Lab # : 5; Lab Name : Data Profiling and System Capacity Planning ; Subject Name : Information Storage and Retrieval; Week #: 3; Lab Duration : 20 to 30 mins

Intro

In this lab, we will be covering two areas: Data Profiling and System Capacity Planning. Data Profiling skills are used when learning about the data set that you will be building the storage and retrieval system for. You will identify the pros and cons of setting up a storage system in various environments like cloud, on premise, and more. System Capacity planning is an important activity for designing a storage system that identifies the requirements of Hardware, CPU, IO, Memory, and even the number of servers needed for your system.

Let's go!

Step-1.Data Profiling : example 1

For the following table you can create a data profiling hql script or run queries in Hive CLI. So, construct queries to know your dataset in your web_log dataset.

```
CREATE TABLE Web_Session_Log(

DATETIME varchar(500), USERID varchar(500), SESSIONID varchar(500),

PRODUCTID varchar(500), REFERERURL varchar(500))

COMMENT 'This is the Twitter streaming data'

PARTITIONED BY(DT STRING)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t'

STORED AS TEXTFILE;
```

You might have this table from previous lab as well. If not, please create and load data as well using S3 data. Once, the table is ready, for profiling the data or to get to know the data you can run various queries. You can find the minimum values of DATETIME column in this table. Log that.

i.e. `select min(datetime)  from web_session_log;`

Step-2. Data Profiling Example 2

From the same table above, you can find out how many PRODUCTIDs are present in this web log table. This will tell you for which exact products are being searched or browsed. You should log the results.

i.e. `select count(*) from web_session_log;`

Step-3. Data Profiling - Example 3

You can also find out how many users actually were active during these session logs using the above table. Please construct your queries for this and execute.

i.e. `select count(distinct userid) from web_session_log;`

You can further profile the data. This way, you will learn about the data you have received so that you can design your reporting layer better with full insight.

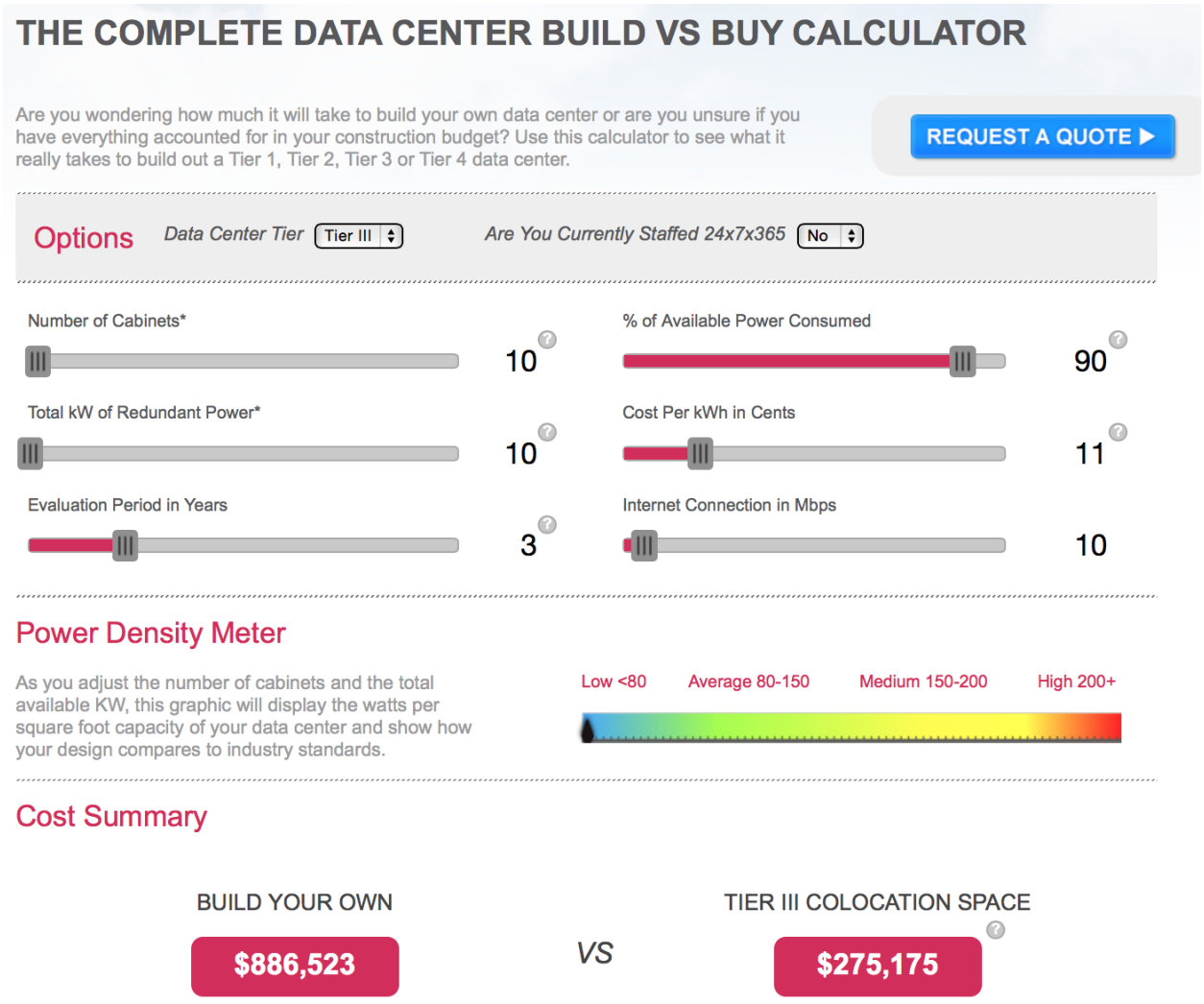
Step-4. Capacity Planning : Data Center : Build vs Buy Calculator

In this step we will use the following website, which will show you how to use different parameters to project capacity requirements.

http://www.thecloudcalculator.com/calculators/disk-raid-and-iops.html

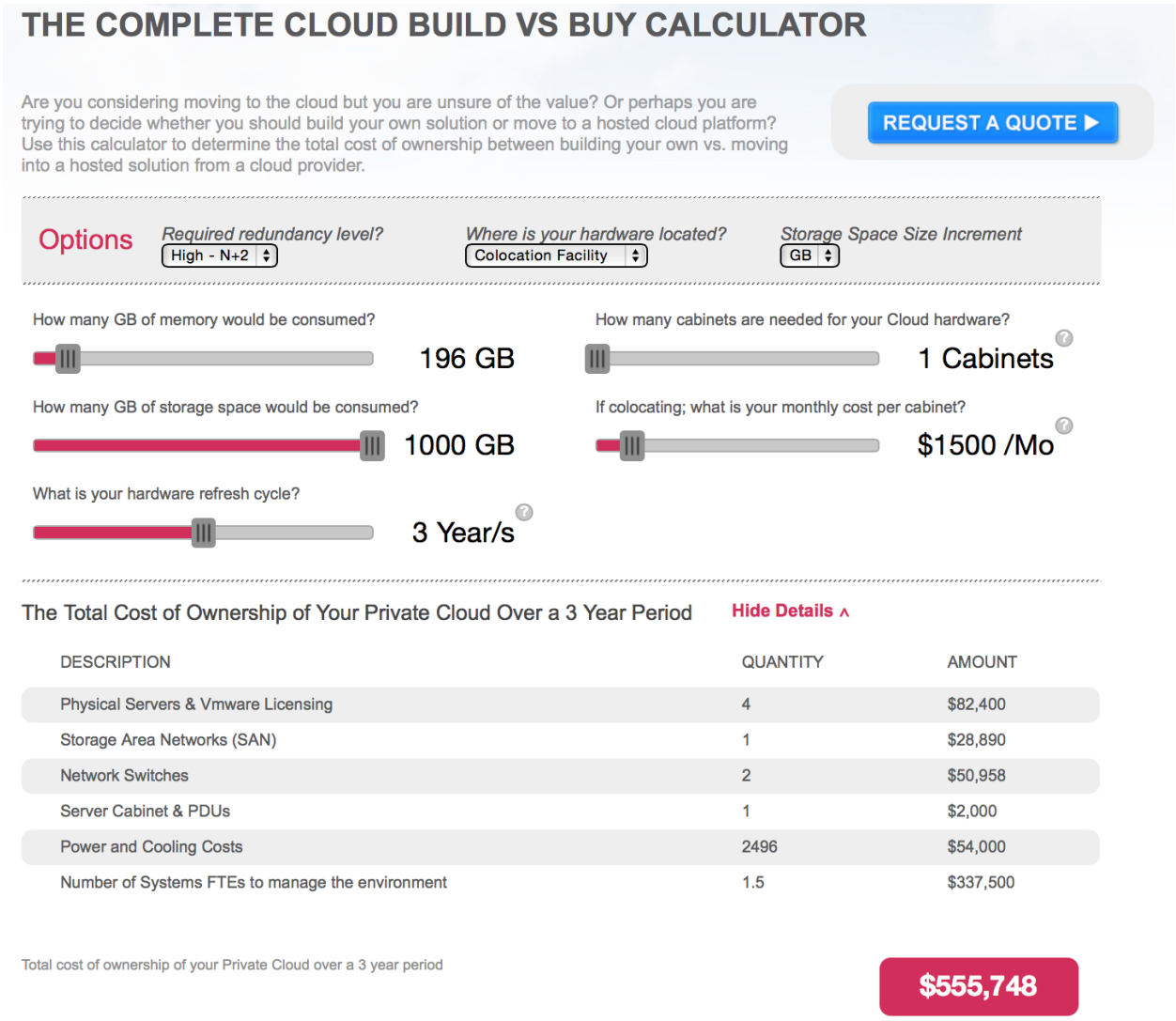
You can use the first tab to determine whether you should build or buy a datacenter.

Adjust the meters below to choose the configuration you need for your datacenter. It will calculate two dollar amounts: one needed for your configuration and one using colocation space instead, so that you could compare the two.



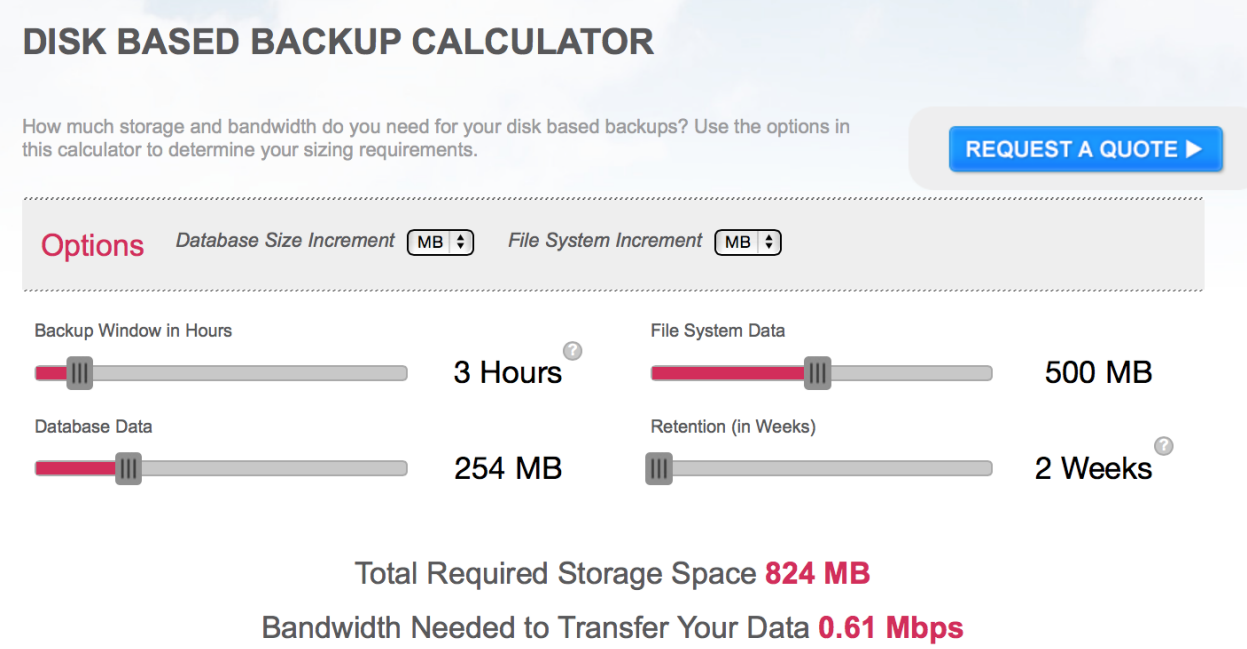
Step-5. Capacity Planning : Cloud Build vs Buy

This is always a big question while hosting your new storage system. Using this tab above, you can estimate the cost by plugging in various parameter values like memory size, storage space, # of cabinets, etc.



Step-6. Capacity Planning : Disk Based Backup Calculator

This tab will give you the estimates on storage and bandwidth needed for your disk based backups.



Here, we have examples only. However, you might have to create your own rules or mathematical formulae for computing and forecasting these various projections.

Questions:

Q1 : Why do you need to do data profiling?
Q2 : What is the difference between data profiling and Data Quality Management?
Q3 : How does Capacity Planning help?
Q4 : For what kind of storage systems would you use cloud based solution?