# Spark with R (기초반)

## ▣ 교육 목표

*○ 오픈 소스 통계툴인 R을 활용하여 Spark를 컨트롤함으로써 빅데이터 연산을 수행한다.*

## ▣ 교육 대상

*○ 사내 빅데이터 분석 관련 담당자*
*○ 오픈소스 R을 사용하여 Spark를 컨트롤 하고 싶은 분*
*○ Spark를 혼자 공부할 때 어려움을 느끼신 분*

## ▣ 선수지식

*○ R 활용 경험 최소 1년 이상*
*○ R을 사용한 기간이 1년 미만이라도 R을 활용하여 데이터 전처리, 시각화, 가능한 분*

## ▣ 준비사항

*○ 노트북 ((OS : Win10 64bit, RAM : 최소 8GB 이상, HDD : 100GB이상, 인터넷 연결가능)*

## ▣ 교육 내용

| 구분 | 시 간 | 교 육 내 용 | |
|---|---|---|---|
| 1일차 | 09:00 ~ 13:00<br>(4시간) | **R 기초 과정 리뷰**<br>(데이터 전처리 및 시각화 작성) | - 데이터 전처리 리뷰<br>filter, select, arrange, group_by, summarise,<br>bind, join 함수 작성<br>- 데이터 시각화 리뷰<br>분포, 비교, 관계, 중요도를 표현하는 그래프 작성 |
| | 14:00 ~ 18:00<br>(4시간) | **Spark 전 과정 리뷰**<br>(스파크 전체 과정 살펴보기) | - 스파크 설치 및 연결<br>- 데이터 스파크에 올리기<br>- 스파크에서 데이터 전처리 및 시각화 (sparklyr package)<br>- 스파크 모델링 및 예측<br>- 데이터 읽기 및 쓰기 |
| 2일차 | 09:00 ~ 13:00<br>(4시간) | **Spark 전처리와 시각화**<br>(데이터 전처리와 시각화) | - 데이터 전처리<br>sparklyr 함수 vs. spark built-in 함수<br>- 데이터 시각화<br>R로 가져와서 시각화 vs. Spark에서 데이터 시각화 |
| | 14:00 ~ 18:00<br>(4시간) | **Spark 모델링**<br>(회귀 또는 분류모형 작성하기) | - Spark 회귀(또는 분류) 모형 작성 (방법 1)<br>- 데이터 분할, 10겹 교차 검증, 더미 변수화, 모델 평가 등<br>- Spark 모형 작성 (방법 2, pipeline)<br>- 모형 deployment |