# Hadoop（入门）

## 大数据概论

### 大数据概念

大数据（big data）：指**无法在一定时间范围**内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的**海量、高增长率和多样化的信息资产**。

主要解决，海量数据的**存储**和海量数据的**分析计算**问题。

按顺序给出数据存储单位：bit、Byte、KB、MB、GB、TB、PB、EB、ZB、YB、BB、NB、DB。1Byte = 8bit 1K = 1024Byte 1MB = 1024K 1G = 1024M 1T = 1024G 1P = 1024T

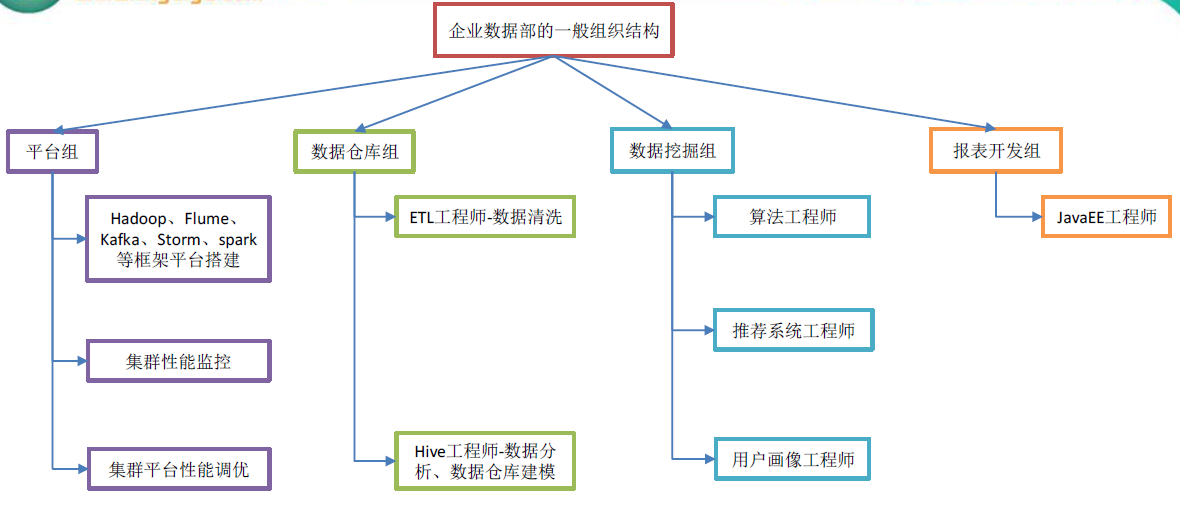
### 大数据的特点

大量、高速、多样、低价值密度。

### 大数据能干什么

O2O、零售、旅游、商品广告推荐、保险、金融、房产、人工智能

### 企业数据部的业务流程分析



## Hadoop基础

### Hadoop是什么

1）Hadoop是一个由Apache基金会所开发的**分布式系统基础架构**

1. 主要解决，海量数据的**存储**和海量数据的**分析计算**问题。
2. 广义上来说，HADOOP通常是指一个更广泛的概念——**HADOOP生态圈**

### Hadoop 三大发行版本

Apache、Cloudera、Hortonworks。

Apache版本最原始（最基础）的版本，对于入门学习最好。

Cloudera在大型互联网企业中用的较多。

Hortonworks文档较好。

### Hadoop 的优势

1. 高可靠性：因为Hadoop假设计算元素和存储会出现故障，因为它维护多个工作数据副本，在出现故障时可以对失败的节点重新分布处理。

2）高扩展性：在集群间分配任务数据，可方便的扩展数以千计的节点。

3）高效性：在MapReduce的思想下，Hadoop是并行工作的，以加快任务处理速度。

4）高容错性：自动保存多份副本数据，并且能够自动将失败的任务重新分配。

### Hadoop组成

1）Hadoop HDFS：一个高可靠、高吞吐量的分布式文件系统。

2）HadoopMapReduce：一个分布式的离线并行计算框架。

3）HadoopYARN：作业调度与集群资源管理的框架。

4）Hadoop Common：支持其他模块的工具模块（Configuration、RPC、序列化机制、日志操作）。

### 2.4.1 HDFS 架构概述

1）NameNode（nn）：存储文件的元数据，如文件名，文件目录结构，文件属性（生成时间、副本数、文件权限），以及每个文件的块列表和块所在的DataNode等。

2）DataNode(dn)：在本地文件系统存储文件块数据，以及块数据的校验和。

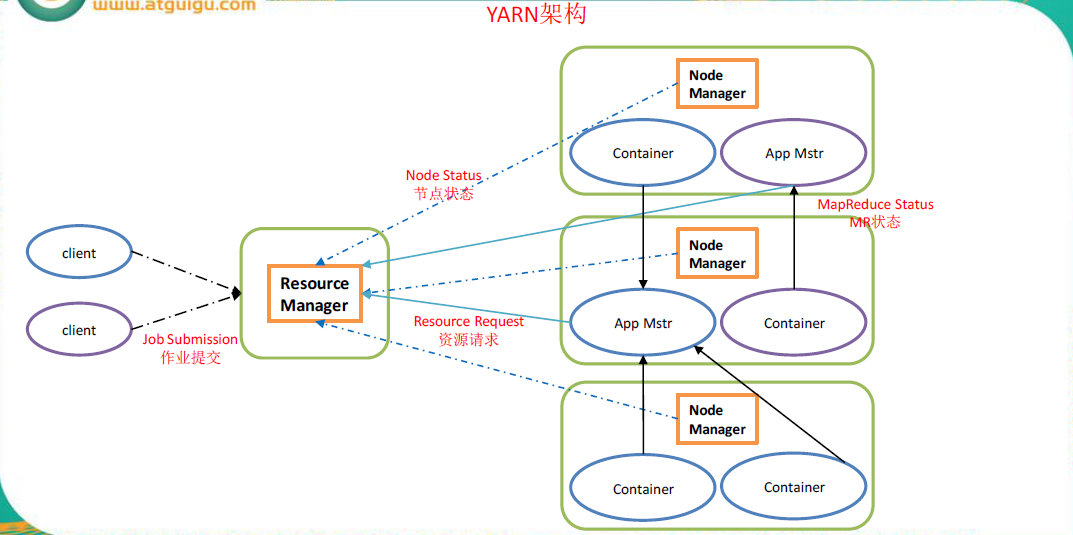
3）Secondary NameNode(2nn)：用来监控HDFS状态的辅助后台程序，每隔一段时间获取HDFS元数据的快照。

### 2.4.2 YARN架构概述

1）ResourceManager(rm)：处理客户端请求、启动/监控ApplicationMaster、监控NodeManager、资源分配与调度；

2）NodeManager(nm)：单个节点上的资源管理、处理来自ResourceManager 的命令、处理来自ApplicationMaster 的命令；

3）ApplicationMaster：数据切分、为应用程序申请资源，并分配给内部任务、任务监控与容错。

4）Container：对任务运行环境的抽象，封装了CPU、内存等多维资源以及环境变量、启动命令等任务运行相关的信息。

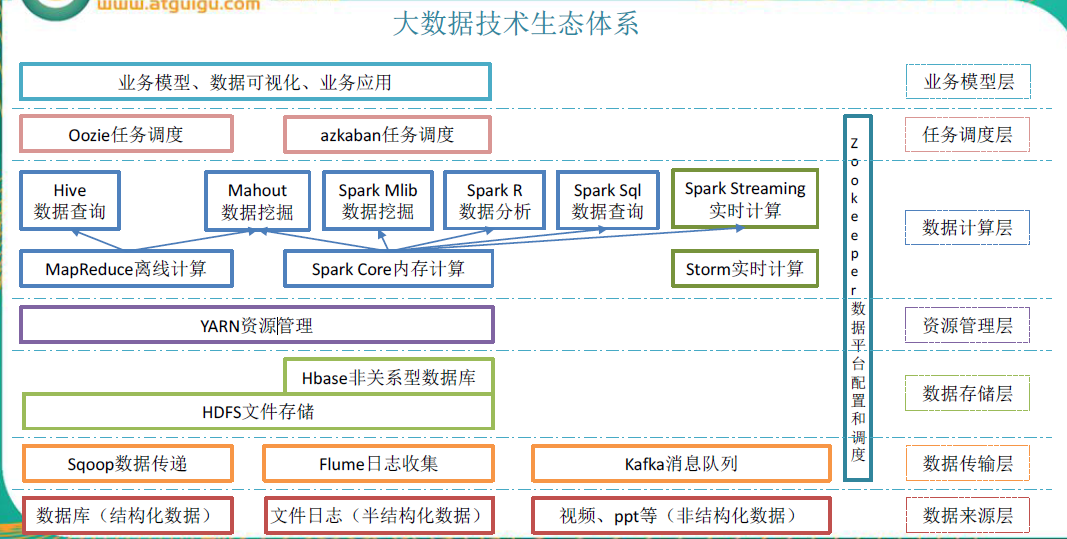
### 2.4.3 MapReduce架构概述

MapReduce将计算过程分为两个阶段：Map和Reduce

1）Map阶段并行处理输入数据

2）Reduce阶段对Map结果进行汇总

### 大数据技术生态体系



图中涉及的技术名词解释如下：

1）Sqoop：sqoop是一款开源的工具，主要用于在Hadoop(Hive)与传统的数据库(mysql)间进行数据的传递，可以将一个关系型数据库（例如：MySQL ,Oracle 等）中的数据导进到Hadoop的HDFS中，也可以将HDFS的数据导进到关系型数据库中。

2）Flume：Flume是Cloudera提供的一个高可用的，高可靠的，分布式的海量日志采集、聚合和传输的系统，Flume支持在日志系统中定制各类数据发送方，用于收集数据；同时，Flume提供对数据进行简单处理，并写到各种数据接受方（可定制）的能力。

3）Kafka：Kafka是一种高吞吐量的分布式发布订阅消息系统，有如下特性：

（1）通过O(1)的磁盘数据结构提供消息的持久化，这种结构对于即使数以TB的消息存储也能够保持长时间的稳定性能。

（2）高吞吐量：即使是非常普通的硬件Kafka也可以支持每秒数百万的消息

（3）支持通过Kafka服务器和消费机集群来分区消息。

（4）支持Hadoop并行数据加载。

4）Storm：Storm为分布式实时计算提供了一组通用原语，可被用于“流处理”之中，实时处理消息并更新数据库。这是管理队列及工作者集群的另一种方式。Storm也可被用于“连续计算”（continuous computation），对数据流做连续查询，在计算时就将结果以流的形式输出给用户。

5）Spark：Spark是当前最流行的开源大数据内存计算框架。可以基于Hadoop上存储的大数据进行计算。

6）Oozie：Oozie是一个管理Hdoop作业（job）的工作流程调度管理系统。Oozie协调作业就是通过时间（频率）和有效数据触发当前的Oozie工作流程。

7）Hbase：HBase是一个分布式的、面向列的开源数据库。HBase不同于一般的关系数据库，它是一个适合于非结构化数据存储的数据库。

8）Hive：hive是基于Hadoop的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供简单的sql查询功能，可以将sql语句转换为MapReduce任务进行运行。其优点是学习成本低，可以通过类SQL语句快速实现简单的MapReduce统计，不必开发专门的MapReduce应用，十分适合数据仓库的统计分析。

10）R语言：R是用于统计分析、绘图的语言和操作环境。R是属于GNU系统的一个自由、免费、源代码开放的软件，它是一个用于统计计算和统计制图的优秀工具。

11）Mahout:

Apache Mahout是个可扩展的机器学习和数据挖掘库，当前Mahout支持主要的4个用例：

推荐挖掘：搜集用户动作并以此给用户推荐可能喜欢的事物。

聚集：收集文件并进行相关文件分组。

分类：从现有的分类文档中学习，寻找文档中的相似特征，并为无标签的文档进行正确的归类。

频繁项集挖掘：将一组项分组，并识别哪些个别项会经常一起出现。

12）ZooKeeper：Zookeeper是Google的Chubby一个开源的实现。它是一个针对大型分布式系统的可靠协调系统，提供的功能包括：配置维护、名字服务、分布式同步、组服务等。ZooKeeper的目标就是封装好复杂易出错的关键服务，将简单易用的接口和性能高效、功能稳定的系统提供给用户。