

ICDS

پایگاه داده‌ها در علم داده

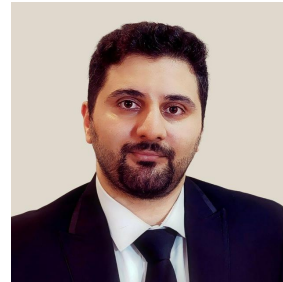
Databases for Data Science

MasoudKaviani.ir

MasoudKaviani.ir

مسعود کاویانی

« دانشمند ارشد داده در صبا ایده (فیلمو، آپارات، سینماتیکت)
« مدرس دانشگاه



ذخیره‌ی داده قدمتی چند هزار ساله دارد



• داده‌ها بر روی لوح گلی ذخیره می‌شد

ذخیره‌ی داده قدمتی چند هزار ساله دارد

• و یا پوست و استخوان حیوانات



ذخیره‌ی داده قدمتی چند هزار ساله دارد

• و با پیشرفت علم دیسک‌های سخت و SSDها اختراع شدند



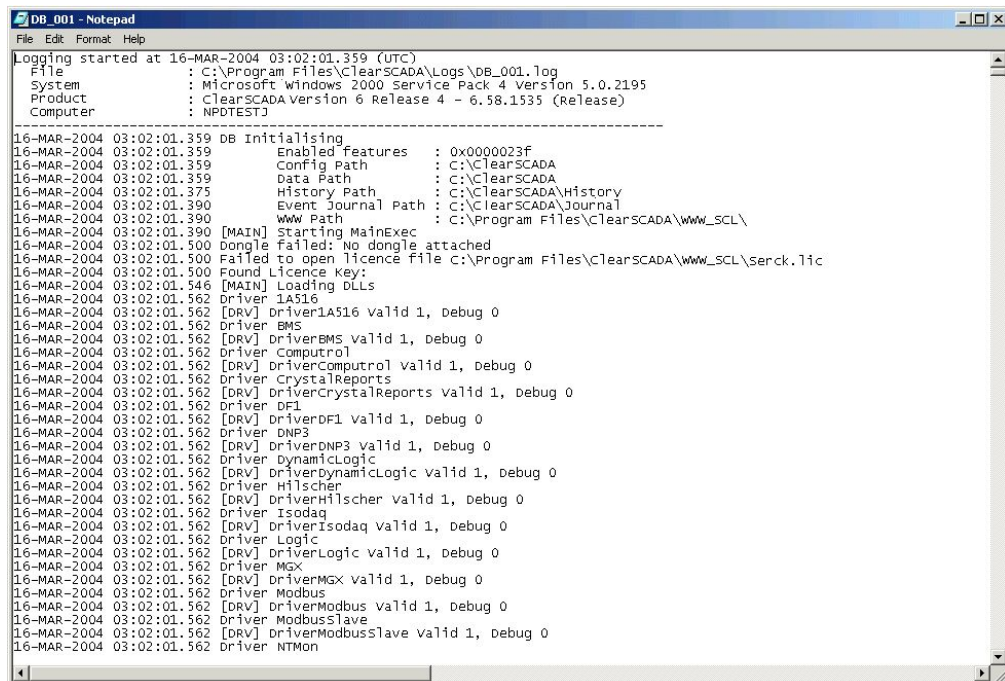
HDD



SSD

ذخیره داده‌ها در فایل

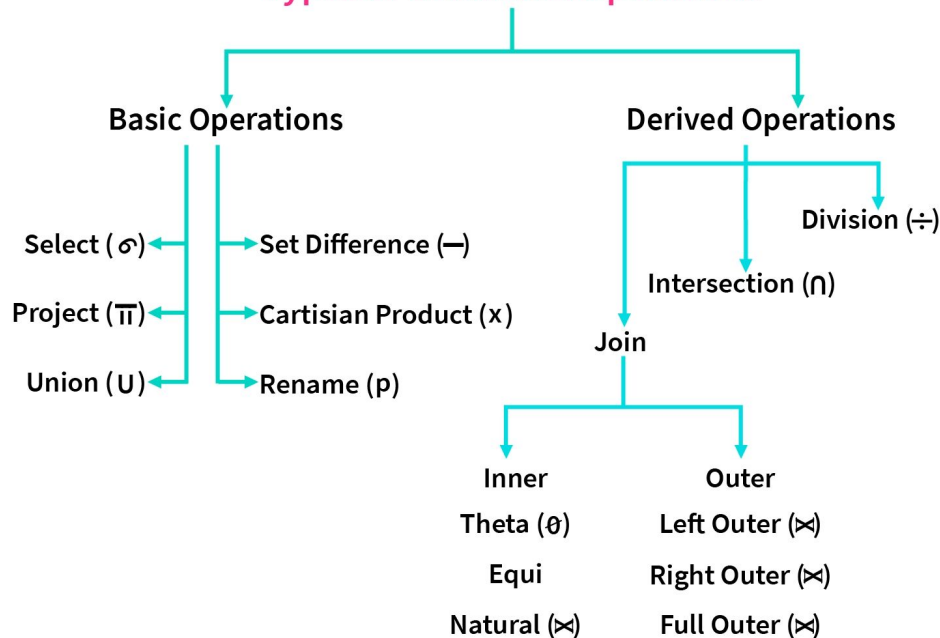
- ساده‌ترین نوع ذخیره‌سازی داده‌ها
- مزیت: سرعت ذخیره‌ی (write) بالا
- عیب: سرعت جستجوی (search) پایین



```
DB_001 - Notepad
File Edit Format Help
Logging started at 16-MAR-2004 03:02:01.359 (UTC)
File      : C:\Program Files\CleasSCADA\Logs\DB_001.log
System    : Microsoft Windows 2000 Service Pack 4 Version 5.0.2195
Product   : CleasSCADA Version 6 Release 4 - 6.58.1535 (Release)
Computer  : NPDTEST
-----
16-MAR-2004 03:02:01.359 DB initialising
16-MAR-2004 03:02:01.359   Enabled Features   : 0x0000023f
16-MAR-2004 03:02:01.359   Config Path      : C:\CleasSCADA
16-MAR-2004 03:02:01.359   Data Path       : C:\CleasSCADA
16-MAR-2004 03:02:01.375   History Path    : C:\CleasSCADA\History
16-MAR-2004 03:02:01.390   Event Journal Path : C:\CleasSCADA\Journal
16-MAR-2004 03:02:01.390   Wwww Path      : C:\Program Files\CleasSCADA\www_SCL\
16-MAR-2004 03:02:01.390 [MAIN] Starting MainExec
16-MAR-2004 03:02:01.500 Dongle failed: No dongle attached
16-MAR-2004 03:02:01.500 Failed to open licence file c:\Program Files\CleasSCADA\www_SCL\serck.lic
16-MAR-2004 03:02:01.500 Found Licence key:
16-MAR-2004 03:02:01.546 [MAIN] Loading DLLs
16-MAR-2004 03:02:01.562 Driver IA516
16-MAR-2004 03:02:01.562 [DRV] DriverIA516 valid 1, Debug 0
16-MAR-2004 03:02:01.562 [DRV] DriverBMS valid 1, Debug 0
16-MAR-2004 03:02:01.562 [DRV] DriverBMS valid 1, Debug 0
16-MAR-2004 03:02:01.562 [DRV] DriverComputrol
16-MAR-2004 03:02:01.562 [DRV] DriverComputrol valid 1, Debug 0
16-MAR-2004 03:02:01.562 [DRV] DriverCrystalReports
16-MAR-2004 03:02:01.562 [DRV] DriverCrystalReports valid 1, Debug 0
16-MAR-2004 03:02:01.562 [DRV] DriverDF1
16-MAR-2004 03:02:01.562 [DRV] DriverDF1 valid 1, Debug 0
16-MAR-2004 03:02:01.562 [DRV] DriverDNP3
16-MAR-2004 03:02:01.562 [DRV] DriverDNP3 valid 1, Debug 0
16-MAR-2004 03:02:01.562 [DRV] DriverDynamicLogic
16-MAR-2004 03:02:01.562 [DRV] DriverDynamicLogic valid 1, Debug 0
16-MAR-2004 03:02:01.562 [DRV] DriverHilscher
16-MAR-2004 03:02:01.562 [DRV] DriverHilscher valid 1, Debug 0
16-MAR-2004 03:02:01.562 [DRV] DriverIsodaq
16-MAR-2004 03:02:01.562 [DRV] DriverIsodaq valid 1, Debug 0
16-MAR-2004 03:02:01.562 [DRV] DriverLogic
16-MAR-2004 03:02:01.562 [DRV] DriverLogic valid 1, Debug 0
16-MAR-2004 03:02:01.562 [DRV] DriverMGX
16-MAR-2004 03:02:01.562 [DRV] DriverMGX valid 1, Debug 0
16-MAR-2004 03:02:01.562 [DRV] DriverModbus
16-MAR-2004 03:02:01.562 [DRV] DriverModbus valid 1, Debug 0
16-MAR-2004 03:02:01.562 [DRV] DriverModbusSlave
16-MAR-2004 03:02:01.562 [DRV] DriverModbusSlave valid 1, Debug 0
16-MAR-2004 03:02:01.562 [DRV] DriverNTMon
```

جبر رابطه‌ای (Relational Algebra)

Types of Relational Operations



- جبر رابطه‌ای توانایی مدل‌سازی اشیا و عناصر دنیای واقعی را دارد
- از جبر رابطه‌ای برای ساخت خانواده‌ی پایگاه داده‌های مبتنی بر SQL استفاده می‌شود

پایگاه داده مبتنی بر SQL

- مثال سیستم اسپاتیفای!

Music

Artists		
ArtistId	ArtistName	Desc
1	AC/DC	One of t...
2	U2	Another...
3	Nelly	When y...
4	Lorde	From N...

Albums		
AlbumId	AlbumName	ArtistId
1	Nellyville	3
2	Black Ice	1
3	Ballbreaker	1
4	October	2

Ratings		
RatingId	AlbumId	Rating
1	2	5
2	1	3.5
3	4	3
4	3	4

مزایا و معایب پایگاه داده‌های رابطه‌ای

مزایا:

- قدیمی‌تر هستند و باگ‌های بیشتری را برطرف کرده‌اند
- استفاده‌کننده‌های بیشتری دارند
- مبتنی بر SQL هستند
- ACID را پشتیبانی می‌کنند

معایب:

- شمای ثابت دارند و تغییر در شما در برخی از مواقع دشوار است
- پیچیدگی محاسباتی به دلیل محدودیت‌های اعمال شده (مانند نرمال‌سازی و رابطه‌ها)
- به راحتی Scale Out نمی‌شوند

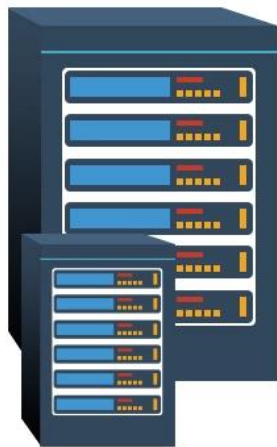
بيگ ديتا

- حجم، سرعت و تنوع بالا



افزایش حجم و سرعت تولید داده‌ها

- افزایش حجم، سرعت تولید و تنوع داده‌ها نیاز به افزایش ظرفیت ظرفیت و سرعت دیسک داشت
- از طرفی حجم و سرعت سخت افزار (دیسک) با سرعت داده‌ها افزایش نیافت

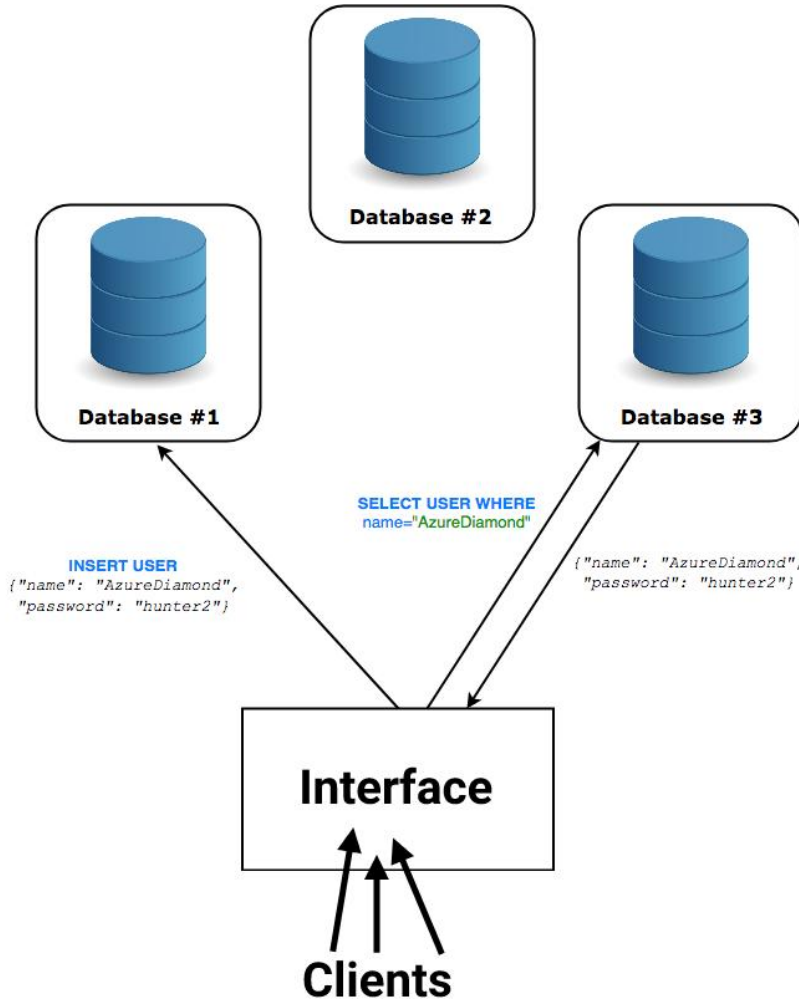


Vertical Scaling
(Scaling up)



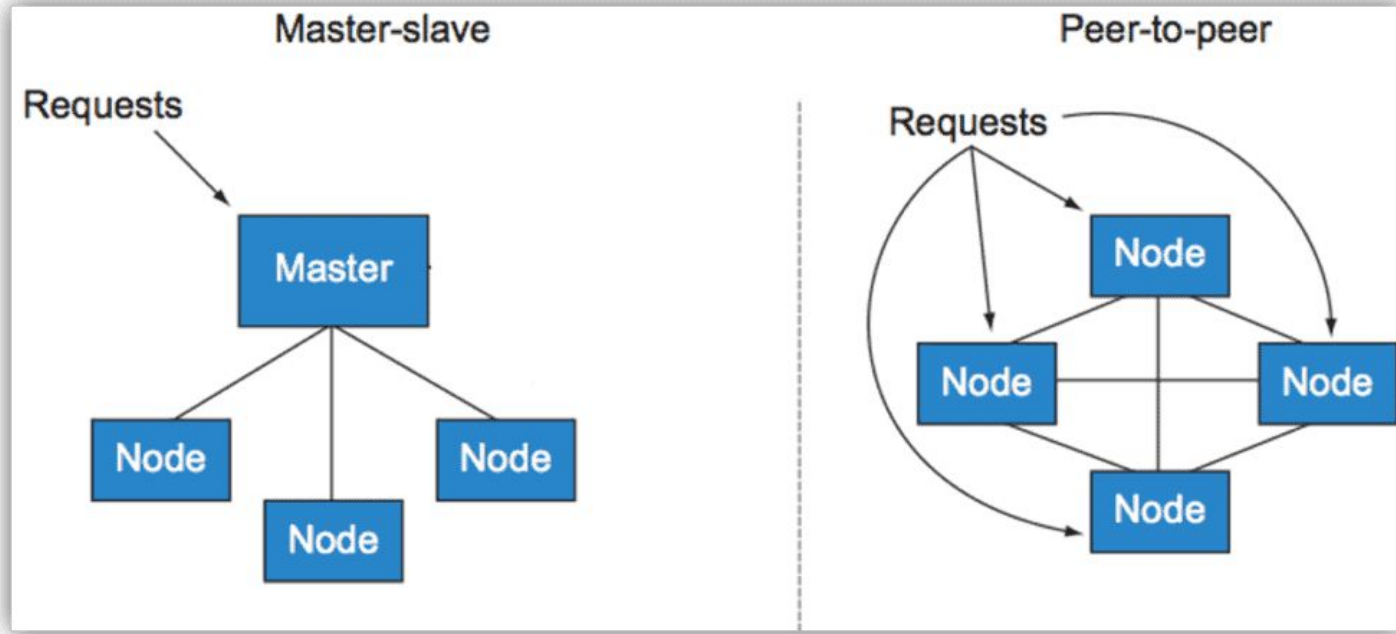
Horizontal Scaling
(Scaling out)

سیستم توزیع شده

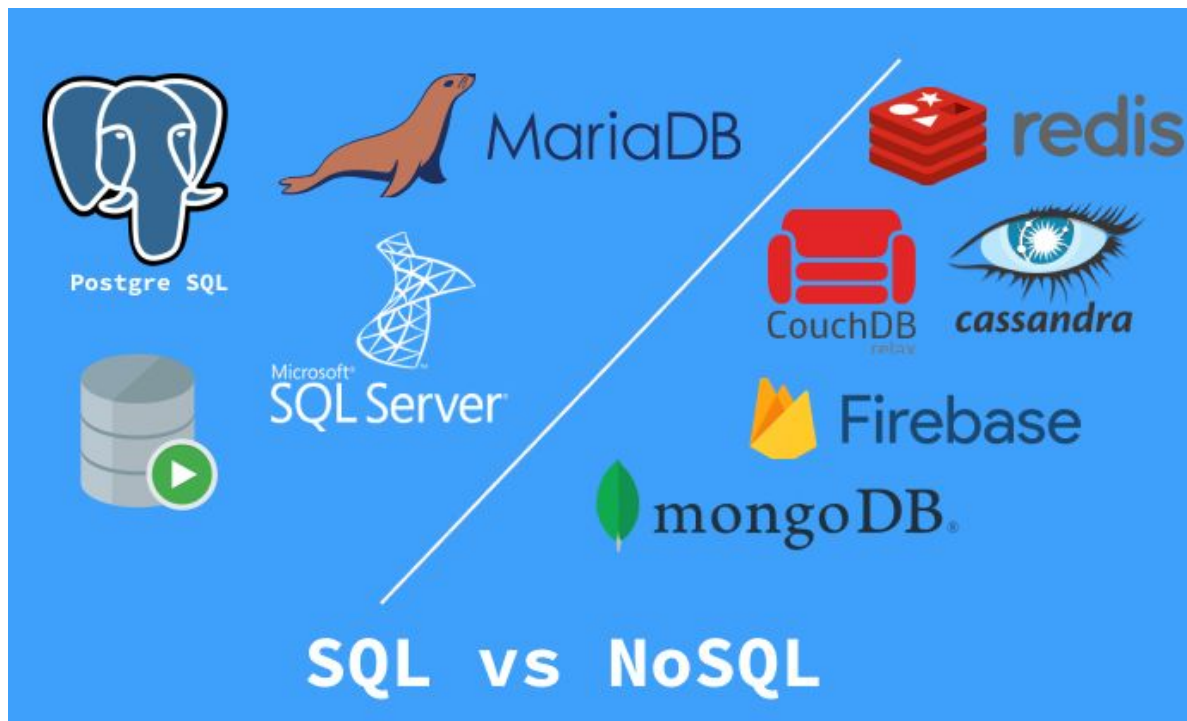


- مجموعه‌ای از کامپیوترها که با تبادل اطلاعات و ارتباطات با یکدیگر، شبیه به یک سیستم منسجم می‌شوند
- برای scale out کردن نیاز به سیستم توزیع شده است

سیستم‌های توزیع شده Master-Slave VS Masterless



خانواده‌ی SQL در مقابل NoSQL ها



خاصیت ACID در مقابل BASE

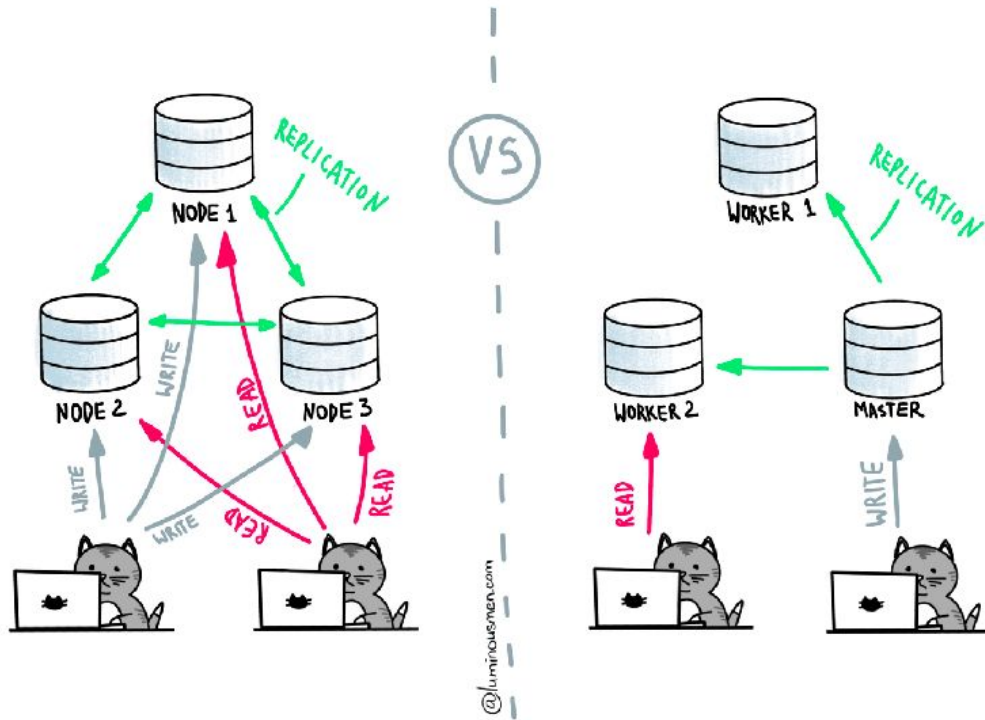
The word "ACID" is written in a colorful, hand-drawn style. Each letter has a different color: 'A' is pink, 'C' is yellow, 'I' is blue, and 'D' is green. The letters are set against a light blue, cloud-like background.

- STRONG CONSISTENCY
- ISOLATION
- TRANSACTIONS
- SCALE-UP (LIMITED)
- PRECISE ANSWERS
- CONSISTENCY FIRST

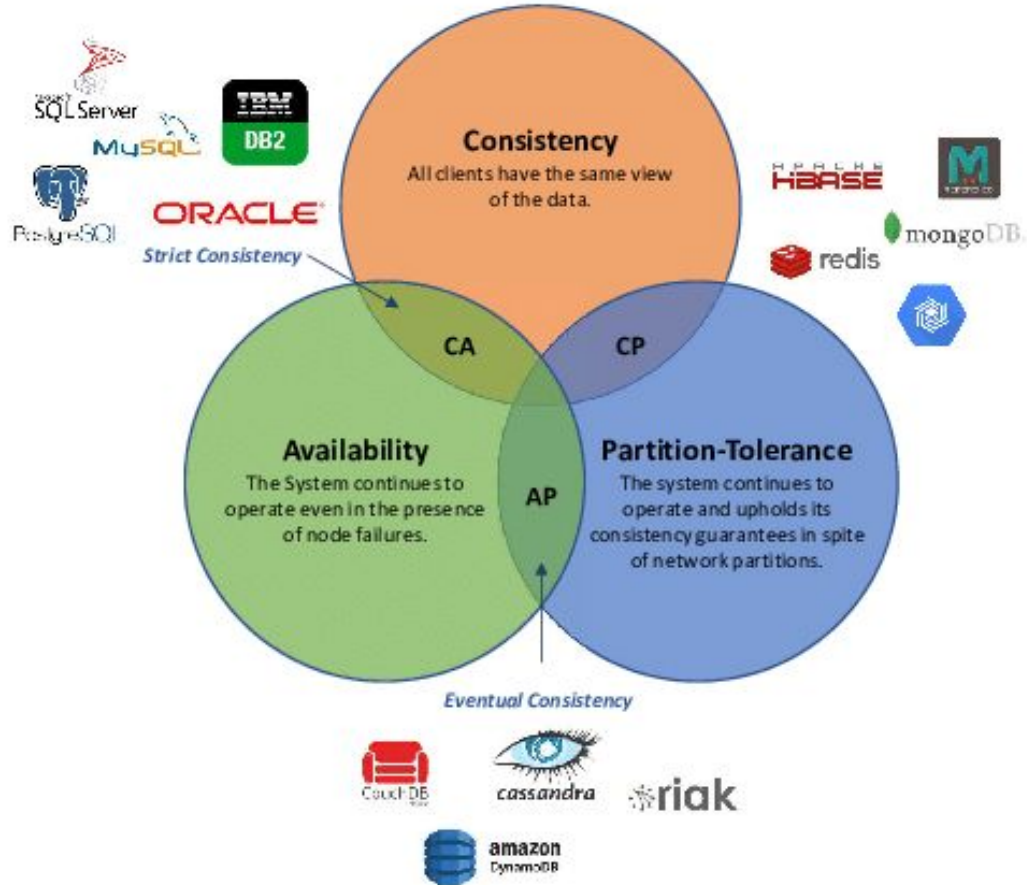
The word "BASE" is written in a colorful, hand-drawn style. Each letter has a different color: 'B' is purple, 'A' is blue, 'S' is green, and 'E' is yellow. The letters are set against a light blue, cloud-like background. The word is underlined with a black horizontal line.

- WEAK CONSISTENCY
- LAST WRITE WINS
- DEVELOPER MANAGED
- SCALE-OUT (UNLIMITED)
- APPROXIMATE ANSWERS
- AVAILABILITY FIRST















خاصیت ACID در مقابل BASE



تئوری CAP



انواع پایگاه داده‌های NoSQL

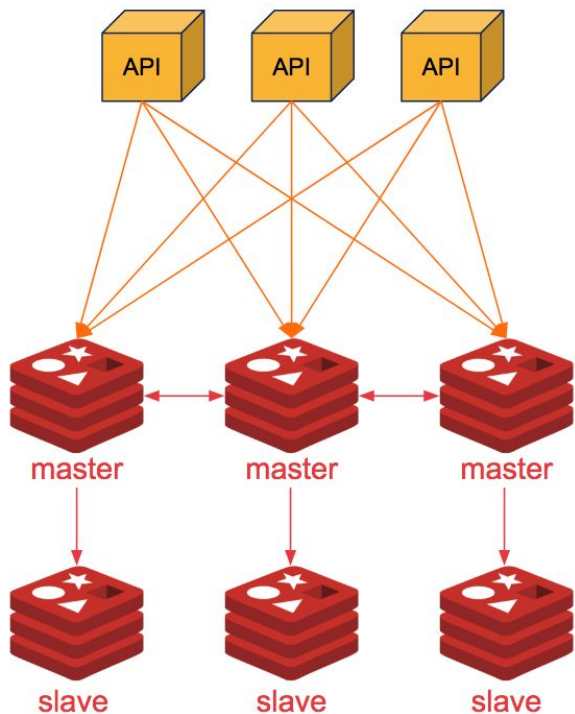
Document Database	Graph Databases
  	 
Key-Value Databases	Wide Column Stores
   	    

پایگاه داده Redis

Key	Value
K1	AAA,BBB,CCC
K2	AAA,BBB
K3	AAA,DDD
K4	AAA,2,01/01/2015
K5	3,ZZZ,5623

- جفت کلید/مقدار
- در حافظه‌ای اصلی (RAM) ذخیره می‌شود
- قابلیت ذخیره در حافظه‌ی دائمی را نیز دارد
- به عنوان حافظه‌ی نهان (cache) نیز مورد استفاده قرار می‌گیرد

پایگاه داده Redis



- از معماری Master/Slave در حالت توزیع شده استفاده می‌کند
- کلیدها توسط تابع درهم‌ساز قابل پیش‌بینی hash می‌شوند
- هر کلید به یک عدد بین ۰ تا ۱۶۳۸۴ تبدیل شده و بر اساس جایگاهش در یکی از کلاسترها قرار می‌گیرد

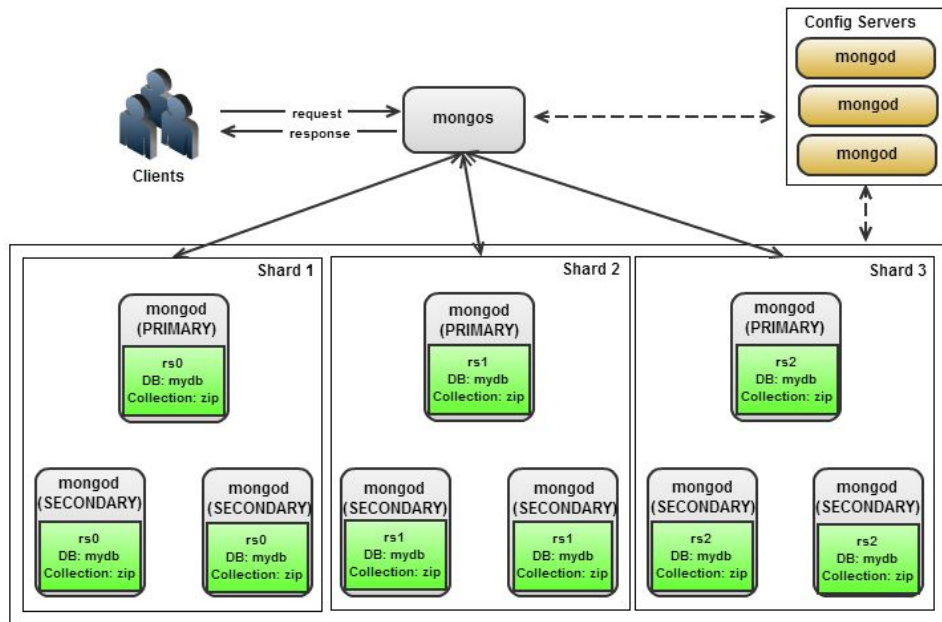
پایگاه داده MongoDB

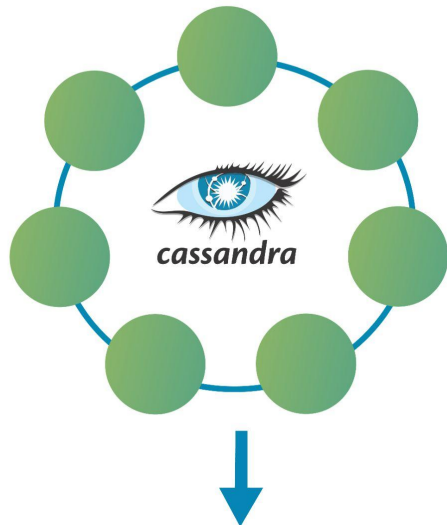
- مبتنی بر سند (document)
- داده‌ها به صورت json قابل ذخیره و بازیابی هستند

```
{
  "_id": "5cf0029caff5056591b0ce7d",
  "firstname": "Jane",
  "lastname": "Wu",
  "address": {
    "street": "1 Circle Rd",
    "city": "Los Angeles",
    "state": "CA",
    "zip": "90404"
  }
  "hobbies": ["surfing", "coding"]
}
```

پایگاه داده MongoDB

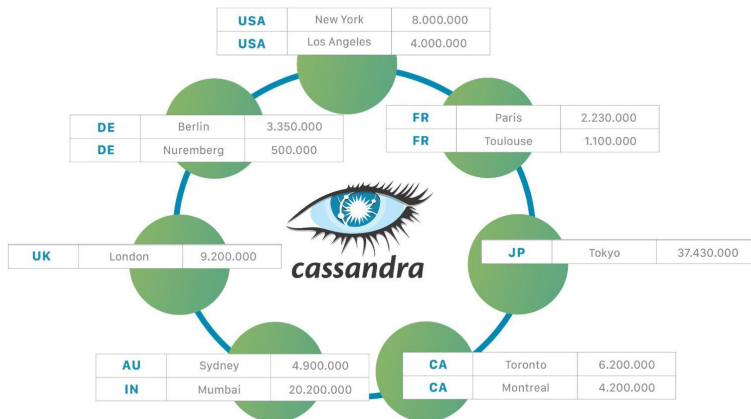
- برای توزیع‌شدگی، داده‌ها به صورت تکه‌هایی در shards قرار می‌گیرند
- کاربر به mongos متصل شده و mongos با استفاده از اطلاعاتی که در config servers ذخیره کرده است، می‌تواند بفهمد که کدام تکه از داده در کدام shard ذخیره شده است





COUNTRY	CITY	POPULATION
USA	New York	8,000,000
USA	Los Angeles	4,000,000
FR	Paris	2,230,000
DE	Berlin	3,350,000
UK	London	9,200,000
AU	Sydney	4,900,000
DE	Nuremberg	500,000
CA	Toronto	6,200,000
CA	Montreal	4,200,000
FR	Toulouse	1,100,000
JP	Tokyo	37,430,000
IN	Mumbai	20,200,000

Partition Key

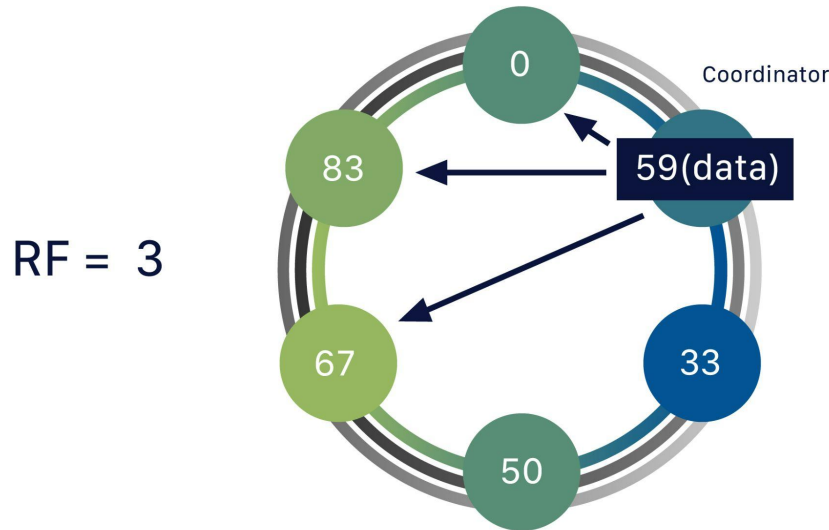


پایگاه داده Cassandra

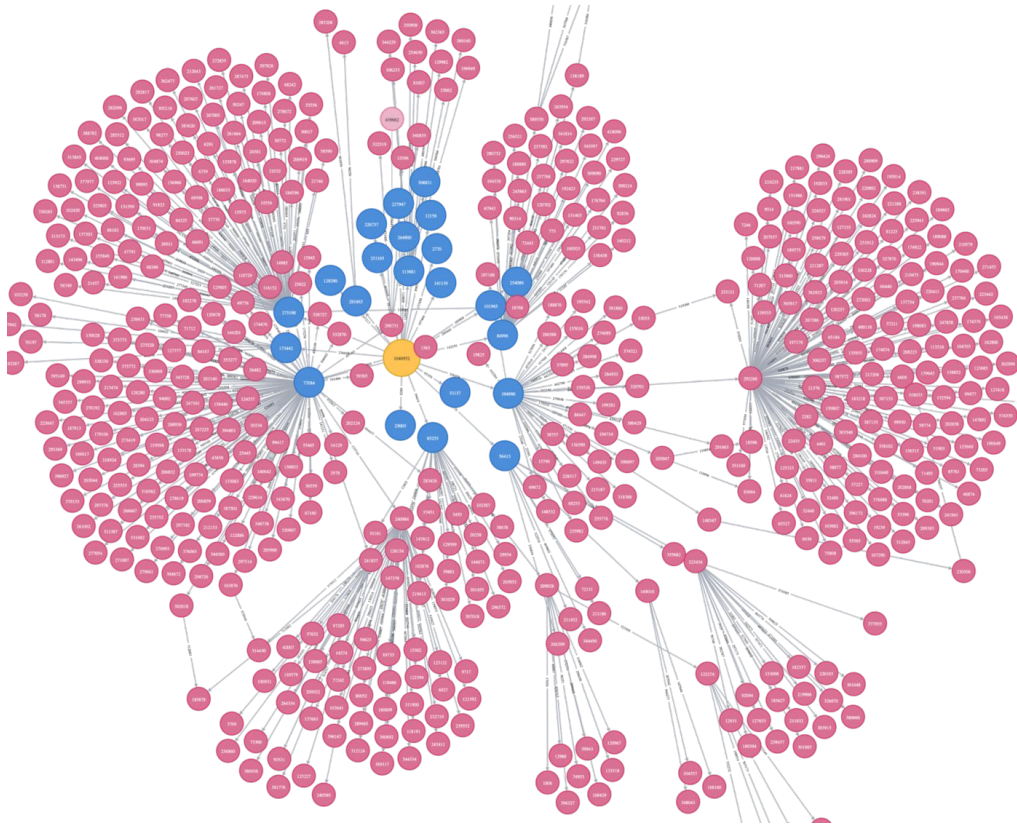
- پایگاه داده مبتنی بر خانواده‌ای از ستون‌هاست
- دسترسی بسیار بالا و سرعت خواندن از ویژگی‌های اصلی این پایگاه داده است

پایگاه داده Cassandra

- توزیع شدگی بدون ارباب (masterless) دارد
- استفاده از قانون quorum در نوشتن و خواندن داده‌ها



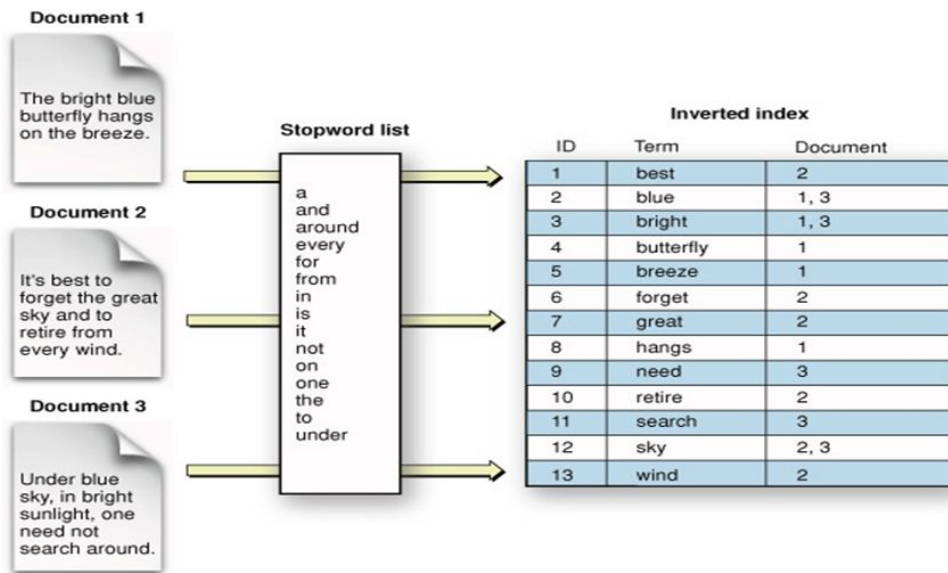
پایگاه داده Neo4J



- پایگاه داده‌ای مبتنی بر رابطه
- هر المان یک گره و هر ارتباط یک یال است
- داده‌ها مبتنی بر ارتباطاتشان ذخیره و پردازش می‌شوند

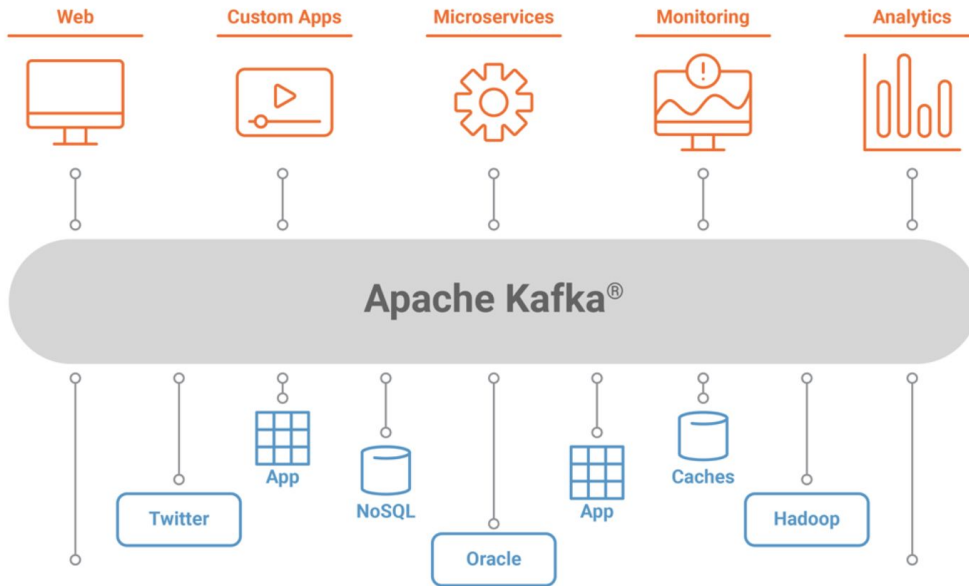
پایگاه داده ElasticSearch

- ذخیره‌ی داده‌های متنی با استفاده از روش شاخص‌گذاری معکوس (inverted index)
- سرعت جستجوی بالا و توانایی توزیع‌شدگی قوی



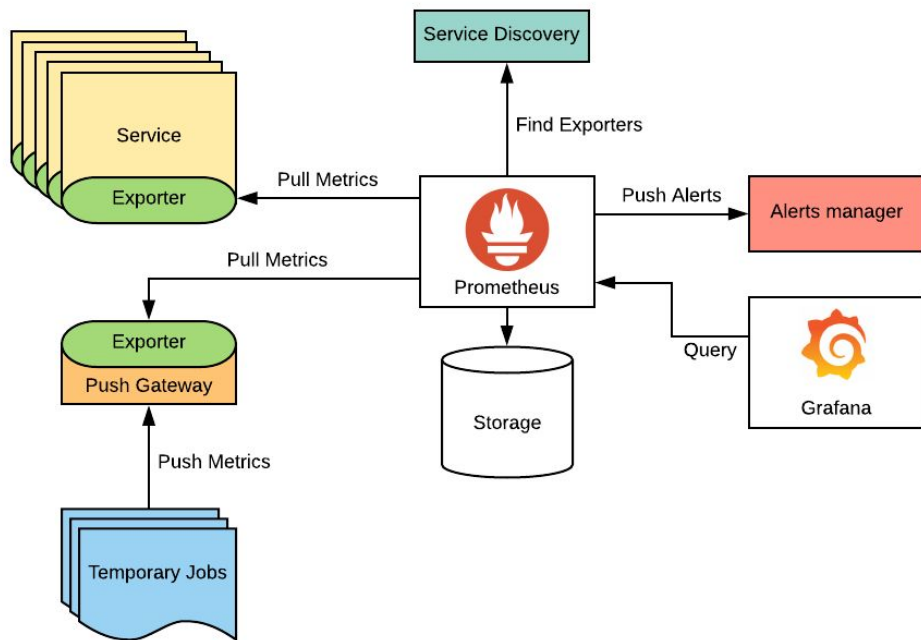
Apache Kafka

- برای ارسال و دریافت پیام با سرعت بالا و به صورت مطمئن
- قابلیت صفبندی پیام به صورت توزیع شده



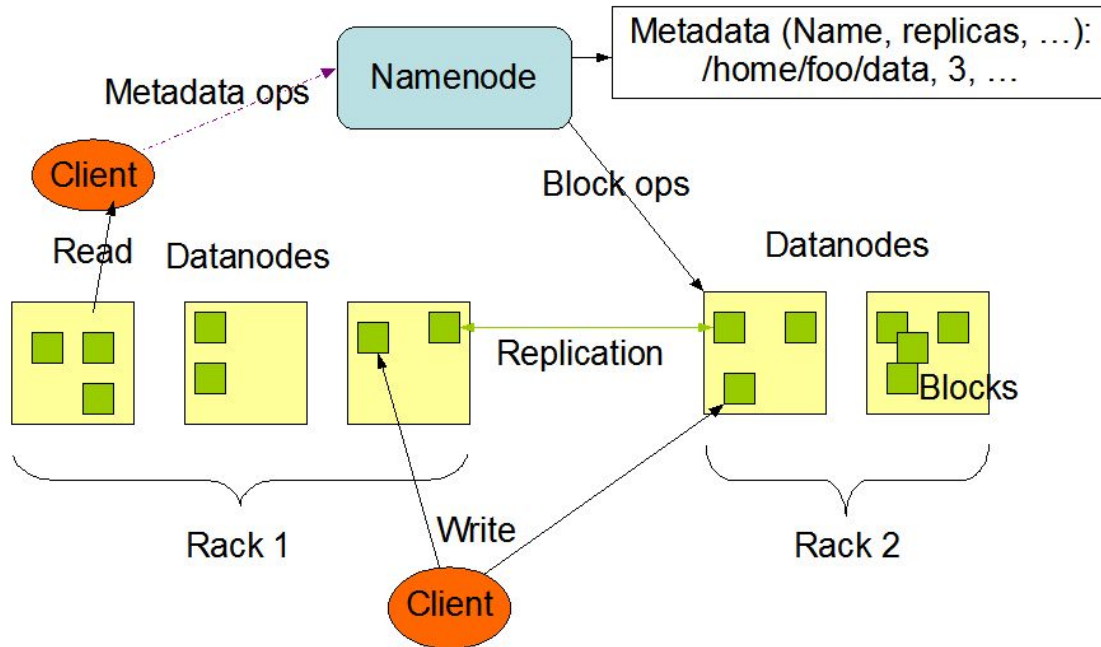
پایگاه داده‌ی Prometheus

- پایگاه داده مبتنی بر زمان (series time)
- از pulling برای دریافت و ذخیره‌ی داده‌ها استفاده می‌کند

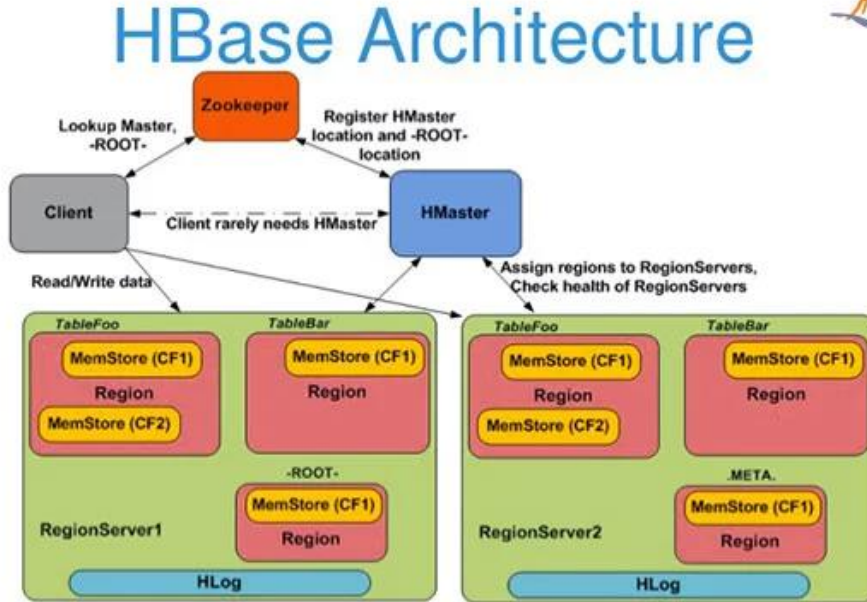


سیستم فایل توزیع شده (DFS) و هادوپ

HDFS Architecture

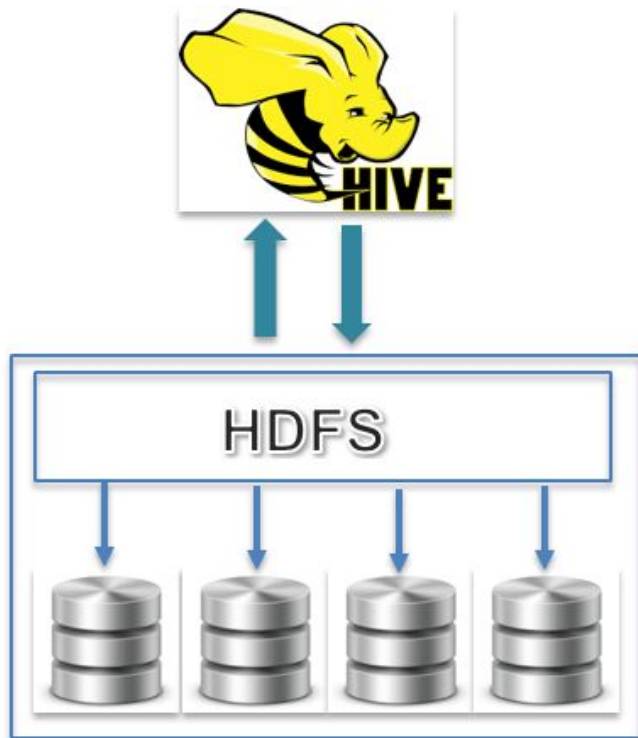


Apache HBase و پایگاه داده‌ی HDFS



- ذخیره داده‌ها بر اساس سیستم‌فایل توزیع شده (DFS)
- استفاده از ساختار ارباب-برده (master-slave) برای توزیع داده‌ها

Apache Hive و انبار داده‌ی HDFS



- انبار داده‌ی Apache Hive برای ذخیره‌سازی داده‌های حجیم مبتنی بر زمان استفاده می‌شود

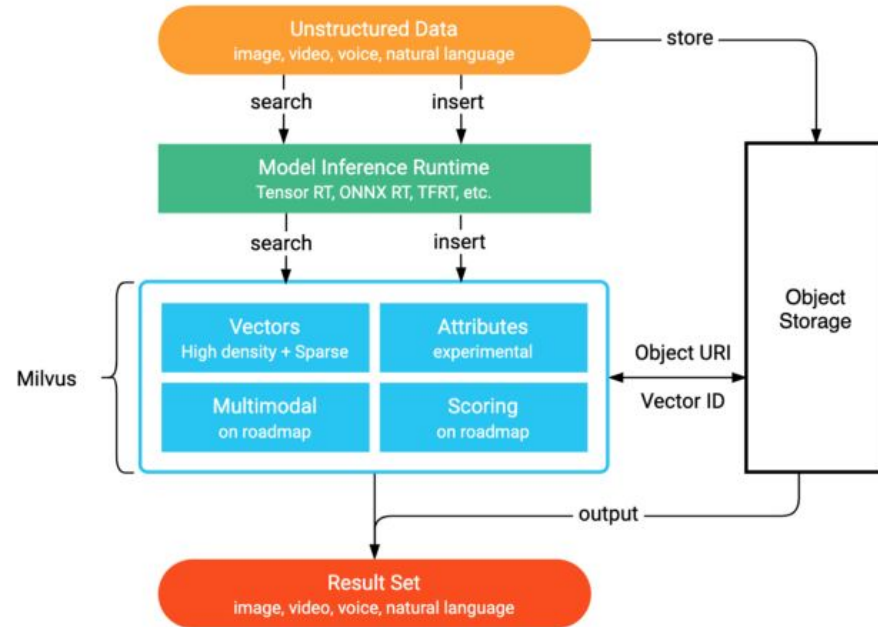
انبار داده‌ی Click House

- ذخیره‌ی داده‌های حجیم مبتنی بر جداول Fact و ابعاد



Learn
ClickHouse

پایگاه داده‌ی Milvus



- ذخیره و شاخص‌گذاری و جستجوی سریع و کتورها
- با استفاده از روش‌های قسمت‌بندی فضا (space partitioning) جستجو به قسمت‌های کوچک‌تر تقسیم‌بندی می‌کند
- جستجوی ANN در مقابل KNN

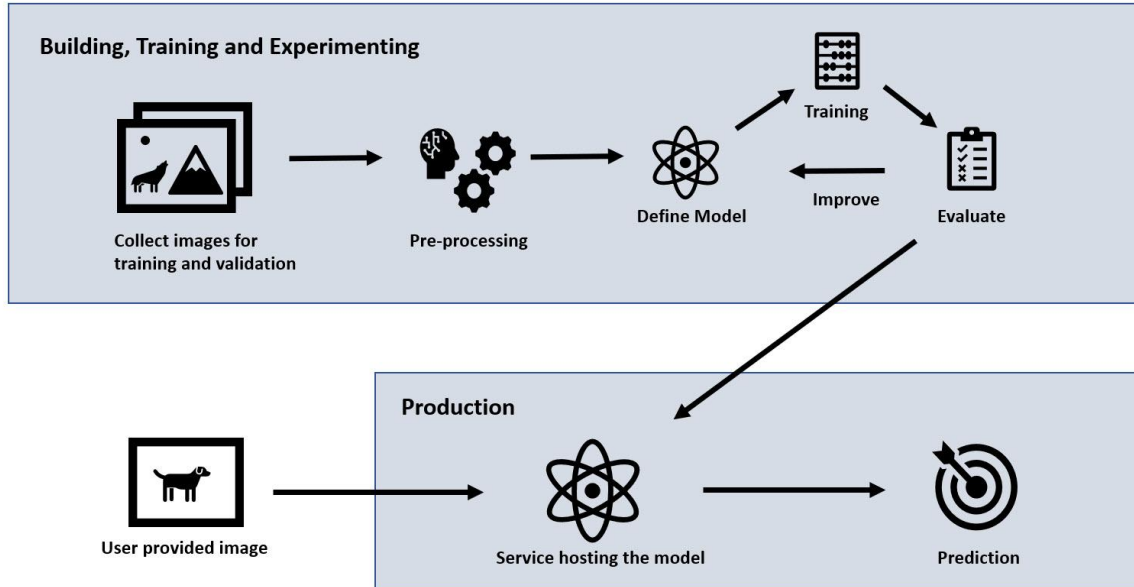
پایگاه داده‌ی Weaviate

- ذخیره و شاخص‌گذاری و جستجوی سریع و کتورها

- با استفاده از روش‌های قسمت‌بندی فضا (space partitioning) فضا را برای جستجو به قسمت‌های کوچک‌تر تقسیم‌بندی می‌کند



یک فرآیند یادگیری ماشین

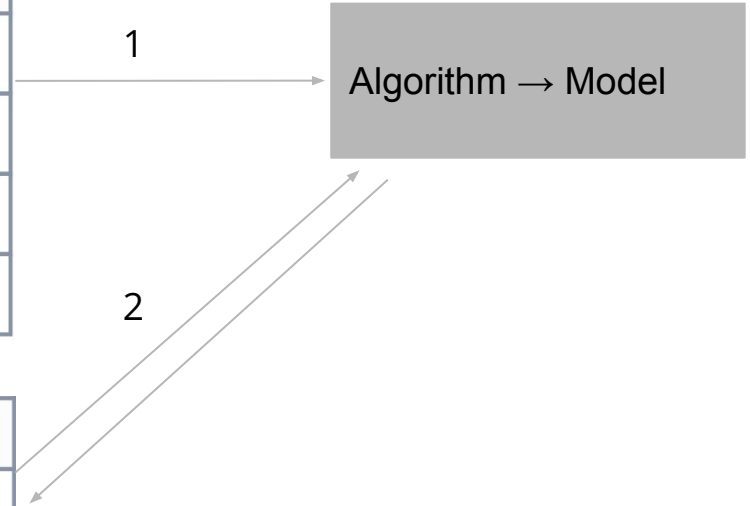


یک فرآیند یادگیری ماشین

مجموعه داده‌ی رانندگان تاکسی اینترنتی

	MKPD	MR	MTPD	...	Age	Gender	Label (Class)
#1	11.6	9.6	23	...	36	0	No
#2	9.1	9.8	53	...	25	0	Yes
#3	9.2	8.6	50	...	60	0	No
...

	MKPD	MR	MTPD	...	Age	Gender	Label (Class)
#new	12.6	8.6	24	...	34	0	?



تمرین...

مجموعه داده‌ی رانندگان تاکسی اینترنتی

	MKPD	MR	MTPD	...	Age	Gender	Label (Class)
#1	11.6	9.6	23	...	36	0	No
#2	9.1	9.8	53	...	25	0	Yes
#3	9.2	8.6	50	...	60	0	No
...

	MKPD	MR	MTPD	...	Age	Gender	Label (Class)
#new	12.6	8.6	24	...	34	0	?

Algorithm → Model

1

2

Resources (Comma Separated!):

<https://iranatlas.info/life/writing5.htm>,
<https://www.yjc.news/fa/news/5269545/%D9%85%D8%AA%D9%86-%D9%82%D8%B1%D8%A2%D9%86-%D8%A8%D8%B1-%D8%B1%D9%88%DB%8C-%D9%BE%D9%88%D8%B3%D8%AA-%D8%A8%D8%B2-%D9%88-%DA%AF%D9%88%D8%B3%D9%81%D9%86%D8%AF>,
<https://www.bbntimes.com/technology/ssd-or-hdd-how-each-type-works-and-what-it-means-for-you>,
<https://www.cloudzero.com/blog/horizontal-vs-vertical-scaling>, <https://www.m-brain.com/la-technologie-m-brain/>,
<https://medium.com/geekculture/sql-vs-nosql-which-database-to-choose-347839f4513f>,
<https://www.scaler.com/topics/dbms/relational-algebra-in-dbms/>,
<https://luminousmen.com/post/acid-vs-base-comparison-of-two-design-philosophies/>,
https://www.debadityachakravorty.com/system_design/captheorem/, <https://intellipaat.com/blog/what-is-apache-hbase/>,
<https://www.complexsql.com/difference-between-sql-and-nosql/>, <https://chathuranga94.medium.com/introduction-to-redis-348d9ccbfd0d>,
<https://medium.com/@ragulan28/redis-in-action-5c6b4706a977>,
<https://www.freecodecamp.org/news/a-thorough-introduction-to-distributed-systems-3b91562c9b3c/>,
<https://k21academy.com/terraform-iac/why-terraform-not-chef-ansible-puppet-cloudformation/>, <https://weaviate.io/>,
<https://www.mongodb.com/document-databases>, <https://chiruideas.weebly.com/mongo-db.html>,
https://cassandra.apache.org/_/cassandra-basics.html, <https://towardsdatascience.com/getting-started-graph-database-neo4j-df6ebc9ccb5b>,
<https://stackoverflow.com/questions/47003336/elasticsearch-index-sharding-explanation>,
<https://clickhouse.com/learn/lessons/whatsnew-clickhouse-21.10>, <https://docs.confluent.io/5.5.1/kafka/introduction.html>,
<https://www.devopsschool.com/blog/what-is-prometheus-and-how-it-works/>,
<https://www.databricks.com/kr/glossary/hadoop-distributed-file-system-hdfs>, <https://logz.io/blog/hive-vs-spark/>,
<https://milvus.io/blog/Thanks-to-Milvus-Anyone-Can-Build-a-Vector-Database-for-1-Billion-Images.md>,
<https://tprojects.schneider-electric.com/GeoSCADAHelp/Geo%20SCADA%202020/Content/ServerAdministrationGuide/DisplayaServerLogFile.htm>