# An Empirical Study of the Out of Memory Errors in Data-parallel Applications

Cause distribution

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Causes  Sources | Large data stored in memory | | Abnormal dataflow | | | Memory-consuming user code | | |
| Large buffered intermediate data | Large cached  data | Improper data partition | Hotspot key | Large single record | Large intermediate results | Large accumulated results | Large results generated /collected by driver |
| StackOverflow  (23/59) | 3 | 2 | 1 | 7 | 5 | 5(5) | 10(5) |  |
| Hadoop mailing list (12/65) | 1 |  | 1 | 1 | 1 | 1 | 8(1) |  |
| Developer’s blogs (3/3) |  |  | 1 | 1 |  |  | 2(1) |  |
| MapReduce books (9/9) |  | 2 |  | 3 |  |  | 9(3) |  |
| Spark user/dev mailing list  (57/152) | 2 | 9 | 8 | 8 | 2 | 2 | 10(2) | 9,9 |
| StackOverflow (20/46) | 2 |  | 5 | 1 |  | 2 | 2 | 2,6 |
| Total (124/334) | 8 | 13 | 16 | 21 | 3 | 10 | 41 | 11,15 |

Cause patterns:

