习题答案

第 1 章　HBASE 介绍

一、选择题

1．A　　2．B　　3．C　　4．C

二、简答题

1．答：

行式存储是指一行中的数据在存储介质中是连续存储的。

列式存储是指一列中的数据在存储介质中是连续存储的；

（1）行数据库适用于读取出少行、多列的情况；

列数据库相反，适用于读取出少列、多行的情况。

（2）列数据库可以节省空间，如果某一行的某一列没有数据，那么在列存储时，就可以不存储该列的值。

2．答：

使用 HBase 作为数据存储，捕获来自于各种数据源的增量数据。比如目前流行的 Kylin、阿里内部的日志同步工具 TT、图组件 Titan、日志收集系统 Flume 等。

3．答：

优点：

（1）高容错性。

（2）适合大数据的处理。

（3）流式文件写入。

（4）可构建在廉价机器上。

缺陷：

（1）不适合低延迟数据访问。

（2）无法高效存储大量的小文件。

（3）不支持多用户写入及任意修改文件。

第 2 章　HBase 模型和系统架构

一、选择题

1．A　　2．A　　3．A　　4．B

二、填空题

1．主从分布式　HDFS

2．Row Key

3．全表扫描

4．三

5．HRegion

三、简答题

1．答：

在表里面，每一行代表着一个数据对象，每一行都是以一个行键（Row Key）进行唯一标识的。HBase 中的行里包含一个 Key 和一个或者多个包含值的列。行键并没有什么特定的数据类型，以二进制的字节来存储。Row Key 只能由一个字段组成而不能由多个字段组合组成，HBase 对所有行按照 Row Key 升序排序，在设计 Row Key 时将经常一起读取的行放到一起。

2．答：

一个行键映射一个列族数组，列族数组中的每个列族又映射一个列标识数组，列标识数组中的每一个列标识又映射到一个时间戳数组，里面是不同时间戳映射下不同版本的值，但是默认取最近时间的值，所以可以看成是列标识和它所对应的值的映射。

3．答：

（1）容量巨大。

（2）面向列。

（3）稀疏性。

（4）数据多版本。

（5）可扩展性。

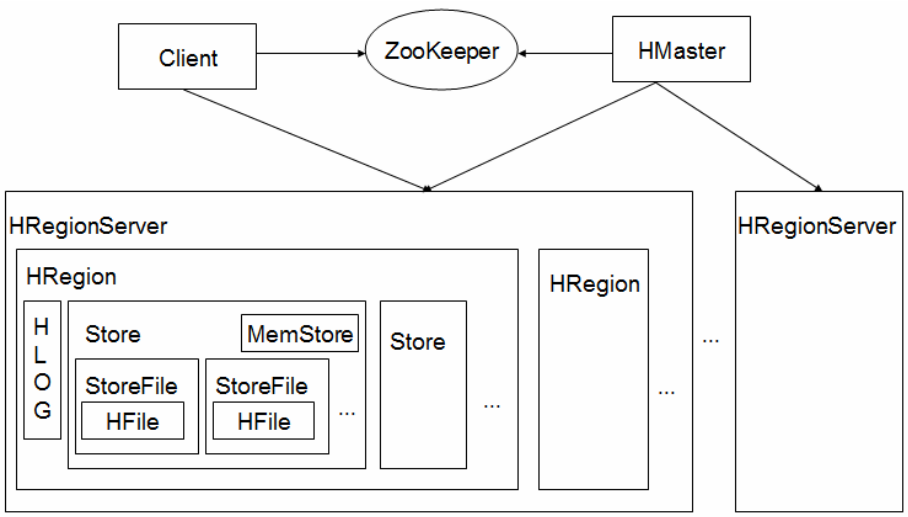
（6）高可靠性。

（7）高性能。

（8）数据类型单一。

4．答：

HBase 同样是主从分布式架构，隶属于 Hadoop 生态系统，由以下组件组成：Client、ZooKeeper、HMaster、HRegionServer 和 HRegion。在底层，它将数据存储在 HDFS 中，总体结构如图所示。



第 3 章　HBase 数据读写流程

一、选择题

1．C　　2．A　　3．D　　4．D　　5．D　　6．B

二、简答题

1．答：

WAL 即 Write Ahead Log，在早期版本中称为 HLog，它是 HDFS 上的一个文件，如其名字所表示的，所有写操作都会先保证将数据写入这个 Log 文件后，才会真正更新到 MemStore，最后写入 HFile 中。

2．答：

数据先顺序写入 WAL，再写入对应的缓存 MemStore，当 MemStore 中数据大小达到一定阈值（128MB）之后，系统会异步将 MemStore 中的数据 Flush 到 HDFS 形成小文件。

3．答：

（1）Client 访问 zookeeper，获取元数据存储所在的 regionserver。

（2）通过刚刚获取的地址访问对应的 regionserver，拿到对应的表存储的 regionserver

（3）去表所在的 regionserver 进行数据的读取。

（4）查找对应的 region，在 region 中寻找列族，先找到 memstore，找不到去 blockcache 中寻找，再找不到就进行 storefile 的遍历。

（5）找到数据之后会先缓存到 blockcache 中，再将结果返回。

第 4 章　HBase 环境搭建

一、选择题

1．A　　2．C　　3．B　　4．C

二、实验题

（略）

第 5 章　HBASE shell

一、选择题

1．A　　2．A　　3．B　　4．B　　5．A　　6．C　　7．D　　8．C　　9．C

二、实验题

（略）

第六章　HBASE 程序开发

一、选择题

1．C　　2．B　　3．A　　4．C　　5．D　　6．D　　7．C

二、实验题

（请参看正文相关程序）

第 7 章　HBase 高级特性

一、选择题

1．D　　2．D　　3．C

二、简答题

1．答：

通常情况下大多数业务都会开启 WAL 机制（默认），但是对于部分业务可能并不特别关心异常情况下部分数据的丢失，而更关心数据写入吞吐量，这类业务即使丢失一部分用户行为数据也并不对推荐结果构成很大影响，但是对写入吞吐量要求很高，不能造成数据队列阻塞。这种场景下可以考虑关闭 WAL，写入吞吐量可以提升 2～3 倍。

2．答：

（1）读请求是否均衡。

（2）BlockCache 是否设置合理。

（3）HFile 文件是否太多。

（4）Compaction 是否消耗系统资源过多。

3．答：

9TB/3/10GB=300 个。

三、设计题

学生表（Student）和课程表（Course）是多对多的关系。因此，可以转换为 HBase 中对应的表如下。

表 1　HBase_Student

	字段	说明
Row Key	S_No	学号，行键，逆排序
Column Family	Stuent:S_Name	学生姓名
	Stuent:S_Sex	学生性别
Column Family	Course: C_No	课程号
	Course:C_Name	课程名
	Course:C_Credit	学分
Column Family	SC：SC_Score	成绩

表 2　HBase_ Course

	字段	说明
Row Key	C_No	课程号，行键，逆排序
Column Family	Course:C_Name	课程名
	Course:C_Credit	学分
Column Family	Stuent: S_No	学号
	Stuent:S_Name	学生姓名
	Stuent:S_Sex	学生性别
Column Family	SC：SC_Score	成绩

第 8 章　MapReduce On HBase

一、选择题

1．D　　2．B　　3．A　　4．B　　5．B

二、简答题

答：

（1）Mapper 类从 HBase 读取数据，Mapper 继承的是 TableMapper 类。作用是：允许 MapReduce 应用，自动从指定的 HBase 表内按行读取数据进行处理。

（2）Reducer 类将数据写入 HBase，Reducer 继承的是 TableReducer 类。作用是：Reducer 输出的数据会自动插入 outputTable 指定的表内，把数据写回到 HBase。

三、实验题

（略）