招商局数据报表架构方案对比

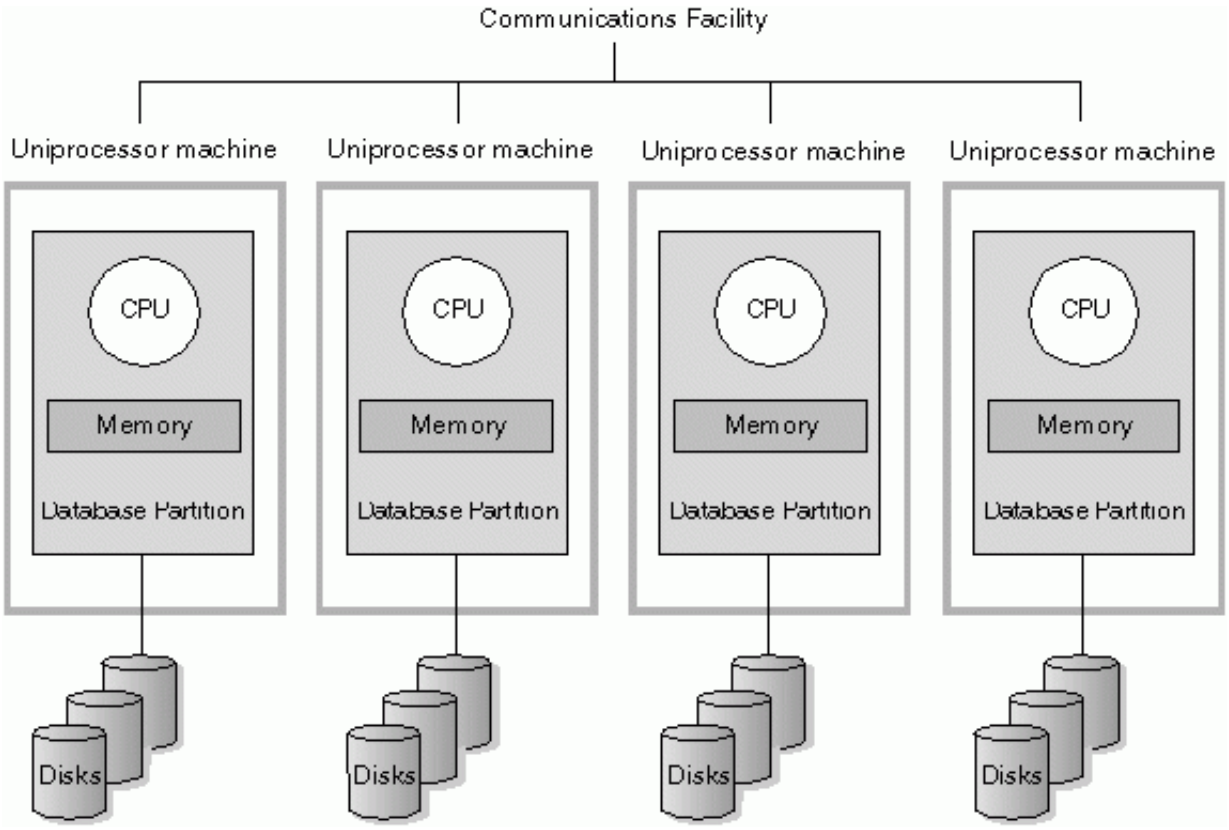
项目背景

对于招商局的数据报表的需求，如今整个团队面临的主要问题是大数据平台的搭建，基于交付日期的临近，考虑到指标统计可能遇到的性能问题、以及所搭建平台的适用性，同时技术的选型迫在眉睫。如今我们深圳大数据组面临两种架构方案的选择：

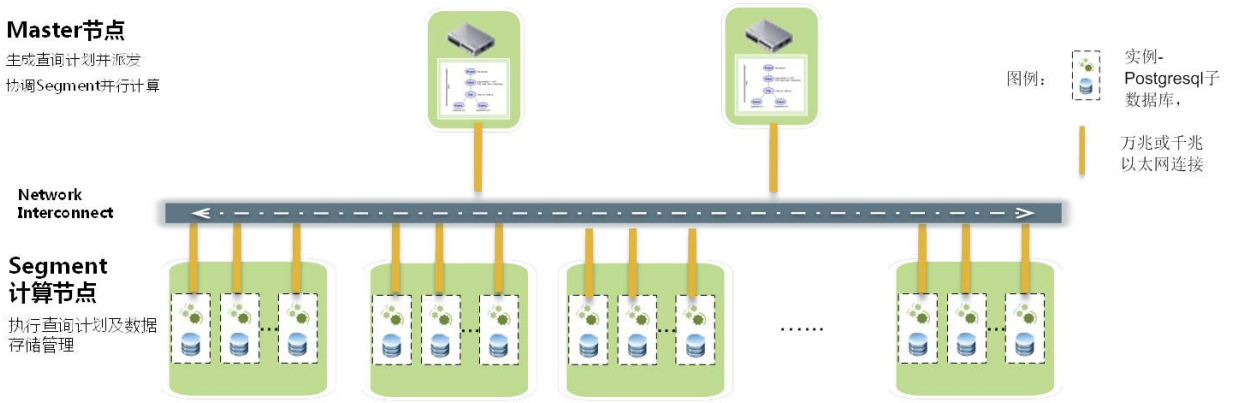
- 基于MPP的并行计算计算架构
- 基于Hadoop生态的大数据架构

MPP概念

MMP(Massively-Parallel-Process)代表大规模并行处理，当集群的所有单独节点都参与协调计算时，这是集群计算中的方法。MPP DBMS是基于此方法构建的数据库管理系统。在这些系统中，您正在注视的每个查询被分解为由MPP集群的节点并行执行的一组协调过程，从而以比传统SMP RDBMS系统更快的速度运行计算。该体系结构为您提供的另一个优势是可伸缩性，因为您可以通过在网格中添加新节点来轻松扩展集群。为了能够处理大量数据，这些解决方案中的数据通常按每个节点仅处理其本地数据的方式在节点之间拆分（分片）。这进一步加快了数据的处理速度，因为将共享存储用于这种设计将是一个巨大的过大杀伤力-更复杂，更昂贵，可伸缩性更低，网络利用率更高，并行性更低。这就是为什么大多数MPP DBMS解决方案都是不共享的，并且不能在DAS存储或在小型服务器组之间共享的一组存储架上工作的原因。 **Teradata**，**Greenplum**，**Vertica**，**Netezza**和其他类似解决方案都采用了这种方法。它们都具有专门为MPP解决方案开发的复杂成熟的SQL优化器。所有这些都可以通过内置语言和围绕这些解决方案的工具集进行扩展，无论是地理空间分析还是数据挖掘的全文搜索，这些工具集几乎可以满足任何客户的需求。它们都是开源的复杂企业解决方案，在该行业已经存在多年了，它们足够稳定，可以运行用户的关键任务工作负载。



MPP（大规模并行处理）是由在程序的不同部分上工作的多个处理器对程序进行的协同处理，每个处理器使用其自己的操作系统和内存。通常，MPP处理器使用某些消息传递接口进行通信。在某些实现中，同一应用程序上最多可以使用200个或更多处理器。数据路径的“互连”设置允许在处理器之间发送消息。通常，MPP的设置更为复杂，需要考虑如何在处理器之间划分公用数据库以及如何在处理器之间分配工作。MPP系统也称为“松散耦合”或“无共享”系统，大致概念如下图：

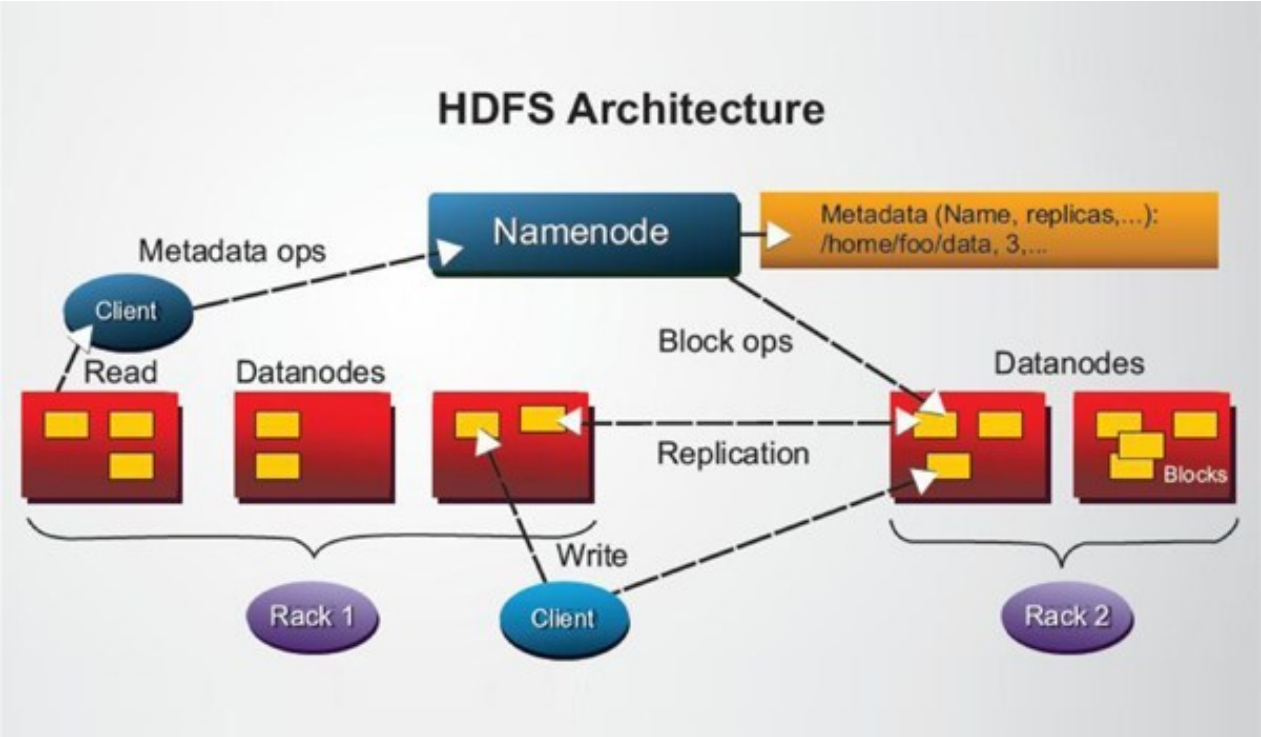


Hadoop概念

Hadoop简介

Hadoop不是一项单独的技术，而是一个相关项目的生态系统，它有其优点和缺点。最大的优点是可扩展性–出现了许多新组件（如Spark），并且它们与基础Hadoop的核心技术保持集成，这可以防止锁定，并可以进一步扩展集群用例。作为一个缺点，我可以说一个事实，那就是，您自己构建单独技术的平台是一项艰巨的工作，并且现在还没有人手动进行，大多数公司都在运行由Cloudera和Hortonworks搭建的平台。

Hadoop存储技术基于完全不同的方法。它不是根据某种密钥来分片数据，而是将数据分块为固定大小（可配置）的块，然后在节点之间进行拆分。这些块很大，它们以及整个文件系统（HDFS）都是只读的。简单来说，将小的100行表加载到MPP中会导致引擎根据表的键将数据分片，这样，在足够大的群集中，每个节点仅存储一个节点的可能性就很大。相反，在HDFS中，整个小表将写入单个块中，该块将表示datanode的文件系统上的单个文件。



集群资源管理

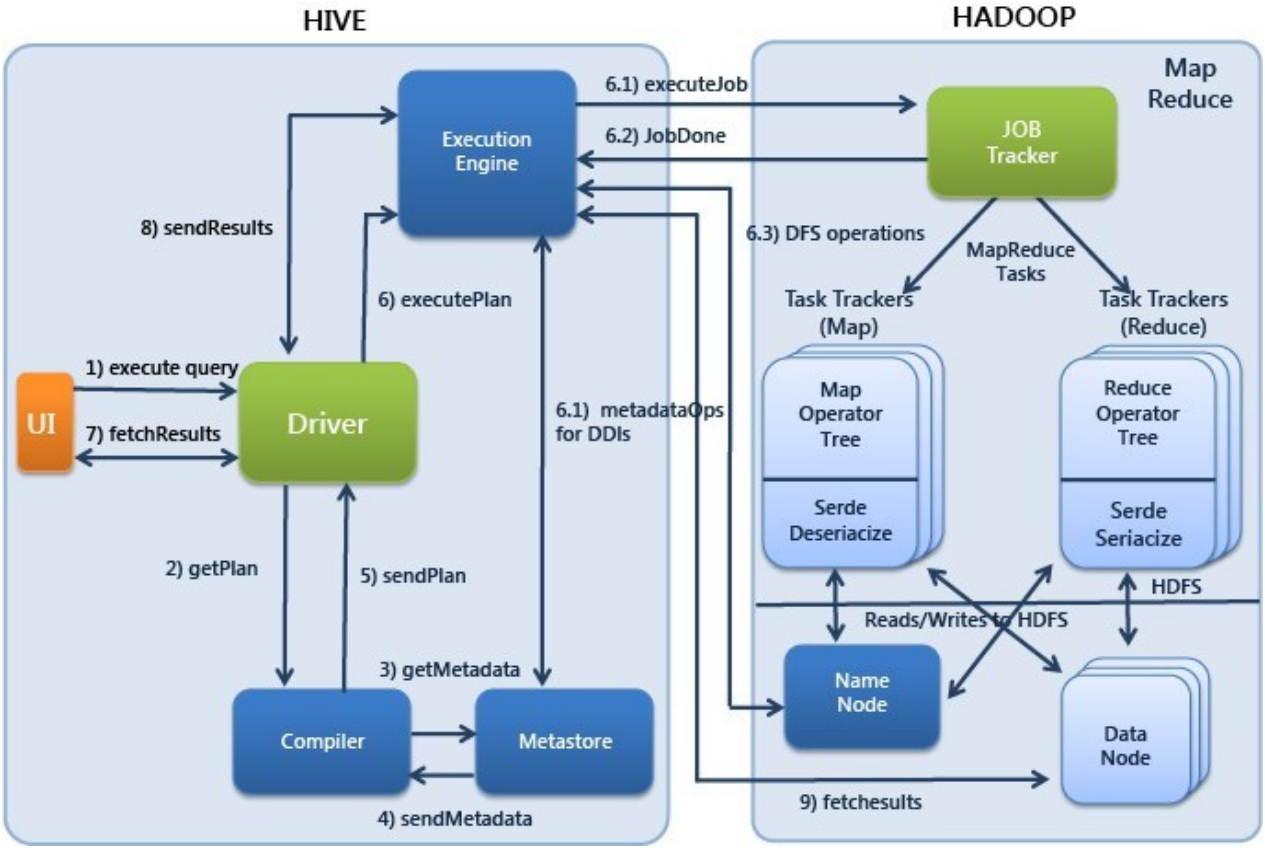
与MPP设计相比，Hadoop资源管理器（YARN）为您提供了更细粒度的资源管理；与MPP相比，MapReduce作业不需要并行运行所有计算任务，因此您甚至可以处理大量的如果完全利用了群集的其他部分，则在单个节点上运行的一组任务中的数据。它还具有一系列不错的功能，例如可扩展性，对长寿命容器的支持等。但是实际上，它比MPP资源管理器要慢，有时在并发性管理方面也不那么好。

Hadoop的SQL接口

Hadoop集成多种工具支持SQL接口，如：SparkSQL、Hive（MR、Tez、Spark）、Impala、Presto等等

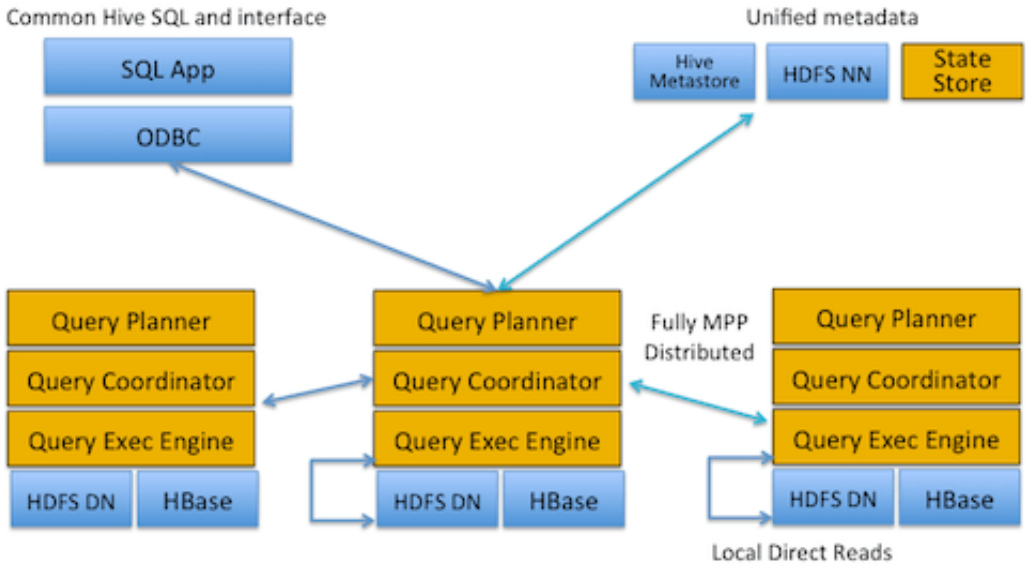
- **hive**

以Hive来说，它是将SQL查询转换为MR / Tez / Spark作业并在群集上执行它们的引擎。所有作业均基于相同的MapReduce概念构建，并为提供了良好的群集利用率选项以及与其他Hadoop堆栈的良好集成。但是缺点也很大-执行查询的延迟大，尤其是对于表联接的性能较低。



- **impala**

诸如Impala的解决方案处于这一优势的另一端，它们是Hadoop之上的MPP执行引擎，可处理HDFS中存储的数据。与其他MPP引擎一样，它们可以为您提供更低的延迟和更少的查询处理时间，但代价是可伸缩性和稳定性较低。



- **SparkSQL**

SparkSQL是介于MapReduce和基于MPP-over-Hadoop的方法之间的另一种野兽，它试图兼顾两者，并有其自身的缺点。与MR相似，它将工作分解为一组单独计划的任务，以提供更好的稳定性。与MPP一样，它尝试在执行阶段之间流式传输数据以加快处理速度。它还使用MPP熟悉的固定执行程序概念来减少查询的延迟。但是它也结合了这些解决方案的缺点：速度不如MPP，不如MapReduce稳定和可扩展。

MMP与Hadoop架构的比较

	MMP	Hadoop
平台开放性	不开源且专有，对于某些技术，非客户甚至无法下载文档	完全开放源代码，供应商和社区资源均可通过Internet免费获得
硬件选项	许多解决方案仅适用于设备，您无法在自己的群集上部署软件。所有解决方案都需要特定的企业级硬件，例如快速磁盘，具有大量ECC RAM的服务器，10GbE / Infiniband等。	任何硬件都可以使用，供应商提供了一些配置准则。大多数建议是将便宜的商品硬件与DAS结合使用
可伸缩性（节点）	平均数十个节点，最大100-200	平均100个节点，最大为数千个
可扩展性（用户数据）	平均值为数十TB，最大为PB	平均数以百亿兆字节为单位，最大为数十PB
查询延迟	10-20毫秒	10-20秒
查询平均运行时间	5-7秒	10-15分钟
查询最大运行时间	1-2小时	1-2周
查询优化	复杂的企业查询优化器引擎一直是最有价值的企业机密之一	没有优化器或功能有限的优化器，有时甚至都不基于成本
查询调试和分析	代表性查询执行计划和查询执行统计信息，解释性错误消息	OOM问题和Java堆转储分析，GC在集群组件上暂停，每个任务的单独日志为您提供了很多方便
技术价格	每个节点数十到十万美元	每个节点免费或高达数千美元

最终用户的可访问性	简单友好的SQL接口和简单可解释的数据库内函数	SQL并不完全符合ANSI，用户应注意执行逻辑，基础数据布局。函数通常需要用Java编写，编译并放在集群中
目标最终用户受众	业务分析师	Java开发人员和经验丰富的DBA
单作业冗余	低，MPP节点失败时作业失败	高，仅当节点管理作业执行时作业才会失败
目标系统	通用DWH和分析系统	专用数据处理引擎
最大并发	数十到数百个查询	最多10-20个job
技术可扩展性	仅使用供应商提供的工具	引入的任何全新开放源代码工具
使用要求	普通RDBMS DBA	一流的Java和RDBMS背景
方案实施的复杂性	中等	高