Distribuição da premiação da Mega-Sena por estados

REIS, E. J.
emerson.oliveira@aluno.ufms.br

TORRES, J. P.
jeanpttorres@gmail.com

Setembro de 2016

Resumo

Neste trabalho apresentamos como foi a distribuição da premiação da Mega-Sena entre os estados do Brasil. Utilizando Hadoop e seu paradigma de programação MapReduce.

Palavras-chave: sistemas distribuídos, hadoop, mega-sena

1 Introdução

Mega-Sena sempre foi um tema cheio de polêmicas envolvidas. Para muito, é a chance de se tornarem milionários. Para outros, é tudo arranjado e apenas tomam o dinheiro do povo. Provavelmente nunca saberemos a verdade, mas neste trabalho apresentamos alguns estudos sobre os dados disponibilizados pela Caixa Federal[1] com todos os resultados da jogatina desde 1996.

2 Descrição do Problema

A partir do conjunto de dados da Mega-Sena, queremos saber como foi distribuída a premiação, dividindo por estados. Repare que neste trabalho usamos apenas os vencedores da sena, sem utilizar os vencedores de quina nem de quadra, já que estes não são disponibilizados pela Caixa Federal em seu conjunto de dados.

Utilizando os dados disponíveis, descobrimos a distribuição dos vencedores de acordo com a quantidade de vezes que ganharam e também de acordo com o valor total recebido por cada estado. Após isso, estudamos a relação dos resultados com as informações demográfica dos estados.

3 Metodologia

Para solucionar nosso problema, utilizamos *Hadoop 1.2.1*, junto com *Hadoop Streaming*, já que programamos nossos algoritmos de *MapReduce* em *Python 2.7* e não em *Java*. Os códigos produzidos foram baseados no aprendizado adquirido

na sala de aula e no curso *Intro to Hadoop and MapReduce*[2], todos os códigos podem ser encontrados no Github[3].

3.1 Conjunto de Dados

O conjunto de dados utilizado não é exatamente igual ao disponibilizado pela Caixa Federal[1], fizemos duas principais modificações.

A primeira foi passar as informações que são apresentadas em *HTML* para texto, isso foi feito simplesmente abrindo o arquivo *HTML*, selecionando toda a página (CTRL + A) e colando em um editor de texto.

A segunda modificação ocorreu para facilitar o nosso *mapper*, pois após a primeira modificação, em casos onde haviam mais de 1 ganhador por sorteio, os outros estados ficavam em linhas diferentes, por causa disso, resolvemos copiar a linha original modificando apenas o estado vencedor, sem perda de informações estatisticamente.

Após as modificações, tivemos o tamanho do nosso cojunto de dados, que foi 237.5 KB. Sabemos que isso é um tamanho muito pequeno para uma tarefa ser executada distribuidamente e utilizando os recursos que tínhamos, já que o *Hadoop* utiliza *chunks* de tamanho 64 MB, então nosso conjunto de dados usaria apenas 1 *chunk*.

Para não mudar o tema do trabalho, resolvemos replicar as informações do conjunto de dados, criando um conjunto estatisticamente equivalente, mas com tamanho muito maior: 4.8 GB. Fizemos isso replicando cada linha do nosso conjunto de dados 25000 (vinte e cinco mil) vezes.

3.2 Configuração do Hadoop

O *Hadoop* foi instalado e configurado no *cluster* do CTEI seguindo os passos disponibilizados durante as aulas e no portal EAD. Para o *masters*, usamos "localhost", e para os *slaves* usamos "compute-3-3".

3.3 Distribuição dos Vencedores

Após a execução das tarefas *Mapper* e *Reducer* sobre os dados, obtivemos tanto a proporção por quantidade de vencedores quanto pelo valor total recebido pelos vencedores para cada estado. Essas informações foram analisadas e comparadas de acordo com a proporção da população do estado em relação a população total do país. O mais correto estatisticamente falando seria utilizar o número de apostas de cada estado, mas essa informação não está disponível, então usamos a população, que de qualquer forma também representa o número de apostas. Os resultados são apresentados na próxima seção.

4 Resultados e Discussão

A execução foi relativamente rápida (cerca de 5 minutos) para um conjunto de dados do tamanho utilizado (4.8 GB), principalmente comparado a um algoritmo que fosse executado localmente que certamente demoraria muito mais tempo do que o obtido utilizando o *cluster* e *Hadoop*.

Todos os dados sobre a população do Brasil foram coletados no site do IBGE (Instituto Brasileiro de Geografia e Estatística)[4]. Vale a pena ressaltar que o estado do Amapá nunca recebeu o prêmio da sena no concurso.

4.1 Resultados por Quantidade de Vencedores

Para a execução dessa tarefa, o *Hadoop* gastou cerca de 4 minutos e 26 segundos, um tempo razoável e bem viável considerando o tamanho do conjunto de dados. na tabela a seguir podemos ver as proporções tanto de população quanto para a distribuição de vencedores por estado. Perceba que a ordem dos estados apresentada é a mesma ordem obtida no site do IBGE[4], que, por motivos desconhecidos, não é a ordem alfabetica.

Estado	População	%	Vencedores	%
AC	819,121.00	0.40%	1	0.17%
AL	3,362,271.00	1.63%	5	0.83%
AP	785,938.00	0.38%	0	0.00%
AM	4,013,428.00	1.94%	6	1.00%
BA	15,290,441.00	7.41%	21	3.50%
CE	8,980,294.00	4.35%	15	2.50%
DF	2,986,168.00	1.45%	24	4.00%
ES	3,978,281.00	1.93%	20	3.33%
GO	6,713,266.00	3.25%	13	2.17%
MA	6,959,009.00	3.37%	5	0.83%
MT	3,312,906.00	1.61%	11	1.83%
MS	2,689,187.00	1.30%	11	1.83%
MG	21,023,628.00	10.19%	58	9.67%
PA	8,287,421.00	4.02%	11	1.83%
PB	4,002,983.00	1.94%	6	1.00%
PR	11,262,970.00	5.46%	54	9.00%
PE	9,422,855.00	4.57%	18	3.00%
PI	3,214,451.00	1.56%	4	0.67%
RJ	16,659,351.00	8.07%	66	11.00%
RN	3,482,162.00	1.69%	7	1.17%
RS	11,300,646.00	5.47%	31	5.17%
RO	1,791,365.00	0.87%	5	0.83%
RR	516,868.00	0.25%	1	0.17%
SC	6,907,788.00	3.35%	19	3.17%
SP	44,838,872.00	21.72%	181	30.17%
SE	2,270,799.00	1.10%	6	1.00%
TO	1,536,544.00	0.74%	1	0.17%
BRASIL	206,409,013.00	100.00%	600	100.00%

Vemos na tabela que proporção tende a ser bastante semelhante, sugerindo a corretude estatística do concurso.

4.2 Resultados por Valor Recebido pelos Vencedores

Nesta tarefa, foram gastos 5 minutos e 31 segundos para a execução do *mapper* e do *reducer* pelo *Hadoop*. Tempo também bastante viável. Veja os resultados obtidos na tabela a seguir.

Estado	População	%	Vencedores	%
AC	819,121.00	0.40%	6,796,888.03	0.07%
AL	3,362,271.00	1.63%	125,285,438.30	1.25%
AP	785,938.00	0.38%	0.00	0.00%
AM	4,013,428.00	1.94%	41,864,334.43	0.42%
BA	15,290,441.00	7.41%	232,575,547.12	2.33%
CE	8,980,294.00	4.35%	199,267,034.41	2.00%
DF	2,986,168.00	1.45%	654,277,358.66	6.55%
ES	3,978,281.00	1.93%	389,068,491.36	3.90%
GO	6,713,266.00	3.25%	376,913,581.80	3.77%
MA	6,959,009.00	3.37%	58,557,059.23	0.59%
MT	3,312,906.00	1.61%	215,489,681.99	2.16%
MS	2,689,187.00	1.30%	163,989,451.66	1.64%
MG	21,023,628.00	10.19%	931,671,186.75	9.33%
PA	8,287,421.00	4.02%	140,013,592.26	1.40%
PB	4,002,983.00	1.94%	17,782,897.70	0.18%
PR	11,262,970.00	5.46%	1,016,012,843.14	10.18%
PE	9,422,855.00	4.57%	191,739,041.18	1.92%
PI	3,214,451.00	1.56%	59,576,916.67	0.60%
RJ	16,659,351.00	8.07%	1,172,860,434.86	11.75%
RN	3,482,162.00	1.69%	63,813,462.76	0.64%
RS	11,300,646.00	5.47%	475,104,919.81	4.76%
RO	1,791,365.00	0.87%	107,035,124.07	1.07%
RR	516,868.00	0.25%	2,820,316.51	0.03%
SC	6,907,788.00	3.35%	242,318,398.76	2.43%
SP	44,838,872.00	21.72%	2,968,173,035.89	29.73%
SE	2,270,799.00	1.10%	129,757,083.38	1.30%
TO	1,536,544.00	0.74%	1,931,474.80	0.02%
BRASIL	206,409,013.00	100.00%	9,984,695,595.53	100.00%

Nesta tabela também podemos perceber certa tendência das proporções a seguirem a proporção da população. É verdade que neste caso a correlação está um pouco pior do que o visto pela primeira tabela, mas ainda assim sugere a corretude estatística do concurso.

5 Conclusões

Podemos concluir a partir dos resultados obtidos que o *Hadoop* se mostra uma ferramenta bastante eficiente, somada ao paradigma de *MapReduce* eles conseguem resolver problemas envolvendo grandes conjuntos de dados em tempo viável, o que seria impossível com ferramentas e paradigmas ordinários.

Sobre o concurso da Mega-Sena, fica difícil tirar muitas conclusões apenas com os dados coletados, principalmente por não estarmos usando a proporção de apostas, e sim a da população. Seria preciso um estudo mais aprofundado no assunto, com várias outras abordagens.

Mas para esse trabalho vimos que a proporção tende a ser semelhante à proporção da população dos estados, o que indicaria uma certa confiança no concurso, pois de certa forma mostra que não há tratamento preferencial para nenhum estado. Por outro lado, os mais céticos diriam que o concurso é manipulado para que os resultados se matenham numa proporção parecida com a proporção da população para que ninguém desconfie do concurso, porém, não podemos provar nada.

Referências

[1] Resultados da Mega-Sena,
`http://loterias.caixa.gov.br/wps/portal/loterias/landing/megasena/`
(visitado em 13/09/2016)

[2] Sarah Sproehnle, Ian Wrigley, and Gundega Dekena.
`https://udacity.com/course/intro-to-hadoop-and-mapreduce--ud617`,
Intro to Hadoop and MapReduce, How to Process Big Data

[3] Emerson Jair's Github,
`https://github.com/dungahk/mega-sena-hadoop`

[4] Projeção da população do Brasil e das Unidades da Federação,
`http://www.ibge.gov.br/apps/populacao/projecao/`
(visitado em 14/09/2016 às 22:40)