The IBM Advantage for Discovery Cloud Architecture

Table of Contents

Executive overview

Data is the fuel of business innovation in proportion to the increase in the amount of data that is available. Sensors, video, news and social media streams, and weather data are only a few of the sources of data that are available to an enterprise, in addition to their private stores. The organization that can tap those sources, separate the valuable information from the noise, see relationships and patterns in the data, and then act upon this knowledge is best prepared to overtake its competitors.

Traditional approaches to data analytics and knowledge management typically help with specific kinds of tasks that are related to structured data. The sheer amount of unstructured data that is being produced means that human physical capacity is quickly overwhelmed by the effort to collect and curate it. New techniques that use natural language processing (NLP), visual recognition, and other elements AI can help to identify and organize unstructured data. This is where computing with artificial intelligence (AI) comes in. IBM's AI services are trained by humans to augment and amplify human cognition. The systems are not designed to replace a human's cognitive capabilities but to enhance them. For example, a system that is trained by a legal expert to sort through thousands of files of unstructured data to identify files that are pertinent to legal claims can do it faster than a person, freeing up the expert for higher value activities.

AI systems can be transformative. A business can change how it operates when the proprietary content and expert knowledge of the organization are extended into the enterprise through applications that include NLP, hypothesis generation, and evidence-based learning. Strategic and day-to-day decisions are better informed, leading to better business outcomes. Best practices encourage the use and embedding of decision making with AI into current processes and into the creation of new processes.

The tools that are used to achieve these results have evolved to meet the expectations of the enterprise. Both highly structured and unstructured data must be used. Especially in the text-heavy, unstructured data domain, there is a natural and cumulative evolution from basic search to AIsearch through NLP and machine learning, with the goal of delivering deeper insights more accurately, faster, and at a greater scale.

IBM® Watson® Discovery is designed to make it more efficient to identify, collect, and curate text-heavy unstructured data. This can simplify the human use of information through more efficient access to large content stores or through the integration of the service in support of larger AI systems.

Before NLP and contextual search applications were available, keyword searches were

the way that users engaged with masses of information. Previous approaches to the enterprise management of information, which were launched under the banner of knowledge management, relied on the creation of complex content topologies, huge internal indexes, and the speed of the keyword search. These projects were not adopted, widely due to the level of effort that was required for basic results. The following diagram shows how value to business increases with the adoption of more sophisticated techniques for search and analysis.

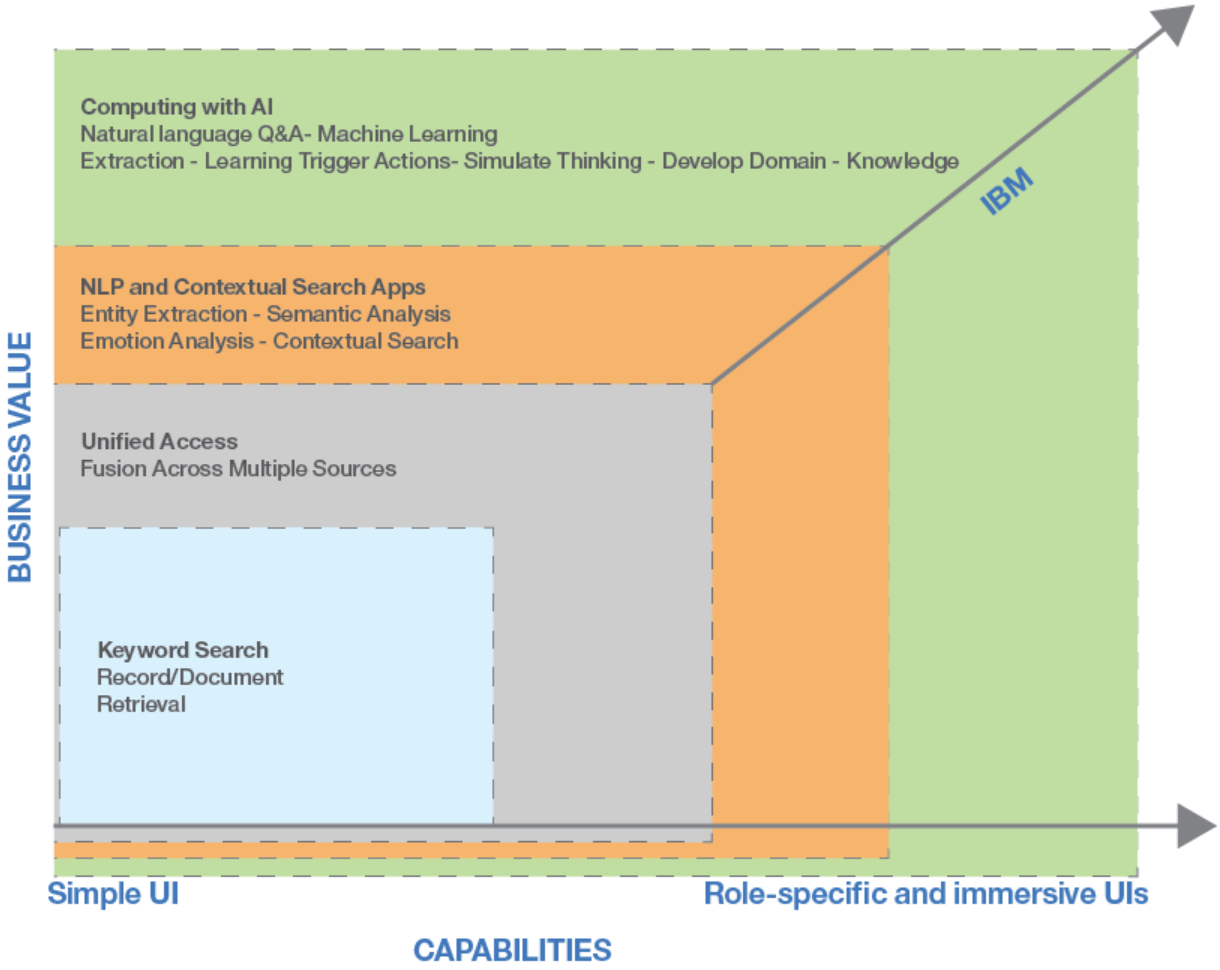


Figure 1: Business value of search and analysis

To create content collections and custom AI applications that follow IBM's approach for curating content and designing AI applications, you can use Watson Discovery on the IBM Cloud® or IBM Cloud Pak®. Before you get started, it's important to understand the relationships between business processes and technical architecture components that use cloud computing infrastructure, platforms, and services. It also describes the kinds of personnel and activities that are required to prepare and implement a Watson Discovery solution that can evolve with the organization. If you are new to AI systems, you can familiarize yourself with

fundamental AI concepts such as ground truth, training set, and test set by reading the AI glossary.

IBM Cloud Customer Architecture for AI

As shown in Figure 2, IBM's AI reference architecture can be categorized into three broad capabilities:

1. **Discovery**: IBM's AI discovery capabilities ingest and enrich information, annotate the information that is stored in multiple documents, and prepare a corpus for discovering insights with ready-to-use AI capabilities for better decision-making. For more information on how these capabilities are realized, see the Data discovery reference architecture.
2. **Conversation**: IBM's AI conversation capabilities are trained to assist in decision-making by using natural language conversation. In situations where there is a conversation or a dialog, IBM Watson Assistant offers an intent-based understanding and a conversation model that is driven by a dialog that can be used to determine the best course of action. For more information, see the Conversational chatbot reference architecture .
3. **Extend:** IBM's conversation and discovery capabilities can be extended by AI services that take broad or unstructured data and create meaningful, actionable, and valuable information that can be domain-specific for users. By using various services or offerings such as IBM Watson Speech to Text, IBM Watson Text to Speech, IBM Watson Tone Analyzer, IBM Watson Visual Recognition, IBM Watson Natural Language Classifier, and IBM Watson Personality Insights, businesses can turn previously "dark" data in the form of contact center recordings, images, unstructured text, and video into valuable, actionable insights and assets.

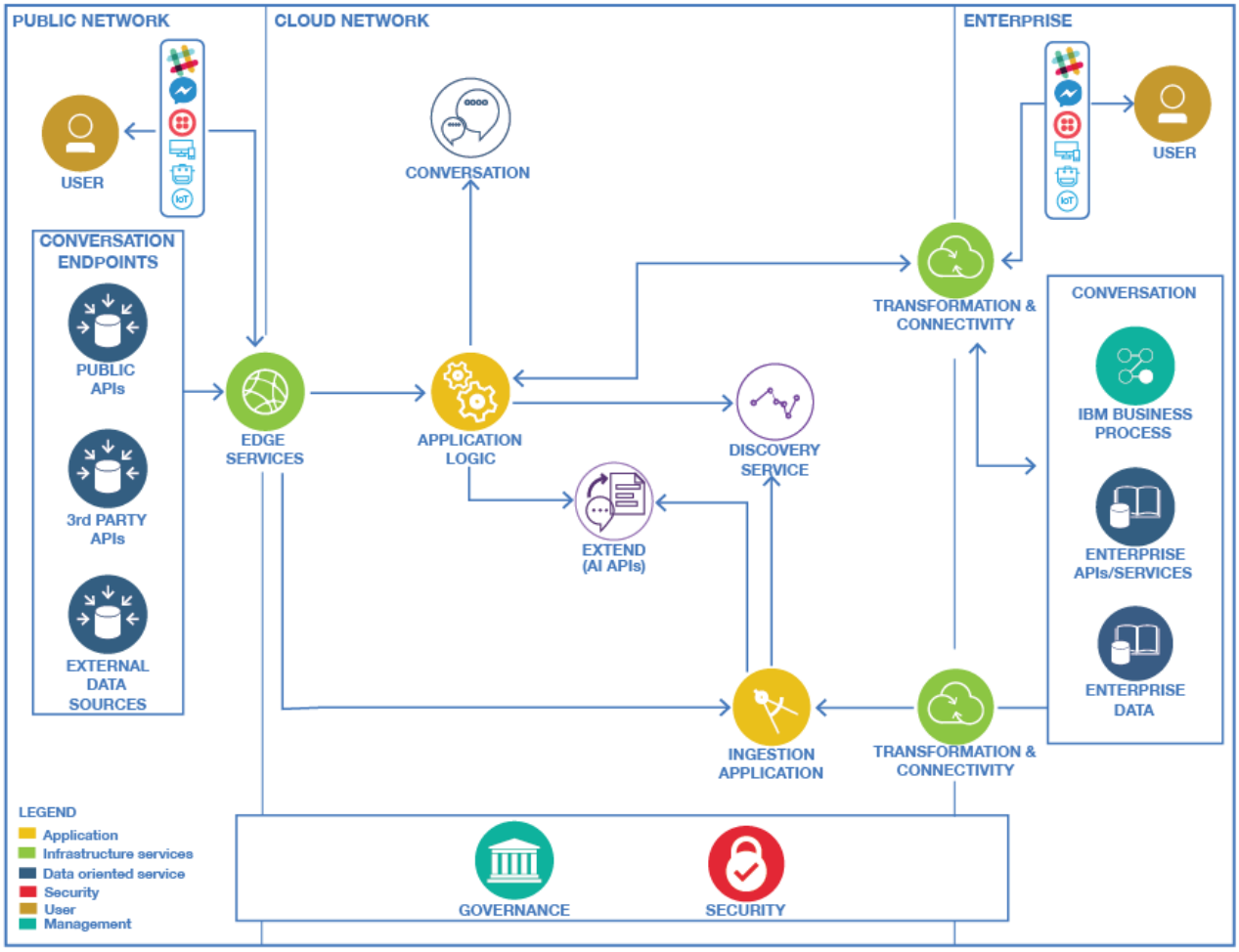


Figure 2: AI reference architecture

The AI reference architecture presents conversation, discovery, and extend capabilities in relation to each other. Watson Discovery extracts value from unstructured data by converting, normalizing, and enriching it. This data can be proprietary, public, or third-party. You can apply various AI-powered information retrieval techniques to identify the best responses to questions on discovery collections. The resulting output of the Discovery service can be used with other services, such as a self-service chatbot or call center agent assistant, automated expert guidance, or self-service knowledge bases, to support business functions or customer support.

IBM AI systems learn from continuous interactions and identified patterns. Watson Discovery is an efficient way to provide both the initial content for a system and the content that is needed to keep it up to date. Watson Discovery is available on IBM Cloud platform and on IBM Cloud Pak for Data, so it can be deployed anywhere to support AI solutions that are designed for any industry.

Watson Discovery is enabled in four phases:

- Phase 1: Data understanding
- Phase 2: Preparation
- Phase 3: Ingestion
- Phase 4: Consumption

Phase 4 is illustrated by a business scenario and sample application, followed by a summary of the practices and components that are necessary for a successful deployment.

What is Watson Discovery?

Watson Discovery, a market leading AI-powered search solution , empowers an organization's experts and knowledge workers with the right information at the right time in the right context. Clients can unlock hidden value in data to find answers, monitor trends, and surface patterns.

Watson Discovery is an insight engine that provides an end-to-end pipeline for ingesting, storing, and enriching data, so that you can query both the original data and the enriched metadata to find relevant answers efficiently. Watson Discovery supports the concepts of projects and collections.

A *project* is a convenient way to build and manage your Watson Discovery application. It can be one of the following types:

- Document retrieval: This project type is used for AI-powered search use cases to search and find the most relevant answers from your data.
- Conversational search: This project type supplies answers to a virtual agent that is built with Watson Assistant.
- Content mining: This project type is used to discover hidden insights, trends, and relationships in the data.
- Custom: You can configure this project type for different purposes if you don't want to select one of the other project types.

A *collection* is set of documents that you upload or crawl into Watson Discovery. It's a mechanism for isolating and organizing documents in Watson Discovery where you can apply consistent configuration and enrichments to all documents within a collection.

Watson Discovery delivers significant business value to clients in these common repeatable use case patterns:

- Contact Center Insights: As the market-leading text analytics platform, Watson Discovery uses AI to generate trends and insights from calls, emails, and chats. Contact center supervisors use those insights to improve operational performance.

- Agent Assist: Watson Discovery assists agents to find the most relevant answers from a large body of complex business documents so that they can guide customers through interactions with the brand.
- Employee Self-Service: Watson Discovery integrates with Watson Assistant to provide quick answers to employee questions and guidance for tasks across a variety of enterprise channels.
- Customer Self-Service: Watson Discovery integrates with Watson Assistant through search skills to provide quick and relevant answers to customer questions from a large corpus of complex business documents.

A few of the key differentiating capabilities of Watson Discovery are as follows:

- An end-to-end AI search function that doesn't require the chaining of multiple APIs.
- Out-of-the-box AI capabilities, such as natural language queries, query expansion, relevancy training, passage retrieval, document similarity, anomaly detection, and more.
- Natural language understanding, machine learning, and deep learning to sort through ubiquitous data and extract insights from complex business documents.
- Continuous relevancy training to improve relevancy of returned results by monitoring user behavior.
- Smart Document Understanding (SDU) to understand the structure of complex documents within minutes of training.
- Integration with Watson Assistant for a powerful conversational experience.
- Point-and-click connectivity to multiple popular data sources such as Salesforce, Microsoft® SharePoint®, web crawl, Open Database Connectivity (ODBC) databases, File Systems, and FileNet.
- Domain customization to learn unique linguistic nuances through integration with IBM Watson Knowledge Studio.
- Element classification to rapidly parse documents to convert, identify, and classify elements of important legal and contractual documents.
- Fully integrated administrative interface and tools that mirror the function with the API.

You can use the passage retrieval, relevancy training, and continuous relevancy training capabilities of Watson Discovery to find relevant answers to complex queries more efficiently. With passage retrieval, you can find information within documents that is relevant to your query. It dynamically selects snippets from within larger documents based on an input query and displays the results. Relevancy training can scale search relevancy by using domain expertise to train the Discovery service on the best ranking of results. Developers and subject matter experts (SMEs) work together to teach the system to find signals in the way that documents and queries are related and to bring the most relevant documents to the top of results.

Developers use continuous relevancy training to automatically update the ranking of returned responses by monitoring user behavior and interaction with the ranked responses from Watson Discovery. If Watson Discovery returns the top 5 responses to a query from the corpus of documents and users consistently choose the third response, the machine-learning ranking model is updated to learn from these interactions so that in future queries, the ranking better reflects the expected

response from users.

Watson Discovery goes beyond search by using AI to enable sophisticated relevancy ranking techniques to return the best results.

Watson Discovery uses IBM Watson Natural Language Understanding, which offers advanced text analysis through NLP, machine learning, and deep learning. This set of APIs can analyze text to extract concepts, entities, keywords, and sentiment.

Watson Discovery can also integrate with IBM Watson Knowledge Studio, which provides an integrated development environment to train domain-specific machine learning models. Watson Knowledge Studio can be used by SMEs who don't have machine learning or data science expertise. SMEs use Watson Knowledge Studio to teach by example, where they annotate representative sample documents with the correct entities and relations that are unique to a domain. Watson Knowledge Studio uses such annotations as ground truth to train machine-learning models that capture the domain-specific knowledge. Watson Knowledge Studio supports several pre-annotators, including dictionary-based and rule-based pre-annotators, to simplify the annotation task of SMEs.

Developers, SMEs, and business experts can use the Smart Document Understanding (SDU) function of Watson Discovery to train Watson Discovery to understand the structure of complex business documents that vary in type, structure, and format. For better retrieval results, it is common to break down business documents into subdocuments based on their structure and then use the subdocuments to return the most relevant answers to queries. Watson Discovery's SDU capability makes this task practically achievable with minimal training effort.

In addition to the AI-powered search function in Watson Discovery, it is expanded to support content-mining use cases. Both AI-powered search and content mining use NLP to enrich documents and extract insights. The key difference is that in search use cases, the user knows what questions to ask while in content-mining use cases, the user doesn't have a specific query but rather wants to understand the patterns that are embedded in the data and extracted enrichments. The content mining solution in Watson Discovery is an analytical tool that helps users explore and discover hidden insights in unstructured data by analyzing anomalies, trends, and relationships in their documents.

Trained IBM AI services or systems can be used in any form factor, including mobile, kiosk, car dashboard, web, voice response unit, or others, for decision assistance. The content collections that are created with Watson Discovery are meant to support any of those form factors.

Data and content collections

An enterprise's unstructured data is proprietary, and with the right tools, the enterprise can harness insights from the right data. But you can't always make informed content decisions based only on this unstructured data. You need to use external sources to supplement the data, such as well-structured enterprise data and any real-time data that comes from sensors and other Internet of Things (IoT) technology. Watson Discovery supports the collection and curation of content from all of these sources.

Organizations aren't limited to static content collections. AI applications interact with people in a natural way to answer questions and provide guidance to help people make decisions. When you develop an AI application for your industry, you want to develop domain expertise from your SMEs, incorporating industry practices and relevant material from relevant data sources.

Watson Discovery can ingest data from various sources. AI enrichments extract insights by using query language. Sources of data might include social media streams, pictures, periodicals, books, and other electronic or print media.

Watson Discovery News, a public data set that is augmented with NLP enrichments, is also included with Watson Discovery. You can use this public, unstructured data set to query for insights that you can integrate into your applications. Watson Discovery News is a data set of primarily English language news sources that is updated continuously, with approximately 300,000 new articles and blogs added daily. This indexed data set is pre-enriched with keywords, entities, concepts, relations, sentiment, and categories. Metadata is also added, including crawl date, publication date, URL ranking, host rank, and anchor text. Historical search is available for the past 60 days of news data.

To support building content-driven solutions, Watson Discovery supports a rich set of popular data types and includes, by default, connectors to popular data sources. Some of the document types supported by Watson Discovery include PDF, Microsoft® Word®, Microsoft PowerPoint®, Microsoft Excel®, TIFF, JPG, and JSON. As for data sources, Watson Discovery includes connectors to Box, Salesforce, Microsoft® SharePoint online, and IBM Cloud Object Storage and a configurable web crawler to crawl and ingest documents from the web.

IBM Discovery reference architecture

The Data discovery reference architecture for discovery shows the two distinct flows--user runtime and data collection—that illustrate a typical use of Watson Discovery.

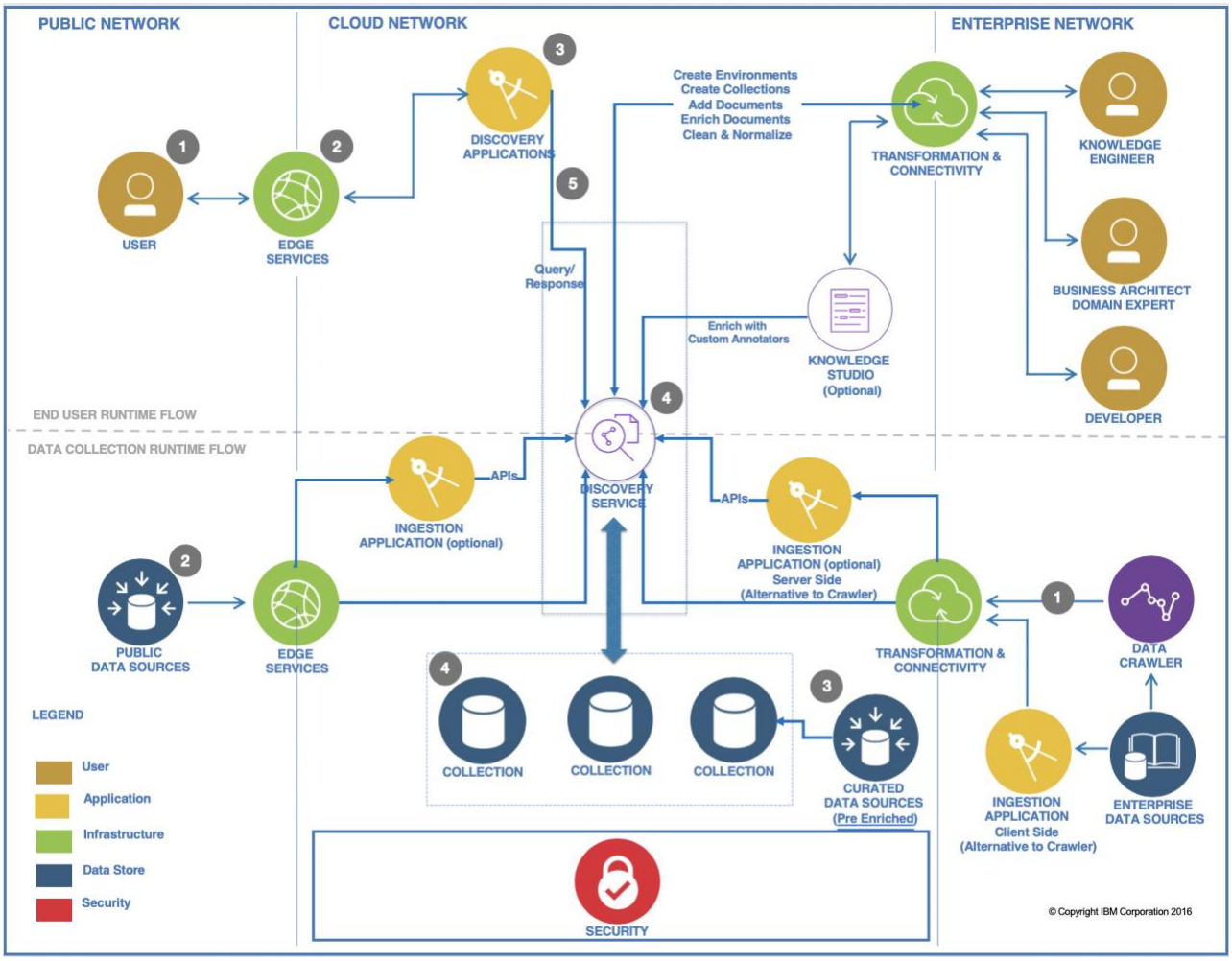


Figure 3: Data discovery reference architecture

User runtime flow

1. The user enters a search query through a web or mobile application user interface.
2. Connectivity from the enterprise network to the cloud is secured through VPN and edge services, which consists of a domain name server, a CDN server, a firewall, and load balancers. This group of services handles the request and gets it to the right destination securely.
3. A custom-developed discovery application orchestrates all of the business flow and internal API calls to Watson Discovery.

4. Watson Discovery processes the query and returns the results to the Discovery application.
5. The Discovery application visualizes the results for the user.

Data collection flow

1. In the back end, internal content is ingested from various enterprise content repositories through the Watson Discovery Service Data Crawler, Watson Discovery tooling to upload files or connect to data sources, or a custom application by using ingestion APIs.
2. External content can be ingested by using the included data source connectors in Watson Discovery or through a custom application by using ingestion APIs.
3. Pre-enriched content, such as news, is available for querying alongside any internal and external content. Watson Discovery News content is available only on IBM Public Cloud.
4. All content is stored and enriched with AI data within collections.

Content collection: Data understanding, preparation, and ingestion

The following diagram shows the top-level and fundamental tasks to deploy Watson Discovery. Step 3 is iterative. The primary roles must become familiar with the data, identify data quality challenges, and verify that the data is aligned to the business use case. Don’t overlook the dependencies and relationships of data. Learn and understand why and how the data will be used after it is ingested into Watson Discovery.

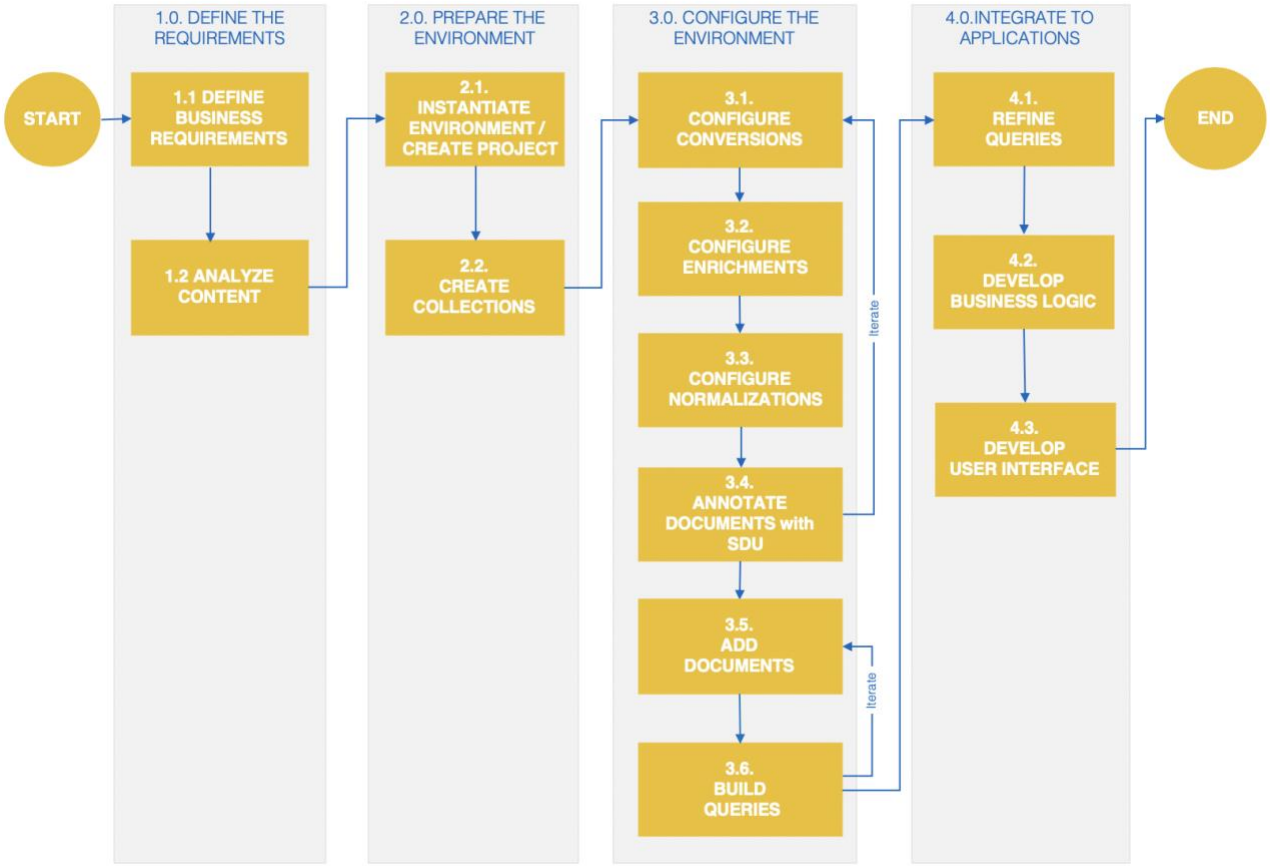


Figure 4: Discovery design time flow

As mentioned, creating a discovery system involves four phases:

- Phase 1: Data understanding
- Phase 2: Preparation
- Phase 3: Ingestion
- Phase 4: Consumption

Phase 1: Data understanding

Figure 3 shows the runtime and design time flows, and the sequence of steps and tasks that are needed to set up Watson Discovery. The architecture refers to three personas that are needed for this phase: the knowledge engineer, domain expert, and developer. These persona names refer to the skills and experience that are necessary to train and support the creation of the models or rules. Job titles or roles that might fulfill that need are data scientist, data analyst, business architect, or business analyst. It might also be someone from a line of business or in a functional role, such as a call center manager or a physician, nurse, lawyer, or other professional. You might want to work with a data architect and someone from the

data governance team in your organization. The activities of each persona are described in detail so that you can map the description to the role in your organization with the right skills and knowledge.

The governance of training content is important. It is accomplished through straightforward methods throughout the life of your project:

- Document the source of subject matter information.
- Record dates when content is created, used, and updated.
- Record the date that the content was used in training.

Apply DevOps practices to all the components in your solution. As you update configuration details, export the configurations to a source control repository with commit logs that explain the configuration changes. Don't make any configuration changes without logging them.

Make sure that you can use all the content in your discovery solution. Verify that the terms and conditions of your data aren't being violated in your solution. Also, make sure that you're taking appropriate safeguards with confidential client information (CCI) or personally identifiable information (PII). Check that your solution conforms to relevant legal standards, such as the Health Insurance Portability and Accountability Act (HIPAA) for healthcare solutions.

Use subset of content

If you have a large subset of information and content, you might be tempted to use it all and create a domain-specific model by using Watson Knowledge Studio or the custom annotator. However, when you create the training model, it's better to select a small subset of documents that have a representative set of vocabulary, information, and concepts to create the content enrichment models. Doing so reduces time, increases speed, and helps you focus on getting at least 80% of the knowledge. It's important to know the content sources and select the small subset for the creation of the custom model. After the training model is created, a larger subset of the documents can be ingested to improve search results and insights.

Continuous improvement

Start small and continue to improve the relevance and the confidence level of the search results. You can continuously improve the Watson Discovery by adding content sources and by updating the already ingested content with updated information. Governance is extremely important to ensure transparency, accuracy, the quality of the content sources that are used for training and the expert guidance that is used for continuous improvement. Historical information is often a good indicator of what contemporary usage might be like.

Analyze usage logs and ratings as a source of information to continuously improve your AI decision solution. Don't rush to ingest thousands or millions of documents if you don't understand how to properly ingest a single document.

Define the success criteria for your use case up front and determine metrics to support it. In a typical discovery scenario, success might be defined as seeing a relevant result in the top 5 results X% of the time. Get buy-in from the technical and business teams on the metrics, and continuously evaluate them. If possible, gather metrics on every code, process, or ground truth change. Without continuous metrics evaluation, you can't measure whether your system is improving or regressing, and you can't understand when you will be "good enough" or "done".

Test automation

Even more important than continuously verifying that your system maintains its performance is ensuring that the system actually functions. Plan on a test automation suite to verify that all components in the system work together. This process includes verifying your pre-processing code, ingestion code, and your application code. The test suite uses a dedicated subset of your corpus and is ideally executable within an hour. The test suite must verify that your latest pre-processing and ingestion code and your configuration are compatible with the downstream queries and integrations that you want to complete on the data.

Plan to automate your entire content ingestion process, especially if your solution has many moving parts. If you have automated the content ingestion, automating the testing is simplified.

Planning for an AI solution

Figure 3 also shows the personas that are typically involved in the planning phase.

- **Business architect or domain expert**: This person knows the source of information, also known as the ground truth. Ground truth can be a training manual, product manual, testing manual, or external, publicly available information. The business architect defines the goals and objectives of the AI-powered search or content mining application, including the channels that the application must support (web, mobile, social, and others). They also provide knowledge and understanding of technical and business domains. Their responsibility is to provide the specific terminology, classifications, and relationships that the Discovery service needs to identify the relevant content. An example of an SME in a use case for an appliance manufacturer is the field technician. Other examples are product experts, call center supervisors, scientists, doctors, and engineers.
- **Data scientist or knowledge engineer**: The data scientist supports the

business architect to understand the right kind of information that is needed to train the discovery application. Data scientists have deep knowledge of information that can be used to extract insights.

- **Developer**: The developer writes functional and test code to automate the ingestion of documents into Watson Discovery. Developers also deliver the discovery application to interact with users and the Watson Discovery service.

This planning phase involves two broad categories of information sources: internal and external. Internal content sources include processes and data sources that are within an enterprise. Typically, they contain the data that is generated and owned by the enterprise as part of its business operations.

- Business processes: These involve enterprise level-business processes that the discovery solution might have to interact with in order to process and respond to the user's queries.
- Enterprise APIs or services: These APIs or services might need to be accessed or invoked to ingest documents and information into Watson Discovery and be made available for future user queries. Most systems of records are likely to involve an API to serve the data that they generate or control.
- Shared file repository: This information is kept in file systems that are shared between users and locations and are accessible through FTP and other mechanisms.
- Content from enterprise systems: This content includes data from various enterprise systems, including but not limited to things like catalogs, order and transaction data, and ECM repositories.

External content sources include public and third-party sources, multimedia content, social data sources, and public API sources.

- Public and third-party sources: These information sources include information sources that are available for public consumption. This set of information is neither owned by the enterprise nor generated by the enterprise as part of the business operations. This information includes both public domain data (that is, available free of cost) and data that is controlled by other parties. Examples include weather data, domain-specific catalogs that are made available by third-party vendors, and point of interest data.
- Multimedia content: This content includes audio, video, or images that are available on the internet.
- Social data sources: This is a subset of public and third-party sources, but specifically involves social media such as Twitter, Facebook, and others.
- Public API Sources: This data is accessible for public consumption but requires the invocation of an API.

The purpose of the first phase is to identify and map answer units or sections from the corpus (public and private) that must be provided as a textual response to the user of the

application that is using Watson Discovery. Another purpose is to identify the processes and the APIs that might need to be invoked to collect relevant documents to be ingested into Watson Discovery. Training the Discovery service is an iterative process.

You must understand your content sources and formats thoroughly. Watson Discovery works best on unstructured text and can also handle tables and images in documents. Consider the content formats that you have: are they Microsoft Word documents, HTML, PDF, JSON, or some other format? If they're PDF files, are they scanned or searchable files? If they are scanned PDF files, they must be of sufficient scan quality that you can use OCR (optical character recognition) to read the output. Content sources that don't have attached metadata, such as document type and date, might require you to infer this information.

If the files aren't formatted to be easily consumed by the default document conversion of Watson Discovery, you must plan extra conversion steps to test them in the data preparation stage. Watson Discovery includes capabilities to increase time-to-value by handling complex document types, but in some cases you might need to allocate time and planning to convert complex document types and formats. Try simple conversions first and then iterate.

Input documents might contain poor formatting even before they run through document conversion. You must address any poor formatting in the conversion stage. Watson NLP tools make special use of sentence and paragraph boundaries. If your converted documents have errant line breaks, they can affect the NLP results.

Phase 2: Preparation

The next step in setting up Watson Discovery to support other AI systems is to prepare the ground truth for consumption at runtime. Before the work of annotation begins, the type of system and dictionaries must be loaded.

Preparing the data requires some level of training the system. This training might be done by humans or machines and is categorized as supervised, unsupervised, and semi-supervised learning. You can find high-level guidance on which type is most appropriate for your needs in the AI glossary. Base the decision on the advice of your SMEs and on your project team's understanding of the desired end state and approach to maintaining the collections.

The preparation is likely to take place by using Watson Knowledge Studio and can include multiple annotation types: human, rule-based, and machine learning. This preparation phase also involves multiple personas, including the knowledge engineer. These personas use the guidance that is provided by the domain experts to develop the rules that enable automation of curation and collection.

Watson Discovery also includes SDU, which enables quick training to better understand the structure of the documents. Use SDU to extract custom fields in your documents. By customizing how your documents are indexed into Watson Discovery, you improve the answers that your application returns.

With SDU, you annotate fields within your documents to train custom conversion models. As you annotate, Watson Discovery learns and starts to predict annotations. You can export SDU models and use them in other collections. SDU supports several document types including PDF, Word, PowerPoint, Excel, JSON, HTML, PNG, TIFF, and others[1]. For more information, see the SDU documentation for supported document types.

Phase 3: Ingestion

You can ingest data into Watson Discovery in three ways:

- If you're integrating the upload of content with an existing application or creating your own custom upload mechanism, use the API.
- If you want to quickly upload locally accessible files, use the Watson Discovery tooling. When you upload documents by using the tooling, all documents must have a unique file name. If two files have the same name, the original is overwritten when the newer version is uploaded. If you would prefer that documents with the same file name coexist in your collection, you must specify the Document ID. You can specify the Document ID if you upload documents by using the API.
- You can use the Watson Discovery tooling or API to connect to Box, Salesforce, Microsoft SharePoint Online, IBM Cloud Object Storage, and Microsoft SharePoint 2016 data sources, or do a web crawl. For more information, see Connecting to data sources.

Whichever upload mechanism you use, you need a configuration that tells Watson Discovery how to ingest your files. The configuration steps include conversion, enrichment, and normalization. Conversion and normalization dictate the data schema that you use for your files.

Watson Discovery comes with reasonable defaults for PDF, Word, HTML, and JSON file conversion. Watson Discovery first converts PDF and Word files to HTML and then converts HTML to JSON. You can configure things such as the expected major heading size in Word documents or what fields to duplicate, merge, or drop in your JSON structure. For enrichment, you can select which fields are enriched and in what way.

[1] JSON and HTML documents are supported by Watson Discovery but cannot be edited by using the SDU editor. To change the configuration of HTML and JSON docs, you need to use the API. For more information, see the API reference.

Phase 4: Consumption

The previous three phases all support the final phase, the consumption of the content, which typically happens through an application. In this example, the content is consumed by various users through a customer web application that is called "Weather Insights".

Watson Discovery runtime architecture for Weather Insights

Earlier, the runtime flow for a typical AI solution was illustrated along with how to plan and prepare Watson Discovery to ingest and annotate content. Before you create an application, it's important to understand the following concepts:

- **Discovery**: Discovery finds the relevant passages in the corpus and answers open- ended questions. It's often used for knowledge expansion or long-tail scenarios. For training purposes, the knowledge engineer loads and annotates unstructured documents to train a ranker model to rank the returned passages for an utterance.

- **Collections:** A collection is the logical grouping of your documents within your environment. Your configuration controls how the documents are ingested and enriched into the collection. Each collection has a unique configuration pipeline, so all documents are converted, enriched, and normalized in a unique way.

- **Ground truth**: Ground truth is the information that is used to train the Discovery service. Content from public and enterprise sources, under guidance from domain SMEs, is ingested into Discovery collections that can be used to respond to user queries. Domain experts or SMEs define the ground truth by providing ranking data to teach Discovery which responses are most relevant to a query. Ground truth is also typically split into training, testing, and evaluation data.

Figure 5 illustrates the architecture for the Weather Insights application, showing the relationship between the capabilities of the data collection application and the functions that are needed for the user. You can see how different users interact with a trained AI system and how the system interacts with other components on the cloud platform. The figure also shows that data can come from public sources or the customer's enterprise, either in the form of raw data or in the form of an API.

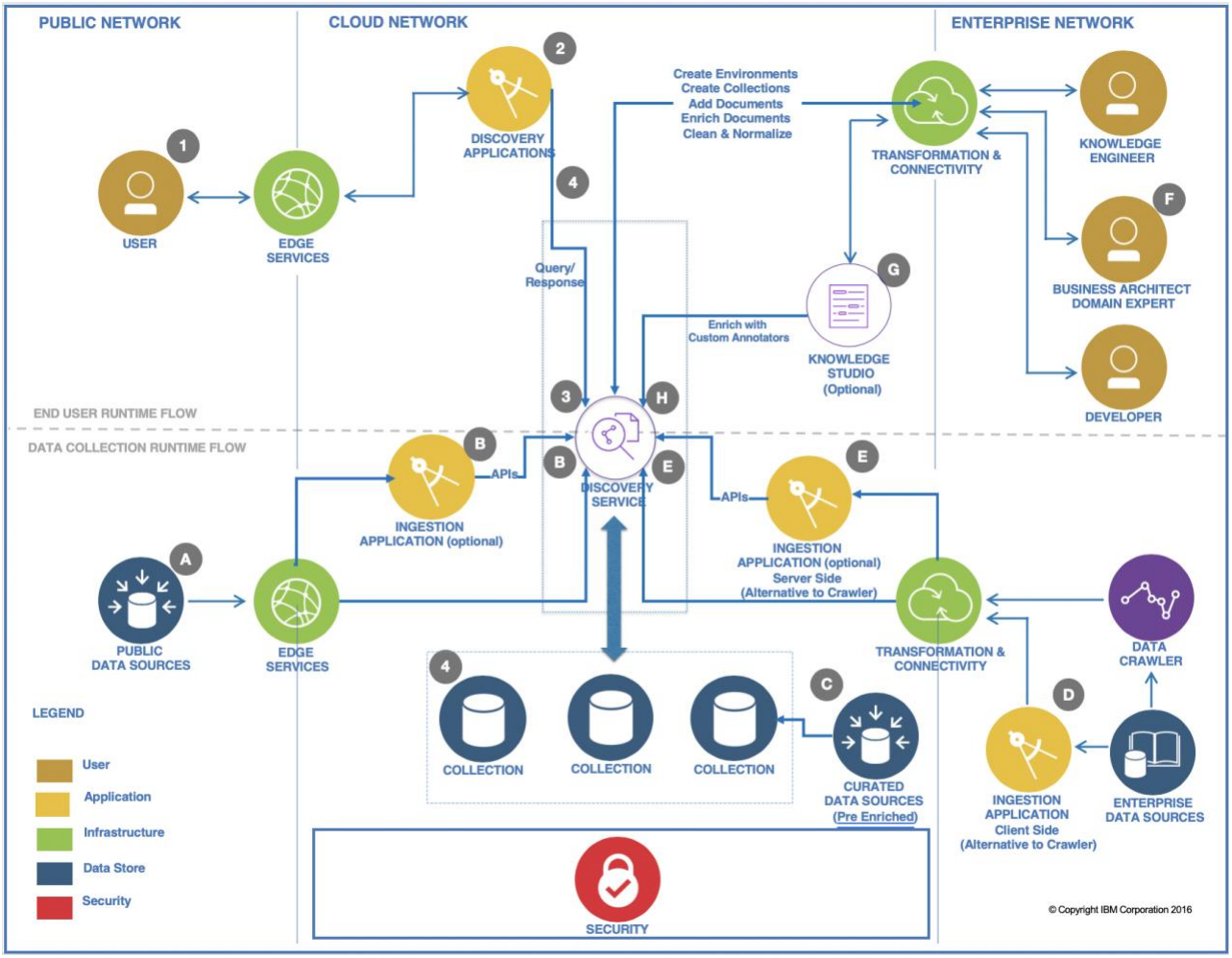


Figure 5: Watson Discovery runtime solution for Weather Insight

In Figure 5, items A – H are the activities that are needed to build the ground truth and content collection for the Weather Insights application. Items 1 – 4 describe the user experience as they interact with the application.

A home improvement store in North America wants to offer an AI decision assistant application for emergency preparedness to its retail customers, professional clientele, and in-store associates.

Residents in counties and cities that are prone to weather-related events like hurricanes rely on news organizations, TV, radio, mobile, web, and their own knowledge to prepare in case they are struck by a hurricane or are in the path of a potential hurricane. Counties deploy temporary emergency workers who respond to common questions that are related to hurricane preparedness. The time that is required to get the most current knowledge is important. Unfortunately, it's common for residents to get the wrong information about a hurricane or about preparations necessary to stay safe. Enterprises also struggle to understand the impact of the

hurricane to their business, such as supply chain disruption. The Discovery service can be a decision assistant that can be trained by ingesting a corpus of knowledge that is related to hurricanes. This knowledge might involve guidance to prepare for hurricanes, such as the procedure to install hurricane shutters, historical supply chain disruption information, and operational guidance for emergency workers.

Data collection flow

Items A – H describe the activities to build the ground truth and content collection for Weather Insights:

A. Public data sources include data sources that are available for public consumption with or without any fee or subscription. These sources include all data that is owned by the third parties. It resides outside the enterprise or the agency that consumes the data. In this example, it includes documents that are published by FEMA (Federal Emergency Management Agency) and other governmental agencies, publicly available information such as blog posts, and information that is published by individuals and various NGOs (Non-government organizations). The subject matter of these documents includes things like weather information, instructions that are related to hurricane preparation, tips, instructions for putting up shutters, various kinds of checklists, and more.
B. The ingestion application or the Watson Discovery tooling uses the identified information in step A to crawl data into a Watson Discovery collection.
C. Pre-enriched curated data sources, such as Discovery News, are available by default in Watson Discovery.
D. Emergency employee response, supply chain disruption, and other documents from private sources are identified.
E. The ingestion application or the discovery tooling uses the identified information in step D to crawl data into a Watson Discovery collection.
F. The business domain expert logs in to the Watson Knowledge Studio and uploads sample representative documents like the ones that were obtained in steps B and D. The domain expert might need to train different domain-specific models for the documents from different sources.
G. The business domain expert (or experts) annotates entities and relations in Watson Knowledge Studio to create ground truth that is specific to this application domain. A domain-specific custom model is then trained in Watson Knowledge Studio by using this ground truth to extract entities and relations that are specific to this application.
H. The custom model is then deployed in the Discovery service.

User interaction flow

Items 1 – 4 describe the user experience:

1. The user selects the option from the drop-down list. A list of default questions is made available. The user selects a question from the list and sends the request to the Discovery application. Alternatively, the user types the question in a free-form text input field.
2. The Discovery application sends the question to the trained Discovery service.
3. The trained Discovery service returns the top N responses of the query. Typically, it returns the top 3 or 5 responses.
4. The Discovery application relays the returned responses to the user.

Best practice approaches for your Discovery services project

The following practices are important for the successful of a Discovery application.

Do your prework

Allow ample time for advance planning. Prework includes the identification of users, queries, and knowledge sources.

1. Users: Who will use the trained Discovery service? When you answer this question thoroughly, you will understand who needs to train the Discovery service.
2. Queries: What are typical queries that you anticipate your users might often submit? Assessing the kinds of queries and searches that your users make can help to train the system for realistic scenarios, resulting in higher accuracy in the results.
3. Corpus of knowledge: Identifying the knowledge sources is important to get the application to work correctly. This process includes identifying documents with potential answers and selecting content sources to use to train the Discovery service.

The final required prework step is to analyze the document content to assess how answers must be extracted.

Understand your content sources

While you can use many publicly available sources of information to train the Discovery service for better results, the most relevant content belongs to the customer. Their understanding of the business domain, operations, and other expert knowledge gained over time is critical to getting the content right. This knowledge also includes the

customer's knowledge of their own database and the way that the data is expressed, such as aliases, abbreviations, acronyms, and operation-specific codes.

For example, a call center manager might know where the historic user conversations are stored and know the definitions for various domain-specific terminology that is used in the conversation. A medical doctor might know which patient cases are better for training. A training or curriculum specialist might know the right sources for education. An insurance agent might know policy details best.

The most important question is whether the content sources that are selected for ingestion and training contain answers to the questions that the user is likely to ask.

Govern the content used for training the discovery service

Governing the content sources that are used for training means that a process is in place to ensure the quality and consistency of documents and other unstructured content. Governance also implies that the training content is versioned and managed, and that trainers have a way to keep their own domain knowledge up to date.

This kind of governance supports the results of initial data pre-work, increases the probability of gaining the best insights, and helps to ensure a continuous improvement in the result set over time. Establishing an AI center of excellence as part of your data and information governance program also ensures that the right set of resources, information, and people are used to train the Discovery service for later projects. A center of excellence can provide guidelines for creating content so that it is most usable by Watson Discovery or other AI services.

You might need to update your machine-learning models due to common business scenarios such as new product launches, an acquisition or a merger, or a product end of life. Do more than anticipate the need to update. Build on the work that you started in your planning and pre-work by making a change management process part of your prework, preparation, and training stages. By creating a plan and implementing it before you need to make changes to your production service, you can simplify the effort to maintain content collections and enrichment models. Don't worry if your process isn't perfect at first. You can modify your initial approach as part of the planning and training iterations.

Include time for adequate iteration and create a sustainability plan

Your solution gets better with usage and continuous training. As existing experts and new experts add more relevant data sources and the knowledge of how to use the

Discovery service improves, the information and insights become more accurate and more useful. Continuous training ensures better decision support for the enterprise.

This usage can drive insights that might lead to the decision to modify ingestion and prepare new data. Continuous usage and training on small sets of data can enable rapid iteration cycles.

Iteration encourages quick wins

Support for customer service staff or customer self-service are common scenarios for building an AI application. Rather than trying to curate and enrich all your proprietary or internal information at once, consider an incremental approach to introducing a solution into your operations environment. For example, in industries such as life science or medical devices where SMEs are in short supply or have high salaries, consider automating the collection and curation of content through machine rules. The SMEs can spend some time training and then move on to higher-value tasks.

Executive sponsorship and business support

AI solutions are often business- or operations-centric. IT, functional groups, and lines of business need to collaborate for both immediate success and for sustainable results. Creating an AI center of excellence for these groups to engage in can help to expand the adoption of AI solutions and provide a framework for sharing lessons learned. Because this will transform the decision-making in an enterprise, executive sponsorship ensures success. Transformation requires executive support for cultural and process changes. It also requires support during the initial bottlenecks, pushbacks, and failures to ensure future successes. The initial projects are business experiments, so executive sponsorship ensures support from the organization. Also, technology decisions like movement of data and current practices like single tenancy and multiple tenancy might need to be changed.

Security architecture: Content and data collection

Security is a critical aspect of the AI reference architecture. Security in the AI reference architecture addresses the following areas:

- Data or content at rest
- Data or content in motion
- Identity of the user
- Authorized access to every task that is part of the Discovery service
- Monitoring events and applying AI capabilities for securing and removing

threats before they happen

Security can be divided into two parts:

1. Security for training the Discovery service
2. Security for usage of the trained Discovery service

Security for training the Discovery service

You must ensure and enforce that only authenticated users who have access privileges to do certain tasks are allowed to do those tasks. The policies for enforcing access control must be maintained in the policy administration system.

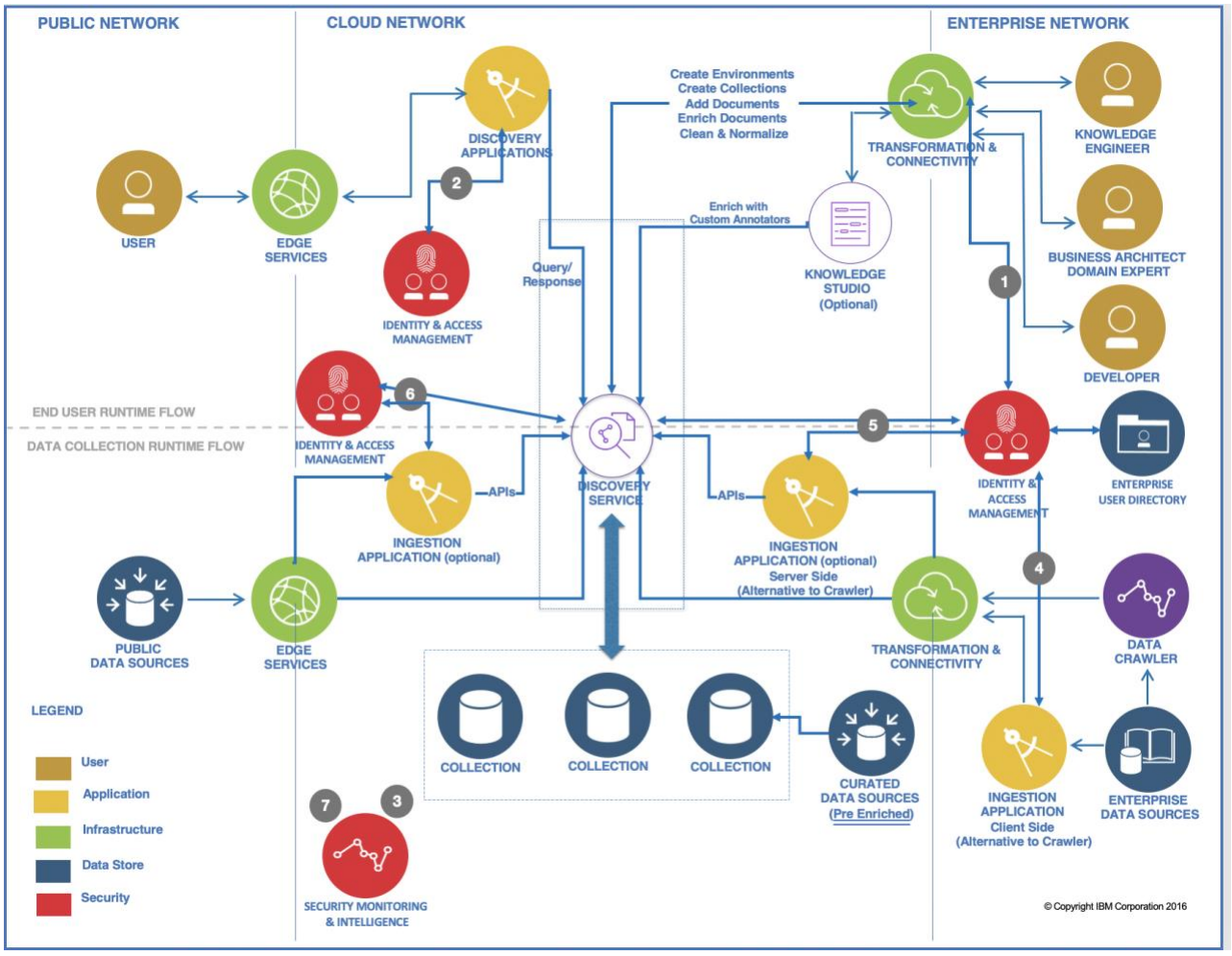


Figure 6: Runtime flow and security access

- In step 1, training the Discovery service for a specific industry domain or business function area requires industry knowledge. You can accomplish this by using Watson Knowledge Studio. The tasks of creating and managing the training

models in Watson Knowledge Studio must be controlled through security access policies. In the user runtime, the identity and access management component validates the access rights and credentials of the business architect or domain expert, the developer, and the knowledge engineer. Only authorized users are allowed to access Watson Knowledge Studio to enrich the ingested data with custom annotators.

- In step 2, the identity and access management component validates the access rights and credentials of the end user who is authorized to access the Discovery application to send query requests to the Watson Discovery service.
- Step 3 shows how the security monitoring and intelligence component continually monitors user access information for advanced threats and to validate that only authorized users have access to the Discovery application and to Watson Knowledge Studio to train custom domain-specific AI models.
- Content such as product manuals, historical call center records, training manuals, legal documents, and insurance policy documents are stored in a content repository in the customer data center. This content must be ingested in the Discovery service for annotation and enrichment. In the data collection runtime, the identity and access management component validates the access rights and credentials of the ingestion applications (step 4 and 5 for content that is in the enterprise data center; step 6 for content that is in the public domain). Only authorized applications and users are allowed to access the enterprise data for ingestion. Similarly, only authorized users should be able to configure Discovery tooling to crawl content in the customer data center to ingest into a Discovery collection.
- The organization should define encryption policies for encrypting the content that is used to train the Discovery service. The policy must address securing the content at rest and in motion. Watson Discovery encrypts the ingested content at rest and in motion.
- The security monitoring and intelligence component continues to monitor user access information, valid encryption of all data assets, and all the components for advanced threats as shown in Step 7.

Security for usage of the trained discovery service

- After the Discovery system is trained, the business can decide the definition and enforcement of the policies of authentication and access. For example, in certain scenarios such as searching product information, all users can have access to general product information, but only privileged users have access to confidential research materials. As shown in step 2 of the diagram, the identity and access management component validates the access rights and credentials

of the user. Only authorized applications and users are allowed to access the Discovery application.

- Watson Discovery provides a capability to encrypt content that is part of its data collection. The data collection of any trained Discovery service can be encrypted.
- The security monitoring and intelligence component, which is referenced in step 3, continues to monitor assets and information for advanced threats.

Components

Examine the individual components that make up the AI architecture:

Public network components

The public network contains elements that exist in the internet: data sources and APIs, users, and the edge services that are needed to access the provider cloud or enterprise network. The public network also includes the conversation endpoints.

User

The user is a customer who uses his device to access the conversation system on the cloud provider platform or enterprise network.

Device

A user uses a mobile device or other form factor that has an application with an embedded chatbot to start a conversation with the AI system.

Cloud network components

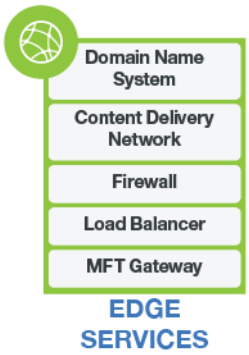
Several components make up the cloud network.

Edge services

Edge services are distinct network components that are a part of the IBM Cloud platform. These services allow data to flow safely from the Internet into the IBM provider cloud and into the enterprise. Edge services also support user applications.

This domain includes the following key capabilities:

- Domain name system server: Resolves the URL for a web resource to the IP address of the system or service that can deliver that resource.
- Content delivery networks (CDN): Supports user applications by providing geographically distributed



systems of servers that are deployed to minimize the response time for serving resources to geographically distributed users. This process ensures that content is highly available and is provided to users with minimum latency. Which servers are engaged depends on server proximity to the user and where the content is stored or cached.

- Firewall: Controls communication access to or from a system, permitting only traffic that meets a set of policies to proceed and blocking any traffic that doesn't meet the policies. You can implement firewalls as separate dedicated hardware, as a component in other networking hardware such as a load balancer or router, or as integral software to an operating system.
- **Load balancers:** Provide distribution of network or application traffic across many resources, such as computers, processors, storage, or network links, to maximize throughput, minimize response time, increase capacity, and increase the reliability of applications. Load balancers can balance loads locally and globally. Load balancers must be highly available without a single point of failure. Load balancers are sometimes integrated as part of the provider cloud analytical system components such as stream processing, data integration, and repositories.
- **Managed file transfer (MFT) gateway:** This gateway is a multi-protocol gateway (AS2, AS4, sftp, ftps, C:D) into and out of the organization. It provides security (encryption and decryption), virus checks, data loss prevention, certificate and key management, monitoring, and auditing.

IBM Cloud platform supports various services for DNS, firewalls, load balancing, and CDN. IBM Security Network Protection (IBM XGS) is a next-generation intrusion prevention system (IPS) that can be used to monitor network traffic and to provide protection from hidden security vulnerabilities. IBM DataPower® provides load balancing and Secure Sockets Layer (SSL) termination. It can help to quickly secure, integrate, control, and optimize access to a range of workloads through a single, extensible, DMZ-ready gateway (a firewall configuration for securing local area networks).

Watson Discovery

Watson Discovery helps users find the most relevant information for their query by using a combination of search and machine-learning algorithms to detect signals in the data. The service can be combined with other natural language services to create custom applications. It can also be integrated with other applications to enhance functionality.

Among the advanced AI functions that come ready for immediate use are natural language queries, passage retrieval, continuous relevancy training, relationship graphs, anomaly detection, SDU, and content mining. All can assist an organization to identify their specific knowledge and data assets and achieve results more efficiently.

By default, IBM Cloud offerings don't share any log information.

Content storage

Watson Discovery includes storage that is provided through collections.

Application logic

Application logic, which might be a Node.js application, first passes the natural language utterance (request) to the Discovery service. When it receives
the response from the Discovery service, the application returns top N responses to the user. The application includes logic to decide how many responses to return and how best to display the returned results depending on the user's application interface.

Transformation and connectivity

Application logic can strengthen the response by supplementing structured data, such as user profile, past orders, and policy information, from the enterprise network. The connection to the enterprise network
is established through the transformation and connectivity component.

In IBM Cloud, you can use the IBM Integration Bus container allows you to integrate applications and infrastructures that are deployed in multiple clouds or in legacy or core applications that are deployed in customers' traditional data centers.

IBM API Connect® is a comprehensive API lifecycle solution that enables the automated creation of APIs, simple discovery of systems of records, self-service access for internal and third-party developers, and built-in security and governance. By using automated, model-driven tools, you can create APIs and microservices that are based on Node.js and Java® runtimes, all managed from a single unified console. Ensure secure and controlled access to the APIs by using a rich set of enforced policies. Drive innovation and engage with the developer community through the self-service developer portal. IBM API Connect provides streamlined control across the API lifecycle and enables businesses to gain deep insights around API consumption from its built-in

analytics.

IBM Secure Gateway for IBM Cloud brings hybrid integration capabilities to your IBM Cloud environment. It provides secure connectivity from IBM Cloud to other applications and data sources that run on premises or in other clouds. A remote client is provided to enable secure connectivity.

Enterprise network components

Several components make up the enterprise network.

Ground truth

Ground truth is the training data for Watson Discovery and related services APIs. It includes various artifacts from public and enterprise sources.

Ground truth is typically split into training, test, and evaluation data. The ground truth for relevancy training is in the form of sample questions, answers, and relevance labels. The questions are referred to as *natural language queries* and the answers are referred to as *answer units*. For each natural language query, you need multiple answer units that are rated with a different relevance score. You upload training data into Watson Discovery for your specific collection and Watson Discovery automatically handles the training of the machine learning model. For more information, see the training data requirements.

IBM capabilities for security for content discovery and management

The following table maps the IBM capabilities and services mapped to the components in Figure 6.

Component	Definition	IBM and other products
Edge services	Edge services include services that are needed to allow data to flow safely from the internet.	DNS, CDN, firewall, load balancer
Transformation and connectivity	This component includes scalable messaging and transformation and secure connectivity.	IBM Integration Bus container, IBM DataPower, IBM API Connect, IBM Secure Gateway

Key management service	The key management service is a cloud-based security service that provides key lifecycle management, including key creation, usage, and deletion, for encryption keys that are used in IBM Cloud services or in customer-built applications, with "root of trust" backed by a hardware security module (HSM).	IBM Key Protect
File encryption service	The file encryption service safeguards data even when network protection fails. It has built-in key management to handle the storage of all the encryption and splitting keys.	IBM Multi-Cloud Data Encryption
Secured connectivity	These services offer security connectivity such as Virtual Private Network (VPN) or TLS-based encryption that ensures the secure transmission of data from enterprise to cloud or vice versa. Social media providers use TLS-based security to perform a single sign-on	VPN providers
Identity and access management	This component identifies and authenticates the user. It determines access levels by using an enterprise security directory such as LDAP.	IBM Security Access Manager
Security monitoring and intelligence	This component provides security and visibility into cloud infrastructures, data, and applications by collecting and analyzing logs in real time across the various components and services in the cloud. It provides risk analysis of the workloads that are hosted in the cloud against the myriad of known vulnerabilities and alerts against zero-day	IBM QRadar® SIEM

Infrastructure security	This component protects against network-level threats and attacks with intrusion prevention and detection, including attacks that tunnel through encrypted web transactions and web applications that are deployed within the system.	IBM Security Server Protection, IBM Security SiteProtector System

Watson Discovery: Planning for success

You can create any number of collections in your Discovery service environment. The limit is the storage size for the environment and the number of documents that are allowed by the subscription. Your index size can't exceed your storage. The size of your index varies based on document size, the number of fields that you index, and how many enrichments your data contains. Your index can be twice or more the size of your input document text. To estimate your needs, you should measure the size of a subset of your corpus and extrapolate that size based on number of documents.

When you decide on the number of collections that you need, the determining factors are your data type, format, and structure, and whether you want logically separate development, test, staging, and production environments. When data sets aren't consistent, you likely need multiple collections. You might need to modify some types of files before ingestion. By using multiple collections, you can test the configuration pipeline without affecting production activities.

The right cloud platform

To create an efficient AI environment, it's important for your enterprise-grade cloud platform to be built on a data-first architecture that gives you the choice between using a public, private, or proprietary hybrid architecture. The cloud platform must be user friendly and created with scalability and resiliency in mind.

The IBM Cloud gives enterprises the following benefits:

- Control over where the customers' data resides
- A level of security that enables the secured movement of content from other cloud providers and customers' existing data centers into the IBM Cloud
- Capabilities to encrypt and store data securely
- Capabilities for secure access of information and systems

IBM Cloud is an industrialized cloud, which enables integration between data and applications and also between public, private, and proprietary clouds. IBM Cloud is an industry-centric cloud, offering capabilities that are designed for industry-specific data or content and regulations. As a result, you have a broad variety in the information that is available to Watson Discovery for use in creating collections. IBM Cloud provides more than 120 services and includes Watson APIs, services, and software that can help you enable your business.

Robust ecosystem

Your business' ecosystem plays a major role in the successful transition to an AI

business. Watson services can be strengthened by content that is captured in the peer cloud of companies that have instrumented the physical world to create a robust ecosystem for solving business challenges.

These examples show how Watson services can be used with ecosystem partners to gain insights into data:

- Content that is captured by car manufacturers can be used with Watson services to provide car-related information to drivers and for self-driving cars.
- Content that is captured by health-monitoring devices such as blood sugar monitoring devices can be used to give doctors recommendations about changes in medication or to remind patients to take medication at the right time.

Deployment considerations

As stated earlier, a critical success factor for creating successful Discovery solutions is a secure, user-friendly cloud platform. The cloud platform provides capabilities for actionable insights. IBM Cloud includes AI services across cloud deployment options.

When you deploy AI systems, isolation, privacy, region and language support, and performance and scalability are important considerations.

Isolation

This consideration involves deploying to customers who carry client-confidential or sensitive private information. In those cases, you might want to use a premium or a dedicated deployment option to support your application or Discovery service, whether it is a multi-tenant or single-tenant implementation. Where you have few or no confidentiality concerns, it might be acceptable to choose a multi-tenant model that is provisioned by using a standard or public deployment.

Privacy

As a general rule, don’t store or pass any confidential information or protected health information (PHI) when you interact with an AI system. However, where data is encrypted end-to-end, this rule doesn’t apply. Watson Discovery in Dedicated and Premium deployments offers encryption at rest and is suitable for PII.

Watson Discovery is available as a premium subscription, which offers developers and organizations a single tenant instance of one or more Watson services for better isolation and security. These plans offer compute-level isolation on the shared platform and end-to-end encrypted data while in transit and at rest.

Region and language support

When you deploy applications that involve multiple geographies and languages, the services might need to be deployed in multiple regions byusing the IBM Cloud region settings.
Additionally, AI systems must be designed and trained against various languages based on the support that the service provides. It is the responsibility of the application or the solution to pass the language parameters to the APIs during runtime.

Performance and scalability

To support a large volume of users, create a testing plan that involves load testing. You can use open source frameworks such as JMeter or third-party services such as Blazemeter in IBM Cloud to create and run load tests. The load test must include submitting various request sizes and concurrent users. Depending on your performance needs, you might need to scale the service instances in IBM Cloud. IBM Cloud offers the capability to scale the services both horizontally and vertically. You can also employ capabilities such as auto-scaling to configure the scaling based on demand, throughput, and memory usage.

References

- ISO/IEC 17788:2014 Information technology -- Cloud computing -- Overview and vocabulary
- ISO/IEC 17789:2014 Information technology -- Cloud computing -- Reference architecture
- IBM Watson Assistant
- IBM Watson Speech to Text
- IBM Watson Discovery portfolio
    - IBM Watson Discovery
    - IBM Watson Knowledge Studio
    - IBM Watson Natural Language Understanding
- IBM Cloud catalog
- AI glossary