## Hadoop生态系统应用于互联网金融领域

Hadoop生态系统包含数据存储HDFS，数据分析HIVE、HBASE、MR等，机器学习库Mahout（以主要被Spark代替）。

HDFS是分布式离线存储服务，可以离线存储大量数据，以小文件的形式存储。HBase是实时查询系统，支撑线上服务，通过构建索引实现。MR是最基础的分布式计算框架，实现简单的统计和ETL任务。HIVE提供了类SQL的语言接口，通过SQL语言操作表结构数据，是最常用的ETL工具。Mahout作为已经停止的项目，所有算法已经迁移到Spark ML和Spark MLlib中，对于大数据机器学习任务的训练预测、聚类和挖掘，可以很方便的实现，因为使用内存计算和DAG，可以使得迭代任务的数据不需要落地并能复用共享数据，计算性能有很大提升。

在互联网金融领域，我们的数据量大概在千万到亿级别。数据类型包括：用户关系网络、用户消费数据、用户信用数据。目前在存储、分析、建模几个环节都在使用Hadoop生态系统的工具。具体使用情况如下：

用户关系网络的数据，主要包含用户通话关系、用户紧急联系人关系等，通过Spark的GraphX离线分析，做fast unfolding社区挖掘和类似PageRank的用户重要度发现。

用户消费数据，主要包含信用卡网银账单数据、淘宝流水等，通过Spark和Hive离线分析数据，做ETL。把用户消费的非结构化明细，清洗成schema数据。

用户信用数据，大部分是结构化数据，也包含上面个两种数据清洗后的结构化数据，还有像用户逾期次数、信用卡额度等数据。通过HIVE可以方便的处理。对于业务要求的风控和反欺诈模型，通过Spark实现的XGBoost做分类预测，或者通过mllib提供的聚类、密度估计等算法挖掘潜在特征。

在线上通过HBase实时服务提供给业务方调用，服务用户。