Southern New Hampshire University

8-1 Assignment: Data Aggregation Pipeline

CS-340-Q7703

Jensen, Bryston
2-29-2024

Using the mongoimport tool, **create the database** “companies” by loading the documents found in the “companies.json” file into the “research” collection. This file is located in the “/usr/local/datasets/” directory in Apporto.

```
brystonjensen_snhu@78e20b13519a:/usr/local/datasets$ mongoimport --username="${MONGO_USER}" --password="${MONGO_PASS}" --port=${MONGO_PORT} --host=${MONGO_HOST} --d
b=companies --collection=research --authenticationDatabase=admin ./companies.json
2024-02-29T00:29:09.478+0000    connected to: mongodb://localhost:27017/
2024-02-29T00:29:11.315+0000    18801 document(s) imported successfully. 0 document(s) failed to import.
```

Verify your load by issuing the following queries:

db.research.find({"name" : "AdventNet"})

```
admin> show dbs
AAC         1.43 MiB
admin     156.00 KiB
city       16.60 MiB
companies  31.33 MiB
config     72.00 KiB
enron       7.64 MiB
local      80.00 KiB
admin> use companies
switched to db companies
companies> db.research.find({"name": "AdventNet"})
[
  {
    _id: ObjectId("52cdef7c4bab8bd675297d8b"),
    name: 'AdventNet',
    permalink: 'abc3',
    crunchbase_url: 'http://www.crunchbase.com/company/adventnet',
    homepage_url: 'http://adventnet.com',
    blog_url: '',
    blog_feed_url: '',
    twitter_username: 'manageengine',
    category_code: 'enterprise',
    number_of_employees: 600,
    founded_year: 1996,
    deadpooled_year: 2,
    tag_list: '',
    alias_list: 'Zoho ManageEngine ',
    email_address: 'pr@adventnet.com',
    phone_number: '925-924-9500',
    description: 'Server Management Software',
    created_at: ISODate("2007-05-25T19:24:22.000Z"),
    updated_at: 'Wed Oct 31 18:26:09 UTC 2012',
    overview: '<p>AdventNet is now <a href="/company/zoho-manageengine" title="Zoho ManageEngine" rel="nofollow">Zoho ManageEngine</a>.</p>\n' +
      '\n' +
      '<p>Founded in 1996, AdventNet has served a diverse range of enterprise IT, networking and telecom customers.</p>\n' +
      '\n' +
      '<p>AdventNet supplies server and network management software.</p>',
    image: {
      available_sizes: [
        [
          [ 150, 55 ],
          'assets/images/resized/0001/9732/19732v1-max-150x150.png'
        ],
        [
          [ 150, 55 ],
          'assets/images/resized/0001/9732/19732v1-max-250x250.png'
companies>
        [
          [ 150, 55 ],
          'assets/images/resized/0001/9732/19732v1-max-450x450.png'
        ]
      ]
    },
    products: [],
    relationships: [
      {
        is_past: true,
        title: 'CEO and Co-Founder',
        person: {
          first_name: 'Sridhar',
          last_name: 'Vembu',
          permalink: 'sridhar-vembu'
        }
      },
      {
        is_past: true,
        title: 'VP of Business Dev',
        person: {
          first_name: 'Neil',
```

```
        last_name: 'Butani',
        permalink: 'neil-butani'
      }
    },
    {
      is_past: true,
      title: 'Usabiliy Engineer',
      person: {
        first_name: 'Bharath',
        last_name: 'Balasubramanian',
        permalink: 'bharath-balasibramanian'
      }
    },
    {
      is_past: true,
      title: 'Director of Engineering',
      person: {
        first_name: 'Rajendran',
        last_name: 'Dandapani',
        permalink: 'rajendran-dandapani'
      }
    },
    {
      is_past: true,
      title: 'Market Analyst',
      person: {
        first_name: 'Aravind',
        last_name: 'Natarajan',
        permalink: 'aravind-natarajan'
      }
    },
    {
      is_past: true,
      title: 'Director of Product Management',
      person: {
        first_name: 'Hyther',
        last_name: 'Nizam',
        permalink: 'hyther-nizam'
      }
    },
    {
      is_past: true,
      title: 'Western Regional OEM Sales Manager',
      person: {
        first_name: 'Ian',
        last_name: 'Wenig',
        permalink: 'ian-wenig'
      }
    }
  ],
  competitions: [],
  providerships: [
    {
      title: 'DHFH',
      is_past: true,
      provider: { name: 'A Small Orange', permalink: 'a-small-orange' }
    }
  ],
  total_money_raised: '$0',
  funding_rounds: [],
  investments: [],
  acquisition: null,
  acquisitions: [],
  offices: [
    {
      description: 'Headquarters',
      address1: '4900 Hopyard Rd.',
      address2: 'Suite 310',
      zip_code: '94588',
      city: 'Pleasanton',
      state_code: 'CA',
      country_code: 'USA',
      latitude: 37.692934,
      longitude: -121.904945
    }
  ],
  milestones: [],
  video_embeds: [],
  screenshots: [
    {
      available_sizes: [
        [
          [ 150, 94 ],
          'assets/images/resized/0004/3400/43400v1-max-150x150.png'
        ],
        [
          [ 250, 156 ],
          'assets/images/resized/0004/3400/43400v1-max-250x250.png'
```

```
        ],
        [
          [ 450, 282 ],
          'assets/images/resized/0004/3400/43400v1-max-450x450.png'
        ]
      ],
      attribution: null
    }
  ],
  external_links: [],
  partners: []
 }
]
```

db.research.find({"founded_year" : 1996},{"name" : 1}).limit(10)

```
companies> db.research.find({"founded_year" : 1996},{"name" : 1}).limit(10)
[
  { _id: ObjectId("52cdef7c4bab8bd675297d8b"), name: 'AdventNet' },
  { _id: ObjectId("52cdef7c4bab8bd675297e24"), name: 'RegOnline' },
  { _id: ObjectId("52cdef7c4bab8bd675297e42"), name: 'LiveWorld' },
  { _id: ObjectId("52cdef7c4bab8bd675297f26"), name: 'Shopzilla' },
  { _id: ObjectId("52cdef7c4bab8bd675297fca"), name: 'LinkShare' },
  { _id: ObjectId("52cdef7c4bab8bd675298032"), name: 'MSNBC' },
  { _id: ObjectId("52cdef7c4bab8bd6752980cf"), name: 'TheStreet' },
  { _id: ObjectId("52cdef7c4bab8bd6752980d0"), name: 'Omniture' },
  { _id: ObjectId("52cdef7c4bab8bd6752980df"), name: 'Blucora' },
  { _id: ObjectId("52cdef7c4bab8bd675298223"), name: 'Alexa' }
]
```

Perform the following tasks **using MongoDB queries**:

List only the first 20 names of companies founded after the year 2010, ordered alphabetically.

```
companies> db.research.find({"founded_year": {$gte: 2010}}, {"_id": 0, "name": 1}).limit(20).sort({"name": 1})
[
  { name: '4shared' },
  { name: 'Abengoa' },
  { name: 'Advaliant' },
  { name: 'Advisor' },
  { name: 'Ajiel' },
  { name: 'AppNeta' },
  { name: 'Aquavation' },
  { name: 'AudioBoo' },
  { name: 'Avenir Medical' },
  { name: 'BASH Gaming' },
  { name: 'Baveo' },
  { name: 'BizEquity' },
  { name: 'Bling Easy' },
  { name: 'Blurtt' },
  { name: 'Carfeine' },
  { name: 'Cellogic' },
  { name: 'Cellogic' },
  { name: 'Chicisimo' },
  { name: 'CircleUp' },
  { name: 'Clowdy' }
]
Type "it" for more
```

List only the first 20 names of companies with offices in either California or Texas, ordered by the number of employees and sorted largest to smallest.

```
companies> db.research.find({$or: [{"offices.state_code": "TX"}, {"offices.state_code": "CA"}]}, {"_id": 0, "name": 1, "number_of_employees": 1}).limit(20).sort({"n
umber_of_employees": -1})
[
  { name: 'PayPal', number_of_employees: 300000 },
  { name: 'Samsung Electronics', number_of_employees: 221726 },
  { name: 'Accenture', number_of_employees: 205000 },
  { name: 'Flextronics International', number_of_employees: 200000 },
  { name: 'Safeway', number_of_employees: 186000 },
  { name: 'Sony', number_of_employees: 180500 },
  { name: 'Intel', number_of_employees: 86300 },
  { name: 'Dell', number_of_employees: 80000 },
  { name: 'Apple', number_of_employees: 80000 },
  { name: 'ExxonMobil', number_of_employees: 76900 },
  { name: 'Affiliated Computer Services', number_of_employees: 74000 },
  { name: 'Cisco', number_of_employees: 63000 },
  { name: 'Sun Microsystems', number_of_employees: 33350 },
  { name: 'Texas Instruments', number_of_employees: 30175 },
  { name: 'Google', number_of_employees: 28000 },
  { name: 'The Walt Disney Company', number_of_employees: 25000 },
  { name: 'Avaya', number_of_employees: 18000 },
  { name: 'AMD', number_of_employees: 16420 },
  { name: 'Experian', number_of_employees: 15500 },
  { name: 'eBay', number_of_employees: 15000 }
]
```

Design and implement a MongoDB aggregation pipeline to show the total number of offices by state for all companies that have offices in the United States.

Okay so in the pipeline there are four main things happening.

- First it is told to collect (`aggregate`) the following. This is what starts the pipeline.
- Second, we **unwind** the `offices` field for everything **matches** the country code `USA`. This means that all the offices are separated into separate documents for proper counting, accounting for different offices in separate states.
- Third, everything is grouped by state code (`field = _id`), and the `number_of_employees` are each added (**summed**) together based on the state and placed in the field `num_of_employees_per_state`.
- Lastly, it sorts the `_id` field by smallest to largest, or alphabetically.

(Screenshot on next page)

```
companies> db.research.aggregate([{ $unwind: "$offices" }, {$match: {"offices.country_code": "USA"}}, {$group: {_id: "$offices.state_code", num_of_employees_per_sta
te: {$sum: "$number_of_employees"}}}, {$sort: {_id: 1}}])
[
  { _id: null, num_of_employees_per_state: 68453 },
  { _id: 'AL', num_of_employees_per_state: 57794 },
  { _id: 'AR', num_of_employees_per_state: 12412 },
  { _id: 'AZ', num_of_employees_per_state: 54181 },
  { _id: 'CA', num_of_employees_per_state: 2149593 },
  { _id: 'CO', num_of_employees_per_state: 69188 },
  { _id: 'CT', num_of_employees_per_state: 1585 },
  { _id: 'DC', num_of_employees_per_state: 37964 },
  { _id: 'DE', num_of_employees_per_state: 1639 },
  { _id: 'FL', num_of_employees_per_state: 62936 },
  { _id: 'GA', num_of_employees_per_state: 333721 },
  { _id: 'HI', num_of_employees_per_state: 173 },
  { _id: 'IA', num_of_employees_per_state: 3268 },
  { _id: 'ID', num_of_employees_per_state: 90135 },
  { _id: 'IL', num_of_employees_per_state: 279012 },
  { _id: 'IN', num_of_employees_per_state: 360 },
  { _id: 'KS', num_of_employees_per_state: 337 },
  { _id: 'KY', num_of_employees_per_state: 1950 },
  { _id: 'LA', num_of_employees_per_state: 335 },
  { _id: 'MA', num_of_employees_per_state: 315187 }
]
Type "it" for more
companies> it
[
  { _id: 'MD', num_of_employees_per_state: 3811 },
  { _id: 'ME', num_of_employees_per_state: 275 },
  { _id: 'MI', num_of_employees_per_state: 244158 },
  { _id: 'MN', num_of_employees_per_state: 81857 },
  { _id: 'MO', num_of_employees_per_state: 11006 },
  { _id: 'MS', num_of_employees_per_state: 158 },
  { _id: 'MT', num_of_employees_per_state: 847 },
  { _id: 'NC', num_of_employees_per_state: 49802 },
  { _id: 'ND', num_of_employees_per_state: 18 },
  { _id: 'NE', num_of_employees_per_state: 1579 },
  { _id: 'NH', num_of_employees_per_state: 1464 },
  { _id: 'NJ', num_of_employees_per_state: 118693 },
  { _id: 'NM', num_of_employees_per_state: 109 },
  { _id: 'NV', num_of_employees_per_state: 2705 },
  { _id: 'NY', num_of_employees_per_state: 671303 },
  { _id: 'OH', num_of_employees_per_state: 11362 },
  { _id: 'OK', num_of_employees_per_state: 538 },
  { _id: 'OR', num_of_employees_per_state: 39144 },
  { _id: 'PA', num_of_employees_per_state: 142780 },
  { _id: 'RI', num_of_employees_per_state: 7714 }
]
Type "it" for more
companies> it
[
  { _id: 'SC', num_of_employees_per_state: 4038 },
  { _id: 'SD', num_of_employees_per_state: 485 },
  { _id: 'TN', num_of_employees_per_state: 14951 },
  { _id: 'TX', num_of_employees_per_state: 393037 },
  { _id: 'UT', num_of_employees_per_state: 10354 },
  { _id: 'VA', num_of_employees_per_state: 334707 },
  { _id: 'VT', num_of_employees_per_state: 187 },
  { _id: 'WA', num_of_employees_per_state: 227730 },
  { _id: 'WI', num_of_employees_per_state: 242 },
  { _id: 'WV', num_of_employees_per_state: 5 },
  { _id: 'WY', num_of_employees_per_state: 229 }
]
```