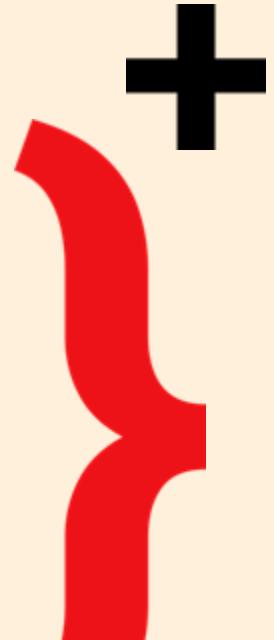




Couchbase Capella Workshop

> GenAI (LLM, RAG)

2시에 시작하겠습니다.



2024.10.30

손 광락, Solutions Engineer

Capella Webinar

구분	일자	Webinar 주제
시리즈 1	2024-09-25	벡터 검색을 활용한 AI Powered 어플리케이션 구축
시리즈 2	2023-10-30	벡터 검색을 활용한 GenAI(LLM/RAG) 어플리케이션 구축
시리즈 3	2024-11-27	벡터 검색을 활용한 Mobile On-Device AI 어플리케이션 구축



Agenda

- 14:00 – 14:10 등록 확인 & 실습 환경 확인
- 14:10 – 14:35 카우치베이스 소개
- 14:35 – 15:00 AI, Vector 와 Vector 검색, GenAI
- 15:00 – 15:10 Break
- 15:10 – 15:25 Capella/Couchbase 사용법 실습
- 15:25 – 15:50 GenAI(LLM,RAG) 실습/데모
- 15:50 – 16:00 Q & A

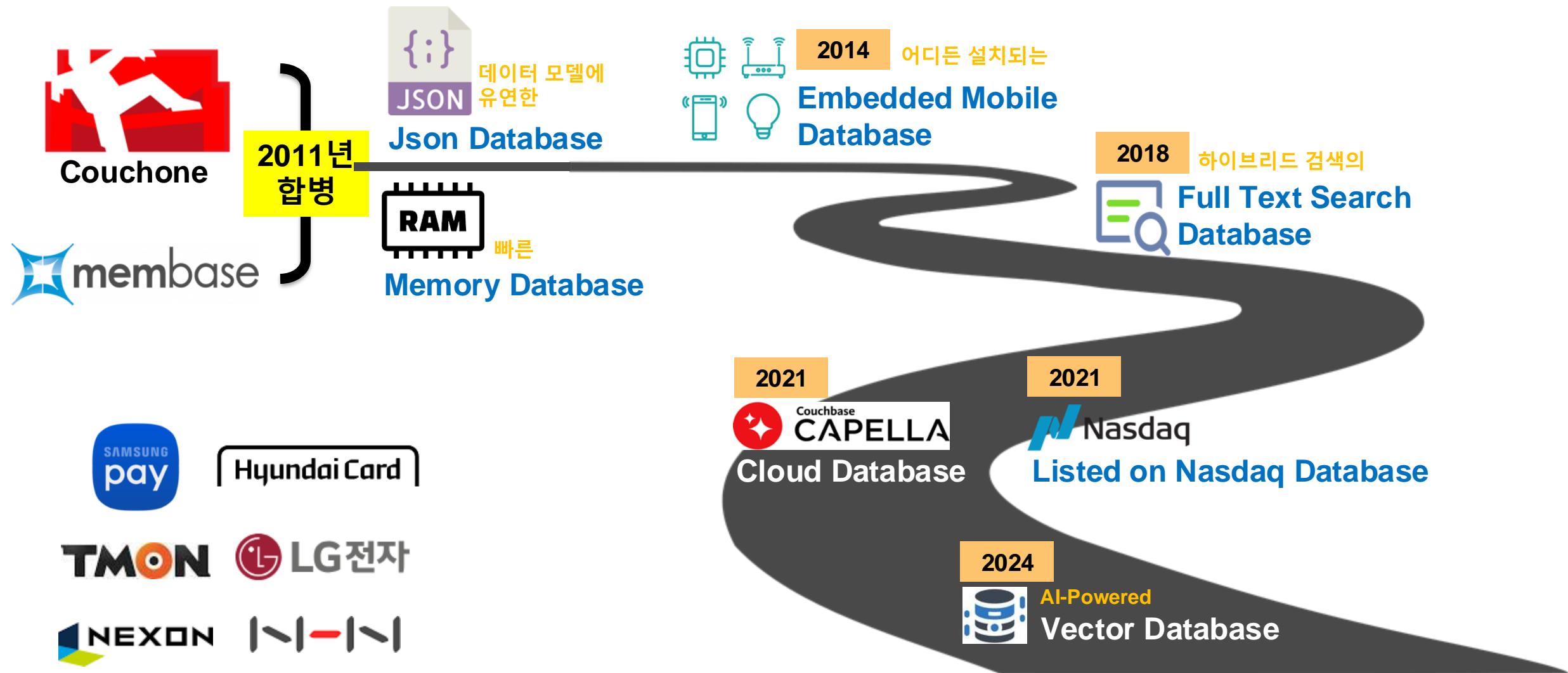
Couchbase Introduction & Architecture

>

T



1 개요 : Couchbase 회사 소개

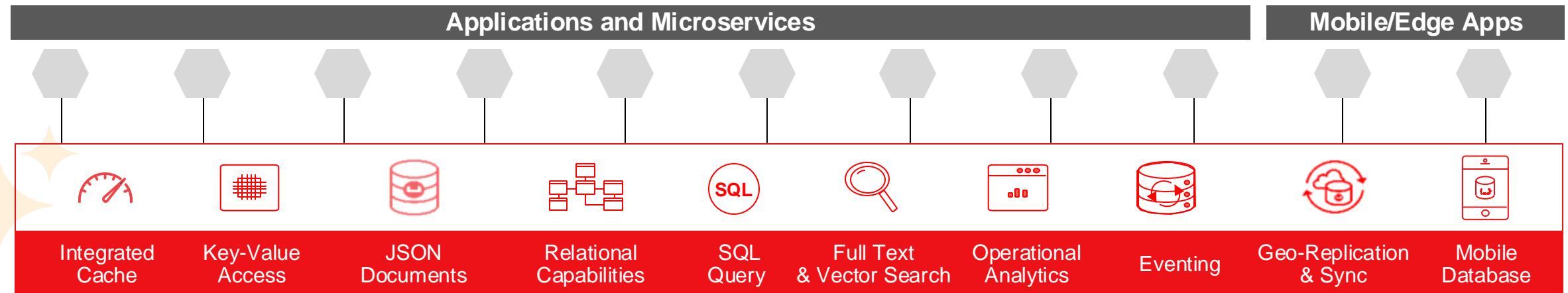
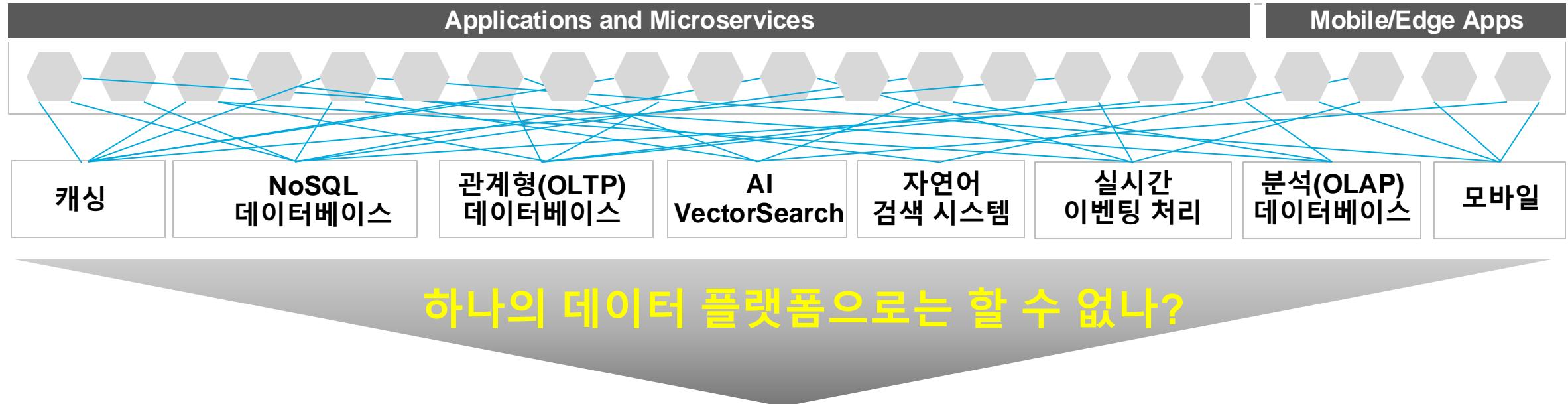


1 개요 : Couchbase 도입 사례

Customers	LinkedIn	TESCO	AMADEUS	COMCAST	UNITED
Application	캐싱 & 싱글뷰를 위한 세션 스토어	리얼 타임 프라이싱, 제품 캐탈로그, 재고관리	비행편 가용성, 예약, 가격분석등	Customer 360 싱글뷰, 'Unified notes'App지원	리얼타임 승무원 분석, 일정 및 리소스 관리
Performance	2백만+ 읽기/초당	1천만+ SKUs	8백만 Ops / 초당	2.1억개 다큐먼트	4.1만 종업원
	1천만 쿼리/초당	3만5천 요청/초당.	<2.5ms 반응시간	10만 사용자	1.5억+ 이용객

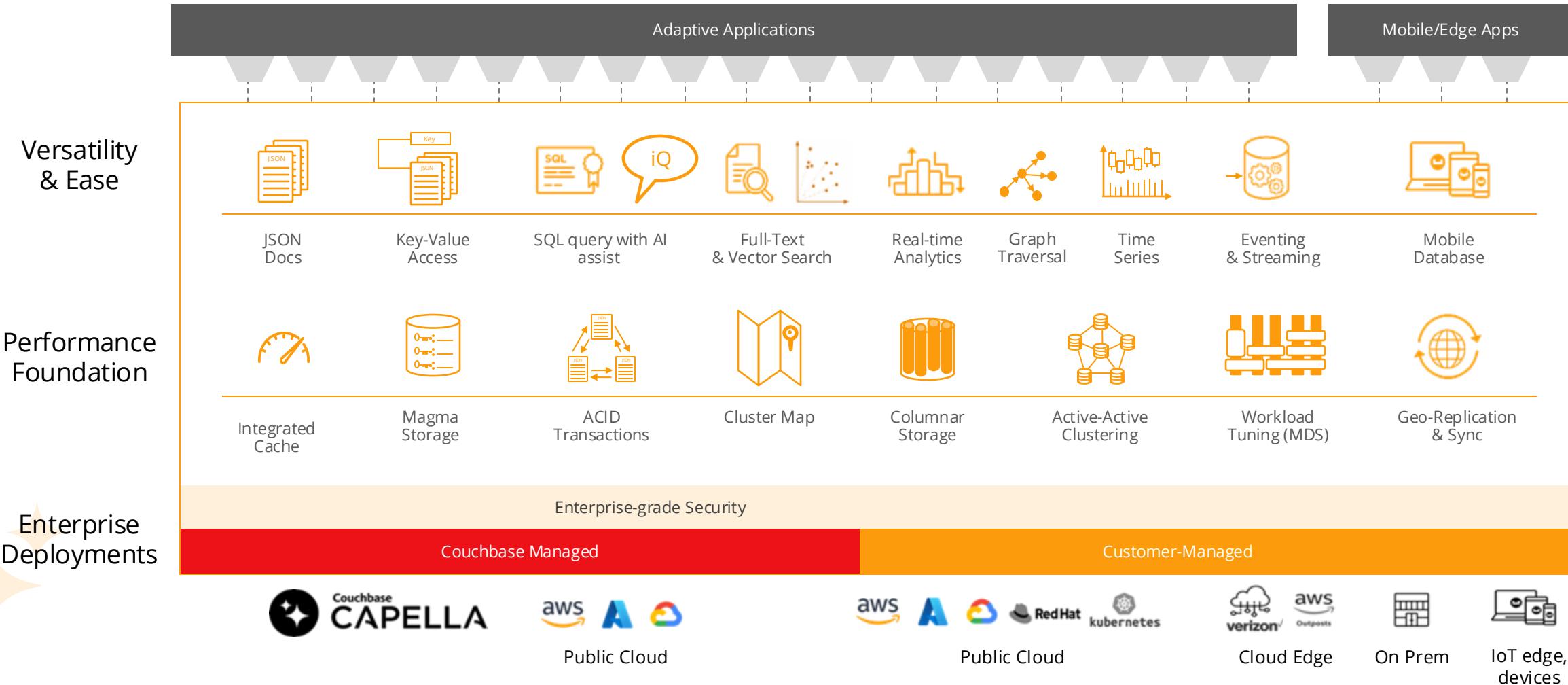


1 개요 : 하나의 App을 만들기 위해 다양한 데이터 관리 솔루션이 필요.

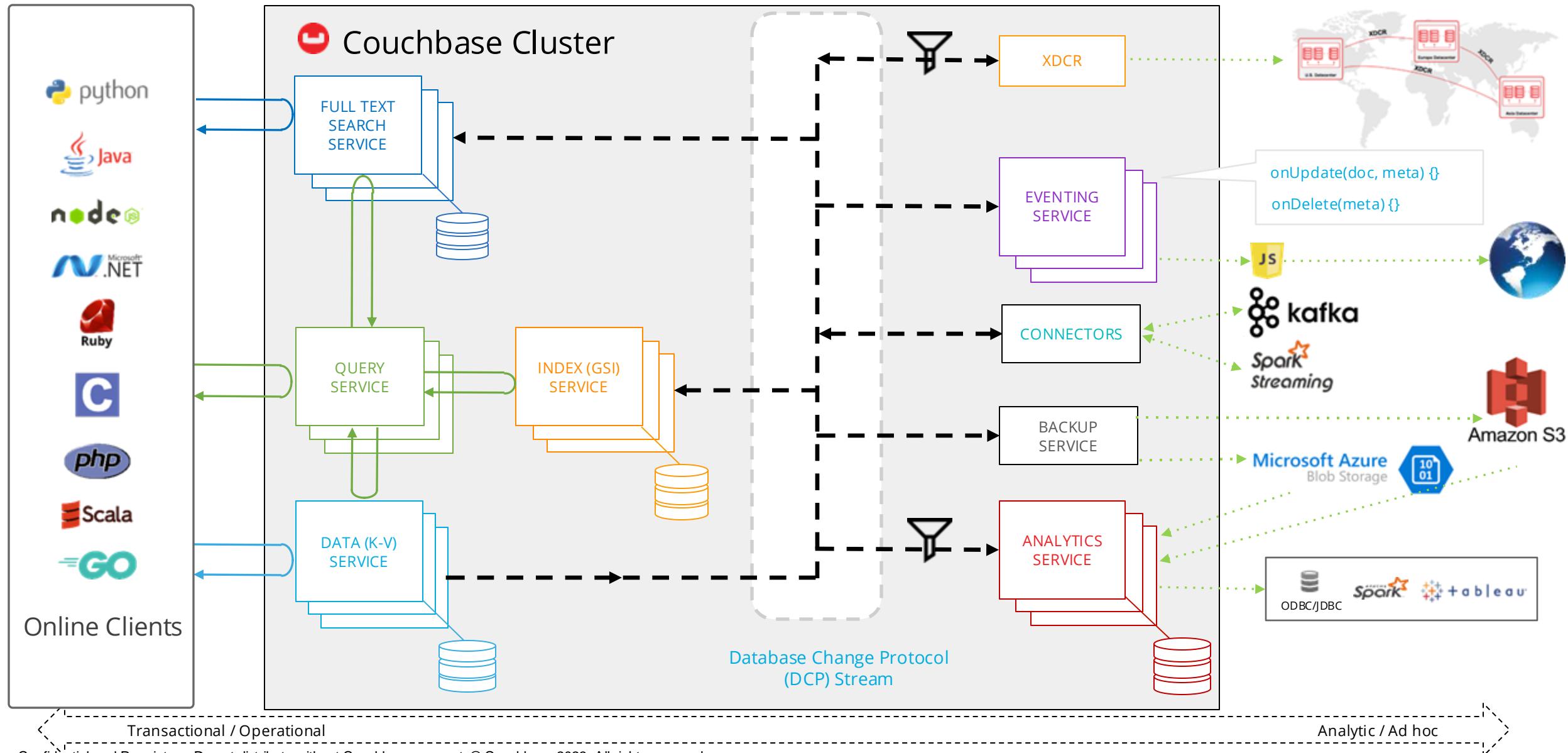


1 개요 : Enterprise 데이터 플랫폼

JSON 도큐먼트 DB, Key-Value 캐시, 표준 SQL, 텍스트 검색, 실시간 분석, 시계열 처리, 고가용성, 재해복구, 모바일 DB 지원하며 AI-Powered 어플리케이션을 위한 데이터 플랫폼입니다.



2 아키텍처 : 메모리 기반 Micro Service 아키텍처



2 아키텍처 : Memory First 아키텍처

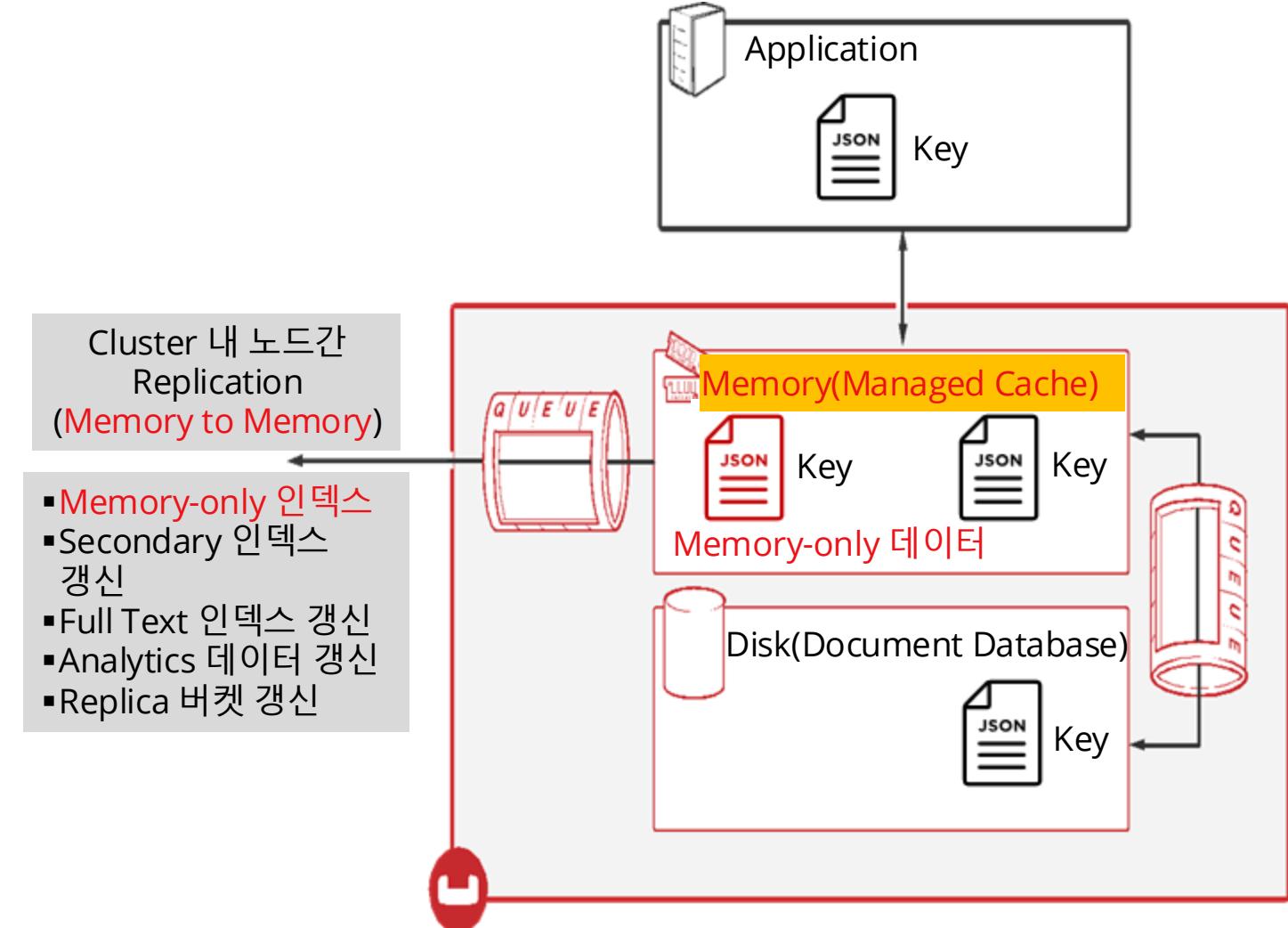
인 메모리 빌트인 캐시를 통해 빠른 Read/Write 업무를 수행하고 데이터 분산 관련 작업도 메모리 기반 프로토콜 사용

- **인메모리 Key-Value 오퍼레이션**

- 특정 Key를 기준으로 데이터를 인메모리에서 처리하는 메카니즘
- 대부분 도큐먼트 데이터베이스는 Read 성능 향상을 위해 별도 솔루션으로 적용

- **Couchbase**

- 인메모리 Key-Value 오퍼레이션의 장점을 구현한 빌트인 캐시 제공
- Value가 단순 수치나 배열이 아닌 JSON 도큐먼트 자체
- JSON 도큐먼트 처리가 메모리 우선 방식



2 아키텍처 : 분산 병렬, Master Node-less 아키텍처

데이터를 다수의 노드에 Key 기반 자동 분산 저장하며, 별도의 마스트 노드없이 모든 노드에서 병렬 처리를 수행함.

• Key 기반 자동 분산 아키텍처

- 대량의 Key-Value 처리를 노드 별로 분산하여 성능 향상
- 최대한 균등하게 분산 저장 가능, 특정 노드에 편중되는 현상(Data Skew) 방지
- 별도의 분산 정책 불필요

• Couchbase

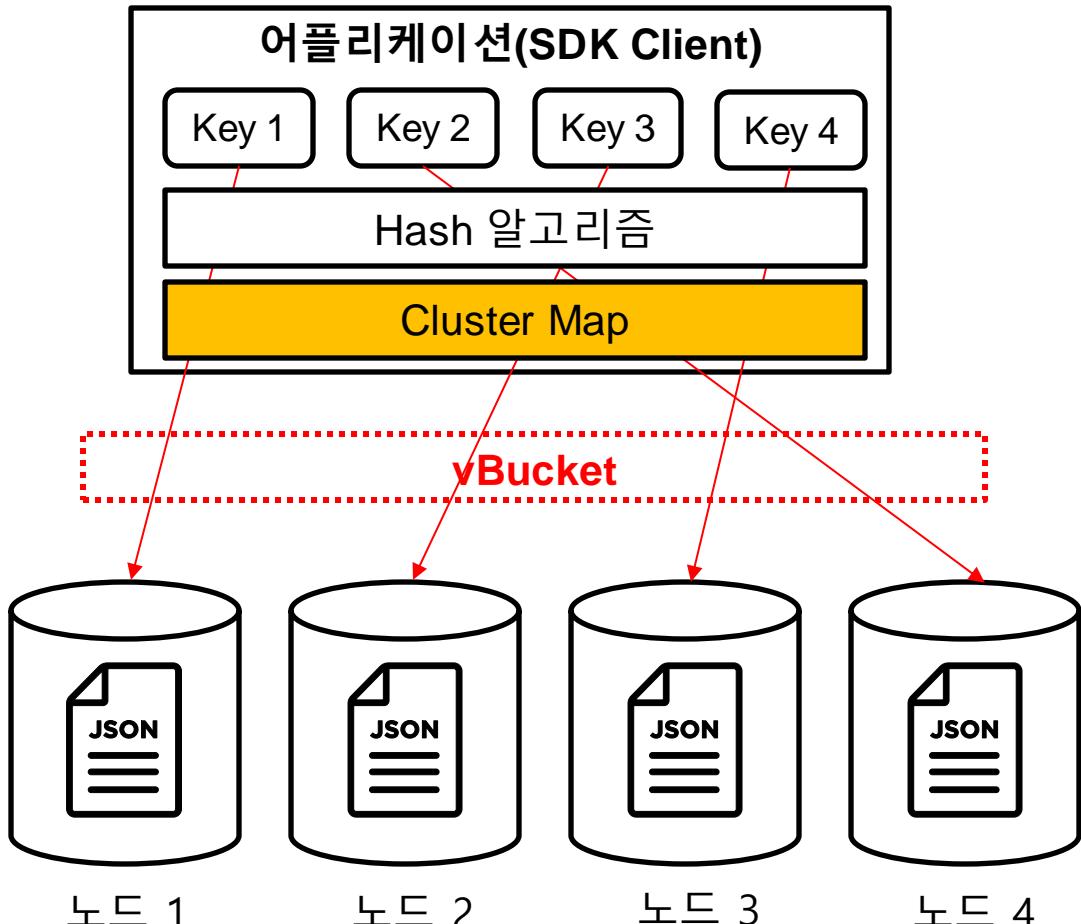
- Key에 대한 Hash 알고리즘 적용으로 자동 분산
- 노드 추가 시, Rebalancing을 통해 Key 재 분산 수행

• Master Node-less 아키텍처

- 어플리케이션의 Key-Value Operation 시, 해당 Key에 맵핑된 특정 노드에 직접 접근
- 모든 노드가 어플리케이션 측면에서 Active 노드 역할

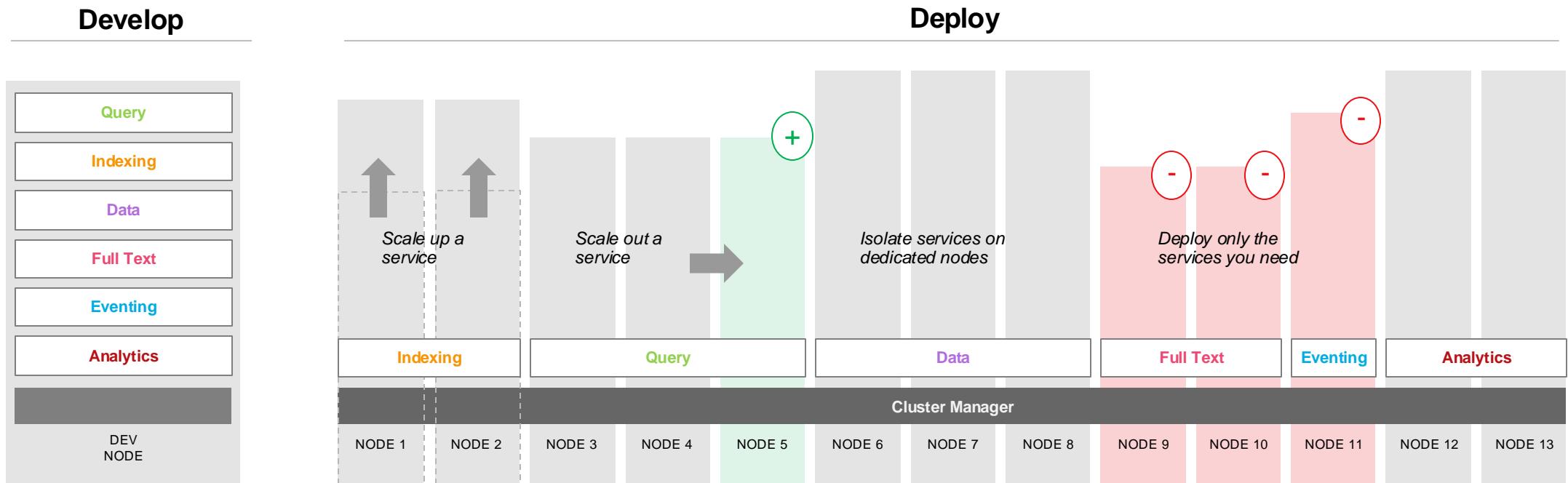
• Couchbase

- 어플리케이션이 데이터 처리를 구현하기 위해 SDK 활용
- 데이터 분산 정보(Cluster Map)를 지속적으로 SDK Client에 Update



2 아키텍처 : 자원 절약형 다차원 독립 확장

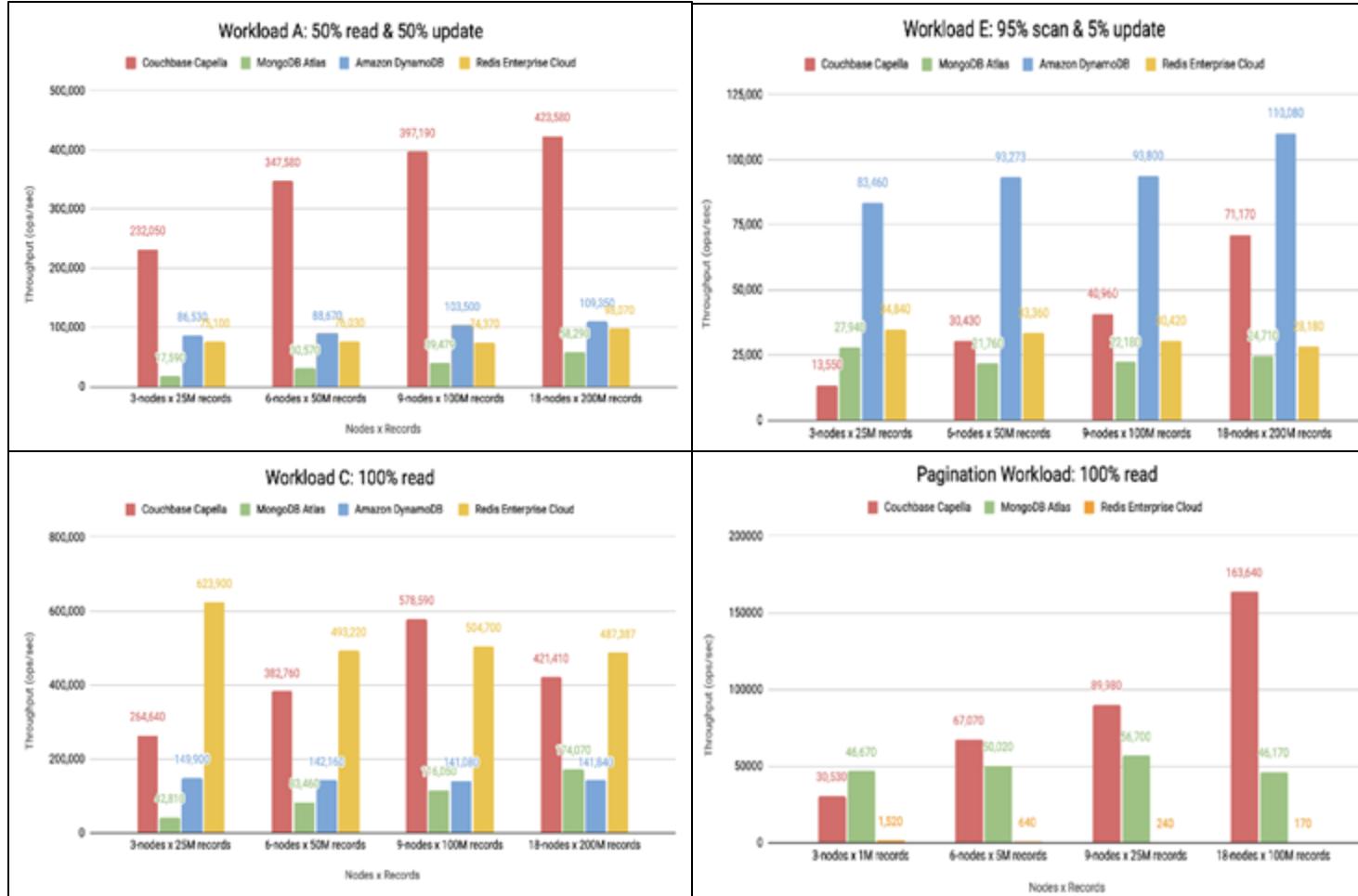
Couchbase는 **Multi-Dimensional Scaling** 기능으로 서비스 별 Workload 분산 및 독립성을 보장합니다.



- **Service 단위 하드웨어 자원 최적화**
 - 각 Service 별 시스템 자원을 독립적으로 할당
 - **각 Service에서 수행되는 작업이 다른 Service에 영향을 최소화**, 예를 들어 Analytics Service에서 복잡한 작업을 수행하여 시스템 자원을 많이 사용해도 Data Service 혹은 Query Service에서 수행하는 Operational 작업에는 영향이 없는 구조

3 성능 : NoSQL Database 벤치마크에서 탁월한 우위

<Yahoo! Cloud Serving Benchmark(YCSB-NoSQL Benchmark)>



출처 : <https://www.altoros.com/blog/couchbase-capella-vs-mongodb-atlas-vs-amazon-dynamodb-vs-redis-enterprise-cloud/>

Altoros, April 2023

MongoDB

- Struggles to scale, strongest at 3 nodes
- Price performance makes it worse
- Weakest performer

DynamoDB

- Excels at scans, but throws excessive errors
- Challenged across other workloads

Redis

- Excels at read-only, Capella meets at scale
- Regularly used as cache for Atlas or DynamoDB
- “Fails” Pagination workload

Capella

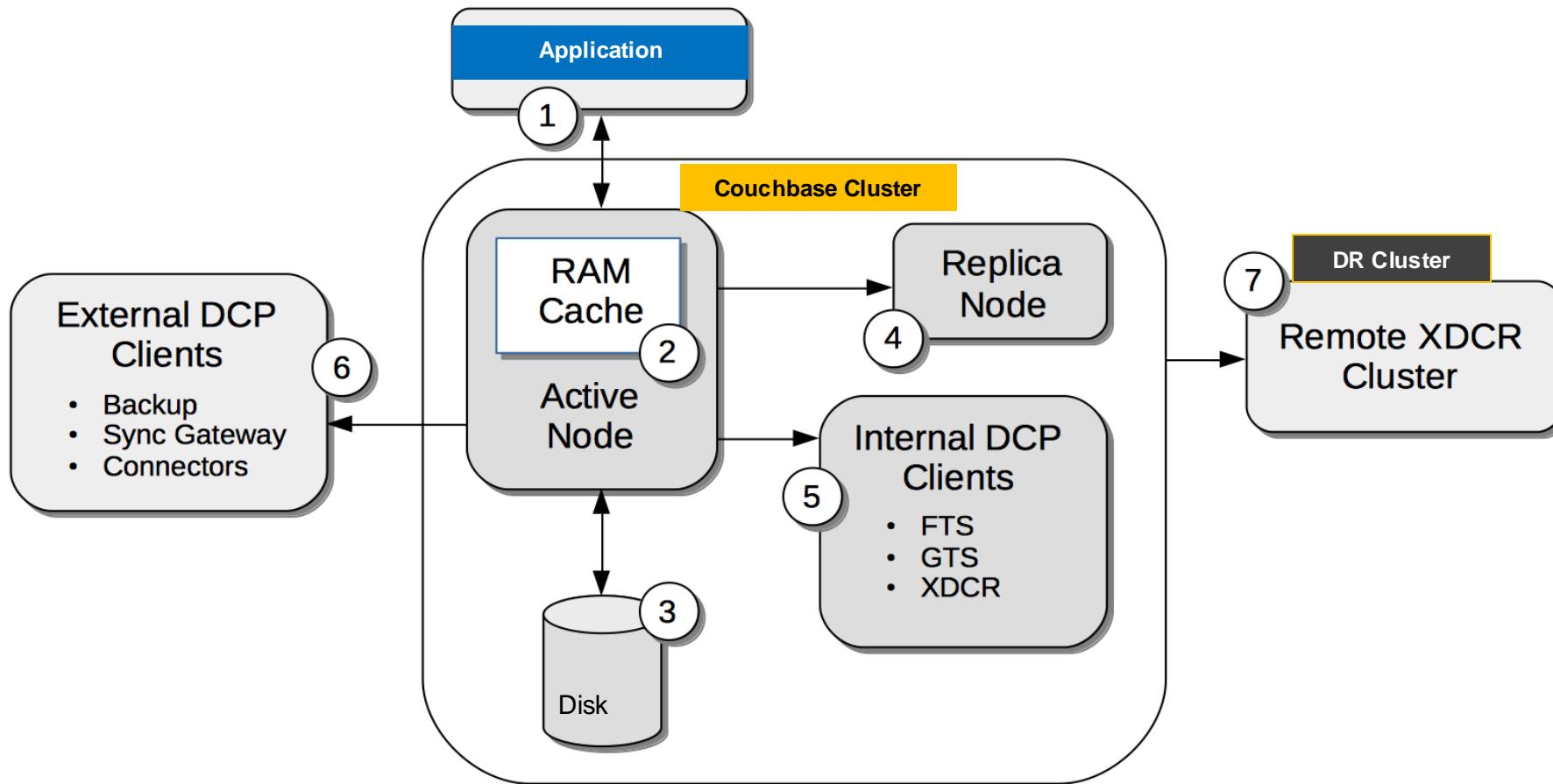
- Excels at Read and Write
- Scales effectively for multiple workloads
- Best price-performance

	Couchbase Capella	MongoDB Atlas	Azure CosmosDB\ (MongoDB API)
DEPLOYMENT	90 /116	84 /116	71 /116
MANAGEMENT	28 /32	27 /32	21 /32
SUPPORT	19 /25	19 /25	19 /25
PERFORMANCE	11 /16	12 /16	9 /16
PRICING	20 /26	13 /26	17 /26
	12 /17	13 /17	5 /17

출처 : <https://benchant.com/navigator/dbaas>

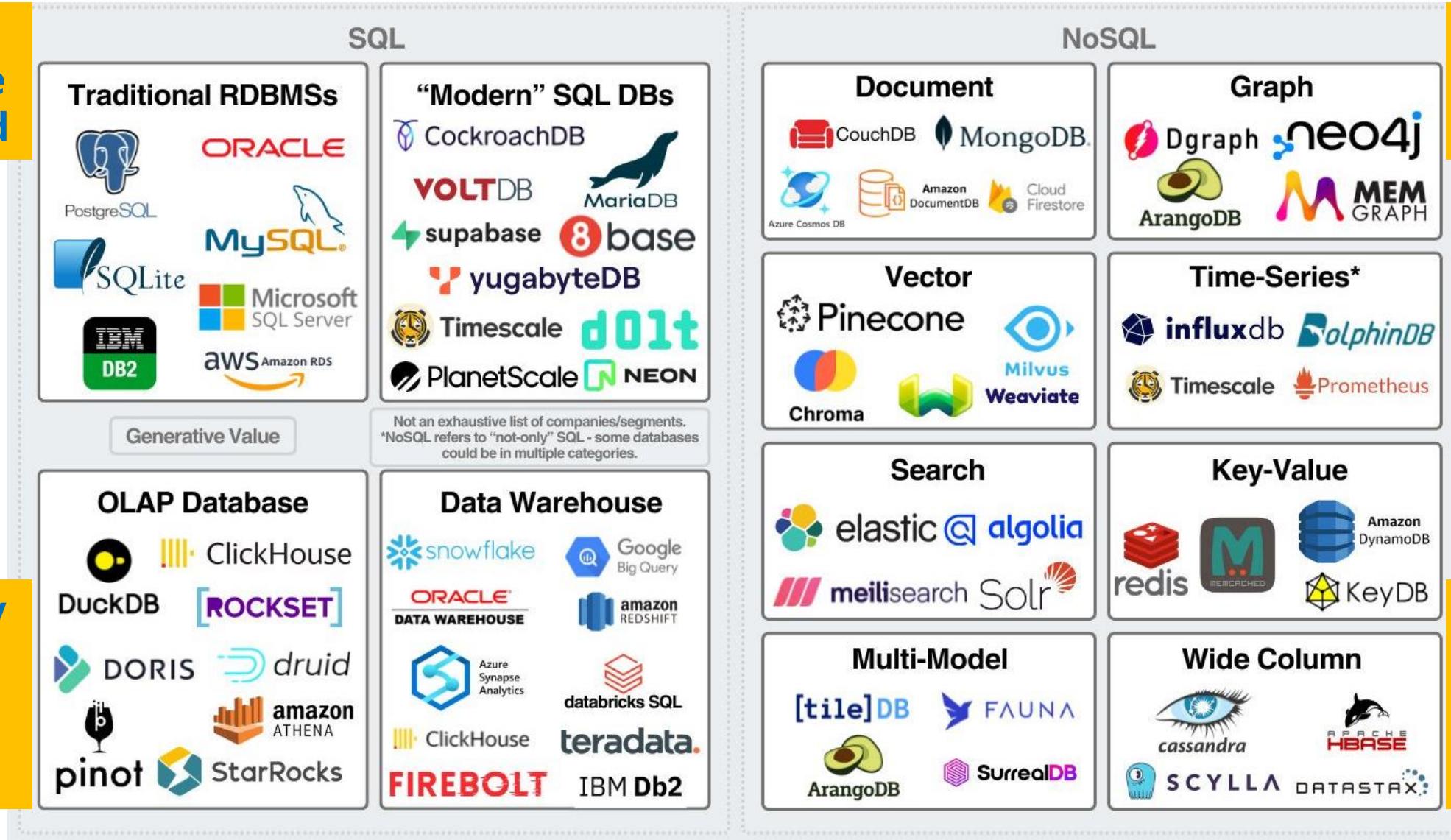
3 성능 : End-to-End 압축으로 네트워크/스토리지 비용 절감

Couchbase는 데이터를 압축하여 전송하며 압축된 형태로 메모리와 디스크에 보관됩니다. 즉, 네트워크 전송량, 메모리 및 디스크 사용량을 줄려 성능 및 운영비용을 절감할 수 있습니다.



4 데이터 모델 : SQL(Table)과 NoSQL(Real World)

1970
Machine
Oriented

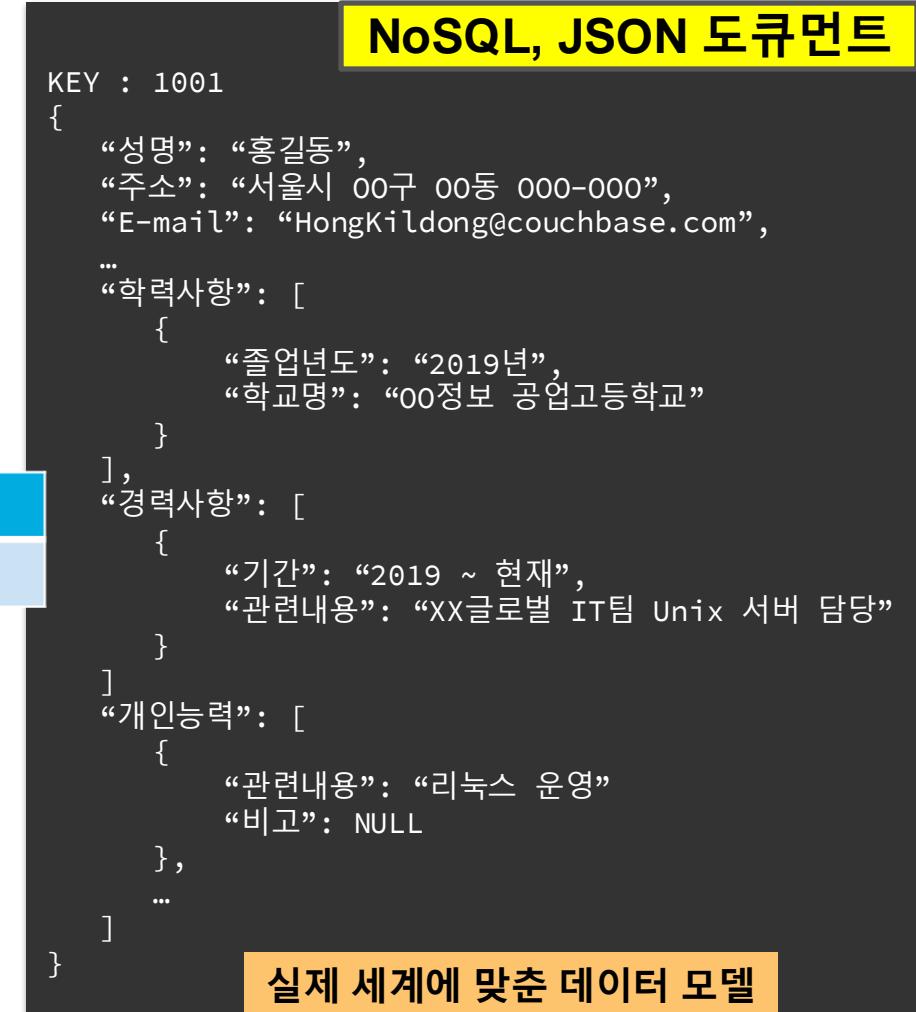
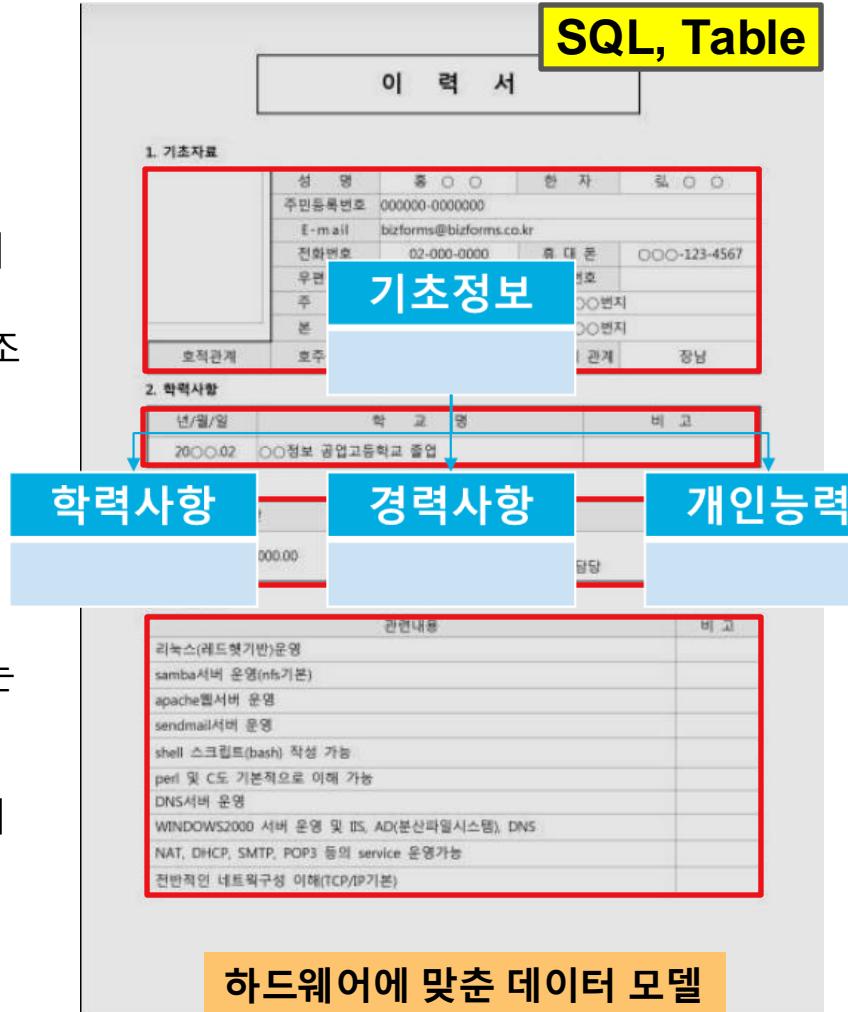


Assembly
Cobol
C
Pascal
Perl

4 데이터 모델 : JSON 도큐먼트

JSON은 텍스트로 이루어져 있으므로, 사람과 기계 모두 읽고 쓰기 쉽다. 프로그래밍 언어와 플랫폼에 독립적이므로, 서로 다른 시스템 간에 객체를 교환하기에 좋다.

- JSON 도큐먼트의 장점
 - 단일 도큐먼트 내에 다양한 정보를 계층 구조를 활용하여 저장
 - 정보 추가/삭제가 유연한 구조 제공
 - 데이터 전달을 위한 표준 인터페이스 역할
- RDB와 차별점
 - 여러 테이블로 분리, 저장되는 데이터를 단일 도큐먼트에 저장
 - 테이블 간 조인을 최소화하여 데이터 처리 속도 향상

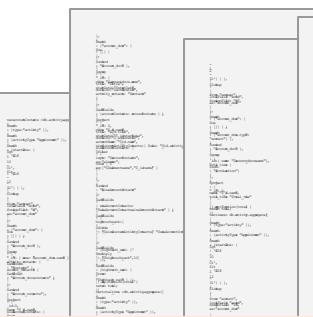


4 데이터 모델 : ANSI 표준 SQL vs. 자체 Query 언어

2018년 Q4 고객에 대해서 각 미팅에 소요된 시간, 해당 고객에 모든 미팅에 소요된 시간 대비 해당 미팅의 소요 시간 비율, 순위 등을 조회하는 Query에 대해 Couchbase와 타 NoSQL 솔루션 비교입니다.

```
SELECT
    c.name,
    a.title,
    a.actduration,
    a.startDate,
    SUM(a.actduration) OVER ( PARTITION BY
        c.name ORDER BY c.name, a.startDate )running_total,
    TRUNC(100*(a.actduration/SUM(a.actduration)
        OVER(PARTITION BY c.name)) ) pct_of_total_time,
    RANK()OVER(PARTITION BY c.name ORDER BY
        (ARRAY_COUNT(a.contacts)/
        ARRAY_COUNT(c.contacts))DESC)hightouch_rank
FROM activity a
    INNER JOIN account c
        ON (a.accid = c.id)
WHERE a.activityType = 'Appointment'
    AND a.startDate BETWEEN '2018-10' AND '2018-12'
GROUP BY c.name,a.title,a.startDate,
        a.actduration,a.contacts,c.contacts
ORDER BY c.name,a.startDate
```

21 lines vs 347 lines



```
db.activity.aggregate([
    { $match : { type: "activity" } },
    { $match : { activityType:"Appointment" } },
    { $match : { startDate:{$gt:'2018-10-01',
                     $lt:'2018-12-31' }}},
    { $lookup: {
        from: "account",
        localField: "accid",
        foreignField: "id",
        as: "account_docs"} },
    { $match : { "account_docs": {$ne: []} } },
    { $unwind: "$account_docs" },
    { $group : { "_id": {
        name: "$account_docs.name",
        title: "$title",
        startDate: "$startDate",
        duration: "$actduration",
        activity_contacts: "$contacts",
        account_contacts: "$account_docs.contacts",
        }}},
    { $addFields: { total_time: total_time } },
    { $addFields: { hightouch_rank: rank_temp } },
    { $addFields:{ running_total: running_total}},
    { $project: {
        "_id": 0,
        name: "$_id.name",
        title: "$_id.title",
        startDate: "$_id.startDate",
        duration: "$_id.duration",
        activity_contacts:
            "$_id.activity_contacts",
        account_contacts:.....}}
```

• Couchbase

- ANSI 기반 SQL
- Join, Windows Function 등 지원

4 데이터 모델 : ANSI 표준 SQL++

- **개요**
 - ANSI 2003 SQL 기반
 - Multi-core 병렬 수행에 최적화된 Query Engine
 - Cost-based Optimizer 지원
- **확장 SQL**
 - NEST/UNNEST : Embedded Object, Arrays 지원
 - IS EMPTY/IS MISSING : Flexible Schema
- **SQL-Like**
 - 99 % 표준 SQL과 동일한 Syntax
 - DML 지원 : Insert/Select/Update/Delete/Upsert
 - **INNER/OUTER JOIN** 지원



4 데이터 모델 : 논리 / 물리 모델

- RDBMS와 유사한 구조의 논리 계층 구조로 구성하여 편리한 데이터 관리
- Data 서비스를 완전 메모리DB로 사용도 가능하며 용도에 따라 물리 저장 방식을 선택할 수 있음

RDBMS	Couchbase
Server	Cluster
Database	Bucket
Schema	Scope
Table	Collection
Row	Document (JSON)
Value	Sub-Document, Array

Feature	Ephemeral Bucket	Couchbase Bucket	Magma Bucket
Bucket memory quota (per node)	Min 256MB	Min 256MB	Min 1024MB
Max Object Size	20MB	20MB	20MB
Persistence	no	yes	yes
Replication and XDCR	yes	yes	yes
Encrypted data access	yes	yes	yes
Rebalance	yes	yes	yes
N1QL, Search, Analytics, Eventing	yes	yes	yes
Indexing	yes	yes	yes
Backup	yes	yes	yes

4 데이터 모델 : Time-Series

Time Series Format

- JSON Array를 사용하여 대용량 데이터를 효율적으로 저장
- ts_start, ts_end, ts_interval, ts_data 규격에 맞게 데이터 구조화 필요

데이터 활용

- _timeseries function을 통해 테이블 형태의 데이터로 전환 가능
- 인덱스 하나로 모든 Query 대응
- 일반 JSON 도큐먼트 대비 사이즈 축소를 통해 이력 데이터 용도로 활용

```
{  
    "equip_id": "1",  
    "lot_id": "1",  
    "ts_start": 1375228800000,  
    "ts_end": 1372636800000,  
    "ts_interval": 100  
    "ts_data": [  
        [10, 27], [10, 27], [10, 27], [10, 27], [10, 30], [10, 30], [10, 30],  
        [10, 30], [10, 30], [10, 30], [10, 30], [10, 30], [10, 30], [10, 30], [10, 30],  
        [10, 30], [10, 30], [10, 30], [10, 30], [10, 30], [10, 30], [10, 30], [10, 30],  
        [10, 23], [10, 23], [10, 23], [10, 23], [10, 23], [10, 23], [10, 23], [10, 23],  
        [10, 23], [10, 23], [10, 23]  
    ]  
}
```

```
SELECT t.*  
FROM Raw_Collection AS d  
UNNEST _timeseries(d, {"ts_ranges":$ts_ranges}) AS t  
WHERE d.equip_id = '1'  
AND (d.ts_start <= $ts_ranges[1]  
AND d.ts_end >= $ts_ranges[0]);
```

5 서비스 : Analytics > Hybrid Transaction Analytical Processing

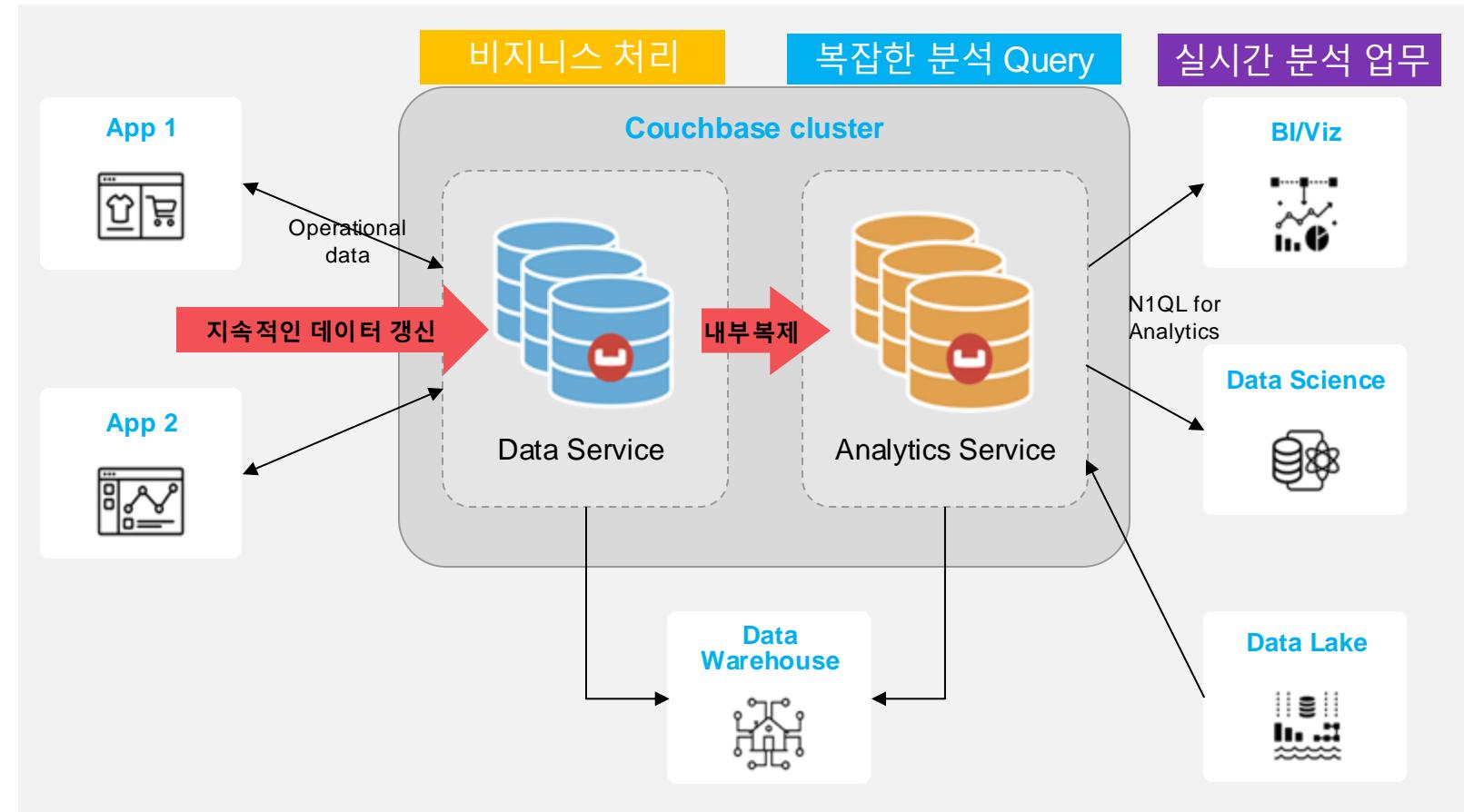
- 데이터를 별도의 시스템으로 추출-변환-로드할 필요 없이 거의 실시간으로 JSON 데이터를 분석
- 운영 데이터나 쿼리 속도를 늦추지 않고 Analytics용 SQL++을 사용하여 Analytics 데이터를 쿼리

• Shadow Copy

- Data Service에 처리되는 데이터를 그대로 Analytics Service로 복제
- 메모리 기반 DCP 사용
- 별도의 CDC/ETL 솔루션 불필요

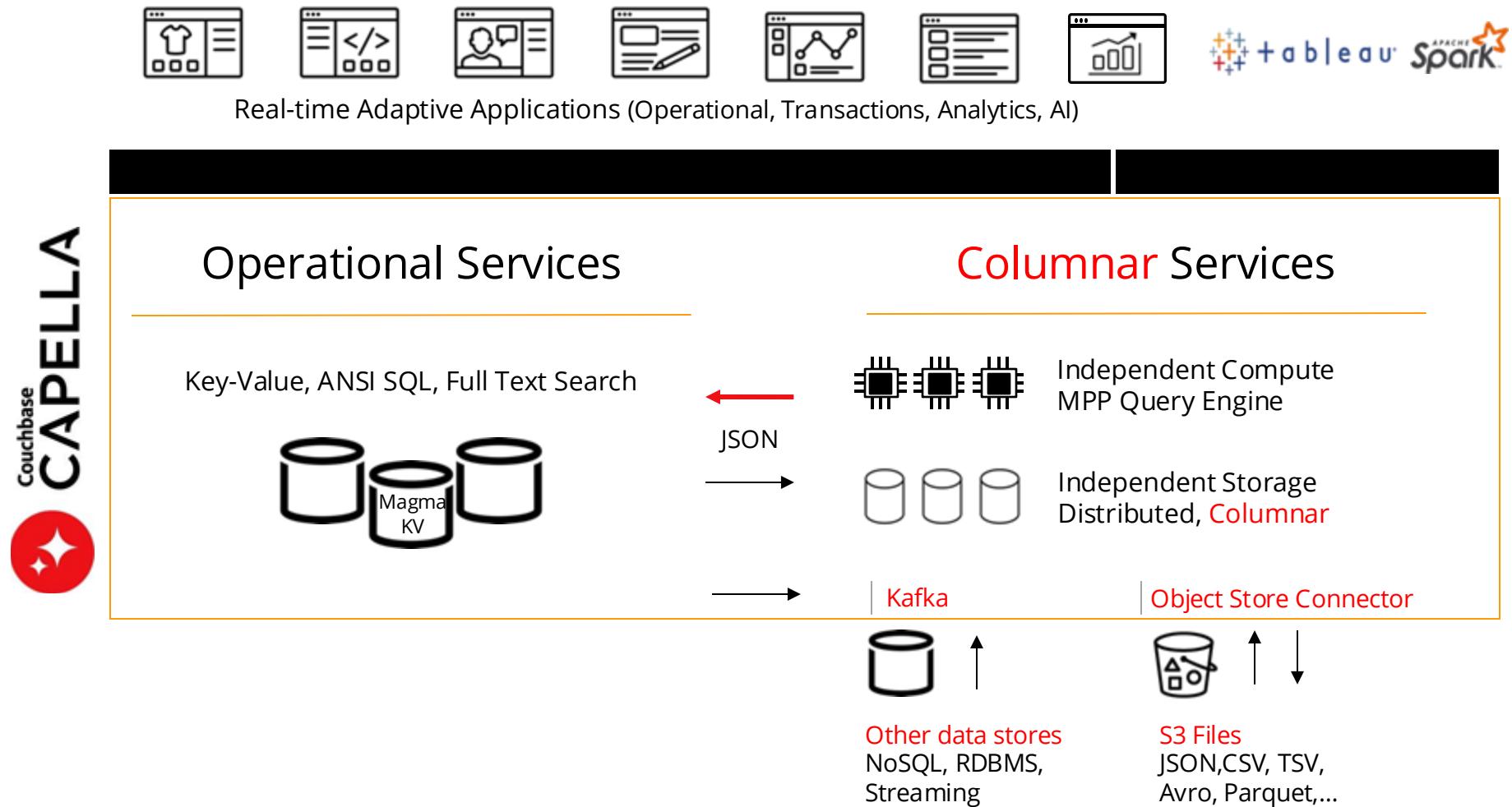
• Parallel Processing

- SQL을 Analytics Service를 구성하는 노드에 분산하여 병렬 처리
- 복잡한 Query 및 ad hoc Query 수행에 적합
- 대용량 데이터 처리
- Tableau Native Connector 제공
- Power BI Native Connector 제공



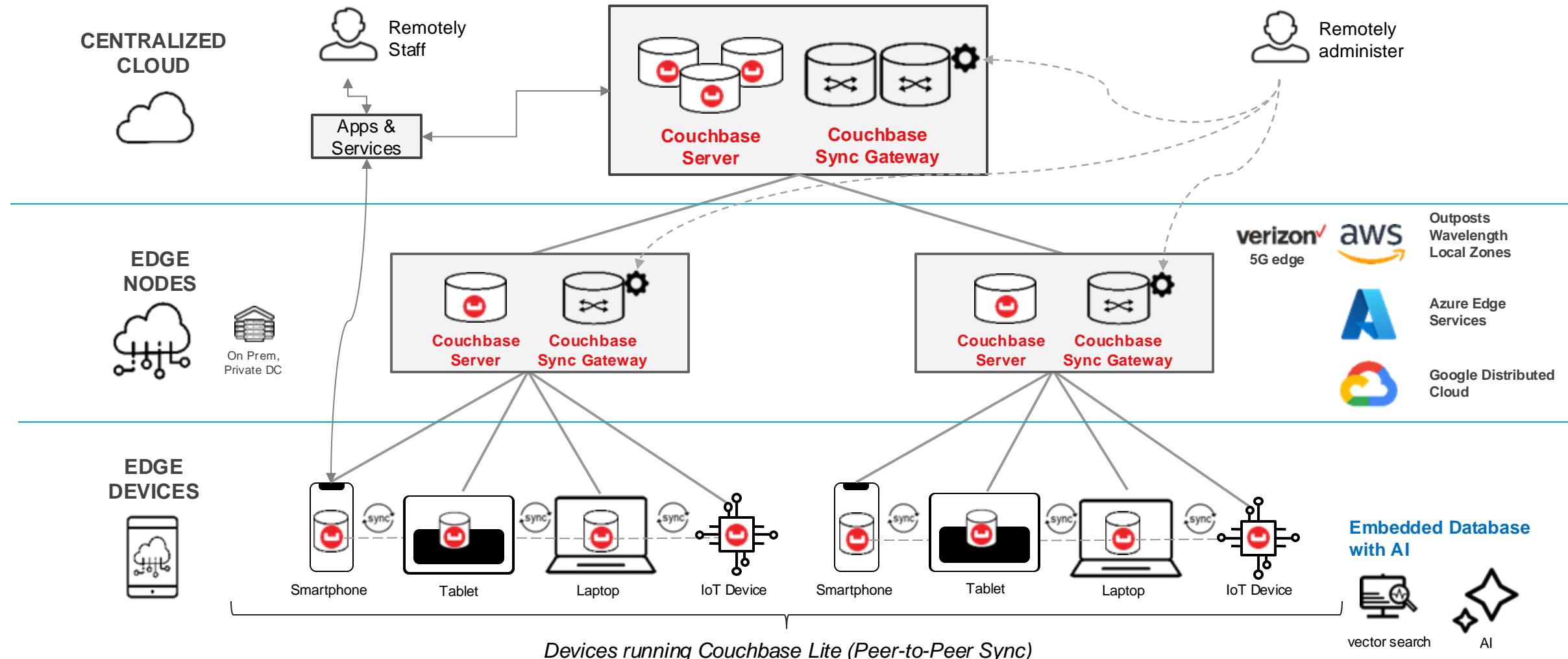
5 서비스 : Capella Columnar Service : Cloud Data Lake

내부 JSON 데이터를 비롯하여 다양한 외부 데이터 소스를 통합하여 분석을 수행하는 컬럼 기반 Data Lake 기능 출시



5 서비스 : Mobile (Couchbase Lite, Sync Gateway)

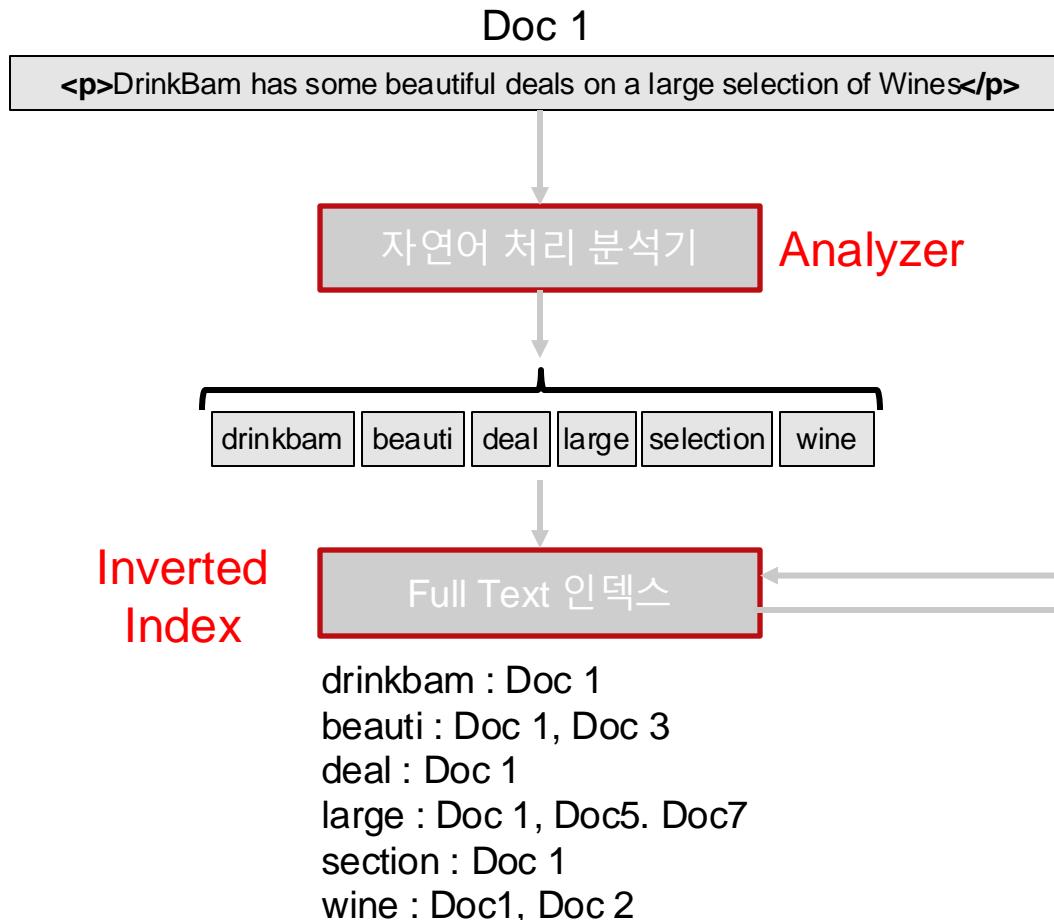
데이터 관리가 필요한 모든 디바이스에 데이터베이스를 적용할 수 있으며, 데이터 센터의 데이터베이스와 손쉬운 일관성을 유지



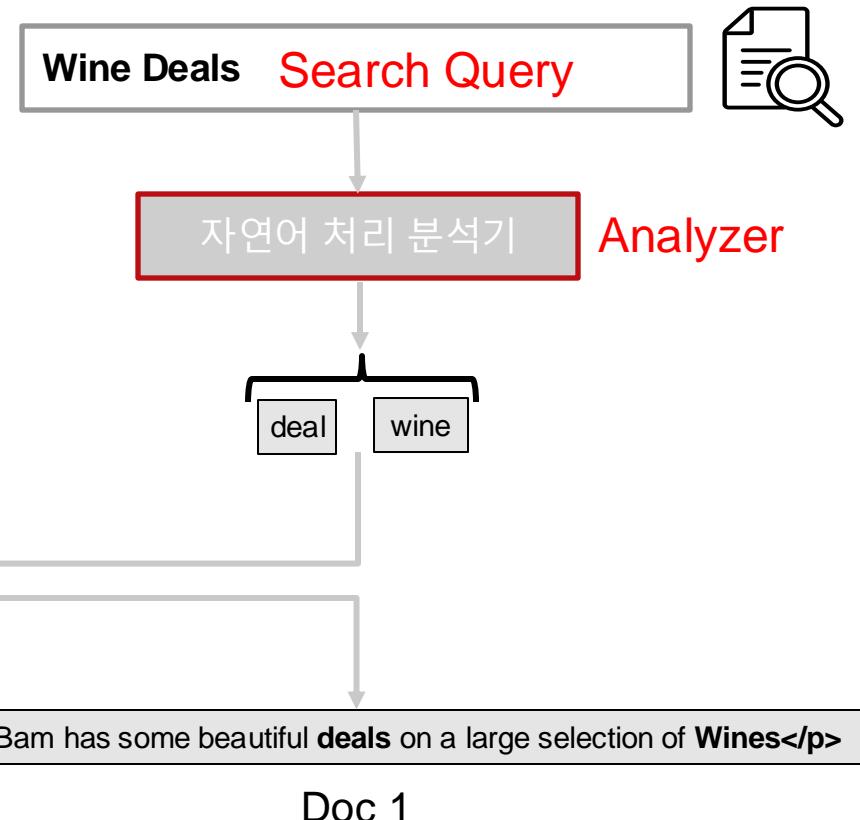
5 서비스 : Full Text Search(자연어 검색)

Data와 Query 서비스에 통합하여 검색 서비스 제공

<1. 검색을 위한 인덱스 구성>



<2. 검색어를 통한 도큐먼트 검색>



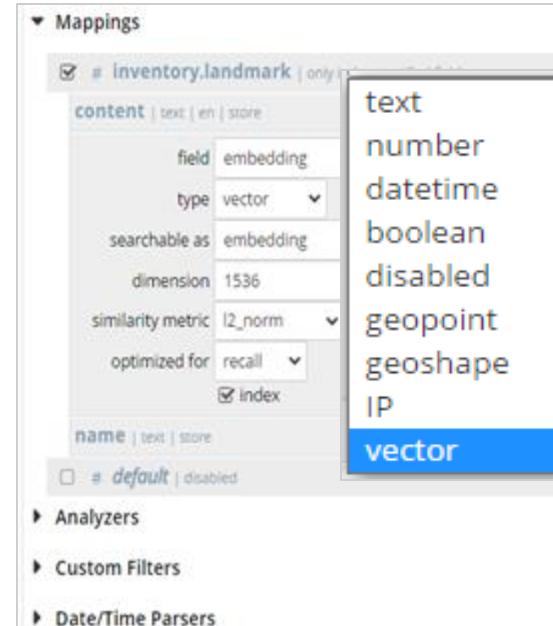
5 서비스 : Full Text Search with AI 벡터 검색

SQL 쿼리에 Full Text 검색 쿼리와 벡터 검색 쿼리를 통합 가능

JSON Storage

```
{ "type": "shoes",
  "productId": "CP123456",
  "category": "Gym Shoes",
  "name": "Beach Sneakers",
  "brand": "Ultimate Surf",
},
"description": "The ultimate companion for beach adventurers, designed to seamlessly transition from sandy shores to urban landscapes. This innovative sneaker features a water-resistant, quick-drying mesh upper, allowing your feet to breathe while keeping them dry.",
"descriptionVector": [0.131, 0.339, -0.611, 0.981,...]
```

Indexes



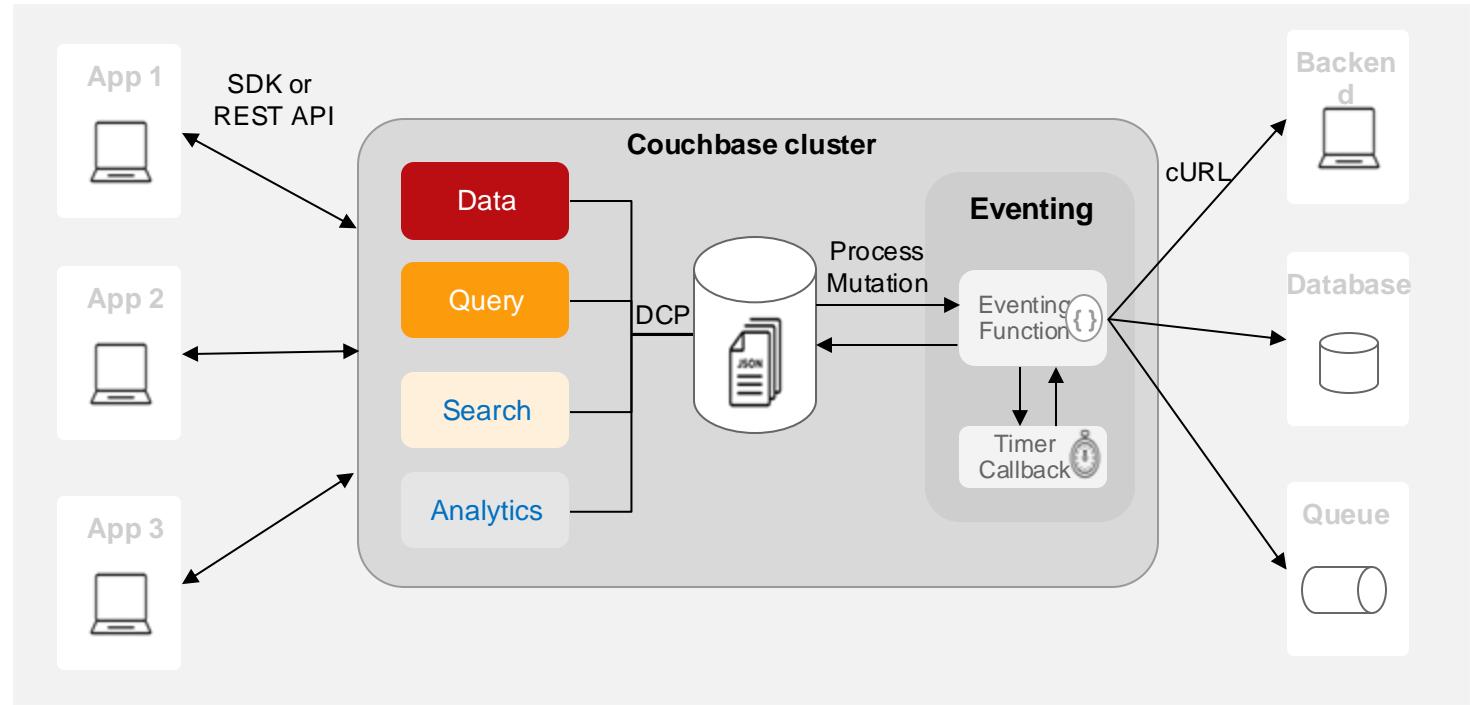
SQL + 자연어 검색 + AI Vector Search

```
SELECT meta().id, t.name
FROM `product` AS t
WHERE
SEARCH(t,
{"query": {"match_phrase": "sneaker", "field": "description"}})
ORDER BY search_score() DESC
LIMIT 10;
```

```
SELECT *
FROM product
WHERE LOWER(product.type) = 'shoes'
AND product.size = 11
AND product.price between 50 and 80
/* desc SIMILAR TO 'blue running shoes' */
ORDER BY GSI_VECTOR_ORDER(desc_embedding, {
  "knn": [
    "field": "desc_embedding",
    "vector": [0.1, 0.334, -0.604, 0.985] ]})
LIMIT 4
```

5 서비스 : Eventing

- **개요**
 - Event-Condition-Action Model
 - JavaScript 기반 : JSON Document의 변경을 분석, 처리에 유리
- **Handler Type : DB Trigger와 유사**
 - onUpdate
 - onDelete
- **주요 Use case**
 - 실시간 Document Enrichment
 - Threshold 기반 Altering
 - Streaming 처리
 - 데이터 변경에 대한 확산 작업
 - 데이터 Cleansing



- **비즈니스 로직을 중앙 집중적 관리**
 - Document 데이터 변경에 실시간으로 Trigger되는 Business Logic 구현
 - Java Script 기반으로 Cluster 내 Key-Value operation, SQL++ Query 지원
 - cURL function을 통해 외부 REST API 호출 가능

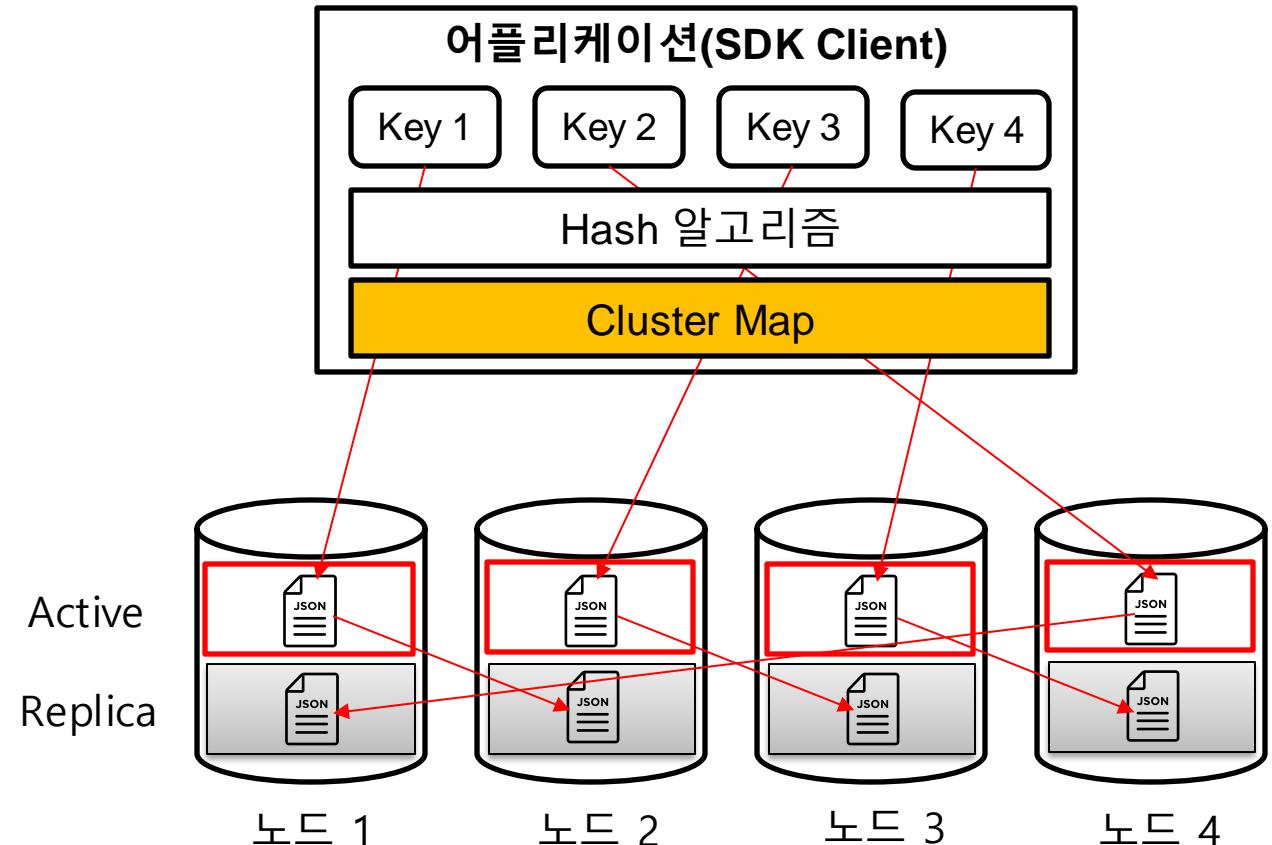
6 가용성 : 빠른 Fail-over를 통한 고 가용성

- **Replica 버켓**

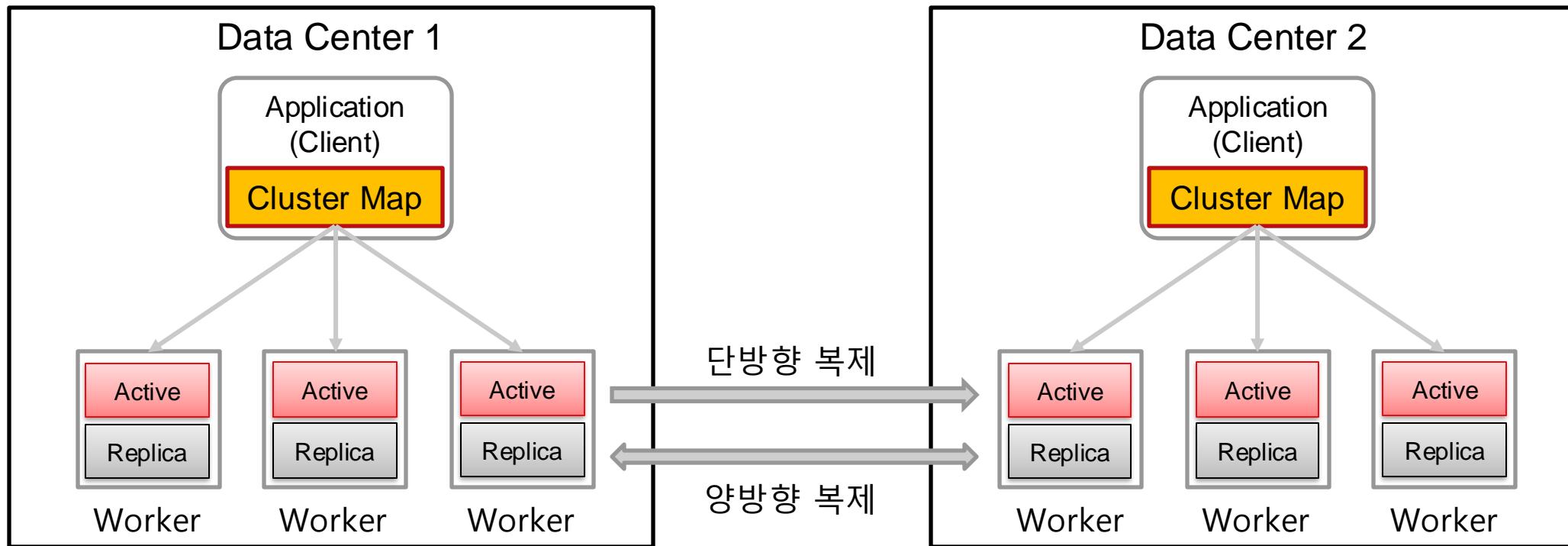
- Auto-sharding을 통한 균등한 데이터 분산
- Active 버켓에 대한 Replica 버켓을 내부 노드 단위에서 관리
- Replica 버켓을 최대 3 까지 가능

- **Couchbase**

- 특정 노드의 장애가 발생하면 장애가 발생한 Active 버켓의 Replica 버켓을 Active 상태로 전환
- 빠른 Fail-over
- 별도의 Replica 노드나 Standby 노드가 필요 없음



6 가용성 : 클러스터 간 Native 복제를 통한 재해 복구

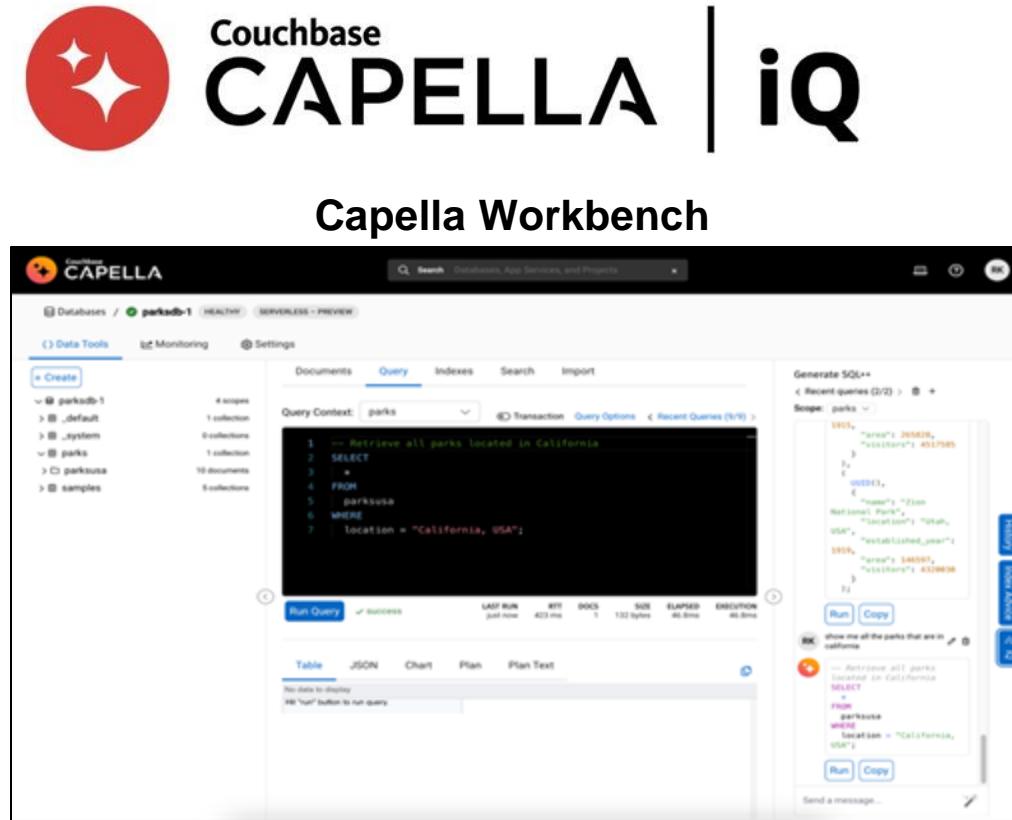


- **XDCR(Cross Data Center Replication)**

- XDCR을 통해 멀티 데이터 센터에 위치한 클러스터 간 데이터 복제
- 단방향 복제 및 양방향 복제 지원
- 복제는 필요한 데이터만 필터링 가능
- 단순 재해 복구 솔루션 이상의 글로벌 워크로드 분산 솔루션

7 개발 편의성 : 생성형 AI 기반 코딩 지원

- Generative AI의 LLM을 활용한 Couchbase Capella 전용 Code Assistant
- 자연어로 SQL 및 소스 코드 코딩 지원
- Couchbase 내부 스키마 정보를 활용하여 실제적인 코딩 지원



8 운영/관리 : 내장 백업복구 솔루션

• Backup Service

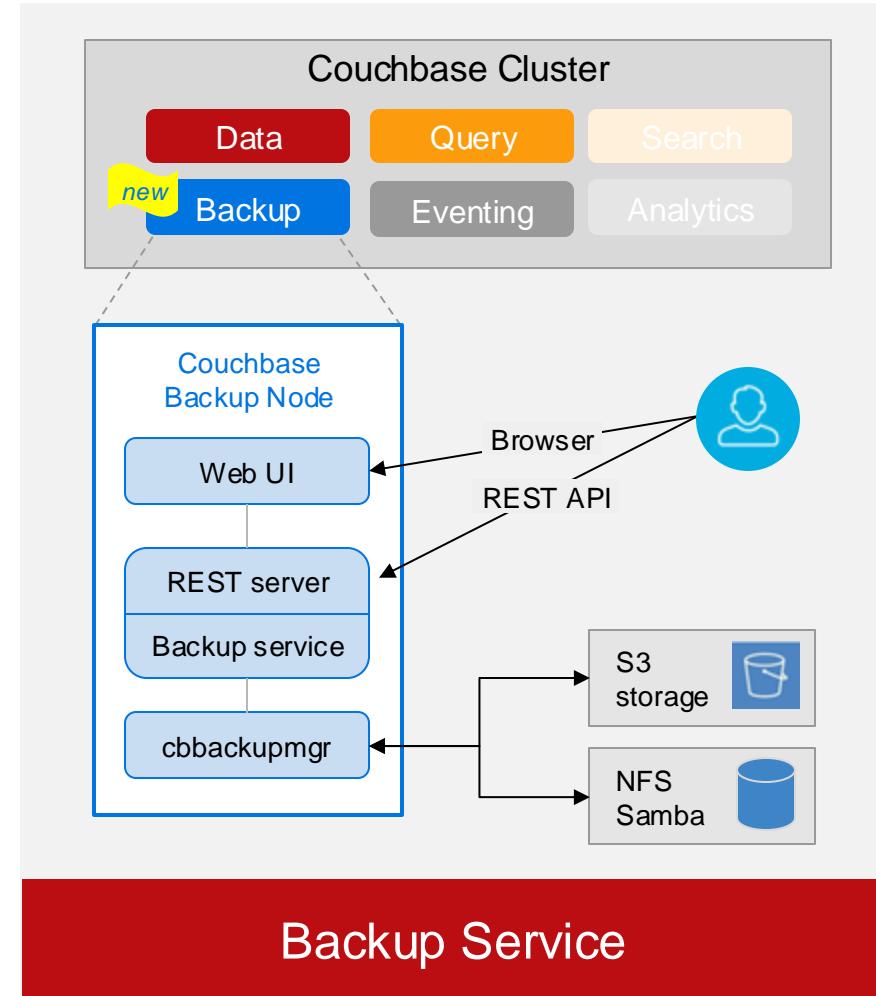
- UI 기반 Backup & Recovery
- 공유 파일 시스템 구성 필수
- Backup Scheduler 제공
- 백업 중 장애를 위한 Resume 기능 제공
- 병렬 백업
- 암호화 저장

• Backup 방식

- Full Backup
- Incremental Backup
- Merge Backup File

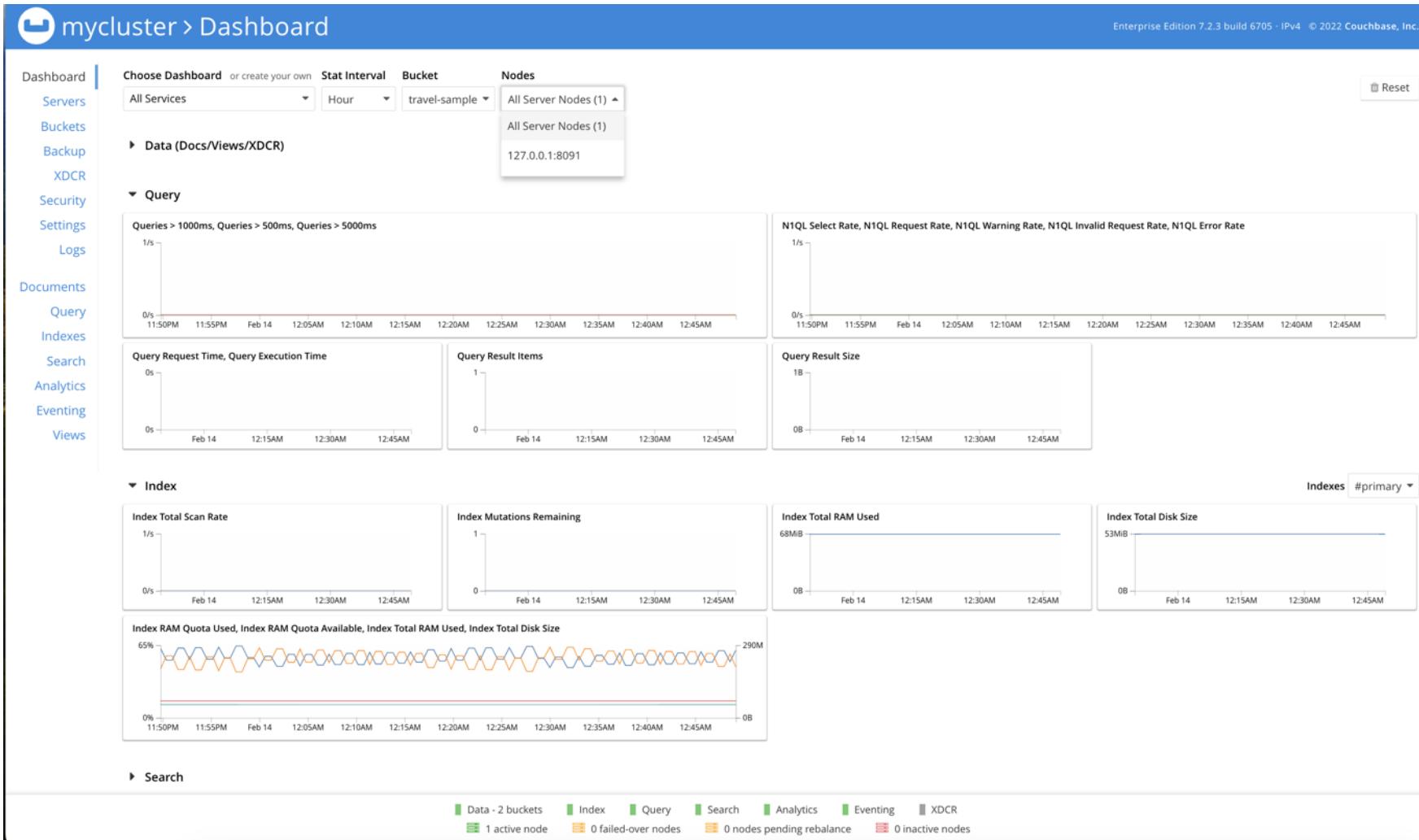
• Restore 방식

- Bucket, Scope, Collection 단위 가능



8 운영/관리 : 성능 모니터링

Couchbase는 다양한 성능 매트릭을 제공합니다. 그리고, Prometheus/Grafana 로도 관리할 수도 있습니다.



8

운영/관리 : 시스템 상태 경고 및 권고

Couchbase는 매트릭이 임계 수준을 넘어서면 경고(Alerts) 보내게 되고, 권고 사항(Recommended actions)도 제시됩니다. 경고 수준(Severity Level)은 **Info**, **Warning**, **Critical** 이 있습니다.

The screenshot shows the Couchbase Settings page with the 'Alerts' tab selected. On the left, there's a sidebar with various navigation options like Dashboard, Servers, Buckets, etc. The main area has sections for enabling email alerts, setting up an email server host (localhost), specifying a port (25), and entering a username and password. There's also a checkbox for 'Require encryption (TLS)'. Below these, there's a 'Sender Email' field set to 'couchbase@localhost' and a 'Recipients' field containing 'root@localhost'. At the bottom is a 'Send Test Email' button. The right side lists 'Available Alerts' with checkboxes for each alert type, such as 'Email UI popup Node was auto-failed-over' and 'Email UI popup Memory usage threshold exceeded'.

Alert Type	Description
Email UI popup	Node was auto-failed-over
Email UI popup	Maximum number of auto-failed-over nodes was reached
Email UI popup	Node was not auto-failed-over as other nodes are down at the same time
Email UI popup	Node was not auto-failed-over as there are not enough nodes in the cluster running the same service
Email UI popup	Node was not auto-failed-over as auto-failover for one or more services running on the node is disabled
Email UI popup	Node's IP address has changed unexpectedly
Email UI popup	Disk space used for persistent storage has reached at least 90% of capacity
Email UI popup	Metadata overhead is more than 50%
Email UI popup	Bucket memory on a node is entirely used for metadata
Email UI popup	Writing data to disk for a specific bucket has failed
Email UI popup	Writing event to audit log has failed
Email UI popup	Approaching full Indexer RAM warning
Email UI popup	Remote mutation timestamp exceeded drift threshold
Email UI popup	Communication issues among some nodes in the cluster
Email UI popup	Node's time is out of sync with some nodes in the cluster
Email UI popup	Disk usage analyzer is stuck; cannot fetch disk usage data
Email UI popup	Memory usage threshold exceeded
Email UI popup	History size threshold exceeded
Email UI popup	Approaching Indexer low resident percentage

9 보안 > End-to-End 보안

Couchbase는 엔드 투 엔드 보안 기능을 제공합니다. 클라이언트 접속 인증에서 네트워크 보안, 서버 스토리지 저장시 암호화할 수 있으며, 다양한 감사 기능 제공합니다.

Authentication 인증	Authorization 권한	Crypto 암호화	Auditing 감사	Operations 운영보안
<ul style="list-style-type: none">App/Data: SASL AuthenticationUsers Database:<ul style="list-style-type: none">LocalLDAP / ADPAM	<ul style="list-style-type: none">RBAC<ul style="list-style-type: none">For AdminFor DataLimits at various levels<ul style="list-style-type: none">ClusterBucketScope (v7.0)Collection (v7.0)	<ul style="list-style-type: none">TLS Admin AccessTLS client-server accessTLS XDCRX.509 certificatesSecret ManagementTLS Protocol and Cipher MgmtData-at-rest Encryption*	<ul style="list-style-type: none">Audit Events:<ul style="list-style-type: none">LoginAdd nodeSettingsN1QLEtc.	<ul style="list-style-type: none">Security management via UI/CLI/RESTLog RedactionNon-Root Install

9 보안 > 사용자/그룹 권한 관리

Couchbase는 그룹과 사용자에 대한 권한 관리를 통해 Admin 관리 기능과 데이터별 쓰기/읽기를 허용할 수 있습니다.

The screenshot shows the Couchbase Security interface with the following details:

- Dashboard:** Shows a summary of the cluster: Sync Gateway Replicator [*:*:*], Sync Gateway Architect [*:*:*], Sync Gateway Application Read Only [*:*:*], Sync Gateway Application [*:*:*], Sync Gateway [*], Sync Gateway Dev Ops.
- Filter:** A search bar with placeholder "filter by username..." and a magnifying glass icon.
- LDAP Status:** A note that "LDAP/SAML is not enabled".
- Users & Groups:** A dropdown menu currently set to "Users".
- Buttons:** ADD GROUP and ADD USER.
- Left Sidebar:** Includes links for Servers, Buckets, Backup, XDCR, Security (selected), Settings, and Logs. A dropdown menu shows "20" items.
- Add New Group Dialog:** Fields for "Group Name" and "Description". A "Map to LDAP Group" section is present. A "Roles" section lists various administrative and data-related roles with checkboxes. Buttons: Cancel and Save.
- Add New User Dialog:** Fields for "Username", "Full Name (optional)", "Password", and "Verify Password". A "Roles" section lists various roles with checkboxes, and a "Groups" section lists existing groups with checkboxes. Buttons: Cancel and Add User.

10 설치/구성 : 지원 플랫폼

Bare-Metal, VM, Container 와 완전관리형 클라우드 데이터베이스 서비스(DBaaS) 사용 가능



Fully Managed

- 완전관리형 데이터베이스 서비스
- AWS, GCP, Azure
- 설치, 구성, 모니터링, 업그레이드 등 모든 운영은 Couchbase가 담당



Enterprise

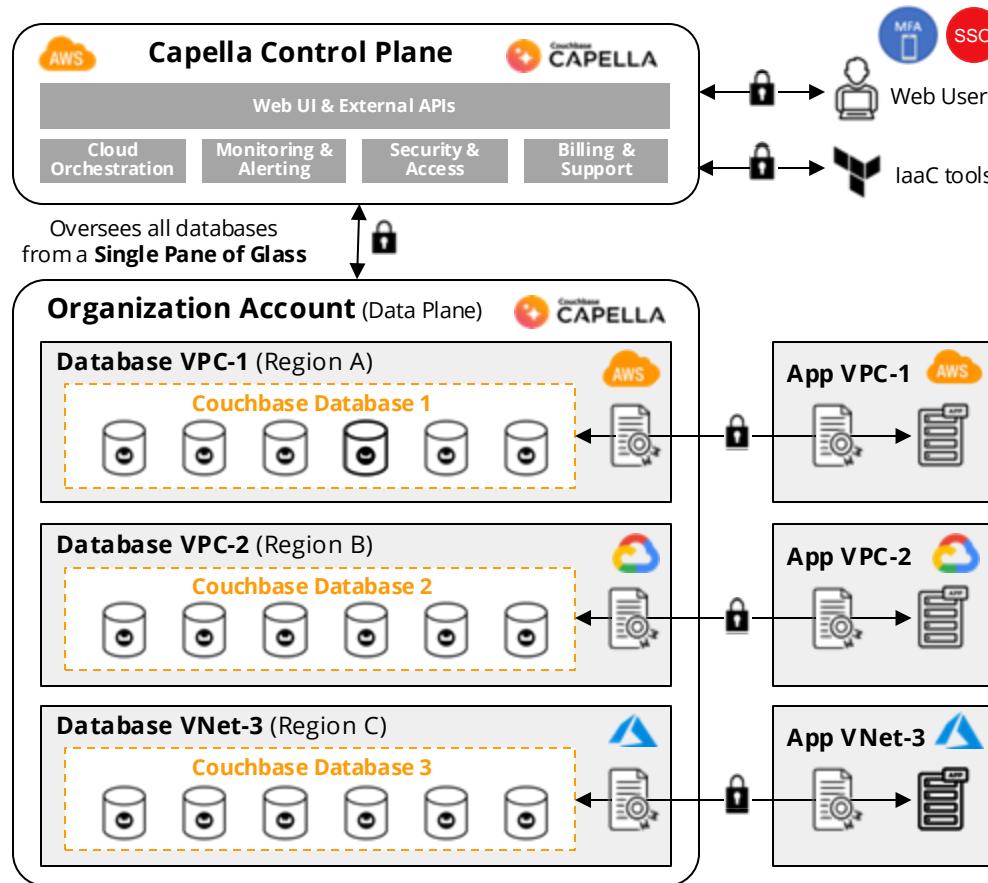
- Bare-metal 서버, 클라우드 IaaS 서버
- Private Cloud 서버, K8S 컨테이너
- 설치, 구성, 모니터링, 업그레이드 등 모든 운영은 고객이 수행



Linux, Windows, MacOS, Intel/AMD, ARM

10 설치/구성 : 카펠라 아키텍처

완전관리형 클라우드 데이터베이스 서비스(DBaaS)인 Capella 아키텍처



Capella Control Plane

- Manages the Cloud Orchestration, Monitoring & Alerting, Security & Access, Billing & Support
- Is the Access Point for Organization Web UI Users
- Allows Infrastructure as Code Tools (e.g. Terraform)

Organization Account (Data Plane)

- Account Isolation:** 1 Account per Organization
- Database Isolation:** 1 VPC per Couchbase Database
- Multiple Clouds:** AWS, Google and Azure

Applications

- Connect directly to Databases
- Multiple connectivity options: over Public Connection, through VPC Peering or Private Link
- All communications are encrypted

11

교육 : Couchbase Academy

<https://learn.couchbase.com/store>

The screenshot shows the Couchbase Academy store page. At the top, there's a banner with two women working at a computer, with the text "Couchbase Academy" overlaid. Below the banner, a message says "Welcome to the Couchbase Academy instructor-led and eLearning training options! Couchbase Certification Exams for 2023, now without a proctor requirement." A "Questions?" button is visible. Below the banner is a search bar with a placeholder "Search by keyword" and a "Search" button, followed by a "or" link and a "All Types and topics" button.

On the left, there's a sidebar with a red circular icon and the word "Couchbase". Below it, a section titled "Upcoming Sessions" lists an event for February 20, 2024, titled "CD410: Advanced N1QL Course: Tuning and Optimization - APAC Virtual (GMT+8)". It includes details like starting and ending times (02/20/2024 @ 09:00 AM (GMT+08:00) Singapore / 02/23/2024 @ 05:00 PM (GMT+08:00) Singapore) and a "Type: Multi-day Session" label.

The screenshot shows the Couchbase Academy course catalog with four entries:

- CB130n: Couchbase Associate Node.js Developer Certification With Capella Course**
This newly revamped course shows how to leverage the full power of Couchbase 7 as a service with Couchbase Capella. The following 8 courses provide a fundamental understanding of the Couchbase NoSQL database and essential functionality. Throughout these courses, we share the basics of SQL vs. NoSQL, how to sign up for Couchbase Capella, modeling data to the benefit of Couchbase, and an example application you will build. Learners will also walk through the basics of Couchbase's N1... [Read More](#)
- CB130j: Couchbase Associate Java Developer Certification With Capella Course**
This newly revamped course leverages the full power of Couchbase 7 and supports Couchbase Capella. The following 8 courses provide a fundamental understanding of the Couchbase NoSQL database and essential functionality. Throughout these courses, we share the basics of SQL vs. NoSQL, obtaining and downloading Couchbase, modeling data to the benefit of Couchbase and an example application you will build. Learners will also walk through the basics of Couchbase's N1... [Read More](#)
- CB131: Couchbase Associate Architect Certification With Capella Course**
This newly revamped course demonstrates the full power of Couchbase 7 and the fully-managed Database as a Service (DBaaS), Couchbase Capella. The Couchbase Associate Architect Course shares a fundamental understanding of the Couchbase NoSQL database and essential functionality as accessed through the Couchbase Capella user interface. It discusses modeling data to the benefit of the database and application, as well as how to write and implement SQL... [Read More](#)
- CB140a: Couchbase Associate Android Developer With Capella Course**
This course showcases and demonstrates how to create a new Android application using Couchbase Mobile and Couchbase Capella, our fully managed DBaaS service. The following 7 modules provide fundamental instruction on building an Android application with or without a pre-existing database. To that end, we walk through the essential functionality of Couchbase Capella, the benefits of a fully managed NoSQL database, and how that database interacts with Couchbase Mobile products t... [Read More](#)

12 요약하면,

Couchbase는

JSON Document
직관적이고

SQL, Generative AI
익숙하고, 쉽게

Data Platform
일관적 적용

- 사람이 인지 하는 세상을 그대로 데이터 모델로 사용
- 복잡한 정규화 과정 없이 직관적인 방식으로 어플리케이션 개발/운영

- NoSQL 이지만 표준 SQL을 지원
- 생성적 AI인 Capella IQ 지원으로 더 손쉬운 개발이 가능

- Key/Value 데이터서비스에서 분석서비스, 모바일 앱서비스까지 일관성있게 업무에 적용 가능
- 센서, 모바일, 퍼스널컴퓨터, 데이터센터 서버, 쿠버네티스, 클라우드에 동일한 데이터플랫폼 적용으로 개발의 일관성 뿐만 아니라 데이터의 일관성도 보장

Enterprise에서 요구하는 성능, 안정성, 보안성

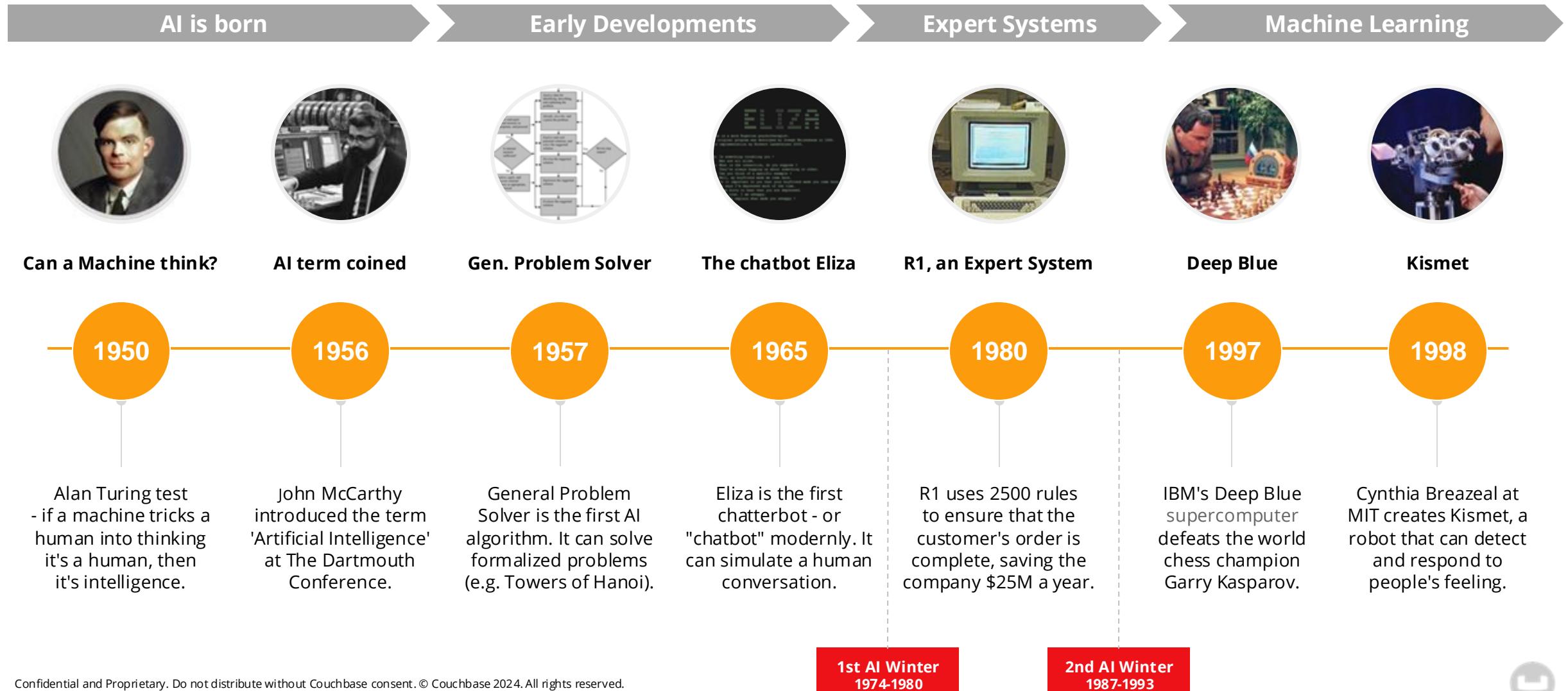
A Brief History of AI

>

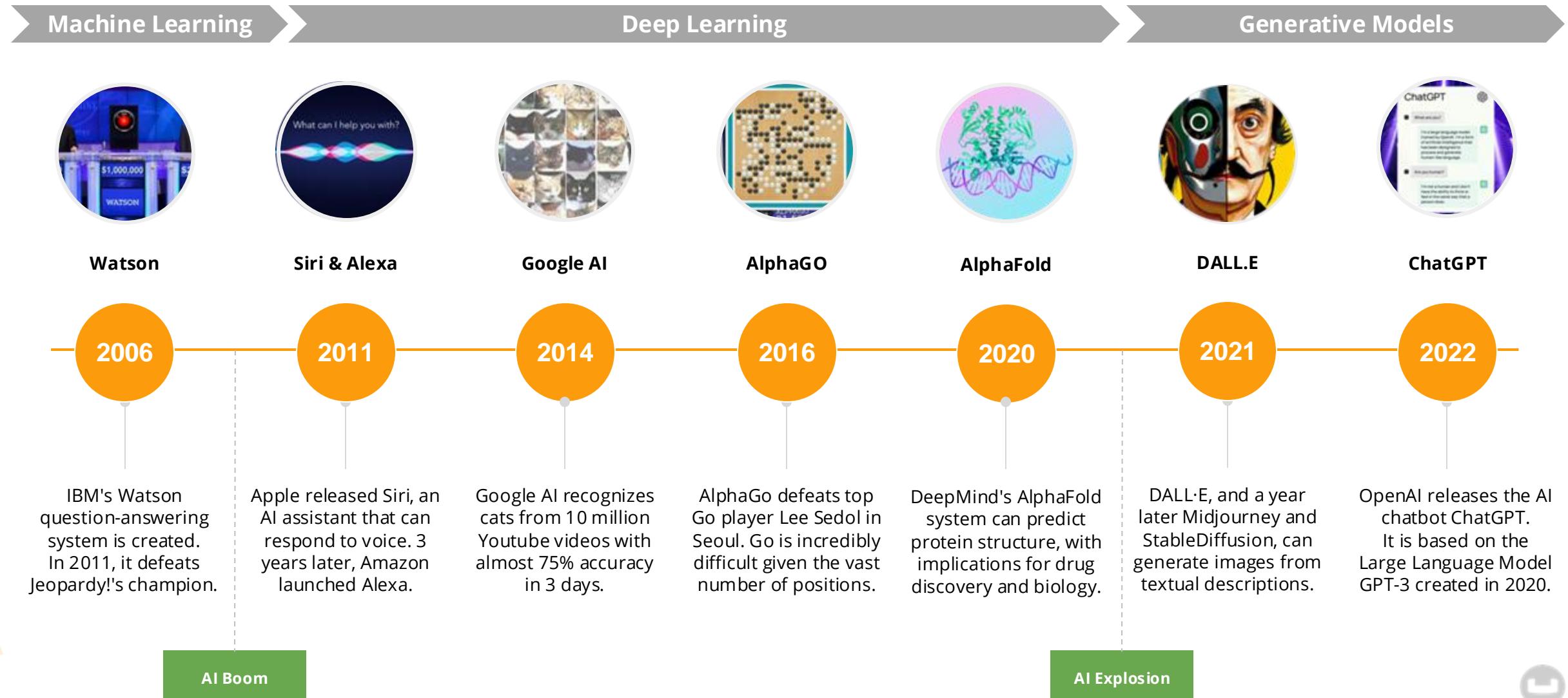
T



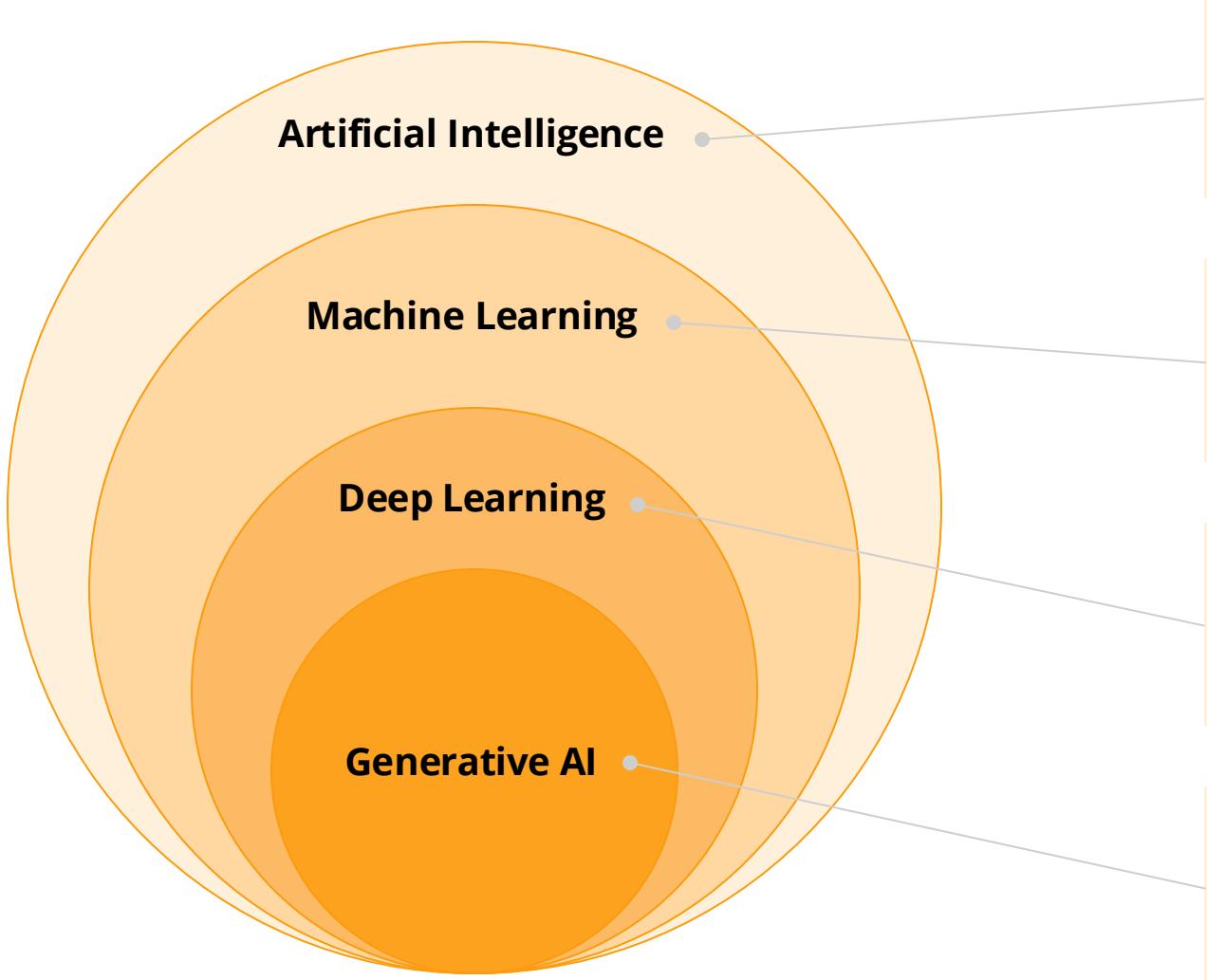
Key Milestones in the History of AI | 20th Century



Key Milestones in the History of AI | 21st Century



The Technology behind AI



Artificial Intelligence (AI)

Techniques that allows computers to emulate human behavior (e.g. learn, recognize patterns, solve complex problems).

Machine Learning (ML)

A subset of AI, using advanced algorithms to detect patterns in large data sets, allowing machines to learn and adapt for prediction or content generation use cases.

Deep Learning (DL)

A subset of ML, using multiple layers of artificial neural networks that simulate human brains for in-depth data processing.

Generative AI (GenAI)

A subset of DL, using models that generate content like text, images, or code based on provided input.

Powering Apps: A Combination of Predictive & Generative AI

Predictive AI

Outcomes and Insights driven by ML



- Predict Outcomes based on historical data
- Utilize ML algorithms for pattern recognition
- Learns patterns and correlations from data
- Drives decision making and Future planning
- High ROI, trained on proprietary data

- Predictive Insights
- Dynamic Pricing
- Fraud Detection
- Inventory Optimization

Generative AI

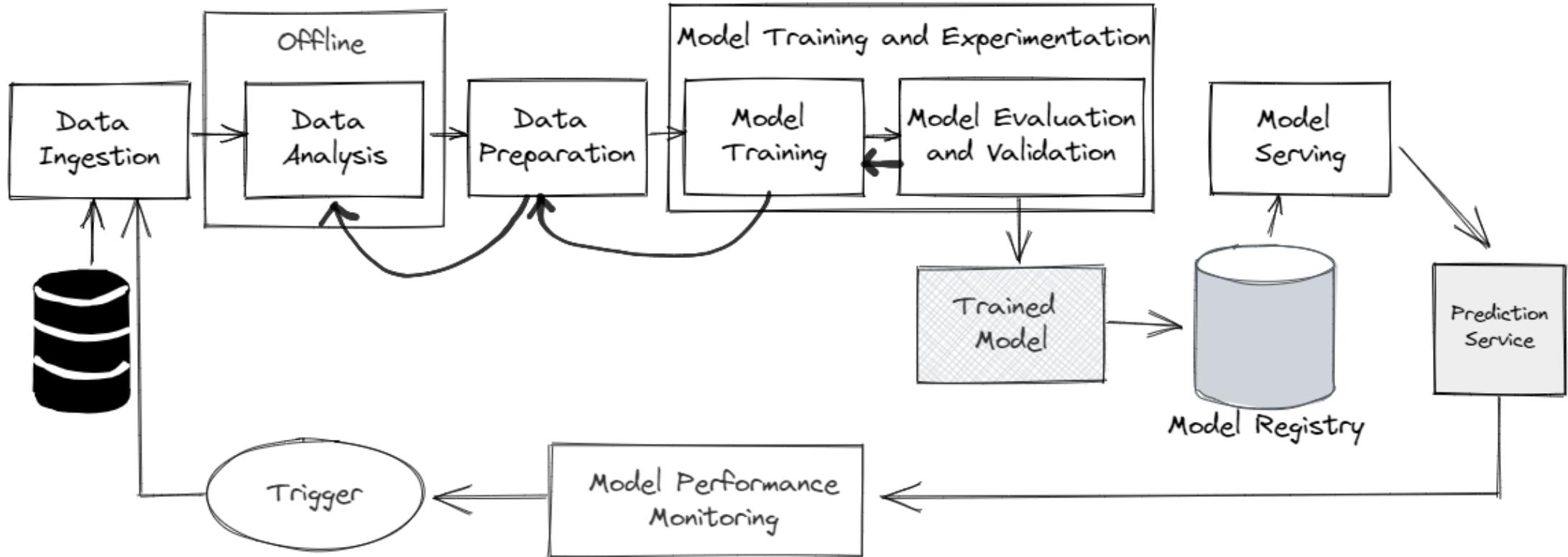
Generate Content and Experiences



- Generate or Synthesize content
- Needs large amounts of unlabeled data for training
- Generates new data probabilistically
- Fosters creativity, innovation
- Accelerates human productivity

- Hyper-personalized experiences
- Contextualized content
- Chatbots and CoPilots
- Synthetic data and Summarization

Model? Machine Learning Workflow



출처 : <https://www.iguazio.com/blog/ml-workflows-what-can-you-automate/>

<https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning>

Vector DB, Search

1. Demystifying Vector Search
2. The power of Hybrid Search

GenAI (LLM, RAG)

3. Vector Search Use Case:
Semantic Search
4. A quick tour of LLMs and Generative AI
5. Vector Search Use Case:
Retrieval-Augmented Generation (RAG)



Demystifying Vector Search



What is a Vector

This is a vector

2.6	11.3	-4.2
-----	------	------



First value



Second value



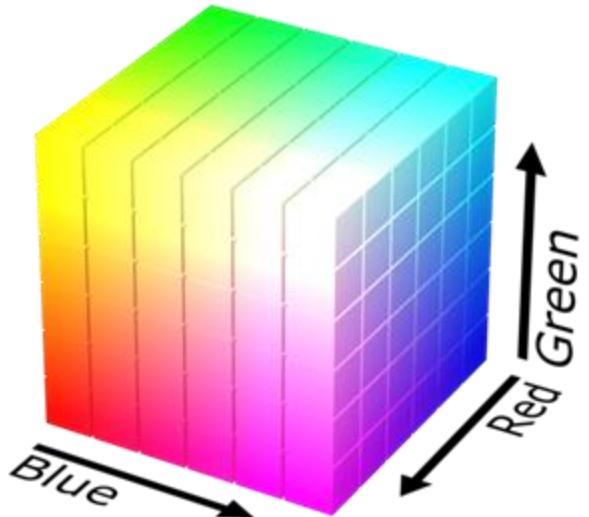
Third value

Here, it contains 3 values
=> its dimension is 3

A Vector is a just an **array of numerical values**

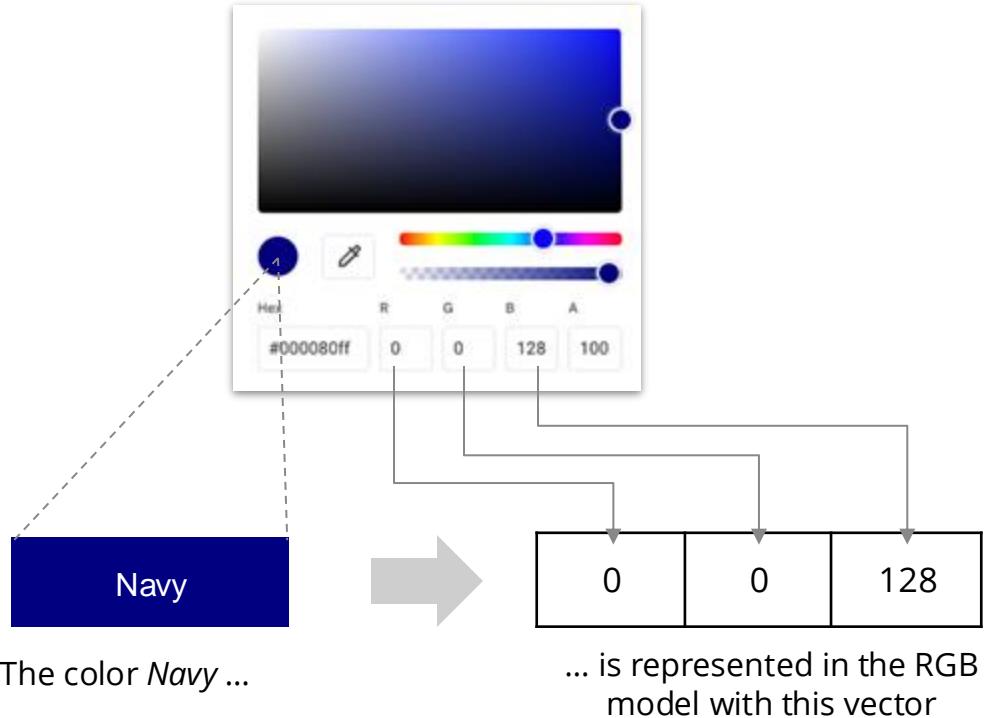
What is a Vector used for | Basic RGB Example

The RGB model example



The model used to create the colors you see on TV and computer screens.
Each color is the addition of a **Red**, **Green** and **Blue** primary colors.

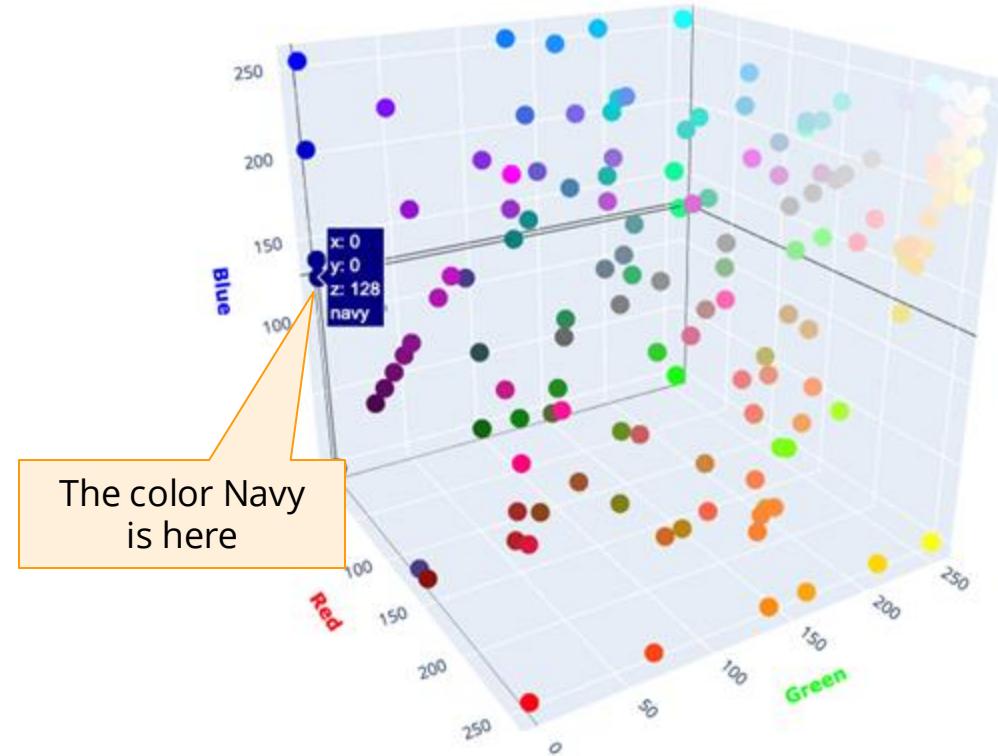
A Vector representation of a Color



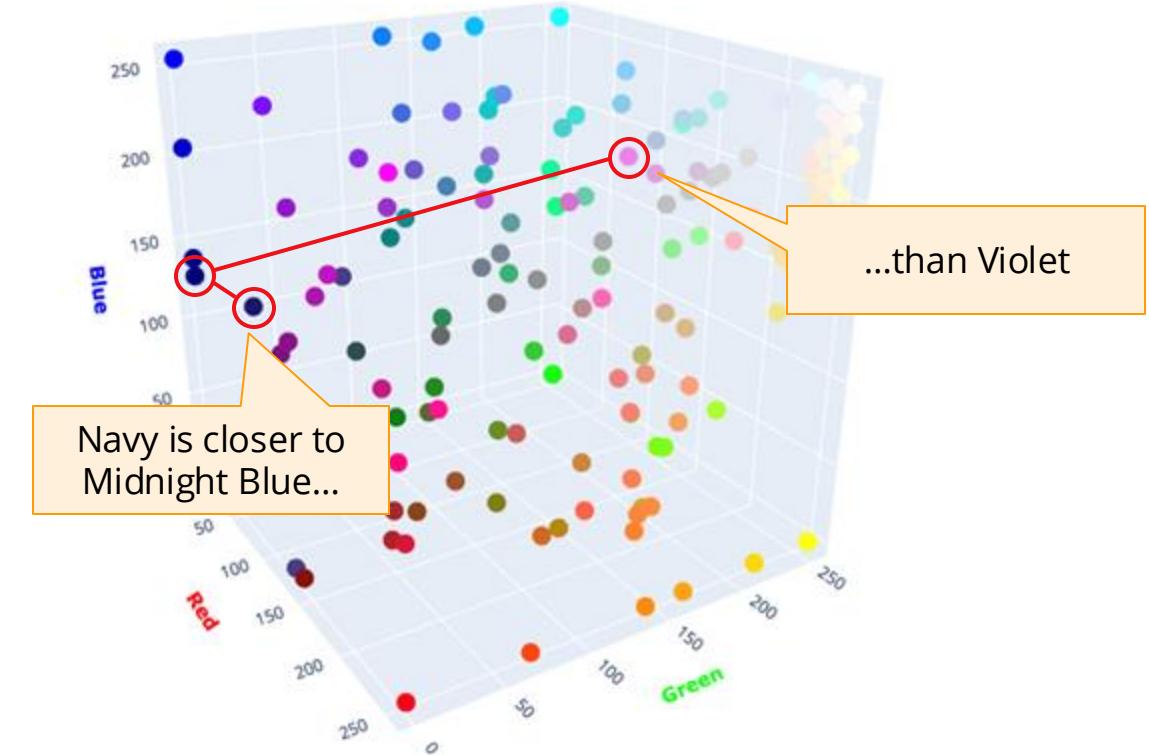
A **Vector** is used to **represent** a more complex object. The **Model** defines the **meaning** of each dimension.

Vectors Similarity

Example of 123 vectors of RGB colors



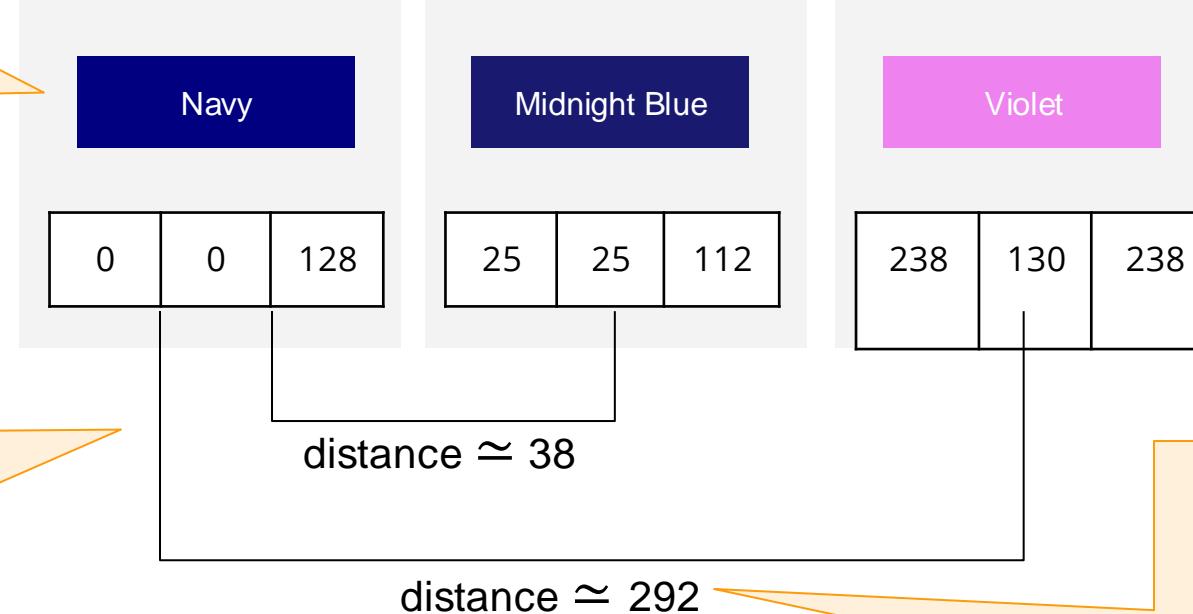
Similar colors are closer to each other



Vectors make it possible to translate **similarity** as perceived by humans to **proximity in a vector space**.

How does Similarity works

To the human eyes,
Navy is closer to Midnight
Blue than Violet



Mathematically,
we got the same result by
comparing the vectors

Vectors are compared using
a similarity distance.

Here the *euclidean distance*
 $292 \approx \sqrt{(238-0)^2 + (130-0)^2 + (238-128)^2}$

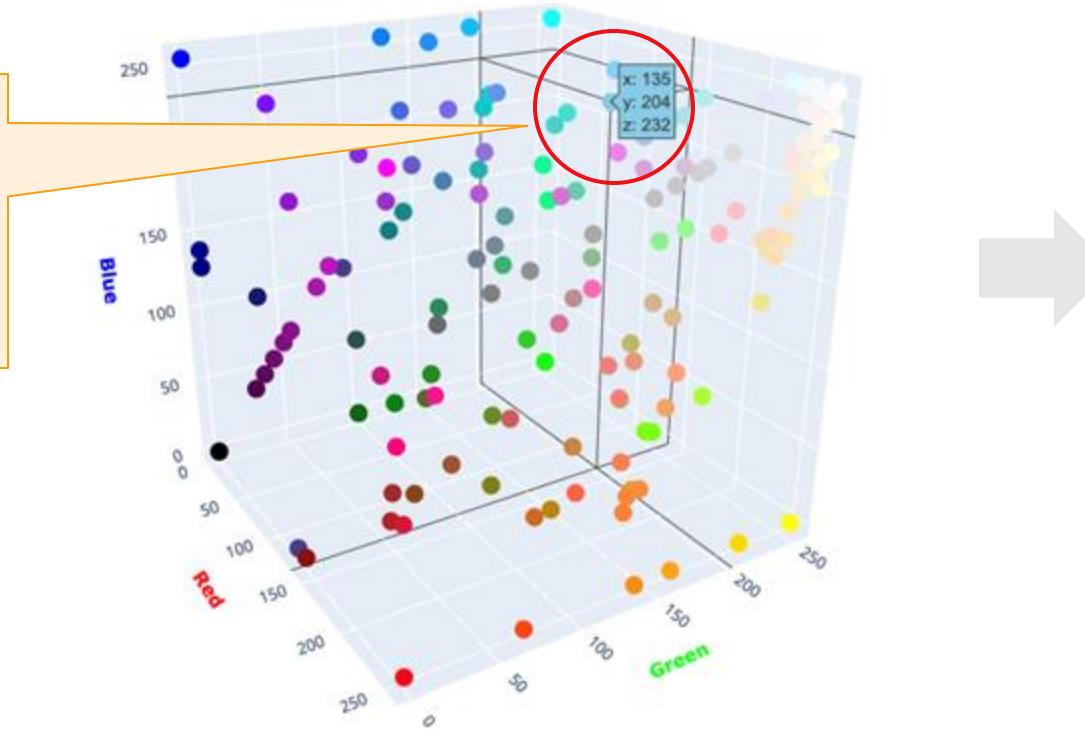
Vectors can easily be compared mathematically using a **similarity distance**

Similarity Search with K-NN (K-Nearest Neighbors)

Which are the top k nearest neighbors to this color?

[135,204,232]

Example of 123 vectors of RGB colors

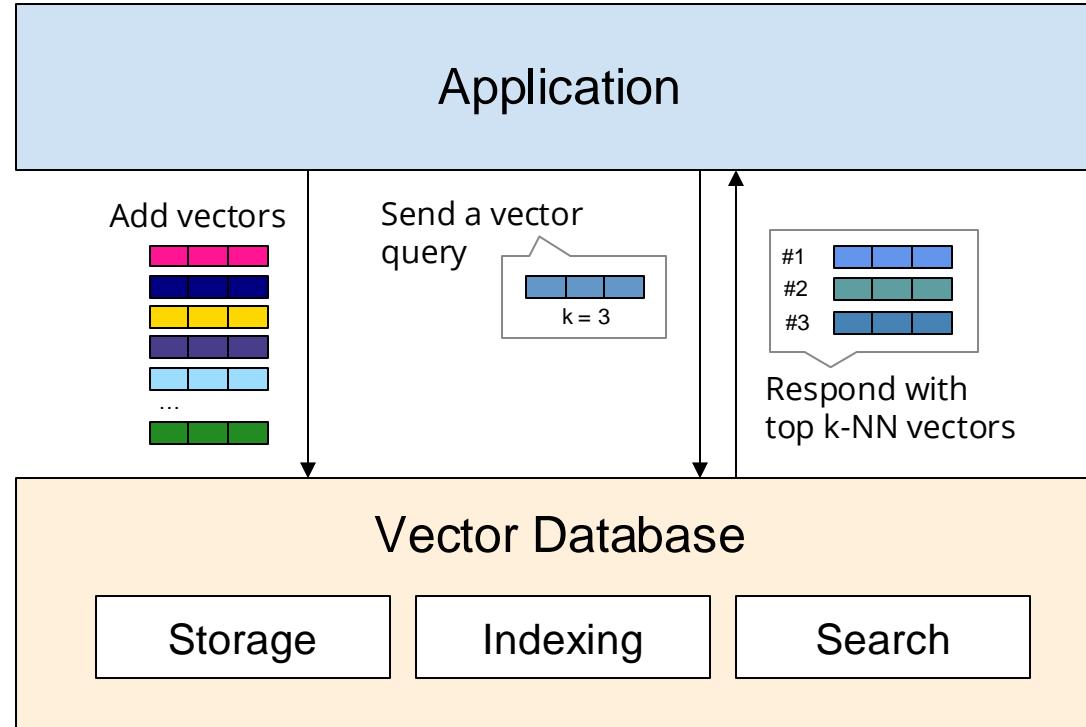


Top k-NN results of the query

- #1 sky blue [135,206,235]
- #2 light sky blue [135,206,250]
- #3 light blue [173,216,230]

A similarity search is a query that **finds the k nearest neighbors to a vector**, as measured by a similarity metric

What is Vector Database



Vector databases provide the ability to **store, index and search vectors** using similarity search

Couchbase Vector Search

The vectors are stored as a **field in JSON** documents

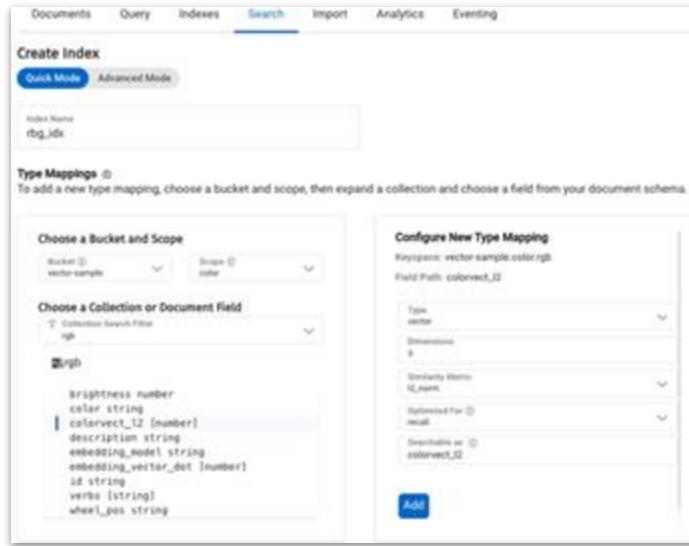
```
{  
  "id": "#000080",  
  "color": "navy",  
  "brightness": 14.592,  
  "colorvect_l2": [0, 0, 128],  
  "description": "Navy is a deep, rich color that  
    exudes sophistication. It is a dark shade of  
    blue that is often associated with authority,  
    stability, and elegance. Navy is a versatile  
    color that can be both bold and understated,  
    making it a popular choice in fashion and  
    interior design. It is a timeless color that  
    never goes out of style and adds a touch of  
    sophistication to any look or space.",  
}
```

JSON Storage

Data Service



A **Vector Index** must be created to allow the vectors to be searched



Vector Index

Search Service

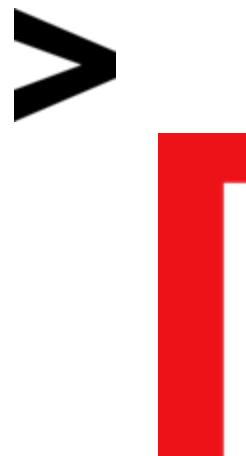
A **Vector Query** can now search for the top k-NN of a color

```
{  
  "query": { "match_none": {} },  
  "knn": [  
    {  
      "field": "colorvect_l2",  
      "vector": [135, 204, 232],  
      "k": 3  
    }  
  ],  
  "fields": ["color"]  
}
```

Vector Query

Couchbase uses the **Data Service to store vectors**, and the **Search Service to index and query vectors**

The power of Hybrid Search



Hybrid SQL++ and Vector Search with Couchbase

This is a **SQL++ query**

Combining Vector Search query

And standard SQL++ criteria

```
SELECT color, brightness
FROM `vector-sample`.color.rgb AS t1
WHERE
  SEARCH(t1,
  {
    "query": { "match_none": {} },
    "knn": [
      "field": "colorvect_l2",
      "vector": [135,204,232],
      "k": 3
    ]
  }
)
AND
  brightness >= 180 AND brightness <= 190
```



SQL++ is easy and familiar to developers



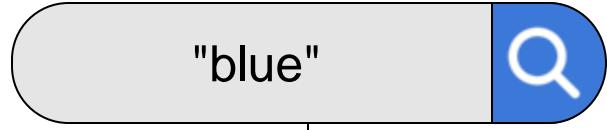
You can filter vector search results with other criteria



You don't have to run 2 separate databases, one for Documents and one for Vector Search!

Couchbase can run hybrid SQL++ and Vector Search queries to **facilitate application development**

Comparison between Keyword Search and Vector Search



Keyword Search on the
description of the colors



Vector Search on the
RGB vectors of the colors

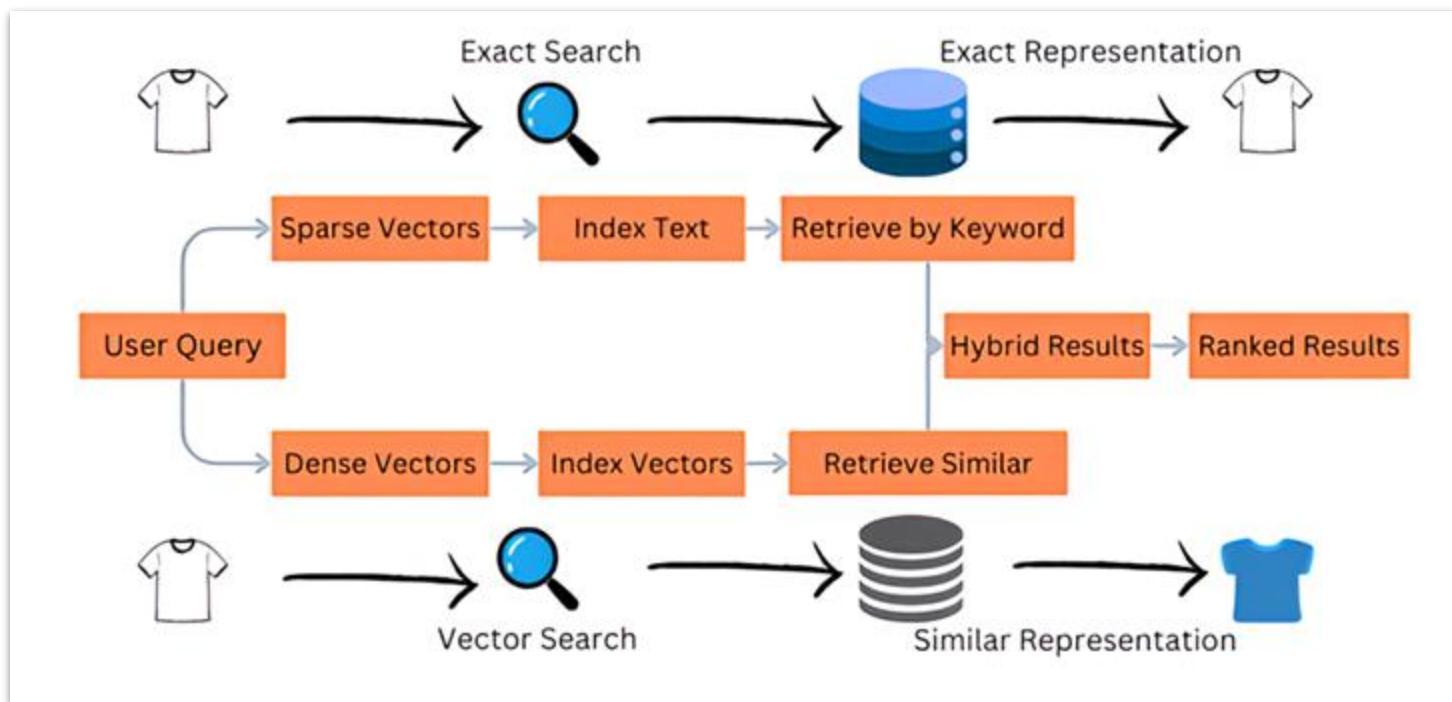


A Keyword search looks for **terms** that match

A Vector search looks for **similarity**

Hybrid Search to get the best of both worlds

Hybrid Search Architecture



Hybrid Search with Couchbase

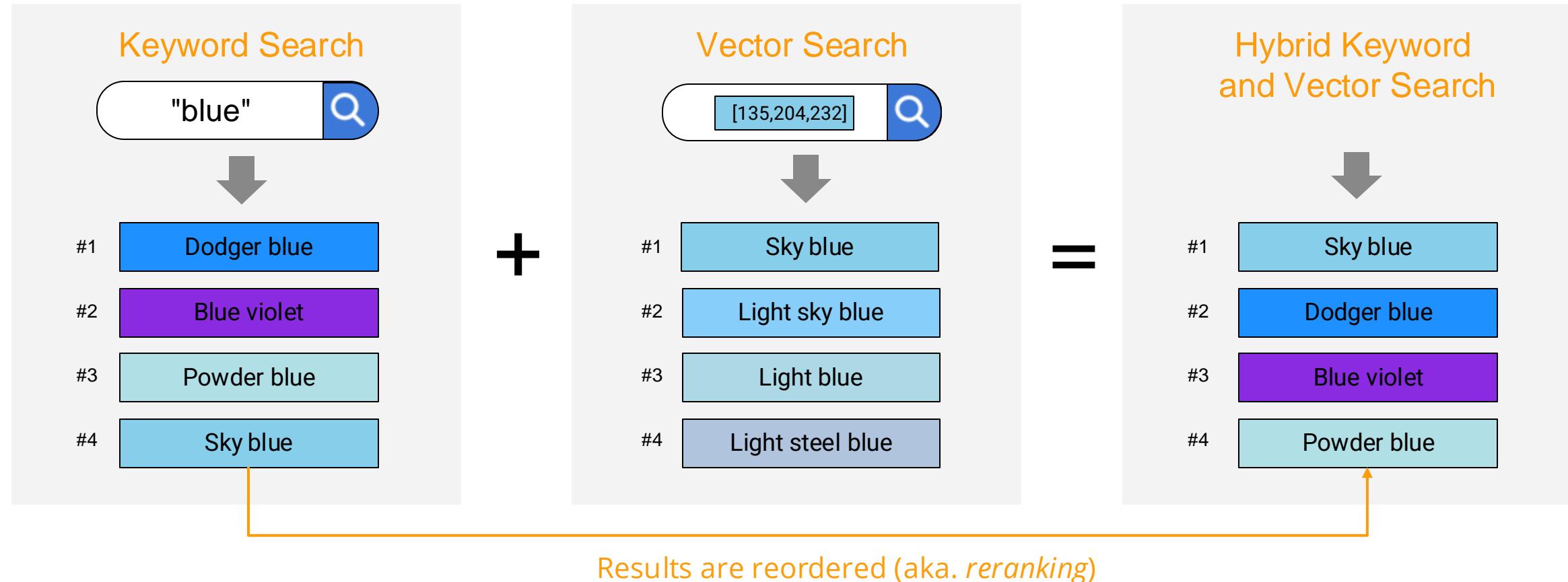
```
{  
  "query": {  
    "match": "blue",  
    "field": "description"  
  },  
  "knn": [  
    {  
      "field": "colorvect_l2",  
      "vector": [135,204,232],  
      "k": 4  
    }  
  ],  
  "fields": ["color","description"],  
  "size": 4  
}
```

Arrows on the right side of the JSON object point to specific fields, indicating their purpose:

- Keyword search:** Points to the "query" field.
- Vector search:** Points to the "knn" field.
- Results to return:** Points to the "size" field.

Vector search in conjunction with traditional Keyword search delivers the most complete and relevant results

Hybrid Keyword and Vector Search Example

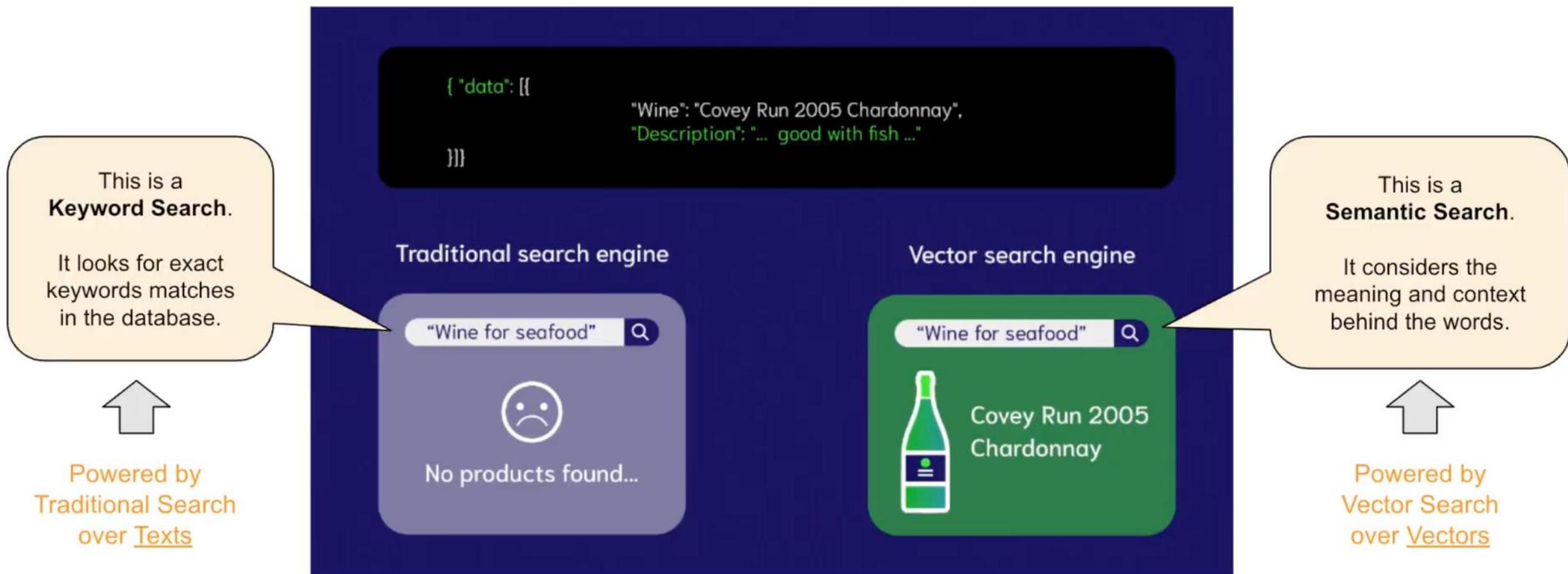


Results from the Keyword search are **boosted** if they appear in the Vector Search results

Vector Search Use Case: Semantic Search

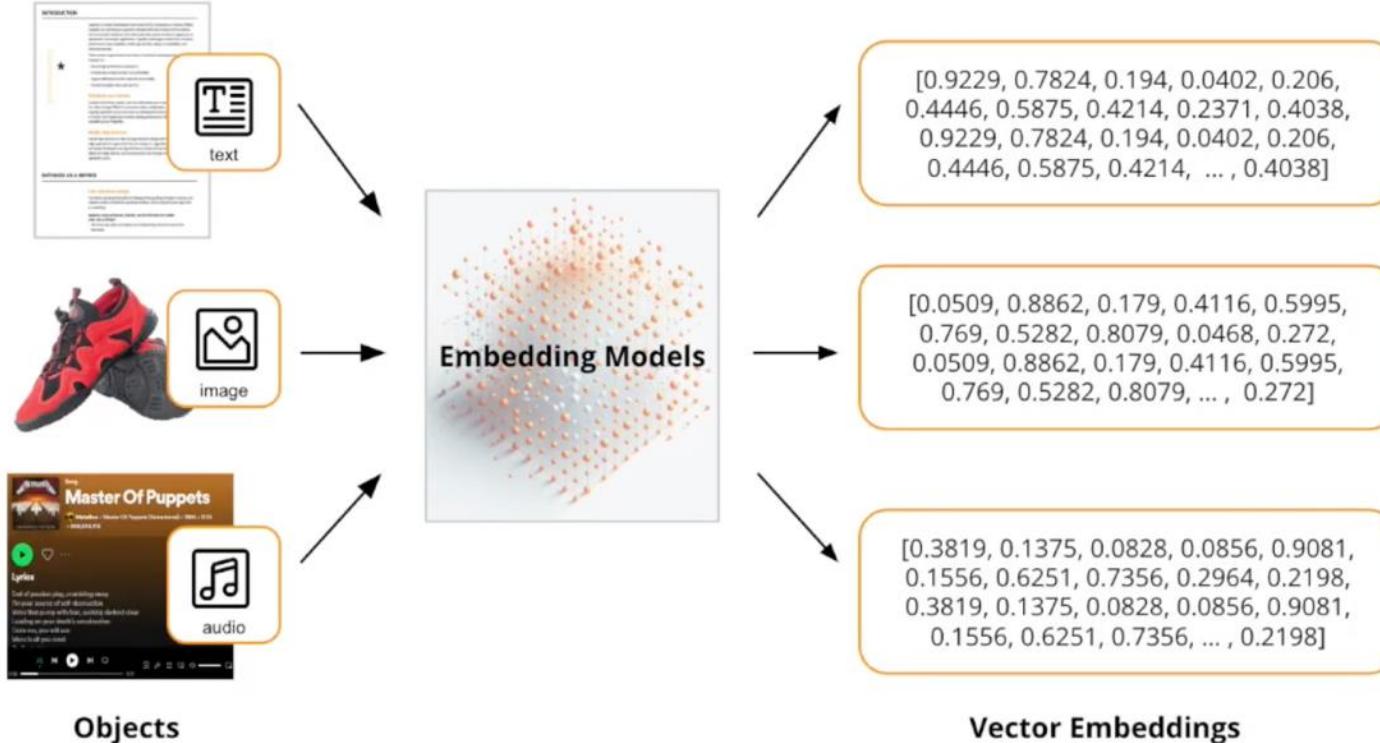


What is a Semantic Search



Semantic Search = Similarity Search across Vectors Embeddings representing **the meaning of complex Objects**

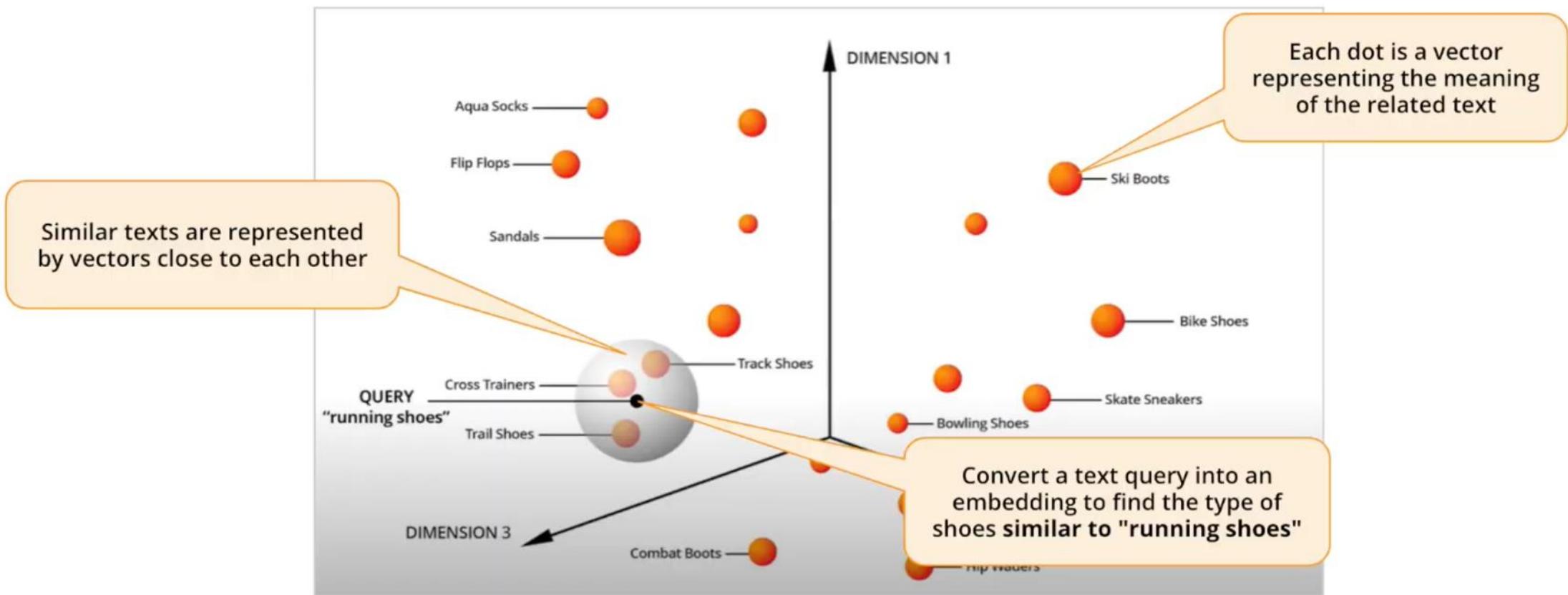
What are Embedding Models



- Embedding Models convert complex objects into vectors using **Machine Learning**
- Embedding Models are trained with very large datasets so they can **preserves the meaning of objects**
- The vectors created are called **Vector Embeddings** - referred as Embeddings for short
- Vector Embedding that can have **large dimensions** (e.g. 1536)

Embedding models are algorithms trained to **convert objects into Vectors Embeddings** (aka. Embeddings)

Example of Texts converted into Vector Embeddings



Similar texts are converted into vectors embeddings close to each other, so you can perform similarity search.

Generating Embeddings is easy for Developers

Example of code to generate Vector Embeddings

```
import os
from openai import OpenAI

openai_api_key = os.getenv("OPENAI_API_KEY")
client = OpenAI()

text = "Your text string goes here"

print(client.embeddings.create(input = [text],
model="text-embedding-ada-002").data[0].embedding)
```



```
[  
    0.024032991379499435,  
    -0.009131478145718575,  
    0.013961897231638432,  
    ...  
    0.0034673146437853575  
]
```

Choose your Embedding Model Platform

- In this example, this is Azure OpenAI

Provide credentials to access the platform

- OpenAI requires API keys for authorization

Provide the text to encode

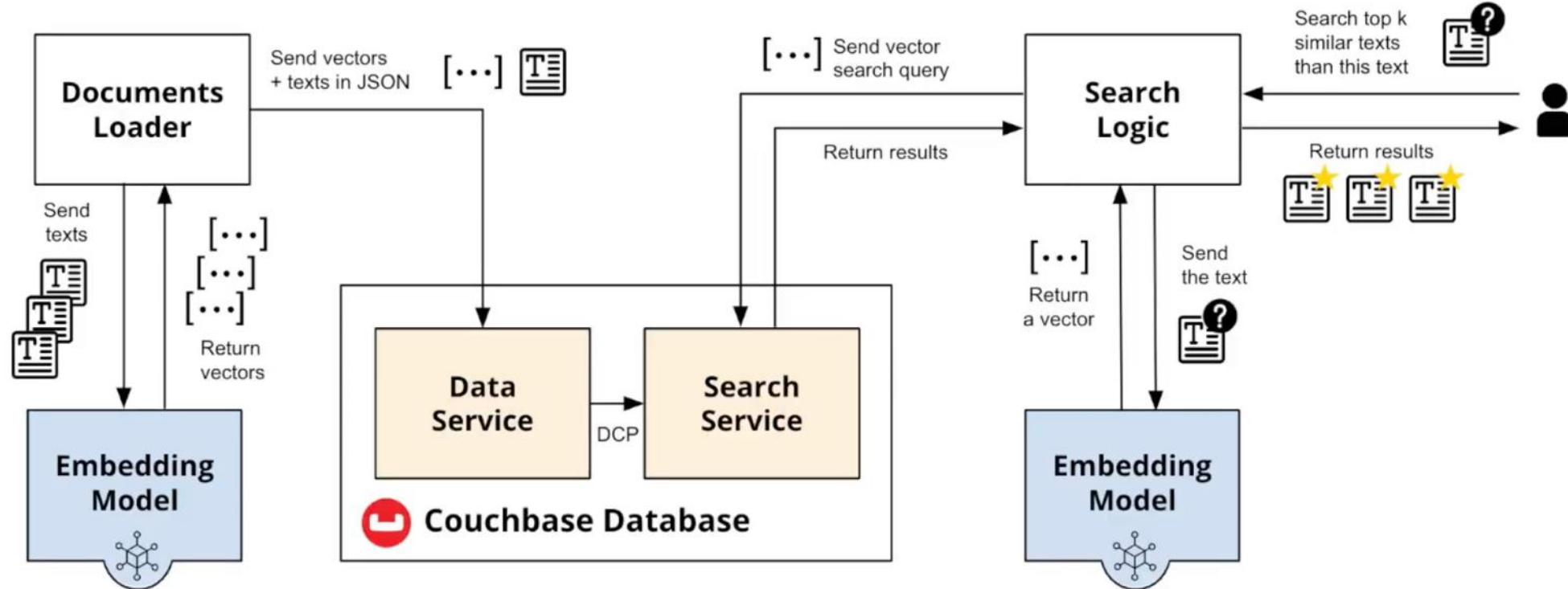
- Either the objects or the query

Send the text to the Embedding Model

- In this example text-embedding-ada-002 model

Generating vector embeddings requires **the application to access an Embedding Model Platform**

Semantic Search With Couchbase



Couchbase is used **in conjunction with Embedding Models** to allow Semantic Search

The number of Embedding Models is growing fast

Open-Source Text Embedding Models

Word2Vec Google project in 2013

Glove Stanford University project - 2014

BERT Adopted by Google Search in 2019

 **txtai** [github link](#)

 **chroma** [github link](#)

Proprietary Text Embedding Models



Text-embedding-ada-002 (2022)
text-embedding-3-small (2024)



embed-english-v3.0
embed-multilingual-v3.0



Vertex AI text-embeddings API



Titan Text Embeddings models

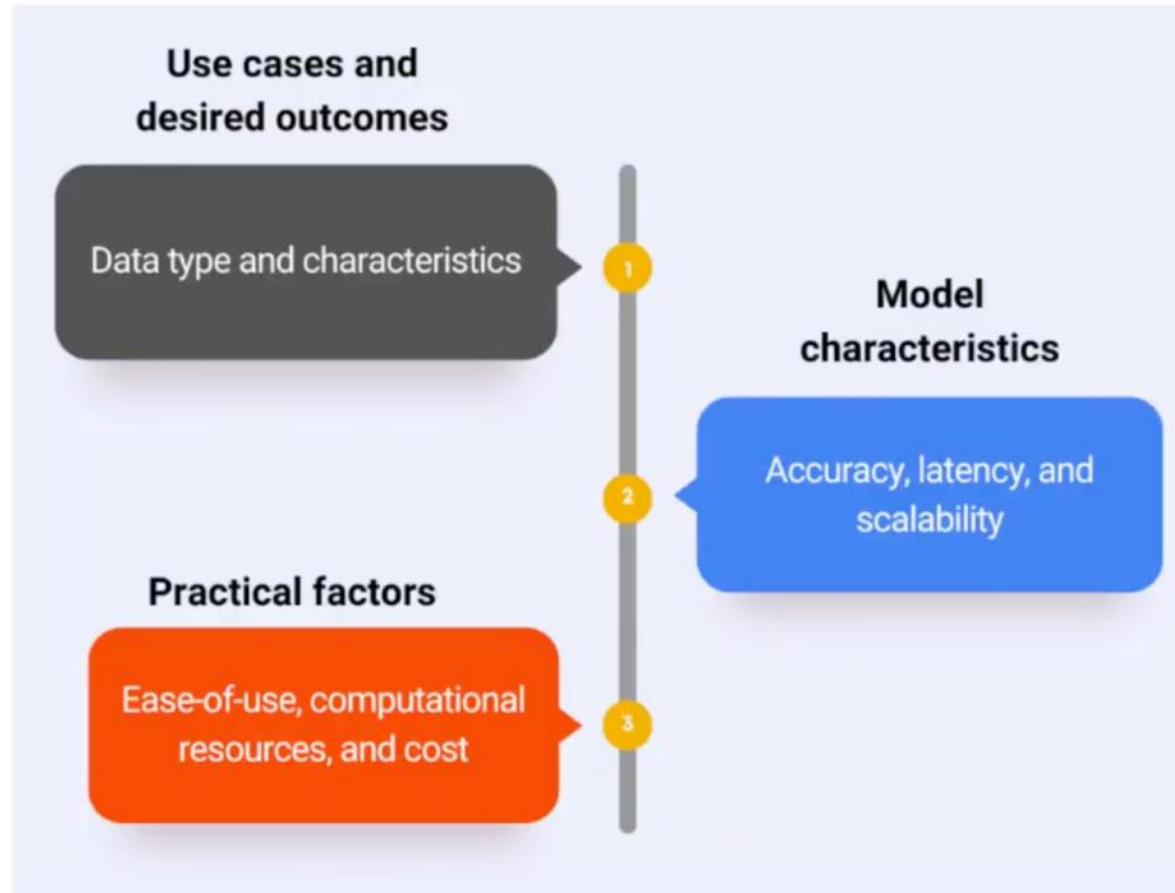
More than 300 Text Embedding Models can be found in the [MTEB leaderboard](#)

How Customers are choosing their Embedding Models

Each Embedding Model can process only a certain type of data (e.g. text, or video, or audio) and trained with a specific set of data.

Review the model's performance in terms of resource requirements.

For example, large vectors can result in significantly higher costs.



The quality of the search depends crucially on the quality of the model.

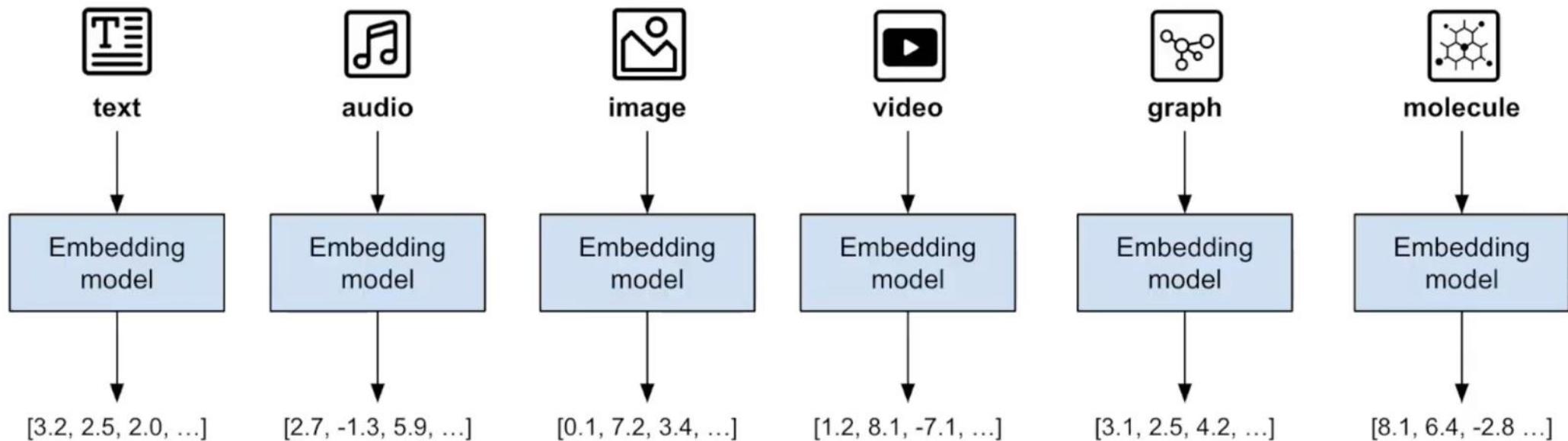
Latency is a key factor to deliver real-time interactions with the app.

This is the responsibility of the customer to choose the Embedding Model based on many factors



Type of Objects that can be Embedded

And many more, as long as there is an Embedding Model to vectorize them



Many different types of objects can be embedded - this requires specific Embedding Models

Semantic Search Use Cases | Examples

Recommendations

Semantic search enhances product recommendations by understanding customer preferences beyond keyword matches.

Content Discovery

Media platforms can leverage semantic search to help users discover relevant articles, videos, or music.

Fraud Detection

Identifying unusual patterns in customer behavior to detect potential fraudulent activities. Customer's behaviors over time are represented as vector embeddings.

Medical Diagnosis

In the medical field, semantic search assists doctors in diagnosing diseases and finding relevant research.

Enterprise Search

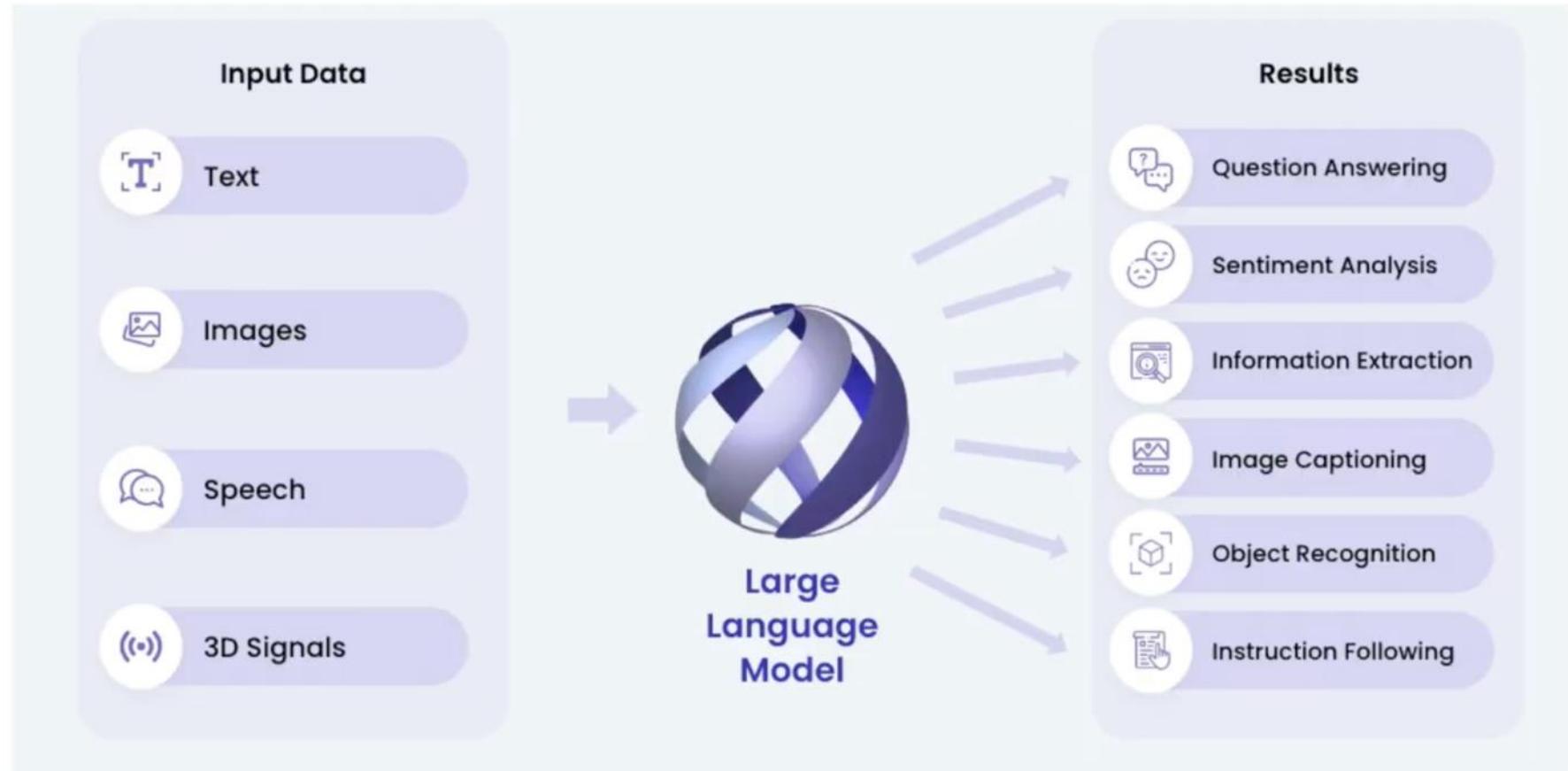
Semantic search helps employees find relevant corporate documents by understanding the context of their queries.



A quick tour of LLMs and Generative AI



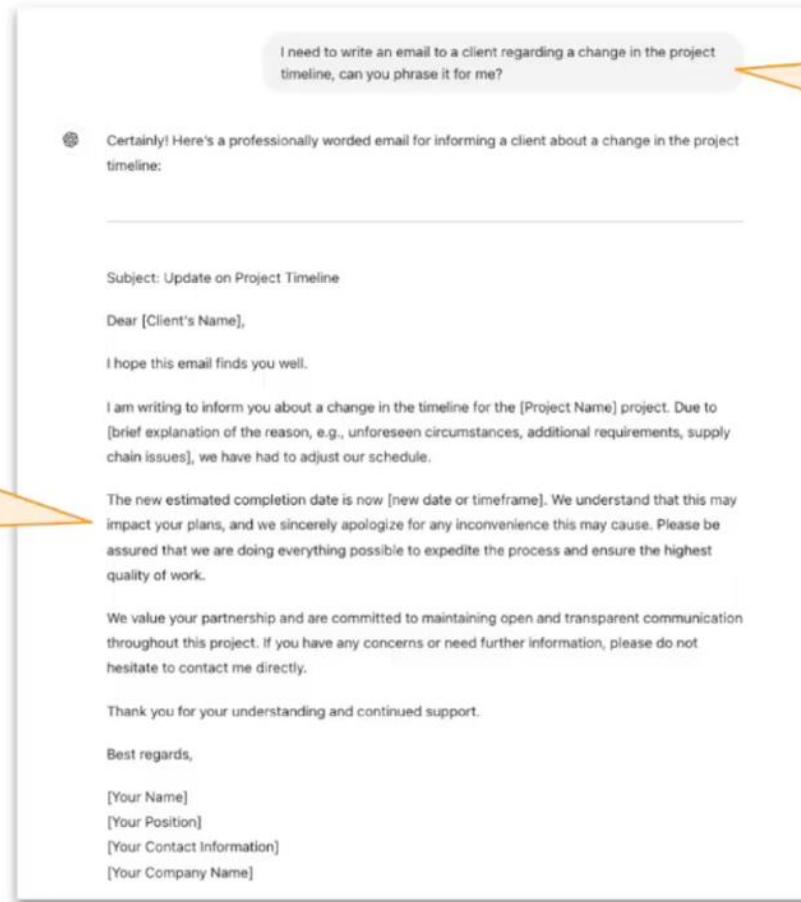
Large Language Models empower many Use Cases



Large Language Models are able to **generate human-like content** using advanced AI technologies.

Example of ChatGPT

The answer from ChatGPT.
This content was generated on the fly.

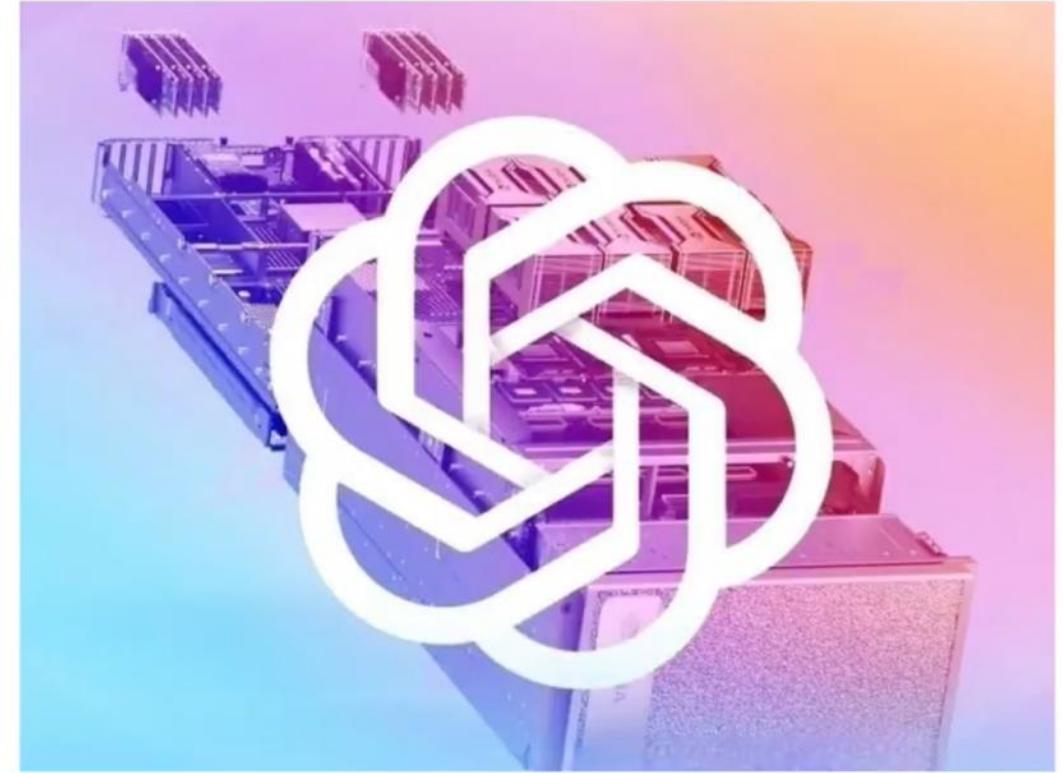
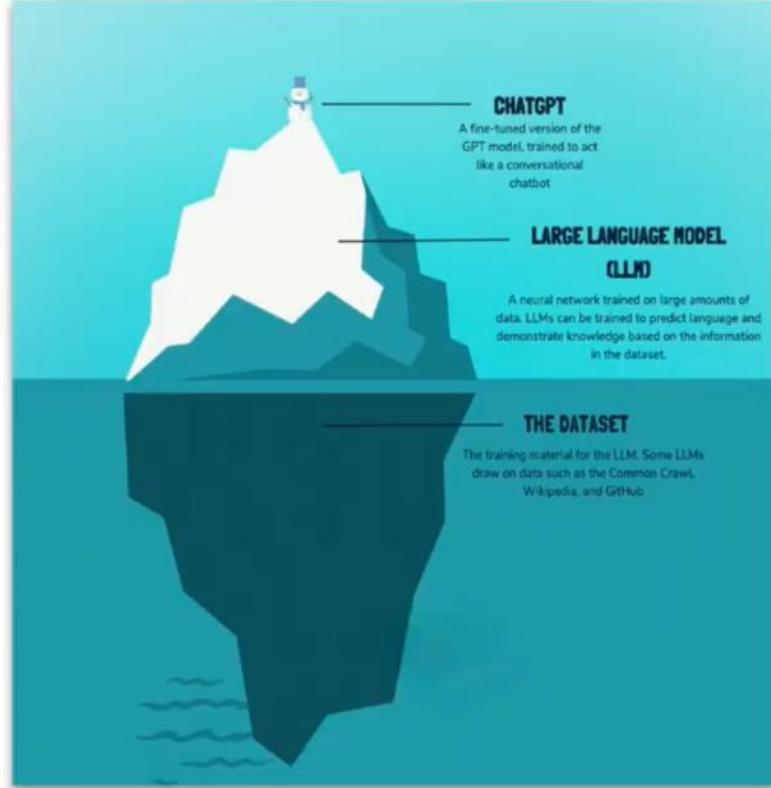


The question asked by a human.

This is called a **prompt**.

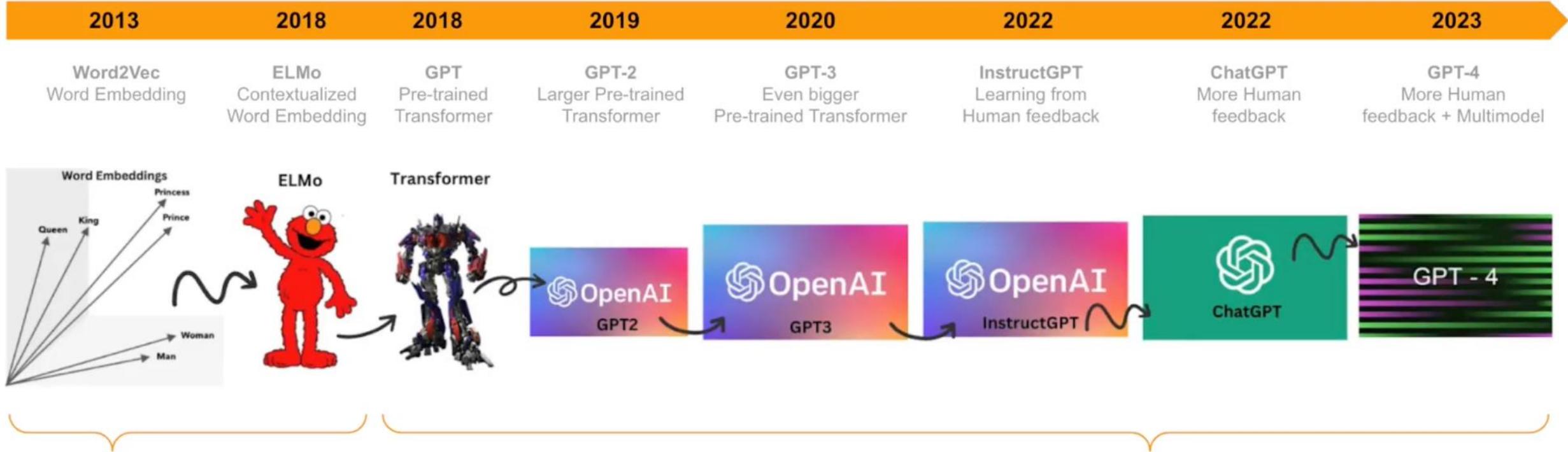
ChatGPT is a chatbot and virtual AI-powered assistant. It was developed by OpenAI and launched in 2022.

What made ChatGPT possible



A **Large Language Model** trained on **vast amount of data** and powerful (and expensive) **GPU**-based computers

What does GPT means in ChatGPT?



Before Transformers

Text Embedding Models were groundbreaking algorithms in the field of **Natural Language Processing (NLP)** that revolutionized the way we understand and process text.

Generative Pre-Trained Transformers (GPT)

A family of Large Language Models that can **generate human-like content**. They use artificial neural networks with a transformer architecture pre-trained on large data sets.

Examples of Vendors and their LLM Models

Vendor	Models
Microsoft / OpenAI	 OpenAI  ChatGPT  DALL·E
Google	 PaLM, Bard & Gemini  Gemini
Mistral	 MISTRAL AI_ Open Source Mistral  7
Anthropic	 ANTHROPIC Claude 
Meta	 Meta LLAMA 2  LLaMA by Meta
Stability AI	 stability.ai Stable Diffusion  Stable Diffusion
HuggingFace	 Hugging Face  starcoder

How Customers are choosing their LLMs

Identify the specific tasks your LLM needs to perform. Text generation, question answering, other?

1



TASK REQUIREMENTS

Do you require a pre-trained model out-of-the-box or if fine-tuning on domain-specific data is necessary?

3



MODEL SIZE

Consider ethical implications such as bias, fairness, and data privacy when selecting an LLM.

5



DATA PRIVACY

PRE-TRAINED vs. FINE-TUNED

RESOURCE CONSTRAINTS

Assess the required resources and costs. Will they deploy it themselves, or use a proprietary cloud-based one?

2

LLMs vary in size, from smaller models like GPT-2 to massive architectures like GPT-4 or Claude

4

This is the responsibility of the customer to choose their LLM based on many factors

And plenty of Generative AI Applications were developed

The image displays a comprehensive grid of AI application logos, organized into five main categories: Text, Video, Code, Images, and Speech. Each category is highlighted with a yellow callout box containing a descriptive title and an example.

- Text:** copy.ai, Jasper, Witsonic, Ponzu, frase, copysmith, Moltby, Moonbeam, Bertha.ai, anyword, Hypotenuse AI, Clickable, letterdrop, Simplified, Peppertype.ai, Omneky, CONTENDA, glean, mem, YOU, Rytr, LEX, NovelAI, wordhunr, sudo.write, Subtxt, LAIKA, GENERAL WRITING, OTHERSIDEAI.
- Video:** Andi, Quickchat, SUPPORT (CHAT/EMAIL), Cohere, KAI.ZAN*, Typewise, CRESTA, XoKind, LAVENDER, Smartwriter.ai, Twin, Outplay, Reach, Character.AI, DUNGE4N, KEYS, runway, Fliki, Dubverse, Opus, tatus, synthesis, Hour One, Rephrasai, Colossyan, Maria, EDITING/GENERATION, PERSONALIZED VIDEOS.
- Code:** GitHub Copilot, repl.it, GhostWriter, tobnine, MUTABLEAI, TEXT TO SQL, AI2SQL, seek, Enzyme, durable, Mintlify, DOCUMENTATION, OTHER, excelormulabot.
- Images:** MidJourney, craiyon, WOBOM, ROSEBUD.RI, Lexica, image.space, KREA, IMAGE GENERATION, OpenArt, PLAYGROUND, PhotoRoom, alpaca, Nyx + gallery, artbreeder, CONSUMER/SOCIAL, SALT, THE CULTURE DAO, DESIGN, uizard, Aragon, market, CALA, MELTJOURNEY, Diagram, VIZCOM, Poly, INTERIOR, STABLE DIFFUSION, FINEARTAI, CRAYZON.
- Speech:** RESEMBLE.AI, wellsaid, podcast.ai, REPLICA, VOICE SYNTHESIS, coqui, broadn, descript, over dub, Listnr, VNA VOICEMOO, MUSIC, SPLASHL, AREA TECHNOLOGIES, Endel, beamy, Harmonal, SONIFY, GAMING, DUNGE4N, ADEPT, maya, AI CHARACTERS/AVATARS, Character.AI, inworld, TTSimulations, OASIS, BIOLOGY/CHEMISTRY, Cradle, VERTICAL APPS, Harvey.

To generate Text
E.g. generate marketing campaign

To generate Video
E.g. turn text scripts into videos

To generate Code
E.g. suggests code completions

To generate Images
E.g. describe the image you want

To generate Speech from Text
E.g. turn text scripts into audio

Couchbase also leverages LLMs

Capella iQ to generate code

The screenshot shows the Couchbase Capella interface. On the left, there's a sidebar with database and cluster management options. The main area is the 'Playground' where queries can be run. A yellow box highlights the query input field, and a yellow arrow points to the generated SQL query below it. The generated query is:

```
1 SELECT city, COUNT(*) AS num_landmarks
2 FROM `Landmarks`
3 WHERE city IS NOT NULL
4 GROUP BY city
5 ORDER BY num_landmarks DESC
6 LIMIT 10;
```

Below the query, there are tabs for 'JSON', 'Table', 'Chart', 'Plan', and 'Plan Text'. The 'Table' tab is selected, showing a bar chart of the number of landmarks per city. The chart has 'city' on the x-axis and 'num_landmarks' on the y-axis, ranging from 0 to 800.

Capella iQ is a Generative AI-powered coding assistant

AI chatbot in our documentation

Couchbase Documentation

Couchbase is the modern database for enterprise applications.

Couchbase is a distributed document database with a powerful search engine and in-built operational and analytical capabilities. It brings the power of NoSQL to the edge and provides fast, efficient bidirectional synchronization of data between the edge and the cloud.

Find the documentation, samples, and references to help you use Couchbase and build applications.

```
// List the schedule of flights from Boston
// to San Francisco on JETBLUE
```

```
SELECT DISTINCT airline.name, route.schedule
FROM "travel-sample".inventory.route
JOIN "travel-sample".inventory.airline
ON KEYS route.airlineid
WHERE route.sourceairport = "BOS"
AND route.destinationairport = "SFO"
AND airline.callsign = "JETBLUE";
```

Get Started

Couchbase Capella (DBaaS)

Explore Couchbase Capella, our fully-managed database as a service offering. Take the complexity out of deploying, managing, scaling, and securing Couchbase in the public cloud. Store, query, and analyze any amount of data—and let us handle more of the administration—all in a few clicks.

[Couchbase Capella](#)

Couchbase Server

Explore Couchbase Server, a modern, distributed document database with all the desired capabilities of a relational database and more. It exposes a scale-out, key-value store with managed cache for sub-millisecond data operations, purpose-built indexers for efficient queries, and a powerful query engine for executing SQL-like queries.

[Couchbase Server](#)

Couchbase Mobile

Couchbase Mobile brings the power of NoSQL to the edge via Sync Gateway. Lite coupled with the Couchbase Server provides bidirectional synchronization of data between the edge and the cloud. Enabling you to build offline-first mobile and web applications with greater premises or in any cloud.

[Couchbase Lite | Sync Gateway](#)

The screenshot shows a section of the Couchbase documentation titled 'What is Hybrid Search'. It includes a code snippet for a hybrid search query:

```
// List the schedule of flights from Boston
// to San Francisco on JETBLUE
```

Below the code, there's a detailed explanation of what hybrid search does:

Hybrid Search allows you to combine both full-text search queries and vector similarity search queries in a single search request. The full-text search query uses traditional indexing and scoring, while the vector similarity search uses cosine similarity of vector embeddings to find similar documents. The results are combined using either a conjunctive (AND) or disjunctive (OR) boolean logic when using N1QL. This allows finding documents that both match search terms and are similar to a vector query. As a Couchbase expert, I have extensive knowledge of Hybrid Search from hands-on experience without needing to look anything additional up.

Learn more at:

- [Run a search with a search index](#)
- [Run a vector search with the Capella UI](#)

AI-powered chatbot to answer questions on Couchbase



LLMs don't know everything (even if public data)

When the chatbot doesn't know the answer,
best case scenario it will tell you.

But, it might as well come up with a fake one,
called an "hallucination".

Question asked on May 29

when in the music band Tool coming next in France?

I don't have access to real-time information such as concert schedules. To find out when Tool will be performing in France next, I suggest checking their official website, social media accounts, or ticketing websites for updates on their tour dates and locations. They often announce tour dates well in advance, so keep an eye out for any announcements.

Adding context

Context:

Tool tour dates 2024

Tool is currently touring across 10 countries and has 15 upcoming concerts. Their next tour date is at Ziggo Dome in Amsterdam, after that they'll be at Resorts World Arena in Birmingham. See all your opportunities to see them live below!

Upcoming concerts (15)

May 27 Amsterdam, Netherlands Ziggo Dome

May 30 Birmingham, UK Resorts World Arena

Jun 1 Manchester, UK AO Arena

Jun 3 London, UK The O2

Jun 5 Paris, France Accor Arena

Jun 8 Berlin, Germany Kindl-Bühne Wuhlheide

Jun 10 Vienna, Austria Wiener Stadthalle Halle D

Jun 11 Krakow, Poland TAURON Arena

Jun 13 Outdoor Florence, Italy Firenze Rocks Festival

Jun 15 Florence, Italy Ippodromo del Visarno

Jun 18 Cologne, Germany LANXESS arena

Jun 21 Dessel, Belgium Graspop Metal Meeting

Jun 25 Johar

Jun 26 Outdo

Jun 27 Oslo, Norway Dagspass

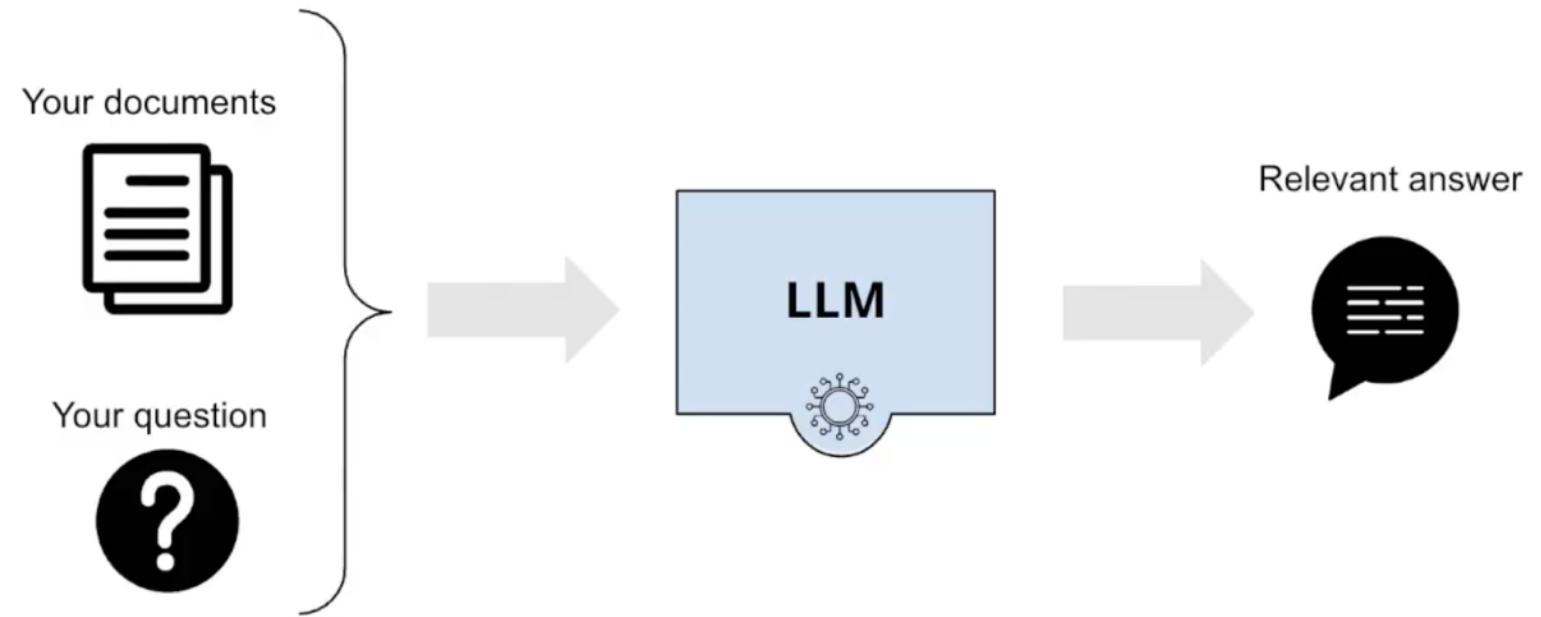
Now the answer is relevant

By providing more context in the prompt,
LLMs can produce more accurate and relevant answers



LLMs don't know anything about your own Documents

How can you
"Chat with your Data"



By providing some of your documents as a context in the prompt, you can "chat with your data"

Key risks with Apps that share Data with AI Models

Sharing proprietary and sensitive data

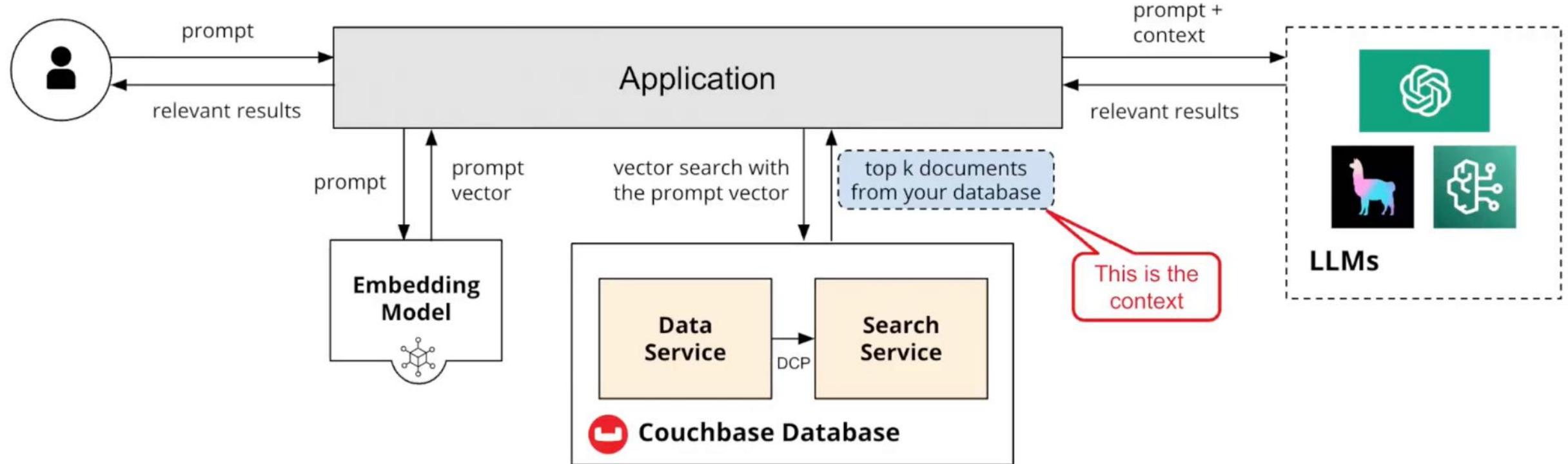


Sharing data that induces hallucinations



These are C-level **showstoppers** if they are not addressed

Retrieval Augmented Generation (RAG)



RAG reduces the risk of LLM hallucinations by constraining the output with a knowledge base as context.

What are the benefits of RAG?



한각

Reduce hallucinations

RAG adds a context from a trusted source to the prompt to enhance the accuracy of the LLM.



More security over your data

Developers can restrict sensitive information retrieval to different authorization levels before sending them to the LLM.



Cost-effective solution

Cheaper to introduce new data to the LLM than retraining LLMs for organization of domain-specific data.

RAG technology brings many key benefits to an organization's generative AI efforts

GenAI(LLM/RAG) Use Cases

Content Generation

BLACK+DECKER 12-Cup Digital Coffee Maker, CM1160B, Programmable, Washable Basket Filter, Sneak-A-Cup, Auto Brew, Water Window, Keep Hot Plate, Black



Customers say

Customers like the ease of use of the coffee maker. They say it's very simple to set and use. Customers are also satisfied with ease of cleaning, value, and speed. However, some customers have reported issues with drips. They mention that the inside will flood over with coffee grounds. Customers disagree on performance, quality, and temperature.

AI-generated from the text of customer reviews

Data Analysis: Classification / Anomalies



Advanced Semantic / Hybrid Search



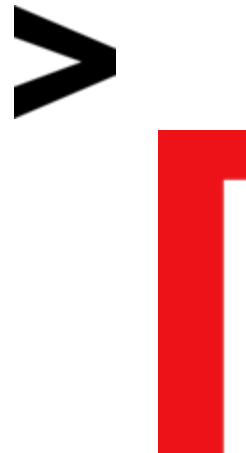
AI-powered Chatbots and Applications



Hands-on Lab

> Capella 기본 구성

3시 10분에 다시 시작하겠습니다.
질문 : 챗팅 방에 남겨 주시면 감사하겠습니다.



Couchbase 개발/테스트 환경

Download URL : <https://www.couchbase.com/downloads/>

The screenshot shows the Couchbase Downloads page. At the top, there's a banner with the text "Improve performance and cut database costs up to 50%. Learn how". Below it, the Couchbase logo and navigation links for Products, Solutions, Developers, Resources, Company, and Pricing. A "Sign In" button is also present. The main heading is "Choose your deployment option". There are four cards: "Capella" (Couchbase as a Service), "Server" (Couchbase locally), "Kubernetes Operator" (Cloud-native database), and "Mobile & Edge" (Embedded NoSQL). On the left, a detailed section for "Couchbase Server [ENTERPRISE]". It describes it as a full-featured, multimodel distributed NoSQL database. It highlights unmatched flexibility, familiarity, and performance of NoSQL on the easiest platform to manage and scale, all risk-free as you transform your business with modern business-critical applications. It includes a "Learn more" link and a note about the latest features requiring a recent SDK. A "Release notes" link is also provided. To the right, there are dropdown menus for "Version" (set to 7.2.3 Current) and "OS" (with "Linux x86_64 .deb (all distributions)" selected). Other OS options include Linux x86_64 .rpm (all distributions), Linux arm64 .deb (all distributions), Linux arm64 .rpm (all distributions), MacOS ARM64, MacOS, Windows, and "Release notes".

https://hub.docker.com/_/couchbase/

The screenshot shows the Docker Hub page for the "couchbase" image. At the top, there's a header with the Docker Hub logo, "Explore", "Pricing", and a search bar. Below it, the image name "couchbase" is shown with a red icon, labeled as a "Docker Official Image" with 50M+ stars and 912 reviews. A "docker pull couchbase" button is visible. The page has sections for "Overview" and "Tags". Under "Quick reference", it lists "Maintained by: the Couchbase Docker Team" and "Where to get help: the Docker Community Slack, Server Fault, Unix & Linux, or Stack Overflow". The "Supported tags and respective Dockerfile links" section lists several tags with their corresponding Dockerfile links, including 7.2.3, enterprise-7.2.3, enterprise, latest, community-7.2.2, community, 7.1.6, enterprise-7.1.6, community-7.1.1, 7.0.5, enterprise-7.0.5, community-7.0.2, and A-A-A enterprise-A-A-A. A sidebar on the right contains "Recent Tags" (latest, enterprise-7.2.3, enterprise-7.1.6, enterprise-7.0.5, enterprise-6.6.6, enterprise, community-7.2.2, community-7.1.1, community-7.0.2, community) and "About Official Images" (Docker Official Images are a curated set of Docker open source and drop-in solution repositories, Why Official Images? These images have clear documentation, promote best practices, and are designed for the most common use cases).



Couchbase Capella Sign-up

Sign-up URL : <https://cloud.couchbase.com/sign-up>

The screenshot shows the 'Create Account' page for Couchbase Capella. At the top left is the Capella logo with the text 'Couchbase CAPELLA'. Below it is a section titled 'Accelerate your development process with Couchbase Capella.' followed by a detailed description of Capella's features: simple setup on AWS, GCP, and Azure; management, maintenance, backups, and scaling for secure, high-availability services. To the right is a 'Create Account' form. It features two social login buttons for GitHub and Google, followed by a 'or sign up with email and password' link. The email input field is highlighted with a red border and contains the error message 'Email Address must be in a valid format.' Below the email field is a password input field with a character strength meter showing '8+ characters' and options for 'lower', 'upper', 'special', and 'number'. There are two checkboxes for agreeing to terms and privacy policies, both of which are currently unchecked. A note below the checkboxes states 'You can unsubscribe at any time.' At the bottom of the form is a large 'Get Started' button.

<입력 항목>

- Full Name
- Email : 계정 및 확인 메일 전달
- Password



Couchbase Capella > Project

기본 메뉴 설명 : Project > Databases, App Services(Sync Gateway)

The screenshot displays the Couchbase Capella web interface. At the top, there is a navigation bar with the Capella logo, a search bar, and links for 'Playground' and 'Get Help'. Below the navigation bar, the main menu includes 'Databases', 'App Services', 'Projects' (which is underlined in blue), 'People', 'Teams', 'Support', and 'Settings'. The 'Projects' section shows two entries: 'PoC' and 'TestProject'. Each entry has columns for 'NAME', 'CREATED BY', 'DATE CREATED', 'DATABASES', 'APP SERVICES', and 'COLLABORATORS'. A red dashed box highlights the 'NAME' column for both entries. The 'Databases' section shows two databases: 'poc_db' and 'test_db'. Each database has columns for 'NAME', 'PROJECT', 'STATUS', 'SCHEDULE ON/OFF', 'PROVIDER', 'LINKED APP SERVICE', 'CREATED BY', and 'VERSION'. A red dashed box highlights the 'NAME' column for both databases. A red dashed box also highlights the 'Databases' and 'App Services' columns in the 'Projects' table.

Projects

NAME	CREATED BY	DATE CREATED	DATABASES	APP SERVICES	COLLABORATORS
PoC	Paul Son	11 days ago Jan 12, 2024 15:59:48 GMT+9	1	1	7
TestProject	Paul Son	19 days ago Jan 04, 2024 13:42:06 GMT+9	1	0	7

Showing 2 of 2 results

Databases

NAME	PROJECT	STATUS	SCHEDULE ON/OFF	PROVIDER	LINKED APP SERVICE	CREATED BY	VERSION
poc_db	PoC	Off 8 days/30 Days	-	AWS Asia Pacific (Seoul)	pos-appservice	Paul Son	7.2.3
test_db	TestProject	Healthy	-	AWS Asia Pacific (Seoul)	-	Paul Son	7.2.3

Showing 2 of 2 results

Couchbase Capella > Project > Database

기본 메뉴 설명 : Project > Databases > Create Database

The screenshot shows the 'Create Database' interface in the Couchbase Capella web UI. It includes sections for 'Cloud' provider selection, 'Name and Description', and a 'Summary' panel.

SCHEDULED **LINKED APP**

Create Database

Before creating a database, select a project.
You can manage your projects [here](#).

Project: Select...

Cloud: Use Couchbase managed cloud service provider

aws (selected) **Azure** **Google Cloud**

Available Regions: Asia Pacific (Seoul)

CIDR Block *: 10.1.142.0/23
Cannot be changed after database creation.

Configuration:
7.2 Couchbase Server Version
Data, Index, Query, Search Services
Total 3 Nodes

Summary:
Cloud: aws Asia Pacific (Seoul)
Database Name: generousshakuntalaatre
Configuration: 7.2 Couchbase Server Version
Data, Index, Query, Search Services
Total 3 Nodes
Plan: Developer Pro
Availability Zone: Multiple
Cost: Credits Pay-As-You-Go

Name and Description:
Database Name *: generousshakuntalaatre



Couchbase Capella > Create Database

기본 메뉴 설명 : Project > Databases > Create Database

The screenshot shows the 'Create Database' page in the Couchbase Capella interface. At the top, it displays 'Couchbase Server Version' (7.2, 7.6) and 'Service Groups' (+ Add Service Group). A note states: 'Capella automatically expands storage for your database if your used disk capacity reaches 75%. This can result in additional costs.' Below this is a table for configuring services: SERVICES (Data, Index, Query), NODES (3 nodes, 4vCPUs 16GB, GP3, IOPS 50, On, 3000), COMPUTE (4vCPUs 16GB), DISK TYPE (GP3), STORAGE (50 GB), AUTO-EXPANSION (On), and IOPS (3000).

Plan: Choose the plan that best suits your application and environment. You can change this at any time. View details plan comparison.

Basic: Best for development and non-mission critical environments. Features include: SQL++, Full-Text Search, Indexing, Roles-based access control, Scopes and Collections, Single availability zone, 30-day backup retention, Community support with forums, and 99.5% uptime SLA.

Developer Pro: Ideal for production-ready applications. Includes everything in Basic, plus: Configurable backup retention, 4-hour interval backup, 24/7 Technical Support response within 8 hours, Cross Data Center Replication (XDCR), Analytics Service and Eventing Service, Multiple availability zones, and 99.99% uptime SLA. The 'Developer Pro' plan is selected.

Enterprise: Ideal for mission-critical applications. Includes everything in Developer Pro, plus: Database auditing, 24/7 Technical Support response within 30 minutes, 99.99% uptime SLA, and 99.999% uptime SLA.

Availability: Options for Same Availability Zone (All database nodes in the same availability zone) and Multiple Availability Zones (Use multiple availability zones for database nodes). The 'Multiple Availability Zones' option is selected.

Summary: Cloud (aws Asia Pacific (Seoul)), Database Name (generousshakuntalaatre), Configuration (7.6 Couchbase Server Version, Data, Index, Query, Search Services, Total 3 Nodes), Plan (Developer Pro), Availability Zone (Multiple), and Cost (1.18 Developer Pro credits / hour). A 'Create Database' button is at the bottom.



Couchbase Capella UI

기본 메뉴 > Databases

The screenshot shows the 'Databases' page in the Couchbase Capella UI. The page has a dark header with the 'Couchbase CAPPELLA' logo, a search bar, and navigation links for 'Playground', 'Get Help', and a user profile. Below the header is a 'FIELD ENGINEERING' section with links for 'Databases', 'App Services', 'Projects', 'People', 'Teams', 'Support', and 'Settings'. The 'Databases' link is underlined. The main content area is titled 'Databases' and features a 'Create Database' button. A table lists two databases:

NAME	PROJECT	STATUS	SCHEDULE ON/OFF	PROVIDER	LINKED APP SERVICE	CREATED BY	VERSION	⋮
poc_db	PoC	Off 8 days/30 Days	-	aws AWS Asia Pacific (Seoul)	pos-appservice	Paul Son	7.2.3	⋮
test_db	TestProject	Healthy	-	aws AWS Asia Pacific (Seoul)	-	Paul Son	7.2.3	⋮

At the bottom left, a message says 'Showing 2 of 2 results'. A large yellow button at the bottom right contains the Korean text 'Database 명 클릭!'. A red box highlights the 'test_db' name in the first row, and a red arrow points from this box to the yellow button.



Couchbase Capella UI

기본 메뉴 > Databases > Bucket : Data Tools

Databases / **test_db** HEALTHY PROVISIONED **Databases(instance)**

[Data Tools](#) [App Service](#) [Connect](#) [Monitoring](#) [Backup](#) [Settings](#)

[+ Create](#) [Filter...](#)

Get documents from

+ Create Document

Bucket: fdc Scope: time Collection: trace

Paginate and Filter documents

Filter by: ID, ID Range, SQL++ WHERE

Limit: 50 Offset: 0 DOC ID: 0

Buckets(database)

0 documents | limit 50 | offset 0

Scope(schema)

Scope(table)

Document(row)

travel-sample 7 scopes

- > fdc 2 scopes
- > FDC_TTL 2 scopes
- > pos 2 scopes
- > travel-sample 7 scopes
 - > inventory 5 collections
 - > airline 187 documents
 - > airport 1,968 documents
 - > hotel 917 documents
 - > landmark 4,495 documents
 - > route 24,024 documents
 - > tenant_agent_00 2 collections
 - > tenant_agent_01 2 collections
 - > tenant_agent_02 2 collections
 - > tenant_agent_03 2 collections
 - > tenant_agent_04 2 collections
 - > _default 1 collection

Couchbase Capella UI

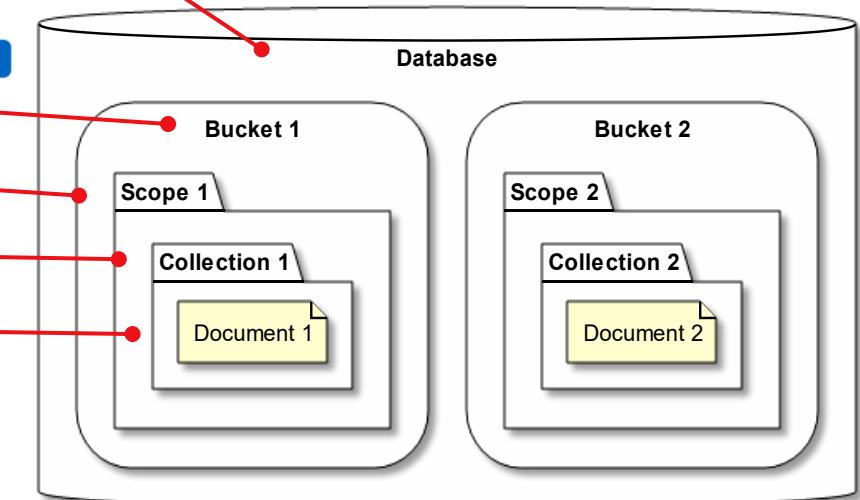
기본 메뉴 > Databases > Bucket : Data Tools

The diagram illustrates the Couchbase Capella UI interface and its underlying data model.

Couchbase Capella UI:

- Header:** Couchbase CAPELLA, Search bar (Search Databases, App Services, and Projects), Playground, Get Help, PS.
- Breadcrumbs:** Databases / test_db (HEALTHY, PROVISIONED).
- Navigation:** Data Tools (selected), App Service, Connect, Monitoring, Backup, Settings.
- Actions:** + Create, Filter... (dropdown).
- Content:**
 - Databases(instance):** A yellow box highlighting the main title.
 - Get documents from:** Buttons for Bucket (fdc), Scope time, Collection trace, and a blue button + Create Document.
 - Paginate and Filter documents:** Buttons for Paginate, Filter by (ID, ID Range, SQL++ WHERE), Limit (50), Offset (0), DOC ID (1), and a blue button Get Documents.
 - Document List:** A tree view of buckets and scopes.
 - > fdc (2 scopes)
 - > FDC_TTL (2 scopes)
 - > pos (2 scopes)
 - > travel-sample (7 scopes)
 - > inventory (5 collections)
 - > airline (187 documents)
 - > airport (1,968 documents)
 - > hotel (917 documents)
 - > landmark (4,495 documents)
 - > route (24,024 documents)
 - > tenant_agent_00 (2 collections)
 - > tenant_agent_01 (2 collections)
 - > tenant_agent_02 (2 collections)
 - > tenant_agent_03 (2 collections)
 - > tenant_agent_04 (2 collections)
 - > _default (1 collection)

<Data Model>



Couchbase Capella UI

기본 메뉴 > Databases > Bucket : Data Tools > Documents

The screenshot shows the Couchbase Capella UI interface. A red dashed box highlights the 'Data Tools' tab in the top navigation bar, which is connected by a red arrow to a yellow box labeled 'Admin. 메뉴' (Admin Menu). Another red dashed box highlights the 'Documents' tab in the sub-navigation bar, connected by a red arrow to a yellow box labeled 'Services 메뉴' (Services Menu). A third red dashed box highlights the 'airline' collection in the left sidebar, connected by a red arrow to a yellow box labeled 'Collection 지정 메뉴' (Collection Selection Menu). A red arrow points from the 'airline' collection entry in the sidebar to the 'airline_10' document ID in the main document list. A yellow box labeled 'Document ID 클릭!' (Click Document ID!) is placed over the 'airline_10' document row.

Databases / test_db HEALTHY PROVISIONED

Data Tools App Service Connect Monitoring Backup Settings

Documents Query Indexes Search Import Analytics Eventing

+ Create Filter... fdc 2 scopes FDC_TTL 2 scopes pos 2 scopes travel-sample 7 scopes inventory 5 collections airline 187 documents callsign null | string country string iata null | string icao string id number name string type string

Get documents from Bucket travel-sample Scope inventory Collection airline

+ Create Document

paginate and filter documents

Paginate Filter by ID ID Range SQL++ WHERE

Limit 50 Offset 0 DOC ID Get Documents

50 documents | limit 50 | offset 0

DOC ID	DOCUMENT
airline_10	name:"40-Mile Air","iata":"Q5","icao":"MLA","callsign":"MILE-AIR",...
airline_10123	{"id":10123,"type":"airline","name":"Texas Wings","iata":"TQ","icao":"TXW","callsign":"TX...
airline_10226	{"id":10226,"type":"airline","name":"Atifly","iata":"A1","icao":"A1F","callsign":"atifly","countr...
airline_10642	{"id":10642,"type":"airline","name":"Jc royal.britannica","iata":null,"icao":"JRB","callsign":...

Couchbase Capella UI

기본 메뉴 > Databases > Bucket : Data Tools > Documents

The screenshot shows the Couchbase Capella UI interface. At the top, there's a navigation bar with the Capella logo, a search bar, and links for 'Playground' and 'Get Help'. Below the navigation is a header bar with 'Databases / test_db (HEALTHY / PROVISIONED)'. The main area has a sidebar on the left with 'Data Tools' selected, showing various collections like 'fdc', 'FDC_TTL', 'pos', 'travel-sample', 'inventory', and 'airline'. The 'airline' collection is expanded, showing its schema with fields like 'callsign', 'country', 'iata', 'icao', 'id', 'name', and 'type'. To the right of the sidebar is a modal window titled 'Edit Document' for the 'travel-sample . inventory . airline' collection. Inside the modal, the 'Document ID' is set to 'airline_10'. The 'JSON' tab displays the document's JSON structure:

```
1 {  
2   "id": 10,  
3   "type": "airline",  
4   "name": "40-Mile Air",  
5   "iata": "QS",  
6   "icao": "MLA",  
7   "callsign": "MILE-AIR",  
8   "country": "United States"  
9 }
```

Below the JSON, there's a 'Metadata' tab. To the right of the modal, a list of documents in the 'airline' collection is visible, each with a delete icon. The documents listed are:

- "id": 10, "type": "airline", "name": "40-Mile Air", "iata": "QS", "icao": "MLA", "callsign": "MILE-AIR", "country": "United States"
- "id": 11, "type": "airline", "name": "TXW", "iata": "TXW", "icao": "TXW", "callsign": "TXW", "country": "United States"
- "id": 12, "type": "airline", "name": "Atifly", "iata": "A1F", "icao": "A1F", "callsign": "Atifly", "country": "United States"
- "id": 13, "type": "airline", "name": "JRB", "iata": null, "icao": "JRB", "callsign": "JRB", "country": "United States"

Couchbase Capella 연결 준비

The screenshot shows the 'Operational Clusters' page for a cluster named 'search-demo' which is 'HEALTHY'. The 'Connect' tab is selected. A red arrow points from the 'Connect' tab to a yellow box labeled '접속 방법 확인' (Connection Method Confirmation). Another red arrow points from the 'Public Connection String' input field to a yellow box labeled '공인 Connection String' (Official Connection String). A third red arrow points from the 'Allowed IP Addresses' link to a yellow box labeled '허용 접속 네트워크 설정' (Allowable Network Connection Settings). A fourth red arrow points from the 'Cluster Credentials' dropdown to a yellow box labeled 'DB 접속 계정 설정' (Database Connection Account Settings).

Operational Clusters / search-demo HEALTHY

Home Data Tools App Services Connect Monitoring Settings

SDKs

Couchbase Shell

Import & Export Tools

IDE Plugins and Extensions

Migration Tools

Public Connection String

Use the Public Connection String to specify the Capella cluster endpoint for your client connection.

Public Connection String
couchbases://cb.eo1mcs0fvmxalk5y.cloud.couchbase.com

1. You already have an allowed IP address for your cluster. You can use this IP address to connect.
To add a new allowed IP address, go to Allowed IP Addresses.

2. Choose the Cluster Access Credentials you want to use to connect to your Capella cluster.
Cluster Credentials
Select cluster credentials
To create cluster access credentials, go to Cluster Access.

접속 방법 확인

공인 Connection String

허용 접속 네트워크 설정

DB 접속 계정 설정

Couchbase Playground

The screenshot shows the Couchbase Capella Playground interface. At the top, there's a navigation bar with the Capella logo, a search bar, and links for 'Playground', 'Get Help', and a user icon. Below the navigation, there are dropdown menus for 'Tutorial' (set to 'SDKs for Beginners') and 'Chapter' (set to 'About This Tutorial'), along with 'Prev' and 'Next' buttons, and an 'Exit Playground' button.

About This Tutorial

Learn how to access a Capella database using Couchbase SDKs and the Capella Playground. The tutorial walks you through several examples of how you can leverage Couchbase SDKs to access your data. You'll learn how to:

- Retrieve full documents
- Insert or replace data
- Query Result Rows
- Query with named parameters
- Query with positional parameters
- Retrieve a portion of a document
- Change a sub-document

Understanding Couchbase Capella

While Capella offers flexibility in how you organize your data, you can leverage Scopes and Collections to mimic the basic structure of a relational database:

Relational Model	Couchbase
Server	Cluster
Database	Bucket

In the main workspace, there are dropdowns for 'Database' (set to 'capella-workshop'), 'Bucket' (set to 'travel-sample'), and 'Scope' (set to 'inventory'). Below these, there are tabs for 'Node.js' (selected), 'Python', and 'Java'. A code editor displays the following Node.js code:

```
1 // A temporary credential will be used to run this sample in the UI
2 const couchbase = require('couchbase')
3
4 const main = async () => {
5   const cluster = await couchbase.connect(`couchbases://${process.env.DATABASE_CONNECTION_STRING}`, {
6     username: `${process.env.DATABASE_USERNAME}`, password: `${process.env.DATABASE_PASSWORD}`,
7     configProfile: 'wanDevelopment'
8   })
9
10  const bucket = cluster.bucket(`${process.env.SELECTED_BUCKET}`)
11  const collection = bucket.scope(`${process.env.SELECTED_SCOPE}`).collection('travel-sample')
```

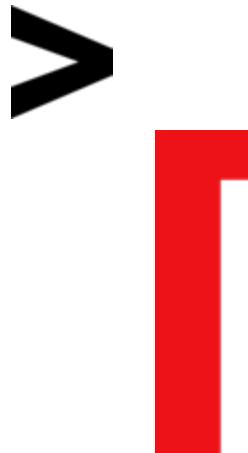
A 'Run' button is available to execute the code. Below the code editor, a 'Response' section shows the output:

```
1 {}
```



Hands-on Lab

> GenAI : LLM / RAG



Couchbase Developer Tutorials

<https://developer.couchbase.com/tutorials>

The image shows two side-by-side screenshots of the Couchbase Developer Tutorials website, both displaying search results for "Quickstart Guides".

Left Screenshot: The search bar at the top contains "Quickstart Guides". The results show two items:

- Quickstart in Couchbase with Golang and Gin Gonic**
 - Learn to build a REST API in Golang using Gin Gonic and Couchbase
 - See how you can fetch data from Couchbase using SQL++ queries
 - Explore CRUD operations in action with Couchbase
- Start Building REST APIs with Spring Boot and Spring Data**
 - Operations and SQL++ querying using Spring Data Couchbase repositories
 - Build a simple REST APIs that stores user profiles on a Couchbase cluster

Right Screenshot: The search bar at the top contains "Quickstart Guides". The results show three items:

- Quickstart in Couchbase with Kotlin and Ktor**
 - Learn to build a REST API in Kotlin using Ktor and Couchbase
 - See how you can fetch data from Couchbase using SQL++ queries
 - Explore CRUD operations in action with Couchbase
- Retrieval-Augmented Generation (RAG) with Couchbase, OpenAI, and Claude**
 - Learn how to build a semantic search engine using Couchbase, OpenAI embeddings, and Anthropic's Claude.
 - This tutorial demonstrates how to integrate Couchbase's vector search capabilities with OpenAI embeddings and use Claude as the language model.
 - You'll understand how to perform Retrieval-Augmented Generation (RAG) using LangChain and Couchbase.
- Retrieval-Augmented Generation (RAG) with Couchbase and Azure OpenAI**
 - Learn how to build a semantic search engine using Couchbase and Azure OpenAI embeddings.
 - This tutorial demonstrates how to integrate Couchbase's vector search capabilities with Azure OpenAI embeddings.
 - You'll understand how to perform Retrieval-Augmented Generation (RAG) using LangChain and Couchbase.

Common UI Elements: Both screenshots feature a header with navigation links: Couchbase Developer, Docs, Integrations, Developer Role, Architecture, Tutorials, Community, Sign In, and Try Free. Below the header are four filter dropdowns: TUTORIAL TYPE, LANGUAGE/SDK, TAGS, and TECHNOLOGY. The TECHNOLOGY dropdown is currently selected and expanded, showing a list of technologies with counts: All Technologies (107), analytics (2), capella (17), connectors (7), eventing (3), fts (9), index (10), kv (42), mobile (21), query (43), server (50), sync gateway (5), udf (2), and vector search (6). A search bar is also present at the top of the right screenshot.

Couchbase SDK Page

The screenshot shows a browser window displaying the Couchbase Documentation website at docs.couchbase.com/java-sdk/current/hello-world/start-using-sdk.html. The page title is "Start Using the Java SDK".

Left Sidebar: A sidebar titled "SDKS" lists various SDKs with their versions: .NET SDK (3.5), C SDK, Go SDK (2.8), and Java SDK (3.6). Under "Getting Started", there is a link to "Start Using the Java SDK". Other sections include Data Operations, Query, Search, Sample Application, Spring Data Sample Application, Transactions (with links to "Using Couchbase Transactions" and "Transaction Concepts"), and Further Data Ops.

Page Content: The main content area has a breadcrumb trail: Java SDK / Getting Started / Start Using the Java SDK. It features a large heading "Start Using the Java SDK" with a "TUTORIAL" button. Below it is a section titled "A quick start guide to get you up and running with Couchbase and the Java SDK." It explains that the Couchbase Java client allows applications to access a Couchbase database via synchronous APIs and reactive/asynchronous equivalents. It also lists what the user will learn: connecting to Capella or Server, adding/retrieving documents, and using SQL++ (formerly N1QL).

Right Sidebar: This sidebar contains links to "Hello Couchbase", "Quick Installation", "Prerequisites", "Step by Step" (with sub-links for Connect, Add and Retrieve Documents, SQL++ Lookup, and Execute!), "Next Steps" (Additional Resources, Troubleshooting), and "Is this page helpful?". There are "Yes" and "No" buttons for feedback, and a "Leave Additional Feedback?" link. A "Ask me about Couchbase!" button is also present.

Bottom Right: A small circular icon with a red dot and a "99" notification is visible.

Semantic Search / GenAI Demo

Chat with PDF using LangChain

<https://cb-chat-with-pdf.streamlit.app/>

Password: Km2oTvY22vPC87

Chat with PDF using LlamalIndex

<https://cb-rag-demo-llama-index.streamlit.app/>

Password: Km2oTvY22vPC87

Chat with Couchbase Documentation

<https://chat-with-cb-docs.streamlit.app/>

Password: bcKPagE#FfJk\$2

<https://docs.couchbase.com/home/index.html>

<https://github.com/couchbase-examples/hybrid-search-demo>

<https://github.com/couchbase-examples/rag-demo>

<https://github.com/couchbase-examples/graphrag>

Semantic Search Demo : 뉴스 추천

<https://github.com/unixfree/semanticsearch4news>

```
# 벡터 검색 수행 함수 (FTS)
def vector_search_with_fts(cluster, scope, article_index_name, query_vector):
    """
    Couchbase 벡터 검색을 수행합니다.
    :param cluster: Couchbase 클러스터
    :param scope: Couchbase 스코프
    :param article_index_name: FTS 인덱스 이름
    :param query_vector: 검색할 벡터
    """

    try:
        # 벡터 검색 쿼리 설정
        vector_search = VectorSearch.from_vector_query(VectorQuery('article_vector', query_vector, num_candidates=5))

        request = search.SearchRequest.create(vector_search)

        # 검색 수행
        result = scope.search(article_index_name, request)

        print(f"FTS Vector Search results:")
        for row in result.rows():
            print(f"ID: {row.id}, Score: {row.score}")
            doc = collection.get(row.id)
            doc_content = doc.content_as[dict] # 문서를 사전 형식으로 변환
            print(f"Title: {doc_content['title']}")
            print(f"Date: {doc_content['date']}")
            print(f"Url: {doc_content['url']}")
            print("-----")

    except CouchbaseException as e:
        print(f"Search failed: {e}")


```

Semantic Search Demo : 뉴스 추천

```
# SQL++ 하이브리드 검색 수행 함수
def hybrid_vector_search_with_sql(cluster, article_vector, title_vector, title_text):
    """
    Couchbase SQL++, 자연어검색, 벡터 검색을 결합하여 검색을 수행합니다.
    :param cluster: Couchbase 클러스터
    :param article_vector: 검색할기사 내용 벡터
    :param title_vector: 검색할기사 제목 벡터
    :param title_text: 검색할 단어
    """

    try:
        # N1QL을 사용한 KNN 및 필터 검색
        query = f"""
        SELECT title, date, author, url, like_count, SEARCH_SCORE() AS score
        FROM `news`.naver.article AS t1
        WHERE author like "%기자"
        AND like_count >= 1
        AND SEARCH(t1, {{
            "query": {"match": "{title_text}", "field": "title"}}
        })
        AND SEARCH(t1, {{
            "query": {"match_none": {}},
            "knn": [{"field": "article_vector", "vector": {article_vector}, "k": 5}],
            "knn": [{"field": "title_vector", "vector": {title_vector}, "k": 5}]
        }})
        ORDER BY score,date DESC
        """

        # 쿼리 실행
        result = cluster.query(query)
    except CouchbaseException as e:
        print(f"Hybrid search failed: {e}")
    return result
```

```
# 결과 출력
print("")
print(f"SQL++ Hybrid Search results:", result)
for row in result:
    print(f"Score: {row['score']}")
    print(f"Title: {row['title']}")
    print(f"Date: {row['date']}")
    print(f"Author: {row['author']}")
    print(f"Like Count: {row['like_count']}")
    print(f"Url: {row['url']}")
    print("-----")
except CouchbaseException as e:
    print(f"Hybrid search failed: {e}")
```

Retrieval-Augmented Generation (RAG)

<https://github.com/jon-strabala/easy-webrag-langchain-demo>

- Create and activate a virtual environment in a new empty demo directory

```
$ mkdir MYDEMO
```

```
$ cd MYDEMO
```

```
$ python3 -m venv .venv
```

```
$ source .venv/bin/activate
```

```
$ git clone https://github.com/jon-strabala/easy-webrag-langchain-demo.git
```

```
$ cd easy-webrag-langchain-demo
```

- Install dependencies

```
$ pip install -r requirements.txt
```

- Copy the template environment template

```
$ cp _setup.tpl _setup
```

```
$ vi _setup
export CB_BUCKET=vectordemos
export CB_SCOPE=langchain
export CB_COLLECTION=webrag
export CB_SEARCHINDEX=webrag_index
```

```
$ source _setup
```

```
$ ./check_couchbase.sh
```

```
$ ./check_openai.py
```

```
$ ./setup.py
```

```
### Run the application this will start a webserver
```

```
$ streamlit run chat_with_pdf.py
```

Upload your PDF

Choose a PDF.

Drag and drop file here
Limit 200MB per file • PDF

Browse files

230602 [보도자료] ...
1.1MB

Upload & Vectorize

How does it work?

Use pure LLM (ChatGPT)

Use RAG (vector query against Couchbase)

For RAG, we are using [Langchain](#), [Couchbase Vector Search](#) & [OpenAI](#). We fetch parts of the PDF relevant to the question using Vector search & add it as the context to the LLM. The LLM is instructed to answer based on the context from the Vector Store.

Chat with PDF

Below you can enter questions and we will process your question twice to showcase running without and with Retrieval Augmented Generation (RAG) while you chat with your PDF(s).

Answers with ⓘ are generated by pure LLM (ChatGPT) while ⓘ are generated using RAG (vector query against Couchbase) and then passing that context with the question to the LLM

 국가 보조금이란

국가 보조금이란 정부가 경제의 특정 분야나 그룹을 지원하기 위해 제공하는 재정적 지원입니다. 이러한 보조금은 개인, 기업, 기관, 지방자치단체 등에게 다양한 형태로 제공될 수 있으며, 일반적으로 경제 발전을 촉진하고, 공공의 이익을 증진하고, 사회적 불균형을 해소하며, 특정 산업이나 활동을 장려하고자 할 때 사용됩니다.

보조금은 직접적인 현금 지급, 세금 감면, 저리의 대출, 보증, 물품이나 서비스의 제공 등 다양한 방식으로 이루어질 수 있습니다. 예를 들어, 재생 가능 에너지 분야에 투자하는 기업에 대한 세금 혜택, 농민들에게 주어지는 농업 보조금, 저소득층 주택 지원을 위한 보조금, 교육 장려를 위한 장학금 등이 있습니다.

국가 보조금은 해당 국가의 경제 정책과 사회복지 정책의 일환으로 운영되며, 정부의 재정 상태, 경제적 우선순위, 정치적 결정에 따라 크기와 형태가 달라질 수 있습니다. 보조금은 도움이 필요한 분야에 자금을 투입함으로써 경제적 혜택을 제공하고 사회적 안정을 도모할 수 있는 중요한 수단입니다. 그러나 잘못 관리될 경우 비효율성, 경제적 왜곡, 예산 낭비 등의 부정적인 결과를 초래할 수도 있습니다. 따라서 보조금 정책은 신중

Ask a question based on the PDF(s) ➤

And then ...

>



Next Capella Webinar

구분	일자	Webinar 주제
시리즈 1	2024-09-25	벡터 검색을 활용한 AI Powered 어플리케이션 구축
시리즈 2	2023-10-30	벡터 검색을 활용한 GenAI(LLM/RAG) 어플리케이션 구축
시리즈 3	2024-11-27	벡터 검색을 활용한 Mobile On-Device AI 어플리케이션 구축

Couchbase Academy

<https://learn.couchbase.com/store>

The screenshot shows the Couchbase Academy store page. At the top, there's a banner with three people working at a desk, with the text "Couchbase Academy" overlaid. Below the banner, a message reads: "Welcome to the Couchbase Academy instructor-led and eLearning training options! Couchbase Certification Exams for 2023, now without a proctor requirement." A "Questions?" button is visible. Below the banner is a search bar with a magnifying glass icon and the placeholder "Search by keyword". To the right of the search bar are buttons for "Search" and "All Types and topics".

Upcoming Sessions

February 20	CD410: Advanced N1QL Course: Tuning and Optimization - APAC Virtual (GMT+8)
Starting: 02/20/2024 @ 09:00 AM (GMT+08:00) Singapore	Ending: 02/23/2024 @ 05:00 PM (GMT+08:00) Singapore
Type: Multi-day Session	

The screenshot shows the Couchbase Academy course catalog. It features four courses arranged in a grid:

- CB130n: Couchbase Associate Node.js Developer Certification With Capella Course**
This newly revamped course shows how to leverage the full power of Couchbase 7 as a service with Couchbase Capella. The following 8 courses provide a fundamental understanding of the Couchbase NoSQL database and essential functionality. Throughout these courses, we share the basics of SQL vs. NoSQL, how to sign up for Couchbase Capella, modeling data to the benefit of Couchbase, and an example application you will build. Learners will also walk through the basics of Couchbase's N1... [Read More](#)
- CB130j: Couchbase Associate Java Developer Certification With Capella Course**
This newly revamped course leverages the full power of Couchbase 7 and supports Couchbase Capella. The following 8 courses provide a fundamental understanding of the Couchbase NoSQL database and essential functionality. Throughout these courses, we share the basics of SQL vs. NoSQL, obtaining and downloading Couchbase, modeling data to the benefit of Couchbase and an example application you will build. Learners will also walk through the basics of Couchbase's N1... [Read More](#)
- CB131: Couchbase Associate Architect Certification With Capella Course**
This newly revamped course demonstrates the full power of Couchbase 7 and the fully-managed Database as a Service (DBaaS), Couchbase Capella. The Couchbase Associate Architect Course shares a fundamental understanding of the Couchbase NoSQL database and essential functionality as accessed through the Couchbase Capella user interface. It discusses modeling data to the benefit of the database and application, as well as how to write and implement SQL... [Read More](#)
- CB140a: Couchbase Associate Android Developer With Capella Course**
This course showcases and demonstrates how to create a new Android application using Couchbase Mobile and Couchbase Capella, our fully managed DBaaS service. The following 7 modules provide fundamental instruction on building an Android application with or without a pre-existing database. To that end, we walk through the essential functionality of Couchbase Capella, the benefits of a fully managed NoSQL database, and how that database interacts with Couchbase Mobile products t... [Read More](#)

<https://docs.couchbase.com/>

<https://developer.couchbase.com/>

<https://couchbase.live/>

<https://query-tutorial.couchbase.com/tutorial/#1>



Thank you!



Paul.Son@couchbase.com

www.couchbase.com

cloud.couchbase.com



Couchbase

