Dimensional Data Modelling for Developers with AWS

Rev 4a (July 2025)

This comprehensive training program is designed to equip developers and engineers with in-depth knowledge and practical skills in dimensional data modeling. Participants will explore foundational concepts, advanced modeling techniques, and the integration of ETL/ELT processes. The curriculum emphasizes hands-on experience with modern tools and real-world scenarios, ensuring attendees can design scalable, efficient, and compliant data models aligned with industry best practices.

Prerequisites:

- Foundational Understanding of Database
- Solid understanding of SQL (basic select, joins, aggregations)
- Basics of AWS

Duration: 5 Days

Module 1: Introduction to Data Warehousing

- Purpose and components of data warehouses (Amazon Redshift, Amazon S3, AWS Glue)
- Differences between OLTP and OLAP systems (Amazon RDS for OLTP, Amazon Redshift for OLAP)
- Data warehouse architecture and concepts (Redshift, S3 for data lake zone, Glue for staging and processing)

Module 2: Fundamentals of ETL and ELT Processes

- Basics of ETL (Extract, Transform, Load) and ELT (Extract, Load, Transform) (AWS Glue, Athena, Lambda)
- Key differences between ETL and ELT (Glue for ETL, Athena for ELT on S3)
- Common tools and technologies in ETL/ELT workflows (AWS Glue, AWS Lambda, Amazon MWAA, AWS Data Pipeline, Amazon DMS)

Module 3: Core Concepts of Dimensional Data Modeling

- Introduction to dimensional modeling (Amazon Redshift)

- Fact tables: types and characteristics (Redshift, S3 as source)
- Dimension tables: types and roles (Redshift, Glue Catalog)
- Grain definition and its impact on data modeling (Redshift table design, distribution style and sort keys)

Module 4: Schema Design Techniques

- Star schema vs. Snowflake schema: advantages and trade-offs (Amazon Redshift)
- Denormalization strategies and when to use them (Redshift, Athena on denormalized S3 data)
- Handling hierarchical data structures (Redshift, use of parent-child keys or bridge tables)

Module 5: Advanced Dimension Modeling

- Slowly Changing Dimensions (SCD) Types 0 through 3 (AWS Glue PySpark Jobs, AWS Glue Bookmarks)
- Junk dimensions, degenerate dimensions, and role-playing dimensions (Amazon Redshift)
- Hierarchical and recursive relationships in dimensions (Redshift, with recursive CTEs or bridge tables)

Module 6: Advanced Fact Table Design

- Transactional, snapshot, and accumulating fact tables (Amazon Redshift)
- Additive, semi-additive, and non-additive measures (Amazon Redshift)
- Handling late-arriving facts and missing data (AWS Glue, Redshift staging tables, MERGE statements)

Module 7: Data Pipeline Integration

- Designing robust ETL/ELT pipelines (AWS Glue, Amazon MWAA, EventBridge)
- Integrating dimensional models into data pipelines (Redshift, Glue Workflows)
- Automating data workflows using orchestration tools (AWS Glue Workflows, AWS Step Functions, Amazon MWAA)

Module 8: Performance Optimization

- Optimizing dimensional models for query performance (Amazon Redshift, Sort/Dist Keys, Materialized Views)

- Implementing indexing, partitioning, and materialized views (Redshift, Redshift Spectrum, Athena on partitioned S3 datasets)
- Query optimization techniques (Redshift Query Plan, Amazon CloudWatch, WLM Queues)

Module 9: Data Quality and Governance

- Ensuring data quality and consistency throughout the pipeline (AWS Glue Data Quality)
- Data validation and cleansing techniques (AWS Glue, Amazon Lambda for lightweight validation)
- Metadata management and data lineage (AWS Glue Data Catalog, AWS Lake Formation)

Module 10: Implementing a Data Extraction Solution

- Introduction to Incremental ETL (AWS Glue Bookmarks, Athena with partition filtering, Amazon DMS)
- Extracting Modified Data (AWS DMS, S3 + Glue Jobs, CDC logic in Glue or Redshift)

Module 11: Hands-on Project

- Designing and implementing a dimensional data model for a sample business process (Amazon Redshift, S3, Glue, Glue Catalog)
- Building and integrating ETL/ELT pipelines (AWS Glue, S3, Athena, MWAA)
- Applying performance optimization and data governance techniques (Redshift tuning, Glue Data Quality, Lake Formation, QuickSight)