Elasticsearch Final Revision

Weicong HUANG, stalwarthuang@outlook.com

December 2024

These two tests are class tests for the course 《Big Data Processing Techniques - Elasticsearch》 at SCAU. They are only for personal research and study exchanges!

1 Class Test 1

1. 关于 Elasticsearch 的描述中, 不正确的是

   A. Elasticsearch 是解决海量数据全文检索的不二之选

   B. Elasticsearch 只能为结构化数据提供搜索和分析服务

   C. ES 是一个基于 Java 语言开发的, 基于 Lucene 的开源分布式搜索引擎

   D. 只要是用到搜索的场景, ES 几乎都可以说是最好的选择

   Elasticsearch 为所有类型的数据提供近乎实时的搜索和分析

2. ES 的典型应用场景不包括

   A. 在线实时日志分析

   B. 物联网数据监控

   C. 事务场景

   D. 文献检索和文献计量

3. 访问 ES 集群服务器的端口地址默认是

   A. 9200

   B. 9220

   C. 9020

   D. 9300

4. ES 集群中, 节点之间相互通信的默认端口号是

   A. 9200

   B. 9300

C. 9100

D. 9230

5. 搜索引擎中的反向索引 (倒排索引) 是指

A. 通过文章找词

B. 通过文章找文章

C. 通过词找词

D. 通过词找文章

6. 关于 ES 集群中索引和分片的描述, 不正确的是

A. 索引是由分片 (shards) 组成, 并且分片可以有副本

B. 分片的种类包括主分片和副本分片

C. 分片可以提高服务的高可用性

D. 主分片和其副本分片可以同时存在于同一个节点上

7. ES 集群节点的角色不包括

A. 主节点 (master node)

B. 数据节点 (data node)

C. 预处理节点 (ingest node)

D. 从节点 (slave node)

8. 关于索引的说法中, 不正确的是

A. 索引名称必须用小写字符

B. 索引的主分片数量定义后不能修改

C. 索引的副本分片数量定义后可修改

D. 从节点当向索引中添加数据的字段是原先未定义的, 数据不可以添加

当向索引中添加数据的字段是原先未定义的, 数据依然可以被成功添加。*ES* 拥有动态映射机制, 会根据数据的内容自动识别对应的字段类型。

9. 关于索引中文档主键的描述, 正确的是

A. 添加文档数据时如果没有指定主键, 则系统会生成一个不重复的字符串作为主键

B. ES 的索引中的文档, 可以没有主键

C. 写入索引数据时, 如果文档主键已存在则会报错

10. 关于索引的健康状态的描述, 正确的是

A. 蓝色表示所有分片（主、副本）都可用

B. 绿色表示至少有一个副本不可用, 但所有主分片都可用

C. 黄色表示至少有一个副本不可用，但所有主分片都可用

D. 红色表示至少有一个主分片不可用，数据不完整

- 绿色：所有分片都可用
- 黄色：至少有一个副本不可用，但是所有主分片都可用，此时集群能提供完整的读写服务，但是可用性较低。
- 红色：至少有一个主分片不可用，数据不完整。此时集群无法提供完整的读写服务。集群不可用

11. 当一个索引数据量太大时，继续写入可能导致分片数据量过大，查询时会因内存不足引起集群崩溃；为避免所有数据都写入同一个索引，可以使用 _____ 技术。
该技术需要配合索引别名一起使用，可实现把原先写入一个索引的数据自动分发到多个索引中。

A. 滚动索引

B. 索引模板

C. 动态映射

D. 字段复制

12. 文本分析的描述中，错误的是：

A. 文本分析器包含：大于等于零个字符过滤器、一个分词器、大于等于零个分词过滤器

B. 文档入库时，任何 text 类型字段都会进行文本分析

C. 检索已入库文档时，对于查询的字段会进行文本分析

D. 文本分析器就是文本分词器

13. 标准分析器 standard 包含

A. 标准分词器和一个小写分词过滤器

B. 简单分词器和一个小写分词过滤器

C. 只包含标准分词器，没有其他过滤器

D. 标准分词器和一个标点符号去除器

14. IK 中文分词器的描述中，不正确的是：

A. 分词两种：ik_smart、ik_max_word

B. 全文检索时文本分析使用 ik_smart 较为常见

C. 索引时文本分析使用 ik_max_word 更加合适

D. 同一中文句子使用 ik_smart 分词后得到的词数量一般比 ik_max_word 分词后得到的词数量多

15. 查询索引 xx 的映射的命令是：

A. get xx/_search

B. get xx/map

C. get xx/_mapping

D. get xx/mapping

16. 查看集群所有索引的命令是:

A. get _cat/indices

B. get cat/indices

C. get all/indices

D. get indices

17. 使用 Kibana 时, 浏览器端的默认端口号是:

A. 6501

B. 5600

C. 5601

D. 5061

18.
```
POST mydata/_search
{
  "query": {
    "match_all": {}
  },
  "size": 10,
  "from": 0
}
```

关于语句的描述中, 错误的是:

A. 索引名称是 mydata

B. _search 表示查询

C. match_all 表示查询所有文档

D. size 为 10 表示查询前 10 个文档

size 表示分页大小, "size":10 表示返回 10 个文档, 而并非前 10 个, 结合了"from":0 才是返回前 10 个

19.
```
POST _analyze
{
  "analyzer": "ik_smart",
  "text": "数据科学与大数据专业"
}
```

字符串"数据科学与大数据专业"使用 ik_smart 分词器得到的分词结果是:

A. 数据, 科学, 与, 大, 专业

B. 数据, 科学, 与, 大, 数据, 专业

C. 数据科学, 大数据, 专业

D. 数据, 科学, 专业

-未设置停用词, 所以包含“与”
-大数据可以拆分为“大”, “数据”
数据
科学
与
大
数据
专业

20. POST test-3-2-1/_search

```
{
  "query": {
    "term": {
      "name.keyword": {
        "value": "张三"
      }
    }
  }
}
```

查询语句的描述中, 错误的是:

A. term 表示术语查询

B. name.keyword 此处不可以改为 name

C. 该语句表示查询姓名为“张三”的文档, 其中包括“张三丰”的文档

D. 该语句是以一个精准查询, 而不是模糊匹配

term 精准查询”张三”, 不包括张三丰

2 Class Test 2

1. 关于 ES 特点的描述, 错误的是:

A. 基于 Java 语言开发

B. 基于 Lucene 框架

C. 原生支持分布式

D. 只支持 TB 级数据量

可支持 PB 级数据量

2. ES 集群的节点有多种类型，其中不包括：

A. master node

B. data node

C. input node

D. coordinating node

数据接入节点：ingest node

3. 关于分片的策略，错误的描述是：

A. 主分片和其副本分片不能同时存在于同一个节点上

B. 每个分片都是一个 Lucene 实例

C. ES 会自动在 nodes 上做分片均衡 shard rebalance

D. 完全相同的副本可以同时存在于同一个节点上

完全相同的副本"不能"不能同时存在于同一个节点上

4. 修改一个索引文档内容时，使用的 REST 方法是

A. POST

B. UPDATE

C. REPLACE

D. UPSERT

REST 方法包括：PUT,GET,POST,DELETE,HEAD。

5. 向索引输入一条文档数据时，其 REST 方法是

A. GET

B. POST

C. DELETE

D. PUTS

6. 索引数据批量录入时的指令方法是

A. BAT

B. BATS

C. BULK

D. INPUT

7. 定义一个字段的类型是不经过文本分析处理的关键字类型是

A. text 类型

B. key 类型

C. keytext 类型

D. keyword 类型

8. 关于索引路由的描述中，错误的是

A. 路由的计算公式与索引主分片数量无关

B. 路由的计算公式与索引主分片数量有关

C. 路由计算公式中的 _routing，默认是文档的 _id 值

D. 路由的本质就是计算新数据属于哪个分片

路由计算公式：$shard\_num = hash(\_routing)\%num\_primary\_shards$
其中 $\_routing$ 与主键有关，$num\_primary\_shards$ 表示主分片的数量。

9. 在 dev-tools 中，查询 ES 集群所有索引的命令正确的是

A. GET _cat/allindex

B. GET _cat/indices

C. GET _cat/index

D. GET _cat/allindices

10. 搜索数据时，术语查询是指 _____ 查询

A. match 查询

B. match_all

C. term

D. regexp

11. kibana 的默认 web ui 端口是

A. 6500

B. 6501

C. 5600

D. 5601

12. 在管道聚集中，如果要定义一个平均管道聚集，则其关键词是

A. sum_bucket

B. min_bucket

C. max_bucket

D. avg_bucket

13. 在父子关联索引数据设计方法中，关于 join 字段方法的描述中，错误的是

A. 父子关系数据写入同一个索引

B. 父子关系数据写入不同的两个索引

C. 父文档和子文档分开输入

D. 每个子文档都是独立的文档

14. 索引的健康状态显示黄色，描述错误的是

A. 有主分片没有分配，索引不可用

B. 主分片都已分配，但有副本分片没分配，索引可用

C. 所有主分片都已分配，索引可用

D. 黄色表示索引可用。当所有主分片和副本分片都分配后，索引状态变为绿色

黄色表示，亚健康，集群可用；
→ 所有主分片都已分配，有副本分片没有分配

15. 获取 ES 集群健康情况的操作指令是

A. GET _cat/index

B. GET _cat/nodes

C. GET _cat/health

D. GET _cat/master

16. 如果需要监控主机性能指标数据，则在 Beats 采集家族中选用的工具是

A. filebeat

B. metricbeat

C. packetbeat

D. Heartbeat

17. 判断一个索引 test_3 是否存在时，使用语句 HEAD test-3，如果返回结果是 200，则表示索引存在 **(T)**

18. 新建、索引和删除请求都是“写”操作，必须在主分片上面完成之后才能被复制到相关的副本分片 **(T)**

19. 根据文档 _id 检索文档时，可以从主分片或者从其它任意副本分片检索文档； **(T)**

20. 使用标准分析器处理文本时，不进行字母的小写转换 **(F)**

标准分析器 *standard analyzer* 会根据 *Unicode* 文本分割算法定义的单词边界将文本划分为术语。它删除了大多数标点符号并将单词转为小写，并支持删除停用词

21. 为某个字段指定 ignore_malformed 为 false 表示：即使该字段输入数据类型不匹配，不影响其他字段的输入 **(F)**

"ignore_malformed":true 才能保证输入数据类型不匹配时不影响其他字段输入

22. 向索引中写入一个文档包含事先未定义的字段内容时，写入操作不成功； **(F)**
ES 有动态映射功能

23. 管道聚集与桶聚集无关 **(F)**
管道聚集必须有一个输入才能正常工作，这个输入是桶聚集生成的结果，故管道聚集与桶聚集有关 (管道聚集第一步就是桶聚集)

24. 采用 Logstash 采集数据时，不添加额外的信息，数据原封不动输出下游 **(F)**
即使不添加额外信息，Logstash 也会自动添加一些元数据信息，比如 @timestamp @version 等等

25. beats 在数据收集层面上并不进行过于复杂的数据处理，只是将数据简单的组织并上报给上游系统 **(T)**

3 Aggregations

1. 既计算了总共的 sales 总和，也计算了每个桶内的 sales 总和。

```
POST index/_search
{
  "query": {
    "match_all": {}
  },
  "size": 0,
  "aggs": {
    "total_sales": {
      "sum": {
        "field": "sales"
      }
    },
    "group_by_category": {
      "terms": {
        "field": "category",
        "size": 5,
        "order": {
          "total_sales": "desc"
        }
      },
      "aggs": {
        "total_sales": {
          "sum": {
```

```
            "field": "sales"
          }
        }
      }
    }
  }
}
```