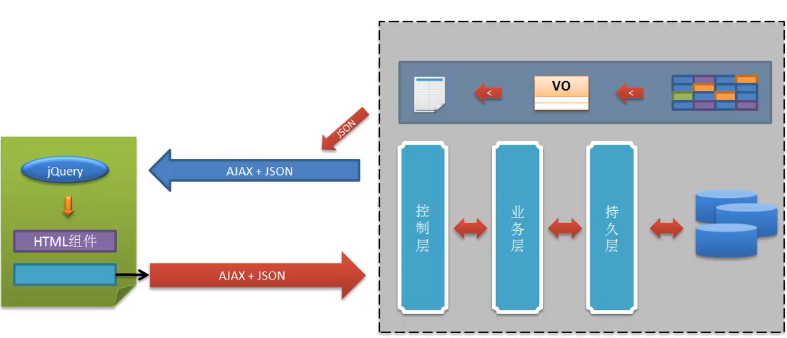
# 大数据入门

## 认识大数据

### 1.1 Web 2.0造就大数据（Ajax）



### 1.2单服务器时代



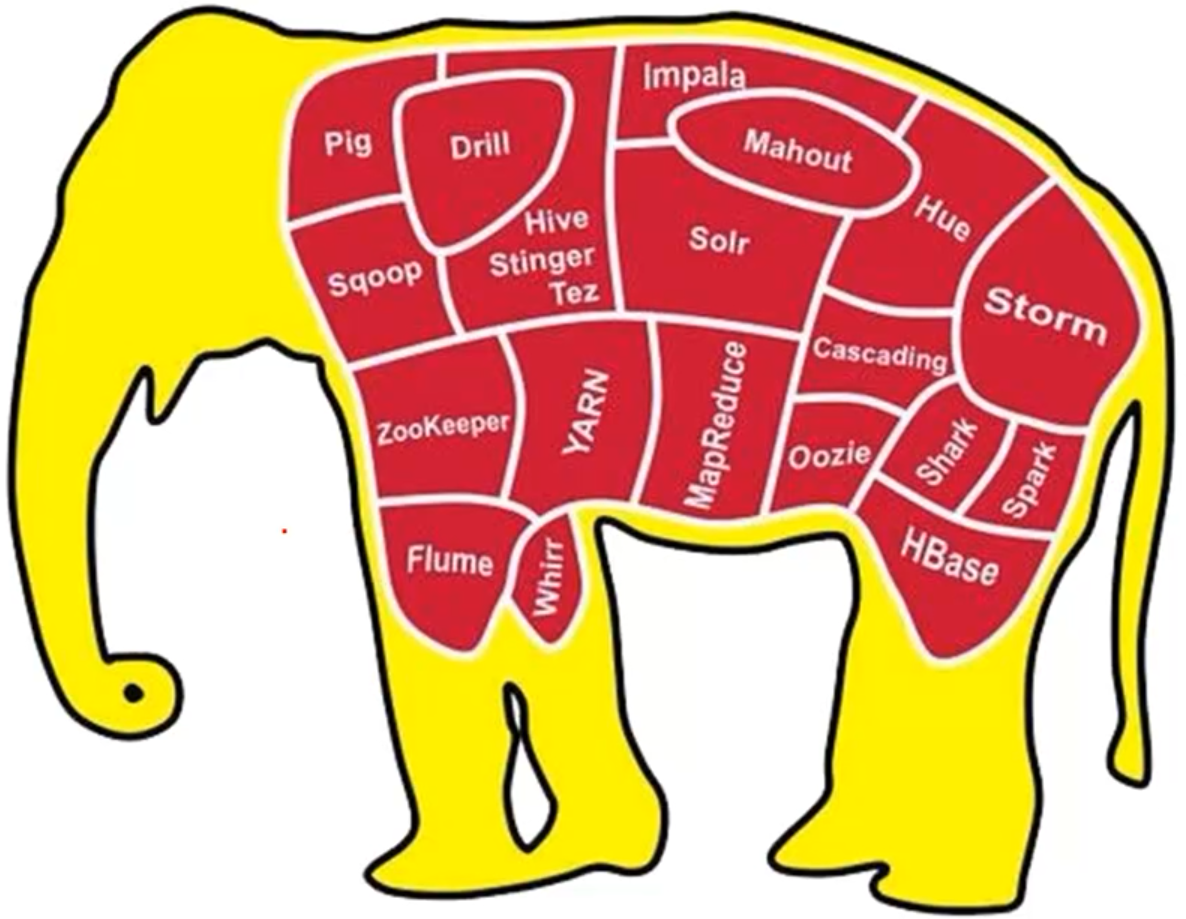
### 1.3数据的价值

企业成长模式以及数据分析的重要性

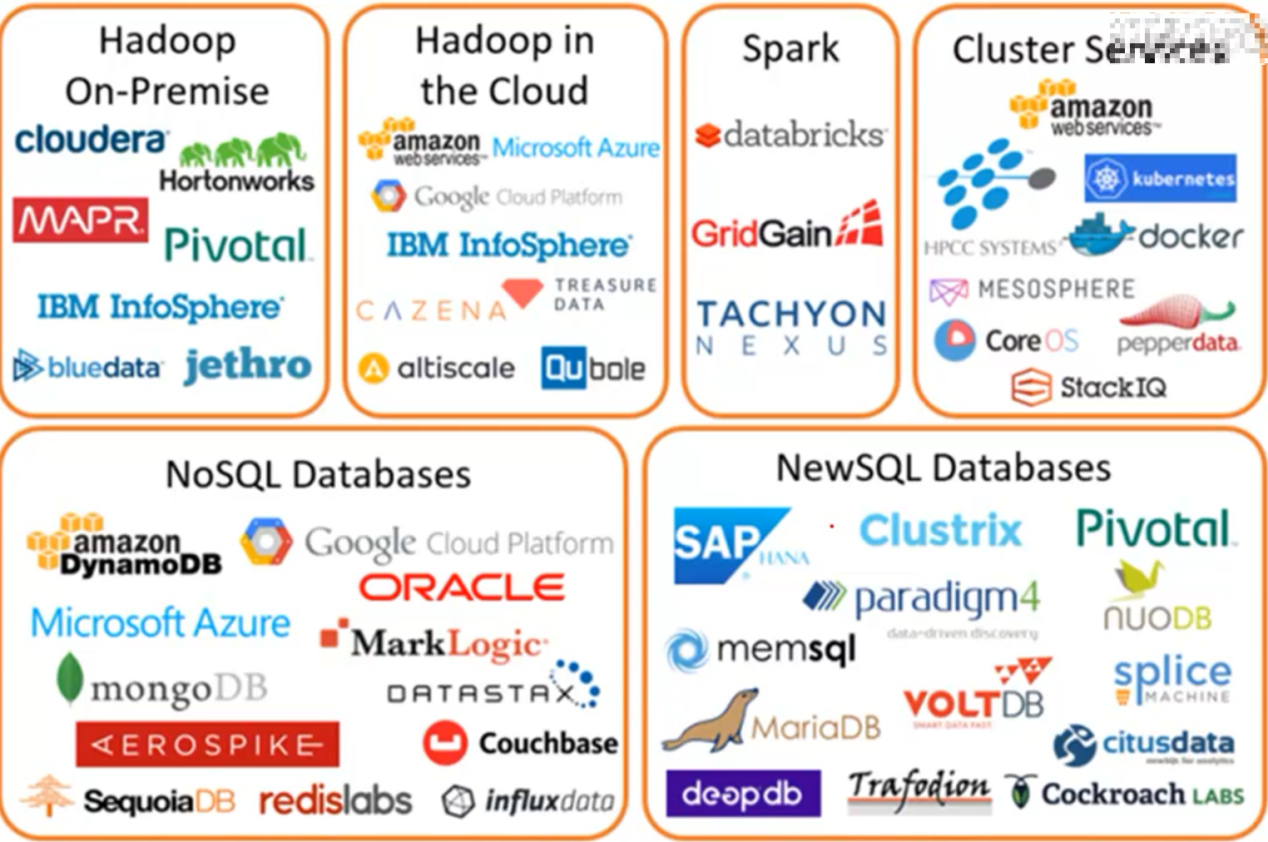
技术 + 分析 + 售前

## 大数据开发技术与架构

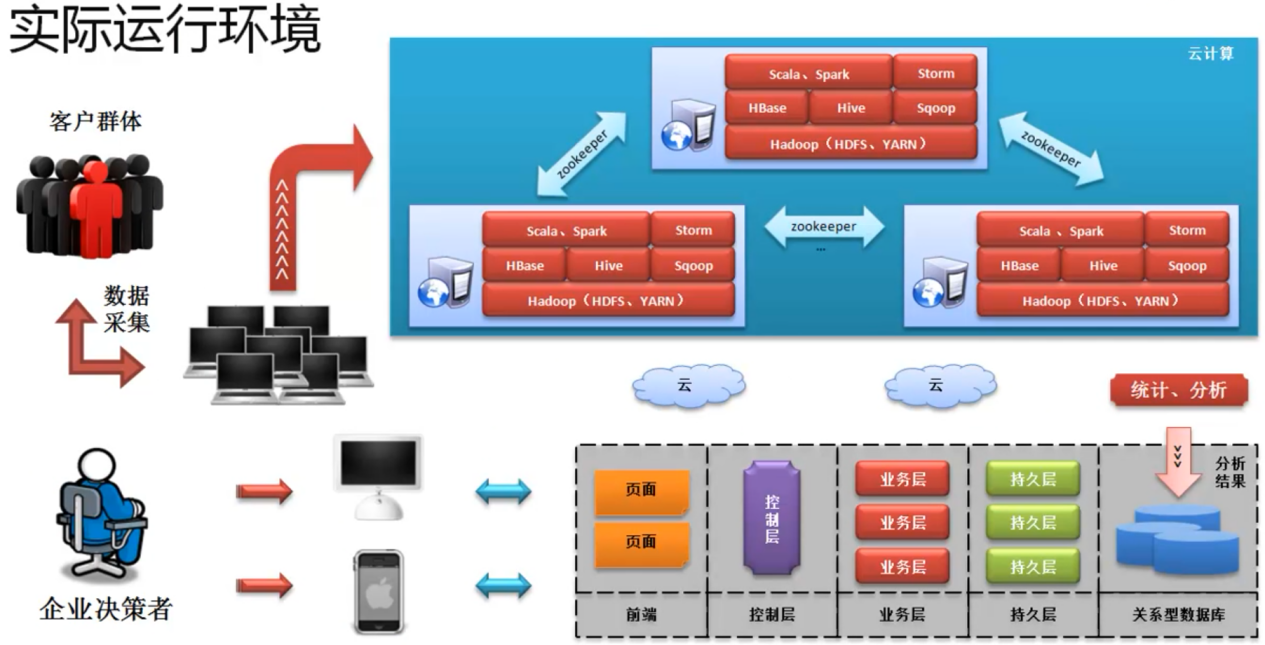
### 2.1大数据生态圈



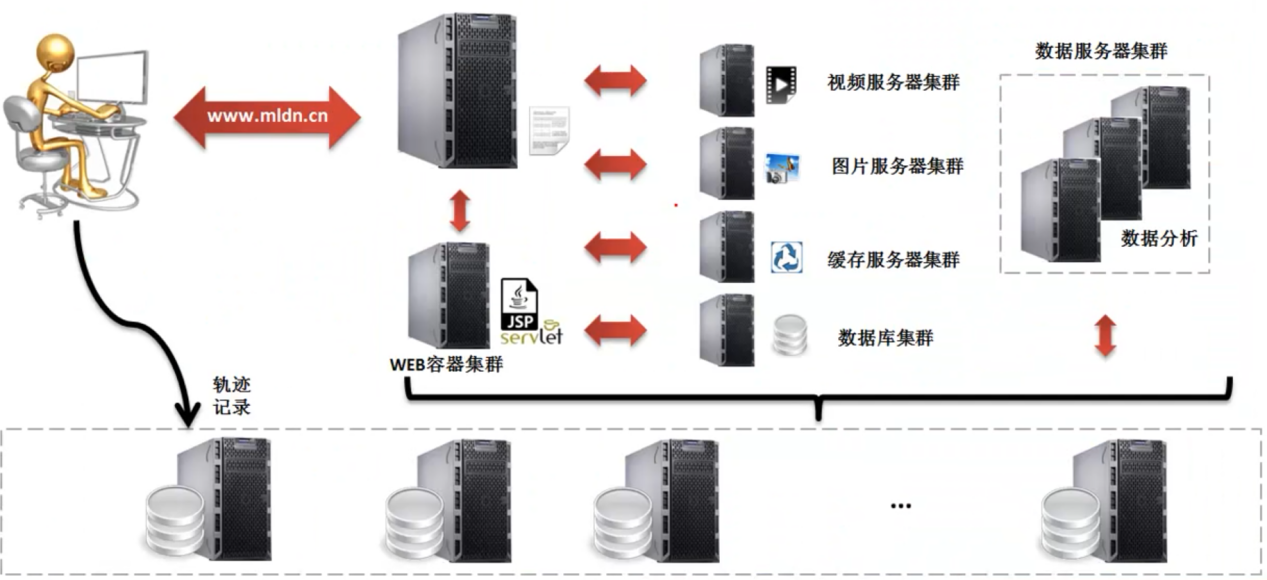
### 2.2大数据版图



### 2.3实际运行环境



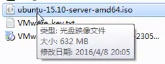
### 2.4大数据与传统项目整合



## 第3节 准备环境搭建

### 3.1Linux操作系统安装

1. 新建虚拟机，选择核数、内存等；



1. 使用NAT网络地址转换模式；
2. 英文，不要使用键盘；
3. 设置主机名：
4. 纯净安装，不要选择额外软件；
5. 超级管理员权限，root
6. 修改字体大小：$ sudo dpkg-reconfigure console-setup
7. 修改root密码：$ sudo passwd root & logout

### 3.2ssh连接控制

1. 如果在你主机IP不固定的情况下，千万不要进行免登录配置。
2. $ apt-get update & apt-get install openssh-server
3. $ vi /etc/ssh/sshd\_config

|-PermitRootLogin yes 这个时候就可以使用CRT、xshell登录

1. $ pkill -9 sshd & /etc/init.d/sshd start
2. $ ifconfig | more
3. $ apt-get remove iptables

### 3.3 配置环境

1）$ apt-get install ntp ntpdate1

2）时间同步服务器配置：

$ ntp-date -u 133.100.11.8

3）$ apt-get install gcc libpcre3 libpcrecpp\* libpcre3-dev libssl-dev

### 3.4配置FTP服务

1. $ apt-get install vsftpd
2. $ passwd ftp “ftp@110”
3. $ chmod 755 /srv/ftp
4. $ vi /etc/vsftpd.conf

|-Write\_enable=YES

|-Chroot\_local\_user=YES

|-Chroot\_list\_enable=YES

|-chroot\_list\_file=/etc/vsftpd.chroot\_list

1. $ vim /etc/vsftpd.chroot\_list

### 3.5JDK安装与配置

1. $ tar xzvf /srv/ftp/jdk-8u73-linux-x64.tar.gz -C /usr/local

2）创建软连接或重命名：ln -s 源文件 目标文件

$ ln -s jdk\_1.8.0\_73 jdk 或者重命名 mv jdk\_1.8.0\_73 jdk

3）$ vi /etc/profile

|-export JAVA\_HOME=/usr/local/jdk

|-export PATH=$PATH:$JAVA\_HOME/bin:

4）$ source /etc/profile

### 3.6MySQL安装配置

1. $ tar zxvf /srv/ftp/mysql-5.6.10-linux-glibc2.5-x86\_64.tar.gz -C /usr/local
2. $ ln -s mysql-5.6.10 mysql 或者重命名mv mysql-5.6.10 mysql
3. vi /etc/profile

|-export MYSQL\_HOME=/usr/local/mysql

|-export PATH=$PATH:$MYSQL\_HIME/bin:

$ source /etc/profile

1. $ apt-get install libaiol libaio-dev
2. /usr/local/mysql/scripts/mysql\_install\_db --user=root --basedir=/usr/local/mysql --datadir=/usr/local/mysql/data
3. $ nohup mysqld\_safe --user=root > /dev/null 2>&1 &
4. mysqladmin 查看mysql版本
5. $ mysql -uroot -p 直接回车（初始密码为空）

mysql>update mysql.user set password=PASSWORD(‘mysqladmin’) where user=’root’;

## 第4节 认识Hadoop

### 4.1Hadoop简介

1. hadoop.apache.org
2. Map/Reduce HDFS YARN；重点Map/Reduce

### 4.2配置SSH连接

1）$ rm -f ~/.ssh

2）$ ssh-keygen -t rsa

3）$ cat ~/.ssh/id\_rsa.pub >> ~/.ssh/authorized\_keys

4）$ ssh root@hostname

### 4.3Hadoop安装

1. $ wget <https://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-2.7.7/hadoop-2.7.7.tar.gz>
2. $ tar zxvf /srv/ftp/hadoop-2.7.7.tar.gz -C /usr/local
3. $ ln -s hadoop-2.7.7 hadoop 或者 mv hadoop-2.7.7. hadoop
4. $ vi /etc/profile

|-Export HADOOP\_HOME=/usr/local/hadoop

|-export PATH=$PATH:$HADOOP\_HOME/sbin:

$ source /etc/profile

1. $ vi /usr/local/hadoop/etc/hadoop/hadoop-env.sh

|-export JAVA\_HOME=/usr/local/jdk

1. $ mkdir -p /usr/test/hadoop/input

$ cp /usr/local/hadoop/\*.txt /usr/local/test/hadoop/input

7）$ hadoop jar /usr/local/hadoop/share/hadoop/mapreduce/sources/hadoop-mapreduce-examples-2.7.7-sources.jar org.apache.hadoop.examples.WordCount /usr/test/hadoop/input /usr/test/hadoop/output

8）cat /usr/test/hadoop/output/part-000000

### 4.4Hadoop伪分布模式

1. hadoop配置文件目录：

$ ll /usr/local/hadoop/etc/hadoop

1. 禁止配置hadoop.tmp.dir为“/tmp”，一旦被清理hadoop的配置就完了；
2. $ vi /usr/local/hadoop/etc/hadoop/core-site.xml

设置hadoop的临时目录；hdfs的默认文件系统的路径

$ mkdir -p /usr/data/hadoop/tmp（重新搭建hadoop需要清空该目录）

|  |
| --- |
| <property>  <name>fs.defaultFS</name>  <value>hdfs://hostname:9000</value>  </property>  <property>  <name>hadoop.tmp.dir</name>  <value>/usr/data/hadoop/tmp</value>  </property> |

4）$ vi /usr/local/hadoop/etc/hadoop/hdfs-site.xml

新建namenode的保存目录：mkdir -p /usr/data/hadoop/hdfs/name（重新搭建hadoop需要清空该目录）

新建datanode的保存目录：mkdir -p /usr/data/hadoop/hdfs/data（重新搭建hadoop需要清空该目录）

“dfs.replication”：文件保存的副本数，副本保存在DataNode中，现在只有一个主机，所以只存1份

“dfs.namenode.name.dir”：保存namenode节点信息的相关操作；

“dfs.datanode.data.dir”：保存真实数据；

“dfs.namenode.http-address”：Hadoop启动之后会会自动启动一个HTTP服务，通过浏览器可以访问

“dfs.namenode.secondary.http-address”：SecondaryNameNode节点的HTTP服务

“dfs.permissions”：表示hdfs操作权限，设置false表示不验证

|  |
| --- |
| <property>  <name>dfs.replication</name>  <value>1</value>  </property>  <property>  <name>dfs.namenode.name.dir</name>  <value>file:///usr/data/hadoop/hdfs/name</value>  </property>  <property>  <name>dfs.datanode.data.dir</name>  <value>file:///usr/data/hadoop/hdfs/data</value>  </property>  <property>  <name>dfs.namenode.http-address</name>  <value>hostname:50070</value>  </property>  <property>  <name>dfs.namenode.secondary.http-address</name>  <value>hostname:50090</value>  </property>  <property>  <name>dfs.permissions</name>  <value>false</value>  </property> |

5）$ vi /usr/local/hadoop/etc/hadoop/yarn-site.xml

所有的配置都可以从官网获得

|  |
| --- |
| <property>  <name>yarn.resourcemanager.admin.address</name>  <value>hostname::8033</value>  </property>  <property>  <name>yarn.nodemanager.aux-services</name>  <value>mapreduce\_shuffle</value>  </property>  <property>  <name>yarn.nodemanager.aux-services.mapreduce\_shuffle.class</name>  <value>org.apache.hadoop.mapred.ShuffleHandler</value>  </property>  <property>  <name>yarn.resourcemanager.resource-tracker.address</name>  <value>hostname:8025</value>  </property>  <property>  <name>yarn.resourcemanager.scheduler.address</name>  <value>hostname:8030</value>  </property>  <property>  <name>yarn.resourcemanager.address</name>  <value>hostname:8050</value>  </property>  <property>  <name>yarn.resourcemanager.webapp.address</name>  <value>hostname:8088</value>  </property>  <property>  <name>yarn.resourcemanager.webapp.https.address</name>  <value>hostname:8090</value>  </property> |

6）$ vi /usr/lcoal/hadoop/etc/hadoop/slaves

#如果有多台主机，需要编写多台主机的hostname

hostname

1. hadoop环境初始化

$ hdfs namenode -format

....

Exiting with status 0

1. hadoop启动，启动路径：/usr/local/hadoop/sbin

$ start-all.sh

...

启动完成后，可能启动失败会自己关闭服务，使用jps查看java进程

NameNode

DataNode

SecondaryNameNode

NodeManager

ResourceManager

1. 使用浏览器查看http://hostname:50070