

Chapter 5 *Advanced Access* Table of Contents

Advanced Access	1
ESE and Local Data.....	2
The Data Warehouse Model	2
Figure 1 – Local Data Data Marts (Grades and Staff Schedule)	3
Figure 2 – Student Dimension Records	4
Table 1 – Behavior of Records Loaded into Student Dim Table	4
Table 2 – Fact Dependencies	5
Step 1: Data Review and Discovery	6
Step 2: Data Extraction, Transforming, and Cleansing	6
Extract File Format Requirements.....	6
Required (Mandatory) Fields.....	7
Table 3 – File Extract Field Information	7
Transforming and Cleansing Data.....	8
Formatting Files for Upload.....	9
Figure 3 – District Data Upload and Staging	9
Step 3: Loading Data Extracts	10
The Staging Process.....	10
Figure 4 – Extract Validation Reports available in staging for username vtester3	11
Figure 5 – Extract Validation Report for Staff Schedule Fact	11
Step 4: Warehouse Testing and Validation	12
Figure 6 – Warehouse Table Validation Reports available in production	13
Figure 7 – Warehouse Table Validation Report for Course Dimension	13
Local Assessments	14
Authoring Reports Using Local Data	14

Advanced Access

This chapter is intended for data administrators responsible for uploading local data to the EDW. If you are the district’s data administrator, you need to have the *Data Warehouse File Exchange* function in **Directory Administration** assigned to you. This security role enables access to the Warehouse File Exchange drop box located on the Security Portal at Drop Box Central. You will also need the *DW – (210) District User* or the *DW – (209) District Report Author* security role to review the staging reports and accept into the warehouse the data you have uploaded.

District technical staff responsible for producing and loading data extracts need to understand the specific format and procedures that are required. The Department is available to support districts in their efforts, but it is important that you read this chapter thoroughly and understand the process prior to creating and loading extract files. Appendices D and E, *District Extract Guides* and the accompanying *Release Notes* provide the specifics about formatting and naming these files. In addition, you should expect to devote considerable time and resources to producing the initial loads and be aware of all of the resources available to answer your questions.

The use and reporting on local data is the responsibility of the district. If your district has advanced to this level of access, it is likely that you already have a report author assigned in your district. The Department has some reports available for displaying local data but if additional reports are needed, you should look to your local report author to develop these. (Please see Chapter 4, *Intermediate Access*, for more information on creating local reports.)

ESE and Local Data

Local data uploaded by a district cannot be viewed by other districts and is accessible only to those district staff assigned data warehouse roles in **Directory Administration**. The state's access to local data is limited to such rare cases as the required review of a district's files, in which case the district would be notified prior to such a review.

The Department will not monitor the content or quality of district-loaded data nor will it use local data for analysis. Each district may use fields differently, and therefore any analyses using local data from multiple districts may not be valid.

The data warehouse is not a data collection system, but rather a historical data archiving, analysis, and reporting system. The data loaded into the data warehouse from districts' local SIS systems must be accurate and complete. If the data collection system data is flawed, the reports generated from warehouse data will be as well.

The Data Warehouse Model

Before creating data extracts, it's important to understand the way data is stored in the data warehouse. There are two types of tables in the data warehouse, *fact* tables and *dimension* tables, and each extract file you generate is loaded into one or the other of these types of tables.

Fact tables contain quantitative measures that are associated with discrete occurrences of specific events. Supporting fact tables are the dimension tables. Dimension tables primarily contain qualitative *attributes* of a person, place, or thing.

Fact tables can sometimes contain *attributes*. Attributes stored in fact tables are non-quantifiable pieces of data that are context-dependent on the measure with which it is associated within the fact table. This data would not easily fit into a dimension table. For example:

- MCAS Performance Level qualifies MCAS Score in Test fact table
- Academic Period (semester/term/grading period) qualifies Grade in Grade fact table
- Class Section qualifies Course in Schedule fact table

Fact tables are *context-dependent* in that the data recorded in the fact tables require a connection to one or more dimension tables in order to be of value in reporting. It would not be useful to produce a list of MCAS scores from a fact table without knowing any details about the scores, such as the people or places associated with them. On the other hand, dimension tables are *context-independent* because reports can be written from the data without using data from the fact tables. For instance, the student dimension table contains all the demographic data pertaining to a student including gender and race-ethnicity, so districts can generate a report counting the students by gender or race-ethnicity.

Each fact table grouped with its associated dimension tables is called a *data mart*. There are five data marts in the EDW:

1. Grades Fact Table and the Student, Course, and Staff Dimension Tables
2. Grades/GPA Fact Table and the Student Dimension Table

3. Student Schedule Fact Table and the Student and Course Dimension Tables as well as the Staff Schedule Fact Table
4. Staff Schedule Fact Table and the Course and Staff Dimension Tables
5. Local Assessment Fact Table

The figure below illustrates two fact tables (grades and staff schedule) and their relationships to supporting dimension tables. (*Figure 1*) The tables are organized in a “star schema” model with the fact table in the center and dimension tables being shared by multiple data marts across the model. Note that the grades and staff schedule data marts connect to each other through the shared course and staff dimension tables.

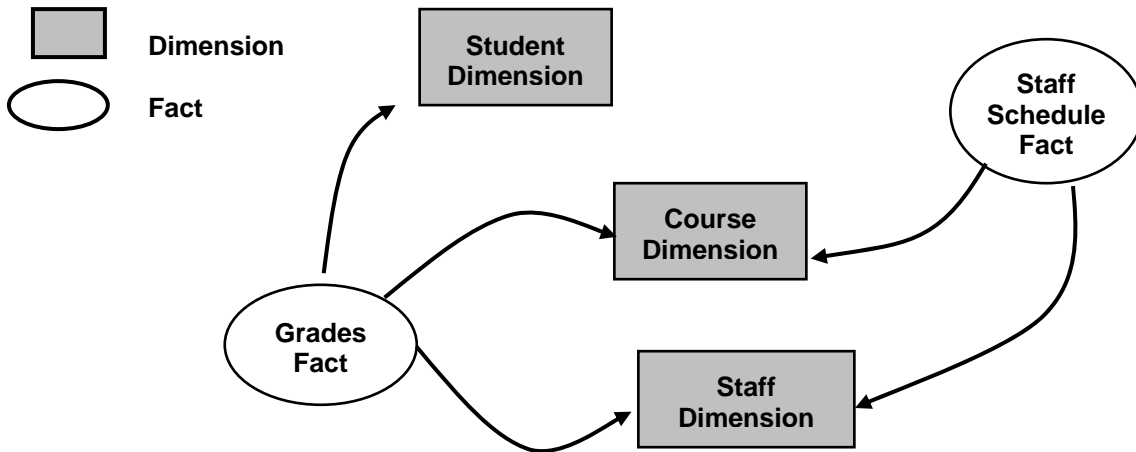


Figure 1 – Data Marts (Grades and Staff Schedule) for Local Data

The link that connects the fact and dimension tables is called the “Key.” Key fields must be consistent over time for the same person, place, or thing. The data warehouse uses primary keys and surrogate keys. Primary keys that link the dimension tables to their various fact tables are:

- SASID (Student Dim to Grades/GPA, Grades, and Student Schedule Facts)
- COURSE_KEY (Course Dim to Grades, Student Schedule, and Staff Schedule Facts)
- STAFF_KEY (Staff Dim to Grades and Staff Schedule Fact)

Key fields link records to other records within tables and across tables. Primary, or business, keys are uploaded with the record. These fields are the unique identifiers for the subject of the record. For instance, the SASID is the primary key in the student dimension table and the MEPID is the primary key in the staff dimension table. Similarly, course records have a unique course ID for each course.

When dimension data is loaded, new dimension records are compared to existing dimension records based on the primary keys, and the records are evaluated for changes in data. Some changes in data are considered worth tracking while others are not. Fields with historically tracked changes are called slowly changing dimensions (SCD). Incoming records are treated differently based on changes in data and whether or not these changes occur in SCD fields. In some cases, new records are added, in other cases, existing records are updated, and in some cases, no action is taken at all. The table below

shows the actions that are taken based on the comparisons of newly loaded records to existing records. (Table 1)

Table 1 – Behavior of Records Loaded into Student Dim Table

IF:	THEN:
SASID does not exist in table	New “Current” record is added
SASID exists, but new data is same as existing data	No action is taken and new record is completely ignored
SASID exists, and new data has differences in non-historically tracked attributes (non-SCD fields)	Existing “Current” record is updated
SASID exists, and new data has differences in historically tracked attributes (SCD fields)	Existing “Current” record is changed to “Historical,” and new “Current” record is added
SASID exists, and new data has differences in both non-historically tracked attributes and historically tracked attributes	Existing “Current” record is changed to “Historical,” and new “Current” record is added

“Current” and “historical” records are determined by the *current* attribute. The current attribute is a system-generated field that is used to identify the current record. In figure 2 below (an excerpt from a student dimension table), each student has one *current* record and zero or more rows of *historical* records. Other fields such as the effective start date, the effective end date, and the surrogate key are also fields that are added to the record by the system when you upload a file.

SASID	F Name	M Name	L Name	School	Eff Start Date	Eff End Date	Current?	Surrogate Key
0001	John	J	Smith	Cherry Lane	8/5/2002	9/13/2003	N	151266
0002	Alice		Jones	Cherry Lane	8/5/2002		Y	151267
0003	Paula	P	Pupil	Cherry Lane	8/5/2002	9/13/2003	N	151268
0001	John	J	Smith	Ambrose Bierce	9/13/2003		Y	187690
0003	Paula		Pupil	Ambrose Bierce	9/13/2003		Y	187691

Figure 2 – Student Dimension Records

In figure 2, new dimension records were added to the student table for students John Smith and Paula Pupil because changes were found in the school field and the school field is an SCD. The middle name, however, is not an SCD. If the data uploaded on 9/13/2003 had not contained new data in a slowly changing dimension field, Paula’s middle initial would have been overwritten in the existing record (with the effective start date of 8/5/2002), no new records would have been added, and there would be no historical data to show that Paula Pupil’s middle initial was once stored as P.

Effective date/time stamps (“Eff Start Date” and “Eff End Date” in figure 2 above) allow for the expiration of student record versions. As fact records, such as grades and schedules are loaded, they locate the correct dimension records via the primary key field. Since many records in a dimension table can contain the same key field and only one is current, the time stamps are used to determine which record containing the primary key the fact record should link to. When the correct “current” dimension record is located, its

unique surrogate key is stored in the fact table forever linking this fact record with this dimension record. This is especially valuable for seeing trends and the effectiveness of applied methodologies over time. For example, the system could identify that a student once belonging to the Free and Reduced Meal Program is no longer part of the program.

Fact records load differently than dimension records. New fact records are added into the warehouse, without making any comparisons to existing fact records. All incoming fact records are treated as new. Mandatory and recommended fields, such as the primary keys, are used to link fact records to dimension records. These links are made at the time of loading, so if a dimension record does not exist to link to, the fact is linked to “UNKNOWN” dimension records. Table 2 below shows the fact table dependencies and the dimension tables that must be loaded first.

Table 2 – Fact Dependencies

<i>If you plan to load these Facts...</i>	<i>You must first load these Dimensions:</i>		
	<i>Student</i>	<i>Staff</i>	<i>Course</i>
<i>Grades/GPA</i>	X		
<i>Grades</i>	X	X	X
<i>Staff Schedule</i>		X	X
<i>Student Schedule</i>	X	Staff Schedule Fact and its 2 related Dimensions (Staff and Course)	

If current data is to remain current and changes to historically tracked data is to be accurately recorded, dimension data should always be loaded in chronological order with the oldest dimension records loaded first.

Student schedule fact is dependent on staff schedule fact and all its related dimensions, as well as the student dimension.

Districts that plan to load historical data must adhere to the following rules:

- Load dimension records in order by date (older records are ignored when newer ones exist).
- Load fact records after the related dimension records. Fact records will be linked to the dimension records with the appropriate effective dates. For example a GPA fact record with a GPA_DATE of June 30, 2003 will link to the Student dimension record that was effective at that time.
- If dates on fact records are omitted, then they will link to the current (latest) dimension record. This can lead to incorrect analyses on the data.

Note: If your district plans to load historical data, it’s essential that you provide snapshots of student and staff dimensions that correspond to the related facts. For instance, it does not make sense to load grades or assessments from 2002 if you do not have student dimension data from 2002. If you are loading historical snapshots, it’s essential that you load these snapshots in the correct order. Failure to provide consistent historical snapshots can lead to erroneous analyses.

Step 1: Data Review and Discovery

Now that you have a clearer understanding of the data warehouse model, you can get together with your district planning team to decide which dimensions and fields will be loaded into the data warehouse.

- Review Appendix D, *District Extract Guides* and decide which of the recommended and optional data elements are valuable to your district's improvement plans.
- Locate and review district data sources and determine if the data is available in the required format.
- Select dimensions and fields to load into the warehouse.
- Determine how far back to go in loading historical data.
- Plan for future uploads/refreshes.

At this point, it's good to make a reality check on your initial district plan and your district's data readiness. As a team, decide which type of data and how much historical data will be loaded, and update your district warehouse plan accordingly.

Step 2: Data Extraction, Transforming, and Cleansing

Extract File Format Requirements

Prior to uploading data, districts must export the data from their local systems into files that meet the following requirements:

- Tab Separated Values
- Column headers as specified in extract guides (There should always be a row for the column headers. If you do not include the headers row, your first line of data will be interpreted as the header row and will not be imported as data.)
- Double-quoted string fields
- 10-character date format: `mm/dd/yyyy`
- School year fields should be populated with the ending calendar year of the academic year. For example, school year 2006–2007 should be specified as 2007.
- File name format: `<distcode>_extract_name.txt` where `<distcode>` is the 4-digit district code.

Examples of valid filenames include:

```
0210_COURSE.TXT
0674_GRADES.TXT
0035_STUDENT_SCHED.TXT
```

A valid filename is required to trigger the warehouse data loading process.

- Macintosh text files are NOT recognized. (Although the student claiming process has been modified to accept Macintosh formatted files, the ETL processes for uploading the local data files defined in the *District Extract Guides* have not.)

Required (Mandatory) Fields

In general, the fields that are required in the data extracts are restricted to the record identifiers (STU_ID, STAFF_KEY, COURSE_KEY, etc.). There are additional fields, such as the dimension keys, or date fields noted above, which are highly recommended, but not required. Each file has a maximum number of fields, listed as mandatory, recommended, and optional. The table below provides the total number and provides the mandatory and recommended field names. (Table 3) For a more complete list, including optional fields, data types, and descriptions of all the fields, please consult Appendix D, *District Extract Guides*.

Table 3 – File Extract Field Information

File Name (Table Type)	Total Possible Fields	Mandatory Fields	Recommended Fields
STUDENT (Dimension)	163	STU_ID (which is the SASID) STU_LOAD_DATE	STU_FULL_NAME STU_LAST_NAME STU_FIRST_NAME STU_MIDDLE_NAME STU_BIRTH_DT STU_CUR_SCHL SCHL_YEAR STU_EXTRACT_DATE
STAFF (Dimension)	27	STAFF_KEY STAFF_LOAD_DATE	STAFF_ID STAFF_FULL_NAME STAFF_LAST_NAME STAFF_FIRST_NAME STAFF_MIDDLE_NAME STAFF_ID_STATE (MEPID)
COURSE (Dimension)	21	COURSE_KEY COURSE_LOAD_DATE	SCHOOL_YEAR COURSE_ID COURSE_DESC COURSE_EXTRACT_DATE
GRADES (Fact)	39	STU_ID (SASID)	COURSE_KEY STAFF_KEY GRADE_DATE SCHOOL_YEAR GRADING_PERIOD GRADE_EXTRACT_DATE
GRADES_GPA (Fact)	13	STU_ID (SASID)	GPA_DATE GRADING_PERIOD SCHOOL_YEAR GPA_EXTRACT_DATE

File Name (Table Type)	Total Possible Fields	Mandatory Fields	Recommended Fields
STAFF_SCHED (Fact)	79	RECORD_KEY	COURSE_KEY STAFF_KEY SCHL_YEAR LOC_KEY SECTION_ID (corresponds to EPIMS field WA11 Class Section) TERM EXTRACT_DATE
STUDENT_SC HED (Fact)	26	RECORD_KEY	SCHOOL_YEAR STU_ID (SASID) LOC_KEY COURSE_KEY SECTION (corresponds to EPIMS field WA11 Class Section) TERM_CODE EXTRACT_DATE

Transforming and Cleansing Data

Data cleansing is an essential step in populating a data warehouse. Due to differences in storage conventions between source systems (SIS, Payroll, Finance, etc.) and the target data warehouse, data from district systems may not conform to requirements of the data warehouse and must be *transformed* to fit. As part of this process, the data should also be cleansed by removing all inaccurate, incomplete, or irrelevant data. Typical data cleansing tasks include:

- Matching records and assigning IDs.
- Checking that unique record IDs are present and accurate (District Code, School Code, SASID, MEPID, and other local identifiers).
- Deleting duplicate records.
- Segmenting columns (e.g., last name, first name, middle name).
- Reformatting the date fields.

As you have learned, extract files load specific fact and dimension tables and records in these files will be linked across tables based on the identifiers, or keys (STU_ID, STAFF_KEY, COURSE_KEY, etc.). The data warehouse uses these identifiers to join tables and display related data on reports. An essential aspect of data preparation is ensuring that record identifiers are correctly aligned across extract files, and that a set of files represents a consistent historical snapshot. For instance, to load 2006 grades data, you must first load student dimension, course dimension, and staff dimension data for

2006. (If you prefer, you could load all of your dimension data for multiple years before any of your fact data, as long as you load all of the dimension data in chronological order.) You should check all of the dimension data to ensure that the STU_IDs, COURSE_KEYS, and STAFF_KEYS are identical between the dimensions and the grade fact records.

Data cleansing is essential to ensuring that warehouse data analysis is reliable and accurate.

Formatting Files for Upload

The ESE has been working with software vendors to enhance their human resource and student information applications to be able to export files that are ready for upload to the EDW. Some vendors are working to comply. (For information on whether your software vendors are EDW compliant, please email the Data Warehouse Team at datawarehouse@doe.mass.edu.)

Determine which fields from your local data systems should map to the fields in each of the DIM and FACT files your district plans to upload. (It is not necessary to upload all of the FACT files, especially the Assessment FACTs, during the first load attempt.)

Remember that in generating your files, they must conform to the requirements identified in the *Guides* and that all of the columns must be present even if they are not being used by your district.

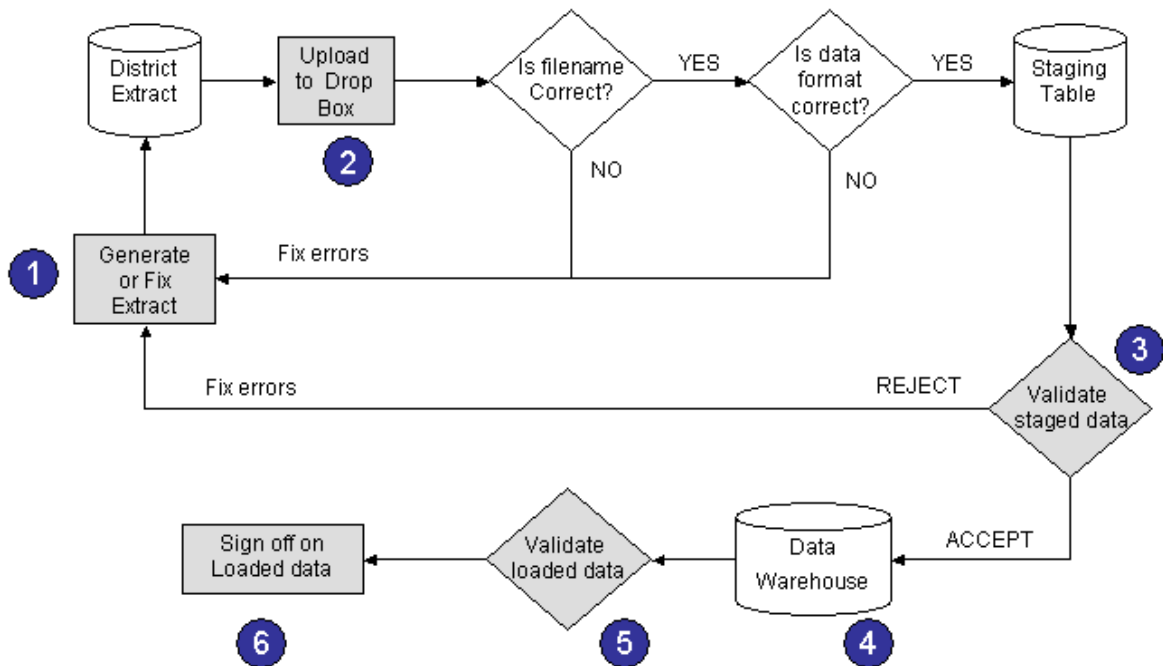


Figure 3 – District Data Upload and Staging

Step 3: Loading Data Extracts

Once the district data has been extracted, transformed, and cleansed, the extract file is ready to be uploaded to the warehouse. If you have been granted the *Data Warehouse File Exchange* security role, you can upload files to the warehouse as follows:

1. Log on to Drop Box Central on the ESE Security Portal and select the Warehouse File Exchange drop box.
2. Select your district from the list of organizations and click the Next button.
3. Click the Browse button to find the file you wish to upload.
4. Click the Upload File button to upload the file.

Receipt of a properly named extract file triggers the warehouse data staging process. (Figure 7) You will be notified by email of the results of this process. If the records fail to load into the staging tables, a file containing the errors will be attached to the email.

The most frequent causes of extract upload failures are:

- Incorrect file name.
- Dates are not in the required format: **mm/dd/yyyy**.
- String fields have an extra set of double quotes.
- Field length exceeds the maximum specified in the *Guides*.
- Incorrect number of columns, that is, the file has more or fewer columns than specified in the *Guides*.

Opening the extract file in *Excel* or *Notepad* and inspecting it for these errors can help identify the problem. Note: *Excel* may conceal some formatting errors that will be visible in *Notepad* or another text editor.

Tip: To view Tab characters, open the extract file using MS-Word, and then select Tools | Options | View | Formatting marks | All.

Once you have corrected any data errors, upload the file again to trigger the staging process.

The Staging Process

If the upload and staging process is successful, an extract validation report is available. The district must review the report and accept or reject the staged records. If the records are accepted, they are loaded into the data warehouse. If the records are rejected, the extract is discarded and the district must correct any errors and re-upload the file.

Note: Districts should not move on to the next upload before ensuring that the first file has been uploaded and validated.

To validate staged records:

1. Log in to the data warehouse.
2. Click on Public Folders > District Data Upload Validation.
3. Click on Extract File Validation (Staging).

A list of extract validation reports displays.

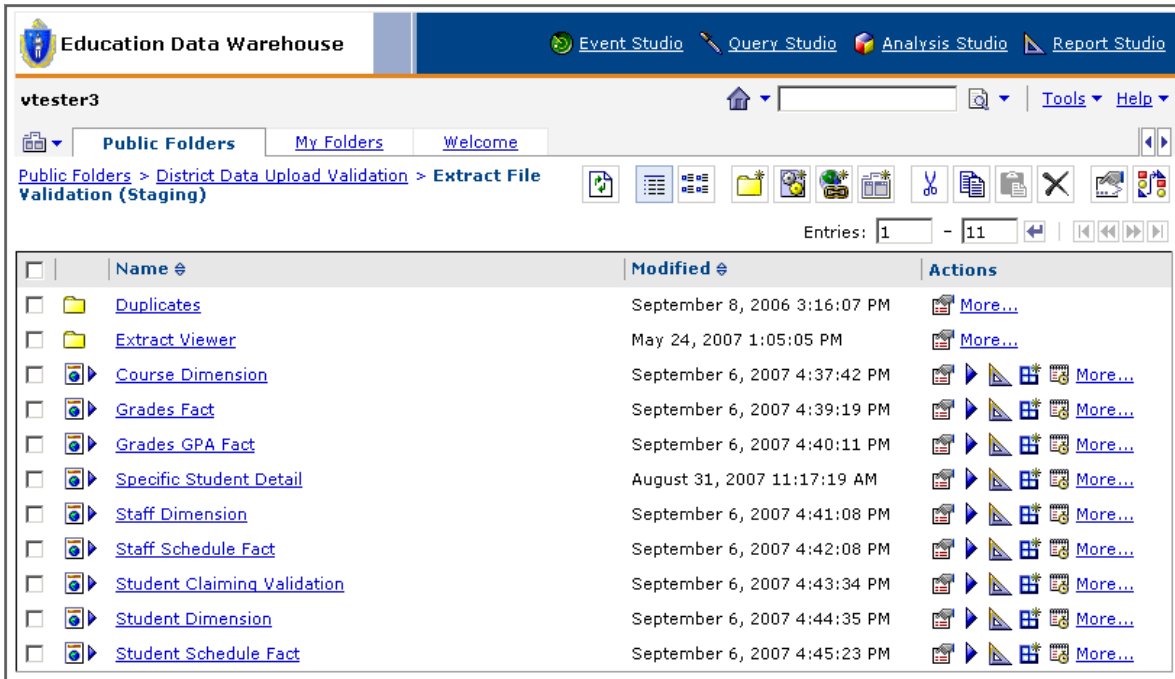


Figure 4 – Extract Validation Reports available in staging for username vtester3

- Run the report(s) corresponding to the extract file(s) you have uploaded. The report includes a summary of the staged records for your district.

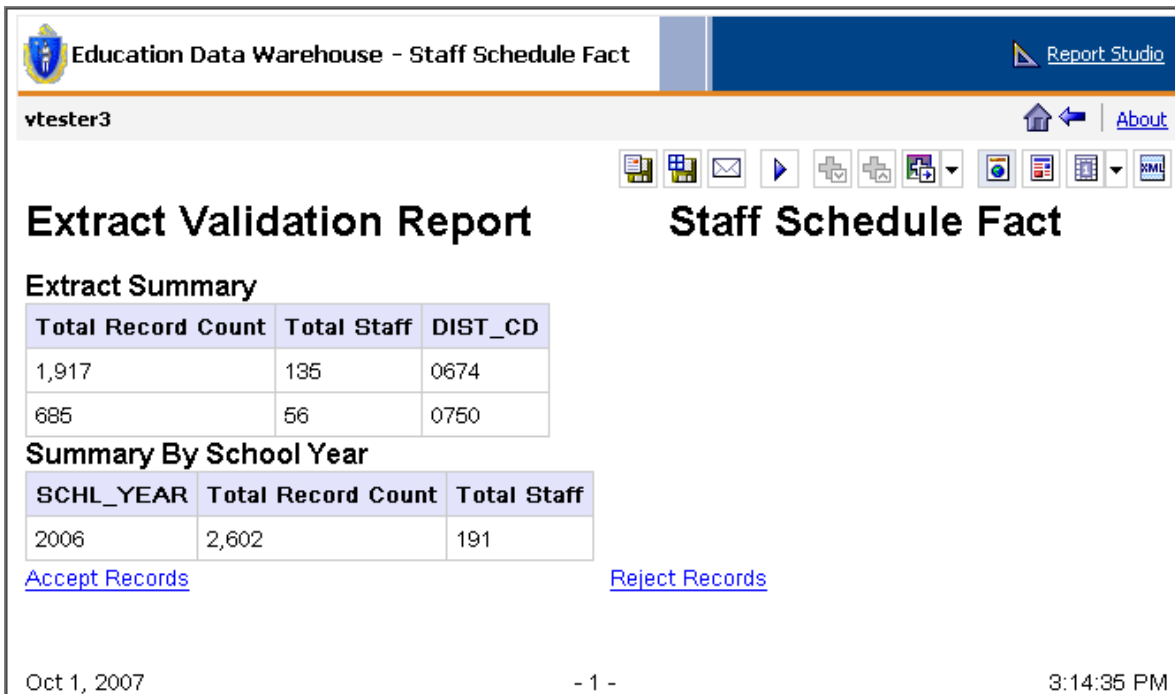


Figure 5 – Extract Validation Report for Staff Schedule Fact

- Review the summary data on the extract validation report, and then select one of the following actions:

- a. Click on **Accept Records** to load the staged extract into the data warehouse.
- b. Click on **Reject Records** to discard the staged extract. Correct any errors in the extract file, and then re-upload the file.
- c. Click on **Return** to exit the report without taking any action. The extract will remain in the staging table until you are ready to take action.

Important Steps in the Data Load Process:

1. Begin with the Earliest Year of Data.
2. Upload Dimensions: Course Dim, Staff Dim, Student Dim
3. Accept Dimensions in any order but accept all the Dimensions before their dependent Facts.
4. Upload Facts: Grades Fact, GPA Fact, Staff Sched. Fact, Student Sched. Fact.
5. Accept Facts: *Staff Sched. Fact must be accepted before Student Sched. Fact.*
6. Review this first year's data in the Table Validation Reports in Production and the Sample Local Data Reports.

Only *when you are satisfied that Facts link to the correct Dimension records* should you begin the process for the next year's data.

Step 4: Warehouse Testing and Validation

Once a district's data extract is loaded into the warehouse, the district is responsible for validating the accuracy and completeness of the loaded data. Prior to rolling out the warehouse across the district, the planning team should perform additional testing and validation of warehouse data.

Districts should perform testing and validation of warehouse reports on their state-loaded data (MCAS and SIMS) as well, not only on their uploaded local data. To submit your feedback and change requests, please email the data warehouse team at datawarehouse@doe.mass.edu.

Validating District-Loaded Data

To validate data loaded into the warehouse tables:

1. Log in to the data warehouse.
2. Click on **Public Folders > District Data Upload Validation**.
3. Click on **Table Load Validation (Production)**.
4. A list of warehouse table validation reports displays.

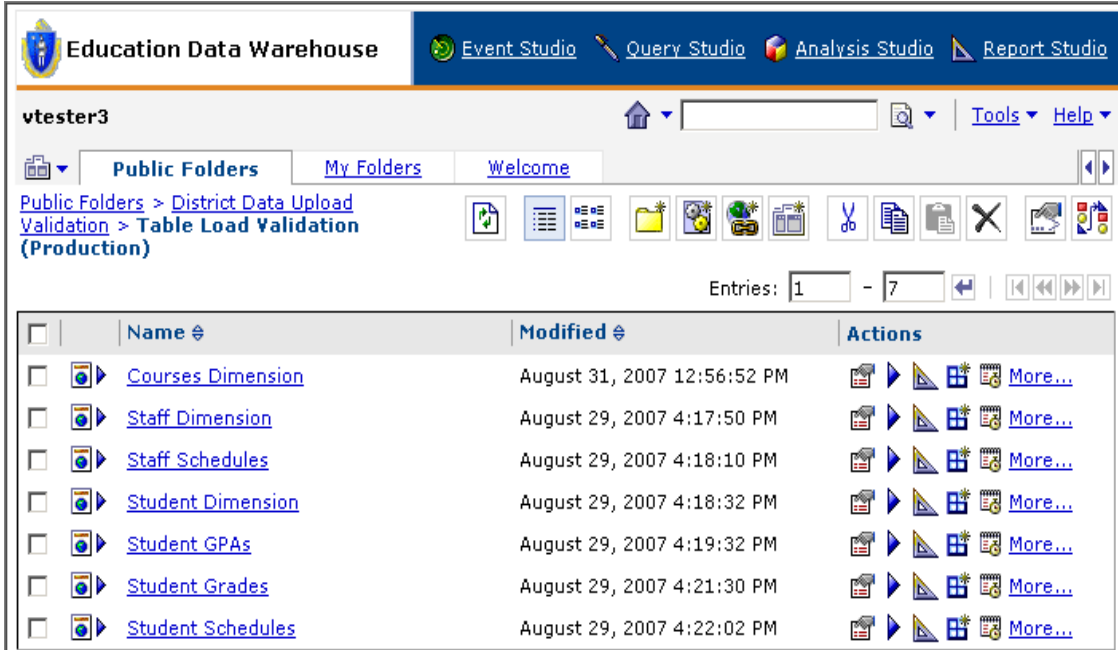


Figure 6 – Warehouse Table Validation Reports available in production

- Run the report(s) corresponding to the data you uploaded, and validate that the data was loaded correctly.

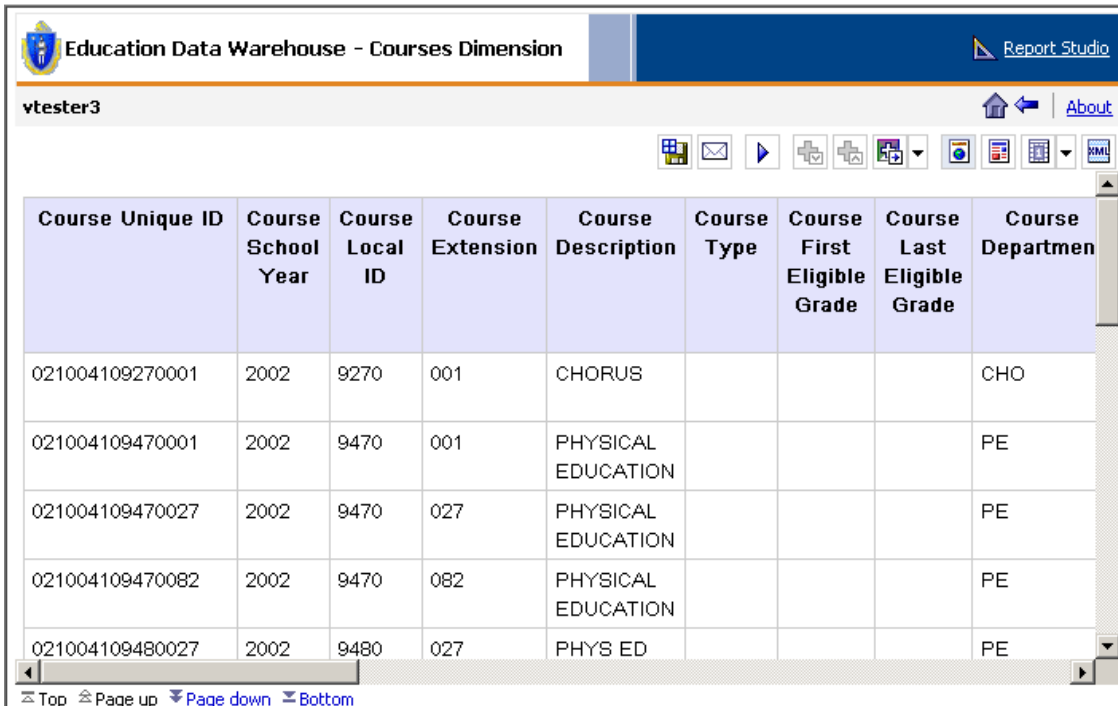


Figure 7 – Warehouse Table Validation Report for Course Dimension

Compare the loaded warehouse data to data in your source systems. Spot check records for accuracy and completeness:

- Validate that you are able to see data for all the districts you are authorized to view.
- All uploaded attributes are populated in the correct fields, for example, student's attributes such as school, district, birth date, etc.
- Data is formatted as expected (e.g., date and number formats are correct).
- Validate that leading zeroes exist as required in character fields that contain only numeric data (e.g., district code, school code).
- When submitting data, districts supply separate extract files for each warehouse table that are related to one another based on identifiers (STU_ID, STAFF_KEY, COURSE_KEY, etc.). The data warehouse uses these identifiers to join tables and display related data on reports. Please validate that records are being joined correctly. For example, validate that a teacher is correctly associated with a class and students in the class.

Note: If you are not satisfied that your district data was successfully loaded and integrated into the warehouse, send an email to the data warehouse team at datawarehouse@doe.mass.edu requesting deletion of the problem records from the warehouse. Once the records are deleted, fix any errors in your SIS extracts and resubmit the extracts.

Local Assessments

A separate fact table has been designed for the storage of local assessment results and several extract files have been created for specific assessments (DIBELS and GRADE) as well as a generic extract file which can be adapted to accommodate all others. A mapping of the Iowa Test of Basic Skills (ITBS) to the generic assessment file. The Department is working with pilot districts to map additional high priority assessments to the warehouse Local Assessment fact table such as Galileo and MAP. These file formats are available in Appendix D, *District Extract Guides*.

There is no architectural limit on the number or type of assessments a district can store. However, there is some effort involved in mapping and transforming the vendor-supplied assessment results to the electronic format needed for loading into the data warehouse. Once the assessment data is loaded, additional effort is required to define meaningful reports based on the generic assessment tables.

Authoring Reports Using Local Data

Although some sample reports exist in the data warehouse that use local data as well as local assessments, the Department has provided these for districts to adapt to their own use. The Department is not requesting feedback on these reports and does not include the updating of these reports in its list of priorities for maintaining and improving the Education Data Warehouse. Districts should hire or train a local report writer for improving these reports or creating others. Chapter 4, *Intermediate Access* addresses the issues that district report writers will need to know.