**CAPSTONE PROJECT 1**

**1)Sales Data analysis using S3, EMRHive, tableau**

**Analyze and generate dashboards using the insights from the provided sales data using AWS S3, EMRHive and Tableau. Some important insights could be but not limited to Units sold by country, Total revenue and cost per country etc.**

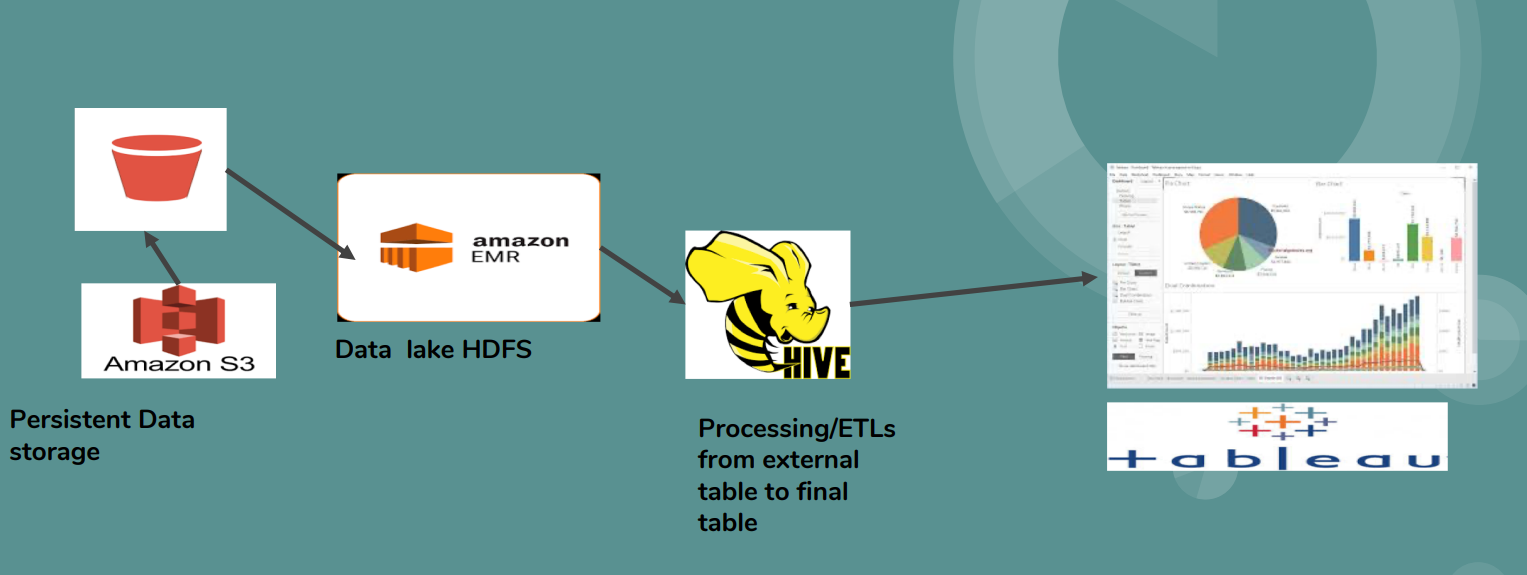
**solution:**

There are 3 stages required in given project

1)data analysis using aws s3 bucket.

2)data load into EMR HIVE .

3)data load into tableau and generate a dashboards.



**stage1:**

--> login to AWS account and open Amazon s3 created new bucket "capstoneprjt".

--> I created new folder name"csv".

--> in that csv folder i uploaded given database"sales\_data.csv".

--> by selecting the sales\_data we can able to see how data is available in that given csv file.

select sales\_data-->actions-->querry-->run the querry.

successfully uploaded data into s3 bucket.

**stage2:**

-->we have to create EMR cluster to connect EMR HIVE.

creation of cluster:

--> opened EMR and created new cluster name"capstone1".i choosed default values emr 5.36.0,m5.large,i kept auto termination 4hours.last we have to give keypair. so open EC2

select keypairs-->create keypairs-->'saleskeypair'-->.pem(here automatically .pem file will be download into our pc)-->open puttygen-->load our .pem file-->saveprivatekey-->.ppk file save.

-->choose the keypair name into creating cluster and hit on create cluster.

successfully EMR cluster created.

-->choose master link and edit inbounded rules(SSH& MYIP)save.

-->copy master node link

-->open putty-->paste the master node link-->add-->auth-->choose .ppk file -->open.

by accepting the window it will directly connect to EMRHIVE.

-->hive-->show databases;(default databse will shown)

-->create databse salesdb;

-->use salesdb;

**Creating external table:**

-->create external table if not exists sales\_table

(

region string,

country string,

item\_type string,

sales\_channel string,

order\_priority string,

order\_date string,

order\_id string,

ship\_date string,

units\_sold string,

unit\_price string,

unit\_cost string,

total\_revenue string ,

total\_cost string,

total\_profit string)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LOCATION "s3://capstoneprjt/csv/";

-->show tables;

-->select \* from sales\_table limit 20;

-->select count(\*) from sales\_table;

-->describe extended sales\_table;

-->select distinct item\_type from sales\_table;

-->select item\_type,count(distinct order\_id)total\_orders from sales\_table group by item-type;

-->set hive.cli.print.header=true;

-->select item\_type,count(distinct order\_id)total\_orders from sales\_table group by item-type having item\_type<>'item\_type" order by total\_orders;

-->select order\_id,item\_type,order\_priority from sales\_table limit 10;

-->select order\_id,item\_type,item\_priority,count(distinct order\_id)total\_orders from sales\_table group by item-type,order\_priority having item\_type<>'item\_type" and order\_priority='H' and total\_priority>20 order by toatl\_orders dsc;

-->select distinct sales\_channel from sales\_table;

**Creating target table:**

create table if not exists target\_table

(

region string,

item\_type string,

sales\_channel string,

order\_priority string,

order\_date string,

order\_id string,

ship\_date string,

units\_sold string,

unit\_price string,

unit\_cost string,

total\_revenue string,

total\_cost string,

total\_profit string)

partitioned by (country string)

stored as ORC;

**To load data into target table:**

SET hive.exec.compress.intermediate=True;

SET hive.exec.dynamic.partition=True;

Set exec.dynamic.partition.mode=nonstrict;

SET hive.mapred.mode=nonstrict;

insert overwrite table target\_table partition(country)

select

cast(region as string),

cast(item\_type as string),

cast(sales\_channel as string),

cast(order\_priority as string),

coalesce(

cast(from\_unixtime(unix\_timestamp(order\_date,'MM/dd/yyyy'),"yyyy")as string),

cast(from\_unixtime(unix\_timestamp(order\_date,'dd/MM/yyyy'),"yyyy")as string)

)as order\_date,

cast(order\_id as int),

coalesce(

cast(from\_unixtime(unix\_timestamp(ship\_date,'MM/dd/yyyy'),"yyyy")as string),

cast(from\_unixtime(unix\_timestamp(ship\_date,'dd/MM/yyyy'),"yyyy")as string)

)as ship\_date,

cast(units\_sold as int),

cast(unit\_price as float),

cast(unit\_cost as float),

cast(total\_revenue as float),

cast(total\_cost as float),

cast(total\_profit as float),

cast(country as string)

from sales\_table;

**To see the loaded data in EMR HIVE:**

quit;

hdfs dfs -ls /user/hive/warehouse/salesdb/

hdfs dfs -ls /user/hive/warehouse/salesdb/target\_table/

hsdf dfs -ls /user/hive//warehouse/salesdb/target\_table/country=8809999934/

hive

use salesdb;

describe extended target\_table;

select \* from target\_table where country='iceland';

**stage3:**

**connecting to tableau and generating dashboards to easy understand the given data in the form of visuals(graphs):**

-->downloded tableau deskktop latest version

-->in tableau -->Data-->new data source-->EMRHadoop Hive connector(if it is not visible then we have to download this connector)-->general-->paste the link(master node link)

-->give the user name-->hive-->next-->sign in

here i am unable connect to tableau .i am getting error .

Data Analysis:

there are 7 distinct Regions available

1.'Asia'-----27 countries

2.'Australia and Oceania'-------15 countries

3.'Central America and the Caribbean' ----20 countries

4.'Europe'------------48 countries

5.'Middle East and North Africa'----- 23 countries

6.'North America'-------------- 4 countries

7.'Sub-Saharan Africa'-----48 countries

NORTH AMERICA:

1.sales\_channel: online=11,offline=8.

2.maximum profit=15,41,620.46.

greenland--household(item)--online(mode)--9302(units)

'North America', 'Greenland', 'Household', 'Online', 'L', '4/17/2014', '192721068', '5/20/2014', '9302', '668.27', '502.54', '6216247.54', '4674627.08', '1541620.46'

3.minimum profit=11,348.69

greenland--fruits--offline--9.33(units)

'North America', 'Greenland', 'Fruits', 'Offline', 'M', '1/31/2012', '469839179', '2/22/2012', '4709', '9.33', '6.92', '43934.97', '32586.28', '11348.69'

4. there are 10 item\_types:

'Vegetables', 'Beverages', 'Fruits'

'Office Supplies', 'Clothes'

'Baby Food', 'Household'

'Personal Care', Meat', 'Snacks'

5.total item types are 12