Adding AI Cloud Services to Your On-Prem Data Workflows for NLP & Content Enrichment

Daniel Wrigley, SHI GmbH

November 25, 2020

Unfortunately not in Vilnius ☺

Search | Analytics | Big Data

Agenda

1. What‘s the problem?
2. How do I get data?
3. How do I store the data?
4. How do I process the data?
5. I want to do ML/AI stuff!
6. What are the challenges?
7. I can just use services?
8. Let me show you!
9. What YOU should take away with you!

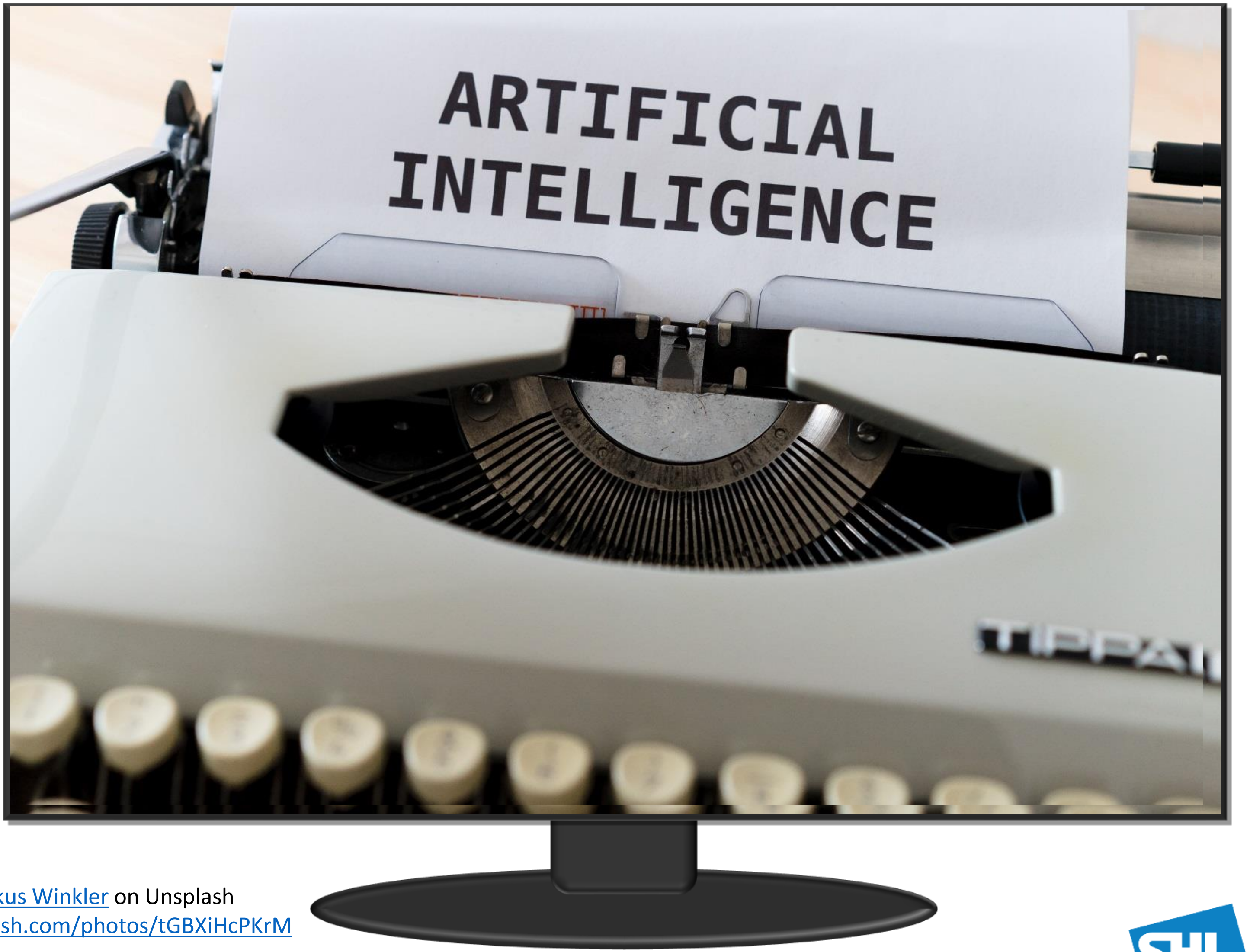


Photo by Markus Winkler on Unsplash
https://unsplash.com/photos/tGBXiHcPKrM

What‘s the problem?

I have a lot of data: Nowadays, everyone (every company) has!

I can store a lot of data: Cost for storage has been decreasing for the last years!

I can process a lot of data: Scalable open source frameworks enable a broad community!

How can I get the data?

- Multi-Tools
  - Apache NiFi
  - StreamSets
  - Apache Flume
  - Vector
- Logs & Metrics
  - Logstash
  - Beats
- Web Crawling
  - Apache Nutch
  - Scrapy
- …

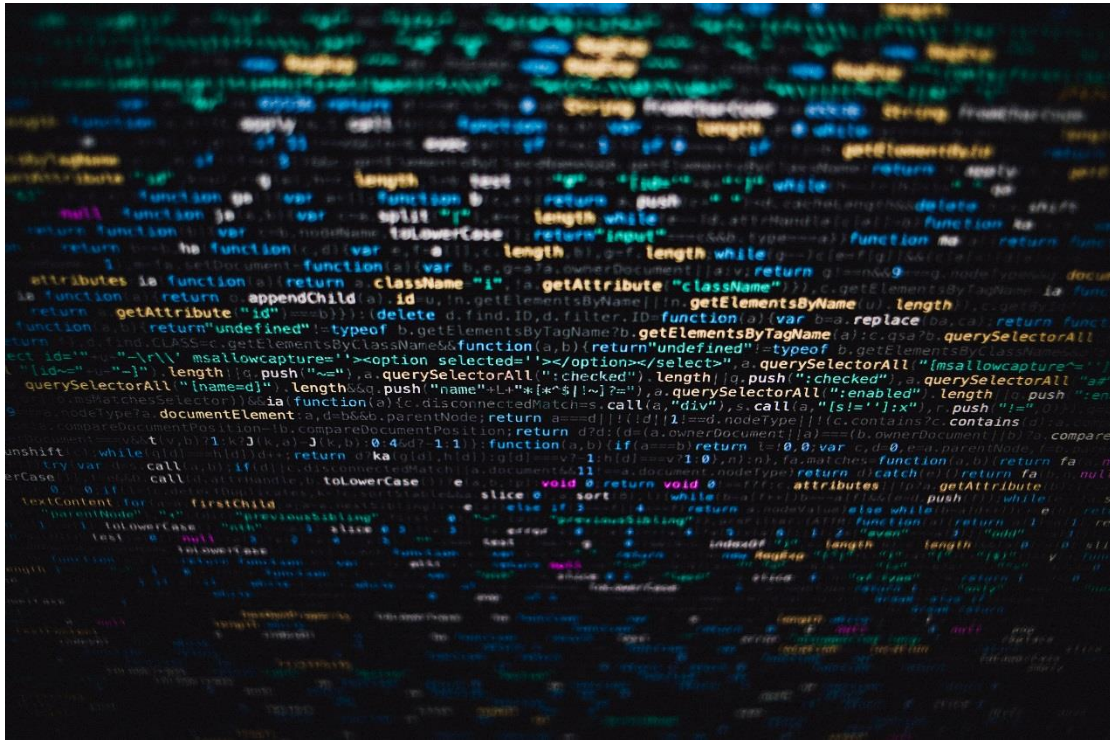


Photo by Markus Spiske on Unsplash
https://unsplash.com/photos/qjnAnF0jIGk

How do I store the data?

- Distributed Filesystems, e.g. HDFS
- NoSQL Datastores
  - Apache Solr
  - Elasticsearch
  - Hive
  - MongoDB
  - …
- Cloud Services
  - AWS S3
  - Google Cloud Storage
  - Azure Storage
  - …
- Private Cloud/Private Data Center

Photo by Denny Müller on Unsplash
https://unsplash.com/photos/1qL31aacAPA

How do I process the data?

- Ingestion Frameworks/ETL Tools
  - Apache NiFi
  - Vector
  - StreamSets
  - Apache Flume
- Apache Spark
- Apache Flink
- Apache Kafka
- …

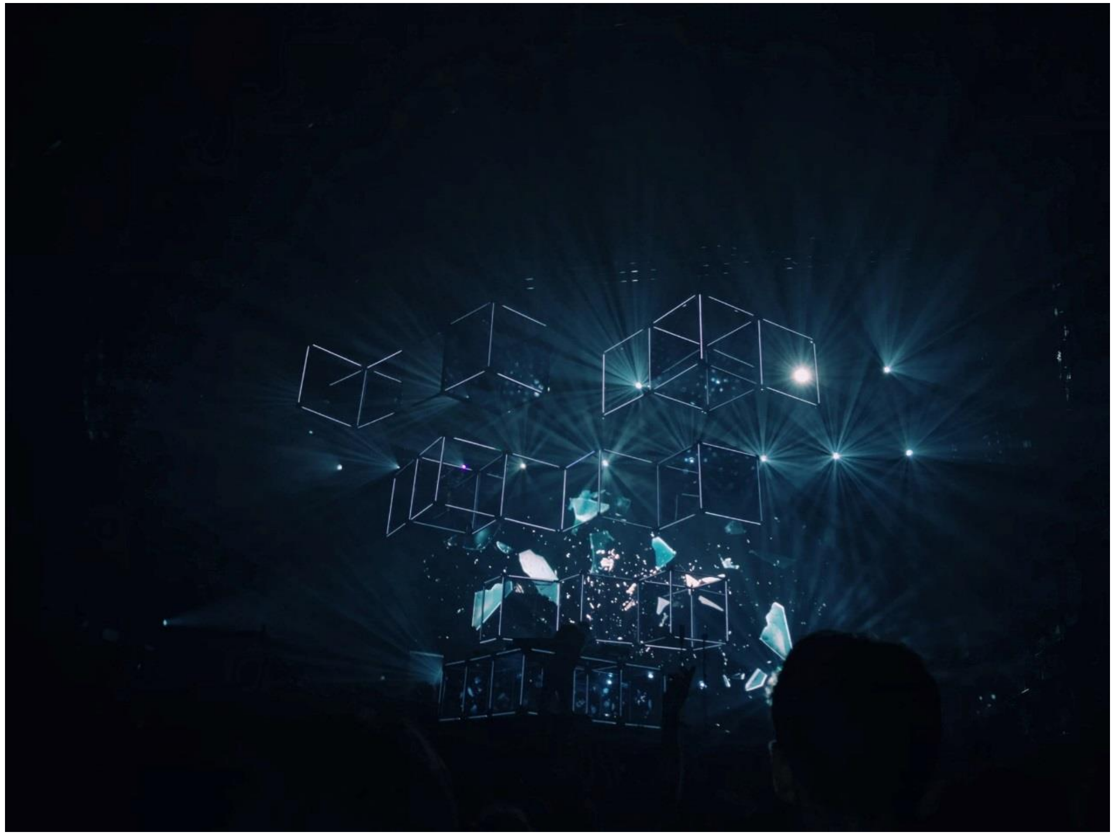
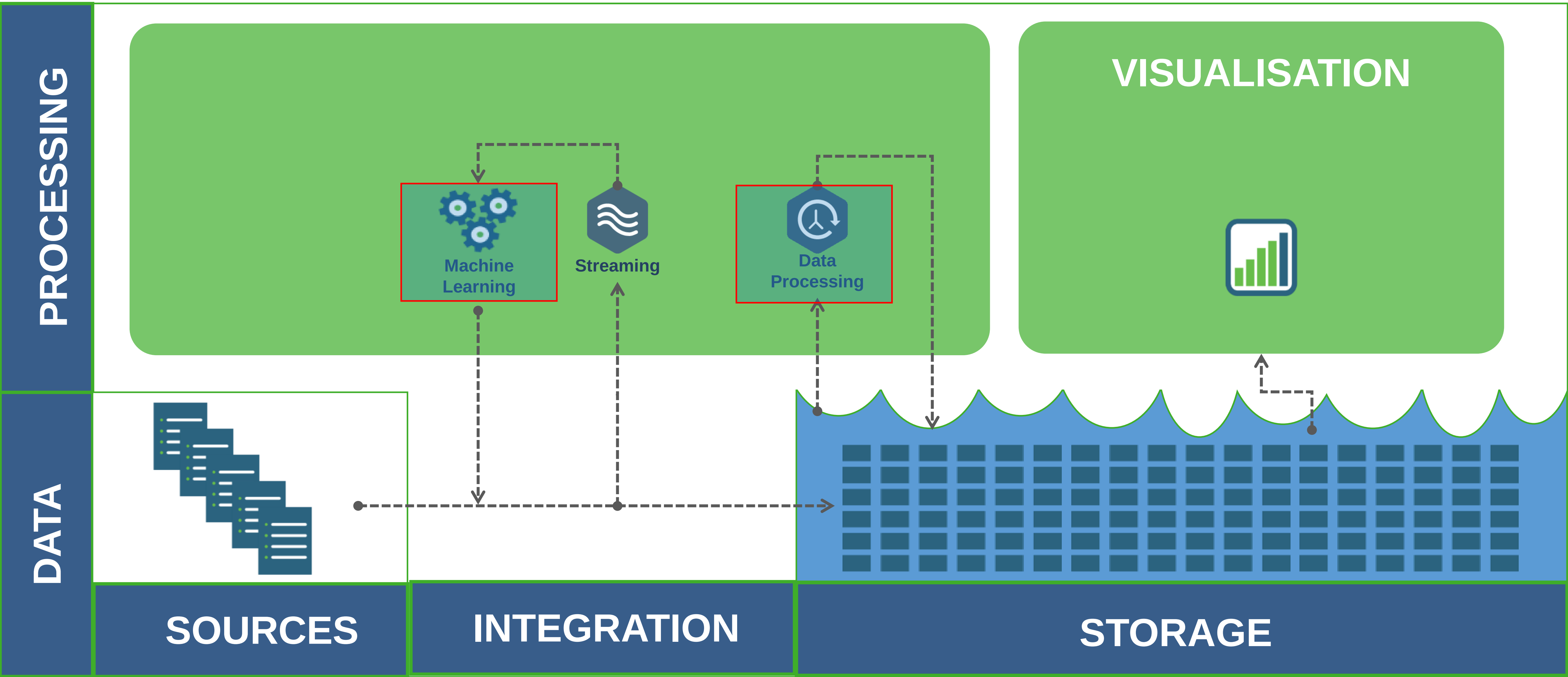


Photo by fabio on Unsplash
https://unsplash.com/photos/oyXis2kALVg

Blueprint Data Processing Architecture



Challenge #1: Lack of Talent

- Working with (unstructured) data requires expertise:
  - Data Engineering
    Extract – Transform – Load
  - Data scientists
    Someone needs to have a deep understanding of the data that we want to work with!
  - Natural language processing:
    We're not dealing with true/false, integers, enums but text!
  - Machine learning engineering
    Which algorithms work for our data and our business/use case?
- Chances are you do not have all of these in your team/company

Challenge #2: Lack of Data

- But we already have a lot of data!?
- For supervised learning you need **labelled data**
- Train a NER Model in Apache OpenNLP: ~15,000 annotated sentences

```
<START:person> Pierre Vinken <END> , 61 years
old , will join the board as a nonexecutive
director Nov. 29 .
```

Photo by Simone Secci on Unsplash
https://unsplash.com/photos/49uySSA678U

Challenge #3: Lack of Resources

- Model training, evaluating, tuning in several iterations can be time-consuming
- Training phases require additional hardware

Photo by Brandon Morgan on Unsplash
https://unsplash.com/photos/3qucB7U2l7I

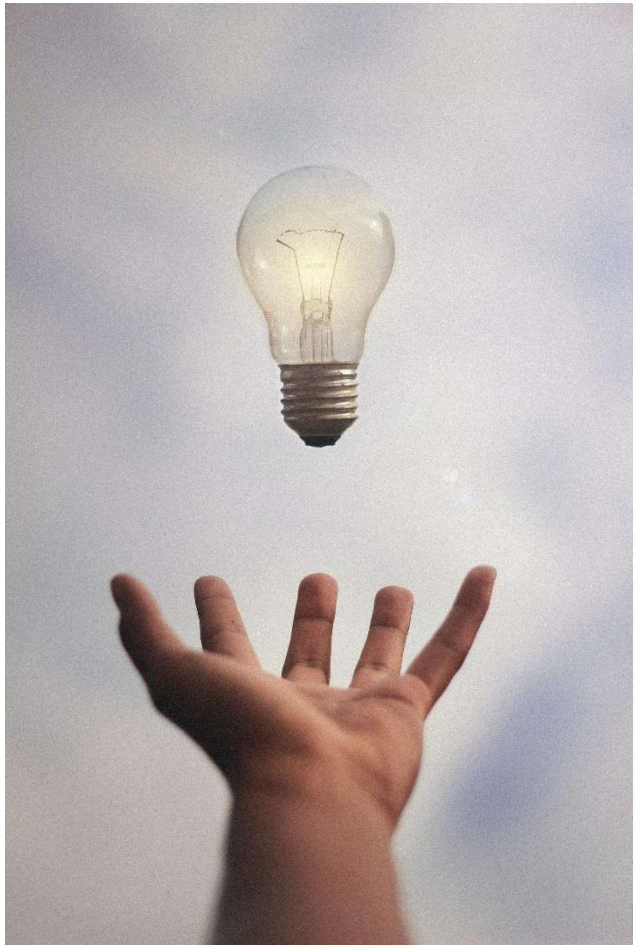
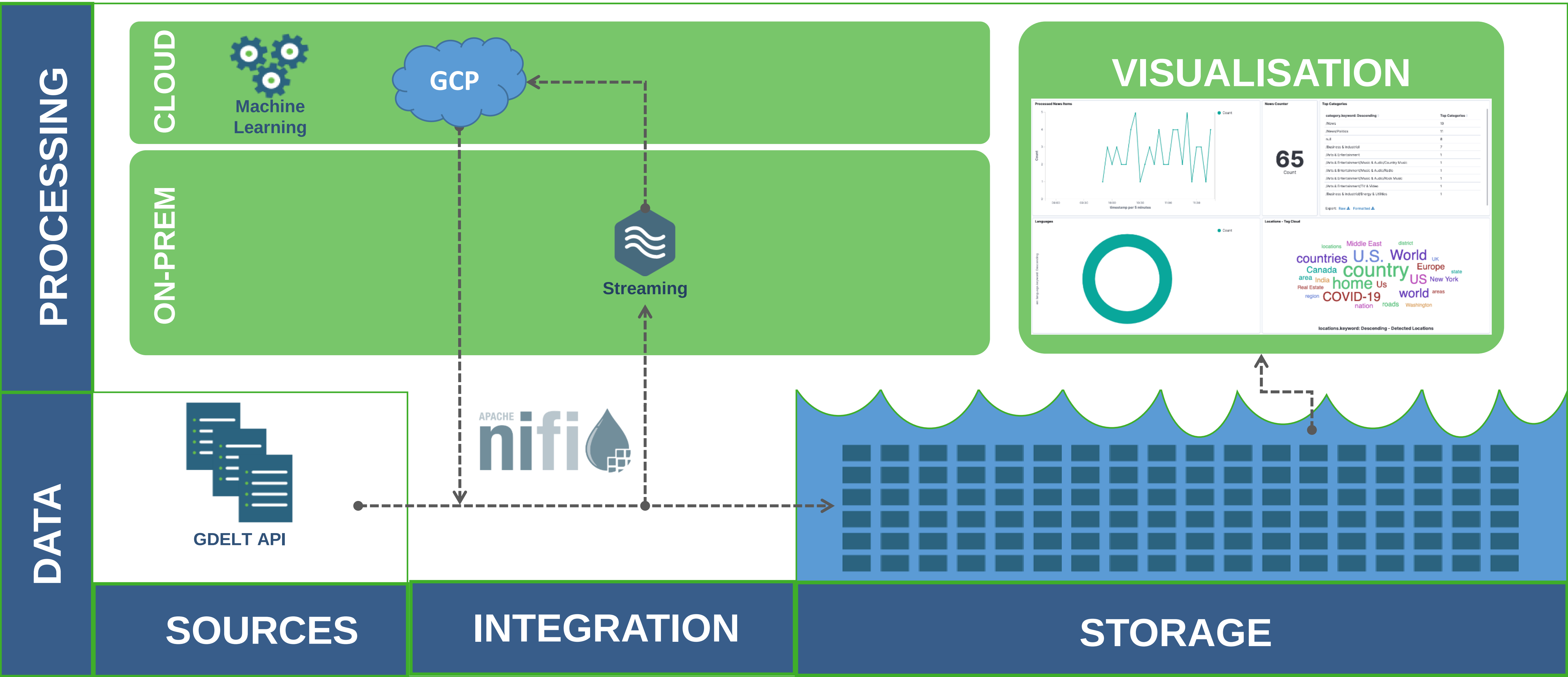


Photo by [Júnior Ferreira](https://unsplash.com/photos/7esRPTt38nI) on Unsplash
https://unsplash.com/photos/7esRPTt38nI

Photo by [C Dustin](https://unsplash.com/photos/K-Iog-Bqf8E) on Unsplash
https://unsplash.com/photos/K-Iog-Bqf8E

Demo – Data Processing Architecture



Top Categories

category.keyword: Descending	Top Categories
/News	19
/News/Politics	11
null	8
/Business & Industrial	7
/Arts & Entertainment	1
/Arts & Entertainment/Music & Audio/Country Music	1
/Arts & Entertainment/Music & Audio/Radio	1
/Arts & Entertainment/Music & Audio/Rock Music	1
/Arts & Entertainment/TV & Video	1
/Business & Industrial/Energy & Utilities	1

Export: Raw Formatted

Demo Setting – Briefly Explained

- Data Source: GDELT – a news monitoring project
- Latest news is requested by NiFi
- News items are extracted and transformed
- Extracted content sent to GCP Natural Language API
- Entities, categories, language are extracted from the response
- Results are sent to Elasticsearch
- Kibana visualizes the results in a dashboard

Now show us already!

Photo by Rob Laughter on Unsplash
https://unsplash.com/photos/WW1jsInXgwM

When should I use Cloud Services?

Pros

- Accelerate Time-to-Market
- Compensate lack of talent
- Less Operational costs
- Less Know-how necessary
- Integrate well in other tools
- Pay-as-you-go
- Rapid Prototyping

Photo by Paul on Flickr
https://www.flickr.com/photos/vegaseddie/5700609302/

When should I NOT use Cloud Services?

Cons

- Black Box: No control
- Generic: Not domain-specific
- High usage → high cost
- Need tooling around (whitelists, blacklists, sanity checks, etc.)

Photo by ApolitikNow on Flickr
https://www.flickr.com/photos/92457334@N04/50119448558/

Key Takeaways

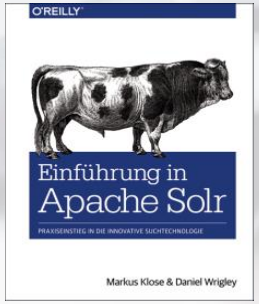
- A lot of technologies, frameworks, services for processing data exist
- Some parts of data-driven projects are easier than others
- In some cases, using or starting with a service makes sense
  - Faster time-to-market
  - No huge team of experts necessary
- For sophisticated use cases: Build your own
  - Gain control
  - Domain knowledge

Resources

- The GDELT Project: https://www.gdeltproject.org/
- Apache NiFi: https://nifi.apache.org/
- Google Cloud Natural Language API: https://cloud.google.com/natural-language
- Apache OpenNLP: https://opennlp.apache.org/
- spaCy: https://spacy.io/
- Hello-NLP: https://github.com/o19s/hello-nlp

Daniel Wrigley

Lead Consultant
Search & Analytics



O'REILLY

Einführung in Apache Solr

PRAXISEINSTIEG IN DIE INNOVATIVE SUCHTECHNOLOGIE

Markus Klose & Daniel Wrigley

daniel.wrigley@shi-gmbh.com

Thank you!
See you next year!
Hopefully in Vilnius!