ICDS

پایگاه داده‌ها در علم داده

Databases for Data Science

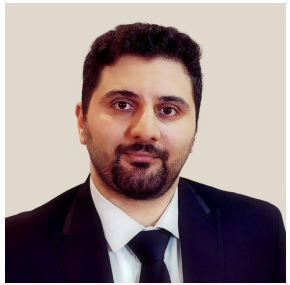
MasoudKaviani.ir

MasoudKaviani.ir

مسعود کاویانی

» دانشمند ارشد داده در صبا ایده (فیلیمو، 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

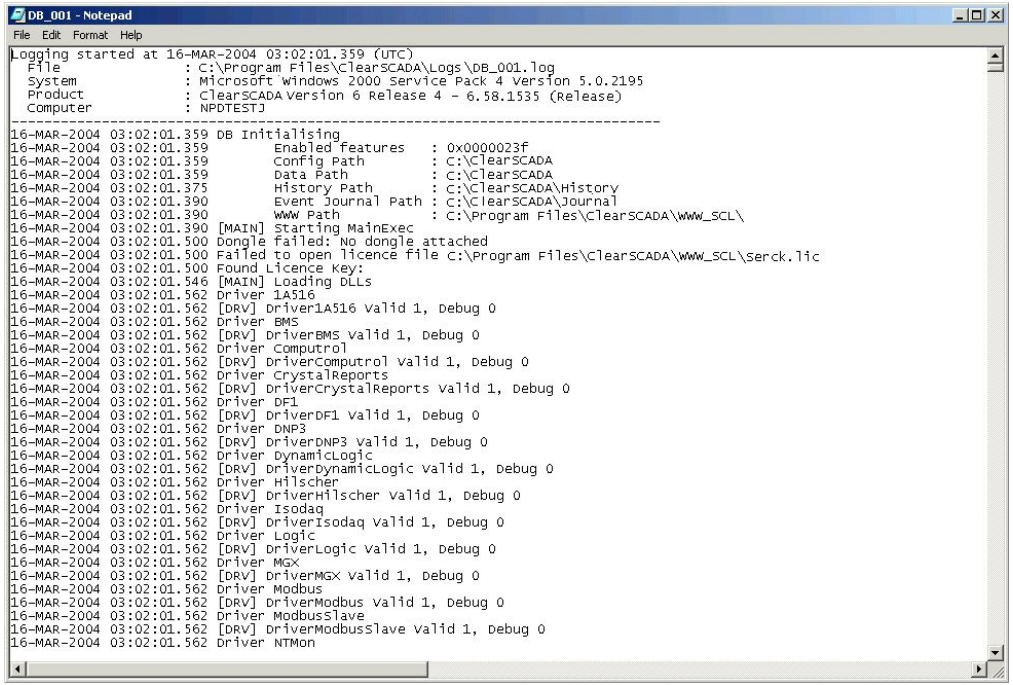
- و با پیشرفت علم دیسک‌های سخت و SSDها اختراع شدند

HDD

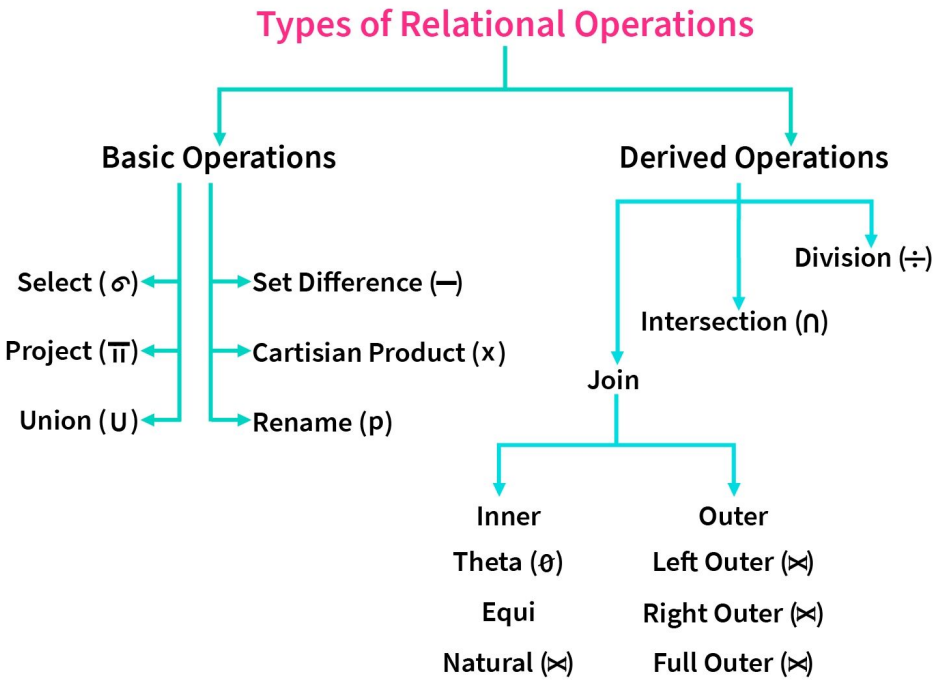
SSD

ذخیره داده‌ها در فایل

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



جبر رابطه‌ای (Relational Algebra)



- 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!

مزایا و معایب پایگاه داده‌های رابطه‌ای

مزایا:

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

معایب:

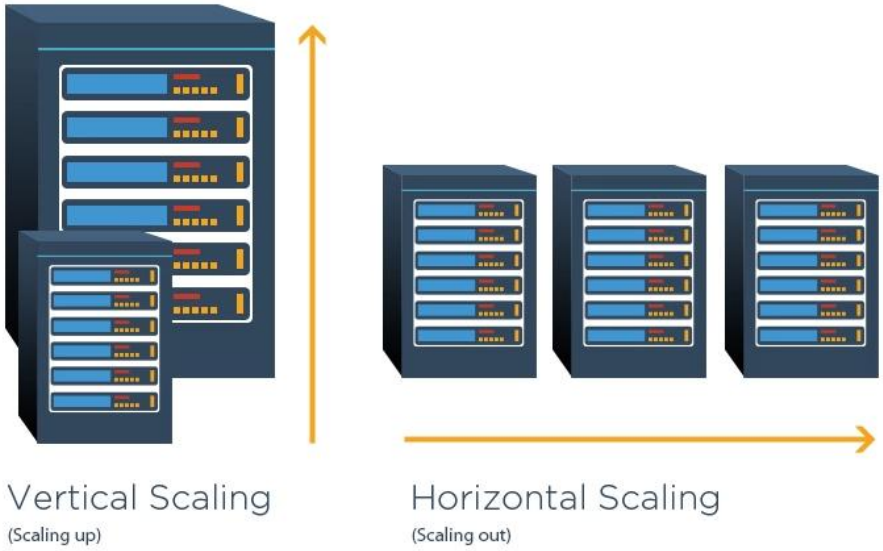
- شمای ثابت دارند و تغییر در شما در برخی از مواقع دشوار است
- پیچیدگی محاسباتی به دلیل محدودیت‌های اعمال شده (مانند نرمال‌سازی و رابطه‌ها)
- به راحتی Scale Out نمی شوند

بیگ‌دیتا

- حجم، سرعت و تنوع بالا

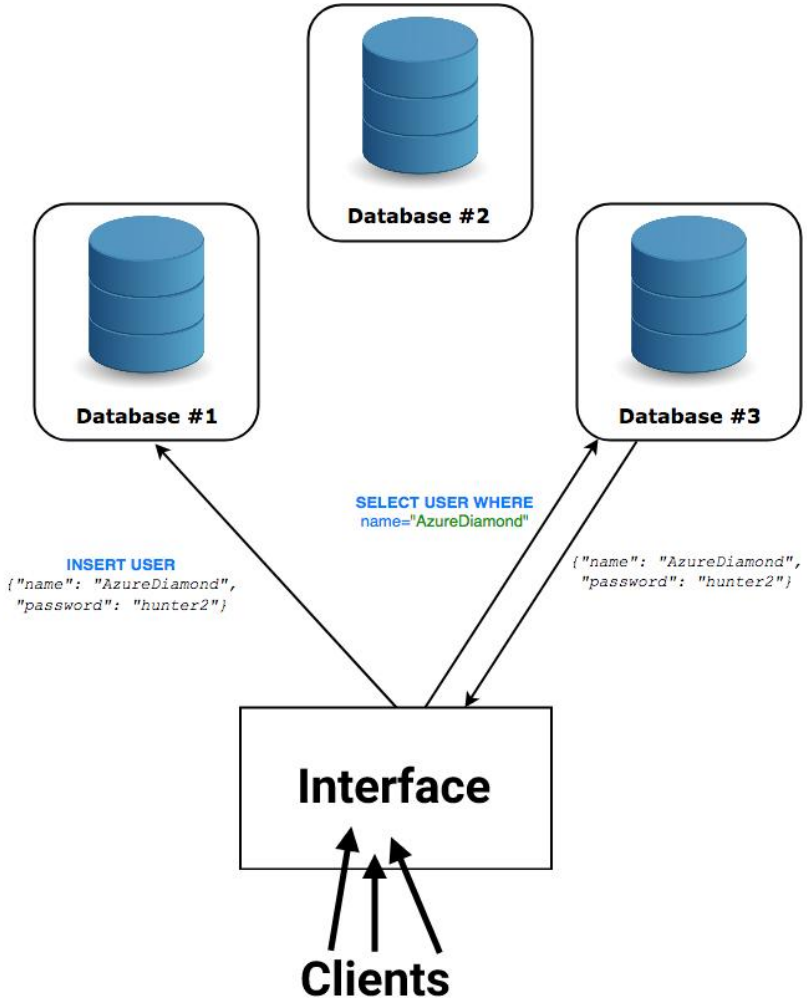
افزایش حجم و سرعت تولید داده‌ها

- 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

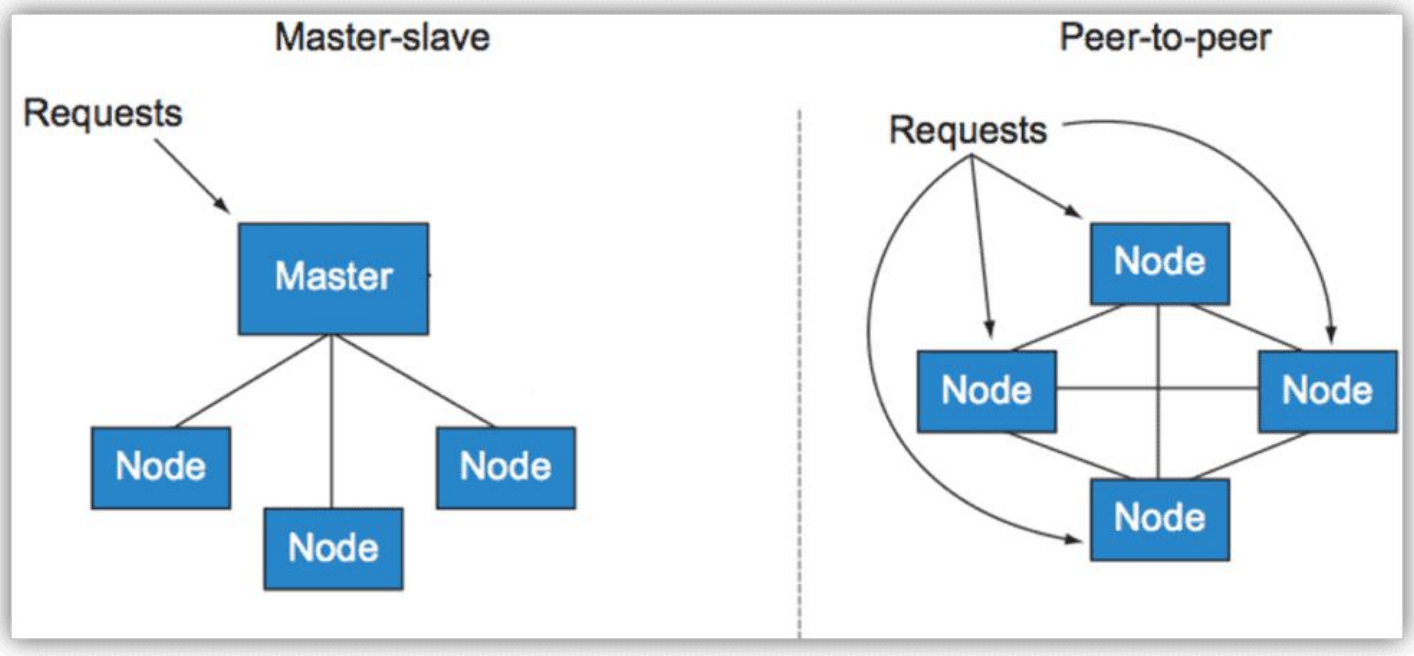


سیستم توزیع شده

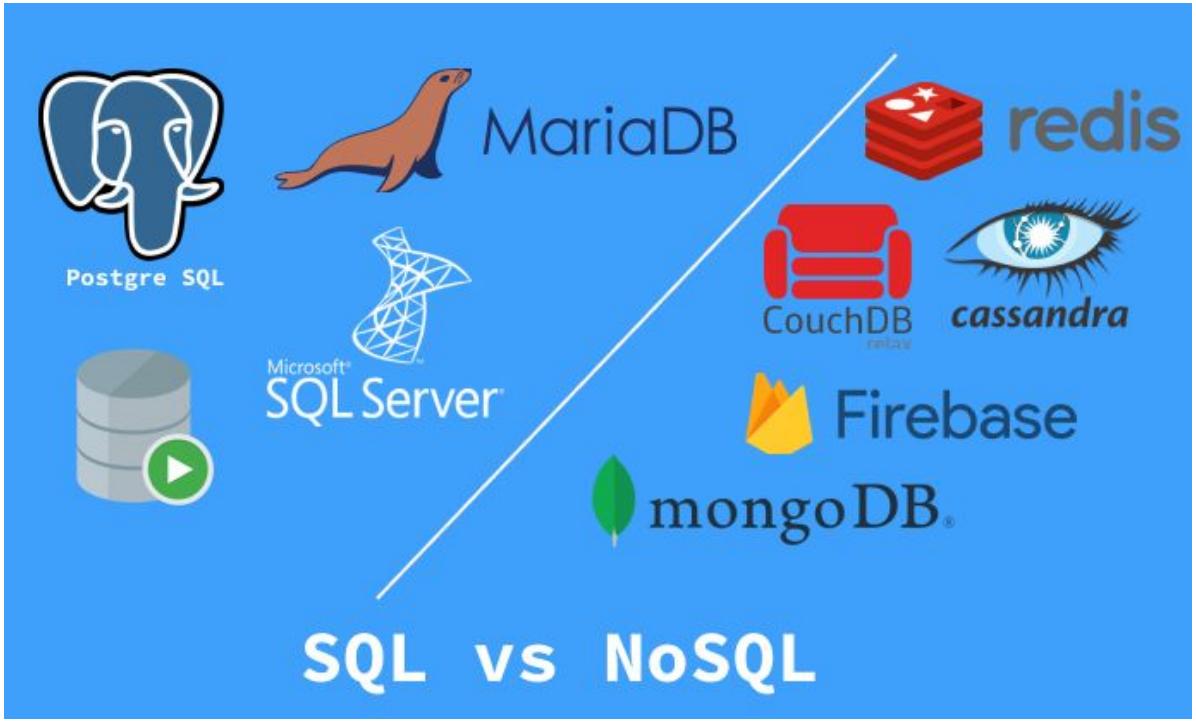
- مجموعه‌ای از کامپیوترها که با تبادل اطلاعات و ارتباطات با یکدیگر، شبیه به یک سیستم منسجم می‌شوند
- برای scale out کردن نیاز به سیستم توزیع شده است



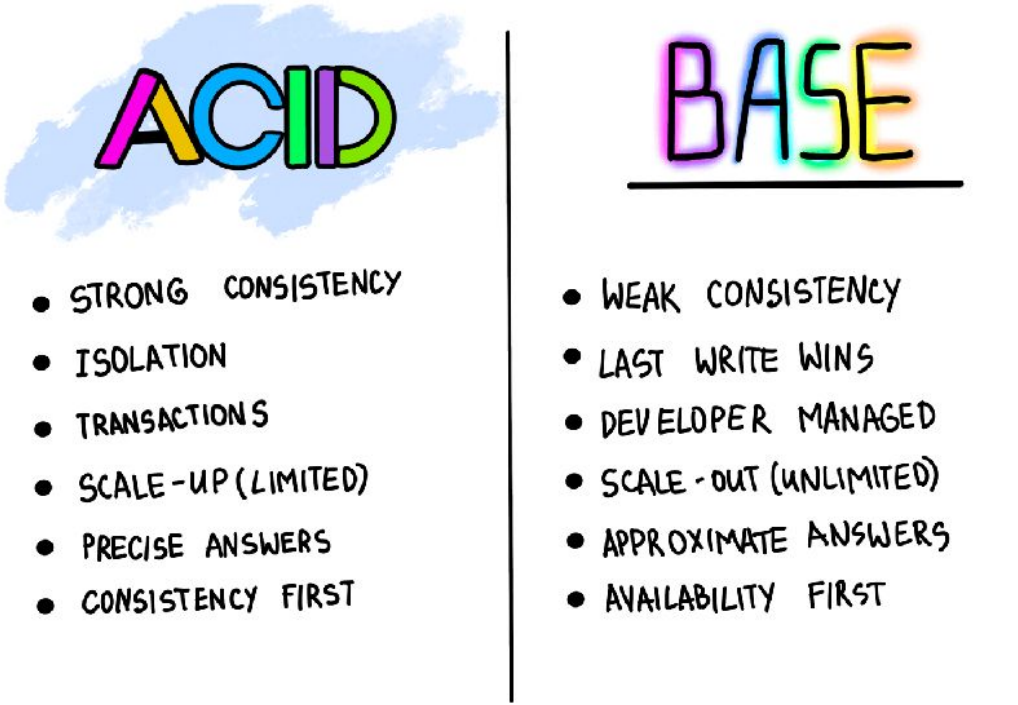
سیستم‌های توزیع شده Master-Slave VS Masterless



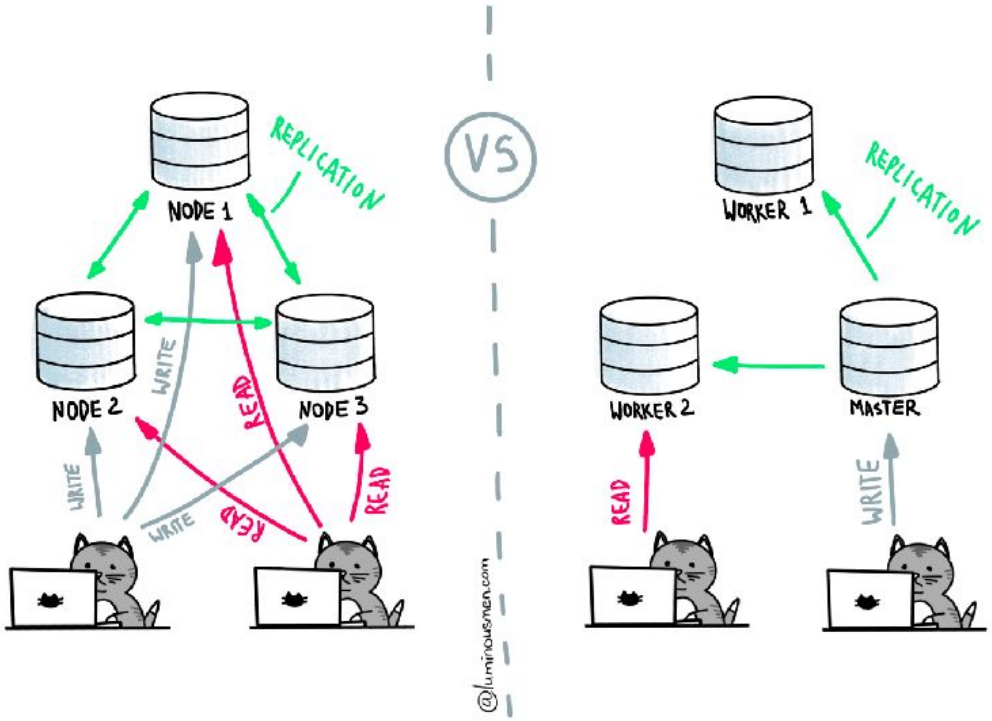
خانواده‌ی SQL در مقابل NoSQLها



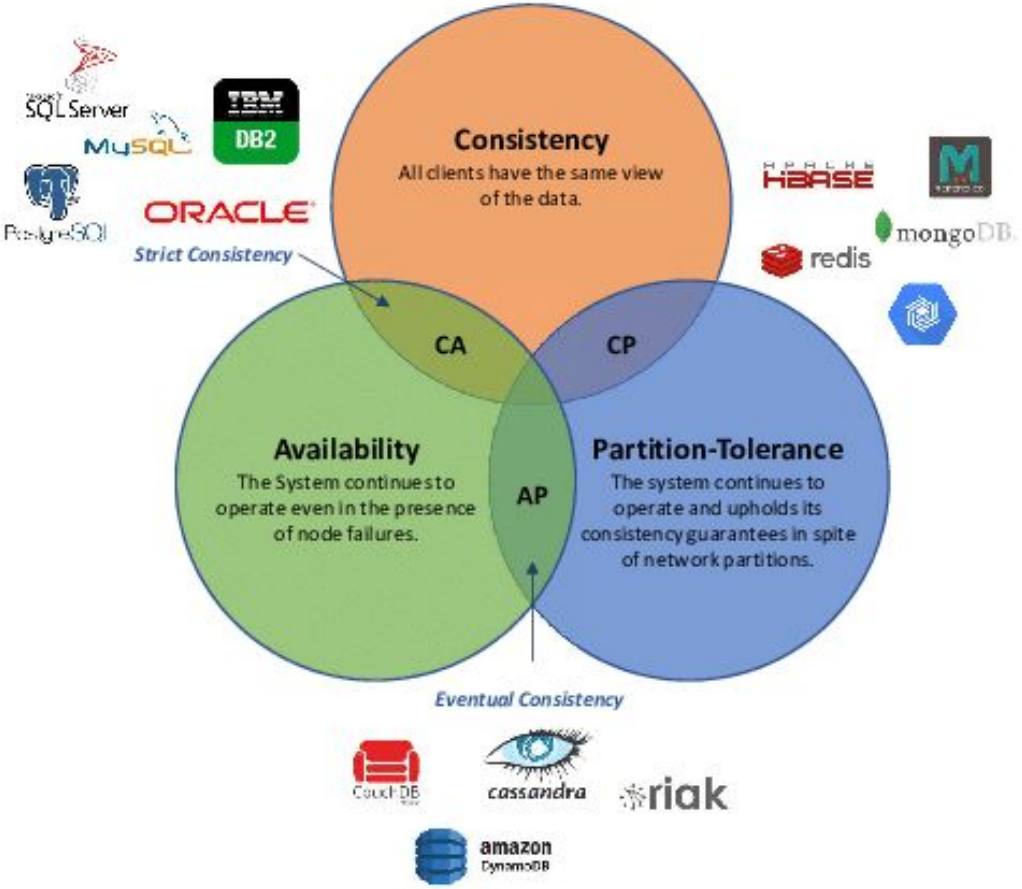
خاصیت ACID در مقابل BASE



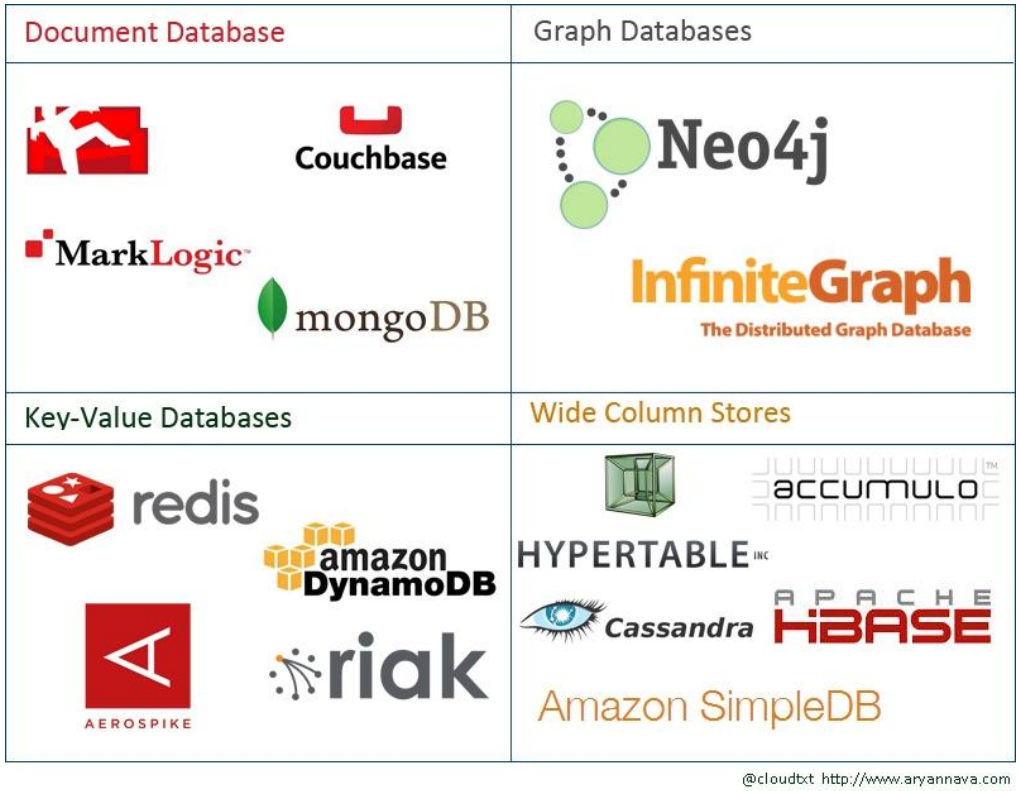
خاصیت ACID در مقابل BASE



تئوری CAP


Consistency
All clients have the same view of the data.
Availability
The System continues to operate even in the presence of node failures.
Partition-Tolerance
The system continues to operate and upholds its consistency guarantees in spite of network partitions.
CA
CP
AP
Strict Consistency
Eventual Consistency
SQL Server
MySQL
DB2
ORACLE
PostgreSQL
HBASE
mongoDB.
redis
CouchDB
cassandra
riak
amazon
DynamoDB

انواع پایگاه داده‌های NoSQL

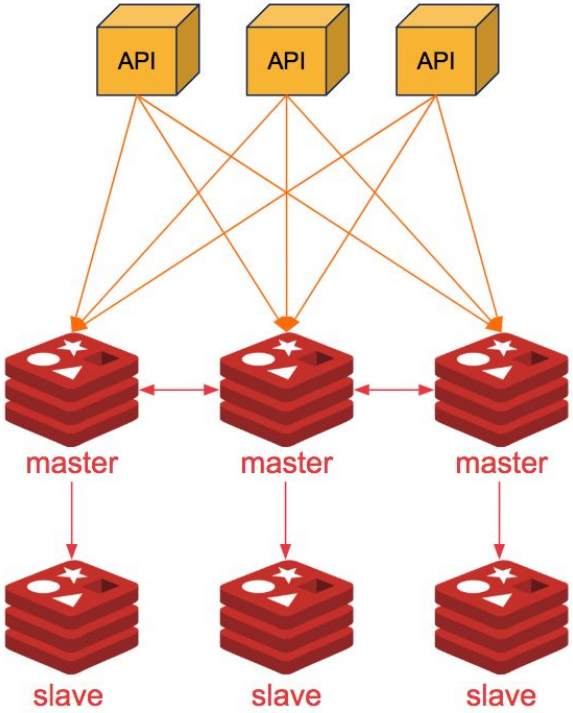


پایگاه داده Redis

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 توزیع‌شده 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

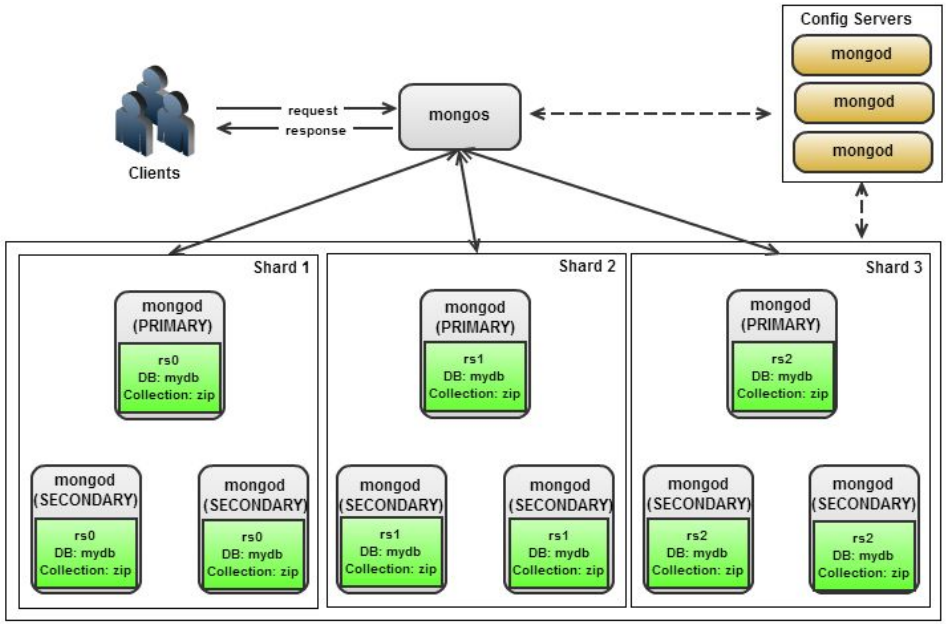
پایگاه داده MongoDB

- 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"]
}
```

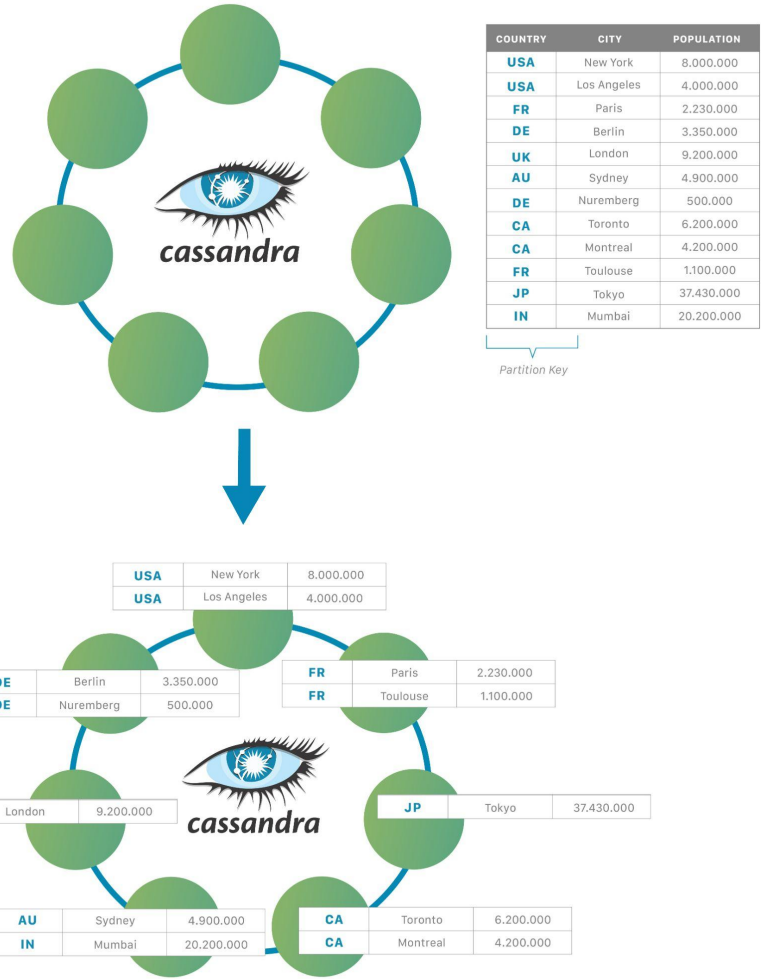
پایگاه داده MongoDB

- 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



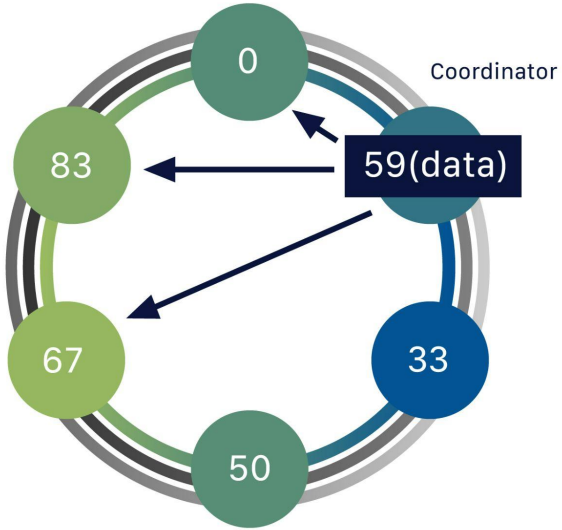
پایگاه داده Cassandra

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



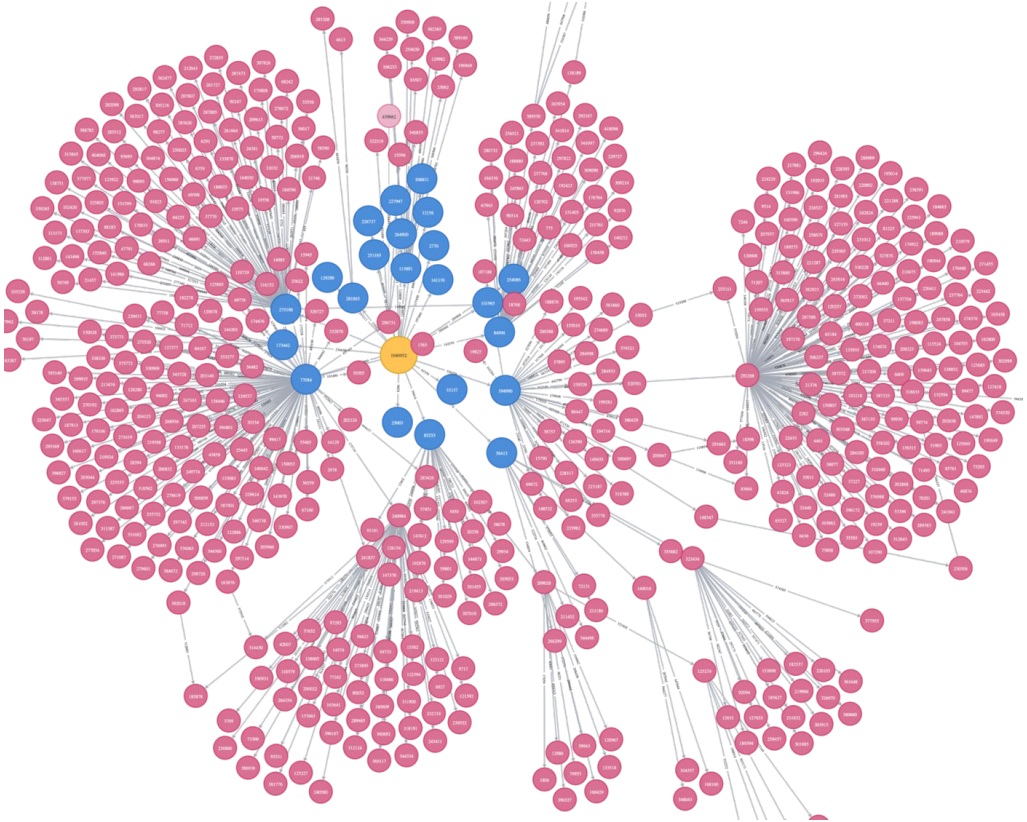
پایگاه داده Cassandra

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



RF = 3

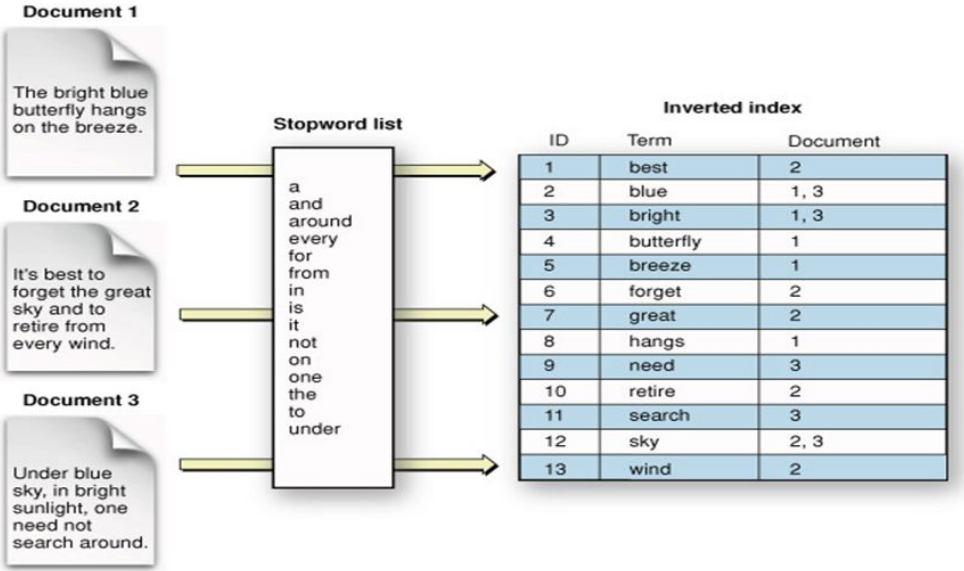
پایگاه داده Neo4J



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

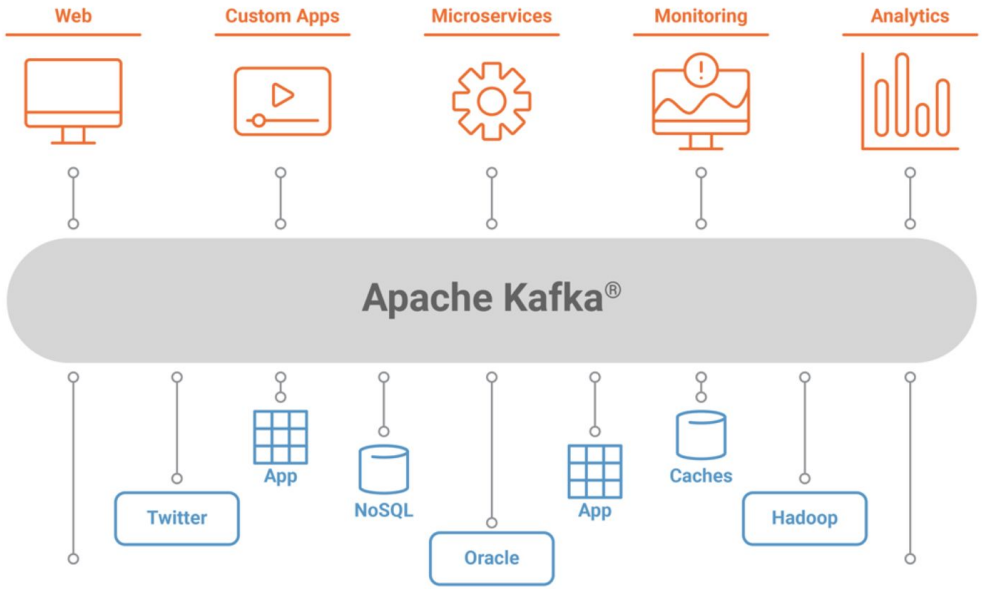
پایگاه داده ElasticSearch

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



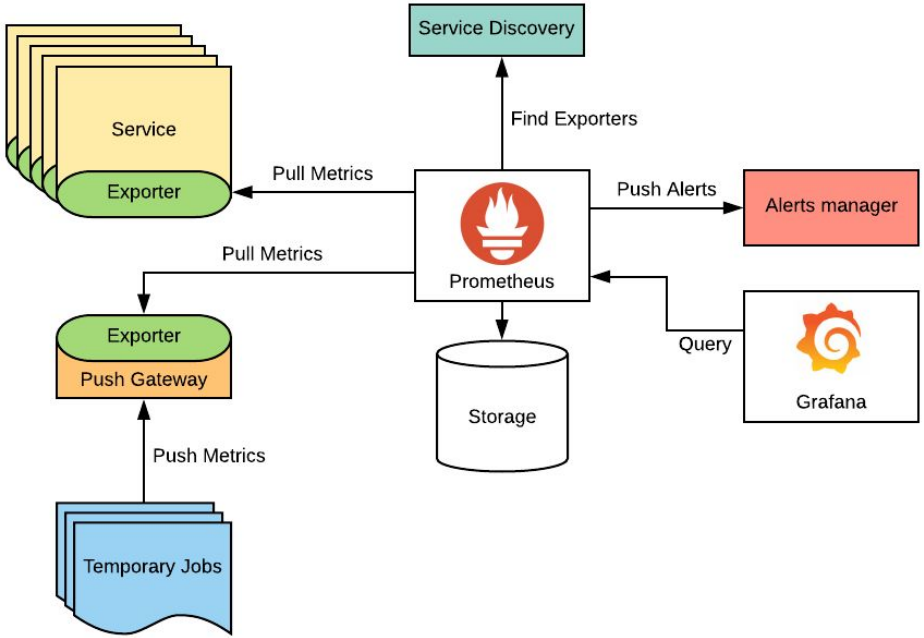
Apache Kafka

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

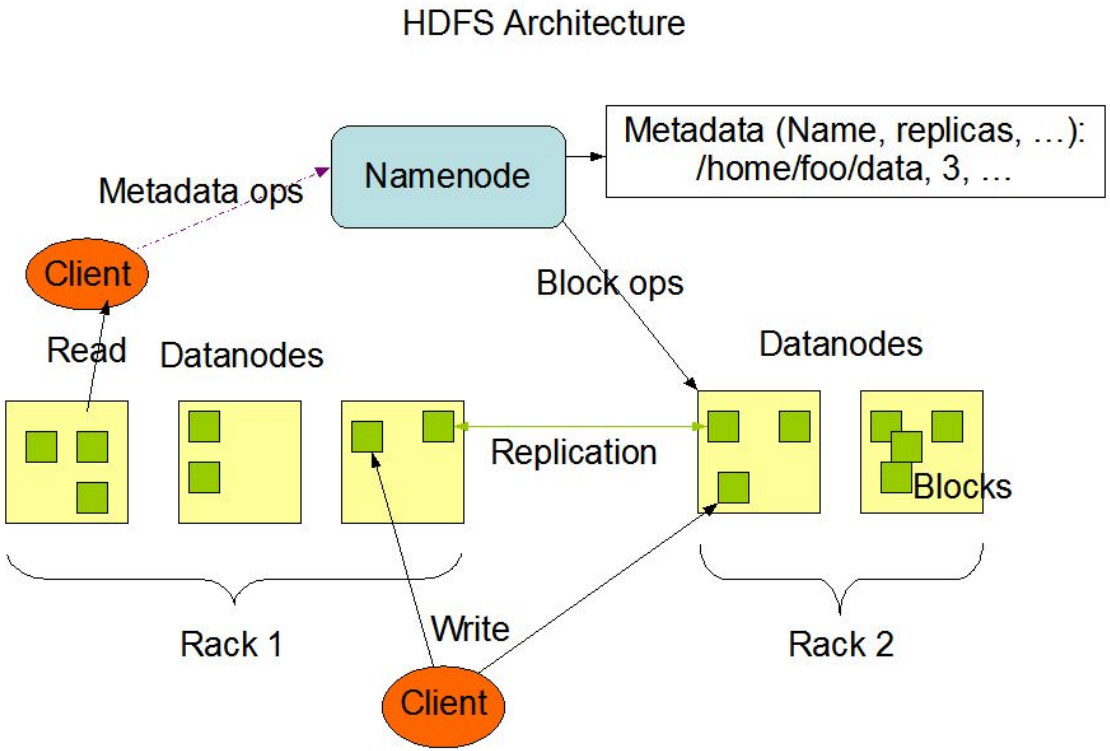


پایگاه داده‌ی Prometheus

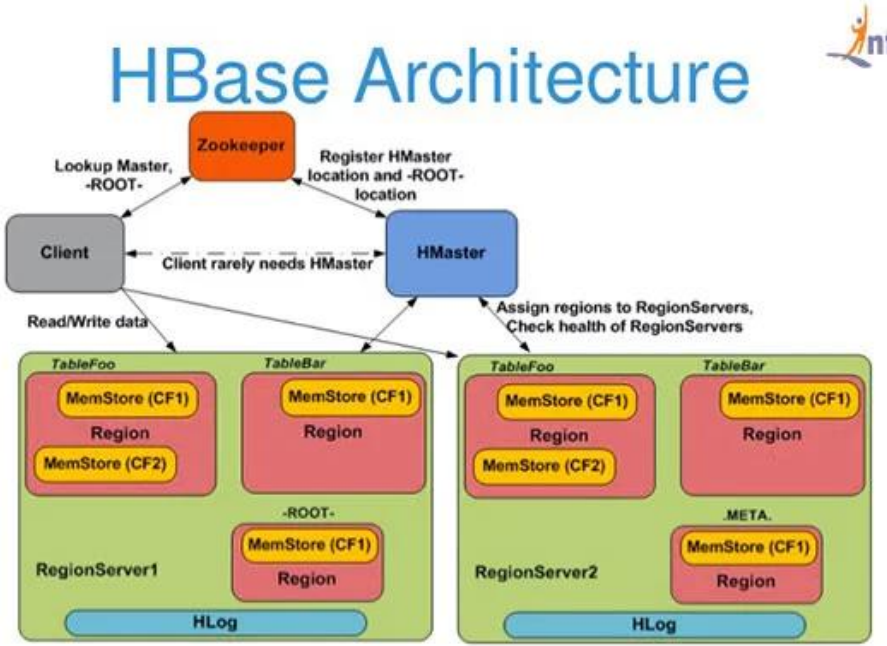
- پایگاه داده مبتنی بر زمان (series time)
- از pulling برای دریافت و ذخیره‌ی داده‌ها استفاده می‌کند



سیستم فایل توزیع شده (DFS) و هدوپ

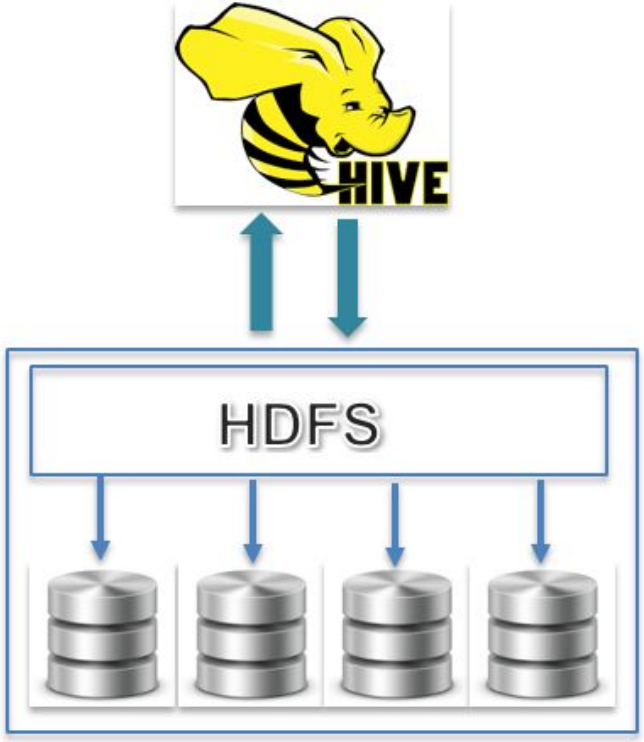
- 

HDFS و پایگاه داده‌ی Apache HBase



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

HDFS و انبار داده‌ی Apache Hive



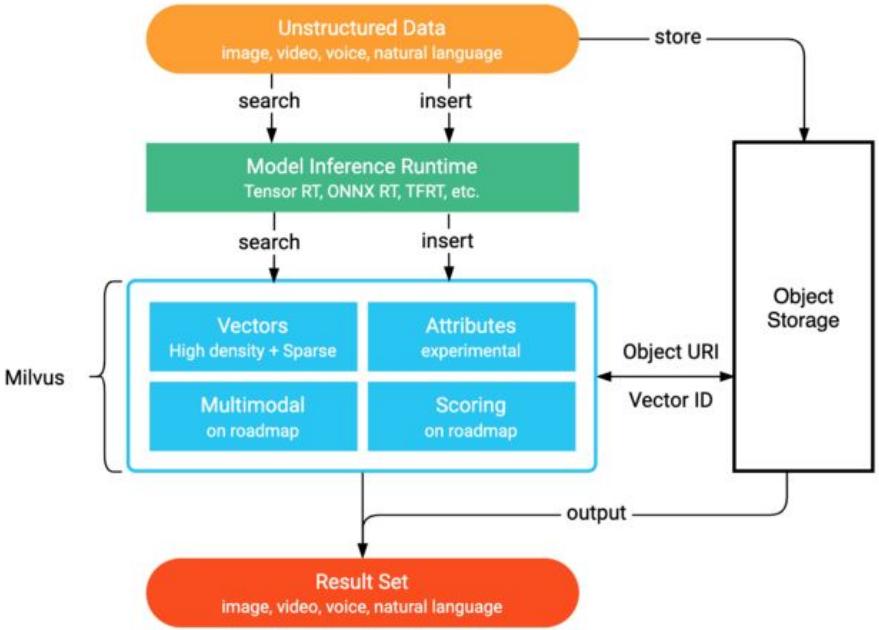
- انبار داده‌ی Apache Hive برای ذخیره‌سازی داده‌های حجیم مبتنی بر زمان استفاده می‌شود

انبار داده‌ی Click House

- ذخیره‌ی داده‌های حجیم مبتنی بر جداول Fact و ابعاد

Learn
ClickHouse

پایگاه داده‌ی Milvus

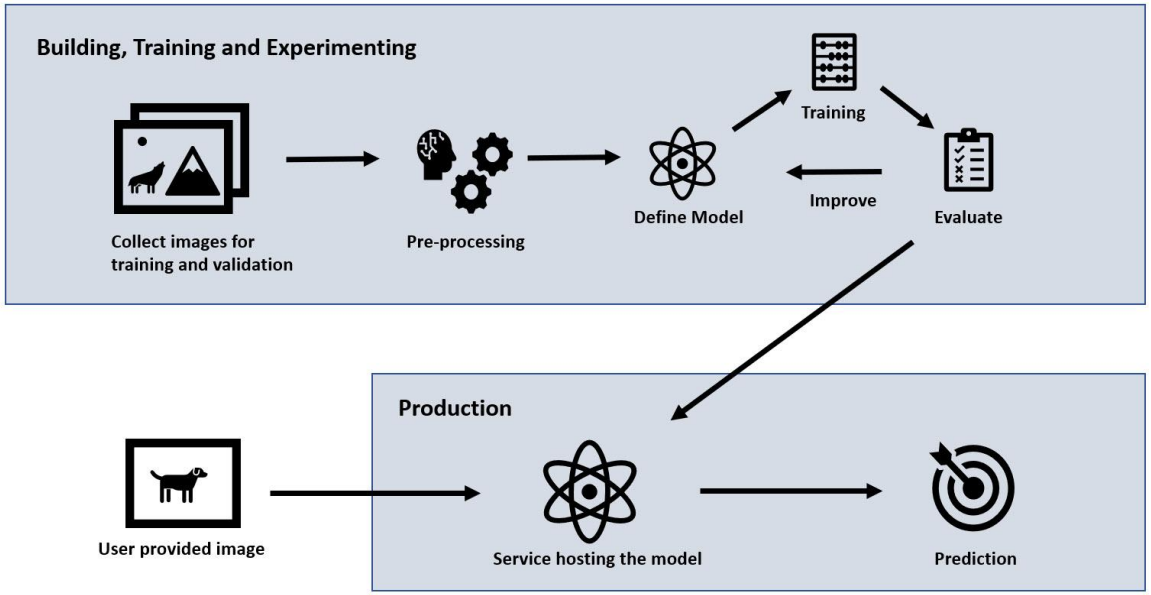


- ذخیره و شاخص‌گذاری و جستجوی سریع وکتورها
- با استفاده از روش‌های قسمت‌بندی فضا (space partitioning) فضا را برای جستجو به قسمت‌های کوچک‌تر تقسیم‌بندی می‌کند
- جستجوی ANN در مقابل KNN

پایگاه داده‌ی Weaviate

- ذخیره و شاخص‌گذاری و جستجوی سریع وکتورها
- 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
...	...	...	...	...	...	...	...

1

Algorithm → Model

2

	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
...	...	...	...	...	...	...	...

1

Algorithm → Model

2

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

Resources (Comma Separated!):

https://iranatlas.info/life/wrtiting5.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