BIG DATA USING SPARK PROGRAM

An eleven-week in-depth program covering the Apache Spark® and how it fits with Big Data

DePaul University’s Big Data Using Spark Program is designed to provide a rapid immersion into Big Data Analytics with Spark. Apache Spark® helps data scientists, data engineers and business analysts more quickly develop the insights that are buried in Big Data and put them to use driving customer interactions, product development, and more. Apache Spark takes users beyond Hadoop, simplifies development and data access, and helps data professionals develop a wide range of algorithms.

IT professionals will be given a broad understanding of the different leading Big Data/Spark technologies along with the technical skills required to successfully implement and maintain Spark clusters. The program begins with an introduction to the Spark cluster and teaches the ways to interact with the Spark Resilient Distributed Datasets (RDD), Data Frame and the job process using different shells. The topics also include different higher level interfaces and tools to manage data process in the Spark cluster, such as SparkSQL, R, etc. The program consists of an effective mix of interactive lecture and extensive use of hands-on lab exercises. Students will build their own Big Data applications using different Spark platforms from commercial-level distributors such as Cloudera, Hortonworks, Databricks, IBM, Microsoft, Amazon, and others.

Classroom lectures and demonstrations will be complemented by hands-on labs, reading assignments, case studies, and projects. In order to maximize learning, students will be required to bring their own laptop computer to every class session. While access to most cloud services explored in the program will be provided to students in class, there may be some cloud services that are only accessible via the use of a student’s own credit card. Students should expect to spend a small fee to access these services.

YOU WILL LEARN:

- Big Data technologies and trends
- Spark ecosystem
- How to perform typical ETL, queries, joins, and aggregations on datasets using Spark and SQL
- Spark RDD and Data Frame as a programming model for distributed processing of large volumes of data
- How to build and deploy applications that utilize the Spark platform (on-premise and Cloud) using different programming languages (Python, Scala, R) and tools (Notebook, Jupyter, Pandas, etc.)
- Showcase real world Spark usage patterns that are commonly used in industries
BIG DATA USING SPARK PROGRAM

Big data is a company’s most valuable asset. It is the driving force for business decisions and processes—helping businesses spot emerging trends, identify new markets, improve customer service and operate more efficiently. Apache Spark® is a compelling multi-purpose platform for use cases that span investigative as well as operational analytics. Spark has enterprise-grade capabilities and componentry for enhanced resiliency and optimal performance. Spark includes Data Streaming, which runs real-time access and analysis of operational data, which is the key to making faster and better business decisions. It also includes Graph, SQL, R and Machine Learning modules. Spark can complement an existing Hadoop environment and leverage Hadoop File System (HDFS). With the ever-increasing volume, velocity, and variety of data, now is the time to move Big Data Analytics to the next level. Spark can capitalize on the opportunity to bridge relational (structured data) and non-relational (unstructured data) data sources to deliver accelerated business and customer insights to drive agility and competitive differentiation.

The Big Data Using Spark Program provides the knowledge to build and deploy applications that utilize the Apache Spark platform with the flexibility of using different programming languages and tools. Students will learn Spark concepts and practices including planning, installation and configuration, load balancing and tuning as well as diagnosing and solving problems in Spark cluster deployment. Students will use example codes and scripts to learn best practices and real-world scenarios. Classroom lectures and demonstrations will be complemented by post-class reading assignments, hands-on exercises, case studies, and projects.

CURRICULUM

BIG DATA OVERVIEW
Review the history of data models and usage, current state, and future trends. Cover Big Data industry and vendor landscape. Discuss Big Data adoption, maturity level, market, and job opportunities.

SPARK ARCHITECTURE AND RUNNING ENVIRONMENT
Introduction to the Apache Software Foundation, the Spark ecosystem, products, tools, user community, and contributors. Review core Spark components. Cover fundamental concepts of parallel and distributed systems.

PLANNING AND DEPLOYING SPARK CLUSTER
General planning considerations including selecting the right hardware, software, storage, and network.

CONFIGURE AND MANAGING SPARK CLUSTER
Spark installation and cluster configuration. Review advanced configuration parameters and configuration management tools.

SPARK ECOSYSTEM AND INTEGRATION TOOLS
Overview of basic Spark RDD and Data Frame concepts. Understand Spark and HDFS architecture. Major features and key components of Spark.

APPLICATION DEVELOPMENT IN SPARK CLUSTER
Overview of HDFS architecture design, features, and security. Understand writing and reading files.

DATA ANALYSIS AND REPORTING IN SPARK
Access, aggregate, and collect assorted data using SQL syntax.
BIG DATA ANALYTICS USING SPARK IN THE CLOUD
Process real-time streams of data using Spark Streaming.

CASE STUDIES AND INDUSTRY
Real-world case studies and best practices.

GENERAL INFORMATION

ADMISSION
Applicants should have some programming experience and understanding of Windows and Linux commands. Basic understanding of Big Data and Hadoop would be a plus. No prior Spark experience is necessary. In addition, students are required to bring their own laptop computers to class.

A substantial commitment of time is required for this intensive course of study. Final admission will be determined by the admissions committee on the basis of an applicant's overall qualifications, including work history and educational background.

FACILITIES
To promote the learning process, the college maintains special-purpose laboratories as well as dedicated classrooms equipped with state-of-the-art audio/visual equipment.

In addition, the college’s Course OnLine (COL) system records classroom lectures that take place on campus, allowing students to replay lectures online. COL captures all classroom elements from audio and video to whiteboard notes and the instructor’s supplemental materials. This system gives students the flexibility to review class content or watch a lecture they may have missed.

SCHEDULE
The Institute offers one section of the program each quarter. Classes meet one day per week. The option to take the program entirely online is also available.

FACULTY
The faculty consists of a team of instructors from the Jarvis College of Computing and Digital Media and experts from industry. Faculty are available throughout the program both in person (for classes that meet on campus) and through e-mail.
The college, through its School of Cinematic Arts, School of Computing, and its School of Design, offers a variety of programs at the undergraduate and graduate levels. More than 3,500 students are enrolled in the college’s bachelor’s programs and more than 2,000 students are enrolled in the master’s and Ph.D. programs making the college’s graduate program one of the largest in the country. The college offers more than 400 courses each quarter, many in the evening, and primarily in the Loop and Lincoln Park Campuses. Most of the degree programs are also available exclusively online.

**Undergraduate programs include:**
- Animation B.A. / B.F.A.
- Computer Science B.S.
- Computing B.A.
- Cyber-Physical Systems Engineering B.S.
- Cybersecurity B.S.
- Data Science B.S.
- Film and Television B.A. / B.F.A.
- Game Design B.S.
- Game Programming B.S.
- Graphic Design B.F.A.
- Industrial Design B.F.A.
- Information Systems B.S.
- Information Technology B.S.
- Math and Computer Science B.S.
- Network Engineering and Security B.S.
- User Experience Design B.S.

**Graduate programs include:**
- Animation M.A. / M.F.A.
- Artificial Intelligence M.S.
- Business Information Technology M.S.
- Computational Finance M.S.
- Computer Science M.S.

**Ph.D. Programs:**
- Computer and Information Sciences
- Human Centered Design

**INSTITUTE FOR PROFESSIONAL DEVELOPMENT**

The Institute for Professional Development was formed by the college in 1984 to assist both individuals and businesses in keeping pace with the rapid development of computer technologies. The Institute currently offers a variety of intensive certificate programs in these areas:

**Career Changers:**
- Career Change Program

**Artificial Intelligence & Machine Learning:**
- Advanced Data Science with Python
- Artificial Intelligence for Enterprise
- Data Science for Business
- Machine Learning and Deep Learning
- Robotic Process Automation

**Big Data Analytics:**
- Big Data and NoSQL
- Big Data Using Spark
- Data Analytics with Excel
- Data Analytics with Excel and Tableau
- Data Visualization with Tableau
- Data Science: Programming with Python
- Fundamentals of R
- Fundamentals of Statistics and Machine Learning Using R

**Cinematic Arts:**
- Web Series Development

**Cloud Computing and DevOps:**
- Cloud Computing Technologies
- DevOps

**Cybersecurity:**
- Cybersecurity Risk Management
- Incident Response and Digital Forensics

**Database Technologies:**
- Advanced SQL
- Introduction to SQL
- SQL Server® Business Intelligence
- SQL Server® Database Administration

**Mobile Application Development:**
- iOS Developer

**Software Development and Software Testing:**
- Advanced Python
- Automated Software Testing
- Modern .NET Web Development
- Web Development with JavaScript and HTML5

**APPLICATION PROCEDURE:**

Prospective students should complete the online application for the program of interest and pay a $40.00 application fee (non-refundable) online during the application process. Alternatively, prospective students may complete the printable application (PDF) and email it to ipd@cdm.depaul.edu, and mail the non-refundable $40.00 application fee (check or money order made payable to DEPAUL UNIVERSITY) to:

DePaul University
Institute for Professional Development
243 S. Wabash Avenue, Room 301
Chicago, IL 60604-2300

The words “Apache Hadoop”, “Apache Hive”, “Apache Pig”, “Apache Spark”, “Amazon Web Services”, “Cloudera”, “SharePoint”, “SQL Server”, “Java” and “Spark” are registered or unregistered trademarks in the United States of America and/or other countries. The Big Data Using SparkProgram at DePaul University is an independent program of study and is not affiliated with, nor has it been authorized, sponsored, or otherwise approved by the Apache Software Foundation or any other external entities.