About AtScale Virtual Cubes

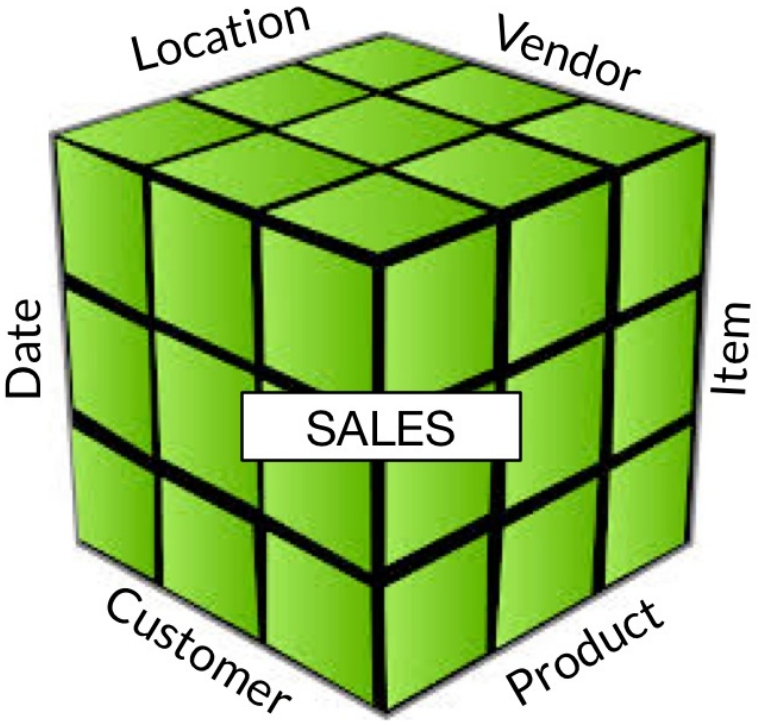
An AtScale virtual cube is a metadata layer that overlays a multi-dimensional cube format on top of the datasets stored in a connected data warehouse, such as Google BigQuery or a Hadoop cluster. The cube is virtual because the data is not moved or processed up front. Instead, the cube contains the logic about how to process and optimize the data at query runtime.

The Purpose Of An OLAP System

The basic purpose for having an Online Analytical Processing (OLAP) system is to allow people to access data, ask questions, and get answers quickly. OLAP is the foundation of business intelligence (BI) - the broader set of tools and methodologies for gaining meaningful insights from raw transactional data.

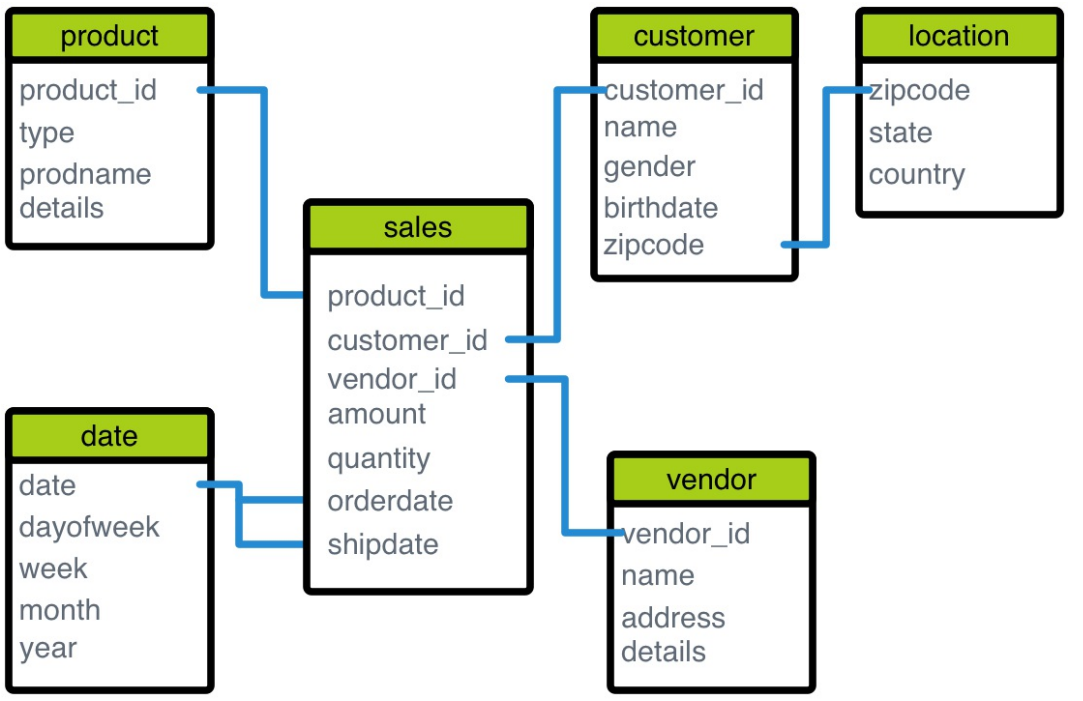
When describing OLAP systems, there are two major approaches in how the data is optimized to support BI analysis - MOLAP (multi-dimensional OLAP) and ROLAP (relational OLAP).

In the MOLAP approach, aggregate (measure) data is pre-calculated for every dimension combination. The result is a multi-dimensional cube, where each cell in the cube represents an intersection of n dimension values.



The benefit of building MOLAP cubes is performance. Since measures are calculated for every possible combination of dimensions ahead of time, slicing and dicing of multi-dimensional data is very fast. But this performance comes at a cost - redundant storage, long build times, time-to-insight latency, and the administration overhead of managing data silos. Plus, when dealing with the size of dimensions in big data, MOLAP cubes cannot scale. Cube build times only increase as the data grows. The latency between when data lands, and when it is ready for analysis can prevent the business from making timely decisions.

In the ROLAP approach, data is not pre-computed ahead of time. Instead data is stored in relational tables that use a star (or snowflake) schema to model multi-dimensional data.



Analysis tools then query these tables using query languages such as SQL, and aggregate results are calculated on-the-fly for the requested measures and dimensions. The benefits of this approach is that it is more scalable in handling large data volumes, and does not require off-loading the data into another storage system for analysis. But ROLAP engines are usually slower when the data is not aggregated ahead of time. To get around this performance hit, it is often necessary to build separate summary tables that contain pre-aggregated data in order to boost query performance.

What Is An AtScale Virtual Cube?

AtScale virtual cubes are a hybrid of the MOLAP and ROLAP approaches. The AtScale platform combines the scalability of ROLAP with the ease-of-use of MOLAP-like data modeling.

Logically, the cube looks like a MOLAP cube to the business intelligence applications. The MOLAP model is easy for BI users to understand, because the data is presented as a simple list of measures and dimensions that can be used to build reports.

However, AtScale does not build MOLAP cubes. AtScale overlays a virtual ROLAP schema on top of the datasets stored in Hive.

How AtScale Optimizes OLAP Queries

BI applications send their queries to a cube hosted on the AtScale engine. This cube metadata is used to interpret the SQL queries sent by the BI tools, optimize them for the best performance, and then execute them directly against the corresponding data warehouse.

AtScale's cost-based query planner and optimizer dynamically builds and maintains aggregates (summary tables) based on the queries issued by BI users. Once aggregates exist, future queries can run against the aggregated data instead of the raw data, dramatically improving query performance.