Query Implementation

ˆ **A basic search query on an attribute value.**

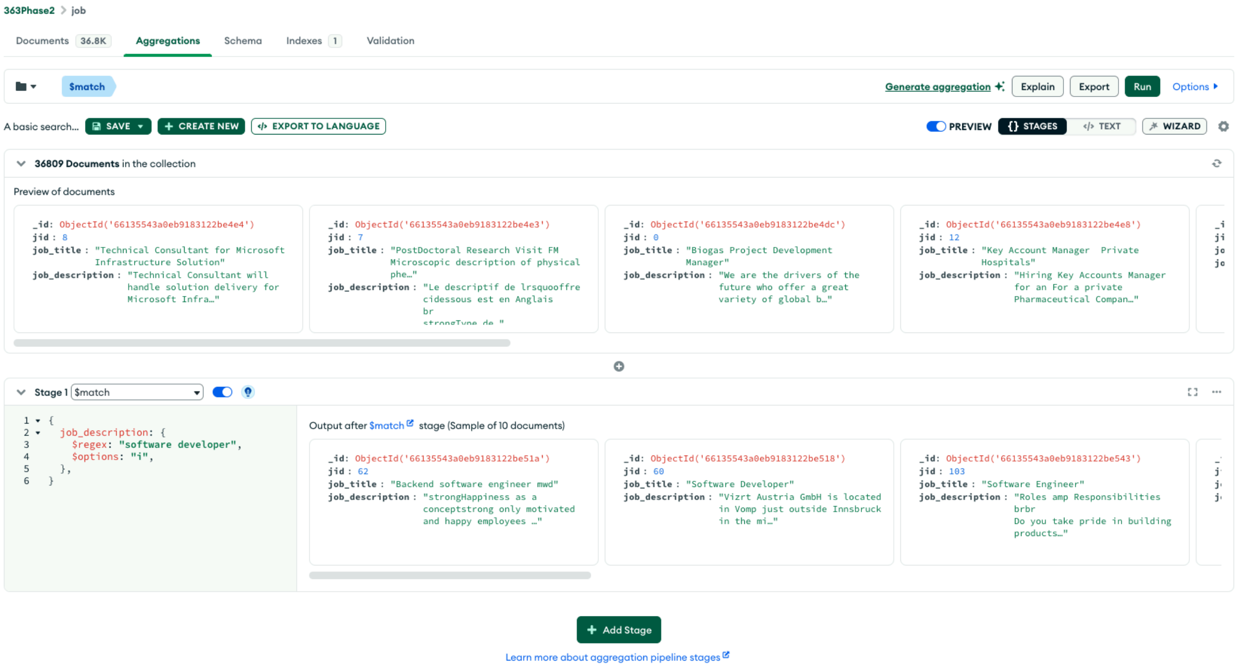
Pipeline saved to Job collection as “A basic search query on job_description value.”

This pipeline finds all jobs that have ‘software developer’ in the job_description.

Here is the query when I export it in JSON:

```
db.getCollection('job').aggregate(
  [
    {
      $match: {
        job_description: {
          $regex: 'software developer',
          $options: 'i'
        }
      }
    }
  ],
  { maxTimeMS: 60000, allowDiskUse: true }
);
```

Stages:



36809 Documents in the collection

Preview of documents

```
_id: ObjectId('66135543a0eb9183122be4e4')
jid : 8
job_title : "Technical Consultant for Microsoft
            Infrastructure Solution"
job_description : "Technical Consultant will
                  handle solution delivery for
                  Microsoft Infra…"
```

```
_id: ObjectId('66135543a0eb9183122be4e3')
jid : 7
job_title : "PostDoctoral Research Visit FM
            Microscopic description of physical
            phe…"
job_description : "Le descriptif de lrsquooffre
                  cidessous est en Anglais
                  br
                  strongType de "
```

```
_id: ObjectId('66135543a0eb9183122be4dc')
jid : 0
job_title : "Biogas Project Development
            Manager"
job_description : "We are the drivers of the
                  future who offer a great
                  variety of global b…"
```

```
_id: ObjectId('66135543a0eb9183122be4e8')
jid : 12
job_title : "Key Account Manager  Private
            Hospitals"
job_description : "Hiring Key Accounts Manager
                  for an For a private
                  Pharmaceutical Compan…"
```

Stage 1 $match

```
{
  job_description: {
    $regex: "software developer",
    $options: "i",
  },
}
```

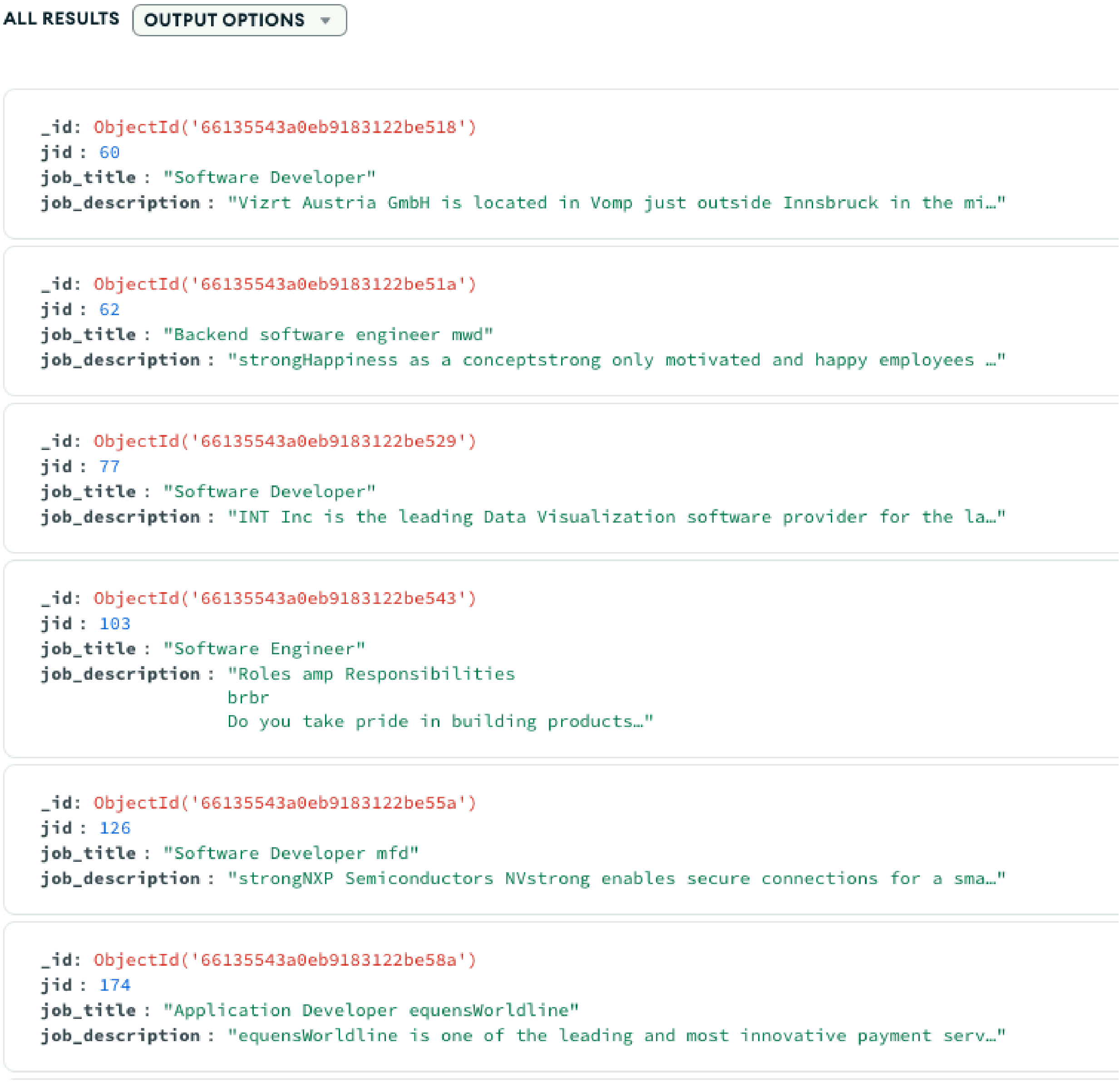
Output after $match stage (Sample of 10 documents)

```
_id: ObjectId('66135543a0eb9183122be51a')
jid : 62
job_title : "Backend software engineer mwd"
job_description : "strongHappiness as a
                  conceptstrong only motivated
                  and happy employees …"
```

```
_id: ObjectId('66135543a0eb9183122be518')
jid : 60
job_title : "Software Developer"
job_description : "Vizrt Austria GmbH is located
                  in Vomp just outside Innsbruck
                  in the mi…"
```

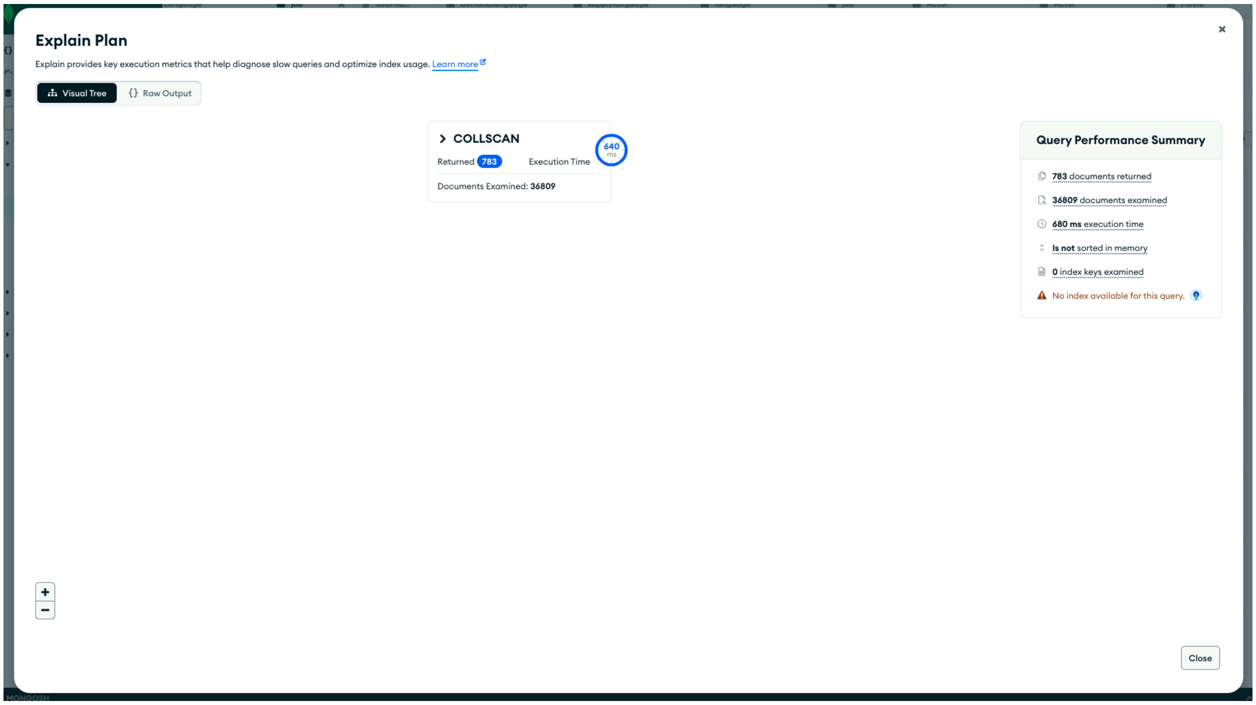
```
_id: ObjectId('66135543a0eb9183122be543')
jid : 103
job_title : "Software Engineer"
job_description : "Roles amp Responsibilities
                  brbr
                  Do you take pride in building
                  products…"
```

Output:



```
_id: ObjectId('66135543a0eb9183122be518')
jid : 60
job_title : "Software Developer"
job_description : "Vizrt Austria GmbH is located in Vomp just outside Innsbruck in the mi…"

_id: ObjectId('66135543a0eb9183122be51a')
jid : 62
job_title : "Backend software engineer mwd"
job_description : "strongHappiness as a conceptstrong only motivated and happy employees …"

_id: ObjectId('66135543a0eb9183122be529')
jid : 77
job_title : "Software Developer"
job_description : "INT Inc is the leading Data Visualization software provider for the la…"

_id: ObjectId('66135543a0eb9183122be543')
jid : 103
job_title : "Software Engineer"
job_description : "Roles amp Responsibilities
                  brbr
                  Do you take pride in building products…"

_id: ObjectId('66135543a0eb9183122be55a')
jid : 126
job_title : "Software Developer mfd"
job_description : "strongNXP Semiconductors NVstrong enables secure connections for a sma…"

_id: ObjectId('66135543a0eb9183122be58a')
jid : 174
job_title : "Application Developer equensWorldline"
job_description : "equensWorldline is one of the leading and most innovative payment serv…"
```

Execution time:



Explain Plan

Explain provides key execution metrics that help diagnose slow queries and optimize index usage. Learn more

COLLSCAN

Returned 783 Execution Time 640 ms

Documents Examined: 36809

Query Performance Summary

783 documents returned

36809 documents examined

680 ms execution time

Is not sorted in memory

0 index keys examined

No index available for this query.

ˆ A query that provides some aggregate data (i.e. number of entities satisfying a criteria)

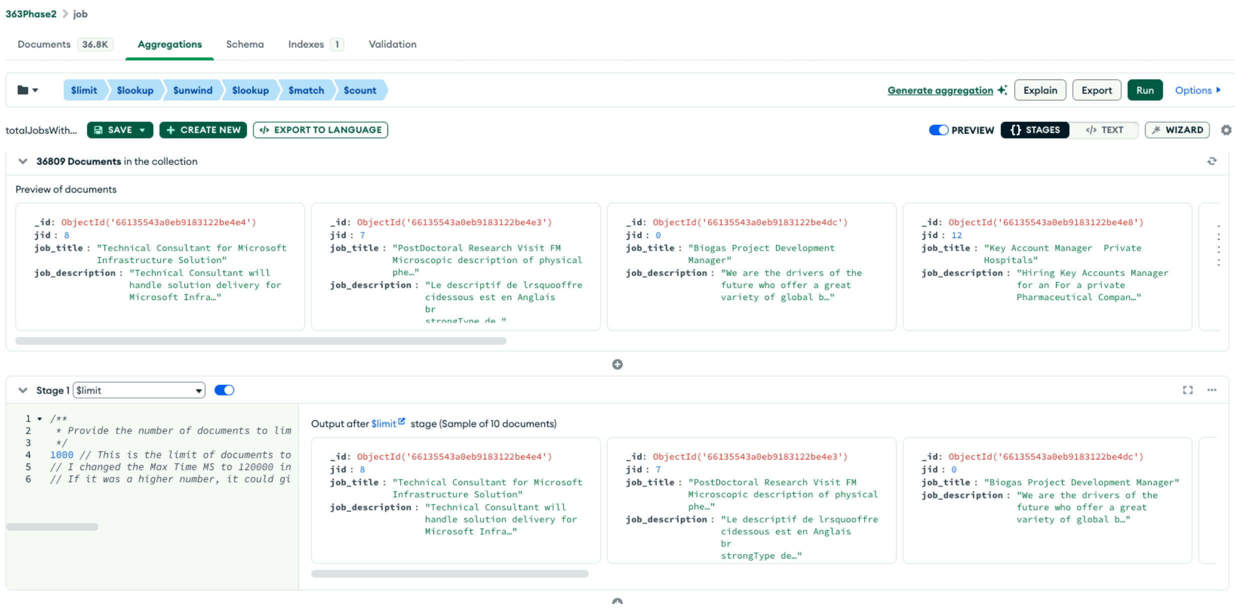
Pipeline saved to Job collection as "totalJobsWithPython A query that provides some aggregate data (i.e. number of entities satisfying a criteria)"

This pipeline finds the total number of jobs that demand python.

Here is the query when I export it in JSON:

```
db.getCollection('job').aggregate(
  [
    { $limit: 1000 },
    {
      $lookup: {
        from: 'demandlanguage',
        localField: 'jid',
        foreignField: 'jid',
        as: 'job_demand'
      }
    },
    { $unwind: '$job_demand' },
    {
      $lookup: {
        from: 'language',
        localField: 'job_demand.lid',
        foreignField: 'lid',
        as: 'job_language'
      }
    },
    {
      $match: {
        'job_language.language': {
          $regex: 'python',
          $options: 'i'
        }
      }
    },
    { $count: 'countOfJobsWithPython' }
  ],
  { maxTimeMS: 120000, allowDiskUse: true }
);
```

Stages:



36809 Documents in the collection

Preview of documents

```
_id: ObjectId('66135543a0eb9183122be4e4')
jid : 8
job_title : "Technical Consultant for Microsoft
            Infrastructure Solution"
job_description : "Technical Consultant will
                  handle solution delivery for
                  Microsoft Infra…"
```

```
_id: ObjectId('66135543a0eb9183122be4e3')
jid : 7
job_title : "PostDoctoral Research Visit FM
            Microscopic description of physical
            phe…"
job_description : "Le descriptif de lrsquooffre
                  cidessous est en Anglais
                  br
                  strongType de…"
```

```
_id: ObjectId('66135543a0eb9183122be4dc')
jid : 0
job_title : "Biogas Project Development
            Manager"
job_description : "We are the drivers of the
                  future who offer a great
                  variety of global b…"
```

```
_id: ObjectId('66135543a0eb9183122be4e8')
jid : 12
job_title : "Key Account Manager  Private
            Hospitals"
job_description : "Hiring Key Accounts Manager
                  for an For a private
                  Pharmaceutical Compan…"
```

Stage 1 $limit

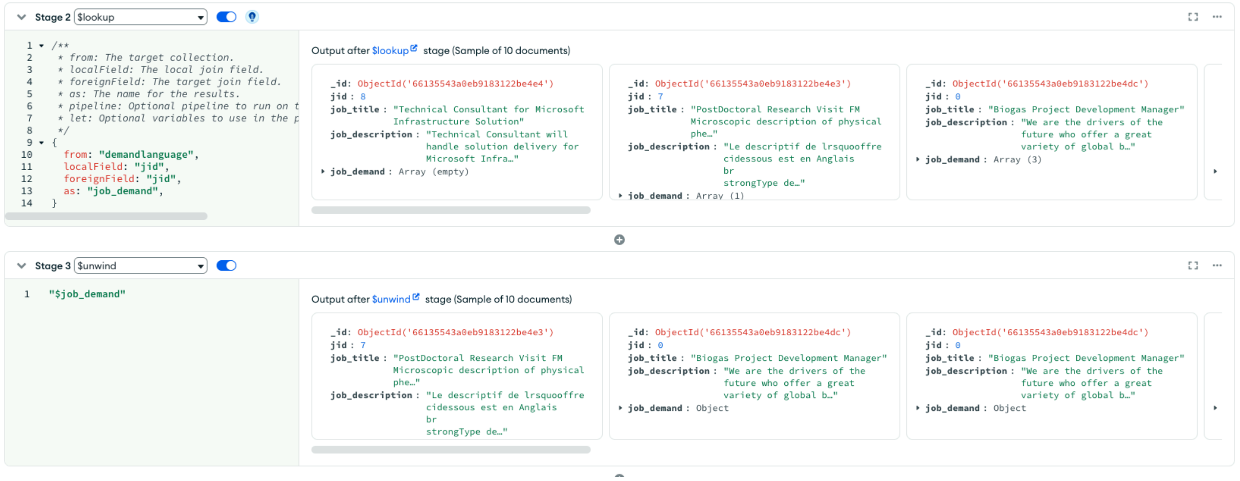
```
/**
  * Provide the number of documents to lim
  */
1000 // This is the limit of documents to
// I changed the Max Time MS to 120000 in
// If it was a higher number, it could gi
```

Output after $limit stage (Sample of 10 documents)

```
_id: ObjectId('66135543a0eb9183122be4e4')
jid : 8
job_title : "Technical Consultant for Microsoft
            Infrastructure Solution"
job_description : "Technical Consultant will
                  handle solution delivery for
                  Microsoft Infra…"
```

```
_id: ObjectId('66135543a0eb9183122be4e3')
jid : 7
job_title : "PostDoctoral Research Visit FM
            Microscopic description of physical
            phe…"
job_description : "Le descriptif de lrsquooffre
                  cidessous est en Anglais
                  br
                  strongType de…"
```

```
_id: ObjectId('66135543a0eb9183122be4dc')
jid : 0
job_title : "Biogas Project Development Manager"
job_description : "We are the drivers of the
                  future who offer a great
                  variety of global b…"
```



Stage 2 $lookup

```
/**
  * from: The target collection.
  * localField: The local join field.
  * foreignField: The target join field.
  * as: The name for the results.
  * pipeline: Optional pipeline to run on t
  * let: Optional variables to use in the p
  */
{
  from: "demandlanguage",
  localField: "jid",
  foreignField: "jid",
  as: "job_demand",
}
```

Output after $lookup stage (Sample of 10 documents)

```
_id: ObjectId('66135543a0eb9183122be4e4')
jid : 8
job_title : "Technical Consultant for Microsoft
            Infrastructure Solution"
job_description : "Technical Consultant will
                  handle solution delivery for
                  Microsoft Infra…"
▸ job_demand : Array (empty)
```

```
_id: ObjectId('66135543a0eb9183122be4e3')
jid : 7
job_title : "PostDoctoral Research Visit FM
            Microscopic description of physical
            phe…"
job_description : "Le descriptif de lrsquooffre
                  cidessous est en Anglais
                  br
                  strongType de…"
▸ job_demand : Array (1)
```

```
_id: ObjectId('66135543a0eb9183122be4dc')
jid : 0
job_title : "Biogas Project Development Manager"
job_description : "We are the drivers of the
                  future who offer a great
                  variety of global b…"
▸ job_demand : Array (3)
```

Stage 3 $unwind

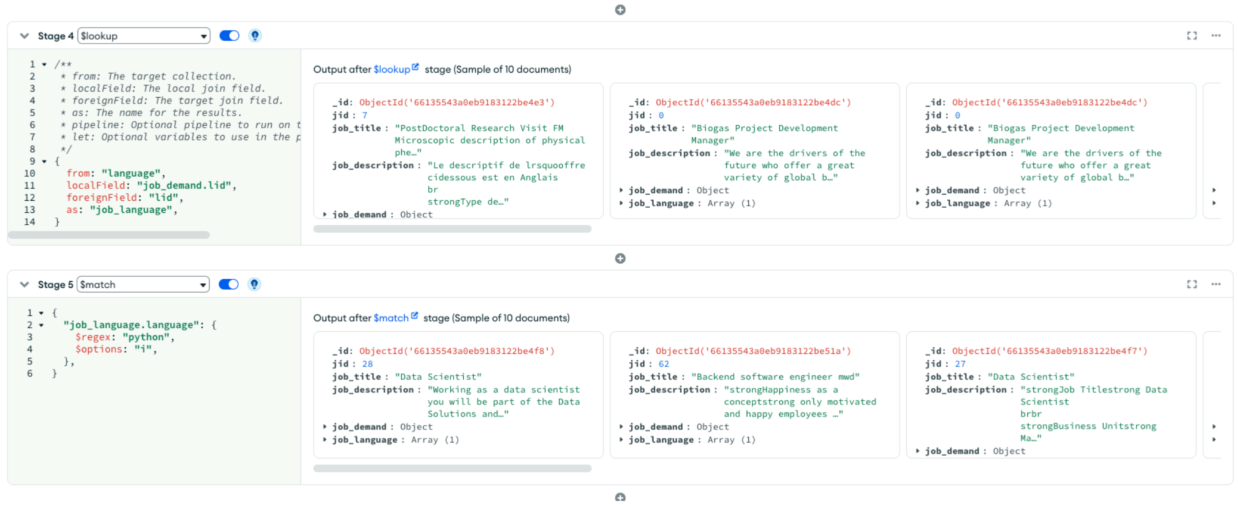
```
"$job_demand"
```

Output after $unwind stage (Sample of 10 documents)

```
_id: ObjectId('66135543a0eb9183122be4e3')
jid : 7
job_title : "PostDoctoral Research Visit FM
            Microscopic description of physical
            phe…"
job_description : "Le descriptif de lrsquooffre
                  cidessous est en Anglais
                  br
                  strongType de…"
```

```
_id: ObjectId('66135543a0eb9183122be4dc')
jid : 0
job_title : "Biogas Project Development Manager"
job_description : "We are the drivers of the
                  future who offer a great
                  variety of global b…"
▸ job_demand : Object
```

```
_id: ObjectId('66135543a0eb9183122be4dc')
jid : 0
job_title : "Biogas Project Development Manager"
job_description : "We are the drivers of the
                  future who offer a great
                  variety of global b…"
▸ job_demand : Object
```



Stage 4 $lookup

```
/**
 * from: The target collection.
 * localField: The local join field.
 * foreignField: The target join field.
 * as: The name for the results.
 * pipeline: Optional pipeline to run on t
 * let: Optional variables to use in the p
 */
{
  from: "language",
  localField: "job_demand.lid",
  foreignField: "lid",
  as: "job_language",
}
```

Output after $lookup stage (Sample of 10 documents)

```
_id: ObjectId('66135543a0eb9183122be4e3')
jid : 7
job_title : "PostDoctoral Research Visit FM
            Microscopic description of physical
            phe…"
job_description : "Le descriptif de lrsquooffre
                  cidessous est en Anglais
                  br
                  strongType de…"
▸ job_demand : Object
```

```
_id: ObjectId('66135543a0eb9183122be4dc')
jid : 0
job_title : "Biogas Project Development
            Manager"
job_description : "We are the drivers of the
                  future who offer a great
                  variety of global b…"
▸ job_demand : Object
▸ job_language : Array (1)
```

```
_id: ObjectId('66135543a0eb9183122be4dc')
jid : 0
job_title : "Biogas Project Development
            Manager"
job_description : "We are the drivers of the
                  future who offer a great
                  variety of global b…"
▸ job_demand : Object
▸ job_language : Array (1)
```

Stage 5 $match

```
{
  "job_language.language": {
    $regex: "python",
    $options: "i",
  },
}
```

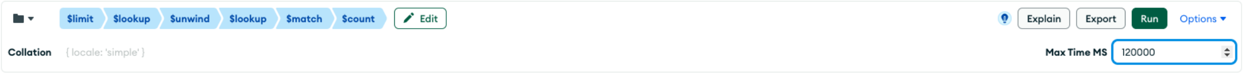
Output after $match stage (Sample of 10 documents)

```
_id: ObjectId('66135543a0eb9183122be4f8')
jid : 28
job_title : "Data Scientist"
job_description : "Working as a data scientist
                  you will be part of the Data
                  Solutions and…"
▸ job_demand : Object
▸ job_language : Array (1)
```

```
_id: ObjectId('66135543a0eb9183122be51a')
jid : 62
job_title : "Backend software engineer mwd"
job_description : "strongHappiness as a
                  conceptstrong only motivated
                  and happy employees …"
▸ job_demand : Object
▸ job_language : Array (1)
```

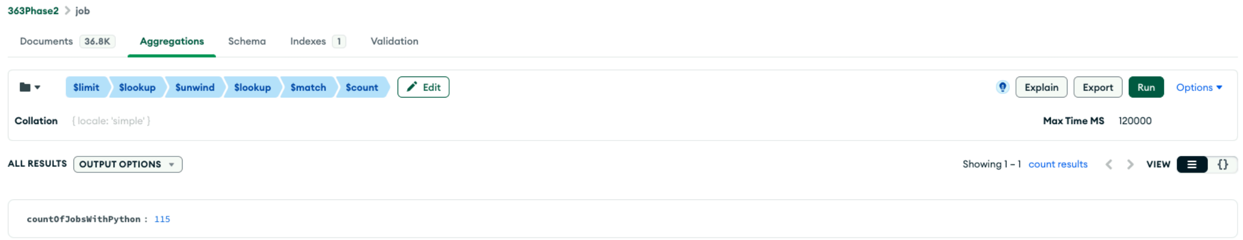
```
_id: ObjectId('66135543a0eb9183122be4f7')
jid : 27
job_title : "Data Scientist"
job_description : "strongJob Titlestrong Data
                  Scientist
                  brbr
                  strongBusiness Unitstrong
                  Ma…"
▸ job_demand : Object
```

Output:



Collation { locale: 'simple' }

Max Time MS 120000



Collation { locale: 'simple' }

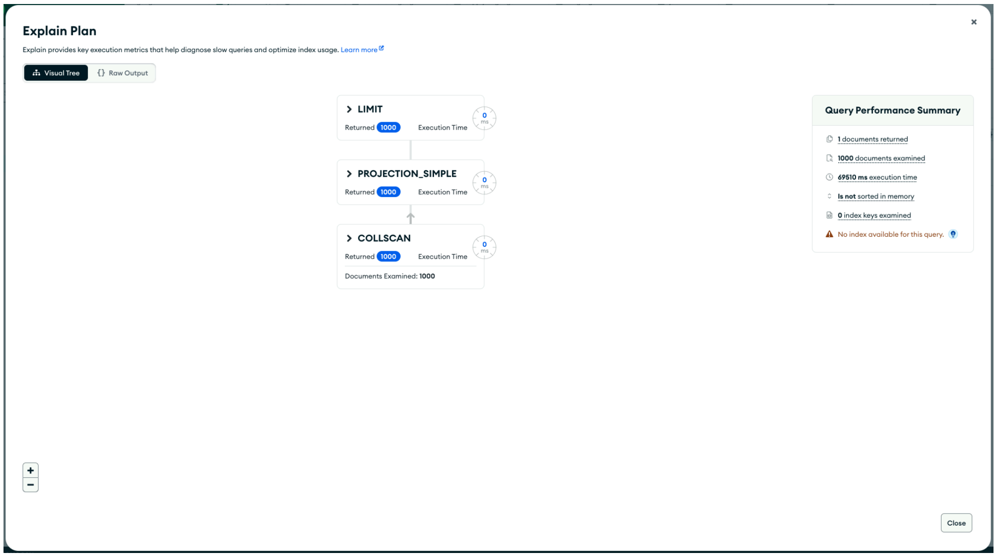
Max Time MS 120000

ALL RESULTS

Showing 1 – 1

```
countOfJobsWithPython : 115
```

Execution time:



Explain Plan

Explain provides key execution metrics that help diagnose slow queries and optimize index usage. Learn more

LIMIT
Returned 1000 Execution Time 0 ms

PROJECTION_SIMPLE
Returned 1000 Execution Time 0 ms

COLLSCAN
Returned 1000 Execution Time 0 ms
Documents Examined: 1000

Query Performance Summary

- 1 documents returned
- 1000 documents examined
- 69510 ms execution time
- Is not sorted in memory
- 0 index keys examined
- No index available for this query.

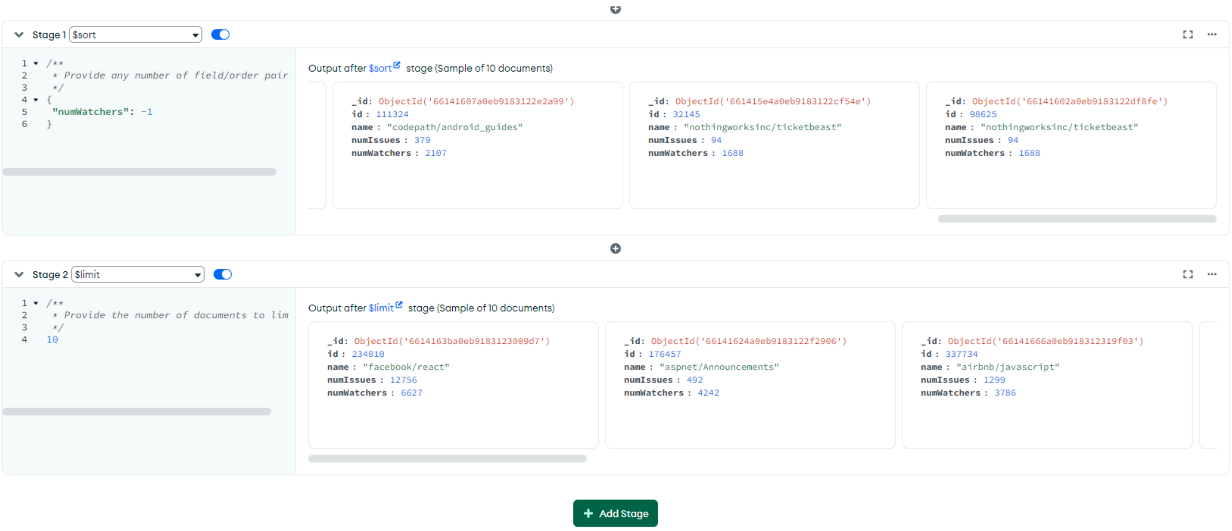
ˆ Find top n entities satisfying a criteria, sorted by an attribute.

Pipeline saved to Repository collection as ”TopNEntities”. This query finds the top 10 repositories with the higher number of watchers.

Here is the query when I export it in JSON:

```
db.getCollection('repository').aggregate(
 [
   { $sort: { numWatchers: -1 } },
   { $limit: 10 }
 ],
 { maxTimeMS: 60000, allowDiskUse: true }
);
```

Stages:



Stage 1 $sort

```
/**
 * Provide any number of field/order pair
 */
{
  "numWatchers": -1
}
```

Output after $sort stage (Sample of 10 documents)

```
_id: ObjectId('66141607a0eb9183122e2a99')
id : 111324
name : "codepath/android_guides"
numIssues : 379
numWatchers : 2107
```

```
_id: ObjectId('661415e4a0eb9183122cf54e')
id : 32145
name : "nothingworksinc/ticketbeast"
numIssues : 94
numWatchers : 1688
```

```
_id: ObjectId('66141602a0eb9183122df8fe')
id : 98625
name : "nothingworksinc/ticketbeast"
numIssues : 94
numWatchers : 1688
```

Stage 2 $limit

```
/**
 * Provide the number of documents to lim
 */
10
```

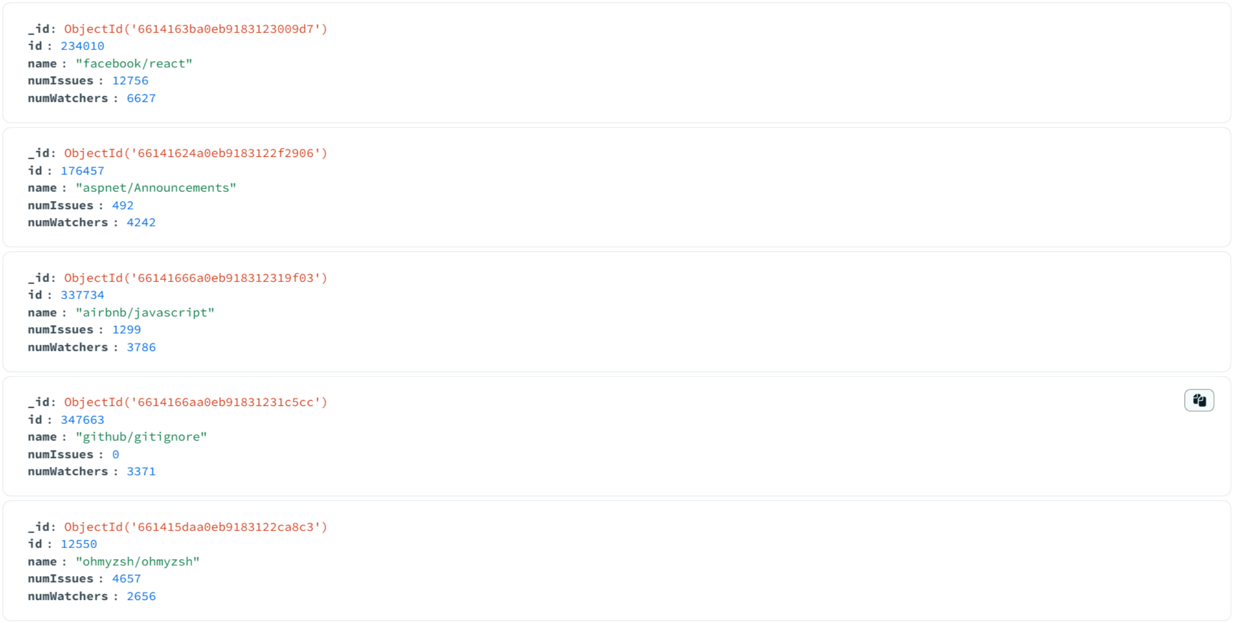
Output after $limit stage (Sample of 10 documents)

```
_id: ObjectId('6614163ba0eb9183123009d7')
id : 234010
name : "facebook/react"
numIssues : 12756
numWatchers : 6627
```

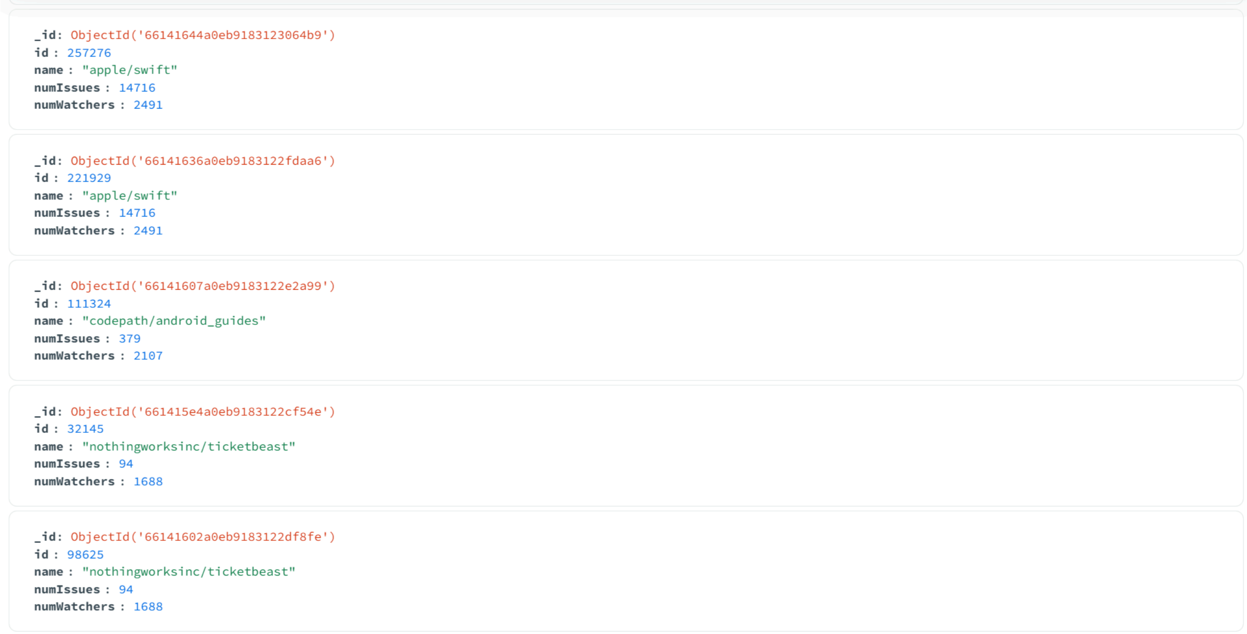
```
_id: ObjectId('66141624a0eb9183122f2906')
id : 176457
name : "aspnet/Announcements"
numIssues : 492
numWatchers : 4242
```

```
_id: ObjectId('66141666a0eb918312319f03')
id : 337734
name : "airbnb/javascript"
numIssues : 1299
numWatchers : 3786
```

Output:

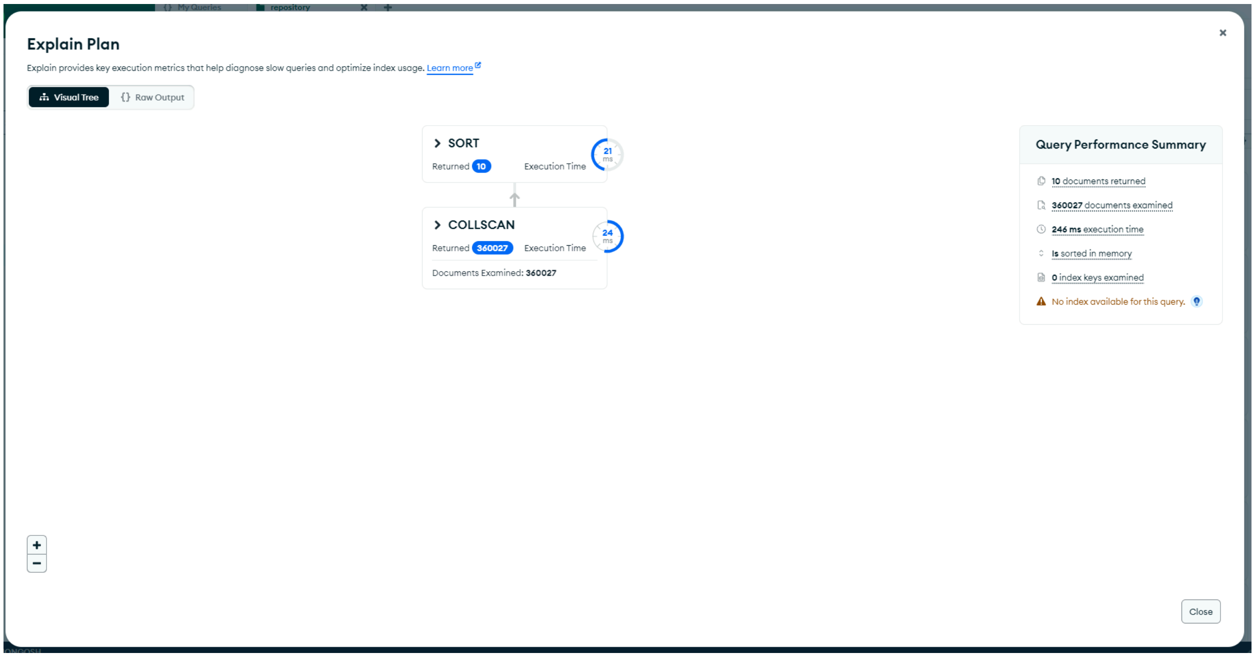


```
_id: ObjectId('6614163ba0eb9183123009d7')
id : 234010
name : "facebook/react"
numIssues : 12756
numWatchers : 6627

_id: ObjectId('66141624a0eb9183122f2906')
id : 176457
name : "aspnet/Announcements"
numIssues : 492
numWatchers : 4242

_id: ObjectId('66141666a0eb918312319f03')
id : 337734
name : "airbnb/javascript"
numIssues : 1299
numWatchers : 3786

_id: ObjectId('6614166aa0eb91831231c5cc')
id : 347663
name : "github/gitignore"
numIssues : 0
numWatchers : 3371

_id: ObjectId('661415daa0eb9183122ca8c3')
id : 12550
name : "ohmyzsh/ohmyzsh"
numIssues : 4657
numWatchers : 2656
```



```
_id: ObjectId('66141644a0eb9183123064b9')
id : 257276
name : "apple/swift"
numIssues : 14716
numWatchers : 2491

_id: ObjectId('66141636a0eb9183122fdaa6')
id : 221929
name : "apple/swift"
numIssues : 14716
numWatchers : 2491

_id: ObjectId('66141607a0eb9183122e2a99')
id : 111324
name : "codepath/android_guides"
numIssues : 379
numWatchers : 2107

_id: ObjectId('661415e4a0eb9183122cf54e')
id : 32145
name : "nothingworksinc/ticketbeast"
numIssues : 94
numWatchers : 1688

_id: ObjectId('66141602a0eb9183122df8fe')
id : 98625
name : "nothingworksinc/ticketbeast"
numIssues : 94
numWatchers : 1688
```

Execution time:



Explain Plan

Explain provides key execution metrics that help diagnose slow queries and optimize index usage. Learn more

SORT

Returned 10 Execution Time 21 ms

COLLSCAN

Returned 360027 Execution Time 24 ms

Documents Examined: 360027

Query Performance Summary

10 documents returned

360027 documents examined

246 ms execution time

Is sorted in memory

0 index keys examined

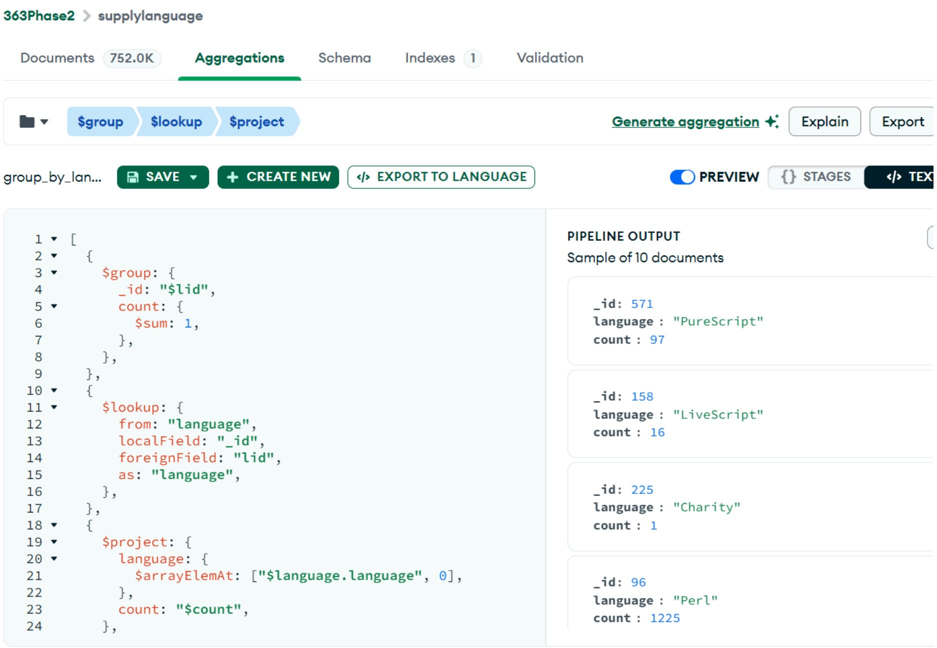
No index available for this query.

ˆ **Simulate a relational group by query in NoSQL (aggregate per category).**

Query to group languages by the number of times they appear in a repository.

Here is the query when I export it in JSON:

```
[
  {
    $group: {
      _id: "$lid",
      count: {
        $sum: 1,
      },
    },
  },
  {
    $lookup: {
      from: "language",
      localField: "_id",
      foreignField: "lid",
      as: "language",
    },
  },
  {
    $project: {
      language: {
        $arrayElemAt: ["$language.language", 0],
      },
      count: "$count",
    },
  },
]
```



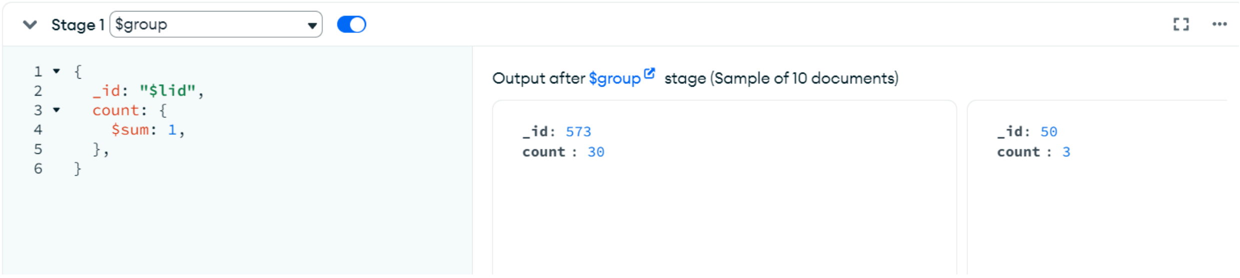
```
[
  {
    $group: {
      _id: "$lid",
      count: {
        $sum: 1,
      },
    },
  },
  {
    $lookup: {
      from: "language",
      localField: "_id",
      foreignField: "lid",
      as: "language",
    },
  },
  {
    $project: {
      language: {
        $arrayElemAt: ["$language.language", 0],
      },
      count: "$count",
    },
```

PIPELINE OUTPUT

Sample of 10 documents

```
_id: 571
language : "PureScript"
count : 97

_id: 158
language : "LiveScript"
count : 16

_id: 225
language : "Charity"
count : 1

_id: 96
language : "Perl"
count : 1225
```

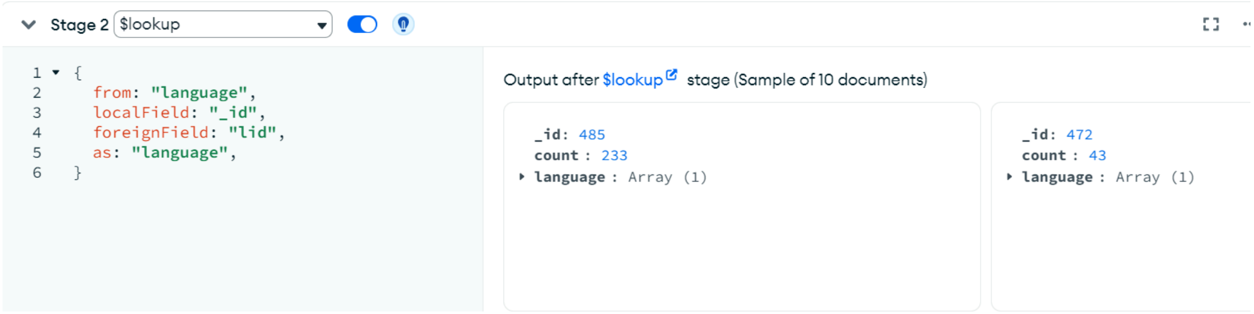
Stages:



```
{
  _id: "$lid",
  count: {
    $sum: 1,
  },
}
```

Output after $group stage (Sample of 10 documents)

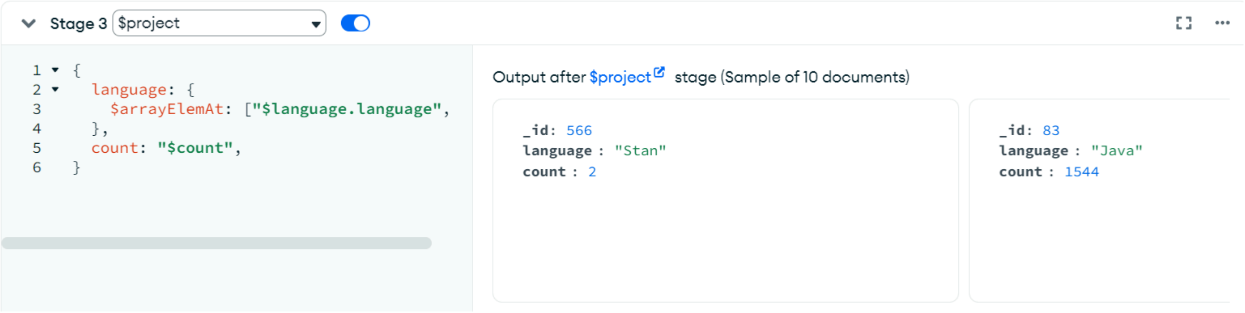
```
_id: 573
count : 30

_id: 50
count : 3
```



```
{
  from: "language",
  localField: "_id",
  foreignField: "lid",
  as: "language",
}
```

Output after $lookup stage (Sample of 10 documents)

```
_id: 485
count : 233
▸ language : Array (1)

_id: 472
count : 43
▸ language : Array (1)
```



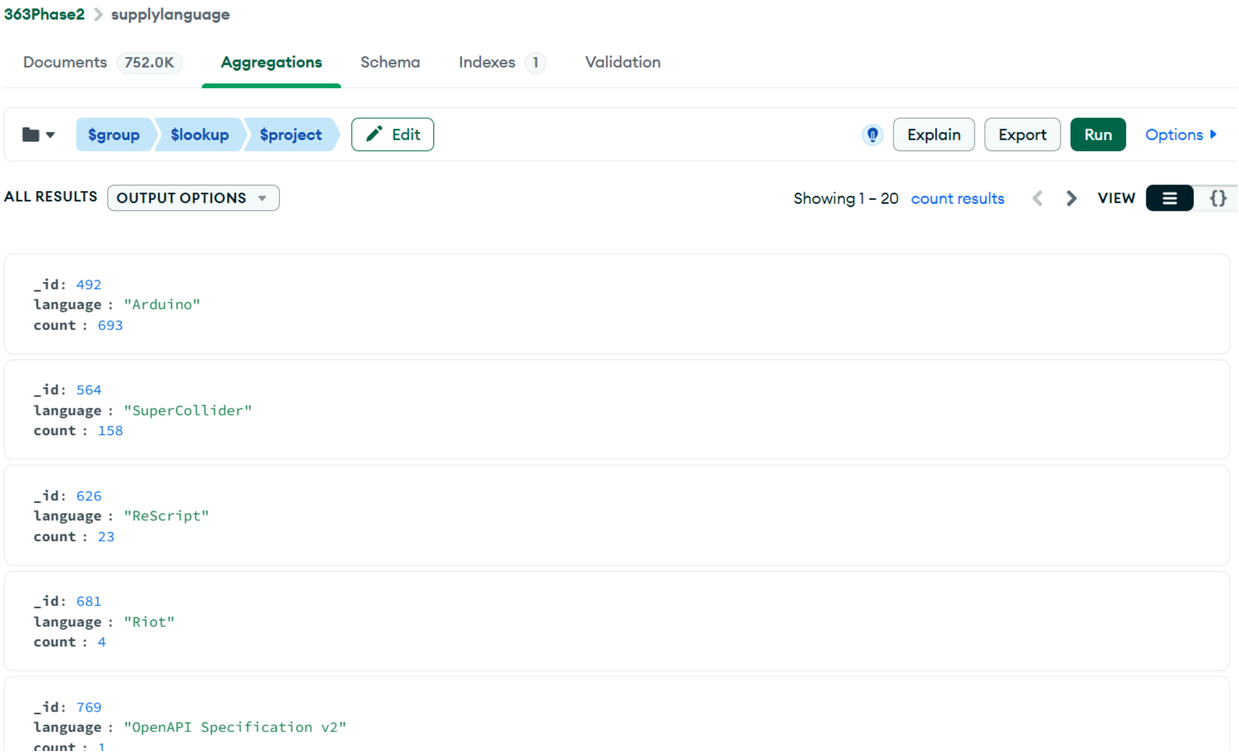
```
{
  language: {
    $arrayElemAt: ["$language.language",
  },
  count: "$count",
}
```

Output after $project stage (Sample of 10 documents)

```
_id: 566
language : "Stan"
count : 2
```

```
_id: 83
language : "Java"
count : 1544
```

Output:



```
_id: 492
language : "Arduino"
count : 693
```

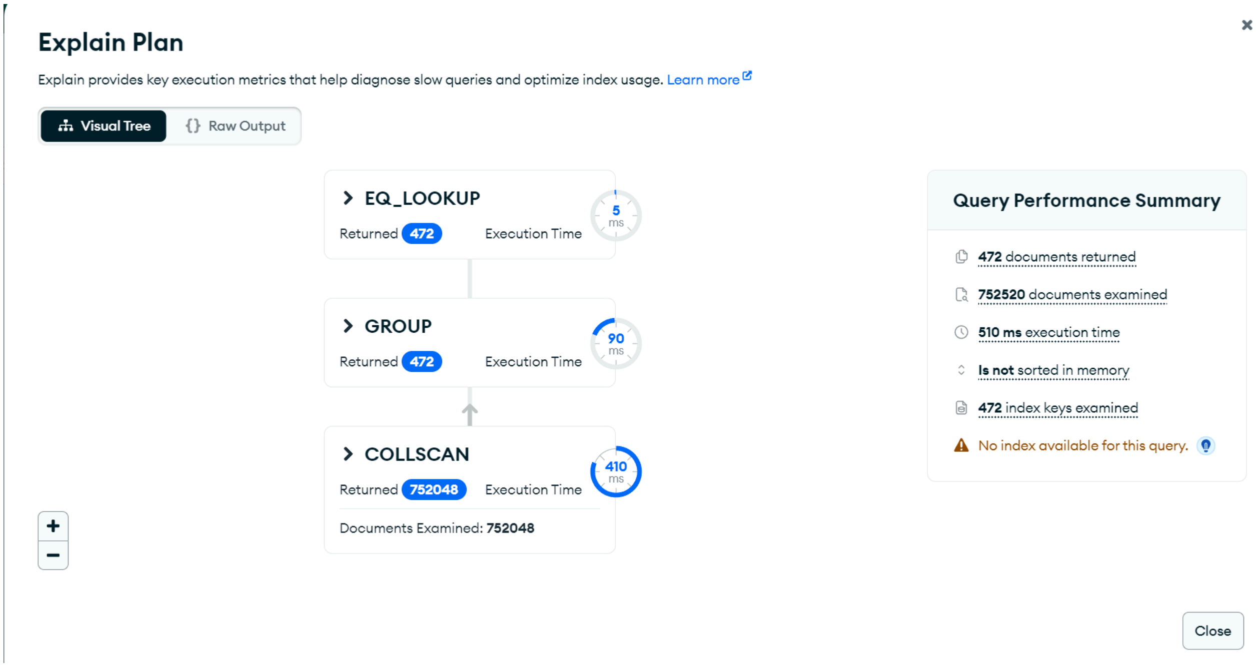
```
_id: 564
language : "SuperCollider"
count : 158
```

```
_id: 626
language : "ReScript"
count : 23
```

```
_id: 681
language : "Riot"
count : 4
```

```
_id: 769
language : "OpenAPI Specification v2"
count : 1
```

Execution time:



Explain Plan

Explain provides key execution metrics that help diagnose slow queries and optimize index usage. Learn more

EQ_LOOKUP
Returned 472 Execution Time 5 ms

GROUP
Returned 472 Execution Time 90 ms

COLLSCAN
Returned 752048 Execution Time 410 ms
Documents Examined: 752048

Query Performance Summary

- 472 documents returned
- 752520 documents examined
- 510 ms execution time
- Is not sorted in memory
- 472 index keys examined
- No index available for this query.

ˆ **Build the appropriate indexes for previous queries, report the index creation statement and the query execution time before and after you create the index.**

A basic search query on an attribute value.

Pipeline saved to Job collection as “INDEX a basic search query on job_description value.”

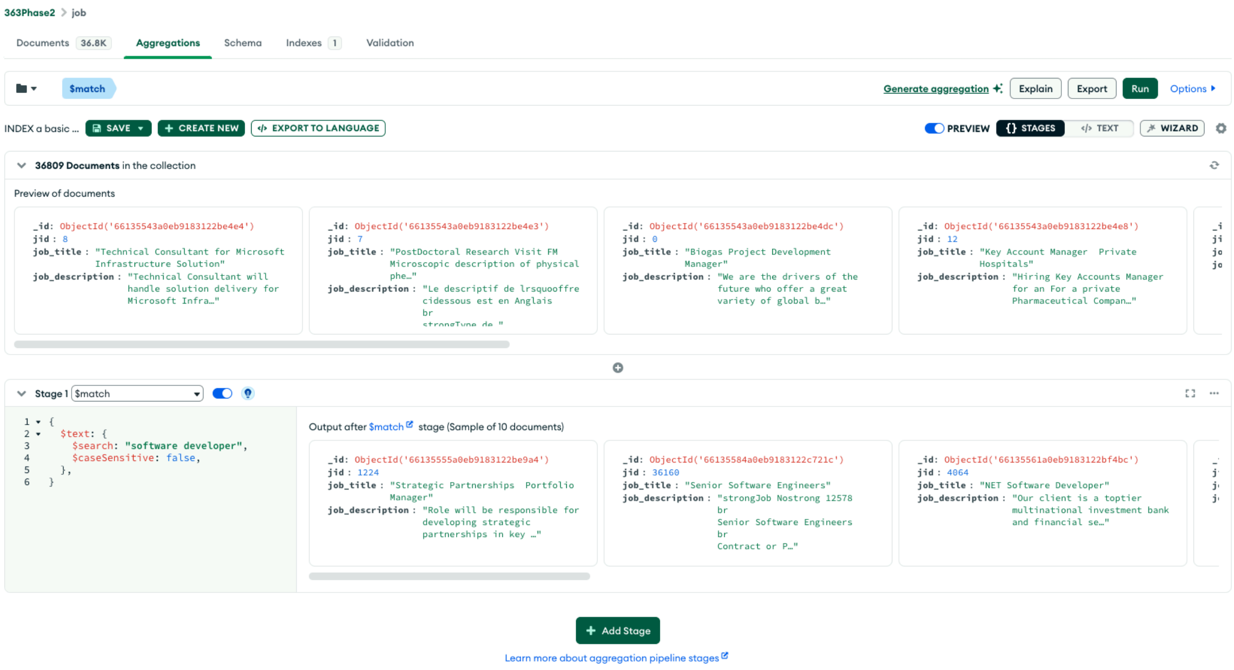
This pipeline finds all jobs that have ‘software developer’ in the job_description using indexes.

Here is the query when I export it in JSON:

```
db.getCollection('job').aggregate(
 [
   {
     $match: {
       $text: {
         $search: 'software developer',
         $caseSensitive: false
```

```
      }
    }
  }
 ],
 { maxTimeMS: 60000, allowDiskUse: true }
);
```

Stages:



36809 Documents in the collection

Preview of documents

```
_id: ObjectId('66135543a0eb9183122be4e4')
jid : 8
job_title : "Technical Consultant for Microsoft
            Infrastructure Solution"
job_description : "Technical Consultant will
                  handle solution delivery for
                  Microsoft Infra…"
```

```
_id: ObjectId('66135543a0eb9183122be4e3')
jid : 7
job_title : "PostDoctoral Research Visit FM
            Microscopic description of physical
            phe…"
job_description : "Le descriptif de lrsquooffre
                  cidessous est en Anglais
                  br
                  strongType de "
```

```
_id: ObjectId('66135543a0eb9183122be4dc')
jid : 0
job_title : "Biogas Project Development
            Manager"
job_description : "We are the drivers of the
                  future who offer a great
                  variety of global b…"
```

```
_id: ObjectId('66135543a0eb9183122be4e8')
jid : 12
job_title : "Key Account Manager  Private
            Hospitals"
job_description : "Hiring Key Accounts Manager
                  for an For a private
                  Pharmaceutical Compan…"
```

Stage 1 $match

```
{
  $text: {
    $search: "software developer",
    $caseSensitive: false,
  },
}
```

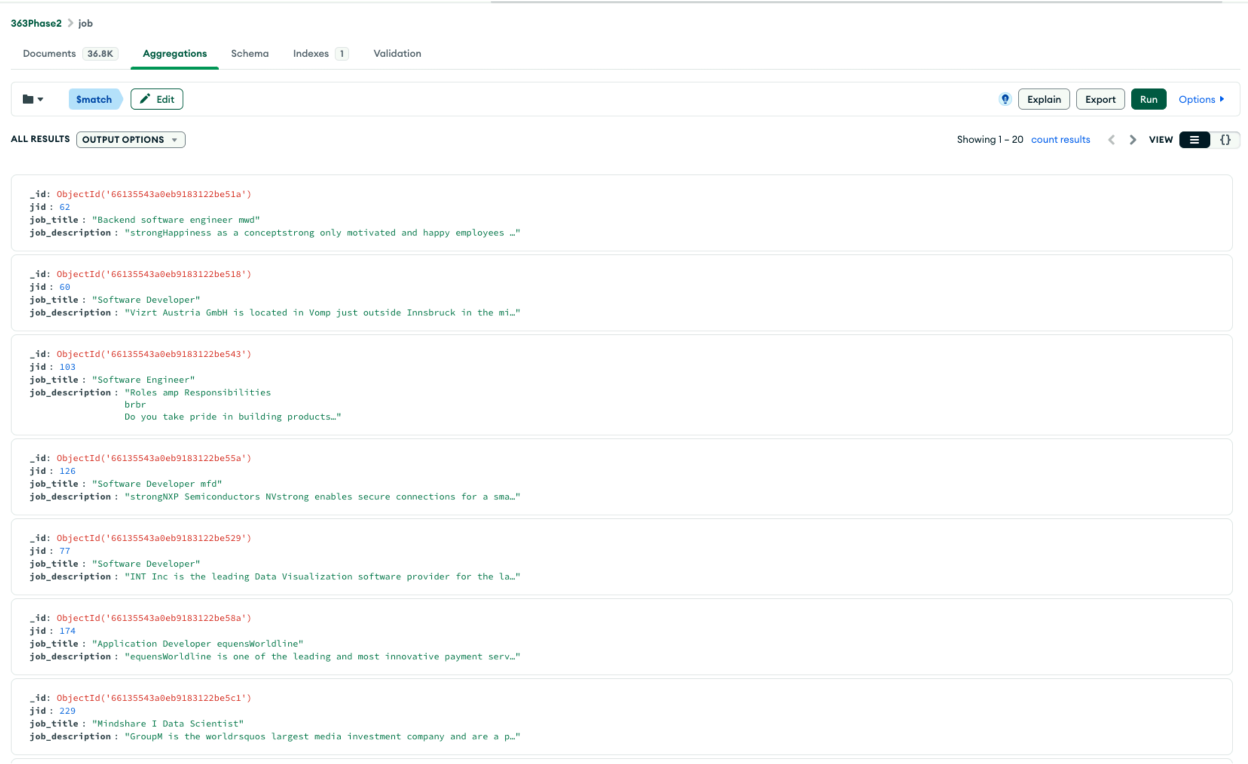
Output after $match stage (Sample of 10 documents)

```
_id: ObjectId('66135555a0eb9183122be9a4')
jid : 1224
job_title : "Strategic Partnerships  Portfolio
            Manager"
job_description : "Role will be responsible for
                  developing strategic
                  partnerships in key …"
```

```
_id: ObjectId('66135584a0eb9183122c721c')
jid : 36160
job_title : "Senior Software Engineers"
job_description : "strongJob Nostrong 12578
                  br
                  Senior Software Engineers
                  br
                  Contract or P…"
```

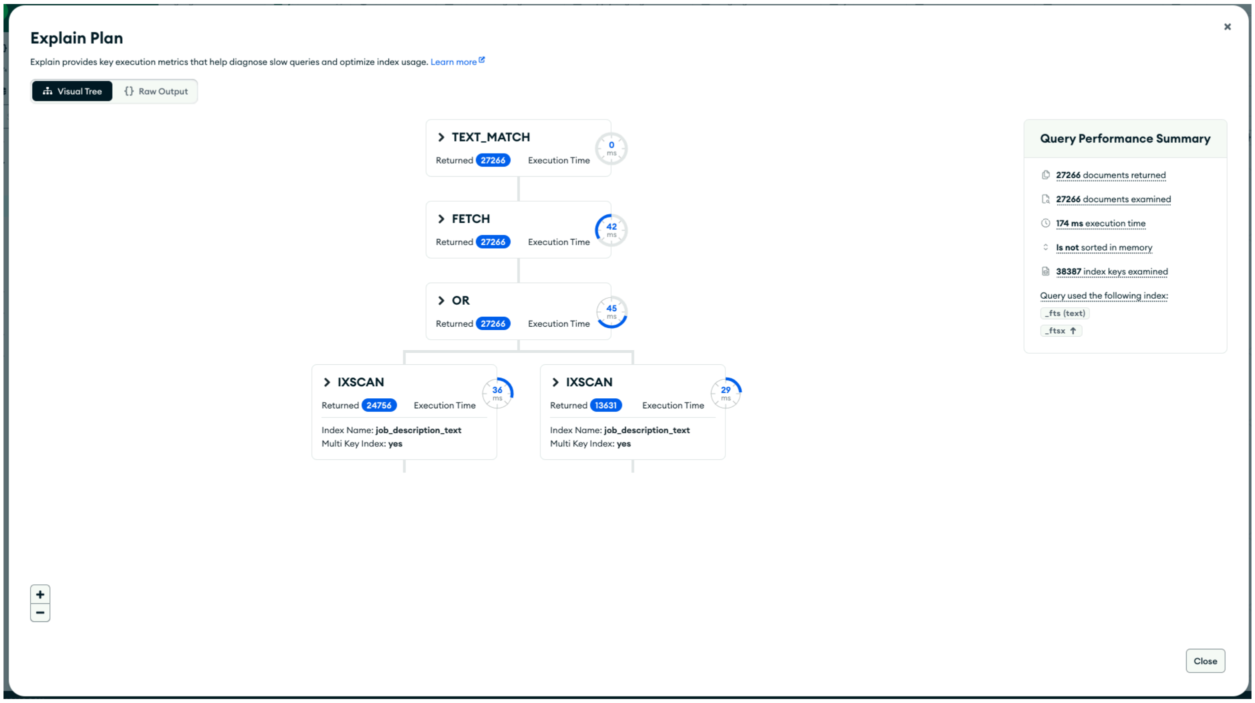
```
_id: ObjectId('66135561a0eb9183122bf4bc')
jid : 4064
job_title : "NET Software Developer"
job_description : "Our client is a toptier
                  multinational investment bank
                  and financial se…"
```

Output:



```
_id: ObjectId('66135543a0eb9183122be51a')
jid : 62
job_title : "Backend software engineer mwd"
job_description : "strongHappiness as a conceptstrong only motivated and happy employees …"

_id: ObjectId('66135543a0eb9183122be518')
jid : 60
job_title : "Software Developer"
job_description : "Vizrt Austria GmbH is located in Vomp just outside Innsbruck in the mi…"

_id: ObjectId('66135543a0eb9183122be543')
jid : 103
job_title : "Software Engineer"
job_description : "Roles amp Responsibilities
                  brbr
                  Do you take pride in building products…"

_id: ObjectId('66135543a0eb9183122be55a')
jid : 126
job_title : "Software Developer mfd"
job_description : "strongNXP Semiconductors NVstrong enables secure connections for a sma…"

_id: ObjectId('66135543a0eb9183122be529')
jid : 77
job_title : "Software Developer"
job_description : "INT Inc is the leading Data Visualization software provider for the la…"

_id: ObjectId('66135543a0eb9183122be58a')
jid : 174
job_title : "Application Developer equensWorldline"
job_description : "equensWorldline is one of the leading and most innovative payment serv…"

_id: ObjectId('66135543a0eb9183122be5c1')
jid : 229
job_title : "Mindshare I Data Scientist"
job_description : "GroupM is the worldrsquos largest media investment company and are a p…"
```

Execution time:



Explain Plan

Explain provides key execution metrics that help diagnose slow queries and optimize index usage. Learn more

TEXT_MATCH — Returned 27266 — Execution Time 0 ms

FETCH — Returned 27266 — Execution Time 42 ms

OR — Returned 27266 — Execution Time 45 ms

IXSCAN — Returned 24756 — Execution Time 36 ms
Index Name: job_description_text
Multi Key Index: yes

IXSCAN — Returned 13631 — Execution Time 29 ms
Index Name: job_description_text
Multi Key Index: yes

Query Performance Summary

- 27266 documents returned
- 27266 documents examined
- 174 ms execution time
- Is not sorted in memory
- 38387 index keys examined

Query used the following index:

_fts (text)

_ftsx ↑

Comparison:

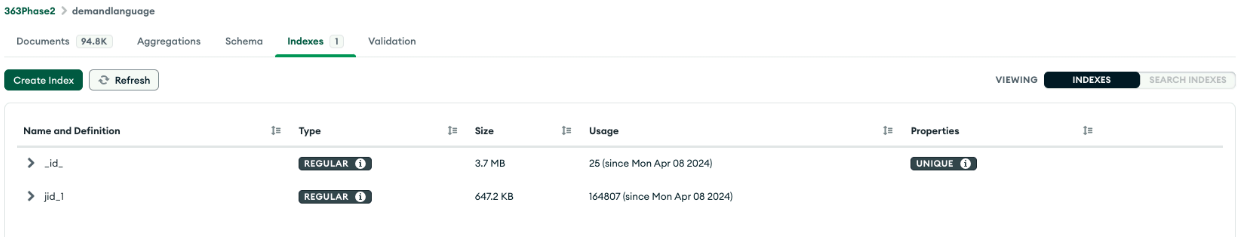
The index shortened the execution time from 860ms to 174ms.

A query that provides some aggregate data (i.e. number of entities satisfying a criteria)

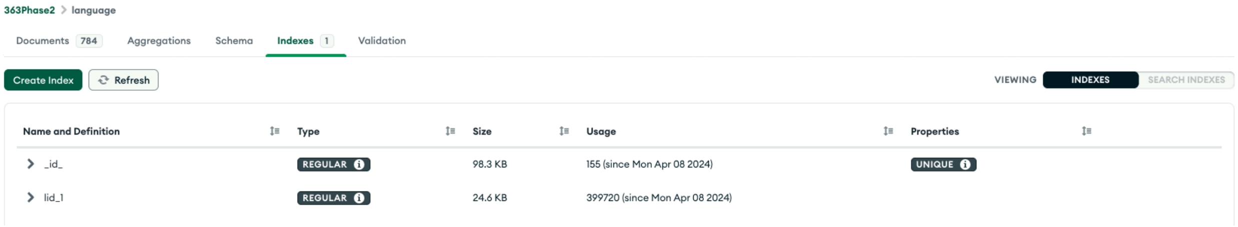
Pipeline saved to Job “INDEX totalJobsWithPython A query that provides some aggregate data (i.e. number of entities satisfying a criteria)”

This pipeline finds the total number of jobs that demand python using indexes.

I created an index on jid in demandlanguage and an index on lid in language collections



Name and Definition	Type	Size	Usage	Properties
_id_	REGULAR	3.7 MB	25 (since Mon Apr 08 2024)	UNIQUE
jid_1	REGULAR	647.2 KB	164807 (since Mon Apr 08 2024)	

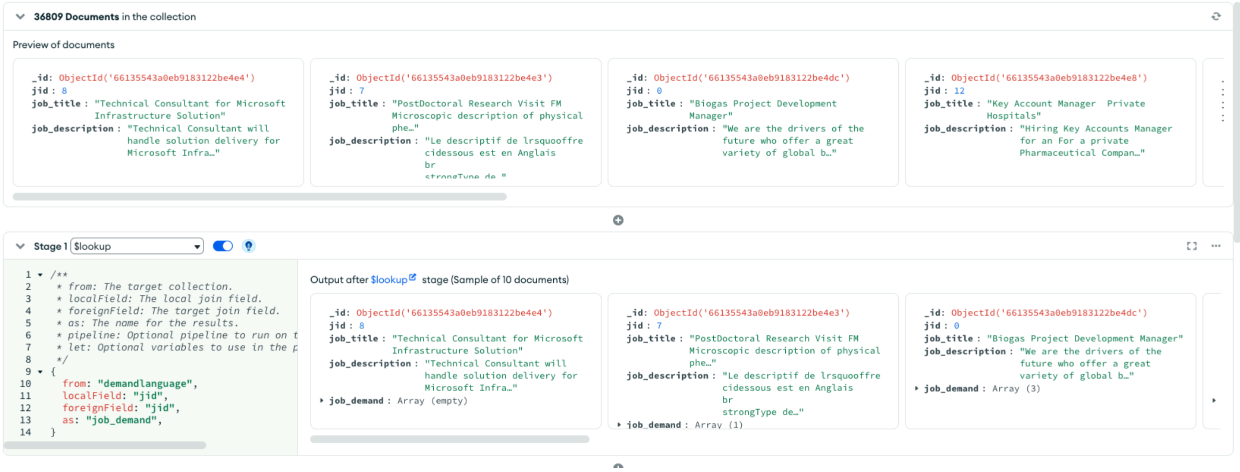


Name and Definition	Type	Size	Usage	Properties
_id_	REGULAR	98.3 KB	155 (since Mon Apr 08 2024)	UNIQUE
lid_1	REGULAR	24.6 KB	399720 (since Mon Apr 08 2024)	

Here is the query when I export it in JSON:

```
db.getCollection('job').aggregate(
  [
    {
      $lookup: {
        from: 'demandlanguage',
        localField: 'jid',
        foreignField: 'jid',
        as: 'job_demand'
      }
    },
    { $unwind: '$job_demand' },
    {
      $lookup: {
        from: 'language',
        localField: 'job_demand.lid',
        foreignField: 'lid',
        as: 'job_language'
      }
    },
    {
      $match: {
        'job_language.language': {
          $regex: 'python',
          $options: 'i'
        }
      }
    },
    { $count: 'totalJobsWithPython' }
  ],
  { maxTimeMS: 60000, allowDiskUse: true }
);
```

Stages:



36809 Documents in the collection

Preview of documents

```
_id: ObjectId('66135543a0eb9183122be4e4')
jid : 8
job_title : "Technical Consultant for Microsoft
            Infrastructure Solution"
job_description : "Technical Consultant will
                  handle solution delivery for
                  Microsoft Infra…"
```

```
_id: ObjectId('66135543a0eb9183122be4e3')
jid : 7
job_title : "PostDoctoral Research Visit FM
            Microscopic description of physical
            phe…"
job_description : "Le descriptif de lrsquooffre
                  cidessous est en Anglais
                  br
                  strongType de…"
```

```
_id: ObjectId('66135543a0eb9183122be4dc')
jid : 0
job_title : "Biogas Project Development
            Manager"
job_description : "We are the drivers of the
                  future who offer a great
                  variety of global b…"
```

```
_id: ObjectId('66135543a0eb9183122be4e8')
jid : 12
job_title : "Key Account Manager  Private
            Hospitals"
job_description : "Hiring Key Accounts Manager
                  for an For a private
                  Pharmaceutical Compan…"
```

Stage 1 $lookup

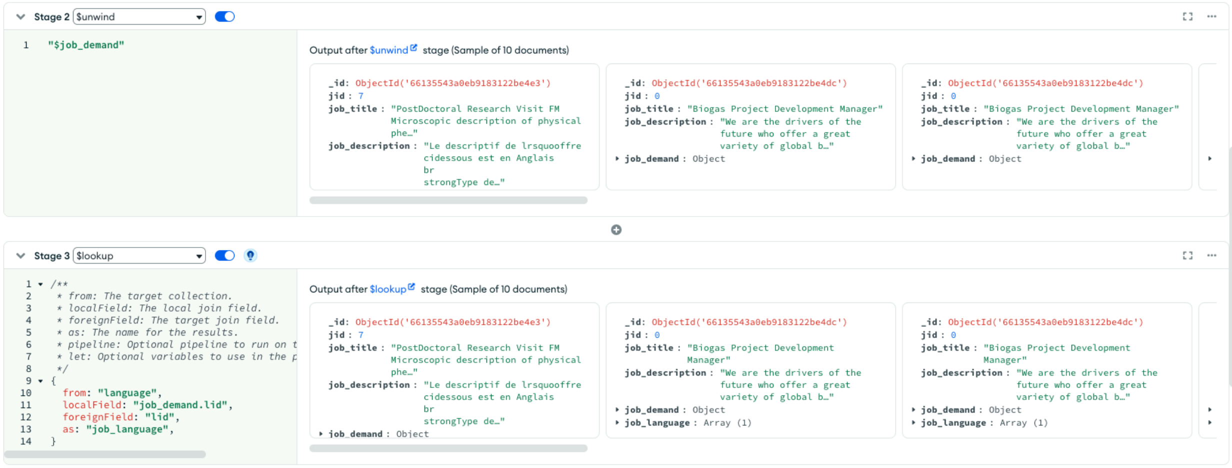
```
/**
 * from: The target collection.
 * localField: The local join field.
 * foreignField: The target join field.
 * as: The name for the results.
 * pipeline: Optional pipeline to run on t
 * let: Optional variables to use in the p
 */
{
  from: "demandlanguage",
  localField: "jid",
  foreignField: "jid",
  as: "job_demand",
}
```

Output after $lookup stage (Sample of 10 documents)

```
_id: ObjectId('66135543a0eb9183122be4e4')
jid : 8
job_title : "Technical Consultant for Microsoft
            Infrastructure Solution"
job_description : "Technical Consultant will
                  handle solution delivery for
                  Microsoft Infra…"
▸ job_demand : Array (empty)
```

```
_id: ObjectId('66135543a0eb9183122be4e3')
jid : 7
job_title : "PostDoctoral Research Visit FM
            Microscopic description of physical
            phe…"
job_description : "Le descriptif de lrsquooffre
                  cidessous est en Anglais
                  br
                  strongType de…"
▸ job_demand : Array (1)
```

```
_id: ObjectId('66135543a0eb9183122be4dc')
jid : 0
job_title : "Biogas Project Development Manager"
job_description : "We are the drivers of the
                  future who offer a great
                  variety of global b…"
▸ job_demand : Array (3)
```



Stage 2 $unwind

```
"$job_demand"
```

Output after $unwind stage (Sample of 10 documents)

```
_id: ObjectId('66135543a0eb9183122be4e3')
jid : 7
job_title : "PostDoctoral Research Visit FM
            Microscopic description of physical
            phe…"
job_description : "Le descriptif de lrsquooffre
                  cidessous est en Anglais
                  br
                  strongType de…"
```

```
_id: ObjectId('66135543a0eb9183122be4dc')
jid : 0
job_title : "Biogas Project Development Manager"
job_description : "We are the drivers of the
                  future who offer a great
                  variety of global b…"
▸ job_demand : Object
```

```
_id: ObjectId('66135543a0eb9183122be4dc')
jid : 0
job_title : "Biogas Project Development Manager"
job_description : "We are the drivers of the
                  future who offer a great
                  variety of global b…"
▸ job_demand : Object
```

Stage 3 $lookup

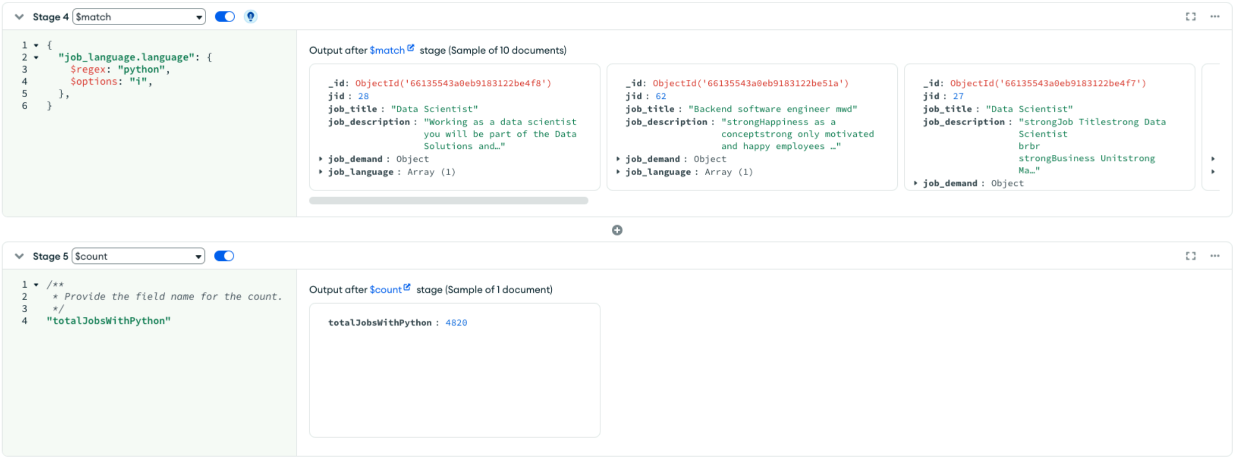
```
/**
 * from: The target collection.
 * localField: The local join field.
 * foreignField: The target join field.
 * as: The name for the results.
 * pipeline: Optional pipeline to run on t
 * let: Optional variables to use in the p
 */
{
  from: "language",
  localField: "job_demand.lid",
  foreignField: "lid",
  as: "job_language",
}
```

Output after $lookup stage (Sample of 10 documents)

```
_id: ObjectId('66135543a0eb9183122be4e3')
jid : 7
job_title : "PostDoctoral Research Visit FM
            Microscopic description of physical
            phe…"
job_description : "Le descriptif de lrsquooffre
                  cidessous est en Anglais
                  br
                  strongType de…"
▸ job_demand : Object
```

```
_id: ObjectId('66135543a0eb9183122be4dc')
jid : 0
job_title : "Biogas Project Development
            Manager"
job_description : "We are the drivers of the
                  future who offer a great
                  variety of global b…"
▸ job_demand : Object
▸ job_language : Array (1)
```

```
_id: ObjectId('66135543a0eb9183122be4dc')
jid : 0
job_title : "Biogas Project Development
            Manager"
job_description : "We are the drivers of the
                  future who offer a great
                  variety of global b…"
▸ job_demand : Object
▸ job_language : Array (1)
```



Stage 4 $match

```
{
  "job_language.language": {
    $regex: "python",
    $options: "i",
  },
}
```

Output after $match stage (Sample of 10 documents)

```
_id: ObjectId('66135543a0eb9183122be4f8')
jid : 28
job_title : "Data Scientist"
job_description : "Working as a data scientist
                  you will be part of the Data
                  Solutions and…"
▸ job_demand : Object
▸ job_language : Array (1)
```

```
_id: ObjectId('66135543a0eb9183122be51a')
jid : 62
job_title : "Backend software engineer mwd"
job_description : "strongHappiness as a
                  conceptstrong only motivated
                  and happy employees …"
▸ job_demand : Object
▸ job_language : Array (1)
```

```
_id: ObjectId('66135543a0eb9183122be4f7')
jid : 27
job_title : "Data Scientist"
job_description : "strongJob Titlestrong Data
                  Scientist
                  brbr
                  strongBusiness Unitstrong
                  Ma…"
▸ job_demand : Object
```

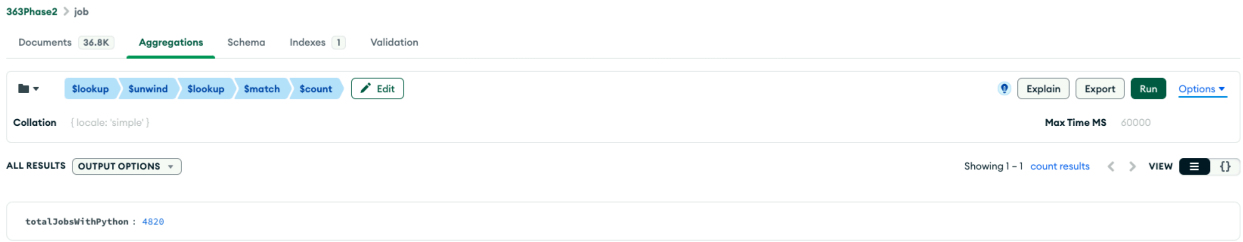
Stage 5 $count

```
/**
 * Provide the field name for the count.
 */
"totalJobsWithPython"
```

Output after $count stage (Sample of 1 document)

```
totalJobsWithPython : 4820
```

Output:

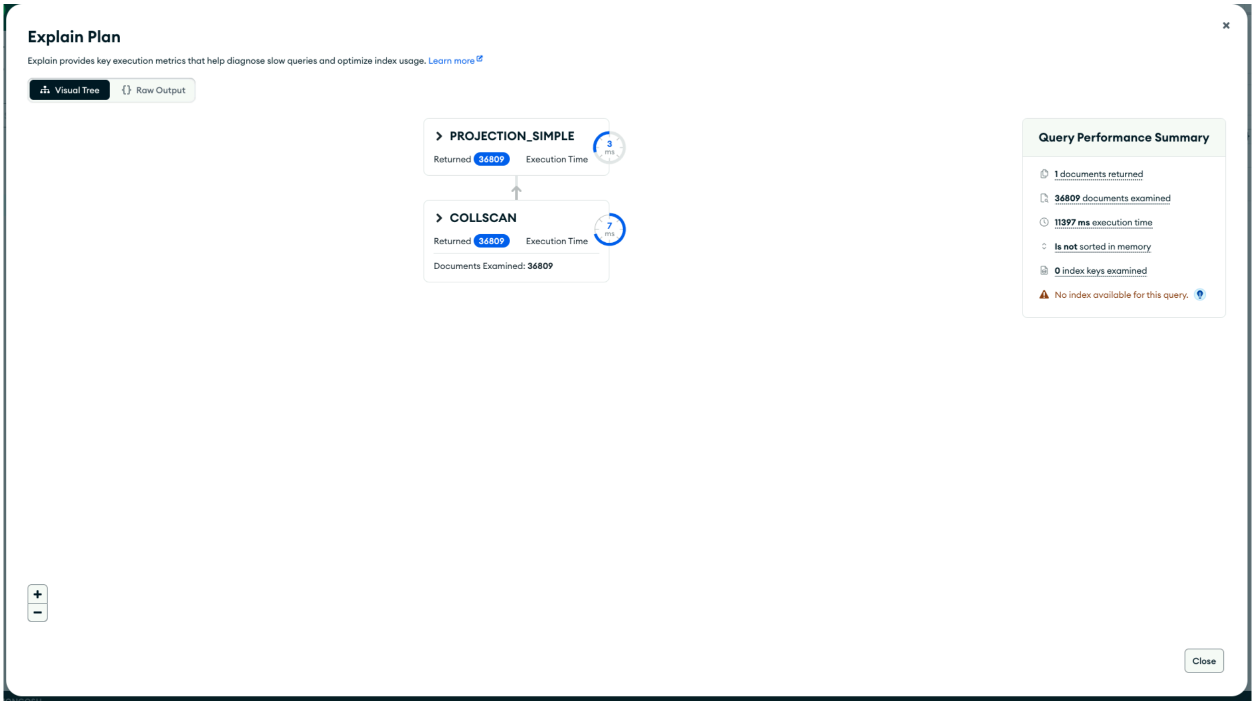


Collation { locale: 'simple' }

Max Time MS 60000

```
totalJobsWithPython : 4820
```

Execution time:



Explain Plan

Explain provides key execution metrics that help diagnose slow queries and optimize index usage. Learn more

Query Performance Summary

1 documents returned

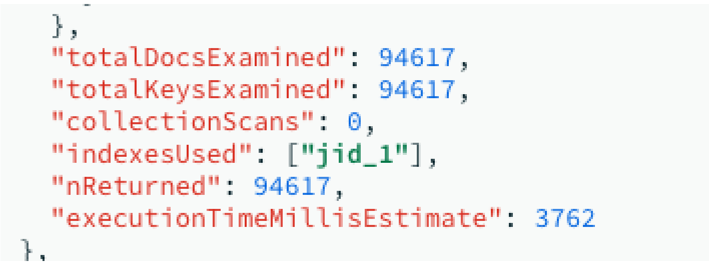
36809 documents examined

11397 ms execution time

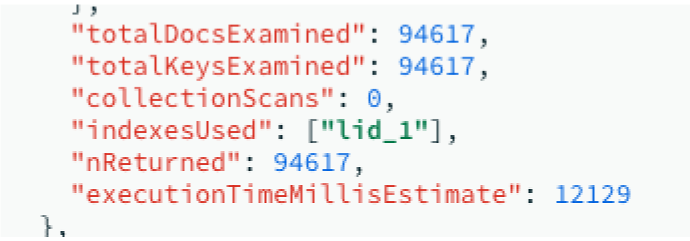
Is not sorted in memory

0 index keys examined

No index available for this query.



```
    },
    "totalDocsExamined": 94617,
    "totalKeysExamined": 94617,
    "collectionScans": 0,
    "indexesUsed": ["jid_1"],
    "nReturned": 94617,
    "executionTimeMillisEstimate": 3762
  },
```



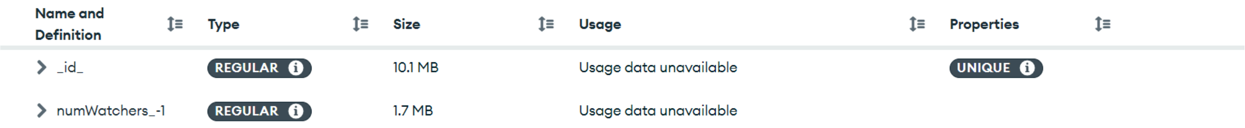
```
    ],
    "totalDocsExamined": 94617,
    "totalKeysExamined": 94617,
    "collectionScans": 0,
    "indexesUsed": ["lid_1"],
    "nReturned": 94617,
    "executionTimeMillisEstimate": 12129
  },
```

Comparison: The indexes shortened the execution time from 69510ms to 11397ms. I also did not need to increase the Max Time MS to 120000 while running the aggregation like I needed to for the one without the indexes.

ˆ **Find top n entities satisfying a criteria, sorted by an attribute.**

I created a pipeline titled “TopNEntitiesIndexed”. This query finds the top 10 repositories with the highest number of watchers.

I created an index on numWatchers in the table Repositories. Image below.

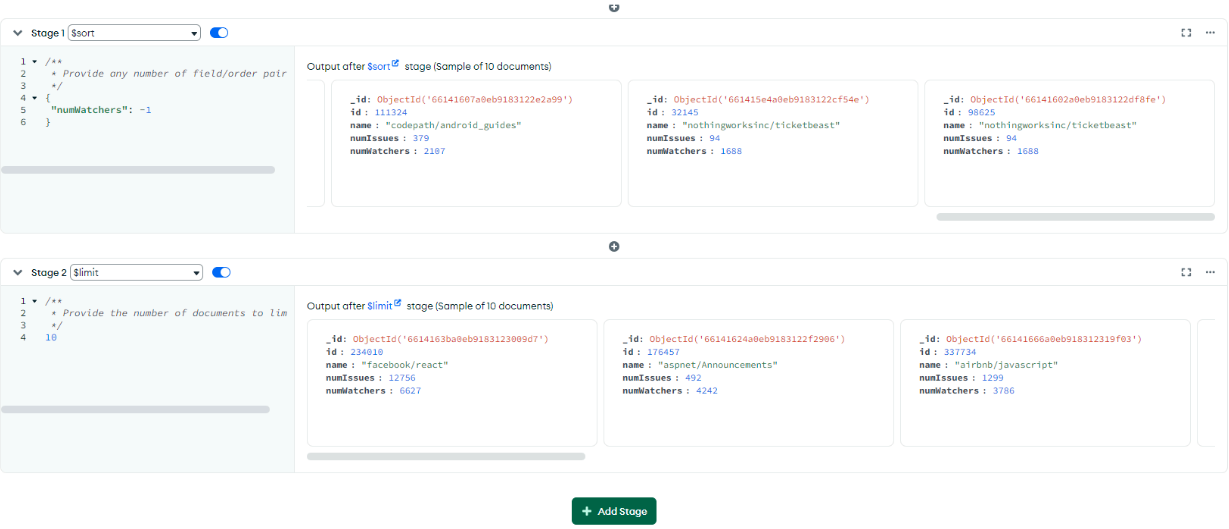


Name and Definition	Type	Size	Usage	Properties
› _id_	REGULAR	10.1 MB	Usage data unavailable	UNIQUE
› numWatchers_-1	REGULAR	1.7 MB	Usage data unavailable	

Now, I ran the same query as before. I will copy it below.

```
db.getCollection('repository').aggregate(
  [
    { $sort: { numWatchers: -1 } },
    { $limit: 10 }
  ],
  { maxTimeMS: 60000, allowDiskUse: true }
);
```

Stages:



Stage 1 $sort

```
/**
 * Provide any number of field/order pair
 */
{
  "numWatchers": -1
}
```

Output after $sort stage (Sample of 10 documents)

```
_id: ObjectId('66141607a0eb9183122e2a99')
id : 111324
name : "codepath/android_guides"
numIssues : 379
numWatchers : 2107
```

```
_id: ObjectId('661415e4a0eb9183122cf54e')
id : 32145
name : "nothingworksinc/ticketbeast"
numIssues : 94
numWatchers : 1688
```

```
_id: ObjectId('66141602a0eb9183122df8fe')
id : 98625
name : "nothingworksinc/ticketbeast"
numIssues : 94
numWatchers : 1688
```

Stage 2 $limit

```
/**
 * Provide the number of documents to lim
 */
10
```

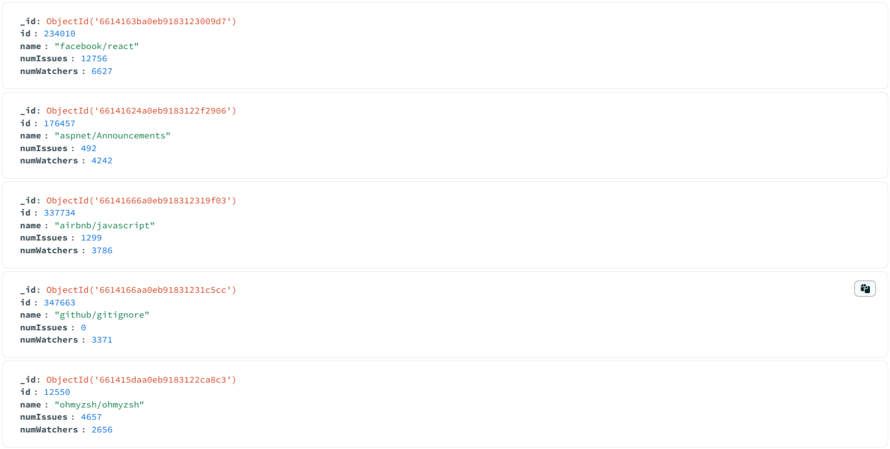
Output after $limit stage (Sample of 10 documents)

```
_id: ObjectId('6614163ba0eb9183123009d7')
id : 234010
name : "facebook/react"
numIssues : 12756
numWatchers : 6627
```

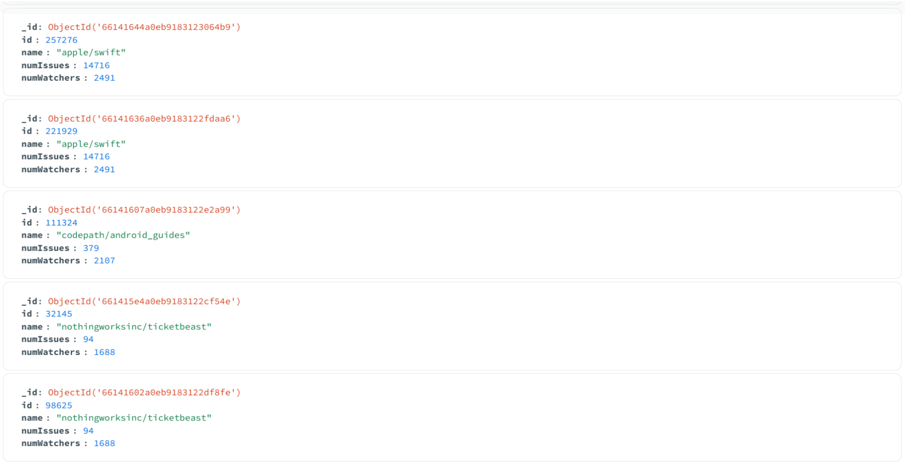
```
_id: ObjectId('66141624a0eb9183122f2906')
id : 176457
name : "aspnet/Announcements"
numIssues : 492
numWatchers : 4242
```

```
_id: ObjectId('66141666a0eb918312319f03')
id : 337734
name : "airbnb/javascript"
numIssues : 1299
numWatchers : 3786
```

Output:

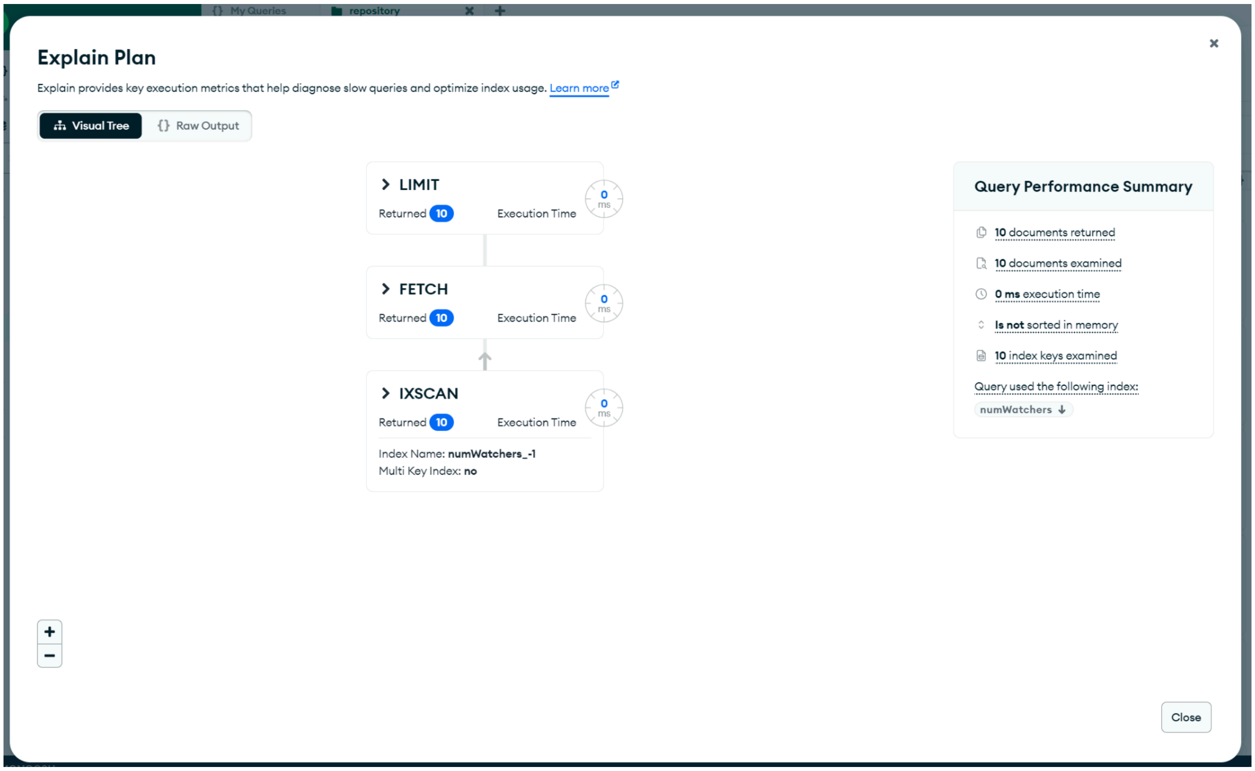


```
_id: ObjectId('6614163ba0eb9183123009d7')
id : 234010
name : "facebook/react"
numIssues : 12756
numWatchers : 6627

_id: ObjectId('66141624a0eb9183122f2906')
id : 176457
name : "aspnet/Announcements"
numIssues : 492
numWatchers : 4242

_id: ObjectId('66141666a0eb918312319f03')
id : 337734
name : "airbnb/javascript"
numIssues : 1299
numWatchers : 3786

_id: ObjectId('6614166aa0eb91831231c5cc')
id : 347663
name : "github/gitignore"
numIssues : 0
numWatchers : 3371

_id: ObjectId('661415daa0eb9183122ca8c3')
id : 12550
name : "ohmyzsh/ohmyzsh"
numIssues : 4657
numWatchers : 2656
```



```
_id: ObjectId('66141644a0eb9183123064b9')
id : 257276
name : "apple/swift"
numIssues : 14716
numWatchers : 2491

_id: ObjectId('66141636a0eb9183122fdaa6')
id : 221929
name : "apple/swift"
numIssues : 14716
numWatchers : 2491

_id: ObjectId('66141607a0eb9183122e2a99')
id : 111324
name : "codepath/android_guides"
numIssues : 379
numWatchers : 2107

_id: ObjectId('661415e4a0eb9183122cf54e')
id : 32145
name : "nothingworksinc/ticketbeast"
numIssues : 94
numWatchers : 1688

_id: ObjectId('66141602a0eb9183122df8fe')
id : 98625
name : "nothingworksinc/ticketbeast"
numIssues : 94
numWatchers : 1688
```

Execution time:



Explain provides key execution metrics that help diagnose slow queries and optimize index usage. Learn more

LIMIT
Returned 10 Execution Time 0 ms

FETCH
Returned 10 Execution Time 0 ms

IXSCAN
Returned 10 Execution Time 0 ms
Index Name: numWatchers_-1
Multi Key Index: no

Query Performance Summary

- 10 documents returned
- 10 documents examined
- 0 ms execution time
- Is not sorted in memory
- 10 index keys examined

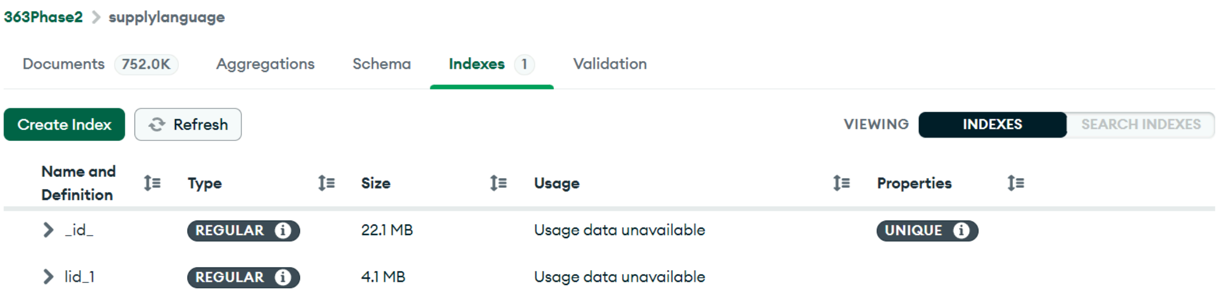
Query used the following index:
numWatchers ↓

Comparison:

As you can see, indexing this query will make the execution time equal to 0. While the other query had searches in the 20s of ms.

Simulate a relational group by query in NoSQL

Index created on lid from supplylanguage

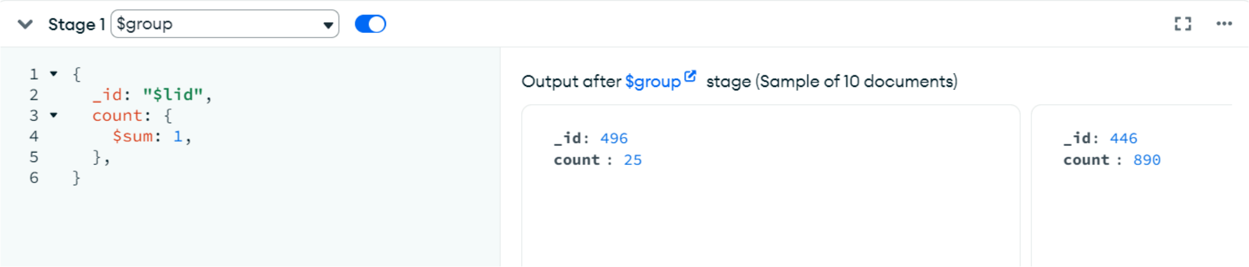


Name and Definition	Type	Size	Usage	Properties
_id_	REGULAR	22.1 MB	Usage data unavailable	UNIQUE
lid_1	REGULAR	4.1 MB	Usage data unavailable	

Query to group languages by the number of times they appear in a repo with index. (same query ran as before)

```
[
  {
    $group: {
      _id: "$lid",
      count: {
        $sum: 1,
      },
    },
  },
  {
    $lookup: {
      from: "language",
      localField: "_id",
      foreignField: "lid",
      as: "language",
    },
  },
  {
    $project: {
      language: {
        $arrayElemAt: ["$language.language", 0],
      },
      count: "$count",
    },
  },
]
```

Stages:

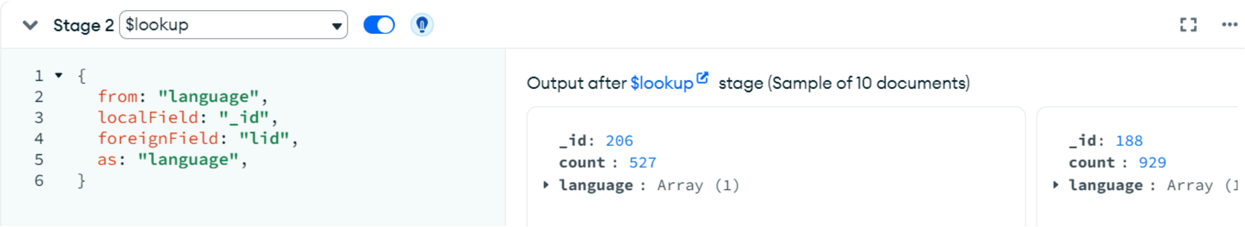


```
{
  _id: "$lid",
  count: {
    $sum: 1,
  },
}
```

Output after $group stage (Sample of 10 documents)

```
_id: 496
count : 25
```

```
_id: 446
count : 890
```

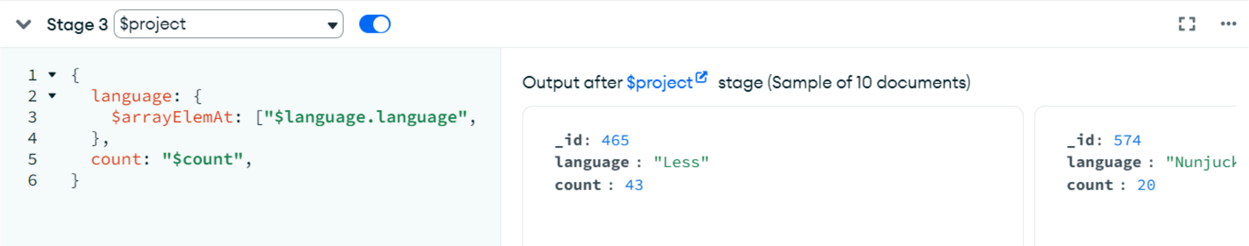


```
{
  from: "language",
  localField: "_id",
  foreignField: "lid",
  as: "language",
}
```

Output after $lookup stage (Sample of 10 documents)

```
_id: 206
count : 527
language : Array (1)
```

```
_id: 188
count : 929
language : Array (1
```



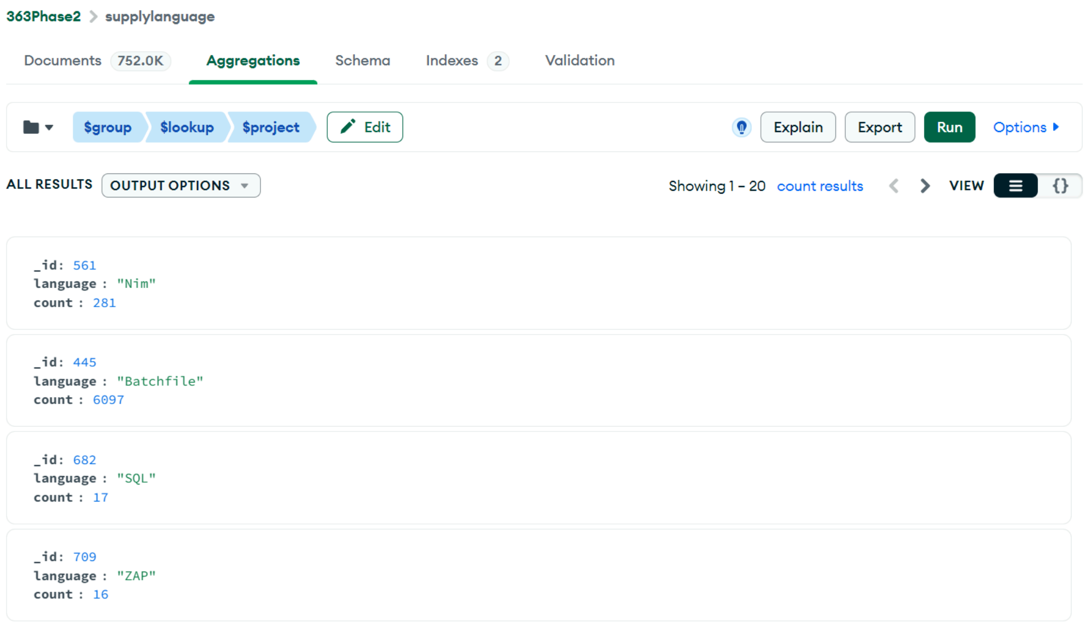
```
{
  language: {
    $arrayElemAt: ["$language.language",
  },
  count: "$count",
}
```

Output after $project stage (Sample of 10 documents)

```
_id: 465
language : "Less"
count : 43
```

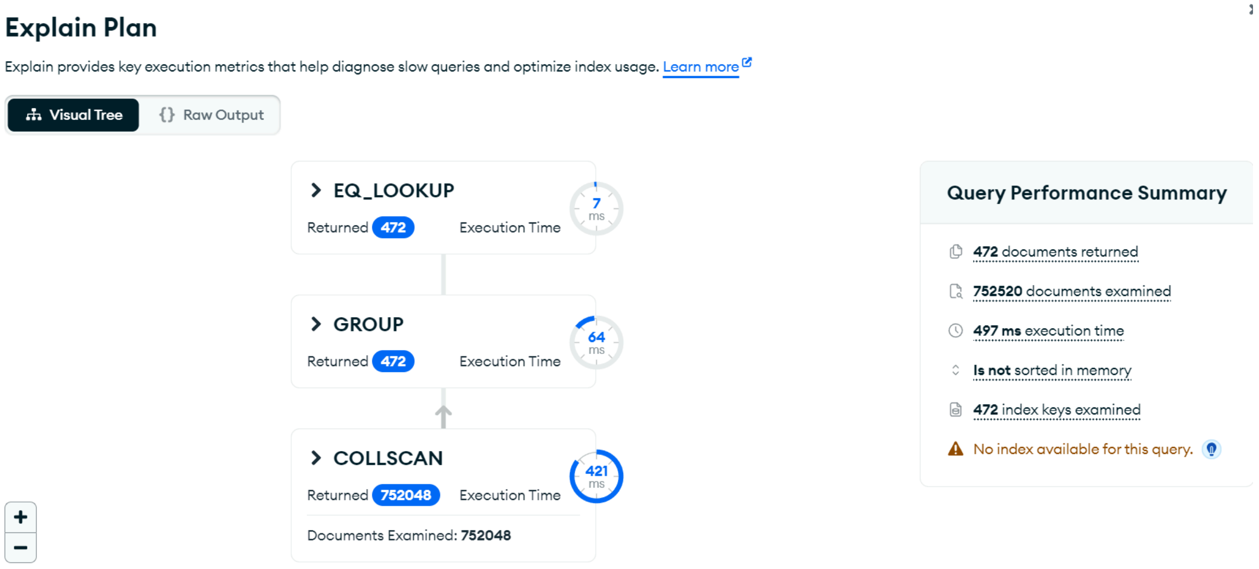
```
_id: 574
language : "Nunjuck
count : 20
```

Output:



```
_id: 561
language : "Nim"
count : 281

_id: 445
language : "Batchfile"
count : 6097

_id: 682
language : "SQL"
count : 17

_id: 709
language : "ZAP"
count : 16
```

Execution time:



Explain Plan

Explain provides key execution metrics that help diagnose slow queries and optimize index usage. Learn more

EQ_LOOKUP
Returned 472 Execution Time 7 ms

GROUP
Returned 472 Execution Time 64 ms

COLLSCAN
Returned 752048 Execution Time 421 ms
Documents Examined: 752048

Query Performance Summary

472 documents returned
752520 documents examined
497 ms execution time
Is not sorted in memory
472 index keys examined
No index available for this query.

Comparison:

We can see from the execution times that without an index it takes 90ms to group and while using an index it goes down to 64ms, almost 30ms faster. Running EG_LOOKUP and COLLSCAN do not have any significant time improvements.

ˆ Demonstrate a full-text search. Show the performance improvement by using indexes.

We performed a full-text search on our MongoDB collection, first using regex and afterwards using a search index. The term we are looking for is “python” under the job_description of our job collection. The code can be found at fullTextSearch.py

```
search_term = "python"
```

Regex

```
regex_pattern = r"{}".format(search_term) # search for word

   results = jobCollection.find({"job_description": {"$regex": regex_pattern, "$options": "i"}})

   print(results.explain())
```

Here, we are searching for the word “python” in a case-insensitive regex search.

The results.explain() of this query gives us the following information:

- nReturned (number of documents returned): 6327
- executionTimeMillis: 342
- totalDocsExamined: 36809

Search index

```
# Perform the search with text index (jobCol.create_index([("job_description", "text")],
name="job_description_text_index", default_language='english'))

   results2 = jobCollection.find({"$text": {"$search": search_term, "$caseSensitive": False}})

   print(results2.explain())
```

Here, we are searching for the word “python” after having created a MongoDB search index.

The results.explain() of this query gives us the following information:

- nReturned (number of documents returned): 5542
- executionTimeMillis: 14
- totalDocsExamined: 5542

Comparison

We can see that using a Search index is a lot faster (14ms vs 342ms) and a lot less documents need to be scanned (5542 vs 36809). Without an index, MongoDB has to scan every single document in the collection to return the query results as opposed to using an index to limit the amount of documents it has to scan.

However, the main drawback of using a Search index is that it doesn't support partial matches. MongoDB indexing uses stemming to match words. For example, if we are searching for "cat", it will match "cats" but not "caterpillar" or "bobcat". Full-text search using regex, on the other hand, allows us to search for partial text. This explains the discrepancy between the number of documents returned (6327 for regex vs 5542 with index).