Otto-Friedrich-Universität Bamberg

Keynote:
First Thoughts on a Data Lake Architecture for an Open Search Infrastructure

Leon Martin, M.Sc.

Prof. Dr. Andreas Henrich

University of Bamberg
Media Informatics

https://www.uni-bamberg.de/minf/

Outline

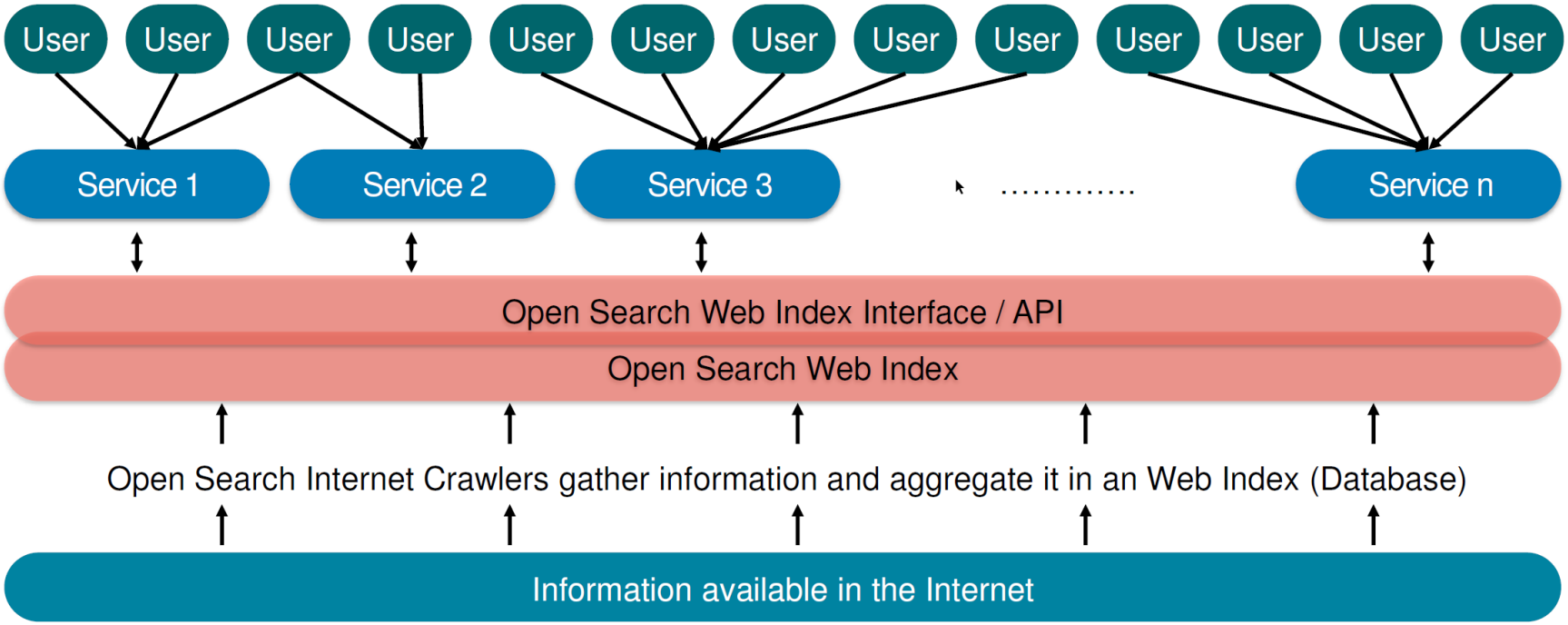
1. Motivation
2. From Data Warehouses to Data Lakes
3. First Thoughts on an Architecture
4. Next steps & Conclusion

Outline

1. Motivation
2. From Data Warehouses to Data Lakes
3. First Thoughts on an Architecture
4. Next steps & Conclusion

An European Open Search Infrastructure

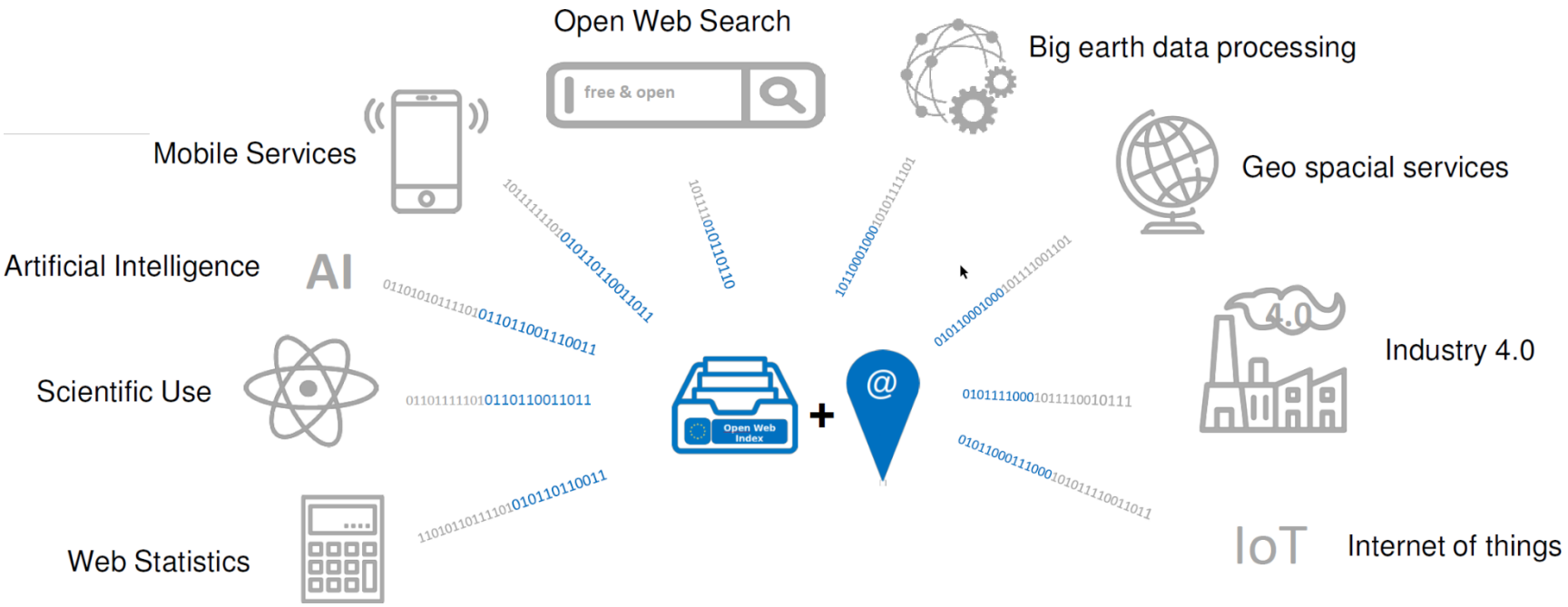
An open and distributed Internet search in Europe bases on an open search ecosystem – The Open Search Web Index



https://opensearchfoundation.org/

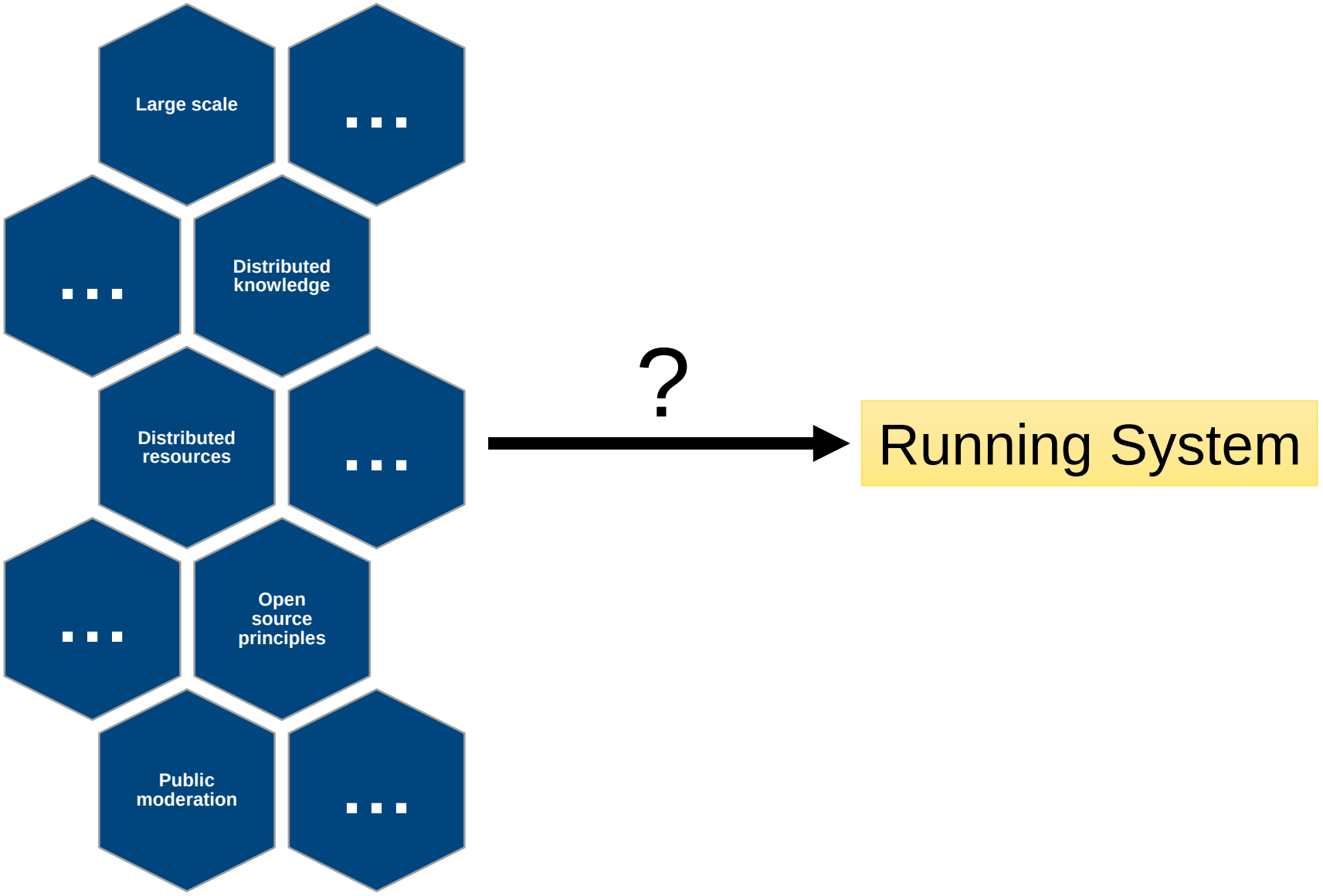
An European Open Search Infrastructure

- An **Open Search Index**, as a fundamental and indispensable basis for a large variety of public and private information services.



https://opensearchfoundation.org/

An European Open Search Infrastructure



An European Open Search Infrastructure

- Technological and computational aspects
  - Distributed crawling, indexing, and search
  - Distributed storage of Big Data
  - Security
  - …
- Societal aspects
  - Right to be forgotten
  - Transparency
  - Access management
  - Fake news detection
  - …
- …

We need a robust architecture!

An Architecture to Rule Them All?

Benefits of a good architecture:

- Standardized schemata
- Clear interfaces / APIs
- Well defined functional blocks

→ Will reduce risk

→ Will attract various players to contribute

→ Will allow for adaptation and specialization in a generic frame

→ Will foster the Open Search idea

An Architecture to Rule Them All?

Problem:

Architecture is everything that is costly to change later ⇔ Big-Design-Up-Front is not feasible at this scale and complexity

Solution:

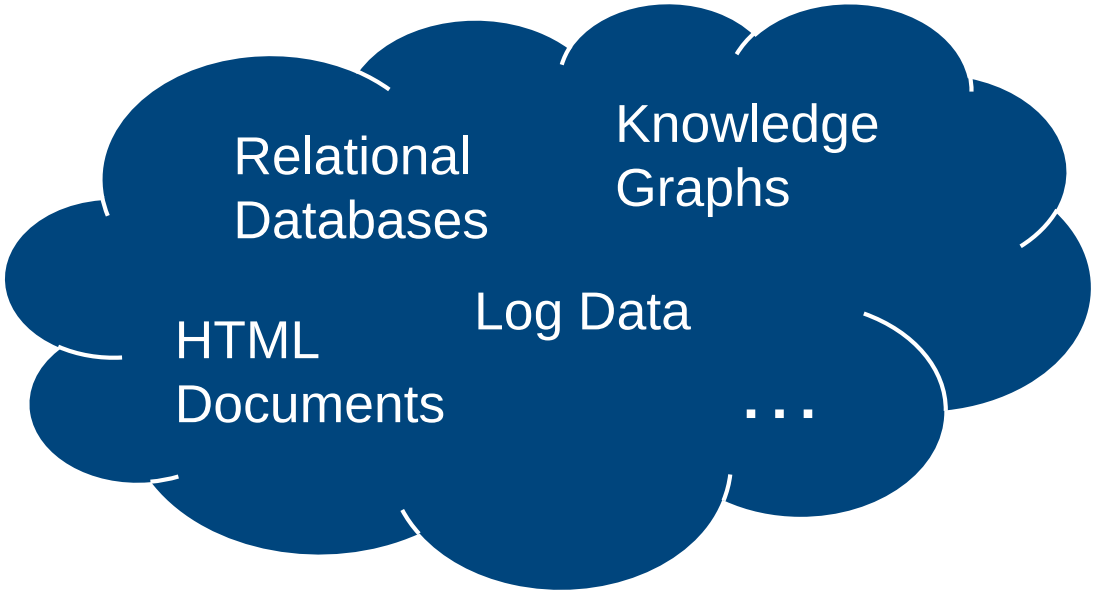
- Design an extensible architecture as a starting point that covers key aspects

Outline

1. Motivation
2. From Data Warehouses to Data Lakes
3. First Thoughts on an Architecture
4. Next steps & Conclusion

Talking Data

- Open search infrastructure will store heterogenous data ranging from deeply structured to totally unstructured



The infrastructure has to handle Big Data!

Handling Data using Data Warehouses

- For data analysis
- Data is only added and read
- Data is never updated or deleted
- Use of rigid data models and schemata tailored to specific data mining purpose

Problems:

- Big Data does not fit into predefined data models and schemata
- We have no specific data mining purpose

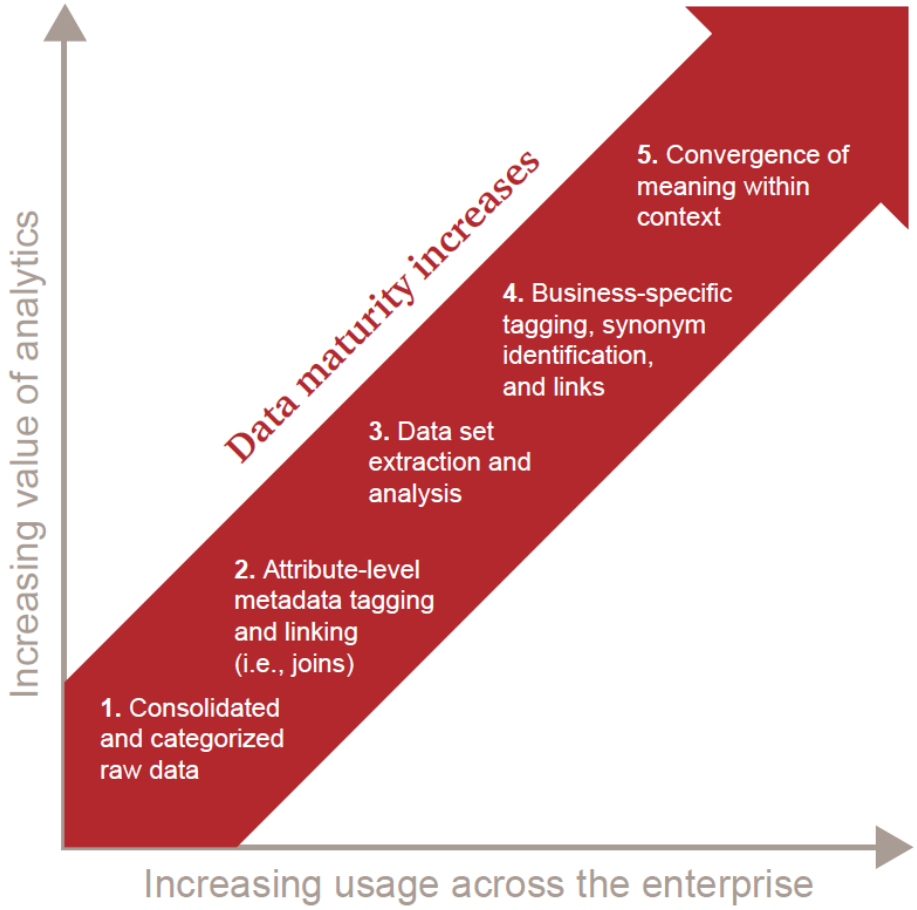
Handling Data using Data Lakes

- Regarding data storage
  - Data is only added and read
  - Data is never updated or deleted
  - Data is stored in their raw format
  - Metadata keeps track of new versions of data

- Regarding data governance
  - Clear-cut componentization and responsibilities
  - Proper use of metadata is mandatory to avoid data swamps
  - A catalog takes inventory and stores management routines

Handling Data using Data Lakes

- Regarding data interaction and maturation
  - Raw data interaction and view-based interaction
  - Data matures through user interaction



Adopted from [6]; in our case enterprise means the open search infrastructure

Handling Data using Data Lakes

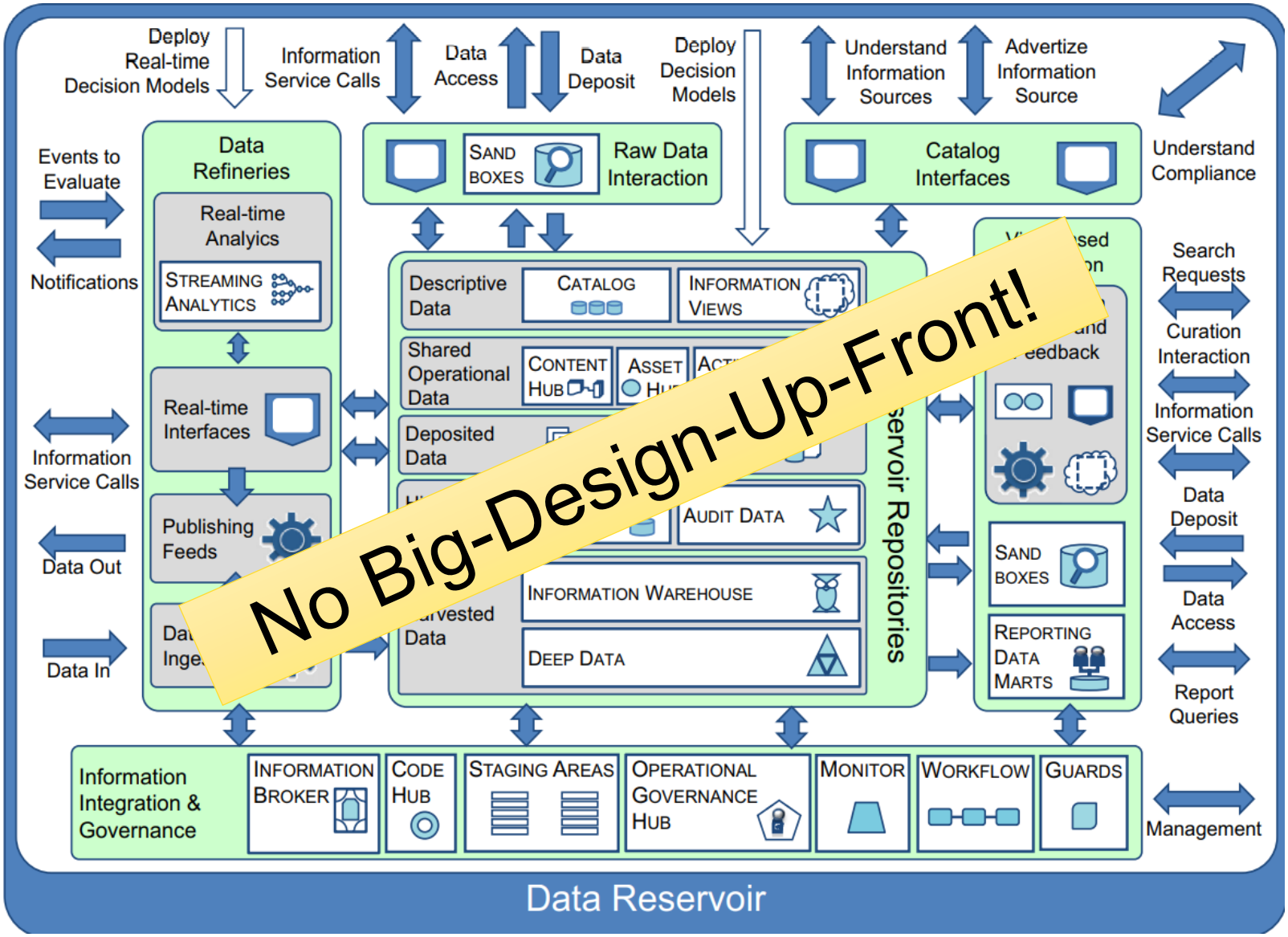
Key benefits:

- Schema-on-read defers data modelling and schema definition
- Data provenance always comprehensible
- High level of data accessibility
- Immediate access to original raw data
- Use case agnostic data management system

Outline

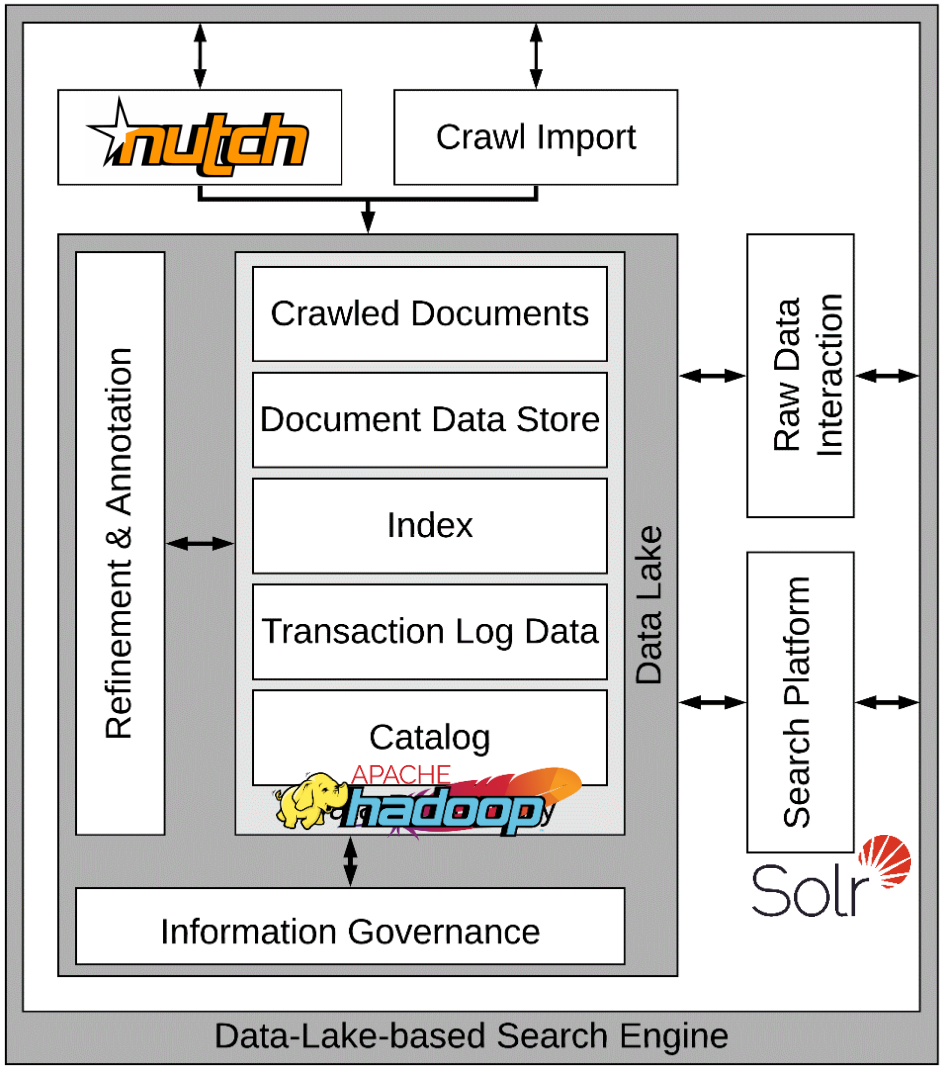
1. Motivation
2. From Data Warehouses to Data Lakes
3. First Thoughts on an Architecture
4. Next steps & Conclusion

Data Reservoir Overview



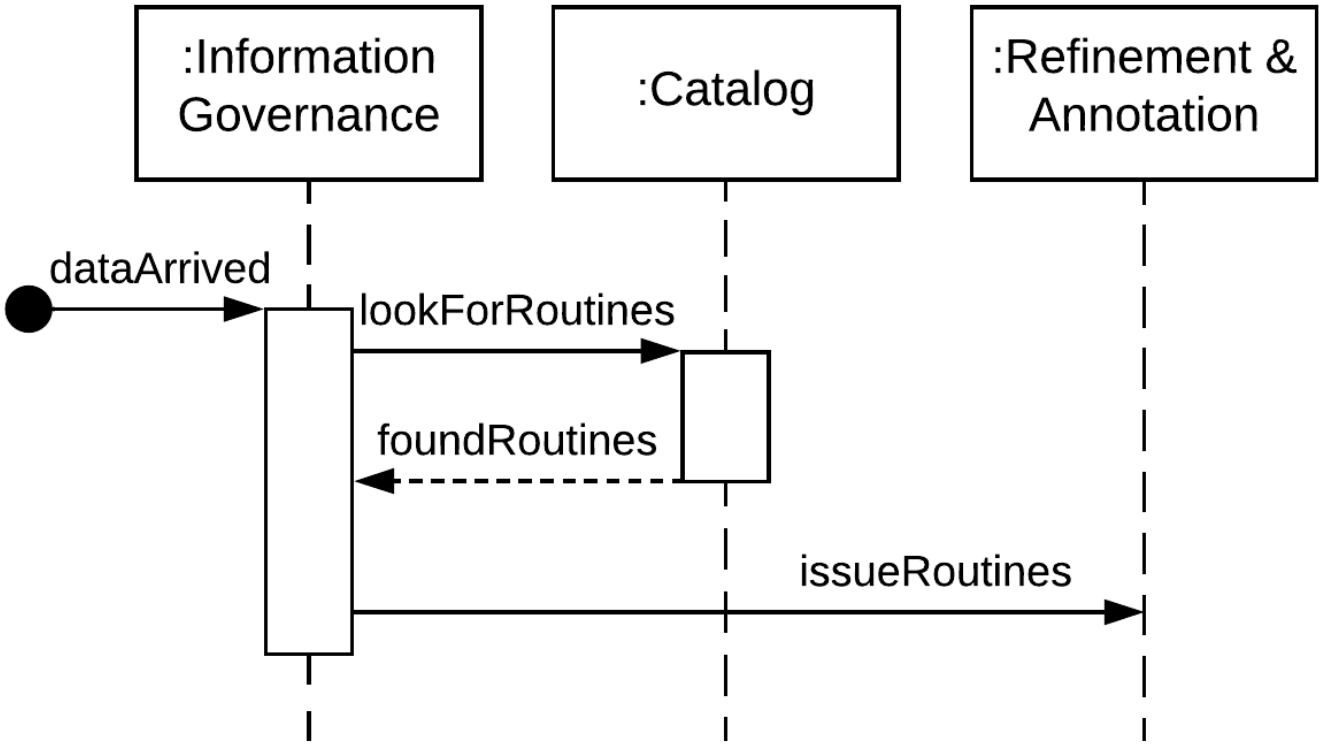
Adopted from [3]; the summary of a data reservoir, a data lake architecture by IBM

A Data-Lake-based Search Engine



- Minimum number of components
- Apache Hadoop, Apache Nutch, and Apache Solr as candidate technologies

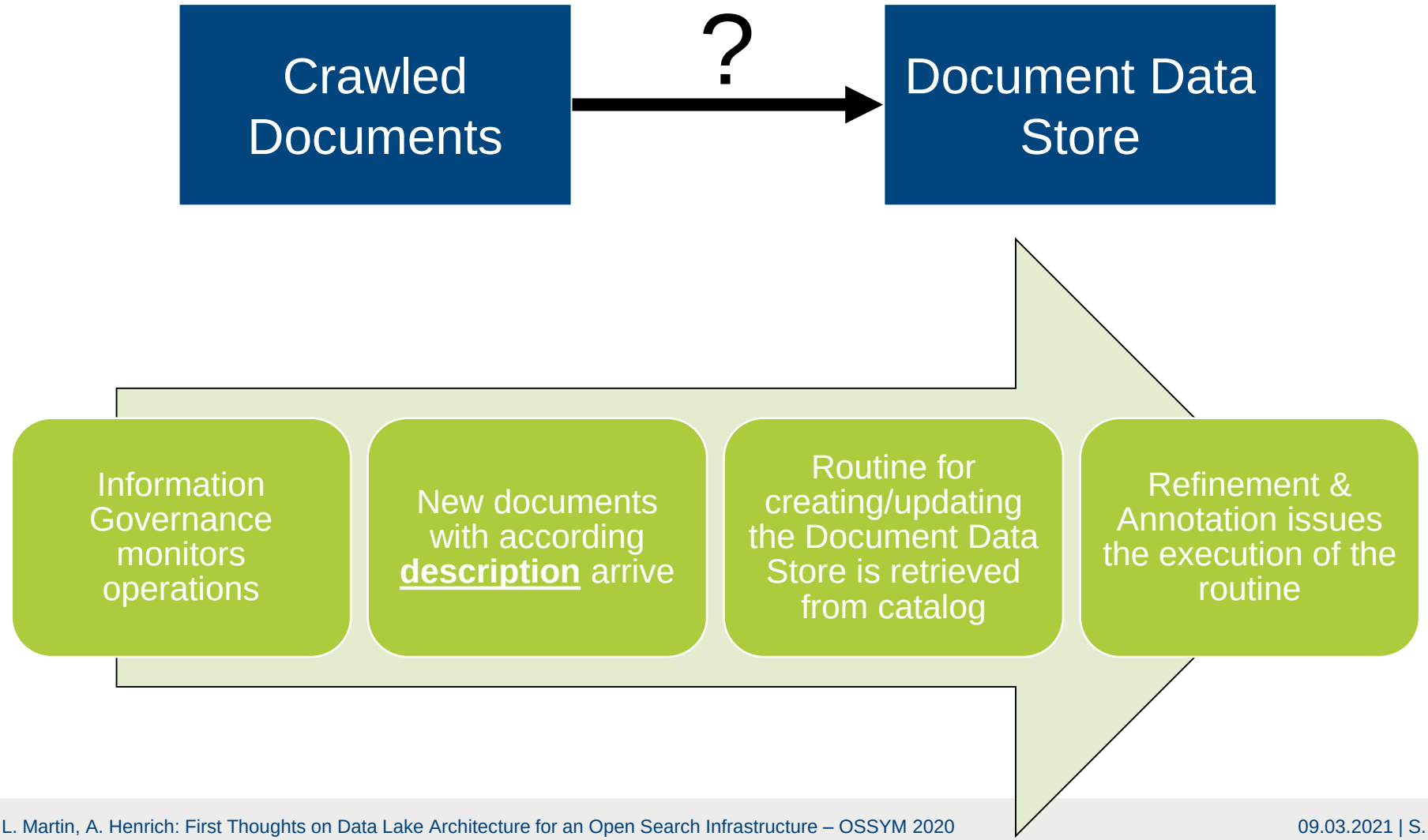
A Data-Lake-based Search Engine



Note the clear-cut componentization and responsibilities!

A Data-Lake-based Search Engine

Simplified example of a management routine:



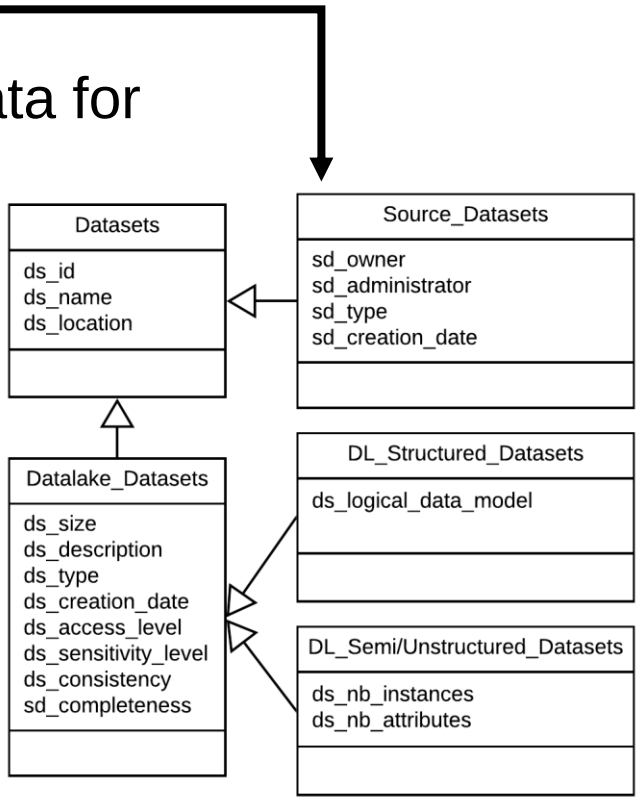
Outline

1. Motivation
2. From Data Warehouses to Data Lakes
3. First Thoughts on an Architecture
4. Next steps & Conclusion

Towards a Data-lake-based Search Infrastructure

We need

- a conceptual model for all metadata
- more (generic) views for accessing the data for other purposes besides search platform (e.g. statistics dashboards)
- more routines for automated refinement (e.g. NLP pipeline)
- statistics, security, …



Again, build a basic extensible architecture first!

Conclusion

- Open search infrastructure → Big Data
- Data lakes could be used as the basis for the envisaged open search infrastructure
- No Big-Design-Up-Front
- Start with a basic extensible architecture, e.g., for a data-lake-based search engine

The points discussed are by no means complete or settled. Instead, our goal is to fuel further discussions.

References

1. Apache Nutch, http://nutch.apache.org/.
2. Apache Solr, https://lucene.apache.org/solr/.
3. M. Chessell, F. Scheepers, N. Nguyen, R. van Kessel, and R. van der Starre, “Governing and managing big data for analytics and decision makers”, in *IBM Redguides*, 2014.
4. C. Madera and A. Laurent, “The next information architecture evolution”, in *Proceedings of the 8th International Conference on Management of Digital EcoSystems - MEDES*, New York, New York, USA, pp. 174-180, 2016.
5. F. Ravat and Y. Zhao, “Metadata Management for Data Lakes”, in *New Trends in Databases and Information Systems*, Cham, pp. 37-44, 2019.
6. B. Stein and A. Morrison, “The enterprise data lake: Better integration and deeper analytics”, in *PwC Technology Forecast: Rethinking integration*, pp. 1-9, 2014.
7. T. White, Hadoop - The Definitive Guide: Storage and Analysis at Internet Scale (3. ed., revised and updated): O’Reilly, 2012.
8. L. Martin, A. Henrich, “First Thoughts on a Data Lake Architecture for an Open Search Infrastructure”, presented at OSSYM 2020, Geneva, Switzerland, October 2020, this conference.

Thank you!

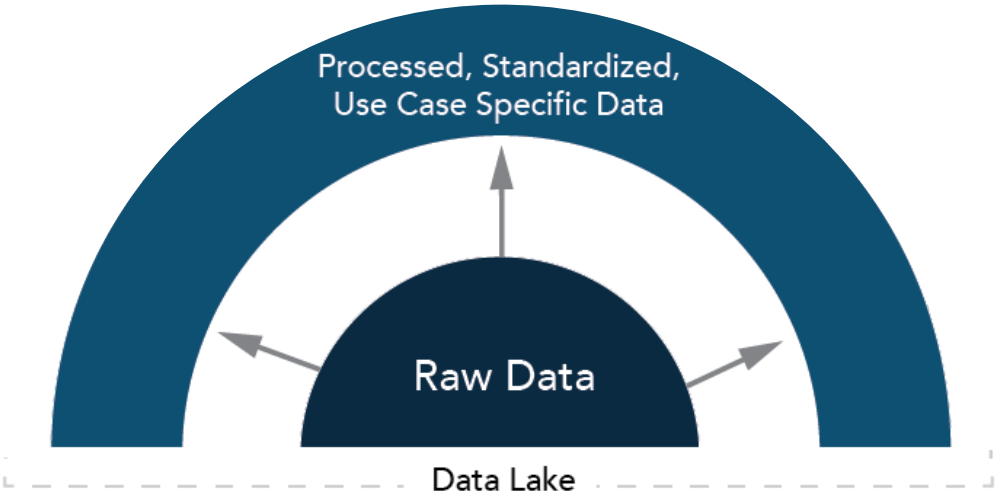
Backup Slides

Data Lake Layers and Consumption Patterns

- Enterprise Data Lake Architecture: What to Consider When Designing
  [Cloud Technology Partners, Sudi Bhattacharya, Neal Matthews
  https://www.cloudtp.com/doppler/how-to-guide-architecture-patterns-to-consider-when-designing-an-enterprise-data-lake/]

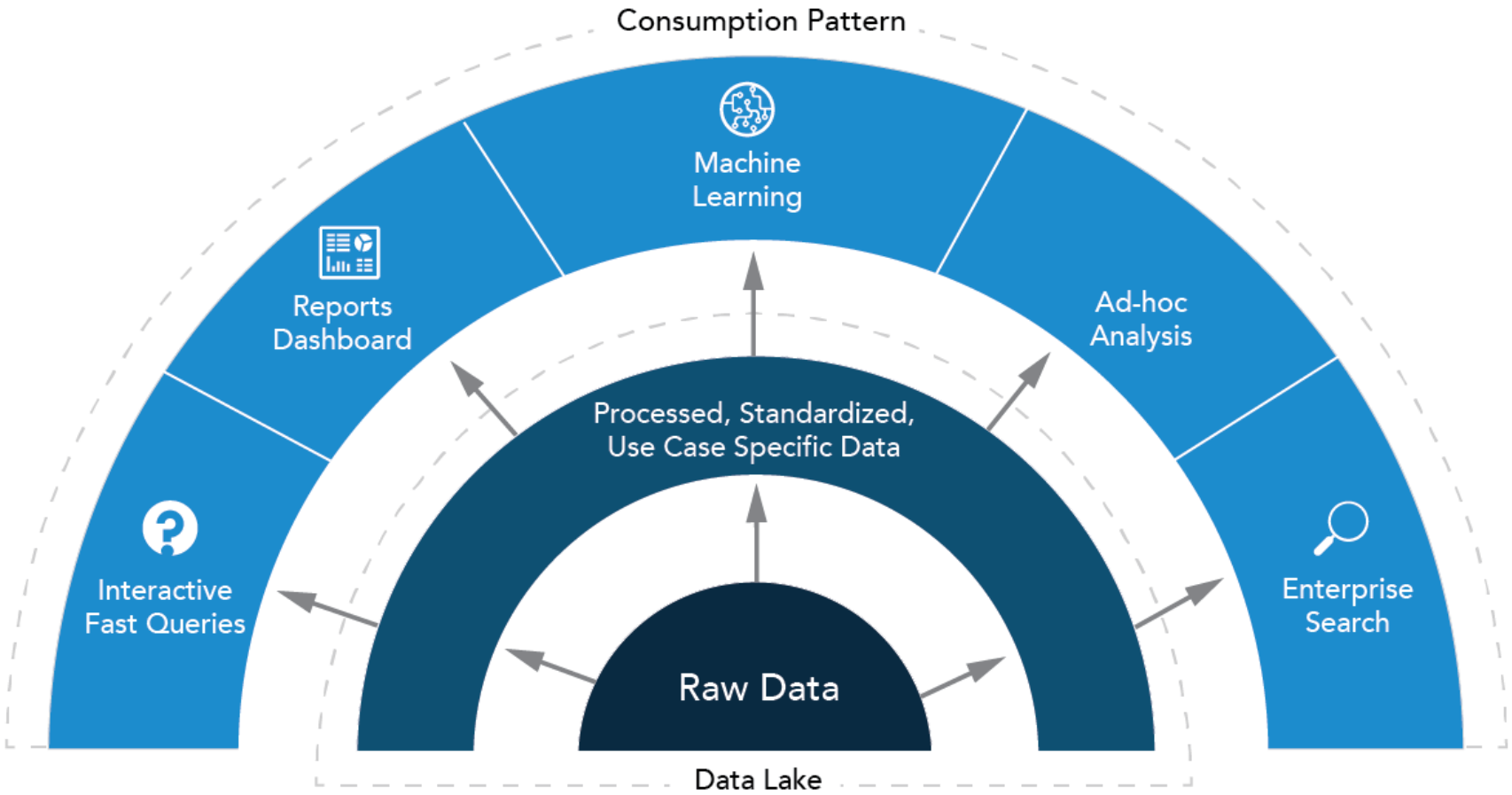
Data Lake Layers and Consumption Patterns

- Enterprise Data Lake Architecture: What to Consider When Designing
  [Cloud Technology Partners, Sudi Bhattacharya, Neal Matthews
  https://www.cloudtp.com/doppler/how-to-guide-architecture-patterns-to-consider-when-designing-an-enterprise-data-lake/]

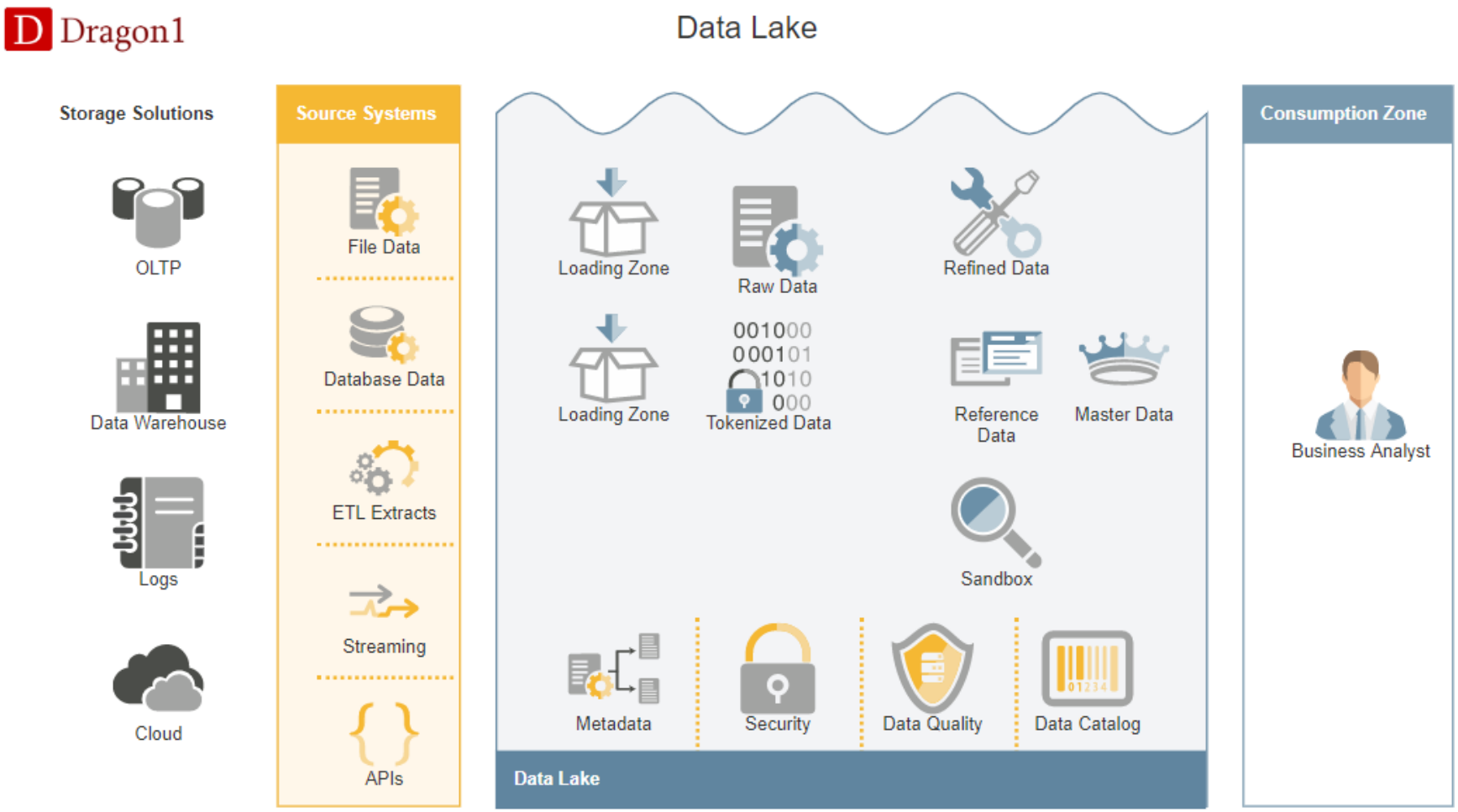


Data Lake Layers and Consumption Patterns

- Enterprise Data Lake Architecture: What to Consider When Designing
  [Cloud Technology Partners, Sudi Bhattacharya, Neal Matthews
  https://www.cloudtp.com/doppler/how-to-guide-architecture-patterns-to-consider-when-designing-an-enterprise-data-lake/]



Data Lake Template for Reference Architecture



https://www.dragon1.com/demo/data-lake

Key Benefits Of Data Lake

1. Scalability
   - storage from disparate sources like multimedia, binary, XML; …
2. High-velocity Data
   - data stream processing and large volumes of historical data
3. Structure
   - unique arena where structure like metadata, speech tagging etc. can be applied on varied datasets
4. Storage
   - iterative and immediate access to the raw data
5. Schema
   - schemaless write and schema-based read

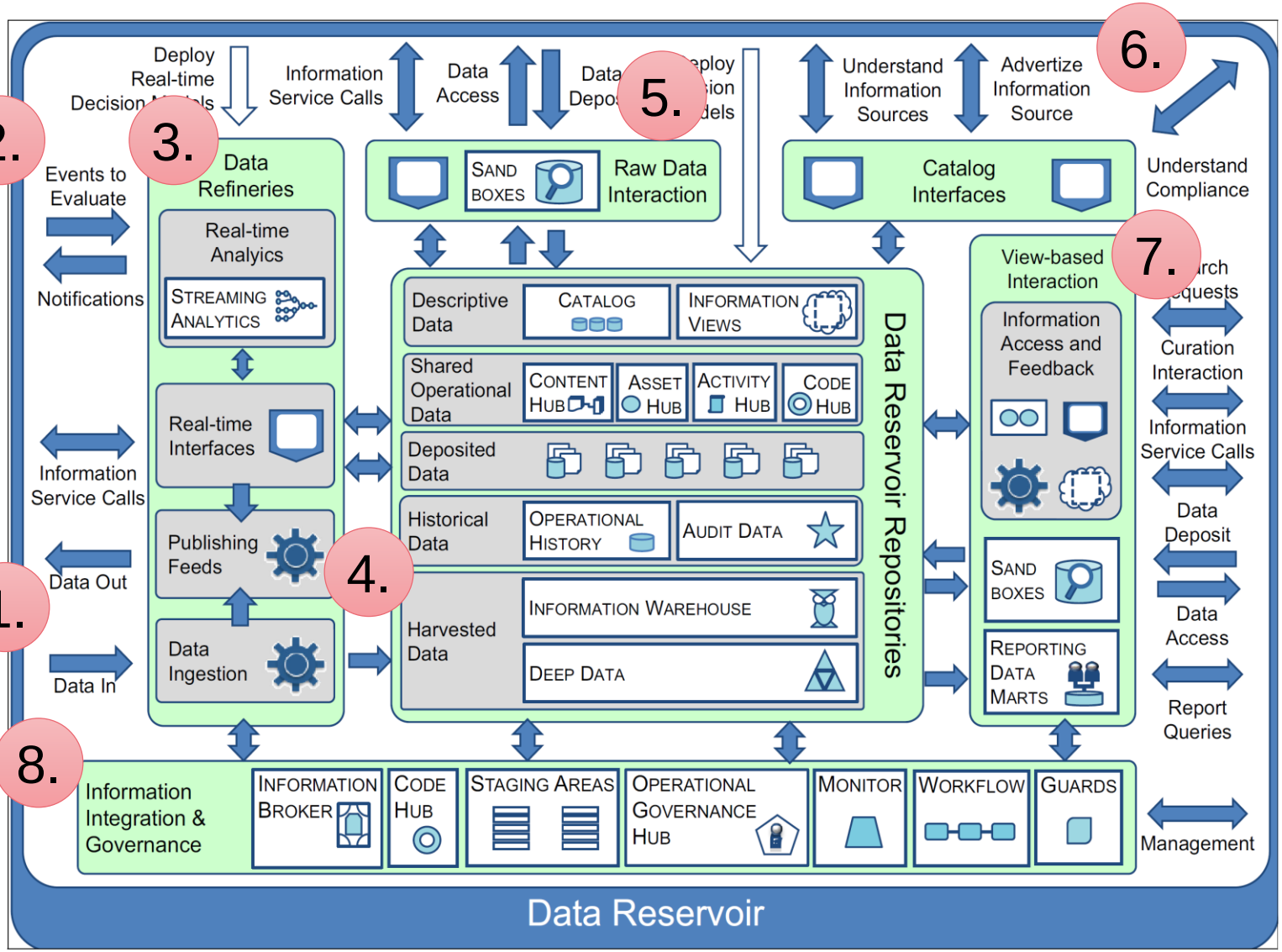
Source: Ajit Singh: *Architecture of Data Lake*, 2019, Data science Foundation, https://datascience.foundation/sciencewhitepaper/architecture-of-data-lake

Architecture of Data Lake

- Factors to consider:
  - Data Governance and Security Layer
  - Metadata Layer
  - Information Lifecycle Management Layer
- Tiers to manage data flows :
  - Intake Tier
  - Management Tier
  - Consumption Tier
- What is needed according to the CAP theorem?
  - Consistency
  - Availability
  - Partition tolerance

Source: Ajit Singh: *Architecture of Data Lake*, 2019, Data science Foundation, https://datascience.foundation/sciencewhitepaper/architecture-of-data-lake

Data Reservoir Overview



Source: M. Chessell, F. Scheepers, N. Nguyen, R. van Kessel, R. v.d. Starre: *Governing and Managing Big Data for Analytics and Decision Makers*, Redguides for Business Leaders , 2014, IBM, http://www.redbooks.ibm.com/redpapers/pdfs/redp5120.pdf