MC970/MO644 – Introdução à Programação Paralela

Laboratório 10 - Spark

Ronnypetson Souza da Silva - RA 211848

Professor: Guido Araújo

IC - Instituto de Computação

UNICAMP - Universidade Estadual de Campinas

r211848@dac.unicamp.br / ronnypetson.silva@students.ic.unicamp.br / rsronnypetson4@gmail.com

1) Introdução

Neste laboratório da disciplina Introdução à Programação Paralela foi implementado um analisador de textos simples, que recebe como entrada dois arquivos de texto, aplica um pré-processamento para "limpar" o texto de caracteres não desejados, conta o número de ocorrência das palavras e obtém as palavras mais frequentes em cada texto e lista as palavras comuns entre os textos com frequência maior que um certo limite. A implementação deste analisador usa conceitos de distribuição de processamento em clusters, através da linguagem de programação Scala e do framework Spark.

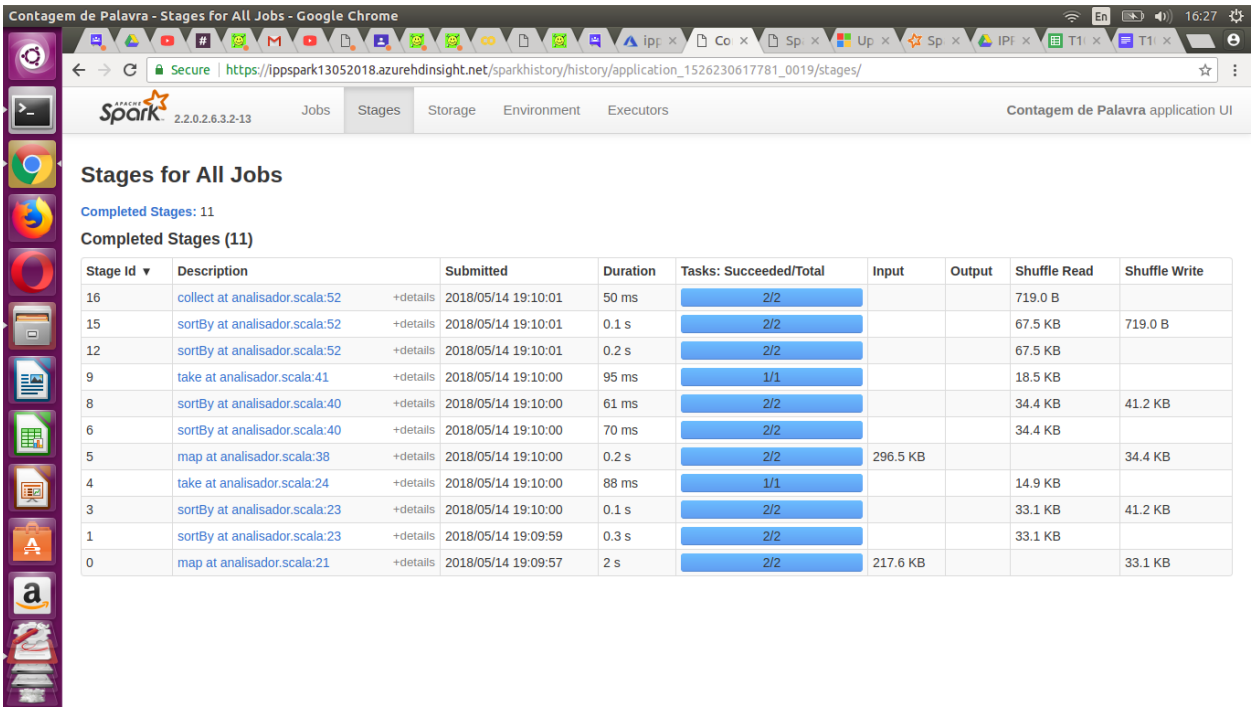
2) Comparação dos tempos de execução

O analisador de textos foi executado localmente de modo serial e na nuvem em um cluster "HDInsight" da Azure, de modo paralelo. Os tempos de execução (soma dos tempos de execução dos jobs e duração total no caso do cluster) são comparados na tabela abaixo para cada uma das 3 combinações de pares possíveis com os arquivos "alice30.txt", "warw10.txt" e "wizoz10.txt".

	Local (duração total)	Cluster (duração total)	Cluster (soma dos jobs)
alice30.txt, warw10.txt	9s	27s	2.13s
alice30.txt, wizoz10.txt	9s	24s	2.12s
warw10.txt, wizoz10.txt	9s	23s	3.11s

Nota-se que apesar do exemplo no cluster ter demorado mais, os textos usados neste laboratório são muito pequenos para que alguma vantagem do paralelismo seja notada.

A imagem a seguir é uma captura de tela da interface gráfica da ferramenta de profiling do Spark. Nela é possível ver a execução em paralelo do analisador nos núcleos do cluster.


Contagem de Palavra - Stages for All Jobs - Google Chrome
16:27
Secure | https://ippspark13052018.azurehdinsight.net/sparkhistory/history/application_1526230617781_0019/stages/
Spark 2.2.0.2.6.3.2-13
Jobs
Stages
Storage
Environment
Executors
Contagem de Palavra application UI
Stages for All Jobs
Completed Stages: 11
Completed Stages (11)
Stage Id ▼ | Description | Submitted | Duration | Tasks: Succeeded/Total | Input | Output | Shuffle Read | Shuffle Write
16 | collect at analisador.scala:52 +details | 2018/05/14 19:10:01 | 50 ms | 2/2 | | | 719.0 B |
15 | sortBy at analisador.scala:52 +details | 2018/05/14 19:10:01 | 0.1 s | 2/2 | | | 67.5 KB | 719.0 B
12 | sortBy at analisador.scala:52 +details | 2018/05/14 19:10:01 | 0.2 s | 2/2 | | | 67.5 KB |
9 | take at analisador.scala:41 +details | 2018/05/14 19:10:00 | 95 ms | 1/1 | | | 18.5 KB |
8 | sortBy at analisador.scala:40 +details | 2018/05/14 19:10:00 | 61 ms | 2/2 | | | 34.4 KB | 41.2 KB
6 | sortBy at analisador.scala:40 +details | 2018/05/14 19:10:00 | 70 ms | 2/2 | | | 34.4 KB |
5 | map at analisador.scala:38 +details | 2018/05/14 19:10:00 | 0.2 s | 2/2 | 296.5 KB | | | 34.4 KB
4 | take at analisador.scala:24 +details | 2018/05/14 19:10:00 | 88 ms | 1/1 | | | 14.9 KB |
3 | sortBy at analisador.scala:23 +details | 2018/05/14 19:10:00 | 0.1 s | 2/2 | | | 33.1 KB | 41.2 KB
1 | sortBy at analisador.scala:23 +details | 2018/05/14 19:09:59 | 0.3 s | 2/2 | | | 33.1 KB |
0 | map at analisador.scala:21 +details | 2018/05/14 19:09:57 | 2 s | 2/2 | 217.6 KB | | | 33.1 KB