Assignment Guidance and Front Sheet

This front sheet for assignments is designed to contain the brief, the submission instructions, and the actual student submission for any WMG assignment. As a result the sheet is completed by several people over time, and is therefore split up into sections explaining who completes what information and when. Yellow highlighted text indicates examples or further explanation of what is requested, and the highlight and instructions should be removed as you populate ‘your’ section.

This sheet is only to be used for components of assessment worth more than 3 CATS (e.g. for a 15 credit module, weighted more than 20%; or for a 10 credit module, weighted more than 30%).

**To be completed by the student(s) prior to final submission:**

Your actual submission should be written at the end of this cover sheet file, or attached with the cover sheet at the front if drafted in a separate file, program or application.

|  |  |
| --- | --- |
| **Student ID or IDs for group work** | **e.g. 1234567** |

**To be completed (highlighted parts only) by the programme administration after approval and prior to issuing of the assessment; to be consulted by the student(s) so that you know how and when to submit:**

|  |  |
| --- | --- |
| **Date set** | 12/01/2022 |
| **Submission date (excluding extensions)** | Before 12 noon UK Time on 14/02/2022 |
| **Submission guidance** | Submit electronically via Tabula |
| **Marks return date (excluding extensions)** | 14/03/2022 |
| **Late submission policy** | If work is submitted late, penalties will be applied at the rate of **5 marks per University working day** after the due date, up to a **maximum of 10 working days** late. After this period the mark for the work will be reduced to 0 (which is the maximum penalty). “Late” means **after the submission deadline time as well as the date** – work submitted after the given time even on the same day is counted as 1 day late.  For **Postgraduate** students only, who started their **current course before 1 August 2019**, the daily penalty is **3 marks** rather than 5. |
| **Resubmission policy** | If you fail this assignment or module, please be aware that the University allows students to remedy such failure (within certain limits). Decisions to authorise such resubmissions are made by Exam Boards. Normally these will be issued at specific times of the year, depending on your programme of study. More information can be found from your programme office if you are concerned. |

**To be completed by the module owner/tutor prior to approval and issuing of the assessment; to be consulted by the student(s) so that you understand the assignment brief, its context within the module, and any specific criteria and advice from the tutor:**

|  |  |
| --- | --- |
| **Module title & code** | WM9B1-15 Big Data Technology & Visualisation |
| **Module owner** | Michael Mortenson |
| **Module tutor** | Michael Mortenson |
| **Assessment type** | Essay |
| **Weighting of mark** | 80% |

|  |
| --- |
| **Assessment brief** |
| You have been appointed as a big data architecture consultant for a start-up in fashion e-commerce using artificial intelligence. The company is initially looking for you provide advice on how they should acquire, store, process and visualise their data – including:  Product information and reviews (text and image data)  Sales and supply chain information (transactional data)  Social media data (mixed)  Other business and operational data they may generate/use.  Your answer should provide justified recommendations for a range of technologies they may use, including:  Databases and/or data stores  Data pipeline and orchestration tools  Data visualisation solutions  Diagrams and descriptions of the end-to-end solution  Consideration of any major risks and identification of appropriate mitigation strategies.  You are expected to use an appropriate report style and support your arguments and recommendations with a mixture of academic and commercial sources. You are free to make any assumptions you need to about the business scenario in order to answer the question.  你已被任命为一家AI初创公司（时尚电子商务）的大数据架构顾问。该公司最初希望你能就他们应该如何获取、存储、处理和可视化他们的数据提供建议--包括  你的答案应该为他们可能使用的一系列技术提供合理的建议，包括。要有理由和证据！  - 数据库和/或数据存储  - 数据管道和协调工具  - 数据可视化解决方案  - 端到端解决方案的图表和描述  - 考虑任何主要的风险并确定适当的缓解策略：新项目在落地实行的时候可能会有风险，多看文献，如何减轻负担  你应该使用适当的报告风格，用学术和商业来源的混合物来支持你的论点和建议。为了回答这个问题，你可以自由地对商业场景做出任何假设。  Prior to working on its O2O commerce strategy, [Beauty Heroes](https://www.shopify.com/plus/customers/beauty-heroes)’ ecommerce sales were soaring. So much so, it decided to crack the offline market and open its first brick-and-mortar store in Novato, California.  在开展O2O商务战略之前，美丽英雄的电商销售额一直在飙升。因此，它决定打开线下市场，并在加州的诺瓦托开设第一家实体店。  Beauty Heroes used Shopify POS to aggregate sales data across both online and offline channels. It also encouraged existing customers to shop in-store by delivering targeted promotions based on purchase history.  Beauty Heroes使用Shopify POS来整合线上和线下渠道的销售数据。它还根据顾客的购买历史，提供有针对性的促销活动，鼓励现有顾客在店内购物。  开设实体店，让客户亲眼看到、触摸并与他们的产品进行互动。  Smart new retail  智慧新零售是指基于互联网的新零售模式，利用大数据、人工智能等先进技术，对生产、物流、销售过程进行升级改造，重构产业结构和生态系统，深入集线上服务、线下体验、现代物流于一体。  O2O运营模式下的智慧新零售注重客户体验的创新、大数据的采集和技术的应用。不仅关注消费便利，更关注消费心理、消费习惯，更加重视基于个体差异的精准营销。  [80% 的购物者](https://www.thinkwithgoogle.com/consumer-insights/consumer-trends/in-store-shopping-statistics/)会在有他们需要或想要的商品时立即去商店购买。    According to [research shared by Google](https://www.thinkwithgoogle.com/consumer-insights/online-to-offline-marketing/) in late 2018, 61% of shoppers would rather shop at a store that had a physical location than one that is online only.  根据谷歌在2018年底分享的研究，61%的购物者宁愿在有实体店的商店购物，而不是只在网上购物。  研究表明，[61% 的购物者](https://www.thinkwithgoogle.com/consumer-insights/consumer-trends/shopping-preference-statistics/" \t "_blank)更愿意购买也有实体店的品牌（而不是那些仅限在线的品牌）。  have a physical storefront when you could have a warehouse and online store  当你可以拥有一个仓库和一个在线商店的时候，却要拥有一个实体店面  线上提高产品和服务的知名度，然后访问商店进行购买  商业智能是电子商务和人工智能的结合，使得电子商务的整个过程都依赖于智能终端机器或操作系统来实现最大的收益。商业智能的出现满足了市场对高效交易、低成本交易和信息交易的需求。  公式 4 给出了 AI 模型的表示。在电子商务平台的应用实践中，如何根据用户的意图、兴趣和特点，对现有的客户信息、库存信息以及大量的数据信息进行智能排列、调整和匹配，从而获得用户满意的检索输出，将成为电子商务未来应用的技术难题。    AI对商业创新的影响：  客户细分、客户关系、价值定位、渠道、关键资源、关键活动、关键合作伙伴关系、收入流和成本结构  2019年以来，新模式、新形式层出不穷 电子商务已经出现，新的合适的术语和 随着研究的成熟，研究方向也逐渐浮出水面， 如直播电子商务、社交电子商务、在线虚拟 技术等，符合当前多元化， 个性化、体验化的消费需求。  分析公司需求是什么，架构整体展示，然后引出后面detail  AI ：feature value，deep learning  定义小服装公司，成本不能太高，以后可扩展性，不能影响其他部门  服装主要是图片很多，类储存在spark里？！文档类在hadoop里  - 1. - 商品：产品信息和评论（文本和图像数据）  用户基本信息：注册，登录  用户行为：- 销售和供应链信息（交易数据）网页点击浏览量等等  - 社交媒体数据（混合）。  - 他们可能产生/使用的其他业务和运营数据。  比如：email，hr system  Limitation：flink  多个数据库  不要比较优缺点，强调决策过程    **图形用户界面, 应用程序  描述已自动生成**  **Extract：**  **Load：**  **Stream：实时流**  **Batch：数据在特定的时间段以块的形式发送：每小时、每天**  **Data warehouse：hadoop，spark**  存储、管理和分析分布式数据，但在拉取数据时不涉及任何查询   架构设计 本文的架构基于 Spark Streaming 流数据处理框架。本文结合了flume、Kafka、HBase、mysql、echarts等技术。每个组件的具体版本如表 2所示。本文分析了电子商务网站的用户行为。    下面用的这种大数据平台架构也叫 Lambda 架构，是构建大数据平台的一种常规架构原型方案。  优点：Lambda 架构使开发人员能够构建大规模分布式数据处理系统。它具有很好的灵活性和可扩展性，也对硬件故障和人为失误有很好的容错性。 分析维度和指标 对于收集到的数据，分以下两种情况进行分析：   * 行为分析：本文针对某个时间点，分析了电子商务网站四种常见行为（点击、收藏、添加购买和支付）的分布情况。从而帮助企业分析用户当前的行为。 * 时间段分析：本文分析用户行为数据在不同时间段的变化，帮助企业掌握用户在不同时间段的行为趋势，并做出合理的预测。涉及数据的索引说明见表1。   **第三节。** 模块分析 模型设计的总体框架分为日志采集与预处理模块、实时数据分析模块和结果展示模块。如图1所示。系统将日志采集和预处理模块分为采集和预处理两个子模块。本系统将实时分析模块分为数据接入、流量计算和数据输出三个子模块。 日志采集和预处理模块（数据采集） 将应用程序产生的数据和日志等同步到大数据系统中，由于数据源不同，这里的数据同步系统实际上是多个相关系统的组合。数据库同步通常用 Sqoop，日志同步可以选择 Flume，打点采集的数据经过格式化转换后通过 Kafka 等消息队列进行传递。  不同的数据源产生的数据质量可能差别很大，数据库中的数据也许可以直接导入大数据系统就可以使用了，而日志和爬虫产生的数据就需要进行大量的清洗、转化处理才能有效使用。  Flume 是一个分布式日志收集系统，用于收集和清理日志。Flume 具有高可用、高可靠的特点，支持在日志系统中定制不同类型的数据发送者。在该系统中，Hume 用于清洗数据，然后将其传输到下一个数据接收器。    为了协调工作，在日志收集过程中引入了zookeeper。Zookeeper 保证了flume中配置的一致性和高可用。当配置数据发生变化时，zookeeper 负责与 Flume 主节点通​​信，并告诉它使用 gossip 协议来同步数据[4]。Agent结构是flume的核心。代理结构包括：source、channel、sink。具体结构如图5所示。  在图6中，源接收来自服务器的数据并将数据以事件格式传递给通道。Channel是一个完整的事务，是source和sink之间的临时存储容器。Channel会缓存传输过程中的数据，等待后续sink的消费。此外，通道保证了发送和接收过程中的数据一致性。最后，sink 会将数据存储在类似于 HDFS [5]的文件系统中。  Kafka 是一个分布式发布订阅消息队列。它是由 Apache 基金会开发的。它是用 Scala 和 Java 编写的。Kafka 具有高吞吐量。在大数据实时处理领域，Kafka作为开源流处理平台，可以很好地缓冲数据流。Kafka由生产者发布，由消费者消费。此外，它还可以根据不同的类别对发布的消息进行分类。在本系统中，使用Kafka作为消息缓冲队列，避免了后续处理逻辑因数据爆炸而停滞。系统首先将消息放入队列，并延迟一定时间处理。  日志采集模块采用JS埋点技术获取用户行为日志信息。该模块在用户点击、收集、添加到购物车和支付时埋点。此外，该模块使用反向代理技术将日志发送到日志收集服务器[2]。Flume 收到这些日志后，会清理不完整或丢失的数据，然后将清理后的数据发送给 Kafka。该模块的处理结果为火花流的实时分析提供了有效的数据。图2显示了模块的具体流程。   数据分析模块 这部分是大数据存储与计算的核心，数据同步系统导入的数据存储在 HDFS。MapReduce、Hive、Spark 等计算任务读取 HDFS 上的数据进行计算，再将计算结果写入 HDFS。   1. Batch离线计算   MapReduce、Hive、Spark 等进行的计算处理被称作是离线计算，HDFS 存储的数据被称为离线数据。在大数据系统上进行的离线计算通常针对（某一方面的）全体数据，比如针对历史上所有订单进行商品的关联性挖掘，这时候数据规模非常大，需要较长的运行时间，这类计算就是离线计算。   1. Stream实时分析   除了离线计算，还有一些场景，数据规模也比较大，但是要求处理的时间却比较短。比如淘宝要统计每秒产生的订单数，以便进行监控和宣传。这种场景被称为大数据流式计算，通常用 Storm、Spark Steaming 等流式大数据引擎来完成，可以在秒级甚至毫秒级时间内完成计算。  数据实时分析模块分为数据接入、流量计算和数据输出三个子模块。具体流程如图3所示。   * 数据访问：由于不同系统之间数据生产/消耗率的差异，本文在实时数据采集预处理模块和数据实时分析模块之间增加了Kafka作为缓冲区。数据经过flume采集和预处理后，会发送到Kafka。然后将数据输出到 spark 流进行统计分析[3]。 * 流计算：本系统采用流数据实时计算框架Spark Streaming（ETL组件），对采集到的用户行为数据进行实时统计分析   Spark Streaming 是 Spark 中的一个实时流数据处理框架。它是火花的延伸。它具有低延迟、高吞吐量、可扩展性和容错机制的特点。Spark Streaming 的工作原理与 Spark 类似。spark streaming的处理逻辑其实就是对接收到的数据流及时进行分段，然后对分段后的文件进行批处理。与 Strom 相比，Spark Streaming 有一定的延迟。Spark Streaming将输入的实时流数据按照一定的时间间隔进行拆分，然后传递给Spark Engine输出一系列batch数据。工作原理如图6 [6]所示。     * 数据输出：分析完成后，系统会将输出数据通过HBase进行持久化。     HBase 是一个面向列的分布式开源数据库。它是稀疏的、多维的和有序的。不同于一般的关系型数据库，HBase 更类似于 NoSQL。HBase 采用基于列的方式存储，支持高速读写。它适用于非结构化和半结构化数据。它可以按行键、列族、列限定符和时间戳来索引。HBase可以在廉价的PC上搭建大型存储集群，降低企业成本。它为用户提供了高可靠、高性能、列存储、可扩展、实时读写的数据库系统。  横向来看，HBase 中的数据由一个行键和一个或多个列族组成。HBase支持列族的动态扩展，由于HBase中存储的值没有数据类型，用户在添加列族或单列时无需提前定义列的数量和类型。HBase 中的所有列都以字符串的形式存储，数据类型需要用户自行转换[7]。 数据输出与展示  1. 数据输出   大数据计算产生的数据还是写入到 HDFS 中，但应用程序不可能到 HDFS 中读取数据，所以必须要将 HDFS 中的数据导出到数据库中。数据同步导出相对比较容易，计算产生的数据都比较规范，稍作处理就可以用 Sqoop 之类的系统导出到数据库。   1. 数据展示   这时，应用程序就可以直接访问数据库中的数据，实时展示给用户，比如展示给用户关联推荐的商品。  除了给用户访问提供数据，大数据还需要给运营和决策层提供各种统计报告，这些数据也写入数据库，被相应的后台系统访问。很多运营和管理人员，每天一上班，就是登录后台数据系统，查看前一天的数据报表，看业务是否正常。如果数据正常甚至上升，就可以稍微轻松一点；如果数据下跌，焦躁而忙碌的一天马上就要开始了。  结果演示模块基于SSM（Spring + spring MVC + mybatis）框架。该模块结合了 MySQL 和 echarts 来展示数据可视化。SSM框架分为三层：  首先是道层。主要封装了大部分与数据库通信的任务；第二层是服务层。用于实现业务逻辑；第三层是控制器层。主要负责与用户通信、传输数据、进行业务调度。最后，模块使用echarts，从用户行为分布和时间段分布的角度，通过饼图和折线图的方式，直观的展示分析结果。  **任务调度管理系统**  将上面三个部分整合起来的是任务调度管理系统，不同的数据何时开始同步，各种 MapReduce、Spark 任务如何合理调度才能使资源利用最合理、等待的时间又不至于太久，同时临时的重要任务还能够尽快执行，这些都需要任务调度管理系统来完成。 |

|  |  |
| --- | --- |
| **Word count** | 4,000 words |
| **Module learning outcomes (numbered)** | Demonstrate an comprehensive understanding of the key differences between Big Data technologies and analysis methods and traditional approaches.  Evaluate real-world scenarios and determine appropriate database solutions (traditional and NoSQL)  Demonstrate a comprehensive understanding of cloud data architectures, the operational risks associated with them, and develop appropriate mitigation strategies  Demonstrate a comprehensive understanding of the core concepts of visual communication and data visualisation.  Practically implement data pipelines and processing in a cloud setting |
| **Learning outcomes assessed in this assessment (numbered)** | All |
| **Marking guidelines** | *20% of marks. Analysis of the business context and identification of the important requirements (functional and non-functional). L2, L3*  **Above expectations**  A detailed explanation of the organisational context with clear requirements, constraints and objectives identified. These elements are re-enforced throughout the report.  **Expectation**  A realistic company description is provided, with sensible particulars that have shaped some of the later decisions. Some elements may be more generic and/or could have been more influential on the recommendations provided.  **Below expectation**  The company description is superficial or overly contrived. Only generic requirements are given and/or requirements are inconsistent with later decisions.  *35% of marks. Development of a data architecture and flow. L1, L2, L3, L4*  **Above expectations**  A realistic and professional design that accounts for business requirements and relevant best practices.  **Expectation**  The proposed solution generally meets the requirements and objectives and is largely consistent with best practices.  **Below expectation**  There are inefficiencies and inconsistencies in the suggested solution that suggest a lack of awareness on current best practices. Some elements may be inconsistent with the organisational description given.  *35% of marks. Selection of an appropriate technology stack. L1, L2, L3, L5*  **Above expectations**  A sensible and comprehensive selection of technologies that are current and relevant to the organisational brief. Decisions have been thoroughly justified throughout.  **Expectation**  Technology choices show some awareness of best practices, organisational requirements and interoperability. Some justification is given for the decisions that have been.  **Below expectation**  Technology choices are arbitrary and poorly justified. Not enough consideration is given to the appropriateness of the choices to the business problems, organisational requirements and/or best practices.  *10% of marks. Overall quality of the report and awareness of the overall risks. L1, L3*  **Above expectations**  Report is well written and professionally presented. Good awareness shown of the limitations and risks associated with the selected technology stack, the complexities of modern data architecture and suitable risk mitigation strategies are suggested.  **Expectation**  Report is reasonably realistic in its presentation and structure. Some consideration is shown of the various risks present, and how they may be mitigated.  **Below expectation**  Report is poorly presented/structured/conceived. Significant risks/limitations are not identified and/or appropriate mitigation strategies have not been identified. |
| **Academic guidance resources** | *Contact the module tutor for guidance* |

**The following is pre-populated for PGT assignments only:**

|  |
| --- |
| **Writing your Post-Module Assignment (PMA): specific additional advice for WMG’s Postgraduate Taught Students** |
| As a postgraduate level student in WMG you may have some concerns about your ability to write at the high standard required. This short guide is intended to provide general guidance and advice. It is important that if you have any questions you discuss them with your module tutor. Remember, in writing your PMA you need to meet the expectations of the reader and university. |
| **A good PMA generally requires you to answer the question and to include**…   1. A title, with your student number, module, lecturer’s name and any other documentation required by the university. 2. A contents page and if appropriate, an abstract. 3. An introduction which acts as a ‘map’ to the rest of the document, describing the aim or purpose of the work and explaining how this aim is achieved. At this point it is usually helpful to paraphrase your conclusion. 4. Evidence of an appropriate level of background reading of relevant texts. 5. Evidence of systematic and clear thinking, indicative of good planning and organisation. 6. Writing which makes sense, is clearly and carefully presented (proof-read and grammar checked). 7. A critical style of writing which compares and contrasts the main theories, concepts and arguments with conclusions that are based in evidence presented. 8. High levels of accurate academic referencing. 9. A logical and well-defined structure with headings and subheadings. 10. Clearly labelled and well-presented diagrams and other graphics that are discussed in the text. 11. Adherence to usual academic standards including length and a timely submission. 12. A reference section in which every source that is cited in the text is listed. |
| **Where to get help:**   1. **Talk to your module tutor if you don’t understand the question or are unsure as to exactly what is required.** 2. Study, Professional and Analytical Skills (SPA) Moodle site – we have a lot of resources on this website with workbooks, links and other helpful tools. <https://moodle.warwick.ac.uk/> 3. The university Academic Writing centre provides workshops and useful tools to help you in all aspects of your work. <https://warwick.ac.uk/services/skills/academicwriting/> 4. Avoiding Plagiarism, the university’s site to help you to reference properly <https://moodle.warwick.ac.uk/course/view.php?id=42224> 5. Wellbeing support services <https://warwick.ac.uk/services/wss> 6. Numerous online courses provided by the University library to help in academic referencing, writing, avoiding plagiarism and a number of other useful resources. <https://warwick.ac.uk/services/library/students/your-library-online/> |

图示

描述已自动生成