#### 三级《数据管理与分析技术》考试大纲（2019版）

##### 考试目标

掌握关系数据库理论和基本应用，掌握大数据基本概念、Hadoop与Spark的核心思想、分布式计算模型，具备基于关系数据库和大数据的数据管理和应用分析能力。

##### 基本要求

1、掌握数据库的基本概念；

2、掌握关系模型、关系模型完整性约束和函数依赖范畴的规范化理论；

3、熟练掌握关系数据库设计方法：包括ER模型的设计、ER模型向关系模型的转换与优化、数据库逻辑结构设计；

4、熟练掌握在MySQL平台的基本用户管理、基本权限管理、SQL基本指令的应用；

5、熟练掌握MySQL存储过程和触发器的设计，理解数据库事务与并发控制机制；

6、理解大数据技术Hadoop与Spark的核心思想与各自的特点；Spark与Hadoop功能定位的区别与联系；理解Spark优于Hadoop的核心技术（RDD，DAG，内存计算，惰性求值）；

7、掌握HDFS分布式文件系统与MapReduce计算模型；

8、熟练掌握常用的Linux命令行操作和Hadoop指令操作；

9、熟练掌握Hadoop分布式计算常见任务的MapReduce实现方法；

10、熟练使用交互式spark-shell编写分布式机器学习的任务。

##### 考试内容

一、关系数据库应用（40%）

1、数据库的基本概念：数据、数据库与数据处理、数据库系统的组成、结构与发展历史；数据库三级模式结构的概念、数据逻辑独立性与物理独立性的含义。

2、数据库系统的数据模型类别：层次、网状、关系、面向对象数据库、NoSQL数据库模型的特点、区别与典型DBMS产品。

3、关系数据库的完整性约束类别：主键约束、外键约束、数据类型约束、(Not)Null约束、Check约束；掌握关系数据理论的函数依赖、1NF、2NF、3NF、BCNF的定义和判断方法。

4、结构化查询语言SQL的基本应用：数据库对象定义语言（DDL）（数据类型、库的创建与删除、表的创建、修改与删除、视图的创建与删除、索引的创建与删除）、数据库查询（单表查询、单表自身连接查询、2表或3表的连接查询(等值连接、自然连接、左外、右外、全外连接的select实现)、不相关与相关嵌套子查询、分组统计查询、查询结果排序）、数据更新（表数据的插入、删除和修改）；其中不相关嵌套子查询要求掌握in、any、all的应用，相关嵌套子查询要求掌握exists谓词的简单应用。

5、MySQL存储过程和触发器的设计：MySQL平台下不带参数、带若干in、out参数的存储过程的设计及触发器的设计；存储过程的调用方法和触发器的测试；理解事务的定义、ACID特性与多用户数据库系统的封锁并发控制技术的基本原理。

6、MySQL用户及权限基本管理：新用户的创建、table对象的授权（select、insert、update、delete权限）。

7、简单数据库应用的数据库设计：需求描述、ER图设计、ER图向关系模型的转换方法、数据模型优化、设计视图、逻辑设计、物理设计；ER图包括实体、属性、联系（1对1、1对多、多对多）、参与度约束（最小min、最大max）的表示方法与含义。

二、大数据管理与分析技术（60%）

1、大数据基本概念：大数据的4V特征、类型（结构化与非结构化大数据）、核心技术（分布式存储和分布式处理）、大数据计算模式（批处理计算、流计算、图计算、查询分析计算）、每类计算模式典型的代表产品。

2、Hadoop框架基础理论： Hadoop特性、核心模块与相应的主要功能（HDFS分布式文件系统、MapReduce计算模型）。

(1)HDFS文件系统基本内容：体系结构、HDFS实现的目标与局限性、HDFS的NameNode和DataNode的功能与模块（NameNode：FsImage与EditLog；DataNode：数据存储与检索）。

(2)MapReduce计算模型基本内容：体系结构（Client、JobTracker、TaskTracker以及Task）、优势（容错性好、硬件要求低、编程难度低、使用场景多等）、设计策略（分而治之、计算向数据靠拢、Master/Slave架构）。

(3)Map/Reduce的输入/输出和工作流程：Input->Map->Reduce->Output。

3、常用的Linux命令行工具和Hadoop操作：

(1)Linux常用操作：cd、mkdir、rmdir、cp、mv、rm、cat、more、head、tail、touch、chown、chmod、find、tar、grep；

(2)Hadoop常用操作：Hadoop启动（所有进程启动、单进程启动）、查看目录（hdfs dfs –ls)、打开文件（hdfs dfs –cat)、本地文件或目录上传到Hadoop（hdfs dfs –put）、从Hadoop下载至本地目录（hadoop dfs -get）、删除Hadoop上文件夹或文件（hdfs dfs –rm||-rmr)、在Hadoop指定目录内创建新目录（hdfs dfs –mkdir）、将Hadoop某个文件重命名（hdfs dfs –mv）、将Hadoop指定目录下所有内容保存为一个文件，同时download至本地（hdfs dfs –getmerge）、kill正在运行的Hadoop作业（hadoop job –kill）、查看PATH目录信息（hdfs dfs –count）、显示文件的内容（hdfs dfs –text）、查看帮助（hdfs dfs -help）。

4、经典统计算法（去重、计数、排序、TopK排序、求最大最小值）与关系运算（选择、投影、分组）的MapReduce实现。

5、Spark基本概念：Spark核心技术（RDD：弹性分布式数据集，RDD的两类操作：Transformation和Action，有向无环图 DAG、内存计算技术、惰性计算）、Spark特性（速度快、丰富的API、高容错性、部署方式多样化）、Spark的架构（驱动器程序、SparkContext对象、集群管理器（Cluster Manager）、工作节点）。

6、Spark应用程序的运行架构与运行流程（集群管理器（Cluster Manager），多个工作节点（Worker Node），每个应用的任务控制节点（Driver）和每个工作结点上负责具体任务执行进程（Executor））。

7、使用Spark MLib库进行机器学习（特征提取、统计、分类、回归、聚类、协同过滤）。

8、Spark经典应用场景分析：SQL查询，文本处理，分析，音乐、视频、广告精准推荐，实时数据分析。