Enterprise Knowledge Mining Solution Microsoft AI Platform

LearnAI Team

January 2019
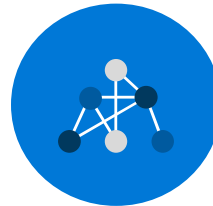
Microsoft

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

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
- 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

- **OR** Azure Search Cost + Free up to 20 documents per day

What Cognitive Search is

Enrichment Pipeline

Form 5471
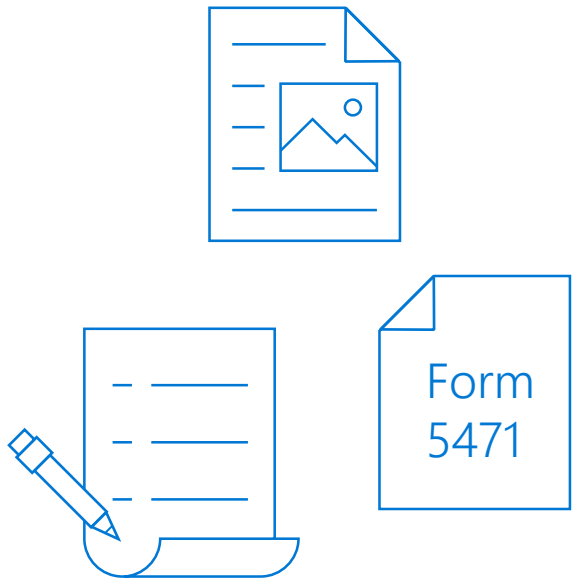
Unlock valuable information lying latent in all your content

What Cognitive Search is

Documents

Enrichment Pipeline

Fully text-searchable rich index

Key Phrase extraction

Location entity extraction

Sentiment analysis

Organization entity extraction

Persons entity extraction

Language detection

Face detection

Celebrity recognition

Tag extraction

Custom skills

Landmark detection

Printed text recognition

Form 5471
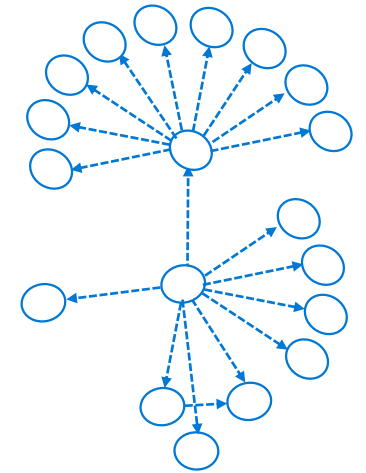
When to use it

Documents Analysis

Understanding engineering plans
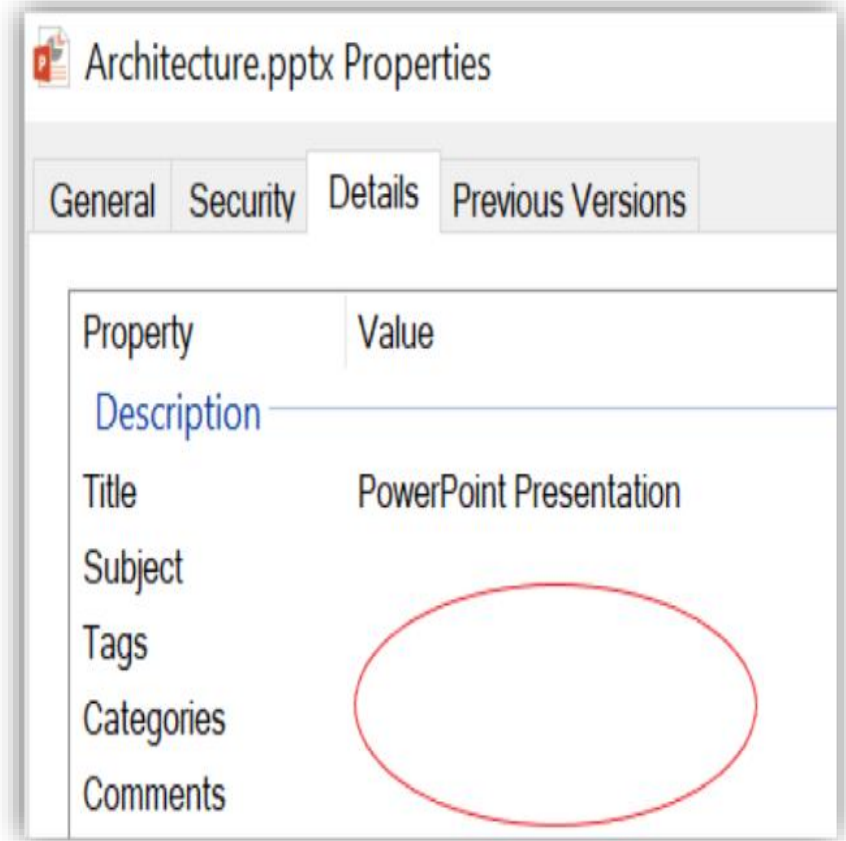
Lack of Metadata

When to use It

Forms
Reading

Handwritten
Information

Image
Analysis

When to use It

Dataset of this training!!

Form 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

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

Ingest
Any Data format
Any Azure Store

Enrich
Cognitive Skills

Explore
Azure Search API
Portal

Cognitive Search Pipeline
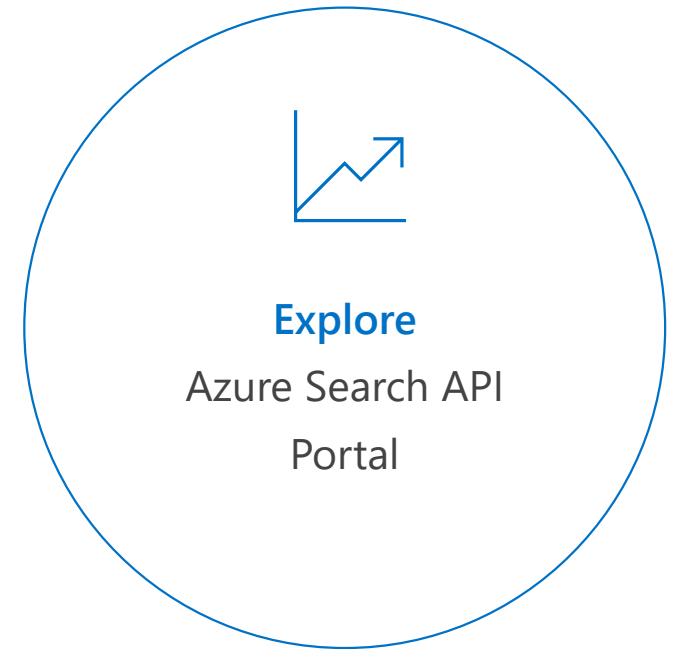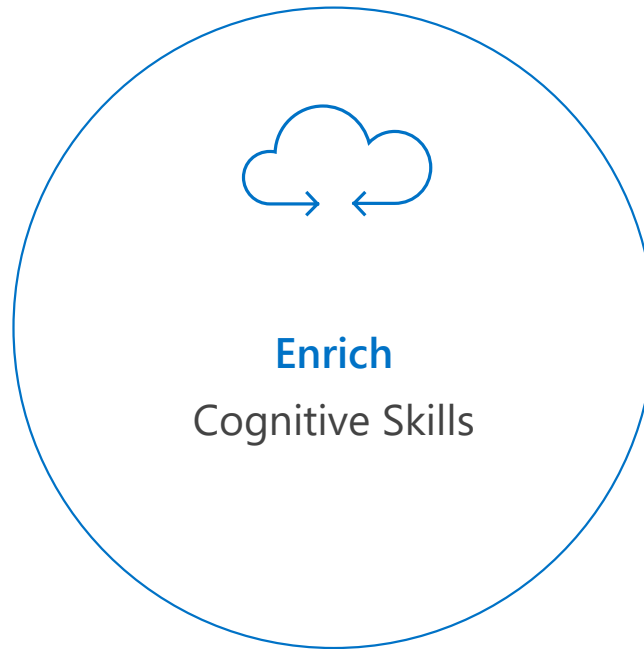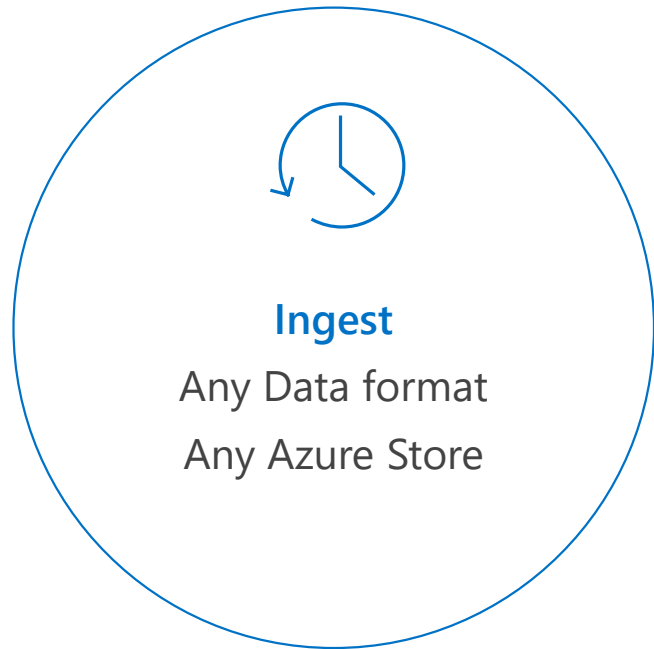
How it works

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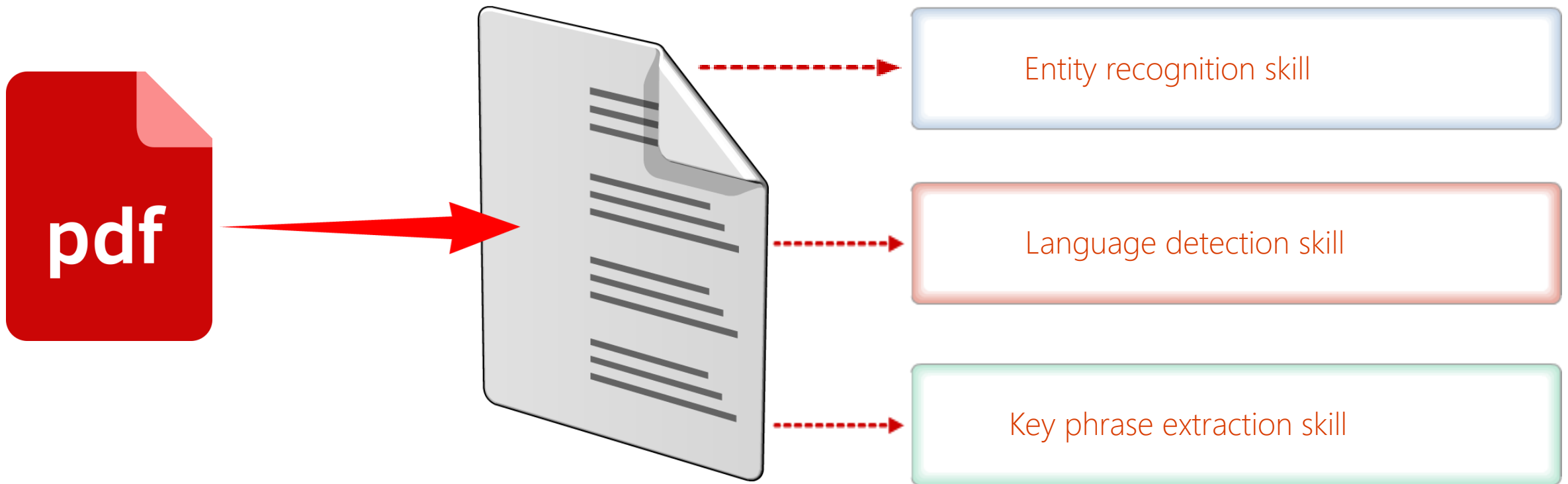
pdf

Entity recognition skill

Language detection skill

Key phrase extraction skill

How it works – Cognitive Search Pipeline

Cognitive Skill

Data Source

Document Cracking → Extensible Enriching Pipeline

Enriched Documents

Search Index

Ingest — Enrich — Explore

How it works – Cognitive Search Pipeline

How it works – Cognitive Skills

Ingest

Enrich

Explore

Create:
- Data source
- Skillset
- Index
- Indexer

Run:
- Indexer
  - Ingest
  - Enrich
  - Index
- Execution Status Check
- Results Check

Query:
- Azure Search
  - API
  - Explorer
- .Net SDK
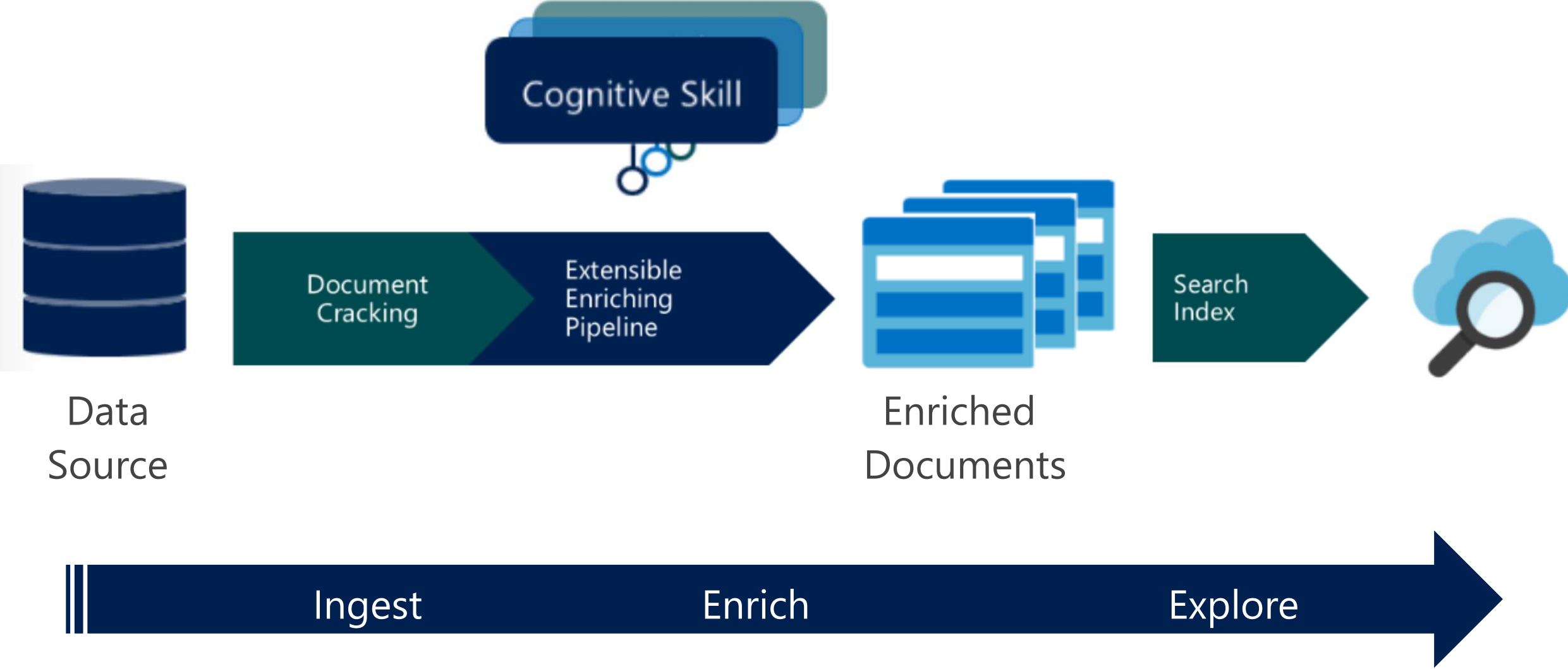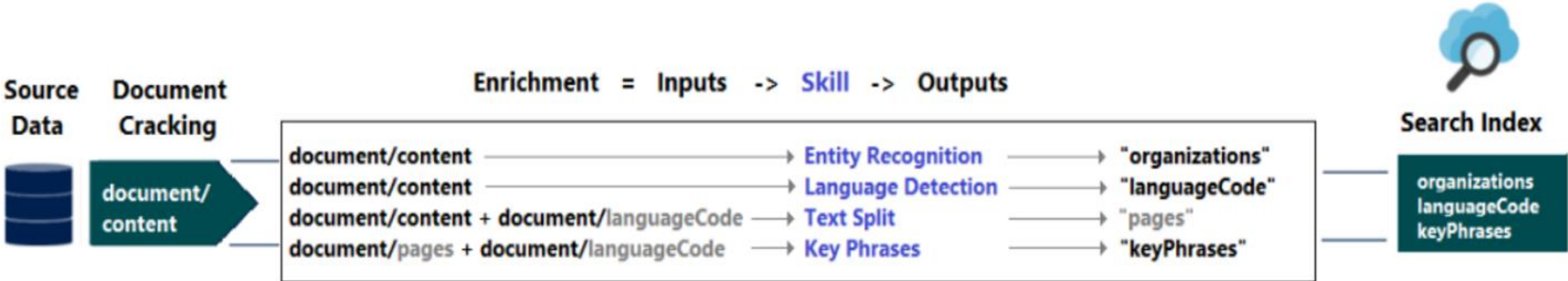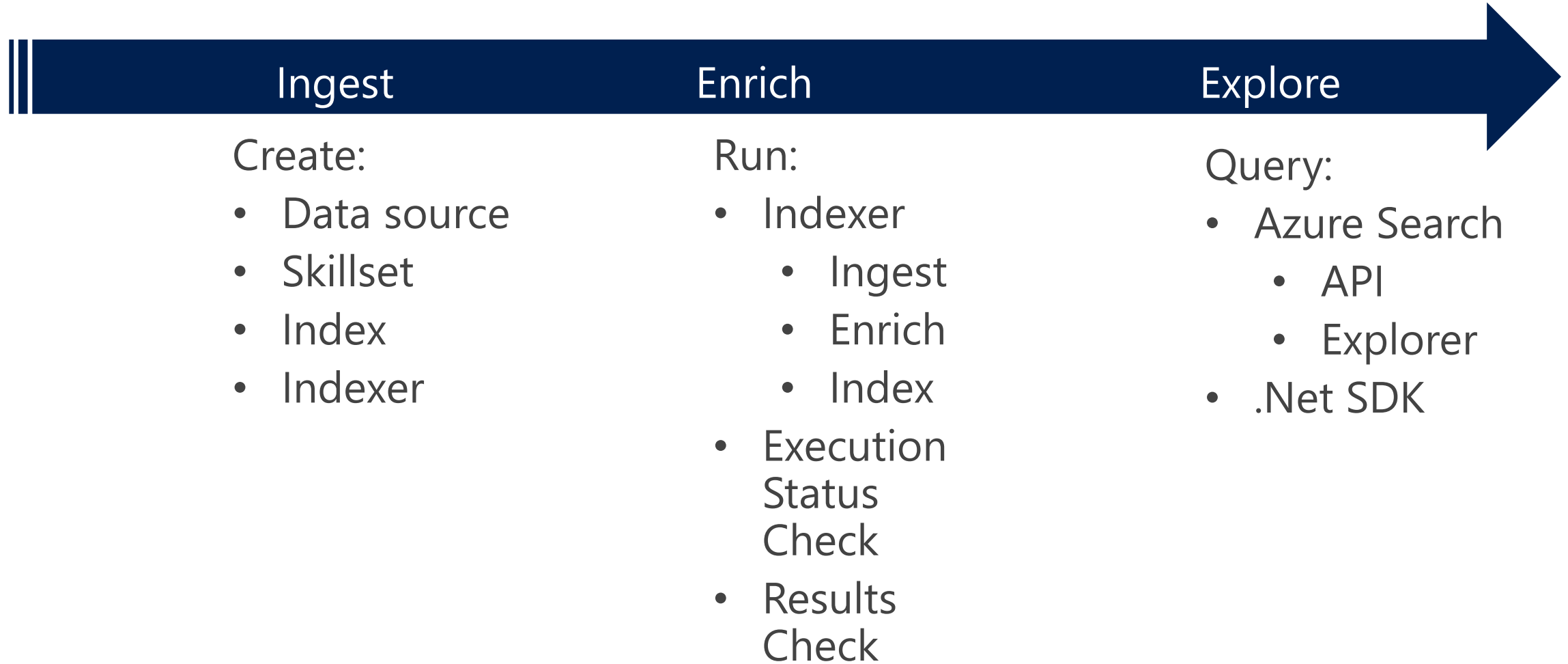
How it works – Old High Level Architecture
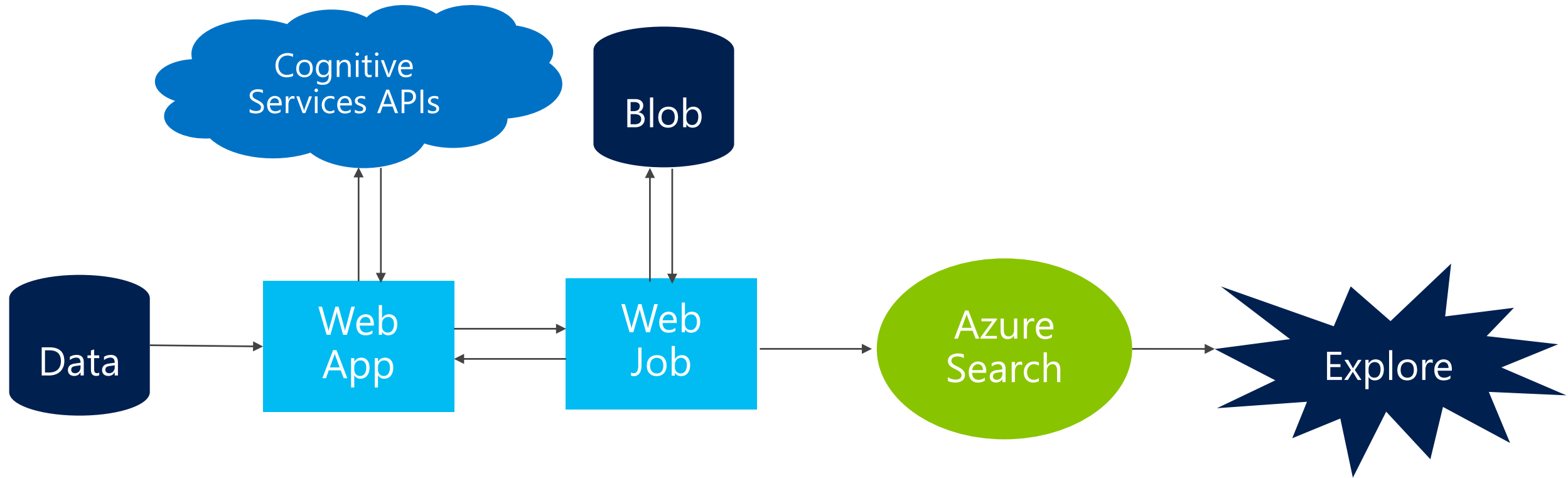
Challenges

- Development
- Costs
- Latency

How it works – New High Level Architecture
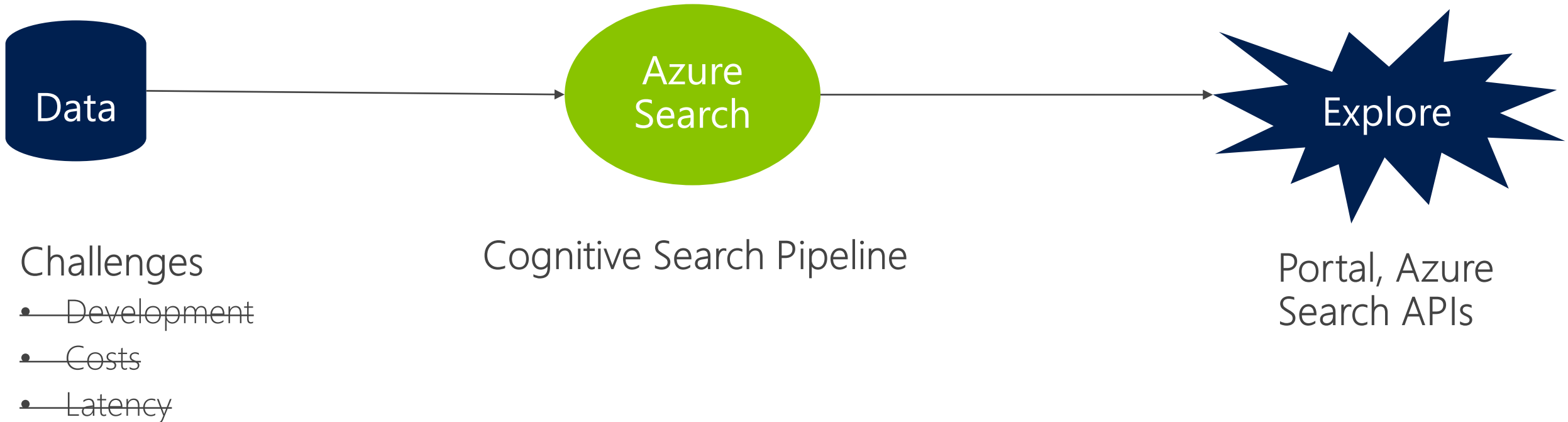
Data → Azure Search → Explore

Challenges
- ~~Development~~
- ~~Costs~~
- ~~Latency~~

Cognitive Search Pipeline

Portal, Azure Search APIs

Good Practices

Supplement the indexes with:

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

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: Metadata Store! 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 are heavier than text skills
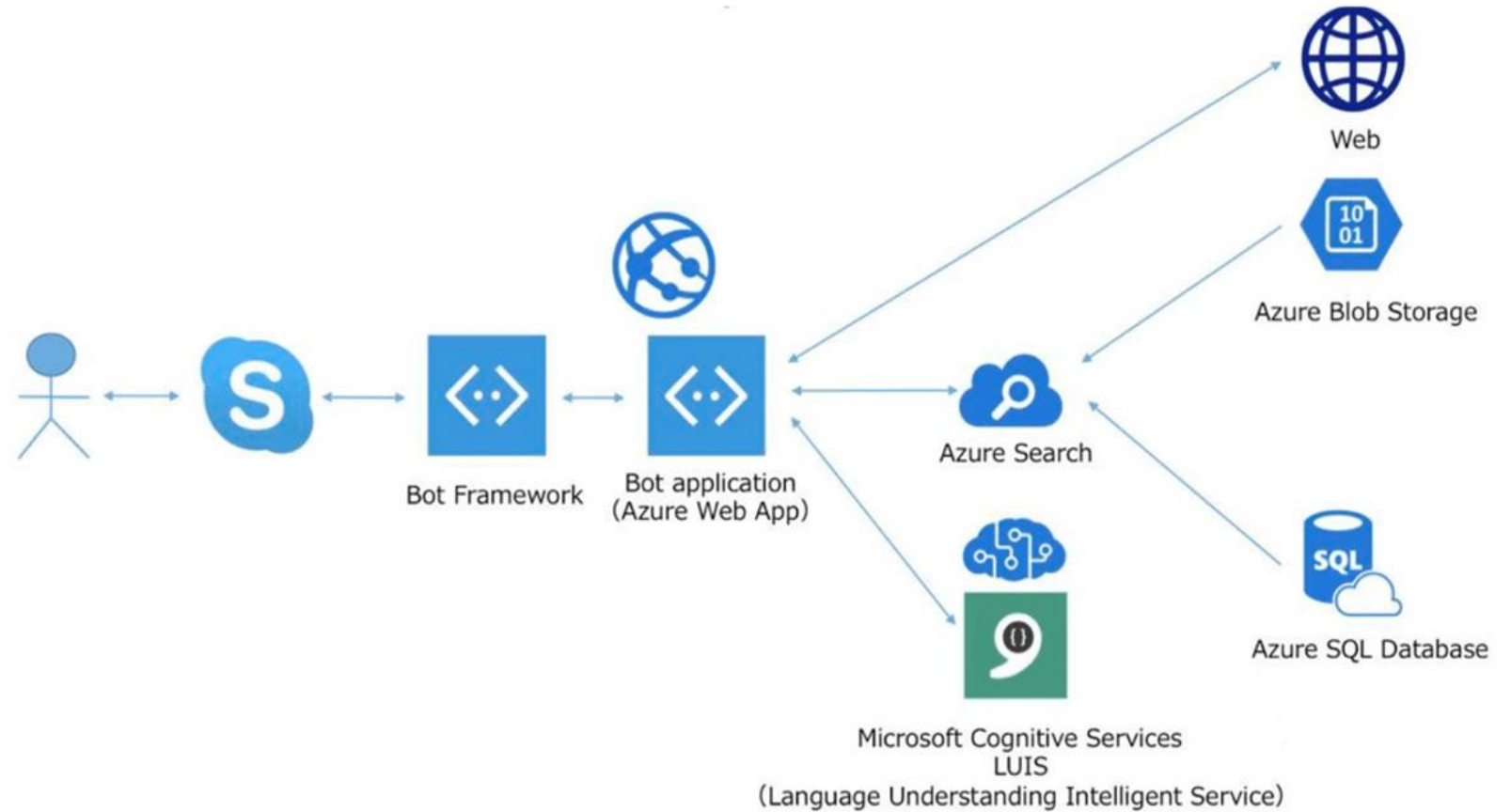
- Parallel Indexing

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

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-Portal

http://aka.ms/LearnAI-GitHub

http://aka.ms/LearnAI-kmb

http://aka.ms/LearnAI-links

LearnAI – Knowledge Mining Bootcamp