

# Solr as an Oak index for AEM

Tommaso Teofili | Computer Scientist



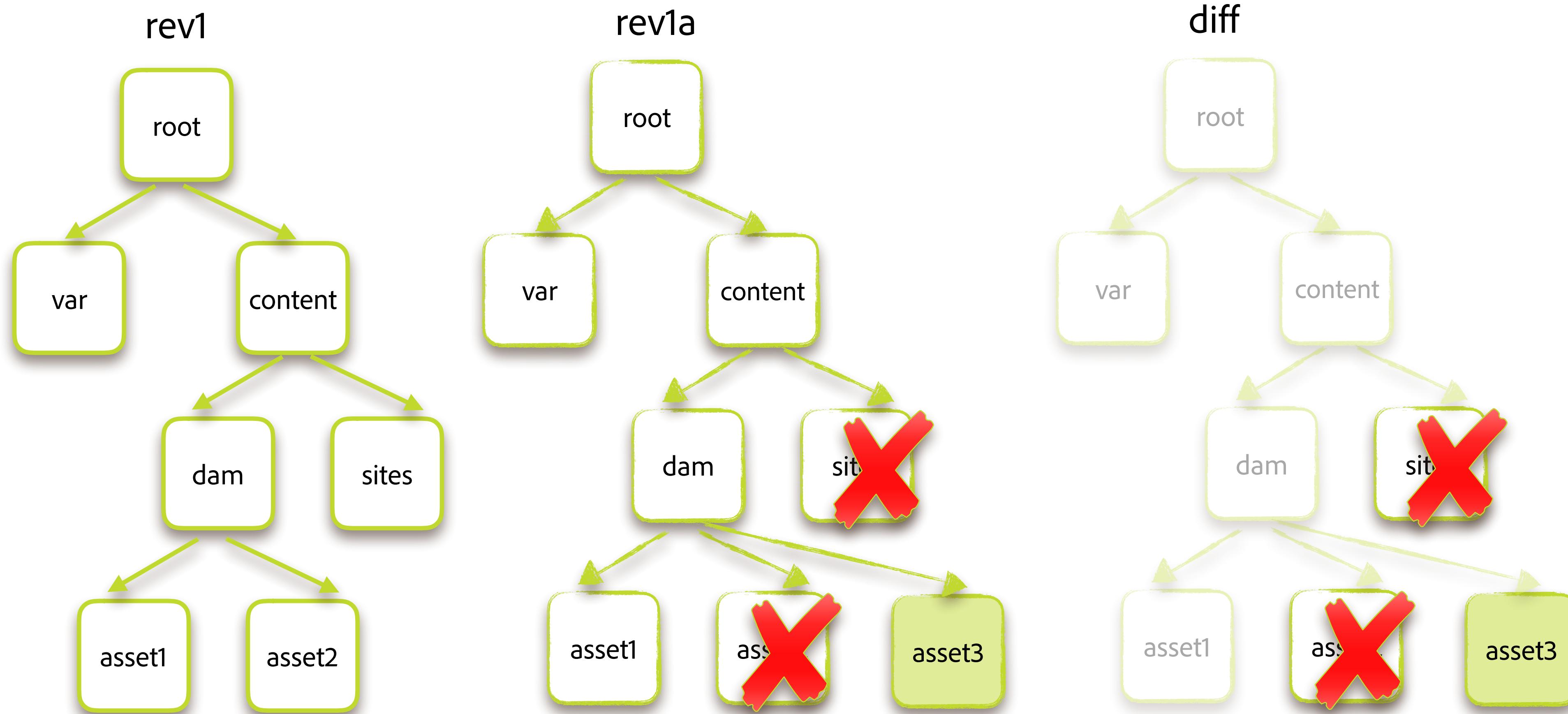
#AdobeRemix

Hiroyuki-Mitsume Takahashi

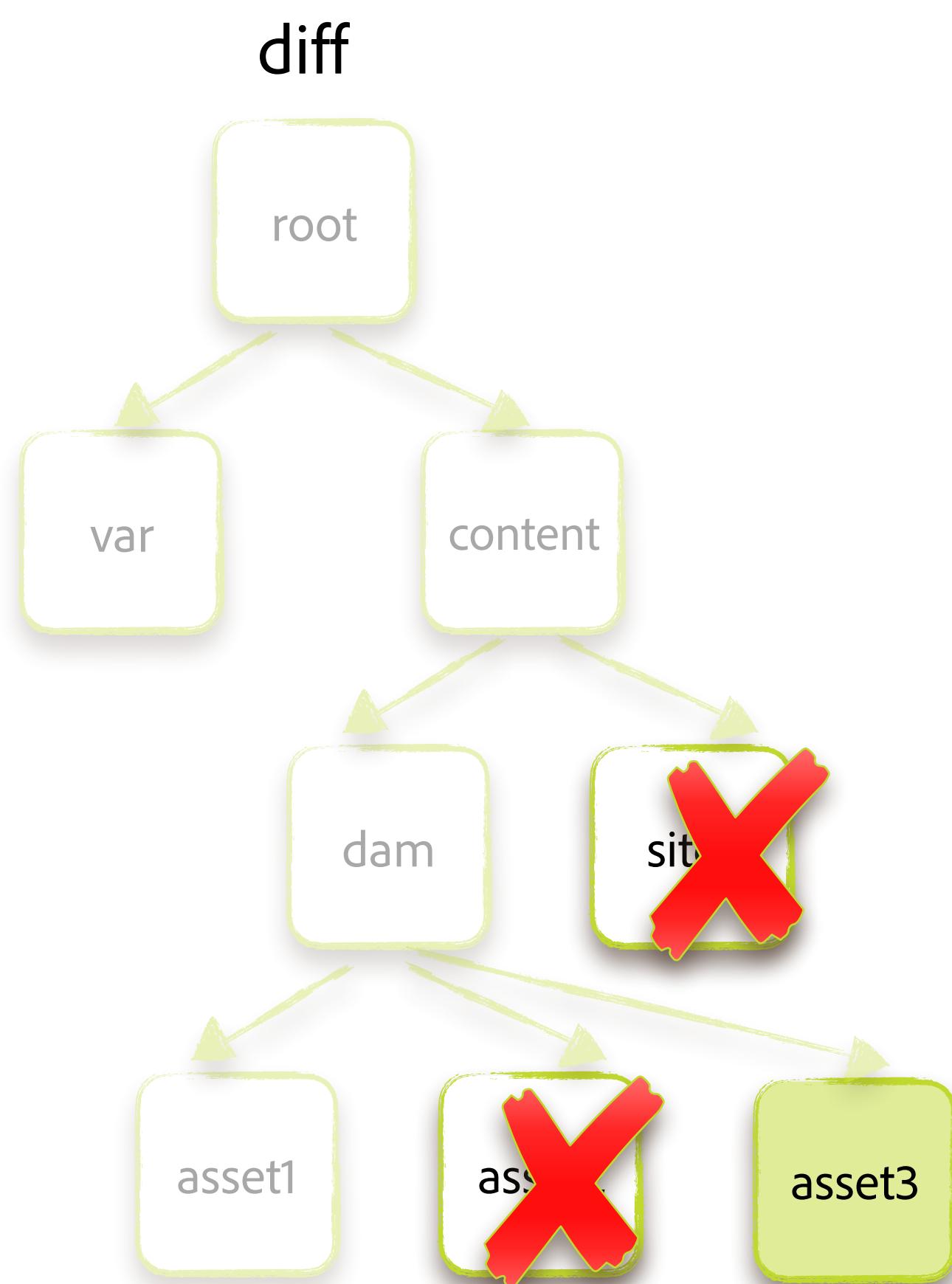
# Agenda

- Oak indexing & search in AEM overview
- Apache Solr
- Oak Solr indexes deployment scenarios for AEM
- Q & A

# Oak indexing in AEM



# Oak indexing in AEM



- Diffs define what changes need to be done in indexes
  - delete /content/sites
  - delete /content/dam/asset2
  - add /content/dam/asset3

# Oak indexing in AEM - index types

- Property
- Lucene
- Solr
- Counter
- Node type

The screenshot shows the Adobe Experience Manager (AEM) Index Manager interface. The left sidebar contains filters for Index Name, Path, and Type. The Type filter is set to "All". The main area displays a table of configured indexes with columns: Index Name, Path, Type, and Async.

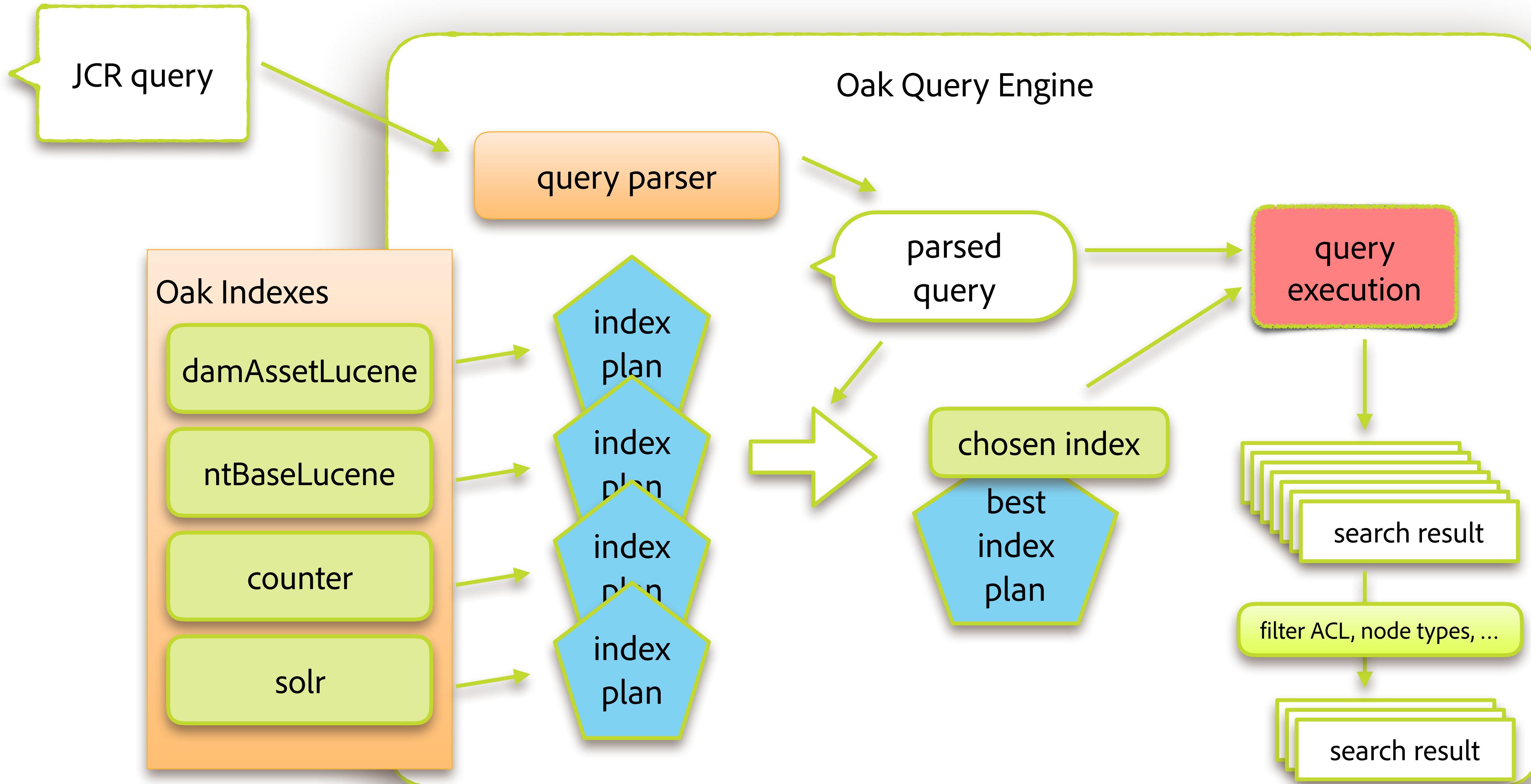
Index Name	Path	Type	Async
socialLucene	/oak:index/socialLucene	lucene	async, nrt
workflowDataLucene	/oak:index/workflowDataLucene	lucene	async, nrt
slingeventJob	/oak:index/slingeventJob	lucene	async, nrt
jcrLanguage	/oak:index/jcrLanguage	property	
versionStoreIndex	/oak:index/versionStoreIndex	lucene	async, sync
repMembers	/oak:index/repMembers	property	
cqReportsLucene	/oak:index/cqReportsLucene	lucene	async
commerceLucene	/oak:index/commerceLucene	lucene	async, nrt
counter	/oak:index/counter	counter	async
authorizables	/oak:index/authorizables	lucene	async, nrt
enablementResourceName	/oak:index/enablementResourceName	lucene	async
externalPrincipalNames	/oak:index/externalPrincipalNames	property	

# Oak indexing in AEM - Index definitions

The screenshot shows the CRXDE Lite interface for managing Oak indexing. On the left, a tree view displays a node structure under 'ntBaseLucene' / 'indexRules' / 'nt:base' / 'properties'. The selected node is 'cqTags', which has properties like 'cqTags', 'slingResource', 'damMetadataSch...', 'damMetadataProf...', 'damVideoProfile', 'damImageProfile', 'damResolvedPath', 'afTemplateVersion', 'damAutoTag', 'cqLastReplication', 'status', 'type', 'offTime', 'onTime', 'analytics\_pageview', and 'analytics\_visitors'. The main area is titled 'CRXDE | Lite' and contains a search bar with placeholder text 'Enter search term to search the repository'. Below the search bar is a table titled 'Properties' with columns 'Name', 'Type', and 'Value'. The table lists the following properties:

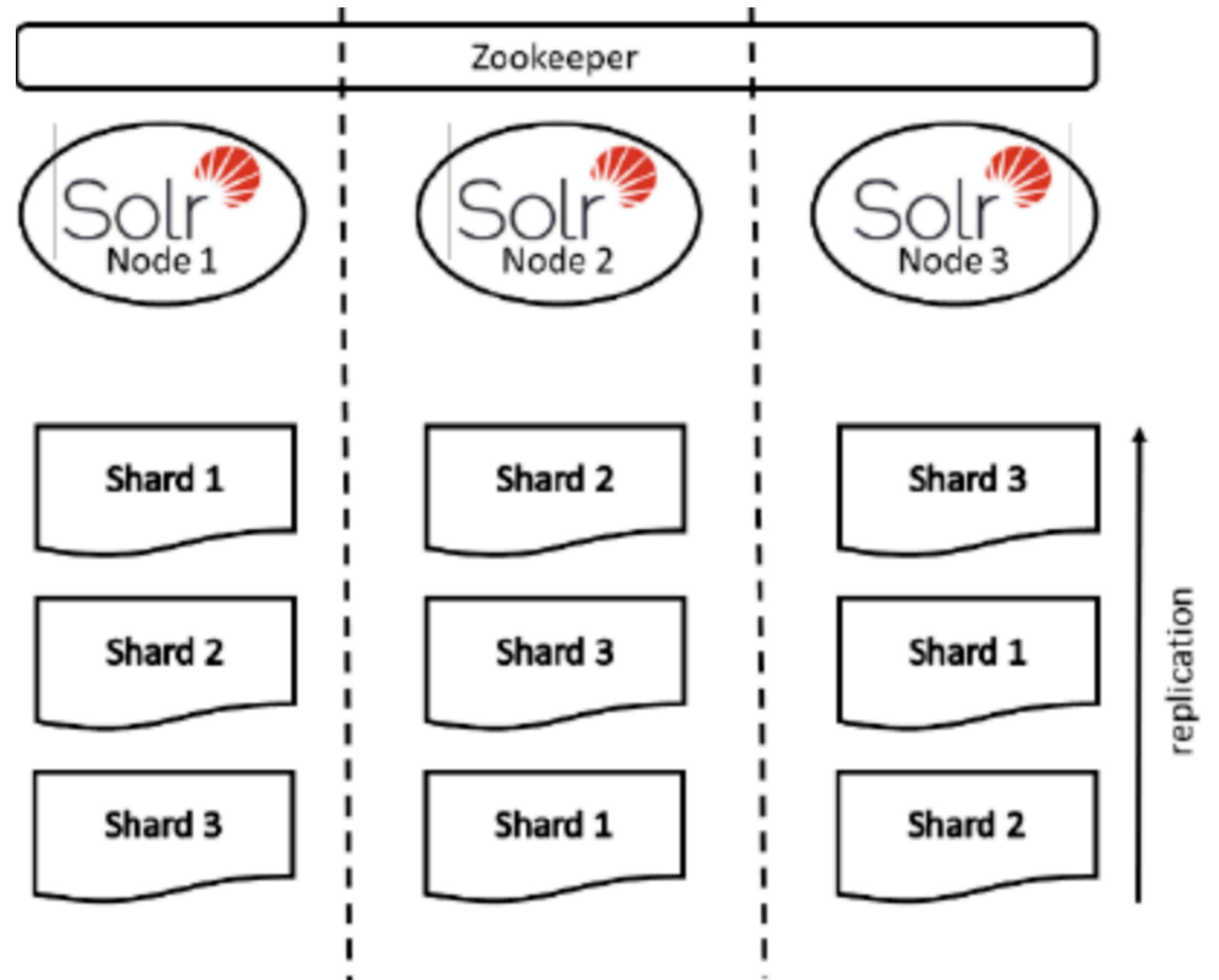
Name	Type	Value
1 async	String[]	async, nrt
2 compatVersion	Long	2
3 evaluatePathRestrictions	Boolean	true
4 jcr:primaryType	Name	oak:QueryIndexDefinition
5 reindex	Boolean	false
6 reindexCount	Long	1
7 supersedes	String[]	/oak:index/cqDefaultFormFor, /oak:in...
8 type	String	lucene

# Oak searching in AEM



# Apache Solr

- Enterprise search server based on Apache Lucene
- Horizontal and vertical scaling through SolrCloud
- Cluster coordination via Apache Zookeeper
- Rich full text query syntax
- Highly configurable relevance and indexing
- Plugin architecture for
  - query parsing
  - searching / ranking
  - indexing
  - ...
- Latest Oak Solr index supports Apache Solr 5.5.5
- <http://jackrabbit.apache.org/oak/docs/query/solr.html>



# Index definition for SolrCloud cluster

1. open CRXDE
2. create a oak:QueryIndexDefinitionNode (e.g. named solrcloud)
3. set 'type' property to 'solr'
4. set 'async' property to 'async'
5. create a 'server' node under 'solrcloud' node (nt:unstructured type)
6. set 'solrServerType' property to 'remote'
7. set 'zkHost' to Solr Zookeeper host(s) (e.g. 10.1.2.10:9983,10.1.2.11:9983,10.1.2.12:9983)
8. more optional properties ...

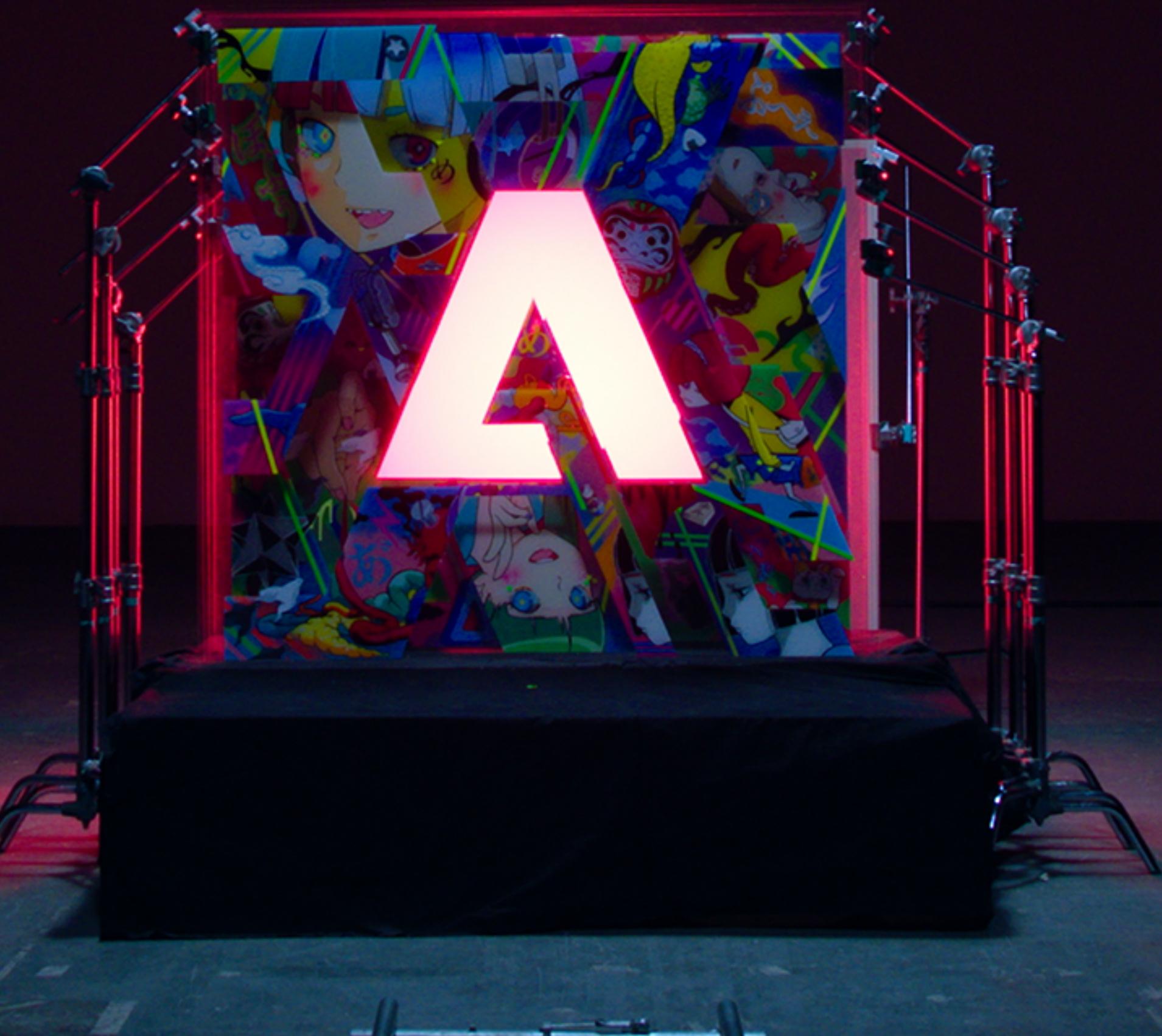
The screenshot shows the CRXDE interface with the following details:

- Node Structure:** The left pane displays a tree view of nodes under the root. The 'solrcloud' node is expanded, showing its children: 'server', 'rep:policy', 'cqPayloadPath', 'nodetypeLucene', 'nodetype', 'ntBaseLucene', 'reference', 'principalName', 'cqTagLucene', 'lucene', and 'repTokenIndex'. The 'server' node is selected.
- Properties View:** The right pane shows the properties for the selected 'server' node. It includes tabs for Properties, Access Control, Replication, and Console. The Properties tab is active, displaying the following table:

Name	Type	Value
1 configurationDirectory	String	/Users/teofili/dev/oak
2 jcr:primaryType	Name	nt:unstructured
3 replicationFactor	Long	1
4 shardsNo	Long	2
5 solrServerType	String	remote
6 zkHost	String	localhost:9983

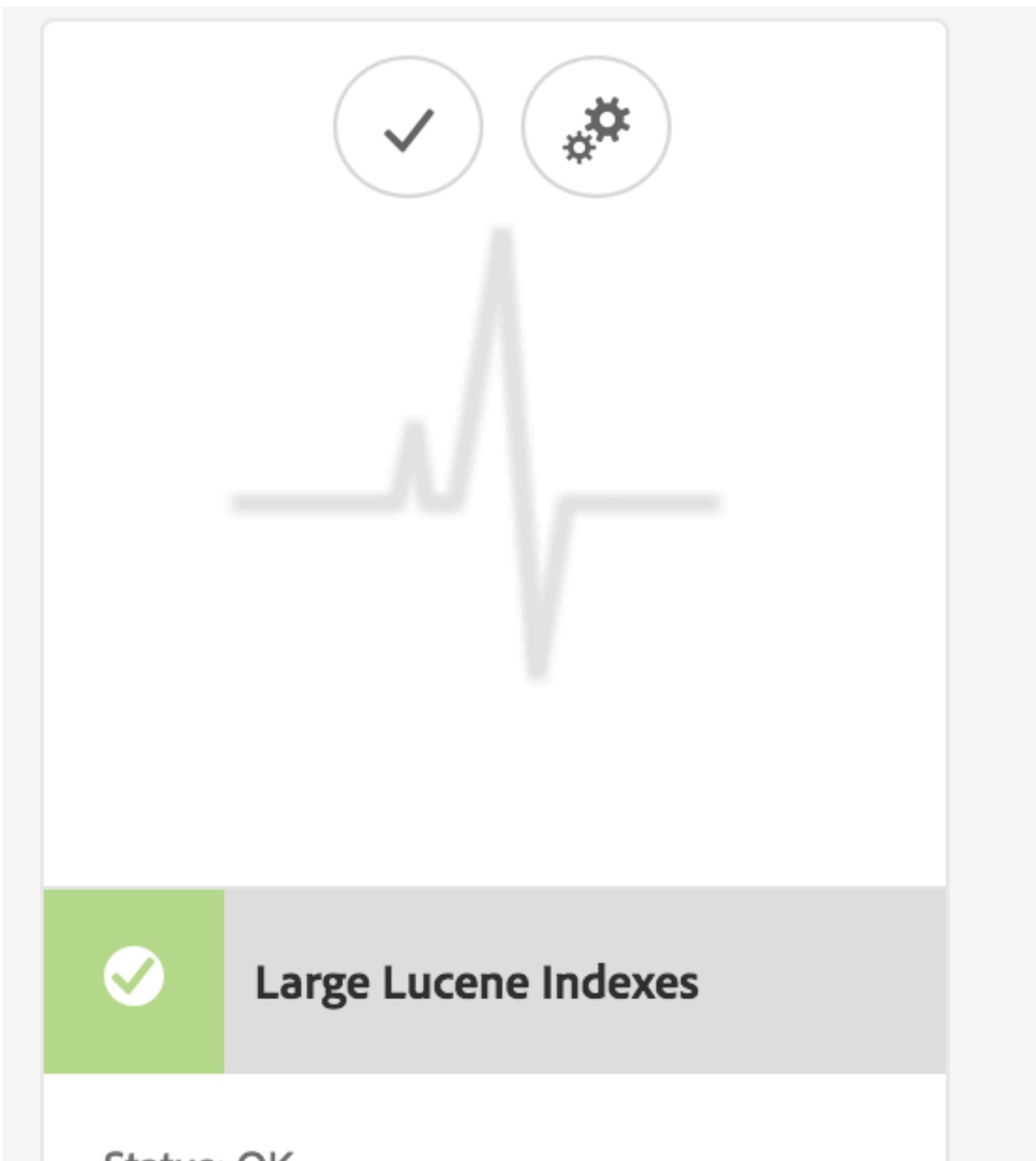
# Why and when

Deployment scenarios for Oak Solr index in AEM



# Huge Lucene indexes

- Number of documents
  - hit the Lucene 2B documents limit
- Size of documents
  - repository size growth
    - for Segment, might be a bottleneck
    - Lucene indexes binaries take lots of space (binaries in DS)
- Offload indexing and search load out of the repo



# Solr features not covered by current AEM Oak Lucene index

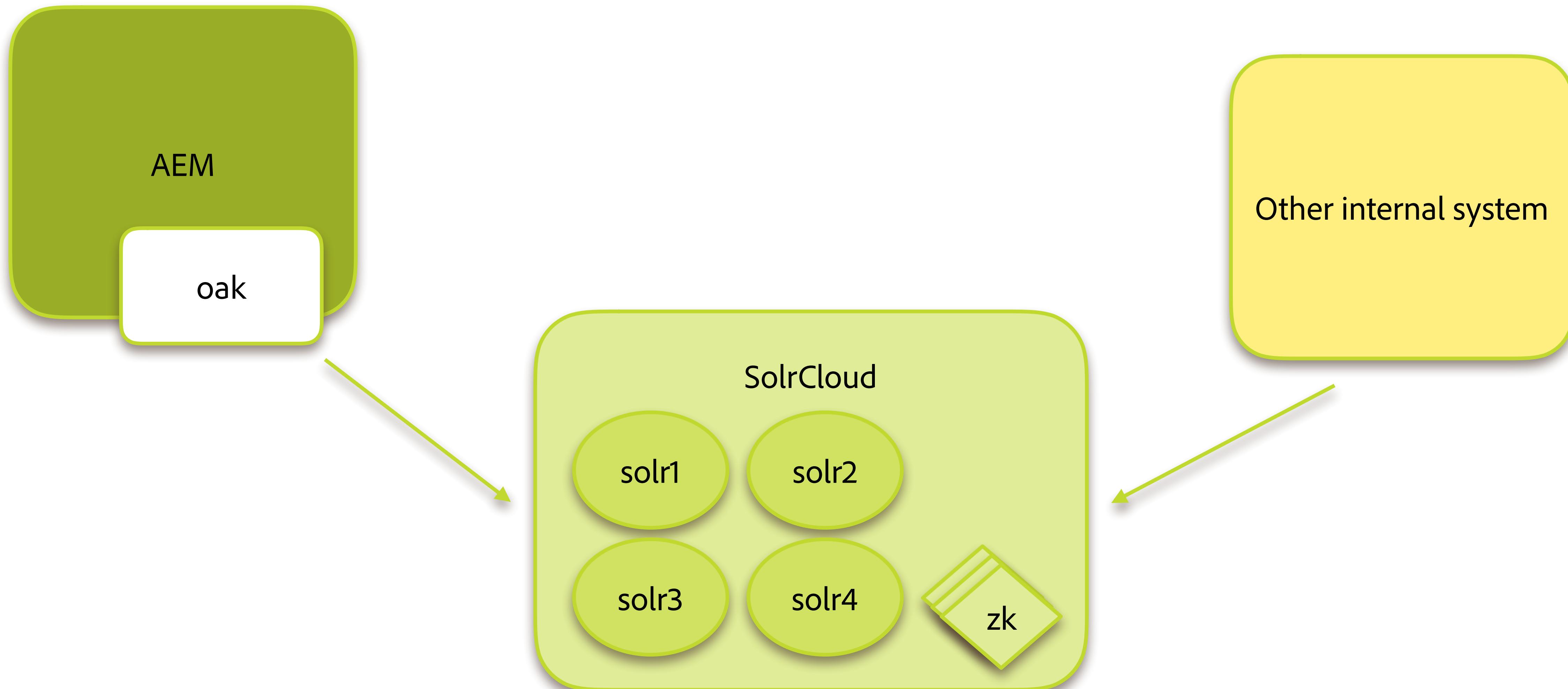
- By configuring Solr clusters
  - Custom indexing and searching via Solr plugins
    - NLP tasks via (e.g. OpenNLP / UIMA plugins)
    - named entity indexing
    - keyword / tags indexing
    - natural language search
  - Relevance tuning
    - query elevation
    - various Solr query parsers
- Via **Oak Native Language Support**
  - search time boosting
  - geospatial search
  - Solr local query params
  - ...

The screenshot shows the AEM Query tool interface. At the top, there are fields for 'Type' (set to 'XPath'), 'Path' (set to '/'), and 'Text'. Below these is a 'Query' field containing the XPath expression: `//*[rep:/native('solr', '(hike^3 OR die')')]`. There are three buttons at the bottom: 'Generate', 'Execute', and 'Show result' (which is checked). The main area displays a table titled 'Path' with four rows of results:

	Path
1	/var/commerce/products/we-retail/me/pants/mehisutrs
2	/libs/settings/community/templates/subscriptions-email/default/social.qna.components.hbs.post/de/jcr:content
3	/libs/settings/community/templates/email/html/social.qna.components.hbs.post/de/jcr:content
4	/content/we-retail/ca/en/experience/hours-of-wilderness/jcr:content/root/responsivegrid/contentfragment

At the bottom of the results table, it says 'Execution Info: 94 results (27msec)'

# Integration with / reuse existing customer infrastructure



# Solr performance bits - Search performance stats

/select

class: org.apache.solr.handler.component.SearchHandler  
description: Search using components: query,facet,facet\_module,mlt,highlight,stats,expand,debug,

src:

version: 5.5.5

---

stats: 15minRateReqsPerSecond: 0.0007996636229865102  
5minRateReqsPerSecond: 0.0000015668443405330119  
75thPcRequestTime: 8.67007825  
95thPcRequestTime: 31.683060399999768  
999thPcRequestTime: 110.194317  
99thPcRequestTime: 110.194317  
avgRequestsPerSecond: 0.00705868524501249  
avgTimePerRequest: 8.120410868421052  
errors: 0  
handlerStart: 1531836136106  
medianRequestTime: 2.3427214999999997  
requests: 38  
timeouts: 0  
totalTime: 308.575613

# Solr general setup rules

- Ideally AEM and Solr should be sitting on a same low latency network
- SolrCloud should always be preferred to single remote Solr instance
- Embedded Solr server should never be used beyond development environment
- Use persisted index definition as:
  - allows for indexing only specific subtrees
  - if needed, it allows for multiple Oak Solr indexes to be used for different portions of the index
  - registering Oak Solr index via OSGi would index (almost) all repository
- A starting setup for a SolrCloud cluster should always have 2 shards with 2 replicas each
- Avoid changing the Oak Solr schema and solrconfig unless you're sure about what you're doing

# Solr performance bits - Indexing performance stats

## /update

class: org.apache.solr.handler.UpdateRequestHandler

description: Add documents using XML (with XSLT), CSV, JSON, or javabin

src:

version: 5.5.5

stats: 15minRateReqsPerSecond: 0.6472559660946228

5minRateReqsPerSecond: 0.00013989022320941766

75thPcRequestTime: 1.43678525

95thPcRequestTime: 5.289718149999996

999thPcRequestTime: 259.1163314600003

99thPcRequestTime: 13.450827660000014

avgRequestsPerSecond: 20.23469612970971

avgTimePerRequest: 0.8324854031285285

errors: 0

handlerStart: 1531836136077

medianRequestTime: 0.8023625

requests: 108933

timeouts: 0

totalTime: 90685.132419

# Solr performance bits - Cache performance stats

## queryResultCache

class: org.apache.solr.search.FastLRUCache  
description: Concurrent LRU Cache(maxSize=40960, initialSize=4096, minSize=36864, acceptableSize=38912, cleanupThread=false, autowarmCount=2048, regenerator=org.apache.solr.search.SolrIndexSearcher\$3@277a20ac)  
src:

version: 1.0

stats: cumulative\_evictions: 0  
cumulative\_hitratio: 0.82  
cumulative\_hits: 27  
cumulative\_inserts: 10  
cumulative\_lookups: 33  
evictions: 0  
hitratio: 0.82  
hits: 27  
inserts: 10  
lookups: 33  
size: 6  
warmupTime: 0

## documentCache

class: org.apache.solr.search.LRUCache  
description: LRU Cache(maxSize=40960, initialSize=1024)  
src:

version: 1.0

stats: cumulative\_evictions: 0  
cumulative\_hitratio: 0.94  
cumulative\_hits: 1074  
cumulative\_inserts: 70  
cumulative\_lookups: 1144  
evictions: 0  
hitratio: 0.94  
hits: 1074  
inserts: 70  
lookups: 1144  
size: 70  
warmupTime: 0

# No **one** index to rule them all

Query Performance ▾				
	Slow Queries	<u>Popular Queries</u>	Explain Query	
	Last Execution	Execution Count / Language	Statement	Duration
<input type="checkbox"/>	2018-07-18 12:33:00	2 XPATH	(/jcr:root/content/dam//element(*, dam:Asset)[(jcr:contains(., 'winter'))]   /jcr:root/content/dam//element(*, nt:folder) [(jcr:contains(., 'winter'))])	<div style="width: 80%; background-color: #0072bc; height: 10px;"></div> 679972 ms

- Solr index can also be used
  - as a property index
  - to index a specific subtree
    - for subtree indexing Solr index configuration needs to be persisted

The screenshot shows the CRXDE Lite interface. On the left, there is a tree view of the repository structure under the 'content' folder, including 'apps', 'bin', 'conf', 'content', 'rep:policy', 'dam', 'assetinsights', 'catalogs', 'collections', 'dam:batch', 'demo', 'formsanddocuments', 'formsanddocuments-theme', 'jer:content', 'oak:index', 'solrcloud', 'server', 'projects', 'rep:policy', 'templates', and 'we-retail'. On the right, there is a search bar with the placeholder 'Enter search term to search the repository'. Below the search bar, there is a 'Properties' table with the following data:

Properties		Access Control	Replication
Name	Type	Value	
1	String	async	
2	Name	oak:Qu	
3	Boolean	false	
4	Long	3	
5	String	solr	

# Security

- Solr is not meant to be exposed to public traffic
  - see SolrSecurity Wiky
    - <https://wiki.apache.org/solr/SolrSecurity>
    - Apache Solr Ref Guide 5.5.5
      - <http://archive.apache.org/dist/lucene/solr/ref-guide/apache-solr-ref-guide-5.5.pdf>

# Q & A

This is an experimental UI. Report bugs [here](#). For the old UI click [here](#) ⓘ

Request-Handler (qt)  
`/select`

common  
q  
`winter`

fq  
 - +

sort

start, rows  
`0` `10`

fl

df

Raw Query Parameters  
`key1=val1&key2=val2`

wt  
`json`

indent

debugQuery

dismax

<http://localhost:7574/solr/oak/select?indent=on&q=winter&wt=json>

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 17,
    "version": 1606247226996162560
  },
  "response": {
    "numFound": 286, "start": 0, "maxScore": 2.9745665, "docs": [
      {
        "path_exact": "/content/dam/we-retail/en/products/activities/winter-sports/jcr:content",
        "suggest": ["Winter Sports"],
        "version": 1606247226996162560
      },
      {
        "path_exact": "/conf/we-retail/settings/wcm/segments/winter/jcr:content",
        "suggest": ["Winter"],
        "version": 1606246822869729280
      },
      {
        "path_exact": "/content/cq:tags/we-retail/season/winter",
        "suggest": ["Winter", ""],
        "version": 1606247184773152768
      },
      {
        "path_exact": "/var/commerce/products/we-retail/eq/winter-sports",
        "suggest": ["Winter Sports"],
        "version": 1606247681359872000
      },
      {
        "path_exact": "/jcr:system/jcr:versionStorage/58/a6/e9/58a6e9e5-88b6-4e7f-a5e5-256ee",
        "version": 1606247038513577984
      },
      {
        "path_exact": "/jcr:system/jcr:versionStorage/b2/fc/54/b2fc54f2-2ee9-49d2-8de1-187af",
        "version": 1606246981140742144
      }
    ]
  }
}
```



Adobe