Note: all the numeric values in equations below are presented as examples only.

y_1 = size of raw data

y_2 = size of processed data

Y = maximum size of data in hadoop

$$\overset{(1)}{\Rightarrow} y_1 + y_2 = Y$$

T = size of temporary data

I = size of intermediate data= $p \times y_1,\ \ p \cong 0.3$

R = replications in hadoop

S = available size required to process data of y_1 size in hadoop

$$\overset{(2)}{\Rightarrow}\ (T + R \times Y + I) \times 1.2 = S$$

c_1 = ***lzo*** compression ratio $\cong 2$

c_2 = ***parquet*** compression ratio $\cong 5$

$$\overset{(3)}{\Rightarrow} T = \max_{1 \leq i \leq 2}(c_i \times y_i\,)$$

$$e = \frac{y_1}{y_2} \cong 50$$

$\overset{(5)}{\Rightarrow}\ c_1 \times y_1 \gg c_2 \times y_2$ (based on our numeric assumptions)

$$(3),(5) \overset{(6)}{\Rightarrow}\ T = c_1 \times y_1$$

$$(1),(2),(6)\ \overset{(7)}{\Rightarrow}\ (c_1 \times y_1 + R \times (y_1 + y_2) + p \times y_1) \times 1.2 = S$$

$$\overset{(8)}{\Rightarrow}\ y_2 = \frac{S}{(e \times c_1 + R \times (e+1) + p \times e) \times 1.2}\ ,\ \ y_1 = \frac{e \times S}{(e \times c_1 + R \times (e+1) + p \times e) \times 1.2}$$

$$\xRightarrow{if\ e=50, c_1=2, c_2=5, R=3, S=3T, p=0.3}\ y_2 \cong 10\ GB, y_1 \cong 500\ GB$$