Intro to Apache Spark Workshop

**Sponsored By:** Clairvoyant LLC

**Abstract**

This workshop introduces participants to Apache Spark, an open source distributed computing framework used to process large volumes of data. We will primarily focus on Core Apache Spark and do a deep dive into the implementation of Resilient Distributed Datasets (RDDs) in the framework. Details will be provided on how to work with the API’s as well as how to develop and execute Spark batch jobs. The session will include examples and exercises to help participants gain hands on knowledge of Spark so that they walk away with a better understanding of how Apache Spark can benefit their organization.

**Objectives**

* Compare MapReduce to Apache Spark
* Introduce the Apache Spark EcoSystem
* Provide an in depth look at Core Apache Spark
  + Implementation
  + API’s
  + Fault Tolerance
  + RDDs
* Give you experience with Apache Spark through examples and exercises

**Pre-requisites**

Knowledge

* Understanding of Hadoop and MapReduce
* Understanding of one of the following Languages:
  + Java
  + Scala
  + Python

Hardware

* Computer that supports Virtualization to allow the installation of a VM
* Minimum of 4 GB of RAM (2GB to be allocated to a VM)

Software

* An IDE (Integrated Developers Environment) that supports your language of choice (Optional)
* Installation of a Virtualization Application. VirtualBox is preferred.

**Schedule**

|  |  |
| --- | --- |
| **TIME** | **TOPIC** |
| 9:00 - 9:30 | Apache Spark History/Background and MapReduce |
| 9:30 - 9:45 | System Setup |
| 9:45 - 10:15 | Apache Spark Context and RDDs |
| 10:15 - 10:45 | Exercise 1 - Running Spark Jobs |
| 10:45 - 11:15 | Apache Spark APIs |
| 11:15 - 11:45 | Exercise 2 - Access Logs |
| 11:45 - 1:00 | Lunch |
| 1:00 - 2:00 | Advanced Apache Spark APIs and Lineage |
| 2:00 - 2:30 | Exercise 3 - Joining Datasets |
| 2:30 - 3:00 | Shared Variables |
| 3:00 - 3:30 | Exercise 4 - Shared Variables |
| 3:30 - 3:45 | Wrap Up |
| 3:45 - 4:15 | Q&A |