印度版的“大众点评”如何将 Food Feed 业务从 Redis 迁移到 Cassandra

Zomato 是一家食品订购、外卖及餐馆发现平台，被称为印度版的“大众点评”。目前，该公司的业务覆盖全球24个国家（主要是印度，东南亚和中东市场）。本文将介绍该公司的 Food Feed 业务是如何从 Redis 迁移到 Cassandra 的。

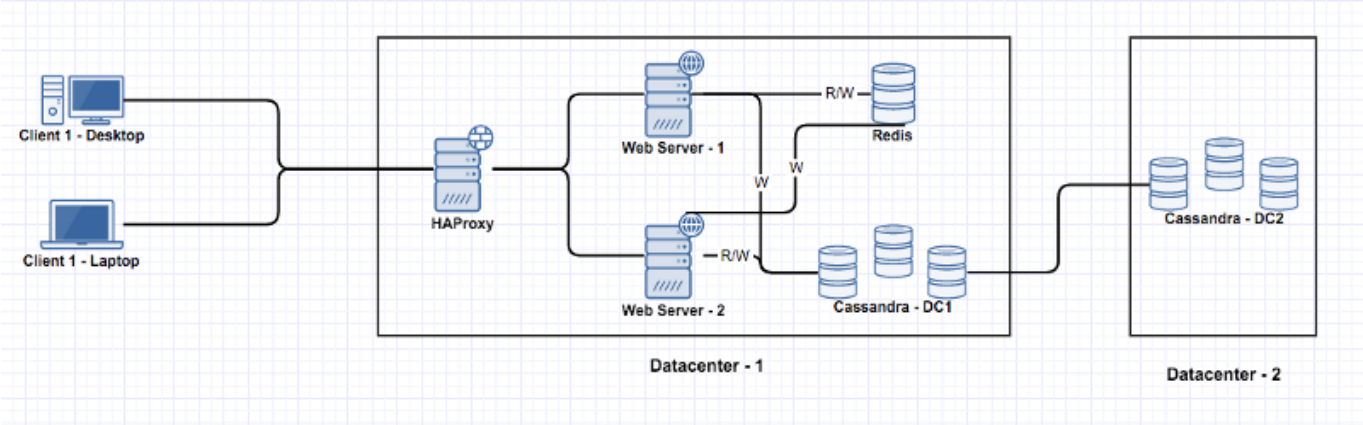
Food Feed 是 Zomato 社交场景中不可或缺的一部分，因为它可以让我们的用户参与其中并与朋友的餐厅评论和图片保持同步，甚至可以通过这个获取餐厅提供的优惠和折扣。开始我们选择 Redis 作为消息 Feed 流的存储引擎，因为在当时的用户场景这是最好的选择。但是随着业务的发展，我们需要更高的可用性和负载支持，而 Redis 不能很好的满足这个需求。虽然我们可以通过丢失一些数据来避免系统的中断，但这不是我们想做的事情。为了确保我们的系统具有高可用性，我们不得不放弃 Redis，而选择 Cassandra 作为其替代品。

Cassandra 非常适合这个用例，因为它是分布式的，就有高可用性等。并且 Cassandra 也可以用于存储时间序列数据 - 这实际上就是我们的Feed 流。在做出这一重大改变之前，我们确实有一些 Cassandra 的使用经验，但对于像 Feed 这样重要的东西来说肯定是不够的。我们必须弄清楚如何顺利的从 Redis 过渡到 Cassandra，并像在 Redis 上那样有效地运行 Feed，并且没有停机时间。

我们开始花时间在 Cassandra 上，在前两周深入探索其配置并调整它以满足我们的要求。接下来，在最终确定 Feed 的架构之前，我们明确了一下两个情况：

- Feed 流信息一般只用于读取而基本上不会修改。使用 Redis 的时候，我们可以同时读取上百个 keys 而不必担心读取延迟，但是对于Cassandra 而言，连接延迟可能是读取请求过程中一个相当重要的部分。
- schema 需要足够灵活，以便将来允许 Feed 中新类型的数据。鉴于我们不断迭代并致力于丰富产品体验，因此在 Feed 中添加元素和功能几乎是不可避免的。

我们花了几天时间用于收集了我们项目的数据模式以及各种用户案例，然后开始使用2个数据中心，每个数据中心有3个节点。 我们从 Redis 中迁移大概 6000万条记录到 Cassandra 中用于测试其性能。由于是测试阶段，我们只将一部分流量切入到 Cassandra ，并准备了两个版本的代码，分别写入到 Cassandra 和 Redis 。架构图如下

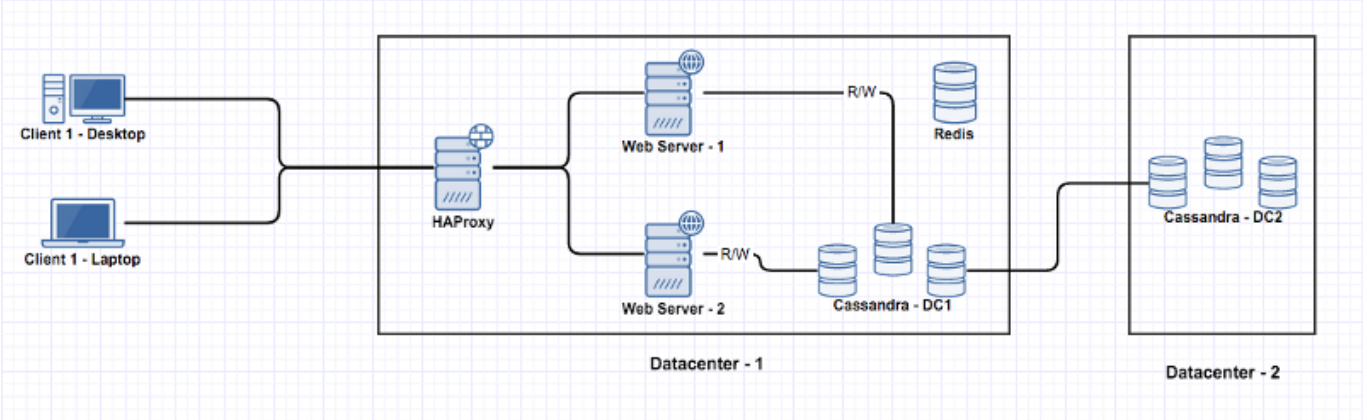


我们监控系统的延迟和其他问题，令人惊讶的是，我们遇到了写入吞吐量的瓶颈问题。 我们知道 Cassandra 的写入能力非常强，但是我们无法实现我们在各种博客文章和文章中阅读的写入吞吐量。 我们知道出了什么问题，但我们不知道是什么。我们从三个节点中获得的最佳结果是每秒1500次写入，这完全不能满足线上的需求，我们不得不在几个小时内回滚并重新评估。

经过几天的排查，我们意识到问题不在于 Cassandra，而在于 Elastic Block Store（EBS）。EBS是安装在每个EC2实例上的网络驱动器，具有10 Gigabits 的共享带宽和网络流量。当在单个EC2实例上的所有用户之间共享时，该带宽成为我们的瓶颈。为了满足这一需求，我们将数据从基于网络的EBS存储移动到同一EC2实例中的磁盘存储。然后我们在每个服务器上逐个部署由 Cassandra 提供支持的新 Food Feed，以便我们控制吞吐量。很高兴的是，这次成功了。

然后我们开始从我们的生产 Redis 服务器迁移数据（我们花了14个小时来迁移所有内容），在迁移过程中我们没有任何故障或额外负载。这就是 Redis 和 Cassandra 的强大功能。今天，我们的 Food Feed 完全运行在 Cassandra

上，我们在没有停机的情况下完成了这项工作。新的架构如下：



总而言之，通过上面这个项目，我们学到了以下几点：

- 在写入期间避免数据的读取。“读取”吞吐量大致保持不变，而“写入”规模与节点数量成比例；
- 避免数据的删除。删除意味着压缩（compaction），当它运行时，节点的资源会被占用；
- 延迟是一个问题。与Redis相比，Cassandra的连接延迟很高，大约是 Redis 的10x-15x。

微信公众号和钉钉群交流

为了营造一个开放的 Cassandra 技术交流，我们建立了微信公众号和钉钉群，为广大用户提供专业的技术分享及问答，定期在国内开展线下技术沙龙，专家技术直播，欢迎大家加入。

如果想及时了解Spark、Hadoop或者HBase相关的文章，欢迎关注微信公众号：iteblog_hadoop