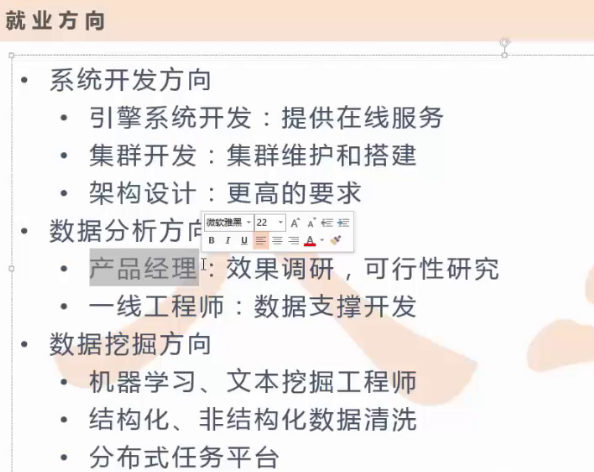
1. ：



就业方向：

系统开发方向：

1. 引擎系统开发：提供在线服务
2. 集群开发：集群维护和搭建
3. 架构设计：更高的要求

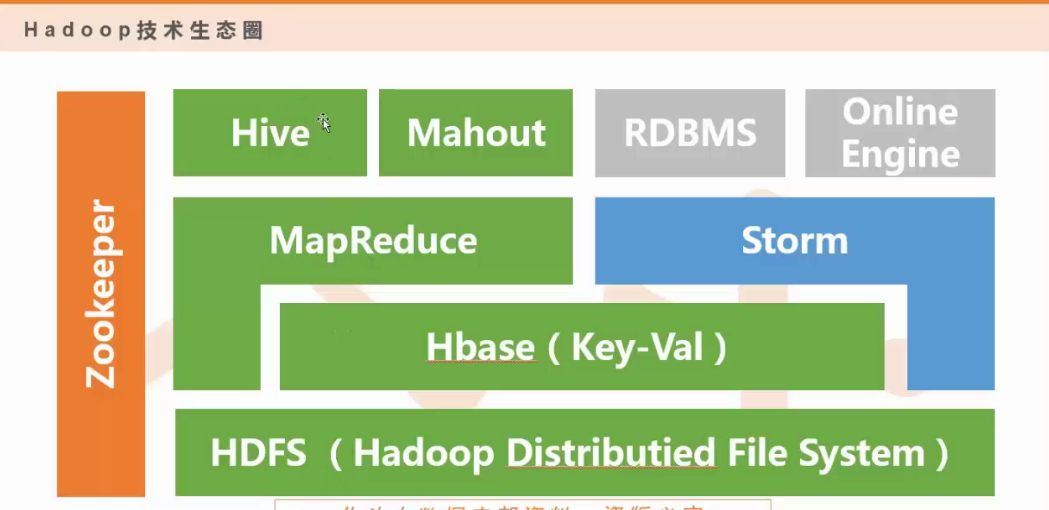
数据分析方向：

1. 产品经理：效果调研，可行性研究
2. 一线工程师：数据支撑开发（产生数据的能力）

数据挖掘方向：

1. 机器学习，文本挖掘工程师
2. 结构化，非结构化数据清洗

生态圈：



解图：

看：从低往上

数据存储：

HDFS（分布式文件系统，文件系统的一些格式）作为一个最基本的文件系统，存储数据的地方

Hbase：列存储（数据库），更侧重于结构化的数据，存储分门归类。Hbase的数据基本存储在HDFS上。

数据应用：

计算框架：

MapReduce：读取HDFS上存储的文件，计算，更适合做一线

Storm：更适合做在线

HIVE：语言转换工具，支持一种sql语句（hsql），本质是MapReduce(启动一个或者多个MapReduce帮你做数据处理)

Mahout：在hadoop里封装了一些算法库，主流算法都支持，包括聚类。本质也是MapReduce

Storm：实时处理数据==MapReduce

Online Engine：给你推荐一个你喜欢的类似的东西，通过Online Engine给你服务

Zookeeper：分布式锁服务，全局协调集群用的。

商业项目体现：



解读：

数据预处理：

HDFS--》HBASE：将hdfs的非结构化数据变成结构化数据存储到hbase

Storm：实时性要求高（spark-streaming实时性要求不高）

Maprereduce：离线数据挖掘（spark是第二代MapReduce的计算框架，也可以批量处理数据）。Storm和

Spark-streaming可以通过流式的方式处理数据

通过在线和离线的数据处理方式，有用数据已经能够对外提供服务

如何对外服务：Nosql：并发度高（一秒内同时处理300个请求（tps），服务响应时间（最多几百毫秒）），对外提供服务

谁来请求这些数据：

Webpy在线引擎：页面存到后面的数据库里。

检索系统--》hdfs：用户在网站上有了收藏，播放等行为，需要将行为存储起来（日志存储到hdfs上）