Oak Lucene Indexes

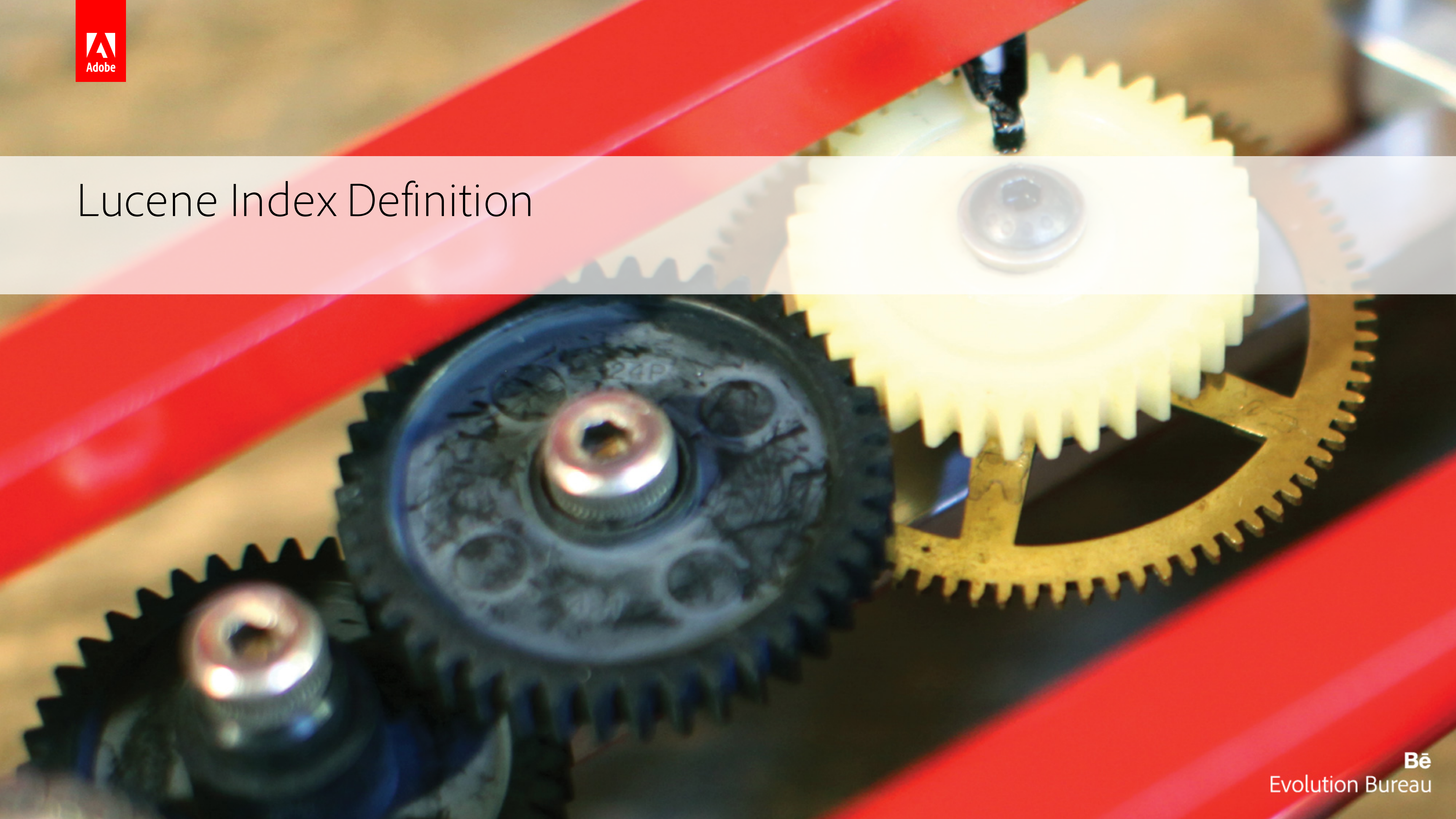
Chetan Mehrotra | Senior Computer Scientist

Alex Parvulescu | Senior Developer

Bē
Evolution Bureau

Content

- Lucene Index Definitions
- Anatomy of a Query (Restrictions, Sorting, Aggregation)
- Query Diagnostics and Troubleshooting
- Lucene Index Internals (Oak Directory, JMX, Luke)
- Asynchronous Indexing
- Q&A



Lucene Index Definition

Index Definition

- Stored under **oak:index** node
- **Define** how content gets indexed
- type **`oak:QueryIndexDefinition`**
- Required properties
  - compatVersion = 2
  - type = “lucene”
  - async = “async”

```
/oak:index/assetType (oak:QueryIndexDefinition)
  - compatVersion = 2
  - type = "lucene"
  - async = "async"
  + indexRules (nt:unstructured)
    + dam:Asset
      + properties (nt:unstructured)
        + assetType
          - propertyIndex = true
          - name = "jcr:content/metadata/type"
```

Index Definition – Index Rules

- Defines which types of node and properties are indexed
- Rules are defined per nodeType
- Rule consist of one or more property definitions
- Index selected based on match between type used in Query and presence of indexRule for that type
- Multiple indexRules in same index
- Order important – nodeType matching honors inheritance

```
SELECT  *
FROM  [dam:Asset] AS a
WHERE ISDESCENDANTNODE([/content/en])
      AND a.[jcr:content/metadata/type] = 'image'
```

```
/oak:index/assetType (oak:QueryIndexDefinition)
  - compatVersion = 2
  - type = "lucene"
  - async = "async"
  + indexRules (nt:unstructured)
    + dam:Asset
      + properties (nt:unstructured)
        + assetType
          - propertyIndex = true
          - name = "jcr:content/metadata/assetType"
```

https://jackrabbit.apache.org/oak/docs/query/lucene.html#Indexing_Rules

Index Definition – Property Definitions

- Defines how a property gets indexed
- One or more property definition per indexRule
- Definition mapping done based on matching property name or regex pattern
- Supports relative property name by there relative paths
- Order important (if regex are used)

```
SELECT  *
FROM   [dam:Asset] AS a
WHERE  ISDESCENDANTNODE([/content/en])
       AND a.[jcr:content/metadata/type] = 'image'
```

```
/oak:index/assetType (oak:QueryIndexDefinition)
  - compatVersion = 2
  - type = "lucene"
  - async = "async"
  + indexRules (nt:unstructured)
    + dam:Asset
      + properties (nt:unstructured)
        + assetType
          - propertyIndex = true
          - name = "jcr:content/metadata/assetType"
```

https://jackrabbit.apache.org/oak/docs/query/lucene.html#Property_Definitions

Index Definition – Best Practices

- **Precise Index Definition** - That indexes just the right amount of content based on your query requirement. Precise index is happy index!

- **Make use of nodetype to achieve a "*cohesive* index"** - This would allow multiple queries to make use of same index and also evaluation of multiple property restrictions natively in Lucene

- **For people familiar with Relational Databases** - Nodetype is your Table in your DB and all the direct or relative properties as columns in that table. Various property definitions are like indexes on those columns.

https://jackrabbit.apache.org/oak/docs/query/lucene.html#Design_Considerations

Sample Content to Query Against

```
/content/dam/assets/december/banner.png (dam:Asset)
    + metadata (dam:AssetContent)
      - dc:format = "image/png"
      - status = "published"
      - jcr:lastModified = "2009-10-9T21:52:31"
      - app:tags = ["properties:orientation/landscape",
                    "marketing:interest/product"]
      - size = 450
      - comment = "Image for december launch"
      - jcr:title = "December Banner"
      + xmpMM:History
        + 1
          - softwareAgent = "Adobe Photoshop"
          - author = "David"
    + renditions (nt:folder)
      + original (nt:file)
        + jcr:content
          - jcr:data = ...
```

Anatomy of Query

```
SELECT
  *
FROM [dam:Asset] AS a
WHERE ISDESCENDANTNODE([/content/public/platform])
  AND a.[jcr:content/metadata/status] = 'published'
  AND CONTAINS([jcr:content/metadata/comment], 'december')
ORDER BY
  a.[jcr:content/metadata/jcr:lastModified] DESC
```

- Nodetype restriction on **dam:Asset**
- Path restriction on **/content/public/platform**
- Property restriction on **jcr:content/metadata/status**
- Fulltext property restriction on **jcr:content/metadata/comment**
- Sorting done on **jcr:content/metadata/jcr:lastModified**

Nodetype Restrictions

```
SELECT
  *
FROM
    [dam:Asset] AS a
WHERE
    ISDESCENDANTNODE([/content/public/platform])
  AND
    a.[jcr:content/metadata/status] = 'published'
  AND
    CONTAINS([jcr:content/metadata/comment], 'december')
ORDER BY
  a.[jcr:content/metadata/jcr:lastModified] DESC
```

```
/oak:index/damAsset (oak:QueryIndexDefinition)
  - compatVersion = 2
  - type = "lucene"
  - async = "async"
  + indexRules (nt:unstructured)
    + dam:Asset (nt:unstructured)
      + properties
        ...
```

Create index definition node at /oak:index/damAsset with **indexRule** for **dam:Asset**

Path Restriction

```
SELECT
  *
FROM
    [dam:Asset] AS a
WHERE
    ISDESCENDANTNODE([/content/public/platform])
  AND
    a.[jcr:content/metadata/status] = 'published'
  AND
    CONTAINS([jcr:content/metadata/comment], 'december')
ORDER BY
  a.[jcr:content/metadata/jcr:lastModified] DESC
```

```
/oak:index/damAsset (oak:QueryIndexDefinition)
  - compatVersion = 2
  - type = "lucene"
  - async = "async"
  - evaluatePathRestrictions = true
  + indexRules (nt:unstructured)
    + dam:Asset (nt:unstructured)
      + properties
        ...
```

Enable **evaluatePathRestrictions** for indexing paths

Bonus Tip – If all indexable content is under /content/public and query always specify the path restriction then it would be better to define index definition under **/content/public/oak:index** (more details)

Property Restriction

```
SELECT
  *
FROM
    [dam:Asset] AS a
WHERE
    ISDESCENDANTNODE([/content/public/platform])
  AND
    a.[jcr:content/metadata/status] = 'published'
  AND
    CONTAINS([jcr:content/metadata/comment], 'december')
ORDER BY
  a.[jcr:content/metadata/jcr:lastModified] DESC
```

```
/oak:index/damAsset (oak:QueryIndexDefinition)
  - compatVersion = 2
  - type = "lucene"
  - async = "async"
  - evaluatePathRestrictions = true
  + indexRules (nt:unstructured)
    + dam:Asset (nt:unstructured)
      + properties
        + status
          - propertyIndex = true
          - name = "jcr:content/metadata/status"
```

Create property definition node with **propertyIndex** enabled and **name** set to **relative path of property**

Fulltext Property Restriction

```
SELECT
  *
FROM
    [dam:Asset] AS a
WHERE
    ISDESCENDANTNODE([/content/public/platform])
  AND
    a.[jcr:content/metadata/status] = 'published'
  AND
    CONTAINS([jcr:content/metadata/comment], 'december')
ORDER BY
  a.[jcr:content/metadata/jcr:lastModified] DESC
```

```
/oak:index/damAsset (oak:QueryIndexDefinition)
  - compatVersion = 2
  - type = "lucene"
  - async = "async"
  - evaluatePathRestrictions = true
  + indexRules (nt:unstructured)
    + dam:Asset (nt:unstructured)
      + properties
        + status
          - propertyIndex = true
          - name = "jcr:content/metadata/status"
        + comment
          - name = "jcr:content/metadata/comment"
          - analyzed = true
```

Create property definition node with **analyzed** enabled

Sorting

```
SELECT
  *
FROM
    [dam:Asset] AS a
WHERE
    ISDESCENDANTNODE([/content/public/platform])
  AND
    a.[jcr:content/metadata/status] = 'published'
  AND
    CONTAINS([jcr:content/metadata/comment], 'december')
ORDER BY
  a.[jcr:content/metadata/jcr:lastModified] DESC
```

```
/oak:index/damAsset (oak:QueryIndexDefinition)
  - compatVersion = 2
  - type = "lucene"
  - async = "async"
  - evaluatePathRestrictions = true
  + indexRules (nt:unstructured)
    + dam:Asset (nt:unstructured)
      + properties
        + status
          - propertyIndex = true
          - name = "jcr:content/metadata/status"
        + comment
          - name = "jcr:content/metadata/comment"
          - analyzed = true
        + lastModified
          - name = "jcr:content/metadata/jcr:lastModified"
          - ordered = true
          - type = Date
          - propertyIndex = true
```

Create property definition node with **ordered** enabled and **type** set to property type. Also enable **propertyIndex** if you plan to have some restrictions on it

Fulltext Node Restriction

```
SELECT * FROM [dam:Asset] WHERE CONTAINS(., 'christmas')
```

- Searches for ‘christmas’ in all nodes of type dam:Asset
- Fulltext index for a node is made up fulltext terms made up from
  - Node properties – Properties with **nodeScopeIndex** set to true
  - Properties of relative nodes defined by Aggregation Rules
- Aggregation Rules
  - Define path patterns for selecting the relative nodes
  - Are bound to specific type
  - Can be recursive – Relative path refers to nt:file and nt:file has its own aggregation rule defined
- For aggregated nodes all properties whose type are part of includePropertyTypes are included unless a property definition is defined with nodeScopeIndex=false

Fulltext - Aggregation

Content

```
/content/dam/assets/december/banner.png
(dam:Asset)
  + metadata (dam:AssetContent)
    - dc:format = "image/png"
    - status = "published"
    - jcr:lastModified = "2009-10-9T21:52:31"
    - app:tags =
          ["properties:orientation/landscape",
          "marketing:interest/product"]
    - size = 450
    - comment = "Image for Christmas launch"
    - jcr:title = "December Banner"
    + xmpMM:History
      + 1
        - softwareAgent = "Adobe Photoshop"
        - author = "David"
  + renditions (nt:folder)
    + original (nt:file)
      + jcr:content
        - jcr:data = ...
```

Aggregation Rules

```
+ aggregates
  + dam:Asset
    + include0
      - path = "jcr:content"
    + include1
      - path = "jcr:content/metadata"
    + include2
      - path = "jcr:content/metadata/*"
    + include3
      - path = "jcr:content/metadata/*/*"
    + include4
      - path = "jcr:content/renditions"
    + include5
      - path = "jcr:content/renditions/original"
  + nt:file
    + include0
      - path = "jcr:content"
```

Extracted Terms for Fulltext Index

```
image/png
Published
properties:orientation/landscape
marketing:interest/product
December Banner
Image for Christmas launch
```

```
Adobe Photoshop
David
```

https://jackrabbit.apache.org/oak/docs/query/lucene.html#Aggregation

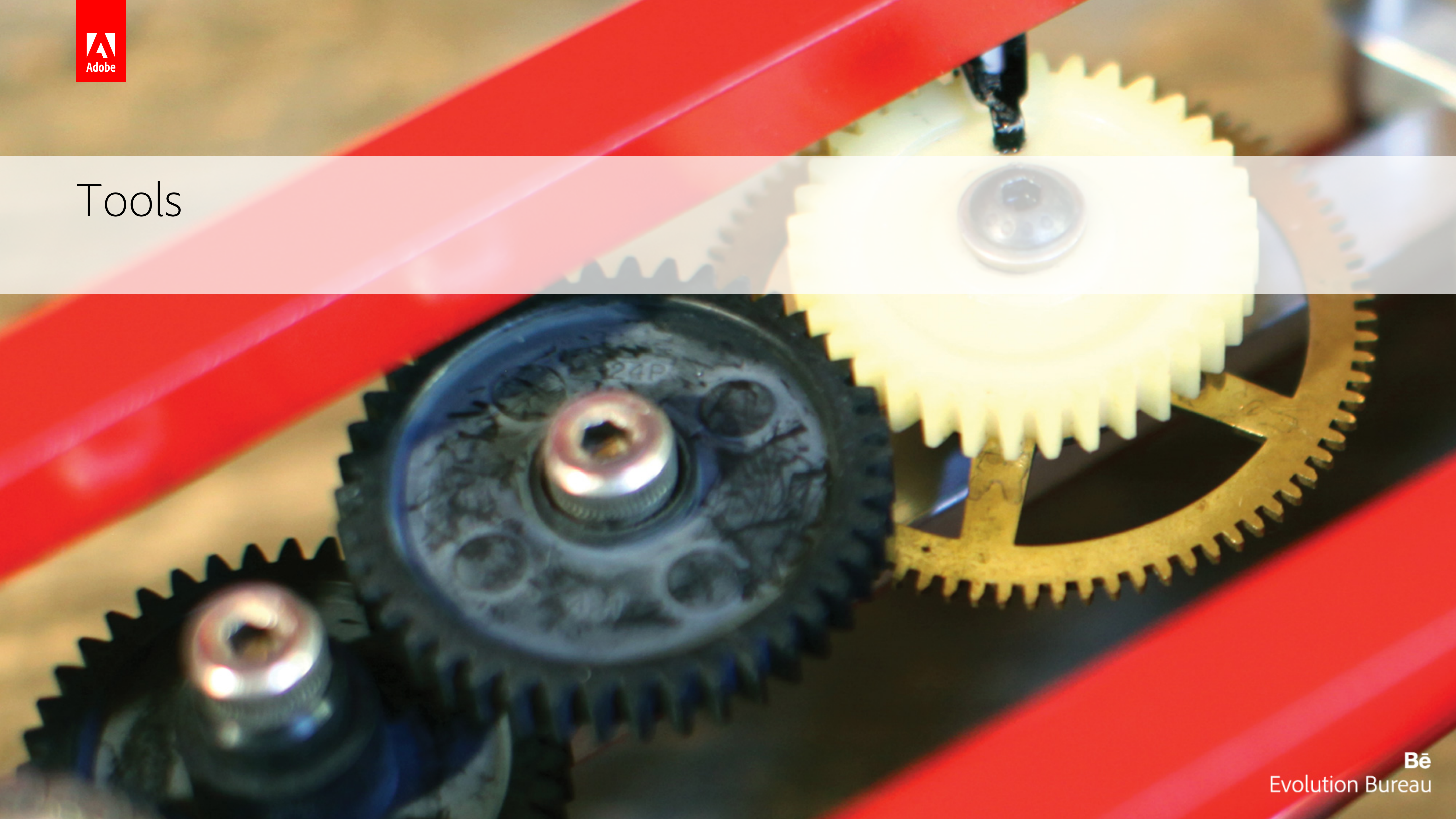
Query Result Size

- Oak Fast Result Size
  - By default NodeIterator.getSize() returns -1 if result is large as size estimate cost is O(n) due to ACL checks
  - ACL Checks can be relaxed (check first 'k' only). Enable via system property **oak.fastQuerySize.**
  - OSGi config support with next release

- AEM Query Builder and Pagination
  - Make use of **p.guessTotal** query parameter to avoid costly operation for determining result size
  - Use progressive pagination

Other Features

- Composing Analyzer – For configuring Stemming, Synonyms, Stop words etc
- Boost – Improving search relevancy
- Tika Config – Control how and which types of binary files are indexed
- Suggestions
- Spell Check
- Pre Extracting Text from Binaries – To speedup reindexing time for repositories having marge number of binaries having text



Tools

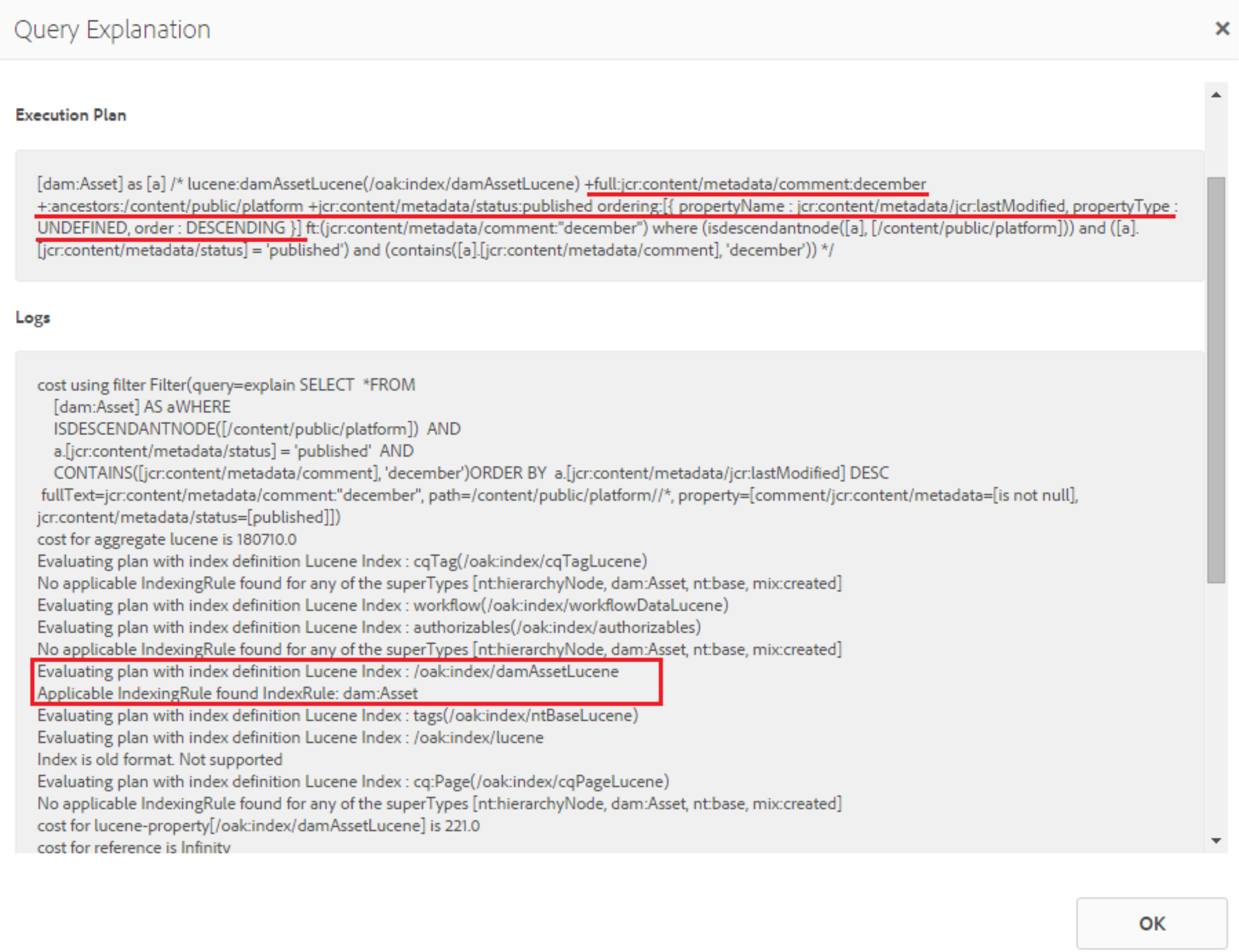
Evolution Bureau

Query Explain Tool

- Shipped with AEM 6.1
  - Tools -> Operations -> Dashboard -> Diagnosis -> Query Performance
  - http://localhost:4502/libs/granite/operations/content/diagnosis/tool.html/_granite_queryperformance
  - Shows Slow Query, Popular Query and Explain Query
- ACS Tools (more upto date)
  - https://adobe-consulting-services.github.io/acs-aem-tools/explain-query.html

Query Explain Tool

- Shows logs from various index consulted
- Shows the actual Lucene query fired
  - Path Restriction
    +:ancestors:/content/public/platform
  - Fulltext Restriction
    +full:jcr:content/metadata/comment:december
  - Property Restriction
    +jcr:content/metadata/status:published
  - Ordering



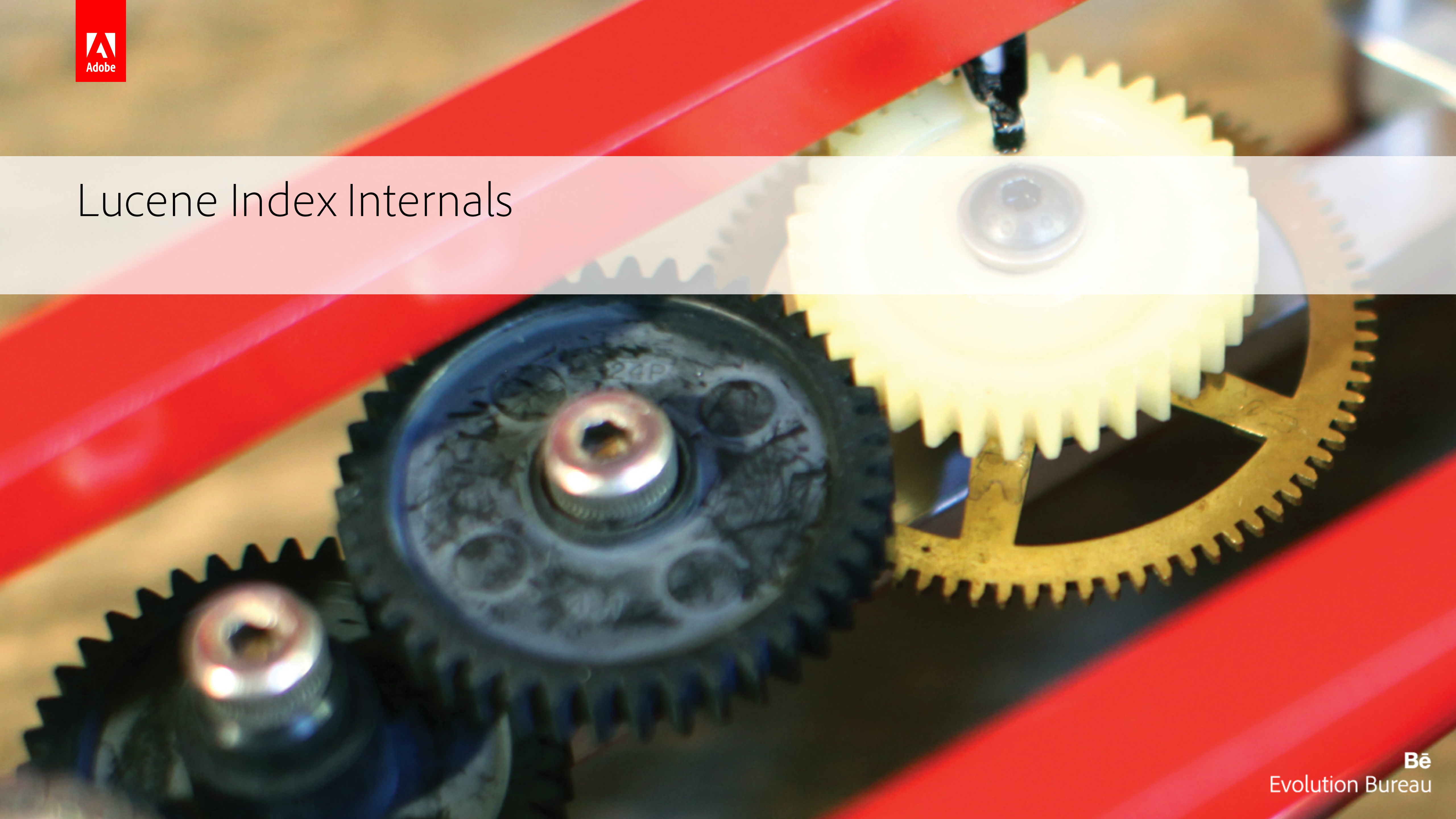
Execution Plan

```
[dam:Asset] as [a] /* lucene:damAssetLucene(/oak:index/damAssetLucene) +full:jcr:content/metadata/comment:december
+:ancestors:/content/public/platform +jcr:content/metadata/status:published ordering:[{ propertyName : jcr:content/metadata/jcr:lastModified, propertyType :
UNDEFINED, order : DESCENDING }] ft:(jcr:content/metadata/comment:"december") where (isdescendantnode([a], [/content/public/platform])) and ([a].
[jcr:content/metadata/status] = 'published') and (contains([a].[jcr:content/metadata/comment], 'december')) */
```

Logs

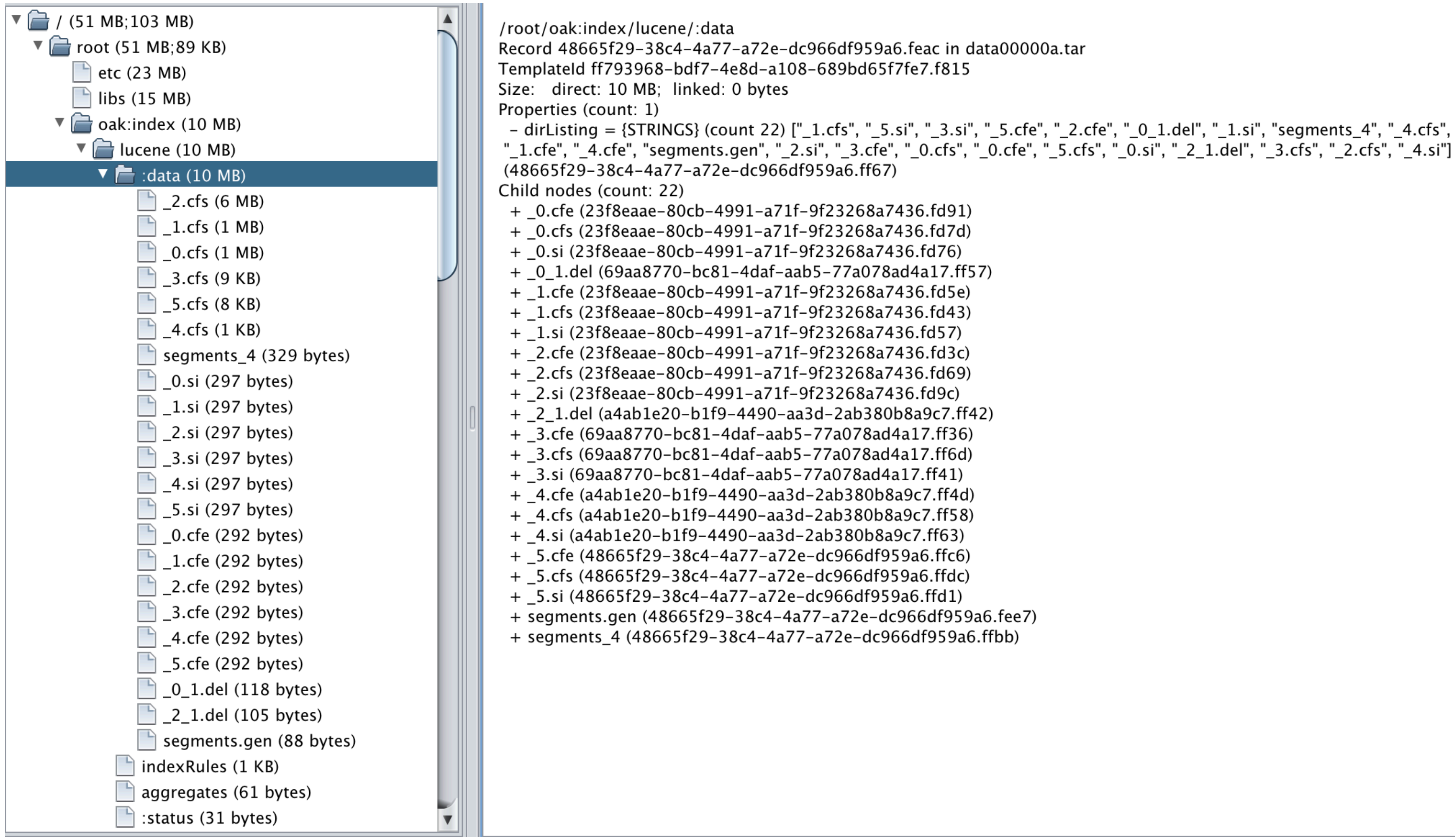
```
cost using filter Filter(query=explain SELECT  *FROM
   [dam:Asset] AS aWHERE
   ISDESCENDANTNODE([/content/public/platform])  AND
   a.[jcr:content/metadata/status] = 'published'  AND
   CONTAINS([jcr:content/metadata/comment], 'december')ORDER BY  a.[jcr:content/metadata/jcr:lastModified] DESC
 fullText=jcr:content/metadata/comment:"december", path=/content/public/platform//*, property=[comment/jcr:content/metadata=[is not null],
jcr:content/metadata/status=[published]])
cost for aggregate lucene is 180710.0
Evaluating plan with index definition Lucene Index : cqTag(/oak:index/cqTagLucene)
No applicable IndexingRule found for any of the superTypes [nt:hierarchyNode, dam:Asset, nt:base, mix:created]
Evaluating plan with index definition Lucene Index : workflow(/oak:index/workflowDataLucene)
Evaluating plan with index definition Lucene Index : authorizables(/oak:index/authorizables)
No applicable IndexingRule found for any of the superTypes [nt:hierarchyNode, dam:Asset, nt:base, mix:created]
Evaluating plan with index definition Lucene Index : /oak:index/damAssetLucene
Applicable IndexingRule found IndexRule: dam:Asset
Evaluating plan with index definition Lucene Index : tags(/oak:index/ntBaseLucene)
Evaluating plan with index definition Lucene Index : /oak:index/lucene
Index is old format. Not supported
Evaluating plan with index definition Lucene Index : cq:Page(/oak:index/cqPageLucene)
No applicable IndexingRule found for any of the superTypes [nt:hierarchyNode, dam:Asset, nt:base, mix:created]
cost for lucene-property[/oak:index/damAssetLucene] is 221.0
cost for reference is Infinity
```

Lucene Index Internals



Lucene Index Internals - Directory

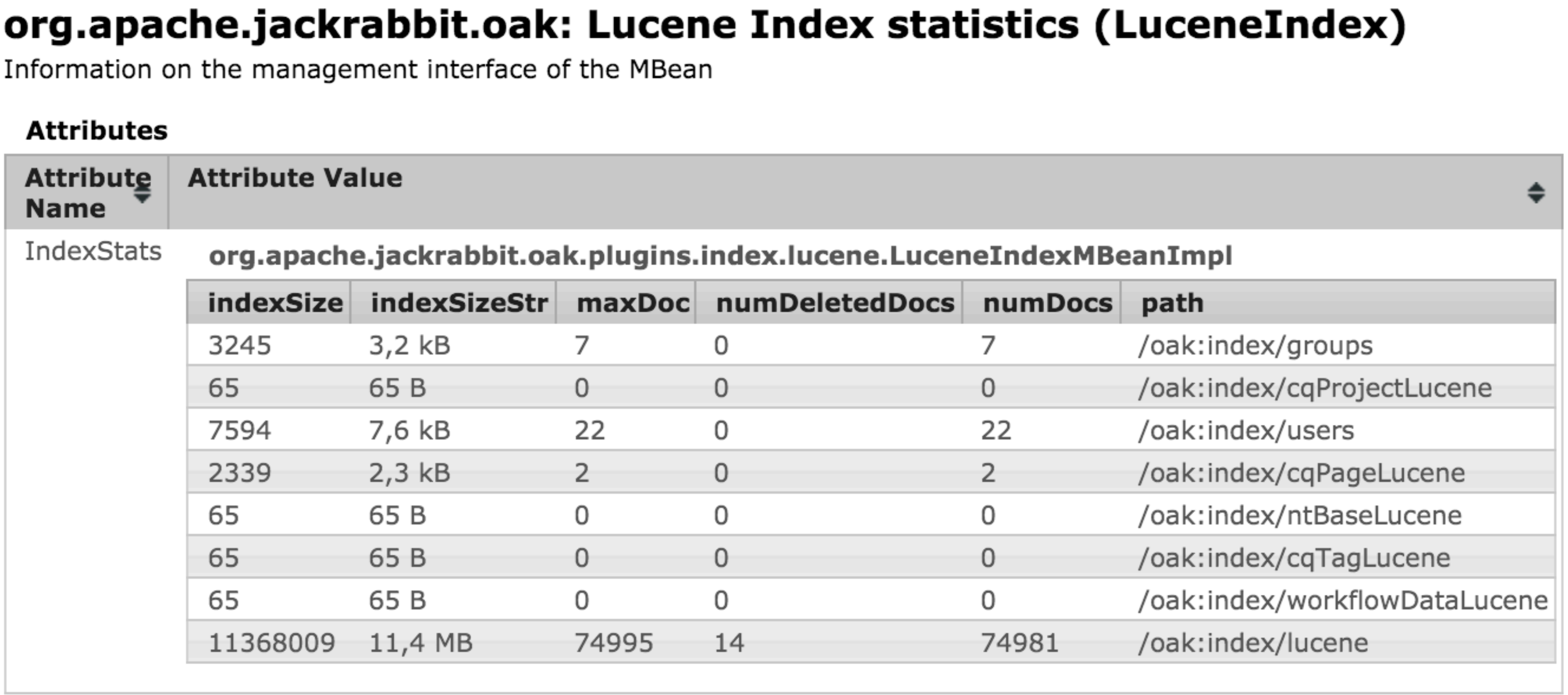
- Lucene Directory is stored in the repository (the source of truth)
- Copy on Read & Copy on Write maintain local copies for faster access (index content to disk location mappings are exposed via JMX)



```
▼ / (51 MB;103 MB)
  ▼ root (51 MB;89 KB)
      etc (23 MB)
      libs (15 MB)
    ▼ oak:index (10 MB)
      ▼ lucene (10 MB)
        ▼ :data (10 MB)
            _2.cfs (6 MB)
            _1.cfs (1 MB)
            _0.cfs (1 MB)
            _3.cfs (9 KB)
            _5.cfs (8 KB)
            _4.cfs (1 KB)
            segments_4 (329 bytes)
            _0.si (297 bytes)
            _1.si (297 bytes)
            _2.si (297 bytes)
            _3.si (297 bytes)
            _4.si (297 bytes)
            _5.si (297 bytes)
            _0.cfe (292 bytes)
            _1.cfe (292 bytes)
            _2.cfe (292 bytes)
            _3.cfe (292 bytes)
            _4.cfe (292 bytes)
            _5.cfe (292 bytes)
            _0_1.del (118 bytes)
            _2_1.del (105 bytes)
            segments.gen (88 bytes)
          indexRules (1 KB)
          aggregates (61 bytes)
          :status (31 bytes)
```

```
/root/oak:index/lucene/:data
Record 48665f29-38c4-4a77-a72e-dc966df959a6.feac in data00000a.tar
TemplateId ff793968-bdf7-4e8d-a108-689bd65f7fe7.f815
Size:   direct: 10 MB;  linked: 0 bytes
Properties (count: 1)
  - dirListing = {STRINGS} (count 22) ["_1.cfs", "_5.si", "_3.si", "_5.cfe", "_2.cfe", "_0_1.del", "_1.si", "segments_4", "_4.cfs",
 "_1.cfe", "_4.cfe", "segments.gen", "_2.si", "_3.cfe", "_0.cfs", "_0.cfe", "_5.cfs", "_0.si", "_2_1.del", "_3.cfs", "_2.cfs", "_4.si"]
 (48665f29-38c4-4a77-a72e-dc966df959a6.ff67)
Child nodes (count: 22)
  + _0.cfe (23f8eaae-80cb-4991-a71f-9f23268a7436.fd91)
  + _0.cfs (23f8eaae-80cb-4991-a71f-9f23268a7436.fd7d)
  + _0.si (23f8eaae-80cb-4991-a71f-9f23268a7436.fd76)
  + _0_1.del (69aa8770-bc81-4daf-aab5-77a078ad4a17.ff57)
  + _1.cfe (23f8eaae-80cb-4991-a71f-9f23268a7436.fd5e)
  + _1.cfs (23f8eaae-80cb-4991-a71f-9f23268a7436.fd43)
  + _1.si (23f8eaae-80cb-4991-a71f-9f23268a7436.fd57)
  + _2.cfe (23f8eaae-80cb-4991-a71f-9f23268a7436.fd3c)
  + _2.cfs (23f8eaae-80cb-4991-a71f-9f23268a7436.fd69)
  + _2.si (23f8eaae-80cb-4991-a71f-9f23268a7436.fd9c)
  + _2_1.del (a4ab1e20-b1f9-4490-aa3d-2ab380b8a9c7.ff42)
  + _3.cfe (69aa8770-bc81-4daf-aab5-77a078ad4a17.ff36)
  + _3.cfs (69aa8770-bc81-4daf-aab5-77a078ad4a17.ff6d)
  + _3.si (69aa8770-bc81-4daf-aab5-77a078ad4a17.ff41)
  + _4.cfe (a4ab1e20-b1f9-4490-aa3d-2ab380b8a9c7.ff4d)
  + _4.cfs (a4ab1e20-b1f9-4490-aa3d-2ab380b8a9c7.ff58)
  + _4.si (a4ab1e20-b1f9-4490-aa3d-2ab380b8a9c7.ff63)
  + _5.cfe (48665f29-38c4-4a77-a72e-dc966df959a6.ffc6)
  + _5.cfs (48665f29-38c4-4a77-a72e-dc966df959a6.ffdc)
  + _5.si (48665f29-38c4-4a77-a72e-dc966df959a6.ffd1)
  + segments.gen (48665f29-38c4-4a77-a72e-dc966df959a6.fee7)
  + segments_4 (48665f29-38c4-4a77-a72e-dc966df959a6.ffbb)
```

Lucene Index Internals - JMX



org.apache.jackrabbit.oak: Lucene Index statistics (LuceneIndex)

Information on the management interface of the MBean

Attributes

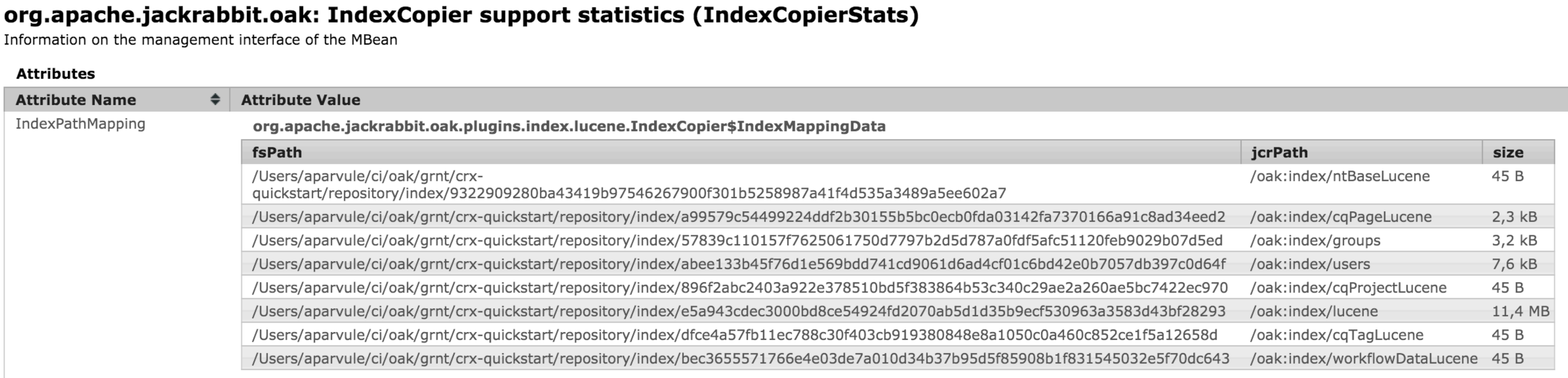
Attribute Name	Attribute Value
IndexStats	org.apache.jackrabbit.oak.plugins.index.lucene.LuceneIndexMBeanImpl

indexSize	indexSizeStr	maxDoc	numDeletedDocs	numDocs	path
3245	3,2 kB	7	0	7	/oak:index/groups
65	65 B	0	0	0	/oak:index/cqProjectLucene
7594	7,6 kB	22	0	22	/oak:index/users
2339	2,3 kB	2	0	2	/oak:index/cqPageLucene
65	65 B	0	0	0	/oak:index/ntBaseLucene
65	65 B	0	0	0	/oak:index/cqTagLucene
65	65 B	0	0	0	/oak:index/workflowDataLucene
11368009	11,4 MB	74995	14	74981	/oak:index/lucene

org.apache.jackrabbit.oak: Lucene Index statistics (LuceneIndex)

- provides a listing of the existing lucene indexes
- http://localhost:4502/system/console/jmx/org.apache.jackrabbit.oak%3Aname%3DLucene+Index+statistics%2Ctype%3DLuceneIndex

Lucene Index Internals - JMX continued



org.apache.jackrabbit.oak: IndexCopier support statistics (IndexCopierStats)

Information on the management interface of the MBean

Attributes

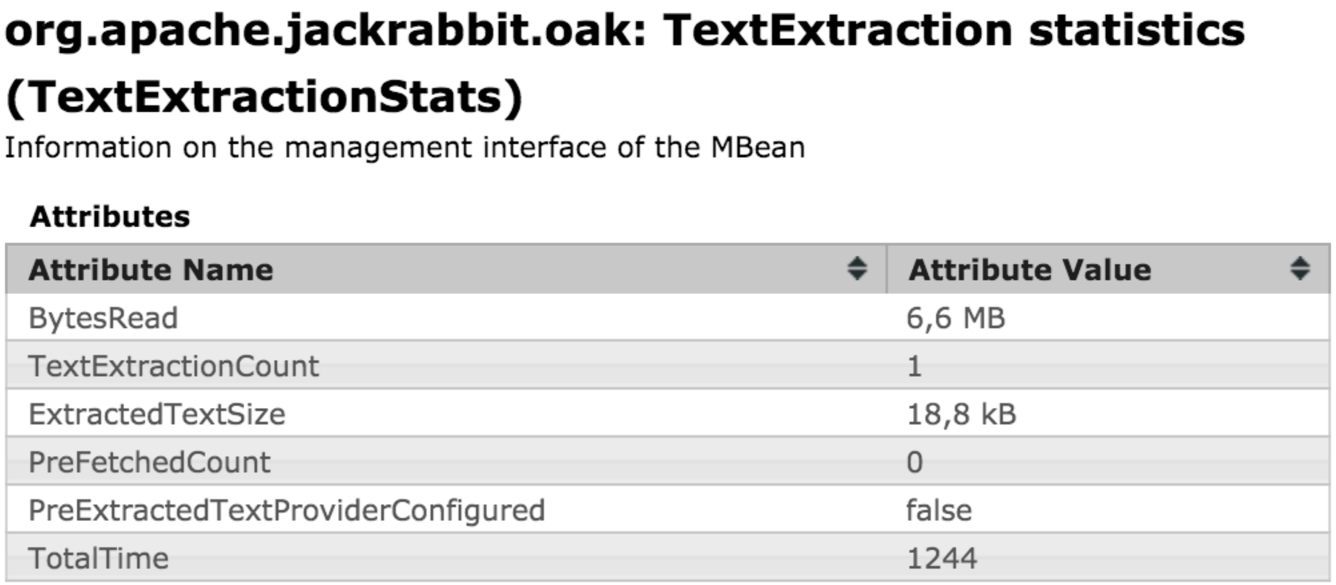
Attribute Name	Attribute Value
IndexPathMapping	**org.apache.jackrabbit.oak.plugins.index.lucene.IndexCopier$IndexMappingData**

fsPath	jcrPath	size
/Users/aparvule/ci/oak/grnt/crx-quickstart/repository/index/9322909280ba43419b97546267900f301b5258987a41f4d535a3489a5ee602a7	/oak:index/ntBaseLucene	45 B
/Users/aparvule/ci/oak/grnt/crx-quickstart/repository/index/a99579c54499224ddf2b30155b5bc0ecb0fda03142fa7370166a91c8ad34eed2	/oak:index/cqPageLucene	2,3 kB
/Users/aparvule/ci/oak/grnt/crx-quickstart/repository/index/57839c110157f7625061750d7797b2d5d787a0fdf5afc51120feb9029b07d5ed	/oak:index/groups	3,2 kB
/Users/aparvule/ci/oak/grnt/crx-quickstart/repository/index/abee133b45f76d1e569bdd741cd9061d6ad4cf01c6bd42e0b7057db397c0d64f	/oak:index/users	7,6 kB
/Users/aparvule/ci/oak/grnt/crx-quickstart/repository/index/896f2abc2403a922e378510bd5f383864b53c340c29ae2a260ae5bc7422ec970	/oak:index/cqProjectLucene	45 B
/Users/aparvule/ci/oak/grnt/crx-quickstart/repository/index/e5a943cdec3000bd8ce54924fd2070ab5d1d35b9ecf530963a3583d43bf28293	/oak:index/lucene	11,4 MB
/Users/aparvule/ci/oak/grnt/crx-quickstart/repository/index/dfce4a57fb11ec788c30f403cb919380848e8a1050c0a460c852ce1f5a12658d	/oak:index/cqTagLucene	45 B
/Users/aparvule/ci/oak/grnt/crx-quickstart/repository/index/bec3655571766e4e03de7a010d34b37b95d5f85908b1f831545032e5f70dc643	/oak:index/workflowDataLucene	45 B

org.apache.jackrabbit.oak: IndexCopier support statistics (IndexCopierStats)

- Copy on Read and Copy on Write related stats, of interest is the mapping between index content and location on disk
- http://localhost:4502/system/console/jmx/org.apache.jackrabbit.oak%3Aname%3DIndexCopier+support+statistics%2Ctype%3DIndexCopierStats

Lucene Index Internals - JMX continued



org.apache.jackrabbit.oak: TextExtraction statistics (TextExtractionStats)

Information on the management interface of the MBean

Attributes

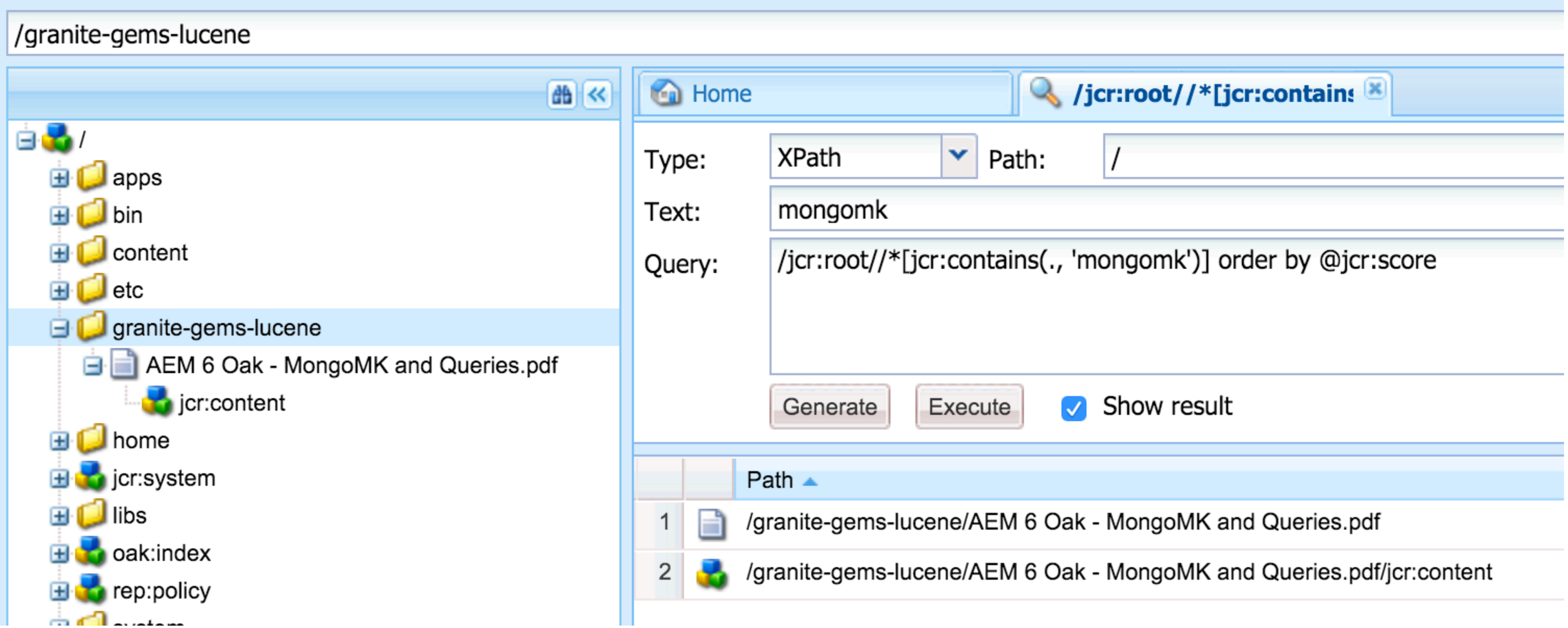
Attribute Name	Attribute Value
BytesRead	6,6 MB
TextExtractionCount	1
ExtractedTextSize	18,8 kB
PreFetchedCount	0
PreExtractedTextProviderConfigured	false
TotalTime	1244

org.apache.jackrabbit.oak: TextExtraction statistics (TextExtractionStats)

- Very relevant stats related to how much work is done extracting text from binaries
- http://localhost:4502/system/console/jmx/org.apache.jackrabbit.oak%3Aname%3DTextExtraction+statistics%2Ctype%3DTextExtractionStats
- *Make sure you remember this one for our experiment with 'Luke'*

Lucene Index Internals - Luke

Let's run a small experiment: upload a pdf file to the repository, verify if full-text search works


/granite-gems-lucene

Type: XPath

Path: /

Text: mongomk

Query: /jcr:root//*[jcr:contains(., 'mongomk')] order by @jcr:score

☑ Show result

		Path
1		/granite-gems-lucene/AEM 6 Oak - MongoMK and Queries.pdf
2		/granite-gems-lucene/AEM 6 Oak - MongoMK and Queries.pdf/jcr:content

It works! And now let's see why…

Lucene Index Internals - Luke

Setting up 'Luke' to look at the Lucene index:

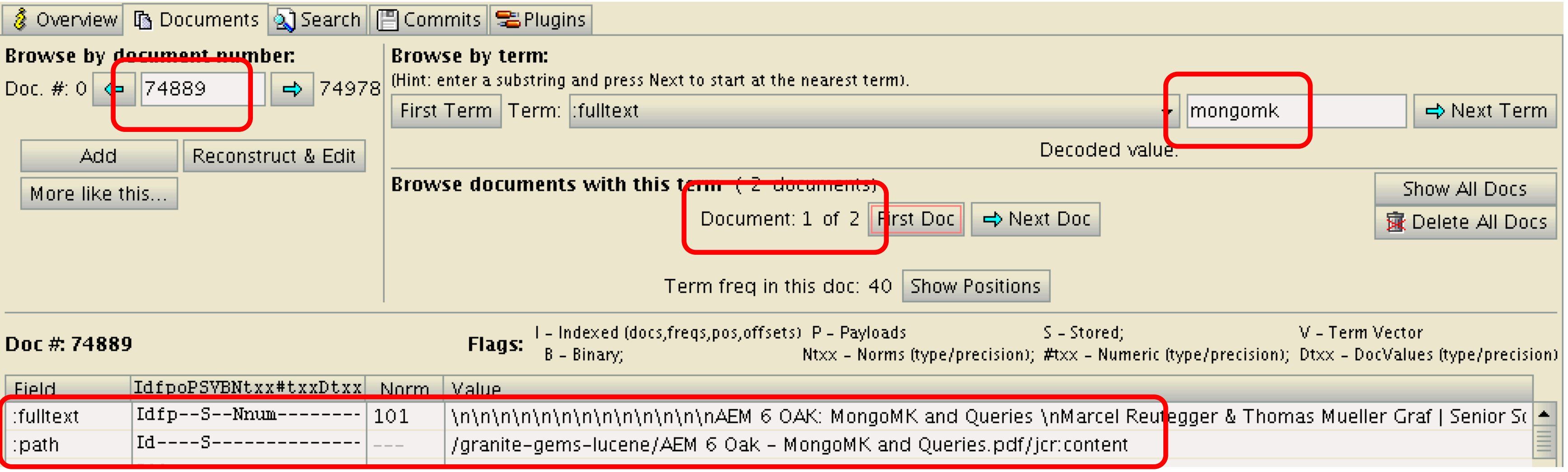
1. *Why Luke?* Luke is a dedicated Lucene index tool, has no alternatives for viewing content

2. Identify which index you want to look at

3. Export Index Contents
   - (easy/online) Lookup the Copy on Read mappings in the JMX console and grab a copy of the index
   - (harder/possibly offline) Use the oak console to export the index to a specific location

4. Open 'Luke' and make sure you pass in the oak-lucene jar as a classpath entry (as documented on the docs)

Lucene Index Internals - Luke



Browse by document number:

Doc. #: 0 74889 74978

Browse by term:

(Hint: enter a substring and press Next to start at the nearest term).

Term: :fulltext mongomk

Decoded value:

Browse documents with this term (2 documents)

Document: 1 of 2

Term freq in this doc: 40

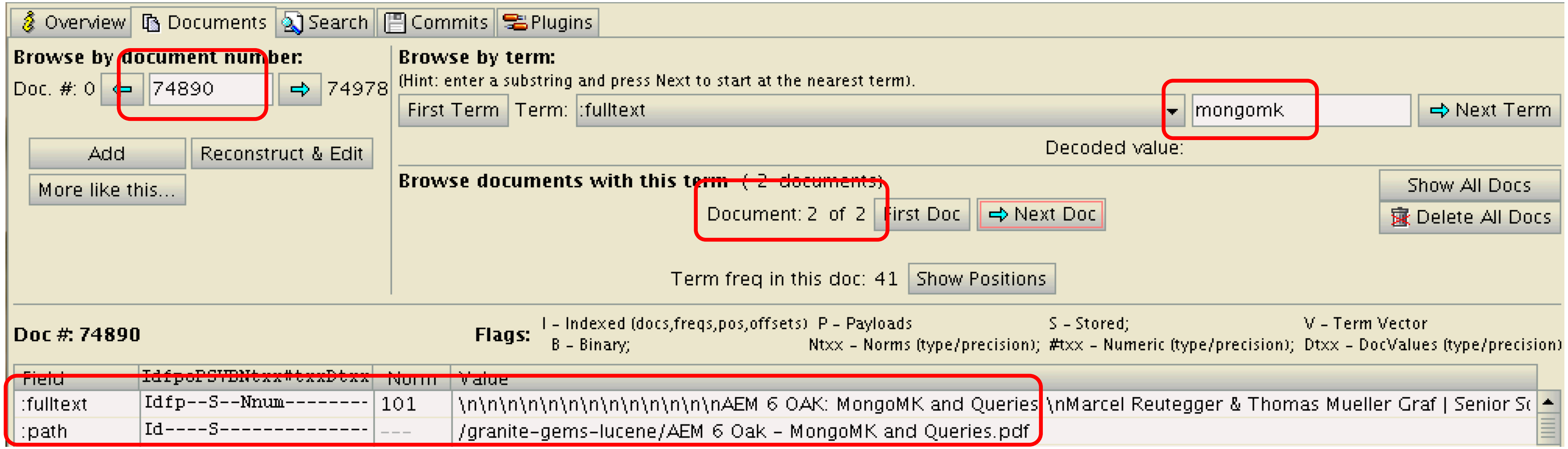
Doc #: 74889

Flags: I – Indexed (docs,freqs,pos,offsets) P – Payloads S – Stored; V – Term Vector B – Binary; Ntxx – Norms (type/precision); #txx – Numeric (type/precision); Dtxx – DocValues (type/precision)

Field	IdfpoPSVBNtxx#txxDtxx	Norm	Value
:fulltext	Idfp--S--Nnum--------	101	\n\n\n\n\n\n\n\n\n\n\n\n\nAEM 6 OAK: MongoMK and Queries \nMarcel Reutegger & Thomas Mueller Graf | Senior Sc
:path	Id----S--------------	---	/granite-gems-lucene/AEM 6 Oak – MongoMK and Queries.pdf/jcr:content

For the given token ‘***mongomk***’ there are 2 matching lucene docs, pointing to the pdf file.

Why 2? Because of index time aggregation: the parent node will inherit the ‘:fulltext’ information from its child node.



Browse by document number:

Doc. #: 0 74890 74978

Browse by term:

(Hint: enter a substring and press Next to start at the nearest term).

Term: :fulltext mongomk

Decoded value:

Browse documents with this term (2 documents)

Document: 2 of 2

Term freq in this doc: 41

Doc #: 74890

Flags: I – Indexed (docs,freqs,pos,offsets) P – Payloads S – Stored; V – Term Vector B – Binary; Ntxx – Norms (type/precision); #txx – Numeric (type/precision); Dtxx – DocValues (type/precision)

Field	IdfpoPSVBNtxx#txxDtxx	Norm	Value
:fulltext	Idfp--S--Nnum--------	101	\n\n\n\n\n\n\n\n\n\n\n\n\nAEM 6 OAK: MongoMK and Queries \nMarcel Reutegger & Thomas Mueller Graf | Senior Sc
:path	Id----S--------------	---	/granite-gems-lucene/AEM 6 Oak – MongoMK and Queries.pdf

Lucene Index Internals - Luke

The default Lucene index defines aggregation for ‘nt:file’s, meaning they will inherit all extracted full-text information from the ‘nt:resource’ child nodes.

This means that the following search

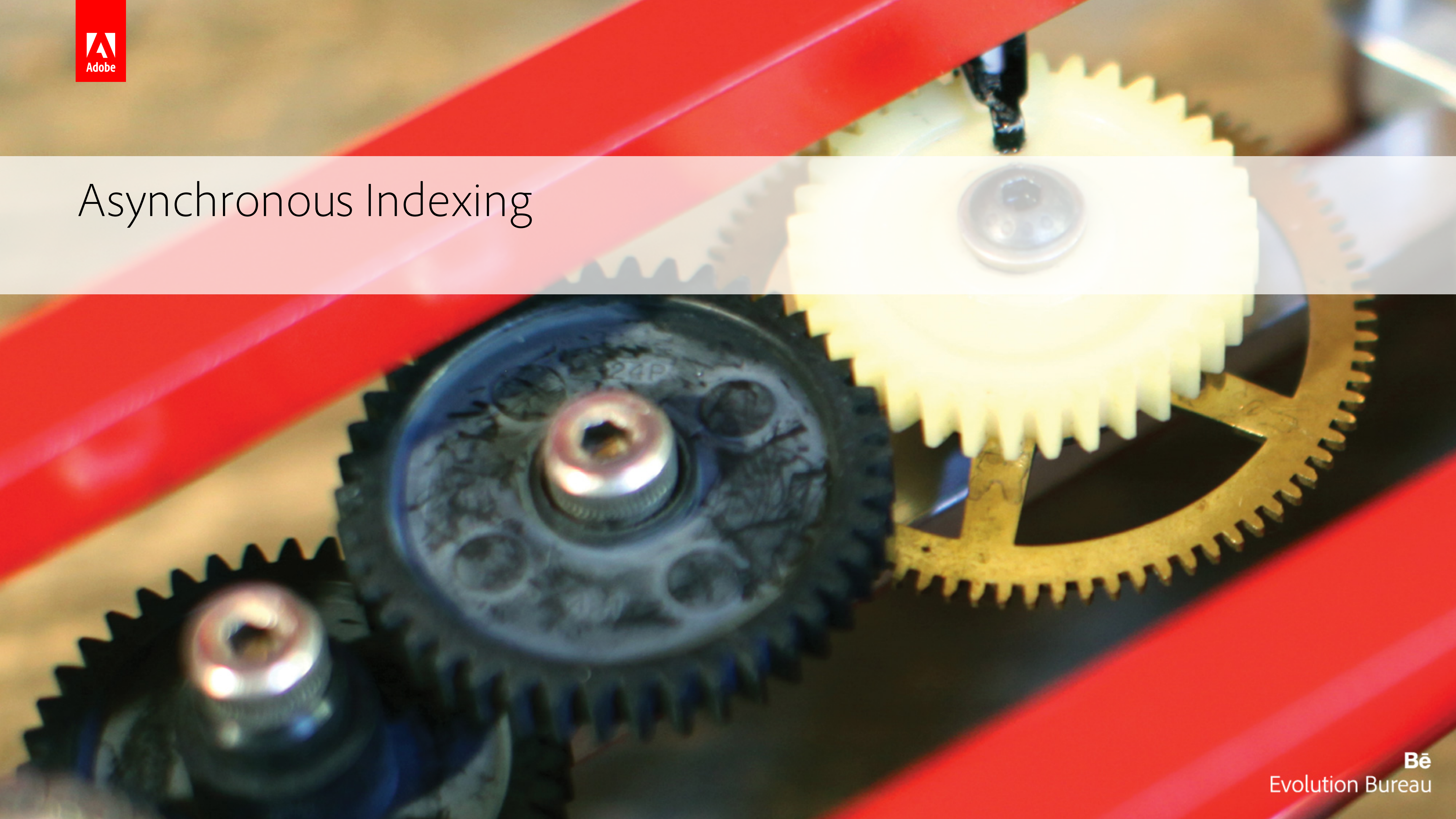
/jcr:root//element(, nt:file)[jcr:contains(., 'mongomk')]*

Will return a single item:

/granite-gems-lucene/AEM 6 Oak - MongoMK and Queries.pdf

even though the *nt:file* node itself contains no full-text information

Asynchronous Indexing



Asynchronous Indexing - Overview

- *AsyncIndexUpdate* class is the glue for all existing index implementations (all logging comes from this place)
- Runs as a background job every 5 seconds, for clusters this runs on a single cluster node
- Used mainly with full-text indexes: lucene/solr, also for ordered property indexes (deprecated)
- Efficient: takes care of processing only new content since last successful cycle, uses a fast diff based on *checkpoints*
- Resilient: in case of error, it will try again on next cycle (no data loss)
- Status exposed via JMX " *IndexStats*"
- You can change an index definition to be asynchronous by setting the *async* property: *async="async"*

Asynchronous Indexing - Checkpoints

- Checkpoints are a form of read-only tagging of the current state of the repository
- Each checkpoint has an expected lifetime provided at creation time, after which it will be removed, as well as some metadata related to its creation

```
/checkpoints/a6fe070e-deef-4582-85fb-b96b57ecd1a9
 - created = 1450285984929
 - timestamp = 1536685984929
 + properties
    - creator = "AsyncIndexUpdate"
    - name = "async"
    - thread = "pool-75-thread-4"
 + root  // entire repository content
    + libs
    + content
    + apps
    ....
```

[SegmentMK representation of a checkpoint]

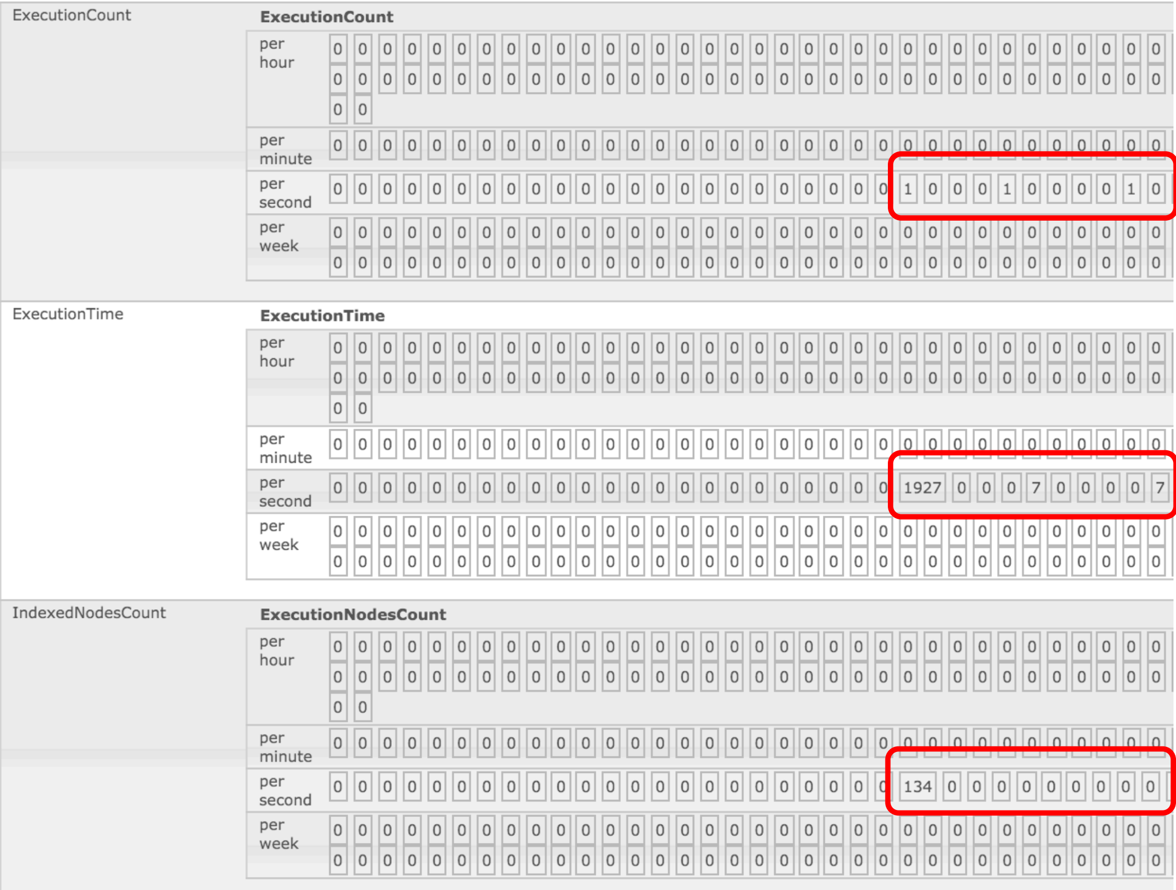
- The link between the async indexing process and a checkpoint is established via the */:async* node
- */:async@async* property must point to an existing checkpoint, otherwise *a full reindex will happen*
- */:async@async-LastIndexedTo* stores the timestamp up to which the repository was indexed
- */:async@async-temp* is the list of checkpoints to be cleaned up after all processing is done

```
/:async
 - async = "a6fe070e-deef-4582-85fb-b96b57ecd1a9"
 - async-LastIndexedTo = 2015-12-16T18:13:04.929+01:00
 - async-temp = ["6766f0ec-600f-4b8e-95d3-9b4d04f5877e",
                 "a6fe070e-deef-4582-85fb-b96b57ecd1a9"]
```

Asynchronous Indexing - JMX

org.apache.jackrabbit.oak: "async" ("IndexStats")

- Start / Done timestamps
- Checkpoints (reference, temp)
- Execution Count & Time, Indexed Nodes Count series
- Errors: failing flag, latest seen error with its timestamp



ExecutionCount	**ExecutionCount**	

	per hour	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
		0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
		0 0

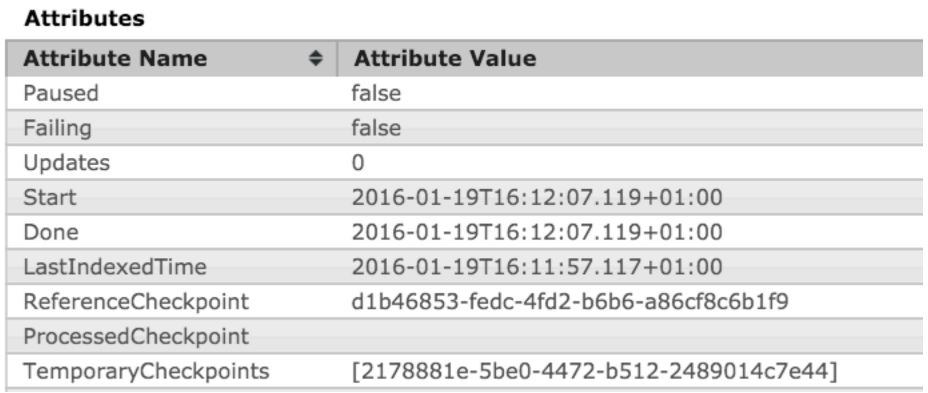
	per minute	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
	per second	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 1 0
	per week	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
ExecutionTime	**ExecutionTime**	

	per hour	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
		0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
		0 0

	per minute	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
	per second	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1927 0 0 0 7 0 0 0 0 7
	per week	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
IndexedNodesCount	**ExecutionNodesCount**	

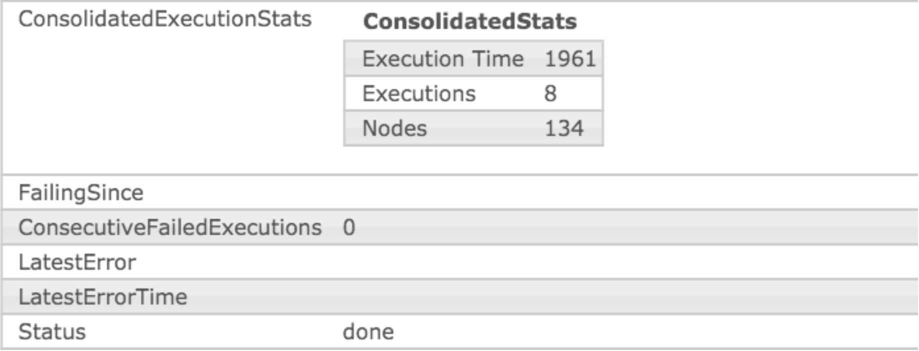
	per hour	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
		0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
		0 0

	per minute	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
	per second	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 134 0 0 0 0 0 0 0 0 0
	per week	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0



Attributes

Attribute Name	Attribute Value
Paused	false
Failing	false
Updates	0
Start	2016-01-19T16:12:07.119+01:00
Done	2016-01-19T16:12:07.119+01:00
LastIndexedTime	2016-01-19T16:11:57.117+01:00
ReferenceCheckpoint	d1b46853-fedc-4fd2-b6b6-a86cf8c6b1f9
ProcessedCheckpoint	
TemporaryCheckpoints	[2178881e-5be0-4472-b512-2489014c7e44]



ConsolidatedExecutionStats	**ConsolidatedStats**	
	Execution Time	1961
	Executions	8
	Nodes	134
FailingSince		
ConsecutiveFailedExecutions	0	
LatestError		
LatestErrorTime		
Status	done	

Useful Links

Oak Lucene Docs

https://jackrabbit.apache.org/oak/docs/query/lucene.html

AEM 6 Oak: MongoMK and Queries Gem session

http://dev.day.com/content/ddc/en/gems/aem-6-oak--mongomk-and-queries.html

AEM Docs on Oak Queries and Indexing

https://docs.adobe.com/docs/en/aem/6-1/deploy/platform/queries-and-indexing.html

https://docs.adobe.com/docs/en/aem/6-1/deploy/best-practices/best-practices-for-queries-and-indexing.html

The Index Manager

https://docs.adobe.com/docs/en/aem/6-1/administer/operations/operations-dashboard.html#The_Index_Manager

Adobe