BD04 - Big Data dla Data Scientist

Harmonogram: łącznie 16 godz.

LP	Temat szkolenia	Data realizacji szkolenia	Godzina rozpoczęcia	Godzina zakończenia	Ilość godzin
Dzień pierwszy					
1	Wprowadzenie a. Czym jest Big Data b. Cele i historia powstania		9:00	9:45	1
2	Wprowadzenie: c. Typowe zastosowania		9:45	10:30	1
	Przerwa kawowa		10:30	11:00	
3	Apache Hadoop a. Wprowadzenie do platformy Hadoop b. HDFS i. Składowanie danych w HDFS ii. Korzystanie z interfejsu WWW iii. Korzystanie z CLI		11:00	11:45	1
4	Apache Hadoop a. HDFS i. Składowanie danych w HDFS ii. Korzystanie z interfejsu WWW iii. Korzystanie z CLI		11:45	12:30	1
	Przerwa obiadowa		12:30	13:30	
5	Apache Hadoop a. MapReduce i YARN i. Wprowadzenie do paradygmatu MapReduce ii. Architektura klastrów obliczeniowych opartych o YARN iii. Tworzenie i uruchamianie zadań MapReduce iv. Hadoop Streaming		13:30	14:15	1
6	Apache Hadoop a. MapReduce i YARN i. Tworzenie i uruchamianie zadań MapReduce ii. Hadoop Streaming		14:15	15:00	1
	Przerwa kawowa		15:00	15:30	

7	Apache Hive a. Wprowadzenie b. Architektura		15:30	16:15	1
8	Apache Hive a. Tabele zewnętrzne i wewnętrzne b. Przetwarzanie danych za pomocą języka HiveQL		16:15	17:00	1
Dzień drugi					
1	Apache Pig a. Wprowadzenie b. Architektura		9:00	9:45	1
2	Apache Pig a. Typy danych b. Tryby pracy c. Przetwarzanie danych za pomocą języka PigLatin		9:45	10:30	1
	Przerwa kawowa		10:30	11:00	
3	HBase a. Wprowadzenie do baz NoSQL na przykładzie HBase b. Model danych		11:00	11:45	1
4	HBase c. Korzystanie za pomocą CLI d. Dostęp do danych za pomocą Hive i Pig		11:45	12:30	1
	Przerwa obiadowa		12:30	13:30	
5	Spark: Wprowadzenie do rozproszonych kolekcji obiektów Resilient Distributed Datasets (RDDs) i porównanie z Hadoop MapReduce		13:30	14:15	1
6	Spark: Tworzenie i uruchamianie zadań Spark SQL i Spark Streaming Spark MLlib i GraphX		14:15	15:00	1
	Przerwa kawowa		15:00	15:30	
7	Praca ze środowiskiem Big Data: Jupyter		15:30	16:15	1
8	Praca ze środowiskiem Big Data: Zeppelin		16:15	17:00	1