1

数据特征：

1. 规模庞大：大数据管理系统处理的数据规模通常非常庞大，包括结构化数据、半结构化数据和非结构化数据。
2. 高速度：大数据管理系统需要能够实时或近实时地处理大量数据流，以满足快速决策和应用需求。
3. 多样性：大数据管理系统需要能够处理各种类型的数据，包括文本、图像、音频、视频等多媒体数据。
4. 高维度：大数据管理系统需要能够处理具有高度维度的数据，例如传感器数据、遥感数据等。

系统特征：

1. 分布式架构：大数据管理系统通常基于分布式计算架构，利用多台计算机节点协同工作，以实现高性能和可扩展性。
2. 并行处理：大数据管理系统利用并行计算技术，将任务分解为多个子任务并行处理，提高数据处理的效率。
3. 容错性：大数据管理系统需要具备容错机制，以应对硬件故障、网络中断等异常情况，并确保数据的完整性和可靠性。
4. 数据存储和处理：大数据管理系统需要提供高效的数据存储和处理能力，包括分布式文件系统、列式存储、内存计算等技术。

应用特征：

1. 数据分析和挖掘：大数据管理系统可用于进行大规模数据分析和挖掘，发现数据中的潜在模式、关联和趋势，从而提供决策支持和洞察力。
2. 实时处理：大数据管理系统可用于实时处理数据流，例如实时监控、广告投放、交易处理等应用场景，要求系统能够以毫秒级响应时间处理数据。
3. 智能推荐和个性化服务：大数据管理系统可用于构建个性化推荐系统，分析用户行为和兴趣，提供个性化的产品推荐、内容推荐等服务。
4. 数据安全和隐私保护：大数据管理系统需要具备数据安全和隐私保护的能力，包括数据加密、访问控制、匿名化处理等技术，以确保数据的保密性和合规性。

2

1. 数据采集与存储层：
    - 数据采集：负责从不同的数据源（传感器、日志文件、社交媒体等）收集原始数据，并进行预处理和清洗。
    - 分布式文件系统：例如Hadoop的Hadoop Distributed File System（HDFS），用于高可靠性和高容量的数据存储。
2. 数据处理与计算层：
    - 批处理：通过批处理技术（如MapReduce）对大规模数据进行离线处理和分析，以发现模式、关联和趋势。
    - 流处理：通过流处理技术（如Apache Kafka、Apache Flink）对实时数据流进行实时处理和分析，以支持实时决策和应用。
    - 图处理：通过图处理技术（如Apache Giraph、Neo4j）对复杂网络和关系进行图分析和计算，以揭示隐藏的关联和结构。

- 机器学习与人工智能：通过机器学习和人工智能算法对大数据进行模型训练和预测，以实现智能化的数据分析和决策支持。

3. 数据管理与存储层：
   - 列式存储：将数据按列存储，提高数据的读取效率和压缩率，例如Apache HBase、Apache Cassandra等。
   - 内存数据库：将数据存储在内存中，提供高速的数据读写和查询能力，例如Redis、MemSQL等。
   - NoSQL数据库：非关系型数据库，适用于大规模、分布式和非结构化数据的存储和查询，例如MongoDB、Couchbase等。
4. 数据分析与应用层：
   - 数据可视化：通过可视化技术（如Tableau、Power BI）将数据转化为图表、仪表盘等形式，以直观展示数据分析结果。
   - 数据挖掘与业务智能：使用数据挖掘技术（如关联规则挖掘、聚类分析）和业务智能工具，从数据中发现有价值的信息和洞察力。
   - 实时监控与预警：通过实时监控和预警系统，对数据进行实时监测和异常检测，及时发现和解决问题。
   - 个性化推荐与智能服务：通过个性化推荐算法和智能化服务，根据用户兴趣和行为，提供个性化的产品推荐、内容推荐等服务。

3

1）关系数据库领域面对半结构化数据和非结构化数据的典型解决思路包括以下几个方面：

- 扩展关系型数据库：传统的关系型数据库可以通过引入新的数据类型、扩展数据模型和查询语言来支持半结构化数据和非结构化数据的存储和查询。例如，引入XML数据类型、JSON数据类型，以及对应的查询语言和索引机制，使得关系型数据库能够存储和查询半结构化数据。
- 文档数据库：文档数据库是一种面向文档型数据的非关系型数据库，能够存储和查询具有层次结构的半结构化数据，如JSON、XML等。文档数据库提供了灵活的数据模型和查询语言，适用于存储和处理半结构化数据。
- 分布式文件系统和对象存储：对于非结构化数据，可以使用分布式文件系统（如HDFS）或对象存储（如Amazon S3）来存储和管理大规模的非结构化数据。这些系统具有高容量、高可靠性和可扩展性的特点，并且能够提供简单的数据访问接口。

2）谷歌提出的“三件套”技术成果是：Google File System（GFS）、MapReduce和Bigtable。

- Google File System（GFS）：GFS是谷歌开发的分布式文件系统，用于存储和管理大规模的数据。它通过将数据分布在多个计算机上，并提供高度容错性和可扩展性，以实现高吞吐量的数据访问和处理。
- MapReduce：MapReduce是一种分布式计算模型和编程模型，用于大规模数据的并行处理。它将数据处理任务分解为多个Map和Reduce阶段，充分利用分布式计算资源，实现高效的数据处理和分析。
- Bigtable：Bigtable是谷歌开发的分布式的、面向列的非关系型数据库。它基于GFS和MapReduce技术构建，适用于大规模结构化和半结构化数据的存储和查询。Bigtable提供了高度可扩展性、高性能的数据访问能力，广泛应用于谷歌的各种服务。

3）在“三件套”的基础上，谷歌提出了以下代表性的新型“数据库”：

- Spanner：Spanner是谷歌开发的全球分布式数据库系统，具备ACID事务特性和外部一致性。Spanner能够在全球范围内提供一致性的数据访问和复制，支持水平扩展和自动故障恢复。
- BigQuery：BigQuery是谷歌云平台上的一项托管式数据分析服务，用于快速查询和分析大规模数据集。它基于列式存储和分布式计算技术，提供高性能的数据查询和聚合功能。

- Cloud Firestore：Cloud Firestore是谷歌云平台上的文档型数据库服务，用于存储和同步实时数据。它支持半结构化数据的存储和查询，并提供实时数据同步和离线访问功能。