CSC 369: Distributed Computing

Alex Dekhtyar

April 15

Day 5: The Algebra Of Data Transformations
Housekeeping

- **LAST DAY TO DROP THE CLASS**

- 28 students enrolled, no more waitlist

⇒ **Slack**: Can I ask every person to send me a private message? Tell me:
  - How the quarter has been so far.
  - What is harder than typically?
  - What is easier than typically?
  - What do you miss the most?
  - 0.5% of the final grade in the class (comes out of "homework" allotment).
Housekeeping

Data Science Fellowship

I will send the flyer around
The most important conversation in the course
Motivating Example

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

{name: "Julie",
 sections: 3,
 totalEnrollment: 112
}
{name: "Kurt V.",
 sections: 4,
 totalEnrollment: 112
}
What shall we do now?
Motivating Example

Find the total enrollment for each CSSE instructor

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

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

{name: "Julie",
sections: 3,
totalEnrollment: 112}

{name: "Kurt V.",
sections: 4,
totalEnrollment: 112}
Motivating Example

Keep only CSSE instructors

Remove unnecessary data

Find the total enrollment for each CSSE instructor

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

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

{name: "Julie",
sections: 3,
totalEnrollment: 112}

{name: "Kurt V.",
sections: 4,
totalEnrollment: 112}
Motivating Example

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

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

{name: "Julie",
  sections: 3,
  totalEnrollment: 112
}

{name: "Kurt V.",
  sections: 4,
  totalEnrollment: 112
}
What Did We Just Do???

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

Problem Decomposition!!!
What Did We Just Do???

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
What Did We Just Do???

**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**
What Did We Just Do???

**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
What Did We Just Do???

- 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

Example:

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

```
{name: "Aaron",
  teaches: ["CSC 369", "DATA 452"],
  department: "CSSE",
  enrollments: [28, 20],
  position: "professor",
  office: {building: 14, room: 210}
}
```
What Did We Just Do???

- 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 Examples:

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

```
{name: "Aaron",
teaches: ["CSC 369", "DATA 452"],
department: "CSSE",
enrollments: [28, 20],
position: "professor",
office: {building: 14, room: 210}}
```
What Did We Just Do???

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
What Did We Just Do???

- 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
What Did We Just Do???

- 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

```
{name: "Julie",
enrollments: [35, 35, 42]}
```

```
{name: "Aaron",
enrollments: [32, 31]}
```

```
{name: "Alex",
enrollments: [28, 20]}
```
What Did We Just Do???

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

- Julie: Enrollment: 112, sections: 3
- Aaron: Enrollment: 63, sections: 2
- Alex: Enrollment: 48, sections: 2
What Did We Just Do???

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

- **Julie**: enrollment 112, sections: 3
- **Aaron**: enrollment 63, sections: 2
- **Alex**: enrollment 48, sections: 2
What Did We Just Do???

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

{name: “Julie”,
enrollment: 112,
sections: 3,
maxEnrollment: 112}

{name: “Aaron”,
enrollment: 63,
sections: 2,
maxEnrollment: 112}

{name: “Alex”,
enrollment: 48,
sections: 2,
maxEnrollment: 112}
What Did We Just Do???

- 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

```
{name: "Julie",
enrollment: 112,
sections: 3,
maxEnrollment: 112}
```

```
{name: "Aaron",
enrollment: 63,
sections: 2,
maxEnrollment: 112}
```

```
{name: "Alex",
enrollment: 48,
sections: 2,
maxEnrollment: 112}
```
What Did We Just Do???

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

{name:”Julie”,
enrollment: 112,
sections: 3,
maxEnrollment: 112}

{name:”Aaron”,
enrollment: 63,
sections: 2,
maxEnrollment: 112}

{name:”Alex”,
enrollment: 48,
sections: 2,
maxEnrollment: 112}
What Did We Just Do???

- 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
{name: "Julie", enrollment: 112, sections: 3, maxEnrollment: 112}
```
What Did We Just Do???

- 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

{name: "Julie",
  enrollment: 112,
  sections: 3
}
Motivating Example #2

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

```json

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

```
Motivating Example #2

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

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

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

{ course: "DATA 452",
  instructors: [{name: "alex", dept: "CSSE"},
                {name: "hunter", dept: "STAT"}]
}
```
What did we just do?

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?
What did we just do?

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?
What did we just do?

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?
What did we just do?

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?
What did we just do?

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?
What did we just do?

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?
What did we just do?

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?
What did we just do?

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?
What did we just do?

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?
What did we just do?

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?
What did we just do?

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?
What did we just do?

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?
What did we just do?

Problem Decomposition!!!!

into atomic operations
What “Atomic Operations”

Problem Decomposition!!!!

into atomic operations

Relational Algebra (hello, CSC 365)
What “Atomic Operations”

Problem Decomposition!!!!

into atomic operations

Relational Algebra
(hello, CSC 365)
What “Atomic Operations”

- Problem Decomposition!!!!
- Algebra of atomic Data operations
- into atomic operations
What “Atomic Operations”

Relational Algebra

Selection
Projection
Set Operations
Join
Grouping/Aggregation
Sort
What “Atomic Operations”

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

Generalized Algebra
- Filtering
- Projection/Transformation
- Join
- Grouping/Aggregation
- Sort
Why Do We Discuss these Operations?

db.collection.find(....).<finishingtouch>()

Selection, Projection, Sort, Skip, Limit

db.collection.aggregate(....)
What “Atomic Operations”

Generalized Algebra

Filtering
Projection/Transformation

Join

Grouping/Aggregation

Sort

Unwind

Limit

Skip
Overview: Selection/Filtering

Given a selection criterion keep objects that match it, Remove objects that don’t.
Overview: Selection/Filtering

Given a selection criterion keep objects that match it, Remove objects that don’t.

Keep only CSSE instructors
Overview: Projection/Transformation

Given an object, transform it into a different object
Overview: Projection/Transformation

Given an object, transform it into a different object

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

Given an object, transform it into a different object

```
{name: "Alex",
 enrollments: [28, 20],
}
```

Remove unnecessary data
Overview: Aggregation

Given an object with arrays, aggregate their content.

Add up enrollments

```
{name: "Alex",
enrollments: [28, 20],
}
```
Overview: Aggregation

Given an object with arrays, aggregate their content.

Add up enrollments

```javascript
{name:”Alex”,
  enrollments:[28,20],
}
```

```javascript
{name:”Alex”,
  enrollments:48,
}
```
Overview: Grouping

Combine information from multiple objects into one, based on common attributes