MLOPS WITH AWS SAGEMAKER

Session 3

Agenda

- Sagemaker DataWrangler Overview
- Understand Datawrangler Flow
- Sagemaker Feature Store overview
- **Demo - Use DataWrangler transform Data to publish to Feature Store for ML Pipeline**

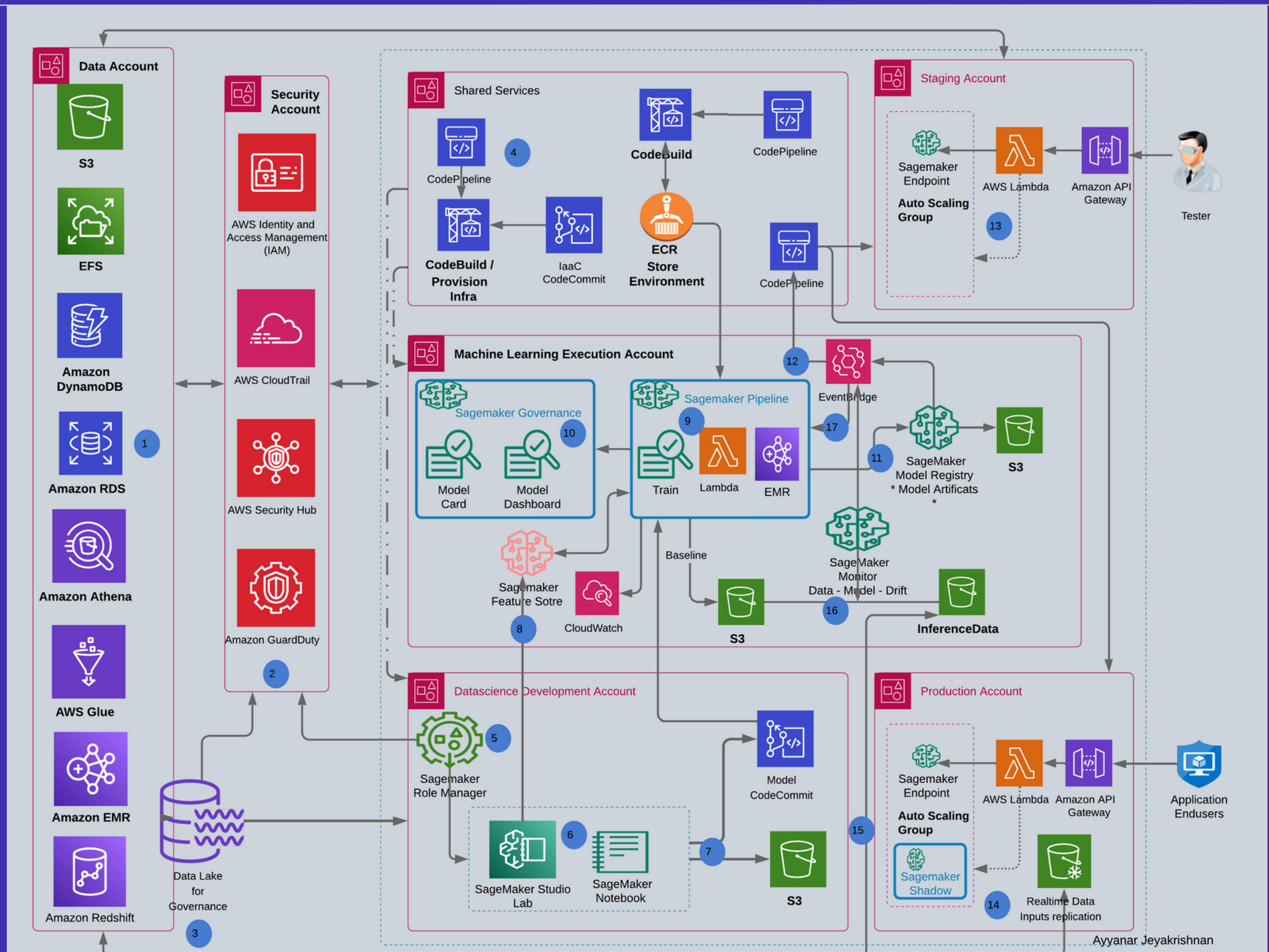
Presented by Ayyanar Jeyakrishnan - AWS Data User Group Bangalore Organizer and Cloundloud Tech Community Leader

RECAP

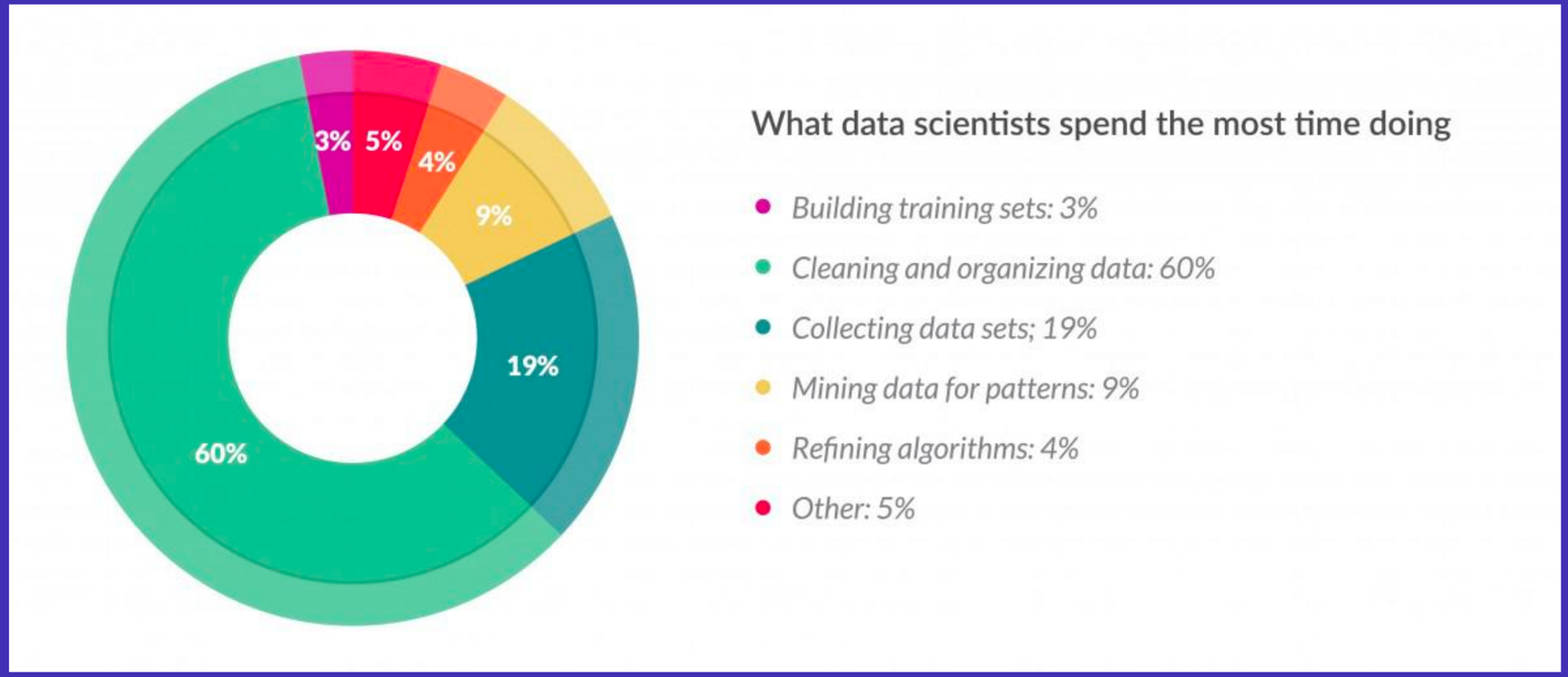
- In previous sessions, We discuss about what is MLOps, MLOps Challenges, MLOps Benefits. and highLevel Overview of AWS Sagemaker.
- Demo -Use the terraform code to Deploy the Sagemaker Studio with User profiles and Spaces.
- Demo - Sagemaker Pipeline
- Demo - Sagemaker Studio - Projects, MLOps Templates

* You can find the All Session Presentation and Recordings in GitHub - https://github.com/aws-data-usergroup-bangalore/sagemaker-mlops/tree/main

LET US RECALL OUR ROADMAP



DATA SCIENTISTS CHALLENGE



DATA PREPROCESSING

CLEAN
PARTITION
SCALE
UNBIAS
BALANCE
AUGMENT

FEATURE ENGINEERING

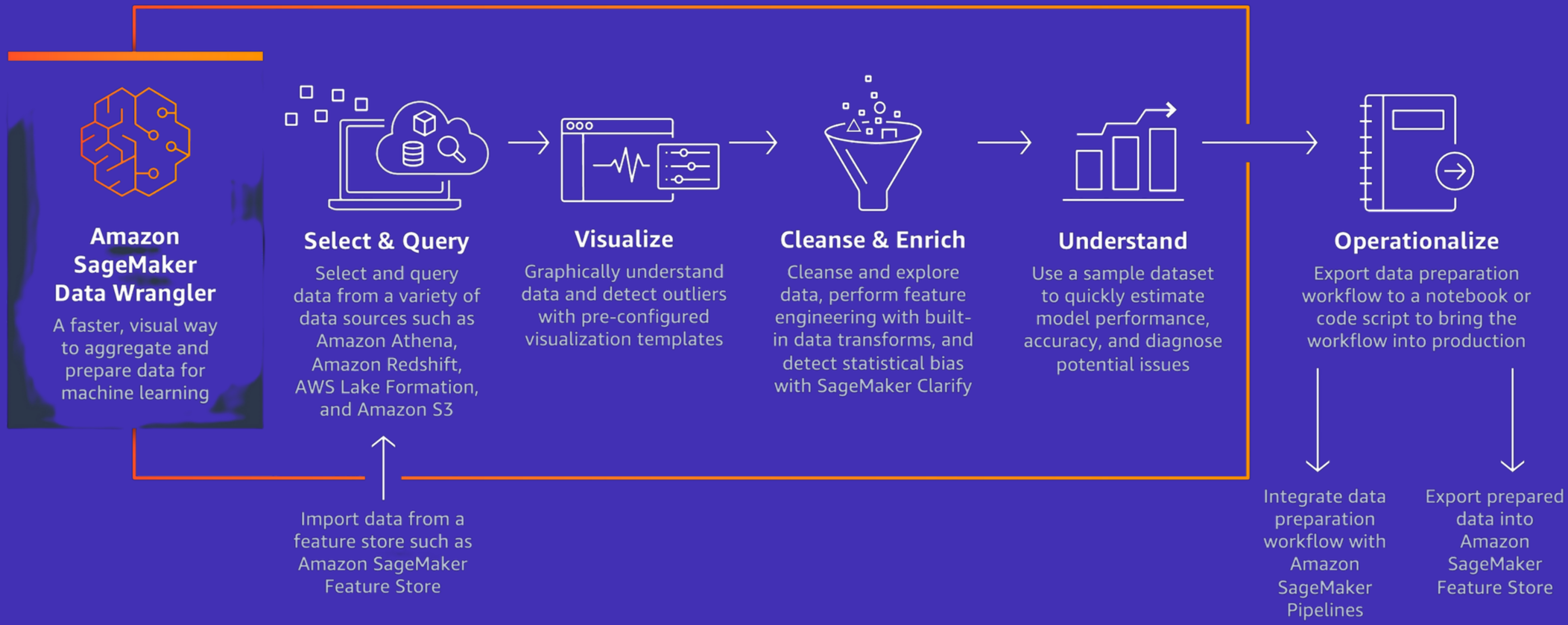
FEATURE CREATION
FEATURE TRANSFORMATION
FEATURE EXTRACTION
FEATURE SELECTION

https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=7ef9bd606f63

Cleaning the raw data is most time consuming and least enjoyable and repetitive Data Science task. But its Must Needed - Because its all start with Clean data

Presented by Ayyanar Jeyakrishnan - AWS Data User Group Bangalore Organizer and Cloundloud Tech Community Leader

SAGEMAKER DATA WRANGLER - INTRODUCTION



SAGEMAKER DATA WRANGLER - HIGH LEVEL OVERVIEW

IMPORT

Connect and Import data from S3, Athena, Redshift, EMR, Snowflake, Databricks, Salesforce and Also use Appflow to import 40+ datasource

DATA FLOW

Create a data flow to define a series of ML data preparation steps. You can use a flow to combine datasets from different data sources, identify the number and types of transformations you want to apply to datasets, and define a data preparation workflow that can be integrated into an ML pipeline.

TRANSFORM

Clean and transform your dataset using standard transforms like string, vector, and numeric data formatting tools. Featurize your data using transforms like text, date/time embedding and categorical encoding.

DATA INSIGHTS

Automatically verify data quality and detect abnormalities in your data with Data Wrangler Data Insights and Quality Report.

ANALYZE

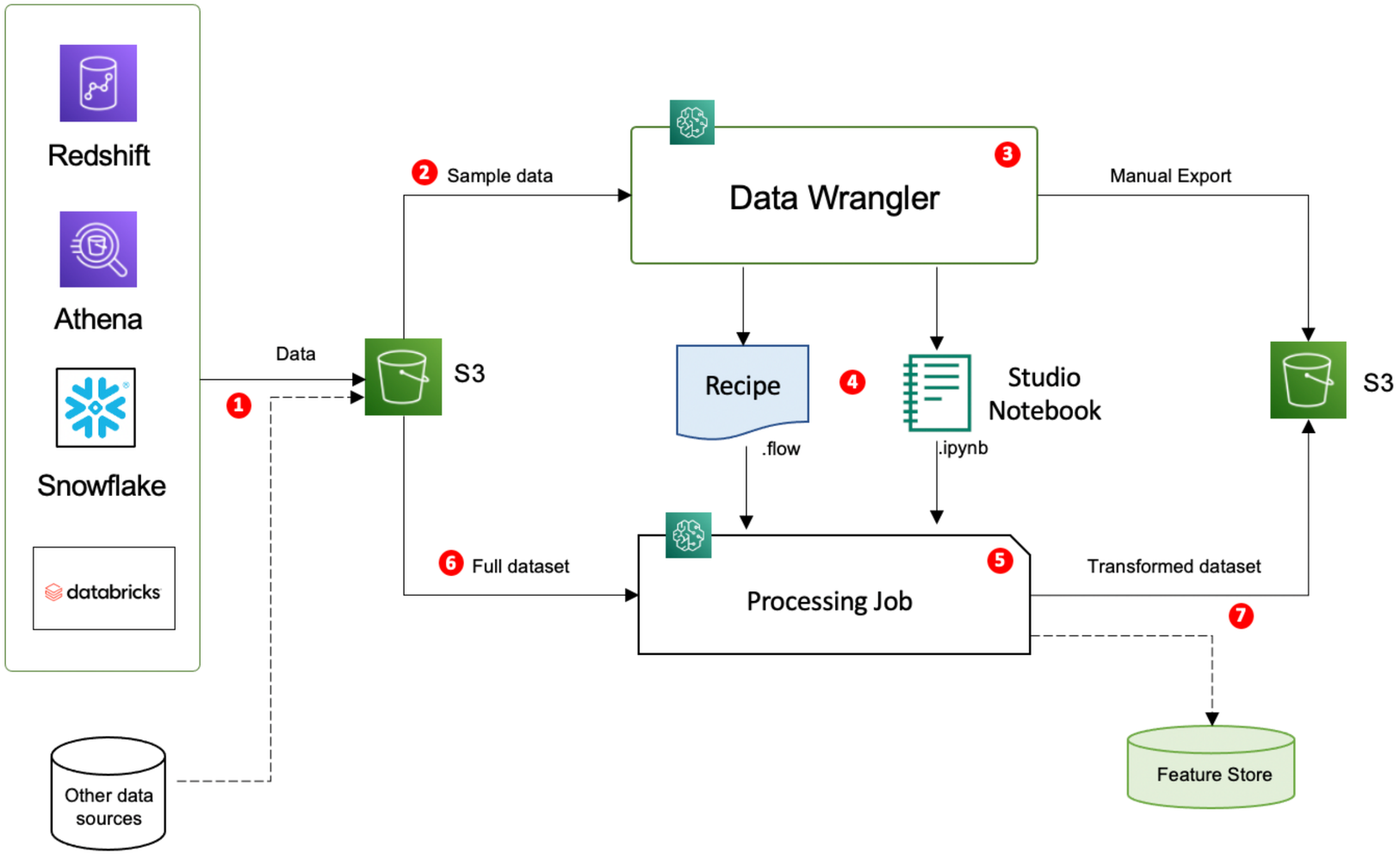
Analyze features in your dataset at any point in your flow. Data Wrangler includes built-in data visualization tools like scatter plots and histograms, as well as data analysis tools like target leakage analysis and quick modeling to understand feature correlation.

EXPORT

Export data preperation workflow to S3, Sagemaker Model Pipeline, Feature store, Python script / Notebook

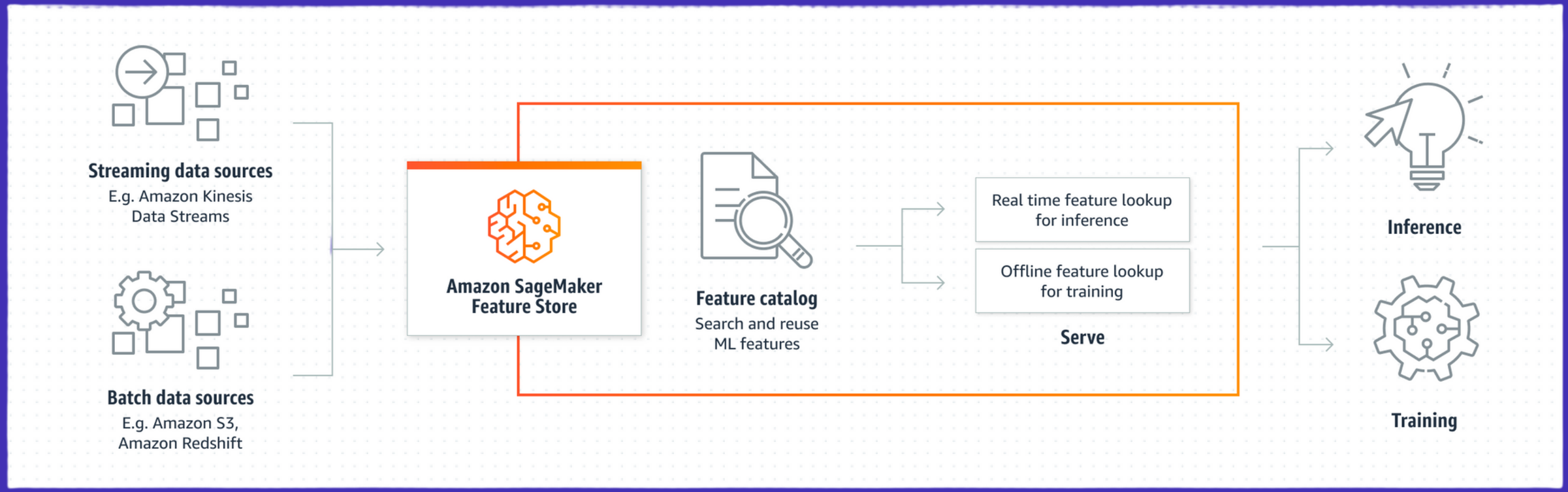
Presented by Ayyanar Jeyakrishnan - AWS Data User Group Bangalore Organizer and Cloundloud Tech Community Leader

SAGEMAKER DATA WRANGLER - LOW LEVEL OVERVIEW



1. Training data is imported into S3 either via Data Wrangler's SQL (from supported sources) or using alternative tools (from any other sources)
2. A sample subset of records from the pre-staged in S3 dataset is ingested into Data Wranger's UI
3. Exploratory Data Analysis is conducted and a set of data transformations is defined. Optionally, sample data can be manually exported
4. The transformations are saved as a "recipe" while a processing job orchestration code is generated and saved as a notebook.
5. Data processing job can either be triggered on-demand or as part of a training/inference pipeline
6. The processing job applies the transformation recipe to the entire dataset. It runs containerized Spark job which allows for parallel processing
7. Transformed data can be either exported into Feature Store or into S3

SAGEMAKER FEATURESTORE - INTRODUCTION



SAGEMAKER FEATURE-STORE - HIGH LEVEL OVERVIEW

INGEST

Include streaming data from sources such as logs, clickstreams and sensors, and tabular data from sources such as Amazon S3, Amazon Redshift, AWS Lake Formation, Snowflake, and Databricks Delta Lake. With Amazon SageMaker Data Wrangler you can publish features directly into SageMaker Feature Store. With the Apache Spark connector, you can batch ingest a high volume of data with a single line of code.

STORAGE

Uses the AWS Glue Data Catalog by default, but allows you to use a different catalog if desired. You can also query features using familiar SQL with Amazon Athena or another query tool of your choice.
SageMaker Feature Store tags and indexes feature groups so they are easily discoverable through the visual interface of Amazon SageMaker Studio. Browsing the feature catalog allows teams to discover existing features they can confidently reuse and avoid duplication of pipelines.

CONSISTENCY

SageMaker Feature Store supports offline storage for training and online storage for real-time inference.
Its ensures that offline and online datasets remain in sync which is critical because if they diverge, it can negatively impact model accuracy.

LINEAGE

To enable feature reuse with confidence, data scientists need to know how features were built and which models and endpoints are using them. SageMaker Feature Store enables data science teams to incorporate lineage tracking into their workflows.

TIME TRAVEL

PITR - Data scientists may need to train models with the exact set of feature values from a specific time in the past without the risk of including data from beyond that time, such as patient medical data before a diagnosis.

EXPORT

MLOps lifecycle. It manage datasets and feature pipelines, speeding up data science tasks and eliminating the duplicate work of creating the same features multiple times. It can be used as a standalone service or together with other SageMaker services in an integrated manner across the MLOps lifecycle.

Presented by Ayyanar Jeyakrishnan - AWS Data User Group Bangalore Organizer and Cloundloud Tech Community Leader

DEMO

How to Create a Data flow usign Data Wrangler

Export to S3 and Feature store

Create a Model using AutoPilot