大数据与人工智能三页纲要

【导读】最近很多朋友都问我要做一个好的数据科学家应该掌握哪些知识或技能？确实，大数据领域犹如当年的大淘金年代，各种新名词日新月异，一不小心就被会被拍到沙滩上。于是就冒充一把专家，整理了这个文档，不求全，但求对朋友们有点帮助。

1. 数据平台（分布式存储、计算技术及安全管理）

(1) Hadoop 系列

HDFS: CF: GFS
Map-Reduce:
HBase: 数据库
Hive: 查询工具
YARN: 资源管理工具
Zookeeper: 一致性管理工具
Avro: 接口
Pig: 高级查询

(2) 数据库（SQL、NOSQL、k-v）

传统关系型数据库：Mysql, Oracle
分布式列存储关系型数据库（永久存储）：HBase, Bigtable, Greenplum,
Ali 的列存储数据库: RDS, PetaData, OceanBase, HybridDB
分布式非关系型数据库（永久存储）：MongoDB（基于文档）, ElasticSearch（以搜索为主）,
Neo4j（图结构数据库）,
内存 Key-Value 数据库（适合高热数据缓存）：Redis, Memcached,

MongoDB: cloudera 出品，使用 c++开发，BSON 结构，
关键词：collection, Document, Schema-free
ElasticSearch: 基于 Lucene（java/json）的搜索服务器，带有存储机制，
在需求稳定时一定程度上可以代替 MongoDB。CF: Solr
Redis: 内存数据库，K-v 型，适合缓存高热数据
Memcached: 内存数据库，k-v 型，出现较早，与 Redis 相比各有优劣。

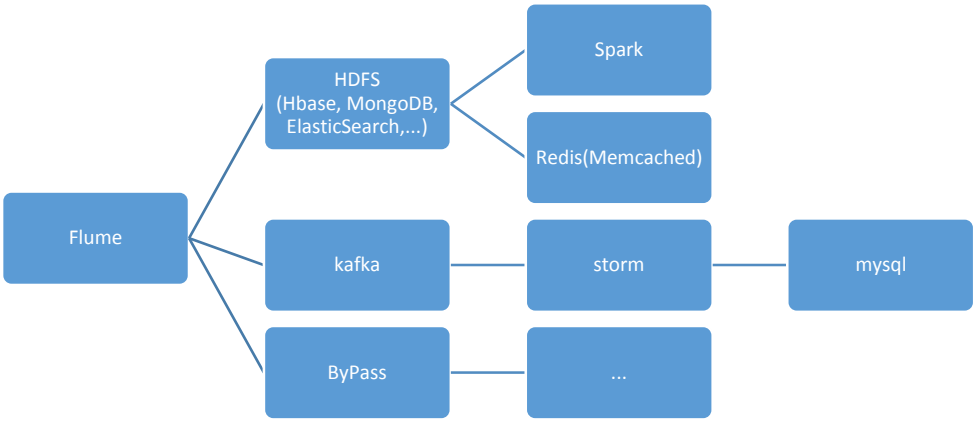
(3) 数据分析处理工具（采集、消息系统、MPP、内存计算、实时流处理）

数据采集工具：Flume,
消息系统: Kafka,
大数据处理（MR 机制）: mahout,H2o, Drill, ODPS
大数据处理（MPP）: Impala,
大数据处理（内存计算）: SparkSQL/SparkMLlib,
实时流数据处理: Storm, SparkStream,
异构数据处理（图计算）: SparkGraphX, Caffeine/Pregel/Dremel（@google）,

Flume: 日志采集工具，关键词：Source, Channel, Sink CF:Scribe
Kafka: 消息系统，可使用 Zookeeper 管理，关键词：Topics, Producer, Consumer
Impala: 快捷 MPP 查询，cloudera 出品，使用 C++，可代替 Spark 但不是内存计算。
Spark: 内存计算，除了 SQL/MR，还带强大的 ML 库和图计算包。
Storm: 实时流数据处理，clojure 语言编写，支持 java, python, ruby 语言，
关键词：Topology, Spouts, Bolt, Nimbus, 使用 Zookeeper 管理，

(4) 大数据平台架构实例

一个实际的架构（从一个听说是美团的方案上做了修改）



2 模型与算法（机器学习、数据挖掘、人工智能）

(1) 机器学习十字真言（标准步骤）

领域、数据、特征、模型、融合

领域：领域知识，业务逻辑
数据：采样、清洗、ETL、整理，高价值数据源、脏数据，增强（ADO Boost），迁移，
特征：特征工程，特征识别，
模型：算法，参数。
融合：成本函数、损失函数、

(2) 模型类别（回归、分类、聚类、挖掘）

回归（排序、评价）: LR, GBRT, FM, LTR, LSVR, Ridge Regression,
分类: GBDT, Random Forest, logistic regression, SVM, Bayes, KNN, deep neural network, Discriminate Analysis (Bayes, Fisher, Mahalanobis distance),
聚类: K-mean, EM,
挖掘（主题识别、关系挖掘、特征分解等）: LDA, Association rule mining, Tensor Analysis, Factor Analysis, PCA (Principal Component Analysis),

(3) 几个经典模型（最常见的和不常见的都不介绍了）

LTR: learning to Rank, 学习排序

LSVR: 线性支持向量回归

FM: Factorization Machine, 因子分解机

LDA: Latent Dirichlet Allocation, 词-主题-文档 三层 Bayes 模型

Bayes: 贝叶斯模型有多种，朴素贝叶斯，贝叶斯分类，贝叶斯判别都比较常用。

EM: Expectation Maximization Algorithm 最大期望算法，用于含有隐变量（latent variable）的概率参数模型的最大似然估计。

3. 应用场景（应用场景加行业应用）

(1) 用户画像（用户行为分析）及个性化推荐模型

业务场景的重要性：在商业领域，理论模型还是必要的绕不开的，黑盒法远远满足不了需求。

数据源的重要性：有时候一个看似无法解决的问题可能可以通过引入新的数据源来解决，比如识别假货可能需要投诉数据，找客户流失可能需要客户常住地点变化的数据。

模型方面，分类模型是用得最好的，但聚类有时也很有效，不仅对用户，还可以在店铺、产品、类别等多粒度多次的进行分析处理。多源异构数据近年来也开始引入（如在线评论，社会关系）。除了常见的用户特征，文本偏好等更多用户特征被引入称为多角度（360 度）建模。

(2) 征信评分、客户声音与信贷风控模型

典型案例：芝麻信用，身份特质、行为偏好、人脉关系、履约能力、信用历史五个维度。

传统的信贷风控模型：按客户类型或保全方式分类（分级）或得到风险评估分值，客观因素包括公司年度审计财务报告和银行流水等，主观因素包括所在行业和管理人的经验等。一般遵循业务定义-风险定义-风险分解-风险策略四个步骤。

经典的大数据信贷风控模型：使用数据挖掘方法（如 K-Means/LDA 等）进行用户分类（预测），建立标签库，为评分提供数据支撑。

互联网金融----大数据风控在小额贷款上的尝试。信用分期，支付宝花呗、借呗小贷等产品。

(3) VR 及文本语音图像视频处理

这是一个黑盒法能够发挥作用的领域，因此与前面两个应用场景差别很大。如果说前面那个领域更多是人文社会科学，需要的是以统计学为基础的统计学习模型，那这里则是纯粹的“物理”问题，传统机器学习的效用更好，所以深度学习在这几个领域都带来了革命性的突破。

近年来，人工智能的成就主要体现在这个领域，语音识别，文本处理，图形处理都已经比较成熟。但也有许多挑战，如语言逻辑的处理，多传感器数据融合等都需要突破。

(4) 机器人、决策系统、专家系统及其它

这是最激动人心的领域，下围棋的阿法狗，能看病的沃森都是各路数据科学家和数据淘金者们努力的明灯。

其核心技术就是基于大数据的人工智能，尤其机器学习、数据挖掘方法的应用。

叶俊杰（无锚），2016/12 于杭州西厂

Email: yejunjie@21cn.com