ArgoCon

Streamlining Big Data Workflows with Memoization & Work Avoidance

Alan Clucas - Pipekit | Julie Vogelman - Intuit

pipekit

INTUIT turbotax creditkarma quickbooks mailchimp

#ArgoCon

November 6, 2023

INTROS

pipekit

Alan Clucas

Senior Software Engineer @ Pipekit.io

Argo Workflows maintainer

INTUIT turbotax creditkarma mailchimp quickbooks

Julie Vogelman

Staff Software Engineer @ Intuit

Argo Workflows maintainer

GOALS

Saving time

Saving Cost

HOW

Don’t repeat work.

If we know the answer already, use that...

WHAT THINGS

Individual steps/ tasks between workflows

Templates

(DAG, steps)

Terminology

Inputs and outputs

Parameters and artifacts

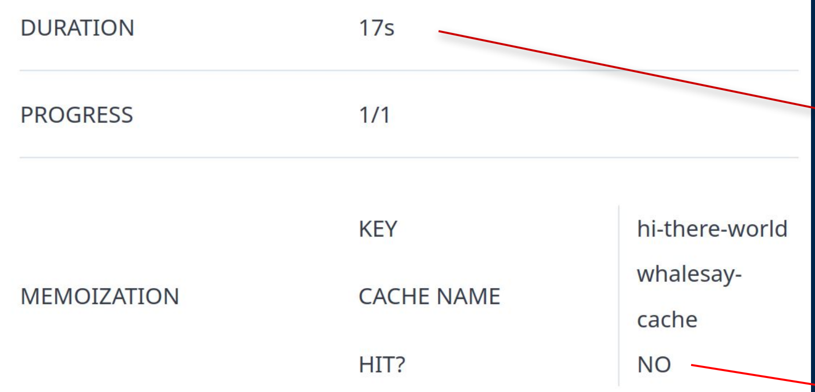
Memoization and Work Avoidance

Introducing memoization

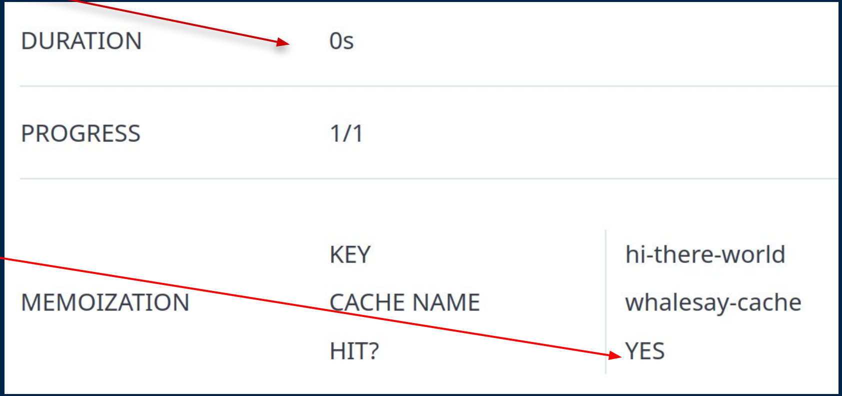
```
apiVersion: argoproj.io/v1alpha1
kind: Workflow
metadata:
   generateName: memoized-workflow-
spec:
   entrypoint: whalesay
   templates:
    - name: whalesay
      memoize:
         key: "{{inputs.parameters.message}}"
         maxAge: "10s"
         cache:
             configMap:
                 name: whalesay-cache
```

What happens?

- Argo looks up keys in a cache
- If they match…
  - Remembers the outputs and provides them instead of running the step

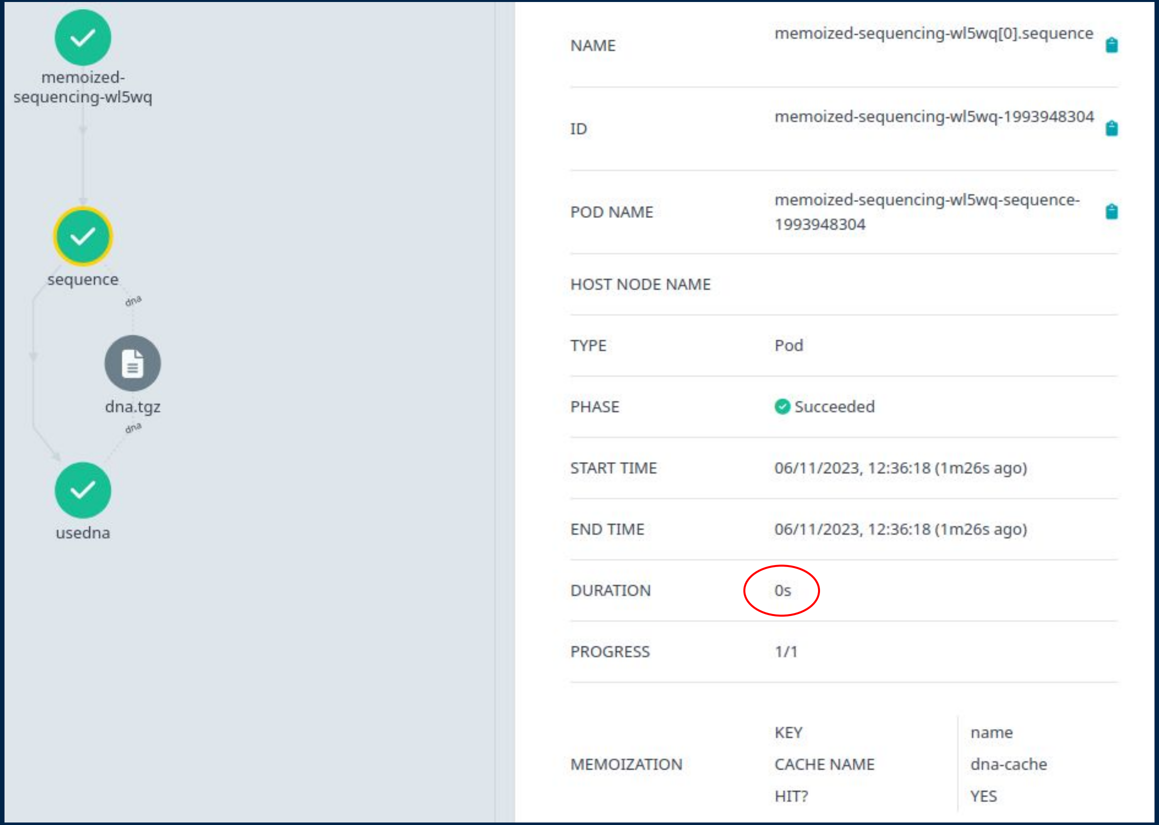


DURATION	17s	
PROGRESS	1/1	
MEMOIZATION	KEY	hi-there-world
	CACHE NAME	whalesay-cache
	HIT?	NO



DURATION	0s	
PROGRESS	1/1	
MEMOIZATION	KEY	hi-there-world
	CACHE NAME	whalesay-cache
	HIT?	YES

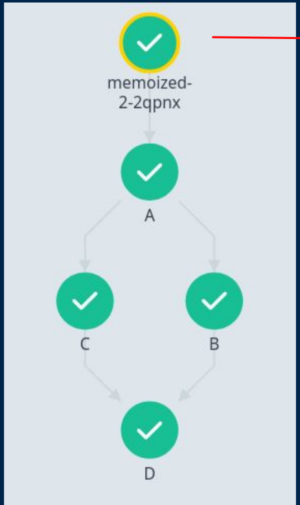
This also works for output artifacts

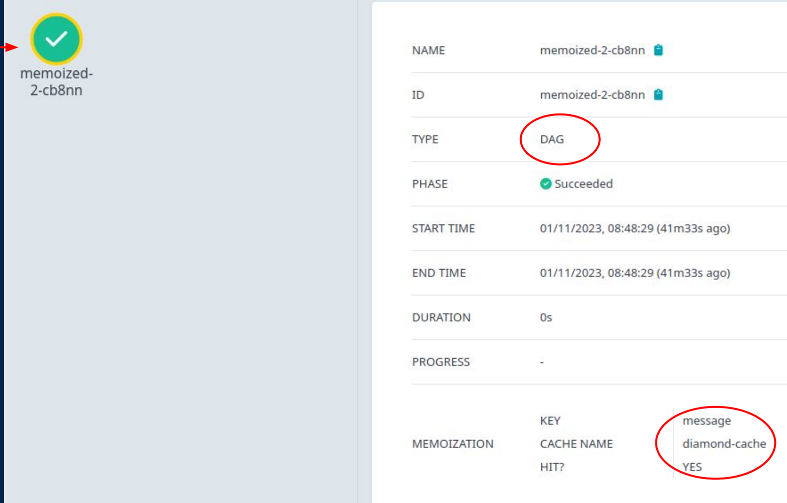


NAME	memoized-sequencing-wl5wq[0].sequence
ID	memoized-sequencing-wl5wq-1993948304
POD NAME	memoized-sequencing-wl5wq-sequence-1993948304
HOST NODE NAME	
TYPE	Pod
PHASE	✓ Succeeded
START TIME	06/11/2023, 12:36:18 (1m26s ago)
END TIME	06/11/2023, 12:36:18 (1m26s ago)
DURATION	0s
PROGRESS	1/1
MEMOIZATION	KEY: name CACHE NAME: dna-cache HIT?: YES

You can memoize whole templates

Workflow skipping an entire DAG template:





NAME	memoized-2-cb8nn	
ID	memoized-2-cb8nn	
TYPE	DAG	
PHASE	✓ Succeeded	
START TIME	01/11/2023, 08:48:29 (41m33s ago)	
END TIME	01/11/2023, 08:48:29 (41m33s ago)	
DURATION	0s	
PROGRESS	-	
MEMOIZATION	KEY	message
	CACHE NAME	diamond-cache
	HIT?	YES

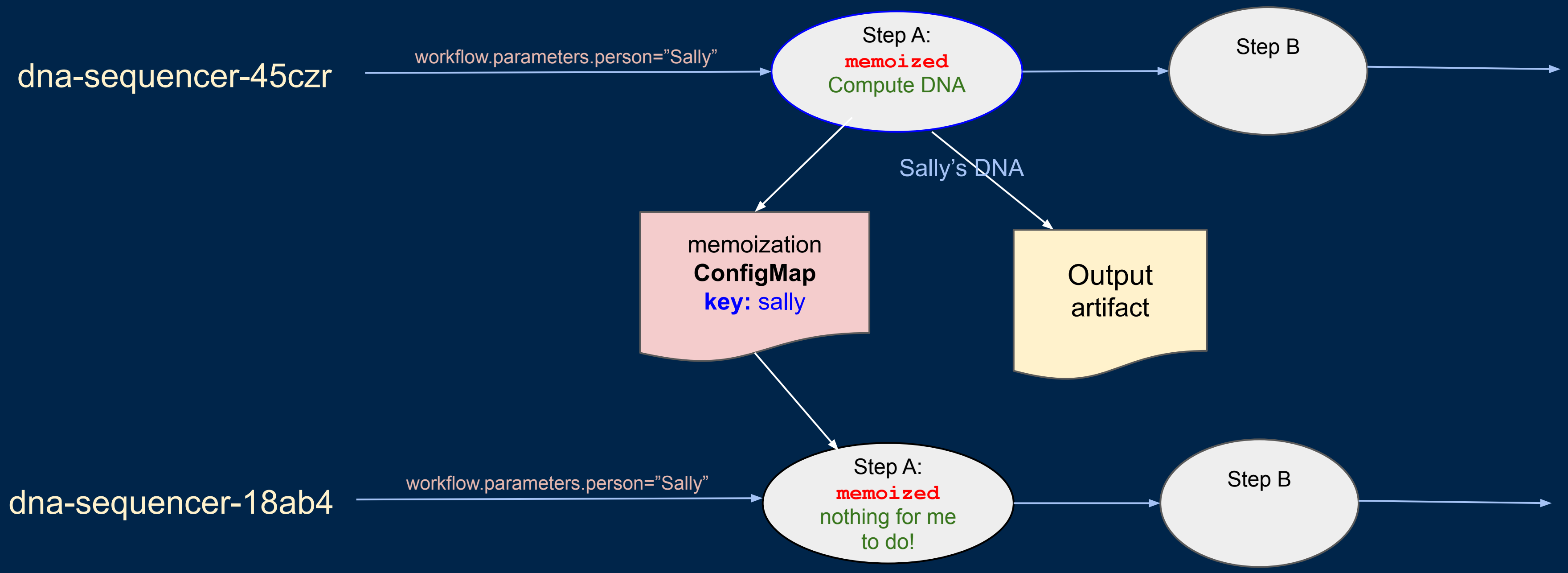
ConfigMap

Created for you by the controller

Usually limited to 1MiB

Contains ‘human readable’ JSON

How memoization works in Argo Workflows



dna-sequencer-45czr

workflow.parameters.person="Sally"

Step A:
`memoized`
Compute DNA

Step B

Sally's DNA

memoization
ConfigMap
key: sally

Output artifact

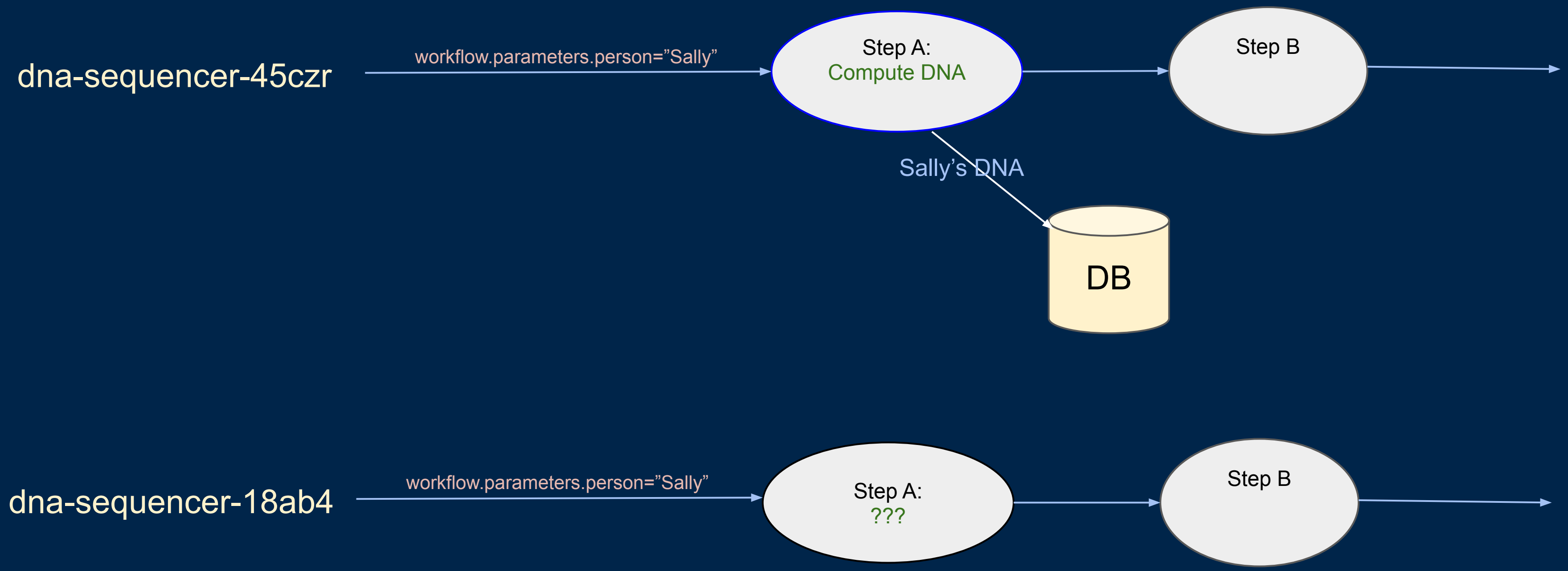
dna-sequencer-18ab4

workflow.parameters.person="Sally"

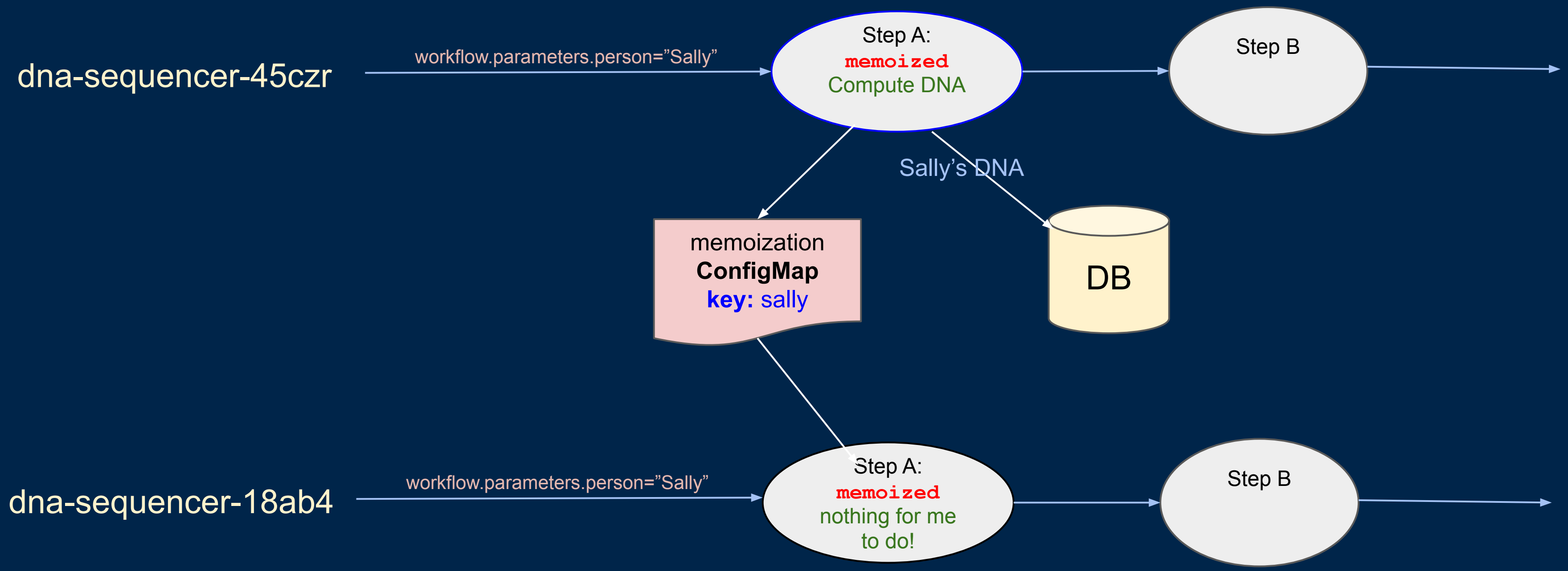
Step A:
`memoized`
nothing for me to do!

Step B

What if my step outputs to somewhere other than an artifact or parameter?



What if my step outputs to somewhere other than an artifact or parameter?



What if my step outputs to somewhere other than an artifact or parameter?

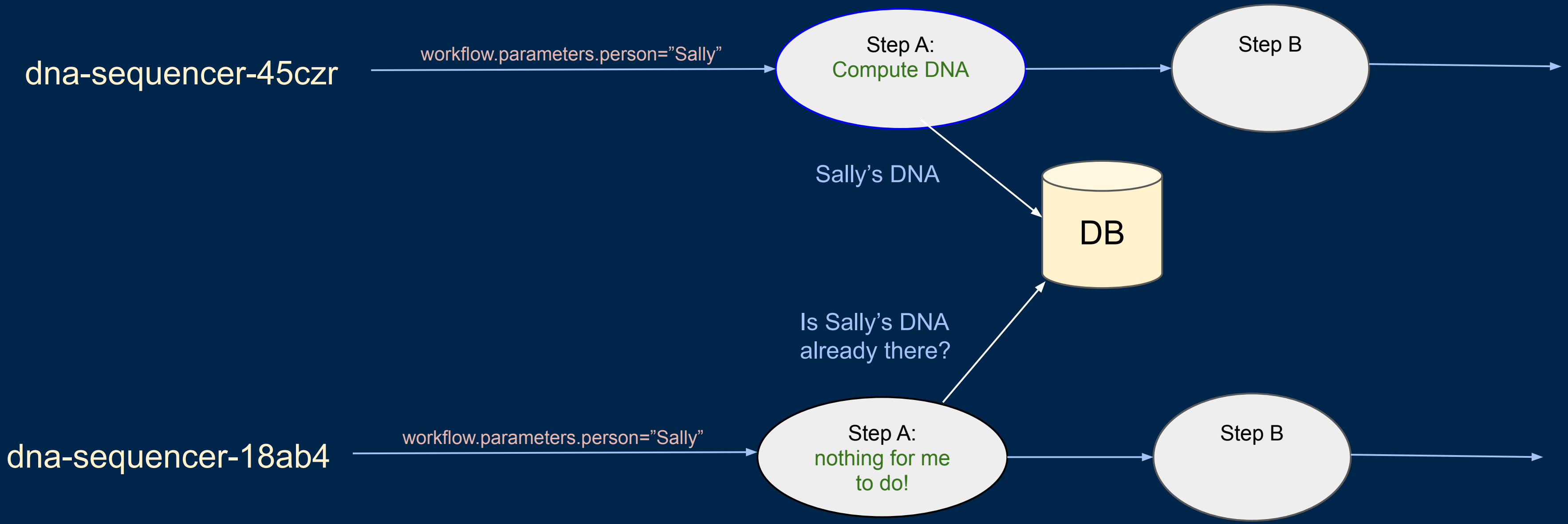
Memoization works for this in latest

Alternative: use a technique called “Work Avoidance” instead

Avoiding work sounds pretty good 😅 …but what is it?

Pod logic checks to see if the data exists before recomputing

Or can use a separate marker file to indicate data written (in PVC, DB, etc) and check for that



Any other time I can’t use memoization?

If your data’s too big for a ConfigMap (> 1 MiB)

Use “Work Avoidance” instead

Caching

There are only two hard things in Computer Science: cache invalidation and naming things.

-- Phil Karlton

Guaranteed safe

Pure steps

- Inputs - all as key
  - Outputs are derived only from inputs
- Has no external interactions
  - Nothing is read from the outside world
  - Changes nothing about the outside world, no side effects

Otherwise - choose keys wisely

Will skip step when key(s) match

And time since storing < maxAge

The End

Come have a chat with Alan about memoization this week.

@Alan Clucas on CNCF Slack

@Joibel on GitHub

Further reading:

https://argoproj.github.io/argo-workflows/memoization/

https://argoproj.github.io/argo-workflows/work-avoidance/

https://s.pipekit.io/chat-argo-wf