

The background is a dark gray collage featuring various logos and mascots associated with big data technologies. At the top left is the Hadoop logo (a red shield with a white 'H'). Next to it is the 'g1c' logo in a bold, lowercase font. To the right is the 'APACHE Spark' logo, with 'APACHE' in small caps and 'Spark' in a large, stylized font. Further right is a colorful geometric logo with a black square containing a white '1'. Below the Hadoop logo is a cartoon bee with the word 'HIVE' written below it. In the center is a cartoon elephant, the mascot for Hadoop. To the right of the elephant is a cartoon pig wearing blue overalls, the mascot for Pig. At the bottom center is a green outline logo consisting of several circles and squares. The text 'Big Data Processing Level-II' is overlaid in the center in a white, sans-serif font.

# Big Data Processing Level-II

# Agenda

2

- ▶ Multinode cluster installation on AWS
- ▶ UDF
- ▶ Various file formats and serde
- ▶ Mini project on Scala and advanced Spark
- ▶ Twitter streaming on Spark
- ▶ End-To-End job scheduling
- ▶ Kafka
- ▶ Cluster management tools

# Fun to learn

- ▶ How to check if the cluster is healthy
- ▶ Disk Usage
- ▶ Quotas
- ▶ Storage Policies
- ▶ How to stop/start/restart the cluster
- ▶ Log analysis and Troubleshooting
- ▶ Namenode stuck in safemode
- ▶ Memory issues
- ▶ Kerberos authentication