Supercharge Athena - With AWS Wrangler

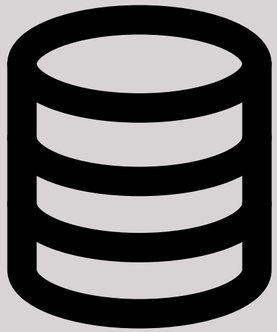
Querying Athena from AWS Wrangler

Made By

Kamalraj M M

What is a Database

LAYMAN DEFINITION



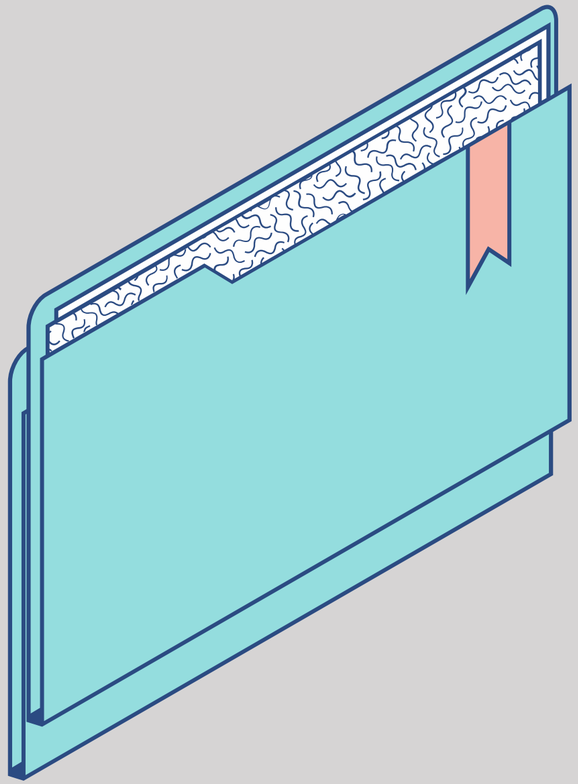
A General Purpose Database

Database is a server + application that stores data of any type and serves the stored data **when it is Queried**

Database Server has many database in it.

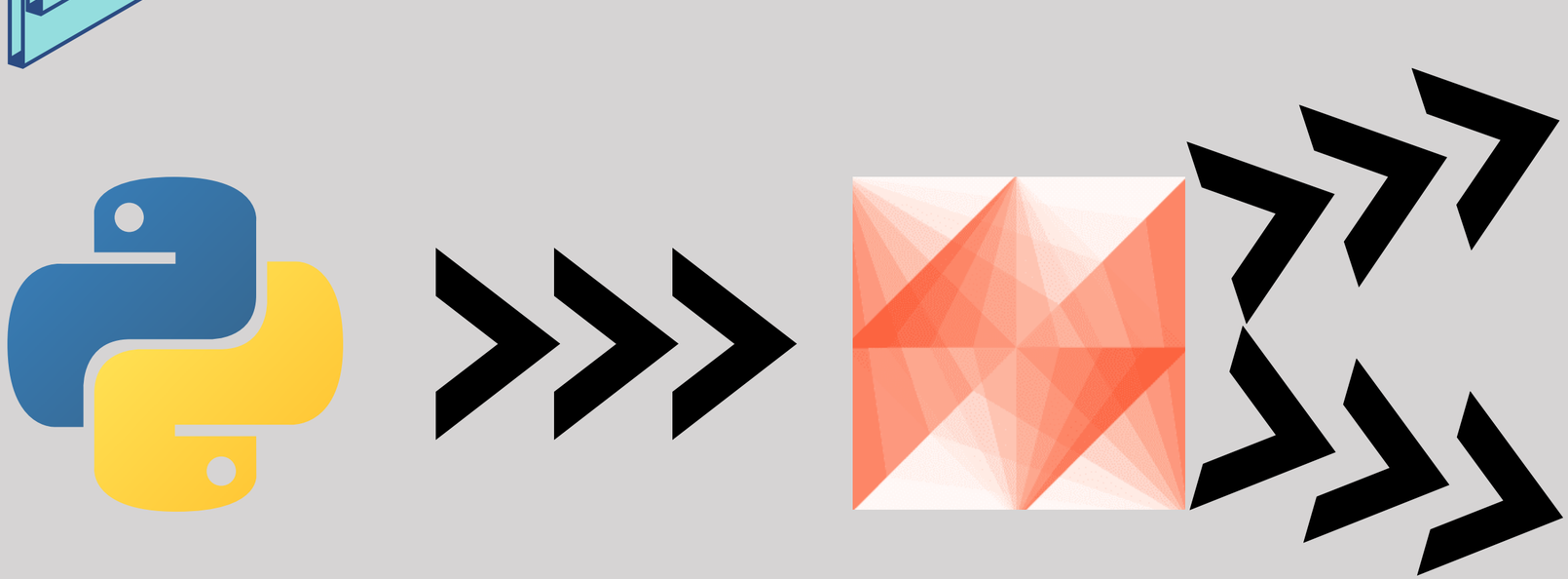
Each Database can have many tables inside them.

Each Tables stores data as Tuples



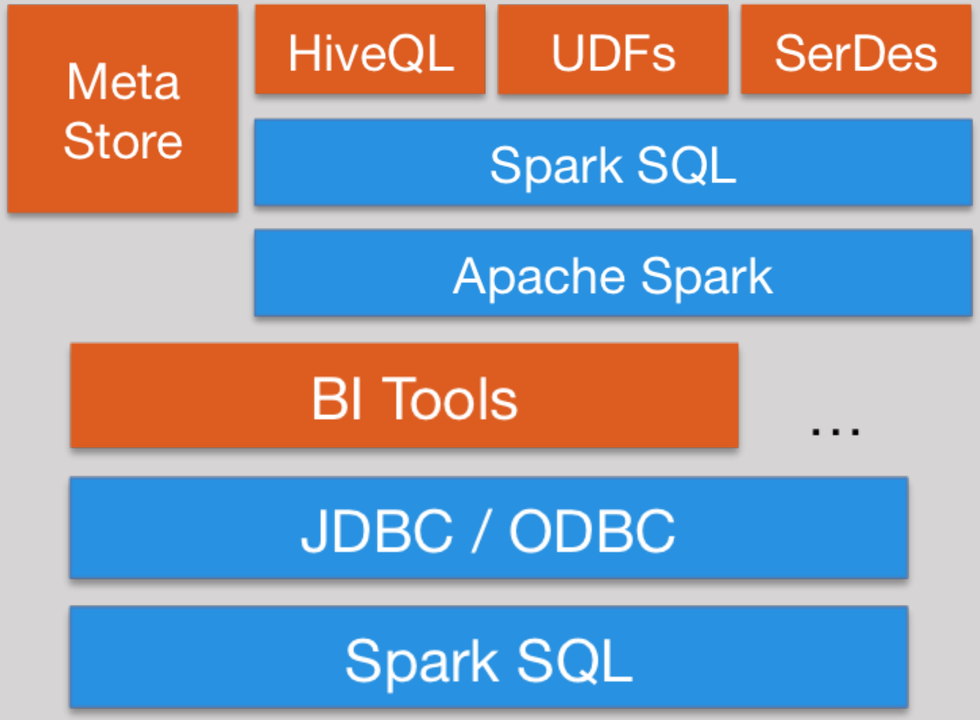
Pre-Requisites

- Tables created in the Glue Tutorial along with the data in S3 bucket.
- Good understanding of the SQL join



presto

AWS Athena is Based on



- **Hue & Presto :** The open source tools for querying the Hadoop file system data.
- Hue is **like SQLWorkbench for Big Data** Ecosystem. Presto helps in connecting different data sources to Meta Store
- **Hive / Spark / SparkSQL:** are other popular HDFS query frameworks in the ecosystem
- **Hive / Spark Metastore :** A regular postgres/ mysql data base server storing schema and data location information

Athena - Beacon of Control

PSEUDOCODE

Connecting and Querying Athena

1. What is Athena?
2. Querying with Athena using AWS Wrangler
3. Describing the Table create queries
4. Deleting Tables from Athena
5. Executing table joins from AWS Wrangler
6. Unloading the query results to the S3 bucket

Athena Methods of Interest

0 - Creating and connecting to AWS Session

wr.athena.describe_table(boto3_session=your_session)

wr.athena.list_query_executions(boto3_session=your_session)

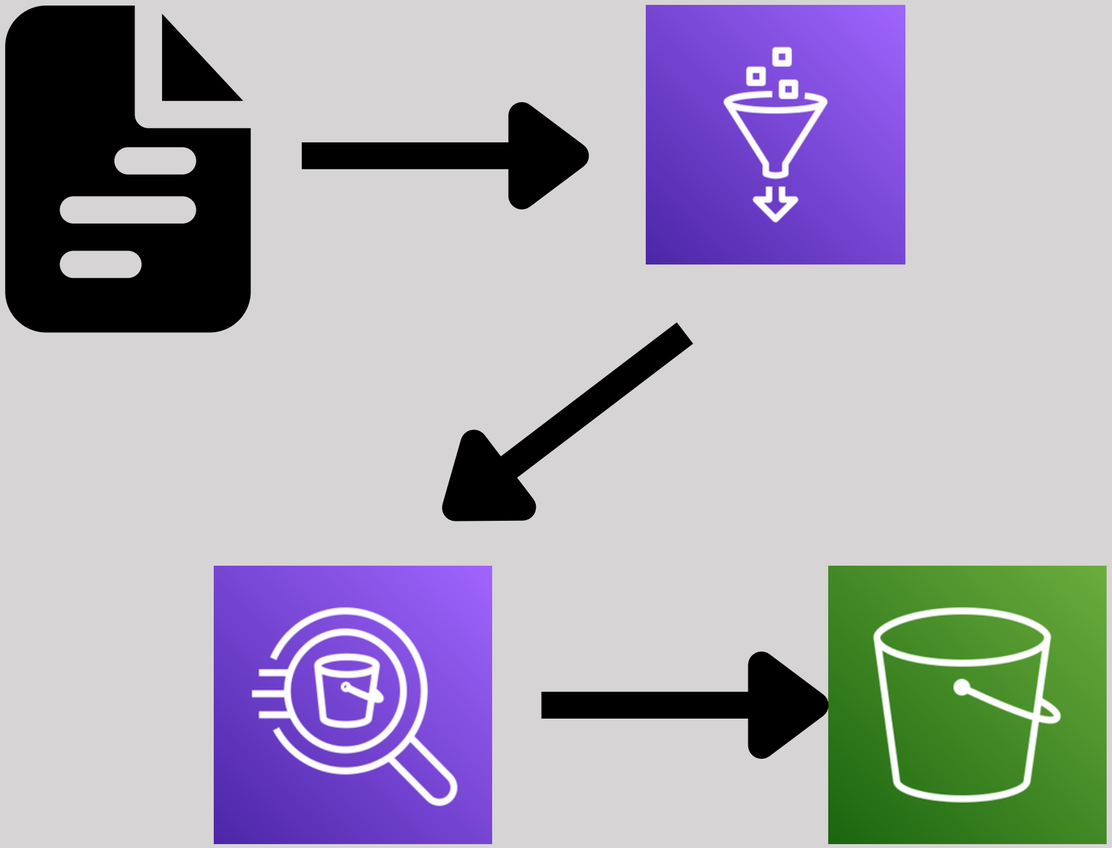
wr.athena.read_sql_query(sql='str',table='tbl_name',database ='database name, boto3_session=your_session)

wr.athena.show_create_table(table='name',boto3_session= your_session)

wr.athena.describe_table(table='name',boto3_session=your_s ession)

LIVE DEMO

SUPER CHARGING ATHENA



Jupyter Notebook + AWS Wrangler Demo

1. Write pandas dataframe to S3 buckets and Glue catalog.
2. Write SQL Query and execute them on Athena
3. Execute the same query on AWS Wrangler on Jupyter notebook
4. Write a SQL join on dataset and unload data to S3 bucket