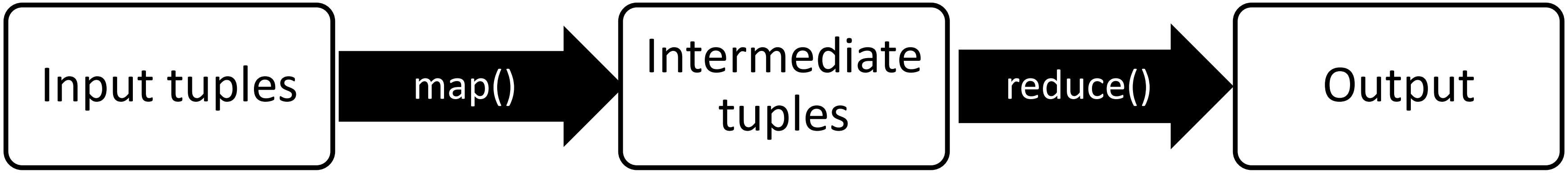
Introduction to MapReduce, using CIEL and Amazon EC2

Data Centric Networking (R202)

MapReduce Basics



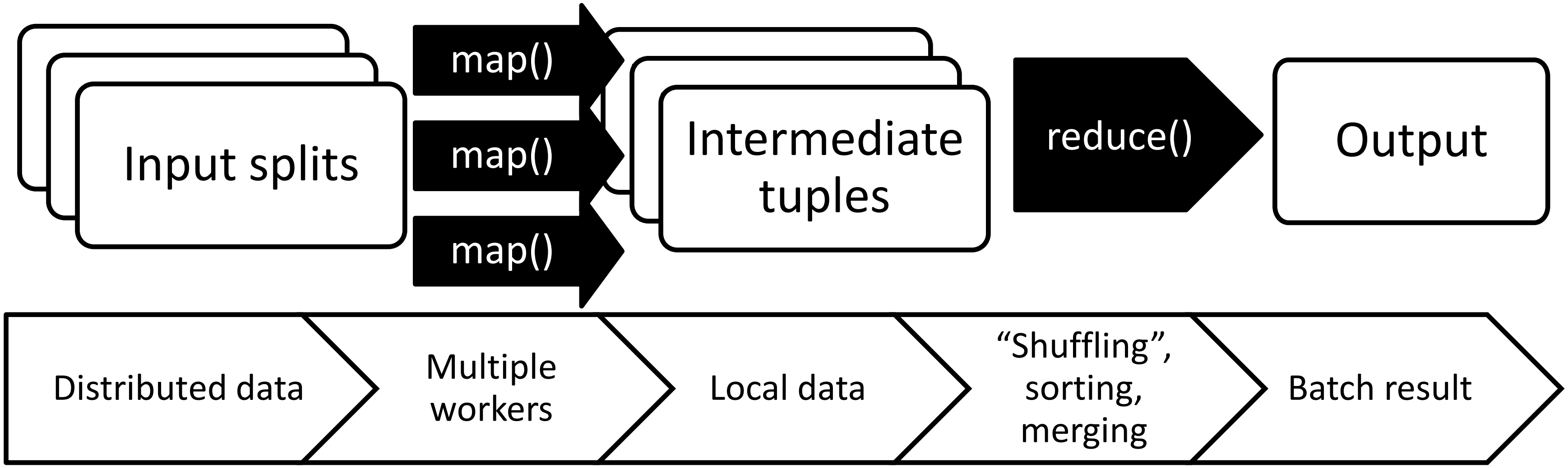
Input tuples

map()

Intermediate tuples

reduce()

Output



Input splits

map()

map()

map()

Intermediate tuples

reduce()

Output

Distributed data

Multiple workers

Local data

“Shuffling”, sorting, merging

Batch result

Task Coordination

- Typical architecture utilises a single master and multiple (unreliable) workers.
- Master holds state of current configuration, detects node failure, and schedules work based on multiple heuristics. Also coordinates resources between multiple jobs.
- Workers perform work! Both mapping and reducing, possibly at the same time.

CIEL: Dynamic Task Graphs

- MapReduce prescribes a “task graph” that can be adapted to many problems.
- Later execution engines such as Dryad allow more flexibility, for example to combine the results of multiple separate computations.
- CIEL takes this a step further by allowing the task graph to be specified at run time – for example:
  - `while (!converged) spawn(tasks);`

Amazon Elastic Compute Cloud

- EC2 = “Infrastructure as a service”
- Key decisions for provisioning instances:
  - Pricing? Reserved, on-demand, spot, geography
  - System? OS, customisations (AMI)
  - Sizing? RAM / CPU based on tiered model
  - Storage? Quantity, type (EBS, instance)
  - Networking / security

Practice makes perfect

- Feel free to ask questions during the session
- Helpful links:
  - http://www.cambridgeplus.net/tutorials/CIEL-DCN/