习题答案

习题 1

一、选择题

1～6　CABCAA

二、填空题

1．Hadoop
2．查询引擎
3．复杂数据类型 数值型
4．tinyint

三、简答题

1．（1）Hive 和关系数据库存储文件的系统不同。Hive 使用的是 Hadoop 的 HDFS，关系数据库使用的则是服务器本地的文件系统。

（2）Hive 使用的计算模型是 MapReduce，而关系数据库使用的则是自己设计的计算模型。

（3）关系数据库都是为实时查询的业务进行设计的，而 Hive 则是为海量数据做数据挖掘设计的。Hive 的实时性很差，实时性的差别导致 Hive 的应用场景和关系数据库有很大的不同。

（4）Hive 很容易扩展自己的存储能力和计算能力，这是继承了 Hadoop 的特性，而关系数据库在这方面要比 Hive 逊色很多。

2．（1）词法分析和语法分析。使用 antlr 将 SQL 语句解析成抽象语法树。

（2）语义分析。从 MetaStore 中获取元数据信息，验证 SQL 语句中的表名、列名、数据类型。

（3）逻辑计划生成。生成逻辑计划得到算子树。

（4）逻辑计划优化。对算子树进行优化，包括列剪枝、分区剪枝、谓词下推等。

（5）物理计划生成。将逻辑计划生产出包含由 MapReduce 任务组成的 DAG 的物理计划。

（6）物理计划执行。将 DAG 发送到 Hadoop 集群进行执行。

（7）将查询结果返回。

3．（1）支持索引，加快数据查询。

（2）支持不同的文件存储类型，例如，纯文本文件、HBase 中的文件。

（3）将元数据保存在关系数据库中，大大减少了在查询过程中执行语义检查的时间。

（4）可以直接使用存储在 Hadoop 文件系统中的数据。

（5）内置大量用户函数 UDF 来操作时间、字符串和其他的数据挖掘工具；支持用户扩展 UDF 函数来完成内置函数无法实现的操作。

（6）类 SQL 的查询方式，将 SQL 查询转换为 MapReduce 的 Job 在 Hadoop 集群上执行。

习题 2

一、选择题

1～5　BCADA

二、填空题

1．MapReduce
2．Hive 用户接口
3．HiveQL
4．Hive CLI
5．表数据　元数据
6．外部表（External Table）
7．二进制

三、简答题

1．（1）Hive CLI（Hive Command Line，Hive 命令行）。
（2）HWI（Hive Web Interface，Hive Web 接口）。
（3）Hive 提供了 Thrift 服务，即 HiveServer。
2．（1）单用户模式
（2）多用户模式
（3）远程服务器模式
3．（1）从本地文件系统中导入数据到 Hive 表。
（2）从 HDFS 上导入数据到 Hive 表。
（3）从其他的表中查询出相应的数据并导入到 Hive 表中。
（4）在创建表的时候通过从其他的表中查询出相应的记录并插入到所创建的表中。
4．TextFile 格式、SequenceFile 格式、RCFile 格式、ORC 格式

习题 3

一、选择题

1～6　ABACAC

二、填空题

1．show tables;
2．select * from test1;
3．drop table test3;
4．external
5．分区列

三、简答题

1．create table test1
(id int,name string,age int,tel string)
row format delimited
fields terminated by ','
stored as textfile;

2．（1）只有逻辑视图，没有物化视图。
（2）视图只能查询数据，不能 Load/Insert/Update/Delete 数据。
（3）视图在创建的时候只是保存了一份元数据，当查询视图的时候，才开始执行视图对应的子查询。

3．create view emp_30000 AS
select * from employee
where salary>30000;

习题 4

一、选择题

1～4　DBAD

二、填空题

1．数据装载
2．数据
3．非分区表
4．local
5．overwrite
6．inpath 子句
7．insert 子句

三、简答题

1．具体实例如下所示：

```
insert overwrite table employees
partition(country='US',state='OR')
select * from employees_tmp et
where et.cnty = 'US' AND et.st = 'OR';
```

2．

```
create table employees_new
as select name, salary, address
from employees em
where em.state = 'US';
```

3．

```
Insert overwrite local directory '/tmp/employee'
select name, salary, address
from employees em
WHERE em.state = 'US'
```

习题 5

一、选择题

1～6　ABCDDB

二、填空题

1．要保存的列以及输出函数需要调用的一个或多个列
2．引号　string
3．NULL 引号
4．函数调用
5．bigint
6．true
7．where
8．内连接（inner join）
9．LEFT outer
10．union all

三、简答题

1．

```
hive> select count(*),avg(salary) from employees;
4 77500.0
2hive> select explode (subordinates) AS sub from employees;
Mary Smith
Todd Jones
Bill King
```

3．

```
hive> select name, address.street
> from employees where address.street rlike '.*(Chicago|Ontario).*';
Mary Smith      100     Ontario St.
Todd Jones      200     Chicago Ave.
```

4．

```
hive> select year(ymd),avg(price_close) from stocks
> where exchange='NASDAQ' and symbol='AAPL'
> group by year(ymd);
1984    25.578625440597534
1985    20.193676221040867
1986    32.46102808021274
    ...
```

习题 6

一、选择题

1～5　DDABD

二、填空题

1．create database financial;
2．show databases like 'f.*';
3．hadoop
4．load data local
5．insert overwrite local directory
6．移动数据
7．drop table　alter table
8．select
9．group by

三、简答题

1．支持的数据类型（书写不区分大小写）
tinyint：1 byte 有符号整数
smallint：2 byte 有符号整数
int：4 byte 有符号整数
bigint：8 byte 有符号整数
boolean：布尔类型
float：单精度浮点数
double：双精度浮点数

string：字符串

timestamp：整数、浮点数或字符串

binary：字节数组

timestamp：日期时间

2.（1）struct。和 C 语言中的 struct 对象类似，都可以通过“.”符号访问元素内容。例如，如果某个列的数据类型是 struct{first,last}，那么第一个元素可通过字段名.first 来引用。

（2）map。例如，有一个 map 的键值对为'first'->'name'，则可通过字段名['first']来访问该元素。

（3）array。数组值为['name']，那么第一个元素可通过数组名[0]来访问。

3.

表类型	语句	是否复制数据到 HDFS	删除表时是否删除数据
内部表	create table	是	是
外部表	create external table	否	否

习题 7

一、填空题

1．UDAF（用户自定义聚合函数）

2．CLASSPATH

3．name，value，extended

4．GenericUDF

5．多行

6．Object Inspectors

二、设计题

略