MAS DSE 260: Capstone Project

İlkay ALTINTAŞ, Ph.D.

Lecture 8: Deploying a Robust and Scalable Solution

Process Roadmap (260 B)

- ✓ Step 6: Modeling and Visualization
  - ✓ REPORT 6: due 4/12
  - ✓ PRESENTATION 3: 4/13
- ✓ Step 7: Evaluating and Interpreting Modeling Results
  - ✓ REPORT 7: due 4/26
- Step 8: Deploying a Robust and Scalable Solution
  - REPORT 8: due 5/10
  - PRESENTATION 4: 5/11
- Step 9: Developing a Communication Plan and Monitoring Dashboard
  - REPORT 9: due 5/24
- Step 10: Optimization
  - FINAL REPORT AND POSTER: due 6/8
  - FINAL PRESENTATION AND DEMO: due 6/9

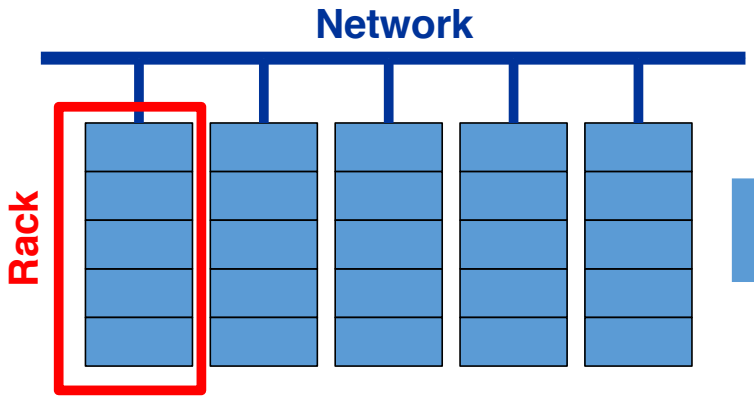
What is needed in this step?

- Explain the requirements of scalability in your problem and why you should care about them
- Update your solution architecture and data pipelines to ensure scalability
- Define metrics for robustness
- Create an evaluation plan for scalability and robustness
- Report on evaluation

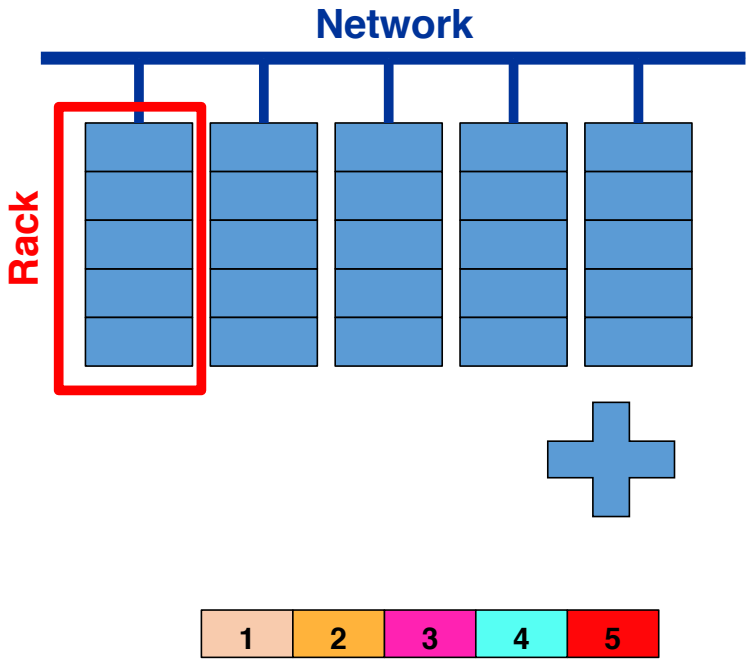
Scalable Data Science

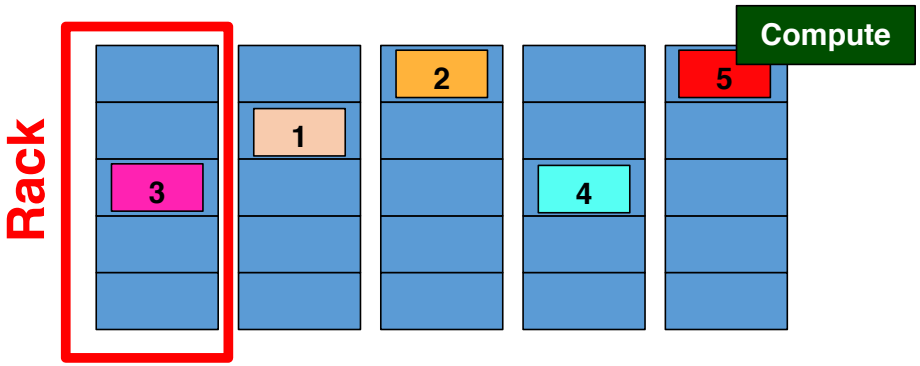
Statistics + Systems + Machine Learning

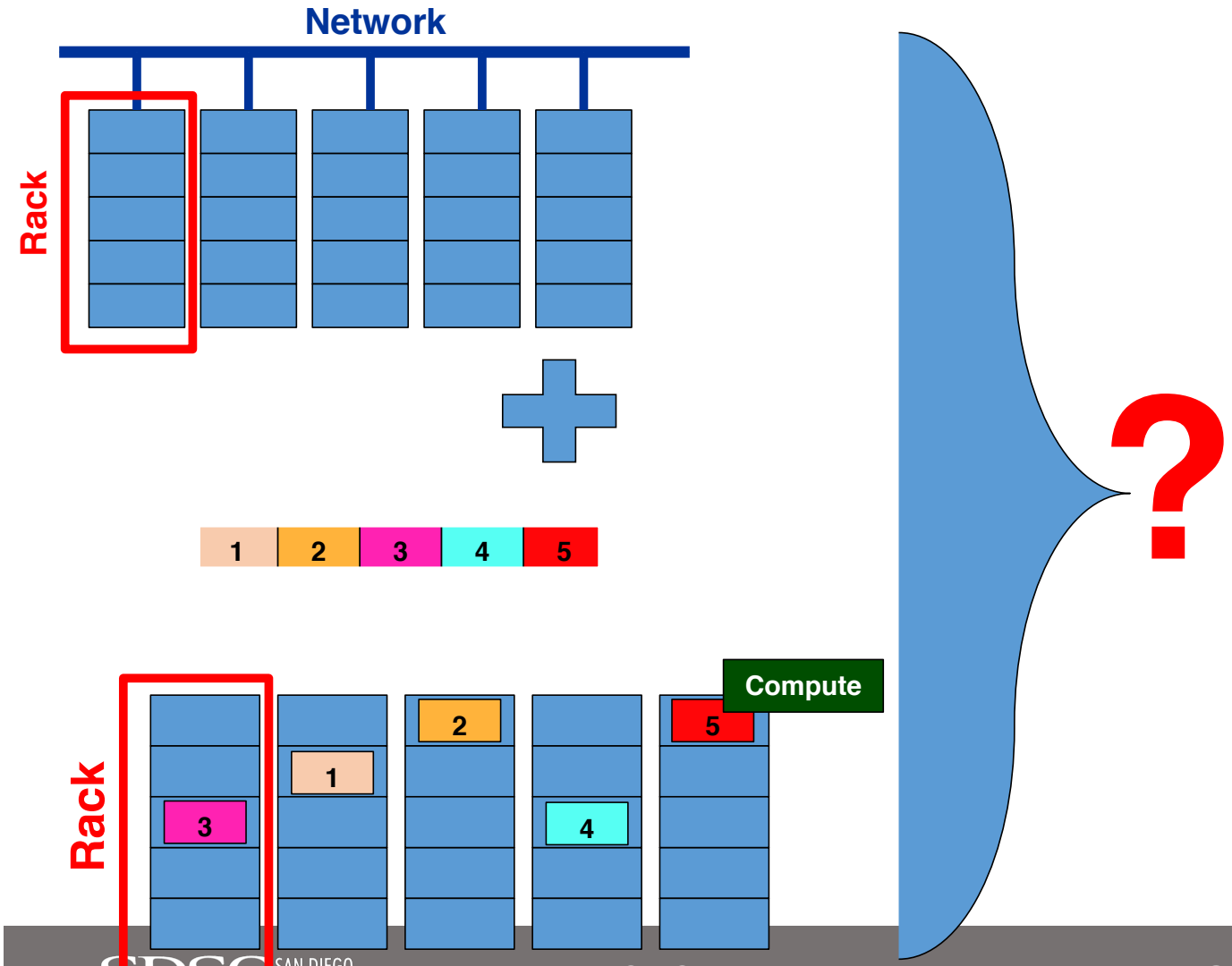
➔ techniques for analyzing large amounts of data at scale...



Data-parallel scalability

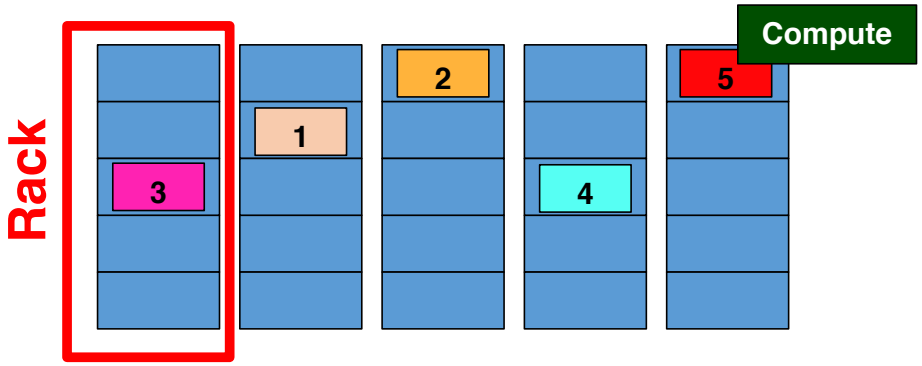

Network
Rack
1
2
3
4
5


Compute
Rack
1
2
3
4
5

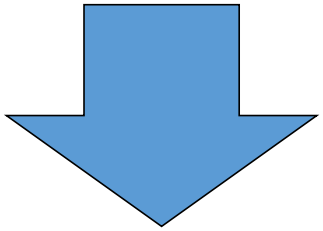

Network
Rack
1
2
3
4
5
?
Compute
Rack
3
1
2
4
5

Programming Model = abstractions

Runtime Libraries + Programming Languages



Programming Model for Big Data



Programmability on top of Distributed File Systems

Requirements for Big Data Programming Models

1. Support Big Data Operations

Split volumes of data

1. Support Big Data Operations

Split volumes of data

Access data fast

1. Support Big Data Operations

Split volumes of data

Access data fast

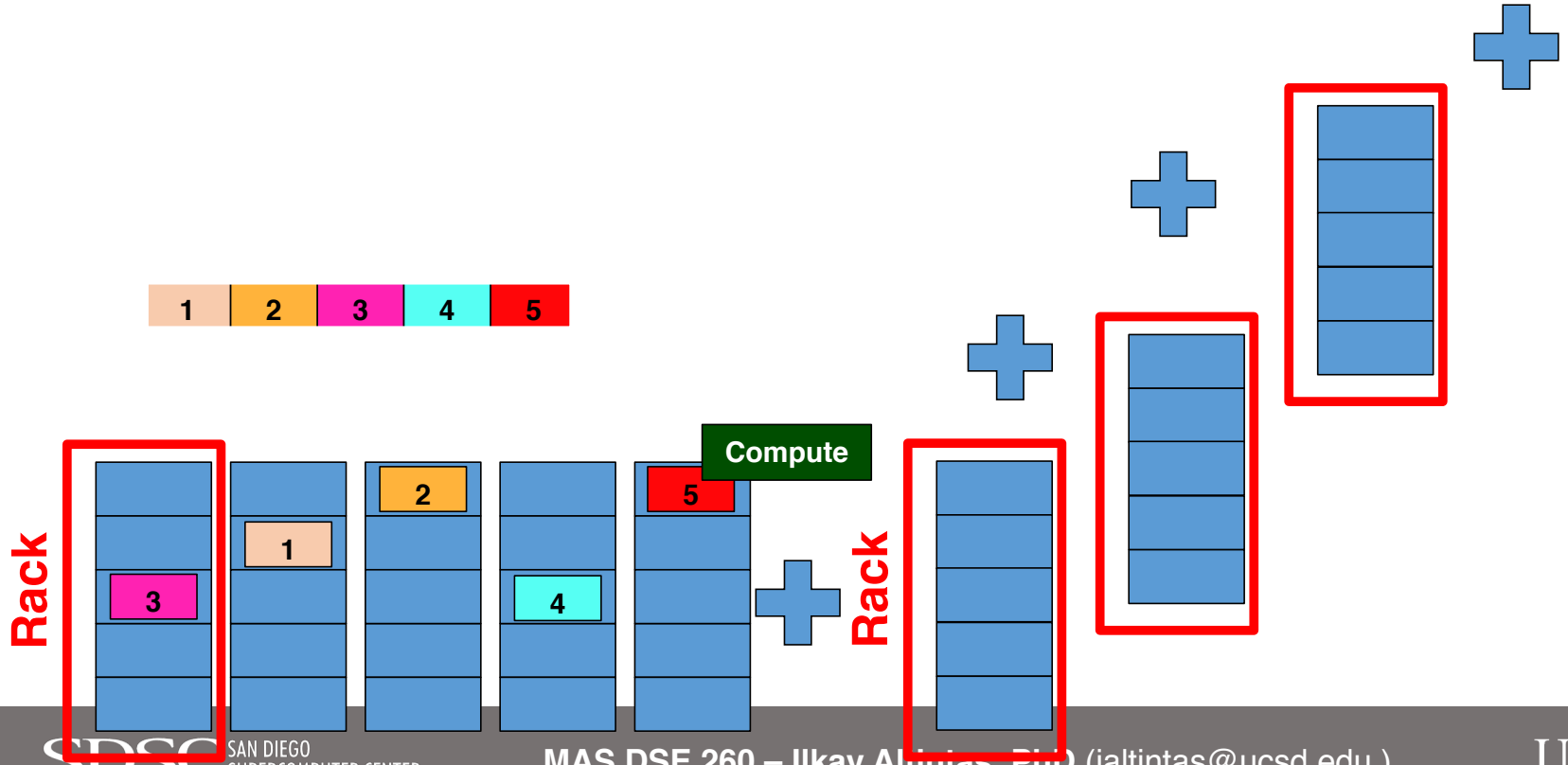
Distribute computations to nodes

2. Handle Fault Tolerance

Replicate data partitions

Recover files when needed

3. Enable Adding More Racks



4. Optimized for specific data types

Document

Table

Graph

Key-value

Multimedia

Stream

Also consider statistical scalability!

Good article: https://arxiv.org/pdf/1309.7804.pdf

Step VIII Report Guidelines

- Title, team members and advisor(s)
- Summary of Scalability and Robustness Requirements
  - Explain what you mean by scalabaility, performance, more datasets, model scalability, …
  - Summarize case studies that includes robustness
  - Illustrate your data pipelines and end-to-end data science process
- Approach
- Evaluation Strategy and Plan
  - How do you define success? What are your metrics?
- Results and Next Steps
- Any major updates to Steps 1 through 7 as a result of Step 8
- Keep it to 4-6 pages
- Due date: 5/10/2018 midnight

Questions?

İlkay Altintas, Ph.D.
Email: ialtintas@ucsd.edu