Big Data Processing Level-II

Agenda

- 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