A high throughput cloud computation architecture for ‘deep’ parsing

Alexandre Rademaker and Henrique Muniz

IBM Research, Brazil

July 15, 2019

What is it?

It is our first steps on the use of emerging technologies for distributed cloud computing for building scalable and high-performance architecture for ‘deep’ parsing with DELPHI-IN tools.

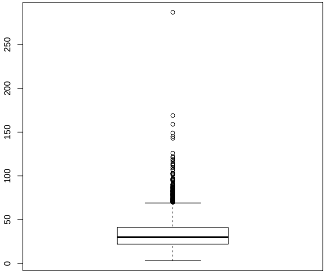
Our goals

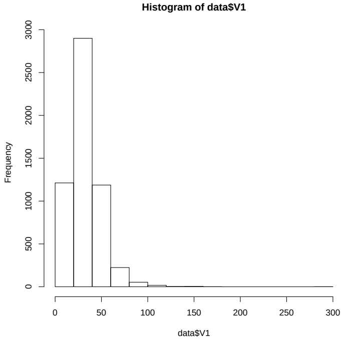
- a high throughput architecture
- as simple as possible
- scalable, pay what you need
- flexibility

The available options

- LOGON's batch parsing script (pvm library last release 2005)
- Heart of Gold middleware
- ACE and Arbiter

Data for experiments





A corpus with 5602 sentences obtained from 155 text passages relevant to petroleum systems extracted from documents randomly selected from a corpus of 1298 publicly available **English** language geological reports, published by the United States Geological Survey (USGS), Geological Survey of Canada (GSC), and British Geological Survey (BGS).

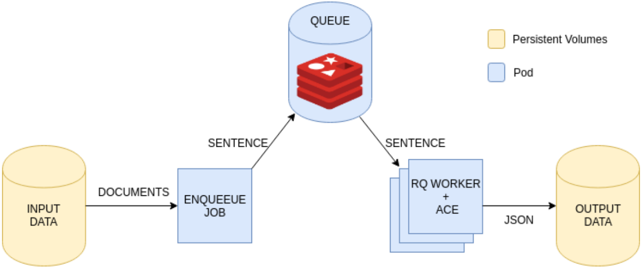
Cluster 1

One node with 40 cores and 126 GB RAM.

LOGON parsing script

- `-count 4` takes 8 hours 30 min
- `-count 8` takes 6 hours 30 min
- `-count 10` takes 4 hours 30 min
- `-count 20` takes 4 hours 40 minutes

Cloud Architecture



Libraries and tools: ACE, PyDephin, Kubernets and Docker, Python RQ (Redis Queues).

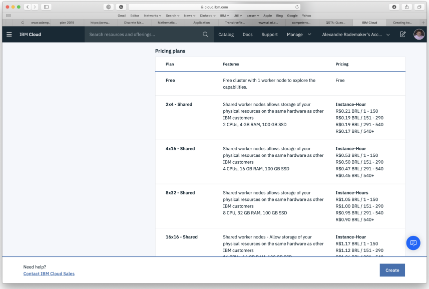
. . . but we are lisp programmers! ;-)

https://github.com/own-pt/k8s-delphin-parsing

IBM Cloud Kubernets Cluster Service

IBM Cloud Kubernets Cluster (RIS): 15 workers, 56 cores, 242 GB RAM.

2273/5602 sentences (many results PyDelphin could not read?!) in 3 hours using 8 workers.



Pricing plans

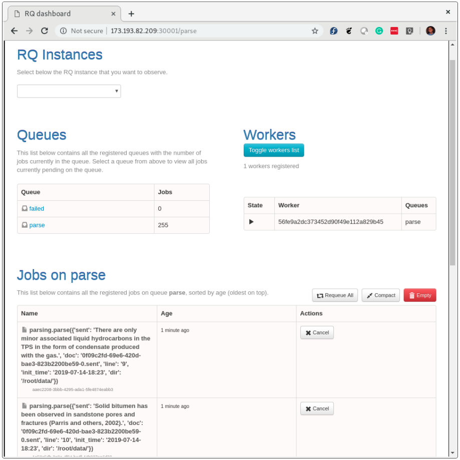
Plan	Features	Pricing
Free	Free cluster with 1 worker node to explore the capabilities.	Free
2x4 - Shared	Shared worker nodes allows storage of your physical resources on the same hardware as other IBM customers 2 CPUs, 4 GB RAM, 100 GB SSD	**Instance-Hour** R$0.21 BRL / 1 - 150 R$0.19 BRL / 151 - 290 R$0.19 BRL / 291 - 540 R$0.17 BRL / 540+
4x16 - Shared	Shared worker nodes allows storage of your physical resources on the same hardware as other IBM customers 4 CPUs, 16 GB RAM, 100 GB SSD	**Instance-Hour** R$0.53 BRL / 1 - 150 R$0.50 BRL / 151 - 290 R$0.47 BRL / 291 - 540 R$0.45 BRL / 540+
8x32 - Shared	Shared worker nodes allows storage of your physical resources on the same hardware as other IBM customers 8 CPU, 32 GB RAM, 100 GB SSD	**Instance-Hours** R$1.05 BRL / 1 - 150 R$1.00 BRL / 151 - 290 R$0.95 BRL / 291 - 540 R$0.90 BRL / 540+
16x16 - Shared	Shared worker nodes - Allow storage of your physical resources on the same hardware as other IBM customers	**Instance-Hour** R$1.17 BRL / 1 - 150 R$1.12 BRL / 151 - 290

Need help?
Contact IBM Cloud Sales

IBM Cloud Kubernets Cluster Service

IBM Cloud Kubernets Cluster (RIS): 15 workers, 56 cores, 242 GB RAM.

2273/5602 sentences (many results PyDelphin could not read?!) in 3 hours using 8 workers.



RQ Instances

Select below the RQ instance that you want to observe.

Queues

This list below contains all the registered queues with the number of jobs currently in the queue. Select a queue from above to view all jobs currently pending on the queue.

Queue	Jobs
failed	0
parse	255

Workers

1 workers registered

State	Worker	Queues
▶	56fe9a2dc373452d90f49e112a829b45	parse

Jobs on parse

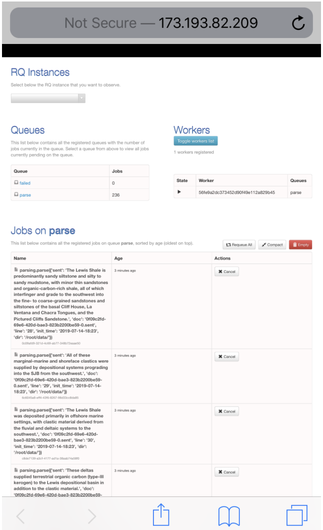
This list below contains all the registered jobs on queue **parse**, sorted by age (oldest on top).

Name	Age	Actions
parsing.parse({'sent': 'There are only minor associated liquid hydrocarbons in the TPS in the form of condensate produced with the gas.', 'doc': '0f09c2fd-69e6-420d-bae3-823b2200be59-0.sent', 'line': '9', 'init_time': '2019-07-14-18:23', 'dir': '/root/data/'}) aaec2208-3bbb-4295-ada1-5fe4874eabb3	1 minute ago	✖ Cancel
parsing.parse({'sent': 'Solid bitumen has been observed in sandstone pores and fractures (Parris and others, 2002).', 'doc': '0f09c2fd-69e6-420d-bae3-823b2200be59-0.sent', 'line': '10', 'init_time': '2019-07-14-18:23', 'dir': '/root/data/'})	1 minute ago	✖ Cancel

IBM Cloud Kubernets Cluster Service

IBM Cloud Kubernets Cluster (RIS): 15 workers, 56 cores, 242 GB RAM.

2273/5602 sentences (many results PyDelphin could not read?!) in 3 hours using 8 workers.



Not Secure — 173.193.82.209

RQ Instances

Select below the RQ instance that you want to observe.

Queues

This list below contains all the registered queues with the number of jobs currently in the queue. Select a queue from above to view all jobs currently pending on the queue.

Queue	Jobs
failed	0
parse	236

Workers

Toggle workers list

1 workers registered

State	Worker	Queues
▶	56fe9a2dc373452d90f49e112a829b45	parse

Jobs on **parse**

This list below contains all the registered jobs on queue **parse**, sorted by age (oldest on top).

Requeue All Compact Empty

Name	Age	Actions
parsing.parse({'sent': 'The Lewis Shale is predominantly sandy siltstone and silty to sandy mudstone, with minor thin sandstones and organic-carbon-rich shale, all of which interfinger and grade to the southwest into the fine- to coarse-grained sandstones and siltstones of the basal Cliff House, La Ventana and Chacra Tongues, and the Pictured Cliffs Sandstone.', 'doc': '0f09c2fd-69e6-420d-bae3-823b2200be59-0.sent', 'line': '28', 'init_time': '2019-07-14-18:23', 'dir': '/root/data/'}) 0c09afd9-321d-4c69-ab77-348b72eaae50	3 minutes ago	✖ Cancel
parsing.parse({'sent': 'All of these marginal-marine and shoreface clastics were supplied by depositional systems prograding into the SJB from the southwest.', 'doc': '0f09c2fd-69e6-420d-bae3-823b2200be59-0.sent', 'line': '29', 'init_time': '2019-07-14-18:23', 'dir': '/root/data/'}) 6c6045a8-eff4-43f6-8267-98d33cc8da85	3 minutes ago	✖ Cancel
parsing.parse({'sent': 'The Lewis Shale was deposited primarily in offshore marine settings, with clastic material derived from the fluvial and deltaic systems to the southwest.', 'doc': '0f09c2fd-69e6-420d-bae3-823b2200be59-0.sent', 'line': '30', 'init_time': '2019-07-14-18:23', 'dir': '/root/data/'}) c8de7139-a3cf-4177-ad1a-38aab74a58f0	3 minutes ago	✖ Cancel
parsing.parse({'sent': 'These deltas supplied terrestrial organic carbon (type-III kerogen) to the Lewis depositional basin in addition to the clastic material.', 'doc': '0f09c2fd-69e6-420d-bae3-823b2200be59-	3 minutes ago	✖ Cancel

Future work

- It should run in any cloud environment that supports Kubernets! But we need to try more cloud environments.
- Standard protocols from DELPH-IN (i.e. ErgApi?)
- Part of an internal text processing pipeline for IE of scientific articles from the O&G domain. More experiments.