

Hydrating a Data Lake to Power Innovation

Enabling Pharmaceutical Companies to Automate
Data De-Identification for Secondary Use

d-wise™

An Instem Company



Kristin McDougall, Director of Transparency Services
Cathal Gallagher, Delivery Lead, Transparency

The Cost of Transparency

RESOURCES NEEDED



Time

Reacting to new regulations and developing new processes takes time.



Money

Both outsourced and internal methods require additional costs.



Knowledge

Keeping up-to-date with current regulations can feel like a full-time job.

ROI



Public Education

Transparency enables the public to make informed decisions regarding their health care.



Trust

Being transparent with trial data builds trust in the clinical trial process.



Secondary Research

Empower medical breakthroughs by utilizing existing data to minimize duplicative efforts and reduce timelines.

Thinking Beyond Regulations



Completed Study Data



**Compliance Standards
Roadblocks**



**Exploratory analysis to
support future studies**

Minimize duplicate efforts

Medical Breakthroughs

Potential

Automating De-Identification at a Large Scale | Case Study

A global sponsor enlisted d-wise to provide a proof-of-concept capability to **accelerate the flow of data to and from an internal data lake** through d-wise's Blur anonymization solution, providing sufficiently anonymized content to achieve re-identification risk levels appropriate for internal secondary use.

Challenges

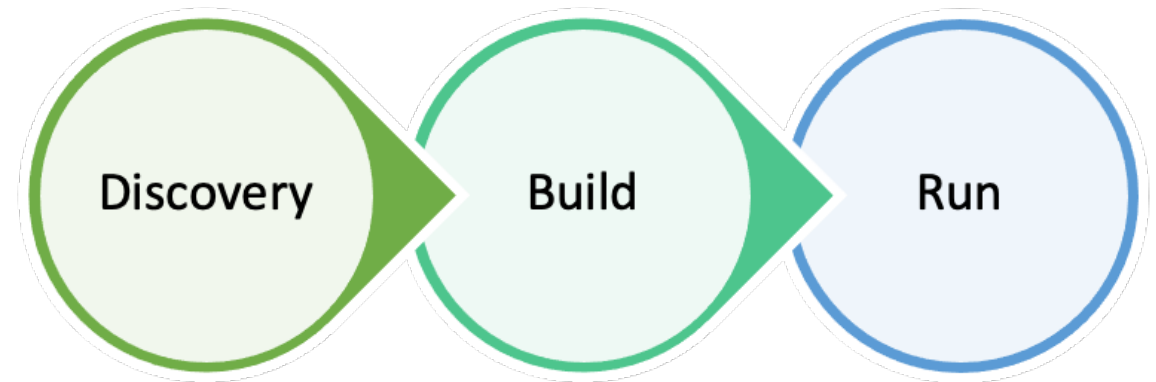
- What to Anonymize
- Internal Process Alignment
- Scalability
- Risk Threshold Strategy

Solution

Automated de-identification
of datasets through Blur

d-wise Approach

- **Discovery, Build, Run Process**
- **Master Template Format**
- **Audit Trail**
- **Variable Analysis**
- **Verbatim Text**
- **Inbound Folder Structure for System Input**
- **Outbound Folder to Receive Final, Resorted Data**



Run Phase

Manual

- Place study folders into Inbound folder

Automated

- Template applied
- Calculating Match Score
- Anonymization Report
- Resorting datasets
- Exporting into study folders in Outbound folder

Results

84

Studies
Anonymized

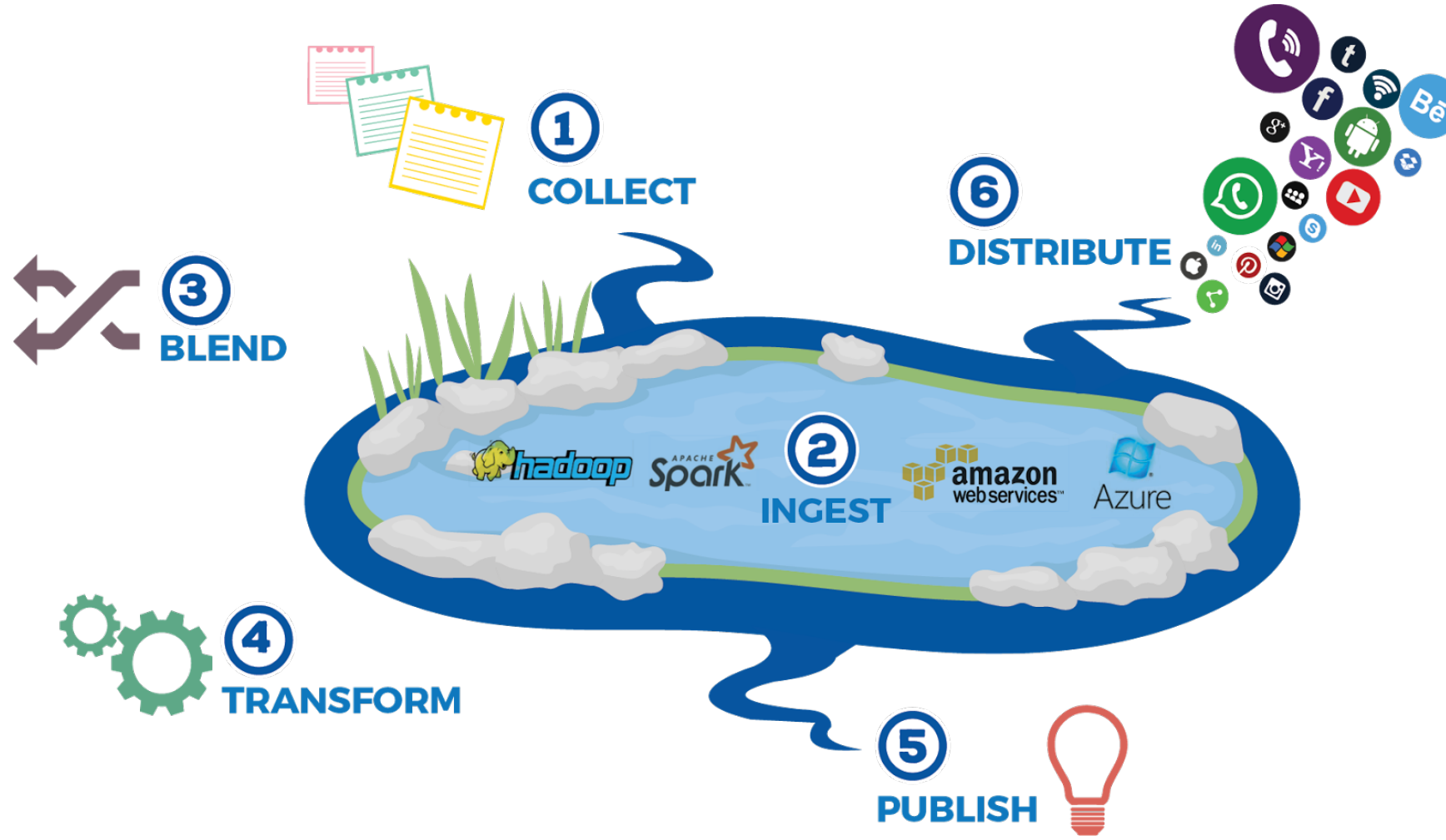
30

Minutes

1

Manual
Task

Data Lake Considerations



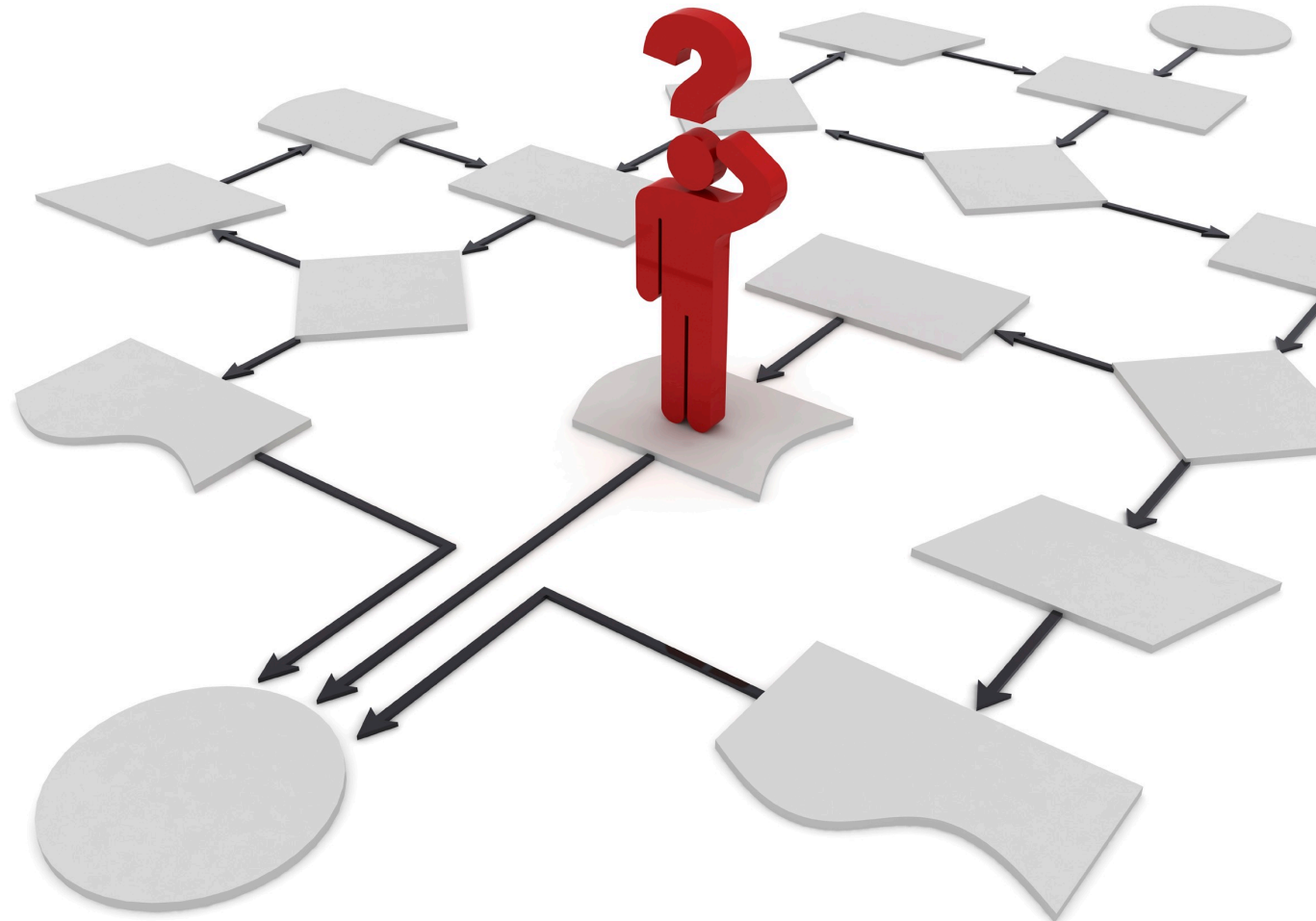
Data Lake Considerations

Access Rights



Data Lake Considerations

Data Mapping



Data Lake Considerations

What's the Point?

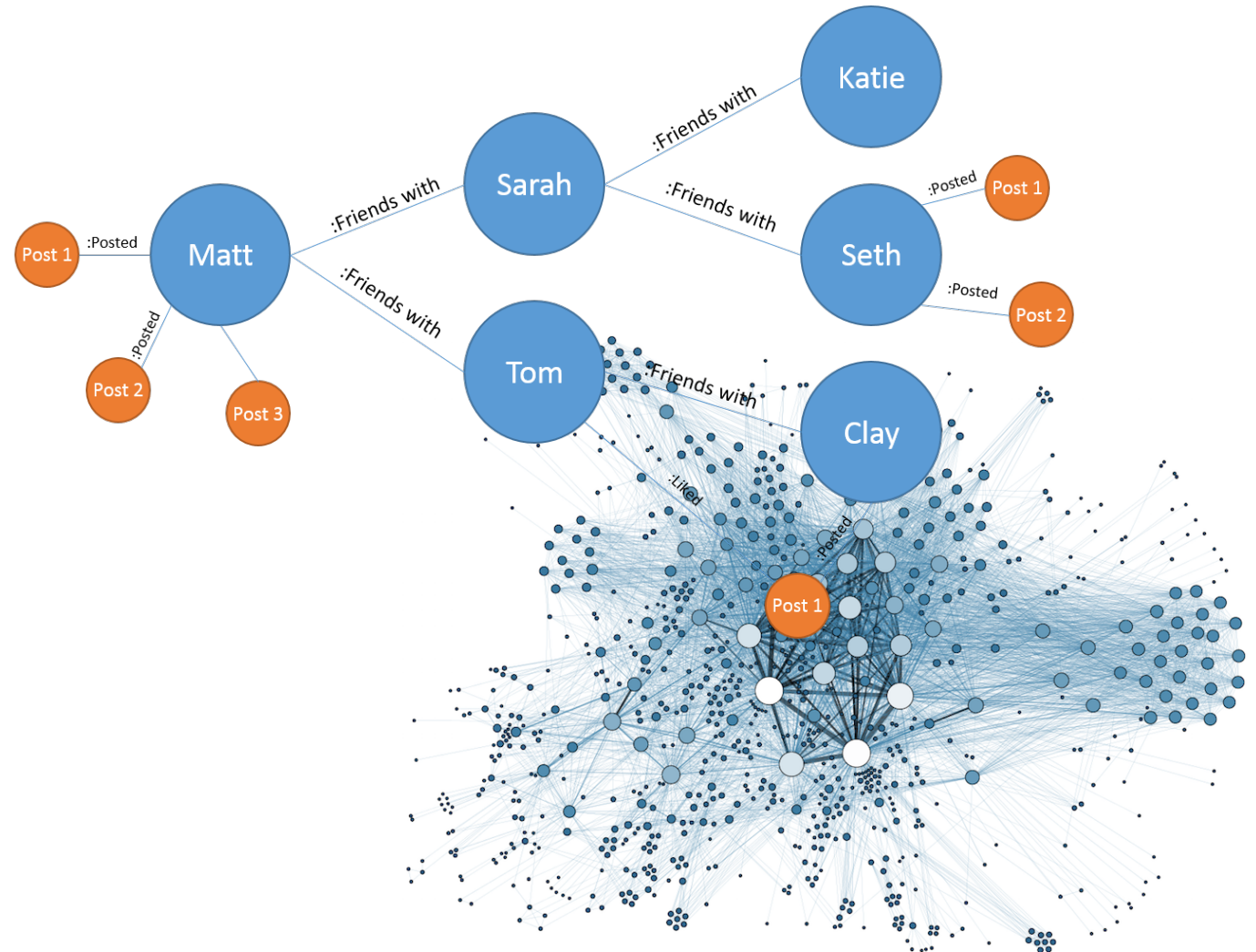


Data Lake Considerations

How will you store it and why?

syview_1_shoes.sas7bdat

	Region	Product	Subsidiary	Stores	Sales	Inventory	Returns
322	United States	Slipper	New York	24	\$252,758	\$1,059,33	\$8,503
323	United States	Sport Sho	New York	23	\$22,190	\$141,140	\$847
324	United States	Women's C	New York	10	\$178,842	\$438,035	\$4,238
325	United States	Women's D	New York	17	\$283,972	\$1,129,69	\$10,257
326	United States	Boot	Seattle	15	\$70,790	\$226,678	\$3,039
327	United States	Men's Cas	Seattle	6	\$66,842	\$112,638	\$3,335
328	United States	Men's Dre	Seattle	10	\$132,397	\$274,310	\$3,150
329	United States	Sandal	Seattle	10	\$6,595	\$18,758	\$181
330	United States	Slipper	Seattle	10	\$155,255	\$53,502	\$5,277
331	United States	Sport Sho	Seattle	29	\$80,408	\$108,266	\$1,021
332	United States	Women's C	Seattle	2	\$17,252	\$28,519	\$730
333	United States	Women's D	Seattle	21	\$133,405	\$392,007	\$4,037
334	Western Europe	Boot	Copenhagen	2	\$1,683	\$4,657	\$129
335	Western Europe	Men's Cas	Copenhagen	1	\$38,492	\$90,672	\$767
336	Western Europe	Men's Dre	Copenhagen	2	\$12,415	\$36,074	\$686
337	Western Europe	Slipper	Copenhagen	4	\$15,908	\$59,133	\$594
338	Western Europe	Sport Sho	Copenhagen	1	\$1,927	\$17,603	\$31
339	Western Europe	Sport Sho	Copenhagen	15	\$101,922	\$327,742	\$4,204



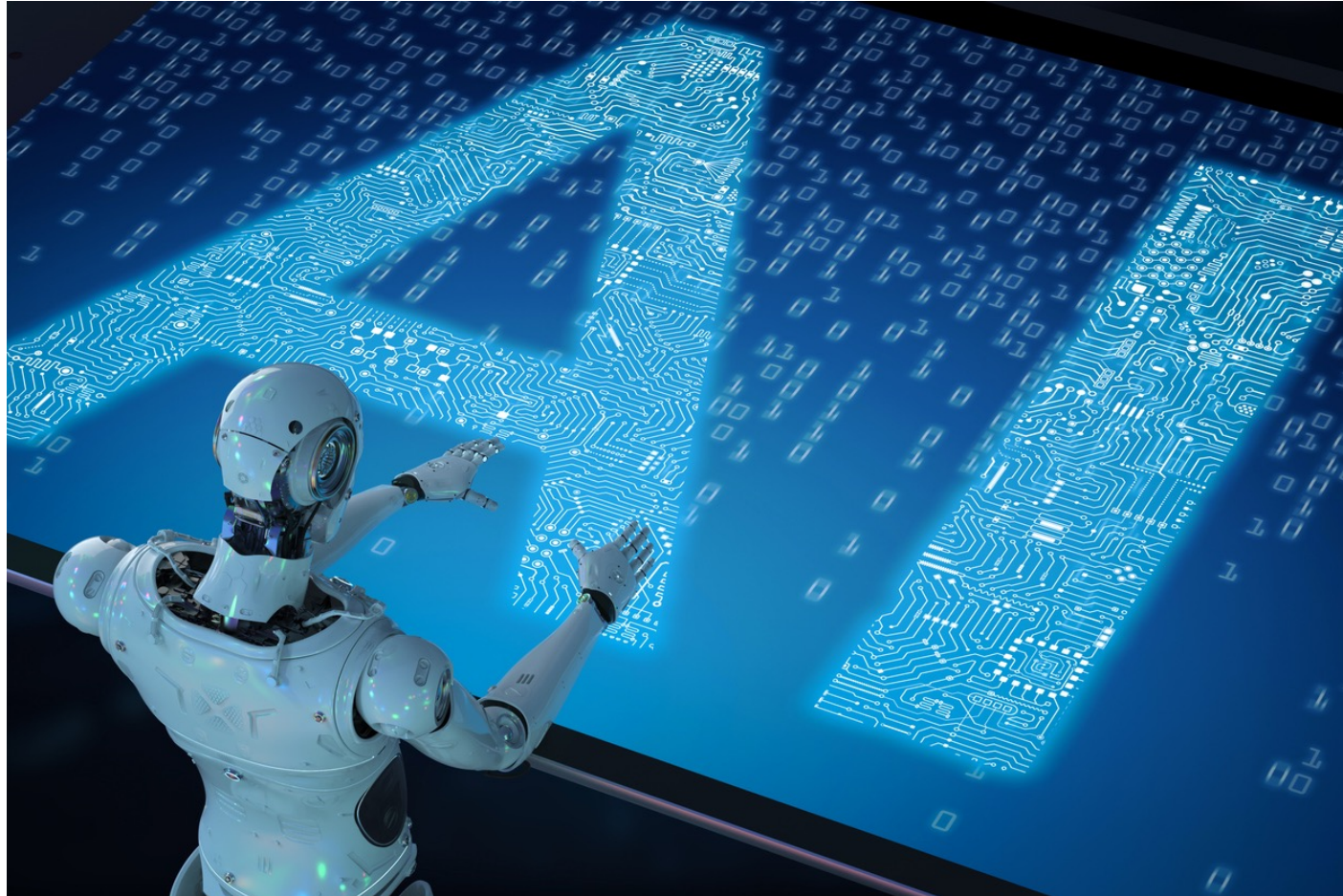
Data Lake Considerations

Visualization software uses flat files



Data Lake Considerations

Machine Learning



Data Lake Considerations

Entry Points



Data Lake Considerations

Real World Evidence Integration



Contact

Kristin McDougall

Director of Transparency

kristin.mcdougall@d-wise.com

www.d-wise.com