2.1 SparkSql  
Spark SQL 是 Spark 处理结构化数据的程序模块。它将 SQL 查询与 Spark 程序无缝集成,可以将结构化数据作为 Spark 的 RDD 进行查询。RDD 的全称为 Resilient Distributed Datasets，即弹性分布式数据集，是 Spark 基本的数据结构。Spark 使用   
RDD 作为分布式程序的工作集合，提供一种分布式共享内存的受限形式。RDD 是只读的，对其只能进行创建、转化和求值等操作。SparkSQL作为Spark生态的一员继续发展，而不再受限于Hive，只是兼容Hive。我们利用hive作为数据源，spark作为计算引擎，通过SQL解析引擎，实现基于hive数据源，spark作为计算引擎的SQL测试方案。  
2.2 Presto  
Presto是一个分布式SQL查询引擎， 它被设计为用来专门进行高速、实时的数据分析。它支持标准的ANSI SQL，包括复杂查询、聚合（aggregation）、连接（join）和窗口函数（window functions)。作为Hive和Pig（Hive和Pig都是通过MapReduce的管道流来完成HDFS数据的查询）的替代者，Presto 本身并不存储数据，但是可以接入多种数据源，并且支持跨数据源的级联查询。Presto是一个OLAP的工具，擅长对海量数据进行复杂的分析；但是对于OLTP场景，并不是Presto所擅长，所以不要把Presto当做数据库来使用。  
<>2.3 Impala  
Impala 是 Cloudera 在受到 Google 的 Dremel   
启发下开发的实时交互SQL大数据查询工具，它拥有和Hadoop一样的可扩展性、它提供了类SQL（类Hsql）语法，在多用户场景下也能拥有较高的响应速度和吞吐量。它是由Java和C++实现的，Java提供的查询交互的接口和实现，C++实现了查询引擎部分，除此之外，Impala还能够共享Hive Metastore，甚至可以直接使用Hive的JDBC jar和beeline等直接对Impala进行查询、支持丰富的数据存储格式（Parquet、Avro等）。此外，Impala 没有再使用缓慢的 Hive+MapReduce 批处理，而是通过使用与商用并行关系数据库中类似的分布式查询引擎（由 Query Planner、Query Coordinator 和 Query Exec Engine 三部分组成），可以直接从 HDFS 或 HBase 中用 SELECT、JOIN 和统计函数查询数据，从而大大降低了延迟。  
2.4 HAWQ  
HAWQ 是一个 Hadoop 上的 SQL 引擎，是以 Greenplum Database 为代码基础逐渐发展起来的。HAWQ 采用 MPP 架构，改进了针对 Hadoop 的基于成本的查询优化器。除了能高效处理本身的内部数据，还可通过 PXF 访问 HDFS、Hive、HBase、JSON 等外部数据源。HAWQ全面兼容 SQL 标准，能编写 SQL UDF，还可用 SQL 完成简单的数据挖掘和机器学习。无论是功能特性，还是性能表现，HAWQ 都比较适用于构建 Hadoop 分析型数据仓库应用。  
2.5 ClickHouse  
Clickhouse由俄罗斯yandex公司开发。专为在线数据分析而设计。Yandex是俄罗斯搜索引擎公司。官方提供的文档表名，ClickHouse 日处理记录数"十亿级"。特性:采用列式存储；数据压缩；基于磁盘的存储，大部分列式存储数据库为了追求速度，会将数据直接写入内存，按时内存的空间往往很小；CPU 利用率高，在计算时会使用机器上的所有 CPU 资源；支持分片，并且同一个计算任务会在不同分片上并行执行，计算完成后会将结果汇总；支持SQL，SQL 几乎成了大数据的标准工具，使用门槛较低；支持联表查询；支持实时更新；自动多副本同步；支持索引；分布式存储查询。

2.6 Hive  
  
Hive是基于Hadoop的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供完整的sql查询功能，可以将sql语句转换为MapReduce任务进行运行。其优点是学习成本低，可以通过类SQL语句快速实现简单的MapReduce统计，不必开发专门的MapReduce应用，十分适合数据仓库的统计分析。Hive是建立在 Hadoop 上的数据仓库基础构架。它提供了一系列的工具，可以用来进行数据提取转化加载（ETL），这是一种可以存储、查询和分析存储在 Hadoop 中的大规模数据的机制。Hive 定义了简单的类 SQL 查询语言，称为 HQL，它允许熟悉 SQL 的用户查询数据。同时，这个语言也允许熟悉   
MapReduce 开发者的开发自定义的 mapper 和 reducer 来处理内建的 mapper 和 reducer 无法完成的复杂的分析工作。  
2.7 GreenPlum  
Greenplum是一个开源的大规模并行数据分析引擎。借助MPP架构，在大型数据集上执行复杂SQL分析的速度比很多解决方案都要快。GPDB完全支持ANSI SQL 2008标准和SQL OLAP 2003 扩展；从应用编程接口上讲，它支持ODBC和JDBC。完善的标准支持使得系统开发、维护和管理都大为方便。支持分布式事务，支持ACID。保证数据的强一致性。做为分布式数据库，拥有良好的线性扩展能力。GPDB有完善的生态系统，可以与很多企业级产品集成，譬如SAS，Cognos，Informatic，Tableau等；也可以很多种开源软件集成，譬如Pentaho,Talend 等。

Sparkql、Impala、Presto均采用的是hive元数据

多表关联presto、impala和hawq查询时间快于SparkSql和ClickHouse，性能约是SparkSql的2-3倍，其中尤其以Presto和Impala性能要好一些。greenplum在多表查询上也有不错的表现；ClickHouse对于多表join效果相比较于Presto、Impala、HAWQ不是很好，并且很多复杂语法支持的不够好，可见并不是为关联分析而设置

单表测试ClickHouse要比其余几种组件查询速度表现突出，测试性能约是其余四种的3-6倍。而Presto相比于HAWQ、Impala、SparkSql、GreenPlum在单表操作方面性能也稍好些。

SparkSQL是Hadoop中另一个著名的SQL引擎，它以Spark作为底层计算框架，Spark使用RDD作为分布式程序的工作集合，它提供一种分布式共享内存的受限形式。在分布式共享内存系统中，应用可以向全局地址空间的任意位置进行读写操作，而RDD是只读的，对其只能进行创建、转化和求值等操作。这种内存操作大大提高了计算速度。SparkSql的性能相对其他的组件要差一些，多表单表查询性能都不突出。  
 Impala官方宣传其计算速度是一大优点，在实际测试中我们也发现它的多表查询性能和presto差不多，但是单表查询方面却不如presto好。而且Impala有很多不支持的地方，例如：不支持update、delete操作，不支持Date数据类型，不支持ORC文件格式等等，所以我们查询时采用parquet格式进行查询，而且Impala在查询时占用的内存很大。  
 Presto综合性能比起来要比其余组件好一些，无论是查询性能还是支持的数据源和数据格式方面都要突出一些，在单表查询时性能靠前，多表查询方面性能也很突出。由于Presto是完全基于内存的并行计算，所以presto在查询时占用的内存也不少，但是发现要比Impala少一些，比如多表join需要很大的内存，Impala占用的内存比presto要多。  
 HAWQ 吸收了先进的基于成本的 SQL 查询优化器，自动生成执行计划，可优化使用hadoop 集群资源。HAWQ 采用 Dynamic pipelining 技术解决这一关键问题。Dynamic pipelining 是一种并行数据流框架，利用线性可扩展加速Hadoop查询，数据直接存储在HDFS上，并且其SQL查询优化器已经为基于HDFS的文件系统性能特征进行过细致的优化。但是我们发现HAWQ在多表查询时比Presto、Impala差一些；而且不适合单表的复杂聚合操作，单表测试性能方面要比其余四种组件差很多，hawq环境搭建也遇到了诸多问题。  
ClickHouse   
作为目前所有开源MPP计算框架中计算速度最快的，它在做多列的表，同时行数很多的表的查询时，性能是很让人兴奋的，但是在做多表的join时，它的性能是不如单宽表查询的。性能测试结果表明ClickHouse在单表查询方面表现出很大的性能优势，但是在多表查询中性能却比较差，不如presto和impala、hawq的效果好。

GreenPlum作为关系型数据库产品，它的特点主要就是查询速度快，数据装载速度快，批量DML处理快。而且性能可以随着硬件的添加，呈线性增加，拥有非常良好的可扩展性。因此，它主要适用于面向分析的应用。比如构建企业级ODS/EDW，或者数据集市等，GREENPLUM都是不错的选择。