# The IBM Advantage for Cognitive Discovery Cloud Architecture

## Table of Contents

- Executive overview ................................................................. 2
- IBM Cloud Customer Architecture for Cognitive ......................... 5
  - What is Watson Discovery? .................................................... 7
    - Data and content collections .............................................. 8
- IBM Cognitive Discovery reference architecture ............................ 9
- Content collection: Data understanding, preparation, and ingestion .... 10
  - Phase 1: Data understanding ............................................... 11
  - Phase 2: Preparation .......................................................... 16
  - Phase 3: Ingestion .............................................................. 17
  - Phase 4: Consumption ......................................................... 17
    - Watson Discovery runtime architecture for Weather Insights .......... 17
- Best practice approaches for your discovery services project ............ 21
  - Do your pre-work ............................................................. 21
  - Understand your content sources ......................................... 21
  - Govern the content used for training the discovery service ............ 22
  - Include time for adequate iteration and create a sustainability plan .... 22
  - Iteration encourages quick wins ......................................... 23
  - Executive sponsorship and business support ............................ 23
- Security architecture: Content and data collection ......................... 23
  - Security for training the discovery service ............................. 24
  - Security for usage of the trained discovery service .................... 25
- Components ............................................................................ 25
  - Public network components .................................................. 26
    - User .................................................................................. 26
    - Device .............................................................................. 26
  - Cloud network components .................................................... 26
    - Edge services ...................................................................... 26
    - Watson Discovery ............................................................ 27
Data has become the fuel of business innovation in proportion to the increase in the amount of data available. Sensors, video, news and social media streams, and weather data are only a few of the sources of data available to an enterprise, in addition to their private stores. The organization that is able to tap these sources, separate out the valuable information from the noise, see relationships and patterns in the data, and then act upon this knowledge is best prepared to overtake their 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 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, visual recognition, and other elements of artificial intelligence can help in identifying and organizing unstructured data. This is where cognitive computing comes in. IBM’s cognitive 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 trained by a legal expert to sort through thousands of files of unstructured data to identify those pertinent to legal claims can do it faster than a person, freeing up the expert for higher value activities.

Cognitive 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 natural language processing, 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 cognitive decision making into existing processes and into the creation of new processes.

The tools 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 cognitive search through natural language processing 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 and structured data. This can simplify human use of information through more efficient access to large content stores or through the integration of the service in support of larger cognitive systems.

Prior to the availability of natural language processing and contextual search applications, keyword searches were the way users engaged with masses of information. Previous approaches to the enterprise management of information, launched under the banner of knowledge management, relied on the creation of complex content topologies, huge internal indices, and the speed of the keyword search. These projects were not adopted widely due to the level of effort required for basic results. The table below shows how value to business increases with the adoption of more sophisticated techniques for search and analysis.
This paper describes how to use Watson Discovery on the IBM Cloud platform to create content collections and custom cognitive applications that follow IBM’s best practices approach for curating content and designing cognitive applications. The relationships between business processes and technical architecture components that use cloud computing infrastructure, platforms, and services are described, along with the kinds of personnel and activities required to prepare and implement a Watson Discovery solution that is capable of evolving with the organization.

If you are new to cognitive systems, you can familiarize yourself with fundamental cognitive concepts such as ground truth, training set, and test set by reading the cognitive glossary.
IBM Cloud Customer Architecture for Cognitive

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

1. **Discovery**: IBM’s cognitive discovery capabilities ingest and enrich information, annotate the information stored in multiple documents, and prepare 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 Cognitive Discovery Reference Architecture.
2. **Conversation:** IBM’s cognitive conversation capabilities are trained to assist in decision-making 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 driven by dialog that can be used to determine the best course of action. See the [Cognitive Conversation Reference Architecture](#) to understand how these capabilities are realized.

3. **Extend:** IBM’s cognitive conversation and discovery capabilities can be extended by cognitive services that take broad or unstructured data and create meaningful, actionable, and valuable information for users (which can be domain-specific). Using a variety of services or offerings such as the Watson speech to text, text to speech, tone analyzer, visual recognition, natural language classifier, and 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.

The cognitive reference architecture presents conversation, discovery, and extend capabilities in relation to each other. In this paper, we discuss the details of the planning, preparation, and execution specific to Watson Discovery. Watson Discovery is designed to extract value from unstructured data by converting, normalizing, and enriching it. This data can be proprietary, public, or third-party. Users can apply various AI-powered information retrieval techniques to identify the best responses to questions on discovery collections. You can use the resulting output of the Discovery service in combination 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 cognitive 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 needed to keep it always up-to-date. Because Watson Discovery is part of the IBM Cloud platform, it can be used in support of a cognitive system designed for any industry.

There are four phases to enable Watson Discovery:

- Phase 1: Data understanding
- Phase 2: Preparation
- Phase 3: Ingestion
- Phase 4: Consumption
Phase 4 is illustrated by a business scenario and sample application. This is followed by a summation of the best practices and an itemization of components necessary for a successful deployment.

**What is Watson Discovery?**

Watson Discovery is an insight engine that provides an end-to-end pipeline for ingesting, storing, and enriching data, allowing you to query both cognitively enriched metadata and content using APIs. Data that is enriched and curated is referred to as a *collection*.

One use of Watson Discovery is to create cognitive applications that can help users find the specific information they need quickly. You can use the relevancy training and passage retrieval capabilities of Discovery to do this more efficiently. Relevancy training can scale search relevancy by using domain expertise to train Discovery on the best ranking of results. It allows developers and subject matter experts to work together to teach the system to find signals in the way documents and queries are related and to bring the most relevant documents to the top of results. Developers can also use the relevancy model to further customize responses using a machine learning model.

Passage retrieval lets you find information within documents that is relevant to your query. It can dynamically select snippets from within larger documents based on an input query to present users with easy-to-display and consume results.

Both of these capabilities allow developers to improve the user experience by displaying answers to queries without any additional configuration or special processing.

Watson Discovery goes beyond just search by leveraging an automated knowledge graph that is created at scale on data that is ingested into collections. Knowledge graphs go beyond just data retrieval by making connections within your data across documents, extracting and disambiguating entities and relationships, and using sophisticated relevancy ranking techniques to return the best results.

Watson Discovery leverages IBM Watson Natural Language Understanding, a collection of APIs on the Watson platform that offers advanced text analysis through natural language processing, machine learning, and deep learning. This set of APIs can analyze text to extract concepts, entities, keywords, sentiment, relation, emotion, and taxonomy.

Watson Discovery can also integrate with IBM Watson Knowledge Studio, which provides an integrated development environment to create and deploy custom annotators infused with specialized domain knowledge. Knowledge Studio can be used by subject matter experts who do not have machine learning or data science expertise. Annotators can be used to drive adaptive learning scenarios.
Trained IBM cognitive services or systems can be used in any form factor (mobile, kiosk, car dashboard, web, voice response unit, or others) for decision assistance. The content collections created with Discovery are meant to support any of these 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 existing data, such as well-structured enterprise data and any real-time data that could come from sensors and other wearable technology. Watson Discovery supports collection and curation of content from all of these sources.

Organizations are not limited to static content collections.

Cognitive applications interact with people in a natural way to answer questions and provide guidance to help people make decisions. When you develop a cognitive application for your industry, you want to develop domain expertise from your subject matter experts, incorporating best industry practices and relevant material from relevant data sources.

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

Watson Discovery News, a public data set that has been pre-enriched with cognitive insights, 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 the following: **keyword extraction, entity extraction, concept tagging, relation extraction, sentiment analysis**, and **taxonomy classification**. The following metadata is also added: 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, IBM Cloud offers access to information like weather data and Twitter streams, and collects content from third-party content providers and industry experts to assist in better decision-making.
IBM Cognitive Discovery reference architecture

The reference architecture for cognitive discovery shows the two distinct flows—user and data collection—that comprise a typical use of Watson Discovery.

![Diagram of Cognitive Discovery reference architecture](image)

**Figure 3: Cognitive Discovery reference architecture**

**User runtime flow**

1. User enters a search query through the web application using business concepts of the domain.
2. Connectivity from the enterprise network to the cloud is secured through VPN and edge services, which consists of domain name server, CDN server, 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 the discovery service.
4. The discovery service 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, or a custom application using ingestion APIs.
2. External content can be ingested through a custom application using ingestion APIs.
3. Pre-enriched content, such as news, will be available for querying alongside any internal and external content.
4. All content is stored and enriched with cognitive data within collections

Content collection: Data understanding, preparation, and ingestion

The diagram below shows the top level and fundamental tasks necessary to a successful deployment of Watson Discovery. Step two is iterative. It is essential that the primary roles become familiar with the data, identify data quality challenges, and validate that the data is aligned to the business use case. Awareness of dependencies and relationships of data cannot be overlooked. Take time to learn and understand why and how the data will be used once it is turned into corpus. You can review the complete Cognitive Discovery design time flow for a deeper understanding of the phases and activities.
As mentioned above, there are four phases in creating a cognitive discovery system. In this section, we will show you reference architectures that relate to all phases. As a reminder, the phases are:

- 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 needed to set up Watson Discovery. The architecture refers to three personas needed for this phase: the knowledge engineer, developer, and domain expert. These persona names refer to the skills and experience necessary to train and support creation of the models or rules. Job titles or roles that might fulfill the need include 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 also want to have a data architect and someone from the data governance team in your organization working with you. The activities of each persona are described in detail so you can map the description to the role at your organization with the right skills and knowledge.

Governance of training content is a best practice and 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 components in your solution. As you update configuration details, export these configurations to a source control repository with commit logs explaining the configuration changes. Do not make any configuration changes without logging them.

Additionally, make sure that you are able to use all content in your discovery solution. Verify that the terms and conditions of your data are not being violated in your solution. Make sure you are taking appropriate safeguards with confidential client information (CCI) or personally identifiable information (PII). Check that your solution conforms to relevant legal standards, such as 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 begin creating a custom training model using Watson Knowledge Studio or the customer annotator. However, a best practice is that when creating the training model, you should select two to six documents that have a representative set of vocabulary, information, and concepts for creation of the content enrichment models. This reduces time, increases speed, and helps you focus on getting at least 80% of the knowledge. Knowing the content sources and selecting the small subset for the creation of the custom model is important. Once 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 discovery service by adding new 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
used for training, and the expert guidance used for continuous improvement. Historical information is often a very good indicator of what contemporary usage will be like. Analyzing usage logs and ratings as a source of information to continuously improve a cognitive decision solution is a best practice.

Do not be in a rush to ingest thousands or millions of documents if you do not understand how to properly ingest a single document.

Define the success criteria for your use case up front and determine metrics to support that. 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, then continuously evaluate them. If possible, gather metrics on every code, process, or ground truth change. Without continuous metrics evaluation, you cannot measure whether your system is improving or regressing, nor can you understand when you will be "good enough" or "done".

**Test automation**

Even more important than continuously verifying that your system is maintaining its performance is making sure that the system actually functions. Plan on a test automation suite to verify that all components in the system work together end-to-end. This includes verifying your pre-processing code, ingestion code, and your application code. This test suite should use a dedicated subset of your corpus and is ideally executable within an hour. The test suite should verify that your latest pre-processing and ingestion code and your configuration are compatible with the downstream queries and integrations you wish to perform on the data.

You should plan on automating your entire content ingestion process, especially if your solution has many moving parts. If you have automated the content ingestion, automating testing is simplified.
Planning for a cognitive solution

Figure 5 shows personas that are typically present in the planning phase.

- **Business architect**: This person knows the source of information, also known as the ground truth. It could be a training manual, product manual, testing manual, or external, publicly available information. The business architect defines the goals and objectives of the conversational application or bot, including the channels that the application needs to support (web, mobile, social, and others).

- **Data scientist or knowledge engineer**: The data scientist supports the business architect to understand the right kind of information that is needed for training the cognitive conversation system. Data scientists have deep knowledge of information that can be used to extract insights.

- **Domain experts**: Domain experts can also be known as subject matter experts (SMEs). This includes resources who have knowledge and understanding of the
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 a subject matter expert in a use case for an appliance manufacturer is the field technician. Other examples are product experts, call center supervisors, scientists, doctors, engineers, or the like.

This planning phase involves two broad categories of information sources:

- **Internal content sources**: This includes processes and data sources that are within a given enterprise. Typically, they contain the data generated (and owned) by the enterprise as part of its business operations.
  - **Business processes**: This involves enterprise level business processes with which the discovery solution might have to interact in order to process and respond to the user’s intent.
  - **Enterprise APIs or services**: APIs or services that might need to be accessed or invoked in response to the user queries and discovery responses. Most systems of records are likely to involve some API to serve the data they generate or control.
  - **Shared file repository**: Information kept in file systems shared between users and locations accessible through FTP and other mechanisms.
  - **Content from enterprise systems**: This includes data from various enterprise systems including but not limited to things like catalogs, order and transaction data, and ECM repositories.

- **External content sources**
  - **Public and third-party sources**: This includes 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 includes both public domain data (that is, available free of cost) and data controlled by other parties. Examples include weather data, domain-specific catalogs made available by third-party vendors, point of interest data, and more.
  - **Multimedia content**: This includes content like audio, video, or images 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 includes data accessible for public consumption that 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 using discovery. Another purpose is to identify the processes and the APIs that
might need to be invoked to fulfill the intent. You must remember that training the discovery service is an iterative process. For example, you might need to normalize content.

You must understand your content sources and formats thoroughly. Watson Discovery works best on unstructured text in documents. If these documents include images and tables, you must decide whether to omit them or to integrate them separately. Consider the content formats you have available: are they Microsoft Word documents, HTML, PDF, JSON, or some other format? If they are PDFs, are they scanned or searchable PDFs? If they are scanned PDFs, they must be of sufficient scan quality that you can use OCR (optical character recognition) to read the output. Content sources that do not have attached metadata, such as document type and date, might require you to infer this information.

If the files are not formatted to be easily consumed by the default document conversion of Watson Discovery, you must plan additional conversation steps to test them in the data preparation stage. Depending on the document format and structure, you may need to allocate significant time and planning for complex conversions. Try simple conversions first and then iterate.

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

**Phase 2: Preparation**

The next step in setting up Watson Discovery to support other cognitive 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.

Preparation of the data requires some level of training the system. This training might be carried out 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 cognitive glossary. The decision should be based on the advice of your SMEs and the understanding of the project team of the desired end state and the approach to maintaining the collections.

The preparation is likely to take place 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 persons
will use the guidance provided by the domain experts to develop the rules that enable automation of curation and collection.

**Phase 3: Ingestion**

There are three ways of ingesting data into Watson Discovery:

- APIs allow the upload of content with an existing application or the creation of your own custom upload mechanism.
- Discovery tooling gives you a quick upload capability of locally accessible files mostly for configuration and testing.
- Data Crawler provides managed upload of a significant number of files from a supported repository (such as an IBM Db2® database).

The Data Crawler is a Java® command line tool for Linux® that can help you take your documents from the repositories where they reside (such as file shares, databases, Microsoft® SharePoint) and push them to the cloud to be used by the Discovery service.

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

Discovery comes with reasonable defaults for PDF, Word, HTML, and JSON file conversion. Discovery first converts PDF and Word to HTML, then converts HTML to JSON. You may 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.

**Phase 4: Consumption**

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

**Watson Discovery runtime architecture for Weather Insights**

Previously we described the runtime flow needed for a typical cognitive solution and showed how to plan and prepare Watson Discovery to ingest and annotate content. You should understand the following concepts when creating an application.
• **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 would often load and annotate unstructured documents to train a ranker model to rank the returned passages for a given utterance.

• **Collections**: A collection is the logical grouping of your documents within your environment and your configuration is what tells the collection how to ingest and enrich this collection. Each collection has a unique configuration pipeline, which means that all documents are converted, enriched, and normalized in a unique way.

• **Ground truth**: This is the information used to train the discovery service. Content from public and enterprise sources, under guidance from domain SMEs, is used to model the intents and entities. Ground truth is also typically split into training, testing, and evaluation data.

Figure 6 illustrates the architecture for the Weather Insights application, showing the relation between the capabilities of the data collection application and the functionality needed for the user. It shows how different users interact with a trained cognitive system and how the system interacts with other components on the cloud platform. It also shows that data can come from public source or the customer’s enterprise, either in the form of raw data or in the form of an API.
In figure 6, items A – I are the activities needed to build the ground truth and content collection for the Weather Insights application. Items 1 – 3 describe the user experience.

A home improvement store in North America wants to offer a cognitive 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 make preparations 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 related to hurricane preparedness. The time required to get the most current knowledge is important. It is not uncommon for residents to get the wrong information about the 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 corpus of knowledge related to hurricanes. This could involve guidance in preparing 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 – I describe the activities needed to build the ground truth and content collection for Weather Insights:

A. Public data sources include data sources available for public consumption with or without any fee or subscription. This includes all data owned by the third parties, and hence resides outside the enterprise (or the agency) that consumes the data. In this specific example, this includes documents published by FEMA (Federal Emergency Management Agency) and other governmental agencies, publicly available information such as blog postings, and information published by individuals and various NGOs. The subject matter of these documents includes things like weather information, instructions related to hurricane preparation, helpful tips, steps for putting up shutters, various kinds of checklists, and more.

B. The ingestion application or the discovery tooling uses the identified information in step A and calls the discovery service.

C. The discovery service uses the ingestion application to ingest the information in the discovery collection.

D. Emergency employee response, supply chain disruption, and other information documents from private sources are identified.

E. The ingestion application or the discovery tooling uses the identified information in step D and calls the discovery service.

F. The discovery service uses the ingestion application to ingest the information in the discovery collection.

G. The business domain expert logs on to the knowledge studio.

H. The business domain expert enriches, annotates, and builds relationships to the already ingested content.

I. The custom model is then deployed in the discovery service.

**User interaction flow**

Items 1 – 3 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 particular question from the list and sends the request to the discovery application.
2. The discovery application sends the question to the trained discovery device.
3. The trained discovery service returns the response of the query.

**Best practice approaches for your discovery services project**

**Do your pre-work**

Pre-work is very important for the success of a discovery service. Allow ample time for advance planning; without it your chances of failure increase by more than 50%. Pre-work should include identification of the following:

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 will often submit? Assessing the kinds of queries and searches that your users will make can help train the system for realistic scenarios, resulting in higher accuracy in the results.
3. Corpus of knowledge: Identification of the knowledge sources is important to get the application working correctly. This includes identification of documents with potential answers and the selection of content sources that will be used to train the discovery service.

The final required pre-work step is to analyze the document content to assess how answers should be extracted.

**Understand your content sources**

While there are many publicly available sources of information that you can use 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 also includes the customer's knowledge of their own database and the way the data is expressed (like aliases, abbreviations, acronyms, operation specific codes, and more). For example, a call center manager might know where the historic user conversations are stored and know the definitions for various domain-specific terminology used in the conversation. A medical doctor might know which patient cases would be better for training; a training or
curriculum specialist might know the right sources for education; and an insurance agent might know policy details.

The most important question of all is whether the content sources 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**

Governance of the content sources used for training is very important. This means that there is a process in place to assure the quality and consistency of documents and other unstructured content. This also implies that the training content is versioned and managed, and that trainers have a means of keeping 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 ensure a continuous improvement in the result set over time. Establishing a cognitive center of excellence as part of your data and information governance program also ensures that the right set of resources, information, and people will be used for training the discovery service for subsequent projects. It also can provide guidelines for creating new content so that it is most usable by the discovery or other cognitive service.

You might need to update your machine learning models due to new product launches, acquisition or merger, product end of life, and other common business scenarios. Do more than anticipate the need to update. Build on the work begun in your planning and pre-work by making a change management process part of your pre-work, preparation, and training stages. By coming up with 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. You do not need to worry if your process is not 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 using 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 a cognitive 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 subject matter experts may be in short supply or have high salaries, automating the collection and curation of content via machine rules allows the SME to spend some time training, then move on to higher value tasks.

Executive sponsorship and business support

Cognitive solutions are often very business or operations-centric. IT, functional groups, and lines of business need to partner for both immediate success and for sustainable results. Creating a cognitive center of excellence for these groups to engage in can help expand adoption of cognitive solutions and provide a framework for sharing lessons learned. Since 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, multiple tenancy might need to be changed.

Security architecture: Content and data collection

Security is a very important aspect of the cognitive reference architecture. Security in the cognitive 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 cognitive 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 with access privileges to perform a certain task are allowed to perform that task. The policies for enforcing access control should be maintained in the policy administration system.

Figure 7: 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 through the use of Watson Knowledge Studio. The task of the creation and management of the training models in Watson Knowledge Studio should 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 applications and users are allowed to access the Knowledge Studio for enriching the ingested data with custom annotators.
• Content like 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 residing in the enterprise data center, step 7 for content residing 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 run the crawler application (step 6) provided by Watson Discovery to crawl for content in the customer data center to use in training Discovery.
• The organization should define encryption policies for encrypting the content used for training the discovery service. The policy should address securing the content at rest and in motion. Watson Discovery provides the capability for encrypting the content at rest and in motion.
• The security monitoring and intelligence component continues to monitor assets and information for advanced threats.

**Security for usage of the trained discovery service**

• Once 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 like searching product information, all users could 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 the capability for encrypting content that is part of its data collection, thus the data collection of any trained discovery service can be encrypted.
• The security monitoring and intelligence component continues to monitor assets and information for advanced threats.

**Components**

Let’s examine the individual components that make up the cognitive architecture.
Public network components

The public network contains elements that exist in the Internet: data sources and APIs, users, and the edge services 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 cognitive 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 cognitive system.

Cloud network components

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.

Key capabilities in this domain include:

- **Domain name system server**: Resolves the URL for a particular 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 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 meeting a set of policies to proceed and blocking any traffic that does not 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 reliability of applications. Load balancers can balance loads locally and globally. Load balancers should be highly available without a single point of failure. Load balancers are sometimes integrated as part of the provider cloud analytical system components like stream processing, data integration, and repositories.

• **Managed file transfer (MFT) gateway:** This is a multi-protocol gateway (AS2, AS4, sftp, ftps, C:D) into and out of the organization that 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 leveraged to monitor network traffic and to provide protection from hidden security vulnerabilities. IBM DataPower® provides load balancing and SSL termination. It can help quickly secure, integrate, control, and optimize access to a range of workloads through a single, extensible, DMZ-ready gateway.

**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 or integrated with other applications to enhance functionality.

Among the advanced AI functions that come ready for immediate use are natural language queries, passage retrieval, relevancy training, relationship graphs, element classification, and anomaly detection. All can assist the organization in identifying their specific knowledge and data assets and in achieving results more efficiently.

IBM Cloud offerings by default do not share any log information.

**Content storage**

Watson Discovery includes storage which is provided via the collections discussed in the previous section.
Application logic

Application logic, which might be a Node.js application, first passes the natural language utterance (request) to the conversation service. When it receives the response from the conversation service, the application logic checks the level of confidence. If the level of confidence is above a set threshold, it returns the response to the user. The application logic may need to invoke APIs to fetch the answers needed to fulfill the intent detected from the utterance. If the confidence levels are low, the application logic checks for possible answers using a discovery service.

IBM Cloud platform provides containers that are portable and allow for consistent delivery of your app without the need to manage the underlying operating systems. IBM Cloud also provides Cloud Foundry services, so you can deploy your application without managing the underlying infrastructure.

The applications built for Cloud Foundry-based services like Node.js or container-based deployments like Liberty for Java are built to orchestrate, choreograph, or enrich decision management or to produce actions that are based on cognitive or analytical insights. These cloud platform services are essential for the success of cognitive systems.

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, the IBM Integration Bus container allows you to integrate applications and infrastructures deployed in multiple clouds or in legacy or core applications deployed in customers’ traditional data center.

IBM API Connect® is a comprehensive end-to-end 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. Using automated, model-driven tools, you can create new APIs and microservices based on Node.js and Java runtimes—all managed from a single unified console. Ensure secure and controlled access to the APIs 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 running on-premises or in other clouds. A remote client is provided to enable secure connectivity.

Enterprise network components

Ground truth

This is the training data for Watson Discovery and related services APIs. This includes a variety of 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 Discovery will automatically handle the training of the machine learning model. Read the requirements for training data that are described here.

IBM capabilities for security for content discovery and management

The table below shows the IBM capabilities and services mapped to the components in Figure 7.

<table>
<thead>
<tr>
<th>COMPONENT</th>
<th>DEFINITION</th>
<th>IBM and other products</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge services</td>
<td>Edge services include services needed to allow data to flow safely from the Internet.</td>
<td>DNS, CDN, firewall, load balancer</td>
</tr>
<tr>
<td>Transformation and connectivity</td>
<td>This includes scalable messaging and transformation and secure connectivity.</td>
<td>IBM Integration Bus container, IBM DataPower, IBM API Connect, IBM Secure Gateway</td>
</tr>
<tr>
<td>Key management service</td>
<td>A cloud-based security service to provide key lifecycle management (key creation, usage, and deletion) for encryption keys</td>
<td>IBM Key Protect</td>
</tr>
<tr>
<td>Service</td>
<td>Description</td>
<td>Provider</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------</td>
</tr>
<tr>
<td>File encryption service</td>
<td>Used in IBM Cloud services or in customer built applications, with &quot;root of trust&quot; backed by a hardware security module (HSM).</td>
<td>IBM Multi-Cloud Data Encryption</td>
</tr>
<tr>
<td>Secured connectivity</td>
<td>Safeguards data even when network protection fails. It has built-in key management for handling storage of all the encryption and splitting keys.</td>
<td>VPN providers</td>
</tr>
<tr>
<td>Identity and access management</td>
<td>Identifies and authenticates the user. Determines access levels by using an enterprise security directory such as LDAP.</td>
<td>IBM Security Access Manager</td>
</tr>
<tr>
<td>Security monitoring and intelligence</td>
<td>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. Provides real-time risk analysis of the workloads hosted in cloud against the myriad of known vulnerabilities and alerts against zero day attacks.</td>
<td>IBM QRadar® SIEM</td>
</tr>
<tr>
<td>Infrastructure security</td>
<td>Protects against network-level threats and attacks with intrusion prevention and detection, including those tunneling through encrypted web transactions and web applications deployed within the system.</td>
<td>IBM Security Server Protection, IBM Security SiteProtector System</td>
</tr>
</tbody>
</table>
Watson Discovery: Planning for success

There is no limit to the number of collections you can create in your services environment. The limit is the storage size for the environment and the number of documents allowed by the subscription. Your index size cannot exceed your storage. The size of your index will vary based on document size, the number of fields you index, and how many enrichments your data contains. Your index can easily 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.

The determining factors when deciding the number of collections needed are your data type, format, and structure, and whether you want logically separate development, test, staging, and production environments. When data sets are not consistent, you will likely need multiple collections. You may need to modify some types of files, such as spreadsheets, prior to ingestion. Multiple collections allow experimentation and testing of the configuration pipeline without affecting production activities.

The right cloud platform

To create an efficient cognitive environment, your enterprise-grade cloud platform should 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 platform gives enterprises the following:

- 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 secured 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 designed for industry-specific data or content and regulations. This gives you broad variety in the information available to Watson Discovery for use in creating collections. IBM Cloud provides 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 a cognitive business. IBM Watson services can be strengthened by content captured in the peer cloud of companies that have instrumented the physical world to create a robust ecosystem for solving business challenges.

Here are some examples of how Watson services can be used in conjunction with ecosystem partners to gain insights into data:

- Content captured by car manufacturers can be used with IBM Watson services to provide car-related information to drivers and for self-driving cars.
- Content captured by health-monitoring devices like 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.

Refer to the Cloud Customer Reference Architecture IBM advantage paper for Internet of Things (IoT) for a thorough discussion and best practices.

Deployment considerations

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

The following are some important considerations for deploying cognitive systems:

Isolation

This consideration involves deploying to customers who carry client confidential or sensitive private information. In these cases, you might want to use a premium or a dedicated deployment options to support your application or discovery service, whether it is a multi-tenant or single tenant implementation. Where there are no or fewer confidentiality concerns, choosing a multi-tenant model provisioned using a standard or public deployment option might be acceptable.

Privacy

A general rule is not to store or pass any confidential information or protected health information (PHI) when interacting with a cognitive system. Where data is encrypted end-
to-end, this rule does not apply. Watson Discovery in Dedicated and Premium deployments offers encryption at rest and is suitable for PII.

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 existing shared platform, as well as end-to-end encrypted data while in transit and at rest.

**Region and language support**

When deploying applications that involve multiple geographies and languages, the services might need to be deployed in multiple regions using the IBM Cloud region settings. Additionally, cognitive systems must be designed and trained against various languages based on the support provided by the service. 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, you should create a testing plan that involves load testing. You can use open source frameworks like JMeter or third-party services such as Blazemeter in IBM Cloud to create and execute load tests. The load test should include submitting various request sizes and concurrent users. Depending upon 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. In addition, you can employ capabilities such as auto-scaling to configure the scaling based on demand, throughput, and memory utilization.

**References**

- The IBM Advantage for Implementing the CSCC Cloud Customer Reference Architecture for Internet of Things (IoT)
- ISO/IEC 17788:2014 Information technology -- Cloud computing -- Overview and vocabulary
- ISO/IEC 17789:2014 Information technology -- Cloud computing -- Reference architecture
- IBM Watson Assistant API
- IBM Watson Speech to Text
- IBM Watson Discovery portfolio
  - IBM Watson Discovery
  - IBM Watson Knowledge Studio
- IBM Watson Natural Language Understanding
- IBM Cloud catalog
- Cognitive glossary