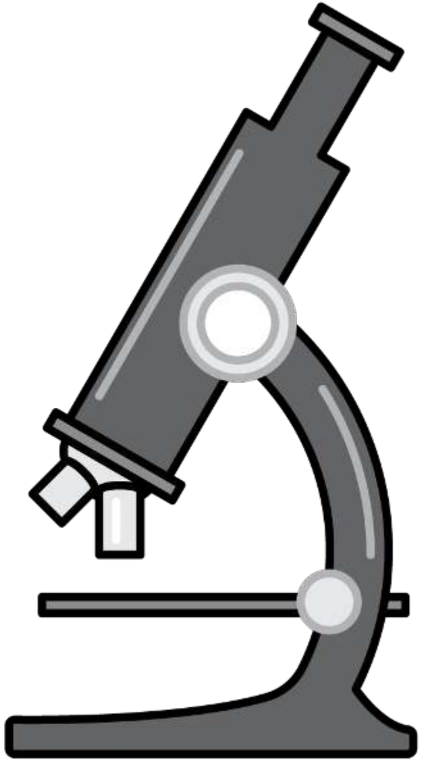
Picking the right AWS backend for your Java application

Julien Simon, Principal Technical Evangelist, AWS

julsimon@amazon.fr
@julsimon

What to expect

- Writing Java apps on AWS
- Databases
  - Amazon RDS
  - Amazon DynamoDB
- Analytics
  - Hive on Amazon EMR
  - Amazon Athena
  - Amazon Redshift
- Conclusion



Code available at https://github.com/juliensimon/aws/tree/master/javabackends

Writing Java apps on AWS

Four deployment options

Amazon EC2

Amazon EC2 Container Service

AWS Lambda

Java 8

Open Source frameworks:

Serverless, Gordon, Apex, ….

AWS ElasticBeanstalk

Java 6/7/8 with Tomcat 7/8

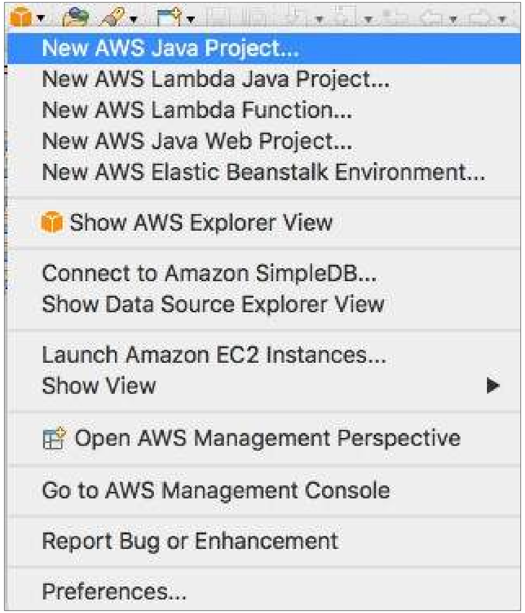
Java 7/8 with Glassfish 4

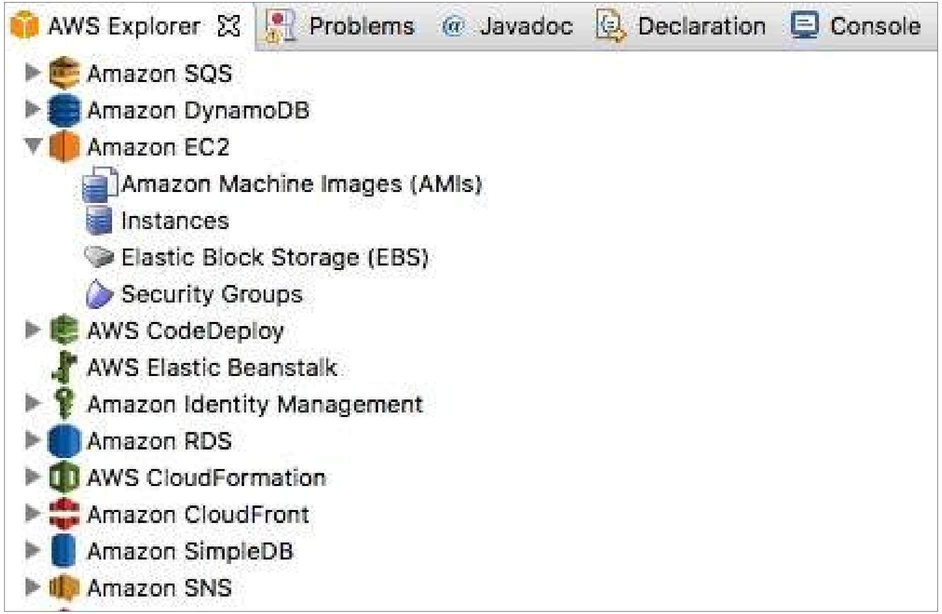
Java SDK for the AWS API (Java 1.6+)

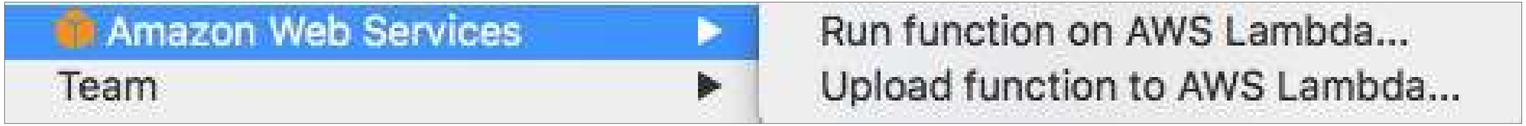
https://docs.aws.amazon.com/sdk-for-java/

https://github.com/aws/aws-sdk-java

AWS plugin for Eclipse


New AWS Java Project...
New AWS Lambda Java Project...
New AWS Lambda Function...
New AWS Java Web Project...
New AWS Elastic Beanstalk Environment...
Show AWS Explorer View
Connect to Amazon SimpleDB...
Show Data Source Explorer View
Launch Amazon EC2 Instances...
Show View
Open AWS Management Perspective
Go to AWS Management Console
Report Bug or Enhancement
Preferences...




Amazon Web Services
Team
Run function on AWS Lambda...
Upload function to AWS Lambda...

https://aws.amazon.com/documentation/awstoolkiteclipse/

3rd party plugins for Intellij IDEA

- AWS Elastic Beanstalk Integrationhttps://plugins.jetbrains.com/plugin/7274-aws-elastic-beanstalk-integration

- AWS CloudFormationhttps://plugins.jetbrains.com/plugin/7371-aws-cloudformation

- AWS Manager – almost 2 years old :-/https://plugins.jetbrains.com/plugin/4558-aws-manager

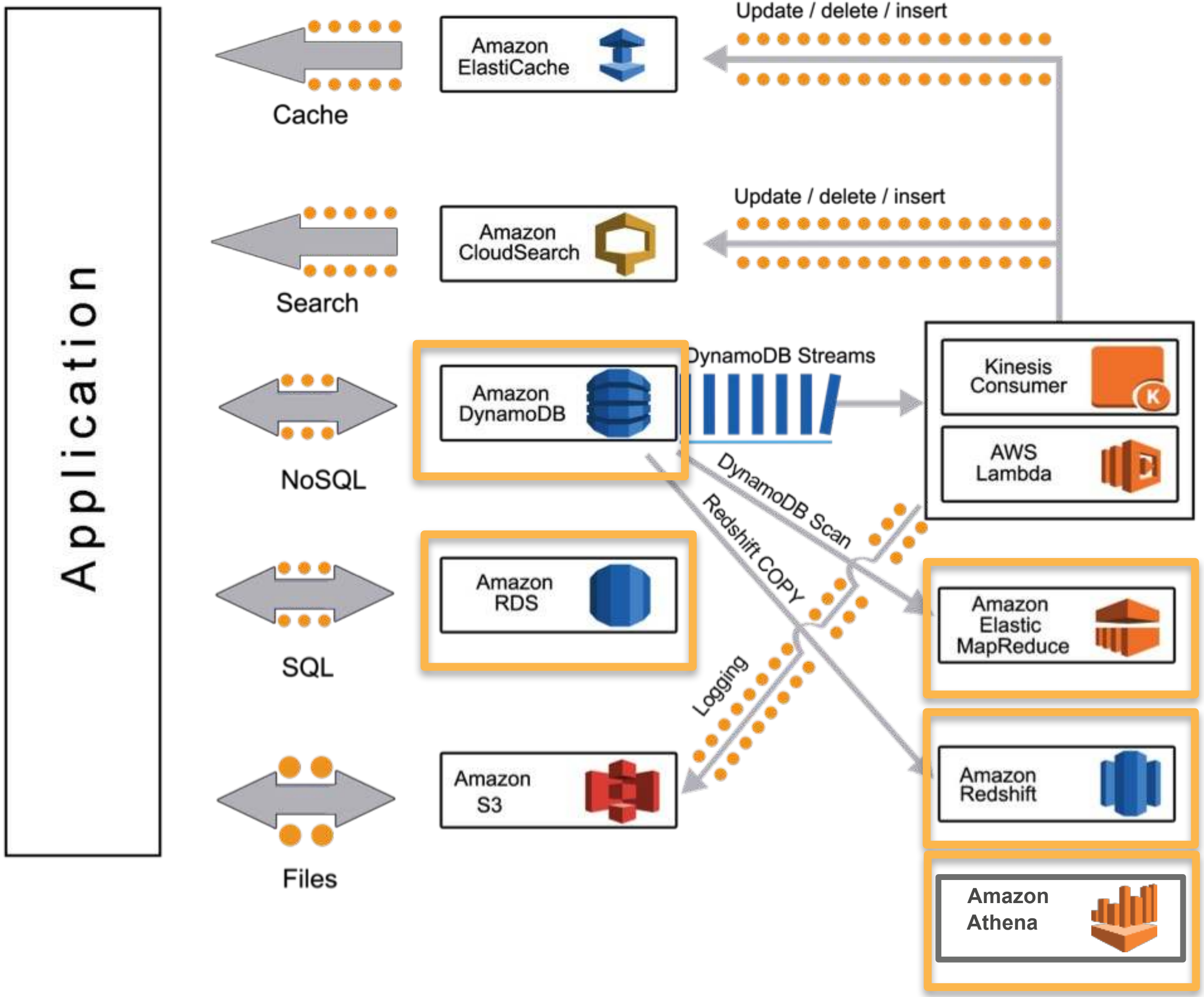
Managing credentials

- Please do not hardcode them in your application
- Please do not store them on EC2 instances
- It WILL end in tears!

- AWS credentials: use IAM Roles
- Backend credentials: use the EC2 SM Parameter Store
  - Automatic encryption with Amazon KMS

https://docs.aws.amazon.com/sdk-for-java/v1/developer-guide/credentials.html
https://docs.aws.amazon.com/AWSJavaSDK/latest/javadoc/index.html?com/amazonaws/auth/AWSCredentialsProvider.html
https://docs.aws.amazon.com/systems-manager/latest/userguide/systems-manager-paramstore.html

Reference architecture



Databases

Amazon Relational Database Service

AWS Free Tier

No infrastructure management

Application compatibility

Instant provisioning

Cost-effective

99.95% SLA

Scale up/down

https://aws.amazon.com/rds/
https://aws.amazon.com/rds/sla/

Amazon RDS – 6 Database Engines

- Amazon Aurora
- MySQL 5.5.46 → 5.7.16
- MariaDB 10.0.17 → 10.1.19
- PostgreSQL 9.3.12-R1 → 9.6.2-R1
- Oracle 11.2.0.4.v1 → 12.1.0.2.v7
- SQL Server 2008 → 2016

Aurora design principles https://dl.acm.org/citation.cfm?id=3056101

Selected Amazon RDS customers

SAMSUNG

vodafone

intuit

LAMBORGHINI

Unilever

airbnb

ZIFF DAVIS

SEGA®

Kempinski
HOTELIERS SINCE 1897

ROVIO

FCBARCELONA
més que un club

OUTBACK
STEAKHOUSE®

OVERSEAS VOTE
FOUNDATION

Newsweek
& THE DAILY BEAST

The Washington Post Company

coursera

C O
D E

ZUMBA®
FITNESS

Trinity Mirror plc

Amazon Aurora demo

Java SDK: https://docs.aws.amazon.com/AWSJavaSDK/latest/javadoc/com/amazonaws/services/rds/AmazonRDSClient.html
JDBC drivers: https://docs.aws.amazon.com/elasticbeanstalk/latest/dg/java-rds.html

Amazon DynamoDB

AWS Free Tier

Fully Managed NoSQL

Document or Key-Value

Scales to Any Workload

Fast and Consistent

Access Control

Event Driven Programming

https://aws.amazon.com/dynamodb/
http://www.allthingsdistributed.com/2007/10/amazons_dynamo.html
http://www.allthingsdistributed.com/2012/01/amazon-dynamodb.html

Case Study – Expedia

> “With DynamoDB we were up and running in a less than day, and there is no need for a team to maintain.”
>
> **Kuldeep Chowhan**
> Principal Engineer, Expedia

Expedia is a leader in the $1 trillion travel industry, with an extensive portfolio that includes some of the world’s most trusted travel brands.

- Expedia’s real-time analytics application collects data for its “test & learn” experiments on Expedia sites.
- The analytics application processes ~200 million messages daily.
- Ease of setup, monitoring, and scaling were key factors in choosing DynamoDB.

https://aws.amazon.com/fr/blogs/big-data/how-expedia-implemented-near-real-time-analysis-of-interdependent-datasets/

Amazon DynamoDB demo

`Low-level API: getItem, putItem, updateItem`
`batchGetItem, batchWriteItem`
`query, scan`

`High-level API:` *`DynamoDBMapper`*

Java SDK: https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/CodeSamples.Java.html
https://docs.aws.amazon.com/sdk-for-java/v1/developer-guide/examples-dynamodb.html
https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/DynamoDBLocal.html

	Amazon ElastiCache	Amazon DynamoDB	Amazon RDS/Aurora	Amazon ElasticSearch	Amazon S3	Amazon Glacier
Average latency	ms	ms	ms, sec	ms,sec	ms,sec,min (~ size)	hrs
Typical data stored	GB	GB–TBs (no limit)	GB–TB (64 TB max)	GB–TB	MB–PB (no limit)	GB–PB (no limit)
Typical item size	B-KB	KB (400 KB max)	KB (64 KB max)	B-KB (2 GB max)	KB-TB (5 TB max)	GB (40 TB max)
Request Rate	High – very high	Very high (no limit)	High	High	Low – high (no limit)	Very low
Storage cost GB/month	$$	¢¢	¢¢	¢¢	¢	¢4/10
Durability	Low - moderate	Very high	Very high	High	Very high	Very high
Availability	High 2 AZ	Very high 3 AZ	Very high 3 AZ	High 2 AZ	Very high 3 AZ	Very high 3 AZ

Hot data Warm data Cold data

Analytics

Amazon Elastic Map Reduce (EMR)

- Apache Hadoop, Spark, Hive,etc.
- Managed service
- Easy to start, resize & terminate clusters
- Cost-efficient, especially with Spot Instances
- Integration with backends

https://aws.amazon.com/emr/

Case study – FINRA

FINRA, the primary regulatory agency for stock brokers in the US, uses AWS extensively in their IT operations and has migrated key portions of its technology stack to AWS including Market Surveillance and Member Regulation.

For market surveillance, each night FINRA loads approximately 35 billion rows of data into Amazon S3 and Amazon EMR (up to 10,000 nodes) to monitor trading activity on exchanges and market centers in the US.

https://aws.amazon.com/solutions/case-studies/finra/

Hive demo

Java SDK: https://docs.aws.amazon.com/emr/latest/ManagementGuide/calling-emr-with-java-sdk.html
JDBC: https://docs.aws.amazon.com//emr/latest/ReleaseGuide/HiveJDBCDriver.html

Amazon Athena

- Run read-only SQL queries on S3 data
- No data loading, no indexing, no nothing
- No infrastructure to create, manage or scale

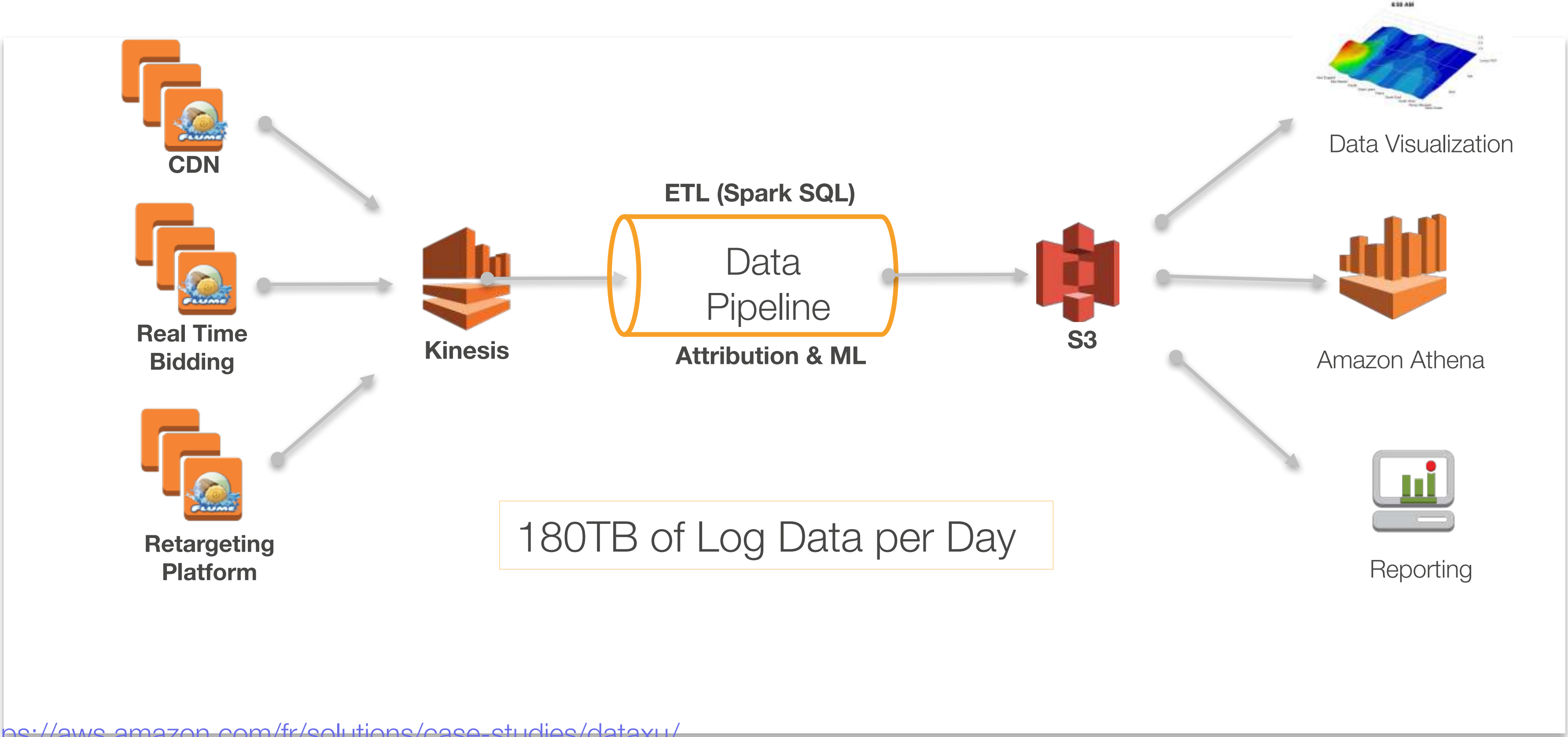
- Service based on Presto
- Table creation: Apache Hive Data Definition Language
- ANSI SQL operators and functions: what Presto supports, with a few exceptions

AWS re:Invent 2016: Introducing Athena (BDA303) https://www.youtube.com/watch?v=DxAuj_Ky5aw
https://prestodb.io

Data formats supported by Athena

- Unstructured
  - Apache logs, with customizable regular expression
- Semi-structured
  - delimiter-separated values (CSV, OpenCSV)
  - Tab-separated values (TSV)
  - JSON
- Structured
  - Apache Parquet https://parquet.apache.org/
  - Apache ORC https://orc.apache.org/
  - Apache Avro https://avro.apache.org/
- Compression (LZO, Snappy, Zlib, GZIP) & partitioning

Case Study – DataXu



180TB of Log Data per Day

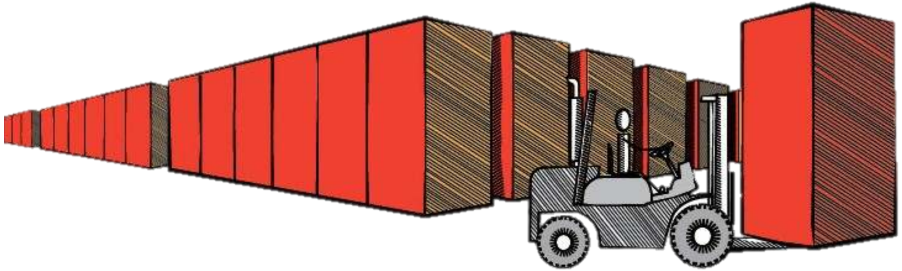
https://aws.amazon.com/fr/solutions/case-studies/dataxu/

Amazon Athena demo

Java SDK: https://docs.aws.amazon.com/athena/latest/ug/code-samples.html
JDBC: https://docs.aws.amazon.com/athena/latest/ug/connect-with-jdbc.html

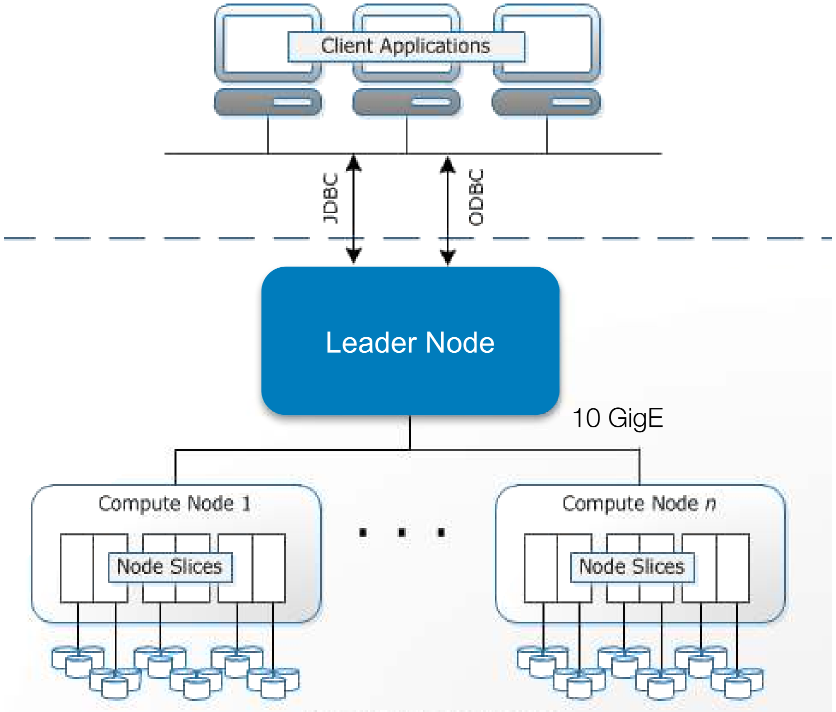
Amazon Redshift AWS Free Tier

- Relational data warehouse
- SQL is all you need to know
- Fully managed
- Massively parallel
- Petabyte scale
- As low as $1000/TB/year
- Athena-like capabilities with Redshift Spectrum



https://aws.amazon.com/redshift
http://www.allthingsdistributed.com/2012/11/amazon-redshift.html
Intro to Amazon Redshift Spectrum https://www.youtube.com/watch?v=gchd2sDhSuY

Amazon Redshift Architecture



Ingestion
Backup
Restore

S3 / EMR / DynamoDB

Spectrum

Query external
tables stored
in S3

What customers says about Amazon Redshift

“Redshift is twenty times faster than Hive” (5x – 20x reduction in query times) link

…[Redshift] performance has blown away everyone here (we generally see 50-100x speedup over Hive). link

We regularly process multibillion row datasets and we do that in a matter of hours. link

“Queries that used to take hours came back in seconds. Our analysts are orders of magnitude more productive.” (20x – 40x reduction in query times) link

“Did I mention it's ridiculously fast? We'll be using it immediately to provide our analysts an alternative to Hadoop.”

“Team played with Redshift today and concluded it is ****** awesome. Un-indexed complex queries returning in < 10s.”

Amazon Redshift demo

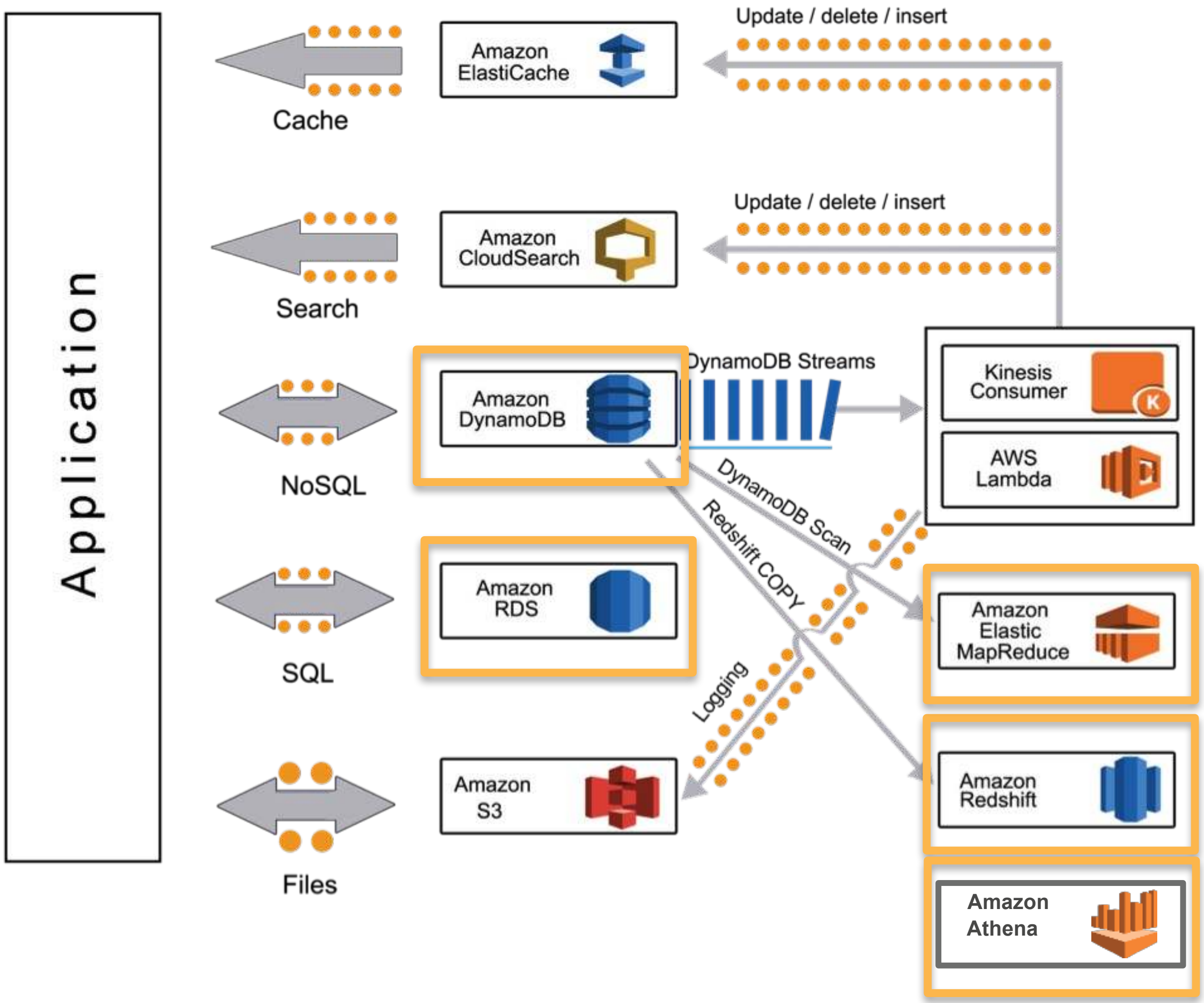
Java SDK: https://docs.aws.amazon.com/redshift/latest/mgmt/using-aws-sdk-for-java.html
JDBC driver: https://docs.aws.amazon.com/redshift/latest/mgmt/configure-jdbc-connection.html

	Amazon Redshift	Amazon Athena	Amazon EMR		
			Presto	Spark	Hive
Use case	Optimized for data warehousing	Ad-hoc Interactive Queries	Interactive Query	General purpose (iterative ML, RT, ..)	Batch
Scale/throughput	~Nodes Automatic (Spectrum)	Automatic / No limits	~ Nodes		
AWS Managed Service	Yes	Yes, Serverless	Yes		
Storage	Local storage Amazon S3 (Spectrum)	Amazon S3	Amazon S3, HDFS		
Optimization	Columnar storage, data compression, and zone maps	CSV, TSV, JSON, Parquet, ORC, Apache Web log	Framework dependent		
Metadata	Amazon Redshift managed	Athena Catalog Manager	Hive Meta-store		
BI tools supports	Yes (JDBC/ODBC)	Yes (JDBC)	Yes (JDBC/ODBC & Custom)		
Access controls	Users, groups, and access controls	AWS IAM	Integration with LDAP		
UDF support	Yes (Scalar)	No	Yes		

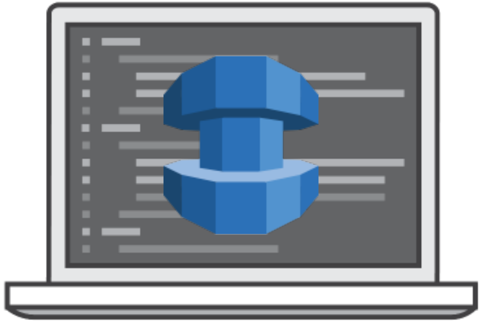
Fast Slow

Conclusion

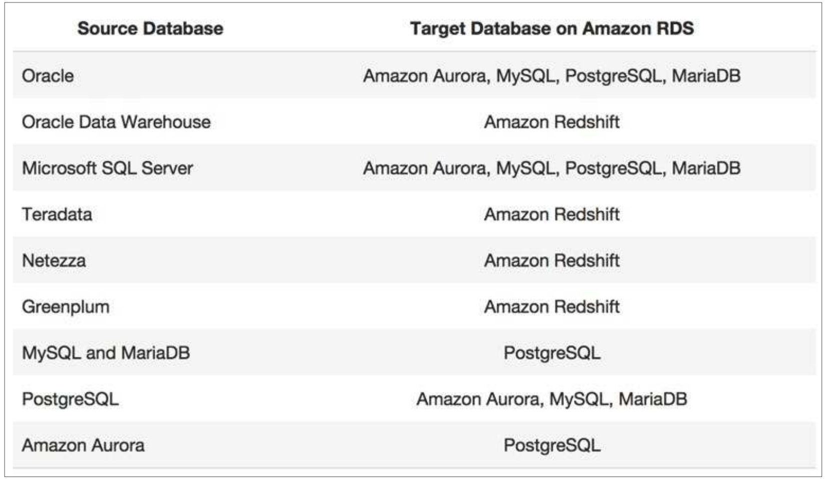
Reference architecture



AWS Schema Conversion Tool



- Converts your tables, views, stored procedures, and data manipulation language to RDS or Amazon Redshift
- Highlights where manual edits are needed



Source Database	Target Database on Amazon RDS
Oracle	Amazon Aurora, MySQL, PostgreSQL, MariaDB
Oracle Data Warehouse	Amazon Redshift
Microsoft SQL Server	Amazon Aurora, MySQL, PostgreSQL, MariaDB
Teradata	Amazon Redshift
Netezza	Amazon Redshift
Greenplum	Amazon Redshift
MySQL and MariaDB	PostgreSQL
PostgreSQL	Amazon Aurora, MySQL, MariaDB
Amazon Aurora	PostgreSQL

https://aws.amazon.com/dms/

AWS Database Migration Service

PostgreSQL Amazon Aurora mongoDB ORACLE

MySQL MariaDB SAP ASE Microsoft SQL Server

- ✓ Move data to the same or different database engine
- ✓ Move data to Redshift, DynamoDB or S3
- ✓ Keep your apps running during the migration
- ✓ Start your first migration in 10 minutes or less
- ✓ Replicate within, to, or from Amazon EC2 or RDS

https://aws.amazon.com/dms/
http://docs.aws.amazon.com/dms/latest/userguide/CHAP_Introduction.Sources.html
http://docs.aws.amazon.com/dms/latest/userguide/CHAP_Introduction.Targets.html
https://aws.amazon.com/blogs/database/database-migration-what-do-you-need-to-know-before-you-start/

AWS is a rich and lively environment for Java platforms

Your choice of backends: relational, NoSQL, Big Data, analytics

The tools you need, with less or no infrastructure drama

Built-in high availability, scalability, security & compliance

Focus on creativity and productivity, not on plumbing

Thank you!

Julien Simon, Principal Technical Evangelist, AWS

julsimon@amazon.fr
@julsimon