Spark and Big Data Processing

Filip Peterek

Technical University of Ostrava

March 2021

Big Table of Big Contents

Big Table of Big Contents

Big Data (Little Penis)

My Hadoop Is Bigger Than Yours

Apache Spark

What is Big Data

- Large volumes of data
- Cannot be processed using traditional methods
- Often unstructured
- Though some form of a structure is highly desirable
- Parallel processing is required
- Big Data is not just the dataset, but also the entire field which focuses on handling of such large datasets

Utilization of Big Data

- Statistics
- Business intelligence
- Machine learning
- Determining common denominators

Big Data, Little Benefit?

So is Big Data just another way of exploiting user data to make more money?

The Big Question of Big Data?

Is Big Data evil?

The Big Answer to the Big Question

No.

The Big Answer to the Big Question

No. Big Data is a tool.

The Big Answer to the Big Question

No. Big Data is a tool. Big Data is nothing more than a tool used to analyze and handle data. It's as evil as any other tool.

The Big Evil

Big Data is as evil as a hammer, a screwdriver, a car, a brick

The Big Danger of Big Data

Is Big Data dangerous?

The Big Danger of Big Data

Is a gun dangerous?

The Big Danger of Big Data

Is a gun dangerous? Depends on who's holding the gun.

The Big Danger of Big Data

Is a gun dangerous? Depends on who's holding the gun.
Is a gun dangerous in the hands of a police officer?

The Big Danger of Big Data

Is a gun dangerous? Depends on who's holding the gun.
Is a gun dangerous in the hands of a police officer? Only if you're black.

Big Data, Big Benefits

Profits of a company and the benetifs of society are not mutually exclusive

Big Data, Big Benefits

Profits of a company and the benetifs of society are not mutually exclusive

- Think Apple Watch

Big Uses of Big Data

- Medicine and Healthcare
- Infrastructure
- Smart cities (think Singapore)
- Environmental issues (think Singapore)
- Improving tools that make life easier (think for profit companies)

Big Table of Big Contents

Big Data (Little Penis)

My Hadoop Is Bigger Than Yours

Apache Spark

Big Data, Big Tools

Hadoop is a word often used when talking about Big Data

Big Data, Big Tools

Hadoop is a word often used when talking about Big Data
So what exactly is Hadoop?

Big Data, Big Cluster

Apache Hadoop

- Framework for distributed processing of large datasets
- Consists of many open source technologies
  - Yarn
  - HDFS
  - MapReduce
- Usually runs on a cluster
  - Master
  - Executors

Big Data, Big Community

There is also a lot of related Big Data Processing tools, which work well in combination with Hadoop

- Cassandra
- HBase
- Avro
- Kafka
- Pig
- Spark

Big Table of Big Contents

Big Data (Little Penis)

My Hadoop Is Bigger Than Yours

Apache Spark

Big Spark, Big Fire

What is Spark?

- Latest and greatest Big Data processing tool
- Two modes – standalone and Hadoop
- Replacement for MapReduce
    - Faster
    - Smarter
    - More readable
    - Easier to use
    - Less terrible language

Big Spark, Big Fire

What is Spark?

- Written in Scala
- Runs on JVM, just like the rest of Hadoop
- Can be used with any JVM language
  - Scala works the best as Spark was built to work with Scala
  - Java is oficially suported, but is a terrible language
  - Kotlin support is being worked on, but is currently subpar (source: experience)
  - Other JVM languages can use Java API

Big Data, Big Transformations

The core of Big Data processing is the application of lazy transformations on immutable datasets

- RDD – Resilient Distributed Dataset
    - Resilient and Fault Tolerant
    - Consists of both data and a set of applied data transformations
    - Immutable
    - Low level, though there are higher level abstractions over RDD
- Functionally pure, lazy transformations
    - Map
    - Filter
    - Fold
    - Reduce
    - Select
    - Group By

Big Data, Big Schema

As mentioned before, structure is highly desirable

- Unstructured data should be given a structure
  - Tabular data
  - JVM objects

Big Data, Big Structures

There are multiple data structures in Spark

- RDD
  - Low level generic structure consisting of data of any non-primitive data type
  - Divided into partitions
- Dataset
  - High level abstraction over RDD
  - Provides higher level API
- DataFrame
  - Typealias for Dataset[Row]
  - Represents tabular data

Big Data, Big Rows

Row is an essential data type for working with tabular data

- Represents a single row of a table
- Data is described by a schema
- Not a template, fields are void*
    - Types are checked at runtime
    - Programmer should make sure to respect the schema to avoid runtime errors
- Structure is not necessarily flat
    - Arrays
    - Row fields inside other Rows
    - Row is more of a struct rather than an untyped array

Big Data, Big APIs

Transformations can be applied in multiple ways

- SQL
  - Programmers can write SQL queries
  - Works well with tabular data
  - Supports SQL constructs and UDFs
- RDD/Dataset API
  - RDD or Dataset methods
  - Uses Scala methods and lambda functions to apply transformations
  - In it's essence very similar to writing SQL queries
  - Can be used with custom classes, not just Rows

Big Data, Big Performance

Performance is critical when working with large datasets

- Plan generation, optimization and execution
  - Transformations are not applied directly
  - Instead, a plan is generated
  - Said plan is then optimized
  - Optimized plan is compiled to JVM bytecode and transferred to executors
  - Executors execute code
  - Driver (master) controls executors
  - When execution is finished, driver collects the results
- Lazy evaluation
  - Allows Spark to avoid unnecessary operations

Big Data, Big Formats

Multiple formats are used when working with Big Data

- CSV
  - Simple, human readable format
  - Nested structures slightly more difficult to implement
  - Slow to process, large files due to plaintext nature
- Avro
  - Binary format
  - Requires schema
  - Custom classes emitted by Avro compiler
- Parquet
  - Binary format
  - Self-described, doesn't require an external schema
  - Parsed into Spark Rows

Big Data, Big Mess of a Presentation

Any questions?

Big Data, Big Thanks

Thank you for at least not disturbing since none of you were paying attention anyway

Big Data, Big Sources

https://hadoop.apache.org
https://spark.apache.org