分　数：	
评卷人：	

华中科技大学

研究生（梁稚媛）课程论文（报告）

题　目：**Albis:High-Performance File Format for Big Data Systems**

学　　号　M201873309

姓　　名　梁稚媛

专　　业　计算机技术

课程指导教师　曾令仿、施展

院（系、所）　计算机科学与技术学院

2018 年　**11** 月　**28** 日

1.论文作者：Animesh Trivedi, Patrick Stuedi, Jonas Pfefferle, Adrian Schuepbach, and Bernard Metzler, IBM Research, Zurich

2.论文发表时间：2018年7月

3.论文所属顶级会议：USENIX。2018 USENIX Annual Technical Conference (ATC)

USENIX成立于1975年,当时的名字叫做"Unix用户群"。主要目的是学习及开发Unix以及类似系统。1977年六月，更名成USENIX。USENIX逐渐发展成一个倍受尊敬的由计算机操作系统用户，开发者和研究者所组成的机构。USENIX每年赞助好几个学术会议和工作室会议，其中最有名的是USENIX操作系统设计和部署座谈会，USENIX联网系统设计和部署座谈会，USENIX安全座谈会，USENIX年度技术会议，USENIX文件和存储技术会议。

4.期刊最新进展：

该团队在这片文章后又发表了与存储相关的文章：Pocket: Ephemeral Storage for Serverless Analytics。

2018年4月，Anchal Gupta等人发表了Performance Analysis of RDBMS and Hadoop Components with Their File Formats for the Development of Recommender Systems。该方法使用四种查询工具对不同文件格式（如text,CSV,AVRO,PARQUET,RC和ORC）存储的数据进行查询，发现使用AVala查询工具在AVRO文件格式保存的数据集上效果最好。

5.该领域主要学者：IBM实验室苏黎世研究院、Anchal Gupta、Michael Stonebraker、Vitaliy Rudnytskiy等。

6.研究机构：IBM实验室

7.主要课题来源：IBM 实验室。IBM 实验室专门从事基础科学研究，并探索与产品有关的技术，创建于1911年，现已发展成为跨国公司。IBM在全球建有8个研发中心，该小组的主要课题来源于位于瑞士的苏黎世研究院，该研究院成立于 1956 年，主要研究领域包括存储系统、半导体技术、纳米科学及技术和业务优化等领域。

8.所属领域发展脉络：

该领域属于存储领域，而且主要是基于文件格式的存储领域。

Apache Hive 数据仓库软件可以使用 SQL 方便地阅读、编写和管理分布在分布式存储中的大型数据集。结构可以投射到已经存储的数据上。提供了一个命令行工具和 JDBC 驱动程序来将用户连接到Hive。

Hive 数据存储的文件格式很大程度上影响了数据的存储、查询效率。因此，实现一个优秀的文件格式存储数据是十分必要的。几种典型的 hive 文件存储格式：

Textfile:存储方式为行存储，可以直接存储，加载数据的速度最高，但它的磁盘开销较大，数据解析也很大，并且压缩的 txt 文件无法进行合并和拆分。一般不使用这种方式。

类似，还有 csv，csv 数据文件属于文本存储方式，spark 默认支持，按照行以文本的方式写到文件中，每行一条记录.一般来说文本存储方式无压缩，性能相对较差。

Sequencefile:存储方式为行存储，属于二进制文件，以<key,value>的形式序列化到文件中，它可以进行分割或压缩，优点是文件和 Haddop api 中的 mapfile 可相互兼容。

Rcfile:存储方式为按行分块，每块按照列存储。和本文的存储方式比较像。它压缩速度快，同时能实现快速的列存取，读取需要的列只需要读取每个 row group 的头部定义。但是读取全量数据的操作性能可能比 Sequencefile 没有明显的优势。

Orc:存储方式为数据按行分块，每块按照列存储。它的压缩速度快，且能实现快速列存取，它和 Rcfile 的存储方式一样，但效率比 Rcfile 高，是 Rcfile 的改良版本。

RC 是一种列式存储引擎，对 schema 演化（修改 schema 需要重新生成数据）支持较差，而 ORC 是对 RC 改进，但它仍对 schema 演化支持较差，主要是在压缩编码，查询性能方面做了优化。RC/ORC 最初是在 Hive 中得到使用，最后发展势头不错，独立成一个单独的项目。Hive 1.x 版本对事务和 update 操作的支持，便是基于 ORC 实现的（其他存储格式暂不支持）。ORC 发展到今天，已经具备一些非常高级的 feature，比如支持 update 操作，支持 ACID，支持 struct，array 复杂类型。你可以使用复杂类型构建一个类似于 parquet 的嵌套式数据架构，但当层数非常多时，写起来非常麻烦和复杂，而 parquet 提供的 schema 表达方式更容易表示出多级嵌套的数据类型。

Parquet 是 Hadoop 上的一种支持列式存储文件格式，Parquet 使用一些自动压缩技术，例如行程编码(run-length encoding,RLE) 和字典编码(dictionary encoding)，基于实际数据值的分析。一当数据值被编码成紧凑的格式，使用压缩算法，编码的数据可能会被进一步压缩。Impala 创建的 Parquet 数据文件可以使用 Snappy, GZip, 或不进行压缩；Parquet 规格还支持 LZO 压缩，但是目前 Impala 不支持 LZO 压缩的 Parquet 文件。

数据读取时，行存储通常将一行数据完全读出，如果只需要其中几列数据的情况，

就会存在冗余列，出于缩短处理时间的考量，消除冗余列的过程通常是在内存中进行的。列存储每次读取的数据是集合的一段或者全部，如果读取多列时，就需要移动磁头，再次定位到下一列的位置继续读取。 再谈两种存储的数据分布。由于列存储的每一列数据类型是同质的，不存在二义性问题。比如说某列数据类型为整型(int)，那么它的数据集合一定是整型数据。这种情况使数据解析变得十分容易。相比之下，行存储则要复杂得多，因为在一行记录中保存了多种类型的数据，数据解析需要在多种数据类型之间频繁转换，这个操作很消耗 CPU，增加了解析的时间。

相比传统的行式存储引擎，列式存储引擎具有更高的压缩比，更少的 IO 操作而备受青睐（注：列式存储不是万能高效的，很多场景下行式存储仍更加高效），尤其是在数据列（column）数很多，但每次操作仅针对若干列的情景，列式存储引擎的性价比更高。在互联网大数据应用场景下，大部分情况下，数据量很大且数据字段数目很多，但每次查询数据只针对其中的少数几行，这时候列式存储是极佳的选择。但是，如本文的 Albis 文件格式相比于 ORC、Parquet 两种列式存储在应对高速输入输出设备就取得了明显的效果。

而在 2017 年 1 月开源的 IndexR 也是一种很好的文件存储格式。IndexR 实现了一种可部署于分布式环境，可并行化处理，带索引的，列式的结构化数据格式。基于这种数据格式，IndexR 构建了一个数据仓库系统(Data Warehouse)，它基于 Hadoop 生态，可以对海量数据集做快速统计分析（OLAP），数据可实时导入并且对于查询零延迟。IndexR 为解决大数据场景下分析缓慢、数据延迟、系统复杂等问题而设计。IndexR 的存储格式是目前查询速度最快的开源大数据格式，扫描速度是 Parquet 的 2~4 倍，在添加索引之后查询速度普遍提升十几倍以上。适合于大数据的各种场景，包括离线和在线的各种统计分析，和快速过滤查询。它目前已经有很多的案例，如用于存放超大量（单表千亿级别）的复杂明细数据，做历史数据的明细查询。

研 究 生 签 字_____