databricks

Databricks 101

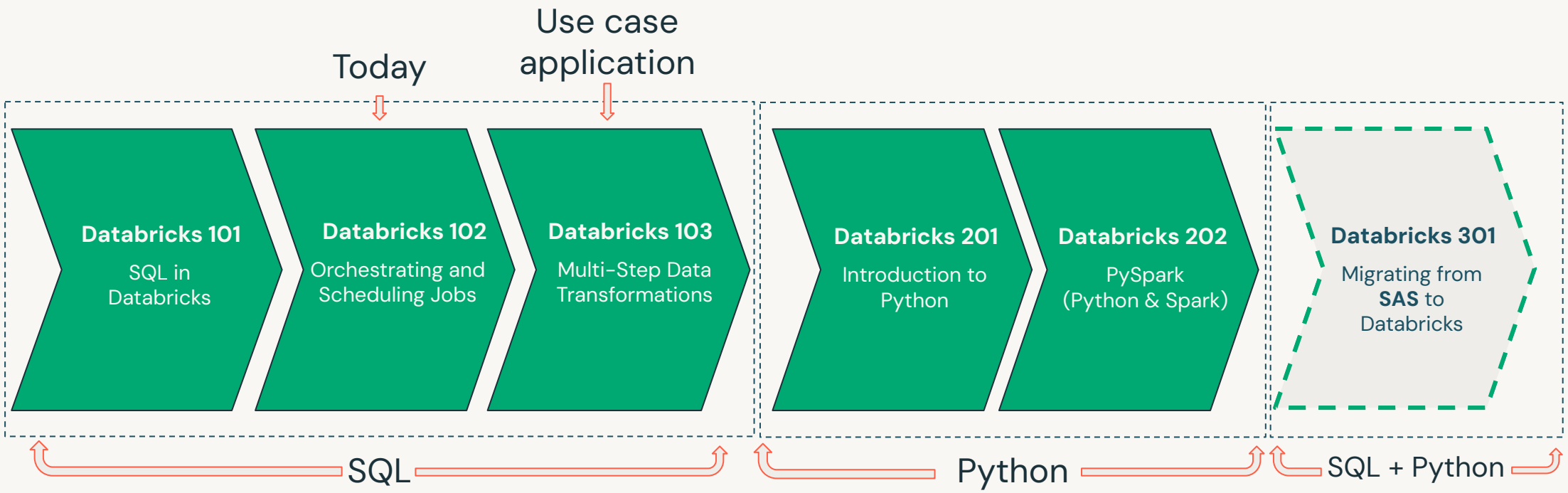
Orchestrating and Scheduling Jobs in Databricks

Dan Davis - Databricks

Agenda

- Review Previous Session
- Questions
- Clusters
- Databricks Workflows
  - Features
  - Use Cases
- Q&A / Exploration
- Next Steps

SAS Enablement Pathway



Databricks Resources

Start here

Start here today: Analyst Enablement Workspace

Start here for resources: D&A Reference Hub

Other Resources

- D&A Databricks Development Team Resources
- **Community**
  - Quick questions and community resources- Databricks Viva Engage page
  - Weekly Q&A with Databricks SMEs - DatabricksUserForum@KP
  - Databricks Community
  - FAQ
- **Training**
  - Recommended New AE Training - Training Link
  - Self-Directed Learning

Review Previous Session

Questions

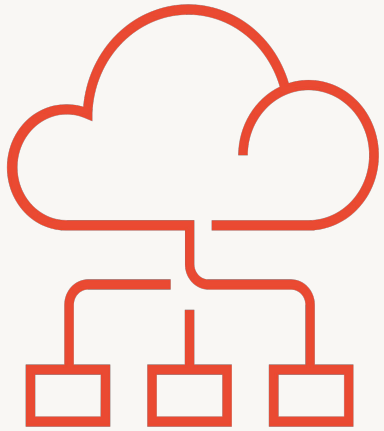
Clusters

Clusters

A collection of compute resources that execute your queries, commands, or jobs. Think of this as **SAS compute engine = Databricks cluster.**

Components of a cluster

Not important day-to-day, but helpful to know!



Driver (single per cluster)
Tells each worker what to process

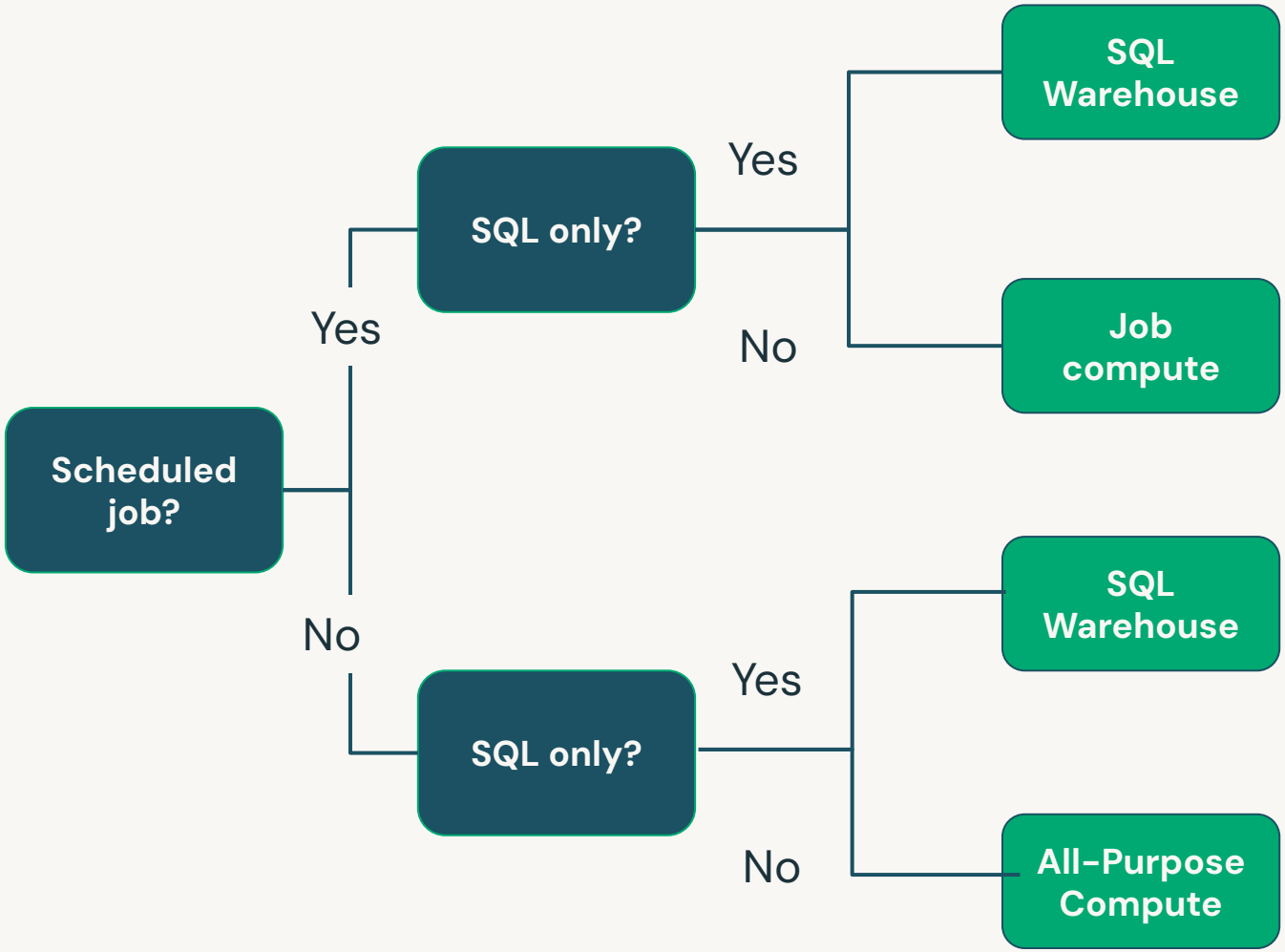
Workers (multiple per cluster)
Process a chunk of data. These scale out (or in) automatically based on how much data you are processing!

Types of Clusters

Cluster: A collection of compute resources that execute your queries, commands, or jobs. Think of this as **SAS compute engine = Databricks cluster. Clusters can be optimized for specific tasks.**

Type	What	Why	Cost
All-purpose	• Existing cluster you choose from. • Go to sleep after inactivity (e.g. 30 minutes) • 5-10 minute startup time when "waking up"	• Use languages other than just SQL (e.g. python, r) • Ad-hoc queries/development • Interactive Development	$$$
Jobs	• Newly created when you create a job	• Jobs using languages other than just SQL (e.g. python, r, etc.)	$
SQL Warehouse	• Existing cluster you choose from • Go to sleep after inactivity (e.g. 5 minutes) • 1-10 second startup time when "waking up"	• SQL-only jobs or development • SQL Queries • Dashboards and BI reporting (e.g. Power BI)	$$

Selecting a cluster



Compute Engine: SAS to Databricks Terminology

SAS	Databricks	Why
Workspace Server	All-Purpose Cluster	**SAS:** When working with SAS 9 in client-server architecture, it is possible to execute code remotely on a Workspace Server form a client tool such as Enterprise Guide. **Databricks:** you run Notebooks mainly on All-Purpose Clusters defined in the Workspace.
Batch Server	Job Cluster	**SAS:** Jobs are configured and scheduled through the SAS Management Console's Scheduling/Jobs interface (often called "**Management Console Scheduling/Jobs**," even though under the covers they execute as Batch Server jobs). **Databricks**: supports processing through Job Clusters - which is **cheaper** compute intended to run for scheduled jobs.
Proc SQL	SQL Warehouse	**SAS:** When you run PROC SQL, Data Steps, or any other PROC, they all execute inside the same SAS session—same memory space, same workspace server, same environment. There isn't a separate "SQL engine" to spin up; it just uses the same compute resources that you'd use for a Data Step or any other procedure. **Databricks:** Optimized compute to run SQL commands similar to a data warehouse. Think Oracle or Teradata.

Questions

Orchestrating and Scheduling Jobs