

# Experiments and Results

## Specs

- Apple M1, 16GB RAM, 8 cores
- Hadoop single node setup
  - Max split size → 40MB

## Dataset Scale

details: <https://github.com/omkarprabhu-98/mining-white-house-visitor-logs/blob/master/DATA.md>

1x → 28 MB

2x → 80 MB

6x → 180 MB

14x → 407 MB

## Application 1

Get top 10 based on the key:

- key 0 → top10 visitors

Sample Result 6x Dataset:

```

190 cat Top10Tmp/output/part-r-*
191 fontenot_yvette_e → 155
192 levitis_jason_a → 156
193 schultz_william_b → 156
194 borzi_phyllis_c → 165
195 khalid_aryana_c → 179
196 brookslasure_chiquita_n → 185
197 tavenner_marilyn_n → 196
198 hoff_james_c → 197
199 oneil_dennis_p → 234
200 hash_michael_m → 315

```

- key 1 → top10 visitee

Sample Result 6x Dataset:

```

cat Top10Tmp/output/part-r-*
raghavan_gautam 3401
matusiak_ari → 3883
_ → 3946
/_potus 3992
mccullough_victoria 4037
lambrew_jeanne → 6426
lierman_kyle → 8023
_potus/flotus → 11060
_potus → 39551
office_visitors 623575

```

- key 2 → top10 visitor-visitee combination

Sample Result 6x Dataset:



```
cat Top10Tmp/output/part-r-*
levitis_jason_a_&&lambrew_jeanne 115
mann_cynthia_r_&&lambrew_jeanne 115
kronick_richard_g_&&lambrew_jeanne 117
fontenot_yvette_e_&&lambrew_jeanne 122
choe_kenneth_y_&&lambrew_jeanne 123
brookslasure_chiquita_n_&&lambrew_jeanne 144
khalid_aryana_c_&&lambrew_jeanne 151
tavenner_marilyn_n_&&lambrew_jeanne 163
hoff_james_c_&&hoff_joanne 177
hash_michael_m_&&lambrew_jeanne 231
```

- key 3 → top 10 locations for meetings (based on no of visitors)

Sample Result 6x Dataset:

```
cat Top10Tmp/output/part-r-*
meeting_loc 2
vpr 3945
neob 22138
oeob 186590
wh 785654
```

**Table 1**

<u>Aa</u> Key No	 Dataset Size	 No of Mappers (Job1, Job2)	 No of Reducers (Job1, Job2)	 Job1 Time spent in mappers	 Job1 Time spent in reducers	 Job2 Time spent in mappers	 Job2 Time spent in reducers
<u>0</u>	1x	1, 3	3, 1	0.1	0.4	0.6	0.1
<u>0</u>	2x	2, 3	3, 1	0.4	0.4	0.6	0.1
<u>0</u>	6x	5, 3	3, 1	2	0.5	0.7	0.1
<u>0</u>	14x	11, 3	3, 1	5.5	0.9	0.6	0.1
<u>1</u>	1x	1, 3	3	0.13	0.40	0.64	0.13
<u>1</u>	2x	2, 3	3	0.35	0.43	0.49	0.12

Aa Key No	≡ Dataset Size	≡ No of Mappers (Job1, Job2)	≡ No of Reducers (Job1, Job2)	≡ Job1 Time spent in mappers	≡ Job1 Time spent in reducers	≡ Job2 Time spent in mappers	≡ Job2 Time spent in reducers
<u>1</u>	6x	5, 3	3	1.98	0.48	0.56	0.15
<u>1</u>	14x	11, 3	3	4.73	0.71	0.51	0.12
<u>2</u>	1x	1, 3	3	<b>0.40</b>	<b>0.15</b>	0.49	0.13
<u>2</u>	2x	2, 3	3	0.44	0.45	0.59	0.12
<u>2</u>	6x	5, 3	3	2.03	0.50	0.60	0.12
<u>2</u>	14x	11, 3	3	4.95	1.00	0.63	0.12
<u>3</u>	1x	1, 3	3	0.14	0.39	0.52	0.12
<u>3</u>	2x	2, 3	3	0.38	0.41	0.60	0.12
<u>3</u>	6x	5, 3	3	1.94	0.44	0.53	0.12
<u>3</u>	14x	11, 3	3	4.87	0.68	0.55	0.12

Execution logs (containing results) can be found in folder  
app<app\_no>\_<key\_no>\_<dataset\_scale>

## Application 2

Monthly Distribution:

- key 0 → of visitors

Sample Result 6x Dataset:

```
cat MonthlyDistTmp/output/part-r-*
2-2009→ 1
3-2014→ 74398
NULL-NULL→ 985668
1-2014→ 39202
6-2009→ 1
4-2009→ 1
```

- key 1 → of no. of visits to the POTUS

Sample Result 6x Dataset:

```
cat MonthlyDistTmp/output/part-r-*
3-2014→ 1284
NULL-NULL→ 39807
1-2014→ 1413
```

**Table 2**

<u>Aa</u> Key No	<u>≡</u> Dataset Size	<u>≡</u> No of Mappers	<u>≡</u> No of Reducers	<u>≡</u> Time spent in mappers	<u>≡</u> Time spent in reducers
<u>0</u>	1x	1	3	0.15	0.37
<u>0</u>	2x	2	3	0.39	0.40
<u>0</u>	6x	5	3	1.70	0.43
<u>0</u>	14x	11	3	4.79	0.74
<u>1</u>	1x	1	3	0.15	0.39
<u>1</u>	2x	2	3	0.38	0.38
<u>1</u>	6x	5	3	1.93	0.40
<u>1</u>	14x	11	3	4.81	0.73
<u>Untitled</u>					

Execution logs (containing results) can be found in folder  
app<app\_no>\_<key\_no>\_<dataset\_scale>