AWS re:Invent

ANT371

Migrate Your On-Premises Data Warehouse to Amazon Redshift with AWS DMS and AWS SCT

Shree Kenghe
Solutions Architect
AWS

Wesley Wilk
Solutions Architect
AWS

Ram Palaniappan
Sr. Director Data Analytics and Insights
TEKsystems

AWS re:Invent

aws

Agenda

- Workshop Introduction
- Lab architecture and environment setup
- Amazon Redshift overview
- AWS Schema Conversion Tool overview
- Migration considerations
- Workshop lab

Workshop Introduction

AWS re:Invent

aws

Workshop Details

- Workshop duration
  - Proposed solution and AWS services presentation - 20 minutes
  - Hands-on workshop - 1.5 hours
- Workshop team
  - Shree Kenghe, AWS Solutions Architect
  - Wesley Wilk, AWS Solutions Architect
  - Arun Kannan, Partner Solutions Architect
  - Ram Palaniappan, Sr. Director Data Analytics and Insights, TEKSystems
- Requirements and expectations
  - Students use their own AWS accounts to run the lab with IAM admin permissions
  - Basic knowledge of AWS services (Amazon RDS, Amazon Simple Storage Service (Amazon S3), AWS Database Migration Service, AWS SCT, Amazon Redshift)
  - Comfortable working on the AWS console and configure AWS services
  - Working knowledge of relational databases (Oracle)

Related sessions

Monday, November 26th Builder's Session

Modernize Your Data Warehouse with Amazon Redshift
11:30am PST | Aria

Tuesday, November 27th Chalk Talk

Migrate from Teradata to Amazon Redshift: Best Practices with McDonald's
9:15am PST | Aria

Thursday, November 29th Chalk Talk

Migrate from Netezza to Amazon Redshift Best Practices with Financial Engines
11:30am PST | MGM

Friday, November 30th Chalk Talk

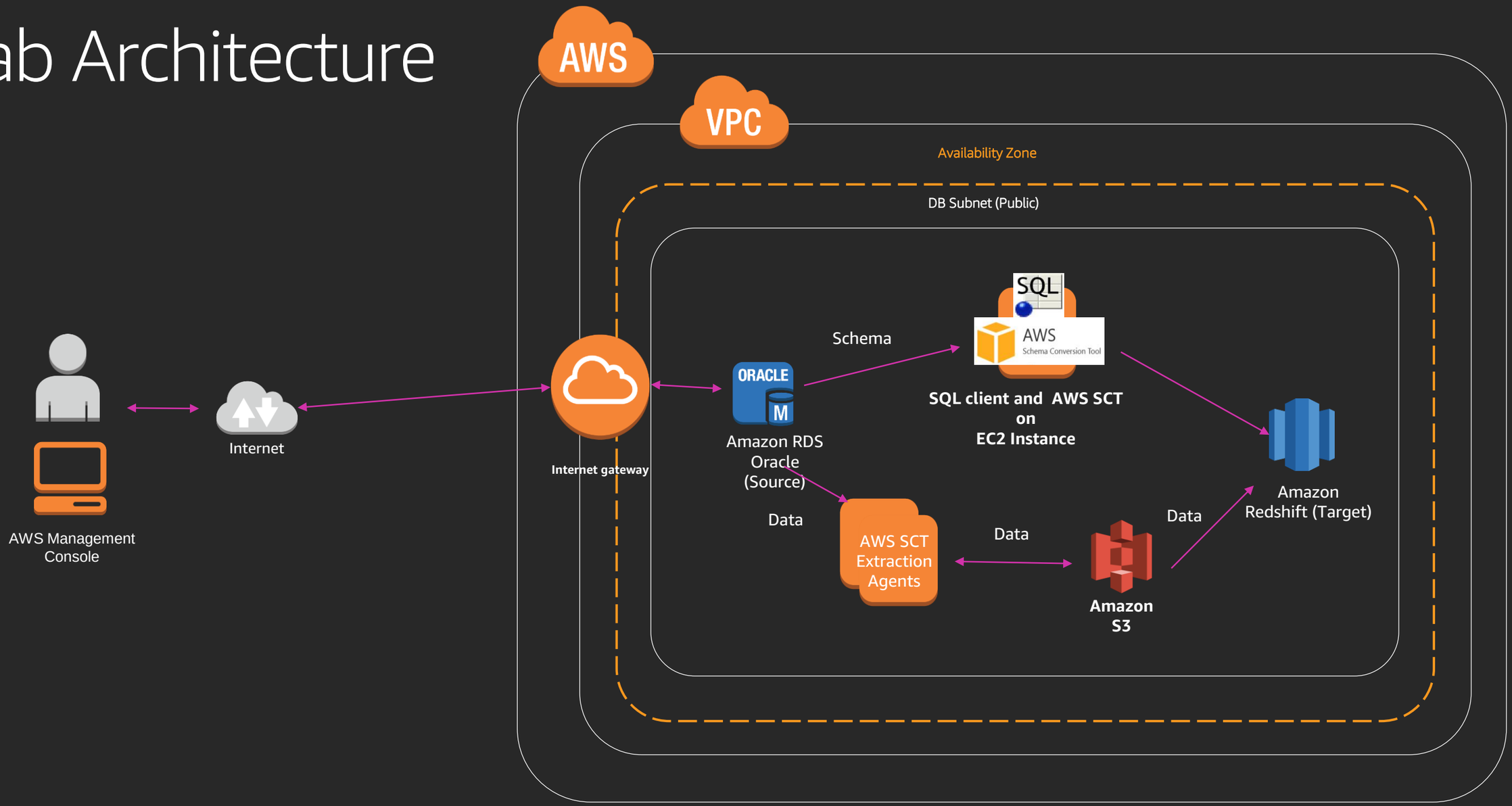
Migrating Workloads from Oracle to Amazon Redshift: Best Practices with Pfizer
11:30am PST | Mirage

Lab architecture and environment setup

AWS re:Invent

aws

Lab Architecture



aws

Lab Setup and Environment

- Download zip file: **https://bit.ly/2TkGyiu**
  - AWS CloudFormation template
  - Lab guide
  - SQL file
  - PowerPoint presentation
  - Policy file
- In AWS Management Console choose AWS Region **eu-west-1** (Ireland)
- Follow lab guide Step 1 - Launch AWS CloudFormation template

Amazon Redshift Overview

AWS re:Invent

aws

Amazon Redshift

10x faster at 1/10th the cost

Fast

Delivers fast results for all types of workloads

Simple

Create and start using a data warehouse in minutes

Cost-effective

No upfront costs, start small, and pay as you go

Scalable

Gigabytes to petabytes to exabytes

Integrated

Integrated with Amazon S3 data lakes, AWS services and third-party tools

Secure

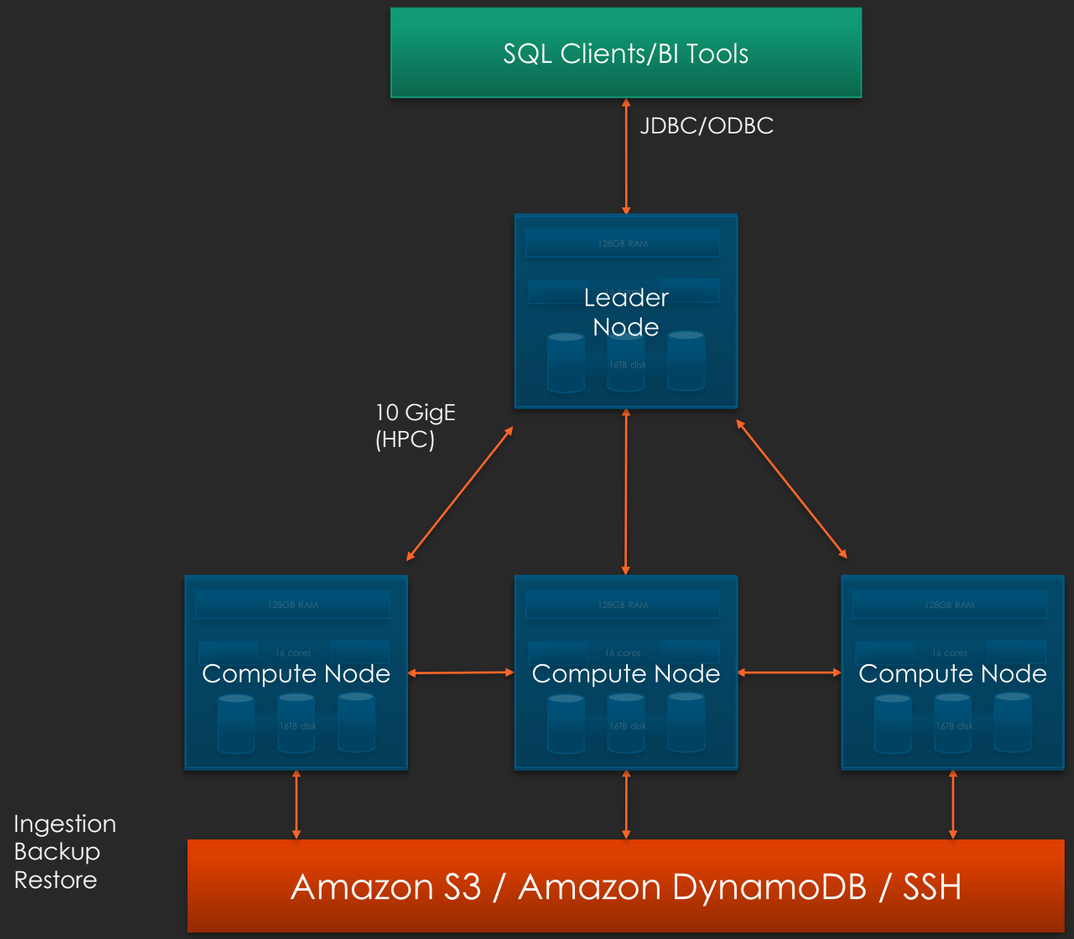
Audit everything; encrypt data end-to-end; extensive certification and compliance

AWS re:Invent

aws

Amazon Redshift Architecture

- Leader Node
  - SQL endpoint, JDBC/ODBC
  - Stores metadata
  - Coordinates query execution
- Compute Nodes
  - Local, columnar storage
  - Execute queries in parallel
  - Load, backup, restore via Amazon S3
  - Load from Amazon DynamoDB or SSH
  - Fault Tolerant
- Two hardware platforms
  - Optimized for data processing
  - DS2: HDD; scale from 2TB to 2PB
  - DC2: SSD; scale from 160GB to 326TB

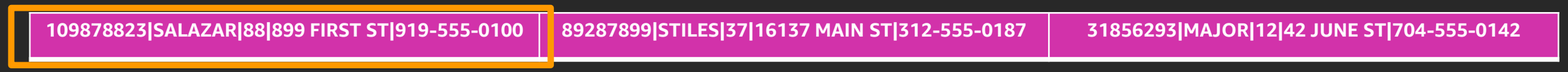


Amazon Redshift

Columnar Storage

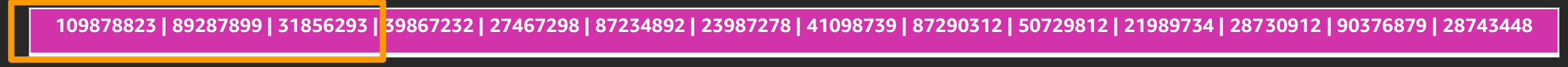
Row Storage

Customer ID	Name	Age	Address	Phone
10987823	SALAZAR	88	899 FIRST ST	919-555-0100
89287899	STILES	37	16137 MAIN ST	312-555-0187
31856293	MAJOR	12	42 JUNE ST	704-555-0142



Columnar Storage

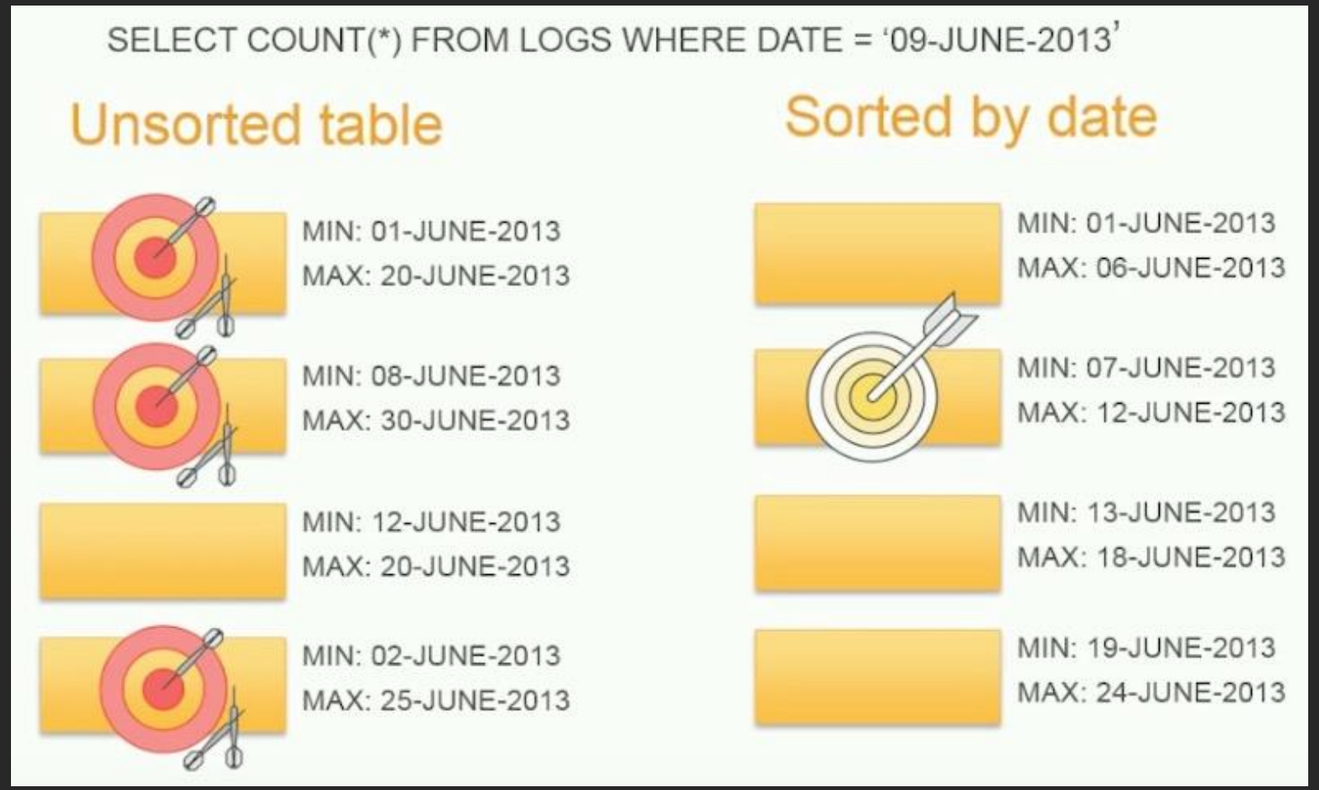
Customer ID	Name	Age	Address	Phone
10987823	SALAZAR	88	899 FIRST ST	919-555-0100
89287899	STILES	37	16137 MAIN ST	312-555-0187
31856293	MAJOR	12	42 JUNE ST	704-555-0142



Amazon Redshift

Zone Maps and Sort Keys

- Track the minimum and maximum value for each block
- Skip over blocks that don't contain relevant data



- Single Column
- Compound
- Interleaved

Amazon Redshift

Sort Keys – Single Column

- Best for Queries that use 1st column as primary filter
- Can Speed up joins and group bys
- Quickest to VACUUM

```
[ SORTKEY ( date ) ]
```

Date	Region	Country
2-JUN-2015	Oceania	New Zealand
2-JUN-2015	Asia	Singapore
2-JUN-2015	Africa	Zaire
2-JUN-2015	Asia	Hong Kong
3-JUN-2015	Europe	Germany
3-JUN-2015	Asia	Korea

Amazon Redshift

Sort Keys – Compound

- Table is sorted by 1st column, then 2nd, etc
- Best for Queries that use 1st column as primary filter, then others
- Can Speed up joins and group bys
- Slower to VACUUM

```
[ SORTKEY COMPOUND ( date, region, country) ]
```

Date	Region	Country
2-JUN-2015	Africa	Zaire
2-JUN-2015	Asia	Korea
2-JUN-2015	Asia	Singapore
2-JUN-2015	Europe	Germany
3-JUN-2015	Asia	Hong Kong
3-JUN-2015	Asia	Korea

Amazon Redshift

Sort Keys - Interleaved

- Equal weight is given to each column
- Best for queries that use different columns in filter
- Queries get faster the more columns used in the filter
- Slowest to VACUUM

```
[ SORTKEY INTERLEAVED ( date, region, country) ]
```

Date	Region	Country
2-JUN-2015	Africa	Zaire
3-JUN-2015	Asia	Singapore
2-JUN-2015	Asia	Korea
2-JUN-2015	Europe	Germany
3-JUN-2015	Asia	Hong Kong
2-JUN-2015	Asia	Korea

Amazon Redshift

Distribution Keys

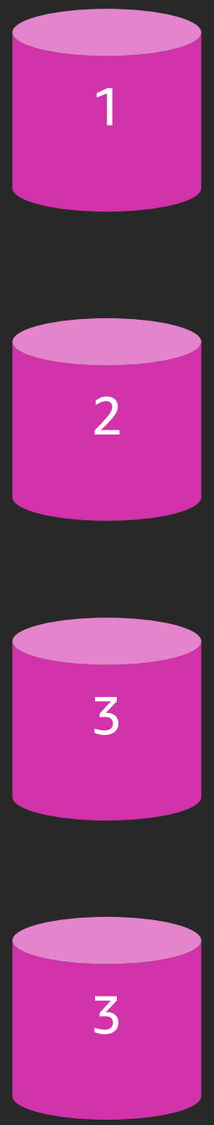
- EVEN
  - Tables with no joins or group by
  - Small Dimension tables (<1000)
- KEY
  - Large Fact tables
  - Large Dimension tables
- ALL
  - Medium Dimension tables (1K-2M)

Amazon Redshift

Distribution Keys - Even

ID	Gender	Name
101	M	Carlos Salazar
292	F	Li Juan
139	M	John Stiles
446	M	Arnav Desai
658	F	Mary Major
164	M	Mateo Jackson
209	M	Zhang Wei
306	F	Wang Xiulan

Round Robin



ID	Gender	Name
101	M	Carlos Salazar
306	F	Wang Xiulan

ID	Gender	Name
292	F	Li Juan
209	M	Zhang Wei

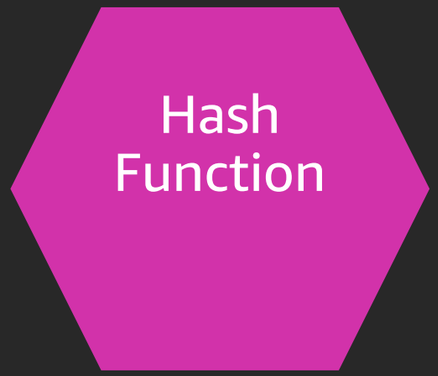
ID	Gender	Name
139	M	John Stiles
164	M	Mateo Jackson

ID	Gender	Name
446	M	Arnav Desai
658	F	Mary Major

Amazon Redshift

Distribution Keys - Even

ID	Gender	Name
101	M	Carlos Salazar
292	F	Li Juan
139	M	John Stiles
446	M	Arnav Desai
658	F	Mary Major
164	M	Mateo Jackson
209	M	Zhang Wei
306	F	Wang Xiulan



ID	Gender	Name
101	M	Carlos Salazar
306	F	Wang Xiulan

ID	Gender	Name
292	F	Li Juan
209	M	Zhang Wei

ID	Gender	Name
139	M	John Stiles
164	M	Mateo Jackson

ID	Gender	Name
446	M	Arnav Desai
658	F	Mary Major

Amazon Redshift

Best Practices & Pointers

- Maximize Load Performance
  - COPY multiple files
  - COPY to multiple nodes
  - Compress source data
  - Use a manifest file
- Amazon Redshift does not enforce primary key constraints
  - If you load data multiple times, Amazon Redshift will not complain
- After Loading
  - Data all added at end of columns for speed
  - Fully functional, but not set for optimum performance
- VACUUM command
  - Massages data to optimum disk organization for performance

AWS SCT Overview

AWS re:Invent

aws

What are AWS Database Migration Service and AWS SCT?

AWS DMS easily and securely migrates and/or replicate your databases *and* data warehouses to AWS

AWS SCT converts your commercial database and data warehouse schemas to open-source engines or AWS-native services, such as Amazon Aurora and Amazon Redshift

When to use AWS SCT?

Modernize

Modernize your database tier

ORACLE

Microsoft SQL Server

IBM DB2

PostgreSQL

MySQL

Modernize and **Migrate** your Data Warehouse to Amazon Redshift

Microsoft SQL Server

NETEZZA

TERADATA

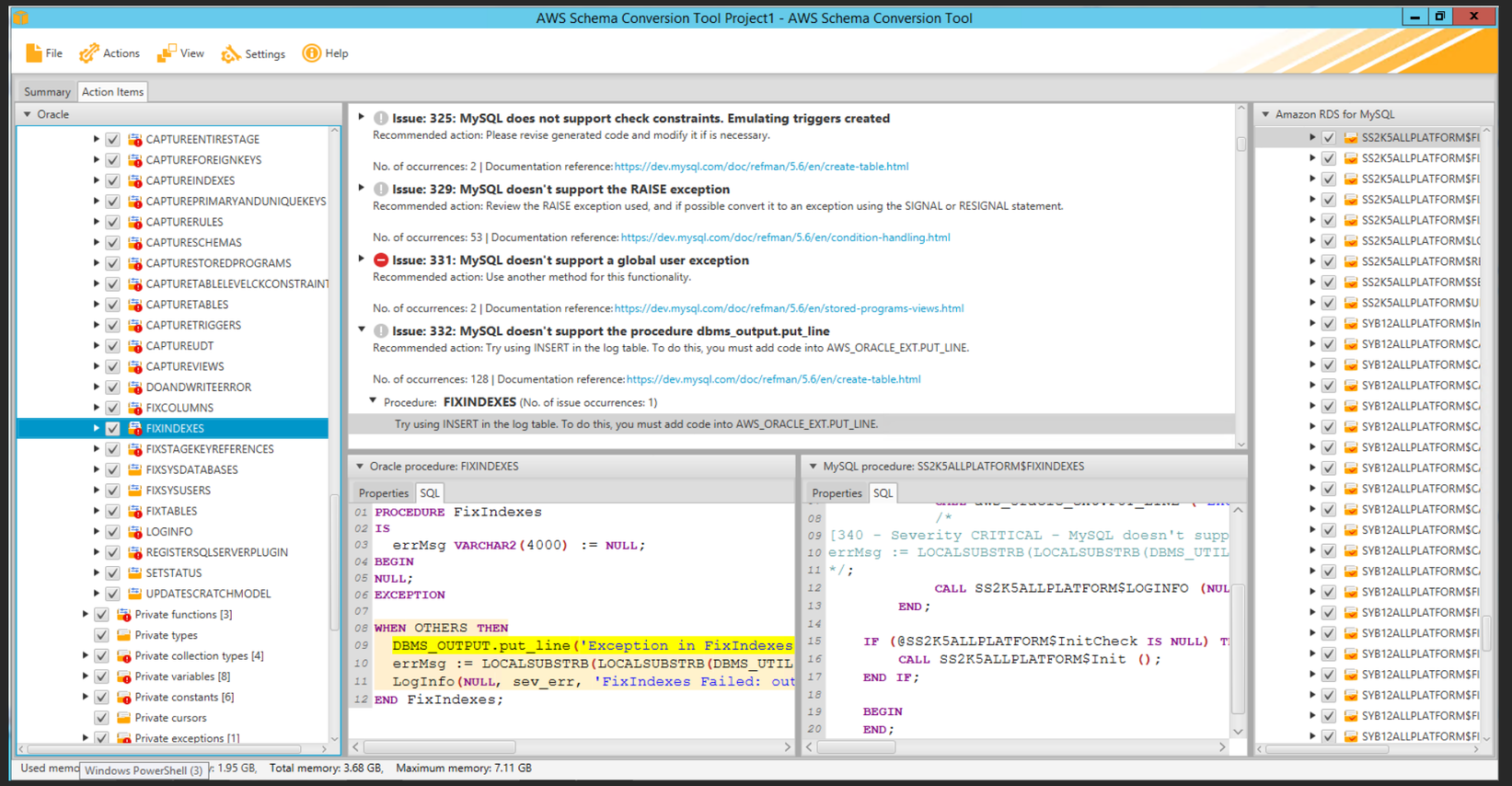
VERTICA

Greenplum

ORACLE

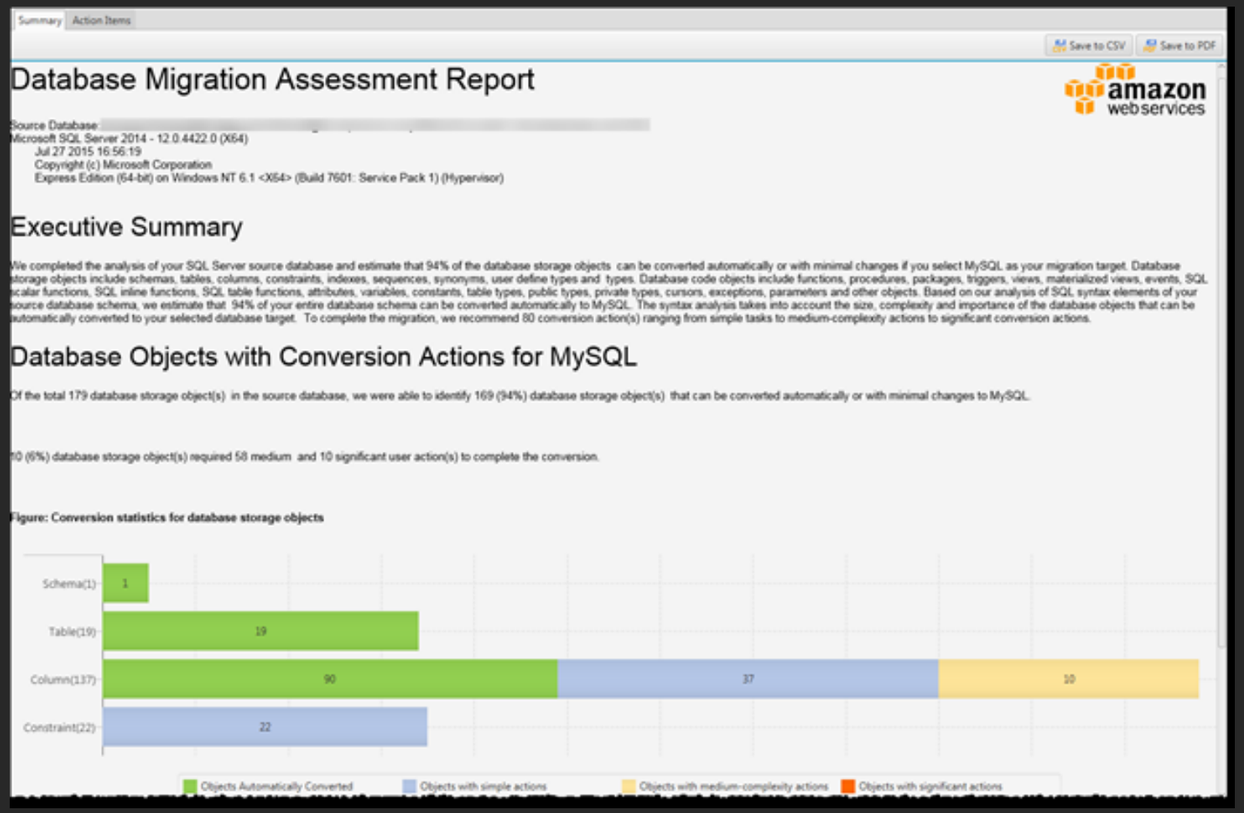
aws

AWS SCT helps with converting tables, views, and code



- Sequences
- User-defined types
- Synonyms
- Packages
- Stored procedures
- Functions
- Triggers
- Schemas
- Tables
- Indexes
- Views
- Sort and distribution keys

AWS SCT Migration Assessment Report



- Assessment of migration compatibility of source databases with open-source database engines – Amazon RDS MySQL, Amazon RDS PostgreSQL and Amazon Aurora
- Recommends best target engine
- Provides details level of efforts to complete migration

AWS SCT Data Extractors

Extract Data from your data warehouse and migrate to Amazon Redshift

- **Extracts** data through local migration agents
- Data is **optimized** for Amazon Redshift and saved in local files
- Files are **loaded** to an Amazon S3 bucket (through network or AWS Snowball) and then to Amazon Redshift

VERTICA

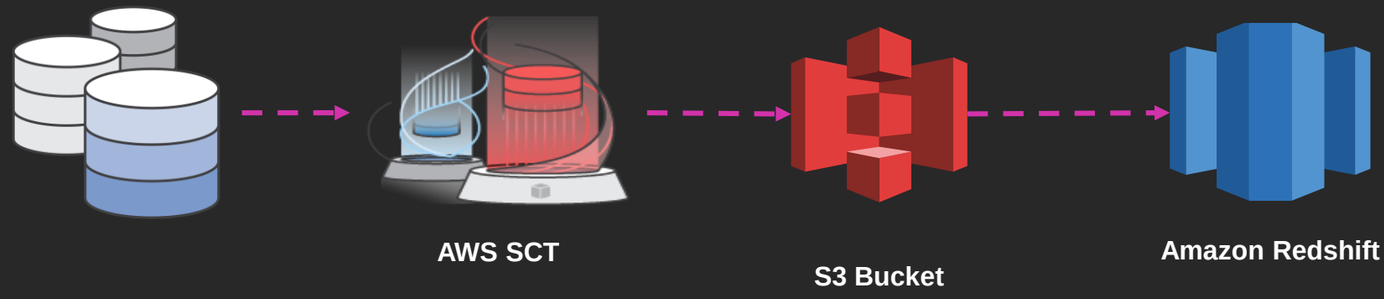
ORACLE

Microsoft SQL Server

Greenplum

TERADATA

NETEZZA

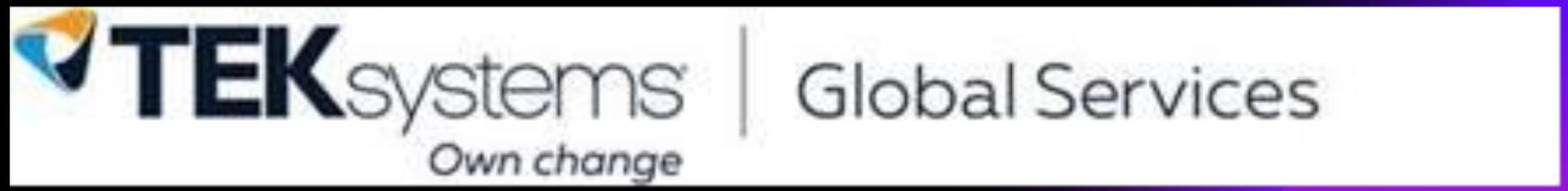


aws

AWS DMS + Snowball

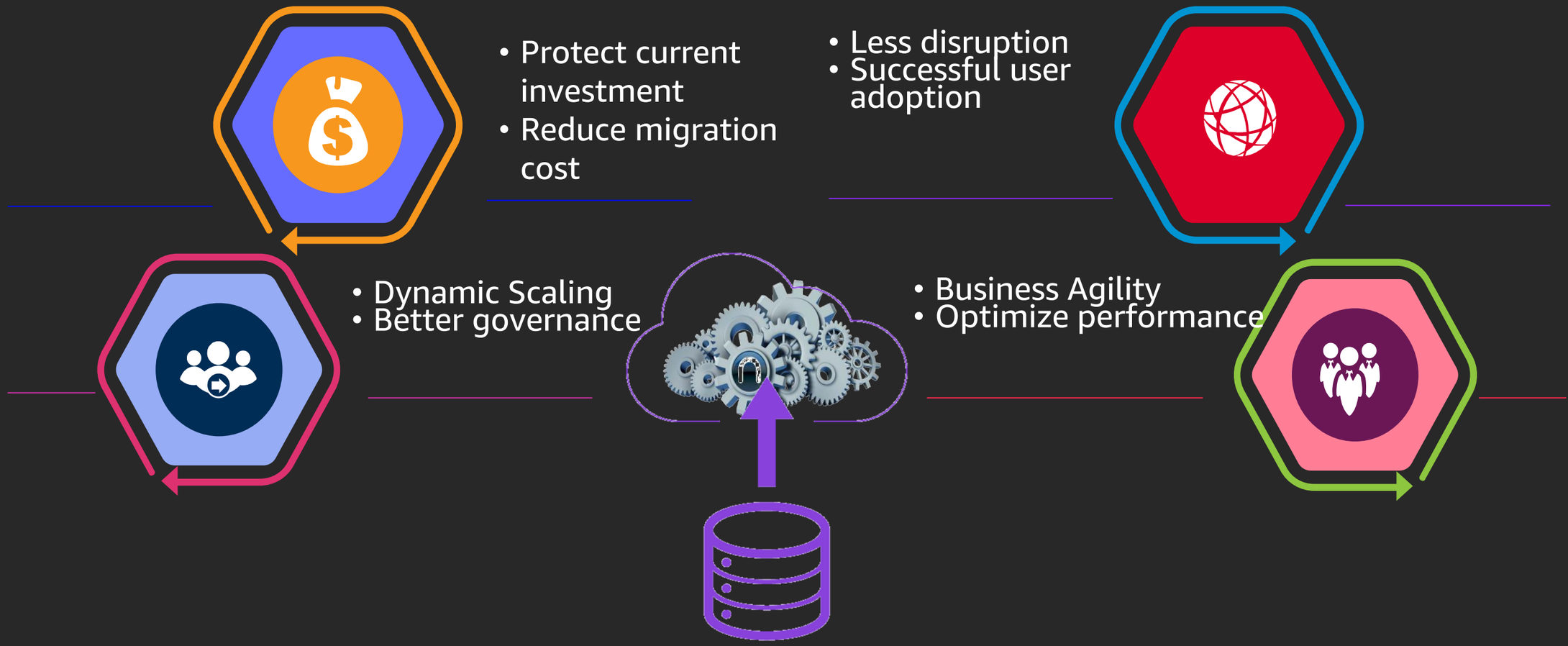
Common use cases

- Migrate large databases (over 5TB)
- Migrate many databases at once
- Migrate over slow network
- Push vs. Pull



Migration Considerations

Migration Objectives



Considerations for Migration

Data Movement – Source to DW	**One Time**: AWS SCT, AWS DMS **Incremental:** Informatica, Attunity, Talend, Alooma, Oracle DI
Star Schema & modeling	Amazon S3, Amazon Redshift, Amazon DynamoDB
Transformation Logic	AWS Data Pipeline, AWS Glue, Informatica, Talend
Aggregates, Snapshots	AWS Lambda function, ELT tools
Data refresh	Scheduling
Data Security.	Tokenization

TEKsystems Approach

Define

- Discover & Profile Current Datawarehouse - Tools: AWS SCT
- Assess & Capture Metadata **- Tools: AWS SCT**

Design

- Establish Reference Architecture - TEKsystems
- Schema Generation - Tools: AWS SCT
- ETL Mapping migration plan – TEKsystems Lineage

Move Data

- Initial data migration – AWS SCT DE, AWS DMS
- Data Validation – TEKsystems DW Scanner

Move Process

- Segregate ETL jobs – TEKsystems Lineage
- Retrofit or migrate to AWS ecosystems – TEKsystems Genie, Glue Converter

Optimize

- Tune for cloud native functions and features – TEKsystems Optimizer

Success Story

Jack in the box®

Technology

aws

tableau

AWS SCT
AWS Lambda
AWS Glue
Amazon Athena
Amazon S3
Amazon Redshift

AWS re:Invent

Vision :

- Build a global view of franchise operations
- Modernize the existing data platform that can be agile and scale
- Drive more self-service reporting for the sales to get better insights

Solution Delivered

- Data engineering using AWS Lambda, AWS Glue, AWS Step Functions and Amazon Athena
- Data warehousing using Amazon Redshift
- Data lake with Amazon S3 and Amazon Athena
- Analytics with Tableau Server on Amazon Elastic Compute Cloud (Amazon EC2) – in a load-balanced secure configuration using AWS WAF

Benefits

Single canonical view of all of the franchise data

Reduced the overall licensing cost and footprint on Oracle, SQL Server workloads

aws

Learnings

- Pilot the migration with migrating incremental data periodically
- Create design patterns based on data type, volume, frequency
- Leverage automation approach to convert transformation and aggregation ETLs to cloud native (AWS Glue)
- Keep your target data model same to retrofit current reporting structure
- Lot to gain when you go cloud native and tune your Amazon Redshift after migration

At a glance

We're partners in transformation

As an industry leader in Full-Stack Technology Services, Talent Services and real-world application, we work with progressive leaders to drive change.

100+
locations worldwide

$4+ billion
annual revenue

$650 million
annual revenue
TEKsystems Global Services

Full-Stack Expertise and Services

Serving 6K+ Clients,
80% of the Fortune 500

98% client **retention**

94% **NPS score** for services delivered

30% **higher client satisfaction** vs. the competition

Global Delivery
seamlessly on shore, off shore, near shore

Industry-specific expertise

Experience the power of real partnership. TEKsystems.com

Lab

https://bit.ly/2TkGyiu

Thank you!

Shree Kenghe
Solutions Architect
AWS

Wesley Wilk
Solutions Architect
AWS

Ram Palaniappan
Sr. Director Data Analytics and Insights
TEKsystems

AWS re:Invent

aws

Please complete the session survey in the mobile app.

AWS re:Invent

aws