Designing Big Data Analytics on the Cloud With Oracle Applications, Redshift & Tableau

Session ID: 10333

An architecture overview of designing a cloud based big data analytic solution - Leverage best practices & avoid common pitfalls.

Prepared by:

Sam Palani
Director, Infrastructure & Cloud Solutions
CTR, Inc.

@samx18

Introduction – Sam Palani

- Engaged with CTR as the Director, Infrastructure & Cloud Solutions.
- 14+ Years working with Oracle and related products: Oracle RDBMS, Oracle Applications, Unix & Linux
- 4+ Years building and supporting enterprise class data driven applications that leverage the cloud.
- Certified Oracle Specialist, Six Sigma Green Belt, Java Programmer, PMP & CSM.
- More about me ? http://samx18.io

Introduction - CTR

- Global Systems Integrator – North Americas & Asia Pacific locations.
- Oracle Platinum Partner.
- AWS & Tableau Partner.
- AWS, Tableau & Oracle certified strategically located across the globe.
- Working with clients since 1998 to Deploy & Support Data Driven Applications both on site and in a cloud model.

More about CTR? http://ctrworld.com

Session Agenda

- Summary – What the session will cover
- Pre-requisites and Assumptions
- The Data Problem
- The Enterprise Data Warehouse - Redshift
- Data Loading & Transformations
- Data Visualizations – Tableau
- Live Data & Data Extracts
- Live Demo
- Additional Resources – Where I can explore more
- Questions / Feedback

Summary

This session will cover how to leverage Amazon Web Services (AWS) Redshift to design a data warehouse on the cloud for Oracle Applications and integrate it with data visualization tools like Tableau.

The session will focus on best practices and tips & tricks to perform fast data analysis as well discover hidden insights in EBS data.

We will also discuss the best data load strategies specific to Oracle Applications.

The session will be aimed mainly at technical audiences.

Prerequisites and Assumptions

The topics discussed in this presentation and related white paper are of an intermediate to an advanced level.

It is assumed that readers and participants have a good understanding of Oracle products and concepts, namely Oracle Database, Oracle Applications as well as a basic understanding of the current enterprise cloud models available with AWS.

In addition to that you will also need an active Oracle Technology Network account and an active Amazon Web Services account and an account with Tableau.

The Data Problem

How did we end up here?

As business expand their Oracle Applications / ERP footprint it starts to get increasingly challenging to manage and consolidate all their reporting data.

Segregation of transactional data from the reporting the data is just one of the challenges.

The Data Problem

Lets briefly look at the challenges below

Data Size	Data Complexity	Multiple Data Sources	Performance	Scalability
• Data Growth • Low storage costs • Analyze & report on big data	• Normalized data • Summarized data • Transformations needed	• ERP data • Interface systems • External sources	• Trade-off between functionally & performance • Transactional system vs. Reporting system	• Non constant reporting needs • Regular vs. peak phases • Respond to changing business demands

Enter The Enterprise Data Warehouse

- ✓ Enterprise data warehouses have attempted to address some or all of the concerns listed.
- ✓ Historically traditional data warehouses have not been easy to design, build and maintain.
- ✓ Often required a significant up front investment in terms and capital as well as resources.
- ✓ Not easy to scale up or ramp down.
- ✓ Significant investment in terms of supporting infrastructure.

Cloud Data Warehouse - Redshift

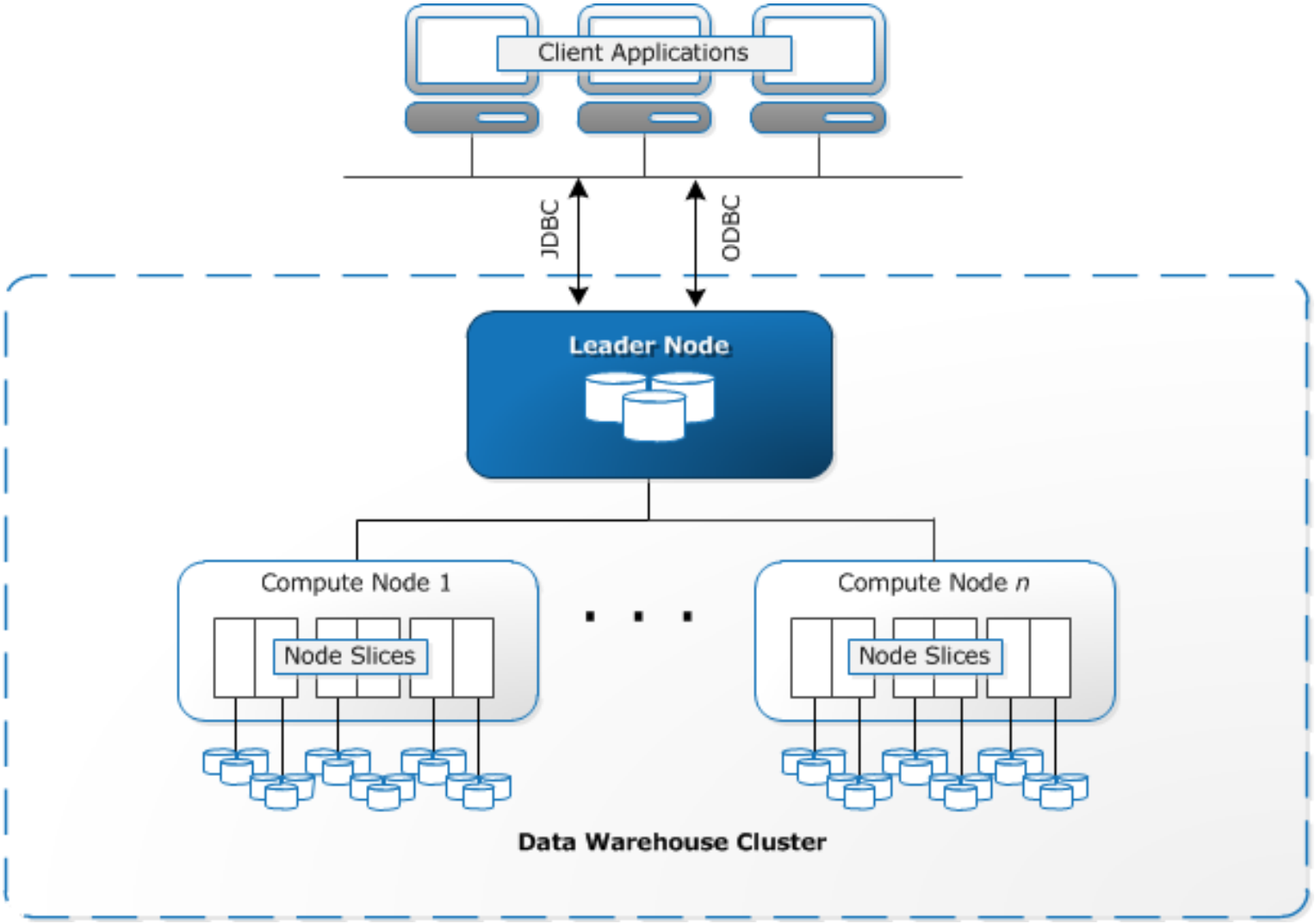
- ✓ AWS Redshift is an on demand data warehouse solution hosted on Amazon AWS cloud.
- ✓ An Amazon Redshift data warehouse is a collection of computing resources called nodes, which are organized into a group called a cluster.
- ✓ Each cluster runs an Amazon Redshift engine and contains one or more databases.

AWS Components

In addition to Redshift, AWS offers a variety of IaaS and PaaS products, however for the purpose of this discussion and the presentation, we will be leveraging these specific tools.

- **S3 Blocks Simple Storage Service (S3)** are blocks of storage that you can access on demand.
- **IAM & Security Groups** AWS Identity and Access Management (IAM) enables you to securely control access to AWS services and resources for your users.
- **VPC** - Amazon Virtual Private cloud (VPC) - A virtual private cloud (VPC) is a virtual network dedicated to your account. It is logically isolated from other virtual networks in the AWS cloud.

Redshift Architecture



Redshift Architecture

Key components of the Redshift architecture

- AWS Region
  - Region where your cluster will be created.
  - Cannot be changed.
- Clusters
  - One or more nodes.
- Leader Nodes
  - Each cluster will have one leader.
  - Responsible for all management & assignment tasks.
- Compute Nodes –
  - One or more compute nodes.
  - Single node cluster the same node acts as leader & compute.

Redshift Architecture

Key components of the Redshift architecture

- Database
  - Clusters have one or more databases.
  - Default DB is created when the cluster is launched.
  - Additional DBs are created manually after connecting to the cluster.
- Access Control
  - Supports database level access control.
  - Also supports AWS IAM level access control.
  - Application level access control via security groups.
- SSL
  - Allows encrypted connection to the cluster.
  - SSL does not cover user authentication or data encryption.

Setup Prerequisites for Redshift

Couple of things you need to take care before you can launch & connect.

- A SQL client
  - Client that supports JDBC connections
  - SQL Workbench J
- Firewall Rules
  - Default port of 5439
  - Inbound & outbound rules to accept TCP for 5439
- Security group settings

Redshift Cluster Launch

CLUSTER DETAILS | NODE CONFIGURATION | ADDITIONAL CONFIGURATION | REVIEW

You are about to launch a cluster with following the following specifications:

Cluster Properties

These attributes specify the name of your cluster, what type of virtual hardware it will run on, how many nodes it will contain, and the availability zone in which it will be located.

Cluster Identifier: oaug16

Node Type: dc1.large

Number of Compute Nodes: 3 (plus a free leader node)

Availability Zone: No Preference

Database Configuration

These properties specify the database name, port, and username you will use to connect to the database. The parameter group contains configuration values used by the database.

Database Name: sampledb

Database Port: 5439

Master User Name: sam

Cluster Parameter Group: default.redshift-1.0

Security, Access, and Encryption

These settings control whether your cluster will be created in an existing VPC to allow for simpler integration with other AWS Services, and the security groups which define access rules to your cluster.

Virtual Private Cloud: Not in VPC

Publicly Accessible: Yes

Elastic IP: Not used

Cluster Security Groups: default

Encrypt Database: No

CloudWatch Alarms

CloudWatch alarms are used to notify if metrics for your cluster are within a certain threshold. All recipients under the SNS topic specified for your alarm will receive notifications once an alarm is triggered.

Basic alarms will not be created for this cluster.

Redshift Best Practices

- Sort Keys
  - Based on the kind of data queried most often.
  - Recently update data, use the timestamp.
  - For joined tables, use joined column.
  - For range filters, use the filer column.
- Distribution Style
  - Minimize redistribution while query execution.
- COPY compression
  - Use native copy command to enable compression during data loads.
- Primary & Foreign Key constrains
  - Informational only.
  - But used by optimizer for efficient query plans.

Redshift Best Practices

- Smallest Possible Column Size
  - Compression by default.
  - Complex queries need temporary space.
- Date & Timestamp columns where appropriate

Redshift Data Load

- Redshift can load data in parallel
  - Almost proportional to the number of nodes
- COPY command to load data from S3
  - Use a manifest file to enforce data load constraints.
  - Use multiple input files.
  - Automatic compression.
- Using a data pipeline

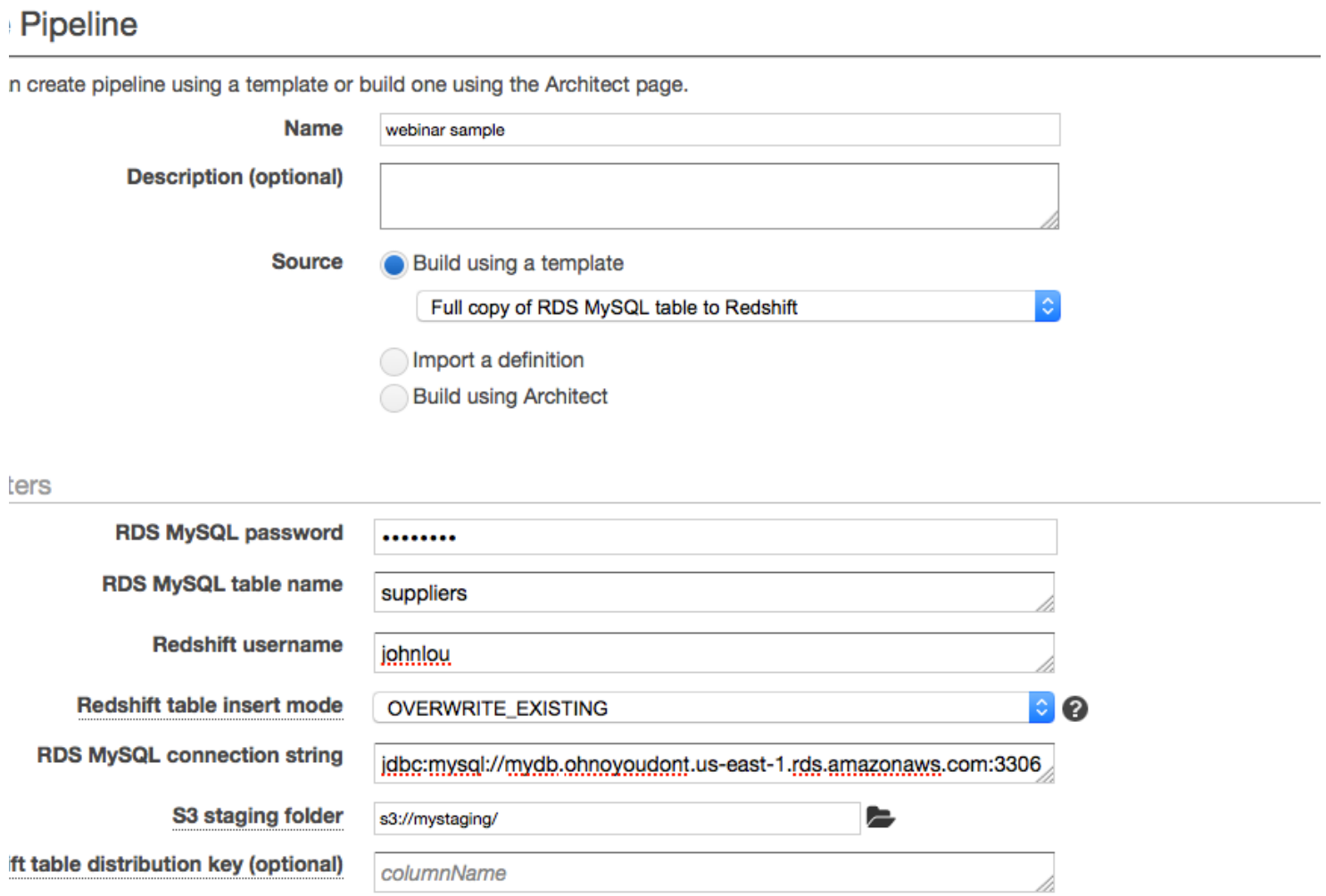
- ETL load from direct data sources

Redshift Data Load – COPY

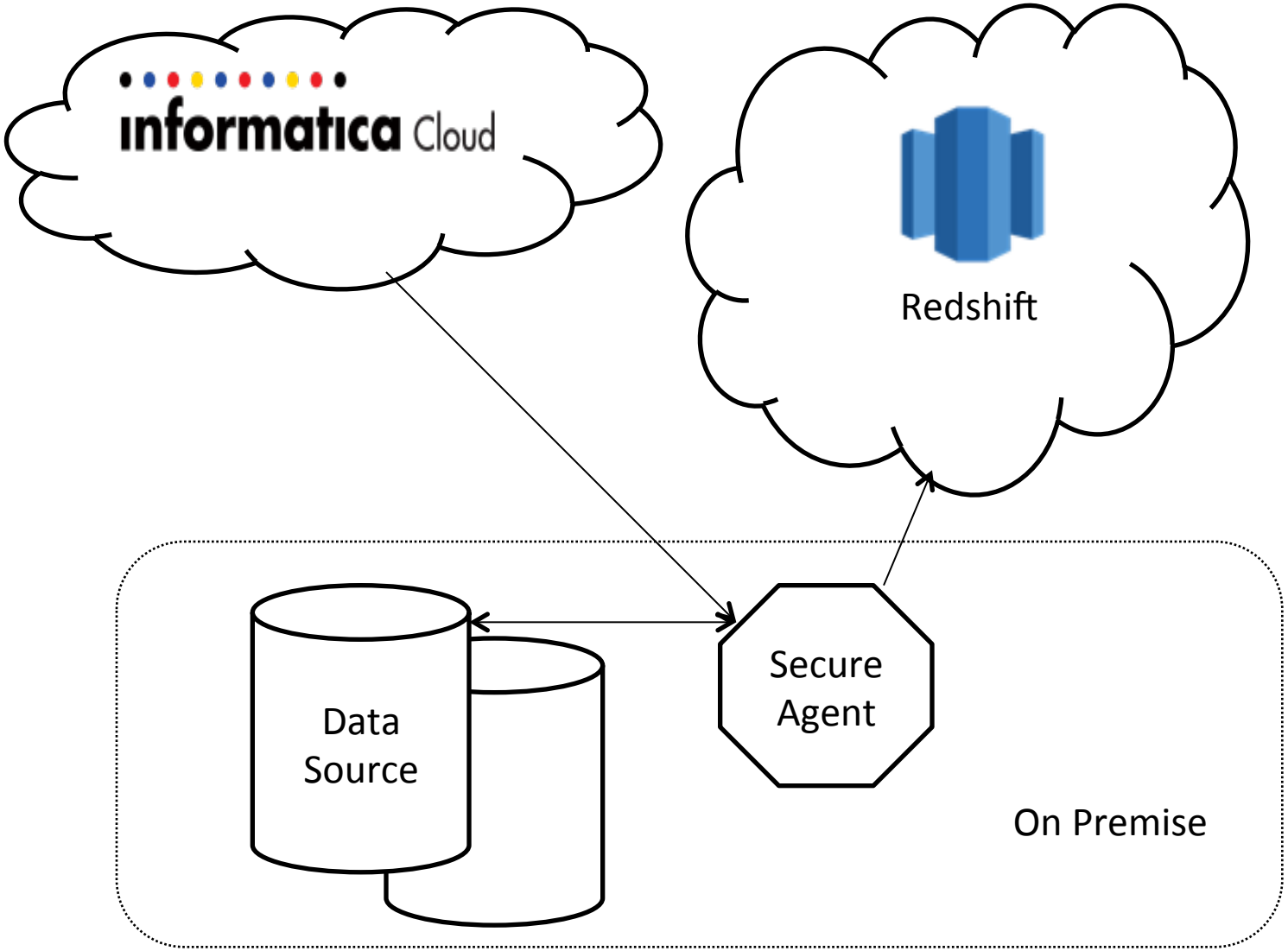
- COPY provides better performance compared to traditional inserts.
- Can be used to load data directly from S3 or an EC2 via SSH.
- Uses automatic compression by default to ensure the best performance.

copy users from 's3://awssampledb/tickit/allusers_pipe.txt'
credentials 'aws_access_key_id=XXXXXXXXXXX;aws_secret_access_key=XXXXXXXXXX'
delimiter '|';

Redshift Data Load – Data Pipeline



Redshift Data Load – Cloud ETL



Data Visualizations – Tableau

Tableau is a visual data exploration tool that can efficiently work with multiple data sources to create interactive data visualizations and dashboards.

- Visual Data Exploration
  - Think visually.
  - Automatically composes the queries and analytical computations needed to create the picture.
- Multiple Data Sources
  - Redshift, Oracle & many more.
  - No support yet for NoSQL.

Data Visualizations – Tableau

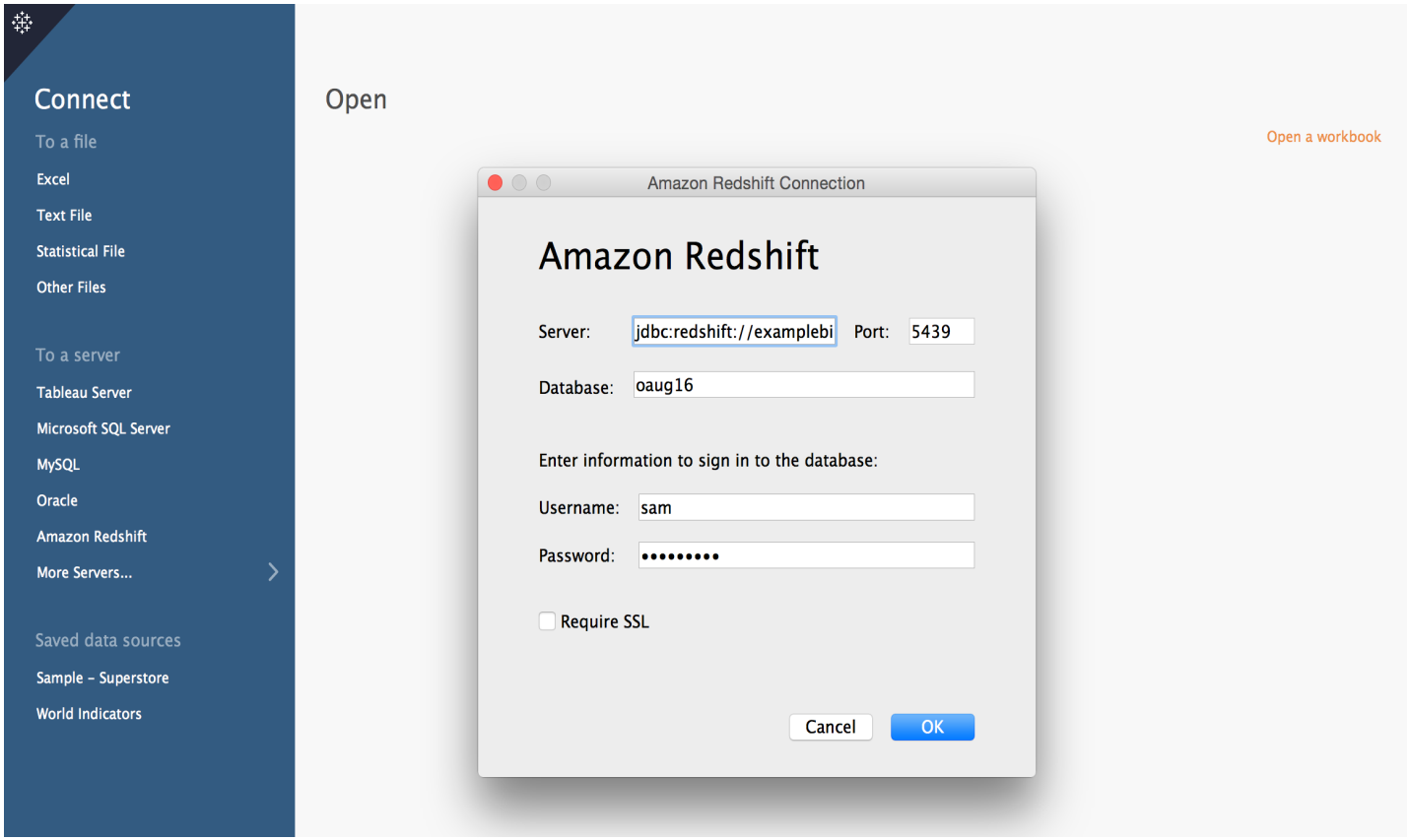
- Database Independence
  - No dependence on how data is stored.
- VIZQL
  - Built upon the industry standard SQL & MDX.
- Publishing Options
  - Interactive reports via server
  - Static reports as PDFs etc.
  - Data exports to excel etc.

Tableau Components

- Tableau Desktop
  - The main the main tool where you will design and create your workbooks and dashboard.
  - Both windows and Mac platforms, however the Mac version has a few feature limitations.
- Tableau Server
  - Used to serve the reports and dashboard to the users.
  - Reports & dashboards are published from the desktop to the server.
  - Can only be installed on a windows server.
- Tableau Online
  - Hosted software as a service solution.
  - All features of Tableau server are available on Tableau online.

In addition to the above you also have tableau public where you can publish your workbooks to a public domain. This is not designed for enterprise use.

Tableau Redshift Connection



Live VS Data Extracts

- Tableau can make both live connection to your data source as well as use data extracts (TDE).
- TDE is a compressed snapshot of data stored on disk and loaded into memory as required.
- TDE has a couple of advantages with Redshift
  - TDE is a columnar store similar to the columnar format that is used in Redshift itself.
  - Key aspect of TDE design is how they are structured which impacts how they are loaded into memory and used by Tableau.
  - Also enables you to add portability to your application.

Live Demo

Putting it all together in a live demo

Additional Resources

Informatica

Getting Started with Informatica Cloud - http://videos.informaticacloud.com/HY6u/getting-started-with-informatica-cloud/

AWS Documentation

Redshift - https://aws.amazon.com/redshift/getting-started/

Computing - http://aws.amazon.com/ec2/

AMI - https://aws.amazon.com/marketplace

Tableau

Getting started with Tableau - http://www.tableau.com/learn/tutorials/on-demand/getting-started

Thank You

Sam Palani

spalani@ctrworld.com

@samx18

http://samx18.io