Changhong Data Intelligence

内容智能洞察软件架构白皮书

2019 年 07 月

目录

1. 摘要

本文档的目的是让用户熟悉内容智能洞察软件的体系架构、功能、设计概念、用例和最佳实践。

本白皮书介绍了内容智能洞察软件的体系结构和功能，旨在帮助相关的产品专家和客户了解底层架构和原理，包括内容智能洞察的操作原则，评估潜在客户环境，并定义实施要求和部署策略。本文档中的信息可用作客户解决方案设计的起点。读者应具备内容智能洞察软件产品的相关知识，并熟悉一般存储、操作系统、Linux 容器技术和网络原理，以及基本的软件工程概念。

2．序言

内容智能洞察软件是一个成熟的搜索和数据处理解决方案。它处理使数据可搜索的所有步骤，而不考虑数据的位置或数据的格式。

收集和处理用于发现和探索的信息可能是一项重大挑战。数据的生成及其重要性可能以不同的速率变化，从而创建一组动态变化的处理规则和要求。数据通常驻留在多个存储库中，包括内部存储库、远程站点和云中。这些存储库可能由不同的团队管理，需要协调数据处理的访问。预计将遵守每个系统上的现有访问控制。元数据可能仅存在于存储数据的平台中，如果没有公开数据的工具，则可能无法从外部访问元数据。此元数据可以具有描述相同存储信息的不同名称（例如 "医生" 与 "内科医生"）。元数据本身可能不完整，可能缺少所需的字段值（例如，缺少 "邮政编码" 的地址）。

此外，这些数据以前可能没有分类，没有元数据来协助处理评价工作。生成这种缺少的元数据是很困难的，因为源数据可能存在于许多不同结构化和非结构化格式的数据中，即使对于相同的数据类型（例如，表示相同信息的不同 xml 架构，或使用不同版本创建的 Word 文档文件也是如此）。最重要的是，数据可能存在于多种语言中，并且可能会显示不同类型的读取错误（由于损坏、加密、OCR 故障等原因）。即使在最简单的内容处理任务中也可能遇到的很多的问题，需要一个策略和工具集来有效地识别和解决每个问题。

内容智能洞察软件为跨多个存储库大规模自动提取、分类、丰富和数据归集提供了工具。这些工具可用于标识、混合、规范化、查询和为数据编制索引，以用于搜索、发现和报告的目的。可以对系统进行培训，以了解整个组织中使用的数据和架构之间的细微差异，从而快速提取发现和探索所需的业务价值。

3．产品架构概要

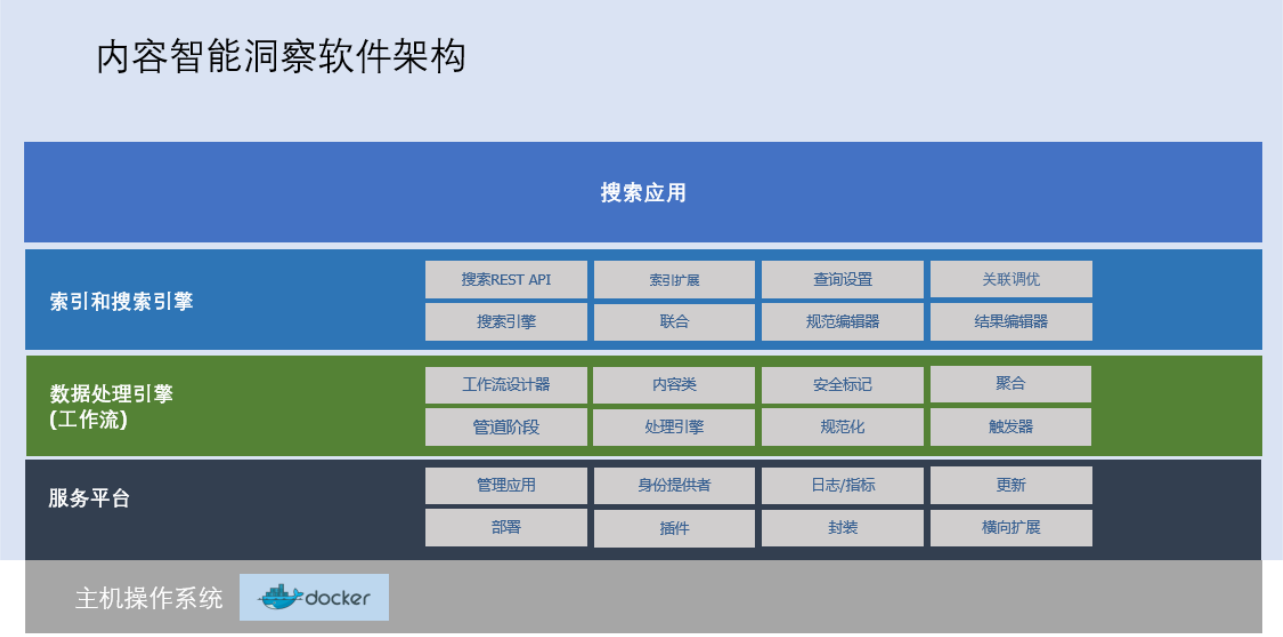
内容智能洞察软件是一种纯软件的内容处理和搜索的整体解决方案，可在各种计算环境中轻松、一致地部署。

内容智能洞察是许多服务（包括工作流执行器和搜索引擎）的集合，嵌入在可扩展服务平台中。该平台提供产品可移植性、可扩展性、打包和标准化的企业级功能集。它允许灵活的产品部署选项，包括裸机、虚拟机，以及在云中利用 linux 容器技术（docker）的可移植性。它旨在从单个实例部署扩展到具有数千个实例的大型群集。群集大小是动态的，允许在需要额外资源时添加实例，并在不需要实例时向下扩展实例。服务可以根据需要在系统实例中单独扩展或负载平衡，以满足不断变化的需求。内置的监视工具允许管理员通过任何接口轻松地管理任何规模的系统（通过 REST API、CLI 和 UI 提供奇偶校验）。提供了系统、实例和服务级颗粒度，可了解系统运行状况和资源利用率（包括 CPU/内存/磁盘/网络等）。

工作流服务用于自动化、扩展和管理数据处理和索引活动。首先，需要在系统中注册数据连接器，使其能够访问数据源。然后，可以通过由阶段组成的一个或多个可自定义的处理管道对这些数据进行分析，每个阶段对数据执行一些转换。内容类还可用于从原始数据流中提取感兴趣的特定结构化或非结构化信息。然后，可以将这些转换的聚合输出保存到可下载的报表中，用于触发其他处理流程，或将其建立索引后给搜索引擎使用，以便对获得的信息进行高级查询和探索。

通过确保只有经过授权的用户才能访问特定功能和搜索结果，并使用文档和元数据字段级别的粒度来实现安全性。该系统将现有标识提供程序（如活动目录 Active Directory 和 LDAP 兼容服务器）的身份验证功能与可自定义的用户角色和访问控制结合起来，以授权使用任何特定的系统功能。搜索引擎查询结果视图可以根据用户或组的需要，通过文档级安全控制和/或基于角色的访问规则访问数据。此外，集中式 SSL 证书管理有助于确保应用程序和数据交换的服务器到服务器通信的安全，而可搜索的访问日志和系统事件则提供强大的审计控制。

使用附带的软件开发工具包，可以专门为正在使用的数据创建自定义数据连接、转换和丰富阶段（Stage）。此组件的可插拔性有助于避免使用过时的技术，同时允许连接和利用任何外部系统、云服务或平台。



4．可扩展的服务平台

内容智能洞察软件是在可扩展的服务平台软件层之上构建的，它抽象底层硬件资源并将它们与产品部署和管理分离。

这个服务平台提供：

- 基于服务产品的纯软件包装和可组合性
- 跨物理、虚拟和云环境的一致产品部署
- 支持群集、扩展、监视、更新、维护和服务管理
- 身份提供者身份认证和授权集成框架
- 通过插件和服务扩展和扩展系统功能的能力
- SDKs 二次开发和可选服务和插件目录

下一节介绍基于底层服务平台的内容智能洞察软件的一些重要概念和功能。

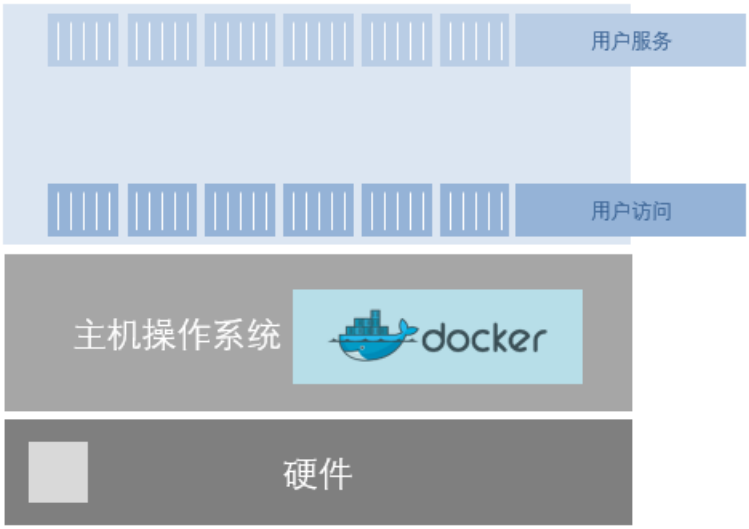
4.1. 实例

实例是运行内容智能洞察软件的任何服务器或虚拟机。

系统可以有一个实例，也可以有多个实例。

单实例系统可用于运行完整的应用程序服务集。单实例部署通常用于开发或测试目的，但也可在生产环境中用于小型用例。

图 1：内容智能洞察软件堆栈 (每个实例)

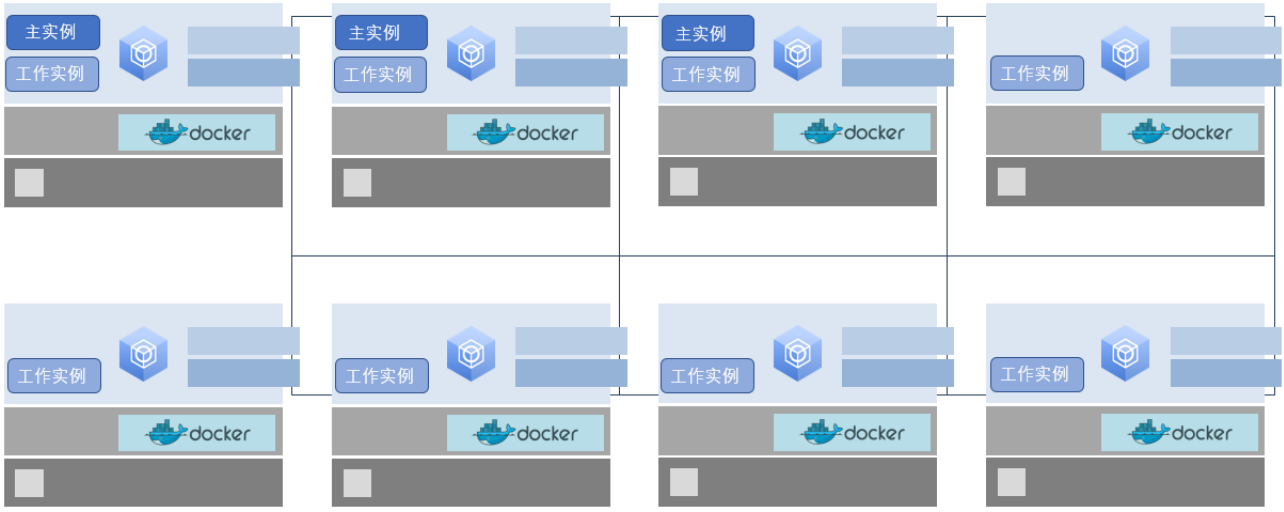


可以部署多个实例来创建群集系统。具有多个实例的系统可以在实例故障的情况下保持更高的可用性。此外，具有更多实例的系统可以同时运行更多的任务，并且通常可以比实例较少或仅使用一个实例的系统更快地处理任务。由于多实例系统可以消除所有单点系统故障，因此强烈建议在生产环境中部署多实例系统。

多实例系统有两种类型的实例：主实例（运行一组基本的核心服务）和非主实例（也称为 "工作实例/工作线程"）。单实例系统将自动同时具有主实例和非主实例的功能。

系统可以有一个主实例，也可以有三个主实例。由于三个主服务器可以满足数千个工作线程实例的管理，因此无需在系统中创建三个以上的主实例。至少需要三个主实例才能执行领导选举和解决冲突等系统操作的仲裁。因此，不允许使用两个主实例系统。

图 2：内容智能洞察软件集群



4.2. 容器

Docker 提供了在称为容器的松散隔离环境中打包进程和运行进程的能力。容器是直接在主机内核中运行的轻量级进程。Docker 平台具有较高的可移植性，因为容器可以在数据中心、云基础架构或混合环境中的便携式计算机、物理或虚拟机上运行。

容器是 Docker 镜像的可运行实例。 图像是一个只读模板，其中包含有关创建 Docker 容器的说明。通常，您可以创建自己的映像，也可以使用其他已在 Docker 注册表中发布的映像。

虽然许多开源工具和软件包用于创建数据洞察，但产品不使用由第三方构建的任何容器映像。内容智能洞察软件的开发人员亲自挑选图像中包含的软件包和版本，并利用自定义构建流程来确保可重复性，并最大限度地减少安全问题。在生成内容智能提供的容器映像期间，不会从 Internet 中提取任何数据。

内容智能洞察软件利用容器的可移植性来部署其服务，将一套 Docker 客户端工具与服务平台的打包，可伸缩性，可扩展性，安全性和更新功能相结合。

4.3. 服务

服务是在内容智能洞察软件平台上运行的一个或多个容器化进程。 服务可单独部署并提供单一功能。许多服务都是内部服务，最终用户可能无法直接使用。

应用程序是在群集外部公开的服务。这些服务为最终用户提供了面向公众的 UI，CLI 和 API 的访问。它们通常受到经过身份验证的用户的保护，通常不会被群集中的代码使用。

内容智能提供两个应用程序：管理应用和搜索应用。 管理应用程序支持产品管理，监控和系统配置，而搜索应用程序则支持针对本地和/或远程搜索引擎的单一或联合查询。

4.4. 网络

与系统中所有最终用户的交互是在可公开访问的网络接口和端口中，通过身份验证服务（应用程序）发起。集群中的其他服务在内部使用所有其他服务。 这些“内部”网络端口必须可以从系统中的每个其他实例访问，但出于安全原因，通常应该使系统外的公共网络不可访问。 应将主机操作系统配置为遵守这些网络规则以最大化系统安全性。

4.5. 时间同步

对于多实例系统，每个实例应将网络时间与系统中的每个其他实例同步。 这通常通过利用 NTP（网络时间协议）和来自每个实例的相同外部源来实现。这有助于确保在群集中的服务之间对最终用户请求进行负载平衡时，可以正确地遵循已配置的登录会话超时。 如果在访问应用程序时立即发生注销问题，则可能未正确配置时间同步。

4.6. 日志

系统中的每个服务都在其自己的命名目录下生成日志文件，该目录位于每个实例的“/ log”文件夹中。系统提供自动日志轮换，当日志文件的大小达到阈值（默认为 10MB）时，它们将被压缩并移动到顶级“/ retired”文件夹。 可以使用递归 grep 命令（例如“grep -R ERROR *”）从顶级日志文件夹全局查询实例上的大多数活动日志文件，这有助于快速识别特定服务的问题。当需要日志下载时，“/ bin”目录下会提供日志下载脚本。此工具可用于收集有关系统的日志和/或诊断信息。选项可用于收集特定服务，特定实例和/或特定日期范围的数据。还可以按需生成诊断报告，其包括有助于分类各种系统故障的信息。

4.7. 指标（度量）

收集各种服务的最后三个月的详细度量信息并将其存储到度量服务中，通过指定索引模式“metric-*”，可以使用仪表板服务进行查询和可视化。此索引包括系统和工作流任务的直方图和详细性能信息，可用于分类服务性能问题。“access- *”索引模式也可用，并包含管理和搜索应用程序的可搜索访问日志事件。

4.8. 部署

内容智能洞察平台的部署涉及以下步骤：

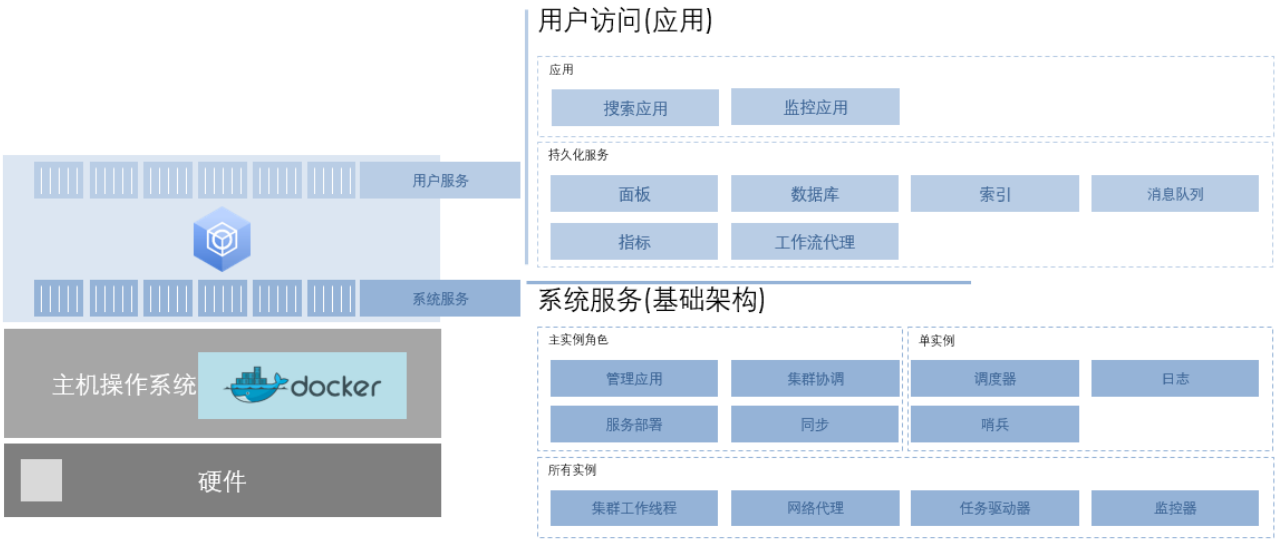
1.首先选择运行主机操作系统（64 位 Linux）和 Docker 的硬件或虚拟机管理程序

2.将产品二进制文件复制到群集中的所有实例

3.执行产品脚本以在群集中的所有实例上启动系统进程

4.跨运行实例集群部署和管理用户服务

图 3：服务的部署



启动内容智能洞察的实例时，首先部署容器化系统服务。这些服务处理系统的基础架构相关任务，支持系统管理，日志记录，网络代理，度量收集，业务对象管理和集群硬件资源协商等功能。 这些系统服务也称为低级服务。

部署低级系统服务后，管理应用程序可供最终用户使用。 在此，管理员可以在部署特定产品服务（也称为高级服务）之前确保系统的所有实例都正确通信并正确配置。高级服务由产品管理员管理，可以通过定制产品服务部署进行扩展和配置，以满足特定用户案例或工作负载。

例如，如果“搜索”应用程序开始指示较高的用户负载，则产品管理员可以使用“管理”应用程序将“搜索”应用程序的其他实例扩展到资源可用的其他实例。由于系统会自动在应用程序实例之间对请求进行负载平衡，因此最终用户可以看到该服务的整体性能和响应时间得到改善。如果随后不再需要额外的服务实例，则还可以相应地缩小它们以重新分配系统资源。

4.9. 监控

Admin 应用程序提供 REST API，CLI 和 UI 仪表板，用于监视系统的当前状态。

这包括：

- 概述系统中的所有实例，其运行状况，负载平均值以及 CPU /内存/磁盘利用率
- 所有已部署实例，健康状况，负载平均值，及其运行状况以及 CPU /内存/磁盘利用率的概况
- 主实例与工作实例的细分
- 特定实例和服务的详细信息
- 关键系统问题的告警

4.10. 安全

身份提供者

内容智能洞察软件在管理应用程序中提供服务，以便自定义管理身份验证和授权。首先，管理员通过选择和配置任何可用的插件实现来向系统注册身份提供者（Identity Providers）。身份提供者可以是组织使用的现有目录服务器系统，或者可以作为基于云的认证服务提供。

接下来，可以使用身份提供程序发现组并将其映射到内容智能洞察平台。使用远程目录服务器管理单个用户的组成员身份，并且只需在内容智能洞察平台中注册。然后可以为这些注册的组分配一个或多个角色。角色是一组或多组权限。每个服务都注册一个由系统强制执行的自定义权限集。管理员将权限组合到他们希望用于特定应用程序的自定义角色中，并将这些角色中的一个或多个分配给已注册的组。

只有已配置了特定角色的组才具有访问系统中相应服务 API 集所需的权限。授予用户权限时，UI 的相应区域将变为可用，并允许 REST API 请求。禁用用户权限时，管理 UI 还将动态删除 UI 的这些部分以防止访问这些部分，并且对这些服务的任何 REST API 或 CLI 请求都将失败并显示错误。

登录到内容智能洞察平台的应用程序时，用户可以选择要使用的安全领域，这将使用相应的身份提供程序进行身份验证。 管理员选择与每个身份提供者关联的每个安全领域名称。

然后该应用程序将：

- 针对选定的身份提供商对用户进行身份验证
- 如果成功，请确定该用户所属的组
- 给定组成员身份，确定已为该用户分配了哪些角色和权限
- 强制该用户可以访问已被授予访问权限的产品功能，并且无法访问未被授予访问权限的功能。

SSL 证书

Admin 应用程序提供集中的 SSL 证书管理，以保护服务器到服务器的通信，以便作为客户端和服务器进行应用程序和数据交换。

内容智能洞察软件附带自己的自签名 SSL 服务器证书，该证书在安装系统时自动生成并安装。 Web 浏览器不会自动信任此证书。管理员可以选择信任此自签名证书，将其替换为证书颁发机构（CA）中的证书，或创建自定义证书。系统可以选择生成并安装新的自签名 SSL 服务器证书。

例如，如果当前证书即将到期并且您正在等待从 CA 检索新证书，则可以执行此操作。还包括用于客户端的单独证书存储，因为内部内容智能服务也利用 SSL 的外部服务。

4.11. 插件

插件是模块化的代码片段，允许内容智能洞察软件执行特定的活动。这些插件可以以插件包的形式上传到系统中，这些插件包将把一个或多个插件的实施与系统解释其内容的详细清单结合起来。插件包可以直接上传到管理应用程序，该应用程序处理所有多实例分发和激活新安装的插件。在上传和激活关联捆绑后，可能会立即使用插件提供的功能。插件允许引入或更新系统提供的功能，而不需要更新或重新启动系统本身或单个服务。

许多系统组件（包括数据连接、管道级联、电子邮件/系统日志通知、聚合、索引器和身份提供程序）都是使用插件实现的。因此，可以通过添加新的现有插件或编写自己的自定义插件来扩展系统功能。提供的 "插件 SDK" 可用于构建和测试数据连接和管道处理级联插件。

4.12. 更新

内容智能洞察软件可以通过使用管理应用程序安装更新包进行更新。此更新包包含用于更新软件组件的产品、服务和版本特定说明，并将根据需要在每个实例上自动执行这些说明。这些更新包可能包含新的或更新的服务和/或插件版本，并且可以扩展系统的功能。更新包是每个新产品版本提供的单独文件。

上传和执行更新包后，系统首先执行一系列特定于版本的预检，以确保给定的系统能够满足更新。如果这些预检查中的任何一个失败，更新操作将失败并出现错误，并且系统将不会更改。但是，如果这些预

检查成功，系统将自动执行更新过程以完成。产品级别的更新是单向过程，系统不能降级到以前的版本。在更新期间，系统将进入只读状态，服务在更新到新版本时可能会定期短暂脱机。报告更新过程中发生的任何错误，如果需要，可以根据需要重试。更新操作完成后，系统将恢复到正常运行状态。

5．数据处理引擎

内容智能洞察平台中的处理引擎可用于:

- 连接到感兴趣的数据源
- 在数据源上执行自定义操作，如读取、写入或删除
- 处理在每个数据源上找到的元数据和原始数据流
- 使用其他结构化元数据对数据进行分类、标记和扩展
- 从非结构化原始数据流中提取感兴趣的信息
- 用于分析和通知目的的聚合内容
- 将处理后的信息发送到输出目标，如搜索引擎

5.1. 数据连接

为了连接和检查数据，内容智能洞察软件平台具有称为数据连接的组件，它使用这些组件访问数据的存储位置（称为数据源)。数据连接包含系统从数据源读取文件所需的所有身份验证和访问信息。许多数据连接实现可执行的可配置操作，这些操作可用于创建、删除和修改数据源中的数据。

内容智能洞察平台包括许多存储平台的内置数据连接器，可以连接所有数据类型，包括结构化数据和非结构化数据源，例如私有云中的对象存储系统、NAS 存储系统、HDFS、数据库（例如 PostgreSQL、MySQL、MariaDB、Oracle、SQL Server)、邮件系统、Windows 服务器、消息队列（例如 Kafka 消息队列)、公有云中的对象存储（例如 AWS S3 存储系统)。如果没有现成的数据源连接器，可以通过使用提供的 "插件 SDK" 编写自定义数据连接插件，为其他类型的存储平台添加支持。

5.2. 文档

数据连接使用称为 "文档" 的数据结构在任何数据源中显示数据。

文档是一段数据及其关联的元数据的表示形式。任何数据都可以用这种形式表示，为高级内容处理提供了一个规范化机制。文档由任意数量的字段和/流组成。

字段是与数据关联的单个元数据键/值对。例如，医学图像可以成为包含字段值对的文档，如 "医生=张山" 和 "位置=省人民医院"。

这些字段是文件的元数据，可用于处理和构造搜索索引。字段可以 (可选) 被强行归类，但仍可以以其原本的字符串形式计算所有字段。

流是指向位于文档本身之外的另一个位置的原始数据字节序列的指针。流通常指向大量数据，这些数据作为文档字段加载到内存中的成本高得令人望而却步，例如大型 pdf 文件的全文内容。内容智能洞察不需要耗费系统资源通过管道传递大量数据，而是使用流访问数据并按需从所在的位置读取数据。这是通过评估每个数据连接中的流元数据来确定要提取到系统中进行流处理的数据来实现的。这些数据流通常在系统中进行分析，而不需要将流的全部内容加载到内存中。

5.3. 管道

从数据连接返回的文档可以表示 pdf 文件、数据库行、图像、视频、传感器读取，甚至还可以表示日志文件。任何搜索应用程序都首先要求对这些不同类型的数据的元数据字段和内容流进行规范化，以支持跨它们的查询。

处理管道从数据源提取的文档执行操作，从而允许对所包含的信息进行转换和规范化。内置管道是开箱即用的，可用于执行基本的搜索内容处理。可以根据需要克隆和自定义这些管道。

处理管道由一个或多个级联组成，每个级联执行特定类型的操作。例如,内容智能洞察平台包括展开.zip 文件 ("zip 扩展" 级联)、向文档中添加新字段 ("标记" 级联) 或在给定纬度和经度坐标 ("地理编码" 级联) 的位置详细信息的阶段。如果您有一个只想影响某些文档的级联，则可以用条件语句包围该级联。不符合指定条件的文档将在处理时绕过该级联。自定义级联插件可以使用 "插件 SDK" 开发。

5.4. 内容类

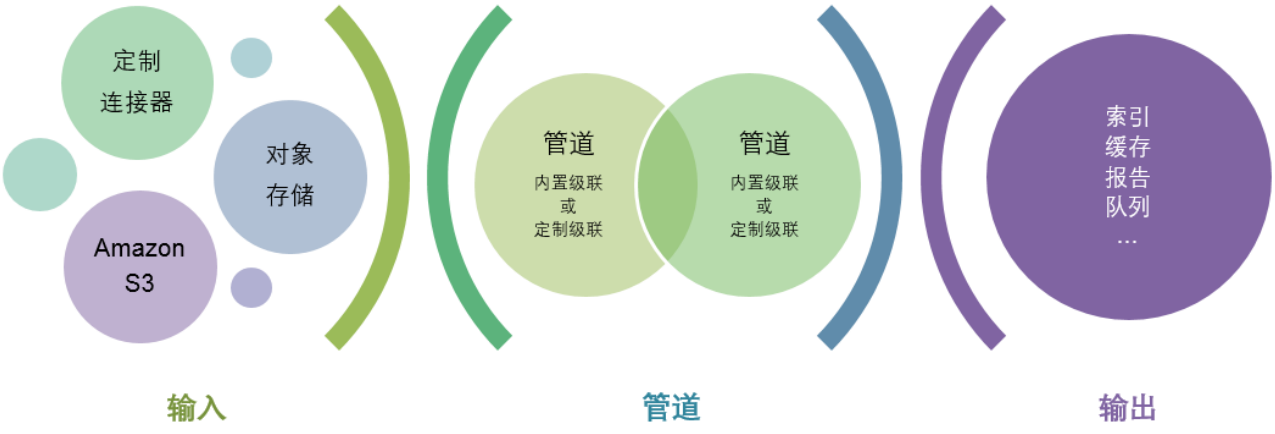
通过使用内容类从原始内容流中提取其他结构化元数据，可以使数据更易于搜索。内容类是处理表达式（称为内容属性）的命名集合，可应用于数据以提取在数据流中也标识的类中定义的任何信息。xml 或 json 文档中包含的信息可以使用内容类轻松提取、混合、转换和规范化，而无需实现任何解析逻辑。正则表达式模式也可用于创建其值与感兴趣的特定信息模式（如社会保障号码或电子邮件地址）相匹配的信息字段。每个内容类标识可能在文档数据流中识别信息位，并将这些信息提取为可在后续处理过程中使用的附加文档字段。然后，还可以对这些字段编制索引，并在搜索中使用这些字段。

在预先知道要提取的所需信息或可以从已知示例生成的情况下，内容类非常有用。提供了自动生成和测试内容类的工具，这些工具提供了 xml 或 json 结构化文档的示例，这些文档很可能会在处理后的数据集中找到。通过利用内容智能导入功能，内容类也可以轻松地移植到其他系统，允许任何系统使用其他人为提取特定信息类型而生成的内容类。

5.5. 工作流

通过工作流，您可以选择要处理的数据、如何处理这些数据以及如何处理结果。它们通过将数据连接、处理管道和索引集合相互关联来实现此目标。

图 4：用户定义工作流输入、管道（用于处理）和输出

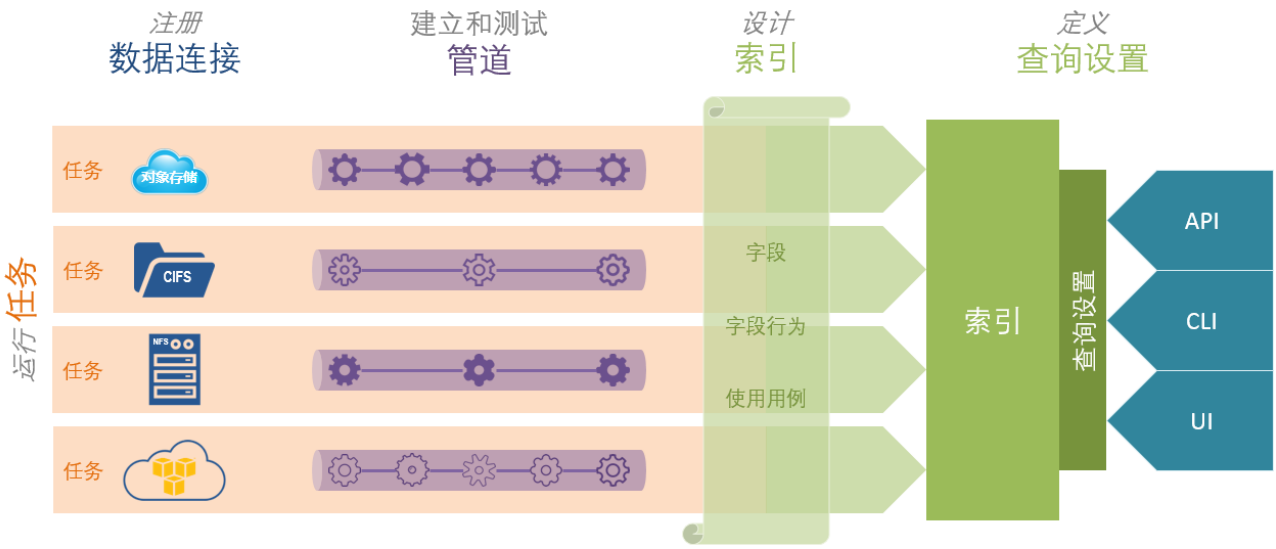


通过将多个管道添加到工作流，可以将它们链接在一起，从而形成工作流管道。

若要让内容智能洞察平台执行工作流定义的工作，需运行工作流任务。工作流任务可以运行一次，也可以计划定期运行。

当工作流任务运行时，"工作流代理" 服务的新实例将在已配置为运行此服务的所有实例上启动。在其中一个实例中，"生成器" 通过利用每个相应的数据连接，发现并提取每个工作流输入中的文档。这些发现的文档中的每一个都通过工作流管道发送进行处理，并将结果发送到每个配置的工作流输出。此工作流任务可以连续运行、按需运行，也可以计划仅在指定时间内运行。

图 5：工作流任务协调数据的收集、处理和索引



5.6. 管道执行模式

可以将工作流配置为以两种执行模式之一运行管道："预处理" 或 "工作流代理"。

工作流代理执行模式的工作原理是将输入文档的批次分发到配置为运行 "工作流代理" 服务的所有实例，从而允许并行执行多个文档的指定处理级联器。运行此服务的实例越多，可以并行处理的文档就越多。这样就可以通过添加或删除系统实例来简化资源的扩展。此模式优化管道以进行大规模并行处理，当内容已处于应搜索的格式时，此模式最有用。例如，如果数据源只包含应单独搜索的图像或音频文件，则

工作流代理模式下的管道将通过跨系统实例并行处理所有文档来提供最佳的处理性能。添加实例线性地扩展处理性能，因为其他实例可以与其他实例并行地参与工作流执行。

相反，通过在并行处理发生之前将数据转换为所需格式，对预处理模式进行了优化，以准备数据。它通常用于将容器文档（zip、tar、pst 等）扩展到其中包含的可搜索内容中。此模式不是并行处理所有输入文档，而是在 "工作流代理" 服务的单个实例上连续执行管道级联，以便快速生成可输入到后续并行处理管道中的文档。产品中提供了 "基本预处理" 管道，可自动处理最常用文档格式的扩展，并且几乎应始终在预处理执行模式中使用。

预处理和工作流代理管道都可以在单个工作流中使用，以优化大规模内容处理的两个方面。例如，如果工作流输入提供了引用大型日志文件的文档，但处理目标是生成可搜索的日志事件，则使用 "读取行" 级联将这些日志文件扩展到单个日志事件通常是最有效的在预处理管道中。此级联器将日志文件文档转换为每个单独的日志行文档，然后可以将这些文档与所有可用实例并行发送，以便通过 "工作流代理" 模式下的管道将其解析为可搜索的字段。

5.7. 聚合和触发器

对数据流执行高级计算可能具有挑战性，尤其是在数据集非常大的情况下。工作流提供了一种内置机制，用于在处理数据时自动计算、聚合和报告信息，而无需最终用户生成自定义代码来执行此工作。

聚合功能允许注册一组可自定义的聚合器，其中包含在处理时更新的工作流。这些聚合器是通过聚合插件实现的，这些插件可以对工作流中的每个输出文档执行任意操作。这些聚合收集的信息将在工作流 UI 和可下载的工作流报告中保留并提供。

内容智能包括几个内置的可自定义聚合插件，包括 "平均值" (计算字段的所有值的平均值) 和 "字符串计数" (保留为特定字段找到的唯一值的直方图)。

默认情况下，所有新创建的工作流都会自动配置 3 个有用的聚合:

- 发现的字段：计算所有已处理文档中出现的字段，并提供建议的类型信息（用于自动创建架构）和用于管道创建的示例值。
- MIME 类型：计算并报告标准 "内容_类型" 字段的所有发现值。
- 扩展名：计算并报告在 "CHDI _文件名" 字段末尾发现的所有发现的文件扩展名。

当聚合值满足用户定义的条件时，工作流触发器允许触发自定义管道的执行。触发器可用于根据数据处理过程中遇到的一个或多个聚合结果启动通知或开始执行其他管道逻辑。利用内置电子邮件或系统日志通知管道级联器时，工作流触发器提供了一种机制，以便在聚合值超过特定阈值或检测到其他感兴趣的指标时通知用户。例如，如果包含温度传感器读数的文档由工作流处理，则在滚动平均值超过配置的最大值或低于预期最小值的情况下，可以使用触发器通知管理员。由于触发的管道可以使用与任何其他处理管道相同的级联器，因此它们可以利用操作级联来自动执行一组纠正措施。

6．高级文本和元数据搜索

内容智能索引和查询引擎可用于:

- 扩展和监控本地管理的搜索引擎
- 创建和管理搜索引擎索引
- 在外部搜索引擎中注册、创建和管理索引
- 设计、管理和优化索引架构
- 将处理后的工作流输出文档索引到搜索引擎
- 对内部和外部搜索引擎索引执行简单、高级和联合查询
- 管理特定用户组的索引、文档和字段级别的访问控制
- 为访问同一索引的不同用户组配置自定义搜索体验
- 动态调整特定用户组的查询结果相关性顺序

6.1. 索引

索引是一组用于生成、配置和管理搜索引擎索引集合的说明。每个内容智能洞察平台的索引都包含一个架构、索引文档集合和一组查询设置。

索引集合通常由 "索引" 服务创建和管理，该服务允许内容智能洞察平台轻松监视、缩放和管理搜索引擎。或者，外部搜索引擎索引也可以在系统中注册。支持由外部 Apache Solr、ElasticSearch 等搜索引擎管理的索引，并可在内容智能洞察平台中注册，以便针对每个搜索引擎启用查询、索引和/或其他管理操作。链接索引。

6.2. 索引架构

索引架构是包含每个字段的配置设置的字段列表。这些设置指定字段的类型、在索引文档上遇到该字段时的分析方式，以及该字段在搜索引擎中支持的查询功能。例如，可以将字段配置为支持层面搜索，并允许用户根据该字段的值对搜索结果进行排序。

搜索引擎索引始终存在架构，即使在使用可用的 "无模式" 模式时也是如此。这些模式尝试自动创建索引架构中的字段，而不特别要求提前定义这些字段。但是，系统仍基于索引架构中定义的字段的配置运行。创建和优化索引架构可能需要在了解基础搜索引擎技术方面进行大量投资。内容智能洞察平台提供了更高级别的管理工具，允许按字段用例（而不是基础搜索引擎的索引字段属性）管理索引架构。这种方法，再加上能够从工作流聚合报告或管道测试结果自动配置优化的索引架构字段，大大简化了这一任务。

系统提供了一组用于应用初始架构的架构模板配置：

- 无模式：当在索引文档上发现字段时，将自动创建字段配置。当您希望通过索引和查询来了解数据中包含的信息而不需要预先进行架构培训时，此模板非常有用。但是，此模板在生产中可能会出现问题，因为它将所有发现的字段添加到架构中（尽管它们可能没有用），并且字段类型检测逻辑可能会选择错误的字段类型来满足所有文档，从而导致查询低效率和指数膨胀。建议仅在入门时使用此模板。
- 默认值：与无模式模板类似，但可缓解自动添加字段所导致的许多问题。此模板还会自动配置不存在的所有字段，但许多公共字段已作为预定义类型添加到架构中，以避免不正确的类型选择问题。如果您正在开始使用工作流中的默认管道，或者如果您的索引字段无法事先知道，则应使用此模板而不是无模式。
- 基本：只预定义唯一身份证件所需的最小字段。字段不会自动添加到索引集合架构中，因此您需要使用提供的内置便利工具手动添加和配置字段（测试文档或运行工作流任务可以生成可批量添

加到索引集合架构)。为文档编制索引时，不会对任何与架构不匹配的字段编制索引。这是生产索引集合的首选初始模板。由于字段是显式定义的，因此只有所需的信息存储在索引中，从而优化了特定用例的索引。

6.3. 索引文档

工作流既可用于处理文档，也可将其发送到搜索引擎索引中。当文档发送到搜索引擎进行索引时，将尝试将这些文档中的任何字段和/或流加载到索引架构上定义的等效索引字段中。根据架构中定义的字段类型，搜索引擎可能会在存储之前对信息进行进一步分析。例如，原始文本流可能会标记为每个单独的单词，删除常见的术语（如 "如果" 或 "和" 以提高搜索相关性），并进一步规范化为小写字母，以更好地匹配查询中的任何用户输入。

当工作流配置索引输出时，处理过的文档会批量发送到搜索引擎以编制索引。批处理的使用消除了与搜索引擎索引通信中的往返瓶颈，并大大提高了整体索引性能。索引请求会自动调整，以确保在成功检查内容智能工作流进度时，索引文档在查询结果中可见。

6.4. 索引扩展

当搜索引擎索引变得太大，无法满足用例的业务需求时，应相应地缩放索引。应缩放索引以允许它超出单个计算机的范围，以满足应用程序的高可用性要求，以满足高索引查询卷，或这些的任何组合。

6.5. 索引复制

为了满足高可用性和查询响应时间要求，应通过复制缩放索引。

此过程涉及在其他服务器上创建索引的其他副本。一旦存在这些副本，系统就可以在每个索引副本之间对查询请求进行负载平衡，以满足较高的查询负载，同时确保抵御单个实例故障的恢复能力。主索引副本接收传入的索引文档，并且副本会随着时间的推移尽快进行异步更新。此方法在索引性能和查询响应时间之间提供了很大的平衡。

在内容智能洞察平台中，索引复制是通过将索引服务配置中每个索引的索引保护级别（IPL）设置从 1 个副本（默认值）更改为多个副本来完成的。此计数确定要在所有索引服务实例中存储的索引副本数。系统会在每个正在运行的索引服务实例之间自动平衡这些副本。可以随时动态地增加或减少该设置，并导致添加或删除索引副本。其他索引副本可提高每个搜索索引的可用性，并导致分布式查询负载（允许在不降低性能的情况下进行更多并发查询)。但是，它确实将每个新索引副本的磁盘和内存要求增加一倍。确保分配了足够的资源来满足所需的保护级别。

6.6. 索引分片

如果一个索引预计将增长到 600 万个或更多的文档，则应通过分片对索引进行缩放。

系统支持将索引拆分为较小的段（称为 "分片")，这些段可以在群集中的实例中动态分布。这使得索引变得非常大，因为如果实例空间不足或负载过重，则可以将分片平衡到新实例。

索引分片计数是在索引创建过程中设置的。建议每个正在运行的索引服务实例至少有一个分片，以最大限度地提高并行查询性能。

在创建索引以允许索引增长非常大时增加分片计数（分片可以平衡到其他索引服务实例)。如果您的系统有望增长，则应过度分片，以便在将来将额外的分片无缝地平衡到其他实例。例如，如果索引实例计数预计会随着时间的推移而增加一倍，则初始分片计数将增加一倍。将此分片计数增加一倍将使系统能够通过将段平衡到这些实例，将运行索引服务的实例数量增加一倍。

分片的一个缺点是，单个查询请求仍然需要访问所有索引分片，以便以所有可能的结果正确响应。分片越多，需要单独查询索引段以响应用户请求的索引段就越多。

因此，建议将分片和复制结合起来，以确保可以适当地吸收查询负载，并消除单点故障。

6.7. 查询设置

成功编制文档索引后，最终用户可以使用搜索应用程序使用提供的 REST API、CLI 或 UI 跨一个或多个索引进行查询。

为了自定义搜索体验，内容智能洞察平台的索引引入了查询设置的概念。

每个索引定义一组或多组查询设置。这些查询设置决定了用户在搜索索引时的总体搜索体验。例如，查询设置可用于指定用户可以在搜索结果中看到哪些字段、他们在结果中看到的返回的名称或搜索结果的显示方式。

索引可以定义多组查询设置。每个查询设置都可以直接映射到一个或多个用户组。每个索引都包含一个内置的公共查询设置，该设置（启用时）定义默认情况下所有用户查看索引的方式。然后，可以针对特定的用户集克隆和自定义这些设置，因为可以使用特定的索引查询设置将每个标识提供程序组映射到。这允许特定的用户组在同一组索引数据中具有不同的自定义搜索体验。查询设置确定公开哪些索引字段、显示哪些方面、支持哪些优化操作、是否遵守访问控制以及许多其他配置。

还可以使用查询设置强制实施文档级安全性的访问控制。查询设置可以定义添加到文档中的访问控制列表（ACL）字段是否在查询中自动获得（例如，通过管道中的 "文档安全" 阶段）。ACL 包含用户和/或组的列表，并指定是否允许或拒绝这些用户或组访问关联的文档。

每个查询设置还可能具有自定义相关性要求。为了支持查询结果的自定义相关性排序，查询设置提供了一个设置，以启用各种相关性提升。推入是增加或减少该查询设置的用户的索引中某些字段匹配的相关性的语句。例如,如果索引文档具有 "高"、"中" 和 "低" 值的 "重要性" 字段，则通过提升标记为非常重要的结果，可以自动将更重要的搜索结果移至结果的顶部。或者，重要性较低的结果可以降低。查询设置相关性只会影响结果的相关性排序，并且不会更改给定用户查询返回的结果列表。

搜索应用程序包括内置的呈现支持缩略图、元数据键值、图像、视频、音频，甚至每个单独的搜索结果中的原始 HTML。可以使用查询设置 "显示" 工具自定义搜索结果布局，而无需额外的 UI 编程。这些工具还可用于自定义搜索索引中的哪些字段驱动搜索结果中的各个字段，如 "标题" 或 "标题链接"。这使得内容智能洞察平台中的内置搜索应用程序能够满足尽可能多的用例。

6.8. 搜索应用

内容智能洞察平台的搜索应用程序提供了可自定义的 REST API、命令行和用户界面，用于查询搜索引擎索引。

搜索应用程序提供多种智能的数据搜索:

- 通过标准的查询解析器执行全文搜索：可以通过输入一个单字、一个词语，或一组短语执行全文搜索，当输入一组短语时，需要在短语的首尾使用双引号进行标记
- 查询单个搜索引擎索引，或并行查询多个索引
- 多种逻辑组合条件的搜索：也就是通过“与、或、非”等逻辑操作符组合多个查询条件进行搜索。使用“与”表示多个条件同时满足，使用“或”表示只需要其中某个条件满足，使用“非”表示不满足该条件的搜索。
- 根据索引中的短语自动完成建议
- 元数据搜索：通过文件和元数据提取器采集对象存储中对象的所有元数据（包括系统元数据和自定义元数据）之后创建索引，执行针对元数据的搜索。
- 使用查询优化构建结构化字段查询
- 模糊搜索：通配符搜索功能，可以通过?（替代单个字符）和/或*（替代连贯的字符串）等通配符进行搜索；范围搜索功能，可通过正则表达式执行时间范围检索、数值范围检索、字符串范围检

索等。当需要查询某个范围之间的子集时，这个功能非常有用。例如，如果只想查询 2012 年 2 月 1 日到 2012 年 8 月 1 日这六个月之间的文档，就可以使用范围搜索；

- 临近搜索：对于很多搜索应用来说，很重要的功能是不仅仅需要精确匹配用户的文本内容。而且还允许一些灵活的变化，比如一些用户的拼写错误或相同单词的其它变体。通过在首尾两个关键词之间指定编辑距离，把中间的所有结果进行查询显示。例如，查询语句“首席 官”~2，编辑距离为 2，将返回“首席执行官”、“首席财务官”、“首席运营官”等系列满足条件的结果
- 显示自定义结果，其中包含元数据、缩略图、图像、视频、音频和 HTML 显示的丰富代码段
- 以**字段:值**的格式输入查询条件，精准返回所有符合条件的结果，例如标题:“季度报告”为查询条件，这个查询条件将输出所有标题是季度报告的文件
- 下载具有特定元数据选择的搜索结果
- 搜索时采用多种智能分词方法，包括基于自然语义的分词和基于词库的分词（支持导入各种行业词库或专业词库），可对分词进行自动的词性标注（例如该分词是动词、名词、副词、形容词等的哪一种）。基于自然语义的新词发现可不断添加到词库，以应用于后续的分词和索引中。此外，支持基于多种语言的分词，例如中文简体、中文繁体、英文等多种语种。
- 对搜索结果排序：可以基于一个或多个信息类对搜索的结果进行排序。例如，基于某个作者创建的所有文件按照文件的创建时间进行排序，也可以同时对所有 pdf 类型的文件按照创建时间进行排序。此外，还可根据被搜索的关键词的相似度进行排序，也就是通过内嵌的相似度算法，根据搜索的关键字词频计算文档相似度，并根据相似度进行升序或降序排序。
- 通过对搜索结果添加过滤条件以缩小结果的范围，可以在初次搜索结果中添加文件类型过滤搜索结果，例如仅显示 pdf 类型的文件。

- 建立索引时基于每个文档的内容采用 MD5 加密算法计算出对应的唯一 hash 值，然后基于该 Hash 值实现搜索的消重，避免对相同的文档执行搜索。
- 实现自动文摘：在搜索结果中可突出显示（高亮度显示）与用户查询相匹配的“文档片段”或“段落”。突出显示是可配置的,可以设置片段的大小、格式、排序，以及更多难以分类的选项。
- 通过 SQL 解析器提供基于 SQL 语句的查询，此外还支持多个并行的 SQL 查询操作。SQL 接口支持从 SQL 客户端以及数据库可视化工具（如 DbVisualizer 和 Apache Zeppelin）发送的查询请求。
- 对多种文件格式的搜索操作，包括 Office 文件（Word/PowerPoint/Excel）、pdf、文本文件、XML 文件、JSON 文件、压缩文件（包括 zip 文件、tar 文件、tar.gz 文件、tar.bz 文件等类型）、eml 文件、pst 文件等
- 通过智能提示简化搜索条件的输入，当输入个别单词的时候，能够自动联想提示，避免冗余输入并提供一定的推荐，从而实现更好的交互效果。
- 语义扩展搜索：提供被搜索词扩展之后的相关内容，例如搜索“Java 读文件”时，可以同时搜索出“Java 读取文件”、“Java 读写文件”、“Java 文件读写操作”等内容
- 查询结果的统计分析：包括对数字型、字符型和日期型字段进行统计分析。例如，对于数字型的字段，支持按照最小值、最大值、所有值之和、个数、空值个数、平均值、标准差、平方差、平方和等进行统计
- 分层搜索：允许搜索用户通过对字段值应用多个筛选器来浏览文档。在索引文档中找到的任何元数据字段都可以通过索引查询设置进行分层搜索。

通过在索引查询设置中配置字段来启用结构化查询生成器，以启用优化。优化允许最终用户基于类型化字段构建结构化查询，而不需要了解查询语言。选择活动优化时，可以通过更改为 "高级" 模式来查看和编辑生成的查询语言字符串。为某些字段类型（如日期选取器）提供了 UI 图形界面小工具。

由于搜索引擎可以轻松地支持结构化和自然语言查询的组合以及近乎实时的响应，因此它们越来越多地参与驱动分析。通过绘制系统中任何方面的图形功能来演示此功能，从而为探索索引数据提供了额外的视觉导航选项。

CHANGHONG

北京

北京市丰台区南四环西路188号18区26号楼长虹科技大厦 邮编：100738 电话：010-58292000 传真：010-58292000

上海

上海市静安区北京西路1701号静安中华大厦602单元 邮编：200040 电话：021-62889117 传真：021-62889115

广州

广州市天河北路898号信源大厦3408室 邮编：510898 电话：020-38182838 传真：020-38182835

深圳

深圳市福田区华强北路群星广场B座2408室 邮编：518000 电话：0755-25327693 传真：0755-83534550