Azure Databricks

Temario

1. Introduction
2. Data Sources
3. Data Manipulation
   3.1 Data manipulation is one of the main roles of Spark
   3.2 Needs of distributed libraries and structures (Spark DataFrames, RDDs…)
4. Spark ML
   4.1 Tensorflow
   4.2 PyTorch
   4.3 Keras
   4.4 XGBoost
5. Streaming
   5.1 Streaming aggregations
   5.2 Event-time Windows
   5.3 Windowed grouped aggregations
6. Cluster types
   6.1 Standard
   6.2 High Concurrency
7. Security
   7.1 Sing in to the workspace
   7.2 Workspace security and object access policies
8. API
   8.1 Databricks offers a REST API accesiblethrough any application
   8.2 Authentication must be specified by user tokens
   8.3 Almost any aspect of Databricks can be managed through API
9. Azure Integration (ADF + GIT)
   9.1 High parallelism
   9.2 Adaptative performance
   9.3 GitHub
   9.4 AureDevOps Services
10. ML Flow + Azure ML SDK
    10.1 Way too many tools and platforms
    10.2 What to track in an experiment? Where?
    10.3 Results should be reproducible