Supporting Complex Analytics by Non-Expert Users

Magdalena Balazinska

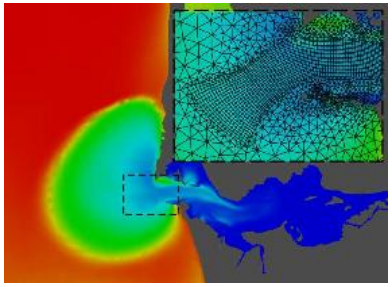
UNIVERSITY OF WASHINGTON

http://www.cs.washington.edu/homes/magda

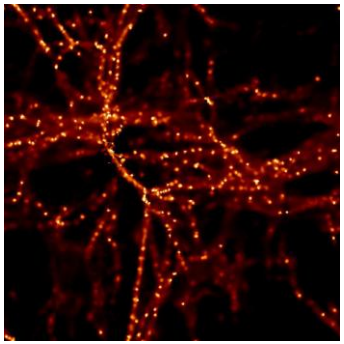
Nuage Project

http://nuage.cs.washington.edu/

Non-Experts Need to Perform Complex Analytics



- Areas
  - Science
  - Business
- Users
  - Experts in their area
  - NOT database experts

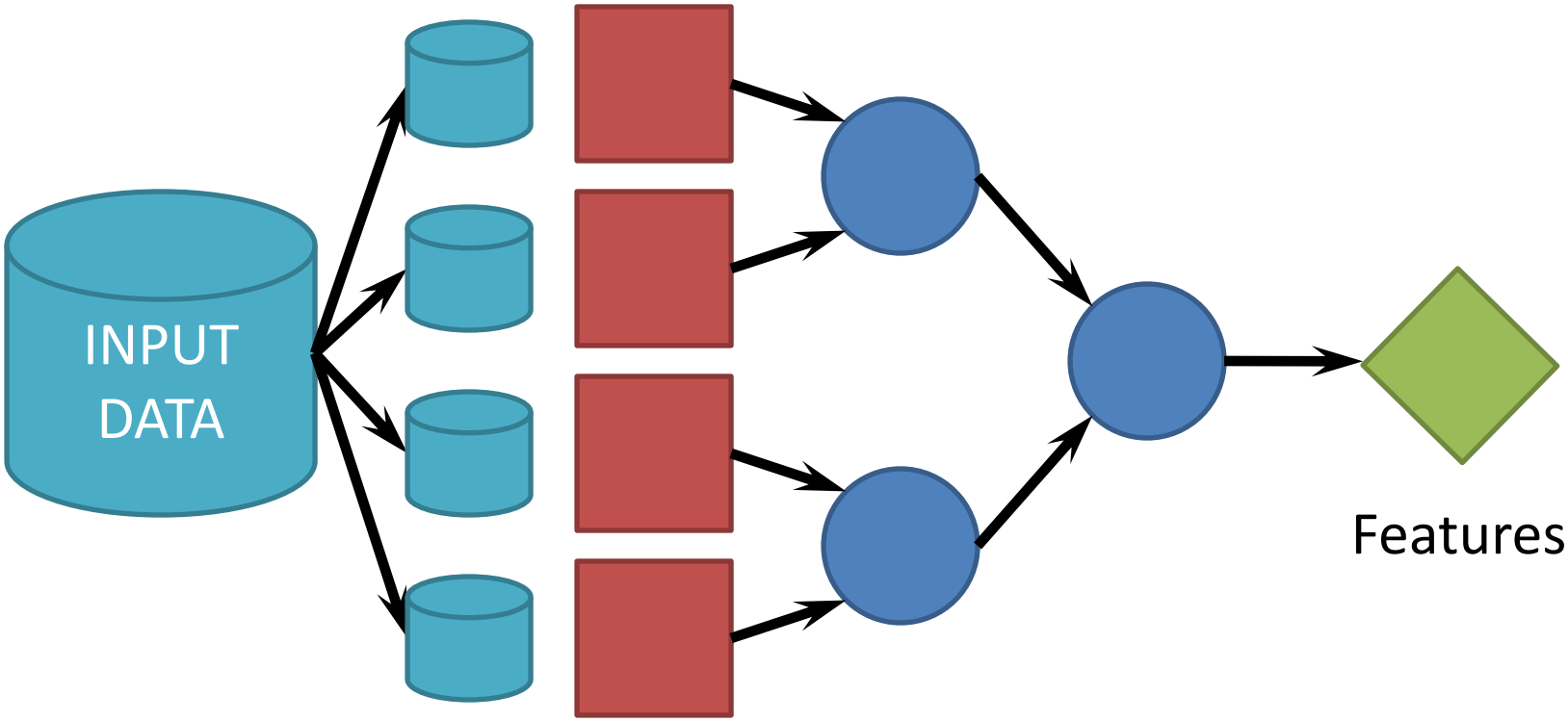


- Analysis
  - Often complex: machine learning, UDFs, etc.
  - Ad-hoc and changing

What is the Challenge?

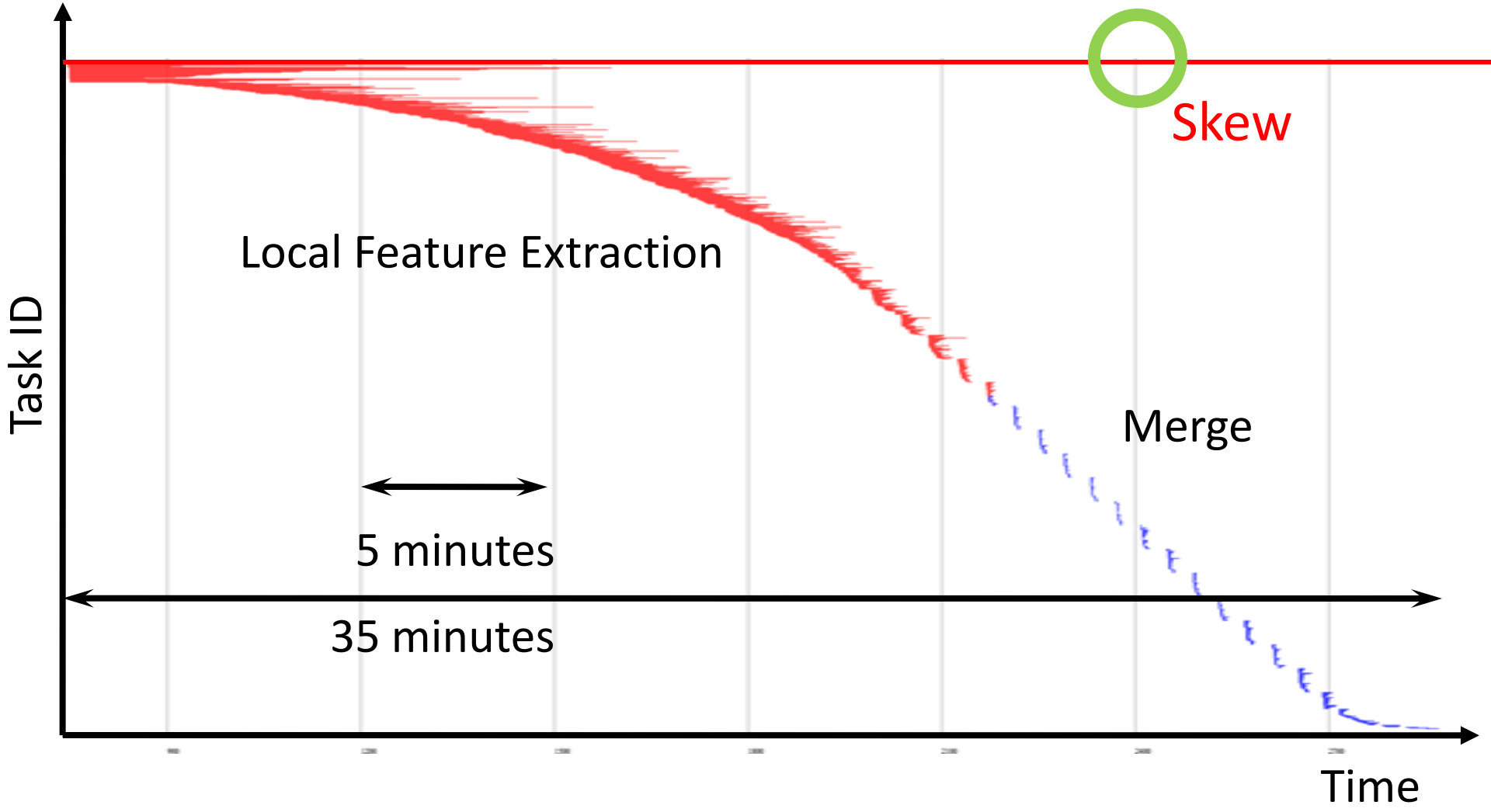
- Not easy to express algorithms using a dataflow-style of processing
- Even more difficult to get high-performance
- And even more difficult for users to understand the performance they are getting
- Not enough administrators to go around

Example: Parallel Feature Extraction



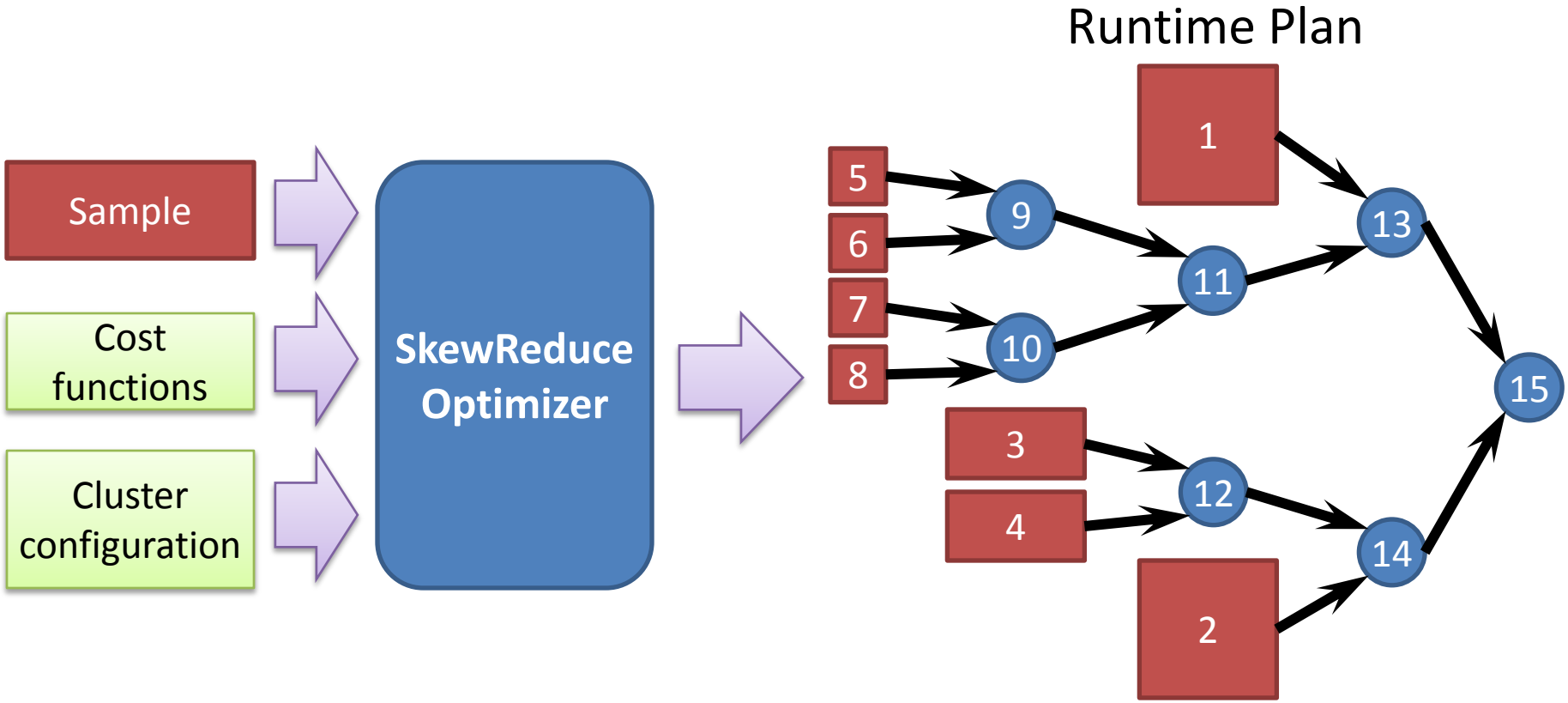
- *Partition* multi-dimensional input data
- *Extract* features from each partition **Map**
- *Merge* (or reconcile) features **Hierarchical Reduce**
- *Finalize* output

Problem: Skew



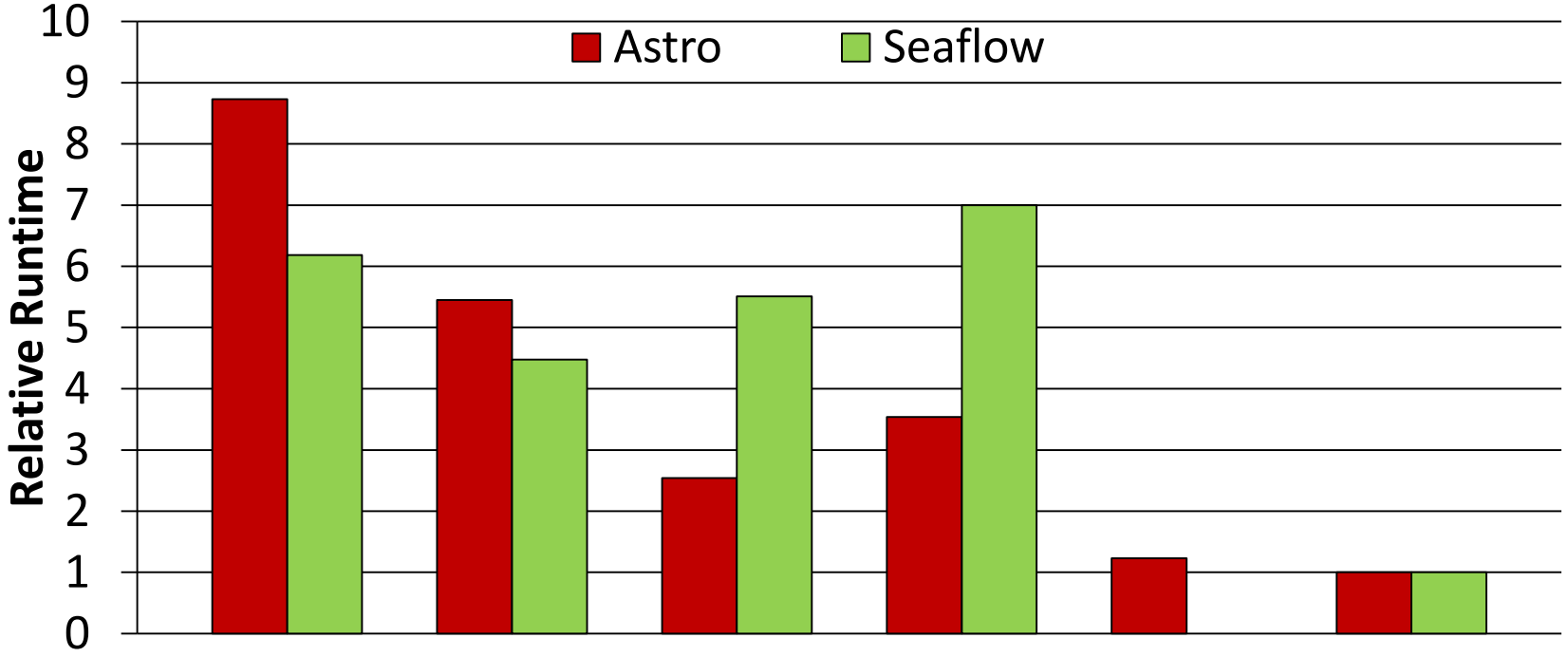
- The top red line runs for 1.5 hours

Our Approach: SkewReduce



- ***Goal**: minimize expected total runtime*
- SkewReduce automatically derives partition plan
- Key idea: leverage user-provided cost functions

Does SkewReduce work?



Coarse	Fine	Finer	Finest	Manual	SkewReduce	
14.1	8.8	4.1	5.7	2.0	1.6	Hours
87.2	63.1	77.7	98.7	-	14.1	Minutes

- Static plan yields 2 ~ 8 times faster running time

More Generally

- Need to help users
  - Expression complex analytics
  - And get high-performance [E.g., SkewReduce]
- Need auto-tuning for high-performance
  - [E.g., FTOpt, fault-tolerance optimizer]
- Need to help users understand performance
  - [E.g., ParaTimer time-oriented progress indicator]
  - [E.g., User-oriented performance explanation]