Working with Spark in Cloud

Mohit Batra

Founder, Crystal Talks

linkedin.com/in/mohitbatra

Overview

Use Spark in Databricks

Use Spark in Azure Synapse Analytics

Using Spark in Databricks

Databricks

Unified Analytics Service

- Built on Apache Spark
- Provides managed Spark environment
- Provides tools for building and deploying large-scale data management and processing solutions
- Can run in Azure, AWS and GCP

To learn about Databricks, check out courses

- Building Your First ETL Pipeline Using Azure Databricks
- Delta Lake with Azure Databricks: Deep Dive

Demo

Prerequisites

- Azure Databricks workspace
  (check Setup document in Exercise Files)

In Azure Databricks workspace

- Setup multi-node cluster
- Write code in Notebooks and execute on cluster

Using Spark in Azure Synapse Analytics

Azure Synapse Analytics

Unified Analytics Service

- Data Ingestion, Data Warehousing, Big Data Analytics and more!

Multiple Compute Engines

- Apache Spark Pool
- Dedicated SQL Pool
- Serverless SQL Pool etc.

To learn about Azure Synapse, check out courses

- Data Literacy: Essentials of Azure Synapse Analytics
- Building Your First Data Lakehouse Using Azure Synapse Analytics

Demo

Prerequisites

- Azure Synapse Analytics workspace
  (check Setup document in Exercise Files)

In Synapse workspace

- Setup multi-node cluster
- Configure Spark session
- Write code in Notebooks and execute on cluster

Summary

Several cloud platforms are built on or support Spark

- Databricks, Azure Synapse Analytics, Azure HDInsight, AWS EMR etc.

Deploy multi-node cluster

Using managed services in cloud for Spark helps to:

- Easily provision infrastructure
- Scale on-demand
- Migrate to newer version quickly
- Use integrated security, management & logging
- Turn-off clusters when not in use to save cost

Thank You…

Keep Learning! ☺