

**Cloudera and Greenplum (IS) - Data Usage Analysis and Data Archival**

Document ID: Cloudera and Greenplum (IS) Data Usage analysis and Data Archival

Document Revision: 1.0

Issue Date:

Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| Revision | Author Name | Date | Description |
| Draft | Azrar Ahmed | 10-April-2019 | Data Usage and Data Archival Team |

TABLE OF CONTENTS

[1 Introduction 4](#_Toc5803627)

[1.1 Purpose 4](#_Toc5803628)

[1.2 Scope 4](#_Toc5803629)

[1.3 Objectives 4](#_Toc5803630)

# 1 Introduction

## Purpose

The purpose of this document is to provide an insight for **Cloudera and Greenplum data usage and Data Archival.** This document provides the specifications from a business user and a technical/functional perspective.

## Scope

This document applies specifically to **Cloudera and Greenplum Data usage and DataArchival** process that we are planning to apporach, will be used to guide the technical team’s activities.

## Objectives

The hot (1-6 months) and cold(6-12 months) tables are classified by the Data usage team and are sent to the Data Archival team for extracting the cold tables and archiving it into the AWS Neptune.

The Steps to achive this are as follows:

1. The Data usage team provides the archival team with the list of Hot, Cold and frozen tables tables.
2. The list of cold tables to be archived and the frozen tables are to be purged.
3. The Identified Cold and frozen tables are to be extracted from HIVE via python scripts.
4. Extract the data from different types of files like parquet, Avro etc that is to be archived to Neptune.
5. The cold tables that has been extracted from the HIVE are to be archived to the AWS Neptune via the python scripts.
6. Apply necessary transformations to eliminate redundancy and duplication.
7. Automate the entire process into an ever green process

**Note:** Required python libraries and packages are to be installed.

S3 Storage (AWS Neptune)

Python scripts

HDFS

Archiving Files

Extracting Files

**Hive**

**File1**

**File2**

**.**

**.**

**.**

Purged

Via Python scripts Purging Files

The files (Archival or purging) are extracted from HDFS (HIVE) using python. Post extraction the files are supposed to be archived will be pushed to the S3 storage (Neptune) via python scripts.

The purged or duplicate files will be deleted and not moved into the S3 storage(Neptune)