

2023 年上海市高等学校信息技术水平考试试卷

四级 大数据与云计算——数据开发与应用（模拟卷）

（本试卷考试时间 150 分钟）

一、单选题（本大题 30 道小题，每小题 1 分，共 30 分），从下面题目给出的 A、B、C、D 四个可供选择的答案中选择一个正确答案。

1. 在 Linux 操作系统中，__命令可以用来改变一个文件的权限。

- A. chmod
- B. chdir
- C. change
- D. file

2. Linux 上查看网络连接、路由表、接口统计使用的命令是__。

- A. netstat
- B. top
- C. ps
- D. iostat

3. NAT 工作在 IOS 模型中的__。

- A. 网络层
- B. 链路层
- C. 传输层
- D. 应用层

4. HTTPS 协议采用的默认 TCP 端口号是__。

- A. 443
- B. 80
- C. 22
- D. 3306

5. __不属于服务器安全测试的项目。

- A. 单元测试
- B. 泄露测试
- C. 接地电阻测试
- D. 耐压测试

6. __是公有云计算基础架构的基石。

- A. 分布式
- B. 虚拟化
- C. 并行计算
- D. 高可用

7. 大大降低用户在硬件上的开销，按需要租用相应的计算能力和存储能力，这是__的特

点。

- A. 基础设施即服务 (IaaS)
- B. 数据即服务 (DaaS)
- C. 平台即服务 (PaaS)
- D. 软件即服务 (SaaS)

8. 云计算按照提供的服务类型进行分类, 包括 IaaS、PaaS、__。

- A. SaaS
- B. Docker
- C. Xen
- D. KVM

9. 在 SQL 语句中, 用于表示任意字符的是__。

- A. %
- B. *
- C. LIKE
- D. _

10. 在 Oracle 中, 可用于提取日期时间类型特定部分 (如年、月、日、时、分、秒) 的函数是__。

- A. TRUNC
- B. DATEPART
- C. EXTRACT
- D. TO_CHAR

11. 有关系 S (SNO, SNAME, SEX), C (CNO, CNAME), SC (SNO, CNO, GRADE)。其中 SNO 是学生号, SNAME 是学生姓名, SEX 是性别, CNO 是课程号, CNAME 是课程名称。要查询选修“数据库”课的全体男生姓名的 SQL 语句是 SELECT SNAME FROM S, C, SC WHERE 子句。这里的 WHERE 子句的内容是__。

- A. S.SNO = SC.SNO and C.CNO = SC.CNO and SEX='男' and CNAME='数据库'
- B. S.SNO = SC.SNO and C.CNO = SC.CNO and SEX in '男' and CNAME in '数据库'
- C. SEX '男' and CNAME '数据库'
- D. S.SEX='男' and CNAME='数据库'

```
12. str = "Hello,Python";  
suffix = "Python";  
print (str.endswith(suffix,2));  
上述代码的输出结果是__。
```

- A. True
- B. False
- C. 语法错误
- D. P

13. 箱形图 (Box-plot) 又称为盒须图、盒式图或箱线图, 是一种用作显示一组数据分散情

况的统计图。常用于以下____分析。

- A. 品质管理
- B. 销售预测
- C. 库存管理
- D. 兴趣度用户管理

14. 作为一家制造类企业，下列____适合用于比较不同产品各年的销售变化情况。

- A. 分组柱形图
- B. 堆叠面积图
- C. 堆叠柱形图
- D. 分组饼形图

15. 去年人员更替主要集中在 30-35 岁的年龄段，体现了____相对关系。

- A. 频率分布
- B. 相关性
- C. 时间序列
- D. 时间

16. 表的两种相关约束是____。

- A. 主键和外键
- B. 主键和索引
- C. 内键和外键
- D. 外键和索引

17. 从宏观角度看，数据可视化不包括下列____功能。

- A. 信息的表示与变化
- B. 支持对信息的推理和分析
- C. 信息传播与协同
- D. 信息记录

18. Hive 数据仓库中，创建了以下外部表：

```
CREATEEXTERNAL TABLE sogou_ext (  
ts STRING, uid STRING, keyword STRING,  
rank INT, order INT, url STRING,  
year INT, month INT, day INT, hourINT  
)  
COMMENT 'This is the sogou search data of extend data'  
ROW FORMATDELIMITED  
FIELDSTERMINATED BY '\t'  
STORED ATEXTFILE  
LOCATION'/sogou_ext/20160508';
```

对于 keyword，给出其频度最高的 20 个词的 HQL 语句是_____。

- A. `select keyword from sogou_ext group by keyword order by order desc limit 20;`
- B. `select keyword from sogou_ext order by keyword order by order desc limit 20;`
- C. `select keyword from sogou_ext group by keyword order by group desc limit 20;`
- D. `select keyword from sogou_ext order by keyword order by order desc limit 20;`

19. 下来关于数据管理体系的建设目标描述，错误的是_____。

- A. 数据仓库设计是一个迭代的过程，因此重复建设是必须且必要的
- B. 构建统一、规范、可共享的全域数据体系
- C. 规避数据烟囱和不一致性，充分发挥大数据在海量和多样性方面的优势
- D. 避免数据不必要的冗余和重复建设

20. 下列不是使用 EChart 制作图表的时的步骤_____。

- A. 调用数据库
- B. 新建 HTML 页面，一般为 HTML5 页面
- C. 在 HTML 页面正文中用 JavaScript 代码实现图表显示
- D. 在 HTML 页面头部中导入 js 文件

21. 视图是从一个或多个表或视图中导出的_____。

- A. 表
- B. 查询
- C. 报表
- D. 数据

22. MongoDB 数据库的基本单位为_____。

- A. Doc
- B. Set
- C. Table
- D. elevation

23. 多维度数据模型简称维度模型，维度模型的核心概念是_____。

- A. 事实和维度
- B. 数量单位
- C. 向量坐标
- D. 数据表头

24. OLAP 的多维分析操作不包括_____。

- A. 抽出
- B. 上卷
- C. 钻取
- D. 旋转

25. 网页的解析方式不包括_____。

- A. Scrapy
- B. Xpath
- C. css
- D. bs4

26. 数据清洗的方法不包括_____。

- A. 逻辑错误处理
- B. 重复数据记录处理
- C. 缺失值处理
- D. 格式内容错误处理

27. 解决数据的质量问题包括确保数据的完整性、唯一性、权威性、合法性和_____。

- A. 一致性
- B. 安全性
- C. 逻辑正确性
- D. 可学习性

28. 假设属性 Age 的最大最小值分别是 0 岁和 100 岁。利用最大最小规范化的方法将属性的值映射到 0 至 1 的范围内, 属性 Age 的 27 岁将被转化为_____。

- A. 0. 270
- B. 0. 133
- C. 0. 729
- D. 1. 073

29. 下面_____不属于数据预处理的方法。

- A. 估计遗漏值
- B. 变量代换
- C. 聚集
- D. 离散化

30. 在数据中遇到缺失值时, 应对缺失值的方法不包括_____。

- A. 极值填补
- B. 极大似然估计
- C. 删除含有缺失值的个例
- D. 均值插补

二、多选题 (本大题 5 道小题 , 每小题 2 分, 共 10 分), 从下面题目给出的 A、B、C、D 四个可供选择的答案中选择所有正确答案。

1. 下列属于热插拔设备的是__。

- A. 硬盘
- B. 电源

- C. PCI 卡
- D. 机箱风扇

2. 云管理平台的作用包括__。

- A. 将各种接口、工具和流程进行组合以提供服务
- B. 以软件和硬件相结合的方式提供服务
- C. 各种自动化的工作流程
- D. 提供云资源的监控、运维、计费等功能

3. 如果要修改元数据合并触发的条件，应该修改下列_____配置。

- A. dfs.namenode.checkpoint.period
- B. dfs.namenode.checkpoint.txns
- C. dfs.metadata.checkpoint.txns
- D. dfs.metadata.checkpoint.period

4. 数据仓库的常见模型有_____。

- A. 星型模型
- B. 雪花模型
- C. 星座模型
- D. 空间模型

5. 国产数据可视化工具有_____。

- A. FineReport
- B. 魔镜 BI
- C. Tableau
- D. PowerBI

三、是非题（本大题 5 道小题，每小题 2 分，共 10 分）。

1. 一般而言，传统软件开发是业务（功能）驱动的，这就使得数据具有一定的开放性。
2. 当两个变量的数据点在散点图上不呈现出直线带状，且无任何其他规律，可以判断两变量不相干。
3. 在关系型数据库的同一个索引的不同类型中，相同名称的字段可以有不同的映射。
4. 在数据库的关系模型中，一个关键字可由一个或多个能惟一标识该关系模式中任何元组的属性组成。
5. 当某一视图的基本表被删除后，由该视图导出的其他视图也将被自动删除。

四、填空题（本大题 5 道小题，每空 2 分，共 10 分）。

1. 大数据的计算模式包括___计算、流处理计算、图计算和查询分析计算。
2. 简单地将数据对象集划分成不重叠的子集，使得每个数据对象恰好在一个子集中，这种

聚类类型称作_____。

3. 有甲乙两种糖果，原价分别为每千克 a 元和每千克 b 元，根据调查，将两种糖果按照甲种糖果 x 千克和乙种糖果 y 千克的比列混合，能取得较好的销售效果。现在糖果价格有了调整，甲种糖果单价下降 15%，乙种糖果单价上涨 20%，但按照原比列混合的糖果单价恰好不变，则 x/y 等于：_____。

4. 如果很多爬虫同时对一个网站全速爬取，那么就是对网站进行了_____攻击。

5. 表的访问方式主要是两种：全表扫描和_____。

五、操作题

以下第（一）题（简答题）、第（二）题（综合题）题目请在文件“C:\KS\大数据与云计算-答题纸.docx”中作答！

（一）、简答题（本大题包括5道小题，每小题4分，共20分）

1. 请描述使用Pyecharts绘制图表的基本方法。

请在答题纸作答！此处答题一律无效！

2. 请描述使用Python3中的pandas库或绘图函数的调用语句。

请在答题纸作答！此处答题一律无效！

3. KETTLE中作业有哪两种执行方式？并说明其运行过程。

请在答题纸作答！此处答题一律无效！

4. 请给出KETTLE性能调优的4个开发注意事项。

请在答题纸作答！此处答题一律无效！

5. 简单描述数据仓库设计的典型步骤。

请在答题纸作答！此处答题一律无效！

（二）、综合题（共20分）

对于大数据的应用开发，我们需要不同的技能，请完成下面任务。

任务1: 根据下面表格中的数据编写SQL语句。计算每个产品购买金额的最大证券，表名data，展示字段为：日期、产品、证券、购买数据和购买金额。

请在答题纸作答！此处答题一律无效！

任务2: 按月份统计产品的购买数量和购买金额，展示内容为：月份、产品、购买数量和购买金额。

请在答题纸作答！此处答题一律无效！

TDATE (日期)	ACC_CD (产品)	INS_CD (证券)	QUANTITY (购买数量)	DIRTYMKTVALUE (购买金额)
2021-01-03	A00002	1100233SH	1000	12300
2021-01-19	A00002	8300002SH	57000	701100
2021-02-11	L00033	8300002SH	211000	2595300
2021-02-19	L00033	8300002SH	225000	2767500
2021-02-03	A00002	8300002SH	71000	873300
2021-02-11	A00002	8300002SH	85000	1045500
2021-01-07	A00002	1100233SH	15000	184500
2021-01-11	A00002	1100233SH	29000	356700
2021-01-15	A00002	1100233SH	43000	528900
2021-01-19	L00033	8300002SH	183000	2250900
2021-02-03	L00033	8300002SH	197000	2423100
2021-02-27	L00033	8300002SH	239000	2939700
2021-02-19	A00002	8300002SH	99000	1217700
2021-02-27	A00002	8300002SH	113000	1389900
2021-01-03	L00033	1100233SH	127000	1562100
2021-01-07	L00033	1100233SH	141000	1734300
2021-01-11	L00033	1100233SH	155000	1906500
2021-01-15	L00033	1100233SH	169000	2078700

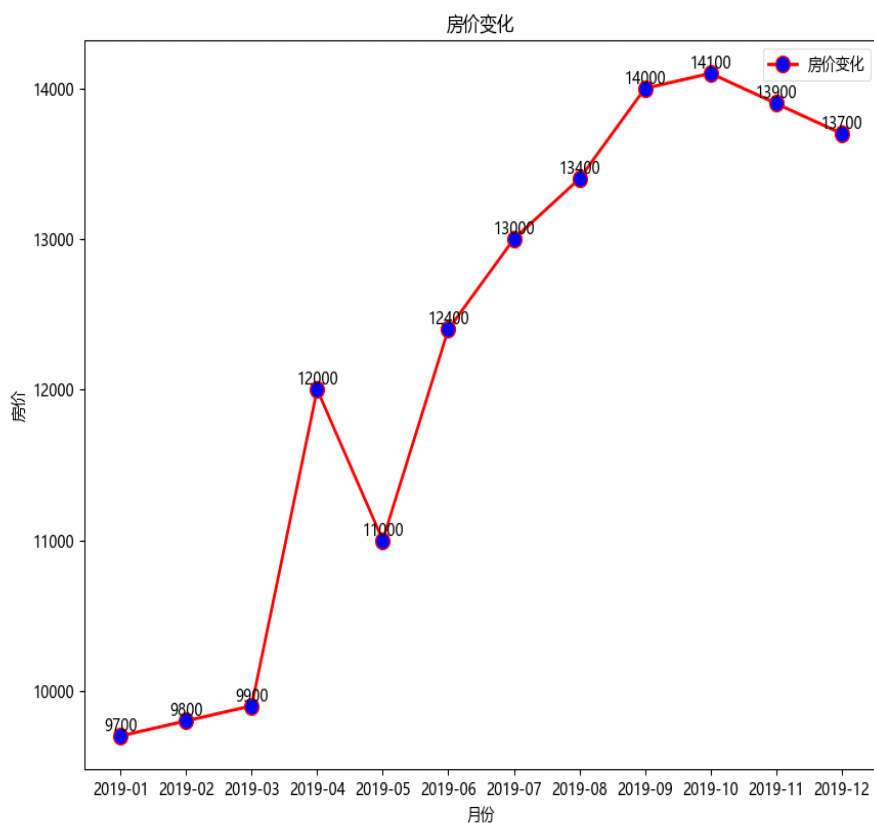
任务3: KETTLE是一个常见的ETL工具,请简单描述KETTLE的各个核心组件的功能。

请在答题纸作答! 此处答题一律无效!

任务4: 有房价数据如下:

月份	房价
2019-01	9700
2019-02	9800
2019-03	9900
2019-04	12000
2019-05	11000
2019-06	12400
2019-07	13000
2019-08	13400
2019-09	14000
2019-10	14100
2019-11	13900
2019-12	13700

利用Python语言完成如下图形的制作:



请在答题纸作答！此处答题一律无效！

上海市教育考试院