

搜索型数据库白皮书

CCSA TC601 大数据技术标准推进委员会

2024 年 7 月

版 权 声 明

本报告版权属于 CCSA TC601 大数据技术标准推进委员会，并受法律保护。转载、摘编或利用其它方式使用本报告文字或者观点的，应注明“来源：CCSA TC601 大数据技术标准推进委员会”。违反上述声明者，本推委会将追究其相关法律责任。

编写委员会

本报告的撰写得到了数据库领域多家企业与专家的支持和帮助，主要参与单位与人员如下。

❖ 主要编写单位（排名不分先后）：

大数据技术标准推进委员会、极限数据（北京）科技有限公司、拓尔思信息技术股份有限公司、中移（苏州）软件技术有限公司、星环信息科技（上海）股份有限公司、北京人大金仓信息技术股份有限公司、达梦数据技术（江苏）有限公司、腾讯云计算（北京）有限责任公司、中移（杭州）信息技术有限公司。

❖ 编写组主要成员（排名不分先后）：

刘蔚、刘思源、齐丹阳、马嘉慧、马鹏玮、闫树、姜春宇、曾勇、罗厚付、张磊、杨帆、廖石阳、李银松、肖锋、江敏、姜康、陶捷、秦华婵、刘永恒、王宝晗、刘熙、谭力鸣、李昆、张欣毅、尹强、王凯龙、刘俊、付新、张静修、鲁剑峰、高攀、任翔、黄超、徐良。

前 言

随着数据智能时代的到来，非结构化数据日益成为各类组织数据的增长主力，蕴含巨大价值，如何高效存储并利用非结构化数据，成为业内关注的焦点。以自动分词、倒排索引、相关度计算、向量检索引擎等技术为核心构建的搜索型数据库，作为高效处理非结构化数据的基础软件，逐渐成为数据智能时代的基石。

搜索型数据库结合全文检索引擎和 NoSQL 数据库的特点，相较于传统的关系型数据库，在处理大规模数据、实现快速检索和智能分析方面展现出显著的优势。近年来，搜索型数据库进一步集成了深度学习、向量搜索和 AI 大模型等尖端技术，使其不仅能够处理基础的文本搜索，还能胜任地理位置信息、图片、音视频等多模态数据的复杂检索任务。然而，搜索型数据库在性能优化、数据一致性保障以及安全性维护等方面仍面临挑战。

本白皮书全面介绍了搜索型数据库的基本概念、发展轨迹和技术架构，同时深入探讨了其在电子商务、金融、医疗健康等关键应用场景中的运用。结合当前搜索型数据库的发展现状和面临的主要问题，本白皮书提出了关于数据库的选择、应用实施的策略，并展望了搜索型数据库未来的发展方向和演进趋势，旨在为行业内外的相关利益方提供宝贵的参考和指导。

目 录

前 言 - 1 -

一、 搜索型数据库概述 - 2 -

 (一) 搜索型数据库概念 - 2 -

 (二) 搜索型数据库发展历程 - 3 -

 (三) 搜索型数据库的挑战 - 6 -

二、 搜索型数据库技术详解 - 8 -

 (一) 搜索型数据库核心技术 - 8 -

 (二) 搜索型数据库关键特性 - 12 -

三、 搜索型数据库应用场景 - 14 -

 (一) 业务场景 - 14 -

 (二) 行业应用 - 17 -

四、 搜索型数据库能力模型 - 19 -

 (一) 基础能力模型 - 19 -

 (二) 能力模型的符合性验证情况 - 21 -

五、 搜索型数据库应用与实施建议 - 22 -

 (一) 数据建模和索引设计 - 23 -

 (二) 性能调优与监控 - 26 -

 (三) 安全性考量 - 30 -

 (四) 应用案例 - 30 -

六、 搜索型数据库演进趋势 - 32 -

 (一) 多模态数据的融合处理和分析 - 32 -

 (二) 人工智能与搜索型数据库的结合 - 33 -

 (三) 云原生架构的演进 - 36 -

 (四) 其它技术对搜索型数据库的影响 - 37 -

七、 总结和展望 - 38 -

图 目 录

图 1 DB-Engines 国际数据库排行榜 - 1 -

图 2 搜索型数据库的发展历程 - 3 -

图 3 DB-Engines 最新搜索引擎排名 - 6 -

图 4 搜索型数据库核心技术模块 - 8 -

图 5 搜索型数据库应用场景一览 - 14 -

图 6 搜索型数据库应用场景 - 19 -

图 7 搜索型数据库应用实施流程图 - 23 -

图 8 中国知识产权大数据与智慧服务系统 - 30 -

图 9 星环科技 Transwarp Scope - 31 -

表 目 录

表 1 搜索型数据库与其它数据库对比 - 3 -

表 2 搜索型数据库产品 - 5 -

表 3 搜索型数据库业务场景 - 14 -

表 4 可信数据库-搜索型数据库基础能力测试通过名单 - 22 -

一、搜索型数据库概述

近几十年来，随着信息技术以及互联网的不断发展，全球数据量呈爆发式增长。数据已经成为信息时代的“石油”，蕴藏的巨大的价值。面对海量的数据，传统关系型数据库逐渐暴露出其局限性，难以满足现代社会对信息获取速度和灵活性的需求。伴随着互联网搜索引擎的成功，作为搜索引擎的底层支撑的搜索型数据库逐渐成为大数据时代的基石。

搜索型数据库是以自动分词、倒排索引、相关度排序、向量检索引擎等前沿技术为核心的数据库系统，依托其卓越的性能和灵活性，满足了现代数据处理的复杂需求。特别是基于分布式架构构建的搜索型数据库，也被称作分布式搜索型数据库，它们作为非关系型数据库中的佼佼者，已成为分析性数据库领域的一个重要分支。在国际数据库流行度排行榜（DB-Engines）前二十五名中，搜索型数据库已经成为除了关系型数据库以外的第二大类数据库。

1.	1.	1.	Oracle	Relational, Multi-model	1221.06	-20.39	-40.23
2.	2.	2.	MySQL	Relational, Multi-model	1101.50	-5.17	-81.29
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model	845.81	-7.76	-76.20
4.	4.	4.	PostgreSQL	Relational, Multi-model	634.91	+5.50	+21.08
5.	5.	5.	MongoDB	Document, Multi-model	424.53	+4.18	-34.25
6.	6.	6.	Redis	Key-value, Multi-model	157.00	-3.71	-15.45
7.	7.	8.	Elasticsearch	Search engine, Multi-model	134.79	-0.95	-4.28
8.	8.	7.	IBM Db2	Relational, Multi-model	127.75	-4.47	-15.17
9.	9.	11.	Snowflake	Relational	125.38	-2.07	+10.98
10.	10.	9.	SQLite	Relational	118.16	+0.88	-15.66
11.	11.	10.	Microsoft Access	Relational	107.93	-5.24	-24.13
12.	12.	12.	Cassandra	Wide column, Multi-model	104.59	-4.69	-9.20
13.	13.	13.	MariaDB	Relational, Multi-model	95.03	-2.20	-1.81
14.	14.	14.	Splunk	Search engine	89.68	-1.97	+1.71
15.	16.	16.	Microsoft Azure SQL Database	Relational, Multi-model	78.51	-1.06	+1.06
16.	15.	15.	Amazon DynamoDB	Multi-model	77.72	-5.18	-3.05
17.	17.	19.	Databricks	Multi-model	74.34	-2.57	+13.48
18.	18.	17.	Hive	Relational	64.82	-0.99	-6.09
19.	19.	21.	Google BigQuery	Relational	62.67	-0.96	+9.23
20.	20.	18.	Teradata	Relational, Multi-model	48.95	-2.29	-14.79
21.	21.	22.	FileMaker	Relational	48.81	-1.67	-2.33
22.	23.	23.	SAP HANA	Relational, Multi-model	45.49	+0.27	-5.36
23.	22.	20.	Neo4j	Graph	44.36	-2.25	-9.15
24.	24.	24.	Solr	Search engine, Multi-model	43.91	-0.69	-3.36
25.	25.	25.	SAP Adaptive Server	Relational, Multi-model	37.49	-1.85	-5.46

数据来源：DB-Engines 官网，2024 年 6 月

图 1 DB-Engines 国际数据库排行榜

根据 IDC 数据预测，随着数据量持续增长，全球数据预计到 2025 年将达到 175ZB，其中 80% 为非结构化数据，相较于传统关系型数据库（RDBMS），以分词器、倒排索引、相似度评分、向量检索引擎等技术为核心构建的搜索型数据库可以更高效管理非结构化数据的处理需求。

接下来本章将从搜索型数据库的概念、发展历程以及优势与挑战三个方面介绍搜索型数据库的基本情况。

(一) 搜索型数据库概念

搜索型数据库是一类专门用于数据内容搜索的 NoSQL 数据库，它结合了全文检索引擎和 NoSQL 数据库的特点，通过对索引和检索过程的优化，以高效处理和检索大规模文本、半结构化或非结构化数据，提供快速的搜索和分析功能，是一种应对大数据时代的数据解决方案。

当前，搜索型数据库的应用范围已经远超传统的长文本数据领域，它们同样擅长处理常规的数值、日期等结构化数据，以及 IP 地址、地理位置信息、图片、音视频等非结构化数据。随着技术的发展和需求的演进，搜索型数据库的应用场景持续扩展，正逐步涵盖业务系统检索加速、IT 运维的可观测性提升、聚合查询分析等多个方面，并朝着支持更广泛场景和多模态数据搜索的方向发展，展现出其在现代数据管理中的多面性和深远潜力。

表 1 搜索型数据库与其它数据库对比

	搜索型数据库	关系型数据库	时序数据库	图数据库	多模数据库
数据结构	多样，支持文本、图片等	表格形式，强结构化	时间序列数据	图形结构，节点和边	支持多种数据模型
查询语言	通常使用自定义的查询语言或 API	SQL	针对时间序列优化的查询语言	图查询语言，如 Gremlin	支持多种查询语言
索引机制	倒排索引、向量索引	B 树、哈希等	通常优化时间戳索引	基于图形的索引，如邻接表	可能集成多种索引机制
读写性能	读优化，快速全文检索	读写均衡，事务支持	读优化，时间序列快速检索	读写性能取决于图形结构	根据数据模型可能不同
事务支持	通常不支持或有限支持	支持 ACID 事务	不适用	不一定，取决于实现	可能支持 ACID 事务
水平扩展	优秀，易于扩展	有限，通常垂直扩展	有限，可能通过分区实现	取决于图形数据库实现	优秀，设计用于多种数据模型
适用场景	大规模文本、多媒体数据搜索	事务处理、结构化数据存储	监控、传感器数据等时间序列分析	社交网络、推荐系统等	多数据类型统一管理

(二) 搜索型数据库发展历程

搜索型数据库的发展主要经历了初始起步期、技术突破（成熟）期、高速发展期及智能转型（融合）期四个阶段：

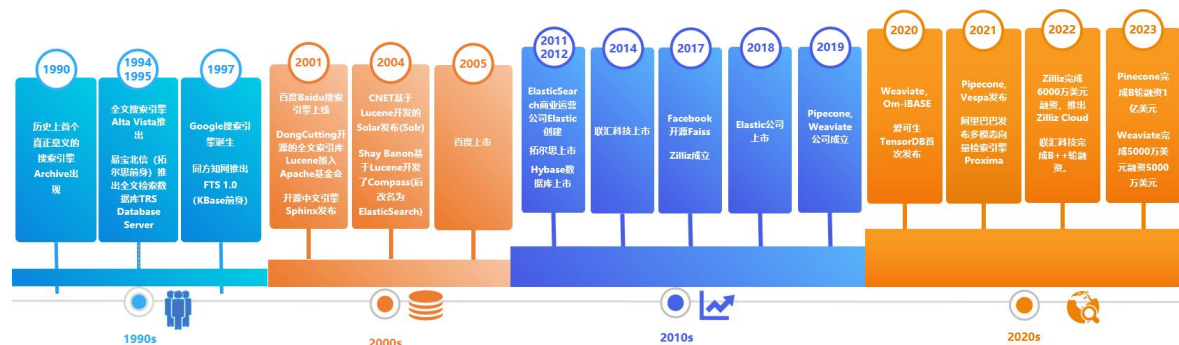


图 2 搜索型数据库的发展历程

1.初始起步期（1990 年代）

搜索数据库的雏形出现于上世纪 90 年代，当时以全文检索为主要技术手段，是为解决传统关系型数据库在处理非结构化数据处理上的低效问题而诞生的，如关系型数据库中长文本检索效率低下的问题。搜索型数据库最初应用于文档检索和网络搜索。随着互联网搜索引擎的兴起和计算机软硬件的不断发展，为满足企业对内部文档、邮件、日志等数据的检索需求，搜索型数据库开始广泛应用于企业的内部应用。典型代表包括 AltaVista、Excite 等。

2.技术突破（成熟）期（2000 年代）

随着互联网的快速发展，搜索型数据库开始应用于更多领域，如电子商务、社交网络等。1999 年，Doug Cutting 创建了 Lucene，它是一个开源的全文搜索引擎库，提供了强大的文本搜索和分析功能，能够快速而高效地处理大规模的文本数据。它以高性能的数据结构、分词分析能力、高扩展性、多语言支持以及良好的社区形态迅速成为搜索型数据库重要的组成部分。除 Lucene 外，Sphinx 等开源搜索引擎的出现也推动了搜索技术的进步。

3.高速发展期（2010 年代）

2010 年前后，搜索数据库进入商业化阶段，以 Elasticsearch 等为代表的商业搜索引擎崭露头角。企业开始大规模应用搜索数据库来管理和检索大量数据。

随着大数据时代的到来和开源软件的流行，以 Solr、ElasticSearch 为代表的开源搜索型数据库逐渐成为主流，为搜索型数据库的发展提供了丰富的技术支持和生态系统。开源社区的贡献促进了搜索型数据库技术的创新和发展，使得更多的企业和组织能够从中受益。ElasticSearch 凭借其优秀的架构设计成为各大云平台厂商的首选。

4.智能转型（融合）期（2020 年代）

2020 年前后，随着大数据和人工智能技术的融合发展，搜索型数据库逐渐向智能化转型。随着云计算的普及，云原生搜索型数据库成为新的发展趋势。这些数据库系统能够充分利用云平台的弹性伸缩和自动化管理能力，为用户提供更加灵活和高效的搜索服务。搜索型数据库开始引入深度学习、向量搜索、自然语言处理等技术，不仅可以处理简单的文本搜索，也可以支持地理位置信息、图片、音视频等多模态数据的检索，为企业和组织提供了更加灵活、快速、个性化和智能化的数据处理和分析能力。

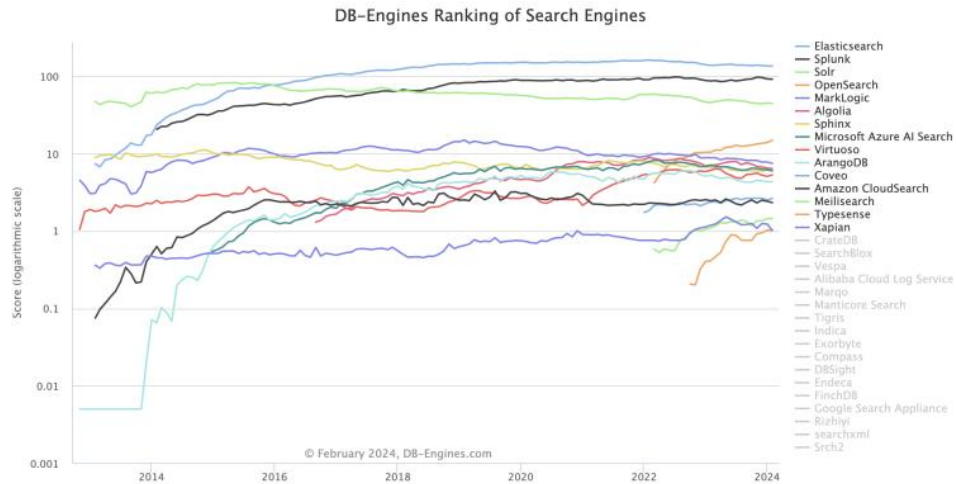
近年来，以 Transformer 为代表的大语言模型把人工智能技术带到了新的高度，这些模型经过大量数据的预先训练，能够理解上下文、生成相关响应，甚至进行对话。搜索型数据针对大模型的数据局限性、幻觉、知识边界的限制也做了交叉验证、预训练模型、评分、搜索增强（RAG）技术层面的探索。

表 2 搜索型数据库产品

年份	国家	公司	产品
1993	中国	拓尔思	TRS HyBase
2001	美国	MarkLogic Corp.	MarkLogic
2003	美国	Splunk Inc.	Splunk

2006	美国	Apache Software Foundation	Solr
2012	荷兰	Elastic	ElasticSearch
2015	中国	百度	Tera
2018	中国	优特捷	Beaver
2021	美国	Amazon Web Services	OpenSearch
2022	中国	星环科技	Scope
2023	中国	国网信通亿力	思极亿搜
2023	中国	极限科技	EasySearch

随着技术的不断进步和完善，搜索型数据库将继续发挥着重要的作用，并为用户带来更多的创新和价值。



数据来源：DB-Engines 官网，2024 年 6 月

图 3 DB-Engines 最新搜索引擎排名

(三) 搜索型数据库的挑战

搜索型数据库作为一种创新性的数据库技术，针对搜索需求设计了专用的数据结构与算法，能够满足用户丰富的搜索需求。同时新技术的发展也给搜索型数据库带来了新的变革,使其成为大数据处理和分析的重要工具。

搜索型数据库以其灵活的检索功能、高效的查询性能、对多样化数据类型的处理能力以及分布式架构下的高可扩展性，在现代数据密集型应用中发挥着至关重要的作用。尽管搜索型数据库具有许多优势，但也面临着诸多挑战。通过克服这些挑战，搜索型数据库可以更好地发挥其在大数据处理方面的优势：

（1）复杂的实现和部署过程

搜索型数据库的实现和部署相对复杂，需要充分考虑数据的结构、索引设计以及查询优化等方面的问题，对系统架构和技术选型有较高要求。

（2）性能优化的挑战

在处理大规模数据时，搜索型数据库需要面临性能优化的挑战，以保证快速的查询响应时间和稳定的性能表现。对于特定的查询场景和业务需求，需要进行系统调优和性能测试。

（3）数据安全和一致性

由于搜索型数据库通常存储敏感信息，安全性成为其面临的另一个重要挑战，需要采取严密的安全措施来保护数据的安全和隐私，如访问控制、数据加密等。另外，在分布式环境下，数据一致性成为搜索型数据库面临的一个重要挑战，需要采取有效的一致性保障策略，确保数据的一致性和完整性。

（4）集群规模与稳定性

随着数据量的爆炸性增长，搜索型数据库的集群规模不断扩大，对可扩展性和稳定性提出了更高要求。用户期望集群能够在 PB 级别的数据量和数十甚至上百节点的规模下，依然保持高效运行。

(5) 流处理支持与读写分离

在实时数据处理日益重要的今天，搜索型数据库需要支持流处理，并能够实现高效的读写分离。这意味着数据库不仅要能够处理实时数据采集和查询，还要在写入和查询之间实现性能的平衡。

二、搜索型数据库技术详解

(一) 搜索型数据库核心技术

搜索型数据库作为一种基于全文检索技术的数据库系统，其核心技术涵盖了多个方面，包括分词、倒排索引、相关度排序等。近年来，随着深度学习技术的发展，稠密向量和向量索引等技术为搜索型数据库在语义搜索、多模态搜索等提供了技术基础。

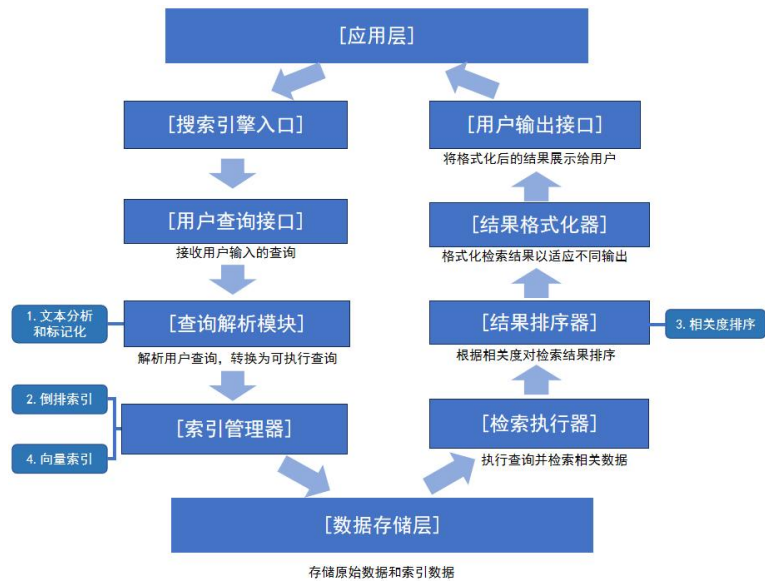


图 4 搜索型数据库核心技术模块

1.文本分析和标记化（Text Analysis and Tokenization）

文本在索引前需要经过处理，包括分词（将文本分解成单词或短语的过程）、文本过滤（移除停用词等）、小写化、词干提取等，以形成可以被索引的标记（tokens）。

分词技术是搜索型数据库中的重要组成部分，用于将文本数据进行分词处理，将连续的文本序列切分成有意义的词语或词项。在中文搜索中，分词技术能够识别出词语的语义边界，从而准确地进行索引和检索。常见的分词技术包括基于规则的分词、基于词典的分词、基于统计的分词等。常用的中文分词器有 IK 分词器、Smart-CN 分词器等等。

2.倒排索引技术（Inverted Index）

倒排索引技术是搜索型数据库的核心索引结构，它将文档中的每个词项映射到包含该词项的文档列表。倒排索引以词项为键，文档 ID 列表为值，实现了从词项到文档的快速反向查找。倒排索引结构能够快速定位到包含查询词的文档，是实现快速文本检索的基础。为了提高倒排索引的压缩率和查询速度，通常会采用数据压缩技术和列存储等技术。

3.相关度排序技术（Ranking and Relevance Scoring）

相关度排序技术是搜索型数据库中的重要功能，它负责根据查询条件和文档内容计算出每个文档与查询之间的相关度，并将结果按相关度从高到低排序。相关度排序通常基于信息检索领域的相关度模型，如 TF/IDF（词频-逆文档频率）、BM25 模型等。这些模型能够通过分析查询词项

在文档中的频率和位置等信息，计算出文档与查询之间的相似度，从而实现结果的排序和排名。

4. 向量索引技术（Vector Index）

向量索引技术是近年来在搜索型数据库中广泛应用的一种技术，这项技术的核心是将数据转换为向量嵌入，本质上是将数据项以多维空间中的数值表示形式呈现，它将文档、图片、音视频等多模态数据表示为高维度的稠密向量，这些表示捕获了数据的语义上下文和含义，使得能够基于相似性而不仅仅是关键词匹配来获得高度相关的搜索结果。向量索引技术通常结合了深度学习技术，使用预训练的模型（如 BERT、ViT 等模型）将多模态数据映射到低维度的稠密向量空间，然后利用向量相似度计算方法（如余弦相似度、欧氏距离等）来进行数据的检索和相关度排序。为实现高效的向量搜索，搜索型数据库往往引入高效的向量索引机制，实现高效的近似最近邻搜索（ANNS）。常见的向量索引算法有局部敏感哈希（Locality Sensitive Hashing，LSH）、最小世界导航图（Hierarchical Navigable Small World，HNSW）等等。

搜索型数据库通过这些核心技术，不仅优化了传统的文本搜索，还拓展了对图像、音频等非文本内容的搜索能力，极大地丰富了搜索的应用场景和提高了搜索的准确性。

5. 数据存储和检索（Storage and Retrieval）

在搜索型数据库的设计和实现中，数据的存储和检索策略是其核心性能的基石。

（1）存储策略：

分段存储：为了提高写入效率和优化索引维护，数据通常会被分成多个较小的块或分段。每个分段独立索引，可以并行处理，从而提高了系统的伸缩性和性能。

数据压缩与编码：为了节省存储空间并提高数据传输效率，搜索型数据库经常使用各种数据压缩和编码技术。这包括对文档 ID 列表进行编码以减少倒排索引的大小，以及使用列式存储时对列数据进行有效压缩。

数据分区和分片：为了进一步提升性能和伸缩性，数据通常会根据某些键（如时间戳或地理位置）进行分区或分片。这允许数据库在物理上跨多台服务器分布，支持水平扩展和高可用性。

内存管理：为了提高索引数据的查询与写入效率，需要考虑索引数据在内存中的缓存淘汰策略、优化的数据结构来减少内存占用空间、故障发生时内存缓存数据的持久化问题。

（2）检索策略

布尔查询处理：搜索型数据库支持复杂的布尔查询逻辑，允许用户执行包含 AND、OR 和 NOT 操作的复合查询。这需要数据库高效处理多个倒排索引的交集、并集和差集操作。

相关性评分和排序：除了找到匹配的文档外，搜索型数据库还会计算每个文档与查询的相关性得分，并根据这个得分对结果进行排序。这通常涉及诸如 TF-IDF、BM25 等评分算法，考虑词频、文档频率和文档长度等因素。

查询优化：为了提高查询性能，搜索型数据库采用各种优化策略，如查询重写、索引剪枝、缓存策略和并行处理。这些技术可以减少需要检查的数据量，降低查询延迟。

分布式查询执行：在分布式搜索型数据库中，查询可能需要跨多个节点执行。这就要求数据库能够有效地分解查询任务，分发到各个节点上，并高效汇总查询结果。

综上，搜索型数据库的设计注重于如何高效地存储和检索大量动态变化的数据，同时保持查询的高速响应。通过利用倒排索引、分段存储、数据压缩和优化的查询执行策略，搜索型数据库能够提供强大的全文搜索能力和高效的数据分析功能。

(二) 搜索型数据库关键特性

搜索型数据库具有多种关键特性，这些特性使得它在处理大规模数据、实现快速检索和智能分析方面具有优势。以下是搜索型数据库的一些关键特性：

1. 全文搜索能力

搜索型数据库的核心特性之一是其对全文搜索的支持，允许用户在大量文本中快速查找包含特定词汇的文档。通过高效的倒排索引实现，这些系统能够在极短的时间内返回查询结果。

2. 相关度排序

搜索型数据库具有高精确度和高召回率的特性，即它能够根据预定义的或自定义的相关性算法，返回与用户查询高度相关的结果，并且能够尽

可能多地检索到与查询相关的文档。高精确度保证了搜索结果的准确性，即返回的结果大部分都是与用户查询相关的信息；高召回率保证了搜索的全面性，即搜索结果中包含了大部分与用户查询相关的信息。这一特性可以提高搜索型数据库的效率和用户体验，确保用户能够快速找到所需信息。

3.实时数据处理

搜索型数据库能够处理实时产生的数据流，并能够在数据发生变化时实时更新索引以反映最新的数据变化，确保搜索结果的实时性。这使得用户能够及时获取最新的数据信息，并进行即时的查询和分析。

4.多样化数据模型支持

与传统关系数据库相比，搜索型数据库通常提供更灵活的数据模型，能够高效处理结构化、半结构化及非结构化数据，如文本、日志、图像、音频等多种数据类型。这种能力使得搜索型数据库在处理各种类型的数据时更加灵活和全面。

5.高并发支持

搜索型数据库具有强大的大并发支持能力，能够处理大规模用户同时发起的查询请求。这意味着搜索型数据库在高流量和大规模数据场景下仍能保持稳定的性能表现，并且能够有效地处理多个用户同时进行的查询操作。为了实现大并发支持，搜索型数据库通常采用了分布式架构、负载均衡、缓存技术、异步处理等手段，以确保系统的可扩展性、稳定性和性能表现。大并发支持使得搜索型数据库能够应对用户数量剧增或突发性流量的情况，保证系统始终保持高效、稳定的运行状态。

搜索型数据库的这些关键特性使其成为大数据处理和实时分析的理想选择,尤其适合于当前数据驱动的业务环境。通过不断优化和技术创新,搜索型数据库将继续在数据管理和信息检索领域发挥重要作用。

三、搜索型数据库应用场景

搜索型数据库通过提供快速、灵活的搜索功能,帮助各种规模和类型的组织提高数据处理能力,优化用户体验,并支持业务决策。随着技术的不断进步,搜索型数据库的应用场景也在扩大。

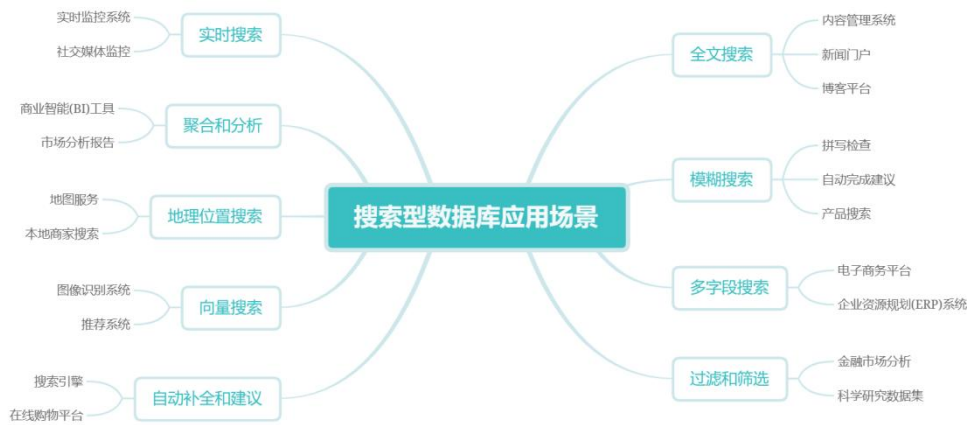


图 5 搜索型数据库应用场景一览

(一) 业务场景

搜索型数据库以其卓越的搜索能力和智能分析功能,在多个业务场景中发挥着关键作用,成为企业数字化转型的强大推动力。

表 3 搜索型数据库业务场景

	企业搜索平台	产品信息管理	内部知识管理	客户服务支持	网络空间治理
功能描述	提供企业内部信息检索	管理企业产品信息	索引和检索企业文档	响应客户查询和反馈	网络舆情监测与分析
实际应用	员工快速查找企业资料	实时更新和检索产品数据	员工访问内部报告、邮件	客户问题快速解决	监测网络言论、信息筛选

1.企业搜索

搜索型数据库在企业搜索领域有着广泛的应用，能够帮助企业高效地管理和利用内部和外部的信息资源，提升工作效率、提高客户满意度，促进企业决策科学化和合规经营。搜索型数据库的应用为企业提供了强大的信息管理和智能分析能力，是企业数字化转型的重要支撑和推动力量。

2.产品信息管理

搜索型数据库可以用于管理企业的产品信息，包括产品规格、特性、价格、库存等。通过搜索型数据库，企业可以实现对产品信息的实时更新和快速检索，帮助销售团队和客户快速找到所需产品信息，提高销售效率和客户体验。

3.内部知识管理

搜索型数据库可以帮助企业进行内部知识管理，将企业内部的文档、报告、会议记录、邮件等信息进行索引和检索。员工可以通过搜索型数据库快速找到所需的信息资源，避免信息孤岛和重复劳动，提高工作效率和协作能力。

4.客户服务与支持

搜索型数据库可以用于企业客户服务和支持领域，帮助企业及时响应客户的需求和问题。企业可以将客户反馈、投诉、建议等信息进行索引和分析，快速识别问题并给予解决方案，提高客户满意度和忠诚度。

5.网络空间治理

搜索型数据库在网络空间治理方面的应用具有重要意义，可以为网络空间治理提供数据支持、智能分析和决策支持等方面的帮助。以下是搜索型数据库在网络空间治理方面的应用场景及作用：

（1）舆情监测与分析

搜索型数据库可以实现对网络舆情的监测和分析，通过收集和分析网络上的言论、评论、情绪等信息，及时了解和掌握社会热点、舆论动向等情况。基于搜索型数据库的数据挖掘和智能分析技术，网络空间治理部门可以从海量信息中发现潜在的问题和风险，及时采取相应的措施。

（2）信息筛选与过滤

搜索型数据库可以帮助网络空间治理部门对信息进行筛选和过滤，识别和清除网络中的虚假信息、有害信息和违法内容。通过建立相关的搜索规则和过滤机制，搜索型数据库能够自动检测和识别不良信息，并进行相应的处理和屏蔽。

（3）数据挖掘与预警

搜索型数据库结合数据挖掘技术，可以发现网络空间中的异常行为、恶意攻击、网络欺诈等问题。通过分析用户行为数据、网络流量数据、安全日志等信息，搜索型数据库可以实现对网络安全事件的预警和预测，为网络空间治理部门提供及时的警报和响应。

总体而言，搜索型数据库的设计注重于高效地存储和检索大量动态变化的数据，同时保持查询的高速响应。通过利用倒排索引、分段存储、数据压缩和优化的查询执行策略，搜索型数据库能够提供强大的全文搜索能

力和高效的数据分析功能，满足企业在信息管理和网络治理方面的复杂需求。

(二) 行业应用

搜索型数据库的影响力远远超出了企业搜索的范畴，其多功能性和灵活性使其成为众多行业的关键技术支柱，包括但不限于以下几个方面：

1. 电子商务与零售业

搜索型数据库在电子商务和零售业中有着广泛的应用。它可以用于商品搜索和推荐，帮助用户快速找到所需的商品，并根据用户的偏好和行为进行个性化推荐。同时，搜索型数据库还可以用于库存管理、订单处理、用户反馈等方面，提升电商平台的运营效率和用户体验。

2. 金融与投资领域

在金融和投资领域，搜索型数据库可以用于数据挖掘、风险管理、投资分析等方面。它可以帮助金融机构对大量的金融数据进行全文检索和分析，发现潜在的投资机会和风险因素，为投资决策提供参考和支持。

3. 健康医疗行业

搜索型数据库在健康医疗行业中也有着重要的应用，主要包括医疗知识管理、医学文献检索、疾病诊断等方面。通过搜索型数据库，医疗工作者可以快速获取医学文献和最新研究成果，辅助临床决策和医疗诊断。

4. 教育与培训领域

在教育和培训领域，搜索型数据库可以用于教学资源管理、学习资料检索、个性化学习推荐等方面。它可以帮助学生和教师快速找到所需的教学资源和学习材料，提升教学效率和学习成效。

5.媒体与娱乐产业

搜索型数据库在媒体和娱乐产业中也有着广泛的应用，主要包括新闻资讯搜索、视频音频检索、内容推荐等方面。通过搜索型数据库，用户可以快速找到感兴趣的新闻报道、影视作品和娱乐节目，丰富了媒体和娱乐内容的传播和使用。

6.物联网和智能设备领域

在物联网和智能设备领域，搜索型数据库可以用于设备管理、数据监控、故障诊断等方面。它可以帮助设备制造商和运营商对大量的设备数据进行管理和分析，发现设备运行状态和异常情况，提升设备运行效率和可靠性。

7.公共服务与政府治理

搜索型数据库还可以应用于公共服务和政府治理领域，包括政府信息公开、公共服务导航、政务查询等方面，为公民和企业提供便捷的政务服务和信息查询渠道，促进政府治理的透明化和高效化。



图 6 搜索型数据库应用场景

搜索型数据库的行业应用广泛且深入，它通过提供快速、准确的数据处理和分析能力，为不同行业的数字化转型和智能化升级提供了坚实的技术基础。

四、搜索型数据库能力模型

(一) 基础能力模型

1. 基本能力

搜索型数据库的主要功能是数据库基础能力是指提供支持非结构化数据导入、用户快速定位所需信息的能力，提供包括数据内容检索、索引、数据导入导出、数据存储、数据一致性、排序、自定义插件、数据类型丰富度等多个方面的功能，具体指标如下：

1) 数据库内容检索：非结构化数据检索、全文检索、向量检索、地理信息查询、组合索引查询、多语种支持、聚合分析等；

2) 索引：索引的定义、修改、删除等操作；

3) 数据库导入导出：通过 API、客户端或其它工具等至少一种数据导入、导出方式；

4) 数据存储：数据和索引的存储压缩；

5) 数据一致性：数据的最终一致性；

6) 排序：按照相关度排序（如 BM25、TF/IDF 以及向量相似度等算法）、基于字段排序、组合排序等；

7) 检索语法：常见的检索语法，如与、或、非等逻辑运算符操作；

8) 自定义插件：通过自定义插件方式扩展数据库功能；

9) 数据类型丰富度：文本、图片、音视频、地理位置信息、数字、日期、二进制数据类型支持。

2.管理能力

数据库管理能力是指支持借助工具或自我管理数据库状态和存储索引的能力，包括监控警告、数据备份与恢复和日志与审计，具体指标如下：

1) 监控告警：数据库指标的监控，如节点健康状态、连接数等，异常告警机制，如提供告警 API、告警日志等对接方式；

2) 数据备份与恢复：增量备份、全量备份以及对数据备份的恢复；

3) 日志与审计：记录系统日志和审计日志等其它相关日志。

3.安全能力

数据库的安全能力是指数据库系统具备权限管理的能力，包括身份认证和访问控制，以确保数据的保密性，具体指标如下：

1) 权限管理：身份认证与访问控制。

4.兼容能力

数据库兼容能力是指数据库系统能够在不同的部署环境中运行和部署，包括硬件平台和操作系统，具体指标如下：

- 1) 硬件兼容：支持在 X86、ARM 等主流硬件上正常运行；
- 2) 操作系统兼容：支持在主流操作系统上正常运行。

5.扩展能力

数据库的扩展能力是指数据库系统能够支持快速、可靠的扩展和缩容，以满足不断增长的数据量和应用负载的需求，具体指标如下：

- 1) 集群扩容：支持集群的扩展；
- 2) 集群缩容：支持集群的动态缩容。

6.高可用能力

数据库的高可用能力是指在面对各种硬件、软件或网络故障时，数据库系统能够保持高可用性和可靠性，包括数据高可用和服务高可用，具体指标如下：

- 1) 数据高可用：支持多副本等机制实现数据高可用；
- 2) 服务高可用：支持在集群部分节点服务异常情况下，其他正常节点仍可提供服务。

(二)能力模型的符合性验证情况

2023 年初，中国信通院云计算与大数据研究所牵头，依托中国通信标准化协会大数据技术标准推进委员会，联合 30 余家企业完成《搜索型数据库技术要求》团体标准编制工作，该标准有效填补了国内关于搜索型

数据库能力模型研究的空白，已成为搜索型数据库行业内技术选型和产品开发的风向标。同时，依托该标准，中国信通院云大所组织开展“可信数据库”-搜索型数据库专项测试工作，拓尔思、极限科技、日志易、爱可生、星环科技、国网亿力、火山引擎、同方知网等 8 家企业的搜索型数据库产品通过测试。关于该评测的详细情况，可联系刘老师，13661023626，liuweil1@caict.ac.cn。

表 4 可信数据库-搜索型数据库基础能力测试通过名单

企业名称	产品名称	版本号	完成测试时间
极限科技	INFINIEasysearch 搜索引擎软件	v1.0.0	2023 年 4 月 19 日
拓尔思	TRS 海贝搜索数据库系统	V10.0	2023 年 5 月 10 日
日志易	日志易 Beaver 数据搜索引擎系统	v1.0	2023 年 5 月 23 日
国网亿力	思极亿搜	V1.0	2023 年 5 月 26 日
星环科技	星环分布式搜索引擎软件 【简称:TranswarpScope】	V9.0	2023 年 6 月 1 日
爱可生	爱可生向量数据库企业版软件 【简称:TensorDB】	V2.0	2023 年 6 月 5 日
火山引擎	火山引擎企业级云搜索服务软件	V2	2023 年 11 月 10 日
同方知网	KBase 数据库	V13.0	2024 年 6 月 4 日

五、搜索型数据库应用与实施建议

搜索型数据库的应用与实施全流程是一个综合性的系统工程，涉及从需求分析到系统部署、优化和维护的多个阶段。本章节将从数据建模与索

引设计、性能调优与监控、安全性配置几个较为重要的环节给出相应的应用实施建议。

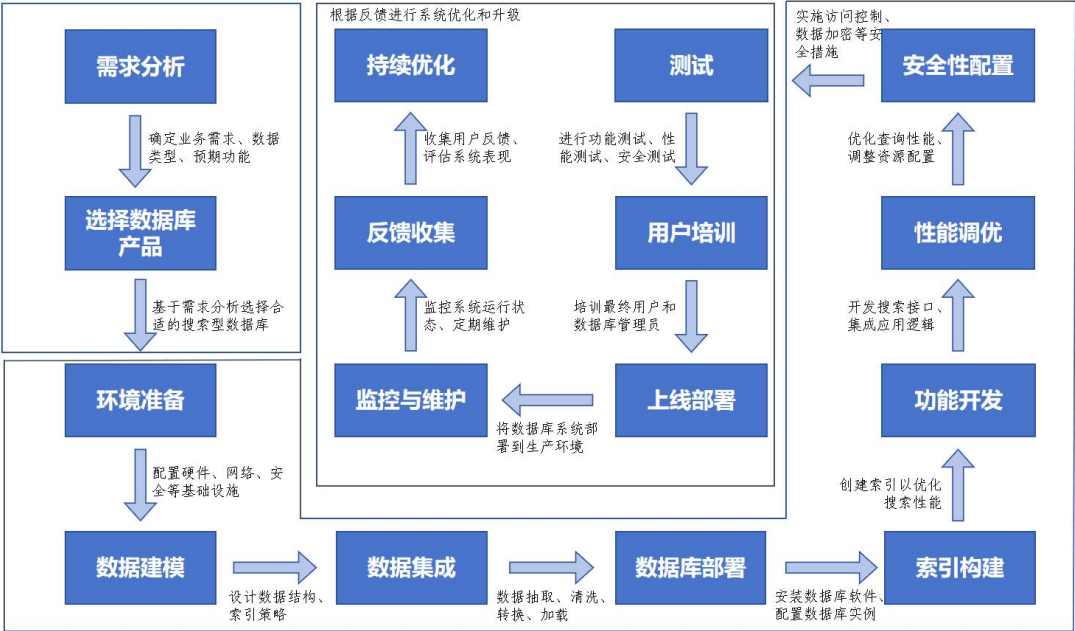


图 7 搜索型数据库应用实施流程图

(一)数据建模和索引设计

在构建搜索型数据库过程中，关键是要紧密围绕搜索和查询性能的优化进行设计，同时保持数据的灵活性和可扩展性。通过精心设计的数据模型和索引策略，可以显著提高搜索效率，满足不同场景下的数据检索需求。

搜索型数据库的数据建模过程稍有不同于传统数据库，因为它更侧重于优化数据以支持高效的搜索、查询性能和数据分析。搜索型数据库如Elasticsearch 通常采用倒排索引来快速定位包含特定词汇的文档，这对于数据建模有一定的影响。以下是搜索型数据库数据建模的步骤：

1) 确定搜索需求分析查询模式: 了解用户如何查询数据, 包括最常见的查询类型、需要全文搜索的字段以及可能的范围查询。识别关键字段: 确定哪些字段是搜索中的关键字段, 如文本内容、日期、标签等。

2) 设计文档结构定义文档模型: 搜索型数据库通常以文档形式存储数据。设计文档结构时, 需要确定文档中将包含哪些字段和字段类型。

3) 设置字段属性: 为每个字段定义类型(如文本、日期、关键字等)和属性(如是否被索引、是否存储、是否参与分词等)。对于需要进行全文搜索的文本字段, 选择合适的分词器。分词器会影响搜索结果的查准率和查全率。

4) 配置索引策略定义索引设置: 包括配置如分片数量、副本数量、分片的调度策略等, 以优化搜索性能和数据的可靠性。根据数据访问模式和查询性能需求, 进行索引优化, 使用索引别名进行灵活的索引管理。

5) 数据规范化与反规范化考虑数据冗余: 为了优化搜索性能, 搜索型数据库中的数据模型可能会比关系型数据库更倾向于数据的反规范化, 减少查询时的关联操作。数据更新策略: 设计数据更新和维护策略, 以保持数据一致性和准确性, 尤其是在反规范化设计中。

6) 性能评估和调优测试查询性能: 基于实际的查询模式测试数据模型的性能, 确保满足应用需求。调整和优化: 根据性能测试结果, 调整文档结构、索引设置或查询策略, 以达到最优的搜索性能和数据处理效率。

索引设计的目标是创建一个既能快速响应查询, 又能高效存储数据的系统。通过精心的索引设计, 可以提升搜索型数据库的整体性能, 满足不同应用场景的需求。

1.全文索引

1) 倒排索引是全文索引的核心,它将文档中的词映射到包含该词的文档列表,极大地加快了文本搜索的速度。

2) 分词器选择对处理特定语种的文本至关重要。根据语言的特性选择或定制分词器,以确保高效准确的文本分析。

3) 配置停用词过滤,减少索引大小并提升搜索效率。同时,通过同义词扩展和相关性评分参数调整,优化搜索结果的准确性和相关性。

2.字段索引

1) 对于需要支持精确匹配和范围查询的字段,B+树索引通常是最高效的选择,它具备优秀的读写性能和范围查询能力。

2) 位图索引适合于值域有限的字段,如状态标志和枚举类型,特别是在查询操作远多于更新操作的场景下。

3) 对于文本字段,考虑是否进行分词处理以及是否需要处理大小写敏感性,这些设置主要适用于全文搜索场景。

3.综合索引策略

1) 确定使用综合索引(组合字段)或单独索引

2) 评估索引稀疏性或选择性,避免过多或过少的索引

4.索引分片和复制

1) 对索引进行分片,分散存储在不同节点以提高吞吐量

2) 根据查询模式设置适当的索引复制因子,平衡读写性能

5.索引生命周期管理

- 1) 建立索引更新、合并、回收等机制
- 2) 优化索引存储,如压缩、文件合并等

(二)性能调优与监控

在搜索型数据库的性能调优与监控方面,以下几个策略对于提升系统效率和用户体验至关重要:

1.列存储

列存储的核心思想是将数据按列组织存储,而不是按行组织存储。例如,在一个包含姓名、年龄和性别的表中,传统的行存储方式会将每一行的数据连续存储,即姓名、年龄和性别的数据依次存储在一起,而列存储则会将所有姓名存储在一起,所有年龄存储在一起,所有性别存储在一起。列存储可以更有效地利用存储空间,减少数据在磁盘上的占用,降低存储成本。由于相同类型的数据在列存储中是连续存储的,因此可以更容易地进行数据压缩,进一步节约存储空间。列存储可以实现对特定列的高效访问,因此适用于大规模数据分析和数据挖掘场景。在这些场景下,往往需要对特定列的数据进行聚合、过滤、计算等操作,列存储可以显著提升数据处理的效率。

2.读写分离

采用读写分离架构是提升搜索型数据库性能的一种重要手段。这种架构将读操作和写操作分别路由到不同的数据库实例或节点上,以实现更好的负载均衡、提高系统的并发处理能力、降低数据库压力、提高数据访问速度和响应性能等目标。将读操作和写操作分离后,可以通过独立部署的

读库来处理大量的读请求，从而有效分担写库的压力，实现负载均衡。这种架构可以通过横向扩展的方式增加读库的数量，从而提高系统的并发处理能力，支持更多的并发读取操作。将读操作分发到读库上，可以减轻主库的负载压力，降低数据入库的响应延迟，提升系统的稳定性和性能表现。

3.提升硬件

搜索型数据库无论是索引还是检索的过程，都需要消耗大量的 CPU、内存以及磁盘的 IO。高主频的 CPU 能够更快地执行计算任务，而增加 CPU 核心数量则可以提高系统的并行处理能力，使得数据库能够同时处理更多的查询请求。通过使用高性能的多核 CPU，可以实现更快速的索引构建和查询响应，提升系统的吞吐量和性能表现。大内存能够缓存更多的数据和索引信息，减少磁盘 IO 操作，加快数据的读取和写入速度。SSD 硬盘具有高速的读写速度和低延迟的特点，能够显著提升搜索型数据库的 IO 性能。与传统的机械硬盘相比，SSD 硬盘具有更快的数据访问速度和更低的数据访问延迟，能够极大地缩短索引构建和数据检索的时间。通过使用 SSD 硬盘作为数据库的存储介质，可以大幅提高系统的响应速度和数据处理效率，提升用户的查询体验和系统的性能表现。通过提高 CPU 的主频和核数、使用更大的内存以及 SSD 硬盘等手段，可以有效地提高搜索型数据库的性能，加快索引构建和数据检索的速度，提升系统的响应速度和处理能力，从而提高用户的查询体验和系统的性能表现。

综上，通过对存储结构的优化、读写操作的分离以及硬件资源的提升，可以显著提高搜索型数据库的性能，满足日益增长的数据处理需求，为用户提供更快速、更可靠的服务。

(三) 安全性考量

近年来，随着搜索型数据库的广泛应用，搜索型数据库已经深入到各行各业，推动了国内信息化的进一步发展。但是，由于部分厂商的安全意识不足，再加上开源搜索型数据库在安全防护这块的缺失，导致各种安全事故频繁发生。数据泄露严重问题，波及范围广、发生频率高、数据规模大。

因此，用户在搜索型数据库选型的时候，要重点考虑搜索型数据库的安全性问题。这需要综合考虑多个方面，包括数据隐私保护、访问控制、加密技术、安全审计等。以下是搜索型数据库安全性考量的一些关键点：

1.数据隐私保护

搜索型数据库需要采取有效措施保护用户数据的隐私安全，包括对数据进行加密存储、隐私脱敏处理、权限访问控制等。确保用户敏感数据不被未授权访问和泄露是搜索型数据库安全性的首要任务。

2.访问控制

搜索型数据库需要建立完善的访问控制机制，包括身份认证、权限管理、访问审计等功能。只有经过授权的用户才能够访问数据库，并且需要根据用户的角色和权限限制其对数据的操作。

3.数据加密

搜索型数据库需要支持数据加密技术，包括数据传输加密、数据存储加密等。通过对数据进行加密，可以有效防止数据在传输和存储过程中被窃取和篡改，保障数据的加密性和完整性。

4.安全审计

搜索型数据库需要具备安全审计功能，记录用户的操作行为和系统的运行状态，以便及时发现安全事件和异常情况。安全审计可以帮助数据库管理员监控系统的安全性，并对安全事件进行调查和处理。

5.安全漏洞修复

搜索型数据库需要定期对系统进行安全漏洞扫描和修复，及时更新补丁和升级版本，以防止黑客利用已知漏洞攻击系统。及时修复安全漏洞是保障搜索型数据库安全性的重要措施之一。

6.高可用性和灾备备份

搜索型数据库需要具备高可用性和灾备备份功能，保障数据在灾难事件发生时的可恢复性和可用性。通过建立灾备备份系统和实现数据冗余存储，可以确保数据不会因意外事件而丢失或损坏。

7.安全培训和意识

搜索型数据库管理员和用户需要接受安全培训，增强安全意识和技能。通过定期的安全培训和演练，提高管理员和用户对安全问题的认识和应对能力，有效防范安全风险和威胁。

综上，搜索型数据库的安全性考量是一个多维度、多层次的过程，需要从技术、管理和培训等多个方面综合施策，以构建一个全面、坚固的安全防护体系。

(四) 应用案例

1. 拓尔思

中国知识产权大数据与智慧服务系统（DI Inspiro™）是由知识产权出版社有限责任公司开发建设的国内第一个知识产权大数据应用服务系统。DI Inspiro™已经整合了国内外专利、商标、法律文书、标准和科技期刊等知识产权数据资源，实现了数据检索、分析、关联、预警和项目管理等多种功能。该系统充分利用了拓尔思海贝搜索数据库的高性能并发能力以及多语言处理方面的优势，实现了中、英、日之间多语种无差别、无障碍的智能扩展检索和及时统计分析。



图 8 中国知识产权大数据与智慧服务系统

2. 极限科技

极限科技是国内一家专做搜索型数据库的厂商，在数据搜索领域有深厚的技术积累。极限科技的产品 Easysearch 在功能和性能上完全满足集团应用的搜索需求。此外，Easysearch 产品支持在国产服务器、操作系统环境运行，目前已获得了多家国产 CPU、操作系统的厂家认证，如华为、飞腾、龙芯中科、麒麟软件、统信、海光等。

通过国产化替代，不仅提高了企业 IT 技术栈的国产化程度，落实了自主可控的战略布局，也实实在在地给企业降低了系统成本。

3. 星环科技

星环科技 Scope 支持实时，微批处理等多种数据写入方式，产品内置自研存储引擎 Search Engine，对于传统日志等半结构化数据以及标准的结构化数据，支持精确/模糊/多维检索等各类检索场景，做到毫秒级查询与检索能力。

Scope 基于 Raft 协议保证数据的一致性，同时提供多副本支持，满足数据的可靠性需求。此外，Scope 支持跨数据中心的多活部署，使得重要业务在多个数据中心的交互和流转；支持主备集群的实时数据同步，满足主备集群的 failover 和 switchover。

Scope 支持适配长城飞腾、华为泰山、龙芯等服务器架构，同时满足麒麟，UOS 等操作系统，在传统搜索业务的功能外，同时提供多租户，联邦检索等多样化功能，充分满足国产化替换需求。

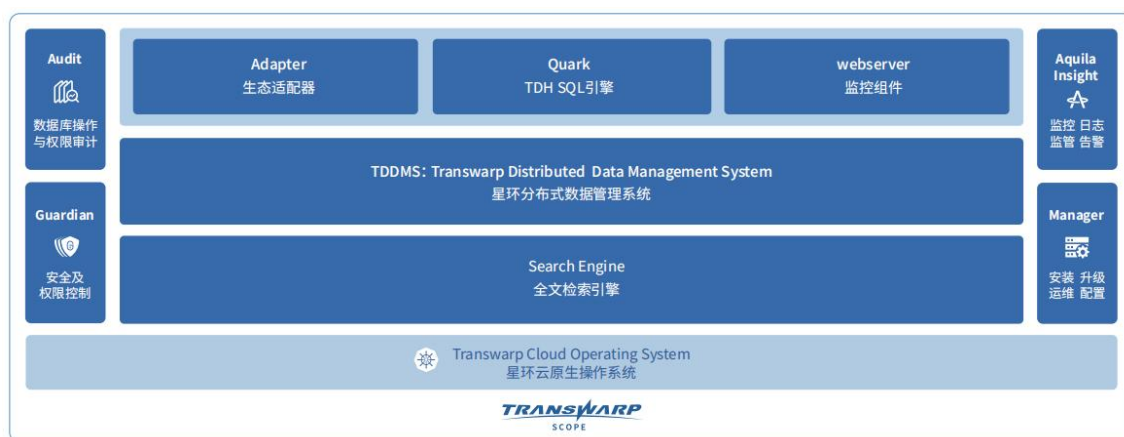


图 9 星环科技 Transwarp Scope

六、搜索型数据库演进趋势

(一) 多模态数据的融合处理和分析

随着数据获取方式和应用场景的多样化，多模态数据（如文本、图像、视频等）融合趋势日益明显，对于多模态数据的处理需求与日俱增。其中，跨模态检索指的是在不同模态的数据之间进行检索的技术，即通过查询一个模态的数据来检索相应的其它模态数据。为了支持跨模态检索，搜索数据库不仅要支持基于语义的文本检索，还需要结合自然语言处理、计算机视觉等技术覆盖图像、视频、语音等多模态数据的检索查询。此外，不同模态间的关联挖掘和协同推理能力成为关键，准确把握数据的语义内涵和上下文信息对于多模态融合至关重要。

通过向量数据的支持，搜索型数据库具有对多种数据类型的支持能力，包括文本数据、结构化数据、半结构化数据、图像数据、音频数据、视频数据等。通过灵活的数据模型和存储结构，搜索型数据库能够处理各种类型的数据，实现多模数据的统一管理和跨模态检索。例如，搜索型数据库通常利用空间对齐的预训练大模型（CLIP、BLIP 等）对数据进行向量抽取实现跨模态搜索。利用这种跨模态搜索技术，可以很方便地实现用语音对手机里面的文字、图片、音视频等进行搜索。

搜索型数据库在处理多模态数据融合方面的能力，已被广泛应用于电子商务、金融服务、医疗健康、智能制造、智慧城市等多个领域和场景。在多模融合的道路上，搜索型数据库的发展方向主要有两个：一是作为全文索引能力的提供者，集成到多模态数据库（如 NoSQL 数据库）中；二

是基于搜索型数据库本身的引擎，通过插件方式构建出其他模型的能力。多模融合的目标是有效降低多模态数据的存储和开发成本。

为了进一步支持多模态数据的融合处理与分析，未来搜索型数据库还将逐步支持对不同模态数据进行联合编码、统一表示，为模态间语义相关性挖掘和跨模态计算奠定基础。

1) 通过表示学习，可以获得不同模态数据的统一语义向量表示，推动模态间的关联挖掘和融合分析；

2) 实现多模态数据的索引关联，构建统一的索引框架捕获模态间的相关联系，从而支持基于语义的跨模态检索和关联查询；

3) 提供模态间协同计算能力，与 AI 计算框架相结合，在数据库内存级别提供针对多模态数据的协同计算能力，直接支持针对多模态场景的智能计算和推理需求；

4) 构建面向特定领域的融合索引库，为领域应用的多模态数据理解和分析提供支持。

随着云上数据和业务场景的不断丰富，各种模态的数据都有着存储、查询、搜索的需求。通过搜索型数据库的多模数据融合处理能力，可以实现数据的全面挖掘和应用，为各行业提供更高效、智能的数据管理和应用解决方案。

(二)人工智能与搜索型数据库的结合

近年来，以 ChatGPT 为代表的大语言模型给人工智能技术注入了新的发展动力，人工智能技术快速迭代。搜索型数据库在人工智能技术发展

中发挥了重要作用，同时大语言模型带来的能力提升也给搜索型数据库技术发展产生重要影响。

1.语义理解

通过利用人工智能技术来实现智能化、语义化和个性化的检索，结合自然语言处理、机器学习等技术分析用户意图，提供更加智能、个性化的搜索服务。搜索输入处理不再局限于文本关键词的分词技术，大模型语义建模与信息生成能力的引入，将会提升语义信息理解与表达能力。大规模通过预训练任务编码语料中所包含的语义信息掌握丰富的知识，进而能够理解自然语言提词，执行复杂的推理任务。

2.长期记忆

在搜索型数据库的业务场景中解决大模型的长期记忆匮乏的问题，其一是解决上下文输入窗口长度限制，有效提升记忆长度，增强搜索相关度，其二是训练数据的缓存，提升微调的处理能力。

3.搜索增强生成（RAG）

大模型训练的计算成本和时间成本都较高，且大模型往往只针对历史数据进行训练，缺乏对时效知识的推理能力。其次，大模型生成的信息较为发散，容易产生幻觉问题，缺乏知识的边界。搜索数据库在检索增强生成中的作用主要有：高效索引和快速检索，利用倒排索引等数据结构，搜索数据库能快速从海量语料中检索与查询相关的片段，并根据相关性评分对检索结果进行排序，提供随时间变化并被持续追踪的信息并融入大模型，丰富上下文输入，确保高质量结果优先被生成模型利用等。

4.融合搜索与评分排序

在向量搜索领域的 BM25 评分算法存在局限性，融合搜索场景的评分也是一个重要的发展方向。向量检索能够具备理解语义相关性的特征，但是在关键词传统搜索场景不太擅长，混合检索能够结合不同检索技术的优势，以获得更好的召回结果。在不同检索模式下，通过把检索结果归一化，引入重排序模型，提升搜索的查准率与召回率。

5.对向量数据库的支持

随着数据处理领域的不断发展，嵌入技术（Embedding）和稠密向量（DenseVector）越来越广泛地被应用于自然语言处理、图像识别、推荐系统、人工智能等领域。嵌入技术是将高维度的离散数据（如文本、图像、音视频等）映射到低维度的连续空间中形成稠密向量，从而更好地捕捉数据之间的潜在关系和语义信息的方法。嵌入技术可以通过无监督或有监督的方式进行训练，其中无监督的训练方式通常使用神经网络模型如 Word2Vec、GloVe 等，有监督的训练方式通常使用深度学习模型，如卷积神经网络（CNN）、循环神经网络（RNN）、注意力机制（Transformer）等。

随着各种近似最近邻搜索算法被提出来，一些专门用于向量搜索的工具被开发出来，如 ANNOY、Faiss 等在图像识别和推荐系统等领域得到了广泛应用。常见的向量索引算法有局部敏感哈希、最小世界导航图等等。然而随着应用的发展以及数据量的不断增长，更加专业的向量数据库逐渐取代向量检索工具。此外，传统的数据库如关系型数据库、内存数据以及搜索型数据库也开始扩展对向量数据的支持。以搜索为核心的搜索型数据

库，在通过插件的形式引入向量搜索以后，增加向量索引算法的支持，并能够适配和集成 LangChain、HuggingFace 等 AI 大模型开发生态。这些数据库支持第三方 Transformer 模型的应用接口接入，提供检索增强生成服务，进一步推动了数据处理技术的革新和发展。

搜索型数据库在人工智能技术发展的浪潮中，不断吸收和融合新技术，以提升其语义理解、长期记忆、生成能力和评分排序等方面的性能，为用户带来了更加智能和个性化的搜索体验。

(三) 云原生架构的演进

云原生是数据库服务与云计算基础设施深度融合的关键，它旨在实现成本效益和效率的最大化。近年来，云原生已成为数据库厂商，尤其是云服务提供商的重要发展方向。

云原生技术发展的核心路线集中在利用 Kubernetes（K8s）进行容器化，以及云计算主机提供的虚拟化、编排和资源管理能力。尽管业界对于数据库是否应运行在容器中存在一定争议，容器化已成为云数据库厂商广泛认可的发展趋势。容器化数据库的优势在于其快速部署、弹性伸缩能力以及全面的运维可观测性，这些特性是其成为技术选型的关键因素。

在计算与存储方面，搜索型数据库正逐步向存算分离架构演进。计算层基于业务流量波峰、低谷的阈值精确检测，提供计算服务的灵活供给与平滑伸缩，并实现租户层的资源隔离与管控。基于存算分离的架构，存储层实现了缓存加速与副本的灵活扩展，实现写入性能的提升的同时，灵活提供本地盘、对象存储、云盘、存储加速多类存储服务，并在存储层维度实现异地多活能力，为数据提供高可用保障。

Serverless 化服务是另一个重要的发展方向，它旨在更精确地匹配业务需求与资源供给，实现用户和云服务提供商的降本增效。Serverless 服务简化了用户对资源购买的评估工作，并避免了因资源评估不准确而导致的浪费或服务降级。Serverless 服务的实现依赖于搜索型数据库的容器化和存算分离架构改造，同时，业务流量的监控、预测和运维支持也在逐步由固定规则向 AI 工具链的方向演进。

总体而言，云原生、存算分离以及 Serverless 化服务的发展，不仅推动了搜索型数据库技术的创新，也为用户带来了更高效、更灵活、更经济的数据处理解决方案。随着技术的不断进步，这些趋势预计将在未来继续塑造数据库服务的新格局。

(四) 其它技术对搜索型数据库的影响

新技术对搜索型数据库的影响显著，推动了数据库系统在多个方面的演进。以下是一些关键的演进趋势：

1. 实时搜索与分析

随着业务应用对即时性需求的增加，搜索型数据库正朝着实时方向演进，通过实时索引和更新机制，能够快速检索和反映最新数据。

2. 新型并发控制和索引技术

研究新型的并发控制算法和索引结构，如基于非易失性内存的索引，以适应现代硬件架构并提升性能。

3. 安全和隐私保护

在数据安全方面，多方安全计算、联邦学习等技术的发展给数据库安全与隐私保护带来了新的挑战 and 机遇。

4. 数据库智能化

AI4DB 可以利用机器学习技术对搜索查询进行优化，通过学习用户的查询模式和数据的使用情况，智能调整索引结构和查询计划，从而提高搜索性能。另外，自动化和智能化功能可以减少搜索型数据库的运维工作量，降低运维成本。

5. 增强现实搜索

随着增强现实技术的发展，搜索技术也将逐渐与增强现实相结合，为用户提供更加直观和沉浸式的搜索体验。增强现实搜索能够将搜索结果无缝融入现实世界，结合 AI 技术为用户提供更加个性化和便捷的搜索服务。

综上，新型技术正在推动搜索型数据库向更高效、更智能、更安全的方向发展，同时也对数据库的设计和实现提出了新的要求。随着技术的不断进步，预计未来搜索型数据库将更加紧密地与 AI、云计算和大数据等技术结合，为用户提供更加丰富和高效的数据服务。

七、总结和展望

随着数字化转型的加速，企业对于能够快速检索、分析和理解数据的系统需求日益增长。作为现代数据管理架构的关键组成部分，搜索型数据库已经证明了其在处理大规模、多样化数据集方面的独特价值。

搜索型数据库通过提供全文检索、模糊匹配、排名算法等功能，使用户能够快速定位到所需信息。它通常结合了倒排索引、向量空间模型等技

术，以优化搜索性能，能够提供对非结构化数据的高效处理能力，以及对实时搜索需求的支持。此外，许多搜索型数据库还提供了强大的分析功能，可以对数据进行深入洞察。

随着人工智能和机器学习等技术的融合，未来的搜索型数据库预计将变得更加智能和自适应，能够提供更加精准和个性化的搜索体验：

（1）通过集成更先进的 AI 算法，搜索型数据库将能够提供更加精准的搜索结果，并能够理解用户的搜索意图。通过集成更先进的 AI 算法，搜索型数据库将能够提供更加精确的搜索结果，并能够理解用户的搜索意图。

（2）随着多云和混合云架构的普及，搜索型数据库需要更好地与各种数据源和平台集成，以实现无缝的数据检索和分析。

（3）随着物联网设备的增多，实时数据处理的需求也在增长。未来的搜索型数据库将更加注重实时性，以支持即时搜索和分析。

（4）数据安全和隐私保护始终为用户关注的重点。搜索型数据库将进一步加强数据加密、访问控制和合规性，以保护用户数据。

（5）随着数据类型的多样化，搜索型数据库将加强对多模数据的支持，包括结构化、半结构化和非结构化数据。

（6）云原生搜索型数据库将获得更多关注，利用容器化、微服务架构等云原生技术，提供更加灵活和高效的服务。

随着技术的不断进步，搜索型数据库将继续演进，以满足不断变化的市场需求。通过持续创新和优化，搜索型数据库将在未来的数据处理领域扮演更加关键的角色。