What is Amazon Redshift?

By Rich Morrow
January 6, 2015

Like any "big data" initiative, deploying and operating a data warehouse of any size used to be limited to only large **enterprises with deep budgets for proprietary hardware and multi-year software licenses. Pay-as-you-go cloud products** like Google's BigQuery and Amazon Redshift change all of that, putting a fully blown, fully managed data warehouse **within reach of even the smallest business. This article addresses what Amazon Redshift is (and is not).**

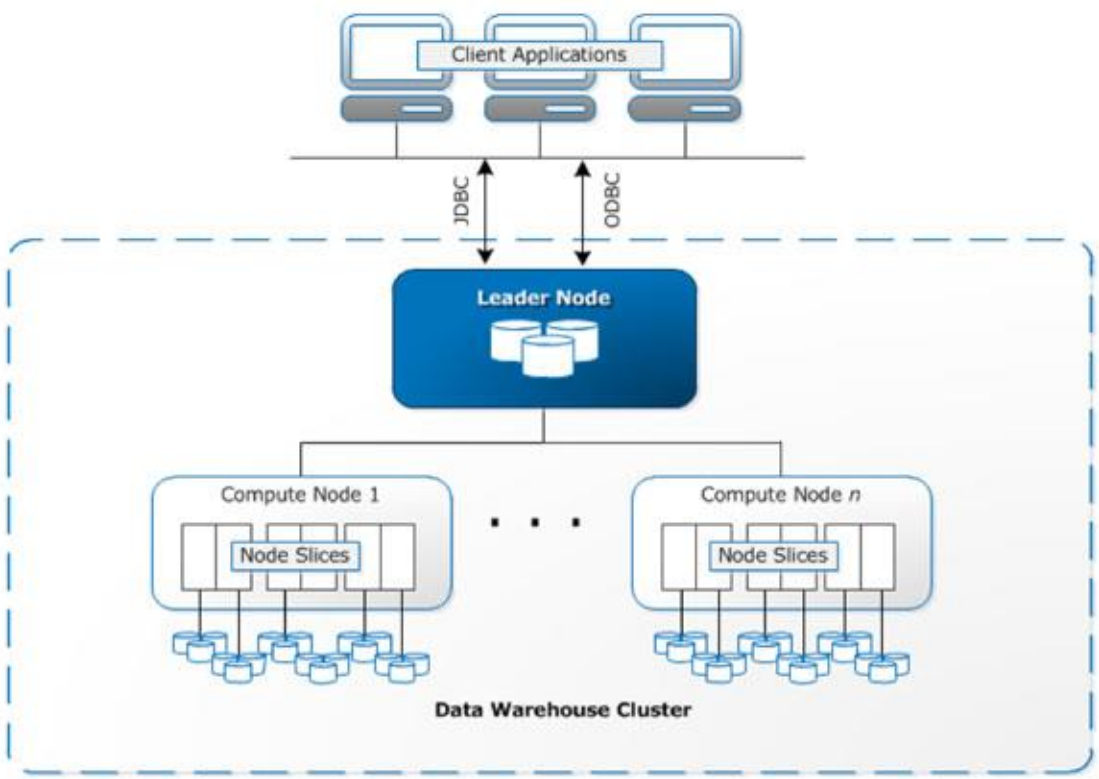
Perhaps one of the most exciting outcomes of the public cloud was addressing the shortcomings of traditional enterprise data warehouse (EDW) storage and processing. The fast provisioning, commodity costs, infinite scale, and pay-as-you-grow pricing of public cloud are a natural fit for EDW needs, providing even the smallest of users the ability to now get valuable answers to BI questions. Amazon Redshift is one such system built to address EDW needs, and it boasts low costs, an easy SQL-based access model, easy integration to other Amazon Web Services (AWS) solutions, and most importantly, high query performance.

Amazon Redshift gets its name from the astronomical phenomenon noticed by Hubble, which explained the expansion of the universe. By adopting the Amazon Redshift moniker, AWS wanted to relay to customers that the service was built to handle the perpetual expansion of their data.

An Amazon Redshift cluster consists of one leader node (which clients submit queries to) and one or more follower (or "compute") nodes, which actuall**y perform the queries on locally stored data. By allowing for unlimited expansion of follower nodes, Amazon Redshift ensures that customers can continue to grow their cluster as their data needs grow.** Customers can start with a "cluster" as small as a sing**le node (acting as both leader and follower), and for the smallest** supported instance type (a DW2), that could be as low cost as $0.25/hour or about $180/month. By using "Reservations" **(paying an up-front fee in exchange for a lower hourly running cost) for the underlying instances, Amazon Redshift can cost as little as $1,000/TB/year** — **upwards of one-fifth to one-tenth of the cost of a traditional EDW.**

Because Amazon Redshift provides native Open Database Connectivity (ODBC) and Database Connectivity (JDBC) connectivity (in addition to PostgresSQL driver support), most third-party BI tools (like Tableu, Qlikview, and MicroStrategy) work right out of the box. Amazon Redshift also uses the ubiquitous Structured Query Language (SQL) language for queries, ensuring that your current resources can quickly and easily become productive with the technology.

Amazon Redshift Architecture:



Amazon Redshift was custom designed from the ParAccel engine — an analytic database which used columnar storage and parallel processing to achieve very fast I/O. Columns of data in Amazon Redshift are stored physically adjacent on disk, meaning that queries and scans on those columns (common in online analytical processing [OLAP] queries) run very

fast. Additionally, Amazon Redshift uses 10GB Ethernet interconnects, and specialized EC2 instances (with between three and 24 spindles per node) to achieve high throughput and low latency. For even faster queries, Amazon Redshift allows customers to use column-level compression to both greatly reduce the amount of data that needs stored, and reduce the amount of disk I/O.

Amazon Redshift, like many of AWS's most popular services, is also fully managed, meaning that l**ow-level, time-consuming administrative tasks like OS patching, backups, replacing failed hardware, and software upgrades are handled automatically and transparently. With Amazon Redshift, users simply provision a cluster, load it with their data, and begin executing queries. All data is continuously, incrementally, automatically backed up in the highly durable S3, and enabling disaster recovery across regions can be accomplished with just a few clicks. Spinning a cluster up can be as simple as a few mouse clicks, and as fast as a few minutes.**

A very exciting aspect of Amazon Redshift, and something that is not possible in traditional EDWs, is the ability to easily scale a provisioned cluster up and down. In Amazon Redshift, this scaling is transparent to the customer—when a resize is requested, data is copied in parallel from the source cluster (which continues to function in read-only mode) to a new cluster, and once all data is live migrated, DNS is flipped to the new cluster and the old cluster is de-provisioned. This allows customers to easily scale up and down, and each scaling event nicely re-stripes the data across the new cluster for a balanced workload. In a traditional, hosted EDW environment, a resize would typically involve weeks of preparation and days of downtime, along with a hefty six- or seven-figure bill.

Amazon Redshift offers mature, native, and tunable security. Clusters can be deployed into a Virtual Private Cloud (VPC), and encryption of data is supported via hardware accelerated AES-256 (for data at rest) and SSL (for data on the wire). Compliance teams will be pleased to learn that users can manage their own encryption keys via AWS's Hardware **Security Module (HSM) service, and that Amazon Redshift provides a full audit trail of all SQL connection attempts, queries, and modifications of the cluster. The AWS CloudTrail service additionally logs all API calls against Amazon Redshift, and both the native SQL logs and AWS CloudTrail logs can be exported and queried.**

Now that we know what Redshift is, let's also clarify what it is not. It is not a NoSQL engi**ne. It is not a suitable solution to search through large collections of text documents. It is not a relational database management system (RDBMS), nor is it intended to serve online transaction processing (OLTP) for external customers. It is not a real-time analysis engine. It is not a good place to store anything but structured data. It is not the fastest way to analyze data, nor is it the cheapest way**

to store data. Instead, it is a cloud-based EDW that allows internal users to quickly perform business analytics on large collections of both rolled-up and granular data.

To learn more about Amazon Redshift, check out my white paper called An Introduction to Amazon Redshift on the Global Knowledge website.