

1. Em relação ao Apache Beam, defina os conceitos abaixo:

- (a) Windowing
- (b) Watermark
- (c) PTransform

2. Ainda no contexto do Apache Beam, e dada a Figura 1, onde *worker 1* processa os dados de entrada e *worker 2* processa o segundo conjunto de dados (resultado do processamento da entrada), se o *Bundle A* falhar, como é que a execução prossegue?

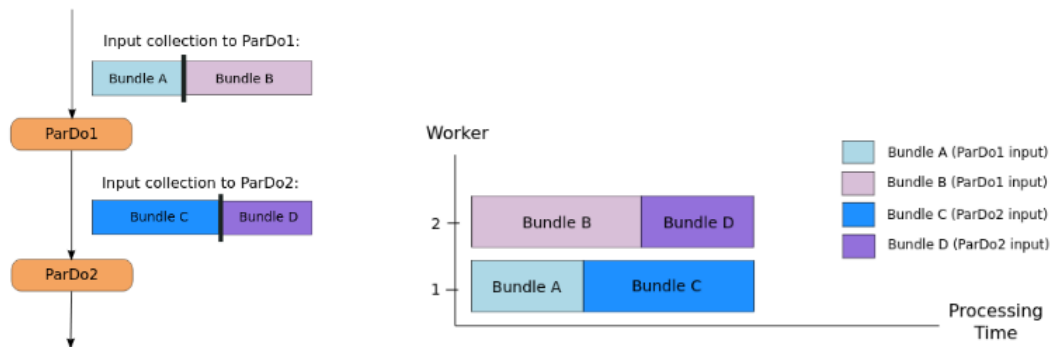


Figure 1: Exemplo Apache Beam

3. Dado o código Beam abaixo, indique:

- (a) o que poderia ser o input
- (b) o que seria o output
- (c) explique os passos de processamento da pipeline

```

1   below, above, marked = (
2       words
3       | beam.ParDo(ProcessWords(), cutoff_length=2,
4                           marker='x').with_outputs(
5                               'above_cutoff_lengths',
6                               'marked strings',
7                               main='below_cutoff_strings'))
8
9   class ProcessWords(beam.DoFn):
10      def process(self, element, cutoff_length, marker):
11          if len(element) <= cutoff_length:
12              yield element
13          else:
14              yield pvalue.TaggedOutput('above_cutoff_lengths', len(
15                  element))
16          if element.startswith(marker):
17              yield pvalue.TaggedOutput('marked strings', element)

```

4. No contexto da biblioteca Dask, para que tipos de dados utilizaria:

(a) bags (b) arrays (c) dataframe

5. Dado o código abaixo, qual é o valor da variável L quando é passada como parâmetro para a função `summarize`?

```
1 from dask import delayed
2 L = []
3 for fn in filenames:
4     data = delayed(load)(fn)
5     L.append(delayed(process)(data))
6
7 result = delayed(summarize)(L)
8 result.compute()
```

6. Em que situações não é adequado utilizar Dask dataframes?

7. Quais são as funcionalidades das bibliotecas Modin e Joblib e em que situações utilizaríamos uma ou outra ou as duas?

8. Indique **uma** vantagem e **uma** desvantagem do PySpark quando comparado com o Apache Beam.

9. Em Graph Neural Networks podemos utilizar pelo menos três métodos para representar os dados de entrada da rede neuronal: *Spectral*, *Spatial* e *Sampling*. Qual é a diferença entre estes três métodos em relação a:

(a) representação dos dados de entrada

(b) significado dos dados de entrada

(c) impacto no número de parâmetros e tempo de treinamento do modelo

10. Quais são as diferenças entre uma Message Passing Neural Network (MPNN) e uma Graph Attention Network (GAT)?

11. Dado o fragmento de código pycuda da Figura 2, e dado que $R=16$ e $C=16$, qual é o efeito de modificarmos a linha 43 para `func(a_gpu, block=(1, R, C))`?

```
34 mod = SourceModule("""
35     __global__ void doublify(float *a)
36     {
37         int idx = threadIdx.x + threadIdx.y*16;
38         a[idx] *= 2;
39     }
40 """)
41 func = mod.get_function("doublify")
42 # now, we define the shape
43 func(a_gpu, block=(R,C,1))
```

Figure 2: Exemplo de código pycuda