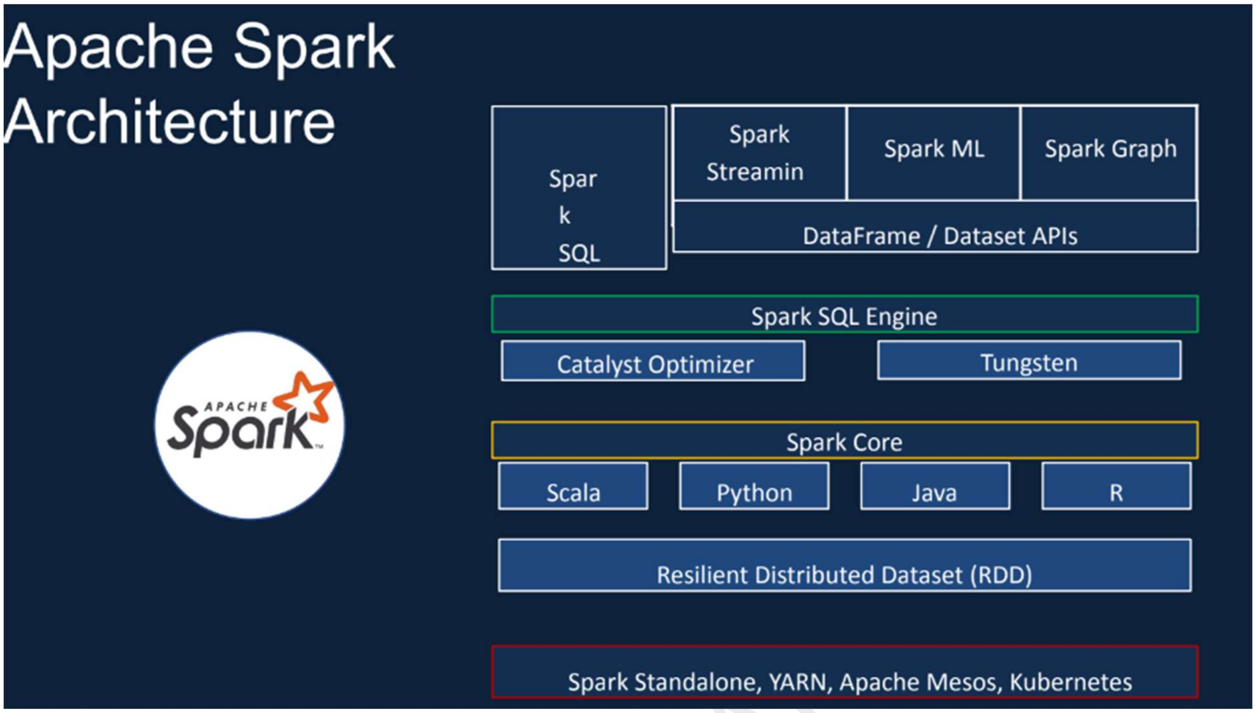
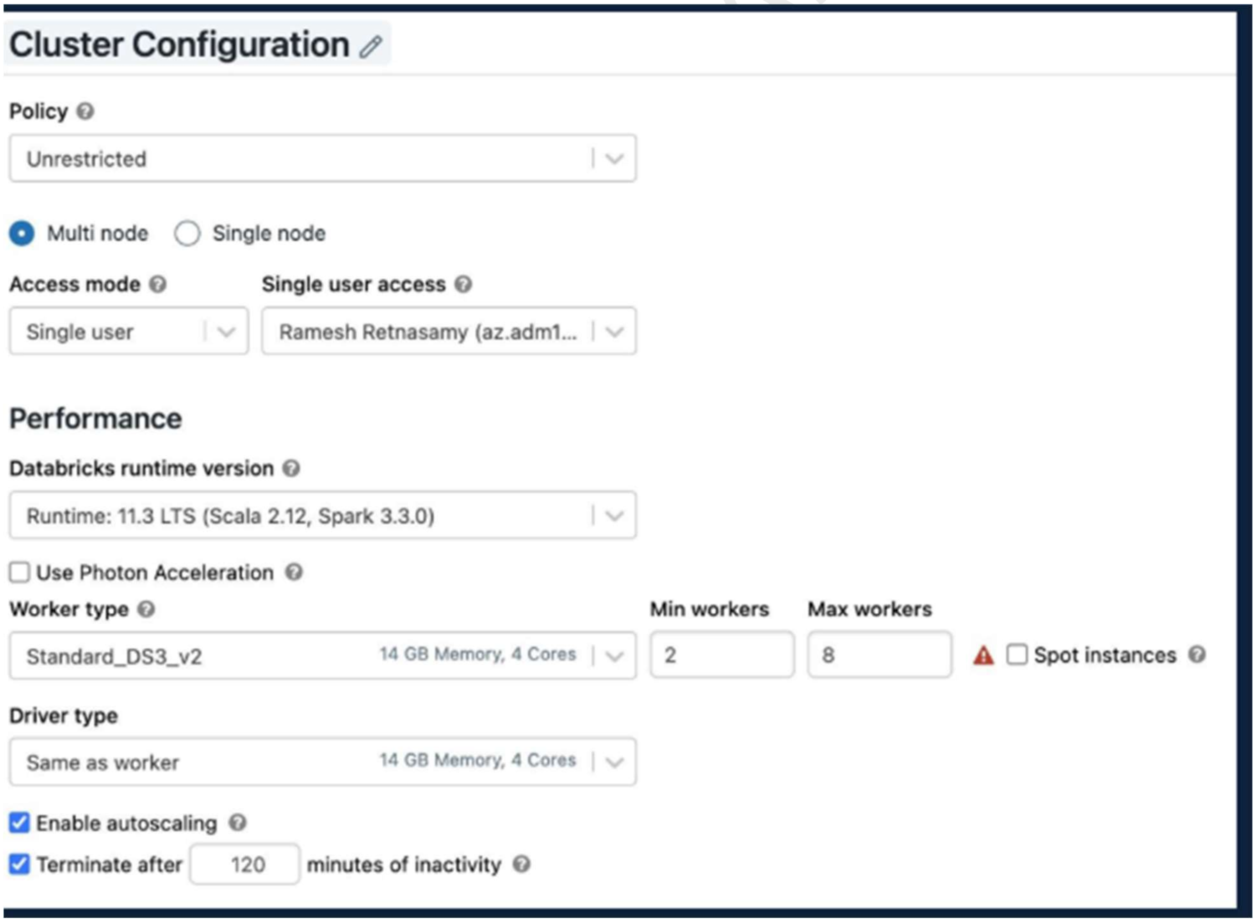


Apache Spark Architecture



Cluster Configuration

Policy

Unrestricted

● Multi node ○ Single node

Access mode	Single user access
Single user	Ramesh Retnasamy (az.adm1...

Performance

Databricks runtime version

Runtime: 11.3 LTS (Scala 2.12, Spark 3.3.0)

□ Use Photon Acceleration

Worker type		Min workers	Max workers	
Standard_DS3_v2	14 GB Memory, 4 Cores	2	8	□ Spot instances

Driver type

Same as worker 14 GB Memory, 4 Cores

☑ Enable autoscaling

☑ Terminate after 120 minutes of inactivity

1. What is Azure Databricks?

- A cloud-based data analytics and machine learning platform built on Apache Spark.
- Integrates with Azure for big data processing, analytics, and AI/ML development.

2. Key Features:

- **Collaborative Workspace:** Supports real-time collaboration for data engineers, analysts, and data scientists.
- **Unified Analytics Platform:** Combines data engineering, data science, and business analytics workflows.
- **Built-in Machine Learning Tools:** Includes ML libraries, automated machine learning, and integration with frameworks like TensorFlow and PyTorch.
- **Scalability:** Automatically scales resources for distributed computing.

3. Benefits:

- **Ease of Use:** Managed environment reduces setup and maintenance.
- **Performance:** Optimized Spark runtime enhances performance for large-scale data processing.
- **Integration:** Natively integrates with Azure services like Azure Data Lake, Azure Blob Storage, and Azure Synapse Analytics.
- **Security:** Enterprise-grade security features, including role-based access and encryption.

4. Use Cases:

- **Data Engineering:** ETL (Extract, Transform, Load) pipelines and real-time streaming.
- **Data Science:** Building and deploying machine learning models.
- **Data Analytics:** Interactive data exploration and business intelligence.
- **AI Applications:** Training and deploying AI models at scale.

5. Pricing:

- Pay-as-you-go pricing model based on usage (compute and storage).

6. Core Components:

- **Databricks Workspace:** Centralized interface for project collaboration.
- **Clusters:** Managed Spark clusters for distributed computing.
- **Notebooks:** Interactive notebooks for coding, visualization, and collaboration.
- **Delta Lake:** Storage layer for reliable, ACID-compliant data lakes.

7. Integration Highlights:

- Works seamlessly with Azure Data Factory, Azure Machine Learning, Power BI, and more.