**HW 05(a) DWBI**

**SHIVAM VERMA**

Ques1) The COVID-19 Nursing Home dataset consists of data reported by the nursing homes to the Centers for Disease Control and Prevention (CDC) or, more specifically, the National Healthcare Safety Network (NHSN) system. The dataset contains about 322k rows and 91 columns. It can be downloaded using various file formats, including the popular CSV and TSV file formats. I will be using CSV formatted file to load the data into the SQL Server.

Before loading data in SQL Server, we must create a SQL Server table to load our data. The table should have metadata and datatypes like the source file. After creating the table in SQL, we can load the data using multiple ways. The two ways I will be loading the data is by using SSIS and BCP.

To load the data using SSIS first, we must create a data integration package in Visual Studio. We must complete a data flow task in our control flow. The data flow task mainly consists of two operations. First, to load the flat file data, CSV file in our case. Second, ingest the data into the table created in SQL by creating an OLE DB connection. We will select our CSV files in our flat-file connection and check whether all our columns are loaded correctly. After our file is loaded successfully, we will connect it with our table in the database and review it for the columns' mappings. We might need to edit the mappings as per the requirements. We also need to check for the columns data type to be loaded without errors. Finally, we run the package and ensure the data flow is executed successfully.

To load the data using BCP, we have to open the command prompt and use the following **"bcp "destination table" in "source file" -S "server name" -d "database name" -T"** BCP command to create a format file. After making the format file, we can use that format file to set our data types, field terminators, and row terminators as required. Then we can use the following **"bcp "destination table" in "source file" -S "server name" -d "database name" -T -f "format file""** command to load our data.

Ques2) Load the data into SQL server using SSIS

Create Table Script:

CREATE TABLE [dbo].[COVID19NursingHome](

[WeekEnding] [varchar](255) NULL,

[FederalProviderNumber] [varchar](255) NULL,

[ProviderName] [varchar](255) NULL,

[ProviderAddress] [varchar](255) NULL,

[ProviderCity] [varchar](255) NULL,

[ProviderState] [varchar](255) NULL,

[ProviderZipCode] [varchar](255) NULL,

[SubmittedData] [varchar](255) NULL,

[PassedQualityAssuranceCheck] [varchar](255) NULL,

[ResidentsWeeklyAdmissionsCOVID19] [varchar](255) NULL,

[ResidentsTotalAdmissionsCOVID19] [varchar](255) NULL,

[ResidentsWeeklyConfirmedCOVID19] [varchar](255) NULL,

[ResidentsTotalConfirmedCOVID19] [varchar](255) NULL,

[ResidentsWeeklySuspectedCOVID19] [varchar](255) NULL,

[ResidentsTotalSuspectedCOVID19] [varchar](255) NULL,

[ResidentsWeeklyAllDeaths] [varchar](255) NULL,

[ResidentsTotalAllDeaths] [varchar](255) NULL,

[ResidentsWeeklyCOVID19Deaths] [varchar](255) NULL,

[ResidentsTotalCOVID19Deaths] [varchar](255) NULL,

[NumberOfAllBeds] [varchar](255) NULL,

[TotalNumberOfOccupiedBeds] [varchar](255) NULL,

[ResidentAccessToTestingInFacility] [varchar](255) NULL,

[LaboratoryTypeIsStateHealthDept] [varchar](255) NULL,

[LaboratoryTypeIsPrivateLab] [varchar](255) NULL,

[LaboratoryTypeIsOther] [varchar](255) NULL,

[AbletoTestOrObtainResourcestoTestAllCurrentResidentsWithinNext7Days] [varchar](255) NULL,

[ReasonForNotTestingResidents\_LackOfPPEForPersonnel] [varchar](255) NULL,

[ReasonForNotTestingResidents\_LackOfSupplies] [varchar](255) NULL,

[ReasonForNotTestingResidents\_LackOfAccessToLaboratory] [varchar](255) NULL,

[ReasonForNotTestingResidents\_LackOfAccessToTrainedPersonnel] [varchar](255) NULL,

[ReasonForNotTestingResidents\_UncertaintyAboutReimbursement] [varchar](255) NULL,

[ReasonForNotTestingResidents\_Other] [varchar](255) NULL,

[DuringPastTwoWeeksAverageTimeToReceiveResidentTestResults] [varchar](255) NULL,

[HasFacilityPerformedResidentTestsSinceLastReport] [varchar](255) NULL,

[TestedResidentsWithNewSignsOrSymptoms] [varchar](255) NULL,

[TestedAsymptomaticResidentsInAUnitOrSectionAfterANewCase] [varchar](255) NULL,

[TestedAsymptomaticResidentsFacilityWideAfterANewCase] [varchar](255) NULL,

[TestedAsymptomaticResidentsWithoutKnownExposureAsSurveillance] [varchar](255) NULL,

[TestedAnotherSubgroupOfResidents] [varchar](255) NULL,

[AbleToTestOrObtainResourcesToTestAllStaffAndOrPersonnelWithinNext7Days] [varchar](255) NULL,

[ReasonforNotTestingStaffAndOrPersonnel\_LackOfPPEForPersonnel] [varchar](255) NULL,

[ReasonforNotTestingStaffAndOrPersonnel\_LackOfSupplies] [varchar](255) NULL,

[ReasonforNotTestingStaffAndOrPersonnel\_LackOfAccessToLaboratory] [varchar](255) NULL,

[ReasonforNotTestingStaffAndOrPersonnel\_LackOfAccessToTrainedPersonnel] [varchar](255) NULL,

[ReasonforNotTestingStaffAndOrPersonnel\_UncertaintyAboutReimbursement] [varchar](255) NULL,

[ReasonforNotTestingStaffAndOrPersonnel\_Other] [varchar](255) NULL,

[DuringPastTwoWeeksAverageTimetoReceiveStaffAndOrPersonnelTestResults] [varchar](255) NULL,

[HasFacilityPerformedStaffAndOrPersonnelTestsSinceLastReport] [varchar](255) NULL,

[TestedStaffAndOrPersonnelWithNewSignsOrSymptoms] [varchar](255) NULL,

[TestedAsymptomaticStaffAndOrPersonnelInAUnitOrSectionAfterANewCase] [varchar](255) NULL,

[TestedAsymptomaticStaffAndOrPersonnelFacility\_WideAfterANewCase] [varchar](255) NULL,

[TestedAsymptomaticStaffAndOrPersonnelWithoutKnownExposureAsSurveillance] [varchar](255) NULL,

[TestedAnotherSubgroupOfStaffAndOrPersonnel] [varchar](255) NULL,

[InHousePointOfCareTestMachine] [varchar](255) NULL,

[COVID19PointOfCareTestsPerformedOnResidentsSinceLastReport] [varchar](255) NULL,

[COVID19PointOfCareTestsPerformedOnStaffAndOrPersonnelSinceLastReport] [varchar](255) NULL,

[EnoughSuppliesToTestAllStaffAndOrPersonnelUsingPointOfCareTestMachine] [varchar](255) NULL,

[StaffWeeklyConfirmedCOVID19] [varchar](255) NULL,

[StaffTotalConfirmedCOVID19] [varchar](255) NULL,

[StaffWeeklySuspectedCOVID19] [varchar](255) NULL,

[StaffTotalSuspectedCOVID19] [varchar](255) NULL,

[StaffWeeklyCOVID19Deaths] [varchar](255) NULL,

[StaffTotalCOVID19Deaths] [varchar](255) NULL,

[ShortageOfNursingStaff] [varchar](255) NULL,

[ShortageOfClinicalStaff] [varchar](255) NULL,

[ShortageOfAides] [varchar](255) NULL,

[ShortageOfOtherStaff] [varchar](255) NULL,

[AnyCurrentSupplyOfN95Masks] [varchar](255) NULL,

[OneWeekSupplyOfN95Masks] [varchar](255) NULL,

[AnyCurrentSupplyOfSurgicalMasks] [varchar](255) NULL,

[OneWeekSupplyOfSurgicalMasks] [varchar](255) NULL,

[AnyCurrentSupplyOfEyeProtection] [varchar](255) NULL,

[OneWeekSupplyOfEyeProtection] [varchar](255) NULL,

[AnyCurrentSupplyOfGowns] [varchar](255) NULL,

[OneWeekSupplyOfGowns] [varchar](255) NULL,

[AnyCurrentSupplyOfGloves] [varchar](255) NULL,

[OneWeekSupplyOfGloves] [varchar](255) NULL,

[AnyCurrentSupplyOfHandSanitizer] [varchar](255) NULL,

[OneWeekSupplyOfHandSanitizer] [varchar](255) NULL,

[VentilatorDependentUnit] [varchar](255) NULL,

[NumberOfVentilatorsInFacility] [varchar](255) NULL,

[NumberOfVentilatorsInUseForCOVID19] [varchar](255) NULL,

[AnyCurrentSupplyOfVentilatorSupplies] [varchar](255) NULL,

[OneWeekSupplyOfVentilatorSupplies] [varchar](255) NULL,

[TotalResidentConfirmedCOVID19CasesPer1000Residents] [varchar](255) NULL,

[TotalResidentCOVID19DeathsPer1000Residents] [varchar](255) NULL,

[TotalResidentsCOVID19DeathsAsAPercentageOfConfirmedCOVID19Cases] [varchar](255) NULL,

[County] [varchar](255) NULL,

[ThreeOrMoreConfirmedCOVID19CasesThisWeek] [varchar](255) NULL,

[InitialConfirmedCOVID19CaseThisWeek] [varchar](255) NULL,

[Geolocation] [varchar](255) NULL

) ON [PRIMARY]

GO

Screenshot of Create Table Command in SQL Server Management Studio:

Graphical user interface, application

Description automatically generated

Screenshot of Adding Flat file as source:

![Graphical user interface

Description automatically generated]()

Screenshot of adding OLE DB as destination:

![Graphical user interface, application

Description automatically generated]()

Screenshot to show columns are mapped accurately:

![Graphical user interface, table

Description automatically generated]()

Screenshot of Package executed successfully:

A screenshot of a computer screen

Description automatically generated

Screenshot of data populated into table using SSIS:

Graphical user interface, application, table

Description automatically generated

Ques3) Load the data using BCP

Create Table command:

CREATE TABLE [dbo].[COVID19NursingHome](

[WeekEnding] [varchar](255) NULL,

[FederalProviderNumber] [varchar](255) NULL,

[ProviderName] [varchar](255) NULL,

[ProviderAddress] [varchar](255) NULL,

[ProviderCity] [varchar](255) NULL,

[ProviderState] [varchar](255) NULL,

[ProviderZipCode] [varchar](255) NULL,

[SubmittedData] [varchar](255) NULL,

[PassedQualityAssuranceCheck] [varchar](255) NULL,

[ResidentsWeeklyAdmissionsCOVID19] [varchar](255) NULL,

[ResidentsTotalAdmissionsCOVID19] [varchar](255) NULL,

[ResidentsWeeklyConfirmedCOVID19] [varchar](255) NULL,

[ResidentsTotalConfirmedCOVID19] [varchar](255) NULL,

[ResidentsWeeklySuspectedCOVID19] [varchar](255) NULL,

[ResidentsTotalSuspectedCOVID19] [varchar](255) NULL,

[ResidentsWeeklyAllDeaths] [varchar](255) NULL,

[ResidentsTotalAllDeaths] [varchar](255) NULL,

[ResidentsWeeklyCOVID19Deaths] [varchar](255) NULL,

[ResidentsTotalCOVID19Deaths] [varchar](255) NULL,

[NumberOfAllBeds] [varchar](255) NULL,

[TotalNumberOfOccupiedBeds] [varchar](255) NULL,

[ResidentAccessToTestingInFacility] [varchar](255) NULL,

[LaboratoryTypeIsStateHealthDept] [varchar](255) NULL,

[LaboratoryTypeIsPrivateLab] [varchar](255) NULL,

[LaboratoryTypeIsOther] [varchar](255) NULL,

[AbletoTestOrObtainResourcestoTestAllCurrentResidentsWithinNext7Days] [varchar](255) NULL,

[ReasonForNotTestingResidents\_LackOfPPEForPersonnel] [varchar](255) NULL,

[ReasonForNotTestingResidents\_LackOfSupplies] [varchar](255) NULL,

[ReasonForNotTestingResidents\_LackOfAccessToLaboratory] [varchar](255) NULL,

[ReasonForNotTestingResidents\_LackOfAccessToTrainedPersonnel] [varchar](255) NULL,

[ReasonForNotTestingResidents\_UncertaintyAboutReimbursement] [varchar](255) NULL,

[ReasonForNotTestingResidents\_Other] [varchar](255) NULL,

[DuringPastTwoWeeksAverageTimeToReceiveResidentTestResults] [varchar](255) NULL,

[HasFacilityPerformedResidentTestsSinceLastReport] [varchar](255) NULL,

[TestedResidentsWithNewSignsOrSymptoms] [varchar](255) NULL,

[TestedAsymptomaticResidentsInAUnitOrSectionAfterANewCase] [varchar](255) NULL,

[TestedAsymptomaticResidentsFacilityWideAfterANewCase] [varchar](255) NULL,

[TestedAsymptomaticResidentsWithoutKnownExposureAsSurveillance] [varchar](255) NULL,

[TestedAnotherSubgroupOfResidents] [varchar](255) NULL,

[AbleToTestOrObtainResourcesToTestAllStaffAndOrPersonnelWithinNext7Days] [varchar](255) NULL,

[ReasonforNotTestingStaffAndOrPersonnel\_LackOfPPEForPersonnel] [varchar](255) NULL,

[ReasonforNotTestingStaffAndOrPersonnel\_LackOfSupplies] [varchar](255) NULL,

[ReasonforNotTestingStaffAndOrPersonnel\_LackOfAccessToLaboratory] [varchar](255) NULL,

[ReasonforNotTestingStaffAndOrPersonnel\_LackOfAccessToTrainedPersonnel] [varchar](255) NULL,

[ReasonforNotTestingStaffAndOrPersonnel\_UncertaintyAboutReimbursement] [varchar](255) NULL,

[ReasonforNotTestingStaffAndOrPersonnel\_Other] [varchar](255) NULL,

[DuringPastTwoWeeksAverageTimetoReceiveStaffAndOrPersonnelTestResults] [varchar](255) NULL,

[HasFacilityPerformedStaffAndOrPersonnelTestsSinceLastReport] [varchar](255) NULL,

[TestedStaffAndOrPersonnelWithNewSignsOrSymptoms] [varchar](255) NULL,

[TestedAsymptomaticStaffAndOrPersonnelInAUnitOrSectionAfterANewCase] [varchar](255) NULL,

[TestedAsymptomaticStaffAndOrPersonnelFacility\_WideAfterANewCase] [varchar](255) NULL,

[TestedAsymptomaticStaffAndOrPersonnelWithoutKnownExposureAsSurveillance] [varchar](255) NULL,

[TestedAnotherSubgroupOfStaffAndOrPersonnel] [varchar](255) NULL,

[InHousePointOfCareTestMachine] [varchar](255) NULL,

[COVID19PointOfCareTestsPerformedOnResidentsSinceLastReport] [varchar](255) NULL,

[COVID19PointOfCareTestsPerformedOnStaffAndOrPersonnelSinceLastReport] [varchar](255) NULL,

[EnoughSuppliesToTestAllStaffAndOrPersonnelUsingPointOfCareTestMachine] [varchar](255) NULL,

[StaffWeeklyConfirmedCOVID19] [varchar](255) NULL,

[StaffTotalConfirmedCOVID19] [varchar](255) NULL,

[StaffWeeklySuspectedCOVID19] [varchar](255) NULL,

[StaffTotalSuspectedCOVID19] [varchar](255) NULL,

[StaffWeeklyCOVID19Deaths] [varchar](255) NULL,

[StaffTotalCOVID19Deaths] [varchar](255) NULL,

[ShortageOfNursingStaff] [varchar](255) NULL,

[ShortageOfClinicalStaff] [varchar](255) NULL,

[ShortageOfAides] [varchar](255) NULL,

[ShortageOfOtherStaff] [varchar](255) NULL,

[AnyCurrentSupplyOfN95Masks] [varchar](255) NULL,

[OneWeekSupplyOfN95Masks] [varchar](255) NULL,

[AnyCurrentSupplyOfSurgicalMasks] [varchar](255) NULL,

[OneWeekSupplyOfSurgicalMasks] [varchar](255) NULL,

[AnyCurrentSupplyOfEyeProtection] [varchar](255) NULL,

[OneWeekSupplyOfEyeProtection] [varchar](255) NULL,

[AnyCurrentSupplyOfGowns] [varchar](255) NULL,

[OneWeekSupplyOfGowns] [varchar](255) NULL,

[AnyCurrentSupplyOfGloves] [varchar](255) NULL,

[OneWeekSupplyOfGloves] [varchar](255) NULL,

[AnyCurrentSupplyOfHandSanitizer] [varchar](255) NULL,

[OneWeekSupplyOfHandSanitizer] [varchar](255) NULL,

[VentilatorDependentUnit] [varchar](255) NULL,

[NumberOfVentilatorsInFacility] [varchar](255) NULL,

[NumberOfVentilatorsInUseForCOVID19] [varchar](255) NULL,

[AnyCurrentSupplyOfVentilatorSupplies] [varchar](255) NULL,

[OneWeekSupplyOfVentilatorSupplies] [varchar](255) NULL,

[TotalResidentConfirmedCOVID19CasesPer1000Residents] [varchar](255) NULL,

[TotalResidentCOVID19DeathsPer1000Residents] [varchar](255) NULL,

[TotalResidentsCOVID19DeathsAsAPercentageOfConfirmedCOVID19Cases] [varchar](255) NULL,

[County] [varchar](255) NULL,

[ThreeOrMoreConfirmedCOVID19CasesThisWeek] [varchar](255) NULL,

[InitialConfirmedCOVID19CaseThisWeek] [varchar](255) NULL,

[Geolocation] [varchar](255) NULL

) ON [PRIMARY]

GO

Screenshot of Creating a table in SQL Server:

Graphical user interface, application

Description automatically generated

Input command:

bcp COVID19NursingHome in COVID-19\_Nursing\_Home\_Dataset.csv -S DESKTOP-S7MQMV4\SS2019 -d HW05 -T -f COVID-19\_Nursing\_Home\_Dataset.fmt

Screenshots of working bcp commands in command prompt:

![Text

Description automatically generated]()

![Text

Description automatically generated]()

Screenshot to show data is copied into SQL Server table:

Graphical user interface, application, table

Description automatically generated

Screenshot of format file (COVID-19\_Nursing\_Home\_Dataset.fmt):

![Table

Description automatically generated]()

Ques4) Analyze the data

This view displays the top 10 counties with most providers that has passed the Quality Assurance Checks

A picture containing graphical user interface, text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Map

Description automatically generated

Ques5) If we get file again next month it would have new data on it the best method to load data from this file would be through incremental load mechanism. In this selective movement of data from one system to another. An incremental load pattern will attempt to identify the data that was created or modified since the last time the load process ran. This differs from the conventional full data load, which copies the entire set of data from a given source. The selectivity of the incremental design usually reduces the system overhead required for the ETL process. The selection of data to move is often temporal, based on when the data was created or most recently updated. Here we can load data Using an upsert operation, such as a merge statement. An upsert (update + insert) will handle both the new and changed data in a single step. It does this by matching the unique key column(s) and comparing the columns you want to check for changes.

Ques6) Here, a test file is generated with the new values it is then passed through a lookup function to look for the new Ids and merge it into the database.

Graphical user interface, application

Description automatically generated

Ques7) There are various ways in which a test data file can be generated. In SQL itself, we can either write a SQL DDL to build a file or we can write hard code values in excel as well. To build a sample copy of the data, there are different algorithms that can produce random numbers and variables. We make sure that the test data type corresponds to our original data. By using a subset of our original data, we can also create a test data file.