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."
"Apache Spark lets you use clusters of tens, hundreds, or thousands of servers to run simulations in a way that is intuitive and scales to meet your needs."

1 Terminology

- **Resilient Distributed Datasets - RDD:** fault-tolerant collection of elements that can be operated on in parallel. Optimized for parallel processing.
- **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. Can provision capacity on demand and pay for it by the minute.

- **Spark vs Beam:** Apache Beam can be classified as a tool in the "Workflow Manager" category, while Apache Spark is grouped under "Big Data Tools" [1].
- **Spark vs Hadoop: Hadoop is more focused towards batching and spark streaming. Spark utilises HDFS for**

Spark Memory Intensive
Streaming builds RDD's then creates micro-batches, processed by Spark engine, outputting processed data. Not per data stream.
Utility being a all-in-one solution for all processing needs.

Booking.com utilises spark for online ML features for real-time prediction of behaviour and preference of users.

Kafka
Data Pipeline

2 Spark Use Cases

- Streaming Data
  - Streaming Extract, Transform, Load
  - Data Enrichment
  - Trigger Event Detection
  - Complex Session Analysis - session activity easy to group and analyse
- ML

3 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

[1] https://stackshare.io/stackups/apache-beam-vs-spark

4 Other Documentation and Links

- Google Cloud Storage Connector with Apache Spark
- Apache Spark Use Cases
- Google Cloud Console Authentication Setup

Comparison Between Spark and Beam

RDD	PCollection	Both Immutable Data Storage Structures
Transformations	PTransform	Both receive their relevant immutable datasets and output the same type

4.1 Hive vs Spark

Hive Horizontally scalable.
SQL Interface, operating on Hadoop.
Built for data warehousing operations