Contents

I&D V14-B(Data Engineering with AWS) - 2025
Lot- Course Structure

I&D (Data Engineering with AWS) Lot provides exposure to a band of data warehousing technologies. It focuses on application development for data warehouses. The following table lists the course structure for I&D Lot.

Sr. No.	Course	Duration (In Days)
1	Discover(Induction)	-
2	Power skills (Behavioural) & Language proficiency -Foundation – session 1	1
3	Oracle SQL/ANSI SQL Overview on RDBMS	4
2	Power skills (Behavioural) & Language proficiency -Foundation – session 2	0.5
5	Oracle SQL/ANSI SQL Overview on RDBMS	1.5
6	Data Warehousing Concepts	0.5
7	ETL and ELT Basics	0.5
8	**Data Modelling for Business Intelligence & Data Vault**	**1**
9	Python	2
2	Power skills (Behavioural) & Language proficiency -Foundation – session 3	0.5
11	Python	2.5
12	Module Assessment	1
13	AWS Fundamentals	1
2	Power skills (Behavioural) & Language proficiency -Foundation – session 4	1
15	**Spark**	**2**
16	Data Ingestion on AWS	0.5
17	Data Storage	0.5
18	**Glue & Athena**	**3**
2	Power skills (Behavioural) & Language proficiency -Foundation – session 5	1
20	Real Time Analytics / AWS Streams	0.5
21	Data Access	0.5
22	AWS Databases (RedShift)	2
2	Power skills (Behavioural) & Language proficiency -Foundation – session 6	1
24	AWS Databases (RedShift)	1
25	Data Migration Utilities on AWS	1
26	**Databricks**	3
2	Power skills (Behavioural) & Language proficiency -Foundation – session 7	1
28	Data Process & Compute Services (Airflow, Stepfunction to be included)	3
29	Sprint - Evaluation	1

30	L1 Preparation	1
31	L1 Assessment (MCQ - Concept & Code-based Qs)	1
	Total	40

I&D Curriculum

Oracle SQL/ANSI SQL Overview on RDBMS

Program Duration: 5.5 days
Contents:

Introduction to Database
- Introduction to DBMS
- Characteristics of DBMS
- DBMS Models
- Relational DBMS
- Data Integrity
- Security in Database

Normalization & Codd's Rules for "FULLY" Functional System
- First Normal Form
- Second Normal Form
- Third Normal Form
- Relational DBMS
- Data Integrity

Structured Query Language
- Interacting SQL using SQL *Plus
- Using SQL *Plus
- What is SQL?
- Rules for SQL statements
- Standard SQL Statement Groups
- Basic DataTypes
- Rules for naming a Table

- Specifying Integrity Constraints
- DDL Statements: Create, Alter, Drop, Truncate
- Regular vs Temporary tables

Data Manipulation Language
- Inserting Rows Into a Table
- Deleting Rows from a Table
- Updating Rows in a Table

Data Control Language
- Grant
- Revoke

Database Objects
- Index
- Synonym
- Sequence
- Views

Data Query Language (Select Statement)
- Select Statement
- Distinct Clause
- Comparison, arithmetic & Logical Operators SQL Operators
- The ORDER BY Clause
- Tips and Tricks

Aggregate Functions, Group By and Having Clause
- Aggregate Functions
- The GROUP BY Clause
- HAVING Clause
- ROLLUP Operation
- CUBE Operation
- Tips and Tricks

Transactions
- Transaction
- Commit Command
- Rollback and Savepoints

Joins and Subqueries
- Inner/Equi Join
- Outer Join
- Self Join
- Subquery
- SUBQUERIES Using Comparison Operators Co-related Subquery
- Exists / Not Exists Operator

Set Operations
- The UNION Operator
- The INTERSECT Operator

- The MINUS Operator
- The UNION Operator
- The INTERSECT Operator
- Tips and Tricks

Data Warehouse Concepts

Program Duration: 0.5 day.

Contents:

Business Intelligence
- Business Intelligence
- Need for Business Intelligence
- Terms used in BI
- Components of BI

General concept of Data Warehouse
- Data Warehouse
- History of Data Warehousing
- Need for Data Warehouse
- Data Warehouse Architecture
- Data Mining Works with DWH
- Features of Data warehouse
- Data Mart
- Application Areas

Dimensional modeling
- Dimension modeling
- Fact and Dimension tables
- Database schema
- Schema Design for Modeling
- Star
- Snow Flake
- Fact Constellation schema

ETL and Metadata
- ETL process
- Metadata used in ETL
- Metadata in Data Warehousing
- Simple Data warehouse model

Online Analytical Processing (OLAP)
- Online Analytical Processing (OLAP)
- Nature of OLAP analysis
- Types of OLAP
- OLAP Tools
- OLTP and OLAP

- OLAP Functional requirements
- OLAP Fast and Selective
- Operational versus Informational System

Data Mining

- Data mining
- The Knowledge Discovery process
- Need of Data Mining
- Use of Data mining
- Data mining and Business Intelligence
- Types of data used in Data mining
- Data Mining applications
- Data Mining products
- Data Mining market

Best Practices for Building Data Warehouse

- Recipe for a Successful data warehouse
- Data warehouse pitfalls
- Popular BI DW tools and suits
- Trends in BIDW

ETL and ELT Basics

Program Duration: 0.5 day.

- Basic Concepts
  - Data warehouse
  - Data warehousing strategies
  - Data warehouse architecture
  - ETL Meaning
  - Need for ETL
  - ETL Process
  - Operational Considerations
- ETL Process
  - Data extraction
  - Data transformation
  - Data Loading
- Operational Considerations
  - Exceptional Handling
  - Alerts and Notification
  - Process restart-ability
  - Job Scheduling and Monitoring

ETL Tools

- Leading ETL tool vendors

ETL tool strengths / weaknesses
Choosing the correct ETL tool

Basic concepts of ELT
Tools used for ELT
Difference between ETL and ELT

Data Modeling for Business Intelligence and Data Vault

Program Duration: 1 day

Contents:

· Introduction to Data Modeling
  - Importance of data modeling
  - Features of a good data model
  - Who should be involved in data modeling
  - Database design stages and deliverables
  - Classification of information

· Understanding Business Requirements
  - Need of Requirement Analysis
  - Characteristics of a Good Requirement
  - The Data Life cycle
  - Methods of Collecting requirement
  - Business Requirement Specification (BRS)

· Conceptual Model
  - Define conceptual model
  - Objectives of conceptual model
  - Components of Conceptual Model
  - Types of Modeling
  - Entity-Relationship (ER) model
  - Types of Attributes
  - Join Problems
  - Steps of dimension modeling
  - Star Schema
  - Snowflake Schema
  - Bill Inmon Vs Ralph Kimball Approach

· Logical Model

Define logical model
List features of a logical model
Transformations required to be done while converting a conceptual model into a Logical model
Activities in table specification
Activities in column specification
Activities in Primary key specification

Datavault or **data vault modeling**

What is Datavault
History of Datavault
Basic Notation
DataVault Vs Dimension modeling

Python

Program Duration: 1 days.
Contents:

Introduction to Python Programming
- Why do we need Python?
- Program structure in Python

Execution steps
- Interactive Shell
- Executable or script files.
- User Interface or IDE

Flow Control
Boolean Operators
Comparison Operators
Binary Boolean Operators
The not Operator

Data Types and Operations
- Numbers
- Strings
- List
- Tuple
- Dictionary

- Other Core Types
- Changing Values in a List with Indexes
- List Concatenation and List Replication
- Using for Loops with Lists
- Removing Values from Lists with del Statements

Pattern Matching with Regular Expressions

- Regular Expression Matching
- Finding Patterns of Text with Regular Expressions
- Grouping with Parentheses
- Matching Multiple Groups with the Pipe
- Matching Zero or More with the Star
- Matching Specific Repetitions with Curly Brackets
- Case-Insensitive Matching

Statements and Syntax in Python

- Assignments, Expressions and prints
- If tests and Syntax Rules
- While and For Loops
- Iterations and Comprehensions
- Break/Continue Statements

Functions in Python

- Function definition and call
- Function Scope
- Return Values and return Statements
- Local and Global Scope
- Arguments
- Function Objects
- Anonymous Functions
- Exception Handling

Modules and Packages-Basic

- Module Creations and Usage
- Package Creation and Importing

Classes in Python

- Classes and instances
- Classes method calls

File Operations

- Backslash on Windows and Forward Slash on OS X and Linux
- Absolute vs. Relative Paths
- Finding File Sizes and Folder Contents
- Open/Read/Write/Append into file
- Using Files
- Copying Files and Folders

AWS Fundamentals

Program Duration: 1 day
Contents:

- What is Cloud Computing?
- Cloud Deployment Models
- Key Cloud Concepts
- Cloud Service Models
- Cloud providers and details
- AWS Overview
- Various AWS Services
- Global Infrastructure – Regions and Availability Zones
- Understanding Identity Access Management of AWS
- EC2 Instance
- Auto Scaling
- Load Balancing
- Object Storage
- Amazon Virtual Private Cloud (VPC)
- Relational Database Service (RDS)
- Monitoring Services
- AWS S3 / Storage Tiers
- EBS
- EFS
- AWS GLUE

Spark

Program Duration: 2 days
Contents:

SPARK Basics

What is SPARK?
History of SPARK
SPARK Architecture
SPARK Shell
Pyspark Introduction

Installation

Prerequisites
SPARK 2 Standalone

- SPARKS 2 on Cloudera
- Python with Anaconda

Understanding SPARK

- SPARK Architecture
- Operations, and Transformations
- Fine Grained Transformations and Scalability
- Parallelism by partitioning Data
- Pipelining
- Lazy Execution, Lineage, Directed Acyclic Graph(DAG), and Fault Tolerance
- SPARK based Libraries and Packages
- Word Count
- Storage and Supported Data Formats
- Low-level and High-level SPARK APIs
- Performance Optimizations: Tunsten and Catalyst
- SparkContext and SparkSession
- Spark Configuration + Client and Cluster Deployment Modes
- Spark on Yarn; Visualizing Your Spark App; Logging in Spark

RDDs

- RDD and PairRDD
- Creating RDDs with Parallelize
- collect(), take(), first()…
- Partitions, Repartition, Coalesce, Saving as Text
- RDDs from External Datasets
- Saving Data as PickleFile, NewAPIHadoopFile

RDDs with Transformations

Lineage and Dependencies

Spark Advanced

- Accumulators
- Broadcast Variables
- Piping to External Programs
- Numeric RDD Operations
- Spark Runtime Architecture
- Deploying Applications
- Functional Programming: Lambda in Spark
- Map, FlatMap, Filter, and Sort
- Actions
- Partition Operations: MapPartitions and PartitionBy
- Sampling of Data
- Set Operations: Join, Union, Full Right, Left Outer, and Cartesian
- Combining, Aggregating, Reducing, and Grouping on PairRDDs
- ReduceByKey vs. GroupByKey

- Grouping Data into Buckets with Histogram
- Regular Expressions
- Caching and Data Persistence
- Shared Variables
- Developing Self-contained PySpark Application, Packages, and Files
- Disadvantages of RDDs
- Dataframes
- Hello DataFrames and Spark SQL
- DataFrames to RDDs and Vice versa
- Loading DataFrames from CSVs
- Schemas
- Loading Parquet and JSON
- Rows, Columns, Expressions, and Operators like Cloning, Renaming, Casting, & Dropping
- Querying, Sorting, and Filtering DataFrames
- Missing or Corrupt Data
- Saving DataFrames
- SPARK with SQL
- Spark SQL Overview
- Spark SQL Architecture
- Catalyst
- Plan Optimization & Execution
- ROW API
- Querying Using Temporary Views
- Loading Files and Views into DataFrames Using Spark SQL
- Saving to Persistent Tables + Spark 2 Known Issue
- Hive Support and External Databases; Aggregating, Grouping, and Joining
- User Defined Functions (UDFs) on Spark SQL
- Spark streaming
- What is Spark streaming?
- Spark streaming : How it works ?
- Spark DStreams
- A Twitter example
- Fault-tolerance
- Stateful Stream Processing

Data Ingestion on AWS

Program Duration: 0.5 day
Contents:

AWS sFTP
AWS CLI
AWS Data pipeline
AWS API Management
AWS SDK

Data Storages

Program Duration: 0.5 days
Contents:

AWS S3 / Storage Tiers
AWS EC2
EBS
EFS

Glue and Athena

Program Duration: 3 days
Contents:

AWS Glue - Architecture

AWS Glue - Architecture
AWS Glue - Terminology
AWS Glue - Applications
AWS Glue - Internals
AWS Glue - Cost
Lab: AWS Glue - Security and Privileges Setup
AWS Glue - Advance Network Configuration
Lab: AWS Glue - Advance Network Configuration
AWS Glue - Data Catalogue
Lab: AWS Glue - Databases
AWS Glue - Tables
AWS Glue - Designing Tables

AWS Glue - Introduction to Crawlers

Lab - Introduction to AWS Glue Classifiers
Lab 1 - AWS Glue - Developing Data Catalog with Crawlers
Lab 2 - AWS Glue - Developing Data Catalog with Crawlers
Lab 3 - AWS Glue - Developing Data Catalog with Crawlers
Lab 4 - AWS Glue - Developing Data Catalog with Crawlers
Lab 5 - AWS Glue - Developing Data Catalog with Crawlers
Lab 6 - AWS Glue - Developing Data Catalog with Crawlers
Lab 7 - AWS Glue - Developing Data Catalog with Crawlers

Introduction to AWS Glue Jobs

Lab 1 - Developing AWS Glue Jobs
AWS Glue Job Properties
Lab 2 - Developing AWS Glue Jobs
Lab 3 - Assignment: Importing Data from Redshift
Lab 4 - Developing AWS Glue Jobs
AWS Glue Job Scripts and Properties
Lab 5 - Developing AWS Glue Jobs
AWS Glue - Built-in ETL Transformations and Job Bookmarks

Introduction to AWS Glue Triggers

Lab 1 - Developing AWS Glue Triggers
Lab: Creating a AWS Glue Development Endpoint
Lab: Installing and configuring Apache Zeppelin
Lab: Port Forwarding Configuration
Lab: Integrating AWS Glue Development Endpoint with Apache Zeppelin
AWS Glue Monitoring

Athena

- AWS Athena - Architecture
- AWS Athena - Features
- AWS Athena - Object Model
- Lab 1 - Developing Data Catalog with AWS Athena
- Lab 2 - Developing Data Catalog with AWS Athena
- AWS Athena - Data Types and DDL Statements
- AWS Athena - SerDe
- Lab 3 - AWS Glue - Developing Data Catalog with Athena
- AWS Athena - Querying AWS Logs
- AWS Athena - Limitations

Athena New Features

- Athena releases support for Views
- Data Lake Solution uses Athena for data analysis
- Athena supports Creating Tables using results of Select Query
- Athena supports resource based policies in AWS Glue Data Catalog
- Athena introduces Workgroups to manage Workloads
- Athena supports resource tagging
- Athena supports AWS Lake Formation for fine-grained permissions

Real Time Analytics / AWS Streams

Program Duration: 0.5 day
Contents:

Kinesis Streams
Kinesis Firehouse
Kinesis Analytics

Data Access

Program Duration: 0.5 day
Contents:

AWS Glue Crawlers / Catalogs
Athena

AWS Databases (RedShift)

Program Duration: 3 days
Contents:

Foundation

- Redshift Architecture
- Use cases
- Features
- Hardware configuration
- Pricing
- Security
- Availability & Fault Tolerance
- Limitations
- Columnar Storage
- Why Redshift is so fast
- Usage with other AWS Services

Database design
Upload and Unloading data
DML operations
Execution Query Plan

Work Load Management
Admin Queries
Redshift Spectrum
Data Share
Query Federation
Materialized Views
Redshift Procedures

Data Migration Utilities on AWS

Program Duration: 1 day
Contents:

AWS DMS
AWS SCT
AWS Snowball
Aws Data Sync

Data Bricks

Program Duration: 3 days
Contents:
Introduction

- Overview of Big Data Architectures
- Top-down vs bottom-up
- What is Databricks?

Databricks concepts

- Workspace
- Interface
- Data Management
- Computation Management
- Model Management
- Authentication and Authorization

Apache Spark

- What is Apache Spark?
- Spark Architecture
- What is Ecosystem of Apache Spark?
- Data Frames and Datasets

Databricks development and Deployment

- Collaborative Workspace
- Perform ETL Operations
- Deploy production jobs and workflows
- Optimized databricks runtime engine

Databricks Jobs & Cluster

- Introduction to Jobs and Cluster
- General Spark Cluster Architecture
- How to Submit Jobs using Job Cluster?
- Pool in Databricks
- Azure Databricks Integration with AAD
- Clusters: Auto Scaling and auto termination

Databricks Data Lake

- Data lake defined
- Hadoop as the data lake

Modern data warehouse

- Federated querying
- Solution in the cloud
- SMP Vs MPP

Data Process & Compute Services

Program Duration: 3 days

Contents:

Airflow

Step functions