**2018年全国职业院校技能大赛**

**赛项申报书**

赛项名称：大数据技术与应用

赛项类别：常规赛项■ 行业特色赛项□

赛项组别：中职组□ 高职组■

涉及的专业大类/类：电子信息大类

方案设计专家组组长：

手机号码：

方案申报单位（盖章）：全国工业与信息化职业教育教学指导委员会

方案申报负责人：

方案申报单位联络人：

联络人手机号码：

电子邮箱：

通讯地址：

邮政编码：

申报日期：2017年8月25日

**2018年全国职业院校技能大赛**

**竞赛项目方案**

**一、赛项名称**

（一）赛项名称

大数据技术与应用

（二）压题彩照

（三）赛项归属产业类型

电子信息产业、战略性新兴产业

（四）赛项归属专业大类

电子信息大类-计算机类-大数据技术与应用（610215）

电子信息大类-计算机类-计算机应用技术（610201）

**二、赛项申报专家组**

本赛项设计团队由全国大数据领域学科带头人、工业和信息化职业教育教学指导委员会牵头负责，由国内高职院校、高校、科研机构以及国内一流大数据产品及应用相关企业共同组成，阵容强大，通过大数据领域专家、教育专家、和知名企业的共同努力，保障本赛项的科学性、前瞻性以及行业实践性。

**三、赛项目的**

本赛项紧贴行业应用和行业需要，模拟大数据项目实施前沿技术和实际生产场景，旨在考察参赛选手在企业真实项目环境下大数据平台搭建、数据采集与清洗、数据建模与预测、数据分析与展现等核心知识、技能，以及团队协作、沟通力、抗压力、职业规范等职场素质，展现职业院校大数据应用专业学生技能与风采，激发学生的求知欲和参与教学活动的热情，以达到“以赛促学”的目的；整合行业、企业和院校资源，搭建校企合作平台，深化产教融合，促进职业院校准确把握大数据产业发展及岗位需求，调整课程结构，更新课程内容，引领大数据技术与应用等相关专业改革与发展，有效开展实践促进产教深度融合的人才培养模式改革，切实培养企业需要的高素质大数据技能型人才，以达到“以赛促改”的目的；培养一批“实践能力强、教学水平高、敬业精神佳”的大数据双师型“种子教师”师资队伍，建设一批高质量、立体化、一体化的专业、课程、项目教学资源，以达到“以赛促教”的目的。

**四、赛项设计原则**

**（一）遵守大赛制度，确保赛项公开、公平、公正。**

确保公开、公平、公正是赛项设计的第一准则。本赛项严格遵循《全国职业院校技能大赛制度汇编》要求，严格遵守大赛制度文件，规范赛项设计、组织、实施各环节流程，确保赛项公开、公平、公正。

**（二）赛项关联职业人才需求量大，并服务于国家重点战略。**

伴随着数据爆炸式增长，大数据时代正向我们逐步展开一场意义深远的数据革命，对全球整合经济时代的业务和服务产生深远的影响。国务院印发的《促进大数据发展行动纲要》和《“十三五”国家战略性新兴产业发展规划》中，认定数据是国家基础性战略资源，大数据定位为国家战略性新兴产业，大数据应用赛项服务大数据人才培养，助力国家抢占大数据产业先机，成为数据强国，推动经济转型。大数据浪潮中，各大企业对大数据人才的需求也越来越紧迫，不论是拓展大数据业务的百度、阿里、奇虎360等互联网公司，还是专业提供数据服务的大数据服务商，以及拥有大量数据的传统行业，都对大数据人才有着迫切的需求。据专业机构的调查预测，中国未来3-5年内的大数据人才需求总量将达到180万人，目前的人才缺口现状至少有150万人。

我国行业大数据总体发展水平受基础环境、数据汇集、行业应用等因素影响，部分行业大数据整体发展情况较好，这些行业也是学生的主要就业方向。目前来看，高职学生就业主要面向：政务、金融、交通、电信、商贸、医疗、教育、旅游、工业、农业等几大行业。从事云计算和大数据中心、信息中心运维管理，大数据研发部门的工作，也能胜任企事业单位系统集成、业务数据分析工程师、数据挖掘工程师和部分开发工程师岗位需求。

本赛项方案设计充分考虑了高职学生大数据就业的10大行业对应的5大就业技术方向。即大数据分析架构类、大数据算法挖掘类、大数据平台运维类、大数据数据管理类、大数据开发类。因此，本赛项在设计中面向就业，充分考核参赛考生5类专业技能掌握情况。具体岗位对应关系如下图：

职业院校开设的数据库管理与开发、软件开发、计算机应用技术、计算机信息管理、物流信息技术、信息安全技术、智能产品开发等专业都能通过本赛项促进教学。

**（三）竞赛内容对应相关职业岗位或岗位群、体现专业核心能力与核心知识、涵盖丰富的专业知识与专业技能点。赛题设计没有理论测试环节，并保证于开赛1个月前在大赛网络信息发布平台上（www.chinaskills-jsw.org)公开全部赛题。**

大数据产业链条分为大数据平台搭建、数据采集、数据清洗、数据分析、数据可视化、数据挖掘等六大环节，其中数据挖掘又包含数据建模和数据预测环节。根据《中国大数据产业发展评估报告》当前我国大数据企业业务范围不断拓展，几乎覆盖了产业链的各个环节，其中以从事**大数据分析挖掘业务**的企业最为集中，所占比例高达63.7%；从事**数据采集业务**的企业占比为37.4%；从事IDC、数据中心租赁等数据**大数据平台搭建与存储业务**的企业比重最低，仅为8.5%；从事数据分类、清洗加工、脱敏等**数据清洗与分析业务**的企业占比为27.8%；从事**数据可视化**相关业务的企业占比14.3%。

针对以上情况，本赛项竞赛选题取材于大数据行业的真实应用需求及案例，针对行业人才需求情况，经过职业教育专家与用人单位的充分论证和浓缩提炼，提取其中核心能力与核心知识，建立了高职学生大数据人才能力图谱。

该图谱涵盖丰富的相关岗位对应的专业知识与专业技能点，确保竞赛工作任务考察重点与真实大数据职业技能相吻合，突出考查选手基础知识水平、大数据平台基础、数据可视化、编程语言、数据挖掘五个方面的核心技能。**以上5大考核点既包括大数据技术与应用初始就业岗位需求，同时也涵盖大数据技术与应用发展岗位需求**，确保竞赛项目涵盖的专业知识与专业技能点可直接实施于大数据高职教育的实训课程中，体现竞赛项目产教融合的设计原则。

本赛项虽然注重考查选手基础知识水平，但却不通过简单的理论考试去考核。赛项设计通过紧贴行业真实案例的技能考核，能更准确了解学生对理论知识真实掌握并融会贯通的情况。所以本赛项没有直接考核理论知识。

整个赛题设计与日常工程项目交付流程高度一致，能够让学生充分实践，并按照项目交付标准完成比赛各个步骤，提升学生上岗实习、就业的竞争力。

通过本赛项的训练和比赛，参赛选手未来可以胜任：**大数据平台的运维工程师、大数据数据分析工程师、大数据应用开发工程师、大数据采集工程师、大数据可视化工程师、大数据业务售后支持、大数据业务售前支持以及大数据产品的销售等相关岗位。**通过大赛的牵引，在竞赛中成绩优异的学生还能胜任诸如**算法研究工程师、大数据建模工程师**等更有挑战的岗位**。**

**（四）采用工业级技术平台，竞赛平台先进、成熟、稳定、社会保有量高。**

根据行业特点，本赛项选择了业内先进、通用性强、社会保有量高的设备与软件，竞赛平台所涉及产品成熟度高，社会应用广泛，拥有**国检证书、工信部、泰尔实验室等多机构检测报告，拥有多项国家专利**。该平台的选择，**有助于让学生在竞赛过程中接触到行业优秀的主流工业产品**，贴近实战，提升就业竞争力。

竞赛平台的设计也充分考虑到行业发展情况及大数据应用案例教学转换，社会应用成熟，该竞赛平台已经在诸如公安部、广东省信息中心、湖南省邮储银行、西安电子科技大学、河南省人民医院、永州市智慧交通、四川省电力局等多个行业核心部门及骨干单位**实现了商用，业界反响好，技术成熟稳定**。并且该赛项相关技术平台该平台使用专业的ETL引擎支撑底层数据抽取及清洗、海量异构数据的存储与计算；使用创新应用开发引擎加速上层大数据应用构建，以应用为核心，既是行业主流需求，同时也更符合职业教育培养学生应用能力的特点。既满足大数据未来发展的趋势需要，同时也能直接应用于各参赛院校后续的日常教学要求，将比赛设备用于日常教学过程，切实提高比赛设备的利用率，培养更多的学生，同时减少校方建设投资。

**（五）夯实资源转化，引领专业建设与发展**

大数据技术与应用专业作为新设专业，专业建设基础薄弱，急需通过大赛引领专业建设与发展。在竞赛同时，将在行业发展与岗位需求分析、专业教学标准建设、课程标准建设、师资队伍建设、实训基地建设、课程与项目资源建设方面开展资源转化工作，真正发挥大赛对专业建设与发展的引领作用。

**五、赛项方案的特色与创新点**

**(一) 竞赛专家队伍强大，竞赛平台及内容选择科学合理并具有前瞻性。**

赛项将邀请北京大学、北京大数据研究院、新华三、阿里巴巴、中科曙光、京东、美团、BOSS直聘等大数据研究机构和行业专家参与赛项的设计、论证，保证竞赛内容的科学合理。

赛项选择的技术平台是行业真实应用的平台，成熟可靠。该技术平台使用专业平台搭建，并非简单的开源平台，专业优势明显，行业应用广，贴合行业主流需求。该技术平台和赛项考核内容更加注重学生的应用能力考核，而不是针对开源平台的开发和优化，以及简单的数据分析操作等。

竞赛内容的选择上，选取了典型、热门的场景。根据社会上对大数据人才需求，以及高职教育教学特点，有针对性地进行赛题设计，竞赛内容覆盖大数据典型工作任务。

**(二) 赛项采用工程项目实施型命题的模式，无限贴近行业实际应用。**

赛项采用紧贴行业的大数据应用方案为素材，学生按照相关应用场景进行大数据项目实施的考核方式。赛项将该项目的几个关键环节“大数据平台部署——数据采集——数据清洗——数据分析——数据可视化——数据挖据及开发”作为考察重点，深度契合高职学生大数据人才能力培养方向。并且，该项目的几个环节链条清晰，基本覆盖常见的大数据应用环节，知识点体现该产业及专业核心能力与核心知识、涵盖丰富的专业知识与专业技能点。能够满足未来3~5年的国赛及专业教学市场就业需求。

例如样题的舆情分析场景，便是政府和互联网行业主流应用场景。该场景的实施过程在行业真实案例中，人才的需求量很大，一个项目有时甚至需要上百人协同工作，就业前景和空间大。

**(三) 竞赛过程安排和结果评判体现公平、公正原则。**

整个赛项过程进行项目式管理支撑，包含选手比赛过程监控、进度展现、结果预判，形成大数据竞赛中观众区、裁判区、竞赛区信息实时发布。按照严格执行大赛执委会的指导规范要求，由专家组对比赛命题和评判规则进行制定和评审。评分以结果评分（95%）为主，在考察学生专业技能的同时，也考察学生的专业素养。评分员互不干涉，充分体现了竞赛的公平、公正、公开的原则。

 **(四) 开展大赛多行业深度合作，提升社会对职业教育关注度，提升技能大赛国内和国际影响力。**

在大赛举办过程中，举办多种形式的同期活动。

举办**大数据人才供需对接会**。邀请京东、美团、BOSS直聘、新华三、联想、中科曙光、东软等行业内从事大数据相关业务的企业参加大数据人才供需对接会，从而使职业院校更加了解企业用人需求，使企业更加了解职业院校人才培养特色，加强校企双方交流，提升企业对职业院校关注度。

举办**大数据应用型人才培养国际论坛**。与北京大数据研究院、参赛企业和赛项承办地政府共同举办大数据应用型人才培养国际论坛，邀请亚、欧、大洋洲等相关国家的专家以及联合国教科文组织高等教育创新中心专家参加本次论坛，借此提升大赛国际影响力。

举办**大数据技术与应用国际邀请赛**。加大宣传力度，计划吸引港澳台及国际队伍参赛，进一步提升全国职业院校技能大赛的国际化水平和国际影响力。

 **(五)大力度夯实资源转化工作，引领专业建设和教学改革。**

大数据技术与应用专业作为新设专业，专业建设基础薄弱，急需通过大赛引领专业建设与发展。在竞赛同时，将在行业发展与岗位需求分析、专业教学标准建设、课程标准建设、师资队伍建设、实训基地建设、课程与项目资源建设方面开展资源转化工作，真正发挥大赛对专业建设与发展的引领作用。

另外，本赛项会通过推广校企合作专业共建等多种形式，为各院校提供专业建设支持，搭建院校与企业间的合作桥梁。

**六、竞赛内容简介（须附英文对照简介）**

大数据技术与应用赛项应用企业真实项目，紧密结合行业实际应用场景，主要考核参赛选手大数据平台搭建、数据采集与清洗、数据建模与预测、数据分析与展现等方面的核心技能，以及现场问题的分析、处理、团队组织及协作等多方面的专业素质,同时考察学生的快速学习和应用能力。参赛选手利用数据分析方法以及大数据创新应用，对原型系统分析数据价值、针对问题提出合理的解决思路。主要内容包括：

1. 大数据平台搭建包括：Hadoop大数据平台的安装、配置与使用；HDFS常用操作命令；HBase基本操作；
2. 数据采集与清洗包括：Python编程语言的使用，数据预处理（基于MapReduce和Spark），数据标记；
3. 数据分析与展现包括：BI工具的使用，统计图表展现，计算分析结果导出；
4. 数据建模与预测包括：通过对已有数据建模，从而对数据建立预测；
5. 数据综合分析包括：分析结果总结和预测。

竞赛指导学生由浅入深，学习如何搭建大数据平台、进行网络数据的采集、数据清洗、完成数据分析并展现，以及大数据应用系统的开发。涉及大数据平台、数据采集与清洗、大数据存储、数据分析与展现、数据挖据与开发等多项知识综合应用。

The Big Data Technology and Application competition, which applies the real cases in company and closely unifies the industry actual application scene, tests the core skills of big data platform construction, data collection and cleaning, data modeling and forecasting, data analysis and presentation, as well as the analysis and handling of on-site problems, team working and collaboration. It also examines the students ' quick learning and application ability. The competitors use the data analysis method and the Big Data Innovation application to proposes the reasonable solution and the prototype system analysis data value.

1. Big data platform construction includes: Hadoop big data platform installation, configuration and use; HDFS common operation command; HBase basic operation.
2. Data collection and cleaning includes: the use of Python programming language, data preprocessing (based on MapReduce and Spark), and data mark;
3. Data analysis and presentation includes: the use of BI tools, presentation of statistical chart, calculation and analysis results derived;
4. Data modeling and forecasting includes: establish a forecast for the data through modeling the existing data;
5. Data synthesis analysis includes: summary and forecast of analysis results.

The competition instructs the students to make great progress in learning how to build big data platform, collect network data, clean data, complete the analysis of data and show them. This process involves comprehensive application of multiple knowledge, such as big data platform, data collection and cleaning, big data storage, data analysis and presentation, data mining and development.

**七、竞赛方式（含组队要求、是否邀请境外代表队参赛）**

（一）竞赛以团队方式进行，每支参赛队由3名选手组成（可设场上队长1名），性别和年级不限，参赛选手为2018年在籍的高职学生，性别不限。不得跨校组队。

（二）本赛项邀请港澳台及国外队伍参加国际邀请赛，欢迎社会各界人士到赛场观摩。

**八、竞赛时间安排与流程**

（一）时间安排

时间安排：4个小时

|  |
| --- |
| **日程安排** |
| 第一天 | 14:00前 | 参赛队报到 |
| 15:00—15:30 | 参赛队参观赛场，熟悉比赛环境 |
| 15:30—16:00 | 领队会 |
| 16:30—17:30 | 检查比赛环境 |
| 17:30 | 赛场封闭 |
| 第二天 | 7:30 | 评分裁判封闭 |
| 7:00—7:30 | 开启赛场，检查比赛环境 |
| 7:30—8:00 | 参赛队检录、一次加密、二次加密 |
| 8:00—12:00 | 比赛 |
| 12:00—14:00 | 申诉仲裁受理 |
| 13:00—14:00 | 三次加密 |
| 13:00—14:00 | 评分裁判培训 |
| 14:00—20:00 | 评分 |
| 第三天 | 8:30—9:30 | 闭赛式 |

（二）竞赛流程

入场检录

一次加密

二次加密

正常进行

完成比赛

成绩复核

成绩发布

裁判培训训训

分组候场

出现问题

裁 判

异 议

调 试

延 时

仲裁委

三次加密

信息解密

**九、竞赛试题**

见附件1。

**十、评分标准制定原则、评分方法、评分细则**

**（一）评分标准制定原则**

竞赛评分严格按照公平、公正、公开的原则，评分标准注重考查参赛选手以下各方面的能力和水平。

1.实施构建大数据基础应用平台的技术技能；

2.实施数据采集与数据清洗的技术技能；

3.实施大数据分析与展现的技术技能；

4.实施大数据挖掘与应用开发的技术技能；

5.业务分析和工作总结的技术职业能力。

**（二）评分方法**

按照《全国职业院校技能大赛成绩管理办法》的相关要求，根据申报赛项自身的特点，制定的评分方法和细则如下。

1.赛卷满分为1000分。参赛队最终成绩按照实际得分的1/10折算成百分制。

2.本竞赛参与赛项成绩管理的组织机构包括裁判组、监督组和仲裁组，受赛项执委会领导。竞赛设置裁判22人，包括裁判长1名，裁判21名。其中加密裁判3人，现场裁判4人，评分裁判14人。加密裁判，现场裁判，评分裁判分别隔离，杜绝人为干预评分可能性，评分裁判评分时，随机抽签组队。监督组3人，对裁判组的工作进行全程监督，并对竞赛成绩抽检复核。仲裁组3人，负责接受由参赛队领队提出的对裁判结果的申诉，组织复议并及时反馈复议结果。

3.竞赛将制定裁判遴选管理办法、赛事保密细则和预案、命题管理办法等制度，保证竞赛的公平公正。赞助企业、参赛院校不安排人员进入裁判团队。

4.竞赛采取三层加密。第一组加密裁判组织参赛队选手第一次抽签，抽取参赛编号，替代选手参赛证等个人信息；第二组加密裁判组织参赛选手进行第二次抽签，确定赛位号，替换选手参赛编号；第三组加密裁判对各参赛队竞赛结果进行加密，替换赛位号。三层加密信息由不同加密裁判密封后保管，在评分结束后开封统计成绩。

5.本赛项的评分方法分为：结果评分和现场评分。本赛项考核环节：平台搭建、数据采集与清洗、数据分析与展现、工程文档的考核采用结果评分，占95%，由评分裁判依据赛项评价标准对结果进行评分。结果评分采取分步得分、累计总分的计分方式。各环节分别计算得分，错误不传递，按规定比例计入团队总分。根据赛题情况划分模块，每两名裁判负责一个模块进行独立评分，取两名评分裁判的平均分作为该参赛队该模块的最后得分。两名裁判评分差距超过5%时，由裁判长主持复核。现场评分是对参赛队的操作规范、职业素养、现场表现等进行评分。评分结果由参赛选手、裁判员、裁判长签字确认。现场评分占比5%。裁判长在竞赛结束18小时内提交评分结果，经复核无误，由裁判长、监督人员和仲裁人员签字确认后公布。

6.裁判长正式提交评分结果并复核无误后，加密裁判在监督人员监督下进行三层解密：竞赛作品编号到工位号解密；工位号到参赛编号解密；参赛编号到参赛队名称解密。

7.为保障成绩评判的准确性，监督组对赛项总成绩排名前30%的所有参赛队伍的成绩进行复核；其余成绩进行抽检复核，抽检覆盖率不低于15%。

8.监督组在复检中发现错误，需以书面形式及时告知裁判长，由裁判长更正成绩并签字确认。如复核、抽检错误率超过5%，裁判组需对所有成绩进行复核。

9.在竞赛过程中，参赛选手如有不服从裁判裁决、扰乱赛场秩序、舞弊等行为的，由裁判长按照规定扣减相应分数，情节严重的将取消比赛资格，比赛成绩计0分。

**（三）评分细则**

评分标准:总分1000分

|  |  |  |  |
| --- | --- | --- | --- |
| **考核环节** | **考核任务** | **考核知识点和技能点** | **分值** |
| 方案分析与理解（5%） | 分析方案要点，理解项目需求，指定需求说明书 | 对方案的理解与分析能力。 | 10 |
| 关键技术的选型说明。 | 10 |
| 根据需求说明书，制定项目计划书 | 工作量分解。 | 20 |
| 人员分工。 | 10 |
| 大数据平台安装部署（5%） | 大数据平台安装部署和规划 | 大数据平台连接测试正常，各虚拟机间网络连通。 | 20 |
| 大数据平台组件安装与部署 | 选择合适的组件进行安装与配置。 | 30 |
| 数据采集与清洗（15%） | 数据爬取 | 爬虫软件的安装和部署 | 10 |
| 建立爬虫任务 | 20 |
| 编辑合适的爬虫代码 | 20 |
| 爬虫作业的运行与下载 | 20 |
| Kafka配置 | 远程登录Kafka的集群节点 | 10 |
| 相关话题的创建 | 20 |
| Flume配置 | 相关配置组的创建 | 10 |
| 配置正确的agent | 20 |
| 配置正确的Source、Channel与Sink | 20 |
| 数据分析与存储（10%） | 数据实时处理 | 在应用开发平台中使用代码样例，高效处理获取数据 | 40 |
| 访问Hbase表数据 | HiveSQL的配置与外部表的创建 | 40 |
| 数据存储 | 将计算结果按照写入HBase非关系型数据库中 | 20 |
| 数据可视化展示（20%） | BI软件的部署 | BI软件的安装 | 20 |
| 数据源的配置 | BI管理页面的配置与数据源的导入 | 40 |
| 业务包的配置 | 业务包的创建与数据表管理 | 30 |
| 数据转义 | 20 |
| 数据关联 | 20 |
| 数据ETL处理 | 50 |
| 可视化图表创建 | 根据要求创建可视化图表 | 20 |
| 数据挖掘-数据建模（20%） | 数据导入 | 数据应用平台的登录与数据导入 | 30 |
| 数据解析 | 数据解析的配置 | 30 |
| 数据切分 | 根据要求对数据进行合适的切分 | 30 |
| 多维分析 | 多维特征分析方法 | 30 |
| 数据建模 | 数据建模的算法选择 | 20 |
| 正确的预测评估 | 20 |
| 建模评估 | 评估或预测模型的可靠性 | 30 |
| 模型下载 | 结果下载 | 10 |
| 数据挖掘-数据预测（15%） | WEB工程下载 | 下载方式正确 | 20 |
| 模型导入 | 模型导入操作正确 | 30 |
| 数据库配置 | 创建正确的用户与数据库 | 40 |
| 数据导入正确 | 20 |
| WEB页面验证 | 登录WEB页面 | 20 |
| 输出预测结果 | 20 |
| 项目文档（5%） | 根据项目需求和实施成果，编写系统使用文档。 | 根据项目需求和实施成果，编写系统使用文档。 | 50 |
| 职业素养（5%） | 项目实施符合企业“5S”（即整理、整顿、清扫、清洁和素养）原则，团队分工明确合理、操作规范、文明竞赛。 | 项目实施符合企业“5S”（即整理、整顿、清扫、清洁和素养）原则，团队分工明确合理、操作规范、文明竞赛。 | 20 |
| 团队分工明确合理、操作规范、文明竞赛。 | 团队分工明确合理、操作规范、文明竞赛。 | 30 |

**十一、奖项设置**

竞赛设参赛选手团体奖，以赛项实际参赛队总数为基数，一等奖占比10%，二等奖占比20%，三等奖占比30%，小数点后四舍五入。

获得一等奖的参赛队指导教师获“优秀指导教师奖”，授予荣誉证书。

**十二、技术规范**

|  |  |  |
| --- | --- | --- |
| **序号** | **标准号/规范简称** | **名称** |
| 1 | ISO/IEC JTC 1 SC32 | 数据管理与交换  |
| 2 | ISO/IEC JTC1/SC27 | 信息安全标准 |
| 3 | GB/T18142-20 00 | 信息技术数据元素值格式记法 |
| 4 | GB/T 20009-2005 | 信息安全技术数据库管理系统安全评估准则 |
| 5 | GB/Z 28828-2012 | 信息安全技术公共及商用服务信息系统个人信息保护指南 |
| 6 | 2010-3324T-SJ | 信息技术元数据质量要求框架 |
| 7 | 20121409-T-4 69 | 非结构化数据表示规范 |
| 8 | 20121411-T-4 69 | 非结构化数据管理系统技术要求 |

**十三、建议使用的比赛器材、技术平台和场地要求**

**（一）赛项所需的技术平台**

在充分考虑到赛项在国赛、省赛中的持续使用，以及全国院校参赛后将相关大数据技术应用到实际教学，所以本着实训与实验并重、教学与应用结合的原则，因此建议赛项技术平台采用基于实际项目式的大数据平台软件，并对其进行虚拟化部署。

**1.竞赛软件平台——标准软件平台**

竞赛将提供计算机，并部分预装Windows 10、Microsoft Office 2013、Adobe Reader等常用软件。

**2.竞赛技术平台——大数据软件平台与系统**

A厂家技术平台

本技术平台采用符合国际大数据标准的、采用主流大数据厂商的大数据框架进行搭建。

硬件系统：主要由高性能x86服务器、网络交换机、个人电脑等组成。

软件系统：大数据软件平台、商业智能系统、虚拟化平台及软件。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **编号** | **类型** | **设备名称** | **型号** | **产品描述** | **数量** |
| 1 | 硬件 | 网络交换机 | LS-5130-28S-EI | S5130-28S-EI L2以太网交换机主机 | 1 |
| 2 | 软件 | 大数据软件平台 | LIS-BD-HDP-1N | H3C DataEngine BD2M1HDP2C-HDP系统1license | 3 |
| 3 | 软件 | DataEngine BI商业智能系统 | LIS-BD-BI-10U | DataEngine BI商业智能系统全功能版--10 user License | 1 |
| 4 | 软件 | 虚拟化平台 | VC-CAS-ENT | CAS-CAS云计算管理平台-纯软件(DVD) | 1 |
| 5 | 软件 | 虚拟化管理系统 | LIS-CAS-CVMA-STD-2 | CAS-CVM虚拟化管理系统标准版软件License费用-管理2个物理CPU | 1 |
| 6 | 软件 | 创新应用开发平台 | LIS-BD-ADE-APP | Data ADE应用开发平台软件License费用-应用构建工厂 | 1 |
| 7 | 硬件 | 服务器 | NaviData 5200 G2 | NaviData 5200 G2 8SFF CTO大数据一体机或其他品牌同等配置服务器（内存不少于256G，CPU支持虚拟化） | 1 |
| 8 | 硬件 | 个人电脑 |  | Windows 10系统，预装Office 2013、Adobe Reader等常用软件 | 3 |

B厂家技术平台

本技术平台采用符合国际大数据标准的、采用主流大数据厂商统一的大数据框架进行搭建。

硬件系统：主要由高性能x86服务器、网络交换机、个人电脑等组成。

软件系统：大数据软件平台、 商业智能系统、虚拟化平台及软件。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **编号** | **类型** | **设备名称** | **产品描述** | **数量** |
| 1 | 硬件 | 网络交换机 | 以太网交换机主机，实现主机之间互联 | 1 |
| 2 | 软件 | 大数据平台 | 大数据存储与计算平台 | 3 |
| 3 | 软件 | 商业智能系统 | 智能构建数据分析图形、图表 | 1 |
| 4 | 软件 | 虚拟化平台 | 云计算管理平台-纯软件 | 1 |
| 5 | 软件 | 数据ETL平台 | 数据抽取、清洗、转化 | 1 |
| 6 | 软件 | 虚拟化管理系统 | 虚拟化管理系统软件License-管理2颗物理CPU | 1 |
| 7 | 软件 | 应用开发平台 | 大数据应用快速开发平台 | 1 |
| 8 | 硬件 | 服务器 | 大数据一体机或其他品牌同等配置服务器（内存不少于256G，CPU支持虚拟化） | 1 |
| 9 | 硬件 | 个人电脑 | Windows 10系统，预装Office 2013、Adobe Reader等常用软件 | 3 |

**（二）竞赛场地和环境标准**

1.竞赛场地应为地面平整、明亮、通风的室内场地，场地面积应不小于1000㎡，场地净高应不低于3.5m。

2.每个竞赛工位应能够提供独立的电源，其供电负荷不小于0.5kw，且含安全的接地保护，每个赛位8-10㎡。

3.每个竞赛工位应提供性能完好的竞赛平台、相关工具和电脑3台，安装竞赛所需的相关软件。

**（三）安全防范措施**

1.参赛选手根据规定确认竞赛设备、工具是否安全完好，严格遵守赛场规章、操作规程，保证人身及设备安全，接受裁判员的监督和警示，文明竞赛；

2.参赛选手安装部署竞赛设备时，请详细了解各设备性能参数，如供电输入等，确保设备的正常使用；

3.参赛选手连接传感器及其他套件时，注意防止正负极短路，避免烧坏；

4.参赛选手如遇设备故障，请及时示意现场裁判，保证竞赛的正常进行；

5.参赛选手请勿触屏和打开实训工位配电箱，注意实训工位后面220V强电使用安全。

**十四、安全保障**

根据《2017年全国职业院校技能大赛安全管理规定》的有关要求，制以下具体的安全管理措施并严格执行。

(一) 成立安全保障工作组，负责本赛项筹备和比赛期间的各项安全工作，赛项执委会主任和安全保障工作组组长均为第一责任人；

(二) 制定安全管理的相应规范、流程和突发事件应急预案，保证比赛筹备和实施工作全过程的安全；

(三) 确保比赛内容涉及的器材、设备符合国家有关安全规定；

(四) 赛项执委会在赛前对本赛项全体裁判员、工作人员、指导老师、参赛选手进行安全培训；

(五) 赛项执委会制定专门方案保证比赛命题、赛题保管、发放、回收和评判过程的安全；

(六) 赛项执委会在赛前组织专人对比赛现场、住宿场所和交通保障进行考察，并对安全工作提出明确要求。赛场的布置，赛场内的器材、设备，应符合国家有关安全规定，组织模拟测试，排除安全隐患；

(七) 赛场周围设立警戒线，防止无关人员进入，发生意外事件。比赛现场内参照相关职业岗位的要求为选手提供必要的劳动保护和医务服务；

(八) 承办院校应提供保障应急预案实施的条件，明确制度和预案，并配备急救人员与设施；

(九) 赛项执委会会同承办院校制定开放赛场和体验区的人员疏导方案。赛场环境中存在人员密集、车流人流交错的区域，除了设置齐全的指示标志外，增加引导人员，并开辟备用通道；

(十) 大赛期间，赛项承办院校须在赛场设置医疗医护工作站；在管理的关键岗位，增加力量，建立安全管理日志。比赛期间安排的住宿地应具有宾馆、住宿经营许可资质，保证住宿、卫生、饮食安全等；

(十一) 比赛期间发生意外事故时，发现者应第一时间报告赛项执委会，同时采取措施，避免事态扩大。赛项执委会应立即启动预案予以解决并向赛区执委会报告。

**十五、经费概算**

参照《2017年全国职业院校技能大赛赛项经费管理规定》的有关要求，制定赛项经费预算如下：

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **项目阶段** | **资金用途** | **费用** |
| 1 | 方案论证 | 专家论证会议 | 3 |
| 2 | 赛前准备 | 3次专家筹备会、（含差旅交通、食宿）、模拟题开发 | 8 |
| 全国赛前说明会 | 3 |
| 3 | 比赛现场 | 竞赛设备 | 合作企业提供 |
| 设备运输、安装调试 | 5 |
| 专家、监考和裁判、现场技术支持、后勤保障劳务费 | 10 |
| 赛场布置、技术展示体验 | 5 |
| 参赛选手奖品 | 5 |
| 竞赛指南印刷、选手服装等 | 5 |
| 竞赛现场办公文具、耗材等 | 5 |
| 4 | 比赛费用 | 总结研讨会 | 5 |
| 5 | 其它费用 | 临时突发事件 | 2 |
| 小计(单位：万元) | 56 |

**十六、比赛组织与管理**

经大赛执委会批准，成立赛项执行委员会、赛项专家组，落实赛项承办院校。

(一) 赛项执委会：全面负责本赛项的筹备与实施工作，接受大赛执委会领导，接受赛项所在分赛区执委会的协调和指导。赛项执委会的主要职责包括：领导、协调赛项专家组和赛项承办院校开展本赛项的组织工作，管理赛项经费，选荐赛项专家组人员及赛项裁判与仲裁人员等；

(二) 赛项专家组：在赛项执委会领导下开展工作，负责本赛项技术文件编撰、赛题设计、赛场设计、设备拟定、赛事咨询、技术评点、赛事成果转化、赛项裁判人员培训、赛项说明会组织等竞赛技术工作；同时负责赛项展示体验及宣传方案设计；

(三) 承办院校：在赛项执委会领导下，负责承办赛项的具体保障实施工作，主要职责包括：按照赛项技术方案要求落实比赛场地及基础设施，赛项宣传，组织开展各项赛期活动，参赛人员接待，生活服务，比赛过程文件存档等工作，赛务人员及服务志愿者的组织，赛场秩序维持及安全保障，赛后搜集整理大赛影像文字资料上报大赛执委会等。赛项承办院校按照赛项预算执行各项支出。承办院校人员不得参与所承办赛项的赛题设计和裁判工作；

(四) 现场裁判、仲裁、监督组：开赛前一周，在裁判员库、仲裁员库、监督员库中随机抽取组成。裁判组负责赛前检查及赛场鉴定、现场执裁和评审比赛结果等工作；仲裁组负责受理各参赛队的书面申诉、对受理的申诉进行深入调查，做出客观、公正的集体仲裁；监督组对指定赛区、赛项执委会的竞赛筹备与组织工作实施全程现场监督，包括赛项竞赛场地和设施的部署、选手抽签、裁判培训、竞赛组织、成绩评判及汇总、成绩发布、申诉仲裁、成绩复核等；

(五) 协办企业：提供竞赛现场设备并设置技术保障组，为竞赛设备、软件与竞赛设施提供保障服务。

**十七、教学资源转化建设方案**

大数据技术与应用专业作为新设专业，各院校在专业建设过程中还存在建设思路不清晰，专业办学基础还较弱，师资队伍欠缺，实训条件不足，教学资源短缺等问题。因此，赛项资源转化作为一项长项工作，统筹规划，真正做到大赛引领专业建设与发展的作用。

**（一）专业咨询服务**

联合北京大数据研究院、新华三、京东、美团、东软、联想、阿里巴巴等行业企业，与职业院校一同成立“大数据产教合作联盟”，发挥联盟专家智慧和行业企业资源优势，开展如下工作：

1.定期发布“中国大数据人才需求白皮书”，为院校提供行业最新的技术发展趋势、岗位需求等信息；

2.根据最新的技术发展和岗位需求，定期更新发布“大数据技术与应用专业建设参考规范”，更好地引领院校专业建设；

3.定期举办“中国大数据应用型人才培养高峰论坛”，搭建校企合作交流和信息互通平台；

4.不定期举办“CXO走进职业院校”活动，邀请新华三集团、东软集团、联想集团、阿里巴巴、美团、京东、BOSS直聘等主流IT企业的CTO、CHO等高管走进职业院校，使职业院校的师生能近距离了解行业发展趋势和企业用人需求，更好的规划职业发展和专业发展。

**（二）教学资源建设**

1.课程标准与课程资源：围绕大数据技术与应用专业核心课程，建设《大数据应用导论》、《大数据基础平台与关键技术》、《大数据分析与工具使用》、《Python语言程序设计》、《数据可视化技术与应用》、《大数据技术与应用综合实训项目》等核心课程标准，以及课程配套的教材、教案、课件、微课等资源。

2.大数据技术与应用生产实际教学案例库：围绕大数据技术与应用实践教学需要，建设大数据技术与应用生产实际教学案例库，覆盖课程、课程群和专业的一体化的项目教学资源，项目案例数量不低于20个，更好地服务院校实践教学。

**（三）师资队伍建设**

师资队伍是专业建设的核心保障条件，计划从专业与课程开发能力、教学能力和实战能力三方面提升职业院校大数据技术与应用专业师资队伍水平。

1.专业与课程开发能力。基于成果导向教育理念（OBE，Outcome Based Education）、CDIO工程教育理念和悉尼协议的相关要求，研究制定相关培训体系，系统化提升职业院校教师的专业与课程开发能力。

2.教学能力。围绕大数据技术与应用专业5—8门核心课程，制定核心课程的课程标准、教学资源，组织核心课程教学能力提升师资培训。

3.实战能力。通过大数据技术与应用专业实战项目，组织师资培训，系统性提升大数据专业教师实战能力。

**十八、筹备工作进度时间表**

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **筹备阶段** | **内　容** | **时间安排** |
| 1 | 申报、立项 | 赛项设计专家研讨会，完成赛项申报方案 | 2017年8月 |
| 确定赛项 | 2017年10月 |
| 成立赛项专家组 |
| 2 | 赛前准备 | 赛项专家会议3-5轮次会议，确定赛项规程、样题、赛项技术方案、赛场方案、体验环节设计方案、开放方案、宣传方案、教学资源转化方案、赛事安全规章、突发事件应急预案等 | 2017年11月～2018年 1 月 |
| 确定分赛区及承办校 | 2018年1月 |
| 全国赛项说明会 | 2018年2月 |
| 命题专家组会议，赛题开发、确定竞赛题库 | 2018年1月～2月 |
| 赛项预报名及报名完成 | 2018年2月～4月 |
| 在大赛网络信息发布平台上公开全部赛题 | 2018年4月上旬 |
| 3 | 比赛阶段 | 比赛设备安装、调试，赛场布置、同期技术展示、体验和活动现场布置；赛项指南印刷、选手服装制作 | 2018年5月～6月中旬 |
| 专家组题库审核，确定评分标准及抽题 |
| 成立裁判组、仲裁组、监督组；培训并验收赛场 |
| 正式比赛、同期技术展示、体验和活动举办；竞赛成绩提交、竞赛过程文档提交、教学资源转化成果与赛项总结 |

**十九、裁判人员建议**

根据教育部及大赛组委会要求，推荐经验丰富的一线教师、行业专家、企业技术专家作为裁判，裁判选取和培训工作将按照大赛组委会的统一安排执行。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **序号** | **裁判分类** | **知识能力要求** | **专业技术职称****（职业资格等级）** | **人数** |
| 1 | 裁判长 | 熟练掌握大数据采集、存储、处理与分析、传输与应用等技术，具备大数据工程项目的系统集成能力、应用软件设计和开发能力，具有一定的大数据科学研究能力及数据科学家的基本能力 | 高级职称以上或企业高级工程师 | 1 |
| 2 | 现场裁判 | 熟悉大数据采集、存储、处理与分析、传输与应用等 | 中级职称以上的教师或企业的工程师 | 4 |
| 3 | 加密裁判 | 无 | 无 | 3 |
| 4 | 评分裁判 | 熟悉大数据采集、存储、处理与分析、传输与应用等 | 中级职称以上的教师或企业的工程师 | 14 |
| 裁判总人数 | 22 |

**二十、其他**

**“2018年全国职业院校技能大赛”高职组**

**大数据技术与应用竞赛**

**样题**

# 第一部分：竞赛说明

一、 竞赛内容分布

* 互联网舆情分析数据采集、清洗、分析与展示
* 互联网舆情分析数据建模与预测

二、 竞赛时间

竞赛时间为4个小时。

三、 竞赛注意事项

1. 竞赛所需的硬件、软件和辅助工具由组委会统一布置，选手不得私自携带任何软件、移动存储、辅助工具、移动通信等进入赛场。
2. 请根据大赛所提供的比赛环境，检查所列的硬件设备、软件清单、材料清单是否齐全，计算机设备是否能正常使用。
3. 操作过程中，需要及时保存设备配置。比赛结束后，所有设备保持运行状态，不要拆动硬件连接。
4. 比赛完成后，比赛设备、软件和赛题请保留在座位上，禁止将比赛所用的所有物品（包括试卷和草纸）带离赛场。
5. 裁判以各参赛队提交的竞赛结果文档为主要评分依据。所有提交的文档必须按照赛题所规定的命名规则命名。

四、 赛题说明

* 赛场提供5台虚拟机，其中3台安装大数据综合平台，2台安装应用开发平台、商业智能系统。
* 考生通过WEB页面登录到大数据综合平台中，对平台进行相应的组件操作。系统用户名和密码均为admin。
* Windows、CentOS的用户密码都是Root2018。
* 竞赛结果文件的制作请参考U盘“竞赛资料\模板”中的 “大数据配置模板.doc”。请注意排版和格式，文档格式和排版得分将计入总成绩。
* 竞赛所需的技术参考文档在PC机桌面的“竞赛资料\网络应用\参考文档”中。

五：设备及文档列表

本竞赛中所使用的设备及软件如下表所示：

设备及软件列表

|  |  |  |
| --- | --- | --- |
| 序号 | 设备类型 | 数量 |
| 1 | 服务器 | 1 |
| 2 | 交换机 | 1 |
| 3 | 大数据综合平台 | 1 |
| 4 | 应用开发平台 | 1 |
| 5 | 商业智能系统 | 1 |
| 6 | 爬虫系统 | 1 |

本竞赛中使用的相关文档如下表所示：

文档列表

|  |  |  |
| --- | --- | --- |
| 序号 | 名称 | 位置 |
| 1 | 大数据服务模板.doc | U盘中的“竞赛资料\模板” |
| 2 | 大数据综合平台开局指导书 | PC机桌面上的“竞赛资料\大数据技术与应用\参考文档” |
| 3 | 应用开发平台使用手册 | PC机桌面上的“竞赛资料\大数据技术与应用\参考文档” |
| 4 | 商业智能系统使用手册 | PC机桌面上的“竞赛资料\大数据技术与应用\参考文档” |

六、 竞赛结果文件的提交

按照题目要求，提交符合模板的WORD文件。

# 第二部分：互联网舆情分析

1.

## 背景

网络舆情是指在互联网上流行的对社会问题不同看法的[网络舆论](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=29208037&ss_c=ssc.citiao.link)，是[社会舆论](http://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=560870&ss_c=ssc.citiao.link)的一种表现形式，是通过互联网传播的公众对现实生活中某些热点、焦点问题所持的有较强影响力、倾向性的言论和观点。网络舆情是以网络为载体，以事件为核心，广大网民情感、态度、意见、观点的表达、传播与互动，以及后续影响力的集合。

利用大数据综合平台优秀的数据采集、数据解析、数据存储功能，实现了热点事件的实时采集、来源追溯、情感分析、热点事件挖掘及状态的及时跟踪，同时借助大数据可视化工具，以图表形式展示出热点事件的发展脉络与趋势，包括产生、发展、消亡全过程。

利用ADE大数据科学平台对于数据进行分析、建模、预测舆情数据的倾向等。比如根据前一个月的热门话题的数据预测以后几天的变化趋势。

## 任务一：大数据平台配置

大数据综合平台初始完成了软件部署，但只添加了基本服务，包括HDFS、YARN、ZooKeeper、Metrics。其它需要使用的组件需要手动添加。

**组件部署**

为了完成互联网舆情的分析，需要规划并配置HDP所需的组件，具体要求如下：

* HDP组件配置合理：HDP组件有30多个，为了合理利用大数据平台资源的开支，合理添加组件；
* 大数据综合平台用户名和密码均为admin；

服务器IP地址信息如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
| **服务器** | **IPv4地址** | **网关** | **操作系统** |
| 大数据综合平台 | 192.168.1.10/24 | 192.168.1.1 | Centos6.5 |
| 192.168.1.11/24 | 192.168.1.1 | Centos6.5 |
| 192.168.1.12/24 | 192.168.1.1 | Centos6.5 |
| 应用开发平台、商业智能系统 | 192.168.1.13/24 | 192.168.1.1 | Centos7.0 |
| 192.168.1.14/24 | 192.168.1.1 | Win7 |

根据上述信息，在大数据综合平台上完成本次互联网舆情分析所需组件的添加。主要任务如下：

* 通过WEB登入到大数据综合平台，用户名和密码均为admin；
* 通过服务页签，添加大数据综合平台所需组件；
* 配置大数据综合平台的HA，设定好系统的容灾机制。

## 任务二：数据采集与清洗

网络舆情分析的数据源来自互联网。为了完成互联网舆情的分析，需要通过爬虫工具从互联网爬取数据，并对数据进行解析，保存在指定目录下；通过Flume定时读取数据推送到Kafka中。具体任务如下：

* 按照指定的网站[www.2018guosai.com](http://www.2018guosai.com)（事先已经建好）爬取数据；
* 完成对爬取数据的解析工作；
* 配置Flume定时读取爬取的数据；
* 配置Kafka完成作为数据缓冲，能被Spark读取数据；
* 大数据综合平台用户名和密码均为admin。

注：为避免数据量小，实验平台原先已采集部分数据。本部分采集的数据作为后续任务输入的一部分增量数据。

根据上述信息，在大数据综合平台上完成数据采集和清洗。主要分为以下三步：

1. **数据爬取**
* 请考生到IP地址为192.168.1.13服务器的/opt/目录下找到爬虫软件并完成安装和部署；
* 配置爬虫数据萃取表达式，把采集到的html内容提取出xml格式，xml格式要求包括标题，简要内容、发生时间、新闻热度、阅读量；
* 配置爬取数据存放在目录，能被Flume读取；
* 存放文件命名格式要求统一：时间+后缀。

数据爬取并下载，具体步骤如下：

* + 通过WEB登入到Datahunter软件，用户名和密码均为admin；
	+ 新建爬虫任务；
	+ 编辑爬虫代码；
	+ 运行爬虫作业；
	+ 查看结果并下载。
1. **定时抽取**

爬虫工具爬下来的数据保存在指定目录下，需要通过Flume定时的抽取到大数据平台。主要任务如下

* 配置Flume定时抽取数据，时间设定为5分钟抽取一次。

数据抽取具体步骤如下：

1、通过WEB登录大数据综合平台Manager管理界面（用户名和密码均为admin），并对Flume进行相关配置。

2、创建配置组；

3、配置flume监测配置的目录下新增的文件，并将文件数据读取出来，sink到Kafka。

4、配置名为a1的agent的Source、Channel与Sink分别为：ss1、c1、sk1；

5、配置名为c1的Channel，包括类型、容量以及传输容量；

6、配置ss1名称的Source，包括类型、channel、监听目录；

7、配置sk1名称的Sink使用Kafka sink，配置输出Kafka的主题和节点。

8、保存配置组成功后，重启服务。

1. **数据缓冲**

采集到的数据送往消息中间件Kafka作为缓冲，以解决采集数据的速度和数据处理的速度不一定同步的问题。考生需要完成Kafka的配置，具体任务如下：

* 创建话题，启动Kafka。

数据缓冲配置具体步骤如下：

1、找到安装有Kafka的集群节点，并以ssh方式远程登录该节点；

2、进入Kafka的bin原生工具命令行的bin目录；

3、创建一个话题，话题topicA

## 任务三：数据分析和存储

完成数据源的采集后，需要Spark定时去Kafka取数据，进行分析。具体任务如下：

* 在应用开发平台构建中下载的实时计算框架Spark Streaming工程的代码样例，开发Spark代码对数据进行分析；
* Spark代码主要内容为；
* 中文分词：将新闻话题进行分词，依据话题的关注、转发、评论的主体内容计算词频；
* 情感分析：依据话题中分词的词性进行话题评论的情感分析，将评论内容进行打分，计算社会舆论的主题情绪；
* 统计：话题量，按时间热点新闻关注度。
* 数据存储：将新闻数据解析、汇总和统计，包含分词、情感分析、热门话题。并将计算结果按照写入HBase非关系型数据库中；
* 使用Hive创建外部表，可通过HiveSQL访问Hbase表数据

数据分析和存储具体步骤如下：

1. 下载Spark样例代码：登录ADE后，点击<开始构建>，进入应用构建工厂页面。点击<舆情分析>，进入舆情分析示例页面，点击<源码下载>，下载源码至本地。
2. 使用Spark样例代码编写本实验所需的代码。
3. 使用Hive命令创建外部表。

## 任务四：可视化展示

完成数据分析和存储后，可通过商业智能系统对现有的数据进行可视化展示。具体任务如下：

* 部署商业智能系统软件，请考生到IP地址为192.168.1.13服务器的/opt/目录下找到BI软件并完成安装。
* 完成可视化图表展示，展示的内容包括如下；
* 在不同时间段热门话题的转发关注量变化（TOP10）；
* 话题情绪变化趋势。
* 当前热点新闻TOP10

数据可视化展示步骤如下：

1、数据源配置

登入商业智能系统管理页面，配置数据源，通过商业智能系统软件定义的数据连接到Hive中取数据，获取到的数据自动保存在Cube中。

2、创建业务包并对其进行配置，具体如下：

创建业务包；数据表管理；数据转义；数据关联；数据ETL处理。

3、可视化图表创建

进行可视化图表创建，要包含如下要素：

a）在不同时间段热门话题的转发关注量变化（TOP10）；

b）话题情绪变化趋势。

c）当前热点新闻TOP10

## 任务五：数据挖掘-数据建模

为了完成对于舆情数据的预测，必须对原有的数据进行分析、建模、评估等任务，本次任务是对热门话题数据进行分析、建模，具体任务如下：

* 对现有的数据导入到《数据科学工厂》模块；
* 对数据进行解析工作，能被数据科学工厂识别；
* 对数据进行切片工作，把数据切分为训练集、验证集和评估集等多种数据集。使用数据的80%来分析建模，20%数据来验证所建立模型的准确性；
* 对数据进行多维分析，查看数据的分布形式，对数据进行关联分析；
* 对数据进行建模，使用数据科学工厂提供的各种算法工具对数据进行建模；
* 建模评估，使用切分的20%数据对建立的模型进行评估，验证模型的准确性；
* 模型下载，作为数据预测使用。

根据上述信息，在应用开发科学平台上完成数据的导入、解析、切分、分析、建模、评估等工作。具体步骤如下：

1、导入数据，登录应用开发平台，进入《数据科学工厂》模块，导入数据。

2、解析数据，对指定的数据文件进行数据解析配置。要求包含如下要素，源文件指定、数据文件格式、分隔符、列名，列名称编辑、数据类型编辑。

3、数据切分，将加载到计算平台中的数据进行切分，切分为训练集、验证集和评估集等多种数据集，每个切分后的数据集的命名为原文件+切分比例(0.8/0.2)。

4、多维分析，配置要使用的多维特征分析方法，包括概率分布统计，关联性分析，占比分析， 趋势分析。要求配置合理，能看到数据的分布形式，并对数据进行关联分析。

5、数据建模，选择合理的算法来建模和预测评估，使用“选择算法”中选择要使用的算法，点击“创建模型”建模；

6、预测评估，选择预测评估的数据集，模型，然后进行预测评估，点击开始预测评估按钮然后等待其结束，完成建模。

7、建模评估，通过未参与训练的数据来评估或预测模型的可靠性，打开预测评估界面并进行配置，查看原先建模的准确性。

8、下载模型，下载模型是指把训练完成的模型导出为模型文件，并下载到本地。

## 任务六：数据挖掘-数据预测

完成数据的建模后，通过WEB工程，根据前一个月的热门话题数据预测后面几天的变化趋势，具体任务如下：

* 在应用开发平台的应用构建工厂下载WEB展示应用；
* 导入创建好的模型；
* 创建数据库、导入WEB初始化数据；
* 预测热门话题在以后几天的变化趋势。

数据预测具体步骤如下：

1、下载WEB工程，登录应用开发平台后，进行全量下载。

2、导入模型，将模型导入应用开发平台及工程中。

3、修改数据库，在安装postgresql的linux系统上，通过xshell登录，并执行以下操作：创建用户，数据库，授权；导入web初始化表数据。

4、预测结果，登录WEB管理页面<http://localhost:8080/program>，录入数据，点击提交，输出预测结果。

## 任务七：竞赛结果文件制作

上述实施完成后，参照U盘“竞赛资料\模板”中的“大数据配置模板.doc”，制作本组的项目配置文档。文档名称为“大数据配置－工位号.doc”。

考生将文件 “大数据配置－工位号.doc”保存到PC桌面上，并在考试结束后拷贝到U盘中提交给现场工作人员。

***注意：考生在U盘中所提交的文件是竞赛结果的唯一依据，请考生一定确保文件确实有效，能够正常读取。如有疑问，可咨询现场工作人员。***