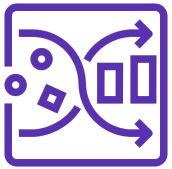
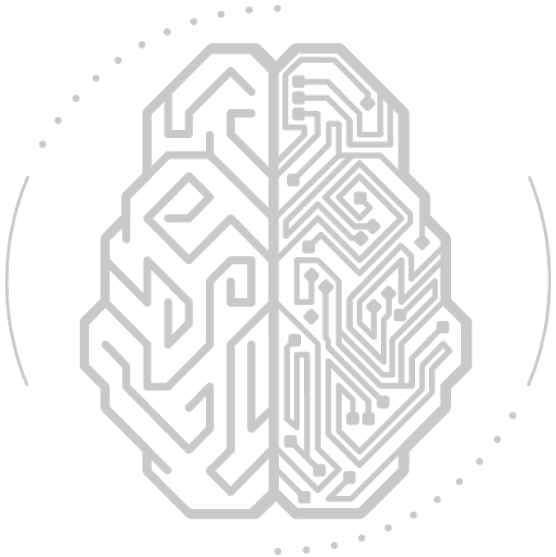


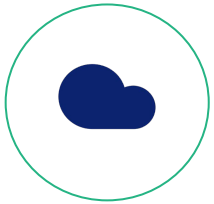
AWS DataBrew Introduction



Agenda

Introductions

Overview on DataBrew



Demo

Q&A

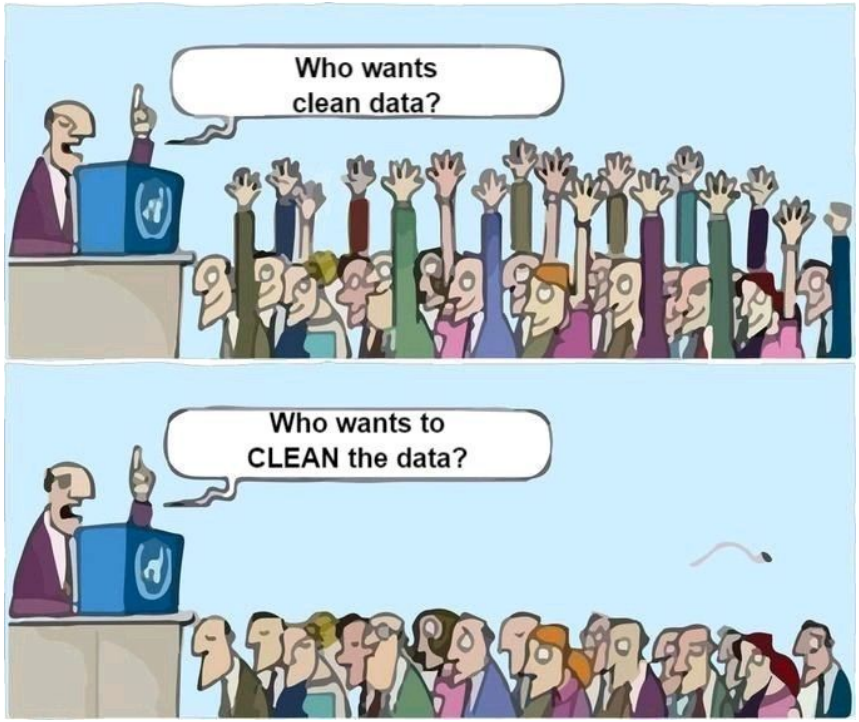
Speaker

Sanchit Jain

Lead Architect - AWS at Quantiphi
AWS APN Ambassador

Why we need DataBrew?

- We are all fascinated with various analytical things like fancy visualization, BI report, Machine learning output.
- But do we know, 50% of the time is just spent in cleansing, understanding & exploring your data?
- We all like the elegant outcome, but really no one wants to invest the time required to clean the data due to the combustion process, multiple rounds of back and forth, and time-consuming, etc.
- With this growing need of the industry, AWS launched the Glue DataBrew service in Nov 2020 with an idea to purely focus on business use-case rather than preparation work required for the same.



What is AWS Glue DataBrew ?

AWS Glue DataBrew is a visual data preparation tool that makes it easy for data analysts and data scientists to prepare data with an interactive, point-and-click visual interface without writing code.

CAPABILITIES OF GLUE DATABREW

- Profile
- Clean and Normalize
- Map Data Lineage
- Automate

NEED FOR DATABREW

"Upto 80% of data analysis time is spent on preparing data"

Time Consuming

- Multi-step process to extract, clean, normalize & load data at scale
- The right tools for the right persona must be integrated

Expensive

- Costly user licenses & siloed tools that cause rework
- Often requires moving large amount of data into silos

Manual

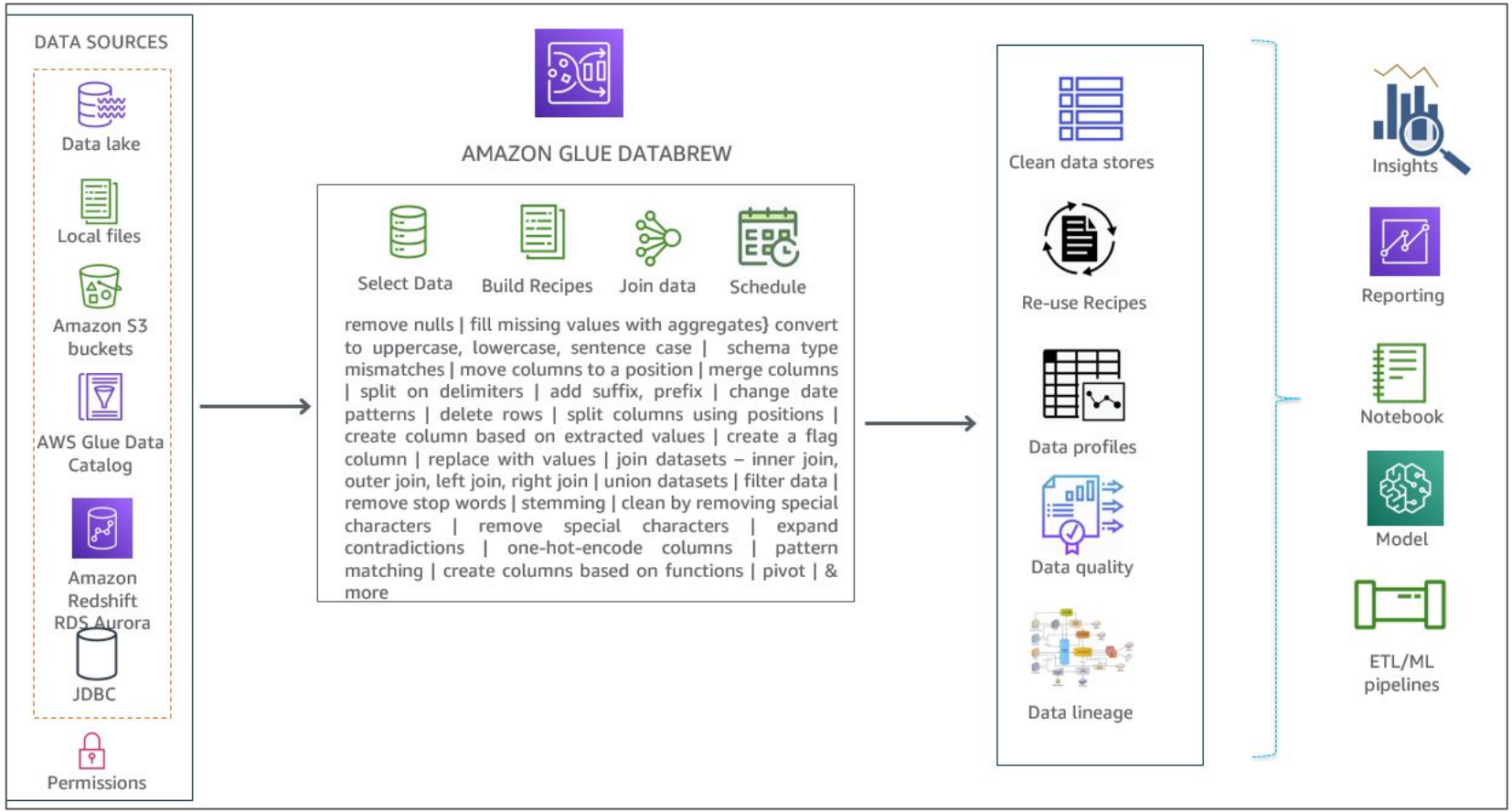
- Needs a lot of code-based heavy lifting to work at scale
- Hard to operationalize & build repeatable workflows

How does DataBrew Work ?

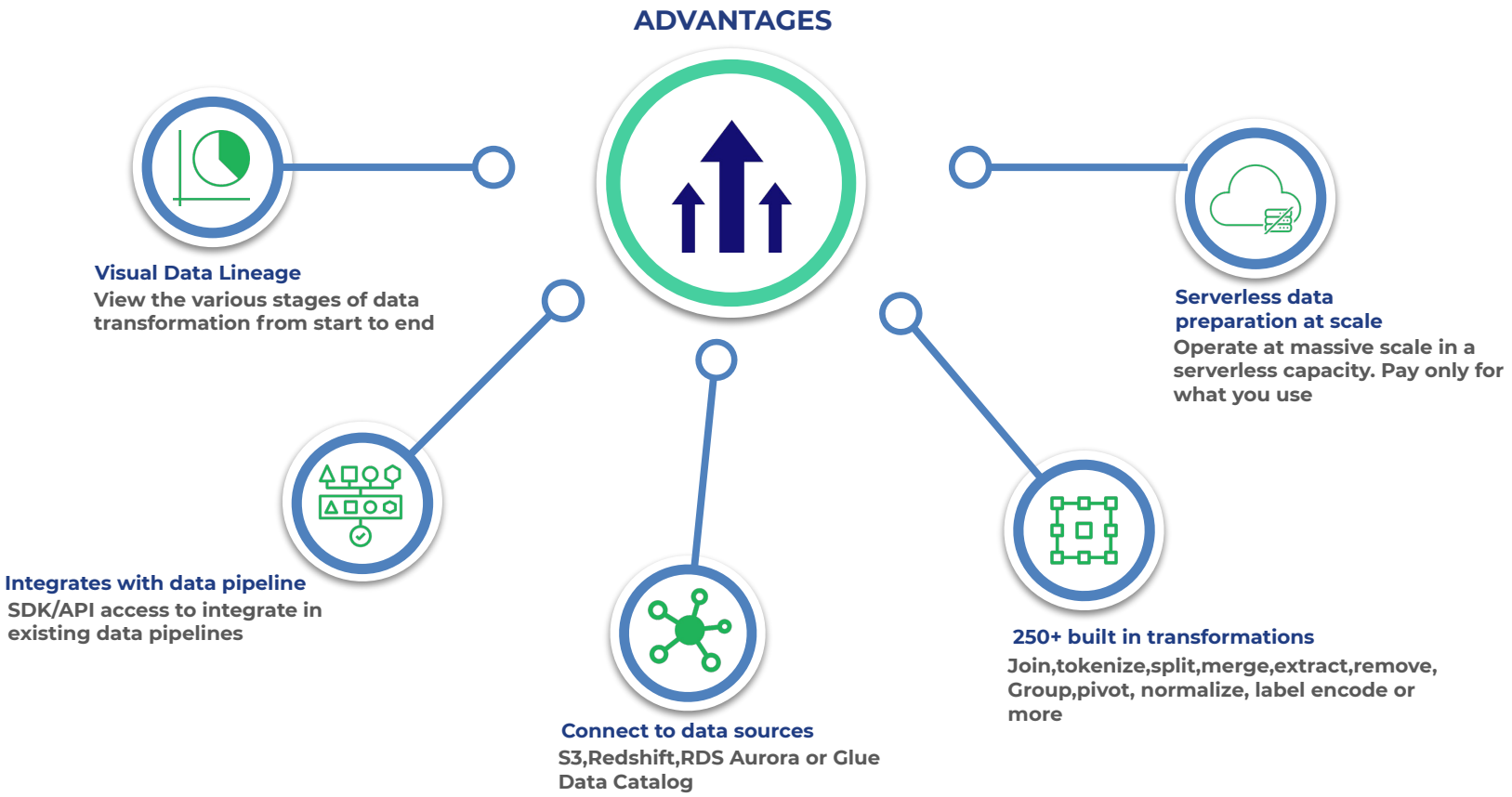
- You can choose from over 250 built-in transformations to combine, pivot, and transpose the data without writing code.
- AWS Glue DataBrew also automatically recommends transformations such as filtering anomalies, correcting invalid, incorrectly classified, or duplicate data, normalizing data to standard date and time values, or generating aggregates for analyses.
- For complex transformations, such as converting words to a common base or root word, DataBrew provides transformations that use advanced machine learning techniques such as Natural Language Processing (NLP).
- You can group multiple transformations together, save them as recipes, and apply the recipes directly to newly incoming data.

For input data, AWS Glue DataBrew supports commonly used file formats, such as comma-separated values (.csv), JSON and nested JSON, Apache Parquet and nested Apache Parquet, and Excel sheets. For output data, AWS Glue DataBrew supports comma-separated values (.csv), JSON, Apache Parquet, Apache Avro, Apache ORC and XML.

Overview of DataBrew



Advantages of DataBrew



Core Concepts

Dataset - a set of data—rows or records that are divided into columns or fields

Project - The interactive data preparation workspace in DataBrew

Recipe - is a set of instructions or steps for data that you want DataBrew to act on

Job - The process of running these instructions when you make the recipe is called a job.

Data Lineage - DataBrew tracks your data in a visual interface to determine its origin, called a data lineage.

Data Profile - A report which summarizes your existing shape of your data

THANK YOU