BDA304

What's New with Amazon Redshift

Vidhya Srinivasan, General Manager, AWS

November 30th, 2016

amazon web services

AWS Big Data Portfolio

Collect


Amazon Kinesis Firehose


Amazon Kinesis Analytics


AWS Direct Connect


Amazon Snowball


Amazon Kinesis Streams

Store


Amazon S3


Amazon Glacier


Amazon CloudSearch


Amazon RDS, Amazon Aurora


Amazon Dynamo DB


Amazon Elasticsearch Service

Analyze


Amazon EMR


Amazon EC2


Amazon Redshift

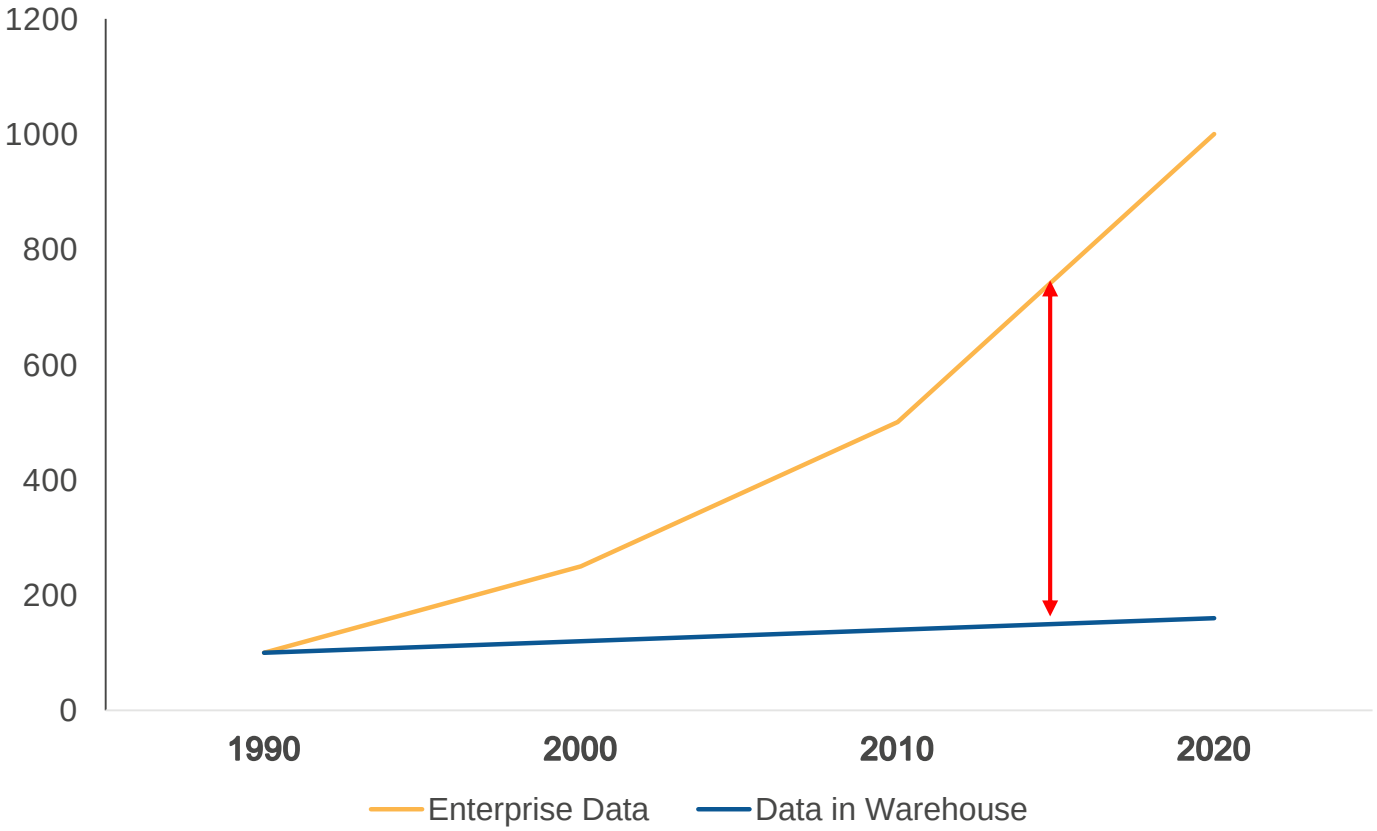

Amazon Machine Learning


Amazon QuickSight


AWS Database Migration Service


AWS Data Pipeline

Legacy architectural models lead to dark data

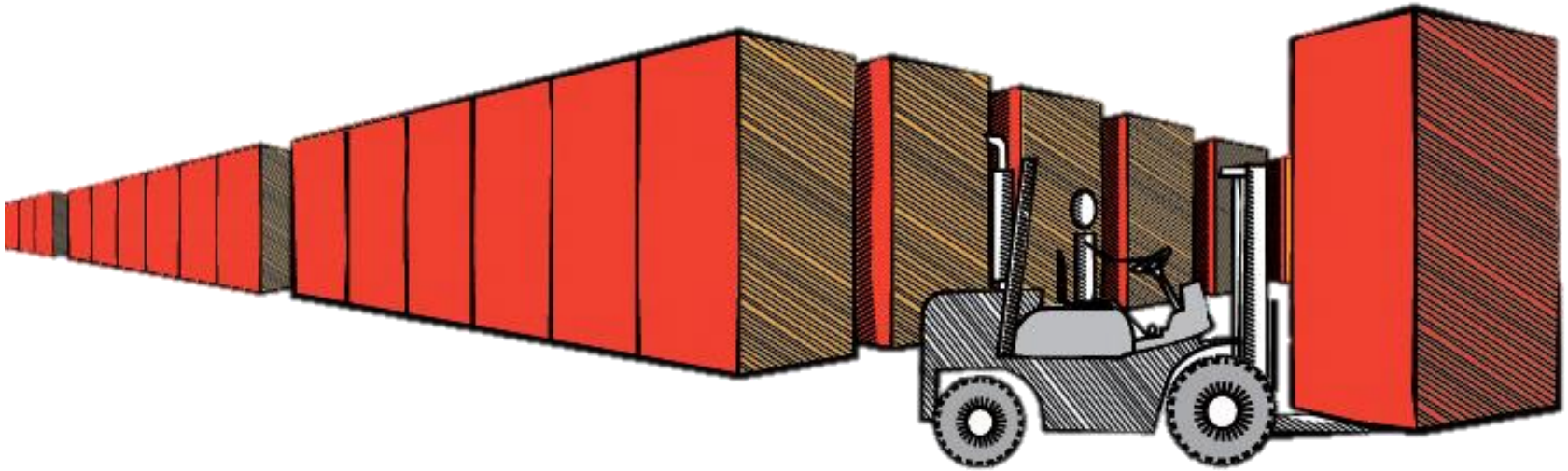


Very Expensive

complicated

Slow at scale

Amazon Redshift



Fast, simple, petabyte-scale data warehousing for $1,000/TB/Year

Amazon Redshift is fast

REDFIN

*"Did I mention that it's **ridiculously fast**? We're using it to provide our analysts with an alternative to Hadoop"*

Periscope DATA

*"After investigating Redshift, Snowflake, and BigQuery, we found that Redshift offers **top-of-the-line performance at best-in-market price points**"*

boingo

*"On our previous big data warehouse system, it took around 45 minutes to run a query against a year of data, but that number went down to **just 25 seconds** using Amazon Redshift"*

Pinterest

*"…[Redshift] performance has blown away everyone here. We generally see **50-100X speedup over Hive**"*

CHANNEL FOUR TELEVISION

*"We **regularly process multibillion row datasets and we do that in a matter of hours**. We are heading to up to 10 times more data volumes in the next couple of years, easily"*

optimum.

*"We saw a **2X performance improvement** on a wide variety of workloads. The **more complex the queries, the higher the performance improvement**"*

Amazon Redshift is easy to use

FOURSQUARE

"With Amazon Redshift and Tableau, ***anyone in the company can set up any queries they like*** *- from how users are reacting to a feature, to growth by demographic or geography, to the impact sales efforts had in different areas"*

vevo

"The doors were blown wide open to ***create custom dashboards for anyone to instantly go in and see and assess what is going in our ad delivery landscape****, something we have never been able to do until now."*

VIVAKI

Provides an easy-to-use mechanism for querying data with quick and uniform response times that analysts can use to run research projects and perform in-depth analysis…We don't have to pre-allocate resources and can easily scale up to meet demand and then scale down for efficiency"

Amazon Redshift is inexpensive

FT FINANCIAL TIMES

"450,000 ***online queries 98 percent faster*** *than previous traditional data center, while* ***reducing infrastructure costs by 80 percent****."*

HAUTELOOK
A NORDSTROM COMPANY

"Annual costs of Redshift are equivalent to just the ***annual maintenance of some of the cheaper on-premises options*** *for data warehouses.."*

kik·

"Most competing data warehousing solutions would have cost us up to $1 million a year. By contrast, Amazon Redshift costs us just $100,000 all-in, representing a total ***cost savings of around 90%****"*

Selected Amazon Redshift customers

Traditional Data Warehousing

Business Reporting

Complex pipelines and queries

Secure and Compliant

Bulk Loads and Updates

Simple Migration – Point & Click using AWS Database Migration Service

Secure & Compliant – End-to-End Encryption. SOC 1/2/3, PCI-DSS, HIPAA and FedRAMP compliant

Large Ecosystem – Variety of cloud and on-premises BI and ETL tools

NTT docomo

Japanese Mobile Phone Provider

SCHOLASTIC

World's Largest Children's Book Publisher

Nasdaq

Powering 100 marketplaces in 50 countries

Log Analysis

Log & Machine IOT Data

Clickstream Events Data

Time-Series Data

Inexpensive – Analyze large volumes of data cost-effectively

Fast – Massively Parallel Processing (MPP) and columnar architecture for fast queries and parallel loads

Near real-time – Micro-batch loading and Amazon Kinesis Firehose for near-real time analytics

Pinterest

Interactive data analysis and recommendation engine

lyft

Ride analytics for pricing and product development

yelp

Ad prediction and on-demand analytics

Business Applications

Multi-Tenant BI Applications

Back-end services

Analytics as a Service

Fully Managed – Provisioning, backups, upgrades, security, compression all come built-in so you can focus on your business applications

Simple Chargeback – Pay as you go, add clusters as needed. A few big common clusters, several data marts

Service Oriented Architecture – Integrated with other AWS services. Easy to plug into your pipeline

Infosys

Infosys Information Platform (IIP)

accenture

Analytics-as-a-Service

Amplitude

Product and Consumer Analytics

Thousands of companies run mission-critical workloads on Amazon Redshift

Redshift is used for mission-critical workloads

Financial and management reporting

Payments to suppliers and billing workflows

Web/Mobile clickstream and event analysis

Recommendation and predictive analytics

Amazon Redshift is available everywhere AWS is

Redshift is fast and has gotten faster...

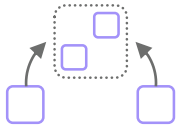
5X Query throughput improvement this year

- Memory allocation (launched)
- Improved commit and I/O logic (launched)
- Queue hopping (launched)
- Query monitoring rules (coming soon)
- Power start (coming soon)
- Short query bias (coming soon)

Fast

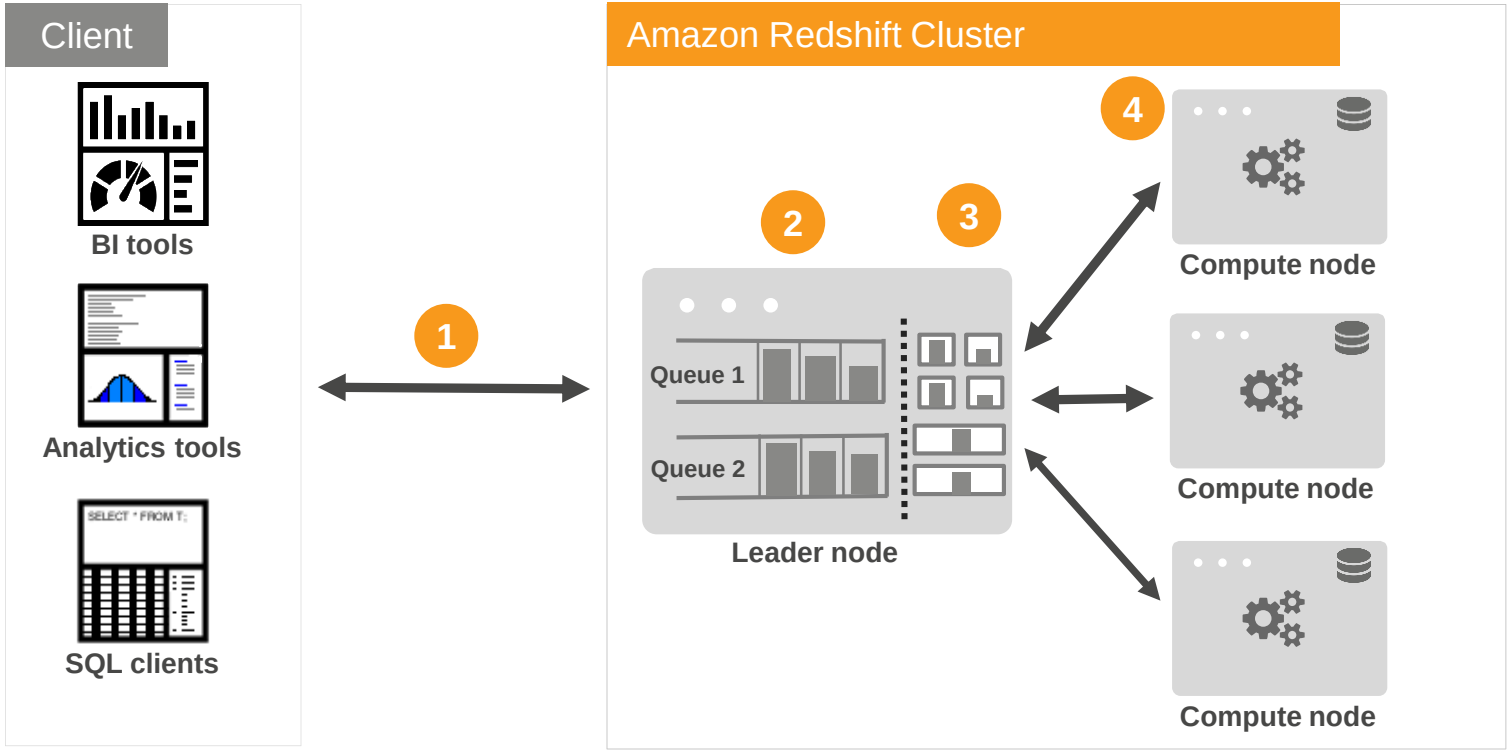
10X Vacuuming performance improvement

- Ensures data is sorted for efficient and fast I/O
- Reclaims space from deleted rows
- Enhanced vacuum performance leads to better system throughput

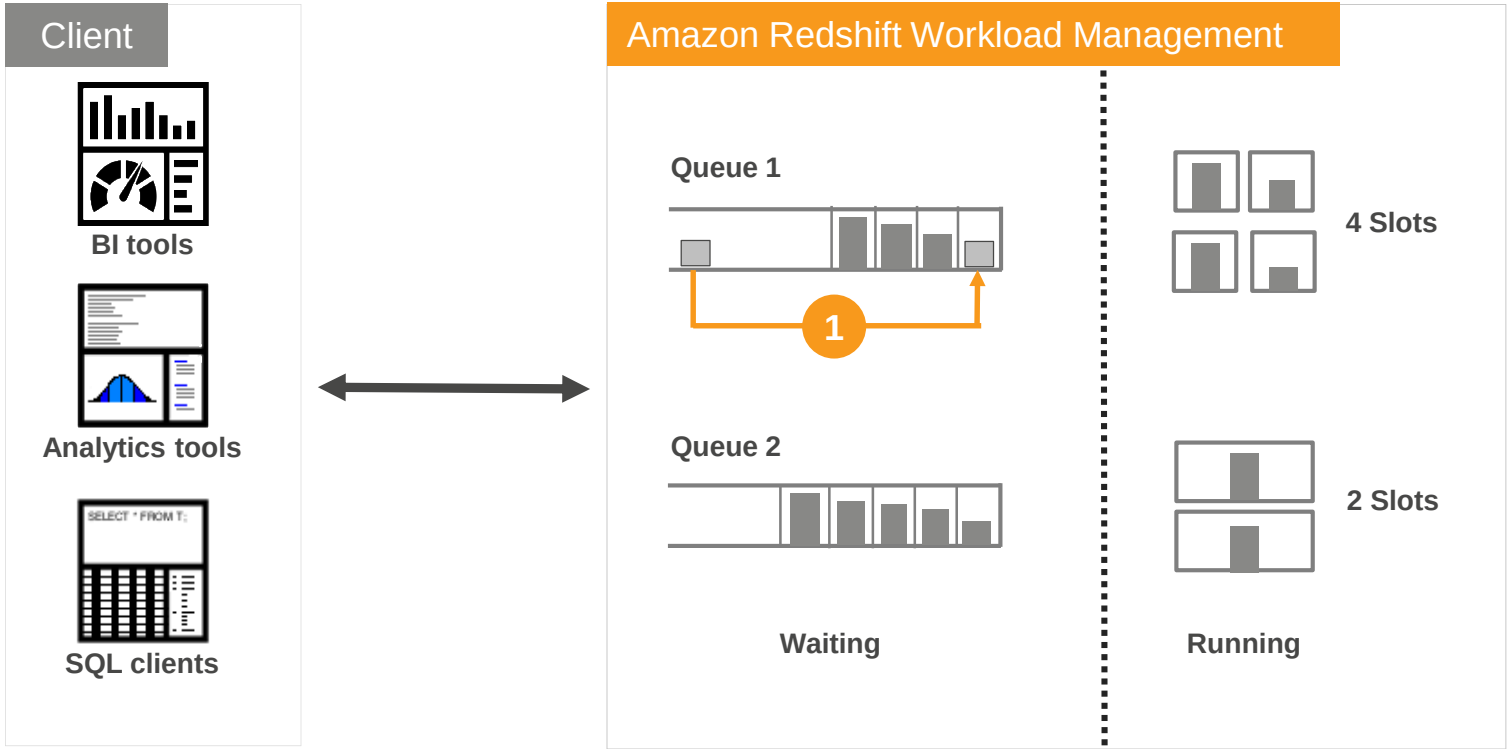


Efficient

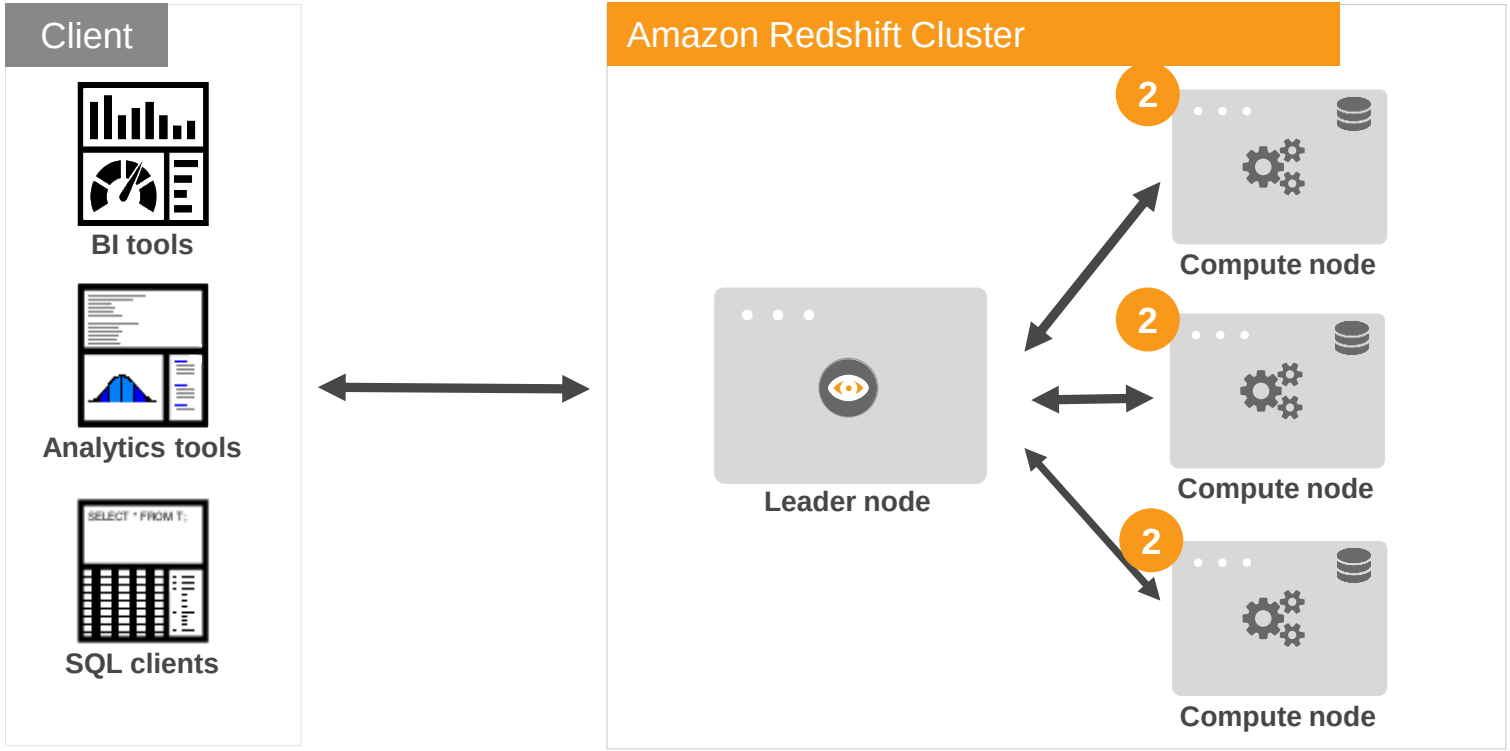
The life of a query



Coming soon: Short query bias

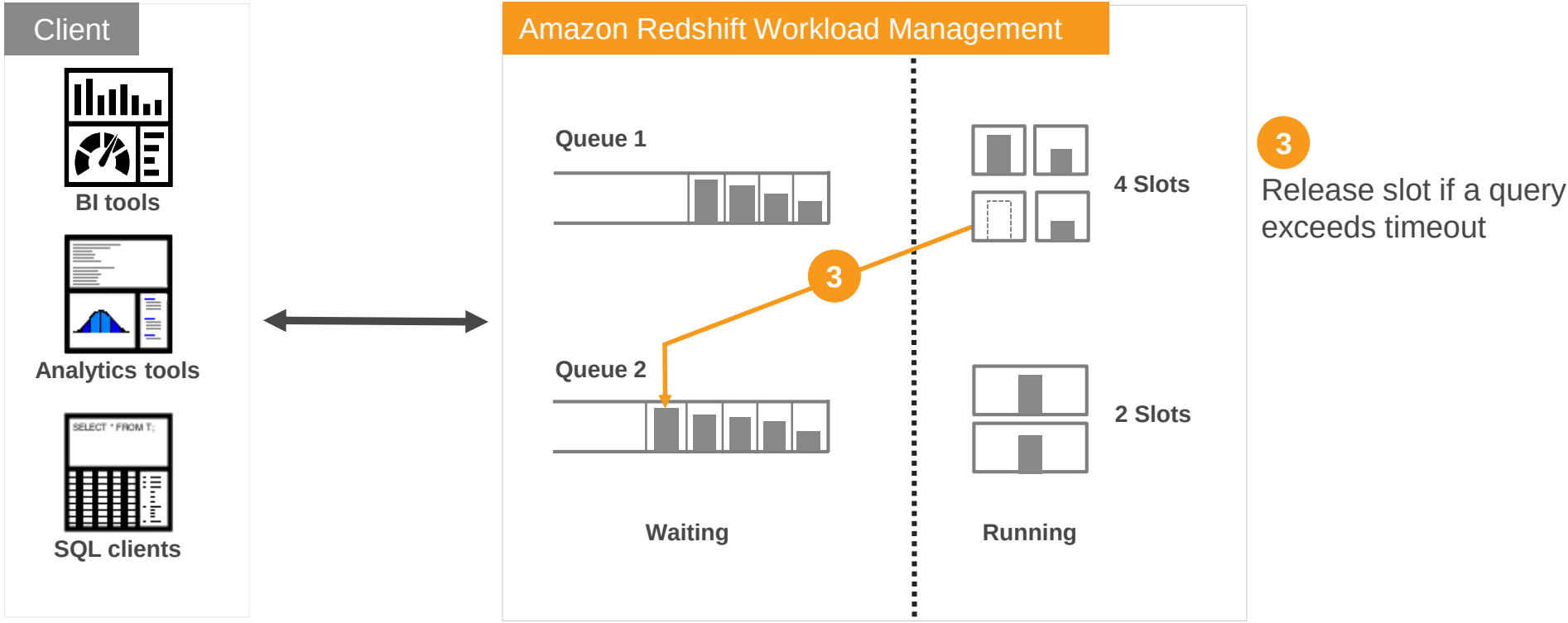


Coming soon: Power start



2 All queries receive a power start. Shorter queries benefit the most

Queue hopping



Coming soon: Query monitoring rules

- Allows automatic handling of runaway (poorly written) queries
- Metrics with operators and values (e.g. query_cpu_time > 1000) create a *predicate*.
- Multiple predicates can be AND-ed together to create a *rule*.
- Multiple rules can be defined for a queue in WLM. These rules are OR-ed together.

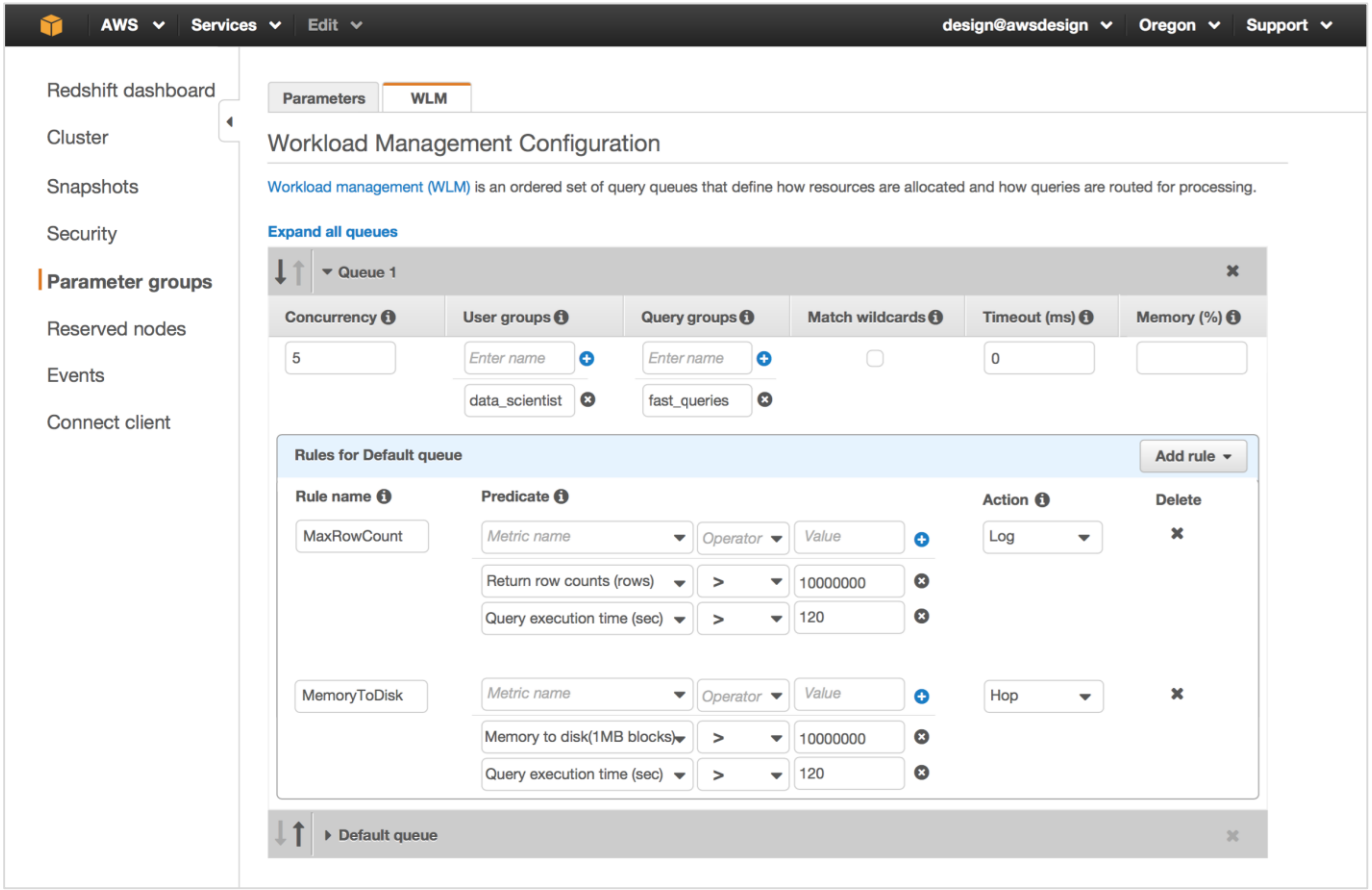
If { *rule* } then [*action*]

{ rule : metric operator value } eg: rows_scanned > 100000

- Metric : cpu_time, query_blocks_read, rows scanned, query execution time, cpu & io skew per slice, join_row_count, etc.
- Operator : <, >, ==
- Value : integer

[*action*] : hop, log, abort

Coming soon: Query monitoring rules



Monitor and control cluster resources consumed by a query

Get notified, abort and reprioritize long-running / bad queries

Pre-defined templates for common use cases

Amazon Redshift is easy to use

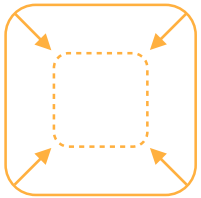
Provisioning in minutes

Automatic Patching

SQL - Data loading

Backups built-in

Security is built-in



Compression is built-in

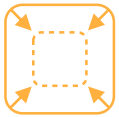
And is getting easier…

Automatic and incremental background VACUUM (coming soon)

- Reclaims space and sorts when Redshift clusters are idle
- VACUUM is initiated when performance can be enhanced
- Improves ETL and query performance

Automatic data compression for CTAS

- CREATE TABLE AS (CTAS) command creates a new table
- The new table leverages compression automatically



Automatic compression for new tables (coming soon)

- All newly created tables will leverage default encoding
- Provides higher compression rates

And is getting easier…

Enhanced localization

- Support for Timestamp with Time zone : New TIMESTAMPTZ data type to input complete timestamp values that include the date, the time of day, and a time zone. Eg: *30 Nov 07:37:16 2016 PST*
- Support for Multi-byte (UTF-8) characters for tables, columns, and other database object names

Resource management

- Connection limits: You can now set a limit on the number of database connections a user is permitted to have open concurrently
- Query Monitoring rules (coming soon)

Table level restore

- Restore tables that you might have dropped accidentally
- Reconcile data from an older table that you might have updated or deleted unintentionally

Coming soon: Ingestion improvements

Faster migrations

- Schema Conversion Tool
  - Already supports schema conversion from Oracle, Teradata, Netezza and Greenplum
  - Coming soon: Support for Vertica and SQL Server
- AWS Database Migration Service
  - Supports data migration with CDC from Oracle, SQL Server, PostgreSQL, MySQL and Aurora
  - Will add more DW engines as sources

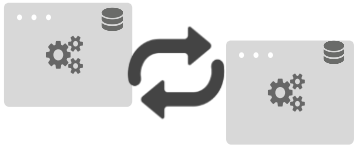
Data Loading

- Enhanced JSON & AVRO ingestion performance

Durability and availability

Automated Backups

Cross-region backups



Cluster-level mirroring

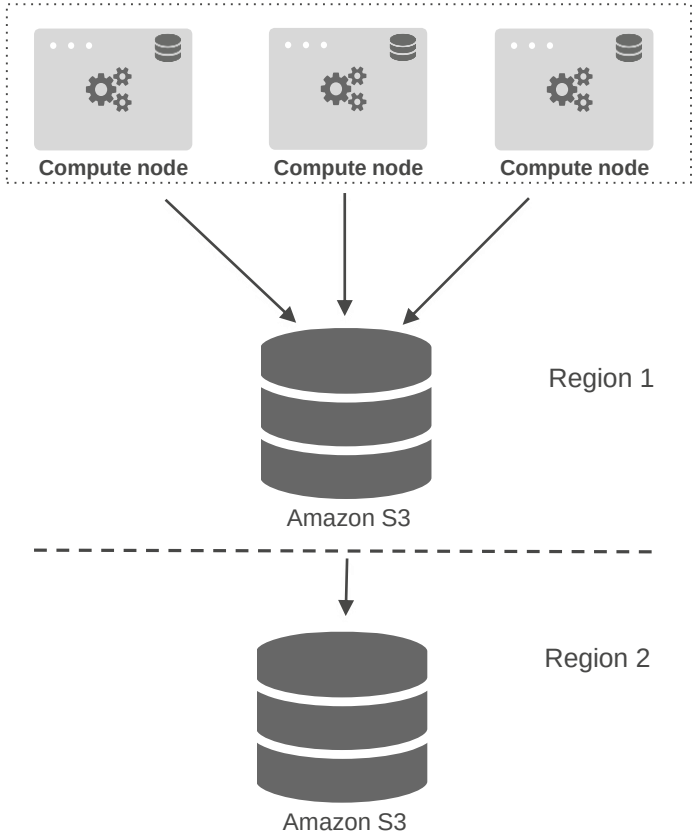
Streaming restore

Monitoring

Ticketing

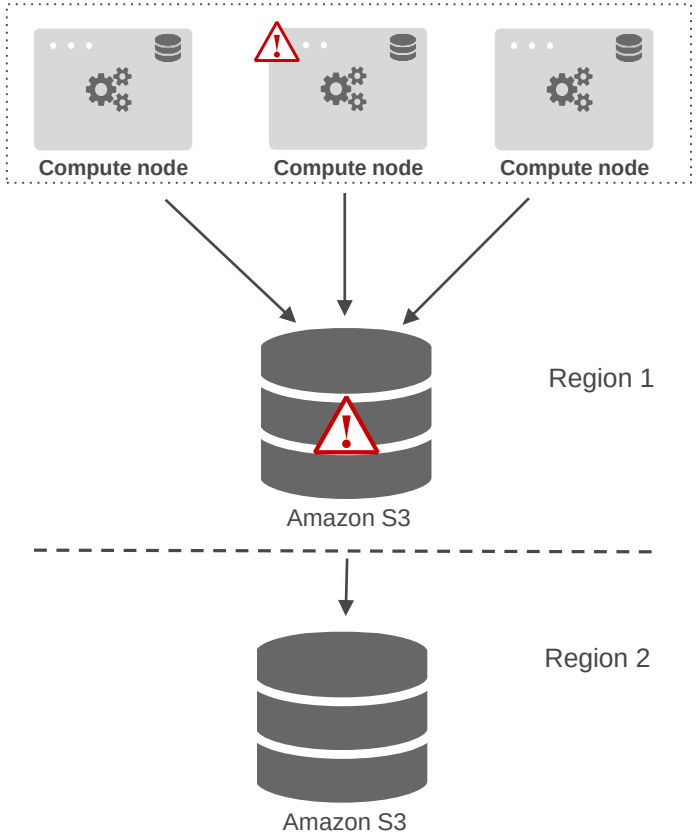
Durability

1. Multiple copies of a block within the cluster
2. Continuous and incremental backups to Amazon S3 (automatic)
3. Continuous and incremental backups across regions (checkbox enable)
4. Streaming restore – database is available for query once metadata restored. Data streams down in background or when queried. Enables cross-AZ restores.



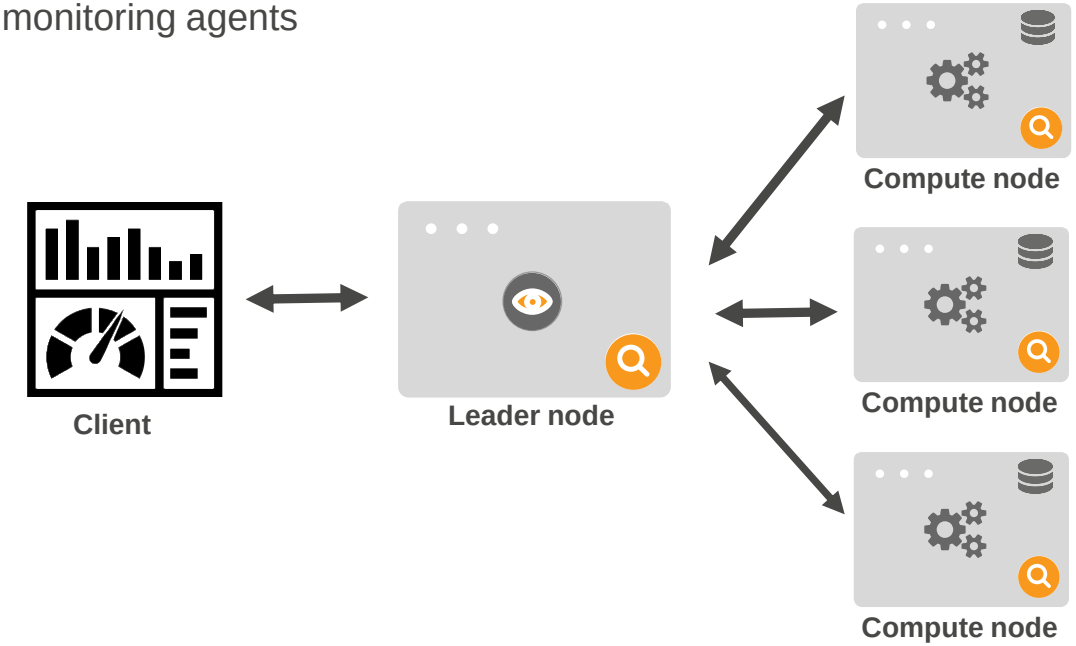
Fault tolerance

1. Disk failures
2. Availability Zone failures
3. Region level failures
4. Node failures



Node fault tolerance

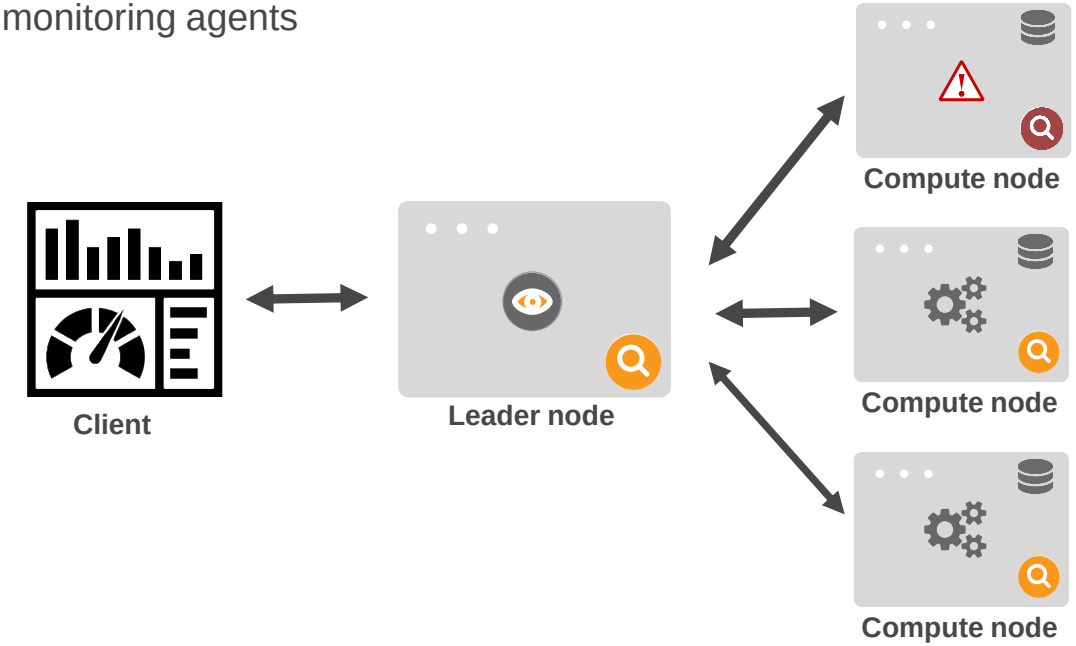
Data-path monitoring agents



Node level monitoring can detect SW/HW issues and take action

Node fault tolerance

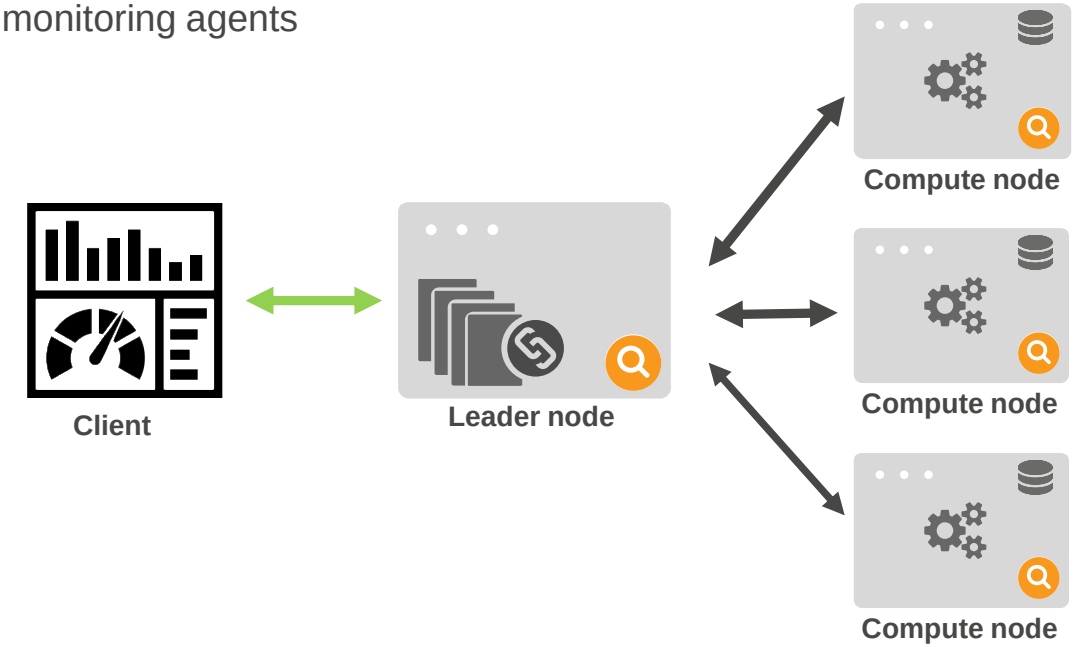
- Data-path monitoring agents



Failure is detected at one of the compute nodes

Node fault tolerance

Data-path monitoring agents

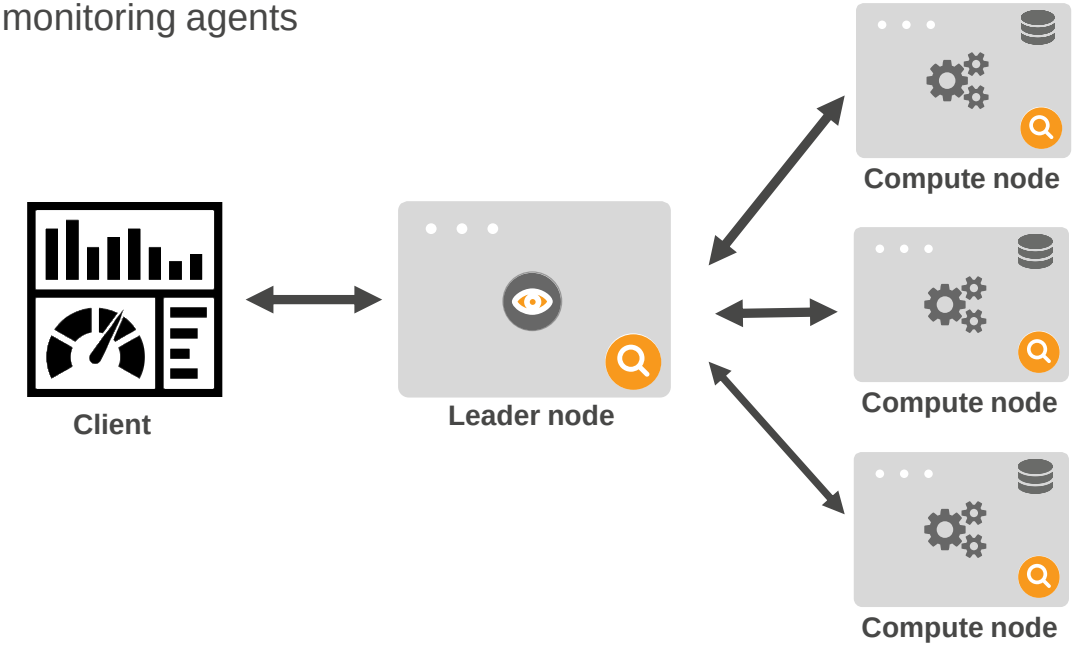


Redshift parks the connections

Next, the node is replaced

Node fault tolerance

Data-path monitoring agents

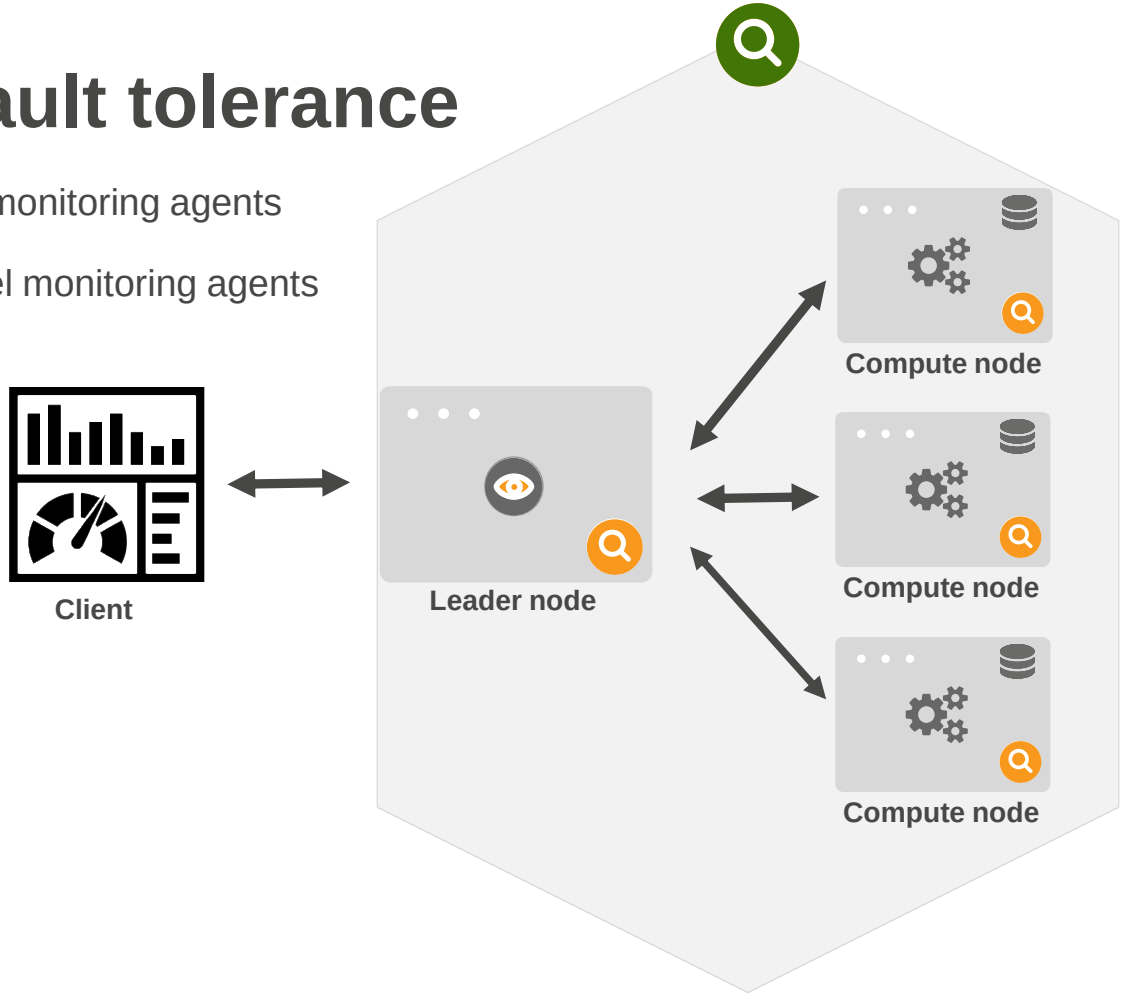


Queries are re-submitted (coming soon)

Node fault tolerance

Data-path monitoring agents

Cluster-level monitoring agents



Additional monitoring layer for the leader node and network

Amazon Redshift is secure

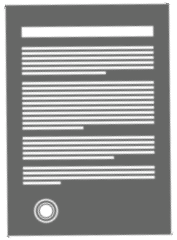
End-to-End
data encryption

Alerts and notifications

Amazon VPC

AWS KMS and HSM

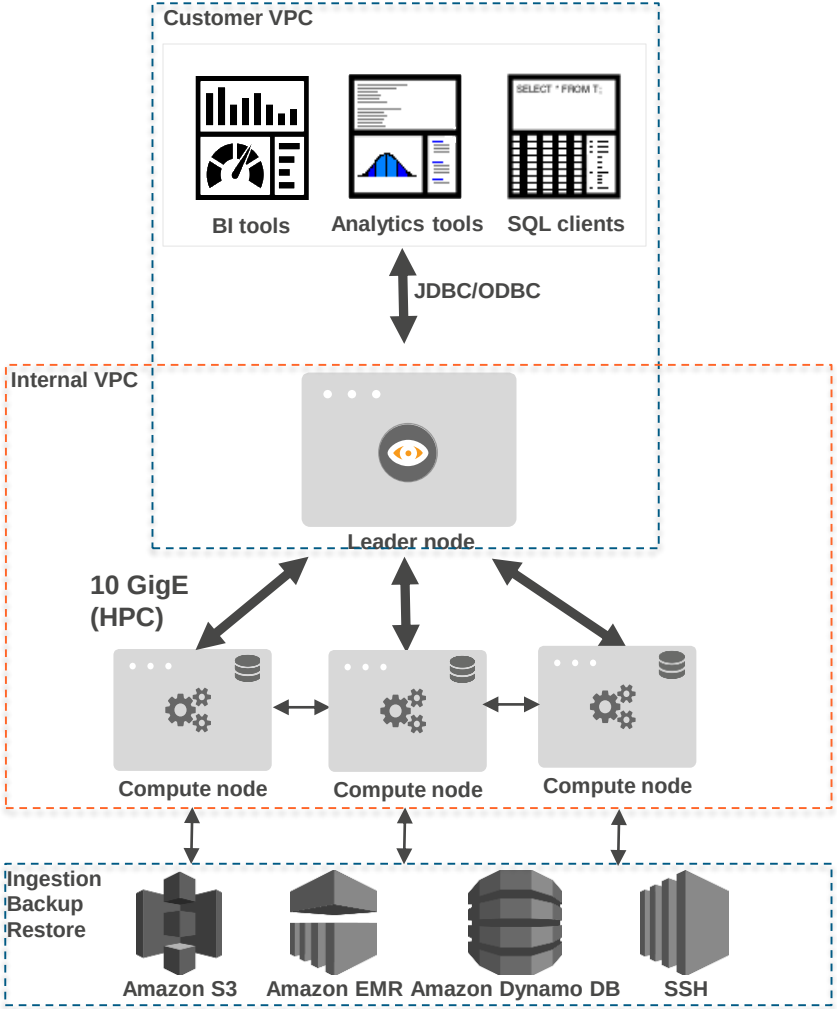
Audit logging



Certifications & Compliance

Amazon Redshift is secure

- Load encrypted from S3
- SSL to secure data in transit
  - ECDHE perfect forward secrecy
- Amazon VPC for network isolation
- Encryption to secure data at rest
  - All blocks on disks and in S3 encrypted
  - Block key, cluster key, master key (AES-256)
  - On-premises HSM & AWS CloudHSM support
- Audit logging and AWS CloudTrail integration
- SOC 1/2/3, PCI-DSS, FedRAMP, BAA



We have added even more security features...

IAM support for data LOAD/UNLOAD

- Use IAM roles to perform LOAD/UNLOAD operations
- Cluster is given IAM credentials
- Cluster can be granted access to specific S3 buckets
- Simplify credentials management
- More secure

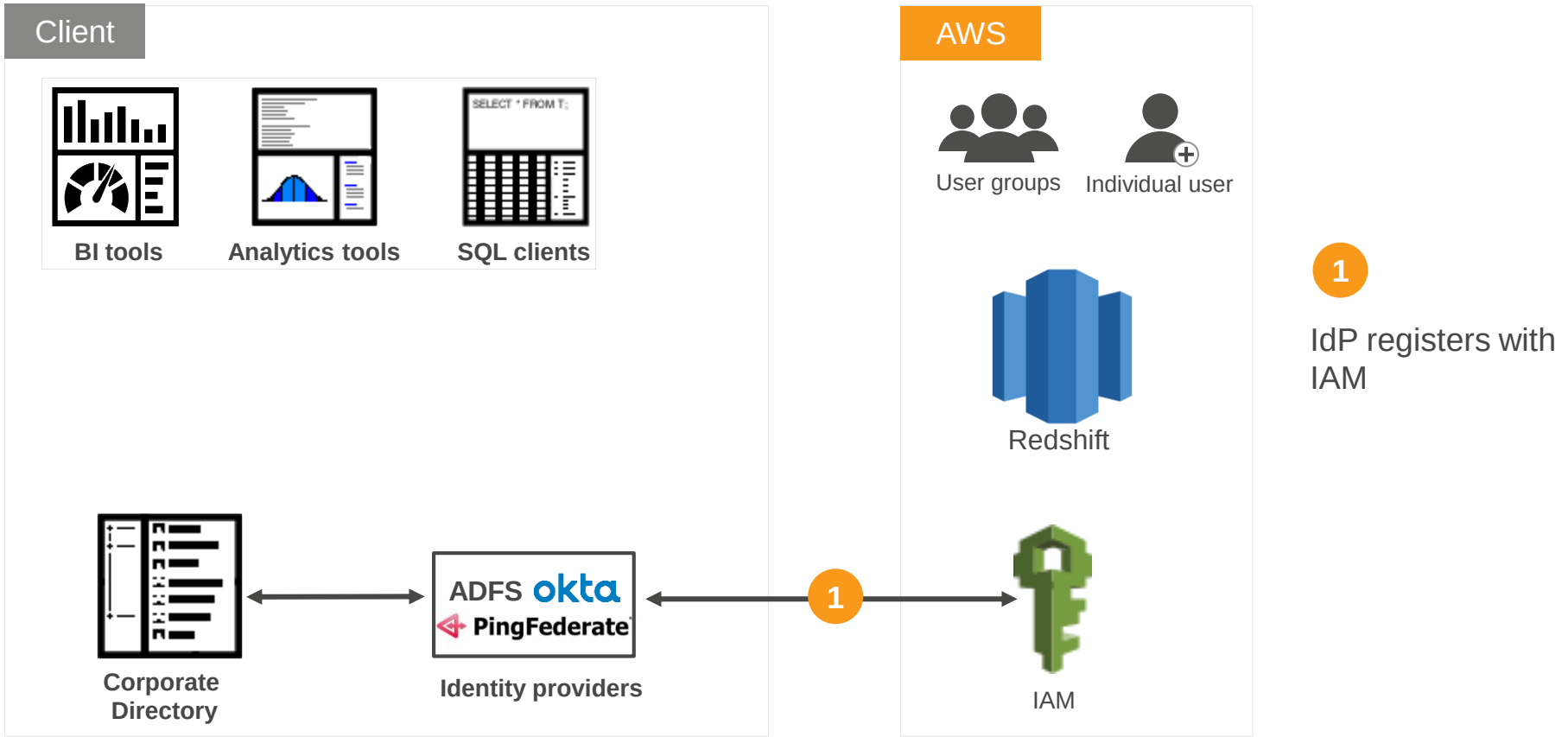
Simple

Enhanced VPC Routing

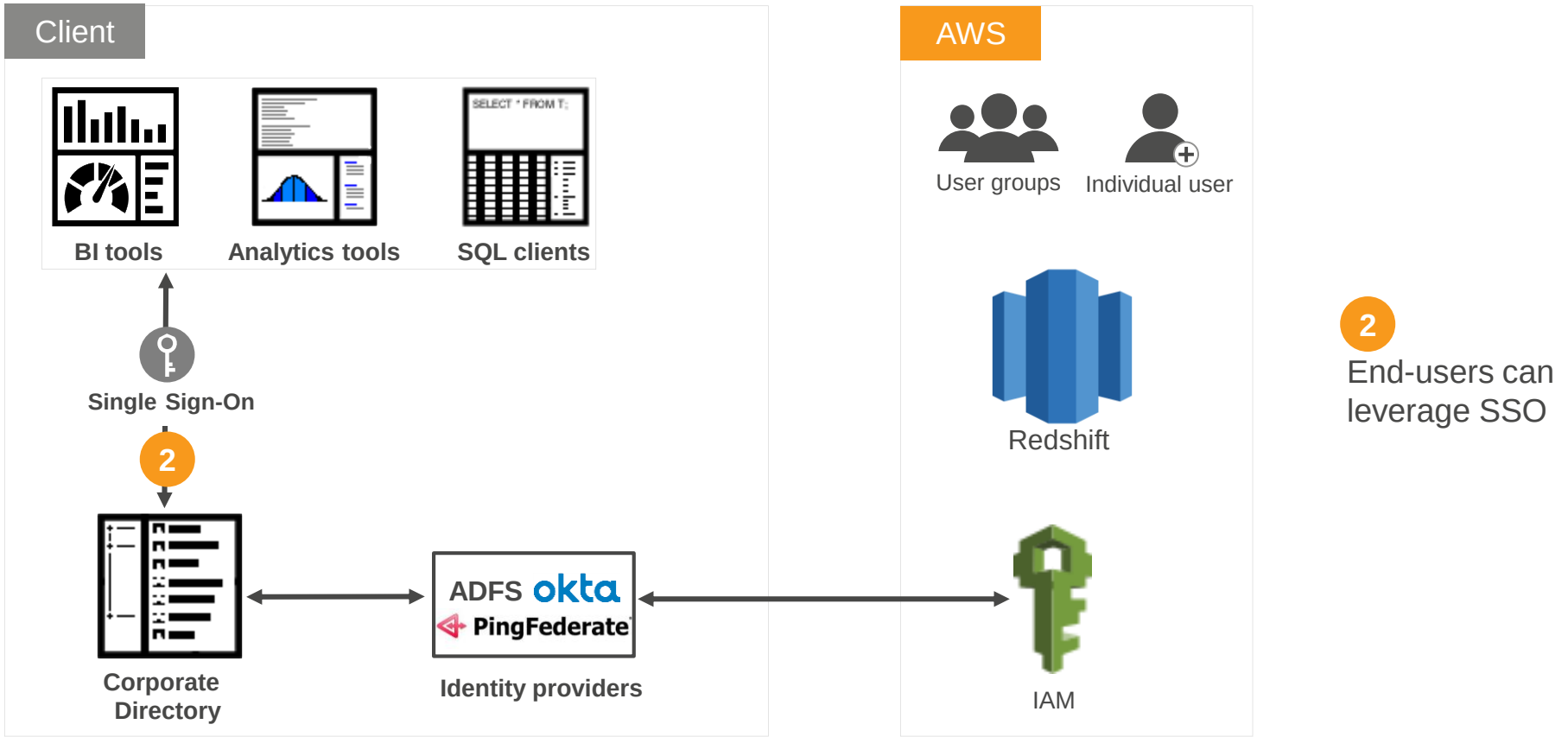
- All query traffic only flows through customer VPC
- Strict data traffic management between Amazon Redshift and other data sources

Secure

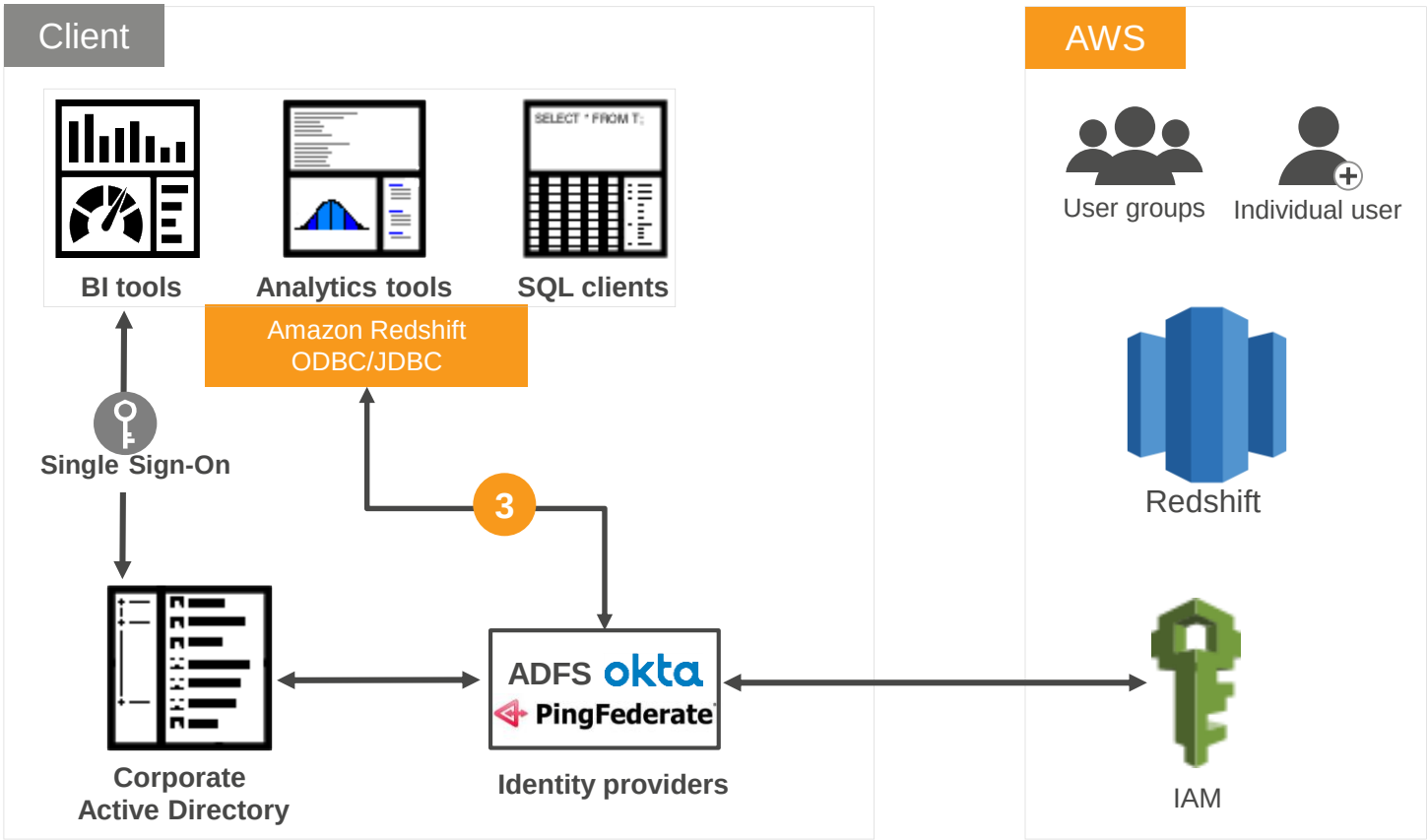
Coming soon: IAM Authentication for DB users



Coming soon: IAM Authentication for DB users



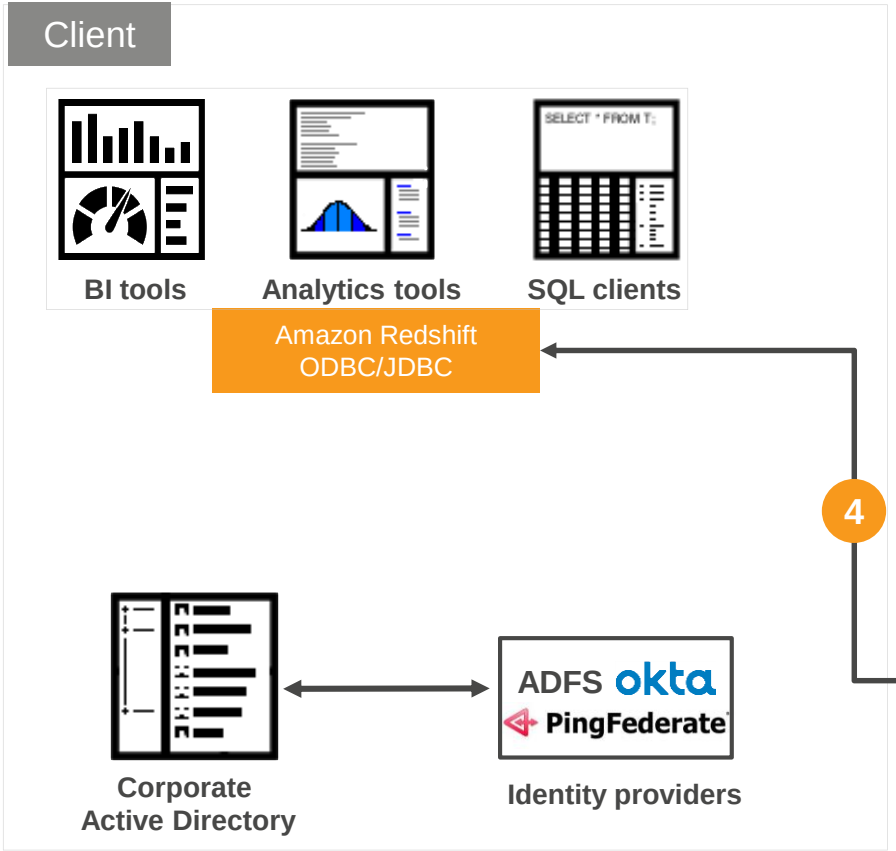
Coming soon: IAM Authentication for DB users



3

New Redshift ODBC/JDBC drivers. Grab the ticket (userid) and get a SAML assertion

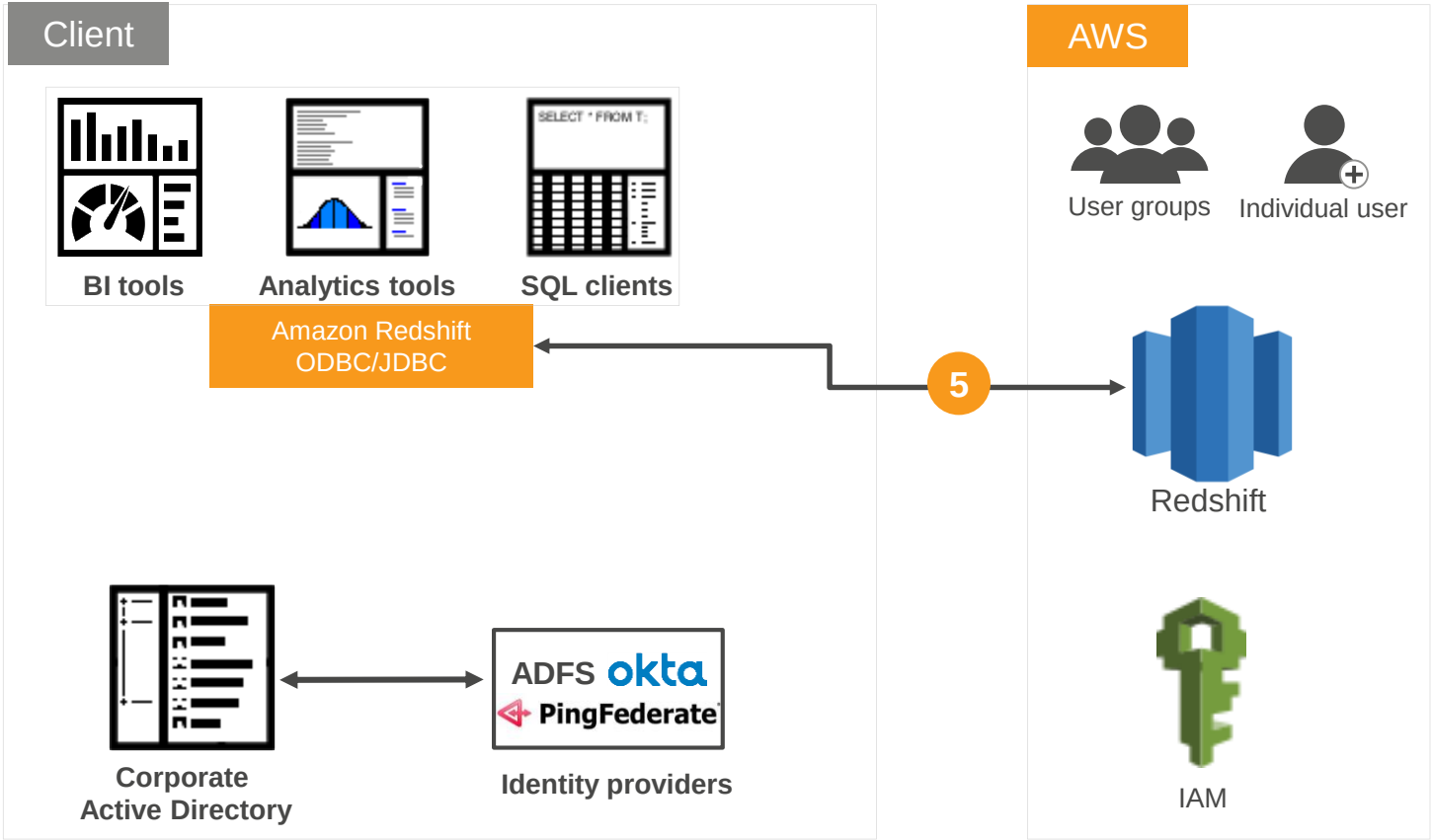
Coming soon: IAM Authentication for DB users



4

Send SAML assertion to IAM. Obtain IAM role credentials

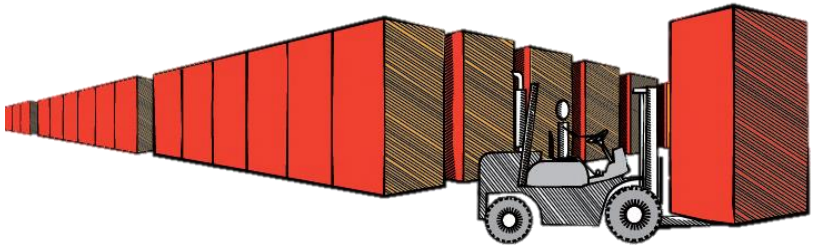
Coming soon: IAM Authentication for DB users



5

Use IAM role to obtain temporary cluster credentials. Use temporary credentials to log into the database

Amazon Redshift



Contact us at redshift-pm@amazon.com

Available now

- Queue hopping
- 10X VACUUM performance improvement
- Node fault tolerance
- Enhanced VPC routing
- IAM support for Copy/Unload
- Auto compression for CTAS
- TimestampTZ datatype

Coming soon

- Query Monitoring rules
- Automatic and incremental background VACUUM
- Short query bias
- Power start
- IAM Authentication for DB users
- Auto compression for new tables
- Enhanced JSON & AVRO ingestion performance

Thank you!

Remember to complete your evaluations!