Azure Cognitive Search - use pure magic to start finding instead of searching

ADIS JUGO • @adisjugo
KORTO CEO

Microsoft REGIONAL DIRECTOR

MVP Microsoft® Most Valuable Professional

BIWUG

DIWUG
Dutch Information Worker User Group

Adis Jugo

CEO and Co-Founder of KORTO (www.korto.io)
Firmly believes in leading people by trust and example
Microsoft Regional Director
Microsoft MVP Office Apps and Services
Microsoft MVP Azure
Mastermind behind European Collaboration Summit (www.collabsummit.eu)
Mastermind behind European Cloud Summit (www.cloudsummit.eu)
In IT for way too long (first money earned with development back in 1991)
Born in Sarajevo, Bosnia, living in Bingen, Germany
Speaker, author. adisjugo.com

EUROPEAN COLLABORATION SUMMIT 2021

MODERN WORKPLACE | MICROSOFT 365 | TEAMS | SHAREPOINT | POWER PLATFORM

WIESBADEN, GERMANY / JUNE 14-16 2021

EUROPEAN CLOUD SUMMIT 2021

AZURE | ARCHITECTURE | DEVELOPMENT | INFRASTRUCTURE | MANAGEMENT | SECURITY

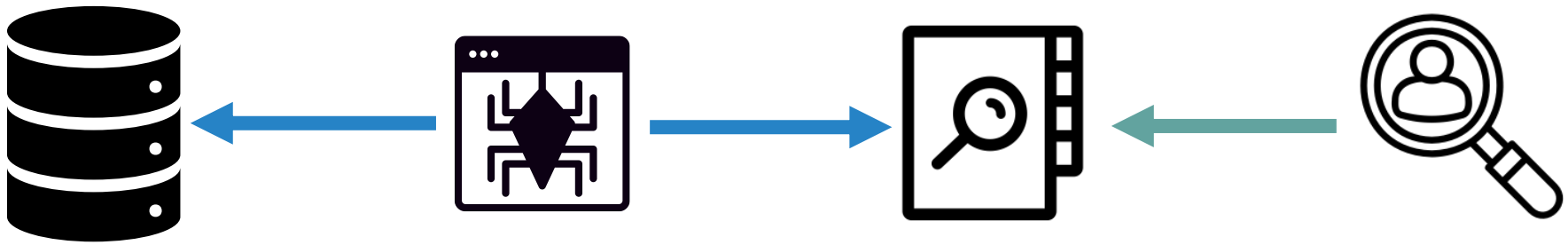
MAINZ, GERMANY / SEPTEMBER 27-29 2021

DÜSSELDORF / GERMANY, 29 NOV 29 – 01 DEC 2021

www.collabsummit.eu

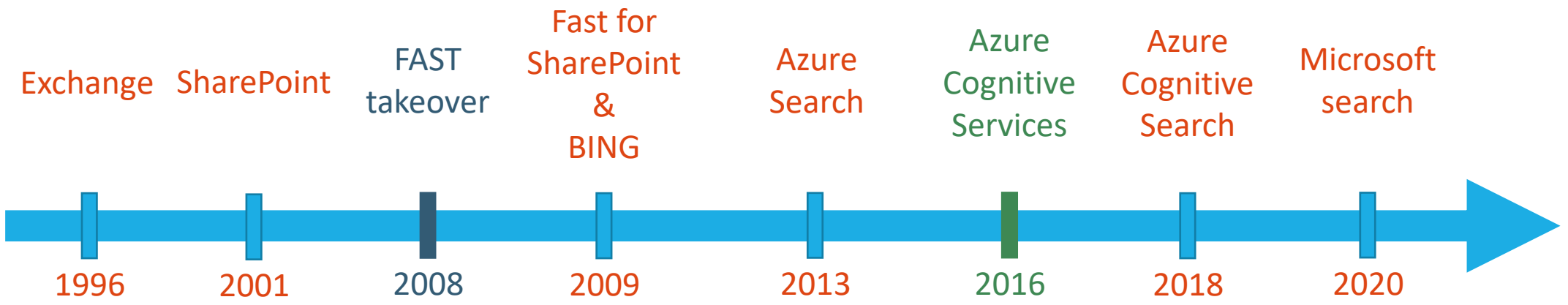
www.cloudsummit.eu

What is search?



KORTO

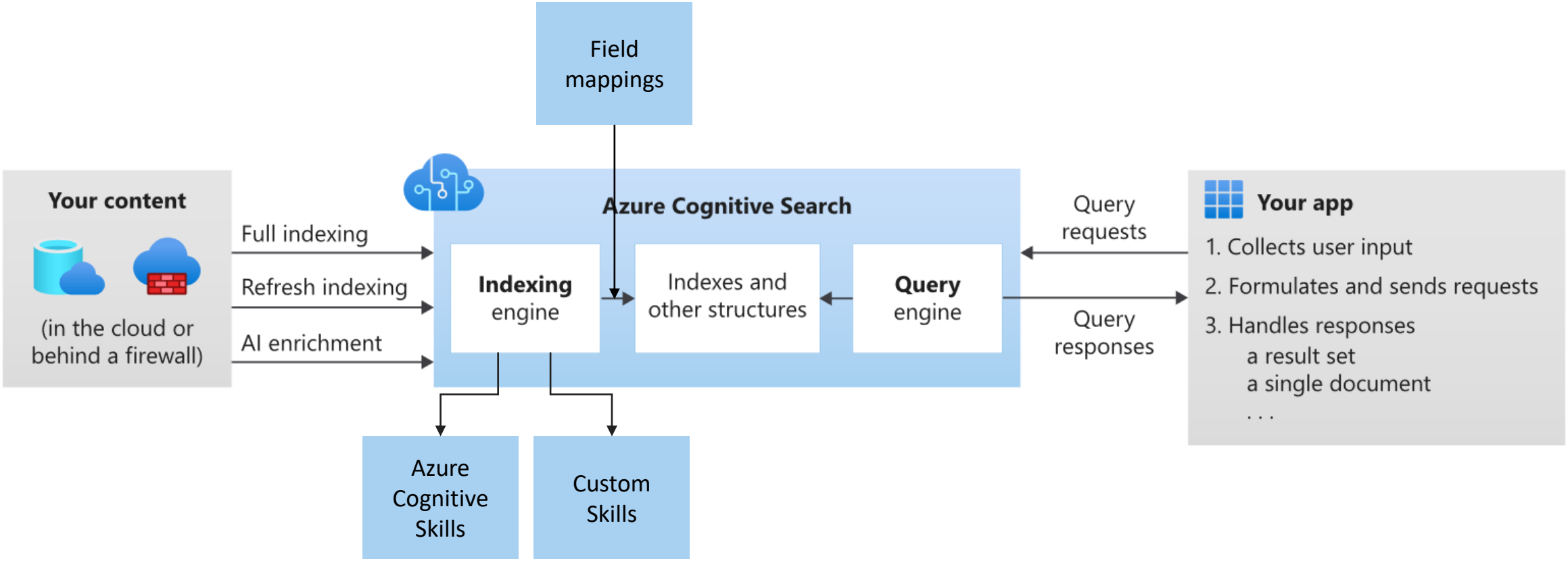
Microsoft and search



KORTO

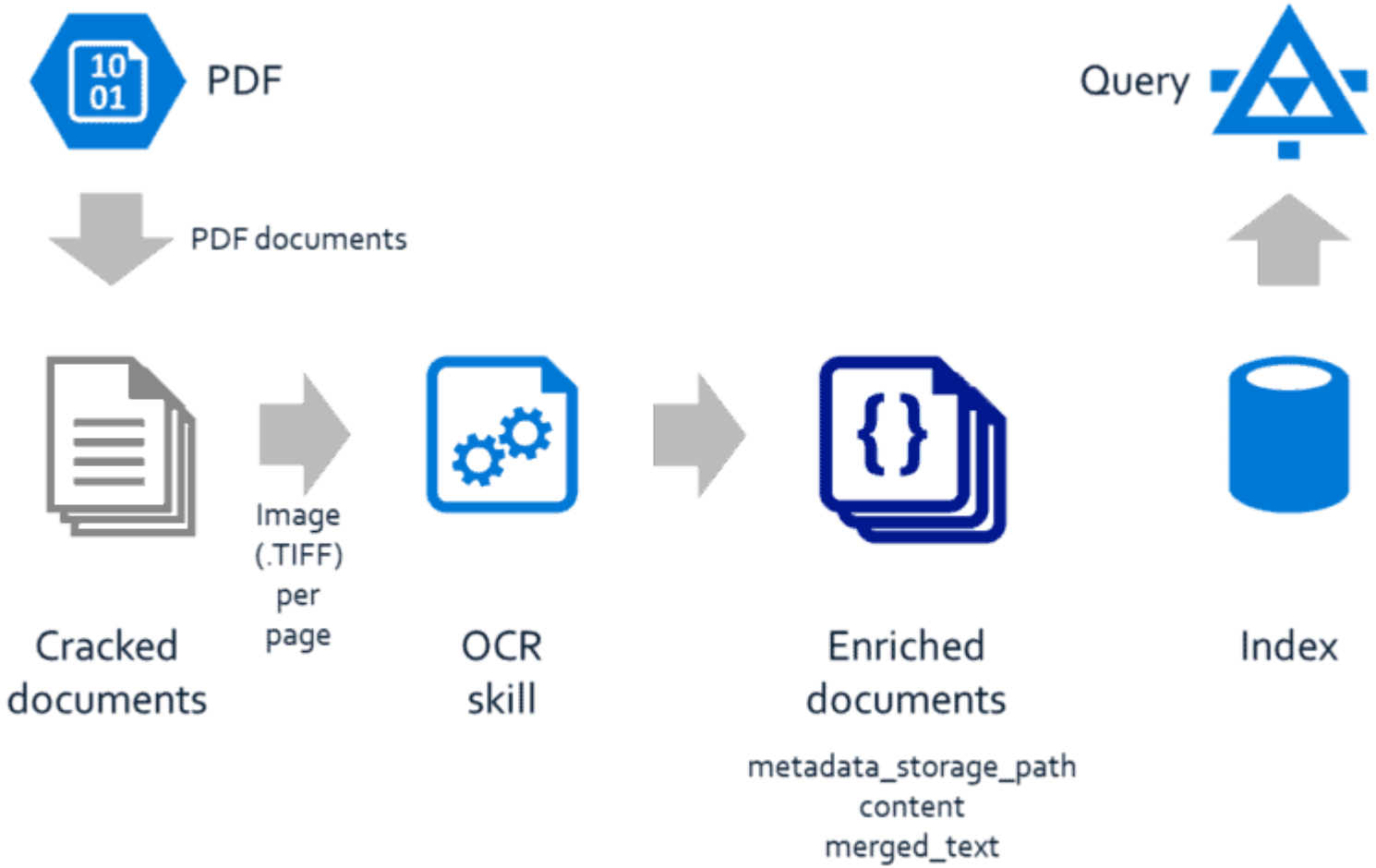
DEMO: Set up Azure Cognitive Services Application

Azure Cognitive Search – Mechanics



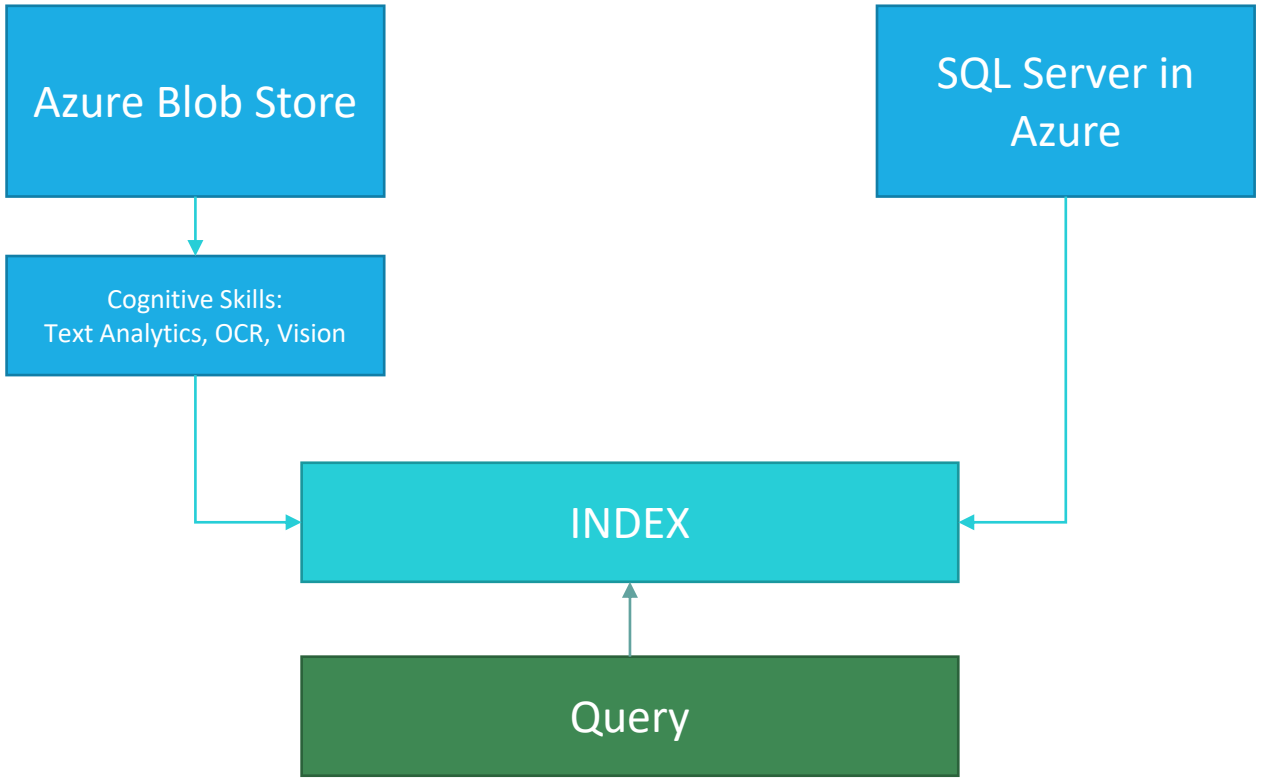
KORTO

Azure Cognitive Search - example



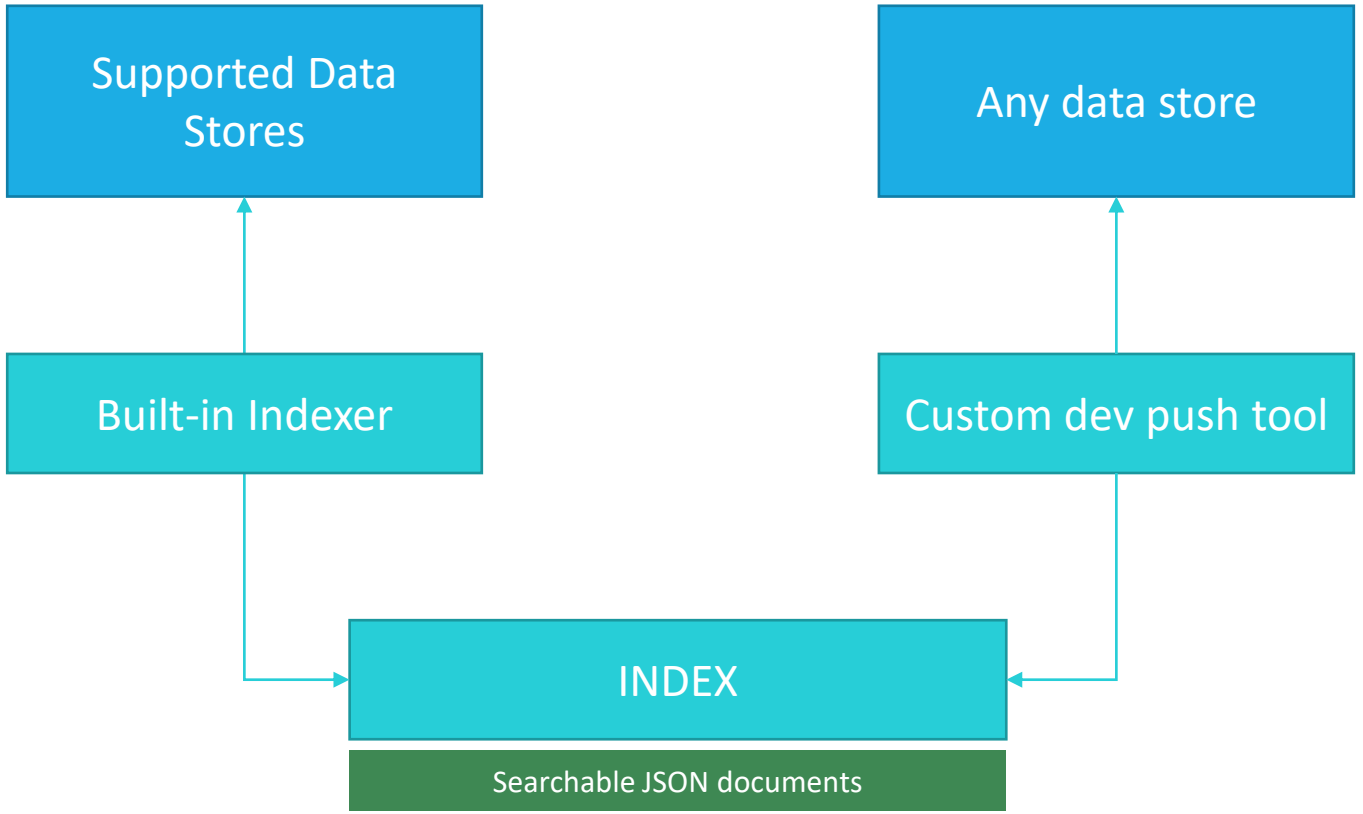
KORTO

KORTO: Data and index structure



KORTO

Data for crunching



KORTO

Supported Data Sources

- Azure Blob Storage
- Azure Data Lake Storage Gen 2
- Azure SQL Database
- Azure Table Storage
- Azure Cosmos DB
- SharePoint Online (Preview)

KORTO

DEMO: Creating a data source through code

KORTO

LET'S TALK ABOUT INDEXES

Index…

…is a collection of JSON documents

…document can be hierarchical (simple types, collections, complex types)

…is independent of data sources

…has a strictly defined schema with fields

…has to be populated

…must have a key, and it must be a string

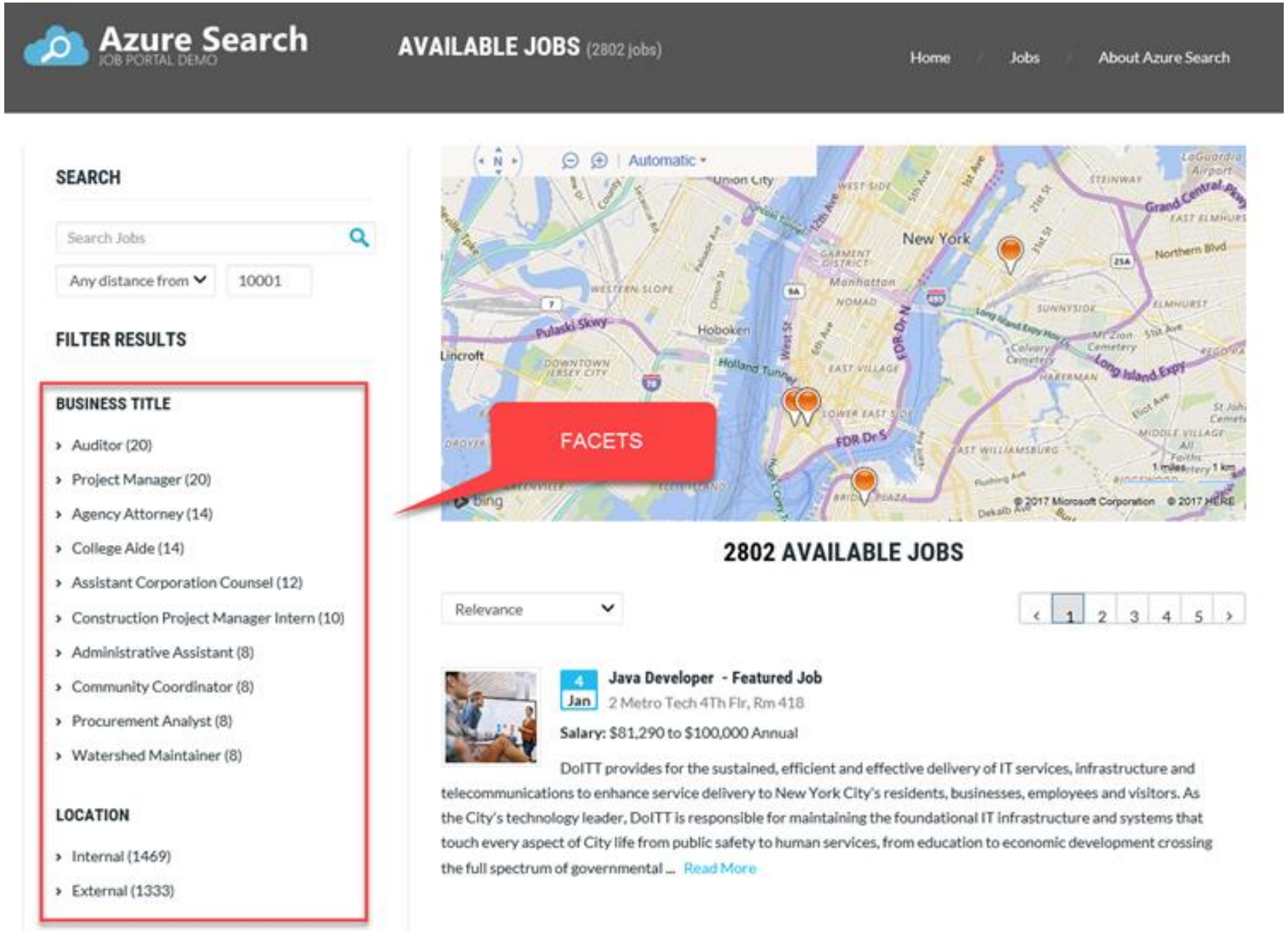
…can be protected using CORS

KORTO

Index fields

- …must have a type
- …can be Retrievable
- …can be Filterable
- …can be Sortable
- …can be Facetable
- …can be searchable
- …can be assigned analyzers

KORTO


Azure Search
JOB PORTAL DEMO
AVAILABLE JOBS (2802 jobs)
Home
Jobs
About Azure Search
SEARCH
Search Jobs
Any distance from
10001
FILTER RESULTS
BUSINESS TITLE
Auditor (20)
Project Manager (20)
Agency Attorney (14)
College Aide (14)
Assistant Corporation Counsel (12)
Construction Project Manager Intern (10)
Administrative Assistant (8)
Community Coordinator (8)
Procurement Analyst (8)
Watershed Maintainer (8)
LOCATION
Internal (1469)
External (1333)
FACETS
Automatic
New York
Manhattan
Hoboken
2802 AVAILABLE JOBS
Relevance
1 2 3 4 5
4 Jan
Java Developer - Featured Job
2 Metro Tech 4Th Flr, Rm 418
Salary: $81,290 to $100,000 Annual
DoITT provides for the sustained, efficient and effective delivery of IT services, infrastructure and telecommunications to enhance service delivery to New York City's residents, businesses, employees and visitors. As the City's technology leader, DoITT is responsible for maintaining the foundational IT infrastructure and systems that touch every aspect of City life from public safety to human services, from education to economic development crossing the full spectrum of governmental ... Read More

Index field types

Edm.String	Collection(Edm.String)
Edm.Boolean	Collection(Edm.Boolean)
Edm.Int32	Collection(Edm.Int32)
Edm.Int64	Collection(Edm.Int64)
Edm.Double	Collection(Edm.Double)
Edm.DateTimeOffset	Collection(Edm.DateTimeOffset)
Edm.GeographyPoint	Collection(Edm.GeographyPoint)
Edm.ComplexType	Collection(Edm.ComplexType)

KORTO

STOP THE COUNT!

- Custom Scoring Profiles for Indexes
- Giving more importance to certain fields
- Field Weights and Functions

"freshness"	Boosts by values in a datetime field (Edm.DataTimeOffset). This function has a boostingDuration attribute so that you can specify a value representing a timespan over which boosting occurs.
"magnitude"	Boosts based on how high or low a numeric value is. Scenarios that call for this function include boosting by profit margin, highest price, lowest price, or a count of downloads. This function can only be used with Edm.Double and Edm.Int fields. For the magnitude function, you can reverse the range, high to low, if you want the inverse pattern (for example, to boost lower-priced items more than higher-priced items). Given a range of prices from $100 to $1, you would set "boostingRangeStart" at 100 and "boostingRangeEnd" at 1 to boost the lower-priced items.
"distance"	Boosts by proximity or geographic location. This function can only be used with Edm.GeographyPoint fields.
"tag"	Boosts by tags that are common to both search documents and query strings. Tags are provided in a tagsParameter. This function can only be used with Edm.String and Collection(Edm.String) fields.

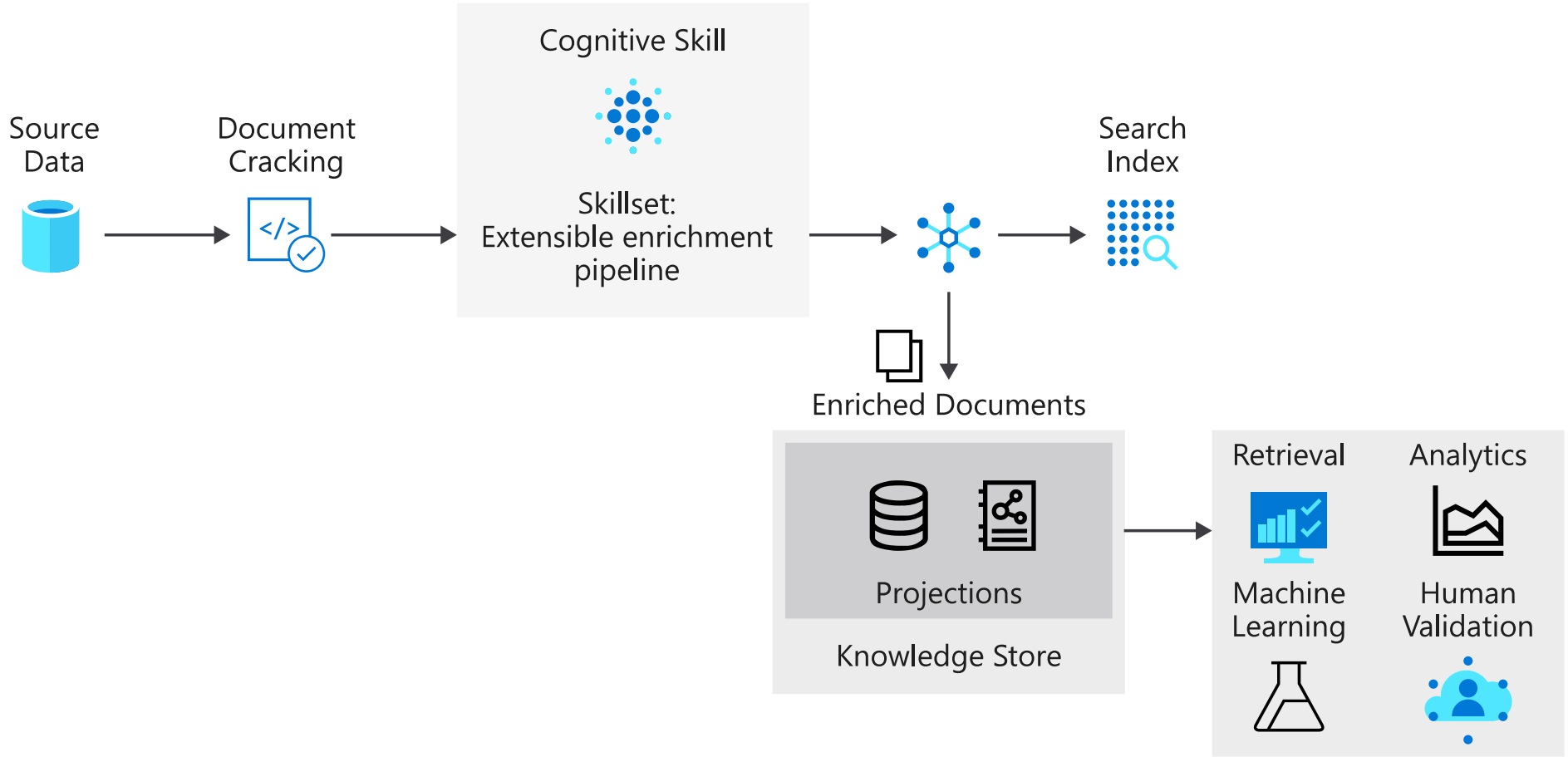
KORTO

DEMO: Creating an index through code

KORTO

YOU'VE GOT SKILLS

You've got skills....



KORTO

Skills have

- …a type
- …a context
- …inputs and outputs that are often chained together

KORTO

An enriched document…

… is temporary tree-like data structure

… is created during skillset execution

… collects all of the changes introduced through skills

… represents changes them in a hierarchy of addressable nodes

… exists for the duration of skillset execution

… can be cached or persisted to a knowledge store

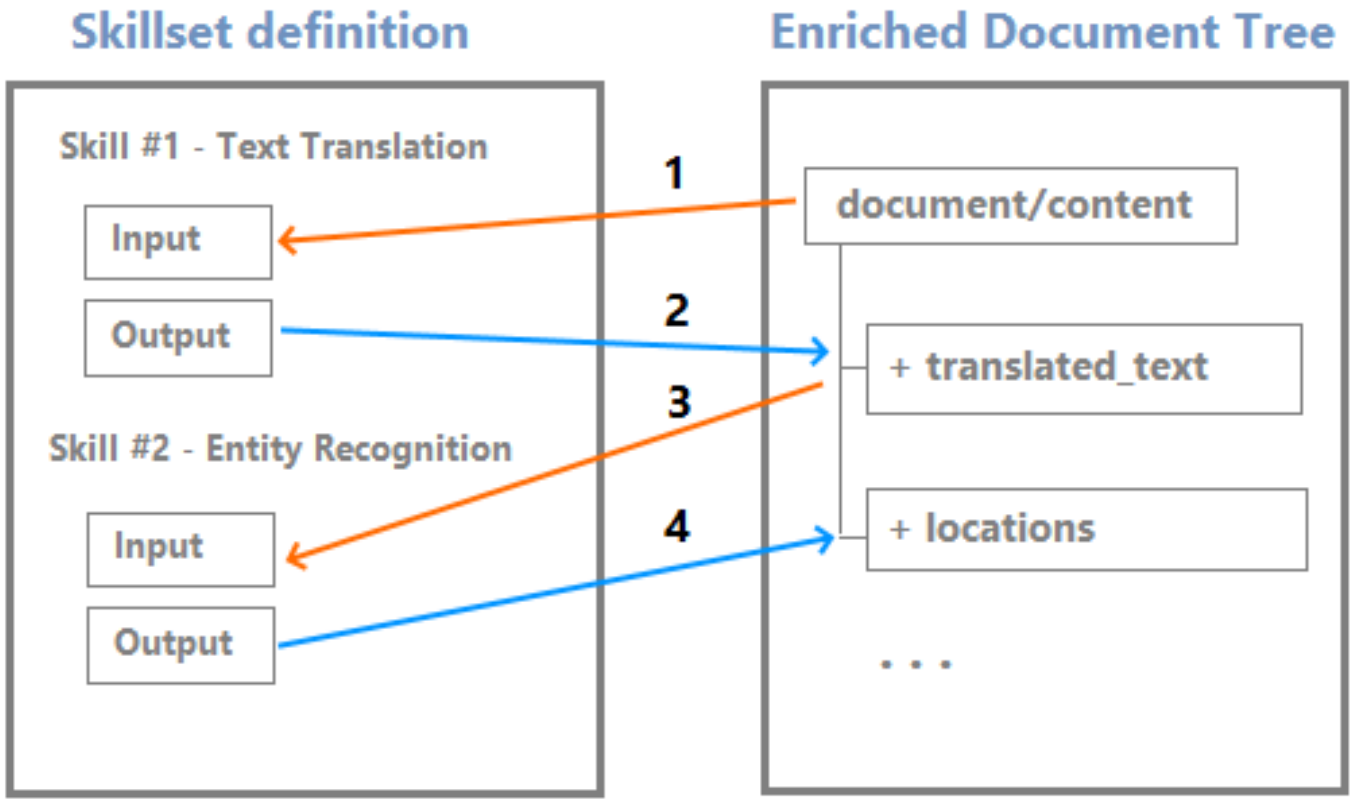
KORTO

Enrichment Tree

Data Source\Parsing Mode	Default	JSON, JSON Lines & CSV
Blob Storage	/document/content /document/normalized_images/* ...	/document/{key1} /document/{key2} ...
Azure SQL	/document/{column1} /document/{column2} ...	N/A
Cosmos DB	/document/{key1} /document/{key2}	N/A

KORTO

Skillset and enriched document example



KORTO

What skills are available?

Text

Detecting language

Understanding context, detecting sentiment

Recognizing People

Recognizing Organizations

Recognizing Locations

Extracting keywords / tags

Identifying Personally Identifiable Information

Translations

- Image
- OCR
- Extracting keywords / tags
- Creating captions
- Identifying celebrities

Heavily depends on Azure Region and Language

KORTO

Wait, that’s not enough!

Really?

No big deal: build a custom skill!

It‘s just a Web Api.

But this would require a separate session.

KORTO

The story of the knowledge store

Wait, once the AI has crunched and enriched the document, those enrichments are gone as soon as the document has been stored into index?

Yes. And no...

...store enriched document structures

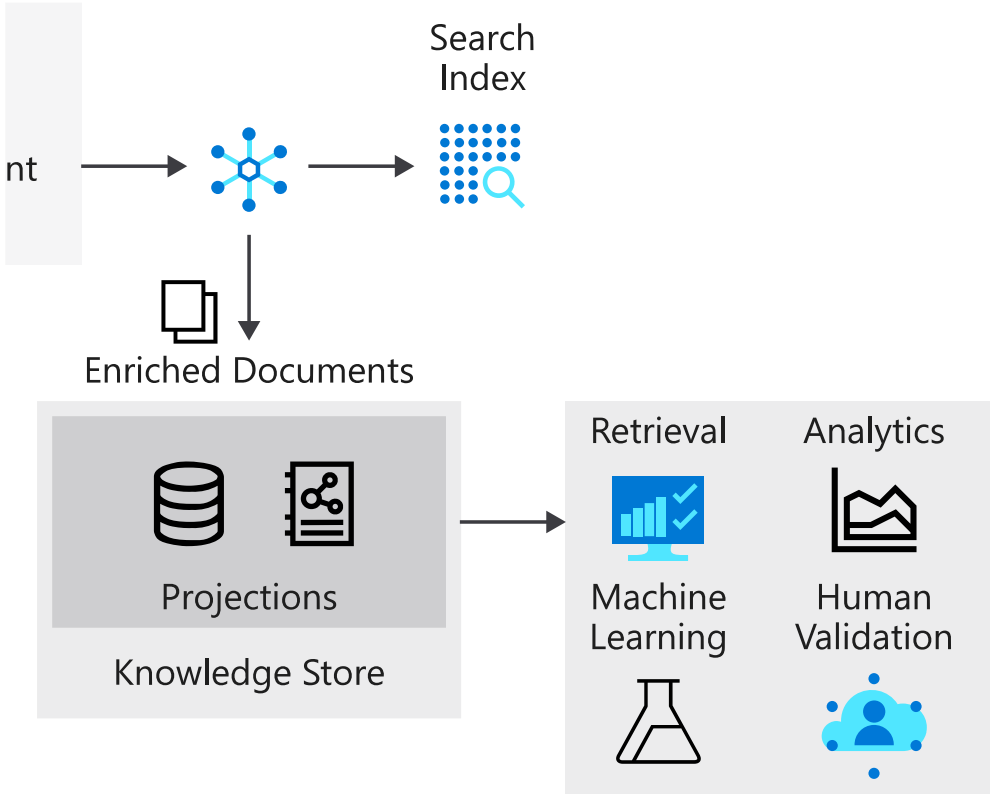
... either in Table Store or Blob Store

...do human validation

...improve enrichments

...run machine learning processes over that

...run analytics reports and create dashboards



Our use case

Get document from blob store

Extract text, OCR-ed text, and content of the images (if any embedded)

Recognize

- persons
- organizations
- locations
- key phrases
- entities
- image captions
- celebrities

KORTO

DEMO: Creating a skillset through code

KORTO

SHOWTIME: INDEXERS

Indexer…

… is a crone job (daemon) that runs in Azure

… crawls the data from ONE data source

… determines the difference (incremental crawl)

… enriches the data with the cognitive skills

… enriches the data with custom enrichments

… maps the fields between source and target (JSON document defined by index)

… stores the document into Index

KORTO

Field mappings

- Standard (simple) mappings
- Mapping fields enriched by cognitive services
- Transformations using Field Mapping Functions

Field Mapping Functions
base64Encode
base64Decode
extractTokenAtPosition
jsonArrayToStringCollection
urlEncode
urlDecode

KORTO

DEMO: Create indexer through code

KORTO

LET'S FIND!

Analysers – Lucene

https://lucene.apache.org/core/6_6_1/queryparser/org/apache/lucene/queryparser/classic/package-summary.html

syntax for specialized query forms:

... wildcard

... fuzzy search

... proximity search

... regular expressions

KORTO

Boolean operators

Operator	Character	Example	Usage
AND	&, +	wifi + luxury	Specifies terms that a match must contain. In the example, the query engine will look for documents containing both wifi and luxury. The plus character (+) is used for required terms. For example, +wifi +luxury stipulates that both terms must appear somewhere in the field of a single document.
OR	|	wifi | luxury	Finds a match when either term is found. In the example, the query engine will return match on documents containing either wifi or luxury or both. Because OR is the default conjunction operator, you could also leave it out, such that wifi luxury is the equivalent of wifi | luxury.
NOT	!, -	wifi –luxury	Returns matches on documents that exclude the term. For example, wifi –luxury will search for documents that have the wifi term but not luxury. The searchMode parameter on a query request controls whether a term with the NOT operator is ANDed or ORed with other terms in the query (assuming there is no + or | operator on the other terms). Valid values include any or all. searchMode=any increases the recall of queries by including more results, and by default - will be interpreted as "OR NOT". For example, wifi -luxury will match documents that either contain the term wifi or those that do not contain the term luxury. searchMode=all increases the precision of queries by including fewer results, and by default - will be interpreted as "AND NOT". For example, wifi -luxury will match documents that contain the term wifi and do not contain the term "luxury". This is arguably a more intuitive behavior for the - operator. Therefore, you should consider using searchMode=all instead of searchMode=any if you want to optimize searches for precision instead of recall, and Your users frequently use the - operator in searches. When deciding on a searchMode setting, consider the user interaction patterns for queries in various applications. Users who are searching for information are more likely to include an operator in a query, as opposed to e-commerce sites that have more built-in navigation structures.

KORTO

Search examples

```
category: budget AND \"Recently renovated\"^3"
```

```
artists:("Miles Davis" "John Coltrane")
```

```
genre: jazz NOT country
```

Fuzzy search:
"blue~" or "blue~1" would return "blue", "blues", and "glue"

Proximity search:
"hotel airport"~5 will find the terms "hotel" and "airport" within 5 words

Regular expressions:
/[mh]otel/ will find both "motel" or "hotel"

KORTO

Wildcard search

You can use generally recognized syntax for multiple (*) or single (?) character wildcard searches. Full Lucene syntax supports prefix, infix, and suffix matching.

Note the Lucene query parser supports the use of these symbols with a single term, and not a phrase.

Affix type	Description and examples
prefix	Term fragment comes before * or ?. For example, a query expression of search=alpha* returns "alphanumeric" or "alphabetical". Prefix matching is supported in both simple and full syntax.
suffix	Term fragment comes after * or ?, with a forward slash to delimit the construct. For example, search=/.*numeric./ returns "alphanumeric".
infix	Term fragments enclose * or ?. For example, search=/.non*al./returns "nonnumerical" and "nonsensical".

KORTO

Contact

Adis Jugo

Germany
Deputy CEO
@adisjugo

Microsoft MVP
Microsoft Regional Director

KORTO

A Südwind Company
Member of Insa Group

For additional questions contact us at
info@korto.io
www.korto.io

KORTO