

来自RISELab实验室解决机器学习新的大数据工具

大家对加州大学伯克利分校的AMPLab可能不太熟悉，但是它的项目我们都耳闻——没错，它就是Spark和Mesos的诞生之地。

AMPLab是加州大学伯克利分校一个为期五年的计算机研究计划，其初衷是为了理解机器和人如何合作处理和解决数据中的问题——使用数据去训练更加丰富的模型，有效的数据清理，以及进行可衡量的数据扩展。

AMPLab于去年11月份结束关闭。新的实验室就此开启——RISELab，另一个该校五年期项目，有着强力的财政支持，将聚焦于提供安全执行的实时人工智能系统。



如果想及时了

解Spark、Hadoop或者Hbase相关的文章，欢迎关注微信公共帐号：iteblog_hadoop

RISELab的团队任务是前瞻性地推动大数据分析到一个更加深入的世界，在那个世界，AI是真实的，世界是可编程的。举个例子：围绕着“小型自主飞行器”管理数据设备，无论是无人机还是气垫汽车，都是数据在被高速且安全地处理着。

其他的挑战还包括安全领域，但不是传统意义上的访问控制。当然也包括类似 "homomorphic" encryption的概念，加密数据可以直接被使用不需要解码。“如果没有对云的理解，我们如何预测云上的数据？” Gonzalez如是说。

尽管实验室还处于初期阶段，一些项目已经浮现在人们面前：

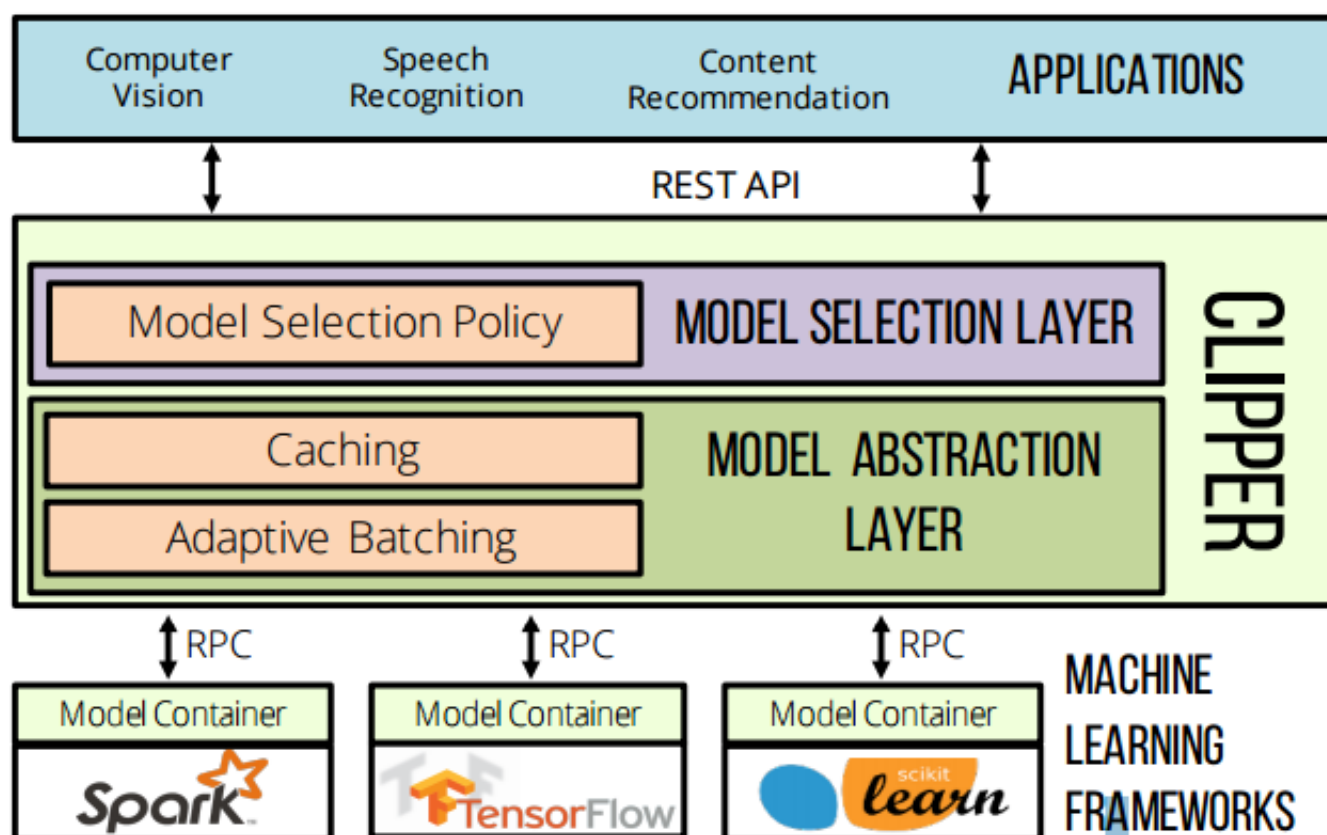
Clipper

机器学习包括两项基本的工作：根据预测建立模型以及从模型提供预测。Clipper专注于后者，是一个多用途、低延迟的预测服务系统，根据机器学习framework以最低的延迟进行预测。

Clipper在机器学习方面主要有三个目的：其一，加速从被训模型中获取预测的速度。其二，在多个机器学习framework上提供一个抽象层，开发者只需要编程一个API即可。其三，Clipper的设计让它可以动态地响应，如单个模型响应请求。举例来说，允许一个给定模型对特定类型的

问题进行优先级的回复。目前还没有明确的机制，但已是未来的趋势。

Clipper的体系结构如下：



如果想及时了

解Spark、Hadoop或者Hbase相关的文章，欢迎关注微信公共帐号：iteblog_hadoop

关于Clipper的论文可以参见[Clipper: A Low-Latency Online Prediction Serving System](#)

Opaque

目前看来RISELab的项目会补足AMPLab余下的工作。Opaque就是其中之一：Opaque和Apache Spark SQL一起为DataFrame提供强力的安全保障。它使用Intel SGX处理器的扩展部分，把DataFrame标记为加密，所有的操作都在"SGX enclave"下执行，数据就地使用AES算法加密，只有通过硬件层保护的应用使用它时才可见。

没有性能损耗的情况下，它提供了同态加密（homomorphic encryption）的优势。使用SGX的性能损耗大概在50%，但是当前最快的同态算法实现起来也要比它慢2000倍。另一方面，SGX的处理器在云上还未提供，尽管Gonzalez表示近期将会实现。最大的障碍在于，为了让它能够工作，“你必须相信Intel”。

Ground

Ground是一个数据湖（data lake）context管理系统。它提供了在Java中实现一个RESTful服务的机制，让用户去推论他们拥有什么数据，数据从哪里来向哪里去，谁在使用数据，数据何时变化，为什么会有这种变化等。

数据聚合（data aggregation）已经从严格的数据仓库型管理中移除，向开放且灵活的数据湖接近，但是也让追踪数据形成变得很难。在某些方面，弄清楚谁改变了给定数据集以及如何改变，比了解数据本身更重要。Ground提供了一个通用API和追踪信息的元模型，可以和很多数据储存库一起工作。

Gonzalez承认定义RISELab的目标并不容易，但是他表示，“它的核心在于从如何建立高级分析模型、如何分析数据到如何利用洞察来做出决定的转型过渡——将Spark和大规模分析的产品与世界连接”。

本文原文：[Spark和Mesos的诞生之地，又在研究什么新技术？](#)

英文原文：[New big data tools for machine learning spring from home of Spark and Mesos](#)

本博客文章除特别声明，全部都是原创！

原创文章版权归过往记忆大数据（[过往记忆](#)）所有，未经许可不得转载。

本文链接：[【】（）](#)