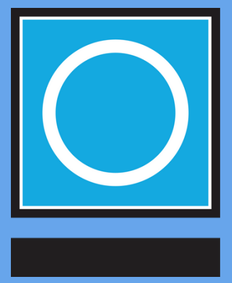
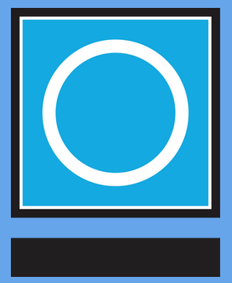
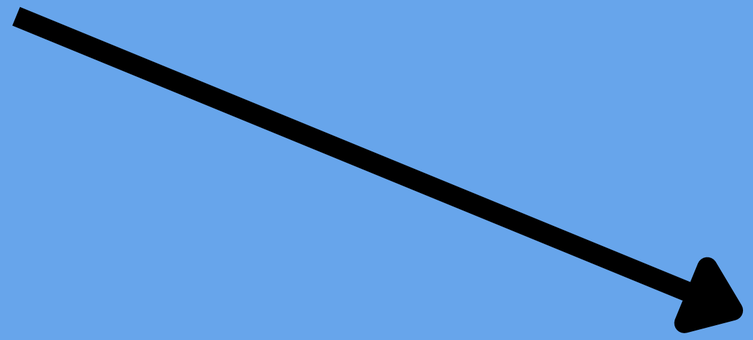


PYSPARK FOR AWS GLUE



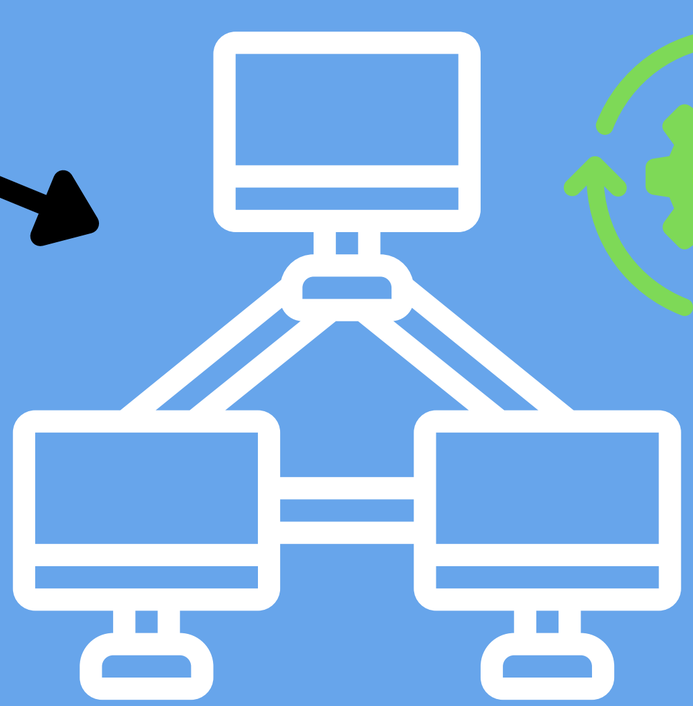
WHAT JUST HAPPENED?



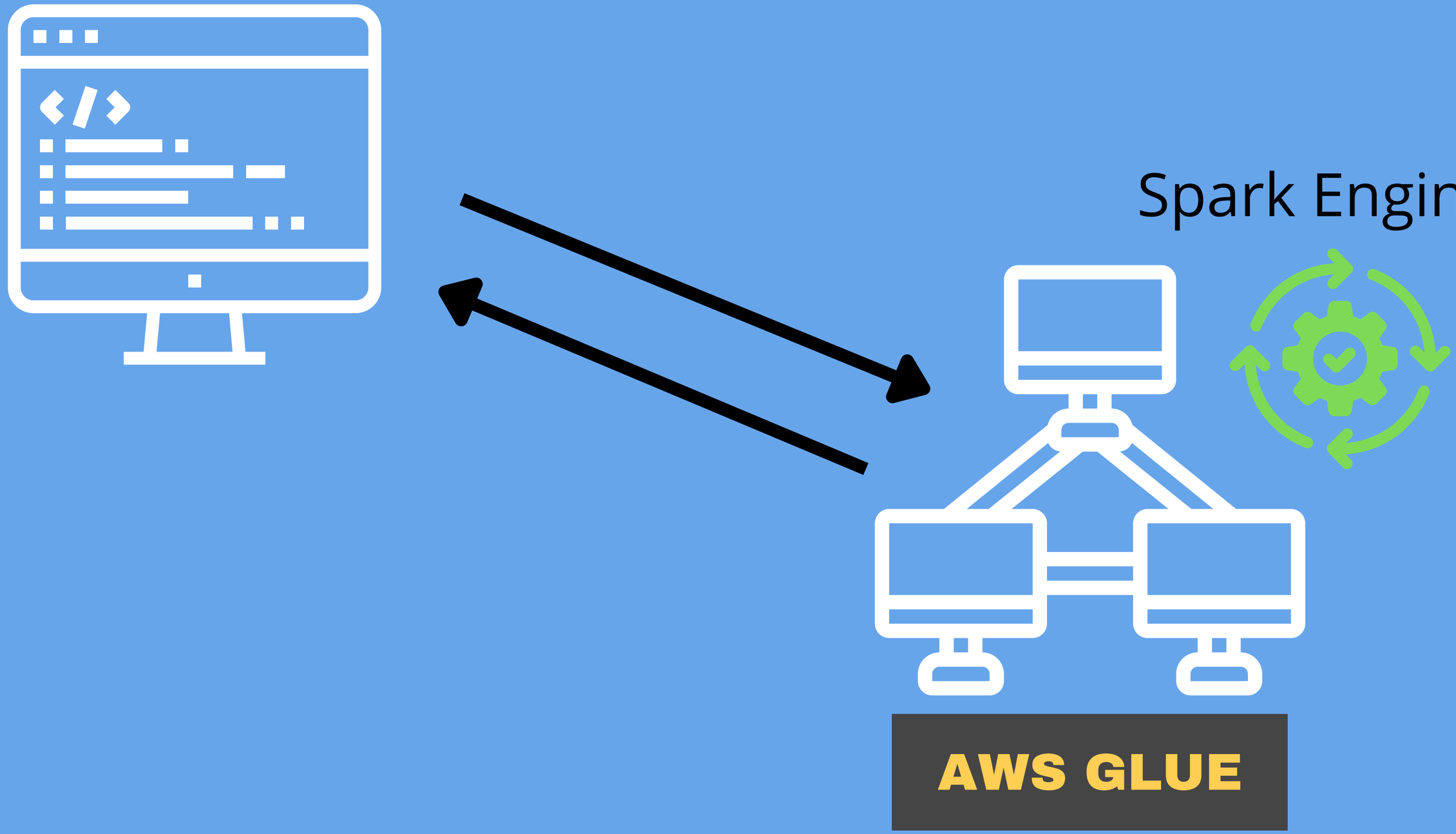
AWS GLUE


AWS GLUE

Spark Engine



AWS GLUE


Spark Engine
AWS GLUE

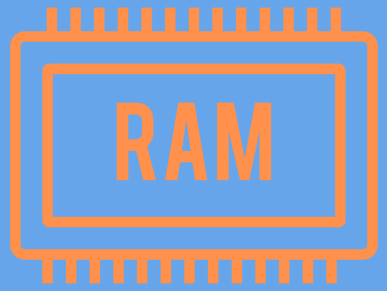
GLUE INTERACTIVE SESSIONS

- A programmatic and visual interface for building and testing extract, transform, and load (ETL) scripts for data preparation.
- Interactive sessions run Apache Spark analytics applications and provide on-demand access to a remote Spark runtime environment.
- AWS Glue transparently manages serverless Spark for these interactive sessions.

FUNDAMENTALS OF SPARK FOR GLUE

Apache Spark is an open-source in memory distributed processing system used for big data workloads

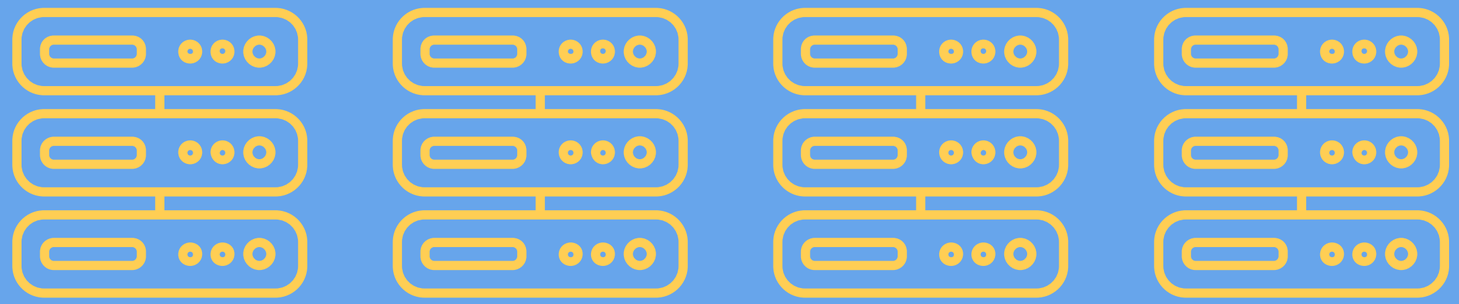
FUNDAMENTALS OF SPARK FOR GLUE



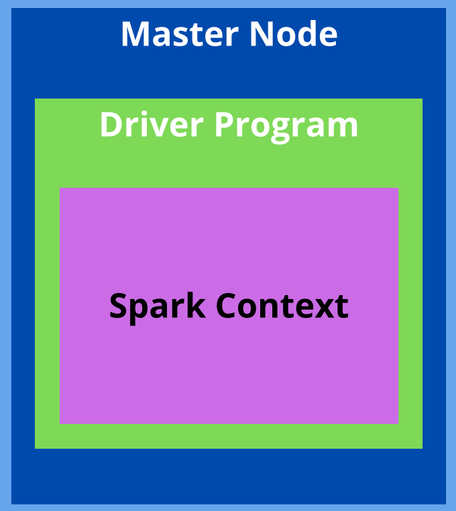
Apache Spark is an open-source **in memory** distributed processing system used for big data workloads

FUNDAMENTALS OF SPARK FOR GLUE

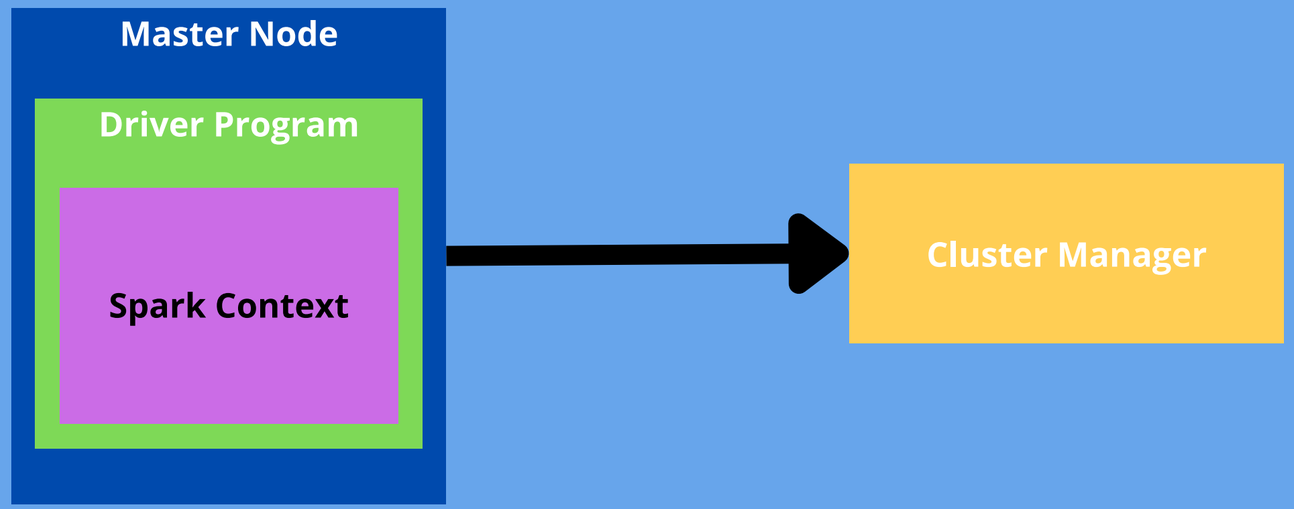
Apache Spark is an open-source in memory **distributed processing** system used for big data workloads



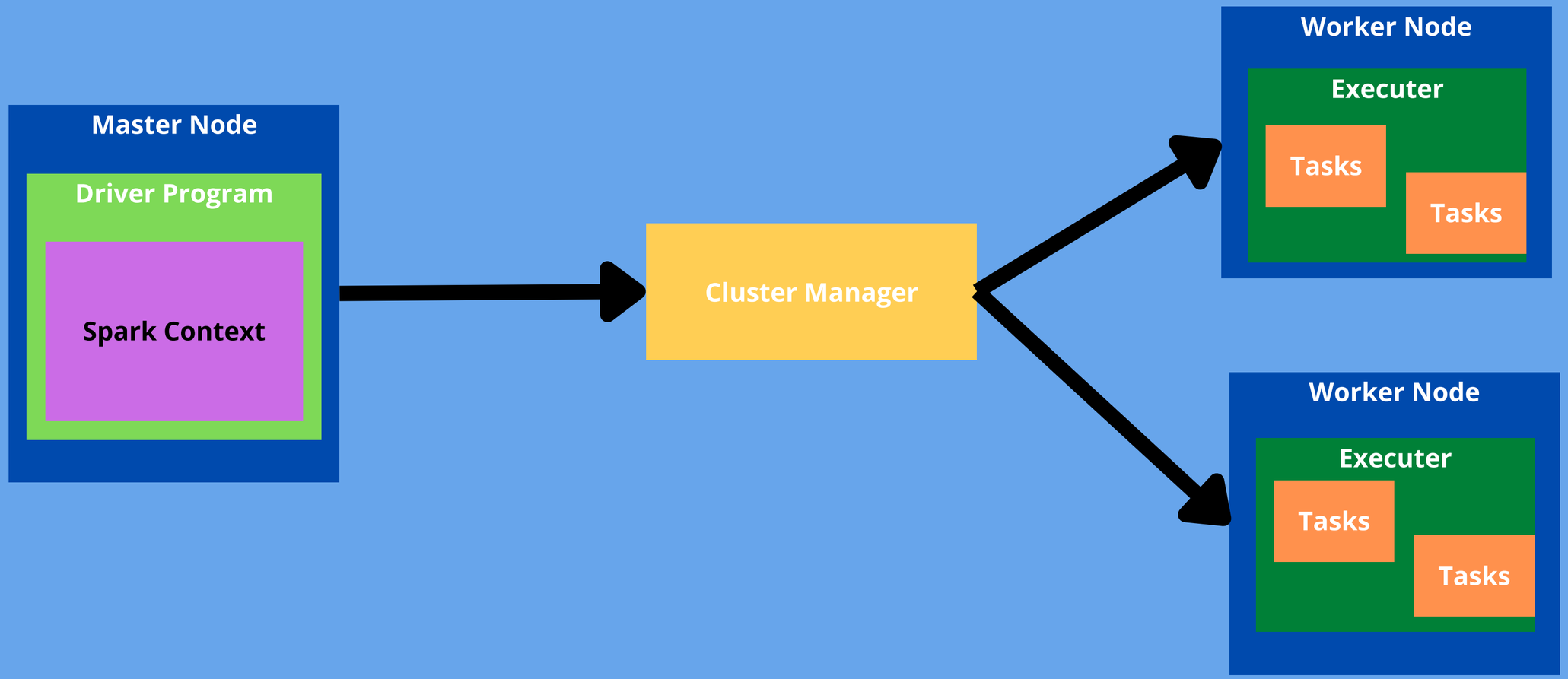
FUNDAMENTALS OF SPARK FOR GLUE



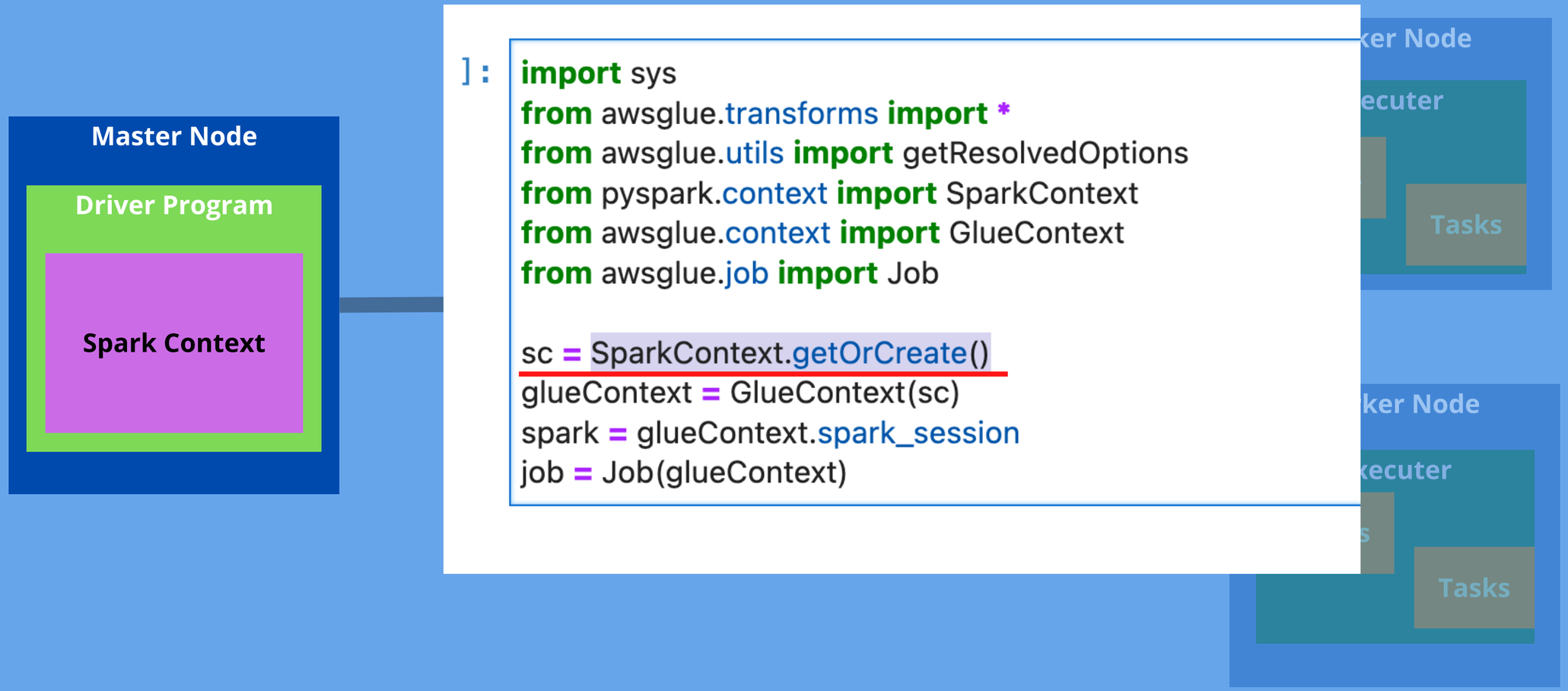
FUNDAMENTALS OF SPARK FOR GLUE



FUNDAMENTALS OF SPARK FOR GLUE



FUNDAMENTALS OF SPARK FOR GLUE

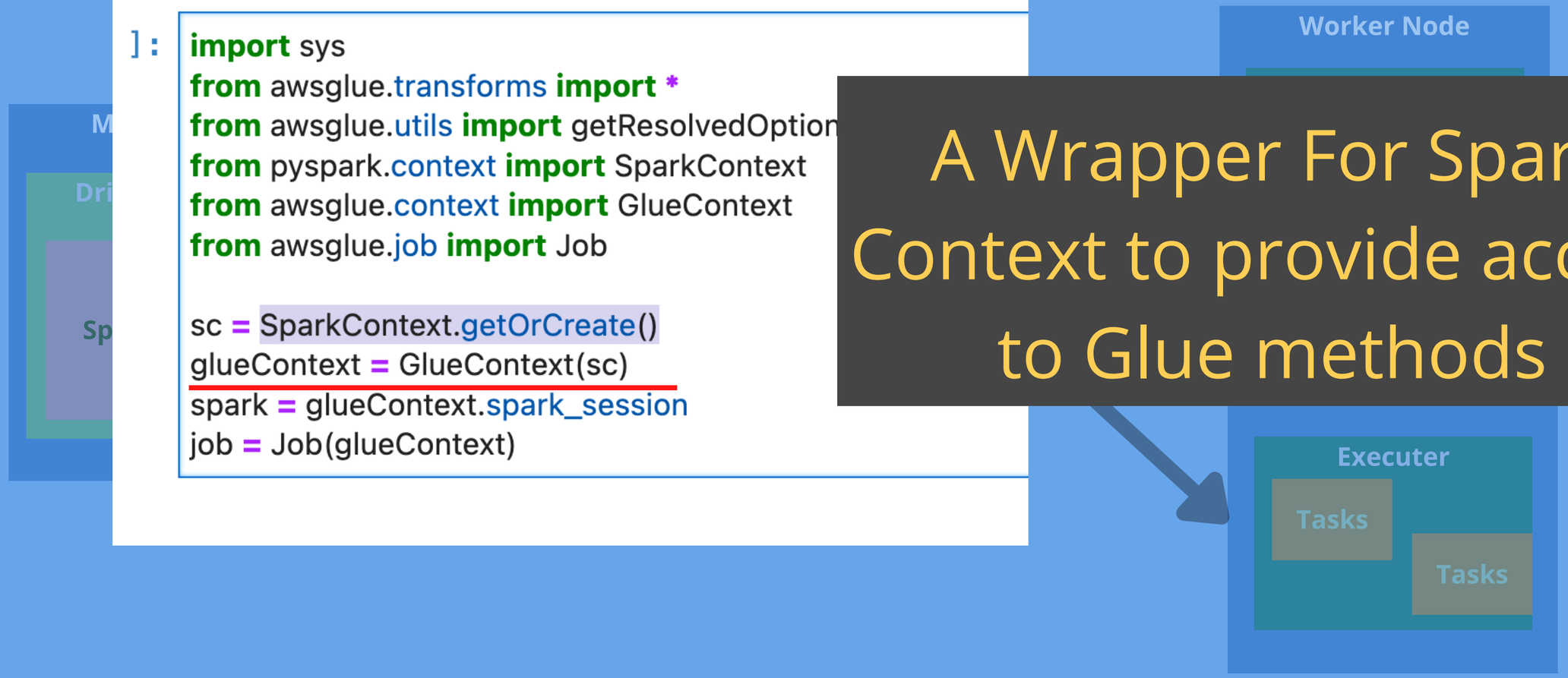


FUNDAMENTALS OF SPARK FOR GLUE

```
]:  import sys
    from awsglue.transforms import *
    from awsglue.utils import getResolvedOption
    from pyspark.context import SparkContext
    from awsglue.context import GlueContext
    from awsglue.job import Job

    sc = SparkContext.getOrCreate()
    glueContext = GlueContext(sc)
    spark = glueContext.spark_session
    job = Job(glueContext)
```

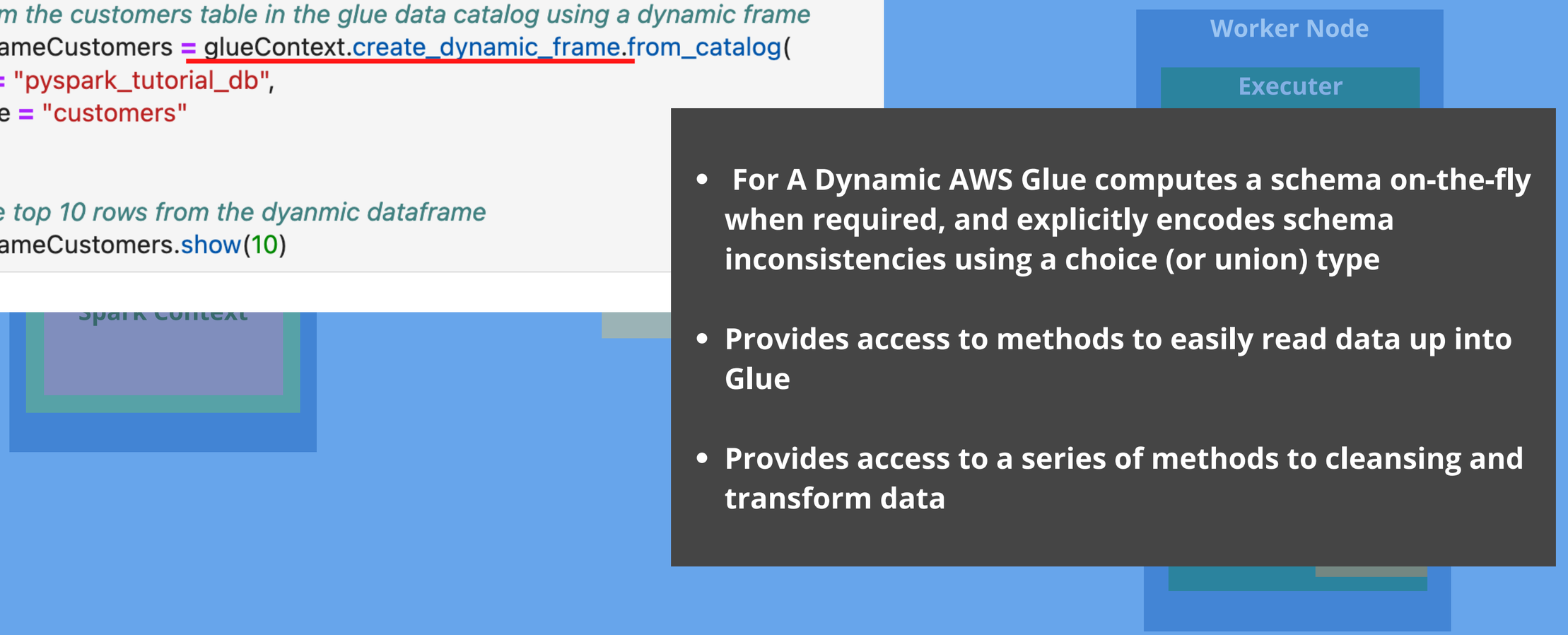
A Wrapper For Spark Context to provide access to Glue methods



GLUE DYNAMIC FRAME

```
# Read from the customers table in the glue data catalog using a dynamic frame
dynamicFrameCustomers = glueContext.create_dynamic_frame.from_catalog(
database = "pyspark_tutorial_db",
table_name = "customers"
)

# Show the top 10 rows from the dyanmic dataframe
dynamicFrameCustomers.show(10)
```



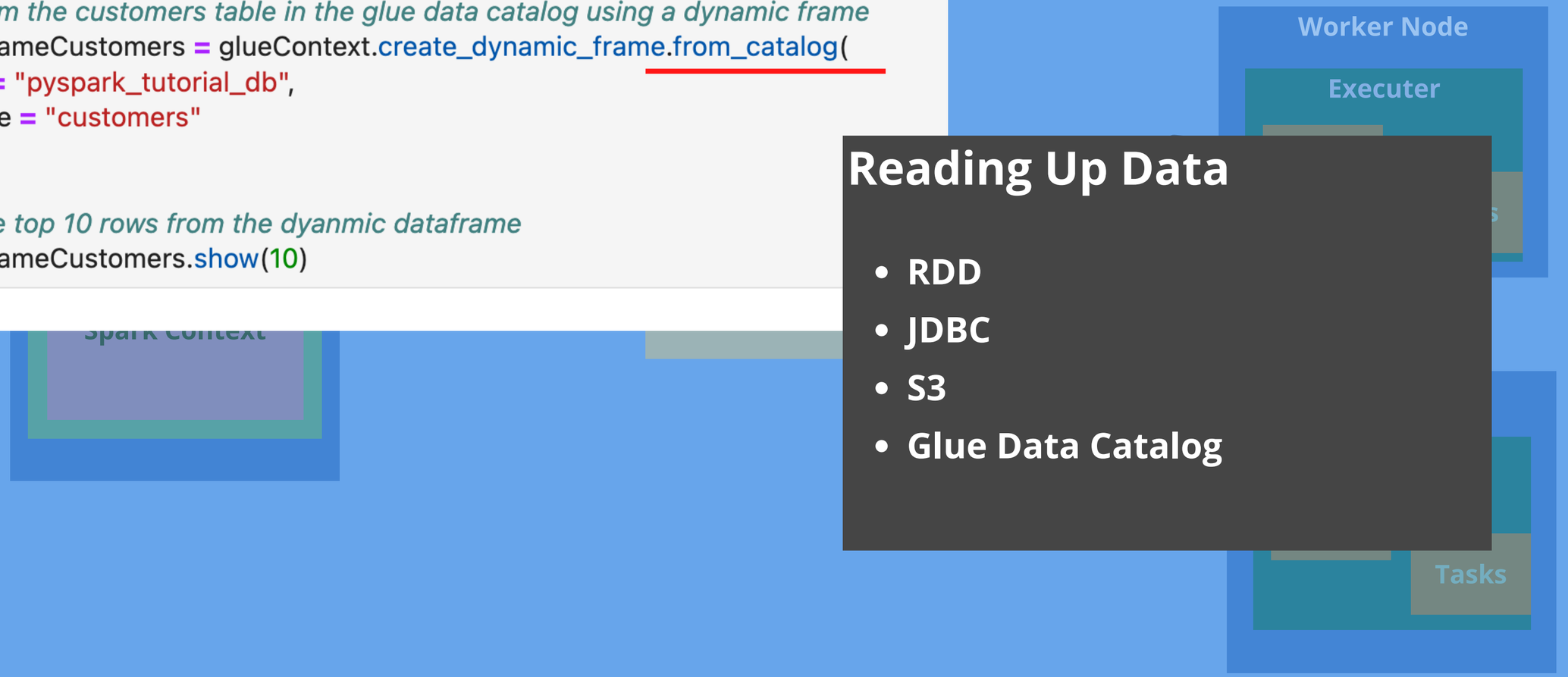
- For A Dynamic AWS Glue computes a schema on-the-fly when required, and explicitly encodes schema inconsistencies using a choice (or union) type
- Provides access to methods to easily read data up into Glue
- Provides access to a series of methods to cleansing and transform data

https://docs.aws.amazon.com/glue/latest/dg/aws-glue-api-crawler-pyspark-extensions-dynamic-frame.html

GLUE DYNAMIC FRAME

```
# Read from the customers table in the glue data catalog using a dynamic frame
dynamicFrameCustomers = glueContext.create_dynamic_frame.from_catalog(
database = "pyspark_tutorial_db",
table_name = "customers"
)

# Show the top 10 rows from the dyanmic dataframe
dynamicFrameCustomers.show(10)
```



Reading Up Data

- RDD
- JDBC
- S3
- Glue Data Catalog

https://docs.aws.amazon.com/glue/latest/dg/aws-glue-api-crawler-pyspark-extensions-dynamic-frame-reader.html

SPARK DATAFRAME

```
# Dynamic Frame to Spark DataFrame
sparkDf = dynamicFrameCustomers.toDF()

#show spark DF
sparkDf.show()
```

Spark Context

```
+----------+----------+-----------+--------------------+
|customerid| firstname|   lastname|            fullname|
+----------+----------+-----------+--------------------+
|       293| Catherine|       Abel|      Catherine Abel|
|       295|       Kim|Abercrombie|      Kim Abercrombie|
|       297|  Humberto|    Acevedo|    Humberto Acevedo|
|       291|   Gustavo|     Achong|      Gustavo Achong|
|       299|     Pilar|   Ackerman|      Pilar Ackerman|
|       305|     Carla|      Adams|         Carla Adams|
|       301|   Frances|      Adams|       Frances Adams|
|       307|       Jay|      Adams|          Jay Adams|
|       309|    Ronald|      Adina|        Ronald Adina|
|       311|    Samuel|   Agcaoili|     Samuel Agcaoili|
|       313|     James|    Aguilar|       James Aguilar|
|       315|    Robert|   Ahlering|     Robert Ahlering|
|       319|       Kim|      Akers|          Kim Akers|
|       441|   Stanley|       Alan|        Stanley Alan|
|       323|       Amy|    Alberts|         Amy Alberts|
|       325|      Anna|   Albright|       Anna Albright|
|       327|    Milton|     Albury|       Milton Albury|
|       329|      Paul|     Alcorn|         Paul Alcorn|
|       331|   Gregory|   Alderson|    Gregory Alderson|
|       333|J. Phillip|  Alexander|J. Phillip Alexander|
+----------+----------+-----------+--------------------+
only showing top 20 rows
```

https://docs.aws.amazon.com/glue/latest/dg/aws-glue-api-crawler-pyspark-extensions-dynamic-frame-reader.html