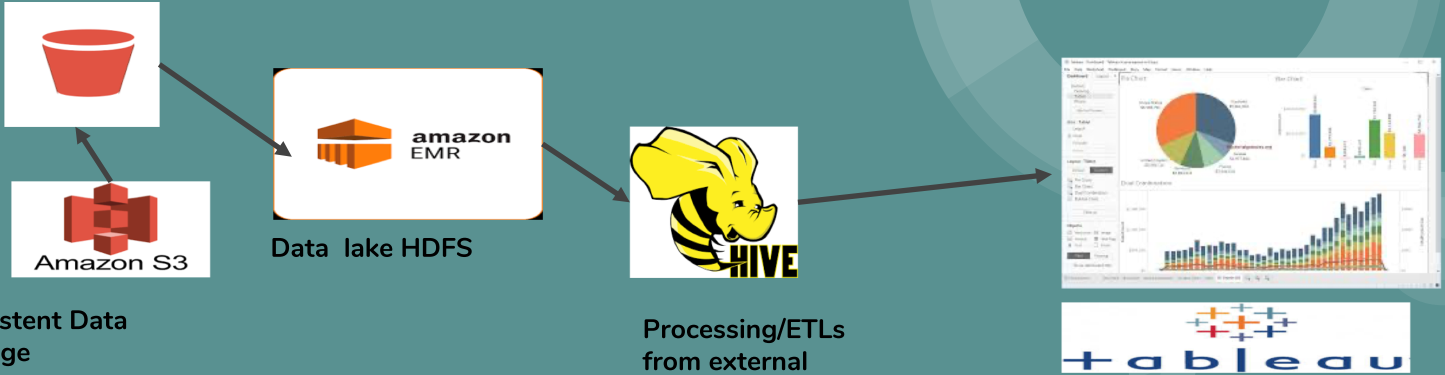
Sales Data analysis using S3,EMRHive,tableau

Data architecture SAAS



Data lake HDFS

Persistent Data storage

Processing/ETLs from external table to final table

Data flow design

1. Create an S3 bucket in AWS
2. Upload sales data into the S3
3. Spin an EMR cluster on AWS which has required services (1 master node 2 core nodes m5.xlarge)
4. Create Hive external table to point to the data in S3
5. Perform ETLs on Hive table and store it in final Hive table
6. Connect Hive final table in AWS EMR to tableau in local and plot the graphs
7. Documentation : https://docs.google.com/document/d/1HFM-TdZo2zVJAEoXx834QYkhRaz6Wx8k86PTR3RQMi0/edit?usp=sharing

S3 and its features

- S3 offers total four class storage solutions, with unlimited data storage capacity.
- S3 Standard offers high durability, availability, and performance object storage for frequently accessed data.
- S3 has a Reduced Redundancy Storage feature
- We store any data into S3, we need to create a Bucket , Amazon S3 creates buckets in a region we specify we can choose any AWS Region that is geographically close to our requirement.
- Amazon S3 also supports features that help maintain data version control, prevent accidental deletions, and replicate data to the same or different AWS Region.
- Static Web Hosting is one of the most powerful features of the AWS S3.
- Amazon S3 offers flexible security features to block unauthorized users from accessing your data. Use VPC endpoints to connect to S3 resources from your Amazon Virtual Private Cloud (Amazon VPC)

EMR and its features

1. Easy to use
2. Low cost
3. Elasticity
4. Reliability
5. Security
6. Flexibility
7. Monitoring

EMR vs PAAS hadoop solution on AWS

1. Auto scaling
2. Dynamic orchestration
3. Access to Amazon S3
4. High availability
5. Ease of use
6. Hadoop management console
7. On-premise and cloud options
8. Debugging features
9. Cost

Hive vs Impala

Apache Hive might not be ideal for interactive computing whereas Impala is meant for interactive computing. Hive is batch based Hadoop MapReduce whereas Impala is more like MPP database. Hive supports complex types but Impala does not. Apache Hive is fault tolerant whereas Impala does not support fault tolerance

Query Execution :

Hive : The output of the query will be produced as Hive is fault tolerant, while a data node goes down during the query execution.

Impala : Impala starts all over again, while a data node goes down during the query execution.

Performance:

Hive: while keeping Hive’s ability to perform well at mid to high query complexity, Hive LLAP gets good performance at the low end.

Impala: Similarly, while Impala struggles as query complexity increases but Impala perform well with less complex queries.

During the Runtime:

Hive : At Compile time, Hive generates query expressions.

Impala : During the Runtime, Impala generates code for “big loops”.

Time consumption

Hive: The dynamic runtime features of Hive LLAP minimizes the overall work. Hence, we can say working with Hive LLAP consumes less time.

Impala: Impala consumes less time for simpler queries, but for complex queries, it needs more time than Hive LLAP.

Tableau connector to EMR

Before you can build a view and analyze your data, you must first connect Tableau to your data. Tableau supports connecting to a wide variety of data, stored in a variety of places. For example, your data might be stored on your computer in a spreadsheet or a text file, or in a big data, relational, or cube (multidimensional) database on a server in your enterprise. Or, you might connect to public domain data available on the web such as U.S. Census Bureau information, or to a cloud database source, such as Google Analytics, Amazon Redshift, or Salesforce.

https://help.tableau.com/current/pro/desktop/en-us/basicconnectoverview.htm
https://help.tableau.com/current/pro/desktop/en-us/exampleconnections_overview.htm

Connecting Tableau to AWS EMR

1. Install the ODBC driver on your machine with Tableau Desktop, required for connecting Tableau Desktop to HIVE on Amazon EMR.
2. Modify the Amazon EMR cluster Master Security Group so Tableau can connect with the AmazonEMRHadoopHive server running on the master node of the Amazon EMR cluster.
3. Follow the steps as directed by Tableau to enable Amazon EMR Hadoop Hive as a data connection option in Tableau.
4. Got to Add a connection -> Click on more-> Select Amazon EMR Hadoop Hive and a pop up window appears to enter details .
5. Give your server DNS , port number as 10000
6. Choose Authentication as Username from the dropdown
7. Specify Username as hive
8. Check the RequireSSL box and click on Sign in

Tableau dashboard best practices

- Know your purpose and audience
- Leverage the most-viewed spot
- Design for the real world
- Limit the number of views
- Be security-savvy
- Add interactivity to encourage exploration
- Show filters
- Enable highlighting