1.**欧洲核子研究组织或欧洲核研究组织收集的大数据的来源是什么**？

答：每秒四千万次，粒子在大型强子对撞机内发生碰撞，每次碰撞产生的粒子往往以复杂的方式衰变成更多的粒子。 LHC周围的精确电子电路记录每个粒子通过检测器作为一系列电子信号的通过，并将数据发送到CERN数据中心（DC）进行记录和数字重建。 数字化的数据摘要记录为“碰撞事件”。 每年生成15个PB左右的数字化摘要数据，由物理学家处理，以确定碰撞是否引发了任何有趣的物理过程。

2. **列出并描述表征大数据的三个主要“V”**

答：

•数量Volume：这显然是大数据最常见的特征。许多因素促成了数据量的指数增长，例如多年来存储的基于交易的数据，不断从社交媒体流入的文本数据，收集的传感器数据量增加，自动生成的RFID和GPS数据等等。

•多样性Variety：今天的数据有各种格式 - 从传统数据库到最终用户和OLAP系统创建的分层数据存储，到文本文档，电子邮件，XML，计量收集，传感器捕获数据，再到视频，音频和股票行情数据。据估计，所有组织的数据中有80％到85％是以某种非结构化或半结构化的形式存在的

•速度Velocity：这指的是数据的生成速度以及必须处理（即捕获，存储和分析）数据的速度以满足需求。 RFID标签，自动传感器，GPS设备和智能电表正在推动越来越需要近乎实时地处理数据流。

3. **列出并描述大数据分析的四个最关键的成功因素**

答：

•**明确的业务需求（与愿景和战略保持一致）**。商业投资应该是为了企业的利益，而不仅仅是为了技术进步。因此，大数据分析的主要驱动力应该是任何级别的业务需求 - 战略，战术和运营。

**•坚定，忠诚的赞助（执行冠军）**。众所周知，如果你没有强大的，坚定的高管赞助，那么很难（如果不是不可能）成功。如果范围是单个或几个分析应用程序，赞助可以在部门级别。但是，如果目标是企业范围的组织转型（大数据计划通常就是这种情况），则赞助需要处于最高级别和组织范围内。

•**业务和IT战略之间的一致性**。确保分析工作始终支持业务战略至关重要，而不是其他方式。分析应该在成功执行业务战略中发挥支持作用。

•**基于事实的决策文化**。在基于事实的决策文化中，数字而不是直觉，直觉或假设驱动决策。还有一种实验文化，看看哪些有效，哪些无效。为了创建一个基于事实的决策文化，高级管理层需要做到以下几点：认识到有些人不能或不会做出调整;是一个声音支持者;强调必须停止过时的方法;要求了解哪些分析可用于决策;将激励和补偿与期望行为联系起来。

•**强大的数据基础架构**。数据仓库为分析提供了数据基础架构。随着新技术的发展，这种基础设施正在改变并在大数据时代得到增强。成功需要将旧的与新的结合起来，以实现协同工作的整体基础设施。

4. **在考虑大数据项目和架构时，列出并描述设计人员应注意的五个挑战，以便使分析能力之旅的压力减轻。**

答：

•数据量：以可接受的速度捕获，存储和处理大量数据的能力，以便决策者在需要时可以获得最新信息。

•数据集成：能够组合结构或来源不相似的数据，并以合理的成本快速完成。

•处理能力：捕获数据时快速处理数据的能力。收集和处理数据的传统方式可能不起作用。在许多情况下，一旦捕获数据需要进行分析以利用最大价值。

•数据治理：能够跟上大数据的安全性，隐私，所有权和质量问题。随着数量，种类（格式和来源）以及数据速度的变化，治理实践的能力也应如此。

•技能可用性：使用新工具利用大数据，并以不同方式进行研究。缺乏具备完成工作技能的数据科学家。

•解决方案成本：由于大数据开辟了一个可能的业务改进世界，因此需要进行大量的实验和发现，以确定重要的模式和转向价值的洞察力。因此，为确保大数据项目的投资回报率具有良好的投资回报率，降低用于发现该价值的解决方案的成本至关重要。

5.**定义MapReduce**

答：正如Dean和Ghemawat（2004）所描述的那样，“MapReduce是一种编程模型，是处理和生成大型数据集的相关实现。以这种函数式编写的程序自动并行化并在大型商用机器集群上执行。这使程序员在没有任何并行和分布式系统的经验的情况下，可以轻松利用大型分布式系统的资源。“

6. **什么是NoSQL用于大数据？ 描述其主要缺点。**

答：

•NoSQL是一种新型数据库，与Hadoop一样，可以处理大量的多结构化数据。 然而，尽管Hadoop擅长支持大规模的批量式历史分析，但NoSQL数据库的目标大部分（尽管有一些重要的例外），用于提供存储在大量多结构数据中的离散数据 最终用户和自动化大数据应用程序。 这种能力严重缺乏关系数据库技术，这种技术根本无法在大数据规模上维持所需的应用程序性能水平。

•当今大多数NoSQL数据库的缺点是它们在性能和可伸缩性方面交换ACID（原子性，一致性，隔离性，持久性）合规性。 许多人还缺乏成熟的管理和监测工具。

7. **什么是数据科学家，工作涉及什么？**

答：数据科学家是经常与大数据或数据科学相关联的角色或工作。 在很短的时间内，它已成为市场上最受欢迎的角色之一。 目前，数据科学家最基本的，当前的技能是编写代码的能力（在最新的大数据语言和平台中）。 一项更持久的技能是数据科学家需要用所有利益相关者都能理解的语言进行交流，并展示讲故事所涉及的特殊技能，包括口头，视觉或理想两者。 数据科学家结合其业务和技术技能，调查大数据，寻找改进当前业务分析实践的方法（从描述性到预测性和规范性），从而改善新业务机会的决策。

8. **为什么今天磁带备份工作负载的某些部分被重定向到Hadoop集群？**

答：

•首先，虽然在磁带上存储数据可能看起来很便宜，但真正的成本却来自于检索的难度。这些数据不仅是离线存储的，还需要数小时甚至数天来恢复，但磁带盒本身也会随着时间的推移而降级，从而使数据丢失成为现实，并迫使公司将这些成本考虑在内。更糟糕的是，磁带格式每隔几年就会发生变化，要求组织执行大规模数据迁移到最新的磁带格式，否则就有可能无法从过时的磁带恢复数据。

•其次，已经证明，保持历史数据在线和可访问是有价值的。与点击流示例一样，将原始数据保留在旋转磁盘上的时间更长，这使得公司可以在上下文更改并需要应用新约束时轻松重新访问数据。使用Hadoop搜索数千个磁盘比旋转数百个磁带快得多，也更容易。此外，随着磁盘密度每18个月持续翻倍，组织在HDFS中持有多年的原始或精炼数据变得经济可行。

9. **流分析和永久分析之间有什么区别？ 你什么时候使用其中一个？**

答：

•在许多情况下，它们同义使用。 但是，在智能系统的背景下，存在差异。 流分析涉及将事务级逻辑应用于实时观察。 适用于这些观察的规则考虑到以前的观察，只要它们发生在规定的窗口中; 这些窗口具有一些任意大小（例如，最后5秒，最后10,000个观察等）。 另一方面，永久分析会根据所有先前的观察结果评估每个进入的观察结果，其中没有窗口大小。 认识到新观察如何与所有先前的观察相关，使得能够发现实时洞察力。

•当事务量很高且决策时间太短时，支持非持久性和小窗口大小，这转化为使用流分析。 但是，当任务至关重要且可以实时管理交易量时，永久分析就是更好的答案。

10. **描述数据流挖掘及其使用方法**

答：数据流挖掘作为流分析的支持技术，是从连续，快速的数据记录中提取新模式和知识结构的过程。 数据流是有序的实例序列的连续流，其在数据流挖掘的许多应用中可以使用有限的计算和存储能力仅读取/处理一次或少量。 数据流的示例包括传感器数据，计算机网络流量，电话对话，ATM交易，网络搜索和财务数据。 数据流挖掘可以被认为是数据挖掘，机器学习和知识发现的子领域。 在许多数据流挖掘应用程序中，目标是在给定关于类成员资格或数据流中先前实例的值的一些知识的情况下预测数据流中的新实例的类或值。