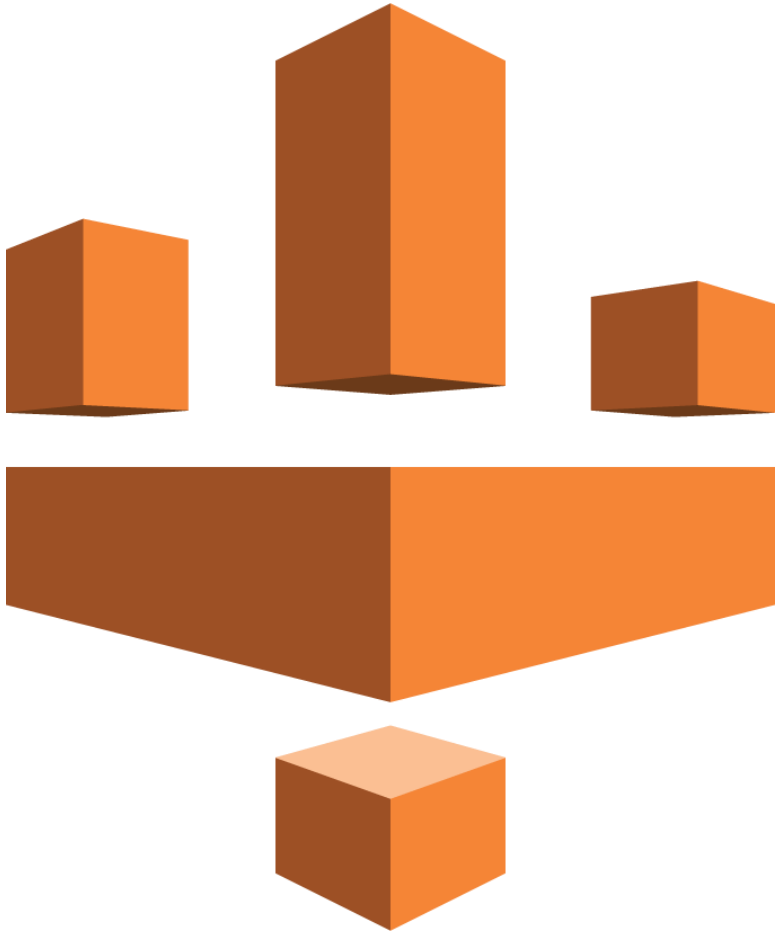


# aws Glue

Let's Get Stuck In!

## Gold Sponsors

dbWatch DATABASE CONTROL

quorum

DATA MOVEMENTS BRINGING DATA INTO FOCUS

adatis

## Silver Sponsors

COZYROC Founding Sponsor

Incremental Group

DLM Consultants

endjin work smarter

## Bronze Sponsors

mbn. people are inspiring

SQL RELAY

## Local Partners

Glasgow pdc Print+Design Centre

orb GROUP Branded Merchandise Solutions

EMILY CHAPPELL

# Chris Taylor

Jarrin Consultancy
SQL Server Specialists

- Worked with SQL Server since 2001
- MCSE – Data Platform
- SQLNE PASS ~~Chapter~~ Group Leader
- SQLRelay Organiser
- Cricket/Football Coaching


@SQLGeordie

github.com/SQLGeordie/

chris.taylor@jarrinconsultancy.com

www.jarrinconsultancy.com\blog
www.chrisjarrintaylor.co.uk

# Agenda

- Session Aim
- The Problem
- What is AWS Glue?
- Use Cases
- Demos
- Costs
- Q&A

# Not on the Agenda

- Comparison with other cloud offerings

# Session Aim

An understanding of the issues faced with ETL Development

Learn by example

Enough of a taste to get the Glue bug and start experimenting!

# The Problem

*"….consumes 70 percent of the resources needed for implementation and maintenance of a typical data warehouse"*

R. Kimball and J. Caserta. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Wiley, 2004.

# The Problem

70% of ETL Jobs are hand-coded
with no use of ETL Tools

https://www.slideshare.net/AmazonWebServices/bda311-introduction-to-aws-glue?qid=b3da6acd-c11b-4576-8f40-88906fb6c3f3&v=&b=&from_search=2

# Why hand-code?

- Flexible
- Powerful
- Unit test
- Deploy with other code
- You know your dev tools

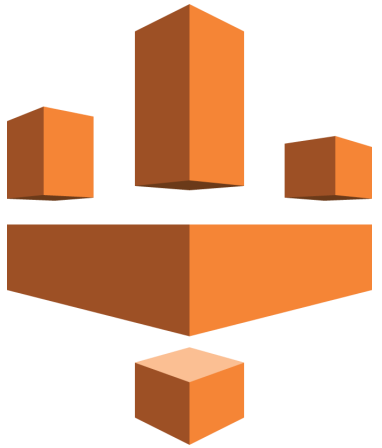
https://www.slideshare.net/AmazonWebServices/bda311-introduction-to-aws-glue?qid=b3da6acd-c11b-4576-8f40-88906fb6c3f3&v=&b=&from_search=2

# Involves a lot of effort

- Data formats change
- Source/target schemas change
- You add sources
- Data volume grows

https://www.slideshare.net/AmazonWebServices/bda311-introduction-to-aws-glue?qid=b3da6acd-c11b-4576-8f40-88906fb6c3f3&v=&b=&from_search=2

# What is AWS Glue?



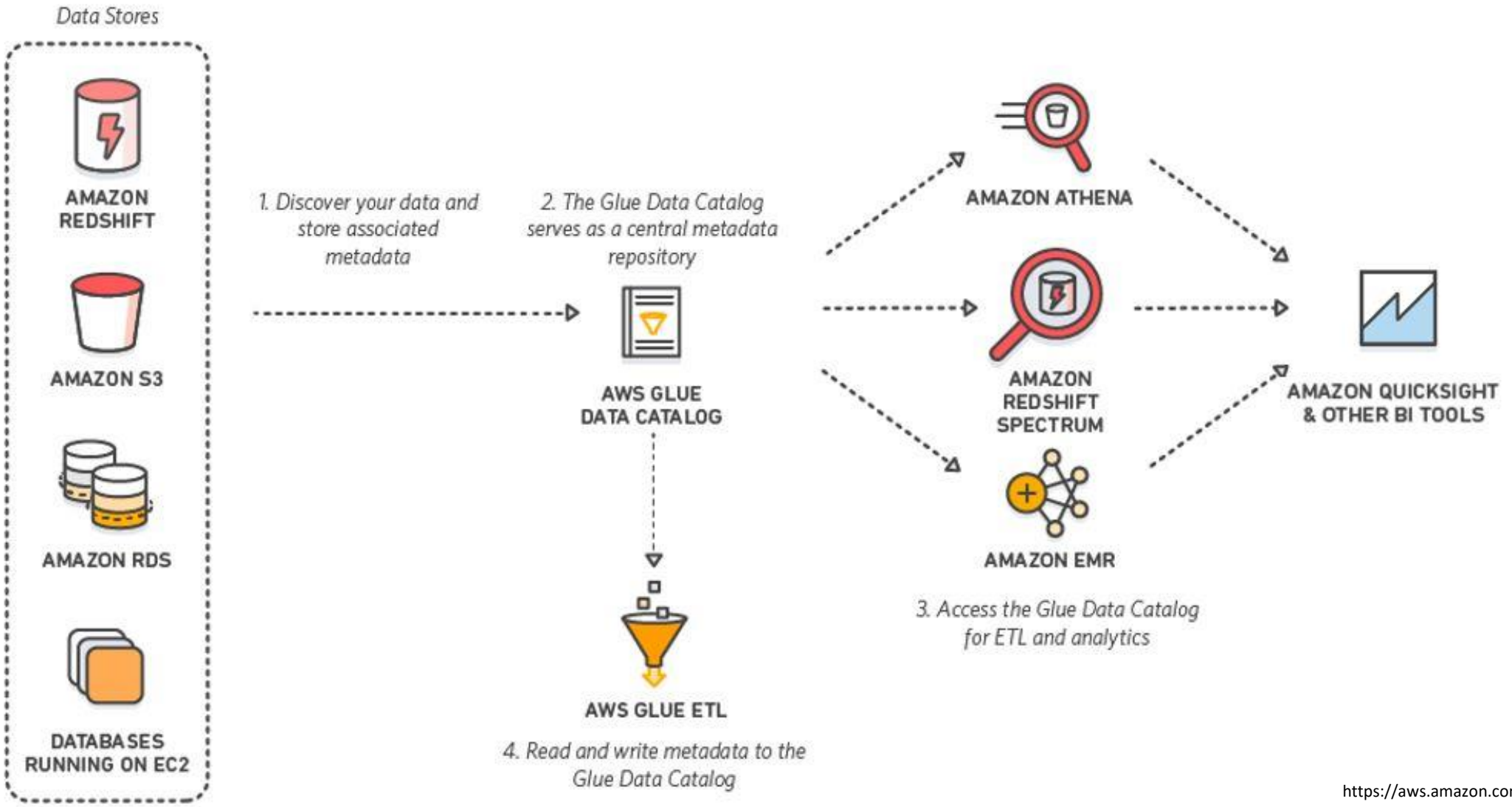
- Fully managed, ETL service
- Serverless
- Automates the undifferentiated heavy lifting of ETL
  - Discover, Develop, Deploy
- For Developers by Developers

# Components

- Data Catalog
  - Hive Metastore compatible
  - Crawlers automatically extracts metadata and creates tables
  - Integrated with Amazon Athena, Amazon Redshift Spectrum
- Job Authoring
  - Auto-generates ETL code
  - Build on open frameworks – Python and Spark
  - Developer-centric
- Job Execution
  - Run jobs on a serverless Spark platform
  - Provides flexible scheduling
  - Handles dependency resolution, monitoring and alerting

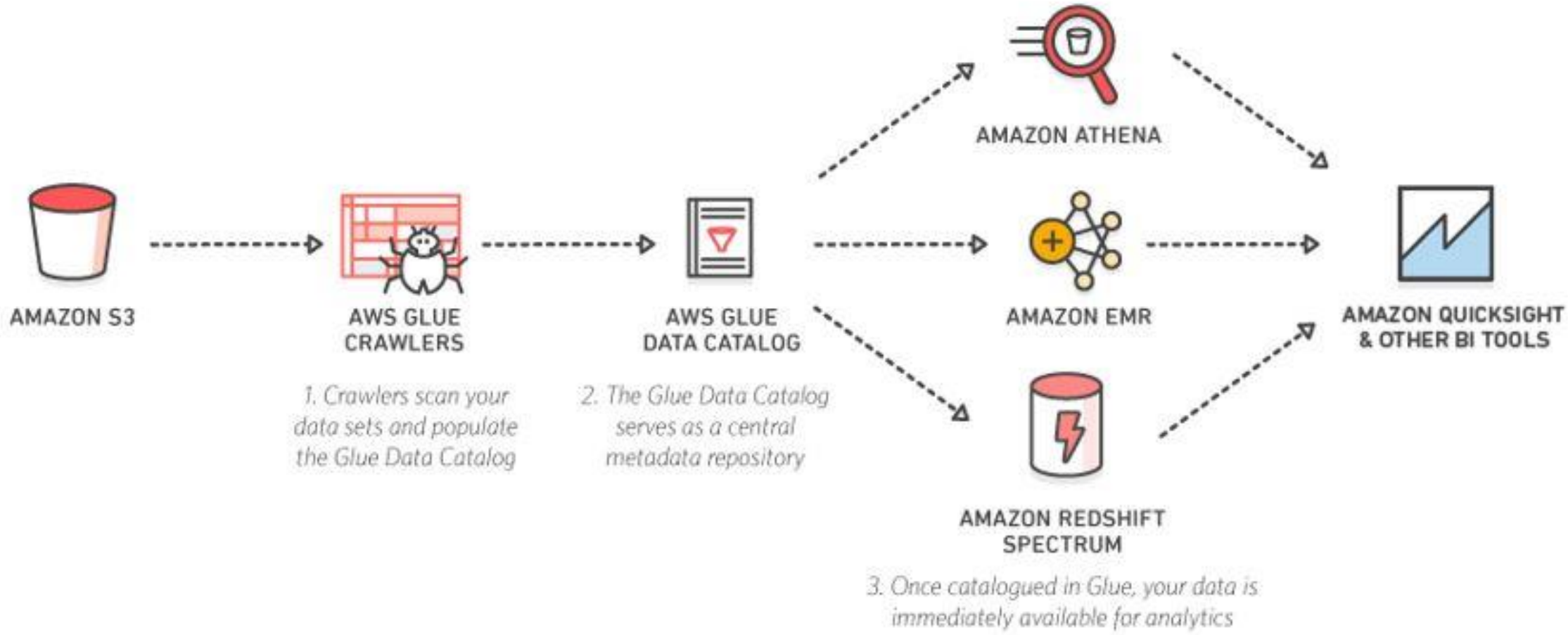
# Use Cases?

# Understand your data



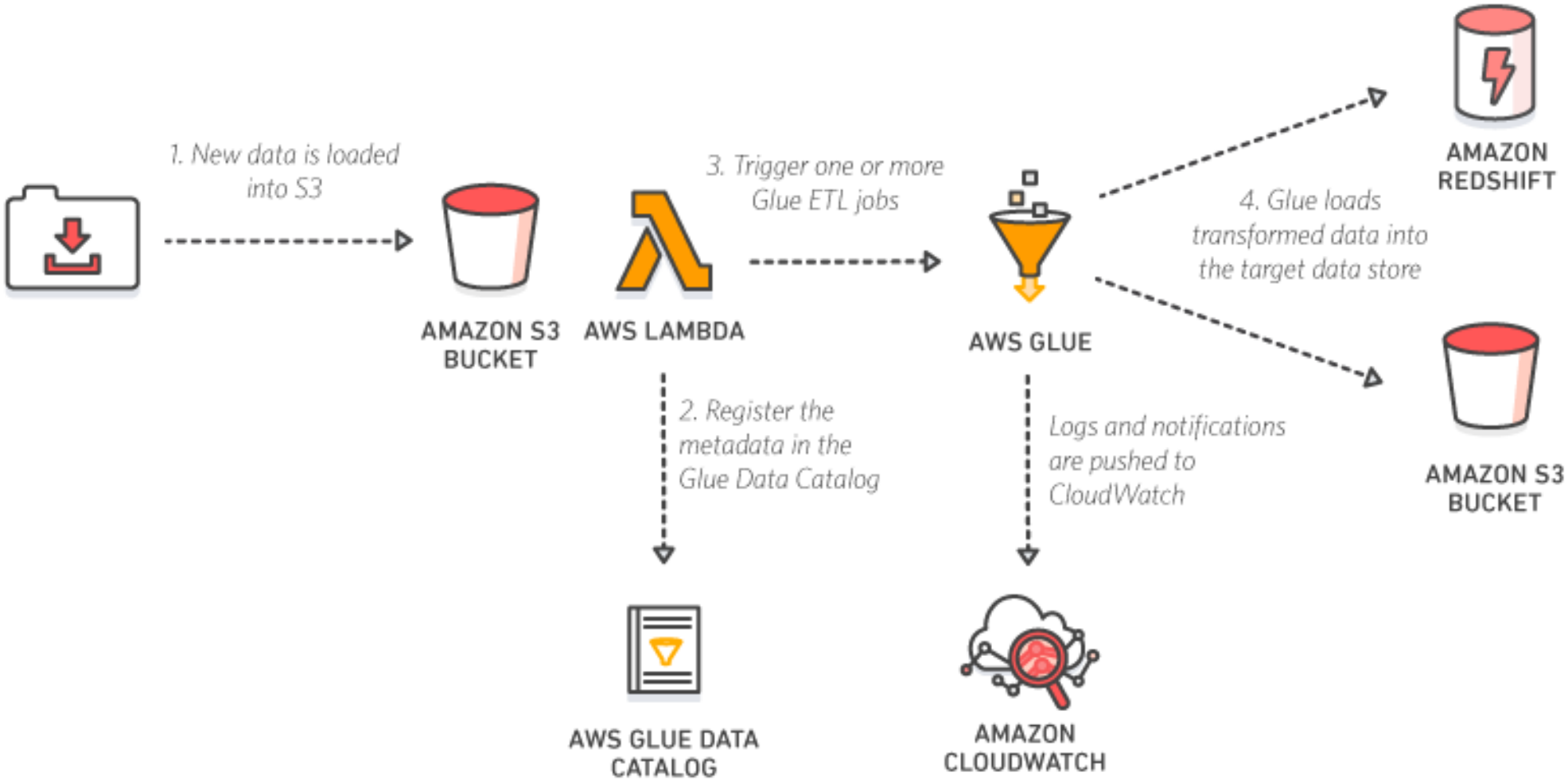

https://aws.amazon.com/glue/

# Query your data lake on Amazon S3




https://aws.amazon.com/glue/

# Build event driven ETL pipelines




https://aws.amazon.com/glue/

# DEMO

# Costs

- https://aws.amazon.com/free/
  - 1 Million objects stored in the AWS Glue Data Catalog**
  - 1 Million requests made per month to the AWS Glue Data Catalog**

** These free tier offers do not automatically expire at the end of your 12 month AWS Free Tier term, but are available to both existing and new AWS customers indefinitely

https://www.slideshare.net/AmazonWebServices/bda311-introduction-to-aws-glue?qid=b3da6acd-c11b-4576-8f40-88906fb6c3f3&v=&b=&from_search=2

# Costs

- DPU
- Compute based usage:
  - AWS Glue pricing ETL jobs, development endpoints, and crawlers $0.44 per DPU-Hour
  - 1 minute increments
  - 10-minute minimum ☹
  - A single DPU Unit = 4 vCPU and 16 GB of memory
- Data Catalog usage:
  - Data Catalog Storage:
    - Free for the first million objects stored $1 per 100,000 objects, per month, stored above 1M
  - Data Catalog Requests:
    - Free for the first million requests per month $1 per million requests above 1M

https://www.slideshare.net/AmazonWebServices/bda311-introduction-to-aws-glue?qid=b3da6acd-c11b-4576-8f40-88906fb6c3f3&v=&b=&from_search=2

# Costs Example #1

- ETL job
  - Ran for 10 minutes on a 6 DPU environment.
  - The price of 1 DPU-Hour in US East (N. Virginia) is $0.44.
  - The cost for this job run = 6 DPUs * (10/60) hour * $0.44 per DPU-Hour or $0.44.
- Development Endpoint
  - Active for 24 min.
  - Each development endpoint is provisioned with 5 DPUs
  - The cost to use the development endpoint = 5 DPUs * (24/ 60) hour * 0.44 per DPU-Hour or $0.88.

https://www.slideshare.net/AmazonWebServices/bda311-introduction-to-aws-glue?qid=b3da6acd-c11b-4576-8f40-88906fb6c3f3&v=&b=&from_search=2

# Costs Example #2

- Store 1 million tables in your Data Catalog in a given month and make 1 million requests to access these tables.
  - You pay $0 for using data catalog.
  - You are covered under the Data Catalog free tier.
- Your requests double to 2 million requests.
  - You will only be paying for one million requests above the free tier, which is $1
  - If you use crawlers to find new tables and they run for 30 min and use 2 DPUs. You will pay for 2 DPUs * (30/60) hour * $0.44 per DPU-Hour or $0.44. Your total monthly bill = $0 + $1 + $0.44 or $1.44

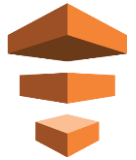
https://www.slideshare.net/AmazonWebServices/bda311-introduction-to-aws-glue?qid=b3da6acd-c11b-4576-8f40-88906fb6c3f3&v=&b=&from_search=2

# Why can’t I just use Data Pipeline?

## Glue

- Managed ETL service
- Discovering unstructured data
- Runs on a serverless Apache Spark environment.
- Takes a data first approach
- Provides an integrated data catalog / metadata
- Querying via Amazon Athena and Amazon Redshift Spectrum
- ETL jobs are Scala or Python based

## Data Pipeline



- Managed orchestration service
- Simple data replication tasks
- Greater flexibility (environment, access and compute resources)
- Launches compute resources in your account allowing you direct access to the Amazon EC2 instances or Amazon EMR clusters.
- Run on a different engine (Hive, Pig)

# Conclusion

## Good

- Fully Managed ETL
- Serverless
- Crawlers for discovering and relationalizing semi / unstructured data
- Developer Endpoints

## Not so good

- 10 minute minimum Job run
- Developer Endpoint £££££££
- AWS Documentation is lacking
- Multiple Files in folder (Athena)
- Complex non-scheduled automation
  - None for Crawlers!

# Summary

- Session Aim
- The Problem
- What is AWS Glue?
- Use Cases
- Demos
- Costs

# Questions?

# Contact

@SQLGeordie

github.com/SQLGeordie/

chris.taylor@jarrinconsultancy.com

www.jarrinconsultancy.com\blog
www.chrisjarrintaylor.co.uk

# Links / Info

- http://aws.amazon.com/documentation/glue
- https://www.slideshare.net/search/slideshow?searchfrom=header&q=aws+glue
- https://www.slideshare.net/AmazonWebServices/building-serverless-etl-pipelines-with-aws-glue-aws-summit-sydney-2018?qid=b3da6acd-c11b-4576-8f40-88906fb6c3f3&v=&b=&from_search=6
- https://www.slideshare.net/MichaelRainey3/going-serverless-an-introduction-to-aws-glue?qid=b3da6acd-c11b-4576-8f40-88906fb6c3f3&v=&b=&from_search=5
- https://aws.amazon.com/blogs/big-data/orchestrate-multiple-etl-jobs-using-aws-step-functions-and-aws-lambda/
- https://gluent.com/access-catalog-query-enterprise-data-gluent-cloud-sync-aws-glue/

# Geospatial in Quicksight

***Important***

*Geospatial charts in Amazon QuickSight currently aren't supported in some geographies, including India and China. We are working on adding support for more regions.*

*For now, automatic geocoding works only for **US locations**. However, you can add latitude and longitude coordinates to your data to make geospatial charts. For help with geospatial issues, see Geospatial Troubleshooting.*

https://docs.aws.amazon.com/quicksight/latest/user/geospatial-charts.html

# Best Practices and Questions

- https://docs.aws.amazon.com/athena/latest/ug/glue-best-practices.html
- https://aws.amazon.com/glue/faqs/
- https://www.accenture.com/us-en/blogs/blogs-kalyani-sayyed-amazon-glue-etl