Functional Programming using Scala

(Application with Spark)

Faaiz Hussain Shah

University of Montpellier
(BforeAI)

14/10/2024

Agenda

1. Introduction

2. Big Data Technology

3. Apache Spark

4. Conclusion

Distributed Computing

What is Distributed Computing?

A computing paradigm or environment in which components of a software system are shared among multiple computers to improve efficiency and performance.

Figure: figure

Distributed Computing [1]

[1] https://cloudxlab.com/blog/introduction-to-big-data-and-distributed-computing/

Distributed Computing: Key Characteristics

- **Scalability**: Systems can easily scale out to accommodate increased load
- **Fault Tolerance**: Systems are designed to continue operating even if parts fail
- **Concurrency**: Multiple components can operate simultaneously

Distributed Computing: Benefits of Distributed Systems

- Increased computational power
- Redundancy and reliability
- Resource sharing across different geographies

Distributed Computing: Challenges

- **Network Issues**: Latency, bandwidth limitations
- **Security Concerns**: More endpoints, more vulnerabilities
- **Complexity in Management**: Difficulty in synchronizing and managing multiple systems

Distributed Processing

Distributed processing architecture, such as Apache Spark, is a software infrastructures designed to process large amounts of data across a **cluster** of interconnected machines

They enable the **distribution** of computational **tasks** across multiple processing nodes **in parallel**, which provides **greater processing** capacity, better **scalability**, and **enhanced performance**.

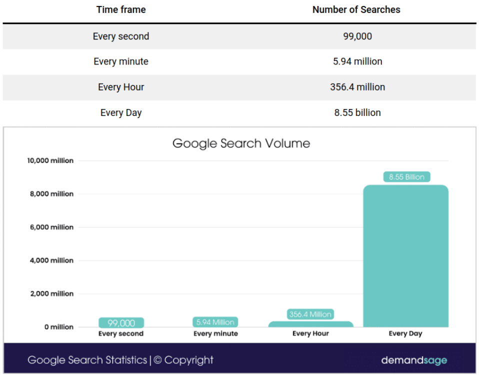
Spark is one of the most popular and powerful distributed processing architectures.

Big Data Technology

- Data has not only become the lifeblood of any organization, but also it is growing exponentially
- The challenge is how to get business value out of this data

What is big data ?

Big Data Technology



Time frame	Number of Searches
Every second	99,000
Every minute	5.94 million
Every Hour	356.4 million
Every Day	8.55 billion

Figure: Google Search Volume [1]

[1] https://blog.hubspot.com/marketing/google-search-statistics

Big Data Technology

- Data has not only become the lifeblood of any organization, but also it is growing exponentially
- The challenge is how to get business value out of this data

What is big data ?

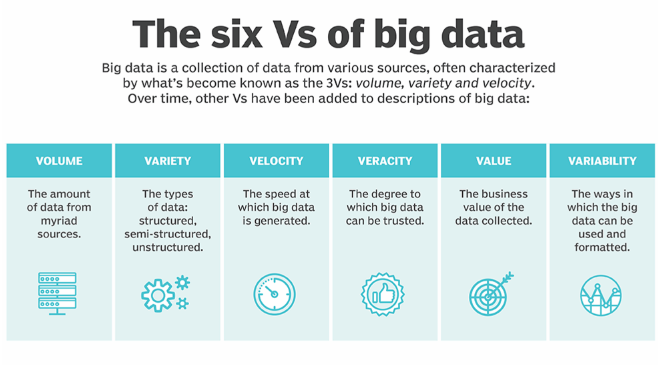
Big Data Technology

What is big data ?

1. Is it a dataset whose volume exceeds petabytes or several terabytes ?
2. A relational database table with billions of rows ?
3. a relational database table with thousands of columns ?

Although the term “big data” is hot, its definition is quite vague

Big Data Technology



The six Vs of big data

Big data is a collection of data from various sources, often characterized by what's become known as the 3Vs: *volume, variety and velocity*. Over time, other Vs have been added to descriptions of big data:

VOLUME	VARIETY	VELOCITY	VERACITY	VALUE	VARIABILITY
The amount of data from myriad sources.	The types of data: structured, semi-structured, unstructured.	The speed at which big data is generated.	The degree to which big data can be trusted.	The business value of the data collected.	The ways in which the big data can be used and formatted.

Figure: The six Vs of big data [1]

[1] https://www.quora.com/What-are-the-six-Vs-of-Big-Data

Big Data Technology

- Standard relational databases could not easily handle big data
- The core technology for these databases was designed several decades ago when few organizations had petabytes or even terabytes of data
- Today it is normal for some organizations to generate terabytes of data every day

Hence there was a need for new technologies that could not only process and analyze large volume of data, but also ingest large volume of data at a fast pace.

Big Data Technology

- Key driving factors for the big data technologies include:
    - **Scalability**
    - **High availability**
    - **Fault tolerance**

What is Apache Spark™?

Apache Spark is a unified analytics engine for large-scale data processing. [1]

It provides high-level APIs in Java, Scala, Python and R, and an optimized engine that supports general execution graphs. [1]

Apache Spark™ is a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters.[2]

[1] https://spark.apache.org/docs/latest/index.html

[2] https://spark.apache.org/

Apache Spark™ Components and Libraries

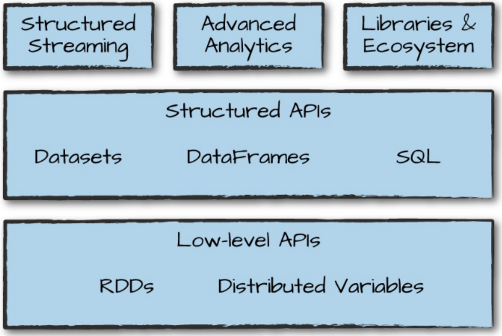


Figure: Spark Components & Libraries [1]

[1] Chambers, Bill, and Matei Zaharia. Spark: The definitive guide: Big data processing made simple

Spark Key Features: Programming Model

Abstract Programming Model

An abstract programming model provides a simplified, high-level interface for programming, abstracting away complex lower-level details

- Spark offers an abstract programming model called **Resilient Distributed Datasets (RDD)**, which allows data to be processed transparently across the cluster
- RDDs are **immutable** and **fault-tolerant** collections, meaning they can be distributed across multiple computing nodes and retrieved in case of failure
- RDD is defined as an abstract class (i.e., it can not be instantiated) in Spark library

Spark Key Features: In-Memory Processing

- Spark uses **RAM** to store intermediate data and computation results, which allows for rapid access to data without having to read from disk
- This enables faster response times and more efficient task execution

Spark’s in-memory processing is particularly advantageous for iterative processing because it significantly reduces the time taken to read from and write to disk, thus speeding up the iterations

Spark Key Features: Batch and Real-Time Processing

- Spark supports both **batch** data processing and **real-time** (streaming) processing
- It enables continuous analysis on real-time data streams, as well as **iterative processing** for machine learning algorithms

Iterative processing in machine learning

It involves repeatedly applying the same steps to refine the model's parameters until the model meets a specific criterion, such as a set number of iterations or a minimum error threshold

Spark Key Features: Extensive Ecosystem

- Spark has a rich ecosystem with a comprehensive library of components, including **Spark SQL** for SQL processing, **Spark Streaming** for real-time processing, **MLlib** for machine learning, **GraphX** for graph processing, and many more.
- It facilitates the development of complex applications using a coherent set of tools

Distributed processing architectures

- Distributed processing architectures like Spark are used for various applications, such as:
  - big data analysis
  - real-time stream processing
  - distributed machine learning
  - personalized recommendation
  - predictive analytics, etc.
- They leverage the parallel computing power of distributed clusters for fast and efficient processing of large volumes of data

Conclusion

1. We studied basics about distributed computing
2. We got a quick overview of big data
3. We learned about key features of Apache Spark using Scala

Thank you