Getting Started with Amazon Redshift

Maor Kleider, Sr. Product Manager, Amazon Redshift

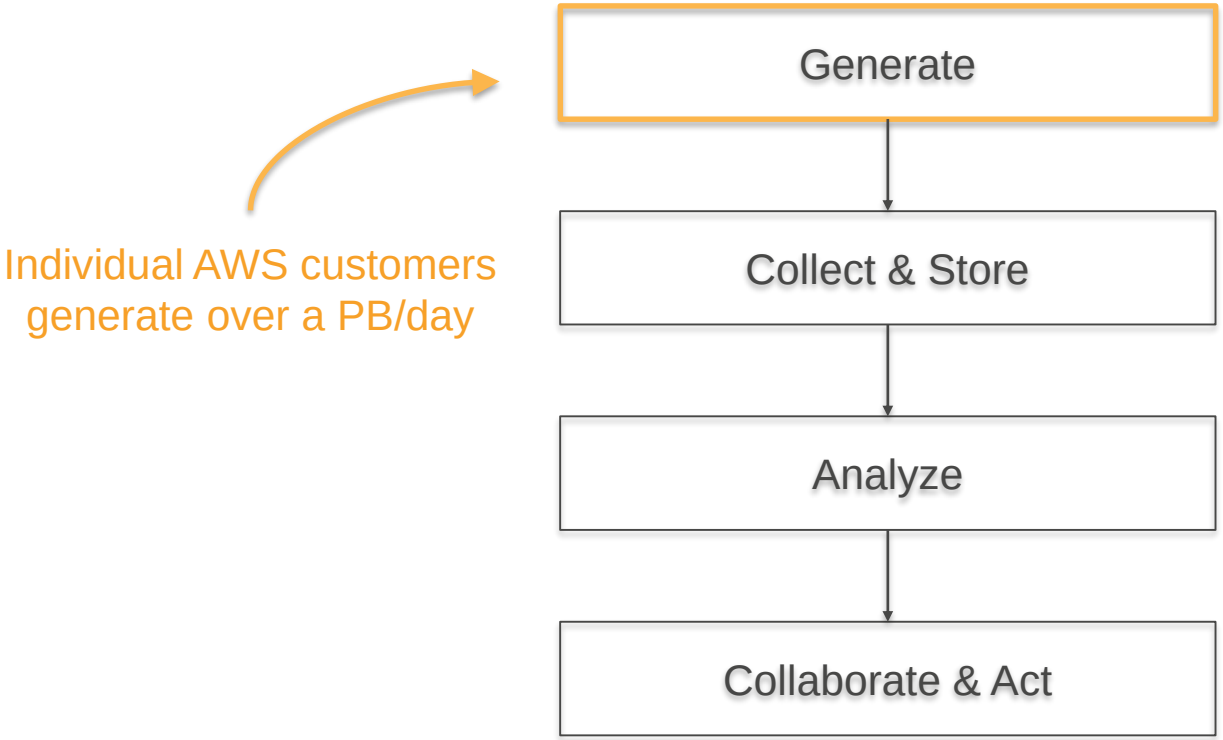
Agenda

- Introduction
- Benefits
- Use cases
- Getting started
- Q&A

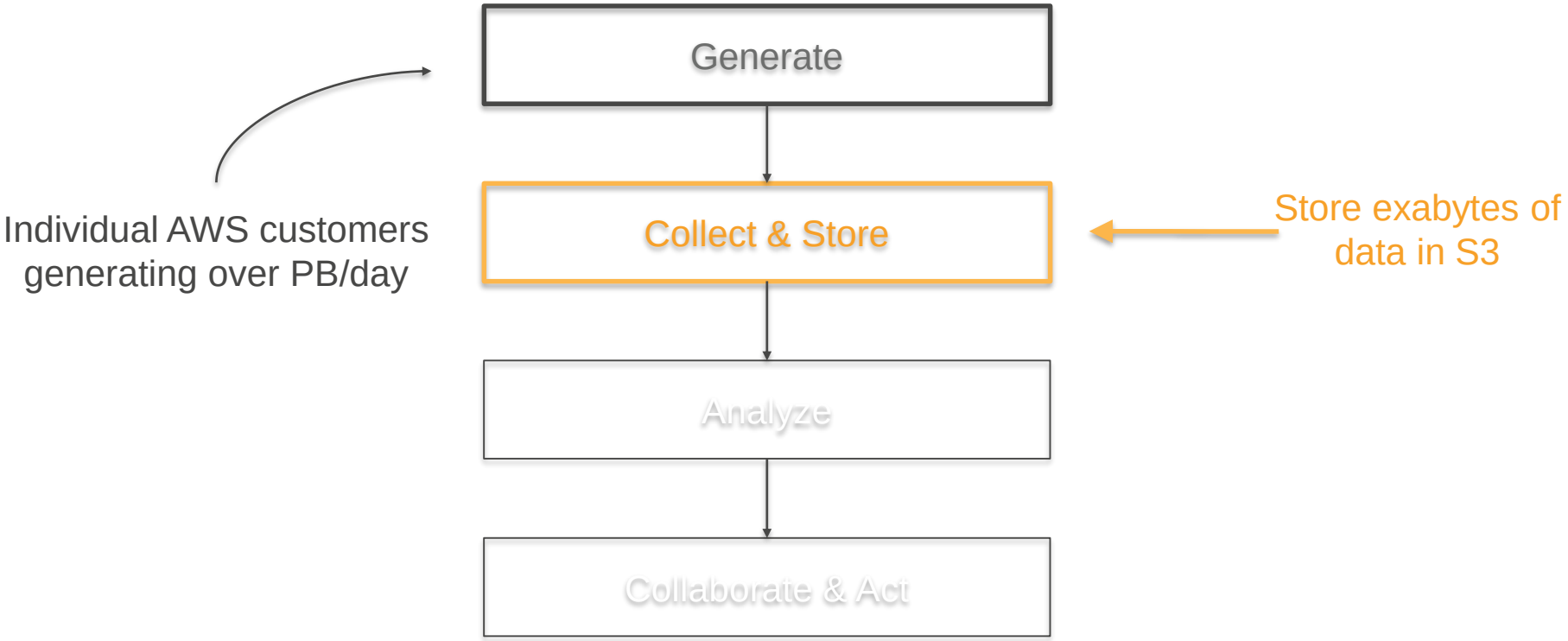
What is Big Data?

When your data sets become so large and diverse that you have to start ***innovating*** around how to collect, store, process, analyze and share them

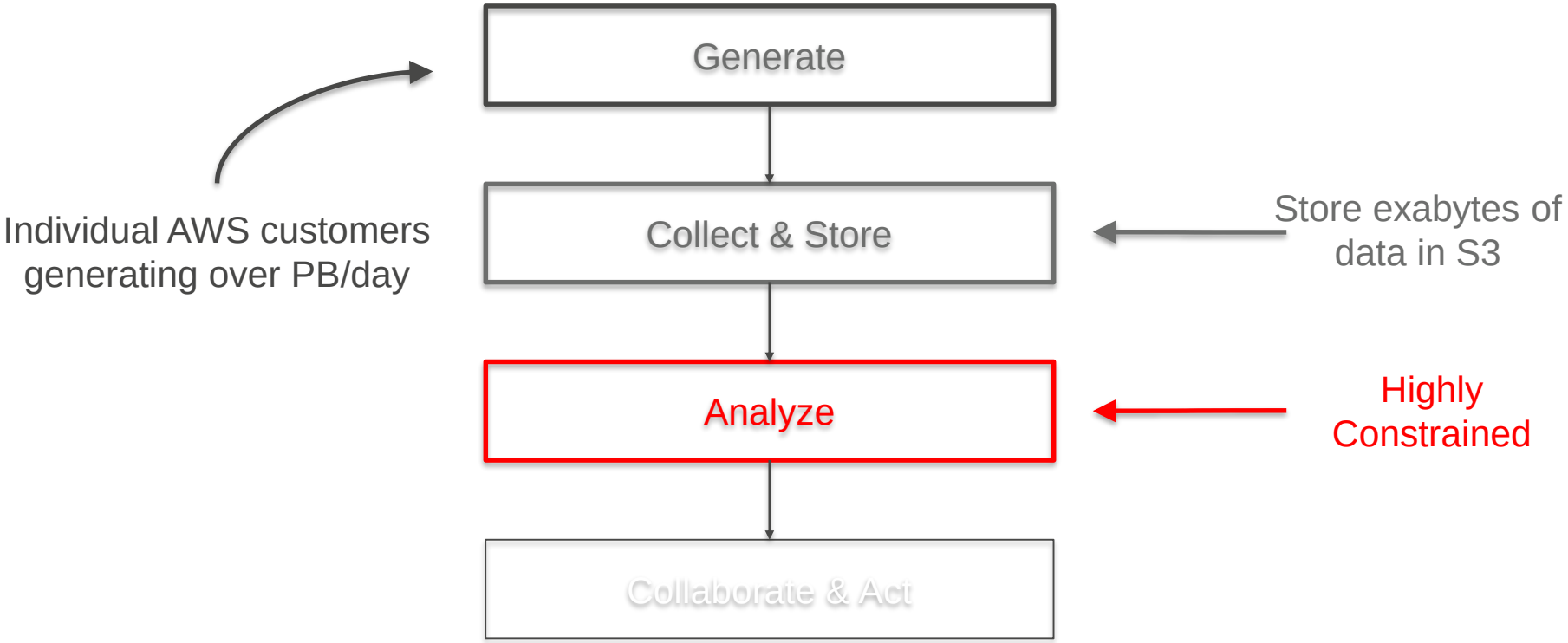
It’s never been easier to generate vast amounts of data



Amazon S3 lets you collect and store all this data

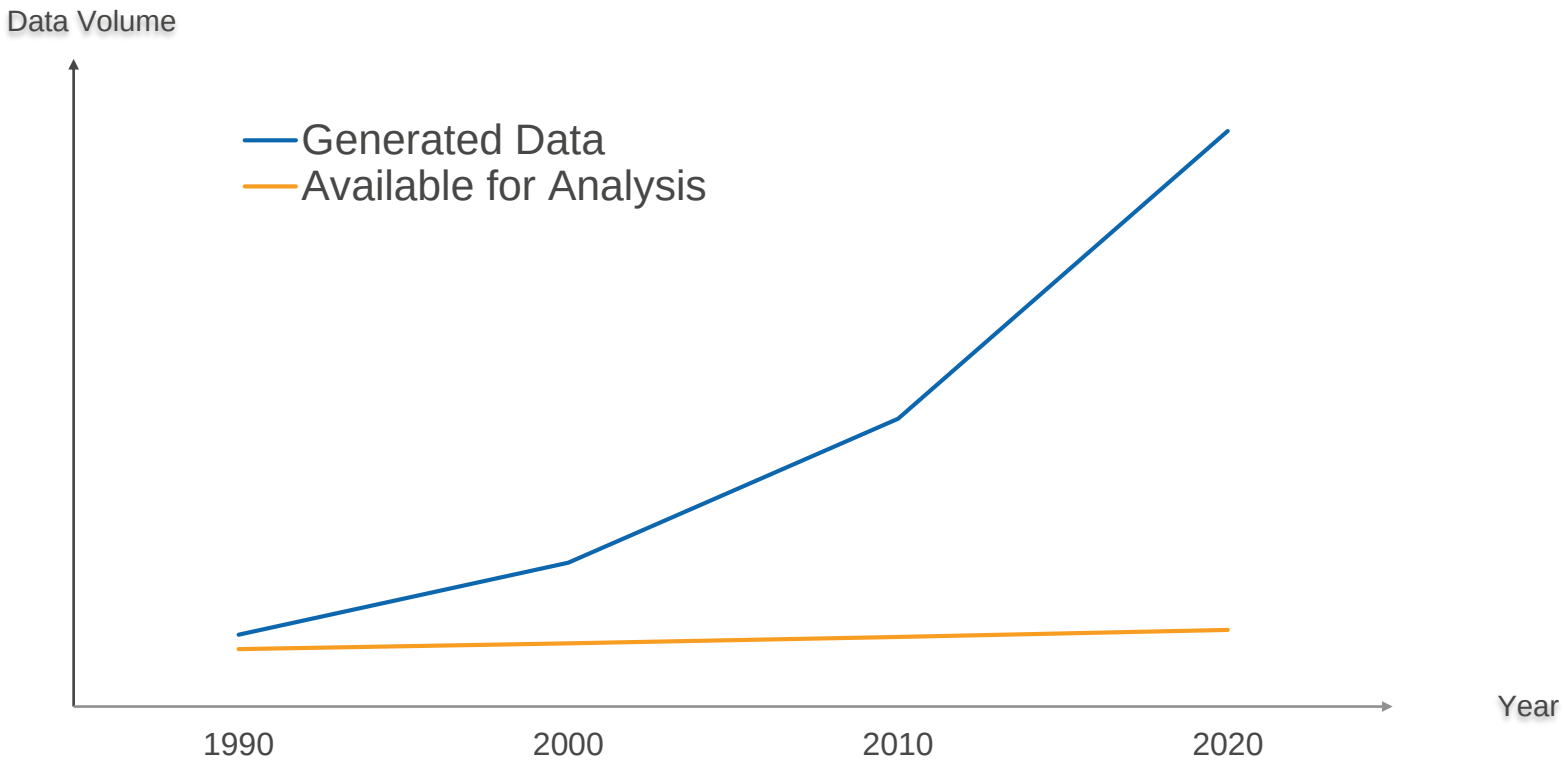


But how do you analyze it?



The Dark Data Problem

Most generated data is unavailable for analysis



Sources:
Gartner: User Survey Analysis: Key Trends Shaping the Future of Data Center Infrastructure Through 2011
IDC: Worldwide Business Analytics Software 2012–2016 Forecast and 2011 Vendor Shares

AWS Big Data Portfolio

Collect

- Amazon Kinesis Firehose
- AWS Direct Connect
- Amazon Kinesis Analytics
- Amazon Snowball
- Amazon Kinesis Streams

Store

- Amazon S3
- Amazon Glacier
- Amazon Dynamo DB
- Amazon RDS, Amazon Aurora
- Amazon CloudSearch
- Amazon Elasticsearch

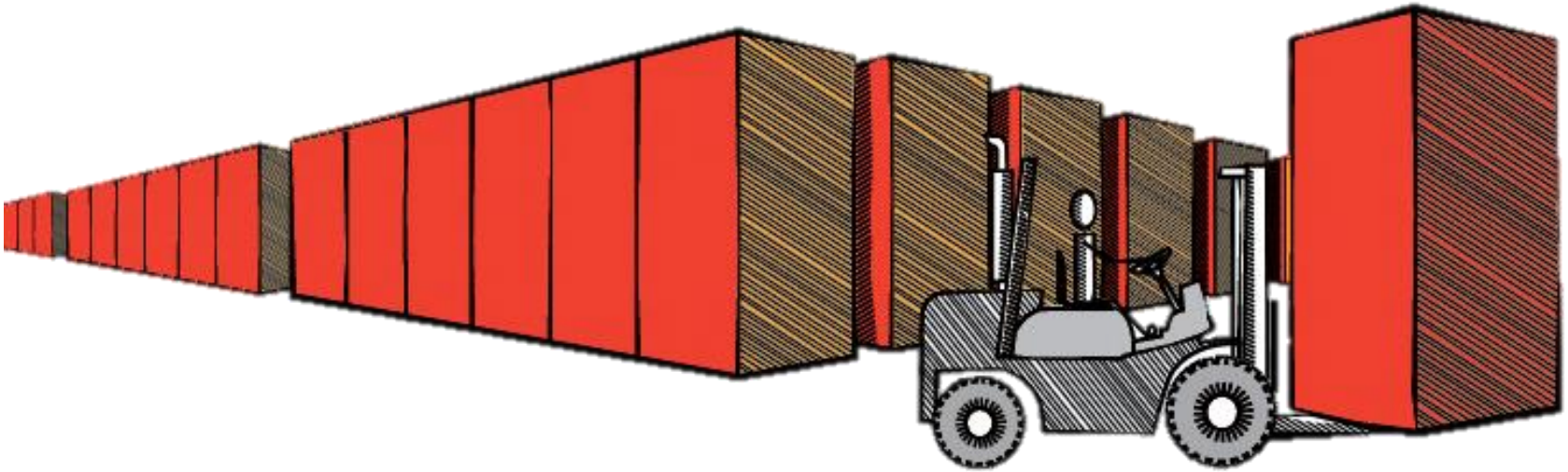
Analyze

- Amazon EMR
- Amazon EC2
- Amazon Redshift
- Amazon Athena
- Amazon QuickSight
- Amazon Machine Learning

AWS Database Migration Service

AWS Glue

Amazon Redshift



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

150+ features

Amazon Redshift

a lot faster
a lot simpler
a lot cheaper

Relational data warehouse

Massively parallel; petabyte scale

Fully managed

HDD and SSD platforms

$1,000/TB/year; starts at $0.25/hour

Selected Amazon Redshift customers

Adobe
Time Inc.
yelp
FT.com FINANCIAL TIMES
BookmyShow
Schumachergroup
coursera
boingo
NOKIA
edmunds.com
finra
SCHOLASTIC
Pinterest
foursquare
Sansan
NASDAQ OMX
SOUNDCLOUD
Albert Optimization technology
salesforce desk
docomo
latentview Actionable Insights • Accurate Decisions
scopely
sling
DataXu
peak GAMES
lyft
BEACHMINT
VIVAKI
has offers
amazon
MINICLIP
UMUC University of Maryland University College
Nintendo
etix
SNOWPLOW
imshealth INTELLIGENCE APPLIED.
euclid
spuul

Use Case: Traditional Data Warehousing

Business Reporting

Advanced pipelines and queries

Secure and Compliant

Bulk Loads and Updates

Easy 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

Use Case: Log Analysis

Log & Machine IOT Data

Clickstream Events Data

Time-Series Data

Cheap – 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

Use Case: 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

Ease of 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

Amazon Redshift architecture

Leader node

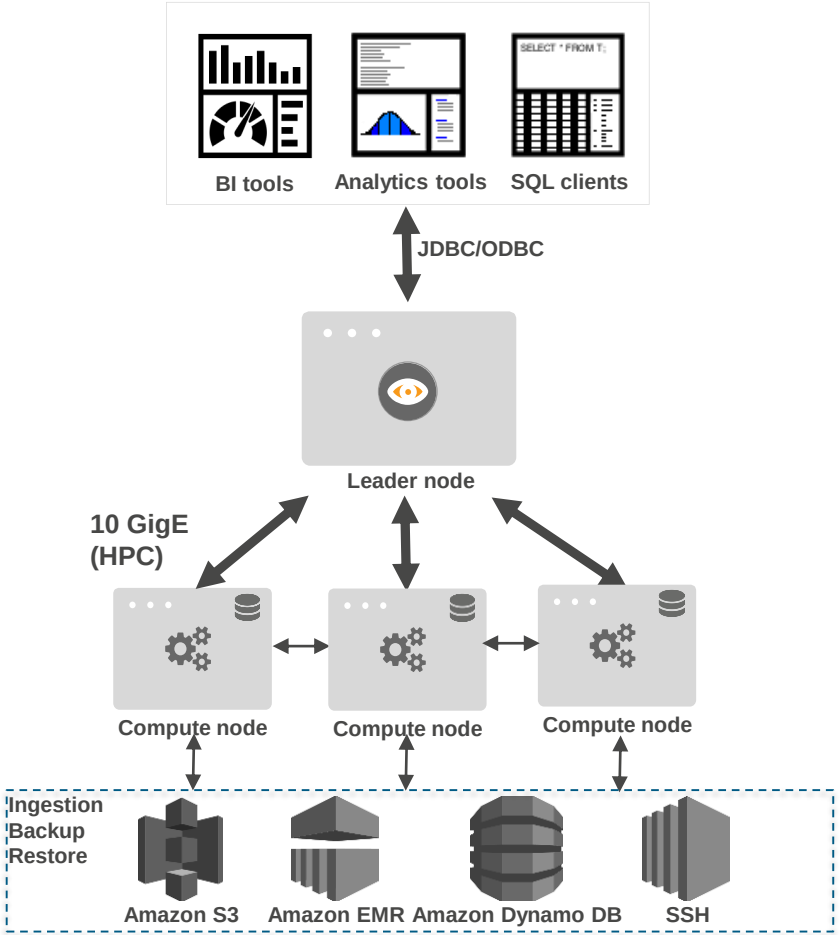
- Simple SQL endpoint
- Stores metadata
- Optimizes query plan
- Coordinates query execution

Compute nodes

- Local columnar storage
- Parallel/distributed execution of all queries, loads, backups, restores, resizes

Start at just $0.25/hour, grow to 2 PB (compressed)

- DC1: SSD; scale from 160 GB to 326 TB
- DS2: HDD; scale from 2 TB to 2 PB



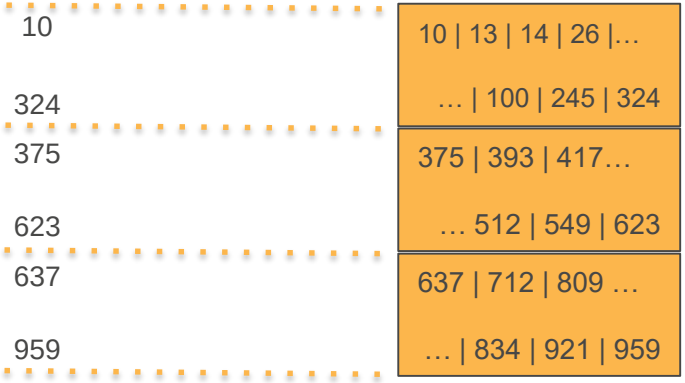
Benefit #1: Amazon Redshift is fast

Dramatically less I/O

- Column storage
- Data compression
- Zone maps
- Direct-attached storage
- Large data block sizes

```
analyze compression listing;

  Table   |      Column      | Encoding
---------+------------------+----------
 listing | listid           | delta
 listing | sellerid         | delta32k
 listing | eventid          | delta32k
 listing | dateid           | bytedict
 listing | numtickets       | bytedict
 listing | priceperticket   | delta32k
 listing | totalprice       | mostly32
 listing | listtime         | raw
```



Benefit #1: Amazon Redshift is fast

Parallel and distributed

Query

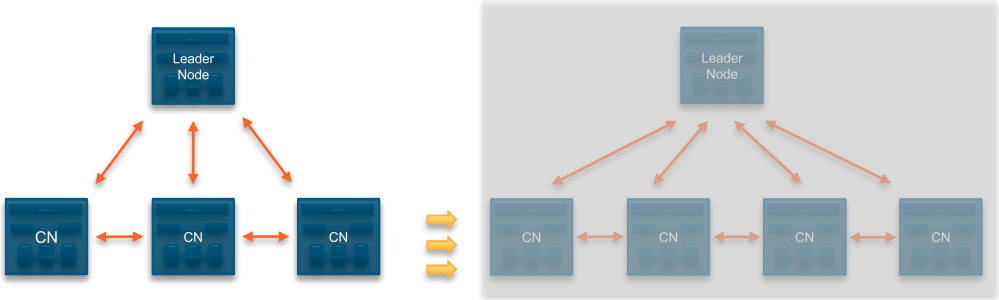
Load

Export

Backup

Restore

Resize



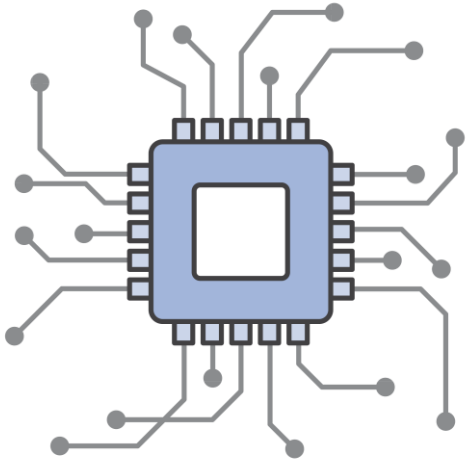
Benefit #1: Amazon Redshift is fast

Hardware optimized for I/O intensive workloads, 4 GB/sec/node

Enhanced networking, over 1 million packets/sec/node

Choice of storage type, instance size

Regular cadence of auto-patched improvements



Benefit #1: 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**"*

4 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**"*

And has gotten faster...

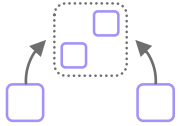
5X Query throughput improvement over the past year

- Memory allocation (launched)
- Improved commit and I/O logic (launched)
- Queue hopping (launched)
- Query monitoring rules (launched)

Fast

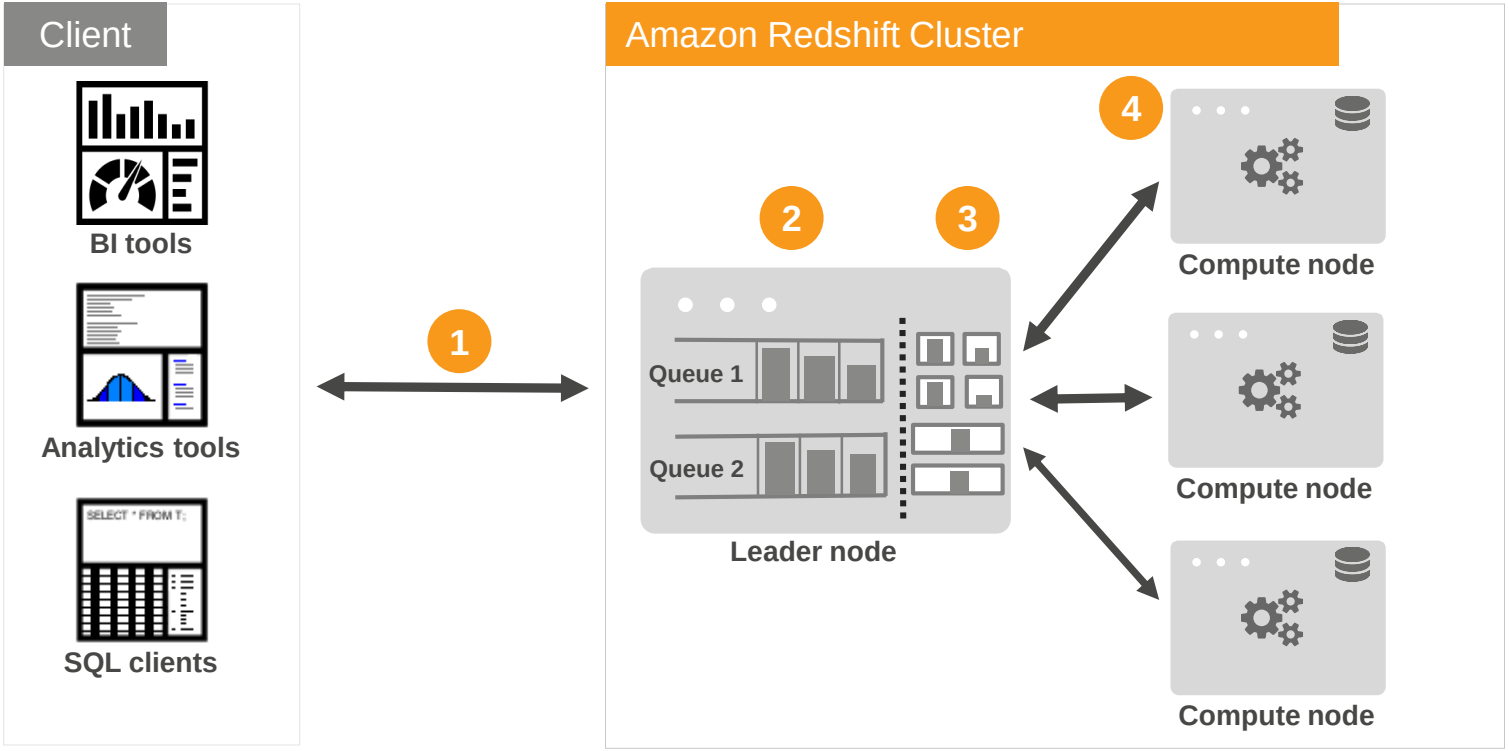
10X Vacuuming performance improvement

- 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



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

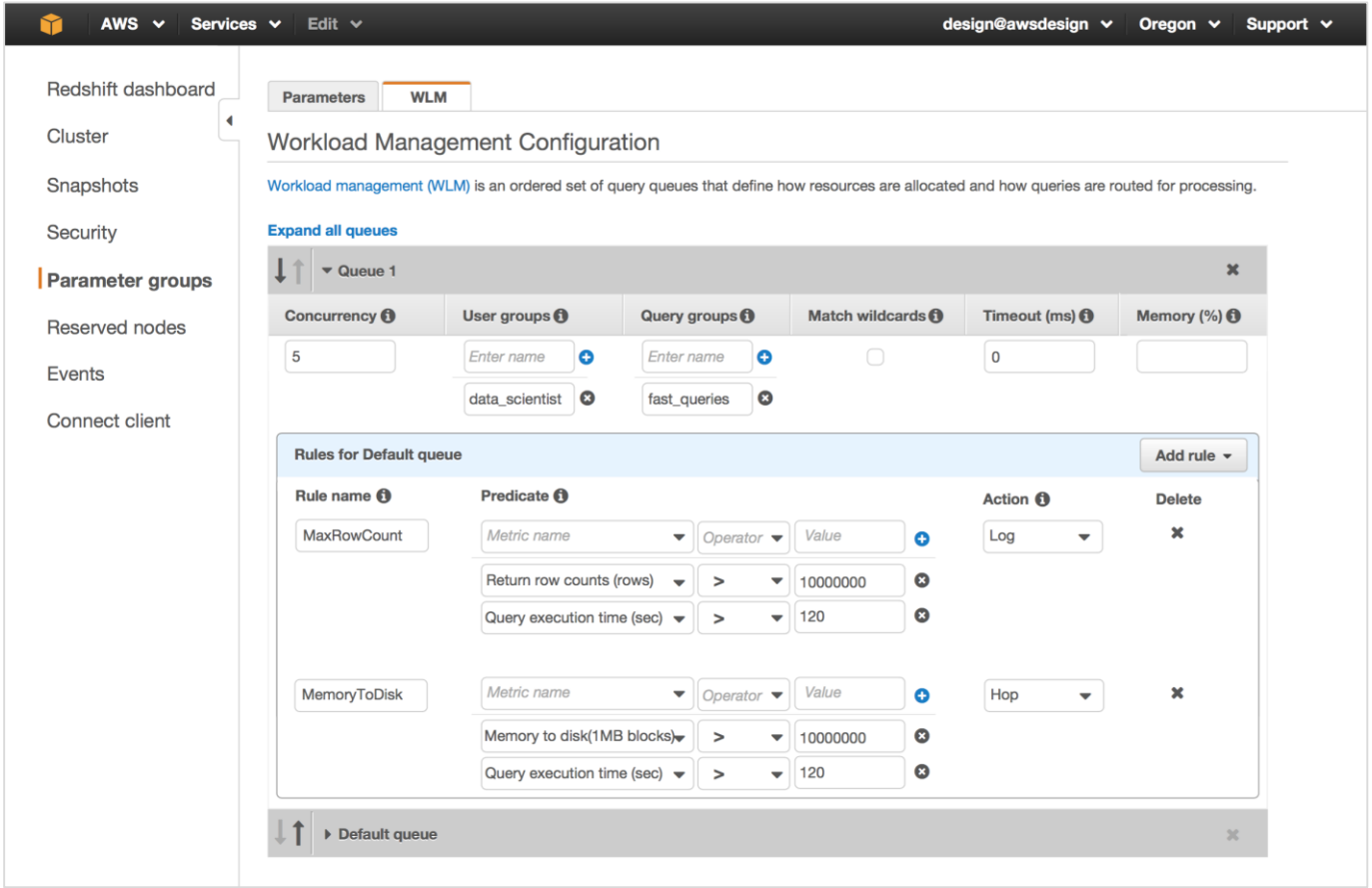
If { *rule* } then [*action*]

{ 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

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

Query monitoring rules

Common use cases:

- Protect interactive queues

```
INTERACTIVE = { “query_execution_time > 15 sec” or
                “query_cpu_time > 1500 uSec” or
                ”query_blocks_read > 18000 blocks” } [HOP]
```

- Monitor ad-hoc queues for heavy queries

```
AD-HOC = { “query_execution_time > 120” or
           “query_cpu_time > 3000” or
           ”query_blocks_read > 180000” or
           “memory_to_disk > 400000000000”} [LOG]
```

- Limit the number of rows returned to a client

```
MAXLINES = { “RETURN_ROWS > 50000” } [ABORT]
```

Benefit #2: Amazon Redshift is inexpensive

DS2 (HDD)	Price per hour for DS2.XL single node	Effective annual price per TB compressed
On-demand	$ 0.850	$ 3,725
1 year reservation	$ 0.500	$ 2,190
3 year reservation	$ 0.228	$ 999

DC1 (SSD)	Price per hour for DC1.L single node	Effective annual price per TB compressed
On-demand	$ 0.250	$ 13,690
1 year reservation	$ 0.161	$ 8,795
3 year reservation	$ 0.100	$ 5,500

<u>Pricing is simple</u>

Number of nodes x price/hour

No charge for leader node

No upfront costs

Pay as you go

Benefit #3: Amazon Redshift is fully managed

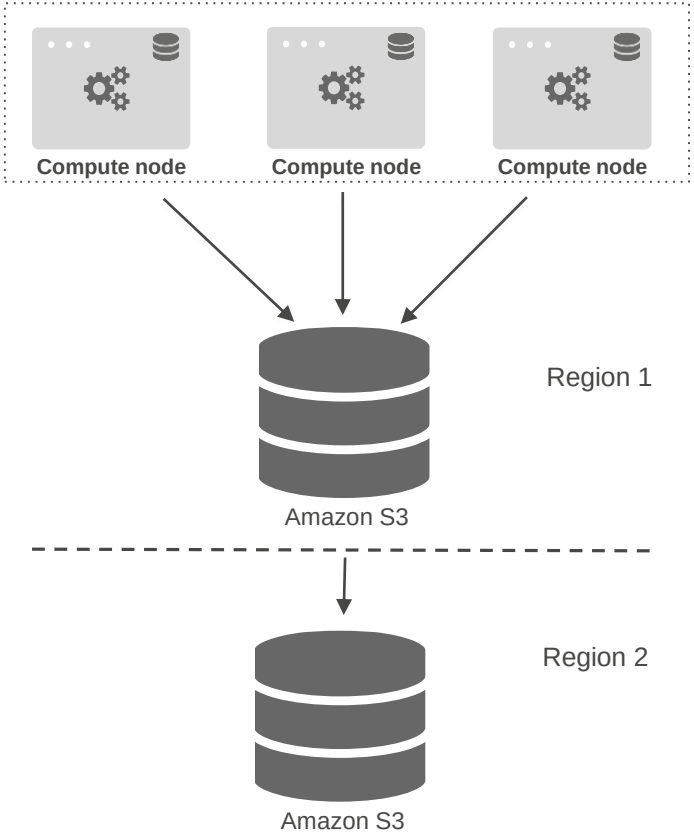
Continuous/incremental backups

Multiple copies within cluster

Continuous and incremental backups to Amazon S3

Continuous and incremental backups across regions

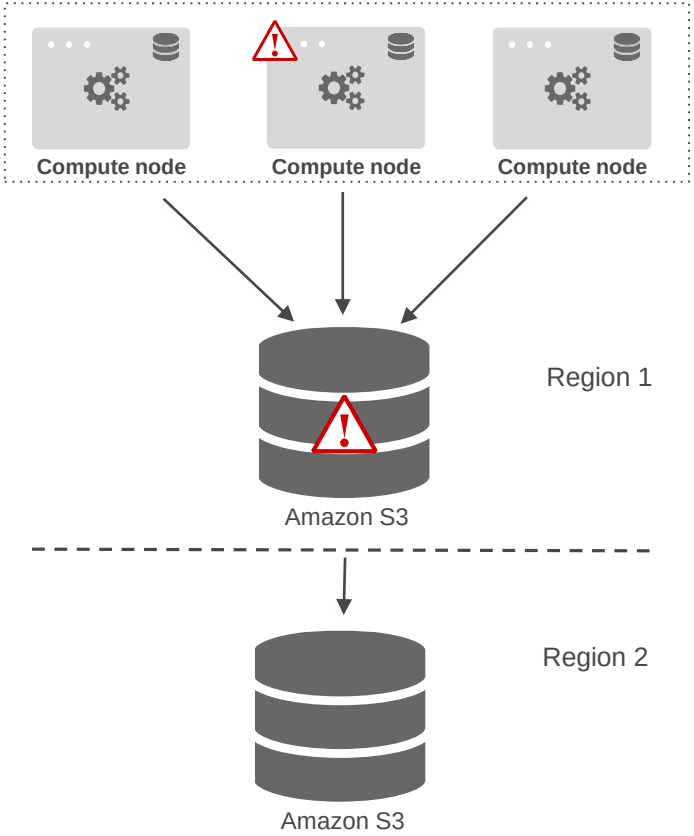
Streaming restore



Benefit #3: Amazon Redshift is fully managed

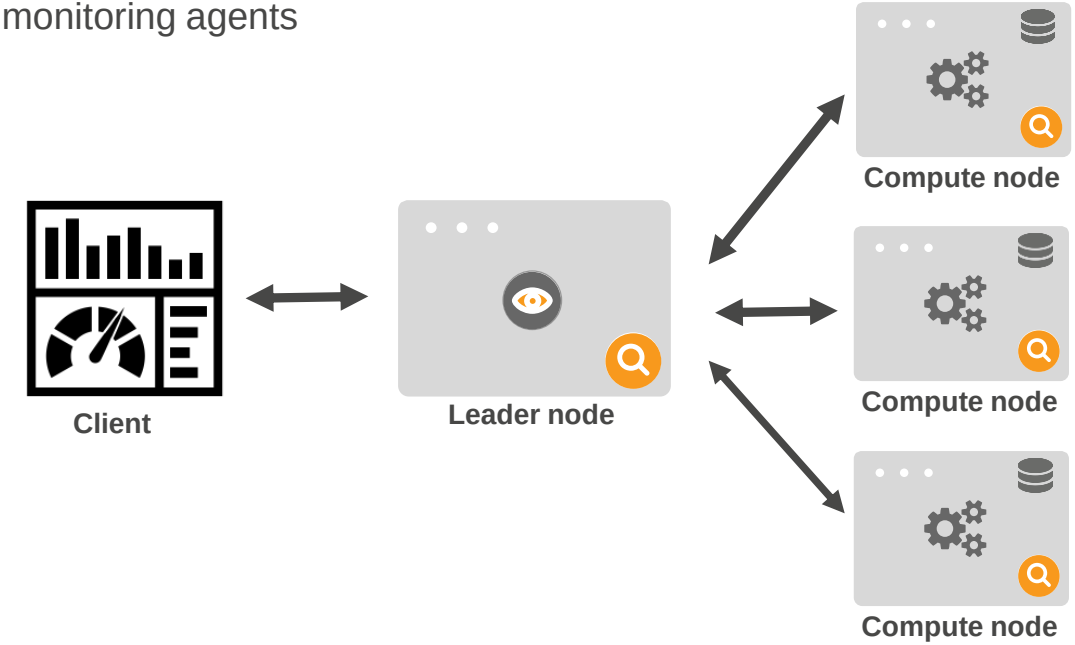
Fault tolerance

- Disk failures
- Node failures
- Network failures
- Availability Zone/region level disasters



Node fault tolerance

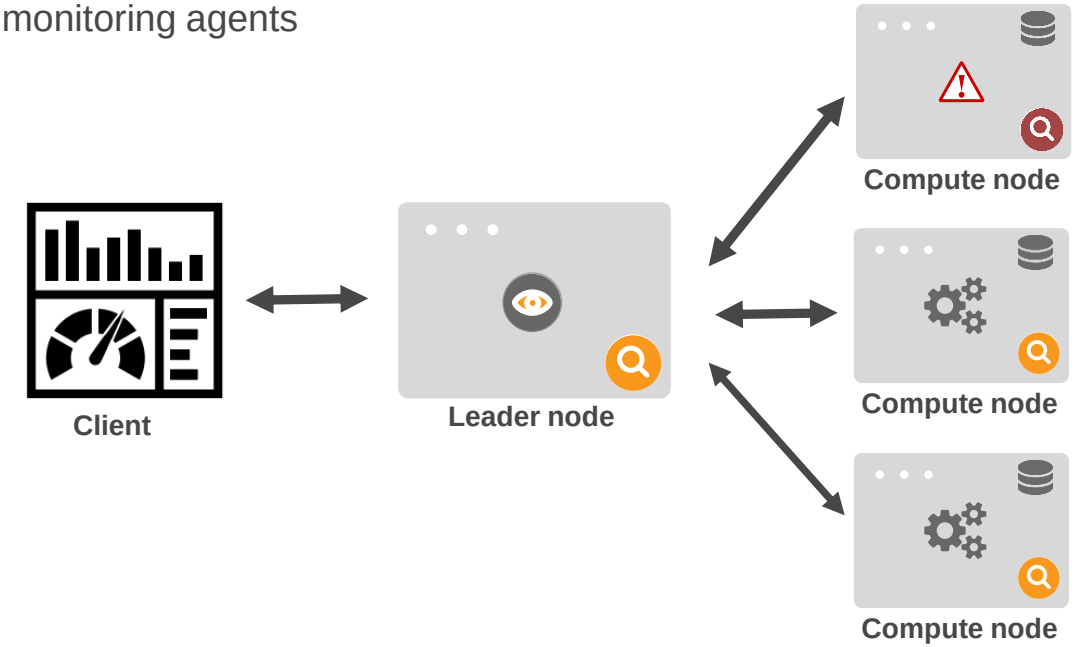
Data-path monitoring agents



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

Node fault tolerance

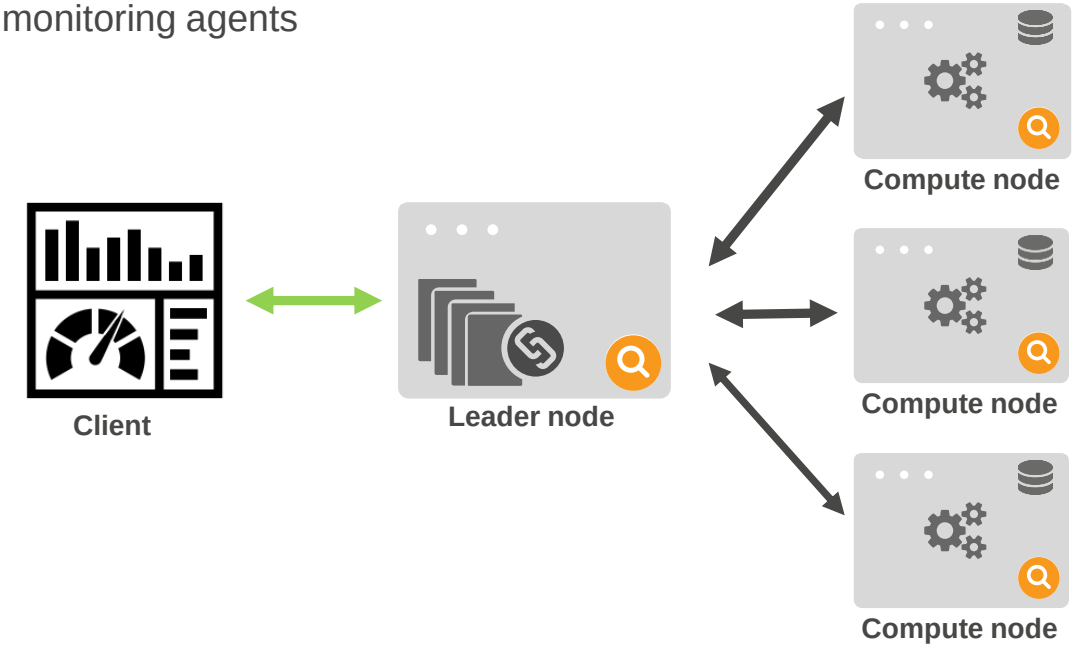
Data-path monitoring agents



Failure is detected at one of the compute nodes

Node fault tolerance

Data-path monitoring agents



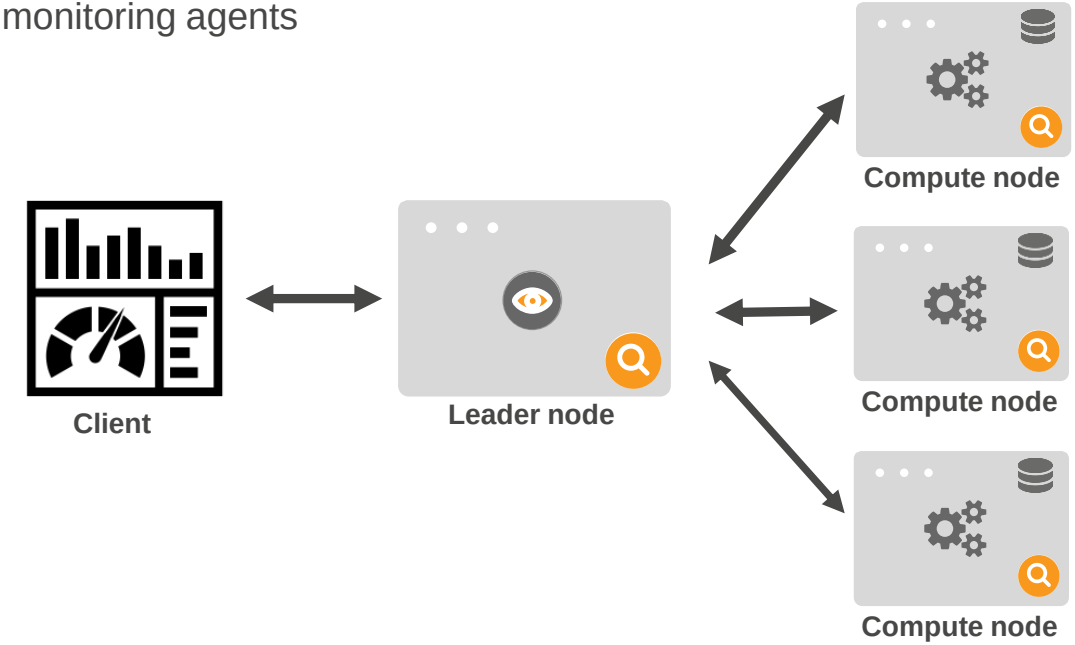
Redshift parks the connections

Next, the node is replaced

Node fault tolerance

Data-path monitoring agents

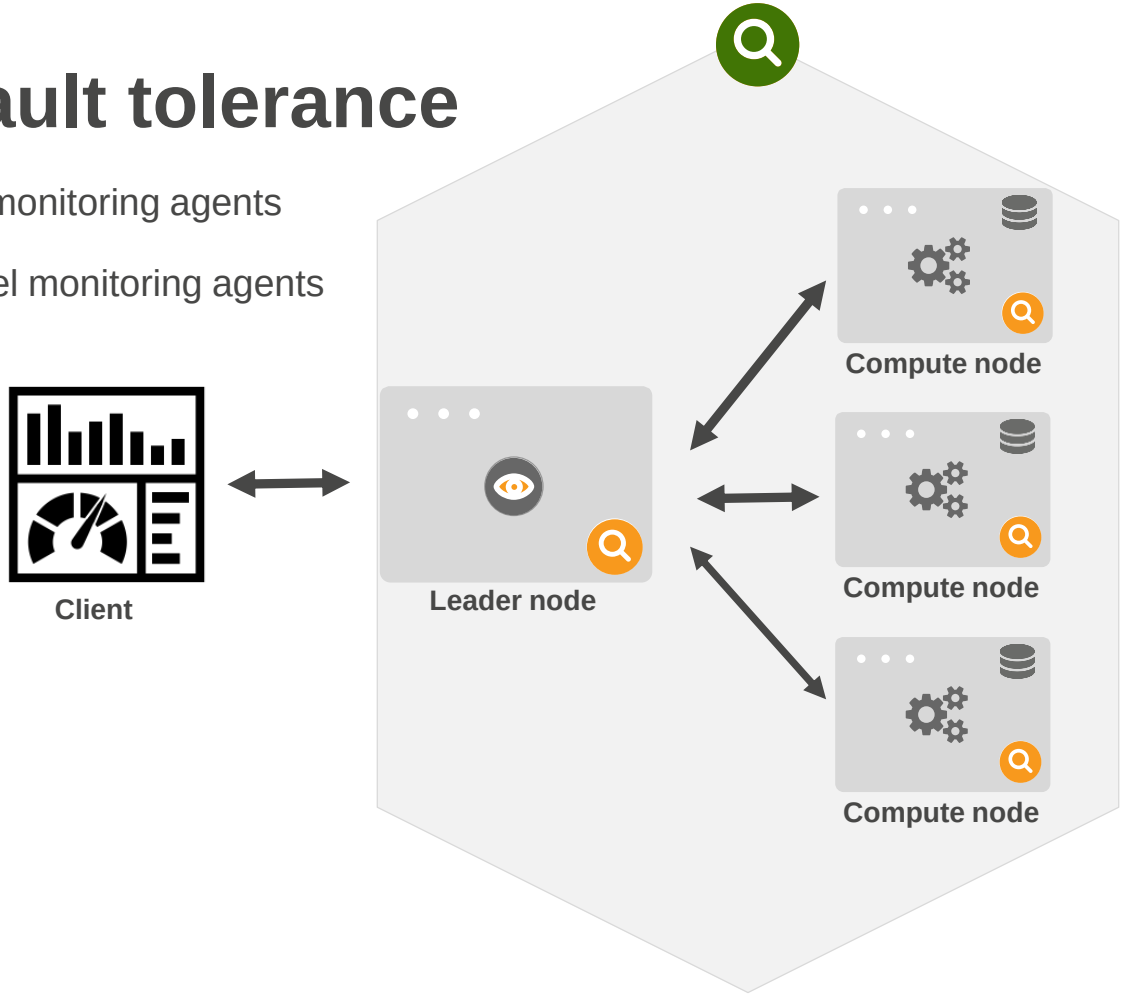
Queries are re-submitted



Node fault tolerance

Data-path monitoring agents

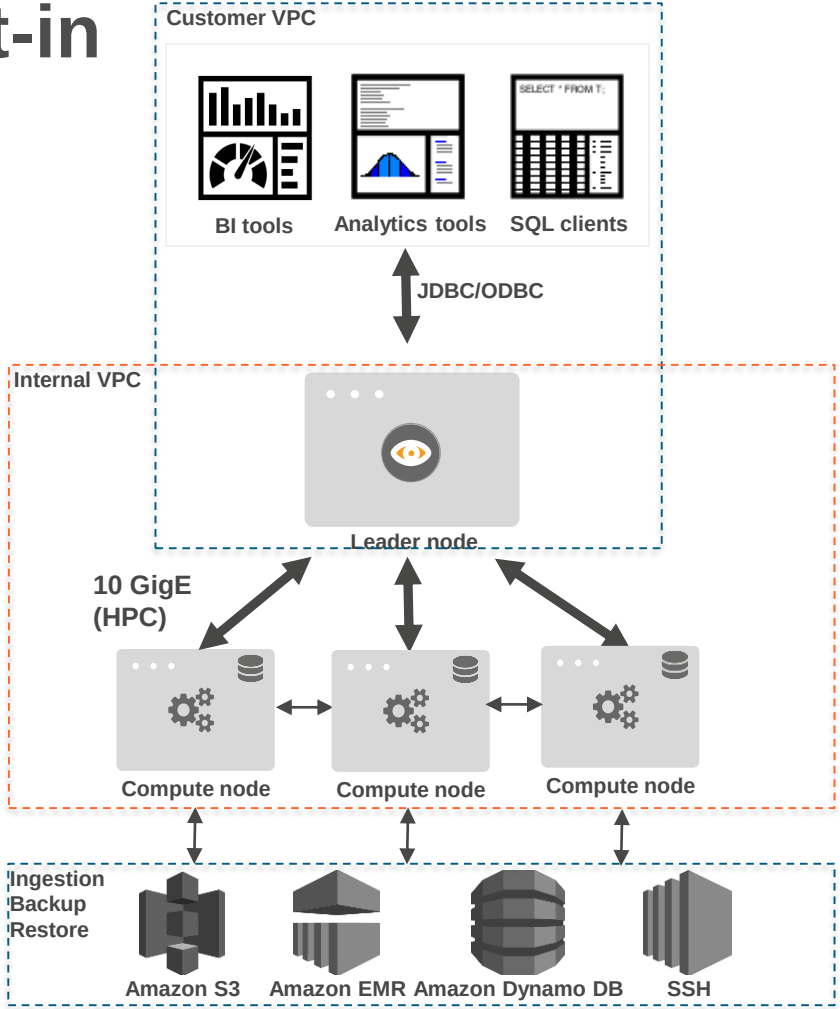
Cluster-level monitoring agents



Additional monitoring layer for the leader node and network

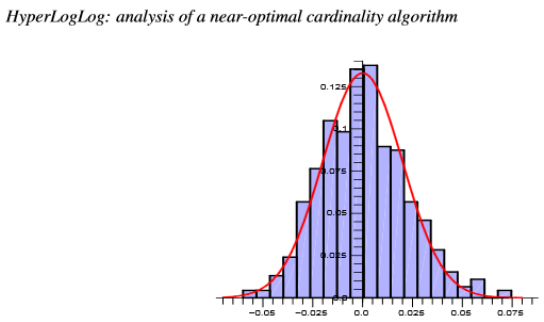
Benefit #4: Security is built-in

- 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



Benefit #5: Amazon Redshift is powerful

- Approximate functions
- User defined functions
- Machine learning
- Data science



Benefit #6: Amazon Redshift has a large ecosystem

Data integration

ATTUNITY CloudBeam
bryte Systems
FlyData
INFORMATICA
Matillion
pentaho
Segment
snapLogic
talend integration at any scale
Qubole
xplenty

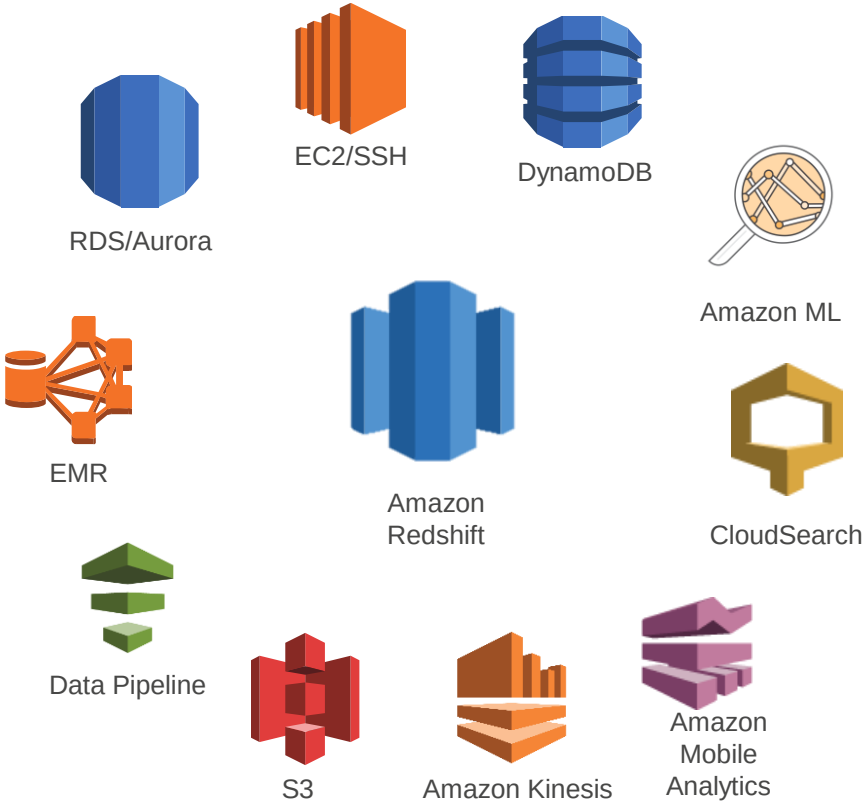
Business intelligence

actian
infor
periscope
alteryx
Qlik
JReport JINFONET SOFTWARE
ACTUATE The BIRT Company
SiSense
Logi ANALYTICS
birst
REDROCK BI
looker
ClearStory DATA Now You See It
tableau SOFTWARE
MicroStrategy
Chart.io Turn data to insight
TIBCO
MODE
Dundas
PANORAMA
Yellowfin
Business Partner IBM
pentaho A Hitachi Data Systems Company

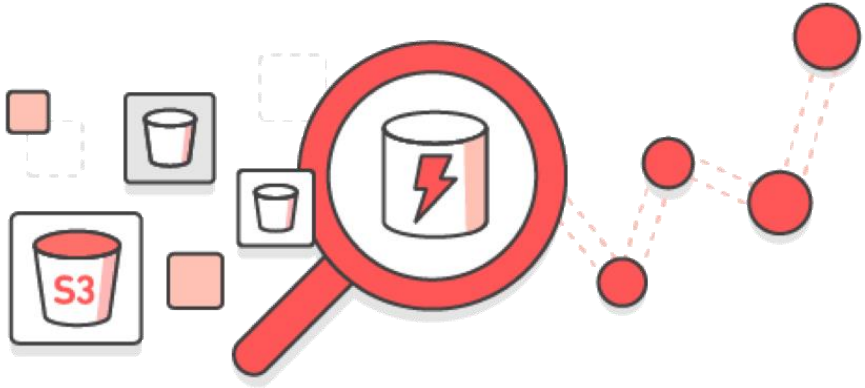
Systems integrators

2ND WATCH
accenture
BEEVA
BRIGHTLIGHT Ask more of your data
Capgemini CONSULTING.TECHNOLOGY.OUTSOURCING
cloudpack
Cognizant
FULL 360
iOLAP
isobar
MAGNUS DATA
NorthBay solutions
SLALOM CONSULTING

Benefit #7: Service oriented architecture



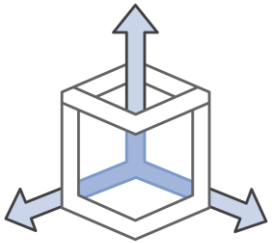
Amazon Redshift Spectrum



Amazon Redshift Spectrum

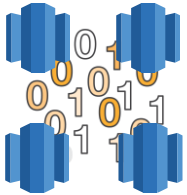
Run SQL queries directly against data in S3 using thousands of nodes

Fast @ exabyte scale



Elastic & highly available

On-demand, pay-per-query



High concurrency: Multiple clusters access same data

No ETL: Query data in-place using open file formats

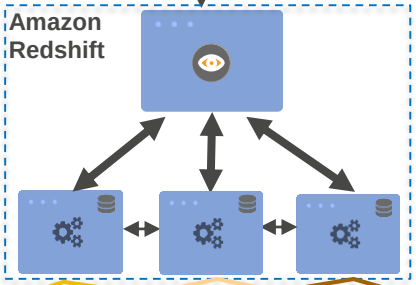
Full Amazon Redshift SQL support

Life of a query

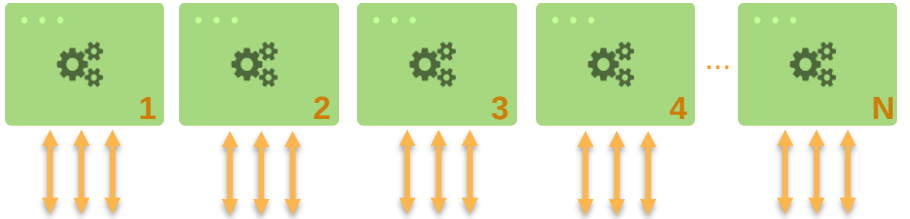
Query
SELECT COUNT(*)
FROM S3.EXT_TABLE
GROUP BY…

JDBC/ODBC

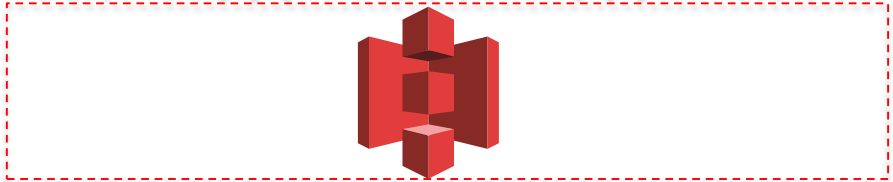
Amazon Redshift



Redshift Spectrum
Fast @ Exabyte scale



Amazon S3
Exabyte-scale object storage



Data Catalog
Apache Hive Metastore

Amazon Redshift Spectrum – Current support

File formats

- Parquet
- CSV
- Sequence
- RCFile
- ORC (coming soon)
- RegExSerDe (coming soon)

Compression

- Gzip
- Snappy
- Lzo (coming soon)
- Bz2

Encryption

- SSE with AES256
- SSE KMS with default key

Column types

- Numeric: bigint, int, smallint, float, double and decimal
- Char/varchar/string
- Timestamp
- Boolean
- DATE type can be used only as a partitioning key

Table type

- Non-partitioned table (s3://mybucket/orders/..)
- Partitioned table (s3://mybucket/orders/date=YYYY-MM-DD/..)

The Emerging Analytics Architecture

Storage

Amazon S3
Exabyte-scale Object Storage

AWS Glue Data Catalog
Hive-compatible Metastore

Serverless Compute

Amazon Kinesis Firehose
Real-Time Data Streaming

AWS Glue
ETL & Data Catalog

Amazon Redshift Spectrum
Fast @ Exabyte scale

AWS Lambda
Trigger-based Code Execution

Data Processing

Amazon EMR
Managed Hadoop Applications

Amazon Redshift
Petabyte-scale Data Warehousing

Amazon Athena
Interactive Query

Over 20 customers helped preview Amazon Redshift Spectrum

yelp

REDFIN

NTT docomo

WB GAMES

RECRUIT
Recruit Technologies Co.,Ltd.

edmunds

Time Inc.

Use cases

NTT Docomo: Japan's largest mobile service provider

68 million customers

Tens of TBs per day of data across a mobile network

6 PB of total data (uncompressed)

Data science for marketing operations, logistics, and so on

Greenplum on-premises

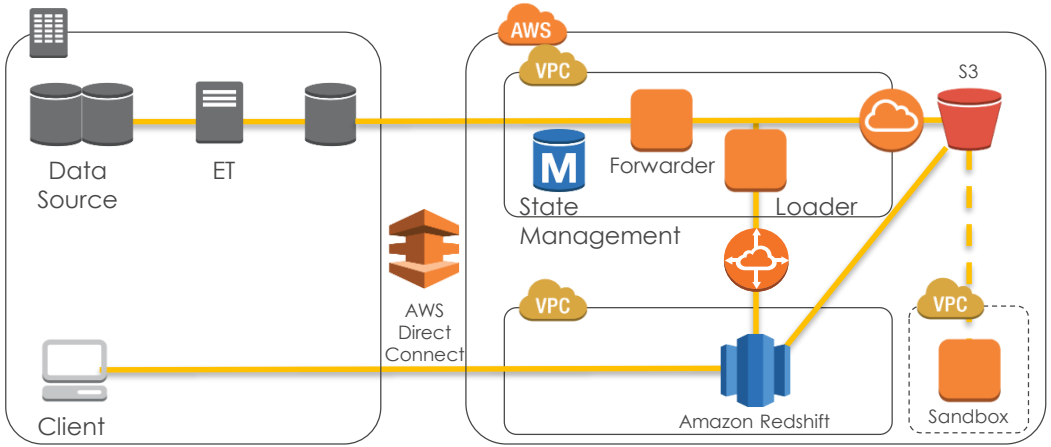
Scaling challenges

Performance issues

Need same level of security

Need for a hybrid environment

NTT Docomo: Japan's largest mobile service provider



125 node DS2.8XL cluster

4,500 vCPUs, 30 TB RAM

2 PB compressed

10x faster analytic queries

50% reduction in time for new BI application deployment

Significantly less operations overhead

Nasdaq: powering 100 marketplaces in 50 countries

Orders, quotes, trade executions, market “tick” data from 7 exchanges

7 billion rows/day

Analyze market share, client activity, surveillance, billing, and so on

Microsoft SQL Server on-premises

Expensive legacy DW ($1.16 M/yr.)

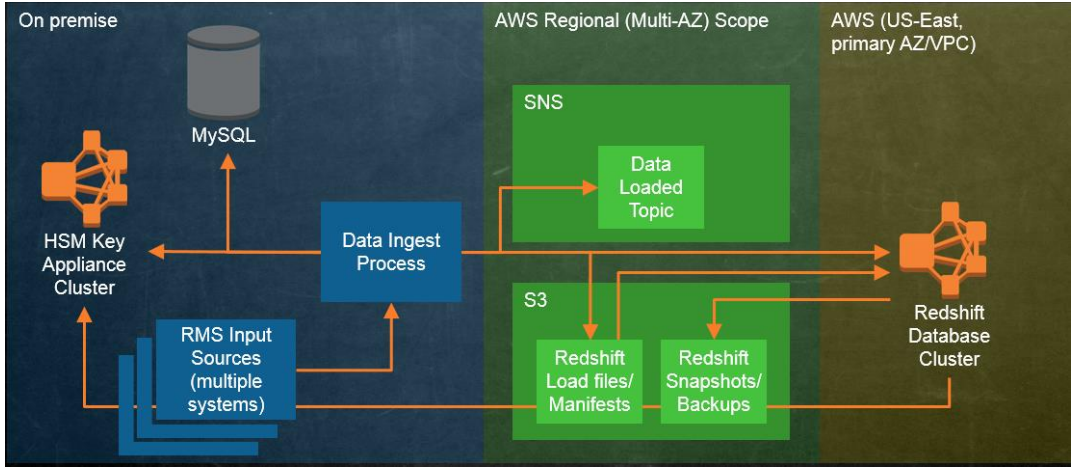
Limited capacity (1 yr. of data online)

Needed lower TCO

Must satisfy multiple security and regulatory requirements

Similar performance

Nasdaq: powering 100 marketplaces in 50 countries



23 node DS2.8XL cluster

828 vCPUs, 5 TB RAM

368 TB compressed

2.7 T rows, 900 B derived

8 tables with 100 B rows

---

7 man-month migration

¼ the cost, 2x storage, room to grow

Faster performance, very secure

Amazon.com clickstream analytics

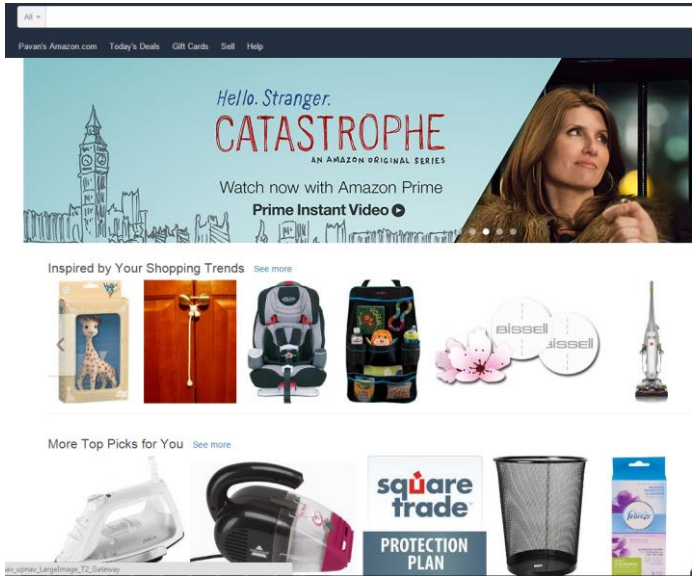
Web log analysis for Amazon.com

- PBs workload, 2TB/day@67% YoY
- Largest table: 400 TB

Understand customer behavior

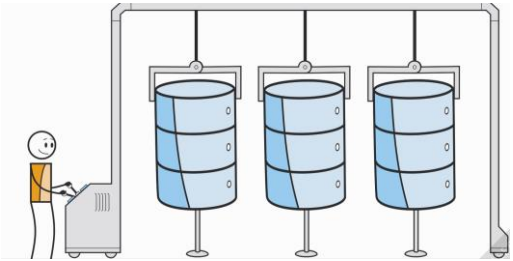
Previous solution

- Legacy DW (Oracle)—query across 1 week/hr
- Hadoop—query across 1 month/hr



Results with Amazon Redshift

- Query 15 months **in 14 min**
- Load 5B rows in **10 min**
- 21B w/ 10B rows: **3 days to 2 hrs** (Hive → Redshift)
- Load pipeline: **90 hrs to 8 hrs** (Oracle → Redshift)



- 100 node DS2.8XL clusters
- Easy resizing
- Managed backups and restore
- Failure tolerance and recovery

- 20% time of one DBA
- Increased productivity

Resources

Detail Pages

- http://aws.amazon.com/redshift
- https://aws.amazon.com/marketplace/redshift/
- https://aws.amazon.com/redshift/developer-resources/
- Amazon Redshift Utilities - GitHub

Best Practices

- http://docs.aws.amazon.com/redshift/latest/dg/c_loading-data-best-practices.html
- http://docs.aws.amazon.com/redshift/latest/dg/c_designing-tables-best-practices.html
- http://docs.aws.amazon.com/redshift/latest/dg/c-optimizing-query-performance.html

Thank you!