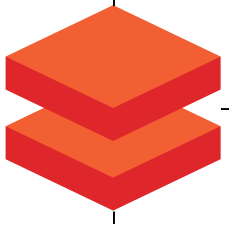
Compiled by Dr T Oberholster
DP-900 version: March 2021

[Website & Blog]
[Amazon Books]
[Teepublic Shirts]
[RedBubble Merch]
[SpreadShirt Apparel]
[Contrado Fashion]

Concepts

- Apache Spark environment
  - Massively Parallel Processing engine
  - Big data streaming/processing and machine learning
  - Libraries include SQL/statistical/machine learning
- GUI => define & test step-by-step processing before submission as batch tasks
  - Create & run R/Python/Scala scripts (Spark notebooks program)
  - Notebook cells = block of code
- Structured stream processing => incremental computations & continuous updates of data



Azure Databricks

Connections

- Azure Blob Storage
- Azure Data Lake
- Hadoop
- Flat files
- Azure SQL Database
- Azure Cosmos DB
- Sensors
- IOT devices

Configuration

- One-click setup
- Streamline workflows
- Interactive workspace
- Collaborate between Data scientists/engineers/analysts

- Extensible architecture drivers
  - driver = code connection data source to read/write
  - drivers are part of libraries loaded
  - Examples = Azure SQL Database/Azure Cosmos DB/Azure Blob Storage/Azure Data Lake/MySQL/Postgres
- Clustered or distributed in-memory models
  - R/Python/Scala/Java/SQL