Il programma sviluppato in 60 ore, giornate da 6 ore

Ogni argomento/obiettivo, sarà affrontato con un'alternanza di introduzione dei concetti teorici, esercitazioni pratiche individuali progressive (facile, intermedia, difficile), esercitazionI di progettazione, esercitazionI di gruppo. Tutti i concetti teorici vengono subito messi in pratica e i concetti più complessi vengono ripetuti più volte con diverse spiegazioni e diverse tipologie di esercizi.

Lezione 1

- Big Data intro
- Structured Data
- Unstructured Data
- Features and limitations of the relational model
- From RDBMS to NoSQL
- Partitioning and sharding
- Main categories of NoSQL DBMS
- Key-value storage (Redis)

Lezione 2

- Document oriented storage (MongoDB)
- Columnar / Big Table (Cassandra, Dynamo DB)
- ACID vs. EC: BASE
- CAP theorem
- Dealing with unstructured and semi-structured data
- Differences between the most commonly used formats (avro, prquet, etc.)
- The importance of correct key modeling

Lezione 3

- Big Data Hetics
- ELT vs ETL
- Alternative to ER diagrams for model representation
- Introducing Hackolade: Agile visual data modeling for NoSQL
- Big Data modeling best practices
- Operations vs Analytics processing
- Vertical vs Horizontal Scalability

Lezione 4

- Big Data Modeling Bootcamp (laboratorio pratico di progettazione)
- Hadoop HDFS, Object Sotarge and Data Lake
- Apache Hadoop, Spark Intro, Distributed Computing Introduction
- Apache Spark Bootcamp
- Scala vs PySpark

Lezione 5

- Load data from Data Lake for use in Spark applications
- Write the results back into Data Lake using Spark
- Functional Programming intro
- Imperative vs Descriptive programming and distributed computing
- Map, Reduce programming paradigms
- Resilient Distributed Dataset
- RDD Actions
- RDD Transformation

Lezione 6

- RDD Partitioning
- RDD Persistence
- Apache SparkSQL Dataframe
- PySpark Dataframe API intro
- Spark Dataframe Row object and RDD
- Spark Dataframe and RDD interoperation
- PySpark Dataframe data loading
- PySpark Dataframe persistence and storage

Lezione 7

- PySpark Dataframe storage format: Parquet, ORC, Snappy, CSV
- PySpark Dataframe select and filters
- PySpark Dataframe join
- PySpark Dataframe order
- PySpark Dataframe aggregation

Lezione 8

- PySpark Dataframe window function
- Over
- Partition by
- Ranking function
- Window Frame

Lezione 9

- SQL for Big Data
- Using SQL with Big Data
- SparkSQL intro
- SparkSQL Select, Filter,
- SparkSQL Analytical Functions

Lezione 10

- Narrow trasformation vs Wide Trasformation
- Use metastore tables as an input source or an output sink for Spark applications
- SparkSQL Window Function
- Execution DAG
- Spark-shell
- Spark-submit