CASE STUDIES WITH PYSPARK AND MACHINE LEARNING

Big data

- ❑ Big Data is a collection of data that is huge in volume, yet growing exponentially with time.
- ❑ There are 4 main V/s of big data
- ❑ Velocity ,Variety ,Volume and Veracity
- ❑ Apache Hadoop is an open source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data
- ❑ Hadoop framework is written in Java.
- ❑ Spark was built on the top of Hadoop MapReduce module
- ❑ It extends the MapReduce model to efficiently use more type of computations which include Interactive Queries and Stream Processing.
- ❑ Spark can interact in Java , Python or Scala languages.
- ❑ HDFS is a distributed file system that handles large data sets running on commodity hardware.
- ❑ The way HDFS works is by having a main NameNode and multiple data nodes on a commodity hardware cluster.
- ❑ All the nodes are usually organized within the same physical rack in the data center.
- ❑ Data is then broken down into separate blocks that are distributed among the various data nodes for storage

Why Pyspark?

- ❑ Python API for Apache spark
- ❑ Open source distributed computing framework
- ❑ Helpful for real time large scale data processing
- ❑ Utilize the spark framework in combination with python
- ❑ The key data type used in PySpark is the Spark dataframe.
- ❑ Pandas run operations on a single machine whereas PySpark runs on multiple machines.
- ❑ provides Py4j library, with the help of this library, Python can be easily integrated with Apache Spark.
- ❑ provides real-time computation on a large amount of data because it focuses on in-memory processing. It shows the low latency.
- ❑ provides powerful caching and good disk constancy.
- ❑ allows us to achieve a high data processing speed, which is about 100 times faster in memory and 10 times faster on the disk.

Case study 1
Predict the number of crew members for a ship company

- ❑ Algorithm Used : Linear Regression
- ❑ Assumes a linear relationship between input variable x and output variable y
- ❑ If there is only one input variable then it is called simple linear regression
- ❑ Features include name of ship, Cruise Line ,age ,tonnage , number of passengers , number of cabins etc.
- ❑ Label is no of crew members.
- ❑ **'Cruise Line'** is an important feature but categorical
- ❑ To convert it into numerical form String Indexer is imported
- ❑ Imported Vector Assembler to convert useful features into vector form
- ❑ Split the final data into training and testing set in 70 and 30 percent respectively.
- ❑ From pyspark.ml.regression imported Linear Regression
- ❑ Fitted the regression model on training data and evaluated on testing data.

Case study 1
Predict the number of crew members for a ship company

Evaluation metrics

- ❑ R Squared value : 0.8547
- ❑ Root mean square error : 1.422

Case study 2
Customer churn prediction for a marketing agency

- ❑ Algorithm used : Logistic Regression
- ❑ Features include name , age , total purchase , whether account manager assigned or not, location , company name , onboard date etc.
- ❑ Label is whether customer will churn or not. It can be 0 or 1
- ❑ Logistic regression is a classification problem.
- ❑ We are converting the useful features into a vector form using vector assembler and created a new data frame using the new feature.
- ❑ From the new data set we are performing training and testing in the ratio 0.7 and 0.3
- ❑ Logistic regression have been imported from pyspark.ml.classification
- ❑ We have imported binary classification evaluator from pyspark.ml.evaluation
- ❑ After creating model deployed it into a new dataset

Case study 2

Customer churn prediction for a marketing agency

Evaluation metrics

- ❑ Area Under Curve : 0.740

Case Study 3
Predicting whether a university is private or not

- ❑ Algorithm used : Random forest classifier ,Decision tree classifier , Gradient boost trees
- ❑ Importing vector assembler to group the necessary features needed
- ❑ The label column private is categorical we are importing String Indexer to convert them into numerical form
- ❑ From pyspark.ml.classification imported Decision Tree Classifier,Gradient Boost trees,random forest classifier
- ❑ Calling the imported classification algorithms where private index column is given as the label and features column is given as the feature
- ❑ Performing fitting on the training data and transforming on the testing data

Case Study 3

Predicting whether a university is private or not

Evaluation metrics

Area under the curve:

- ❖ Random Forest Classifier : 0.9913
- ❖ Gradient Boost Trees : 0.97
- ❖ Decision Tree Classifier : 0.922
- ❑ Accuracy when using random forest classifier : 0.9613733
- ❑ F1 score when using random forest classifier : 0.961167

Case Study 4

Analyzing dog food spoilage

- ❑ Algorithm used : Random forest classifier .
- ❑ The objective of the case study is to find out which preservative contribute to the most in food spoilage.
- ❑ Use vector assembler to convert the useful features into a vector form.
- ❑ Transforming the data using particular assembler.
- ❑ Random forest Classifier is imported from pyspark.ml.classification
- ❑ The classifier is called. The Spoiled column is assigned as label and the vector column named features is assigned as featuresCol.
- ❑ From the data frame we are selecting features column and Spoilage column.
- ❑ Then assigning them to new data frame created called final data.
- ❑ The next step is to fit the classifier into the final data.
- ❑ An attribute called featureImportance helps to evaluate the importance of each and every feature . Helps to predict which preservative is highly important and contribute the most in spoilage.

Case Study 5

Creating a movie recommendation engine using collaborative filtering

- ❑ Algorithm used : ALS algorithm which helps in collaborative filtering
- ❑ Alternative least square algorithm / ALS algorithm is a matrix factorization algorithm.
- ❑ Imported ALS from pyspark.ml.recommendation
- ❑ Imported Regression Evaluator from pyspark.ml.evaluation
- ❑ The data is split into 80 and 20 %
- ❑ Called the ALS algorithm.Assigned the value of maxIter , userCol, ratingCol and itemCol.
- ❑ Fitted the ALS into training data and created the model
- ❑ Transformed the testing data using the model created
- ❑ Regression evaluator is called to calculate the rmse value
- ❑ Deployed the model by selecting the movieId of a particular user.
- ❑ Evaluated by how much percentage the user may like the particular movie

Case Study 5

Creating a movie recommendation engine using collaborative filtering

Evaluation metrics

Root mean square error : 1.788

Case study 6

Analysis of patterns in hacking

- ❑ Algorithm used : Kmeans
- ❑ Vector assembler is imported from pyspark.ml.feature
- ❑ We are transforming the data using particular vector assembler
- ❑ Standard Scaler is imported from pyspark.ml.feature
- ❑ In Machine Learning, StandardScaler is used to resize the distribution of values so that the mean of the observed values is 0 and the standard deviation is 1.
- ❑ We are fitting the standard Scaler into the final data and creating the model
- ❑ Then transforming the final data using the model created
- ❑ The new data frame consists of column called features and scaled features
- ❑ The next phase is to call the K Means algorithm by specifying values of k as 2 and 3 respectively.
- ❑ Analyzing the difference in output for various k values

Case study 7
Creating spam filter Natural Language Processing techniques

- ❑ The data set consists of a feature called message and a label called class
- ❑ The class label shows whether the message is spam or not
- ❑ We have imported length from pyspark.sql.funcions
- ❑ The next phase is to import Tokenizer,Stop words remover,Count Vectorizer,IDF,String Indexer from pyspark.ml.feature
- ❑ Vector assembler is imported to convert the useful input features to an output column called features
- ❑ Naïve Bayes algorithm is imported from pyspark.ml.classification
- ❑ Since various stages are involved Pipeline is imported
- ❑ Added the various stages involving tokenizing ,Stop words removing , count vectorizer , IDF, String indexer and vector assembler to the pipeline.
- ❑ We are fitting the pipeline to the data
- ❑ Then transforming the data using the pipeline model
- ❑ Next step is to select the label and features from data and assigning them to a new data frame called cleaner data
- ❑ Then splitting the cleaner data into training and testing data set
- ❑ Fitting the naïve Bayes classifier to training data and creating a model
- ❑ Transforming the testing data using the model created

Case study 7

Creating spam filter Natural Language Processing techniques

Evaluation Metrics

- ❑ Accuracy : 0.929

About the organization

- ❑ Cognizant is an American multinational information technology services and consulting company.
- ❑ It is headquartered in Teaneck, New Jersey, United States.
- ❑ Cognizant is part of the NASDAQ-100 and trades under CTSH.
- ❑ Like many other IT services firms, Cognizant follows a global delivery model based on offshore software R&D and offshore outsourcing.
- ❑ Cognizant is organized into several verticals and horizontal units.
- ❑ The vertical units focus on specific industries such as Banking & Financial Services, Insurance, Healthcare, Manufacturing and Retail.
- ❑ The horizontals focus on specific technologies or process areas such as Analytics, mobile computing, BPO and Testing.

About the internship

- ❑ Cognizant internship is a pre requisite skill development program which makes the selected candidates ready for employment .
- ❑ The internship program consists of learning curriculum as per the learning track assigned.
- ❑ The learning path will include in-depth sessions, hands on exercise and project work.
- ❑ There will also be a series of webinars, quizzes, mentor connects, leader connect ,code challenges, assessments etc. to accelerate learning.

- ❑ The outcomes during the Internship would be monitored through formal evaluations.

Internship Structure

Period of 4-6 months

Designation:

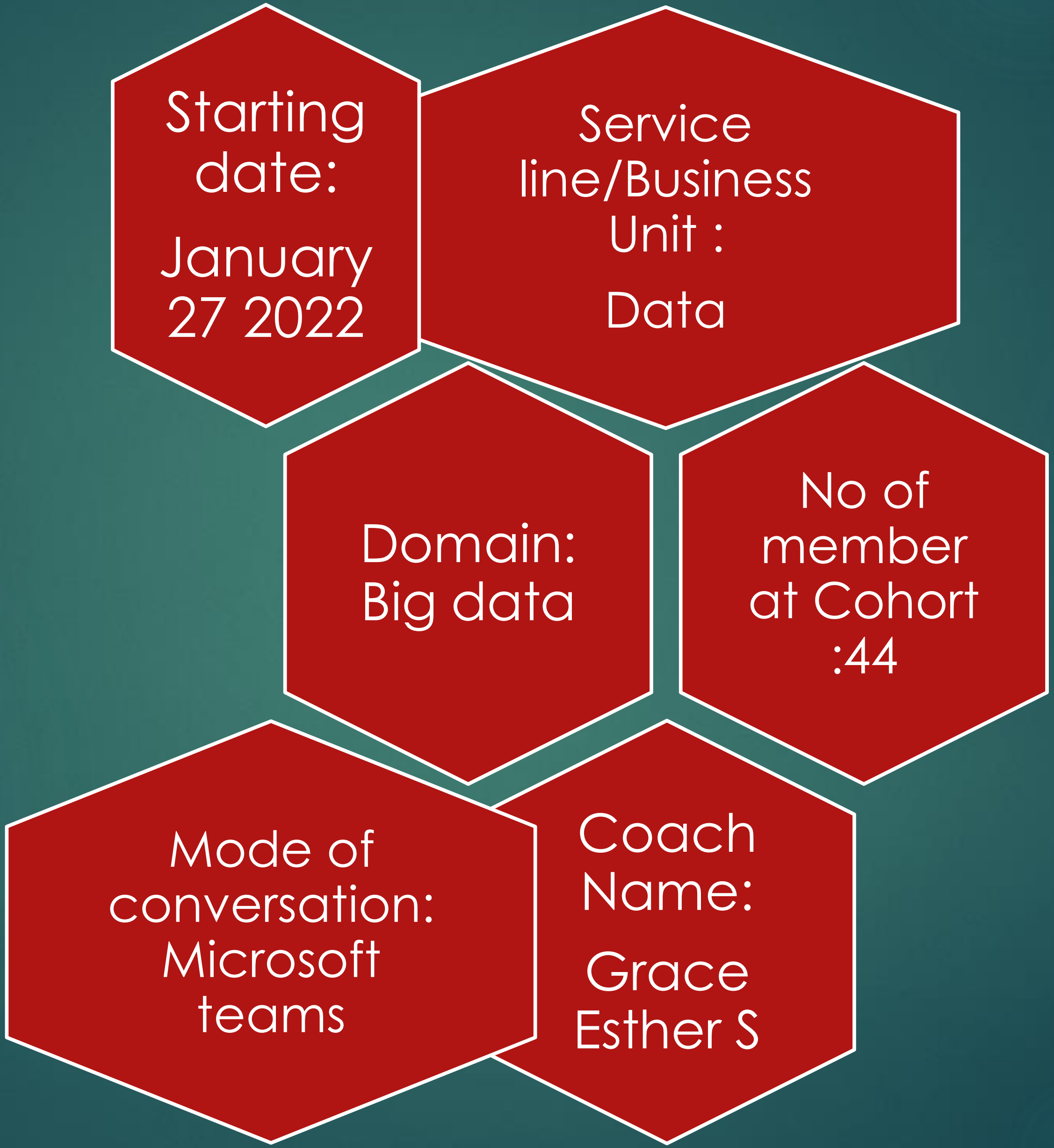
Programmer Analyst Trainee

Working hours

9-7

Each candidate will be mapped to batches where they further get divided into cohorts

My internship overview



Starting date:

January 27 2022

Service line/Business Unit :

Data

Domain: Big data

No of member at Cohort :44

Mode of conversation: Microsoft teams

Coach Name:

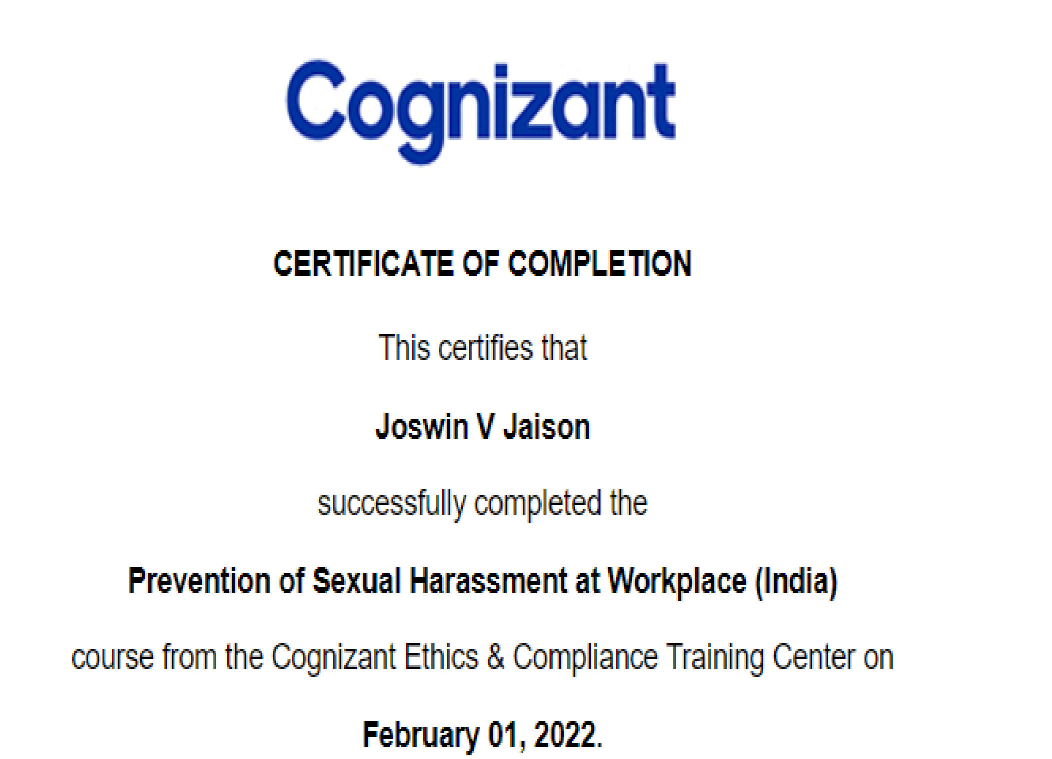
Grace Esther S

Weekly overview of Internship activities

Week 1-3 : January 27-February 11

- ❑ The first 2 weeks of my internship was mainly for induction activities. It involves the completion of onboarding formalities, completion of courses, participation in introductory sessions and leadership connects at cognizant.
- ❑ The introductory sessions were conducted on topics like importance of professional etiquette in corporate environment
- ❑ Security training helped me to gain knowledge on security best practices in cognizant
- ❑ This practices include protecting our credentials and data, protecting our assets and network systems, protecting our workspace.
- ❑ There were also leadership connects where directors of various domains in cognizant took session

Introductory courses



Cognizant

CERTIFICATE OF COMPLETION

This certifies that

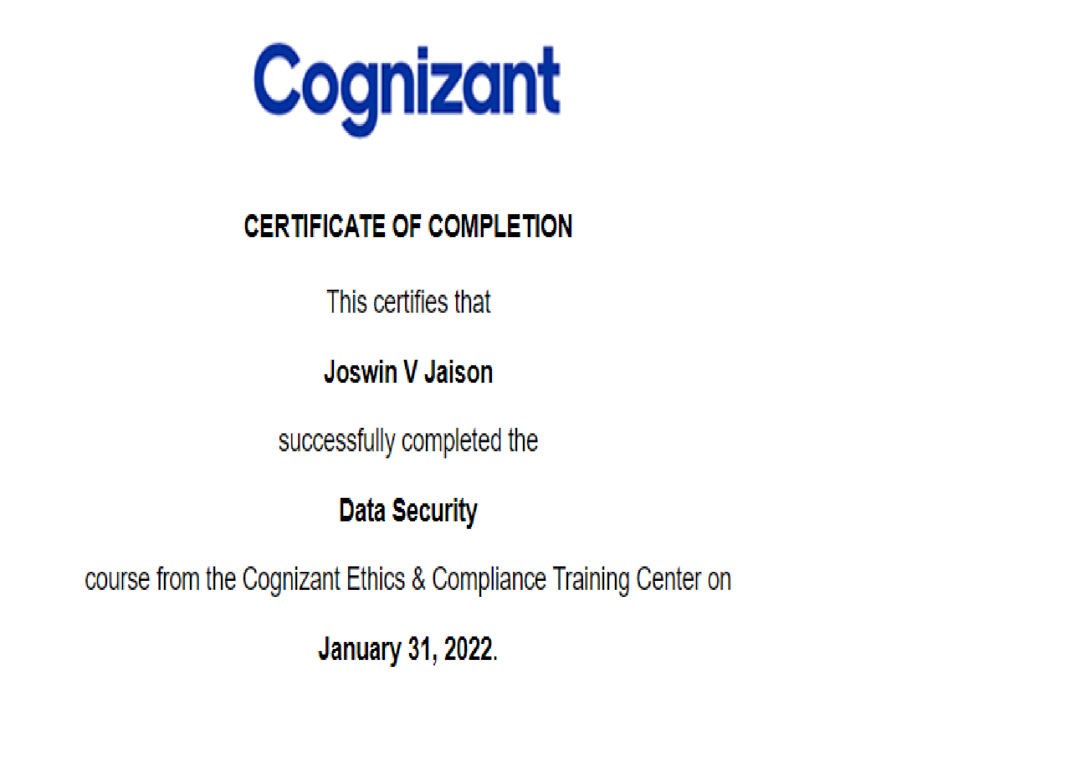
Joswin V Jaison

successfully completed the

Prevention of Sexual Harassment at Workplace (India)

course from the Cognizant Ethics & Compliance Training Center on

February 01, 2022.



Cognizant

CERTIFICATE OF COMPLETION

This certifies that

Joswin V Jaison

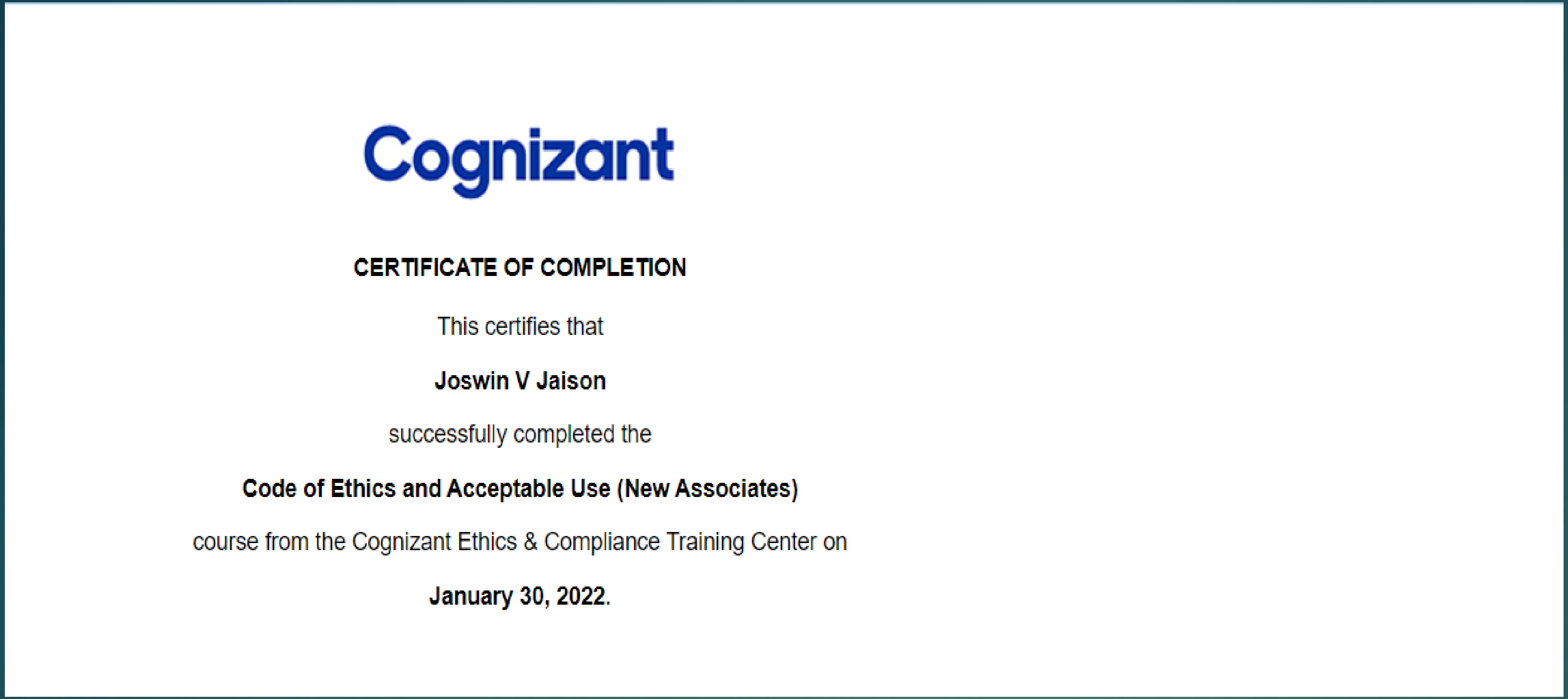
successfully completed the

Data Security

course from the Cognizant Ethics & Compliance Training Center on

January 31, 2022.

Introductory courses



Cognizant

CERTIFICATE OF COMPLETION

This certifies that

Joswin V Jaison

successfully completed the

Code of Ethics and Acceptable Use (New Associates)

course from the Cognizant Ethics & Compliance Training Center on

January 30, 2022.

Weekly overview of Internship activities

Week 3 : February 14-February 21

Database fundamentals

1. *Couse 1: Relational Database design*

- ❑ how to create an effective relational database design using proven concepts and industry knowledge.
- ❑ helped in understanding normalization and type of normal forms.
- ❑ how to identify tables and how to create relationships
- ❑ gave me an insight on the naming convention we need to follow while designing tables.

2.Course 2:Understand SQL using the MySQL database. Learn Database Design and Data Analysis with Normalization and Relationships

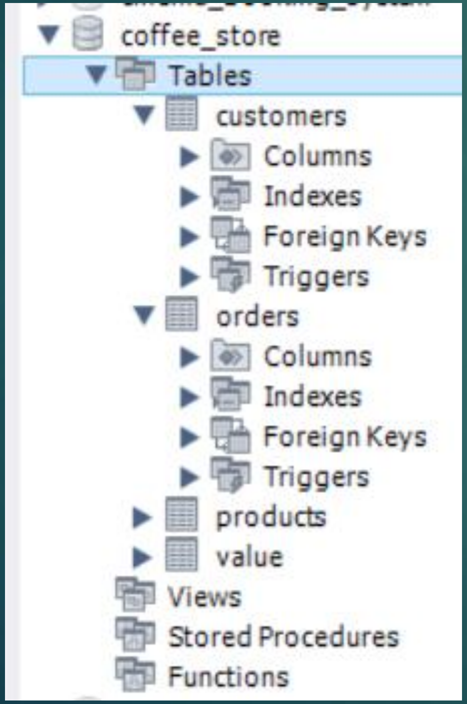
□ This course provide an in depth knowledge on data base design .

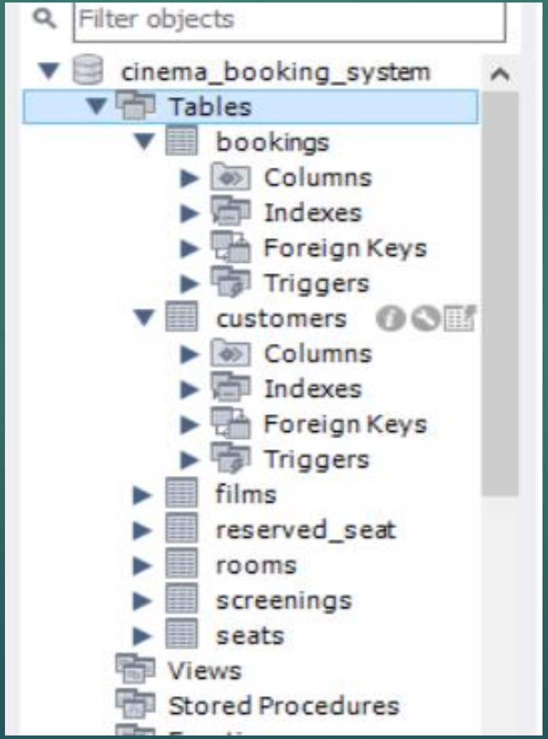
□ Worked on MySql work bench

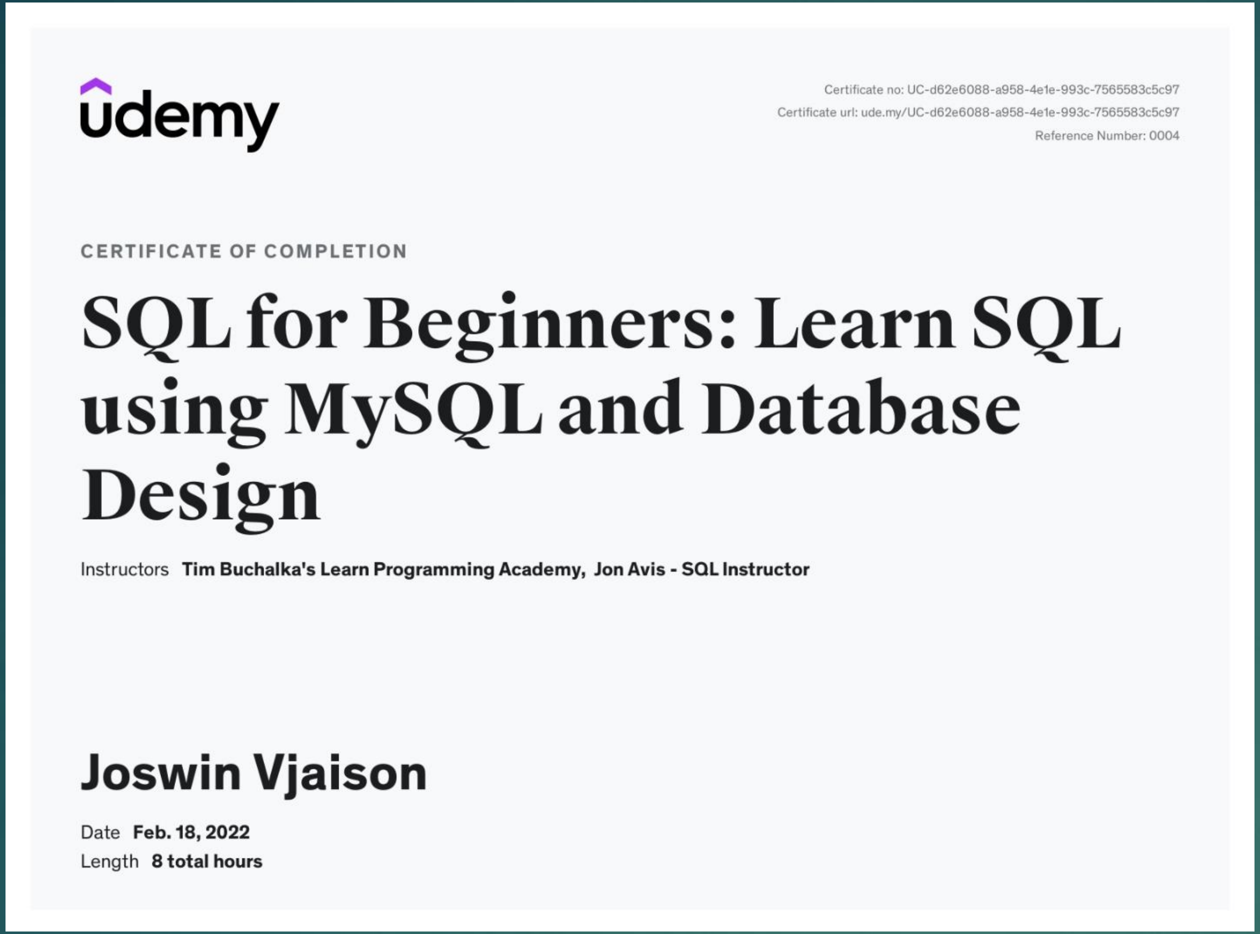
□ The initial phase of this course focus mainly on data definition language , data manipulation language , Transaction control language

□ Then it is shifted to aggregate functions , sub queries , functions etc.

□ Performed 2 case studies with Coffee store data base and cinema booking database.







udemy

Certificate no: UC-d62e6088-a958-4e1e-993c-7565583c5c97
Certificate url: ude.my/UC-d62e6088-a958-4e1e-993c-7565583c5c97
Reference Number: 0004

CERTIFICATE OF COMPLETION

SQL for Beginners: Learn SQL using MySQL and Database Design

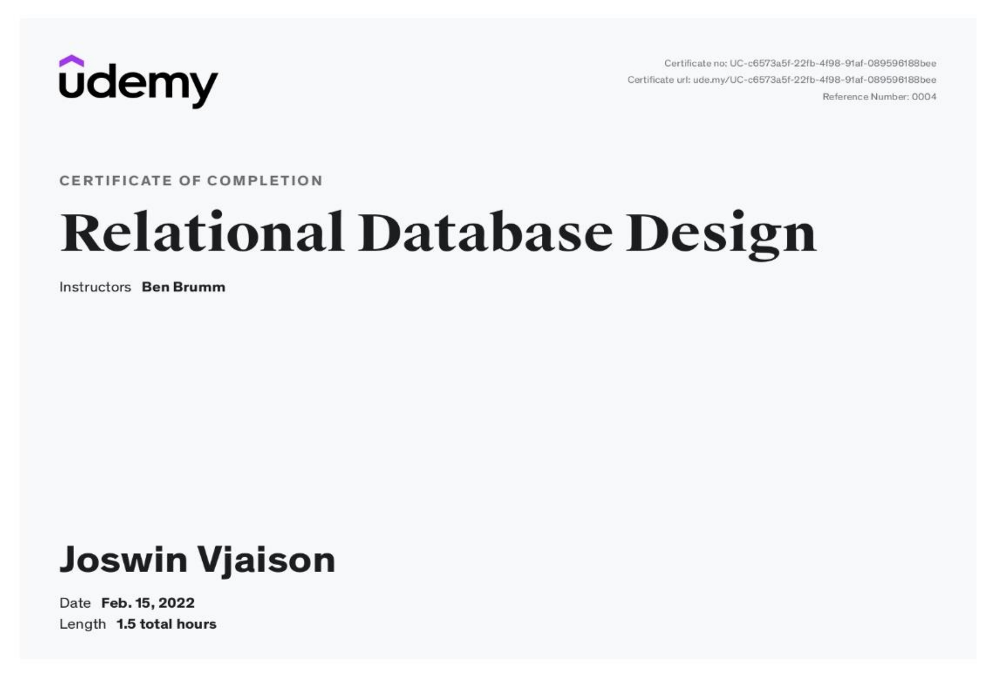
Instructors **Tim Buchalka's Learn Programming Academy, Jon Avis - SQL Instructor**

Joswin Vjaison

Date **Feb. 18, 2022**
Length **8 total hours**

Data Modelling and relational database design using Erwin

- ❑ theory course which help us to learn how to develop data models and maintain them using the data modelling tool called Erwin.
- ❑ It helped me in creating entity relationship diagrams by identifying entities, attributes, relationships and constraints from a set of requirements.



udemy

Certificate no: UC-c6573a5f-22fb-4f98-91af-089596188bee
Certificate url: ude.my/UC-c6573a5f-22fb-4f98-91af-089596188bee
Reference Number: 0004

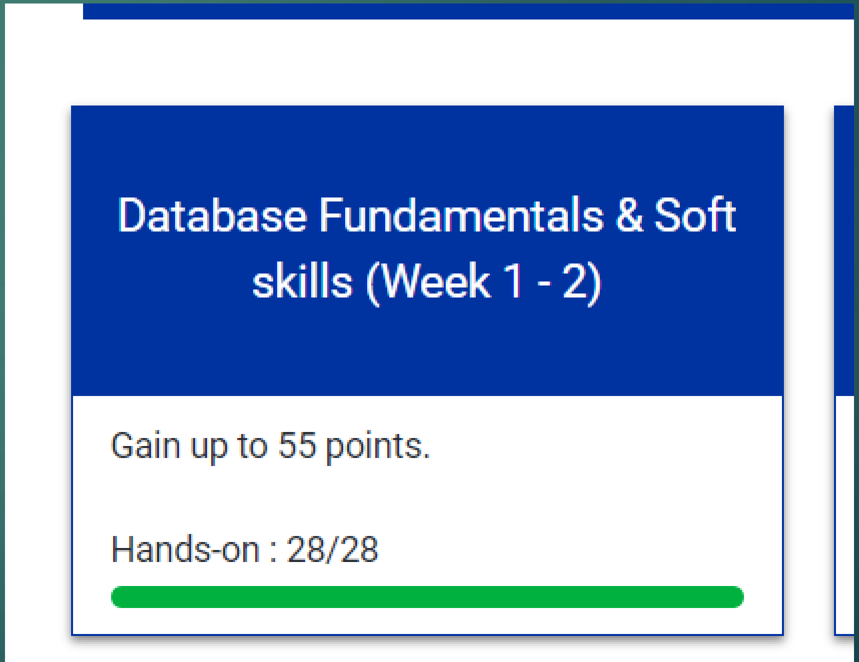
CERTIFICATE OF COMPLETION

Relational Database Design

Instructors **Ben Brumm**

Joswin Vjaison

Date **Feb. 15, 2022**
Length **1.5 total hours**



Database Fundamentals & Soft skills (Week 1 - 2)

Gain up to 55 points.

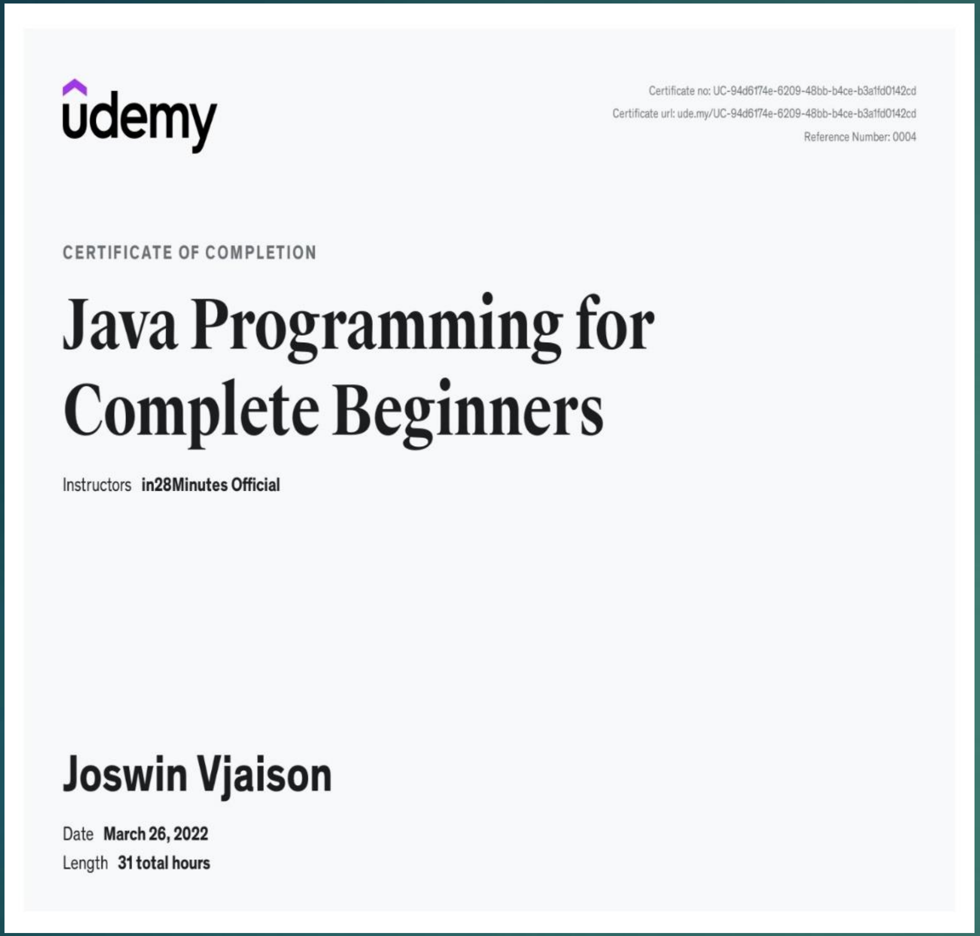
Hands-on : 28/28

Weekly overview of Internship activities

Week 4-6 : February 21- March 7

Java programming

- ❑ focus mainly on basic java concepts , object oriented programming in java, advanced java concepts ,collections in java etc.
- ❑ assigned with a 31 hours java course which helped me in enhancing my java skills and helped me in learning advanced java concepts
- ❑ come across basic java concepts like conditional statements , primitive types ,loops etc.
- ❑ the next stage was mainly based on object oriented programming.
- ❑ implemented the concepts like inheritance , abstraction etc in java.
- ❑ have come across object composition concept in Java which will help in programming.
- ❑ The next phase was mainly on collections .
- ❑ learnt about implementing collections like List interface, Set interface , Map interface etc.
- ❑ The next phase was mainly based on functional programming and multi threading.
- ❑ In multithreading I have learnt about creating thread,placing priority requests in thread,thread utility methods , executor service etc.
- ❑ The next phase was mainly based on exception handling .
- ❑ In that I have come across basics of exception handling ,creating our own exception,throwing a checked exception etc.



udemy

Certificate no: UC-94d6174e-6209-48bb-b4ce-b3a1fd0142cd
Certificate url: ude.my/UC-94d6174e-6209-48bb-b4ce-b3a1fd0142cd
Reference Number: 0004

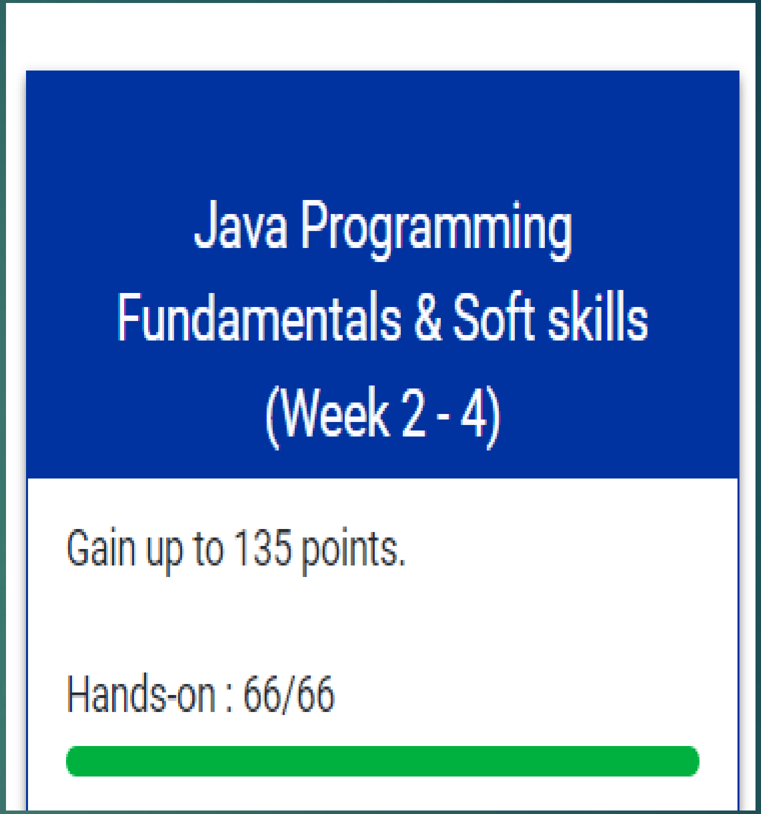
CERTIFICATE OF COMPLETION

Java Programming for Complete Beginners

Instructors **in28Minutes Official**

Joswin Vjaison

Date **March 26, 2022**
Length **31 total hours**



Java Programming Fundamentals & Soft skills (Week 2 - 4)

Gain up to 135 points.

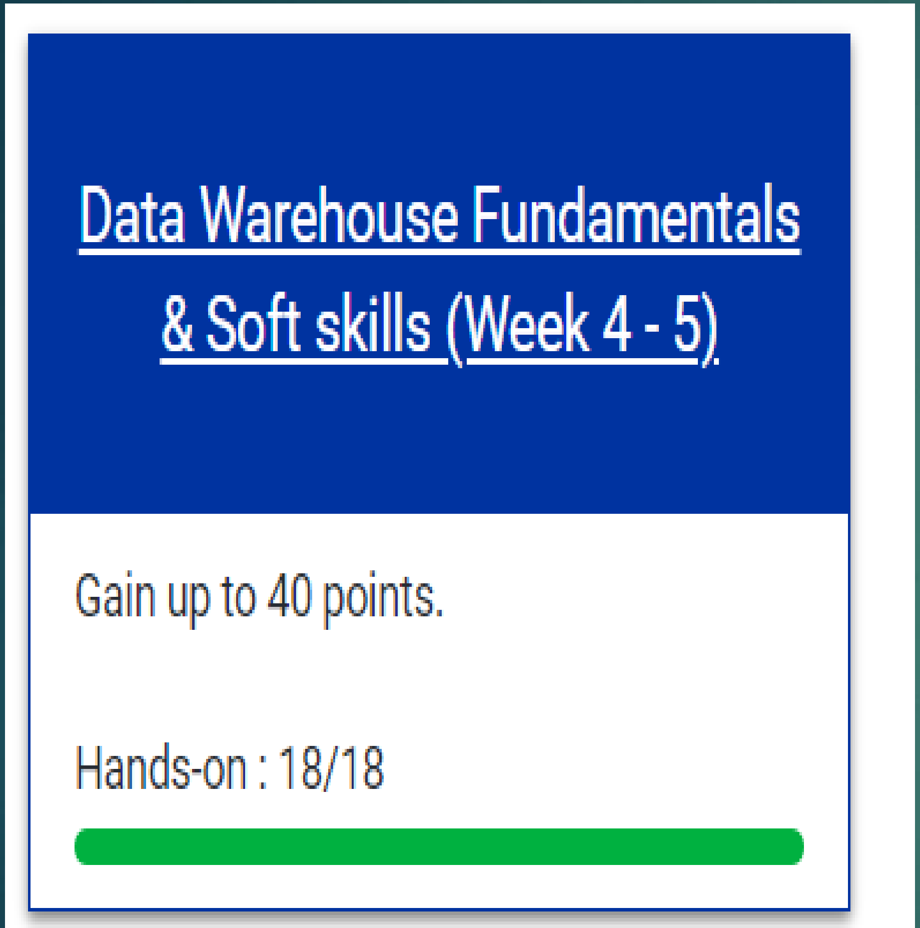
Hands-on : 66/66

Weekly overview of Internship activities

Week 6-8: March 7-March 14

Data warehouse fundamentals

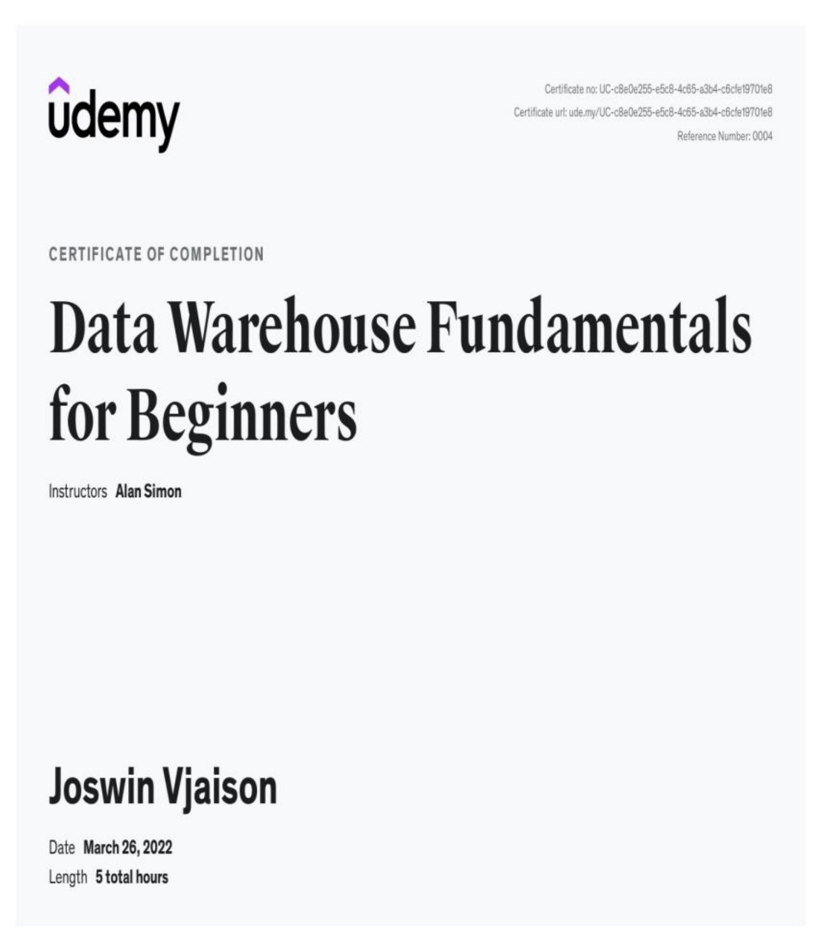
- ❑ focus on data warehouse fundamentals and Unix commands.
- ❑ completed a course on data warehousing and completed the hands on exercises on Unix
- ❑ I have come across topics like data warehousing architecture ,ETL ,dimension tables , fact table , star schema , snow flake etc in data ware house architecture.
- ❑ During Unix phase I have worked on areas including directory creation , copying file , grep ,tail command , redirect command , pattern printing etc.



Data Warehouse Fundamentals & Soft skills (Week 4 - 5)

Gain up to 40 points.

Hands-on : 18/18



udemy

Certificate no: UC-c8e0e255-e5c8-4c65-a3b4-c6cfe19701e8
Certificate url: ude.my/UC-c8e0e255-e5c8-4c65-a3b4-c6cfe19701e8
Reference Number: 0004

CERTIFICATE OF COMPLETION

Data Warehouse Fundamentals for Beginners

Instructors **Alan Simon**

Joswin Vjaison

Date **March 26, 2022**
Length **5 total hours**

Weekly overview of Internship activities

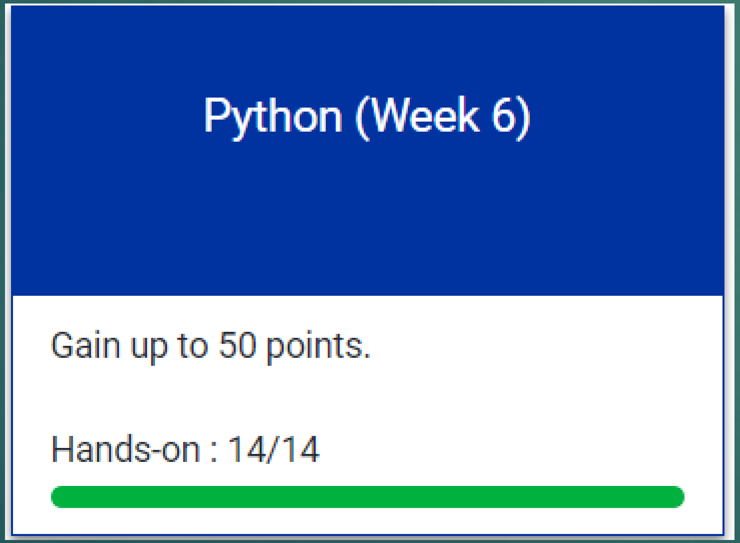
Week 8: March 14-March 21

Python

❑ Was focused on developing python programming skills

❑ Assigned to complete hands on exercises in python which was mainly based on python data types , functions ,collections in python , operators ,modules , file handling

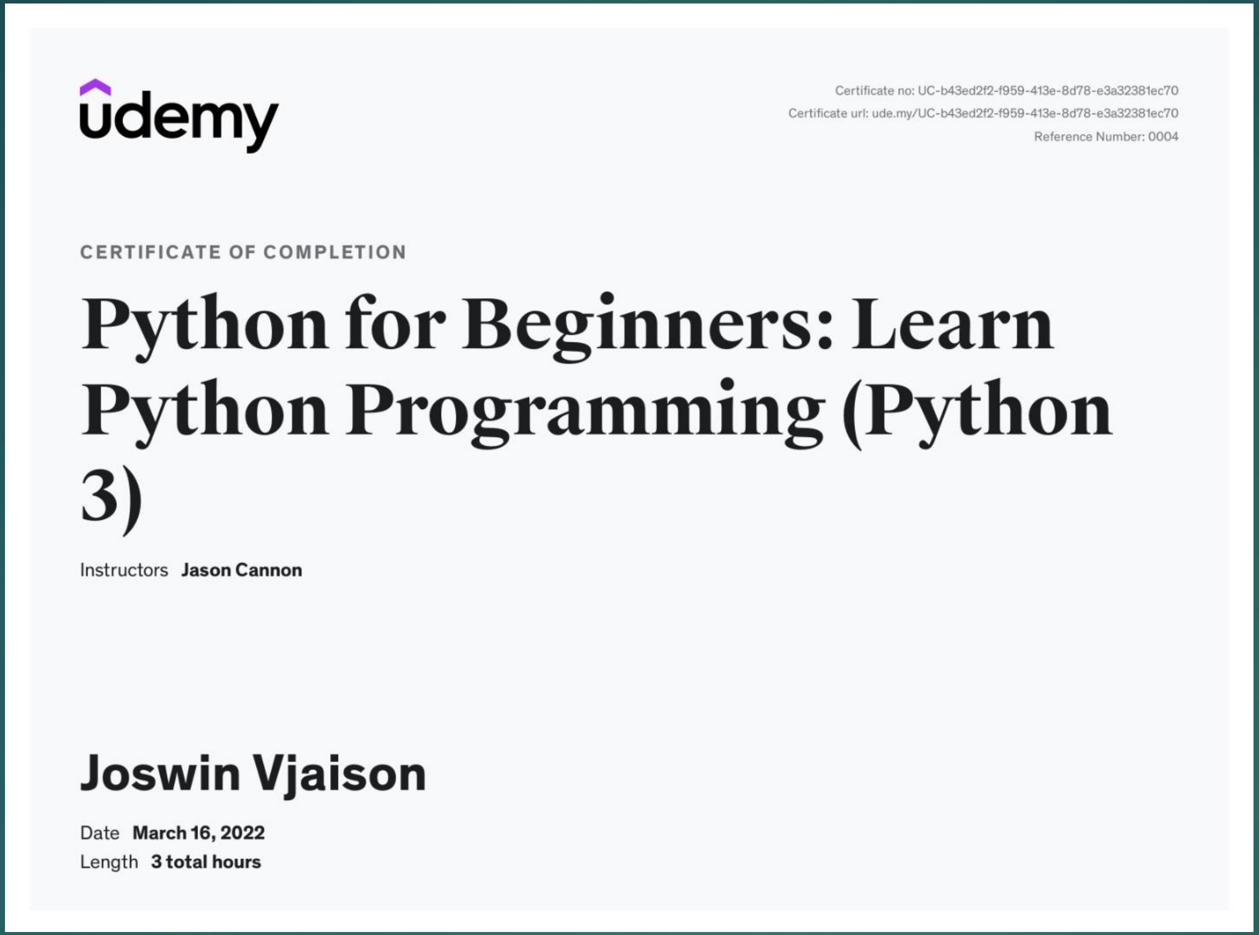
❑ During python collection stage I have worked on collections like list , tuple and dictionary



Python (Week 6)

Gain up to 50 points.

Hands-on : 14/14



udemy

Certificate no: UC-b43ed2f2-f959-413e-8d78-e3a32381ec70
Certificate url: ude.my/UC-b43ed2f2-f959-413e-8d78-e3a32381ec70
Reference Number: 0004

CERTIFICATE OF COMPLETION

Python for Beginners: Learn Python Programming (Python 3)

Instructors **Jason Cannon**

Joswin Vjaison

Date **March 16, 2022**
Length **3 total hours**

Weekly overview of Internship activities

Week 9: March 21-present

Big data and cloud fundamentals

- ❑ Assigned with a course on Big data with spark
- ❑ Working on Ubuntu
- ❑ During the starting phase I have worked on a case study with Walmart stock dataset.
- ❑ learnt about data frame basics.
- ❑ worked with group by, order by ,head, collect,describe etc
- ❑ also had hands on experience on date and timestamps in spark which is helpful in data analysis.
- ❑ In machine learning module I have started with linear regression
- ❑ worked on a case study with ecommerce customer dataset .
- ❑ In that case study yearly amount spent by the customers are assigned as labels.
- ❑ I have also done a consulting project on building a predictive model for a ship company.
- ❑ The objective is to predict the number of crew members considering various features like no of cabins,no of passengers etc.

Weekly overview of Internship activities

Week 9: March 21-present

Big data and cloud fundamentals

❑ In logistic regression phase I performed a case study on titanic dataset

❑ Here we need to create a logistic regression model to predict the total number of passengers survived.

❑ I have done string indexing ,one hot encoding converting into training and testing dataset etc.

❑ I have also done a project called customer churn prediction

❑ The objective of the project is to predict whether a particular customer will churn or not

Software and Operating System used

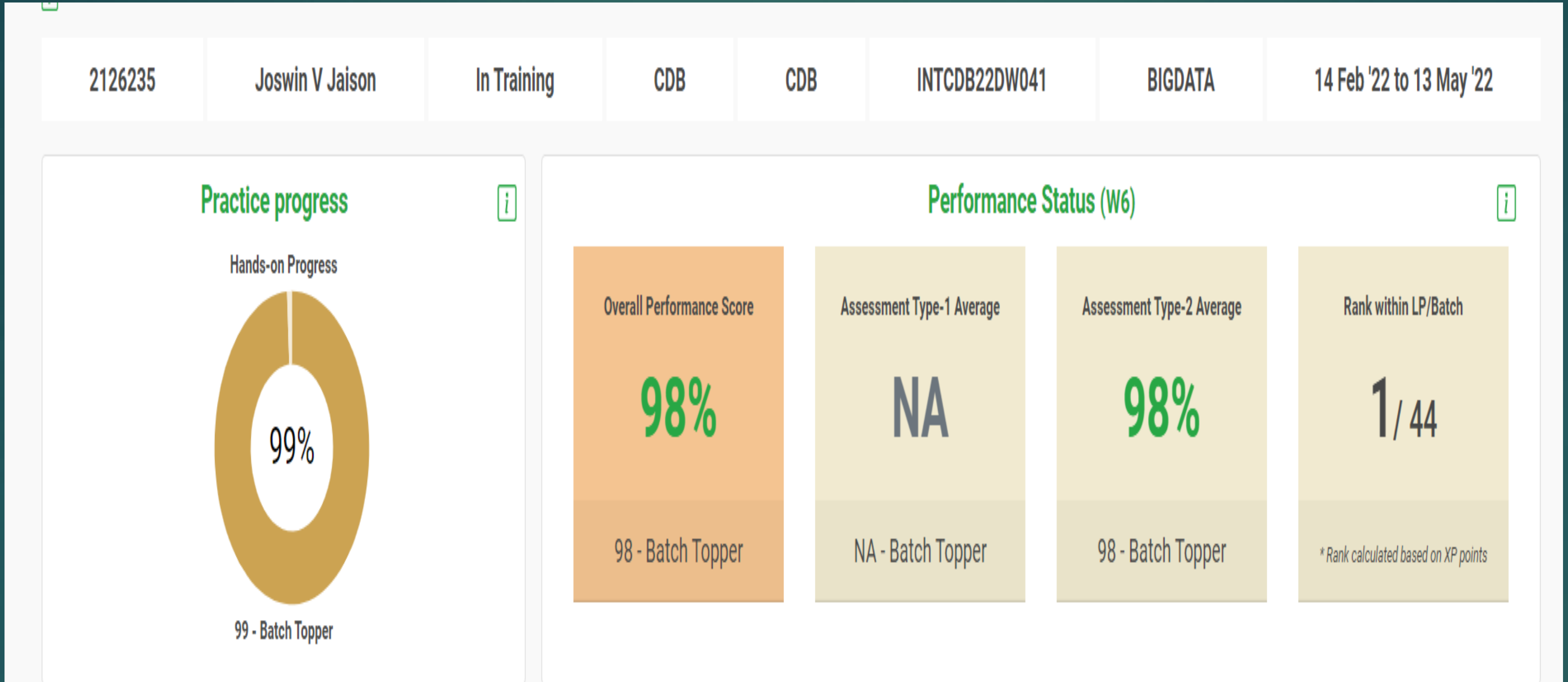
- ❑ SQL Work Bench
- ❑ Eclipse
- ❑ Jupyter Notebook
- ❑ Oracle VM Virtual Box
- ❑ Windows Operating System
- ❑ Ubuntu

Tools and Technologies Used

- ❑ Apache Spark
- ❑ Python Programming language
- ❑ Java Programming language
- ❑ Unix commands
- ❑ SQL
- ❑ My SQL
- ❑ Spring Framework

Perfomance and evaluation

Overall perfomance score



2126235	Joswin V Jaison	In Training	CDB	CDB	INTCDB22DW041	BIGDATA	14 Feb '22 to 13 May '22

Practice progress

Hands-on Progress

99%

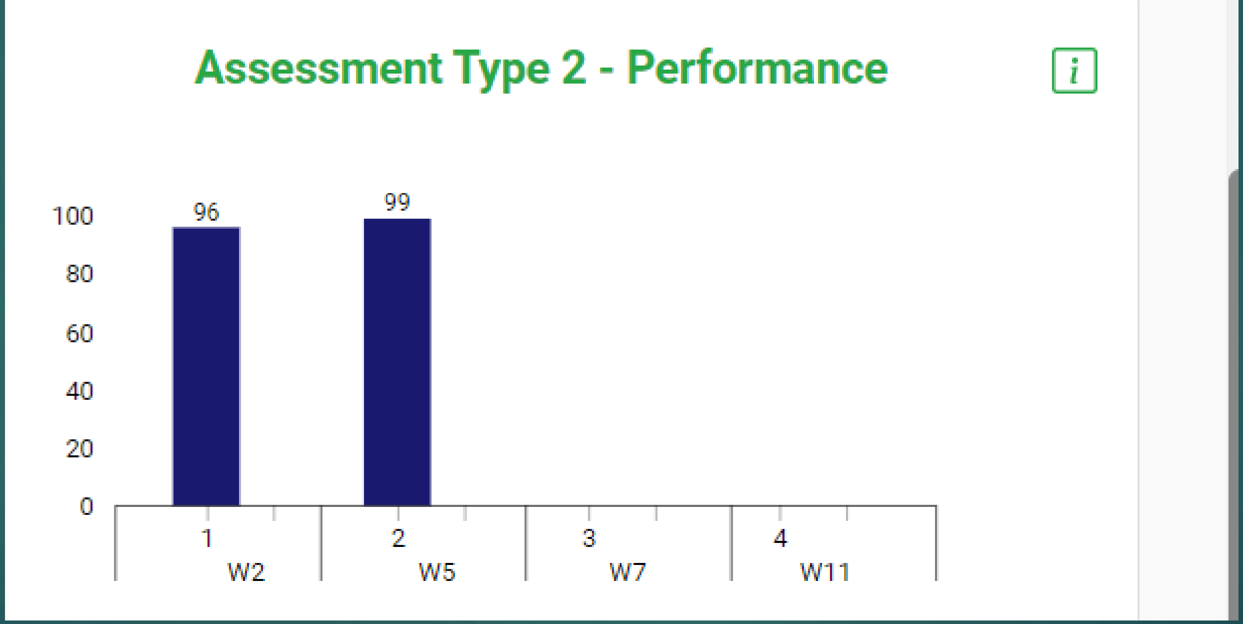
99 - Batch Topper

Performance Status (W6)

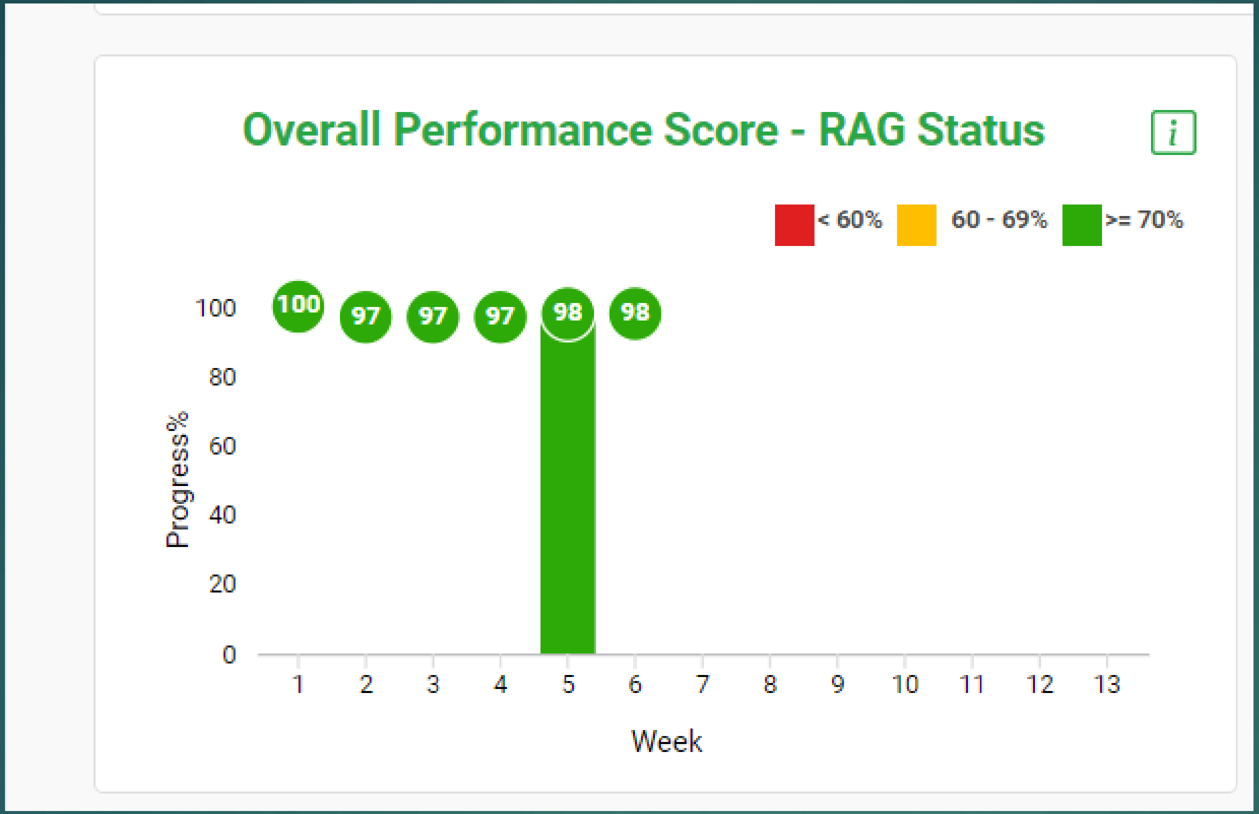
Overall Performance Score	Assessment Type-1 Average	Assessment Type-2 Average	Rank within LP/Batch
98%	NA	98%	1/ 44
98 - Batch Topper	NA - Batch Topper	98 - Batch Topper	* Rank calculated based on XP points

Perfomance and evaluation

The candidates are required to complete Integrated capability test . An average score of 70 is mandatory for the candidate to complete the internship successfully .
1.Genc AVM ANSI-SQL Skill Based Assessment percentage : 96/100
2.Genc Core Java –Skill Based Assessment percentage : 99/100



Perfomance and evaluation



Thank you