Big Data

	MON	TUE	WED	THU	FRI	ENVIRONMENT
Week 1 **Python**	• Interpreter vs Compiler • REPL • Identifiers & Keywords • Simple/Compound statements • Types & Values	• Namespaces • String Literals • Operators • Lambdas • Functions • Exceptions	• Lists • Sets • Tuples • Dictionaries • Classes & Objects • Variable Scope/lifetime • Simple/Compound Statements	• File Handling: open()/close() • Logging • Collections • Datatime • RegEx	• PyPI & pip • Modules • Unit Testing • Pylint	• VS Code
Week 2 **Database-MongoDB**	Written evaluation Trainer interviews Quality Control Audit • Intro to RDBMS • DML, DDL, DQL • Basic SQL queries • Aggregate functions • Multiplicity	• NoSQL • MongoDB • PyMongo Driver • Connection String	• Mongo Collections • Mongo Documents • Insert • Find	• Query • Sort • Limit	• Update • Delete • Indexing	• MongoDB
Week 3 **UNIX/Hadoop Fundamentals**	Written evaluation Trainer interviews Quality Control Audit • Intro to Open Source Software • Linux commands brainstorm • Root [/] vs Home [~] • Commands: mkdir, rm, cp, mv, cd, ls, cat, grep, echo	• Commands: df, fdisk, sfdisk, cfdisk, lsblk, blkid, mdadm • File Editors - vim, nano • Intro to SSH (credentials/private key) • Intro to data evolution • Intro to big data	• Hadoop Ecosystem Introduction • Intro to HDFS • Evolution of Hadoop • HDFS Commands	• Introduction to MapReduce • Mapper/Intermediate/Reducer phases • Partitioners • Combiners	• YARN Overview • InputFormats	
Week 4 **AVRO/Sqoop/Hive**	Written Evaluation Trainer interviews Quality Control Audit • Introduction to AVRO File Format • Schema declaration in AVRO	• Primitive and complex types • AVRO and MapReduce	• Importing/exporting data using Sqoop • Introduction to Sqoop- basic commands	• Introduction to Hive • Basic Hive Queries • Hive Commands	• Data Types • Managed vs External tables • Partitions and Buckets	

	MON	TUE	WED	THU	FRI	ENVIRONMENT
Week 5 **Pig/Spark Fundamentals**	Written Evaluation Trainer Interviews Quality Control Audit • Introduction to Pig Latin • Commands • Datatypes	• sorting/filtering • Describe/Explain/Illustrate • local mode vs MapReduce mode	• Introduction to Spark • Hadoop vs Spark • Spark Setup	• Introduction to RDDs • Basic RDD operations • Local vs Cluster mode • Working with Key/Value pairs	• Transformations • Actions • Shared variables • pySpark	
Week 6 **Spark Fundamentals**	Quality Control Audit Trainer interviews Written Evaluation • Accumulators	• Creating Spark EMR cluster • Spark Cluster mode • Introduction to YARN	• Spark Cluster Manager • Running Spark job on EMR • Driver class configuration • Executors	• Configure number of executors • Spark cluster configuration • Configure memory: Driver & executors	• Spark caching • Memory management	• AWS EMR
Week 7 **Spark SQL/DataFrames**	Quality Control Audit Trainer interviews Written Evaluation • Introduction to Spark SQL • Introduction to DataSets	• Introduction to DataFrames • Entry point: SparkSession • Creating DataFrames	• Creating DataSets • Working with RDDs • Using DataFrame aggregate functions	• Bucketing • Sorting and Partitioning	• Working with JSON Datasets • Working with Parquet Files	
Week 8 **Streaming/Kafka**	Quality Control Audit Trainer interviews Written Evaluation • Introduction to Streaming • Introduction to Kafka • Kafka Fundamentals • Topics/Brokers/Consumer/Producer	• Apache Kafka architecture • Pub-Sub messaging • Creating a Kafka topic • Retrieve list of topics	• Create producer and consumer • Send messages from Producer • Producer API	• Introduction to Spark Streaming • Spark engine • Sending data stream from Kafka to Spark	• Processing data stream using Spark streaming	• Kafka
Week 9	Project 3 Written evaluation Trainer interviews Quality Control Audit	Project 3	Project 3	Project 3	Project 3	
Week 10	Project 3 QC Audit - Cumulative	Project 3	Project 3	Project 3	Project 3	

	MON	TUE	WED	THU	FRI	ENVIRONMENT
Week 11	Project 3	Project 3	Project 3	Project 3 Project showcase		

PROJECT	TECHNOLOGIES
Project 3	Spark, Spark SQL, Kafka, Spark Streaming