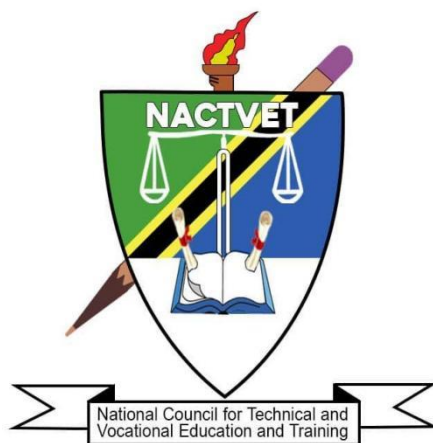


**NATIONAL COUNCIL FOR TECHNICAL AND VOCATIONAL EDUCATION AND
TRAINING**



FEBRUARY 2023

PROPOSED OCCUPATIONAL STANDARDS

OCCUPATION: BIG-DATA ENGINEER

LEVEL: NTA 8

TABLE OF CONTENT

CONTENTS

ABBREVIATIONS	ii
GLOSSARY OF TERMS.....	iii
1.0. INTRODUCTION	1
2.0. OCCUPATIONAL STANDARD DEVELOPMENT PROCESS.....	2
3.0. THE SCOPE AND OVERVIEW OF THE OCCUPATION STANDARDS FOR BIG-DATA ENGINEERS	2
4.0. VALIDITY PERIOD.....	3
5.0. OCCUPATIONAL STANDARDS	3
5.1 OCCUPATIONAL STANDARDS FOR FOR BIG-DATA ENGINEER - NTA 8	4

ABBREVIATIONS

BDSE	Big Data Software Engineer
CBET	Competency Based Education and Training
NACTVET	National Council for Technical and Vocational Education and Training
NOS	National Occupational Standards
OS	Occupational Standards
PL	Production Line
SOP	Standard Operating Procedures
TET	Technical Education and Training
DE	Data Engineer
TVET	Technical and Vocational Education and Training

GLOSSARY OF TERMS

Circumstantial Knowledge:	Detailed knowledge, which allows the decision-making in regard to different circumstances and cross cutting issues.
Competence:	The ability to use knowledge, understanding, practical, and thinking skills to perform effectively to the workplace standards required in employment.
Competency:	A description of the ability one possesses when able to perform a given occupational task effectively and efficiently.
Competency-based Education:	An instructional programme that derives its content from validated tasks and bases assessment on the learner's performance.
Curriculum:	A description or composite of statements about "what is to be learned" by the trainee/student in a particular instructional programme; a product that states the "intended learning outcomes".
Educational/Training Programme:	The complete curriculum and instruction (what and how) that is designed to prepare a person for employment in a job or other particular performance situation.
Occupation:	A specific position requiring the performance of specific tasks – essentially the same tasks are performed by all employees having the same title. (Example: baker)
Occupational Area:	The extensive grouping of related work, such as big data development.
Occupational Competence:	The application of knowledge and skills that consistently meet the standards required by the work context.
Occupational Standards:	Specific requirements of competences people are expected to demonstrate in a particular occupational area, including knowledge and relevant attitudes. They also act as a performance tool of assessment of the prescribed outcomes.
Occupational/Job Analysis:	A process used to identify the tasks that are important to employees in any given occupation.
Performance Criteria:	Indicate expected end results or outcomes in the form of evaluative statements.
Skills:	The ability to perform occupational tasks with a high degree of proficiency within a given occupation. Skill is conceived of as a composite of three completely interdependent components: cognitive, affective, and psychomotor.

Standards:	A set of statements, which if proved true under working conditions, means that an individual is meeting an expected level and type of performance.
Task Analysis:	The process of analysing each task to determine the steps, circumstantial knowledge, attitudes, performance standards, tools and materials needed, as well as safety concerns required for the employees performing it.
Task:	A work activity that has a definite beginning and ending, is observable or measurable, and consists of two or more definite steps that leads to a product, service, or decision.
Underpinning Knowledge:	Crucial knowledge that an individual must acquire in order to demonstrate competences that are associated in performing a given task.
Verification Process:	The process of having experts review and confirm the importance of the task (competency) statements identified through occupational analysis. Other questions, such as the degree of task learning difficulty are also frequently asked. This process is also sometimes referred to as validation.

1.0. INTRODUCTION

Technical Education and Training (TET) is one of the most important education sub-sectors in Tanzania, responsible for developing a skilled workforce to support the country's industrialization economic agenda. Tanzania's *Development Vision 2025* intends to raise the country's economy to a middle-income status. This requires a skilled workforce that is aligned with the needs of the public and private sectors of the economy. The National Council for Technical Education has begun the job of drafting Occupational Standards that will eventually be adopted as National Occupational Standards for TET in order to ensure that it meets the needs of the labour market and the country's economic agenda.

National Occupational Standards (NOS) are performance criteria that are matched with labour market demands. Each National Occupation Standard describes functions, performance standards, and knowledge/understanding for one important function or task. They combine skills, knowledge, and attitudes to describe best practice. They are useful tools for establishing job roles, personnel recruiting, supervision, and appraisal, as well as TET standards. They're also helpful for benchmarking and harmonizing qualifications on a national and international level. Standards, in general, provide a solid framework for high-quality TET that is labour market-relevant, current, and consistent in delivery across all public and private institutions.

However, it must be noted that, Occupational Standards and Training standards/qualifications standards are different. Occupational standards are defined in terms of activities performed by a person in a selected occupation and they are usually defined by employers following procedures agreed upon by all stakeholders. Education and training standards are developed from the activities defined in occupational standards, and they include learning objectives to ensure that the necessary skills and knowledge are developed by a person to enable him or her to function at an agreed level in an occupation. Education and Training standards are used to define curricula in training institutions. It is however critical that there must be a direct link between the occupational standards and the training standards to respond to the demands of the labour market.

In TET delivery, Tanzania adopted the Competence Based Education and Training (CBET) approach. The CBET approach focuses on providing learners with the skills and knowledge required to meet the occupational standards. Occupational standards are thus the starting point for developing competency-based training (CBET) programmes. TET institutions will be required to benchmark their curricula with relevant occupational standards.

Occupational Standards are developed based on a given occupation's current and future demands. As a result, they serve as a means of bridging the gap between the worlds of employment and technical education and training (TET).

The Big-Data Engineer Occupation has its own set of occupational standards. The document explains how the occupational standards were developed, as well as the scope, the occupational profile in the form of DACUM charts, and the Occupational Standards.

2.0. OCCUPATIONAL STANDARD DEVELOPMENT PROCESS

The Occupational standards development process began with an examination of major documents that guide Tanzanian skill development. The 10-year National Skills Development Strategy (2016~2026) was one of the documents reviewed, and it outlined six (6) economic sectors that should be prioritized when developing skills development programmes.

These sectors include: Transport and Logistics, Tourism and Hospitality, Agribusiness, Construction, Energy and ICT. NACTE labour market reports were also used in the literature review to determine the skills demand in the Tanzanian labour market as a whole.

After the literature review, a workshop comprised of expert workers and educators with substantial knowledge and experience in the occupation conducted an occupational analysis utilizing the DACUM approach to produce the occupational profile. The analysis resulted in DACUM Charts, which are attached as Appendix 1 to this document.

The occupational standards were then developed. Experts in Occupational Analysis and the Development of Occupational Standards facilitated the workshop. Interviews, online surveys, and a stakeholder forum were used to validate the Occupational Standards. Engineers, supervisory technicians on the job, and experienced Big-Data Engineers were key informants in the survey to discover occupational trends. This information was used to gain insight from the workplaces regarding trends and changes in the profession, including how well graduates are prepared for working in the occupation. A total of ... online surveys were completed by experts from the labour market across the country. Apart from the surveys aiding in defining the scope for the occupational analysis, they also served to engage a wide cross-section of experts in the occupation. Apart from this, the stakeholders' forum was attended by ... participants from different parts of the country representing various companies.

3.0. THE SCOPE AND OVERVIEW OF THE OCCUPATION STANDARDS FOR BIG-DATA ENGINEERS

The standards cover a broad range of duties and tasks that can be performed by a Big-Data Engineer. However, the occupational standards are not meant to replace individual job descriptions. Instead, they are to be used for guidance in defining skill levels and knowledge for the technician in specific settings or positions. The Big-Data Engineer may perform tasks in a number of key areas of the

occupational standards, but not necessarily in all areas. For example, in large operations, other individuals may be employed or designated to perform specific tasks.

Big-Data Engineers perform production management, complex technical activities and so on. They also undertake tasks such as production process management, application, operation and maintenance of complex data equipment, troubleshooting of faults in complex electronic equipment in the workshop. Generally, the Big-Data Engineer performs the following responsibilities:

- a. Data collection and processing
- b. Big data analysis
- c. Data warehouse management
- d. Data visualization
- e. Big data architecture design
- f. Data security and privacy
- g. Technical research and innovation

The Occupational Standards have been clustered into the NTA qualification level, i.e. NTA level 8.

4.0. VALIDITY PERIOD

Due to the rapid development of technology, the validity period of occupational standards is 3-5 years. The review will proceed in the same manner as the one before it, with new occupational standards being developed based on current trends of the labour market.

5.0. OCCUPATIONAL STANDARDS

5.1 OCCUPATIONAL STANDARDS FOR BIG-DATA ENGINEER - NTA 8

OCCUPATION	BIG-DATA ENGINEER	OCCUPATION CODE	
DUTY TITLE	COMPETENCE TO DIAGNOSE AND TUNE BIG DATA PLATFORMS AND COMPONENTS	DUTY NO.	801
TASK TITLE	PLANNING OF HARDWARE SYSTEM DEPLOYMENT SCHEMES	TASK NO.	8011
PERFORMANCE CRITERIA	The person performing this task must be able to plan hardware system deployments in accordance with the construction requirements of a big data platform.		
RANGE STATEMENT	The hardware system deployments can be completed under the supervision of management authorities. The equipment or tools to be used include: 1. The number of server nodes; 2. Central processing unit (CPU); 3. Solid state drive; 4. Dual-port gigabit Ethernet card; 5. Database server; 6. Backup server; 7. EVEREST.		
EVIDENCE REQUIREMENT			
PRACTICAL PERFORMANCE		UNDERPINNING KNOWLEDGE	
The person performing this task must be able to do the following: 1. Plan the hardware conditions required for the cluster according to the construction requirements of the big data platform; 2. Complete the preparation of cluster configuration documentation according to professionals' instructions; 3. Correctly use cluster resources, such as cloud servers, object storage services, cloud network services, and cloud database services, according to the operational documentation; 4. Independently complete the installation and setup of an operating system; 5. Operate using both the graphical and command line interfaces.		Detailed knowledge about: 1.0 Methods The person performing this task must be able to explain how to: 1.1 Explain the fundamentals, functions and application scenarios of big data technologies; 1.2 Explain the architecture of a big data system, including the components of data collection, storage, processing, and analysis; 1.3 Explain the requirements and limitations of big data processing on hardware systems; 1.4 Explain the advantages and disadvantages of different hardware selections and configurations; 1.5 Explain the scaling and optimization schemes for hardware systems, including the principles and operation methods of scale-out, scale-up, load balancing, fault-tolerant design, etc. 2.0 Principles The person performing this task must be able to explain the following principles:	

	2.1 Principles and applications of big data technologies; 2.2 Architecture and components of big data systems; 2.3 Principles of hardware selection and configuration. 3.0 Theories The person performing this task must be able to explain the following: 3.1 Server parameter configuration; 3.2 Virtual environment of the system. 4.0 Essential Skills 4.1 Communication skills; 4.2 Computer network skills; 4.3 Document editing skills; 4.4 Teamwork skills; 4.5 Linux system commands. 5.0 Digital Skills 5.1 Skills in algorithms and models; 5.2 Skills in pre-processing algorithms; 5.3 Optimization algorithms; 5.4 Network design; 5.5 Big data storage; 5.6 Scripting.
CIRCUMSTANTIAL KNOWLEDGE	Detailed knowledge about: 1. Operations for system deployment; 2. Operations for service deployment; 3. Testing hardware system information.

OCCUPATION	BIG-DATA ENGINEER	OCCUPATION CODE	
DUTY TITLE	COMPETENCE TO DIAGNOSE AND TUNE BIG DATA PLATFORMS AND COMPONENTS	DUTY NO.	801
TASK TITLE	DEVELOPMENT OF STRATEGIES FOR DEPLOYMENT, UPGRADE AND OPTIMIZATION OF CLUSTER COMPONENTS	TASK NO.	8012
PERFORMANCE CRITERIA	The person performing this task must be able to develop the strategies for deployment, upgrade and optimization of cluster components in accordance with customer demands.		
RANGE STATEMENT	The task can be performed in offices, requiring the communication with the Operation & Maintenance Engineer. The materials and tools to be used include: 1. Cluster configuration and stress test documents; 2. Component deployment planning documents; 3. Documents on component version features; 4. Text editing tools; 5. Software or tools for building the Hortonworks Data Platform (HDP).		
EVIDENCE REQUIREMENT			
PRACTICAL PERFORMANCE		UNDERPINNING KNOWLEDGE	
The person performing this task must be able to do the following: 1. Build the Ambari big data management platform according to the requirements of the tasks of big data platform component installation and configuration; 2. Determine the cluster component configuration strategy by referring to the cluster configuration and component deployment planning documents; 3. Determine the compatibility between components according to the documents on component version features; 4. Conduct a stress test after deploying the service and develop an upgrade and optimization strategy.		Detailed knowledge about: 1.0 Methods The person performing this task must be able to explain how to: 1.1 Deploy cluster components; 1.2 Upgrade and optimize cluster components; 1.3 Explain the requirements and limitations of big data processing on hardware systems. 2.0 Principles The person performing this task must be able to explain the following principles: 2.1 Inner working principles of big data platforms and components; 2.2 Architecture and components of big data systems; 2.3 Performance indicators and monitoring tools. 3.0 Theories The person performing this task must be able to explain the following: 3.1 Inner working principles of platforms and components;	

	3.2 Methods for collecting and analyzing these indicators; 3.3 Competence in analyzing and diagnosing performance issues of big data platforms and components. 4.0 Essential Skills 4.1 Communication skills; 4.2 Computer network skills; 4.3 Document editing skills; 4.4 Teamwork skills; 4.5 Linux system commands. 5.0 Digital Skills 5.1 Skills in algorithms and models; 5.2 Skills in pre-processing algorithms; 5.3 Optimization algorithms; 5.4 Network design; 5.5 Big data storage; 5.6 Scripting.
CIRCUMSTANTIAL KNOWLEDGE	Detailed knowledge about: 1. Operations for deployment of big data cluster components; 2. Operations of big data management platforms; 3. Cluster performance stress test.

OCCUPATION	BIG-DATA ENGINEER	OCCUPATION CODE	
DUTY TITLE	COMPETENCE TO DIAGNOSE AND TUNE BIG DATA PLATFORMS AND COMPONENTS	DUTY NO.	801
TASK TITLE	DEVELOPMENT OF HIGH AVAILABILITY AND DISASTER TOLERANCE STRATEGIES	TASK NO.	8013
PERFORMANCE CRITERIA	The person performing this task must be able to develop high availability configuration and disaster tolerance strategies in accordance with customer demands.		
RANGE STATEMENT	The task can be performed under the supervision of the Project Manager, requiring the communication with the Architect. The equipment and software to be used include: 1. Hadoop High Availability Deployment Guide; 2. Data backup software; 3. Dual-active disaster tolerance software; 4. Disaster tolerance migration service management terminal; 5. Database server; 6. Application server; 7. Disaster tolerance server; 8. Virtual disk library; 9. Firewall; 10. Optical fiber switch.		
EVIDENCE REQUIREMENT			
PRACTICAL PERFORMANCE		UNDERPINNING KNOWLEDGE	
The person performing this task must be able to do the following: 1. Independently configure the high availability cluster Hadoop file parameters according to the requirements of the task of big data platform high availability deployment and the High Availability Deployment Guide for Big Data Platforms; 2. Determine the impact of the construction of the disaster tolerance system on the original business system; 3. Determine the latency imposed on the system I/O by data replication techniques; 4. Determine the stress imposed on the day-to-day business processing system by application data synchronization; 5. Perform remote data replication based on the storage system;		Detailed knowledge about: 1.0 Methods The person performing this task must be able to explain how to: 1.1 Configure a high availability cluster; 1.2 Construct data disaster tolerance; 1.3 Understand the concepts of high availability and disaster tolerance; 1.4 Analyze and assess possible risks and potential impacts on the system. 2.0 Principles The person performing this task must be able to explain the following principles: 2.1 Principles and applications of big data technologies; 2.2 Architecture and components of big data systems; 2.3 Implementation and testing strategies.	

6. Back up data in real time to an off-site server on the local server by timing.	3.0 Theories The person performing this task must be able to explain the following: 3.1 Server parameter configuration; 3.2 Virtual environment of the system; 3.3 Data backup; 3.4 Data migration. 4.0 Essential Skills 4.1 Communication skills; 4.2 Documentation skills; 4.3 Computer programming skills; 4.4 Teamwork skills; 4.5 Program debugging and problem solving ability; 4.6 Problem analysis and positioning skills. 5.0 Digital Skills 5.1 Skills in algorithms and models; 5.2 Skills in pre-processing algorithms; 5.3 Optimization algorithms; 5.4 Network design; 5.5 Big data storage; 5.6 Scripting.
CIRCUMSTANTIAL KNOWLEDGE	Detailed knowledge about: 1. Operations of big data cluster high availability; 2. Construction of a data disaster tolerance center; 3. Construction of a data backup system.

OCCUPATION	BIG-DATA ENGINEER	OCCUPATION CODE	
DUTY TITLE	COMPETENCE TO DIAGNOSE AND TUNE BIG DATA PLATFORMS AND COMPONENTS	DUTY NO.	801
TASK TITLE	DEVELOPMENT OF MONITORING SCRIPTS AND TROUBLESHOOTING AND REVIEW OF FAULTS AND ACCIDENTS	TASK NO.	8014
PERFORMANCE CRITERIA	The person performing this task must master the scripting language and develop scripts to monitor cluster faults.		
RANGE STATEMENT	The task can be performed in offices or machine rooms. The equipment and software to be used include: 1. Apache Ambari; 2. Visual Studio Code; 3. Zabbix.		
EVIDENCE REQUIREMENT			
PRACTICAL PERFORMANCE		UNDERPINNING KNOWLEDGE	
The person performing this task must be able to do the following: 1. Be familiar with the planning paths for big data platform components; 2. Develop scripts on development tools using the Shell programming language; 3. Provision, manage and monitor clusters using web tools; 4. Trace the fault in the cluster monitoring system; 5. Review and retain documents after troubleshooting.		Detailed knowledge about: 1.0 Methods The person performing this task must be able to explain how to: 1.1 Have programming and script development skills; 1.2 Have problem troubleshooting skills; 1.3 Review the fault to gain insight into its cause and root cause; 1.4 Realize automation and self-healing ability. 2.0 Principles The person performing this task must be able to explain the following principles: 2.1 Principles of development of monitoring scripts; 2.2 Principles of fault review; 2.3 Principles of automation and self-healing ability. 3.0 Theories The person performing this task must be able to explain the following: 3.1 Monitoring scripts; 3.2 Review of fault causes. 4.0 Essential Skills 4.1 Communication skills; 4.2 Documentation skills;	

	4.3 Computer programming skills; 4.4 Teamwork skills; 4.5 Program debugging and problem solving ability; 4.6 Problem analysis and positioning skills. 5.0 Digital Skills 5.1 Skills in algorithms and models; 5.2 Skills in pre-processing algorithms; 5.3 Optimization algorithms; 5.4 Network design; 5.5 Big data storage; 5.6 Scripting.
CIRCUMSTANTIAL KNOWLEDGE	Detailed knowledge about: 1. Shell scripting using development tools; 2. View of the cluster status using monitoring platform tools; 3. Troubleshooting and review.

OCCUPATION	BIG-DATA ENGINEER	OCCUPATION CODE	
DUTY TITLE	DATA COLLECTION AND DEVELOPMENT OF PRE-PROCESSING SCHEMES	DUTY NO.	802
TASK TITLE	DEVELOPMENT OF DATA COLLECTION STRATEGIES AND MONITORING OF THE COLLECTION	TASK NO.	8021
PERFORMANCE CRITERIA	The person performing this task must master data collection technologies and be able to develop data collection strategies and monitor the collection.		
RANGE STATEMENT	The task can be performed with the cooperation of the ETL data engineer. The equipment and software to be used include: 1. Flume; 2. Scribe; 3. Chukwa monitoring system; 4. Ganglia; 5. Text editing tool.		
EVIDENCE REQUIREMENT			
PRACTICAL PERFORMANCE		UNDERPINNING KNOWLEDGE	
The person performing this task must be able to do the following: 1. Be familiar with the planning paths for big data platform components; 2. Use different data collection technologies for data collection tasks; 3. Use a data monitoring system to monitor and analyze the collection; 4. Analyze the characteristics of the data and formulate a data collection strategy through the position of the data on the disk.		Detailed knowledge about: 1.0 Methods The person performing this task must be able to explain how to: 1.1 Apply data collection rules; 1.2 Conduct data monitoring; 1.3 Understand the business requirements and data models; 1.4 Have knowledge of a variety of data sources and be able to select the appropriate data source based on factors such as type, source, and reliability of the data. 2.0 Principles The person performing this task must be able to explain the following principles: 2.1 Data quality monitoring; 2.2 Principles of data processing and conversion; 2.3 Principles of data source selection. 3.0 Theories The person performing this task must be able to explain the following: 3.1 Collection technologies for all types of data; 3.2 Operations of the data collection monitoring system.	

	4.0 Essential Skills 4.1 Communication skills; 4.2 Skills in data sensitivity; 4.3 Documentation skills; 4.4 Computer programming skills; 4.5 Teamwork skills. 5.0 Digital Skills 5.1 Skills in algorithms and models; 5.2 Skills in pre-processing algorithms; 5.3 Optimization algorithms; 5.4 Network design; 5.5 Big data storage; 5.6 Scripting.
CIRCUMSTANTIAL KNOWLEDGE	Detailed knowledge about: 1. Application scenarios for data collection technologies; 2. Operations of the data collection monitoring system; 3. Indicator dimensions of data collection.

OCCUPATION	BIG-DATA ENGINEER	OCCUPATION CODE	
DUTY TITLE	DATA COLLECTION AND DEVELOPMENT OF PRE-PROCESSING SCHEMES	DUTY NO.	802
TASK TITLE	DATA MIGRATION STRATEGY DEVELOPMENT AND MIGRATION MONITORING	TASK NO.	8022
PERFORMANCE CRITERIA	The person performing this task must be able to perform data migrations correctly and monitor the migrations.		
RANGE STATEMENT	The task can be performed with the cooperation of the ETL data engineer. The software and tools to be used include: 1. Sqoop; 2. MySQL database; 3. ClickHouse database; 4. Clickhouse-copier; 5. Text editing tool.		
EVIDENCE REQUIREMENT			
PRACTICAL PERFORMANCE		UNDERPINNING KNOWLEDGE	
The person performing this task must be able to do the following: 1. Be familiar with the location of data storage; 2. Understand the indicators of data migration; 3. Use tools to migrate data; 4. Consider the risk of data migration in business scenarios; 5. Use text tools to summarize the data migration process into a document.		Detailed knowledge about: 1.0 Methods The person performing this task must be able to explain how to: 1.1 Apply data migration rules; 1.2 Monitor data migration; 1.3 Select the appropriate data source based on factors such as the type, source and reliability of the data; 1.4 Explain the technologies and tools used (such as ETL and programming language). 2.0 Principles The person performing this task must be able to explain the following principles: 2.1 Data migration strategy; 2.2 Monitoring and analysis of data migration; 2.3 Principles of big data resource allocation. 3.0 Theories The person performing this task must be able to explain the following: 3.1 Data migration rules; 3.2 Data migration operations; 3.3 Monitoring of the data migration process.	

	4.0 Essential Skills 4.1 Communication skills; 4.2 Skills in data sensitivity; 4.3 Documentation skills; 4.4 Computer programming skills; 4.5 Teamwork skills. 5.0 Digital Skills 5.1 Skills in algorithms and models; 5.2 Skills in pre-processing algorithms; 5.3 Optimization algorithms; 5.4 Network design; 5.5 Big data storage; 5.6 Scripting.
CIRCUMSTANTIAL KNOWLEDGE	Detailed knowledge about: 1. Application scenarios for data migration technologies; 2. Operations of the data migration monitoring system; 3. Rules of data falling off the database.

OCCUPATION	BIG-DATA ENGINEER	OCCUPATION CODE	
DUTY TITLE	DATA COLLECTION AND DEVELOPMENT OF PRE-PROCESSING SCHEMES	DUTY NO.	802
TASK TITLE	DESIGN OF STORAGE STRATEGIES	TASK NO.	8023
PERFORMANCE CRITERIA	The person performing this task must be able to design data storage strategies correctly.		
RANGE STATEMENT	The task can be performed with the cooperation of the database engineer. The software and tools to be used include: 1. Databases; 2. Message-oriented middleware; 3. Text editing tool; 4. Software or tools for building the Hortonworks Data Platform (HDP).		
EVIDENCE REQUIREMENT			
PRACTICAL PERFORMANCE		UNDERPINNING KNOWLEDGE	
The person performing this task must be able to do the following: 1. Be familiar with the location of data storage; 2. Understand the data source dictionary; 3. Use commands for data storage; 4. Analyze the database to be used for data in different business scenarios; 5. Ensure the integrity of data without omission; 6. Use text tools to summarize the data storage scheme into a document.		Detailed knowledge about: 1.0 Methods The person performing this task must be able to explain how to: 1.1 Apply data storage formats; 1.2 Perform data storage operations; 1.3 Understand the business requirements and data characteristics; 1.4 Provide appropriate storage architecture design schemes; 1.5 Partition and fragment data. 2.0 Principles The person performing this task must be able to explain the following principles: 2.1 Storage requirement analysis; 2.2 Design of storage architecture; 2.3 Storage performance optimization. 3.0 Theories The person performing this task must be able to explain the following: 3.1 Storage types of various databases; 3.2 Data storage formats; 3.3 Data lifecycle. 4.0 Essential Skills 4.1 Communication skills;	

	4.2 Skills in data sensitivity; 4.3 Documentation skills; 4.4 Database programming skills; 4.5 Teamwork skills. 5.0 Digital Skills 5.1 Skills in algorithms and models; 5.2 Skills in pre-processing algorithms; 5.3 Optimization algorithms; 5.4 Network design; 5.5 Big data storage; 5.6 Scripting.
CIRCUMSTANTIAL KNOWLEDGE	Detailed knowledge about: 1. Database application scenarios; 2. Operation of data integration; 3. Structured data and unstructured data.

OCCUPATION	BIG-DATA ENGINEER	OCCUPATION CODE	
DUTY TITLE	DATA COLLECTION AND DEVELOPMENT OF PRE-PROCESSING SCHEMES	DUTY NO.	802
TASK TITLE	FORMULATION OF DATA CLEANING PROCESS SPECIFICATIONS	TASK NO.	8024
PERFORMANCE CRITERIA	The person performing this task must be able to standardize the data format and formulate the data cleaning process.		
RANGE STATEMENT	The task can be performed with the cooperation of the ETL engineer. The software and tools to be used include: 1. Kettle; 2. Database; 3. Message-oriented middleware; 4. Text editing tool.		
EVIDENCE REQUIREMENT			
PRACTICAL PERFORMANCE		UNDERPINNING KNOWLEDGE	
The person performing this task must be able to do the following: 1. Independently complete the design of data table structures; 2. Select the appropriate data structure according to the characteristics of the application data not used and the characteristics of the commonly-used data structures. 3. Save data as different types of files according to the characteristics of various file storage formats; 4. Ensure data integrity and no data loss; 5. Ensure data traceability and security; 6. Use text tools to summarize the data cleaning process into a document.		Detailed knowledge about: 1.0 Methods The person performing this task must be able to explain how to: 1.1 Follow the process of data cleaning; 1.2 Conduct data cleaning; 1.3 Consider issues such as data migration and data consistency; 1.4 Set the appropriate cleaning monitoring indicators. 2.0 Principles The person performing this task must be able to explain the following principles: 2.1 Principles and applications of big data technologies; 2.2 Architecture and components of big data systems; 2.3 Principles of data cleaning selection and configuration. 3.0 Theories The person performing this task must be able to explain the following: 3.1 Data cleaning rules; 3.2 Data structure; 3.3 Data quality.	

	4.0 Essential Skills 4.1 Communication skills; 4.2 Skills in data sensitivity; 4.3 Documentation skills; 4.4 Computer programming skills; 4.5 Teamwork skills. 5.0 Digital Skills 5.1 Skills in algorithms and models; 5.2 Skills in pre-processing algorithms; 5.3 Optimization algorithms; 5.4 Network design; 5.5 Big data storage; 5.6 Scripting.
CIRCUMSTANTIAL KNOWLEDGE	Detailed knowledge about: 1. Data channel and storage format; 2. Data cleaning operation; 3. Structured data and unstructured data.

OCCUPATION	BIG-DATA ENGINEER	OCCUPATION CODE	
DUTY TITLE	DATA COLLECTION AND DEVELOPMENT OF PRE-PROCESSING SCHEMES	DUTY NO.	802
TASK TITLE	DESIGN OF UNIFIED DATA PRE-PROCESSING SCHEMES	TASK NO.	8025
PERFORMANCE CRITERIA	The person performing this task must be able to process different types of data according to the requirements of data collection and pre-processing.		
RANGE STATEMENT	This task can be performed on the data collection components/tools used on big data platforms. The components to be used include: 1. Business data in MYSQL database; 2. Flume components; 3. KAFKA components; 4. Log data on the local disk; 5. Big data platform;		
EVIDENCE REQUIREMENT			
PRACTICAL PERFORMANCE		UNDERPINNING KNOWLEDGE	
The person performing this task must be able to do the following: 1. Understand the log format information by analyzing the log data on the local disk; 2. Design the pre-processing of the dirty read data in the operation log; 3. Complete the collection and cleaning of operation logs by writing pre-processing programs; 4. Put the cleaned log data on the designated storage platform; 5. Use tools to clean business data in the database; 6. Put the cleaned operation data on the designated storage platform;		Detailed knowledge about: 1.0 Methods The person performing this task must be able to explain how to: 1.1 Design pre-processing for different types of data; 1.2 Clean data; 1.3 Perform data conversion, including data format conversion, data normalization, data standardization, etc.; 1.4 Perform data segmentation and sampling for model training and evaluation. 2.0 Principles The person performing this task must be able to explain the following principles: 2.1 Feature selection and extraction; 2.2 Data segmentation and sampling 2.3 Data pre-processing process. 3.0 Theories The person performing this task must be able to explain the following: 3.1 Architecture and configuration of Flume components; 3.2 Architecture and configuration of KAFKA	

	components. 4.0 Essential Skills 4.1 Basic operation skills of LINUX system; 4.2 Flume component design skills; 4.3 KAFKA component design skills; 4.4 HDFS system operation skills; 4.5 Programming and writing skills; 4.6 MYSQL database operation skills; 4.7 Skills in using KETTLE; 4.8 Skills in using MAXWELL; 4.9 Skills in using DATAX. 5.0 Digital Skills 5.1 Skills in algorithms and models; 5.2 Skills in pre-processing algorithms; 5.3 Optimization algorithms; 5.4 Network design; 5.5 Big data storage; 5.6 Scripting.
CIRCUMSTANTIAL KNOWLEDGE	Detailed knowledge about: 1. Design and construction of Flume-KAFKA-HDFS log data collection system; 2. Writing of Flume interceptor; 3. Operation of log data collection channel; 4. Design and construction of MYSQL-HDFS business data collection system; 5. Operation of log data collection channel.

OCCUPATION	BIG-DATA ENGINEER	OCCUPATION CODE	
DUTY TITLE	DATA COLLECTION AND DEVELOPMENT OF PRE-PROCESSING SCHEMES	DUTY NO.	802
TASK TITLE	DEVELOPMENT OF EXCEPTION PRE-PROCESSING MECHANISMS AND OPTIMIZATION OF PERFORMANCE INDICATORS	TASK NO.	8026
PERFORMANCE CRITERIA	The person performing this task must be able to adjust and optimize the parameters of the collection system.		
RANGE STATEMENT	This task can be performed on the data collection components/tools used on big data platforms to adjust and optimize the parameters. The components to be used include: 1. Flume components; 2. KAFKA components; 3. DataX components.		
EVIDENCE REQUIREMENT			
PRACTICAL PERFORMANCE		UNDERPINNING KNOWLEDGE	
The person performing this task must be able to do the following: 1. Adjust and optimize the parameters of the Flume components according to the operational document provided; 2. Optimize the configuration of the core parameters of the KAFKA components according to the operational document provided; 3. Optimize the use of the DATAX components according to the operational document provided; 4. Detect and identify abnormal situations in the data pre-processing process; 5. Develop appropriate exception handling strategies, including methods such as repair, removal and substitution; 6. Develop and optimize appropriate performance indicators according to the requirements and objectives of the pre-processing process.		Detailed knowledge about: 1.0 Methods The person performing this task must be able to explain how to: 1.1 Optimize data collection components; 1.2 Select appropriate processing strategies to ensure data reliability and consistency; 1.3 Learn about common performance indicators such as running time, memory usage and throughput, and select appropriate optimization methods such as algorithmic improvement and parallel processing based on the actual situation; 1.4 Learn about the principles and technologies of streaming data processing and be able to design and implement appropriate real-time exception handling methods. 2.0 Principles The person performing this task must be able to explain the following principles: 2.1 Optimization of performance indicators; 2.2 Exception handling strategies; 2.3 Detection and identification of exceptions. 3.0 Theories The person performing this task must be able to explain the following:	

	3.1 Working principles of Flume components; 3.2 Working principles of KAFKA components; 3.3 Working principles of DATAX components. 4.0 Essential Skills 4.1 Skills in using Flume components; 4.2 Skills in using KAFKA components; 4.3 Skills in using DATAX components. 5.0 Digital Skills 5.1 Skills in algorithms and models; 5.2 Skills in pre-processing algorithms; 5.3 Optimization algorithms; 5.4 Network design; 5.5 Big data storage; 5.6 Scripting.
CIRCUMSTANTIAL KNOWLEDGE	Detailed knowledge about: 1. Flume parameter tuning and transaction mechanisms; 2. Optimization of DATAX usage; 3. Speed of each channel; 4. Number of Channel concurrency in DataX Job; 5. Configuration of speed limits of global Byte and single Channel Byte; 6. JVM heap memory; 7. KAFKA parameter tuning. (1) Optimization of producer core parameter configuration; (2) Optimization of Broker core parameter configuration; (3) Optimization of consumer core parameter configuration.

OCCUPATION	BIG-DATA ENGINEER	OCCUPATION CODE	
DUTY TITLE	BIG DATA ANALYSIS AND APPLICATION	DUTY NO.	803
TASK TITLE	CONSTRUCTION OF ANALYTICAL MODELS AND OPTIMIZATION OF FITTING RESULTS	TASK NO.	8031
PERFORMANCE CRITERIA	The person performing this task must be able to construct analytical models corresponding to data according to different characteristics of the data.		
RANGE STATEMENT	This task can be performed on the big data platform constructed to construct analytical models for the data warehouse. The designs to be implemented include: 1. Core architecture of the data warehouse; 2. Data warehouse modeling; 3. Cluster configuration and stress test documents; 4. Component deployment planning documents; 5. Documents on component version features; 6. Text editing tools; 7. Software or tools for building the Hortonworks Data Platform (HDP).		
EVIDENCE REQUIREMENT			
PRACTICAL PERFORMANCE		UNDERPINNING KNOWLEDGE	
The person performing this task must be able to do the following: 1. Have knowledge of the concept and core architecture of a data warehouse; 2. Master the theory of data warehouse modelling methods; 3. Extract meaningful features from original data and perform feature engineering to transform the features into a form usable by the model; 4. Understand various analytical models and be able to select appropriate models to solve specific problems; 5. Apply appropriate algorithms and technologies for model training and tuning models to improve their performance.		Detailed knowledge about: 1.0 Methods The person performing this task must be able to explain how to: 1.1 Construct a data warehouse; 1.2 Evaluate the model constructed; 1.3 Deploy the model constructed to the actual production environment and maintain and update it; 1.4 Consider factors such as scalability and efficiency. 2.0 Principles The person performing this task must be able to explain the following principles: 2.1 Data cleaning and pre-processing; 2.2 Feature extraction and engineering; 2.3 Model selection and construction. 3.0 Theories The person performing this task must be able to explain the following: 3.1 Server parameter configuration; 3.2 Virtual environment of the system;	

	3.3 Concept of data warehouse; 3.4 Significance of data warehouse modeling. 4.0 Essential Skills 4.1 Communication skills; 4.2 Computer network skills; 4.3 Document editing skills; 4.4 Teamwork skills; 4.5 Linux system commands; 4.6 Core architecture of the data warehouse; 4.7 Data warehouse modeling methods. 5.0 Digital Skills 5.1 Skills in algorithms and models; 5.2 Skills in pre-processing algorithms; 5.3 Optimization algorithms; 5.4 Network design; 5.5 Big data storage; 5.6 Scripting.
CIRCUMSTANTIAL KNOWLEDGE	Detailed knowledge about: 1. Core architecture of the data warehouse; 2. ER model; 3. Dimensional model; 4. Fact table for dimensional modeling theory; 5. Dimension table for dimensional modeling theory; 6. Essentials of the dimensional design.

OCCUPATION	BIG-DATA ENGINEER	OCCUPATION CODE	
DUTY TITLE	BIG DATA ANALYSIS AND APPLICATION	DUTY NO.	803
TASK TITLE	ANALYSIS OF DATA CHARACTERISTICS, SUMMARIZATION OF RESULTS AND OUTPUT OF ANALYSIS REPORTS	TASK NO.	8032
PERFORMANCE CRITERIA	The person performing this task must be able to analyze data characteristics using tools.		
RANGE STATEMENT	The task can be performed using MYSQL and other databases to analyze the data. The tools to be used include: 1. Sqoop; 2. MySQL database; 3. ClickHouse database; 4. Clickhouse-copier; 5. Text editing tools; 6. MYSQL database; 7. EZDML design tools.		
EVIDENCE REQUIREMENT			
PRACTICAL PERFORMANCE		UNDERPINNING KNOWLEDGE	
The person performing this task must be able to do the following: 1. Install and use mysql; 2. Install and use the EZDML design tool; 3. Process raw data, including missing value processing, abnormal value detection and processing, data standardization, etc., to facilitate subsequent accurate analysis; 4. Perform operations such as dimensionality reduction and extraction of useful features regarding the data to improve the accuracy and interpretability of the analysis; 5. Analyze and model the data.		Detailed knowledge about: 1.0 Methods The person performing this task must be able to explain how to: 1.1 Explain the reasons for using the EZDML design tool; 1.2 Deploy cluster components; 1.3 Upgrade and optimize cluster components; 1.4 Explain the requirements and limitations of big data processing on hardware systems. 2.0 Principles The person performing this task must be able to explain the following principles: 2.1 Principles and applications of big data technologies; 2.2 Architecture and components of big data systems; 2.3 Principles of hardware selection and configuration. 3.0 Theories The person performing this task must be able to	

	explain the following: 3.1 Server parameter configuration; 3.2 Virtual environment of the system; 3.3 Use of the EZDML tool to sort out the complex relationships between business tables. 4.0 Essential Skills 4.1 Installation and use of the Mysql database; 4.2 Installation and use of EZDML. 5.0 Digital Skills 5.1 Skills in algorithms and models; 5.2 Skills in pre-processing algorithms; 5.3 Optimization algorithms; 5.4 Network design; 5.5 Big data storage; 5.6 Scripting.
CIRCUMSTANTIAL KNOWLEDGE	Detailed knowledge about: 1. Data import to the mysql database; 2. EZDML installation and configuration; 3. Model setting; 4. Importing of data from mysql into EZDML; 5. Database configuration; 6. Table relationship establishment.

OCCUPATION	BIG-DATA ENGINEER	OCCUPATION CODE	
DUTY TITLE	BIG DATA ANALYSIS AND APPLICATION	DUTY NO.	803
TASK TITLE	CONSTRUCTION OF THE DATA MODELING ARCHITECTURE	TASK NO.	8033
PERFORMANCE CRITERIA	The person performing this task must be able to model data warehouses according to the theory of data warehouse construction.		
RANGE STATEMENT	This task can be performed on the big data platform constructed to construct the data warehouse. The analyses to be conducted include: 1. HIVE data warehouse components; 2. HADOOP big data platform; 3. ZOOKEEPER big data components.		
EVIDENCE REQUIREMENT			
PRACTICAL PERFORMANCE		UNDERPINNING KNOWLEDGE	
The person performing this task must be able to do the following: 1. Install and use the HIVE data warehouse according to the operational document provided; 2. Install, operate and maintain the HADOOP big data platform according to the operational document provided; 3. Configure and install the Spark components according to the operational document provided..		Detailed knowledge about: 1.0 Methods The person performing this task must be able to explain how to: 1.1 Make hierarchical planning for a data warehouse; 1.2 Know about business requirements, identify and collect relevant data, and understand data availability and feasibility; 1.3 Design and develop data collection and cleaning processes; 1.4 Design and construct data warehouses and data lakes and determine appropriate data models and architectures. 2.0 Principles The person performing this task must be able to explain the following principles: 2.1 Performance optimization and tuning; 2.2 Data visualization and report; 2.3 Data security and privacy protection; 2.4 Data modeling methods. 3.0 Theories The person performing this task must be able to explain the following: 3.1 Data research; 3.2 Clarification of data fields; 3.3 Construction of the business bus matrix;	

	3.4 Clarification of statistical indicators; 3.5 Dimensional modeling. 4.0 Essential Skills 4.1 HIVE library operations; 4.2 HIVE window function; 4.3 Skill in eliminating HIVE data skew; 4.4 Spark programming skills. 5.0 Digital Skills 5.1 Skills in algorithms and models; 5.2 Skills in pre-processing algorithms; 5.3 Optimization algorithms; 5.4 Network design; 5.5 Big data storage; 5.6 Scripting.
CIRCUMSTANTIAL KNOWLEDGE	Detailed knowledge about: 1. HIVE environment setup; 2. HIVE ON Spark configuration; 3. HIVE ON Spark test; 4. YARN environment configuration; 5. Data warehouse ODS modeling; 6. Data warehouse DWD modeling; 7. Data warehouse DWS modeling; 8. Data warehouse ADS modeling.

OCCUPATION	BIG-DATA ENGINEER	OCCUPATION CODE	
DUTY TITLE	BIG DATA ANALYSIS AND APPLICATION	DUTY NO.	803
TASK TITLE	DEVELOPMENT OF DATA QUERY OPERATING PROCEDURES AND SPECIFICATIONS	TASK NO.	8034
PERFORMANCE CRITERIA	The person performing this task must be able to perform the specified query on data via query components.		
RANGE STATEMENT	This task can be performed on the big data platform constructed to construct the data warehouse. The query components to be used include: 1. HIVE data warehouse; 2. Kylin components; 3. Presto components.		
EVIDENCE REQUIREMENT			
PRACTICAL PERFORMANCE		UNDERPINNING KNOWLEDGE	
The person performing this task must be able to do the following: 1. Install and use the Kylin components according to the operational document provided; 2. Install and use the Presto components according to the operational document provided.		Detailed knowledge about: 1.0 Methods The person performing this task must be able to explain how to: 1.1 Identify and collect relevant data and understand the availability and feasibility of the data; 1.2 Ensure the efficiency of data storage and management and support various data processing operations; 1.3 Design and implement data integration and ETL (extraction, conversion and loading) processes to integrate data from multiple data sources into a unified data model to support data analysis and modeling; 1.4 Select and apply appropriate data modeling methods, such as dimensional, relational and graphical modeling, to create a data structure that facilitates data analysis and prediction. 2.0 Principles The person performing this task must be able to explain the following principles: 2.1 Principles and applications of big data technologies; 2.2 Architecture and components of big data systems; 2.3 Data integration and ETL processes; 2.4 Data warehouse and data lake design.	

	3.0 Theories The person performing this task must be able to explain the following: 3.1 Cube construction principles in Kylin theory; 3.2 Kylin component architecture; 3.3 Kylin component-derived dimensions; 3.4 Presto component architecture. 4.0 Essential Skills 4.1 Deployment and use of Kylin components; 4.2 KylinBI integration; 4.3 Deployment and use of Presto components; 4.4 Teamwork skills; 4.5 Linux system commands. 5.0 Digital Skills 5.1 Skills in algorithms and models; 5.2 Skills in pre-processing algorithms; 5.3 Optimization algorithms; 5.4 Network design; 5.5 Big data storage; 5.6 Scripting.
CIRCUMSTANTIAL KNOWLEDGE	Detailed knowledge about: 1. Environmental preparation; 2. Use of Kylin; 3. Analysis of the Kylin duplicate key problem; 4. Kylin aggregation group; 5. Presto Server; 6. Presto Client.

OCCUPATION	BIG-DATA ENGINEER	OCCUPATION CODE	
DUTY TITLE	BIG DATA ANALYSIS AND APPLICATION	DUTY NO.	803
TASK TITLE	DEVELOPMENT OF EXTERNAL DATA INTERFACE SPECIFICATIONS AND AUTHORITIES	TASK NO.	8035
PERFORMANCE CRITERIA	The person performing this task must be able to complete the authentication service by deploying Kerberos components.		
RANGE STATEMENT	This task can be performed on the data warehouse platform constructed to configure the authentication service. The components to be configured include: 1. HADOOP big data platform; 2. HIVE data warehouse platform; 3. Kerberos components; 4. Ranger components; 5. Atlas components.		
EVIDENCE REQUIREMENT			
PRACTICAL PERFORMANCE		UNDERPINNING KNOWLEDGE	
The person performing this task must be able to do the following: 1. Install and use the Kerberos components according to the operational document provided; 2. Install and use the Ranger components according to the operational document provided; 3. Install and use the Atlas components according to the operational document provided; 4. Evaluate and optimize the performance of data interfaces to improve the efficiency and responsiveness of data transmission; 5. Write and maintain documents on data interfaces, including API documents, instructions, etc.		Detailed knowledge about: 1.0 Methods The person performing this task must be able to explain how to: 1.1 Explain the authentication principles of Kerberos components; 1.2 Explain the principles of the architecture of the Ranger components; 1.3 Monitor the operation of data interfaces to identify and solve problems in time. 2.0 Principles The person performing this task must be able to explain the following principles: 2.1 Data interface performance optimization; 2.2 Data interface documents and monitoring; 2.3 Data interface security protection. 3.0 Theories The person performing this task must be able to explain the following: 3.1 Kerberos authentication principles; 3.2 Ranger architecture theory; 3.3 Atlas architecture theory. 4.0 Essential Skills	

	4.1 Linux operation skills; 4.2 Computer network skills; 4.3 Document editing skills; 4.4 Teamwork skills. 5.0 Digital Skills 5.1 Skills in algorithms and models; 5.2 Skills in pre-processing algorithms; 5.3 Optimization algorithms; 5.4 Network design; 5.5 Big data storage; 5.6 Scripting.
CIRCUMSTANTIAL KNOWLEDGE	Detailed knowledge about: 1. Kerberos deployment; 2. Use of Kerberos; 3. Kerberos integration with Hadoop: Create system users and Kerberos principals; 4. Kerberos integration with hadoop: Configure HTTPS; 5. Kerberos integration with hadoop: Configure LinuxContainerExecutor; 6. Kerberos integration with hadoop: Start HDFS; 7. Kerberos integration with hadoop: Start Yarn and the history server.

OCCUPATION	BIG-DATA ENGINEER	OCCUPATION CODE	
DUTY TITLE	BIG DATA ANALYSIS AND APPLICATION	DUTY NO.	803
TASK TITLE	PREPARATION OF BATCH AND REAL-TIME DATA COMPUTATION TASKS	TASK NO.	8036
PERFORMANCE CRITERIA	The person performing this task must be able to realize the computation tasks by writing a computation program.		
RANGE STATEMENT	This task can be performed on the big data platform constructed to construct the data warehouse. The computational components to be used include: 1. HIVE data warehouse; 2. Flink components; 3. Spark components.		
EVIDENCE REQUIREMENT			
PRACTICAL PERFORMANCE		UNDERPINNING KNOWLEDGE	
The person performing this task must be able to do the following: 1. Install and use the Spark components according to the operational document provided; 2. Install and use the Flink components according to the operational document provided; 3. Have knowledge and proficiency in the use of popular big data computing frameworks; 4. Write logic for data computation tasks using programming languages (such as Java and Python) according to business requirements; 5. Demonstrate competence in optimizing the performance of big data computation tasks, and identifying and fixing potential performance bottlenecks.		Detailed knowledge about: 1.0 Methods The person performing this task must be able to explain how to: 1.1 Recognize offline and real-time business architectures in data warehouses; 1.2 Select the appropriate framework according to the actual needs; 1.3 Adjust resource allocation and data partitioning, and reasonably set the degree of parallelism, etc.; 1.4 Prepare batch data processing tasks, and understand and use the batch processing mode for data computation and analysis. 2.0 Principles The person performing this task must be able to explain the following principles: 2.1 Writing of the data processing logic; 2.2 Batch data processing; 2.3 Troubleshooting and debugging. 3.0 Theories The person performing this task must be able to explain the following: 3.1 Spark framework processing flow; 3.2 Internal structure of Spark-RDD; 3.3 Spark-SQL components;	

	3.4 Flink framework processing flow; 3.5 Evolution of the data processing framework. 4.0 Essential Skills 4.1 JAVA programming language; 4.2 SCALA programming language; 4.3 Document editing skills; 4.4 Teamwork skills; 4.5 Linux system commands. 5.0 Digital Skills 5.1 Skills in algorithms and models; 5.2 Skills in pre-processing algorithms; 5.3 Optimization algorithms; 5.4 Network design; 5.5 Big data storage; 5.6 Scripting.
CIRCUMSTANTIAL KNOWLEDGE	Detailed knowledge about: 1. Installation of Spark cluster; 2. Spark runtime architecture; 3. Spark-RDD operator operations; 4. Broadcast variables and accumulators; 5. Spark-SQL core programming; 6. Flink cluster installation; 7. Flink reading data; 8. Flink functions; 9. Flink read-in data.

OCCUPATION	BIG-DATA ENGINEER	OCCUPATION CODE	
DUTY TITLE	BIG DATA MINING	DUTY NO.	804
TASK TITLE	FEATURE ENGINEERING PROCESSING	TASK NO.	8041
PERFORMANCE CRITERIA	The person performing this task must be able to pre-process data to adapt the data to the model and match the requirements of the model; be able to process and construct numerical, textual, and time features to transform raw data into features that are more representative of the potential problems of the prediction model.		
RANGE STATEMENT	The task can be performed for data mining based on big data technologies. The tools and equipment to be used include: 1. Computers with CPU at I5 or higher level, RAM at 8GB or above and hard disk storage capacity of 500GB or above; 2. Computers with Unix/Linux operating systems installed; 3. Computers with various virtualized environments installed; 4. Computers with commonly-used scripting tools installed, such as Shell/Python/Scala/Java; 5. Computers with commonly-used big data mining tools installed, such as Apache Spark, Hadoop MapReduce and Qlik.		
EVIDENCE REQUIREMENT			
PRACTICAL PERFORMANCE		UNDERPINNING KNOWLEDGE	
1. Be able to remove noise from data and correct inconsistencies; 2. Combine multiple data sources into consistent data storage to constitute a complete data set, such as a data warehouse; 3. Compress data by aggregating, removing redundant attributes, or clustering; 4. Convert data from one format to another (such as standardization); 5. Extract new information as features from other unstructured data such as text, images and sounds, and evaluate the availability of the features; 6. Combine existing features (including categories, values, time features, etc.) or compute them against each other to construct new features; 7. Select features by filtering, packing, embedding and various dimensionality reduction algorithms; 8. Analyze the importance and weight of features; 9. Monitor major features to prevent		Detailed knowledge about: 1.0 Methods The person performing this task must be able to explain how to: 1.1 Deploy cluster components; 1.2 Upgrade and optimize cluster components; 1.3 Explain the requirements and limitations of big data processing on hardware systems. 2.0 Principles The person performing this task must be able to explain the following principles: 2.1 Principles and applications of big data technologies; 2.2 Architecture and components of big data systems; 2.3 Principles of hardware selection and configuration. 3.0 Theories The person performing this task must be able to explain the following: 3.1 Data mining methods and processes;	

degradation of their quality.	3.2 Basis for simple statistical analysis; 3.3 Knowledge of data preparation and pre-processing; 3.4 Knowledge of feature extraction, representation and selection; 3.5 Commonly-used machine learning algorithms; 3.6 Knowledge of databases; 3.7 Knowledge of data storage. 4.0 Essential Skills 4.1 Communication skills; 4.2 Computer network skills; 4.3 Document editing skills; 4.4 Teamwork skills; 4.5 Linux system commands. 5.0 Digital Skills 5.1 Skills in algorithms and models; 5.2 Skills in pre-processing algorithms; 5.3 Optimization algorithms; 5.4 Network design; 5.5 Big data storage; 5.6 Scripting.
CIRCUMSTANTIAL KNOWLEDGE	Detailed knowledge about: 1. Operating system; 2. Basic theory of big data; 3. Algorithms and data structure; 4. Distributed computing framework; 5. Software engineering; 6. Databases; 7. Statistical learning methods; 8. Data collection and pre-processing; 9. Communication and document writing; 10. Teamwork; 11. Occupational health and safety.

OCCUPATION	BIG-DATA ENGINEER	OCCUPATION CODE	
DUTY TITLE	BIG-DATA TECHNICAL SERVICES	DUTY NO.	805
TASK TITLE	BIG-DATA TECHNICAL CONSULTING	TASK NO.	8051
PERFORMANCE CRITERIA	The person performing this task must be able to collect all kinds of related information based on customer needs and accomplish the technical consulting work; be able to provide product presentations and display based on sales needs; be able to participate in projects and provide recommendations for product or architecture design.		
RANGE STATEMENT	The task can be performed to provide big-data technical consulting services at the company or customers' companies. The tools and equipment to be used include: 1. Computers with CPU at I5 or higher level, RAM at 8GB or above and hard disk storage capacity of 500GB or above; 2. Computers with commonly-used operating systems such as Windows, Mac and Linux installed; 3. Computers with various virtual machine environments installed; 4. Computer with big data development tools and related frameworks installed; 5. Computers with office software installed.		
EVIDENCE REQUIREMENT			
PRACTICAL PERFORMANCE		UNDERPINNING KNOWLEDGE	
1. Collect information on market objectives in accordance with the established plan of the team; 2. Work with the sales team to produce presentation materials and solution demonstration materials; 3. Work with the engineer to resolve customer technical inquiries and provide reference information; 4. Collect target market information and analyze industry needs; 5. Work with the sales team to deliver product presentations and solution demonstrations; 6. Independently resolve customers' technical inquiries and provide technical schemes; 7. Participate in the design of the project architecture and provide reference ideas; 8. Build a target market analysis model to analyze industry needs; 9. Output the product solution as a whole;		Detailed knowledge about: 1.0 Methods The person performing this task must be able to explain how to: 1.1 Clearly explain concepts, terms and principles in the technical field of big data; 1.2 Explain the overall architecture of a big data system; 1.3 Explain commonly-used tools and technologies in big data processing; 1.4 Provide technical consulting suggestions based on specific customer needs. 2.0 Principles The person performing this task must be able to explain the following principles: 2.1 Data processing tools and technologies; 2.2 Explanation of data models and algorithms; 2.3 Concepts of big data technologies. 3.0 Theories The person performing this task must be able to	

10. Independently solve technical problems consulted by customers and provide technical solutions; 11. Participate in the design of the project architecture and products and provide constructive suggestions;	explain the following: 3.1 Relational databases, NoSQL database, machine learning algorithms; 3.2 Recommendations for appropriate technology selection, system architecture design, performance optimization, etc.; 3.3 Relationships and functions of components such as data collection, data storage, data processing and data presentation; 3.4 Theoretical knowledge of big data storage, data processing, data analysis, etc. 4.0 Essential Skills 4.1 Applied knowledge of the big data industry; 4.2 Knowledge of big data technology systems; 4.3 Knowledge of big data architecture; 4.4 Knowledge of big data technology trends; 4.5 Knowledge of big data architecture analysis; 4.6 Knowledge of big data product design. 5.0 Digital Skills 5.1 Skills in algorithms and models; 5.2 Skills in pre-processing algorithms; 5.3 Optimization algorithms; 5.4 Network design; 5.5 Big data storage; 5.6 Scripting.
CIRCUMSTANTIAL KNOWLEDGE	Detailed knowledge about: 1. Operating system; 2. Basic theory of big data; 3. Big data industry development; 4. Big-data software architecture; 5. Software engineering; 6. Process management; 7. Operation management; 8. Project management; 9. Communication and document writing; 10. Teamwork; 11. Occupational health and safety.

OCCUPATION	BIG-DATA ENGINEER	OCCUPATION CODE	
DUTY TITLE	BIG-DATA TEAM BUILDING AND ABILITY TRAINING	DUTY NO.	806
TASK TITLE	BIG-DATA TECHNICAL TEAM BUILDING AND DEVELOPMENT OF VOCATIONAL ABILITY TRAINING	TASK NO.	8061
PERFORMANCE CRITERIA	The person performing this task must be able to research advanced big data technologies, undertake big data technology team building, and conduct vocational ability training for technicians and assistant engineers.		
RANGE STATEMENT	The task can be performed within a data team under the supervision of a big data director. The tools and equipment to be used include: 1. Computer; 2. Database and data warehouse; 3. Conference room; 4. Instant communication tools and information management platforms, such as shared documents, planning items and schedule management.		
EVIDENCE REQUIREMENT			
PRACTICAL PERFORMANCE		UNDERPINNING KNOWLEDGE	
The person performing this task must be able to do the following: 1. Research the advanced development trends and application scenarios of big data technologies to provide enterprises with strategic planning and suggestions for technologies; 2. Plan the building and management of the big data technology team based on the long-term strategy and needs of the enterprise and department to ensure that team members remain at the forefront of technology research and development; 3. Ensure that the team's development tasks are completed on time and that the product quality meets the requirements; 4. Train and mentor technicians and assistant engineers in their vocational ability to ensure that they have the skills and knowledge that match the development needs of the enterprise.		Detailed knowledge about: 1.0 Methods The person performing this task must be able to explain how to: 1.1 Develop specific training programs, including the determination of training topics, schedules, training methods, etc.; 1.2 Select appropriate training resources and methods; 1.3 Clearly explain the content and objectives of the training, including key points on specific technical fields, application and operation modes of tools and technologies; 1.4 Provide support and guidance in the training process. 2.0 Principles The person performing this task must be able to explain the following principles: 2.1 The principle of the importance of team building; 2.2 The principle of the objectives of vocational ability training; 2.3 Principles of training program development. 3.0 Theories The person performing this task must be able to	

	explain the following: 3.1 Big-data system architecture; 3.2 Data warehouse construction; 3.3 Data mining; 3.4 Machine learning. 4.0 Essential Skills 4.1 Proficiency in big data technologies and applications; 4.2 Familiarity with various big data processing technologies and tools; 4.3 Good communication skills and strong collective awareness; 4.4 Experience in education and training. 5.0 Digital Skills 5.1 Skills in algorithms and models; 5.2 Skills in pre-processing algorithms; 5.3 Optimization algorithms; 5.4 Network design; 5.5 Big data storage; 5.6 Scripting.
CIRCUMSTANTIAL KNOWLEDGE	Detailed knowledge about: 1. Proficiency in technologies and applications in the big data field, and experience in big data system architecture, data warehouse construction, data mining, machine learning, etc.; 2. Familiarity with various big data processing technologies and tools, including but not limited to Hadoop, Spark, Flink, Hive and Pig; 3. Extensive team management experience, specifically in personnel recruitment, team building, performance management, decision-making management, etc.; 4. Good communication skills, strong collective awareness, and the ability to collaborate and communicate effectively with different team members; 5. Experience in education and training, and the ability to develop and implement training programs and to effectively improve the specialized skills and knowledge of team members.

OCCUPATION	BIG-DATA ENGINEER	OCCUPATION CODE	
DUTY TITLE	DATA LIFECYCLE MANAGEMENT	DUTY NO.	806
TASK TITLE	DEVELOPMENT OF DATA LIFECYCLE MANAGEMENT STANDARDS AND DATA AFFINITY MANAGEMENT STANDARDS	TASK NO.	8062
PERFORMANCE CRITERIA	The person performing this task is responsible for developing the company's data lifecycle management standards and data affinity management standards to ensure that data is adequately managed and protected throughout its lifecycle.		
RANGE STATEMENT	The task can be performed within an enterprise under the supervision of a big data director. The tools and equipment to be used include: 1. Data warehouse and database management system; 2. Storage devices; 3. Data backup and recovery tools; 4. Data classification and archiving tools; 5. Data encryption and security tools; 6. Data destruction tools; 7. Data monitoring and management tools.		
EVIDENCE REQUIREMENT			
PRACTICAL PERFORMANCE		UNDERPINNING KNOWLEDGE	
The person performing this task must be able to do the following: 1. Study and understand best practices in data lifecycle management and data lineage management to develop the most appropriate standards. Manage, control and optimize data to ensure data reliability, security and consistency, and reasonably retain, archive and destroy data depending on different data types and application scenarios, from data creation to data destruction, including all phases of data acquisition, storage, processing, analysis, sharing and protection; 2. Work with relevant departments to collect and analyze the data needs and usage within the company to determine appropriate strategies for data lifecycle management and data lineage management; 3. Design, develop and implement data lifecycle management standards and data lineage management standards and ensure that those standards meet the		Detailed knowledge about: 1.0 Methods The person performing this task must be able to explain how to: 1.1 Make a clear description of the objectives of data lifecycle management; 1.2 Explain the main processes and different stages of data lifecycle management; 1.3 Present the best practices in data lifecycle management; 1.4 Explain the importance of data affinity management standards, including ensuring the accuracy, traceability, and stability of data relationships, to support data analysis and decision-making. 2.0 Principles The person performing this task must be able to explain the following principles: 2.1 Best practices in data lifecycle management; 2.2 Architecture and components of big data systems; 2.3 Significance and objectives of the data affinity management standards.	

company's needs and the requirements of regulations; 4. Provide training and support to ensure that the company's employees understand and comply with the data lifecycle management standards and the data lineage management standards; 5. Monitor and evaluate the effectiveness of the implementation of the data lifecycle management standards and the data lineage management standards, and make adjustments and improvements in time.	3.0 Theories The person performing this task must be able to explain the following: 3.1 Data management tools and technologies for data storage, backup, recovery, security protection, etc.; 3.2 Data management regulations and standards such as ISO27001 Information Security Management System certification; 3.3 Communication and coordination skills; 3.4 Organizational and project management skills. 4.0 Essential Skills 4.1 Communication skills; 4.2 Computer network skills; 4.3 Document editing skills; 4.4 Teamwork skills; 4.5 Linux system commands. 5.0 Digital Skills 5.1 Skills in algorithms and models; 5.2 Skills in pre-processing algorithms; 5.3 Optimization algorithms; 5.4 Network design; 5.5 Big data storage; 5.6 Scripting.
CIRCUMSTANTIAL KNOWLEDGE	Detailed knowledge about: 1. Familiarity with the concepts and principles of data lifecycle management and data affinity management, and understanding of best practices; 2. Good data analysis skills and the ability to collect, analyze and evaluate the needs and effectiveness of data lifecycle management and data affinity management; 3. Familiarity with data management tools and technologies such as data storage, backup, recovery and security protection; 4. Familiarity with data management regulations and standards, such as ISO27001 Information Security Management System certification; 5. Good communication and coordination skills and the ability to work with all levels of management and technical staff to promote the implementation and improvement of the standards;

	6. Good organizational and project management skills, and the ability to deliver on time, manage multiple projects and handle multiple tasks.
--	---

OCCUPATION	BIG-DATA ENGINEER	OCCUPATION CODE	
DUTY TITLE	DATA ASSET MANAGEMENT	DUTY NO.	806
TASK TITLE	DEVELOPMENT OF DATA ASSET MANAGEMENT RULES AND MAINTENANCE OF DATA ASSETS	TASK NO.	8063
PERFORMANCE CRITERIA	The person performing this task is responsible for developing data asset management rules, registering warehoused data asset information, and performing data asset maintenance.		
RANGE STATEMENT	The task can be performed within an enterprise under the supervision of a big data director. The tools and equipment to be used include: 1. Data warehouse and database management system; 2. Data quality tool; 3. Data classification and management tools; 4. Data cost analysis tool; 5. Data monitoring and reporting tools.		
EVIDENCE REQUIREMENT			
PRACTICAL PERFORMANCE		UNDERPINNING KNOWLEDGE	
The person performing this task must be able to do the following: 1. Develop data asset management rules: develop data asset management rules for data classification, security, storage, etc., according to enterprise needs and relevant laws and regulations; 2. Register warehoused data asset information: be responsible for collecting, sorting out, registering and warehousing data asset information, including information on data sources, types, volumes, quality, etc.; 3. Perform data asset maintenance: be responsible for the daily maintenance of data assets, including data backup, cleaning, archiving, etc. to ensure data security and integrity; 4. Research and analyze best practices for data governance and cost optimization and provide solutions.		Detailed knowledge about: 1.0 Methods The person performing this task must be able to explain how to: 1.1 Make a clear description of the objectives of data asset management rules; 1.2 Explain methods of data classification and standardization, including the classification methods based on business requirements and data attributes, to ensure proper management and use of data assets; 1.3 Explain methods of data asset monitoring and tracking; 1.4 Explain measures for data security and privacy protection, including data encryption, access control, authentication, and authority management. 2.0 Principles The person performing this task must be able to explain the following principles: 2.1 Concepts of data classification, security and cost; 2.2 Principles of data storage; 2.3 Principles of data fusion; 2.4 Methods and tools for data analysis; 2.5 Principles of data evaluation and cleaning.	

	3.0 Theories The person performing this task must be able to explain the following: 3.1 Knowledge of data management; 3.2 Ability to use data management tools; 3.3 Data analysis competence; 3.4 Teamwork skills; 3.5 Communication skills; 3.6 Awareness of continuous learning and self-improvement. 4.0 Essential Skills 4.1 Communication skills; 4.2 Computer network skills; 4.3 Document editing skills; 4.4 Teamwork skills; 4.5 Linux system commands. 5.0 Digital Skills 5.1 Skills in algorithms and models; 5.2 Skills in pre-processing algorithms; 5.3 Optimization algorithms; 5.4 Network design; 5.5 Big data storage; 5.6 Scripting.
CIRCUMSTANTIAL KNOWLEDGE	Detailed knowledge about: 1. Familiarity with data management rules: good knowledge of data management, including data classification, cost, security and reliability, storage, fusion, etc., as well as relevant laws and regulations; 2. Proficiency in data management tools: proficiency in data management tools, such as data management software and database systems, and ability to perform data management tasks independently; 3. Data analysis competence: be familiar with data analysis methods and tools and be able to analyze, evaluate and clean data; 4. Collective awareness: Be able to work collaboratively with team members, and have good communication skills and strong collective awareness; 5. Continuous learning skills: Demonstrate the ability to learn and awareness of self-improvement and be able to continually learn new technologies and knowledge and

	enhance self-competence.
--	--------------------------

OCCUPATION	BIG-DATA ENGINEER	OCCUPATION CODE	
DUTY TITLE	DEVELOPMENT OF A MANAGEMENT SYSTEM FOR DATA STANDARDS	DUTY NO.	806
TASK TITLE	DEVELOPMENT OF DATA STANDARD MANAGEMENT SYSTEMS AND IMPROVEMENT OF CORPORATE DATA GOVERNANCE	TASK NO.	8064
PERFORMANCE CRITERIA	The person performing this task must be able to be responsible for developing a data standard management system to improve corporate data governance.		
RANGE STATEMENT	The task can be performed within an enterprise under the supervision of a big data director. The tools and equipment to be used include: 1. Computer and office software; 2. Conference room; 3. Project management tools; 4. Flowchart software.		
EVIDENCE REQUIREMENT			
PRACTICAL PERFORMANCE		UNDERPINNING KNOWLEDGE	
The person performing this task must be able to do the following: 1. Study and analyze the types, sources, storage, sharing and use of the data on the platform, and develop an appropriate data standard management system from the perspectives of development specifications, indicator management, data cost, quality and security, etc.; 2. Design and implement the data standard management procedures, including formulation, issuance, examination and approval, updating and archiving of standards, to ensure timely updating and effective implementation of the data standards; 3. Instruct the enterprise to strictly follow the data standard management system in big data governance tasks, including data collection, pre-processing, analysis and visualization, data standard management, data quality control, data security management, data fusion and sharing management, etc.; 4. Instruct and support departments to follow the data standard management system in practical work, and assist in addressing issues and disputes relating		Detailed knowledge about: 1.0 Methods The person performing this task must be able to explain how to: 1.1 Make a clear description of the objectives of the data standard management system; 1.2 Explain the definition of data standards, including naming, formats, classification, and quality requirements of data, and classify data by importance and sensitivity for better management and protection; 1.3 Explain the process of developing and reviewing data standards; 1.4 Explain the methods of implementing and monitoring data standards. 2.0 Principles The person performing this task must be able to explain the following principles: 2.1 Update and change management of data standards; 2.2 Training and communication on data standards; 2.3 Methods of implementing and monitoring data standards. 3.0 Theories The person performing this task must be able to explain the following:	

to data standards, to improve the data quality and governance of the platform.	3.1 Characteristics and usage scenarios of various data types; 3.2 Development and management procedures for data standards; 3.3 Laws, regulations and policies on data standards. 4.0 Essential Skills 4.1 Project management competence; 4.2 Legal literacy and risk awareness; 4.3 Good communication skills and strong collective awareness. 5.0 Digital Skills 5.1 Skills in algorithms and models; 5.2 Skills in pre-processing algorithms; 5.3 Optimization algorithms; 5.4 Network design; 5.5 Big data storage; 5.6 Scripting.
CIRCUMSTANTIAL KNOWLEDGE	Detailed knowledge about: 1. Familiarity with national, industry and corporate standards and business department data; Understanding of the principles of data management and the basic theory of data standard development; 2. Familiarity with the characteristics and usage scenarios of various data types and understanding of technical methods and tools for data sourcing and storage; 3. Familiarity with data standard development and management procedures, and strong project management skills; 4. Familiarity with laws, regulations and policies on data standards, as well as certain legal literacy and risk consciousness; 5. Good communication skills and strong collective awareness, and ability to communicate and coordinate with various departments to promote the implementation of the data standard management system.

OCCUPATION	BIG-DATA ENGINEER	OCCUPATION CODE	
DUTY TITLE	DATA SECURITY MANAGEMENT AND PRIVACY PROTECTION	DUTY NO.	806
TASK TITLE	DATA SECURITY MANAGEMENT AND PRIVACY PROTECTION	TASK NO.	8065
PERFORMANCE CRITERIA	The person performing this task must be able to be responsible for managing, maintaining, and protecting the data security and user privacy in big data systems.		
RANGE STATEMENT	The task can be performed within an enterprise under the supervision of a big data director. The tools and equipment to be used include: 1. Data warehouse and database management system; 2. Encryption software 3. Virtual private network (VPN); 4. Data authority management tools; 5. Security gateway.		
EVIDENCE REQUIREMENT			
PRACTICAL PERFORMANCE		UNDERPINNING KNOWLEDGE	
The person performing this task must be able to do the following: 1. Design and implement big data security strategies and security management standards. Establish security audit requirements to ensure that data activities and related operations comply with security requirements; 2. Clarify data authorities based on business requirements and regulate the scope of authorities for different users on different data; 3. Be responsible for security vulnerability analysis, assessment and risk control in big data systems; 4. Ensure that data security and privacy protection policies comply with legal and industry regulations; 5. Coordinate and communicate with various departments and personnel involved in data security and privacy protection; 6. Identify and analyze data security and privacy protection issues, and propose solutions and implement them; 7. Understand the concepts of big data security management and privacy protection, and their significance in big data governance. Have knowledge of		Detailed knowledge about: 1.0 Methods The person performing this task must be able to explain how to: 1.1 Make a clear description of the requirements and objectives of data security; 1.2 Explain the relevant laws, regulations and policies applicable to data protection and fully understand and comply with the relevant requirements; 1.3 Explain the principles of data classification and authorization of access on demand, including clear partitioning of sensitive data and detailed management of access permissions for different users or roles; 1.4 Identify and explain possible security vulnerabilities and potential risks, and perform risk assessments and security threat modeling; 1.5 Explain the methods of developing and implementing data backup and disaster recovery strategies. 2.0 Principles The person performing this task must be able to explain the following principles: 2.1 Theory of data privacy; 2.2 Theory of authentication and authorization; 2.3 Cryptological theory;	

laws, regulations, policies and standards for big data security and privacy protection.	2.4 Computer network security theory; 2.5 Privacy computing and federated learning. 3.0 Theories The person performing this task must be able to explain the following: 3.1 Big data storage and processing technologies; 3.2 Data security and privacy protection standards and technologies; 3.3 Security setup and management of operating systems and database systems; 3.4 Big data security protection technologies; 3.5 Competence in risk assessment and risk control; 3.6 Good communication, coordination and teamwork skills. 4.0 Essential Skills 4.1 Communication skills; 4.2 Computer network skills; 4.3 Document editing skills; 4.4 Teamwork skills; 4.5 Linux system commands. 5.0 Digital Skills 5.1 Skills in algorithms and models; 5.2 Skills in pre-processing algorithms; 5.3 Optimization algorithms; 5.4 Network design; 5.5 Big data storage; 5.6 Scripting.
CIRCUMSTANTIAL KNOWLEDGE	Detailed knowledge about: 1. Familiarity with big data storage and processing technologies, including but not limited to Hadoop, Spark, etc.; 2. Familiarity with common data security and privacy protection standards and technologies, including but not limited to AES, SHA, SSL, IPSec, etc.; 3. Familiarity with the security setup and management of mainstream operating systems and database systems; Familiarity with big data security protection technologies, including anonymous protection for data release and social networks, data traceability technologies, access control technologies, etc.; Familiarity with big data privacy protection technologies, including those for threat detection, big data

	authentication, data authenticity analysis, data distortion processing, data encryption, restricted release, etc.; 4. Familiarity with advanced technologies such as privacy computing and federated learning; 5. Strong risk assessment and risk control capabilities, and good communication, coordination and teamwork skills.
--	--