(Big) Data Integration and Pipelines

Andrew Brust
FOUNDER & CEO, BLUE BADGE INSIGHTS

@andrewbrust www.bluebadgeinsights.com

Movement and Transformation: Glue

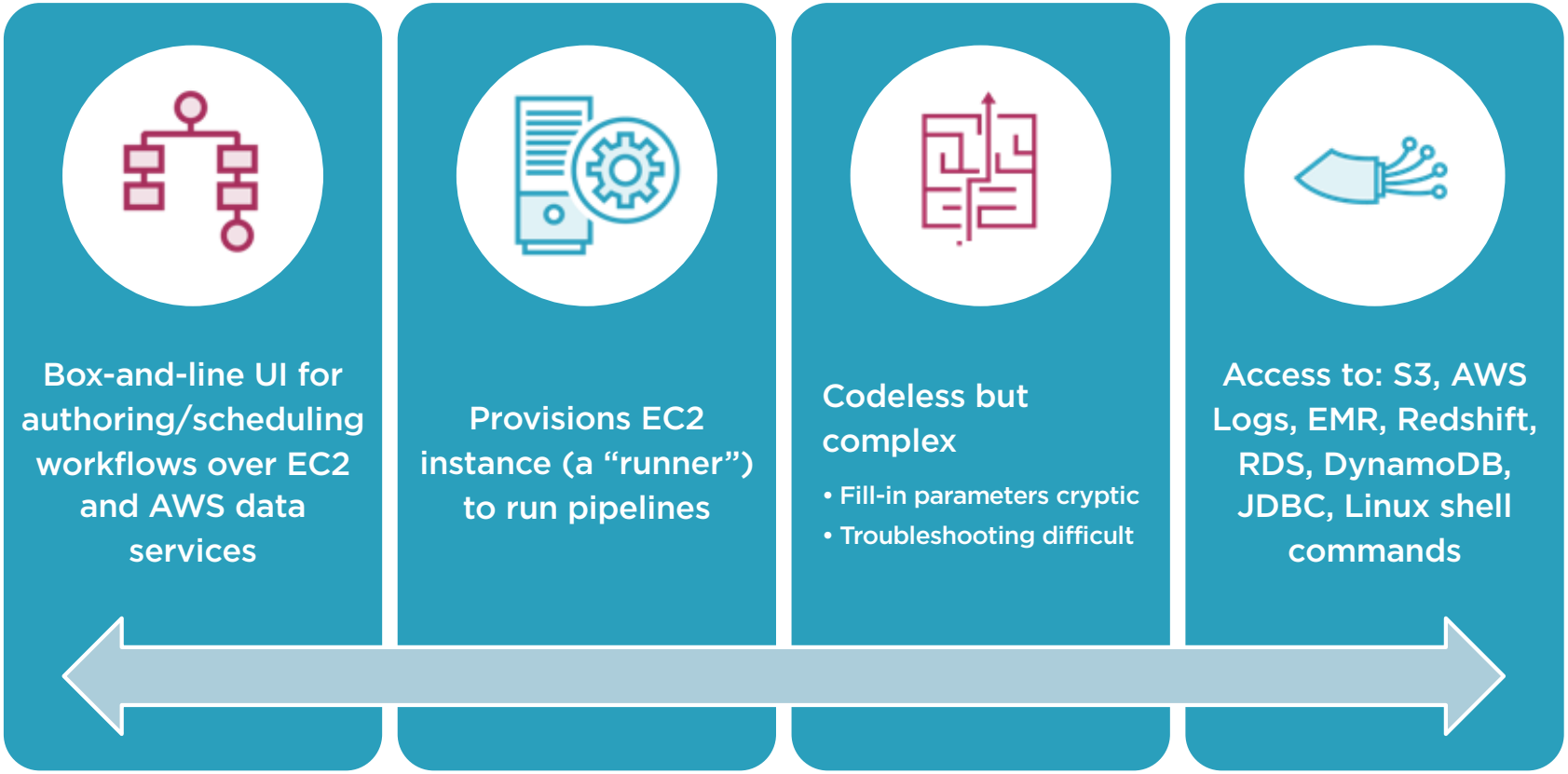
Visual interface that generates code

Runs on Spark, in serverless environment

Requires source, destination schemas must be in Glue Data Catalog

Access to: S3, DynamoDB, Redshift, RDS (including Aurora), JDBC

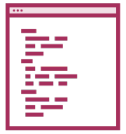
Movement and Transformation: Data Pipeline



Spark and Pig

Core Spark, Spark SQL all about data engineering

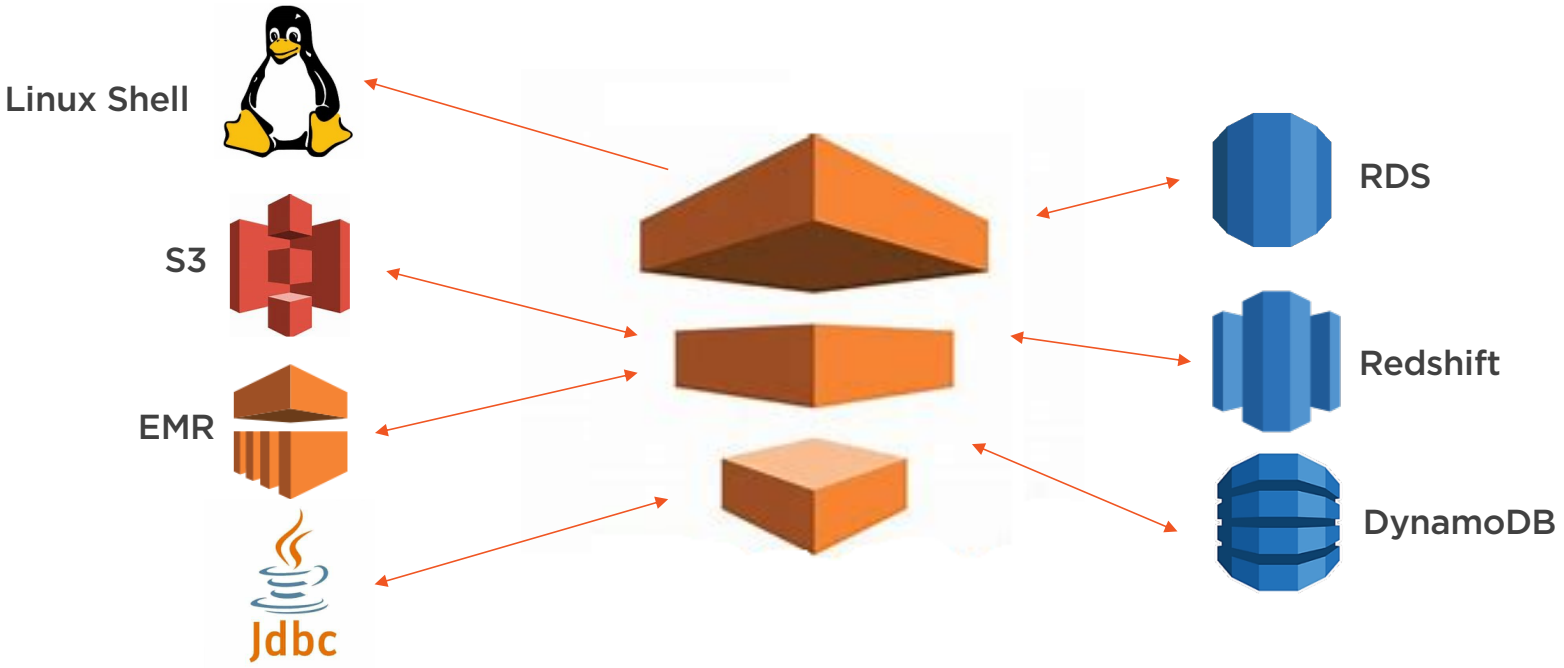
Apache Pig all about data transformation
(and can run atop MapReduce, Tez or Spark)



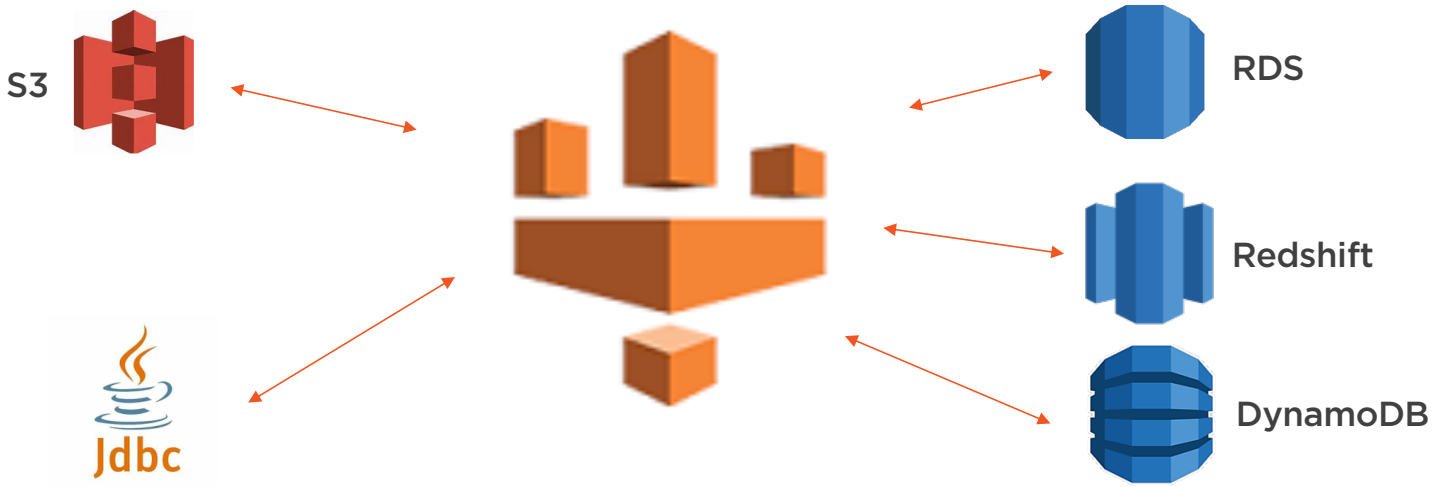
Lots of techies prefer hand-coding data pipelines to using graphical, declarative tools and platforms

Python, Scala, HiveQL or Pig Latin: all about scripts for ETL

Data Pipeline: Sources and Destinations



Glue: Sources and Destinations



Demo

Data Pipeline

- Moving data between S3 and Redshift