云计算与大数据期末考试整理

> ## 免责声明：
>
> 1. 整理是我的个人行为，查看是你的个人行为。如果因为本份整理中的内容导致考试丢分，概不负责。
> 2. 所有内容均来自百度，想要书本原文的，请自行翻书。

1 概述

1.1 大数据的特征

> 主要特征：大量、多样、高速、价值

- 容量（Volume）：数据的大小决定所考虑的数据的价值和潜在的信息；
- 种类（Variety）：数据类型的多样性；
- 速度（Velocity）：指获得数据的速度；
- 价值（value）：合理运用大数据，以低成本创造高价值。
- 可变性（Variability）：妨碍了处理和有效地管理数据的过程。
- 真实性（Veracity）：数据的质量。
- 复杂性（Complexity）：数据量巨大，来源多渠道。

1.2 大数据技术

> 从大数据的生命周期来看，无外乎四个方面：**大数据采集、大数据预处理、大数据存储、大数据分析**，共同组成了大数据生命周期里最核心的技术

1.2.1 大数据采集

大数据采集，即对各种来源的**结构化**和非结构化海量数据，所进行的采集。

- **数据库采集：**流行的有Sqoop和ETL；传统的关系型数据库MySQL和Oracle；开源的Kettle和Talend可实现hdfs，hbase和主流Nosq数据库之间的数据同步和集成。
- **网络数据采集：**一种借助网络爬虫或网站公开API，从网页获取非结构化或半结构化数据，并将其统一结构化为本地数据的数据采集方式。
- **文件采集：**包括实时文件采集和处理技术flume、基于ELK的日志采集和增量采集等等。

1.2.2 大数据预处理

大数据预处理，指的是在进行数据分析之前，先对采集到的原始数据所进行的诸如“清洗、填补、平滑、合并、规格化、一致性检验”等一系列操作，旨在提高数据质量，为后期分析工作奠定基础。

- **数据清理：**指利用ETL等清洗工具，对有遗漏数据(缺少感兴趣的属性)、噪音数据(数据中存在着错误、或偏离期望值的数据)、不一致数据进行处理。
- **数据集成：**是指将不同数据源中的数据，合并存放到统一数据库的，存储方法，着重解决三个问题：模式匹配、**数据冗余**、数据值冲突检测与处理
- **数据转换：**是指对所抽取出来的数据中存在的不一致，进行处理的过程。它同时包含了数据清洗的工作，即根据业务规则对异常数据进行清洗，以保证后续分析结果准确性
- **数据规约：**是指在最大限度保持数据原貌的基础上，最大限度精简数据量，以得到较小数据集的操作，包括：数据方聚集、维规约、数据压缩、**数值规约**、概念分层等。

1.2.3 大数据存储

大数据存储，指用存储器，以数据库的形式，存储采集到的数据的过程，包含三种典型路线：

1、基于MPP架构的新型数据库集群

2、基于Hadoop的技术扩展和封装

3、大数据一体机