BigQuery for Big data engineers

Duration: 4 days (32 hours)

What you'll learn

• Learn Full In & Out of Google Cloud BigQuery with proper HANDS-ON examples from scratch.
• Get an Overview of Google Cloud Platform and a brief introduction to the set of services it provides.
• Start with Bigquery core concepts like understanding its Architecture, Dataset, Table, View, Materialized View, Schedule queries, Limitations & Quotas.
• ADVANCE Big query topics like Query Execution plan, Efficient schema design, Optimization techniques, Partitioning, Clustering, etc.
• Build Big data pipelines using various Google Cloud Platform services - Dataflow, Pub/Sub, BigQuery, Cloud storage, Beam, Data Studio, Cloud Composer/Airflow.
• Learn to interact with Bigquery using Web Console, Command Line, Python Client Library etc.
• Learn Best practices to follow in Real-Time Projects for Performance and Cost saving for every component of Big query.
• Bigquery Pricing models for Storage, Querying, API requests, DMLs and free operations.
• Data-sets and Queries used in lectures are available in resources tab. This will save your typing efforts.

Day 1: Introduction to BigQuery Basics

Introduction to BigQuery (1.5 hours)

- Conventional Data Warehouse Problems.
- What is BigQuery?
- BigQuery Out-of-the-Box Features.
- Architecture of BigQuery.

Dataset and Table Creation (2 hours)

- Setting up a GCP account.
- Creating a Project.
- BigQuery UI Tour.
- Region vs Multi-region Datasets.
- Creating a Dataset.
- Creating a Table.

Using BigQuery Dashboard Options (1.5 hours)

- Caching features and limitations.
- Querying Wildcard Tables and their limitations.
- Schedule, Save, and Share a Query.
- Schema Auto-detection.
- Exploring Data Canvases.
- Using the Query Scheduler and Data Forms.

SQL for BigQuery (1 hour)

- Core SQL syntax and functions.
- Optimizing queries for performance.
- Advanced functions and features (e.g., window functions, arrays, structs).
- Introduction to BigQuery Stored Procedures.

Disaster Recovery (1 hour)

- Automated recovery features for BigQuery datasets and queries.

Day 2: Advanced Data Management

Efficient Schema Design in BigQuery (1 hour)

- Designing efficient schemas for BigQuery tables.
- Nested and repeated columns.

Operations on Datasets and Tables (1 hour)

- Copying datasets.
- Transfer service for scheduling copy jobs.
- Native schema modification and manual operations on tables.

Execution Plans in BigQuery (1 hour)

- Understanding how BigQuery creates execution plans for queries.
- Analyzing execution plans in the BigQuery UI dashboard.

Partitioned Tables in BigQuery (2 hours)

- Partitioning concepts and benefits.
- Types: Ingestion time, Date column, Integer-based partitioning.
- ALTER and COPY operations on partitioned tables.
- Best practices for partitioning.

Clustered Tables in BigQuery (2 hours)

- Clustering concepts and when to use clustering vs partitioning.
- Creating clustered tables and best practices.

Querying External Data Sources (1 hour)

- Creating a Cloud Storage bucket.
- Querying permanent tables linked to external data sources.
- Limitations of external data source querying.

Day 3: BigQuery Command-Line Tool and Machine Learning

BigQuery Command-Line Tool (bq) (2 hours)

- Installing and configuring the bq command-line tool.
- Running queries, managing datasets, and scripting tasks.

BigQuery ML (BQML) Basics (2 hours)

- Introduction to BQML: What is BQML and its use cases?
- Supported ML models and their advantages.
- Limitations of BQML.

Creating and Training Models (2 hours)

- Using SQL to train ML models (e.g., linear regression, logistic regression).
- Evaluating model performance with metrics.
- Hyperparameter tuning for optimization.

Using Trained Models (1 hour)

- Generating predictions using SQL.
- Evaluating models with new data.
- Exporting models for external use.

Day 4: Advanced Topics and Latest Updates

BigQuery ML with Advanced Techniques and Gemini 2.0

1. **Introduction to Gemini 2.0 (1 hour)**
    - Overview of improvements in BQML.
2. **BigQuery ML for Time Series Forecasting (1.5 hours)**

    - Using Gemini 2.0 for enhanced time series modeling.
    - Practical example: Building ARIMA models.
3. **BigQuery ML for Recommendation Systems (1 hour)**
    - Building recommendation systems with Gemini 2.0.

Integration with TensorFlow and External Frameworks (1 hour)

- Exporting models to TensorFlow.
- Combining TensorFlow models with BigQuery predictions.

BigQuery Performance Optimization (1 hour)

- Latest storage optimization techniques.
- Advanced partitioning and clustering features.
- Using the **BigQuery Analysis Hub**.

BigQuery Security and Access Control (1 hour)

- Updates in IAM roles and permissions.
- Column-level encryption and masking improvements.