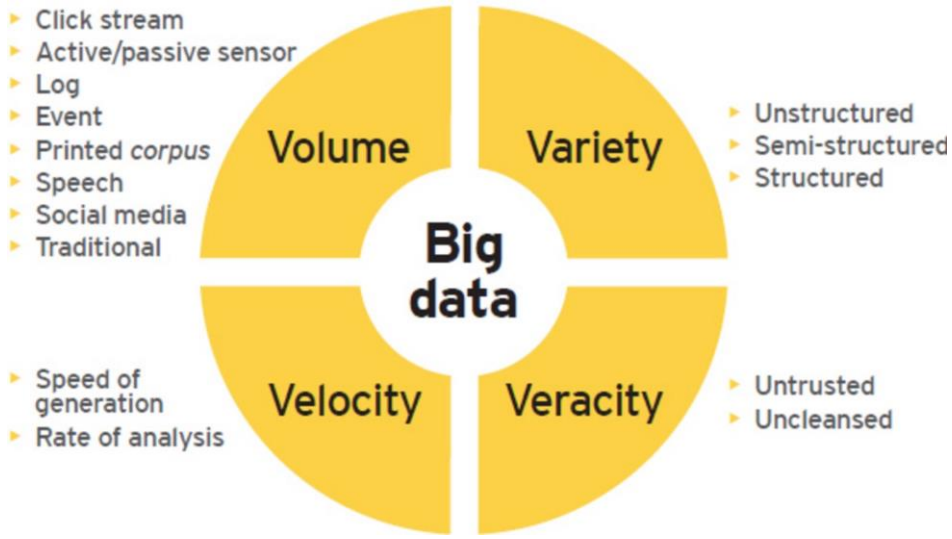
Big Data Analytics Fundamental

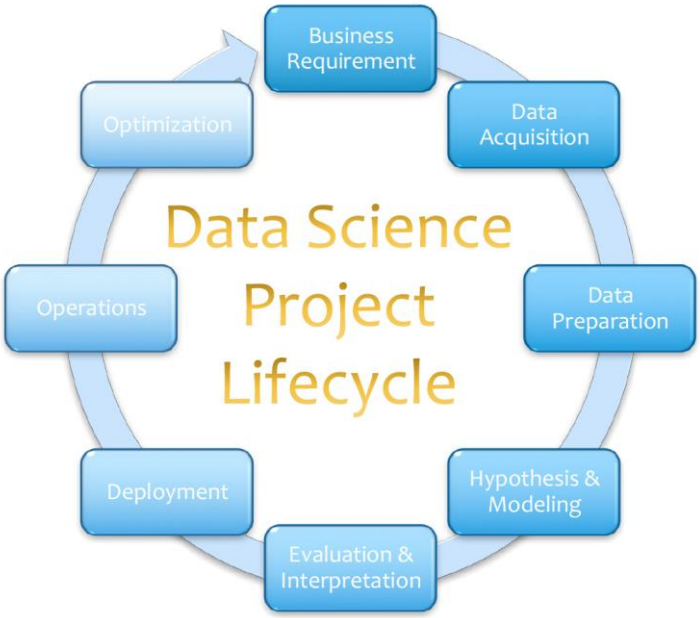
1 What is "big data"

核心要点：4 个 V



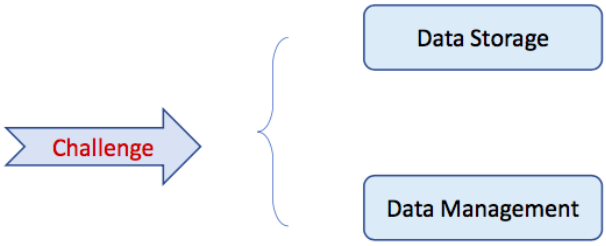
- Volume："Big data"是大规模数据集（large scale data set）。
- Variety：处理各种类型的数据。
  - Unstructured：音频、图像、视频、产品评价（product review）
  - Structured：TSV，CSV 数据
  - Semi-structured：混合类型数据（hybrid）
- Velocity：模型的实时性（process customers' requests in a real time manner）。
  - **e.g.** Amazon 每天面对百万消费者的快速实时的产品、关键词（keyword）推荐系统
- Veracity：数据质量对模型训练非常重要。
  - **e.g.** 使用基本 supervised 机器学习算法训练模型并预测贷款申请者贷款是否被批准，当训练模型的数据（training data）的标签（labeling）有错误（如，高信用记录者因系统故障或超时而被拒绝贷款），会对预测的结果的准确性有很大影响。

2 Data Science Project Life Cycle



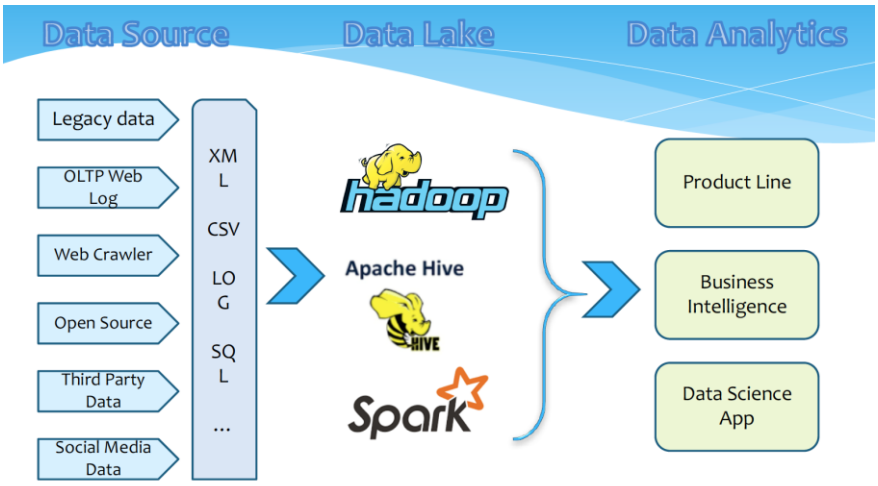
- **Business Requirement**
  - 和相关领域专家合作，理解获取数据的意义，并进一步明确所分析的需求。
  - e.g. 以一个 Health Care 的数据分析为例，我们通过数据准备预测出院后的病人三十天内再次返院的几率。从而通过预测给医院节省费用。
  - e.g. 以另一个游戏公司面试题为例：公司给出四个数据表格, User, Session, Purchase, Usage. 我们通过这些表格去分析如何促进不花钱的用户去消费，从而使游戏公司获得更多利润。

- **Data Collection/Acquisition**



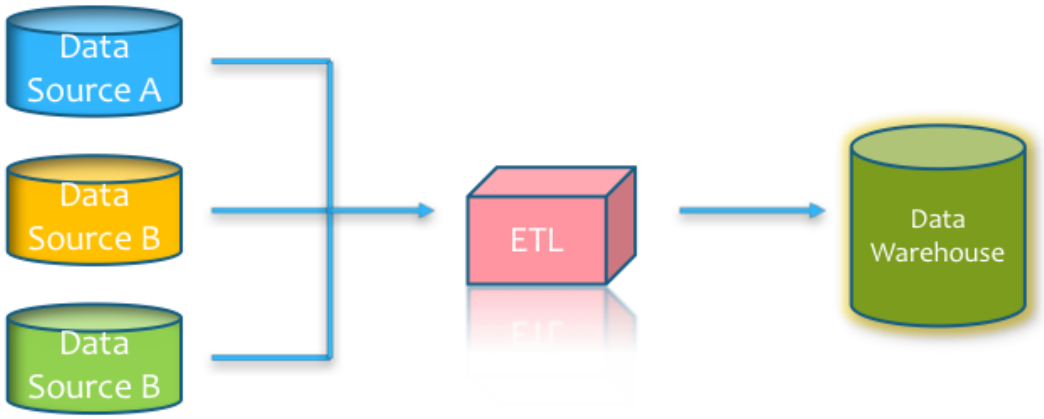
  - 关键任务是数据存储（storage）以及管理（management）。

- 数据来源（data source）：Data from profuct line, Purchase third party data, Social Media (Facebook, LinkedIn), Web Crawler, Open Source 等等。
- 在数据处理中，我们会接触到很多有帮助的处理工具，我们从各种来源处收集到数据后会把它们存储到 Data Lake 里，之后把这些数据进行分析。

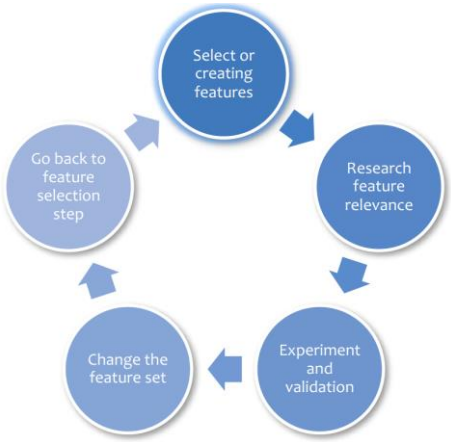


- 数据湖泊（data lake）：可以保存大数据的分布式并行系统、能够在数据不移动的情况下进行计算的系统，比如 Hadoop, Apache Hive, Spark.
- 数据分析（data analysis）：Product Line, Business Intelligence, Data Science App。其中将数据从平台取出进行分析，主要是 Data Science App。

- **Data Preparation**
  - 我们拿到数据之后便要开始准备数据在 workflow 中占绝大多数的时间（80%）。
  - 主要任务：
    - 数据清理（data clean-up）：语义错误（semantic errors）、条目缺失（missing entries）、格式不一致（inconsistent formatting）等等。
    - 数据整合（data integration）：将历史数据与新获取的数据进行整合，对建模有很多帮助。

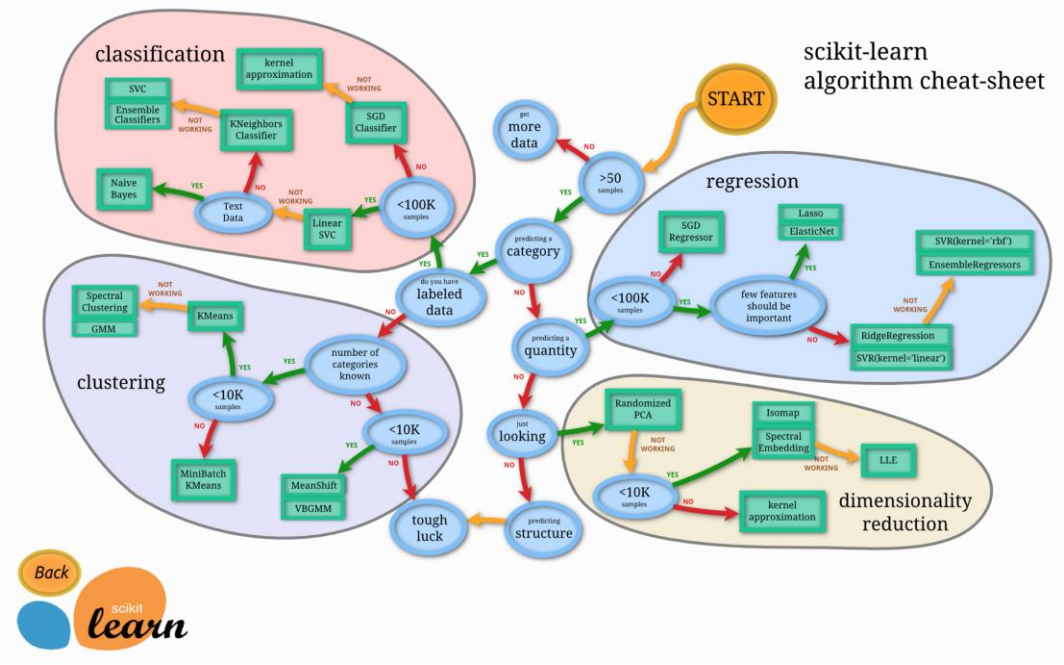


- **Feature Engineering**



  - 其中关于 Selecting/Creating features 主要包括: 对现有数据进行分析，或者 dimension reduction（分析 correlation 等等）.
  - 计算新的 features（例如，关于某产品订单数据，想获得一个新的 feature：每个用户购买某产品的平均间隔时间，此时需要对数据进行相关的分组处理等等）

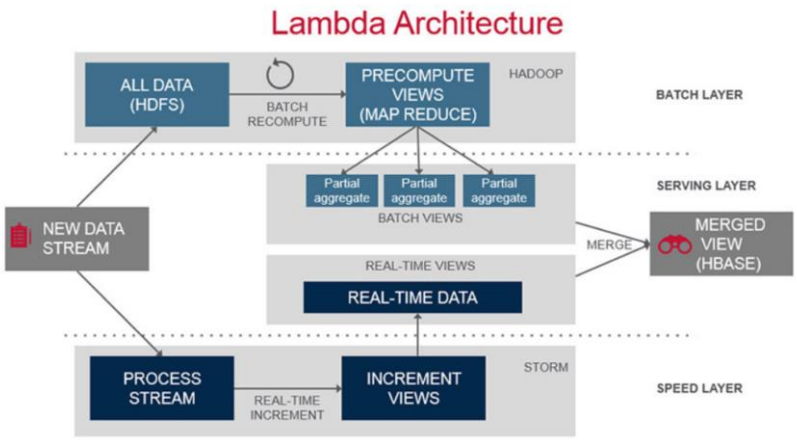
- **Modeling**
  - 后续课程会详细讲解。可以参考资料：http://scikit-learn.org/stable/tutorial/machine_learning_map/



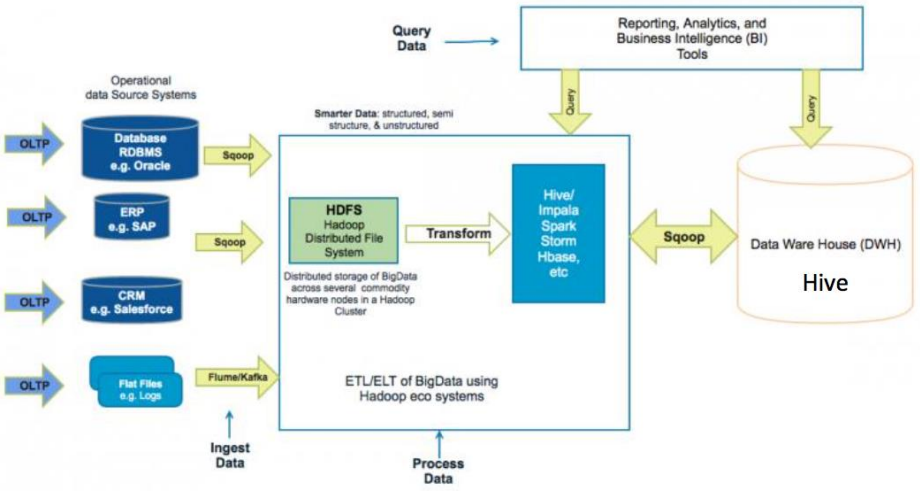
- **Work Flow**

  - 这是前学员在面试时做的项目展示的截图之一。大致上我们做每一个 DS 项目时，都是按照这套流程来做的。

- **Deployment**
  - 我们可以把做好的 Model 放在 Lambda 里。
  - Lambda 结构有三层：
    - 批处理层（Batch Layer）：按批把数据加载到数据仓库中。
    - 加速层（Real Time/Speed Layer）：应用 e.g. Amazon search keyword 实时推荐系统。
    - 服务层（Serving Layer）：通过建立的模型分析后，将服务提供给用户。



- **Hadoop Data Warehouse**



  - 我们之后要进行的大数据的操作都会在这个集群里
  - 我们最常用的会是 Hive，一些比较大的项目都会放在这里。

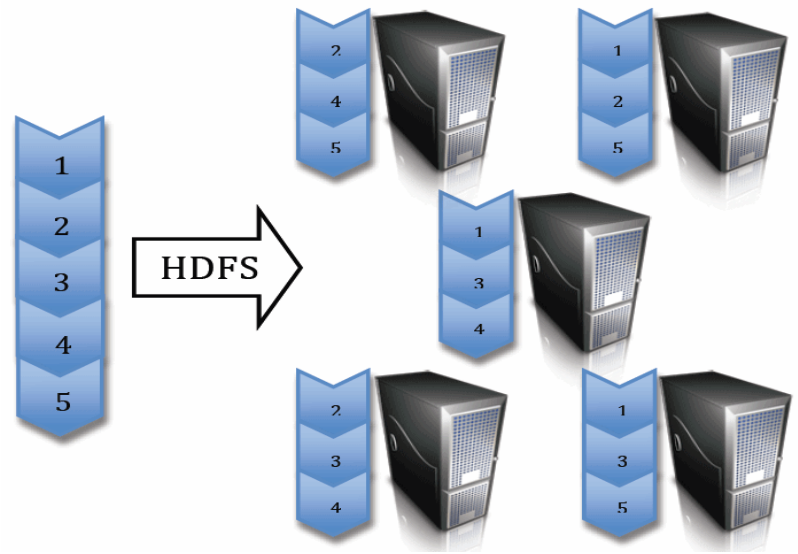
- **Optimization**
  - 和运营部门合作，收集用户反馈，进行优化。

3 Cluster Operations

集群计算（Cluster）

将许多计算机被网络连接到一起形成集群，文件系统存储、数据处理在集群中而不是个体计算机。整体是一个分布式并行计算过程。关于 Hadoop 分布式平台运作原理，建议阅读 Google 的三篇论文：Bigtable: A Distributed Storage System for Structured Data，The Google File System，MapReduce: Simpliyed Data Processing on Large Clusters。

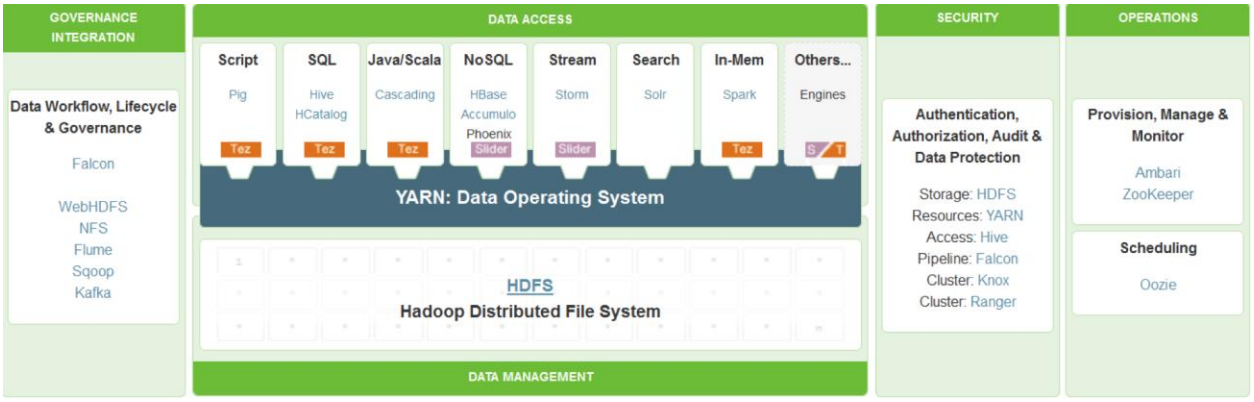
- **优势：**
  - 分布式并行计算速度大大提升。e.g. 一份数据分成许多数据块，每份用一台计算机来计算，每台计算机计算的数据量很小，进而速度很快。
  - 可以存储相当大规模的数据，只需要根据需求增加机柜即可。
  - 提示：在写简历时要注明自己对具体大数据平台的操作技能，如 Hadoop 和 Spark 系统。



基本操作

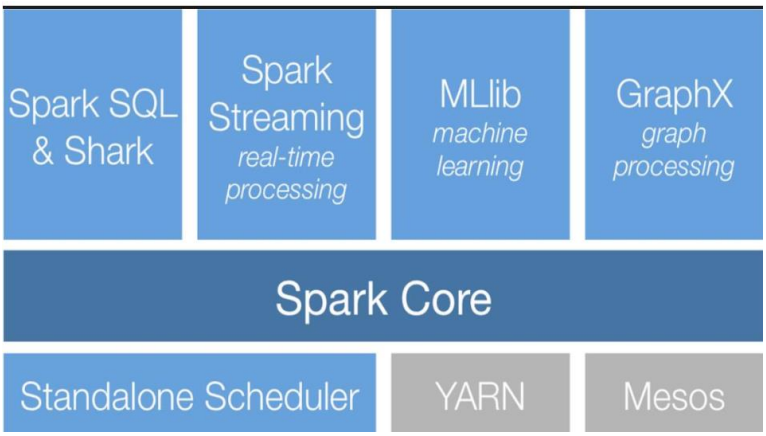
- **参考视频教程：**
  - How to install putty on Windows https://www.youtube.com/watch?v=a4K9mvKxrwI
  - How to use winscp https://www.youtube.com/watch?v=e7AgOFS_g8Q
  - How to use SSH on Mac https://www.youtube.com/watch?v=J_8ZsXP1EYk
  - How to use scp on Mac https://www.youtube.com/watch?v=EJOoiYtyPTE
  - VI tutorial https://www.youtube.com/watch?v=TBu6qxd5uAc
- 登入 montana.dataapplab.com

- **MacOS 的 ssh 远程登录指令：**ssh username@master1.internal.datalaus.net -P 49233（提示：ssh 登录的是整个集群中某一台机器）
- **常用工具**
  - HDFS（Hadoop Distributed File System）：分布式文件系统。文件的 I/O 操作。HDFS 将文件分成（split）许多数据块（通常每份 128M），将每一份数据块（Data Block）拷贝到三台机器上存储。在其中一台或几台机器出现故障时，在其他机器上有同样数据块的拷贝。四家大数据系统提供商：Hortonworks，MapR，Cloudera，Databrick（将开源平台封装好后，帮助进行安装运行调试）。本课程的 Cluster 由 Hortonworks 提供。
    - 优点 1 大大提高数据的安全性，防止数据丢失。
    - 优点 2 可以存储大量的数据。增加机柜数量便可增加存储量。
  - HDFS Design Goal:
    - Recover from hardware failure 即使有一两台机器坏掉，成千台机器仍可以工作
    - Streaming data access 流动数据可以通过
    - Large data file/dataset 可存储海量数据
    - Write-once-read-many IO model 写入一次数据之后可以反复读取
    - Move computation to data 可以在集群里进行计算和分析
    - Commodity hardware 硬件消费
    - Do not conflict with OS file system 对自己的电脑操作系统没有影响
    - Low-latency and many small files 低延迟，小文件组成



  - Hive：数据仓库框架，基于 sql 语法的大数据处理工具。
  - Pig（脚本语言，做数据 loading 用），Spark，后续课程会讲解。
  - MapReduce：不是 DS 的工作范围。可以了解工作原理。

- Spark, 我们会用 Spark 去操作 SQL & ML



- Hadoop 和 Spark 的区别：都是 distributed data platform；最大区别：spark 原理：computation in memory（内存计算比硬盘计算要快），使用的语言也略有不同。
- 如果每个数据块有三个拷贝，到哪台机器取？从离最近的的机器中取。

- **HDFS 操作指令**
  - 提示：分清楚是操作的是其中一台机器，还是操作 HDFS（一个集群，也就是所有主机），例如，一个新的文件是在一台主机上建立目录还是在 HDFS 上建立。二者在目录上有区别：HDFS 的目录为/user/yourusername；对其中一台主机进行本地操作的目录为/home/yourusername。
  - 基本指令：
    - hdfs dfs - ls 打印当前目录内容
    - e.g. hdfs dfs -ls /user/hadoop/file
    - hdfs dfs – put localsrc dst 将本地文件传送到 HDFS 中
    - e.g. hdfs dfs – put /home/jason/test2.csv /user/jason/temp2
    - hdfs dfs – mkdir 创建新目录
    - e.g. hdfs dfs -mkdir /user/hadoop/dir1 /user/hadoop/dir2
    - hdfs dfs – rmdir 移除目录
    - e.g. hadoop fs -rmdir /user/hadoop/emptydir
    - hdfs dfs – rm 移除文件（提示：删除后很难恢复）
    - e.g. hadoop fs -rm /user/hadoop/emptydir/file
    - hdfs dfs -get 将文件从 HDFS 拷贝到本地
    - e.g. hadoop fs -get /user/hadoop/file localfile

- 更多指令集參考： https://hadoop.apache.org/docs/r2.4.1/hadoop-project-dist/hadoop-common/FileSystemShell.html