

DATA AND ARTIFICIAL INTELLIGENCE



Big Data Hadoop and Spark Developer

DATA AND ARTIFICIAL INTELLIGENCE

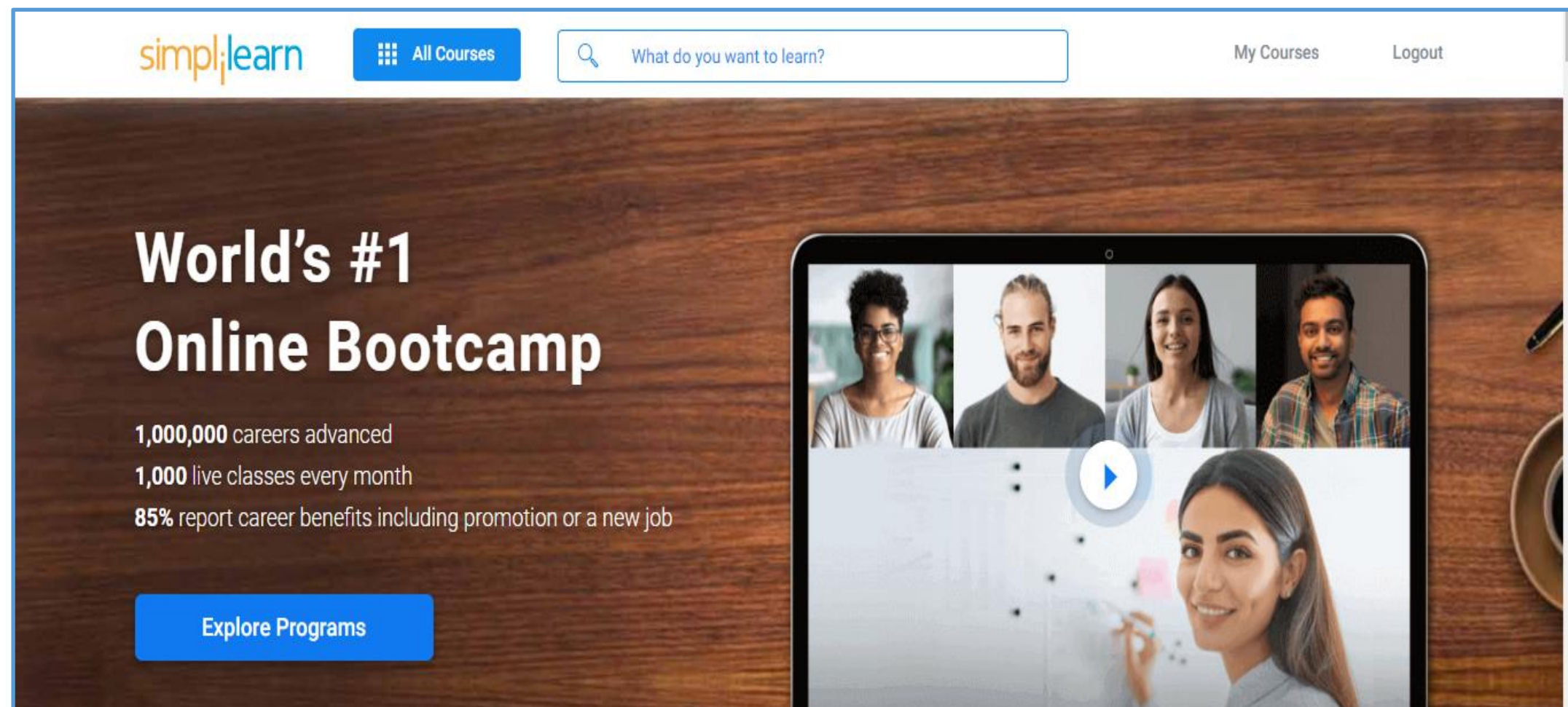


Course Introduction

About Simplilearn

Simplilearn

For over a decade, Simplilearn has focused on digital economy skills.
Now, Simplilearn has become the **World's #1 Online Bootcamp**.

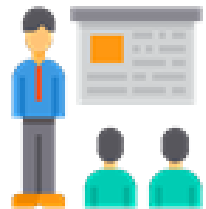


Simplilearn

Simplilearn
provides:



Live virtual classes (LVCs)



Self-paced
learning content



Interactive labs

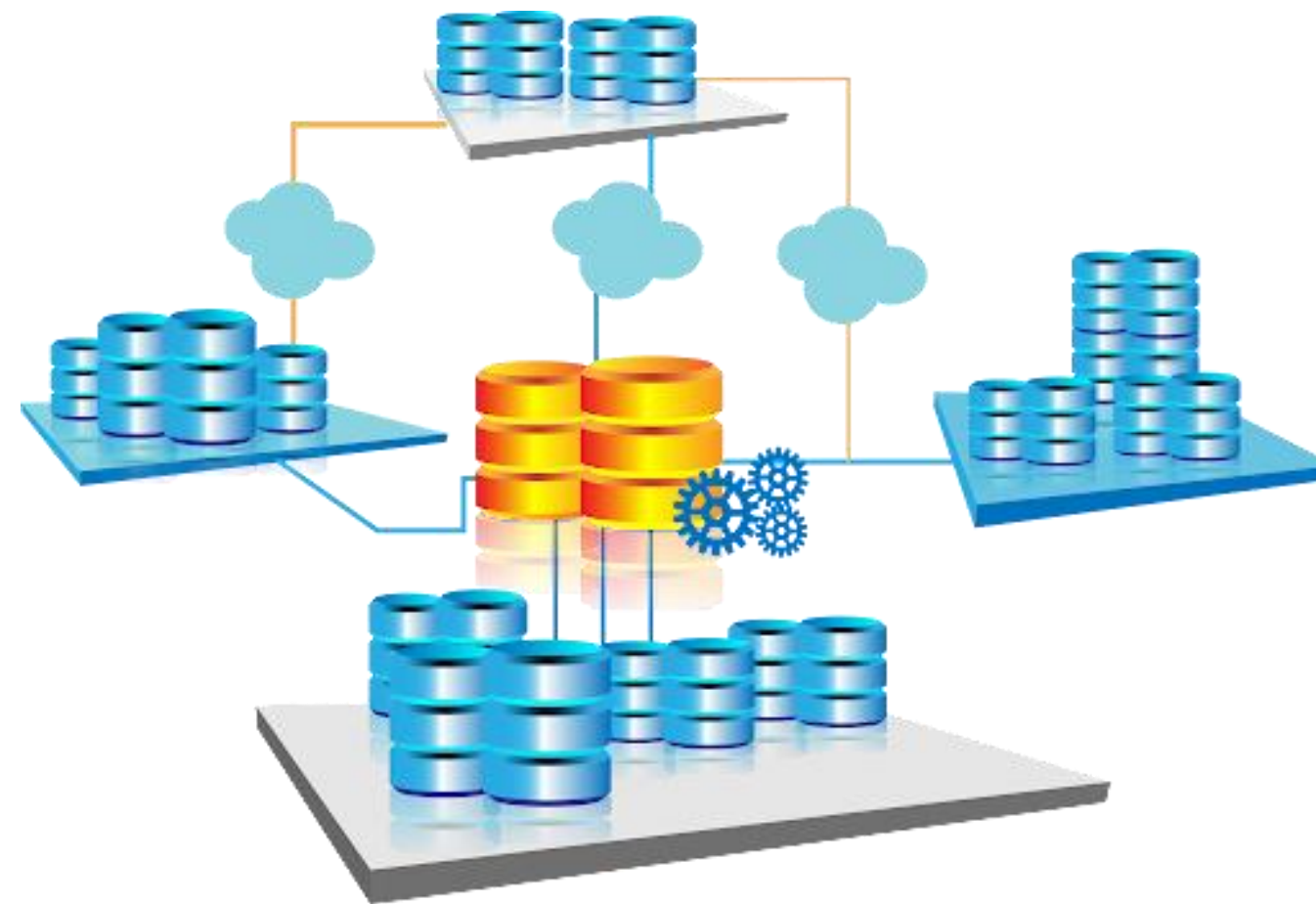


Real-time,
scenario-based projects

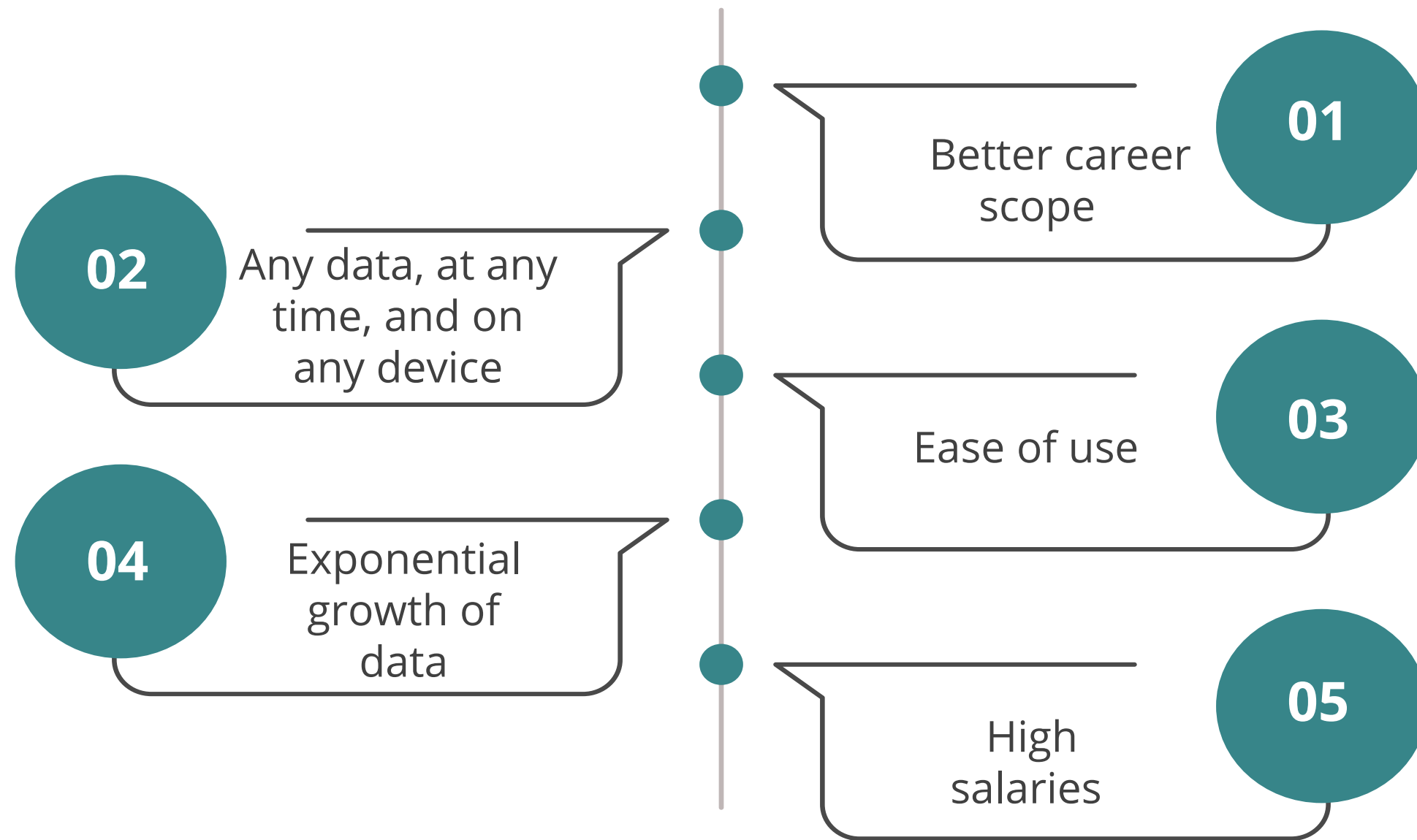


What Is Big Data?

Big data is an open-source software framework for storing data and executing applications on commodity hardware clusters.

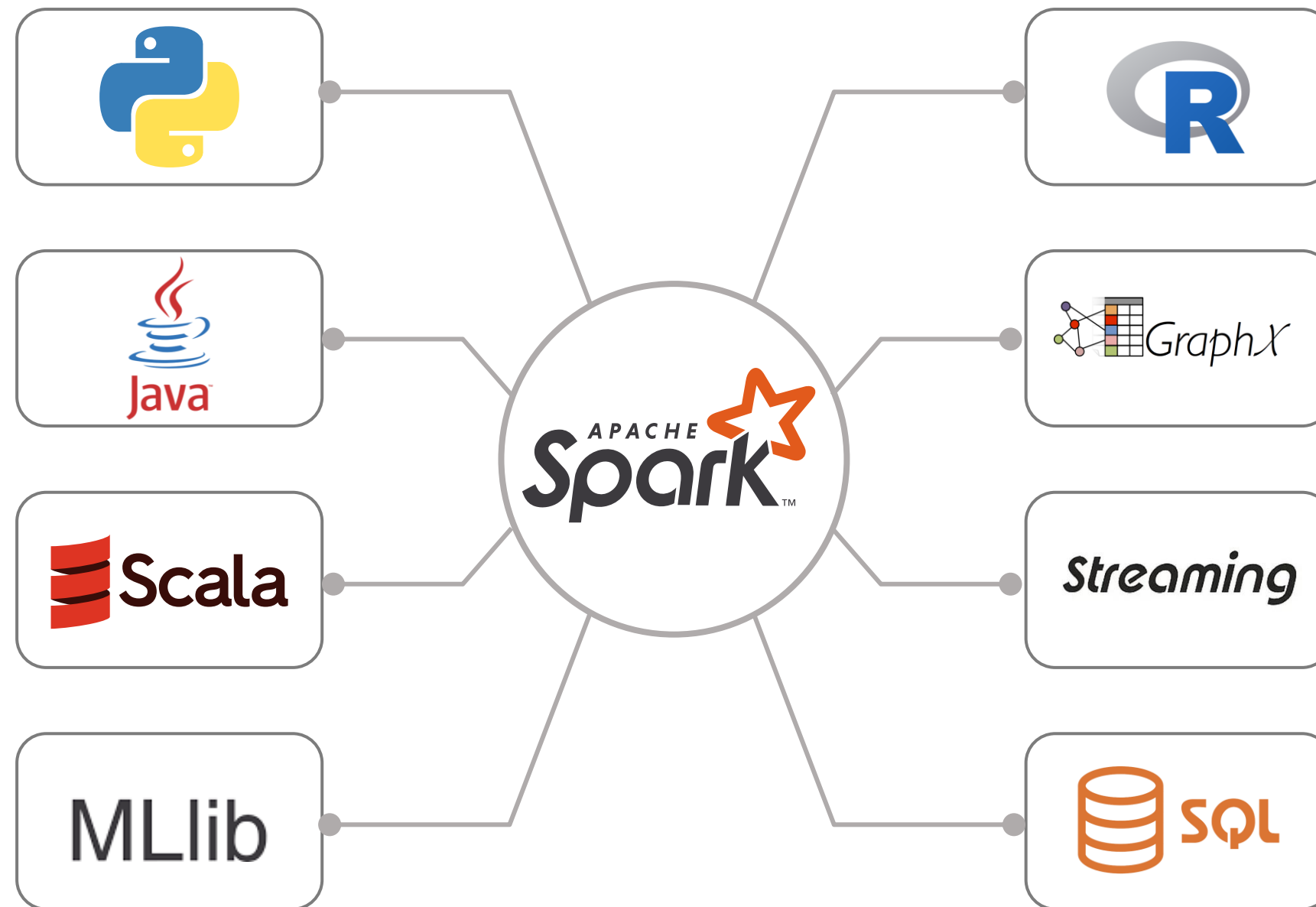


Why Big Data?



Apache Spark

Apache Spark is an open-source cluster computing framework for real-time data processing.
It contains the following components:

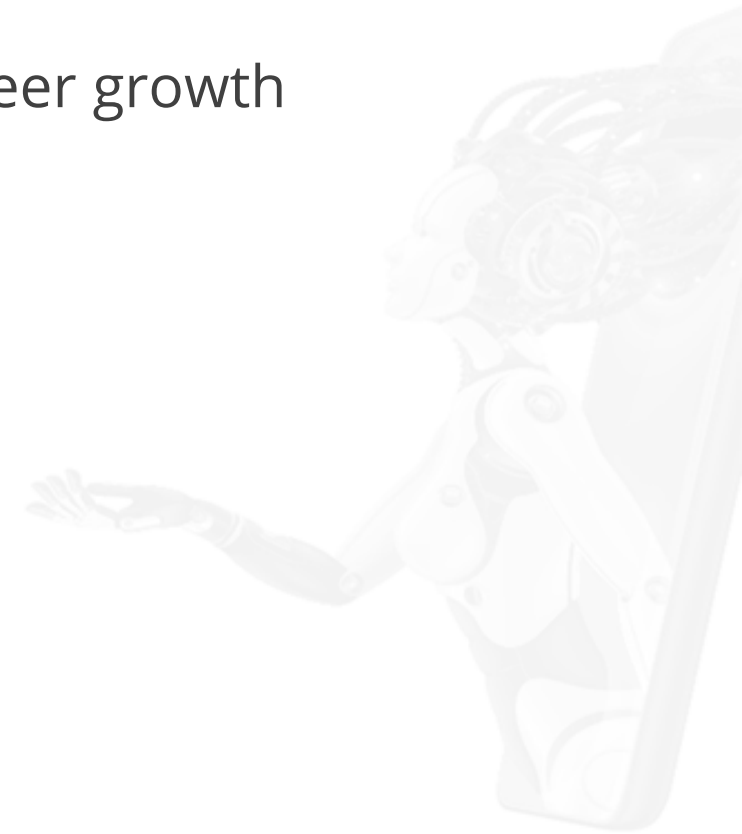


Why Apache Spark?

More than 91% of companies use Apache Spark because of its performance gains. It has:

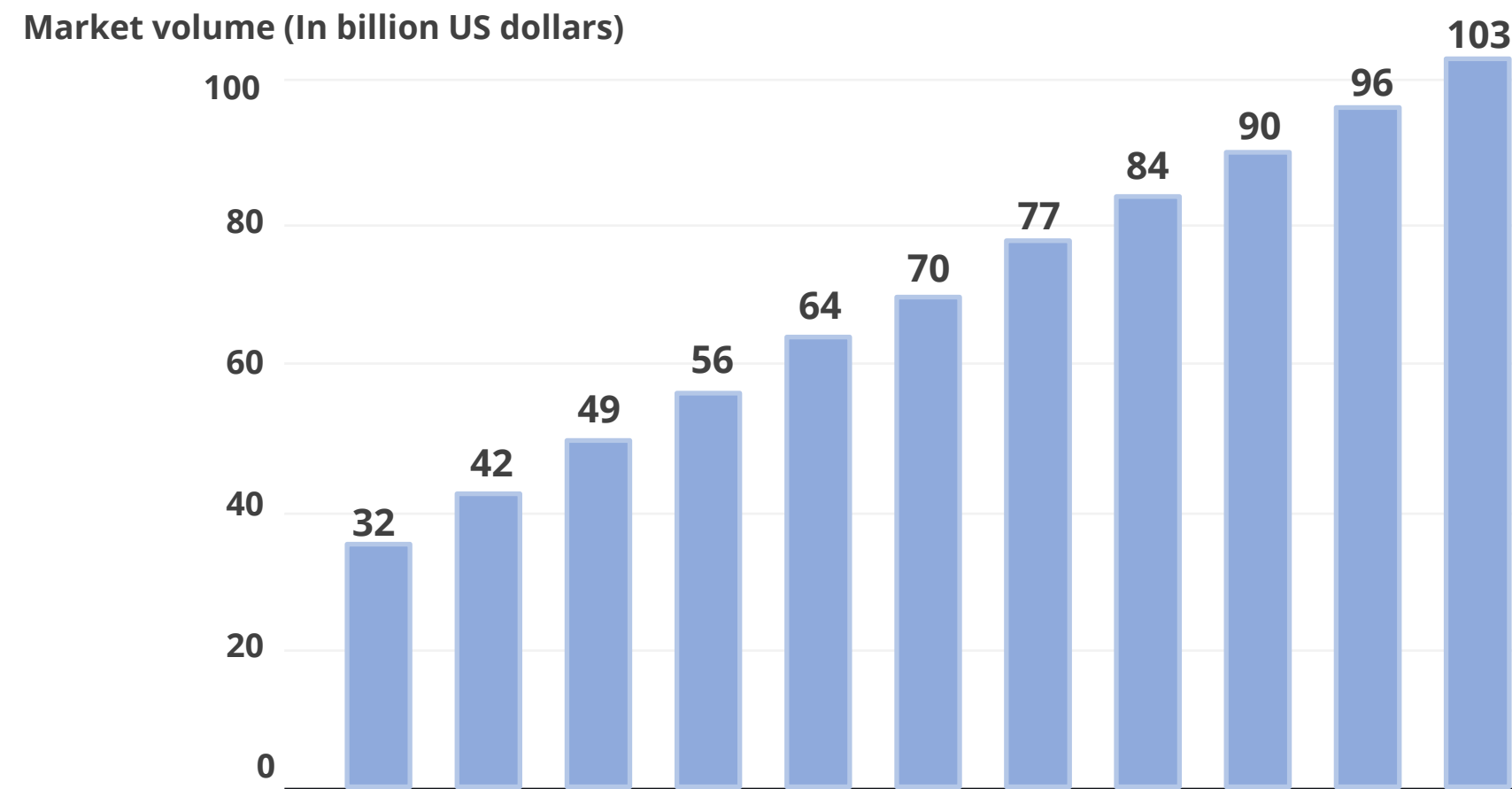


Demand for Big Data and Apache Spark



Demand for Big Data and Apache Spark

The demand for Big data is increasing in various data science fields. In the future, it is expected that this demand will continue to grow significantly.



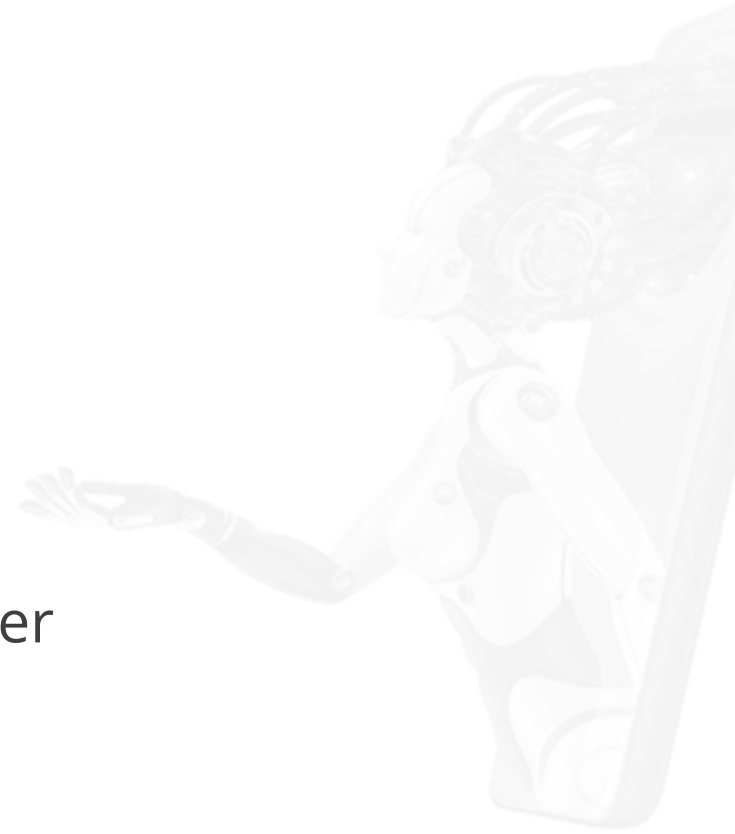
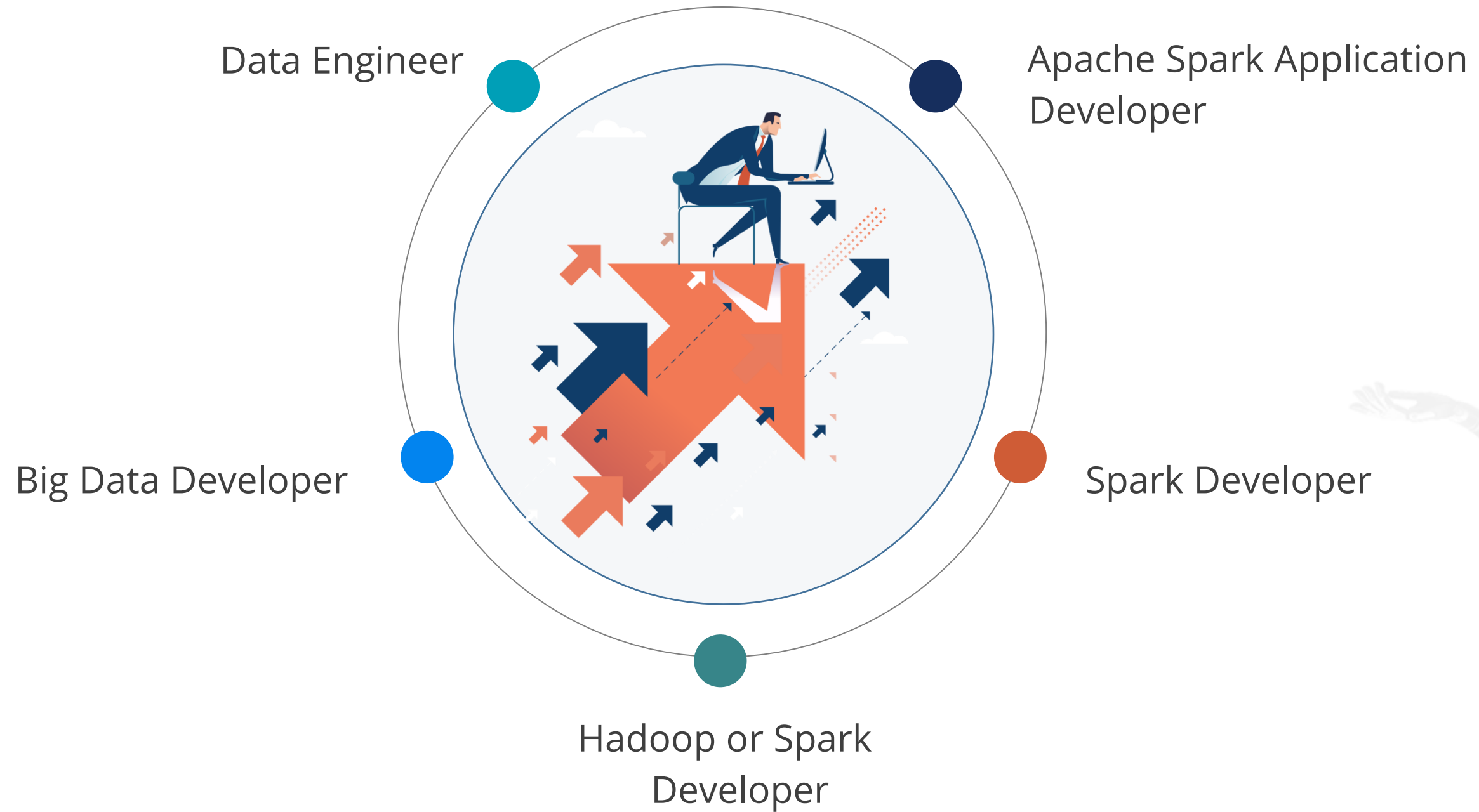
Source: <https://appinventiv.com/blog/spark-vs-hadoop-big-data-frameworks/>

Companies Hiring Data Engineers

Many companies around the world hire data engineers. These include:



Career Opportunities



Prerequisites

Prior knowledge and understanding of the following languages:



JAVA



SQL

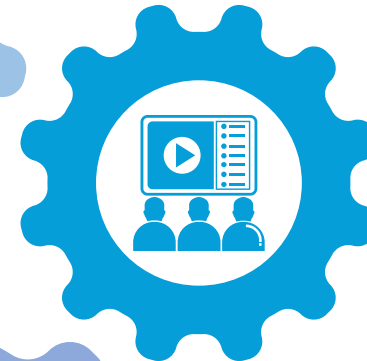
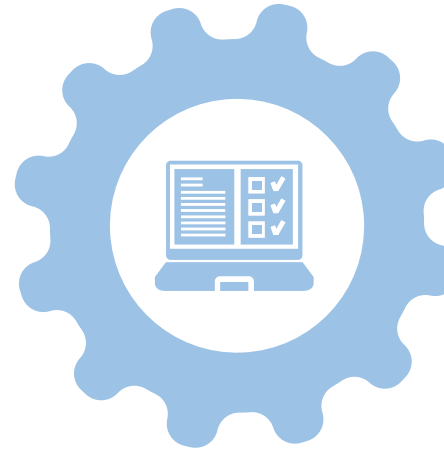


Simplilearn Program Features

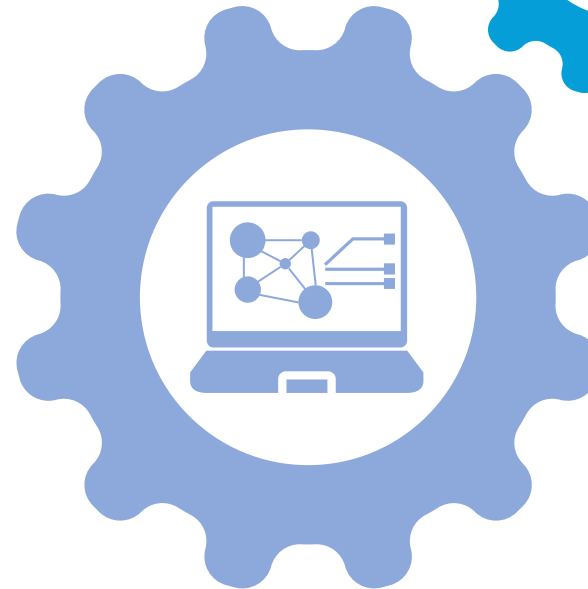
Program Features

The blended learning program is a combination of:

Self-paced learning
content



Live virtual classes
(LVCs)



Hands-on exercises



Program Features

The program contains the following features:



Theoretical concepts



Case studies



Integrated labs



Projects



Program Features

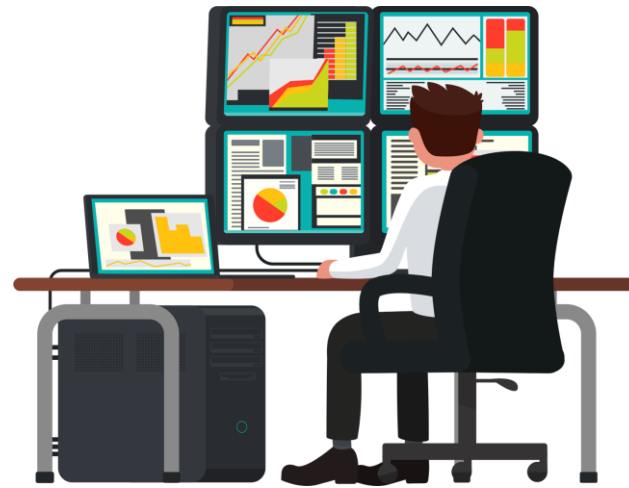
The class sizes are limited to foster maximum interaction.



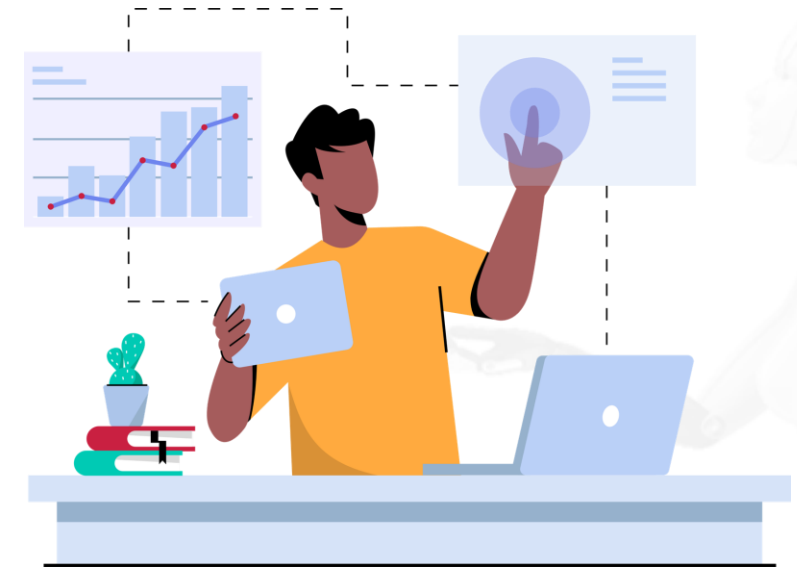
Target Audience



Students



IT Professionals

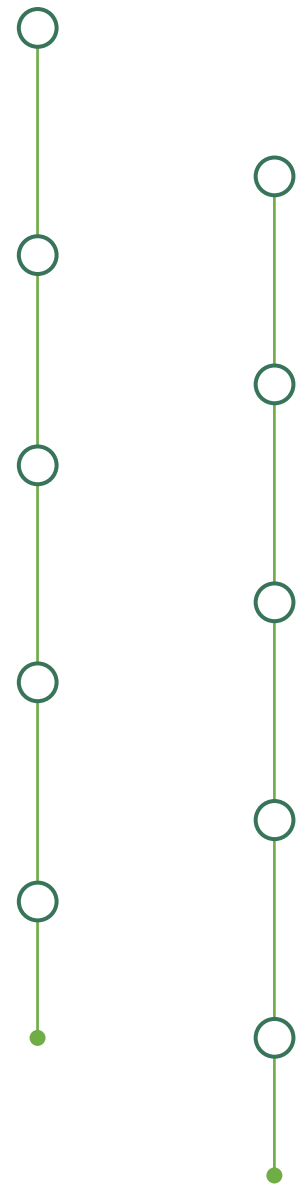


Data Engineers

Learning Path

Course Outline

The outline of the course helps to understand the path of Big data Hadoop and Spark developers.

- 
1. Course Introduction
 2. Introduction to Big Data and Hadoop
 3. HDFS: The Storage Layer
 4. Distributed Processing: MapReduce Framework
 5. MapReduce: Advanced Concepts
 6. Apache Hive
 7. Pig-Data Analysis Tool
 8. NoSQL Databases: HBase
 9. Data Ingestion into Big Data Systems and ETL
 10. YARN Introduction

Course Outline

11. Introduction to Python for Apache Spark

12. Functions, OOPS, and Modules in Python

13. Big Data and the Need for Spark

14. Deep Dive into Apache Spark Framework

15. Working with Spark RDDs

16. Spark SQL and Data Frames

17. Machine Learning Using Spark ML

18. Stream Processing Frameworks and Spark Streaming

19. Spark Structured Streaming

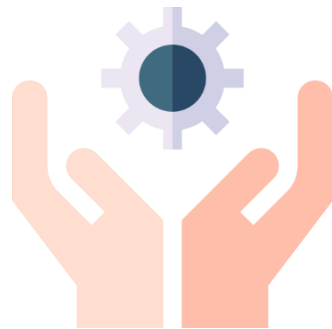
20. Spark GraphX

Course Components

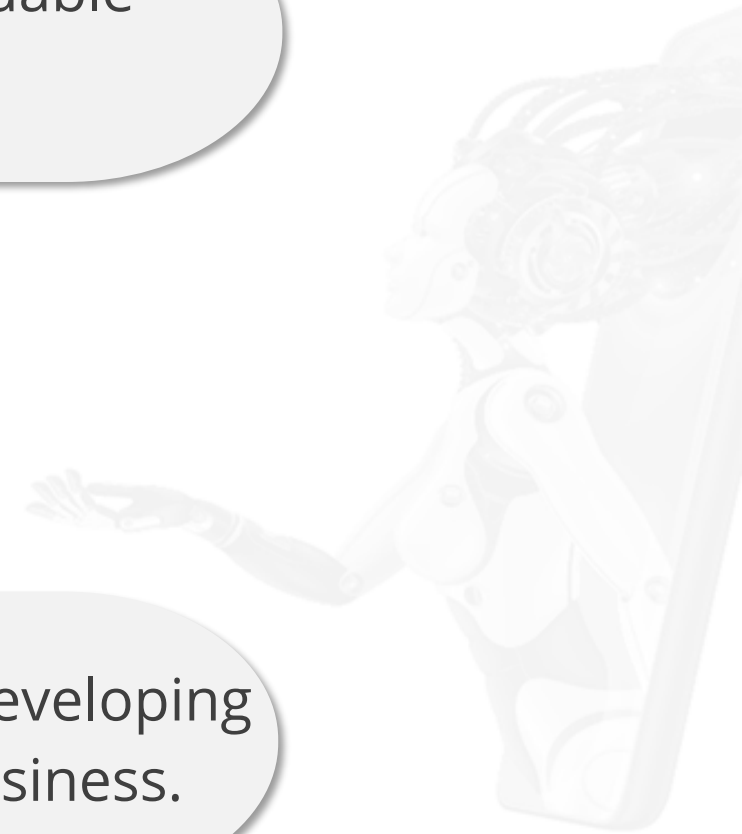
Course Components



E-books: All lessons are available as downloadable PDF files for quick reference guides.



Assisted practices: These will assist you in developing abilities that will make you an asset to any business.



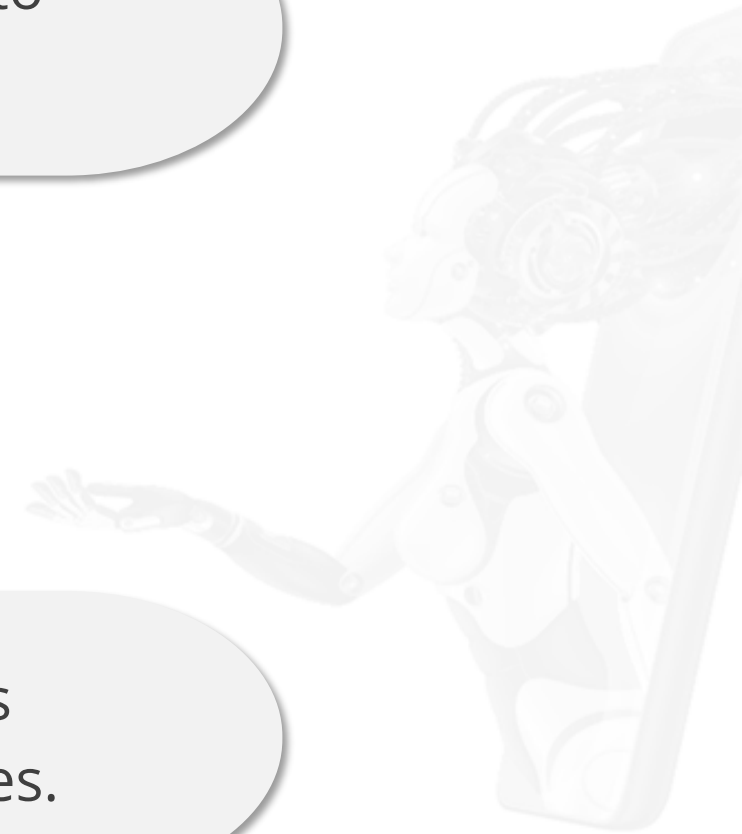
Course Components



Assessments: There are over 100 questions to assess your knowledge.



Projects: Lesson-end and course-end projects provide real-time and industry-based examples.

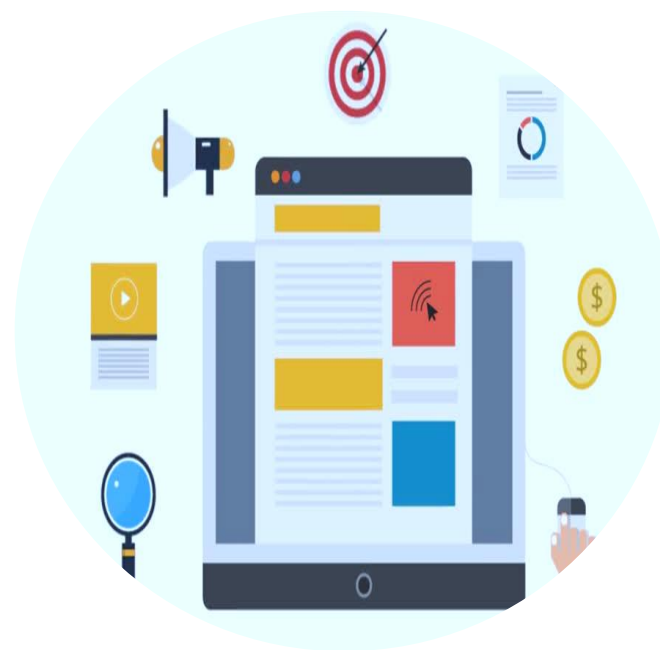


Course Completion Criteria

The learner needs to complete:



85% OSL or 80%
LVC classes



Course-end
assessment



At least one project

Course Outcomes



By the end of this course, you will be able to:

- Create an interaction between users and Hadoop Distributed File System using Hive
- Create an internal and external Hive table structure to read data from different formats
- Execute batch jobs using MapReduce frameworks
- Work with real-time streaming data pipelines and applications using Kafka



Course Outcomes



By the end of this course, you will be able to:

- Create Spark applications using Spark 3.x cluster and client mode
- Determine the components of Spark machine learning and GraphX
- Create and execute a real-time pipeline using Spark streaming and structured streaming
- Analyze the appropriate tools based on the data trends



DATA AND ARTIFICIAL INTELLIGENCE

Let's get started!