Towards a Cross-Disciplinary Pedagogy for Big Data

Joshua Eckroth

Math/CS Department
Stetson University

CCSC-Eastern 2015

What is big data?

“Data mining and analysis require ‘big data’ techniques when the data have such **high volume** or **high velocity** that **more than one machine** are required to store and/or process the data.”

"Data mining and analysis require 'big data' techniques when the data have such **high volume** or **high velocity** that **more than one machine** are required to store and/or process the data."

- Data volume: The size of the data when collected and stored.
- Data velocity: The speed at which data is acquired.

Who has big data?

Marketing

Monitor Twitter for keywords about specific companies, products, and services. Apply sentiment analysis to estimate the general sentiment from the population.

Emily!
@Emk4ever

Follow

Had an awesome Hatter Saturday!!! Can not wait to be at my new home! 😍😍😍 #NewHatter #stetson #stetsonuniversity pic.twitter.com/YLuREhiowd

3:58 PM - 18 Apr 2015

1

Ali Ernest
@AliErnest97

Follow

At the #newhatter orientation! I'm super excited to see the beautiful campus #Stetson

9:30 AM - 18 Apr 2015

2 6

Colleen Hamilton
@Colleen_ham1

Follow

You can bring your dog to school at #Stetson. How cool is that? pic.twitter.com/1x4ICYuX1f

9:11 AM - 4 Apr 2015

2

Sonni Abatta
@SonniAbatta

Follow

BREAKING: Bomb threat on Deland campus of #Stetson University. McMahan, Elizabeth, Presser Halls evacuated. Volusia Co investigating. #FOX35

10:20 PM - 24 Mar 2015

6 1

Healthcare

Combine numerous sources of patient data in order to predict readmissions. Possibly also combine HealthData.gov data, e.g., records of hospital-acquired infections, and records of readmissions across 17,000 hospitals.

Criminology

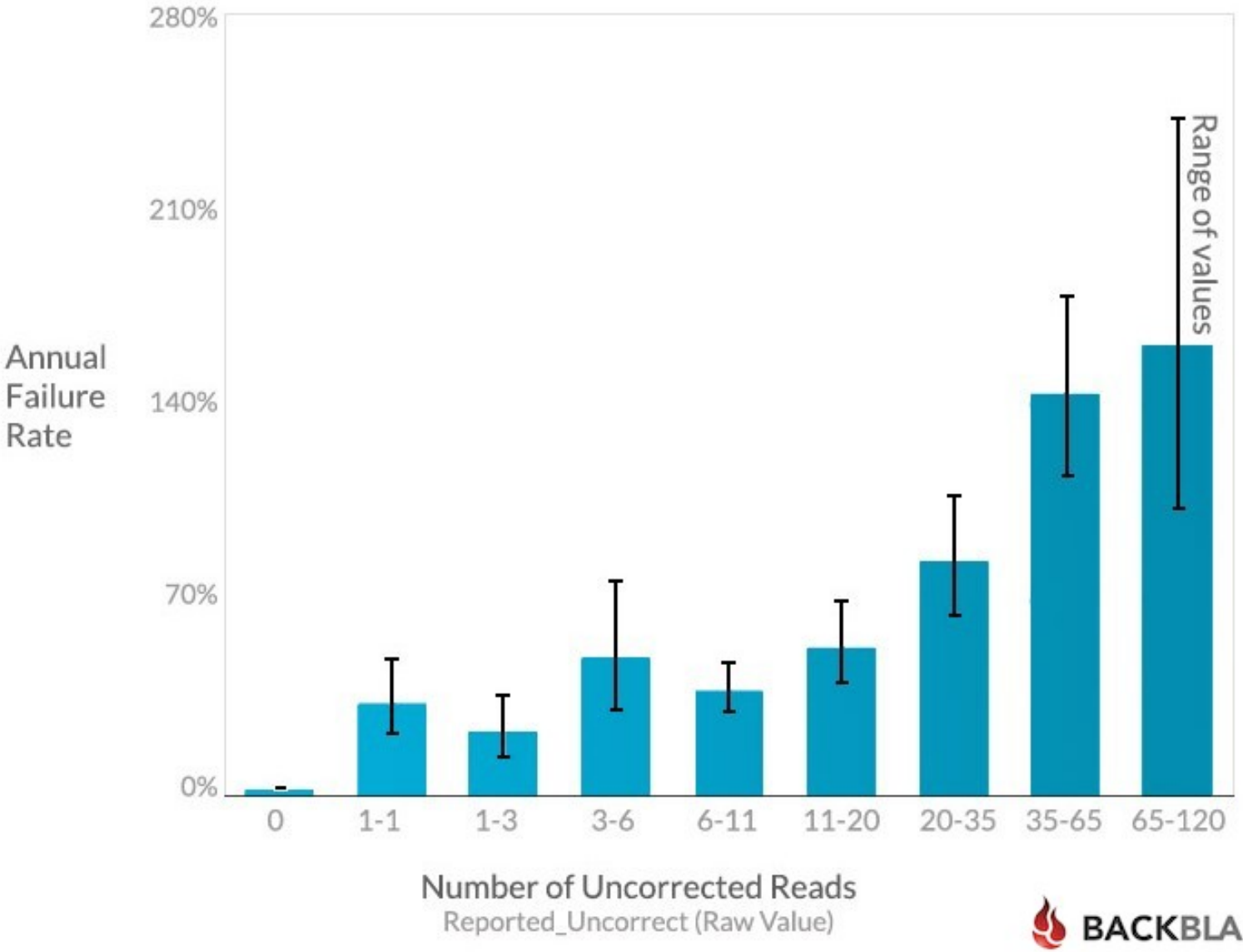
Gather abundant statistics about crimes, demographics, transportation, and pedestrian flow from data.gov, census.gov, and other sources to test theories about crime locality and flow.

Information Technology

Extract specific metrics from hard disk self-monitoring data streams in order to find significant indicators of imminent hard disk failure.

SMART 187

Correlated to Annual Failure Rate. As the number of read errors increase, it is more likely the drive will experience a failure.



BACKBLAZE

Goal

- Develop curricula for both CS and non-CS students
  - with applications from several domains
  - that teaches how to **design** a big data processing task
  - using common language and paradigms
  - so that computer scientists may be able to build and execute the job

Learning Outcomes

- LO1: Determine whether a data processing task requires "big data" tools and techniques
- LO2: Identify appropriate big data paradigms to solve a problem
- LO3: Design an abstract representation of the solution using computation graphs

Simple example

Find the maximum value of two metrics in a data set with one trillion records, stored across several machines.

Simple example

Find the maximum value of two metrics in a data set with one trillion records, stored across several machines.

BAD IDEAS:

- Use Excel
- First, create a subset of the data keeping only the first metric. Then sort the list of values. The maximum will be the last value. Then repeat with the second metric.
- Find the maximum value of both metrics on data stored on machine 1, then on machine 2, …, then find the maximum of the maximums after processing on each machine sequentially.

Simple example

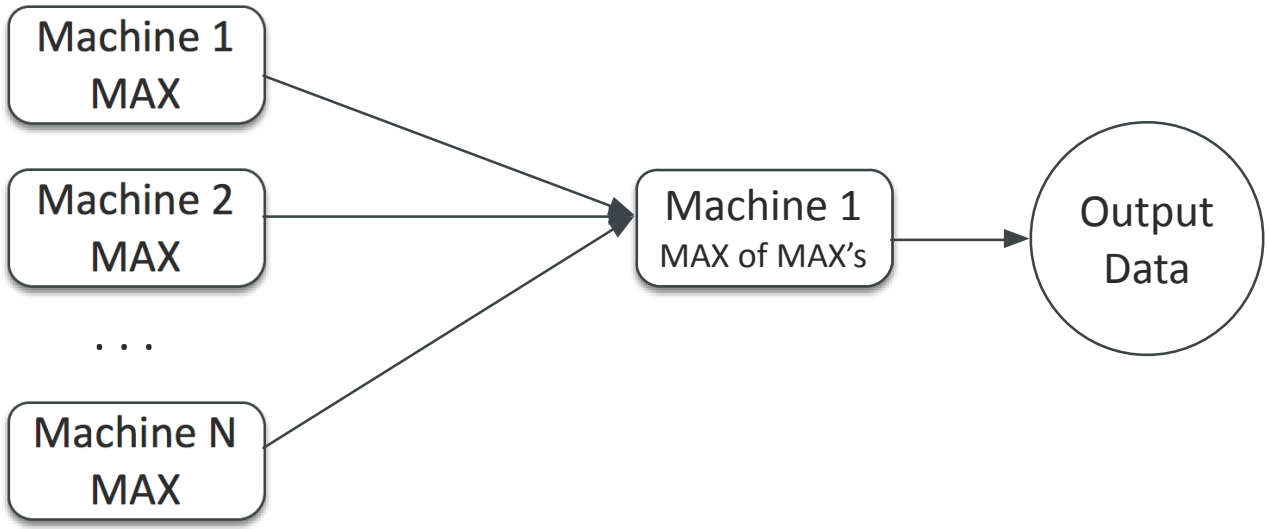
Find the maximum value of two metrics in a data set with one trillion records, stored across several machines.

GOOD IDEA:

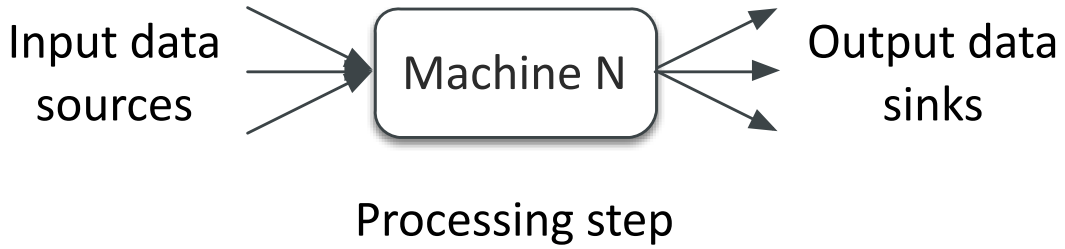
- Split the data into chunks, transfer each chunk to one machine
  - For each chunk:
    - Find the maximum value of both metrics **just for the data stored on that machine**; do these computations in parallel
  - Then combine the results:
    - Find the maximum of the maximums (using whatever machines are available)

Simple example

Find the maximum value of two metrics in a data set with one trillion records, stored across several machines.



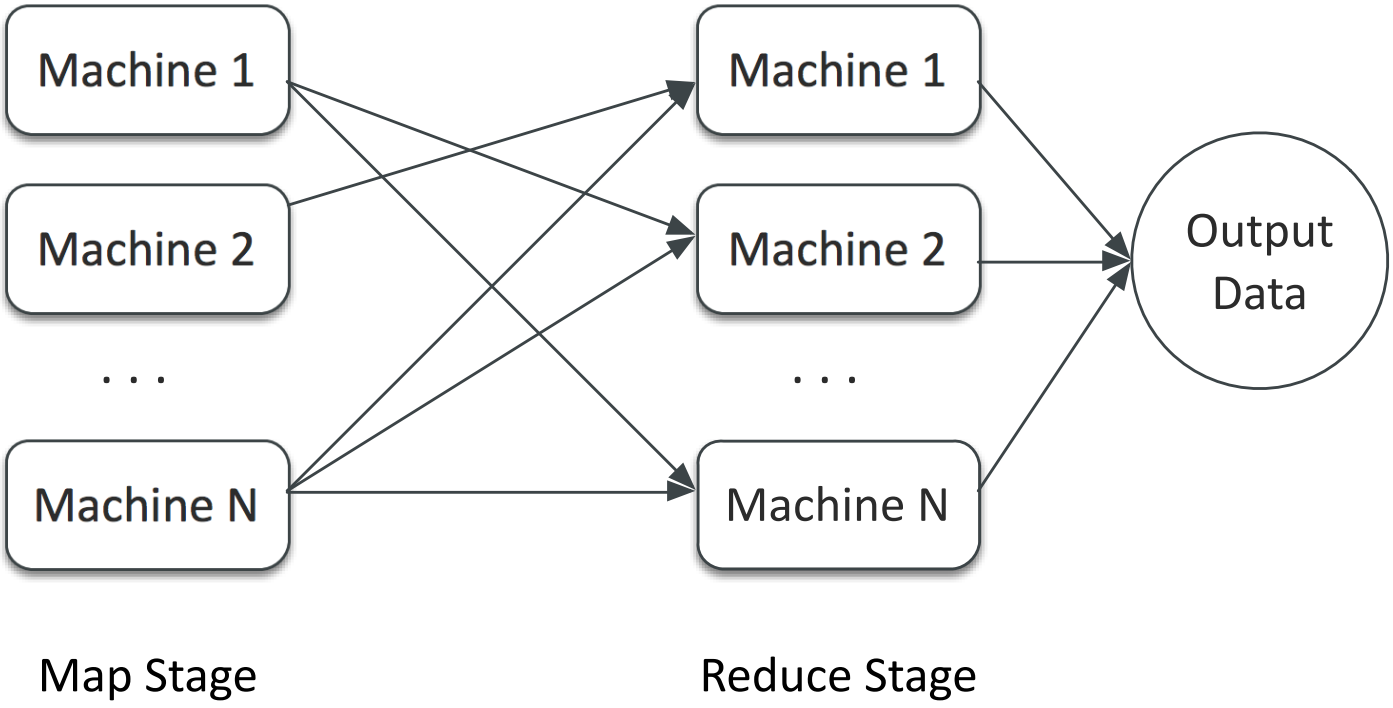
Computation graphs



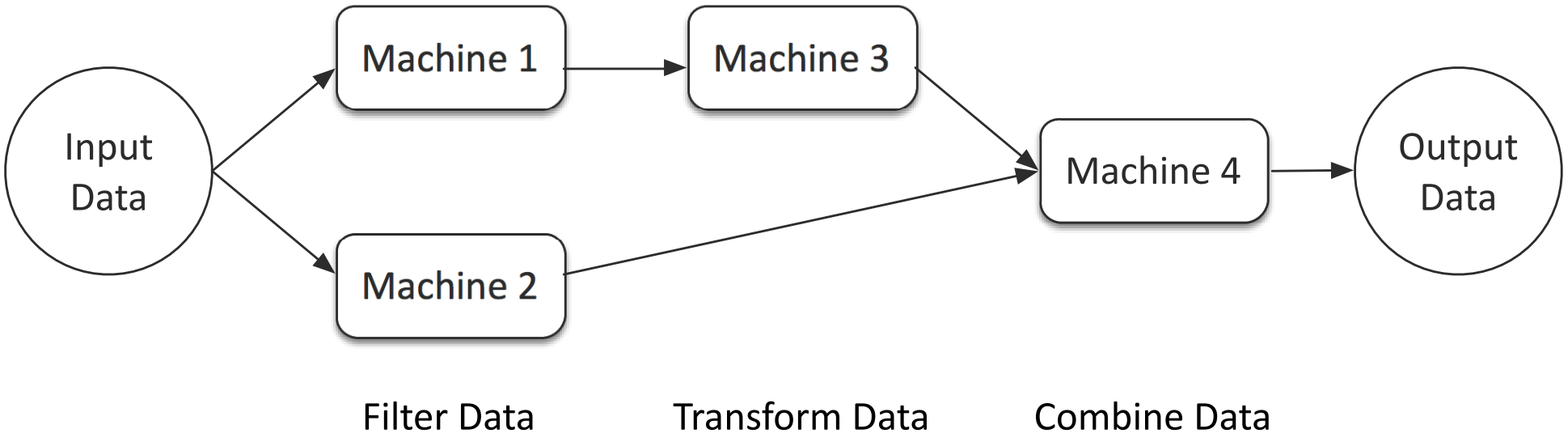
- Each processing step (machine) can only process the data the machine has stored locally or are made available from input arrows
- Assume each machine already holds 0-1TB of data (redefine the upper limit as necessary)
- Input and output data (arrows) are able to support relatively low data velocity

Big data paradigms

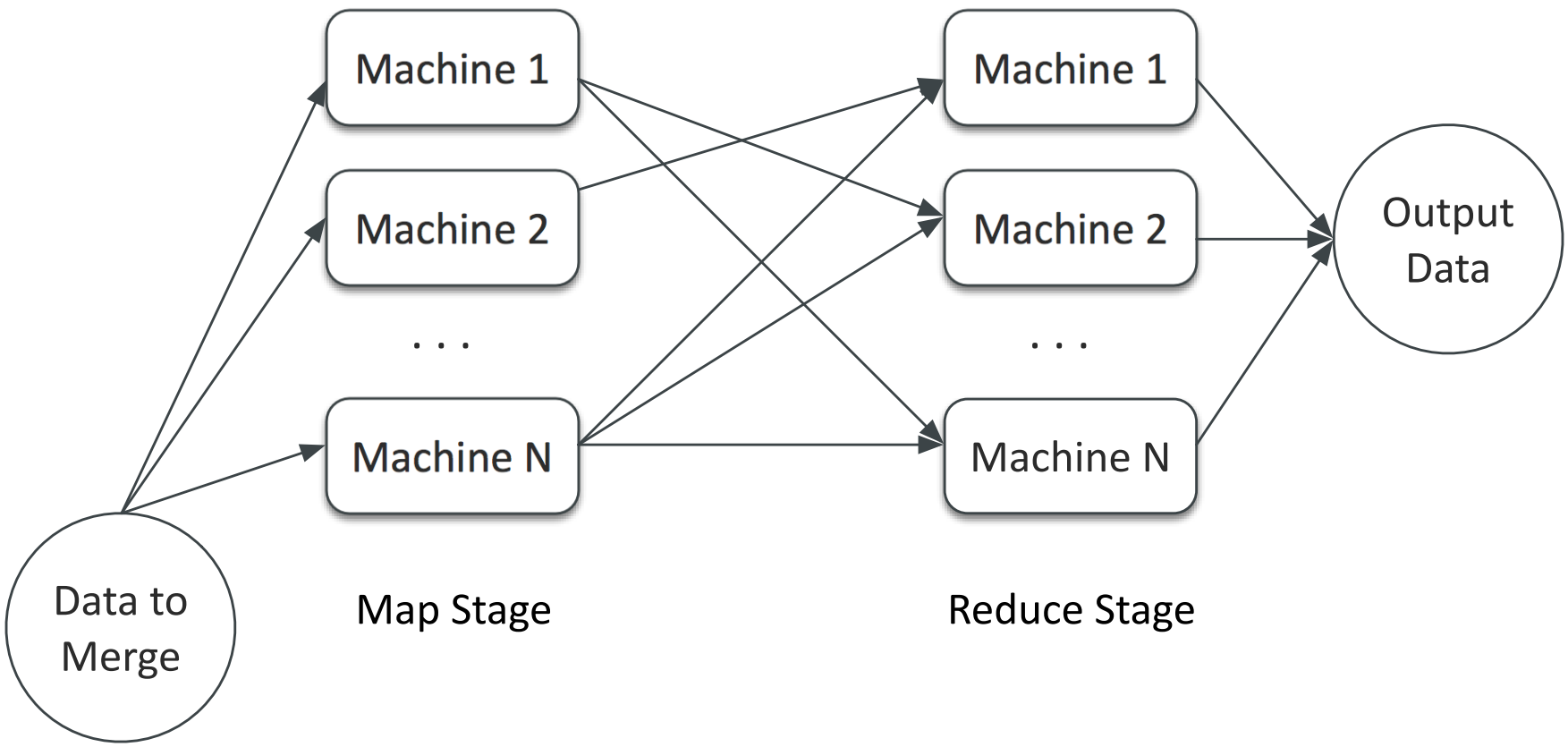
Batch processing



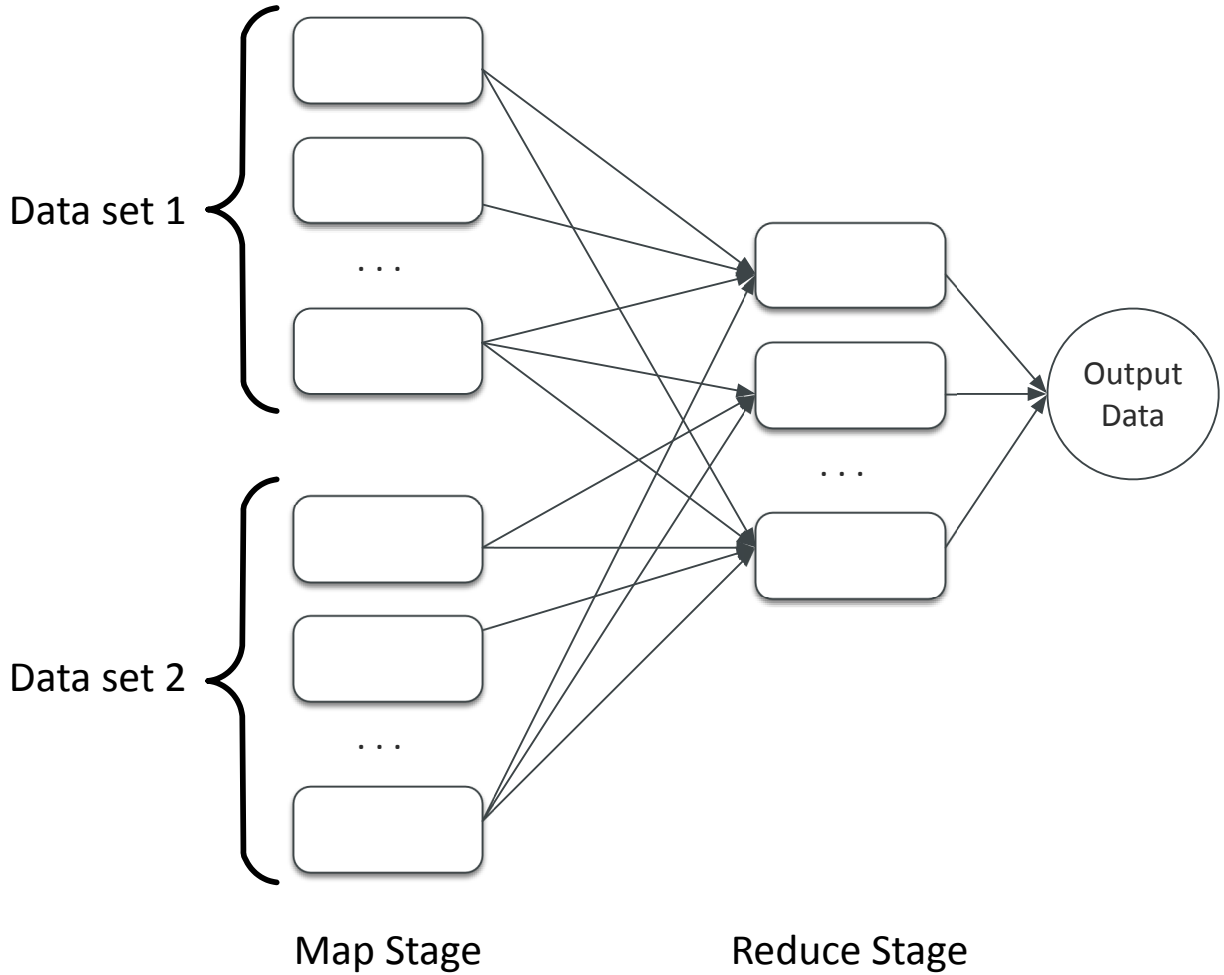
Real-time processing



Map-stage Merge



Reduce-stage Merge



More complex example

A large grocery chain has been monitoring purchases of various nutrition bars in their stores. They also have click data for a series of web advertisements for these same nutrition bars.

The data analysts want to find out if web advertisements have an impact on actual purchases.

Both datasets are “big.”

More complex example

It should be clear that we must filter through purchase data to find purchases of nutrition bars.

Then, these purchase data must be combined with web click data.

These combined data must then be summarized, e.g., by computing the statistical correlation, for each nutrition bar, between number of purchases and number of web clicks between the start and end times of the advertisement campaign.

More complex example

From purchase data, we need this kind of record for each purchase:

("purchase", time, nutrition bar type)

From click data, we need this kind of record for each click:

("click", time, nutrition bar type)

More complex example

From purchase data, we need this kind of record for each purchase:

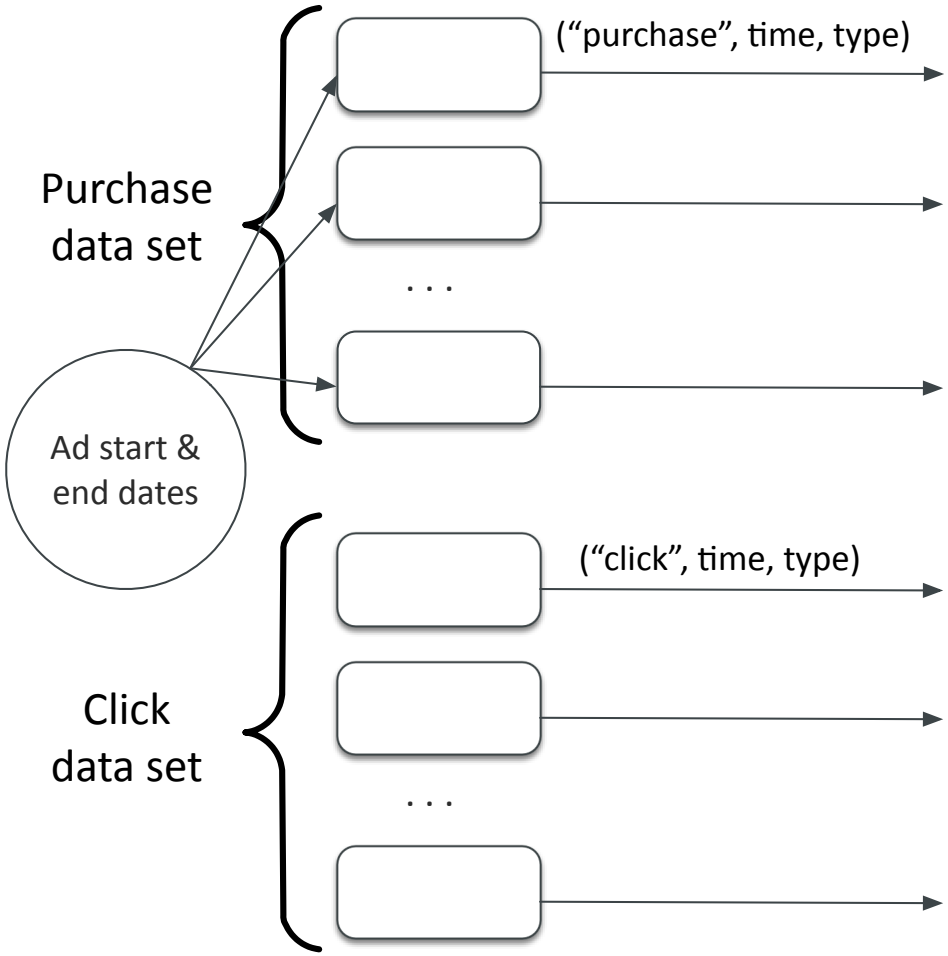
("purchase", time, nutrition bar type)

From click data, we need this kind of record for each click:

("click", time, nutrition bar type)

From a third (small) data set, we can identify advertisement campaign start/end dates. We will only generate records for purchases that occur within the corresponding ad campaign date range.

More complex example



More complex example

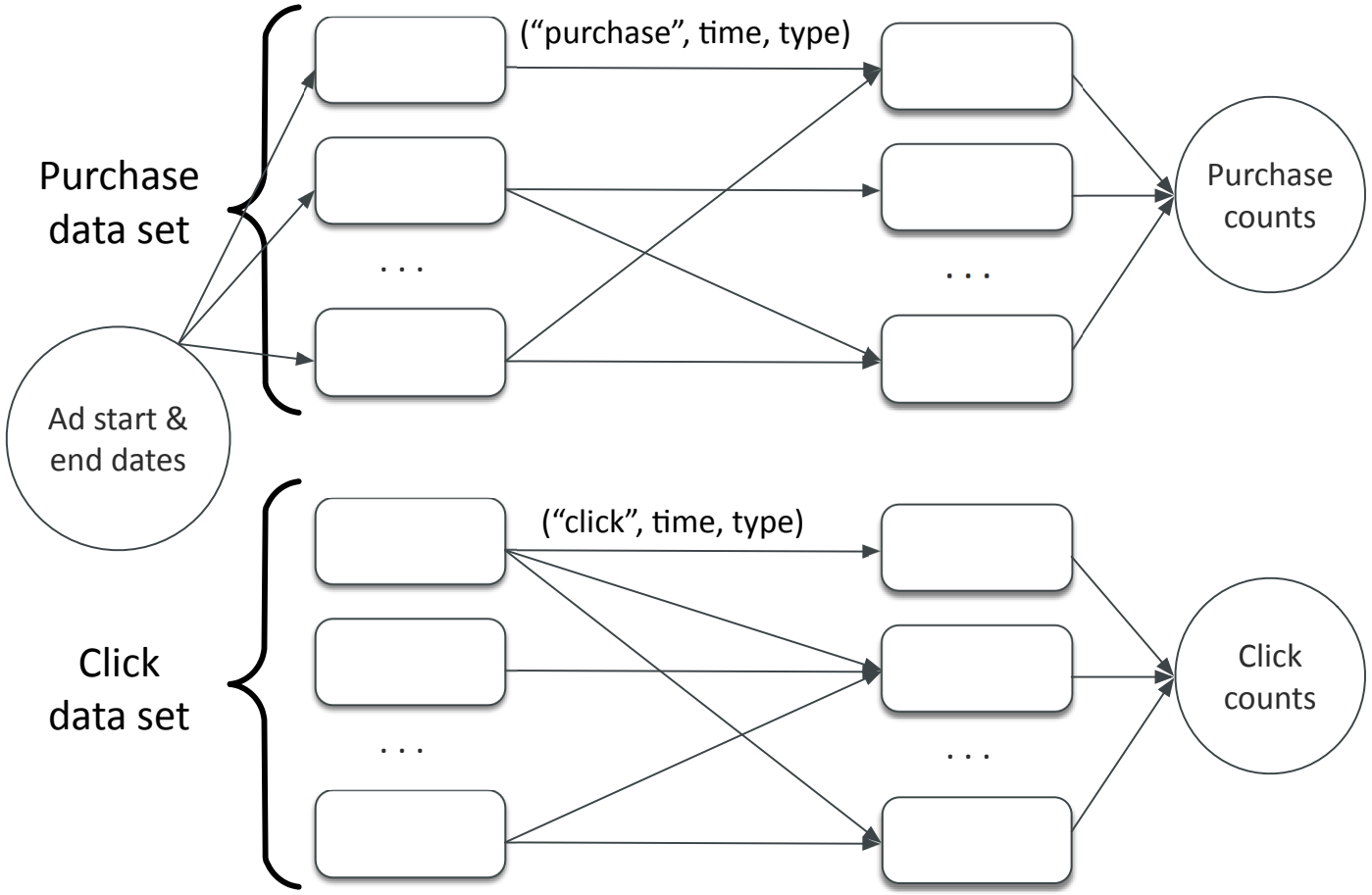
Next, we can aggregate and count purchases and clicks per day (or per week, or …), for each nutrition bar.

The result would be a list of records:

(event, type, day, count)

Where “event” is either “purchase” or “click.”

More complex example



Summary

- We've looked at LO2 and LO3:

  LO2: identify appropriate paradigm

  LO3: design solution with computation graphs

- Next: LO1, do we have a big data problem?

Decision tree

- Can the data be processed in R, SPSS, Stata, etc.?
  - YES: Avoid big data technologies, use traditional methods.
  - NO: …

Decision tree (big data)

- Are the data high volume or high velocity?
  - VOLUME: …
  - VELOCITY: Decompose the processing into transform/filter stages
  - BOTH: Design a real-time processing job that saves the resulting data in splits across multiple machines; then use batch processing

Decision tree (high volume)

- Does more than one data set need to be merged?
  - YES: Is more than one data set “big”?
    - YES: Join in the Reduce stage of Map-Reduce
    - NO: Join in the Map stage of Map-Reduce
  - NO: Design a simple Map-Reduce job

Summary

- We have some tools for achieving learning outcomes LO1-LO3
- We have evidence that analysts from various fields have or will have big data problems and need big data solutions
- Future work: develop learning modules; evaluate with non-computer scientists; formalize computation graphs and integrate into software tools

Experience in the Classroom

- These tools grew out of classroom experience, but did not exist at the time

- Some observations:
  - As projects grew in size (data size), students realized on their own that traditional tools (Excel, R) were not sufficient
  - Students composed multi-stage batch processing jobs but did not discover or implement merges without assistance

Experience in the Classroom

- Some more observations:
  - Existing tools are still very difficult to use
    - We used Hadoop, Hive, Mahout, R
    - Programming experience required, Linux experience very beneficial
  - Actually executing big data jobs was important to the students
    - They tried ideas by running their code
    - I suspect non-CS students will also need some way to test their ideas

Questions?