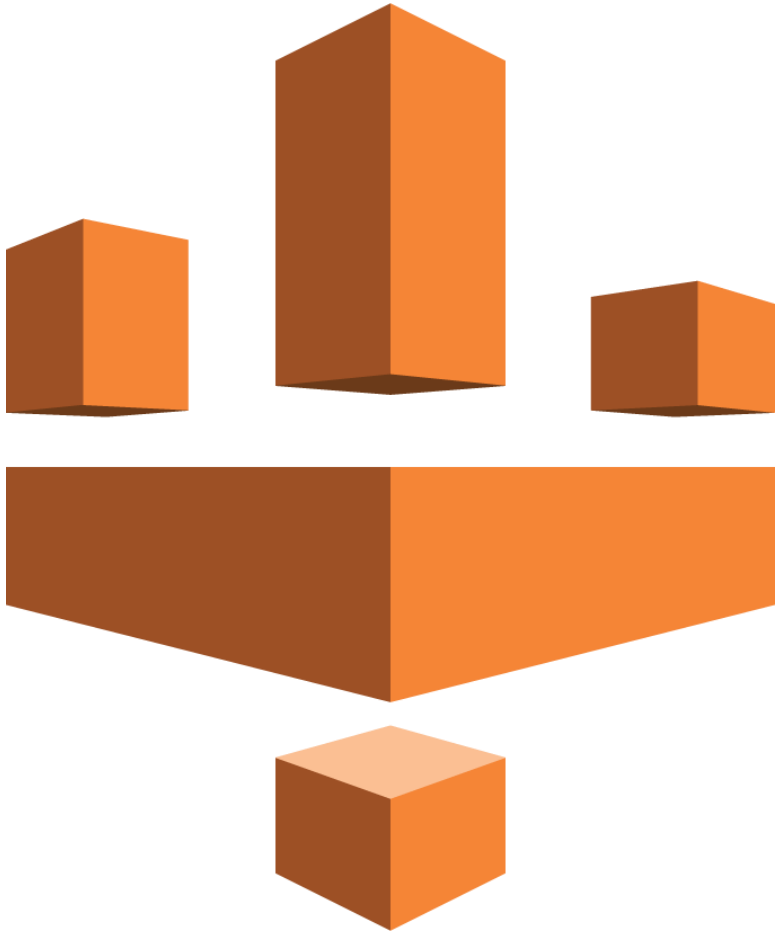


aws Glue

Let's Get Stuck In!

Chris Taylor

Jarrin Consultancy
SQL Server Specialists

- Worked with SQL Server since 2001
- MCSE – Data Platform
- Exceptional DBA Award finalist
  - *Damn that Jeff Moden with his RBAR and Tally tables ☺*
- SQLNE PASS ~~Chapter~~ Group Leader
- SQLRelay Organiser
- Formerly one of those “dirty devs”

@SQLGeordie

github.com/SQLGeordie/

chris.taylor@jarrinconsultancy.com

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

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

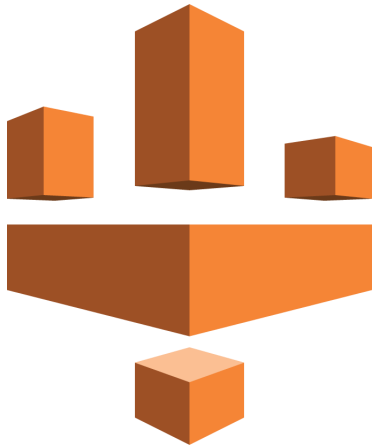
https://www.slideshare.net/AmazonWebServices/bda311-introduction-to-aws-glue?qid=b3da6acd-c11b-4576-8f40-88906fb6c3f3&v=&b=&from_search=2

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

DEMO

Conclusion

Good

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

Not so good

- Complex costing
- 10 minute minimum Job run
- Developer Endpoint Costs
- AWS Documentation is lacking
- Complex non-scheduled automation
  - None for Crawlers!

Contact

@SQLGeordie

github.com/SQLGeordie/

chris.taylor@jarrinconsultancy.com

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

Questions?