Also,

for the labs, I am writing the followings:

Labs:

1. Exercise env set up

2. Intuition for size

3. Intuition for IOPS

4. Hive - Fact based v.s transactional

5. HIve - transactional v.s analytivs data

6. Impala - Data Access and reporting

## Labs:

|  |  |  |  |
| --- | --- | --- | --- |
| **Module** | **Lab Prposal** | **Description/Goal** | **Aparajeeta’s Comment** |
| 1 | Exercise env set up | The goal is to have the student set the software and tools environment that they need for Exercise 1. | This would be for using the AMI and launch a server, identify the software tools needed for Exercises |
| 2 | Intuition for size | Write a simple app in Python that counts distinct elements in a data set. Uses RDBMS and HDFS Adapters. Time the calculation for 1 M, 10 M and 100 M elements. | Will use Postgres and HDFS using Python script  I will have to keep an AMI with pre-loaded data |
| 2 | Intuition for IOPS | Plot IOPS with http://www.thecloudcalculator.com/calculators/disk-raid-and-iops.html | Will instruct to scope for 10TBs of data and in memory processing |
|  |  |  | ?? |
| 4 | Fact based v.s transactional | Have a fact based db and a transactional db. Try to answer a set of questions with both. Is is possible? Difficult? Come up with a question that can not be answered with a fact based db. | They can query against Postgres vs. Hive Semantic Tables  Even in Postgres, Fact based table could be built |
| 5 | transactional, master data v.s. analytics data | Is this transactional data : http://www.census.gov/hhes/families/data/cps2014F.html  Is this transactional data: http://catalog.data.gov/dataset/consumer-complaint-database#topic=consumer\_navigation | Handle Transaction data and build analytics data |
| 6 | Data Access and reporting | Use of Hive - Hive Labs |  |
|  |  |  |  |
|  |  |  |  |
| 10 | Cleaning data | The goal is to clean some data sets with a tool such as Google Refine. |  |