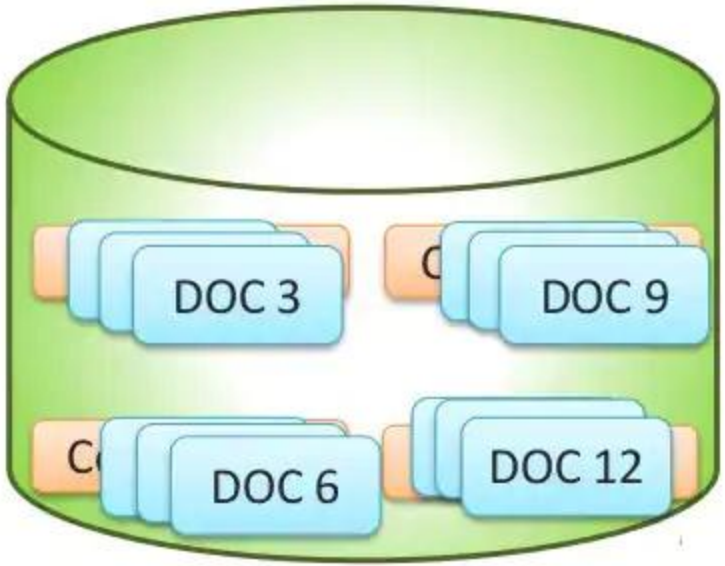
MongoDB Collection

- ✓ Collection is a group of MongoDB documents.
- ✓ It is the equivalent of an RDBMS table.
- ✓ A collection exists within a single database.
- ✓ Collections do not enforce a schema.
- ✓ Documents within a collection can have different fields.
- ✓ Typically, all documents in a collection are of similar or related purpose.



MongoDB Document

- ✓ A document is a set of key-value pairs.
- ✓ Documents have dynamic schema.



MongoDB Sample Document

edurek

```
{
  _id: ObjectId(7df78ad8902c)
  title: 'edureka',
  description: 'Leading Training Provider Across Glob',
  by: 'edureka',
  url: 'http://www.edureka.in',
  tags: ['mongodb', 'database', 'NoSQL'],
  likes: 100,
  comments: [
    {
      user:'user1',
      message: 'My first comment',
      dateCreated: new Date(2011,1,20,2,15),
      like: 0
    },
    {
      user:'user2',
      message: 'My second comments',
      dateCreated: new Date(2011,1,25,7,45),
      like: 5
    }
  ]
}
```

RDBMS Terminology with MongoDB

edurek

RDBMS	MongoDB
Database	Database
Table	Collection
Tuple/Row	Document
Column/Attribute/Variable	Field
Table Join	Embedded Documents
Database Server and Client	
Primary Key	Primary Key (Default key _id provided by mongodb itself)
Mysqld/Oracle	mongod
mysql/sqlplus	mongo

MongoDB Data types

edurek

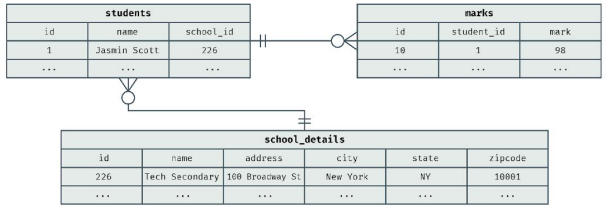
String: Used to store the string data. String in mongodb must be UTF-8 valid.	**Arrays**: This type is used to store arrays or list or multiple values into one key.	**Object**: This datatype is used for embedded documents.	**Code**: To store javascript code into document.	**Time Stamp**: For recording when a document has been modified or added.
Integer: For numerical value. Integer can be 32 bit or 64 bit depending upon your server.	**Boolean**: For a Boolean (true/ false) value.	**Double**: For floating point values.	**Regular expression**: To store regular expression.	**Null**: This type is used to store a Null value.
Date: For current date or time in UNIX time format. Can specify your own date time by creating object of Date and passing day, month, year into it.	**Min/ Max keys**: To compare a value against the lowest and highest BSON elements.	**Symbol**: It's generally reserved for languages that use a specific symbol type.	**Object ID**: Store the document's ID.	**Binary data**: To store binay data.

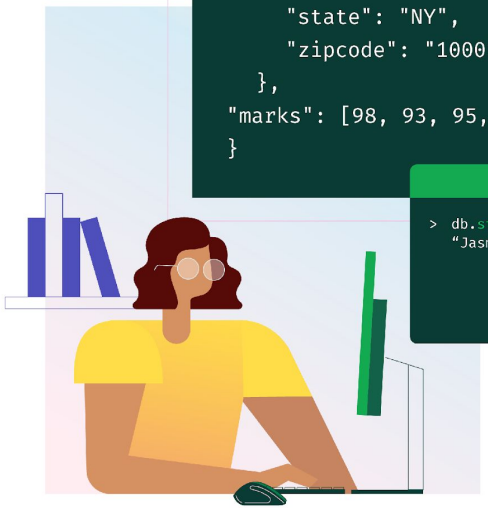
SQL vs MongoDB

MongoDB

```
{
"_id": 1,
"student_name": "Jasmin Scott",
  "school": {
    "school_id": 226,
    "name": "Tech Secondary",
    "address": "100 Broadway St",
    "city": "New York",
    "state": "NY",
    "zipcode": "10001"
  },
"marks": [98, 93, 95, 88, 100],
}
```

SQL





```
mongo
> db.students.find({"student_name":
  "Jasmin Scott"})
```

Results

name	mark	school_name	city
Jasmin Scott	98	Tech Secondary	New York
...	...	...	...

```
sql
SELECT s.name, m.mark, d.name as "school name",
d.city
FROM students s
INNER JOIN marks m ON s.id = m.student_id
INNER JOIN school_details d ON s.school_id = d.id
WHERE s.name = "Jasmin Scott";
```

MongoDB CRUD Syntax & Queries

edurek

- ✓ Insert Documents
- ✓ Query Documents
- ✓ Limit Fields to Return from a Query
- ✓ Iterate a Cursor in the mongo Shell
- ✓ Analyze Query Performance
- ✓ Modify Documents
- ✓ Remove Documents
- ✓ Perform Two Phase Commits
- ✓ Create Tailable Cursor
- ✓ Isolate Sequence of Operations
- ✓ Create an Auto-Incrementing Sequence Field
- ✓ Limit Number of Elements in an Array after an Update

JavaScript Object Notation

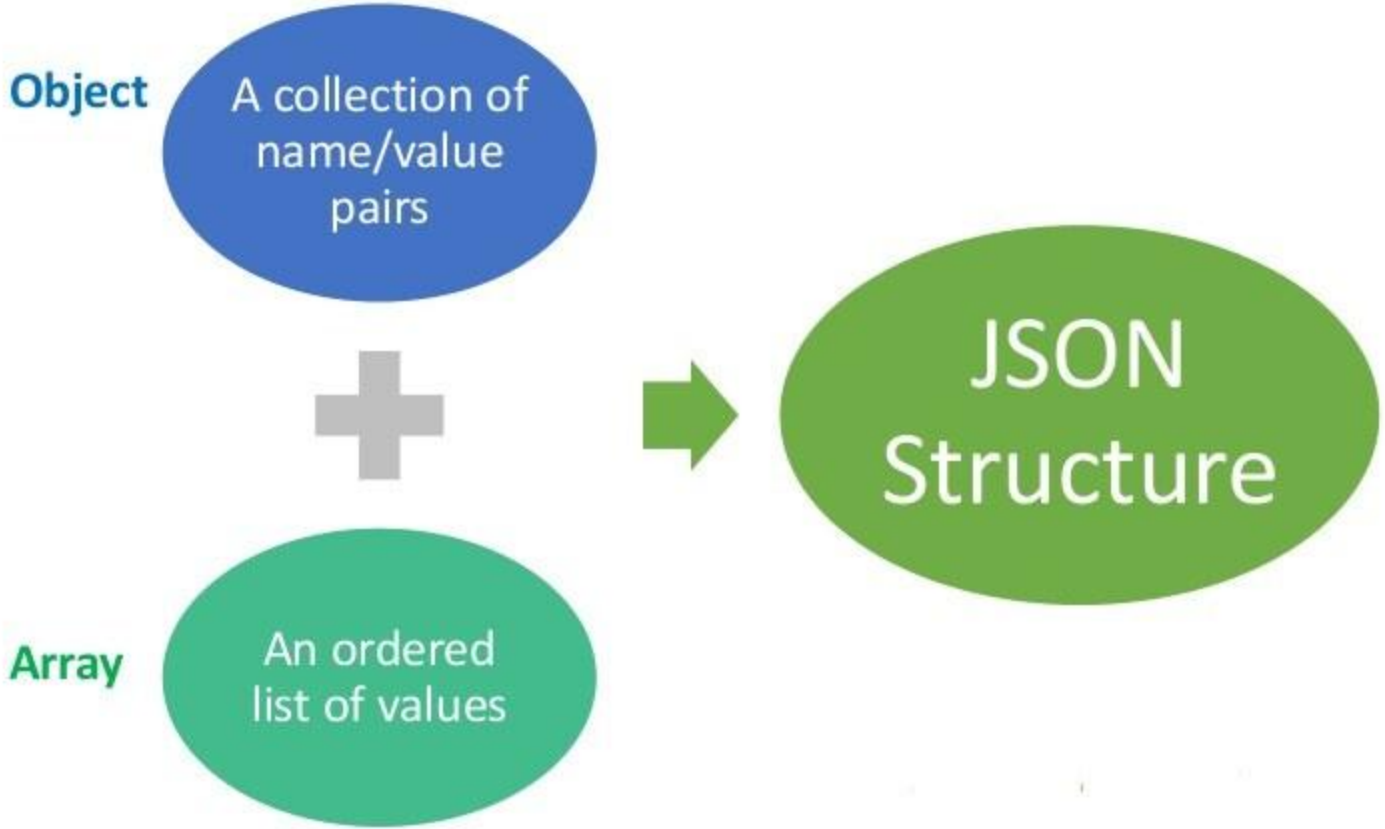
JSON Abbreviation

Lightweight data-interchange format

Easy for humans to read and write

Easy for machines to parse and generate

Text format that is completely language independent


Object
A collection of name/value pairs
Array
An ordered list of values
JSON Structure

BSON

Binary JavaScript Object Notation

BJSON Abbreviation

Supports the embedding of documents and arrays within other documents and arrays

Contains extensions that allo representation of data types t are not part of the JSON spe

Easy for machines to parse and generate

Text format that is completely language independent

JSON and BSON format

{"hello": "world"} →
\x16\x00\x00\x00 // total document size
\x02 // 0x02 = type String
hello\x00 // field name
\x06\x00\x00\x00world\x00 // field value
\x00 // 0x00 = type EOO ('end of object')

{"BSON": ["awesome", 5.05, 1986]} →
\x31\x00\x00\x00
\x04BSON\x00
\x26\x00\x00\x00
\x02\x30\x00\x08\x00\x00\x00awesome\x00
\x01\x31\x00\x33\x33\x33\x33\x33\x33\x14\x40
\x10\x32\x00\xc2\x07\x00\x00
\x00
\x00

	JSON	BSON
Type	JSON files are written in text format.	BSON files are written in binary.
Speed	JSON is fast to read but slower to build.	BSON is slow to read but faster to build and scan.
Space	JSON data is slightly smaller in byte size.	BSON data is slightly larger in byte size.
Encode and Decode	We can send JSON through APIs without encoding and decoding.	BSON files are encoded before storing and decoded before displaying.
Parse	JSON is a human-readable format that doesn't require parsing.	BSON needs to be parsed as they are machine-generated and not human-readable.
Data Types	JSON has a specific set of data types—string, boolean, number for numeric data types, array, object, and null.	Unlike JSON, BSON offers additional data types such as bindata for binary data, decimal128 for numeric.
Usage	Used to send data through the network (mostly through APIs).	Databases use BSON to store data.

Assignments

edurek

Tasks for you

- **Attempt the following Assignments using the documents present in the LMS:**
  - Write a JSON document which can have all data types supported by JSON?
  - What all core differences are there in MongoDB, Hadoop, HBase and Cassandra?
  - How can you define Horizontal & Vertical Scalability?
  - Can we design a Social Media App with MongoDB, if yes then how?
  - To design a content management system what all databases can be used and why?
  - I want to create a solution for Data Hub and I have choice of MySQL, Hadoop, Cassandra, MongoDB, HBase, which one is more suitable and why?
  - What is Online & Offline Big Data?
  - What is Agility, What is tailored and elastic?

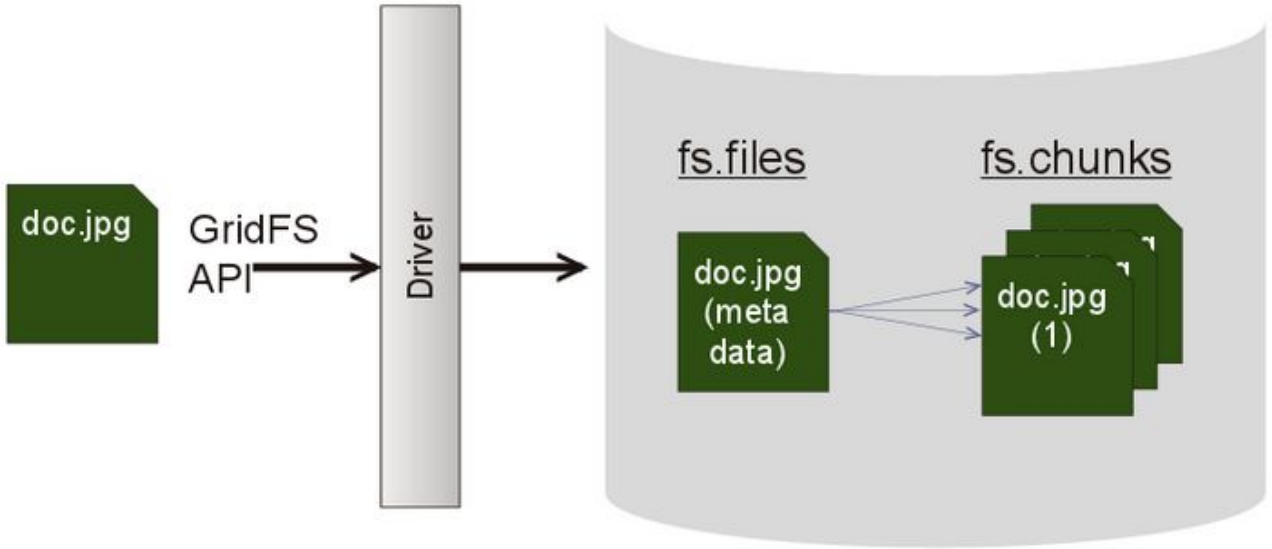
MongoDB Installation – Live Demo

MongoDB Installation – Live Demo

- ✓ Running MongoDB on Windows
- ✓ Installation of MongoDB on Windows as a Service
- ✓ Running of MongoDB on Linux (CentOS)
- ✓ Installation of MongoDB on CentOS

Further Reading

- ✓ **More on MongoDB**
  - http://docs.mongodb.org/manual/
  - http://www.tutorialspoint.com/mongodb/
  - http://en.wikipedia.org/wiki/MongoDB
- ✓ **Connection with PHP to MongoDB**
  - http://www.php.net/manual/en/mongo.tutorial.connecting.php
- ✓ **NoSQL Databases**
  - http://www.w3resource.com/mongodb/nosql.php
  - http://www.dataversity.net/acid-vs-base-the-shifting-ph-of-database-transaction-processing/
- ✓ **Big Data**
  - http://en.wikipedia.org/wiki/Big_Data



Sharding and partitioning are techniques to divide and scale large databases. **Sharding distributes data across multiple servers while partitioning splits tables within one server**.