Cloudera-Data Usage Analysis & Footprint

**Definitions of Hot/Cold/Frozen:**

Hot & Cold Data – Here is an example of what is being followed in Teradata (ICW), Definition of Cloudera will be finalized with 1-2 weeks of program kick off.

Hot – Data accessed within 6 months

Cold – between 6 to 12 months (Archive Data - to have provision for restore on demand)

Frozen – Data accessed beyond 1 yr , rarely used .(Data to be purged)

Oracle Source:

JPMC, Consumer & Community banking (CCB) has around 10 LOB’s.

1. Auto/Student Loan
2. CBB
3. CBB OPS
4. CTO
5. Enterprise Tech
6. GTI
7. IB
8. IRC
9. Mortgage Bank
10. Risk Management

If we look at Hot & Cold data analysis, the definition might vary from LOB to LOB.

Since the Data usage analysis is at CDH level, we shall maintain one common definition for Data usage analysis.

E.g.

1-6 months – Hot

6-12 Months – Cold

More than 12 months - Frozen

**Objective:**

**Hot/Cold (Purge, Archive) Data Analysis –**

* Read the data from hive tables for hot/cold analysis.
* Read the data from files (text, parquet, Avro) if required
* Apply transformations to segregate Hot/cold analysis data.
* Write a query to identify duplicate tables, columns, rows etc. and remove the same.
* Write a scripts to get final result into new hive external tables.
* Automate the entire process

