Hadoop & Spark 峰会杂谈 － 关于演讲和参会

2015-07-03 Luke Han [立方体](javascript:void(0);)

今年5月和6月我们参加了两次全球Hadoop的盛会：Strata＋Hadoop World 2015 London及Hadoop Summit 2015 San Jose。我们的项目Apache Kylin也分别在这两次大会上作了专题演讲。机缘巧合，在San Jose的时候总部的同事转让了一张Spark Summit的门票，得以参加三个全世界最热的大数据处理和分析大会，在此总结一些我们演讲和参会的经验和感受，希望能给大家一些参考和帮助。

**三个大会的背景**

先分享下几个大会的背景，以便我们对几个大会有更深入的了解：

1. Strata＋Hadoop World是由O’Reilly和Cloudera主办，全年有四个会，分别在北美（硅谷和纽约），欧洲（伦敦）及亚洲（新加坡）举行， 以在硅谷（San Jose）举行的为主，官方网址是：http://strataconf.com

我们参加的是在伦敦的大会，在伦敦市区的希尔顿大都会举行，来自Hadoop世界的主要玩家，包块Cloudera，Hortonworks，MapR以及大厂如Teradata等都积极参与，各种专题演讲，客户案例及展台等非常的丰富，特别是来自欧洲的一些初创公司，演讲专题等非常的有意思，其中有一个关于物联网的专题会场中有来自西门子等各种实际案例，非常值得一看，可以从附录的链接中找到相关视频和Slides。

2. Hadoop Summit则是由Hortonworks和Yahoo主办，并由Apache基金会作为社区合作伙伴，全年有两个会，分别在北美（硅谷）和欧洲（布鲁塞尔）举行，以在硅谷（San Jose）举行的为主，官方网址是：http://hadoopsummit.org。

今年的参会人数突破4000人，比去年的峰会增长了30%，在San Jose高大上的会议中心举行。除Keynotes外还有9个Tracks总共超过165个专题演讲，涵盖了几乎所有Hadoop 生态圈的内容，从Committer到Hadoop各个项目，数据科学，实际应用到治理和安全，运维以及其他等。有客户的实际使用案例到核心功能，各种围绕Hadoop的开源项目，大小公司的创新等不一而足，可以从附录中的链接中找到相关视频和Slides。

3. Spark Summit则是由Databricks，Intel及IBM主办，全年三个会，分别在旧金山，纽约和阿姆斯特丹举行，以旧金山为主，官方网址是：https://spark-summit.org/

参会人数从2014年的不到500人暴涨到2015年的2000多人，可见Spark社区之火，特别加上会议开始前IBM宣布的3500名开发人员投入到Spark相关的研发更加引爆了整个社区和业界。今年的会议在旧金山希尔顿联合广场酒店举行，可能组委会没有预料到如此多的参会者，举办地略显拥挤。Keynotes之外分为三个Tracks（开发者，数据科学及应用）约60个左右的专题演讲，并加一天全天的Spark培训。而且几乎场场都爆满，没有参加的同学可以从附录中的链接中找到相关视频和Slides。

那么商业敏感的同学应该已经知道相关的背景了，特别是Hadoop的两个大会。由于这是全世界最大的两个Hadoop大会，因此都吸引了众多的参会企业和个人，但由于是分别由Couldera和Hortonworks举办，相对的参会议题，企业和方向都略有不同，Hadoop Summit更关注在Hadoop核心，社区，生态及应用上，加上Apache基金会的合作，Hadoop生态圈相关的内容非常丰富，而Strata＋Hadoop World更关注应用，业务，行业解决方案等。当然大部分是重叠的。可能其中最憋火和开心的是MapR，作为三大Hadoop分发版之一，只能两边跑，但却也可以开心的两边都有很好的合作。

三者中组织的最好的应该是Strata＋Hadoop World，全程安排，引导，用餐等都非常到位，各个专题内容也错落有致，O’Reilly不愧是搞媒体和宣传的好手。特别是会后的跟进，媒体资源的引入等对于希望去提高曝光度，品牌宣传等是非常不错的选择。

Hadoop Summit今年也是非常的高大上，在整体规模和参会人数，演讲内容上都略胜一筹，几乎涵盖了整个Hadoop神态圈的各个方面，但由于同时举行的会议太多，如何挑选则成了非常头痛的事情。

相比较而言，Spark Summit由于比较新，各个方面还有待完善，比如过小的展厅和简陋的午餐等，演讲主题方面客户的实际案例不过多不够丰富，大多都集中在机器学习和Streaming上，SparkSQL相关内容估计会在下一次的峰会上称为热点。不过即使如此，也是几乎场场爆满，可见Spark的炙热：）

**关于专题演讲**

参加这几个世界级大会对于提升公司和项目的知名度非常有帮助，特别是对于在从事Hadoop／Spark相关的工作的朋友，通过这几个会议可以让更多的人甚至整个社区和业界了解正在做的事情，分享一些经验和教训等，也是非常好的平台可以与其他公司，社区及作者进行交流。如果你的议题足够吸引人或有趣，你会发现原来你的工作可以让更多人获益，这对于自身的成就感不言而喻。上面是李扬和我在伦敦演讲开始前的照片。

几个大会对于主题演讲的提交时间都有比较严格的要求，一半都要求提前半年以上。可以去他们的网站上查询具体的提交截止日期。一般来说，需要准备以下几方面：

1. 标题，起个切合题意并吸引眼球的标题很重要，对于入选及吸引现场听众非常关键

2. 简介，一般在400字左右的Abstract，简单明了描述演讲内容

3. 演讲内容，可以放更多更详细的内容来帮助组委会筛选

4. 演讲者个人介绍

提交之后可以不断的更新，基本上在峰会开始前都可以不断更新版本。如果要参会建议尽早准备，在众多的议题中入选并让更多的人来参会需要做不少的工作。今年Hadoop Summit提交了有超过650个演讲专题，最后只有165个被选中得以入选，竞争也是相当激烈，认真准备货真价实的内容是关键。

特别提醒的是Strata＋Hadoop World会要求提交一段2～3分钟的英文演讲视频，如果有以前的演讲内容则是最好的，否则可以自己录一段后上传到Youtube上。

Hadoop Summit有一个社区投票的过程，每个专题都会挑选社区挑选中最高的两个，其他则由组委会决定，非常遗憾的是Apache Kylin在今年的社区投票中排名第三，幸运的是我们最终还是被选上了。

Strata+Hadoop World会负责演讲者的差旅费用，包块飞机票，酒店，当地交通及用餐等费用（有限额），以及参会门票。

Hadoop Summit需要讲师自己承担差旅费用，但会提供参会门票给演讲者。

由于门票都要上千美金，并且为了保证演讲效果，这两个会议的主办方都不会允许一个议题超过两个以上的演讲者。

Spark Summit由于没有参加演讲，不是很清楚，如果有兴趣可以去联系阿里的明风同学。

**关于参会**

动辄上千美金的门票加上出国的差旅费用导致了很多国内的同学难以参加，但这样的机会对于工程师，负责人等开拓眼界，与社区和作者交流，了解行业动态等都很有裨益。很是推荐有能力和实力的公司多送人出去，参与演讲或者参会。在几个大会上听了很多用案例，都是来自一线企业的实践，比如西门子，介绍了他们如何利用物联网＋Machine Learning来收集他们提供的铁路及设备信息，及早预测设备磨损，从而保证客户项目的顺利使用，德国人的严谨在这个领域又一次展现实力了。来自Netflix分享如何使用Presto服务PB级分析，LinkedIn构建的Pinot框架等。来自社区和生态圈中的演讲，特别是实践类演讲比较有参考价值，和演讲者进一步的交流的效果也会很好，这次和Apache Lens，Apache Tajo，Apache Zeppelin的交流很有收获，大家未来可以看到一些与Apache Kylin在这方面的合作和整合。而来自大小供应商的专题一般都有点水，既然赞助了，卖货是唯一目的，当然，商业版也很有价值，就看在实际中是否能应用到。

**关于展台**

每个会议都有赞助商，每个赞助商都有展台，每个展台都希望能抓住更多人的吸引力，并建立联系。虽然大部分都是为了达到品牌宣传和寻找潜在客户的机会，但多去和他们聊聊确实也获益匪浅，特别是一些创新型的企业，比如H2O.ai，基于ARM的迷你集群等。很多创新企业能给你带来相当不错的思维开拓。在伦敦和美国都看到了几个和我们一样做类似OLAP on Hadoop的解决方案，经过一番了解，都没有我们Apache Kylin完备和经受过超大规模数据的实际检验，感觉信心满满的：）当然也非常的有紧迫感，深深的感受到了和人赛跑的感觉，虽然我们有一点点优势，但有更多的不足，无论如何不能放松。

**关于活动**

每个大会的会前会后都会有一些不同的活动，包块参会者Network形式，Party，Meetup等。Hadoop Summit还有一个Bike Ride活动，今年是第三次，可以租个自行车与一行人骑行在硅谷最棒的风景线，今年报名参加了这次活动，作为一个胖子第一次骑行并顺利到达，完成了又一次的自我挑战。

Spark Summit的第一天晚上还举办了一个Spark Meetup，介绍了Spark1.4的新特性，之后几个主要的Spark Committer往前一坐，轻松面对狂轰烂炸似的提问却应答如流，一种学霸们集体舌战群儒的即视感相当强烈，问了他们几个关于SparkSQL相关的问题。正好Apache Kylin下半年正在做Spark和SparkSQL相关的整合，希望我们能有机会参与并贡献一些到社区。

**Big Data & China**

参加Hadoop大会的时候发朋友圈说怎么感觉华人这么少，更不用说国内企业去演讲了，但到Spark大会上就发现完全不一样，估计40%以上的参会人员是华人，而且非常多来自国内的同学，当晚组了好大一群华人，可惜小肥羊没有那么大的位置，错过了刷美国小肥羊的机会。

百度，淘宝等这次有不错的专题分享，加上Intel中国研发中心的专题中提到的大规模应用案例中大部分都是国内企业，包块优酷，京东等，说大数据，咱们的数据和挑战确实足够大，特别是Spark Summit的Keynote中提到的最大集群：腾讯8000个节点，最大的单个Job：阿里巴巴1PB，恐怕很多西方人都不太敢相信。同时也看到了华为的展台，华为在相关硬件，Hadoop和Spark上投入了非常多的资源，在业界有了不错的立足之地。很多国内的实践，项目和应用其实都走在很前面，非常期待能够有更多的公司和个人去这些大会上发出我们的声音，希望在未来看到更多的来自中国的演讲，展台和参会人员，Big Data & China，我们正处在一个好时代。

**趋势和热点**

从这几个会议中，明显感受到以下的一些趋势和热点：

1. Hadoop － Enterprise Ready, Spark - the Future

众多的议题和讨论中，可以发现传统数据库，数据仓库，ETL，商业智能，数据挖掘等行业中的一些重要特性，优点等都在不断出现在Hadoop之上，整个Hadoop生态圈也越来越成熟，除了几个Hadoop发行版的公司之外，Teradata，SAS，Informatica，EMC，微软等提供了更多重量级的支持，加上各种新创企业和项目，使得之前很多人顾忌的企业级品质也越来越完备，特别是运维和监控，安全性和稳定性，分析应用的整合等，Hadoop已经不再是互联网企业才会玩的先锋项目，可以看到Hadoop和Spark在包括西门子，丰田，迪斯尼，NASA以及石油，政府等传统企业和行业也在不断使用，应用也更多样化，涉及到更多的方面。

相对来说Spark在企业级应用上还略显年轻，也更缺乏相应的生态圈支持，更多的都还是在机器学习，Streaming等领域，但却是星星之火，越来越旺，所有的Hadoop发行版都说支持Spark，几乎所有的大小公司都在使用或者研究使用Spark，其必然是发展的未来。

2. Realtime & Streaming

Realtime和Streaming成了全年的关键字，随着数据量，种类等的飞速增长，实时的数据处理几乎成了必备的需求。从物联网，交互式在线分析，机器学习等都有涉及，几个大会中也充分体现了这个，每个大会都有超过一，二十个以上以实时为标题的议题，还不包括内容中涉及实时和流式处理的议题。

从技术架构上，几乎所有的项目都宣称支持Lambda架构，即同时处理历史数据和实时数据流的体系架构。Hadoop正从Batch走向Realtime，于是相关的几个项目都获得了长足关注，包括Storm，Kafka，SparkStreaming等。

3. Spark, Spark and Spark

你在玩大数据？玩不玩Spark？不玩你就out了。两个Hadoop大会上都有相当数量的的Spark议题，更不用说Spark自己的峰会，以前还觉得是Spark贴着Hadoop的生态圈玩，这次感觉完全倒过来了，从Coudera，到Hortonworks到MapR到其他所有大小玩家，都在向Spark靠拢，感觉都怕别人不带他们玩儿似的。而从相关的专题分享来看，Spark目前还更多的集中在数据科学，Machine Learning，Streaming，预测等领域，但随着相关项目成熟，更好的稳定性和性能，从MapReduce等Hadoop生态迁移去Spark生态会越来越多，个人感觉，下一个热点应该是SparkSQL和Tungsten项目。

**后记**

有幸参加了几个大会，也贡献了两个专题演讲，听了很多不错的演讲，也有很多太水的内容，见识了其他有趣的项目和应用，对未来的看法和项目发展有了更清晰的认知。

而更重要的是认识了很多不错的朋友，包括Apache Kylin相关的朋友，Hadoop和Spark社区的作者，业界的朋友等等，来自国内的，硅谷的，其他地区的等等，非常高兴和你们认识：）