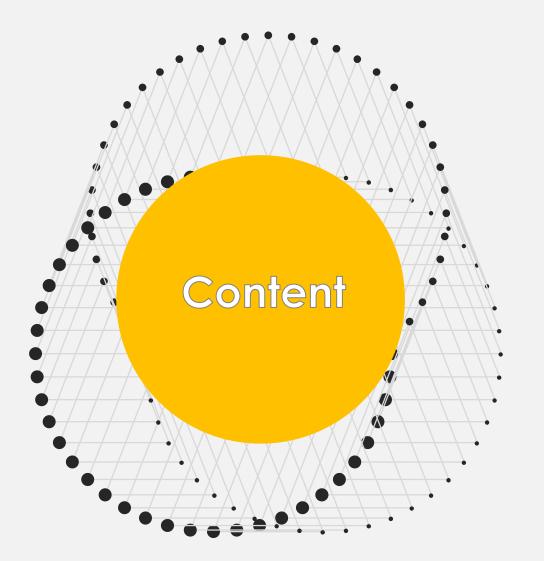


Starfish: A Self-tuning System for Big Data Analytics

Authors: Herodotos Herodotou, Harold Lim, Fei Dong, Shivnath Babu



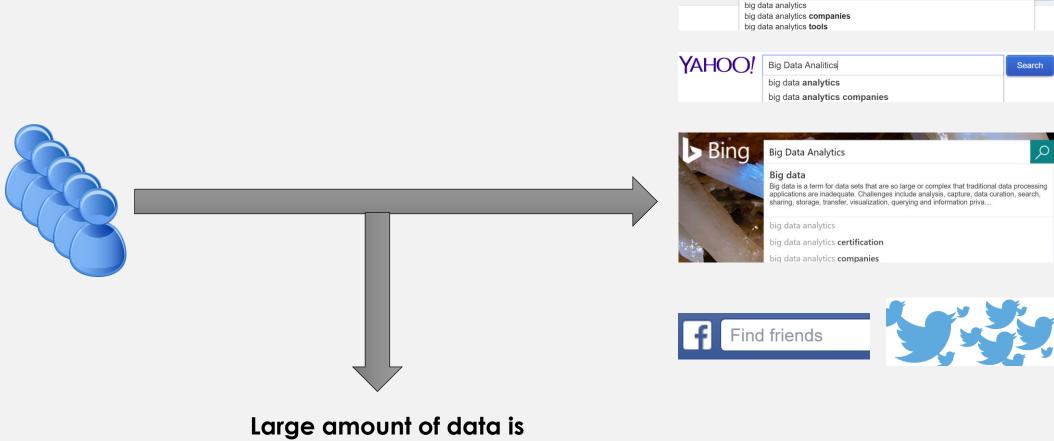


- Part 1 Introduction
- Part 2 Overview
- Part 3 Job-level Tuning
- Part 4 Workflow-aware tuning
- Part 5 Workload-level Tuning

Part 6 Critique



Big Data is EVERYWHERE!



Google

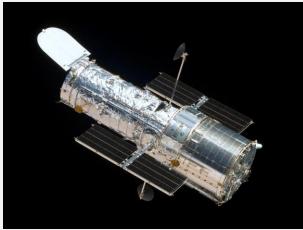
Big Data Analytics

J Q

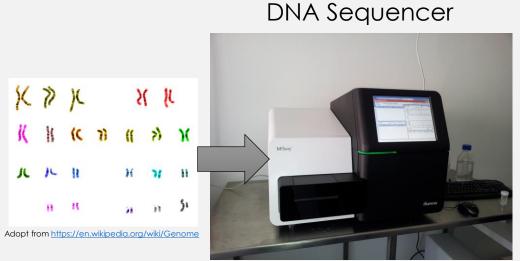
captured and analyzed

Big Data is EVERYWHERE!

Hubble Space Telescope

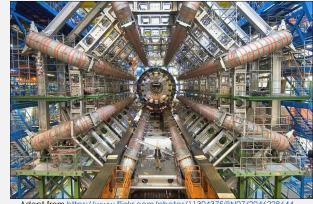


Adopt from https://en.wikipedia.org/wiki/Hubble Space Telescope



Adopt from https://en.wikipedia.org/wiki/DNA_sequencing

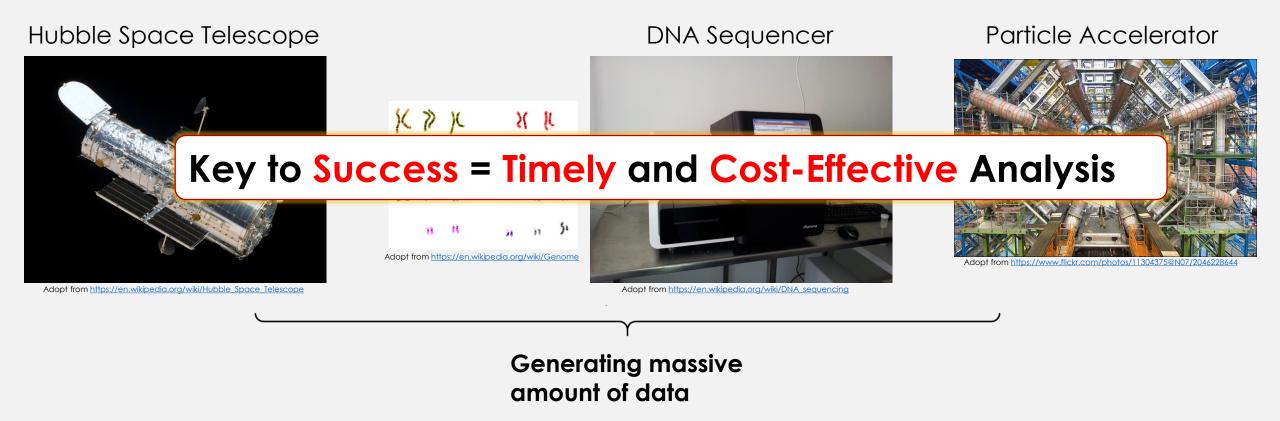
Particle Accelerator



Adopt from https://www.flickr.com/photos/11304375@N07/2046228644

Generating massive amount of data

Big Data is EVERYWHERE!



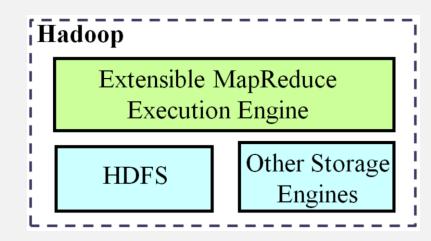
What we need is a MAD system

Magnetism "Attracts" or welcomes all sources of data, regardless of structure, values, etc.

Agility Adaptive, remains in sync with rapid data evolution and modification

Depth More than just your typical analytics, we need to support complex operations like statistical analysis and machine learning





- Copying files into the Distributed File System is all it takes to get data into Hadoop
- Interpret data at processing time, and not at loading time
- Supports different kinds of programming languages

eded

art

ill

lugin

Want best performance? Non-trivial!

mapreduce.jobtracker.jobhistory.location mapreduce.job.reduces mapreduce.map.cpu.vcores mapreduce.jobtracker.jobhistory.task.num mapreduce.jobtracker.restart.recover mapreduce.reduce.memory.mb berprogresssplits mapreduce.jobtracker.jobhistory.block.size mapreduce.reduce.cpu.vcores mapreduce.job.userhistorylocation mapreduce.jobtracker.taskscheduler mapreduce.jobtracker.retiredjobs.cache.siz mapreduce.jobtracker.jobhistory.complete mapreduce.job.running.map.limit d.location mapreduce.tasktracker.outofband.heartbeat mapreduce.job.running.reduce.limit mapreduce.job.committer.setup.cleanup.ne mapreduce.jobtracker.jobhistory.lru.cache mapreduce.job.reducer.preempt.delay.sec size mapreduce.task.io.sort.factor mapreduce.job.max.split.locations mapreduce.jobtracker.instrumentation mapreduce.task.io.sort.mb mapreduce.job.split.metainfo.maxsize mapred.child.java.opts mapreduce.map.sort.spill.percent mapreduce.jobtracker.taskscheduler.maxru mapred.child.env nningtasks.neriob mapreduce.jobtracker.address mapreduce.admin.user.env manreduce man maxattempts mapreduce.local.clientfactory.class.name mapreduce.map.log.level mapreduce.reduce.maxattempts mapreduce.jobtracker.http.address mapreduce.reduce.log.level mapreduce.reduce.shuffle.fetch.retry.enabl mapreduce.jobtracker.handler.count mapreduce.map.cpu.vcores mapreduce.tasktracker.report.address mapreduce.reduce.shuffle.fetch.retry.interv mapreduce.reduce.cpu.vcores manreduce.cluster.local.dir al-ms mapreduce.reduce.merge.inmem.threshold mapreduce.jobtracker.system.dir mapreduce.reduce.shuffle.fetch.retry.timeo mapreduce.reduce.shuffle.merge.percent ut-ms mapreduce.jobtracker.staging.root.dir mapreduce.reduce.shuffle.input.buffer.perc mapreduce.reduce.shuffle.retrymapreduce.cluster.temp.dir delay.max.ms mapreduce.tasktracker.local.dir.minspacest mapreduce.reduce.input.buffer.percent mapreduce.reduce.shuffle.parallelcopies mapreduce.reduce.shuffle.memory.limit.pe mapreduce.reduce.shuffle.connect.timeout mapreduce.tasktracker.local.dir.minspacek rcent manreduce reduce shuffle read timeout mapreduce.shuffle.ssl.enabled mapreduce.jobtracker.expire.trackers.inter mapreduce.shuffle.connection-keepmapreduce.shuffle.ssl.file.buffer.size alive.enable mapreduce.shuffle.max.connections mapreduce.shuffle.connection-keepmapreduce.tasktracker.instrumentation alive.timeout mapreduce.shuffle.max.threads manreduce.tasktracker.resourcecalculatorp mapreduce.task.timeout mapreduce.shuffle.transferTo.allowed mapreduce.tasktracker.taskmemorymanage mapreduce.tasktracker.map.tasks.maximu mapreduce.shuffle.transfer.buffer.size r.monitoringinterval mapreduce.reduce.markreset.buffer.percer mapreduce.tasktracker.tasks.sleeptimebefo mapreduce.tasktracker.reduce.tasks.maxim t resigkill um mapreduce.map.speculative mapreduce.job.maps mapreduce.map.memory.mb

mapreduce.reduce.speculative	mapreduce.output.fileoutputformat.compre	mapreduce.jobtracker.persist.jobstatus.hou
mapreduce.job.speculative.speculative-	55	rs
cap-running-tasks	mapreduce.output.fileoutputformat.compre	mapreduce.jobtracker.persist.jobstatus.dir
mapreduce.job.speculative.speculative-	ss.type	mapreduce.task.profile
cap-total-tasks	mapreduce.output.fileoutputformat.compre	
mapreduce.job.speculative.minimum-	ss.codec	mapreduce.task.profile.maps
allowed-tasks	mapreduce.map.output.compress	mapreduce.task.profile.reduces
		mapreduce.task.profile.params
mapreduce.job.speculative.retry-after-no-	mapreduce.map.output.compress.codec	1
speculate	map.sort.class	mapreduce.task.profile.map.params
mapreduce.job.speculative.retry-after-	mapreduce.task.userlog.limit.kb	mapreduce.task.profile.reduce.params
speculate	mapreduce.task.useriog.timit.ko	mapreduce.task.skip.start.attempts
mapreduce.job.map.output.collector.class	yarn.app.mapreduce.am.container.log.limit	1.000 00 10 00 10
	.kb	mapreduce.map.skip.proc.count.autoincr
mapreduce.job.speculative.slowtaskthresho	yarn.app.mapreduce.task.container.log.bac	mapreduce.reduce.skip.proc.count.autoincr
ld	kups	1 MILL (22)
mapreduce.job.jvm.numtasks	varn.app.mapreduce.am.container.log.back	mapreduce.job.skip.outdir
mapreduce.job.ubertask.enable	ups	mapreduce.map.skip.maxrecords
		mapreduce.reduce.skip.maxgroups
mapreduce.job.ubertask.maxmaps	yarn.app.mapreduce.shuffle.log.separate	
mapreduce.job.ubertask.maxreduces	yarn.app.mapreduce.shuffle.log.limit.kb	mapreduce.ifile.readahead
mapreduce.job.ubertask.maxbytes	yam.app.mapreduce.shuffle.log.backups	mapreduce.ifile.readahead.bytes
mapreduce.job.emit-timeline-data	mapreduce.job.userlog.retain.hours	mapreduce.jobtracker.taskcache.levels
mapreduce.input.fileinputformat.split.mins	mapreduce.jobtracker.hosts.filename	mapreduce.job.queuename
ize		mapreduce.job.tags
	mapreduce.jobtracker.hosts.exclude.filena	8 6 6
mapreduce.input.fileinputformat.list-	me	mapreduce.cluster.acls.enabled
status.num-threads	mapreduce.jobtracker.heartbeats.in.second	mapreduce.job.acl-modify-job
mapreduce.jobtracker.maxtasks.perjob	mapreduce.jobtracker.tasktracker.maxblac	mapreduce.job.acl-view-job
mapreduce.input.lineinputformat.linesperm	klists	mapreauce.job.aci-view-job
ap		mapreduce.tasktracker.indexcache.mb
	mapreduce.job.maxtaskfailures.per.tracker	mapreduce.job.token.tracking.ids.enabled
mapreduce.client.submit.file.replication	mapreduce.client.output.filter	
mapreduce.tasktracker.dns.interface	mapreduce.client.completion.pollinterval	mapreduce.job.token.tracking.ids
mapreduce.tasktracker.dns.nameserver		mapreduce.task.merge.progress.records
	mapreduce.client.progressmonitor.pollinter	mapreduce.task.combine.progress.records
mapreduce.tasktracker.http.threads	val	
mapreduce.tasktracker.http.address	mapreduce.jobtracker.persist.jobstatus.acti	mapreduce.job.reduce.slowstart.completed
mapreduce.task.files.preserve.failedtasks	ve	maps
maps consectable.mes.preserve.tarteufasks		mapreduce.job.complete.cancel.delegation.
		tokens

mapreduce.tasktracker.taskcontroller	yarn.app.mapreduce.am.job.client.port-	mapreduce.jobhistory.done-dir
mapreduce.tasktracker.group	range	mapreduce.jobhistory.cleaner.enable
mapreduce.shuffle.port	yarn.app.mapreduce.am.job.committer.can cel-timeout	mapreduce.jobhistory.cleaner.interval-ms
mapreduce.job.reduce.shuffle.consumer.pl	yarn.app.mapreduce.am.job.committer.co	mapreduce.jobhistory.max-age-ms
ugin.class	mmit-window	mapreduce.jobhistory.client.thread-count
mapreduce.tasktracker.healthchecker.script .path	mapreduce.fileoutputcommitter.algorithm.	mapreduce.jobhistory.datestring.cache.size
mapreduce.tasktracker.healthchecker.inter	version	mapreduce.jobhistory.joblist.cache.size
val	yarn.app.mapreduce.am.scheduler.heartbea Linterval-ms	mapreduce.jobhistory.loadedjobs.cache.siz
mapreduce.tasktracker.healthchecker.script .timeout	yarn.app.mapreduce.client-am.ipc.max-	mapreduce, jobhistory, move, interval-ms
	retries	
mapreduce.tasktracker.healthchecker.script .args	yarn.app.mapreduce.client-am.ipc.max-	mapreduce.jobhistory.move.thread-count
mapreduce.job.counters.limit	retries-on-timeouts	mapreduce.jobhistory.store.class
mapreduce.framework.name	yarn.app.mapreduce.client.max-retries	mapreduce.jobhistory.minicluster.fixed.por ts
yam.app.mapreduce.am.staging-dir	yarn.app.mapreduce.am.resource.mb	mapreduce.jobhistory.admin.address
mapreduce.am.max-attempts	yarn.app.mapreduce.am.resource.cpu-	
mapreduce.am.max-attempts	veores	mapreduce.jobhistory.admin.acl
mapreduce.job.end-notification.url	yarn.app.mapreduce.am.hard-kill-timeout-	mapreduce.jobhistory.recovery.enable
mapreduce.job.end-	ms	mapreduce.jobhistory.recovery.store.class
notification.retry.attempts	yarn.app.mapreduce.client.job.max-retries	mapreduce.jobhistory.recovery.store.fs.uri
mapreduce.job.end-	yarn.app.mapreduce.client.job.retry-	mapreduce.jobhistory.recovery.store.leveld
notification.retry.interval	interval	b.path
mapreduce.job.end-	mapreduce.application.classpath	mapreduce.jobhistory.http.policy
notification.max.attempts	mapreduce.app-submission.cross-platform	varn.app.mapreduce.am.containerlauncher.
mapreduce.job.log4j-properties-file	mapreduce.application.framework.path	threadpool-initial-size
mapreduce.job.end- notification.max.retry.interval	mapreduce.job.classloader	
yam.app.mapreduce.am.env	mapreduce.job.classloader.system.classes	
yam.app.mapreduce.am.admin.user.env	mapreduce.jobhistory.address	
	mapreduce.jobhistory.webapp.address	
yam.app.mapreduce.am.command-opts	mapreduce.jobhistory.keytab	
yarn.app.mapreduce.am.admin-command- opts		
	mapreduce.jobhistory.principal	
yarn.app.mapreduce.am.job.task.listener.th	mapreduce.jobhistory.principal mapreduce.jobhistory.intermediate-done-	

mapreduce.jobtracker.jobhistory.location	mapreduce.job.reduces			manneduce reduce speculative	manaduce output fileoutputformat comme	manroduco inhtrackar parcist inhetatus hou			ort-	mapreduce.jobhistory.done-dir
mapreduce.jobtracker.jobhistory.task.num	mapreduce.jobtracker.restart.recover									mapreduce.jobhistory.cleaner.enable
berprogresssplits	mapreduce.jobtracker.jobhistory.block.size		100	1	• • • • • • • • • • • • •				er.can	mapreduce.jobhistory.cleaner.interval-ms
mapreduce.job.userhistorylocation	mapreduce.jobtracker.taskscheduler	Over	IYU	CONT	Idurat	n noi	ara	meter	2	mapreduce.jobhistory.max-age-ms
mapreduce.jobtracker.jobhistory.complete d.location	mapreduce.job.running.map.limit				goiai				V • • • • • • • • • •	mapreduce.jobhistory.client.thread-count
mapreduce.job.committer.setup.cleanup.ne	mapreduce.job.running.reduce.limit					mapreduce.task.profile.params		.path	gorithm.	mapreduce.jobhistory.datestring.cache.size
eded	mapreduce.job.reducer.preempt.delay.sec	mapreduce.jobtracker.jobhistory.lru.cache. size		mapreduce.job.speculative.retry-after-no- speculate	mapreduce.map.output.compress.codec	ALC: STATE AND A		mapreduce.tasktracker.healthchecker.inter	version	mapreduce.jobhistory.joblist.cache.size
mapreduce.task.io.sort.factor	mapreduce.job.max.split.locations	mapreduce.jobtracker.instrumentation		mapreduce.job.speculative.retry-after-	map.sort.class	mapreduce.task.profile.map.params mapreduce.task.profile.reduce.params		val	yarn.app.mapreduce.am.scheduler.heartbea Linterval-ms	mapreduce.jobhistory.loadedjobs.cache.siz
mapreduce.task.io.sort.mb	mapreduce.job.split.metainfo.maxsize	mapred.child.java.opts		speculate	mapreduce.task.userlog.limit.kb			mapreduce.tasktracker.healthchecker.script		e
mapreduce.map.sort.spill.percent	mapreduce.jobtracker.taskscheduler.maxru	mapred.child.env		mapreduce.job.map.output.collector.class	yarn.app.mapreduce.am.container.log.limit	mapreduce.task.skip.start.attempts		.timeout	yarn.app.mapreduce.client-am.ipc.max- retries	mapreduce.jobhistory.move.interval-ms
mapreduce.jobtracker.address	nningtasks.perjob	mapreduce.admin.user.env		mapreduce.job.speculative.slowtaskthresho	.kb	mapreduce.map.skip.proc.count.autoincr		mapreduce.tasktracker.healthchecker.script	varn.app.mapreduce.client-am.ipc.max-	mapreduce.jobhistory.move.thread-count
mapreduce.local.clientfactory.class.name	mapreduce.map.maxattempts	mapreduce.map.log.level		ld	yarn.app.mapreduce.task.container.log.bac kups	mapreduce.reduce.skip.proc.count.autoincr		.args	retries-on-timeouts	mapreduce.jobhistory.store.class
mapreduce.jobtracker.http.address	mapreduce.reduce.maxattempts	mapreduce.reduce.log.level		mapreduce.job.jvm.numtasks	yarn.app.mapreduce.am.container.log.back	mapreduce.job.skip.outdir		mapreduce.job.counters.limit	yarn.app.mapreduce.client.max-retries	mapreduce.jobhistory.minicluster.fixed.por
mapreduce.jobtracker.handler.count	mapreduce.reduce.shuffle.fetch.retry.enabl ed	mapreduce.nap.cpu.vcores		mapreduce.job.ubertask.enable	ups	mapreduce.map.skip.maxrecords		mapreduce.framework.name	yarn.app.mapreduce.am.resource.mb	ts
mapreduce.tasktracker.report.address				mapreduce.job.ubertask.maxmaps	yam.app.mapreduce.shuffle.log.separate	mapreduce.reduce.skip.maxgroups		yarn.app.mapreduce.am.staging-dir	yarn.app.mapreduce.am.resource.cpu-	mapreduce.jobhistory.admin.address
mapreduce.cluster.local.dir	mapreduce.reduce.shuffle.fetch.retry.interv al-ms	mapreduce.reduce.cpu.vcores		mapreduce.job.ubertask.maxreduces	yarn.app.mapreduce.shuffle.log.limit.kb	mapreduce.ifile.readahead		mapreduce.am.max-attempts	veores	mapreduce.jobhistory.admin.acl
mapreduce.jobtracker.system.dir	mapreduce.reduce.shuffle.fetch.retry.timeo	mapreduce.reduce.merge.inmem.threshold		mapreduce.job.ubertask.maxbytes	yarn.app.mapreduce.shuffle.log.backups	mapreduce.ifile.readahead.bytes		mapreduce.job.end-notification.url	yarn.app.mapreduce.am.hard-kill-timeout-	mapreduce.jobhistory.recovery.enable
mapreduce.jobtracker.staging.root.dir	ut-ms	mapreduce.reduce.shuffle.merge.percent		mapreduce.job.emit-timeline-data	mapreduce.job.userlog.retain.hours	mapreduce.jobtracker.taskcache.levels		mapreduce.job.end-	ms	mapreduce.jobhistory.recovery.store.class
mapreduce.cluster.temp.dir	mapreduce.reduce.shuffle.retry-	mapreduce.reduce.shuffle.input.buffer.perc ent		mapreduce.input.fileinputformat.split.mins	mapreduce.jobtracker.hosts.filename	mapreduce.job.queuename		notification.retry.attempts	yarn.app.mapreduce.client.job.max-retries	mapreduce.jobhistory.recovery.store.fs.uri
mapreduce.tasktracker.local.dir.minspacest	delay.max.ms	mapreduce.reduce.input.buffer.percent		ize	mapreduce.jobtracker.hosts.exclude.filena	mapreduce.job.tags		mapreduce.job.end-	yarn.app.mapreduce.client.job.retry- interval	mapreduce.jobhistory.recovery.store.leveld
art	mapreduce.reduce.shuffle.parallelcopies	mapreduce.reduce.shuffle.memory.limit.pe		mapreduce.input.fileinputformat.list-	me	mapreduce.cluster.acls.enabled		notification.retry.interval		b.path
mapreduce.tasktracker.local.dir.minspacek	mapreduce.reduce.shuffle.connect.timeout	rcent		status.num-threads	mapreduce.jobtracker.heartbeats.in.second	mapreduce.job.acl-modify-job		mapreduce.job.end- notification.max.attempts	mapreduce.application.classpath	mapreduce.jobhistory.http.policy
ill	mapreduce.reduce.shuffle.read.timeout	mapreduce.shuffle.ssl.enabled		mapreduce.jobtracker.maxtasks.perjob	mapreduce.jobtracker.tasktracker.maxblac	mapreduce.job.acl-view-job		mapreduce.job.log4j-properties-file	mapreduce.app-submission.cross-platform	yarn.app.mapreduce.am.containerlauncher.
mapreduce.jobtracker.expire.trackers.inter	mapreduce.shuffle.connection-keep- alive.enable	mapreduce.shuffle.ssl.file.buffer.size		mapreduce.input.lineinputformat.linesperm ap	klists	mapreduce.tasktracker.indexcache.mb		mapreduce.job.end-	mapreduce.application.framework.path	threadpool-initial-size
maproduce.tasktracker.instrumentation	mapreduce.shuffle.connection-keep-	mapreduce.shuffle.max.connections		mapreduce, client, submit, file, replication	mapreduce.job.maxtaskfailures.per.tracker	mapreduce.job.token.tracking.ids.enabled		notification.max.retry.interval	mapreduce.job.classloader	
mapreduce.tasktracker.resourcecalculatorp	alive.timeout	mapreduce.shuffle.max.threads		mapreduce.tasktracker.dns.interface	mapreduce.client.output.filter	mapreduce.job.token.tracking.ids		yam.app.mapreduce.am.env	mapreduce.job.classloader.system.classes	
mapreduce.tasktracker.resourcecalculatorp lugin	mapreduce.task.timeout	mapreduce.shuffle.transferTo.allowed		mapreduce.tasktracker.dns.nameserver	mapreduce.client.completion.pollinterval	mapreduce.task.merge.progress.records		yam.app.mapreduce.am.admin.user.env	mapreduce.jobhistory.address	
mapreduce.tasktracker.taskmemorymanage	mapreduce.tasktracker.map.tasks.maximu	mapreduce.shuffle.transfer.buffer.size			mapreduce.client.progressmonitor.pollinter val	mapreduce.task.combine.progress.records		yam.app.mapreduce.am.command-opts	mapreduce.jobhistory.webapp.address	
r.monitoringinterval	m	mapreduce.reduce.markreset.buffer.percen		mapreduce.tasktracker.http.threads		mapreduce.job.reduce.slowstart.completed		yam.app.mapreduce.am.admin-command-	mapreduce.jobhistory.keytab	
mapreduce.tasktracker.tasks.sleeptimebefo	mapreduce.tasktracker.reduce.tasks.maxim	t		mapreduce.tasktracker.http.address	mapreduce.jobtracker.persist.jobstatus.acti ve	maps		opts	mapreduce.jobhistory.principal	
resigkill	um	mapreduce.map.speculative		mapreduce.task.files.preserve.failedtasks		mapreduce.job.complete.cancel.delegation.		yam.app.mapreduce.am.job.task.listener.th	mapreduce.jobhistory.intermediate-done-	
mapreduce.job.maps	mapreduce.map.memory.mb					tokens		read-count	dir	

mapreduce.jobtracker.jobhistory.location	mapreduce.job.reduces		manzaduce raduce enculative	manaduce output fileoutputformat comme	manreduce inhtracker narrist inhetatus hou			mapreduce.jobhistory.done-dir
mapreduce, jobtracker, jobhistory, task.num	mapreduce.job.reduces mapreduce.jobtracker.restart.recover						con-	
berprogresssplits	mapreduce.jobtracker.jobhistory.block.size			•			er.can	mapreduce.jobhistory.cleaner.enable mapreduce.jobhistory.cleaner.interval-ms
mapreduce.job.userhistorylocation	mapreduce.jobtracker.taskscheduler	()Ver	190 conf	Idurdi	rion na	Irametei	rc I	mapreduce.jobhistory.max-age-ms
mapreduce.jobtracker.jobhistory.complete d.location	mapreduce.job.running.map.limit			iguia			19• • • 1	mapreduce.jobhistory.client.thread-count
mapreduce, job committer, setup, cleanup, ne edd mapreduce, task, io, sort, flictor mapreduce, task, io, sort, mb mapreduce, task, io, sort, mb mapreduce, task, io, sort, mb mapreduce, task, io, sort, mb mapreduce, jobtracker, address mapreduce, jobtracker, handler, count mapreduce, jobtracker, handler, count mapreduce, tasktracker, report, address mapreduce, cluster, local, dir	mapreduce, job running, reduce. Jimi mapreduce, job reducer preempt delay see mapreduce, job split, metainfo. maxize mapreduce, job split, metainfo. maxize mapreduce, job tracker, taskscheduler, maxru mapreduce, metaic, taskscheduler, maxru mapreduce, metaic, taskscheduler, maxru mapreduce, metaic, taskscheduler, maxru mapreduce, metaic, taskscheduler, maxru ed mapreduce, reduce, taskscheduler, maxru ed mapreduce, reduce, taskscheduler, maxru ed	mapreduce jobtracker jobbistory.htu cache. size mapreduce jobtracker instrumentation mapred.child.env mapreduce.admin.user.env mapreduce.admin.user.env mapreduce.admin.user.env	mapreduce. job. speculative.netry-after-no- speculate mapreduce. job. speculative.netry-after- speculate mapreduce. job. map.output.collector.class mapreduce. job.speculative.skowtuskthresho id mapreduce. job.jvm.numtasks	mapreduce.map.output.compress.codee map.sort.class mapreduce.task.userlog.limit.kb yarn.app.mapreduce am.container.log.limit kb yarn.app.mapreduce task.container.log.bae kups	mapreduce task profile params mapreduce task profile nup params mapreduce task profile reduce params mapreduce task profile reduce params mapreduce may skip proc count autoiner mapreduce reduce skip proc count autoiner mapreduce job skip outdir	-path mapredace taskracker healthchecker inter val mapredace taskracker healthchecker script .inncost mapreduce taskracker healthchecker script .args mapreduce job.counters.limit	version yara ago magreduce.client-um.ipc.max- retries yara ago magreduce.client-um.ipc.max- retries yara ago magreduce.client-um.ipc.max- retries-on-timoouts yara ago magreduce.client.max-retries	mapreduce jobhistory datestring.eache size mapreduce jobhistory jobhit cache size mapreduce jobhistory loadeljobs.cache.siz e mapreduce jobhistory move interval-ms mapreduce jobhistory move interval-ms mapreduce jobhistory move fraed-count mapreduce jobhistory store class mapreduce jobhistory store class
mapreduce.jobtracker.staging.root.dir	mapreduce.reduce. ut-ms	ск тпе	e expertis	se to i	iune sy	stem int	ernals	uce.jobhistory.admin.acl uce.jobhistory.recovery.enable uce.jobhistory.recovery.store.class
	mapreduce.reduce.sit				une sy			ace jobhistory.recovery.store.class
mapreduce.jobtracker.staging.root.dir	mapreduce.reduce.sh delay.max.ms	CK TNC	нара садее простистврополны зравлица ize	SE TO I Inspective jour acce acoust measure Imageeduce jobracker Josts, exclude filena	IUNE SY	STEM INT	yarn.app.mapreduce.client.job.retry- interval	mee.jobhistory.recovery.store.class
mapreduce.jobtracker.staging.root.dir mapreduce.cluster.temp.dir mapreduce.tasktracker.local.dir.minspacest art	mapreduce.reduce.sit	ent mapreduce.reduce.input.buffer.percent mapreduce.reduce.shuffle.memory.limit.pe	наргенке.праг.пеаритотна.spit.nins	mapreduce jostracker.nosts.mename mapreduce jobtracker.hosts.exclude.filena me	mapreduce.job.tags mapreduce.cluster.acls.enabled	mapredace, job.end- notification.retry.interval mapredace, job.end-	yarn.app.mapreduce.client.job.retry-	uce.jobhistory.recovery.store.class
mapreduce.jobtracker.staging.root.dir mapreduce.eluster.temp.dir	mapreduce.reduce.st delay.max.ms mapreduce.reduce.shuffle.parallelcopies	ent mapreduce.reduce.input.buffer.percent	mapressee, input, incompationmat, Spin, nims ize mapreduce, input, fileinputformat, list-	mapreduce.joon.acker.nosts.mename mapreduce.jobtracker.hosts.exclude.filena	mapreduce job.tags	mapredace, job.end- notification.retry.interval mapredace, job.end- notification.max.attempts	yarn.app.mapreduce.client.job.retry- interval	mapreduce jobhistory.recovery.store.leveld b.path
mapreduce.jobtracker.staging.root.dir mapreduce.cluster.temp.dir mapreduce.tasktracker.local.dir.minspacest art	mapreduce reduce and delay.max.ms mapreduce.reduce.shuffle.parallelcopies mapreduce.reduce.shuffle.connect.timeout	est maproduce robace input buffer percent maproduce zobace.shuffle.memory.limit.pe reent maproduce shuffle.sal.enabled maproduce.shuffle.sal.file.buffer.size	impressee input interspectrostant spin muss ize mapredace, input fileinputformat.list- status.num-threads	Inspective geomacker answermentite mapreduce jobtracker hosts exclude filtena me mapreduce jobtracker heartbeats in second	mapreduce.job.tags mapreduce.cluster.acis.embiled mapreduce.job.aci-modify-job mapreduce.job.aci-view-job mapreduce.taskracker.indexesche.mb	mapredace, job.end- notification.refry.interval mapredace, job.end- notification.max.attempts mapredace, job.logiJ.geoperties-file mapredace, job.logiJ.geoperties-file	yarn.app.mapreduce.client.job.retry- interval mapreduce.application.classpath	nce jobhistory recovery store class opticalize jobhistory recovery store. Is uri mapreduce jobhistory covery store leveld b path mapreduce jobhistory http policy
mopreduce, jobtracker, staging, root, dir mopreduce, cluster, temp, dir mopreduce, tasktracker Jocal dir, minopacest art mopreduce, tasktracker Jocal dir, minopacek il mopreduce, jobtracker, espire, trackers, inter val mopreduce, tasktracker, instrumentation	mapreduce reduces delay, max.ms mapreduce reduce.shuffle.parallelcopies mapreduce.reduce.shuffle.connect.timeout mapreduce.reduce.shuffle.read.timeout mapreduce.shuffle.connection-keep-	ent mapreduce.reduce.input.buffer percent mapreduce.reduce.shuffle.memory.limit.pe reent mapreduce.shuffle.anl.enabled	inspectace input intemperatures spacinities ize mapreduce input fileinputformat.list- status num-threads mapreduce.jobtracker.maxtasks.perjob mapreduce.input lineinputformat, linesperm ap mapreduce.client.aubmit.file.replication	impreniece postracter mine meanie mapreniece jobtracker hosts, exclude, filena me mapreniece jobtracker heartbeats, in second mapreniece jobtracker, tasktracker, maxblac kliets	mapreduce job.tags mapreduce cluster.acls.embled mapreduce job.acl-modify-job mapreduce job.acl-view-job	mapredace, job.end- notification.retry.interval mapredace, job.end- notification.max.attempts mapredace.job.logij.groperties-file	yarn app mapreduce.elient job retry- interval mapreduce application.elasspath mapreduce app-submission.cross-platform mapreduce application.framework.path	nee jobbistory recovery store, class mercluce jobbistory recovery store. Is ari mapreduce jobbistory recovery store. Jeveld b path mapreduce jobbistory. http policy yarn app. mapreduce am. containerlauncher.
mapreduce, jobtracker, staging, root, dir mapreduce, cluster, temp, dir mapreduce, tasktracker, local, dir, minspacest art mapreduce, tasktracker, local, dir, minspacek ill mapreduce, jobtracker, espire, trackers, inter val	umm methods roduces delay max ms maproduce roduce shuffle parallelcopies maproduce roduce shuffle connect timoout maproduce and the connection-keep- alive emble maproduce shuffle connection-keep-	est mapreduce reduce input buffer percent mapreduce reduce shuffle memory limit pe reent mapreduce shuffle ast enabled mapreduce shuffle ast. file buffer size mapreduce shuffle max connections	impreduce: input fileinputformat.iss- ize mapreduce.input fileinputformat.iss- status:num-threads mapreduce.input.inciputformat.incsperm ap mapreduce.client.submit.file.replication mapreduce.tuktricker.dms.interface	impreduce jobracker notice menne mapreduce jobracker hosts exclude filena me mapreduce jobracker heartbeats in second mapreduce jobracker tasktracker, maxblac kliets mapreduce jobracker skifalures per tracker	mapreduce.job.tags mapreduce.chaster.ach.embled mapreduce.job.acl-modify-job mapreduce.job.acl-sicw-job mapreduce.indextracker.indexcasche.mb mapreduce.job.token.tracking.ids.embled	mapreduce_job.end- notification.refy.interval mapreduce_job.end- notification.max.ttempts mapreduce_job.logij.groperties-file mapreduce_job.end- notification.max.retry.interval	yarn app mapreduce.elient job retry- interval mapreduce application.classpath mapreduce app-submission.cross-platform mapreduce.application.framework.path mapreduce.application.framework.path	nce jobhistory recovery store class apreduce jobhistory necovery store. Is ari mapreduce jobhistory necovery store. Ieveld b path mapreduce jobhistory.http.policy yarn app.mapreduce.am.containerlauncher.
mapreduce jobtracker staging.root.dir mapreduce cluster temp dir mapreduce tasktracker local.dir minspacest art mapreduce tasktracker local.dir minspacest ill mapreduce isobtracker expire trackers inter val mapreduce tasktracker instrumentation mapreduce tasktracker nesourcealculutop	umm delay.max.ms mapreduce.reduce.shuffle.paralleleopies mapreduce.reduce.shuffle.connect.timeout mapreduce.thuffle.connection-keep- alive.enable mapreduce.shuffle.connection-keep- alive.timeout	est mapreduce reduce input buffer percent mapreduce reduce shuffle anemory limit pe reent mapreduce shuffle ast enabled mapreduce shuffle ast, clubed mapreduce shuffle max connections mapreduce shuffle max threads	inspectace input intemperatures spacinities ize mapreduce input fileinputformat.list- status num-threads mapreduce.jobtracker.maxtasks.perjob mapreduce.input lineinputformat, linesperm ap mapreduce.client.aubmit.file.replication	Impreduce jobracker mitter mennte mapreduce jobracker hosts exclude filena me mapreduce jobracker heartbeats in second mapreduce jobracker tasktracker matblac klists mapreduce jobrackatskfallures, per tracker mapreduce client output.filter	mapreduce.job.tags mapreduce.chster.acls.emabled mapreduce.job.acl-wicw-job mapreduce.job.acl-vicw-job mapreduce.taktracker.indexcsche.mb mapreduce.job.token.tracking.ids.emabled mapreduce.job.token.tracking.ids	mapreduce job.end- notification.retry.interval mapreduce job.end- notification.msa.attempts mapreduce.job.logij.properties-file mapreduce job.end- notification.msa.retry.interval yam.app.mapreduce.am.env	yarn.app.mapreduce.elient.job.retry- interval mapreduce.application.classpath mapreduce.application.cross-platform mapreduce.application.framework.path mapreduce.application.framework.path mapreduce.job.classloader mapreduce.job.classloader.system.classes	nee jobhistory recovery store class aproduce jobhistory recovery store level b path mapreduce jobhistory.http.policy yarn app.mapreduce am.containerlauncher

mapreduce jobtracker jobhistory, location mapreduce jobtracker jobhistory task.num berprogresssplits mapreduce job aserhistory location mapreduce jobtracker jobhistory, complete d.location	mapreduce.job.reduces mapreduce.jobtracker.restart.recover mapreduce.jobtracker.jobiistory.block.size mapreduce.jobtracker.tuskocheduler mapreduce.job.running.reduce.limit mapreduce.job.running.reduce.limit	Over 1	90 conf	igurat	tion pai	ramete	rs	mapreduce jobhistory.done-dir mapreduce jobhistory.cleaner.enable mapreduce jobhistory.cleaner.interval-ms mapreduce jobhistory.max-age-ms mapreduce jobhistory.altent.thread-count
marerduce job committer setup, cleanup ne eded mapreduce task, io. sort, factor mapreduce, task, io. sort, mb mapreduce, jobtracker, address mapreduce, jobtracker, address	mapreduce job reducer preempt.delay.sec mapreduce job.split.netainfo.maxsize mapreduce job.split.netainfo.maxsize mapreduce_jobtracker.tuskscheduler.maxru mingtasks_perjob mapreduce.map.maxattempts mapreduce.netuce.maxattempts	mapreduce.jobtracker.jobtinistry.lm.cache. size mapreduce.jobtracker.instrumentation mapred.child.java.opts mapreduchild.env mapreduce.map.log.level mapreduce.map.log.level mapreduce.map.log.level	mapreduce, job speculative, zetry-after-no- speculate mapreduce, job speculative, zetry-after- speculate mapreduce, job speculative, advortadatheresho id mapreduce job speculative, advortadatheresho id	mapreduce: map.output.compress.codec map.sort.class mapreduce: task.userlog.limit.kb yarn.app.mapreduce: ann container.log.limit kb yarn.app.mapreduce.task.container.log.bac kups yarn.app.mapreduce.ann container.log.back	maproduce tank profile params maproduce tank profile map params maproduce tank profile reduce params maproduce tank akip start attempts maproduce tank akip start attempts maproduce map akip proc. count autoiner maproduce reduce skip, proc. count autoiner maproduce job skip outdir	path mapreduce tasktracker.healthchecker.inter val mapreduce tasktracker.healthchecker.script .inneout mapreduce tasktracker.healthchecker.script .args mapreduce.job.counters.limit	version yarn app, mapreduce am scheduler heartbea timterval-ms yarn app, mapreduce client-am ipc max- retries yarn app, mapreduce client-am ipc max- retries-on-timeosts yarn app, mapreduce client max-retries	mapreduce jobhistory datestiring, each asi mapreduce jobhistory Jobhic cache size mapreduce jobhistory Joadedjobs, cache si e mapreduce jobhistory.move.interval-ms mapreduce jobhistory.move.thread-count mapreduce jobhistory.store.class mapreduce jobhistory.store.class
mapreduce.jobtracker.handler.count	mapreduce.reduce.sh							
mapreduce, cluster, local dir mapreduce, jobtracker, system, dir mapreduce, jobtracker, staging, root, dir mapreduce, cluster, temp, dir	maproduce reduce al-ms maproduce reduce ut-ms maproduce reduce as delay max ms	ent	expertis	паря спосе јора за кет лозах, полкане		mapreduce.job.end-	yarn.app.mapreduce.client.job.retry-	uce.jobhistory.recovery.store.class
mapreduce.cluster.local.dir mapreduce.jobtracker.system.dir mapreduce.jobtracker.staging.root.dir	al-ms mapreduce.reduce. ut-ms mapreduce.reduce.st	ext maproduce reduce, layut buffer, percent maproduce reduce, dauffe, memory, limit, per rent maproduce sharffe, sal, enabled maproduce sharffe, sal, faile buffer, site maproduce sharffe, sal, faile buffer, site maproduce sharffe max, connections maproduce sharffe max, threads	maprence.mpar.memputtorman.spin.mins	see to to the second se	-			uee.jobhistory.admin.acl uee.jobhistory.recovery.enable uee.jobhistory.recovery.store.class

Introduction







Data-lifecycleDo more than just queries, optimize the movement, storage, and processingAwarenessof big data

Elasticity Dynamically adjust resource usage and operational costs based on workload and user requirements

Robustness Provide storage and querying services even in the event of some failures





Enable Hadoop users and applications to get good performance automatically throughout

the data lifecycle in analytics; without any need on their part to understand and manipulate

the many tuning knobs available.

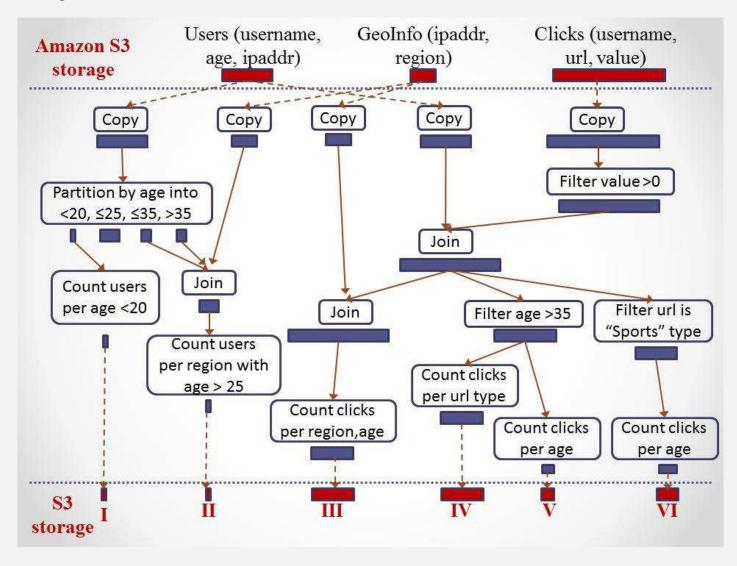
Purpose of this paper

Using experimental results to illustrate the challenges in each component and to motivate

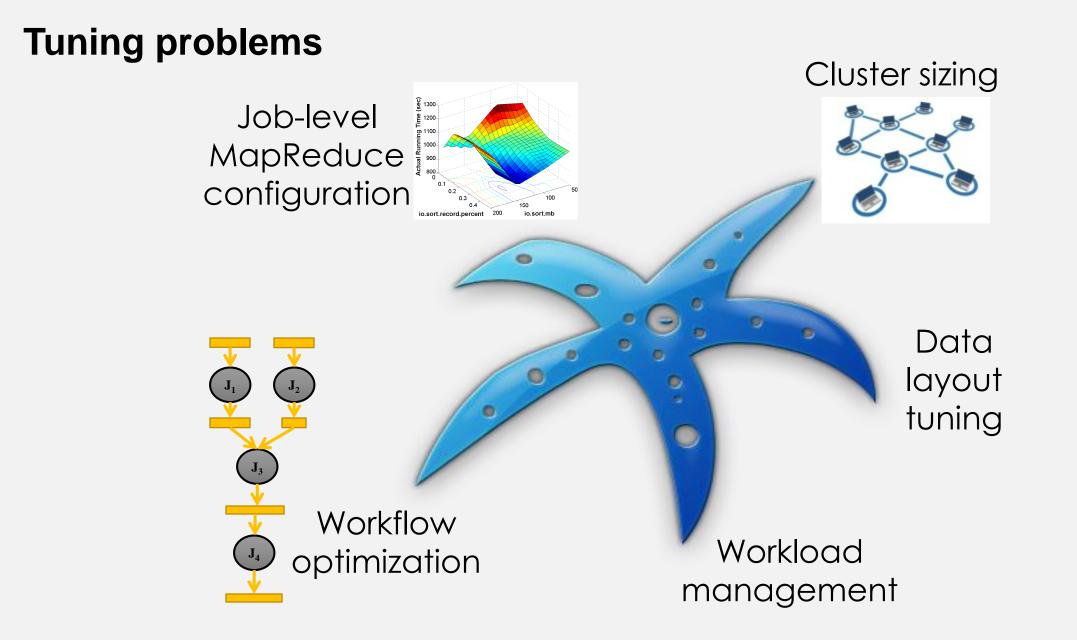
Starfish's solution approach.



Example analytics workload



Overview



Core Approaches to Tuning

Profiler

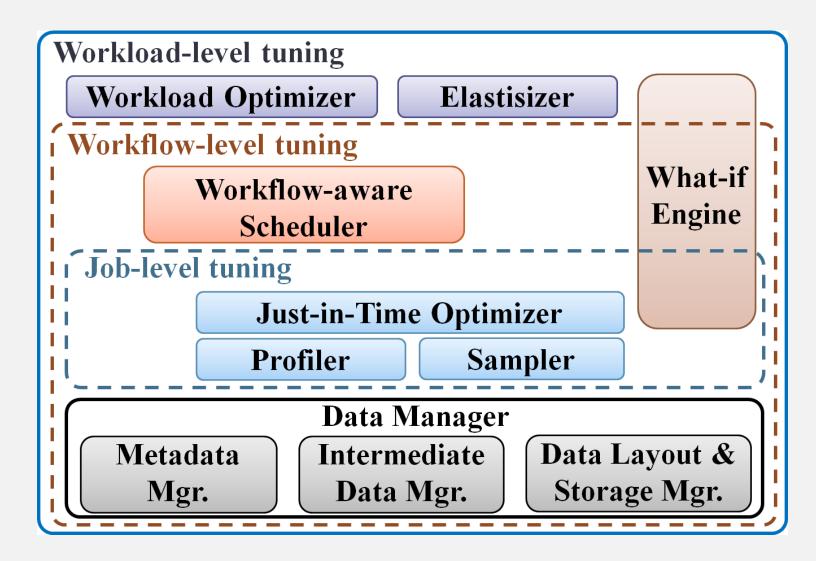
Collects concise summaries of execution

What-if Engine

Estimates impact of hypothetical changes on execution

Optimizers Search through space of tuning choices

Architecture



Basic idea of each level

Job-level	The Just-in-Time Optimizer optimizes parameters based on the
Tuning	results of the Profiler , the Sampler , and the What-if Engine .

Workflow-levelWorkflow-aware Scheduleraddresses concerns, such asTuningunbalanced data layout, in conjunction with the What-if Engine and
the Data Manager.

Workload-levelThe Workload Optimizer generates optimized workflows with theTuninghelp of the Elastisizer and the What-if Engine.

LASTWORD:

Language for Starfish Workloads and Data

Clients submit WORKLOADS expressed in Lastword. It is not a human-interactable language.

Starfish provides language translators to convert workloads specified in high level languages

to Lastword.



Three Components



Just-in-TimeAutomatically selects efficient execution techniques forOptimizerMapReduce jobs.

Profiler Collects detailed summaries of jobs on a task-by-task basis.

SamplerCollects statistics about input, intermediate, and output data of a
MapReduce job.

Job-level Tuning

What Controls the Execution of a MR Job?

• Number of map tasks

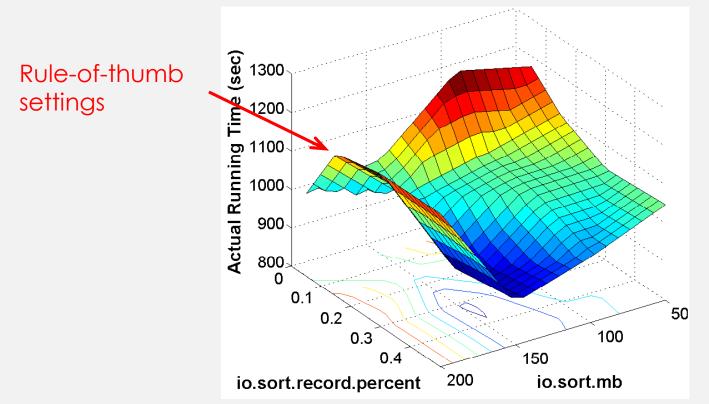
٠

. . .

- Number of reduce tasks
- Whether combine function should be used
- Partitioning of map outputs to reduce tasks
- Memory allocation to task-level buffers
- Whether output data from tasks should be compressed

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

What's Wrong with the Rule-of-thumb Settings?



Response surface of a Word Co-occurrence

MapReduce program

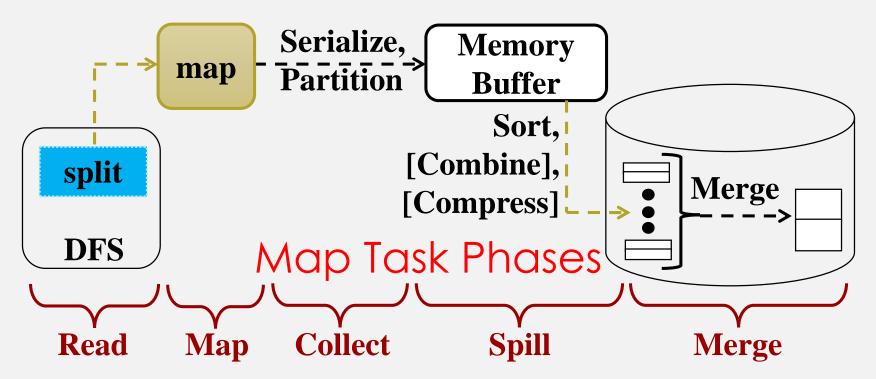
Dataset: 10 GB Wikipedia

MapReduce Job Tuning in a Nutshell

- **Goal**: $perf = F(p, d, r, c), c_{opt} = \arg\min_{c \in S} F(p, d, r, c)$
- **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 <p,d1,r1,c1>, estimates virtual profile for <p,d2,r2,c2>
- 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-level Tuning

Job Profile Fields

Dataflow: amount of data flowing through task phases

- Map output bytes
- Number of spills
- Number of records in buffer per spill
- ...

Dataflow Statistics: statistical information about dataflow

- Width of input key-value pairs
- Map selectivity in terms of records
- Map output compression ratio
- • •

2

4

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
- ...

3

Cost Statistics: statistical information about resource costs

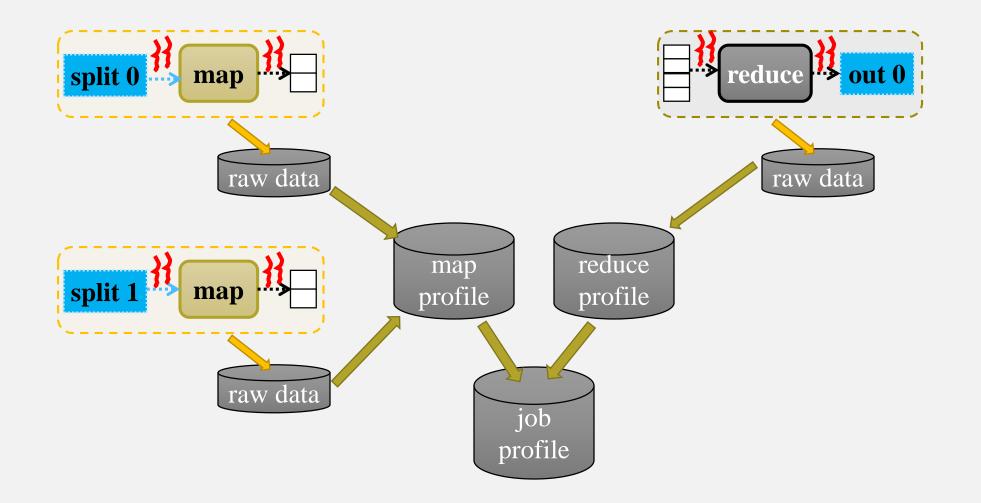
- I/O cost for reading from local disk per byte
- CPU cost for executing the Mapper per record
- ..

Job-level Tuning

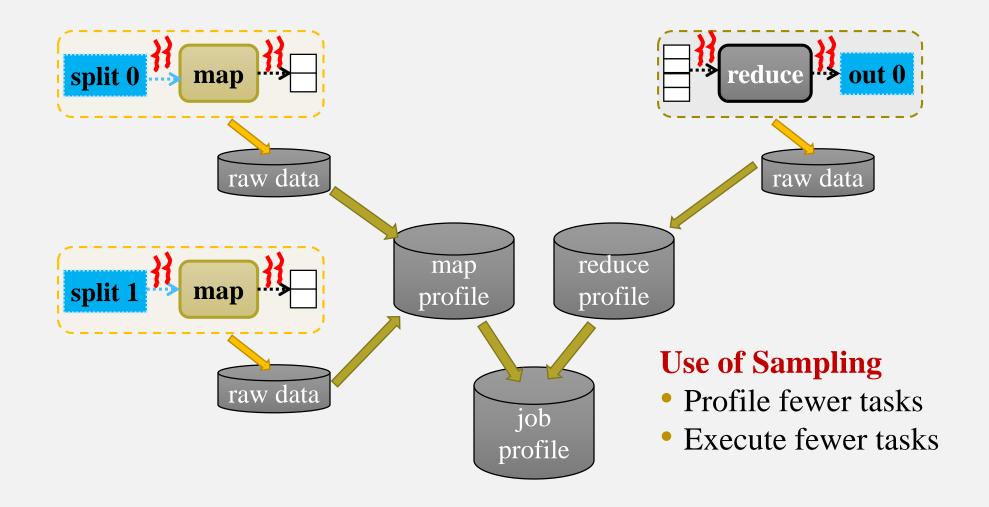
Profiling Using Dynamic Instrumentation

- **Dynamic Instrumentation** collects run-time monitoring information from unmodified MapReduce programs running on Hadoop.
- **BTrace** is the current implementation of the Profiler.
 - A safe, dynamic tracing tool for the Java platform
- Three views are exposed to capture the features of a job's execution:
 - **Timings view**: The amount of time spent in each sub-phases
 - Data-flow view: The amount of data processed during each sub-phases
 - Resource-level view: The usage trends of CPU, memory, I/O, and network resources
 during each sub-phases

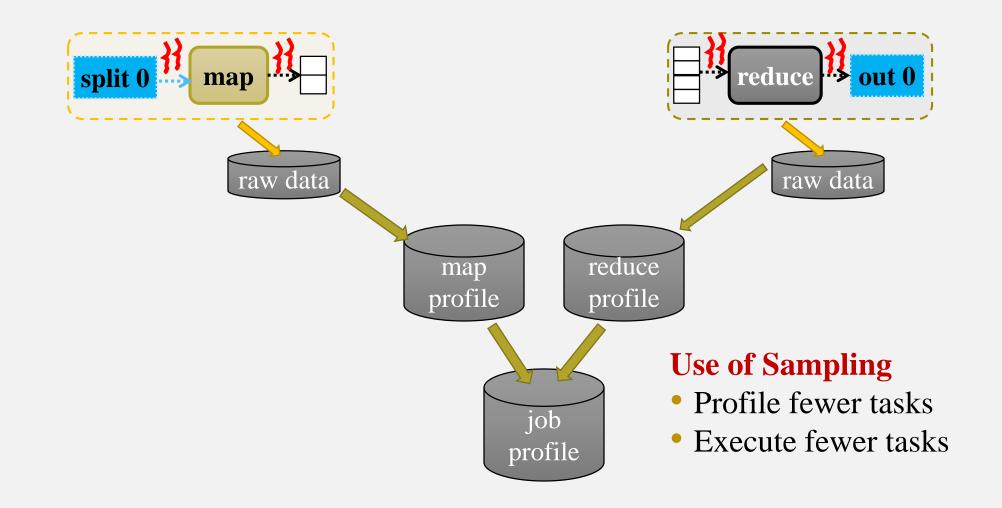
Generating Profiles



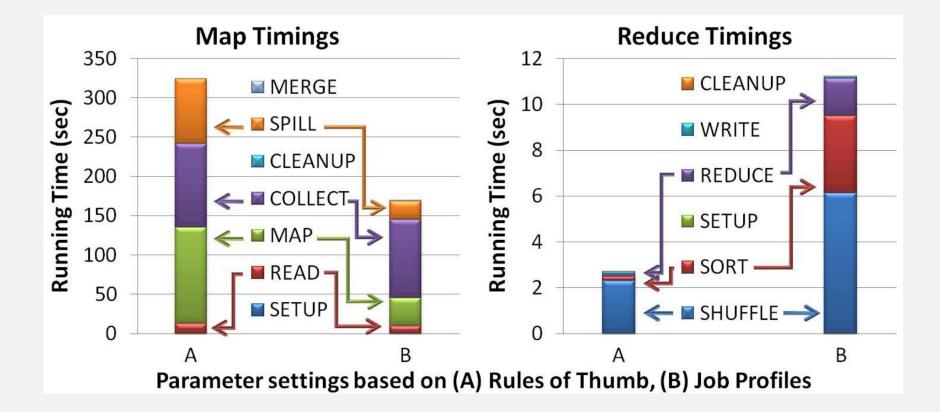
Generating Profiles



Generating Profiles



Example of a Timings View

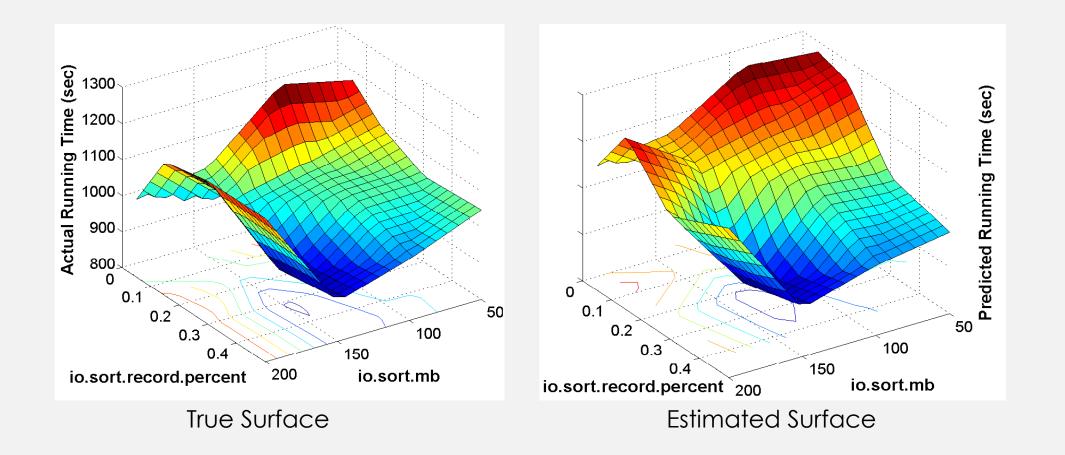


Predicting Job Performance in Hadoop

The What-if Engine

- Uses the job profile and a set of performance models to estimate the new profile if the job were to be run using S
- Four inputs are required:
 - The **job profile** $\langle p, d_1, r_1, c_1 \rangle$ generated for this job J by the Profiler
 - The **new setting** c_2 of the job configuration parameters using which Job J will be run
 - The size, layout, and compression information d₂ of the input dataset on which Job J will be run
 - The cluster setup and resource allocation r_2 that will be used to run Job J
- Virtual job profile <p, d₂, r₂, c₂>: Contains the predicted Timings and Data-flow views of the job when run with the new parameter settings

Estimate Result of the What-if Engine



Experiment Results

	WordCount		TeraSort	
	Rules of	Based on	Rules of	Based on
	Thumb	Job Profile	Thumb	Job Profile
io.sort.spill.percent	0.80	0.80	0.80	0.80
io.sort.record.percent	0.50	0.05	0.15	0.15
io.sort.mb	200	50	200	200
io.sort.factor	10	10	10	100
mapred.reduce.tasks	27	2	27	400
Running Time (sec)	785	407	891	606

- Single-rack Hadoop cluster running on 16 Amazon EC2 nodes of the c1.medium type.
- Each node runs at most 3 map tasks and 2 reduce tasks concurrently.
- WordCount processes 30GB of data generated using the RandomTextWriter program in Hadoop.
- TeraSort processes 50GB of data generated using Hadoop's TeraGen program.

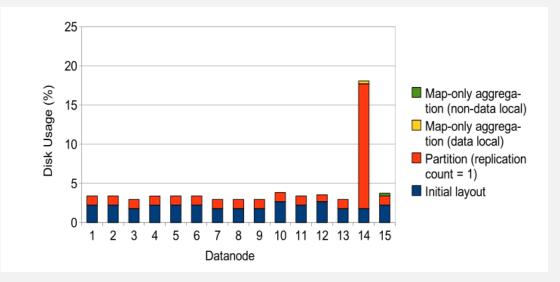


Skewed Data

- Data-Layerout-unaware Tasks Scheduling
- Addition or dropping of nodes Without data rebalancing

An Example Easy to Understand

Consequence of one execution of a large job



Main Problem: Unbalanced Data Layouts in Hadoop

A Slightly More Complicated Example Showing that

Unbalanced Data Layout is Not Only A Coincidence

Tasks are usually scheduled

where data is stored.



HDFS always write the first replica of any block on the same node the writer runs. (Overutilize)

Unbalanced data layouts are serious problems in big data analytics because they are prominent causes of tasks failure and performance degradation.

Dilemma faced:

Exploiting data locality pros:

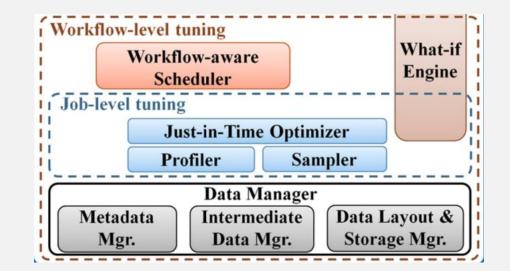
1. Less disk I/O. Save shuffling time.

Exploiting data locality cons (in this context):

- 1. Performance degradation due to reduced parallelism.
- 2. Making data layout further unbalanced.

Need a tuning method to achieve automatic optimization.

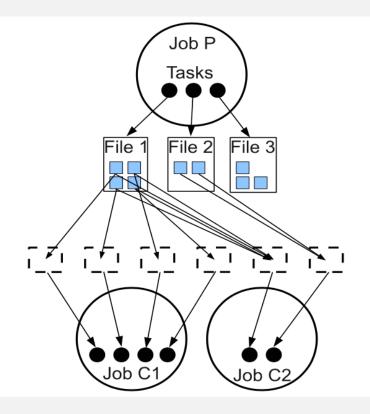
A Workflow-aware Scheduler can ensure that job-level optimization and scheduling policies are **coordinated tightly** with the policies for data placement employed by the underlying distributed filesystem.



Rather than making decisions that are locally optimal for individual MapReduce jobs, Starfish's Workflow-aware Scheduler makes decisions by considering producer-consumer relationships among jobs in the workflow. Some information this tuning needs:

1. What parts of the data output by a job are used by downstream jobs in the workflow?

In this example, both file 1 and file 2 will be used by downstream jobs



2. What is the unit of data-level parallelism in each job that reads the data output by a job?

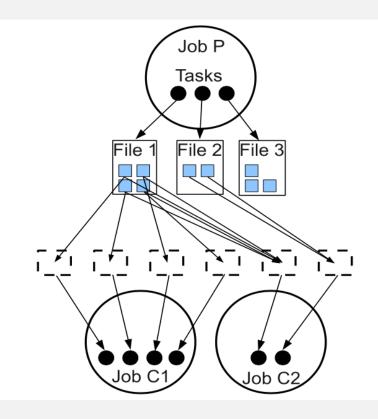
Data-parallel reader tasks of Job C1 read and process one data block each. One file each for Job C2.

Workflow-aware Tuning

. . .

Some information this tuning needs (cont'd):

- 3. What is the expected running time of Job P if the Round Robin block placement policy is used for P 's output files?
- 4. What will the new data layout in the cluster be if the Round Robin block placement policy is used for P 's output files?
- 5. What is the expected running time of Job C1 (C2) if its input data layout is the one in the answer to Question 4?

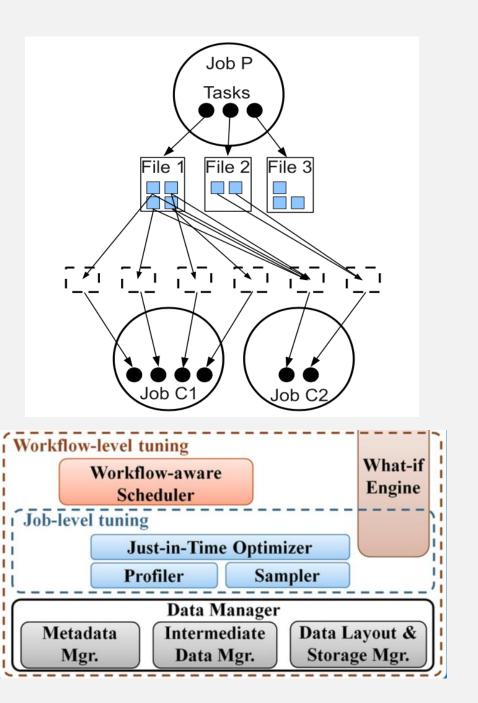


Some information this tuning needs (cont'd):

Questions in the previous slide are answered by the What-if Engine based on a simulation of the main aspects of workflow execution.

Action Involves Simulating:

- MapReduce job execution
- Task scheduling
- HDFS block placement policies



This system answers these questions:

- What block placement policy to use in the distributed filesystem for the output file of a job? Implemented a Round Robin block placement policy in HDFS.
- How many replicas to store for the blocks of a file?

Replication helps tasks including heavily-accessed files

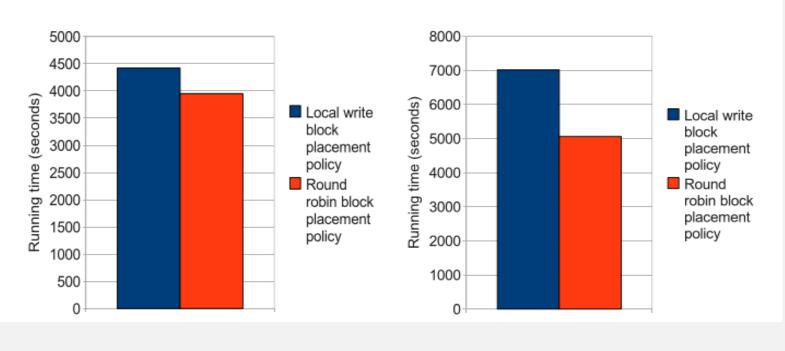
- What size to use for blocks of a file?
- Should a job's output files be compressed for storage?

Transfer additional CPU cost for saving local disks I/O and network cost.

Experiment Result

Parameters:

- Replication Factor: 1
- Block Size: 128MB
- Compression: None



A partition job

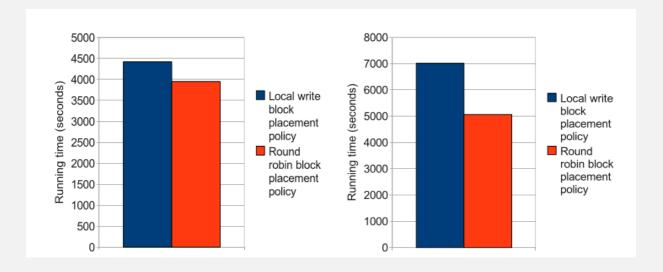
A two-job workflow

Reason for the winning of Round Robin

The local I/O within a node becomes the bottleneck before the parallel writes of data blocks to other storage nodes over the network.

PLUS

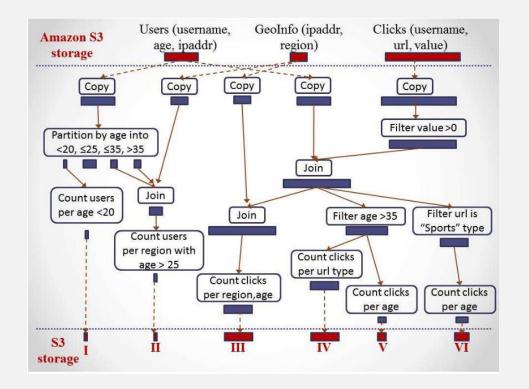
The Round Robin policy spreads the blocks over the cluster so that maximum data-level parallelism of sort processing can be achieved while performing data-local computation.





Workload Optimizer

This optimizer uses the What-if Engine to do a cost-based estimation of whether the graph-to-graph transformation will improve performance.

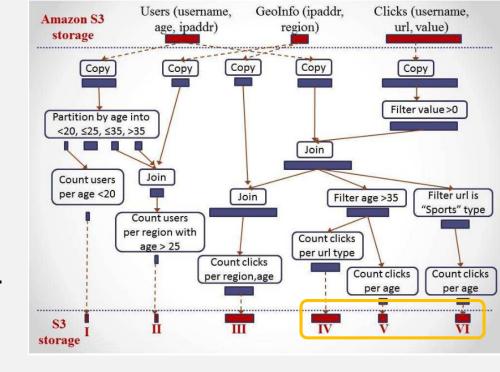


Workload-level Tuning

Workload Optimizer

Introducing Jumbo Operator

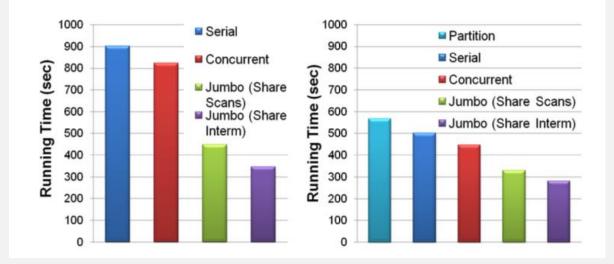
The results IV, V and VI can be represented as a Select-Project-Aggregation (SPA) expression over the join. Jumbo operator can process any number of logical SPA Nodes over the same table in a single job.



The Jumbo operator enables sharing of all or some of the map-side scan and computation, sorting and shuffling, as well as the reduce-side scan, computation, and output generation. At the same time, the Jumbo operator can help the scheduler to better utilize the bounded number of map and reduce task slots in a Hadoop cluster.

Workload Optimizer

If possible, Running a MapReduce job to partition Users based certain attribute will enable the four workflows to prune out irrelevant partitions efficiently.



Processing multiple SPA workflow nodes on the same input dataset

Generating the partitions has significant. But possibilities exist to hide or reduce this overhead by combining partitioning with a previous job like data copying. Partition pruning improves the performance of all MapReduce jobs in our experiment. At the same time, partition pruning decreases the performance benefits provided by the Jumbo operator.

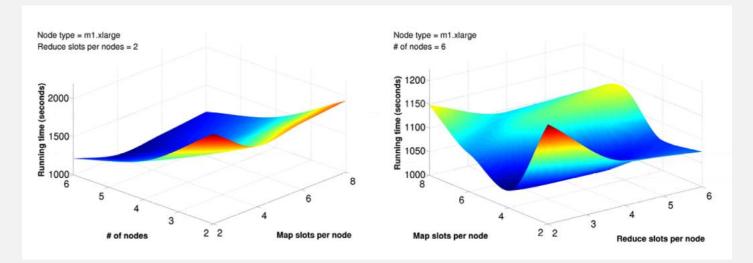
Workload-level Tuning

Elastisizer

Introducing an especially made optimizer for Amazon Web Service.

It automatically free users from the burden of setting the number and type of EC2 nodes.

These parameters are Hadoop-level configuration parameters.

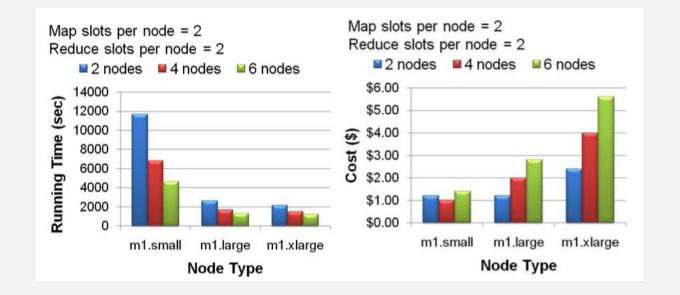


Workload performance under various cluster and Hadoop configurations on Amazon Elastic MapReduce

Workload-level Tuning

Elastisizer

The user could have multiple preferences and constraints for the workload, which poses a multi-objective optimization problem. For example, the goal may be to minimize the monetary cost incurred to run the workload, subject to a maximum tolerable workload completion time.



Performance Vs. costs for a workload on Amazon Elastic MapReduce



Critique of this paper

Good:

- Able to See the impact of various settings
- Good Visualization Tools

Bad:

- This paper talks a lot about what starfish can do but not how it can achieve that
- Experiments constrained many variants (possibly related to each other) fixed.



THANK YOU FOR WATCHING

