XALTIUS
Towards Higher Excellence

It's all about Analytics!

# Big Data Analytics using Spark

Big Data has evolved over the last few years and has become mainstream for many big organizations. Corporates, Academicians, Students who want to learn more about the Big Data Ecosystem and how it can be leveraged for predictive and streaming analytics should attend this workshop.

www.xaltiusacademy.com

## Start your data science journey with us.

Basics understanding of Python is required to attend this course.

XALTIUS
Towards Higher Excellence

**A short course in 16 hours.**

# What You'll Learn

This course covers the fundamentals of Big Data using PySpark. Spark is a "fast cluster computing framework" for Big Data Processing. It lets you run programs and operations up-to 100x faster in memory. You will be exposed to various libraries in PySpark for Data Processing and Machine Learning. You will have a chance to work with various datasets through guided hands-on training. At the end of this course, participants will gain an in-depth understanding of PySpark and its application to general Big Data analysis. This workshop will be conducted using a tool called Databricks which is used today to run big data loads on spark.

**Learn how you can put your data to good use.**

# About the Course

These are the key takeaways that participants will gain:

- Basic understanding of Big Data processing
- How to play with data and programming
- Understanding of Distributed dataset
- Exposure to Dataframes
- Leveraging Spark to do Analytics
- Work with various business cases

It's about your goals & your future!

**Contact us |** info@xaltius.tech or +65 8303 9150 / +65 9138 9813

# Course Outline

## Module 1: Introduction to Big Data Ecosystem and Databricks

The participants will be introduced to the basics of Big Data as well as the various concepts and different frameworks for processing Big Data. They will learn major tools and frameworks available in the market for Big Data, Pros and Cons of the tools and frameworks.

## Module 2: Introduction to Big Data Analysis using PySpark

The participants will be exposed to the basics of spark with python, playing with data and functional programming.

## Module 3: Programming in PySpark

The participants will be exposed to the backbone of Spark, resilient distributed dataset RDDs. We will learn how RDDs are created, executed and various transformations and actions (map, reduce, collect among others) using RDDs.

## Module 4: PySpark SQL and Dataframes

Structured data processing is important when profiling and understanding data. Spark provides an elegant method for the above using Spark SQL. The participants will be exposed to dataframes, the distributed SQL query engine, operations using Spark SQL and data visualization using PySpark.

## Module 5: Machine Learning with PySpark

Participants will be exposed to various machine learning methods and algorithms and will work will different datasets to perform regression, clustering, classification among other such operations. Participants will also understand and learn to go through the process of model training and evaluation.

## Module 6: Streaming Analytics

Participants will be exposed to real time streaming data and how spark can be leveraged to deal with real time data and perform real time analytics.

## Module 7: Practice and Extra Hands-on Workshops

Participants will have the opportunity to work with various datasets and practice all the operations and techniques learnt over various modules. Trainers and Assistant Trainers will help the participants through their exercises and practice sessions. This will give the participants an opportunity to get a thorough grasp of programming using PySpark.

# Join us & launch your career in data science.

It's time to upskill for the Industry 4.0

**www.xaltiusacademy.com**

**Contact us |** info@xaltius.tech or +65 8303 9150 / +65 9138 9813