

Starfish: A Self-tuning System for Big Data Analytics

Analysis in the Big Data Era

Explore Y! Santa Clara [New](#) My Yahoo! | Mobile

YAHOO! SITES [Edit](#)

TODAY - July 28, 2011

The skin condition these celebs share

Kim Kardashian and LeAnn Rimes suffer from psoriasis, which ups the chance of heart attack. [Often misdiagnosed](#)

Learn about psoriasis
Recovering from sunburn
The beatable cancer

NFL QB's pricey wedding Malady these stars share 'All-you-can-eat' diet foods Key test for GOP debt plan

1 - 4 of 56

NEWS WORLD LOCAL FINANCE

- House GOP sets vote on revamped debt-limit bill
- Chris Christie taken to hospital after difficulty breathing
- Ex-wife: Ill. congressman owes \$117K child support
- Indonesian fishermen find body of American surfer
- Bones found in Louisiana bank chimney identified
- Tim Gunn mocks Hillary Clinton's pantsuits
- Brothers arrested for attacking pilot on Miami-to-SFO flight - CBS
- Obama attracts new California bundlers - C/W
- Pet store troubles continue in San Francisco - Curbed
- MLB · NFL · NASCAR · UFC · Golf · NCAA Football

updated 10:37 am More: [News](#) [Popular](#) [Sports](#)

Markets: Dow: 12,338.87 0.29% Nasdaq: 2,791.96 0.98%

Enter stock symbol [Get Quotes](#) [Scottrade](#) [Open An Account](#)

FEATURED PARTNERS

[Netflix](#) [Univ. of Phoenix](#)

INTERACTIVE

get 25% off.

Sign In | New here? **Sign Up** | Page Options ▾

TRENDING NOW

- Gene Simmons
- Kate Bosworth
- Stevie Nicks
- Harrison Ford
- Selma Blair
- Walmart video
- Jalen Rose
- Lee DeVyze
- NFL signings
- Unemployment ben...

[Watch the show >>](#)

Old Navy **The Semi-Annual STOCK UP SALE** AdChoices ▾

IN-STORE & ONLINE [SHOP NOW](#)

Old Navy Stock Up Event - Ad Feedback

VIDEO PICKS

Go to Video

Extreme hair trends involve feathers, chalk

Study: Crying at work doesn't hinder success

Controversy over facial recognition device for police

Lockout is off. Trash talking is on. [Sports](#)

Massive Data

Data Analysis

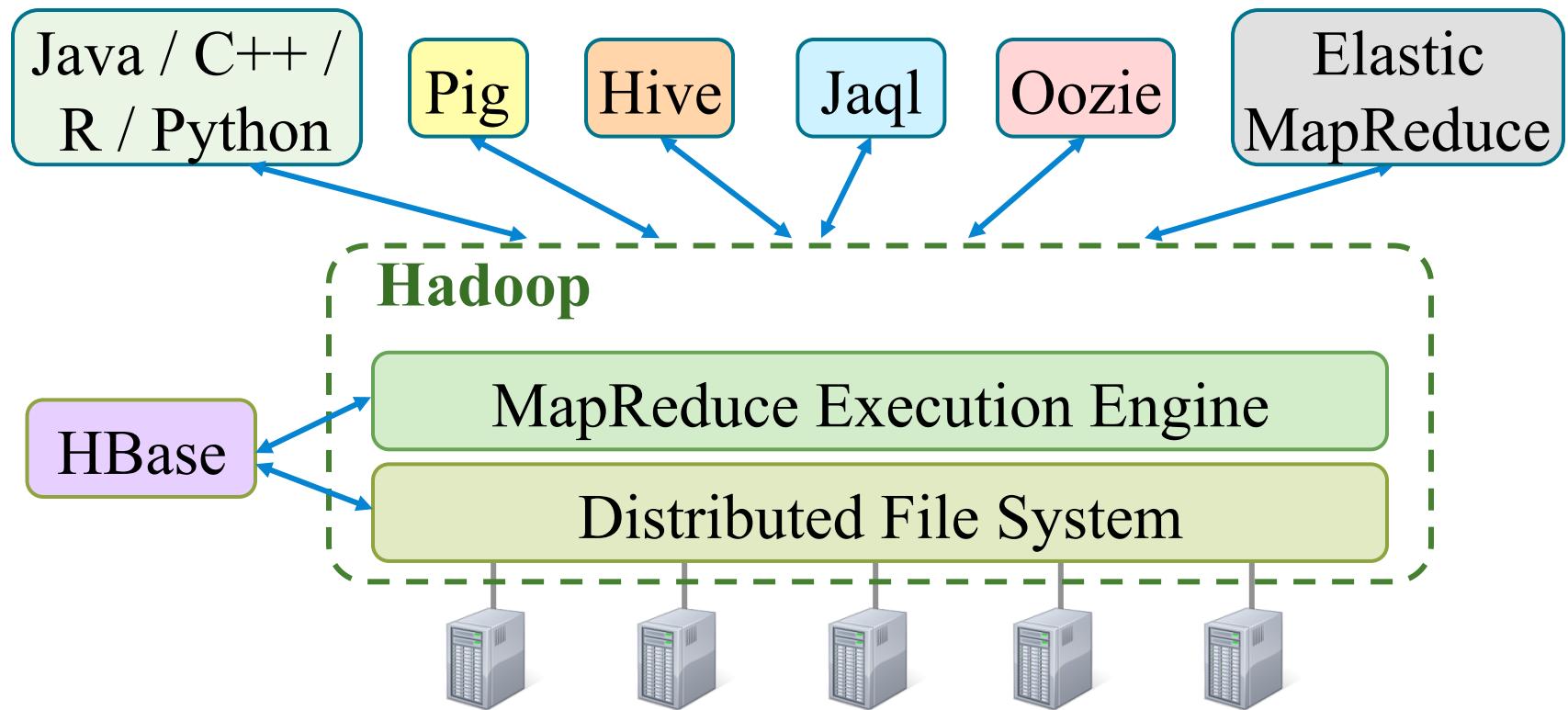


Insight

Key to Success = Timely and Cost-Effective Analysis

Hadoop MapReduce Ecosystem

- Popular **solution** to Big Data Analytics



Practitioners of Big Data Analytics

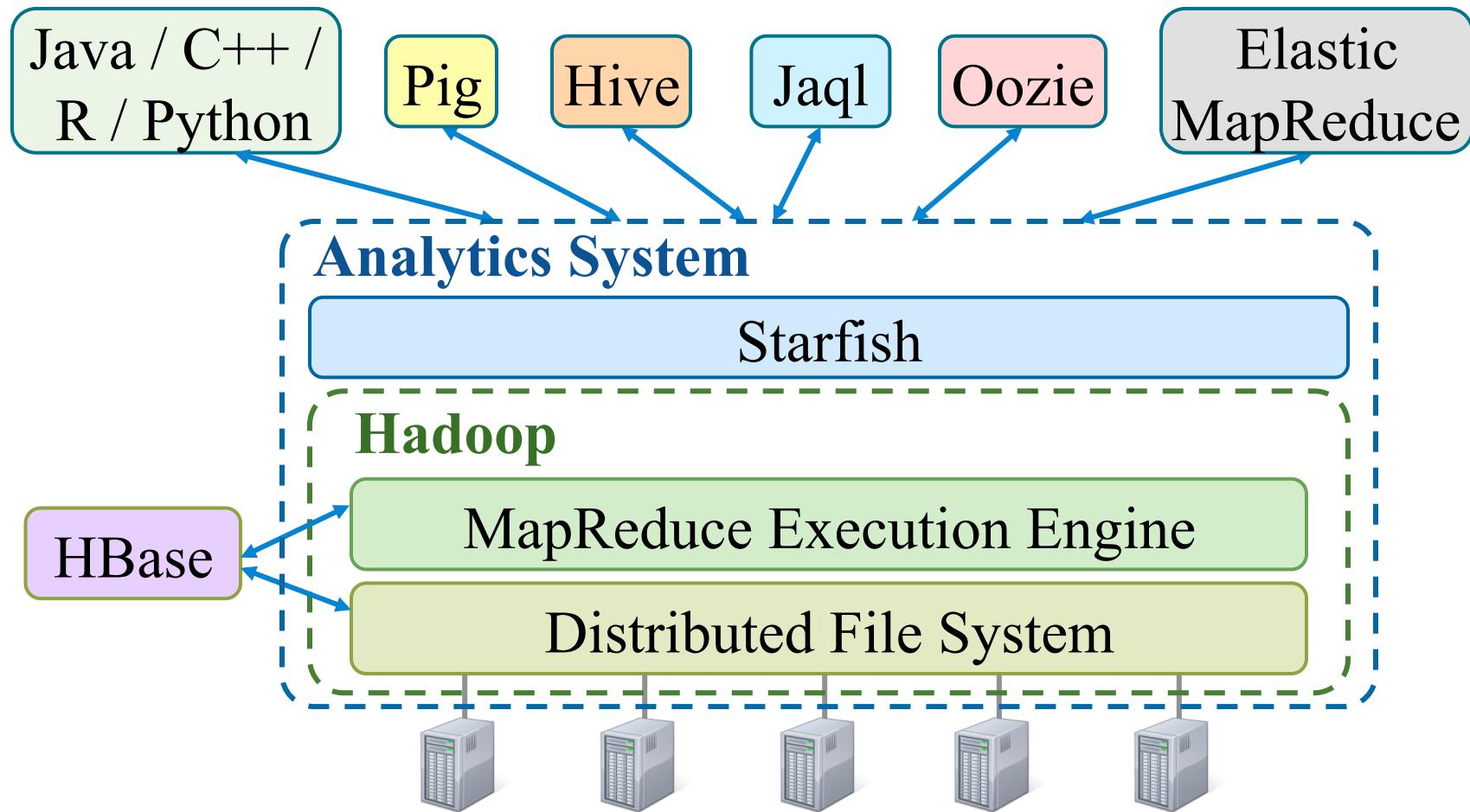
- Who are the users?
 - Data analysts, statisticians, computational scientists...
 - Researchers, developers, testers...
 - You!
- Who performs setup and tuning?
 - The users!
 - Usually lack expertise to tune the system

Tuning Challenges

- Heavy use of **programming languages** for MapReduce programs (e.g., Java/python)
- Data loaded/accessed as **opaque files**
- Large space of tuning choices
- **Elasticity** is wonderful, but hard to achieve (Hadoop has many useful mechanisms, but policies are lacking)
- **Terabyte-scale** data cycles

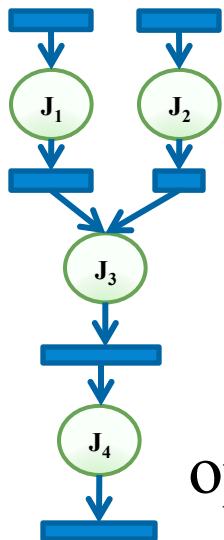
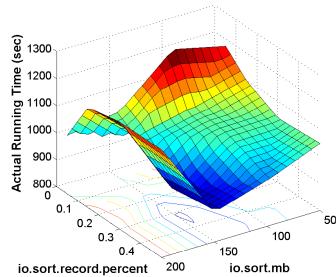
Starfish: Self-tuning System

- Our goal: Provide good performance automatically



What are the Tuning Problems?

Job-level
MapReduce
configuration

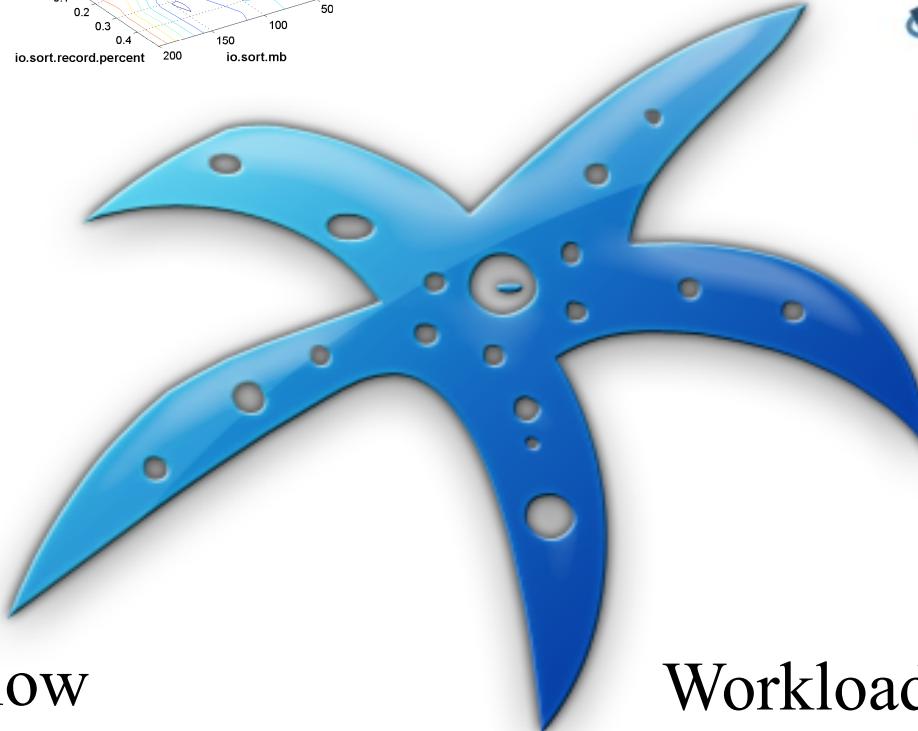


Workload
management

Cluster sizing



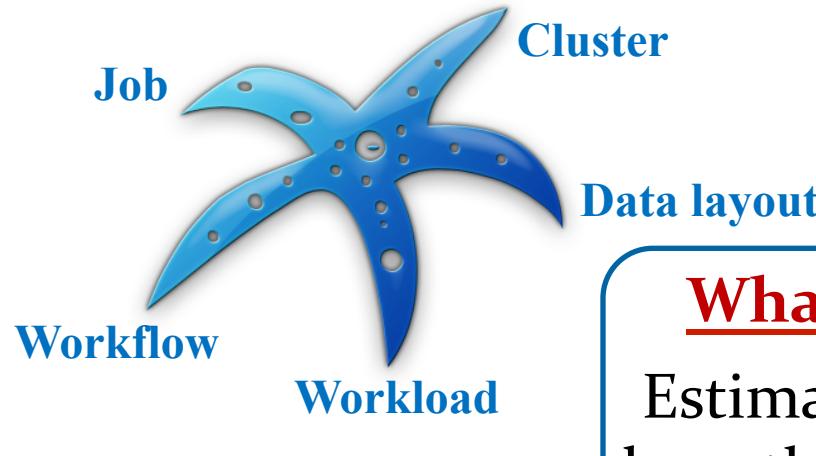
Data
layout
tuning



Starfish's Core Approach to Tuning

Optimizers

Search through space of tuning choices



Profiler

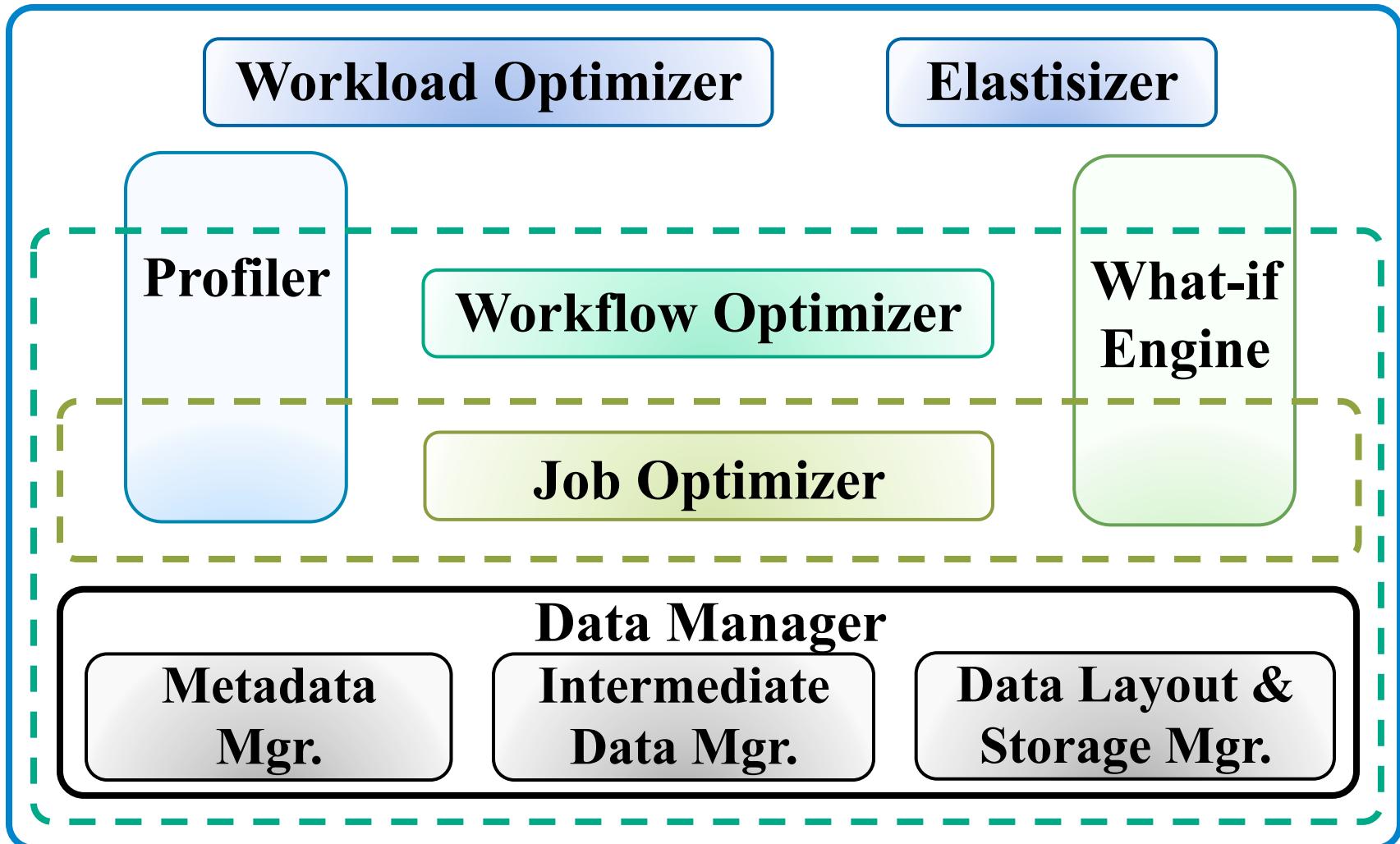
Collects concise summaries of execution

What-if Engine

Estimates impact of hypothetical changes on execution

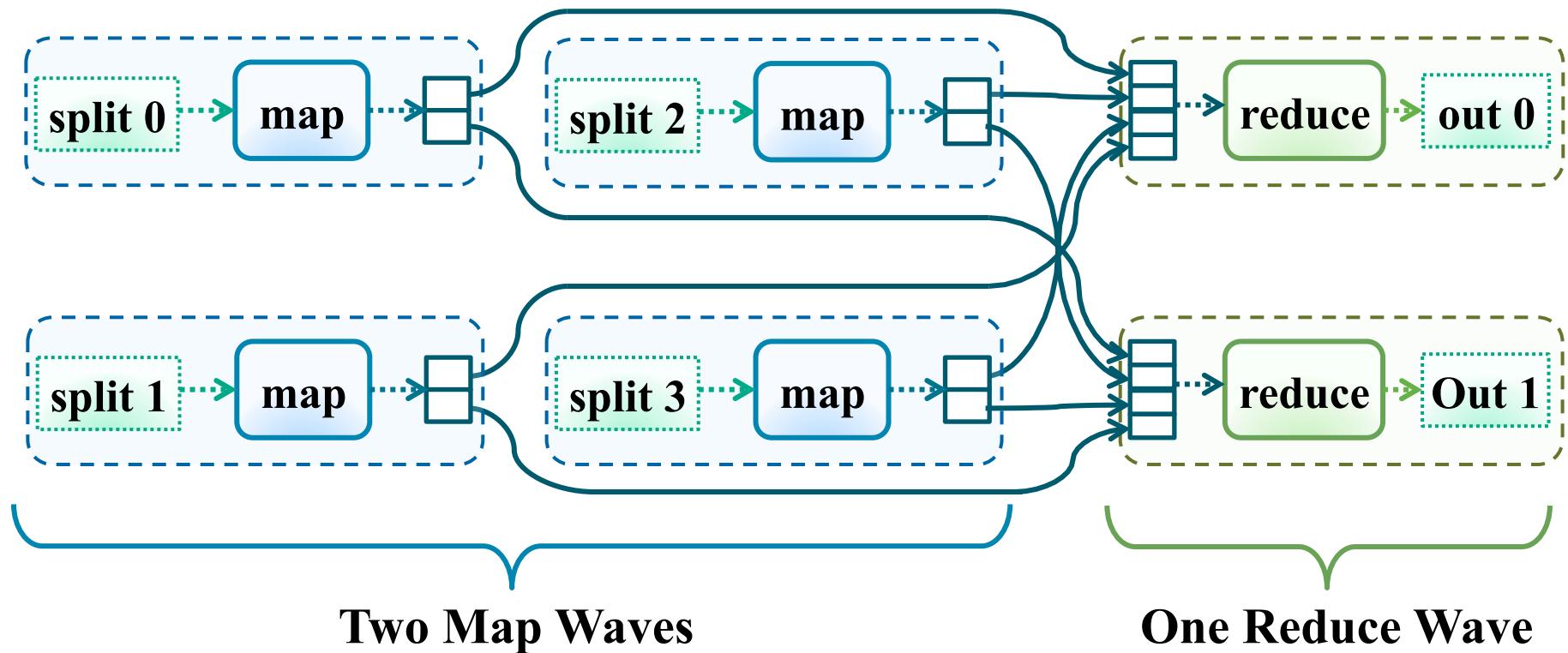
- 1) if $\Delta(\text{conf. parameters})$ then what ...?
- 2) if $\Delta(\text{data properties})$ then what ...?
- 3) if $\Delta(\text{cluster properties})$ then what ...?

Starfish Architecture



MapReduce Job Execution

job $j = \langle \text{program } p, \text{ data } d, \text{ resources } r, \text{ configuration } c \rangle$



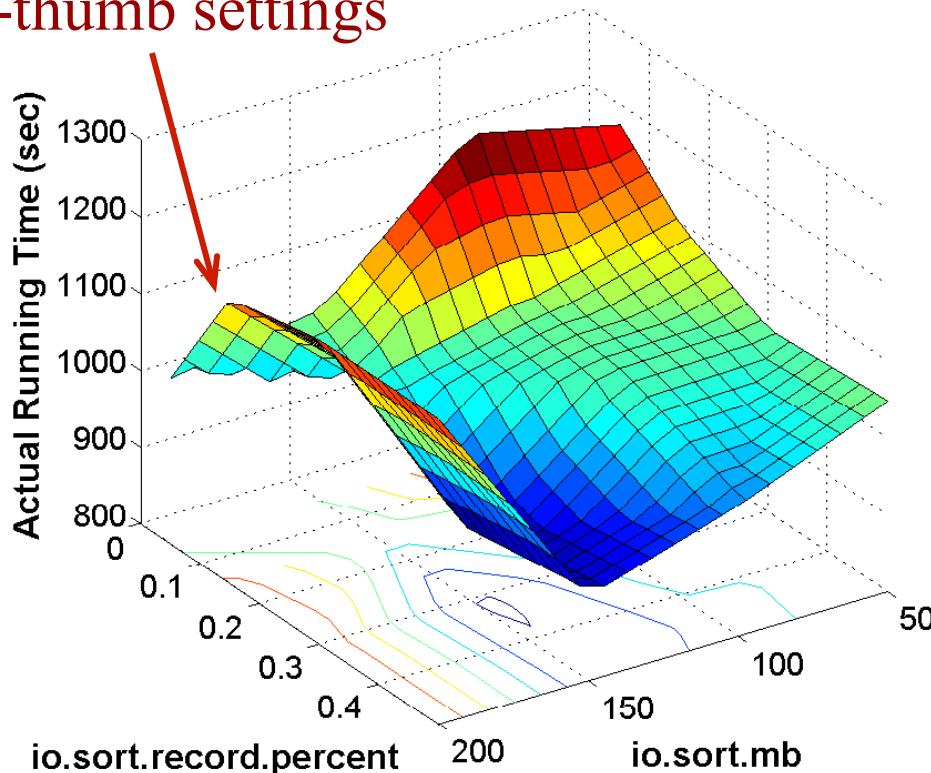
What Controls MR Job Execution?

$\text{job } j = \langle \text{program } p, \text{ data } d, \text{ resources } r, \text{ configuration } c \rangle$

- Space of configuration choices:
 - Number of map tasks
 - Number of reduce tasks
 - Partitioning of map outputs to reduce tasks
 - Memory allocation to task-level buffers
 - Multiphase external sorting in the tasks
 - Whether output data from tasks should be compressed
 - Whether combine function should be used

Effect of Configuration Settings

Rules-of-thumb settings



Two-dimensional projection of a multi-dimensional surface
(Word Co-occurrence MapReduce Program)

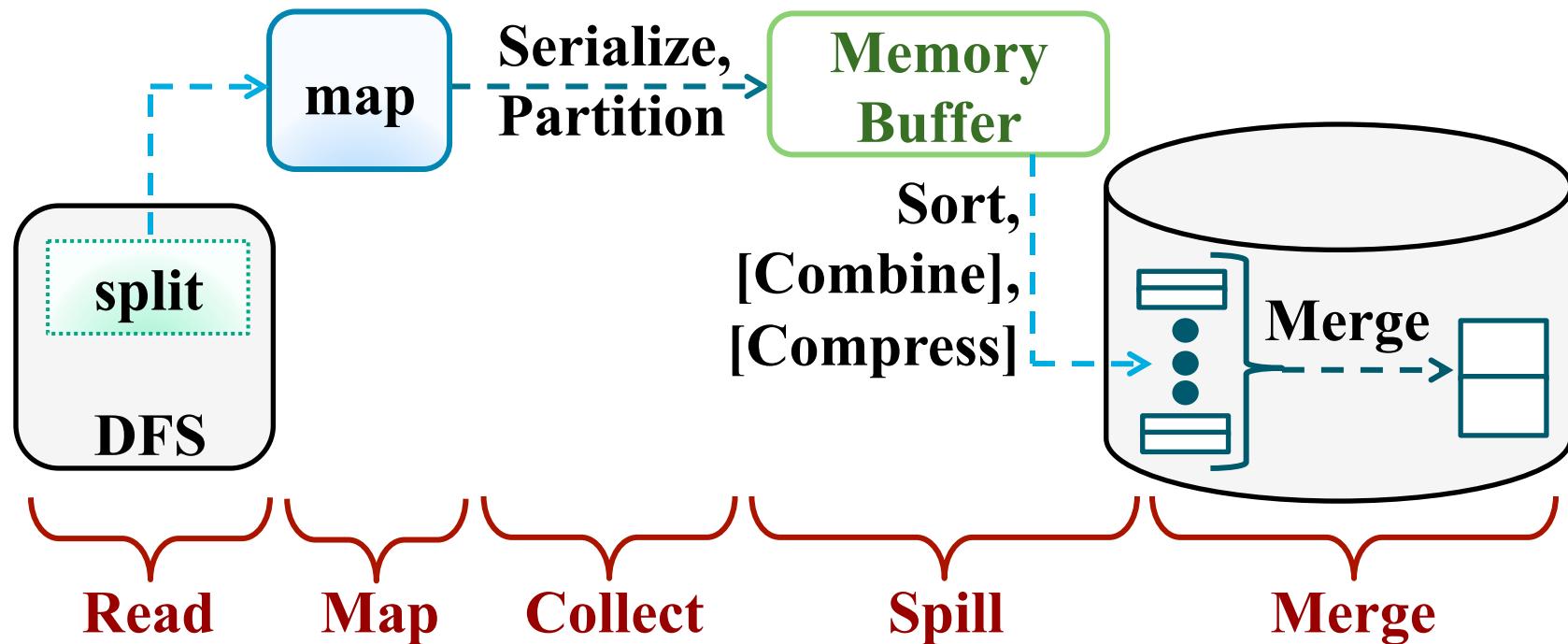
- Use **defaults** or set **manually** (rules-of-thumb)
- Rules-of-thumb **may not** suffice

MapReduce Job Tuning in a Nutshell

- Goal: $\text{perf} = F(p, d, r, c)$
 $c_{opt} = \arg \min_{c \in S} F(p, d, r, c)$
- Challenges: p is an arbitrary MapReduce program; c is high-dimensional; ...
- Profiler Runs p to collect a ***job profile*** (concise execution summary) of $\langle p, d_1, r_1, c_1 \rangle$
- What-if Engine Given profile of $\langle p, d_1, r_1, c_1 \rangle$, estimates ***virtual profile*** for $\langle p, d_2, r_2, c_2 \rangle$
- Optimizer Enumerates and searches through the ***optimization space S*** efficiently

Job Profile

- Concise representation of program execution as a job
- Records information at the level of “task phases”
- Generated by Profiler through measurement or by the What-if Engine through estimation



Job Profile Fields

Dataflow: amount of data flowing through task phases

Map output bytes

Number of spills

Number of records in buffer per spill

⋮

Costs: execution times at the level of task phases

Read phase time in the map task

Map phase time in the map task

Spill phase time in the map task

⋮

Dataflow Statistics: statistical information about dataflow

Width of input key-value pairs

Map selectivity in terms of records

Map output compression ratio

⋮

Cost Statistics: statistical information about resource costs

I/O cost for reading from local disk per byte

CPU cost for executing the Mapper per record

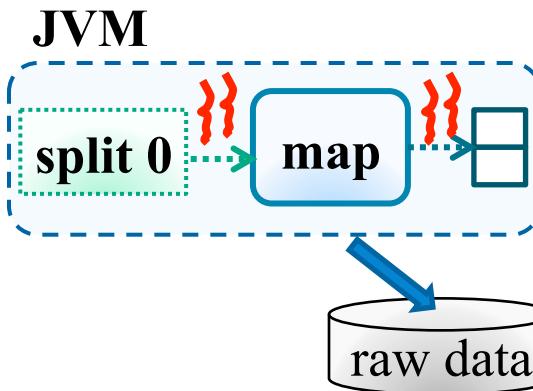
CPU cost for uncompressed the input per byte

⋮

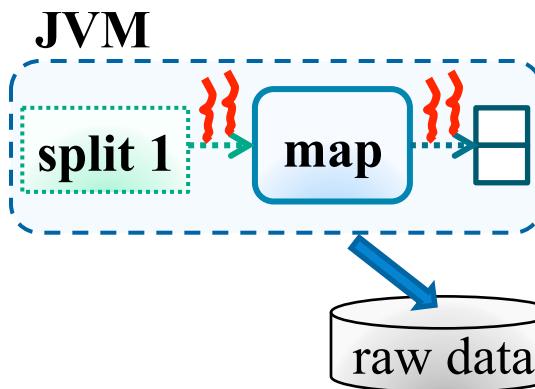
Generating Profiles by Measurement

- Goals
 - Have zero overhead when profiling is turned off
 - Require no modifications to Hadoop
 - Support unmodified MapReduce programs written in Java or Hadoop Streaming/Pipes (Python/Ruby/C++)
- Approach: Dynamic (on-demand) instrumentation
 - Event-condition-action rules are specified (in Java)
 - Leads to run-time instrumentation of Hadoop internals
 - Monitors task phases of MapReduce job execution
 - We currently use Btrace (Hadoop internals are in Java)

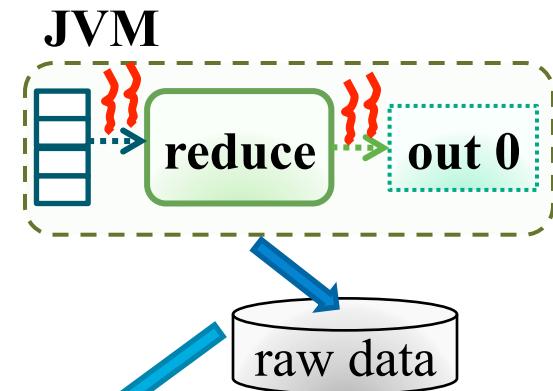
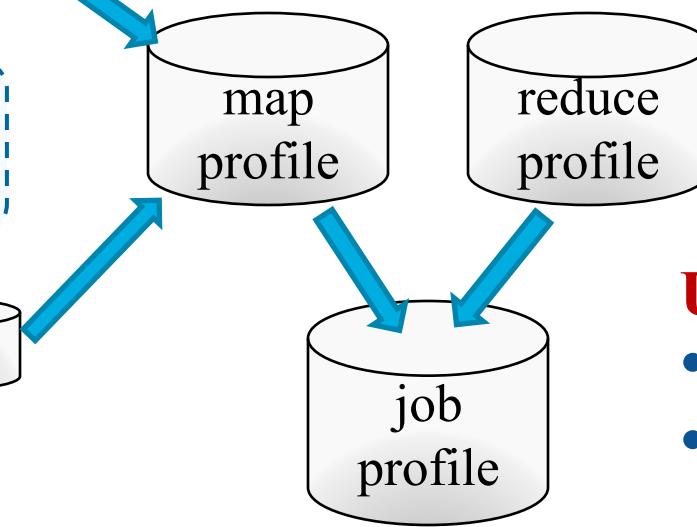
Generating Profiles by Measurement



Enable Profiling



ECA rules



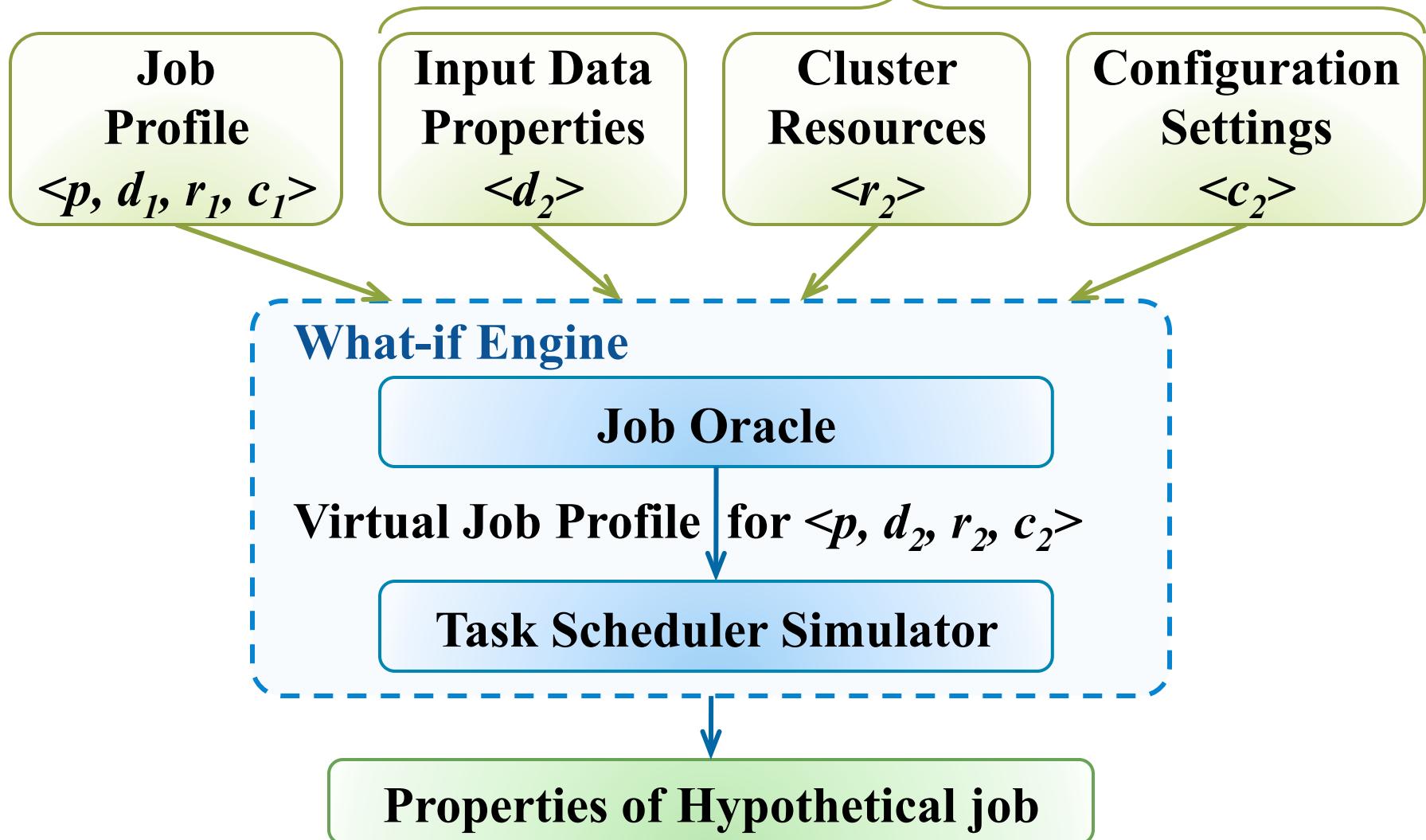
Use of Sampling

- Profile fewer tasks
- Execute fewer tasks

JVM = Java Virtual Machine, ECA = Event-Condition-Action

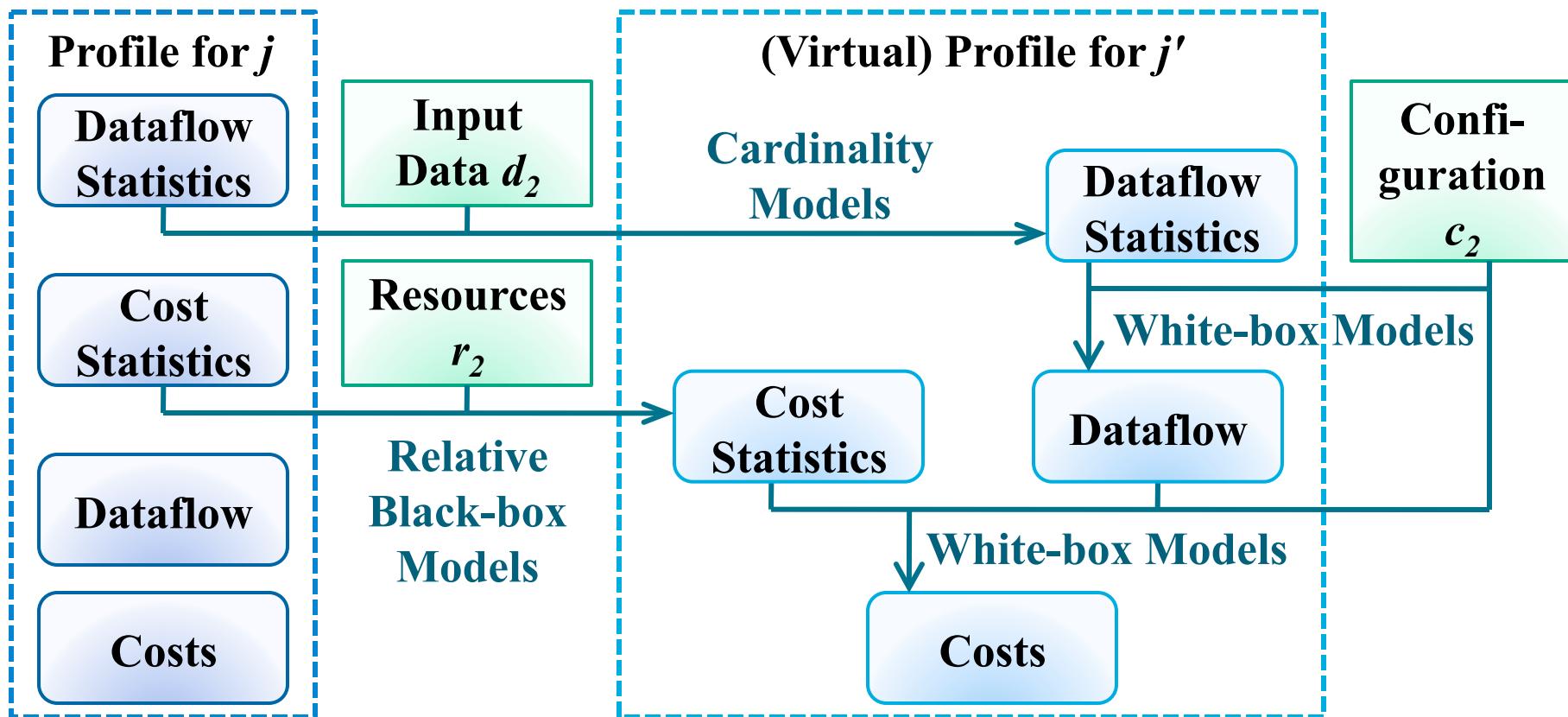
What-if Engine

Possibly Hypothetical

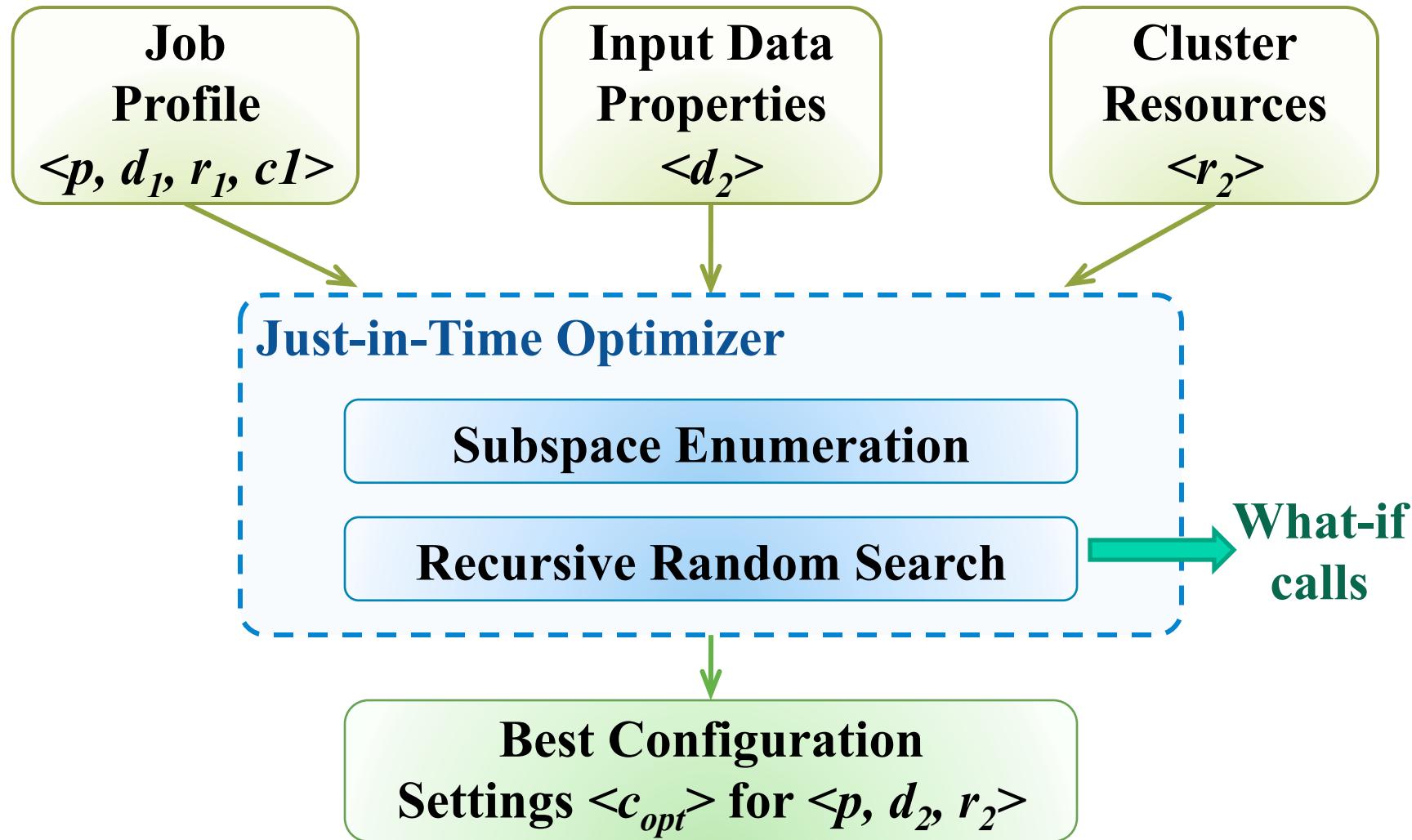


Virtual Profile Estimation

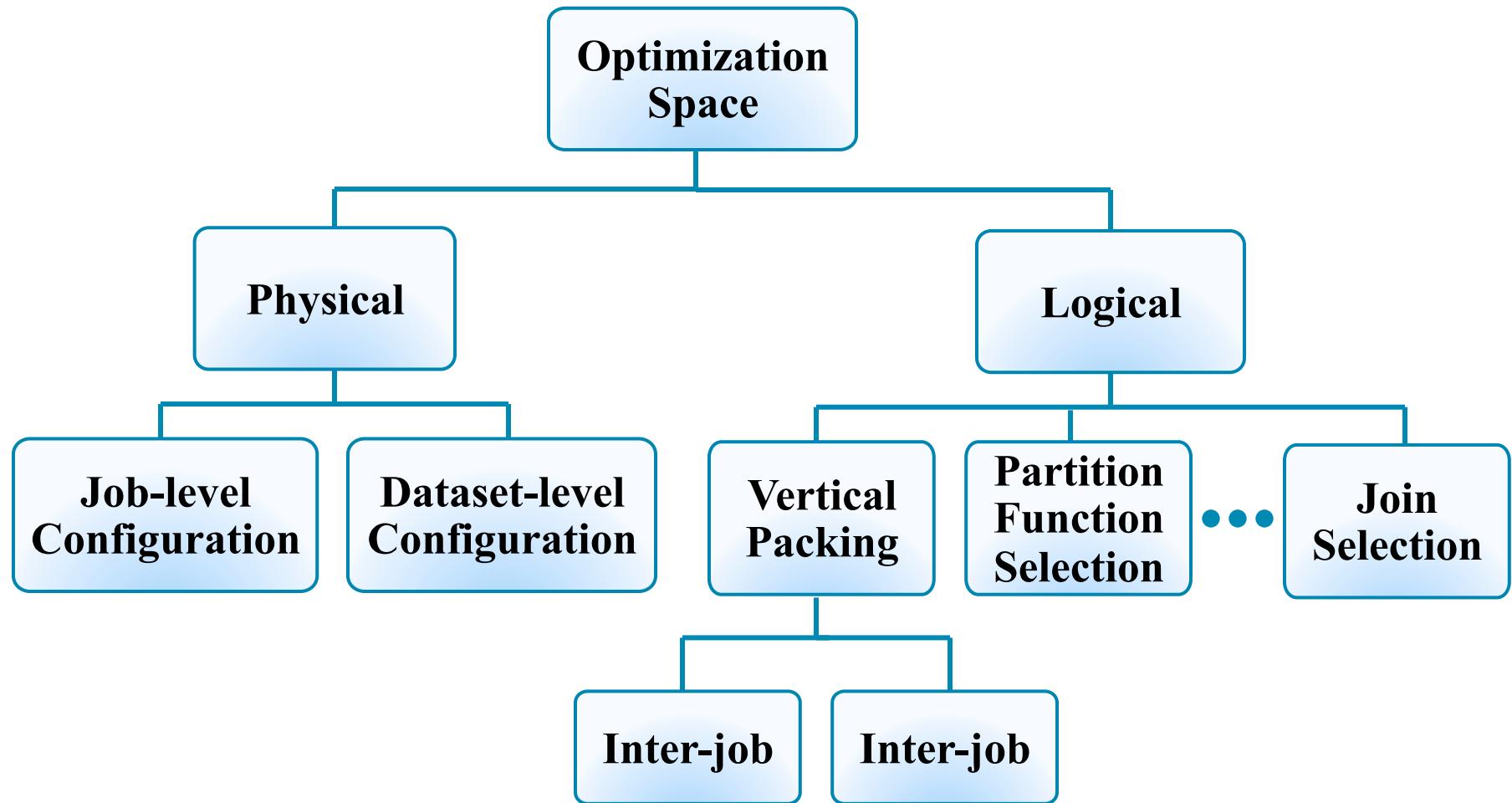
Given profile for job $j = \langle p, d_1, r_1, c_1 \rangle$
estimate profile for job $j' = \langle p, d_2, r_2, c_2 \rangle$



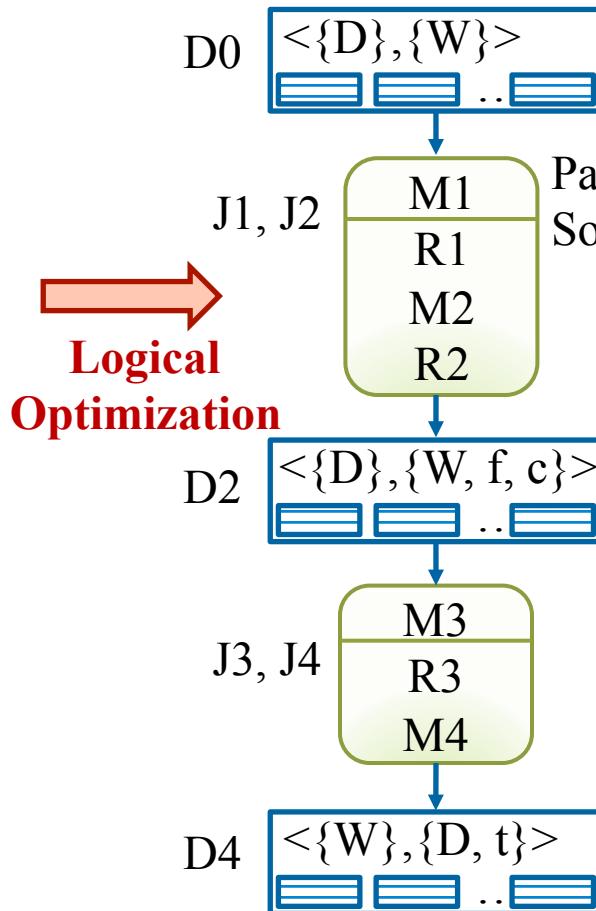
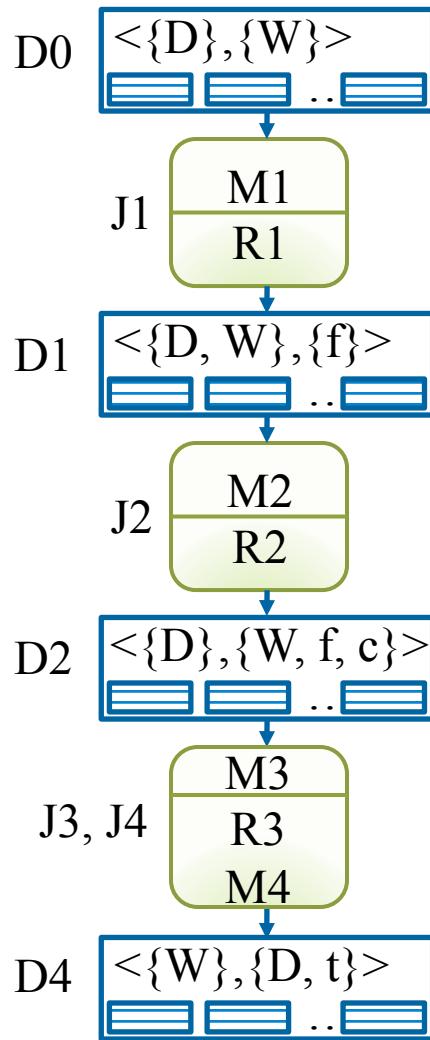
Job Optimizer



Workflow Optimization Space

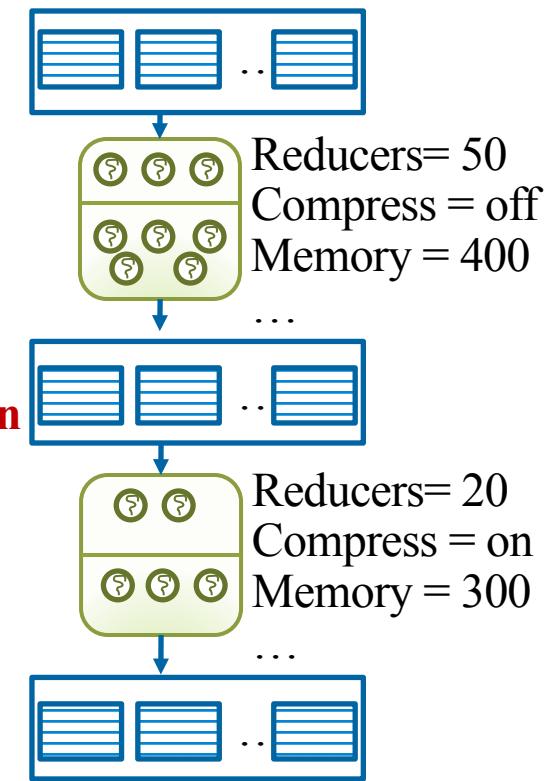


Optimizations on TF-IDF Workflow



Partition: {D}
Sort: {D, W}

Physical Optimization

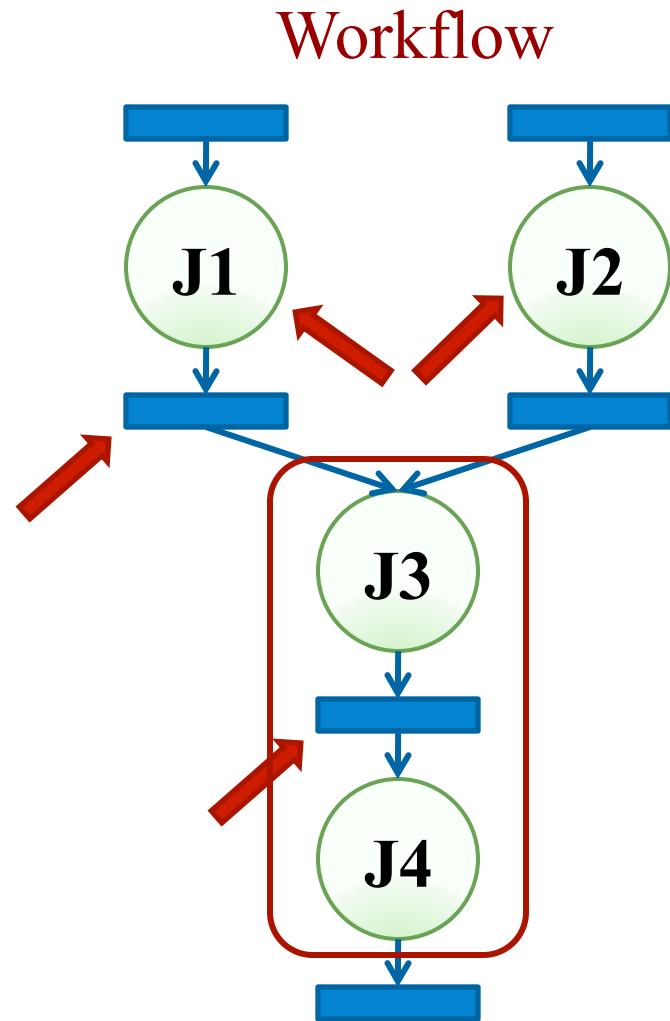


Legend

D = docname f = frequency
W = word c = count
t = TF-IDF

New Challenges

- What-if challenges:
 - Support concurrent job execution
 - Estimate intermediate data properties
- Optimization challenges
 - Interactions across jobs
 - Extended optimization space
 - Find good configuration settings for individual jobs

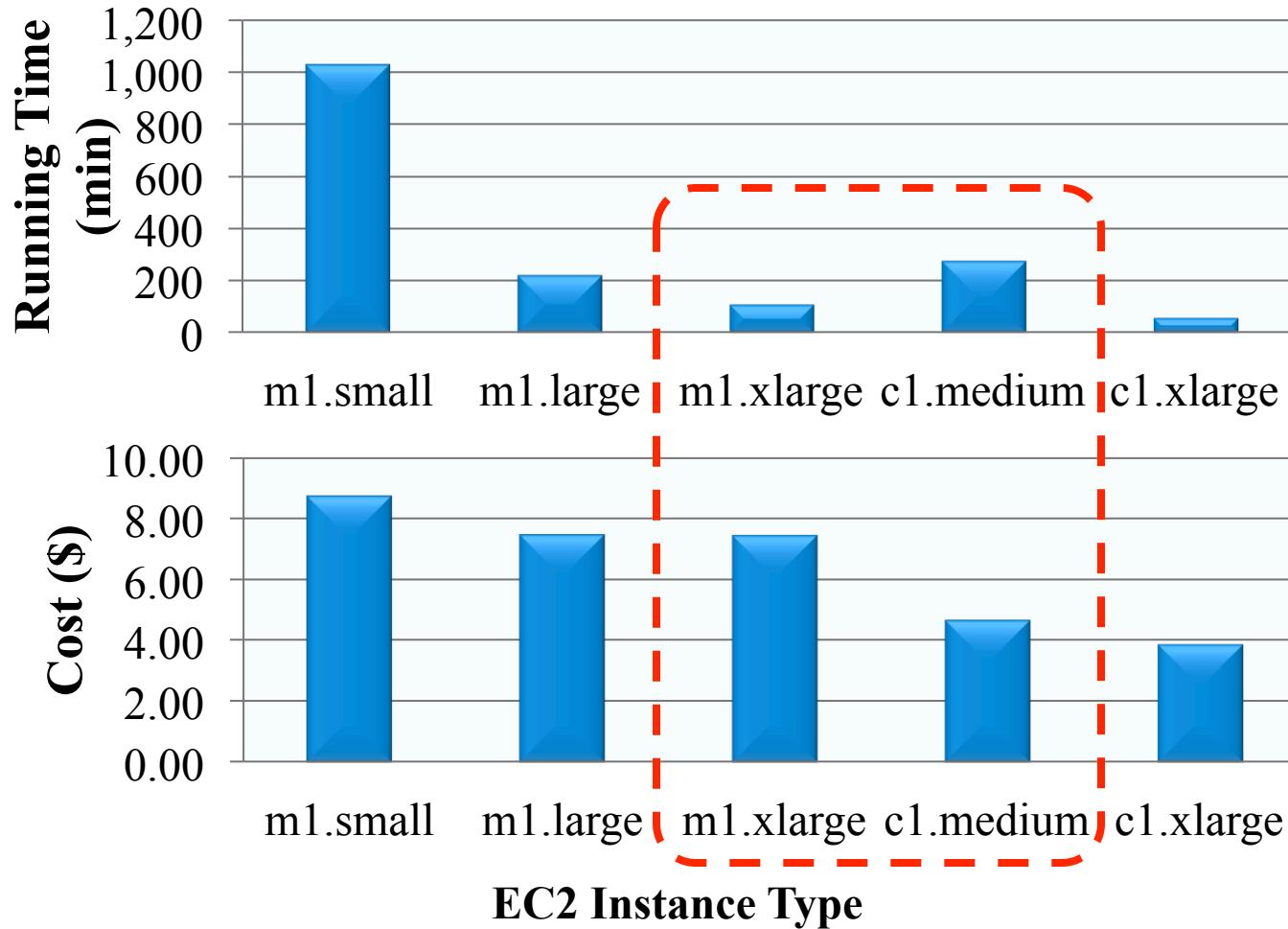


Cluster Sizing Problem

- Use-cases for cluster sizing
 - Tuning the cluster size for elastic workloads
 - Workload transitioning from development cluster to production cluster
 - Multi-objective cluster provisioning
- Goal
 - Determine cluster resources & job-level configuration parameters to meet workload requirements

Multi-objective Cluster Provisioning

- Cloud enables users to provision clusters in minutes



Experimental Evaluation

- Starfish (versions 0.1, 0.2) to manage Hadoop on EC2
- Different scenarios: Cluster × Workload × Data

EC2 Node Type	CPU: EC2 units	Mem	I/O Perf.	Cost / hour	#Maps /node	#Reds /node	MaxMem /task
m1.small	1 (1 x 1)	1.7 GB	moderate	\$0.085	2	1	300 MB
m1.large	4 (2 x 2)	7.5 GB	high	\$0.34	3	2	1024 MB
m1.xlarge	8 (4 x 2)	15 GB	high	\$0.68	4	4	1536 MB
c1.medium	5 (2 x 2.5)	1.7 GB	moderate	\$0.17	2	2	300 MB
c1.xlarge	20 (8 x 2.5)	7 GB	high	\$0.68	8	6	400 MB
cc1.4xlarge	33.5 (8)	23 GB	very high	\$1.60	8	6	1536 MB

Experimental Evaluation

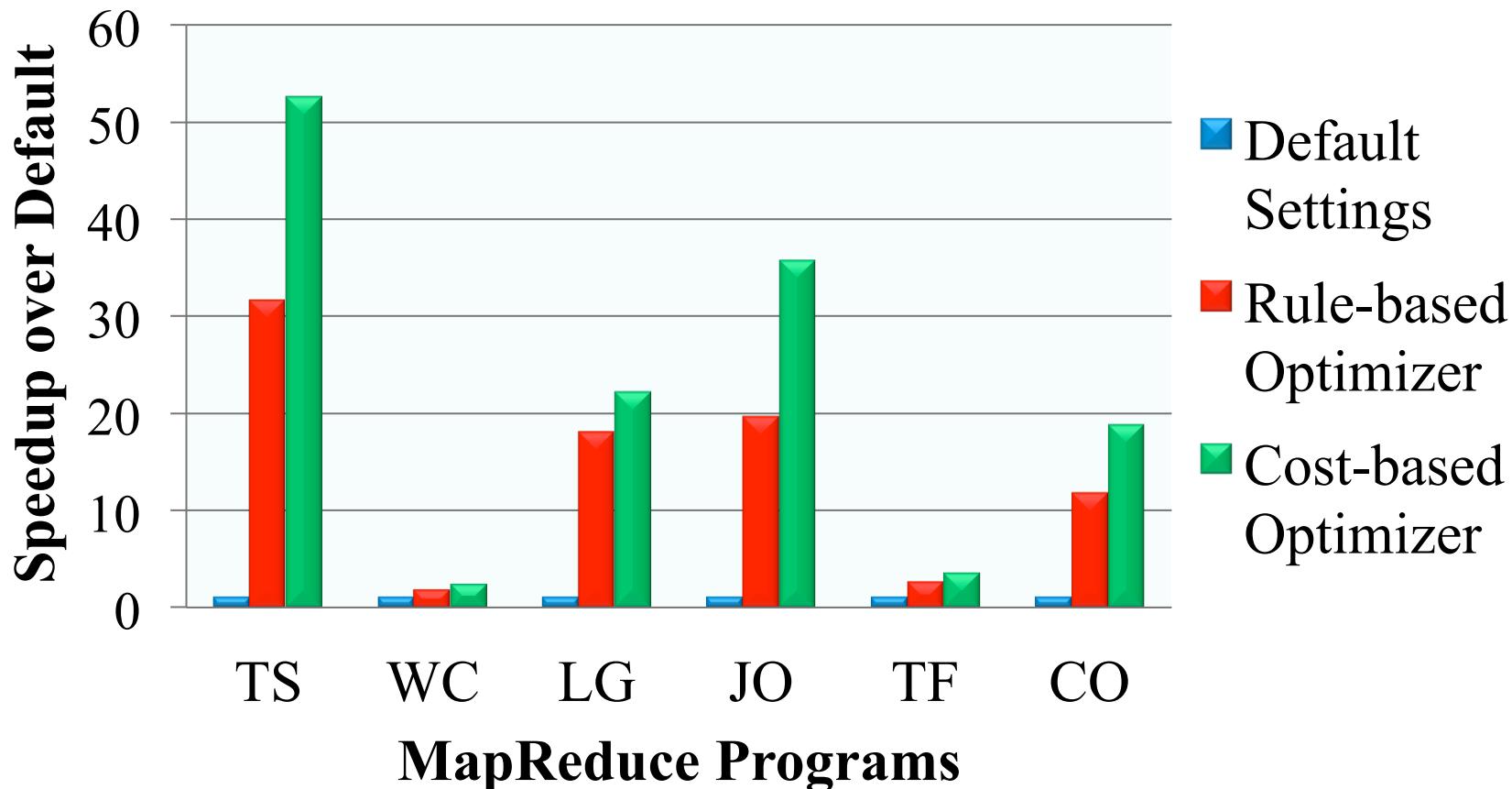
- Starfish (versions 0.1, 0.2) to manage Hadoop on EC2
- Different scenarios: Cluster × Workload × Data

Abbr.	MapReduce Program	Domain	Dataset
CO	Word Co-occurrence	Natural Lang Proc.	Wikipedia (10GB – 22GB)
WC	WordCount	Text Analytics	Wikipedia (30GB – 1TB)
TS	TeraSort	Business Analytics	TeraGen (30GB – 1TB)
LG	LinkGraph	Graph Processing	Wikipedia (compressed ~6x)
JO	Join	Business Analytics	TPC-H (30GB – 1TB)
TF	Term Freq. - Inverse Document Freq.	Information Retrieval	Wikipedia (30GB – 1TB)

Job Optimizer Evaluation

Hadoop cluster: 30 nodes, m1.xlarge

Data sizes: 60-180 GB

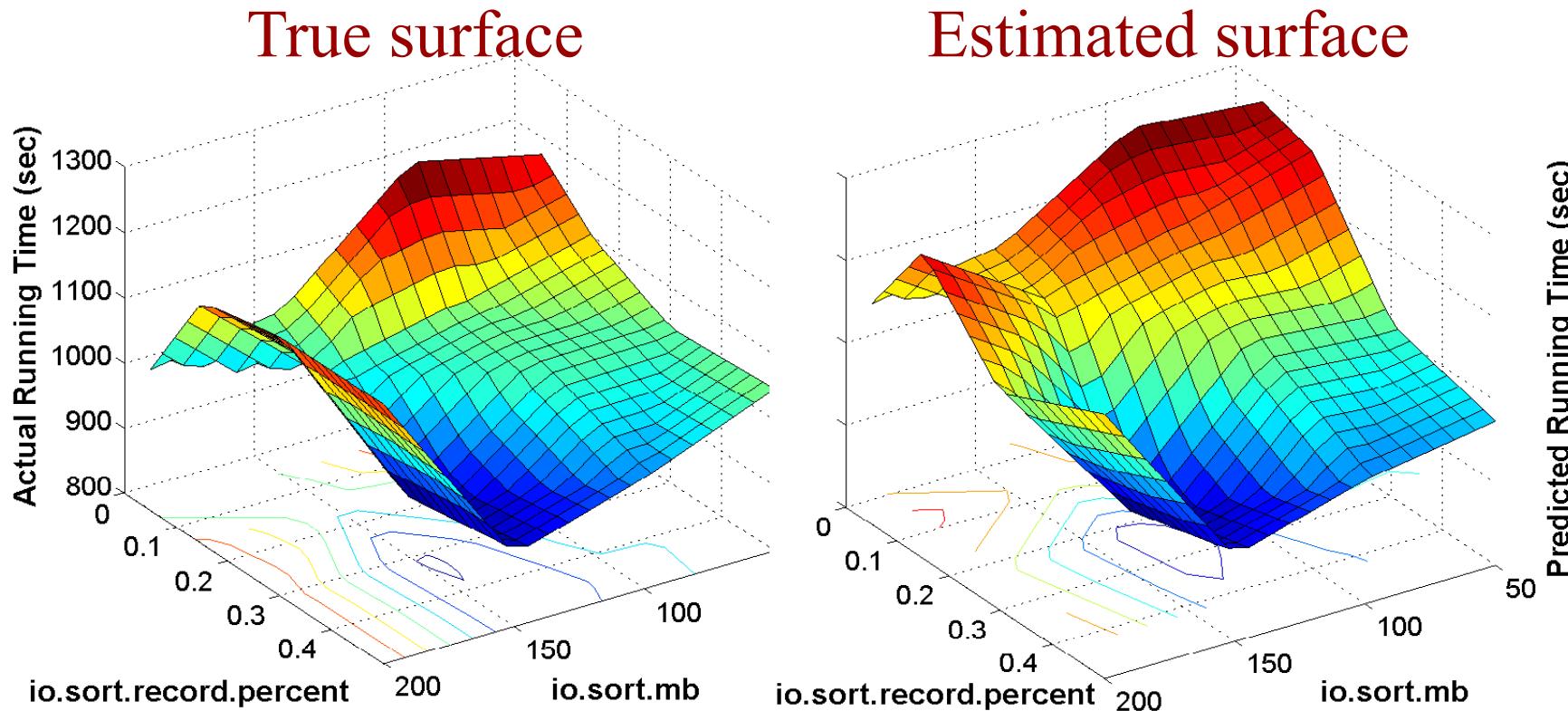


Estimates from the What-if Engine

Hadoop cluster: 16 nodes, c1.medium

MapReduce Program: Word Co-occurrence

Data set: 10 GB Wikipedia

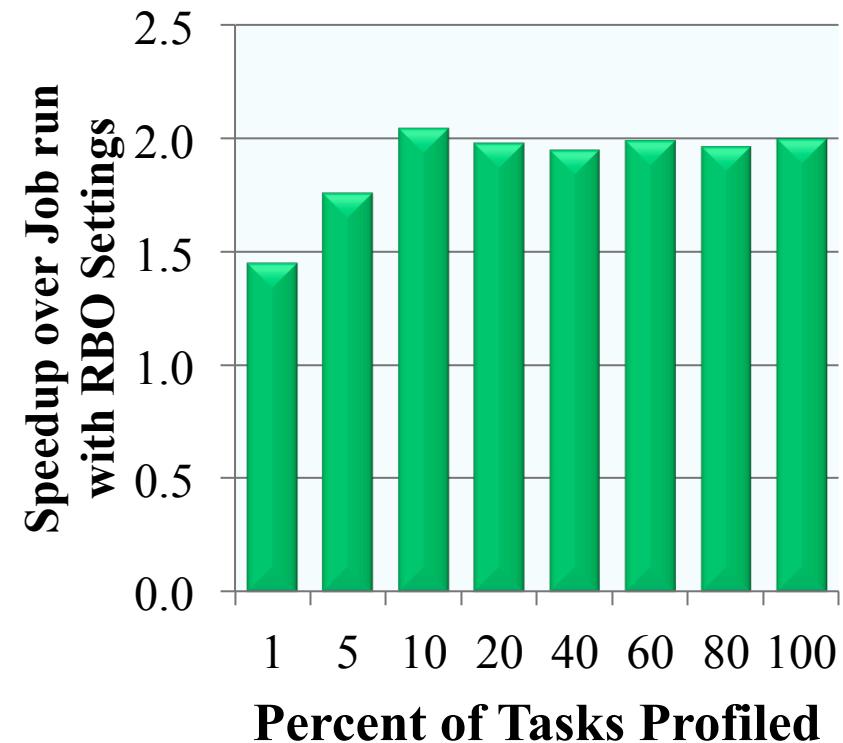
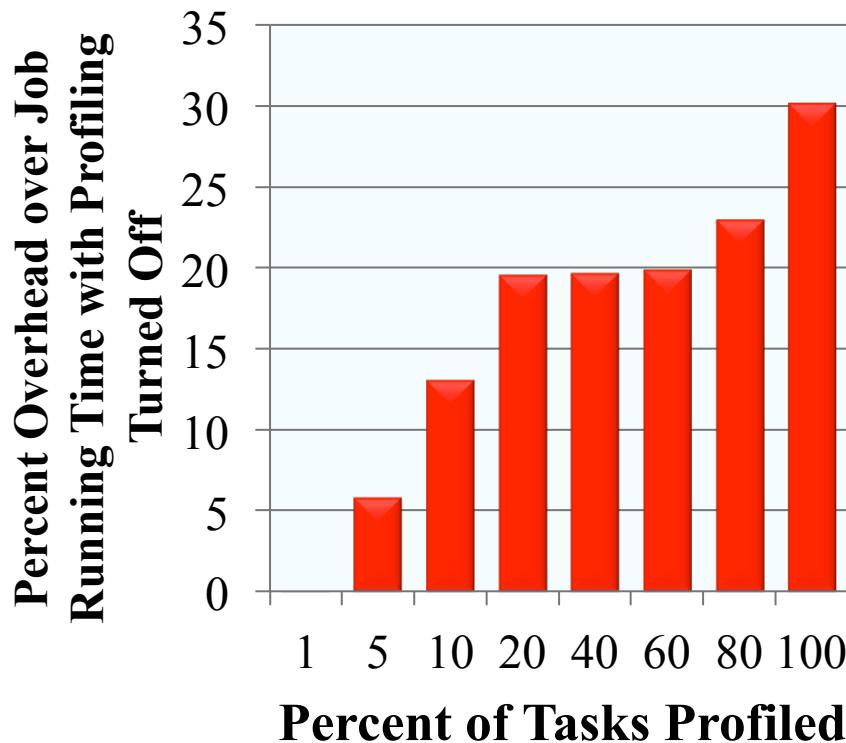


Profiling Overhead Vs. Benefit

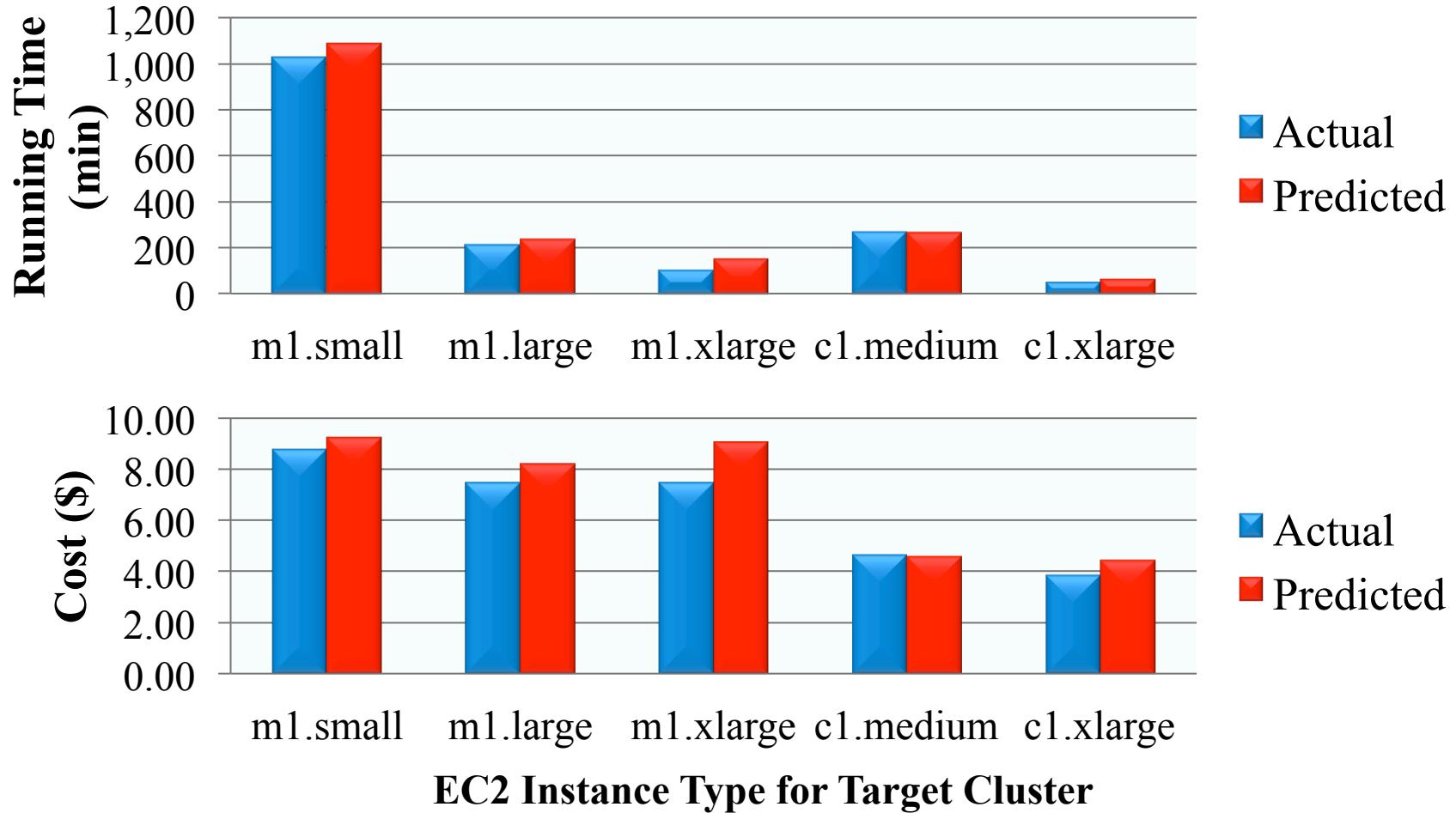
Hadoop cluster: 16 nodes, c1.medium

MapReduce Program: Word Co-occurrence

Data set: 10 GB Wikipedia



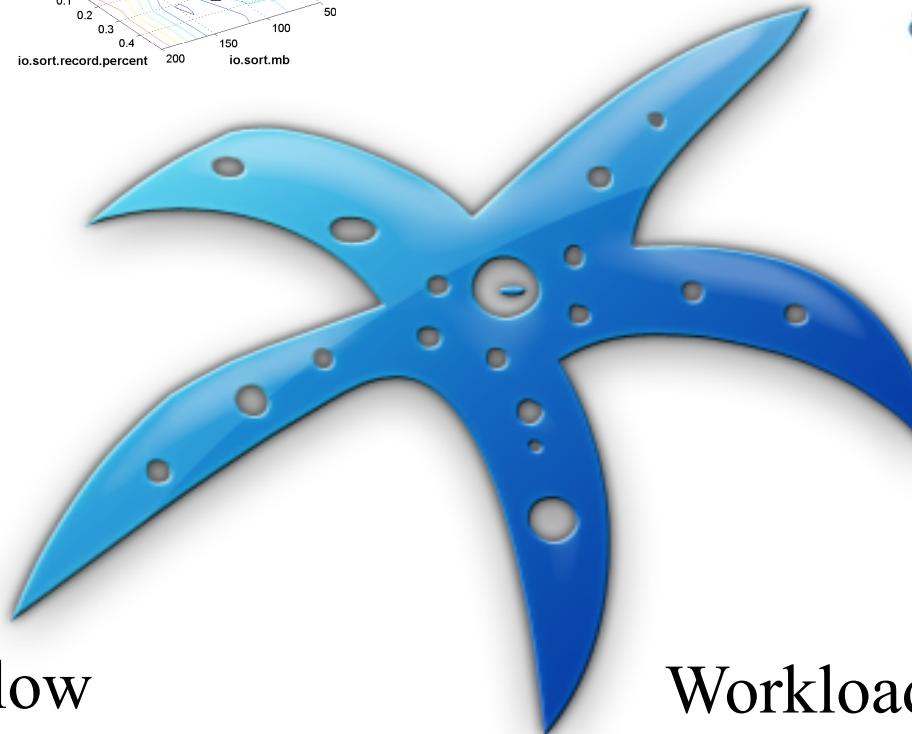
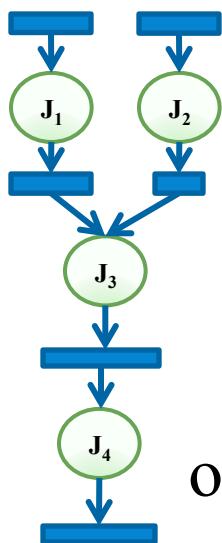
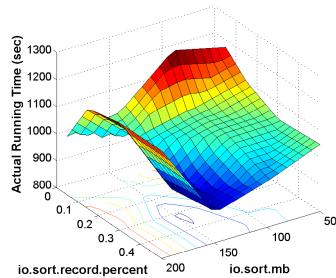
Multi-objective Cluster Provisioning



Instance Type for Source Cluster: m1.large

More info: www.cs.duke.edu/starfish

Job-level
MapReduce
configuration



Cluster sizing



Data
layout
tuning

Workload
management