

Revised Application Documentation: Version 4 /22 April, 2015

QUALIFICATION FILE – CONTACT DETAILS OF SUBMITTING BODY

Name and address of submitting body:

NIELIT Chandigarh,

Plot No. C-134, Phase VIII, Industrial Area, Sector 72, Mohali. 160071.

Name and contact details of individual dealing with the submission

Name: Dr. Manish Arora

Position in the organisation Joint Director (Systems)

Address if different from above NA

Tel number(s) 0172-2236470, 7087235351

E-mail address manisharora@nielit.gov.in

List of documents submitted in support of the Qualifications File

1. Detailed Curriculum
2. Industry Validation (Attached at Annexure I)
3. Placement Records (Attached at Annexure II)

QUALIFICATION FILE SUMMARY

Qualification Title	Certified Data Scientist
Body/bodies which will assess candidates	Examination Cell, National Institute of Electronics and Information Technology, 6-CGO Complex, Electronics Niketan, Lodhi Road, New Delhi. 110003.
Body/bodies which will award the certificate for the qualification.	National Institute of Electronics and Information Technology 6-CGO Complex, Electronics Niketan, Lodhi Road, New Delhi. 110003.
Body which will accredit providers to offer the qualification.	National Institute of Electronics and Information Technology 6-CGO Complex, Electronics Niketan, Lodhi Road, New Delhi. 110003. Presently, Accreditation is not prescribed; affiliation is one of the models.
Occupation(s) to which the qualification gives access	Data Scientist
Proposed level of the qualification in the NSQF.	6
Notional Learning Hours	240 hours.
Entry requirements / recommendations.	Graduate/Undergraduate and knowledge of any computer language, Internet Concepts and Database.
Progression from the qualification.	<u>Data Scientist – Next Job Role – Next Job Role</u> 1) <u>In Academic</u> After completion of this course, students can go for course that integrates Cloud Computing and Big Data analysis. Many companies are providing Big Data as a Service solution to individuals and organization for analysing large data using Big Data tools over Cloud Computing infrastructure. 2) <u>Professional</u> Initially, candidate can work as Data Scientist/Programmer whose role would be to identify analytical requirement and provide solution. Later on, they can be Big Data Solutions Architect. Big Data Solutions Architect is a very skilled architect with cross-industry and cross-functional know-how
Planned arrangements for RPL.	<ul style="list-style-type: none"> • Presently only candidates who undergo training shall be assessed. • It will be incorporated once RPL strategy is finalized

Formal structure of the qualification

Title of unit or other component (include any identification code used)	Mandatory/ Optional	Estimated size (learning hours)	Level
Configure Deployment Platform	Mandatory	15	6
Analyze and Define Business Requirements	Mandatory	15	
Design and Develop Presentation Layer	Mandatory	45	
Analyze Big Data in Cluster Environment	Mandatory	30	
Analyze Data using Big Data Analytic Tools	Mandatory	60	
Manage Real World Data Analytic Application	Mandatory	60	
Enhancing Communication Skill	Mandatory	15	

Please attach any document giving further detail about the structure of the qualification – e.g. a Curriculum or Qualification Pack. Detailed Curriculum attached at Annexure III.

SECTION 1 **ASSESSMENT**

Name of assessment body:

Examination Cell,

National Institute of Electronics and Information Technology
6-CGO Complex, Electronics Niketan,
Lodhi Road, New Delhi. 110003.

Will the assessment body be responsible for RPL assessment?

Give details of how RPL assessment for the qualification will be carried out and quality assured.

Presently, only candidates undergoing training shall be assessed. Later on, candidates having experience and knowledge shall be assessed. The information will be provided on finalization of such procedure.

Describe the overall assessment strategy and specific arrangements which have been put in place to ensure that assessment is always valid, consistent and fair and show that these are in line with the requirements of the NSQF:

The emphasis is on practical demonstration of skills & knowledge based on the performance criteria. Each OUTCOME is assessed & marked separately. Student is required to pass in all OUTCOMES individually and marks are allotted. Following assessment methodologies are used.

- A. Written Assessment (Multiple Choice Questions)
- B. Practical Assessment
- C. Viva Voce Assessment

The assessment results are backed by following evidences.

- 1 The assessor collects a copy of the attendance for the training done under the scheme. The attendance sheets are signed and stamped by the In charge / Head of the Training Centre.
- 2 The assessor verifies the authenticity of the candidate by checking the photo ID card issued by the institute as well as any one Photo ID card issued by the Central/Government. The same is mentioned in the attendance sheet.
- 3 The assessor assigns roll number.
- 4 The assessor takes photograph of all the students along with the assessor standing in the middle and with the centre name/banner at the back as evidence.

Please attach any documents giving further information about assessment and/or RPL.

ASSESSMENT EVIDENCE

Complete the following grid for each grouping of NOS, assessment unit or other component as listed in the entry on the structure of the qualification on page 1.

Job Role

Data Scientist

Title of Unit/Component:

(Detailed Curriculum attached As Annexure-III)

Assessable Outcomes	Assessment criteria for the outcome	Total Mark	Written	Practical	Vivo-voce
1. Configure Deployment Platform	Preparation of platform for analyzing big data	75	5	5	5
	Acquiring skills to interact with prepared platform		10	10	10
	Acquiring skills to secure files by managing users, groups and their privilege		10	10	10
	Total		25	25	25
2 Analyze and Define Business Requirements	Selection of database based on Requirements	100	5	5	0
	Acquiring the skills on designing database		5	5	0
	Acquiring the advanced skills on database designing		5	5	5
	Acquiring the skills to manipulate data		5	5	5
	Acquiring the advanced skills to manipulate data		10	10	5
	Acquiring the skills to manage large scale data warehouse and eliciting hidden information		10	10	5
Total	Total	40	40	20	

3 Design and Develop Presentation Layer	Acquiring fundamental software developing skills	150	10	10	5
	Acquiring skills on handling unusual situations at runtime		10	10	5
	Acquiring skills on development software with latest practices		10	10	5
	Acquiring skills on architecture of front-end application		10	10	5
	Acquiring skills on developing front-end application		10	10	5
	Attaining skills on integrating application with backend database		10	10	5
	Total		60	60	30
4 Analyze Big Data in Cluster Environment	Acquiring skills on platform preparation for managing big data	150	5	5	5
	Acquiring skills on platform preparation for managing big data in grid		10	10	5
	Acquiring skills on interacting with big data file system		10	10	10
	Acquiring skills on analyzing data using conventional style of programming		10	10	10
	Acquiring advance skills on analyzing data using conventional style of programming		20	20	10
	Total		55	55	40
5 Analyze Data using Big Data Analytic Tools	Use of data warehouse facility for analyzing big data	200	5	5	5
	Use of Programming language to Analyze big data stored at data warehouse		10	10	10
	Use of column database for analyzing big data		10	10	5
	Use of programming language to analyze stored in column database		10	10	5
	Use of high level tool to analyze big data		10	10	5
	Use of programming language to analyze big data stored in high level tool		10	10	5
	Use of Big Data Analytic tool to analyze semi-structured/unstructured data		10	10	10
	Use of programming language to analyse Semi Structured/ unstructured data		10	10	5
	Total		75	75	50
6 Manage Real World Data Analytic Application	Identify big data Requirements	325			25
	Document big data requirements				25
	Design big data application				100
	Develop big data application				50
	Test big data application				50
	Steps to implement the developed application				50
	Demonstrate big data application				25

		Total			325
7 Enhancing Communication Skill	Acquiring Communication Skill				10
	Managing career, staff and professional relationships				20
	Preparing for interview				20
		Total			50
	Grand Total	1050	255	255	540

Means of assessment 1

Proctored online assessments (LAN and Web based), carried out using a variety of question formats applicable for linear / adaptive methodologies; performance criteria being assessed via situation judgement tests, simulations, code writing, psychometrics and multiple choice questions etc.

SECTION 2

EVIDENCE OF NEED

What evidence is there that the qualification is needed?

Big Data is a term used to describe the process of collecting, organizing, and analyzing large sets of data to discover hidden patterns, unknown correlations, and other useful information. Big Data helps you understand the information contained within the data and helps identify that which is most important for future business decisions. Big data has become an essential factor for the success of business in various verticals. Many developing countries such as China, Brazil, and France are also focusing on adopting this new technology.

The major driving force of this technology is the rising unstructured data from several sources and the constant need of enterprises to optimize large workloads of data to enhance the overall efficiency of system. However, there are few restraints in the advancement of this technology such as lack of skilled personnel and lack of security measures and solutions. However, with the extensive efforts of large vendors in the ecosystem, they have established advanced training and learning centers, and have launched various certification courses to cater to these issues.

(Source: <http://www.micromarketmonitor.com/market-report/big-data-reports-4483038255.html?gclid=CLnf6JvS3coCFVAXaAodtboF6w>)

The global big data market is expected to grow to \$46.34 billion by 2018 at an estimated CAGR of 25.5%, during the forecast period. Rapid increase in data generation from different industry verticals is one of the major factors driving this market.

The demand for Analytics skill is going up steadily but there is a huge deficit on the supply side. This is happening globally and is not restricted to any part of geography. In spite of Big Data Analytics being a 'Hot' job, there is still a large

number of unfilled jobs across the globe due to shortage of required skill. A McKinsey Global Institute study states that the US will face a shortage of about 190,000 data scientists and 1.5 million managers and analysts who can understand and make decisions using Big Data by 2018.

Currently, India has the highest concentration of analytics globally. In spite of this, the scarcity of data analytics talent is particularly acute and demand for talent is expected to be on the higher side as more global organizations are outsourcing their work.

According to Srikanth Velamakanni, co-founder and CEO of Fractal Analytics, there are two types of talent deficits: Data Scientists, who can perform analytics and Analytics Consultant, who can understand and use data. The talent supply for these job titles, especially Data Scientists is extremely scarce and the demand is huge.

(Source: <http://www.edureka.co/blog/10-reasons-why-big-data-analytics-is-the-best-career-move>)

What is the estimated uptake of this qualification and what is the basis of this estimate?

Student uptake from Industry

What steps were taken to ensure that the qualification(s) does/do not duplicate already existing or planned qualifications in the NSQF?

As the understanding and adoption models of QPs evolve in the industry and across its sub-sectors, we foresee consolidation of qualification packs as a natural progression. The Qualification does not exist as per information available in public domain.

What arrangements are in place to monitor and review the qualification(s)? What data will be used and at what point will the qualification(s) be revised or updated?

The Qualification is to be monitored and reviewed every two years.

The following data will be used

1. Results of assessments
2. Employer feedback will be sought post-placement
3. Student feedbacks
4. Workshops and seminar for reviewing the qualifications
5. Industry Requirements
6. Consultation/ Tie-up with Industries or Expert for review of the Curriculum.

Please attach any documents giving further information about any of the topics above.
NIL

SECTION 3

SUMMARY EVIDENCE OF LEVEL

Level of qualification: 6

Summary of Direct Evidence:

Justify the NSQF level allocated to the QP by building upon the five descriptors of NSQF. Explain the reasons for allocating the level to the QP.

Generic NOS is/are linked to the overall authority attached to the job role.

Data Scientist					
Process required	Professional knowledge	Professional skill	Core skill	Responsibility	Level
<p>Data Scientists carries out the job to identify requirements of business analyzes which are helpful in making business decisions</p> <p>Data Scientists acquire wide range of theoretical practical skills to provide analytic solution to business analytic problem</p> <p>Their job is to prepare abstract model based on requirement to propose solution</p>	<p>After acquiring professional knowledge on Big Data tools and Techniques, the Data Scientist will be competent to identify technical requirements in terms of hardware, software and other IT related devices.</p> <p>They can prepare detailed design of the proposed solution for Big Data Analytics</p>	<p>They are proficient in developing solution based on detailed design and practical knowledge gained during course</p> <p>They plan tests, prepare tests cases, generate test data and perform testing on test data</p>	<p>Data Scientist after acquiring skills both managerial and technical of this level are able to interact with different stakeholders involved like vendors, clients and users.</p> <p>They are able to make independent decision involved in providing solution.</p>	<p>They are able to lead team as well as work in team.</p>	6
<u>7</u>	<u>5</u>	6	9	6	

SECTION 4

EVIDENCE OF RECOGNITION OR PROGRESSION

What steps have been taken in the design of this or other qualifications to ensure that there is a clear path to other qualifications in this sector?

This qualification comprises both technical and analytic skills and can be linked to any qualification higher than this one, existing or to come.

Cloud providers are now started to provide services for Big data analytics. Big Data as a Service (BDaaS) like traditional cloud services, IaaS, PaaS and SaaS, is next paradigm shift in analytics. Amazon Web Service's Elastic MapReduce (EMR) is the most prominent core BDaaS available. NIELIT has recently signed MoU with Amazon for imparting training on AWS (Amazon, Web Services)

Similarly, Altiscale named one of top 5 Big Data Cloud providers provides solution for BDaaS.

Please attach any documents giving further information about any of the topics above.

Sources:

<https://aws.amazon.com/elasticmapreduce/>

<https://www.altiscale.com/>

SECTION 5

EVIDENCE OF INTERNATIONAL COMPARABILITY

List any comparisons which have been established.

Big Data University is an IBM initiative to spread big data literacy. The university offers course on 'Data Scientist Fundament'

(Source: <http://bigdatauniversity.com/>)

IBM offers three courses namely :

- IBM Certified Solution Advisor - Big Data & Analytics V1
- IBM Certified Data Architect - Big Data
- IBM Certified Data Engineer - Big Data

(Source: http://www-03.ibm.com/certify/certs/bigdata_index.shtml)

Industry Validation

The global big data market is expected to grow to \$46.34 billion by 2018 at an estimated CAGR of 25.5%, during the forecast period. Rapid increase in data generation from different industry verticals is one of the major factors driving this market.

The demand for Analytics skill is going up steadily but there is a huge deficit on the supply side. This is happening globally and is not restricted to any part of geography. In spite of Big Data Analytics being a 'Hot' job, there is still a large number of unfilled jobs across the globe due to shortage of required skill. A McKinsey Global Institute study states that the US will face a shortage of about 190,000 data scientists and 1.5 million managers and analysts who can understand and make decisions using Big Data by 2018.

India currently has the highest concentration of analytics globally. In spite of this, the scarcity of data analytics talent is particularly acute and demand for talent is expected to be on the higher side as more global organizations are outsourcing their work.

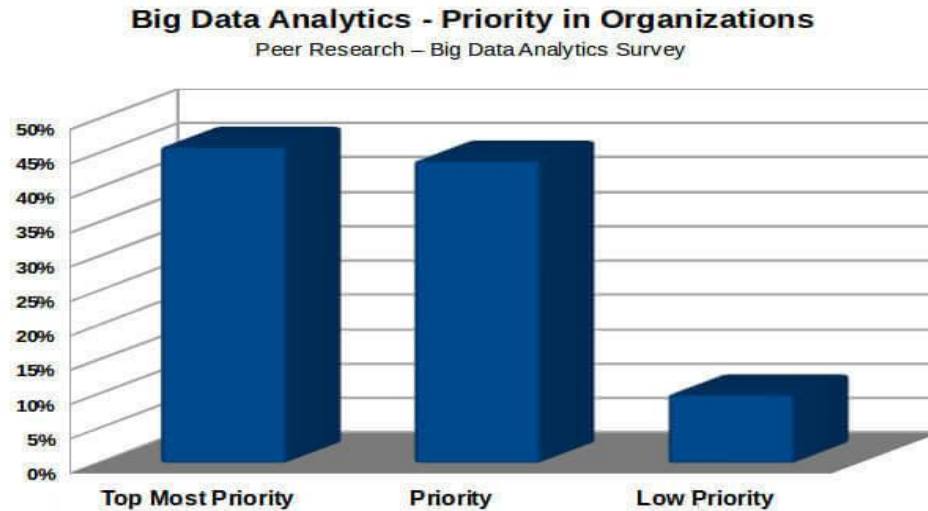
According to Srikanth Velamakanni, co-founder and CEO of Fractal Analytics, there are two types of talent deficits: Data Scientists, who can perform analytics and Analytics Consultant, who can understand and use data. The talent supply for these job titles, especially Data Scientists is extremely scarce and the demand is huge.

(Source: <http://www.edureka.co/blog/10-reasons-why-big-data-analytics-is-the-best-career-move>)

IBM, Cisco and Oracle together advertised 26,488 open positions that required big data expertise in 2015. DELL has 25.1% of all available big data positions that WANTED Analytics tracks.

(Source: <http://www.forbes.com/sites/louiscolombus/2015/11/16/where-big-data-jobs-will-be-in-2016/#4f8a1261f7f1>)

According to the ‘Peer Research – Big Data Analytics’ survey, it was concluded that Big Data Analytics is one of the top priorities of the organizations participating in the survey as they believe that it improves the performances of their organizations.



Based on the responses, it was found that approximately 45% of the surveyed believe that Big Data analytics will enable much more precise business insights, 38% are looking to use Analytics to recognize sales and market opportunities. More than 60% of the respondents are depending on Big Data Analytics to boost the organization’s social media marketing abilities. The QuinStreet research based on their survey also backs the fact that Analytics is the need of the hour, where 77% of the respondents consider Big Data Analytics a top priority.

(Source: <http://www.datasciencecentral.com/profiles/blogs/10-reasons-why-big-data-analytics-is-the-best-career-move>)

Placement Record

Following is summary of the batches conducted at NIELIT Chandigarh on Big Data

Sr. No.	Batch Code	Start Date	Duration	No of Students
1	ST322	1/6/2015	2 months	17
2	ST336	22/6/2015	2 months	7
3.	ST344	18/1/2016	2 months	4
4.	PJ095	18/1/2016	6 months	7
			Total	35

As per information, 3 students are placed, 3 are undergoing further study. The candidates are working in Outline Software Solution, Chandigarh (2) and High Tech ILS Cloud Solution Pvt. Ltd Chandigarh (1)

Annexure -III

Detailed Curriculum

Name of Unit of Qualification : Configure Deployment Platform
Duration : 15 Hours
Topics : Ubuntu

Outcome	Contents	Hrs.
Preparation of platform for analysing big data	Introduction to Virtual Machine, creating and configuring Virtual Machine, Installing Ubuntu Operating System on Virtual Machine	3
Acquiring skills to interact with prepared platform	Operating System Concepts: Linux History, Benefits of Linux, Different Flavors of Linux, Introducing Ubuntu, Installing Ubuntu: Starting Up, Logging in, Exploring the Desktop, Ubuntu Basics, Browsing the File System, Understanding File System Concept, Managing Files, Real and Virtual Files, Mounting, File Searches, File Size, File Space Understanding Linux Files/Directories: Viewing Text Files, Using a Command Line Text Editor, Creating Files, Searching through Files, Comparing Text Files, Copying, Moving, Managing Files. Ubuntu Commands, Running Basic commands, Piping and Filtering Commands, Directory and File handling commands	6
Acquiring skills to secure files by managing users, groups and their privilege	Users, Groups and Permissions, Root and Other Users, Adding and Deleting Uses and Groups, Adding and Changing Passwords, Users and File Permissions Managing and Handling Processes: Viewing Processes, Controlling Processes, Controlling Jobs, Background and Foreground Jobs. Scheduling Tasks, Installing and Removing Software	6

Name of Unit of Qualification : Analyse and Define Business Requirements
Duration : 30 Hours
Topics : Data Base Management Systems and MySQL/Oracle

Performance Criteria(OUTCOME) No.	Contents	Hrs.
Selection of database based on Requirements	Introduction to database, characteristics of data in database, DBMS, advantages of DBMS, file-oriented approach versus Database-oriented approach to Data Management, disadvantages of file-oriented approach. A brief overview of relational model. Definition of relation, properties of relational model, Concept of keys: candidate key, primary key, alternate key, Foreign key, Fundamental integrity rules: entity integrity, referential integrity.	3
Acquiring the skills on designing database	Entity-relationship model as a tool for conceptual design-entities, attributes and relationships, ER-diagram, Cardinality Ratio, strong and weak entities, converting an E-R model into relational schema, Examples of E-R Model	3
Acquiring the advanced skills on database designing	Normalization concepts in relational model, normal forms (1NF, 2NF, 3NF)	3
Acquiring the skills to manipulate data	SQL, characteristics & Advantages of SQL, SQL data types, SQL constructs: select-from-where, insert, delete, and update. SQL constructs: group by, having, order by.	9
Acquiring the advanced skills to manipulate data	Nested queries, joins, union, intersection, correlated nested queries, views and indexes. Practical hands on SQL statements using MySQL/Oracle database	6

Acquiring the skills to manage large scale data warehouse and eliciting hidden information	Knowledge Discovery in Databases, Data Mining, Data warehouse. Migrating data from source to data warehouse, cleaning, aggregation operations.	6
--	--	---

Name of Unit of Qualification : Design and Develop Presentation Layer
Duration : 60 Hours
Topics : Java Programming

Learning Outcome (NO)	Topics	Hours
Acquiring fundamental software developing skills	OOPS Principles, an Overview of Java Object-Oriented Programming, Data Types, Variables, and Arrays, Operators-Arithmetic Operators, The Bitwise Operators ,Relational Operators, Boolean ,Logical Operators, Programming Constructs, Methods and Inheritance	15 hrs
Acquiring skills on handling unusual situations at runtime	The basic Java I/O Classes and String Handling Exception-Handling Fundamentals, Exception Types ,Uncaught Exceptions , Using try and catch , Displaying a Description of an Exception ,Multiple catch Clauses , Nested try Statements , Throw throws finally Java's Built-in Exceptions	6 hrs
Acquiring skills on development software with latest practices	Packages, Access Protection, Importing Packages and Interfaces	15
Acquiring skills on architecture of front-end application	Applet Fundamentals Applet Architecture An Applet Skeleton Applet Initialization and Termination ,Simple Applet Display ,	6
Acquiring skills on developing front-end application	Controls: Labels , TextField and Buttons .Handling Buttons and TextField and basic Layout Managers	6
Attaining skills on integrating application with backend database	Java Database Connectivity JDBC-ODBC Bridge JDBC Drivers Creating DSN DriverManager, Connection, Statement, ResultSet ,ODBC Database URL Statement and it usage with Applet	12

Name of Unit of Qualification : Analyze Big Data in Cluster Environment
Duration : 30 Hours
Topics :Hadoop and Map Reduce Programming

Performance Criteria(OUTCOME) No.	Contents	Hrs.
Acquiring skills on platform preparation for managing big data	Big Data Concepts, Need for analyzing Big Data, its roles in Business Intelligence and decision making. Big Data, Hadoop Architecture, Hadoop ecosystem components, storage, Hadoop Distributed File System (HDFS), Single node installation.	6
Acquiring skills on platform preparation for managing big data in grid	Multi node installations. Cluster Architecture, Cluster configuration files Hadoop commands, Hadoop Server Role, name Node, secondary node, data node, file write and read.	6
Acquiring skills on interacting with big data file system	Shell commands, Accessing files on HDFS and local machine	3
Acquiring skills on analysing data using conventional style of programming	Map Reduce Framework, Examples, Developing Map Reduce Programs, structure of Map Reduce program, Map and Reduce Tasks, Life cycle methods, Data types, data loading, Map and Reduce Tasks, partitioner, combiners, input formats, output formats, Custom input format, Error Handling, Tuning Advance Map Reduce, Fair and Capacity Scheduler. Running Map Reduce on local and Distributed modes.	9
Acquiring advance skills on analysing data using conventional style of programming	Sorting, reverse sorting, secondary sorting, Compression Techniques, Working with sequential files,	6

Name of Unit of Qualification : Analyze Data using Big Data Analytic Tools
Duration : 60 Hours
Topics :Big Data Analysis using HBase, HIVE and PIG

Performance Criteria(OUTCOME) No.	Contents	Hrs.
Use of data warehouse facility for analysing big data	Introduction to HIVE, installing HIVE, Data types, HIVE shell, HIVE commands, Complex Data types,	9
Use of Programming language to Analyse big data stored at data warehouse	UDF in Hive using Java	6
Use of column database for analysing big data	HBASE introduction and installation in Ubuntu, integration with Hadoop, HBase Shell, HBase storage techniques, HBase-Java connectivity,	9
Use of programming language to analyse stored in column database	Writing programs in Java using HBase to handle big data	6
Use of high level tool to analyse big data	PIG installation and configuration in Ubuntu, data types, commands: group, filter, order, Distinct, Join, union.	9
Use of programming language to analyse big data stored in high level tool	UDF in Pig using Java	6
Use of Big Data Analytic tool to analyze semi-structured/unstructured data	JAQL data model, Jaql syntax, jaqlshell, core operators, input / output adapter, jaql build in function, jaql statement: assignment, explains, import, quit etc.	9
Use of programming language to analyse Semi Structured/ unstructured data	Embedding jaql in java	6

Name of Unit of Qualification : Manage Real World Data Analytic Application
Duration : 60 Hours
Topics :Project

Performance Criteria(OUTC OME) No.	Contents	Hrs.
Identify big data Requirements	Identification of requirements of analytics	9
Document big data requirements	Requirement Analysis Preparation	9
Design big data application	Design of Real World Data Analytic Application	9
Develop big data application	Develop Real World Data Analytic Application	18
Test big data application	Test Real World Data Analytic Application	9
Steps to implement the developed application	Implement Real World Data Analytic Application	3
Demonstrate big data application	Demonstration	3

Name of Unit of Qualification : Enhancing Communication Skill
Duration : 15 Hours
Topics :Soft Skills and Communication

Performance Criteria(OUTCOME) No.	Contents	Hrs.
Acquiring Communication Skill	Communication , verbal and non-verbal communication	5
Managing career, staff and professional relationships	Building professional relationship, Relationship at work , Making the most of personal and professional relationships, Competency Description, Managing Difficult Business Relationships	5
Preparing for interview	<p>Interview Techniques: Planning For The Interview, Preparing for an Interview, Interview Formats, Stages Of The Interview, Types Of Interview Questions</p> <p>Best Bet for Interview Preparation: Mock Interviews, The Benefits of Mock Interviews Experience & Skills,</p> <p>Curriculum Vitae:Overview, types of CV, Covering letter, Writing a Resume, Acceptance Letter, Thank You Letter</p>	5