

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.

THE COMPLETE DATA CENTER BUILD VS BUY CALCULATOR

Are you wondering how much it will take to build your own data center or are you unsure if you have everything accounted for in your construction budget? Use this calculator to see what it really takes to build out a Tier 1, Tier 2, Tier 3 or Tier 4 data center.

[REQUEST A QUOTE ►](#)

Options

Data Center Tier **Tier III** ▼

Are You Currently Staffed 24x7x365 **No** ▼

Number of Cabinets*



10 ?

% of Available Power Consumed



90 ?

Total kW of Redundant Power*



10 ?

Cost Per kWh in Cents



11 ?

Evaluation Period in Years



3 ?

Internet Connection in Mbps



10

Power Density Meter

As you adjust the number of cabinets and the total available KW, this graphic will display the watts per square foot capacity of your data center and show how your design compares to industry standards.

Low <80 Average 80-150 Medium 150-200 High 200+



Cost Summary

BUILD YOUR OWN

\$886,523

VS

TIER III COLOCATION SPACE ?

\$275,175

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.

THE COMPLETE CLOUD BUILD VS BUY CALCULATOR

Are you considering moving to the cloud but you are unsure of the value? Or perhaps you are trying to decide whether you should build your own solution or move to a hosted cloud platform? Use this calculator to determine the total cost of ownership between building your own vs. moving into a hosted solution from a cloud provider.

[REQUEST A QUOTE ►](#)

Options

Required redundancy level?

High - N+2

Where is your hardware located?

Colocation Facility

Storage Space Size Increment

GB

How many GB of memory would be consumed?



196 GB

How many cabinets are needed for your Cloud hardware?



1 Cabinets

How many GB of storage space would be consumed?



1000 GB

If colocating; what is your monthly cost per cabinet?



\$1500 /Mo

What is your hardware refresh cycle?



3 Year/s

The Total Cost of Ownership of Your Private Cloud Over a 3 Year Period

[Hide Details ^](#)

| DESCRIPTION | QUANTITY | AMOUNT |
|--|----------|-----------|
| Physical Servers & VMware Licensing | 4 | \$82,400 |
| Storage Area Networks (SAN) | 1 | \$28,890 |
| Network Switches | 2 | \$50,958 |
| Server Cabinet & PDUs | 1 | \$2,000 |
| Power and Cooling Costs | 2496 | \$54,000 |
| Number of Systems FTEs to manage the environment | 1.5 | \$337,500 |

Total cost of ownership of your Private Cloud over a 3 year period

\$555,748

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.

DISK BASED BACKUP CALCULATOR

How much storage and bandwidth do you need for your disk based backups? Use the options in this calculator to determine your sizing requirements.

[REQUEST A QUOTE ►](#)

Options

Database Size Increment

MB ▾

File System Increment

MB ▾

Backup Window in Hours



3 Hours [?]

File System Data



500 MB

Database Data



254 MB

Retention (in Weeks)



2 Weeks [?]

Total Required Storage Space **824 MB**

Bandwidth Needed to Transfer Your Data **0.61 Mbps**

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?