Apache Spark Quick Guide
Python

Joshua Jansen Van Vueren

December 3, 2019

Introduction

This document seeks to outline and describe the model utilised by Apache Spark.

"Apache Spark is a fast and general-purpose cluster computing system."

1 Terminology

- **Resilient Distributed Datasets - RDD:** fault-tolerant collection of elements that can be operated on in parallel. (Old Programming Interface - pre Spark 2.0)
- **Dataset:** Distributed collection of data, newer programming interface, better performance over RDD's (supported in Java and Scala but currently not in Python).
- **DataFrame:** is a DataSet organized into named columns, conceptually equivalent to a relational DB table (supported in Scala,Java,Python,R).
- **Parallelization:** members from collection are copied to form a distributed dataset, to be operated on in parallel. Slicing the data into a number of partitions to be used in the cluster.
- **External Datasets** spark can create distributed datasets from any storage source supported by Hadoop.
- **Cluster:** ...
- **Cloud Dataproc:** Google Cloud Dataproc lets you provision Apache Hadoop clusters and connect to underlying analytic data stores.

2 RDD Operations

- **Transformations** create new dataset from an existing one
- **Actions** return a value to the driver program after running a computation on the dataset.
- **Note** all transformations are lazy, therefore they do not compute all results immediately, only when required. One would need to run a `collect()` or some other action.
- **Models** ML models can be run on data through the spark structure, example here

3 Other Documentation and Links

- Google Cloud Storage Connector with Apache Spark