**Hadoop**

**1. Job type: User-defined (StackOverflow)**

**Causes: Large framework buffer (identified by expert)**

**Fix suggestions: lower buffer size (accepted)**

**2. Job type: User-defined (StackOverflow)**

**Causes: Large framework buffer (identified by user)**

**Fix suggestions: lower buffer size (user)**

**3. Job type: Pig (StackOverflow)**

**Causes: Large framework buffer (identified by user)**

**Fix suggestions: lower buffer size (user), use “skewed”**

**4. Job type: User-defined (StackOverflow)**

**Causes: Large framework buffer (reproduced)**

**Fix suggestions: lower buffer size (us)**

**5. Job type: User-defined (StackOverflow)**

**Causes: Large framework buffer (reproduced)**

**Fix suggestions: lower buffer size (us)**

**6. Job type: Apache Hive (Mailing list)**

**Causes: Large framework buffer (Expert)**

**Fix suggestions: lower buffer size (Expert)**

**7. Job type: Pig (Mailing list)**

**Causes: Improper data partition (Expert)**

**Fix suggestions: no**

**8. Job type: Pig script (StackOverflow)**

**Causes: Improper data partition (user)**

**Fix suggestions: add reduce number and lower framework buffer (User)**

**9. Job type: User-defined (Mailing list)**

**Causes: Improper data partition (Reproduced)**

**Fix suggestions: add reduce number (us)**

**10. Job type: User-defined (StackOverflow)**

**Causes: Hotspot key (Expert)**

**Fix suggestions: Filter the useless the hot key (Expert)**

**11. Job type: Pig (Book, MapReduce Design Patterns)**

**Causes: Hotspot key (Expert)**

**Fix suggestions: no**

**12.Job type: User-defined (StackOverflow)**

**Causes: Large single record (User)**

**Fix suggestions: no**

**13. Job type: User-defined (StackOverflow)**

**Causes: Large single record (Expert)**

**Fix suggestions: no**

**14. Job type: User-defined (StackOverflow)**

**Causes: Large single record (Expert)**

**Fix suggestions: read data chunk by chunk, avoid load the whole file (Expert)**

**15. Job type: User-defined (StackOverflow)**

**Causes: Large single record (Expert)**

**Fix suggestions: break up the large record into smaller records (Expert)**

**16. Job type: User-defined (Mailing list)**

**Causes: Large single** **record (Expert)**

**Fix suggestions: no**

**17. Job type: Apache Mahout (StackOverflow)**

**Causes: Large external data (large training model) (Expert)**

**Fix suggestions: Add memory space (Expert)**

**18. Job type: Apache Hive (StackOverflow)**

**Causes: Large external data (large external table) (User)**

**Fix suggestions: decrease the dataset (User)**

**19. Job type: Pig join (Book, MapReduce Design Pattern)**

**Causes: Large external data (large table for join) (Expert, reproduced)**

**Fix suggestions: use reduce join (Expert)**

**Interesting: Yes**

**20. Job type: Pig join (Book, MapReduce Design Pattern)**

**Causes: Large external data (user data) (Expert)**

**Fix suggestions: no**

**21. Job type: Apache Mahout (StackOverflow)**

**Causes: Large external data (large model) (us, reproduced)**

**Fix suggestions: no**

**22. Job type: User-defined (StackOverflow)**

**Causes: Large external data (external table) (us, reproduced)**

**Fix suggestions: no**

**23. Job type: Apache Mahout (StackOverflow)**

**Causes: Large external data (external table) (us, reproduced)**

**Fix suggestions: no**

**24. Job type: Apache Pig (Mailing list)**

**Causes: Large external data (User)**

**Fix suggestions: no**

**25. Job type: User-defined with third library (StackOverflow)**

**Causes: Large intermediate results + large single record (Expert & Reproduced)**

**Fix suggestions: split long records to multiple small records (Expert, accepted)**

**26. Job type: User-defined (StackOverflow)**

**Causes: Large intermediate results (User)**

**Fix suggestions: no**

**27. Job type: User-defined (StackOverflow)**

**Causes: Large accumulated results + Hotspot key (Expert)**

**Fix suggestions: avoid accumulation (Expert)**

**28. Job type: User-defined (StackOverflow)**

**Causes: Large accumulated results + Hotspot key (User and Expert, reproduced)**

**Fix suggestions: use “skew join” (Expert)** create artificial keys and spread the key distribution

**29. Job type: User-defined (StackOverflow)**

**Causes: Large accumulated results (User)**

**Fix suggestions: no**

**30. Job type: User-defined (StackOverflow)**

**Causes: Large accumulated results + Hotspot key (Expert)**

**Fix suggestions: Job split, do the operation in several passes (Expert)**

**31. Job type: Apache Hive (StackOverflow)**

**Causes: Large accumulated results (Expert)**

**Fix suggestions: set hive.map.aggr.hash.percentmemory (Expert)**

**32. Job type: Pig UDF (StackOverflow)**

**Causes: Large accumulated results (Expert)**

**Fix suggestions: no**

**33. Job type: User-defined (StackOverflow)**

**Causes: Large accumulated results + Hotspot key (Expert)**

**Fix suggestions: add combiner to reduce the values for a key, in-memory combing (Expert)**

**34. Job type: Apache Mahout (StackOverflow)**

**Causes: Large accumulated results (User)**

**Fix suggestions: no**

**35. Job type: User-defined (StackOverflow)**

**Causes: Large accumulated results (User)**

**Fix suggestions: no**

**36. Job type: User-defined (StackOverflow)**

**Causes: Large accumulated results + Hotspot key (Expert)**

**Fix suggestions: multiple outputs (Expert)**

**37. Job type: Pig script (Mailing list)**

**Causes: Large accumulated results (Expert)**

**Fix suggestions: Filter the unrelated columns (TOP), use two groups (Expert)**

**38. Job type: Pig script (Mailing list)**

**Causes: Large accumulated results (Expert)**

**Fix suggestions: no**

**39. Job type: Cloud9 (Mailing list)**

**Causes: Large accumulated results (Expert, Reproduced)**

**Fix suggestions: Change sort by using Hadoop secondary sort (Expert)**

**40. Job type: Pig script (Mailing list)**

**Causes: Large accumulated results (Expert, Reproduced)**

**Fix suggestions: two groups (Expert)**

**41. Job type: User-defined (developer’s blog)**

**Causes: Large accumulated results (User, Reproduced)**

**Fix suggestions: Map aggregation => Reduce aggregation, avoid in-memory aggregation (us)**

**42. Job type: User-defined (developer’s blog)**

**Causes: Large accumulated results + Hotspot key (User, Reproduced)**

**Fix suggestions: output several records for the same key (us)**

**43. Job type: User-defined (book, MapReduce Design Patterns)**

**Causes: Large accumulated results (Expert, Reproduced)**

**Fix suggestions: no**

**Interesting: Yes**

**44. Job type: User-defined (book, MapReduce Design Patterns)**

**Causes: Large accumulated results (Expert)**

**Fix suggestions: no**

**45. Job type: User-defined (book, Data-Intensive Text Processing with MapReduce)**

**Causes: Large accumulated results (Expert)**

**Fix suggestions: control the sort order (Expert)**

**46. Job type: User-defined (book, Data-Intensive Text Processing with MapReduce)**

**Causes: Large accumulated results (Expert)**

**Fix suggestions: Use framework sort mechanism (Expert)**

**Interesting: Yes**

**47. Job type: User-defined (book, Data-Intensive Text Processing with MapReduce)**

**Causes: Large accumulated results (Expert)**

**Fix suggestions: no**

**48. Job type: User-defined (book, Data-Intensive Text Processing with MapReduce)**

**Causes: Large accumulated results (Expert)**

**Fix suggestions: no**

**49. Job type: User-defined (book, Data-Intensive Text Processing with MapReduce)**

**Causes: Large accumulated results + Hotspot key(Expert)**

**Fix suggestions: no**

**50. Job type: User-defined (book, Data-Intensive Text Processing with MapReduce)**

**Causes: Large accumulated results + Hotspot key(Expert)**

**Fix suggestions: no**

**51. Job type: Pig (StackOverflow)**

**Causes: Large accumulated results + Hotspot key (us, reproduced)**

**Fix suggestions: no**

**52. Job type: User-defined (Mailing list)**

**Causes: Large accumulated results (Expert)**

**Fix suggestions: no**

**53. Job type: User-defined (Mailing list)**

**Causes: Large accumulated results + Hotspot Key (Expert, reproduced)**

**Fix suggestions: Write the computing results continuously (Expert)**

**54. Job type: User-defined (Mailing list)**

**Causes: Large accumulated results + Hotspot Key (Expert)**

**Fix suggestions: Use combiner (User)**

**55. Job type: Apache Pig (Mailing list)**

**Causes: Large accumulated results - Count(distinct) + Hotspot Key (Expert)**

**Fix suggestions: Add reduce number (User)**

**Spark:**

1. **Job type: User-defined (StackOverflow)**

**Causes: Large buffered data (User)**

**Fix suggestions: lower the buffer size (user)**

2. **Job type: User-defined (Mailing list)**

**Causes: Large buffered data (User)**

**Fix suggestions: decrease the rank value, lower the buffer size, lower the cache limit (Expert)**

1. **Job type: User-defined (Mailing list)**

**Causes: Large cached data (User)**

**Fix suggestions: no**

2. **Job type: User-defined (Mailing list)**

**Causes: Large cached data (User)**

**Fix suggestions: no**

**3. Job type: User-defined (Mailing list)**

**Causes: Large cached data (Expert)**

**Fix suggestions: Try TorrentBroadcast**

**4. Job type: Streaming (Mailing list)**

**Causes: Large cached data (Expert)**

**Fix suggestions: Change storage level, increase reduce number, add combineByKey()**

**5. Job type: User-defined (Mailing list)**

**Causes: Large cached data (User)**

**Fix suggestions: no**

**6. Job type: GraphX (Mailing list)**

**Causes: Large cached data (User)**

**Fix suggestions: no**

**7. Job type: User-defined (Mailing list)**

**Causes: Large cached data (Expert)**

**Fix suggestions: lower the cache size**

**1. Job type: User-defined (StackOverflow)**

**Causes: Improper data partition (Expert)**

**Fix suggestions: add partition number**

**2. Job type: User-defined (StackOverflow)**

**Causes: Improper data partition (Expert)**

**Fix suggestions: add partition number**

**3. Job type: User-defined (StackOverflow)**

**Causes: Improper data partition (User)**

**Fix suggestions: repartition() (User)**

**4. Job type: User-defined (StackOverflow)**

**Causes: Improper data partition (User)**

**Fix suggestions: add partition number (User)**

**5. Job type: User-defined (StackOverflow)**

**Causes: Improper data partition (User)**

**Fix suggestions: add partition number (User)**

**6. Job type: User-defined (StackOverflow)**

**Causes: Improper data partition (Expert)**

**Fix suggestions: increase the partition number**

**7. Job type: MLlib (StackOverflow)**

**Causes: Improper data partition (User)**

**Fix suggestions: add partition number**

**8. Job type: User-defined (Mailing list)**

**Causes: Improper data partition (us, reproduced)**

**Fix suggestions: set spark.spill = true**

**9. Job type: User-defined (Mailing list)**

**Causes: Improper data partition (us, reproduced)**

**Fix suggestions: no**

**10. Job type: User-defined (Mailing list)**

**Causes: Improper data partition (Expert)**

**Fix suggestions: repartition (add partition number)**

**11. Job type: User-defined (Mailing list)**

**Causes: Improper data partition (User)**

**Fix suggestions: change partition function**

**12. Job type: User-defined (Mailing list)**

**Causes: Improper data partition (User)**

**Fix suggestions: repartition (add partition number)**

**13. Job type: User-defined (Mailing list)**

**Causes: Improper data partition (User)**

**Fix suggestions: no**

**14. Job type: User-defined (Mailing list)**

**Causes: Improper data partition (Expert)**

**Fix suggestions: repartition (add partition number)**

**1. Job type: User-defined (Mailing list)**

**Causes: Improper data partition (Expert)**

**Fix suggestions: repartition (add partition number)**

**2. Job type: User-defined (Mailing list)**

**Causes: Hotspot Key (User, reproduced)**

**Fix suggestions:  no**

**3. Job type: User-defined (Mailing list)**

**Causes: Hotspot Key (Expert)**

**Fix suggestions:  change key, cut down the data associated to a single key (Expert)**

**4. Job type: User-defined (Mailing list)**

**Causes: Hotspot Key (Expert, reproduced)**

**Fix suggestions: change key, sub-dividing any very large groups, aggregating partial values for each group (Expert)**

**5. Job type: User-defined (Mailing list)**

**Causes: Hotspot Key (User, reproduced)**

**Fix suggestions: no**

**6. Job type: User-defined (Mailing list)**

**Causes: Hotspot Key (User, reproduced)**

**Fix suggestions: no**

**7. Job type: User-defined (Mailing list)**

**Causes: Hotspot Key (us, reproduced)**

**Fix suggestions: no**

**1. Job type: User-defined (Mailing list)**

**Causes: Large single record (Expert, reproduced)**

**Fix suggestions: write multiple times (split the single large record)**

**1. Job type: User-defined (Mailing list)**

**Causes: Large intermediate results (User)**

**Fix suggestions: add partition number, change storage level**

**2. Job type: MLlib (Mailing list)**

**Causes: Large intermediate results (Expert, reproduced)**

**Fix suggestions: adjust the parameters**

**1. Job type: User-defined (StackOverflow)**

**Causes: Large accumulated results (Expert, Reproduced)**

**Fix suggestions:  moving the byte array allocation outside the iterator, so it gets reused by all partition elements**

**2. Job type: User-defined (StackOverflow)**

**Causes: Large accumulated results (Expert)**

**Fix suggestions:  increase the number of partitions**

**3. Job type: User-defined (StackOverflow)**

**Causes: Large accumulated in accumulableCollection (Expert, reproduced)**

**Fix suggestions: no**

**4. Job type: User-defined (Mailing list)**

**Causes: Large accumulated results (us, reproduced)**

**Fix suggestions: no**

**5. Job type: User-defined (Mailing list)**

**Causes: Large accumulated results (Expert)**

**Fix suggestions: no**

**6. Job type: User-defined (Mailing list)**

**Causes: Large accumulated results (Expert)**

**Fix suggestions: add partition number**

**7. Job type: User-defined (Mailing list)**

**Causes: Large accumulated results (us, reproduced)**

**Fix suggestions: no**

**8. Job type: User-defined (Mailing list)**

**Causes: Large accumulated results (Expert)**

**Fix suggestions: no**

**9. Job type: User-defined (Mailing list)**

**Causes: Large accumulated results (User)**

**Fix suggestions: no**

**10. Job type: User-defined (Mailing list)**

**Causes: Large accumulated results + Hotspot Key (Expert, reproduced)**

**Fix suggestions: no**

**1. Job type: MLlib (StackOverflow)**

**Causes: Driver generate large results (Top eigenvectors) (User)**

**Fix suggestions: Add driver’s memory space**

**2. Job type: User-defined (StackOverflow)**

**Causes: Driver generates large results (Expert)**

**Fix suggestions: no**

**3. Job type: MLlib (Mailing list)**

**Causes: Driver generates large results (Expert, reproduced)**

**Fix suggestions: reduce the partition number**

**4. Job type: User-defined (Mailing list)**

**Causes: Driver generates/broadcasts large results (Expert)**

**Fix suggestions: no**

**5. Job type: User-defined (Mailing list)**

**Causes: Driver generates/broadcasts large results (User)**

**Fix suggestions: no**

**6. Job type: MLlib (Mailing list)**

**Causes: Driver generates large results (User, reproduced)**

**Fix suggestions: no**

**7. Job type: User code (Mailing list)**

**Causes: Driver generates large results (Expert)**

**Fix suggestions: no**

**8. Job type: User code (Mailing list)**

**Causes: Driver generates/broadcasts large results (User)**

**Fix suggestions: no**

**9. Job type: User code (Mailing list)**

**Causes: Driver generates/broadcasts large results (User)**

**Fix suggestions: no**

**1. Job type: User-defined (StackOverflow)**

**Causes: Driver collect results (Expert)**

**Fix suggestions: no**

**2. Job type: User-defined (StackOverflow)**

**Causes: Driver collects large results (User)**

**Fix suggestions: Multi-level tree aggregation (Expert)**

**3. Job type: GraphX (StackOverflow)**

**Causes: Driver collects large results (User, reproduced)**

**Fix suggestions:  Change collect().count() to count() (User)**

**4. Job type: User-defined (StackOverflow)**

**Causes: Driver collects large results (User)**

**Fix suggestions: no**

**5. Job type: User-defined (StackOverflow)**

**Causes: Driver collects large results (Expert)**

**Fix suggestions: remove collect()**

**6. Job type: User-defined (Mailing list)**

**Causes: Driver collects large results (Expert)**

**Fix suggestions: no**

**7. Job type: User-defined (Mailing list)**

**Causes: Driver collects large results (User)**

**Fix suggestions: no**

**8. Job type: User code (Mailing list)**

**Causes: Driver collects large results (Expert)**

**Fix suggestions: reduce task number**

**9. Job type: User code (Mailing list)**

**Causes: Driver collects large results (User)**

**Fix suggestions: no**

**10. Job type: User code (Mailing list)**

**Causes: Driver collects large results (Expert)**

**Fix suggestions: no**

**11. Job type: User code (Mailing list)**

**Causes: Driver collects large results (Expert)**

**Fix suggestions: no**

**12. Job type: User code (Mailing list)**

**Causes: Driver collects large results (User)**

**Fix suggestions: no**

**13. Job type: MLlib (Mailing list)**

**Causes: Driver collects large results (User, reproduced)**

**Fix suggestions: use a small k (parameter)**

**14. Job type: MLlib (Mailing list)**

**Causes: Driver collects large results (User)**

**Fix suggestions: tree reduce**

**15. Job type: User-defined (Mailing list)**

**Causes: Driver collects large results (User)**

**Fix suggestions: no**