CSC 369: Distributed Computing

Alex Dekhtyar

April 17

Day 6: The Algebra Of Data Transformations Part II

db.collection.aggregate()
Housekeeping

- Lab 2:
  - Submit from unix1-2-3-4-5
  - handin dekhtyar lab02 <files>
- Lab 3:
- Lab 4: Python application, teams of 2.
Recall from last class

{name:"Alex",
teaches: ["CSC 369", "DATA 452"],
department: "CSSE",
enrollments: [28, 20],
position: "professor",
office: {building: 14, room: 210}}

Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

Decomposition into atomic operations in “Generalized” Data Algebra
Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

- Keep only CSSE instructors
- Remove unnecessary data
- Find the total enrollment for each CSSE instructor and number of sections taught
- Find the largest total enrollment for a CSSE instructor
- Compare each instructor’s total enrollment to the largest; keep only instructors with largest enrollment
Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment.

1. Keep only CSSE instructors.
2. Remove unnecessary data.
3. Find the total enrollment for each CSSE instructor and number of sections taught.
4. Find the largest total enrollment for a CSSE instructor.
5. Compare each instructor’s total enrollment to the largest; keep only instructors with largest enrollment.

Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

1. Deconstruct “teaches” arrays, create one object per instructor-course pairing.
2. Keep information about only “CSC”, “CPE”, and “DATA” courses.
3. Remove unnecessary data.
4. For each course, combine instructors teaching it into a list.
5. Sort?
Express using these operations....

**Relational Algebra**
- Selection
- Projection
- Set Operations
- Join
- Grouping/Aggregation
- Sort

**Generalized Algebra**
- Filtering
- Projection/Transformation
- Join
- Grouping/Aggregation
- Sort
Very Tersely

Filtering
Given a condition - keep only objects that satisfy it

Projection Transformation
Modify the contents of its object based solely on what’s in the object itself

Grouping
Break collection into groups, each representing objects with same values of some keys

Aggregation
Compute an aggregate value over a set of objects

Join
Combine objects from two different collections based on matches in values of some keys

Sort
Return objects in a specific order
... and a few more

<table>
<thead>
<tr>
<th>Ungrouping</th>
<th>Opposite of grouping - build an object for each element of an array</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unwinding</td>
<td></td>
</tr>
<tr>
<td>Limit</td>
<td>Return a specific number of documents</td>
</tr>
<tr>
<td>Skip</td>
<td>Return documents after skipping a specified number</td>
</tr>
<tr>
<td>Sample</td>
<td>Return a random sample of documents</td>
</tr>
</tbody>
</table>
Let’s use our intuition
Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

Remove unnecessary data

Find the total enrollment for each CSSE instructor and number of sections taught

Find the largest total enrollment for a CSSE instructor

Compare each instructor’s total enrollment to the largest; keep only instructors with largest enrollment

Exercise Time!!

File posted to Slack/chat.
Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

1. Keep only CSSE instructors
2. Remove unnecessary data
3. Find the total enrollment for each CSSE instructor and number of sections taught
4. Find the largest total enrollment for a CSSE instructor
5. Compare each instructor’s total enrollment to the largest; keep only instructors with largest enrollment
Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

Deconstruct “teaches” arrays, create one object per instructor-course pairing

Keep information about only “CSC”, “CPE”, and “DATA” courses.

Remove unnecessary data

For each course, combine instructors teaching it into a list

Sort?
Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

Deconstruct “teaches” arrays, create one object per instructor-course pairing

- Keep information about only “CSC”, “CPE”, and “DATA” courses.
- Remove unnecessary data
- For each course, combine instructors teaching it into a list
- Sort?
Now, let’s learn all this for real...

`db.collection.aggregate(<aggregation pipeline>)`
Now, let’s learn all this for real...

db.collection.aggregate({$operation:{<doc>}},
  {$operation:{<doc>}},
  ...
  )
Now, let’s learn all this for real...

db.collection.aggregate({$operation:{<doc>}},
    {$operation:{<doc>}},
    ...
})

**In Lecture:** basic ideas  
**In Lab:** all the syntax you can handle
Let’s Learn By Doing

Query 2 first (it is simpler)

Query 1 second (it has layers)
Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

Deconstruct “teaches” arrays, create one object per instructor-course pairing

Keep information about only “CSC”, “CPE”, and “DATA” courses.

Remove unnecessary data

For each course, combine instructors teaching it into a list

Sort?
Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

Deconstruct “teaches” arrays, create one object per instructor-course pairing

Keep information about only “CSC”, “CPE”, and “DATA” courses.

Remove unnecessary data

For each course, combine instructors teaching it into a list

Sort?
Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

Deconstruct “teaches” arrays, create one object per instructor-course pairing.

$unwind

Keep information about only “CSC”, “CPE”, and “DATA” courses.

$match

Remove unnecessary data.

$project

For each course, combine instructors teaching it into a list.

$group

Sort?
Q2: Report a list of instructors for each "CSC", "CPE" and "DATA" course. For each instructor, list name and department.

Deconstruct “teaches” arrays, create one object per instructor-course pairing

\$unwind

{\$unwind: "\$teaches"}

Keep information about only “CSC”, “CPE”, and “DATA” courses.

\$match

Remove unnecessary data

\$project

For each course, combine instructors teaching it into a list

\$group

Sort?
Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

Deconstruct “teaches” arrays, create one object per instructor-course pairing

$unwind

{$unwind: "$teaches"}

Keep information about only “CSC”, “CPE”, and “DATA” courses.

$match

{$match: 
  {teaches: {$in: ["CSC", "CPE", "DATA"]}}
}

Remove unnecessary data

$project

For each course, combine instructors teaching it into a list

$group

Sort?

$sort
Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

Deconstruct “teaches” arrays, create one object per instructor-course pairing

$unwind

Keep information about only “CSC”, “CPE”, and “DATA” courses.

$match

Remove unnecessary data

$project

For each course, combine instructors teaching it into a list

$group

Regular Expressions!!!
Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

Deconstruct “teaches” arrays, create one object per instructor-course pairing

$\text{Deconstruct “teaches” arrays, create one object per instructor-course pairing}$

$\text{Remove unnecessary data}$

For each course, combine instructors teaching it into a list

$\text{For each course, combine instructors teaching it into a list}$

$\text{Sort?}$

Keep information about only “CSC”, “CPE”, and “DATA” courses.

$\text{Keep information about only “CSC”, “CPE”, and “DATA” courses.}$

$\text{Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.}$

$\text{Deconstruct “teaches” arrays, create one object per instructor-course pairing}$

$\text{Remove unnecessary data}$

$\text{For each course, combine instructors teaching it into a list}$

$\text{Sort?}$
Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

Deconstruct “teaches” arrays, create one object per instructor-course pairing

- $unwind

Keep information about only “CSC”, “CPE”, and “DATA” courses.

- $match

Remove unnecessary data

- $project

For each course, combine instructors teaching it into a list

- $group

Sort?

- $sort

```javascript
{$unwind: "$teaches"}

{$match:
  {teaches: {$in: ["/CSC/", "/CPE/", "/DATA/"]}}
}

{$project:
  {_id:0,
    Name:1, department:1,
    course:"$teaches"}
}

"office" : {
  "building" : 14,
  "room" : 210
}
```
Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

Deconstruct “teaches” arrays, create one object per instructor-course pairing

$unwind

Keep information about only “CSC”, “CPE”, and “DATA” courses.

$match

Remove unnecessary data

$project

For each course, combine instructors teaching it into a list

$group

Sort?
Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

Deconstruct “teaches” arrays, create one object per instructor-course pairing

Keep information about only “CSC”, “CPE”, and “DATA” courses.

Remove unnecessary data

For each course, combine instructors teaching it into a list

Sort?

{$unwind: "$teaches"}$

{$match:
  {teaches: {$in: [/^CSC/,
    /^CPE/,/^DATA/]}}}}

{$project:{_id:0,
  Name:1, department:1,
  course:"$teaches"}}

{$group:{_id:"$course",
  instructors:{$push: {
    name:"$name",
    department:"$department"}}}
}

{$project:{_id:0, instructors:1,
  $course:"$_id"}}
Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

Deconstruct “teaches” arrays, create one object per instructor-course pairing

Keep information about only “CSC”, “CPE”, and “DATA” courses.

Remove unnecessary data

For each course, combine instructors teaching it into a list

Sort?

```javascript
{$unwind: "$teaches"}

{$match:
  {teaches: {$in: ['/^CSC/','/^CPE/','/^DATA/']}}}

{$project:
  {_id:0,
   Name:1, department:1,
   course:'$teaches'}}

{$group:
  {_id:'$course',
   instructors:{$push: {
     name:'$name',
     department:'$department'}}}
  }

{$project:
  {_id:0, instructors:1, $course:'$_id'}}
```
Q2: Report a list of instructors for each “CSC”, “CPE” and “DATA” course. For each instructor, list name and department.

Deconstruct “teaches” arrays, create one object per instructor-course pairing

Keep information about only “CSC”, “CPE”, and “DATA” courses.

Remove unnecessary data

For each course, combine instructors teaching it into a list

Sort?

```json
{$unwind: "$teaches"}

{$match:
  {teaches: {$in: [/^CSC/,
                 /^CPE/,
                 /^DATA/]}}}

{$project:
  {_id:0,
   Name:1, department:1,
   course:"$teaches"}}

{$group:
  {_id:"$course",
   instructors:[
     {name:"$name",
      department:"$department"}
   ]}
}

{$project:
  {_id:0, instructors:1,
   course:"$_id"}}

{$sort: {course:1}}
```
db.spring.aggregate(
    {
        $unwind: "$teaches",
        $match: {teaches: {$in: ["/^CSC/", "/^CPE/", "/^DATA/"]}}
    },
    {
        $project: {
            _id: 0,
            name: 1, department: 1,
            course: "$teaches"
        }
    },
    {
        $group: {
            _id: "$course",
            instructors: {
                $push: {name: "$name",
                    department: "$department"}
            }
        }
    },
    {
        $project: {
            _id: 0, instructors: 1,
            course: "$_id"
        }
    },
    {
        $sort: {course: 1}
    }
)
Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment.

- Keep only CSSE instructors
- Remove unnecessary data
- Find the total enrollment for each CSSE instructor and number of sections taught
- Find the largest total enrollment for a CSSE instructor
- Compare each instructor’s total enrollment to the largest; keep only instructors with largest enrollment
Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment.

1. Keep only CSSE instructors
2. Remove unnecessary data
3. Find the total enrollment for each CSSE instructor and number of sections taught
4. Find the largest total enrollment for a CSSE instructor
5. Compare each instructor’s total enrollment to the largest; keep only instructors with largest enrollment
Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

- Keep only CSSE instructors
  - \$match
  - \{department: "CSSE"\}

- Remove unnecessary data
  - \$project

- Find the total enrollment for each CSSE instructor and number of sections taught
  - \$project

- Find the largest total enrollment for a CSSE instructor
  - \$group

- Compare each instructor’s total enrollment to the largest; keep only instructors with largest enrollment
  - \$match
Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

Remove unnecessary data

Find the total enrollment for each CSSE instructor and number of sections taught

Find the largest total enrollment for a CSSE instructor

Compare each instructor’s total enrollment to the largest; keep only instructors with largest enrollment
Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

Remove unnecessary data

Find the total enrollment for each CSSE instructor and number of sections taught

Find the largest total enrollment for a CSSE instructor

Compare each instructor’s total enrollment to the largest; keep only instructors with largest enrollment

{$match: {department: "CSSE"}}

{$project: {_id: 0, name: 1, enrollments: 1} }
Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

Remove unnecessary data

Find the total enrollment for each CSSE instructor and number of sections taught

Find the largest total enrollment for a CSSE instructor

Compare each instructor’s total enrollment to the largest; keep only instructors with largest enrollment

```
{$match: {department: "CSSE"}}

{$project: {_id: 0, name: 1, enrollments: 1}}

{$group}

{$project: {name: 1, enrollments: {$sum: "$enrollments"}}}

{ "name": "Alex", "enrollments": [ 28, 20 ] }
```
Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

Remove unnecessary data

Find the total enrollment for each CSSE instructor and number of sections taught

Find the largest total enrollment for a CSSE instructor

Compare each instructor’s total enrollment to the largest; keep only instructors with largest enrollment
Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

Remove unnecessary data

Find the total enrollment for each CSSE instructor and number of sections taught

Find the largest total enrollment for a CSSE instructor

Compare each instructor’s total enrollment to the largest; keep only instructors with largest enrollment

Keep only CSSE instructors

Remove unnecessary data

Find the total enrollment for each CSSE instructor and number of sections taught

Find the largest total enrollment for a CSSE instructor

Compare each instructor’s total enrollment to the largest; keep only instructors with largest enrollment

Keep only CSSE instructors

Remove unnecessary data

Find the total enrollment for each CSSE instructor and number of sections taught

Find the largest total enrollment for a CSSE instructor

Compare each instructor’s total enrollment to the largest; keep only instructors with largest enrollment

```json
{ $match: { department: "CSSE" } }

{ $project: { _id: 0, name: 1, enrollments: 1 } }

{ $project: { name: 1, enrollments: { $sum: "$enrollments" } } }

{ $group: { _id: "1", mEnr: { $max: "$enrollments" } } }

{ "name": "Alex", "enrollments": 48 }  
{ "name": "Kirsten", "enrollments": 108 } 
```
Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

Remove unnecessary data

Find the total enrollment for each CSSE instructor and number of sections taught

Find the largest total enrollment for a CSSE instructor

Compare each instructor’s total enrollment to the largest; keep only instructors with largest enrollment
Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

Remove unnecessary data

Find the total enrollment for each CSSE instructor and number of sections taught

Find the largest total enrollment for a CSSE instructor

Compare each instructor’s total enrollment to the largest; keep only instructors with largest enrollment
Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

Remove unnecessary data

Find the total enrollment for each CSSE instructor and number of sections taught

Find the largest total enrollment for a CSSE instructor

Compare each instructor’s total enrollment to the largest; keep only instructors with largest enrollment
Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

**Keep only CSSE instructors**

- **$match**
  - \{\text{department: "CSSE"}\}

**Remove unnecessary data**

- **$project**
  - \{\_id: 0, name: 1, enrollments: 1 \}

**Find the total enrollment for each CSSE instructor and number of sections taught**

- **$project**
  - \{name: 1, enrollments: \{\$sum: "$enrollments"\}\}

**Find the largest total enrollment for a CSSE instructor**

- **$group**
  - \{\_id: "1", data: \{\$push: \{name: "$name", enr: "$enrollments"\}\}, mEnr: \{$max: "$enrollments"\}\}, \{\$unwind: "$data"\}

**Compare each instructor’s total enrollment to the largest; keep only instructors with largest enrollment**

- **$match**
  - maxEnrollment: "$data.enrollments"
Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

Remove unnecessary data

Find the total enrollment for each CSSE instructor and number of sections taught

Compare each instructor's total enrollment to the largest; keep only instructors with largest enrollment

Keep only CSSE instructors

Remove unnecessary data

{ $match: { department: "CSSE" } }

{ $project: { _id: 0, name: 1, enrollments: 1 } }

{ $project: { name: 1, enrollments: { $sum: "$enrollments" } } }

{ $project: { diff: { $subtract: [ "$maxEnrollment", "$data.enrollments" ] } } }

{ $match: { maxEnrollment: "$data.enrollments" } }
Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

Remove unnecessary data

Find the total enrollment for each CSSE instructor and number of sections taught

Compare each instructor's total enrollment to the largest; keep only instructors with largest enrollment

Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

Remove unnecessary data

Find the total enrollment for each CSSE instructor and number of sections taught

Compare each instructor's total enrollment to the largest; keep only instructors with largest enrollment

```json
{$match: {department: "CSSE"}}

{$project: {_id: 0, name: 1, enrollments: 1}}

{$project: {name: 1, enrollments: {$sum: "$enrollments"}}}

{$group: {_id: 1, data: {$push: {name: "$name", enr: "$enrollments"}}, mEnr: {$max: "$enrollments"}}}

{$unwind: "$data"}

{$project: {
  diff: {$subtract: ["$maxEnrollment", "$data.enrollments"],
  _id: 0,
  name: "$data.name",
  enrollments: "$data.enrollments"
}}

{$match: {
  maxEnrollment: "$data.enrollments"
}}
```
Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

Remove unnecessary data

Find the total enrollment for each CSSE instructor and number of sections taught

Compare each instructor's total enrollment to the largest; keep only instructors with largest enrollment

{$match: {department: "CSSE"}}

{$project: {
  _id: 0,
  name: 1,
  enrollments: 1
}}

{$project: {
  name: 1,
  enrollments: {$sum: "$enrollments"}
}}

{$group: {
  _id: 1,
  data: {$push: {name: "$name", enr: "$enrollments"}},
  mEnr: {$max: "$enrollments"}
}}

{$unwind: "$data"}

{$project: {
  diff: {$subtract: ["$maxEnrollment", "$data.enrollments"]},
  _id: 0,
  name: "$data.name",
  enrollments: "$data.enrollments"
}}

{$match: { diff: 0 } }
Q1: Find all CSSE faculty with highest total enrollments, report name, number of sections taught, total enrollment

Keep only CSSE instructors

Remove unnecessary data

Find the total enrollment for each CSSE instructor and number of sections taught

Compare each instructor's total enrollment to the largest; keep only instructors with largest enrollment

$match: {department: "CSSE"}

$project: {
  _id: 0,
  name: 1,
  enrollments: 1
}

$project: {
  name: 1,
  enrollments: {
    $sum: "$enrollments"
  }
}

$match: {
  _id: 0,
  name: "$data.name",
  enrollments: "$data.enrollments"
}

$match: {
  diff: 0
}

$project: {
  diff: 0
}

$group: {
  _id: "1",
  data: {
    $push: {
      name: "$name",
      enr: "$enrollments"
    }
  },
  mEnr: {
    $max: "$enrollments"
  }
}

$unwind: "$data"

$project: {
  name: "$data.name",
  enrollments: "$data.enrollments"
}
db.spring.aggregate(
    {
        $match: {department: "CSSE"},
        $project: { _id: 0, name: 1, enrollments: 1} //cleaning
    },
    {
        $project: { name: 1, enrollments: { $sum: "$enrollments"} } //transformation
    },
    {
        $group: { _id: "1",
            maxEnrollment: { $max: "$enrollments"},
            data: { $push: { name: "$name", enrollments: "$enrollments"} }
        }
    },
    {
        $unwind: "$data"
    },
    {
        $project: { _id: 0,
            diff: { $subtract: [ "maxEnrollment", "$data.enrollments"] },
            name: "$data.name",
            enrollments: "$data.enrollments"
        }
    },
    {
        $match: { diff: 0 }
    },
    {
        $project: { diff: 0 }
    }
)