

Enterprise Knowledge Mining Solution Microsoft AI Platform

LearnAI Team
October 2018

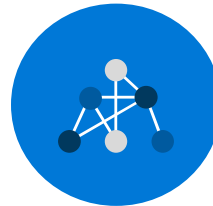
Azure AI

AI apps & agents



Azure Bot Service
Azure Cognitive Services

Machine learning



Azure Databricks
Azure Machine Learning
Azure AI infrastructure

Knowledge mining



Azure Cognitive Search

AI uses **cognitive skills** to simulate human perception for:

- Vision
- Speech
- Language
- Knowledge

Customized language understanding Text-to-speech
Content moderation Spell check
Speech translation
Custom image classification
Speaker recognition Entity linking
Sentiment analysis, & augmentation
key phrase extraction **Image tagging**
Custom voice Object detection Text translation Intend
OCR handwriting analysis
Emotion detection **recognition** Custom translation
Video insights **Face** Custom speech Assisted text moderation
identification Speech transcription

Knowledge Mining is the process of discovering actionable information from large sets of **unstructured data**, like text or images.

It uses **Artificial Intelligence** to detect hidden patterns and information. It can be used to guarantee compliance, enrich search, automate processes, among other **business processes**.



Cognitive Search is the Microsoft solution for Knowledge Mining.

AI approach for content understanding. Ingests unstructured content to create rich metadata into an **Azure Search** index, with the power of **Cognitive Services**.



What Azure Search is

- AI-Powered cloud search service for web and mobile app development
- Enterprise Search as a Service – Azure PaaS product
- Enrich and extract insights through cognitive skills (Cognitive Search)
- Easily scale up and down
- Natural Language Processing for web search grade experience
- Connect search results to business goals with great control over search ranking
- 99.9% SLA, GDPR, Standard Azure OST (Online Service Terms)
- Creates indexed metadata about your data

What Azure Search is

The screenshot displays the Azure Search Job Portal Demo interface. At the top, the header includes the Azure Search logo, the text 'AVAILABLE JOBS (180 jobs)', and navigation links for 'Home', 'Jobs', and 'About Azure Search'.

Search and Suggestions: The search bar on the left shows the input 'analis'. A dropdown menu provides suggestions, including 'DATA AND BUSINESS ANALYST Strategy & Analytics', 'DATA AND BUSINESS ANALYST Performance Mgmt. & Analytics', 'Technology and Analytics Trainer Performance Mgmt. & Analytics', 'Workforce Data Analyst/Statistician Data Analytics Center', and a list of specific roles like 'Policy Analyst (6)', 'Procurement Analyst (6)', 'Change Order Analyst (4)', 'Claiming Analyst, Bureau of Budget and Revenue (4)', 'Junior Claiming Analyst, Bureau of Budget and Revenue (4)', 'Junior Tax Credit Analyst (4)', 'PMO ANALYST (4)', '421 - a Analyst (2)', and 'ANCP Project Manager / Junior Underwriter / Analyst (2)'. A blue callout box labeled 'Spelling Mistakes' points to the search bar.

Geospatial: A map of New York City is displayed, showing various locations. A blue callout box labeled 'Geospatial' points to the map.

Ranking: The results are ranked by 'Relevance'. A blue callout box labeled 'Ranking' points to the 'Relevance' dropdown menu.

Paging: The results are paginated, showing 180 available jobs. A blue callout box labeled 'Paging' points to the pagination controls (1, 2, 3, 4, 5).

Facets: The 'LOCATION' facet is visible on the left, showing 'Internal (92)' and 'External (99)'. A blue callout box labeled 'Facets' points to the 'LOCATION' section.

Hit Highlighting: The first job listing is 'Budget Analyst, Family and Child Health Administration'. The text is highlighted in blue. A blue callout box labeled 'Hit Highlighting' points to the highlighted text.

Job Listings: The first job listing is 'Budget Analyst, Family and Child Health Administration' with a salary range of \$45,358 to \$61,754 Annual. The second job listing is 'PMO ANALYST - Featured Job' with a salary range of \$45,174 to \$62,370 Annual.

What Azure Search is

Formats

- Microsoft Office
- PDF
- HTML
- XML, JSON
- ZIP
- EML, RTF, TXT
- JPEG, PNG, JPG
- CSV

Data Sources

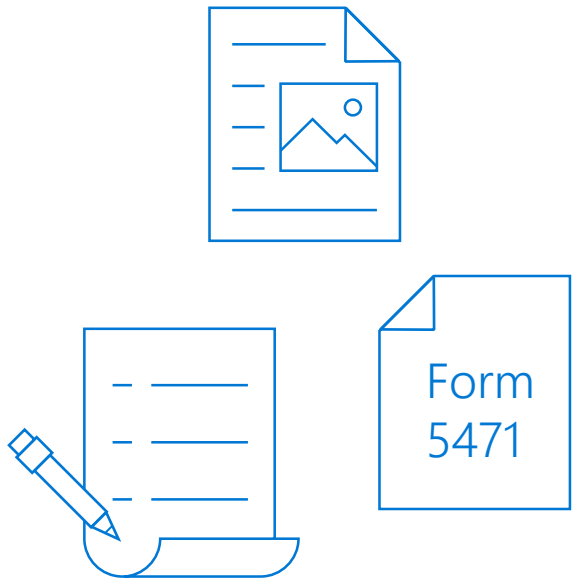
- Azure SQL
 - Text columns
 - JSON/XML columns
- SQL Server IaaS on Azure
- CosmosDB
- Azure blob storage
- Azure Table
- MySQL (PaaS and IaaS)
- Azure Files (Private Preview)

What Cognitive Search is

- **Azure Search feature**, Announced in May 2018 (MS Build)
- It uses AI to create searchable metadata, transforming unstructured data into information
- Data Enrichment != Data Integration. Original data isn't moved, changed or copied
- Results are always loaded into an Azure Search Index
- 10+ regions, including South Central US, West Europe, North Europe, Brazil South, and Southeast Asia
- Azure Search Cost + Cost per image + Cost per Cognitive Service used
- Free up to 20 documents per day

What Cognitive Search is

Enrichment Pipeline



Unlock valuable
information lying latent
in all your content

What Cognitive Search is

Documents



Enrichment Pipeline



Key Phrase extraction



Organization entity extraction



Face detection



Custom skills



Location entity extraction



Persons entity extraction



Celebrity recognition



Landmark detection



Sentiment analysis



Language detection

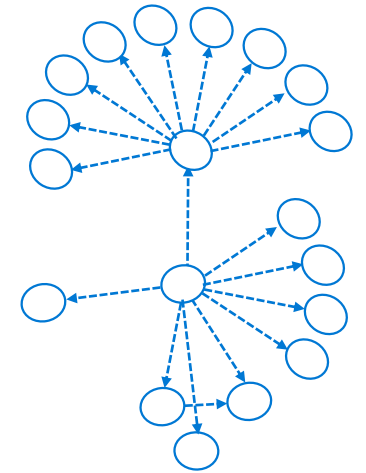


Tag extraction



Printed text recognition

Fully text-searchable
rich index



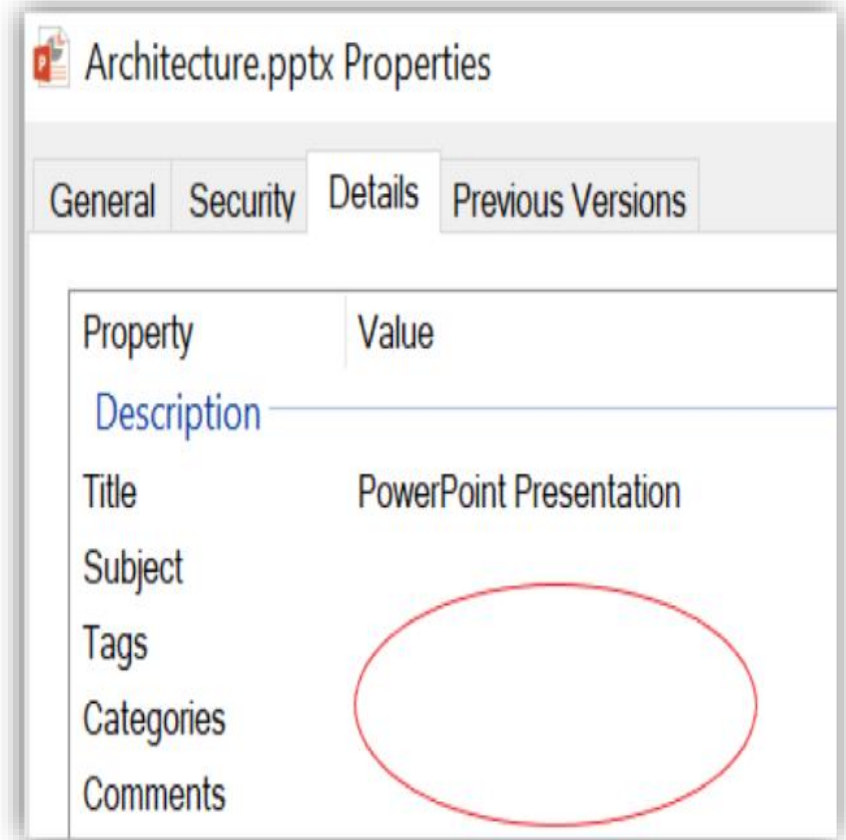
When to use it



Documents
Analysis



Understanding
engineering plans



Lack of
Metadata

When to use It

Enagic USA, Inc.
Headquarters
4115 Spencer St., Torrance, CA 90503
Phone: (310) 542-7700 / FAX: (310) 542-1700
Toll Free: (866) 261-9500 / cc@enagic.com

Product Order Form & Distributor Application

Kangen Water®

PRINT CLEARLY

Distributor ID # <Do NOT Fill In>

Applicant Information

Legal Name (First, Middle Initial, Last) or Company Name

Application Date:

Nguyen Van Le

Driver's License #

State

Date of Birth

Are you currently an Enagic Distributor?

C4574938

CA

02-03-1960

No ☒ Yes ☐ Enagic ID#:

Shipping Address (must match US)

City

State

Zip Code

1234 Senter Road

San Jose

CA

95051

SSN

Phone Number

City

State

Zip Code

523-45-6789

(408)765-4321

Cell Number

Fax Number

Email Address

(408)345-6789

nguyenle36@gmail.com

Billing Address (if different from mailing address)

City

State

Zip Code

Same above

Alternate Shipping Address

City

State

Zip Code

Sponsor Information

Sponsor Name

REGISTER THIS APPLICANT AS YOUR [4] A

Son Le

Under Sponsor

Phone Number

ID Number

(408)234-5678

7296044

ITEM ORDERED

SD501

PAYMENT METHOD

☒ SINGLE PAYMENT

\$ 3,980.00

Unit Price

+ 348.26

Tax

+ 23.00

Shipping

= \$ 4,351.26

Total

☐ ENAGIC PAYMENT:

☐ 3 months

☐ 6 months

☐ 10 months

☐ 16 months

Product Retail Price

\$ 3,980.00

ENAGIC PAYMENT:

☐ 3 months

☐ 6 months

☐ 10 months

☐ 16 months

Finance Amount

Monthly Payment Amount

Withdrawal Date (Circle One)

First Payment Date

\$

\$

1st / 15th

/ /

Payment Information: CREDIT CARD or CHECKING ACCOUNT

Void check needed for Checking Account Payment

☒ Visa

☐ Master Card

☐ Amex

☐ DISCOVER

Credit Card Number / Checking Account Number

Expiration Date / Checking Account Routing Number

CW #

No Owner's Card#

2345-6789-1234-5678

03/17

376

Card Holder Name (Please Print)

Card Holder Signature

Nguyen Van Le

Note: An applicant will be able to become a distributor with the purchase of 10 units Sales Kit.

I certify that I have been furnished a copy of, and have read, understand, and agree to the provisions in Enagic USA, Inc.'s Policies and Procedures manual, which documents (with any amendments or restatements furnished by Enagic USA after this date) are hereby incorporated by reference as if fully set forth herein and set forth the exclusive terms and conditions of my agreement with Enagic USA, Inc.

I hereby certify that the information provided on this form is complete and accurate to the best of my knowledge. I authorize ENAGIC USA, INC to debit the amount I have indicated above from my bank account or credit card. This agreement will remain in effect until the balance is paid in full. \$20 late fee will be applied to your account every time payment is missed. By signing the line below, you are acknowledging that you have read and understand the terms and conditions. Terms and conditions are subject to change without notice. If you fail to make a monthly payment, Enagic may offset the payment amount from your commissions. FOR ALTERNATE PAYERS: By signing Alternate Payer Form, you will be jointly responsible for any and all balance owing on the account. This agreement is governed by the laws of California and proper venue will be in a court of competent jurisdiction located nearest to the Company's headquarters.

Print Applicant Name

Print Sponsor Name

Nguyen Van Le

Son Le

Applicant Signature

Date

Sponsor Signature

Date

03-17-15

03-17-15

Change Your Water...
Change Your Life™

Revised 10/20/12

SHIP

PICKUP

"Golden Rules" for OP Amps.

1. No current actually flows into the - and + input terminals.
(Very large input Resistance).

2. The feedback makes the voltage at the input terminals the SAME.

$$V_0 = (V_1 + V_2 + V_3) \frac{R_F}{R}$$

$$\frac{R_F}{R} = \text{Gain}$$

$$R_F = 1M\Omega$$
$$R = 10k\Omega$$

Forms
Reading

Handwritten
Information

Image
Analysis

When to use It

Enagic USA, Inc.
Headquarters
4115 Spencer St., Torrance, CA 90503
Phone: (310) 542-7700 / FAX: (310) 542-1700
Toll Free: (888) 261-9500 / cc@enagic.com

Product Order Form & Distributor Application *Kangen Water®*

PRINT CLEARLY

Legal Name (First, Middle Initial, Last) or Company Name: **Nguyen Van Le** Application Date: _____
Distributor ID #: _____

Driver's License #: **C4574938** State: **CA** Date of Birth: **02-03-1960** Are you currently an Enagic Distributor? ☒ Yes ☐ No

Shipping Address (must match US):
1234 Senter Road San Jose CA 95128
Phone Number: **(408)765-4321**

Cell Number: **(408)345-6789** Fax Number: _____ Email Address: **nguyenle36@gmail.com**

Billing Address (if different from mailing address):
Same above City: _____ State: _____ Zip Code: _____

Alternate Shipping Address: _____ City: _____ State: _____ Zip Code: _____

Sponsor Information
Sponsor Name: **Son Le** REGISTER THIS APPLICATION: _____
Phone Number: **(408)234-5678**

ITEM ORDERED
SD501 ☒ SINGLE PAYMENT \$ **3,980.00**
Unit Price: _____
Product Retail Price: _____ ☐ ENAGIC PAYMENT: ☐ 3 MONTHS
\$ **3,980.00** + Handling + Tax = _____
Finance Amount: _____ Monthly Payment Amount: _____ Withdrawal Date: _____
\$ _____ 1st / _____

Payment Information: **CREDIT CARD or CHECKING ACCOUNT** ☒ Visa ☐ Master Card ☐ Amex ☐ Discover
Credit Card Number / Checking Account Number: **2345-6789-1234-5678** Expiration Date: _____
Card Holder Name (Please Print): **Nguyen Van Le**

Note: An application fee of \$20 late fee will be applied to your account if you do not pay the amount I have indicated above within 45 days of purchase date. Terms and conditions apply. This agreement is for any and all balance owing on the account. This agreement is subject to the terms and conditions of the Enagic USA, Inc. Standard Terms and Conditions of Sale, which are available at www.enagic.com. Jurisdiction located nearest to the Company's headquarters.

Print Applicant Name: **Nguyen Van Le**
Applicant Signature: *Nguyen Van Le* Date: **03-17-15** Sponsor: _____

Change Your Water...
Change Your Water...

Dataset of this training!!



For
Reading

Handwritten
Information

Image
Analysis

When to use It

How to identify Cognitive Search Opportunities

- Every company has ppts, pdfs, docx, html, images...
- Every company has contracts, forms, plans, memos, emails
- Every company needs compliance, risk detection
- Client needs a better search experience on top of business documents
- Global companies with enterprise documents in multiple languages
- Local file server is out of space
- Client already has data on Azure
- Client needs to apply AI to business documents: data != information

When to use It

How to ~~identify~~ **CREATE** Cognitive Search Opportunities

- **Every company** has ppts, pdfs, docx, html, images...
- **Every company** has contracts, forms, plans, memos, emails
- **Every company** needs compliance, risk detection
- Client needs a better search experience on top of business documents
- Global companies with enterprise documents in multiple languages
- Local file server is out of space
- Client already has data on Azure
- Client needs to apply AI to business documents: data != information

When to use It

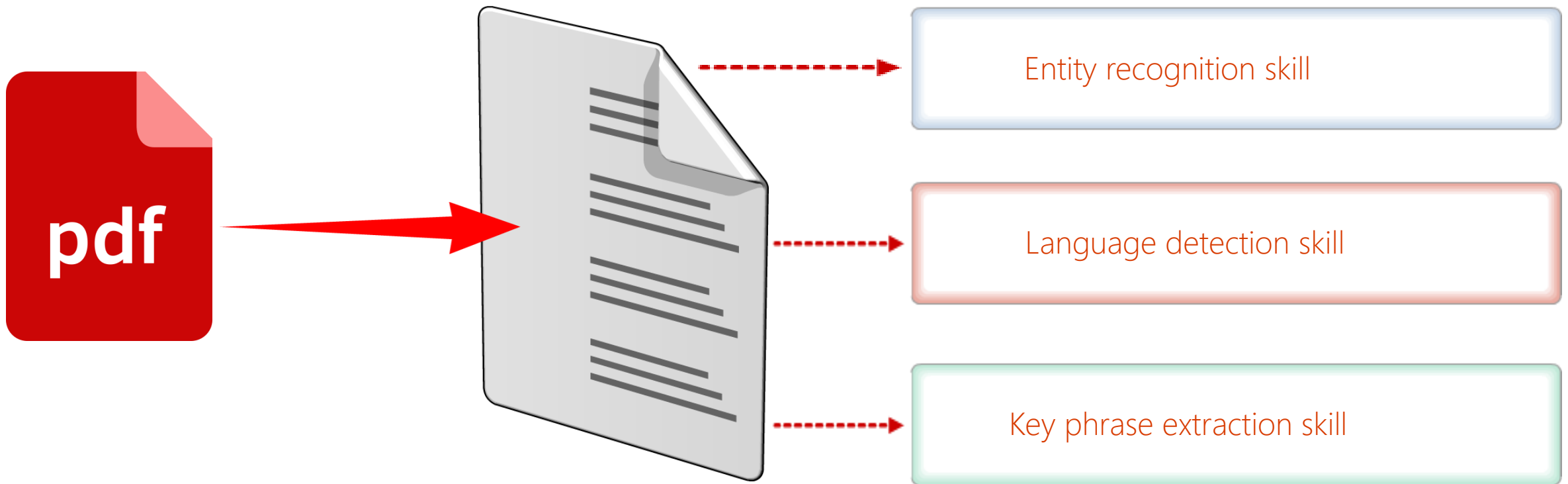
How to identify **CREATE** Cognitive Search Opportunities



80% of relevant
business
information is
unstructured,
usually text.

How it works

Cognitive Search **enrichment pipeline**: Atomic cognitive processes, aka **Cognitive Skills**, applied for each document, creating metadata



How it works

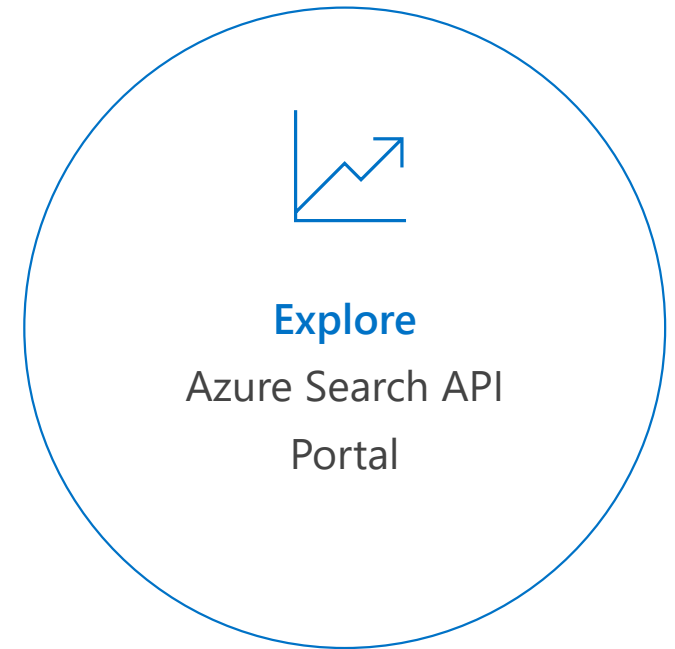
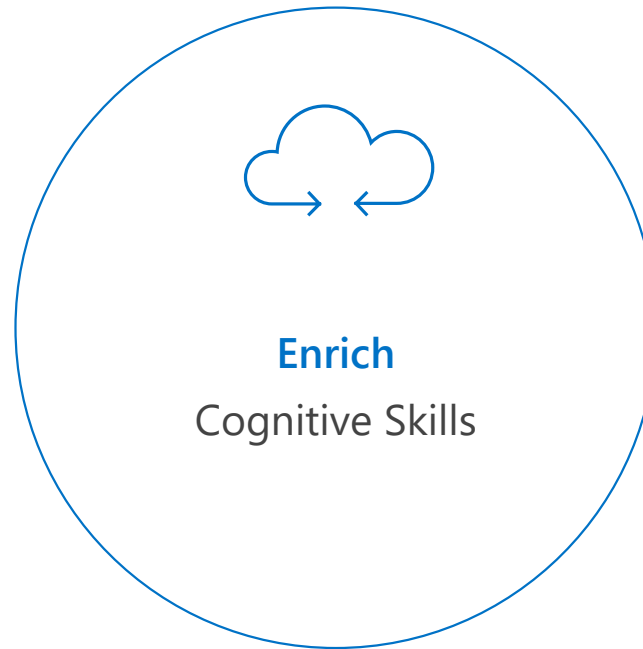
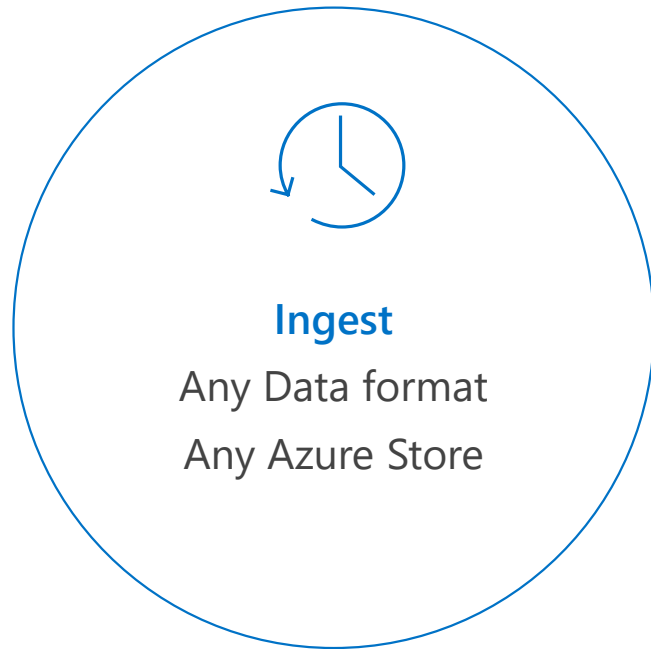
Predefined Cognitive Skills

- Key Phrases
- Language Detection
- Text Merger
- Entity Recognition: Names, Locations, Organizations
- Sentiment Analysis
- Text Splitter
- Image Analysis: categories, tags, description, faces, type, colors, adult content
- OCR
- Shaper: Complex types (matrix) – Private Preview

Custom Skills

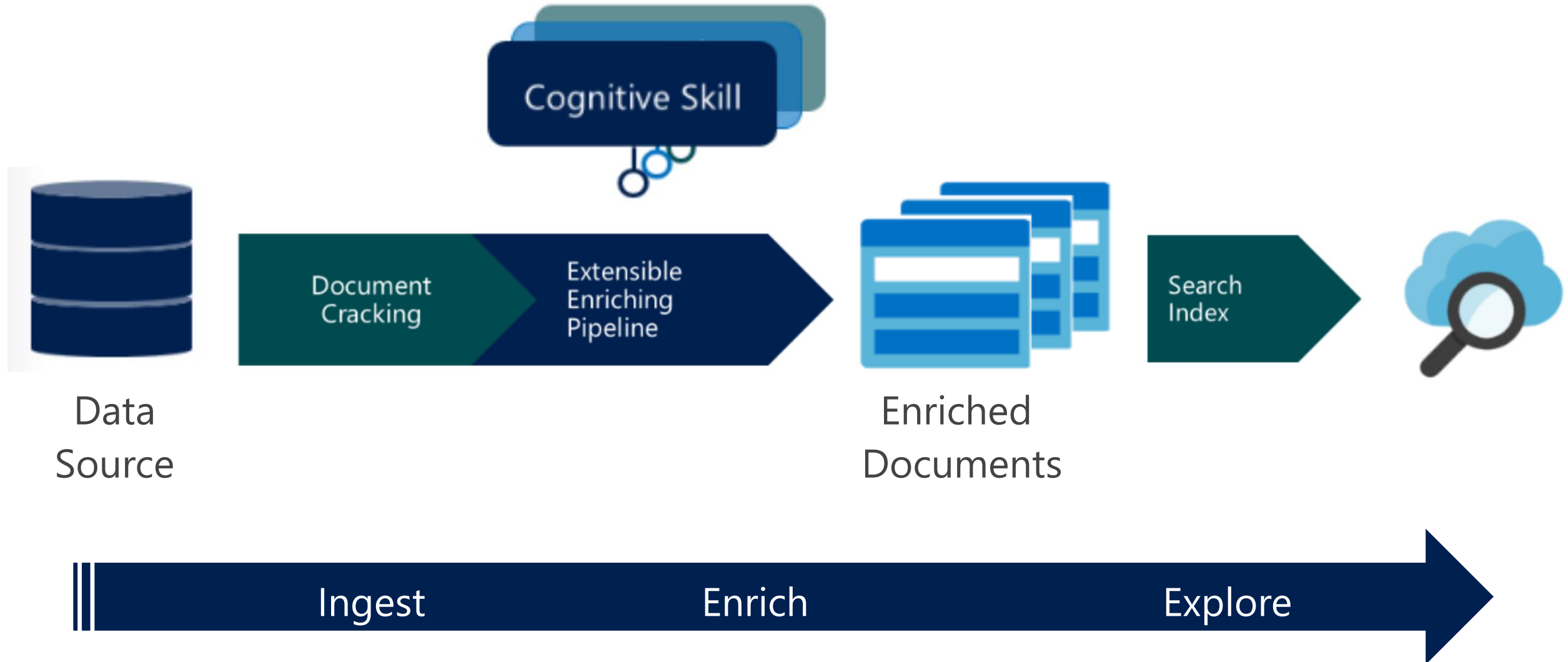
- Any REST API
 - Public
 - Client Specific
 - Microsoft Cognitive Services
- Create your own API
 - Industry specific
 - Hosted on Azure
 - Azure Functions
 - Web APP

How it works

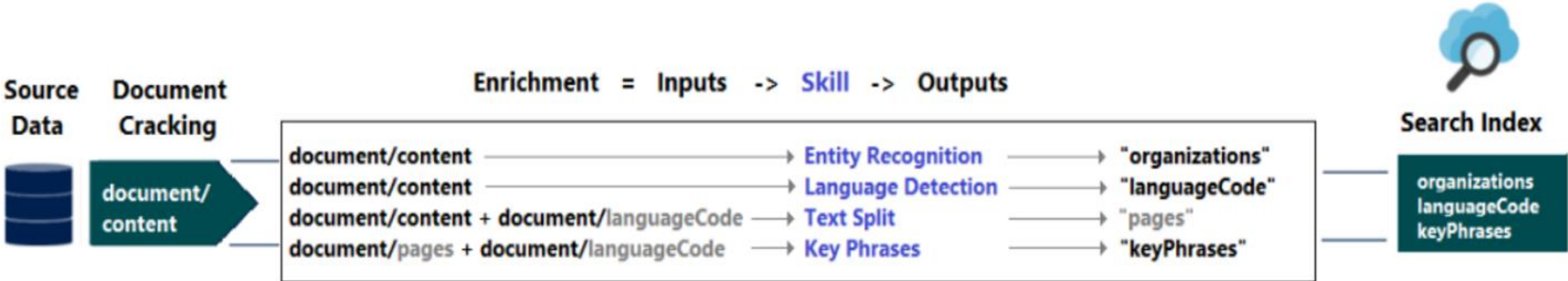


Cognitive Search Pipeline

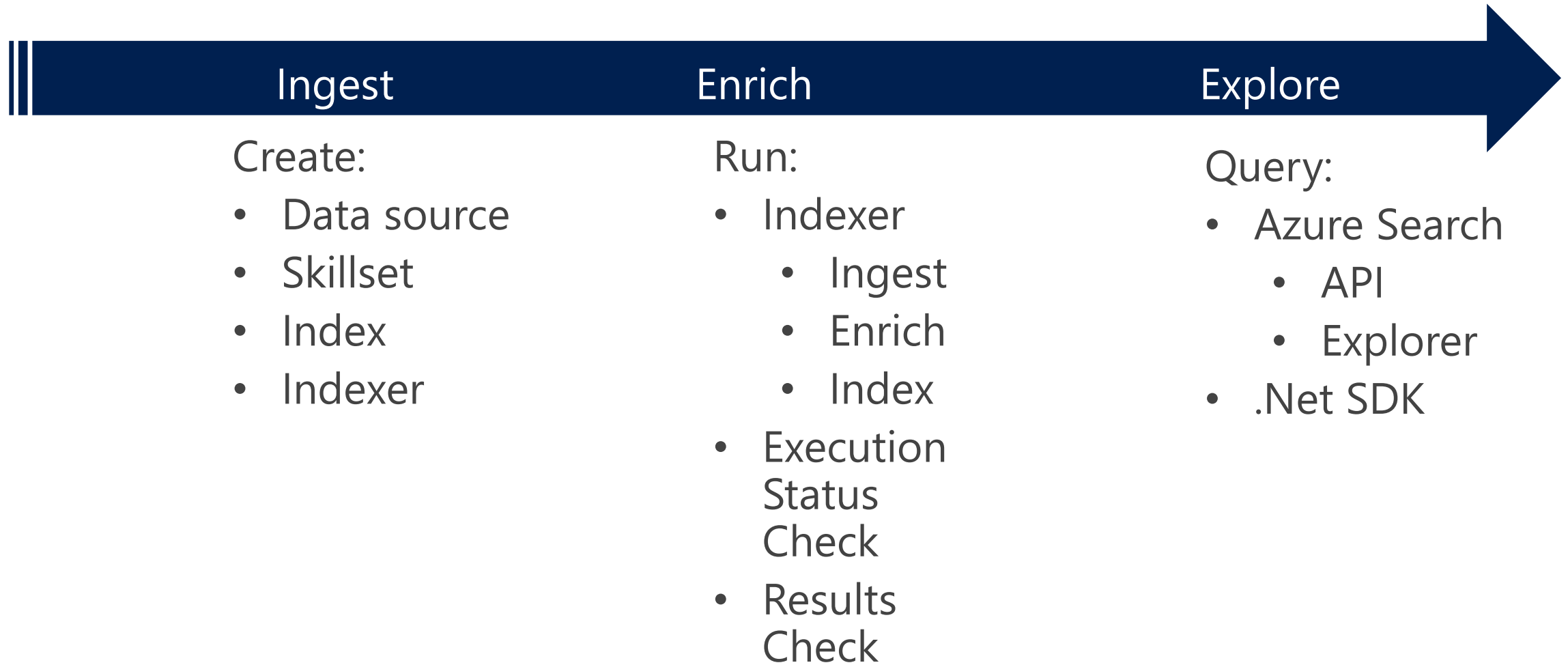
How it works – Cognitive Search Pipeline



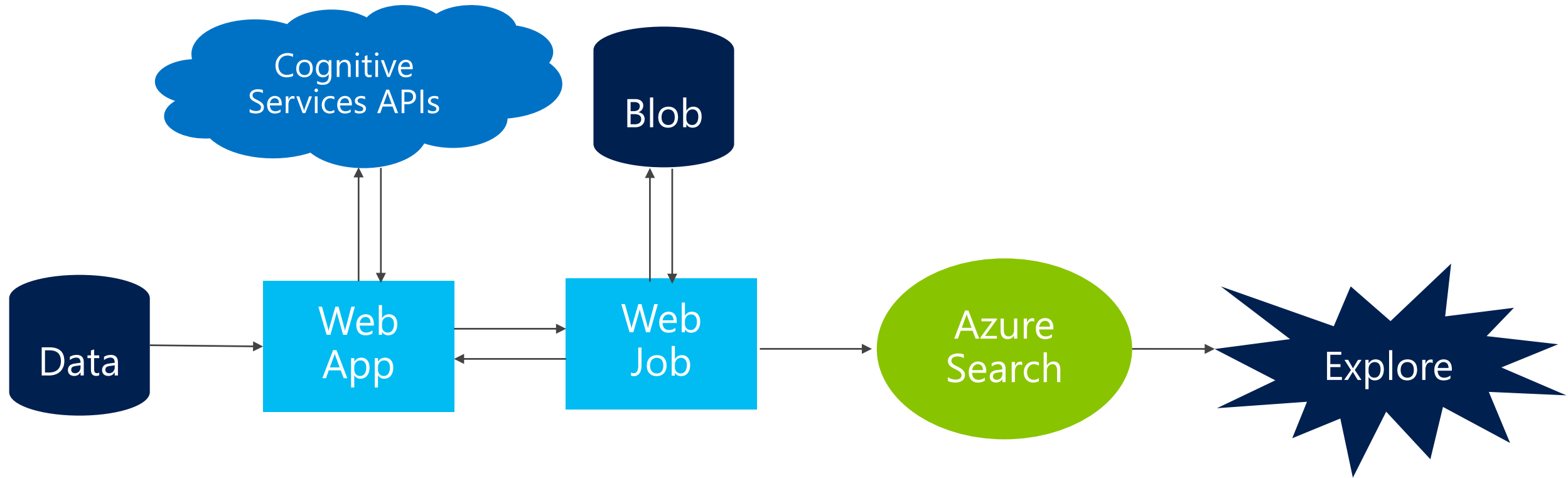
How it works – Cognitive Search Pipeline



How it works – Cognitive Skills



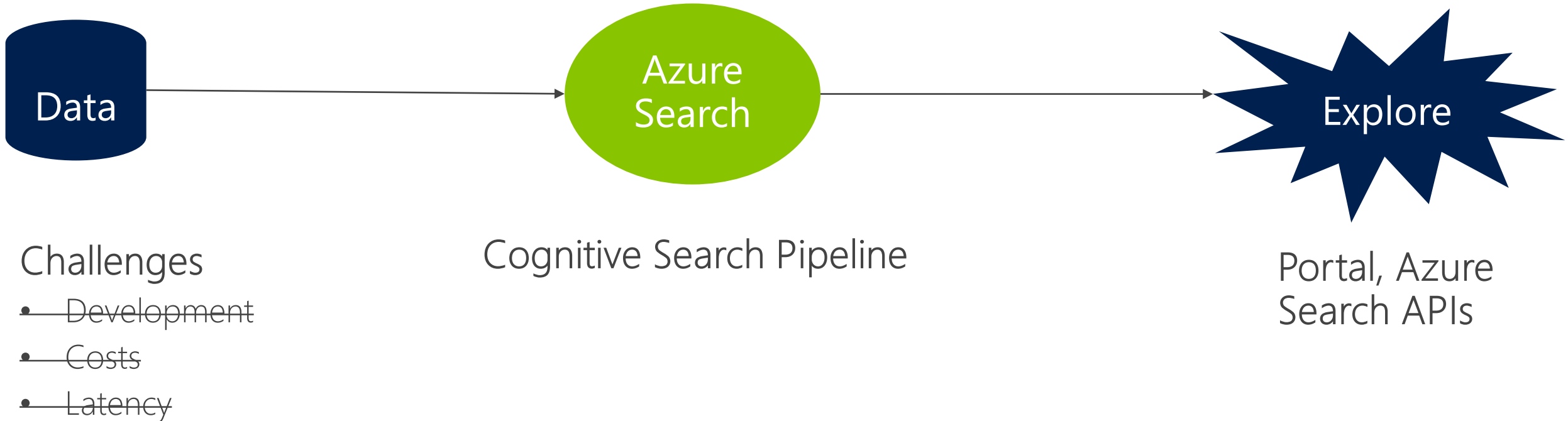
How it works – Old High Level Architecture



Challenges

- Development
- Costs
- Latency

How it works – New High Level Architecture



Good Practices

Supplement the indexes with:

- Natural Language Processing
- Custom Analyzers (50+ Languages)
- Fuzzy Search
- Filtered Search
- Scoring Profiles
- Autocomplete (new – June 2018)
- Synonymous (new – June 2018)


The screenshot displays the Azure Search Job Portal Demo interface. The top navigation bar includes the Azure Search logo, 'AVAILABLE JOBS (180 Jobs)', and links for 'Home', 'Jobs', and 'About Azure Search'. The main content area is divided into a left sidebar and a right main panel. The sidebar contains a 'SEARCH' section with a search bar and a list of job titles, and a 'LOCATION' section with a list of locations. The main panel features a map of New York City with several orange location pins, a list of '180 AVAILABLE JOBS', and a detailed view of a specific job listing. Annotations with blue callout boxes point to various features: 'Spelling Mistakes' points to the search bar; 'Geospatial' points to the map; 'Suggestions' points to the job title list; 'Ranking' points to the 'Relevance' dropdown; 'Paging' points to the page numbers; 'Hit Highlighting' points to the job title; and 'Facets' points to the location list.

Annotations:

- Spelling Mistakes
- Geospatial
- Suggestions
- Ranking
- Paging
- Hit Highlighting
- Facets

Good Practices

- Results are **always** loaded into an Azure Search Index. But you can do extra Analytics: Create a Custom Skill to save the enriched metadata:
 - CosmosDB
 - Azure SQL json column
 - Blob Storage
 - Others



Roadmap: Automatic
export of the enriched
metadata

Good Practices

- Clear Business requirements
- SQL Server on-prem is supported but not recommended
 - Performance
 - SQL Server open to the internet
 - Use ADF to copy data to Azure
- Co-locate the Azure Search Service with your data: Costs, Latency

Good Practices

- Leverage Azure Free Account and Azure Search Free Tier
 - 50 MB
 - 5 Cognitive Skills, 3 Indexes
 - Keep it up and running – demo in minutes with clients own data
- Security Trimming to filter content based on user identity
- Multiple Indexes for the same dataset may be required
 - Ranking, languages, filters...

Good Practices

- Incremental Updates for new files (portal or command line)
 - On Demand, File Watch, Schedule
 - Execution Overlap: The schedule essentially tells the indexer to proceed where it left off at the next hour
- Image Skills Performance
 - Deep Learning behind the sciences
 - While in public preview, performance will be better if you break documents with multiple images on multiple documents. When GA this will be done automatically

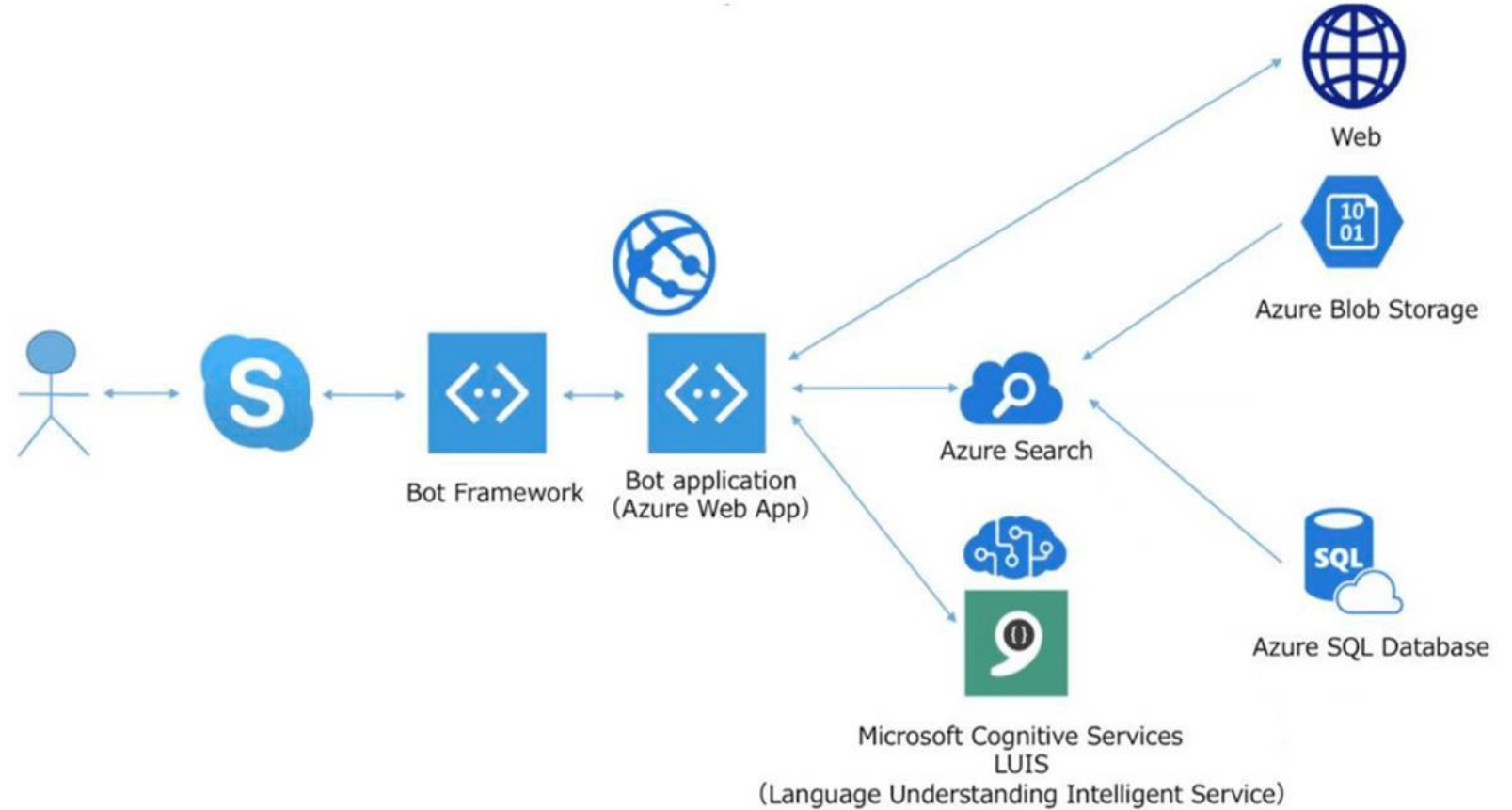
Good Practices

- Troubleshooting
 - Start with a small dataset
 - Make Sure your credentials are correct
 - See what works even if there are some failures
 - Check Status
 - Error Parameters
 - Enriched documents under the hood: saving the raw transformations into the index
 - Free and basic tears have limits for document sizes
 - Parallel Indexing
 - Maximum runtime

Good Practices

End to End Solutions

- .NET SDK
- REST API
- BOT + LUIS



Demos – Cognitive Search

Enrichment Pipeline

<https://text-analytics-demo-dev.azurewebsites.net/>

JFK Files

<https://jfk-demo.azurewebsites.net/>

Business Documents Demo - Wolters Kluwer

<https://wolterskluwereap.azurewebsites.net/>

Oil & Gas Demo - Exxon Mobile

<http://seismicsearch.azurewebsites.net/>

Healthcare - CTakes

<http://webmedsearch.azurewebsites.net/>

Additional Resources

<http://aka.ms/LearnAI-trainings>

<http://aka.ms/LearnAI-csw>

<http://aka.ms/LearnAI-kmb>

<http://aka.ms/LearnAI-links>

LearnAI – Knowledge Mining Bootcamp

