Spark

A quick intro to Spark

Spark

- “Apache Spark is a fast and general engine for large-scale data processing” - Spark Website
- Provides interactive response times to large amounts of data
- Written in Scala, but can also be used from Java python Python and R
- Fault tolerant

Key concepts

- RDD (Resilient Distributed Dataset)
- Transformations (on RDDs)
- Broadcast variables
- Accumulator variables
- Tasks
- Executors

Resilient Distributed Dataset

RDD

RDD - Overview

- Immutable representation of a dataset
- Deterministic instantiation and transformation
- Distributed (partitions)
- Instantiated by
  - transforming another RDD
  - from an input source, like a file on HDFS
- Computation close to the data
- Fault tolerant (based on lineage)

RDD - Persistence

- Caching is handled by the developer
- An RDD can be cached in memory by calling the cache() method
- The persist() method lets you persist an RDD to
  - Memory
  - Memory and disk
  - Disk only
- Variable methods of serialization and replication

RDD – Transformations and actions

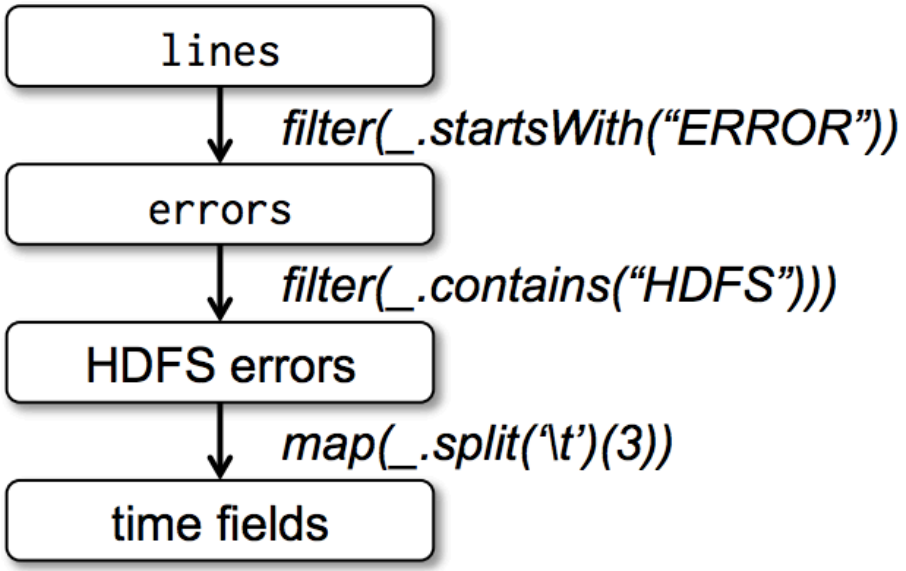
Method	Signature
map(f: T => U)	RDD[T] => RDD[U]
filter(f: T => Bool)	RDD[T] => RDD[T]
groupByKey()	RDD[(K, V)] => RDD[(K, Seq[V])]
join()	(RDD[K,V],RDD[K,W]) => RDD[(K, (V, W))]
partitionBy(p: Partitioner[K])	RDD[(K, V)] => RDD[(K, V)]

Method	Signature
count()	RDD[T] => Long
collect()	RDD[T] => Seq[T]
reduce(f: (T, T) => T)	RDD[T] => T
save(path: String)	Outputs RDD to a storage system, e.g., HDFS, Amazon S3

RDD - Example

```
val lines = spark.textFile("hdfs://...")
val errors = lines.filter(_.startsWith("ERROR"))
errors.persist()


// Returns Seq[String]
errors.filter(_.contains("HDFS"))
    .map(_.split('\t')(3))
    .collect()
```



(taken from Spark paper)

SHARED VARIABLES

Broadcast variable

- Immutable variable that is broadcasted to all nodes
- Useful for pre-computed tables, etc..

Accumulator variable

- A variable that supports an “add” operation
- Useful for implementing counters

EXECUTION

Tasks

- A task is the unit of execution in Spark
- Each partition of an RDD is mapped to a task
- Each task is executed by an executor

Spark ecosystem

- Spark Streaming: Live stream processing with Spark
- Shark SQL: SQL interface to RDDs (compatible with Apache Hive)
- MLib: Machine learning library built on top of Spark
- GraphX: Graph analysis on Spark

Resources

- https://spark.apache.org/research.html

QUESTIONS