日研システムにおける分散処理システムの構築実験について

1．実験内容

ビッグデータ時代と呼ばれて久しい中で、代表的存在の Hadoop で分散環境の構築が実施できること、また、実際に稼働している何百分の 1 以下の環境でも、分散環境の特徴である「複数のマシンを使ってスケールアウトする「分散処理」」が体現できるか、下記の内容で実験を行った。

今回の実験は Hadoop で代表的な「WordCount」を用いて行った。

2．分散処理システム環境

OS	Hadoop	CDH	Java
Centos6.8	2.5.0	5.2.1	1.8.0_144

3．スレーブノードの仕様

名称	ＣＰＵ	メモリ	ディスク	台数
S-type1	Core I3-5010U 2.10 GHz	12 GB	500 GB	4 台
S-type2	Core I5-6200U 2.30 GHz	12 GB	500 GB	2 台

4．Hadoop 基盤環境

ケース	スレーブノード数	説明
1	1	S-type1
2	2	S-type1
3	6	S-type1（4 台）、S-type2（2 台）

5．データ

無作為に単語を並べた大容量ファイル（1.06GB）

6．結果

ケース	実行時間	オーバヘッドを 30 秒と仮定した時間	備考
1	３ｍ２７ｓ（２０７ｓ）	２ｍ５７ｓ（１７７ｓ）	
2	１ｍ５８ｓ（１７７ｓ）	１ｍ２８ｓ（　８８ｓ）	ケース１の約１／２
3	１ｍ０２ｓ（　６２ｓ）	３２ｓ（　３２ｓ）	ケース２の役１／３

7．まとめ

オーバヘッドを考慮した数値で考慮すると、今回の数値が分散環境で聞いていた数値となったことでひとまず安心できた。やはり今回のような小さい環境では特徴を生かしきることができなかったが、次は 10 台程度に増やしてみてやって見るとどうだろう等の興味が出てきて、なかなか面白いものとなった。

次回は、自分たちで開発をしてみよう！ということで、自作のプログラムが動く環境の作成を目指す。