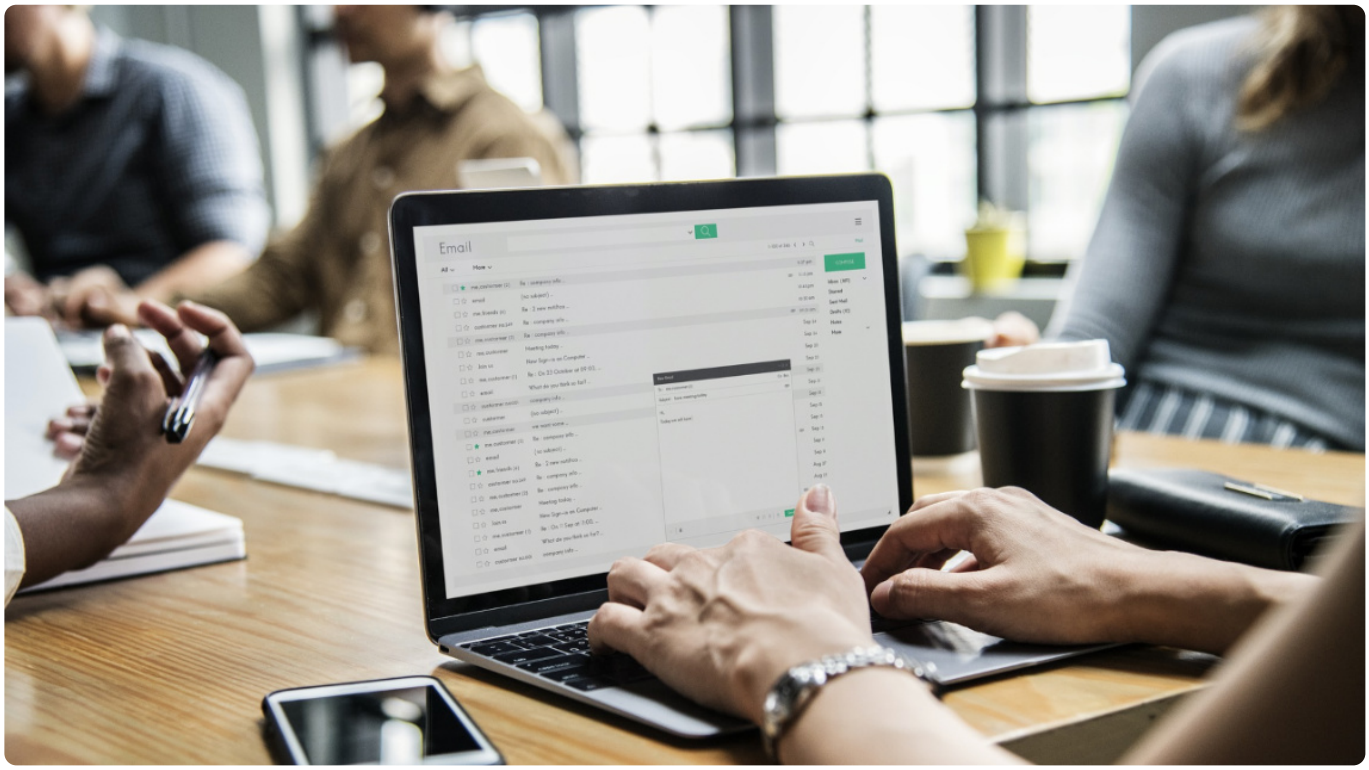
21 | 从阿里内部产品看海量数据处理系统的设计（上）：Doris的立项

2018-12-15 李智慧

从0开始学大数据 进入课程 >



讲述：李智慧
时长 08:45 大小 8.03M

从今天开始，我会分两期内容来讨论阿里巴巴的一个海量数据处理系统的设计，这个系统的名字叫 Doris，它是阿里巴巴的一个内部产品。前面专栏曾经提到过，2010 年前后是各种 NoSQL 系统爆发的一个时期，各种开源 NoSQL 在这个时期发布出来，当时阿里巴巴也开发了自己的 NoSQL 系统 Doris。

Doris 的设计目标是支持海量的 KV 结构的数据存储，访问速度和可靠性要高于当时主流的 NoSQL 数据库，系统要易于维护和伸缩。和当时众多 NoSQL 系统相比，Doris 在架构设计上颇具独特，路由算法、失效转移、集群扩容也有自己的创新之处，并成功申请三项技术专利。

在我们开始讨论 Doris 项目前，我想先跟你聊聊大公司是如何看待内部技术产品这件事。事实上，阿里巴巴内部底层技术产品的研发决策思路也颇有值得借鉴之处，你可以吸收其中好的经验，并把它转化到你所开发的产品上。

我们知道一家互联网公司主要靠自己的互联网产品盈利，比如阿里巴巴主要靠淘宝、天猫、阿里巴巴 B2B 网站等产品赚钱，而公司的工程师主要也是开发这些产品，但是这些产品通常都需要处理海量的用户请求和大规模的数据存储，所以在系统底层通常用到很多基础技术产品，比如分布式缓存、分布式消息队列、分布式服务框架、分布式数据库等。这些基础技术产品可以选择开源技术产品，也可以选择自己研发。自己研发的优点是可以针对业务场景进行定制开发，同时培养提高自己工程师的技术实力；缺点是投入大、风险高。

通常公司到了一定规模，都会开始逐渐自主研发一些基础技术产品，既可以提升自己的产品研发能力，又可以提高自身在业界的地位，吸引更优秀的人才并提高竞争门槛，形成自己的竞争壁垒。

但是公司的资源毕竟是有限的，主要的资源又投入到业务产品开发去了，那剩下的资源到底应该投入到哪里呢？这需要形成公司内部一套竞争策略，以使优秀的项目能够得到资源。

另一方面，对工程师而言，业务产品的开发技术难度相对较低，如果要想更快提高自己的技术水平，去开发基础技术产品更能得到提升和锻炼，所以优秀的工程师更愿意去开发有难度有挑战的创新性基础技术产品，而不是去开发那些千篇一律的业务产品。

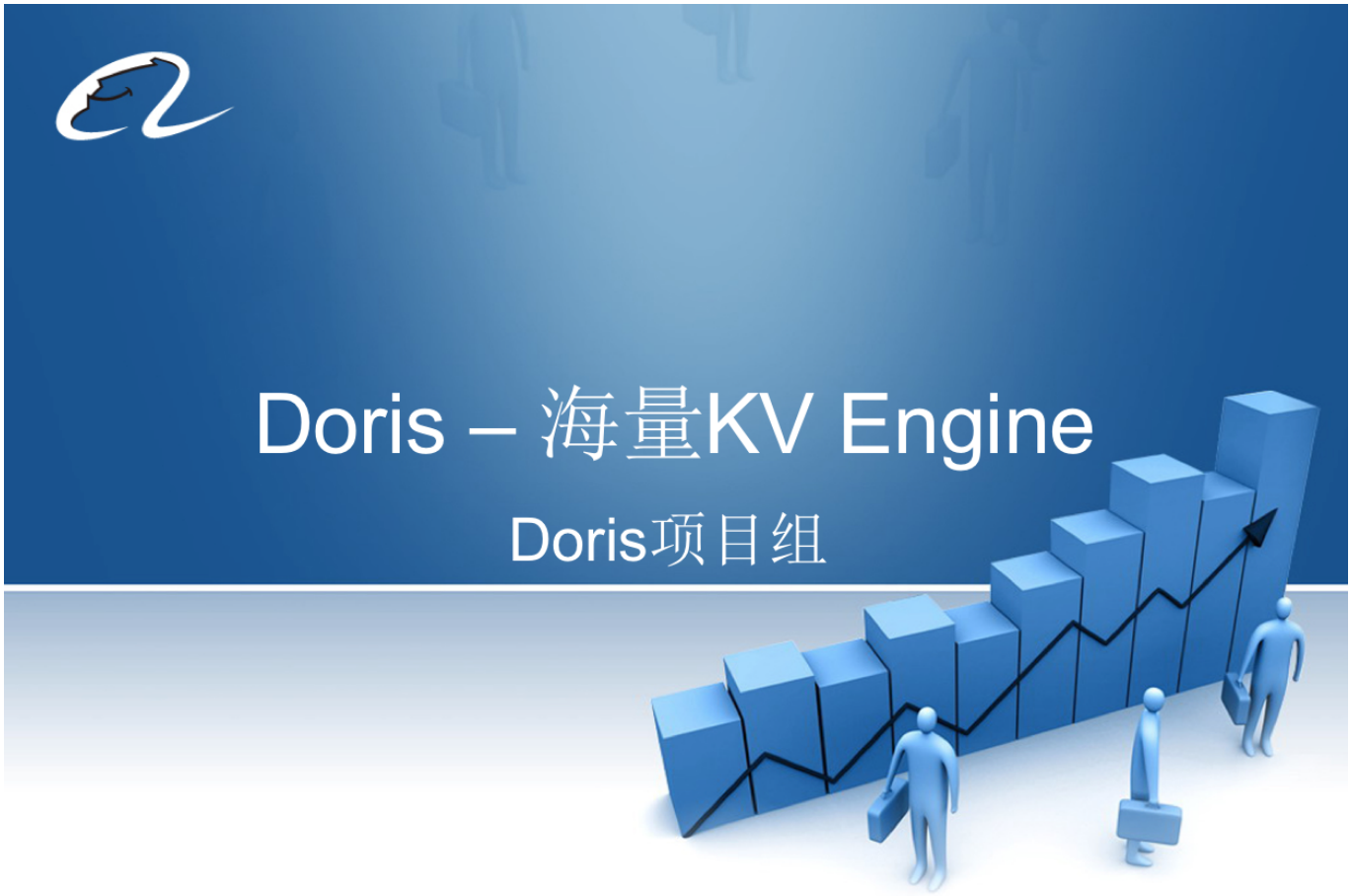
这样，在工程师和公司之间就形成了一种博弈：工程师想要开发基础技术产品，但是必须要得到公司管理层的支持；管理层资源有限，只愿意支持开发那些对业务有价值、技术有创新、风险比较低的基础技术产品。

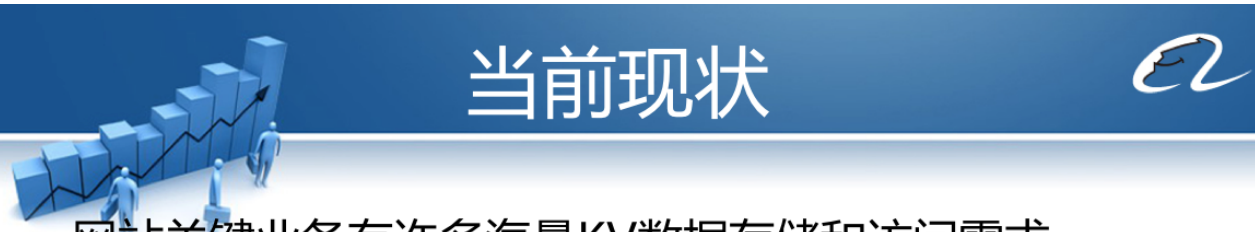
所以事情就变成工程师需要说服公司管理层，想要做的就是对业务有价值、技术有创新、风险比较低的基础技术产品；而管理层则要从这些竞争者中选出最优秀的项目。

通过这种博弈，公司的资源会凝聚到最有价值的技术产品上，优秀的工程师也会被吸引到这些项目上，最后实现了公司价值和员工价值的统一和双赢。

下面我们进入正题，我会拿出当时 Doris 开发立项时说服管理层用的 PPT，向你解读个中技巧以及 Doris 的创新设计。 需要提醒你的是，你在学习这两期专栏时可以试着想象一个场景，假设是在 Doris 项目的立项启动会，今天你是老板，看看你最关注一个项目的哪些

技术指标；又或者你是 Doris 项目的工程师，可以想想哪些指标是老板关注的，并且从技术上是可以实现的。这样把自己带入到一个角色中，对于你更好理解这个数据处理系统很有帮助。





- 网站关键业务有许多海量KV数据存储和访问需求
- 国际站UDAS使用
  - 存在问题：扩容困难、写性能较低、实时性低等
- 网站有多套KV方案, 接口不统一, 运维成本高
  - 国际站 UDAS - BDB
  - 中文站 ： TT
- 飞天KV Engine(Aspara)问题
  - 使用复杂
  - 性能较低

PPT 开篇就是当前现状，当时阿里巴巴没有统一的大数据 NoSQL 解决方案，有的产品是自己在业务代码中实现数据分区逻辑，从而实现海量 KV 数据的存储访问。这样做的主要问题有

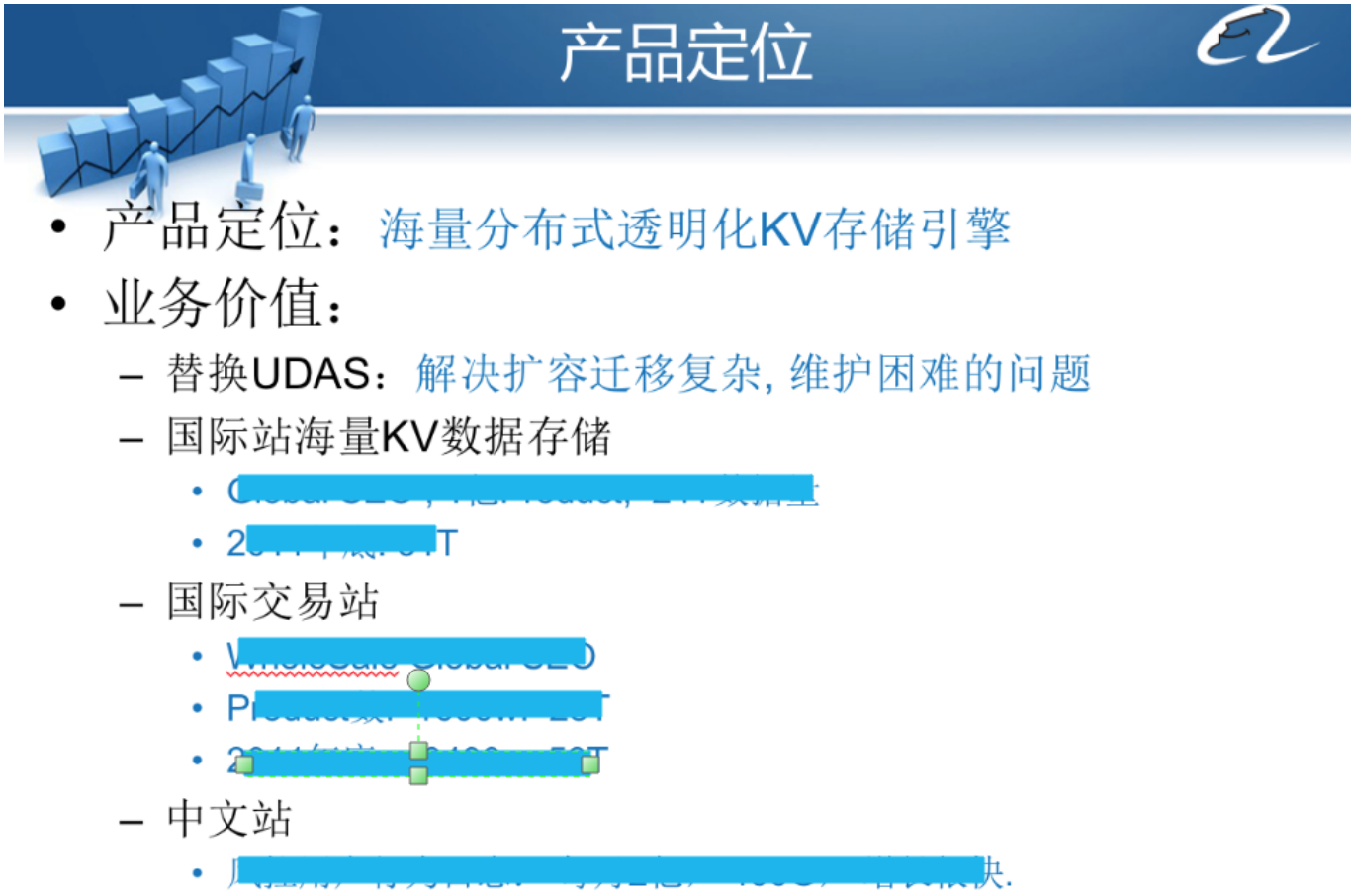
开发困难。程序员在开发时要知道自己存储的数据在哪台服务器。

运维困难。增加服务器的时候，需要开发配合，故障的时候也很难排查问题。

现状一定是有问题的，需要我们去解决。有没有现成的解决方案？有，但是现成的方案也有问题，所以我们必须要自己开发一套系统才能解决问题。这样，后面想做的一切才能顺理成章。

当你想做一个新东西，它必须要能解决当前的问题，这是人类社会的基本运行规律。如果当前没有问题呢？你相信我，这个世界不可能没有问题的，重要的是你要能发现问题。就像你做的东西将来也一定会有问题，因为现在的产品在将来一定会落伍，但那已经不再是你的问题。

技术只是手段，技术不落在正确的问题上一点用也没有，而落在错误的问题上甚至会搬起石头砸了自己的脚。而什么是正确的问题，你需要自己去思考和发现。



前一页说完了当前存在的问题，引出了我们必须要自己开发一个海量数据处理系统，这一页就要说明这个产品的定位，也就是“海量分布式透明 KV 存储引擎”，这个引擎能够实现的业务价值就是能够支撑阿里巴巴未来各个主要产品的海量数据存储访问需求。

这两页是整个 PPT 的灵魂，管理层如果对第一页提出的问题不认可，又对第二页产品要实现的价值不以为然，那基本上这个项目也就凉凉了。

如果到这里没有问题，得到认可，那下一步就要趁热打铁，**突出项目的创新和特点。**

产品目标

- 功能目标:
  - KV存储Engine
  - 逻辑管理：Namespace
  - 二级索引
- 非功能目标：
  - 海量存储：透明集群管理，存储可替换
  - 伸缩性：线性伸缩，平滑扩容
  - 高可用：自动容错和故障转移
  - 高性能：低响应时间，高并发
  - 扩展性：灵活扩展新功能
  - 低运维成本
    - 易管理
    - 可监控
- 约束
  - 一致性：最终一致性

产品的功能目标和非功能目标要清晰、要有亮点，和业界主流产品比要有竞争优势（用红色字体标出），要更贴合公司的业务场景。Doris 的主要功能目标是提供 KV 存储，非功能目标包括在运维上要实现集群易于管理，具有自我监控和自动化运维功能，不需要专业运维人员维护；要支持集群线性伸缩，平滑扩容；具有自动容错和故障转移的高可用性；高并发情况下快速响应的高性能性；支持未来功能持续升级的可扩展性。

目标	指标	说明
集群规模	100+ Machine	
容量	1000T+ (取决于硬件规模)	B2B所有KV存储场景
可用性	99.99+7%	
持久性	10个9	
伸缩性、平滑扩容	不停机扩容完成时间 约= 单Node迁移时间 (10台扩1台场景) 总数据=2.4T 单Node迁移量=240G/10 = 24G 迁移时间=24G/33M = 12分钟	10+1 场景
高性能	Read： < 8 ms （Aspara:10ms) Write： < 10 ms （ Aspara: 10ms) （高于Aspara,国际站SEO需求,高并发场景）	

技术指标也要亮眼，至少不能明显低于当前主流同类产品的指标。当时 Doris 根据阿里巴巴的内部使用需求场景，支持所有的 B2B 业务的 KV 存储，因此设计目标是未来部署一个 100～10000 台服务器的集群规模，并不支持无限伸缩。如果前面说过别的产品的缺点，这里也要对应说明自己强在哪里。

设计指标的设定，既不能低，如果比目前主流同类产品的指标还要差，自己再开发这样的产品就没有意义；也不能太高，如果设定太高，过度承诺，让老板、用户对你未来交付的产品抱有太高的期望，将来稍有不慎，无法达到期望，不但对产品的发展造成不良影响，甚至大家对你的人品都会产生怀疑。做好对别人的期望管理，让大家对你既充满期待，又不至不切实际，不但对你的职业发展大有帮助，应用到生活中也会获益良多。

到这里，问题也说了、方向也有了、设计指标也定了，究竟能不能开发出满足设计目标的产品，就看后面的 PPT 把核心架构和关键设计讲清楚，要证明自己有把握、有能力做到。

到底如何证明自己能做到，且听下回分解。

思考题

在你的工作环境中，哪些工作是更有技术挑战和难度的工作？现在是否有人在做？如果你想做，该如何说服上司支持你？

欢迎你点击“请朋友读”，把今天的文章分享给好友。也欢迎你写下自己的思考或疑问，与我和其他同学一起讨论。

上一篇　20 | Spark的性能优化案例分析（下）

下一篇　22 | 从阿里内部产品看海量数据处理系统的设计（下）：架构与创新

精选留言 (11)

写留言

黄海峰

2018-12-15

4

当时没出现memcached和redis吗？比这两个流行的有什么优势

展开

作者回复: 缓存的数据持久性（永久保存）和可靠性不能满足需求，缓存对内存的需求也不符合应用场景（当时需要存储千T级的数据）

风中有个肉...
2018-12-29

2

我目前负责公司产品开发迭代，角色类似团队小组长，我们依赖的数据源来源于大数据达标计算，我认为大数据技术是我的技术栈薄弱的一块，我想参与该块开发并提升自己的能力。
但是按照部门领导的意思，一个纽扣，一个洞，专业的人做专业的事。
难度在这，我想的几块解决方案如下：...
展开

作者回复: 念念不忘，必有回响，不要放弃，寻找机会

纯洁的憎恶
2018-12-20

1

我很想知道用现有产品，如一些NoSql开源产品、或者付费产品，为什么无法解决现有问题，这对作出自开发的决策时也是十分重要的。

杰之7
2018-12-19

1

通过这一节的学习，技术在伴随着业务的发展而逐步完善的，技术是手段，不落在正确的事情上一点用也没有，所以，在以后的工作中，我们需要知道技术不是目的，对个人和公司能解决实际问题的还 需要我们更多的去关注。

在这篇文章中，老师讲述了针对外部产品扩容低，写性能较低，实时性低的问题，提出...
展开

2018-12-17

1

李老师，之前在公司听过您布道，很受鼓舞。后来有次您指导我们的bi产品，因为有事错过了交流机会。在这里一样跟您学习了很多，一般这类存储的最底层都会基于leveldb或者

改进后的rocketdb进而做分布式和API包装吧

展开

作者回复: 这是一种分布式存储系统开发的捷径，也有很多全部自己实现的。

一块跑跑 1

2018-12-17

技术指标都是经过如何评估计算出来的呢?

展开

作者回复: 参考需求和业界指标，根据自己的设计方案评估。

行者

2019-04-21

教知识又教人，值了!

展开

小老鼠

2019-01-17

现在该产品如何何了?

展开

clairec

2019-01-02

最近在筹备公司级专项，看您的指导，茅塞顿开。

展开

lanpay

2018-12-24

当时业内标杆应该是hbase和Cassandra，不知道Doris设计上借鉴哪个多些😁

纯洁的憎恶

2018-12-20

1.通过现状分析发现问题与瓶颈。

2.通过市场研究，结合自身资源与手段，确定宏观路线图（产品定位）。

3.根据路线图，确定具体的解决方案（产品目标）。

4.基于具体目标，明确合理、可量化的业绩指标。