+33 (0)1 48 10 09 46
formation@ambient-it.net
140 Avenue Jean Lolive 93500 Pantin - France

Updated on 29/07/2025

Sign up

Greenplum Parallel SQL Training

3 days (21 hours)

Presentation

Master the full power of Greenplum Parallel SQL with this expert course, designed for data engineers, SQL developers and architects wishing to fully exploit Greenplum's massively parallel processing capabilities for large-scale analysis.

The course begins with the fundamentals of Greenplum and the MPP architecture, with a focus on data distribution, partitioning and modeling strategies designed to maximize the engine's native parallelism.

You'll learn how to write high-performance SQL queries in a distributed context: optimized joins, parallel aggregates, advanced analytical functions, and fine-tuned exploitation of execution plans through concrete tuning cases.

Practical modules also cover massive ingestion (COPY, gpload, external tables), distributed updates, the use of UDF, and the integration of Greenplum into a wider BI or data engineering ecosystem.

As with all our training courses, this one will be presented with the latest Greenplum Parallel updates.

Objectives

- Understand Greenplum's massively parallel architecture and its specific distribution and partitioning features.
- Model, distribute and partition data schemas optimized for processing
  SQL
- master methods for writing high-performance SQL queries in a parallel environment
- Be able to ingest massive volumes of data via COPY, gpload or external tables
- Exploit the full potential of the Greenplum engine in concrete BI and data engineering cases.

Target audience

- Data architects
- SQL developers

Prerequisites

- Basic knowledge of SQL language

Greenplum Parallel SQL training program

Introduction to Greenplum and Massively Parallel SQL

- Introduction to Greenplum
- PostgreSQL as a foundation
- Typical use cases
- Segments, masters and interconnections
- Mirroring and fault tolerance
- Differences with SMP and Hadoop
- Key concepts of parallel SQL
- Data partitioning
- Distribution vs. replication
- Distributed SQL query processing

Data modeling for parallelism

- Parallel table creation
- Greenplum-specific SQL syntax
- Distribution parameters
- Comparison with local tables
- Choice of distribution keys
- Unique key vs. frequent key
- Impact on performance
- Practical cases of poor distribution
- Logical partitioning
- Partitioning syntax
- Partitioning by range or list
- Examples of optimized queries with partitions

Parallel SQL queries

- SELECT and parallel aggregation
- Distributed aggregates
- Specific aggregation functions
- Group By on distributed data
- Parallel joins
- Broadcast vs Hash Join
- Optimizing joins on distributed keys
- High Shuffle cost
- Analysis of execution plans
- Using EXPLAIN and EXPLAIN ANALYZE
- Reading distributed plans
- Interpreting intersegment costs

Writing and inserting parallel data

- INSERT, UPDATE, DELETE
- Distributed writing behavior
- Restrictions on UPDATEs with joins
- Bulk sorting and insertion
- COPY and bulk loading
- Using COPY in parallel
- Error and log management
- Loading external files
- Constraints and indexing
- Constraints supported in parallel
- Index bitmap vs. BTree
- Impact of indexes on distributed performance

Analytical and advanced functions

- Window functions
- Use of ROW_NUMBER, RANK, LEAD, LAG
- Partitioning windows on segments
- Typical BI use cases
- Complex expressions and types
- Derived tables, CTE, parallel subqueries
- ARRAY and JSON types in Greenplum
- User functions (UDF/UDT)
- Creating functions in SQL or PL/pgSQL
- Portability from PostgreSQL
- Parallelism considerations

Optimizing and tuning parallel SQL queries

- Statistics and analysis
- Using ANALYZE on distributed tables
- Global vs. local statistics
- Greenplum performance analysis tools
- Rewriting and intelligent indexing
- Redesigning costly queries
- Using indexes on join columns
- Rewriting case studies
- Transaction optimization
- Isolating and managing locks
- Locks in parallel environments
- Best practices for concurrent writes

Integration, import/export and external access

- External tables
- Declaration of flat files or HDFS
- Integration with gpload and gpfdist
- Parallel loading of large volumes
- gpload and gpexpand
- YAML file configuration for gpload
- Add new segments dynamically
- Extend a cluster without redeployment
- Access from external tools
- Connection via ODBC/JDBC
- SQL queries from Python, R, Java
- BI compatibility (Tableau, Power BI)

Case studies and final project

- BI case study (high-volume reporting)
- Creating distributed schemas
- Optimized analytical queries
- Connected dashboard
- Data Engineering case study
- Massive data ingestion with partitioning
- SQL pipelines with complex joins
- Performance monitoring
- Synthesis project
- Choice of distribution strategy
- Query, load, analyze
- Performance reporting

Companies concerned

This training course is aimed at both individuals and companies, large or small, wishing to train their teams in a new advanced computer technology, or to acquire specific business knowledge or modern methods.

acquire specific business knowledge or modern methods.

Positioning on entry to training

Positioning at the start of training complies with Qualiopi quality criteria. As soon as registration is finalized, the learner receives a self-assessment questionnaire which enables us to assess his or her estimated level of proficiency in different types of technology, as well as his or her expectations and personal objectives for the forthcoming training course, within the limits imposed by the selected format. This questionnaire also enables us to anticipate any connection or security difficulties within the company (intra-company or virtual classroom) which could be problematic for the follow-up and smooth running of the training session.

Teaching methods

Practical training: 60% hands-on, 40% theory. Training material distributed in digital format to all participants.

Organization

The course alternates theoretical input from the trainer, supported by examples, with brainstorming sessions and group work.

Validation

At the end of the session, a multiple-choice questionnaire verifies the correct acquisition of skills.

Certification

A certificate will be awarded to each trainee who has completed the entire course.