SERVERLESS ARCHITECTURE

What is it?

Serverless computing is a cloud execution model where providers dynamically manage infrastructure while developers focus solely on business logic[1]. Applications execute as event-triggered functions that automatically scale based on demand with pay-per-use pricing, eliminating server provisioning and maintenance overhead[1][2].

Decoupled Storage and Compute

Serverless platforms separate storage from compute, enabling independent scaling of each component[1]. Unlike traditional clusters that tightly couple these resources, serverless systems persist storage independently while compute allocates dynamically only when needed, then scales to zero during idle periods[1][2]. BigQuery demonstrates this: data resides in Colossus storage while Dremel compute workers allocate only during query execution, eliminating idle infrastructure costs[2].

Event-Driven Architecture

Serverless excels at event-driven patterns where functions trigger on file uploads, API requests, database changes, or scheduled events[2]. This enables asynchronous processing, loose coupling between services, and elastic scaling where each event spawns compute instances as needed.

Serverless Data Warehousing

BigQuery enables querying petabyte-scale datasets without cluster management[1]. The Dremel engine parallelizes SQL across thousands of workers with automatic resource allocation. Google's Jupiter network enables rapid data movement between storage and compute, eliminating data locality constraints. Users pay only for data processed, contrasting with traditional warehouses requiring continuously-running clusters[1].

Challenges

Serverless introduces specific trade-offs: **cold start latency** when functions scale from zero, **execution time limits** (typically 5-15 minutes) requiring workload decomposition, **vendor lock-in** as implementations vary across providers, and **debugging complexity** in distributed ephemeral environments[1][2]. Despite these constraints, serverless reduces operational overhead and enables cost-efficient scaling for event-driven workloads.

References

[1] Jonas, E., Schleier-Smith, J., Sreekanti, V., et al. (2019). Cloud Programming Simplified: A Berkeley View on Serverless Computing. UC Berkeley Technical Report UCB/EECS-2019-3.

[2] Google Cloud (2020). Serverless at Scale: From Design to Production. Google Cloud Whitepaper.